SOME THEORETICAL ASPECTS OF ECONOMETRIC INERENCE
WITH HETEROSKEDASTIC MODELS

By

Hernando Carlos Luis Sabau García

A thesis submitted for the degree of
Doctor of Philosophy
of
The Australian National University

April, 1988.
DECLARATION

The contents of this thesis are my own work,
except where otherwise indicated.

Hernando C. L. Sabau
ACKNOWLEDGEMENTS

I am deeply indebted to my Supervisors, Dr. Trevor S. Breusch and Dr. Anthony D. Hall, for their comments, guidance and constructive criticism. Their help was not restricted to academic matters and their doors were always open for advice and discussion.

Dr. Adrian R. Pagan introduced me to the topic of the Thesis and put me in contact with a rich and interesting literature. To him I want to express my most sincere gratitude. The many references to his work in this Thesis show only a very small fraction of my intellectual debt to him.

Many others in the Department of Statistics, Faculty of Economics and Commerce, ANU, contributed to a stimulating environment. In particular, I would like to thank Professor R. Deane Terrell without whose support my doctoral studies would not have been possible, and Dr. John J. Beggs who always offered constructive criticism and was enthusiastic in undertaking joint work with me. My thanks also go to Mrs. Danièle Elliot who typed the manuscript with great skill and care.

I gratefully acknowledge the financial support of an ANU Scholarship during the period of my study, and an educative credit from Banco de México. I also want to express my gratitude for the institutional support of the Centro de Investigación y Docencia Económicas, A. C. (CIDE).

To my daughters Lucía and Ana I thank for putting up with a not always understanding father during this period, and to my Mother I thank especially for the time she spent with us in Canberra.

This Thesis is dedicated to Mila, with all my love, for her continuous encouragement, enthusiasm and unconditional support.
ABSTRACT

This Thesis is concerned with econometric inference in parametric heteroskedastic models. Each moment of the conditional distribution can be seen as a source of information which provides an estimating equation for the parameter vector. Different issues arise in the different moments concerning the identifiability of parameters, the observability of the dependent variable of the estimating equation, and the positivity restrictions implicit in even order moments. Estimators of the identifiable functions of the parameter vector are obtained from orthogonality conditions in each moment. Under symmetry of the distribution, the sources of information corresponding to the first two conditional moments are independent, at least asymptotically, and the information about common parameters is combined in estimation by constructing a matrix weighted average. Estimation procedures under normality are viewed in a maximum likelihood framework, and generalized method of moments estimation provides the setup for the analysis of more general distributions. The separation of the information into its moment source constitutes a basic element for diagnostic testing of the model. The implications of different forms of misspecification are analyzed and robustness properties are established for some leading cases, especially the ARCH class of models. A general framework is presented for diagnostic testing of heteroskedastic models, which includes tests of the coherency of the information contributed by the two moments, a family of 'consistency tests' which concentrates on the assessment of the first two moments, and a family of 'efficiency tests' which concentrates on checking the specification of moments of order three and higher. The consistency and efficiency tests may be constructed without using information external to the model and thus may be reported with standard computer output, but these families also include many LM tests against specific departures by suitable choice of the test parameters. Tests for autocorrelation, dynamics, parameter stability, different types of
exogeneity, and normality, are analyzed in particular. The estimation and
diagnostic testing framework is extended to the inclusion of latent variables in
the conditional mean, such as parametric risk measures and varying
coefficients, and also to a multivariate setting. Finally, the problem of
extracting information from higher order moments is considered by looking at
the information that each moment contributes in addition to what has already
been contributed by the lower order moments. Information is extracted from
orthogonality conditions and a sequential strategy proposed which analyzes the
efficiency gains and the coherency of the available information with the new
information obtained from incorporating an additional moment into the model.
TABLE OF CONTENTS

DECLARATION (ii)

ACKNOWLEDGEMENTS (iii)

ABSTRACT (iv)

CHAPTER 1. INTRODUCTION 1

CHAPTER 2. THE BASIC MODEL 11
   § 2.1. Basic model and notation 11
   § 2.2. Assumptions 18
   § 2.3. Some special cases 24
      § 2.3.1. Simple heteroskedasticity 24
      § 2.3.2. Amemiya model 26
      § 2.3.3. The Poisson and Poisson-type models 27
      § 2.3.4. The ARCH class of models 29
   § 2.4 Objective and design of Monte Carlo experiments 34

CHAPTER 3. ESTIMATION AND LIKELIHOOD FACTORIZATION 39
   § 3.1. Extracting information from the conditional mean 40
      § 3.1.1. The parametric approach 42
      § 3.1.2. A note on semi-parametric approaches 44
   § 3.2. Extracting information from the conditional variance 47
      § 3.2.1. Two-stage estimators for $\alpha$ 53
      § 3.2.2. Estimating the identifiable functions in the variance equation 56
   § 3.3. Combining information from the two conditional moments 59
      § 3.3.1. Combining information from orthogonality conditions 60
      § 3.3.2. Likelihood factorization: the case of normality 65
   § 3.4. Analyzing the contributions to efficiency 71
   § 3.5. Some comments on Monte Carlo evidence 74
CHAPTER 4. THE ROBUSTNESS OF THE QUASI-MAXIMUM LIKELIHOOD ESTIMATOR

§ 4.1. Some conditions on the likelihood function
§ 4.2. Specification error in the conditional mean
§ 4.3. Specification error in the conditional variance
§ 4.4. Specification error in higher order moments
§ 4.5. Some comments on Monte Carlo evidence
  § 4.5.1. Specification error in the conditional mean
  § 4.5.2. Specification error in the conditional variance

CHAPTER 5. GENERAL SPECIFICATION TESTS

§ 5.1. Coherency tests
§ 5.2. Consistency tests
  § 5.2.1. Consistency tests and other tests in the literature
  § 5.2.2. The distribution of consistency test-statistics
  § 5.2.3. Some power considerations
  § 5.2.4. A simple example
§ 5.3. Efficiency tests for higher order moments
§ 5.4. Comments on some Monte Carlo evidence
  § 5.4.1. Consistency and coherency tests
  § 5.4.2. Efficiency tests

CHAPTER 6. TESTING THE MODEL AGAINST SPECIFIC ALTERNATIVES

§ 6.1. The LM test for variable addition
§ 6.2. LM tests for autocorrelation and dynamics
  § 6.2.1. Autocorrelation in the conditional mean
  § 6.2.2. Dynamics in the conditional mean
  § 6.2.3. Autocorrelation in the variance equation
  § 6.2.4. Dynamics in the conditional variance
§ 6.3. Parameter stability
  § 6.3.1. Structural break
  § 6.3.2. Prediction error tests
§ 6.4. LM tests for weak and strong exogeneity
§ 6.5. Testing normality
§ 6.6. Some comments on the Monte Carlo evidence
  § 6.6.1. LM tests for autocorrelation
  § 6.6.2. LM tests for normality
CHAPTER 7. RISK MODELS

§ 7.1. Parametric risk models and testing for risk effects 190
§ 7.2. Estimation and diagnostic testing of y-risk models 195
  § 7.2.1. Estimation of y-risk models 195
  § 7.2.2. Consistency and efficiency tests for y-risk models 200
§ 7.3. Estimation and diagnostic testing of x-risk models 202
  § 7.3.1. Estimation of x-risk models 202
  § 7.3.2. Consistency and efficiency tests for x-risk models 206

CHAPTER 8. VARYING COEFFICIENT HETEROSKEDASTIC MODELS 215

§ 8.1. Varying coefficients and heteroskedastic models 216
§ 8.2. Further tests for parameter stability and diagnostic tests for varying coefficient heteroskedastic models 224
§ 8.3. Testing for superexogeneity and invariance 227

CHAPTER 9. MULTIVARIATE HETEROSKEDASTIC MODELS 239

§ 9.1. Notation and assumptions 240
§ 9.2. Multivariate ARCH and some other multivariate heteroskedastic models 243
§ 9.3. Estimation and likelihood factorization 247
§ 9.4. A note on specification error 254
§ 9.5. Consistency tests 256
§ 9.6. Efficiency tests 260
  Appendix to Chapter 9 265

CHAPTER 10. AN EXPLORATION INTO HIGHER ORDER MOMENTS 272

§ 10.1. Estimating the first two moments in the presence of asymmetry 273
§ 10.2. Extracting information from higher order moments 280
§ 10.3. Diagnostic testing 294

CHAPTER 11. CONCLUSIONS 298

REFERENCES 307
ERRATA FOR THE Ph.D. THESIS:

SOME THEORETICAL ASPECTS OF ECONOMETRIC INFERENCE
WITH HETEROSKEDASTIC MODELS,

By Hernando C.L. Sabau

Page Line

19 21 Says: ... of the variance equation. ...
Should say: ... of the variance equation. Care should be exercised when combining lagged dependent variables and mixing conditions (Athreya and Pantula [1986]). ...

20 2 Says: ... 2 on \( \theta \), uniformly in \( t \). Further, \( \mu_t \) and \( h_t \) and their first two derivatives with respect to \( \theta \) are bounded from above and \( h_t \) is bounded away from zero almost everywhere in \( \Theta \), uniformly in \( t \).
Should say: ... 2 on \( \theta \), and \( h_t \) is bounded away from zero almost everywhere in \( \Theta \).

20 7 Says: ... Given this smoothness ... in the classical linear model). ...
Should say: ... ...

20 16 Says: introduce
Should say: introduce a condition of uniform integrability (White and Domowitz [1984], page 147):

21 13 Says: ... of \( \theta_0 \).
Should say: ... of \( \theta_0 \), which is identifiably unique.

21 last Says: ... \( \psi(\lambda)' A_T \psi(\lambda) \)
Should say: ... \( \psi(\lambda)' A_T^{-1} \psi(\lambda) \)

22 18 Says: ... (Q0) - (Q7) are sufficient for (A1) - (A9).
Should say: ... (Q0) - (Q7) are sufficient for (A1) - (A9), with the addition that the continuous differentiability in (Q3) be uniformly in \( t \). Furthermore, the applicability of Theorems 3.1 and 3.2 of White and Domowitz [1984]
remains valid without this continuous differentiability condition.

22 19 Says: ... (Q2) and (Q3) ⇒ (A1) and (A5) ...  
Should say: ... (Q2) and (Q3) with the continuous differentiability  
being uniformly in t ⇒ (A1) and (A5) ...

22 24 Says: ... in (5) - (7).  
Should say: ... in (5) - (7). Now, (Q2) ⇒ A1 of Andrews [1987]  
(Andrews in what follows); (Q4) - (Q5) ⇒ A5 of Andrews; the mixing conditions in (Q1) ⇒ B1 of Andrews; and (Q4) - (Q6) ⇒ B2 of Andrews. Corollaries 1 and 2 of Andrews then ensure the  
applicability of his uniform Law of Large Numbers and guarantee the validity of Theorems 3.1 and 3.2 of White and Domowitz without continuity uniformly in t.

49 7 Says: and  
Should say: and, provided the expectations exist,

53 8 Says: ... Weiss [1984, 1986a]).  
Should say: ... Weiss [1984, 1986a], Davidian and Carroll [1987]).

60 25 Says: which evaluated at some root-T ...  
Should say: where for any set of matrices D_1,..., D_m we define the  
block diagonal matrix

$$D = \text{diag} \{ D_1, ..., D_m \} = \begin{pmatrix} D_1 & 0 & ... & 0 \\ 0 & D_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & D_m \end{pmatrix}.$$

The weighting matrix $A_T(\hat{\theta})$ evaluated at some  
root-T...

61 last Says: ... $V(\hat{\theta}_j) = \mathcal{E} \{ T^{-1} G' \Sigma^{-1} G \}.$  
Should say: ... $V(\hat{\theta}_j) = \mathcal{E} \{ T^{-1} G' \Sigma^{-1} G \}^{-1}.$

82 19 Says: ... limit $\theta^*$.  
Should say: ... limit $\theta^*$, which is identifiably unique.

Andrews, D.W.K. [1987], Consistency in nonlinear econometric models: a

Athreya, K.B. and Pantula, S.G. [1986], Mixing properties of Harris chains and

Page 310, between "DasGupta, S. and Perlman, M.D. [1974]" and

Davidian, M. and Carroll, R.J. [1987], Variance function estimation, *Journal of
the American Statistical Association* 82, 1079-1091.
CHAPTER 1

INTRODUCTION

The presence of heteroskedastic disturbances in regression models has been traditionally regarded as a problem. Modelling the mean alone has proved to be a difficult enough task and at the same time has permitted a large number of useful applications in many areas of economics. Applied econometricians have seldom have a priori information to specify the variance but have been well aware of the implications of ignoring heteroskedasticity, namely inefficiency in estimation and the risk of drawing incorrect inferences because the covariance matrix of the least squares estimator is incorrectly estimated by the usual formula. As a protection against these undesirable outcomes, testing for heteroskedasticity has become a well established routine in applied work, and a vast literature was generated during the sixties and seventies to produce alternative testing procedures. Some of the most relevant examples are Goldfeld and Quandt [1965], Glejser [1969], Harvey [1976], and Breusch and Pagan [1979].

More recently, White [1980b] pursued the ideas of Eicker [1967] and produced an estimate of the covariance matrix which is robust to heteroskedasticity, and this development relieved applied econometricians from the most serious of the two consequences cited above. But if the interest of the researcher goes beyond the estimation of the parameter vector in the mean of the distribution, heteroskedasticity has further implications. Forecast intervals are incorrect because using White's robust covariance matrix corrects only one component of the variance of prediction errors, namely that which arises from the imperfect knowledge of the parameters, but it does not
correct the component arising from the inherent randomness of the model which is changing with the variance itself. Thus any forecasting exercise is still jeopardized by the presence of heteroskedasticity. Besides, there may exist an explicit interest in studying the change in the variability of the dependent variable in response to changes in the independent variables and in analyzing the dynamics intrinsic in the variance, which may be substantially different from the dynamics in the mean of the process. We should be more concerned about the implications of policy actions on the variability of the main indicators of economic activity, and exercises of policy design such as those proposed by Tinbergen [1952] and Chow [1975,1981] might produce interesting results if the variances are considered amongst the targets. When analyzing economic policy, for example, one of the contentions of the rational expectations school has been that unpredicted policy changes will have a much stronger effect on the variance than on the mean of the process, mainly creating uncertainty.

Some areas of economic modelling, mainly those dealing with situations in which risk plays an important role in explaining economic behavior such as financial markets, have felt the need for a more constructive approach to heteroskedasticity. Models treating jointly the first two moments of the distribution have been constructed for a long time (Prais and Houthakker [1955]), but it has not been until recently with the appearance of Engle's [1982a] paper presenting the ARCH model that heteroskedasticity has been more systematically incorporated in applied work. The ARCH model has filled several gaps by providing a way to approximate many conditionally heteroskedastic patterns much in the same way that ARMA models can approximate the conditional mean of linear processes, by having theoretical plausibility and empirical success, and by generating lines of research which have produced a more complete body of inferential procedures when the heteroskedasticity is accounted for in the model.
The aim of this Thesis is to contribute to a constructive treatment of heteroskedasticity by analyzing some theoretical aspects of econometric inference with heteroskedastic models, constructing a coherent general framework for the estimation and testing of such models. Our view of heteroskedasticity is that it provides a second equation about the subject of study — and hence an additional source of information — which in many cases constitutes directly an additional set of observations to improve the efficiency of estimators of the parameters of the mean — usually the focus of interest of applied econometricians. The implication of this view is that we can clearly separate what information is being contributed to the process of inference by each of the two moments. Along the way we find that estimation procedures are not substantially complicated by taking the heteroskedasticity into account, with most of the results being capable of interpretation in terms of the generalized classical linear model, and that the better known diagnostic tests under homoskedasticity are easily extended to this situation. To fulfill our goal, the presentation of the material can be divided into two parts. The first part deals with the estimation and diagnostic testing of the univariate heteroskedastic model (Chapters 2 - 6), and in the second part this basic model is extended to more general settings (Chapters 7 - 10).

We start by introducing the notation and characterizing the model as a two-equation system composed of a mean equation and a variance equation with (possibly) cross-equation restrictions in Chapter 2. This chapter is essentially a reference chapter for the rest of the Thesis. A basic set of assumptions is presented which conforms to the hypotheses underlying general method of moments estimation (Hansen [1982]) and nonlinear estimation with dependent observations (White and Domowitz [1984]), and this enables the use of very powerful results in econometric inference in the rest of the Thesis. The most relevant special cases of heteroskedastic models in the literature are reviewed, and their characteristics and regularity conditions
discussed. These special cases include the simple heteroskedasticity model, the Amemiya model (Amemiya [1973]), the Poisson model and continuous models with a Poisson structure (Cameron and Trivedi [1985, 1986]), and the ARCH class of models presented by Engle [1982a] and extended by Bollerslev [1986], Engle and Bollerslev [1986], and Weiss [1984, 1986a].

We then move to the problem of estimation under correct specification, and this is the subject of Chapter 3. The central issue is that we can extract information separately from each of the two conditional moments by means of orthogonality conditions and then form matrix weighted averages of the separate estimators to combine the information and obtain joint estimators. Further, the contribution to efficiency of each of the moments can be measured using simple statistics. These developments are closely related to the joint generalized least squares estimator presented by Jobson and Fuller [1980]. Under conditional normality, the likelihood can be locally factorized and one of the factors contains the information contributed by the conditional mean while the other factor contains the information contributed by the conditional variance. Of course there is no computational need to separate the estimation problem in this fashion, as full maximum likelihood can be implemented in microcomputers. But the presentation of separate estimators clarifies the structure of joint estimators, and we argue that it constitutes an important tool to assess model specification. To estimate the mean equation, generalized least squares with known conditional variance is set as an efficiency benchmark, and using parametric estimates of the conditional variance produces estimators with the same asymptotic distribution (Carroll and Ruppert [1982b]). Alternative semi-parametric approaches are briefly reviewed: least squares (White [1980b]) is seen to be inefficient even when the model is unconditionally homoskedastic, and an improvement is provided by partially generalized least squares (Amemiya [1983] and Cragg [1983]), while Carroll [1982] and Robinson [1987] have provided semi-parametric estimators of the parameters of the
conditional mean which are fully efficient with respect to the information contributed by the mean of the process. To estimate the variance equation we must solve the problem posed by the unobservability of its dependent variable (the squared mean innovations), but this is solved trivially using residuals from the mean equation. There may also exist identifiability problems for the parameters of this equation, so that in general we can only estimate a function of the parameters of lower dimension, and we find the form of the identifiable parameters for some leading cases. Generalized least squares with known conditional kurtosis is set as an efficiency benchmark for the identifiable parameters, and using parametric estimates of the conditional kurtosis results in an estimator with the same asymptotic distribution. The possibilities of semi-parametric approaches are considered, and the commonly used two-stage estimators of a subset of the parameters are studied (Amemiya [1977]).

The "axiom of correct specification" is certainly very restrictive (Leamer [1978]), but it provides a useful benchmark for the analysis. In Chapter 4 we analyze the effects of specification error on the properties of the quasi-maximum likelihood estimator (the estimator obtained from maximizing the likelihood function assuming normality). We assume that the pseudo-true value of the parameters exists (Domowitz and White [1982]), and conditions for the consistency of estimators under misspecification are characterized in terms of expectations of the score evaluated at this pseudo-true value. These conditions are used to analyze different forms of specification error, and in general we find that misspecification of the first two moments causes inconsistency in the estimators, misspecification of the third and fourth moments does not affect consistency but results in incorrect estimates for the covariance matrix of the estimators, and misspecification in higher order moments does not affect the asymptotic distribution of the estimators but results in inefficiency. Thus as usual a bid to improve efficiency introduces the risk of inconsistency. Conditions under which serial correlation does not
introduce inconsistency in the estimation of mean parameters are derived, and the ARCH model is seen to preserve the consistency of estimators of the mean parameters under some classes of specification error in the conditional variance.

The consequences of specification error call for a careful evaluation of at least the first four conditional moments, and for an assessment of whether there may be substantial loss of information due to incorrect distributional assumptions. Chapters 5 and 6 are devoted to this task. In Chapter 5 we derive general diagnostic tests which do not use information external to the model. To evaluate the first two moments we produce tests of the coherency of the information of the two moments, and a more general class of tests which we call consistency tests can be derived from residual analysis of both the mean and the variance equations. This class includes the tests of coherency of information and is related to other procedures in the literature, mainly those of Hausman [1978], White [1980a], Pagan and Hall [1983], Ruud [1984], Cameron and Trivedi [1985], and Breusch and Godfrey [1986]. The distribution of consistency test-statistics is derived under the null hypothesis and sequences of local parametric alternatives by relating this family of tests to conditional moment tests (Newey [1985a, 1985b], Tauchen [1985]), and the tests can be computed from the coefficient of determination of a double-length auxiliary regression with the residuals of the two equations as dependent variables. An example is given by applying some simple tests to the GARCH specification for foreign exchange rates of Engle and Bollerslev [1986]. To evaluate the specification of moments of order third and higher, we develop a general class of tests which we call efficiency tests. Efficiency tests are also based on the analysis of the residuals, and the distribution of the efficiency test-statistics is also derived by relating them to conditional moment tests.

In Chapter 6 we consider model evaluation by testing the model against specific alternatives, thus using external information to enhance power in the
desirable directions. The LM test for variable additions (Breusch and Pagan [1980], Engle [1982b, 1984], Pagan [1984a]) in the mean or variance equations is shown to be a consistency test by suitable choice of the moment restrictions. The results of Engle [1982b] are extended to the case of a non-diagonal covariance matrix, and the principle is used to generalize some typical departures considered in applied work to a heteroskedastic setup. We consider testing for autocorrelation, lag orders and common factors in both the mean and variance equations; testing for structural breaks (Chow [1960]) and prediction error tests (Salkever [1976], Pagan and Nicholls [1984]); and testing for weak and strong exogeneity (Engle et al. [1983], Wu [1973], Hausman [1978], Geweke [1978]). As to testing the higher order moments, the normality test of Jarque and Bera [1980] against the Pearson family is generalized to heteroskedastic models and shown to be an efficiency test by suitable choice of the conditional moment restrictions.

The developments mentioned above constitute a more complete theoretical body of inferential procedures than has been provided so far in the literature on heteroskedastic models, and in order to have some means to assess the performance in small to moderate samples of the asymptotic approximations derived, some Monte Carlo experiments are reported at the end of each of Chapters 2 to 6.

Our next task is to try to cover a wider range of situations, and we proceed to generalize the results obtained so far. This is the objective of the four chapters which constitute the second part of the Thesis. In these chapters the estimation and evaluation framework of the first part of the Thesis is adapted to more general situations, and since the generalizations are made in different directions there is not a sequential path to the reading of these four chapters.

In Chapter 7 the model is extended to allow for latent variables which represent measures of risk (Pagan and Ullah [1986]). We concentrate on
parametric risk measures, derive tests for risk effects, and classify parametric risk models in terms of the relationship between the risk measure and the variables in the model. When the risk term is a function of the conditional variance of the dependent variable we term the model a $y$-risk model (as the ARCH-M model of Engle et al. [1987]). The regularity conditions and factorization of the likelihood are reconsidered in this framework, and it is seen that the consistency and efficiency tests derived in Chapters 5 and 6 apply directly to $y$-risk models. The risk premia model of Engle et al. [1987] is analyzed as an illustration. When the risk measure is a parametric function of the conditioning variables we term the model an $x$-risk model (Hansen and Hodrick [1983]). Inference may be conducted on $x$-risk models from the joint likelihood, or from the conditional likelihood using a two-stage approach (Pagan [1984b, 1986]). Two-stage estimators are derived and their properties are analyzed, and the efficiency and consistency tests are extended to the two-stage procedure by taking into account the additional source of randomness introduced by the extraneous estimates obtained in the first stage.

In Chapter 8 we consider varying coefficients in the conditional mean. The coefficients may vary deterministically (Belsley [1974a, 1974b]), randomly (Swamy [1971]), or they may evolve randomly (Nicholls and Pagan [1985]). We discuss the conditions under which the results available for homoskedastic models extend to models with implicit sources of heteroskedasticity, and the estimation and evaluation procedures of Chapters 2 to 6 are easily generalized because the likelihood has been cast in state space form from the beginning. Further tests for parameter stability are derived which generalize well known results in homoskedastic models (e.g. Breusch and Pagan [1979], Nicholls and Quinn [1982]), and an evolving coefficient model for the joint parameters of the conditional and marginal models is proposed to represent a learning mechanism on the part of economic agents and policy makers, providing a natural environment to test the Lucas [1976] critique. The Kalman filter in this
model must introduce new information in stages, and LM tests for superexogeneity are derived.

In many cases researchers are interested in modelling more than one variable, and we face the task of generalizing our results to a multivariate framework in Chapter 9. The model is interpreted as a system including mean, variance and covariance equations with (possibly) cross-equation restrictions, and we extend the principle of extracting information separately from orthogonality conditions and of combining the information in matrix weighted averages. The likelihood function of the multivariate normal heteroskedastic model can be locally factorized and the consistency and efficiency tests are generalized to evaluate the model, using vectorizations of the higher order moments.

Once the move has been made to incorporate the second moment into the modelling process, the question arises as to the possibility of considering higher order moments as well. In Chapter 10 we explore this situation by viewing the problem as a generalization of extracting information from the variance under heteroskedasticity: each moment provides an additional equation to the system introducing the potential to improve efficiency in estimation, and we can extract the information from orthogonality conditions. To motivate this exploration we use the two-moment case when the distribution is not symmetric and the estimators may not be combined by simple matrix weighted averages. We use a modified version of the variance equation that contains only the information not already contributed by the mean, and with this we recover the asymptotic independence between the separate estimators. This approach is then generalized to produce a sequential search for information in higher order moments in which use is made of r-th order equations free from the information already contributed by the first r-1 moments, and conditions to stop the search for further information are provided. The unobservability of the dependent variables in higher order
equations is solved by the use of residuals from the mean equation. When the conditional distribution is symmetric this raises no further problems, but when the distribution is not symmetric, substituting innovations by residuals introduces a new source of uncertainty which must be accounted for. The tests for coherency of information are generalized and constitute an integral part of the search for information in higher order moments.

Finally, in Chapter 11 we review two lists of questions: those which have been answered and those which have not been answered. As in most research projects, we find that the second list is substantially longer and constitutes an interesting set of topics awaiting further research.
CHAPTER 2

THE BASIC MODEL

This Chapter is a reference for the main core of the Thesis contained in Chapters 3 to 6. The model and notation are presented in section § 2.1 and the basic assumptions discussed in section § 2.2. It is not our purpose to weaken the set of assumptions to the last consequence, nor do we want to make such assumptions at an excessively high level. We follow Hansen [1982] and White and Domowitz [1984] and relate their hypotheses to the specific case of heteroskedastic models. In section § 2.3 we analyze some of the more important particular cases in the literature. The Chapter concludes with the presentation of the experimental design for the Monte Carlo evidence that will be discussed in Chapters 3 to 6.

§ 2.1 Basic model and notation

We start from the relation between a (scalar) variable \( Y \) and a vector of variables \( X^* \). The primitive theoretical proposition is that \( Y \) is a function of \( X^* , Y = Y(X^*) \). The variables correspond to measurable and observable concepts and are thus related to specific time periods or units. For the \( t \)-th period, \( Y \) and \( X^* \) are represented by \( y_t \) and \( x_t^* \), respectively. Throughout the Thesis we place the emphasis on time-series models, but many of the results we derive apply as well to cross-section models and to panel data. However, some of the concepts introduced would not make sense in a cross-section framework and would require modification.

We gather information on \(( Y , X^* )\) and denote the information sets for a given number of periods from \( s \) to \( t \), \( s \leq t \), as
\[ Y_t^g = \sigma( y_s, y_{s+1}, ..., y_t ), \quad \text{and} \quad X_t^{*g} = \sigma( x_s^*, x_{s+1}^*, ..., x_t^* ), \]

where for any set of observations \( A \), \( \sigma(A) \) represents the corresponding \( \sigma \)-algebra (White [1984], Spanos [1986]). We use \( Y_t = Y_t^{\infty} \) and \( X_t^* = X_t^{*\infty} \) to designate the past information sets of \( Y \) and \( X^* \) respectively, up to and including period \( t \).

The data generating process (DGP henceforth) for \( (y_t, x_t^*) \) is the time-dependent probability mechanism that underlies the stochastic structure of the variables (see Hendry and Richard [1982,1983]), and is given by the joint probability density function (pdf) conditional on past information \(^{(1)}\). This joint DGP can always be factorized as

\[ D( y_t, x_t^* | \mathcal{G}_t ) = D( y_t | \mathcal{F}_t ) D( x_t^* | \mathcal{G}_t ), \]

where \( \mathcal{G}_t = \sigma( Y_{t-1} U X_{t-1}^* ) \) is the past information set, and \( \mathcal{F}_t = \sigma( Y_{t-1} U X_t^* ) = \sigma( \mathcal{G}_t U x_t^* ) \) is the conditioning information set, and we have that \( \mathcal{G}_t \subset \mathcal{F}_t \).

Corresponding to the primitive theoretical proposition \( Y = Y( X^* ) \), our interest is to construct a statistical model for the conditional DGP \( D( y_t | \mathcal{F}_t ) \). This constitutes the conditional model. At the extreme of generality we may build a fully non-parametric conditional model, and at the extreme of specificity we can construct a fully parametric one. In between these two extremes we have the semi-parametric alternative in which part of the conditional model is characterized by a finite-dimensional parameter set and part by an infinite-dimensional one. In the literature on heteroskedastic models Amemiya [1973,1977], Jobson and Fuller [1980], Engle [1982a], and Bollerslev [1986] are examples of parametric models, Fuller and Rao [1978], White [1980b], Carroll [1982], Carroll and Ruppert [1982b], Amemiya [1983], Cragg [1983], and Robinson [1987] are examples of semi-parametric specifications, and Hall and Carroll [1987] analyze a fully non-parametric

---

\(^{(1)}\) We use the term DGP to denote the distribution in general or its corresponding pdf.
model. Our approach is mainly parametric though we consider the possibility of specifying only a few conditional moments which is itself a semi-parametric proposition. Also in Chapter 3 we briefly review some semi-parametric approaches to inference and explore the possibilities of mixed parametric/semi-parametric strategies.

A random variable is fully characterized by its distribution or its characteristic function. Our weakest proposition is a partial description of the characteristic function by the parametric specification of the first two moments of the conditional DGP. The conditional mean is characterized by a $k$-dimensional parameter vector $\beta$, $\beta \in \mathbb{R}^k$, that is,

$$E \left[ y_t \mid \mathcal{F}_t \right] = \mu_t = \mu_t(\beta) = \mu_t(\beta; \mathcal{F}_t).$$

The information set $\mathcal{F}_t$ and the parameter vector $\beta$ will not be shown as explicit arguments unless necessary. For ease of reference to the most common case in which $\mu_t$ is linear in $\beta$ we denote

$$x_t = x_t(\beta) = x_t(\beta; \mathcal{F}_t) = \frac{\partial \mu_t}{\partial \beta},$$

which does not depend on $\beta$ in the linear case. The conditional variance is characterized by the $p$-dimensional parameter vector $\theta$, $\theta \in \Theta \subset \mathbb{R}^p$, as in

$$\text{Var} \left[ y_t \mid \mathcal{F}_t \right] = h_t = h_t(\theta) = h_t(\theta; \mathcal{F}_t).$$

The class of models which is of our particular interest is that in which $\mu_t$ and $h_t$ have common parameters. To allow for this we let $p \geq k$ and introduce the vector $\alpha$, $\alpha \in \mathcal{Q} \subset \mathbb{R}^{p-k}$, such that $\theta$ may be partitioned as $\theta = (\beta', \alpha')'$. In most cases $\Theta = \mathcal{B} \times \mathcal{Q}$, but this need not be the case. Thus $p$ is the total number of parameters for the two conditional moments. Also for ease of reference to the most common case, denoted the linear-in-$\alpha$ model, in which $h_t$ is a linear function in $\alpha$, we define

$$z_t = z_t(\theta) = z_t(\theta; \mathcal{F}_t) = \frac{\partial h_t}{\partial \alpha},$$

(1)
so that when \( h_t \) is linear-in-\( \alpha \) we have \( h_t = z_t \alpha \), where \( z_t = z_t(\beta) \) in general. The other partial derivative of \( h_t \) is

\[
wt = wt(\theta) = \frac{\partial h_t}{\partial \beta},
\]

(2) and \( w_t \) and \( z_t \) are stacked into

\[
s_t = s_t(\theta) = s_t(\theta;F_t) = \left( \begin{array}{c} w_t \\ z_t \end{array} \right).
\]

The parameter space \( \Theta \) and/or the function \( h_t \) and its variable arguments in \( F_t \) need to incorporate restrictions to ensure \( h_t > 0 \) for all \( t \). This can be achieved if \( h_t = h(\xi_t) \), where \( \xi_t = \xi_t(\theta;F_t) \), and \( h(\xi) > 0 \) for \( \xi \), which generalizes the formulation of Breusch and Pagan [1979]. For example, Harvey [1976], Hausman et al [1984], Gourieroux et al [1984b] and Geweke [1986a] choose \( h(\cdot) = \exp(\cdot) \). Other functional forms may require restrictions to be imposed on the parameter space (mainly on \( \Theta \)), but for most well-known variance specifications such restrictions can be enforced through reparameterization of the model. In linear-in-\( \alpha \) models we assume that \( z_t > 0 \), so that reparameterizing in terms of \( \alpha^* = \alpha^{1/2} \) ensures positivity.

In parallel to the specification of the conditional moments \( \mu_t \) and \( h_t \) we have the corresponding sequences of innovations. The innovations in the conditional mean of \( y_t \) are

\[
\epsilon_t = \epsilon_t(\theta) = \mu_t - \mathbb{E}[y_t | F_t] = y_t - \mu_t,
\]

(3) and the innovations in the conditional variance of \( y_t \) are

\[
\epsilon_t = \epsilon_t(\theta) = \mu_t - \mathbb{E}[\mu_t^2 | F_t] = \mu_t^2 - h_t.
\]

(4) The joint modelling of \( \mu_t \) and \( h_t \) requires new information to accumulate on both moments and this is reflected in the two sets of innovations \( u_t \) and \( \epsilon_t \). In

(2) We make the convention that scalar functions and expressions referred to vectors apply element by element.
other words, there are two distinct, though possibly related, relationships underlying the heteroskedastic model. From (3) we derive the usual regression function

\[ y_t = \mu_t(\beta) + u_t, \]

which will be referred to as the mean equation. And from (4) we obtain the second relation

\[ u_t^2 = h_t(\beta, \alpha) + \varepsilon_t = h_t(\theta) + \varepsilon_t, \]

which will be referred to as the variance equation. The disturbances \( u_t \) and \( \varepsilon_t \) in these equations are not 'added errors', but are derived from the proposed model for the DGP (see Hendry and Richard [1982,1983], Spanos [1986]).

In Chapter 3 we gain insight into the problem of inference in heteroskedastic models by treating the mean and variance equations as a set of seemingly unrelated regression equations (SURE, Zellner [1962]). Let \( g_t = g_t(\theta) = g_t(\theta; \mathcal{F}_t) = (\mu_t, h_t)' \), \( \eta_t = (y_t, u_t)' \), and \( \nu_t = (u_t, \varepsilon_t)' \). The two-equation system is given by

\[ \eta_t = g_t(\theta) + \nu_t, \]

and has conditional covariance matrix

\[ \Sigma_t = \text{var}[\nu_t | \mathcal{F}_t] = E[\nu_t \nu_t' | \mathcal{F}_t] = \begin{pmatrix} h_t & \lambda_t \\ \lambda_t & \kappa_t \end{pmatrix}, \]

where \( \lambda_t = E[(y_t - \mu_t)^2 | \mathcal{F}_t] \), \( \kappa_t = \text{var}[\varepsilon_t | \mathcal{F}_t] = \kappa_t^* - h_t^2 \), and \( \kappa_t^* = E[(y_t - \mu_t)^4 | \mathcal{F}_t] \), and of course \( \lambda_t, \kappa_t^* \in \mathcal{F}_t(3) \). The parameterization of these higher order moments will be explicitly introduced when required. When the conditional distribution is symmetric we have that \( \lambda_t = 0 \) and \( \Sigma_t \) is diagonal.

---

*(3) The notation 's stands for 'is a measurable function of \( \mathcal{F}_t \).'*
Our strongest proposition is the complete parametric specification of the conditional DGP,

\[ D (y_t | \mathcal{F}_t) = f(y_t | \mathcal{F}_t; \theta, \pi), \]

where \( f(\cdot) \) is a pdf and \( \pi \) is the vector of (possible) additional parameters in moments of order higher than two. In this case we use the likelihood function

\[ \mathcal{L}(\theta, \pi) = \prod_{t=1}^{T} f(y_t | \mathcal{F}_t; \theta, \pi) f(y_0 | \mathcal{F}_0; \theta, \pi), \]

where \( f(y_0 | \mathcal{F}_0; \theta, \pi) \) represents the information from the initial conditions \( y_0 \). The corresponding log-likelihood is given by

\[ \mathcal{L}(\theta, \pi) = T^{-1} \sum_{t=1}^{T} \log f(y_t | \mathcal{F}_t; \theta, \pi), \]

where the conventions of normalizing by sample size, neglecting constants, and the assumption that the term \( T^{-1} \log f(y_0 | \mathcal{F}_0; \theta, \pi) \) has no asymptotic effect for inferences on \( \theta \) and \( \pi \), are utilized. The subvector of the score for \( \theta \) is

\[ d_{\theta}(\theta, \pi) = T^{-1} \sum_{t=1}^{T} \frac{\partial \mathcal{L}(\theta, \pi)}{\partial \theta} = T^{-1} \sum_{t=1}^{T} \frac{\partial \log f(y_t | \mathcal{F}_t; \theta, \pi)}{\partial \theta} = T^{-1} \sum_{t=1}^{T} d_{\theta t} = \left(d_{\beta}(\theta, \pi) \right), \]

in obvious notation and partition, and the information submatrix for \( \theta \) is

\[ J_{\theta\theta}(\theta, \pi) = -E \left[ \frac{\partial^2 \mathcal{L}(\theta, \pi)}{\partial \theta \partial \theta'} \right] = \left( \begin{array}{cc} J_{\beta\beta}(\theta, \pi) & J_{\beta\alpha}(\theta, \pi) \\ J_{\alpha\beta}(\theta, \pi) & J_{\alpha\alpha}(\theta, \pi) \end{array} \right). \]

The most common full parametric model is the normal model which assumes conditional normality, that is,

\[ y_t | \mathcal{F}_t \sim N[\mu_t, h_t], \]

and it has been used successfully in heteroskedastic econometric models in many cases. See for example Engle [1982a,1983], Weiss [1984], Domowitz and Hakkio [1985], Bollerslev [1986], Diebold and Nerlove [1986], and Engle et al [1987]. The normal model has also been the basic building block in theoretical developments such as Amemiya [1973], Jobson and Fuller [1980], Carroll and Ruppert [1982a], Engle [1982a], Weiss [1984,1986a], and Bollerslev [1986], inter
Other distributions that have played an important role in the subject either theoretically or empirically and that will receive mention in this Thesis are the t-distribution (Bollerslev [1985], and Engle and Bollerslev [1986]), the gamma distribution (Amemiya [1973]), and the Poisson distribution and other distributions compounded with the Poisson (Hausman et al [1984]), Gourieroux et al [1984b], and Cameron and Trivedi [1985, 1986]). For some specific purposes other more general distributions such as the Pearson family (Kendall and Stuart [1968]) will be used as well.

Since the normal model will play an important role in our analysis, it is convenient to present the likelihood, score and information matrix for this case (see Engle [1982a]). These are given by

\[ \mathcal{L}(\theta) = \left[ \prod_{t=1}^{T} h_t \right]^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} h_t^{-1} u_t^2 \right\} \left[ (2\pi)^{-T/2} f(y_0 | \mathcal{F}_0 ; \theta) \right], \]  

- (9)

\[ L(\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log h_t - \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-1} u_t^2, \]  

- (10)

\[ d_\theta(\theta) = T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} s_t \varepsilon_t \]  

\[ = T^{-1} X' \Omega^{-1} u + \frac{1}{2} T^{-1} S' \Omega^{-2} \varepsilon \]  

\[ = T^{-1} G' \Sigma^{-1} u, \]  

- (11a)

\[ d_\beta(\theta) = T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} w_t \varepsilon_t \]  

\[ = T^{-1} X' \Omega^{-1} u + \frac{1}{2} T^{-1} W' \Omega^{-2} \varepsilon, \]  

- (11b)

\[ d_\alpha(\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t \varepsilon_t = \frac{1}{2} T^{-1} Z' \Omega^{-2} \varepsilon, \]  

- (11c)

and
\[ j(\theta) = E \left[ T^{-1} G' \Sigma^{-1} G \right] = E \begin{pmatrix} T^{-1}X' \Omega^{-1} X + \frac{1}{2} T^{-1} W' \Omega^{-2} W \\ \frac{1}{2} T^{-1} Z' \Omega^{-2} Z \end{pmatrix} \]  

where \( \bar{x}_t = (x_t', 0)' \), \( y = (y_1, \ldots, y_T)' \), \( \mu = (\mu_1, \ldots, \mu_T)' \), \( u = (u_1, \ldots, u_T)' \), \( h = (h_1, \ldots, h_T)' \), \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)' \), \( X = (x_1, \ldots, x_T)' \), \( \bar{X} = (\bar{x}_1, \ldots, \bar{x}_T)' = (X, 0) \), \( W = (w_1, \ldots, w_T)' \), \( Z = (z_1, \ldots, z_T)' \), \( S = (s_1, \ldots, s_T)' = (W, Z) \), \( \eta = (y', u^2)' \), \( g = (\mu', h')' \), \( v = (u', \varepsilon')' \), \( \Omega = \text{diag} \{ h_1, \ldots, h_T \} \), \( \Delta = \text{diag} \{ \Delta_1, \ldots, \Delta_T \} \), \( K = \text{diag} \{ \kappa_1, \ldots, \kappa_T \} \), \( G = \begin{pmatrix} X \\ S \end{pmatrix} = \begin{pmatrix} X \\ 0 \end{pmatrix} \), and \( \Sigma = \begin{pmatrix} \Omega & \Delta \\ \Delta^* & K \end{pmatrix} \). In (9) - (12) we have used the properties of the normal distribution that \( \Delta = 0 \) and \( K = 2 \Omega^2 \), and the functions depend only on \( \theta \) because the normal distribution is completely characterized by its first two moments. For this reason we have also deleted the '\( \theta \theta \)' subscript from the information matrix.

§ 2.2 Assumptions

The previous section gives form to the model and produces our fundamental 'correct specification' assumption

\((Q0)\) The conditional mean of \( y_t \) is \( \mu_t(\theta_0) = E [ y_t | F_t ] \), its conditional variance is \( h_t(\theta_0) = \text{Var} [ y_t | F_t ] \), and the third and fourth moments are given by \( \Lambda_t = E [ (y_t - \mu_t(\theta_0))^3 | F_t ] \) and \( \kappa_t^* = E [ (y_t - \mu_t(\theta_0))^4 | F_t ] \) which exist and are finite conditional on \( F_t \).

A full parametric alternative to \((Q0)\) is

\((Q0')\) \( y_t | F_t \sim N [ \mu_t(\theta_0), h_t(\theta_0) ] \).

Other possibilities are the gamma, the Poisson, or Student's t distributions, or simply the parameterization of \( \Lambda_t \) and \( \kappa_t \) without further specification of the conditional pdf.
There are now available many powerful results to establish the asymptotic properties of estimators in nonlinear models under a variety of circumstances, see Jennrich [1969], Malinvaud [1970], Hannan [1971], Burguete et al [1982], Hansen [1982], White and Domowitz [1984], inter alia. The last two references are of particular interest and will allow us to establish the main asymptotic results under different conditions of heterogeneity and dependence of the processes involved. The generalized method of moments (GMM) theory of Hansen [1982] permits relatively mild memory restrictions for economic variables in the form of ergodicity, and requires a degree of homogeneity which is fulfilled by strict stationarity. The general theory for nonlinear regression with dependent observations presented by White and Domowitz [1984] allows for more heterogeneity by strengthening the conditions on the memory of the process to uniform or strong mixing of some given order. In line with this, our next assumption is

(Q1) \((y_t, x_t^*)\) is stationary and ergodic,

or

\((y_t, x_t^*)\) is mixing with \(\phi(m)\) of size \(r/(2r - 1)\), \(r \geq 1\), or \(\alpha(m)\) of size \(r/(r - 1)\), \(r > 1\).

Note that \((u_t^2, x_t^*)\) has the same heterogeneity and memory structure than \((y_t, x_t^*)\) e.g. Lemma 2.1 of White and Domowitz. This is important for the estimation of the variance equation. The parameter space is restricted by

(Q2) \(\Theta\) is a compact subspace of \(\mathbb{R}^p\) and \(\theta_0\) is an interior point of \(\Theta\).

Not allowing \(\theta_0\) to lie on the boundary of the parameter space is required for a well-behaved (normal) asymptotic distribution of the estimators.

Some structure must be imposed on the conditional moment functions \(\mu_t\) and \(h_t\), and this is done in
\( (\text{Q3}) \) \( \mu_t \) and \( h_t \) are measurable functions of \( \mathcal{F}_t \) for all \( \theta \in \Theta \), and are continuously differentiable of order 2 on \( \Theta \), uniformly in \( t \).

Further, \( \mu_t \) and \( h_t \) and their first two derivatives with respect to \( \theta \) are bounded from above and \( h_t \) is bounded away from zero almost everywhere in \( \Theta \), uniformly in \( t \).

This assumption states the smoothness conditions which are characteristic of parametric work in heteroskedastic models in econometrics. Given this smoothness and the compactness of \( \Theta \), boundedness relates to functions of moments of the variables which are generally assumed to hold (e.g. the convergence of \( T^{-1} X'X \) in the classical linear model). Positivity of the conditional variance typically imposes some restrictions on the parameter space \( \Theta \), as discussed in the previous section.

We must ensure the existence of the first few unconditional moments and also that we can take expectations of the likelihood function or other estimation criteria (quadratic in nature) over the parameter space. For this purpose we introduce

\( (\text{Q4}) \) \( \varepsilon_t^2 = \varepsilon_t(\theta_0)^2 \) is uniformly \((r + \delta)\)-integrable \( r \geq 1, \delta > 0 \), while

\[ [ \mu_t(\beta_0) - \mu_t(\beta) ]^2 \text{ and } [ h_t(\theta_0) - h_t(\theta) ]^2 \] are dominated by uniformly \((r + \delta)\)-integrable functions .

Note that the uniform integrability of \( \varepsilon_t^2 \) implies the existence of at least four moments of \( y_t \). The first two are usually assumed for mean models, and the last two are required for the proper estimation of the variance equation. Also observe that this imposes conditions on dynamic models which in general imply further restrictions on the parameter space \( \Theta \).

Two additional regularity assumptions must be made to ensure the applicability of the central limit theory of Hansen and of White and Domowitz.
These assumptions are required to ensure the proper asymptotic behavior for the Hessians of the relevant criterion functions.

\[(\text{C5})\]

The matrix function of \(\theta\) given by \(\frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} \frac{\partial g_t}{\partial \theta'}\) is dominated by uniformly \(r\)-integrable functions, \(r > 1\), and has finite expectation at \(\theta_0\).

\[(\text{C6})\]

\(\{ h_t^{-1} x_t x_t' - h_t^{-1} u_t(\beta) \frac{\partial x_t}{\partial \beta'} \} \) and \(\{ \kappa_t^{-1} s_t s_t' - \kappa_t^{-1} \varepsilon_t(\theta) \frac{\partial s_t}{\partial \theta'} \} \) are dominated by uniformly \(r\)-integrable functions, \(r > 1\).

Observe that \((\text{C5})\) is sufficient for the same condition to apply separately to \(h_t^{-1} x_t x_t'\) and \(\kappa_t^{-1} s_t s_t'\).

Finally, identifiability is ensured by means of

\[(\text{C7})\]

\[\text{Var} [ T^{-1/2} X' \Omega^{-1} u ] = E [ T^{-1} X' \Omega^{-1} X ] \quad \text{and} \quad \text{Var} [ T^{-1/2} G' \Sigma^{-1} u ] = E [ T^{-1} G' \Sigma^{-1} G ] \]

are uniformly positive definite in an open neighborhood of \(\theta_0\).

The equalities in \((\text{C7})\) are obtained using iterated expectations. In other words, \((\text{C7})\) ensures the identifiability of \(\beta_0\) in the mean equation, and that of \(\theta_0\) in the two-equation system. Because the mean equation does not contain information about \(\alpha_0\), then given \(\beta\), this vector must be identifiable in the variance equation. Therefore \((\text{C7})\) implies that \(\text{Var}[ T^{-1/2} Z' K^{-1} \varepsilon ] = E [ T^{-1} Z' K^{-1} Z ]\) is uniformly positive definite in an open neighborhood of \(\theta_0\).

In Chapter 3 we make repeated use of Theorems 3.1 and 3.2 of White and Domowitz, and Theorems 2.1 and 3.1 of Hansen. A summary of these results can be stated as

**Theorem 2.1.** Suppose a parameter vector \(\lambda \in \Lambda \subset \mathbb{R}^n\) is to be estimated by

\[\hat{\lambda} = \min_{\lambda \in \Lambda} \psi(\lambda)^T A_T \psi(\lambda)\]
where $\psi (\lambda) = T^{-1} \sum_{t=1}^{T} \psi_t (\lambda) , E [ \psi_t (\lambda_0) ] = 0 ,$ and $A_T \overset{\text{as}}{\to} \mathcal{E} \{ T \psi(\lambda_0) \psi(\lambda_0)' \} \ (4)$, where $\lambda_0$ is the true value of $\lambda$. Assume that the regularity conditions of Hansen and/or White and Domowitz are fulfilled for this problem. Then

$$T^{1/2} (\widehat{\lambda} - \lambda_0) \overset{d}{\to} N \{ 0 , V(\widehat{\lambda}) \} ,$$

where (5)

$$V(\widehat{\lambda}) = \left[ \mathcal{E} \{ \frac{\partial \psi(\lambda_0)'}{\partial \lambda} \} \mathcal{E} \{ T \psi(\lambda_0) \psi(\lambda_0)' \}^{-1} \mathcal{E} \{ \frac{\partial \psi(\lambda_0)'}{\partial \lambda} \} \right]^{-1} .$$

**Proof:** See Theorems 3.1 and 3.2 of White and Domowitz and Theorems 2.1 and 3.1 of Hansen.

A unified treatment of both types of estimators is considered by Burguete et al [1982], and Chamberlain [1987] discusses the selection of the optimal weighting matrix. In order to avoid later repetition we establish here the sufficiency of the assumptions (C10) - (C17) for the application of these results to our specific problem.

**Proposition 2.2.** Suppose the mean equation in (5), or the variance equation in (6), or both equations jointly as in (7), are to be estimated respectively by

$$\min_{\theta} \{ u' Q^{-1} u \} , \min_{\theta} \{ s' K^{-1} s \} , \text{ or } \min_{\theta} \{ n' I_{n} \} ,$$

where $Q$, $K$ and $\Sigma$ are given. Let (A1) - (A9) denote Assumptions 1 to 9 of White and Domowitz [1984]. Then (C10) - (C17) are sufficient for (A1) - (A9).

**Proof:** (C12) and (C13) $\Rightarrow$ (A1) and (A5) ; (C14) $\Rightarrow$ (A2) ; (C10) $\Rightarrow$ (A3) ; (C17) $\Rightarrow$ (A4) and (A9) ; the martingale assumption in (C10) with Exercise 5.26 of White [1984] ensure that (C17) also implies (A7) ; (C15) $\Rightarrow$ (A6) ; and (C16) $\Rightarrow$ (A8). The mixing conditions of Theorems 3.1 and 3.2 of White and Domowitz are guaranteed by (C11), and therefore we can readily apply these results to generalized least squares (GLS) estimation of the equations in (5) - (7).

---

(4) For any random sequence $X_T$ we define $\mathcal{E} \{ X_T \} = \lim_{T \to \infty} E \{ X_T \}$, provided such limit exists.

(5) We denote $\frac{\partial f(x)}{\partial x}$ evaluated at $x_0$ by $\frac{\partial f(x_0)}{\partial x}$.
Proposition 2.3.- Suppose the mean equation in (5), or the variance equation in (6), or both equations jointly as in (7), are to be estimated by the GMM method with orthogonality conditions given by equating to zero the parametric functions $T^{-1}X'\Omega^{-1}u$, $T^{-1}S'K^{-1}\varepsilon$, or $T^{-1}G'\Sigma^{-1}v$, respectively, where $\Omega$, $K$ and $\Sigma$ are given. Let (A2.1)-(A2.5) denote Assumptions 2.1 - 2.5, and (A3.1)-(A3.6) denote assumptions 3.1 - 3.6 of Hansen [1982]. Then (Q0) - (Q7) are sufficient for (A2.1) - (A2.5) and also for (A3.1) - (A3.6).

Proof: (Q1) $\Rightarrow$ (A2.1) and (A3.1); (Q2) and (Q3) $\Rightarrow$ (A2.2), (A2.3), (A3.2), and (A3.3). The martingale difference assumption in (Q0) together with (Q5) and (Q7) are sufficient for (A2.4), (A3.4) and (A3.5), by Theorem 5.24 of White [1984].

Since there are as many orthogonality conditions as parameters in all three cases, (A2.5) and (A3.6) are trivially satisfied. Finally, note that for Theorem 2.1 of Hansen condition (i) is guaranteed by (Q5) and his Lemma 3.1, (ii) is ensured by (Q2), and the identifiability condition in (iii) is fulfilled in view of (Q7). Therefore, we can also readily apply Theorems 2.1 and 3.1 of Hansen to the heteroskedastic setting under (Q0) - (Q7).

The GLS estimators of White and Domowitz and the GMM estimator of Hansen cannot make claims of efficiency other than in relation to the orthogonality conditions involved\(^{(6)}\). Such relative efficiency follows trivially in this case from Theorem 3.2 of Hansen and also from Chamberlain [1987] because the number of orthogonality conditions in Proposition 2.3 equals the number of parameters, and hence the definition of a weighting matrix is superfluous. Under some circumstances, however, stronger efficiency claims can be made. This generally entails the knowledge of the form of the conditional DGP up to the parameter vector $\theta$, and for full efficiency in the Cramer-Rao sense in this context we will require the additional assumption

\(^{(6)}\) All efficiency propositions in this Thesis relate to asymptotic efficiency.
The variables $x_t^*$ are weakly exogenous for $\theta$ in the sense of Engle, Hendry and Richard [1983].

This ensures that the conditional likelihood in (8) contains all the relevant information about $\theta$.

§ 2.3 Some special cases

In this section we present some particular cases of heteroskedastic models which have been studied and applied in the literature. Subsection § 2.3.1 considers models in which $h_t$ is not a function of $\beta$, while the remaining subsections deal with cases in which $h_t$ is parameterized as a function of $\beta$.

§ 2.3.1 Simple heteroskedasticity

In cross-section models it may be too restrictive to assume that all units can be represented by a fixed parameter set. Hildreth and Houck [1968] and Swamy [1971] have proposed a more general model in which the coefficients vary randomly across units as drawings from a common parent distribution. This has led to the random coefficients model which results in a specific heteroskedastic pattern. An alternative is to allow for changing variances and to try to model these changes as functions of other observables. In time series models confusion in the relationships between nominal and real measures induces heteroskedasticity (see Theil [1971]), and the same happens in general if the $\beta$ coefficients evolve stochastically through time. These are cases of the simple heteroskedastic model which is defined when

$$h_t = h_t(\alpha; \mathcal{F}_t),$$

so that $\beta$ does not affect the conditional variance and therefore $w_t = 0$ from (2). If conditional normality is assumed $\langle \alpha \beta \rangle = 0$, as can be seen from (12), and it
follows that the maximum likelihood estimators (MLE's) of \( \beta \) and \( \alpha \) are asymptotically independent.

The most common form given to \( h_t \) has been a linear one, \( h_t = z_t' \alpha \), where \( z_t \in \mathcal{F}_t \), (Amemiya [1977], Pagan [1984a], Pagan and Hall [1983], inter alia). Other possible models include the simple quadratic \( h_t = (z_t' \alpha)^2 \) (Glejser [1969]), and the multiplicative \( h_t = \exp (z_t' \alpha) \) (Harvey [1976]). All these are encompassed by the parameterization of Breusch and Pagan [1979] in which \( h_t = h(z_t' \alpha) \), for some continuous function \( h(\cdot) \) taking positive values only. A constant is usually included in \( z_t \) so that homoskedasticity is nested within this model. If \( z_t \) is strongly exogenous no further conditions are required in general for the existence of moments. But if \( z_t \) is weakly, but not strongly, exogenous so that \( z_t \) is Granger-caused by \( y_t \) (see Granger [1969], Engle et al [1983]), restrictions may be needed on the parameter space for the existence of fourth order moments. Even a simple model like this can produce a very complicated dynamic structure involving both the conditional mean and the conditional variance, and these dynamics will certainly have to be non-explosive for the existence of fourth order moments. Dominance conditions are trivially satisfied if \( h_t \) and \( \mu_t \) are linear. If this is not the case, some structure on the functions \( \mu_t (\cdot) \) and \( h_t (\cdot) \) may be required.

We must also mention in this section the attempts in the literature to allow for heteroskedasticity without going through the burden of specifying the conditional variance. This has led to semi-parametric models in which \( h_t \) is only specified as

\[
h_t = h_t(\mathcal{F}_t).
\]

These models have been studied by Goldfeld and Quandt [1965], Fuller and Rao [1978], White [1980b], Amemiya [1983], Carroll [1982], Carroll and Ruppert [1982b], Cragg [1983] and Robinson [1987], among others. Here no form for \( h_t \)
is put forward other than the variables involved as arguments, and maybe some ranking relations.

§ 2.3.2 Amemiya Model

One of the earliest references to a model in which the heteroskedasticity is made dependent upon $\beta$ is the Prais and Houthakker [1955] study of family budgets (see also Theil [1951]). There they propose a cross-section model in which the variance is proportional to the square of the mean,

$$h_t = \alpha \mu_t^2,$$

and therefore $h_t = h_t(\beta, \alpha)$. They did not capitalize, however, on the information provided by the variance. Amemiya [1973] considered this model in a likelihood framework and proved that the GLS estimator of $\beta$ would be inefficient if $y_t$ was distributed normally or log-normally conditional on $\mathcal{F}_t$, thus showing the potential importance of variance information. He also proved that if $y_t | \mathcal{F}_t \sim \Gamma(\alpha, \alpha^{-1} \mu_t)$, then GLS for $\beta$ would be fully efficient.

This model is clearly linear-in-$\alpha$, and from (1) and (2) we have

$$z_t = \mu_t^2,$$

and

$$w_t = 2 \alpha \mu_t x_t,$$

so that using (12) we have under conditional normality

$$J_{\beta\alpha}(\theta) = \alpha^{-1} E \left[ \frac{1}{T} \sum_{t=1}^{T} \mu_t^{-1} x_t \right],$$

which will be nonzero in general. Therefore there is asymptotic dependence between the MLE's of $\beta$ and $\alpha$ in the Amemiya model.

If $\mu_t$ depends on strongly exogenous variables only, then the model is a static one and the conditions for the existence of moments depend only upon the
nature of $x_t^*$. Positivity is ensured by $\alpha > 0$ and this can be implicitly incorporated parameterizing in terms of $\alpha^2$ i.e. $h_t = \alpha^2 \mu_t^2$. But the parameter space has to be restricted to meet the moment conditions when dynamics are allowed in the conditional mean. Dominance conditions, on the other hand, are straightforward when $\mu_t$ is linear.

Homoskedasticity is not nested within the proposition $h_t = \alpha \mu_t^2$, but it can be incorporated with the obvious generalization to $h_t = \alpha_0 + \alpha_1 \mu_t^2$, where $\alpha_0 > 0$ and $\alpha_1 \geq 0$, so that homoskedasticity obtains when $\alpha_1 = 0$ (Jobson and Fuller [1980]). A simple further generalization to other functions of $\mu_t$ is

$$h_t = \alpha_0 + \alpha_1 h_1(\mu_t; \alpha_2)$$  \hspace{1cm} (13)

where $h_1(\cdot)$ is a non-negative function of $\mu_t$ and must obey some smoothness conditions to satisfy the regularity requirements.

§ 2.3.3 The Poisson and Poisson-type models

There are many economic examples in which the dependent variable is discrete rather than continuous (e.g. Maddala [1983], McFadden [1984]). In many instances discrete economic variables can be characterized as Poisson or Poisson compounded processes. Some examples are Chatfield et al [1966], Chatfield and Goodhardt [1970] and Ehrenberg [1972] in analyzing purchasing behavior, and Hausman et al [1984] in the study of the relationship between patents and R&D expenditure, while Cameron and Trivedi [1985, 1986] survey the literature and propose diagnostic tests for these models.

The mean and variance of a Poisson variate are equal and therefore

$$\mu_t = \mu_t(\beta; \mathcal{F}_t) = h_t.$$  \hspace{1cm} (14)

As a model for the conditional variance this could be considered as a special case of (13) but it merits a separate analysis in view of the discreteness of the
dependent variable and its importance in modelling count data. Also observe that there are no $\alpha$ parameters in (14) and so $p = k$ and $\theta = \beta$.

Rather than restricting the parameter space to ensure $h_t > 0$ (and also $\mu_t > 0$ in this case) Hausman et al. [1984] and Gourieroux et al. [1984b] argue for a parameterization that implicitly incorporates positivity, and they suggest $\mu_t = h_t = \exp (x_t^* \beta)$, which implies a null $z_t$ and $x_t = w_t = \mu_t x_t^*$. Dominance conditions are ensured by the compactness of $\mathcal{B}$ with this proposition. The required moment conditions and whether the parameter space needs to be further restricted depends on the dynamic characteristics of the model.

A more general version of the Poisson model compounds this distribution with a random distributional parameter. Hausman et al. have introduced random effects by these means, while Gourieroux et al. have allowed for specification error. A gamma-distributed parameter has led to negative binomial models (see also Ehrenberg [1972]). Such extension of the Poisson model allows for overdispersion ($h_t > \mu_t$) and produces a natural diagnostic for the Poisson model. Cameron and Trivedi [1985] propose alternative regression-based tests for over- and under-dispersion parameterizing the conditional variance under the alternative hypothesis as

$$h_t = \mu_t + \alpha h_1 (\mu_t),$$

for some positive function $h_1$. Then $\alpha < 0$ ($\alpha > 0$) results in under- (over-) dispersion and $\alpha = 0$ represents the Poisson null hypothesis.

Using Poisson characteristics with a continuous dependent variable results in Poisson-type models in which the central issue is the moment relationship in (14). Cameron and Trivedi [1985] and also Pagan and Sabau [1987a] consider the model

$$y_t | \mathcal{F}_t \sim N [ \mu_t (\beta), \mu_t (\beta) ],$$
and the former authors consider alternative assumptions by specifying the first four moments. Here again the fulfillment of the assumptions in § 2.2 depends on the nature of $\mu_t$ and its variable arguments in $\mathcal{F}_t$.

§ 2.3.4 The ARCH class of models

Engle [1982a] proposed the ARCH model to account for inflation uncertainty, and subsequently this model has been successfully applied to study many financial variables (see Engle and Bollerslev [1986] for a survey of applications). The conditional variance of an ARCH process depends on past information, drawing a clear distinction between the conditional and the unconditional second moment of the variable under study. The (linear) ARCH(q) process is characterized by

$$h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j u_{t-j}^2 = z_t' \alpha,$$

where $z_t = (1, u_{t-1}^2, ..., u_{t-q}^2)'$ and $\alpha = (\alpha_0, \alpha_1, ..., \alpha_q)'$, and therefore has a linear-in-$\alpha$ structure. The derivative with respect to $\beta$ is $w_t = -2 \sum_{j=1}^{q} \alpha_j u_{t-j} x_{t-j}$. The positivity constraint requires $\alpha_0 > 0$ and $\alpha_j \geq 0$, $j = 1, ..., q$, which may be implicitly incorporated by parameterization in terms of the $\alpha_j^2$. An alternative approach has been to restrict the lag structure to reduce the number of parameters and hence the probability of obtaining negative values. Although this procedure does not formally exclude negative estimates it has been successfully applied (see Engle [1982a, 1983], Engle et al [1987]).

The moment structure of the ARCH(q) was analyzed by Engle for some cases, and Milhoj [1985] has produced general results. For wide-sense stationarity we require $\alpha_0 > 0$ and $\sum_{j=1}^{q} \alpha_j < 1$ (Theorem 1 of Milhoj), and then $E[h_t] = \alpha_0 [1 - \sum_{j=1}^{q} \alpha_j]^{-1}$. The condition for the existence of fourth order moments is that $3 \alpha_1' (I_q - A) \alpha_1 < 1$, where $\alpha_1 = (\alpha_1, ..., \alpha_q)'$, $A = \| a_{ij} \| = \ldots$
\[ b_{i+j} + b_{i-j} \] , and \( b_i = \alpha_i \) for \( 1 \leq i \leq q \) and \( b_i = 0 \) otherwise. Thus for example for the ARCH(1) the condition is \( \alpha_1^2 < 1/3 \) (Theorem 3 of Milhoj). The ARCH(q) implies a leptokurtic distribution for \( y_t \), and this is one of its attractive features for modelling interest rates, exchange rates, inflation, stock prices and other financial data.

Another interesting aspect of the ARCH model is that \( \beta \alpha = 0 \) under conditional normality and so the ML estimates of \( \alpha \) and \( \beta \) are asymptotically independent (Engle [1982a]). Thus although the conditional variance is a function of \( \beta \) through the presence of \( u_{t-j} \), the ARCH model retains a stochastic structure which is similar in several aspects to the simple heteroskedastic model.

The variance memory of an ARCH(q) dies after \( q \) periods and this led Bollerslev [1986] to put forward the GARCH(q1,q2) process whose conditional variance is

\[
h_t = \alpha_0 + \sum_{j=1}^{q_1} \alpha_j h_{t-j} + \sum_{j=1}^{q_2} \alpha_{q_1+j} u_{t-j}^2,
\]

or

\[
\alpha_1(L) h_t = \alpha_0 + \alpha_2(L) u_t^2,
\]

where \( L \) is the lag operator, \( \alpha_1(L) = 1 - \sum_{j=1}^{q_1} \alpha_j L^j \), and \( \alpha_2(L) = \sum_{j=1}^{q_2} \alpha_{q_1+j} L^j \), for \( q_2 > 0 \). The boundedness of \( h_t \) away from zero still requires \( \alpha_0 > 0 \) and \( \alpha_j \geq 0 \) for \( j > 0 \), but since (15) allows for a long memory with a more parsimonious parameterization than the simple ARCH, it is also less demanding in terms of achieving positivity in estimation. The conditional variance is not linear-in-\( \alpha \) in the GARCH model.

Bollerslev provided the necessary and sufficient conditions for wide-sense stationarity of the GARCH(q1,q2) process, which is that \( \alpha_1(1) - \alpha_2(1) > 0 \), or equivalently that

\[
\sum_{j=1}^{q_1+q_2} \alpha_j < 1.
\]

He also gave the necessary and sufficient

(7) \( A = \| a_{ij} \| \) is used to denote that \( a_{ij} \) is the typical \((i,j)\)-th element of \( A \).
conditions for the existence of $2m$-th order moments in the GARCH(1,1), and in particular for the fourth moment we require that $\alpha_1^2 + 2\alpha_1\alpha_2 + 3\alpha_2^2 < 1$, and similar conditions may be established for higher order processes following the lines of Milhoj [1985] in obtaining the autocorrelation structure for $u_t^2$.

There is a close resemblance of the GARCH variance specification to ARMA models for the mean of a process. This is better appreciated by substituting the variance equation (6) in (15b) to get

$$\alpha_{12}(L) u_t^2 = \alpha_0 + \alpha_1(L) \varepsilon_t,$$

where $\alpha_{12}(L) = \alpha_1(L) - \alpha_2(L)$. Because of this similarity Bollerslev suggested the use of the autocorrelation and partial autocorrelation functions of $u_t^2$ as tools for model identification (see also MacLeod and Li [1983]). The $\varepsilon_t$ follow a heteroskedastic pattern and have changing support (see Engle and Bollerslev [1986]), but nevertheless Bollerslev's suggestion is entirely appropriate if fourth order moments exist because then $\varepsilon_t$ is unconditionally homoskedastic. An interesting alternative for model identification which is also valid in these circumstances is the Hannan and Rissanen [1982] procedure with some information criterion such as AIC (Akaike [1974]) or BIC (Schwarz [1978], see also Geweke and Meese [1981]), and the advantage of such procedure is that it relies more on direct analytical results and less on visual inspection and subjective judgement.

We assume that the polynomials $\alpha_1(L)$ and $\alpha_2(L)$ do not have any common factors. The wide-sense stationarity and positivity conditions ensure the invertibility of $\alpha_1(L)$ and $\alpha_{12}(L)$. Thus solving (16) for $u_t^2$ we obtain

$$u_t^2 = \alpha_{12}(L)^{-1} \alpha_0 + \alpha_{12}(L)^{-1} \alpha_1(L) \varepsilon_t,$$

which provides the MA($\infty$) representation of $u_t^2$. Also, solving (15) for $h_t$ we can write
\[ h_t = \alpha_1(1)^{-1} \alpha_0 + \alpha_1(L)^{-1} \alpha_2(L) u_t^2, \]

which expresses the GARCH\((q_1, q_2)\) as an ARCH\((\infty)\). Thus the normal GARCH process is also leptokurtic, and Bollerslev [1985] and also Engle and Bollerslev [1986] have proposed the use of the conditional t-distribution for fatter tails. Similarly, the block-diagonality of the information matrix of the normal ARCH extends to the normal GARCH.

ARMA models provide a rational approximation to the undeterministic component of Wold's decomposition of wide-sense stationary time series. Diebold [1986a] has argued that GARCH models may play a similar role in modelling conditional variances with time series data, and this suggestion is already implicit in Engle's [1982a] paper. Diebold argues that the undeterministic component of wide-sense stationary series allows for a changing conditional variance and hence GARCH processes are not excluded. But the power of the GARCH parameterization as an approximation to many heteroskedastic patterns is better understood from its ARMA form: if \( y_t \) possesses fourth order moments then \( u_t^2 \) is wide-sense stationary and (16) provides a rational approximation to its undeterministic component, while its changing conditional mean is \( h_t = E [ u_t^2 | \mathcal{F}_t ] \).

Our main concern in econometric modelling is to construct statistical models of economic propositions. ARMA models are atheoretical propositions in general and from the previous argument the same could be said about GARCH models for the conditional variance. What is important about these models is that they permit us to make statistical propositions even in the absence of a priori knowledge. As theory says more about the second moment we expect a movement towards a mixture of the different models presented in this section. Dependence of \( h_t \) on \( \mu_t \) and exogenous variables may represent the theoretical propositions about the variance, while a GARCH component can represent the dynamics in \( h_t \). This is the variance analogy of transfer
function models (see Harvey [1981]) and leads to the generalization of the GARCH($q_1,q_2$) which we will call the GARCHZ($q_1,q_2,q_3$) model

$$\alpha_1(L) h_t = \alpha_3(L)' z_t^\ast + \alpha_2(L) u_t^2,$$

where $z_t^\ast$ is a vector of measurable functions of $\mathcal{F}_t$ representing the variables theoretically affecting the conditional variance, and $\alpha_3(L)$ is a vector of polynomials of orders given by the integer vector $q_3$. We take $z_t^\ast$ to include a constant, so that the GARCH($q_1,q_2$) model is nested within this model. The simple heteroskedasticity, Amemiya and Poisson models may also be nested in this structure. Such a model for $h_t$ is in the spirit of the proposition by Weiss [1984, 1986a, 1986b], where he also allows for MA errors though lagged $h_t$'s are not included in his parameterization. The variance equation can now be expressed as

$$\alpha_{12}(L) u_t^2 = \alpha_3(L)' z_t^\ast + \alpha_1(L) \varepsilon_t,$$

or in transfer function form as

$$u_t^2 = \frac{\alpha_3(L)'}{\alpha_{12}(L)} z_t^\ast + \frac{\alpha_1(L)}{\alpha_{12}(L)} \varepsilon_t.$$

Wide-sense stationarity of $y_t$ requires that $x_t^\ast$ and $z_t^\ast$ be wide-sense stationary, and further depends on the nature of $z_t^\ast$. If $z_t^\ast$ is strongly exogenous then $\alpha_1(1) - \alpha_2(1) > 0$ is still the necessary and sufficient condition, and Weiss [1986b] has given conditions for the wide-sense stationarity of $y_t$ in the GARCHZ model in some cases in which $z_t^\ast$ includes squared lagged dependent variables.

The elements of $z_t^\ast$ must also be strongly exogenous to preserve the block diagonal structure of the information matrix between $\beta$ and $\alpha$ of pure GARCH processes. The GARCHZ model is attractive under these circumstances both because of its generality and tractability. However, in a pure time series context Weiss [1984] has found significant relationships for the
conditional variance of many economic variables with $y_{t-1}^2$, which argues against the empirical plausibility of simple combinations of strongly exogenous variables and GARCH effects. In such case $\alpha \neq 0$ in general, and the same occurs when functions of $\mu_t$ affect $h_t$.

For further reference in this Thesis the terms 'ARCH class' or 'ARCH model' denote the more general family of GARCH parameterizations that preserve a block diagonal information matrix, as this turns out to be a central issue for the properties of this class.

§ 2.4 Objective and design of Monte Carlo experiments

The theoretical results derived in the Thesis are based on asymptotic theory. In practical situations with economic data the sample sizes are typically small, but asymptotic theory makes complex general problems tractable and thus provides an approximation to the sampling properties of estimators and tests. In the absence of operative exact results, this approximation makes possible empirical work in a wide range of situations and provides a benchmark for further theoretical developments. The cost of using asymptotic results in small sample situations is, of course, that we are dealing only with approximations whose quality may differ substantially from one problem to another.

At the other extreme we have Monte Carlo methods, which permit a more precise assessment of the sampling properties of estimators and test-statistics in specific situations, but are limited in scope to the specific forms of data generating processes and models used and thus lack generality. In some cases, the availability of results on invariant properties of the exact distribution of some statistics gives more general validity to simulation results. For example, the distribution of least squares residuals in the classical linear model does not depend on the true value of the parameter vector, amplifying
the scope of simulation experiments in this context (see for example White and MacDonald [1980]). In generalized regression models, a topic of particular relevance for our purposes, Breusch [1980] proved a similar result and thus established the invariance of various statistics to the true value of the mean parameters. Unfortunately, these invariance properties do not extend to the type of models that we consider because generally the dependence of the conditional variance on mean parameters and the nonlinear nature of the problem cause the distribution of the residuals of both the mean and the variance equations to depend on the full parameter vector $\theta$.

A different approach to extend the conclusions obtained from Monte Carlo experimentation has been put forward by Hendry [1979,1984], who advocates the use of response surfaces. These surfaces represent attempts to estimate the distribution of statistics conditionally on the key parameters of the problem, using a specification coherent with the theoretical asymptotic distribution. Maasoumi and Phillips [1982] have expressed a skeptical view on the possibilities of response surfaces, but nevertheless the approach seems to be gaining popularity, and Engle et al [1985] provide an application of this approach to the study of the ARCH model.

As a complement to the theory presented in the next few chapters, we have performed some Monte Carlo experiments. The nature of this evidence is rather limited because we cannot even claim the applicability of Breusch's results in our framework, and therefore the aim of our experiments is simply to analyze in a few models whether the asymptotic results provide reasonable approximations to the properties of estimators and test-statistics in small to moderate samples, or whether there are obvious departures from the asymptotic theory which might suggest situations where special care is required in the application of the theoretical body of the Thesis. The results of the simulations are presented and discussed in Chapters 3 to 6.
The experiments consider the estimation and testing of two different models, the Poisson-N model,
\[ y_t \mid \mathcal{F}_t \sim N \left[ x_t' \beta, x_t' \beta \right], \]
and the ARCH(1) model,
\[ y_t \mid \mathcal{F}_t \sim N \left[ x_t' \beta, \alpha_0 + \alpha_1 u_{t-1}^2 \right], \]
where \( x_t = (1, x_{1t})' \).

We have conducted two types of experiments. The first type looks at the behavior of estimators, and the second type looks at the size and power of tests. The DGP's are designed to provide either correct specification or a misspecified model. Specification error may take the form of a misspecified conditional mean, a misspecified conditional variance, or a misspecified distributional form. The regressor \( x_{1t} \) was generated as an independent uniform variate in \((-1, 0)\), and an additional regressor \( x_{2t} \) generated as independent uniform in \((0, 1)\) was also included in some experiments to induce misspecification in the conditional mean. The correlation between \( x_{1t} \) and \( x_{2t} \) was set to 0.5, the regressors were fixed in repeated samples, and all pseudo-random numbers were generated using NAG Library routines. Most experiments consider four sample sizes (\( T = 20, 50, 100, 200 \)), and in order to reduce experiment variance the smaller samples are always taken as subsamples of the larger ones. The number of replications for each simulation experiment was 500.

It was thought convenient to have the simplest possible DGP's. The true mean parameter vector \( \beta_0 \) was set to \( \beta_0 = (0, 0)' \) for the ARCH model, and to \( \beta_0 = (1, 1)' \) for the Poisson-N simulations. However some of the departures from the null hypothesis analyzed for the ARCH model have no consequence if \( \beta_0 = 0 \) (e.g. \( h_t = \alpha_0 + \alpha_1 y_{t-1}^2 \) and \( h_t = \alpha_0 + \alpha_1 u_{t-1}^2 \) are identical), and a true parameter vector \( \beta_0 = (1, 1) \) was also considered for some ARCH estimation experiments, and for the simulations to assess the power of the consistency tests presented in
Chapter 5. The additional regressor $x_{2t}$, whenever included, was assigned a coefficient of unity.

There are no $\alpha$ parameters in the Poisson-N model, and in an ARCH(1) process the strength of the conditional variance effect is directly related to the ratio $r = \alpha_1 / \alpha_0$. We considered a mild effect ($r = 0.25$), a strong effect ($r = 4$), and a regular effect ($r = 1$). For the mild ARCH effect the parameters were set at $\alpha_0 = 0.8$ and $\alpha_1 = 0.2$, so that $y_t$ would possess moments of order eight. This DGP is referred to as ARCH I in the Monte Carlo sections and tables, or simply as the 'mild model'. For the strong ARCH effect the parameters were set at $\alpha_0 = 0.2$ and $\alpha_1 = 0.8$, so that $y_t$ would not possess moments of order four. This DGP is referred to as ARCH II in the Monte Carlo sections and tables, or simply as the 'strong model'. For the regular ARCH effect the parameters were set at $\alpha_0 = \alpha_1 = 0.5$, so that $y_t$ would possess moments of order four but not of order six. This DGP is referred to as ARCH III in the Monte Carlo sections and tables, or simply as the 'regular model'. An additional ARCH DGP (referred to as ARCH IV in the tables) was a second order ARCH process with parameter values of $\alpha_0 = 0.5$, $\alpha_1 = 0.3$, and $\alpha_2 = 0.2$, so that $\alpha_0 = \alpha_1 + \alpha_2 = 0.5$, and the strength of the conditional variance effect would be similar to that in the regular model. The parameters of the conditional variance of other DGP's used in the experiments to induce misspecification in $h_t$ were assigned in similar form and are detailed in the tables reporting the simulation results (e.g. for $h_t = \alpha_0 + \alpha_1 y_{t-1}^2$ the same parameter values of the corresponding ARCH process were used).

Most DGP's used were conditionally normal, and in such case the mean innovations $u_t$ were calculated as $u_t = h_t^{1/2} \xi_t$, where the $\xi_t$ were generated as $N I I [0, 1]$. To analyze the power of the efficiency tests presented in Chapter 5 we also considered generating the $\xi_t$ as independent drawings from the following distributions: uniform in (-1, 1), Student's t with fifteen degrees of freedom, Student's t with five degrees of freedom, $\chi^2$ with two degrees of
freedom, beta with parameters (0.5, 0.5), lognormal with parameters (0, 1), and Cauchy with parameters (0, 1). In these cases $u_t$ was calculated as $u_t = h_t^{1/2} \left[ \xi_t - \bar{\xi} \right]$, where $\bar{\xi} = \mathbb{E} [ \xi_t ]$ for distributions other than the Cauchy, in which case $\bar{\xi}$ was the median of the distribution.
CHAPTER 3

ESTIMATION AND LIKELIHOOD FACTORIZATION

Our concern in this chapter is the estimation of the parameters in the model presented in Chapter 2. We concentrate initially on the extraction of information about \( \theta \) from each conditional moment separately using orthogonality conditions which arise from the specification of the conditional moments and some efficiency considerations. Later in the chapter conditional normality is assumed and maximum likelihood (ML) estimation undertaken.

This organization serves several purposes. First, it separates clearly the different sources of information that contribute to the overall process of inference. A complicated problem is broken down into two simpler ones and this may be helpful for model specification. Second, proofs are greatly simplified by referring to Hansen [1982] and to White and Domowitz [1984] using Theorem 2.1 and Propositions 2.2 and 2.3. Many of our propositions are given in a more general environment than usually found in the literature on heteroskedastic models and are simple corollaries to the GMM theory of estimation. Finally, the extent to which the properties of different estimators depend on the normality assumption is made evident. In the classical linear model the ML estimator (MLE) under normality is best linear unbiased even when normality does not hold (e.g. Dhrymes [1970], Theil [1971]). In the more general framework of a nonlinear heteroskedastic model the properties of the MLE under normality (which is referred to as the quasi-MLE or QMLE when normality is not assumed) are similar to those obtained in the linear homoskedastic case. The QMLE becomes asymptotically a matrix weighted average (MWA) of two independent Gauss-Markov type estimators. Besides,
the GMM approach is suitable for problems not restricted to conditional normality, or even to symmetry. This will be exploited later in Chapter 10.

Throughout the chapter we assume 'correct specification' in the form of \((Q0)\) or \((Q0')\), and that the set of regularity conditions given by \((Q1) - (Q7)\) or \((Q1) - (Q8)\) are fulfilled. We also assume the conditional distribution of \(y_t\) to be symmetric.

The organization of the chapter follows. In section § 3.1 we discuss estimation using the information provided by the conditional mean, in section § 3.2 we discuss estimation using the information provided by the conditional variance, and in section § 3.3 we use all the information, first combining the orthogonality conditions and later considering ML estimation, and show the two estimators to be asymptotically equivalent under normality. The joint estimator is given a MWA interpretation, where the conditional moments are the separate sources of information. In section § 3.4 we suggest ways of measuring the relative contributions to efficiency of each conditional moment. Finally, in section § 3.5 we discuss some Monte Carlo evidence on the performance of several of the estimators reviewed in the chapter.

§ 3.1 Extracting information from the conditional mean

In this section we consider estimating the parameter vector \(\beta\) using only the information contributed by the mean equation

\[
y_t = \mu_t(\beta) + u_t.
\]  

-(1)

To extract the information from the mean innovations \(u_t\) we consider orthogonality conditions of the form \(T^{-1} \sum_{t=1}^{T} f_t u_t\), for some vector \(f_t \in \mathcal{F}_t\) of dimension greater or equal to \(k\). Since the martingale sequence \(u_t\) has conditional variance \(h_t\), the efficiency benchmark for the estimation of \(\beta\) in (1)
is obtained setting \( f_t = h_t^{-1} x_t \) (White [1984], Chapter 4). Assuming that the conditional variance \( h_t \) is known this produces the generalized least squares (GLS) 'estimator' (see for example Amemiya [1985]: 181-185). Denote this 'estimator' by \( \hat{\beta}_m^* \). Thus

\[
\hat{\beta}_m^* = \min_{\beta} \sum_{t=1}^{T} h_t^{-1} u_t^2 = \min_{\beta} u' \Omega^{-1} u,
\]

for known \( h_t \). The asymptotic properties of \( \hat{\beta}_m^* \) are well known: it is a strongly consistent estimator of \( \beta_0 \) and has asymptotic distribution given by

\[
T^{1/2} (\hat{\beta}_m^* - \beta_0) = T^{1/2} (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} u + o_p(1)
\]

\[\xrightarrow{d} N \left[ 0 , \Omega^{-1} \right]. \tag{2}\]

It is convenient to represent \( \hat{\beta}_m^* \) as a GMM estimator with orthogonality conditions

\[
\psi_m^*(\beta) = T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t = T^{-1} X' \Omega^{-1} u. \tag{3}\]

Since there are as many orthogonality conditions as parameters, the definition of a weighting matrix is not necessary. The distribution of \( \hat{\beta}_m^* \) follows from Theorem 2.1 and Propositions 2.2 and 2.3 by observing that \( E [ \psi_m^*(\beta_0) ] = 0 \), and \( E [ T \psi_m^*(\beta_0) \psi_m^*(\beta_0)' ] = -E [ \partial \psi_m^*(\beta_0) / \partial \beta' ] = E [ T^{-1} X' \Omega^{-1} X ]. \) Most proofs of the asymptotic distribution of \( \hat{\beta}_m^* \) found in the literature on heteroskedastic models are given under more restrictive conditions. For example, Jobson and Fuller [1980] and Carroll and Ruppert [1982b] consider the case of a linear mean and independent observations.

In the remaining of this section we consider the more realistic case in which \( h_t \) is not known. In § 3.1.1 we consider the parametric approach and in § 3.1.2 we briefly discuss semi-parametric approaches.
§ 3.1.1 The parametric approach

Suppose \( h_t = h_t(\theta) \) is known except for the values of the parameter vector \( \theta \), and we have a root-\( T \) consistent estimator \( \tilde{\theta} \) of the true value \( \theta_0 \) \(^{(1)}\). Let \( \tilde{h}_t = h_t(\tilde{\theta}) \) and \( \tilde{\Omega} = \Omega(\tilde{\theta}) \) and define the estimator \( \tilde{\beta}_m \) as the feasible GLS estimator

\[
\tilde{\beta}_m = \min_{\beta} \sum_{t=1}^{T} \tilde{h}_t^{-1} u_t^2 = \min_{\beta} \tilde{\Omega}^{-1} u , \tag{4}
\]
or equivalently as the GMM estimator with orthogonality conditions

\[
\psi_m(\beta) = T^{-1} \sum_{t=1}^{T} \tilde{h}_t^{-1} x_t u_t = T^{-1} X' \tilde{\Omega}^{-1} u . \tag{5}
\]

Carroll and Ruppert [1982b] have shown that in the case of linear \( \mu_t \) and independent observations \( \tilde{\beta}_m \) has the same asymptotic distribution as \( \tilde{\beta}_m^* \). For the more general case we prove

**Lemma 3.1.** - Under the assumptions \((C0) - (C7)\), and \( \tilde{\theta} \) being a root-\( T \) consistent estimator of \( \theta_0 \), then

(i) (a) \( \psi_m(\beta) - \psi_m^*(\beta) \xrightarrow{a.s.} 0 \) uniformly in \( \beta \), and

(b) \( T^{1/2} [ \psi_m(\beta_0) - \psi_m^*(\beta_0) ] \xrightarrow{a.s.} 0 \),

and

(ii) (a) \( T [ \psi_m(\beta_0) \psi_m(\beta_0)' - \psi_m^*(\beta_0) \psi_m^*(\beta_0)' ] \xrightarrow{a.s.} 0 \), and

(b) \( \frac{\partial \psi_m(\beta_0)}{\partial \beta'} - \frac{\partial \psi_m^*(\beta_0)}{\partial \beta'} \xrightarrow{a.s.} 0 \).

**Proof:** By the Mean Value Theorem for random functions (MVT) (Jennrich [1969]) we can express \( h_t(\tilde{\theta})^{-1} = h_t(\theta_0)^{-1} + [ \partial h_t(\tilde{\theta}) / \partial \theta ] ( \tilde{\theta} - \theta_0 ) \), where \( \tilde{\theta} \in [\tilde{\theta}, \theta_0] \), and so we have that

\[
\psi_m(\beta) = \psi_m^*(\beta) + \Psi_{\theta}(\tilde{\theta}; \beta) ( \tilde{\theta} - \theta_0 ) , \tag{6}
\]

\(^{(1)}\) \( \tilde{\theta} \) is root-\( T \) consistent if it is strongly consistent and \( \tilde{\theta} - \theta_0 \) is \( O_p(T^{-1/2}) \).
where \( \psi_\theta(\beta) = \frac{1}{T} \sum_{t=1}^{T} u_t x_t [ \partial h_t^{-1}(\theta) / \partial \theta ] \). From (4.3) - (4.5) the sequence \( u_t x_t \partial h_t^{-1}/\partial \theta \) obeys a Strong Law of Large Numbers uniformly in \( \Theta \), and therefore \( \psi_\theta(\beta) \) is \( O_p(1) \) uniformly in \( \Theta \), a.e. in \( \Theta \). Moreover, at \( \beta = \beta_0 \) the sequence has zero mean and obeys a Central Limit Theorem and so \( T^{1/2} \psi_\theta(\beta; \beta_0) \) is \( O_p(1) \) a.e. in \( \Theta \). Hence the strong consistency of \( \tilde{\theta} \) ensures that the second term in (6) is \( o_p(1) \) uniformly in \( \Theta \), and that it is \( o_p(T^{1/2}) \) at \( \beta_0 \). This establishes (i).

Forming \( T \psi_m(\beta_0) \psi_m(\beta_0)' \) with (6) and using the fact that \( T^{1/2} \psi_\theta(\beta; \beta_0) \) is \( O_p(1) \) a.e. in \( \Theta \) results in

\[
T \psi_m(\beta_0) \psi_m(\beta_0)' = T \psi_m^*(\beta_0) \psi_m^*(\beta_0)' + o_p(1),
\]

which establishes (ii) (a). The use of iterated expectations in the derivative of (3) shows that \( E \left[ \frac{\partial \psi_m^*(\beta_0)}{\partial \beta'} \right] = -E \left[ T \psi_m^*(\beta_0) \psi_m^*(\beta_0)' \right] \), while from (5) we get

\[
E \left[ \frac{\partial \psi_m(\beta_0)}{\partial \beta'} \right] = -E \left[ T \psi_m(\beta_0) \psi_m(\beta_0)' \right] + E \left[ T^{-1} \sum_{t=1}^{T} h_t^{-1} u_t \frac{\partial x_t}{\partial \beta'} \right],
\]

and a simple expansion of \( h_t^{-1} \) around \( \theta_0 \) and repeating the argument below (6) shows that the second term converges to zero. This completes the proof.

Condition (i) (a) ensures that \( \beta_m^* - \beta_0 \xrightarrow{as} 0 \), condition (i) (b) establishes that \( T^{1/2} (\beta_m^* - \beta_0) = T^{1/2} (\beta_m - \beta_0) + o_p(1) \), and (ii) guarantees that both covariance matrices converge to the same limit. Therefore, using (2),

\[
T^{1/2} (\beta_m - \beta_0) = T^{1/2} (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} u + o_p(1)
\]

\[
\xrightarrow{d} N \left[ 0, \mathcal{S} \{ T^{-1} X' \Omega^{-1} X \} \right].
\]

The crucial assumption here is that of a correctly specified conditional variance and so it is worth stressing the need for a careful assessment of the specification of \( h_t \). If this is the case then the availability of \( \tilde{\theta} \) to compute \( \beta_m \) is ensured, as we shall see in § 3.2.
§ 3.1.2 A note on semi-parametric approaches

The argument that Economic Theory rarely provides propositions about the variance calls for an attempt to make inferences about the mean parameters $\beta$ without imposing a parametric structure on the variance. The least-squares (LS) estimator does not require knowledge of $h_t$ for its computation and constitutes the simplest semi-parametric estimator of $\beta$ under heteroskedasticity. It is well known nevertheless that its covariance matrix is incorrectly estimated from LS output and this may result in incorrect inferences about the parameters. The LS estimator $\hat{\beta}_L$ is obtained from the orthogonality conditions $\psi(L) = T^{-1}X'u$. With iterated expectations we can see that $E[\psi(L)] = 0$, $E[T\psi(L)\psi(L)'] = E[T^{-1}X'\Omega X]$, and $E[\partial\psi(L)/\partial\beta'] = -E[T^{-1}X'X]$. From Theorem 2.1 and Propositions 2.2 and 2.3, $\hat{\beta}_L$ is strongly consistent for $\beta_0$ with asymptotic distribution given by

$$T^{1/2}(\hat{\beta}_L - \beta_0) = T^{1/2}(X'X)^{-1}X'u + o_p(1) \xrightarrow{d} N[0, \Omega(X'X)^{-1}X'\Omega X(X'X)^{-1}],$$

which shows that $V(\hat{\beta}_L)$ depends on the form of the heteroskedasticity.

Eicker [1967] and more recently White [1980b] have produced an estimated covariance matrix robust to heteroskedasticity of unknown form. White's principle has by now been extended to the more general settings of lagged dependent regressors (Nicholls and Pagan [1983]), misspecified general models (White [1980a], White [1982a], Domowitz and White [1982] *inter alia*), and nonlinear models, instrumental variables and GMM estimation (White and Domowitz [1984], Hansen [1982], Burguete *et al.* [1982], White [1982b, 1984], Chamberlain [1987], Duncan [1987], *inter alia*). With this $V(\hat{\beta}_L)$ is consistently estimated by

$$\hat{V}(\hat{\beta}_L) = T(\hat{X}'\hat{X})^{-1}\hat{X}'\hat{U}^2\hat{X}(\hat{X}'\hat{X})^{-1},$$

(2) The operator $V(\cdot)$ denotes the covariance matrix of the asymptotic distribution.
where $\hat{X} = X(\hat{\beta}_L)$ and $\hat{U} = \text{diag} \{ u_1(\hat{\beta}_L), \ldots, u_T(\hat{\beta}_L) \}$.

It is well known that $V(\hat{\beta}_L) - V(\hat{\beta}_m)$ is positive semidefinite (psd) (see for example Amemiya [1985]: 182-183, who also provides conditions under which $V(\hat{\beta}_L) = V(\hat{\beta}_m)$). If the data generating process (DGP) is only conditionally heteroskedastic but unconditionally homoskedastic $E[h_t] = \sigma^2$ for all $t$, and in this case the covariance matrix of $\hat{\beta}_L$ reduces to

$$V(\hat{\beta}_L) = \sigma^2 \mathcal{C} \{ T^{-1} X' X \}^{-1}.$$  \hspace{1cm} -(8)

There does not seem to be in the literature a comparison of $V(\hat{\beta}_L)$ and $V(\hat{\beta}_m)$ in these circumstances. In fact, $\hat{\beta}_m$ remains efficient relative to $\hat{\beta}_L$ as shown in

**Lemma 3.2.-** Under the assumptions (C1)-(C7) and $E[h_t] = \sigma^2$ for all $t$,

$V(\hat{\beta}_L) - V(\hat{\beta}_m)$ is psd.

**Proof:** $V(\hat{\beta}_L) - V(\hat{\beta}_m)$ is psd $\iff V(\hat{\beta}_m)^{-1} - V(\hat{\beta}_L)^{-1}$ is psd. But

$$V(\hat{\beta}_m)^{-1} - V(\hat{\beta}_L)^{-1} = \mathcal{C} \{ T^{-1} X' \Omega^{-1} X - \sigma^{-2} T^{-1} X' X \}$$

$$= \mathcal{C} \{ T^{-1} X' (\Omega^{-1} - \sigma^{-2} I_T) X \}$$

$$= \mathcal{C} \{ T^{-1} \sum_{t=1}^T (h_t^{-1} - \sigma^{-2}) x_t x_t' \},$$

and since $\sigma^{-2} = E[h_t]^{-1}$ and $h_t^{-1}$ is a convex function of $h_t$ it follows from Jensen's inequality (White [1984], Proposition 2.38) that $E[h_t^{-1} - \sigma^{-2}] \geq 0$, so that the last expectation is psd.

This result differs from the usual argument for the efficiency of GLS over LS which assumes unconditional heteroskedasticity. It is interesting because it illustrates the possible gains from using conditional information which is lost unconditionally, and in fact Engle's [1982a] argument for ARCH models was made along these lines.
A simplified covariance matrix for $\hat{\beta}_2$ of the form in (8) may occur in other circumstances apart from unconditional homoskedasticity. Cragg [1983] has noted that when the heteroskedasticity is unrelated to the variability of the regressors in the sense that

$$\sum_{t=1}^{T} (h_t - \bar{h}) x_t x_t' = 0,$$

where $\bar{h} = \frac{1}{T} \sum_{t=1}^{T} h_t$, then

$$V(\hat{\beta}_2) = \mathcal{C} (T^{-1}XX)' (\Omega - \bar{h} I_T) X (XX)' \mathcal{C} (T^{-1}XX)^{-1}.$$

There is no advantage in using the heteroskedasticity-robust covariance matrix in this case, and the biases of $\hat{V}(\hat{\beta}_2)$ in small samples suggest that care must be exercised (see Chesher and Jewitt [1987]).

The next step in the semi-parametric estimation of $\beta$ is to try to improve upon the efficiency of $\hat{\beta}_2$. Amemiya [1983] and Cragg [1983] have proposed the partially generalized least squares (PGLS) estimator (see also Chamberlain [1982, 1987] and White [1982b]), defined by the orthogonality conditions

$$\psi_p(\beta) = T^{-1} \sum_{t=1}^{T} q_t u_t = T^{-1} Q' u,$$

where $q_t$ includes the derivatives of $\mu_t$ with respect to $\beta$, $q_t = (x_t', p_t')'$ for some $p_t \in \mathcal{F}_t$ of fixed dimension. Using iterated expectations it is seen that

$$E [ T \psi_p(\beta_0) \psi_p(\beta_0)' ] = E [ T^{-1} Q' \Omega Q ],$$

and that $E [ \partial \psi_p(\beta_0) / \partial \beta' ] = - E [ T^{-1} Q' X ]$, and it follows from Theorem 2.1 and Propositions 2.2 and 2.3 that

$$T^{1/2} ( \hat{\beta}_p - \beta_0 ) \overset{d}{\to} N [ 0, \mathcal{C} (T^{-1} X'Q (Q' \Omega Q)^{-1} Q'X)^{-1} ].$$

Consistent estimation of $V(\hat{\beta}_p)$ is obtained by evaluating all functions at $\hat{\beta}_p$ and substituting $\Omega$ by $\hat{U}^2$. Note that although Amemiya’s proposal for $p_t$ is not of a measurable function of $\mathcal{F}_t$, this has no effect on the asymptotic distribution of $\hat{\beta}_p$. Both Amemiya and Cragg show that PGLS has intermediate efficiency between LS and GLS, and Balestra [1983] provides
conditions under which PGLS and GLS are equivalent. Cragg provides a sampling experiment which effectively produces efficiency gains, but the problem of selecting the additional instruments $p_t$ is not completely resolved and requires further research.

An alternative approach to improve efficiency upon LS is feasible GLS as in (4), using non-parametric estimates $\tilde{h}_t$ of $h_t$. Different forms of obtaining such non-parametric estimates have been suggested in the literature (see for example Fuller and Rao [1978] and the references therein) but most of them produce a semi-parametric estimator of $\beta$ which is inefficient relative to $\hat{\beta}_m$. Carroll [1982] and Robinson [1987], however, have provided non-parametric estimates of the conditional variance such that the semi-parametric estimator of $\beta$ attains full efficiency with respect to the information in the conditional mean (i.e. is as efficient as $\hat{\beta}_m$). Carroll uses kernel estimation for $h_t$, while Robinson uses non-parametric nearest neighbor regression and produces more general results. These efficient semi-parametric estimators have not yet attained a high level of generality and their behavior in small samples is not known. Still with some more research they may prove an important tool for specification and estimation in heteroskedastic models since they permit an almost perfect separation of the problems of extracting information from the conditional mean and from the conditional variance. This point becomes evident after the next two sections.

§ 3.2 Extracting information from the conditional variance

In this section we concentrate on the estimation of the parameter vector $\theta$ using only the information contributed by the variance equation

$$u_t^2 = h_t(\theta) + \epsilon_t.$$ - (9)
Since we are assuming that the distribution of $y_t$ is conditionally symmetric it follows that $E [u_t \varepsilon_t] = 0$ and hence the mean and variance equations represent unrelated sources of information. In principle the estimation problem for $\theta$ in the variance equation is of the same form as the estimation of $\beta$ in the mean equation, but there are some major differences:

(i) the positivity of $u_t^2$ implies a changing support for $\varepsilon_t$,

(ii) $u_t^2$ is not observable whereas $y_t$ is, and

(iii) there may be problems associated to the identifiability of $\theta$ in the variance equation, whereas $\beta$ is identifiable by assumption in the mean equation.

Ignoring the changing support for $\varepsilon_t$ may result in negative predictors for $u_t^2$. But this does not prevent us from obtaining the proper asymptotic distribution of GMM estimators from orthogonality conditions in $\varepsilon_t$ because these are proper innovations under (Q0). The positivity of $h_t$ must be ensured by restrictions on $\Theta$ and/or $h_t$, and in all well known models this can be achieved by reparameterization. When the form of the conditional pdf is known, this prevents us from producing negative predictors. For example, under conditional normality $u_t^2$ is conditionally gamma distributed.

The unobservability of $u_t^2$ can be solved if $\beta$ can be estimated consistently in the mean equation by $\tilde{\beta}$, say. The squared residuals $\tilde{u}_t^2 = [y_t - \mu_t(\tilde{\beta})]^2$ can then be used in place of $u_t^2$, producing an alternative form for the variance equation,

$$\tilde{u}_t^2 = h_t(\theta) + e_t,$$

where $e_t = \varepsilon_t + (\tilde{u}_t^2 - u_t^2)$. The following Lemma generalizes some of the results in the Appendix of Pagan and Hall [1983]:

**Lemma 3.3.** Given (Q0) - (Q7) and
(a) the existence of moments of $y_t$ up to order $r$,

(b) a root-$T$ consistent estimator $\tilde{\beta}$ of $\beta$, and

(c) a measurable function $f_t(\pi)$ of $\mathcal{F}_t$, where $\pi$ is a parameter vector for which a root-$T$ consistent estimator $\tilde{\pi}$ is available,

then for all positive integer $r$,

(i) $T^{-1} \sum_{t=1}^{T} f_t(\tilde{u}_t^r - u_t^r) \xrightarrow{as} 0$,

and

(ii) $T^{-1/2} \sum_{t=1}^{T} f_t(\tilde{u}_t^2 - u_t^2) \overset{d}{\to} N[0, r^2 \mathbb{E}(T^{-1} X' \Omega_r X) V(\beta) \mathbb{E}(T^{-1} X' \Omega_r F)]$.

For $r = 2$ we have that

(iii) $T^{-1} \sum_{t=1}^{T} f_t(\tilde{u}_t^2 - u_t^2) \xrightarrow{as} 0$,

and moreover, if the conditional distribution of $y_t$ is symmetric then

(iv) $T^{-1/2} \sum_{t=1}^{T} f_t(\tilde{u}_t^2 - u_t^2) \xrightarrow{as} 0$ for $r$ even,

where $\tilde{u}_t = y_t - \mu_\pi(\tilde{\beta}) = y_t - \mu_\beta$, $\Omega_s = \text{diag} \{ \mathbb{E}[u_t^s | \mathcal{F}_t] \}$ for $s \leq r$, $\tilde{f}_t = f(\tilde{\pi})$, and $F = (f_1, \ldots, f_T)'$.

Proof: We have that $\tilde{u}_t = u_t - (\mu_t - \mu_\beta)$, and so

$$\tilde{u}_t^r = \sum_{j=0}^{r} \binom{r}{j} (-1)^j (\mu_t - \mu_\beta)^j u_t^{r-j} = u_t^r + \sum_{j=1}^{r} \binom{r}{j} (-1)^j (\mu_t - \mu_\beta)^j u_t^{r-j}.$$ 

Therefore,

$$T \sum_{t=1}^{T} \tilde{f}_t(\tilde{u}_t - u_t) = T \sum_{j=1}^{r} \binom{r}{j} \sum_{t=1}^{T} \tilde{f}_t(\mu_t - \mu_\beta)^j u_t^{0} = \sum_{k=1}^{r} \binom{r}{j} (-1)^j \tilde{\xi}(j),$$

where $\tilde{\xi}(j) = \sum_{t=1}^{T} \tilde{f}_t(\mu_t - \mu_\beta)^j u_t^{r-j}$. We can express $\mu_t - \mu_\beta = x_t(\beta)'(\beta - \beta_0)$, using the MVT, where $\beta \in [\beta_0, \beta]$. Then

$$\tilde{\xi}(j) = T \sum_{t=1}^{T} \tilde{f}_t(\tilde{u}_t - u_t) u_t^{r-j} = \left( \sum_{t=1}^{T} \tilde{f}_t(\mu_t - \mu_\beta)^j u_t^{r-j} x_t(\beta)'(\beta - \beta_0) \right)(\beta - \beta_0).$$
and from (a) and the boundedness of $x_t$ in (Q3), the term in brackets is of the same order of magnitude as $\sum_{t=1}^{T} f_t (\tilde{\mu}_t - \mu_t)^{-1} u_t^{(j-1)}$, $j = 1, \ldots, r - 1$, so that from the consistency of $\tilde{\beta}$ it follows that $O_p[\xi(j)] = O_p[T^{-1/2} \xi(j-1)]$, $j = 2, \ldots, r$, and $\xi(1)$ dominates $\xi(j)$ for $j > 1$. Because $\xi(1)$ is $O_p(T^{1/2})$, we have that

$$\sum_{t=1}^{T} f_t (\tilde{u}_t - u_t^r) = -r \sum_{t=1}^{T} f_t (\tilde{\mu}_t - \mu_t) u_t^{r-1} + O_p(1) = -r \sum_{t=1}^{T} f_t u_t^{r-1} x_t'(\tilde{\beta} - \beta_0) + O_p(1),$$

and using (a) and (c) we obtain $T^{-1} \sum_{t=1}^{T} f_t u_t^{r-1} x_t' \xrightarrow{d} N[0, r^2 \mathbb{E} \{ T^{-1} \mathbb{E}^{\prime} \Omega_{r-1} X\}]$, so that (i) follows from (b), and

$$T^{-1/2} \sum_{t=1}^{T} f_t (\tilde{u}_t - u_t^r) = -r (T^{-1} \sum_{t=1}^{T} f_t u_t^{r-1} x_t') T^{1/2} (\tilde{\beta} - \beta_0) + O_p(T^{-1/2})$$

$$\xrightarrow{d} N[0, r^2 \mathbb{E} \{ T^{-1} \mathbb{E}^{\prime} \Omega_{r-1} X\} V(\tilde{\beta}) \mathbb{E} \{ T^{-1} X^{\prime} \Omega_{r-1} \mathbb{E} \}],$$

proving (ii). Also, since $\mathbb{E}[u_t | \mathcal{F}_t] = 0$ we have that $\Omega_1 = 0$ and hence the covariance matrix is null for $r = 2$, which establishes (iii). Finally, when the distribution is symmetric $\Omega_{r-1} = 0$ for $r$ even and this completes the proof.

Therefore using $\tilde{u}_t^2$ in place of $u_t^2$ in (9) is innocuous asymptotically to the order of $T^{1/2}$ and (10) is the 'operative' version of the variance equation.

The issue of identifiability of $\theta$ in the variance equation now follows. First note that (Q7) ensures that $\alpha$ is identifiable in the variance equation given $\beta$. But the identifiability of the whole vector $\theta$ depends on the parameterization of $h_t$. Consider for example the linear Amemiya model in which $h_t = \alpha (x_t' \beta)^2$, where $\theta$ is not identifiable because $h_t = (x_t' \beta^*)^2$ for $\beta^* = \alpha^{1/2} \beta$ provides an observationally equivalent parameterization (Fisher [1966]). As a contrast, consider the ARCH class of models under normality, in which

$$\mathbb{E} \{ \sum_{t=1}^{T} h_t^2 w_t w_t^{\prime} \} = 4 \mathbb{E} \{ \sum_{t=1}^{T} h_t^2 \sum_{i=1}^{q} \alpha_i \alpha_j u_{t-i} u_{t-j} x_{t-i} x_{t-j} \}$$

$$= 4 \sum_{j=1}^{q} \alpha_j^2 \mathbb{E} \{ \sum_{t=1}^{T} h_t^2 u_{t,j} x_{t,j} x_{t,j}^{\prime} \},$$
where the expectations for $i \neq j$ vanish because these terms are conditionally odd functions of $u_{t-j}$ (see Lemma 3 of Pagan and Sabau [1987a]). It follows that $T^{-1} W' \Omega^{-2} W$ is uniformly positive definite in an open neighborhood of $\theta_0$, and so is $T^{-1} Z' \Omega^{-2} Z$ in view of (37). Engle [1982a] has shown that $T^{-1} W' \Omega^{-2} W \rightarrow 0$, and therefore $T^{-1} S' \Omega^{-2} S - \text{diag} \{ T^{-1} W' \Omega^{-2} W, T^{-1} Z' \Omega^{-2} Z \} \rightarrow 0$, and it follows that $T^{-1} S' \Omega^{-2} S$ is uniformly positive definite in an open neighborhood of $\theta_0$, and hence that $\theta$ is identifiable in the variance equation. In general, however, only a function

$$\phi = \phi(\theta) \in \Phi \subseteq \mathbb{R}^{p*}, \ 0 \leq p^* \leq p,$$

is identifiable in the variance equation. We assume that $p^*$ is the maximum number of identifiable functions of $\theta$ and that the function $\phi$ is continuous in $\Theta$. When $\theta$ is identifiable we have $\phi = \theta$.

After these considerations we can exploit the similarities between the mean and the variance equations. Since $E[\varepsilon_t^2 | \mathcal{F}_t] = \kappa_t$ the efficiency benchmark for the estimation of $\phi$ is the GLS 'estimator' $\hat{\phi}_v$ assuming the $\kappa_t$ known, that is,

$$\hat{\phi}_v^* = \min_{\phi} \sum_{t=1}^{T} \kappa_t^{-1} \varepsilon_t^2 = \min_{\phi} \varepsilon' K^{-1} \varepsilon,$$

with the equation reparameterized in terms of the identifiable functions $\phi$. Equivalently, $\hat{\phi}_v^*$ may be seen as a GMM 'estimator' with orthogonality conditions

$$\psi_v^*(\phi) = T^{-1} \sum_{t=1}^{T} \kappa_t^{-1} s_{\phi t} \varepsilon_t = T^{-1} S_\phi' K^{-1} \varepsilon,$$

where $s_{\phi t} = \partial h_t / \partial \phi$, and $S_\phi = (s_{\phi 1}, ..., s_{\phi T})'$. Using iterated expectations it can be seen that $E[\psi_v^*(\phi_0)] = 0$, and also that $E[T \psi_v^*(\phi_0) \psi_v^*(\phi_0)'] = -E[\partial \psi_v^*(\phi_0) / \partial \phi'] = E[T^{-1} S_\phi' K^{-1} S_\phi]$, where $\phi_0 = \phi(\theta_0)$. The latter matrix is uniformly positive definite in an open neighborhood of $\phi_0$ because $\phi$ is identifiable by assumption. It follows from Theorem 2.1 and Propositions 2.2 and 2.3 that $\hat{\phi}_v^*$ is a strongly
consistent estimator of \( \phi_0 \) and has asymptotic distribution

\[
T^{1/2}(\hat{\phi}_v - \phi_0) = T^{1/2}(S_\phi' K^{-1} S_\phi)^{-1} S_\phi' K^{-1} \varepsilon + o_p(1)
\]

\[
\xrightarrow{d} N\left[0, \mathcal{C}\left(T^{-1} S_\phi' K^{-1} S_\phi\right)^{-1}\right],
\]

- (11)

where the expectation is evaluated at \( \phi = \phi_0 \).

When more information is available on the conditional distribution a specific form can be given to the kurtosis factors \( \kappa_t \). For example, if the conditional distribution is normal then \( \kappa_t = 3 h_t^2 - h_t^2 = 2 h_t^2 \), so that \( K = 2 \Omega^2 \), and \( V(\phi_v^*) = 2 \mathcal{C}\left(T^{-1} S_\phi' \Omega^{-2} S_\phi\right)^{-1} \), and if the conditional distribution is Student's t with \( \pi \) degrees of freedom (Bollerslev [1985], Engle and Bollerslev [1986]) then \( \kappa_t = \left[3(\pi - 2)/(\pi - 4) - 1\right] h_t^2 \) for \( \pi > 4 \) and so \( K = \left[3(\pi - 2)/(\pi - 4) - 1\right] \Omega^2 \), and \( V(\phi_v^*) = \left[3(\pi - 2)/(\pi - 4) - 1\right] \mathcal{C}\left(T^{-1} S_\phi' \Omega^{-2} S_\phi\right)^{-1} \).

When \( \theta \) is identifiable in the variance equation such as in the ARCH case then \( \phi_v^* = \hat{\theta}_v^* \), say, and

\[
T^{1/2}(\hat{\theta}_v^* - \theta_0) = T^{1/2}(S' K^{-1} S)^{-1} S' K^{-1} \varepsilon + o_p(1)
\]

\[
\xrightarrow{d} N\left[0, \mathcal{C}\left(T^{-1} S' K^{-1} S\right)^{-1}\right].
\]

- (12a)

Partitioning \( \hat{\theta}_v = \left(\hat{\beta}_v^*, \hat{\alpha}_v^*\right)' \) and using partitioned inversion we get

\[
T^{1/2}(\hat{\beta}_v^* - \beta_0) = T^{1/2}(W' K^{-1/2} Q_z K^{-1/2} W)^{-1} W' K^{-1/2} Q_z K^{-1/2} \varepsilon + o_p(1)
\]

\[
\xrightarrow{d} N\left[0, \mathcal{C}\left(T^{-1} W' K^{-1/2} Q_z K^{-1/2} W\right)^{-1}\right],
\]

- (12b)

and

\[
T^{1/2}(\hat{\alpha}_v^* - \alpha_0) = T^{1/2}(Z' K^{-1/2} Q_w K^{-1/2} Z)^{-1} Z' K^{-1/2} Q_w K^{-1/2} \varepsilon + o_p(1)
\]

\[
\xrightarrow{d} N\left[0, \mathcal{C}\left(T^{-1} Z' K^{-1/2} Q_w K^{-1/2} Z\right)^{-1}\right],
\]

- (12c)

where \( Q_z \) and \( Q_w \) are the projection matrices, respectively, onto the spaces orthogonal to the ranges of \( K^{-1/2} Z \) and \( K^{-1/2} W \).
§ 3.2.1 Two-stage estimators for $\alpha$

The idea of estimating all the identifiable functions in the variance equation has seldom been considered. Exceptions are Jobson and Fuller [1980] and Amemiya [1977], though the latter paper only deals with a simple heteroskedastic model. The emphasis has been on ML estimation of $\theta$, and the variance equation on its own has mainly been considered to get a two-stage initial estimator of $\alpha$ given one of $\beta$, say $\tilde{\beta}$ (see for example Pagan [1984a], Engle [1982a, 1983], Weiss [1984, 1986a]).

In this subsection we consider two alternative two-stage estimators for $\alpha$: the simple LS estimator (SLS) obtained from the regression of $\tilde{u}_t^2$ on $h_t(\tilde{\beta}, \alpha)$ (e.g. Jobson and Fuller [1980]) and the simple GLS estimator (SGLS) obtained from the weighted regression of $\tilde{u}_t^2$ on $h_t(\tilde{\beta}, \alpha)$ taking into account the heteroskedastic nature of $\epsilon_t$ (Amemiya [1977]). These estimators are attractive for their simplicity, especially in linear-in-$\alpha$ models for which the regressions are linear.

Two-stage estimators have been studied by Pagan [1984b, 1986], and in the case of conditional normality we can view the SGLS estimator $\tilde{\alpha}_g$, say, as a straight-forward application of these results. The pseudo-log-likelihood function for $\alpha$ given $\tilde{\beta}$ is then from (2.10)

$$\ell(\tilde{\beta}, \alpha) = -\frac{1}{2} \sum_{t=1}^{T} \log h_t(\tilde{\beta}, \alpha) - \frac{1}{2} \sum_{t=1}^{T} h_t(\tilde{\beta}, \alpha)^{-1} \tilde{u}_t^2,$$

which has the form of a log-likelihood from a gamma parent, $\Gamma\left[\frac{1}{2}, 2 h_t(\tilde{\beta}, \alpha)\right]$. This pseudo-log-likelihood is maximized by the SGLS estimator (Amemiya [1973]). Conditions (i) - (vi) of Theorems 1 and 3 of Pagan [1986] are fulfilled under ($\mathcal{G}0'$) - ($\mathcal{G}7$), provided that $\tilde{\beta}$ is a root-T consistent estimator of $\beta_0$, and hence $\tilde{\alpha}_g$ is a strongly consistent estimator of $\alpha_0$ with asymptotic distribution

$$T^{1/2} (\tilde{\alpha}_g - \alpha_0) \xrightarrow{d} N\left[0, V(\tilde{\alpha}_g)\right].$$

To obtain an expression for the covariance matrix we assume that $\tilde{\beta}$ uses only
information from the mean equation, so that it is asymptotically independent of functions of the $\varepsilon_t$. Then, using the score and information matrix in (2.11) and (2.12), respectively, we have that

$$\lim_{T \to \infty} \text{Cov} [ T^{1/2} (\hat{\beta} - \beta_0) , d_\alpha (\theta) ] \to 0 ,$$

and so

$$V(\hat{c}_g) = 2 \epsilon (T^{-1} Z' \Omega^{-2} Z)^{-1} + \epsilon (Z' \Omega^{-2} Z)^{-1} Z' \Omega^{-2} W) V(\hat{\beta}) \epsilon (W' \Omega^{-2} Z (Z' \Omega^{-2} Z)^{-1}) .$$

(14)

The more efficient the estimate $\hat{\beta}$ used from the first stage, the more efficient $\hat{c}_g$ is. But for practical purposes the LS estimator $\hat{\beta}_L$ is the interesting one to consider because $\hat{c}_g$ is needed before we can get $\hat{\beta}_m$. When the information matrix is block-diagonal between $\beta$ and $\alpha$ the second term in (14) vanishes and $\hat{c}_v$ and $\hat{c}_g$ have the same asymptotic distribution. This is Amemiya's [1977] result for the simple heteroskedasticity model and it extends to the ARCH class.

If conditional normality is relaxed while retaining symmetry, the distribution in (13) still holds but the covariance matrix in (14) is modified to

$$V(\hat{c}_g) = \epsilon (T^{-1} Z' K^{-1} Z)^{-1} + \epsilon (Z' K^{-1} Z) Z' K^{-1} W) V(\hat{\beta}) \epsilon (W' K^{-1} Z (Z' K^{-1} Z)^{-1}) .$$

For these more general cases one would need to estimate the conditional kurtosis implicit in $\kappa_t$. This could be done by parameterizing the fourth conditional moment or following a non-parametric approach.

The SLS estimator $\hat{\alpha}_s$, though less efficient, overcomes these problems and provides a simpler initial estimator of $\alpha$, which is obtained from the first order conditions

$$T^{-1} \sum_{t=1}^T z_t (\hat{\beta} , \hat{\alpha}_s) [ \hat{u}_t^2 - h_t (\hat{\beta} , \hat{\alpha}_s) ] = T^{-1} Z(\hat{\beta} , \hat{\alpha}_s)' [ \hat{u}_t^2 - h (\hat{\beta} , \hat{\alpha}_s) ] = 0 ,$$

(15)

and using the operative variance equation we get

$$Z(\hat{\beta} , \hat{\alpha}_s)' ( \varepsilon + (\hat{u}_t^2 - u_t^2) - (h (\hat{\beta} , \hat{\alpha}_s) - h (\beta_0 , \alpha_0)) ) = 0 .$$
From the MVT for random functions (Jennrich [1969]) we have

\[ h_t(\tilde{\beta}, \tilde{\alpha}_s) - h_t(\beta_0, \alpha_0) = \overline{w}_t'(\tilde{\beta} - \beta_0) + \overline{z}_t'(\tilde{\alpha}_s - \alpha_0), \]

where \( \overline{w}_t = w_t(\tilde{\beta}, \tilde{\alpha}) \) and \( \overline{z}_t = z_t(\tilde{\beta}, \tilde{\alpha}) \) for \( \tilde{\beta} \in [\tilde{\beta}, \beta_0] \) and \( \tilde{\alpha} \in [\tilde{\alpha}_s, \alpha_0] \), and substituting in (15) produces

\[ \hat{Z}'\hat{Z}(\hat{\alpha}_s - \alpha_0) = \hat{Z}'\varepsilon + \hat{Z}'(u^2 - u^2) - \hat{Z}'\overline{W}(\tilde{\beta} - \beta_0), \]  

where \( \hat{Z} = Z(\tilde{\beta}, \tilde{\alpha}_s) \), \( \overline{Z} = Z(\tilde{\beta}, \tilde{\alpha}) \), and \( \overline{W} = W(\tilde{\beta}, \tilde{\alpha}) \). Lemma 3.3 ensures that \( T^{-1/2}\hat{Z}'(u^2 - u^2) \xrightarrow{a.s.} 0 \), whereas \( T^{-1}\hat{Z}'\overline{W} \) converges a.s. and \( T^{-1}\hat{Z}'\overline{Z} \) is uniformly positive definite in an open neighborhood of \( \theta_0 \). Strong consistency follows since \( T^{-1}\hat{Z}'\varepsilon \xrightarrow{a.s.} 0 \). Further the central limit theory of White and Domowitz [1984] applies to \( T^{-1/2}\hat{Z}'\varepsilon \), which converges in distribution to \( N\{0, \mathbb{C}\{T^{-1}Z'KZ\}\} \), and premultiplying (16) by \( T^{1/2}(Z'Z)^{-1} \) and taking limits we get

\[ T^{1/2}(\hat{\alpha}_s - \alpha_0) \xrightarrow{d} N\{0, \mathbb{C}\{T(Z'Z)^{-1}Z'KZ(Z'Z)^{-1}\} + \mathbb{C}\{(Z'Z)^{-1}Z'W\}V(\tilde{\beta})\mathbb{C}(W'Z(Z'Z)^{-1})\} \]

where the expectations are evaluated at \( \theta = \theta_0 \).

Unless \( \kappa_t \) is parametrized in terms of \( \theta \), we now have a problem estimating \( V(\hat{\alpha}_s) \). White's [1980b] covariance matrix does not provide a consistent estimator unless \( T^{-1}Z'W \xrightarrow{a.s.} 0 \), as in the simple heteroskedasticity and ARCH models. Nevertheless, \( \hat{\alpha}_s \) is root-T consistent so that \((\hat{\beta}_s', \hat{\alpha}_s')'\) constitutes a root-T consistent estimator for \( \theta \), and its covariance matrix need not be computed if it is only required as an initial estimator (e.g. to obtain \( \hat{\beta}_m \)).

Now if there exists a GARCH component in the conditional variance, we may write \( \alpha_1(L) h_t = \alpha_0 + \alpha_2(L) u^2_t \), as in (2.15b). The dynamic component in the unobservable \( h_t \) makes the above estimators \( \hat{\alpha}_s \) and \( \hat{\alpha}_g \) unfeasible. But consider the ARMA-type form \( \alpha_{12}(L) u^2_t = \alpha_0 + \alpha_1(L) \varepsilon_t \) of (2.16). When fourth order unconditional moments of \( y_t \) exist, the unconditional variance of \( \varepsilon_t \) is
constant and we may apply the procedure of Hannan and Rissanen [1982] to get an initial set of consistent estimates. This involves fitting a long autoregression for $\tilde{u}_t^2$, taking the residuals $\tilde{\epsilon}_t$, say, and then getting estimates of $\alpha_1(L)$ and $\alpha_2(L)$ from the regression of $\tilde{u}_t^2$ on $\tilde{u}_{t-1}^2, \ldots, \tilde{u}_{t-q}^2$ and $\tilde{\epsilon}_{t-1}, \ldots, \tilde{\epsilon}_{t-q}$, where $q = \max (q_1, q_2)$. This procedure can also be used for model identification, in conjunction with some information criterion such as AIC (Akaike [1974]) or BIC (Schwarz [1978]).

§ 3.2.2. Estimating the identifiable functions in the variance equation

Similarly to the estimation of $\beta$ in the mean equation, estimating $\phi$ in the variance equation presents the problem that $\kappa_t$ is not known in general and thus $\hat{\phi}_v$ is not a feasible estimator. As in previous sections we need estimates of the $\kappa_t$, say $\tilde{\kappa}_t$, to define the feasible estimator

$$\hat{\phi}_v = \min_{\phi} \sum_{t=1}^{T} \kappa_t^{-1} e_t^2 = \min_{\phi} e' \tilde{K}^{-1} e,$$

which can also be interpreted as a GMM estimator with orthogonality conditions

$$\psi_v(\phi) = T^{-1} \sum_{t=1}^{T} \tilde{\kappa}_t^{-1} s_{et} e_t = T^{-1} S_{\phi}' \tilde{K}^{-1} e.$$

If $\kappa_t$ is specified parametrically as $\kappa_t = \kappa_t(\theta, \pi)$ and $\tilde{\pi}$ is a root-$T$ consistent estimator of the parameter vector $\pi$ we have a result similar to Lemma 3.1 in

Lemma 3.4. Under (Q0) - (Q7), and if $\tilde{\theta}$ and $\tilde{\pi}$ are root-$T$ consistent estimators of $\theta$ and $\pi$, respectively, then

(i) (a) $\psi_v(\phi) - \psi_v^*(\phi) \rightarrow 0$ uniformly in $\Phi$, and

(b) $T^{1/2} [ \psi_v(\phi_0) - \psi_v^*(\phi_0) ] \rightarrow 0$,

and
Proof: Let $\psi_v^*(\phi) = T^{-1} S^{-1} K^{-1} \varepsilon$. If we substitute $\psi_v(\phi)$ by $\psi_v^*(\phi)$ in (i) and (ii), the proof of this proposition is exactly as that of Lemma 3.1, and Lemma 3.3 ensures that the terms in $\tilde{u}_t^2 - u_t^2$ in $\psi_v(\phi)$ do not have asymptotic effect. □

It follows that under these conditions, using (11),

$$T^{1/2} (\hat{\phi}_v - \phi_0) = T^{1/2} (S^{-1} K^{-1} S^{-1} K^{-1} \varepsilon + o_p(1)) \xrightarrow{d} N [ 0, \Sigma (T^{-1} S^{-1} K^{-1} S^{-1})^{-1} ].$$

The problem here is the parameterization and estimation of the conditional kurtosis function. If conditional normality is assumed the parameterization is automatic as it yields $\kappa_t = 2 h_t^2$ and there are no extra parameters apart from $\theta$. Moreover, initial estimates from the mean equation (for example $\hat{\beta}_2$) and a two-stage estimate of $\alpha$ (for example $\hat{\alpha}_a$) provide the required $\tilde{\theta}$. Because $u_t^2 \mid \mathcal{F}_t \sim \Gamma(1/2, 2 h_t)$ the GLS estimator $\hat{\phi}_v$ is the MLE using variance information only and therefore it uses all the information about $\theta$ that has not already been used by the conditional mean. If a conditional $t$-distribution with $\pi$ degrees of freedom is assumed (Bollerslev [1985], Engle and Bollerslev [1986]), then $\kappa_t = [3 (\pi - 2)/(\pi - 4) - 1] h_t^2$, or $K = [3 (\pi - 2)/(\pi - 4) - 1] \Omega^2$. Because the degrees-of freedom parameter enters $K$ only as a scalar factor, we can obtain $\tilde{\theta}$ and $\hat{\phi}_v$ exactly as in the normal case. We need $\tilde{\pi}$, though, to estimate the covariance matrix $V(\hat{\phi}_v)$. This problem is easily solved because the residual variance of the generalized regression that produces $\hat{\phi}_v$ is a consistent estimator of $[3 (\pi - 2)/(\pi - 4) - 1]$ if $\pi > 4$, and we have the following

Lemma 3.5.- Assume (C0) - (C7), and that the distribution of $y_t$ conditional on $\mathcal{F}_t$ is Student's $t$ with conditional mean $\mu_t(\beta)$, conditional variance $h_t(\theta)$, and $\pi$ degrees of freedom. Consider
\[ \pi = \frac{2 - 4 s^2}{2 - s^2}, \]

where \( s^2 \) is the residual variance estimate from the nonlinear regression of \( u_t^2 \) on \( h_t(\phi) \) in the metric of \( h_t(\tilde{\theta}) \). Then if \( \pi > 4 \), \( \tilde{\pi} \) is a strongly consistent estimator of \( \pi \) given that \( \tilde{\theta} \) is a root-\( T \) consistent estimator of \( \theta_0 \).

**Proof:** If \( y_t \mid \mathcal{F}_t \) is t-distributed, then \( \kappa_t = [3(\pi - 2)/(\pi - 4) - 1] h_t^2 = E(\varepsilon_t^2 \mid \mathcal{F}_t) \).

From standard LS theory it follows that the residual variance \( s^2 \) of the generalized regression in (9) is a strongly consistent estimator of the constant \( \kappa_t / h_t^2 \) if \( \pi > 4 \). Lemmas 3.3 and 3.4 ensure that substituting \( \tilde{u}_t^2 \) and \( \kappa_t(\tilde{\theta}) \) does not affect the asymptotic properties of estimators. Finally, solve \( s^2 = [3(\tilde{\pi} - 2)/(\tilde{\pi} - 4) - 1] \) for \( \tilde{\pi} \).

For the ARCH model with these symmetric distributions the whole vector \( \theta \) is identifiable in the variance equation, so \( \hat{\phi}_v = \hat{\theta}_v \), and further \( \hat{\alpha}_v \) and \( \hat{\alpha}_v \) are asymptotically independent because \( T^{-1/2} W' K^{-1} Z \cdot d \rightarrow N(0, 5) \) (Engle [1982a]). It follows immediately that \( \hat{\alpha}_v \) has the same asymptotic distribution as \( \hat{\alpha}_g \) of the previous subsection, that is,

\[ T^{1/2}(\hat{\alpha}_v - \alpha_0) = T^{1/2} (Z' K^{-1} Z)^{-1} Z' K^{-1} \varepsilon + o_p(1) \]

\[ d \rightarrow N [ 0, \Sigma \{ T^{-1/2} Z' K^{-1} Z \}^{-1} ], \]

and for the variance estimate of \( \beta \) we have

\[ T^{1/2}(\hat{\beta}_v - \beta_0) = T^{1/2} (W' K^{-1} W)^{-1} W' K^{-1} \varepsilon + o_p(1) \]

\[ d \rightarrow N [ 0, \Sigma \{ T^{-1} W' K^{-1} W \}^{-1} ]. \]

In the simple heteroskedasticity model \( \phi = \alpha \), so the asymptotic distribution of \( \hat{\alpha}_v \) is also given by (17) and is the same as that of \( \hat{\alpha}_g \).

For distributions other than Student's t or normal it may be harder to parameterize \( \kappa_t \). If no assumption on the conditional distribution is available, we are left with the semi-parametric possibilities described in § 3.1.2 to estimate the variance equation without parameterizing the conditional fourth
moment, and the simplest semi-parametric alternative is to use LS. Jobson and Fuller [1980] have provided a proof of the properties of such an estimator with iid observations when $\theta$ is not identifiable. They implicitly assume that the identifiable functions are $\alpha$ and a subvector of $\beta$. This is not sufficiently general for our purposes and, paradoxically, the model they use to illustrate their argument is the Amemiya model which, because $\alpha$ is not identified, does not possess identifiable functions of this type. The extension, however, is a minor one and using the same reasoning as with LS in the mean equation we get for the LS estimator $\hat{\phi}_2$ that

$$T^{1/2}(\hat{\phi}_2 - \phi_0) = T^{1/2} (S_\phi' S_\phi)^{-1} S_\phi' \varepsilon + o_p(1) \xrightarrow{d} N [0, \mathcal{E}(T(S_\phi' S_\phi)^{-1} S_\phi' K S_\phi (S_\phi' S_\phi)^{-1})],$$

and $V(\hat{\phi}_2)$ is consistently estimated using White's [1980b] procedure by

$$\hat{V}(\hat{\phi}_2) = T (\hat{S}_\phi' \hat{S}_\phi) \hat{S}_\phi' \hat{\varepsilon}^2 \hat{S}_\phi (\hat{S}_\phi' \hat{S}_\phi)^{-1},$$

where $\hat{S}_\phi = S_\phi(\hat{\theta})$ for any root-$T$ consistent $\hat{\theta}$, and $\hat{\varepsilon} = \text{diag} (\hat{\varepsilon}_1,...,\hat{\varepsilon}_T)$ where $\hat{\varepsilon}_t$ are the LS residuals. Thus when all the identifiable functions are estimated, the problem of estimating the covariance matrix of variance estimators found in the two-stage estimators of the previous subsection is overcome. We can also use the PGLS approach of Amemiya [1983] and Cragg [1983] provided we solve the problem of selecting the instruments, but the more interesting alternative is to extend the adaptive approach of Carroll [1982] and Robinson [1987].

§ 3.3 Combining information from the two conditional moments

The estimators for $\beta$ and $\alpha$ analyzed in the preceding sections are not the most efficient we can get, in general. This is because they exploit only one source of information when there are more sources available. An exception is the simple heteroskedasticity model under conditional normality when both $\hat{\beta}_m$ and $\hat{\alpha}_m$ are asymptotically equivalent to the MLE. There are other exceptions
for subvectors of $\theta$ such as the ARCH model when $\hat{\alpha}_v$ is asymptotically MLE but $\hat{\beta}_m$ is not, and Amemiya's [1973] model under the conditional gamma distribution when $\hat{\beta}_m$ is asymptotically MLE but $\hat{\alpha}_v$ is not (though $\hat{\alpha}_g$ is the MLE for $\alpha$).

In this section we consider the estimation of $\theta$ using jointly the two equations. This improves efficiency but, as usual, the bid to improve efficiency introduces the risk of inconsistency because the additional structure imposed on the problem may be in error. The full information (joint moment) estimators commonly used for heteroskedastic models are interpreted as combinations of limited information (single moment) estimators, and understanding this structure is an important tool for specification in this class of models because separating the sources of information provides the means to assess the importance of efficiency gains and the coherency of the information in different sources.

§ 3.3.1 Combining information from orthogonality conditions

When the conditional variance $h_t$ depends on mean parameters the identifiable parameters $\phi$ of the variance equation in general depend on $\beta$. Therefore the orthogonality conditions $\psi_m(\beta)$ and $\psi_v(\phi)$ share information about common parameters and the natural thing is to consider them jointly. Thus let $\psi(\theta) = [ \psi_m(\beta)' , \psi_v(\phi)' ]'$ define the orthogonality conditions for the joint estimator $\hat{\theta}_J$, say. Because the number of orthogonality conditions in $\psi(\theta)$ is in general greater than the number of parameters $p$ in $\theta$ we need to specify the weighting matrix, and we consider

$$A_T(\theta) = \text{diag} \left( \begin{array}{cc} T^{-1} X' \Omega^{-1} X \end{array} \right)^{-1} , \begin{array}{cc} (T^{-1} S_\phi K^{-1} S_\phi)^{-1} \end{array} \right) ,$$

which evaluated at some root-T consistent $\tilde{\theta}$ produces the optimal GMM estimator for the given orthogonality conditions because $A_T(\tilde{\theta}) \xrightarrow{as} A(\theta_0)$, and
\[ \mathcal{E} \{ T \, \psi(\theta_0) \, \psi(\theta_0)' \} = A(\theta_0) \] (see Hansen [1982], and Chamberlain [1987]). The properties of \( \hat{\theta}_J \) are given in

**Theorem 3.6.** Under (C10) - (C7) and \( y_t \) conditionally symmetrically distributed, the GMM estimator \( \hat{\theta}_J \) with orthogonality conditions \( \psi(\theta) \) and weighting matrix \( A_T(\tilde{\theta}) \) is strongly consistent for \( \theta_0 \) and has asymptotic distribution

\[ T^{1/2} ( \hat{\theta}_J - \theta_0 ) \xrightarrow{d} N \{ 0, \mathcal{E} \{ T^{-1} G' \Sigma^{-1} G \}^{-1} \}, \]

where the expectation is evaluated at \( \theta = \theta_0 \).

**Proof:** Lemmas 3.1 and 3.4 ensure that using \( \psi(\theta) \) and \( A_T(\tilde{\theta}) \) produces estimators with the same asymptotic distribution than those obtained using orthogonality conditions \( \psi^*(\theta) = [ \psi_m^*(\beta)', \psi_v^*(\phi)' ]' \) and weighting matrix \( A_T(\theta_0) \), and hence it suffices to prove the theorem for the latter conditions. We have

\[ \frac{\partial [h_t^{-1} x_t u_t]}{\partial \theta'} = u_t \frac{\partial (h_t^{-1} x_t)}{\partial \theta'} - h_t^{-1} x_t x_t', \]

and

\[ \frac{\partial [\kappa_t^{-1} s_{xt} e_t]}{\partial \theta'} = e_t \frac{\partial (\kappa_t^{-1} s_{xt})}{\partial \theta'} - \kappa_t^{-1} s_{xt} s_t', \]

and using iterated expectations we obtain

\[ E \left[ \frac{\partial \psi(\theta_0)'}{\partial \theta} \right] = - T^{-1} ( \bar{X}' \Omega^{-1} X, S' K^{-1} S \phi). \]

Strong consistency and asymptotic normality follow from Theorem 2.1 and Propositions 2.2 and 2.3. The asymptotic covariance matrix is obtained from Theorem 2.1 by simple algebra using (18) and (20), and is given by

\[ V(\hat{\theta}_J) = \mathcal{E} \{ T^{-1} \bar{X}' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \bar{X} + T^{-1} S' K^{-1} S \phi' (S \phi' K^{-1} S \phi)'^{-1} S \phi' K^{-1} S \}^{-1}. \]

Now since \( \bar{X} = (X, 0) \) it is easily seen that the first term in the expectation is simply \( T^{-1} \bar{X}' \Omega^{-1} \bar{X} \). To complete the proof we show that the second term is \( T^{-1} S' K^{-1} S \), so that \( V(\hat{\theta}_J) = \mathcal{E} \{ T^{-1} G' \Sigma^{-1} G \} \).
Let \( \phi_c = \phi_c(\theta) \) be such that \( \phi^* = (\phi', \phi_c')' = \phi^*(\theta) \) is a one-to-one transformation. Let \( \Delta = \partial \theta / \partial \phi^* = (\Delta_1, \Delta_2) \). By the chain rule we have \( S \Delta = (S \Delta_1, S \Delta_2) = (S\phi, S \Delta_2) \), and since \( \Delta \) is nonsingular and \( \text{rank}(S) = \text{rank}(S\phi) \) we can assume without loss of generality that \( S \Delta_2 = 0 \). Using this, premultiply the second term in the expectation by \( \Delta^{-1} \Delta' \) and postmultiply by \( \Delta \Delta^{-1} \) to obtain

\[
\Delta^{-1} \begin{pmatrix}
S\phi' K^{-1} S\phi & 0 \\
0 & 0
\end{pmatrix} \Delta^{-1} = \Delta^{(1)'},
\]

where \( \Delta^{-1} = (\Delta^{(1)'}, \Delta^{(2)'})' \). Now from \( \Delta \Delta^{-1} = I_p \) we have \( \Delta_1 \Delta^{(1)} = I_p - \Delta_2 \Delta^{(2)} \), and thus \( S\phi \Delta^{(1)} = S \Delta_1 \Delta^{(1)} = S \), using again \( S \Delta_2 = 0 \) and \( S \Delta_1 = S\phi \). This completes the proof upon substitution.

Several points are of interest in this Theorem. First, \( \hat{\theta}_J \) is constructed as the optimal GMM estimator for the orthogonality conditions \( \psi(\theta) \). When the two subsets of orthogonality conditions have some optimal properties we should expect optimality for \( \hat{\theta}_J \) with respect to the information in the first two conditional moments. Indeed under normality \( \hat{\theta}_J \) has the same asymptotic distribution as the MLE, as we show in § 3.3.2. Secondly, \( \hat{\theta}_J \) can be obtained as the joint GLS estimator that minimizes \( \nu' \Sigma^{-1} \nu \) and thus is also asymptotically equivalent to Jobson and Fuller's JGLS estimator. Under normality this is the two-step estimator from the method of scoring. Thirdly, here we have used orthogonality conditions that produce GLS estimates, but the same principle of combination may be applied to other estimators obtained from the separate moments. Finally, \( \hat{\theta}_J \) combines the information in a very natural way. To see this, note that underlying (19) we have

\[
T^{1/2} (\hat{\theta}_J - \theta_0) = T^{1/2} (G' \Sigma^{-1} G)^{-1} G' \Sigma^{-1} \nu + o_p(1).
\]

For simplicity, consider the case when \( \theta \) is identifiable in the variance equation. Using simple results of partitioned inversion we get

**Corollary 3.7.-** Under the assumptions of Theorem 3.6,
\[ \hat{\beta}_J - \beta_0 = (I_k - \Pi)(\hat{\beta}_m - \beta_0) + \Pi(\hat{\beta}_v - \beta_0) + o_p(T^{-1/2}), \]  

and

\[ V(\hat{\beta}_J)^{-1} = V(\hat{\beta}_m)^{-1} + V(\hat{\beta}_v)^{-1}, \]

where

\[ \Pi = V(\hat{\beta}_J)V(\hat{\beta}_v)^{-1} = I_k - V(\hat{\beta}_J)V(\hat{\beta}_m)^{-1}, \]

and the covariance matrices are

\[ V(\hat{\beta}_J) = \xi (T^{-1}X' \Omega^{-1}X + T^{-1}W' K^{-1/2}Q_z K^{-1/2}W)^{-1}, \]

\[ V(\hat{\beta}_m) \text{ is given in (7), and } V(\hat{\beta}_v) \text{ is given in (12b)}. \]

**Proof:** Partition \( G = (G_\beta, G_\alpha) \) conformably to \( \theta = (\beta', \alpha')' \). The usual partitioned regression results applied to (21) (e.g. Theil [1971]) yield

\[ T^{1/2}(\hat{\beta}_J - \beta_0) = T^{1/2}(G_\beta \Sigma^{-1/2}Q_\alpha \Sigma^{-1/2}G_\beta)'(G_\beta \Sigma^{-1/2}Q_\alpha \Sigma^{-1/2}G_\beta)^{-1}(G_\beta \Sigma^{-1/2}Q_\alpha \Sigma^{-1/2}u + W'K^{-1/2}Q_z K^{-1/2}\epsilon) + o_p(1), \]

where \( Q_\alpha \) is the projection matrix onto the space orthogonal to \( \Sigma^{-1/2}G_\alpha \). But

\[ G_\beta = (X', W')' \text{ and } G_\alpha = (0, Z')', \]

so the expression reduces to

\[ T^{1/2}(\hat{\beta}_J - \beta_0) = T^{1/2}(X' \Omega^{-1}X + W'K^{-1/2}Q_z K^{-1/2}W)^{-1}(X' \Omega^{-1}u + W'K^{-1/2}Q_z K^{-1/2}\epsilon) + o_p(1), \]

where \( Q_z \) is the projection matrix onto the space orthogonal to \( K^{-1/2}Z \), and the covariance matrix is

\[ V(\hat{\beta}_J) = \xi (T^{-1}X' \Omega^{-1}X + T^{-1}W' K^{-1/2}Q_z K^{-1/2}W)^{-1}. \]

The partition of \( \hat{\theta}_v \) is given in (12). Collecting these results and substituting back in the expression for \( \hat{\beta}_J \) using (7) produces after simple algebra

\[ T^{1/2}(\hat{\beta}_J - \beta_0) = V(\hat{\beta}_J) [ V(\hat{\beta}_m)^{-1} T^{1/2}(\hat{\beta}_m - \beta_0) + V(\hat{\beta}_v)^{-1} T^{1/2}(\hat{\beta}_v - \beta_0) ] + o_p(1), \]

and the decomposition of the inverse covariance matrix in (23).

Therefore \( \hat{\beta}_J \) is asymptotically a matrix weighted average (AMWA) of \( \hat{\beta}_m \) and \( \hat{\beta}_v \). MWA's have been studied extensively in the Bayesian literature (e.g. Zellner [1971], Chamberlain and Leamer [1976] and Leamer [1978] *inter alia*) and they are known to be the optimal way of combining estimators from independent sources of information in the sense of achieving the smallest
variance. Thus (22) - (24) are not surprising results and, given symmetry, we can combine any two estimators - one from each moment - optimally by means of an AMWA. The AMWA structure of $\hat{\beta}_j$ is rich in intuition and it clearly shows how the information extracted from each of the conditional moments is combined to improve efficiency in estimation. One of the central issues of heteroskedastic models is that efficiency can be improved over GLS when the conditional variance depends on $\beta$, and (22) - (24) separate clearly the efficiency gain. Moreover, we can break the complicated problem of joint specification and estimation of the two moments into the simpler problems of specification and estimation of each moment separately.

Although our central interest is $\beta$ because this is the subvector of the parameters with more than one source of information, it is also important to look at the properties of the estimator for $\alpha$ whose only source of information is the variance. We do this in

**Corollary 3.8** - Under the assumptions of Theorem 3.6

$$T^{1/2}(\hat{\alpha}_j - \alpha_0) \overset{d}{\to} N\left(0, V_\alpha + \text{Cov}(\hat{\alpha}_j, \hat{\beta}_j) V(\hat{\beta}_j)^{-1} \text{Cov}(\hat{\beta}_j, \hat{\alpha}_j)\right), \quad (25)$$

where $V_\alpha = E\{ T^{-1} Z' K^{-1} Z \}^{-1}$ and $\text{Cov}(\hat{\alpha}_j, \hat{\beta}_j) = -V_\alpha E\{ T^{-1} Z' K^{-1} W \} V(\hat{\beta}_j)$, and

$$T^{1/2}(\hat{\alpha}_j - \hat{\alpha}_v) = \text{Cov}(\hat{\alpha}_j, \hat{\beta}_j) V(\hat{\beta}_j)^{-1} T^{1/2}(\hat{\beta}_j - \hat{\beta}_v) + o_p(1). \quad (26)$$

**Proof:** Using the partition of $G = (G_\beta, G_\alpha)$ in (21) yields for $\hat{\alpha}_j$

$$T^{1/2}(\hat{\alpha}_j - \alpha_0) = T^{1/2}(Z' K^{-1} Z)^{-1} Z' K^{-1} \left[ \epsilon - W (\hat{\beta}_j - \beta_0) \right] + o_p(1),$$

and the covariance matrix in (25) is obtained by simple partitioned inversion of $V(\hat{\theta}_j)$. Similarly, the partition of $\hat{\theta}_v$ in (12) for $\alpha$ may be rewritten as

$$T^{1/2}(\hat{\alpha}_v - \alpha_0) = T^{1/2}(Z' K^{-1} Z)^{-1} Z' K^{-1} \left[ \epsilon - W (\hat{\beta}_v - \beta_0) \right] + o_p(1),$$

and the result for $T^{1/2}(\hat{\alpha}_j - \hat{\alpha}_v)$ follows from the difference of the two expressions, using the form of $\text{Cov}(\hat{\alpha}_j, \hat{\beta}_j)$. \qed
The difference between \( \hat{\alpha}_J \) and \( \hat{\alpha}_\nu \) is that the former takes advantage of the more efficient joint estimator of \( \beta \) and is therefore more efficient. When \( \mathcal{E} \{ T^{-1} Z' K^{-1} W \} = 0 \) as in ARCH class or simple heteroskedastic models, we get asymptotic independence between \( \hat{\beta}_J \) and \( \hat{\alpha}_J \). It follows from (26) that in these cases the variance equation alone provides \( \hat{\alpha}_J \) with covariance matrix \( V_\alpha, \hat{\beta}_\nu \) and \( \hat{\alpha}_\nu \) are independent, and therefore \( \hat{\alpha}_J \) is asymptotically equivalent to \( \hat{\alpha}_g \).

§ 3.3.2 Likelihood factorization: the case of normality

In this subsection we assume the conditional distribution to be normal and take the more traditional ML approach. We show that the MLE is an AMWA of the estimators from the conditional mean and conditional variance. The log-likelihood, score and information matrix are (see (2.10) - (2.12))

\[
\mathcal{L}(\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log h_t - \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-1} u_t^2 ,
\]

\[
d_\theta(\theta) = T^{-1} X' \Omega^{-1} u + \frac{1}{2} T^{-1} S' \Omega^{-2} \varepsilon = T^{-1} G' \Sigma^{-1} u ,
\]

\[
d_\theta(\theta) = T^{-1} X' \Omega^{-1} u + \frac{1}{2} T^{-1} W' \Omega^{-2} \varepsilon ,
\]

\[
d_\alpha(\theta) = \frac{1}{2} T^{-1} Z' \Omega^{-2} \varepsilon ,
\]

and

\[
\mathcal{J}(\theta) = E \left[ T^{-1} G' \Sigma^{-1} G \right] = E \left( \begin{array}{c}
T^{-1} X' \Omega^{-1} X + \frac{1}{2} T^{-1} W' \Omega^{-2} W \\
\frac{1}{2} T^{-1} W' \Omega^{-2} Z \\
\frac{1}{2} T^{-1} Z' \Omega^{-2} Z
\end{array} \right) .
\]

We start with

**Theorem 3.9.** Under \((Q0') - (Q8)\) the MLE \( \hat{\theta} \) of \( \theta \) is strongly consistent and has asymptotic distribution given by

\[
T^{1/2} ( \hat{\theta} - \theta_0 ) \overset{d}{\to} N \left[ 0 , \mathcal{J}(\theta_0)^{-1} \right] .
\]
Proof: The MLE is the GMM estimator with orthogonality conditions
d_{\theta}(\hat{\theta}) = T^{-1} \hat{G}' \hat{\Sigma}^{-1} \hat{v} = 0. The result follows from Theorem 2.1 and Propositions 2.2 and 2.3.

The weak exogeneity assumption is not required for the proof but is included to ensure the full efficiency of \( \hat{\theta} \). The score and information matrix show that the MLE is formed by combining additively the two sets of orthogonality conditions for the mean and variance equations. In fact, simple inspection of the information matrix establishes

**Corollary 3.10.** Under the assumptions of Theorem 3.9 the MLE \( \hat{\theta} \) and the GMM estimator \( \hat{\theta}_j \) have the same asymptotic distribution.

Proof: Under normality \( K = 2 \Omega^2 \), and the distributions in (19) and (28) are identical.

This Theorem and Corollary establish that the MLE is an AMWA of the estimators obtained separately from the two conditional moments. In likelihood terms this suggests that we can factorize the likelihood function, at least locally.

To see this let \( \hat{h}_t^{-1} \) and \( \hat{u}_t^2 \) be functions of the data alone. Using the Mean Value Theorem we can write

\[
\hat{h}_t^{-1} \hat{u}_t^2 = \hat{h}_t^{-1} \hat{u}_t^2 + \hat{h}_t^{-1} \hat{u}_t^2 + (\hat{h}_t^{-1} \hat{u}_t^2 - \hat{h}_t^{-1} \hat{u}_t^2 - \hat{h}_t^{-1} \hat{u}_t^2),
\]

for \( \hat{u}_t^2 \in [u_t^2, \tilde{u}_t^2] \) and \( \hat{h}_t^{-1} \in [h_t^{-1}, \tilde{h}_t^{-1}] \). Then, neglecting the term in brackets,

\[
\ell(\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log h_t - \frac{1}{2} T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-1} \hat{u}_t^2 - \frac{1}{2} T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-1} \hat{u}_t^2,
\]

and if we choose \( \hat{h}_t^{-1} \) and \( \hat{u}_t^2 \) to approximate \( h_t^{-1} \) and \( u_t^2 \) in the sense given in Theorem 3.11 below, then maximizing the log likelihood

\[
\ell^*(\theta) = \ell_m(\theta) + \ell_v(\theta),
\]

where
\[ \mathcal{L}_m(\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \tilde{h}_t^{-1} u_t^2, \]
and
\[ \mathcal{L}_v(\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} \log h_t - \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-1} \tilde{u}_t^2, \]
yields estimators with identical asymptotic distribution to those obtained by maximizing \( \mathcal{L}(\theta) \). This equivalence is proven in

**Theorem 3.11.** The log-likelihoods \( \mathcal{L}^*(\theta) \) and \( \mathcal{L}(\theta) \) produce estimators of \( \theta \) with identical asymptotic distribution to the order of \( T^{1/2} \) if

(i) \( T^{-1/2} \sum_{t=1}^{T} x_t u_t (\tilde{h}_t^{-1} - h_t^{-1}) \xrightarrow{as} 0 \), and \( T^{-1/2} \sum_{t=1}^{T} s_t h_t^{-2} (\tilde{u}_t^2 - u_t^2) \xrightarrow{as} 0 \),

and

(ii) \( T^{-1} \sum_{t=1}^{T} u_t \frac{\partial x_t}{\partial \theta'} (\tilde{h}_t^{-1} - h_t^{-1}) \xrightarrow{as} 0 \), and \( T^{-1} \sum_{t=1}^{T} \frac{\partial h_t^{-2} s_t}{\partial \theta'} (\tilde{u}_t^2 - u_t^2) \xrightarrow{as} 0 \).

**Proof.** Since \( \tilde{u}_t^2 \) and \( \tilde{h}_t^{-1} \) are functions of the data alone, the score for \( \mathcal{L}^*(\theta) \) is

\[ d^*(\theta) = T^{-1} X' \tilde{\Omega}^{-1} u + \frac{1}{2} T^{-1} S' \tilde{\Omega}^{-2} \varepsilon = d_0 + T^{-1} X' (\tilde{\Omega}^{-1} - \Omega^{-1}) u + \frac{1}{2} T^{-1} S' \tilde{\Omega}^{-2} (\tilde{u}^2 - u^2), \]
because \( d_0(\theta) = T^{-1} X' \tilde{\Omega}^{-1} u + \frac{1}{2} T^{-1} S' \tilde{\Omega}^{-2} \varepsilon \), and \( \varepsilon = \varepsilon + (\tilde{u}^2 - u^2) \), and therefore

\[ T^{1/2} [d^*(\theta) - d(\theta)] \xrightarrow{as} 0 \] follows from (i). Taking derivatives again we have

\[ \frac{\partial^2 \mathcal{L}^*(\theta)}{\partial \theta \partial \theta'} = - T^{-1} G' \Sigma^{-1} G + T^{-1} \sum_{t=1}^{T} \tilde{h}_t^{-1} u_t \frac{\partial x_t}{\partial \theta'} + T^{-1} \sum_{t=1}^{T} \frac{\partial h_t^{-2} s_t}{\partial \theta'} e_t, \]
and since at \( \theta_0 \)

\[ E \left[ T^{-1} \sum_{t=1}^{T} h_t^{-1} u_t \frac{\partial x_t}{\partial \theta'} + T^{-1} \sum_{t=1}^{T} \frac{\partial h_t^{-2} s_t}{\partial \theta'} e_t \right] = 0, \]
it follows from (ii) that \( \mathcal{J}^*(\theta_0) \xrightarrow{as} \mathcal{J}(\theta_0). \)

Given root-\( T \) consistent estimators \( \tilde{\theta} \) and \( \tilde{\beta} \), the obvious choices for \( \tilde{h}_t^{-1} \) and \( \tilde{u}_t^2 \) are given by \( \tilde{h}_t = h_t(\tilde{\theta}) \) and \( \tilde{u}_t = y_t - \tilde{\mu}_t = y_t - \mu_t(\tilde{\beta}) \), respectively, and then

Lemmas 3.1, 3.3 and 3.4 ensure that the conditions for Theorem 3.11 are met.
Ruud [1984] has provided a Likelihood Factorization Theorem which we can apply to the likelihood function \( \mathcal{L}(\theta) \). This Theorem is generalized to unidentifiable factors in

**Theorem 3.12.** Suppose a regular likelihood \( \mathcal{L}_0(\theta) \) for the parameter vector \( \theta \in \Theta \subset \mathbb{R}^p \) factorizes into the product of two (pseudo-)likelihoods \( \mathcal{L}_0(\theta) = \mathcal{L}_1(\theta) \mathcal{L}_2(\theta) \) which are regular for their identifiable parameters \( \phi_1 = \phi_1(\theta) \) and \( \phi_2 = \phi_2(\theta) \), respectively, where \( \phi_j \in \Phi_j \subset \mathbb{R}^{p_j} \) for \( 0 \leq p_j \leq p \). Suppose that there exists a jointly identifiable function \( \gamma = \gamma(\theta) \) so that we may partition \( \phi_1 = (\gamma, \phi_1') \) and \( \phi_2 = (\gamma, \phi_2') \) in such a way that \( \mathcal{L}_1 \) does not depend on \( \phi_2 \) and \( \mathcal{L}_2 \) does not depend on \( \phi_1 \). Then

(i) The (pseudo-)MLE's \( \hat{\phi}_1 \) and \( \hat{\phi}_2 \) of \( \phi \) obtained from each of the factors separately are asymptotically independent,

(ii) The MLE \( \hat{\gamma} \) of \( \gamma \) is an AMWA of \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \), the (pseudo-)MLE's from the separate factors, that is,

\[
V(\hat{\gamma})^{-1} (\hat{\gamma} - \gamma_0) = V(\hat{\gamma}_1)^{-1} (\hat{\gamma}_1 - \gamma_0) + V(\hat{\gamma}_2)^{-1} (\hat{\gamma}_2 - \gamma_0) + o_p(T^{-1/2}),
\]

where \( \gamma_0 \) is the true value of \( \gamma \), and

\[
V(\hat{\gamma})^{-1} = V(\hat{\gamma}_1)^{-1} + V(\hat{\gamma}_2)^{-1},
\]

(iii) The MLE's \( \tilde{\phi}_j \) of \( \phi_j \) are efficient relative to the (pseudo-)MLE's \( \hat{\phi}_j \) from the factors, for \( j = 1, 2 \).

**Proof.** Let \( \phi = (\gamma, \phi_1', \phi_2') \). The identifiability of \( \theta \) in \( \mathcal{L}_0 \) implies a one-to-one correspondence between \( \theta \) and \( \phi \), and by the invariance of the likelihood principle (see e.g. Cox and Hinkley [1974]) it is equivalent to conduct inference on \( \theta \) or \( \phi \). Taking logs of the likelihoods we get \( \mathcal{L}_0(\phi) = \mathcal{L}_1(\phi_1) + \mathcal{L}_2(\phi_2) \), where \( \mathcal{L}_j = \log \mathcal{L}_j, j = 0, 1, 2 \). Differentiating produces in obvious notation

\[
d_0(\phi) = \frac{\partial \mathcal{L}_0}{\partial \phi} = \begin{pmatrix} \frac{\partial \mathcal{L}_0}{\partial \phi} \\ \frac{\partial \mathcal{L}_0}{\partial \phi} \\ \frac{\partial \mathcal{L}_0}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathcal{L}_1}{\partial \phi} + \frac{\partial \mathcal{L}_2}{\partial \phi} \\ \frac{\partial \mathcal{L}_1}{\partial \phi} + \frac{\partial \mathcal{L}_2}{\partial \phi} \\ \frac{\partial \mathcal{L}_1}{\partial \phi} + \frac{\partial \mathcal{L}_2}{\partial \phi} \end{pmatrix} = d_1(\phi) + d_2(\phi), \quad (29)
\]
and differentiating again and taking expected values at $\phi_0$, the true value of $\phi$, we get

$$J_0(\phi_0) = -E \left[ \frac{\partial^2 \ell_0}{\partial \phi \partial \phi'} \right] = \begin{pmatrix} \int_{\gamma \gamma} & \int_{\gamma 1} & \int_{\gamma 2} \\ \int_{1 \gamma} & \int_{1 1} & 0 \\ \int_{2 \gamma} & 0 & \int_{2 2} \end{pmatrix} = \begin{pmatrix} \int_{1 \gamma} & \int_{1 1} & 0 \\ \int_{1 \gamma} & \int_{1 1} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \int_{2 \gamma} & 0 & \int_{2 2} \\ 0 & 0 & 0 \end{pmatrix}$$

$$= -E \left[ \frac{\partial^2 \ell_1}{\partial \phi \partial \phi'} \right] - E \left[ \frac{\partial^2 \ell_2}{\partial \phi \partial \phi'} \right] = J_1(\phi_0) + J_2(\phi_0). \quad -(30)$$

Using the Mean Value Theorem for random functions (Jennrich [1969]) on the first order conditions $d_j(\hat{\phi}_j) = 0$ yields for $j = 0, 1, 2$

$$d_j(\phi_0) = d_j(\phi_0)(\hat{\phi}_j - \phi_0) + o_p(T^{-1/2}). \quad -(31)$$

The independence assertion in (i) follows by taking the variance of $d_0(\phi_0)$ in (29) and comparing to (30) exactly as in Ruud [1984]. Substituting (29) and (30) in (31) results in

$$d_0(\phi_0) = J_{0 \gamma}(\hat{\gamma} - \gamma_0) + J_{1 \gamma}(\hat{\phi}_1 - \phi_1^0) + J_{2 \gamma}(\hat{\phi}_2 - \phi_2^0) + o_p(T^{-1/2})$$

$$= J_{1 \gamma}(\hat{\gamma}_1 - \gamma_0) + J_{2 \gamma}(\hat{\gamma}_2 - \gamma_0) + J_{1 \gamma}(\hat{\phi}_1 - \phi_1^0) + J_{2 \gamma}(\hat{\phi}_2 - \phi_2^0) + o_p(T^{-1/2}), \quad -(32)$$

where $\phi_0 = (\gamma_0', \phi_1^0, \phi_2^0)'$ and

$$d_j(\phi_0) = J_{j \gamma}(\hat{\gamma} - \gamma_0) + J_{j \gamma}(\hat{\phi}_j - \phi_j^0) + o_p(T^{-1/2})$$

$$= J_{j \gamma}(\hat{\gamma}_j - \gamma_0) + J_{j \gamma}(\hat{\phi}_j - \phi_j^0) + o_p(T^{-1/2}), \quad -(33)$$

for $j = 1, 2$. Solving subsequently (33) for $(\hat{\phi}_j - \phi_j^0)$ and $(\hat{\phi}_j - \phi_j^0)$ we get

$$\hat{\phi}_j - \phi_j^0 = \frac{1}{J_{j \gamma}} [ d_j(\phi_0) - J_{j \gamma}(\hat{\gamma}_j - \gamma_0) ] + o_p(T^{-1/2}),$$

and

$$\hat{\phi}_j - \phi_j^0 = \frac{1}{J_{j \gamma}} [ d_j(\phi_0) - J_{j \gamma}(\hat{\gamma}_j - \gamma_0) ] + o_p(T^{-1/2}),$$

which upon substitution in (32) yields
\[
( \mathbf{I}_{2\gamma} - \mathbf{I}_{2\gamma}^1 \mathbf{I}_{1\gamma}^1 \mathbf{I}_{1\gamma}^{-1} \mathbf{I}_{2\gamma}^2 \mathbf{I}_{2\gamma}^{-1} ) ( \hat{\gamma} - \gamma_0 ) + o_p(T^{-1/2}) \\
= ( \mathbf{I}_{1\gamma} - \mathbf{I}_{1\gamma}^1 \mathbf{I}_{1\gamma}^{-1} ) ( \hat{\gamma}_1 - \gamma_0 ) + ( \mathbf{I}_{2\gamma} - \mathbf{I}_{2\gamma}^1 \mathbf{I}_{2\gamma}^{-1} ) ( \hat{\gamma}_2 - \gamma_0 ) + o_p(T^{-1/2}) ,
\]
and partitioned inversion in (30) together with standard ML theory produces
\[
V(\hat{\gamma})^{-1} ( \hat{\gamma} - \gamma_0 ) = V(\hat{\gamma}_1)^{-1} ( \hat{\gamma}_1 - \gamma_0 ) + V(\hat{\gamma}_2)^{-1} ( \hat{\gamma}_2 - \gamma_0 ) + o_p(T^{-1/2}) ,
\]
and
\[
V(\hat{\gamma})^{-1} = V(\hat{\gamma}_1)^{-1} + V(\hat{\gamma}_2)^{-1},
\]
which establishes (ii). Finally, (iii) follows from (33) and the efficiency of \( \hat{\gamma} \) relative to \( \hat{\gamma}_1 \).

The proof follows closely that of Ruud, except that it uses only the consistency of the identifiable functions in the factors. It can also be generalized to cover more than two factors. To conform to this Theorem we must select the identifiable functions in the variance equation so that \( \phi \) may be partitioned as \( \phi = (\gamma', \phi_1')' \), where \( \gamma = \gamma(\beta) \) and \( \phi_1 \) does not depend on \( \beta \). Denote by \( k^* \) the dimension of \( \gamma \) and by \( m^* = p^* - k^* \) that of \( \phi_1 \). Clearly, \( 0 \leq k^* \leq k \) and \( 0 \leq m^* \leq m \). Thus we have the estimator \( \hat{\phi}_v = (\hat{\gamma}', \hat{\phi}_1v')' \) from the variance factor. Let \( \hat{\gamma}_m = \gamma(\hat{\beta}_m) \) and use (7) to obtain
\[
T^{1/2} ( \hat{\gamma}_m - \gamma_0 ) \xrightarrow{d} N \left[ 0 , \Gamma(\beta_0) V(\hat{\beta}_m) \Gamma(\beta_0)' \right],
\]
where \( \Gamma(\beta) = \partial \gamma(\beta) / \partial \beta' \). Using Theorems 3.11 and 3.12 we obtain
\[
\hat{\gamma} - \gamma_0 = ( \mathbf{I}_{k^*} - \Pi_u ) ( \hat{\gamma}_m - \gamma_0 ) + \Pi_u ( \hat{\gamma}_v - \gamma_0 ) + o_p(T^{-1/2}) , \tag{34}
\]
where
\[
\Pi_u = V(\hat{\gamma}) V(\hat{\gamma}_v)^{-1} = \mathbf{I}_{k^*} - V(\hat{\gamma}) V(\hat{\gamma}_m)^{-1},
\]
and
\[
V(\hat{\gamma})^{-1} = V(\hat{\gamma}_m)^{-1} + V(\hat{\gamma}_v)^{-1}, \tag{35}
\]
and the AMWA interpretation applies to the jointly identifiable functions.
As an example consider the linear Amemiya model. The k-1 dimensioned function \( \gamma = (\beta_2/\beta_1, \ldots, \beta_k/\beta_1) \) is identified in the variance equation (the normalizing coefficient may be any nonzero \( \beta_j \)). There is one more identifiable parameter in the variance equation, given by \( \beta^*_1 = \alpha^{1/2} \beta_1 \), and the normalizing coefficient \( \beta_1 \) is an additional parameter in the mean equation reparameterized in terms of \((\gamma', \beta_1')'\). This conforms to the parameter structure in Theorem 3.12 because \( \lambda_m \) does not contain any information about \( \beta_1^* \) and, conversely, \( \lambda_\nu \) does not contain any information about \( \beta_1 \).

When \( \theta \) is identified in \( \lambda_\nu \) we have that \( \gamma = \beta \) as in the ARCH model and the Poisson-type parameterization. In the ARCH model \( \text{Cov}(\alpha_\nu, \hat{\beta}_\nu) = 0 \) and so \( \hat{\beta}_\nu \) can be obtained from the generalized regression of \( \hat{u}_t^2 \) on \( h_t(\beta, \alpha) \), while the covariance matrix of Corollary 3.7 simplifies to \( V(\hat{\beta}_\nu) = 2 \xi \{ T^{-1} W' \Omega^{-2} W \}^{-1} \). In the Poisson model \( \alpha \) is not present and so \( \hat{\beta}_\nu \) can be obtained from the generalized regression of \( \hat{u}_t^2 \) on \( \mu_t(\beta) \), and has covariance matrix \( V(\hat{\beta}_\nu) = 2 \xi \{ T^{-1} X' \Omega^{-2} X \}^{-1} \).

§ 3.4 Analyzing the contributions to efficiency

To complete the picture, we must assess the information content of each of the moments. This allows us to discern whether the information in both moments is coherent, when one or both sources are not informative enough, and when alternative parameterizations can be more fruitful.

To measure the relative contributions to efficiency from each of the two conditional moments we construct a quadratic in the (inverse) variance decomposition (35) and divide by the left-hand-side to get

\[
\frac{\xi' V(\hat{\gamma}_m)^{-1} \xi}{\xi' V(\hat{\gamma})^{-1} \xi} + \frac{\xi' V(\hat{\gamma}_\nu)^{-1} \xi}{\xi' V(\hat{\gamma})^{-1} \xi} = \phi_m + \phi_\nu = 1 ,
\]  
- (36)
with \( \varphi_m \) and \( \varphi_v \) defined in obvious way. These quantities measure the relative contributions to efficiency from the conditional mean and from the conditional variance, respectively. These relative contributions can be estimated consistently by substitution of the corresponding estimators, but because the relation holds only asymptotically the covariance matrices must be evaluated at the same estimator. Simple choices for \( \zeta \) to employ in an efficiency comparison may be the parameters themselves \( \gamma \), and also a vector of ones.

Consider the related measure of relative efficiency given by \( \varphi = \varphi_v / \varphi_m \). As \( \varphi \) falls (as \( \varphi \) grows) the conditional mean (variance) becomes more informative relative to the conditional variance (mean). We can rewrite \( V(\hat{\gamma}) \) using (35) as

\[
V(\hat{\gamma}) = \varphi_m^{-1} \left[ V_m^{-1} + \varphi V_v^{-1} \right]^{-1} = \varphi_v^{-1} \left[ \varphi^{-1} V_m^{-1} + V_v^{-1} \right]^{-1},
\]

where \( V_j = \varphi_j V(\hat{\gamma}_j) \) for \( j = m, v \), and from (34) we obtain

\[
\hat{\gamma} = \left[ V_m^{-1} + \varphi V_v^{-1} \right]^{-1} \left[ V_m^{-1} \hat{\gamma}_m + \varphi V_v^{-1} \hat{\gamma}_v \right] + o_p(T^{-1/2}) = \\
\left[ \varphi^{-1} V_m^{-1} + V_v^{-1} \right]^{-1} \left[ \varphi^{-1} V_m^{-1} \hat{\gamma}_m + V_v^{-1} \hat{\gamma}_v \right] + o_p(T^{-1/2}),
\]

so that as \( \varphi \to 0 \) we have \( \hat{\gamma} - \hat{\gamma}_m \xrightarrow{as} 0 \) and \( V(\hat{\gamma}) - V(\hat{\gamma}_m) \to 0 \), and thus all the information is coming from the mean equation. Similarly as \( \varphi \to \infty \) (\( \varphi^{-1} \to 0 \)) we have \( \hat{\gamma} - \hat{\gamma}_v \xrightarrow{as} 0 \) and \( V(\hat{\gamma}) - V(\hat{\gamma}_v) \to 0 \), and all the information is coming from the variance equation.

The above discussion suggests that we may extract information from the variance equation in relative independence from the mean equation, and this point deserves a more careful examination. If such is the case it appears that assumption (C.7), which requires that \( \beta \) be identifiable in the mean equation alone, is unnecessarily strong because the information contained in the variance equation could be used to identify \( \beta \). To assess this proposition observe that
\[
\text{rank } \left( T^{-1} G' \Sigma^{-1} G \right) = \text{rank } \begin{pmatrix} T^{-1} X' \Omega^{-1} X + W' K^{-1} W & T^{-1} W' K^{-1} Z \\ T^{-1} Z' K^{-1} W & T^{-1} Z' K^{-1} Z \end{pmatrix}
\]
\[
= \text{rank } \begin{pmatrix} T^{-1} X' \Omega^{-1} X + W' K^{-1/2} Q_z K^{-1/2} W & 0 \\ T^{-1} Z' K^{-1} W & T^{-1} Z' K^{-1} Z \end{pmatrix},
\]

where \( Q_z \) is the projection matrix onto the space orthogonal to \( K^{-1/2} Z \), and therefore the identifiability of \( \beta \) depends on the rank of the matrix

\[
V_T = T^{-1} X' \Omega^{-1} X + W' K^{-1/2} Q_z K^{-1/2} W.
\]

Let us consider the case when \( h_t = h_t(\mu_t, y_{t-j}, u_{t-j} ; 1 \leq j \leq n, \alpha) \), so that the \( \beta \) coefficients affect the variance only through current predictions and past errors. All well known models of the conditional variance fall into this description, and it seems hard to think of cases where the coefficients of the conditional mean would be affecting the conditional variance in a different way. We can then express \( w_t = \partial h_t / \partial \beta = \sum_{j=0}^{n} \lambda_{jt} x_{t-j} \), where \( \lambda_{ot} = \partial h_t / \partial \mu_t \), and \( \lambda_{jt} = \partial h_t / \partial \mu_{t-j} + \partial h_t / \partial u_{t-j} \), for \( 1 \leq j \leq n \), and if we ignore the first \( n \) observations we can write \( W = AX \), where \( A \) is a \( T \times T \) matrix which accommodates properly the \( \lambda_{jt} \). Substituting \( W \) in the expression for \( V_T \) we get

\[
V_T = T^{-1} X' ( \Omega^{-1} + A' K^{-1/2} Q_z K^{-1/2} A ) X,
\]

and it is immediate that \( \text{rank } (V_T) \leq \text{rank } (T^{-1} X' \Omega^{-1} X) \), and therefore no aid in the identifiability of the mean parameters can be obtained from the information in the conditional variance. It is unlikely that the first \( n \) observations can break the collinearity lock in small to moderate samples, but even if this were the case one would expect very imprecise estimates. Therefore we require that \( \beta \) be identifiable in the mean equation alone, assumption (Q.7) cannot be relaxed, and the variance information may play only a quantitative and not a qualitative role in the estimation of \( \beta \). Nevertheless, this quantitative role of the variance information may be very important, especially when working with time series data in which dynamics and the highly
collinear nature of economic time series may result in imprecise estimates for \( \beta \) obtained from the mean equation alone.

§ 3.5 Some comments on Monte Carlo evidence

Tables 3.1 to 3.4 present simulation results on the performance of different estimators of \( \theta \) and measures of the contribution to efficiency of the two moments under correct specification. For the mean parameters we report the estimators \( \hat{\beta}_A \) (OLS), \( \hat{\beta}_m \) (GLS), \( \hat{\beta}_v \) (VAR), the matrix weighted average of \( \hat{\beta}_m \) and \( \hat{\beta}_v \) (MWA), and \( \hat{\beta} \) (ML). For the variance parameters we report the estimators \( \hat{\alpha}_g \) (SLS), \( \hat{\alpha}_g \) (GLS), \( \hat{\alpha}_v \) (VAR), and \( \hat{\alpha} \) (ML). The parameter vector \( \theta \) is identifiable in the variance equation in both the Poisson-N and ARCH models. The row 'Inf. meas.' contains the sample means of the information measures \( \phi_m \) for the conditional mean, and \( \phi_v \) for the conditional variance, given in (36). The sampling variances of these statistics were in all cases very small and are not reported, and the quadratic forms and covariance matrix estimates were evaluated at the MLE, without restricting the measures to add to unity. Below the information measures there are two blocks of information for the estimators of the coefficients. The first block reports the sample mean and standard error for the bias \( (\hat{\theta} - \theta) \) of different estimators, and the second block reports the skewness coefficient \( \nu b_1 \) and the kurtosis coefficient \( b_2 \), which are tabulated in Pearson and Hartley [1962]. For a sample of size 500 the one-tailed 5 % and 1 % critical values for \( \nu b_1 \) are ± 0.179 and ± 0.255, respectively, and for \( b_2 \) the one-tailed upper (lower) 5 % and 1 % critical values are 3.37 (2.67) and 3.60 (2.57), respectively.

All biases show a clear tendency to vanish as the sample size grows, in agreement with the consistency of the estimators, but convergence to normality does not seem to be fast. Most of the estimators show significant skewness and significant departures from mesokurtosis. The vast majority of the estimators
show thinner tails than the normal distribution, and in almost all cases the variance estimators VAR show thicker tails than the mean estimators OLS and GLS. The latter effects are transmitted to the mixed estimators MWA and ML depending on the relative contributions of information of the two moments.

The Poisson-N model reported in Table 3.1 is interesting because, having equal mean and variance, it should show whether the information can be extracted equally well from both moments. The information measures suggest that this is the case by showing that the contribution of both moments to overall efficiency is roughly the same, and this is in agreement with the substantial improvement in efficiency of the MLE over the GLS estimator in all sample sizes and both coefficients. In most cases, mean biases are small in magnitude and they are statistically insignificant (at the 5% level) in those estimators which use mean information only, but they are statistically significant in those estimators that use variance information. This may be due to the effect of some extreme values because of more kurtosis in the variance estimators. The efficiency gain in GLS over OLS is small and both estimators behave similarly, and GLS dominates VAR in both bias and efficiency. The latter fact casts some doubt on whether the contributions to efficiency are equal from the two moments, but still the variance contribution is clear. Except for the smallest sample, the MWA estimator performs reasonably well but it is poorer than the MLE, especially in efficiency considerations.

The ARCH models are reported in Tables 3.2 - 3.4. In contrast with the Poisson-N model in which β enters both μt and ht in identical form, in the ARCH model μt is a linear function of β, while ht is a more complex nonlinear function of β, and this naturally makes information more difficult to extract in the variance equation. The contribution of the variance information to efficiency increases with the strength of the ARCH effect as one would expect, and in the largest sample it registers 5% in the mild model, 18% in the regular model, and 30% in the strong model. The consequence is that the
efficiency gains in ML over GLS are very small in the mild model, but become clearer in the regular model and are important in the strong model. Likewise, OLS and GLS are very similar in the mild model, but GLS becomes substantially more efficient as the ARCH effect is stronger for $T = 50$ or larger. Almost all biases are statistically insignificant for the $\beta$ coefficients, but the mean biases of variance estimators are very large in magnitude. GLS is clearly superior to VAR in all cases, and the MWA estimator is dominated by the ML estimator except for the mild model in which they are very similar.

With respect to the $\alpha$ parameters, a distinctive feature of the simulation results is that the mean bias is substantially smaller in magnitude and significance for the MLE in relation to the other estimators in almost all cases. Remember that for this model GLS, VAR and ML are asymptotically equivalent, and more efficient than SLS, but the latter effect is not clear when the ARCH effect is mild. With a stronger ARCH effect, the improvement in efficiency of GLS, VAR and ML over SLS is evident for the $\alpha_0$ coefficient, but not for $\alpha_1$ in which the SLS is in general more efficient. The performance of VAR and GLS is similar, and ML performs remarkably well in the very small sample.
**TABLE 3.1 - ESTIMATION (Poisson-N, correct specification).**

Model: $y_t \sim \mathcal{N}(\beta_0 + \beta_1 x_{it}, \beta_0 + \beta_1 x_{it})$

DGP: $y_t \sim \mathcal{N}(1 + x_{it}, 1 + x_{it})$

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th></th>
<th>$T = 50$</th>
<th></th>
<th>$T = 100$</th>
<th></th>
<th>$T = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
</tr>
<tr>
<td>Inf.meas.</td>
<td>0.492</td>
<td>0.498</td>
<td>0.495</td>
<td>0.509</td>
<td>0.486</td>
<td>0.520</td>
<td>0.493</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>OLS</td>
<td>-0.014</td>
<td>0.017</td>
<td>-0.007</td>
<td>0.011</td>
<td>-0.009</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.031</td>
<td>0.016</td>
<td>-0.012</td>
<td>0.011</td>
<td>-0.014</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>0.097</td>
<td>0.030</td>
<td>0.067</td>
<td>0.015</td>
<td>0.085</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>0.053</td>
<td>0.019</td>
<td>0.021</td>
<td>0.009</td>
<td>0.034</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.048</td>
<td>0.011</td>
<td>0.032</td>
<td>0.007</td>
<td>0.020</td>
<td>0.006</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>OLS</td>
<td>-0.028</td>
<td>0.021</td>
<td>-0.019</td>
<td>0.014</td>
<td>-0.019</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.038</td>
<td>0.021</td>
<td>-0.019</td>
<td>0.015</td>
<td>-0.021</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>0.153</td>
<td>0.056</td>
<td>0.076</td>
<td>0.018</td>
<td>0.101</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>0.073</td>
<td>0.037</td>
<td>0.015</td>
<td>0.011</td>
<td>0.037</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.079</td>
<td>0.013</td>
<td>0.045</td>
<td>0.008</td>
<td>0.028</td>
<td>0.006</td>
</tr>
<tr>
<td>$\sqrt{b_1}$</td>
<td>b1</td>
<td>0.254</td>
<td>1.309</td>
<td>-0.409</td>
<td>1.312</td>
<td>0.459</td>
<td>1.328</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>1.344</td>
<td>0.327</td>
<td>1.312</td>
<td>0.459</td>
<td>1.328</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>OLS</td>
<td>-0.440</td>
<td>1.338</td>
<td>0.254</td>
<td>1.309</td>
<td>-0.409</td>
<td>1.312</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.484</td>
<td>2.856</td>
<td>0.076</td>
<td>1.313</td>
<td>-0.381</td>
<td>1.339</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>1.823</td>
<td>1.383</td>
<td>0.807</td>
<td>1.534</td>
<td>1.423</td>
<td>2.441</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>1.831</td>
<td>2.958</td>
<td>-0.304</td>
<td>1.430</td>
<td>1.330</td>
<td>2.179</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.677</td>
<td>1.357</td>
<td>0.513</td>
<td>1.327</td>
<td>0.691</td>
<td>1.320</td>
</tr>
<tr>
<td>$\sqrt{b_1}$</td>
<td>b1</td>
<td>0.472</td>
<td>1.361</td>
<td>0.087</td>
<td>1.312</td>
<td>-0.349</td>
<td>1.279</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>1.341</td>
<td>0.327</td>
<td>1.279</td>
<td>0.472</td>
<td>1.319</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\sqrt{b_1}$</th>
<th>b</th>
<th>$\sqrt{b_1}$</th>
<th>b</th>
<th>$\sqrt{b_1}$</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_3$</td>
<td>OLS</td>
<td>-0.472</td>
<td>1.483</td>
<td>0.087</td>
<td>1.312</td>
<td>-0.349</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.590</td>
<td>2.413</td>
<td>0.724</td>
<td>1.553</td>
<td>1.347</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>2.455</td>
<td>4.000</td>
<td>0.724</td>
<td>1.553</td>
<td>1.347</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>2.455</td>
<td>4.102</td>
<td>-0.590</td>
<td>1.461</td>
<td>1.160</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.806</td>
<td>1.402</td>
<td>0.708</td>
<td>1.354</td>
<td>0.818</td>
</tr>
</tbody>
</table>
### TABLE 3.2 - ESTIMATION (ARCH I, correct specification)

- **Model:** \( y_t \mid \varepsilon_t \sim N(\beta_0 + \beta_1 x_{it}, \sigma_0 + \alpha_1 \varepsilon_{it}^2) \)
- **DGP:** \( y_t \mid \varepsilon_t \sim N(0, 0.8 + 0.2 \varepsilon_{it}^2) \)

<table>
<thead>
<tr>
<th></th>
<th>( T = 20 )</th>
<th>( T = 50 )</th>
<th>( T = 100 )</th>
<th>( T = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
</tr>
<tr>
<td>Inf.meas.</td>
<td>0.925</td>
<td>0.079</td>
<td>0.935</td>
<td>0.069</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>OLS -0.012</td>
<td>0.019</td>
<td>-0.008</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>GLS -0.020</td>
<td>0.019</td>
<td>-0.006</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>VAR 2.459</td>
<td>4.031</td>
<td>8.562</td>
<td>6.858</td>
</tr>
<tr>
<td></td>
<td>MWA -0.017</td>
<td>0.019</td>
<td>0.001</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>ML -0.022</td>
<td>0.019</td>
<td>-0.004</td>
<td>0.012</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>OLS -0.033</td>
<td>0.031</td>
<td>-0.028</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>GLS -0.033</td>
<td>0.034</td>
<td>-0.020</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>MWA -0.032</td>
<td>0.034</td>
<td>-0.011</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>ML -0.046</td>
<td>0.033</td>
<td>-0.019</td>
<td>0.020</td>
</tr>
<tr>
<td>( \sigma_0 )</td>
<td>SLS 0.044</td>
<td>0.016</td>
<td>0.041</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>GLS 0.022</td>
<td>0.016</td>
<td>0.023</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>VAR 0.009</td>
<td>0.019</td>
<td>0.031</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>ML -0.040</td>
<td>0.016</td>
<td>-0.013</td>
<td>0.011</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>SLS -0.153</td>
<td>0.011</td>
<td>-0.074</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>GLS -0.137</td>
<td>0.014</td>
<td>-0.058</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>VAR -0.069</td>
<td>0.009</td>
<td>-0.033</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>ML 0.030</td>
<td>0.015</td>
<td>0.017</td>
<td>0.010</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>OLS -0.534</td>
<td>1.389</td>
<td>0.215</td>
<td>1.314</td>
</tr>
<tr>
<td></td>
<td>GLS -0.547</td>
<td>1.367</td>
<td>0.200</td>
<td>1.323</td>
</tr>
<tr>
<td></td>
<td>VAR 2.507</td>
<td>4.331</td>
<td>-2.719</td>
<td>4.540</td>
</tr>
<tr>
<td></td>
<td>MWA -0.493</td>
<td>1.331</td>
<td>0.291</td>
<td>1.319</td>
</tr>
<tr>
<td></td>
<td>ML -0.416</td>
<td>1.335</td>
<td>-0.288</td>
<td>1.315</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>OLS -0.634</td>
<td>1.397</td>
<td>-0.351</td>
<td>1.339</td>
</tr>
<tr>
<td></td>
<td>GLS -0.608</td>
<td>1.364</td>
<td>-0.472</td>
<td>1.357</td>
</tr>
<tr>
<td></td>
<td>VAR 2.705</td>
<td>4.561</td>
<td>-2.810</td>
<td>4.716</td>
</tr>
<tr>
<td></td>
<td>MWA -0.532</td>
<td>1.340</td>
<td>0.196</td>
<td>1.367</td>
</tr>
<tr>
<td></td>
<td>ML -0.626</td>
<td>1.358</td>
<td>-0.492</td>
<td>1.368</td>
</tr>
<tr>
<td>( \sigma_0 )</td>
<td>SLS 1.015</td>
<td>1.539</td>
<td>0.613</td>
<td>1.281</td>
</tr>
<tr>
<td></td>
<td>GLS 0.977</td>
<td>1.504</td>
<td>0.704</td>
<td>1.278</td>
</tr>
<tr>
<td></td>
<td>VAR 1.109</td>
<td>1.568</td>
<td>1.034</td>
<td>1.632</td>
</tr>
<tr>
<td></td>
<td>ML 0.925</td>
<td>1.468</td>
<td>0.629</td>
<td>1.271</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>SLS 0.791</td>
<td>1.458</td>
<td>0.886</td>
<td>1.438</td>
</tr>
<tr>
<td></td>
<td>GLS 0.805</td>
<td>1.422</td>
<td>0.667</td>
<td>1.272</td>
</tr>
<tr>
<td></td>
<td>VAR 1.266</td>
<td>1.671</td>
<td>0.987</td>
<td>1.348</td>
</tr>
<tr>
<td></td>
<td>ML 1.267</td>
<td>1.674</td>
<td>1.102</td>
<td>1.521</td>
</tr>
</tbody>
</table>
TABLE 3.3 - ESTIMATION (ARCH II, correct specification).

Model: $y_t \sim N(\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2)$
DGP: $y_t \sim N(0, 0.2 + 0.8 u_{t-1}^2)$

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th></th>
<th>$T = 50$</th>
<th></th>
<th>$T = 100$</th>
<th></th>
<th>$T = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
</tr>
<tr>
<td>Inf.meas.</td>
<td>0.783</td>
<td>0.225</td>
<td>0.723</td>
<td>0.277</td>
<td>0.711</td>
<td>0.289</td>
<td>0.703</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th></th>
<th>$\beta_0$</th>
<th></th>
<th>$\beta_0$</th>
<th></th>
<th>$\beta_0$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>-0.003</td>
<td>0.017</td>
<td>-0.007</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.010</td>
<td>-0.004</td>
<td>0.007</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.008</td>
<td>0.018</td>
<td>-0.008</td>
<td>0.009</td>
<td>-0.010</td>
<td>0.006</td>
<td>-0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.863</td>
<td>0.485</td>
<td>0.633</td>
<td>0.439</td>
<td>-1.626</td>
<td>1.784</td>
<td>-0.265</td>
<td>0.259</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.012</td>
<td>0.017</td>
<td>0.006</td>
<td>0.010</td>
<td>-0.008</td>
<td>0.008</td>
<td>-0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>ML</td>
<td>-0.008</td>
<td>0.012</td>
<td>0.001</td>
<td>0.007</td>
<td>-0.001</td>
<td>0.005</td>
<td>-0.001</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SLS</td>
<td>0.350</td>
<td>0.107</td>
<td>0.309</td>
<td>0.042</td>
<td>0.282</td>
<td>0.024</td>
<td>0.279</td>
<td>0.021</td>
</tr>
<tr>
<td>GLS</td>
<td>0.243</td>
<td>0.073</td>
<td>0.063</td>
<td>0.007</td>
<td>0.031</td>
<td>0.004</td>
<td>0.010</td>
<td>0.002</td>
</tr>
<tr>
<td>VAR</td>
<td>0.237</td>
<td>0.063</td>
<td>0.259</td>
<td>0.141</td>
<td>0.061</td>
<td>0.008</td>
<td>0.029</td>
<td>0.003</td>
</tr>
<tr>
<td>ML</td>
<td>0.031</td>
<td>0.008</td>
<td>0.006</td>
<td>0.004</td>
<td>0.001</td>
<td>0.003</td>
<td>0.000</td>
<td>0.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SLS</td>
<td>-0.574</td>
<td>0.016</td>
<td>-0.438</td>
<td>0.014</td>
<td>-0.333</td>
<td>0.008</td>
<td>-0.354</td>
<td>0.007</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.463</td>
<td>0.024</td>
<td>-0.192</td>
<td>0.020</td>
<td>-0.096</td>
<td>0.012</td>
<td>-0.032</td>
<td>0.009</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.500</td>
<td>0.014</td>
<td>-0.282</td>
<td>0.016</td>
<td>-0.164</td>
<td>0.011</td>
<td>-0.082</td>
<td>0.008</td>
</tr>
<tr>
<td>ML</td>
<td>-0.164</td>
<td>0.028</td>
<td>-0.043</td>
<td>0.017</td>
<td>-0.011</td>
<td>0.011</td>
<td>-0.002</td>
<td>0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.752</td>
<td>2.058</td>
<td>0.721</td>
<td>1.661</td>
<td>1.097</td>
<td>2.065</td>
<td>1.038</td>
<td>1.866</td>
</tr>
<tr>
<td>GLS</td>
<td>1.352</td>
<td>2.346</td>
<td>-0.209</td>
<td>1.441</td>
<td>-0.477</td>
<td>1.351</td>
<td>0.575</td>
<td>1.371</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.725</td>
<td>2.014</td>
<td>2.508</td>
<td>4.219</td>
<td>-2.776</td>
<td>4.664</td>
<td>-2.807</td>
<td>4.712</td>
</tr>
<tr>
<td>MWA</td>
<td>0.576</td>
<td>2.301</td>
<td>-0.627</td>
<td>1.561</td>
<td>-0.610</td>
<td>1.663</td>
<td>0.607</td>
<td>2.260</td>
</tr>
<tr>
<td>ML</td>
<td>-0.652</td>
<td>1.473</td>
<td>-0.487</td>
<td>1.392</td>
<td>0.357</td>
<td>1.322</td>
<td>0.503</td>
<td>1.308</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
<th>$\sqrt{\beta_1}$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>-0.652</td>
<td>1.922</td>
<td>0.605</td>
<td>1.613</td>
<td>0.940</td>
<td>1.852</td>
<td>0.820</td>
<td>1.612</td>
</tr>
<tr>
<td>GLS</td>
<td>1.728</td>
<td>2.911</td>
<td>-0.840</td>
<td>1.660</td>
<td>-0.300</td>
<td>1.335</td>
<td>0.576</td>
<td>1.355</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.672</td>
<td>2.172</td>
<td>2.714</td>
<td>4.544</td>
<td>-2.768</td>
<td>4.652</td>
<td>-2.711</td>
<td>4.557</td>
</tr>
<tr>
<td>MWA</td>
<td>1.577</td>
<td>2.998</td>
<td>-0.594</td>
<td>1.678</td>
<td>0.695</td>
<td>1.780</td>
<td>1.510</td>
<td>2.504</td>
</tr>
<tr>
<td>ML</td>
<td>-0.678</td>
<td>1.460</td>
<td>-0.641</td>
<td>1.481</td>
<td>-0.396</td>
<td>1.301</td>
<td>0.403</td>
<td>1.284</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
<th>$\alpha_0$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GLS</td>
<td>2.591</td>
<td>4.262</td>
<td>-0.735</td>
<td>2.262</td>
<td>1.146</td>
<td>1.706</td>
<td>-0.922</td>
<td>1.706</td>
</tr>
<tr>
<td>VAR</td>
<td>2.655</td>
<td>4.436</td>
<td>2.792</td>
<td>4.685</td>
<td>2.326</td>
<td>3.872</td>
<td>1.794</td>
<td>2.859</td>
</tr>
<tr>
<td>ML</td>
<td>1.159</td>
<td>1.572</td>
<td>1.031</td>
<td>1.541</td>
<td>0.867</td>
<td>1.421</td>
<td>0.694</td>
<td>1.310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
<th>$\alpha_1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SLS</td>
<td>1.486</td>
<td>2.347</td>
<td>1.830</td>
<td>2.941</td>
<td>0.655</td>
<td>1.362</td>
<td>0.430</td>
<td>1.254</td>
</tr>
<tr>
<td>GLS</td>
<td>1.082</td>
<td>2.071</td>
<td>1.754</td>
<td>2.946</td>
<td>0.618</td>
<td>1.348</td>
<td>0.785</td>
<td>1.484</td>
</tr>
<tr>
<td>VAR</td>
<td>1.065</td>
<td>1.445</td>
<td>1.416</td>
<td>2.345</td>
<td>-0.130</td>
<td>1.324</td>
<td>0.539</td>
<td>1.462</td>
</tr>
<tr>
<td>ML</td>
<td>1.000</td>
<td>1.401</td>
<td>0.553</td>
<td>1.311</td>
<td>0.465</td>
<td>1.375</td>
<td>-0.263</td>
<td>1.356</td>
</tr>
<tr>
<td></td>
<td>T = 20</td>
<td></td>
<td>T = 50</td>
<td></td>
<td>T = 100</td>
<td></td>
<td>T = 200</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
<td>mean</td>
<td>var.</td>
</tr>
<tr>
<td>Inf. meas.</td>
<td>0.854</td>
<td>0.148</td>
<td>0.829</td>
<td>0.171</td>
<td>0.828</td>
<td>0.172</td>
<td>0.823</td>
<td>0.177</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>OLS</td>
<td>-0.009</td>
<td>0.018</td>
<td>-0.010</td>
<td>0.013</td>
<td>-0.015</td>
<td>0.010</td>
<td>-0.005</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.020</td>
<td>0.018</td>
<td>-0.012</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.009</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>-5.274</td>
<td>3.755</td>
<td>-0.226</td>
<td>0.305</td>
<td>-0.209</td>
<td>0.195</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>-0.020</td>
<td>0.017</td>
<td>-0.001</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.008</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.021</td>
<td>0.017</td>
<td>-0.002</td>
<td>0.010</td>
<td>-0.007</td>
<td>0.008</td>
<td>-0.002</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>OLS</td>
<td>0.201</td>
<td>0.020</td>
<td>0.167</td>
<td>0.012</td>
<td>0.144</td>
<td>0.009</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>0.126</td>
<td>0.016</td>
<td>0.070</td>
<td>0.014</td>
<td>0.037</td>
<td>0.006</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>0.134</td>
<td>0.017</td>
<td>0.080</td>
<td>0.009</td>
<td>0.047</td>
<td>0.006</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.023</td>
<td>0.014</td>
<td>0.004</td>
<td>0.008</td>
<td>0.002</td>
<td>0.006</td>
<td>0.000</td>
</tr>
<tr>
<td>$\sqrt{\beta_1}$</td>
<td>OLS</td>
<td>-0.691</td>
<td>1.479</td>
<td>-0.042</td>
<td>1.388</td>
<td>0.385</td>
<td>1.470</td>
<td>0.659</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>OLS</td>
<td>-0.602</td>
<td>1.424</td>
<td>-0.437</td>
<td>1.332</td>
<td>-0.402</td>
<td>1.322</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>-2.531</td>
<td>4.127</td>
<td>-1.424</td>
<td>2.416</td>
<td>-1.752</td>
<td>2.898</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>-0.512</td>
<td>1.366</td>
<td>-0.720</td>
<td>1.450</td>
<td>-0.393</td>
<td>1.339</td>
<td>0.332</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.547</td>
<td>1.369</td>
<td>-0.485</td>
<td>1.350</td>
<td>-0.403</td>
<td>1.346</td>
<td>0.513</td>
</tr>
<tr>
<td>$\sqrt{\alpha_0}$</td>
<td>OLS</td>
<td>-0.763</td>
<td>1.502</td>
<td>-0.291</td>
<td>1.398</td>
<td>0.465</td>
<td>1.417</td>
<td>0.602</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>-0.668</td>
<td>1.407</td>
<td>-0.571</td>
<td>1.384</td>
<td>-0.386</td>
<td>1.291</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>-2.682</td>
<td>4.502</td>
<td>-0.989</td>
<td>2.225</td>
<td>1.284</td>
<td>2.992</td>
<td>2.404</td>
</tr>
<tr>
<td></td>
<td>MWA</td>
<td>-0.535</td>
<td>1.372</td>
<td>0.281</td>
<td>1.396</td>
<td>-0.437</td>
<td>1.320</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.672</td>
<td>1.396</td>
<td>-0.605</td>
<td>1.432</td>
<td>-0.499</td>
<td>1.317</td>
<td>0.419</td>
</tr>
<tr>
<td>$\sqrt{\theta_1}$</td>
<td>OLS</td>
<td>1.454</td>
<td>2.150</td>
<td>0.833</td>
<td>1.751</td>
<td>1.049</td>
<td>1.558</td>
<td>1.063</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>1.260</td>
<td>1.867</td>
<td>2.198</td>
<td>3.684</td>
<td>0.887</td>
<td>1.389</td>
<td>0.730</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>1.299</td>
<td>1.836</td>
<td>0.940</td>
<td>1.454</td>
<td>0.901</td>
<td>1.394</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>1.150</td>
<td>1.706</td>
<td>0.830</td>
<td>1.333</td>
<td>0.853</td>
<td>1.386</td>
<td>0.633</td>
</tr>
<tr>
<td>$\sqrt{\theta_4}$</td>
<td>OLS</td>
<td>1.142</td>
<td>1.849</td>
<td>1.382</td>
<td>2.245</td>
<td>0.681</td>
<td>1.292</td>
<td>0.608</td>
</tr>
<tr>
<td></td>
<td>GLS</td>
<td>0.944</td>
<td>1.535</td>
<td>-1.843</td>
<td>3.135</td>
<td>0.411</td>
<td>1.298</td>
<td>-0.446</td>
</tr>
<tr>
<td></td>
<td>VAR</td>
<td>1.083</td>
<td>1.475</td>
<td>0.839</td>
<td>1.345</td>
<td>0.494</td>
<td>1.301</td>
<td>-0.456</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>1.066</td>
<td>1.435</td>
<td>0.764</td>
<td>1.302</td>
<td>0.596</td>
<td>1.341</td>
<td>-0.351</td>
</tr>
</tbody>
</table>
CHAPTER 4

THE ROBUSTNESS OF THE QMLE.¹

In this chapter we consider misspecification of the model and analyze its implications on the asymptotic distribution of the QMLE \( \hat{\theta} \) of \( \theta \). The main issues here are the consequences of specification error on the consistency of the QMLE and on the possibility of drawing asymptotically correct inferences based on its distribution. Sufficient regularity is preserved for \( \hat{\theta} \) to remain asymptotically normal (e.g. Domowitz and White [1982]) and so we need only concentrate on the first two moments of the asymptotic distribution of \( \hat{\theta} \).

Various aspects of estimating under misspecification have been analyzed by Burguete et al [1982], Domowitz and White [1982], Gourieroux et al [1984a], and White [1982a] inter alia. Here we use their results on conditionally heteroskedastic models, exploiting the insights gained on the estimation of such models in Chapter 3.

Our basic framework is that of estimating on the presumption that (Q0') holds when this is not the case. We replace this "correct specification" assumption with different possibilities of misspecification. Assumptions (Q1) - (Q8) are preserved to maintain regularity, though (Q2), (Q4), (Q5) and (Q7) must now refer to the pseudo-true value \( \theta^* = (\beta^*, \alpha^*)' \) rather than to the "true" value \( \theta_0 \) which, depending on the nature of the misspecification, may be void of meaning. The pseudo-true value is such that \( \hat{\theta} \rightarrow_{\text{a.s.}} \theta^* \).

¹ In this Chapter Sections § 4.1 and § 4.3 are based on joint work done with A.R. Pagan, reported in Pagan and Sabau [1987a].
In section § 4.1 we study the conditions that the likelihood function must obey to preserve consistency of the QMLE. These conditions are used in the following sections together with the MWA decomposition given in Chapter 3. Section § 4.2 is devoted to specification error in the conditional mean, while § 4.3 analyzes misspecification of the conditional variance. Cases are presented in which the estimators are robust in the presence of certain classes of departures, and the ARCH model is particularly interesting in this respect. Section § 4.4 is then concerned with misspecification of third and higher order moments—departures from normality—where it is made clear that misspecification of such moments does not affect the consistency of the QMLE. Specification error in the third and fourth moments, however, may produce incorrect inferences. Some Monte Carlo evidence on the behavior of estimators under specification error is discussed in § 4.5.

§ 4.1 Some conditions on the likelihood function

We start this section with the additional assumption

$$(\text{B0})$$

The solution $\theta^*_T$ of the normal equations $\mathbb{E}[d_\theta(\theta)] = 0$ exists for sufficiently large $T$, where the score under normality is given by

$$d_\theta(\theta) = T^{-1} G' \Sigma^{-1} v.$$ Further the sequence $\{\theta^*_T\}$ converges almost surely to the non-stochastic limit $\theta^*$. This assumption implies that the solutions, $\beta^*_m$ of $\mathbb{E}[X' \Omega^{-1} u(\beta)] = 0$, and $\phi^*_v$ of $\mathbb{E}[S_\phi' \Omega^{-2} \epsilon(\theta)] = 0$ exist, and the sequences $\{\beta^*_m\}$ and $\{\phi^*_v\}$ converge almost surely to nonstochastic limits given by $\beta^*_m$ and $\phi^*_v$, respectively.

$$(\text{C7})$$ implies that $\theta_0$ is identifiably unique, and Theorem 2.2 and Corollary 3.3 of Domowitz and White [1982] ensure that $\hat{\theta} \overset{\text{as}}{\to} \theta^*$ and $T^{1/2}(\hat{\theta} - \theta^*)$ is
asymptotically normal. Similarly, \( \hat{\beta}_m \overset{as}{\rightarrow} \beta_m^* \) and \( \hat{\phi}_v \overset{as}{\rightarrow} \phi_v^* \), and the asymptotic normality of \( T^{1/2}(\hat{\beta}_m - \beta_m^*) \) and \( T^{1/2}(\hat{\phi}_v - \phi_v^*) \) also follow.

The concept of the pseudo-true value of \( \theta \) to which the QMLE \( \hat{\theta} \) converges almost surely is now clear. An alternative equivalent statement is

\[
\mathbb{E} \{ d_\theta (\theta^*) \} = 0 ,
\]

where the expectation is taken with respect to the true probability measure. This is Lemma 1 in Pagan and Sabau [1987a] (PS in the remaining of this Chapter), and also follows from Domowitz and White [1982] and Gourieroux et al [1984a] inter alia. Thus if \( \theta^* = \theta_0 \), \( \hat{\theta} \) is a consistent estimator under misspecification and (1) states the conditions that the likelihood function must satisfy for this purpose.

It is convenient to analyze separately the consistency of the parameter subvectors \( \beta \) and \( \alpha \), and for this purpose we use

**Lemma 4.1.** - If \( d_\beta (\beta_0 , \alpha^*) - \mathbb{E} [ d_\beta (\beta_0 , \alpha^*) ] \overset{as}{\rightarrow} 0 \) and \( J_{\theta \theta} (\hat{\theta}) + J_{\theta \alpha} (\hat{\theta}) \overset{as}{\rightarrow} 0 \), where \( J_{\theta \theta} = \partial^2 \ell /\partial \theta \partial \theta' \), and \( J_{\theta \alpha} = - \mathbb{E} ( J_{\theta \theta} ) \) is positive definite, a necessary and sufficient condition for \( \hat{\beta} \) to consistently estimate \( \beta_0 \) is that \( \mathbb{E} \{ d_\beta (\beta_0 , \alpha^*) \} = 0 \).

**Proof:** Necessity follows from (1). For sufficiency expand \( d_\beta (\hat{\beta} , \hat{\alpha}) \) around \( (\beta_0 , \alpha^*) \) to get

\[
d_\beta (\hat{\beta} , \hat{\alpha}) = 0 = d_\beta (\beta_0 , \alpha^*) + J_{\beta \beta} (\overline{\theta}) (\hat{\beta} - \beta_0 ) + J_{\beta \alpha} (\overline{\theta}) (\hat{\alpha} - \alpha^*) ,
\]

where \( \overline{\theta} = (\overline{\beta}' , \overline{\alpha}') \) lies between \( \theta^* = (\beta_0' , \alpha^*) \) and \( \hat{\theta} \). Taking limits we get

\[
0 = d_\beta (\beta_0 , \alpha^*) - J_{\beta \beta} (\beta_0 , \alpha^*) (\hat{\beta} - \beta_0 ) - J_{\beta \alpha} (\beta_0 , \alpha^*) (\hat{\alpha} - \alpha^*) + o_p (1) .
\]

Since \( \hat{\alpha} \overset{as}{\rightarrow} \alpha^*, \) \( \hat{\beta} - \beta_0 \overset{as}{\rightarrow} 0 \) provided \( J_{\beta \beta} (\theta^*) \) is positive definite and \( d_\beta (\beta_0 , \alpha^*) - \mathbb{E} [ d_\beta (\beta_0 , \alpha^*) ] \overset{as}{\rightarrow} 0 \), so \( \mathbb{E} \{ d_\beta (\beta_0 , \alpha^*) \} = 0 \) is a sufficient condition. \(\square\)
This is Lemma 2 in PS. By a symmetric argument $\mathcal{E} \{d_{\alpha}(\beta^*, \alpha_0)\} = 0$ is a necessary and sufficient condition for the consistency of $\hat{\alpha}$. These conditions are applied to normal heteroskedastic models by substituting the pseudo-score $d_{\theta}(\theta) = T^{-1} G' \Sigma^{-1} u$, as we do in

**Lemma 4.2.** - Let

$$\varphi_\beta(\theta) = \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t E [ u_t | \mathcal{F}_t ] \right\} + \frac{1}{2} \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} h_t^{-2} w_t E [ \varepsilon_t | \mathcal{F}_t ] \right\},$$

$$\varphi_\alpha(\theta) = \frac{1}{2} \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t E [ \varepsilon_t | \mathcal{F}_t ] \right\},$$

and

$$\varphi_\theta(\theta) = (\varphi_\beta(\theta)', \varphi_\alpha(\theta)')'.$$

Then under (\text{H0}) and (\text{Q1}) - (\text{Q7}),

(i) $\hat{\beta}$ is a consistent estimator of $\beta_0$ if, and only if, $\varphi_\beta(\beta_0, \alpha^*) = 0$,

(ii) $\hat{\alpha}$ is a consistent estimator of $\alpha_0$ if, and only if, $\varphi_\alpha(\beta^*, \alpha_0) = 0$,

and

(iii) $\hat{\theta}$ is a consistent estimator of $\theta_0$ if, and only if, $\varphi_\theta(\theta_0) = 0$.

**Proof:** (\text{H0}) and (\text{Q1}) - (\text{Q7}) are sufficient for the conditions of Lemma 4.1 (see White and Domowitz [1982]). Using iterated expectations on the pseudo-scores $d_{\beta}(\theta) = T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} w_t \varepsilon_t$ and $d_{\alpha}(\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t \varepsilon_t$ and using Lemma 4.1 shows that $\varphi_\beta(\beta_0, \alpha^*) = 0$ and $\varphi_\alpha(\beta^*, \alpha_0) = 0$ are the necessary and sufficient condition for the consistency of $\hat{\beta}$ and $\hat{\alpha}$, respectively. The joint result proves (iii).

This result generalizes Theorem 2 of PS and will be used repeatedly in the sections to follow. The two terms of $\varphi_\beta(\theta)$ in (2) are associated with consistent estimation of $\beta$ (or the estimable functions of $\beta$) in the mean equation and in the variance equation, respectively. This corresponds to the MWA decomposition of $\hat{\beta}$ (or of $\hat{\gamma}$). If the first term of $\varphi_\beta(\theta)$ is nonzero, $\hat{\beta}_m$ is
inconsistent. If the second term is nonzero \( \hat{\gamma} \) is inconsistent. These inconsistencies are transmitted to \( \hat{\gamma} \), and thus to \( \hat{\beta} \), by the MWA in (3.34)

\[
\gamma^* = (I_k^* - \Pi_u^*) \gamma_m^* + \Pi_u^* \gamma_v^*,
\]

and either \( \gamma_m^* \neq \gamma_0 \) or \( \gamma_v^* \neq \gamma_0 \) results in \( \gamma^* \neq \gamma_0 \) in general. For \( \hat{\alpha} \), it is clear from (3) that all inconsistency arises through \( \hat{\alpha}_v \).

These conditions are related to more specific forms of specification error in the following three sections.

\section*{§4.2 Specification error in the conditional mean}

The consequences of misspecification of the mean function in regression have been the subject of a vast literature and one of systematic treatment in econometrics textbooks (e.g. Theil [1971], Intriligator [1978], Amemiya [1985], Spanos [1986], \textit{inter alia}). In general, the regression parameters are inconsistent. Thus in our model estimators obtained from the mean equation, \( \hat{\beta}_m \) in particular, are inconsistent.

Suppose the true conditional mean of \( y_t \) is \( \mu_t \neq \mu_t \). Then

\[
E[u_t \mid F_t] = E[y_t - \mu_t \mid F_t] = \mu_t - \mu_t \neq 0,
\]

and the first term of \( \varphi_{\beta}(\theta) \) is nonzero in general. Substitute (C0) by

(C0-m) \( y_t \mid F_t \sim N(\mu_t, h_t), \mu_t \neq \mu_t(\beta) \) for any \( \beta \in \mathcal{B} \).

There are many factors which may produce misspecification of the conditional mean:

- incorrect parameterization of the \textit{conditional} mean, resulting in violating of the exogeneity assumptions and causing \( \mu_t \neq \mu_t \).

- incorrect functional form or dynamic specification,
- autocorrelation so that $\mu_t = \mu_t + \sum_j \rho_j u_{t-j}$

- the parameters are not constant through time, so $\mu_t = \mu_t (\beta_t)$ and by the MVT $\mu_t - \mu_t = x_t (\beta_t)'b_t$, where $b_t = \beta_t - \beta$ and $B_t \in [\beta, \beta_t]$,

for some definition of $\beta$ (see Chapter 8).

Consider autocorrelation for example. In homoskedastic models, it is well known that autocorrelation does not affect consistency unless the regressors include lagged dependent variables (e.g. Durbin [1970]). To allow for heteroskedasticity, we also require that $h_t$ not be a function of lagged $y$'s and $u$'s so we can have $E[h_t^{-1} x_t (\mu_t - \mu_t)] = E[h_t^{-1} x_t \sum_j \rho_j u_{t-j}] = 0$, and the first term of $\varphi_\beta(0)$ vanishes. This preserves the consistency of $\hat{\beta}_m$.

If $E[u_t | \mathcal{F}_t] \neq 0$ then the $u_t$ are not the innovations in the conditional mean and so $\text{Var}[y_t | \mathcal{F}_t] = h_t = E[u_t^2 | \mathcal{F}_t]$. Thus specification error in the conditional mean induces specification error in the variance equation even when the conditional variance is correctly specified. In fact we have that $u_t = y_t - \mu_t = (y_t - \mu_t) + (\mu_t - \mu_t)$, and because $y_t - \mu_t$ are the true innovations and $\mu_t, \mu_t \in \mathcal{F}_t$ we get

$$E[u_t^2 | \mathcal{F}_t] = h_t + (\mu_t - \mu_t)^2.$$  

A second problem in the variance equation appears when $h_t$ depends on $\mu_{t-j}$, $j \geq 0$, directly or through dependence on $u_{t-j}$, $j \geq 1$. In this case specification error in $\mu_t$ affects also the specification of the conditional variance even when $h_t$ is correctly specified. The true conditional variance is $\mu_t = h_t (\mu_{t-j})$, and so

$$E[u_t^2 | \mathcal{F}_t] = \mu_t + (\mu_t - \mu_t)^2,$$

producing

$$E[e_t | \mathcal{F}_t] = E[u_t^2 - h_t | \mathcal{F}_t] = (\mu_t - \mu_t)^2 + (\mu_t - h_t).$$  

- (6)
This renders the second term in $\varphi_{\beta}(\theta)$ nonzero in general, and thus misspecification of the conditional mean has a double effect on the consistency of $\hat{\beta}$. It causes $\beta_m^* \neq \beta_0$, and it induces specification error in the variance equation making $\phi_{v}^* \neq \phi_0$. This clearly renders $\beta^* \neq \beta_0$, and $\alpha^* \neq \alpha_0$.

There may be conditions under which $\hat{\beta}$ may remain consistent. We have already considered a case in which $\beta_m \overset{as}{\rightarrow} \beta_0$. Note that when there is autocorrelation and $h_t$ is not a function of $\beta$, the second term of $\varphi_{\beta}$ vanishes because $w_t = \partial h_t / \partial \beta = 0$. Thus noting that (BO), (Q0-m) and (Q1) - (Q7) are sufficient for Lemma 4.2, this completes the proof of

Lemma 4.3.- Under (BO), (Q0-m) and (Q1) - (Q7) the QMLE $\hat{\beta}$ of $\beta$ is consistent when the misspecification in (Q0-m) is due only to the presence of serial correlation of the errors of the mean equation if

(i) $\mu_t$ and $h_t$ are functions of strongly exogenous variables only,

and (ii) $h_t$ does not depend on $\beta$.

Condition (ii) avoids any problem with the estimation of $\beta$ in the variance equation by simply excluding it from the parameterization of the conditional variance. This may be weakened in a way that consistency of the mean equation estimator $\hat{\beta}_m$ is retained, as we show in

Lemma 4.4.- Under (BO), (Q0-m) and (Q1) - (Q7) and the misspecification in (Q0-m) is due only to the presence of serial correlation of errors in the mean equation, then if $\mu_t$ is a functions of strongly exogenous variables only and $h_t$ is an even function of $u_{t-j}$ conditional on $\mathcal{F}_t$, $\mathbb{E} \{ T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t \mathbb{E} \{ u_t | \mathcal{F}_t \} \} = 0$.

Proof: $\mathbb{E} \{ h_t^{-1} x_t u_t \} = \mathbb{E} \{ h_t^{-1} x_t \sum_{j} \rho_j u_{t-j} \}$, and under the assumptions $h_t^{-1} x_t$ is a conditionally even function of $u_{t-j}$, whereas $u_{t-j}$ is a conditionally odd function. The distribution of $u_t$ is symmetric because $u_t = (y_t - \bar{\mu}_t) + \sum_{j} \rho_j u_{t-j}$, and therefore $\mathbb{E} \{ h_t^{-1} x_t u_t \} = 0$ from Lemma 3 of PS.
For the ARCH class $h_t$ is conditionally even as required by the lemma, and so a sufficient condition for $\hat{\beta}_m$ to remain consistent under autocorrelation in this class is that $\mu_t$ be a function of strongly exogenous variables only, and this may be further relaxed to require only that $x_t$ be a conditionally even function of $u_{t-j}$. To preserve $\hat{\gamma}_v$ from inconsistency when $h_t$ depends on $\beta$ is much harder and will seldom be the case. If $w_t = \partial h_t / \partial \beta$ is an odd function of $u_{t-j}$ conditional on $\mathcal{F}_t$, the first term in (6) does not introduce inconsistency in $\hat{\gamma}_v$ because $h_t^{-2} w_t (\bar{\mu}_t - \mu_t)^2$ is conditionally odd and its expectation vanishes. This is the case in the ARCH model, where $w_t = -2 \sum_{j=1}^{q} \alpha_j u_{t-j} x_{t-j}$, but the term in $h_{t} - h_t$ is not a conditionally even function of $u_{t-j}$ under autocorrelation and thus will in general induce inconsistency in $\hat{\beta}_v$.

The expectation in (6) also illustrates that $\varphi_{\alpha}(\theta) \neq 0$ in general and thus the QMLE $\hat{\alpha}$ is inconsistent. Indeed it may be very difficult to preserve consistency here. In general, $z_t$ is an even function of $u_{t-j}$ conditional on $\mathcal{F}_t$. This is clear when $h_t$ has the linear-in-$\alpha$ structure $h_t = z_t(\beta)' \alpha$. The evenness of $z_t$ is practically a requirement of $h_t > 0$. If $z_t$ is conditionally even then $\varphi_{\alpha}(\theta) = 0$ because $h_t^{-2} (\bar{\mu}_t - \mu_t)^2$ is also conditionally even.

§ 4.3 Specification error in the conditional variance

Carroll and Ruppert [1982a] and PS have investigated the robustness of $\hat{\beta}$ — in the consistency sense — under misspecification of $h_t$. In general $\hat{\beta}$ is inconsistent. An intuitive explanation is that specification error in the variance equation renders $\hat{\gamma}_v$ inconsistent and this inconsistency is transmitted through the MWA to $\hat{\gamma}$ and hence to $\hat{\beta}$. By a similar argument, $\hat{\alpha}$ is also inconsistent in general.

(2) For our purposes it is equally relevant that a function be an odd function $(f(-x) = -f(x))$, or that it be a linear combination of odd functions in view of the linearity of the operator $E[\cdot]$, and thus we will denote either case by simply saying that "$f$ is an odd function."
Suppose the true conditional variance of $y_t$ is $\Gamma_t \neq h_t$, and replace (C10) by

(C10-v) $y_t \mid \mathcal{F}_t \sim N(\mu_t, \Gamma_t)$, $\Gamma_t \neq h_t$.

Then $E[ u_t \mid \mathcal{F}_t ] = 0$ still holds, but

$$E[ \varepsilon_t \mid \mathcal{F}_t ] = \Gamma_t - h_t,$$

and therefore

$$\varphi_\theta(\theta) = \frac{1}{2} \mathcal{E} \left( \sum_{t=1}^{T} h_t^{-2} w_t ( \Gamma_t - h_t ) \right),$$

and

$$\varphi_\alpha(\theta) = \frac{1}{2} \mathcal{E} \left( \sum_{t=1}^{T} h_t^{-2} z_t ( \Gamma_t - h_t ) \right).$$

A well known result is

**Lemma 4.5.** Under (B0), (C10-v) and (C11) - (C7) the QMLE $\hat{\beta}$ is consistent when $h_t$ does not depend on $\beta$.

**Proof:** $w_t = \frac{\partial h_t}{\partial \beta} = 0$ and therefore $\varphi_\theta(\theta) = 0$.

An immediate consequence of this lemma is that $\hat{\beta}_m$ remains consistent under variance misspecification. Thus (5) becomes

$$\gamma^* = \gamma_0 + \Pi_u^* ( \gamma_v^* - \gamma_0 ),$$

and $\gamma^* = \gamma_0$ only if $\gamma_v^* = \gamma_0$ because $\Pi_u^*$ is nonsingular. But it must be noted that the covariance matrix of $\hat{\beta}_m$ is not $\mathcal{E} \left( T^{-1} X^T \Omega^{-1} X \right)$ and so basing inferences on the GLS output for $\hat{\beta}_m$ may lead to incorrect inferences when $h_t$ is misspecified. White's [1980b] heteroskedasticity-robust covariance matrix must be used to produce correct asymptotic standard errors, and it must be kept in mind that biases in small samples may be substantial (Chesher and Jewitt [1987]). Clearly, the correct asymptotic distribution of $\hat{\beta}_m$ is

$$T^{1/2} (\hat{\beta}_m - \beta_0) \overset{d}{\rightarrow} N(0, \mathcal{E} \left( T(X^T \Omega^{-1} X)^{-1} X \Omega^{-1} \Gamma \Omega^{-1} X^T (X^T \Omega^{-1} X)^{-1} \right)), $$

where $\Omega^* = \Omega(\theta^*)$ and $\Gamma = \text{diag} \left( \Gamma_t \right)$.
For the Amemiya [1973] and the Poisson-type models (Cameron and Trivedi [1985]) it is very unlikely that the conditions for $\varphi_0(\theta) = 0$ will be met. These cases are studied in detail in Theorems 3 and 4 of PS. This suggests that one should also consider the alternative estimators $\hat{\beta}_L$ and $\hat{\beta}_m$, which are robust, as the risk of inconsistency in $\hat{\beta}$ is high. Further, the semi-parametric approaches of Carroll [1982] and Robinson [1987] appear as attractive alternatives not requiring specification of the conditional variance, although their small sample performance remains to be investigated. Careful diagnostic testing of the specification is the least one can suggest and this will be tackled in Chapters 5 and 6. The same happens in these models with $\hat{\alpha}$ since it is also unlikely that $\varphi_\alpha(\theta) = 0$.

The most interesting case from the consistency-robustness point of view seems to be that of the ARCH class of models. Robustness of $\hat{\beta}$ results when $u_t$ is symmetrically distributed and $\overline{h}_t$ is a conditionally even function of $u_{t-j}$, as we show in

**Lemma 4.6.** Under (B0), (C0-v) and (C1) - (C7) the QMLE $\hat{\beta}$ is consistent when $h_t$ is parameterized as GARCH, $\overline{h}_t$ is a conditionally even function of $u_{t-j}$ and $u_t$ is symmetrically distributed around zero.

**Proof:** Since $h_t$ and $\overline{h}_t$ are conditionally even, it follows that $h_t^{-2}(\overline{h}_t - h_t)$ is conditionally even. But $w_t = \partial h_t / \partial \beta$ is conditionally odd for the GARCH class and it follows from Lemma 3 of PS that $\varphi_\beta(\theta) = 0$. 

This lemma is Theorem 5 in PS. $\hat{\beta}$ remains consistent, for example, when the conditional variance is driven by simple heteroskedasticity that is a function of strongly exogenous variables, or when the true variance is member of the symmetric ARCH class put forward by Engle [1982a] (see also Engle and Bollerslev [1986], and Geweke [1986]). A comforting result is that $\hat{\beta}$ remains consistent if the orders of the GARCH specification are incorrect. However, symmetry of the conditional distribution is quite crucial. Also observe that
\( \varphi_\alpha(\theta^*) \neq 0 \) and so \( \hat{\alpha} \) is inconsistent in these cases, thus absorbing the cost of the specification error. And, of course, the covariance matrix of \( \hat{\beta} \) reported in standard output is inconsistent.

A more surprising result, although empirically less important, appears when \( \mathbf{h}_t \) can be decomposed into a component that is a function of strongly exogenous variables and a component that is a conditionally odd function of the lagged errors. Then we have

**Lemma 4.7.** Suppose (B0), (C0-v) and (Q1) - (Q8) hold. If \( h_t \) is parameterized as ARCH and \( \mathbf{h}_t = \mathbf{h}_{xt} + \mathbf{h}_{ot} \), where \( \mathbf{h}_{xt} \) is function of strongly exogenous variables only and \( \mathbf{h}_{ot} \) is an odd function of \( u_{t-j} \), conditional on \( \mathcal{F}_t \), then the QMLE \( \hat{\beta} \) is consistent if the conditional distribution of \( y_t \) is symmetric.

**Proof:** Partition \( z_t = (1, z_{t-1}^\prime) \) and \( \alpha = (\alpha_0, \alpha_1^\prime) \) and observe that the derivative \( w_t = -2 (u_{t-1} x_{t-1}, \ldots, u_{t-q} x_{t-q}) \alpha_1 \) vanishes when \( \alpha_1 = 0 \). Therefore, \( \alpha_1^* = 0 \) is a sufficient condition for \( \varphi_\beta(\theta) = \mathcal{E} \{ T^{-1} \sum_{t=1}^T h_t^2 w_t E[\varepsilon_t|\mathcal{F}_t] \} = 0 \).

Because \( z_t \) is conditionally even in ARCH models, it follows from Lemma 3 of PS that \( T^{-1} \sum_{t=1}^T h_t^2 z_t \mathbf{h}_{ot} \toas 0 \), and therefore,

\[
\varphi_\alpha(\theta) = \frac{1}{2} \mathcal{E} \{ T^{-1} \sum_{t=1}^T h_t^2 z_t (\mathbf{h}_{xt} - h_t) \}.
\]

Suppose \( \alpha_1^* = 0 \), so that \( h_t^* = \mathbf{h}_t^* (\beta_0, \alpha^*) = \alpha_0^* \). Then

\[
\varphi_\alpha(\beta_0, \alpha^*) = \frac{1}{2\alpha_0^*} \mathcal{E} \{ T^{-1} \sum_{t=1}^T z_t (\mathbf{h}_{xt} - \alpha_0^*) \} = \zeta \mathcal{E} \{ T^{-1} \sum_{t=1}^T (\mathbf{h}_{xt} - \alpha_0^*) \}
\]

where \( \zeta = \frac{1}{2\alpha_0^*} \mathcal{E} \{ z_t \} \), which can be factorized because of the strong exogeneity of \( \mathbf{h}_{xt} \) and the strict stationarity assumption on \( y_t \). Now consider

\[
\hat{\alpha}_0 = T^{-1} \sum_{t=1}^T \hat{h}_t^1 u_{t-1}^2 - T^{-1} \sum_{t=1}^T \hat{h}_t^1 \hat{z}_{t-1} \hat{\alpha}_1.
\]
If \( \alpha_1^* = 0 \), then \( \hat{\alpha}_0 \xrightarrow{as} \mathcal{C} \{ T^{-1} \sum_{t=1}^{T} \Gamma_{xt} \} = \alpha_0^* \), and so \( \varphi_{\alpha}(\beta_0, \alpha^*) = 0 \), showing that the pseudo-true value of \( \alpha \) has \( \alpha_1^* = 0 \). \( \square \)

This is Theorem 6 of PS. Under these circumstances \( \beta \) is not identifiable in the variance equation because \( \mathcal{C} \{ T^{-1} W^* \Omega^2 W^* \} = 0 \). Indeed, only \( \alpha \) is identifiable in the variance equation and hence the \( \hat{\beta} - \hat{\beta}_m \xrightarrow{as} 0 \). Observe again the crucial role played by the symmetry of the distribution together with the fact that \( \Gamma_t \) does not include an even term in \( u_{t-j} \) (other than a function of strongly exogenous variables). On the other hand, the inconsistency of \( \hat{\alpha} \) is obvious. The Lemma also shows that the true conditional variance must have a conditionally even component in \( u_{t-j} \) for the ARCH model to be able to extract any information about \( \beta \) from the variance equation. If \( \Gamma_t \) consists of only this even component, Lemma 4.6 establishes that \( \hat{\beta} \) is consistent as long as \( u_t \) is symmetric. But if the true variance combines conditionally even and conditionally odd components then \( \hat{\beta} \) becomes inconsistent in general (e.g. Theorem 7 of PS). As an example, consider a process with linear conditional mean \( \mu_t = x_t' \beta \), and conditional variance of the form

\[
\Gamma_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j y_{t-j}^2,
\]

and rewrite \( y_{t-j}^2 = \mu_{t-j}^2 + 2 \mu_{t-j} u_{t-j} + u_{t-j}^2 \). Then

\[
\Gamma_t - h_t = \sum_{j=1}^{q} \alpha_j (\mu_{t-j}^2 + 2 \mu_{t-j} u_{t-j}) = \sum_{j=1}^{q} \alpha_j \mu_{t-j}^2 - w_t' \beta,
\]  

because \( w_t = -2 \sum_{j=1}^{q} \alpha_j u_{t-j} x_{t-j} \). Then, apart from \( o_p(T^{-1/2}) \) terms,

\[
(\hat{\beta}_v - \beta_0) = (W' \Omega^2 W)^{-1} W' \Omega^2 \varepsilon = (W' \Omega^2 W)^{-1} W' \Omega^2 [\varepsilon + \sum_{j=1}^{q} \alpha_j \mu_{t-j}^2 - W\beta_0],
\]

where \( \varepsilon \) is the vector of true variance innovations, and \( \mu_{t-j}^2 = (\mu_{1-j},...,\mu_{T-j})' \).

Now, \( T^{-1} W' \Omega^{-2} [\varepsilon + \sum_{j=1}^{q} \alpha_j \mu_{t-j}^2] \xrightarrow{as} 0 \) because \( \varepsilon_t \) has zero mean and the term in \( \mu_{t-j}^2 \) vanishes asymptotically through even/odd effects. It follows that \( \hat{\beta}_v \xrightarrow{as} 0 \).
and therefore $\beta^* = (I_k - \Pi) \beta_0$, so the size of the inconsistency depends on the relative informativeness of the two moments about $\beta$. If the mean is very informative in relation to the variance, then $V(\hat{\beta}) = V(\hat{\beta}_m)$ and the inconsistency in $\hat{\beta}$ is negligible.

Many applications have been made of the ARCH model (see Engle and Bollerslev [1986]) and the above robustness results add a further attraction to the ARCH class. But concern has also arisen over whether the class is too restrictive, and a number of alternatives have been proposed. Weiss [1984] for example estimated ARMA models with heteroskedasticity of the form presented in § 2.3.4,

$$h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j u_{t-j}^2 + \sum_{j=1}^{r} \alpha_{q+j} y_{t-j}^2 + \alpha_{q+r+1} \mu_t^2$$

and found $\alpha_j \neq 0$ for $j > q$ frequently for economic time series which casts some doubt on the above consistency-robustness of the ARCH class because departures in the direction of (9) render the QMLE of $\beta$ inconsistent.

§ 4.4 Specification error in higher order moments

When the parameterized moments of an econometric model are correctly specified, the consequences on the QMLE of a non-normal distribution may range from mild losses in efficiency to severe inconsistencies, depending on the nature of the model. In the classical linear model, for example, because the Gauss-Markov Theorem does not depend on normality the OLS estimator may be inefficient only in relation to some other nonlinear estimator. As a contrasting example, the QMLE may be inconsistent in limited dependent variable models (e.g. Bera et al [1984]).

In conditionally heteroskedastic models the QMLE is not rendered inconsistent by the misspecification of any moment of order higher than the second, as shown in
Lemma 4.8. - Under (H0) and (C1) - (C7) the QMLE $\hat{\theta}$ is a consistent estimator of $\theta_0$ if the conditional mean and conditional variance are correctly specified, regardless of the form of the true conditional distribution.

Proof: When the conditional mean and the conditional variance are correctly specified $E[u_t | \mathcal{F}_t] = E[v_t | \mathcal{F}_t] = 0$. Therefore $\varphi(\theta_0) = 0$ from (2) - (4), and the result follows from Lemma 4.2.

However non-normalities may be causing a loss in efficiency in relation to estimators that take into account the proper structure of the information contained in higher order conditional moments. There is also the risk of drawing incorrect inferences if the estimated covariance matrix of $\hat{\theta}$ does not take into account the correct form of the third and fourth moments.

To analyze the implications of asymmetry consider the two-equation system (see (2.7)),

$$\eta_t = g_t(\theta) + v_t.$$  

If the distribution is symmetric, the conditional covariance matrix $\Sigma_t$ of $v_t$ is diagonal. Therefore estimating the two equations separately (by $\hat{\beta}_m$ and $\hat{\phi}_v$) and imposing the cross-equation restrictions by means of the MWA structure is asymptotically efficient with respect to the information provided by the first two conditional moments. If $\Sigma_t$ is not diagonal this procedure ignores the information in the covariances between the equations, resulting in inefficiency. An alternative GMM argument is that using the weighting matrix in (3.18) with the off-diagonal blocks improperly set to zero does not produce a GMM estimator that attains its variance bound (e.g. Hansen [1982], Chamberlain [1987]). The correct asymptotic distribution of the QMLE $\hat{\theta}$ is then

$$T^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N \left[ 0, \Sigma \left( \frac{1}{T} G \Sigma_0^{-1} G' \Sigma G (G' \Sigma_0^{-1} G')^{-1} \right) \right],$$

where $\Sigma_0$ is the covariance matrix under symmetry, and
where we retain mesokurtosis but this is easily dispensed with using $K$ instead of $2 \Omega^2$. It is evident that using the covariance matrix from MLE estimation may lead to erroneous inferences. However, $V(\theta)$ may be estimated using White's [1980b] procedure in the two-equation system. The marginal distributions of $\hat{\beta}_m$ and $\hat{\phi}_v$ are not affected, but these estimators are now asymptotically dependent and hence $\gamma$ is no longer a MWA of $\gamma_m$ and $\gamma_v$. The factorization of the likelihood function does not correspond to the true likelihood.

The consistency-robustness properties for the ARCH class of models cannot be established without symmetry, and thus assessing the symmetry assumption in this class of models becomes an evaluation of the potential for robustness. Also observe that the asymptotic independence in these models between $\hat{\beta}$ and $\hat{\alpha}$ is lost without symmetry (Engle [1982a]).

Another interesting property of the ARCH class of models is that, under conditional normality, it produces a leptokurtic unconditional distribution. This is in agreement with the empirical characteristics of many economic series as exchange rates, rates of return and stock prices data. Bollerslev [1985] and Engle and Bollerslev [1986] have proposed to fatten the tails further by allowing a leptokurtic conditional distribution such as the $t$ distribution. In empirical work with exchange rates and rates of return they rejected the hypothesis of conditional mesokurtosis (normality) in favour of conditional leptokurtosis ($t$-distribution).

When we use the likelihood function derived from conditional normality we impose $\kappa = 2 \mathcal{h}_t^2$, and if this is not the case the variance equation in (3.9) is weighted with an incorrect conditional variance. The QMLE $\hat{\phi}_v$ obtained under conditional normality remains consistent (Lemma 4.5), though it is inefficient and its correct distribution is given by
leading to

$$\mathbb{T}^{1/2} (\hat{\phi}_v - \phi_0) \xrightarrow{d} \mathcal{N} [0, \mathcal{C} \{ T (S_0' \Omega^2 S_0)^{-1} S_0' \Omega^2 K \Omega^{-2} S_0 (S_0' \Omega^{-2} S_0)^{-1} \} ] ,$$

where

$$\Sigma_k = \Sigma_0^{-1} \begin{pmatrix} \Omega & 0 \\ 0 & K \end{pmatrix} \Sigma_0^{-1} = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \frac{1}{4} \Omega^{-2}K \end{pmatrix} .$$

Thus both $\hat{\beta}$ and $\hat{\alpha}$ become inefficient and the ML covariance matrix may lead to incorrect inferences. The inefficiency can be decomposed in two effects. There is the inefficiency associated to the use of an incorrect covariance matrix, and if the conditional distribution is not normal, we are not using the potential additional information in the fourth moment (and possibly in higher moments). In section § 3.2, for example, we saw that it is the fourth moment that provides information to estimate the degrees of freedom parameter in the $t$ distribution, and it also adds information for the estimation of $\beta$ and $\alpha$.

The asymptotic distribution of $\hat{\beta}_m$ is determined by the first two moments, and that of $\hat{\phi}_v$ by the first four moments. Thus if moments of higher order than these are misspecified, this does not affect the asymptotic distribution of the QMLE $\hat{\theta}$ given in (3.28). However, there may be additional information to be extracted from these higher moments to improve efficiency in the estimation of $\theta$. This issue will be considered in Chapter 10.

§ 4.5 Some comments on Monte Carlo evidence

We have conducted simulation experiments with misspecified models. The results are presented in Tables 4.1 - 4.11 for the same estimators that were considered in the experiments under correct specification in section § 3.5. Tables 4.1 - 4.11 report the sample mean and standard error of the bias for the estimators in the same format as Tables 3.1 - 3.4, though in this chapter we do
not report information measures and the coefficients of skewness and kurtosis.

§ 4.5.1 Specification error in the conditional mean

Tables 4.1 - 4.4 refer to models where the conditional mean is misspecified by the exclusion of the regressor \( x_{2t} \), so that \( E [ u_t \mid \mathcal{F}_t ] = x_{2t} \), while the conditional variance is correctly specified in nature (e.g. it is Poisson-type for a Poisson-type DGP, or ARCH for an ARCH DGP). The excluded regressor is correlated with the included one and hence the estimators are rendered inconsistent.

The results for the Poisson-N model are reported in Table 4.1. Because of the Poisson-type nature of both the model and the DGP the nature of the conditional variance is correctly specified, but the specification error in the conditional mean implies the same type of specification error in the conditional variance (both \( \mu_t \) and \( h_t \) are excluding \( x_{2t} \)), so that using (6) the disturbances of the variance equation have expectation

\[
E [ \varepsilon_t \mid \mathcal{F}_t ] = ( \overline{h_t} - h_t ) + ( \overline{\mu_t} - \mu_t )^2 = x_{2t} + x_{2t}^2,
\]

and hence \( h_t \) is omitting two regressors. With a few exceptions, the biases are substantial both in magnitude and significance from the smallest sample and contrast sharply with the results under correct specification given in Table 3.1. The results show that the mean biases in ML are not substantially different from the mean biases in the estimators using only mean information (OLS and GLS). This is clearer in the estimates of \( \beta_0 \), where the bias of VAR is similar to that of GLS, especially as the sample size grows. For \( \beta_1 \), the bias of VAR is not as close to that of GLS, causing the MLE to move away from the GLS estimator, but still the difference is not large.
Tables 4.2 - 4.4 report the results for the mild, strong and regular ARCH models, respectively. Using (6), the variance equation disturbances have expectation

$$E \left[ \varepsilon_t \mid \mathcal{F}_t \right] = -2 \alpha_1 u_{t-1} x_{2t-1} + \alpha_1 x_{2t-1}^2 + x_{2t}^2,$$

and because the $x_{2t}$ are not autocorrelated, they are also uncorrelated with $u_{t-1}$ and we would expect the last term not to have an effect on the estimation of the variance equation. The biases in the estimation of the $\beta$ coefficients are large and significant from the smallest sample for all the estimators using mean information (OLS, GLS, MWA and ML) as one would expect. The biases of VAR, although in many cases large in absolute value, are insignificant in the mild model for all sample sizes, while the regular and strong models require the largest sample to show significant biases. This causes the biases of mean and mixed estimators to be very similar, especially those of GLS and ML.

The effect of the misspecification on estimators using variance information is better captured by the estimators of the $\alpha$ parameters, whose biases in most cases are statistically significant. The biases in estimating $\alpha_0$ are not much larger in general than those in the correctly specified model for the smallest sample size, but they do not show the same tendency to vanish as sample size increases. The biases for the estimates of $\alpha_1$ show a clear dependence on the strength of the ARCH effect, being small when the ARCH effect is mild, and increasing in the regular and strong cases. However, it does not seem that the misspecified conditional mean will have a dramatic effect on the biases in the estimators of $\alpha_1$ which are still decreasing clearly with sample size at $T = 200$.

We have conducted another experiment with the ARCH model with a misspecified conditional mean, by introducing first order autocorrelation in the DGP, so that $E \left[ u_t \mid \mathcal{F}_t \right] = \rho u_{t-1}$. Because the conditional mean includes only strongly exogenous variables, this has no effect on the consistency of the
estimators of $\beta$ using only mean information (Lemma 4.4), but it affects estimation in the variance equation as we have from (6) that

$$E [ \varepsilon_t | \mathcal{F}_t ] = -2 \alpha_1 \rho u_{t-1} u_{t-2} + \alpha_1 \rho^2 u_{t-2}^2 + \rho^2 u_{t-1}^2 ,$$

which illustrates how autocorrelation induces an ARCH component in the conditional variance (i.e. in a homoskedastic model $E [ u_t^2 | \mathcal{F}_t ] = \alpha_0 + \rho^2 u_{t-1}^2$, see Engle et al. [1985]). The last two terms of this expression are conditionally even functions of lagged errors and hence will not affect the consistency of variance estimators of $\beta$ (see the discussion after Lemma 4.4).

Results for the mild and strong models are reported in Tables 4.5 and 4.6, for $\rho = 0.8$. All estimators of $\beta$ using mean information have insignificant biases as expected. The variance estimator of $\beta_0$ does not appear to be strongly affected by the term $-2 \alpha_1 \rho u_{t-1} u_{t-2}$, and biases are insignificant in the mild model and become significant only at the largest sample size of the strong model, where they behave erratically with sample size. ML and MWA present insignificant biases as a consequence. The biases of the estimators of $\alpha_0$ are large and significant in both models, and though they are decreasing with sample size, this effect is very slow. The biases in $\alpha_1$ are smaller but significant in the mild model and do not show a tendency to vanish. In the strong model these biases are much larger and significant, and are still decreasing with sample size at $T = 200$, and this effect is very clear for the MLE which shows evident signs of consistency. An intuitive, though not fully satisfactory, explanation for these effects in the estimation of the $\alpha$ parameters is that the first term in $E [ \varepsilon_t | \mathcal{F}_t ]$ has opposite sign to the remaining two and thus there may be a cancelling effects in the expectation when multiplied by the even component $h_t^2 z_t$. 
§ 4.5.2 Specification error in the conditional variance

Tables 4.7 - 4.11 report experiments in which the conditional mean has been kept correctly specified so that $\mathbb{E}[u_t | \mathcal{F}_t] = 0$, but the conditional variance is misspecified and hence the $\varepsilon_t$ are not proper innovations. It follows from Lemma 4.5 that estimators using only mean information should remain consistent, but estimators using variance information will be affected in general.

For the Poisson-N model, the DGP has conditional variance

$$\bar{h}_t = 0.5 + 0.7 x_{1t},$$

and the results are reported in Table 4.7. The biases for the OLS and GLS estimators are small, mostly insignificant, and decrease evidently with sample size. In contrast, the variance and mixed information estimators show large and significant biases (all t-ratios are over 100 for $T = 200$).

For the ARCH model we have conducted two experiments: one that should induce inconsistency in the MLE and one that should not do so. The first experiment is reported in Tables 4.8 and 4.9 for the mild and strong ARCH effects, respectively, and in this the DGP has conditional variance

$$\bar{h}_t = \alpha_0 + \alpha_1 y_{t-1}^2 = h_t + \alpha_1 \mu_{t-1}^2 - w_t \beta,$$

where the last equality follows from (8), and the DGP has $\beta_0 = \beta_1 = 1$. The combination of even and odd effects in the true conditional variance must induce inconsistency in the MLE of $\beta$. The estimators using only mean information have indeed small and insignificant biases which shrink appropriately with sample size. The estimator using only variance information presents very large biases but it is very inefficient, especially in the mild model, and the biases in $\beta_1$ are in many cases insignificant. This estimator should converge to zero as the sample size increases, but this is not
all that clear in the simulation results though in many cases the bias is not significantly different from zero. When the information is mixed to compute the MWA and ML estimators, these show large and significant biases as expected, which of course are larger for the strong process. Thus the experiment shows nicely how the MLE of $\beta$ is not robust to misspecification of $h_t$ of this sort. The $\alpha$ estimators also show the effects of the modelling error, and for $\alpha_0$ the biases are clear in size and significance and do not shrink as the sample grows. The effect on $\alpha_1$ is much smaller in the mild model, but it is evident in the strong model.

The second experiment considers a departure in the conditional variance to which the MLE of $\beta$ is robust, namely misspecification of the order of the ARCH process. The DGP has conditional variance

$$h_t = 0.5 + 0.3 u_{t-1}^2 + 0.2 u_{t-2}^2,$$

and the results are reported in Table 4.10. The biases of the $\beta$ estimators are in most cases small and insignificant and show a clear tendency to vanish with sample size, in agreement with Lemma 4.6. The weight of the specification error falls, of course, on the estimators of the $\alpha$ parameters which have significant biases. An interesting fact is that the omitted regressor in the variance equation affects mainly $\alpha_0$ and not $\alpha_1$, thus showing essentially the effect on the 'unconditional' component of the variance (the constant), rather than in the 'conditional' component ($u_{t-1}^2$). This experiment was performed with the DGP having a zero conditional mean, and was then repeated with $\mu_t = 1 + x_{1t}$ to make an assessment of the effect of different parameter values on the small sample distribution of the estimators. The results of this second experiment are reported in Table 4.11, and are essentially identical to those of Table 4.10, allowing for a few natural round-up errors. This suggests that Breusch's [1980] invariance results might carry through to the ARCH model under correct specification, and maybe even under the sort of misspecifications
considered in Lemmas 4.6 and 4.7. This point has been used without proof by Diebold [1985] in reporting some simulation experiments, and it is an interesting topic for further research.

**TABLE 4.1. ESTIMATION (Poisson-N, misspecified mean).**

<table>
<thead>
<tr>
<th>Model: $y_t \sim N[\beta_0 + \beta_1 x_{1t}, \beta_0 + \beta_1 x_{1t}]$</th>
<th>DGP: $y_t \sim N[1 + x_{1t} + x_{2t}, 1 + x_{1t} + x_{2t}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T = 20</strong></td>
<td><strong>T = 50</strong></td>
</tr>
<tr>
<td><strong>$\beta_0$</strong></td>
<td><strong>bias</strong></td>
</tr>
<tr>
<td>OLS</td>
<td>0.744</td>
</tr>
<tr>
<td>GLS</td>
<td>0.703</td>
</tr>
<tr>
<td>VAR</td>
<td>0.539</td>
</tr>
<tr>
<td>MWA</td>
<td>0.630</td>
</tr>
<tr>
<td>ML</td>
<td>0.742</td>
</tr>
<tr>
<td><strong>$\beta_1$</strong></td>
<td><strong>bias</strong></td>
</tr>
<tr>
<td>OLS</td>
<td>0.458</td>
</tr>
<tr>
<td>GLS</td>
<td>0.389</td>
</tr>
<tr>
<td>VAR</td>
<td>0.180</td>
</tr>
<tr>
<td>MWA</td>
<td>0.253</td>
</tr>
<tr>
<td>ML</td>
<td>0.459</td>
</tr>
</tbody>
</table>
TABLE 4.2 - ESTIMATION (ARCH I, misspecified mean).

<table>
<thead>
<tr>
<th></th>
<th>Model: $y_t \sim \mathcal{N} (\beta_0 + \beta_1 x_{t1}, \alpha_0 + \alpha_1 u_{t1}^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DGP: $y_t \sim \mathcal{N} (x_{t2}, 0.8 + 0.2 u_{t2}^2)$</td>
</tr>
<tr>
<td></td>
<td>$T = 20$</td>
</tr>
<tr>
<td></td>
<td>bias</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.749</td>
</tr>
<tr>
<td>GLS</td>
<td>0.730</td>
</tr>
<tr>
<td>VAR</td>
<td>1.271</td>
</tr>
<tr>
<td>MWA</td>
<td>0.730</td>
</tr>
<tr>
<td>ML</td>
<td>0.730</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.466</td>
</tr>
<tr>
<td>GLS</td>
<td>0.403</td>
</tr>
<tr>
<td>VAR</td>
<td>2.368</td>
</tr>
<tr>
<td>MWA</td>
<td>0.414</td>
</tr>
<tr>
<td>ML</td>
<td>0.409</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>0.097</td>
</tr>
<tr>
<td>GLS</td>
<td>0.076</td>
</tr>
<tr>
<td>VAR</td>
<td>0.070</td>
</tr>
<tr>
<td>MWA</td>
<td>0.013</td>
</tr>
<tr>
<td>ML</td>
<td></td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>-0.163</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.151</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.073</td>
</tr>
<tr>
<td>MWA</td>
<td>0.015</td>
</tr>
</tbody>
</table>

TABLE 4.3 - ESTIMATION (ARCH II, misspecified mean).

<table>
<thead>
<tr>
<th></th>
<th>Model: $y_t \sim \mathcal{N} (\beta_0 + \beta_1 x_{t1}, \alpha_0 + \alpha_1 u_{t1}^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DGP: $y_t \sim \mathcal{N} (x_{t2}, 0.2 + 0.8 u_{t2}^2)$</td>
</tr>
<tr>
<td></td>
<td>$T = 20$</td>
</tr>
<tr>
<td></td>
<td>bias</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.757</td>
</tr>
<tr>
<td>GLS</td>
<td>0.732</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.035</td>
</tr>
<tr>
<td>MWA</td>
<td>0.734</td>
</tr>
<tr>
<td>ML</td>
<td>0.743</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.474</td>
</tr>
<tr>
<td>GLS</td>
<td>0.401</td>
</tr>
<tr>
<td>VAR</td>
<td>0.359</td>
</tr>
<tr>
<td>MWA</td>
<td>0.421</td>
</tr>
<tr>
<td>ML</td>
<td>0.414</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>0.308</td>
</tr>
<tr>
<td>GLS</td>
<td>0.223</td>
</tr>
<tr>
<td>VAR</td>
<td>0.220</td>
</tr>
<tr>
<td>ML</td>
<td>0.124</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>-0.638</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.560</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.549</td>
</tr>
<tr>
<td>ML</td>
<td>-0.319</td>
</tr>
</tbody>
</table>
### TABLE 4.4 . - ESTIMATION (ARCH III, misspecified mean).

Model: \( y_t \sim \mathcal{N}(\beta_0 + \beta_1 x_{it}, \sigma_0 + \sigma_1 u_{it}^2) \)

\[ DGP: y_t \sim \mathcal{N}(x_{it}, 0.5 + 0.5 u_{it}^2) \]

<table>
<thead>
<tr>
<th>Model: ( y_t \sim \mathcal{N}(\beta_0 + \beta_1 x_{it}, \sigma_0 + \sigma_1 u_{it}^2) )</th>
<th>( T = 20 )</th>
<th>( T = 50 )</th>
<th>( T = 100 )</th>
<th>( T = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
</tr>
<tr>
<td>OLS</td>
<td>0.754</td>
<td>0.021</td>
<td>0.681</td>
<td>0.014</td>
</tr>
<tr>
<td>GLS</td>
<td>0.728</td>
<td>0.021</td>
<td>0.662</td>
<td>0.013</td>
</tr>
<tr>
<td>VAR</td>
<td>0.346</td>
<td>0.784</td>
<td>-0.127</td>
<td>0.815</td>
</tr>
<tr>
<td>MWA</td>
<td>0.734</td>
<td>0.021</td>
<td>0.674</td>
<td>0.014</td>
</tr>
<tr>
<td>ML</td>
<td>0.740</td>
<td>0.019</td>
<td>0.662</td>
<td>0.012</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
</tr>
<tr>
<td>OLS</td>
<td>0.471</td>
<td>0.033</td>
<td>0.338</td>
<td>0.023</td>
</tr>
<tr>
<td>GLS</td>
<td>0.396</td>
<td>0.034</td>
<td>0.298</td>
<td>0.021</td>
</tr>
<tr>
<td>VAR</td>
<td>0.748</td>
<td>1.358</td>
<td>-0.121</td>
<td>0.921</td>
</tr>
<tr>
<td>MWA</td>
<td>0.415</td>
<td>0.034</td>
<td>0.322</td>
<td>0.022</td>
</tr>
<tr>
<td>ML</td>
<td>0.415</td>
<td>0.032</td>
<td>0.301</td>
<td>0.020</td>
</tr>
</tbody>
</table>

### TABLE 4.5 . - ESTIMATION (ARCH I, autocorrelation).

Model: \( y_t \sim \mathcal{N}(\beta_0 + \beta_1 x_{it}, \sigma_0 + \sigma_1 u_{it}^2) \)

\[ DGP: y_t \sim \mathcal{N}(1 + x_{it} + 0.8 u_{it}, 0.8 + 0.2 u_{it}^2) \]

<table>
<thead>
<tr>
<th>Model: ( y_t \sim \mathcal{N}(\beta_0 + \beta_1 x_{it}, \sigma_0 + \sigma_1 u_{it}^2) )</th>
<th>( T = 20 )</th>
<th>( T = 50 )</th>
<th>( T = 100 )</th>
<th>( T = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.017</td>
<td>0.025</td>
<td>-0.006</td>
<td>0.016</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.017</td>
<td>0.025</td>
<td>-0.005</td>
<td>0.016</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.062</td>
<td>1.014</td>
<td>-1.534</td>
<td>1.868</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.023</td>
<td>0.025</td>
<td>0.004</td>
<td>0.017</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.048</td>
<td>0.032</td>
<td>-0.032</td>
<td>0.021</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.043</td>
<td>0.035</td>
<td>-0.026</td>
<td>0.022</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.708</td>
<td>2.107</td>
<td>-0.754</td>
<td>2.358</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.047</td>
<td>0.036</td>
<td>-0.008</td>
<td>0.023</td>
</tr>
<tr>
<td>ML</td>
<td>-0.034</td>
<td>0.033</td>
<td>-0.022</td>
<td>0.021</td>
</tr>
<tr>
<td>( \sigma_0 )</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
</tr>
<tr>
<td>SLS</td>
<td>0.538</td>
<td>0.029</td>
<td>0.490</td>
<td>0.019</td>
</tr>
<tr>
<td>GLS</td>
<td>0.491</td>
<td>0.030</td>
<td>0.434</td>
<td>0.019</td>
</tr>
<tr>
<td>VAR</td>
<td>0.504</td>
<td>0.034</td>
<td>0.430</td>
<td>0.021</td>
</tr>
<tr>
<td>ML</td>
<td>0.345</td>
<td>0.030</td>
<td>0.337</td>
<td>0.019</td>
</tr>
<tr>
<td>( \sigma_1 )</td>
<td>bias</td>
<td>std.err.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>-0.147</td>
<td>0.010</td>
<td>-0.022</td>
<td>0.008</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.120</td>
<td>0.013</td>
<td>0.013</td>
<td>0.010</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.063</td>
<td>0.009</td>
<td>0.019</td>
<td>0.009</td>
</tr>
<tr>
<td>ML</td>
<td>0.084</td>
<td>0.021</td>
<td>0.113</td>
<td>0.013</td>
</tr>
</tbody>
</table>
### TABLE 4.6 . - ESTIMATION (ARCH II, autocorrelation).

Model: $y_t | \mathcal{F}_t \sim N[\beta_0 + \beta_1 x_{it}, \sigma_0 + \sigma_1 u_{it}^2]$  
DGP: $y_t | \mathcal{F}_t \sim N[1 + x_{it} + 0.8 u_{t-1}, 0.2 + 0.8 u_{t-1}^2]$  

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th></th>
<th>$T = 50$</th>
<th></th>
<th>$T = 100$</th>
<th></th>
<th>$T = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.000</td>
<td>0.027</td>
<td>-0.001</td>
<td>0.017</td>
<td>-0.005</td>
<td>0.014</td>
<td>0.004</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.011</td>
<td>0.026</td>
<td>0.001</td>
<td>0.015</td>
<td>0.002</td>
<td>0.011</td>
<td>0.008</td>
</tr>
<tr>
<td>VAR</td>
<td>0.040</td>
<td>0.934</td>
<td>-0.785</td>
<td>1.778</td>
<td>0.444</td>
<td>1.595</td>
<td>-2.124</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.012</td>
<td>0.027</td>
<td>-0.003</td>
<td>0.019</td>
<td>-0.020</td>
<td>0.017</td>
<td>0.001</td>
</tr>
<tr>
<td>ML</td>
<td>0.008</td>
<td>0.025</td>
<td>-0.009</td>
<td>0.014</td>
<td>-0.012</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>-0.023</td>
<td>0.033</td>
<td>-0.033</td>
<td>0.024</td>
<td>-0.026</td>
<td>0.020</td>
<td>-0.006</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.023</td>
<td>0.039</td>
<td>-0.018</td>
<td>0.021</td>
<td>-0.003</td>
<td>0.016</td>
<td>0.004</td>
</tr>
<tr>
<td>VAR</td>
<td>0.312</td>
<td>1.666</td>
<td>-0.044</td>
<td>3.004</td>
<td>0.232</td>
<td>1.087</td>
<td>-4.254</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.010</td>
<td>0.041</td>
<td>-0.010</td>
<td>0.032</td>
<td>-0.021</td>
<td>0.028</td>
<td>0.026</td>
</tr>
<tr>
<td>ML</td>
<td>0.011</td>
<td>0.038</td>
<td>-0.013</td>
<td>0.019</td>
<td>-0.015</td>
<td>0.013</td>
<td>0.003</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.867</td>
<td>0.217</td>
<td>0.825</td>
<td>0.099</td>
<td>0.787</td>
<td>0.063</td>
<td>0.773</td>
</tr>
<tr>
<td>GLS</td>
<td>0.785</td>
<td>0.230</td>
<td>0.596</td>
<td>0.098</td>
<td>0.503</td>
<td>0.052</td>
<td>0.427</td>
</tr>
<tr>
<td>VAR</td>
<td>0.916</td>
<td>0.333</td>
<td>0.803</td>
<td>0.259</td>
<td>0.597</td>
<td>0.106</td>
<td>0.492</td>
</tr>
<tr>
<td>ML</td>
<td>0.570</td>
<td>0.216</td>
<td>0.392</td>
<td>0.092</td>
<td>0.288</td>
<td>0.015</td>
<td>0.280</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>-0.690</td>
<td>0.012</td>
<td>-0.540</td>
<td>0.010</td>
<td>-0.490</td>
<td>0.008</td>
<td>-0.454</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.639</td>
<td>0.017</td>
<td>-0.382</td>
<td>0.018</td>
<td>-0.289</td>
<td>0.015</td>
<td>-0.213</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.610</td>
<td>0.012</td>
<td>-0.439</td>
<td>0.014</td>
<td>-0.340</td>
<td>0.013</td>
<td>-0.251</td>
</tr>
<tr>
<td>ML</td>
<td>-0.258</td>
<td>0.035</td>
<td>-0.076</td>
<td>0.032</td>
<td>-0.031</td>
<td>0.030</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

### TABLE 4.7 . - ESTIMATION (Poisson-N, misspecified variance).

Model: $y_t | \mathcal{F}_t \sim N[\beta_0 + \beta_1 x_{it}, \beta_0 + \beta_1 x_{it}]$  
DGP: $y_t | \mathcal{F}_t \sim N[1 + x_{it} + 0.5 + 0.7 x_{it}]$  

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th></th>
<th>$T = 50$</th>
<th></th>
<th>$T = 100$</th>
<th></th>
<th>$T = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
<td>std.err.</td>
<td>bias</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>-0.010</td>
<td>0.011</td>
<td>-0.005</td>
<td>0.007</td>
<td>-0.006</td>
<td>0.005</td>
<td>-0.004</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.014</td>
<td>0.010</td>
<td>-0.006</td>
<td>0.006</td>
<td>-0.007</td>
<td>0.005</td>
<td>-0.003</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.622</td>
<td>0.008</td>
<td>-0.658</td>
<td>0.005</td>
<td>-0.673</td>
<td>0.003</td>
<td>-0.683</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.308</td>
<td>0.007</td>
<td>-0.316</td>
<td>0.004</td>
<td>-0.324</td>
<td>0.003</td>
<td>-0.332</td>
</tr>
<tr>
<td>ML</td>
<td>-0.323</td>
<td>0.008</td>
<td>-0.358</td>
<td>0.005</td>
<td>-0.374</td>
<td>0.003</td>
<td>-0.375</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>-0.015</td>
<td>0.014</td>
<td>-0.010</td>
<td>0.009</td>
<td>-0.010</td>
<td>0.007</td>
<td>-0.007</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.022</td>
<td>0.011</td>
<td>-0.011</td>
<td>0.007</td>
<td>-0.012</td>
<td>0.005</td>
<td>-0.006</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.593</td>
<td>0.009</td>
<td>-0.639</td>
<td>0.005</td>
<td>-0.657</td>
<td>0.004</td>
<td>-0.673</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.284</td>
<td>0.008</td>
<td>-0.294</td>
<td>0.005</td>
<td>-0.302</td>
<td>0.004</td>
<td>-0.318</td>
</tr>
<tr>
<td>ML</td>
<td>-0.287</td>
<td>0.009</td>
<td>-0.332</td>
<td>0.005</td>
<td>-0.350</td>
<td>0.004</td>
<td>-0.359</td>
</tr>
</tbody>
</table>
TABLE 4.8 - ESTIMATION (ARCH I, misspecified variance).

<table>
<thead>
<tr>
<th>Model: $y_t \sim \mathcal{N}[\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2]$</th>
<th>DGP: $y_t \sim \mathcal{N}[1 + x_t, 0.8 + 0.2 y_{t-1}^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 20$</td>
<td>$T = 50$</td>
</tr>
<tr>
<td><strong>$\beta_0$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.012</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.013</td>
</tr>
<tr>
<td>VAR</td>
<td>-2.396</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.049</td>
</tr>
<tr>
<td>ML</td>
<td>-0.042</td>
</tr>
<tr>
<td><strong>$\beta_1$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.035</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.033</td>
</tr>
<tr>
<td>VAR</td>
<td>-3.625</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.070</td>
</tr>
<tr>
<td>ML</td>
<td>-0.067</td>
</tr>
</tbody>
</table>

TABLE 4.9 - ESTIMATION (ARCH II, misspecified variance).

<table>
<thead>
<tr>
<th>Model: $y_t \sim \mathcal{N}[\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2]$</th>
<th>DGP: $y_t \sim \mathcal{N}[1 + x_t, 0.2 + 0.8 y_{t-1}^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 20$</td>
<td>$T = 50$</td>
</tr>
<tr>
<td><strong>$\alpha_0$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>0.112</td>
</tr>
<tr>
<td>GLS</td>
<td>0.095</td>
</tr>
<tr>
<td>VAR</td>
<td>0.067</td>
</tr>
<tr>
<td>MWA</td>
<td>0.015</td>
</tr>
<tr>
<td>ML</td>
<td>-0.155</td>
</tr>
<tr>
<td><strong>$\alpha_1$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.141</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.074</td>
</tr>
<tr>
<td>VAR</td>
<td>0.042</td>
</tr>
</tbody>
</table>

TABLE 4.8 - ESTIMATION (ARCH I, misspecified variance).

<table>
<thead>
<tr>
<th>Model: $y_t \sim \mathcal{N}[\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2]$</th>
<th>DGP: $y_t \sim \mathcal{N}[1 + x_t, 0.8 + 0.2 y_{t-1}^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 20$</td>
<td>$T = 50$</td>
</tr>
<tr>
<td><strong>$\beta_0$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.008</td>
</tr>
<tr>
<td>GLS</td>
<td>0.008</td>
</tr>
<tr>
<td>VAR</td>
<td>-2.673</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.185</td>
</tr>
<tr>
<td>ML</td>
<td>-0.137</td>
</tr>
<tr>
<td><strong>$\beta_1$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.032</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.014</td>
</tr>
<tr>
<td>VAR</td>
<td>1.223</td>
</tr>
<tr>
<td>MWA</td>
<td>-0.190</td>
</tr>
<tr>
<td>ML</td>
<td>-0.156</td>
</tr>
<tr>
<td><strong>$\alpha_0$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>0.716</td>
</tr>
<tr>
<td>GLS</td>
<td>0.511</td>
</tr>
<tr>
<td>VAR</td>
<td>0.518</td>
</tr>
<tr>
<td>ML</td>
<td>0.254</td>
</tr>
<tr>
<td><strong>$\alpha_1$</strong></td>
<td>bias</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.552</td>
</tr>
<tr>
<td>GLS</td>
<td>-0.444</td>
</tr>
<tr>
<td>VAR</td>
<td>-0.500</td>
</tr>
<tr>
<td>ML</td>
<td>-0.058</td>
</tr>
</tbody>
</table>
### TABLE 4.10 - ESTIMATION (ARCH IV, misspecified variance).

Model: \( y_t | \mathcal{F}_{t-1} \sim N(\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2) \)

\[
\begin{array}{ccccccccc}
T = 20 & T = 50 & T = 100 & T = 200 \\
\hline
\beta_0 & & & & \\
OLS & -0.018 & 0.018 & -0.011 & 0.013 & -0.016 & 0.010 & -0.006 & 0.007 \\
GLS & -0.029 & 0.019 & -0.009 & 0.012 & -0.015 & 0.009 & -0.005 & 0.006 \\
VAR & -0.771 & 0.916 & 1.890 & 0.993 & 0.294 & 0.350 & 0.133 & 0.179 \\
MWA & -0.030 & 0.018 & 0.003 & 0.012 & -0.019 & 0.009 & -0.008 & 0.006 \\
ML & -0.030 & 0.018 & -0.004 & 0.012 & -0.010 & 0.009 & -0.004 & 0.006 \\
\alpha_0 & & & & \\
SLS & 0.256 & 0.022 & 0.251 & 0.014 & 0.234 & 0.010 & 0.220 & 0.008 \\
GLS & 0.225 & 0.022 & 0.203 & 0.013 & 0.181 & 0.009 & 0.166 & 0.006 \\
VAR & 0.209 & 0.022 & 0.207 & 0.014 & 0.187 & 0.009 & 0.170 & 0.006 \\
MWA & 0.138 & 0.018 & 0.144 & 0.012 & 0.144 & 0.009 & 0.145 & 0.006 \\
ML & 0.002 & 0.019 & 0.036 & 0.014 & 0.041 & 0.010 & 0.052 & 0.007 \\
\alpha_1 & & & & \\
SLS & -0.218 & 0.013 & -0.109 & 0.010 & -0.066 & 0.007 & -0.031 & 0.006 \\
GLS & -0.190 & 0.016 & -0.070 & 0.012 & -0.016 & 0.009 & 0.020 & 0.007 \\
VAR & -0.138 & 0.010 & -0.051 & 0.010 & -0.020 & 0.008 & 0.016 & 0.007 \\
MWA & 0.002 & 0.019 & 0.036 & 0.014 & 0.041 & 0.010 & 0.052 & 0.007 \\
\end{array}
\]

### TABLE 4.11 - ESTIMATION (ARCH IV, misspecified variance).

Model: \( y_t | \mathcal{F}_{t-1} \sim N(\beta_0 + \beta_1 x_t, \alpha_0 + \alpha_1 u_{t-1}^2) \)

\[
\begin{array}{ccccccccc}
T = 20 & T = 50 & T = 100 & T = 200 \\
\hline
\beta_0 & & & & \\
OLS & -0.018 & 0.018 & -0.011 & 0.013 & -0.016 & 0.010 & -0.006 & 0.007 \\
GLS & -0.029 & 0.019 & -0.009 & 0.012 & -0.015 & 0.009 & -0.005 & 0.006 \\
VAR & -0.771 & 0.916 & 1.890 & 0.993 & 0.294 & 0.350 & 0.133 & 0.179 \\
MWA & -0.030 & 0.018 & 0.003 & 0.012 & -0.019 & 0.009 & -0.008 & 0.006 \\
ML & -0.029 & 0.018 & -0.004 & 0.012 & -0.010 & 0.009 & -0.004 & 0.006 \\
\alpha_0 & & & & \\
SLS & 0.256 & 0.022 & 0.251 & 0.014 & 0.234 & 0.010 & 0.220 & 0.008 \\
GLS & 0.225 & 0.022 & 0.203 & 0.013 & 0.181 & 0.009 & 0.166 & 0.006 \\
VAR & 0.209 & 0.022 & 0.207 & 0.014 & 0.187 & 0.009 & 0.170 & 0.006 \\
MWA & 0.140 & 0.018 & 0.144 & 0.012 & 0.144 & 0.009 & 0.145 & 0.006 \\
ML & 0.002 & 0.019 & 0.036 & 0.014 & 0.041 & 0.010 & 0.052 & 0.007 \\
\alpha_1 & & & & \\
SLS & -0.218 & 0.013 & -0.109 & 0.010 & -0.066 & 0.007 & -0.031 & 0.006 \\
GLS & -0.190 & 0.016 & -0.070 & 0.012 & -0.016 & 0.009 & 0.020 & 0.007 \\
VAR & -0.138 & 0.010 & -0.051 & 0.010 & -0.020 & 0.008 & 0.016 & 0.007 \\
MWA & 0.002 & 0.019 & 0.036 & 0.014 & 0.041 & 0.010 & 0.052 & 0.007 \\
\end{array}
\]
CHAPTER 5

GENERAL SPECIFICATION TESTS

The situation presented in Chapter 4 shows the implications of specification error and calls for a careful evaluation of the model. We have seen misspecification directions in which the quality of the estimators is fragile. In particular, applied workers using parametric specification of \( h_t \) need to be concerned that any failure to specify \( h_t \) correctly may contaminate the estimates of \( \beta \), which is normally their primary interest. One possibility is to rely on robust estimation procedures, particularly GLS, which remain consistent in the presence of misspecified heteroskedasticity. There are now semi-parametric GLS estimators that are asymptotically as efficient as the GLS yet presume no knowledge of the heteroskedasticity (see § 3.1.2). Consequently, if researchers are going to continue to use MLE's in heteroskedastic models, it is important that some information be provided along with the estimates to assess the validity of the chosen specification.

There are, of course, many diagnostic tests for incorrect specifications in the conditional mean and variances of the general linear model – Pagan [1984a] surveys these – and all might be applied. Engle [1983], Engle et al [1987], Kraft and Engle [1982], and Weiss [1984,1986a] inter alia have proposed diagnostics for specific departures in the context of heteroskedastic models. But it is always useful to have some tests that utilize information that is available solely within the estimated model since this can be provided with computer output, and this Chapter analyzes some possibilities.

(1) In this Chapter Section § 5.2 is based on joint work with A.R. Pagan, reported in Pagan and Sabau [1987b].
In sections § 5.1 and § 5.2 we concentrate on departures that render the QMLE \( \hat{\theta} \) inconsistent. The MWA interpretation of Chapter 3 produces naturally one such test by checking the coherency of the information about common parameters in the mean and variance equations. This test is analyzed in section § 5.1. In section § 5.2 we argue for what we call a set of "consistency tests". Simple examples are that the MLE residuals sum to zero and that the postulated heteroskedastic pattern conforms to the evidence in the squared residuals. These are functions of the output of any MLE program and could be supplied as standard results. The tests in these two sections are related to other procedures in the literature and their power functions under local parametric alternatives are obtained. In section § 5.3 we concentrate on departures that render the QMLE inefficient or may lead to incorrect inferences, but do not affect the estimator consistency. Some Monte Carlo evidence on test performance is presented in section § 5.4.

To consider local alternatives we follow Newey [1985a,b] and restate the basic assumption \((C0)\) allowing for an extra set of parameters \(\lambda_T\) depending on the sample size, such that \(\lambda_T = \lambda_0 + T^{-1/2} \delta\) for fixed \(\lambda_0\) and \(\delta\), and at \(\delta = 0\) we obtain the null hypothesis of "correct specification". That is

\[
y_t \mid \mathcal{F}_t \sim N \left[ \mu_t(\beta_0, \lambda_T), h_t(\theta_0, \lambda_T) \right], \quad (1)
\]

where \(\mu_t = \mu_t(\beta) = \mu_t(\beta, \lambda_0)\) and \(h_t = h_t(\theta) = h_t(\theta, \lambda_0)\) for all \(\theta \in \Theta\), or \(g_t = g_t(\theta) = g_t(\theta, \lambda_0)\) with \(g_t(\theta, \lambda_T) = (\mu_t(\beta, \lambda_T), h_t(\theta, \lambda_T))'\). The derivatives of both \(\mu_t\) and \(h_t\) with respect to the \(\lambda\) parameters obey the same assumptions as those with respect to \(\theta\).

As \(T\) increases, (1) approaches correct specification at the required rate to keep the noncentrality parameters finite in the distribution of the test-statistics that we consider in the following sections.
Corresponding to the sequence of local alternatives are the mean and variance innovations \( u_t(\lambda_T) = u_t(\beta, \lambda_T) = y_t - \mu_t(\theta, \lambda_T) \) and \( \varepsilon_t(\lambda_T) = \varepsilon_t(\theta, \lambda_T) = u_t(\lambda_T)^2 - h_t(\theta, \lambda_T) \), with \( u_t = u_t(\lambda_0) \) and \( \varepsilon_t = \varepsilon_t(\lambda_0) \) under the null hypothesis. Similarly, \( v_t(\lambda_T) = (u_t(\lambda_T), \varepsilon_t(\lambda_T))^\prime \), and \( v_t = v_t(\lambda_0) \) under the null hypothesis. The derivatives of the log-likelihood function at observation \( t \) are

\[
d_{\theta t}(\theta, \lambda_T) = \frac{\partial g_t(\theta, \lambda_T)'}{\partial \theta} \Sigma_t(\lambda_T)^{-1} u_t(\lambda_T) , \tag{2}
\]

and

\[
d_{\lambda t}(\theta, \lambda_T) = \frac{\partial g_t(\theta, \lambda_T)'}{\partial \lambda} \Sigma_t(\lambda_T)^{-1} u_t(\lambda_T) , \tag{3}
\]

where \( \Sigma_t(\lambda_T) = \Sigma_t(\theta, \lambda_T) = \text{diag} \{ h_t(\lambda_T), 2h_t(\lambda_T)^2 \} \), and \( \Sigma_t = \Sigma_t(\theta, \lambda_0) \) under the null hypothesis. The score for the \( t \)-th observation under the null hypothesis is \( \frac{d_{\theta t} = d_{\theta t}(\theta, \lambda_0)}{d_{\theta}(\theta) = T^{-1} \sum_{t=1}^{T} d_{\theta t}} . \) The negative of the Hessian with respect to \( \theta \) is

\[
J_{\theta t}(\theta, \lambda_T) = \frac{\partial g_t(\theta, \lambda_T)'}{\partial \theta} \Sigma_t(\lambda_T)^{-1} \frac{\partial g_t(\theta, \lambda_T)}{\partial \theta'} + A_{\theta t} ,
\]

where

\[
A_{\theta t} = [ u_t(\lambda_T) \otimes I_p ] \frac{\partial}{\partial \theta'} \Sigma_t(\lambda_T)^{-1} .
\]

Using the law of iterated expectations \( A_{\theta t} \) is seen to have zero expected value at \( \theta_0 \) under the local alternatives in (1). Therefore, the information matrix is given by

\[
J(\theta_0) = \mathbb{E} \{ T^{-1} \sum_{t=1}^{T} d_{\theta t} d_{\theta t}' \} = \mathbb{E} \{ T^{-1} \sum_{t=1}^{T} J_{\theta t} \} = \mathbb{E} \{ T^{-1} G' \Sigma^{-1} G \} , \tag{4}
\]

where all terms are evaluated at \( \theta_0 \) and \( \lambda_0 \). \( J(\theta_0) \) is the appropriate information matrix for \( \theta \) under both the null hypothesis and the sequences of local parametric alternatives. To see this we use the MVT for random functions to express

\[
\frac{\partial g_t(\theta, \lambda_T)'}{\partial \theta} = \frac{\partial g_t(\theta, \lambda_0)'}{\partial \theta} + T^{-1/2} B_T(\theta, \lambda) \delta ,
\]
where $B_T(\theta, \lambda)$ properly accommodates the second derivatives of $g_t$ with respect to $\theta$ and $\lambda$, and $\bar{\lambda} \in [\lambda_0, \lambda_T]$. Since $B_T$ is adequately bounded by assumption, the quantities involving the second term of this expression vanish asymptotically when constructing $\psi(\theta_0)$ in view of the factor $T^{-1/2}$.

§ 5.1 Coherency tests

We have seen in Chapters 3 and 4 that the $k^*$ jointly identifiable functions $\gamma$ can be estimated consistently from the information on either moment if the model is correctly specified, while inconsistencies arise in general in $\gamma_m$ or $\gamma_v$ if $\mu_t$ or $h_t$ are misspecified. Thus in general, direct comparison of the alternative estimators of these common estimable functions may warn of possible inconsistencies arising through misspecification of either conditional moment. More formally if $q = \hat{\gamma}_m - \hat{\gamma}_v$ then in general $q$ will converge to a nonzero quantity if there is misspecification in either conditional moment, while it will differ from zero due only to sampling variation when the model is correctly specified. The corresponding test-statistic is $\tau = T q' \hat{V}(q)^{-1} q$ and we have

**Theorem 5.1.** - Under (C0") - (C7) the (asymptotic) covariance matrix of

$q = \hat{\gamma}_m - \hat{\gamma}_v$ is given by $V(q) = V(\hat{\gamma}_m) + V(\hat{\gamma}_v)$, which is positive definite. The asymptotic distribution of the test-statistic is

$$
\tau \xrightarrow{d} \chi^2(k^*; \delta' D' V(q)^{-1} D \delta), \tag{5}
$$

where $D = \Gamma V(\hat{\beta}_m) \mathcal{S} \{ T^{-1} \Omega^{-1} M_\lambda \} - \frac{1}{2} (I_{k^*}, 0) V(\hat{\phi}_v) \mathcal{S} \{ T^{-1} S_{v'} \Omega^{-2} H_\lambda \}$, $\Gamma = \partial \gamma / \partial \beta'$, $M_\lambda = \partial \mu / \partial \lambda'$, and $H_\lambda = \partial h / \partial \lambda'$, and all expectations are evaluated at the true values under correct specification.

**Proof:** Using the Mean Value Theorem for random functions we can write

$$
\mu_t(\beta, \lambda_T) = \mu_t(\beta) + \frac{\partial \mu_t(\beta, \lambda)}{\partial \lambda'} T^{-1/2} \delta, \text{ or } \mu(\beta, \lambda_T) = \mu(\beta) + T^{-1/2} M_\lambda \delta,
$$
where $\bar{\lambda} \in [\lambda_0, \lambda_T]$ and $M_{\lambda} = \partial \mu / \partial \lambda'$ evaluated at $\bar{\lambda}$. The first order conditions now are

$$\hat{X}' \Omega^{-1} \hat{X} (\hat{\beta}_m - \beta_0) = \hat{X}' \Omega^{-1} (\bar{u} + T^{-1/2} M_{\lambda} \delta),$$

where $\bar{u}_t = y_t - \mu_t (\beta, \lambda_T)$ are the true innovations. But $T^{-1} \hat{X}' \Omega^{-1} \hat{X} \xrightarrow{as} \mathcal{E} (T^{-1} X' \Omega^{-1} X) = V(\hat{\beta}_m)^{-1}$, and $T^{-1/2} \hat{X}' \Omega^{-1} \bar{u} \xrightarrow{d} N [0, V(\hat{\beta}_m)^{-1}]$, and hence

$$T^{1/2} (\hat{\beta}_m - \beta_0) \xrightarrow{d} N [B \delta, V(\hat{\beta}_m)],$$

where $B = V(\hat{\beta}_m) \mathcal{E} (T^{-1} X' \Omega^{-1} M_{\lambda})$, and so for $\gamma_m = \gamma(\beta_m)$ it follows that

$$T^{1/2} (\gamma_m - \gamma_0) \xrightarrow{d} N [\Gamma B \delta, \Gamma V(\beta_m) \Gamma'] .$$

For the variance equation the term in $\hat{u}_t^2 - u_t^2$ still vanishes in view of Lemma 3.3, and using the Mean Value Theorem for random functions we can write

$$h_t (\theta, \lambda_T) = h_t (\theta) + \frac{\partial h_t (\theta, \bar{\lambda})}{\partial \lambda'} T^{-1/2} \delta, \text{ or } h(\theta, \lambda_T) = h(\theta) + T^{-1/2} H_{\lambda} \delta,$$

where $\bar{\lambda} \in [\lambda_0, \lambda_T]$ and $H_{\lambda} = \partial h / \partial \lambda'$ evaluated at $\bar{\lambda}$. The first order conditions now are

$$\hat{S}_\phi' \Omega^{-2} \hat{S}_\phi (\hat{\phi}_v - \phi_0) = \hat{S}_\phi' \Omega^{-2} (\bar{\varepsilon} + T^{-1/2} H_{\lambda} \delta),$$

where $\bar{\varepsilon}_t = u_t^2 - h_t (\theta, \lambda_T)$ are the true variance innovations. But $\frac{1}{2} T^{-1} \hat{S}_\phi' \Omega^{-2} \hat{S}_\phi \xrightarrow{as} \frac{1}{2} \mathcal{E} (T^{-1} S_{\phi'} \Omega^{-2} S_{\phi}) = V(\hat{\phi}_v)^{-1}$, and $\frac{1}{2} T^{-1/2} \hat{S}_\phi' \Omega^{-2} \bar{\varepsilon} \xrightarrow{d} N [0, V(\hat{\phi}_v)^{-1}]$, and partitioning $\phi = (\gamma', \phi_1')'$ we get

$$T^{1/2} (\hat{\gamma}_v - \gamma_0) \xrightarrow{d} N [C \delta, V(\hat{\gamma}_v)],$$

where $C = \frac{1}{2} (I_{k*}, 0) V(\hat{\phi}_v) \mathcal{E} (T^{-1} S_{\phi'} \Omega^{-2} H_{\lambda})$. From the independence of $\gamma_m$ and $\gamma_v$ it follows that

$$T^{1/2} q = T^{1/2} (\hat{\gamma}_m - \hat{\gamma}_v) \xrightarrow{d} N [\Gamma B \delta - C \delta = D \delta, V(\hat{\gamma}_m) + V(\hat{\gamma}_v) = V(q)].$$
where \( V(q) \) is positive definite because both \( V(\hat{\gamma}_m) \) and \( V(\hat{\gamma}_v) \) are so. The distribution in (5) is obtained by constructing the quadratic form and substituting a consistent estimator of the covariance matrix.

Thus, under the null hypothesis of correct specification \( \tau \) has asymptotically a central \( \chi^2(k^*) \) distribution, and the local power function of the test is given by the noncentral \( \chi^2(k^*) \) distribution. The test may be inconsistent since the noncentrality parameter may vanish under the alternative. An example is the ARCH model when the alternative belongs to the robust class described in § 4.3.

It is worth noting that in models with unidentified parameters in the variance equation the parameterization of the \( \phi \) function is not unique, and the test depends in finite samples on the parameterization chosen (Gregory and Veall [1985] and Breusch and Schmidt [1986]). The parameterization of \( \phi \) must therefore be selected a priori for inferences to have any meaning, and it should be a natural selection if it is to convince anyone apart from the researcher. This problem is not present when the variance equation is identified for the test is based directly on \( \hat{\beta}_m - \hat{\beta}_v \), as shown in

**Corollary 5.2.** Under the assumptions of Theorem 5.1 and \( \theta \) is identifiable in the variance equation, the (asymptotic) covariance matrix of \( q = \hat{\beta}_m - \hat{\beta}_v \) is given by \( V(q) = V(\hat{\beta}_m) + V(\hat{\beta}_v) \) which is positive definite. The asymptotic distribution of the test-statistic is given by

\[
\tau \xrightarrow{d} \chi^2(k^*; \delta' D' V(q)^{-1} D \delta),
\]

where \( D = V(\hat{\beta}_m) \mathcal{C}(T^{-1} X' \Omega^{-1} M_\Gamma) - \frac{1}{2} (I_k, 0) V(\hat{\beta}_v) \mathcal{C}(T^{-1} S' \Omega^{-2} H) \), and all expectations are evaluated at the true values under correct specification.

**Proof:** When \( \theta \) is identifiable in the variance equation \( \gamma = \beta \), \( k^* = k \), \( \Gamma = I_k \), \( S_\phi = S \), and \( \phi = \theta \), and the result follows from Theorem 5.1.
The Hausman interpretation of the test follows from Ruud [1984]. In fact, (3.22) shows that $\hat{\beta}_m - \hat{\beta}_v$ is asymptotically a nonsingular transformation of both $\hat{\beta} - \hat{\beta}_v$ and $\hat{\beta} - \hat{\beta}_m$, relating the test directly to Hausman's [1978] original proposition. Equivalent tests are also the test of coherency of information Theil [1971] and the Chow [1960] test, an interpretation also noted by Ruud. That is, the test of $\beta_1 = \beta_2$ in

$$y_t = \mu_t(\beta_1) + u_t \text{ and } u_t^2 = h_t(\beta_2, \alpha) + \varepsilon_t,$$

using $\hat{\Omega} = \Omega(\hat{\theta})$ as estimate of the covariance matrix, with $\hat{\theta}$ being a root-$T$ consistent estimator of $\theta$ under the null hypothesis. Using the Mean Value Theorem we can write $h_t(\beta_2, \alpha) = h_t(\beta_1, \alpha) + \tilde{w}_t'(\beta_2 - \beta_1)$, where $\tilde{w}_t = w_t(\beta, \alpha)$, and $\beta \in [\beta_1, \beta_2]$. Under the null hypothesis $b = \beta_2 - \beta_1 = 0$, and the test may be constructed as a variable addition test by including $\tilde{w}_t = w_t(\tilde{\theta})$ in the variance equation and testing its coefficient to be zero with residuals replacing the dependent variables in both equations in (6). Moreover, using the same argument as Breusch and Godfrey [1986], the LM test under normality in the latter case is based on the subvector of the score $h_t^2 w_t \varepsilon_t$, and because $E[h_t^2 w_t \varepsilon_t | \mathcal{F}_t] = 0$ under the null hypothesis, the test also has a conditional moment interpretation (Newey [1985a,b] and Tauchen [1985]). This interpretation is interesting because it is the condition for consistency of the MLE given in Theorem 2 of Pagan and Sabau [1987a] (our Lemma 4.2 when $E[u_t | \mathcal{F}_t] = 0$). It also provides the link of the coherency test analyzed here with the consistency tests of the next section.

For the unidentified variance equation case we can reparameterize in terms of jointly identifiable functions, or use a generalized inverse estimator for $\beta$ from the variance equation. We can see from (3.34) that $q = \gamma_m - \gamma_v$ is asymptotically a nonsingular transformation of $\gamma - \gamma_v$ and also of $\gamma - \gamma_m$. If a generalized inverse estimator $\hat{\beta}_v$ is used, then $q = \beta_m - \hat{\beta}_v$ is asymptotically a nonsingular transformation of $\hat{\beta} - \hat{\beta}_v$ and a singular transformation of $\hat{\beta} - \hat{\beta}_m$. In both cases,
of course, \( V(q) \) has rank \( k^* \) only and a generalized inverse must be used for \( V(q) \). Holly [1982] has established that Hausman tests are invariant to the choice of generalized inverse, but this invariance property does not apply to the choice of the jointly estimable functions \( \gamma \) or, equivalently, to the generalized solution for \( \hat{\beta}_v \). Therefore the tests based on \( \hat{\beta} - \hat{\beta}_v \) and \( \hat{\beta} - \hat{\beta}_m \) may not be equivalent. The latter is in fact invariant in view of the unique solution for \( \hat{\beta}_m \) and thus constitutes a natural choice.

§ 5.2 Consistency tests

The possibility that the MLE of \( \beta \) might inconsistently estimate \( \beta_0 \) motivates the development of tests for the adequacy of the assumed heteroskedastic specification. To appreciate our choices consider the score for \( \theta \) (see (2.11))

\[
d_\beta(\theta) = T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} w_t \varepsilon_t , \tag{7}
\]

and

\[
d_\alpha(\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t \varepsilon_t . \tag{8}
\]

The necessary and sufficient condition for the consistency of \( \hat{\beta} \) is

\[
E [ d_\beta(\beta_0, \alpha^*) ] = 0 \text{ (Lemma 4.1). Unfortunately, } d_\beta(\hat{\theta}) = 0 \text{ by construction and no test can be based on it, but from (7) sufficient conditions for } E [ d_\beta(\beta_0, \alpha^*) ] = 0 \text{ are that } E [ T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t u_t ] = 0 \text{ and } E [ T^{-1} \sum_{t=1}^{T} h_t^{-2} w_t \varepsilon_t ] = 0 , \text{ making it appropriate that tests for the adequacy of the assumed model be based upon these separate orthogonality conditions which are not imposed on the data by the QMLE.}

At a more basic level, correct specification of the two conditional moments requires \( E [ u_t | \mathcal{F}_t ] = 0 \) and \( E [ \varepsilon_t | \mathcal{F}_t ] = 0 \), implying that \( E [ \sum_{t=1}^{T} u_t ] = 0 \) and \( E [ \sum_{t=1}^{T} \varepsilon_t ] = 0 \), and this makes \( \sum_{t=1}^{T} u_t \) and \( \sum_{t=1}^{T} \varepsilon_t \) attractive as tests for the adequacy of the model in view of their simplicity. No external information is needed in their construction, unlike most diagnostic tests which introduce new data...
supplied by postulating alternative models. Replacing the unknowns $\beta$ and $\alpha$ by their MLE's, the sample quantity $T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t$ for example should tend in probability to zero under the null hypothesis; if it does not, the evidence in the squared residuals $\hat{u}_t^2$ is inconsistent with that in $\hat{\theta}_t$.

It is convenient to embed the simple tests described above in a wider class derived from sets of $n$ first order conditions of the form

$$m(\Phi, \theta) = T^{-1} \sum_{t=1}^{T} m_t(\Phi_t, \theta) = T^{-1} \sum_{t=1}^{T} \Phi_t \nu_t = T^{-1} \Phi' \nu,$$  

(9)

where the $\Phi_t = (\Phi_{1t}, \Phi_{2t})$ are $n \times 2$ matrices of measurable functions of $\mathcal{F}_t$, $\Phi_j = (\Phi_{j1}, ..., \Phi_{jT})$, $j = 1, 2$, and $\Phi = (\Phi_1', \Phi_2')'$. The class of tests in (9) is referred henceforth as the class of "consistency tests". The $\Phi_t$ must obey some regularity conditions which are clarified in § 5.2.2. The $m_t(\Phi_t, \theta)$ are $n$-vectors, and under correct specification we have

$$E[ m(\Phi, \theta) ] = 0,$$

because $E[ m_t | \mathcal{F}_t ] = \Phi_t E[ \nu_t | \mathcal{F}_t ] = 0$. The corresponding covariance matrix is

$$\text{Var}[ T^{1/2} m(\Phi, \theta) ] = E[ T^{-1} \Phi' \Sigma \Phi ],$$

(10)

because $E[ \Phi_t \nu_t \nu_s' \Phi_s' | \mathcal{F}_t ] = \delta_{ts} \Phi_t \Sigma_t \Phi_t'$, where $\delta_{ts}$ is the Kronecker delta. The two simple consistency tests described above are obtained by setting $\Phi_t = (1, 0)$ $\forall t$ and $\Phi_t = (0, 1)$ $\forall t$, respectively, and a joint consistency test is obtained making $\Phi_t = I_2$ $\forall t$.

Corresponding to the theoretical moments $m(\Phi, \theta)$ are sample moments $m(\hat{\Phi}, \hat{\theta})$ and these are suitable "consistency test-statistics", since they would tend to be close to zero if the model specification was adequate. These consistency tests are related to other procedures in the literature in § 5.2.1, their distribution is derived in § 5.2.2, their local power considered in § 5.2.3, and a simple application is presented in § 5.2.4.
§ 5.2.1 Consistency tests and other tests in the literature

Before we obtain the distribution of the sample moments \( m(\hat{\Phi}, \hat{\theta}) \), it is useful to look at these tests in somewhat greater detail and to relate them to other procedures in the literature in order to put them into the proper perspective.

The simple consistency test, say \( m_\mu \), based on the sum of ML residuals is an interesting one because many diagnostic tests for specification error (e.g. RESET, Ramsey [1969]) arose because such a criterion was not available in the general linear model as the sum of OLS residuals is identically zero whenever an intercept appears among the regressors. It does not seem to have been fully appreciated in the literature that the residuals defined by other estimators need not share this property. Consider the OLS estimator \( \hat{\beta}_2 \) of \( \beta \) and let \( \hat{u}_{\lambda t} \) be the corresponding residuals. Because \( \sum_{t=1}^{T} \hat{u}_{\lambda t} = 0 \) we have that

\[
\begin{align*}
    m_\mu &= T^{-1} \sum_{t=1}^{T} \hat{u}_t = T^{-1} \sum_{t=1}^{T} \hat{u}_t - T^{-1} \sum_{t=1}^{T} \hat{u}_{\lambda t} = T^{-1} \sum_{t=1}^{T} x_t' (\hat{\beta}_2 - \hat{\beta}) = x' (\hat{\beta}_2 - \hat{\beta}),
\end{align*}
\]

and consequently \( m_\mu \) can be viewed as a specific weighted average of the difference between the ML and OLS estimators, when the weights are the sample means of the regressors. Since \( \hat{\beta}_2 \) is consistent irrespective of the specification of the variance, the test-statistic focuses directly upon the inconsistency in the MLE of \( \beta \).

An alternative strategy for assessing the adequacy of the maintained model would be to directly compare \( \hat{\beta} \) and \( \hat{\beta}_2 \), i.e. to conduct a specification test of the form given by Hausman [1978]. In fact this was White's [1980a] suggestion, and it can be regarded as a special case of the test-statistics considered in this section viz. when \( \Phi_t = (x_t', 0) \). To see this, observe that

\[
\begin{align*}
    m(X, \hat{\beta}) &= T^{-1} \sum_{t=1}^{T} x_t \hat{u}_t = T^{-1} X' \hat{u}.
\end{align*}
\]

But \( T^{-1} X' \hat{u}_{\lambda t} = 0 \) and therefore


\[
m(\mathbf{X}, \hat{\beta}) = T^{-1} \mathbf{X}' \tilde{\mathbf{u}} = T^{-1} \mathbf{X}' (\mathbf{u} - \hat{\mathbf{u}}_t) = T^{-1} \mathbf{X}' \mathbf{X} (\hat{\beta}_t - \hat{\beta}), \tag{12}
\]

showing that \(m(\mathbf{X}, \hat{\beta})\) is a nonsingular transformation of \((\hat{\beta}_t - \hat{\beta})\) and hence the derived test-statistics are identical.

More generally, we can use the MVT for random functions (Jennrich [1969]) on \(m(\Phi, \theta)\), and after expansion around \(\theta_0\) and grouping terms of \(O_p(T^{-1})\) and smaller, we get

\[
m(\Phi, \hat{\theta}) = T^{-1} \sum_{t=1}^{T} \Phi_t \tilde{u}_t = T^{-1} \sum_{t=1}^{T} \Phi_t \tilde{u}_t + T^{-1} \sum_{t=1}^{T} \Phi_t \frac{\partial g_t(\theta_0)}{\partial \theta'} (\hat{\theta} - \theta_0) + A_T (\hat{\theta} - \theta_0) + O_p(T^{-1}),
\]

where \(A_T = T^{-1} \sum_{t=1}^{T} (\mathbf{v}_t' \otimes \mathbf{I}_n) \partial \mathbf{vec} \Phi_t / \partial \theta'\). \(A_T \overset{\text{as}}{\to} 0\) because \(\partial \mathbf{vec} \Phi_t / \partial \theta'\) is a measurable function of \(\mathcal{F}_t\), and therefore the term \(A_T (\hat{\theta} - \theta_0)\) is \(O_p(T^{-1})\) and can be relegated to the remainder. Suppose that the \(p\) orthogonality conditions \(T^{-1} \sum_{t=1}^{T} \Phi_t^* \tilde{u}_t = 0\) define the consistent estimator \(\hat{\theta}\) where \(\Phi_t^*\) includes \(\Phi_t\) as a submatrix i.e. a GMM estimator as in Hansen [1982]. We can expand similarly \(m(\Phi, \hat{\theta})\) which equals zero from the definition of \(\hat{\theta}\). Then under the assumption of a correctly specified model we have

\[
m(\Phi, \hat{\theta}) = m(\Phi, \hat{\theta}) - m(\Phi, \hat{\theta}) = T^{-1} \sum_{t=1}^{T} \Phi_t \hat{u}_t - T^{-1} \sum_{t=1}^{T} \tilde{\Phi}_t \tilde{u}_t \tag{13}
\]

\[
= T^{-1} \sum_{t=1}^{T} \Phi_t \frac{\partial g_t(\theta_0)}{\partial \theta'} (\hat{\theta} - \hat{\theta}) + O_p(T^{-1}) = T^{-1} \Phi' G (\hat{\theta} - \hat{\theta}) + O_p(T^{-1}).
\]

When \(\Phi^* = \Phi\) so that \(T^{-1} \Phi' G\) converges to a nonsingular matrix (a regularity condition for the GMM procedure), we have an asymptotically equivalent Hausman test. When \(\Phi\) is a proper submatrix of \(\Phi^*\) the consistency test is being based asymptotically on linear combinations of \(\tilde{\theta} - \hat{\theta}\).

Another special case of (13) of particular interest is when \(\Phi_t^* = \Phi_t = (h_t^2 x_t', 0)\). Then \(T^{-1} \Phi' \mathbf{v} = 0\) defines the GLS estimator \(\hat{\beta}_m\), and we have

\[
T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-1} x_t \hat{u}_t - T^{-1} \sum_{t=1}^{T} \hat{h}_m x_t \hat{u}_m = (T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t x_t') (\hat{\beta}_m - \hat{\beta}) + O_p(T^{-1}), \tag{14}
\]

which also relates to Hausman's [1978] procedure but now with GLS as the
consistent estimator under the alternative. Because \( \hat{\beta}_m - \hat{\beta} \) is asymptotically a nonsingular transformation of \( \hat{\beta}_m - \hat{\beta}_v \) when \( \beta \) is identifiable in the variance equation, the test based on this choice is also equivalent to the coherency test of Corollary 5.2. If \( \beta \) is not identifiable in the variance equation, the distribution of the statistic in (14) has a singular covariance matrix and a generalized inverse will need to be used. Alternatively, Theorem 5.1 provides an asymptotically equivalent test which is based on the maximum number of linear combinations that define a proper asymptotic distribution in (14). If there is a constant among the regressors then \( T^{-1} \sum_{t=1}^{T} \hat{h}_{mt} \hat{u}_{mt} = 0 \), and defining \( \Phi_t = (h_t^{-1}, 0) \) we get

\[
T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-1} \hat{u}_t - T^{-1} \sum_{t=1}^{T} \hat{h}_{mt} \tilde{u}_{gt} = \left[ T^{-1} \sum_{t=1}^{T} h_t^{-1} x_t \right] (\hat{\beta}_m - \hat{\beta}) + O_p(T^{-1}),
\]

which again centers on the inconsistency of \( \hat{\beta} \), since \( \hat{\beta}_m \) is robust to variance misspecification. Here the differences of estimators are weighted by the sample mean of \( h_t^{-1}x_t \).

In the above illustrations, the estimator \( \tilde{\beta} \) for \( \beta \) came from a set of first order conditions involving \( u_t \) only, and no attention was paid to the second orthogonality condition involving \( \varepsilon_t \). This meant that the consistency tests related to \( \tilde{\beta} - \hat{\beta} \). Similar relations can be found for the \( \alpha \) parameters. For simplicity, take the case where \( h_t = z_t(\beta)' \alpha \), which is the most common in applied work. Consider the SLS \( \hat{\alpha}_a \) of \( \alpha \) described in § 3.2.1. The residuals are \( \hat{\varepsilon}_{st} = \hat{u}_t^2 - z_t(\tilde{\beta})' \alpha_a \). Assuming a constant in \( z_t \) we have \( T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_{st} = 0 \) and, letting \( \Phi_t = (0, 1) \), we have for \( m_h = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t \),

\[
m_h = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t - T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_{st} = T^{-1} \sum_{t=1}^{T} z_t'(\hat{\alpha}_a - \alpha) + O_p(T^{-1})
\]

\[= \tilde{z}'(\hat{\alpha}_a - \alpha) + O_p(T^{-1}), \]

and when \( \Phi_t = (0, \bar{z}_t) \), then

\[
m(\tilde{Z}, \alpha) = T^{-1} \tilde{Z}' \varepsilon = T^{-1} Z'(\varepsilon - \hat{\varepsilon}_a) + O_p(T^{-1}) = T^{-1} Z' Z (\hat{\alpha}_a - \alpha) + O_p(T^{-1}),
\]

- (17)
which are the counterparts for \( \alpha \) of the weighted and full difference of OLS and ML estimators in (11) and (12). In the expansions for \( m_h \) and \( m(\tilde{Z}, \tilde{\alpha}) \) we have taken \( \beta \) as given because we want to concentrate on the \( \alpha \) parameters. Substituting any consistent estimator of \( \beta \) in the construction of \( \sqrt{T} \) or \( T^{-1} Z'Z \) has no effect on the asymptotic distribution of the statistics.

We can also consider the SGLS estimator \( \hat{\alpha}_g \) for \( \alpha \), so setting \( \Phi_t = (0, h_t^2) \) and \( \Phi_t = (0, h_t^2 z_t') \) we get, respectively,

\[
T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-2} \varepsilon_t - T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-2} \hat{e}_g t = [T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t'] (\hat{\alpha}_g - \hat{\alpha}) + O_p(T^{-1}) , \quad (18)
\]

and

\[
T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-2} z_t \varepsilon_t - T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-2} \hat{e}_g t = [T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t z_t'] (\hat{\alpha}_g - \hat{\alpha}) + O_p(T^{-1}) , \quad (19)
\]

as counterparts for \( \alpha \) of (15) and (14). We can also use joint mean/variance consistency statistics e.g. consider \( m_h \) and \( m_h \) jointly.

The simple consistency tests also arise from residual analysis to diagnose the model ([Pagan and Hall [1983]]). The constraints \( E [ u_t | \mathcal{F}_t ] = 0 \) and \( E [ \varepsilon_t | \mathcal{F}_t ] = 0 \) may be written as

\[
E [ u_t | \mathcal{F}_t ] = \xi_1 ,
\]

and

\[
E [ \varepsilon_t | \mathcal{F}_t ] = \xi_2 ,
\]

where \( \xi_1 = \xi_2 = 0 \). From these expectations we derive the estimating equations

\[
\hat{u}_t = \xi_1 + u_t + (\hat{u}_t - u_t) , \quad (20a)
\]

and

\[
\hat{\varepsilon}_t = \xi_2 + \varepsilon_t + (\hat{\varepsilon}_t - \varepsilon_t) , \quad (20b)
\]
Regressing $\hat{u}_t$ and $\hat{e}_t$ against a constant yields $\hat{\xi}_1 = T^{-1} \sum_{t=1}^{T} \hat{u}_t = m_u$, and $\hat{\xi}_2 = T^{-1} \sum_{t=1}^{T} \hat{e}_t = m_e$. The argument extends to the general class of consistency tests $m(\Phi, \theta)$, where the restriction $E [ u_t | F_t ] = 0$ may be expressed as

$$E [ u_t | F_t ] = \Phi_t' \xi,$$

when $\xi = 0$, and we derive the estimating equations

$$\hat{u}_t = \Phi_t' \xi + u_t + \{ \hat{u}_t - u_t \} + \{ \Phi_t - \hat{\Phi}_t \}' \xi,$$

so now the regression of $\hat{u}_t$ would be against $\hat{\Phi}_t$. To ensure that $\Phi_t$ is a function only of information available after MLE it will need to be made a function of either $h_t$ or perhaps past values of $\hat{u}_t$ and $\hat{e}_t$. For example selection of $\Phi_t = I_2 \otimes h_t^{-1}$ produces tests based upon $\sum_{t=1}^{T} h_t^{-1} \hat{u}_t$ and $\sum_{t=1}^{T} (h_t^{-1} \hat{e}_t^2 - 1)$, statistics that appear as outputs in many ARCH programs. Setting $\Phi_t = (0, h_t)$ would effectively produce a test based upon whether the coefficient of $\hat{h}_t$ in the regression of $\hat{u}_t$ against $\hat{h}_t$ was equal to unity, the value the coefficient would be if the model was correctly specified. Thus we focus attention upon the innovations $\hat{u}_t$ and $\hat{e}_t$. An analysis of these quantities should be the primary mode of detecting specification errors in heteroskedastic models. The idea is closely linked to the tests for over- and under-dispersion in Cameron and Trivedi [1985].

The consistency tests are also related to the LM test for variable addition (Breusch and Pagan [1980], Engle [1982b, 1984], Pagan [1984a] inter alia). This test is considered in Chapter 6 where we show that such LM tests amount to choosing a specific $\Phi_t$ normally from data outside the model, and it has the potential to be the best diagnostic if the chosen $\Phi_t$ correlates well with the specification error.
§ 5.2.2 The distribution of consistency test-statistics

Our major problem with implementing the consistency tests lies in the derivation of the limiting distribution of $T^{1/2} m(\hat{\Phi}, \hat{\theta})$. Because $\hat{\theta}$ is obtained as the solution to a set of first order conditions, $d_0(\hat{\theta}) = 0$, it is apparent that $m(\hat{\Phi}, \hat{\theta})$ are conditional moment restrictions of the sort analyzed in Tauchen [1985] and Newey [1985a,b]. The last paper is particularly relevant here and Theorem 5.3 below is extracted from it.

The matrix of derivatives of $m(\Phi, \theta)$ with respect to $\theta$ has an outer product form which under both the null and the sequence of local alternatives is given by (see Newey [1985a, p. 1051-1052])

$$ M = M(\theta_0) = \mathcal{E}\left\{ T^{-1} \sum_{t=1}^{T} \frac{\partial m_t(\theta_0)}{\partial \theta'} \right\} = -\mathcal{E}\left\{ T^{-1} \sum_{t=1}^{T} m_t(\theta_0) \, d_0 \right\}. $$

Using (9) and (2) in the last equality we get

$$ M(\theta_0) = -\mathcal{E}\left\{ T^{-1} \sum_{t=1}^{T} \Phi_t \left( \frac{\partial g_t(\theta, \lambda_t)}{\partial \theta'} \right) \right\} = -\mathcal{E}\left\{ T^{-1} \Theta' G \right\}, $$

by the law of iterated expectations, evaluated under $H_0$. We now prove

Theorem 5.3.- Under (Q1"') - (Q8) and the sequence $\{\Phi_t\}$, $\Phi_t \in \mathcal{F}_t$, being such that the function $g^* = (m(\Phi, \theta)', d_0')'$ obeys the regularity, continuity, dominance and mixing conditions in assumptions (1) - (6) of Newey [1985b], then

$$ T^{1/2} m(\Phi, \theta) \xrightarrow{d} N [\psi, Q_\Phi], $$

where

$$ \psi = [\mathcal{E}\left\{ T^{-1} \Theta' G_\lambda \right\} - M(\theta_0) \, d(\theta_0)^{-1} \, \mathcal{E}\left\{ T^{-1} G' \, \Sigma^{-1} G_\lambda \right\}] \delta, $$

$$ Q_\Phi = V_0 - M(\theta_0) \, d(\theta_0)^{-1} \, M(\theta_0)' , $$

$G_\lambda = \partial g / \partial \lambda'$, and $V_0$ is the limit of the covariance matrix in (10), with all expectations evaluated under $H_0$. A consistent estimator of $Q_\Phi$ is
\[ \hat{Q}_\Phi = T^{-1/2} \hat{\Phi}^* \hat{\Sigma}^{1/2} \hat{m} \hat{\Sigma}^{1/2} \hat{\Phi}, \]  

where \[ \hat{m} = I_{2T} - \Sigma^{-1/2} G ( G' \Sigma^{-1} G )^{-1} G' \Sigma^{-1/2} \] and all estimates are under \( H_0 \).

**Proof.** The result is Lemma 1 in Newey [1985b] with some specialization. In our context \( d_{\theta t} (\theta_0) \) and \( m_t (\Phi, \theta) \) are martingale differences with respect to the \( \sigma \)-field \( \mathcal{F}_t \), and so the matrix \( V \) simplifies from Newey's corresponding expression to

\[
V = \begin{pmatrix}
V_0 & -M(\theta_0) \\
-M(\theta_0)' j(\theta_0)
\end{pmatrix}.
\]

Let \( L \) be the matrix selecting \( m(\Phi, \theta) \) from \( g^* \), \( L = (I_n, 0) \), \( C = \mathbb{E}[\frac{\partial g^* (\theta_0)}{\partial \theta}'] \), \( \mathcal{C} = (0, I_p) \), and \( P_c = I_{p+n} - C (C' C)^{-1} C' \). Then using (4) and (22), \( L P_c = (I_n, -M(\theta_0) j(\theta_0)^{-1}) \), and hence the covariance matrix of \( T^{1/2} m(\Phi, \theta) \) is given by

\[ Q_\Phi = L P_c V P_c' L' = V_0 - M(\theta_0) j(\theta_0)^{-1} M(\theta_0)' \cdot \]

Again using (4) and (22) together with (10), and following the argument in Newey [1985a, p.1052], especially that connected with his equation (2.11), yields the consistent estimator \( \hat{Q}_\Phi \) of \( Q_\Phi \) in (24). Finally, note that

\[ \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} m_t d_{\lambda t} \right] = \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} \Phi_t u_t' \Sigma_t^{-1} \frac{\partial g_t}{\partial \lambda} \right] = \mathbb{E} \left[ T^{-1} \Phi' G_{\lambda} \right] , \]

using iterated expectations and, similarly,

\[ \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} d_{\theta t} d_{\lambda t} \right] = \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} \frac{\partial g_t}{\partial \theta} \Sigma_t^{-1} u_t u_t' \Sigma_t^{-1} \frac{\partial g_t}{\partial \lambda} \right] = \mathbb{E} \left[ T^{-1} G' \Sigma^{-1} G_{\lambda} \right] . \]

Because these two matrices form the matrix \( K \) in Newey [1985b, p.233], the noncentrality parameter follows directly.

The limiting distribution of \( T^{1/2} m(\hat{\Phi}, \hat{\theta}) \) when the model is correctly specified is established by setting \( \delta = 0 \) in Theorem 5.3, a result given in

**Corollary 5.4.** Under the assumptions of Theorem 5.3, when the maintained model is correctly specified (\( \delta = 0 \)),
Proof: Set $\delta = 0$ in Theorem 5.3.

Cases of particular interest are when $\Phi_t = (1, 0)$ and $\Phi_t = (0, 1)$, making the simple consistency statistics $m_\mu$ and $m_h$. The next corollary specializes Theorem 5.3 to these situations:

**Corollary 5.5.** Under the assumptions of Theorem 5.3,

\[
T^{1/2} m_\mu \overset{d}{\to} N \left[ (\mu_\lambda - \bar{\mu}_\theta' V(\hat{\theta}) G_{\theta \lambda} ) \delta, \sigma^2 - \bar{\mu}_\theta' V(\hat{\theta}) \bar{\mu}_\theta \right],
\]

and

\[
T^{1/2} m_h \overset{d}{\to} N \left[ (\bar{\mu}_{\lambda'} - \bar{\mu}_\theta' V(\hat{\theta}) G_{\theta \lambda} ) \delta, 2\tau^2 - \bar{\mu}_\theta' V(\hat{\theta}) \bar{\mu}_\theta \right],
\]

where $\bar{\mu}_\lambda = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} \partial \mu_t / \partial \lambda \right)$, $\bar{\mu}_\theta = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} \partial \mu_t / \partial \theta \right)$, $G_{\theta \lambda} = \mathcal{E} \left( \frac{T}{T} G' \Sigma^{-1} G_\lambda \right)$, $\sigma^2 = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} h_t \right)$, $\bar{\mu}_{\lambda'} = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} \partial h_t / \partial \lambda \right)$, $\bar{\mu}_\theta = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} \partial h_t / \partial \theta \right)$, and $\tau^2 = \mathcal{E} \left( \frac{T}{T} \sum_{t=1}^{T} h_t^2 \right)$.

Proof: For $\Phi_t = (1, 0)$ note that $\Phi_t (\partial g_t / \partial \theta') = \partial \mu_t / \partial \theta'$, $\Phi_t (\partial g_t / \partial \lambda') = \partial \mu_t / \partial \lambda'$, and $\Phi_t \Sigma_t \Phi_t' = h_t$, and use these expressions in Theorem 5.3. For $\Phi_t = (0, 1)$ note that $\Phi_t (\partial g_t / \partial \theta') = \partial h_t / \partial \theta'$, $\Phi_t (\partial g_t / \partial \lambda') = \partial h_t / \partial \lambda'$, and $\Phi_t \Sigma_t \Phi_t' = 2 h_t^2$, and use these expressions in Theorem 5.3.

In the models we are considering the conditional mean does not depend on $\alpha$, and therefore the variance of $T^{1/2} m_\mu$ simplifies to $\sigma^2 - \bar{x}' V(\hat{\beta}) \bar{x}$, where $\bar{x}$ is the limit of the sample mean of $x_t$. Similarly, when $h_t$ does not depend on $\beta$ the variance of $T^{1/2} m_h$ simplifies to $2 \tau^2 - \bar{z}' V(\hat{\alpha}) \bar{z}$, where $\bar{z}$ is the limit of the sample mean of $z_t$. When the information matrix is block-diagonal between $\beta$ and $\alpha$ the variance of $T^{1/2} m_h$ is given by $2 \tau^2 - \bar{z}' V(\hat{\alpha}) \bar{z} - \bar{w}' V(\hat{\alpha}) \bar{w}$, where $\bar{w}$ is the limit of the sample mean of $w_t$. An important case is the ARCH model, for then $\var(\theta_0)$ is diagonal and also

\[
\bar{w} = -T^{-1} \sum_{j=1}^{q} \alpha_j u_{t-j} x_{t-j} \overset{as}{\to} 0,
\]

- (25)
resulting in the simple variance $2 \tau^2 - \overline{Z}' V(\widehat{\alpha}) \overline{Z}$. For the joint test with $\Phi_t = I_2$, proceeding as in Corollary 5.5 and noting that $V_0$ is diagonal by construction, it is seen that $\text{cov} [ m_\mu, m_h ] = - \overline{\mu}_0' V(\theta) \overline{\mu}_\theta$. The statistics are rarely independent, but here again the ARCH class is an exception. This follows from combining $\overline{w} \overset{as}{\rightarrow} 0$ with $\jmath_{\beta_\theta}(\theta_0) = 0$ and $\overline{\mu}_\alpha = 0$. Of course many other test-statistics might be used e.g. there are some advantages to employing the sum of standardized residuals $T^{-1} \sum_{t=1}^T \hat{h}_t^{-1} u_t$ or $T^{-1} \sum_{t=1}^T \hat{h}_t^{-2} \hat{e}_t$ which relate to the differences between ML and GLS estimators and thus have smaller sampling variation under $H_0$ for large $T$. All of these can be handled with Theorem 5.3 by appropriate choices of $\Phi_t$, and the asymptotic distributions obtained as in Corollary 5.5. The variances are consistently estimated by replacing expectations with sample moments.

It is important to note that the fact that $\theta$ needs to be estimated means that the asymptotic variance of the test-statistics is less than it would be if $\theta$ were known, so that inference needs to allow for this extra source of uncertainty. The situation is therefore analogous to that which occurs when testing for serial correlation in the presence of lagged dependent variables (Durbin [1970]). It is useful to reconsider the tests in the framework of Pagan and Hall [1983]. The variances of the $\widehat{\xi}_j$ in (20) and (21) depend upon two factors: the errors $u_t$ and $\epsilon_t$ and the difference between these and their estimated values. As shown in Corollary 5.5, the OLS standard errors accompanying $\widehat{\xi}_j$ overstate the true standard errors. This means that any $t$-statistic for $\xi_j = 0$ from the regressions in (20) and (21) is biased in favour of the null hypothesis. Still, such regressions can provide quick checks of the adequacy of the variance and mean specifications since, if $H_0: \xi_j = 0$ is rejected, this conclusion would not be reversed if the correct standard errors were employed.

It is also possible to produce a simplified calculation of the statistics using the correct estimate of the variance, by means of an uncentered coefficient of
determination of an auxiliary regression as in Engle [1982b, 1984], Newey [1985a], and Davidson and MacKinnon [1984], and this is especially attractive when using multidimensional tests. For this purpose we produce

**Theorem 5.6.** Under the assumptions of Theorem 5.3, 
\[ s = T \text{m}(\Phi, \hat{\theta})' Q^{-1}_\Phi \text{m}(\Phi, \hat{\theta}) \xrightarrow{d} \chi^2(n; \lambda^2_\Phi), \]  
where \( \lambda^2_\Phi = \delta' \mathcal{C} \{ T^{-1} G_\lambda' \Sigma^{-1/2} \eta_\Phi \Sigma^{-1/2} G_\lambda \} \delta, \) and \( \eta_\Phi = \eta \Sigma^{1/2} \Phi (\Phi' \Sigma^{1/2} \eta) \Sigma^{-1/2} \Phi' \Sigma^{1/2} \Phi \). Moreover, if \((G, \Sigma \Phi)\) has full column rank then \( s - s* \xrightarrow{as} 0 \) where \( s* = 2 T R_0^2 \), and \( R_0^2 \) is the uncentered coefficient of determination of the regression of \( \hat{\mu} \) on \( \hat{G} \) and \( \hat{\Sigma} \Phi \) in the metric of \( \hat{\Sigma} \).

**Proof:** From Theorem 5.3 \( s = T \text{m}(\Phi, \hat{\theta})' Q^{-1}_\Phi \text{m}(\Phi, \hat{\theta}) \xrightarrow{d} \chi^2(n; \psi Q^{-1}_\Phi \psi) \). But using (9) and (24) we get
\[ s = \hat{\mu}' \hat{\Phi} (\hat{\Phi}' \hat{\Sigma}^{1/2} \eta \hat{\Sigma}^{1/2} \hat{\Phi})^{-1} \hat{\Phi}' \hat{\mu}. \]  
From the definition of \( \hat{\theta} \) we have \( T^{-1} \hat{G}' \hat{\Sigma}^{-1} \hat{\mu} = 0 \), and using an argument identical to Engle [1982b,p.102] it follows that \( s \) is the explained sum of squares (ESS) of the auxiliary regression, and because \( \frac{1}{2} T^{-1} \hat{\mu}' \hat{\Sigma}^{-1} \hat{\mu} \xrightarrow{as} 1 \), it follows that \( s \xrightarrow{as} 2 TR_0^2 = s* \). The noncentrality parameter is obtained using (4) and (22) in the expression for \( \psi \), so that 
\[ \psi = \mathcal{C} \{ T^{-1} [ \Phi' G_\lambda - \Phi' G (G' \Sigma^{-1} G)^{-1} G' \Sigma^{-1} G_\lambda ] \} \delta \]
\[ = \mathcal{C} \{ T^{-1} \Phi' \Sigma^{1/2} \eta \Sigma^{-1/2} G_\lambda \} \delta, \]  
and the result follows using (24).

The immediate result under the null hypothesis is

**Corollary 5.7.** Under the assumptions of Theorem 5.6 and the model being correctly specified \((\delta = 0)\), \( s \xrightarrow{d} \chi^2_n \) and \( s - s* \xrightarrow{as} 0 \).

**Proof:** Setting \( \delta = 0 \) in Theorem 5.6 results in \( \lambda^2_\Phi = 0 \).
Theorem 5.6 is general enough to accommodate wide classes of heteroskedastic models, allowing for separate or joint testing of either conditional moment. For example, the additional variables to the bivariate regression are \((h_t, 0)\) for (11), or \((h_t x_t', 0)\) for (12). This establishes a relation between variable addition and variable transformation tests as in Breusch and Godfrey [1986].

The auxiliary regression of Theorem 5.6 uses the model residuals as dependent variable, and incorporates the specific form of the information matrix in heteroskedastic models. Thus the testing procedure has more structure than the more general auxiliary regressions of Newey [1985a] and Davidson and MacKinnon [1984].

Because many tests in OLS and GLS regression can be expressed asymptotically as \(T \mathbf{R}_0^2\) from a single-length auxiliary regression with the residuals as dependent variable (Engle [1982a,1982b,1984]), it is of some importance to analyze under what conditions the consistency tests can be equivalently obtained from single-length auxiliary regressions. The first requirement is, of course, that the test focuses on only one set of residuals (\(u\) or \(e\)). Hence we consider the partition \(\Phi = (\Phi_1', \Phi_2')\), so that when \(\Phi_2 = 0\) the consistency test is based on \(\hat{u}\), and when \(\Phi_1 = 0\) the consistency test is based on \(\hat{e}\). Using this partition in (27) and denoting \(S = S(\Phi) = S(\Phi_1, \Phi_2)\) it is seen that

\[
s(\Phi_1, 0) = \hat{u}' \hat{\Phi}_1 (\hat{\Phi}_1' \hat{\Omega}_1 \hat{\Phi}_1 - \hat{\Phi}_1' \hat{X} (\hat{X}' \hat{\Omega}_1^{-1} \hat{X})^{-1} \hat{X}' \hat{\Phi}_1)^{-1} \hat{\Phi}_1' \hat{u},
\]

and

\[
s(0, \Phi_2) = \hat{e}' \hat{\Phi}_2 (2 \hat{\Phi}_2' \hat{\Omega}_2 \hat{\Phi}_2 - \hat{\Phi}_2' \hat{S} (\hat{S}' \hat{\Sigma}_1^{-1} \hat{S})^{-1} \hat{S}' \hat{\Phi}_2)^{-1} \hat{\Phi}_2' \hat{e}.
\]

As an alternative we can use an estimator \(\hat{\beta}\) and residuals \(\hat{u}\), and obtain the ESS of the regression of \(\hat{u}\) on \(\hat{X}\) and \(\hat{\Omega}\hat{\Phi}_1\) in the metric of \(\hat{\Omega}\), that is,

\[
s^*(\hat{\beta}) = \hat{u}' \hat{\Omega}_1^{-1/2} \hat{N} \hat{\Omega}_1^{-1/2} \hat{u} = \hat{u}' \hat{\Omega}_1^{-1/2} \hat{N}_1 \hat{\Omega}_1^{-1/2} \hat{u} + \hat{u}' \hat{\Omega}_1^{-1/2} \hat{N}_2 \hat{\Omega}_1^{-1/2} \hat{u},
\]

where \(\hat{N}, \hat{N}_1\), and \(\hat{N}_2\) are respectively the projection matrices onto the range
spaces of \(( \tilde{\Omega}^{1/2} \tilde{X}, \tilde{\Omega}^{1/2} \tilde{F}_1 )\), \(\tilde{\Omega}^{-1/2} \tilde{X}\), and \(( I_T - \tilde{N}_1 ) \tilde{\Omega}^{1/2} \tilde{F}_1 \). Hence \(\tilde{N} = \tilde{N}_1 + \tilde{N}_2\).

When \(\tilde{\beta} = \tilde{\beta}_m\) it follows that \(\tilde{X}' \tilde{\Omega}^{-1} \tilde{u} = 0\), and then

\[
s*(\tilde{\beta}) = \tilde{u}' \tilde{g}_1 ( \tilde{\Phi}_1' \tilde{\Omega} \tilde{\Phi}_1 - \tilde{\Phi}_1' \tilde{X} ( \tilde{X}' \tilde{\Omega}^{-1} \tilde{X} )^{-1} \tilde{X}' \tilde{g}_1 )^{-1} \tilde{g}_1' \tilde{u},
\]

which is of similar form to \(s( \Phi_1, 0 )\). If \(\tilde{\beta} \neq \tilde{\beta}_m\) then \(\tilde{X}' \tilde{\Omega}^{-1} \tilde{u} \neq 0\), and the latter expression for \(s*(\tilde{\beta})\) is not correct. Thus a limited information version of consistency tests based on residuals from the mean equation can be constructed using the GLS estimator. Such tests have adequate size but are less powerful than \(s( \Phi_1, 0 )\) because they use less information, unless \(\tilde{\beta}_m = \tilde{\beta}\) (i.e. simple heteroskedasticity).

Similarly we can use an estimator \(\tilde{\alpha}\) and residuals \(\tilde{e}\), and obtain the ESS of the regression of \(\tilde{f}\) on \(\tilde{Z}\) and \(2 \tilde{\Omega}^{2} \tilde{\Phi}_2\) in the metric of \(2 \tilde{\Omega}^2\), that is,

\[
s*(\tilde{\alpha}) = \frac{1}{2} \tilde{e}' \tilde{\Omega}^{-1} \tilde{N}^* \tilde{\Omega}^{-1} \tilde{u} = \frac{1}{2} \tilde{e}' \tilde{\Omega}^{-1} \tilde{N}_1^* \tilde{\Omega}^{-1} \tilde{e} + \frac{1}{2} \tilde{e}' \tilde{\Omega}^{-1} \tilde{N}_2^* \tilde{\Omega}^{-1} \tilde{e},
\]

where \(\tilde{N}^*, \tilde{N}_1^*,\) and \(\tilde{N}_2^*\) are respectively the projection matrices onto the range spaces of \(( \frac{1}{\sqrt{2}} \tilde{\Omega}^{-1} \tilde{Z}, \sqrt{2} \tilde{\Omega} \tilde{g}_1 )\), \(\frac{1}{\sqrt{2}} \tilde{\Omega}^{-1} \tilde{Z}\), and \(\sqrt{2} \left( I_T - \tilde{N}_1^* \right) \tilde{\Omega} \tilde{F}_1\). Hence \(\tilde{N}^* = \tilde{N}_1^* + \tilde{N}_2^*\). When \(\tilde{\alpha} = \hat{\alpha}\) (SGLS) it follows that \(\tilde{Z}' \tilde{\Omega}^{-2} \tilde{e} = 0\), and then

\[
s*(\hat{\alpha}) = \frac{1}{2} \tilde{e}' \tilde{f}_2 ( \tilde{f}_2' \tilde{\Omega}^{2} \tilde{f}_2 - \tilde{\Phi}_2' \tilde{Z} ( \tilde{Z}' \tilde{\Omega}^{2} \tilde{Z} )^{-1} \tilde{Z}' \tilde{\Phi}_2 )^{-1} \tilde{\Phi}_2' \tilde{e}.
\]

In what sense are \(s*(\hat{\alpha})\) and \(s( \Phi_2, \Phi_2 )\) similar? When \(\lambda_{\Phi}(\theta_0) = 0\), SGLS is asymptotically the MLE, and we can write

\[
\Phi_2' S V(\theta) S' \Phi_2 = \Phi_2' W V(\tilde{\beta}) W' \Phi_2 + \Phi_2' Z V(\hat{\alpha}) Z' \Phi_2,
\]

and if \(T^{-1} \Phi_2' W \xrightarrow{as} 0\) then \(s*(\hat{\alpha})\) and \(s( \Phi_2, \Phi_2 )\) are asymptotically equivalent.

For the simple heteroskedasticity model \(W = 0\) and \(s*(\hat{\alpha}) = s( \Phi_2, \Phi_2 )\). For the ARCH model a sufficient condition for the asymptotic equivalence is that \(\Phi_2t\) be a conditionally even function of \(u_{t,j}\) (Lemma 3 of Pagan and Sabau [1987a]).

The above relations refer to the original form of the statistic in (26) or (27). If an \(R^2\) computation is used the same asymptotic equivalences apply but the
exact relations in the case of simple heteroskedasticity cease to hold.

§ 5.2.3 Some power considerations

In order to analyze the local power of the tests, it is convenient to decompose $\psi$ in (23b) into $\psi = \psi_{\mu} + \psi_{h}$, where $\psi_{\mu}$ summarizes the inconsistency arising from a misspecified conditional mean, and $\psi_{h}$ summarizes the inconsistency arising from a misspecified conditional variance. Partitioning $\Phi_t = (\Phi_{1t}, \Phi_{2t})$ and after simple algebraic manipulation we get

$$\psi_{\mu} = \mathcal{N} \left\{ \sum_{t=1}^{T} (\Phi_{1t} - M(\theta_0)) V(\hat{\theta}) \frac{1}{\partial \theta} \left( \frac{\partial \mu_t}{\partial \lambda} \right) \delta, \right\} \quad (29a)$$

and

$$\psi_{h} = \mathcal{N} \left\{ \sum_{t=1}^{T} (\Phi_{2t} - \frac{1}{2} M(\theta_0)) V(\hat{\theta}) \frac{1}{\partial \lambda} \left( \frac{\partial h_t}{\partial \lambda} \right) \delta, \right\} \quad (29b)$$

We also decompose $M(\theta_0) = M_{\mu}(\theta_0) + M_{h}(\theta_0)$, where $M_{\mu}(\theta_0) = \mathcal{N} \left\{ \sum_{t=1}^{T} \frac{\partial \mu_t}{\partial \theta} \right\}$, and $M_{h}(\theta_0) = \mathcal{N} \left\{ \sum_{t=1}^{T} \frac{\partial h_t}{\partial \theta} \right\}$.

Since $Q_{\theta}$ is positive definite, the tests have power against misspecification in $\mu_t$ whenever $\psi_{\mu} \neq 0$, and against misspecification in $h_t$ whenever $\psi_{h} \neq 0$. It is of particular interest to examine the power of tests designed for one conditional moment when misspecification appears only in the other.

Let us consider first a mean test with a correctly specified conditional mean. Then $\Phi_{2t} = 0$ and so $M(\theta_0) = M_{\mu}(\theta_0)$, and also $\partial \mu_t / \partial \lambda = 0$. From (29) we get $\psi_{\mu} = 0$ and therefore $\psi = \psi_{h} = \mathcal{N} \left\{ \sum_{t=1}^{T} h_t^2 \frac{\partial h_t}{\partial \theta} \right\} \delta$. Only in special cases will $\psi_{h}$ vanish. Hence when there is inconsistency in $\hat{\beta}$ arising from conditional variance error this induces power through $\psi_{h}$. In the cases in Chapter 4 where we established consistency of $\hat{\beta}$ despite a misspecified $h_t$, the expected value in $\psi_{h}$ vanishes and the test has no power.

The other case is that of a variance test with a correctly specified conditional variance. Then $\Phi_{1t} = 0$ and $\partial h_t / \partial \lambda = 0$, so $M(\theta_0) = M_{h}(\theta_0)$ and $\psi_{h} = 0$.
This produces $\psi = \psi_\mu = - M_h(\theta_0) V(\hat{\theta}) \& (T^{-1} \sum_{t=1}^T h_t^{-1} \frac{\partial u_t}{\partial \theta} \frac{\partial u_t}{\partial \lambda}) \delta$ which is zero only in very special cases. For example, when the misspecification is in the form of autocorrelated errors and neither conditional moment depends on lagged y's. Hence depending upon the context the variance test is capable of detecting misspecification in the conditional mean.

The above cases place properly into perspective our terming the tests as "consistency" tests, for any departure not affecting the consistency of the subset of parameters on which the test focuses are part of the implicit null hypothesis.

Performing groups of consistency tests singly and jointly may provide a valuable tool for assessing the model and, because of Theorem 5.6, performing a wide range of tests has a small computational cost.

Another point of interest in terms of power is the comparison of the one-degree-of-freedom and the n-degrees-of-freedom consistency tests presented in (11) - (19) (n = k for mean tests and n = p - k for variance tests). The one-degree-of-freedom tests are favored for having less degrees of freedom (see DasGupta and Perlman [1974]), but the noncentrality parameter (NCP henceforth) is larger for the n-degrees-of-freedom, as we show in

**Theorem 5.8.** Suppose the hypotheses of Theorem 5.3 hold and let $m(\Phi, \theta)$ denote a consistency statistic with n degrees of freedom, and A be a $n \times n_1$ matrix of rank $n_1 \leq n$. Then the asymptotic distribution of the test-statistic based on $m(\Phi, \theta)$ has NCP no smaller than the one based on $m(\Phi A, \theta)$.

**Proof.** Let $D = \Phi \Sigma^{1/2} \Phi$. From Theorem 5.6, the NCP for the test based on $m(\Phi, \theta)$ is $\lambda_{\Phi}^2 = \delta' \& (T^{-1} G_{\lambda} \Sigma^{-1/2} D (D' D)^{-1} D' \Sigma^{-1/2} G_{\lambda}) \delta$, and that for the test based on $m(\Phi A, \theta)$ is $\lambda_{\Phi A}^2 = \delta' \& (T^{-1} G_{\lambda} \Sigma^{-1/2} D A (A' D' D A)^{-1} A' D' \Sigma^{-1/2} G_{\lambda}) \delta$.

But

$B = D(D' D)^{-1} D' - DA(A' D' D A)^{-1} A' D' = D R^{-1} [I_n - R A (A' R' R A)^{-1} A' R'] R^{-1} D'$,

where $D' D = R' R$ for nonsingular R, and the matrix in the middle is a projection matrix. Hence B is positive semidefinite, so that $\lambda_{\Phi A}^2 \leq \lambda_{\Phi}^2$. \(\square\)
In the cases of (11) - (19) the matrix $A$ simply selects the element of $x_t$ or $z_t$ assumed to be unity. Therefore we cannot obtain a general conclusion as to which tests have better power and this depends on the structure of $x_t$ and $z_t$ in each case, and on the form of the departure from the null hypothesis.

It is also of interest to compare the power of the simple consistency tests based on direct sums of residuals as in (11), (12), (16), and (17), with that of tests based on weighted sums of residuals as in (14), (15), (18), and (19). If we let $m(\Phi, \theta)$ be the basic statistic for any of the direct sum tests, the basic statistic for the corresponding weighted sum test is $m(\Sigma^{-1}\Phi, \theta)$. The NCP for $m(\Phi, \theta)$ is $\lambda_\Phi^2$ as given in (26), and $\lambda_{\Sigma^{-1}\Phi}^2 = \delta^2 \delta' \left( \Sigma^{-1/2} \Sigma^{-1/2} \Phi \Sigma^{-1/2} \Gamma \right)$ is the NCP for $m(\Sigma^{-1}\Phi, \theta)$, where $\Sigma^{-1}\Phi \equiv \Sigma^{-1/2} \Phi (\Phi' \Sigma^{-1/2} \Sigma^{-1/2} \Phi - 1) \Phi' \Sigma^{-1/2} \Sigma^{-1/2} \Phi$. Since both tests have the same degrees of freedom, the question is whether $\lambda_{\Sigma^{-1}\Phi} - \lambda_\Phi^2$ is positive semidefinite to result in $\lambda_\Phi^2 \leq \lambda_{\Sigma^{-1}\Phi}^2$. We have not been able to prove this in general. If both OLS and GLS remain consistent under the alternative hypothesis there is a heuristic argument stemming from the relative efficiency of GLS to suggest that $\lambda_\Phi^2 \leq \lambda_{\Sigma^{-1}\Phi}^2$. But this would not apply to the one-degree-of-freedom tests because the two estimators are weighted differently. If the departure from the null hypothesis renders OLS or GLS (or both) inconsistent, their differences with the MLE will have different sizes and affect differently the NCP's.

The power of general consistency tests against specific departures may be assessed by comparison to the relevant LM tests, which are derived and analyzed in Chapter 6. There we make clear that to evaluate the power of a consistency test in a specific direction all that needs to be done is to assess how well the relevant matrix $\Phi$ projects onto the space spanned by $\Sigma^{-1} \Gamma$. The smaller the distance of $\Phi$ from this space, the greater the power in the specific direction under analysis. The situation resembles that of choice of instruments to achieve efficiency in estimation.
To illustrate the argument let us consider the linear ARCH(1) model

\[ y_t \mid \mathcal{F}_t \sim N [ x_t' \beta, h_t = \alpha_0 + \alpha_1 u_{t-1}^2 = z_t' \alpha ] , \]

where \( z_t = (1, u_{t-1}^2)' \) and \( \alpha = (\alpha_0, \alpha_1)' \). Suppose the true conditional variance \( h_t \) departs locally in the direction of \( h_t^* = \alpha_0 + \alpha_1 y_{t-1}^2 \). Using (4.8) we may write

\[ h_t^* = h_t - w_t' \beta + \alpha_1 (x_{t-1}' \beta)^2 , \]

and we may consider separately the two departure directions by making

\[ \bar{h}_t = h_t - T^{-1/2} \delta_1 w_t' \beta + T^{-1/2} \delta_2 (x_{t-1}' \beta)^2 = h_t - T^{-1/2} z_{at}' \delta , \]

where \( \bar{\delta} = (\delta_1, \delta_2)' \) and \( z_{at} = (-w_t' \beta, (x_{t-1}' \beta)^2)' \). Weiss [1984] has found statistical evidence for the presence of \( y_{t-j}^2 \) in the conditional variance of time series models for many economic variables, and so the alternative \( h_t^* \) is interesting empirically. It is also interesting theoretically because the two departure directions in (32) are, respectively, conditionally odd and conditionally even functions of \( u_{t-1} \) (see Chapter 4). Consider the simple tests based on \( m_\mu \) and \( m_h \). From Corollary 5.5 and (25) the variances are \( \sigma_\mu^2 = \bar{\Sigma}' \alpha - \bar{\Sigma}' V(\hat{\beta}) \bar{\Sigma} \) and \( \sigma_h^2 = 2 \tau^2 - \bar{\Sigma}' V(\hat{\alpha}) \bar{\Sigma} \), respectively, where \( \tau^2 = \alpha' \mathcal{E}(T^{-1} \bar{Z}' \bar{Z}) \alpha \).

Because the mean is correctly specified the term \( \psi_\mu \) in (29) is zero, and to obtain the term \( \psi_h \) we need

\[ T^{-1} \sum_{t=1}^{T} h_t^2 \frac{\partial h_t}{\partial \theta} \frac{\partial h_t}{\partial \lambda'} = T^{-1} \begin{pmatrix} - \sum_{t=1}^{T} h_t^2 w_t w_t' \beta & \sum_{t=1}^{T} h_t^2 w_t (x_{t-1}' \beta)^2 \\ - \sum_{t=1}^{T} h_t^2 z_t w_t' \beta & \sum_{t=1}^{T} h_t^2 z_t (x_{t-1}' \beta)^2 \end{pmatrix} , \]

and because \( w_t \) is conditionally odd in \( u_{t-1} \) while \( h_t^2 z_t \) and \( h_t^2 (x_{t-1}' \beta)^2 \) are conditionally even in \( u_{t-1} \), it follows from Lemma 3 of Pagan and Sabau [1987a] that the off-diagonal elements of this matrix have zero expectation. Therefore,

\[ \frac{1}{2} \mathcal{E} \left( T^{-1} \sum_{t=1}^{T} h_t^2 \frac{\partial h_t}{\partial \theta} \frac{\partial h_t}{\partial \lambda'} \right) = \text{diag}( V(\hat{\beta}_v)^{-1} \beta_0, d ) , \]

where \( d = \mathcal{E}( \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t^2 z_t (x_{t-1}' \beta_0)^2 ) \) and the expectation has been evaluated.
under correct specification. Use has been made of $V(\hat{\beta}_v) = 2 \mathcal{C} \{ T^{-1} W' \Omega^{-2W} \}^{-1}$. Also, $\beta_0 = \delta_1 c_1$, where $c_1 = \mathcal{C} \{ T^{-1} \sum_{t=1}^{T} (x_{t-1} \beta_0)^2 \}$, in view of (25). Finally, using the diagonality of the information matrix between $\beta$ and $\alpha$, we get

$$s_\mu = T \frac{m_\mu^2}{\sigma_\mu^2} \frac{d}{\chi^2} (1; \lambda_\mu = \delta_1^2 \{ \bar{X}' V(\hat{\beta}) V(\hat{\beta}_v)^{-1} \beta_0 \}^2 / \sigma_\mu^2), \quad - (33a)$$

and

$$s_h = T \frac{m_h^2}{\sigma_h^2} \frac{d}{\chi^2} (1; \lambda_h = \delta_2^2 (c_1 - \bar{Z}' V(\hat{\alpha}) d)^2 / \sigma_h^2). \quad - (33b)$$

The power for the mean test comes from the odd misspecification direction. This is in agreement with our previous findings because the even term does not induce any inconsistency in $\hat{\beta}$, while an odd term produces inconsistency when combined with an even term. From section § 3.4, the factor $V(\hat{\beta}) V(\hat{\beta}_v)^{-1}$ is just the (generalized) proportional contribution to the efficiency of the MLE for $\beta$ obtained from the information in the conditional variance. Therefore, the power of $s_\mu$ increases with the informativeness of the variance relative to the mean (about $\beta$). This is the natural thing to expect because only when the signal from the variance is very clear we can get substantial information from it. Thus power is good in the cases when there may be a substantial gain in efficiency over estimating $\beta$ by OLS. The power of the variance test, on the other hand, comes completely from the even term. The odd term vanishes in the expectation when multiplied by the even function $z_t$ of $u_{t-1}$. Clearly, power grows as the presence of the lagged squared mean becomes clearer in the true conditional variance.

§ 5.2.4 A simple example

Engle and Bollerslev [1986] study the evolution of the exchange rate between the US dollar and the Swiss franc. They fit a pure GARCH model (zero conditional mean) to the difference of logs of the exchange rate. Weekly data from July 1973 to August 1985 is used, with a total of 632 observations. The data
is reported as an appendix to their paper, and Engle and Bollerslev present three estimated specifications in their equations (24), (26), and (34). The first two are modelled as conditionally normal and the last as conditionally t-distributed, and variance integration is imposed on the second and third models. A range of LM tests were performed by the authors to assess the specification of the models.

As a quick check, we have regressed the variance residuals on a constant, on $h_t$, and on $h_t^{-1}$, to produce consistency tests based on $T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t$, $T^{-1} \sum_{t=1}^{T} \hat{h}_t \hat{\varepsilon}_t$, and $T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-1} \hat{\varepsilon}_t$, respectively. Significance of these statistics with the usual regression t-statistics would then constitute a warning of possible misspecification. The t-ratios were -4.31, -6.92, and -2.55, respectively, for their equation (24); 0.01, -3.47, and 5.14 for their equation (26); and 0.01, -3.89, and 5.23 for their equation (36). This suggests possible misspecification in these models. The interesting thing to notice is that in all these cases the LM tests applied by Engle and Bollerslev did not detect any source of specification error and thus highlights the potential of consistency tests.

§ 5.3 Efficiency tests for higher order moments

The tests of the previous two sections focus on the inconsistency of $\hat{\theta}$ induced by specification error in the first two conditional moments. Here we concentrate on higher order moments. In Chapter 4 we established that misspecification in conditional moments of order three or more would not induce inconsistency in $\hat{\theta}$, but only inefficiency and the risk of drawing incorrect inferences. Thus by looking at the third and fourth moments we are checking whether the covariance matrix of $\hat{\theta}$ is correctly estimated, whereas by looking at fifth and higher order moments we are considering whether there is still room for improving on the efficiency of $\hat{\theta}$. 
Simple intuitive diagnostics for this purpose are provided by conditional moment restrictions of the form

$$m_e(\theta, s, r; \theta) = T^{-1} \sum_{t=1}^{T} m_{et}(\phi_t, s, r; \theta) = T^{-1} \sum_{t=1}^{T} \phi_t \left[ u_t^s e_t^r - q_t(s, r) \right],$$

-(34)

for the \((2r+s)\)-th moment, where \(m_{et}(\phi_t, s, r; \theta) = \phi_t \left[ u_t^s e_t^r - q_t(s, r) \right],\) and \(q_t(s, r) = E \left[ u_t^s e_t^r \mid \mathcal{F}_t \right],\) so that \(E \left[ m_e(\theta, s, r; \theta) \right] = 0.\) The \(n \times 1\) vector \(\phi_t\) is a measurable function of \(\mathcal{F}_t\) and it may be a function of \(\theta,\) and \(\theta = (\phi_1, \ldots, \phi_T)'.\)

The test is designed for a single moment, and using a multivariate statistic searches for power in different directions. The form of the function \(q_t(s, r)\) is given in

**Lemma 5.9.-** Under \((\mathcal{Q}0'), q_t(s, r) = E \left[ u_t^s e_t^r \mid \mathcal{F}_t \right] = \zeta(s, r) h_t^{s/2},\) where

$$\zeta(s, r) = \begin{cases} 0 & \text{for } s \text{ odd} \\ \sum_{j=0}^{r} \binom{r}{j} (-1)^r c_{j+\frac{s}{2}} & \text{for } s \text{ even} \end{cases},$$

-(35)

and \(c_r = \prod_{i=1}^{r}(2i - 1).\)

**Proof:** \(E \left[ u_t^s e_t^r \mid \mathcal{F}_t \right] = E \left[ u_t^s (u_t^2 - h_t)^r \mid \mathcal{F}_t \right] = E \left[ u_t^s \sum_{j=0}^{r} \binom{r}{j} u_t^{2j} (-h_t)^{r-j} \mid \mathcal{F}_t \right],\)

$$= \sum_{j=0}^{r} \binom{r}{j} (-h_t)^{r-j} E \left[ u_t^{s+2j} \mid \mathcal{F}_t \right],$$

and if \(s\) is odd, \(s + 2j\) is odd and the expectation is zero for all \(j,\) in view of normality. Also, if \(s\) is even \(E \left[ u_t^{s+2j} \mid \mathcal{F}_t \right] = c_{j+\frac{s}{2}} h_t^{j+\frac{s}{2}}\) and substitution yields the result.

Note that the tests in (34) are not consistency tests as in § 5.2. Because they focus on the efficiency of the QMLE we refer to them as 'efficiency tests'.

The basic statistic is the sample moment

$$m_e(\hat{\theta}, s, r; \hat{\theta}) = T^{-1} \sum_{t=1}^{T} m_{et}(\hat{\phi}_t, s, r; \hat{\theta}) = T^{-1} \sum_{t=1}^{T} \hat{\phi}_t \left[ \hat{u}_t^s \hat{e}_t^r - \hat{q}_t(s, r) \right],$$

-(36)
with all functions evaluated at the QMLE $\hat{\theta}$. These efficiency tests can also be seen as an extension of residual analysis (Pagan and Hall [1983]) for higher order moments, since they focus on the residuals $\hat{u}_t$ and $\hat{e}_t$ to detect misspecification. The variance of $m_\phi (\phi, s, r; \theta)$ is given by

$$\text{var} \left[ T^{1/2} m_\phi (\phi, s, r; \theta) \right] = \left[ \zeta (2s, 2r) - \zeta (s, r)^2 \right] E \left[ \sum_{t=1}^{T} h_t^{2r+s} \phi_t \phi'_t \right], \quad (37)$$

where we have used the fact that the $m_{et}$ are martingale differences, and

$$E \left[ m_{et} (\phi_t, s, r; \theta) m_{et} (\phi_t, s, r; \theta)' | \mathcal{F}_t \right] = [q_t (2s, 2r) - q_t (s, r)^2] \phi_t \phi'_t.$$

Note that for $s$ odd $\zeta (s, r)$ is zero, and define $V_\phi (\phi, s, r; \theta_0)$ as the limit of the covariance matrix in (37), that is

$$V_\phi (\phi, s, r; \theta_0) = \left[ \zeta (2s, 2r) - \zeta (s, r)^2 \right] \mathcal{O} \left( T^{-1} \phi' \Omega^{2r+s} \phi \right).$$

We also require the equivalent of matrix $M (\theta_0)$ in (22), which we denote $M_\phi (\phi, s, r; \theta_0)$ for the efficiency tests. Using the outer product form,

$$E \left[ m_{et} (\phi_t, s, r; \theta) d_{et}' | \mathcal{F}_t \right] = \phi_t E \left[ u_t^s \varepsilon_t^r \left( h_t^{1-\alpha} x_t u_t + \frac{1}{2} h_t^{-2} s_t \varepsilon_t \right) | \mathcal{F}_t \right]$$

$$= \begin{cases} \zeta (s+1, r) h_t^{r+(s-1)/2} \phi_t x_t' & \text{for } s \text{ odd} \\ \frac{1}{2} \zeta (s, r+1) h_t^{-1+s/2} \phi_t s_t' & \text{for } s \text{ even} \end{cases}$$

and so

$$M_\phi (\phi, s, r; \theta_0) = \begin{cases} \zeta (s+1, r) \mathcal{O} \left( T^{-1} \phi' \Omega^{r+(s-1)/2} X \right) & \text{for } s \text{ odd} \\ \frac{1}{2} \zeta (s, r+1) \mathcal{O} \left( T^{-1} \phi' \Omega^{-1+s/2} S \right) & \text{for } s \text{ even} \end{cases}, \quad (38)$$

where all the expectations are evaluated under $H_0$.

Sequences of local alternatives here are not specified in much detail as they are not linked to the arguments of the pdf but rather to the form of the pdf itself. We consider
The conditional distribution of $y_t$ is $f(y_t \mid \mathcal{F}_t, \theta_0, \lambda_T)$ where
\[ \lambda_T = \lambda_0 + T^{-1/2} \delta \text{ for fixed } \lambda_0 \text{ and } \delta. \] At $\delta = 0$ $f(\cdot \mid \cdot)$ is the pdf of the normal distribution with mean $\mu_t(\beta_0)$ and variance $h_t(\theta_0)$.

Denote $d_{\lambda_T} = \partial \log f / \partial \lambda_T$. The following Theorem is a special case of Lemma 1 of Newey [1985b].

**Theorem 5.10.** Under $(Q 0''')$ - $(Q 8)$ and the sequence $\{ \varphi_t \}, \varphi_t \in \mathcal{F}_t$, being such that the function $g^* = (m_e(\vartheta, s, r; \theta)'', d_{\vartheta}'')'$ obeys the regularity, continuity, dominance and mixing conditions in assumptions (1) - (6) of Newey [1985b], then

\[ T^{1/2} m_e(\hat{\vartheta}, s, r; \hat{\theta}) \overset{d}{\rightarrow} N(\psi_e, Q_\theta), \] 
where
\[ \psi_e = \mathcal{E}\left\{ T^{-1} \sum_{t=1}^T m_{et}(\varphi_t, s, r; \theta_0) d_{\lambda_t}(\theta_0)' \right\} \delta 
- M_e(\vartheta, s, r; \theta_0) J(\theta_0)^{-1} \mathcal{E}\left\{ T^{-1} \sum_{t=1}^T d_{\vartheta t}(\theta_0) d_{\lambda t}(\theta_0)' \right\} \delta, \]

\[ Q_\theta = V_e(\vartheta, s, r; \theta_0) - M_e(\vartheta, s, r; \theta_0) J(\theta_0)^{-1} M_e(\vartheta, s, r; \theta_0)' . \]

A consistent estimator of $Q_\theta$ is $\hat{Q}_\theta$, which substitutes all expectations with sample moments evaluated at $\hat{\theta}$.

**Proof.** The proof is identical to that of Theorem 5.3 replacing the relevant functions $V_0$ and $M_0$ of consistency tests by $V_e$ and $M_e$ of efficiency tests. 

Note that for $s$ odd the matrix $M_e$ has a $n \times (p - k)$ zero submatrix corresponding to $\alpha$ and thus its covariance matrix depends only on $V(\hat{\beta})$. The test-statistic follows in

**Corollary 5.11.** Under the assumptions of Theorem 5.10, the efficiency test-statistic is
Proof: Construct the quadratic form in (39) and substitute the consistent estimator of $Q_\theta$.

This provides a wide range of tests for a single moment. Omnibus tests can be constructed as joint moment tests. Since the joint tests are also conditional moment tests, the joint distribution of basic statistics as (36) is asymptotic normal, and we only require the covariance between an arbitrary pair to determine the complete joint asymptotic distribution of any combination of efficiency statistics. This covariance is given in

**Theorem 5.12.** If $m_\hat{\theta}(s, r; \hat{\theta})$ and $m_\theta(\hat{\theta}^*, s^*, r^*; \hat{\theta})$ are efficiency tests satisfying the conditions of Theorem 5.10, then

$$\lim_{T \to \infty} \text{cov} \left[ T^{1/2} m_\hat{\theta}(s, r; \hat{\theta}), T^{1/2} m_\theta(\hat{\theta}^*, s^*, r^*; \hat{\theta}) \right] =$$

$$\left[ \zeta(s + s^*, r + r^*) - \zeta(s, r) \right] \Omega_{r + r^* + (s + s^*)/2} \zeta(s, r)^* - M_\theta(s, r; \theta_0) V(\hat{\theta}) M_\theta(\hat{\theta}^*, s^*, r^*; \theta_0)'.$$

Proof: We need the off-diagonal component of the covariance matrix of the joint efficiency statistic. Following Theorem 5.10 (or Theorem 5.3), this component is given by

$$\lim_{T \to \infty} \text{cov} \left[ T^{1/2} m_\hat{\theta}(s, r; \theta_0), T^{1/2} m_\theta(\hat{\theta}^*, s^*, r^*; \theta_0) \right] =$$

$$- M_\theta(s, r; \theta_0) V(\hat{\theta}) M_\theta(\hat{\theta}^*, s^*, r^*; \theta_0)'.$$

Now

$$E \left[ (u_t^s \xi_t^s - q_t(s, r)) (u_t^{s^*} \xi_t^{s^*} - q_t(s^*, r^*)) \mid \mathcal{F}_t \right] = [q_t(s + s^*, r + r^*) - q_t(s, r)] q(t(s^*, r^*)) =$$

$$h_t^{r + r^* + (s + s^*)^2} \left[ \zeta(s + s^*, r + r^*) - \zeta(s, r) \zeta(s^*, r^*) \right] ,$$
using Lemma 5.9. The expected value of sums of this elements multiplied by \( \phi_t \phi_t^* \) provides the first term in the covariance and completes the proof. 

Note that the first term of this covariance is always null when one of the moments is odd and the other is even, because \( s + s^* \) is odd and this makes both \( \zeta(s + s^*, r + r^*) \) and \( \zeta(s, r) \zeta(s^*, r^*) \) equal to zero. The second term is not zero in any well defined model with non-observable mean innovations. The covariance can vanish in tests combining an odd and an even moments only when the disturbances are themselves observable (i.e. \( k = 0 \) and \( X = \partial \mu / \partial \gamma = 0 \)), because then \( M_0 (\theta, s, r; \theta_0) = 0 \) for \( s \) odd. Thus, to get an omnibus test having independent components in a regression model it is necessary that its components be linear combinations of different moments. This is used in the LM test for normality against the Pearson family in Chapter 6.

§ 5.4 Comments on some Monte Carlo evidence

§ 5.4.1 Consistency and coherency tests

We have considered eight different consistency tests in the simulation experiments with the ARCH and Poisson-N models, and all test-statistics were calculated from the coefficient of determination of the auxiliary double-length regression. Six of the test-statistics produce one-degree-of-freedom tests, and the remaining two have \( k \) degrees of freedom (\( k = 2 \) in the experiments). The direct and weighted sums of residuals provide the basic statistics for four of the univariate tests, namely \( \sum u_t \), \( \sum \hat{e}_t \), \( \sum \hat{h}_t^{-1} \hat{u}_t \), and \( \sum \hat{h}_t^{-2} \hat{e}_t \). Observe that the weighted sums of mean and variance residuals produce identical tests in the Poisson-N model because the orthogonality conditions are in this case \( X' \Omega^{-1} u + \frac{1}{2} X' \Omega^{-2} e \). The two remaining one-degree-of-freedom tests are of the RESET type (Ramsey [1969]): if the mean auxiliary regression is augmented with \( \hat{\mu}_t^2 \), from Theorem 5.6 this produces the consistency test statistic with \( \Phi_t = (h_t^{-1} \hat{\mu}_t^2, 0) \), or basic consistency statistic \( \sum h_t^{-1} \hat{\mu}_t^2 \hat{u}_t \); if the variance auxiliary
regression is augmented with $h_t^2$, this produces the consistency test-statistic with $\Phi_t = \frac{1}{2}(0, h_t^{-1})$, or basic consistency statistic $\sum_{t=1}^{T} \hat{h}_t^{-1} \hat{e}_t$. The $k$-degrees-of-freedom tests are the usual Hausman statistic based on $\hat{\beta} - \hat{\beta}_m$, or equivalently on $X^\prime \hat{u}$ as seen in (12), and the coherency test based on $\hat{\beta} - \hat{\beta}_m$, or equivalently on $X^\prime \hat{\Omega}^{-1} \hat{u}$ as seen in (14).

Tables 5.1 and 5.2 report proportion of rejections for these consistency and coherency test-statistics when tests are performed using the asymptotic distribution, by taking the 5% significance points of the $\chi^2$ distribution with the appropriate degrees of freedom. In the Poisson-N model (Table 5.1) the nominal test size mostly understates the real size obtained in the simulations. The largest estimated size is 9.2%, so that the chances of over-rejection of true hypotheses do not seem dramatic. For the ARCH model with mild and strong effects (Table 5.2) there is a more marked tendency to have a conservative test based on the asymptotic distribution, especially in the smallest sample. This effect is clearer in the test based on the sum of variance residuals, whose estimated size is always close to 1%. On the whole, we think that test-sizes are reasonable and there is no major case for concern in the use of consistency test-statistics that arises from this evidence.

To assess the power of the tests using the true (empirical) and asymptotic distributions of the test-statistics, 5% significance points were obtained from 1000 simulation replications under the null hypothesis of correct specification. The results are reported in Tables 5.3 - 5.5 for various misspecified models, with proportion of rejections using the asymptotic distribution reported in the columns headed as 'rejec.', and proportion of rejections using the simulation critical values reported in the column headed as 'power'. Each block of figures corresponds to one DGP and is headed by the source of the misspecification, with the remaining characteristics of the DGP being correctly specified. $\mu_t$ and $h_t$ denote the moments under the null hypothesis, and $\mu_t$ and $h_t$ denote the true moments of the DGP. We do not discuss an ordering of tests in terms of
power because any conclusion will be very sensitive to the specific type of
departure. Rather we are concerned with the general ability of the tests to
detect specification errors of different types. If we have a specific alternative in
mind we can design the consistency test to have the most power in such a
direction i.e. the LM test.

Table 5.3 presents the estimated powers for the Poisson-N model. Real
powers are in general smaller than rejections based on the asymptotic
distribution, in agreement with the difference in nominal and real probabilities
of type I errors. The first experiment corresponds to the misspecified mean by
the exclusion of the regressor $x_{2t}$, and it is clear that all the tests have
esentially no power against this departure. In the simulations of Chapter 4
we noted that the mean biases induced by this type of misspecifications were
very similar in the OLS, GLS and ML estimators. Several of the consistency
tests used here are transformations of differences of these estimators, and thus
inconsistencies of the same size result in no power. This suggests that the
relation between the biases of the different estimators is not just in their
means, but on a case by case basis, and highlights the possibility of inconsistent
tests even in the presence of inconsistent estimators.

The second experiment also involves a misspecified conditional mean, but
now in the form of first order autocorrelation ($\rho = 0.8$). All tests detect this
error reasonably well, with the smallest (largest) power being 0.126 (0.216) at
$T = 20$, 0.460 (0.626) at $T = 50$, 0.776 (0.900) at $T = 100$, and 0.950 (0.996) at $T = 200$.

The third experiment is concerned with variance misspecification of the
sort also analyzed in the simulations of Chapter 4: $h_{t}$ depends on the same
variables as $h_{t}$ but with parameter values different to $\beta_0$. Again power is
satisfactory in all cases with the smallest (largest) values being 0.400 (0.742) at
$T = 20$, 0.608 (1.0) at $T = 50$, 0.824 (1.0) at $T = 100$, and 0.914 (1.0) at $T = 200$. 
The last experiment with the Poisson-N model also involves variance misspecification, but now this takes the form of an Amemiya conditional variance for the DGP. Powers are good except for the test based on the sum of variance residuals which reaches only 0.178 at \( T = 200 \). Without this test, the smallest (largest) powers are 0.304 (0.504) at \( T = 20 \), 0.506 (0.804) at \( T = 50 \), 0.560 (0.948) at \( T = 100 \), and 0.776 (0.998) at \( T = 200 \).

The results for the mild and strong ARCH models are presented in Tables 5.4 and 5.5, respectively, and the differences between power and rejections using the asymptotic distribution does not cause major concern. Using Corollary 3.8 we can see that in the ARCH model \( \hat{\alpha} \), \( \hat{\alpha}_v \) and \( \hat{\alpha}_g \) have the same asymptotic distribution under correct specification, and also under misspecifications that preserve the block diagonality of the information matrix between \( \alpha \) and \( \beta \). Furthermore, even if the specification error renders the information matrix non diagonal, we still have that \( \hat{\alpha} - \hat{\alpha}_v \) and \( \hat{\alpha} - \hat{\alpha}_g \) converge to zero whenever \( \hat{\beta} - \hat{\beta}_v \) converges to zero, and therefore we would not expect tests based on differences of these estimators to have power in these conditions. The difference between \( \hat{\alpha} \) and \( \hat{\alpha}_g \) in the ARCH model is analogous to the difference between \( \hat{\beta}_m \) and \( \hat{\beta}_z \) in the mean equation (i.e. the difference of OLS and a weighted least-squares estimator as in White [1980a]), and the RESET-type test is based on the difference \( \hat{\alpha} - \tilde{\alpha} \), where \( \tilde{\alpha} = (\hat{Z}' \hat{\Omega}^{-1} \hat{Z})^{-1} \hat{Z}' \hat{\Omega}^{-1} \hat{u}^2 \) (another weighted least squares estimator), and these differences would not be too large when all \( \beta \) estimators remain consistent. Hence simple variance tests in the ARCH model are not expected to be very powerful in a wide range of situations.

The first two experiments refer to a misspecified conditional mean by first excluding a regressor, and then having autocorrelated errors. The picture is very much the same for the omitted regressor case as in the Poisson-N model and merits no further discussion. In the case of autocorrelation we have again a lack of power in the tests. This is the natural thing to expect from tests based
on mean residuals because the GLS and OLS estimators are consistency-robust to this type of misspecification and we saw in Chapter 4 that the biases in variance estimators which affect the MLE were rather small and in most cases insignificant.

The third experiment with the ARCH model considers a variance misspecification which is a conditionally even function of $u_{t-j}$, namely a Poisson-type in the presence of strongly exogenous regressors, and therefore $\hat{\beta}$ is robust to this type of misspecification and the consistency tests have no power in this situation.

Finally, the fourth experiment refers to misspecification of the conditional variance which combines even and odd effects and renders the MLE of $\beta$ inconsistent. The size of the inconsistency depends on the strength of the process, and the simulation results are in agreement with this, with a substantial increase in the power of the strong process over the mild process.

§ 5.4.2 Efficiency tests

To analyze the performance of efficiency tests we have chosen two tests for symmetry, two tests for kurtosis, and two omnibus tests in the third and fourth moments. The two tests in each direction differ essentially in the use of residuals from only one equation or mixing residuals from the two equations. The symmetry tests are based on $m_1 = \sum_{t=1}^{T} h_t^{-1} u_t \varepsilon_t$ and $m_2 = \sum_{t=1}^{T} h_t^{-2} u_t^2$, while $m_3 = \sum_{t=1}^{T} h_t^{-2} (u_t^2 \varepsilon_t - 2 h_t^2)$ and $m_4 = \sum_{t=1}^{T} h_t^{-2} (\varepsilon_t^2 - 2 h_t^2)$ provide the basic statistics for the tests of the fourth moment. The first joint test (Joint 1 in the Tables) is based on $(m_1, m_3)'$, and the second joint test (Joint 2 in the tables) is based on $(m_2, m_4)'$. Observe that the components of these omnibus tests are not asymptotically independent.
The conditional mean and conditional variance (or their quantile analogous in the Cauchy distribution) were kept correctly specified, and the misspecification appears through an erroneous form given to the conditional distribution in the model (see § 2.4). Four of the distributions considered under the alternative hypothesis are symmetric: the uniform distribution in \((-1, 1)\), the Student's t distribution with fifteen and five degrees of freedom, and the \(\beta\) distribution with parameters \(0.5\) and \(0.5\). Two other distributions considered under the alternative are asymmetric: the \(\chi^2\) distribution with two degrees of freedom, and the lognormal distribution, and a final distribution does not possess moments of any order, namely the Cauchy distribution. For this experiment only three sample sizes were simulated. Following a suggestion of Weisberg [1980] that the number of regressors might have a considerable effect on the performance of tests for normality, additional experiments were undertaken including three additional regressors, but these are not reported since we did not find any qualitative difference with the basic experiments other than a natural reduction in power because more information is being used to estimate a larger number of parameters. Tables 5.6 - 5.8 report proportion of rejections of the efficiency tests using the asymptotic distribution (5% significance points of the \(\chi^2\) distribution with the appropriate degrees of freedom), and each block of figures corresponds to one test-statistic. The headings in the columns denote sample size (T), and the conditional distribution of the DGP, as follows: normal (N), uniform (U), Student's \(t_{15}\) \(t_{15}\), Student's \(t_{5}\) \(t_{5}\), \(\chi^2\) \(\chi^2\), beta (\(\beta\)), lognormal (LN), and Cauchy (C).

The simulation results for the Poisson-N model are presented in Table 5.6. The first column (N) gives the proportion of rejections under the null hypothesis using the asymptotic distribution at the 5% level, and thus provides estimates of the size of the tests, which are reasonable except perhaps for the kurtosis tests in the smallest sample which are conservative. All tests reject clearly the Cauchy alternative from the smallest sample. The tests for the
third moment do not detect two of the symmetric distributions (U and \( \beta \)), but their power increases as leptokurtosis is introduced (\( t_{15} \) and \( t_{5} \)), and they show good power to detect the asymmetric alternatives. The tests for kurtosis have problem in detecting the uniform and \( \beta \) distributions in the smallest sample, but both reject clearly these alternatives when the sample is increased to \( T = 100 \), and as expected rejection of the \( t \) distribution is clearer as the tails become thicker. The \( \chi^2 \) and lognormal departures are well detected by these tests. The performance of the omnibus tests is in agreement with that of their components, showing good power except for the U and \( \beta \) alternatives in the smaller samples.

The results for the mild and strong ARCH models are given in Tables 5.7 and 5.8, respectively. The Cauchy alternative is not presented for either model and the lognormal alternative is not presented for the strong model. These departures proved to be strong enough to produce a large proportion of convergence failures in the estimation algorithm and the corresponding experiments were suspended. Both models show that basing the tests on the asymptotic distribution tends to produce conservative tests because the real sizes are well below the nominal 5%, and thus the powers reported (proportion of rejections using the asymptotic distribution) would be enhanced by this fact. The overall picture is similar to that in the Poisson-N model, but now both tests for kurtosis show a better detection of the uniform and \( \beta \) departures when \( T = 50 \).
**Table 5.1.** Estimated Size of Consistency and Coherency Tests in the Poisson-N Model.

\[ y_t | F_t \sim N [1 + x_{it}, 1 + x_{it}] \]

<table>
<thead>
<tr>
<th></th>
<th>T = 20</th>
<th>T = 50</th>
<th>T = 100</th>
<th>T = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma u_t )</td>
<td>0.076</td>
<td>0.058</td>
<td>0.064</td>
<td>0.056</td>
</tr>
<tr>
<td>( \Sigma u_{it}/h_t )</td>
<td>0.042</td>
<td>0.034</td>
<td>0.042</td>
<td>0.032</td>
</tr>
<tr>
<td>( \Sigma \mu^2_{it}/h_t )</td>
<td>0.076</td>
<td>0.066</td>
<td>0.066</td>
<td>0.058</td>
</tr>
<tr>
<td>( \Sigma \epsilon_t )</td>
<td>0.092</td>
<td>0.080</td>
<td>0.060</td>
<td>0.060</td>
</tr>
<tr>
<td>( \Sigma \epsilon_{it}/h_t^2 )</td>
<td>0.042</td>
<td>0.034</td>
<td>0.042</td>
<td>0.032</td>
</tr>
<tr>
<td>( \Sigma \epsilon_{it}/h_t )</td>
<td>0.092</td>
<td>0.072</td>
<td>0.066</td>
<td>0.060</td>
</tr>
<tr>
<td>( \Sigma x_{it} u_t )</td>
<td>0.056</td>
<td>0.058</td>
<td>0.060</td>
<td>0.064</td>
</tr>
<tr>
<td>( \Sigma x_{it} u_{it}/h_t )</td>
<td>0.052</td>
<td>0.044</td>
<td>0.052</td>
<td>0.046</td>
</tr>
</tbody>
</table>

**Table 5.2.** Estimated Size of Consistency and Coherency Tests in the ARCH Model.

\[ y_t | F_t \sim N [1 + x_{it}, 0.8 + 0.2 u_{it}^2] \]

\[ y_t | F_t \sim N [1 + x_{it}, 0.2 + 0.8 u_{it}^2] \]

<table>
<thead>
<tr>
<th></th>
<th>T = 20</th>
<th>T = 50</th>
<th>T = 100</th>
<th>T = 200</th>
<th>T = 20</th>
<th>T = 50</th>
<th>T = 100</th>
<th>T = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma u_t )</td>
<td>0.020</td>
<td>0.036</td>
<td>0.032</td>
<td>0.040</td>
<td>0.026</td>
<td>0.032</td>
<td>0.034</td>
<td>0.042</td>
</tr>
<tr>
<td>( \Sigma u_{it}/h_t )</td>
<td>0.008</td>
<td>0.028</td>
<td>0.042</td>
<td>0.042</td>
<td>0.030</td>
<td>0.054</td>
<td>0.054</td>
<td>0.040</td>
</tr>
<tr>
<td>( \Sigma \mu^2_{it}/h_t )</td>
<td>0.012</td>
<td>0.042</td>
<td>0.044</td>
<td>0.050</td>
<td>0.022</td>
<td>0.060</td>
<td>0.058</td>
<td>0.048</td>
</tr>
<tr>
<td>( \Sigma \epsilon_t )</td>
<td>0.010</td>
<td>0.012</td>
<td>0.016</td>
<td>0.012</td>
<td>0.012</td>
<td>0.008</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>( \Sigma \epsilon_{it}/h_t^2 )</td>
<td>0.062</td>
<td>0.070</td>
<td>0.052</td>
<td>0.071</td>
<td>0.056</td>
<td>0.065</td>
<td>0.049</td>
<td>0.046</td>
</tr>
<tr>
<td>( \Sigma \epsilon_{it}/h_t )</td>
<td>0.054</td>
<td>0.048</td>
<td>0.062</td>
<td>0.042</td>
<td>0.042</td>
<td>0.050</td>
<td>0.050</td>
<td>0.060</td>
</tr>
<tr>
<td>( \Sigma x_{it} \epsilon_t )</td>
<td>0.016</td>
<td>0.032</td>
<td>0.034</td>
<td>0.052</td>
<td>0.020</td>
<td>0.026</td>
<td>0.054</td>
<td>0.034</td>
</tr>
<tr>
<td>( \Sigma x_{it} \epsilon_{it}/h_t )</td>
<td>0.018</td>
<td>0.032</td>
<td>0.034</td>
<td>0.052</td>
<td>0.052</td>
<td>0.054</td>
<td>0.058</td>
<td>0.038</td>
</tr>
</tbody>
</table>
TABLE 5.3 - REJECTION FREQUENCIES OF CONSISTENCY AND COHERENCY TESTS IN THE POISSON-N MODEL.

<table>
<thead>
<tr>
<th></th>
<th>( T = 20 )</th>
<th>( T = 50 )</th>
<th>( T = 100 )</th>
<th>( T = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rejec. power</td>
<td>rejec. power</td>
<td>rejec. power</td>
<td>rejec. power</td>
</tr>
</tbody>
</table>

\( \mu_t = \mu_t + x_{2t} \)

- \( \Sigma u_t \)
- \( \Sigma u_t/b_t \)
- \( \Sigma \mu^2 u_t/b_t \)
- \( \Sigma c_t \)
- \( \Sigma c_t/b_t^2 \)
- \( \Sigma x_t u_t \)
- \( \Sigma x_t u_t/b_t \)

\( \mu_t = \mu_t + 0.8 u_{t-1} \)

- \( \Sigma u_t \)
- \( \Sigma u_t/b_t \)
- \( \Sigma \mu^2 u_t/b_t \)
- \( \Sigma c_t \)
- \( \Sigma c_t/b_t^2 \)
- \( \Sigma x_t u_t \)
- \( \Sigma x_t u_t/b_t \)

\( h_t = 0.5 + 0.7 x_{1t} \)

- \( \Sigma u_t \)
- \( \Sigma u_t/b_t \)
- \( \Sigma \mu^2 u_t/b_t \)
- \( \Sigma c_t \)
- \( \Sigma c_t/b_t^2 \)
- \( \Sigma x_t u_t \)
- \( \Sigma x_t u_t/b_t \)

\( h_t = (1 + x_{1t})^2 \)

- \( \Sigma u_t \)
- \( \Sigma u_t/b_t \)
- \( \Sigma \mu^2 u_t/b_t \)
- \( \Sigma c_t \)
- \( \Sigma c_t/b_t^2 \)
- \( \Sigma x_t u_t \)
- \( \Sigma x_t u_t/b_t \)

[147]
<table>
<thead>
<tr>
<th>( \mu_t = \mu_t + x_{2t} )</th>
<th>( \mu_t = \mu_t + 0.8 u_{t-1} )</th>
<th>( \bar{h}<em>t = 1 + x</em>{1t} )</th>
<th>( \bar{h}<em>t = 0.8 + 0.2 y</em>{t-1}^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma u_t )</td>
<td>( \Sigma u_t )</td>
<td>( \Sigma u_t )</td>
<td>( \Sigma u_t )</td>
</tr>
<tr>
<td>0.010</td>
<td>0.052</td>
<td>0.012</td>
<td>0.014</td>
</tr>
<tr>
<td>0.036</td>
<td>0.104</td>
<td>0.040</td>
<td>0.060</td>
</tr>
<tr>
<td>0.026</td>
<td>0.080</td>
<td>0.020</td>
<td>0.012</td>
</tr>
<tr>
<td>0.038</td>
<td>0.118</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>( \Sigma u_t )</td>
<td>( \Sigma u_t / h_t )</td>
<td>( \Sigma u_t / h_t )</td>
<td>( \Sigma u_t / h_t )</td>
</tr>
<tr>
<td>0.016</td>
<td>0.008</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>0.058</td>
<td>0.032</td>
<td>0.042</td>
<td>0.048</td>
</tr>
<tr>
<td>0.032</td>
<td>0.036</td>
<td>0.022</td>
<td>0.014</td>
</tr>
<tr>
<td>0.052</td>
<td>0.060</td>
<td>0.024</td>
<td>0.016</td>
</tr>
<tr>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
</tr>
<tr>
<td>0.012</td>
<td>0.020</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>0.044</td>
<td>0.040</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>0.034</td>
<td>0.060</td>
<td>0.024</td>
<td>0.016</td>
</tr>
<tr>
<td>0.050</td>
<td>0.090</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
<td>( \Sigma \mu_t^2 / h_t )</td>
</tr>
<tr>
<td>0.008</td>
<td>0.020</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>0.036</td>
<td>0.042</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>0.036</td>
<td>0.060</td>
<td>0.024</td>
<td>0.016</td>
</tr>
<tr>
<td>0.060</td>
<td>0.090</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>( \Sigma \epsilon_t )</td>
<td>( \Sigma \epsilon_t / h_t )</td>
<td>( \Sigma \epsilon_t / h_t )</td>
<td>( \Sigma \epsilon_t / h_t )</td>
</tr>
<tr>
<td>0.003</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>0.042</td>
<td>0.040</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>0.044</td>
<td>0.060</td>
<td>0.024</td>
<td>0.016</td>
</tr>
<tr>
<td>0.060</td>
<td>0.090</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>( \Sigma \epsilon_t / h_t^2 )</td>
<td>( \Sigma \epsilon_t / h_t^2 )</td>
<td>( \Sigma \epsilon_t / h_t^2 )</td>
<td>( \Sigma \epsilon_t / h_t^2 )</td>
</tr>
<tr>
<td>0.066</td>
<td>0.066</td>
<td>0.066</td>
<td>0.066</td>
</tr>
<tr>
<td>0.042</td>
<td>0.040</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>0.064</td>
<td>0.064</td>
<td>0.064</td>
<td>0.064</td>
</tr>
<tr>
<td>( \Sigma x_t u_t )</td>
<td>( \Sigma x_t u_t )</td>
<td>( \Sigma x_t u_t )</td>
<td>( \Sigma x_t u_t )</td>
</tr>
<tr>
<td>0.012</td>
<td>0.026</td>
<td>0.006</td>
<td>0.016</td>
</tr>
<tr>
<td>0.048</td>
<td>0.094</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>0.030</td>
<td>0.058</td>
<td>0.016</td>
<td>0.020</td>
</tr>
<tr>
<td>0.058</td>
<td>0.084</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>( \Sigma x_t u_t / h_t )</td>
<td>( \Sigma x_t u_t / h_t )</td>
<td>( \Sigma x_t u_t / h_t )</td>
<td>( \Sigma x_t u_t / h_t )</td>
</tr>
<tr>
<td>0.020</td>
<td>0.022</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>0.044</td>
<td>0.050</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>0.034</td>
<td>0.060</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>0.058</td>
<td>0.084</td>
<td>0.054</td>
<td>0.054</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( T = 20 )</th>
<th>( T = 50 )</th>
<th>( T = 100 )</th>
<th>( T = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
</tr>
<tr>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
</tr>
<tr>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
</tr>
<tr>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
<td>( \text{rejec. power} )</td>
</tr>
</tbody>
</table>

**TABLE 5.4:** REJECTION FREQUENCIES OF CONSISTENCY AND COHERENCY TESTS IN THE ARCH MODEL.
<table>
<thead>
<tr>
<th></th>
<th>T = 20</th>
<th>T = 50</th>
<th>T = 100</th>
<th>T = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rejec.</td>
<td>power</td>
<td>rejec.</td>
<td>power</td>
</tr>
<tr>
<td>$\mu_t = \mu_t + x_{2t}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Sigma u_t$</td>
<td>0.008</td>
<td>0.038</td>
<td>0.024</td>
<td>0.032</td>
</tr>
<tr>
<td>$\Sigma u_t/h_t$</td>
<td>0.026</td>
<td>0.068</td>
<td>0.032</td>
<td>0.026</td>
</tr>
<tr>
<td>$\Sigma \mu^{2}u/h_t$</td>
<td>0.020</td>
<td>0.036</td>
<td>0.042</td>
<td>0.030</td>
</tr>
<tr>
<td>$\Sigma e_t$</td>
<td>0.006</td>
<td>0.042</td>
<td>0.002</td>
<td>0.040</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t^2$</td>
<td>0.060</td>
<td>0.052</td>
<td>0.051</td>
<td>0.041</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t$</td>
<td>0.054</td>
<td>0.036</td>
<td>0.054</td>
<td>0.056</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t$</td>
<td>0.018</td>
<td>0.036</td>
<td>0.020</td>
<td>0.032</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t/h_t$</td>
<td>0.046</td>
<td>0.046</td>
<td>0.052</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_t = \mu_t + 0.8 u_{t-1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Sigma u_t$</td>
<td>0.036</td>
<td>0.094</td>
<td>0.064</td>
<td>0.088</td>
</tr>
<tr>
<td>$\Sigma u_t/h_t$</td>
<td>0.020</td>
<td>0.040</td>
<td>0.038</td>
<td>0.034</td>
</tr>
<tr>
<td>$\Sigma \mu^{2}u/h_t$</td>
<td>0.034</td>
<td>0.050</td>
<td>0.038</td>
<td>0.028</td>
</tr>
<tr>
<td>$\Sigma e_t$</td>
<td>0.010</td>
<td>0.054</td>
<td>0.008</td>
<td>0.034</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t^2$</td>
<td>0.086</td>
<td>0.071</td>
<td>0.066</td>
<td>0.066</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t$</td>
<td>0.066</td>
<td>0.042</td>
<td>0.050</td>
<td>0.060</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t$</td>
<td>0.024</td>
<td>0.052</td>
<td>0.036</td>
<td>0.048</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t/h_t$</td>
<td>0.038</td>
<td>0.036</td>
<td>0.052</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_t = 1 + x_{1t}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Sigma u_t$</td>
<td>0.012</td>
<td>0.028</td>
<td>0.016</td>
<td>0.030</td>
</tr>
<tr>
<td>$\Sigma u_t/h_t$</td>
<td>0.006</td>
<td>0.026</td>
<td>0.028</td>
<td>0.023</td>
</tr>
<tr>
<td>$\Sigma \mu^{2}u/h_t$</td>
<td>0.014</td>
<td>0.014</td>
<td>0.020</td>
<td>0.012</td>
</tr>
<tr>
<td>$\Sigma e_t$</td>
<td>0.012</td>
<td>0.054</td>
<td>0.010</td>
<td>0.038</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t^2$</td>
<td>0.094</td>
<td>0.074</td>
<td>0.073</td>
<td>0.057</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t$</td>
<td>0.070</td>
<td>0.048</td>
<td>0.080</td>
<td>0.092</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t$</td>
<td>0.006</td>
<td>0.018</td>
<td>0.012</td>
<td>0.024</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t/h_t$</td>
<td>0.022</td>
<td>0.016</td>
<td>0.042</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_t = 0.2 + 0.8 y_{t-1}^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Sigma u_t$</td>
<td>0.020</td>
<td>0.060</td>
<td>0.058</td>
<td>0.088</td>
</tr>
<tr>
<td>$\Sigma u_t/h_t$</td>
<td>0.082</td>
<td>0.198</td>
<td>0.456</td>
<td>0.446</td>
</tr>
<tr>
<td>$\Sigma \mu^{2}u/h_t$</td>
<td>0.148</td>
<td>0.192</td>
<td>0.586</td>
<td>0.532</td>
</tr>
<tr>
<td>$\Sigma e_t$</td>
<td>0.004</td>
<td>0.036</td>
<td>0.008</td>
<td>0.050</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t^2$</td>
<td>0.060</td>
<td>0.043</td>
<td>0.047</td>
<td>0.041</td>
</tr>
<tr>
<td>$\Sigma e_t/h_t$</td>
<td>0.060</td>
<td>0.046</td>
<td>0.040</td>
<td>0.048</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t$</td>
<td>0.030</td>
<td>0.070</td>
<td>0.074</td>
<td>0.090</td>
</tr>
<tr>
<td>$\Sigma x_{t}u_t/h_t$</td>
<td>0.078</td>
<td>0.072</td>
<td>0.470</td>
<td>0.424</td>
</tr>
</tbody>
</table>
### Table 5.6 - Rejection Frequencies of Efficiency Tests in the Poisson-N Model

<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>U</th>
<th>t₁₅</th>
<th>t₅</th>
<th>β</th>
<th>χ²</th>
<th>LN</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Σ u₁e/hₜ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.054</td>
<td>0.000</td>
<td>0.096</td>
<td>0.286</td>
<td>0.000</td>
<td>0.490</td>
<td>0.188</td>
<td>0.892</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.000</td>
<td>0.160</td>
<td>0.448</td>
<td>0.000</td>
<td>0.940</td>
<td>0.960</td>
<td>0.994</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.000</td>
<td>0.178</td>
<td>0.568</td>
<td>0.000</td>
<td>1.000</td>
<td>0.860</td>
<td>0.994</td>
</tr>
<tr>
<td>Σ uᵢ²hₜ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.058</td>
<td>0.000</td>
<td>0.102</td>
<td>0.270</td>
<td>0.000</td>
<td>0.334</td>
<td>0.146</td>
<td>0.896</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.000</td>
<td>0.156</td>
<td>0.444</td>
<td>0.000</td>
<td>0.816</td>
<td>0.384</td>
<td>0.994</td>
</tr>
<tr>
<td>100</td>
<td>0.060</td>
<td>0.022</td>
<td>0.188</td>
<td>0.568</td>
<td>0.000</td>
<td>0.922</td>
<td>0.624</td>
<td>0.992</td>
</tr>
<tr>
<td>Σ(uᵢ² - 2hᵢ²)yᵢ²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.018</td>
<td>0.004</td>
<td>0.072</td>
<td>0.272</td>
<td>0.000</td>
<td>0.454</td>
<td>0.374</td>
<td>0.982</td>
</tr>
<tr>
<td>50</td>
<td>0.036</td>
<td>0.414</td>
<td>0.159</td>
<td>0.555</td>
<td>0.068</td>
<td>0.902</td>
<td>0.910</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td>0.058</td>
<td>0.976</td>
<td>0.264</td>
<td>0.826</td>
<td>0.998</td>
<td>1.000</td>
<td>0.966</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table 5.7 - Rejection Frequencies of Efficiency Tests in the ARCH Model I

<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>U</th>
<th>t₁₅</th>
<th>t₅</th>
<th>β</th>
<th>χ²</th>
<th>LN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Σ u₁e/hₜ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.028</td>
<td>0.000</td>
<td>0.018</td>
<td>0.078</td>
<td>0.000</td>
<td>0.420</td>
<td>0.194</td>
</tr>
<tr>
<td>50</td>
<td>0.022</td>
<td>0.000</td>
<td>0.056</td>
<td>0.182</td>
<td>0.002</td>
<td>0.904</td>
<td>0.774</td>
</tr>
<tr>
<td>100</td>
<td>0.026</td>
<td>0.000</td>
<td>0.084</td>
<td>0.282</td>
<td>0.000</td>
<td>0.992</td>
<td>0.990</td>
</tr>
<tr>
<td>Σ uᵢ²hₜ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.030</td>
<td>0.000</td>
<td>0.013</td>
<td>0.078</td>
<td>0.000</td>
<td>0.406</td>
<td>0.178</td>
</tr>
<tr>
<td>50</td>
<td>0.024</td>
<td>0.000</td>
<td>0.054</td>
<td>0.170</td>
<td>0.002</td>
<td>0.852</td>
<td>0.648</td>
</tr>
<tr>
<td>100</td>
<td>0.026</td>
<td>0.000</td>
<td>0.076</td>
<td>0.272</td>
<td>0.000</td>
<td>0.978</td>
<td>0.976</td>
</tr>
<tr>
<td>Σ(uᵢ² - 2hᵢ²)yᵢ²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.010</td>
<td>0.000</td>
<td>0.012</td>
<td>0.052</td>
<td>0.000</td>
<td>0.318</td>
<td>0.098</td>
</tr>
<tr>
<td>50</td>
<td>0.014</td>
<td>0.234</td>
<td>0.056</td>
<td>0.268</td>
<td>0.742</td>
<td>0.844</td>
<td>0.500</td>
</tr>
<tr>
<td>100</td>
<td>0.016</td>
<td>0.948</td>
<td>0.122</td>
<td>0.554</td>
<td>1.000</td>
<td>0.982</td>
<td>0.844</td>
</tr>
<tr>
<td>Σ(eᵢ² - 2hᵢ²)yᵢ²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.010</td>
<td>0.000</td>
<td>0.012</td>
<td>0.054</td>
<td>0.000</td>
<td>0.312</td>
<td>0.096</td>
</tr>
<tr>
<td>50</td>
<td>0.010</td>
<td>0.162</td>
<td>0.058</td>
<td>0.270</td>
<td>0.702</td>
<td>0.842</td>
<td>0.496</td>
</tr>
<tr>
<td>100</td>
<td>0.016</td>
<td>0.944</td>
<td>0.122</td>
<td>0.562</td>
<td>1.000</td>
<td>0.984</td>
<td>0.840</td>
</tr>
</tbody>
</table>

**Notes:**
- Table 5.6 and 5.7 present rejection frequencies of efficiency tests for different models.
- The tables include columns for T, N, U, t₁₅, t₅, β, χ², and LN.
- The data is presented in a tabular format, showing frequencies for different time periods and conditions.
<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>N</th>
<th>U</th>
<th>t₁₅</th>
<th>t₅</th>
<th>β</th>
<th>χ²₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Σ u₁ε₁/h₁</td>
<td>20</td>
<td>0.020</td>
<td>0.002</td>
<td>0.030</td>
<td>0.058</td>
<td>0.004</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.006</td>
<td>0.000</td>
<td>0.024</td>
<td>0.066</td>
<td>0.002</td>
<td>0.384</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.018</td>
<td>0.000</td>
<td>0.046</td>
<td>0.130</td>
<td>0.000</td>
<td>0.654</td>
</tr>
<tr>
<td>Σ u₁³/h₁</td>
<td>20</td>
<td>0.020</td>
<td>0.002</td>
<td>0.028</td>
<td>0.058</td>
<td>0.002</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.010</td>
<td>0.000</td>
<td>0.022</td>
<td>0.048</td>
<td>0.002</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.012</td>
<td>0.000</td>
<td>0.040</td>
<td>0.112</td>
<td>0.002</td>
<td>0.448</td>
</tr>
<tr>
<td>Σ( u₁²ε₁-2h₁²ε₁²)²</td>
<td>20</td>
<td>0.010</td>
<td>0.000</td>
<td>0.034</td>
<td>0.066</td>
<td>0.000</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.008</td>
<td>0.134</td>
<td>0.044</td>
<td>0.216</td>
<td>0.692</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.016</td>
<td>0.920</td>
<td>0.100</td>
<td>0.510</td>
<td>1.000</td>
<td>0.954</td>
</tr>
<tr>
<td>Σ(ε₁²-2h₁²ε₁²)²</td>
<td>20</td>
<td>0.010</td>
<td>0.000</td>
<td>0.034</td>
<td>0.064</td>
<td>0.000</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.008</td>
<td>0.108</td>
<td>0.044</td>
<td>0.216</td>
<td>0.644</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.016</td>
<td>0.920</td>
<td>0.100</td>
<td>0.508</td>
<td>1.000</td>
<td>0.954</td>
</tr>
<tr>
<td>Joint 1</td>
<td>20</td>
<td>0.018</td>
<td>0.002</td>
<td>0.038</td>
<td>0.072</td>
<td>0.000</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.008</td>
<td>0.002</td>
<td>0.044</td>
<td>0.192</td>
<td>0.082</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.014</td>
<td>0.416</td>
<td>0.100</td>
<td>0.478</td>
<td>0.996</td>
<td>0.972</td>
</tr>
<tr>
<td>Joint 2</td>
<td>20</td>
<td>0.016</td>
<td>0.002</td>
<td>0.036</td>
<td>0.072</td>
<td>0.000</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.008</td>
<td>0.000</td>
<td>0.042</td>
<td>0.190</td>
<td>0.018</td>
<td>0.762</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.014</td>
<td>0.412</td>
<td>0.094</td>
<td>0.464</td>
<td>0.996</td>
<td>0.960</td>
</tr>
</tbody>
</table>
TESTING THE MODEL AGAINST SPECIFIC ALTERNATIVES

The tests of Chapter 5 are designed without a specific alternative in mind and therefore do not require information external to the model. However if information is available on the likely form of specification error, using this information will enhance power in the desirable directions. We consider such tests in this Chapter. The null hypothesis is $$H_0: y_t | \mathcal{F}_t \sim N[\mu_t(\beta_0), h_t(\theta_0)]$$, and we use the Lagrange multiplier (LM) principle (Breusch and Pagan [1980], Engle [1982b,84]) as the basic tool, presuming that we are testing the model for misspecification in different directions and we are not in principle interested in estimating any of the alternative models. The LM principle has the advantage of requiring estimation only under the null hypothesis and thus is in general more attractive as a diagnostic.

In section § 6.1 we derive the LM test for variable additions (Pagan [1984a]) in either conditional moment and show that it is a member of the class of consistency tests described in § 5.2. Thus the distribution of the test-statistic is immediate and the results of Engle [1982b,1984] are extended to the case of a non block-diagonal covariance matrix between $\beta$ and $\alpha$. The LM test for variable additions constitutes a powerful tool for testing many alternatives in either or both conditional moments. It includes misspecified functional forms, dynamics, coefficient stability, etc. Some of these deserve more careful analysis.

(1) In this Chapter section § 6.1 is based on joint work with A. R. Pagan, reported in Pagan and Sabau [1987b].
and are considered in sections § 6.2 to § 6.4. In section § 6.5 the LM test for a non-normal conditional distribution is derived. The alternative is taken to belong to the Pearson family as in Jarque and Bera [1980] and Bera and Jarque [1982], and concentrates on the third and fourth moments. This test is not a consistency test but rather it belongs to the class of efficiency tests described in § 5.3. Finally, some Monte Carlo evidence is produced in section § 6.6.

§ 6.1 The LM test for variable addition

Suppose we want to test the maintained hypothesis in (1) against the alternative hypothesis

\[(\Omega^0-a) \quad y_t \mid \mathcal{F}_t \sim N [\mu_t(\beta, \beta_A), h_t(\theta, \theta_A)], \quad - (2)\]

where \(\beta_A\) and \(\theta_A\) represent the additional parametric dimensions in the conditional moments. When the conditional variance depends on present and lagged values of the conditional mean, as in the Poisson, Amemiya and ARCH models, a natural consequence of modelling the second moment as a function of the first is to spread any error in the specification of \(\mu_t\) to \(h_t\), so that \(\theta_A\) includes \(\beta_A\) and hence \(\theta_A\) may be partitioned as \(\theta_A = (\beta_A', \alpha_A')'\). It is clear from \(d_\theta(\theta)\) in (2.11) that the subvector of the score for \(\theta_A\) is

\[d_A(\theta, \theta_A) = T^{-1} \sum_{t=1}^{T} h_t(\theta, \theta_A)^{-1} x_{At}^* u_{At} + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t(\theta, \theta_A)^{-2} s_{At}^* e_{At},\]

where \(u_{At} = y_t - \mu_t(\beta, \beta_A), \quad x_{At}^* = \frac{\partial \mu_t}{\partial \theta_A} = (\frac{\partial \mu_t}{\partial \beta_A'}, 0)' = (x_{At}', 0)', \quad s_{At}^* = \frac{\partial h_t(\theta, \theta_A)}{\partial \theta_A}, \quad e_{At} = u_{At} - h_t(\theta, \theta_A)\). Then under \(H_0: \theta_A = 0\),

\[d_A(\theta) = T^{-1} \sum_{t=1}^{T} h_t(\theta)^{-1} x_{At} u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} h_t(\theta)^{-2} s_{At} e_t,\]

where \(x_{At} = x_{At}^*\) and \(s_{At} = s_{At}^*\) are evaluated at \(\theta_A = 0\). The LM test for \(H_0\) is based upon (3) and it is a particular case of the general consistency test in (5.9) by making \(\Phi_t = (h_t^{-1} x_{At}, \frac{1}{2} h_t^{-2} s_{At})\). Thus denoting \(G_A = \frac{\partial g(\theta, \theta_A)}{\partial \theta_A}\) we have

**Theorem 6.1.** Under the assumptions of Theorem 5.3 the LM test for
\( H_0: \theta_A = 0 \) (i.e. \((C_0)^{-1}\) and \((C_1) - (C_8)\)) against the alternative \( H_1: \theta_A \neq 0 \)
(i.e. \((C_0-a)\) and \((C_1) - (C_8)\)), is given by

\[
slm = \hat{\Sigma}^{-1} G_A [ \hat{G}_A' \hat{\Sigma}^{-1/2} \hat{m} \hat{\Sigma}^{-1/2} \hat{G}_A ]^{-1} \hat{G}_A' \hat{\Sigma}^{-1} \hat{v},
\]

and its distribution under parametric local alternatives in the direction of \( H_1 \) is

\[
slm \overset{d}{\rightarrow} \chi^2(\mbox{n}; \lambda_{LM}^2 = \delta' \mathcal{C} \left( T^{-1} G_A' \Sigma^{-1/2} m \Sigma^{-1/2} G_A \right) \delta),
\]

where \( n = \dim(\theta_A) \), and \( m \) is the projection matrix onto the space orthogonal to \( \Sigma^{-1/2} G \). An asymptotically equivalent statistic is given by \( s_* = 2 T R_0^2 \) from the regression of \( \hat{v} \) on \( \hat{G} \) and \( \hat{G}_A \) in the metric of \( \hat{\Sigma} \), with all matrices evaluated at the MLE under \( H_0 \).

\textbf{Proof}: The subvector of the score for \( \theta_A \) in (2) is \( T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta_A} \hat{v}_t = T^{-1} G_A' \Sigma^{-1} \hat{v} \) under \( H_0 \). Therefore, set \( \Phi = \Sigma^{-1} G_A \) in Theorem 5.6 to establish the asymptotic distribution of \( s_{LM} \) and the asymptotic equivalence of \( s_{LM}^* \). For local parametric alternatives in the direction of \( H_1 \) we have \( G_\lambda = G_A \), and substituting this in (5.24) we get \( Q_\Phi = \mathcal{C} \left( T^{-1} G_A' \Sigma^{-1/2} m \Sigma^{-1/2} G_A \right) \), and substituting in (5.28) we obtain \( \psi = \mathcal{C} \left( T^{-1} G_A' \Sigma^{-1/2} m \Sigma^{-1/2} G_A \right) \delta \). Therefore, \( \lambda_{LM}^2 = \psi' Q_\Phi^{-1} \psi = \delta' \mathcal{C} \left( T^{-1} G_A' \Sigma^{-1/2} m \Sigma^{-1/2} G_A \right) \delta \).

The \( R_0^2 \) construct of the Theorem requires a double-length auxiliary regression to incorporate the information about \( \beta \) in the variance and to allow for a possible non-diagonal information matrix between \( \beta \) and \( \alpha \), a case explicitly excluded from Engle's [1982b] Theorem 1. The LM statistic and its \( 2 T R_0^2 \) version use directly the residuals and the specific form of the information matrix, and therefore introduce more structure into the testing procedure in heteroskedastic models than the also asymptotically equivalent calculations based on \( T R_0^2 \) from the auxiliary regression of unity on the score (e.g. Engle et al [1987]).

It follows from the discussion in § 5.2.2 that in simple heteroskedastic
models $T R_0^2$ from single-length auxiliary regressions provides asymptotically equivalent tests for a single moment. This fact underlies the construction of LM tests for heteroskedasticity (Breusch and Pagan [1979], Engle [1982a,1983]). In the ARCH class equivalent tests on the conditional variance can be obtained from a single length auxiliary regression with $\hat{e}_t$ as the dependent variable, provided $\partial g_t / \partial \theta_A$ is a conditionally even function of $u_{t-1}$ (e.g. Kraft and Engle [1982]). Single-length tests for the conditional mean have the correct size if constructed with GLS residuals (Theorem 1 of Engle [1982b]), but they are not the LM tests in this context because they ignore the information about $\beta$ in the conditional variance.

The tests of Theorem 6.1 have the usual optimal properties of LM, LR and Wald tests and therefore to evaluate the power of a consistency test in a specific direction all that needs to be done is to assess how well the relevant $\Phi$ projects onto the space spanned by $\Omega^{-1} G_A$. For example, in the linear ARCH(1) model of (5.30) the LM test against the alternative $h_t^*$ in (5.31) can be obtained as $T R_0^2$ from the regression of $\hat{e}_t$ on $(\hat{z}_t', \hat{z}_{at}')$ in the metric of $2 (\hat{z}_t' \hat{\alpha})^2$ and has NCP

$$\lambda_{LM}^2 = \delta_1^2 \beta' V(\hat{\beta}_v)^{-1} [ V(\hat{\beta}_v) - V(\hat{\beta}) ] V(\hat{\beta}_v)^{-1} \beta + \delta_2^2 \{ c_2 - d' V(\hat{\alpha}) d \}, \quad (5)$$

where $c_2 = \delta \{ T^{-1} \sum_{t=1}^{T} h_t^2 (x_{t-1}' \beta)^2 \}$, and $d = \delta \{ T^{-1} \sum_{t=1}^{T} h_t^2 z_t (x_{t-1}' \beta)^2 \}$, as used in (5.33). The first term comes from the odd misspecification direction and the second term comes from the even misspecification direction. It is now apparent that the power of the LM test that comes from the odd direction contrast sharply with that of the mean consistency test $s_\mu$ given in (5.33a): when the conditional variance provides most of the information (so that $V(\hat{\beta}_v)$ approaches $V(\hat{\beta})$), the power of $s_\mu$ increases to its maximum while the first term in (5) vanishes. In fact, the $\Phi$ matrix for this simple mean test has a null projection on the space spanned by $\Sigma^{-1} G_A$ and thus its power may be far from optimal, though it has the attraction of a wide range. For the variance test, the
projection of $\Phi$ onto the space spanned by $\Sigma^{-1}G_A$ is that of a vector of ones onto the space spanned by the vector with typical element $h_t^2(x_{t-1}\beta)^2$, and has a much closer relation with the power of the LM test arising from the even misspecification direction.

In the sections to follow we refer to the $2T R_0^2$ version as the LM test.

§ 6.2 LM tests for autocorrelation and dynamics

§ 6.2.1 Autocorrelation in the conditional mean

Diebold [1986b] notes that tests based on the autocorrelation function (Bartlett's confidence bounds as well as the Box-Pierce [1970] and Ljung-Box [1978] statistics) over-reject the null hypothesis of no autocorrelation in the presence of heteroskedasticity of the ARCH type. He puts forward asymptotic corrections for these statistics and assesses the size of the tests by means of a Monte Carlo experiment. Domowitz and Hakkio [1985] test for autocorrelation in the presence of an ARCH variance using a heteroskedasticity robust version of Godfrey's [1978] LM test, referring to a previous paper (Domowitz and Hakkio [1983]). Rather than looking for a heteroskedasticity-robust test, the spirit of this Thesis is to use the information in the variance. Thus we look here at testing (1) against the alternative

\[ H_1 : y_t \mid \mathcal{F}_t \sim N \left[ \mu_t(\beta) + \sum_{j=1}^{n} \rho_j v_{t-j}, h_t(\theta, \rho) \right], \quad (6) \]

where $\rho = (\rho_1, ..., \rho_n)'$, and $v_t = y_t - \mu_t(\beta)$, while we retain $u_t = y_t - E[y_t \mid \mathcal{F}_t]$ to denote the innovations. A more familiar expression of $H_1$ is in regression form,

\[ y_t = \mu_t(\beta) + v_t, \quad \text{and} \quad v_t = \sum_{j=1}^{n} \rho_j v_{t-j} + u_t. \]
Let $\bar{\mu}_t$ denote the true conditional mean, $\bar{\mu}_t = \mu_t + \sum_{j=1}^{n} \rho_j \nu_{t-j} = \mu_t + \rho_{\gamma_{t-1}}$, where $\gamma_{t-1} = (\nu_{t-1}, ..., \nu_{t-n})'$. Thus under $H_0$, $\nu_t = \nu_t$ and $\mu_t = \bar{\mu}_t$.

Clearly (6) is a particular case of (2), with the addition of $\gamma_{t-1}$ to the conditional mean and thus we require

$$\frac{\partial \bar{\mu}_t}{\partial \rho} = \gamma_{t-1} = \nu_{t-1},$$  \hspace{1cm} (7)

and

$$\frac{\partial h_t}{\partial \rho} = \sum_{i \geq 0} \frac{\partial h_t}{\partial \mu_{t-i}} \cdot \frac{\partial \mu_{t-i}}{\partial \rho} = \sum_{i \geq 0} \frac{\partial h_t}{\partial \mu_{t-i}} \nu_{t-i-1},$$ \hspace{1cm} (8)

where both equalities are valid only under $H_0$, and $\gamma_{t-1} = (\nu_{t-1}, ..., \nu_{t-n})'$. The testing procedure for autocorrelation is given in

**Corollary 6.2.** Under the assumptions of Theorem 6.1 the LM for n-th order autocorrelation in the mean equation is given by $s_{mac} = 2 T R_0^2$, where $R_0^2$ is the uncentered coefficient of determination of the regression of

$$\left( \hat{\nu}_t \right) \text{ on } \begin{pmatrix} \hat{x}_t' & 0 & \hat{u}_{t-1} & \ldots & \hat{u}_{t-n} \\ \hat{w}_t' & \mathbb{Z}_t' & \sum_{i \geq 0} \frac{\partial h_t}{\partial \mu_{t-i}} \hat{u}_{t-i-1} & \ldots & \sum_{i \geq 0} \frac{\partial h_t}{\partial \mu_{t-i}} \hat{u}_{t-i-n} \end{pmatrix},$$ \hspace{1cm} (9)

in the metric of $\Sigma_t$, with all functions evaluated at $\hat{\theta}$. Under $H_0$, $s_{mac} \xrightarrow{d} \chi^2_n$.

**Proof.** Use (7) and (8) in Theorem 6.1.

Two cases of particular interest in view of their empirical importance are the simple heteroskedasticity and ARCH models. For the former we have $w_t = 0$ and $\partial h_t / \partial \mu_{t-i} = 0$ for all $i$. Thus an asymptotically equivalent test may be computed as $s_{mac} = T R_0^2$ from the single-length regression of $\hat{\nu}_t$ on

$(\hat{x}_t', \hat{u}_{t-1}, ..., \hat{u}_{t-n})'$, in the metric of $H_t$. For the ARCH(q) model we have

$$\frac{\partial h_t}{\partial \mu_{t-i}} = -2 \sigma_i u_{t-i},$$

under $H_0$, and so $s_{mac} = 2 T R_0^2$ from the regression of
in the metric of $\hat{\Sigma}_t$.

§ 6.2.2 Dynamics in the conditional mean

One of the most complex methodological issues in time series econometrics is the dynamic specification of the model (see Davidson et al [1978], Hendry and Richard [1982, 1983], Hendry et al [1984]). In this subsection we consider tests for order of lags and the closely related issue of the existence of common factors in the lag operator polynomials (Hendry and Mizon [1978], Mizon and Hendry [1980]). Consider the (linear) mean equation

$$[1 - \beta_1(L)] y_t = \beta_2(L)' x_t^* + u_t, \quad (10a)$$

where the polynomials $\beta_1(L)$ and (the vector) $\beta_2(L)$ have degrees $N_1$ and $N_2$, respectively, and the zero order coefficient of $\beta_1(L)$ is null. The equation is derived from the model

$$y_t \mid F_t \sim N[\beta_1(L) y_t + \beta_2(L)' x_t^*, h_t]. \quad (10b)$$

Suppose the polynomial $b(L)$ of order $n \leq \min\{N_1, N_2\}$ is a common factor of $1 - \beta_1(L)$ and $\beta_2(L)$, so that $1 - \beta_1(L) = b(L) [1 - b_1(L)]$ and $\beta_2(L) = b(L) b_2(L)$, where $b_1(L)$ is of degree $n_1 = N_1 - n$, and $b_2(L)$ is of degree $n_2 = N_2 - n$. (10) may be re-expressed in this case as

$$[1 - b_1(L)] y_t = b_2(L)' x_t^* + b(L)^{-1} u_t, \quad (11a)$$

from

$$y_t \mid F_t \sim N[b_1(L) y_t + b_2(L)' x_t^* + b(L) v_t, h_t], \quad (11b)$$
provided the roots of $b(z)$ lie outside the unit circle. If the common roots are null then $b(L) = 1$ and (11) reduces to a form like (10) with lower order polynomials,

$$[1 - b_1(L)] y_t = b_2(L)' x_t^* + u_t,$$

from

$$y_t \ | \ F_t \sim N[b_1(L) y_t + b_2(L)' x_t^*, h_t].$$

To test for lag orders in the conditional mean, (12) is the null hypothesis and (10) represents the alternative, and the LM test is constructed by fitting (12) and computing $2 T R_0^2$ from the regression of

$$(\hat{u}_t) \quad \text{on} \quad \begin{pmatrix} \hat{x}_t' & 0 & y_{t-n_1} & \ldots & y_{t-N_1} & x_{t-n_2}' & \ldots & x_{t-N_2}' \\ \hat{w}_t' & \hat{z}_t' & y_{t-n_1} & \ldots & y_{t-N_1} & x_{t-n_2}' & \ldots & x_{t-N_2}' \end{pmatrix},$$

in the metric of $\hat{\sigma}_t$, where $y_{t-j} = \sum_{i=1}^{j} \delta \mu_{t-i} y_{t-i, j}$, and $x_{t-j} = \sum_{i=1}^{j} \delta \mu_{t-i} x_{t-i, j}$. This statistic is asymptotically $\chi^2(\text{dim}(x^*) + 1)$ under the null.

Testing for zero roots in the common factors ($b(L) = 1$) considers (12) against (11), and the LM test is the test for n-th order autocorrelation of Corollary 6.2.

The test for common factors is that of (11) against (10) and there is no problem in principle to produce the LM test for this situation, but the situation is complex because the restrictions that need to be applied to (10) to obtain (11) are difficult to state explicitly. Since the alternative model is easy to estimate, to use the likelihood ratio test appears as a simpler procedure, though it has the disadvantage that both models need to be estimated. Sargan [1980] has provided an algorithm (COMFAC) that requires estimates only under the alternative

---

(2) The lag orders do not have to be reduced in the same number and the way to approach this problem is methodologically important. For our technical purpose we lose no generality by posing the problem in this way.
and applies a sequence of Wald tests for successive common factors. This procedure is very attractive because it only uses the initial unrestricted estimates and their covariance matrix to produce the sequence of Wald tests, and it is directly applicable in heteroskedastic models because it does not depend on the choice of estimator. The LM and LR tests require new estimates every time the existence of further factors is to be tested, and they also have other disadvantages, as discussed in Hendry et al [1984].

A strategy for the dynamic specification of the conditional mean in heteroskedastic models is to determine the initial lag lengths using some information criteria such as AIC (Akaike [1974]) or BIC (Schwarz [1978], see also Geweke and Meese [1981]), and to use the test for lag orders in (13) and the test for autocorrelation as diagnostics for the chosen specification. Once this is settled, we can proceed to use COMFAC to test for further common factors and zero roots to achieve a more parsimonious parameterization.

§ 6.2.3 Autocorrelation in the variance equation

Autocorrelation may also be present in the variance equation. Suppose the true conditional variance is

$$h_t = h_t + \sum_{j=1}^{n} \rho_j (u_{t-j}^2 - h_{t-j}) = h_t + \rho(L) [ u_t^2 - h_t ] = \rho(L) h_t + \rho(L) u_t^2 ,$$

where $\rho(L) = \sum_{j=1}^{n} \rho_j L^j$, and $\rho(L) = 1 - \rho(L)$. The alternative hypothesis is

$$H_1 : \gamma_t | \mathcal{F}_t \sim N [ \mu_t(\beta) , \bar{h}_t(\theta, \rho) ] .$$

Autocorrelation in the variance equation is a sign of improper dynamic specification and an interpretation is suggested in the next subsection in terms of common factors in transformations of the GARCH polynomials. The test is simpler than that for autocorrelation in the conditional mean because $\partial \mu_t / \partial \rho = 0$ in this case, and since under $H_0$
where \( \xi_{t-1} = (\varepsilon_{t-1}, ..., \varepsilon_{t-n})' \), we have

**Corollary 6.3.** Under the assumptions of Theorem 6.1 the LM test for \( n \)-th order autocorrelation in the variance equation is given by \( svac = 2 T R_0^2 \), where \( R_0^2 \) is the uncentered coefficient of determination of the regression of

\[
\begin{pmatrix}
\hat{u}_t \\
\hat{\varepsilon}_t
\end{pmatrix}
\quad \text{on} \quad \begin{pmatrix}
\tilde{x}_t' & 0 & 0 & \ldots & 0 \\
\tilde{w}_t' & \tilde{z}_t' & \hat{c}_{t-1} & \ldots & \hat{c}_{t-n}
\end{pmatrix},
\]

in the metric of \( \tilde{\Sigma}_t \), with all functions evaluated at \( \hat{\theta} \). Under \( H_0 \), \( svac \xrightarrow{d} \chi^2_n \).

**Proof:** Use (14) in Theorem 6.1.

A joint test for autocorrelation may be constructed by adding simultaneously the variables in (9) and (15).

§ 6.2.4 **Dynamics in the conditional variance**

Consider the conditional variance of a GARCH\((q_1, q_2)\) process,

\[
\alpha_1(L) h_t = \alpha_0 + \alpha_2(L) u_t^2.
\]

- (16a)

If there are common factors in \( \alpha_1(L) \) and \( \alpha_2(L) \) these cancel out and we are left with a lower order model. Since the equality in (16a) is preserved multiplying both sides by the same quantity this shows that common factors in \( \alpha_1(L) \) and \( \alpha_2(L) \) are not identifiable. \( \alpha_1(L) \) has a unit coefficient associated to \( L^0 \), and \( \alpha_2(L) \) has a zero coefficient associated to \( L^0 \). Let \( \overline{\alpha}_1(L) = 1 - \alpha_1(L) \) (i.e. excludes \( L^0 \)), and \( \overline{\alpha}_2(L) = 1 - \alpha_2(L) \) (i.e includes \( L^0 \) with unit coefficient). Substituting these polynomials in (16a) and rearranging terms using \( \varepsilon_t = u_t^2 - h_t \) we obtain

\[
\overline{\alpha}_2(L) u_t^2 = \alpha_0 + \overline{\alpha}_1(L) h_t + \varepsilon_t,
\]

- (16b)
which is the variance equation for a GARCH process. Suppose the polynomial
\( a(L) \) of degree \( q \leq \min ( q_1, q_2 ) \) is a common factor of \( \bar{\alpha}_1(L) \) and \( \bar{\alpha}_2(L) \), so that
\( \bar{\alpha}_1(L) = a(L) \bar{a}_1(L) \) and \( \bar{\alpha}_2(L) = a(L) \bar{a}_2(L) \). Then (16b) may be rewritten as
\[
\begin{align*}
\bar{a}_2(L) u_t^2 &= a_0 + \bar{a}_1(L) h_t + a(L)^{-1} \varepsilon_t ,
\end{align*}
\]
where \( a_0 = a(1)^{-1} \alpha_0 \), provided the roots of \( a(z) \) lie outside the unit circle. This imposes restrictions on the parameterization of the conditional variance which are better seen by writing (17a) as
\[
\begin{align*}
[1 - a(L) \bar{a}_1(L)] h_t &= \alpha_0 + [1 - a(L) \bar{a}_2(L)] u_t^2 ,
\end{align*}
\]
which shows how autocorrelation may arise in the variance equation. If the common roots are zero we have \( a(L) = 1 \), so that the variance equation is
\[
\begin{align*}
\bar{a}_2(L) u_t^2 &= \alpha_0 + \bar{a}_1(L) h_t + \varepsilon_t ,
\end{align*}
\]
where \( a_j(L) = 1 - a_j(L), j = 1, 2 \), and the conditional variance is defined by
\[
\begin{align*}
a_1(L) h_t &= \alpha_0 + a_2(L) u_t^2 ,
\end{align*}
\]
which define the GARCH( \( q_1 - q \), \( q_2 - q \) ) process.

Suppose we have identified the model for \( h_t \) (see § 2.3.4 and § 3.2.1). A suitable diagnostic for the chosen parameterization is the LM test of (18) against (16), which is constructed by fitting the lower order model and using
\[
2 T R_0^2 \text{ from the regression of }
\begin{pmatrix}
\begin{array}{c}
\hat{u}_t \\
\hat{\varepsilon}_t \\
\end{array}
\end{pmatrix}
on \begin{pmatrix}
\begin{array}{ccccccc}
\hat{x}'_t & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
\hat{w}'_t & \hat{z}'_t & \hat{h}_{t-q_1+q} & \ldots & \hat{h}_{t-q_1} & \hat{u}_{t-q_2+q} & \ldots & \hat{u}_{t-q_2}
\end{array}
\end{pmatrix},
\]
and an asymptotically equivalent test may be constructed as \( T R_0^2 \) from the variance auxiliary regression deleting \( \hat{w}_t \) because \( h_{t-j} \) and \( u_{t-j}^2 \) are conditionally even functions of \( u_{t-j} \). Another suitable diagnostic is the test for autocorrelation of Corollary 6.3 (i.e. testing for misspecified orders with
common factor restrictions). Finally, we can use COMFAC to test for the existence of further factors in the chosen model.

§ 6.3 Parameter stability

The importance of stable parameters in econometric relations has long been established, see Chow [1960], Fisher [1970], Hendry [1979], Hendry and Richard [1982, 1983], Engle et al [1983], inter alia. In this section we follow the structural break approach of Chow [1960] to test the constancy of parameters in heteroskedastic models (see Pesaran et al [1986] for a recent survey). Other forms of assessing parameter constancy are considered in Chapter 8. The main issues here are to take advantage of the full information in both conditional moments, and to allow for breaks in either of them. In § 6.3.1 we extend Chow's test by breaking the sample information, and in § 6.3.2 we consider testing prediction errors using one-step forecasts.

§ 6.3.1 Structural break

Toyoda [1974] analyzed the Chow [1960] test using an approximation to its distribution when the structural break was accompanied by a change in variance. His basic conclusion was that the test would have a poor size unless one of the subsamples at least was very large. Schmidt and Sickles [1977] analyzed Toyoda's approximation and their numerical calculations suggested that Toyoda's conclusion for the size of the test appeared to be too pessimistic. In view of Toyoda's findings, Jayatissa [1977] produced a version of the Chow test that would be robust to heteroskedasticity, also in the form of a single change of variance coinciding with the structural break. With White's [1980b] heteroskedasticity-robust covariance matrix, to produce a robust version of the Chow test is now a trivial matter (see Pesaran et al [1986]), but it is more in the
spirit of this Thesis to provide tests that use the information in the conditional variance, and the Chow test is easily extended to heteroskedastic models.

Partition the sample (index) $\mathcal{T} = \{1,\ldots, T\}$ into two subsamples $\mathcal{T}_0$ and $\mathcal{T}_1$ such that $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$. The conditional moments given by $g_t(\theta)$ are assumed to have the same parameterization all through $\mathcal{T}$, but the true value $\theta_0$ for $t \in \mathcal{T}_0$ changes to $\theta_1 = \theta_0 + \Delta\theta$ for $t \in \mathcal{T}_1$. Thus $g_t = g_t(\theta_0)$ for $t \in \mathcal{T}_0$ and $g_t = g_t(\theta_0, \Delta\theta)$ for $t \in \mathcal{T}_1$, and $\Delta\theta = 0$ under the null hypothesis of parameter constancy. When a structural break affects the conditional mean so that $\mu_t = \mu_t(\beta_0)$ for $t \in \mathcal{T}_0$ and $\mu_t = \mu_t(\beta_0, \Delta\beta)$ for $t \in \mathcal{T}_1$ we get

$$\frac{\partial \mu_t}{\partial \Delta\beta} = \begin{cases} 0 & \text{for } t \in \mathcal{T}_0 \\ x_t & \text{for } t \in \mathcal{T}_1 \end{cases}.$$  

If $h_t$ depends on $\beta$ then $h_t = h_t(\theta_0)$ for $t \in \mathcal{T}_0'$, and $h_t = h_t(\theta_0, \Delta\beta)$ for $t \in \mathcal{T}_1'$, so that

$$\frac{\partial h_t}{\partial \Delta\beta} = \begin{cases} 0 & \text{for } t \in \mathcal{T}_0' \\ w_t & \text{for } t \in \mathcal{T}_1' \end{cases}.$$  

where $\mathcal{T} = \mathcal{T}_0' \cup \mathcal{T}_1'$, and the distinction in the partition simply takes into account that the dependence of $h_t$ on $\mu_t$ may be lagged. For example in the ARCH(1) model a break in $\mu_t$ in period $t$ shows in $h_t$ in period $t + 1$.

For a structural break in the conditional variance we have $h_t = h_t(\theta_0)$ for $t \in \mathcal{T}_0^*$ and $h_t = h_t(\theta_0, \Delta\alpha)$ for $t \in \mathcal{T}_1^*$, $\mathcal{T} = \mathcal{T}_0^* \cup \mathcal{T}_1^*$. Therefore,

$$\frac{\partial h_t}{\partial \Delta\alpha} = \begin{cases} 0 & \text{for } t \in \mathcal{T}_0^* \\ z_t & \text{for } t \in \mathcal{T}_1^* \end{cases}.$$  

and $\mu_t = \mu_t(\beta_0)$ for $t \in \mathcal{T}$ because there is no feedback from $h_t$ to $\mu_t$.

*Corollary 6.4.* Under the assumptions of Theorem 6.1, consider the test of parameter constancy against the alternative that there has been a structural break in the conditional mean during $\mathcal{T}_1$ and a structural break in the
conditional variance in $\mathcal{I}_1^*$. The LM test is based on $s_{ab} = 2 \, T \, R_0^2$, where $R_0^2$ is the uncentered coefficient of determination in the regression of

$$
\begin{pmatrix}
\hat{u}_t \\
\hat{e}_t
\end{pmatrix}
on
\begin{pmatrix}
\hat{x}_t' \\
\hat{e}_t'
\end{pmatrix}
\begin{pmatrix}
0 & \partial \mu_t / \partial \Delta \beta' & 0 \\
\partial \mu_t / \partial \Delta \beta' & \partial \sigma_t / \partial \Delta \beta' & \partial \sigma_t / \partial \Delta \alpha'
\end{pmatrix},
$$

in the metric of $\Sigma_t$, with all functions evaluated at $\hat{\theta}$. Under $H_0$, $s_{ab} \rightarrow \chi_n^2$, where $n = p$ if both $\mathcal{I}_1$ and $\mathcal{I}_1^*$ are non-empty, $n = k$ if $\mathcal{I}_1^*$ is empty and $n = p - k$ if $\mathcal{I}_1$ is empty.

Proof: Use (19) - (21) in Theorem 6.1.

The test is easily extended to multiple breaks and/or to partial structural breaks in which only a subvector of $\theta$ is affected.

§ 6.3.2 Prediction error tests

In this subsection the full sample $\mathcal{T}$ of $T$ observations is used to estimate $\theta$, but we assume that an extra set $\mathcal{T}_f$ of $n$ observations has been set aside to assess parameter constancy. We use one-step predictors, that is, the information set $\mathcal{F}_f$ is assumed available to predict for $t \in \mathcal{T}_f$. Under these conditions the optimal forecasts are given by the conditional expectations $\mu_t$ and $h_t$, and so the predictors are

$$
\hat{y}_f = \mu_f(\hat{\beta}) \quad \text{and} \quad \hat{h}_f = h_f(\hat{\theta}),
$$

where $y_f \in \mathcal{T}_f$ is an $n$-vector, and so are $h_f$ and $\mu_f$. The prediction error for the conditional mean is

$$
\hat{u}_f = y_f - \hat{y}_f = \mu_f(\beta_0) - \mu_f(\hat{\beta}) + u_f - X_f(\hat{\beta} - \beta_0) + o_p(T^{-1/2}), \quad (22)
$$

by use of the Mean Value Theorem. If $\mathcal{T}_f$ is fixed while $T \rightarrow \infty$ (something which is odd but is the usual way of getting an operative approximation), the second term in (22) approaches zero and we are left only with $u_f$, with
covariance matrix $\Omega_f = \text{diag} \{ h_f \}$. Thus as in Davidson et al [1978] this suggests the test

$$s_{\text{mpe}}^* = \hat{\delta}_f \hat{\Omega}_f^{-1} \tilde{u}_f \xrightarrow{d} \chi^2_n,$$

under $H_0$. Pagan and Nicholls [1984] suggest that using the variance of the term in $\hat{\beta} - \beta_0$ to approximate the covariance matrix of $\tilde{u}_f$ may produce better results in small samples. This would lead to the modified statistic

$$s_{\text{mpe}} = \hat{u}_f' [ \hat{\Omega}_f + T^{-1} \tilde{X}_f V(\hat{\beta}) \tilde{X}_f' ]^{-1} \hat{u}_f \xrightarrow{d} \chi^2_n,$$

whose computation is more complicated than (23), but Salkever [1976] has devised a simple method to produce the information for $s_{\text{mpe}}$ by the simple addition of the observations in $\mathcal{Y}_f$ and some dummy variables. Pagan and Nicholls [1984] have extended Salkever's results in several directions, and from their equations (4) and (5) we need to look at the estimation problem

$$\min_{\theta, \delta_m} \{ v' \Sigma^{-1} v + u_{\text{f}}^* \Omega_f^{-1} u_{\text{f}}^* \},$$

where $u_{\text{f}}^* = y_f - \mu_f(\beta) - \delta_m$. Because $\delta_m \in \mathbb{R}^n$, the second term is annihilated by making $\delta_m = y_f - \mu_f = u_f$. The first term is minimized at the MLE $\hat{\theta}$.

Therefore $\hat{\delta}_m = \hat{u}_f$ and its covariance matrix from the GLS estimation in (25) is the one used in (24) for the construction of $s_{\text{mpe}}$, and we must add the observations $y_f = \mu_f(\beta) + \delta_m + \hat{u}_{\text{f}}^*$ with metric $\Omega_f$ to the mean equation of the two-equation system in order to obtain $s_{\text{mpe}}$. Considering only the mean equation as in Pagan and Nicholls would produce prediction errors in terms of $\hat{\beta}_m$ and not of $\hat{\beta}$, with the corresponding loss of the variance information and hence power, though the test has the correct size.

For the conditional variance the prediction error is $\hat{e}_f = h_f - \hat{h}_f$, where $\hat{h}_f = h(\hat{\theta})$, but this has no diagnostic use because $h_f$ is not observable. We must turn to the operative prediction error

$$\hat{e}_f = \hat{u}_f^2 - \hat{h}_f = (\hat{u}_f^2 - h_f) + (h_f - \hat{h}_f) = e_f + \hat{e}_f,$$
which using $\hat{\epsilon}_f^2 = \hat{u}_f^2 + (\hat{u}_f^2 - u_f^2)$ may be rewritten as

$$\hat{\epsilon}_f = \epsilon_f - (\hat{\mu}_f - h_f) + (\hat{u}_f^2 - u_f^2) = \epsilon_f - S_f(\theta - \theta_0) + o_p(T^{1/2}),$$  

(26)

where the last inequality has been obtained using the Mean Value Theorem and Lemma 3.3. The second term in (26) is $o_p(1)$, and so as $T \to \infty$

$$s_{\text{vpe}}^* = \frac{1}{2} \hat{\epsilon}_f' \Omega_f^{-2} \hat{\epsilon}_f \xrightarrow{d} \chi_n^2,$$

(27)

under $H_0$. Following the argument of Pagan and Nicholls to consider a higher order approximation to the covariance matrix we get

$$s_{\text{vpe}} = \hat{\epsilon}_f' [2 \Omega_f^{-2} + T^{-1} \hat{S}_f V(\theta) \hat{S}_f']^{-1} \hat{\epsilon}_f \xrightarrow{d} \chi_n^2,$$

under $H_0$, and the information for the calculation of $s_{\text{vpe}}$ is provided by the estimation criterion

$$\min_{\theta, \delta_v} \{ \nu' \Sigma^{-1} \nu + \frac{1}{2} \epsilon_f' \Omega_f^{-2} \epsilon_f \},$$

where $\epsilon_f^* = \hat{u}_f^2 - h_f(\theta) - \delta_v$. That is, add $\hat{u}_f^2 = h_f(\theta) + \delta_v$ in the metric of $2 \Omega_f^2$ to the variance equation of the two-equation system.

A joint mean-variance prediction error test is produced from (23) and (27),

$$s_{\text{pe}}^* = \hat{\nu}_f' \Sigma_f^{-1} \hat{\nu}_f = s_{\text{mpe}}^* + s_{\text{vpe}}^* \xrightarrow{d} \chi_{2n}^2,$$

under $H_0$, where $\nu_f = (u_f', \epsilon_f')'$ and $\Sigma_f = \text{diag}(\Omega_f, 2 \Omega_f^2)$. The absence of correlation between $u_f$ and $\epsilon_f$ implies asymptotic independence for $s_{\text{mpe}}^*$ and $s_{\text{vpe}}^*$, but this is lost at the higher order approximation because the term in $\hat{\theta} - \theta_0$ in (26) is correlated with both terms in (22). The correct statistic is

$$s_{\text{pe}} = \hat{\nu}_f' [\hat{\Sigma}_f + T^{-1} \hat{G}_f V(\hat{\theta}) \hat{G}_f']^{-1} \hat{\nu}_f,$$

which may be obtained by adding jointly the observations for the prediction errors in both equations.

We have only treated one-step prediction errors. Pagan and Nicholls have
also considered the case of multi-step prediction errors and their results can be applied to our two-equation system to produce the more demanding dynamic prediction error tests.

§ 6.4 LM Tests for weak and strong exogeneity

Weak exogeneity of a set of conditioning variables for a parameter vector requires (see Engle, Hendry and Richard [1983], EHR henceforth)

(i) an orthogonality condition between these variables and the disturbances of the conditional model, and

(ii) that the conditional likelihood contains all the relevant information for the estimation of the parameter vector.

Condition (i) suffices for consistent estimation and condition (ii) establishes efficiency. Suppose the DGP for \((y_t, x^*_t)\) can be expressed parametrically as

\[
D(y_t, x^*_t \mid \vartheta_t; \xi) = D(y_t \mid \mathcal{F}_t; \xi_1) D(x^*_t \mid \vartheta_t; \xi_2),
\]

where \(\xi\) is the parameter vector and both \(\xi_1\) and \(\xi_2\) are functions of \(\xi\). The proposed model for the conditional distribution of \(y_t\) is \(f(y_t \mid \mathcal{F}_t; \theta)\) and the parameters of interest are \(\pi = \pi(\theta)\). The orthogonality condition is fulfilled if the model represents the proper conditional pdf in (28) in the sense that for some \(\theta_0 \in \Theta\), \(f(y_t \mid \mathcal{F}_t; \theta_0) = D(y_t \mid \mathcal{F}_t; \xi_1)\). Condition (ii) requires all information about \(\pi\) to be provided by \(f(y_t \mid \mathcal{F}_t; \theta_0)\), and is presented by EHR as having variation free \(\xi_1\) and \(\xi_2\). We find it convenient to separate this into two concepts: the "complete information" and the "sufficient information" of the conditional likelihood \(Z(\theta)\) about the parameters of interest. By complete information we mean that all relevant information about \(\pi\) be contained in \(Z(\theta)\), or equivalently that the marginal likelihood \(Z(\xi_2) = \prod_{t=1}^T f(x^*_t \mid \vartheta_t; \xi_2)\) does not contain any information about \(\pi\). By sufficient information we mean that
\( \mathcal{L}(\theta) \) does not depend on parameters which do not properly belong to the conditional model.

To illustrate this situation let \( x_t^* \) be scalar and consider a first order bivariate ARCH model (see Kraft and Engle [1982]). The mean is assumed zero for simplicity, and the conditional variances are

\[
\text{var} ( y_t | \mathcal{G}_t ) = h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 v_{t-1}^2 + \alpha_3 u_{t-1} v_{t-1}, \tag{29a}
\]

and

\[
\text{var} ( x_t^* | \mathcal{G}_t ) = h_{xt} = \alpha_{x0} + \alpha_{x1} u_{t-1}^2 + \alpha_{x2} v_{t-1}^2 + \alpha_{x3} u_{t-1} v_{t-1}, \tag{29b}
\]

where \( v_t = x_t^* - \mathbb{E} [ x_t^* | \mathcal{G}_t ] \). We assume joint normality conditional on \( \mathcal{G}_t \) with zero conditional covariance between \( y_t \) and \( x_t^* \). The parameters of interest are \( \alpha = ( \alpha_0 , \alpha_1 )' \). The model \( y_t | \mathcal{F}_t \sim N \left[ 0, h_t \right] \) provides the conditional likelihood, while the submodel \( x_t^* | \mathcal{G}_t \sim N \left[ 0, h_{xt} \right] \) produces the marginal likelihood.

Observe that an ARCH(1) process for \( y_t \) does not provide the adequate conditional likelihood and results in inconsistent estimates for \( \alpha \). The marginal likelihood contains information about \( \alpha \) through the presence of \( u_{t-1} \) in \( h_{xt} \). Thus \( \mathcal{L}(\theta) \) does not have complete information about \( \alpha \). For complete information we require \( \alpha_{x1} = \alpha_{x3} = 0 \). At the same time, the conditional likelihood depends on the \( \alpha_x \)'s through the presence of \( v_{t-1} \) in \( h_t \). Thus \( \mathcal{L}(\theta) \) does not have sufficient information about \( \alpha \). Sufficient information is achieved if \( \alpha_2 = \alpha_3 = 0 \). In this case sufficient information also makes the ARCH(1) the correct model.

To formalize these concepts partition \( \theta = (\theta_1', \theta_2')' \) in such a way that the parameters of interest are a function of \( \theta_1 \) only, \( \pi = \pi(\theta_1) \), and \( \theta_2 \) is a set of nuisance parameters. To avoid arbitrariness other than the selection of parameters of interest — always a subjective matter — partition \( \theta \) such that all elements of \( \theta_1 \) are required for \( \pi \), i.e. \( \partial \pi / \partial \theta_{1j} \neq 0 \) for all \( \theta_{1j} \) in \( \theta_1 \). Then we have

**Definition 6.5.** The conditional likelihood \( \mathcal{L}(\theta) \) is said to be information-
complete for inference on \( \pi \) (denoted i-complete) if \( \xi_1 \) and \( \xi_2 \) are variation free, and it is said to be information-sufficient for inference on \( \pi \) (denoted i-sufficient) if \( \xi_2 \) and \( \xi_2' \) are variation free.

The following result re-interprets EHR's definition of weak exogeneity:

**Lemma 6.6.** The conditioning variables \( x_t^* \) are weakly exogenous for \( \pi = \pi(\theta) \) if, and only if,

(i) \( \mathcal{F}(\theta) \) is derived from a proper conditional/marginal factorization,
(ii) \( \mathcal{F}(\theta) \) is i-sufficient (for inference on \( \pi \)), and
(iii) \( \mathcal{F}(\theta) \) is i-complete (for inference on \( \pi \)).

To test for weak exogeneity we produce separate tests for each of the three conditions of Lemma 6.6 given the remaining two conditions, so that each test is easily interpreted, and the implicit assumptions are clear when weak exogeneity is assessed from a subset of the three conditions.

Tests for the orthogonality condition (i) have been put forward in homoskedastic linear models by Wu [1973, 1974], Revankar and Hartley [1973], Revankar [1978], Hausman [1978], Hwang [1980] and Spencer and Berk [1981], *inter alia*. The relations between some of these tests have been analyzed by Nakamura and Nakamura [1981], and EHR have shown that under certain assumptions about the structure of the implicit simultaneous model, the tests can be interpreted as tests for the weak exogeneity of a subset of the endogenous variables for the parameters of the structural equations of the remaining endogenous variables. Alternative sets of such assumptions are presented in their Theorem 4.3. Engle [1982b, 1984], making explicit reference to weak exogeneity, provides more general versions of the Wu and Hausman tests and establishes their asymptotic optimality by showing them to be asymptotically or numerically equivalent to the LM test in a limited information framework.

The tests mentioned above are essentially tests of the adequacy of the
proposed conditional/marginal factorization of the joint DGP for \((y_t, x_t^*)\). To produce a similar test in a heteroskedastic context the model in (1) is completed with the distribution of \(x_t^*\) conditional on \(\mathcal{G}_t\), assumed to be

\[ x_t^* | \mathcal{G}_t \sim N \left[ \mu_{xt}, H_{xt} \right], \tag{30} \]

and possibly a set of structural relations defining implicitly the conditional mean of \(x_t^*\), say \(F_t(\mu_{xt}) = 0\), with \(\mu_{xt}, H_{xt}, F_t \in \mathcal{G}_t\). Thus we are implicitly assuming that the structural equations for \(x_t^*\) do not depend on \(y_t\).

Alternatively, we may allow \(F_t\) to depend on \(y_t\) provided that \(F_t\) represents a set of exactly identified equations and hence imposes no restrictions on the parameters of (30) (see EHR). To analyze whether (1) and (30) properly define the conditional/marginal factorization for the joint DGP for \((y_t, x_t^*)\) we generalize this joint pdf to

\[
D \left( y_t, x_t^* | \mathcal{G}_t \right) = (2\pi)^{(n_x + 1)/2} |H_t|^{-1/2} \exp \left\{ -\frac{1}{2} v_t' H_t^{-1} v_t \right\}, \tag{31}
\]

where \(v_t = (y_t - \mu_t, (x_t^* - \mu_{xt})')'\), \(n_x = \dim(x_t^*)\), and \(H_t = \begin{pmatrix} h_t & c_t' \\ c_t & H_{xt} \end{pmatrix}\). Although \(D \left( y_t, x_t^* | \mathcal{G}_t \right)\) has the form of the normal pdf, the distribution of \((y_t, x_t^*)\) conditional on \(\mathcal{G}_t\) need not be normal because \(\mu_t\) and \(h_t\) may not be measurable functions of \(\mathcal{G}_t\) (they may be functions of \(x_t^*\)). Therefore \((\mu_t, \mu_{xt}')'\) and \(H_t\) may not be the moments of the distribution conditional on \(\mathcal{G}_t\). When \(c_t = 0\) for all \(t\), \(D \left( y_t, x_t^* | \mathcal{G}_t \right)\) collapses into the conditional/marginal factorization given by (1) and (30). By the usual completing of squares for the terms in \(u_t\) of the exponent in (31) (see e.g. Zellner [1971]), we can factorize the joint pdf into the factors

\[
f \left( y_t | \mathcal{F}_t \right) = (2\pi)^{-1/2} \sigma_t^{-1} \exp \left\{ -\frac{1}{2\sigma_t^2} \left( u_t - c_t' H_{xt}^{-1} v_t \right)^2 \right\}, \tag{32a}
\]

and

\[
f \left( x_t^* | \mathcal{G}_t \right) = (2\pi)^{-n_x/2} |H_{xt}|^{-1/2} \exp \left\{ -\frac{1}{2} v_{xt}' H_{xt}^{-1} v_{xt} \right\}, \tag{32b}
\]

where \(\sigma_t^2 = (h_t - c_t' H_{xt}^{-1} c_t)'\), and \(v_{xt} = x_t^* - \mu_{xt}\). To test for proper factorization we parameterize the function \(c_t = c_t(\alpha_{12})\), in such a way that \(\alpha_{12} = 0\) implies \(c_t = 0\). The log-likelihood for the conditional model under the alternative hypothesis is
then from (32)
\[ \mathcal{L}(\theta, \alpha_{12}, \theta_x) = -T^{-1} \sum_{t=1}^{T} \log \sigma_t - T^{-1} \sum_{t=1}^{T} \frac{1}{2\sigma_t} (u_t - c_t'H_{xt} v_{xt})^2 , \]
while the marginal likelihood is
\[ \mathcal{L}_x(\theta, \alpha_{12}, \theta_x) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log |H_{xt}| - \frac{1}{2} T^{-1} \sum_{t=1}^{T} v_{xt}'H_{xt}^{-1} v_{xt} , \]
where \( \theta_x \) are the parameters of the marginal distribution. The joint log-likelihood is simply
\[ \mathcal{L}_j(\theta, \alpha_{12}, \theta_x) = \mathcal{L}(\theta, \alpha_{12}, \theta_x) + \mathcal{L}_x(\theta, \alpha_{12}, \theta_x) . \]

We then have

**Corollary 6.7.** Under the assumptions of Theorem 6.1 and

(i) \( x_t^* | \theta_t \sim N[\mu_{xt}, H_{xt}] \) and the functions \( \mu_{xt} \) and \( H_{xt} \) obey the same regularity conditions as \( \mu_t \) and \( h_t \), and

(ii) \( \mathcal{L}(\theta) \) is i-sufficient and i-complete,

the LM test for proper conditional/marginal factorization is \( s_{pf} = 2T \mathcal{R}_0^2 \) from the regression of

\[
\begin{pmatrix}
\hat{u}_t \\
\hat{e}_t
\end{pmatrix}
\text{ on }
\begin{pmatrix}
\hat{x}_t' \\
\hat{v}_{xt}' \hat{H}_{xt}^{-1} \frac{\partial \hat{c}_t}{\partial \alpha_{12}}
\end{pmatrix}, \text{ in the metric of } \hat{\Sigma}_t ,
\]

where all functions are evaluated at \( \hat{\theta}, \hat{\theta}_x, \) and \( \alpha_{12} = 0 \). Under \( H_0: \alpha_{12} = 0 \),

\[ s_{pf} \xrightarrow{\text{d}} \chi^2_{\text{dim}(\alpha_{12})} . \]

**Proof:** From (32a) we have \( \mu_t(\theta, \theta_x, \alpha_{12}) = \mu_t(\theta) + c_t(\alpha_{12})' H_{xt}(\theta_x)^{-1} v_{xt}(\theta_x) \), and \( h_t(\theta, \theta_x, \alpha_{12}) = h_t(\theta) - c_t(\alpha_{12})' H_{xt}(\theta_x)^{-1} c_t(\alpha_{12}) \), using the i-completeness and i-sufficiency of \( \mathcal{L}(\theta) \). Therefore under \( H_0 \),

\[ \frac{\partial \mu_t(\theta, \theta_x, \alpha_{12})}{\partial \alpha_{12}} = \frac{\partial c_t'}{\partial \alpha_{12}} H_{xt}^{-1} v_{xt} , \]
and 
\[ \frac{\partial h_t(\theta, \theta_x, \alpha_{12})}{\partial \alpha_{12}} = -2 \frac{\partial c_t'}{\partial \alpha_{12}} H_{xt}^{-1} c_t = 0, \]
and the Theorem is proved using these expressions in Theorem 6.1.

This result is a generalization of Wu's test to a heteroskedastic setting. The test almost as simple to construct as the test under homoskedasticity (see Engle [1982b, 1984]). The differences are the double-length auxiliary regression to incorporate variance information, and that the distribution for \( x_t^* \) is allowed to be heteroskedastic. At a diagnostic level researchers may prefer to ignore the latter fact and assume \( x_t^* \) homoskedastic, so that the residuals and covariance matrix may be obtained by simple multivariate regression. The most remarkable feature is that no regressors need to be augmented to the auxiliary variance equation. This can be explained because satisfying the orthogonality condition for the mean equation fulfills the condition for the variance equation: when the moments are correctly specified \( E[ut|\mathcal{F}_t] = 0 \) implies \( E[\varepsilon_t|\mathcal{F}_t] = 0 \) because \( \varepsilon_t = u_t^2 - h_t \). As an example consider the bivariate ARCH(1) model in (29) completed with a nonzero conditional covariance given by

\[ c_t(\alpha_{12}) = \text{cov}(y_t, x_t^* | \mathcal{G}_t) = \alpha_{yx0} + \alpha_{yx1} u_{t-1}^2 + \alpha_{yx2} v_{t-1}^2 + \alpha_{yx3} u_{t-1} v_{t-1}, \]

where \( \alpha_{12} = (\alpha_{yx0}, \alpha_{yx1}, \alpha_{yx2}, \alpha_{yx3})' \). Then \( \partial c_t / \partial \alpha_{12} = (1, u_{t-1}^2, v_{t-1}^2, u_{t-1} v_{t-1})' \), and therefore the variables to add to the auxiliary mean equation are given by the vector \( \hat{h}_{xt} v_{xt}(1, \hat{u}_{t-1}^2, \hat{v}_{t-1}^2, \hat{u}_{t-1} \hat{v}_{t-1})' \).

To analyze both i-sufficiency and i-completeness we define \( g_{xt}(\theta, \theta_x) = (\mu_{xt}', \text{vech } H_{xt})' \), where vech is the vectorization of the lower triangle of \( H_{xt} \) (see Henderson and Searle [1979]). To test for i-sufficiency the dependence of \( g_t \) on \( \theta_x \) must be made explicit. We take this to arise through dependence on the conditional moments of \( x_t^* \) as

\[ \mu_t = \mu_t(\beta, \tau_y; g_{xt,j}, j \geq 0), \text{ and } h_t = h_t(\theta, \tau_y; g_{xt,j}, j \geq 0), \]

- (33)
where \( \tau_y = (\tau_{y1}', \tau_{y2}')' \), so that \( \mu_t \) does not depend on \( g_{xt-j} \) when \( \tau_{y1} = 0 \), and the same applies to \( h_t \) when \( \tau_y = 0 \). The partition of \( \tau_y \) allows for \( h_t \) to depend on \( \theta_x \) either directly (\( \tau_{y2} \)) or indirectly through its dependence on \( \mu_t (\tau_{y1}) \). The presence of \( g_{xt} \) in \( \mu_t \) may be due to cross-autocorrelation or risk terms defined on the distribution of \( x_t^* \) (e.g. Pagan [1984b], Pagan and Ullah [1986]), and the presence in \( h_t \) may be caused by multivariate ARCH effects (Kraft and Engle [1982]). We then have

**Corollary 6.8.** - Under the assumptions of Theorem 6.1 and

(i) \( x_t^* | \Theta_t \sim N [\mu_{xt}, H_{xt}] \) and the functions \( \mu_{xt} \) and \( H_{xt} \) obey the same regularity conditions as \( \mu_t \) and \( h_t \), and

(ii) \( \mathcal{I}(\theta) \) is properly factorized and i-complete,

(iii) \( \theta \) and \( \theta_x \) are variation free, and

(iv) the dependence of \( g_t \) on \( \theta_x \) is as given in (33),

the LM test for i-sufficiency of \( \mathcal{I}(\theta) \) is given by \( s_{is} = 2 T R_0^2 \) from the regression of

\[
\begin{pmatrix}
\hat{u}_t \\
\hat{e}_t
\end{pmatrix}
\text{ on }
\begin{pmatrix}
\hat{x}_t' & 0 & \frac{\partial \mu_t}{\partial \tau_{y1}} & 0 \\
\hat{w}_t' & \frac{\partial h_t}{\partial \tau_{y1}} & \frac{\partial h_t}{\partial \tau_{y2}}
\end{pmatrix},
\]

in the metric of \( \hat{\Sigma}_t \),

where all functions are evaluated at \( \hat{\theta}, \hat{\theta}_x \), and \( \tau_y = 0 \). Under \( H_0: \tau_y = 0 \),

\[
s_{is} \overset{d}{\to} \chi_{\text{dim} (\tau_y)}^2 .
\]

**Proof:** If \( \theta \) and \( \theta_x \) are variation free we only require \( g_t \) not to depend on \( \theta_x \).

Given that this dependence takes the form (33) and \( \alpha_{12} = 0 \), the test is a variable addition test and thus a particular case of Theorem 6.1, and the additions to the auxiliary regression follow obviously.

As an illustration consider again the bivariate ARCH in (29). For this case \( \tau_{y1} = 0 \) and \( \tau_{y2} = (\alpha_2, \alpha_3)' \), and we simply add \( (v_{t-1}, u_{t-1}, v_{t-1})' \) to the auxiliary variance equation.
Similarly, to test i-completeness the dependence of $g_{xt}$ on $\theta$ must be made explicit, and we take this to arise through dependence on lagged conditional moments of $y_t$ as

$$\mu_{xt} = \mu_t (\beta_x, \tau_{x1}; g_{t-j}, j > 0), \quad \text{and} \quad H_{xt} = H_{xt} (\theta_x, \tau_x; g_{t-j}, j > 0),$$

(34)

where $\tau_x = (\tau_{x1}, \tau_{x2})'$, and $\mu_{xt}$ does not depend on $g_{t-j}$ when $\tau_{x1} = 0$ and the same applies to $H_{xt}$ when $\tau_x = 0$. The interpretation of $\tau_x$ may be given along the same lines of that of $\tau_y$ for i-sufficiency. The next corollary is included here for the sake of completeness, though it uses the multivariate version of Theorem 6.1 which is presented in Chapter 9.

**Corollary 6.9.** Under the assumptions of Theorem 6.1 and

(i) $x_t^* | \theta_t \sim N (\mu_{xt}, H_{xt})$ and the functions $\mu_{xt}$ and $H_{xt}$ obey the same regularity conditions as $\mu_t$ and $h_t$, and

(ii) $\mathcal{I}(\theta)$ is properly factorized and i-sufficient,

(iii) $\theta$ and $\theta_x$ are variation free, and

(iv) the dependence of $g_{xt}$ on $\theta$ is as given in (34),

the LM test for i-completeness of $\mathcal{I}(\theta)$ is given by $s_{ic} = \frac{1}{2} n_x (n_x + 3) T R_0^2$ from the regression of

$$\begin{pmatrix} \hat{v}_t \\ \text{vech}[\hat{v}_t \hat{v}_t'] - \hat{H}_{xt} \end{pmatrix} \quad \text{on} \quad \begin{pmatrix} \frac{\partial \mu_{xt}}{\partial \beta_{x}'} & 0 & \frac{\partial \mu_{xt}}{\partial \tau_{x1}'} & 0 \\ \frac{\partial \text{vech} H_{xt}}{\partial \beta_{x}'} & \frac{\partial \text{vech} H_{xt}}{\partial \alpha_{x}'} & \frac{\partial \text{vech} H_{xt}}{\partial \tau_{x1}'} & \frac{\partial \text{vech} H_{xt}}{\partial \tau_{x2}'} \end{pmatrix},$$

in the metric of $\text{diag} (\hat{H}_{xt}^{-1}, P (\hat{H}_{xt}^{-1} \otimes \hat{H}_{xt}^{-1}) P')$, where $P$ is the matrix such that vec $H_{xt} = P' \text{vech} H_{xt}$, and all functions are evaluated at $\hat{\theta}, \hat{\theta}_x$, and $\tau_x = 0$.

Under $H_0: \tau_x = 0$, $s_{ic} \overset{d}{\rightarrow} \chi^2_{\text{dim}(\tau_x)}$.

**Proof:** If $\theta$ and $\theta_x$ are variation free we only require $g_{xt}$ not to depend on $\theta$. Given that this dependence takes the form (34) and $\alpha_{12} = 0$, the test is a variable addition test and is a particular case of Theorem 9.14. The additions to the
The number of auxiliary regressions depends on the dimension of $x_t^*$. If no restrictions are imposed \textit{a priori} on $H_{xt}$, the number of parameters increases geometrically (see Kraft and Engle [1982], Diebold and Nerlove [1986]), and thus the test may only be feasible for small $n_x$. For the bivariate ARCH in (29) we have $\tau_{x1} = 0$, while $\tau_{x2} = (\alpha_{x1}, \alpha_{x3})'$, so the additions to the auxiliary equation for the variance of $x_t^*$ are given by $(u_{t-1}, u_{t-1} v_{t-1})'$.

Collecting the three corollaries we may construct a full test for weak exogeneity. Under weak exogeneity $s_{ic}$ is asymptotically independent from both $s_{pf}$ and $s_{is}$, though the latter two statistics are dependent. A joint test of proper factorization and sufficiency, say $s_{pf-is}$, may be constructed by adding jointly the variables of Corollaries 6.7 and 6.8 to the auxiliary regression. The statistic to test for full weak exogeneity is then $s_{we} = s_{pf-is} + s_{ic}$, which is asymptotically $\chi^2$ under the null hypothesis, with $\dim (\alpha_{12}) + \dim(\tau_y) + \dim(\tau_x)$ degrees of freedom.

Weak exogeneity allows for consistent and efficient estimation. The next question is whether extraneous predictors of the weakly exogenous variables can be used in order to predict $y_t$. For this purpose, the conditioning variables $x_t^*$ must be strongly exogenous for the parameters of interest, that is, they must be weakly exogenous and not G-caused (caused in the sense of Granger [1969]) by $y_t$. Therefore to test for strong exogeneity we must add a test for G-causality to the testing procedure considered above. Tests for G-causality have been considered by Geweke [1978, 1984], and the null hypothesis is that the conditional distribution of $x_t^*$ does not depend on the past of $y_t$. If this is the case we can write $f(\ x_t^* \mid \mathcal{G}_t) = f(\ x_t^* \mid X_{t-1}, Y_0)$, conditioning on initial values of the dependent variable. In a heteroskedastic environment we must ensure that neither conditional moment is being affected by $Y_{t-1}$. Suppose we can decompose the conditional moments of $x_t^*$ as
\[ \mu_{xt}(\beta_x; \theta_t) = \mu_{1xt}(\beta_{x1}; X_{t-1}, Y_0) + \mu_{2xt}(\beta_{x2}; \theta_t), \]  
and
\[ \text{vech} \, H_{xt}(\theta_x; \theta_t) = \text{vech} \, H_{1xt}(\beta_x, \alpha_{x1}; \theta_t) + \text{vech} \, H_{2xt}(\alpha_{x2}; \theta_t), \]

where \( \beta_x = (\beta_{x1}', \beta_{x2}')', \alpha_x = (\alpha_{x1}', \alpha_{x2}')', \mu_{2xt}(\beta_{x2} = 0; \theta_t) = 0, \)
\[ H_{2xt}(\alpha_{x2} = 0; \theta_t) = 0, \text{ and } H_{1xt}(\beta_{x1} = 0, \beta_{x2} = 0, \alpha_{x1}; \theta_t) = H_{1xt}(\beta_{x1}, \alpha_{x1}; X_{t-1}, Y_0). \]
Thus when \( \beta_{x1} = 0 \) neither \( \mu_{xt} \) nor \( \text{vech} \, H_{1xt} \) depend on \( Y_{t-1} \), and when \( \alpha_{x2} = 0 \) in addition, \( \text{vech} \, H_{xt} \) does not depend on \( Y_{t-1} \). The next corollary also uses Theorem 9.14 and is included here for the sake of completeness.

**Corollary 6.10.** Under the assumptions of Theorem 6.1 and

(i) \( x_t^* \) is weakly exogenous for \( \theta \), and

(ii) \( x_t^* \mid \theta_t \sim N[\mu_{xt}, H_{xt}] \), the functions \( \mu_{xt} \) and \( H_{xt} \) obey the same regularity conditions as \( \mu_t \) and \( h_t \), and have the form in (35),

the LM test for strong exogeneity of \( x_t^* \) for \( \theta \) is given by
\[ s_{se} = \frac{1}{2} n_x (n_x + 3) T R_0^2 \]
from the regression of
\[ \begin{pmatrix} \hat{v}_t \\ \text{vech} [\hat{v}_t \hat{v}_t' - \hat{H}_{xt}] \end{pmatrix} \]
on
\[ \begin{pmatrix} \frac{\partial \mu_{1xt}}{\partial \beta_{x1}'} & 0 & \frac{\partial \mu_{2xt}}{\partial \beta_{x2}'} & 0 \\ \frac{\partial \text{vech} \, H_{1xt}}{\partial \beta_{x1}'} & \frac{\partial \text{vech} \, H_{1xt}}{\partial \alpha_{x1}'} & \frac{\partial \text{vech} \, H_{1xt}}{\partial \beta_{x2}'} & \frac{\partial \text{vech} \, H_{2xt}}{\partial \alpha_{x2}'} \end{pmatrix} \]
in the metric of
\[ \text{diag} \{ \hat{H}_{xt}^{-1}, P' (\hat{H}_{xt}^{-1} \otimes \hat{H}_{xt}^{-1}) P' \}, \]
where \( P \) is the matrix such that \( \text{vec} \, H_{xt} = P' \text{vech} \, H_{xt} \), and all functions are evaluated at \( \hat{\theta}, \hat{\theta}_x, \beta_{x2} = 0, \) and \( \alpha_{x2} = 0. \) Under \( H_0, s_{se} \xrightarrow{d} \chi^2 \dim(\beta_{x2}) + \dim(\alpha_{x2}). \)

**Proof:** Given weak exogeneity, we need only test that the conditional moments of \( x_t^* \) do not depend on the past of \( y_t \), and this is accomplished when \( \beta_{x2} = 0 \) and \( \alpha_{x2} = 0 \) in view of (35). The test is a variable addition test and is a particular case of Theorem 9.14. The additions to the auxiliary regression follow obviously.
To construct a full test for strong exogeneity observe that \( s_{ic} \) and \( s_{se} \) are asymptotically dependent in general, but the auxiliary regression that incorporates the additional variables in Corollaries 6.9 and 6.10 (avoiding redundancies if necessary) provides a statistic \( s_{ic-se} \) that is asymptotically independent of \( s_{pf-is} \), and thus \( s_{fse} = s_{pf-is} + s_{ic-se} \) gives the appropriate statistic. An alternative approach to measure the feedback between time series rather than constructing G-causality tests has been put forward by Geweke [1982,1986b], and this may be adapted to the heteroskedastic case given fourth order stationarity.

§ 6.5 Testing normality

White and MacDonald [1980] proposed constructing well known tests for normality using LS residuals and compare different tests in some Monte Carlo experiments. An alternative approach is adopted by Jarque and Bera [1980], by embedding normality into a more general class of distributions - the Pearson family - and using the LM principle. The approach has been extended to other situations by Lee [1982, 1984a, 1984b] and Bera et al [1984].

We follow the approach of Jarque and Bera, and thus the LM test derived below is a simple generalization of theirs that allows for heteroskedasticity under the null and for additional evolution of the conditional third and fourth moments. The Pearson family pdf for \( y_t \) conditional on \( \mathcal{F}_t \) is given by

\[
\begin{align*}
    f(y_t | \mathcal{F}_t) &= \exp \left\{ \varphi_t(u_t) \right\} / \phi_t, \\
    \varphi_t(u_t) &= \int \frac{b_{1t}(b_{1,\theta}) - u_t}{h_t(\theta) - b_{1t}(b_{1,\theta})u_t + b_{2t}(b_{2,\theta})u_t^2} \, du_t, \\
    \phi_t &= \int_{-\infty}^{\infty} \exp \left\{ \varphi_t(u_t) \right\} \, du_t,
\end{align*}
\]

where

\[
\begin{align*}
    b_{1t}(b_{1,\theta}) &= b_{1t}(b_{1,\theta}) - u_t, \\
    b_{2t}(b_{2,\theta}) &= b_{2t}(b_{2,\theta})u_t^2.
\end{align*}
\]
and the $b_{jt}$ are measurable functions of $\mathcal{F}_t$ such that $b_{jt} = 0$ when $b_j = 0$, $j = 1, 2$. This parameterization is similar to the linear-in-$\alpha$ structure for the variance, where $\alpha = 0$ implies homoskedasticity. The distributional parameter $b_{1t}$ is closely related to symmetry because $b_{1t} = h_t(\theta)^{1/2} S_k(y_t)$, $S_k$ being Pearson’s measure of skewness (Kendall and Stuart [1968], pp.85 and 149). Given symmetry, $b_{2t}$ is a monotonically increasing function of the kurtosis measure $\gamma_2(y_t)$ (fourth cumulant), and has the same sign. Therefore, symmetry implies $b_{1t} = 0$, and adding mesokurtosis implies $b_{2t} = 0$. It is easily seen that $b_{1t} = b_{2t} = 0$ results in the conditional normal distribution. These relations between distributional parameters and the third and fourth moments suggests some plausible parameterizations for the $b_{jt}$ in coherency with the specification of the first two moments. For example, in the ARCH model we might propose $b_{1t} = b_{1t}(u_{t,j}^3; j > 0)$ and $b_{2t} = b_{2t}(u_{t,j}^4; j > 0)$.

If we define $b = (b_1', b_2')'$ and $\xi = (\theta', b')'$, the log-likelihood function is

$$L(\xi) = \sum_{t=1}^{T} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{T} \frac{1}{T} \sum_{t=1}^{T} \log \phi_t(\xi).$$

The null hypothesis of normality in (1) can then be expressed as

$$H_0 : b_1 = 0 \text{ and } b_2 = 0,$$

and if we denote by $n_j$ the dimension of $b_j$ and $r_{jt} = r_{jt}(\xi) = \partial b_{jt}/\partial b_j$, $j = 1, 2$, and follow the appendix to Bera and Jarque [1982] we find that under $H_0$

$$d_{1t} = \frac{\partial L_t}{\partial b_1} = r_{1t}(h_t^{-1} u_t - \frac{1}{T} h_t^{-2} u_t^2),$$

and

$$d_{2t} = \frac{\partial L_t}{\partial b_2} = \frac{1}{T} r_{2t}(h_t^{-2} u_t^4 - 3),$$

The LM statistic is based on the subvector of the score $d = (d_1', d_2')'$

$$= \sum_{t=1}^{T} d_t,$$

where $d_t = (d_{1t}', d_{2t}')'$, which shows that this test is an efficiency test as those considered in § 5.3. We then have

**Theorem 6.11.** Under the assumptions of Theorem 5.10,
\[ s_s = \frac{3}{2} T \hat{d}_1' \left[ T^{-1} \hat{R}_1' \hat{Q}^{-1} \hat{R}_1 \right]^{-1} \hat{d}_1 \to \chi^2_{n_1}, \]

and

\[ s_k = T \hat{d}_2' \left[ 6 T^{-1} \hat{R}_2' \hat{R}_2 - \frac{9}{4} T^{-1} \hat{R}_2' \hat{Q}^{-1} \hat{S} \hat{V}(\hat{\theta}) T^{-1} \hat{S}' \hat{Q}^{-1} \hat{R}_2 \right]^{-1} \hat{d}_2 \to \chi^2_{n_2}, \]

under the null hypothesis of conditional normality, where \( R_j = (r_{j1}, \ldots, r_{jT})' \), \( j = 1, 2 \), \( \hat{V}(\hat{\theta}) \) is a consistent estimator of \( V(\theta) \), and all evaluations are made at \( \hat{\theta} \) under \( H_0 \). Moreover \( s_s \) and \( s_k \) are asymptotically independent and the LM test for normality is given by

\[ s_n = s_s + s_k \to \chi^2_{n_1 + n_2}, \]

under \( H_0 \).

**Proof:** In the notation of § 5.3, let \( m_a = m(\hat{\Omega}^{-1} \hat{R}_1, 1, 0 ; \hat{\theta}) \), \( m_b = m(\hat{\Omega}^{-2} \hat{R}_1, 3, 0 ; \hat{\theta}) \), and \( m_c = m(\hat{\Omega}^{-2} \hat{R}_2, 4, 0 ; \hat{\theta}) \). Let \( V_j = \lim_{T \to \infty} \text{var} \left[ T^{1/2} m_j \right] \), and \( V_{ij} = \lim_{T \to \infty} \text{cov} \left[ T^{1/2} m_i, T^{1/2} m_j \right] \), for \( i, j = a, b, c \). From Theorem 5.10 and using (5.35), (5.37) and (5.38) we obtain

\[ V_a = 6 \mathcal{E}(T^{-1} R_1' \Omega^{-1} R_1)^{-1} - 6 \mathcal{E}(T^{-1} R_1' \Omega^{-1} X) V(\hat{\beta}) \mathcal{E}(T^{-1} X' \Omega^{-1} R_1), \]  \( - (36a) \)

\[ V_b = 15 \mathcal{E}(T^{-1} R_1' \Omega^{-1} R_1)^{-1} - 9 \mathcal{E}(T^{-1} R_1' \Omega^{-1} X) V(\hat{\beta}) \mathcal{E}(T^{-1} X' \Omega^{-1} R_1), \]  \( - (36b) \)

and

\[ V_c = 96 \mathcal{E}(T^{-1} R_2' R_2)^{-1} - 36 \mathcal{E}(T^{-1} R_2' \Omega^{-1} S) V(\hat{\theta}) \mathcal{E}(T^{-1} S' \Omega^{-1} R_2), \]  \( - (37) \)

and from Theorem 5.12 we get

\[ V_{ab} = 3 \mathcal{E}(T^{-1} R_1' \Omega^{-1} R_1)^{-1} - 3 \mathcal{E}(T^{-1} R_1' \Omega^{-1} X) V(\hat{\beta}) \mathcal{E}(T^{-1} X' \Omega^{-1} R_1), \]  \( - (38) \)

\[ V_{ac} = -6 \mathcal{E}(T^{-1} R_1' \Omega^{-1} X) V(\hat{\theta}) \mathcal{E}(T^{-1} S' \Omega^{-1} R_2), \]  \( - (39a) \)

and

\[ V_{bc} = -18 \mathcal{E}(T^{-1} R_1' \Omega^{-1} X) V(\hat{\theta}) \mathcal{E}(T^{-1} S' \Omega^{-1} R_2). \]  \( - (39b) \)

Therefore, from (36) and (38),

\[ V(\hat{d}_1) = V_a + \frac{1}{9} V_b - \frac{2}{3} V_{ab} = \frac{2}{3} \mathcal{E}(T^{-1} R_1' \Omega^{-1} R_1), \]
while $V(\hat{d}_2)$ follows from $\hat{d}_2 = \frac{1}{4} m_e$. Independence is established using (39) in

$$\text{cov}[\hat{d}_1, \hat{d}_2] = \frac{1}{4} V_{ac} - \frac{1}{12} V_{bc} = 0,$$

and the asymptotic distribution of the statistics under $H_0$ follows from Corollary 5.11.

The theorem provides three test-statistics to assess normality: $s_s$ emphasizes departures from symmetry, $s_k$ emphasizes on departures from mesokurtosis, and $s_n$ is an omnibus test in the third and fourth moments. The theorem is sufficiently general to allow for different parameterizations of $b_{1t}$ and $b_{2t}$ reflecting alternative propositions about conditional skewness and kurtosis. However, if the objective of the researcher is simply that of producing a diagnostic to evaluate normality, it may appear that the specification of such functions consumes too much time and this is a drawback to the use of the test. For this reason, it may be of interest to have a standard statistic which gives a good indication of whether more careful thought should be given to non-normality before proceeding with further inference. To produce such a standard statistic we make $b_{1t}$ and $b_{2t}$ constants, so that $r_{1t} = r_{2t} = 1$ and $n_1 = n_2 = 1$, and we have

**Corollary 6.12.** Under the assumptions of Theorem 6.11 with $b_{1t} = b_1$ and $b_{2t} = b_2$,

$$s_s = \frac{3}{2} \sum_{t=1}^{T} (\hat{h}_t^{-1} \hat{u}_t - \frac{1}{3} \hat{h}_t^{-2} \hat{u}_t^2)^2 \quad \Rightarrow \chi^2,$$

and
under the null hypothesis of conditional normality, where all evaluations are
made at \( \hat{\theta} \) under \( H_0 \) and \( \hat{V}(\hat{\theta}) \) is a consistent estimator of \( V(\hat{\theta}) \). Moreover \( s_s \)
and \( s_k \) are asymptotically independent and the LM test for normality is

\[
s_n = s_s + s_k \xrightarrow{d} \chi^2_2,
\]

under \( H_0 \).

**Proof:** Set \( r_{1t} = r_{2t} = 1 \) in Theorem 6.11.

This provides a test which does not require the specification of any
moment further than the variance, being analogous to the Jarque-Bera
statistic except for the heteroskedasticity under the null hypothesis. It is easy
to see that it is indeed the Jarque-Bera test when \( h_t \) is constant for all \( t \).
Making the \( b_{jt} \) constants does not affect the size of the test because they are
indeed constants under the null.

In the linear-in-\( \alpha \) case the statistic \( s_k \) may be simplified because using \( h_t = z_t'\alpha \) we get

\[
T^{-1} \sum_{t=1}^{T} h_t^{-1} s_t = T^{-1} \sum_{t=1}^{T} h_t^{-2} s_t z_t' \alpha \xrightarrow{d} 2 ( \text{ Cov}(\alpha, \alpha') \text{ Cov}(\alpha)\alpha )
\]

and since \( \hat{V}(\hat{\theta}) = \hat{\theta}(\hat{\theta})^{-1} \) and \( ( \text{ Cov}(\alpha, \alpha') \text{ Cov}(\alpha)\alpha ) \) the variance of \( \hat{d}_2 \) may
be simply expressed as \( (6 - 9 \alpha' \text{ Cov}(\alpha)\alpha) \). By using \( J_{aa} = \mathbb{E} \left( T^{-1} \sum_{t=1}^{T} h_t^{-2} z_t z_t' \right) \), we
can then express the test-statistic as

\[
s_k = \frac{ \left[ T^{-1} \sum_{t=1}^{T} ( \hat{h}_t^{-2} \hat{u}_t^4 - 3 ) \right]^2 }{ 24 \left[ 4 - 3 \alpha' ( T^{-1} \sum_{t=1}^{T} \hat{h}_t^{-2} \hat{z}_t \hat{z}_t' ) \alpha \right] }
\]
When the information matrix is block diagonal, then $J_{\alpha\alpha} = V(\hat{\alpha})^{-1}$ and the variance is easily estimated from regression output for the estimation of $\alpha$.

Bera et al. [1984] note that some members of the Pearson family violate regularity conditions which are required for the asymptotic properties of ML, such as having supports depending on the parameters. Then, no claim can be made about optimal properties for the LM test. The distribution under the null, however, is not affected.

§ 6.6 Some comments on the Monte Carlo evidence

We have conducted Monte Carlo experiments to assess the performance of the LM tests discussed in this Chapter in small to moderate samples for the cases of autocorrelation and normality, and the results are discussed in the next two subsections. The results are presented in Tables 6.1 - 6.6 which have the same format than Tables 5.6 - 5.8. In Tables 6.1 - 6.3 the column headings are the values of the first order autocorrelation coefficient $\rho$.

§ 6.6.1 LM tests for autocorrelation

Tables 6.1 - 6.3 report the power (proportion of rejections using the asymptotic distribution at the 5% level of significance) of tests for first order autocorrelation as the autocorrelation coefficient $\rho$ is varied from 0 (no autocorrelation) to 0.8. Two alternative tests have been considered: the LM test $s_{mac}$ given in Corollary 6.2, and a consistency test which ignores the variable additions to the auxiliary variance equations, though it still uses a double-length auxiliary regression and evaluates all functions at the MLE. The latter test is denoted $s_{mac}$ in the tables and the objective of including it was to form an idea of whether we should expect substantial loss in power by ignoring the misspecification induced in the conditional variance by the specification error in the conditional mean.
The results for the Poisson-N model are reported in Table 6.1. The size of the test is reasonable for the smaller samples, but it increases in the samples of size 100 and 200. There seems to be a tendency of test size to increase with sample size and this fact is hard to explain under the null hypothesis. Source programs were checked thoroughly to see if some other source of misspecification was being induced in the DGP, but we did not find any, and we would expect this effect to disappear as the sample size is further increased. However, we did not conduct experiments with larger samples. The power of the tests is good and there does not seem to be any important loss in power when ignoring the induced misspecification in the variance equation. In fact, the $s_{wac}$ test appears to have more power unless the autocorrelation is very strong ($\rho = 0.8$).

Powers based on the asymptotic distribution are reported for the mild and strong ARCH models in Tables 6.2 and 6.3, respectively. Test sizes are relatively close to the nominal size, except for the smallest sample, and the 'power' effect under the null found in the Poisson-N model is not present here. Power is good, and the most remarkable feature of the results is that the non-LM test dominates the LM test. This fact cannot be attributed to differences in size of the tests because these differences are small, and though unaccounted autocorrelation might be confused with ARCH effects, we do not find that this is a convincing explanation for this effect and the matter deserves further study.

§ 6.6.2 LM tests for normality

In Tables 6.4 - 6.6 we present power calculations (proportion of rejections using 5% significance points from the asymptotic distribution) for the LM tests for non-normalities in the Poisson-N, and mild and strong ARCH models, respectively. The asymptotically independent components for symmetry and
kurtosis are presented separately, and we also include figures showing the performance of the Jarque-Bera [1980] test which is constructed under the assumption of homoskedasticity under the null hypothesis.

The component of the LM test in the third moment differs from the efficiency tests reported in § 5.4.2 in essentially two ways: the component in third powers of $u_t$ is weighted by $h_t^2$ rather than by $h_t^{-1}$, and an additional element has been introduced in $h_t^{-1} u_t$ to achieve asymptotic independence with the component of the LM test in the fourth moment.

The size of the test is in the three models smaller than the nominal size thus producing a conservative test. This effect is also present in the efficiency tests for the ARCH model, but not in the Poisson-N model. The superiority of $s_3$ over the efficiency tests for symmetry of section § 5.4.2, though not large in general, is clear in both ARCH models, but the $t$ and Cauchy distributions seem to be better detected by the efficiency tests in the Poisson-N model, with a mixed result for the $\chi^2$ distribution. It appears, however, that these results in the Poisson-N model can be attributed to the difference in the size of the tests and that the LM test would in fact dominate the efficiency tests of the previous Chapter.

The component of the LM test in the fourth moment is very similar to the efficiency tests considered in § 5.4.2. There we considered a test based on variance residuals and a test based on a mix of mean and variance residuals. The LM test completes the picture by providing a test based on mean residuals only. The performance of the tests is not substantially different, but nevertheless the LM test dominates the efficiency tests in the three models. In terms of size there is no major difference between the efficiency and LM tests.

The LM test for normality combines independent components, and thus it is in this test that we would expect a clear dominance over the omnibus tests considered in § 5.4.2. This dominance is clear in both ARCH models, while in
the Poisson-N model we have an exception in the $t$ distribution. But here again this effect can be essentially attributed to the different size of the LM and efficiency tests and thus $s_n$ seems to be a more powerful option to test for non-normalities.

The Jarque-Bera test is denoted JB in the tables, and it clearly shows the effect of ignoring heteroskedasticity when testing normality, which may result in substantial over-rejection of true hypotheses.

**TABLE 6.1.- REJECTION FREQUENCIES OF LM TESTS FOR AUTOCORRELATION IN THE POISSON-N MODEL.**

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$T$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_{mac}$</td>
<td>20</td>
<td>0.064</td>
<td>0.086</td>
<td>0.166</td>
<td>0.312</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.040</td>
<td>0.242</td>
<td>0.498</td>
<td>0.796</td>
<td>0.894</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.094</td>
<td>0.476</td>
<td>0.864</td>
<td>0.986</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.132</td>
<td>0.736</td>
<td>0.990</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>$s_{ sac}$</td>
<td>20</td>
<td>0.062</td>
<td>0.174</td>
<td>0.264</td>
<td>0.346</td>
<td>0.436</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.060</td>
<td>0.392</td>
<td>0.612</td>
<td>0.786</td>
<td>0.852</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.108</td>
<td>0.614</td>
<td>0.892</td>
<td>0.984</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.136</td>
<td>0.826</td>
<td>0.990</td>
<td>0.998</td>
<td>0.998</td>
</tr>
</tbody>
</table>

**TABLE 6.2.- REJECTION FREQUENCIES OF LM TESTS FOR AUTOCORRELATION IN THE ARCH MODEL I.**

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$T$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_{mac}$</td>
<td>20</td>
<td>0.048</td>
<td>0.082</td>
<td>0.086</td>
<td>0.126</td>
<td>0.166</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.058</td>
<td>0.114</td>
<td>0.242</td>
<td>0.404</td>
<td>0.442</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.056</td>
<td>0.236</td>
<td>0.532</td>
<td>0.762</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.060</td>
<td>0.398</td>
<td>0.868</td>
<td>0.958</td>
<td>0.970</td>
</tr>
<tr>
<td>$s_{ sac}$</td>
<td>20</td>
<td>0.086</td>
<td>0.132</td>
<td>0.160</td>
<td>0.294</td>
<td>0.344</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.056</td>
<td>0.208</td>
<td>0.492</td>
<td>0.780</td>
<td>0.846</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.052</td>
<td>0.390</td>
<td>0.858</td>
<td>0.982</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.052</td>
<td>0.686</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
TABLE 6.3 - REJECTION FREQUENCIES OF LM TESTS FOR AUTOCORRELATION IN THE ARCH MODEL II.

<table>
<thead>
<tr>
<th>T</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_{mac}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.082</td>
<td>0.090</td>
<td>0.124</td>
<td>0.180</td>
<td>0.188</td>
</tr>
<tr>
<td>50</td>
<td>0.058</td>
<td>0.140</td>
<td>0.386</td>
<td>0.473</td>
<td>0.478</td>
</tr>
<tr>
<td>100</td>
<td>0.054</td>
<td>0.276</td>
<td>0.640</td>
<td>0.806</td>
<td>0.778</td>
</tr>
<tr>
<td>200</td>
<td>0.062</td>
<td>0.470</td>
<td>0.914</td>
<td>0.970</td>
<td>0.934</td>
</tr>
<tr>
<td>$s_{sac}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.084</td>
<td>0.104</td>
<td>0.142</td>
<td>0.256</td>
<td>0.282</td>
</tr>
<tr>
<td>50</td>
<td>0.048</td>
<td>0.164</td>
<td>0.476</td>
<td>0.572</td>
<td>0.662</td>
</tr>
<tr>
<td>100</td>
<td>0.054</td>
<td>0.334</td>
<td>0.716</td>
<td>0.876</td>
<td>0.904</td>
</tr>
<tr>
<td>200</td>
<td>0.048</td>
<td>0.522</td>
<td>0.942</td>
<td>0.984</td>
<td>0.974</td>
</tr>
</tbody>
</table>

TABLE 6.4 - REJECTION FREQUENCIES OF LM TESTS FOR NORMALITY IN THE POISSON-N MODEL.

<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>U</th>
<th>$t_{15}$</th>
<th>$t_5$</th>
<th>$\beta$</th>
<th>$\chi^2$</th>
<th>LN</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_s$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.002</td>
<td>0.006</td>
<td>0.030</td>
<td>0.098</td>
<td>0.116</td>
<td>0.504</td>
<td>0.330</td>
<td>0.796</td>
</tr>
<tr>
<td>50</td>
<td>0.018</td>
<td>0.014</td>
<td>0.080</td>
<td>0.178</td>
<td>0.376</td>
<td>0.914</td>
<td>0.866</td>
<td>0.964</td>
</tr>
<tr>
<td>100</td>
<td>0.046</td>
<td>0.074</td>
<td>0.110</td>
<td>0.246</td>
<td>0.626</td>
<td>0.990</td>
<td>0.956</td>
<td>0.986</td>
</tr>
<tr>
<td>$s_k$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.018</td>
<td>0.004</td>
<td>0.072</td>
<td>0.276</td>
<td>0.000</td>
<td>0.472</td>
<td>0.412</td>
<td>0.984</td>
</tr>
<tr>
<td>50</td>
<td>0.038</td>
<td>0.774</td>
<td>0.159</td>
<td>0.557</td>
<td>0.994</td>
<td>0.908</td>
<td>0.944</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.982</td>
<td>0.268</td>
<td>0.820</td>
<td>0.998</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>$s_n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.014</td>
<td>0.008</td>
<td>0.076</td>
<td>0.260</td>
<td>0.187</td>
<td>0.562</td>
<td>0.462</td>
<td>0.984</td>
</tr>
<tr>
<td>50</td>
<td>0.036</td>
<td>0.280</td>
<td>0.167</td>
<td>0.529</td>
<td>0.932</td>
<td>0.968</td>
<td>0.964</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>0.068</td>
<td>0.978</td>
<td>0.248</td>
<td>0.788</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>JB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.201</td>
<td>0.403</td>
<td>0.218</td>
<td>0.347</td>
<td>1.000</td>
<td>0.168</td>
<td>0.046</td>
<td>0.572</td>
</tr>
<tr>
<td>50</td>
<td>0.343</td>
<td>0.996</td>
<td>0.419</td>
<td>0.666</td>
<td>1.000</td>
<td>0.646</td>
<td>0.216</td>
<td>0.968</td>
</tr>
<tr>
<td>100</td>
<td>0.452</td>
<td>1.000</td>
<td>0.546</td>
<td>0.878</td>
<td>1.000</td>
<td>0.942</td>
<td>0.378</td>
<td>1.000</td>
</tr>
</tbody>
</table>
### TABLE 6.5 - REJECTION FREQUENCIES OF LM TESTS FOR NORMALITY IN THE ARCH MODEL I.

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_a$</td>
<td>$0.026$</td>
<td>$0.022$</td>
<td>$0.028$</td>
</tr>
<tr>
<td>$S_k$</td>
<td>$0.012$</td>
<td>$0.018$</td>
<td>$0.018$</td>
</tr>
<tr>
<td>$S_n$</td>
<td>$0.016$</td>
<td>$0.016$</td>
<td>$0.032$</td>
</tr>
<tr>
<td>$J_B$</td>
<td>$0.050$</td>
<td>$0.076$</td>
<td>$0.138$</td>
</tr>
</tbody>
</table>

### TABLE 6.6 - REJECTION FREQUENCIES OF LM TESTS FOR NORMALITY IN THE ARCH MODEL II.

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_a$</td>
<td>$0.024$</td>
<td>$0.018$</td>
<td>$0.022$</td>
</tr>
<tr>
<td>$S_k$</td>
<td>$0.010$</td>
<td>$0.008$</td>
<td>$0.018$</td>
</tr>
<tr>
<td>$S_n$</td>
<td>$0.016$</td>
<td>$0.010$</td>
<td>$0.018$</td>
</tr>
<tr>
<td>$J_B$</td>
<td>$0.174$</td>
<td>$0.444$</td>
<td>$0.756$</td>
</tr>
</tbody>
</table>
In the preceding chapters we have worked with the implicit assumption that the mean equation is a conventional (possibly nonlinear) regression with heteroskedastic disturbances. By conventional we mean that $\mu_t$ is a function of observables and (unknown) constant parameters. In such case the derivative $x_t = \partial \mu_t / \partial \beta$ is easily calculated for given values of $\beta$, or further it is a direct observable if $\mu_t$ is linear in $\beta$. In this chapter and the next we allow for certain forms of unobservables. The estimation and evaluation problem depicted in Chapters 3 to 6 is then complicated by the fact that we require to extract the signal, use two-stage estimation, or embed the problem into a more general model, depending on the circumstances.

The extension that we introduce here to the basic model of Chapter 2 is a risk measure affecting the conditional mean. In recent years many authors have used risk variables to explain economic phenomena. The way of measuring risk, although essentially presented in the form of a second moment from a distribution, has been diverse. Pagan and Ullah [1986] have classified the alternative proposals of measuring risk into four types, analyzing the relative merits of each of them: moving average measures (Klein [1977], Ibrahim and Williams [1978], Minford and Brech [1978], Gylfason [1981]); relative price measures (Pagan et al [1983]); survey measures (Levi and Makin [1979]); and parameterized measures (Hansen and Hodrick [1983], Domowitz and Hakkio [1985], Bollerslev et al [1985], Engle et al [1987]). Pagan [1984b] showed that when the risk variable is an expectation the simple strategy of substituting it by an observable random variable will almost invariably produce
inconsistent estimates, and he proposed alternative consistent IV estimators. Using a result similar to Lemma 3.3 he also proved that using parametric predictors would not affect the asymptotic distribution of the IV estimators. Pagan and Ullah [1986] have further considered this problem and extended the IV estimator to the case where the random variable is estimated nonparametrically, though they note that the rate of convergence will be slower than in the parametric case and therefore very large samples may be needed to produce good results.

We concentrate on the case where the risk measure is parameterized and show that, provided there exists a root-T consistent estimator of the parameters underlying the risk measure, there is no need to resort to IV estimation. The main argument is developed in a likelihood framework but the results can be related to the more general setup in which the pdf is not given a specific form and the information is extracted by means of orthogonality conditions from each moment. The basic model of Chapter 2 is reformulated to include parametric risk measures in section § 7.1. Parametric risk models are then classified into y-risk models (the risk measure is a function of the conditional variance of \( y_t \) itself), and x-risk models (the risk measure is a function of the conditional moments of other variables), and simple tests for the presence of such risk effects are derived. The problems of estimating and diagnostic testing are analyzed in section § 7.2 for y-risk models and in section § 7.3 for x-risk models.

§ 7.1 Parametric risk models and testing for risk effects

The conditional mean is now modelled as

\[
E [ y_t | \mathcal{F}_t ] = \mu_t ( \beta ; r_t , \pi )
\]  

-(1)
where \( r_t \in \mathcal{F}_t \) is a risk measure, parameterized as a function of the vector \( \pi \in \Pi \). Although more general forms can be allowed, the parameterization can be thought of as essentially linear in \( r_t \), so that by partitioning \( \beta = (\beta_1', \beta_2')' \) the conditional mean is not a function of \( r_t \) when \( \beta_2 = 0 \). The conditional variance is, accordingly,

\[
\text{var} \left[ y_t \mid \mathcal{F}_t \right] = h_t(\theta, \pi),
\]

- (2)

and it is not a direct function of \( r_t \) but depends on \( \pi \) indirectly through \( \mu_{t-j} \), \( j \geq 0 \). If conditional normality is assumed, (1) and (2) are subsumed into

\[
y_t \mid \mathcal{F}_t \sim N \left[ \mu_t(\beta; r_t, \pi), h_t(\theta, \pi) \right].
\]

- (3)

The risk measure is taken to be a function of the conditional moments of some random variable. We define a y-risk model when the risk measure is parameterized as a function of \( h_t = \text{var} \left[ y_t \mid \mathcal{F}_t \right] \), and we define an x-risk model when the risk measure is parameterized as a function of the conditional moments of \( x_t^* \) (the set of conditioning variables). The main example of y-risk is the ARCH-M model of Engle et al [1987], Domowitz and Hakkio [1985] and Bollerslev et al [1985], while Hansen and Hodrick [1983] provide an example of an x-risk model.

We restrict y-risk models to use the conditional variance of \( y_t \) because this is the usual measure of risk and also because this is the natural development in our heteroskedastic framework. The y-risk model is characterized by

\[
(\text{y1}) \mu_t(\beta; h_t, \theta) = \mu_t(\theta), \quad \text{and} \quad h_t = h_t(\theta),
\]

or

\[
(\text{y2}) y_t \mid \mathcal{F}_t \sim N \left[ \mu_t(\theta), h_t(\theta) \right],
\]

and the distinctive feature with our previous model is that \( \mu_t \) now depends on the whole parameter vector \( \theta \).
We do not restrict x-risk models to use a conditional variance to measure risk because then our conclusions readily extend to other situations (e.g. parametric cases of generated regressors, Pagan [1984b]). The risk measures are restricted to be functions of the conditional moments of the conditioning variables for simplicity. The x-risk model is characterized by

\[ (C_{0-xr}) \quad \mu_t = \mu_t (\beta ; \theta_x) = \mu_t (\theta, \theta_x), \quad \text{and} \quad h_t = h_t (\theta, \theta_x), \]

or

\[ (C_{0'-xr}) \quad y_t | \mathcal{F}_t \sim N [ \mu_t (\theta, \theta_x), h_t (\theta, \theta_x) ], \]

where \( \theta_x \) is the parameter vector underlying the distribution of \( x_t^* \).

Before undertaking the more complex task of estimating the risk models it may be convenient to test for risk effects (i.e. \( \beta_2 = 0 \)). When \( \beta_2 = 0 \) the model in (3) is the heteroskedastic model of the previous chapters, and therefore the test for risk effects is a simple variable addition test of the sort analyzed in Chapter 6. We then have

**Corollary 7.1.** - Under the assumptions of Theorem 6.1 the LM test for risk effects (\( H_0 : \beta_2 = 0 \) vs \( H_1 : \beta_2 \neq 0 \)) is given asymptotically by \( s_r = 2T \mathbb{R}_0^2 \) from the regression of \( \delta_t \) on \( (x_t, \delta_t)' \) and \( (x_{rt}, \delta_{rt})' \) in the metric of \( \mathbb{S}_t \), where \( x_{rt} = \partial \mu_t / \partial \beta_2 \), \( s_{rt} = \partial h_t / \partial \beta_2 \), and all functions are evaluated at the MLE \( (\hat{\theta}, \hat{\pi}) \) under \( H_0 \). Under local parametric alternatives,

\[ s_r \xrightarrow{d} \chi^2 [\dim (\beta_2) ; \delta' \mathcal{E} \{ T^{-1} G_r' \Sigma^{-1/2} \mathcal{M}_r \Sigma^{-1/2} G_r \} \delta], \]

where \( G_r = (X_r', S_r')', \quad X_r = (x_{r1}, ..., x_{rT})' \), and \( S_r = (s_{r1}, ..., s_{rT})' \).

**Proof:** That the test for \( \beta_2 = 0 \) is a test for variable addition in a heteroskedastic model is evident from (3) and the parameterization of \( \mu_t \) and \( h_t \). When \( r_t = h_t \) this completes the proof because \( \hat{\pi} = \hat{\theta} \). When \( r_t = r_t (\theta_x) \), \( x_t^* \) is weakly exogenous for \( \theta \) under \( H_0 \). Therefore \( f(y_t | \mathcal{F}_t) \) does not contain any
information about \( \theta \) and \( f(x_t^* | \mathcal{F}_t) \) does not contain information about \( \theta \), which allows the test to be constructed from the double-length regression.

If there is evidence for the existence of risk effects we must proceed to the more general risk model and it is convenient to re-assess the assumptions (\( C_0 \)) - (\( C_8 \)) of Chapter 2. First (\( C_0 \)) is replaced by either (\( C_0 \)-yr) or (\( C_0 \)-xr), and we assume symmetry throughout this chapter. Likewise, (\( C_0' \)) is replaced by either (\( C_0' \)-yr) or (\( C_0' \)-xr). Assumptions (\( C_1 \)) and (\( C_2 \)) remain intact, and so does (\( C_3 \)) which imposes smoothness restrictions to the way in which the risk measure enters \( \mu_t \). For proper estimation of \( \theta \), and hence of \( \tau_t \), all assumptions must be extended to the conditional distribution of \( x_t^* \) given \( \mathcal{G}_t \).

Assumptions (\( C_5 \)) and (\( C_6 \)) remain the same except that (\( C_6 \)) must extend to derivatives with respect to the whole of \( \theta \) for the mean function. The identifiability assumption in (\( C_7 \)) is retained as such, but observe that \( X \) has now a more complex structure. In the y-risk model this means that \( \beta \) is identifiable given \( \alpha \) in the mean equation, and \( \alpha \) is identifiable given \( \beta \) in the variance equation. This suffices for the identifiability of \( \theta \) in the full model.

We also preserve (\( C_8 \)) in the y-risk model so that we may produce global efficiency propositions. This weak exogeneity assumption can not hold in the x-risk case because the conditional distribution of \( y_t \) depends on \( \theta_x \) and hence the likelihood is not i-sufficient (see § 6.4). We retain in this case the remaining implicit assumptions in (\( C_8 \)) , namely the proper factorization of the conditional distribution and the i-completeness of \( \mathcal{L}(\theta, \theta_x) \), so that \( \mathcal{L}_x = \mathcal{L}_x(\theta_x) \) does not depend on \( \theta \). Finally, we retain (\( C_4 \)) exactly, but it is worth analyzing the implications on risk models of the existence of at least fourth order moments. Consider a normal y-risk model with

\[
\mu_t = x_t^* \beta_1 + \beta_2 h_t^{1/2},
\]

and assume that second order moments of all variables exist. Then

\[
\mu_y = \mathbb{E}[y_t] = \mathbb{E}[\mu_t] = \mu_x^* \beta_1 + \beta_2 \sigma,
\]
where \( \mu_x = E[ x_t^*] \), and \( E[ h_t^{1/2}] = \sigma \neq t \), and so
\[
y_t - \mu_y = (x_t^* - \mu_x)'\beta_1 + \beta_2 (h_t^{1/2} - \sigma) + u_t.
\]

Now suppose that the moments of \( y_t \) exist up to order \( r - 1 \) and consider
\[
E[(y_t - \mu_y)^r] = E\left[ \sum_{j=0}^{r} \binom{r}{j} (u_t + \beta_2 (h_t^{1/2} - \sigma))^j ((x_t^* - \mu_x)'\beta_1)^{r-j} \right],
\]
which using the weak exogeneity of \( x_t^* \) and iterated expectations can be decomposed into a term whose existence is guaranteed by that of moments of order \( r - 1 \) or lower, and terms in the expectations of \( r \)-th powers of \( (x_t^* - \mu_x)'\beta_1 \) and \( u_t + \beta_2 (h_t^{1/2} - \sigma) \). This shows that for the existence of \( r \)-th moments of \( y_t \), \( x_t^* \) must be \( r \)-th order stationary and \( E[ h_t^r] \) must exist because under normality \( E[u_t^r] \) is proportional to \( E[h_t^{r/2}] \). Thus in this case the risk and non-risk models place similar restrictions on the parameter subspace \( \mathcal{Q} \), but this is dependent on the way in which \( h_t \) enters \( \mu_t \) and the form of the distribution. For example, if \( \mu_t \) is a linear function of \( h_t \) the restrictions on \( \mathcal{Q} \) are essentially 'doubled' because in this case \( E[h_t^{2r}] \) must exist, and then the \( y \)-risk model requires that \( \mathcal{Q} \) be restricted by conditions which are equivalent to those for the existence of \( 2r \)-th moments in the corresponding no-risk model.

Similarly, in the \( x \)-risk model the stationarity requirements on \( x_t^* \) will depend on the way in which its moments appear in \( \mu_t \). In the sections to follow we refer to the assumptions \( (\mathcal{Q}0)-(\mathcal{Q}8) \) with the modifications discussed above as \( (\mathcal{Q}0-yr)-(\mathcal{Q}8-yr) \) or \( (\mathcal{Q}0-xr)-(\mathcal{Q}8-xr) \) for \( y \)-risk and \( x \)-risk models, respectively.

We conclude this section by showing that parametric estimates of a risk measure have, asymptotically, the "strong property" of Pagan and Ullah [1986], provided a root-\( T \) consistent estimator \( \hat{\pi} \) of \( \pi_0 \) is available. The mean equation can be written as
\[
y_t = \mu_t (\beta_0 ; \pi_0) + u_t = \mu_t (\beta_0 ; \hat{\pi}) + \{ u_t - [\mu_t (\beta_0 ; \hat{\pi}) - \mu_t (\beta_0 ; \pi_0)] \}, \quad (4)
\]
so if we consider the two-stage estimator $\tilde{\beta}$ of $\beta$ from the regression of $y_t$ on $\mu_t(\beta; \pi)$, its consistency depends on the condition

$$T^{-1} \sum_{t=1}^{T} x_t(\pi) \{ u_t - [\mu_t(\pi) - \mu_t(\pi_0)] \} \overset{a.s.}{\to} 0. \quad (5)$$

Using the MVT for random functions (Jennrich [1969]) we have

$$\tilde{r}_t = r_t(\pi) = r_t(\pi_0) + \frac{\partial r_t(\pi)}{\partial \pi'} (\pi - \pi_0), \quad (6)$$

and so for fixed $t$, $\tilde{r}_t - r_t \overset{a.s.}{\to} 0$ as $T \to \infty$, but the infinite length of the vector $r = (r_1, \ldots, r_T)'$ is a problem. Nevertheless, consider for simplicity the case of a linear mean in which $\mu_t = x_t' \beta_1 + \beta_2' r_t$, so that $\mu_t(\beta_0; \pi) - \mu_t(\beta_0; \pi_0) = [r_t(\pi) - r_t(\pi_0)]' \beta_2$. Using this and (6) in (5) we get

$$T^{-1} \sum_{t=1}^{T} x_t' \{ u_t - \frac{\partial r_t(\pi_0)}{\partial \pi'} (\pi - \pi_0) \} + o_p(1),$$

which is $o_p(1)$, and therefore (5) holds. The argument is easily extended to the nonlinear case using the MVT to expand (5).

§ 7.2 Estimation and diagnostic testing of y-risk models

§ 7.2.1 Estimation of y-risk models

With an ARCH parameterization for $h_t$ the y-risk model has been applied by Domowitz and Hakkio [1985] to analyze risk premia in the foreign exchange market, by Engle et al [1987] to study risk premia in the term structure of interest rates, and by Bollerslev et al [1985] in a multivariate CAPM model. All of these authors use a conditionally normal model and refer to the regularity conditions of Crowder [1976] or Domowitz and White [1982] for the usual properties of ML estimators. It must be noted, however, that many of their results do not conform to the conditions for existence of fourth order moments and there is still a need to analyze the properties of the estimators under these conditions. The likelihood and the log-likelihood functions of heteroskedastic
models retain their form with the addition of \( y \)-risk (e.g. as in (2.9) and (2.10)). Similarly, redefining \( \tilde{X} = (X, X') \) where \( X' = \partial \mu / \partial \alpha' \) the score is still given by \( d_{\phi}(\theta) = T^{-1} G' \Sigma^{-1} v \), and the information matrix by \( i(\theta) = E \left[ T^{-1} G' \Sigma^{-1} G \right] \), where \( G = (\tilde{X}', S')' \). The form for the derivatives \( \tilde{x}_t \) and \( s_t \) is now more complex due to the feedback between \( \mu_t \) and \( h_t \). Domowitz and Hakkio report a recursive calculation of these quantities while Engle et al and Bollerslev et al prefer to avoid this complication and resort to numerical calculation in conjunction with the Berndt, Hall, Hall and Hausman [1974] algorithm to maximize the likelihood. The possibility of a block-diagonal information matrix is essentially lost because the off-diagonal block of the information matrix is now

\[
J_{\beta \alpha}(\theta) = E \left[ T^{-1} \left( X' \Omega^{-1} X + \frac{1}{2} W' \Omega^{-2} Z \right) \right],
\]

and this will not be zero in general even when \( W = 0 \). In particular this destroys the robustness properties of simple heteroskedasticity and ARCH models studied in Chapter 4.

Another complication of \( y \)-risk models is that no simple initial consistent estimates of \( \theta \) are available. Because \( h_t \) is not observable it cannot be used as a regressor in the mean equation, and ignoring its presence results in inconsistent estimates for \( \beta_1 \). Lemma 3.3 cannot be applied to the squared residuals obtained from the mean equation ignoring \( h_t \) and this in turn prevents us from using LS in the variance equation to estimate the \( h_t \). Nevertheless, the regularity conditions implied by (C.0′-yr) - (C.8-yr) ensure that the MLE \( \hat{\theta} \) obtained from full convergence to the maximum of \( \mathcal{L}(\theta) \) is root-\( T \) consistent (Theorem 3.9). Using \( \hat{\theta} \) to obtain \( \hat{h}_t \) and \( \hat{u}_t \), the conditions of Theorem 3.11 are satisfied and we can still factorize the likelihood as

\[
\mathcal{L}^*(\theta) = \mathcal{L}_m(\theta) + \mathcal{L}_v(\theta),
\]
with the only difference that \( \lambda_m \) now depends on \( \theta \) and not only on \( \beta \). The question is whether the likelihood factorization may still be useful to separate the information contributed by each moment because \( \lambda_m(\theta) \) still depends on the unobservable \( h_t \). Theorem 3.11 permits the use of \( \tilde{h}_t \) in place of \( h_t \) in the variance of the equation, but not in the regression function. Substitute \( \tilde{h}_t \) in \( \mu_t \) to obtain

\[
y_t = \mu_t(\beta_0; \tilde{h}_t) + \{ u_t - [\mu_t(\beta_0; \tilde{h}_t) - \mu_t(\beta_0; h_t)] \},
\]

and the GLS estimator \( \hat{\beta}_m \) for this equation is such that

\[
T^{1/2}(\hat{\beta}_m - \beta_0) = T^{1/2}(X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \{ u - X_h(\hat{\theta} - \theta_0) \} + o_p(1),
\]

where \( x_{ht} = \frac{\partial u_t}{\partial h_t} = \frac{\partial u_t}{\partial \theta} \), and the MVT has been used to express \( \mu_t(\beta_0; \tilde{h}_t) - \mu_t(\beta_0; h_t) = \frac{\partial \mu_t(\beta_0; \tilde{h}_t)}{\partial \theta}(\hat{\theta} - \theta_0) \), where \( \theta \in [\theta_0, \hat{\theta}] \). Let \( \tilde{\beta}_m \) be the theoretical MLE from maximizing \( \lambda_m(\beta; h_t) \) and observe that \( \tilde{\beta}_m \) is the two-stage estimator (2SE) which maximizes \( \lambda_m(\beta; \tilde{h}_t) \). Then,

\[
T^{1/2}(\tilde{\beta}_m - \beta_0) = T^{1/2}(\hat{\beta}_m - \beta_0) - V(\tilde{\beta}_m) \mathcal{E} \{ T^{-1} X' \Omega^{-1} X_h \} T^{1/2}(\hat{\theta} - \theta_0) + o_p(1),
\]

where the expectation is evaluated at \( \theta_0 \). The term in \( (\hat{\theta} - \theta_0) \) introduces dependence between the estimators from the two factors of the likelihood and hence the use of 2SE's destroys the attractive feature of the factorization, which is to separate the information by moment source. An alternative interpretation can be given to the likelihood factorization by analyzing the mean equation in the same manner as the variance equation was treated in section § 3.2. \( \lambda_m \) is now a function of the whole \( \theta \) vector, and substituting explicitly the parametric form of \( h_t \) in \( \mu_t \) we can estimate the identifiable functions of \( \theta \) in the mean equation. Some examples may clarify this situation. Consider the y-risk model with simple heteroskedasticity and \( r_t = h_t \), so that

\[
\mu_t = \beta_0 + x_t^* \beta_1 + \beta_2 h_t, \quad \text{and} \quad h_t = \alpha_0 + z_t^* \alpha_1,
\]

- (8)
and substituting $h_t$ in the mean equation we can write

$$y_t = b_0 + x_t^* b_1 + z_t^* b_2 + u_t,$$

where $b_0 = \beta_0 + \beta_2 \alpha_0$, $b_1 = \beta_1$, and $b_2 = \beta_2 \alpha_1$. If we assume $(X^*, Z^*)$ to have full column rank $b = (b_0, b_1, b_2')$ is identifiable in the mean equation. In the Amemiya risk model with $r_t = h_t^{1/2}$, so that

$$\mu_t = x_t^* \beta_1 + \beta_2 h_t^{1/2}, \quad \text{and} \quad h_t = \alpha \mu_t^2,$$

substituting $h_t$ in $\mu_t$ and solving for $\mu_t$ gives

$$\mu_t = (1 - \alpha^{1/2} \beta_2)^{-1} x_t^* \beta_1 = x_t^* b_1,$$

where $b_1 = (1 - \alpha^{1/2} \beta_2)^{-1} \beta_1$ is identifiable provided $\alpha^{1/2} \beta_2 \neq 1$. From these examples it is clear that almost certainly $\theta$ will not be identifiable in $\mathcal{L}_m$. Let $\phi_m = \phi_m(\theta)$ be the identifiable functions of $\theta$ in $\mathcal{L}_m = \mathcal{L}_m(\phi_m)$, so that following the same argument as in section § 3.2 and defining $X_\phi = \partial \mu / \partial \phi_m'$ we have

$$T^{1/2}(\hat{\phi}_m - \phi_m^0) = T^{1/2}(X_\phi', \Omega^{-1} X_\phi) - 1 X_\phi', \Omega^{-1} u + o_p(1) \; \overset{d}{\rightarrow} \; \mathcal{N}[0, \Omega \{T^{-1} X_\phi', \Omega^{-1} X_\phi \}^{-1}],$$

where $\hat{\phi}_m$ is the maximizer of $\mathcal{L}_m(\phi_m)$ and $\phi_m^0 = \phi_m(\theta_0)$. Denote by $\phi_v = \phi_v(\theta)$ the identifiable functions in the variance equation and let $\hat{\phi}_v$ be the maximizer of $\mathcal{L}_v(\phi_v)$. It follows that

$$T^{1/2}(\hat{\phi}_v - \phi_v^0) = T^{1/2}(S_\phi', \Omega^{-2} S_\phi)^{-1} S_\phi', \Omega^{-2} \varepsilon + o_p(1) \; \overset{d}{\rightarrow} \; \mathcal{N}[0, 2 \varepsilon \{T^{-1} S_\phi', \Omega^{-2} S_\phi \}^{-1}],$$

where $\phi_v^0 = \phi_v(\theta_0)$ (see § 3.2.2). We can now apply the Factorization Theorem 3.12 provided that the identifiable functions are selected in such a way that $\phi_m = (\gamma, \phi_{m2}')'$ and $\phi_v = (\gamma, \phi_{v2}')'$, where $\gamma$ represents the jointly identifiable functions and $\gamma, \phi_{m2}$ and $\phi_{v2}$ are variation free. Then $\hat{\phi}_m$ and $\hat{\phi}_v$ are asymptotically independent and

$$V(\hat{\gamma})^{-1} T^{1/2}(\hat{\gamma} - \gamma_0) = V(\hat{\phi}_m)^{-1} T^{1/2}(\hat{\phi}_m - \phi_m^0) + V(\hat{\phi}_v)^{-1} T^{1/2}(\hat{\phi}_v - \phi_v^0) + o_p(1),$$

where $\hat{\gamma} = \gamma(\hat{\theta}), \gamma_0 = \gamma(\theta_0), \hat{\phi}_m = (\hat{\phi}_m', \hat{\phi}_{m2}'), \hat{\phi}_v = (\hat{\phi}_v', \hat{\phi}_{v2}'),$ and
\[ V(\gamma)^{-1} = V(\hat{y}_m)^{-1} + V(\hat{y}_v)^{-1}. \]

To illustrate this situation consider the simple risk model in (8) and (9), where \( \alpha = (\alpha_0, \alpha_1)' \) is identifiable in the variance equation. Partition \( b_2 = (b_{21}, ..., b_{2n})' \) and \( \alpha_1 = (\alpha_{11}, ..., \alpha_{1n})' \) and suppose without loss of generality that \( b_{21} \neq 0 \) and \( \alpha_{11} \neq 0 \). Then \( \gamma = b_{21}^{-1} b_2 = \alpha_{11}^{-1} \alpha_1 \) gives the jointly identifiable parameters, while \( \phi_{m2} = (b_0, b_{21}, \beta_1')' \) and \( \phi_{v2} = (\alpha_0, \alpha_{11})' \). For the Amemiya risk model in (10) and (11), substitute (11) in \( h_t \) to obtain

\[ h_t = \alpha (x_t^* b_1)^2 = (x_t^* a_1)^2, \]

where \( a_1 = \alpha^{1/2} b_1 \). Partition \( b_1 = (b_{11}, ..., b_{1n})' \) and \( \alpha_1 = (\alpha_{11}, ..., \alpha_{1n})' \) and suppose without loss of generality that \( b_{11} \neq 0 \). Then \( \gamma = b_{11}^{-1} b_1 = a_{11}^{-1} a_1 \) gives the jointly identifiable parameters, while \( \phi_{m2} = b_{11} \) and \( \phi_{v2} = a_{11} \). \( \alpha \) is recovered from \( a_{11} / b_{11} = \alpha^{1/2} \), but a further restriction is required to identify \( \beta \), and \( \theta_0 \) is not identifiable in this model without such a restriction. A very similar situation arises in a Poisson risk model when \( r_t = h_t = \mu_t \), but because now \( \alpha = 1 \) \textit{a priori}, the jointly identifiable functions are \( b_1 = (1 - \beta^2)^{-1} \beta_1 \).

These are simple examples. In many cases the definition of \( \mu_t \) after substitution of \( h_t \) will be an implicit or recursive relation and finding the jointly identifiable parameters may be a very difficult task. Nevertheless the theoretical MWA structure in \( y \)-risk models is of interest in its own right. If we can separate the information by moment source, we can measure relative contributions to efficiency as in section § 3.4 and the information arising from both moments can be contrasted with a coherency test such as

\[ T \left( \hat{y}_m \right)^2 \left[ V(\hat{y}_m) + V(\hat{y}_v) \right]^{-1} \left( \hat{y}_m - \hat{y}_v \right) \overset{d}{\rightarrow} \chi^2_{\text{dim } \gamma}, \]

under the null hypothesis of correct specification. The noncentrality parameter for the power function under local parametric alternatives can be derived as in Theorem 5.1. Another interesting by-product of the likelihood
factorizations is that there may exist simple consistent initial estimates. These would arise from least-squares on the mean equation to obtain an estimate of \( \phi_m \). This procedure provides adequate residuals in the sense of Lemma 3.3 to estimate \( \phi_v \) by LS in the variance equation, and from \( \phi_m \) and \( \phi_v \) we can recover a consistent initial estimate of \( \theta \).

Finally, note that the arguments of this section under conditional normality can be extended to the non-normal symmetric case. We only need to substitute \( 2h_t^2 \) by \( \kappa_t \) and consider the orthogonality conditions defining \( \phi_m \) and \( \phi_v \) to produce the more general GMM estimators.

§ 7.2 Consistency and efficiency tests for y-risk models (1)

Depending on the complexity of the model the coherency tests suggested above may be hard to construct. A simpler procedure is to resort to the consistency and efficiency tests of Chapter 5 to assess the model in general, and to the LM tests of Chapter 6 to test it against specific departures. The introduction of \( h_t \) as an argument of \( \mu_t \) complicates the calculation of derivatives, the availability of initial consistent estimates, and the factorization of the likelihood. But the consistency and efficiency tests do not depend on any of these characteristics because they are evaluated always at the MLE. Thus all the Theorems of sections § 5.2 and § 5.3 apply equally well to y-risk models with the sole proviso of redefining the nonzero derivative \( \partial \mu_t / \partial \alpha \). The same can be said about the tests for specific directions of Chapter 6, though the specific form of the variables to add to the auxiliary regressions may differ. These tests incorporate more of the structure specific to risk models into the testing procedure than those in Engle, Lilien and Robins [1987] (ELR henceforth), while the tests for risk in Domowitz and Hakkio [1985] are in fact

---

(1) This section relies on joint work with A.R. Pagan, and the example is taken from Pagan and Sabau [1987b].
ELR investigated the influence of risk premia upon the excess holding yield for 60-day Treasury bills. \( y_t \) is the excess holding yield and \( \mu_t(\theta) = x_t' \beta \), where \( x_t \) contains an intercept, the yield differential between 30 and 60-day bills, and a risk premium given by \( \log h_t \). The variable \( h_t \) is modelled as the ARCH process \( \alpha_0 + \alpha_1 \sum_{j=1}^{4} \left( \frac{5}{10} \right) u_{t-j}^2 \). MLE was applied to this model and a number of LM tests were performed to assess the adequacy of the specification.

A range of consistency tests was computed for ELR's preferred equation (eq. (22), p. 402). The ratios of \( m_{h} \) and \( m_{\mu} \) to their asymptotic standard errors were respectively 0.42 and -1.28, which does not suggest that ELR's chosen model is incorrect. However, inspection of the residuals \( \hat{u}_t \) revealed that they were highly non-normal, making a robust estimator of \( \text{var}(m_{h}) \) desirable. Because the term involving the covariance matrix of \( \hat{\theta} \) had already been computed using robust estimators of its components, the only modification needed was the replacement of \( 2\hat{h}_t \) by \( \hat{u}_t^4 - \hat{u}_t^2 \) in the variance formula. When this is done the t-statistic for \( m_{h} \) becomes 2.15, providing some marginal evidence that ELR's selected equation is deficient. A second set of consistency tests was then performed. These involved testing if the coefficient of \( \hat{h}_t \) in the regressions of \( \hat{u}_t \) and \( \hat{e}_t \) against \( \hat{h}_t \) were zero. The regression t-statistics, which are biased in favor of the null hypothesis that the coefficients are zero, were -2.82 and -8.64 respectively, which constitutes stronger evidence against the specification adopted by ELR. Hence, it appears that their decision to model the risk premium as an ARCH process is in error.

What is particularly interesting about this situation is that the consistency tests have disclosed a problem in the specification of ELR's model that was not apparent from the broad range of LM tests that ELR employed in their paper. In fact, the estimated coefficient of \( \hat{h}_t \) in the regression of \( \hat{e}_t \) against \( \hat{h}_t \) was -0.67 instead of zero i.e. the regression of \( \hat{u}_t^2 \) against \( \hat{h}_t \) would yield a coefficient of
0.32 rather than the theoretical value of unity it should have if the ARCH specification was correct. Exactly what one might do to improve the specification of the risk premium for the excess holding yield is an open question, but this example highlights the fact that the consistency tests advocated in this Thesis can provide crucial information about the adequacy of any modelling exercise involving heteroskedastic error terms.

§ 7.3 Estimation and diagnostic testing of x-risk models

§ 7.3.1 Estimation of x-risk models

The conditional mean of $y_t$ is now dependent on the parameter vector $\theta_x$ that characterizes the marginal DGP of $x_t^*$. This opens two alternative ways to conduct inference: we may use the joint pdf for $(y_t, x_t^*)$ to ensure full efficiency, or we may choose to work with the conditional pdf alone plus a consistent estimator of $\theta_x$ in a two-stage estimation (2SE) setup (see Pagan [1986]). The latter is simpler in general at the cost of a loss in efficiency. If we use the joint pdf the model becomes a multivariate y-risk model and can be analyzed as in the previous section, so that there is little to be added of qualitative importance (see also Chapter 9). Because of this we concentrate in this section on the alternative 2SE approach to inference in x-risk models, as this may have wider applicability. The main argument is developed around the conditionally normal model, applying the results of Pagan [1986]. Newey [1984] has shown how 2SE estimators may be interpreted as GMM estimators, and with this result our conclusions extend to the more general cases substituting $2 h_t^2$ by the appropriate conditional kurtosis function.

The log-likelihood function of the conditional model has the same form as in heteroskedastic models, but now it also depends on $\theta_x$ and so we write $\ell = \ell(\theta, \theta_x)$. The log-likelihood function of the marginal model is $\ell_x = \ell_x(\theta_x)$. 
because proper factorization and i-completeness ensure that it is not a function of $\theta$ (see § 6.4). The joint log-likelihood is

$$\mathcal{L}_J (\theta, \theta_x) = \mathcal{L} (\theta, \theta_x) + \mathcal{L}_x (\theta_x) .$$

The score for $\theta$ under conditional normality is $d_\theta(\theta, \theta_x) = T^{-1} G' \Sigma^{-1} \nu$. Define $\theta_F = (\theta', \theta_x')'$, and let $\hat{\theta}_F = (\hat{\theta}_F', \hat{\theta}_x')'$ be the full information MLE obtained from $\mathcal{L}_J (\theta, \theta_x)$. The regularity conditions are extended to the joint DGP, so that

$$T^{1/2} (\hat{\theta}_F - \theta_F^0) \xrightarrow{d} N\left[ 0, J_J(\theta_F^0)^{-1} \right] ,$$

where $\theta_F^0$ is the true value of $\theta_F$, and $J_J(\theta_F)$ is the information matrix of $\mathcal{L}_J$ which using iterated expectations may be seen to partition as

$$J_J(\theta_F) = \begin{pmatrix} J_{\theta \theta} & J_{\theta \theta_x} \\ J_{\theta_x \theta} & J_{\theta_x \theta_x} \end{pmatrix} = \begin{pmatrix} \mathbb{E} [ T^{-1} G' \Sigma^{-1} G ] & \mathbb{E} [ T^{-1} G' \Sigma^{-1} G_x ] \\ \mathbb{E} [ T^{-1} G_x' \Sigma^{-1} G ] & \mathbb{E} [ \frac{\partial^2 \mathcal{L}_x(\theta_x)}{\partial \theta_x \partial \theta_x'} ] \end{pmatrix} ,$$

where $X_x = \partial \mu / \partial \theta_x'$, $S_x = \partial \nu / \partial \theta_x'$, and $G_x = (X_x', S_x')'$. Using $G = (\bar{X}', S')'$ and $\bar{X} = (X, 0)$, $J_{\theta \theta}$ can be further partitioned as

$$J_{\theta \theta} (\theta, \theta_x) = [ J_{\theta \theta \theta} (\theta, \theta_x) , J_{\theta \theta \alpha} (\theta, \theta_x) ]$$

$$= (\mathbb{E} \{ T^{-1} X_x' \Omega^{-1} X + \frac{1}{2} T^{-1} S_x' \Omega^{-2} W \} , \frac{1}{2} \mathbb{E} \{ T^{-1} S_x' \Omega^{-2} Z \} ) ,$$

and using (13) and (14) the covariance matrix of $\hat{\theta}$ is

$$V(\hat{\theta}^*) = \left[ J_{\theta \theta} (\theta_F^0) - J_{\theta \theta_x} (\theta_F^0) J_{\theta_x \theta} (\theta_x^0)^{-1} J_{\theta_x \theta} (\theta_F^0) \right]^{-1} .$$

Now it is not clear whether heteroskedastic models in which the MLE's of $\beta$ and $\alpha$ are independent lose this property with the introduction of the $x$-risk term. When $J_{\theta \theta}$ is block diagonal, $\text{Cov} (\hat{\beta}^*, \hat{\alpha}^*) = 0$ if, and only if, the matrix

$$J_{\beta \beta} (\theta_F^0) J_{\beta \beta} (\theta_x^0) J_{\beta \alpha} (\theta_F^0) = \frac{1}{2} \mathbb{E} (T^{-1} X' \Omega^{-1} X_x + \frac{1}{2} T^{-1} W' \Omega^{-2} S_x ) J_{\theta \theta} (\theta_F^0)^{-1} \mathbb{E} (T^{-1} S_x' \Omega^{-2} Z)$$
equals zero, but whether this condition holds depends on the form in which \( \theta_x \) enters the conditional moments and on the form of the conditioning pdf itself, and even in simple heteroskedastic and ARCH models it will be rarely satisfied without imposing more structure on the problem.

Let us now turn to the 2SE

\[
\tilde{\theta} = \max_{\theta} \mathcal{L}(\theta, \tilde{\theta}_x),
\]

where \( \tilde{\theta}_x \) is some root-T consistent estimator of \( \theta_x \). We assume that \( \tilde{\theta}_x \) is the MLE from the marginal likelihood, so that

\[
T^{1/2} (\tilde{\theta}_x - \theta_x) \rightarrow N \left[ 0, \mathbb{J}_{\theta_x \theta_x}(\theta_x^0)^{-1} \right],
\]

but the results can be extended to other estimators. Under \( C(0',-x) \) we have

\[
\mathbb{E} \left[ d_{\theta}(\theta_0, \theta_0^0) \right] = 0,
\]

and since \( \mathbb{J}_{\theta \theta}(\theta_0^0) \) is positive definite and \( \tilde{\theta}_x \) is consistent, Theorem 1 of Pagan establishes the consistency of \( \tilde{\theta} \). Misspecification of the conditioning model in general produces inconsistency in \( \tilde{\theta} \), but Pagan's Theorem requires only that the condition in (17) hold at the pseudo-true value of \( \theta_x \), and there exists the possibility of finding cases in which misspecification of the conditioning model does not render \( \tilde{\theta} \) inconsistent. The availability of initial consistent estimates of \( \theta \) depends only on the availability of a consistent estimator of \( \theta_x \).

The next issue we must consider is whether ML output from (16) produces correct inferences. Because the likelihood is properly factorized

\[
\text{cov}[\tilde{\theta}_x, d_{\theta}(\theta_0, \theta_x^0)] = 0,
\]

and it follows from Theorem 3 of Pagan that

\[
T^{1/2} (\tilde{\theta} - \theta_0) \rightarrow N \left[ 0, \mathbb{J}_{\theta \theta}^{-1} + \mathbb{J}_{\theta \theta x}^{-1} \mathbb{J}_{\theta x \theta}^{-1} \right].
\]

If \( \theta_x^0 \) is known \( V[\tilde{\theta} | \theta_x^0] = \mathbb{J}_{\theta \theta}(\theta_F)^{-1} \), and this is the covariance matrix reported in ML output, evaluated at \( \tilde{\theta}_x \). In general the second term of \( V(\tilde{\theta}) \) is nonzero and \( V(\tilde{\theta}) - V[\tilde{\theta} | \theta_x^0] \) is positive semidefinite, so that for fixed \( \theta_0 \),
(\tilde{\theta} - \tilde{\theta}_0)' V [\theta|\theta_x^0]^{-1} (\tilde{\theta} - \tilde{\theta}_0) \geq (\tilde{\theta} - \tilde{\theta}_0)' V(\tilde{\theta})^{-1} (\tilde{\theta} - \tilde{\theta}_0),

and hence inferences based on $V [\theta|\theta_x^0]$ lead to over-rejection of the null hypothesis $\theta = \theta_0$. White's [1980b] covariance matrix does not produce the required correction and we need $\tilde{\theta}_x$ to estimate $V(\tilde{\theta})$ correctly.

Turning now to efficiency, we see from (12) that

$$\frac{\partial \mathcal{L}(\theta, \tilde{\theta}_x)}{\partial \theta} = \frac{\partial \mathcal{L}_f(\theta, \tilde{\theta}_x)}{\partial \theta},$$

which is the condition in Theorem 5 of Pagan. But without imposing more structure on the problem the sufficient condition for efficiency of the 2SE ($\gamma_{\theta X} = 0$) does not hold, and $\tilde{\theta}_x$ is strictly efficient relative to $\tilde{\theta}_x$, so that $\gamma_{\theta X} T^{1/2} (\tilde{\theta}_x - \tilde{\theta}_x)$ is not $o_p(1)$. The latter condition is also necessary for efficiency and thus we may conclude that in general $\tilde{\theta}$ is inefficient relative to $\tilde{\theta}^*$. The loss can be evaluated from (15) and (18).

Finally, there is the question of whether the two-stage likelihood can be factorized. Using (4) we get a relation similar to (7), namely,

$$T^{1/2} (\hat{\beta}_m - \beta_0) = T^{1/2} (\hat{\beta}_m - \beta_0) - V(\hat{\beta}_m) \& \{ T^{-1} X'\Omega^{-1} X_x \} T^{1/2} (\tilde{\theta}_x - \theta_x^0) + o_p(1),$$

- (19a)

and by analogous procedure in the variance equation,

$$T^{1/2} (\hat{\theta}_v - \theta_0) = T^{1/2} (\hat{\theta}_v - \theta_0) - \frac{1}{2} V(\hat{\theta}_v) \& \{ T^{-1} S'\Omega^{-2} S_x \} T^{1/2} (\tilde{\theta}_x - \theta_x^0) + o_p(1),$$

- (19b)

or jointly, using (14a),

$$T^{1/2} (\hat{\theta} - \theta_0) = T^{1/2} (\hat{\theta} - \theta_0) - \gamma_{\theta X}(\theta_0, \theta_x^0)^{-1} \gamma_{\theta X}(\theta_0, \theta_x^0) T^{1/2} (\tilde{\theta}_x - \theta_x^0) + o_p(1),$$

- (19c)

where we assume the variance equation identified for (19b), and $\hat{\theta}$, $\hat{\beta}_m$ and $\hat{\theta}_v$ are the estimators assuming $\theta_x$ known, that is,

$$T^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N [ 0, \gamma_{\theta X}(\theta_0, \theta_x^0)^{-1} ],$$
Therefore, $\text{cov} [\tilde{\beta}_m, \tilde{\beta}_r] \neq 0$ and the two-stage likelihood cannot be easily factorized. Observe that the presence of $\tilde{\theta}_x$ in the estimators in (19) is closely related to the form of $g_{\theta_0x}$ in (14), and we may find cases in which adding structure permits the factorization of $\mathcal{L}(\theta, \tilde{\theta}_x)$.

### § 7.3.2 Consistency and efficiency tests for $x$-risk models

It is important to bear in mind that the estimator $\tilde{\theta}$ of $\theta$ may be rendered inconsistent by specification error in either the conditional or the marginal model. One possibility for diagnosis is to consider the joint likelihood, which takes us to the case of a multivariate $y$-risk model and we may proceed as in § 7.2.2 (see also Chapter 9). A second possibility is to evaluate the two models separately, testing first the consistency of $\tilde{\theta}_x$ and its estimated covariance matrix using consistency and efficiency tests, and proceeding then to evaluate the conditional model. A third and simpler possibility is to concentrate the diagnostic testing on $\mathcal{L}(\theta, \tilde{\theta}_x)$ alone. It follows from Proposition 4.1 of Pagan [1986] that consistency tests in this setup are testing the more general version of condition (17) that

$$\mathcal{C} \{ d_{\theta} (\theta_0, \theta_x^*) \} = 0, \quad - (20)$$

where $\theta_x^*$ is the pseudo-true value of $\theta_x$. Because this condition is sufficient for the consistency of $\tilde{\theta}$ we would indeed be testing for inconsistencies arising from specification error in either the conditional or marginal models. One problem with this procedure is that in case of rejection of the null hypothesis we have no information as to which of the models is causing the problem, and in such case we need to assess the marginal model as in the second possibility above. Another problem is the possibility of incorrect inferences arising from the marginal model. If $\tilde{\theta}_x$ is consistent but we are underestimating its covariance matrix, this will induce over-rejection of true hypotheses because
the tests depend on \( V(\tilde{\theta}_x) \). If the model is not rejected the situation would not be reversed by using the correct covariance matrix. To test the conditional model estimated by the two-stage procedure we use the following

**Lemma 7.2** - Suppose a function \( m_a(\theta, \theta_x) \) is such that

\[
(i) \quad T^{1/2} m_a(\hat{\theta}, \theta^0_x) \overset{d}{\rightarrow} N[\psi_a, Q_a],
\]

and

\[
(ii) \quad \mathbb{E} \left\{ \frac{\partial m_a(\theta_0, \theta^0_x)}{\partial \theta'} \right\} \text{ exists and is finite.}
\]

Further assume that \( T^{1/2} (\tilde{\theta}_x - \theta^0_x) \overset{d}{\rightarrow} N[0, V(\tilde{\theta}_x)] \) and \( \tilde{\theta}_x \) is asymptotically independent from \( \hat{\theta} \). Then

\[
T^{1/2} m_a(\tilde{\theta}, \tilde{\theta}_x) \overset{d}{\rightarrow} N[\psi_a, Q_a + M_{\max}(\theta_0, \theta^0_x) V(\tilde{\theta}_x) M_{\max}(\theta_0, \theta^0_x)'],
\]

where \( M_{\max}(\theta, \theta_x) = \mathbb{E} \left\{ \frac{\partial m_a(\theta, \theta_x)}{\partial \theta'} \right\} A(\theta, \theta_x) \), and

\[
A(\theta, \theta_x) = \begin{pmatrix} -J_{\theta \theta}(\theta, \theta_x)^{-1} J_{\theta x}(\theta, \theta_x) & 0 \\ 0 & I \end{pmatrix}.
\]

**Proof:** We use the MVT for random functions (Jennrich [1969]) to expand \( m_a(\hat{\theta}, \tilde{\theta}_x) \) about \((\hat{\theta}', \tilde{\theta}_x')\), getting

\[
T^{1/2} m_a(\tilde{\theta}, \tilde{\theta}_x) = T^{1/2} m_a(\hat{\theta}, \theta^0_x) + \mathbb{E} \left\{ \frac{\partial m_a(\theta, \theta_x)}{\partial \theta'} \right\} T^{1/2} (\tilde{\theta}_x - \theta^0_x) + o_p(1),
\]

where we have used the consistency of \( \hat{\theta} \), \( \tilde{\theta} \) and \( \tilde{\theta}_x \). But using (19c) we can write

\[
T^{1/2} (\tilde{\theta}_x - \theta^0_x) = -J_{\theta \theta}(\theta_0, \theta^0_x)^{-1} J_{\theta x}(\theta_0, \theta^0_x) T^{1/2} (\tilde{\theta}_x - \theta^0_x) + o_p(1),
\]

and substituting this in the above expression and rearranging terms results in

\[
T^{1/2} m_a(\tilde{\theta}, \tilde{\theta}_x) = T^{1/2} m_a(\hat{\theta}, \theta^0_x) + M_{\max}(\theta_0, \theta^0_x) T^{1/2} (\tilde{\theta}_x - \theta^0_x) + o_p(1), \quad -(21)
\]
and the asymptotic distribution of \( T^{1/2} m_a(\tilde{\theta}, \tilde{\theta}_x) \) follows from the distribution of \( \tilde{\theta}_x \) and the independence of \( \tilde{\theta}_x \) and \( \tilde{\theta} \).

The lemma shows how we would be understating the covariance matrix of \( T^{1/2} m_a(\tilde{\theta}, \tilde{\theta}_x) \) when ignoring the fact that \( \theta_x \) is estimated and not known, and we find the use of this lemma simpler to obtain the distribution of consistency and efficiency test-statistics than to cast the 2SE as a GMM estimator (Newey [1984]) and then to reproduce Theorems 5.4 and 5.10 and their corollaries.

For consistency tests we assume
\[
(y_t \mid \mathcal{F}_t \sim \mathcal{N} \left[ \mu_t(\theta, \theta_x, \lambda_T), h_t(\theta, \theta_x, \lambda_T) \right],
\]
where \( \lambda_T = \lambda_0 + T^{-1/2} \delta \) for fixed \( \delta \) and \( \lambda_0 \), and correct specification obtains when \( \delta = 0 \). The consistency statistic is
\[
m(\Phi, \tilde{\theta}; \tilde{\theta}_x) = T^{-1} \sum_{t=1}^{T} m_t(\tilde{\Phi}_t, \tilde{\theta}; \tilde{\theta}_x) = T^{-1} \tilde{\Phi}' \tilde{\psi} ,
\]
where all functions are defined as in § 5.2 and are now evaluated at \( \tilde{\theta} \) and \( \tilde{\theta}_x \).

We then have

**Theorem 7.3.** - Under (i) \( (\mathcal{Q}_0'-xr) \), (ii) the sequence \( \{ \Phi_t \} \), \( \Phi_t \in \mathcal{F}_t \), is such that the function \( g* = (m(\Phi, \theta; \theta_x)', \varphi(\theta, \theta_x)')' \) obeys the regularity, continuity, dominance and mixing conditions in assumptions (1) - (6) of Newey [1985b], and (iii) \( T^{1/2} (\tilde{\theta}_x - \theta_0) \overset{d}{\rightarrow} \mathcal{N} [ 0, V(\tilde{\theta}_x) ] \), where \( \tilde{\theta}_x \) is independent of \( \tilde{\theta} \), then
\[
T^{1/2} m(\Phi, \tilde{\theta}; \tilde{\theta}_x) \overset{d}{\rightarrow} \mathcal{N} [ \psi, Q_{dr} = Q_{\phi} + M_x(\theta_0, \theta_x^0) V(\tilde{\theta}_x) M_x(\theta_0, \theta_x^0)' ] ,
\]
where \( \psi \) and \( Q_{\phi} \) are given in (5.23), and
\[
M_x(\theta_0, \theta_x^0) = - E \{ T^{-1} \Phi' B \} ,
\]
where \( B = B(\theta, \theta_x) = (G, G_x) A(\theta, \theta_x) \), and the expectation is evaluated at \( \theta_0 \) and \( \theta_x^0 \). A consistent estimator for \( Q_{dr} \) is
\[
\tilde{Q}_{\phi} = \tilde{Q}_\phi + T^{-1} \tilde{\Phi}' \tilde{B} \tilde{V}(\tilde{\theta}_x) T^{-1} \tilde{B}' \tilde{\Phi},
\]
where \( \tilde{Q}_\phi \) is given in (5.24), \( \tilde{\Phi} = \Phi(\tilde{\theta}, \tilde{\theta}_x), \) \( \tilde{B} = B(\tilde{\theta}, \tilde{\theta}_x), \) and \( \tilde{V}(\tilde{\theta}_x) \overset{\text{as}}{\rightarrow} V(\tilde{\theta}_x). \)

**Proof:** For known \( \theta_x^0 \) the distribution of \( T^{1/2} m(\tilde{\Phi}, \tilde{\theta}; \theta_x^0) \) is given by (5.23) of Theorem 5.3. Using iterated expectations

\[
E \left[ \frac{\partial m(\Phi, \theta_0^0 ; \theta_x^0)}{\partial \theta_F'} \right] = E \left[ \frac{\partial (T^{-1} \Phi' \nu)}{\partial \theta_F'} \right] = -E \left[ T^{-1} \Phi'(G, G_x) \right],
\]
and (22) follows from Lemma 7.2 while (23) is immediate from Theorem 5.3 and (22b).

There is an abuse of notation in the definition of \( \tilde{B} \), which requires evaluating \( A(\theta, \theta_x) \) at the estimators. To avoid confusion, we define \( A(\tilde{\theta}, \tilde{\theta}_x) \) by substituting the submatrices of the information matrix by their consistent estimates, obtained by deleting the expectations in (14a) and evaluating the functions at the consistent estimators. The test statistic is given in

**Corollary 7.4.** - Under the assumptions of Theorem 7.3,

\[
s = T m(\tilde{\Phi}, \tilde{\theta}; \tilde{\theta}_x)' \tilde{Q}_{\phi} m(\tilde{\Phi}, \tilde{\theta}; \tilde{\theta}_x) \overset{\mathcal{D}}{\rightarrow} \chi^2[ n ; \psi' Q_{\phi}^{-1} \psi ].
\]

**Proof:** Construct the quadratic form in (22a) and substitute the consistent estimate of \( Q_{\phi} \).

An important aspect of Theorem 7.3 is how misspecification in the marginal model affects the test. As long as a pseudo-true value \( \theta_x^* \) exists and \( T^{1/2}(\tilde{\theta}_x - \theta_x^*) \) is normal (e.g. Burguete et al [1982], Domowitz and White [1982], Gourieroux et al [1984a]), this will not show in the "added" element to the test induced by the randomness of \( \tilde{\theta}_x \) (e.g. the second term of the expansion in (21)). Misspecification of the marginal model induces power in the consistency test only if it is strong enough to cause violation of (20). The covariance matrix \( Q_{\phi} \) is under-estimated when \( V(\tilde{\theta}_x) \) is under-estimated. The independence of \( \tilde{\theta}_x \) is not a problem: as long as the likelihood is properly factorized, any estimator
of $\theta_x$ from the marginal model alone is independent of $\hat{\theta}$. The best power for the test is attained using the MLE $\tilde{\theta}_x$. Because of the form of the expansion in (21), the relevant covariance matrix that enters $Q_0$ is $V(\tilde{\theta})$ and not $V(\hat{\theta})$.

When $\theta_x$ is known $V(\tilde{\theta}) = V[\tilde{\theta} \mid \theta_x^0]$, and this matrix is consistently estimated evaluating $V(\tilde{\theta})$ at $\tilde{\theta}$ and $\tilde{\theta}_x$, which we denote as $V(\tilde{\theta} \mid \tilde{\theta}_x)$. The test cannot be calculated from a coefficient of determination because of the extra term in the covariance matrix of the test-statistic induced by the two-stage procedure.

However non-rejection using a $2 T R_0^2$ calculation cannot be reversed by using the correct covariance matrix. LM-type tests for specific departures in the specification of the first two moments obtained are members of the class of consistency tests of Theorem 7.3 and therefore we only need to obtain the correct derivatives to make them operative. Because $\lambda(\theta, \tilde{\theta}_x)$ is not the likelihood of the conditional model we cannot claim optimality for these tests as some power is being lost by the uncertainty about $\theta_x$.

To extend the efficiency tests we introduce the assumption 

$(\mathcal{Q}_0''\text{-xr})$  

*The conditional pdf of* $y_t$ *is* $f(y_t \mid \mathcal{F}_t ; \theta, \theta_x, \lambda_T)$,  

where $\lambda_T = \lambda_0 + T^{-1/2} \delta$ for fixed $\lambda_0$ and $\delta$, and at $\delta = 0$, $f(\cdot \mid \cdot)$ is the normal pdf.

The efficiency tests are defined by the statistic

$$m_e(\tilde{\theta}, s, r; \tilde{\theta}, \tilde{\theta}_x) = T^{-1} \sum_{t=1}^{T} m_{et}(\varphi_t, s, r; \tilde{\theta}, \tilde{\theta}_x) = T^{-1} \sum_{t=1}^{T} \varphi_t [ \bar{u}_t^2 \bar{v}_t - q_t(s,r) ],$$

where all functions are as defined in § 5.3 and are now evaluated at $\tilde{\theta}$ and $\tilde{\theta}_x$.

The basic result is then

**Theorem 7.5.** - Under (i) $(\mathcal{Q}_0''\text{-xr}) - (\mathcal{Q}_7\text{-xr})$, (ii) the sequence $\{\varphi_t\}$, $\varphi_t \in \mathcal{F}_t$, is such that the function $g^*(m_e(\varphi, s, r; \theta, \theta_x), d_0(\theta, \theta_x))'$ obeys the regularity, continuity, dominance and mixing conditions in assumptions (1) - (6) of Newey [1985b], and (iii) $T^{1/2}(\tilde{\theta}_x - \theta_x^0) \overset{d}{\rightarrow} N[0, V(\tilde{\theta}_x)]$, where $\tilde{\theta}_x$ is independent of $\hat{\theta}$, then
\[ T^{1/2} m_e(\tilde{\theta}, s, r; \bar{\theta}, \hat{\theta}_x) \xrightarrow{d} N[\psi_e, Q^{\tilde{\theta}} = Q_{\phi} + M_{xe}(\theta_0, \theta_x^0) V(\theta_x) M_{xe}(\theta_0, \theta_x^0)'], \]

where \( \psi_e \) and \( Q_{\phi} \) are given in Theorem 5.10, and

\[
M_{xe}(\theta_0, \theta_x^0) = \begin{cases} 
\zeta(s + 1, r) \mathcal{C} \left\{ T^{-1} \phi' \Omega^{(s-1)/2} B_1 \right\} & \text{for } s \text{ odd}, \\
\frac{1}{2} \zeta(s, r) \mathcal{C} \left\{ T^{-1} \phi' \Omega^{-1+\pi/2} B_2 \right\} & \text{for } s \text{ even},
\end{cases}
\]

where \( B = B(\theta, \theta_x) \) has been partitioned as \( B = (B_1', B_2')' \), and the expectations are evaluated at \((\theta_0, \theta_x^0)\). A consistent estimator of \( Q^{\tilde{\theta}} \) is

\[
\hat{Q}_{\tilde{\theta}} = \hat{Q}_{\phi} + T^{-1} \hat{\phi} \hat{G}_{rs} \hat{V}(\theta_x) T^{-1} \hat{G}_{rs}' \hat{\phi},
\]

where \( \hat{Q}_{\phi} \) is given in Theorem 5.10, \( \hat{\phi} = (\tilde{\phi}, \hat{\theta}_x) \), \( \hat{V}(\theta_x) \) as \( V(\theta_x) \),

\[
\hat{G}_{rs} = G_{rs}(\tilde{\theta}, \hat{\theta}_x),
\]

and

\[
G_{rs}(\theta, \theta_x) = \begin{cases} 
\zeta(s + 1, r) \Omega^{(s-1)/2} B_1(\theta, \theta_x) & \text{for } s \text{ odd}, \\
\zeta(s, r + 1) \Omega^{-1+\pi/2} B_2(\theta, \theta_x) & \text{for } s \text{ even},
\end{cases}
\]

Proof: For \( \theta_x^0 \) known, the distribution of \( T^{1/2} m_e(\tilde{\theta}, s, r; \bar{\theta}, \hat{\theta}_x^0) \) is given in (5.39). (24b) is obtained as \( M_e \) in (5.38) from the outer product form of \( \mathcal{C} \left\{ T^{-1} \partial m_e/\partial \phi' \right\} \), and then multiplying by \( A(\theta, \theta_x) \). Then (24a) follows from Lemma 7.2, and (25) is a direct consequence of (24b) and the consistency of \( \hat{Q}_{\phi} \).

And we have the test-statistic in

**Corollary 7.6.** - Under the assumptions of Theorem 7.5,

\[ T m_e(\tilde{\theta}, s, r; \bar{\theta}, \hat{\theta}_x) \hat{Q}^{-1}_{\tilde{\theta}} m_e(\tilde{\theta}, s, r; \bar{\theta}, \hat{\theta}_x) \xrightarrow{d} \chi^2 [n; \psi_e, Q^{-1}_{\tilde{\theta}} \psi_e]. \]

Proof: Construct the quadratic form in (24a) and substitute the consistent estimate of \( Q_{\tilde{\theta}} \).

The same comments apply to Theorem 7.5 in relation to the role of \( \hat{\theta}_x \) as in Theorem 7.3. These simple efficiency tests provide the means to assess the adequacy of single conditional moments in \( \ell(\theta, \theta_x) \). For omnibus tests we
also need the covariance matrix between statistics of different moments, which is given in

**Theorem 7.7.** - If \( m_e(\tilde{\theta}_x) = m_0(\tilde{\theta}, s, r; \tilde{\theta}_x) \) and \( m_e^*(\tilde{\theta}_x) = m_0(\tilde{\theta}, s^*, r^*; \tilde{\theta}_x) \) are efficiency statistics satisfying the conditions of Theorem 7.5, then

\[
\text{cov} [m_e(\tilde{\theta}_x), m_e^*(\tilde{\theta}_x)] = \text{cov} [m_e(\tilde{\theta}_x), m_e^*(\tilde{\theta}_x)] + M_{xe}(\theta_0, \theta_x) V(\tilde{\theta}_x) M_{xe}^*(\theta_0, \theta_x)'
\]

where \( \text{cov} [m_e(\tilde{\theta}_x), m_e^*(\tilde{\theta}_x)] \) is given in (5.40) of Theorem 5.12, and \( M_{xe} \) and \( M_{xe}^* \) are the expected matrices of derivatives in (24b) for \( m_e \) and \( m_e^* \), respectively.

**Proof:** The result follows expressing \( m_e(\tilde{\theta}_x) \) and \( m_e^*(\tilde{\theta}_x) \) as in (21) and taking the cross-product expectation using the independence of \( \tilde{\theta} \) and \( \tilde{\theta}_x \).

In section § 5.3 we discussed the possibility of constructing omnibus tests with independent components and concluded that this would require introducing linear combinations of efficiency statistics into the components of the omnibus test. The situation is now further complicated by the additional terms in the covariance matrices \( Q_{re} \) and \( \text{cov} [m_e(\tilde{\theta}_x), m_e^*(\tilde{\theta}_x)] \). But these additional terms are usually annihilated by the same linear combinations that produce independent components in the no-risk case, and the LM-type test for normality against the Pearson family has a structure almost as simple as the corresponding LM test in no-risk heteroskedastic models. This is shown in

**Theorem 7.8.** - Under the assumptions of Theorem 7.7,

\[
s_s = \frac{3}{2} T d_1' [ T^{-1} \tilde{R}_1' \tilde{\Omega}^{-1} \tilde{R}_1 ] d_1 \rightarrow \chi^2_{\alpha_1},
\]

and

\[
s_k = T d_2' [ 6 T^{-1} \tilde{R}_2' \tilde{\Omega}^{-1} \tilde{R}_2 - \frac{9}{4} T^{-1} \tilde{R}_2' \tilde{\Omega}^{-1} \tilde{S} \tilde{V}(\tilde{\theta} | \tilde{\theta}_x) T^{-1} \tilde{S}' \tilde{\Omega}^{-1} \tilde{R}_2 ] d_2 \rightarrow \chi^2_{\alpha_2},
\]
under the null hypothesis of conditional normality, where the $d_j$ and $R_j$ are defined in Theorem 6.11, $\hat{V}(\theta \mid \theta_x) \to V(\theta \mid \theta_x^0)$, and $\hat{V}(\theta_x) \to V(\theta_x)$. Moreover, $s_s$ and $s_k$ are asymptotically independent and the LM-type test for normality against the Pearson family is

$$s_n = s_s + s_k \xrightarrow{d} \chi_{n_1+n_2}^2,$$

under $H_0$.

**Proof:** The proof is almost identical to Theorem 6.11. To account for the two-stage procedure add to (6.36) - (6.39) a term equal to the last term of each equation with opposite sign, replacing $X$ by $B_1$, $S$ by $B_2$, and both $V(\tilde{\beta})$ and $V(\tilde{\theta})$ by $V(\tilde{\theta}_x)$. It is obvious that these terms will behave exactly as their similars. 

We end up this chapter by providing a more practical test that introduces no additional asymmetry and kurtosis (i.e. $b_{1t} = b_1$ and $b_{2t} = b_2$, see § 6.5), in

**Corollary 7.9.** - Under the assumptions of Theorem 7.8 and $b_{1t} = b_{2t} = 1$,

$$s_s = \frac{3 \left[ \sum_{t=1}^{T} \frac{\tilde{h}_t \tilde{u}_t}{\tilde{h}_t^2 + \tilde{u}_t^2} \right]^2}{2 \sum_{t=1}^{T} \tilde{h}_t^{-1}} \xrightarrow{d} \chi_1^2,$$

and

$$s_k = \frac{\left[ \sum_{t=1}^{T} (\tilde{h}_t^2 \tilde{u}_t^4 - 3) \right]^2}{4 T \left[ 24 - 9 T^1 \sum_{t=1}^{T} \tilde{h}_t \tilde{u}_t \tilde{V}(\tilde{\theta} \mid \tilde{\theta}_x) T^{-1} \sum_{t=1}^{T} \tilde{h}_t \tilde{u}_t + 9 T^{-1} \sum_{t=1}^{T} \tilde{V}(\tilde{\theta}_x) T^{-1} \tilde{B}_2 \tilde{V}(\tilde{\theta}_x) T^{-1} \tilde{B}_2' \tilde{\Omega}^{-1} \right]} \xrightarrow{d} \chi_1^2,$$

under the null hypothesis of conditional normality. Moreover $s_s$ and $s_k$ are asymptotically independent and the LM-type test for normality is

$$s_n = s_s + s_k \xrightarrow{d} \chi_2^2,$$
under $H_0$.

**Proof.** Set $r_{1t} = r_{2t} = 1$ in Theorem 7.8.

This result generalizes the Jarque-Bera [1980] test to x-risk models.
The relationship between varying coefficients and heteroskedastic models is a very close one. Indeed, varying coefficient models can be cast as members of the heteroskedastic class we have studied so far. This is hardly surprising if we take account of two facts. First, any unaccounted variation in the mean coefficients becomes part of the error terms. This introduces heteroskedasticity in the model and, depending on the sort of parameter change implicit in the DGP, it also introduces inconsistencies in the estimation of $\theta$. Even when the variation in the mean coefficients is accounted for, heteroskedasticity may arise from randomness in the coefficient evolution. Researchers have been well aware of this for some time and testing procedures for heteroskedasticity have power against coefficient variation and vice versa (e.g. Breusch and Pagan [1979]). The second fact is that the log-likelihood of the conditional model is naturally constructed in terms of prediction errors (Harvey [1984,1985]), and thus the introduction of an extra source of variation can be accommodated within the same framework, as has been noted by Engle and Watson [1985]. The assumptions underlying the parameter variation induce a specific structure into the heteroskedastic model which must be taken into account to derive proper inferences, and regularity conditions may be difficult to interpret. In section § 8.1 we analyze some varying coefficient models and try to assess these conditions in the presence of additional change in the conditional variance. Section § 8.2 considers further tests for parameter stability and diagnostic testing of heteroskedastic varying coefficient models.
Assuming constant parameters has proved very useful in applied econometrics though the limitations it imposes have been a concern of econometricians for a long time (see Nicholls and Pagan [1985]). An important challenge to this assumption and to the utility of using econometric models for policy analysis was put forward by Lucas [1976], who argued that the rationality of economic agents would induce structural changes in the face of alterations in economic policy. The Lucas critique caused a good deal of debate and different sectors of the economics profession derived their own conclusions as to its relevance (see Begg [1982], Sims [1980]). Yet the econometrics literature seems still to be lacking in means to assess and empirically contrast this critique. Engle et al [1983] have proposed a concept of exogeneity which is closely related to the Lucas critique, namely that of superexogeneity, and Engle and Hendry [1986] have provided, to our knowledge, the only proposal for testing this concept. In section § 8.3 we take a different approach to the problem by using a varying coefficient model to test for superexogeneity.

§ 8.1 Varying coefficient and heteroskedastic models

We consider the model

\[ y_t | \mathcal{F}_t, \beta_t \sim N \left( \mu_t^*(\beta_t), h_t^*(\beta_t, \alpha) \right), \]

where we allow for variation only in the mean parameters \( \beta_t \) and keep \( \alpha \) constant. The conditioning information set is now augmented with \( \beta_t \) to cover the possibility that this vector be a random variable conditional on \( \mathcal{F}_t \). If this is the case we need to compound the distributions of \( y_t \) and \( \beta_t \) conditionally on \( \mathcal{F}_t \) to obtain the likelihood function. If \( \beta_t \in \mathcal{F}_t \) (though it may still be a random variable) the likelihood function is directly obtained from the normal pdf above. We assume the mean to be linear in \( \beta_t \), \( \mu_t^* = x_t^* \beta_t \), \( x_t^* \in \mathcal{F}_t \), to keep matters simple and enable the derivation of expressions that avoid repeated use of
linear approximations. The conditional variance \( h_t^* \) may in principle depend on \( \beta_t \), but when \( \beta_t \notin \mathcal{F}_t \) explicit solutions cannot generally be obtained, and so we maintain \( h_t^* \in \mathcal{F}_t \). We then rewrite the model as

\[
y_t \mid \mathcal{F}_t, \beta_t \sim N \left[ x_t' \beta_t, h_t^*(\alpha, \mathcal{F}_t) \right],
\]

where the explicit parameterization for both functions depends on the model introduced to drive the coefficients.

The simplest assumption about the evolution of the coefficients is to make them a function of \( \mathcal{F}_t \), which adopting linearity results in

\[
\beta_t = \beta_1 + B_2 x_t^{++},
\]

where \( x_t^{++} \in \mathcal{F}_t \), and \( \beta_1 \) and \( B_2 \) are respectively a vector and a matrix of constant parameters. Models of this sort have been analyzed by Belsley [1974a, 1974b], though he also allows for conditional randomness as in the models we analyze below. If the \( x_t^{++} \) are composed of dummy variables this proposition represents the basic model of structural change which has been widely used in the literature and is common textbook material. With some stochastic structure determining the dummies, it may be generalized to the switching regression model (Quandt [1972], Goldfeld and Quandt [1974], Richard [1980]), though such stochastic structure usually results in \( x_t^+ \) not belonging to \( \mathcal{F}_t \), and a more general treatment is required. Also, the imposition of some smooth structure to transfer between regimes would result in the spline approach to structural change (Poirier [1976]). Substituting (2) in (1),

\[
y_t \mid \mathcal{F}_t \sim N \left[ x_t' \beta, h_t(\theta) \right],
\]

where \( x_t = (x_t^+, x_t^{++} \otimes x_t^{++})' \), \( \beta = (\beta_1, [\text{vec } B_2]')' \), \( h_t = h_t^* \), and \( \theta = (\beta', \alpha')' \) as usual. This *trending coefficients model* is a particular case of the heteroskedastic model discussed in Chapters 2 - 6, and no major issue arises
as to its treatment. The conditional variance $h_t^*$ is not affected by coefficient variation of this sort.

If the change in the mean coefficients is to affect the variance, we require $\beta_t$ to be a random variable with respect to the conditioning information set. This takes us to the random coefficients model (Swamy [1971]), in which it is proposed that

$$\beta_t \mid \mathcal{F}_t \sim N(\beta, \Omega_\beta),$$

where $\Omega_\beta$ is positive semidefinite and $\beta$ and $\Omega_\beta$ are constant parameters, so that $\beta_t$ is conditionally homoskedastic. The assumption of a constant conditional mean can be replaced, without much consequence, by trending coefficients. By construction $\beta_t$ is independent of $y_t - x_t^* \beta_t$, and since $h_t^*$ does not depend on $\beta_t$ it is seen that (see for example Chow [1984])

$$y_t \mid \mathcal{F}_t \sim N(x_t^* \beta, h_t(\theta)), $$

where $h_t = h_t^* + (x_t^* \otimes x_t^*) \text{vec} \Omega_\beta$, and $\theta = (\beta', \alpha', [\text{vech} \Omega_\beta]')'$. This is also a particular case of the heteroskedastic model in Chapters 2 - 6. The random variability of the coefficients is now an extra source of heteroskedasticity, so even when $h_t^*$ is constant the resulting model is heteroskedastic. The identifiability of $\beta$ is guaranteed from the mean equation as usual, and conditions need to be imposed for the identifiability of the remaining parameters $\alpha^* = (\alpha', [\text{vech} \Omega_\beta]')'$ in the variance equation, given $\beta$. This requires $\mathcal{C} \left( T^{-1} \sum_{t=1}^{T} \frac{\partial h_t}{\partial \alpha^*} \right)$ to exist and be positive definite in a neighborhood of $\theta_0$, and a necessary condition for this is that $R_0$ be positive definite, where

$$R_j = \mathcal{C} \left( T^{-1} \sum_{t=1}^{T} (x_t^* x_t^* \otimes x_t^* x_t^*') \right), \quad j = 0, 1, \ldots, \tag{3}$$

The main problem in estimating this model is that we now have two sources of positivity restrictions: the positivity of $h_t^*$ and the positive semidefiniteness of $\Omega_\beta$. In most well known formulations of $h_t^*$ its positivity can be guaranteed by
reparameterization (see Chapter 2). Swamy [1971] has discussed the problem of estimating $\Omega_\beta$, and this matrix can also be reparameterized to ensure positive semidefiniteness e.g. $\Omega_\beta = R R'$ for some lower triangular $R$ such that rank($R$) = rank($\Omega_\beta$). Nicholls and Quinn [1982] (see also Nicholls and Pagan [1985]) have studied these models when lagged dependent variables are present in $x_t^+$, providing conditions for wide-sense stationarity and the usual properties of the MLE. These conditions are satisfied by (Q0) - (Q7) of Chapter 2.

A richer class of models may be obtained allowing $\beta_t$ to have changing conditional moments, that is,

$$\beta_t \mid F_t \sim N [ \beta_{t|t-1} , V_{t|t-1} ] ,$$

where $\beta_{t|t-1} = E [ \beta_t \mid F_t ]$, and $V_{t|t-1} = V [ \beta_t \mid F_t ]$ is positive semidefinite. The added subscript ' $t-1$ ' to a dated random variable is used in general to denote its expectation conditional on $F_t$. Compounding (1) and (4) we obtain

$$y_t \mid F_t \sim N [ x_t^* \beta_{t|t-1} , h_t ] ,$$

where

$$h_t = h_t^* + ( x_t^* \otimes x_t^* ) \text{vec} V_{t|t-1} = h_t^* + x_t^* V_{t|t-1} x_t^* .$$

These models follow from the contributions of Kalman [1960] and Kalman and Bucy [1961], and were introduced to the econometrics literature, after the work of Schweppe [1965] on the evaluation of the likelihood function, by Rosenberg [1974] and Cooley and Prescott [1974, 1976]. Nicholls and Pagan [1985] and Chow [1984] review the literature. The model must be completed with some assumption about the evolution of the coefficients, and a common specification is an ARIMA formulation of some sort. For example, Cooley and Prescott [1974, 1976], Garbade [1977], Pagan and Tanaka [1979], and Engle and Watson [1985] use a random walk, while Pagan [1980] also considers a stationary general ARMA model. All these can be re-expressed as first order Markovian
processes through a companion form (Pagan [1980], Chow [1984]), and so we
can write without loss of generality

\[ b_t = M b_{t-1} + v_t, \]

where \( b_t = \beta_t - \beta \), and \( \beta \) is the unconditional mean of \( \beta_t \) if it exists or zero if the
expectation does not exist, \( M \) is the transition matrix and \( v_t \) is white noise with
covariance matrix \( \Omega_\beta \). Thus we get a random walk for \( M = I_k \), a stationary
ARMA if the roots of \( M \) lie inside the unit circle, and the random coefficients
model when \( M = 0 \). A trending mean can also be incorporated but this raises
no additional issues and thus will be ignored. Taking conditional expectations,

\[ b_{t\mid t-1} = M b_{t-1\mid t-1}, \]

and

\[ V_{t\mid t-1} = M V_{t-1\mid t-1} M' + \Omega_\beta, \]

constitute one part of the updating equations of the Kalman filter, and are
completed after compounding the distributions and obtaining the conditional
moments of \( \beta_t \) given \( y_t \) with

\[ b_{t\mid t} = b_{t\mid t-1} + h_t^{-1} V_{t\mid t-1} x_t^* u_t, \]

and

\[ V_{t\mid t} = V_{t\mid t-1} + h_t^{-1} V_{t\mid t-1} x_t^* x_t'^* V_{t\mid t-1}, \]

(e.g. Harvey [1985]), where we retain the notation \( u_t \) for the mean innovations
with respect to \( \mathcal{F}_t \), and so \( u_t = y_t - y_{t\mid t-1} = y_t - x_t'^* \beta_{t\mid t-1} = y_t - x_t'^* \beta - x_t'^* b_{t\mid t-1} \). Let \( \theta = (\beta', \alpha', \pi')' \), where \( \pi \) is the vector of distinct and unknown elements in \( M \)
and \( \Omega_\beta \). Then we can rewrite (5) as

\[ y_t \mid \mathcal{F}_t \sim N \left[ \mu_t (\theta), h_t (\theta) \right], \]

which again takes the form of the heteroskedastic model of Chapters 2 - 6,
given the extensions introduced for y-risk models in Chapter 7. The
dependence of the conditional mean in the whole parameter vector is seen from
the Kalman filter updating equations. Underlying this model is the fact that the log-likelihood function is already expressed in state space form, and this has been noted by Harvey [1984,1985] and Engle and Watson [1985], who allow for heteroskedasticity directly in (1). Provided the regularity conditions are met, we can treat the evolving coefficients model — as Nicholls and Pagan [1985] refer to this structure — with basically the same tools we have developed. The problem of evaluating the conditional moments and hence the likelihood function is solved by the Kalman filter which provides the necessary algorithm. Watson and Engle [1983] analyze the performance of the scoring and EM algorithms for the computation. Note that a changing $h_t^*$ does not contaminate the problem of initializing the filter because this is done purely on considerations that stem from the evolution of the coefficients. Thus if the evolution is stationary we may use the unconditional moments and set $b_{010} = 0$ and $V_{010} = M V_{010} M' + \Omega_\beta$, or vec $V_{010} = [I_k^2 - (M \otimes M)]^{-1}$ vec $\Omega_\beta$ (Pagan [1980], Harvey [1985]). If the coefficient evolution is nonstationary both unconditional moments do not exist and can be started from zero.

Theoretically, we can factorize the likelihood and obtain separate estimators from the two moments, but just recognizing what is identifiable in each of the two equations may prove far too complicated. However, it is of interest that in the homoskedastic case with stationary coefficients the parameters can be estimated consistently (directly or indirectly) from equations that resemble the mean and variance equations. From Theorem 4 of Pagan [1980] consider the "mean equation"

$$y_t = x_t^* \beta + u_t^* ,$$

the "variance equation"

$$\tilde{u}_t^2 = \sigma^2 + (x_t^* \otimes x_t^* ) \gamma_0 + \epsilon_{t0}^* ,$$

and the "autocovariance equations"
\[ \tilde{u}_t^+ \tilde{u}_{t+j} = (x_{t+j}^+ \otimes x_t^+) \gamma_j^+ + \tilde{e}_{t+j}^+, \quad j \in J, \quad - (8) \]

where \( u_t^+ = x_t^+ b_t + u_t, \tilde{u}_t^+ \) are the residuals from the 'mean equation', \( \sigma^2 \) is the constant variance (\( h_t^+ = \sigma^2 \)), \( \gamma_j = \text{vec} \Gamma_j \) when \( \Gamma_j \) is the \( j \)-th autocovariance of \( b_t \), the \( \tilde{e}_{t+j}^+ \) are regression errors with zero mean, and \( J \) is a set of indices such that there exists a one-to-one mapping between the parameters of the ARMA model for the coefficients and the set of autocovariances \( \Gamma_j \) for \( j \in J \), given the identification conditions in Hannan [1968]. Then \( \beta \) can be estimated consistently by OLS in (6) under the usual conditions on \( x_t^+ \). If the \( R_j \) defined in (3) are nonsingular for \( j \in J \), then consistent estimators of the \( \gamma_j \) can be obtained by OLS in (8), and from them we can derive consistent estimates for \( M \) and \( \Omega_{\beta} \). As to (7), \( \sigma^2 \) is not identifiable when a constant is present amongst the regressors, but it is identifiable given \( M \) and \( \Omega_{\beta} \). These conditions are sufficient for the existence of consistent estimators for the parameters and hence for their identifiability (see Deistler and Seifert [1978]). Pagan's results depend on the regressors being nonstochastic and bounded, a necessary condition to obtain an Aitken form for the likelihood function equivalent to the state space representation. These restrictions have been relaxed by Solo [1983, 1984] who, using a covariance Kalman filter, shows that the state space form can be reparameterized as a function of \( \beta \) and the autocovariances, and thus obtains the identifiability conditions more directly. Expressing the mean equation as in (6) causes the \( u_t^+ \) not to be innovations but rather autocorrelated through the presence of \( b_t \). This means that the variance equation does not contain all the information not contributed by the mean, and thus some autocovariance equations must be introduced. Under heteroskedasticity we only need to substitute the "variance equation" by

\[ \tilde{u}_t^2 = h_t^+ + (x_t^+ \otimes x_t^+) \gamma_0 + \tilde{e}_{t0}^+, \quad - (7') \]

and though the errors \( \tilde{e}_{t+j}^+ \) will have a more complicated structure, this does not affect the identifiability of the parameters.
As to the regularity conditions, Pagan restricts the parameter space in such a way that the state space form is uniformly completely observable (UCO) and uniformly completely controllable (UCC), and he relates these concepts to more familiar conditions in econometrics. With this the likelihood conforms to Crowder's [1976] conditions and we have the usual result of consistency and asymptotic normality of the MLE with covariance matrix equal to the inverse of the information matrix at \( \theta_0 \). This is valid for stationary parameters and also when \( M \) has unit roots, provided that the elements of \( M \) are known and not estimated, which is usually the case (e.g. \( M = I_k \)). Weiss [1982] has allowed for random regressors. The existence of fourth order moments is required as we have already assumed, and we can safely propose that the heteroskedasticity, so long as it has at least finite second order expectation, will not require additional conditions.

At a practical level, one of the main problems in the estimation of evolving coefficient models is that the number of parameters grows geometrically with the number of varying coefficients because \( M \) is \( k \times k \) and \( \Omega_\beta \) has \( \frac{1}{2} k (k+1) \) distinct elements. The number of parameters in \( \Omega_\beta \) can be reduced using its canonical form and imposing zero restrictions on the eigenvalues. This implies that the coefficient evolution is being driven by a lower order process, and it is a basic element of signal extraction models (see Fernández-Macho et al. [1986], Engle and Watson [1981,1985], Snyder [1985]). This approach has the disadvantage that LM tests for the restrictions on the eigenvalues cannot be computed because there are unidentifiable parameters under the null hypothesis (see Watson [1980] and Engle [1984]), but it has the advantage of ensuring the positive semidefiniteness of \( \Omega_\beta \). The dimensionality of \( M \) is typically reduced by scattering zeroes in this matrix because using a canonical form without symmetry becomes more complex.
§ 8.2 Further tests for parameter stability and diagnostic tests for varying coefficient heteroskedastic models

In section § 6.3 we have already considered some tests for parameter stability against the alternative of a structural break. We expect these tests to have some power in the presence of smooth parameter changes such as those described in the previous section, and it is possible to devise more powerful tests for specific patterns of parameter change by considering

\[ H_0 : \text{constant coefficients (with } \mu_t \text{ linear in } \beta), \]
vs \[ H_1 : \text{varying coefficients according to trending, random, or evolving coefficients}. \]

All these tests can be treated as variable addition tests.

Let us consider first the trending coefficients model. The test is a generalization of the Chow [1960] test in a heteroskedastic environment, and the additional variables for the auxiliary regression are provided by

\[
\frac{\partial \mu_t}{\partial \beta_2} = x_t^+ \otimes x_t^+, \]

for the mean equation, where \( \beta_2 = \text{vec } B_2 \), and

\[
\frac{\partial h_t}{\partial \beta_2} = \sum_{j \geq 0} \frac{\partial h_t}{\partial \mu_{t-j}} (x_{t+j}^+ \otimes x_{t+j}^+),
\]

for the variance equation.

In the random coefficients model parameter constancy obtains when \( \Omega_\beta = 0 \). Hence we need not augment the auxiliary mean equation, and the additional variables for the auxiliary variance equation are simply

\[
\frac{\partial h_t}{\partial \alpha_2} = P (x_t^+ \otimes x_t^+), \quad - (9)
\]

where \( \alpha_2 = \text{vech } \Omega_\beta \), and \( P \) is such that \( \text{vec } \Omega_\beta = P' \text{vec } \Omega_\beta \) (see Henderson and Searle [1979], and also Chapter 9). This is a simple extension of the Breusch-Pagan [1979,1980] test in which the heteroskedasticity is tested only partially.
and the remaining change in $h^*_t$ must be taken into account to produce a test with correct size. This LM test ignores the non-negativity restrictions on $\Omega_\beta$ and thus loses power. The null hypothesis is on the boundary of the parameter space and the alternative likelihood ratio test has a complicated distribution under the null (see Nicholls and Quinn [1982], Nicholls and Pagan [1985]). Nicholls and Quinn propose a more powerful test than the one based in adding (9) to the auxiliary regression, by considering the truncated estimator of $\Omega_\beta$.

To test for coefficient evolution we require the additional variables

$$\frac{\partial \mu_t}{\partial \pi} = \frac{\partial b_{t+1}}{\partial \pi} x_t,$$

in the mean equation, and

$$\frac{\partial h_t}{\partial \pi} = \frac{\partial b_{t+1}}{\partial \pi} \frac{\partial h_t^*}{\partial b_{t+1}} + \frac{\partial v_{t+1}}{\partial \pi} \frac{\partial h_t^*}{\partial v_{t+1}} + \frac{\partial b_{t+1}}{\partial \pi} x_t,$$

in the variance equation, where $v_{t+1} = \text{vech } V_{t+1}$. The derivatives of $b_{t+1}$ and $V_{t+1}$ must be computed recursively using the Kalman filter equations. If the model is stationary and $M$ contains unknown parameters, testing for the whole of $\pi$ to be zero involves some redundancy, as noted by Nicholls and Pagan [1985], and also by Pagan and Hall [1983] who discuss the problem as residual analysis relating it to the regressions in (7) and (8). In this context Watson [1980] has found unidentifiable parameters under the null. If the value of $M$ is fixed to avoid this problem, then Pagan and Tanaka [1979] have provided the statistic to test $\Omega_\beta = 0$ when $M = I_k$. The alternative likelihood ratio test has been studied in the same situation by Garbade [1977]. However, Tanaka [1981] has shown that the LM statistic does not have an asymptotic $\chi^2$ distribution under the null because there are convergence problems in the estimated information matrix. Pagan and Hall [1983] offer an explanation and propose an alternative approach, but their conclusion is that it seems more convenient to base diagnostic tests for potential coefficient evolution simply on tests for $\Omega_\beta = 0$, and the same is proposed by Nicholls and Pagan [1985]. It is easy to see
that, since the intrinsic variance $h_t^*$ only enters the variance equation (7') and not the autocovariance equations (8), Pagan and Hall's argument readily extends to the heteroskedastic setting. They suggest a test based on the joint estimator of $\Omega_3$ obtained from (7) and a fixed number of autocovariance equations. However, since this becomes increasingly complicated with the dimension of $x_t$ and the complexity of the coefficient evolution, it appears again that a reasonable initial diagnostic test is that for random coefficients.

An alternative graphic approach to assessing coefficient evolution in homoskedastic models has been put forward by Brown et al. [1975] in the form of the CUSUM and CUSUMSQ tests. One of the attractions of this method is the simple straight-line form of the bounds on the cumulative sums, though the CUSUM test seems to have very little power to detect parameter instability (see for example the simulations in Garbade [1977]). The problem with this method under heteroskedasticity is that there remains an implicit source of variation in the regression, given by the conditional variance. If the $h_t$ were known this could be solved easily by looking at the cumulative sums of recursive residuals in the mean equation standardized dividing through by $h_t^{1/2}$. The best we can do is to use parametric estimates of the conditional variances $\tilde{h}_t = h_t(\tilde{\theta})$. The method loses appeal because this has an effect on the distribution of the recursive residuals, introducing dependence, but it may still be worth looking at the CUSUM and CUSUMSQ graphics as an informal check on likely coefficient evolution. The uncertainty introduced by the unknown conditional variances disappears at a fast rate when the sample size grows. As a complement to the CUSUMSQ test, we can also apply this method to the variance equation to assess how well the parameterization of $h_t$ is capturing the changes in the conditional variance. The problem here is that unless $h_t$ follows a linear simple heteroskedasticity pattern the updating formulae of Brown et al. are not valid and the problem may involve a good deal of computation. Under the null of parameter stability the recursive estimates
from sample size $t^*$ should provide good starting values for the estimates for sample size $t^*+1$.

Finally, if coefficient variation has been accepted and introduced into the model, we still need to assess the adequacy of the chosen parameterization. Provided the regularity conditions are met for the MLE to have its usual normal asymptotic distribution as discussed in the previous section, and the matrices $\Phi$ and/or $\vartheta$ of the efficiency and consistency tests also fulfill the regularity required by Theorems 5.3 and 5.10, respectively, we can use these tests. This poses no major problem when the maintained model has trending or random coefficients. If the model has evolving coefficients the proper mean and variance equations are

$$y_t = x_t^T \beta + x_t^T b_{t|t-1} + u_t,$$

and

$$u_t^2 = h_t^* (\theta) + (x_t^T \otimes x_t^T) vec V_{t|t-1} + \varepsilon_t,$$

respectively. Under the null hypothesis of a correctly specified model, the filtered coefficients and variances $b_{t|t-1}$ and $V_{t|t-1}$ are a by-product of estimation and thus are available for computation of the statistics. Therefore general tests of misspecification can be easily performed from auxiliary double-length regressions. When a specific alternative is being entertained the computation of the additional variables for the auxiliary regressions may still need use of the updating equations of the Kalman filter. The diagnostics proposed by Nicholls and Pagan [1985] to deal with autocorrelation are in fact consistency tests of the form proposed here.

§ 8.3 Testing for superexogeneity and invariance

The Lucas [1976] critique is a serious challenge to the usefulness of econometric models in policy analysis. Engle et al [1983] have proposed the
concept of superexogeneity, which is closely related to this critique. A set of (conditioning) variables is said to be superexogenous for the parameters of interest in a conditional model if it is weakly exogenous and the parameters of interest are invariant with respect to changes in the conditioning distribution.

The lack of tests for superexogeneity may arise from either confusion with the problem of parameter constancy, or from recognition that testing for superexogeneity complicates substantially the parameter constancy framework. It is therefore convenient to distinguish parameter constancy from invariance, and to analyze whether the approaches to the former can be useful in analyzing the latter. Parameter constancy is neither necessary nor sufficient for invariance. It is not necessary because the conditional model may, on its own, have time varying parameters and still be invariant to changes in the conditioning model. Thus Hendry's [1985] view of the necessity of parameter constancy for invariance does not seem adequate. It may be sufficient if it is known, in addition, that the conditioning model has been subject to change and constancy remains. Invariance does not have to be restricted to simultaneous changes in both the conditioning and conditional models because it may well be the case that agents learn and react through time to changes in policy, a point which is emphasized in the rational expectations literature. Thus, even if we have been able to detect a structural break in the conditioning model, this may induce one or more structural breaks in the (not invariant) conditional model at any time from then onwards. This renders structural break methods like those in § 6.3 of little use and, as has been suggested by Engle and Watson [1985], a varying coefficient approach seems to be better equipped for testing superexogeneity.

The only reference we have come across in testing for superexogeneity is Engle and Hendry [1986], where they propose to augment the conditional moments with characteristics from the conditioning distribution. This must be the basic principle for this type of test, and here we put forward an alternative
view of the problem by specifying in more detail the learning mechanism on the part of economic agents and policy makers that would induce the changes in the conditional distribution. We start by extending the model of the previous sections to include both a behavioral equation and a policy rule. The behavioral equation is basically an evolving coefficient regression, but we must now take into account that the distribution needs to condition on the parameters of the policy rule as well. We partition the conditioning variables into policy and environmental variables, namely \( x_t^e = (x_t^{p'} , x_t^{e'})' \), where \( x_t^{p} \) is the \( kp \times 1 \) vector of policy variables, and \( x_t^{e} \) is the \( ke \times 1 \) vector of environmental variables, including lagged dependent variables if necessary. The behavioral equation is

\[
y_t \mid \mathcal{F}_t, \lambda_t \sim N \left( x_t^{p'} \beta_t + x_t^{e'} \beta_2 , h^*_t(\alpha , \mathcal{F}_t) \right),
\]

where for simplicity we take the coefficients associated to \( x_t^{e} \) as constants, and \( \lambda_t \) is defined after (11) below. The elements other than lagged dependent variables of \( x_t^{e} \in \mathcal{G}_t \) are assumed weakly exogenous for all the parameters of the model. The policy rule is represented by

\[
x_t^p \mid \mathcal{G}_t, \lambda_t \sim N \left( R_{1t} \beta_{xt} + R_{2t} \beta_{x2} , H^*_t(\alpha_{x} , \mathcal{G}_t) \right),
\]

where the partition of the variables \( R_t = (R_{1t} , R_{2t}) \in \mathcal{G}_t \), which are \( k_p \times (k_{x1} + k_{x2}) \) follows a similar argument to that of \( x_t^+ \), and \( \lambda_t = (\beta_{xt}' , \beta_{xt}' \gamma) \) is the vector of varying coefficients in the joint model. This model can be interpreted along the lines of Chow [1975, 1981], though no claim of optimality is made for the policy rule as this would impose a very tight structure on the evolution of \( \beta_{xt} \).

To construct the likelihood function for the model we need to marginalize \( \lambda_t \) to obtain \( f( y_t , x_t^p \mid \mathcal{G}_t ) \), and we also require a Kalman filter to compute the conditional moments in terms of the evolution of the moments of \( \lambda_t \). We assume that

\[
\lambda_t \mid \mathcal{G}_t \sim N \left( \lambda_{t|t-1} , V_{\lambda t|t-1} \right),
\]
where \( \lambda_{t|t-1} = (\beta_{t|t-1}^{\prime}, \beta_{x|t-1}^{\prime})^{\prime} \), and \( V_{\lambda_{t|t-1}} = \begin{pmatrix} V_{\lambda_{t|t-1}} & C_{\lambda_{t|t-1}} \\ C_{\lambda_{t|t-1}}^{\prime} & V_{x_{t|t-1}} \end{pmatrix} \), and for the argument that follows we assume that \( |V_{\lambda_{t|t-1}}| \neq 0 \) so that this distribution is proper, but this can be later relaxed. From (10) - (12) we have

\[
f(y_t, x_t^P, \lambda_t | \mathcal{G}_t) = f(y_t, x_t^P | \mathcal{G}_t, \lambda_t) f(\lambda_t | \mathcal{G}_t) 
= f(y_t | \mathcal{F}_t, \lambda_t) f(x_t^P | \mathcal{G}_t, \lambda_t) f(\lambda_t | \mathcal{G}_t),
\]

and to obtain the likelihood function and Kalman filter we factorize this as

\[
f(y_t, x_t^P, \lambda_t | \mathcal{G}_t) = f(\lambda_t | \mathcal{G}_{t+1}) f(y_t, x_t^P | \mathcal{G}_t) 
= f(\lambda_t | \mathcal{G}_{t+1}) f(y_t | \mathcal{F}_t) f(x_t^P | \mathcal{G}_t),
\]

so that the likelihood can be separated into a conditional/marginal parameterization as required, and the remaining factor provides the updating information for the filter. In order to obtain the factorization we must be careful about the way in which new information is made available so that it conforms to the conditional/marginal factorization of the likelihood. Once the information \( \mathcal{G}_t \) is available, the next step is to incorporate \( x_t^P \) and condition on \( \mathcal{F}_t \). Then we incorporate \( y_t \) and condition on \( \mathcal{G}_{t+1} \). To distinguish clearly the conditioning information sets in the formation of expectations, for any random variable \( \xi_t \) we denote \( \xi_{t|t-1} = E[\xi_t | \mathcal{G}_t] \), \( \xi_{t|t} = E[\xi_t | \mathcal{G}_{t+1}] \), and the intermediate expectation \( \xi_{t|t-1}^{\mathcal{F}} = E[\xi_t | \mathcal{F}_t] \). Information is incorporated according to the following diagram

\[
\begin{array}{c}
\mathcal{G}_t \\
\lambda_{t|t-1} \text{ and } V_{\lambda_{t|t-1}} \\
\uparrow \\
\mathcal{F}_t \\
\lambda_{t|t} \text{ and } V_{\lambda_{t|t}}
\end{array} \rightarrow \begin{array}{c}
\mathcal{G}_{t+1} \\
\lambda_{t|t} \text{ and } V_{\lambda_{t|t}}
\end{array}
\]

To introduce \( x_t^P \) we consider

\[
f(x_t^P, \lambda_t | \mathcal{G}_t) = f(x_t^P | \mathcal{G}_t, \lambda_t) f(\lambda_t | \mathcal{G}_t),
\]
which we want to factorize as

\[ f(\text{x}_t^p, \lambda_t | \mathcal{G}_t) = f(\lambda_t | \mathcal{F}_t) f(\text{x}_t^p | \mathcal{G}_t). \]

This density is the normal pdf because \( H_{xt}^* \in \mathcal{G}_t \), and we can write

\[ \text{x}_t^p = R_{1t} \beta_x + R_{2t} \beta_x^2 + u_t^p = R_{1t} \beta_{xt_l1} + R_{2t} \beta_x^2 + (u_t^p + R_{1t} (\beta_{xt} - \beta_{xt_l1})), \]

where \( u_t^p \) are the innovations with respect to \( \{ \mathcal{G}_t, \lambda_t \} \), and thus are orthogonal to \( \lambda_t \). Then it is seen that

\[
\begin{pmatrix}
\text{x}_t^p \\
\lambda_t
\end{pmatrix}
\mid \mathcal{G}_t 
\sim \mathcal{N}
\left[
\begin{pmatrix}
\text{x}_{t_l1}^p \\
\lambda_{t_l1}
\end{pmatrix},
\begin{pmatrix}
H_{xt} & R_{1t} A_2 V_{\lambda t_l1} \\
V_{\lambda t_l1} A_2' R_{1t}' & V_{\lambda t_l1}
\end{pmatrix}
\right],
\]

where

\[
\begin{align*}
\text{x}_{t_l1}^p &= R_{1t} \beta_{xt_l1} + R_{2t} \beta_x^2, \\
H_{xt} &= H_{xt}^* + R_{1t} V_{xt_l1} R_{1t}',
\end{align*}
\]

and \( A_2 = (0, I_{k_x}'). \) The marginal pdf for \( \text{x}_t^p \) is obvious in these expressions, and conditioning on \( \text{x}_t^p \) we get

\[ \lambda_t \mid \mathcal{F}_t \sim \mathcal{N}[\lambda_{t_l1}^\mathcal{F}, V_{\lambda t_l1}^\mathcal{F}], \]

where

\[
\begin{align*}
\lambda_{t_l1}^\mathcal{F} &= \lambda_{t_l1} + V_{\lambda t_l1} A_2' R_{1t}' H_{xt}^{-1} (\text{x}_t^p - \text{x}_{t_l1}^p), \\
V_{\lambda t_l1}^\mathcal{F} &= V_{\lambda t_l1} - V_{\lambda t_l1} A_2' R_{1t}' H_{xt}^{-1} R_{1t} A_2 V_{\lambda t_l1}.
\end{align*}
\]

The next step is to incorporate \( y_t \) into the information set, and for this we consider

\[ f(\text{y}_t, \lambda_t | \mathcal{F}_t) = f(\text{y}_t | \mathcal{F}_t, \lambda_t) f(\lambda_t | \mathcal{F}_t), \]

which we want to factorize as

\[ f(\text{y}_t, \lambda_t | \mathcal{F}_t) = f(\lambda_t | \mathcal{G}_{t+1}) f(\text{y}_t | \mathcal{F}_t). \]
This density is the normal pdf because $h_t^* \in \mathcal{F}_t$, and we now write

$$y_t = x_t^P \beta_t + x_t^o \beta_2 + u_t^* = x_t^P \beta_{t|t-1}^F + x_t^o \beta_2 + \{ u_t^* + x_t^P(\beta_t - \beta_{t|t-1}^F) \},$$

where $u_t^*$ are the innovations with respect to $(\mathcal{F}_t, \lambda_t)$, and hence are orthogonal to $\lambda_t$. By similar compounding procedure we get

$$\begin{pmatrix} y_t \\ \lambda_t \end{pmatrix} \mid \mathcal{F}_t \sim \mathcal{N} \left[ \begin{pmatrix} y_{t|t-1}^F \\ \lambda_{t|t-1}^F \end{pmatrix}, \begin{pmatrix} h_t & x_t^P A_1 V_{\lambda t|t-1}^F \\ V_{\lambda t|t-1}^F A_1' x_t^P & V_{\lambda t|t-1}^F \end{pmatrix} \right],$$

where

$$y_{t|t-1}^F = x_t^P \beta_{t|t-1}^F + x_t^o \beta_2,$$

$$h_t = h_t^* + x_t^P V_{t|t-1}^F x_t^P = h_t^* + (x_t^P \otimes x_t^P') \text{vec} V_{t|t-1}^F,$$

and $A_1 = (I_{k_p}, 0)$. The marginal pdf for $y_t$ follows from these expressions, and conditioning on $y_t$ we obtain

$$\lambda_t \mid \mathcal{G}_{t+1} \sim \mathcal{N} [ \lambda_{t|t}, V_{t|t} ],$$

where

$$\lambda_{t|t} = \lambda_{t|t-1}^F + h_t^{-1} V_{\lambda t|t-1}^F A_1' x_t^P (y_t - y_{t|t-1}^F),$$

and

$$V_{\lambda t|t} = V_{\lambda t|t-1}^F - h_t^{-1} V_{\lambda t|t-1}^F A_1' x_t^P x_t^P A_1 V_{\lambda t|t-1}^F.$$

We can now collect results for the factorization in (13). The conditional likelihood is constructed from $f( y_t \mid \mathcal{F}_t )$ in (16). The marginal likelihood is constructed from $f( x_t^P \mid \mathcal{G}_t )$ in (14). And the conditional moments for these distributions can be obtained using the Kalman filter equations in (15) and (17). These updating equations are not affected if the distribution of $\lambda_t$ is singular, though $f( \lambda_t \mid \mathcal{G}_t )$ is not defined.

We still need one more element to complete the updating procedure of the filter: the conditional moments of $\lambda_t$. For this purpose we need to specify the
learning mechanism underlying (12) on the part of both agents and policy makers. We start by assuming that there exists a constant (unconditional) expected behavior in the reactions of both agents and policy makers, so that

\[ E[\lambda_t] = \lambda = (\beta_1', \beta_{x1}')', \]

and we define the (unconditionally) zero mean stochastic components

\[ \bar{\lambda}_t = (\bar{\beta}_t', \bar{\beta}_{xt}')' = \lambda_t - \lambda = (\beta_t' - \beta_1', \beta_{xt}' - \beta_{x1}')', \]

so we can now rewrite the linear model in (10) - (12) as

\[ y_t \mid \mathcal{F}_t, \lambda_t \sim N [ x_t^T \beta + x_t^P \bar{\beta}_t, h^*_t(\alpha, \mathcal{F}_t) ] , \]

where \( \beta = (\beta_1', \beta_2')' \), and

\[ x_t^P \mid \mathcal{G}_t, \lambda_t \sim N [ R_t \beta_x + R_{1t} \bar{\beta}_{xt}, H^*_x(\alpha_x, \mathcal{G}_t) ] , \]

where \( \beta_x = (\beta_{x1}', \beta_{x2}')' \), and \( R_t = (R_{1t}, R_{2t}) \). The learning process is completed with a multivariate stationary ARMA for the stochastic component of \( \lambda_t \), and since this can be re-expressed as a first order Markovian process through a companion form we lose no generality by considering the AR(1) learning process

\[ \bar{\lambda}_t = M_\lambda \bar{\lambda}_{t-1} + v_t , \] - (18a)

or in partitioned form,

\[ \begin{pmatrix} \bar{\beta}_t \\ \bar{\beta}_{xt} \end{pmatrix} = \begin{pmatrix} M & M_{yx} \\ M_{xy} & M_x \end{pmatrix} \begin{pmatrix} \bar{\beta}_{t-1} \\ \bar{\beta}_{xt-1} \end{pmatrix} + \begin{pmatrix} v_{yt} \\ v_{xt} \end{pmatrix} , \] - (18b)

where \( v_t \) is a white noise with constant conditional variance

\[ V_\lambda = \begin{pmatrix} V_\beta & C \\ C' & V_x \end{pmatrix} . \] - (18c)

The updating equations follow as
\[ \bar{x}_{t+1} = M_x \bar{x}_{t+1} \]  
- (19a)

and

\[ V_{\lambda t+1} = M V_{\lambda t+1} M_x' + V_\lambda , \]  
- (19b)

and the initial values assuming stationarity are given by the unconditional moments \( \bar{x}_{00} = 0 \), and

\[ V_{\lambda 010} = M_x V_{\lambda 010} M_x' + V_\lambda , \]  
\[ \text{or vec } V_{\lambda 010} = ( I - M_x \otimes M_x )^{-1} \text{ vec } V_\lambda . \]  
- (20)

For propositions about invariance to make sense, the concept of "parameters" must be clarified. When we allow for varying coefficients in a conditional model, the parameters may be taken to be either the conditional moments of \( \lambda_t \) or the constant parameters underlying the mechanism driving the coefficients. If we take the latter concept, then it is clear that superexogeneity is equivalent to weak exogeneity with respect to these parameters. But if we are interested in testing for invariance, more insight into the problem is gained by taking the conditional moments of \( \lambda_t \) as the parameters of the conditional distribution and we do so in what follows. Thus we say that the conditional model is invariant with respect to changes in the conditioning model if the conditional moments \( \mu_t \) and \( h_t \) do not depend on the parameters of the conditioning distribution (i.e. the moments of \( \beta_{xt} \)).

The learning process employed is well known in the econometric literature (e.g. Chow [1975], Harvey [1981]), but nevertheless it may be of interest to highlight its main properties. There is an exogenous component in the determination of the \( \bar{x}_t \), as agents and policy makers act through their corresponding innovations in \( v_t \). But to any such actions, an endogenous learning process recurs through time in three ways. Firstly, there is a contemporaneous effect as agents and policy makers perceive each other's actions. This response is transmitted through the covariances between the coefficients. Secondly, there is an institutional behavior, represented by \( M \) and \( M_x \), which represents the memory of own actions. Finally, there is a direct
learning from the other's previous decisions through $M_{yX}$ and $M_{xy}$. The last two effects constitute the feedback mechanism. Since the process is stationary, an isolated intervention eventually dies out. A permanent intervention is gradually endogeneized by shifting the non-stochastic mean components $\lambda$ and returning the mean value of $v_t$ to zero. Likewise, any systematic behavior through the innovations will tend to be endogeneized by the learning process, provided it can be represented by an ARMA model. This is attractive because of the wide range of data sets that can be well approximated by models of this kind. It is also possible to introduce other variables in the learning process, but this does not introduce any qualitatively different aspect into the analysis.

This learning process suggests natural concepts of invariance. If agents do not learn from the actions of policy makers we define the \textit{natural conditional invariance} hypothesis

$$H_{CI} : M_{yx} = C = 0 ,$$

which eliminates contemporaneous learning through the covariance and also the lagged reactions in $\beta_t$ to changes in $\beta_{xt}$. Medium and long term covariances between the evolving coefficients of the two equations remain because $M$ is block (lower) triangular and $V_\lambda$ is block diagonal, and so $M_\lambda V_\lambda M_\lambda'$ is not block diagonal. These medium and long term covariances are just reflecting the learning which is taking place by the policy makers through $M_{xy}$. This leads us to define the hypothesis of \textit{natural policy invariance}, that is, that policy makers do not learn from agents reactions,

$$H_{PI} : M_{xy} = C' = 0 .$$

We also define the joint hypothesis of \textit{natural invariance} that no learning takes place by either agents or policy makers, and this is given by

$$H_{NI} = H_{CI} \cup H_{PI} : M_{yx} = M_{xy}' = C = 0 .$$
To relate these concepts of natural invariance to invariance proper as in Engle et al. [1983] we prove

**Theorem 8.1.** - If the true DGP is given by (10) - (12) and (18), and \( \{ \beta, V_\beta, M \} \) and \( \{ \beta_x, V_x, M_x, M_{xy} \} \) are variation free, the conditional model is invariant with respect to changes in the conditioning model if, and only if,

(i) \( M_{yx} = 0 \),

(ii) \( C = 0 \),

(iii) \( M_{xy} = 0 \) or \( M = 0 \).

**Proof:** From (16) the conditional moments are invariant if \( \beta_{t|t-1}^\xi \) and \( \psi_{t|t-1}^\xi \) do not depend on moments of \( \beta_{xt} \). From (15) we see that for \( \beta_{t|t-1}^\xi \) this requires (a) \( \beta_{t|t-1}^\xi \) depends only on moments of \( \beta_t \), and (b) \( V_{\lambda t|t-1} A_2^\prime R_{|t} \ H_{xt}^1 ( x_t^p - x_{t|t-1}^p ) \) has zeroes in its first \( k_p \) elements. Inspection of (19) shows that (a) holds for all \( t \) iff (i) \( M_{yx} = 0 \), whereas (b) holds for all \( t \) iff \( C_{t|t-1} = 0 \), because \( V_{\lambda t|t-1} A_2^\prime = ( C_{t|t-1}^\prime, V_{xt|t-1}^\prime ) \).

It can also be seen with (15) that if these conditions hold then \( \psi_{t|t-1}^\xi = \psi_{t|t-1}^\xi = V_{t|t-1} \).

Since \( C_{t|t-1} = M V_{t-1|t-1} M_{xy}^\prime + M C_{t-1|t-1} M_{x}^\prime + C \), and this is null for all \( t \) iff (ii) \( C = 0 \), and (iii) \( M = 0 \) or \( M_{xy} = 0 \). The second term in \( C_{t|t-1} \) vanishes when (i) - (iii) are met. If \( M = 0 \) this is trivial. If \( M_{xy}^\prime = M_{yx} = C = 0 \), the initial conditions in (20) say that \( C_{0|10} = M C_{0|10} M_{x}^\prime \), for which \( C_{0|10} = 0 \) is always a solution, and since under stationarity (20) has a unique solution it follows that \( C_{0|10} = 0 \). This makes \( C_{1|10} = 0 \), and partitioning in (17) it is seen that \( C_{t|t} = 0 \) if \( C_{t|t-1} = 0 \), which is then guaranteed by the other conditions.

Let us denote the two ways of obtaining invariance in the conditional model by

\[ H_{11} : M_{yx} = C = M_{xy}^\prime = 0 , \]

and

\[ H_{12} : M_{yx} = C = 0 \text{ and } M = 0 . \]
It is immediate that natural conditional invariance is necessary, but not sufficient, for invariance in the conditional model, and also $H_{NI}$ is equivalent to $H_{12}$, and therefore natural invariance – no learning at all – is sufficient, but not necessary for invariance of the conditional model. Further note that if $H_{NI}$ and the hypotheses of Theorem 8.1 hold, then $x^p_t$ is weakly exogenous for the parameters of the conditional model, and since the latter is also invariant, it follows that $x^p_t$ is superexogenous for the parameters of the conditional model. Under $H_{12}$ the conditional model is a random coefficients model and it is invariant, but $x^p_t$ is not weakly exogenous because the conditional likelihood is not sufficient for inferences in view of the appearance of $\beta_t$ in the marginal likelihood (see §6.4). Thus we have the following

**Corollary 8.2.** Under the assumptions of Theorem 8.1 $x^p_t$ is superexogenous for the parameters of the conditional model if, and only if, $H_{NI}$ holds.

**Proof:** Follows from the above argument.

$H_{12}$ implies that having constant coefficients in the conditional model is sufficient for invariance, though not for superexogeneity. This arises because of the implicit assumption that the conditioning model is changing. If this is not the case, we will not find evidence against invariance because of a lack of experimentation (i.e. all coefficients remained constant during the sample). In any case this gives grounds for some confidence if both parameter stability and weak exogeneity are not rejected when diagnosing the conditional model alone. Of course one runs the risk that the parameter constancy tests may have little power against the joint evolution of the parameters. Corollary 8.2 then suggests the test for superexogeneity: fit the conditional and marginal models separately with evolving coefficients, and form the auxiliary equations by adding the derivatives with respect to $M_{xy}$, $M_{yx}$ and $C$ evaluated under the null, basing the test on the uncentered coefficient of determination of the auxiliary regression. This involves a good deal of computation and a more practical test
may be constructed under the assumption that $M_\lambda = 0$. Then both the conditional and marginal models are separate random coefficient models under the null hypothesis of superexogeneity (invariance under $H_{12}$ also holds), which can be stated as $H_0: C = 0$. When $M_\lambda = 0$ we have from (19) that $\lambda_{t_{t-1}} = \lambda$ and $V_{\lambda t_{t-1}} = V_\lambda$, and simple algebra establishes that the conditional moments are

$$\mu_t = x_t^P \beta + \{ (x_t^P - x_{t_{t-1}}^P)' H_{xt}^{-1} R_{1t} \otimes x_t^P \} \text{vec } C,$$

and

$$h_t = h_t^* + x_t^P V_\beta x_t^P - x_t^P C R_{1t}' H_{xt}^{-1} R_{1t} C' x_t^P,$$

so that a test for superexogeneity can be constructed by adding to the double-length regression of the conditional model the variables

$$\frac{\partial \mu_t}{\partial \text{vec } C} = \{ (x_t^P - x_{t_{t-1}}^P)' H_{xt}^{-1} R_{1t} \otimes x_t^P \},$$

to the mean auxiliary regression, and

$$\frac{\partial h_t}{\partial \text{vec } C} = \sum_{j \geq 0} \frac{\partial h_t^*}{\partial \mu_{t-j}} \{ (x_{t-j}^P - x_{t_{t-j-1}}^P)' H_{xt-j}^{-1} R_{1t-j} \otimes x_{t-j}^P \},$$

where the last derivative is valid only under the null hypothesis. The additional variables are provided by fitting the marginal model with random coefficients.
CHAPTER 9

MULTIVARIATE HETEROSKEDASTIC MODELS

This chapter extends the results of Chapters 3 to 6 to the case in which \( y_t \) is an n-dimensional random vector. The estimation theory for the multivariate normal ARCH model was developed by Kraft and Engle [1982] (KE in this Chapter), who extended their results to the multivariate simple heteroskedasticity model. We use their main results and extend them to more general heteroskedastic models, and at the same time we provide a wide range of diagnostic tests for these models both against general and specific alternatives. The tests contemplate the assessment of the specification of any conditional moment of the distribution. We use extensively the results on Kronecker products, vec and vech operators in Henderson and Searle [1979] (HS in this Chapter). In section § 9.1 we extend the notation and assumptions to the multivariate model. The multivariate ARCH class and other models are briefly discussed in section § 9.2. In section § 9.3 information is extracted from the mean and covariance matrix of \( y_t \) by means of orthogonality conditions, and is combined using a MWA as in the univariate case. The corresponding likelihood factorization under conditional normality is also derived. The implications of specification error are discussed in section § 9.4, and in section § 9.5 the family of consistency tests is generalized and the LM test for variable addition is seen to belong to this family. Finally, the efficiency tests are considered in section § 9.6. In many cases proofs are simple extensions of the arguments used for the univariate model, and in some others they distract from the main argument. These proofs are omitted in the text and presented in appendix at the end of the chapter.
§ 9.1 Notation and assumptions

We retain the notation of the univariate model as much as possible, re-dimensioning the required variables. The conditional mean

$$\mu_t(\beta) = E[y_t | \mathcal{F}_t]$$

is now an $n \times 1$ vector, the conditional covariance matrix is

$$H_t(\theta) = H_t(\beta, \alpha) = E[u_t u_t' | \mathcal{F}_t]$$

of dimension $n \times n$, and $u_t = y_t - \mu_t(\beta)$ is the vector of innovations in the conditional mean which defines the (system of) mean equation(s)

$$y_t = \mu_t(\beta) + u_t$$

exactly as in the univariate case. For the second moment we now have the matrix of innovations

$$E_t = u_t u_t' - H_t$$

but because of symmetry only the lower (or upper) triangle of $E_t$ provides new information. Thus we use the vech operator which vectorizes the lower triangle of a square matrix, and take the variance innovations as

$$\varepsilon_t = \nu_t - h_t$$

where $\varepsilon_t = \text{vech} E_t$, $\nu_t = \text{vech} u_t u_t'$, and $h_t = \text{vech} H_t$. This defines the (system of co-) variance equation(s)

$$\nu_t = h_t(\theta) + \varepsilon_t$$

Combining the mean and variance innovations we get

$$\psi_t = \eta_t - g_t$$
where \( u_t = (u_t', \varepsilon_t')' \), \( \eta_t = (y_t', v_t')' \), and \( g_t = (\mu_t', h_t')' \), which produces the system of \( n + \frac{n}{2} (n + 1) \) equations for the first two moments,

\[
\eta_t = g_t(\theta) + u_t.
\]

Throughout this chapter we use constantly the relationship between the vec and vech operators. Suppose \( C \) is an \( n \times n \) symmetric matrix. From HS there exist \( \frac{1}{2} n (n + 1) \times n^2 \) matrices \( P \) and \( Q \) such that

\[
\text{vec } C = P' \text{vech } C, \quad - (2)
\]

\[
\text{vech } C = Q \text{vec } C, \quad - (3)
\]

\[
Q P' = PQ' = I_{n(n+1)/2}, \quad - (4)
\]

and

\[
\text{vec } C = P' Q \text{vec } C.
\]

For unique definition of \( Q \), we take it to be the Moore-Penrose generalized inverse of \( P' \),

\[
Q = (P P')^{-1} P. \quad - (5)
\]

The assumptions (\( Q_0 \)) - (\( Q_8 \)) are easily recast in the multivariate context and will not be discussed. We simply refer to them as (\( Q_0-M \)) - (\( Q_8-M \)). Observe that \( H_t \) must be bounded uniformly from below by a positive definite matrix, and that this in general implies complex restrictions on the parameter subspace \( \mathfrak{C} \) (see KE and also Engle et al. [1984] for an example with the bivariate ARCH model). Whenever possible, it seems that the best thing to do is to reparameterize \( H_t \) in such a way that these restrictions are implicit. \( \mathfrak{B} \) and \( \mathfrak{C} \) are also restricted by the moment existence conditions. For example, if \( \mu_t \) is linear the polynomial matrix in the lag operator for \( y_t \) must have all its roots outside the unit circle for first order stationarity, thus imposing structure on \( \mathfrak{B} \). Likewise, for wide-sense stationarity in the ARCH model KE show that the same condition must apply to the relevant matrix polynomial defining the
multivariate ARCH variance and this imposes structure on $\mathcal{Q}$. The multivariate (4.4) requires the existence of fourth moments, and this places further restrictions on $\Theta$. Throughout this chapter we assume the conditional distribution of $y_t$ to be symmetric, and the kurtosis of this distribution can be characterized in terms of

$$
\text{Var} \left[ v_t | \mathcal{F}_t \right] = E \left[ \varepsilon_t \varepsilon_t' | \mathcal{F}_t \right] = K_t,
$$

which is $\frac{1}{2} n (n+1) \times \frac{1}{2} n (n+1)$ and assumed to be positive definite for all $t$. For example, when the conditional distribution is normal,

$$
\text{Var} \left[ \text{vec} \ u_t \ u_t' | \mathcal{F}_t \right] = (H_t \otimes H_t) (I_{n_2} + I_{(n,n)}) ,
$$

(see HS), where $I_{(n_1,n_2)}$ is the $n_1 n_2 \times n_1 n_2$ permutation matrix such that for any $n_1 \times n_2$ matrix $C$, $\text{vec} \ C = I_{(n_1,n_2)} \text{vec} \ C'$, and this implies using (3) that

$$
K_t = 2 Q (H_t \otimes H_t) Q' , \tag{6}
$$

because $Q I_{(n,n)} = Q$, see HS who also establish that for square $C$, $P' Q (C \otimes C) = (C \otimes C) P' Q$, and using this with (4) and (5) it is seen that

$$
K_t^{-1} = \frac{1}{2} P (H_t^{-1} \otimes H_t^{-1}) P' , \tag{7}
$$

under conditional normality.

The MLE under normality - seen as MLE or QMLE - plays a similar role in this multivariate context as the one it had in the univariate case, and it is convenient to introduce here for later reference the relevant functions of the likelihood, as derived by KE (see their Theorem 4). The likelihood function is

$$
\mathcal{L} (\theta) = (2 \pi)^{-n T/2} \left[ \prod_{t=1}^{T} \left| H_t \right| \right]^{1/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} u_t' H_t^{-1} u_t \right\} f( y_0 | \mathcal{F}_t ) ,
$$

and the log-likelihood (with the conventions from Chapter 2) is

$$
\mathcal{L} (\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log \left| H_t \right| - \frac{1}{2} T^{-1} \sum_{t=1}^{T} u_t' H_t^{-1} u_t .
$$
The score is given by
\[ d_\theta (\theta) = T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} S_t' P (H_t^{-1} \otimes H_t^{-1}) P' \varepsilon_t \]
\[ = T^{-1} X' \Omega^{-1} u + T^{-1} S' K^{-1} \varepsilon = T^{-1} G' \Sigma^{-1} \nu , \quad - (8a) \]
where \( X_t = \frac{\partial u_t}{\partial \beta'} \), \( X_t = \frac{\partial \mu_t}{\partial \phi'} = (X_t, 0) \), \( S_t = \frac{\partial h_t}{\partial \theta'} \), \( X = (X_1', ..., X_T')', X = (X, 0) \), \( S = (S_1', ..., S_T')', S = (S', S') \), \( u = (u_1', ..., u_T')', \varepsilon = (\varepsilon_1', ..., \varepsilon_T') \), \( \Omega = \text{diag} \{ H_t \} \), \( K = \text{diag} \{ K_t \} = 2 (I_T \otimes Q) \text{diag} \{ H_t \otimes H_t \} (I_T \otimes Q') \), \( G = (X', S') \), \( \Sigma = \text{diag} \{ \Omega, K \} \) and \( \nu = (u', \varepsilon')' \). Note that \( K^{-1} = \text{diag} \{ K_t^{-1} \} = \frac{1}{2} (I_T \otimes P) \text{diag} \{ H_t^{-1} \otimes H_t^{-1} \} (I_T \otimes P') \), from (7). The partitioned form of the score \( d_\theta (\theta) = (d_\beta (\theta)', d_\alpha (\theta)')' \) is given by
\[ d_\beta (\theta) = T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} W_t' P (H_t^{-1} \otimes H_t^{-1}) P' \varepsilon_t \]
\[ = T^{-1} X' \Omega^{-1} u + T^{-1} W' K^{-1} \varepsilon , \quad - (8b) \]
and
\[ d_\alpha (\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} Z_t' P (H_t^{-1} \otimes H_t^{-1}) P' \varepsilon_t = T^{-1} Z' K^{-1} \varepsilon , \quad - (8c) \]
where we have partitioned \( S_t = (\frac{\partial h_t}{\partial \beta'}, \frac{\partial h_t}{\partial \alpha'}) = (W_t, Z_t) \), and \( S = (W, Z) \).

Finally, the information matrix is
\[ J (\theta) = E \left[ T^{-1} G' \Sigma^{-1} G \right] = E \begin{pmatrix} T^{-1} X' \Omega^{-1} X + T^{-1} W' K^{-1} W & T^{-1} W' K^{-1} Z \\ T^{-1} Z' K^{-1} W & T^{-1} Z' K^{-1} Z \end{pmatrix} \]
\[ - (9) \]

The similarities with the univariate functions in (2.9) - (2.12) are obvious.

§ 9.2 Multivariate ARCH and some other multivariate heteroskedastic models

The multivariate ARCH model proposed by KE appears to be the most important multivariate heteroskedastic model, and applications may be found in Engle et al [1984], Bollerslev et al [1985], and Diebold and Nerlove [1986],
while other developments in semi-nonparametric estimation also allow for multivariate ARCH effects (Gallant and Tauchen [1986]). The ARCH(q) model is given by

$$H_t = A_0 + \sum_{j=1}^{q} \left( I_n \otimes u_{t-j} \right) A_j^* \left( I_n \otimes u_{t-j} \right),$$

where $A_0$ is $n \times n$, symmetric and positive definite, and the $A_j^*$ are $n^2 \times n^2$, symmetric and positive semidefinite, $j = 1, ..., q$. Each $n \times n$ submatrix of $A_j^*$ is also symmetric. $H_t$ may be written (see KE) in vech form as

$$h_t = \alpha_0 + \sum_{j=1}^{q} A_j v_{t-j} = A z_t = Z_t \alpha,$$

where $\alpha_0 = \text{vech } A_0$, the $A_j$ are $\frac{1}{2} n (n+1) \times \frac{1}{2} n (n+1)$ and accommodate the distinct elements of the $A_j^*$, respectively, for $j = 1, ..., q$, $A = (\alpha_0, A_1, ..., A_q)$, $z_t = (1, v_{t-1}', ..., v_{t-q}')'$, $Z_t = I_{n(n+1)/2} \otimes z_t'$, and $\alpha = \text{vec } A'$.

The multivariate ARCH class retains the most interesting properties of its univariate version, as shown by KE:

- $\beta \alpha = 0$ and thus $\hat{\beta}$ and $\hat{\alpha}$ are asymptotically independent. This follows from the fact that $W_t$ is conditionally odd and $Z_t$ is conditionally even in $\mathcal{F}_t$ (Theorem 5 of KE),

- $h_t$ is linear-in-$\alpha$, which is clear from (10),

- wide sense stationarity obtains if the roots of $\mid I - \sum_{j=1}^{q} A_j z_j \mid$ lie outside the unit circle, given first order stationarity (Theorem 2 of KE).

- the normal ARCH is leptokurtic.

Now dim $(\alpha) = \frac{1}{2} n (n+1) [1 + \frac{1}{2} q n (n+1)]$, which for a first order process grows with $n$ as $(2, 12, 42, 110, 240, 462, 812, ...)$, and estimation soon becomes unfeasible for sample sizes typical in economics, posing a problem which is similar to that of VAR models (Sims [1980], Doan et al [1984]). This problem may be solved by reducing the dimensionality of the process driving the
heteroskedasticity (Diebold and Nerlove [1986], Engle [1987]), or by imposing zero restrictions (Bollerslev et al [1985]). The reparameterization to reduce the number of parameters may be used at the same time to ensure the positive definiteness of $H_t$.

The GARCH($q_1$, $q_2$) model is better presented in vech form as

$$h_t = \alpha_0 + \sum_{j=1}^{q_1} A_j h_{t-j} + \sum_{j=1}^{q_2} A_{q_1+j} v_{t-j}, \quad -(11)$$

for $q_2 > 0$, or

$$A_1(L) h_t = \alpha_0 + A_2(L) v_t,$$

where $A_1(L) = I_{m(n+1)/2} - \sum_{j=1}^{q_1} A_j L^j$, and $A_2(L) = \sum_{j=1}^{q_2} A_{q_1+j} L^j$, or using (1) in multivariate ARMA-type form

$$A_{12}(L) v_t = \alpha_0 + A_1(L) \epsilon_t,$$

where $A_{12}(L) = A_1(L) - A_2(L)$. For wide-sense stationarity we require the roots of $|A_{12}(z)|$ to lie outside the unit circle. The multivariate GARCH, if wide-sense stationary, can be expressed as an infinite order ARCH and thus it is also leptokurtic under normality and retains the asymptotic independence between $\hat{\beta}$ and $\hat{\alpha}$. To establish conditions for fourth order stationarity we first prove

**Lemma 9.1.** - If $y_t | \mathcal{F}_t \sim N[\mu_t(\beta), H_t]$, then

$$\text{vec} \ E[v_t v'_t | \mathcal{F}_t] = (2R + I_N) \text{vec} h_t h'_t,$$

where $R = (Q \otimes Q)(I_N \otimes I_{(n^2, n^2)})(P' \otimes P')$, and $N = \frac{1}{T} n^2 (n + 1)^2$. \hfill \Box

This is a simple extension of the 3 $\sigma^4$ kurtosis result for the univariate normal distribution, and with it we can prove
Theorem 9.2. - If the multivariate normal GARCH(1,1) is wide-sense stationary, a necessary and sufficient condition for fourth order stationarity is that the characteristic roots of

\[ A_1 \otimes A_1 + A_1 \otimes A_2 + A_2 \otimes A_1 + (2R + I)(A_2 \otimes A_2) \]

be smaller than unity, where \( R \) and \( I \) are of order \( [\frac{1}{2}n(n+1)]^2 \), and \( R \) is given in Lemma 9.1.

Proof: Using Lemma 9.1 and (11) we have that

\[ \text{vec } E [v_t v_t' \mid \mathcal{F}_t] = (2R + I) \text{vec } (\alpha_0 + A_1 h_{t-1} + A_2 v_{t-1}) (\alpha_0 + A_1 h_{t-1} + A_2 v_{t-1})'. \]

Let \( \kappa = \text{vec } E [v_t v_t'] \) exist. Then by simple algebra and conditional expectations

\[ \kappa = (2R + I) \text{vec } \{ \alpha_0 [\alpha_0 + (A_1 + A_2) \sigma]' + (A_1 + A_2) \sigma \alpha_0' \} + \]

\[ (2R + I)(A_1 \otimes A_1 + A_1 \otimes A_2 + A_2 \otimes A_1) \text{vec } E [h_{t-1} h_{t-1}'] + (2R + I)(A_2 \otimes A_2) \kappa \]

\[ = (2R + I) \text{vec } \{ \alpha_0 [\alpha_0 + (A_1 + A_2) \sigma]' + (A_1 + A_2) \sigma \alpha_0' \} + \]

\[ [A_1 \otimes A_1 + A_1 \otimes A_2 + A_2 \otimes A_1 + (2R + I)A_2 \otimes A_2] \kappa, \]

where \( \sigma = E [h_t] \) and the last equality is based on Lemma 9.1 and conditional expectations. Bringing the terms in \( \kappa \) to the left-hand-side then establishes necessity, and sufficiency is obvious by a similar argument.

Not surprisingly, this result is of the same form as the one established by Bollerslev [1986] for the univariate case.

Other models are easily generalized to the multivariate case preserving most of their properties. The multivariate linear simple heteroskedasticity model is in vech form (see KE)

\[ h_t = A z_t = Z_t \alpha, \]

where \( z_t \in \mathcal{F}_t, Z_t = I_{n(n+1)/2} \otimes z_t', \) and \( \alpha = \text{vec } A' \).
The Poisson-type model can be generalized with

\[ h_t = \alpha_0 + A_1 \mu_t = A z_t = Z_t \alpha, \]

where \( A = (\alpha_0, A_1) \) and \( z_t = (1, \mu_t')' \).

The Amemiya model can also be extended by making

\[ H_t = A_0 + (I_n \otimes \mu_t') A_1^* (I_n \otimes \mu_t), \]

so that similarly to the ARCH class each element of \( H_t \) is allowed to depend on a symmetric quadratic form in \( \mu_t \). This can also be written in vech form as

\[ h_t = \alpha_0 + A \text{vech} \mu_t \mu_t' = A z_t = Z_t \alpha, \]

where \( \alpha_0 = \text{vech} A_0 \), \( A = (\alpha_0, A_1) \), and \( z_t = (1, [\text{vech} \mu_t \mu_t']')' \).

### § 9.3 Estimation and likelihood factorization

Consider first extracting the information from the mean equation by using the orthogonality conditions

\[ \psi_m (\beta) = T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} u_t = T^{-1} X' \Omega^{-1} u, \]

and using Hansen's [1982] Theorems 2.1 and 3.1 or White and Domowitz's Theorems 3.1 and 3.2 we have that \( \hat{\beta}_m \) such that \( \psi (\hat{\beta}_m) = 0 \) is a strongly consistent estimator for \( \beta_0 \), the true value of \( \beta \), and has asymptotic distribution.

\[ T^{1/2} (\hat{\beta}_m - \beta_0) \Rightarrow T^{1/2} (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} u + o_p (1) \quad \Rightarrow \quad N \left[ 0, \sigma \left( T^{-1} X' \Omega^{-1} X \right)^{-1} \right]. \]

Moreover, using a parametric estimate \( H_t (\tilde{\theta}) \) to construct \( \hat{\beta}_m \) does not alter this distribution provided \( \tilde{\theta} \) is root-T consistent.

Similarly, to extract the information from the variance equation we use the orthogonality conditions
\[ \psi(v) = T^{-1} \sum_{t=1}^{T} S_t' K_t^{-1} \varepsilon_t = T^{-1} S' K^{-1} \varepsilon, \]
and provided \( \theta \) is identifiable we have that

\[ T^{1/2} (\hat{\psi} - \psi_0) = T^{1/2} (S' K^{-1} S)^{-1} S' K^{-1} \varepsilon + o_p(1) \]

\[ \mathcal{L} \rightarrow N[0, \mathcal{C}(T^{-1} S' K^{-1} S)^{-1}], \]

where \( \psi_v(\hat{\psi}_v) = 0 \) and \( \psi_0 \) is the true value of \( \psi \). For simplicity we retain the assumption of \( \theta_0 \) uniquely identifiable in the variance equation. Substituting parametric root-\( T \) consistent estimates of \( K_t \) does not affect the asymptotic distribution. An alternative for both \( \xi_t \) in \( \psi_m(\beta) \) and \( K_t \) in \( \psi_v(\theta) \) might be to generalize Carroll's [1982] or Robinson's [1987] nonparametric estimates, but whether this may be of practical use remains to be seen. The unobservability of \( v_t \) is solved by means of

**Lemma 9.3.** - Given (C10-M) - (C17-M) and

(i) a root-\( T \) consistent estimator \( \hat{\beta} \) of \( \beta_0 \),

(ii) a bounded function \( F_t(\pi) \in \mathcal{F}_t \), where \( \pi \) is a parameter vector

for which a root-\( T \) consistent estimator \( \tilde{\pi} \) is available, such that

\[ \tilde{F}_t = F_t(\tilde{\pi}) \in \mathcal{F}_t \]

is bounded uniformly in \( t \),

then

\[ T^{1/2} \sum_{t=1}^{T} \tilde{F}_t \text{vec}(\tilde{u}_t \tilde{u}_t' - u_t u_t') \xrightarrow{as} 0. \]

Therefore we can substitute \( \tilde{u}_t \tilde{u}_t' \) in \( \psi_v(\theta) \) without asymptotic penalty to the order used in the distribution of \( \hat{\theta}_v \). Equivalently \( \tilde{v}_t = \text{vech} \tilde{u}_t \tilde{u}_t' \) takes the place of \( v_t \) in the variance equation to produce its operative version

\[ \tilde{v}_t = h_t(\theta) + e_t, \]

where \( e_t = \varepsilon_t + (\tilde{v}_t - v_t) \). The issue of having \( \tilde{\theta} \) available to obtain \( \tilde{H}_t \) for \( \psi_m(\beta) \) is solved by least-squares in the mean equation and simple least squares in the variance equation, and the existence of GARCH components can be dealt with by using the Hannan-Rissanen procedure as in the univariate case. If, as in
the normal case, the conditional pdf is fully characterized by its first two moments, the availability of \( \hat{\theta} \) suffices to construct \( \tilde{K}_t \) for the computation of \( \hat{\theta}_V \). In the t-distribution we have that the fourth-moment matrix is proportional to the corresponding normal form and thus \( \hat{\theta} \) is sufficient to calculate \( \tilde{\theta}_V \), though not its covariance matrix. In other cases we need to parameterize \( K_t \), extend Lemma 9.3 to fourth order moments, and use the corresponding simple least squares estimators. Alternatively, we may attempt using the semi-parametric approach of Carroll [1982] or Robinson [1987].

Given the properties of the QMLE - the MLE assuming normality - that will be detailed below, a sensible strategy for practical purposes is to use an estimator constructed as a MWA, using White's [1980b] robust covariance matrix in the variance equation. Moreover, comparing such an estimator with the QMLE may provide useful information about the specification of the kurtosis of the conditional distribution.

To combine the information in both moments we first consider the conditionally normal case. Here again the score suggests that the likelihood can be locally factorized. Indeed, let \( \tilde{u}_t \tilde{u}_t' \) and \( \tilde{H}_t^{-1} \) be functions of the data alone, and note that

\[
\tilde{u}_t' \tilde{H}_t^{-1} \tilde{u}_t = \text{vec} \ u_t \tilde{H}_t^{-1} u_t = ( \tilde{u}_t' \odot \tilde{u}_t' ) \text{vec} \ H_t^{-1}.
\]

We use the Mean Value Theorem to obtain a first order expansion of this function (of \( u_t \odot u_t \) and \( \text{vec} \ H_t^{-1} \)) around \( \tilde{u}_t \odot \tilde{u}_t \) and \( \text{vec} \ \tilde{H}_t^{-1} \),

\[
\tilde{u}_t' \tilde{H}_t^{-1} \tilde{u}_t = ( \tilde{u}_t' \odot \tilde{u}_t' ) \text{vec} \ \tilde{H}_t^{-1} + ( \tilde{u}_t' \odot \tilde{u}_t' ) ( \text{vec} \ H_t^{-1} - \text{vec} \ \tilde{H}_t^{-1} )
\]

\[
+ [ ( \tilde{u}_t' \odot \tilde{u}_t' ) - ( \tilde{u}_t' \odot \tilde{u}_t' ) ] \text{vec} \ \tilde{H}_t^{-1}
\]

\[
= [ \tilde{u}_t' \tilde{H}_t^{-1} \tilde{u}_t - \tilde{u}_t' \tilde{H}_t^{-1} \tilde{u}_t - \tilde{u}_t' \tilde{H}_t^{-1} \tilde{u}_t + \tilde{u}_t' H_t^{-1} \ u_t + u_t' \tilde{H}_t^{-1} u_t ]
\]
where $\bar{u}_t \in [u_t, \tilde{u}_t]$ and $\text{vec} \bar{H}_t^{-1} \in [\text{vec} H_t^{1}, \text{vec} \tilde{H}_t^{1}]$. Let $\tilde{u}_t = u_t(\tilde{\beta})$ and $\tilde{H}_t = H_t(\tilde{\theta})$ for root-$T$ consistent $\tilde{\beta}$ and $\tilde{\theta}$ and consider approximating the log-likelihood by

$$\mathcal{L}^*(\theta) = \mathcal{L}_m(\beta) + \mathcal{L}_v(\theta),$$

where

$$\mathcal{L}_m(\beta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} u_t' \tilde{H}_t^{-1} u_t,$$

and

$$\mathcal{L}_v(\theta) = -\frac{1}{2} T^{-1} \sum_{t=1}^{T} \log |H_t^{1}| - \frac{1}{2} T^{-1} \sum_{t=1}^{T} \tilde{u}_t' H_t^{1} \tilde{u}_t.$$

It is clear that

$$T^{-1/2} \sum_{t=1}^{T} X_t' (\tilde{H}_t^{-1} - H_t^{-1}) u_t \xrightarrow{a.s.} 0, \quad -(13a)$$

and this is what makes $\tilde{\beta}_m$ feasible. Also,

$$T^{-1/2} \sum_{t=1}^{T} S_t' (H_t^{1} \otimes H_t^{1}) \text{vech} (\tilde{u}_t \tilde{u}_t' - u_t u_t') \xrightarrow{a.s.} 0, \quad -(13b)$$

from Lemma 9.3. We then have

**Theorem 9.4.** - Under $(\mathcal{C}0\text{-}M) - (\mathcal{C}8\text{-}M)$ the log-likelihoods $\mathcal{L}^*(\theta)$ and $\mathcal{L}(\theta)$ produce estimators of $\theta$ with the same asymptotic distribution. \hfill $\blacksquare$

Therefore the QMLE $\hat{\theta}$ may be obtained alternatively from maximizing $\mathcal{L}(\theta)$ or $\mathcal{L}^*(\theta)$, and its distribution under correct specification is given in

**Theorem 9.5.** - The QMLE $\hat{\theta}$ under $(\mathcal{C}0'\text{-}M) - (\mathcal{C}8\text{-}M)$ is a strongly consistent estimator of $\theta_0$ and has asymptotic distribution

$$T^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N \left[ 0, \mathbf{f}(\theta_0)^{-1} \right]. \quad -(14) \quad \square$$

KE proved this result for the ARCH and multivariate simple heteroskedasticity model by conforming to the conditions in Crowder [1976]. Our proof allows for more heterogeneity of the process and covers a wider
range of multivariate heteroskedastic models. An immediate consequence is
that the MWA structure of the MLE is preserved in a multivariate setting
because $\mathcal{L}^\ast(\theta)$ fulfills the conditions of Theorem 3.12, and we have

**Corollary 9.6.** - Under the assumptions of Theorem 9.4,

\[
T^{1/2}(\hat{\beta} - \beta_0) = (I_k - \Pi)^{1/2}(\hat{\beta}_m - \beta_0) + \Pi T^{1/2}(\hat{\beta}_v - \beta_0) + o_p(1)
\]

\[\xrightarrow{d} \mathcal{N}[0, \{V(\hat{\beta}_m)^{-1} + V(\hat{\beta}_v)^{-1}\}^{-1}], \quad -(15)\]

where $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')$, $\theta_0 = (\beta_0', \alpha_0')$, and

\[\Pi = V(\hat{\beta})V(\hat{\beta}_v)^{-1} = I_k - V(\hat{\beta})V(\hat{\beta}_m)^{-1}. \quad \square\]

The covariance matrix of $\hat{\beta}$ may be expressed in terms of limits of
expectations of data matrices using partitioned inversion in (14) or alternatively
in (12) combined with (15). This matrix has the same form as the one given in
Corollary 3.7, and the covariance matrix of $\hat{\alpha}$ has the same form as in
Corollary 3.8. The likelihood factors are the generalization of those in the
univariate case. Now the conditional distribution of $y_t$ is normal and that of
$u_t u_t'$ is Wishart, $W(H_t; n, 1)$ (e.g. Zellner [1971]). The force of the QMLE
resides on its having the same asymptotic distribution as the estimator
obtained from the joint orthogonality conditions when normality holds. If
normality is not maintained, this shows in the covariance matrix of $\hat{\alpha}$, but
using a robust estimator to compute this covariance matrix and hence the
MWA appears as an attractive option. We have

**Theorem 9.7.** - Under $(Q0-M) - (Q7-M)$ and symmetry of the conditional
distribution, the estimator $\hat{\theta}_T$ obtained from the orthogonality conditions
$\psi(\theta) = (\psi_m(\theta)', \psi_v(\theta)')'$ and weighting matrix $A_T = \text{diag}(T(X'\Omega^{-1}X)^{-1},$

\[T(S'K^{-1}S)^{-1})\] , is strongly consistent for $\theta_0$ and has asymptotic distribution

\[T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}[0, \Sigma(T^{-1}\Sigma^{-1}G)^{-1}]. \quad \square\]
It follows that \( T^{1/2} (\hat{\theta} - \theta_0) = T^{1/2} (\hat{\theta}_j - \theta_0) + o_p(1) \) under normality, and the optimality of \( \hat{\theta}_j \) for the given orthogonality conditions follows by analogy with the weighting matrix in (3.18) (see Chamberlain [1987]). Because of the MWA structure of the joint estimator, we can separate the information about \( \beta \) arising from each of the moments.

We finish this section by discussing the conditions under which we would obtain equivalent estimators using the full system information and the limited information of each equation. In homoskedastic systems, Zellner [1962] proved that full information (FI) and limited information (LI) estimators of \( \beta \) would coincide if either

(a) \( \partial \mu_{ij} / \partial \beta_j = x_t \neq j \), so that \( X_t = (I_n \otimes x_t') \) and \( k = n k^* \), \( k^* = \dim (x_t) \); or

(b) the conditional covariance matrix is diagonal.

Consider (a) under heteroskedasticity. Then

\[
\psi_m(\beta) = T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} u_t = T^{-1} \sum_{t=1}^{T} (I_n \otimes x_t') H_t^{-1} u_t
\]

\[
= T^{-1} \sum_{t=1}^{T} (H_t^{-1} \otimes I_k^*) (I_n \otimes x_t) u_t.
\]

When \( H_t = H \neq t \), the nonsingular matrix \( H_t^{-1} \otimes I_k^* \) can be factored out of the sum and the orthogonality conditions are equivalently given by

\[
\psi_m(\beta) = T^{-1} \sum_{t=1}^{T} (I_n \otimes x_t) u_t = T^{-1} \sum_{t=1}^{T} X_t' u_t,
\]

which define the single equation OLS estimators. This does not carry through to the heteroskedastic case because \( H_t^{-1} \otimes I_k^* \) cannot be factored out of the sum.

Now consider (b) under heteroskedasticity. Partition \( \beta = (\beta_1', \ldots, \beta_n')' \) and \( X_t = \text{diag} \{x_{1t}', x_{2t}', \ldots, x_{nt}'\} \), where \( x_{jt} = \partial \mu_{jt} / \partial \beta_j \) and we assume \( \partial \mu_{jt} / \partial \beta_j = 0 \) for \( i \neq j \), so no cross equation restrictions are present. Then it is seen that

\[
\psi_m(\beta) = T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} u_t = (T^{-1} \sum_{t=1}^{T} h_{1t} x_{1t}' u_{1t}, \ldots, T^{-1} \sum_{t=1}^{T} h_{nt} x_{nt}' u_{nt}').
\]
where \( \psi_{mj}(\beta) \) are the LI orthogonality conditions for the j-th equation. Thus it appears that Zellner's condition (b) also applies under heteroskedasticity. The statement needs to be qualified, though. Suppose there are cross-equation restrictions in the variance equations. These might appear because of cross-restrictions between the \( \alpha \) parameters, or also because of \( h_{ijt} = h_{ijt}(\beta_j) \), amongst other arguments, for \( i \neq j \). The latter would be in fact the natural way to link the variances across equations. For example, in the multivariate ARCH model this would happen if \( h_{ijt} \) depends on \( u_{jt}^2 \). When this is the case, the orthogonality conditions \( \psi_m(\beta) \) use limited information with respect to the mean equation, but the estimates of \( h_t \) require full information with respect to the variance equation. Therefore, for \( \psi_m(\beta) \) to define real single equation estimators we need the variance equation to be equivalently estimated by LI or FI. Partition \( \alpha = (\alpha_1', ..., \alpha_n') \) and \( S_t = \text{diag}(s_1t',...,s_nt') \), where \( s_{jt} = \partial h_{tij}/\partial \theta_j = 0 \) for \( i \neq j \), then

\[
\psi_v(\theta) = (\psi_{v1}(\theta_1)',...,\psi_{vn}(\theta_n)')', \tag{17}
\]

where \( \psi_{vj}(\theta_j) = T^{-1} \sum_{t=1}^T \xi_{jt} s_{jt} \varepsilon_{jt} \), provided \( K_t \) is diagonal, by analogous argument to \( \psi_m(\beta) \).

We then have

**Theorem 9.8.** - If both \( H_t \) and \( K_t \) are diagonal and there are no cross-equation restrictions in the mean and variance equations, then the FI estimator \( \hat{\theta}_J \) is equivalent to the set of LI estimators \( \hat{\theta}_{ji} \), \( i = 1, ..., n \).

**Proof:** The result follows from the above argument on (16) and (17).

Once we have assumed \( H_t \) to be diagonal, it does not appear very restrictive to take the kurtosis function \( K_t \) (not \( K_t^* \)) as diagonal, too. Take, for example, a conditionally normal or \( t \) distribution. For these \( K_t \) is diagonal if,
and only if, $H_t$ is diagonal. Note that in the $t$ distribution there are cross-
restrictions in the kurtosis equations, since they all depend on the degrees-of-
freedom parameter, but the parameters of the variance equations remain
variation-free. In the normal case when $H_t$ is diagonal,

$$ f(y_t | \mathcal{F}_t) = (2\pi)^{-n/2} \left( \prod_{j=1}^{n} h_{jt}^{-1/2} \right) \exp \left\{ - \frac{1}{2} \sum_{j=1}^{n} h_{jt}^{-1} u_{jt}^2 \right\} = \prod_{j=1}^{n} f(y_{jt} | \mathcal{F}_0), $$

and therefore

$$ \ell(\theta) = \sum_{j=1}^{n} \ell_j(\theta), $$

so the likelihood can be factorized in terms of each component of $y_t$. If there
are no cross-equation restrictions $\ell_j(\theta)$ is the sole provider of information
about $\theta_j$ and the equivalence of the FI and LI MLE's is immediate. If there are
cross-equation restrictions, then from Ruud [1984] or from Theorem 3.12 the FI
estimators of jointly identifiable parameters are MWA's of the corresponding
LI estimators.

§ 9.4 A note on specification error

Let us consider briefly the effects of misspecification on the QMLE $\hat{\theta}$. We
assume that the pseudo-true value $\theta^*$ exists (i.e. multivariate version of (30)
of Chapter 4 referred to as (30-M)). Lemma 4.1 does not depend on the
dimension of $y_t$ and thus applies equally well here. Let

$$ \varphi_\beta(\theta) = \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} X_t' H_t^{-1} E [u_t | \mathcal{F}_t] \right\} + \frac{1}{2} \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} W_t' P (H_t^{-1} \otimes H_t^{-1}) P' E [\epsilon_t | \mathcal{F}_t] \right\}, $$

$$ \varphi_\alpha(\theta) = \frac{1}{2} \mathcal{E} \left\{ T^{-1} \sum_{t=1}^{T} S_t' P (H_t^{-1} \otimes H_t^{-1}) P' E [\epsilon_t | \mathcal{F}_t] \right\}, $$

and

$$ \varphi_{\theta}(\theta) = (\varphi_\beta(\theta)', \varphi_\alpha(\theta))'. $$
Then Lemma 4.2 extends obviously to the multivariate case, and we have

**Lemma 9.9.** - Under (B0-M) and (C1-M) - (C7-M),

(i) \( \hat{\beta} \) is a consistent estimator of \( \beta_0 \) if, and only if, \( \varphi_{\beta} (\beta_0, \alpha^*) = 0 \),

(ii) \( \hat{\alpha} \) is a consistent estimator of \( \alpha_0 \) if, and only if, \( \varphi_{\alpha} (\beta^*, \alpha_0) = 0 \),

and

(iii) \( \hat{\theta} \) is a consistent estimator of \( \theta_0 \) if, and only if, \( \varphi_{\theta} (\theta_0) = 0 \).

When there is specification error in either \( \mu_t \) or \( H_t \) the conditional expectations \( E[ u_t | F_t] \) and \( E[ \varepsilon_t | F_t] \) are non-zero and, unless the DGP has structure for these expectations to vanish unconditionally when combined with the other functions in \( \varphi_{\theta} \), the result will be inconsistency of \( \hat{\theta} \). This is generally the case, but as in the univariate model there may be instances in which consistency is preserved. For example, Lemma 4.3 for the presence of autocorrelation in \( \mu_t \) extends to the multivariate setting. Similarly, \( \hat{\beta} \) remains consistent when \( H_t \) is not parameterized as a function of \( \beta \) regardless of the form of the true variance, and essentially the same conditions that combine even and odd effects make the multivariate ARCH class robust to certain departures from the null of correct specification.

The consequences of specification error in moments of order higher than second is analyzed in

**Lemma 9.10.** - Under (B0-M) and (C1-M) - (C7-M), the QMLE \( \hat{\theta} \) is consistent regardless of the form of the DGP provided that \( \mu_t \) and \( H_t \) are correctly specified.

Therefore misspecifying the conditional symmetry or kurtosis, or any conditional moment of higher order, does not impinge upon the consistency of \( \hat{\theta} \). When misspecification is present in third or fourth moments, the true conditional covariance matrix of \( v_t \) is not \( \Sigma_t = \text{diag} \{ H_t, 2Q(H_t^{-1} \otimes H_t^{-1})Q' \} \),
but rather a more general form $\Sigma_{M_t}$, say. In such cases, the correct asymptotic distribution of $\hat{\theta}$ is

$$T_{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N \left[ 0, \Sigma \left( T (G' \Sigma^{-1} G)^{-1} G' \Sigma^{-1} \Sigma_{M} \Sigma^{-1} G (G' \Sigma^{-1} G)^{-1} \right) \right],$$

where $\Sigma_{M} = \text{diag} \left( \Sigma_{M_t} \right)$. Thus the covariance matrix of $\hat{\theta}$ is incorrect from ML output, but a correct estimate of $V(\hat{\theta})$ can be obtained using White's [1980b] procedure. Observe that the first $n \times n$ block of $\Sigma_t$ is correct because $H_t$ is the conditional covariance matrix of $y_t$, and thus the correction needs only apply to the variance equation, and to the off-diagonal block in case of asymmetry. If specification error is present in conditional moments of order five and higher, the distribution of $\hat{\theta}$ remains as in (14), but efficiency may be improved.

§ 9.5 Consistency tests

The situation depicted in § 9.4 is very much the same as that for the univariate model given in Chapter 4 and thus calls for careful assessment of model specification. Researchers usually concentrate on evaluating $\mu_t$, but the need to assess $H_t$ with equal effort must be emphasized because it may have a rebound effect on $\hat{\beta}$. In this section we extend the coherency and consistency tests of § 5.1 and § 5.2 to the multivariate setting. To treat local parametric alternatives ($\mathcal{Q}0''$) of Chapter 5 is extended as

$$\mathcal{Q}0''-M \quad y_t \mid \mathcal{F}_t \sim N \left[ \mu_t(\beta, \lambda_T), H_t(\beta_0, \lambda_T) \right],$$

where $\lambda_T = \lambda_0 + T^{-1/2} \delta$, for fixed $\lambda_0$ and $\delta$, and correct specification obtains when $\delta = 0$. We define $M_{\lambda t} = \frac{\partial \mu_t}{\partial \lambda_t'}$, $H_{\lambda t} = \frac{\partial h_t}{\partial \lambda_t'}$, $M_{\lambda} = (M_{\lambda 1}', ..., M_{\lambda T}')'$, and $H_{\lambda} = (H_{\lambda 1}', ..., H_{\lambda T}')'$. Coherency tests have an intuitive appeal. There are now available two asymptotically independent estimators of $\beta$, namely $\hat{\beta}_m$ and $\hat{\beta}_v$. Define
q = \hat{\beta}_m - \hat{\beta}_v, so that V(q) = V(\hat{\beta}_m) + V(\hat{\beta}_v), and this matrix is positive definite.

We then have

**Theorem 9.11.** - Under \((\Omega^{0''-M}) - (\Omega^{8-M})\),

\[
\tau = T q' \hat{V}(q)^{-1} q \overset{d}{\rightarrow} \chi^2_k \{ k; \delta' D' V(q)^{-1} D \delta \},
\]

where \(D = V(\hat{\beta}_m) \in \{ T^{-1} X' \Omega^{-1} M_\lambda \} - (I_k, 0) V(\hat{\theta}_v) \in \{ T^{-1} S' K^{-1} H_\lambda \} \).

Therefore we have a central \(\chi^2_k\) distribution under the null hypothesis \((\delta = 0)\), and a non-central \(\chi^2_k\) under local parametric alternatives. The test may be inconsistent because it may be that \(q \not\sim 0\) under the alternative. This happens, for example, when \(\hat{\beta}\) remains consistent in the presence of variance misspecification. The test in Theorem 9.11 is a full information test that simultaneously contrasts all the information about \(\beta\) in the system of mean equations with all the information in the system of variance and covariance equations. We can also perform limited information tests from paired mean and variance single equations, contrasting their jointly identifiable functions as in Theorem 5.1. The limited information tests may be useful to suggest which equations are more likely to be causing trouble when \(\tau\) rejects the null hypothesis.

Consider now the more general class of consistency tests, which are defined from the first order conditions

\[
m(\Phi, \theta) = T^{-1} \sum_{t=1}^{T} m_t(\Phi_t, \theta) = T^{-1} \sum_{t=1}^{T} \Phi_t u_t = T^{-1} \Phi' v,
\]

where now the \(\Phi_t = (\Phi_{1t}, \Phi_{2t}) \in \mathcal{F}_t\) are \(r \times N\) matrices, where \(r\) is the number of orthogonality conditions being tested and \(N = n + \frac{1}{2} n(n+1) = \frac{1}{2} n(n+3)\). Also, \(\Phi_j = (\Phi_{j1}, ..., \Phi_{jT})'\), \(j = 1, 2\) and \(\Phi = (\Phi_1', \Phi_2')'\), so \(\Phi_1\) is \(nT \times r\), \(\Phi_2\) is \((N-n)T \times r\), and \(\Phi\) is \(NT \times r\). Clearly, \(E[ m(\Phi, \theta) ] = 0\) and

\[
V_{cT} = V_{cT}(\Phi, \theta) = \text{Var}[ T^{1/2} m(\Phi, \theta) ] = E[ T^{-1} \Phi' \Sigma \Phi ].
\]
The consistency test-statistics are based on the asymptotic distribution of the corresponding sampling moments $m(\hat{\Phi}, \hat{\theta})$. Tests designed for the mean equation have $\Phi_2t = 0 \neq t$, and tests that wish to concentrate on the variances and covariances have $\Phi_1t = 0$. The simplest consistency tests are based on $m_u = T^{-1} \sum_{t=1}^{T} \hat{u}_t$ and $m_h = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_t$, by setting $\Phi_t = (I_n, 0)$ and $\Phi_t = (0, I_{n,n})$, respectively. The simple joint test considers $\Phi_t = I_N$.

To obtain the distribution of $m(\hat{\Phi}, \hat{\theta})$ we need the matrix of expected derivatives of the orthogonality conditions, which takes the form

$$M(\theta_0) = \mathcal{E} \left\{ \frac{d}{d\theta} m_\theta(\theta_0) \right\} = -\mathcal{E} \left\{ T^{-1} \Phi' G \right\},$$

as in (5.22). We then have the generalization of Theorem 5.3.

**Theorem 9.12.** Under $(\mathcal{Q}0^0{-}\text{M})$ - $(\mathcal{Q}8{-}\text{M})$ and the sequence $\{\Phi_t\}$, $\Phi_t \in \mathcal{F}_t$, being such that $g^* = (m(\Phi, \theta)', d_\theta')'$ obeys the regularity, continuity, dominance and mixing conditions in Assumptions (1) - (6) of Newey [1985b], then

$$T^{1/2} m(\hat{\Phi}, \hat{\theta}) \xrightarrow{d} \mathcal{N} [\psi, Q_\phi],$$

where

$$\psi = \left\{ \mathcal{E} \left\{ T^{-1} \Phi' G_\lambda \right\} - M(\theta_0) V(\hat{\theta}) \mathcal{E} \left\{ T^{-1} \Phi' \Sigma^{-1} G_\lambda \right\} \right\} \delta,$$

$$Q_\phi = V_c - M(\theta_0) V(\hat{\theta}) M(\theta_0)',$$

$G_\lambda = (M_\lambda', H_\lambda')'$, and $V_c = \lim V_{cT}$, with all expectations evaluated under $H_0$. $Q_\phi$ is consistently estimated by

$$\hat{Q}_\phi = T^{-1} \Phi' \hat{\Sigma}^{1/2} \hat{m} \hat{\Sigma}^{1/2} \hat{\Phi},$$

where $\hat{m} = I_{NT} - \Sigma^{-1/2} G (G' \Sigma^{-1} G')^{-1} G' \Sigma^{-1/2}$, and all estimates are under $H_0$. □

The test-statistic and its distribution follow immediately and a simplified calculation is available in a wide range of cases, as is shown in
Theorem 9.13. - Under the assumptions of Theorem 9.12,

\[ s = T m (\hat{\Phi}, \hat{\theta})' \hat{Q}_\phi^{-1} m (\hat{\Phi}, \hat{\theta}) \overset{d}{\rightarrow} \chi^2 \left[ r ; \psi' Q_\psi^{-1} \psi \right], \tag{18} \]

and if \((G, \Sigma \Phi)\) has full column rank then \(s - s^* \overset{\text{as}}{\rightarrow} 0\), where \(s^* = N^T R_0^2, N = n + \frac{1}{2} n (n + 1) = \frac{1}{2} n (n + 3)\), and \(R_0^2\) is the uncentered coefficient of determination of the regression of \(\hat{\nu}\) on \(\hat{G}\) and \(\hat{\Sigma} \Phi\) in the metric of \(\hat{\Sigma}\).

Theorems 9.12 and 9.13 provide a wide range of tests for the first two conditional moments, extending the results of KE. By appropriate choice of \(\Phi_t\) the tests may be based on the whole system of equations or on a single equation. The dimension of the auxiliary regression may be reduced provided that the excluded mean or variance equations do not contain any information about the parameters in the moments being tested and thus it may be feasible to construct limited information tests. This happens, for example, in the ARCH tests considered by KE. Also under some circumstances (independence of estimators of the same parameters in the included and excluded auxiliary regressions and diagonal information matrix between the parameters of the included equations and the rest), tests with asymptotic correct size can be constructed from a reduced auxiliary regression, but they will be less powerful than those obtained from the full auxiliary regression.

Many consistency test statistics may be constructed without any information external to the model and thus can be provided easily with MLE output. If there exists external information as to the source of potential departures from the null hypothesis, this can also be used for diagnostic purposes because the LM test for variable additions is, again, a member of the consistency tests family. Suppose we have the multivariate version of \((C_0' - a)\),

\[(C_0'\cdot aM) \quad y_t | \mathcal{F}_t \sim N \left[ \mu_t (\beta, \beta_A), H_t (\theta, \theta_A) \right],\]

then it is clear from (8) that the subvector of the score for \(\theta_A\) evaluated under \(\theta_A = 0\) is
\[ d_A(\theta) = T^{-1} \sum_{t=1}^{T} X_{At}'H_t^{-1}u_t + \frac{1}{2} T^{-1} \sum_{t=1}^{T} S_{At}'P(H_t^{-1} \otimes H_t^{-1})P'e_t, \quad (19) \]

where \( X_{At} = \frac{\partial \mu_t}{\partial \theta_A'} \) and \( S_{At} = \frac{\partial h_t}{\partial \theta_A'} \). Therefore let \( \Phi_t = (X_{At}', S_{At}') \Sigma_t^{-1} \), \( \Phi = \Sigma^{-1} G_A \), where \( G_A = (X_A', S_A')' \), \( X_A = (X_{A1}', \ldots, X_{At}')' \), and \( S_A = (S_{A1}', \ldots, S_{At}')' \). The LM test for \( \theta_A = 0 \) is the consistency test with this choice of \( \Phi \), and we have

**Theorem 9.14** - Under the assumptions of Theorem 9.12 the LM test for \( H_0 : \theta_A = 0 \) against the alternative \( H_1 : \theta_A \neq 0 \) can be obtained as \( s_{LM} = NTR_0^2 \) from the regression of \( \hat{\nu} \) on \( \hat{G} \) and \( \hat{G}_A \) in the metric of \( \hat{\Sigma} \), with all estimates under \( H_0 \). Under local parametric alternatives in the direction of \( H_1 \)

\[ s_{LM} \overset{d}{\rightarrow} \chi^2[r; \delta' \in \{ T^{-1} G_A' \Sigma^{-1/2} \mathcal{M}_r \Sigma^{-1/2} G_A \} \delta], \]

where \( r = \text{dim } \theta_A \).

Theorems 9.12 - 9.14 constitute a basic tool to diagnose multivariate heteroskedastic models both against specific alternatives and as general tests of misspecification of the first two moments without use of information external to the model.

§ 9.6 Efficiency tests

In this section we assume correct specification of \( \mu_t \) and \( H_t \) and are concerned with departures from conditional normality. To evaluate the model in the face of these possibilities we extend the efficiency tests of section § 5.3.

We need to introduce some notation and conventions in order to handle the growing dimensionality of moments in a multivariate situation. All the \( r \)-order products of elements of a vector \( \zeta \in \mathbb{R}^n \) are contained in its \( r \)-th Kronecker power which we denote by \( \zeta^{(r)} \). That is,

\[ \zeta^{(r)} = \otimes_{j=1}^{r} \zeta = \zeta \otimes \zeta \otimes \ldots \otimes \zeta. \]
The vector $\zeta^{(r)}$ is $n^r \times 1$ and can be defined recursively as

$$\zeta^{(r)} = \text{vec } \zeta^{(r-1)} \zeta^r = \zeta \otimes \zeta^{(r-1)}.$$ 

Thus, for example, $\text{vec } E_t = u_t^{(2)} - \text{vec } H_t$. $\zeta^{(r)}$ has repeated elements, and thus if $\zeta$ is a random vector $\zeta^{(r)}$ will have a degenerate distribution. To correct this situation we define $\zeta^{[r]}$ as the vector of distinct elements of $\zeta^{(r)}$. For example, $\zeta^{[2]} = \text{vech } \zeta \zeta^r$, and therefore $e_t = u_t^{[2]} - h_t$. The transformation from $\zeta^{(r)}$ to $\zeta^{[r]}$ is a generalization of the vech operator and is thus a vecp operator (see HS) acting on vec $\zeta^{(r-1)} \zeta^r$. Let $n_r$ be the dimension of $\zeta^{[r]}$, $n_r < n^r$ for $r > 1$. There exists a unique matrix $P_{nr}$ of dimension $n_r \times n^r$ such that

$$\zeta^{(r)} = P_{nr} \zeta^{[r]}, \tag{20}$$ 

and so $P_{nr}$ reallocates the unique elements of $\zeta^{[r]}$ into $\zeta^{(r)}$. There also exists a $n_r \times n^r$ matrix $Q_{nr}$ that selects the elements of $\zeta^{(r)}$ and accommodates them in $\zeta^{[r]}$, that is,

$$\zeta^{[r]} = Q_{nr} \zeta^{(r)}, \tag{21}$$ 

but $Q_{nr}$ is not unique in view of the repetition of elements in $\zeta^{(r)}$. Substituting (20) into (21) we get $\zeta^{[r]} = Q_{nr} P_{nr} \zeta^{[r]}$, and given the non-repetition of the elements of $\zeta^{[r]}$ this holds for arbitrary $\zeta^{[r]}$. As in HS, by making $\zeta^{[r]}$ successively equal to the columns of $I_{nr}$ it follows that

$$Q_{nr} P_{nr} = P_{nr} Q_{nr} = I_{nr},$$ 

and substituting (21) into (20) yields $\zeta^{(r)} = P_{nr} Q_{nr} \zeta^{(r)}$, but this does not imply $P_{nr} Q_{nr}$ to be the identity matrix because $\zeta^{(r)}$ has a fixed structure and thus is not arbitrary. The non-uniqueness problem of $Q_{nr}$ is solved noting that it is a left inverse of $P_{nr}$, and thus choosing the Moore-Penrose generalized inverse, so that

$$Q_{nr} = (P_{nr} P_{nr})^{-1} P_{nr}.$$
The relation between $\zeta^{(r)}$ and $\zeta^{(r)}$ has been given the same structure as that between $\text{vech}$ and $\text{vec}$, and we can now proceed to define the generalized efficiency tests. We consider conditional moment restrictions of the form

$$m_e(\theta, s; \theta) = T^{-1} \sum_{t=1}^{T} m_{et}(\varphi_t, s; \theta) = T^{-1} \sum_{t=1}^{T} m_{et}(\varphi_t, s; \theta) = T^{-1} \sum_{t=1}^{T} \varphi_t \{ u_t^{[s]} - q_t[s] \},$$

for the $s$-th moment, where $m_{et}(\varphi_t, s; \theta) = \varphi_t \{ u_t^{[s]} - q_t[s] \}, q_t[s] = E[u_t^{[s]} | \mathcal{F}_t], \varphi_t$ is an $r \times n_s$ matrix of measurable functions of $\mathcal{F}_t$, and $\theta = (\varphi_1, ..., \varphi_T)'$. In contrast to the restrictions in the univariate case, we are not explicitly considering the variance residuals in $m_e(\theta, s; \theta)$. This is done for simplicity in the presentation, and the statistic may be constructed using a mix of mean and variance residuals. Just observe that using $\varepsilon_t = u_t^{[2]} - h_t$ together with (21) and (2) we may rewrite for $s \geq 2$,

$$u_t^{[s]} = Q_{ns} \{ u_t^{(s-2)} \otimes P'(\varepsilon_t + h_t) \},$$

and this may be used to introduce variance residuals without affecting the asymptotic distribution of $m_e(\theta, s; \theta)$. The problem of evaluating the conditional expectations $q_t[s]$ is solved with the following

**Lemma 9.15.** If $x \sim N[0, \Sigma], x \in \mathbb{R}^n$, then for $s$ even

$$q(s) = E[x^{(s)}] = R_s \text{vec}[\sigma^2 q(s - 2)'], \quad s \geq 2,$$

where $\sigma^2 = \text{vec} \Sigma, R_s = \text{I}_{(n+1), n} + \text{I}_n \otimes R_{s-1}$ so that it is $n^s \times n^s$, and the initial conditions are given by $q(0) = 1$, and $R_0 = 0_{n \times n}$.

The lemma provides a recursive formula to obtain all even order moments of the multivariate normal distribution which generalizes the well-known expression for the univariate case $q(s) = (s - 1)q(s - 2)\sigma^2$ (e.g. Engle [1982a]). Observe that $R_s = s - 1$ when $n = 1$. Thus we have that for $s$ even,

$$q_t(s) = E[u_t^{(s)} | \mathcal{F}_t] = R_s \{ q_t(s - 2) \otimes \text{vec} H_t \}, \quad s \geq 2,$$

and since $u_t^{[s]} = Q_{ns} u_t^{(s)}$ then
\[ q_t [s] = Q_{ns} q_t (s), \] - (23)

that is, \( q_t [s] \) selects the corresponding elements of \( q_t (s) \).

It is evident that \( E \left[ m_e (\theta, s; \theta) \right] = 0 \), and the covariance matrix of \( m_e (\theta, s; \theta) \) is given by

\[
V_{eT} = \text{Var} \left[ T^{1/2} m_e (\theta, s; \theta) \right] = E \left[ T^{-1} \sum_{t=1}^{T} \Phi_t H_t^{[2s]} \Phi_t' \right],
\]

where

\[
H_t^{[2s]} = \text{Var} \left[ u_{t[s]} - q_t [s] \mid \mathcal{F}_t \right] = E \left[ u_{t[s]} u_{t[s]}' | \mathcal{F}_t \right] - q_t [s] q_t [s]'.
\]

can be obtained from \( q_t [2s] \) and \( q_t [s] \) using (22) and (23).

The efficiency test-statistics are based on the corresponding sampling moments \( m_e (\hat{\theta}, s; \theta) \), and to obtain their distribution we need

\[
M_e (\hat{\theta}, s; \theta_0) = - \mathcal{C} \left\{ T^{-1} \sum_{t=1}^{T} m_{et} (\Phi_t, s; \theta_0) d_{et} (\theta_0)' \right\}.
\]

Using (8) and iterated expectations we get

\[
M_e (\hat{\theta}, s; \theta_0) = \begin{cases} 
- \mathcal{C} \left\{ T^{-1} \varphi' H_1 (s) X \right\} & \text{for } s \text{ odd}, \\
- \mathcal{C} \left\{ T^{-1} \varphi' H_2 (s) S \right\} & \text{for } s \text{ even},
\end{cases}
\]

where

\[
H_1 (s) = \text{diag} \left\{ E \left[ u_{t[s]} u_{t[s]}' | \mathcal{F}_t \right] H_t^{-1} \right\}, \quad s \text{ odd},
\]

and

\[
H_2 (s) = \frac{1}{2} \text{diag} \left\{ E \left[ u_{t[s]} u_{t[2s]}^T | \mathcal{F}_t \right] P \left( H_t^{-1} \otimes H_t^{-1} \right) P' \right\}, \quad s \text{ even},
\]

and the expectations can be evaluated using \( q_t [s + 1] \) and \( q_t [s + 2] \).

Consider the multivariate version of (C0'') from § 5.3,

(C0'''-M) \hspace{1cm} \text{The conditional distribution of } y_t \text{ is } f (y_t | \mathcal{F}_t, \theta_0, \lambda_T),

where \( \lambda_T = \lambda_0 + T^{-1/2} \delta \), for fixed \( \lambda_0 \) and \( \delta \), and at \( \delta = 0 \), \( f (\cdot, 1, \cdot) \) is the normal pdf with mean \( \mu_t (\beta_0) \) and covariance matrix \( H_t (\theta_0) \). We can now produce
Theorem 9.16. - Under (Q0''-M) - (Q8-M) and the sequence \( \{ \varphi_t \} \), \( \varphi_t \in \mathcal{F}_t \), being such that \( g^* = (m_0 (\hat{\theta}, s ; \theta_0)' , d_{\theta}' \) obeys the regularity, continuity, dominance and mixing conditions in assumptions (1) - (6) of Newey [1985b],

\[
T m_0 (\hat{\theta}, s ; \hat{\theta}') Q_\varphi^{-1} m_0 (\hat{\theta}, s ; \hat{\theta}) \overset{d}{\to} \chi^2 [r ; \psi_e' Q^{-1}_\varphi \psi_e],
\]

where

\[
\psi_e = \mathcal{C} \left\{ T^{-1} \sum_{t=1}^{T} m_{st} (\varphi_t , s ; \theta_0) d_{\theta} (\theta_0)' \right\} - M_e (\hat{\theta}, s ; \theta_0) V (\hat{\theta}) \mathcal{C} \left\{ T^{-1} \sum_{t=1}^{T} d_{\theta t} (\theta_0) d_{\theta} (\theta_0)' \right\} \delta,
\]

\[
Q_\varphi = V_e (\hat{\theta}, s ; \theta_0) - M_e (\hat{\theta}, s ; \theta_0) V (\hat{\theta}) M_e (\hat{\theta}, s ; \theta_0)', \quad -(24)
\]

and \( V_e (\hat{\theta}, s ; \theta_0) = \lim V_{eT} (\hat{\theta}, s ; \theta_0) \). A consistent estimator \( \hat{Q}_\varphi \) of \( Q_\varphi \) is constructed replacing the expectations in (24) with sample moments evaluated at \( \hat{\theta} \).

This Theorem provides a wide range of tests of higher order moments. A symmetry test is obtained for \( s = 3 \) and a kurtosis test for \( s = 4 \). Note that we can also concentrate on the moments of a reduced number of elements of \( y_t \) by appropriate choice of the \( \varphi_t \). All these are single moment tests, but using Lemma 9.15 and the results of Newey [1985b] we can easily obtain the covariance between basic efficiency statistics and construct omnibus tests involving several moments. Thus for example, an omnibus test based on the third and fourth moments would be a generalization of the Jarque-Bera [1980] test for normality to a multivariate heteroskedastic setting.
APPENDIX TO CHAPTER 9

The following Lemma is used in other proofs:

**Lemma A1.** For any squared symmetric matrix $A$ of order $n$, and any vector $y$ of order $m$,

(i) $\text{vec} (y \otimes A) = I_{(mn, n)} (y \otimes \text{vec} A)$, and

(ii) $\text{vec} (A \otimes A) = (I_n \otimes I_{(n^2, n)})(\text{vec} A) \otimes (\text{vec} A)$.

**Proof:** $\text{vec} (y \otimes A) = I_{(mn, n)} \text{vec} (y \otimes A)' = I_{(mn, n)} \text{vec} (y' \otimes A)$,

because $y \otimes A$ is $mn \times n$ and $A$ is symmetric. But

$\text{vec} (y' \otimes A) = \text{vec} (y_1 A, ..., y_m A) = (y_1 (\text{vec} A)', ..., y_m (\text{vec} A)')' = y \otimes \text{vec} A$,

and substituting back establishes (i). Let $A_j$ be the $j$-th column of $A$. Then

$\text{vec} (A \otimes A) = \text{vec} (A_{1} \otimes A, ..., A_{n} \otimes A) = ([\text{vec}(A_{1} \otimes A)', ..., [\text{vec}(A_{n} \otimes A)'])'$,

and

$(\text{vec} A) \otimes (\text{vec} A) = ((A_{1} \otimes \text{vec} A)', ..., (A_{n} \otimes \text{vec} A)')'$.

But using (i), $\text{vec} (A_j \otimes A) = I_{(n^2, n)} (A_j \otimes \text{vec} A)$, and therefore

$\text{vec} (A \otimes A) = (I_{(n^2, n)} (A_{1} \otimes \text{vec} A)', ..., [I_{(n^2, n)} (A_{n} \otimes \text{vec} A)']')$

$= (I_n \otimes I_{(n^2, n)})(((A_{1} \otimes \text{vec} A)', ..., (A_{n} \otimes \text{vec} A)')'$

$= (I_n \otimes I_{(n^2, n)})(\text{vec} A) \otimes (\text{vec} A)$,

which establishes (ii).

**Proof of Lemma 9.1:** From (1) and (6)

$E \left[ v_t v_t' | \mathcal{F}_t \right] = 2 Q (H_t \otimes H_t) Q' + h_t h_t'$,
and therefore, using $\text{vec } ABC = (C' \otimes A) \text{vec } B$, we get

$$\text{vec } E [v_t v'_t \mid \mathcal{F}_t] = 2 (Q \otimes Q) \text{vec } (H_t \otimes H_t) + \text{vec } h_t h_t'.$$

But using Lemma A1 we obtain

$$(Q \otimes Q) \text{vec } (H_t \otimes H_t) = (Q \otimes Q) (I_n \otimes I(n^2, n)) \text{vec } (H_t \otimes H_t)$$

$$= (Q \otimes Q) (I_n \otimes I(n^2, n)) (P' \text{vech } H_t) \otimes (P' \text{vech } H_t)$$

$$= (Q \otimes Q) (I_n \otimes I(n^2, n)) (P' \otimes P') (h_t \otimes h_t) = R \text{vec } h_t h_t',$$

where $R = (Q \otimes Q) (I_n \otimes I(n^2, n)) (P' \otimes P')$, and we have used $\text{vec } x y' = y \otimes x$, $\text{vec } H_t = P' \text{vech } H_t$, and $h_t = \text{vech } H_t$. The Lemma follows after substituting the last expression in $\text{vec } E [v_t v'_t \mid \mathcal{F}_t]$.

**Proof of Lemma 9.3**: We have that $\tilde{u}_t = u_t - (\tilde{\mu}_t - \mu_t)$, and using the Mean Value Theorem for random functions (Jennrich [1969]) we can write

$$\tilde{\mu}_t - \mu_t = X_t (\tilde{\beta} - \beta) + O_p (T^{-1})$$,

so

$$T^{-1/2} \sum_{t=1}^T F_t \text{vec } (\tilde{u}_t \tilde{u}_t' - u_t u_t') = T^{-1/2} \sum_{t=1}^T F_t \text{vec } X_t (\tilde{\beta} - \beta_0)(\tilde{\beta} - \beta_0)' X_t'$$

$$- T^{-1/2} \sum_{t=1}^T F_t \text{vec } [u_t (\tilde{\beta} - \beta_0)' X_t' + X_t (\tilde{\beta} - \beta_0) u_t'] + O_p (T^{-1/2})$$.

Now

$$T^{-1/2} \sum_{t=1}^T F_t \text{vec } X_t (\tilde{\beta} - \beta_0)(\tilde{\beta} - \beta_0)' X_t'$$

$$= [T^{-1} \sum_{t=1}^T F_t (X_t \otimes X_t)] T^{1/2} \text{vec } (\tilde{\beta} - \beta_0)(\tilde{\beta} - \beta_0)' = o_p (1)$$

in view of the consistency of $\tilde{\beta}$. Similarly,

$$T^{-1/2} \sum_{t=1}^T F_t \text{vec } u_t (\tilde{\beta} - \beta_0)' X_t' = [T^{-1/2} \sum_{t=1}^T F_t (X_t \otimes u_t)] (\tilde{\beta} - \beta_0) = o_p (1),$$

because the first factor is stochastically bounded. Therefore,
When \( \tilde{F}_t \) is used in place of \( F_t \) the result still holds as in Lemma 3.3.

Proof of Theorem 9.4:

\[
T^{-1/2} \sum_{t=1}^{T} F_t \text{vec} \left( \hat{u}_t \hat{u}_t' - u_t u_t' \right) \xrightarrow{as} 0.
\]

\[
\text{Proof of Theorem 9.4:}
\]

\[
d_{m\alpha} (\theta) = 0 ,
\]

\[
d_{m\beta} (\theta) = T^{-1} \sum_{t=1}^{T} X_t' \tilde{H}_t^{-1} u_t ,
\]

and

\[
d_{v\theta} (\theta) = \frac{1}{2} T^{-1} \sum_{t=1}^{T} S_t' P \left( H_t^{-1} \otimes H_t^{-1} \right) P' (\tilde{v}_t - h_t) ,
\]

using the differentiation rules in KE, and also

\[
J_m (\theta_0) = \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} X_t' \tilde{H}_t^{-1} X_t \right] .
\]

and

\[
J_v (\theta_0) = \frac{1}{2} \mathbb{E} \left[ T^{-1} \sum_{t=1}^{T} S_t' P \left( H_t^{-1} \otimes H_t^{-1} \right) P' S_t \right] + o_p (T^{-1/2}).
\]

Now since \( d^*_\theta (\theta) = d_{m\theta} (\theta) + d_{v\theta} (\theta) \) and \( J^* (\theta) = J_m (\theta) + J_v (\theta) \), using (8), (9) and (13) it follows that

\[
T^{1/2} \left[ d_\theta (\theta_0) - d^*_\theta (\theta_0) \right] \xrightarrow{as} 0
\]

and

\[
J(\theta_0) - J^* (\theta_0) \xrightarrow{as} 0 ,
\]

which establishes the Theorem.

Proof of Theorem 9.5: From the orthogonality conditions \( d_\theta (\hat{\theta}) = 0 \) in (8) apply Theorems 2.1 and 3.1 of Hansen [1982] of Theorems 3.1 and 3.2 of White and Domowitz [1984].
Proof of Corollary 9.6: From Theorem 9.4, \( \hat{\theta} \) may be obtained from \( \lambda^*(\theta) \), and the result follows because this log-likelihood conforms to the hypotheses of Theorem 3.12.

Proof of Theorem 9.7: We have that
\[
\frac{\partial}{\partial \theta'} \left[ X_t' H_t^{-1} u_t \right] = \frac{\partial}{\partial \theta'} \left[ \text{vec}(X_t' H_t^{-1}) \right] = (u_t' \otimes I_k) - X_t' H_t^{-1} X_t,
\]
and
\[
\frac{\partial}{\partial \theta'} \left[ S_t' K_t^{-1} \varepsilon_t \right] = \frac{\partial}{\partial \theta'} \left[ \text{vec}(S_t' K_t^{-1}) \right] = (\varepsilon_t' \otimes I_p) - S_t' K_t^{-1} S_t,
\]
so that using iterated expectations we obtain
\[
E \left[ \frac{\partial \psi(\theta_0)}{\partial \theta} \right] = -T^{-1}(X' \Omega^{-1} X, S' K^{-1} S).
\]
Strong consistency and asymptotic normality follow from Theorem 2.1.

The asymptotic covariance matrix is obtained by simple algebra using the weighting matrix and \( E[\partial \psi(\theta_0)/\partial \theta'] \), and is given by
\[
V(\hat{\theta}_J) = \mathcal{C} \left( T^{-1} \bar{X}' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \bar{X} + T^{-1} S' K^{-1} S (S' K^{-1} S)^{-1} S' K^{-1} S \right)^{-1},
\]
but because \( \bar{X} = (X, 0) \), it is easily seen that the first term in the expectation is simply \( T^{-1} \bar{X}' \Omega^{-1} \bar{X} \), so that \( V(\hat{\theta}_J) = \mathcal{C} \left( T^{-1} G' \Sigma^{-1} G \right) \).

Proof of Lemma 9.9: Use iterated expectations in the scores (8) and apply Lemma 4.1.

Proof of Lemma 9.10: When the first two conditional moments are correct \( E[u_t | \mathcal{F}_t] = 0 \) and \( E[\varepsilon_t | \mathcal{F}_t] = 0 \), and therefore \( \varphi_\theta(\theta_0) = 0 \).
Proof of Theorem 9.11: The proof is identical to that of Theorem 5.1 for an identified variance equation (Corollary 5.2), substituting the multivariate MVT expansions.

Proof of Theorem 9.12: Substitute the multivariate expressions in Theorem 5.3.

Proof of Theorem 9.13: (18) follows immediately constructing the quadratic form in Theorem 9.12 and substituting the consistent estimate of $Q_\theta$. Because $\hat{\Gamma}'\hat{\Sigma}_{1/2}\hat{\nu} = 0$, $\hat{\phi}'\hat{\nu} = \hat{\phi}'\hat{\Sigma}_{1/2}\hat{\gamma}$, and therefore substituting $m(\hat{\phi}, \hat{\theta})$ and $\hat{Q}_\theta$ in (18) $s$ can be rewritten as

$$ s = \hat{\nu}'\hat{\Sigma}_{1/2}\hat{\gamma}\hat{\Sigma}_{1/2}\hat{\phi}(\hat{\phi}'\hat{\Sigma}_{1/2}\hat{\gamma}\hat{\Sigma}_{1/2}\hat{\phi})^{-1}\hat{\phi}'\hat{\Sigma}_{1/2}\hat{\gamma}\hat{\Sigma}_{1/2}\hat{\nu}, $$

which is the ESS of the regression of $\hat{\nu}$ on $\hat{G}$ and $\hat{\Sigma}$ in the metric of $\hat{\Sigma}$. The total SS of this regression is $\hat{\nu}'\hat{\Sigma}_{1/2}\hat{\nu} = \hat{w}'\hat{\Omega}_{1/2}\hat{w} + \hat{e}'\hat{K}_{1/2}\hat{e}$, and $E[\hat{w}'\hat{\Omega}_{1/2}\hat{w} | F_t] = n$, while $E[\hat{e}'\hat{K}_{1/2}\hat{e} | F_t] = \frac{1}{2}n(n+1)$, so that $(T\cdot n)^{-1}\hat{\nu}'\hat{\Sigma}_{1/2}\hat{\nu} \overset{\text{as}}{\longrightarrow} 1$, and the Theorem follows.


Proof of Lemma 9.15: Let $\zeta (\zeta) = E[\exp(i\zeta'x)] = \exp(-\frac{1}{2}\zeta'\Sigma \zeta)$ denote the characteristic function of $x$. Let $\zeta_0 (\zeta) = \zeta (\zeta)$ and define, for $s > 0$,

$$ \zeta_s (\zeta) = \text{vec}\frac{\partial \zeta_{s-1} (\zeta)}{\partial \zeta'}, $$

which is $n^s \times 1$. - (A1)

First we use induction to prove that, for $s \geq 2$,

$$ \zeta_s (\zeta) = -[\zeta_{s-1} (\zeta) \otimes \Sigma \zeta] - R_{n} [\zeta_{s-2} (\zeta) \otimes \text{vec} \Sigma], $$

where $R_{n} = I_{n+1,n} + I_{n} \otimes R_{n-1}$ so that it is $n^s \times n^s$, and $R_0 = 0_{n \times n}$. - (A2)
We have that $\mathcal{C}_1(\zeta) = \frac{\partial \mathcal{C}_0(\zeta)}{\partial \zeta} = - \mathcal{C}_0(\zeta) \Sigma \zeta$, and therefore

$$\mathcal{C}_2(\zeta) = \text{vec} \frac{\partial \mathcal{C}_1(\zeta)}{\partial \zeta'} = \text{vec} \left\{ - \zeta \frac{\partial \mathcal{C}_0(\zeta)}{\partial \zeta'} - \mathcal{C}_0(\zeta) \Sigma \right\}$$

$$= - (I_n \otimes \Sigma \zeta) \mathcal{C}_1(\zeta) - \mathcal{C}_0(\zeta) \text{vec} \Sigma$$

$$= - \left[ \mathcal{C}_1(\zeta) \otimes \Sigma \zeta \right] - \mathcal{C}_0(\zeta) \text{vec} \Sigma,$$

where we have used (A1) and well-known properties of Kronecker products. Observe that $\mathcal{C}_s(\zeta)$ is always a vector, which has been used in the last equality. (A2) holds for $s = 2$ because $\mathcal{C}_0(\zeta)$ is scalar, $R_2 = I(n, n)$, and $I(n, n) \text{vec} \Sigma = \text{vec} \Sigma$ in view of the symmetry of $\Sigma$.

Now suppose that

$$\mathcal{C}_{s-1}(\zeta) = - \left[ \mathcal{C}_{s-2}(\zeta) \otimes \Sigma \zeta \right] - R_{s-1} \left[ \mathcal{C}_{s-3}(\zeta) \otimes \text{vec} \Sigma \right].$$

Taking derivatives we get

$$\mathcal{C}_s(\zeta) = \text{vec} \frac{\partial \mathcal{C}_{s-1}(\zeta)}{\partial \zeta'} = \text{vec} \left\{ - \left[ \mathcal{C}_{s-2}(\zeta) \otimes \Sigma \zeta \right] - R_{s-1} \left[ \mathcal{C}_{s-3}(\zeta) \otimes \text{vec} \Sigma \right] \right\}$$

$$= - \text{vec} \left[ \mathcal{C}_{s-2}(\zeta) \otimes \Sigma \zeta \right] - (I_n \otimes R_{s-1}) \text{vec} \left[ \mathcal{C}_{s-3}(\zeta) \otimes \text{vec} \Sigma \right].$$

Because $\mathcal{C}_{s-2}(\zeta) \otimes \Sigma \zeta = [I_{n^{s-2}} \otimes \Sigma \zeta] \mathcal{C}_{s-2}(\zeta) = [\mathcal{C}_{s-2}(\zeta) \otimes \Sigma] \zeta$, it follows that

$$\text{vec} \frac{\partial \left[ \mathcal{C}_{s-2}(\zeta) \otimes \Sigma \zeta \right]}{\partial \zeta'} = \text{vec} \left\{ [I_{n^{s-2}} \otimes \Sigma \zeta] \frac{\partial \mathcal{C}_{s-2}(\zeta)}{\partial \zeta'} + \left[ \mathcal{C}_{s-2}(\zeta) \otimes \Sigma \right] \right\}.$$

$$= \left[ \mathcal{C}_{s-1}(\zeta) \otimes \Sigma \zeta \right] + I_{(n^{s-1}, n)} (\mathcal{C}_{s-2}(\zeta) \otimes \text{vec} \Sigma).$$

The last equality is obtained using

$$\text{vec} \left\{ [I_{n^{s-2}} \otimes \Sigma \zeta] \frac{\partial \mathcal{C}_{s-2}(\zeta)}{\partial \zeta'} \right\} = (I_n \otimes I_{n^{s-2}} \otimes \Sigma \zeta) \text{vec} \frac{\partial \mathcal{C}_{s-2}(\zeta)}{\partial \zeta'}$$

$$= (I_{n^{s-1}} \otimes \Sigma \zeta) \mathcal{C}_{s-1}(\zeta) = \mathcal{C}_{s-1}(\zeta) \otimes \Sigma \zeta,$$

and

$$\text{vec} (\mathcal{C}_{s-2}(\zeta) \otimes \Sigma) = I_{(n^{s-1}, n)} (\mathcal{C}_{s-2}(\zeta) \otimes \text{vec} \Sigma),$$
which follows from Lemma A1. Also, \( \mathcal{G}_{s-3} (\zeta) \otimes \text{vec } \Sigma = [I_{n_{s-3}} \otimes \text{vec } \Sigma ] \mathcal{G}_{s-3} (\zeta) \), and so

\[
\text{vec } \frac{\partial [\mathcal{G}_{s-3} (\zeta) \otimes \text{vec } \Sigma]}{\partial \zeta'} = \text{vec } [I_{n_{s-3}} \otimes \text{vec } \Sigma ] \frac{\partial \mathcal{G}_{s-3} (\zeta)}{\partial \zeta'} \\
= [I_{n_{s-2}} \otimes \text{vec } \Sigma ] \mathcal{G}_{s-2} (\zeta) = \mathcal{G}_{s-2} (\zeta) \otimes \text{vec } \Sigma .
\]

- (A5)

Substituting (A4) and (A5) in (A3) establishes (A2).

Now, by the properties of the characteristic function we have that

\[
q (s) = E [ x^{(s)} ] = \frac{1}{i^s} \mathcal{G}_{s} (0) = - \frac{1}{i^s} R_s [ \mathcal{G}_{s-2} (0) \otimes \text{vec } \Sigma ] ,
\]

using (A2), and because \( i^2 = -1 \) and the alternating sign of \( \mathcal{G}_{s} (0) \), for \( s \) even we take absolute values to obtain

\[
q (s) = R_s [ q (s - 2) \otimes \text{vec } \Sigma ] = R_s \text{vec } [ \sigma^2 q (s - 2)' ] ,
\]

where \( \sigma^2 = \text{vec } \Sigma \).

\[ \square \]

**Proof of Theorem 9.16:** Substitute the multivariate expressions in Theorem 5.10. \[ \square \]
AN EXPLORATION INTO HIGHER ORDER MOMENTS

In the previous chapters we have concentrated on the estimation of the first two conditional moments, and many of the results depend crucially on the symmetry of the distribution. In Chapter 3, in particular, we showed that the independence (asymptotically) of the information contributed by the mean and the variance equations is a direct consequence of symmetry, and so is the matrix weighted average interpretation of the joint estimators. This Chapter is devoted to explore some ideas for treating specification and misspecification of higher order moments, a topic which has not receive much attention in the econometric literature. We proceed naturally from the ideas developed in Chapters 3 to 6 by trying to extract information from the higher order moments using sets of orthogonality conditions. This constitutes an alternative for the treatment of complex distributional problems to the approximation of the true DGP by Hermite polynomial expansions as in Gallant and Nychka [1987] and Gallant and Tauchen [1986].

We showed in Chapter 3 that the QMLE \( \hat{\theta} \) remains consistent in the presence of third or higher order moment misspecification, and even the asymptotic covariance matrix of \( \hat{\theta} \) remains consistent if the specification error occurs in moments of order five or more. Thus our main interest lies in the possibility of increasing efficiency in estimation by extracting information from higher order moments. A related development has been put forward by Newey [1986], who proposed using the information in the high order odd moments in symmetric distributions.
To motivate the Chapter, in section § 10.1 we consider the estimation of a simple heteroskedastic model in the presence of asymmetry. In section § 10.2 we move to parametric estimation of higher order moments, and we consider some diagnostic tests in section § 10.3. In writing this chapter our main interest is to show that the theoretical structure developed in the main body of the Thesis is powerful enough to cover a wide range of situations. Moreover, we make the point that a non-linear regression package with matrix manipulations is sufficient for the computations required.

§ 10.1 Estimating the first two moments in the presence of asymmetry

To motivate our treatment of asymmetry, let us consider a set of two linear correlated equations with cross-equation restrictions, that is,

\[ y_1 = X_{11} \beta + X_{12} \gamma_{12} + u_1 , \quad (T_1 \times 1) \] -(1)

and

\[ y_2 = X_{21} \beta + X_{22} \gamma_{22} + u_2 , \quad (T_2 \times 1) \] -(2)

or jointly,

\[ y = X \gamma + u , \quad -(3) \]

where \( X = \begin{pmatrix} X_{11} & X_{12} & 0 \\ X_{21} & 0 & X_{22} \end{pmatrix} \), and \( \gamma = (\beta', \gamma_{12}', \gamma_{22}')' \). The covariance matrix is \( \Omega = \| \Omega_{ij} \| = \| \mathbb{E} [ u_i u_j'] \| \), which we take as known for simplicity. The GLS estimator is

\[ \hat{\gamma} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y , \]

but \( \hat{\beta} \) cannot be obtained as a matrix weighted average of the estimators from the two sources unless \( \Omega_{12} = 0 \). If we transform (3) by a nonsingular matrix \( A \), say, and apply GLS to the resulting system we get
\[ \hat{\gamma}_A = [X' A' (A \Omega A')^{-1} A X]^{-1} X' A' (A \Omega A')^{-1} Ay = \gamma, \]
because \(|A| \neq 0\). Consider \(A = \begin{pmatrix} I_{T_1} \\ -\Omega_{21} \Omega_{11}^{-1} \\ I_{T_2} \end{pmatrix} \), which is nonsingular, preserves (1) and transforms (2) into
\[
y_2 - \Omega_{21} \Omega_{11}^{-1} y_1 = (X_{21} - \Omega_{21} \Omega_{11}^{-1} X_{11}) \beta + X_{22} \gamma_{22} - \Omega_{21} \Omega_{11}^{-1} X_{12} \gamma_{12} + u_2 - \Omega_{21} \Omega_{11}^{-1} u_1,
\]
or more compactly,
\[
y_{2-1} = X_{21} \beta + X_2 \gamma_2 + u_{2-1}, \quad - (4)
\]
where \(y_{2-1} = y_2 - \Omega_{21} \Omega_{11}^{-1} y_1, X_{21} = X_{21} - \Omega_{21} \Omega_{11}^{-1} X_{11}, u_{2-1} = u_2 - \Omega_{21} \Omega_{11}^{-1} u_1, X_2 = (-\Omega_{21} \Omega_{11}^{-1} X_{12}, X_{22}), \) and \(\gamma_2 = (\gamma_{12}', \gamma_{22}')\). The GLS estimates for \(\gamma\) in (3) are identical to those from (1) and (4). But \(E[u_1 u_{2-1}'] = E[u_1 (u_2 - \Omega_{21} \Omega_{11}^{-1} u_1)'] = 0\), so that the two equations are now uncorrelated. Hence \(\hat{\beta}\) may be obtained as the MWA of the separate GLS estimators from these two equations, that is,
\[ \hat{\beta} = V(\hat{\beta}) [V(\hat{\beta}_1)^{-1} \hat{\beta}_1 + V(\hat{\beta}_2)^{-1} \hat{\beta}_2], \]
where
\[ \hat{\beta}_1 = (X_{11}' Q_1 X_{11})^{-1} X_{11}' Q_1 y_1, \]
\[ \hat{\beta}_2 = (X_{21-1}' Q_{2-1} X_{21-1})^{-1} X_{21-1}' Q_{2-1} y_{2-1}, \]
\[ V(\hat{\beta}_1) = (X_{11}' Q_1 X_{11})^{-1}, \quad V(\hat{\beta}_2) = (X_{21-1}' Q_{2-1} X_{21-1})^{-1}, \]
\[ V(\hat{\beta})^{-1} = V(\hat{\beta}_1)^{-1} + V(\hat{\beta}_2)^{-1}, \]
\[ Q_1 = \Omega_{11}^{-1} - \Omega_{11}^{-1} X_{12} (X_{12}' \Omega_{11}^{-1} X_{12})^{-1} X_{12}' \Omega_{11}^{-1}, \]
\[ Q_2 = \Omega_{22-1} - \Omega_{22-1}^{-1} X_2 (X_2' \Omega_{22-1}^{-1} X_2)^{-1} X_2' \Omega_{22-1}^{-1}, \]
and
\[ \Omega_{22-1} = \Omega_{22} - \Omega_{21} \Omega_{11}^{-1} \Omega_{12}. \]
The effect of the transformation has been to remove from the second equation the information already contained in the first one. For this reason we will say
that (4) is the *conditioned version* of (2). We are assuming that \((X_{21.1}, X_2)\) has full rank so that the conditioned equation still retains enough information to produce a full estimator of \(\gamma\). Depending on the relation between the two equations, identification problems may arise that permit only the estimation of a function of \(\gamma\) of lower dimension in the conditioned equation. In such case the estimates of the jointly identifiable functions are MWA's. The extreme situation is that \(X_{21.1} = 0\), when no information about \(\beta\) can be extracted from the conditioned equation and \(\hat{\beta} = \hat{\beta}_1\).

Ignoring the fact that \(\Omega_{12} \neq 0\) and estimating \(\beta\) as the MWA of the GLS estimators does not affect unbiasedness, but this estimator is inefficient and its covariance matrix is not correct. The proper covariance matrix is the relevant submatrix of

\[
(X' \Omega_D^1 X)^{-1} X' \Omega_D^1 \Omega \Omega_D^1 X (X' \Omega_D^1 X)^{-1},
\]

where \(\Omega_D = \text{diag} \{\Omega_{11}, \Omega_{22}\}\), see White [1980b]. It follows that ignoring asymmetry when estimating a heteroskedastic model produces inefficiency and incorrect estimation of the covariance matrix of estimators in the mean and variance two-equation system because \(\lambda_t \neq 0\) and \(\Sigma_t\) is not diagonal. But given the fact that the \(\nu_t\) are martingale differences, the structure of the submatrices \(\Omega_{ij}\) is diagonal, and thus the *conditioned variance equation* (CVE) is obtained by contemporaneous transformations only, as

\[
\frac{u_t^2 - \lambda_t}{h_t} y_t = h_t(\theta) - \lambda_t \mu_t(\beta) + \epsilon_t^*,
\]

where \(\epsilon_t^* = \epsilon_t - \lambda_t \frac{u_t}{h_t}\). \(E[\epsilon_t^* | \mathcal{F}_t] = 0\) and \(E[u_t \epsilon_t^* | \mathcal{F}_t] = 0\), while

\[
\text{Var}[\epsilon_t^* | \mathcal{F}_t] = \kappa_t - \frac{\lambda_t^2}{h_t}.
\]

Suppose we have parametric estimates of \(\lambda_t / h_t\) and \(\kappa_t - \lambda_t^2 / h_t\) from root-T consistent estimators of their underlying parameters. By an argument identical to that given for \(\hat{\phi}_\nu\) in § 3.2.2 the estimators of the estimable functions...
of $\theta$ in the CVE have the same asymptotic distribution using these parametric estimates as if such functions were known.

We now redefine $\phi$ as the identifiable functions of $\theta$ in the CVE, rather than in the variance equation as in the symmetric case, and take $\lambda_t/h_t$ and $\kappa_t - \lambda_t^2/h_t$ as given. The estimator $\hat{\phi}_v$ is defined by the orthogonality conditions

$$
\psi_v(\phi) = T^{-1} \sum_{t=1}^{T} \left[ \tilde{\kappa}_t - \tilde{h}_t^{-1} \tilde{\lambda}_t^2 \right]^{-1} \left[ \tilde{s}_{\phi t} - \tilde{h}_t^{-1} \tilde{\lambda}_t x_{\phi t} \right] e_t^* 
$$

$$
= T^{-1} \left( S_\phi - \delta \Omega^{-1} X_\phi \right)' (K - \delta \Omega^{-1} \delta)^{-1} e^* ,
$$

where $e_t^* = e_t^* + (u_t^2 - u_t^2)$, $s_{\phi t} = \partial h_t / \partial \phi$, and $x_{\phi t} = \partial \mu_t / \partial \phi$. By Lemma 3.3 the term in $(u_t^2 - u_t^2)$ vanishes asymptotically, and Theorem 2.1 establishes that the GMM/GLS estimator $\hat{\phi}_v$ of $\phi$ from $\psi_v(\phi)$ is strongly consistent and has asymptotic distribution

$$
T^{1/2} (\hat{\phi}_v - \phi_0) \overset{d}{\rightarrow} N [ 0, \Sigma \{ T^{-1} \left( S_\phi - \delta \Omega^{-1} X_\phi \right)' (K - \delta \Omega^{-1} \delta)^{-1} \left( S_\phi - \delta \Omega^{-1} X_\phi \right) \}^{-1} ] ,
$$

where $S_\phi = (s_{\phi 1}, \ldots, s_{\phi T})'$, $X_\phi = (x_{\phi 1}, \ldots, x_{\phi T})'$, and the expectation for $V(\hat{\phi}_v)$ is evaluated at $\phi_0$. If we consider now using $\psi_v(\phi)$ and $\psi_m(\beta) = T^{-1} X' \Omega^{-1} u$ jointly to estimate $\theta$, we have

**Theorem 10.1.** - Under (A0) - (A7), the GMM estimator $\hat{\theta}_J$ obtained from the orthogonality conditions $\psi(\theta) = (\psi_m(\beta)' , \psi_v(\phi)' )'$ and weighting matrix $A_T = \text{diag} \left( T (X' \Omega^{-1} X)^{-1} , T (S_\phi - \delta \Omega^{-1} X_\phi)' (K - \delta \Omega^{-1} \delta)^{-1} (S_\phi - \delta \Omega^{-1} X_\phi) \right)^{-1}$ is strongly consistent for $\theta_0$ and has asymptotic distribution

$$
T^{1/2} (\hat{\theta}_J - \theta_0) \overset{d}{\rightarrow} N [ 0, \Sigma \{ T^{-1} G' \Sigma^{-1} G \}^{-1} ] ,
$$

where the expectation for $V(\hat{\theta}_J)$ is evaluated at $\theta_0$ and $\Sigma = \left( \begin{array}{cc} \Omega & \delta \\ \delta & K \end{array} \right)$.

**Proof:** Apply Theorem 3.6 to the orthogonality conditions $\psi(\theta)$ and weighting matrix $A_T$ and observe that this estimator is asymptotically equivalent to that obtained directly from the mean and variance equations. □
If $\phi = (\gamma', \phi_1')'$, where $\gamma = \gamma(\beta)$, and $\phi_1$ does not depend on $\beta$, the vector $\gamma$ represents the jointly identifiable functions in the mean and conditioned variance equations, and its joint estimator obeys

**Corollary 10.2.** - Under the assumptions of Theorem 10.1,

$$\hat{\gamma}_J - \gamma_0 = (I_{k^*} - \Pi_u)(\hat{\gamma}_m - \gamma_0) + \Pi_u(\hat{\gamma}_v - \gamma_0) + o_p(T^{-1/2}),$$

and

$$V(\hat{\gamma}_J)^{-1} = V(\hat{\gamma}_m)^{-1} + V(\hat{\gamma}_v)^{-1},$$

where $k^* = \text{dim}(\gamma)$, $\hat{\theta}_J = (\hat{\beta}_J', \hat{\alpha}_J')'$, $\hat{\gamma}_J = \gamma(\hat{\beta}_J)$, $\phi_v = (\hat{\gamma}_v', \phi_1')'$, $\hat{\gamma}_m = \gamma(\hat{\beta}_m)$, and

$$\Pi_u = V(\hat{\gamma}_J)V(\hat{\gamma}_v)^{-1} = I_{k^*} - V(\hat{\gamma}_J)V(\hat{\gamma}_m)^{-1}.$$

**Proof:** Define a nonsingular transformation from $\theta$ into $(\gamma, \gamma_c')$ where $\gamma_c$ completes the transformation, and apply Corollary 3.7.

The most relevant features of Theorem 10.1 and Corollary 10.2 are the efficiency of $\hat{\theta}_J$ over the estimator which ignores symmetry (as is apparent from (6) and (5)), the fact that the covariance matrix in (6) produces the correct standard errors, and that the simplicity of the MWA structure of the symmetric case is preserved by using the conditioned variance equation. Of course there is now the additional problem of parameterizing the conditional symmetry and kurtosis, but observe that a parametric estimate for $h_t$ can be obtained by LS in the mean equation and SLS in the variance equation as in Chapter 3, or the semiparametric approach of Carroll [1982] and Robinson [1987] can be used to provide equivalent nonparametric estimates of $h_t$ and $\kappa_t$.

The estimation of $\alpha_t$ will be discussed in § 10.2. The identifiability of $\alpha$ given $\beta$ is not affected in the conditioned variance equation because the mean equation does not contain information about $\alpha$. Another interesting aspect of this approach to the estimation of asymmetric heteroskedastic models is that it explains the cases when $\hat{\beta}_m$ is fully efficient because there is no information left in the conditioned variance equation. This is considered in
Theorem 10.3.- Under (C0) - (C7) the feasible GLS estimator $\hat{\beta}_m$ of $\beta$ is efficient with respect to information in the first two moments if, and only if, 
$$\text{rank } \mathcal{E} \left\{ T^{-1} ( S - \beta \Omega^{-1} \bar{X})' ( K - \beta \Omega^{-1} \beta )^{-1} ( S - \beta \Omega^{-1} \bar{X}) \right\} = p - k .$$

Proof: Let $V = \mathcal{E} \left\{ T^{-1} ( S - \beta \Omega^{-1} \bar{X})' ( K - \beta \Omega^{-1} \beta )^{-1} ( S - \beta \Omega^{-1} \bar{X}) \right\}$, and note that $S - \beta \Omega^{-1} \bar{X} = ( W - \beta \Omega^{-1} X , Z )$, because $\bar{X} = ( X , 0 )$. Therefore, 
$$\text{rank } [V] \geq \text{rank } \mathcal{E} \left\{ T^{-1} Z' ( K - \beta \Omega^{-1} \beta )^{-1} Z \right\} = p - k ,$$
where the last equality follows from the identifiability of $\alpha$ given $\beta$ in the variance equation. If $\text{rank } [V] = p - k$ there are only $p - k$ identifiable functions of $\theta$ in the CVE, and because the mean equation does not contain information about $\alpha$ and $\theta$ is identifiable in the whole model, the information in the CVE serves to identify $\alpha$ only, and therefore $V(\hat{\beta}_j) = V(\hat{\beta}_m)$. Conversely, if $V(\hat{\beta}_j) = V(\hat{\beta}_m)$ the CVE does not provide information about $\beta$ and $\text{rank } [V] \leq p - k$, following that $\text{rank } [V] = p - k . \qed$ 

Checking rank conditions like the one given in Theorem 10.3 is usually tedious, and a simpler, sufficient, condition is provided in

Corollary 10.4.- Under the assumptions of Theorem 10.3, the feasible GLS estimator $\hat{\beta}_m$ of $\beta$ is efficient with respect to information in the first two moments if 
$$\frac{\partial h_t}{\partial \beta} - \frac{\partial \hat{\mu}_t}{\partial \beta}$$
vanishes at $\theta_0$, for all $t$.

Proof: When 
$$\frac{\partial h_t}{\partial \beta} - \frac{\partial \hat{\mu}_t}{\partial \beta}$$
vanishes at $\theta_0$, for all $t$, the matrix $V$ of Theorem 10.3 has zeroes everywhere except for the submatrix $\mathcal{E} \{ T^{-1} Z'( K - \beta \Omega^{-1} \beta )^{-1} Z \}$, and hence $\text{rank } [V] = p - k . \qed$

Observe that a sufficient condition for $V$ to have the structure of this corollary is that $\mathcal{E} \left\{ T^{-1} ( W - \beta \Omega^{-1} X) Y ( K - \beta \Omega^{-1} \beta )^{-1} ( W - \beta \Omega^{-1} X ) \right\} = 0$, but this happens if, and only if, 
$$\frac{\partial h_t}{\partial \beta} - \frac{\partial \hat{\mu}_t}{\partial \beta}$$
vanishes at $\theta_0$, for all $t$, because the expectation involves a quadratic in a positive definite matrix. A particular case of interest is the gamma distributed model of Amemiya [1973] analyzed in
Theorem 10.5.- If $y_t \mid \mathcal{F}_t \sim \Gamma (\alpha^{-1}, \alpha \mu_t)$ and (Q1)-(Q8) hold, $\hat{\beta}_m$ is a fully efficient estimator of $\beta$.

Proof: The first three conditional moments of $y_t$ are $\mu_t$, $h_t = \alpha \mu_t^2$, and $\lambda_t = 2 \alpha^2 \mu_t^3$, respectively (e.g. Zellner [1971]: 370). Therefore, $\frac{\partial h_t}{\partial \beta} = 2 \alpha \mu_t x_t$, and since $h_t^{-1} \lambda_t = 2 \alpha \mu_t$ it follows that $\frac{\partial h_t}{\partial \beta} - \frac{\lambda_t}{h_t} \frac{\partial \mu_t}{\partial \beta} = 0$, and the result follows from Corollary 10.4.

Another interesting case is the Poisson model discussed in § 2.3.3,

$$y_t \mid \mathcal{F}_t \sim P [ \mu_t(\beta) ],$$

whose log-likelihood is

$$\ell (\beta) = T^{-1} \sum_{t=1}^{T} y_t \log \mu_t - T^{-1} \sum_{t=1}^{T} \mu_t,$$

and score

$$d_{\beta} (\beta) = T^{-1} \sum_{t=1}^{T} y_t \mu_t^{-1} x_t - T^{-1} \sum_{t=1}^{T} x_t = T^{-1} \sum_{t=1}^{T} \mu_t^{-1} x_t (y_t - \mu_t)$$

$$= T^{-1} X' \Omega^{-1} u ,$$

using the fact that $h_t = \mu_t$. The information matrix is easily seen to be

$$i (\beta) = E [ T^{-1} X' \Omega^{-1} X ] .$$

Hausman et al [1984] and Gourieroux et al [1984b] consider $\mu_t = \exp (x_t' \beta)$, so that $x_t = \mu_t x_t^*$, and their expressions for the score and information matrix are

$$d_{\beta} (\beta) = T^{-1} \sum_{t=1}^{T} x_t^* (y_t - \mu_t) ,$$

and $i (\beta) = E [ T^{-1} X' \Omega X' ]$.

Now (7) shows that $\hat{\beta}_m$ is asymptotically the MLE, and indeed we have

Theorem 10.6.- If $y_t \mid \mathcal{F}_t \sim P [ \mu_t(\beta) ]$ and (Q1)-(Q8) hold, $\hat{\beta}_m$ is a fully efficient estimator of $\beta$. 
Proof: The first three conditional moments of $y_t$ are $\mu_t = h_t = \lambda_t$. Therefore $\frac{\partial h_t}{\partial \beta} = x_t$ and $h_t^{-1} \lambda_t = 1$, and it follows that $\frac{\partial h_t}{\partial \beta} - \frac{\lambda_t}{h_t} \frac{\partial \mu_t}{\partial \beta} = 0$, and the result follows from Corollary 10.4.

§ 10.2 Extracting information from higher order moments

Now suppose that we want to extract information from higher order moments. It would be presumptuous to believe that theoretical considerations could suggest parameterizations for these moments, so an empirical exploration might start from three considerations:

- Interpret the proposition $Y = f(X^*)$ as "all existing moments of the conditional distribution of $Y$ given $X^*$ are functions of $X^*$". Thus use functions of $X^*$ as arguments.

- Although we are talking about 'free' moments (i.e. not determined by lower order ones), it might be sensible to relate higher to lower order moments. This can permit, among other things, tests of whether the conditional DGP belongs to a simpler class that can be characterized by fewer moments.

- As information is made available, expectations must be revised not only for the first two moments, but for all (existing) moments, introducing dynamics by relating $E[ u_t^r | \mathcal{F}_t ]$ to its past and to lagged values of $u_t^r$, for all $r$. Engle [1982a] notes that this argument for $r = 1$ has been the force behind the Box-Jenkins [1971] methodology, and no doubt it is also an important foundation of GARCH processes. Our suggestion is that the same argument may be used for higher order moments.

These considerations can be summarized in a proposition for the $r$-th moment of the form
\[ h_t^{(r)} = E \left[ (y_t - \mu_t)^r \mid \mathcal{F}_t \right] = h_t^{(r)} [x_t^*, \mu_t, h_t^{(2)}] \text{, } ..., \text{ } h_t^{(r-1)} ; u_{t-j}^r, h_{t-j}^{(r)} \text{, } j > 0 \] - (8a)

If we parameterize the conditional moments sequentially in this form, a new set of parameters \( \alpha_r \), say, is incorporated with each additional moment. Note that \( h_t^{(1)} = 0 \) and \( h_t^{(2)} \) is the conditional variance, and let \( \alpha_1 = \beta \). Then

\[ h_t^{(r)} = E \left[ u_t^r \mid \mathcal{F}_t \right] = h_t^{(r)} (\beta, \alpha_2, \ldots, \alpha_r ; \mathcal{F}_t) \] - (8b)

Suppose \( N \) moments are to be explored and let \( k_r = \text{dim} (\alpha_r) \). The parameter vector is \( \theta = (\beta', \alpha_2', \ldots, \alpha_N') \), with dimension \( p = \sum_{r=1}^{N} k_r \). Let \( \theta_r = \beta', \alpha_2', \ldots, \alpha_r' \), so that \( \theta_1 = \beta, \ldots, \theta_N = \theta \), and we can write \( h_t^{(r)} = h_t^{(r)} (\theta_r) \). A sequential parameterization like (8) has the advantage that estimating the \( r \)-th moment produces a first estimate of \( \alpha_r \) and may improve the efficiency of the estimates of \( \beta, \alpha_2, \ldots, \alpha_r \) obtained from the first \( r - 1 \) moments. The disadvantage is that it does not incorporate higher-to-lower moment effects like the risk terms of the ARCH-M model (Engle et al [1987]). The results can be generalized as was done for the case \( N = 2 \) in Chapter 7, but this would have a cost in terms of modelling strategy, as we will see below. Define the innovations in the \( r \)-th moment,

\[ \varepsilon_t^{(r)} = u_t^r - h_t^{(r)} \]

so that \( E [\varepsilon_t^{(r)} \mid \mathcal{F}_t] = 0 \), and the \( r \)-th moment equation is naturally defined as

\[ u_t^r = h_t^{(r)} (\theta_r) + \varepsilon_t^{(r)} \]

Note that \( \varepsilon_t^{(1)} = u_t \). The covariance between innovations of different moments is

\[ E [\varepsilon_t^{(r)} \varepsilon_t^{(s)} \mid \mathcal{F}_t] = h_t^{(r+s)} - h_t^{(r)} h_t^{(s)} \]

and using iterated expectations we have \( E [\varepsilon_t^{(r)} \varepsilon_t^{(s)}] = 0 \) for \( t \neq t' \). Stacking the \( N \) equations we have the system innovations

\[ \nu_t = \eta_t - E [\eta_t \mid \mathcal{F}_t] = \eta_t - g_t (\theta) \] - (9)
and the system of equations

$$\eta_t = g_t(\theta) + \nu_t,$$

where $\eta_t = (y_t, u_t^2, \ldots, u_t^N)'$, $g_t = (\mu_t, h_t^{(2)}, \ldots, h_t^{(N)})'$, and $\nu_t = (\varepsilon_t^{(2)}, \ldots, \varepsilon_t^{(N)})'$. Assuming the conditional distribution possesses finite $2N$-th moments, the conditional covariance matrix is

$$\Sigma_t = E[\nu_t \nu_t' | \mathcal{F}_t] = \sigma_{\text{trf}} = \|h_t^{(r+s)} - h_t^{(r)}h_t^{(s)}\|.$$

From a theoretical perspective, the estimation of (10) is qualitatively the same as that of the mean and variance two-equation system of heteroskedastic models. The identifiability problems of $\theta_r$ in the $r$-th moment equation can be treated essentially as in Chapter 3 and will be ignored for the sake of simplicity. The non-observability of $u_t^r$ introduces qualitatively different issues for symmetric and non-symmetric distributions and we defer its treatment for the sake of generality, and thus estimators are starred to denote that they are not feasible estimators. Assumptions (C10) - (C7) of Chapter 2 must be modified according to the new structure. We do not discuss them in detail but note that smoothness and dominance assumptions must now apply to the first $N$ moments, existence assumptions must include $2N$ moments, and all even moments must be bounded away from zero. We will refer to this set of assumptions as (C10-N) - (C7-N).

The conditional covariance matrix $\Sigma_t$ is not diagonal and hence the equations in (10) are correlated. To eliminate this correlation we proceed sequentially as in § 10.1 to produce a conditioned version of the $r$-th moment equation, given the information in lower order moments. For this purpose, let $\tilde{e}_t^{(r)}$ be the innovations in the conditioned $r$-th moment equation, which are obtained sequentially by defining $e_t^{(1)} = \varepsilon_t^{(1)} = u_t$, and for $r = 2, \ldots, N$

$$c_t(r,s) = \text{cov} [\tilde{e}_t^{(r)}, \tilde{e}_t^{(s)}] \quad \text{for} \quad s \leq r,$$
\[ v_t(r) = \text{var} \left[ \tilde{\varepsilon}_t^{(r)} \right], \]

and

\[ \tilde{\varepsilon}_t^{(r)} = \tilde{\varepsilon}_t^{(s)} - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)} \tilde{\varepsilon}_t^{(j)}. \]  \hspace{1cm} -(12)

We then have

**Lemma 10.7.** \( \text{cov} \left[ \tilde{\varepsilon}_t^{(r)} , \tilde{\varepsilon}_t^{(s)} \right] = 0 , \text{for } r , s = 1 , \ldots , N , r \neq s . \)

**Proof:** We use induction over \( r \), and because \( \text{cov} \left[ \tilde{\varepsilon}_t^{(r)} , \tilde{\varepsilon}_t^{(s)} \right] = \text{cov} \left[ \tilde{\varepsilon}_t^{(s)} , \tilde{\varepsilon}_t^{(r)} \right] \) it suffices to consider \( s < r \). For \( r = 2 \), \( c_t(2,1) = \text{cov} \left[ \tilde{\varepsilon}_t^{(2)} , u_t \right] = h_t^{(3)} = \lambda_t \), and \( v_t(1) = \text{var} \left[ u_t \right] = h_t^{(2)} = h_t \). Then \( \tilde{\varepsilon}_t^{(2)} = \tilde{\varepsilon}_t^{(2)} - h_t^{-1} \lambda_t u_t \), and it follows from section § 10.1 that \( \text{cov} \left[ \tilde{\varepsilon}_t^{(2)} , \tilde{\varepsilon}_t^{(1)} \right] = 0 \). Now suppose that \( \text{cov} \left[ \tilde{\varepsilon}_t^{(r-1)} , \tilde{\varepsilon}_t^{(s)} \right] = 0 , s < r-1 \). Then, using (12), we have for \( s < r \),

\[ \text{cov} \left[ \tilde{\varepsilon}_t^{(r)} , \tilde{\varepsilon}_t^{(s)} \right] = \text{cov} \left[ \left( \tilde{\varepsilon}_t^{(r)} - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)} \tilde{\varepsilon}_t^{(j)} \right) , \tilde{\varepsilon}_t^{(s)} \right] = c_t(r,s) - \frac{c_t(r,s)}{v_t(s)} v_t(s) = 0 . \]

The functions \( c_t(r,s) \) and \( v_t(r) \) can be computed recursively for \( r = 2 , \ldots , N \) using (11) and the expression

\[ c_t(r,s) = \text{cov} \left[ \tilde{\varepsilon}_t^{(r)} , \tilde{\varepsilon}_t^{(s)} \right] - \sum_{j=1}^{s-1} \frac{c_t(s,j)}{v_t(j)} \tilde{\varepsilon}_t^{(j)} = \text{cov} \left[ \tilde{\varepsilon}_t^{(r)} , \tilde{\varepsilon}_t^{(s)} \right] - \sum_{j=1}^{s-1} \frac{c_t(s,j) c_t(r,j)}{v_t(j)} , \]

for \( s \leq r \), noting that \( v_t(r) = c_t(r,r) \) in view of the conditioned nature of \( \tilde{\varepsilon}_t^{(j)} \).

The conditioned r-th moment equation is then

\[ u_t^r = \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)} u_t^j = h_t^{(r)} (\theta_r) - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)} h_t^{(j)} (\theta_j) + \tilde{\varepsilon}_t^{(r)}, \]

\hspace{1cm} -(13)

where we take the \( c_t(r,s) \) and \( v_t(r) \) as given. This can be done because the estimation of \( \theta_r \) is not affected to the order of \( T^{1/2} \) if root-T consistent estimates are used to construct the \( c_t(r,s) \) and \( v_t(s) \), and root-T consistent estimates of \( \theta \) may be obtained using least-squares in the mean equation, and simple least squares sequentially in the r-th moment equations, that is, least squares in

\[ u_t^r = h_t^{(r)} (\alpha_r ; \theta_r^{(r-1)}) + \tilde{\varepsilon}_t^{(r)}, \]
where \( \hat{e}_t^{(r)} = \xi_t^{(r)} + (h_t^{(r)}(\theta_r) - h_t^{(r)}(\alpha_r; \tilde{\theta}_r^{(r-1)})) \), and \( \tilde{\theta}_r^{(r-1)} = (\hat{\beta}_2^{(r)}, \hat{\theta}_2^{(r)}, \ldots, \hat{\theta}_r^{(r-1)})' \), for \( r = 2, \ldots, N \). To compact the notation, let the dependent variable in the left-hand-side of (13) be \( u_t^{(r-1)} \), and the regression function (the right-hand-side except \( \tilde{e}_t^{(r)} \)) be \( h_t^{(r-1)}(\theta_r) \). We can then rewrite

\[
\begin{align*}
u_t^{(r-1)} &= h_t^{(r-1)}(\theta_r) + \tilde{e}_t^{(r)}.
\end{align*}
\]

To estimate \( \theta_r \) efficiently from information on the \( r \)-th equation only, consider the set of orthogonality conditions

\[
\psi_r^*(\theta_r) = T^{-1} \sum_{t=1}^{T-1} v_t^{(r)} s_{tr} \tilde{e}_t^{(r)} = T^{-1} S_r' \Omega_r^{-1} \tilde{e}^{(r)},
\]

where \( s_{tr} = \partial h_t^{(r-1)} / \partial \theta_r \), \( S_r = (s_{1r}, \ldots, s_{Tr})' \), \( \Omega_r = \text{diag} \{ v_t^{(r)} \} \), and \( \tilde{e}^{(r)} = (\tilde{e}_1^{(r)}, \ldots, \tilde{e}_T^{(r)})' \). Under (C0-N) - (C7-N) we use Theorem 2.1 to establish that \( \hat{\theta}_r^* \) obtained from \( \psi_r^*(\theta_r) \) is a strongly consistent estimator of the true value \( \theta_r^0 \) of \( \theta_r \), and has asymptotic distribution

\[
T^{1/2}(\hat{\theta}_r^* - \theta_r^0) \xrightarrow{d} N[0, S (T^{-1} S_r' \Omega_r^{-1} S_r)^{-1}],
\]

where the expectation for the covariance matrix is evaluated at \( \theta_0 \).

For the estimation of \( \theta_r \) for \( r > N/2 \) we require conditional moments of order higher than \( N \) to construct \( \Omega_r \). In all we need moments of order up to \( 2N \), and there are essentially three possibilities. One is that the pdf be fully characterized by the first \( N \) moments, and the relationships of these with moments of order higher than \( N \) is known, so that the OLS/SLS estimators of \( \theta \) allow the calculation of the covariance matrices. Example of this are the kurtosis function under normality, and the Student's t distribution and Pearson family of distributions for which four moments suffice. A second possibility is to use a semi-parametric approach of Carroll [1982] or Robinson [1987] so that higher order moments are estimated non-parametrically. It is not easy to see how this should be done and what conditions would need to be imposed, and the sample size required for reasonable approximations to the
asymptotic distribution might be too large. The third possibility is to parameterize the moments up to order $2N$. Because we need root-$T$ consistent estimators of the parameters of these moments, this in general requires the existence of moments of order up to $4N$.

Because there exists information about common parameters in different equations it is of interest to consider joint estimation based on the orthogonality conditions
\[
\psi^* (\theta) = (\psi_m (\beta)' , \psi_2^* (\theta_2)' , \ldots , \psi_N^* (\theta)' )',
\]
and weighting matrix
\[
A_T = \text{diag} \{ T (X' \Omega^{-1} X)^{-1} , T (S_2' \Omega_2^{-1} S_2)^{-1} , \ldots , T (S_N' \Omega_N^{-1} S_N)^{-1} \}.
\]

We then have

**Theorem 10.8.** - Under $(\zeta 0$-N) - $(\zeta 7$-N) the GMM estimator $\hat{\theta}_N^*$ obtained from the orthogonality conditions $\psi^* (\theta)$ and weighting matrix $A_T$ is strongly consistent and has asymptotic distribution
\[
T^{1/2} (\hat{\theta}_N^* - \theta_0) \xrightarrow{d} N[0 , \sigma \{ T^{-1} \sum_{r=1}^N \bar{S}_r' \Omega_r^{-1} \bar{S}_r \}^{-1} ],
\]
where $S_j \equiv X_j$, $\Omega_j \equiv \Omega$, and $\bar{S}_j = (S_j , 0 )$ are $T \times p$ matrices, $j = 1 , \ldots , N$, and the expectation for the covariance matrix is evaluated at $\theta_0$.

**Proof:** Our assumptions conform to Theorem 2.1. Because the equations are uncorrelated,
\[
E [ T \psi^* (\theta) \psi^* (\theta)' ] = E [ A_T (\theta_0) ] = \text{diag} \{ E [ T^{-1} S_r' \Omega_r^{-1} S_r ] \},
\]
and using iterated expectations it is seen that $E [ \partial \psi^* (\theta)' / \partial \theta ] = - E [ T^{-1} S_r' \Omega_r^{-1} \bar{S}_r ]$, so that
\[
E [ \frac{\partial \psi^* (\theta_0)' }{\partial \theta} ] = - ( E [ T^{-1} X' \Omega^{-1} X ] , \ldots , E [ T^{-1} S_N' \Omega_N^{-1} S_N ] ),
\]
and the covariance matrix of \( \hat{\theta} \) is therefore

\[
V(\hat{\theta}_N^*) = \mathcal{E} \left( \sum_{r=1}^{N} \mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r (\mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r)^{-1} \mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r \right)^{-1}.
\]

But \( \mathcal{S}_r = \mathcal{I} \), and therefore

\[
\mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r (\mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r)^{-1} \mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r = \mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r,
\]

which completes the proof.

If we define the \( p \times p_r \) matrix \( B_r = (I_{p_r}, 0)' \), where \( p_r = \sum_{j=1}^{r} k_j \), then

\[
V(\hat{\theta}_N^*) = \left( \sum_{r=1}^{N} B_r V(\hat{\theta}_r^*) B_r' \right)^{-1},
\]

and because \( T^{1/2} \mathcal{S}_r \Omega_r^{-1} \mathcal{S}_r \mathcal{E}(0) = \left( T^{1/2} \mathcal{E}(0) \Omega_r^{-1} \mathcal{S}_r, 0 \right)' = B_r V(\hat{\theta}_r^*) T^{1/2} \left( \hat{\theta}_r^* - \theta_0 \right) + \mathcal{O}_p(1) \), it is further seen that

\[
T^{1/2} (\hat{\theta}_N^* - \theta_0) = V(\hat{\theta}_N^*) \sum_{r=1}^{N} B_r V(\hat{\theta}_r^*) T^{1/2} \left( \hat{\theta}_r^* - \theta_0 \right) + \mathcal{O}_p(1), \tag{15}
\]

which shows the implicit MWA structure of the joint estimator. An interesting strategy for estimation is to proceed sequentially exploring the availability of information in one additional moment at the time, until no further improvement in efficiency is achieved. This can be done in a simple way because each time information from a new moment is incorporated by means of a conditioned equation, the new estimator is a MWA of the estimator from the previous moments and the estimator arising from the new information. To see this partition \( \theta_r = (\theta_r', \alpha_r')' \), \( \hat{\theta}_r^* = (\hat{\theta}_r'^*, \hat{\alpha}_r'^*)' \) (the estimator of \( \theta_r \) obtained from information in the conditioned \( r \)-th moment equation alone), and \( \hat{\theta}_r^* = (\hat{\theta}_r'^*, \hat{\alpha}_r'^*)' \) (the estimator of \( \theta_r \) obtained from all the information up to and including the \( r \)-th moment). We prove

**Theorem 10.9.-** Under \((\mathcal{Q}_0^N) - (\mathcal{Q}_7^N)\),

\[
\hat{\theta}_{r-1}^* = V(\hat{\theta}_{r-1}^*) \left[ V(\hat{\theta}_{r-1}^*)^{-1} \hat{\theta}_{r-1}^* + V(\hat{\theta}_{r-1}^*)^{-1} \hat{\theta}_{r-1}^* \right] + \mathcal{O}_p(T^{-1/2}),
\]

\[
V(\hat{\theta}_{r-1}^*)^{-1} = V(\hat{\theta}_{r-1}^*)^{-1} + V(\hat{\theta}_{r-1}^*)^{-1},
\]

\[
\hat{\alpha}_r^* = \alpha_r + \text{cov} [\hat{\alpha}_r^*, \hat{\theta}_{r-1}^*] V(\hat{\theta}_{r-1}^*)^{-1} \left[ \hat{\theta}_{r-1}^* - \theta_{r-1}^* \right] + \mathcal{O}_p(T^{-1/2}),
\]

\[
\text{cov} [\hat{\alpha}_r^*, \hat{\theta}_{r-1}^*] = \text{cov} [\alpha_r, \theta_{r-1}^*] V(\hat{\theta}_{r-1}^*)^{-1} \left[ \hat{\theta}_{r-1}^* - \theta_{r-1}^* \right] + \mathcal{O}_p(T^{-1/2}),
\]

\[
\text{cov} [\hat{\theta}_{r-1}^*, \hat{\theta}_{r-1}^*] = \text{cov} [\theta_{r-1}^*, \theta_{r-1}^*] V(\hat{\theta}_{r-1}^*)^{-1} \left[ \hat{\theta}_{r-1}^* - \theta_{r-1}^* \right] + \mathcal{O}_p(T^{-1/2})
\]
\[ V(\hat{\alpha}_r^*, \tilde{\theta}_{r-1}^*) = V_r + \text{cov}[\hat{\alpha}_r^*, \tilde{\theta}_{r-1}^*] V(\tilde{\theta}_{r-1}^*)^{-1} \text{cov}[\tilde{\theta}_{r-1}^*, \hat{\alpha}_r^*], \]

and

\[ \text{cov}[\hat{\alpha}_r^*, \tilde{\theta}_{r-1}^*] = -V_r \mathcal{E}\{T^{-1} S_{rT} \Omega_r^{-1} S_{r-1}^* \} V(\tilde{\theta}_{r-1}^*), \]

where \(V_r = \mathcal{E}\{T^{-1} S_{rT} \Omega_r^{-1} S_{rT}\}\), and \(S_r = (S_{r-1}^*, S_{rT})\).

**Proof:** In view of the sequential parameterization of \(h_t^{(r)}\) and the conditioned nature of the equations, the problem of incorporating the information in the r-th moment once the first \(r - 1\) moments have been accounted for has the same structure as that of incorporating the variance information in the symmetric heteroskedastic case, and we can apply to \(\tilde{\theta}_r = (\tilde{\theta}_{r-1}^*, \hat{\alpha}_r)\)' Corollaries 3.7 and 3.8 to relate it to \(\theta_{r-1}\) (the estimator of \(\theta_{r-1}\) from the first \(r - 1\) moments) and \(\tilde{\theta}_r = (\tilde{\theta}_{r-1}^*, \tilde{\alpha}_r)\)' (the estimator from the conditioned r-th moment).

When there is no more information about \(\theta_{r-1}\) in the r-th moment equation, this can be detected by a criterion similar to that of Theorem 10.3, as we see in

**Theorem 10.10.** Under \((C0-N) - (C7-N)\), \(\hat{\theta}_{r-1}^*\) is efficient with respect to the information contained in the first \(r\) moments if, and only if,

\[ \text{rank} \mathcal{E}\{T^{-1} (S_r - \sum_{j=1}^{r-1} C_{rj} \Omega_j^{-1} S_{jr}) \} (\Omega_r^{-1} C_{rj}^{-1} C_{rj} \Omega_r^{-1} S_{jr} \} = k_{r}, \]

where \(C_{rj} = \text{diag}\{c_t(r,j)\}\), and \(S_{jr} = (S_j, 0)\) so that it is \(T \times \sum_{j=1}^{r-1} k_j, j = 1, \ldots, r - 1\).

**Proof:** Let \(V\) be the matrix in the Theorem and observe that \(S_r - \sum_{j=1}^{r-1} C_{rj} \Omega_j^{-1} S_{jr} = (S_{r-1}^*, \sum_{j=1}^{r-1} C_{rj} \Omega_j^{-1} S_{jr}^*), S_{rT})\), where \(S_{jr} = (S_j^*, 0)\). The proof follows with the same argument as Theorem 10.3.

A simpler sufficient condition is given in

**Corollary 10.11.** Under the assumptions of Theorem 10.10, \(\hat{\theta}_{r-1}^*\) is efficient with respect to the information contained in the first \(r\) moments if

\[ \frac{\partial h_t^{(r)}}{\partial \theta_{r-1}} - \sum_{j=1}^{r-1} v_t(j)^{-1} c_t(r,j) \frac{\partial h_t^{(j)}}{\partial \theta_{r-1}} \text{ vanishes at } \theta_0, \text{ for all } t. \]
Proof: When \( \frac{\partial h_t^{(r)}}{\partial \theta_{r-1}} - \sum_{j=1}^{r-1} v_t(j)^{-1} c_t(r,j) \frac{\partial h_t^{(i)}}{\partial \theta_{r-1}} \) vanishes at \( \theta_0 \), for all \( t \), the matrix \( V \) of Theorem 10.10 has zeroes everywhere except for the submatrix 
\[ \mathcal{C}(T^{-1} S_r' \sum_j C_{rj} \Omega_j^{-1} C_{rj}^{-1} S_r) \], and hence \( \text{rank}[V] = k_r \).

As an illustration consider the conditional Student’s t distribution with mean \( \mu_t(\beta) \), variance \( \sigma_t(\beta_2) = \sigma_t(\beta, \alpha_2) \), and degrees of freedom \( \alpha_4 \) (e.g. Bollerslev [1985], Engle and Bollerslev [1986]). The even order moments are (Zellner [1971]: 366)

\[
h_t^{(2r)} = c_r d_r(\alpha_4) \sigma_t^{(2r)}, \quad 2r < \alpha_4,
\]

where \( c_r = \prod_{j=1}^{r} (2j-1) \) and \( d_r(\alpha_4) = (\alpha_4 - 2)^r / \prod_{j=1}^{r} (\alpha_4 - 2j) \).

Then for the fourth moment

\[
\frac{\partial h_t^{(4)}}{\partial \theta_2} = 2 c_2 d_2(\alpha_4) \frac{\partial h_t^{(2)}}{\partial \theta_2} = \frac{6(\alpha_4 - 2)}{(\alpha_4 - 4)} \frac{\partial h_t^{(2)}}{\partial \theta_2},
\]

and

\[
c_t(4,2) \frac{\partial h_t^{(6)}}{\partial \theta_2} - \frac{\partial h_t^{(4)}}{\partial \theta_2} - \frac{\partial h_t^{(2)}}{\partial \theta_2} = \frac{12(\alpha_4 - 2)}{(\alpha_4 - 4)(\alpha_4 - 6)} \frac{\partial h_t^{(2)}}{\partial \theta_2}.
\]

As \( \alpha_4 \to \infty \) this quantity goes to zero thus showing that no information is available in the fourth moment of the conditional normal distribution. For the conditional t distribution, the fourth moment always contains information about the degrees-of-freedom parameter \( \alpha_4 \), and the information it contains about \( \theta_2 \) will depend inversely on \( \alpha_4 \).

Up to now we have ignored the unobservability of \( u_t \), and so the estimators are not feasible. To obtain practical results, we must use residuals from the mean equation to construct the dependent variables of the higher order moment equations. The operative version of the \( r \)-th moment equation is
\[ \tilde{u}_t^{(r)} = h_t^{(r)}(\theta_t) + \epsilon_t^{(r)}, \]

where \( \epsilon_t^{(r)} = \tilde{\epsilon}_t^{(r)} + \{ \tilde{u}_t^{(r)} - u_t^{(r)} \} \), and the operative version of the conditioned r-th moment equation is

\[ \tilde{u}_t^{(r/r-1)} = h_t^{(r/r-1)}(\theta_t) + \epsilon_t^{(r/r-1)}, \]

where \( \tilde{\epsilon}_t^{(r/r-1)} = \tilde{\epsilon}_t^{(r)} + \{ \tilde{u}_t^{(r/r-1)} - u_t^{(r/r-1)} \} \), and \( \tilde{u}_t^{(r/r-1)} \) and \( \tilde{u}_t^{(r/r-1)} \) are obtained by substituting root-T consistent estimates \( \tilde{\beta} \) of \( \beta \). \( \tilde{\beta}_m \) can be obtained without information from the third and higher order moments, and so we assume in what follows that \( \tilde{\beta} = \tilde{\beta}_m \). Because \( V(\tilde{\beta}) \) will affect directly the covariance matrices of feasible estimators, using \( \tilde{\beta}_m \) is the best device for efficiency purposes. The orthogonality conditions \( \psi_t^{*}(\theta_t) \) are replaced to obtain the feasible estimators \( \tilde{\theta}_t \) by

\[ \psi_t(\theta_t) = T^{-1} \sum_{t=1}^{T} v_t^{(r)-1} s_{tr} \tilde{\epsilon}_t^{(r)} = T^{-1} S_r \Omega_r^{-1} \tilde{e}^{(r)}, \]

and we have

**Theorem 10.12.** Under (C0-N) - (C7-N),

\[ T^{1/2} ( \tilde{\theta}_t - \theta_t^0 ) = T^{1/2} ( \tilde{\theta}_t^* - \theta_t^0 ) + V ( \tilde{\theta}_t^* ) A_r T^{1/2} ( \tilde{\beta}_m - \beta_0 ) + o_p(1), \]

where \( A_r = \mathcal{C} \{ T^{-1} \sum_{t=1}^{T} v_t^{(r)-1} \left[ \frac{1}{h_t^{(r-1)}} - \frac{r-1}{v_t^{(r-1)}} \frac{c_t^{(r,j)}}{v_t^{(r,j)}} h_t^{(r-1)} \right] s_{tr} x_t \} \) for \( r > 2 \), and \( A_r = 0 \) for \( r \leq 2 \).

**Proof:** Using \( \tilde{\epsilon}_t^{(r)} \) instead of \( \tilde{\epsilon}_t^{(r)} \) in the orthogonality conditions \( \psi_t(\theta_t) \) results in

\[ T^{1/2} ( \tilde{\theta}_t - \theta_t^0 ) = T^{1/2} ( \tilde{\theta}_t^* - \theta_t^0 ) + V ( \tilde{\theta}_t^* ) T^{1/2} \sum_{t=1}^{T} v_t^{(r)-1} s_{tr} \{ \tilde{u}_t^{(r/r-1)} - u_t^{(r/r-1)} \} + o_p(1), \]

where the second term in the right-hand-side is the contribution due to the unobservability of \( u_t^{(r)} \). Substituting \( u_t^{(r/r-1)} \) from (13) and (14) and using Lemma 3.3 this produces

\[ T^{1/2} ( \tilde{\theta}_t - \theta_t^0 ) = T^{1/2} ( \tilde{\theta}_t^* - \theta_t^0 ) + V ( \tilde{\theta}_t^* ) A_r T^{1/2} ( \tilde{\beta}_m - \beta_0 ) + o_p(1), \]
where \( A_r = \mathcal{E} \left( T^{-1} \sum_{t=1}^{T} v_t (r)^{-1} \left[ h_t^{(r-1)} - \sum_{j=3}^{r-1} \frac{c_t (r,j)}{v_t (j)} h_t^{(j-1)} \right] s_{rt} x_t' \right) \) for \( r > 1 \). \( A_1 = 0 \) because the mean equation does not have unobservability problems, and \( A_2 = 0 \) because \( h_t^{(1)} = 0 \).

The Theorem shows that the feasible estimators remain consistent. This is due to the fact that \( T^{-1} \sum_{t=1}^{T} \tilde{f}_t \left[ \tilde{u}_t^r - u_t^r \right] \rightarrow 0 \) as long as \( E [ u_t^r ] \) exists, for any \( r \) (see Lemma 3.3), but the distribution of the feasible estimators will not be equal to that of the nonfeasible estimators unless the second term in (16) vanishes fast enough. For this we need the stronger condition \( T^{1/2} \sum_{t=1}^{T} \tilde{f}_t \left[ \tilde{u}_t^r - u_t^r \right] \rightarrow 0 \), which is satisfied by the first two moments, and for higher order moments this suggests that the feasible estimators have a different asymptotic properties depending on whether the conditional distribution of \( y_t \) is symmetric or asymmetric. In fact we have

**Corollary 10.13.** Under the assumptions of Theorem 10.12 and if the conditional distribution of \( y_t \) is symmetric, then for \( r \) even,

\[
T^{1/2} ( \tilde{\theta}_r - \theta_r^0 ) = T^{1/2} ( \tilde{\theta}_r^* - \theta_r^0 ) + o_p (1).
\]

**Proof:** When the distribution is symmetric \( h_t^{(r)} = 0 \) for \( r \) odd and the odd order equations for \( r > 1 \) are deleted from the system expressions in (9) - (11). Hence the conditioned equations only remove the correlation with lower order even moments, but for these the matrix \( A_r \) is function of \( h_t^{(j)} \) only for odd \( j \), and it follows that \( A_r = 0 \).

Hence under a symmetric conditional distribution the feasible and nonfeasible estimators have the same asymptotic distribution and in this case the odd order equations for \( r > 1 \) are deleted from the system expressions in (9) - (11). These odd order equations could be used as conditional moment restrictions to improve efficiency in estimation as in Newey [1986]. For asymmetric conditional distributions we have
**Corollary 10.14.** Under the assumptions of Theorem 10.12 and if the conditional distribution of $y_t$ is asymmetric, for $r > 2$,

\[ T^{1/2} \left( \hat{\theta}_r - \theta_r^0 \right) \overset{d}{\rightarrow} N \left[ 0, V(\hat{\theta}_r^*) + V(\hat{\theta}_r^*) A_r V(\hat{\beta}_m) A_r' V(\hat{\theta}_r^*) \right]. \quad (17) \]

**Proof:** By construction \( \text{cov} (\hat{\theta}_r^*, \hat{\theta}_s^*) = 0 \) for \( r \neq s \), and in particular \( \text{cov} (\hat{\theta}_r^*, \hat{\beta}_m) = 0 \) for \( r \geq 2 \). Therefore the two terms in (16) are asymptotically independent and the covariance matrix in (17) follows. \( \square \)

The first component of \( V(\hat{\theta}_r) \) is provided by a regression package after GLS estimation of the operative conditioned \( r \)-th moment equation, but White's [1980b] covariance matrix estimator does not account for the second term and this has to be computed separately.

Let us now consider extracting information from the whole set of moment equations. Substituting the feasible estimators in (15) we get the feasible estimator \( \hat{\theta}_N \) of \( \theta_0 \),

\[ T^{1/2} (\hat{\theta}_N - \theta_0) = V(\hat{\theta}_N^*) \sum_{r=1}^{N} B_r V(\hat{\theta}_r^*)^{-1} T^{1/2} (\hat{\theta}_r - \theta_r^0) + o_p(1), \quad (18) \]

for which we have

**Theorem 10.15.** Under (Q0-N) - (Q7-N), if the conditional distribution of \( y_t \) is symmetric,

\[ T^{1/2} (\hat{\theta}_N - \theta_0) = T^{1/2} (\hat{\theta}_N^* - \theta_0) + o_p(1), \]

and if the conditional distribution of \( y_t \) is asymmetric,

\[ T^{1/2} (\hat{\theta}_N - \theta_0) \overset{d}{\rightarrow} N \left[ 0, V(\hat{\theta}_N^*) B V(\hat{\theta}_N^*) \right], \quad (19) \]

where

\[ B = \left( \sum_{r=3}^{N} B_r A_r \right) V(\hat{\beta}_m) \left( \sum_{r=3}^{N} A_r' B_r' \right) + B_1 \left( \sum_{r=3}^{N} A_r' B_r' \right) + \left( \sum_{r=3}^{N} B_r A_r \right) B_1'. \quad (20) \]
Proof: For symmetric distributions the result follows from Corollary 10.13.

For asymmetric distributions, substitute (16) in (18) to get

\[ T^{1/2} (\hat{\theta}_N - \theta_0) = T^{1/2} (\hat{\theta}_N^* - \theta_0) + V (\hat{\theta}_N^*) \left( \sum_{r=1}^{N} B_r A_r \right) T^{1/2} (\hat{\beta}_m - \beta_0) + o_p (1). \]

Now,

\[ \text{cov} (\hat{\theta}_N^*, \hat{\beta}_m) = V (\hat{\theta}_N^*) \sum_{r=1}^{N} B_r V (\hat{\theta}_r^*)^{-1} \text{cov} (\hat{\theta}_r^*, \hat{\beta}_m) = V (\hat{\theta}_N^*) B_1, \]

because \( T^{1/2} (\hat{\beta}_m - \hat{\theta}_1^*) = o_p (1) \) and \( \text{cov} (\hat{\theta}^*, \hat{\beta}_m) = 0 \) for \( r > 1 \) by construction. Therefore,

\[ T^{1/2} (\hat{\theta}_N - \theta_0) \overset{d}{\to} N [ 0, V (\hat{\theta}_N^*) + V (\hat{\theta}_N^*) B V (\hat{\theta}_N^*) ], \]

where \( B = ( \sum_{r=3}^{N} B_r A_r ) V (\hat{\beta}_m) ( \sum_{r=3}^{N} A_r B_r' ) + B_1 ( \sum_{r=3}^{N} A_r B_r' ) + ( \sum_{r=3}^{N} B_r A_r ) B_1'. \]

In asymmetric problems, the first term of the covariance matrix would be directly produced by the joint estimation of the \( N \) moments from the set of conditioned equations. Observe that the sums for the construction of \( B \) start from 3 because the mean and variance equations have no observability problem.

The attractive feature of constructing the joint estimator in this fashion is that we can follow a sequential search of information in higher order moments, stopping the search when there are no further payoffs. An alternative procedure is to proceed to full estimation of the system without conditioning the equations after initial SLS estimation (plus a second round in the first moment to make \( \hat{\beta}_m \) available and improve efficiency). Because the system of \( N \) moment equations is transformed into the system of conditioned moment equations by means of a nonsingular transformation, the nonfeasible joint estimator \( \hat{\theta}_N^* \) is equivalently obtained from the orthogonality conditions

\[ \psi (\theta) = T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} v_t = T^{-1} G' \Sigma^{-1} v, \]

where we now define \( G = ( \frac{\partial g_1'}{\partial \theta}, ..., \frac{\partial g_T'}{\partial \theta} )', \Sigma = \text{diag} \{ \Sigma_t \}, \) and \( v = (v_1', ..., v_T')' \).

Hence an equivalent expression for \( V (\hat{\theta}_N^*) \) is
\[ V(\hat{\theta}_N^*) = \mathcal{E}\left\{ T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} \frac{\partial g_t}{\partial \theta} \right\} = \mathcal{E}\left\{ T^{-1} G' \Sigma^{-1} G \right\}, \]
evaluating the expectation at \( \theta_0 \). Also,

\[ T^{1/2}(\hat{\theta}_N^* - \theta_0) = V(\hat{\theta}_N^*) T^{-1/2} G' \Sigma^{-1} \nu + o_p(1). \]  

Let \( \tilde{\eta}_t = (y_t, u_t^2, \ldots, u_t^N)' \) and \( \tilde{\nu}_t = \tilde{\eta}_t - \tilde{g}_t = \nu_t + \{ \tilde{\eta}_t - \eta_t \} \), and stacking in obvious form \( \tilde{\nu} = \tilde{\eta} - \bar{g} = \nu + (\tilde{\eta} - \eta) \). Substituting \( \tilde{\nu} \) in \( \psi(\theta) \) produces the feasible estimator \( \hat{\theta}_N^* \), that is,

\[ T^{1/2}(\hat{\theta}_N^* - \theta_0) = T^{1/2}(\hat{\theta}_N^* - \theta_0) + V(\hat{\theta}_N^*) T^{-1/2} G' \Sigma^{-1} (\tilde{\eta} - \eta) + o_p(1), \]  

where use has been made of (22). Defining \( \bar{g}_t = (0, 0, h_t^{(2)}, \ldots, h_t^{(N-1)})' \) and using Lemma 3.3 it is seen that

\[ T^{-1/2} G' \Sigma^{-1} (\tilde{\eta} - \eta) = T^{-1/2} \sum_{t=1}^{T} \left( \frac{\partial g_t'}{\partial \theta} \right) \Sigma_t^{-1} (\tilde{\eta}_t - \eta_t) = C T^{1/2}(\hat{\beta}_m - \beta_0) + o_p(1), \]

where \( C = \mathcal{E}\left\{ T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} \bar{g}_t \bar{x}_t' \right\} \). Substituting back in (23) and using (21) we finally get

\[ T^{1/2}(\hat{\theta}_N^* - \theta_0) \xrightarrow{d} N\left[ 0, V(\hat{\theta}_N^*) + V(\hat{\theta}_N^*) D V(\hat{\theta}_N^*) \right], \]  

where

\[ D = C V(\hat{\beta}_m) C' + B_1 C' + C B_1'. \]  

Inspecting (19) and (24) shows that the two procedures with and without conditioning are equivalent asymptotically if \( B = D \), where \( B \) is defined in (20) and \( D \) in (25). Since \( V(\hat{\beta}_m) = B_1' V(\hat{\theta}_N^*) B_1 \) and \( B_1 = (I_k, 0)' \), a sufficient and necessary condition is that \( C = \sum_r B_r A_r \). This is evidently the case when the conditional distribution is symmetric, and then \( B = D = 0 \). Although we attempt no formal proof for the asymmetric case, it is reasonably clear that \( B = D \) holds because of the nonsingular transformation from the unconditioned to the conditioned system. This result is important because it permits using the sequential strategy and allows for simple system estimation as a final stage.
§ 10.3 Diagnostic Testing

The consequences of specification error in each of the conditional moments can be induced from those in heteroskedastic models studied in Chapter 4. First observe that we have constructed a parameterization which is crucially dependent on the specification of the conditional mean. If $\mu_t$ is misspecified this in general induces misspecification in all other equations because even when $h_t^{(r)}$ is correctly specified $\varepsilon_t^{(r)} = u_t - h_t^{(r)}$ are not the appropriate innovations. Moreover, since $h_t^{(r)}$ is a function of $h_t^{(r-1)}$, specification error in the latter propagates to the former. Misspecifying the $r$-th order moment also induces inconsistency in general in all equations of order $s \leq r$ for which $\alpha_s$ appears non trivially in $\Theta_r$. It follows that, as in most econometric exercises, a bid to improve on efficiency by imposing more structure on the problem brings with it a risk of introducing inconsistency.

The situation is similar to the heteroskedastic case:

a) misspecification in the first $N$ conditional moments results in inconsistency in general,

b) misspecification of the second $N$ conditional moments produces inefficiency and incorrect inferences unless robust estimates of the covariance matrix are used, and

c) failure to consider properly conditional moments of order higher than $2N$ prevents further gains in efficiency.

We can associate a family of consistency tests to (a) and a family of efficiency tests to (b) and (c) as we did in Chapter 5 with heteroskedastic models. Here we will develop only a particular form of consistency tests, namely those checking the coherency of the information about the same parameters in different equations, but it does not seem bold to conjecture that other consistency tests, and in particular variable addition tests, may be
constructed as \( N R^2_0 \) where \( R^2_0 \) is the uncentered coefficient of determination of the regression of \( \hat{\sigma}_t \) on \( \partial g_t / \partial \theta \) and the additional variables, in the metric of \( \hat{\Sigma}_t \). This could be seen by deriving the asymptotic distribution of the consistency and efficiency test statistics from the joint distribution of the orthogonality conditions and moment restrictions without imposing normality, rather than from the joint distribution for the score and the moment restrictions as was done in Chapter 5.

Coherency tests have the attraction of being simple to construct, easy to interpret, and fit nicely in the sequential strategy described in the previous section. Each time a new moment is explored an analysis is made of its contribution in efficiency terms. If this contribution is significant a check of its coherency with the information in lower order moments is performed. When fitting the \( r \)-th moment equation - with or without conditioning - a first estimate of \( \alpha_r \) is made available. Moving then to subsequent higher order moments produces other estimates of \( \alpha_r \) which are then checked for coherency with that from the \( r \)-th moment.

For the symmetric distribution case define

\[
q_{rs} = \tilde{\alpha}^{(r)}_r - \tilde{\alpha}^{(s)}_r, \quad s > r,
\]

where \( \tilde{\alpha}^{(s)}_r \) is the estimator of \( \alpha_r \) obtained from the conditioned \( s \)-th moment equation, \( s \geq r \). The quantity \( q_{rs} \) tends to differ from zero due only to sampling variation under the null hypothesis of a correctly specified model, and tends to a nonzero limit under the alternative of specification error. Thus again the test is a Hausman [1978] type test of the form considered by Ruud [1984]. Because of the asymptotic independence of \( \tilde{\alpha}^{(r)}_r \) and \( \tilde{\alpha}^{(s)}_r \),

\[
T^{1/2} q_{rs} \overset{d}{\rightarrow} N \left[ 0, V [\tilde{\alpha}^{(r)}_r] + V [\tilde{\alpha}^{(s)}_r] \right],
\]

and the test-statistic is
\[ \tau_{rs} = T q \hat{V}(q)^{-1} q \xrightarrow{d} \chi^2 \]

under the null hypothesis, with the noncentral \( \chi^2 \) providing the power of the test under local alternatives. Because both terms in the variance in (26) are positive definite, so is \( V(q_{rs}) \).

Joint tests of several equations can also be constructed. Define

\[ q_r = (q_{r+1}, \ldots, q_{rN})' = (1 \otimes \tilde{\alpha}^{(r)}) - a_r, \]

where \( a_r = (\tilde{\alpha}^{(r+1)}, \ldots, \tilde{\alpha}^{(N)})' \) and \( 1 \) is a vector of ones of dimension \( N - r \). Then under the null,

\[ T^{1/2} q_r \xrightarrow{d} N[0, 1' \otimes \text{Diag}(V[\tilde{\alpha}^{(r)}], \ldots, V[\tilde{\alpha}^{(N)}])], \]

and so

\[ \tau_r = T q_r \hat{V}(q_r)^{-1} q_r \xrightarrow{d} \chi^{(N-r)k_r}, \]

while locally under the alternative the distribution in \( q_r \) has nonzero mean in general and \( \tau_r \) follows a noncentral \( \chi^2 \) with the appropriate noncentrality parameter. The positive definiteness of \( V(q_r) \) is guaranteed by similar argument to that used for \( V(q_{rs}) \).

We can also form joint tests of different parameter subvectors. For example, \((q_{1r'}, q_{2r'}, \ldots, q_{r-1r'})' \) produces a test of the coherency of all information in the \( r \)-th moment with the previous \( r - 1 \) moments. The covariance matrix is obtained using

\[
\text{cov} [T^{1/2} q_{rs}, T^{1/2} q_{rs'}] = \text{cov} [\tilde{\alpha}^{(r)}, \tilde{\alpha}^{(r')}] - \text{cov} [\tilde{\alpha}^{(r)}, \tilde{\alpha}^{(s)}] - \text{cov} [\tilde{\alpha}^{(s)}, \tilde{\alpha}^{(r')}] + \text{cov} [\tilde{\alpha}^{(s)}, \tilde{\alpha}^{(s')}]
\]

\[ = \delta_{rr'} V[\tilde{\alpha}^{(r)}] - \delta_{rs'} \text{cov} [\tilde{\alpha}^{(r)}, \tilde{\alpha}^{(s)}] - \delta_{sr'} \text{cov} [\tilde{\alpha}^{(s)}, \tilde{\alpha}^{(r)}]
\]

\[ - \delta_{ss'} \text{cov} [\tilde{\alpha}^{(s)}, \tilde{\alpha}^{(s)}], \]
where $\delta_{ij}$ is the Kronecker delta. With this expression we can form the covariance matrix for any coherency test and compute any joint test-statistic. The possibility of inconsistent tests exists as discussed in § 5.1.

In the non-symmetric case with $N > 2$ so that the coefficient estimators are not asymptotically independent, the covariance matrices of the $q$ statistics must take this into account and there is now the possibility that these not be positive definite, so that resort to a generalized inverse may be required. The invariance of Hausman tests to choice of generalized inverse has been established by Holly [1982]. Apart from this necessary adjustment, the details are similar to the symmetric case.
CHAPTER 11

CONCLUSIONS

We are now faced with the task of drawing some concluding remarks from the contents of the Thesis, and of pointing at some of the many questions that remain unanswered. It is always easier — and shorter — to account for what has been done than for what remains to be done. Therefore, we will start by outlining what we consider to be the main developments, and end with a review of some of the topics that we identify as requiring further research.

The separate estimation of the variance equation is certainly not a new idea. Amemiya [1977] considered the GLS estimation of such equation in simple heteroskedasticity models, Jobson and Fuller [1980] studied the properties of LS estimators in the more complex models in which the conditional variance depends on \( \beta \), and the two-stage estimators for \( \alpha \) are a common use, see for example Engle [1982a], or Pagan [1984a]. However, Chapter 3 presents the first systematic and general treatment of the identification and estimation of the variance equation. The most important result is the separation of the problem of estimating heteroskedastic models into two generalized least squares regressions.

A conceptual advantage of such an approach is that it allows the simple extension of most procedures for diagnostic testing available for generalized regression models. This has been shown to be the case in Chapters 5 and 6, where we have started from the basic principle of diagnosing the model by analyzing its residuals, and have derived the classes of tests we call the consistency and the efficiency tests. By using the principle of conditional moment testing proposed by Newey [1985a, 1985b] and Tauchen [1985], these
classes are general enough to include and extend to heteroskedastic settings the variable addition (LM) tests of Breusch and Pagan [1980], Engle [1982b, 1984] and Pagan [1984a], the estimator difference tests of Hausman [1978] and White [1980a], the data transformation tests of Plosser et al. [1982] and Breusch and Godfrey [1986], the RESET tests of Ramsey [1969] and Ramsey and Schmidt [1976], and the tests against non-normality of Jarque and Bera [1980]. In testing the model against specific alternatives we have considered some of the most important departures analyzed in applied work, and have produced a basic framework for the construction of tests in other directions and for the construction of multidirectional tests such as those suggested by Bera and Jarque [1982].

Another important implication of looking at estimation by decomposition of the model into the two-equation system is that it clearly separates the information arising from each conditional moment. This is useful for model evaluation, as it allows a simple sequential strategy to model building. The mean equation is estimated in the first place, and the specification search proceeds with the variance equation. The sensitivity of the mean estimators to alternative variance specifications can be analyzed from the difference between OLS and GLS estimators, the relative contributions to efficiency can be measured, and the coherency of the information in the two moments can be tested. This produces valuable information as to the potential gains in efficiency obtained from modelling the heteroskedasticity, and the severeness of the inherent risk of introducing inconsistency from imposing more structure on the model.

In the context of estimation, the role of the QMLE has been made clear and it is comparable to that of the Gauss-Markov OLS estimator in the classical linear model: it retains optimal properties with respect to information in the first two moments provided that the first four moments are correctly specified. The QMLE can be made robust to departures in the kurtosis of the distribution
by using White's [1980b] covariance matrix in the variance equation just as the
OLS estimator can be made robust against departures from homoskedasticity.
Furthermore, the consequences of specification error in the different moments
of the conditional distribution for the QMLE have been analyzed, and this calls
for careful evaluation of the model and at the same time provides the basic sect
of symptoms that can be used to devise proper diagnostic tools.

We have produced some Monte Carlo evidence to assess the asymptotic
approximations for the estimation and testing of the univariate heteroskedastic
model in small to moderate samples, and the results can be regarded as
satisfactory in general, though of course the evidence produced is limited.

Along with the general developments of Chapters 2 to 6 we have analyzed
some special cases, and in particular we have unveiled some interesting
properties of the ARCH class of models. The identifiability of the full
parameter vector in the variance equation and the robustness of the QMLE of β
to classes of variance misspecifications add to the already long list of
attractions of the ARCH model, and reinforce Engle's [1982a, p. 990] argument
that ARCH may act as an approximation for other types of heteroskedasticity or
model misspecification. It is our conviction that in the absence of any a priori
idea about the variance of the process, the minimum that applied researchers
working with time series data should consider in relation to heteroskedasticity
should be the GARCH model. A more desirable treatment would be to have
some theoretical propositions with which the steady state of the variance must
be compatible, and to let the GARCH components provide the dynamics of the
second moment. An interesting feature of the ARCH class of models is that its
robustness properties may render general tests of specification inconsistent in
many situations, and this stresses the need to test the model both in general
and specific directions to assess model adequacy. In particular, it may be of
interest to test for lagged squared values of the dependent variable in the
conditional variance in view of the results of Weiss [1984], and also to test for
functions of the regressors, either directly or embedded in functions of the conditional mean.

The results obtained for the basic heteroskedastic model have been generalized in several directions. The introduction of parametric risk measures has been studied in Chapter 7 when it relates either to the conditional moments of the dependent or the conditioning variables. In the latter case the framework of two-stage estimation of Pagan (1984b, 1986) has been extended partially to models with heteroskedasticity, and in the former we have provided a framework for models such as the ARCH-M of Engle et al (1987) which allows for general tests of the specification and, in the case of given alternatives, by using the residuals of the two equations, incorporates more structure specific to the problem than the asymptotically equivalent procedure of regressing a constant on the score. We have also analyzed in Chapter 8 the possible implications of allowing the coefficients of the mean equation to vary, and concluded that the theoretical framework for inference developed for the basic model extends to this situation. The problems of fitting these models are the numerical aspects and the requirement of very large samples to obtain meaningful results. The main development of Chapter 8 is the derivation of a procedure to test for superexogeneity as a means for assessing the Lucas (1976) critique to the use of econometric models to analyze economic policy. In Chapter 9, most of the results have been generalized to the case of multivariate distributions, and the strategy to model the first two moments has been extended to produce a sequential search for information in higher order moments in Chapter 10.

For reasons of time and space, we have left many topics for further research. Although the proofs of the asymptotic properties of estimators and test-statistics are given in an environment which is more general than usually found in the literature on heteroskedastic models, the restriction imposed by the existence of moments of order fourth contrasts with empirical findings
using the ARCH model, and which have motivated Engle and Bollerslev [1986] to propose variance integration. Similarly, we have not touched the topic of cointegration in the mean equation (Engle and Granger [1987]), or in variance equations (Engle [1987]), and the extension of these and other developments in integrated systems (such as Phillips [1987]) constitute important lines for future work. It is to be noticed that in the simulation experiments we have included an ARCH model which does not possess fourth moments, and the performance of the statistics in this case has not been qualitatively different from the other models which do possess at least fourth order moments.

We have concentrated on parametric approaches to inference. We have pointed out, however, the possibilities of semi-parametric methods and these would indeed fit very nicely into the suggested sequential strategy of modelling heteroskedasticity. Using estimators such as those proposed by Carroll [1982] and Robinson [1987], we would be able to treat the two moments in almost total isolation, and this would make the specification search more robust. However, many questions remain to be answered about the assumptions required to use such a procedure in the variance equation, and even in the mean equation when the heteroskedasticity has a GARCH form.

Another aspect of the variance equation to which we have not given much attention has been the positivity restrictions. We have worked on the assumption that the conditional variance can be reparameterized in such a way that the positivity restrictions are implicitly incorporated in the parameter space, as happens with all the special cases analyzed. However, we think that this problem deserves further attention if only for its possible numerical implications in small and moderate samples.

One topic that has not received mention in the Thesis is the estimation of the covariance matrices of the estimators. These take familiar forms and are simple to calculate. However, in some Monte Carlo experiments which we
have not reported, we found that the distribution of the coherency test-statistics can change dramatically when evaluating the covariance matrices at different estimators, thus suggesting a high degree of sensitivity. There is some intuition as to the cause of this problem because the weighted sums by the inverse of the conditional variances may become very sensitive to a few small values of the latter quantities, but this also deserves further attention. This problem is also linked to the information measures suggested in Chapter 3 for the contributions to efficiency of each moment, which are functions of the covariance matrices directly. For these measures it remains to develop approximations to their asymptotic distributions and to further enquire into their properties in small to moderate samples.

Bollerslev [1986] proposed the use of the autocorrelation and partial autocorrelation functions to identify the orders of GARCH processes, and in Chapters 2 and 3 we have suggested that the Hannan-Rissanen [1982] procedure might constitute an alternative. The formal validity of both procedures depends on the existence of fourth order moments, but Bollerslev has used the procedure with success in a context where the estimated model does not conform to such restrictions. It remains to be seen whether the empirical application of the Hannan-Rissanen procedure can produce similarly good results.

The Monte Carlo evidence presented is in no way exhaustive and further evidence is required, as well as analytical results, for a more compete assessment of the properties of estimators and tests in small samples. The analytical results may be hard to derive, but some progress might be made by analyzing the conditions under which the invariance properties of exact distributions put forward by Breusch [1980] for the simple heteroskedasticity model carry through to more complicated settings, and in particular to the ARCH model in which we think there are grounds for expecting good results. Of course linearity of the conditional mean seems to be a first requirement
because otherwise the covariance matrix of the estimators immediately depends on the parameters. An alternative approach to exact inference in the ARCH model has been put forward by Geweke [1986c], using Monte Carlo integration techniques to explore the likelihood function over the parameter space. This method has the advantage of avoiding the problem of the positivity restrictions in the variance parameters, but has the disadvantage that we cannot produce from it a general theory for inference and each case has to be dealt with in particular.

The testing procedures of Chapters 5 and 6 have been developed under the null hypothesis of conditional normality. It is evident that the coherency tests with the simple substitution of the relevant kurtosis functions apply equally well in non-normal but symmetric environments. It also seems clear from the results of Newey [1985b] that the extension of consistency and efficiency tests to the more general framework of GMM estimation without normality is reasonably straightforward, but proofs are required.

The efficiency tests considered in the simulations referred to third and fourth moments only, and it would be interesting to analyze their performance in moments of higher order. For example in the presence of a DGP possessing Tukey's symmetric λ distribution which has the first four moments equal to those of the normal distribution (Joiner and Rosenblatt [1971]). In the multivariate models, we have not produced an LM test against the multivariate Pearson family, though this problem does not seem hard to solve.

We have concentrated mainly on diagnostic tests, either producing statistics which do not require information external to the model, or using the LM principle to test for specific departures. The alternative Wald and likelihood ratio tests, and testing parametric restrictions in general, have received little mention in the Thesis. Calculation of the Wald statistics and likelihood ratio is reasonably straightforward, and the general form derived for
the LM test against variable additions can be used to test the parametric restrictions, though the model may require reparameterization for this purpose. The treatment of the asymptotic theory given in Engle [1984] is sufficiently general for this purpose, and our assumptions ensure the asymptotic equivalence of the three principles. The relative small sample performance of the different tests, however, is a topic on which it may be worth conducting an extensive simulation study, and this study might also include the alternative forms of the consistency and coherency tests.

Testing non-nested hypotheses has also gone without mention, but some of the testing principles such as the J test of Davidson and MacKinnon [1981] are immediately applicable in the context of the two-equation system, and may be used to test non-nested hypotheses in the conditional mean, the conditional variance, or both. We have only produced one-step prediction error tests, and generalizing these tests to multiperiod dynamic forecasts as in Pagan and Nicholls [1984] is a topic for further work.

The discussion of varying coefficient models is rather informal and requires a more rigorous treatment. The procedure for testing superexogeneity based on evolving coefficients appears to be too complicated computationally to provide a practical test. Restricting the coefficient variation to random coefficients is better in this sense, but it loses the dynamic dimension of the learning procedure. Thus it appears that more work is needed to produce a test which may be reasonably simple yet incorporates a rich structure for the potential relationship between the coefficients of the conditional and conditioning models.

For the multivariate heteroskedastic model we have extended the asymptotic theory from the univariate case, but we have paid very little attention to some of the main problems in the application of such models, namely achieving a parsimonious parameterization without exercising a large
degree of arbitrariness, and enforcing the positive definiteness of the conditional covariance matrix at all time periods. Some proposals towards these ends have been put forward by Diebold and Nerlove [1986] and Engle [1987].

Simultaneous equations heteroskedastic models have not been treated in the Thesis. One of the reasons is that we have not been able to establish a framework in correspondence with the rest of the models considered in the Thesis, that is, one that allows the separation of the system of structural equations defining implicitly the conditional mean of the process, and a system of equations that represents the structural covariance matrix. The main problem we have found here is that, although the likelihood function under conditional normality appears amenable to a local factorization, the structural innovations do not obey a result like that of Lemmas 3.2.1 and 9.3.1 because the presence of simultaneity prevents convergence at the required speed.

For the sequential search for information in higher moments we need to develop more diagnostic procedures, mainly those of consistency and efficiency tests. These, however, would be a by-product of the extension of the families of consistency and efficiency tests in the basic model to non-normal maintained models to which we have already referred above. The main problem here appears to be that of finding parameterizations with empirical plausibility, and this is something which cannot be solved without a serious attempt to apply the procedure to real data.

There are many more aspects which have not received attention, and it is impossible to attempt a complete list. As a final word we must acknowledge the lack of empirical content in the Thesis, and thus suggest that the main topic for further research lies in the field of application of the theoretical aspects of econometric inference with heteroskedastic models which are contained in this work.
REFERENCES


Amemiya, T. [1973], Regression Analysis when the variance of the dependent variable is proportional to the square of its expectation, *Journal of the American Statistical Association* 68, 928-934.


Breusch, T. and Schmidt, P. [1986], Alternative forms of the Wald test: how long is a piece of string?, Manuscript, *ANU and Michigan State University* manuscript.


Cameron, A.C. and Trivedi, P.K. [1986], Econometric models based on count data: comparisons and applications of some estimators and tests, *Journal of Applied Econometrics* 1, 29-53.


Diebold, F.X. [1986b], Testing for serial correlation in the presence of ARCH, *University of Pennsylvania* manuscript.


Durbin, J. [1970], Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables, *Econometrica* 38, 410-421.


Engle, R.F. [1987], Multivariate ARCH with factor structures – Cointegration in variance, *UCSD* manuscript.


Engle,R.F. and Watson,M. [1985], Applications of Kalman filtering in econometrics, paper presented to the Fifth World Congress of the Econometric Society.


Geweke, J. [1986c], Exact inference in models with autoregressive conditional heteroscedasticity, *Duke University* manuscript.


Godfrey, L.G. [1978], Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables, *Econometrica* 46, 1303-1310.


Hall, P. and Carroll, R.J. [1987], Variance function estimation in regression: the effect of estimating the mean, *ANU* manuscript.


Harvey, A.C. [1984], Dynamic models, the prediction error decomposition and state space, Chapter 3 in Hendry,D.F. and Wallis,K.F. Eds.: *Econometrics and Quantitative Economics*, Basil Blackwell.

Harvey, A.C. [1985]: Applications of the Kalman filter in econometrics, paper presented to the Fifth World Congress of the Econometric Society.


Jobson, J.D. and Fuller, W.A. [1980], Least squares estimation when the
covariance matrix and parameter vector are functionally related, *Journal of the American Statistical Association* 75, 176-181.

Joiner, B.L. and Rosenblatt, J.R. [1971], Some properties of the range in


Pagan, A.R. and Sabau, H. [1987b], Consistency tests for heteroskedastic and risk models, *University of Rochester and ANU* manuscript.


Salkever, D.S. [1976], The use of dummy variables to compute predictions, prediction errors, and confidence intervals, *Journal of Econometrics* 4, 393-397.


Solo, V. [1983], Varying parameter regression modelling through covariances, *Harvard University* mimeo.


Zellner, A. [1971], *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons.