Essays on missing data problems: MSL estimation in the analysis of censored data and 'doubly robust' estimation in the analysis of treatment effects

Sanghyeok Lee

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

August 2019

Declaration

This thesis consists of four research papers and is within the 100,000 word limit by the ANU. Chapters 2 and 3 are joint work with my primary supervisor Dr Tue Gørgens. Chapters 4 and 5 are joint work with my supervisor Professor Myoung-jae Lee. My contribution to each of Chapters 2 and 3 was 75 percent and my contribution to each of Chapters 4 and 5 was 50 percent. With these exceptions and where otherwise acknowledged I certify that this thesis is my own work.

Sanghyeok Lee	Date

To my parents, Jangwhan Lee and Yoonsik Park

Acknowledgements

Foremost, I would like to express my deep and sincere gratitude to my supervisors Dr Tue Gørgens (primary), Professor Myoung-jae Lee, Professor Robert Breunig, and the unofficial fourth member of supervisory panel Professor Xin Meng, for their continuous support of my PhD study and research. Especially, without the guidance and support of my primary supervisor, my academic journey in Australia would have been much poorer. Certainly, I cannot thank him enough. Also, I thank many seminar participants at the Research School of Economics (Australian National University), the SEF Applied Econometric Workshop (Victoria University of Wellington), the Econometric Society Australasian Meeting (University of Technology Sydney), the Australian PhD Conference (University of New South Wales), the Australasian Workshop on Econometrics and Health Economics for their helpful comments. Last but not least, I would like to thank my parents for fully understanding and supporting my career choice.

Abstract

In Chapter 2, we consider estimation of dynamic models of recurrent events (event histories) in continuous time using censored data. We develop maximum simulated likelihood estimators where missing data are integrated out using Monte Carlo and importance sampling methods. We allow for random effects and integrate out the unobserved heterogeneity using a quadrature rule. In Monte Carlo experiments, we find that maximum simulated likelihood estimation is practically feasible and performs better than both listwise deletion and auxiliary modelling of initial conditions. In an empirical application, we study ischaemic heart disease events for male Maoris in New Zealand.

Chapter 3 describes how the risk of experiencing heart attacks varies across gender and ethnicity in New Zealand. We analyse administrative data and estimate dynamic hazard models using maximum simulated likelihood methods to deal with left-censoring. The models allow risk to vary with age, previous heart attack history, and unobserved individual heterogeneity. We find that the risk of subsequent events is far higher than the risk of the first event, and particularly high within 1 year after an event. In most cases, male Maoris have the highest risk, followed by female Maoris, then male Europeans, while female Europeans have the lowest risk.

Differently from the well-known propensity score (PS), the lesser known '*prognostic score (PGS)*' balances the potential untreated response. Chapter 4 shows that 'double robustness' can be achieved by controlling both PS and PGS in various ways in a method-blind manner.

In Chapter 5, we compare various treatment effect estimators through an extensive simulation study using 64 designs and two empirical examples mimicking experiments. In total, we examine 24 estimators based on matching, weighting, double robustness, regression imputation/adjustment, 'complete pairing', and 'propensity-score residual'.

Our results show that, contrary to the common perception, doubly robust estimators are not necessarily the best. In fact, our findings recommend a couple of non-doubly-robust estimators, with a simple propensity-score-residual-based estimator being the nearly dominant best estimator.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Missing data problems are omnipresent in data analysis, even nowadays when available data abound. Data are missing for natural, administrative, economic or behavioural reasons. If the way data are missing is non-systematic and the proportion of missing data is not considerable, ignoring missing data in analysis can be a first resort. However, if the proportion of missing data is considerable, ignoring them in analysis can be too costly although maybe not harmful. More importantly, if data are missing in a systematic manner, analysing only complete cases and ignoring missing data could provide misleading results. In many cases, therefore, handling these missing data is of great importance in the analysis of available data.

There is a large literature on the problem of missing data. As of June 2019, the online bibliographic database Scopus lists 14,782 journal articles with 'missing data' in their title, keywords, or abstract (and written in English). Figure 1.1 shows the trend by using the same list of journal articles. It is clear that the problem of missing data has been discussed over several decades and shows no sign of being phased out.

Rubin [1976] formally provides the weakest conditions for when missing data can be ignored. If these conditions hold, ignoring the missing data process is always innocuous in the inference of the distribution of the data. Rubin's formal definitions correspond to the following statements with notation such that θ is the parameter of the data generating process and ϕ is the parameter of the missing data process.

> 'The missing data are missing at random if for each possible values of the parameter ϕ, the conditional probability of the observed pattern of missing

data, given the missing data and the value of the observed data, is the same for all possible values of the missing data. The observed data are at random if for each possible value of the missing data and the parameter ϕ, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, is the same for all possible values of the observed data'.

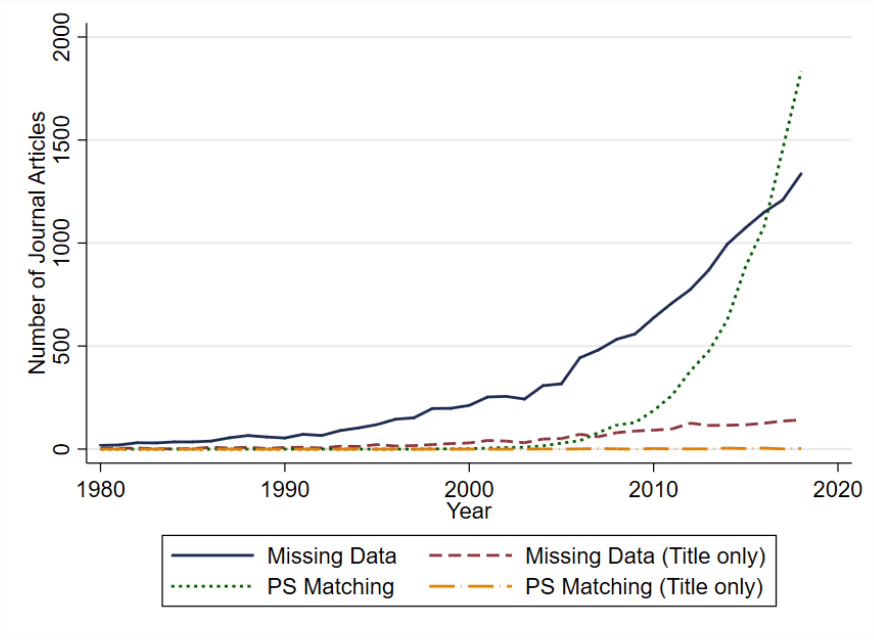

Figure 1.1: Search results from scopus.com using 'missing data'

PS Matching means propensity score matching and is provided for comparison. The figures are the number of journal articles including missing data or PS matching in title, keywords, abstract or only in title (noted in the legend) and written in English.
Data source: publication data searched at online database www.scopus.com on June 2019

This thesis considers the cases where the distribution of the data is independent of the missing data process. In particular, this thesis focuses on two important contexts of missing data problems: censoring in the analysis of event history data and 'missing at random' in the analysis of treatment effects.

This thesis considers the case where the distribution of the data is independent of the missing data process but missing data occur in a systematic manner. In particular, this thesis focuses on two important contexts of missing data problems: censoring in the analysis of event history data and ‘missing at random’ in the analysis of treatment effects. Systematically missing data are handled by various modelling in the former and by controlling observables in the latter.

First, censoring means a situation that in any periods of time an individual is at risk of experiencing an event but is not under observation. If these periods are before the start of observation and after the end of observation, the event histories are left- and right-censored, respectively. If events are independent of each other, a maximum likelihood approach where the likelihood function computed from available data is straightforward and provides consistent estimates under standard assumptions in the literature. However, when dynamic models are estimated with censored data, especially left-censored data, a maximum likelihood approach is not straightforward since the likelihood function is not analytically computable. While analysing a subsample without missing data or maximising an approximate likelihood function using a reduced-form model of missing data is suggested in the literature, consistent estimates come at the cost of efficiency or additional assumptions. In this thesis, we consider maximum simulated likelihood approach to overcome left-censoring problem.

Second, there are two potential outcomes when treatments are binary. Naturally, only one of treated and untreated outcomes is observed, which is the problem of missing data. Commonly, ‘missing at random’ (MAR), also known as selection-on-observables, is assumed to hold in the analysis of treatment effects. The MAR assumption implies that the process of data being missing depends on observables. Under the MAR assumption, consistent estimates of treatment effects can be obtained by comparing treated and untreated groups after controlling observables. A well-known balancing score for controlling observables is propensity score and less known one is prognostic score. Although it is possible in a nonparametric manner, often parametrically estimated balancing scores are controlled. If misspecified, however, these parametric balancing scores result in inconsistent estimates. In this thesis, we consider estimators doubly robust to possible misspecification.

This thesis includes four self-contained research papers. Chapters 2 and 3 (joint work with Dr Tue Gørgens) concern estimation of dynamic models of recurrent events using censored data, while Chapters 4 and 5 (joint work with Professor Myoung-jae Lee) concern 'doubly robust' estimation by controlling both propensity score and prognostic score at the same time.

In Chapter 2, we consider a general framework of censoring where multiple periods are under observation or not under observation in an alternating fashion and event histories are available only on periods under observation. We consider estimation of dynamic models of recurrent events using censored data. In general, consistent estimates can easily be obtained by maximising the likelihood if complete data are available. Due to missing prior histories, however, the likelihood is not analytically computable. We suggest using the maximum simulated likelihood (MSL) method where missing history data are integrated out in the likelihood function using simulation techniques. In the MSL method, we approximate the exact likelihood of observed histories via simulation and maximise the arbitrarily accurate simulated likelihood function. Chapter 2 confirms that our proposed method is feasible in the context of continuous-time dynamic models of recurrent events and is a substantial efficiency improvement on other alternatives. In an empirical application, we estimate a dynamic model of ischaemic heart disease, using New Zealand data on hospital admissions and deaths. Consistent with Monte Carlo results, we find that a substantial efficiency gain can be achieved via MSL methods although the size of the gain varies.

In Chapter 3, we consider dynamic models of recurrent heart attack event in New Zealand. The research question in this chapter is how heart attack risk varies with age and prior history and how patterns in the risk differ across gender and ethnic groups. We compare four groups: male and female people of Maori and European descent. We use high-quality New Zealand administrative data on hospital admissions and deaths and estimate dynamic models of heart attack using the MSL method developed in

Chapter 2. Our main finding is that prior history affects the risk pattern of heart attack through changes in age dependence and dynamic effect as well as changes in the basic risk level of heart attack. Overall, experiencing a heart attack increases the risk of subsequent heart attacks, and it is particularly high within the first year after a heart attack. In most cases, male Maoris have the highest risk of the first heart attack, followed by female Maoris and then male Europeans, while female Europeans have the lowest risk.

In Chapter 4, we consider binary treatment and control observables to estimate the treatment effect. In treatment effect analysis, the propensity score measures the probability of being treated given observables and controlling the propensity score is a very popular way of controlling for observed heterogeneity. In contrast, the prognostic score measures baseline potential outcomes – the untreated potential outcome in a binary treatment. In this chapter, we propose 'doubly robust estimation' in a method-blind manner. Double robustness means that estimation is consistent if either the propensity or the prognostic score is correctly specified. In the literature, doubly robust estimation is based on weighting. We theoretically prove that doubly robust estimation can be achieved by controlling both propensity score and prognostic score, regardless of the way this is done.

In Chapter 5, we provide comprehensive Monte Carlo simulation where 26 estimators, only some of which are doubly robust, are compared and two empirical applications. In particular, we discover that estimators by controlling both propensity score and prognostic score are doubly robust and that doubly robust estimators are not necessarily better than estimators with propensity score or prognostic score alone controlled.

This thesis concludes with a brief summary of the four research papers and a discussion of the future researches.

Chapter 2

Estimation of dynamic models of recurrent events with censored data

2.1 Introduction

Data censoring is a pervasive problem in the analysis of the occurrence and timing of events. Often the observation process is such that some individuals are not under observation continuously during the time they are at risk, and therefore some events may be missing in the data available for analysis. For example, the observation period may begin and end at fixed calendar times and only events that occur within this window are available for analysis. The event histories are said to be left-censored or right-censored if events before the start or after the end of the observation period are missing, respectively. In some longitudinal surveys, participants provide information annually about events that have occurred in the previous year, and participants who skip an interview will have a gap in their recorded event histories. The event histories are said to be middle-censored if there is a gap in the middle of the recorded event histories.

In practice, event history models are estimated by the method of maximum likelihood (ML). Usually it is assumed that the observation process is independent of the event process (and the former is not modelled). In this case, it is straightforward to include right-censored event histories, and gaps can be handled by artificially right-censoring the histories at the start of the gap. If there are not too many gaps, the data loss may be acceptable. However, left-censoring remains a difficult problem in most applications, especially in dynamic models where prior events affect the timing of subsequent events. Since consistent estimates can be obtained from the non-left-censored histories, a common solution is simply to drop all left-censored histories from the analysis. For example, Doiron and Gørgens [2008] and Cockx and Picchio [2012, 2013] studied transitions between labour force states and avoided the left-censoring issue by focusing on young people who first entered the labour force during the observation period (so their initial labour market outcomes are observed). Similarly, Bhuller, Brinch, and Königs [2017] studied dynamic aspects of the receipt of welfare benefits, and selected a sample of individuals who turned 18 and thus became eligible for the first time during the study period. Dropping left-censored histories from the analysis comes at the cost of a smaller sample size. For example, by restricting their sample to school leavers Doiron and Gørgens [2008] used only one third of the total sample.

The problem of left-censoring in event history analysis is related to the well-known problem of initial conditions in discrete-time dynamic panel data models of binary responses or other limited dependent variables. In these models, the ‘structural’ equation involves lagged dependent variables whose coefficients (or partial effects) are parameters of interest. The dilemma is that the structural equation cannot be evaluated for the initial observations since lagged information is not available, but conditioning on the initial observations leads to inconsistent estimates in the presence of unobserved heterogeneity. In the context of a first-order Markov model of binary responses, Heckman [1981] proposed to supplement the structural model with an approximate reduced-form model for the initial conditions, based on exogenous information available for the initial periods, a flexible specification of the influence of unobserved heterogeneity, and imposing no parameter restrictions across submodels. Heckman’s method has been applied for example in continuous-time duration analysis by Gritz [1993] and in discrete-time duration analysis by Ham and LaLonde [1996], Cappellari, Dorsett, and Haile [2010], and Gørgens and Hyslop [2019].

In this chapter we consider estimation of continuous-time dynamic event history models with censored data by maximising a simulated likelihood function using all available data. The likelihood function is specified in terms of observed and unobserved events, and unobserved events are then 'integrated out' using Monte Carlo and importance sampling methods. We allow for unobserved heterogeneity in the form of so-called random effects and integrate out unobserved heterogeneity using a Gaussian quadrature rule. Our maximum simulated likelihood (MSL) estimator uses all available data and does not involve additional functional-form assumptions or additional ad hoc parameters. The method is applicable when the times during which individuals are at risk of experiencing events are known.[1] For simplicity, we focus on recurrent events. This class of models covers a wide range of applications: purchases of specific goods or services, health events such as heart attacks or dental fillings, child births, time between earth quakes or geyser eruptions, etc.

The method of maximum simulated likelihood estimation has been successfully applied in other contexts. For example, Lerman and Manski [1981] were the first econometricians to consider the frequency simulator of (multinomial probit) choice probabilities. Keane [1994] studied MSL estimation of binary response models with serially correlated errors, with the multinomial probit model as the leading case. McCulloch [1997] considered latent class (mixture) models. Kamionka [1998] sketched a general framework for continuous-time transition models and provided some simulation results for estimating continuous-time time-homogeneous Markov processes using data measured on a discrete time scale. Keane and Sauer [2010] developed a method for estimating discrete-time dynamic panel data models with unobserved endogenous state variables. Their method assumed that the dependent variables are measured with error. Some authors have compared MSL estimation with estimation using the

[1] In a study of transitions into and out of female headship, Moffitt and Rendall [1995] were able to integrate out unobserved events analytically because the distribution of missing data was discrete in their model.

EM algorithm and found that the latter performed better. Brinch [2012] argued that the negative assessment of MSL estimation among some authors is at least partly due to suboptimal choices made in the implementation.

The MSL approach has both advantages and disadvantages over the alternatives. As mentioned above, dropping left-censored histories and middle-censored histories from the analysis (listwise deletion) makes for easy ML estimation but can be very costly in terms of sample size. Specifying auxiliary models for the distribution of the initial conditions in terms of unobserved heterogeneity also allow for standard ML estimation, but specification error potentially affects the bias and consistency of the estimates and the additional parameters lead to a loss of degrees of freedom. The MSL approach is expected to have higher efficiency, because the full data set can be used and because no auxiliary parameters are involved. By increasing the number of simulations, MSL estimates can be made arbitrarily close to the exact ML estimates. Since ML estimation is asymptotically efficient, MSL estimates can also be asymptotically efficient.

A potential disadvantage of the MSL approach is computational difficulties. First, numerical integration in high dimensions is known to be difficult, whether by quadrature rules or Monte Carlo methods. In practice, limits on computing capacity may restrict the level of accuracy that can be achieved within reasonable time. Second, when the integration is carried out using Monte Carlo methods the simulated likelihood function is discontinuous, which causes trouble for standard maximisation algorithms such as Newton's method. However, importance sampling methods can be used to smooth the simulated likelihood function (see e.g. Gouriéroux and Monfort, 1991).

The present study contributes to the literature by showing how MSL estimation can be applied in the context of dynamic models of recurrent events in continuous time with censored data. We provide Monte Carlo evidence to show that MSL estimation is practically feasible, and we confirm that MSL estimation can provide substantial efficiency gains over listwise deletion and Heckman's approximate reduced-form modelling. Finally, we provide an empirical study of ischaemic heart disease events for male

Maoris in New Zealand. The application shows that MSL estimation can help to deal with a large (63%) left-censoring problem, and that MSL estimators can have much smaller standard errors than alternative estimators.

The chapter is organised as follows. Section 2.2 sets up the notation and discusses maximum likelihood estimation. Section 2.3 presents the results of our Monte Carlo experiments. Section 2.4 provides our empirical application. Section 2.5 concludes.

2.2 Maximum likelihood estimation

2.2.1 The likelihood function

When analysing censored data, it is necessary to distinguish between the underlying event process and the observation process. For example, the statistics literature talks about time at risk and time under observation. Let time be partitioned into j_i periods, $(c_{ij-1}, c_{ij}]$ for $j = 1, 2, \ldots, j_i$, such that c_{i0} is the time individual i becomes at risk, c_{ij_i} is the last time individual i is both at risk and under observation, and the individual is alternatingly either under observation or not during each period. Thus, individuals are either under observation in all odd periods or in all even periods. Analysis time is defined by normalising $c_{i0} = 0$.

The interaction between the event process and the observation process necessitates notation which number event times within observation periods. Hence, let k_{ij} denote the number of (observed or unobserved) events during individual i's period j, and let b_{ijk} for $k = 1, \ldots, k_{ij}$ denote event times within period j. For convenience, define the vector $\boldsymbol{b}_{ij} = (b_{ijk_{ij}}, \ldots, b_{ij1})$; if $k_{ij} = 0$ then $\boldsymbol{b}_{ij}$ denotes a zero-dimensional vector. To simplify certain expressions, define also b_{ij0} by setting $b_{ij0} = c_{ij-1}$. We assume that the event process and the observation process are independent. We postpone the discussion of observed and unobserved heterogeneity until later.[2]

In general, the likelihood of an event at any given time may depend on the history of events prior to that time. Let $s_i(t)$ denote all individual i's history at time t. That

[2] See Figure 2.1 for examples of event history data.

is, $s_i(t)$ includes all event times until and including t, the fact that no events occurred between the most recent event and time t, and the observation period boundaries. Let $h(t|s(t'),\theta)$ for $t > t'$ denote the conditional hazard function for events evaluated at time t given the history until time t', $s(t')$, where θ is the unknown parameter vector to be estimated. Also let $H(t|s(t'),\theta)$ for $t > t'$ denote the associated value of the cumulative hazard function from time t' until time t. That is, H is defined by $H(t|s(t'),\theta) = \int_{t'}^{t} h(y|s(t'),\theta)\,dy$. Furthermore, let $f(t|s(t'),\theta)$ denote the conditional event density at t given the history $s(t')$, and let F denote the corresponding cumulative distribution function. Then we have the well-known result (see e.g. Lancaster, 1990) that

$$f(t|s(t'),\theta) = h(t|s(t'),\theta)\exp\Big(-H(t|s(t'),\theta)\Big), \quad t > t'. \tag{2.1}$$

Here the exponential term on the right-hand side captures the non-occurrence of events during $(t', t]$. Finally, let g_j be the conditional joint density of events during period j given previous events. Using $\boldsymbol{b}_j$ without subscript i to denote a generic vector of event times in period j and using k_j for the corresponding number of events, we have

$$g_j(\boldsymbol{b}_j|\boldsymbol{b}_{j-1},\ldots,\boldsymbol{b}_1,\theta) = \left(\prod_{k=1}^{k_j} f(b_{jk}|s(b_{jk-1}),\theta)\right)\exp\Big(-H(c_j|s(b_{jk_j}),\theta)\Big). \tag{2.2}$$

The exponential term on the right-hand side represents the fact that no events occurred during $(b_{jk_j}, c_j]$ if $k_j > 0$ or during $(c_{j-1}, c_j]$ if $k_j = 0$. (Recall that we have defined $b_{j0} = c_{j-1}$.) By convention the product of the sequence on the right-hand side of (2.2) is defined to be 1 if $k_j = 0$ (and $\boldsymbol{b}_j$ is zero-dimensional).

The likelihood contribution for individual i in terms of observed and unobserved terms (i.e. the complete-data likelihood contribution, apart from right-censoring) is[3]

$$L_i^\star(\theta) = \prod_{j=1}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1},\ldots,\boldsymbol{b}_{i1},\theta). \tag{2.3}$$

[3] This ignores the likelihood contribution of the entry and exit times, c_{ij-1} and c_{ij}, which leads to valid inference under the maintained assumption that these are independent of the event times. To focus on computational aspects we assume θ is identified and do not further discuss this issue.

The full complete-data likelihood function is defined as the product of $L_i^\star(\theta)$ over i

$$L^\star(\theta) = \prod_{i=1}^{N} \prod_{j=1}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1}, \ldots, \boldsymbol{b}_{i1}, \theta), \tag{2.4}$$

where N is the sample size.

The complete-data likelihood function given in (2.4) cannot be evaluated when the data are not complete. Simply omitting terms that involve missing data in (2.3) and maximising the computable part of the likelihood function generally does not yield a consistent estimator of θ. This is because the resulting truncated sample may not be representative of the population (see e.g. Moffitt and Rendall, 1995).

To get the likelihood contribution of the observed events, the unobserved events must be integrated out. For an individual who is under observation during odd-numbered periods (so j_i is odd), the incomplete-data likelihood contribution is[4]

$$\begin{aligned} L_i(\theta) = \iint \cdots \int & \left(\prod_{j=1:j \,\text{odd}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{j-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_2, \boldsymbol{b}_{i1}, \theta) \right) \\ & \times \left(\prod_{j=1:j \,\text{even}}^{j_i} g_j(\boldsymbol{b}_j|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_2, \boldsymbol{b}_{i1}, \theta) \right) d\boldsymbol{b}_{j_i-1} \ldots d\boldsymbol{b}_4 \, d\boldsymbol{b}_2 \\ = \mathsf{E}_{\boldsymbol{B}^\theta_{i2}} & \left[\cdots \mathsf{E}_{\boldsymbol{B}^\theta_{ij_i-1}} \left[\prod_{j=1:j \,\text{odd}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{B}^\theta_{ij-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{B}^\theta_{i2}, \boldsymbol{b}_{i1}, \theta) \right.\right. \\ & \left.\left. \middle| \, \boldsymbol{B}^\theta_{ij_i-2} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}^\theta_{i2}, \boldsymbol{B}^\theta_{i1} = \boldsymbol{b}_{i1} \right] \cdots \middle| \, \boldsymbol{B}^\theta_{i1} = \boldsymbol{b}_{i1} \right], \end{aligned} \tag{2.5}$$

where $\boldsymbol{B}^\theta_{ij}$ denotes a random vector of potential event times for individual i in period j, whose conditional probability density function given prior history is given in (2.2), taking individual i's realised observation period endpoints $c_{i0}, \ldots, c_{j_i}$ as given. The superscript θ serves as a reminder that this distribution is governed by the θ at which the likelihood contribution is evaluated, not the so-called true value behind the realised events $\boldsymbol{b}_{ij}$.

[4] Admittedly the notation is sloppy here, since the dimension of the terms integrated out are random, and the limits of the definite integrals are omitted. The notation could be made formally correct by conditioning on and summing over the possible dimensions of the vectors.

Similarly, for an individual who is under observation during even-numbered periods (so j_i is even), the incomplete-data likelihood contribution is

$$
\begin{aligned}
L_i(\theta) = \iint \cdots \int \Bigg(\prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_j | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \Bigg) \\
\times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{j-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \Bigg) \, d\boldsymbol{b}_{j_i-1} \ldots d\boldsymbol{b}_3 \, d\boldsymbol{b}_1 \\
= \mathsf{E}_{\boldsymbol{B}^\theta_{i1}} \Bigg[\cdots \mathsf{E}_{\boldsymbol{B}^\theta_{ij_i-1}} \Bigg[\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{B}^\theta_{ij-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}^\theta_{i1}, \theta) \qquad (2.6) \\
\Bigg| \, \boldsymbol{B}^\theta_{ij_i-2} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}^\theta_{i2} = \boldsymbol{b}_{i2}, \boldsymbol{B}^\theta_{i1} \Bigg] \cdots \Bigg].
\end{aligned}
$$

Note the outermost expectation is unconditional here, since there is no history prior to period 1.

The full incomplete-data likelihood function is defined as the product of $L_i(\theta)$ over i. Since this is the exact likelihood function for the observed data, the maximiser is a consistent and asymptotically efficient estimator of θ. However, computing this function is hampered by the fact that in general the integrals (expectations) cannot be solved analytically. In typical model specifications, the event density function depends non-linearly on previous events, and the integrals are not separable.

2.2.2 Monte Carlo integration

Our proposal is to use Monte Carlo simulation to integrate out the unobserved terms. For each individual we draw R independent pseudo-histories for periods with missing information. For a given value of θ, we then approximate the likelihood function by averaging over the R pseudo-histories. That is, for an individual who is under observation during odd-numbered periods, we compute

$$
L_i^{MC}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}^r_{ij-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}^r_{i2}, \boldsymbol{b}_{i1}, \theta), \qquad (2.7)
$$

and for an individual who is under observation during even-numbered periods, we compute

$$L_i^{MC}(\theta) = \frac{1}{R}\sum_{r=1}^{R}\prod_{j=1:j\,\mathrm{even}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1}^r, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta), \tag{2.8}$$

where for each $r = 1, \ldots, R$ and $j = 1, \ldots, j_i$ the $\boldsymbol{b}_{ij}^r$ are sequences of simulated event times specific to individual i's period j, compatible with the individual's observed and simulated event history, and compatible with the density evaluated at θ. That is, each $\boldsymbol{b}_{ij}^r$ is drawn from the conditional distribution g_j given in (2.2), with simulated prior event times replacing actual times when the latter are unobserved, and using the θ at which the likelihood function is evaluated. (For simplicity, the dependence of $\boldsymbol{b}_{ij}^r$ on θ is suppressed in the notation.) Let k_{ij}^r denote the dimension of $\boldsymbol{b}_{ij}^r$. Standard arguments (the law of large numbers) imply that L_i^{MC} converges to L_i pointwise as R diverges to infinity.

The dynamic nature of the density function g_j means that the simulation must be done sequentially. Recall that f denotes the conditional density of events, and F is the corresponding cumulative distribution function. For common parametric specifications of the hazard function, f, F and F^{-1} are easily evaluated using closed-form formulae. Pseudo-histories can therefore be created using the inversion method.[5]

Suppose first that $(c_{i0}, c_{i1}]$ is a period where individual i is not under observation. To simulate a first event time for this individual, we draw a pseudo-random number u_{i11}^r from the uniform distribution and then compute a candidate event time by $b_{i11}^r = F^{-1}(u_{i11}^r|s_i(c_{i0}), \theta)$. If $b_{i11}^r > c_{i1}$, we decide that no events happened during $(c_{i0}, c_{i1}]$ and set $k_{i1}^r = 0$. If $b_{i11}^r \le c_{i1}$, we keep b_{i11}^r and draw a second candidate event time. In general, having drawn $b_{i1k-1}^r, \ldots, b_{i11}^r$ with $b_{i1k-1}^r \le c_{i1}$, we draw a candidate for the kth event time by $b_{i1k}^r = F^{-1}(u_{i1k}^r|s_i^r(b_{i1k-1}^r), \theta)$, where u_{i1k}^r is another (independent) draw from the uniform distribution and where $s_i^r(b_{i1k-1}^r)$ includes the simulated previous

[5] Admittedly, misspecification of the models would deteriorate the performance of the MSL method relative to common ML methods since the simulation procedure in the MSL method is based on the models that are estimated.

events $b^r_{i1k-1}, \dots, b^r_{i11}$. If $b^r_{i1k} > c_{i1}$, the rth pseudo-history is complete with $k^r_{i1} = k - 1$ and $\boldsymbol{b}^r_{i1} = (b^r_{i1k^r_{i1}}, \dots, b^r_{i11})$. If $b^r_{i1k} \leq c_{i1}$, we increment k and consider the next candidate event time.

The simulation procedure is similar for other periods where an individual is not under observation. The only difference is that the history includes the observed event times during prior periods where the individual is under observation as well as simulated event times during prior periods where the individual is not under observation. For example, if individual i is under observation during $(c_{i0}, c_{i1}]$ but not during $(c_{i1}, c_{i2}]$, then $s^r_i(b^r_{i2k-1})$ includes the simulated events b^r_{i2k-1}, …, b^r_{i21} as well as the observed events $\boldsymbol{b}_{i1}$.

As pointed out by several authors (see e.g. Stern, 1997; Brinch, 2012), it is essential for successful numerical maximisation to use the same underlying draws from the uniform distribution in all the evaluations of the likelihood function (including computation of numerical derivatives).

The full incomplete-data simulated likelihood function is defined as the product of $L^{MC}_i(\theta)$ over i. Maximising the simulated likelihood function yields a consistent and asymptotically efficient estimator under standard conditions provided $\sqrt{N}/R \to 0$ as $N \to \infty$ where N is the number of individuals in the sample [Gouriéroux and Monfort, 1991].

2.2.3 Importance sampling

The simulated likelihood contributions described above are not everywhere continuous. Discontinuities occur when a small change in θ leads to a switch in the decision of whether to retain or discard a candidate event time (b^r_{ijk}). These discontinuities mean that standard maximisation methods for differentiable functions such as Newton's method may not work well.

Since the magnitude of the discontinuities are of order $1/R$, one approach to numerical maximisation of the likelihood function is to use a standard derivative-based method with R very large, and increase R whenever a discontinuity is causing problems. Another approach is to use a non-gradient method. These approaches will generally lead to convergence, but are expected to be slow.

An appealing method is to smooth the likelihood contributions using importance sampling techniques. In the present context, an importance sampling distribution for b_{ij} can be any given conditional distribution of events during period j given previous events. For concreteness, we choose g_j evaluated at some fixed value θ^*. For an individual who is under observation during even-numbered periods (the odd-numbered case is similar), the incomplete-data likelihood contribution can be written as

$$\begin{aligned}
L_i(\theta) = \iint \cdots \int \Bigg(\prod_{j=1:j\,\text{odd}}^{j_i} & g_j(\boldsymbol{b}_j | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \\
& \times \frac{g_j(\boldsymbol{b}_j | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta^*)}{g_j(\boldsymbol{b}_j | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta^*)} \Bigg) \\
& \times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{j-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \Bigg) \, d\boldsymbol{b}_{j_i-1} \ldots d\boldsymbol{b}_3 \, d\boldsymbol{b}_1 \\
= \mathsf{E}_{\boldsymbol{B}_{i1}^{\theta^*}} \Bigg[\cdots \mathsf{E}_{\boldsymbol{B}_{ij_i-1}^{\theta^*}} \Bigg[\Bigg(& \prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{B}_{ij}^{\theta^*} | \boldsymbol{b}_{ij-1}, \boldsymbol{B}_{ij-2}^{\theta^*}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta)}{g_j(\boldsymbol{B}_{ij}^{\theta^*} | \boldsymbol{b}_{ij-1}, \boldsymbol{B}_{ij-2}^{\theta^*}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta^*)} \Bigg) \\
& \times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{B}_{ij-1}^{\theta^*}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta) \Bigg) \\
& \qquad \Bigg| \; \boldsymbol{B}_{ij_i-2}^{\theta^*} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}_{i2}^{\theta^*} = \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*} \Bigg] \cdots \Bigg].
\end{aligned} \tag{2.9}$$

The corresponding simulated likelihood contribution is

$$\begin{aligned}
L_i^{IS}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \Bigg(& \prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{b}_{ij}^r | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^r, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta)}{g_j(\boldsymbol{b}_{ij}^r | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^r, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta^*)} \Bigg) \\
& \times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{ij-1}^r, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta) \Bigg),
\end{aligned} \tag{2.10}$$

where $\boldsymbol{b}_{ij}^r$ for $r = 1, \ldots, R$ and $j = 1, \ldots, j_i$ are drawn from the importance sampling distribution $g_j(\cdot|\cdot, \theta^*)$ instead of the 'correct' distribution $g_j(\cdot|\cdot, \theta)$. The principle underpinning importance sampling is that the 'error' can be fixed by reweighting using the ratio of correct density over the importance sampling density.

One of the advantages of the importance sampling approach is that the simulated event times do not depend on the value of θ at which the likelihood contribution is evaluated, and hence the simulated likelihood function is continuous and differentiable. A potential drawback is that a very large R may be needed in order to achieve a good approximation to the likelihood function. Keane and Sauer [2010] suggest that it may be advantageous to scale the importance sampling weights to sum to R over r.[6]

2.2.4 Covariates

So far we have ignored covariates, in order to focus on missing event times. In practice, covariates can be time-invariant or time-varying. Incorporating covariates is straightforward when the covariate paths are completely observed. Usually covariates with incompletely observed paths can also be incorporated, using an extended simulation procedure. For example, in some cases the observation process is such that time-varying covariates are missing during the same periods when the event times are not observed. These covariates can be incorporated by specifying an auxiliary model for their paths, and using this model to integrate out the missing parts of the covariate paths.[7]

2.2.5 Unobserved heterogeneity

Allowing for individual-specific time-invariant effects is standard in the literature. These effects capture correlation across event times. It is well-known that omitting individual-specific time-invariant effects can lead to a bias towards negative duration dependence (see e.g. Elbers and Ridder, 1982; Heckman and Singer, 1984a). The effects are usually assumed to be independent of covariates ('random effects' in the econometrics literature, 'frailty' in the statistics literature). The distribution of the random effects is specified either as discrete (following Heckman and Singer, 1984b) or

[6] Hesterberg [1995] compare unnormalised importance sampling with several normalised importance samplers for the problem of estimating certain aspects of a normal distribution. He finds that there is no uniformly best method.

[7] See e.g. Keane and Sauer [2010] for a similar approach in a discrete-time setting.

as continuous such as a normal distribution with mean 0.

Let v_i denote the realised unobserved random effect for individual i, and consider the complete-data likelihood function given in (2.3). Including and integrating out the random effects gives

$$L_i^\star(\theta) = \int_{-\infty}^{\infty} \left(\prod_{j=1}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{ij-1}, \ldots, \boldsymbol{b}_{i1}, v, \theta) \right) dZ(v), \tag{2.11}$$

where Z denotes the cumulative distribution function of v_i, and implicitly θ has been augmented to include unknown parameters of the distribution of v_i. For simplicity, we also reuse the symbols g_j, f, h, and H to denote the corresponding functions which depend on the random effect. The modification required to include a random effect is similar in the other likelihood contributions given above.

In practice, if Z is continuous then the integration is carried out using Gaussian quadrature. While straightforward, this increases the computational burden somewhat. For example, with Q evaluation points $v_1, \ldots, v_Q$ and weights $w_i, \ldots, w_Q$, the simulated likelihood contribution in (2.10) becomes

$$\begin{aligned} L_i^{IS} = \sum_{q=1}^{Q} w_q \frac{1}{R} \sum_{r=1}^{R} & \left(\prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{b}_{ij}^{qr} | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^{qr}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta)}{g_j(\boldsymbol{b}_{ij}^{qr} | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^{qr}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta^*)} \right) \\ & \times \left(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{ij-1}^{qr}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta) \right), \end{aligned} \tag{2.12}$$

where $\boldsymbol{b}_{ij}^{qr}$ for $q = 1, \ldots, Q$, $r = 1, \ldots, R$ and $j = 1, \ldots, j_i$ are drawn from the importance sampling distribution $g_j(\cdot|\cdot, v_q, \theta^*)$ instead of the 'correct' distribution $g_j(\cdot|\cdot, v_q, \theta)$. Note that the same underlying random draws from the uniform distribution can be used for each q, but the simulated event times, and even the number of compatible simulated event times, k_{ij}^{qr}, will be different.

2.2.6 Estimation based on Heckman's method

The likelihood contribution for individual i's period j given in (2.2) is made up of subcontributions representing each of the events, and a term representing the final right-censored period when no events occurred. In general, the hazard function at any given time may depend on the entire previous history of events. However, in many applications it can be assumed that the hazard function depends only on recent history. For example, the hazard rate for an event occurring at time t may depend only on whether or not an event occurred (or the number of events that occurred) in the period $(t - \tau, t)$ for some fixed τ. In applications where the influence of history is limited, missing data may affect only some and not all of the event subcontributions. If so, then the terms in the likelihood function that do not depend on missing data are 'computable', and it may be feasible to handle the 'uncomputable' parts by adapting the idea of Heckman [1981].

We compare MSL estimation with an implementation of Heckman's method in our Monte Carlo experiments and in our empirical application. To describe how Heckman's idea can be adapted, define d_{ijk} to be 1 if $h(b_{ijk}|s_i(b_{ijk-1}), v_i, \theta)$ is computable, and define d_{ijk} to be 0 otherwise. Define also $d_{ijk_{ij}+1}$ so that $\exp\Big(-H(c_{ij}|s_i(b_{ijk_{ij}}), v_i, \theta)\Big)$ is computable if and only if $d_{ijk_{ij}+1} = 1$.

It is helpful to begin with a simple two-period observation process, so suppose individual i is under observation in period 2 but not in period 1. By definition, the computable terms are those that do not depend on the unobserved events in period 1. Since they don't depend on period 1 events, they can be factored out of the integral in the incomplete-data likelihood contribution for individual i. Allowing for unobserved heterogeneity, we have from (2.6) that

$$\begin{aligned}L_i(\theta) &= \int_{-\infty}^{\infty}\left\{\int g_2(\boldsymbol{b}_{i2}|\boldsymbol{b}_1, v, \theta) g_1(\boldsymbol{b}_1|v,\theta)\, d\boldsymbol{b}_1\right\} dZ(v)\\ &= \int_{-\infty}^{\infty}\Bigg\{\Bigg[\int\left(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}), v, \theta)^{1-d_{i2k}}\right)\\ &\qquad\qquad \times \exp\Big(-H(c_{i2}|s_i(b_{i2k_{i2}}), v,\theta)\Big)^{1-d_{i2k+1}} g_1(\boldsymbol{b}_1|v,\theta)\, d\boldsymbol{b}_1\Bigg] \qquad (2.13)\\ &\qquad\qquad \times\left(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}), v, \theta)^{d_{i2k}}\right)\\ &\qquad\qquad \times \exp\Big(-H(c_{i2}|s_i(b_{i2k_{i2}}), v,\theta)\Big)^{d_{i2k+1}}\Bigg\} dZ(v).\end{aligned}$$

The integral with respect to $\boldsymbol{b}_1$ is uncomputable, because the necessary history is not observed. Heckman's idea was to approximate this using a reduced-form density that is based on as much predetermined information as is available, incorporates unobserved heterogeneity, and uses a flexible parametric specification. How much information is available depends on the details of how the hazard rate depends on previous history.

Let $h^\dagger(t|s(t'), v, \xi)$ for $t > t'$ be an approximate conditional hazard function evaluated at time t given the event history until time t'. For simplicity, we do not introduce new notation for the observed history itself. The principle is that $h^\dagger$ is parameterised so that it depend only on the part of $s(t')$ that is observed at time t'. Hence, $h^\dagger(t|s(t'), v, \xi)$ is computable even though $s(t')$ is not fully observed. For example, in our empirical application no part of $s(t')$ is observed for left-censored histories, so we parameterise $h^\dagger$ in terms of t and v only. Let $H^\dagger$ denote the corresponding cumulative hazard function from time t' to time t, and define $f^\dagger$ by

$$f^\dagger(t|s(t'), v, \xi) = h^\dagger(t|s(t'), v, \xi) \exp\Big(-H^\dagger(t|s(t'), v, \xi)\Big), \quad t > t'. \qquad (2.14)$$

Then the hope is that given θ for some ξ we have that

$$\begin{aligned}
&\int \left(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}), v, \theta)^{1-d_{ijk}} \right) \\
&\qquad\qquad \times \exp\Big(-H(c_{i2}|s_i(b_{i2k_{i2}}), v, \theta)\Big)^{1-d_{ijk+1}} g_1(\boldsymbol{b}_1|v, \theta)\, d\boldsymbol{b}_1 \\
&\quad \approx \left(\prod_{k=1}^{k_{i2}} f^{\dagger}(b_{i2k}|s_i(b_{i2k-1}), v, \xi)^{1-d_{ijk}} \right) \exp\Big(-H^{\dagger}(c_{i2}|s_i(b_{i2k_{i2}}), v, \xi)\Big)^{1-d_{ijk+1}}.
\end{aligned} \tag{2.15}$$

Substituting the approximation into (2.13) gives an approximate likelihood contribution as a function of (θ, ξ).

In the general multi-period case, the approximate likelihood contribution for an individual who is under observation during even-numbered periods (the odd-numbered case is similar) is

$$\begin{aligned}
L_i^{\dagger}(\theta, \xi) = \int_{-\infty}^{\infty} \Bigg\{ &\prod_{j=1:j\,\text{even}}^{j_i} \left(\prod_{k=1}^{k_{ij}} f(b_{ijk}|s_i(b_{ijk-1}), v_i, \theta)^{d_{ijk}} \right. \\
&\left. \times f^{\dagger}(b_{ijk}|s_i(b_{ijk-1}), v_i, \xi)^{1-d_{ijk}} \right) \exp\Big(-H(c_{ij}|s_i(b_{ijk_{ij}}), v_i, \theta)\Big)^{d_{ijk_{ij}+1}} \\
&\times \exp\Big(-H^{\dagger}(c_{ij}|s_i(b_{ijk_{ij}}), v_i, \xi)\Big)^{1-d_{ijk_{ij}+1}} \Bigg\}\, dZ(v).
\end{aligned} \tag{2.16}$$

Maximising the corresponding full likelihood function yields a consistent estimator of θ, provided the approximate reduced-form model is in fact correctly specified. Generally the hope is that the approximation is good enough that the magnitude of the inconsistency is acceptable.

2.3 Monte Carlo experiments

To investigate the performance of the MSL approach, we carried out a small set of Monte Carlo experiments. The designs feature mixed proportional hazards with a

Weibull baseline hazard function, a single time-invariant covariate, x_i, and a continuous random effect, v_i. The covariate and the random effect are realisations from a standard normal distribution.

Separate models are specified for the first event and for subsequent events. Current duration dependence is captured in the baseline hazards. After the first event, the hazard rates also depend on whether an event occurred or not during a recent period of fixed length (i.e. a moving window). Specifically, the hazard function for the first event is

$$h_1(t|s(0), x, v, \theta) = \alpha_1 t^{\alpha_1 - 1} \exp(x\beta_1 + \mu_1 + v\sigma_1), \quad t > 0. \tag{2.17}$$

With t' representing the most recent event time before t, the hazard function for subsequent events is

$$h_2(t|s(t'), x, v, \theta) = \alpha_2 t^{\alpha_2 - 1} \exp(1(t < t' + \tau)\gamma + x\beta_2 + \mu_2 + v\sigma_2), \quad t > t', \tag{2.18}$$

where $\theta = (\alpha_1, \beta_1, \mu_1, \sigma_1, \alpha_2, \gamma, \beta_2, \mu_2, \sigma_2)'$, and τ is a constant that varies across experiments. We normalise $\sigma_1 \geq 0$ and $\sigma_2 \geq 0$. The parameters used in the data-generating processes are fixed at $\alpha_1 = 1$, $\beta_1 = 0.2$, $\mu_1 = -0.5$, $\alpha_2 = 1$, $\gamma = 0.5$, $\beta_2 = 0.2$, and $\mu_2 = -0.5$, while either $\sigma_1 = 0$, $\sigma_2 = 0$ (known) or $\sigma_1 = 1$, $\sigma_2 = 1$ (estimated) as indicated in the tables.

Note that baseline time does not reset after an event in these designs. Alternatively, the baseline hazard rate can be specified in terms of $t - t'$. More flexible models can be obtained by specifying separate hazard functions for second events, third events, etc. Less flexible models can be obtained by assuming $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\mu_1 = \mu_2$, and $\sigma_1 = \sigma_2$. In this case, the model effectively consists of a single hazard specification since (2.17) is simply (2.18) with $\gamma = 0$. Such a specification was adopted for example by Keane and Sauer [2010]. Our designs satisfy these restrictions, but we do not impose them in the estimation.

The observation process mimics a sampling procedure where analysis time is age and data are collected from the population stock over a fixed calendar period. Specifically, half the sample are observed over the age range $(0, 1]$ while the other half is observed over $(1, 2]$. That is, the former is right-censored at time 1 (and not left-censored), while the latter is left-censored at time 1 and right-censored at time 2. The number of non-left-censored individuals in the samples is $N_1 = 250$ and while the number of left-censored individuals is either $N_2 = 250$ or $N_2 = 500$ as indicated in the tables.

Across all designs, about half of the individuals in a sample do not have any events during their observation period. For those who do have observed events, the mean time until the first event is about 0.38. Since $\alpha_1 = 1$ and $\alpha_2 = 1$ imply memoryless exponential hazard functions, these statistics apply to both the left-censored and the non-left-censored.

We compute several estimators to compare the MSL approach with simple estimators that may be considered in practice. Estimator ISU is an MSL estimator which uses importance sampling techniques without scaling of the weights, while estimator ISN has the weights normalised to sum to one. For simplicity, we use the true data-degenerating process as the importance sampling distribution, and we set $R = 100$.

Estimator NLC uses only individuals with non-left-censored data (listwise deletion); that is, half the sample in the experiments with $N_2 = 250$ and a third of the sample when $N_2 = 500$.

Estimator HKM uses the approximate reduced-form idea of Heckman [1981] to handle the left-censoring problem. For the designs considered here, the only uncomputable term in the likelihood contribution for the left-censored individuals concerns the first observed event in period 2, b_{i21}. This is because we do not know whether or not the first observed is the first actual event, while for subsequent observed events there is no ambiguity. Since no useful information is available in $s(1)$, we specify the auxiliary hazard function for b_{i21} as

$$h_3(t|s(1), x, v, \xi) = \alpha_3 t^{\alpha_3 - 1} \exp(x\beta_3 + \mu_3 + v\sigma_3), \quad t > 0. \tag{2.19}$$

The literature on dynamic panel data models usually does not distinguish between the start of the event process and the start of the observation period, although these are associated with conceptually distinct problems: at the start of the event process lags cannot exist so logically a different structural equation is required, whereas at the start of the observation period lags may exist so a method for dealing with missing data is required. Here we maintain the distinction between left-censoring and genuine first events. That is, our HKM implementation estimates the parameters of all three hazard functions.

There are 1000 samples in each experiment.[8] In designs with random effects, unobserved heterogeneity is integrated out using Gauss-Hermite quadrature with $Q = 10$ evaluation points.

Table 2.1 shows root mean square errors (RMSEs) for the four estimators for designs without random effects. The likelihood function is separable in the parameters pertaining to the first and subsequent events, respectively. Consequently, the NLC and HKM estimates for the parameters of the first hazard function are identical. The RMSEs for the IS estimates are slightly lower. For the second hazard functions, the HKM estimates improve dramatically on the NLC estimates. This is because the usable sample is twice as large, and the HKM involve only a few more parameters. The RMSEs for the IS estimates are lower again, especially for γ and μ_2.

The value of τ does not affect the first hazard function, but the higher τ, the more history data are needed to estimate the second hazard function. The problem of missing data therefore becomes more severe and higher RMSEs are expected. This is confirmed in Table 2.1. The results for the first hazard function do not change, because the same data are used. For the second hazard function, the RMSEs for $\ln \alpha_2$ and β_2 also remain roughly constant, while the RMSEs for γ and μ_2 increase. The increase occurs because the number of individuals with no recent events becomes small when τ is large, and hence it becomes difficult to estimate μ_2 accurately.[9] Since individuals

[8] The results omit a few samples (max 3 per experiment) where the estimation procedure did not converge in a sense that the estimates of σ_1 and σ_2 diverged to the negative infinity.

[9] In the extreme, if these individuals experience no further events, the estimated hazard should be

who have recent events identify the sum $\gamma + \mu_2$, the uncertainty in the estimates of μ_2 is mirrored in the estimates of γ. However, the HKM estimator is better than the NLC estimator, since it uses much more of the sample, and the two IS estimators are better than the HKM estimator, since they use the sample efficiently.

Table 2.2 shows results for designs with random effects. Looking first at the case where $\tau = 0.3$ and $N_2 = 250$, the patterns are similar to those without random effects. The HKM estimator improves on the NLC estimator and the IS estimators perform better than the HKM estimator. Estimation of distributions of random effects is notoriously difficult, so it is not surprising to find much higher RMSEs for $\ln \sigma_1$ and $\ln \sigma_2$.

As τ increases, the results for the first-event parameters and for $\ln \alpha_2$ and β_2 do not change much. Similar to the designs without random effects, estimation of γ and μ_2 becomes more difficult when τ is large, so the RMSEs for those parameters increase for all estimators. The increase is very large for the NLC and HKM estimators but only modest for the IS estimators, so the efficiency gain of the latter becomes more substantial. The patterns for the RMSEs of $\ln \sigma_1$ and $\ln \sigma_2$ are complex and not entirely intuitive. For example, the RMSEs for the NLC estimator of $\ln \sigma_2$ tend to increase with τ, but decrease for the HKM estimator. Presumably this is because the 'practical identification' of these parameters is weak, so small approximation errors in the simulated likelihood function can have large effects of the estimates.

When the number of left-censored individuals is increased from $N_2 = 250$ to $N_2 = 500$, the results for the first-event parameters hardly change, while there is some improvement for the parameters relating to the second hazard function. This is particularly true for the difficult parameters $\ln \sigma_1$ and $\ln \sigma_2$, and to a lesser extent for γ and μ_2.

To conclude, it is clear that there are potentially large efficiency gains in using MSL estimation over methods based on listwise deletion or Heckman's approximate reduced-form modelling of initial conditions. The gains are particularly high for parameters that

zero, which means $\hat{\mu}_2 = -\infty$.

are difficult to estimate. The fact that the results for the ISU and ISN estimators are not identical reveal a disadvantage of MSL estimation; namely, that numerical integration inevitably involves some approximation error. As a practical guide, we suggest computing several MSL estimates, using different importance sampling distributions with and without scaling of the weights. If the estimates are too different, then the values of R and Q can be increased until all estimates are in sufficient agreement.

2.4 Empirical application

2.4.1 Modelling ischaemic heart disease risk

To investigate the performance of the MSL approach in a practical setting, we apply the MSL estimation methods and the two alternative methods to a dynamic model of ischaemic heart disease events (IHDs) for males of Maori descent in New Zealand.[10]

We combine nationwide administrative data on hospital admissions and death registrations during the period 2002–2012 with census data from 2001. The combined data set is essentially representative of the population of New Zealand in 2002, except that we exclude people with type 1 diabetes. For each IHD event (hospitalisation or death), we have information on gender, ethnicity, date of birth, date of admission and diagnoses if admitted, and date of death and cause of death if died. Since IHD events are rare before age 40, we define analysis time 0 as age 40. We do not model risk after age 85, because the population over age 85 is very small. However, the full population is large, so our estimation sample is a randomly drawn subset consisting of 50,000 individuals.[11]

Table 2.3 shows summary statistics for the estimation sample. The number of

[10] In related research, we present a thorough investigation of the heart attack (acute myocardial infarction) risk for New Zealanders of Maori and European descent using similar data [Lee and Gørgens, 2019], which is Chapter 3 of the thesis.

[11] IHD events appear in the data as codes I20–I25 according to the *International Classification of Diseases 10 Australian Modification.* We treat events that occur within 29 days of each other as a single event. Since the cause of death is in the register, death is not associated with underreporting events; however, some cases are not acute and may not lead to a hospital admission, so it is likely that some less severe events do not appear in the data.

people decreases with age, and there are not many people aged 80–84 in the sample. The total time at risk is 376,739 years, which is 7.5 years per person on average. The total number of observed IHDs is 7,974. Whether looking at incidence rates or the number of observed IHDs, it is clear that the IHD risk increases with age. The amount of left-censoring in the estimation sample is very large, with about 63% of histories being left-censored.

We consider a dynamic model of IHDs similar to the one in the Monte Carlo study, except that the events follow Gompertz instead Weibull distributions in order to better fit exponentially increasing risk. Let t denote the elapsed time since age 40 measured in decades (i.e. $t = (\text{age} - 40)/10$), and let v denote the standardised random effect. Then the hazard function for the first IHD event is

$$h_1(t|v, \theta) = \exp(t\alpha_1 + \mu_1 + v\sigma_1), \quad t > 0, \tag{2.20}$$

and the hazard function for subsequent IHDs is

$$h_2(t|t', v, \theta) = \exp(t\alpha_2 + 1(t \leq t' + \tau)\gamma + \mu_2 + v\sigma_2), \quad t > t', \tag{2.21}$$

where t' is the event time of the most recent IHD. The length of the high-risk period is fixed at $\tau = 0.1$ decade in our main estimates, but we also consider higher values. In addition, the HKM estimator requires an auxiliary model for the left-censored event times, which we specify as

$$h_3(t|v, \xi) = \exp(t\alpha_3 + \mu_3 + v\sigma_3), \quad t > 0. \tag{2.22}$$

Below we discuss estimates from both models with and models without random effects.

The focus on this investigation is to compare the ISU and ISN estimators with each other and with the NLC and HKM estimators. For the two MSL estimators, we initially set $R = 100$ but report on other values later. In models without random effects, we set the parameters of the importance sampling distributions, θ^*, equal to the HKM estimates. In models with random effects, we set $Q = 10$ and θ^* equal to the HKM estimates with σ_1^* and σ_2^* reset to low values as indicated in the table notes. The modification of σ_1^* and σ_2^* reduces the number of events in the simulated pseudo-histories for large values of v, which conserves computer memory and reduces computing time. All reported standard errors are computed as the outer product of the relevant score functions.

2.4.2 Estimated models without random effects

Table 2.4 shows the estimated parameters for models without random effects, and Figure 2.2 shows the estimated hazard functions. The ISU and ISN estimates are practically identical. The HKM estimates are virtually identical to the NLC for the first event (as they should be), but similar to the ISU and ISN estimates for subsequent events. The NLC estimates for the risk of subsequent events are absurd, as they suggest risk declines with age.[12]

To summarise the ISU/ISN findings, note first that the risk of the first IHD event is quite small around age 40 ($\mu_1 \approx -3.1$), but increases with age ($\alpha_1 \approx 0.5$). The risk of a subsequent IHD event is generally much higher than for first IHD ($\mu_2 \approx 0.1$), and also increasing in age albeit at a slower rate ($\alpha_2 \approx 0.1$). Having had an event, the risk of another event during the following year is more than three times as large as having it at any time later ($\gamma \approx 1.2$).

The four estimation methods provide statistically similar results, in the sense that the 95% confidence intervals overlap. It is not surprising that both the NLC and the HKM estimates of μ_1 and α_1 are somewhat different from the ISU and ISN estimates, since they rely on much less data. This is reflected in their standard errors, which are large. For the risk of subsequent events, the NLC estimator stands out with a large estimate of μ_2 and a negative estimate of α_2, but again these estimates are not statistically significantly different from 0. The NLC estimator can only utilise the data

[12] ISU and ISN estimation of the model without random effects takes about 4 minutes each with our computers and our code when $R = 100$ and 40 minutes when $R = 1000$, while NLC and HKM estimation takes about 1 minute.

for the age range 40–49, and this is a time when people are relatively healthy and few experience multiple events.

The standard errors of the HKM, ISU and ISN estimators are very similar for the parameters relating to the subsequent events. However, there are substantial efficiency gains for the MSL estimators relative to the alternative methods for the parameters of the risk of the first event. In fact, the standard errors for the ISU and ISN estimates of the risk of the first event are as small as those of subsequent events. The efficiency gain is expected, since the left-censored histories (about 63% of the estimation sample) contribute fully in the MSL estimation, while in the alternative estimation methods they contribute only if events occur during the observation period.

To examine the sensitivity of the MSL estimators to the number of simulated histories, Table 2.5 shows MSL estimation results for different values of R. (The first column in Table 2.5 has the same estimates as shown in Table 2.4.) The estimates appear remarkably insensitive. Looking across the columns, there are no practical difference between the estimates, nor between the standard errors. (The difference between the estimates for $R = 100$ and $R = 1000$ is at most 0.7 standard errors.) Clearly $R = 100$ is sufficient for this application.

Including an indicator function to capture the elevated risk for a period immediately following an event is a simple way to distinguish between short-term and long-term risks. In the health and medical literatures it is common to focus on outcomes during a fixed period of a month or a year after an event. However, the one-year cutoff between the two regimes for the risk of subsequent events is essentially arbitrary. Therefore we briefly consider other values of τ.

Table 2.6 compares estimation results for the risk of subsequent events across different values of τ. The precise interpretation of the parameters changes when τ changes, so we expect to get different estimates across the different model specifications. As previously discussed, the estimation of dynamic models is more difficult when τ is large. The table shows that the overall pattern is the same for all values of τ: ISU and

ISN estimates are nearly identical. The HKM estimates are similar to the NLC for the first event, but similar to the ISU/ISN estimates for subsequent events. The NLC estimates for the risk of subsequent events are absurd. In terms of standard errors, the efficiency advantage in estimating the risk of the first event appears for all values of τ. For the risk of subsequent events, notice that the standard errors are smaller for the MSL estimators than for the HKM estimator, with a slightly larger gap for larger values of τ.

2.4.3 Estimated models with random effects

We now turn to models which include random effects. Table 2.7 shows the estimated parameters, and Figure 2.3 shows the corresponding hazard functions. Overall, the patterns are similar to those for models without random effects. The exceptions are that the ISU and ISN estimates are now numerically different from each other, but not practically different, and they are substantially different from the HKM estimates for the risk of subsequent events.[13]

To summarise the ISU/ISN findings, we find that the median risk of the first IHD event is initially quite low ($\mu_1 \approx -3.3$), but increases with age ($\alpha_1 \approx 0.6$). The risk of subsequent events is higher ($\mu_2 \approx -0.9$), and increases with age at a slower rate ($\alpha_2 \approx 0.2$). The short-term increase in risk after an event is about two and half times as high as the long-term increase ($\gamma \approx 0.9$). The estimates suggest that there is considerable unobserved heterogeneity in IHD risk, especially in the risk of subsequent events ($\ln \sigma_1 \approx -0.5$, $\ln \sigma_2 \approx 0.2$). The largest differences between the ISU and ISN estimates occur for $\ln \sigma_1$ and $\ln \sigma_2$. However, as shown in Figure 2.3, the risks at the median of the estimated distribution of unobserved heterogeneity are similar.

Statistically, the different estimators are not completely in agreement in that the 95% confidence intervals do not overlap for some parameters. The standard errors of the ISU and ISN estimates in Table 2.7 tend to be smaller than those of the alternative

[13] ISU and ISN estimation of the model with random effects takes about 90 minutes each on our systems when $R = 100$ and about 25 hours when $R = 1000$, while NLC and HKM estimation takes about 2 minutes each.

methods, especially for the parameters relating to the first IHD event. This pattern is consistent with the efficiency advantage of full ML estimation. The exception is the standard errors for the estimates of $\ln \sigma_1$ and $\ln \sigma_2$, which are larger for ISU and ISN estimates.

The different results for the different estimation methods is likely driven by differences in the estimates of $\ln \sigma_1$ and $\ln \sigma_2$. If the influence of the random effects is large, then more observed IHD events are attributed to high values of the random effects (high innate risk) than to base level risk, dynamic effects, and age. As mentioned, it is generally difficult to estimate random effects distributions and, not unexpectedly, the standard errors for the estimates of $\ln \sigma_1$ and $\ln \sigma_2$ tend to be larger than for the other parameters.

The differences between the ISU and ISN estimates should diminish if the number of simulated pseudo-histories is increased. Table 2.8 compares the MSL estimation results for different values of R. (The first column in Table 2.8 is the same as ISU and ISN estimates in Table 2.7.) Unfortunately, the differences between the ISU and ISN parameter estimates remain nontrivial even for $R = 1000$, especially for the estimates of $\ln \sigma_1$ and $\ln \sigma_2$. Note that the ISN estimates are more stable than the ISU estimates, suggesting they are more reliable. However, this argument would be more convincing if the ISU estimates moved towards the ISN estimates for larger R. Figures 2.4 and 2.5 show that the risks at the medians of the estimated distribution of random effects. The stability of the ISN estimates is striking, while the jump in the ISU estimates for $R = 1000$ is a concern. We leave the reconciliation of these findings to future research.

Table 2.9 shows estimation results for different model specifications with different values of τ. The patterns from Table 2.7 are repeated: The NLC estimates are unreliable and their standard errors are high, while the HKM, ISU and ISN estimates are qualitatively similar if not exactly identical. The standard errors tend to be slightly lower for the ISU and ISN estimators. As before, it is the estimates of $\ln \sigma_1$ and $\ln \sigma_2$ that differ the most across methods, and their standard errors are relatively large.

The problem of estimating models with random effects is not specific to MSL estimation, and it is not uncommon to find that different approaches give somewhat different answers. For our random effects models of IHD events, we have more faith in the MSL estimates on the grounds that theoretically full ML estimation is expected to provide better results and practically the ISU and ISN estimates are substantially similar.

2.5 Concluding remarks

This chapter considers ML estimation of dynamic models of recurrent events in continuous time using censored data. We propose to deal with censoring by integrating out missing data from the likelihood function using Monte Carlo simulation and importance sampling techniques. We compare MSL estimation with estimators that either ignore left-censored individuals and middle-censored individuals (listwise deletion) or deal with censoring using ad hoc modifications to the likelihood function (Heckman's method). The Monte Carlo results show that there can be substantial efficiency gains in maximising the full simulated likelihood function. In an empirical application, we study the risk of ischaemic heart disease using models with and without random effects. We find that the MSL estimators typically have smaller standard errors, especially for parameters relating to the risk of having the first event. In models without random effects, the MSL estimators are clearly preferable. In models with random effects, we find some contradictory patterns, but the MSL estimators are most likely preferable.

There is a large literature that is concerned with the choice of importance sampling distributions in a variety of estimation problems. The question is difficult and the answer tends to be model specific. We use importance sampling distributions that are intuitively reasonable in that they tend to place most weight on outcomes that are most likely. It is a topic for future research to investigate the trade off between the choice of importance sampling distribution and the number of pseudo-histories needed for reliable inference.

We assume that the censoring and the event processes are independent, and we focus on settings where time origins and covariate paths are known. We anticipate that these assumptions can be relaxed, at the costs of further computational complications. Given the encouraging results for models of recurrent events, it is also likely that similar efficiency gains are available for example in multi-state transition models.

Table 2.1: RMSE for designs without random effects

		$N_2 = 250$		
Parameter	NLC	HKM	ISU	ISN
$\tau = 0.3$				
$\ln \alpha_1$	0.083	0.083	0.073	0.076
β_1	0.113	0.113	0.098	0.099
μ_1	0.096	0.096	0.088	0.088
$\ln \alpha_2$	0.281	0.149	0.148	0.149
γ	0.334	0.224	0.206	0.206
β_2	0.168	0.101	0.092	0.092
μ_2	0.338	0.252	0.238	0.239
$\tau = 0.5$				
$\ln \alpha_1$	0.083	0.083	0.073	0.077
β_1	0.113	0.113	0.098	0.098
μ_1	0.096	0.096	0.088	0.088
$\ln \alpha_2$	0.265	0.144	0.143	0.145
γ	0.467	0.299	0.246	0.247
β_2	0.157	0.092	0.086	0.086
μ_2	0.482	0.329	0.285	0.287
$\tau = 0.7$				
$\ln \alpha_1$	0.083	0.083	0.075	0.079
β_1	0.113	0.113	0.098	0.098
μ_1	0.097	0.096	0.089	0.089
$\ln \alpha_2$	0.256	0.140	0.139	0.140
γ	4.273	1.661	0.348	0.349
β_2	0.151	0.091	0.084	0.084
μ_2	4.267	1.665	0.380	0.382

The parameters used in the DGPs are fixed at $\alpha_1 = 1$, $\beta_1 = 0.2$, $\mu_1 = -0.5$, $\alpha_2 = 1$, $\gamma = 0.5$, $\beta_2 = 0.2$, and $\mu_2 = -0.5$. RMSE indicates root mean square errors. NLC and HKM indicate listwise deletion and Heckman's approximate reduced-form modelling, respectively. ISU and ISN indicate MSL estimators which uses importance sampling techniques without and with scaling of the weights, respectively. See text for DGP and implementation of estimators. Results for the parameters in the HKM auxiliary equation not shown.

Table 2.2: RMSE for designs with random effects

		$N_2 = 250$			$N_2 = 500$		
Parameter	NLC	HKM	ISU	ISN	HKM	ISU	ISN
$\tau = 0.3$							
$\ln \alpha_1$	0.139	0.132	0.126	0.123	0.131	0.128	0.120
β_1	0.150	0.145	0.126	0.128	0.146	0.116	0.113
μ_1	0.152	0.147	0.126	0.123	0.145	0.121	0.117
$\ln \sigma_1$	1.248	1.120	0.662	0.662	1.376	0.517	0.617
$\ln \alpha_2$	0.135	0.112	0.107	0.110	0.117	0.112	0.104
γ	0.282	0.193	0.177	0.175	0.164	0.142	0.138
β_2	0.153	0.137	0.090	0.114	0.107	0.074	0.086
μ_2	0.340	0.267	0.225	0.242	0.249	0.204	0.211
$\ln \sigma_2$	0.336	1.593	0.304	0.344	0.532	0.284	0.316
$\tau = 0.5$							
$\ln \alpha_1$	0.128	0.125	0.121	0.117	0.123	0.123	0.114
β_1	0.147	0.143	0.125	0.125	0.144	0.116	0.113
μ_1	0.151	0.146	0.126	0.124	0.143	0.118	0.116
$\ln \sigma_1$	1.379	1.000	0.490	0.491	0.939	0.491	0.452
$\ln \alpha_2$	0.124	0.131	0.105	0.105	0.115	0.111	0.104
γ	0.439	0.300	0.261	0.256	0.240	0.221	0.213
β_2	0.147	0.164	0.090	0.107	0.124	0.073	0.089
μ_2	0.491	0.356	0.284	0.289	0.300	0.259	0.251
$\ln \sigma_2$	0.339	0.794	0.300	0.338	0.518	0.284	0.317
$\tau = 0.7$							
$\ln \alpha_1$	0.163	0.121	0.119	0.115	0.120	0.121	0.111
β_1	0.169	0.144	0.124	0.125	0.144	0.112	0.110
μ_1	0.160	0.143	0.127	0.126	0.141	0.119	0.117
$\ln \sigma_1$	6.663	1.101	1.004	0.949	0.747	0.484	0.451
$\ln \alpha_2$	0.133	0.101	0.104	0.103	0.120	0.109	0.102
γ	4.369	1.077	0.440	0.429	0.712	0.370	0.352
β_2	0.193	0.096	0.089	0.103	0.117	0.072	0.092
μ_2	4.359	1.094	0.444	0.433	0.727	0.391	0.364
$\ln \sigma_2$	0.584	0.338	0.304	0.335	1.538	0.288	0.322

The parameters used in the DGPs are fixed at $\alpha_1 = 1$, $\beta_1 = 0.2$, $\mu_1 = -0.5$, $\sigma_1 = 1$, $\alpha_2 = 1$, $\gamma = 0.5$, $\beta_2 = 0.2$, $\mu_2 = -0.5$, and $\sigma_2 = 1$. RMSE indicates root mean square error. NLC and HKM indicate listwise deletion and Heckman's approximate reduced-form modelling, respectively. ISU and ISN indicate MSL estimators which uses importance sampling techniques without and with scaling of the weights, respectively. See text for DGP and implementation of estimators. Results for the parameters in the HKM auxiliary equation not shown.

Table 2.3: Summary statistics for the estimation sample

Age on 1 July 2002	30–39[†]	40–49	50–59	60–69	70–79	80–84	Total
Number of people	18,349[‡]	15,030	8,957	5,257	2,104	303	50,000
Total time at risk	90,500	145,196	82,251	43,915	14,150	728	376,739
Number of IHDs	528	1,833	2,467	2,173	908	65	7974
Incidence rate (×100)	0.58	1.26	3.00	4.95	6.42	8.93	2.12
Distribution of people by the number of observed IHDs (%)							
0	98.03	91.98	83.79	74.97	71.86	82.18	90.04
1	1.44	5.40	9.87	15.33	18.16	13.86	6.38
2	0.30	1.52	3.25	4.91	5.47	2.31	1.91
3+	0.23	1.10	3.09	4.79	4.52	1.65	1.67

See text for abbreviations. The unit for total time at risk is 1 year. The incidence rate is the number of IHDs divided by total time at risk. [†]The 30–39-year-olds become at risk when they turn 40; [‡]Non-left-censored histories.

Table 2.4: Estimates for models without random effects

Parameter	NLC	HKM	ISU	ISN
First IHD event				
α_1	0.786	0.783	0.467	0.465
	(0.212)	(0.212)	(0.018)	(0.018)
μ_1	−3.486	−3.485	−3.141	−3.138
	(0.095)	(0.095)	(0.032)	(0.032)
Subsequent IHD events				
α_2	−0.147	0.117	0.104	0.103
	(0.211)	(0.011)	(0.010)	(0.010)
γ	1.214	1.201	1.176	1.180
	(0.142)	(0.028)	(0.027)	(0.027)
μ_2	0.167	0.013	0.056	0.056
	(0.167)	(0.034)	(0.032)	(0.033)

See text for abbreviations. Results for the parameters in the HKM auxiliary equation not shown. An analysis time unit is 10 years. MSL estimation is implemented with $R = 100$, $\tau = 0.1$ decade, and θ^* equal to the HKM estimates.

Table 2.5: MSL estimates for models without random effects: different R

Parameter	$R=100$	$R=200$	$R=300$	$R=500$	$R=1000$
			ISU estimates		
First IHD event					
α_1	0.467	0.460	0.458	0.455	0.453
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
μ_1	−3.141	−3.131	−3.129	−3.124	−3.123
	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)
Subsequent IHD events					
α_2	0.104	0.104	0.104	0.104	0.104
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
γ	1.176	1.173	1.172	1.171	1.169
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
μ_2	0.056	0.064	0.067	0.067	0.071
	(0.032)	(0.033)	(0.033)	(0.033)	(0.033)
			ISN estimates		
First IHD event					
α_1	0.465	0.458	0.457	0.454	0.452
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
μ_1	−3.138	−3.129	−3.128	−3.123	−3.121
	(0.032)	(0.032)	(0.033)	(0.033)	(0.033)
Subsequent IHD events					
α_2	0.103	0.101	0.101	0.103	0.103
	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
γ	1.180	1.175	1.174	1.174	1.171
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
μ_2	0.056	0.071	0.074	0.070	0.074
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)

See text for abbreviations. An analysis time unit is 10 years. MSL estimation is implemented with $\tau = 0.1$ decade and θ^* equal to the HKM estimates.

Table 2.6: Estimates for models without random effects: different τ

Parameter	NLC	HKM	ISU	ISN
	$\tau = 0.5$ *decade*			
First IHD event				
α_1	0.784	0.784	0.529	0.528
	(0.213)	(0.212)	(0.018)	(0.018)
μ_1	−3.485	−3.485	−3.192	−3.190
	(0.095)	(0.095)	(0.032)	(0.032)
Subsequent IHD events				
α_2	−0.693	0.108	0.088	0.086
	(0.182)	(0.010)	(0.009)	(0.009)
γ	0.940	1.246	1.228	1.231
	(0.487)	(0.080)	(0.073)	(0.074)
μ_2	0.175	−0.557	−0.521	−0.517
	(0.499)	(0.083)	(0.076)	(0.077)
	$\tau = 0.7$ *decade*			
First IHD event				
α_1	0.783	0.785	0.560	0.557
	(0.216)	(0.212)	(0.019)	(0.019)
μ_1	−3.485	−3.485	−3.231	−3.230
	(0.098)	(0.095)	(0.032)	(0.032)
Subsequent IHD events				
α_2	−0.789	0.098	0.079	0.076
	(0.180)	(0.010)	(0.009)	(0.009)
γ	8.850	1.558	1.782	1.743
	(∞)	(0.181)	(0.160)	(0.165)
μ_2	−7.713	−0.913	−1.127	−1.080
	(∞)	(0.183)	(0.162)	(0.167)

See text for abbreviations. Results for the parameters in the HKM auxiliary equation not shown. An analysis time unit is 10 years. MSL estimation is implemented with $R = 100$ and θ^* equal to the HKM estimates.

Table 2.7: Estimates for models with random effects

Parameter	NLC	HKM	ISU	ISN
First IHD event				
α_1	0.833	1.051	0.614	0.591
	(0.228)	(0.231)	(0.021)	(0.023)
μ_1	−3.891	−4.725	−3.357	−3.296
	(0.489)	(0.272)	(0.066)	(0.070)
$\ln\sigma_1$	0.242	0.804	−0.645	−0.403
	(0.594)	(0.106)	(0.358)	(0.242)
Subsequent IHD events				
α_2	−0.073	0.321	0.222	0.192
	(0.348)	(0.025)	(0.026)	(0.029)
γ	0.741	0.920	0.921	0.856
	(0.191)	(0.033)	(0.032)	(0.033)
μ_2	−1.077	−1.781	−0.958	−0.778
	(0.671)	(0.102)	(0.145)	(0.147)
$\ln\sigma_2$	0.401	0.332	0.134	0.259
	(0.160)	(0.034)	(0.034)	(0.041)

See text for abbreviations. Results for the parameters in the HKM auxiliary equation not shown. An analysis time unit is 10 years. MSL estimation is implemented with $R = 100$, $Q = 10$, $\tau = 0.1$ decade, and θ^* equal to the modified HKM estimates with $\ln\sigma_1^* = \ln\sigma_2^* = -2$.

Table 2.8: MSL estimates for models with random effects: different R

Parameter	$R=100$	$R=200$	$R=300$	$R=500$	$R=1000$
		ISU estimates			
First IHD event					
α_1	0.614	0.588	0.581	0.574	0.552
	(0.021)	(0.020)	(0.021)	(0.023)	(0.021)
μ_1	−3.357	−3.303	−3.290	−3.292	−3.241
	(0.066)	(0.060)	(0.061)	(0.072)	(0.062)
$\ln\sigma_1$	−0.645	−0.849	−0.872	−0.708	−1.018
	(0.358)	(0.473)	(0.512)	(0.465)	(0.712)
Subsequent IHD events					
α_2	0.222	0.214	0.209	0.202	0.189
	(0.026)	(0.025)	(0.025)	(0.025)	(0.024)
γ	0.921	0.903	0.915	0.911	0.898
	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)
μ_2	−0.958	−0.832	−0.798	−0.788	−0.655
	(0.145)	(0.145)	(0.148)	(0.154)	(0.161)
$\ln\sigma_2$	0.134	0.093	0.074	0.057	0.029
	(0.034)	(0.035)	(0.035)	(0.036)	(0.036)
		ISN estimates			
First IHD event					
α_1	0.591	0.573	0.583	0.575	0.567
	(0.023)	(0.023)	(0.025)	(0.023)	(0.026)
μ_1	−3.296	−3.261	−3.316	−3.289	−3.303
	(0.070)	(0.070)	(0.081)	(0.070)	(0.085)
$\ln\sigma_1$	−0.403	−0.512	−0.299	−0.402	−0.292
	(0.242)	(0.302)	(0.237)	(0.241)	(0.251)
Subsequent IHD events					
α_2	0.192	0.187	0.204	0.189	0.193
	(0.029)	(0.029)	(0.030)	(0.027)	(0.030)
γ	0.856	0.845	0.852	0.847	0.832
	(0.033)	(0.033)	(0.033)	(0.032)	(0.033)
μ_2	−0.778	−0.703	−0.827	−0.777	−0.786
	(0.147)	(0.155)	(0.158)	(0.145)	(0.166)
$\ln\sigma_2$	0.259	0.235	0.251	0.254	0.243
	(0.041)	(0.041)	(0.042)	(0.041)	(0.043)

See text for abbreviations. An analysis time unit is 10 years. MSL estimation is implemented with $Q = 10$, $\tau = 0.1$ decade, and θ^* equal to the modified HKM estimates with $\ln\sigma_1^* = \ln\sigma_2^* = -2$.

Table 2.9: Estimates for models with random effects: different τ

Parameter	NLC	HKM	ISU	ISN
	$\tau = 0.5$ *decade*			
First IHD event				
α_1	0.835	0.916	0.697	0.646
	(0.223)	(0.217)	(0.026)	(0.022)
μ_1	−3.917	−4.299	−3.679	−3.426
	(0.290)	(0.193)	(0.075)	(0.062)
$\ln \sigma_1$	0.273	0.591	0.243	−0.121
	(0.311)	(0.108)	(0.072)	(0.115)
Subsequent IHD events				
α_2	−0.435	0.296	0.274	0.200
	(0.310)	(0.025)	(0.022)	(0.024)
γ	0.188	0.782	0.844	0.803
	(0.516)	(0.085)	(0.076)	(0.078)
μ_2	−1.034	−2.325	−1.836	−1.556
	(0.714)	(0.131)	(0.117)	(0.136)
$\ln \sigma_2$	0.580	0.563	0.269	0.431
	(0.155)	(0.025)	(0.028)	(0.032)
	$\tau = 0.7$ *decade*			
First IHD event				
α_1	0.835	0.893	0.706	0.661
	(0.226)	(0.215)	(0.026)	(0.021)
μ_1	−3.918	−4.204	−3.636	−3.381
	(0.291)	(0.178)	(0.071)	(0.053)
$\ln \sigma_1$	0.273	0.529	0.154	−0.330
	(0.310)	(0.109)	(0.079)	(0.135)
Subsequent IHD events				
α_2	−0.429	0.269	0.258	0.154
	(0.300)	(0.025)	(0.022)	(0.025)
γ	8.032	0.937	1.176	1.181
	(∞)	(0.186)	(0.156)	(0.158)
μ_2	−8.884	−2.512	−2.194	−1.827
	(∞)	(0.212)	(0.178)	(0.188)
$\ln \sigma_2$	0.581	0.607	0.301	0.467
	(0.152)	(0.024)	(0.027)	(0.031)

See text for abbreviations. Results for the parameters in the HKM auxiliary equation not shown. An analysis time unit is 10 years. MSL estimation is implemented with $R = 100$, $Q = 10$ and θ^* equal to the modified HKM estimates with $\ln \sigma_1^* = \ln \sigma_2^* = 0$.

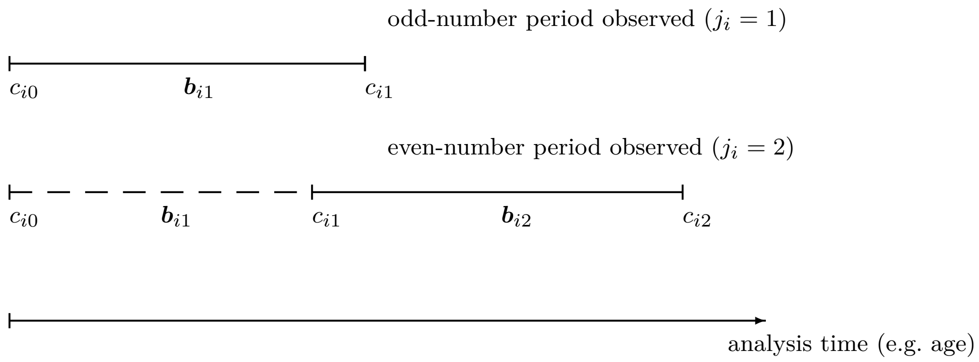

Figure 2.1: Examples of event history data

The two lines in the top describe example event histories, while the line in the bottom indicates analysis time. Dotted line indicates unobserved period, while solid line indicates observed period. The history in the top is right-censored at c_{i1}, while the history in the bottom is left-censored at c_{i1} and right-censored at c_{i2}.

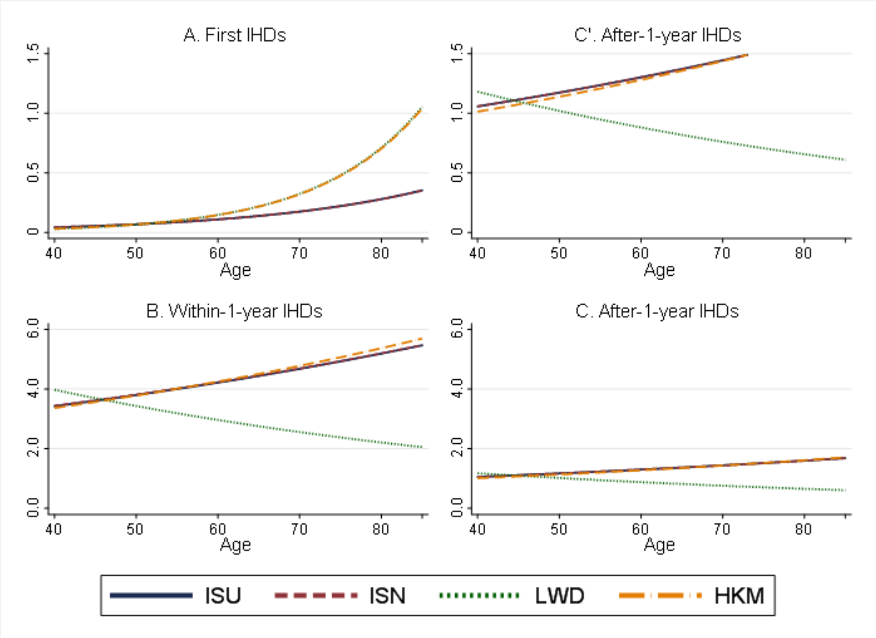

Figure 2.2: Estimated hazard functions for main models without random effects

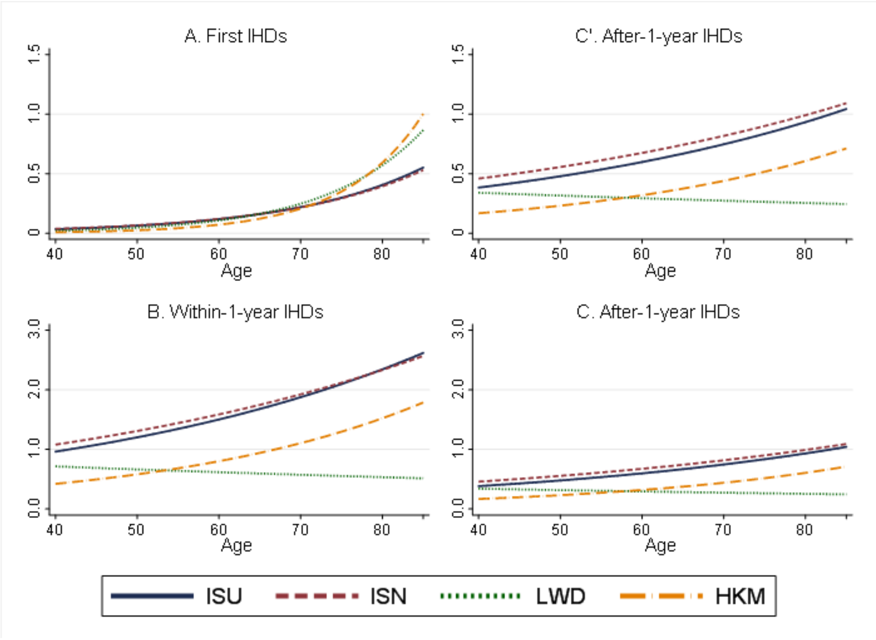

Figure 2.3: Estimated hazard functions for main models with random effects

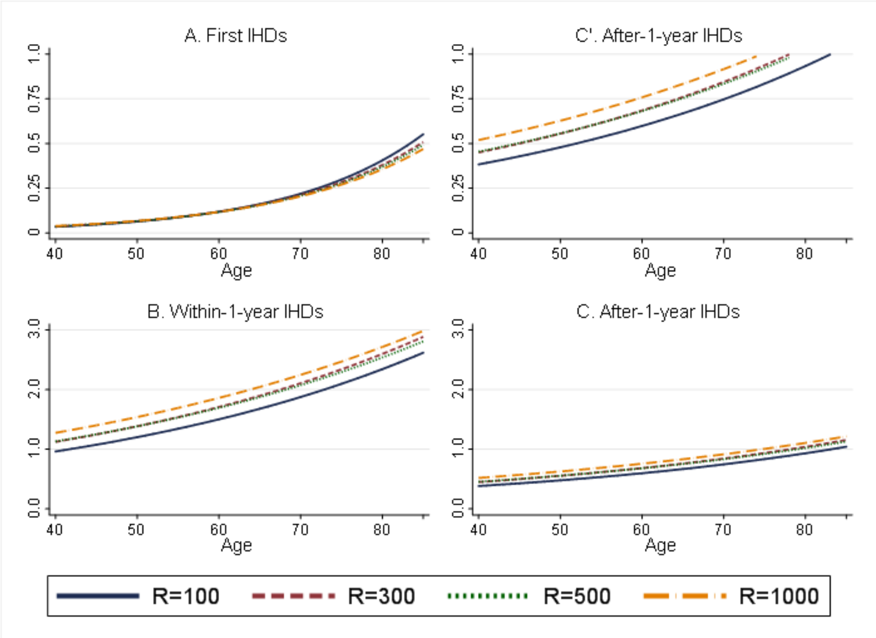

Figure 2.4: Comparing ISU estimated hazard functions for models with random effects

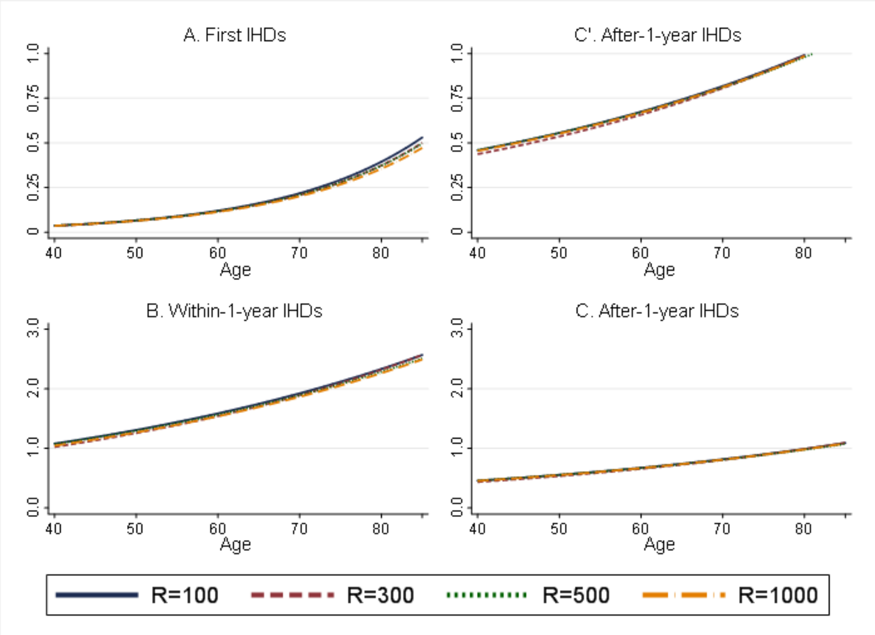

Figure 2.5: Comparing ISN estimated hazard functions for models with random effects

Chapter 3

Heart attack risk in New Zealand: gender, ethnicity, age, and previous heart attacks

3.1 Introduction

In the medical literature, a heart attack is referred to as an acute myocardial infarction (AMI). AMIs are an important public health issue. It has been estimated that the cost of the initial hospitalisation for an AMI in New Zealand is about 4,500 USD for 1999–2001, which is more than twice as large as the total health expenditure per capita of about 1,600 USD in 1999/2000 and about 1,700 USD in 2000/2001.[1] AMI is the most important subclass of the ischemic heart diseases, followed by angina [Morrow, 2017]. In New Zealand, $228 million was spent on treating ischemic heart diseases in hospitals in 2002/2003 [National Health Committee, 2013]. Ischemic heart diseases are a leading cause of death in all of the developed world [Naghavi et al., 2015]. Using administrative data on admissions and death registrations for the period 2002–2012, we estimate that about 10% of all deaths in New Zealand were directly caused by AMI.

This chapter describes and quantifies how the risk of experiencing AMI events varies

[1] Ministry of Health [2012] estimates that health expenditure per capita in New Zealand was 1,600 USD in 1999/2000 and 1,700 USD in 2000/2001 using the concept of purchasing power parities; they do not report figures specifically for AMIs. Based on information from nine countries about diagnosis-related group codes, length of stay, and physician effort, Kauf et al. [2006] estimate that hospitalisation costs was 4,500 USD per event in New Zealand in 1999–2001 with all costs adjusted for 2002 purchasing power parities.

across gender and ethnicity in New Zealand. Our data to be analysed is constructed by combining nationwide administrative data on hospital admissions and death registrations during 2002–2012 with census data from 2001 such that our analysis data have the same population by age as the 2001 census. The data are a kind of unbalanced panel data. Each observation corresponds to an AMI of some person and it provides information on an exact date of the AMI event including the gender and ethnicity of the person.

We consider several aspects of risk. Using event history (hazard) models, we decompose the risk into contributions from age, previous AMI history, and unobserved individual heterogeneity. The decomposition indicates whether the risk is distributed evenly within the population or concentrated among relatively few people, and whether inequality is driven mainly by age, by history dependence, or by unobserved individual heterogeneity. Furthermore, we consider three regimes of risk: the risk of experiencing the first AMI, the risk of a subsequent AMI within 1 year following a previous attack, and the risk of a subsequent AMI more than 1 year after a previous attack. We discuss age-specific risk, and we also compute cumulative outcomes over the age range 40–80.

In general, the estimation of dynamic event history models is hampered by problems of missing data. Often the data available concern the events that happened for a given population during a given period, and there is no information about events that happened before or after the observation period. These problems are called left-censoring and right-censoring, respectively. In the present study, left-censoring means that we do not know which regime an individual is in during the first part of their observation period, because we do not know whether the individual experienced any events prior to their observation period, and if they did, whether that event is within 1 year of becoming under observation. We overcome the left-censoring problem by estimating the models using maximum simulated likelihood (MSL) methods developed by Lee and Gørgens [2017].[2] The idea is that the same model that is used to explain the observed patterns in the data can be used to simulate events that are unobserved.

[2] Lee and Gørgens [2017] is an earlier version of Chapter 2 in the thesis.

Intuitively, the missing data are integrated out of the so-called complete-data likelihood function by simulating pseudo-histories for each person whose history is left-censored. In principle, the simulation technique allows us to compute an arbitrarily good approximation to the exact likelihood function for the data that are observed. In practice, available computing resources and time restrict the feasible accuracy. Regarding right-censoring, we follow common practice and assume that individual observation periods are exogenously determined.

We analyse AMI history data for four groups, namely male and female people of Maori and European descent. Our main finding is that there are large gender and ethnic disparities in AMI risk. The general ranking is that male Maoris tend to have the highest risk, followed by female Maoris, then male Europeans, and finally female Europeans have the lowest risk. The exceptions are that female Maoris and male Europeans have similar risk-levels for the first event, and that the Europeans catch up and overtake the Maoris after age 75 for the risk of events within 1 year of a previous event and after age 80 for the risk of events more than 1 year after. The risk increases strongly with age. This partly reflects biological effects as bodies become older and partly time effects as different cohorts have been exposed to different environments, made different life style choices, and had access to different medical technologies. Regarding history dependence, in terms of the three regimes the risk is lowest for the first AMI event, highest for events within 1 year after an event, while still high for events more than 1 year after an event. In particular, for people below the age of 70, the risk of a subsequent event is at least twice the risk of the first event. Finally, it is notoriously difficult to obtain reliable estimates of the influence of unobserved heterogeneity, but our results suggest that the within-group variation in risk is far greater than the between-group differences since variations in risk from random effects is larger than variations from the other sources.

Our modelling framework permits us to consider life-time perspectives, in addition to age-specific outcomes. Using the estimated models, we run dynamic simulations of individual cumulative outcomes between ages 40 and 80, assuming no one dies. For the extensive margin, we find that the overall proportion of people experiencing at least one AMI event by age 80 is about 35% for male Maoris, 27% for female Maoris, 28% for male Europeans, and about 16% for female Europeans. For the intensive margin, we find that the overall average number of AMI events between ages 40 and 80 for those who have at least one event is 3.7 for male Maoris, 4.7 for female Maoris, 3.1 for male Europeans, and about 2.8 for female Europeans. The high average for female Maoris is due to a small proportion of female Maoris with extremely high risk. When we compare cumulative outcomes for people at the first, second, or third quartiles of the distributions of unobserved heterogeneity, we find that male Maoris expect the highest number of events, followed by female Maoris and male Europeans whose outcomes are similar, and finally female Europeans expect the lowest number of events.

A limitation of our study is that the data do not allow us to explore the factors that cause disparities across gender and ethnicity, such as biological, socioeconomic, behavioural factors, etc. However, we allow for unobserved individual heterogeneity ('random effects' in the econometrics literature, 'frailty' in the statistics literature), and this may capture some of these factors so that the estimated risk distribution is representative. Since we aim to understand the incidence of AMI rather than case fatality, we simplify the analysis by assuming that the individual observation periods are exogenous. In most cases, the observation period ends when the study period ends on 30 June 2012, but some individuals die before this date and mortality risk and AMI risk are likely to be correlated. However, the death rate is small, so we expect the resulting bias to be negligible. Note that since we observe the cause of death, AMI events are not systematically underreported in the data.

The rest of the chapter is organised as follows. Section 3.2 provides a review of related literature. In Section 3.3, we explain the data construction and describe our analysis data. In Section 3.4, we discuss the model specification and the estimation methods. In Section 3.5, we present and discuss the estimation results. Section 3.6

concludes.

3.2 Literature

The literature on gender and ethnic differences in the incidence and reoccurrence of AMIs is relatively small. Wang et al. [2012] compare AMI incidence rates in the US and discover that male whites have the highest risk, followed by male blacks, then female blacks, and female whites have the lowest risk. Smolina et al. [2012] find a gender gap for first AMIs but no gap for subsequent AMIs in the UK. For New Zealand, Chan et al. [2008b] also distinguish between first and subsequent AMIs and find that the rate of AMI readmissions per 100,000 population increases during the 1990s; however, they do not consider gender and ethnic differences. The policy implications of high incidence rates for first and subsequent AMI are different: the former supports primary prevention and the latter supports secondary prevention [Avendano and Soerjomataram, 2008]. The distinction between the first and subsequent AMIs is becoming more important as more people survive AMIs and are at risk of having subsequent AMIs. Chan et al. [2008a] study differences in AMI prevalence across gender and ethnic groups in New Zealand, but do not distinguish between first and subsequent AMIs.

The literature on gender and ethnic differences in mortality during the first 30 days or 1 year after an AMI event is larger than the literature on AMI incidence itself. Gender disparities in mortality have been studied for many countries using different kinds of data sets; e.g. nationwide data on Finland [Kytö et al., 2015], Israel [Gottlieb et al., 2000], Scotland [MacIntyre et al., 2001], and England [Smolina et al., 2012], city-level data in Germany [Herman et al., 1997], and hospital-level data in Vietnam [Nguyen et al., 2014].[3] Ethnic disparities in mortality have been studied for pertinent countries; for example, disparities between blacks and whites for the US [Vaccarino et al., 2005], between people of European, Chinese, and South Asian descent for Canada [Anand

[3] Further, some studies in the literature compare across countries; e.g. Abildstrom et al. [2003] compare Denmark and Sweden and Tunstall-Pedoe et al. [1994] compare 21 countries. Gender and ethnic disparities are also found in other areas of public health; e.g. health care utilisation [Card et al., 2008], provider practice [Currie et al., 2016], and obesity [Zhang and Wang, 2004].

et al., 2000], and between Chinese, Malay, and Indians for Singapore [Mak et al., 2003].

While we do not consider mortality in this chapter, it is interesting to note that the patterns in AMI incidence and AMI case fatality are not necessarily the same across gender, ethnicity, and age. For example, Alderman et al. [2000] find that young black males in the US have lower risk of AMI events than do young white males, but higher 30-day case fatality rates. Comparing US studies that focus on AMI incidence (e.g. Wang et al., 2012) with those on case fatality (e.g. Manhapra et al., 2004) reveals further instances where relatively low/high AMI incidence rates are associated with opposite high/low case fatality rates.

The statistical methods used in this literature include mean comparison, logistic regression, Kaplan-Meier estimation, and Cox regression. These methods are appropriate for summarising outcomes when there are no issues of missing data other than right-censoring. For this reason, the literature has generally focused on outcomes that are fully observed during a relatively short period, say 30 days or 1 year, following an AMI event (e.g. Chang et al., 2006; Pokorney et al., 2012). A few small-scale follow-up studies have been able to track patients for several decades (e.g. Klein et al., 1992). Sometimes models of AMI risk allow for unobserved heterogeneity among families and hospitals. We have found only one study that has considered individual-level unobserved heterogeneity [Hougaard, 1986].

Our study contributes to the literature in several dimensions. This is the first study to examine gender and ethnic disparities in AMI risk utilising a hazard approach. We analyse high-quality nationally representative data from New Zealand. We show that hazard models, in general, can be estimated using the MSL method, despite overwhelming left-censoring. We use the estimated models to discuss and compare different aspects of risk including age dependence, dynamic effects of the AMI history, and the role of unobserved individual heterogeneity. We also use the estimated models to examine the extensive and intensive margins of cumulative lifetime outcomes for representative persons and for synthetic cohorts.

3.3 Data

Our primary data source is the National Minimum Data Set (NMDS) provided by Ministry of Health New Zealand. The NMDS includes administrative data on all hospital admissions in New Zealand (including both in-patients and day-patients). The original NMDS was created in 1993. The current format with 20 diagnosis entries was introduced in June 2002 [National Health Board Business Unit, 2011]. Our study period is between 1 July 2002 and 30 June 2012. We merge the NMDS with death registrations.[4] The latter include information about the cause of death. These administrative data are well suited for studying the incidence of AMI events, because everyone who experiences an event usually present at an emergency department and are admitted to the hospital, or they die on the way to the hospital and so appear in the death data.

About 60% of New Zealanders appear in the administrative data during the study period. For example, there are 2.7 million people alive in the data in 2012 compared to 4.4 million estimated total population in 2012. We use the age distribution in the census conducted on 6 March 2001 to add records for people with no hospital admission during the study period. As a result, the combined data set has the same age distribution in 2002 as the 2001 census.[5]

Diagnosis and cause of death codes follow the *International Classification of Diseases 10 Australian Modification* (ICD-10-AM). According to ICD-10-AM, codes I21 and I22 are 'acute' and 'subsequent' myocardial infarctions, respectively, where subse-

[4] The NMDS and the death registrations share unique (confidentialised) individual identification numbers.

[5] The New Zealand censuses report population figures by age, gender, and ethnicity in five-year age intervals until age 85 plus a single interval for those older than 85. We use the age distribution in the 2001 census to add records representing people who were not admitted to a hospital during the study period. We first convert the age intervals at the census date 6 March 2001 to age intervals in 1 July 2002. We then compute the difference between the census and the administrative data for each interval, and add entries for people with no administrative records. The birthdays for the added entries are uniformly distributed within each age interval. For the open age interval, birthdays are uniformly distributed between age 86 years and 4 months to age 91 years and 4 months (their ages on 1 July 2002). People born before 6 March 1911 are dropped from the administrative data, and people aged 85 years or older are assumed to be less than 90 in the census data.

quent here means within 4 weeks. We do not distinguish between I21 and I22, because apparently the codes are not used consistently in our data. The hospital data contain up to 20 diagnostic entries (for recording complications) for each admission, while the death data have a single entry for the cause of death. We regard an admission and a death with a code I21 or I22 in any of the entries as a distinct AMI event if there is no AMI event within the last 29 days. That is, AMIs that occur within 29 days of each other are considered a single event. There are three reasons for this. First, the risk of subsequent AMIs are highly elevated during the first 29 days after an AMI event [Lee et al., 1995]. Second, according to the pathological classification an infarct is considered healed after 29 days [Steg et al., 2012]. Third, many studies in the literature view any hospital or death record within 30 days after an AMI occurs as related to the same AMI (e.g. Smolina et al., 2012).

The combined data set includes date of birth, gender, ethnicity, date of admission and diagnosis codes if admitted, and date of death and cause of death code if died. Gender is the biological sex reported. Ethnicity is self-identified.[6] Date of admission is when patients are first seen by clinicians at a hospital. With the 29-day caveat, we use time of admission or time of death as the timing of AMI events.[7]

To define the study population, we restrict the combined data set as follows. First, we consider only European and Maori people.[8] In the 2001 census, people of European and Maori descent constitute 83% and 14%. Second, we exclude individuals with type 1 diabetes (about 0.1% of the Maoris and Europeans in the combined data set), whose experiences are expected to be different. For estimation, we further restrict the data to people over the age of 40 and under the age of 85. AMIs are quite rare before age 40. These rare events are less important from a public health perspective, and

[6] In the few cases where date of birth, gender, or ethnicity change between admissions we use the values reported at the last admission. For ethnicity, less than 5% of people in the combined data have different codes across admission; for birth dates and gender, less than 0.1% have different codes.

[7] When multiple AMI events are treated as a single event with the 29-day caveat, we use the earliest date as the timing of the combined event.

[8] The underlying question used to obtain ethnicity is the same in the NMDS data and 2001 census (see Cormack and Robson, 2010). Healthcare users and census respondents can choose up to three ethnicities. In our data, the responses have been prioritised roughly in the order of Maori, Pacific, Asian, others, European.

ignoring them simplifies the analysis. There is no detailed information about the age distribution for those over 85 in the 2001 census and the sample sizes are small. It is therefore difficult to estimate risk for people over age 85. We refer to the study population over 40 and under 85 as the estimation sample.

Migration in and out of New Zealand is substantial, but we expect that this is less of an issue for our analysis. Temporary emigration is common especially among young people, but they are less likely to suffer AMIs. Also immigration of people over the age of 40 of European and Maori descent is relatively low.[9]

Table 3.1 shows summary statistics for our study population and the estimation sample. By construction, the study population is essentially representative of the New Zealand population of Maori and European descent as of the census date, 6 March 2001, except people born between 6 March 2001 and 1 July 2012 are added (only) if they have been admitted to a hospital after 1 July 2002. These added people are younger than age 40 by 2012 and hence are not used for estimating the hazard models.

The first panel in Table 3.1 shows the study population by age on 1 July 2002. It is clear that there are relatively fewer older people so the estimates for older people will be more noisy. The next two panels show the number of people under age 40 and over age 40 during their observed period by their number of observed AMIs. While AMI events are quite rare for people under age 40, this is not the case for people over age 40. Note that these are only the events that occur during the study period, and many people will have experienced events before 2002 and after 2012. These events are not observed in the data; these are the left- and right-censoring problems mentioned in the Introduction. The next panel shows the average number of observed AMIs for people who experienced at least one AMI in their observed period. The average number increases as people age, from about 1.1–1.3 for people aged 30–39 to 1.5–1.6 for people over 70. The last panel shows the number of people who died during their observed period; overall, it is about 8.6% of people in the estimation sample. As mentioned,

[9] According to the 2001 census, about 7% of people aged 30–65 and 2% of people over age 65 lived overseas five years ago. Figures by ethnicity are not available.

right-censoring is less of an issue in this study, since we have data for most people until the end of the study period on 30 June 2012 and we observe the cause of death for those who die before that date.

Figure 3.1 compares the age-specific AMI incidence rate across gender and ethnic groups. The incidence rate is computed as the ratio of the number of observed AMIs to the total time 'at risk' (in years) in two-year age intervals. For ages under 80, Maoris have higher AMI incidence rates than Europeans and males have higher rates than females. In particular, male Maoris have the highest rates, while female Europeans have the lowest rates. Female Maoris and male Europeans have similar rates. After age 80, however, the gender and ethnic disparities become less clear. Partly, this is because the number of people over 80 in the sample is small, so the estimates are noisy.

The incidence rates shown in Figure 3.1 provide a snapshot of the age-specific risk of having an AMI event across the four subpopulations. We now turn to econometric modelling in order to shed light on how events are distributed within each subpopulation.

3.4 Model and estimation

In this section, we discuss our model specification and estimation methods. The most important risk factor is age. As mentioned, age also captures time and cohort effects. Therefore, we use age as analysis time. Specifically, we define analysis time as $t = (\text{age} - 40)/10$ for age > 40. (Normalising the time unit to a decade makes the scale of certain parameters more readable.) Also, it is well known that there are dynamic patterns in risk. Those having experienced an AMI event are more likely to experience subsequent events, and the risk is particularly high for some time immediately after that event. Therefore, we specify separate models for the first and subsequent AMI events, and the equation for subsequent AMIs is allowed to depend on the timing of the most recent event. Heterogeneity in risk can be considerable. Unfortunately, we do not have risk markers in our data, be they biological, socioeconomic, or behavioural factors. Therefore, we include so-called random effects (frailty) to capture the effect of unobserved individual heterogeneity. Specifically, we include an unobserved random variable $v \sim N(0, 1)$ in the model specifications. To allow for more flexibility, we estimate a separate model for each gender and ethnic group. For notational simplicity, we suppress subscripts indicating the groups in the following.

Each model consists of two equations. The first equation represents the hazard function, h_1, of the first AMIs:

$$h_1(t|v, \theta) = \exp(t\alpha_1 + \mu_1 + v\sigma_1), \tag{3.1}$$

where θ denotes the entire unknown parameter vector to be estimated. Parameter α_1 captures age dependence in the risk, parameter μ_1 captures the median overall level of risk for first AMIs, and parameter σ_1 is the influence of the random effect. The second equation represents the hazard function h_2, of subsequent AMIs:

$$h_2(t|t^-, v, \theta) = \exp(t\alpha_2 + \mathit{Recent}\gamma + \mu_2 + v\sigma_2), \tag{3.2}$$

where t^- is the timing of the most recent AMI and the variable *Recent* is defined by $\mathit{Recent} = 1(t \leq t^- + \tau)$ where the value of τ corresponds to 1 year (i.e. $\tau = 0.1$). Parameter α_2 captures age dependence in the risk, parameter γ indicates the dynamic effect of the most recent AMI, parameter μ_2 embodies the median overall level of risk of subsequent AMIs, and parameter σ_2 is the influence of the random effect.

The Gompertz specifications embodied in (3.1) and (3.2) assume that the hazard function progresses exponentially with age. The law of exponential progression is suitable for many common age patterns in actuarial, biological, and demographic applications (e.g. Wienke, 2010). It is also appropriate in the context of AMI risk until age 85, as shown in Figure 3.1. We expect positive signs of α_1 and α_2 given that AMI risk increases as people age. Note that analysis time is not reset after an AMI event.

We capture history dependence partly by distinguishing between h_1 and h_2 and partly by including the time-varying covariate *Recent*. The latter allows for elevated risk proportional to e^{γ} within 1 year following the most recent event. The cutoff between the two regimes for the risk of subsequent events is somewhat arbitrary but follows the literature.[10] There is no theoretical basis for assuming an abrupt change in risk after 1 year, but this specification allows us to distinguish short-term and long-term risks in a simple way.

Since we estimate separate models for each group, effectively all parameters are interacted with gender and ethnicity. In particular, gender and ethnicity are not assumed to have a simple proportional effect on risk. Group-specific parameters mean that differences in outcomes can arise because of a combination of differences in age dependence, in the dynamic effect of the most recent AMI, and in the distribution of the random effects.

As mentioned, the main problem for estimating the models is left-censoring. For people whose histories are left-censored, we cannot tell whether the first observed AMI is the first experienced AMI or a subsequent AMI, nor do we know the value of *Recent* for the first 1 year of the observation period. Comparing the number of people aged 30–39 and 30–85 years in Table 3.1 reveals that about 75% of the AMI histories in our analysis data are left-censored, so the problem is substantial.

Left-censoring and history dependence mean that the likelihood function for the observed data is analytically intractable. Therefore, we estimate the models using the maximum simulated likelihood (MSL) method developed by Lee and Gørgens [2017]. To discuss this method some additional notation is needed. Let $C_i = 0$ indicate that the history for individual i is not left-censored, and let $\mathbf{b}_{i1} = (b_{i1k_{i1}}, \ldots, b_{i11})$ denote their event history where each b_{i1k} is the analysis time when person i had event k and k_{i1} is their total number of events (possibly 0). Persons with $C_i = 0$ are under age 40 on 1 July 2002, and $\mathbf{b}_{i1}$ is the analysis time of their AMI events from the date they turn 40 until 30 June 2012 or until the analysis time of their death, whichever

[10] Many studies in the literature consider mortality during 1 year after an AMI event (see Introduction).

is earlier. Let $C_i = 1$ indicate that the history for individual i is left-censored, and let $(\mathbf{b}_{i2}, \mathbf{b}_{i1}) = (b_{i2k_{i2}}, \dots, b_{i21}, b_{i1k_{i1}}, \dots, b_{i11})$ denote their event history, where $\mathbf{b}_{i2}$ is observed and $\mathbf{b}_{i1}$ is unobserved. Persons with $C_i = 1$ are those who are over age 40 on 1 July 2002, and $\mathbf{b}_{i2}$ is the analysis time of their AMIs from 1 July 2002 until 30 June 2012 or until their analysis time of death, while $\mathbf{b}_{i1}$ is the analysis time of their AMIs from age 40 to 1 July 2002.

Let g_1 and g_2 denote density functions of $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$. They can be derived from the hazard functions given in Equations (3.1) and (3.2). To state the expressions formally, let b_{i10} denote the beginning of analysis time, let b_{i20} denote analysis time on 1 July 2002, and let b_{i30} denote analysis time on 30 June 2012 or on the date of death. Furthermore, let $H_1(t|t^-, v, \theta) = \int_{t^-}^{t} h_1(y|v, \theta)\, dy$ for $t > t^-$ denote the value of the cumulative hazard function from time t^- until time t. Similarly, define $H_2(t|t^-, v, \theta) = \int_{t^-}^{t} h_2(y|t^-, v, \theta)\, dy$ for $t > t^-$. Then the density g_1 of $\mathbf{b}_{i1}$ evaluated at $\mathbf{b}_1$ when $k_1 > 1$ is

$$\begin{aligned} g_1(\mathbf{b}_1|v, \theta) &= h_1(b_{11}|v, \theta) \exp\Big(-H_1(b_{11}|b_{10}, v, \theta)\Big) \\ &\times \left(\prod_{k=2}^{k_1} h_2(b_{1k}|b_{1,k-1}, v, \theta) \exp\Big(-H_2(b_{1k}|b_{1,k-1}, v, \theta)\Big)\right) \exp\Big(-H_2(b_{20}|b_{1k_1}, v, \theta)\Big). \end{aligned} \tag{3.3}$$

If $k_1 = 1$, the product over k in the middle is void, and if $k_1 = 0$, then only the very last exponential term is present with H_1 replacing H_2. When $k_1 = 0$ (so $\mathbf{b}_1$ is empty) and $k_2 > 1$, the conditional density g_2 of $\mathbf{b}_{i2}$ given $\mathbf{b}_{i1} = \mathbf{b}_1$ evaluated at $\mathbf{b}_2$ is

$$\begin{aligned} g_2(\mathbf{b}_2|\mathbf{b}_1, v, \theta) &= h_1(b_{21}|v, \theta) \exp\Big(-H_1(b_{21}|b_{20}, v, \theta)\Big) \\ &\times \left(\prod_{k=2}^{k_2} h_2(b_{2k}|b_{2,k-1}, v, \theta) \exp\Big(-H_2(b_{2k}|b_{2,k-1}, v, \theta)\Big)\right) \exp\Big(-H_2(b_{30}|b_{2k_2}, v, \theta)\Big), \end{aligned} \tag{3.4}$$

and when $k_1 > 0$ and $k_2 > 1$ we have

$$\begin{aligned} g_2(\mathbf{b}_2|\mathbf{b}_1, v, \theta) &= h_2(b_{21}|v,\theta)\exp\Big(-H_2(b_{21}|b_{1k_1}, v, \theta)\Big) \,/\, \exp\Big(-H_2(b_{20}|b_{1k_1}, v, \theta)\Big) \\ &\times \left(\prod_{k=2}^{k_2} h_2(b_{2k}|b_{2,k-1}, v, \theta)\exp\Big(-H_2(b_{2k}|b_{2,k-1}, v, \theta)\Big)\right)\exp\Big(-H_2(b_{30}|b_{2k_2}, v, \theta)\Big). \end{aligned} \tag{3.5}$$

The modifications for individuals with other values of k_2 are relatively straightforward; see Lee and Gørgens [2017] for details.[11]

With these definitions, and letting Φ denote the standard normal cumulative distribution function, the log likelihood function for N observed histories can be written[12]

$$\begin{aligned} L(\theta) = \sum_{i=1}^{N} \Bigg[&(1 - C_i) \ln \int_{\mathbb{R}} g_1(\mathbf{b}_{i1}|v, \theta)\, d\Phi(v) \\ &+ C_i \ln \int_{\mathbb{R}} \int_{\text{Support}(\mathbf{b}_1)} g_2(\mathbf{b}_{i2}|\mathbf{b}_1, v, \theta)\, g_1(\mathbf{b}_1|v, \theta)\, d\mathbf{b}_1\, d\Phi(v) \Bigg]. \end{aligned} \tag{3.6}$$

The first term in the sum on the right-hand side of Equation (3.6) is the likelihood contribution if individual i is non-left-censored. The integral here is over the random effect. The second term is the likelihood contribution if individual i is left-censored. Here the outer integral is over the random effect and the inner integral is over the unobserved history.

In Equation (3.6), we assume that right-censoring and AMIs events are conditionally independent given previous event history, and we do not model the process of right-censoring explicitly. As mentioned, most individuals are right-censored because the study period ends on 30 June 2012, but a small number are right-censored when they die before 30 June 2012. If mortality risk and AMI risk are correlated (e.g. competing risks

[11] It is necessary to keep track of whether (h_1, H_1) or (h_2, H_2) applies as well as the timing of the most recent event t^-.

[12] Lee and Gørgens [2017] consider a more general setup than is necessary here. For example, they allow for multiple observation periods for each individual. In the present application, there is a single observation period from 1 July 2002 until the earlier of 30 June 2012 and date of death. For simplicity, we here use a simple indicator variable C_i to represent observed and unobserved periods. In the terminology of Lee and Gørgens [2017], $C_i = 0$ corresponds to the case where odd-numbered periods are observed and $C_i = 1$ the case where even-numbered periods are observed.

correlated through the random effects), then the likelihood function is misspecified. However, the misspecification bias is likely to be small since the death rate is small. Note that right-censoring due to death here does not cause missing data, as the cause of death is observed in all cases.

There are no closed-form solutions to the integrals and analytical evaluation of the likelihood function is not possible. The solution investigated by Lee and Gørgens [2017] is to use a combination of quadrature and simulation methods to evaluate $L(\theta)$. The integrals over the random effects are one-dimensional and can be handled by e.g. Gaussian quadrature. The integral over the unobserved history is difficult to evaluate, essentially because the dimension of $\mathbf{b}_1$ is unknown. To handle that, we consider two importance sampling simulation methods, unnormalised (ISU) and normalised (ISN). For the ISU method, the simulated log likelihood function that we maximise is

$$\begin{aligned} L(\theta) \approx \sum_{i=1}^{N} \Bigg[&(1 - C_i) \ln\left(\sum_{q=1}^{Q} w_q \, g_1(\mathbf{b}_{i1} | v_q, \theta) \right) \\ &+ C_i \ln\left(\sum_{q=1}^{Q} w_q \, \frac{1}{R} \left\{ \sum_{r=1}^{R} g_2(\mathbf{b}_{i2} | \mathbf{b}_{i1}^{qr}, v_q, \theta) \, \frac{g_1(\mathbf{b}_{i1}^{qr} | v_q, \theta)}{g_1(\mathbf{b}_{i1}^{qr} | v_q, \theta^*)} \right\} \right) \Bigg], \end{aligned} \quad (3.7)$$

where the v_qs are Gauss-Hermite quadrature points and the w_qs are the corresponding weights, and the $\mathbf{b}_{i1}^{qr}$s are simulated pseudo-event histories. The idea of importance sampling is to draw $\mathbf{b}_{i1}^{qr}$ from $g_1(\cdot|v_q, \theta^*)$ using a fixed θ^* instead of drawing from $g_1(\cdot|v_q, \theta)$ using the θ at which the likelihood function is evaluated, and correct the 'mismatch' through the adjustment factor $g_1(\mathbf{b}_{i1}^{qr}|v_q, \theta)/g_1(\mathbf{b}_{i1}^{qr}|v_q, \theta^*)$.[13] One of the advantages of using importance sampling is that the simulated likelihood function is continuous in θ, so gradient-based algorithms can be used to find the maximum. Since event timings depend on prior history, it is not possible to draw an entire history $\mathbf{b}_{i1}^{qr}$ from $g_1(\cdot|v_q, \theta^*)$ in a single step. Instead, it is necessary to draw the individual

[13] In our empirical analysis, we set $Q = 10$ and $R = 100$. For θ^*, we use estimates obtained using Heckman's approach as discussed in Lee and Gørgens [2017], with the modification that $\ln \sigma_1^* = \ln \sigma_2^* = 0$. When estimating a model without random effects for female Europeans, Heckman's approach resulted in non-sensible estimates, so we substituted the estimates for female Maoris. (Using estimates for male Europeans gave similar final estimates.)

pseudo-event timings sequentially; see Lee and Gørgens [2017] for details.

For the ISN method, the simulated log likelihood function is essentially the same; the only difference is that the adjustment factors are normalised so that they sum to R. That is, $g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta)/g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta^*)$ in Equation (3.7) is replaced by

$$\frac{g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta)}{g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta^*)}\bigg/\frac{1}{R}\sum_{r=1}^{R}\frac{g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta)}{g_1(\mathbf{b}_{i1}^{qr}|v_q,\theta^*)}. \tag{3.8}$$

Since ISU and ISN are just different ways of approximating the exact likelihood function in Equation (3.6), both methods should provide similar results. In practice, there may be some differences, and we report estimates from both methods in the discussion of the results.

The MSL estimation method is computationally burdensome, because a large number of draws is required in order to obtain a satisfactory approximation to the exact likelihood function for the observed data (large Q and large R). However, as we show in related research, MSL estimation is more efficient in handling the left-censoring problem than the ad hoc solutions that previously have been considered in the literature [Lee and Gørgens, 2017]. In the present context, this is particularly true for the risk of the first event.

3.5 Results

3.5.1 Age-specific risk

Table 3.2 reports parameter estimates using both the ISU and ISN methods.[14] The two methods give estimates that are very similar, if not exactly the same. We begin with a discussion of the ISU estimates, and comment on the ISN estimates at the end.

Recall that α captures differences in the levels of AMI risk across age groups, which here represent the effects of biological age as well as time and cohort effects. The

[14] Since the full samples are too large for our limited computing resources, we base our estimation on random sub-samples of 50,000 individuals in each gender and ethnic group. Standard errors are computed as the outer product of the score functions.

estimates of α_1 in Table 3.2 are all positive, implying that older people have higher risk than younger people in the same group. The magnitude of the age effects is broadly similar for all groups, although largest for female Europeans. Recall that μ captures the median overall levels of AMI risk at age 40. The estimates of μ_1 in Table 3.2 are small (large negative), reflecting the low but not quite zero incidence rates around age 40 (c.f. Figure 3.1). The estimates of μ_1 are not the same, however, and imply that male Maoris have the highest risk of experiencing the first event, followed by male Europeans and female Maoris whose risk is similar, while young female Europeans have the lowest risk. Since the estimate of μ_1 is smallest and the estimate of α_1 is largest for female Europeans, it is possible that the risk gap between female Europeans and the other groups vanishes for older people. However, as shown below, the differences in the estimates of μ_1 are too large and the differences in α_1 are too small for this to happen within ordinary human life times. Recall that σ captures the influence of unobserved heterogeneity. The estimates of $\ln \sigma_1$ are largest for male Europeans and smallest for female Europeans.

The estimates of α_2 indicate that the risk for Europeans increases with age at about double the rate than the risk for Maoris. The estimates of γ are similar for all four groups and suggest that the risk during the first 1 year after an event is more than twice as large as the risk more than 1 year after the event ($e^{0.9} \approx 2.5$). The parameter μ_2 essentially reflects the median risk of a subsequent event at age 40 for people who had an event at or before age 39. Although this is out of sample, note that the estimates tend to be a bit higher than the estimates for μ_1. The estimate of $\ln \sigma_2$ is largest for female Maoris and smallest for male Maoris.

Table 3.2 also reports χ^2 statistics of the joint null hypotheses that the respective parameters are the same across all four groups. The null is rejected for α_1, α_2, μ_1 and μ_2, but not for γ, $\ln \sigma_1$ and $\ln \sigma_2$.

Since the model is non-linear and it is difficult to interpret some of the parameters, we translate the estimates in Table 3.2 into hazard rates. Figure 3.2 shows the estimated hazards for the four groups, evaluated at the median of the random effects.[15] A general ranking appears: male Maoris tend to have the highest risk, followed by female Maoris, then male Europeans, and finally female Europeans have the lowest risk. The exceptions are that female Maoris and male Europeans have similar risk-levels for the first event, and that the Europeans catch up and overtake the Maoris after age 75 for the risk of events within 1 year of a previous event and after age 80 for the risk of events more than 1 year after.

Regarding the three regimes of risk, it is clear from Figure 3.2, that the risk of subsequent events within 1 year of an event is much larger than the risk of the first event. The risk after 1 year is also higher, although the difference seems small for older Maoris.

A different view is provided in Figure 3.3, which shows the risk of subsequent events relative to the risk of the first event for the four groups, evaluated at the median of the random effects. In our flexible two-equation systems, the effect of having the first event is not restricted to be proportional to a given baseline risk. The relative change in risk before and after the first event depends on age, history, and unobserved heterogeneity:

$$\frac{h_2(t|t^-, v, \theta)}{h_1(t|v, \theta)} = \exp\Big(t\{\alpha_2 - \alpha_1\} + \mathit{Recent}\gamma + \{\mu_2 - \mu_1\} + v\{\sigma_2 - \sigma_1\}\Big). \tag{3.9}$$

Figure 3.3 shows that the relative risk of subsequent events is extremely large for young Maoris, but falls steeply with age. The effect of age is negative, because the estimates imply $\alpha_2 < \alpha_1$. The relative risk is also large for Europeans, and falls with age at a much slower rate. The nearly-proportional effect arises because the estimates imply $\alpha_2 \approx \alpha_1$. These conclusions are supported by Wald tests of the restrictions $\alpha_2 = \alpha_1$, which are rejected for Maoris ($p = 0.00$) while not rejected for Europeans ($p > 0.20$).

It is well known that random effect distributions are difficult to estimate, so the

[15] Note that panels C and C' in Figure 3.2 are the same except for the scale of the vertical axis. Comparing the hazard functions for subsequent events across ages is slightly tricky. For example, in panel C in Figure 3.2 the hazard rate at age 70 assumes a person has experienced at least one previous event and that event is before age 69. The rate at age 80 assumes the person's most recent event is any time before 79.

estimates of $\ln \sigma_1$ and $\ln \sigma_2$ may not be very reliable. Nevertheless, the estimates may provide a ballpark measure of the influence of unobserved heterogeneity. Figures 3.4, 3.5, and 3.6 show the median hazard rates together with the first and third quartile of the random effects. The main impression from the graphs is that there are enormous differences between high- and low-risk people. Presumably it would be possible to explain or reduce this variation if data were available on individual biology, life style choices, etc.

The flexibility of our models means that the ranking of risk across regimes and across groups may not be the same throughout the distribution of random effects. Figure 3.7 provides graphs equivalent to Figure 3.2, but evaluated at the 95th percentile of the distribution of random effects. Not surprisingly, the risk-level is very high. While the risk of a subsequent event within 1 year is still highest in all cases, the risk of the first event is higher than that of subsequent events more than 1 year after for male Maoris after age 65 and for male Europeans at all ages. Also, the risk of subsequent events more than 1 year after a previous event for female Maoris exceeds that of the other groups.

Finally, comparing the ISU and ISN estimates in Table 3.2, we see that the differences are very small for male and female Maoris. Importantly, this is also true for the median estimated hazard functions, see Figure 3.8. For Europeans, the differences in the parameter estimates for the risk of the first event are larger, but the median estimated hazard functions are actually quite similar. In particular, the large differences in the estimates of $\ln \sigma_1$ do not affect the medians. For Europeans' risk of subsequent events, the ISN estimates of μ_2 tend to be higher and the estimates α_2 tend to be lower than the ISU estimates. The net result is that the ISN estimates of the median estimated hazard functions tend to be slightly higher, especially between ages 65 and 85. As in our earlier work [Lee and Gørgens, 2017], apparently the differences between the ISU and ISN estimates are caused by the inherent difficulty in estimating random effects models. When we estimate models without random effects, the ISU and ISN estimates are identical for all practical purposes (see Appendix Table 3.A1). The estimated models without random effects also imply unrealistically large risk of subsequent events, about three times as large as the median risk in models with random effects, which is why we do not further discuss them in this chapter.

3.5.2 Cumulative risk

So far we have compared outcomes across gender and ethnic groups in terms of age-specific risk. The estimated models also allow us to compare cumulative, life-time outcomes. We now discuss estimates of the average number of AMI events that a representative person or a population will experience between ages 40 and 80.

Such an exercise raises two questions. First, the model is estimated on a cross-section and may not be representative of any particular birth cohort. There may have been significant changes in both life styles that affect risk and in the medical know-how in treating symptoms and preventing disease. However, as in other areas of demographic analysis, synthetic cohort analysis provides a useful summary of outcomes during a given period of time. Besides, we are comparing four subpopulations which have lived in similar environments and experienced similar changes.

The second question is what to do about differences in mortality across the four groups. In reality, people do not live forever and mortality is likely to be related to both a person's innate frailty and to their previous medical history. It would be possible to augment the current model with an equation representing mortality risk. However, here we proceed by assuming no one dies until age 80, because this provides a better foundation for comparing the risk of AMI events. Were we to implement group-specific non-independent mortality, the internal composition of the four groups in terms of high-risk and low-risk people would vary over time at differential rates. As a result, the predicted outcomes would partly reflect differences in AMI risk and partly differences in mortality (essentially a 'selection' effect).

We compute cumulative outcomes by dynamically simulating AMI events for each gender and ethnic group over the age range from 40 to 80. Technically, the simulation method is the same as that used internally for computing the likelihood function when we estimate the models. Here we generate 50,000 individual pseudo-histories for each group using the ISU estimates.

The simulation results are summarised for age range 40–80 in Table 3.3. Part A of Table 3.3 shows results from simulations of the entire subpopulation, under the assumed normal distribution of the random effects. The first panel shows the distribution of the cumulative number of AMI events. It is clear that a large fraction of the populations do not experience any events. However, male Maoris are most likely to ever have an AMI event, followed by female Maoris and male Europeans whose outcomes are similar, while female Europeans are least likely to experience any events. All four distributions are heavily skewed to the right, with a small fraction of people experiencing a large number of events. This is particularly true for Maoris, where about 2.5% have 10 or more events. The skewness implies that the mean number of events are much larger than the medians (which are all 0). Specifically, the average number of AMIs for those who have at least one (before age 80) is about 3.7 for male Maoris, 4.7 for female Maoris, 3.1 for male Europeans, and 2.8 for female Europeans. Note here that the ranking between male and female Maoris is switched for the intensive margin. This happens because, as discussed in relation to Figure 3.7, the risk of subsequent events is highest for the 5 percent female Maoris with the highest random effect values (innate risk) than for the top 5 percent people in the other groups. In other words, a small proportion, 5–10 percent, of female Maoris experience a comparatively large number of events. The familiar ranking holds when the random effects are fixed at their quartile values, as shown in panel B.

To get an idea of the range of experience within each group, part B of Table 3.3 shows results from simulations that hold the random effects constant at the first quartile, the median, and the third quartile of the distribution. These quartiles represent low-risk, middle-risk, and high-risk persons. The main conclusion is that the differences within each subpopulation are much larger than the differences between them.

Slightly more detailed views of the average outcomes are provided in Figure 3.9. The left graph shows the proportion of people at different ages who have ever had an AMI event (the extensive margin). The proportions are highest for male Maoris, since they have the highest age-specific AMI risks. Then follow female Maoris and male Europeans whose outcomes are similar, and finally female Europeans are least likely to have had any AMI events. Note that the lines do not cross. The right graph shows the average number of AMI events people have previously experienced when they reach certain ages, for those who have experienced at least one AMI at that age (the intensive margin). Male Maoris tend to have more events than Europeans, but not as many as female Maoris. As mentioned along with Figure 3.7, female Maoris with high random effect values have comparatively high risk.

Simulating histories from the estimated models can also be used as an informal check of their within-sample fit. For this, we draw a random history for each of the 50,000 individuals in the estimation sample, taking their date of birth and their observation period as given (including the date of death if they die before 30 June 2012). We then compute summary statistics for the simulated sample and compare them with the actual estimation sample. Table 3.4 shows the actual and the simulated distributions of AMI events people experience between ages 40 and 85 during their observation period. The fit is good, although all the estimated models tend to underpredict the proportion of people with no events and overpredict the proportion with multiple events.

3.6 Conclusion

In this study, we investigate gender and ethnic disparities in several aspects of AMI risk in New Zealand. Our study complements the literature on gender and ethnic disparities in other areas of public health and health economics. AMIs, commonly known as heart attacks, are one of the leading causes of disability and mortality and therefore an important issue in public health.

We estimate hazard models of AMI risk separately for male and female people of Maori and European descent using high-quality administrative data on hospital admissions and deaths combined with census data. Recent advances in econometric methodology allow us to overcome a large left-censoring problem. Using these hazard models, we examine how AMI risk is distributed within each subpopulation and we compare patterns across subpopulations.

We find, as expected, that older people have much higher risk than younger people. This partly reflects biological effects and partly time effects as different cohorts have been exposed to different environments, made different life style choices, and had access to different medical technologies. In terms of the three risk regimes, we find that there are important dynamic effects in that the risk of experiencing the first AMI event is much lower than the risk of having subsequent events, and the risk is particularly high in the first year following an event. In addition, we find that there is considerable within-group variation in risk, as measured by the influence of random effects. These two aspects, history dependence and unobserved heterogeneity, pull in the same direction of concentrating risk among relatively fewer people within the subpopulations.

Comparing the four gender and ethnic groups, we discover large disparities in AMI risk between male and female people of Maori and European descent. Generally the ranking is that male Maoris tend to have the highest risk, followed by female Maoris, then male Europeans, and finally female Europeans have the lowest risk. However, there are some exceptions. For example, the risk of subsequent events increases with age more for Europeans than for Maoris, and it seems that the risk may actually be higher for Europeans after age 75–80. The models allow the effect of having the first event to be non-proportional. Comparing the risk of subsequent AMI events relative to the risk of having the first event across the four groups, Maoris have larger relative risk of subsequent events than Europeans before age 70. The level of the relative risk for Maoris decreases as they grow older, so that Maoris have smaller relative risk of subsequent events than Europeans after age 70.

Using the estimated hazard models, we compute cumulative outcomes for synthetic populations aged 40–80 in a world where nobody dies. Here, the gender and ethnic disparities are also clear, at least for the extensive margin, where the proportion who experience any event is highest for male Maoris, followed by female Maoris and male Europeans who have similar proportions, while the proportion of female Europeans who experience at least one event is the smallest. For the intensive margin, female Maoris who have at least one event experience more events on average than the other groups. This arises because the estimated proportion who experience a very large number of events is largest for female Maoris.

Our findings motivate further research on gender and ethnic disparities. For example, if the data are found or become available, it would be interesting and useful to investigate potential biological, socioeconomic, or environmental factors which may explain some of the heterogeneity in risk.

Table 3.1: Summary statistics for the study population

	Male Maoris		Male Europeans		Female Maoris		Female Europeans	
	N	%	*N*	%	*N*	%	*N*	%
Number of people by age†								
After 2002	71,284	21.4	170,452	10.8	66,219	19.5	161,935	9.9
Age 0–29	164,581	49.3	565,775	36.0	163,101	48.1	554,967	33.8
Age 30–39	36,283	10.9	198,814	12.6	41,691	12.3	220,047	13.4
Age 40–49	29,117	8.7	207,019	13.2	32,309	9.5	216,836	13.2
Age 50–59	17,420	5.2	175,447	11.2	18,299	5.4	178,981	10.9
Age 60–69	10,145	3.0	121,389	7.7	10,852	3.2	125,875	7.7
Age 70–79	4,045	1.2	91,753	5.8	4,896	1.4	105,395	6.4
Age 80+	916	0.3	42,364	2.7	1,536	0.5	78,423	4.8
Total	333,791	100.0	1,573,013	100.0	338,903	100.0	1,642,459	100.0
Number of people under age 40 on 30 June 2012 by observed AMIs								
None	235,763	100.0	736,005	100.0	229,277	100.0	716,847	100.0
1	94	0.0	207	0.0	41	0.0	50	0.0
2	6	0.0	13	0.0	2	0.0	4	0.0
3+	2	0.0	2	0.0	0	0.0	1	0.0
Total	235,865	100.0	736,227	100.0	229,320	100.0	716,902	100.0
Number of people over age 40 and under 85 on 30 June 2012 by observed AMIs								
None	91,626	94.4	767,461	93.8	104,609	96.3	852,156	96.2
1	4,579	4.7	42,576	5.2	3,280	3.0	28,275	3.2
2	658	0.7	6,028	0.7	514	0.5	3,779	0.4
3+	211	0.2	1,934	0.2	203	0.2	1,304	0.1
Total	97,074	100.0	817,999	100.0	108,606	100.0	885,514	100.0
Average number of observed AMIs for those with at least one AMI by age†								
Age 30–39	1.29		1.14		1.14		1.12	
Age 40–49	1.34		1.19		1.47		1.15	
Age 50–59	1.47		1.25		1.53		1.26	
Age 60–69	1.41		1.38		1.52		1.36	
Age 70–79	1.47		1.60		1.56		1.56	
Age 80–84	1.19		1.35		1.30		1.30	
Number of deaths in study population between ages 40–85								
By AMI‡	1,391		8,885		720		5,197	
AMI 29 days‡	1,426		9,060		750		5,305	
Total	11,609		80,882		9,767		62,861	

The study population excludes people born before 6 March 1911, ethnic groups other than Maoris and people of European descent, and people with type 1 diabetes. †Age defined on 1 July 2002. ‡'By AMI': people whose direct cause of death is AMI; 'AMI 29 days': people who have an AMI within 29 days of (and including) their date of death.

Table 3.2: Parameter estimates for models with random effects

	Male	Male	Female	Female	χ^2
	Maoris	Europeans	Maoris	Europeans	
			ISU estimates		
α_1	0.765	0.872	0.899	1.117	8.27
	(0.076)	(0.102)	(0.076)	(0.045)	[0.04]
μ_1	−4.672	−5.499	−5.571	−6.962	12.61
	(0.347)	(0.541)	(0.409)	(0.265)	[0.01]
$\ln\sigma_1$	0.620	0.703	0.633	0.519	0.26
	(0.199)	(0.253)	(0.212)	(0.147)	[0.97]
α_2	0.363	0.791	0.455	1.024	45.36
	(0.058)	(0.072)	(0.084)	(0.083)	[0.00]
γ	0.967	1.006	0.765	0.896	4.22
	(0.079)	(0.078)	(0.084)	(0.096)	[0.24]
μ_2	−2.761	−4.610	−3.247	−5.641	26.25
	(0.317)	(0.411)	(0.513)	(0.425)	[0.00]
$\ln\sigma_2$	0.227	0.309	0.395	0.318	1.79
	(0.101)	(0.089)	(0.118)	(0.095)	[0.62]
			ISN estimates		
α_1	0.708	0.728	0.895	1.057	6.31
	(0.078)	(0.042)	(0.080)	(0.034)	[0.10]
μ_1	−4.404	−4.661	−5.544	−6.407	7.70
	(0.368)	(0.239)	(0.433)	(0.190)	[0.05]
$\ln\sigma_1$	0.456	0.008	0.623	0.020	3.99
	(0.280)	(0.396)	(0.229)	(0.268)	[0.26]
α_2	0.345	0.680	0.481	0.937	24.17
	(0.074)	(0.071)	(0.094)	(0.091)	[0.00]
γ	0.952	0.938	0.776	0.865	2.62
	(0.081)	(0.081)	(0.088)	(0.100)	[0.45]
μ_2	−2.737	−3.936	−3.446	−5.073	9.60
	(0.413)	(0.459)	(0.570)	(0.474)	[0.02]
$\ln\sigma_2$	0.293	0.294	0.470	0.367	1.89
	(0.100)	(0.088)	(0.112)	(0.093)	[0.59]

Standard errors in parentheses and p-values in brackets. χ^2: test statistic for the null hypothesis that the four parameter estimates are the same.

Table 3.3: Summary of simulated cumulative outcomes

	Male Maoris	Male Europeans	Female Maoris	Female Europeans
Part A: random effects $v \sim N(0,1)$				
Proportion of people by number of AMI events (%)				
None	65.3	71.9	72.8	83.9
1	15.0	13.5	11.6	8.4
2	6.8	5.8	5.0	3.2
3	3.9	2.8	2.7	1.6
4	2.2	1.7	1.8	0.8
5	1.6	1.1	1.1	0.5
6	1.1	0.7	0.9	0.4
7	0.7	0.5	0.6	0.3
8	0.5	0.4	0.4	0.2
9	0.4	0.3	0.3	0.1
10+	2.5	1.4	2.6	0.7
Proportion of people with at least one AMI (%)				
$v \sim N(0,1)$	34.7	28.1	27.2	16.1
Average number of AMIs for people with at least one AMI				
$v \sim N(0,1)$	3.73	3.11	4.71	2.77
Part B: random effects fixed				
Proportion of people with at least one AMI (%)				
$v = -0.674$	6.8	3.8	4.1	2.4
$v = 0$	22.2	14.1	14.3	7.1
$v = 0.674$	58.0	44.4	41.5	20.1
Average number of AMIs for people with at least one AMI				
$v = -0.674$	1.12	1.06	1.07	1.05
$v = 0$	1.27	1.17	1.19	1.12
$v = 0.674$	1.72	1.46	1.58	1.32

Events between age 40 and age 80 simulated using ISU estimates.

Table 3.4: Summary of simulated within-sample outcomes

	Male Maoris	Male Europeans	Female Maoris	Female Europeans
Proportion of people by number of AMI events (%)				
Estimation sample				
None	94.5	94.3	96.4	97.0
1	4.6	4.7	2.9	2.5
2	0.7	0.7	0.4	0.4
3+	0.2	0.2	0.2	0.1
Simulated sample				
None	91.7	91.6	94.4	95.6
1	5.2	5.7	3.2	2.8
2	1.6	1.4	1.0	0.7
3+	1.5	1.3	1.3	0.9

Events between age 40 and age 85 simulated using ISU estimates.

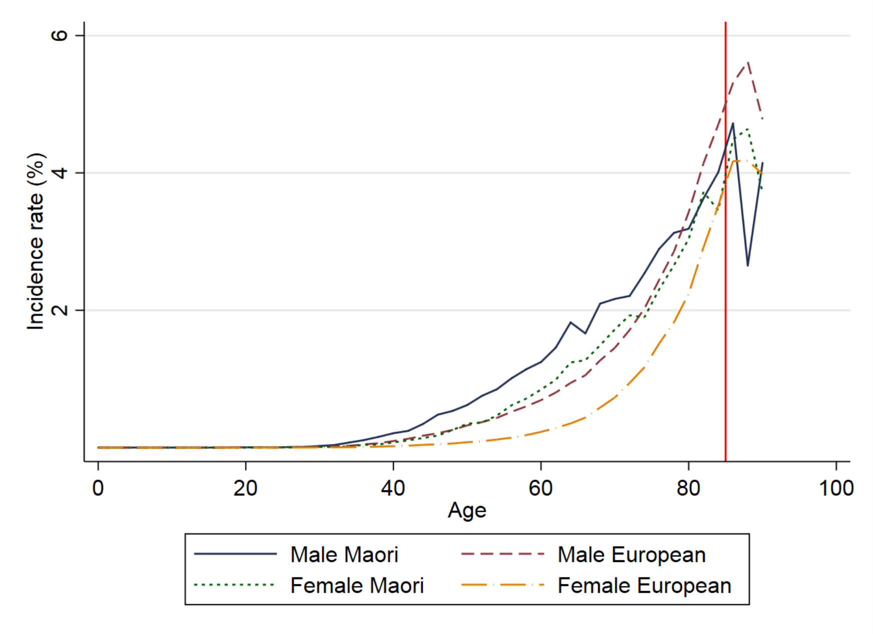

Figure 3.1: Incidence rates (per year)

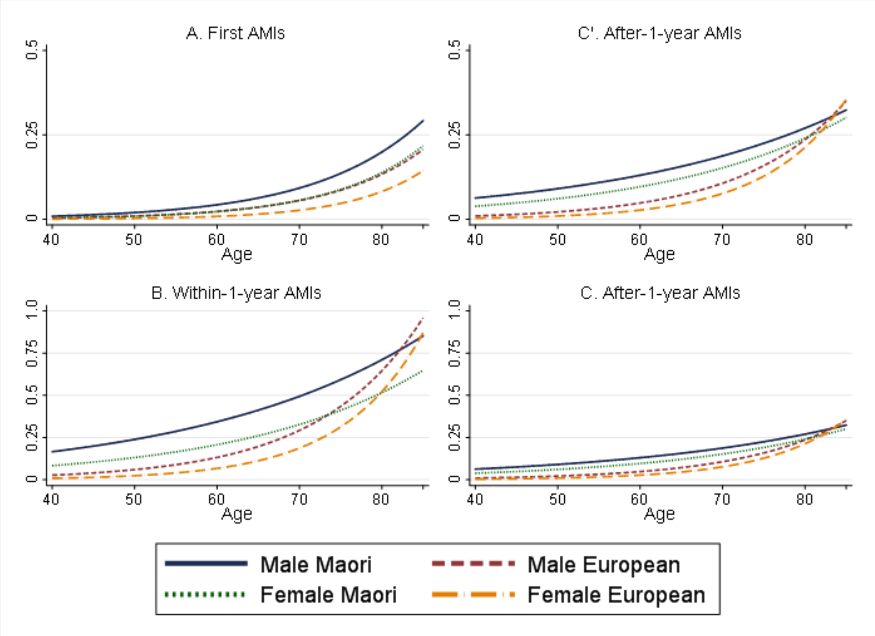

Figure 3.2: Estimated median hazard functions (per 10 years)

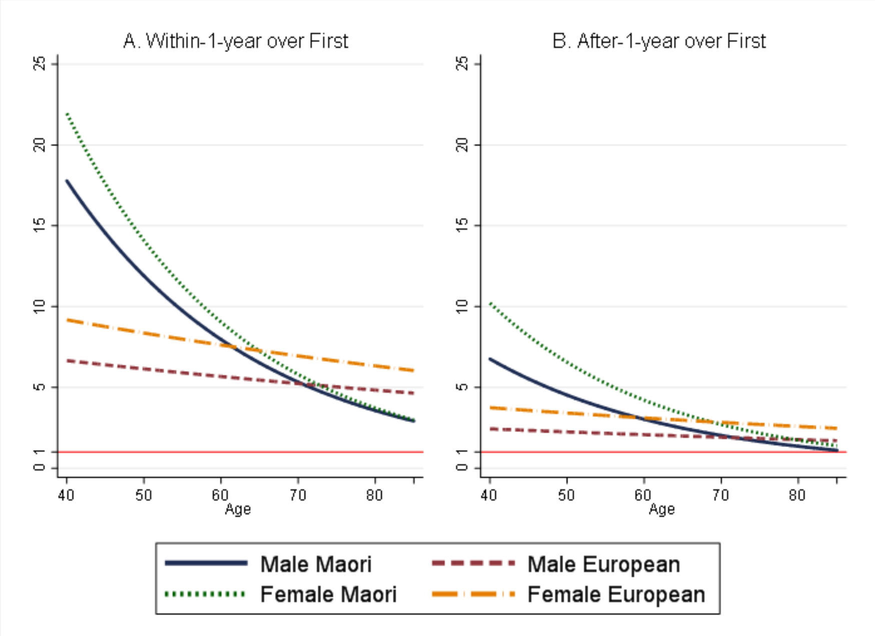

Figure 3.3: Ratio of median hazard functions for subsequent events over first event

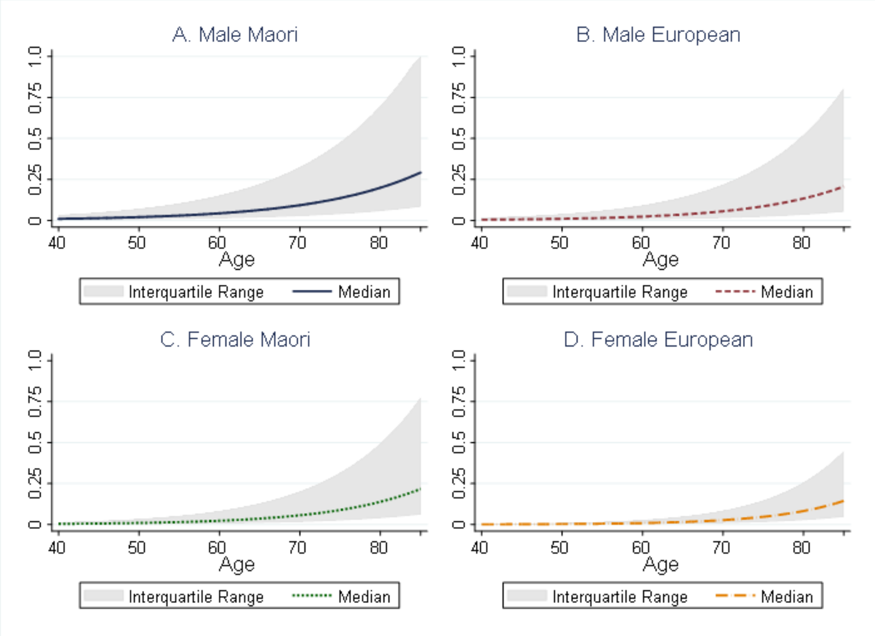

Figure 3.4: Hazard functions for the first AMI (per 10 years)

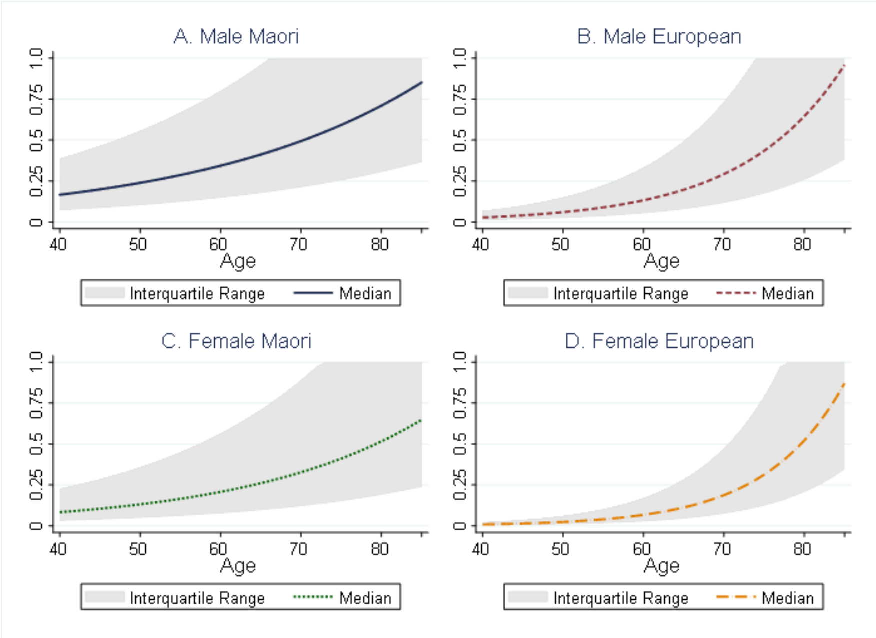

Figure 3.5: Hazard functions for within-1-year AMIs (per 10 years)

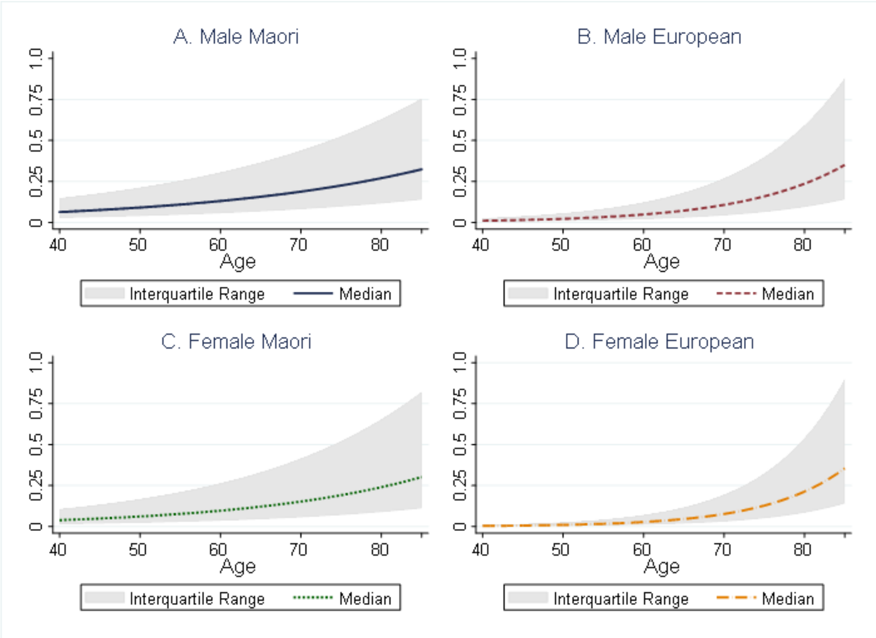

Figure 3.6: Hazard functions for after-1-year AMIs (per 10 years)

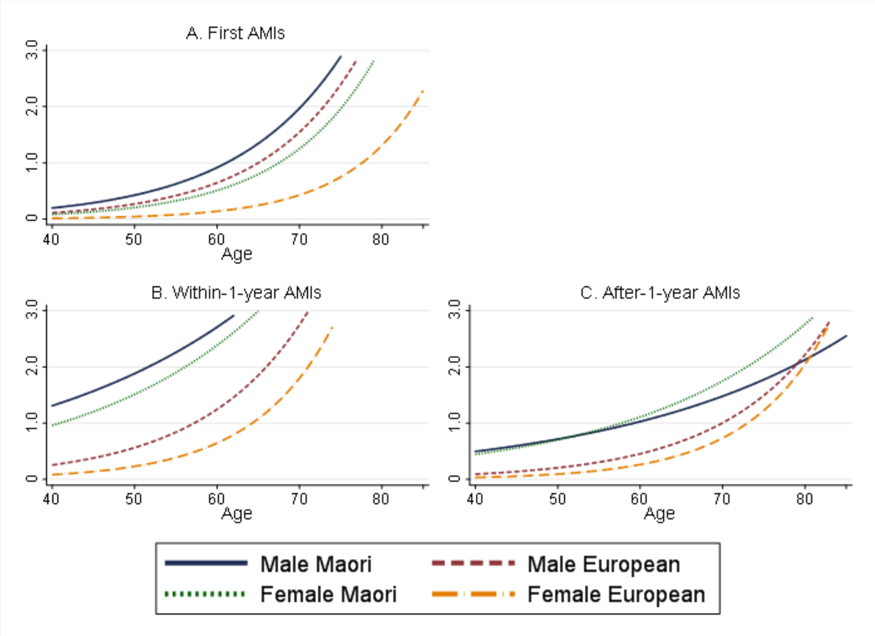

Figure 3.7: Estimated hazard functions at the 95th percentile random effect (per 10 years)

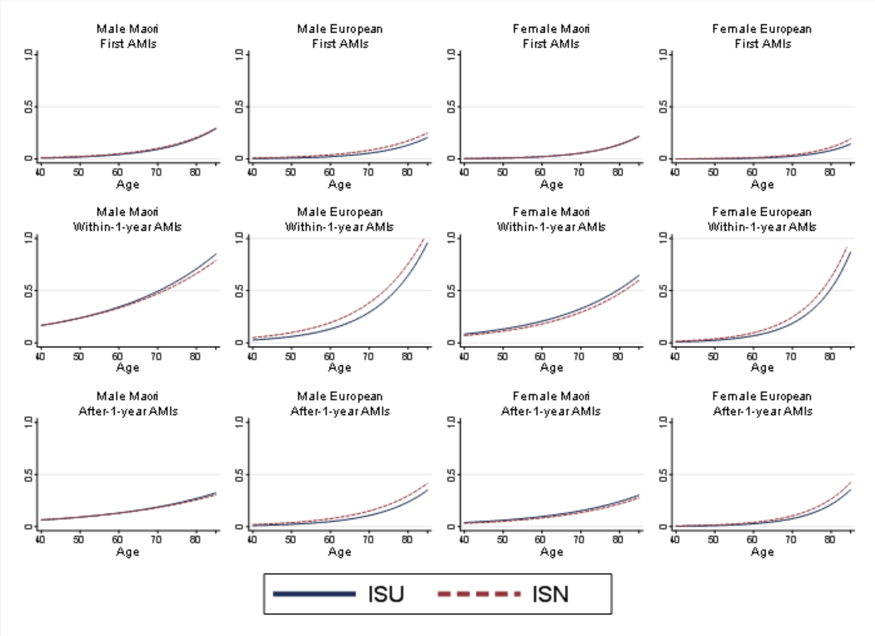

Figure 3.8: ISU and ISN estimates of median hazard functions (per 10 years)

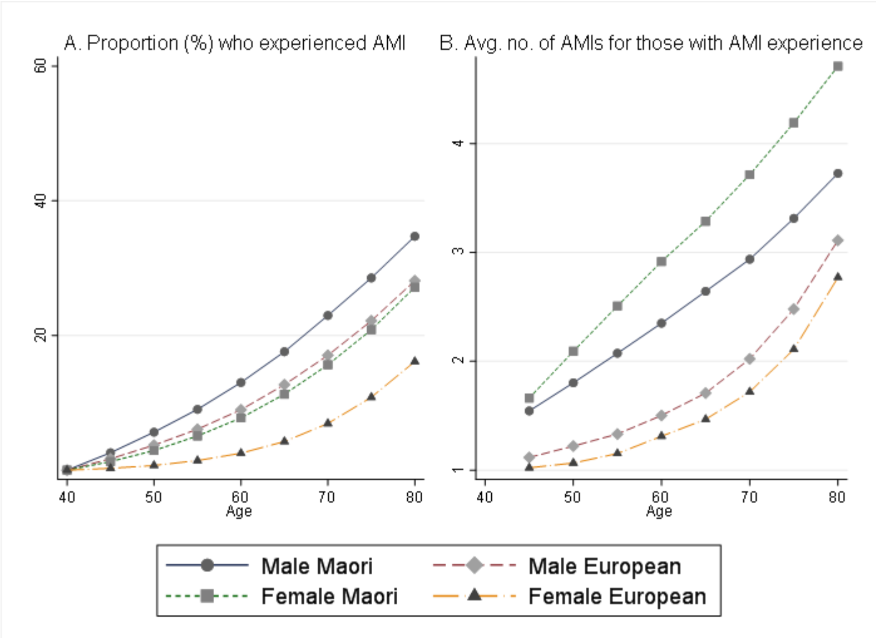

Figure 3.9: Simulated cumulative outcomes

Appendix

Table 3.A1: Parameter estimates for models without random effects

	Male Maoris	Male Europeans	Female Maoris	Female Europeans	χ^2
			ISU estimates		
α_1	0.549	0.659	0.714	0.981	75.69
	(0.024)	(0.022)	(0.026)	(0.032)	[0.00]
μ_1	−3.707	−4.338	−4.556	−6.009	364.50
	(0.045)	(0.057)	(0.059)	(0.099)	[0.00]
α_2	0.196	0.583	0.266	0.759	114.73
	(0.033)	(0.034)	(0.034)	(0.065)	[0.00]
γ	1.080	1.187	1.075	1.084	1.02
	(0.079)	(0.071)	(0.079)	(0.093)	[0.80]
μ_2	−1.120	−2.612	−1.164	−3.213	148.54
	(0.090)	(0.127)	(0.104)	(0.252)	[0.00]
			ISN estimates		
α_1	0.546	0.656	0.711	0.982	76.06
	(0.024)	(0.022)	(0.026)	(0.032)	[0.00]
μ_1	−3.703	−4.335	−4.553	−6.024	364.58
	(0.045)	(0.058)	(0.060)	(0.099)	[0.00]
α_2	0.195	0.591	0.265	0.764	117.54
	(0.033)	(0.034)	(0.034)	(0.065)	[0.00]
γ	1.088	1.187	1.083	1.073	0.86
	(0.080)	(0.072)	(0.080)	(0.093)	[0.84]
μ_2	−1.121	−2.640	−1.163	−3.220	149.97
	(0.091)	(0.128)	(0.105)	(0.252)	[0.00]

Standard errors in parentheses and p-values in brackets. χ^2: test statistic for the null hypothesis that the four parameter estimates are the same.

Chapter 4

Double Robustness without Weighting

4.1 Introduction

In treatment effect analysis (Rosenbaum, 2002; Lee, 2005; and Imbens and Rubin, 2015, among others), matching is widely used. However, given a control group with $D = 0$ and a treatment group with $D = 1$, to find the effects of the treatment D on a response variable Y, matching has two well-known problems: the dimension problem and the support problem. The former describes the dimension of X being too high, which often occurs when we nonparametrically match individuals on covariates X to make sure that two similar individuals are compared. In the latter, the supports of X may not overlap well across the two groups.

Of the two problems, the dimension problem has been overcome using the propensity score (PS) instead of X per se:

$$\pi(X) \equiv E(D|X).$$

PS matching [Rosenbaum and Rubin, 1983] has been very popular; see Stuart [2010], Imbens and Rubin [2015], and references therein. PS matching works because $(Y^0, Y^1) \perp D|X$ implies $(Y^0, Y^1) \perp D|\pi(X)$, where (Y^0, Y^1) are the potential versions of Y corresponding to $D = 0, 1$ and '$\perp$' stands for independence.

Yet another useful, but little known, score is the *prognostic score (PGS)* (for Y^0), say $\psi(X)$, which satisfies $Y^0 \perp X|\psi(X)$, where $\psi(X)$ can be a vector. A prime candidate for PGS is $E(Y^0|X)$; see Hansen [2008] and references therein for PGS in general. This chapter shows that *controlling both PS and PGS makes an estimator doubly robust (DR)* in the sense that the correct parametric specification of either PS or PGS, not necessarily both, makes the estimator consistent for the treatment effect. This finding matters considerably because such a double protection holds for non-weighting estimators, whereas the DR estimators in the literature are weighting-based and thus tend to be numerically unstable due to near-zero denominators in the weighting. See Robins et al. [1994, 2007], Scharfstein et al. [1999], Bang and Robins [2005], Kang and Schafer [2007], Cao et al. [2009], Tan [2010], Rotnitzky et al. [2012], Vermeulen and Vansteelandt [2015], and references therein for DR.

The DR property for controlling both PS and PGS without weighting was first noted by Hu et al. [2012, 2014], only for the 'regression imputation/adjustment'. We establish the DR property for any approach controlling (i.e., conditioning on) both PS and PGS, which includes matching, regression imputation, and complete pairing in Lee [2009, 2012].

We assume two key conditions. First, for simplicity, assume the support overlap

$$0 < P(D = 1|X) < 1, \tag{4.1}$$

although what is necessary is conditioning on only either PS or PGS, not on X. If (4.1) is violated for some values of X, trim those values and redefine X; better yet, trim only those values of X that make $P(D = 1|\cdots)$ almost zero or one, given the conditioning function(s) of X in use. Second, assume no unobserved confounder

$$\begin{aligned}(i) &: \; Y^0 \perp D|X \quad \text{for the effect on the treated } E(Y^1 - Y^0|D = 1); \qquad (4.2)\\ (ii) &: \; Y^0 \perp D|X \text{ and } Y^1 \perp D|X \quad \text{for the effect on the population } E(Y^1 - Y^0).\end{aligned}$$

Although we use independence because Rosenbaum and Rubin [1983] and Hansen [2008] used independence, our proofs in the next section essentially hold under mean independence.

Is the finding that controlling both PS and PGS provides double robustness new? The answer is yes and no. It is no because Rosenbaum and Rubin [1983] considered controlling functions of X other than $\pi(X)$ and because Hansen [2008] wrote in his concluding section "An attractive possibility is to match or subclassify on both propensity and prognostic scores". The answer is yes, however, because the motivation to control anything other than $\pi(X)$ in Rosenbaum and Rubin [1983] is to balance the covariates that remain unbalanced despite $\pi(X)$ being controlled and because Hansen [2008] never contemplated double robustness underlying the idea of controlling both propensity score and prognostic score.

The next section establishes the aforementioned DR property when both PS and PGS are controlled.

4.2 Double Robustness Controlling PS and PGS

Let the parametrically specified $\pi(X)$ and $\psi(X)$ be $\pi(X;\alpha)$ and $\psi(X;\beta)$ for parameters α and β. We first deal with '$\pi(X;\alpha) = \pi(X)$ but possibly $\psi(X;\beta) \neq \psi(X)$' and then the opposite case, '$\psi(X;\beta) = \psi(X)$ but possibly $\pi(X;\alpha) \neq \pi(X)$'. To be precise, we should write $\tilde{\pi}(X;\alpha) = \pi(X)$ for a specified parametric function $\tilde{\pi}(X;\alpha)$, where $\tilde{\pi}(X;\alpha) \neq \pi(X)$ means no α in the parameter space A making $\tilde{\pi}(X;\alpha) = \pi(X)$. '$\pi(X;\alpha) = \pi(X)$ and $\pi(X;\alpha) \neq \pi(X)$' are shorthands for these, and this way of simplifying notation applies to other functions, including $\psi(X)$. We present two main theorems, with the proofs following each theorem.

Theorem 1 *Suppose $Y^0 \perp D|X$ and (4.1) hold. If $\pi(X;\alpha) = \pi(X)$, then regardless of $\psi(X;\beta) = \psi(X)$, $E(Y^1 - Y^0|D = 1)$ is identified by conditioning on $\{\pi(X;\alpha), \psi(X;\beta)\}$ first and then integrating them out; if $Y^1 \perp D|X$ holds additionally, then $E(Y^1 - Y^0)$ is identified.*

Suppose $\pi(X;\alpha) = \pi(X)$ but possibly $\psi(X;\beta) \neq \psi(X)$. Under $Y^0 \perp D|X$, '$Y^0 \perp D|\pi(X)$' holds [Rosenbaum and Rubin, 1983], and '$Y^0 \perp D|\{\pi(X), g(X)\}$' holds as well for any function $g(X)$, as proven by Rosenbaum and Rubin [1983]. Observe now

$$\begin{aligned} & E\{Y|D=1,\pi(X;\alpha),\psi(X;\beta)\} - E\{Y|D=0,\pi(X;\alpha),\psi(X;\beta)\} \\ = \; & E\{Y^1|D=1,\pi(X;\alpha),\psi(X;\beta)\} - E\{Y^0|D=1,\pi(X;\alpha),\psi(X;\beta)\} \\ = \; & E\{Y^1 - Y^0|D=1,\pi(X;\alpha),\psi(X;\beta)\}. \qquad (4.3) \end{aligned}$$

Integrating out $\{\pi(X;\alpha),\psi(X;\beta)\}|D=1$ gives $E(Y^1 - Y^0|D=1)$. If '$Y^1 \perp D|X$' holds additionally, then '$Y^1 \perp D|\{\pi(X;\alpha),\psi(X;\beta)\}$' holds as well. This then turns (4.3) into $E\{Y^1 - Y^0|\pi(X;\alpha),\psi(X;\beta)\}$, which in turn leads to $E(Y^1 - Y^0)$.

Theorem 2 ***(i)*** *Suppose $Y^0 \perp D|X$, $Y^0 \perp X|\psi(X)$ and (4.1) hold. If $\psi(X;\beta) = \psi(X)$, then $E(Y^1 - Y^0|D=1)$ is identified by conditioning on $\{\pi(X;\alpha),\psi(X;\beta)\}$ and then integrating them out; if $Y^1 \perp D|X$ and $Y^1 \perp X|\psi(X)$ additionally, $E(Y^1 - Y^0)$ is identified.* ***(ii)*** *Suppose $Y^0 \perp D|X$, $Y^0 \perp X|\psi(X)$, $Y^1 \perp D|X$, (4.1) and $Y^1 \perp X|\{\psi(X),\mu(X)\}$ hold for some $\mu(X)$. If $\psi(X;\beta) = \psi(X)$ and $\mu(X;\theta) = \mu(X)$, then $E(Y^1 - Y^0)$ is identified by conditioning on $\{\pi(X;\alpha),\psi(X;\beta),\mu(X;\theta)\}$ and then integrating them out.*

'$Y^0 \perp X|\psi(X)$' means that $\psi(X)$ is a PGS. Then, '$Y^0 \perp D|\psi(X)$' holds because

$$\begin{aligned} & E\{D|Y^0,\psi(X)\} = E\{\ E(D|Y^0,X)\ |Y^0,\psi(X)\} \\ & = E\{\ E(D|X)\ |Y^0,\psi(X)\} = E\{\ E(D|X)\ |\psi(X)\} = E\{D|\psi(X)\}, \end{aligned}$$

using $Y^0 \perp D|X$ for the second equality and $Y^0 \perp X|\psi(X)$ for the third.

Suppose $\psi(X;\beta) = \psi(X)$ but possibly $\pi(X;\alpha) \neq \pi(X)$. Since '$Y^0 \perp X|\psi(X)$' means that the distribution of $Y^0|\psi(X)$ does not further depend on X, '$Y^0 \perp X|\{g(X),\psi(X;\beta)\}$'

holds for any $g(X)$. Hence, setting $g(X) = \pi(X;\alpha)$, $\{\pi(X;\alpha), \psi(X;\beta)\}$ is also a PGS, which implies $Y^0 \perp D|\{\pi(X;\alpha), \psi(X;\beta)\}$. We thus have (4.3) again. If $Y^1 \perp D|X$ and $Y^1 \perp X|\psi(X)$ additionally, then the PGS $\psi(X)$ for Y^0 is also a PGS for Y^1, which means that $D = 1$ can be dropped from the last term of (4.3) to lead to $E(Y^1 - Y^0)$.

Turning to THEOREM 2 (ii), $\{\psi(X;\beta), \mu(X;\theta)\}$ is now a PGS for both Y^0 and Y^1, which implies $\{\pi(X;\alpha), \psi(X;\beta), \mu(X;\theta)\}$ is a PGS regardless of $\pi(X;\alpha) = \pi(X)$. This gives

$$\begin{aligned} & E\{Y|D=1, \pi(X;\alpha), \psi(X;\beta), \mu(X;\theta)\} - E\{Y|D=0, \pi(X;\alpha), \psi(X;\beta), \mu(X;\theta)\} \\ = \; & E\{Y^1 - Y^0|\pi(X;\alpha), \psi(X;\beta), \mu(X;\theta)\}, \end{aligned}$$

and integrating out $\{\pi(X;\alpha), \psi(X;\beta), \mu(X;\theta)\}$ yields $E(Y^1 - Y^0)$.

To get an idea about PGS in practice, suppose (Y^0, Y^1) depends on X only through $E(Y^0|X)$ and $E(Y^1|X)$. Then, $\{E(Y^0|X), E(Y^1|X)\}$ can be used for $\{\psi(X), \mu(X)\}$; e.g. $E(Y^0|X) = \psi(X)$ and $E(Y^1|X) = \psi(X) + \mu(X)$. Linear models may be used for $\{E(Y^0|X), E(Y^1|X)\}$.

Chapter 5

Comparison of Treatment Effect Estimators: Matching, Regression Imputation, Doubly Robust Ones and More

5.1 Introduction

Finding the effect of a binary treatment D on a response variable Y is something that is done in almost all disciplines of science. Unless $D = 0, 1$ is randomised, the treatment group ('T group') with $D = 1$ tends to differ from the control group ('C group') with $D = 0$ in observed covariates X and unobserved covariates ε, which can confound the effect of D on Y. To prevent such a confounding, X can be controlled, but not the unobserved ε. Hence, the treatment effect analysis typically proceeds under the assumption of 'no unobserved confounder':

$$(Y^0, Y^1) \perp D|X \tag{5.1}$$

where (Y^0, Y^1) are the potential versions of Y corresponding to $D = 0, 1$ and '$\perp$' stands for independence. See Rosenbaum [2002], Lee [2005, 2016], Pearl [2009] and Imbens and Rubin [2015], among others, for treatment effect analysis in general.

Under (5.1), X can be controlled in various ways, which then leads to various treatment effect estimators. Among them, the most popular approach in practice seems to be matching. Matching has, however, two well-known problems: the dimension problem and the support problem. The former is that, when we match individuals on covariates X to make sure that two similar individuals are compared, the dimension of X is often too high. The latter is that the supports of X may not overlap well across the two groups. Although there is no good solution to the support problem, the dimension problem has been overcome using the propensity score (PS) $\pi(X) \equiv E(D|X)$ instead of X per se. Propensity score matching (PSM) [Rosenbaum and Rubin, 1983] has been very popular; see, e.g. Stuart [2010], Imbens and Rubin [2015], and references therein. PSM works because (5.1) implies $(Y^0, Y^1) \perp D|\pi(X)$.

As well known, however, PSM requires the user to make several arbitrary decisions, such as how many subjects in the opposite group to match, and how to set the value of the 'caliper' (the tolerance threshold of mismatch), whether to match with or without replacement, and so on. Also inference with PSM is difficult despite an advance made in Abadie and Imbens [2016].

Call PSM with an upper bound on the PS distance 'caliper PSM', i.e., if the PS distance between two subjects is greater than the caliper, then the two subjects are regarded as non-matched; in no-caliper PSM, the closest subject is matched regardless of the PS distance. Wu et al. [2015] noted that PSM has been applied 55 times in the four best medical journals. Among the 55 studies, 21 studies used caliper PSM, 10 of which used $0.2 \times Sd(\text{logit PS})$ as the caliper, and 4 of which used $0.6 \times Sd(\text{logit PS})$, where Sd stands for standard deviation; 1:1 (i.e., pair) matching was the most popular. Reviewing 47 earlier medical studies with PSM, Austin [2008] noted "only two studies used appropriate statistical methods both for assessing balance in the matched sample and for assessing the statistical significance of the treatment effect". More recently, Nayan et al. [2017] searched 114 studies with PS in urology to find that PSM was the most popular (62 studies, 54.4%), the majority (77.4%) of which, however, used inappropriate statistical methods. These show that PSM is an important research tool, and yet the degree of arbitrariness and inappropriate use is high.

Other than PSM (and matching in general), there are several types of treatment effect estimators: inverse probability weighting (see Hirano et al., 2003 and references therein), regression imputation/adjustment (to be explained in detail later), complete pairing [Lee, 2009, 2012], PS-residual-based ordinary least squares estimator [Lee, 2018], and various doubly robust (DR) estimators. It is puzzling then why PSM is so popular while the others are not. The reason could be lack of strong evidence that some of the other estimators do much better than matching.

In the literature, there are many small-scale simulation studies whose designs differ in complexity: Linden [2017] with a single regressor and its polynomial functions, Waernbaum [2012] and Linden et al. [2016] with two regressors and their polynomial functions, Kang and Schafer [2007] and Imai and Ratkovic [2014] with four regressors, and Kreif et al. [2016] with eight regressors and their polynomial functions to imitate a real data set. In this chapter, we compare 24 estimators in total through an extensive simulation study with 64 designs and two empirical analyses mimicking experiments.

This chapter is organised as follows. Section 5.2 reviews ‘prognostic score (PGS)’, because Section 5.3 introduces estimators using PS and PGS. Section 5.4 conducts a simulation study, and Section 5.5 provides empirical analyses. Finally, Section 5.6 concludes.

5.2 Prognostic Score

A function $\psi(X)$ is a PGS if $\psi(X)$ satisfies $Y^0 \perp X|\psi(X)$; see Hansen [2008] and references therein. To understand PGS, consider a simple ‘intercept-shift model’:

$$Y_i^0 = X_i'\beta_0 + U_i \qquad \text{and} \qquad Y_i^1 = \beta_* + Y_i^0, \qquad U \perp X \tag{5.2}$$

where β's are parameters and U is an error. Here, $\psi(X) = X'\beta_0$ due to $Y^0 \perp X|X'\beta_0$, and $Y^1 \perp X|X'\beta_0$ holds as well because Y^1 is an intercept-shifted version of Y^0. When Y^1 is not an intercept-shifted version, we may need 'an effect modifier' $\mu(X)$ to achieve $Y^1 \perp X|\{\psi(X), \mu(X)\}$ so that $(Y^0, Y^1) \perp X|\{\psi(X), \mu(X)\}$. For instance, if

$$Y_i^0 = X_i'\beta_0 + U_i \quad \text{and} \quad Y_i^1 = X_i'\beta_1 + Y_i^0, \qquad U \perp X \tag{5.3}$$

where β_1 is a parameter, then $\mu(X) = X'\beta_1$, and we have $Y^1 \perp X|(X'\beta_0, X'\beta_1)$ so that $(Y^0, Y^1) \perp X|(X'\beta_0, X'\beta_1)$.

'$(Y^0, Y^1) \perp D|X$' in (5.1) allows (Y^0, Y^1) and D to be related only through X. If $(Y^0, Y^1) \perp X|\psi(X)$ holds as in (5.2), we then have

$$(Y^0, Y^1) \perp D|\psi(X): \tag{5.4}$$

the possible relation between (Y^0, Y^1) and D through X is "severed" by conditioning on $\psi(X)$. Hence (Y^0, Y^1) *is balanced across the two groups given* $\psi(X)$, *just as* (Y^0, Y^1) *is so given* $\pi(X)$. If (5.4) does not hold, but $(Y^0, Y^1) \perp D|\{\psi(X), \mu(X)\}$ does, then (Y^0, Y^1) is balanced across the two groups given $\{\psi(X), \mu(X)\}$.

Although we used '$\perp$' above in introducing PGS to be "faithful" to the literature, in fact, we can use conditional independence throughout. To see why, consider $E(Y^0|D, X) = E(Y^0|X)$ that is weaker than $Y^0 \perp D|X$ in (5.1), and $E\{Y^0|X, \psi(X)\} = E\{Y^0|\psi(X)\}$ in defining PGS that is weaker than $Y^0 \perp X|\psi(X)$. Then, $E(Y^0|D, X) = E(Y^0|X)$ and $E(Y^0|X)$ $[= E\{Y^0|X, \psi(X)\}] = E\{Y^0|\psi(X)\}$ give

$$E\{Y^0|D, \psi(X)\} = E\{E(Y^0|D, X)|D, \psi(X)\} = E\{E(Y^0|X)|D, \psi(X)\} = E\{Y^0|\psi(X)\}:$$

given $\psi(X)$, Y^0 is balanced across the two groups.

A prime candidate for PGS in practice is $E(Y^0|X)$ as illustrated in (5.2), because

$$E\{Y^0|\psi(X)\} = E\{E(Y^0|X)|\ \psi(X)\} = E(Y^0|X) \quad \text{when} \quad \psi(X) = E(Y^0|X).$$

Considering (5.3), we can also see that setting $\mu(X) = E(Y^1|X)$ gives $E(Y^1|X) = E\{Y^1|\psi(X), \mu(X)\}$ in case $E(Y^1|X) = E\{Y^1|\psi(X)\}$ fails. In short, two prime candidates for PGS to balance Y^0 and Y^1 across the two groups are $E(Y^0|X)$ and $E(Y^1|X)$.

Y^0 and Y^1 being balanced given $\psi(X)$ (or $\{\psi(X), \mu(X)\}$) gives the key equation:

$$\begin{aligned} & E\{Y|D=1,\psi(X)\} - E\{Y|D=0,\psi(X)\} \\ = \ & E\{Y^1|D=1,\psi(X)\} - E\{Y^0|D=0,\psi(X)\} \\ = \ & E\{Y^1|\psi(X)\} - E\{Y^0|\psi(X)\} = E\{Y^1 - Y^0|\psi(X)\}. \end{aligned} \tag{5.5}$$

Integrating out $\psi(X)$ gives $E(Y^1 - Y^0)$. If only Y^0 is balanced given $\psi(X)$, we get

$$E\{Y|D=1,\psi(X)\} - E\{Y|D=0,\psi(X)\} = E\{Y^1 - Y^0|D=1,\psi(X)\}. \tag{5.6}$$

Integrating out $\psi(X)|D=1$ then gives $E(Y^1 - Y^0|D=1)$.

5.3 Review of Estimators to Be Compared

This section reviews five types of estimators to be compared. Certainly, this chapter alone cannot cover all treatment effect estimators that ever appeared, as there are more elaborate estimators than to be reviewed here. The hope is that our selection of estimators is wide enough to address various issues/concerns with treatment effect estimators. Although we can go fully nonparametric, most estimators in practice specify either PS $\pi(X)$ or the PGS $E(Y|X, D=d)$; we will examine only such estimators. Let the pooled sample be indexed by $i = 1, ..., N$, the treatment group by $t = 1, ..., N_1$, and the control group by $c = 1, ..., N_0$; $N = N_0 + N_1$.

To ease comparing estimators below, we impose some restrictions. First, we use probit for $\pi(X)$. Second, we mostly examine $\gamma \equiv E(Y^1 - Y^0)$, not $E(Y^1 - Y^0|D=d)$; nevertheless, γ equals $E(Y^1 - Y^0|D=d)$ for a constant effect (i.e. parallel shift), for which 'effect-on-treated' estimators can be used as well. Third, we use the linear specification $E(Y|X, D = d) = X'\beta_d$ for a parameter β_d. Fourth, for matching, we consider only 1:1 and 1:5 matchings following Rubin and Thomas [2000], and use a caliper of size 0.2 with the matching PS (or PGS) standardised, again following Rubin and Thomas [2000]. Fifth, for weighting with PS, we trim the observations so that $0.001 \leq \pi(X) \leq 0.999$. Sixth, we use two bandwidths for kernel smoothing based on a 'rule-of-thumb' bandwidth $N^{-1/5}$, with the smoothing variable(s) standardised: 0.5 and 1.5 times $N^{-1/5}$ to have relative under-smoothing and over-smoothing.

5.3.1 Regression Imputation (RI)

Under $E(Y|X, D = d) = X'\beta_d$, $d = 0, 1$, the simplest treatment effect estimator is

$$g_{ri-lin} \equiv \frac{1}{N}\sum_{i=1}^{N} X_i'\hat{\beta}_1 - \frac{1}{N}\sum_{i=1}^{N} X_i'\hat{\beta}_0 = \frac{1}{N}\sum_i X_i'(\hat{\beta}_1 - \hat{\beta}_0) \tag{5.7}$$

where $\hat{\beta}_d$ is the ordinary least squares estimator (OLS) of Y on X using the $D = d$ subsample. This 'regression imputation/adjustment (RI)' estimator is consistent for

$$E\{E(Y|X, D = 1) - E(Y|X, D = 0)\} = E\{E(Y^1|X) - E(Y^0|X)\} = \gamma.$$

An alternative to g_{ri-lin} specifying $E(Y|X, D = d)$ as linear is replacing $E(Y|X, D = d)$ with $E\{Y|\pi(X), D = d\}$ after specifying $\pi(X)$ as probit/logit as in Imbens [2000]. A series-approximation estimator for this (using an estimator $\hat{\pi}(X)$ for $\pi(X)$) is

$$g_{ri2-ps} \equiv \frac{1}{N}\sum_i \{\hat{\xi}_{10} + \hat{\xi}_{11}\hat{\pi}(X_i) + \hat{\xi}_{12}\hat{\pi}(X_i)^2\} - \frac{1}{N}\sum_i \{\hat{\xi}_{00} + \hat{\xi}_{01}\hat{\pi}(X_i) + \hat{\xi}_{02}\hat{\pi}(X_i)^2\}$$

where $(\hat{\xi}_{d0}, \hat{\xi}_{d1}, \hat{\xi}_{d2})$ is the OLS of Y on $\{1, \hat{\pi}(X), \hat{\pi}(X)^2\}$ for the $D = d$ subsample; using $\hat{\pi}(X)^3$ additionally gives g_{ri3-ps}. Note the critical difference between g_{ri-lin} and g_{ri2-ps} (and g_{ri3-ps}): the former specifies the PGS's as linear functions of X, whereas the latter specifies PS. As (5.6) shows, we can also use $\psi(X)$ instead of $\pi(X)$: let the analog of g_{ri2-ps} and g_{ri3-ps} using an estimator $\hat{\psi}(X)$ for $\psi(X)$ be $g_{ri2-pgs}$ and

$g_{ri3-pgs}$.

Recently, Lee and Lee [2019] showed that controlling both PS and PGS makes an estimator DR.[1] Hence, let $g_{ri2-ppgs}$ be the DR estimator analogous to g_{ri2-ps} and $g_{ri2-pgs}$ which uses the second order polynomials of $\hat{\pi}(X)$ and $\hat{\psi}(X)$ including the interaction $\hat{\pi}(X)\hat{\psi}(X)$; we do not consider the cubic version, as this results in too many terms compared with g_{ri3-ps} and $g_{ri3-pgs}$.

RI may look like matching, but it is not, because the two terms in each RI estimator are separately averaged for $E(Y^1)$ and $E(Y^0)$. Overall, we use 6 RI's: *RI-lin g_{ri-lin} in (5.7) which is the simplest, RI2-ps g_{ri2-ps} and RI3-ps g_{ri3-ps} specifying PS, RI2-pgs $g_{ri2-pgs}$ and RI3-pgs $g_{ri3-pgs}$ specifying PGS, and RI2-ppgs $g_{ri2-ppgs}$ that is DR specifying both PS and PGS.*

5.3.2 Matching and Bias Correction

Let $\delta_i = 1$ if observation i meets the "caliper condition" for a chosen caliper (and 0 otherwise): having one nearest subject in the opposite group within the caliper distance for 1:1 matching, and five subjects for 1:5 matching; let $N_\delta \equiv \sum_{i=1}^{N} \delta_i$. A "with-replacement" (i.e., a single subject can match multiple subjects in the opposite group) pair-matching estimator using PS is

$$g_{m1-ps} \equiv \frac{1}{N_\delta}\sum_{i=1}^{N}\delta_i(\hat{Y}_i^1 - \hat{Y}_i^0) \text{ with } \hat{Y}_i^1 \equiv D_iY_i + (1-D_i)Y_{t(i)},\ \hat{Y}_i^0 \equiv (1-D_i)Y_i + D_iY_{c(i)}$$

where $t(i)$ is the matched treated subject for control i, and $c(i)$ is the matched control for treated i; $t(i)$ and $c(i)$ are chosen to minimise $|\hat{\pi}(X_i) - \hat{\pi}(X_{t(i)})|$ and $|\hat{\pi}(X_i) - \hat{\pi}(X_{c(i)})|$. Define g_{m1-pgs} analogously using $\hat{\psi}(X)$ instead of $\hat{\pi}(X)$. Going further, define DR $g_{m1-ppgs}$ analogously using both $\hat{\psi}(X)$ and $\hat{\pi}(X)$ to choose $t(i)$ minimising

$$\max\{\frac{|\hat{\pi}(X_i) - \hat{\pi}(X_{t(i)})|}{Sd\{\hat{\pi}(X)\}}, \frac{|\hat{\psi}(X_i) - \hat{\psi}(X_{t(i)})|}{Sd\{\hat{\psi}(X)\}}\}. \tag{5.8}$$

[1] The introduction and the theoretical proof in Lee and Lee [2019] is presented in Chapter 4 of the thesis.

For 1:5 matching, $Y_{t(i)}$ and $Y_{c(i)}$ are replaced with the average of the five closest subjects.

A bias-corrected version of g_{m1-ps} [Abadie and Imbens, 2011] is

$$g_{m1-bc} \equiv \tfrac{1}{N_\delta}\textstyle\sum_{i=1}^{N}\delta_i(\tilde{Y}_i^1 - \tilde{Y}_i^0) \quad \text{with} \quad \begin{aligned}\tilde{Y}_i^1 &\equiv D_iY_i + (1-D_i)(Y_{t(i)} + X_i'\hat{\beta}_1 - X_{t(i)}'\hat{\beta}_1),\\ \tilde{Y}_i^0 &\equiv (1-D_i)Y_i + D_i(Y_{c(i)} + X_i'\hat{\beta}_0 - X_{c(i)}'\hat{\beta}_0);\end{aligned}$$

we consider only 1:1 matching for bias-correction. Combining matching with bias-correction makes g_{m1-bc} DR as Abadie and Imbens [2011] noted. Our g_{m1-bc} differs from Abadie and Imbens' [2011] original formulation in two aspects. First, we use linear models for $E(Y|X, D = d)$ in bias correction, whereas Abadie and Imbens [2011] used nonparametric estimators. Second, we use $\hat{\pi}(X)$ in selecting $t(i)$ and $c(i)$, whereas Abadie and Imbens [2011] used X per se. Abadie et al. [2004] implemented a version of g_{m1-bc} in STATA using linear models for the matched samples only, whereas g_{m1-bc} estimates linear models for the full control and treatment samples.

Instead of matching on population, we also do matching on the treated, for which we do 'without-replacement' matching, as this may better reflect the way matching is done in practice. Let $\delta_{1i} = 1$ if treated subject i satisfies the caliper condition; $N_{\delta 1} \equiv \sum_{i\in\{D=1\}}\delta_{1i}$. Define

$$g_{mt1-ps} \equiv \frac{1}{N_{\delta 1}}\sum_{i\in\{D=1\}}\delta_{1i}(Y_i - Y_{c(i)}),$$

and its PGS and PPGS versions $g_{mt1-pgs}$ and $g_{mt1-ppgs}$. For matching on the treated, $Sd\{\hat{\pi}(X)|D=0\}$ and $Sd\{\hat{\psi}(X)|D=0\}$ are used instead of $Sd\{\hat{\pi}(X)\}$ and $Sd\{\hat{\psi}(X)\}$. Overall, we use 9 matchings: *M1-ps* g_{m1-ps}*, M1-pgs* g_{m1-pgs} *, M1-ppgs* $g_{m1-ppgs}$ *that is DR, M5-ps* g_{m5-ps} *, M5-pgs* g_{m5-pgs}*, M1-bc* g_{m1-bc} *that is DR, Mt1-ps* g_{mt1-ps}*, Mt1-pgs* $g_{mt1-pgs}$*, and Mt1-ppgs* $g_{mt1-ppgs}$ *that is DR.*

5.3.3 Weighting and Doubly Robust Estimators

A weighting (or inverse probability weighting) estimator for γ is

$$\frac{1}{N}\sum_i\{\frac{D_i}{\hat{\pi}(X_i)}-\frac{1-D_i}{1-\hat{\pi}(X_i)}\}Y_i \to^p E\{\frac{DY}{\pi(X)}-\frac{(1-D)Y}{1-\pi(X)}\}=E(Y^1-Y^0)=\gamma.$$

Let $\delta_{\pi i}\equiv 1[0.001\le\hat{\pi}(X_i)\le 0.999]$, where $1[B]=1$ if B holds and 0 otherwise. A normalised version of the weighting estimator with δ_π is

$$g_{wgt}\equiv\sum_i\delta_{\pi i}\frac{D_iY_i}{\hat{\pi}(X_i)}\Big/\sum_i\delta_{\pi i}\frac{D_i}{\hat{\pi}(X_i)}\;-\;\sum_i\delta_{\pi i}\frac{(1-D_i)Y_i}{1-\hat{\pi}\,(X_i)}\Big/\sum_i\delta_{\pi i}\frac{1-D_i}{1-\hat{\pi}\,(X_i)}$$

which is our weighting estimator 'Wgt' to be used in simulation.

A "canonical" DR estimator obtains by modifying the first weighting estimator:

$$g_{dr-c}\equiv\hat{E}(Y^1)-\hat{E}(Y^0)\text{ where }\hat{E}(Y^1)\equiv\frac{1}{N_\pi}\sum_i\delta_{\pi i}\{\frac{D_iY_i}{\hat{\pi}(X_i)}\;-\;\frac{D_i-\hat{\pi}(X_i)}{\hat{\pi}(X_i)}X_i'\hat{\beta}_1\},$$
$$N_\pi\equiv\sum_i\delta_{\pi i},\qquad\hat{E}(Y^0)\equiv\frac{1}{N_\pi}\sum_i\delta_{\pi i}\{\frac{(1-D_i)Y_i}{1-\hat{\pi}(X_i)}\;-\;\frac{\hat{\pi}(X_i)-D_i}{1-\hat{\pi}(X_i)}X_i'\hat{\beta}_0\}$$

and DR-c stands for 'DR-canonical'. DR estimation was proposed originally by Robins et al. [1994], and many variations of DR-c can be seen in Scharfstein et al. [1999], Bang and Robins [2005], Robins et al. [2007], Cao et al. [2009], Tan [2010], Rotnitzky et al. [2012], Vermeulen and Vansteelandt [2015], Lee and Lee [2019], and references therein. Overall, we consider 2 weighting-based estimators: *Wgt* g_{wgt} *and DR-c* g_{dr-c}.

5.3.4 Complete Pairing (CP)

Instead of selecting a few matched individuals, one may wonder why not use all possible pairs across the two groups; N_0N_1 pairs in total. This is the idea of complete pairing (CP) in Lee [2009, 2012]. Whereas Lee [2009, 2012] followed a two-sample framework for more generality, we adopt the usual one-sample framework.

Suppose X is discretely distributed. A marginal effect estimator using all pairs is

$$L_N \equiv \frac{(N_0N_1)^{-1}\sum_{t\in T}\sum_{c\in C}1[X_c = X_t](Y_t - Y_c)}{(N_0N_1)^{-1}\sum_{t\in T}\sum_{c\in C}1[X_c = X_t]}$$

where '$t \in T$' means belonging to the treatment group, and '$c \in C$' the control group. Let x_r, $r = 1, ..., R$, be the common support points across the two groups. With $p_{dr} \equiv P(X = x_r|D = d)$, it holds that

$$L_N \to^p L \equiv \sum_r E(Y^1 - Y^0|X = x_r)\omega(r) \text{ as } N \to \infty \quad \text{where } \omega(r) \equiv \frac{p_{1r}p_{0r}}{\sum_r p_{1r}p_{0r}}; \tag{5.9}$$

L is a marginal effect defined as the "ω-weighted average" of the X-conditional effects.

When $p_{0r} = 0 \neq p_{1r}$, an additive weight such as $(p_{1r} + p_{0r})/\sum_r(p_{1r} + p_{0r})$ is not zero, but $p_{1r}p_{0r}/\sum_r p_{1r}p_{0r}$ in (5.9) is. The latter is preferable because the two groups are not comparable on $X = x_r$. The product weight $p_{1r}p_{0r}$ for L_N thus ensures comparing $E(Y^1 - Y^0|X = x)$ only on the common support, which is a built-in feature of CP to guard against the support problem.

If X is continuous with dimension $k \times 1$, then instead of L_N, we can use

$$M_N \equiv \frac{(N_0N_1)^{-1}\sum_{t\in T}\sum_{c\in C}h^{-k}K\{(X_c - X_t)/h\}(Y_t - Y_c)}{(N_0N_1)^{-1}\sum_{t\in T}\sum_{c\in C}h^{-k}K\{(X_c - X_t)/h\}}$$

where K is a kernel such as $N(0,1)$ density. Since this has the nonparametric dimension problem, we use PS or PGS instead of X per se. M_N is similar to 'kernel matching' as in Heckman et al. [1997], but the difference is that kernel matching uses a kernel-weighed average of matched subjects to construct the counter-factual, which is not the case in M_N. In total, we consider 6 CP estimators where the number after 'CP' indexes bandwidths 1 and 2 for relative under- and over-smoothing: *CP1-ps* g_{cp1-ps} *and CP2-ps* g_{cp2-ps} *with* $\hat{\pi}(X)$, *CP1-pgs* $g_{cp1-pgs}$ *and CP2-pgs* $g_{cp2-pgs}$ *with* $\hat{\psi}(X)$, *and CP1-ppgs* $g_{cp1-ppgs}$ *and CP2-ppgs* $g_{cp2-ppgs}$ *with* $\hat{\pi}(X)$ *and* $\hat{\psi}(X)$ *that are DR.*

5.3.5 OLS with PS-Residual

Lee [2018] proposed a simple OLS using PS residual: obtain the probit estimator $\hat{\alpha}$ under $\pi(X) = \Phi(X^T\alpha)$ where Φ is the $N(0,1)$ distribution function, and then do the simple OLS of $Y - \bar{Y}$ on $D - \Phi(X^T\hat{\alpha})$ to estimate the treatment effect γ, where $\bar{Y}$ is the sample mean of Y. Lee [2018] generalised this simple OLS by obtaining the OLS $(\hat{\zeta}_0, \hat{\zeta}_1, \ ..., \ \hat{\zeta}_q)$ of Y on $\{1, \ X^T\hat{\alpha}, \ ..., \ (X^T\hat{\alpha})^q\}$ and then doing the OLS of

$$Y - \sum_{j=0}^{q} \hat{\zeta}_j (X^T\hat{\alpha})^j \qquad \text{on} \qquad D - \Phi(X^T\hat{\alpha})$$

which includes the simple OLS as a special case when $q = 0$. All estimators with q require the correctly specified PS, but when the PS is misspecified, those with $q \neq 0$ are likely to be less biased than the simple OLS with $q = 0$. Lee [2018] proposed to set $q = 2, 3$, and we use the estimator *OLS-ps* g_{ols-ps} *with* $q = 2$.

The main advantage of OLS-ps is its simplicity in requiring only probit (or logit) and OLS, and the ease in estimating its asymptotic variance. Among the estimators we consider, the only other estimator as simple as OLS-ps in getting the asymptotic variance is RI-lin in (5.7). However, the critical difference between the two estimators is that OLS-ps specifies PS whereas RI-lin specifies PGS.

OLS-ps has further advantages. First, OLS-ps does not require any tuning constant such as a caliper—the q above is not a tuning constant, because OLS-ps is consistent for any value of $q = 0, 1, 2, ...$ Second, it is numerically stable, unlike Wgt and DR-c. Third, it works for non-continuously distributed responses as well, and can be easily extended to multiple treatments. In a small-scale simulation study in Lee [2018], OLS-ps performed far better than the other estimators compared there.

5.4 Simulation Study

5.4.1 Designs and Main Findings

Since we compare as many as 24 estimators, we keep the basic design simple with three regressors, but vary its parameters in diverse ways to accommodate 32 or 64 designs. With $N = 400, 800$ and the number of Monte Carlo data sets $10,000$, our base design is:

$$D = 1[\alpha_1 \;+\; (\alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4)/\sqrt{\alpha_2^2 + \alpha_3^2 + \alpha_4^2} \;+\; \varepsilon > 0];$$

$$X \equiv (X_2, X_3, X_4)' \text{ are iid } N(0,1), \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2) \perp X \text{ with } \sigma_\varepsilon = 1,\ 2;$$

$$\alpha_1 = -0.674,\ 0, \qquad \alpha_2 = 1, \quad \alpha_3 = 1, \qquad \alpha_4 = -1,\ 1;$$

$$Y^0 = \psi(X) + U,\ Y^1 = Y^0 + 1,\ \psi(X) = \beta_1 + (\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)/\sqrt{\beta_2^2 + \beta_3^2 + \beta_4^2};$$

$$U \sim N(0, \sigma_u^2) \perp X \text{ with } \sigma_u = 0.5, \quad \beta_1 = 0, \quad \beta_2 = 2, \quad \beta_3 = 1, \quad \beta_4 = 1.$$

As can be seen in this display, we normalise the regression function for D (other than α_1) and $\psi(X)$ so that their mean and Sd are always 0 and 1, because, otherwise, changes in parameters affect not only the PS and PGS overlaps, but also the variations explained by X in the treatment and response models. In some designs, $Z \equiv (Z_2, Z_3, X_4)'$ is observed instead of X, where $Z_2 = \{1 + \exp(X_2)\}^{-1}$ and $Z_3 = \exp(X_3/2)$. As in Kang and Schafer [2007], $\pi(Z)$ and $\psi(Z)$ are then moderate misspecifications of $\pi(X)$ and $\psi(X)$.

The above setting results in 32 $(= 2 \times 2 \times 2 \times 2 \times 2)$ factorial designs: (1) the control group size being three times greater than or the same as the treatment group size with $\alpha_1 = -0.674,\ 0$; (2) poor or good PS overlap with $\sigma_\varepsilon = 1,\ 2$; (3) good or poor PGS overlap with $\alpha_4 = -1,\ 1$; (4) $\pi(X)$ or $\pi(Z)$ used; and (5) $\psi(X)$ or $\psi(Z)$ used.[2] We use the 32 designs first, and then consider an effect modifier, which results in 64 designs in total. The simulation results not presented below (including all $N = 800$ results) are either in the appendix or available from the authors upon request.

[2] We include a constant term in the regression of PS and PGS models.

About (2), the PS overlap is good when $\sigma_\varepsilon = 2$ because X becomes less relevant for $D = 0, 1$, compared with when $\sigma_\varepsilon = 1$. Regarding (3), although α_4 appears in the D equation, not in the Y equation, α_4 affects the PGS overlap by altering the treatment and control groups; the PGS overlap is good when $\alpha_4 = -1$ (compared with when $\alpha_4 = 1$) because large X_2 and X_3 values get cancelled by a large X_4, which also makes X less relevant for $D = 0, 1$. Figure 5.A1 in the appendix illustrates (2) and (3), using simulated data with $N_0 \simeq N_1$. As for (4) and (5), we consider the four cases of PS and PGS correctly or wrongly specified depending on whether X or Z is observed. Figure 5.A2 in the appendix shows with the same simulated data that Z looks good enough in explaining Y so that the researcher may not detect the misspecification.

Presenting our main simulation findings in advance, first, OLS-ps performs overall the best, followed by RI-lin; both are easy to obtain with simple asymptotic variance estimators. Second, the DR estimators do not perform well when both PS and PGS are misspecified, and when only one is misspecified, the DR estimators work well, but hardly ever better than OLS-ps or RI-lin; *among the DR estimators, we recommend RI2-ppgs and CP1-ppgs.* Third, OLS-ps dominates all the other PS-based estimators. Fourth, RI-lin does mostly better than the other PGS-based estimators, but occasionally RI-lin does worse than some. Fifth, multiple matching performs better than the popular pair matching, but overall, matching is inferior to OLS-ps and RI-lin.

5.4.2 Simulation Tables

Table 5.1 shows results for four base designs with good PS and PGS overlaps, where PS and PGS are correctly specified or moderately misspecified as described above: column (1) for both PS and PGS correct; (2) for PGS wrong; (3) for PS wrong; and (4) for both wrong. The PS-based estimators are in the upper third, the PGS-based estimators are in the middle third, and the DR ones are in the lower third. For all tables in this chapter, *the reported bias, Sd, and root mean squared error (rmse) are scaled up 100 times to avoid too many zeros.*

5.4.2.1 Results for base designs

In column (1) with both scores correct, most estimators perform not much differently in terms of rmse, with almost no bias except for CP2; the best performing estimators are the two simplest ones, OLS-ps and RI-lin, followed by RI2-ppgs and DR-c. In column (2) with PGS misspecified, OLS-ps does clearly best, followed by RI2-ps and RI2-ppgs; despite the misspecified PGS, RI-lin still performs hardly any worse than these two. In column (3) with PS misspecified, RI-lin does best, followed by RI2-ppgs and RI2-pgs; although all PS-based estimators are heavily biased, OLS-ps does best in that group, and better than several PGS-based or DR estimators. In column (4) with both PS and PGS misspecified, OLS-ps and RI-lin perform best, followed by RI2-pgs and RI2-ppgs.

In the following, we make comments on each type (RI, matching, weighting, ...) and each group (PS-based, PGS-based and DR) of estimators. The comments are based, not just on Table 5.1, but also on Tables 5.2–5.4 below. This way, we do not have to make similar comments again for Tables 5.2–5.4, where we focus on a few best-performing estimators.

Examining each type of estimators in turn, for regression imputation, RI2 and RI3 do similarly, with RI3 less biased but more variable than RI2, and RI2 and RI3 are mostly dominated by RI-lin. For matching, MT1 performs mostly a little better than M1 or similarly to M1; M1-ps and M1-pgs do worse than M5-ps and M5-pgs, respectively (see also Table 5.A1 in the appendix); and M1-bc performs overall comparably to M1-ppgs and MT1-ppgs. For Wgt and DR-c, DR-c performs mostly better than Wgt. For complete pairing, CP1 and CP2 show a trade-off between bias and efficiency, but CP1 does overall better than CP2.

Looking at the three groups in turn, among the PS-based estimators which are understandably biased when the PS is misspecified in columns (3) and (4), OLS-ps dominates by a big margin. Among the PGS-based estimators which are understandably biased when the PGS is misspecified in columns (2) and (4), RI-lin mostly dominates, but RI2-pgs comes close. Among the DR estimators, RI2-ppgs dominates in Table 5.1, but CP1-ppgs mostly dominates in Tables 5.2–5.3 where the support overlaps are not good; recall that CP has a built-in protection against the support problem. The DR estimators are not biased when only one of the PS and PGS is misspecified, but when both PS and PGS are misspecified in column (4), the DR estimators are as biased as the other estimators.

Summarising Table 5.1 with good PS and PGS overlaps, OLS-ps, RI-lin and RI2-ppgs are the three best estimators.

5.4.2.2 Results for poor PGS or PS overlap

Table 5.2 is for the poor overlap of only one of PS and PGS, and there are four different designs: column (1) for poor PGS overlap with both scores correct; (2) for poor PGS overlap with both scores wrong; (3) for poor PS overlap with both scores correct; and (4) for poor PS overlap with both scores wrong. The appendix presents rmse's for the omitted cases such as only one score wrong.

In column (1) for poor PGS overlap with both scores correct, most estimators perform similarly with the only exception being CP2 that tends to be more biased. Compared with column (1) of Table 5.1 with good PGS overlap, PGS overlap does not seem to matter much, because columns (1) in Tables 5.1 and 5.2 are similar. The best performing estimators are OLS-ps and RI-lin, closely followed by RI2-pgs and RI2-ppgs.

In column (2) for poor PGS overlap with both scores wrong, all estimators exhibit a bias, although the Sd's do not differ much from those in column (1). OLS-ps and RI-lin do best, closely followed by CP1-ppgs, RI2-pgs, M5-pgs, and RI2-ppgs.

In column (3) for poor PS overlap with both scores correct, OLS-ps and RI-lin still perform best, with no other estimator coming close.

In column (4) for poor PS overlap with both scores wrong, compared with column (4) of Table 5.1, PGS-based estimators exhibit only minor deterioration, whereas the

PS-based estimators show much worse performance except OLS-ps and CP1-ps. DR estimators do reasonably well, except DR-c that is the worst performing. RI-lin performs best, closely followed by RI2-pgs, M5-pgs and CP1-pgs, and then by RI3-pgs, OLS-ps and CP2-ppgs.

Summarising Table 5.2 with only one score poorly overlapping, OLS-ps and RI-lin perform best.

5.4.2.3 Results for poor PGS and PS overlap

Whereas Table 5.2 is for poor overlap in only one of PGS and PS, Table 5.3 is for poor overlap in both scores, with the control group size being as big as (in the left half) or three times bigger than (in the right half) the treatment group size. There are four columns in Table 5.3: column (1) for both scores correct and $N_0 \simeq N_1$, (2) for both scores wrong and $N_0 \simeq N_1$, (3) for both scores correct and $N_0 \simeq 3N_1$, and (4) for both scores wrong and $N_0 \simeq 3N_1$. The other omitted cases such as only one score wrong can be found in the appendix where only rmse's are presented.

In column (1) for both scores correct and $N_0 \simeq N_1$, the PS-based estimators do worse than the PGS-based ones that perform comparably to the DR estimators. OLS-ps and RI-lin perform best, and no other estimator comes close.

In column (2) with both scores wrong and $N_0 \simeq N_1$, OLS-ps and CP1-ppgs do best, followed by RI-lin. No other estimator comes close.

In columns (3) and (4) with $N_0 \simeq 3N_1$, the number of the treated remains the same whereas the control group reservoir goes up from about 200 to 600, which is supposed to give an advantage to MT1. Note that PS and PGS overlaps in columns (3) and (4) are worse than in columns (1) and (2). Compared to columns (1) and (2), indeed, MT1 in columns (3) and (4) does better with smaller rmse's, but MT1 is never best-performing. In column (3), OLS-ps does best followed by CP1-pgs. In column (4), again OLS-ps does best, closely followed by MT1-pgs and CP1-ppgs, and then by M5-pgs, CP1-pgs and MT1-ppgs. Differently from the other designs, RI-lin does not do as well.

Summarising Table 5.3 with both scores overlapping poorly, OLS-ps performs best, and sometimes best by far, to be followed by RI-lin, CP1-pgs and CP1-ppgs.

5.4.2.4 Results with/without modifier $\mu(X)$

So far, Y^1 has been an intercept-shifted version of Y^0, with no effect modifier. In Table 5.4, we consider an effect modifier $\mu(X)$ in two settings: with $\kappa_2 = 1, \kappa_3 = 1$,

$$\begin{aligned} \text{Heterogeneous effect} &: \quad Y^0 = \psi(X) + U, \quad Y^1 = Y^0 + 1 - \kappa_2 X_2; \\ \text{Heteroskedasticity} &: \quad Y^0 = \psi(X) + U, \quad Y^1 = Y^0 + 1 + (\kappa_2 X_2 + \kappa_3 X_3)U; \end{aligned}$$

the heteroskedastic error is normalised so that its marginal Sd becomes one. In Table 5.4, column (1) does not use $\mu(X) = E(Y^1 - Y^0|X)$ in estimation. In contrast, except for the PS-based estimators, column (2) uses $\mu(X)$ along with PS or PGS, and its ‘ratio’ column shows the ratio of the rmse of column (2) (not shown separately) relative to the rmse of column (1). Columns (3) and (4) can be understood analogously, with the only difference being $\mu(X) = Sd(Y^1|X)$.

In column (1) for heterogeneous effect, all estimators do worse than in column (1) of Table 5.1, and all PGS-based estimators except RI-lin are much biased. RI-lin does best, followed by DR-c and RI2-ppgs.

In column (2) for heterogeneous effect with using $\mu(X)$, using $\mu(X)$ additionally improves the PGS-based estimators by removing the biases, except for RI-lin and M5-pgs; it makes no difference for RI-lin, and it worsens M5-pgs by increasing its Sd much. Surprisingly, judging from the ratio column, the performance of M1-ppgs, MT1-ppgs and CP1-ppgs deteriorates much by using $\mu(X)$. Since the biases are almost zero for most estimators except for CP2, we can take the Sd’s as the rmse’s. RI-lin still does best, closely followed by RI2-pgs and RI3-pgs, and then by RI2-ppgs and DR-c.

In column (3) for heteroskedasticity, ignoring $\mu(X)$ makes hardly any difference from column (1) of Table 5.1, and OLS-ps, RI-lin and CP2-ppgs perform best.

In column (4) for heteroskedasticity with using $\mu(X)$, judging from the ratio column, using $\mu(X)$ makes little difference for the PS-based estimators, improves some PGS-based estimators, but worsens some DR estimators. RI-lin does best, closely followed by OLS-ps, CP2-ppgs, and then by CP1-pgs.

Summarising Table 5.4 with heterogeneity or heteroskedasticity, regardless of using modifiers or not, RI-lin does best, followed by OLS-ps, RI2-ppgs and DR-c.

5.5 Empirical Analyses

This section presents two empirical examples, for which we drop 7 estimators among the 24 estimators whose performance were almost dominated by others: RI3-ps, Wgt, CP2-ps, RI3-pgs, CP2-pgs, M1-bc, and CP2-ppgs. The two empirical example data were originally used for 'fuzzy regression discontinuity (RD)' design, where the treatment of interest involves a known cutoff, say c. In fuzzy RD, the main concern is how to overcome the endogeneity of the treatment using a local sample around c. The 17 estimators cannot, however, address treatment endogeneity, as they all require treatment exogeneity. Hence, we put aside the endogeneity issue, and instead focus on generating artificial experiment settings using the two data sets as follows.

Since we do not know the true model in real data, we set up an OLS model for Y and apply a nonparametric model specification test in Stute [1997] to ensure that the model is not rejected, and then we check out how close the 17 estimates are to the OLS effect estimate. The model for Y is also used for PS and PGS specifications, which may put the PS-based estimators at a disadvantage. In the first empirical example, the treatment is not binary, but we transform the treatment into binary in 8 different ways and use the full data. In the second example, the treatment is already binary, and we use 8 different bandwidths h around c to generate 8 local samples.

5.5.1 Class Size Effect on Reading Score

Our first example uses the same data as used in Angrist and Lavy [1999], where the treatment is class size, Y is a reading test score of fifth graders in Israeli public schools, and the observation unit is a class. We transform class size into a binary variable using a threshold τ such that $D = 1[\text{class size} \leq \tau]$. This amounts to assuming that the class size effect is constant with a jump at τ. Since this may result in a misspecification, we apply the Stute test to set $\tau = 28 \sim 35$, for which the OLS model is not rejected; out of this range of τ, the OLS model is either rejected, or one group is too small/big relative to the other group. There are two covariates: the number of enrolled students in the school ('enrol') and the percentage of poor students ('poor'). Table 5.5 provides descriptive statistics and the OLS result for $\tau = 35$ with the regressors 1, D, enrol, enrol^2/100, enrol^3/10000, poor, poor^2/100, poor^3/10000, enrol×poor/10000.

The p-values of the Stute test for $\tau = 28, 29, ..., 35$, and the OLS effect of D and its t-value are in Table 5.6. The OLS effects are all small (0.11∼0.47) and statistically insignificant. The p-value of the Stute test is computed with a bootstrap following Stute et al. [1998], with the bootstrap repetition number 2000. The tests are barely non-rejecting for some values of τ, which is understandable because we are using only a binary transformation of class size, not the class size itself.

Table 5.7 presents the 17 estimates and their t-values for $\tau = 30, 35$, where the 'Mean bias' column is the average of 8 proportional biases defined as |effect−OLS effect|/|OLS effect| for $\tau = 28 \sim 35$. The t-values are computed using the bootstrap Sd based on 500 repetitions (i.e., the Sd of 500 pseudo bootstrap estimates), except for OLS-ps and RI-lin whose t-values use asymptotic variance estimators. Taking the mean bias column as the main performance criterion, OLS-ps has the smallest number (0.16), which is much smaller than the other mean biases, and OLS-ps is followed by RI2-ppgs (0.39), CP1-ppgs (0.40) and CP1-pgs (0.44). RI-lin does not do well with mean bias 0.93. M1-ps does worst (3.98), and the performances of M5-ps (2.86) and DR-c (1.95) are also poor. As a group, the PS-based estimators do worst with highly varying mean biases, and the PGS-based estimators do best with low varying mean biases; if we exclude DR-c, however, the DR estimators do best.

5.5.2 Retirement Effect on Home Food Expenditure

In our second example, we examine the effect of retirement on monthly home food expenditure in Euros, the logarithm of which is Y. Our data are drawn from the Survey of Health, Ageing and Retirement in Europe (SHARE) which is cross-national panel data on health and socioeconomic status of individuals aged 50 or higher. SHARE covers 27 European countries and consists of 6 waves from 2004 to 2015, and we use Estonia for the last two waves because Estonia entered SHARE late. Although our data are panel data, we pool the data to use them as a single cross-section.

The observation unit is a single-earner household. $D = 1$ if the household head is retired, where retirement is defined as non-working and receiving the pension. We control 3 covariates: household size (size), monthly household income including the pension in Euros (income), and marital status of the household head (married). Although other covariates may affect Y, controlling only these three is justified due to the RD nature that using a local sample around c tends to balance most covariates.

We use the localising bandwidth $h = 1.1 \sim 1.8$ years around the retirement age $c = 63$, which gives 8 experiments depending on h; the number of the local observations seems too small for $h < 1.1$, and the Stute test rejects the OLS model for $h > 1.9$. Table 5.8 presents descriptive statistics and the OLS result for $h = 1.8$ with the OLS regressors 1, D, ln(income), married, size and married×size. The OLS R^2 is 0.42, similar to the R^2 in the first empirical example.

Table 5.9 shows h, the local sample size with h, Stute test p-value based on 2000 bootstrap repetitions, OLS effect and its t-value. The local sample sizes are $195 \sim 307$. The Stute test is non-rejecting with the p-value 0.10 or greater, and the p-value decreases as h increases. The OLS effects are all negative and significant: as the household head retires, the household home food expenditure drops by about $13 \sim 20\%$.

Table 5.10 presents the 17 estimates and their t-values for $h = 1.4, 1.8$. Differently from Table 5.7, the bootstrap to get the *Sd*'s for the 17 estimates ran into troubles due to the small sample sizes. It happened sometimes that (i) the PS could not be estimated, when the pseudo sample was highly unbalanced in D; (ii) there were no matching subjects in the opposite group for matching, which occurred mostly for the 1:5 matchings; and (iii) the OLS could not be implemented for RI estimators because the regressor matrix was not invertible. To avoid the problem (i), we drew pseudo samples separately from the $D = 0$ and $D = 1$ groups so that $P(D = 1)$ stays the same. As for (ii) and (iii), whenever they occurred, the pseudo sample was dropped and then redrawn. As in Table 5.7, the *Sd*'s for OLS-ps and RI-lin were obtained with their asymptotic variance estimators.

In Table 5.10, OLS-ps has the smallest mean bias (0.29) by far, followed by DR-c (0.39), MT1-pgs (0.48), M5-pgs (0.50) and RI-lin (0.51). Differently from Table 5.7 for the first empirical example, RI-lin does almost third best, and surprisingly, DR-c does second best. As a group, the PGS-based estimators do best with the group-averaged mean bias 0.58, whereas the PS-based estimators and the DR estimators perform similarly with the group-averaged mean biases 0.88 and 0.89, respectively.

Combining the Tables 5.7 and 5.10 findings with the summaries of the simulation tables, the best performing estimator is clearly OLS-ps, next to which RI-lin, RI2-ppgs, CP1-pgs and CP1-ppgs come. Bear in mind though that there are other factors to take into account in choosing an estimator, such as the requirement of choosing a tuning constant and the ease in estimating the asymptotic variance. Since CP1 needs a bandwidth, *all in all, we would recommend OLS-ps, followed by RI-lin and RI2-ppgs, which are simple estimators with the asymptotic variances easily estimable*; although the order 2 in RI2-ppgs is a bandwidth, practitioners may not see it that way.

5.6 Conclusions

This chapter compared various treatment effect estimators of different types, depending on whether the estimator specifies PS or PGS; the estimators examined in this chapter specifying both PS and PGS are doubly robust (DR). Broadly viewed, we compared five approaches: regression imputation (RI), matching, weighting, complete pairing, and OLS with PS residual. We then carried out an extensive simulation study to compare 24 estimators in 32 or 64 factorial designs to see which estimator performs best. Overall, the OLS with PS residual in Lee [2018] did best which is PS-based. Next to this estimator, RI with linear regression functions did well, which is PGS-based. Both estimators are not DR, but they are simple to obtain with easy-to-compute asymptotic variance estimators. Next to these two, we recommend RI using second-order polynomial functions of PS and PGS, which is DR. These conclusions notwithstanding, a particular method may work better than others in a given data set, which may be seen by a real-data mimicking simulation.

Table 5.1: Bias, Sd, Rmse (×100) for Base Design ($N = 400$)

	(1) $\pi(X),\ \psi(X)$			(2) $\pi(X),\ \psi(Z)$			(3) $\pi(Z),\ \psi(X)$			(4) $\pi(Z),\ \psi(Z)$		
	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse
Estimators with only PS controlled												
RI2-ps	−0.9	5.9	6.0	−0.9	5.8	5.9	−2.7	6.0	6.6	−2.7	6.1	6.7
RI3-ps	0.1	6.2	6.2	0.0	6.1	6.1	−2.7	6.3	6.9	−2.7	6.4	7.0
M1-ps	−0.1	8.8	8.8	−0.2	8.8	8.8	−2.8	9.0	9.4	−2.7	9.1	9.5
M5-ps	−0.1	6.6	6.6	−0.1	6.5	6.5	−2.6	6.8	7.2	−2.5	6.9	7.3
MT1-ps	−0.3	7.9	7.9	−0.4	7.7	7.7	−2.7	8.0	8.4	−2.7	8.0	8.4
Wgt	0.0	6.4	6.4	−0.1	6.3	6.3	−2.0	7.4	7.7	−2.0	7.6	7.8
CP1-ps	−0.3	6.5	6.5	−0.4	6.5	6.5	−2.6	6.6	7.1	−2.6	6.7	7.2
CP2-ps	−3.3	6.2	7.0	−3.4	6.2	7.0	−5.5	6.3	8.4	−5.4	6.4	8.4
OLS-ps	0.2	5.4	5.5	0.2	5.4	5.4	−2.3	5.6	6.1	−2.3	5.7	6.2
Estimators with only PGS controlled												
RI-lin	0.1	5.4	5.4	−2.3	5.7	6.1	−0.1	5.3	5.3	−2.3	5.8	6.2
RI2-pgs	0.0	5.7	5.7	−1.4	6.0	6.2	−0.1	5.6	5.6	−1.5	6.1	6.3
RI3-pgs	0.0	5.7	5.7	−1.3	6.1	6.3	−0.1	5.7	5.7	−1.4	6.2	6.4
M1-pgs	−0.1	6.5	6.5	−1.7	6.8	7.0	−0.2	6.5	6.5	−1.8	6.9	7.1
M5-pgs	−0.1	6.0	6.0	−1.6	6.2	6.4	−0.2	5.9	5.9	−1.7	6.2	6.5
MT1-pgs	−0.3	6.1	6.1	−1.8	6.4	6.7	−0.5	6.0	6.0	−1.9	6.4	6.7
CP1-pgs	−0.4	6.1	6.1	−1.6	6.2	6.4	−0.5	6.0	6.0	−1.7	6.2	6.4
CP2-pgs	−3.2	6.0	6.8	−4.4	6.1	7.5	−3.3	5.9	6.8	−4.5	6.1	7.6
Doubly robust estimators												
RI2-ppgs	0.1	5.6	5.6	−0.3	5.9	5.9	−0.1	5.5	5.5	−2.1	5.9	6.3
M1-ppgs	0.1	6.6	6.6	0.0	6.8	6.8	−0.1	6.5	6.5	−1.9	6.8	7.1
MT1-ppgs	0.1	6.8	6.8	0.0	7.0	7.0	−0.1	6.7	6.7	−1.9	7.1	7.3
M1-bc	0.1	6.4	6.4	0.0	6.7	6.7	−0.1	6.3	6.3	−2.4	6.8	7.2
DR-c	0.1	5.6	5.6	0.1	6.0	6.0	−0.1	5.7	5.7	−3.4	7.4	8.2
CP1-ppgs	0.1	6.4	6.4	−0.1	6.5	6.5	−0.1	6.4	6.4	−1.8	6.6	6.8
CP2-ppgs	−0.3	5.9	6.0	−0.6	6.0	6.0	−0.6	5.9	5.9	−2.2	6.1	6.5

The size of the effect to be estimated is 1. On average, the control group is as large as the treatment group ($N_0 \simeq N_1$). $\pi(X)$ and $\psi(X)$ indicate correct specifications of PS and PGS while $\pi(Z)$ and $\psi(Z)$ indicate mild misspecifications of PS and PGS. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a 'canonical' DR estimator.

Table 5.2: Bias, Sd, Rmse (×100) for Poor PGS or PS Overlap ($N = 400$)

	poor PGS but good PS overlap						poor PS but good PGS overlap					
	(1) $\pi(X)$, $\psi(X)$			(2) $\pi(Z)$, $\psi(Z)$			(3) $\pi(X)$, $\psi(X)$			(4) $\pi(Z)$, $\psi(Z)$		
	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse
Estimators with only PS controlled												
RI2-ps	−1.7	5.8	6.0	−3.2	6.0	6.8	−7.1	8.9	11.4	−8.4	8.7	12.1
RI3-ps	0.1	5.7	5.7	−2.4	6.1	6.5	0.0	10.3	10.3	−4.8	10.2	11.3
M1-ps	−0.1	7.0	7.0	−2.4	7.2	7.6	−1.4	12.9	13.0	−5.7	12.5	13.7
M5-ps	−0.1	5.8	5.8	−1.9	6.1	6.4	−0.2	7.7	7.7	−4.3	7.9	9.0
MT1-ps	−0.5	6.4	6.4	−2.3	6.6	7.0	−0.4	9.9	9.9	−4.2	10.0	10.8
Wgt	−0.2	7.1	7.1	−2.1	8.1	8.4	−1.8	15.3	15.4	−3.0	18.7	18.9
CP1-ps	−0.7	5.8	5.8	−2.2	6.0	6.4	−1.1	6.6	6.7	−4.8	6.8	8.4
CP2-ps	−6.6	5.7	8.7	−8.2	5.9	10.1	−7.0	6.5	9.5	−10.1	6.6	12.0
OLS-ps	0.3	5.4	5.4	−1.8	5.7	5.9	0.1	6.2	6.2	−3.9	6.5	7.6
Estimators with only PGS controlled												
RI-lin	0.1	5.4	5.4	−2.2	5.7	6.1	0.0	6.1	6.1	−3.5	6.4	7.3
RI2-pgs	0.1	5.5	5.5	−2.1	6.0	6.3	−0.1	6.9	6.9	−1.3	7.4	7.5
RI3-pgs	0.1	5.7	5.7	−1.8	6.1	6.4	−0.1	6.9	6.9	−1.0	7.5	7.6
M1-pgs	−0.1	6.4	6.4	−2.1	6.8	7.1	−0.2	7.5	7.5	−1.7	8.0	8.1
M5-pgs	−0.2	5.8	5.8	−2.1	6.0	6.3	−0.3	7.0	7.0	−1.6	7.3	7.5
MT1-pgs	−0.6	6.1	6.1	−2.6	6.3	6.8	−0.7	7.2	7.3	−1.8	7.6	7.8
CP1-pgs	−0.8	5.8	5.9	−2.9	5.9	6.6	−0.8	7.0	7.1	−1.6	7.3	7.5
CP2-pgs	−6.9	5.8	9.0	−8.6	5.9	10.4	−5.4	6.8	8.7	−6.3	7.0	9.4
Doubly robust estimators												
RI2-ppgs	0.2	5.6	5.6	−1.9	5.9	6.3	0.0	7.2	7.2	−3.3	7.6	8.3
M1-ppgs	0.1	6.3	6.3	−1.7	6.5	6.7	0.0	7.5	7.5	−3.3	7.7	8.4
MT1-ppgs	0.1	6.4	6.4	−1.4	6.4	6.6	0.0	8.0	8.0	−3.0	8.2	8.8
M1-bc	0.1	6.5	6.5	−2.4	6.7	7.1	0.0	8.5	8.5	−4.9	8.9	10.1
DR-c	0.1	5.6	5.6	−3.3	7.5	8.2	0.0	9.8	9.8	−9.3	22.1	24.0
CP1-ppgs	0.0	6.1	6.1	−1.5	6.0	6.2	−0.1	7.1	7.1	−3.1	7.3	7.9
CP2-ppgs	−2.2	5.7	6.1	−3.9	5.8	7.0	−0.8	6.5	6.6	−3.6	6.7	7.6

The size of the effect to be estimated is 1. On average, the control group is as large as the treatment group ($N_0 \simeq N_1$). $\pi(X)$ and $\psi(X)$ indicate correct specifications of PS and PGS while $\pi(Z)$ and $\psi(Z)$ indicate mild misspecifications of PS and PGS. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a 'canonical' DR estimator.

Table 5.3: Bias, Sd, Rmse (×100) for Poor PGS and PS Overlap ($N_1 = 200$)

	poor PGS,PS overlap; $N_0 \simeq N_1$						poor PGS,PS overlap; $N_0 \simeq 3N_1$					
	(1) $\pi(X)$, $\psi(X)$			(2) $\pi(Z)$, $\psi(Z)$			(3) $\pi(X)$, $\psi(X)$			(4) $\pi(Z)$, $\psi(Z)$		
	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse
	Estimators with only PS controlled											
RI2-ps	−14.0	8.4	16.3	−14.9	8.5	17.2	−20.2	7.9	21.7	−20.0	7.6	21.4
RI3-ps	0.3	8.2	8.2	−5.3	8.5	10.0	−6.6	8.1	10.5	−10.9	7.9	13.5
M1-ps	−2.6	9.4	9.8	−7.3	9.8	12.3	−5.9	12.5	13.9	−9.1	11.4	14.6
M5-ps	−0.4	6.8	6.8	−3.8	7.0	7.9	−4.5	8.3	9.5	−7.0	7.9	10.5
MT1-ps	−0.9	7.7	7.8	−3.8	7.8	8.7	−0.3	6.2	6.3	−4.3	6.4	7.7
Wgt	−3.8	19.4	19.8	−5.6	21.4	22.1	−3.9	21.9	22.3	−8.7	20.5	22.2
CP1-ps	−2.2	6.3	6.7	−5.2	6.5	8.4	−4.9	6.4	8.1	−6.9	6.4	9.4
CP2-ps	−14.0	6.4	15.3	−16.3	6.5	17.6	−18.5	6.2	19.5	−20.7	6.2	21.6
OLS-ps	0.2	6.2	6.2	−3.4	6.4	7.2	0.0	4.9	4.9	−4.3	5.2	6.7
	Estimators with only PGS controlled											
RI-lin	0.1	6.1	6.1	−3.5	6.5	7.4	0.0	5.9	5.9	−6.2	6.0	8.6
RI2-pgs	−0.1	7.1	7.1	−3.6	7.9	8.6	−0.1	7.3	7.3	−2.5	7.8	8.2
RI3-pgs	−0.1	8.1	8.1	−2.7	9.1	9.5	−0.1	9.6	9.6	−6.8	9.3	11.6
M1-pgs	−0.6	7.6	7.6	−3.4	8.1	8.7	−0.5	7.1	7.1	−3.9	7.3	8.2
M5-pgs	−0.5	6.7	6.7	−3.5	6.9	7.7	−0.5	5.6	5.6	−3.8	5.7	6.9
MT1-pgs	−1.1	7.2	7.3	−4.3	7.4	8.6	−0.7	5.7	5.7	−3.3	6.0	6.8
CP1-pgs	−1.8	6.5	6.7	−5.0	6.6	8.3	−1.4	5.2	5.3	−4.4	5.3	6.9
CP2-pgs	−13.3	6.4	14.8	−15.8	6.5	17.1	−10.9	5.1	12.1	−13.3	5.2	14.2
	Doubly robust estimators											
RI2-ppgs	0.1	7.4	7.4	−3.5	7.9	8.6	0.0	7.9	7.9	−2.6	9.0	9.4
M1-ppgs	0.1	7.6	7.6	−3.1	7.6	8.2	−0.2	7.3	7.3	−3.8	7.3	8.3
MT1-ppgs	0.0	7.5	7.5	−2.5	7.6	8.0	−0.1	6.0	6.0	−3.1	6.1	6.9
M1-bc	0.2	8.5	8.5	−4.7	8.9	10.1	0.0	10.4	10.4	−6.4	10.3	12.1
DR-c	0.1	8.9	8.9	−9.5	24.2	26.0	−0.1	10.0	10.0	−9.7	17.1	19.7
CP1-ppgs	−0.1	6.7	6.7	−2.6	6.7	7.2	−0.2	6.0	6.0	−3.1	6.1	6.8
CP2-ppgs	−4.3	6.3	7.6	−7.1	6.3	9.5	−3.8	5.5	6.6	−7.0	5.5	8.9

The size of the effect to be estimated is 1. On average, the control group is as large as the treatment group in columns (1) and (2) ($N_0 \simeq N_1$), while the control group is three times larger than the treatment group in columns (3) and (4) ($N_0 \simeq 3N_1$). $\pi(X)$ and $\psi(X)$ indicate correct specifications of PS and PGS while $\pi(Z)$ and $\psi(Z)$ indicate mild misspecifications of PS and PGS. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a ‘canonical’ DR estimator.

Table 5.4: Bias, Sd, Rmse (×100) with/without modifier $\mu(X)$

	$\mu(X) = 1 - X_2$ for $E(Y^1{-}Y^0\|X)$						$\mu(X) = \|1 + X_2{+}X_3\|$ for $Sd(Y^1\|X)$					
	(1) $\mu(X)$ not used			(2) $\mu(X)$ used			(3) $\mu(X)$ not used			(4) $\mu(X)$ used		
	bias	sd	rmse	bias	sd	ratio	bias	sd	rmse	bias	sd	ratio
	Estimators with only PS controlled											
RI2-ps	−0.5	7.7	7.8	−0.3	7.8	1.00	−0.9	5.0	5.1	−0.9	5.0	0.99
RI3-ps	−0.2	7.9	7.9	0.1	7.9	1.00	0.0	5.3	5.3	0.0	5.2	0.98
M1-ps	−0.3	9.9	9.9	0.1	9.8	0.99	−0.2	8.0	8.0	−0.1	7.8	0.98
M5-ps	−0.2	8.7	8.7	0.0	8.6	0.99	−0.2	5.6	5.6	−0.1	5.6	0.99
MT1-ps	−0.1	9.4	9.4	0.1	9.3	0.99	−0.3	7.2	7.2	−0.1	7.1	0.99
Wgt	−0.2	8.2	8.2	0.0	8.2	1.00	−0.1	5.7	5.7	−0.1	5.6	0.98
CP1-ps	−0.3	8.4	8.4	−0.1	8.4	1.00	−0.4	5.4	5.5	−0.4	5.4	0.99
CP2-ps	−1.4	8.1	8.2	−1.3	8.1	0.99	−3.4	5.2	6.2	−3.4	5.2	1.00
OLS-ps	0.0	7.9	7.9	0.2	7.8	1.00	0.1	4.4	4.4	0.2	4.4	1.00
	Estimators with only PGS controlled											
RI-lin	−0.1	7.3	7.3	0.0	7.3	1.00	0.0	4.4	4.4	0.0	4.3	1.00
RI2-pgs	6.9	7.8	10.4	0.1	7.4	0.71	0.0	4.9	4.9	0.0	4.7	0.95
RI3-pgs	6.9	7.9	10.5	0.1	7.4	0.71	−0.1	5.0	5.0	0.0	4.9	0.99
M1-pgs	6.8	8.9	11.2	0.1	9.2	0.82	−0.1	5.6	5.6	0.3	5.1	0.92
M5-pgs	6.8	8.6	10.9	0.1	15.2	1.39	−0.2	5.0	5.0	0.3	5.0	1.01
MT1-pgs	6.8	8.9	11.2	0.1	9.5	0.85	−0.5	5.4	5.4	0.5	5.4	1.01
CP1-pgs	6.6	8.8	11.0	−0.1	9.6	0.88	−0.5	4.8	4.8	0.5	4.5	0.93
CP2-pgs	5.0	8.5	9.9	−1.2	8.7	0.89	−3.3	4.8	5.8	0.8	4.6	0.80
	Doubly robust estimators											
RI2-ppgs	0.1	7.6	7.6	0.0	7.5	0.99	0.0	4.6	4.6	0.0	4.7	1.01
M1-ppgs	−0.1	9.5	9.5	0.4	15.3	1.61	−0.1	5.2	5.2	0.0	5.9	1.12
MT1-ppgs	0.0	10.0	10.0	0.3	16.3	1.63	−0.1	5.7	5.7	0.1	6.6	1.14
M1-bc	−0.1	8.3	8.3	0.1	8.2	0.99	0.0	5.2	5.2	0.1	5.2	1.00
DR-c	−0.1	7.4	7.5	0.0	7.5	1.00	0.0	4.7	4.7	0.0	4.8	1.00
CP1-ppgs	0.0	9.6	9.6	0.2	12.7	1.32	0.0	4.7	4.7	0.1	5.2	1.11
CP2-ppgs	0.5	8.6	8.7	0.0	9.3	1.07	−0.4	4.3	4.4	0.3	4.4	1.01

The size of the effect to be estimated is 1 although the effect is heterogeneous in columns (1) and (2) and the error term in Y^1 is heteroskedastic in columns (3) and (4). On average, the control group is as large as the treatment group ($N_0 \simeq N_1$). '$\mu(X)$ not used' means that a modifier $\mu(X)$ is not controlled in estimation and vice versa. $\pi(X)$ and $\psi(X)$ indicate correct specifications of PS and PGS while $\pi(Z)$ and $\psi(Z)$ indicate mild misspecifications of PS and PGS. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a 'canonical' DR estimator.

Table 5.5: Descriptive Statistics & Estimate (t-value) for OLS with $\tau = 35$

Variable	Mean (Sd)	Min,Max	Regressor	Est. (tv)	Regressor	Est. (tv)
Score Y	74 (7.7)	35, 94	D	0.47 (1.4)	poor	-0.95 (-13)
Class size	30 (6.6)	5, 47	enrol	-0.016 (-0.42)	poor$^2/10^2$	1.8 (7.3)
Enrol	77 (37)	5, 208	enrol$^2/10^2$	-0.019 (-0.48)	poor$^3/10^4$	-1.4 (-5.1)
Poor	14 (14)	0, 76	enrol$^3/10^4$	0.015 (1.2)	enrol×poor$/10^4$	11 (2.2)

The outcome variable Y is reading test score and the treatment variable is $D = 1[\text{class size} \leq \tau]$. Enrol indicates # enrolled in the school; Poor indicates the percentage of poor students in the school; the OLS estimate of an intercept is not shown; Est indicates estimates; tv indicates t-value. $R^2 = 0.41$ and $N = 1963$.

Table 5.6: Nonparametric Specification Test & OLS Effect

τ	28	29	30	31	32	33	34	35
Stute test p-value	0.053	0.086	0.12	0.12	0.12	0.11	0.076	0.050
OLS effect of D	0.26	0.11	0.11	0.11	0.35	0.39	0.42	0.47
effect t-value	0.72	0.32	0.35	0.36	1.12	1.26	1.35	1.41

Class size is transformed into a binary variable at threshold τ (i.e. $D = 1[\text{class size} \leq \tau]$). OLS effect of D indicates the OLS estimates of the effect of the treatment D.

Table 5.7: Class Size Effect Estimates & T-Value (tv)

	Effect (tv): $\tau = 30$	Effect (tv): $\tau = 35$	Mean bias
	Estimators with only PS controlled		
RI2-ps	0.03 (0.06)	0.54 (1.00)	0.78
M1-ps	1.65 (2.10)	2.02 (2.44)	3.98
M5-ps	0.85 (1.42)	1.52 (2.12)	2.86
MT1-ps	0.11 (0.24)	-0.07 (-0.17)	0.71
CP1-ps	-0.13 (-0.36)	0.26 (0.52)	1.10
OLS-ps	0.14 (0.44)	0.43 (1.29)	0.16
	Estimators with only PGS controlled		
RI-lin	0.37 (0.86)	0.71 (1.39)	0.93
RI2-pgs	0.01 (0.02)	0.63 (1.03)	0.73
M1-pgs	0.31 (0.66)	0.66 (1.08)	1.04
M5-pgs	0.29 (0.68)	0.55 (1.03)	0.79
MT1-pgs	0.04 (0.08)	0.35 (0.74)	0.58
CP1-pgs	0.19 (0.51)	0.61 (1.36)	0.44
	Doubly robust estimators		
RI2-ppgs	0.21 (0.54)	0.80 (1.41)	0.39
M1-ppgs	0.23 (0.50)	0.49 (0.88)	0.89
MT1-ppgs	0.31 (0.71)	0.48 (1.05)	0.96
DR-c	0.61 (1.28)	0.75 (1.35)	1.95
CP1-ppgs	0.21 (0.52)	0.56 (1.19)	0.40

Class size is transformed into a binary variable at threshold τ (i.e. $D = 1[\text{class size} \leq \tau]$). RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; $D = 1[\text{class size} \leq \tau]$; mean bias indicates the average of 8 proportional biases.

Table 5.8: Descriptive Statistics & Estimate (t-value) for OLS with $h = 1.8$

Variable	Mean (Sd)	Min, Max	Regressor	Est. (tv)
exp(Y)	219 (198)	57.5, 3196	D	-0.16 (-2.86)
Retired (D)	0.78 (0.42)	0, 1	ln(income)	0.14 (2.76)
Income	724 (850)	128, 9587	married	0.87 (5.13)
Married	0.56 (0.50)	0, 1	size	0.35 (5.10)
Size	1.79 (0.66)	1, 5	married×size	-0.31 (-3.42)

Descriptive statistics are computed from a local sample with a bandwidth 1.8 years around retirement age of 63 ($h = 1.8$ and $c = 63$) and the size of the local sample is 307 (N=307). Outcome variable Y is ln(food expenditure) and the treatment is the retirement of the household head; size indicates household size; the OLS estimate of an intercept is not shown; Est indicates estimates; tv indicates t-value; R^2=0.42.

Table 5.9: Nonparametric Specification Test & OLS Effect

h	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
Local sample size	195	210	221	234	261	275	291	307
Stute test p-value	0.65	0.56	0.43	0.28	0.14	0.12	0.13	0.10
OLS effect of D	-0.15	-0.13	-0.18	-0.20	-0.19	-0.20	-0.18	-0.16
effect t-value	-2.0	-1.9	-2.6	-3.0	-3.1	-3.3	-3.0	-2.9

The figures in the table are computed from local samples with different bandwidths ($h's$) from 1.1 to 1.8 years around retirement age of 63. OLS effect of D indicates the OLS estimates of the effect of the treatment D.

Table 5.10: Retirement Effect Estimates & T-Value (tv)

	Effect (tv): $h = 1.4$	Effect (tv): $h = 1.8$	Mean bias
	Estimators with only PS controlled		
RI2-ps	-0.02 (-0.22)	-0.04 (-0.51)	0.89
M1-ps	0.04 (0.40)	-0.06 (-0.68)	1.26
M5-ps	0.05 (0.40)	-0.03 (-0.26)	1.00
MT1-ps	-0.01 (-0.06)	-0.04 (-0.36)	0.77
CP1-ps	0.02 (0.18)	-0.04 (-0.43)	1.06
OLS-ps	-0.12 (-2.19)	-0.13 (-2.46)	0.29
	Estimators with only PS controlled		
RI-lin	-0.10 (-1.75)	-0.09 (-1.66)	0.51
RI2-pgs	-0.08 (-1.05)	-0.06 (-1.04)	0.66
M1-pgs	-0.06 (-0.84)	-0.10 (-1.53)	0.70
M5-pgs	-0.05 (-0.46)	-0.10 (-1.26)	0.50
MT1-pgs	-0.10 (-1.10)	-0.07 (-0.93)	0.48
CP1-pgs	-0.08 (-0.84)	-0.05 (-0.71)	0.65
	Doubly robust estimators		
RI2-ppgs	-0.02 (-0.34)	-0.00 (-0.01)	0.99
M1-ppgs	0.04 (0.42)	-0.02 (-0.26)	1.17
MT1-ppgs	0.02 (0.19)	0.01 (0.08)	0.98
DR-c	-0.11 (-0.52)	-0.08 (-0.82)	0.39
CP1-ppgs	-0.03 (-0.27)	-0.02 (-0.25)	0.94

Outcome variable Y is ln(food expenditure) and the treatment is the retirement of the household head. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; mean bias indicates the average of 8 proportional biases.

Appendix

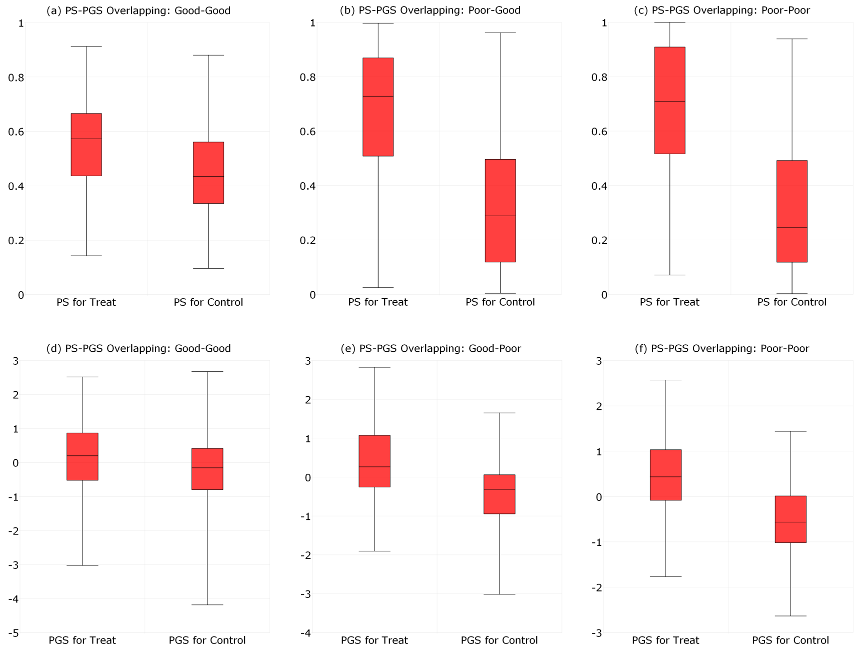

Figure 5.A1: PS and PGS overlaps depending on $\sigma_\varepsilon = 1, 2$ and $\alpha_4 = -1, 1$

Good PS overlap indicates that $\sigma_\varepsilon = 2$ (the variance of error terms in D), while poor PS overlap indicates that $\sigma_\varepsilon = 1$. Good PGS overlap indicates that $\alpha_4 = -1$ (the slope of X_4 in D), while poor PGS overlap indicates that $\alpha_4 = 1$. Poor PS overlap is shown whereas PGS overlap is good in (b); poor PGS overlap is shown whereas PS overlap is good in (e). On average, $E(\pi(X)|D = 0)$ and $E(\pi(X)|D = 1)$ (Sd's) are about 0.4 and 0.6 (0.03 and 0.03) for good PS overlap; $E(\pi(X)|D = 0)$ and $E(\pi(X)|D = 1)$ (Sd's) are about 0.3 and 0.7 (0.03 and 0.03) for poor PS overlap; $E(\psi(X)|D = 0)$ and $E(\psi(X)|D = 1)$ (Sd's) are about -0.2 and 0.2 (0.1 and 0.1) for good PGS overlap; $E(\psi(X)|D = 0)$ and $E(\psi(X)|D = 1)$ (Sd's) are about -0.3 and 0.3 (0.1 and 0.1) for poor PGS overlap. The exceptions are that the differences of $\pi(X)$ and $\psi(X)$ between $D = 0, 1$ increase with the combination of poor PS and poor PGS overlap and with $N_0 \simeq 3N_1$.

In Table 5.A1 with four panels, we compare M1 to M5 and RI2 to RI3 using relative rmse ratios; the left half is for $N_0 \simeq N_1$, and the right half is for $N_0 \simeq 3N_1$. In the first panel for good PGS and PS overlap, the column 'False↓' shows what is misspecified, and the two columns for M1/M5 show the rmse ratios for matching using PS or PGS; the other three panels can be analogously understood. When $N_0 \simeq N_1$ in the left half, M5 is better, which is also the case when $N_0 \simeq 3N_1$ in the right half. Regarding RI2 v. RI3, RI3 is mostly better when PS is used, but RI2 is better when PGS is used. Gaps

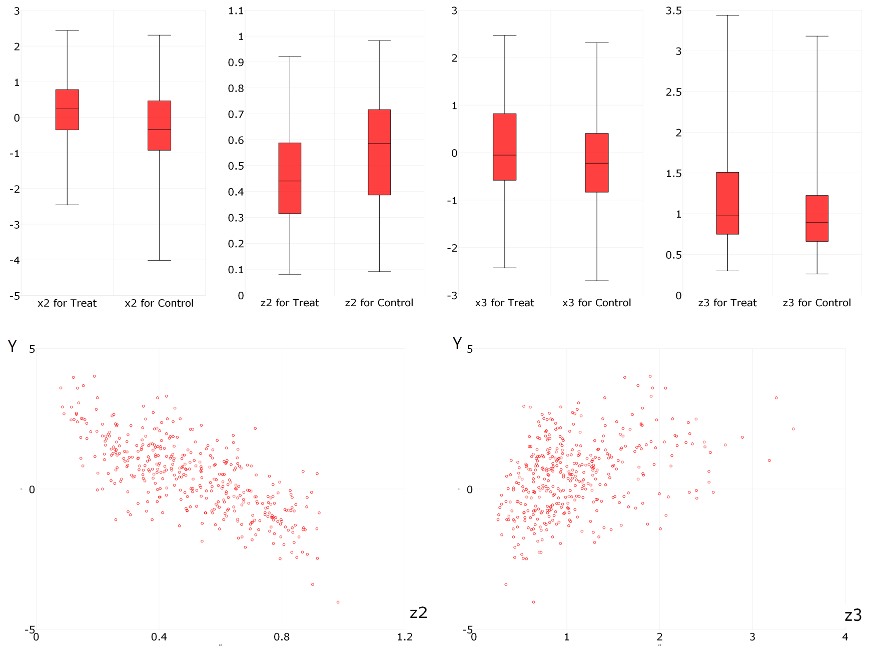

Figure 5.A2: Similarity between X & Z (upper) and Z explaining Y well (lower)

Outcome variable Y is linear in (X_2, X_3). Variables (Z_2, Z_3) are observed, while variables (X_2, X_3) are not observed. The relationship between X's and Z's is that $Z_2 = \{1 + \exp(X_2)\}^{-1}, Z_3 = \exp(X_3/2)$.

between RI2 and RI3 become larger when both scores overlap poorly.

Tables 5.A2–5.A5 show rmse's for the 32 designs where no effect modifier appears. These tables supplement the tables in the main text, because results for certain designs were not presented in the main text to save space; the four columns for 'Base Design' repeat the rmse columns in Table 5.1. For example, no result was shown for when only one of PGS and PS is misspecified in Table 5.2 with only PGS overlapping poorly, and if the reader desires the rmse's when only PGS is wrong and $N_0 \simeq N_1$ with only PGS overlapping poorly, then the desired rmse's can be found in the PGS column within the 'Poor PGS overlap' column of Table 5.A2.

Table 5.A1: Rmse Ratio Comparison of 1:1 v. 1:5 Matchings & RI2 v. RI3

		$N_0 \simeq N_1$					$N_0 \simeq 3N_1$			
		M1/M5		RI2/RI3			M1/M5		RI2/RI3	
	False↓	PS	PGS	PS	PGS	False↓	PS	PGS	PS	PGS
For		1.33	1.09	0.97	1.00		1.45	1.12	0.98	0.99
good	PGS	1.34	1.09	0.97	0.99	PGS	1.44	1.12	0.99	0.99
pgs,ps	PS	1.30	1.09	0.96	0.99	PS	1.34	1.11	1.01	0.99
overlap	Both	1.30	1.10	0.96	0.99	Both	1.33	1.11	1.01	0.99
For		1.20	1.11	1.05	0.98		1.26	1.17	1.19	0.95
poor	PGS	1.19	1.11	1.06	0.99	PGS	1.26	1.15	1.19	0.95
pgs	PS	1.19	1.11	1.04	0.98	PS	1.24	1.15	1.16	0.95
overlap	Both	1.19	1.12	1.05	1.00	Both	1.22	1.13	1.15	0.95
For		1.69	1.07	1.11	0.99		1.77	1.11	1.21	0.98
poor	PGS	1.68	1.08	1.10	0.99	PGS	1.75	1.13	1.20	0.99
ps	PS	1.53	1.07	1.07	0.99	PS	1.50	1.12	1.21	0.98
overlap	Both	1.52	1.08	1.07	0.99	Both	1.49	1.12	1.22	0.99
For		1.44	1.14	1.98	0.87		1.46	1.27	2.07	0.76
poor	PGS	1.44	1.14	2.03	0.91	PGS	1.44	1.21	2.07	0.69
pgs,ps	PS	1.53	1.14	1.74	0.87	PS	1.35	1.25	1.58	0.76
overlap	Both	1.54	1.13	1.72	0.91	Both	1.39	1.20	1.59	0.71

The first column in the table indicate the overlap of PS and PGS. PS and PGS in the head of the table indicate scores to be controlled, while PS and PGS in the second and seventh columns indicate scores with misspecifications. RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively.

Table 5.A2: Rmse (×100) for good PS/PGS overlap and poor PGS overlap ($N_0 \simeq N_1$)

	Base Design				Poor PGS Overlap			
False:		PGS	PS	Both		PGS	PS	Both
			Estimators with only PS controlled					
RI2-ps	6.0	5.9	6.6	6.7	6.0	6.1	6.8	6.8
RI3-ps	6.2	6.1	6.9	7.0	5.7	5.7	6.5	6.5
M1-ps	8.8	8.8	9.4	9.5	7.0	6.9	7.5	7.6
M5-ps	6.6	6.5	7.2	7.3	5.8	5.8	6.4	6.4
MT1-ps	7.9	7.7	8.4	8.4	6.4	6.4	7.0	7.0
Wgt	6.4	6.3	7.7	7.8	7.1	7.1	8.8	8.4
CP1-ps	6.5	6.5	7.1	7.2	5.8	5.9	6.4	6.4
CP2-ps	7.0	7.0	8.4	8.4	8.7	8.8	10.0	10.1
OLS-ps	5.5	5.4	6.1	6.2	5.4	5.4	5.9	5.9
			Estimators with only PGS controlled					
RI-lin	5.4	6.1	5.3	6.2	5.4	6.2	5.4	6.1
RI2-pgs	5.7	6.2	5.6	6.3	5.5	6.4	5.6	6.3
RI3-pgs	5.7	6.3	5.7	6.4	5.7	6.5	5.7	6.4
M1-pgs	6.5	7.0	6.5	7.1	6.4	7.0	6.4	7.1
M5-pgs	6.0	6.4	5.9	6.5	5.8	6.4	5.7	6.3
MT1-pgs	6.1	6.7	6.0	6.7	6.1	6.9	6.1	6.8
CP1-pgs	6.1	6.4	6.0	6.4	5.9	6.6	5.8	6.6
CP2-pgs	6.8	7.5	6.8	7.6	9.0	10.5	9.0	10.4
			Doubly robust estimators					
RI2-ppgs	5.6	5.9	5.5	6.3	5.6	5.9	5.6	6.3
M1-ppgs	6.6	6.8	6.5	7.1	6.3	6.5	6.4	6.7
MT1-ppgs	6.8	7.0	6.7	7.3	6.4	6.5	6.3	6.6
M1-bc	6.4	6.7	6.3	7.2	6.5	6.9	6.4	7.1
DR-c	5.6	6.0	5.7	8.2	5.6	6.1	5.9	8.2
CP1-ppgs	6.4	6.5	6.4	6.8	6.1	6.0	6.1	6.2
CP2-ppgs	6.0	6.0	5.9	6.5	6.1	6.3	6.3	7.0

RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a ‘canonical’ DR estimator.

Table 5.A3: Rmse (×100) for poor PS overlap and poor PS/PGS overlap ($N_0 \simeq N_1$)

	Poor PS Overlap				Poor PGS, PS Overlap			
False:		PGS	PS	Both		PGS	PS	Both
				Estimators with only PS controlled				
RI2-ps	11.4	11.3	12.1	12.1	16.3	16.4	17.2	17.2
RI3-ps	10.3	10.3	11.3	11.3	8.2	8.1	9.9	10.0
M1-ps	13.0	12.9	13.7	13.7	9.8	9.8	12.2	12.3
M5-ps	7.7	7.7	9.0	9.0	6.8	6.8	8.0	7.9
MT1-ps	9.9	9.8	10.8	10.8	7.8	7.7	8.7	8.7
Wgt	15.4	14.8	18.3	18.9	19.8	19.3	22.2	22.1
CP1-ps	6.7	6.6	8.4	8.4	6.7	6.7	8.3	8.4
CP2-ps	9.5	9.5	12.0	12.0	15.3	15.4	17.6	17.6
OLS-ps	6.2	6.3	7.6	7.6	6.2	6.2	7.2	7.2
				Estimators with only PGS controlled				
RI-lin	6.1	7.4	6.1	7.3	6.1	7.3	6.1	7.4
RI2-pgs	6.9	7.5	6.9	7.5	7.1	8.8	7.1	8.6
RI3-pgs	6.9	7.6	7.0	7.6	8.1	9.7	8.1	9.5
M1-pgs	7.5	8.2	7.6	8.1	7.6	8.8	7.6	8.7
M5-pgs	7.0	7.6	7.1	7.5	6.7	7.7	6.7	7.7
MT1-pgs	7.3	7.8	7.3	7.8	7.3	8.6	7.3	8.6
CP1-pgs	7.1	7.5	7.2	7.5	6.7	8.3	6.8	8.3
CP2-pgs	8.7	9.4	8.7	9.4	14.8	17.1	14.9	17.1
				Doubly robust estimators				
RI2-ppgs	7.2	7.7	7.2	8.3	7.4	8.4	7.3	8.6
M1-ppgs	7.5	7.8	7.5	8.4	7.6	7.7	7.4	8.2
MT1-ppgs	8.0	8.4	8.0	8.8	7.5	7.6	7.5	8.0
M1-bc	8.5	9.0	8.4	10.1	8.5	9.1	8.2	10.1
DR-c	9.8	11.9	11.9	24.0	8.9	10.7	13.4	26.0
CP1-ppgs	7.1	7.4	7.1	7.9	6.7	6.7	6.8	7.2
CP2-ppgs	6.6	6.8	6.6	7.6	7.6	8.0	8.0	9.5

RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a ‘canonical’ DR estimator.

Table 5.A4: Rmse ($\times 100$) for good PS/PGS overlap and poor PGS overlap ($N_0 \simeq 3N_1$)

		Base Design				Poor PGS Overlap		
False:		PGS	PS	Both		PGS	PS	Both
			Estimators with only PS controlled					
RI2-ps	6.0	5.9	6.9	6.9	6.3	6.3	7.1	7.1
RI3-ps	6.1	6.0	6.8	6.8	5.3	5.3	6.1	6.2
M1-ps	8.7	8.8	9.5	9.3	6.4	6.5	7.0	6.9
M5-ps	6.0	6.1	7.1	7.0	5.1	5.1	5.6	5.7
MT1-ps	7.2	7.2	8.0	7.9	5.6	5.6	6.3	6.3
Wgt	7.4	7.5	7.4	7.5	9.1	9.4	8.6	8.7
CP1-ps	5.7	5.6	6.6	6.7	5.0	5.0	5.5	5.6
CP2-ps	6.5	6.4	8.2	8.3	8.5	8.5	10.0	10.1
OLS-ps	4.4	4.4	5.3	5.3	4.3	4.3	5.1	5.1
			Estimators with only PGS controlled					
RI-lin	4.7	5.9	4.7	6.0	4.7	6.0	4.6	6.0
RI2-pgs	4.5	5.1	4.5	5.1	4.9	5.7	4.8	5.6
RI3-pgs	4.5	5.1	4.5	5.1	5.1	6.0	5.1	5.9
M1-pgs	5.2	5.8	5.2	5.8	5.5	6.3	5.4	6.2
M5-pgs	4.6	5.2	4.6	5.2	4.7	5.5	4.7	5.4
MT1-pgs	5.2	5.8	5.2	5.8	5.1	5.8	5.1	5.9
CP1-pgs	4.7	5.1	4.7	5.1	4.7	5.5	4.7	5.6
CP2-pgs	5.3	6.2	5.4	6.3	7.3	8.8	7.3	8.9
			Doubly robust estimators					
RI2-ppgs	4.9	5.1	4.8	5.5	5.0	5.2	4.9	5.9
M1-ppgs	5.4	5.5	5.4	5.9	5.3	5.5	5.4	5.9
MT1-ppgs	5.6	5.7	5.6	6.1	5.3	5.4	5.4	5.8
M1-bc	5.8	6.2	5.8	6.7	5.8	6.2	5.8	6.6
DR-c	5.1	5.6	4.9	6.6	5.0	5.6	4.9	6.7
CP1-ppgs	5.3	5.4	5.3	5.7	5.0	5.0	5.1	5.4
CP2-ppgs	5.0	5.0	5.0	5.6	5.1	5.3	5.3	6.1

RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a 'canonical' DR estimator.

Table 5.A5: Rmse ($\times 100$) for poor PS overlap and poor PS/PGS overlap ($N_0 \simeq 3N_1$)

	Poor PS Overlap				Poor PGS, PS Overlap			
False:		PGS	PS	Both		PGS	PS	Both
Estimators with only PS controlled								
RI2-ps	13.6	13.5	15.0	15.0	21.7	21.7	21.4	21.4
RI3-ps	11.2	11.2	12.4	12.3	10.5	10.4	13.6	13.5
M1-ps	18.7	18.7	18.5	18.4	13.9	13.7	14.4	14.6
M5-ps	10.6	10.7	12.3	12.4	9.5	9.5	10.7	10.5
MT1-ps	8.3	8.2	9.6	9.7	6.3	6.3	7.9	7.7
Wgt	17.7	17.8	17.6	17.1	22.3	21.9	22.4	22.2
CP1-ps	8.1	8.1	10.2	10.3	8.1	8.1	9.5	9.4
CP2-ps	11.5	11.5	14.5	14.5	19.5	19.5	21.7	21.6
OLS-ps	5.0	5.0	7.1	7.1	4.9	5.0	6.8	6.7
Estimators with only PGS controlled								
RI-lin	5.8	8.6	5.9	8.6	5.9	8.6	5.9	8.6
RI2-pgs	5.2	5.8	5.2	5.8	7.3	8.1	7.2	8.2
RI3-pgs	5.3	5.8	5.3	5.9	9.6	11.6	9.5	11.6
M1-pgs	5.9	6.5	5.8	6.5	7.1	8.3	7.0	8.2
M5-pgs	5.3	5.8	5.2	5.8	5.6	6.9	5.6	6.9
MT1-pgs	5.7	6.2	5.7	6.2	5.7	6.9	5.8	6.8
CP1-pgs	5.3	5.6	5.3	5.7	5.3	6.9	5.4	6.9
CP2-pgs	6.7	7.5	6.7	7.6	12.1	14.2	12.1	14.2
Doubly robust estimators								
RI2-ppgs	6.4	7.0	6.3	7.2	7.9	7.8	7.5	9.4
M1-ppgs	7.0	7.3	6.9	8.1	7.3	7.5	7.3	8.3
MT1-ppgs	6.4	6.7	6.5	7.4	6.0	6.2	6.1	6.9
M1-bc	10.4	11.1	9.9	11.8	10.4	11.0	9.8	12.1
DR-c	9.9	11.6	9.9	19.4	10.0	11.5	9.5	19.7
CP1-ppgs	6.5	6.7	6.5	7.4	6.0	6.0	6.2	6.8
CP2-ppgs	5.8	6.1	5.9	7.2	6.6	7.2	7.1	8.9

RI# indicates regression imputation with polynomial #; M1 and M5 indicate 1:1 and 1:5 matching, respectively; MT1 indicates 1:1 matching for the effect on the treated; Wgt indicates weighting; CP# indicates complete pairing with bandwidth #; OLS-ps indicates an OLS estimator with PS residual; bc indicates bias-corrected version; DR-c indicates a ‘canonical’ DR estimator.

Chapter 6

Conclusion

This thesis has discussed MSL estimation when dynamic models of recurrent events are estimated with censored data and doubly robust estimation when treatment effects are estimated with missing data at random.

In Chapter 2, we develop MSL estimation in the context of estimation of continuous-time dynamic models of recurrent events using censored data. In MSL estimation, missing data due to censoring are integrated out of the likelihood function via Monte Carlo and importance sampling techniques. In particular, we focus on the importance sampling method and we consider an idea of normalising the likelihood ratio of the true distribution to importance sampling distributions. The main difficulty in this context is the unknown dimension of missing data as well as the unknown values of missing data. For comparison, we consider ML estimation using only the data that are complete until the end of the observation period or using a reduced form approximation for missing data. We conduct a small Monte Carlo study with information on the true parameter. In an empirical application, we analyse New Zealand administrative data to estimate a dynamic model of an IHD event. We find that MSL estimation is feasible in this context and that there is substantial efficiency gain from MSL estimation relative to alternative methods in both the Monte Carlo study and the empirical application.

In Chapter 3, we describe and quantify the risk of experiencing AMI event among male and female people of European and Maori descent in New Zealand. We analyse high-quality administrative data on hospital admissions and death registrations and estimate dynamic models of AMI events. The analysis data include plenty of left-censored histories so we employ the MSL estimation method developed in Chapter 2. The models allow risk to vary with age, previous AMI history, and unobserved heterogeneity. Our main findings are as follows. The risk of subsequent events is far higher than the risk of the first event, and particularly high within 1 year after an event. In most cases, male Maoris have the highest risk, followed by female Maoris, then male Europeans, while female Europeans have the lowest risk. The risk increases strongly with age. The large influence of the random effects and the dynamic effects of previous AMI history imply that the risk tends to concentrate on the small proportion of high risk people.

In Chapter 4, we develop the formal theory of 'doubly robust' estimation. Formally, we show that 'double robustness' can be achieved by controlling both PS and PGS in various ways, regardless of controlling methods.

In Chapter 5, we compare various treatment effect estimators through an extensive simulation study using 64 designs and two empirical examples mimicking experiments. In total, we examine 24 estimators based on matching, weighting, double robustness, regression imputation/adjustment, 'complete pairing', and 'propensity-score residual'. Our results show that contrary to the common perception, doubly robust estimators are not necessarily the best. In fact, our findings recommend a couple of non-doubly-robust estimators, with a simple propensity-score-based estimator being the nearly dominant best estimator.

6.1 Future Work

In future research, I am considering extending MSL estimation in Chapter 2 in two directions. One relates to the choice of good importance sampling distributions in the same context. In the present study, the choice of importance sampling distribution is heuristic. There exists a substantial literature on the choice of importance sampling distributions. General ideas in the literature, however, do not seem to work in a complex setting, which is the case in the present study. A methodical approach to choose a good importance sampling distribution is likely to improve MSL estimation in this context.

The other extension I am considering relates to multi-state models. I am considering extending MSL estimation to continuous-time or discrete-time multi-state event history models. Multi-state event history models are important in empirical research as they are widely used to model employment status, poverty status, welfare status, etc. My hope is that there will be substantial efficiency gains in multi-state models similar to those we found for recurrent event models in Chapter 2. In particular, while there are many studies on discrete-time duration models in the literature, the extension to continuous-time multi-state dynamic models will be the first paper to apply MSL method in the context of continuous-time multi-state dynamic event history models.

As for Chapter 3, I am considering augmenting the dynamic models of an AMI event with mortality models. In the present study, we focus on the distribution of AMI risk across gender and ethnic groups and highlight cumulative life-time outcomes in a so-called experimental setting where no one dies using the estimated models. However, if mortality models are augmented, the dynamic models of an AMI event will be more useful from the perspective of policy makers. Further, I am considering estimating a competing risk model of death using the same data, probably with more variables. The research question is to describe how the cause-specific death rates differ across gender and ethnic groups and explain how the differences in the cause-specific risks contribute to differences in the overall distribution of cause of death.

Regarding doubly robust estimation, I am considering extending the present studies to multi-valued treatments. In reality, treatments are often multi-valued. Also, comparing potential outcomes for multi-valued treatments is not as simple as for binary treatments. Therefore, the extension to multi-valued treatments is worth investigating. In addition, the prognostic score itself has an interesting feature: it does not involve treatment variables. Therefore, the prognostic score approach may be applicable to regression discontinuity designs where controlling the propensity score is difficult or infeasible. I believe this topic may also be interesting and worthwhile investigating in the future.

Bibliography

Abadie, A.; Drukker, D.; Herr, J. L.; and Imbens, G. W., 2004. Implementing matching estimators for average treatment effects in Stata. *Stata Journal*, 4, 3 (2004), 290–311. (cited on page 94)

Abadie, A. and Imbens, G. W., 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29, 1 (2011), 1–11. (cited on page 94)

Abadie, A. and Imbens, G. W., 2016. Matching on the estimated propensity score. *Econometrica*, 84, 2 (2016), 781–807. (cited on page 88)

Abildstrom, S. Z.; Rasmussen, S.; Rosen, M.; et al., 2003. Trends in incidence and case fatality rates of acute myocardial infarction in Denmark and Sweden. *Heart*, 89, 5 (2003), 507–511. (cited on page 51)

Alderman, M. H.; Cohen, H. W.; and Madhavan, S., 2000. Myocardial infarction in treated hypertensive patients: The paradox of lower incidence but higher mortality in young blacks compared with whites. *Circulation*, 101, 10 (2000), 1109–1114. (cited on page 52)

Anand, S. S.; Yusuf, S.; Vuksan, V.; et al., 2000. Differences in risk factors, atherosclerosis, and cardiovascular disease between ethnic groups in Canada: The Study of Health Assessment and Risk in Ethnic groups (SHARE). *Lancet*, 356, 9226 (2000), 279–284. (cited on page 51)

Angrist, J. D. and Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 2 (1999), 533–575. (cited on page 105)

Austin, P. C., 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 12 (2008), 2037–2049.

(cited on page 88)

Avendano, M. and Soerjomataram, I., 2008. Monitoring trends in acute coronary syndromes: Can we use hospital admission registries? *Heart*, 94, 12 (2008), 1524–1525. (cited on page 51)

Bang, H. and Robins, J. M., 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 4 (2005), 962–973. (cited on pages 82 and 95)

Bhuller, M.; Brinch, C. N.; and Königs, S., 2017. Time aggregation and state dependence in welfare receipt. *Economic Journal*, 127, 604 (2017), 1833–1873. (cited on page 8)

Brinch, C. N., 2012. Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics*, 27 (2012), 13–28. (cited on pages 10 and 16)

Cao, W.; Tsiatis, A. A.; and Davidian, M., 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 3 (2009), 723–734. (cited on pages 82 and 95)

Cappellari, L.; Dorsett, R.; and Haile, G., 2010. State dependence and unobserved heterogeneity in the employment transitions of the over-50s. *Empirical Economics*, 38, 3 (2010), 523–554. (cited on page 9)

Card, D.; Dobkin, C.; and Maestas, N., 2008. The impact of nearly universal insurance coverage on health care utilization: Evidence from Medicare. *American Economic Revew*, 98, 5 (2008), 2242–58. (cited on page 51)

Chan, W. C.; Wright, C.; Riddell, T.; et al., 2008a. Ethnic and socioeconomic disparities in the prevalence of cardiovascular disease in New Zealand. *New Zealand Medical Journal*, 121, 1285 (2008). (cited on page 51)

Chan, W. C.; Wright, C.; Tobias, M.; et al., 2008b. Explaining trends in coronary heart disease hospitalisations in New Zealand: Admissions and incidence can trend in opposite directions. *Heart*, 94, 12 (2008), 1589–1593. (cited on page

51)

Chang, W.-C.; Kaul, P.; Fu, Y.; et al., 2006. Forecasting mortality: Dynamic assessment of risk in ST-segment elevation acute myocardial infarction. *European Heart Journal*, 27, 4 (2006), 419–426. (cited on page 52)

Cockx, B. and Picchio, M., 2012. Are short-lived jobs stepping stones to long-lasting jobs? *Oxford Bulletin of Economics and Statistics*, 74, 5 (2012), 646–675. (cited on page 8)

Cockx, B. and Picchio, M., 2013. Scarring effects of remaining unemployed for long-term unemployed school-leavers. *Journal of the Royal Statistical Society Series A*, 176, 4 (2013), 951–980. (cited on page 8)

Cormack, D. and Robson, C., 2010. Classification and output of multiple ethnicities: considerations for monitoring māori health. Report, Te Rōpū Rangahau Hauora a Eru Pōmare. (cited on page 54)

Currie, J.; MacLeod, W. B.; and Van Parys, J., 2016. Provider practice style and patient health outcomes: The case of heart attacks. *Journal of Health Economics*, 47 (2016), 64–80. (cited on page 51)

Doiron, D. and Gørgens, T., 2008. State dependence in youth labor market experiences, and the evaluation of policy interventions. *Journal of Econometrics*, 145 (2008), 81–97. (cited on page 8)

Elbers, C. and Ridder, G., 1982. True and spurious duration dependence: The identifiability of the proportional hazards model. *Review of Economic Studies*, 49 (1982), 402–411. (cited on page 18)

Gørgens, T. and Hyslop, D., 2019. The specification of dynamic discrete-time two-state panel data models. *Econometrics*, 7, 1 (2019), 1–16. (cited on page 9)

Gottlieb, S.; Harpaz, D.; Shotan, A.; et al., 2000. Sex differences in management and outcome after acute myocardial infarction in the 1990s: A prospective observational community-based study. *Circulation*, 102, 20 (2000), 2484–2490. (cited on page 51)

GOURIÉROUX, C. AND MONFORT, A., 1991. Simulated based inference in models with heterogeneity. *Annales d'Economie et de Statistique*, 20/21 (1991), 69–107. (cited on pages 10 and 16)

GRITZ, R. M., 1993. The impact of training on the frequency and duration of employment. *Journal of Econometrics*, 57, 1–3 (1993), 21–51. (cited on page 8)

HAM, J. C. AND LALONDE, R. J., 1996. The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica*, 64, 1 (1996), 175–205. (cited on page 9)

HANSEN, B. B., 2008. The prognostic analogue of the propensity score. *Biometrika*, 95, 2 (2008), 481–488. (cited on pages 82, 83, and 89)

HECKMAN, J. J., 1981. The incidental parameter problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications* (Eds. C. F. MANSKI AND D. MCFADDEN), 179–195. MIT Press, Cambridge, Massachusetts. (cited on pages 8, 20, and 24)

HECKMAN, J. J.; ICHIMURA, H.; AND TODD, P. E., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64, 4 (1997), 605–654. (cited on page 96)

HECKMAN, J. J. AND SINGER, B., 1984a. Econometric duration analysis. *Journal of Econometrics*, 24 (1984), 63–132. (cited on page 18)

HECKMAN, J. J. AND SINGER, B., 1984b. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52 (1984), 271–320. (cited on page 18)

HERMAN, B.; GREISER, E.; AND POHIABELN, H., 1997. A sex difference in short-term survival after initial acute myocardial infarction: The MONICA-Bremen Acute Myocardial Infarction Register, 1985–1990. *European Heart Journal*, 18, 6 (1997), 963–970. (cited on page 51)

HESTERBERG, T., 1995. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37, 2 (1995), 185–194. (cited on page 18)

Hirano, K.; Imbens, G. W.; and Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 4 (2003), 1161–1189. (cited on page 89)

Hougaard, P., 1986. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73, 2 (1986), 387–396. (cited on page 52)

Hu, Z.; Follmann, D. A.; and Qin, J., 2012. Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *Journal of the American Statistical Association*, 107, 497 (2012). (cited on page 82)

Hu, Z.; Follmann, D. A.; and Wang, N., 2014. Estimation of mean response via the effective balancing score. *Biometrika*, 101, 3 (2014), 613–624. (cited on page 82)

Imai, K. and Ratkovic, M., 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B-statistical Methodology*, 76, 1 (2014), 243–263. (cited on page 89)

Imbens, G. W., 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 3 (2000), 706–710. (cited on page 92)

Imbens, G. W. and Rubin, D. B., 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. (cited on pages 81, 87, and 88)

Kamionka, T., 1998. Simulated maximum likelihood estimation in transition models. *Econometrics Journal*, 1, 1 (1998), C129–C153. (cited on page 9)

Kang, J. D. Y. and Schafer, J. L., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 4 (2007), 523–539. (cited on pages 82, 89, and 98)

Kauf, T. L.; Velazquez, E. J.; Crosslin, D. R.; et al., 2006. The cost of acute myocardial infarction in the new millennium: Evidence from a multinational registry. *American Heart Journal*, 151, 1 (2006), 206–212. (cited on page 47)

Keane, M. P., 1994. A computationally practical simulation estimator for panel data. *Econometrica*, 62, 1 (1994), 95–116. (cited on page 9)

Keane, M. P. and Sauer, R. M., 2010. A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables. *International Economic Review*, 51, 4 (2010), 925–958. (cited on pages 9, 18, and 23)

Klein, J. P.; Moeschberger, M.; Li, Y. H.; et al., 1992. Estimating random effects in the Framingham Heart Study. In *Survival Analysis: State of the Art*, 99–120. Springer. (cited on page 52)

Kreif, N.; Gruber, S.; Radice, R.; Grieve, R.; and Sekhon, J. S., 2016. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research*, 25, 5 (2016), 2315–2336. (cited on page 89)

Kytö, V.; Sipilä, J.; and Rautava, P., 2015. Gender and in-hospital mortality of ST-segment elevation myocardial infarction (from a multihospital nationwide registry study of 31,689 patients). *American Journal of Cardiology*, 115, 3 (2015), 303–306. (cited on page 51)

Lancaster, T., 1990. *The Econometric Analysis of Transition Data.* Econometric Society Monographs No. 17. Cambridge University Press, Cambridge; New York. (cited on page 12)

Lee, K. L.; Woodlief, L. H.; Topol, E. J.; et al., 1995. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: Results from an international trial of 41 021 patients. *Circulation*, 91, 6 (1995), 1659–1668. (cited on page 54)

Lee, M.-j., 2005. *Micro-econometrics for policy, program and treatment effects.* Oxford University Press. (cited on pages 81 and 87)

Lee, M.-j., 2009. Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society Series B-statistical*

Methodology, 71 (2009), 243–264. (cited on pages 82, 89, and 95)

Lee, M.-j., 2012. Treatment effects in sample selection models and their nonparametric estimation. *Journal of Econometrics*, 167, 2 (2012), 317–329. (cited on pages 82, 89, and 95)

Lee, M.-j., 2016. *Matching, regression discontinuity, difference in differences, and beyond*. Oxford University Press. (cited on page 87)

Lee, M.-j., 2018. Simple least squares estimator for treatment effects using propensity score residuals. *Biometrika*, 105, 1 (2018), 149–164. (cited on pages 89, 97, and 108)

Lee, M.-j. and Lee, S., 2019. Double robustness without weighting. *Statistics & Probability Letters*, 146 (2019), 175–180. (cited on pages 93 and 95)

Lee, S. H. and Gørgens, T., 2017. Estimation of dynamic models of recurring events with censored data. ANU Working Papers in Economics and Econometrics #655, Australian National University. Revised 2019. (cited on pages 48, 58, 60, 61, 62, and 65)

Lee, S. H. and Gørgens, T., 2019. Heart attack risk in new zealand: gender, ethnicity, age, and previous heart attacks. Unpublished manuscript, Australian National University. (cited on page 27)

Lerman, S. R. and Manski, C. F., 1981. On the use of simulated frequencies to approximate choice probabilities. In *Structural Analysis of Discrete Data with Econometric Applications* (Eds. C. F. Manski and D. McFadden), 303–319. MIT Press, Cambridge, Massachusetts. (cited on page 9)

Linden, A., 2017. Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting. *Journal of Evaluation in Clinical Practice*, 23, 4 (2017), 697–702. (cited on page 89)

Linden, A.; Uysal, S. D.; Ryan, A.; and Adams, J. L., 2016. Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine*, 35, 4 (2016), 534–552. (cited on page 89)

MacIntyre, K.; Stewart, S.; Capewell, S.; et al., 2001. Gender and survival: A population-based study of 201,114 men and women following a first acute myocardial infarction. *Journal of the American College of Cardiology*, 38, 3 (2001), 729–735. (cited on page 51)

Mak, K.-H.; Chia, K.-S.; Kark, J. D.; et al., 2003. Ethnic differences in acute myocardial infarction in Singapore. *European Heart Journal*, 24, 2 (2003), 151–160. (cited on page 52)

Manhapra, A.; Canto, J. G.; Vaccarino, V.; et al., 2004. Relation of age and race with hospital death after acute myocardial infarction. *American Heart Journal*, 148, 1 (2004), 92–98. (cited on page 52)

McCulloch, C. E., 1997. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 437 (1997), 162–170. (cited on page 9)

Ministry of Health, 2012. Health expenditure trends in New Zealand 1997–2007. Wellington. (cited on page 47)

Moffitt, R. A. and Rendall, M. S., 1995. Cohort trends in the lifetime distribution of female family headship in the United States, 1968–1985. *Demography*, 32, 3 (1995), 407–424. (cited on pages 9 and 13)

Morrow, D. A., 2017. *Myocardial Infarction: A Companion to Braunwald's Heart Disease E-Book*. Elsevier Health Sciences. (cited on page 47)

Naghavi, M.; Wang, H.; et al., 2015. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 385, 9963 (2015), 117–171. (cited on page 47)

National Health Board Business Unit, 2011. National Minimum Dataset (hospital events) data dictionary. Wellington: Ministry of Health. (cited on page 53)

National Health Committee, 2013. Strategic overview: Cardiovascular disease in New Zealand (working draft). (cited on page 47)

Nayan, M.; Hamilton, R. J.; Juurlink, D. N.; Finelli, A.; Kulkarni, G. S.;

and Austin, P. C., 2017. Critical appraisal of the application of propensity score methods in the urology literature. *BJU international*, 120, 6 (2017), 873–880. (cited on page 88)

Nguyen, H. L.; Ha, D. A.; Phan, D. T.; et al., 2014. Sex differences in clinical characteristics, hospital management practices, and in-hospital outcomes in patients hospitalized in a Vietnamese hospital with a first acute myocardial infarction. *PloS one*, 9, 4 (2014), e95631. (cited on page 51)

Pearl, J., 2009. *Causality*. Cambridge University Press. (cited on page 87)

Pokorney, S. D.; Rodriguez, J. F.; Ortiz, J. T.; et al., 2012. Infarct healing is a dynamic process following acute myocardial infarction. *Journal of Cardiovascular Magnetic Resonance*, 14, 1 (2012), 62. (cited on page 52)

Robins, J.; Sued, M.; Lei-gomez, Q.; and Rotnitzky, A., 2007. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22, 4 (2007), 544–559. (cited on pages 82 and 95)

Robins, J. M.; Rotnitzky, A.; and Zhao, L. P., 1994. Estimation of regression-coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 427 (1994), 846–866. (cited on pages 82 and 95)

Rosenbaum, P. R., 2002. *Observational studies*. Springer. (cited on pages 81 and 87)

Rosenbaum, P. R. and Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 1 (1983), 41–55. (cited on pages 81, 82, 83, 84, and 88)

Rotnitzky, A.; Lei, Q.; Sued, M.; and Robins, J. M., 2012. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 2 (2012), 439–456. (cited on pages 82 and 95)

Rubin, D. B., 1976. Inference and missing data. *Biometrika*, 63, 3 (1976), 581–592. (cited on page 1)

Rubin, D. B. and Thomas, N., 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical*

Association, 95, 450 (2000), 573–585. (cited on page 92)

Scharfstein, A. O.; Rotnitzky, A.; and Robins, J. M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 448 (1999), 1096–1120. (cited on pages 82 and 95)

Smolina, K.; Wright, F. L.; Rayner, M.; et al., 2012. Incidence and 30-day case fatality for acute myocardial infarction in England in 2010: National-linked database study. *European Journal of Public Health*, 22, 6 (2012), 848–853. (cited on pages 51 and 54)

Steg, P. G.; James, S. K.; Atar, D.; et al., 2012. ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *European Heart Journal*, 33, 20 (OCT 2012), 2569–2619. (cited on page 54)

Stern, S., 1997. Simulation-based estimation. *Journal of Economic Literature*, 35, 4 (1997), 2006–2039. (cited on page 16)

Stuart, E. A., 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1 (2010), 1–21. (cited on pages 81 and 88)

Stute, W., 1997. Nonparametric model checks for regression. *Annals of Statistics*, 25, 2 (1997), 613–641. (cited on page 104)

Stute, W.; Manteiga, W. G.; and Quindimil, M. P., 1998. Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, 93, 441 (1998), 141–149. (cited on page 105)

Tan, Z., 2010. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97, 3 (2010), 661–682. (cited on pages 82 and 95)

Tunstall-Pedoe, H.; Kuulasmaa, K.; Amouyel, P.; et al., 1994. Myocardial infarction and coronary deaths in the World Health Organization MONICA project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. *Circulation*, 90, 1 (1994), 583–612. (cited on page 51)

Vaccarino, V.; Rathore, S. S.; Wenger, N. K.; et al., 2005. Sex and racial differences in the management of acute myocardial infarction, 1994 through 2002. *New England Journal of Medicine*, 353, 7 (2005), 671–682. (cited on page 51)

Vermeulen, K. and Vansteelandt, S., 2015. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110, 511 (2015), 1024–1036. (cited on pages 82 and 95)

Waernbaum, I., 2012. Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Statistics in Medicine*, 31, 15 (2012), 1572–1581. (cited on page 89)

Wang, O. J.; Wang, Y.; Chen, J.; et al., 2012. Recent trends in hospitalization for acute myocardial infarction. *American Journal of Cardiology*, 109, 11 (2012), 1589–1593. (cited on pages 51 and 52)

Wienke, A., 2010. *Frailty Models in Survival Analysis.* Chapman and Hall/CRC. (cited on page 57)

Wu, S.; Ding, Y.; Wu, F.; Hou, J.; and Mao, P., 2015. Application of propensity-score matching in four leading medical journals. *Epidemiology*, 26, 2 (2015), e19–e20. (cited on page 88)

Zhang, Q. and Wang, Y., 2004. Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Social Science & Medicine*, 58, 6 (2004), 1171–1180. (cited on page 51)