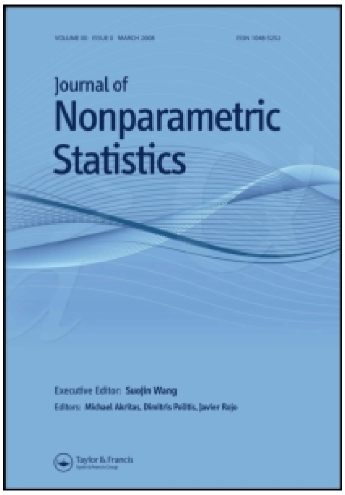

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

Jackknife-after-bootstrap regression influence diagnostics

Michael A. Martin [a]; Steven Roberts [a]
[a] School of Finance and Applied Statistics, Australian National University, Canberra, Australia

First published on: 02 November 2009

To cite this Article Martin, Michael A. and Roberts, Steven(2009) 'Jackknife-after-bootstrap regression influence diagnostics', Journal of Nonparametric Statistics,, First published on: 02 November 2009 (iFirst)

To link to this Article: DOI: 10.1080/10485250903287906

URL: http://dx.doi.org/10.1080/10485250903287906

Journal of Nonparametric Statistics
Vol. 00, No. 0, Month 2009, 1–13

Jackknife-after-bootstrap regression influence diagnostics

Michael A. Martin* and Steven Roberts

School of Finance and Applied Statistics, Australian National University, Canberra, ACT 0200, Australia

(*Received 9 August 2007; final version received 26 August 2009*)

We propose a bootstrap approach to gauging the size of regression influence measures. The bootstrap cut-offs generated are based on approximating the sampling distribution of the respective measures under resampling, work well for small samples, and allow for features such as asymmetric cut-offs. The bootstrap method uses Efron's jackknife-after-bootstrap idea to deal with the issue of an influential point contaminating the resamples from which cut-offs are calculated. The method is illustrated through both real-world examples and a simulation study, the results of which suggest that the bootstrap method provides a reliable alternative to traditional methods particularly in small to moderate samples.

Keywords: Cook's distance; COVRATIO; DFBETAS; DFFITS; leverage; bootstrap; regression; diagnostics

1. Introduction

The identification and assessment of influential points in regression modelling is one of the most important and common data analysis tasks. Research on this problem was furious in the 1970s and 1980s when the advent of fast, cheap computing made routine examination of the data not only possible, but also a *de rigueur* part of the modelling process. Prior to that time, statisticians were still keenly aware of the effect that individual data points could have on an analysis, but the existing computing conditions made the assessment of such points a laborious activity, most examinations carried out 'by hand', with experience and finely honed judgement key elements of successful analyses. The rise of computers in the 1970s presented unprecedented opportunities for data exploration, and early papers by Cook (1977, 1979), the seminal monograph of Belsley, Kuh, and Welsch (1980), and the book by Cook and Weisberg (1982) laid the foundation for the modern approach to diagnosing regression models. The methods proposed reflected the computing culture of the time, with statistics termed DFFITS, DFBETAS, and COVRATIO entering the statistical lexicon, but a quarter century on, these ideas have become ubiquitous, such measures calculated automatically by virtually all commercial statistical packages, and regression diagnostics forming a crucial part of almost every regression text. Chatterjee and Hadi (1986) provide an excellent overview of research into regression diagnostics, with their paper promoting a lively discussion

*Corresponding author. Email: Michael.Martin@anu.edu.au

ISSN 1048-5252 print/ISSN 1029-0311 online
© 2009 Taylor & Francis
DOI: 10.1080/10485250903287906
http://www.informaworld.com

among leading researchers in this area – including some entertaining insights into the nomenclature that characterises the topic. Other useful references include Davison and Tsai (1992), Weisberg (1983), Fox (1991), Pena and Yohai (1995), and Brown and Lawrance (2000). More recently, there has been renewed interest in regression influence diagnostics, with Fung, Zhu, Wei, and He (2002) considering influence diagnostics for semiparametric mixed models, and Pena (2005) developing a new influence measure as an alternative to the common measures discussed here. Seminal among the work on regression diagnostics is Belsley et al. (1980). This book, along with Cook's (1977) work on Cook's distance, set the basis for the measures commonly used to assess influence in regression models. The measures proposed, DFFITS, DFBETAS, COVRATIO, and Cook's D, are all based on the simple idea of measuring the effect of a data point by considering models both including and omitting the point, and computing a (scaled) difference or ratio. This idea was itself not new, but its use in a regression context was path breaking, and the measures produced were evocative, if oddly named. But, how the size of these measures should be assessed with respect to the appraisal of influence was a key question. The advice provided by Cook (1977, 1979) and Belsley et al. (1980) was simple but somewhat *ad hoc*. The cut-offs suggested had a reasonable basis in theory, but the actual distributions of the quantities under study were complex, depending, for instance on sample size, and the size of the model under consideration, and so, appropriately, the judgement on influence remained reasonably subjective supported by simple, objective rules. They argue, for example, that DFFITS and DFBETAS are 't-like', but that the declining effect of individual points as sample sizes grow means that familiar 'normal' cut-offs such as ±2 should be modified by factors depending on the number of regression parameters, p, and sample size, n. Hence, they proposed 'size-adjusted' cut-offs for DFFITS as $\pm 2\sqrt{p/n}$ and for DFBETAS as $\pm 2/\sqrt{n}$. These cut-offs use so-called 'external scaling', based on notions of the nature of the sampling distributions of the relevant measures, and on their dependence on 'size' variables p and n. For Cook's D, the advice given is less prescriptive, some authors advocating using 1 as a cut-off, others the median of an $F_{p,n-p}$ distribution, while others still merely look at the relative sizes among the values of Cook's D for all data points and subjectively designate some as 'large'. The latter approach involves 'internal scaling', the use of all available values of the measures to judge size relatively. The use of internal scaling is also suggested by Belsley et al. (1980) for DFFITS and DFBETAS. These rules have some of the flavour of hypothesis tests – for example, Cook's D is often compared with a quantile of an F distribution – but the comparison falters, as the rules do not explicitly attach significance levels. Internal scaling ideas, such as identifying points whose influence measures look unusual among those of other data points, are also in popular use, but these methods seem more subjective and suffer in small samples because of the small number of comparators available.

While the traditional usage of regression influence diagnostics is straightforward, the cut-offs suggested remain somewhat *ad hoc*. Based on large sample theory and rough approximations, the cut-offs may not adequately allow for small sample sizes, or cases where model error distributions exhibit significant skewness or heavy tails. For example, the cut-offs for DFFITS and DFBETAS are symmetric, and while these quantities may have close to normal distributions asymptotically, for small samples there seems less justification to use symmetric cut-offs. In this article, we propose that a variation of Efron's (1979) bootstrap, itself born when computing power began to significantly impact statistical practice, be used to approximate the sampling distributions of these influence measures. Appropriate bootstrap distributions of these quantities allow for easy specification of cut-offs, allowing naturally for asymmetry and small samples.

The use of a naïve bootstrap approach to generating sampling distributions of influence measures is problematic, because highly influential points may appear multiple times in resamples. For example, the approximate proportion of resamples in which any given data point will appear j times is $(j!e)^{-1}$, meaning that a particular point fails to appear in about $(1 - n^{-1})^n \to \mathrm{e}^{-1} \simeq$ 36.79% of resamples, appears only once in about e^{-1} of resamples, but appears multiple times in the remaining $1 - 2/e \approx 26.4\%$ of resamples. As a result, if a particular data point is highly influential, then for resamples in which that data point appears, perhaps multiple times, the values calculated for the corresponding diagnostic measures will not necessarily be representative of values one might anticipate arising from the sampling distributions of those quantities under 'null' circumstances. This situation would artificially lead to the appearance of more bootstrap replicates in the tails of the target distributions than is reasonable, inflating the estimated cut-offs, and effectively hiding the suspect point. An approach that protects against this situation occurring is to construct individual influence cut-offs for each data point, basing the cut-offs on sampling distributions estimated using resamples that do not include the point in question. While this strategy at first seems very costly, requiring repeated resampling from a sequence of reduced data sets, we propose a bootstrap method based on Efron's (1992) jackknife-after-bootstrap technique. The method is relatively fast and easy to implement, allowing practitioners to quickly isolate suspicious data points for further investigation. Our findings, supported by both analysis of real data and a simulation study, are very encouraging. The bootstrap method successfully identifies influential observations, generally with cut-offs larger and more asymmetric than traditional proposals. For large data sets with normal errors, the bootstrap cut-offs are fairly close to the traditional cut-offs, but for small data sets and non-normal errors, the bootstrap cut-offs provide an automatic adjustment to the traditional cut-offs to account for features of the data.

For Cook's D, we find that the bootstrap method provides more stringent cut-offs than the traditional advice, and in our examples and simulations, the bootstrap method flags reasonably obvious points that the traditional method misses. The traditional cut-off for Cook's D appears to be overly liberal, and textbook examples routinely flag points *inside* the cut-off as warranting further investigation – for example, Ramsey and Schafer (2002) and Neter, Wasserman, and Kutner (1990). Our method avoids such subjective judgements, flagging points as worthy of attention if they fall too far into the tail of the relevant bootstrap distribution. Our approach yields cut-offs for Cook's D that align its advice more closely with that provided by other influence measures for coefficients, and provides cut-offs that seem more reasonable than those found in many standard regression texts. There has been significant recent interest in Cook's D – see, for example, Kim (1996), Muller and Mok (1997), and Pena (2005).

The bootstrap method for assessing influence in regression models has several key advantages over traditional methods, as well as one obvious drawback. First, traditional cut-offs are based on large sample theory and arguments related to sample and model size, while the new methods approximate the actual sampling distribution of the relevant measures regardless of sample size. Second, the new methods allow naturally for asymmetry in the sampling distributions of measures such as DFFITS and DFBETAS, whereas the traditional method assumes the distributions are symmetric – while this assumption seems adequate for large samples, we do not believe it credible in small or moderate samples, or when the underlying error distribution is non-normal, and our numerical results support this view. Third, the traditional cut-offs are invariant to the model selected, whereas the bootstrap methods incorporate model information into the values of the influence measures computed for each resample in establishing the cut-offs. While it is important that the detection of influential points should always be in the context of a sensibly chosen model, often several competing models are close, and incorporating the model information into decisions about influence seems desirable. The principal drawback of our method is that it is slower than the traditional approach. In our computations – carried out using R on a Intel Xeon 3.0 GHz PC – for a data set with 50 data points and four covariates, the traditional method ran in seconds, while our method took about 2 min, still easily fast enough for routine, interactive exploration of the data.

The problem of outlier detection is somewhat more straightforward than that of detecting influential observations since the distributions of diagnostic measures for detecting outliers, such as studentised deleted residuals, are reasonably well understood compared with the distributions of influence measures such as DFFITS and Cook's distance. In recent work, Martin and Roberts (2006) compared several bootstrap-based approaches with other common approaches for outlier detection in least squares regression problems, including a jackknife-after-bootstrap technique as well as residual-based resampling. Other important issues in the detection of influential observations such as masking and swamping are not dealt directly in this article, although they are clearly important concerns. Rather, our focus here is deliberately simple, proposing, and assessing a straightforward but reliable alternative to methods in common usage.

2. Methodology

Our study of regression diagnostics is in the context of multiple regression models $E(Y_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ji}$, $i = 1, \ldots, n$, relating a response, Y, to predictors $X_1, \ldots, X_{p-1}$ for a set of n data points. Belsley et al. (1980) describe a number of diagnostic measures designed to isolate data points that potentially exert influence on the aspects of the model fit. Each of the influence measures compares an aspect of model fit between the scenarios of including and excluding a particular data point. The aspects of model fit considered include effects on fitted values (DFFITS), model coefficients (DFBETAS and an omnibus measure, Cook's D), scale estimates (COVRATIO), and leverage, a measure depending only on $\mathbf{X}$ that appears associated with influence in many cases. For the ith case, the formulas for the aspects of model fit considered are

Influence measure	Formula	Traditional cut-off
Leverage$_i$ (h_{ii})	$\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}_i$	$2\frac{p}{n}$
DFFITS$_i$	$\frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{\mathrm{MSE}_{(i)} h_{ii}}}$	$\pm 2\sqrt{\frac{p}{n}}$
DFBETAS$_{ki}$	$\frac{b_k - b_{k(i)}}{\sqrt{\mathrm{MSE}_{(i)} c_{kk}}}$	$\pm \frac{2}{\sqrt{n}}$
Cook'sD_i	$\frac{(\boldsymbol{b} - \boldsymbol{b}_{(i)})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{b}_{(i)})}{p\mathrm{MSE}}$	1, the median of an $F_{p,n-p}$, or internal scaling
COVRATIO	$\left(\frac{\mathrm{MSE}_{(i)}}{\mathrm{MSE}}\right)\left(\frac{1}{1-h_{ii}}\right)$	$\lvert\mathrm{COVRATIO} - 1\rvert \geq 3\frac{p}{n}$

In these formulas, $\hat{Y}_i$, MSE, b_k, and b are, respectively, the ith fitted value, the mean squared error, estimate of β_k, and vector of estimated coefficients from the regression model fit to all cases. The subscript notation (i) denotes values calculated from the regression model fit with the ith case removed. $\mathbf{X}_i$ is the ith row of the design matrix $\boldsymbol{X}$.

The forms taken by these measures are evocative. For example, DFFITS and DFBETAS resemble t-statistics in their construction, but neither quantity possesses a t-distribution, and the form of Cook's D resembles that of an F-statistic, but the quantity does not have an F distribution. Moreover, distribution theory for these constructs is complex, so although these measures are appealing from a philosophical point of view, their use in practical circumstances has relied on rules adjusting the large-sample cut-offs for sample and model size. Note that the cut-offs for DFFITS, DFBETAS, and COVRATIO are symmetric, and follow a familiar 'plus or minus a number of standard errors' pattern. The traditional cut-offs work quite well, therefore, when sample sizes are large, in particular, when n is much larger than p, and when the model errors are normal. Of course, in many real-world settings, neither of these conditions is satisfied, and so the influence cut-offs are not as effective in influence assessment as in 'ideal' cases. Bootstrap methods provide a way to estimate the sampling distributions of the various regression diagnostics and, therefore, to establish relevant cut-offs beyond which a point might be investigated as influential. Cut-offs found this way can differ markedly from the traditional cut-offs, first in providing reasonable small-sample thresholds, and second in allowing for asymmetry in the sampling distributions of the influence measures.

The bootstrap measures reflect a balance between internal and external scaling measures. By using the data itself to generate the relevant sampling distributions, the bootstrap approach allows features of the data – such as skewness in the error distribution – to modify size judgements for the measures in question, consistent with internal scaling in this regard. The bootstrap also generates an estimate of the entire sampling distribution of an influence measure, so that judgements on whether the influence measure related to a specific data point is 'unusual' can be based on a comparison of that measure with the extreme quantiles of the generated bootstrap distribution. This approach mirrors that of comparing a specific measure against, say, a cut-off generated from the quantiles of some underlying theoretical, asymptotic distribution.

Bootstrapping for regression models can be carried out in two basic ways: residual-based resampling, where resamples are constructed by first fitting an appropriate model, obtaining residuals and using resampled residuals to build resampled data; and case-based resampling, where the data $(\boldsymbol{X}, Y)$ are resampled as p-tuples so that $\mathbf{X}$ and Y cases remain together in resamples. The relative merits of the two schemes have been discussed regularly in the bootstrap literature – see, for example, Efron and Tibshirani (1993) and Davison and Hinkley (1997) – but a simple distinction is that residual-based methods are appropriate when the model is credible and when the covariates $\mathbf{X}$ are non-random, while case-based methods are more favoured in observational studies (when $\mathbf{X}$ is random) or when the form of the model is in question. In the present context, however, the residual-based approach offers no obvious way to 'delete' a data point from consideration, meaning that in a case-deletion context, there would never be bootstrap samples with missing cases, resulting in no meaningful case-deletion diagnostics associated with a particular point. Also, the formation of residuals from which resamples might be drawn involves the use of models potentially unduly influenced by individual data points. For example, a single highly influential data point may affect the model so as to have a zero residual, effectively concealing the effect of that data point in resamples. As a result, our method relates to only case-based resampling.

The bootstrap procedure we propose uses Efron's (1992) jackknife-after-bootstrap sampling lemma, described below, to approximate the sampling distributions of the influence measures of interest. This lemma allows the sampling distributions for a given case to be estimated using resamples that do not include the case in question as part of a single resampling operation.

Jackknife-After-Bootstrap Sampling Lemma (Efron 1992) *A bootstrap sample drawn with replacement from* $z_1, z_2, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ *has the same distribution as a bootstrap sample drawn from* $z_1, z_2, \dots, z_n$ *in which none of the bootstrap values equals* z_i.

In practice, this method costs about e times as much as a 'regular' bootstrap. If 1000 resamples are desired within which a particular data point does not appear, then about $1000e \simeq 3000$ total resamples are required to generate the required set of 'reduced' resamples. For each individual data point, roughly 1000 resamples among 3000 originally generated resamples that do not contain that point can be used to construct the sampling distribution of the respective influence measures to generate influence cut-offs. This method produces separate influence cut-offs for each data point, but only a single set of about $3B$ original resamples is needed to obtain all n sets of B 'case-deleted' resamples for each data point.

Our procedure is based on the idea that within each jackknife-after-bootstrap subgroup of resamples (groups indexed via the missing case), delete-1 diagnostics for each resample can be calculated and the bootstrap distribution of the relevant diagnostic approximated using resamples not contaminated by the point under consideration. We next describe our algorithm in the context of finding appropriate bootstrap cut-offs for a specific diagnostic, say Cook's distance D:

Step 1. Fit the proposed model to the entire original data set, and compute D_i, $i = 1, \ldots, n$.
Step 2. Construct B resamples using uniform resampling from the original cases.
Step 3. (Jackknife-after-bootstrap step). For each data point, $i = 1, \ldots, n$, consider the group of roughly B/e resamples that *do not* contain that data point, and for each resample within the group, calculate the n values of Cook's distance, D_j^*, where the index j runs across all cases within a resample. Collect all *nB/e* values of Cook's distance into a single vector. The rationale behind this approach is to generate a 'null' bootstrap distribution of Cook's distance under the hypothesis that the ith data point is not influential. Since the ith data point is not present in any of the resamples from which this bootstrap distribution is generated, it cannot exert influence, and thus the distribution generated is free from the influence of this point.
Step 4. A suitable quantile (say 97.5%) of the bootstrap distribution generated by the *nB/e* values of Cook's distance calculated under resampling can be used as a cut-off for determining the potential presence of point's influential on the model parameter estimates.

The rationale behind the use of a jackknife-after-bootstrap approach to this problem arises from considering how one might approximate the sampling distribution of an influence measure assuming the data arose from a specific multivariate parent distribution. In that case, independent of the sample data, samples could be generated from the underlying parent distribution, and 'generic' values of the influence measure calculated for each sample point. These values of the influence measure could then be collated over all samples to form a sampling distribution of the statistic, independent of the original data. Of course, in our case, we operate under no specific distributional assumptions, so the jackknife-after-bootstrap resamples offer sets of resamples that allow the generation of bootstrap distributions of the influence measure independent of specific individual data points under assessment, instead based on the remaining data points.

Remark 1 Issues such as masking and swamping – see, for example, Atkinson (1986) and Lawrance (1995) – remain serious challenges for single-case deletion methods such as those discussed here. Our approach suggests a potential way forward through the use of resamples that do not contain larger subsets of data points to establish reasonable influence 'norms'. Computationally, such a strategy is more demanding than the present technique, but relentlessly increasing computing power and the promise of parallel computing systems offer hope for feasible implementation in future. An alternative to the use of bootstrapping in this context is the use of multiple-case deletion jackknife subsamples of the original data to reveal subtle masking effects. This technique has recently been proposed and investigated by Martin, Roberts, and Zheng (in press).

Remark 2 Our focus in this article is only on linear regression models, but the methodology is clearly extensible to other types of models, such as generalised linear models, for which influence measures can be obtained. Our focus here is deliberately simple so as to set the ideas out most clearly, and to explore the performance of the methods in this important common case.

3. Numerical results

The performance of the bootstrap influence diagnostics and a comparison with traditional techniques were studied in a real-data example and a simulation study involving data from several parent distributions and sample size/number of covariates combinations.

Example (the life cycle savings data) Belsley et al. (1980) described data for 50 countries relating the savings ratio (aggregate personal savings divided by disposable income) averaged over the decade 1960–1970 to four covariates: percent of population under 15, percent of population over 75, real per capita disposable income, and percent growth rate in per capita disposable income. Under Modigliani's life-cycle savings hypothesis, a linear regression model relates the response to the four covariates. For this data set, $n = 50$ and $p = 5$, and we wish to identify which of the 50 countries may have undue influence on the model fit. Table 1 summarises the use of the bootstrap and traditional methods for assessing influence. For this example, 3100 resamples were drawn from the original data, thereby creating 50 sets of jackknife-after-bootstrap resamples, each set excluding a particular data point, and each set containing a minimum of 1000 such resamples. For each of the 50 data points, cut-offs for the respective regression diagnostics were generated as 2.5 and 97.5% percentage points of the bootstrap distributions of the diagnostics based on the appropriate set of jackknife-after-bootstrap resamples.

A number of key features of our methods are evident. First, the asymmetry of the sampling distributions of DFBETAS, DFFITS, and COVRATIO are clear from the cut-offs suggested by the bootstrap method. In each case, one cut-off is reasonably close to the corresponding traditional cut-off, but the other cut-off is significantly larger, suggesting that the traditional cut-offs may be needlessly flagging points in one tail of the distribution. For example, the traditional method flags Zambia (point 46) as influential with respect to β_2, but the bootstrap method does not, as the upper cut-off is substantially larger in magnitude than the lower cut-off. Strikingly, the bootstrap method applies much more stringent cut-offs than the traditional methods for the Cook's D, with Japan (23), Zambia (46), and Libya (49) flagged by the bootstrap method as influential with respect to the beta coefficients, whereas the traditional method flags *no* points – our method appears more reasonable here, as these countries were variously noted with respect to individual betas by the traditional cut-offs for DFBETAS. Moreover, an initial visual investigation of the data suggested that these points warranted further inspection.

3.1. *Simulation study results*

A simulation study was conducted to investigate the relative behaviour and performance of the traditional and bootstrap cut-offs under various modelling scenarios. We considered data simulated from a regression model with sample size n and with p-1 covariates for the cases $(n, p) = (20, 2)$ and (50,5), and for three error distributions: normal (N(0,0.5625)), t(3) (heavy-tailed), and centred log-normal (1.5[exp$\{N(0, 0.5625)\}$ − exp(1/2)]; skewed). For brevity, the normal error cases and the case of t(3) errors with $n = 50$ are presented only in summary, in Table 2, while full results are presented for the other cases in Tables 3–5. The scenarios are designed to reflect the behaviour of the bootstrap method under realistic data settings, where errors often appear heavy-tailed or skewed. We considered two basic scenarios, the first where no clear influential points were deliberately generated, and a second scenario where a potential influential data point was inserted into the data set. For the case $p = 2$, the base model was $Y = 1 + 2X + \varepsilon$, with $\mathbf{X}$ generated as i.i.d. N(2,1) variates, and ε having one of the three error distributions listed above. Where relevant, the deliberately inserted influential point was at $x = 5$, $y = 2$. For $p = 5$, the base model was $Y = 1 + 2X_1 + 4X_2 + 3X_3 + 2X_4 + \varepsilon$, with each $\mathbf{X}$ generated as i.i.d. N(2,1) variates, and ε as above. Where relevant, the deliberately inserted influential point kept x_1, x_3, x_4 at their original simulated values but had $x_2 = 10$ and $y = 10$. In all tables, standard errors of the simulated quantities are presented in brackets.

For the simulations based on normal data (not presented) for DFBETAS and DFFITS, when no clear influential point was inserted into the data set, a consistent pattern emerged where the cut-offs for the bootstrap method are roughly symmetric but larger in magnitude than the traditional

Table 1. Regression influence diagnostics for life cycle savings data, $n = 50,\ p = 5$.

Method		DFBETAS ($\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$)					Cook's D	DFFITS	COVRATIO	Leverage
Traditional	Low cut-off	−0.283	−0.283	−0.283	−0.283	−0.283		−0.632	0.7	
	High cut-off	0.283	0.283	0.283	0.283	0.283	0.883	0.632	1.3	0.2
	Points below	21	23, 49	23, 46, 49	None	33, 47, 49		49	7, 46	
	Points above	23, 49	10, 21	21	None	23	None	23, 46	6, 37, 44, 49	21, 23, 44, 49
Bootstrap	Low cut-off	−0.272 (0.006)	−0.361 (0.013)	−0.371 (0.012)	−0.247 (0.010)	−0.339 (0.008)	−0.588 (0.009)	−0.650 (0.009)		
	High cut-off	0.367 (0.012)	0.277 (0.006)	0.283 (0.007)	0.202 (0.005)	0.285 (0.009)	0.0769 (0.002)	0.686 (0.014)	1.521 (0.020)	0.230 (0.004)
	Points below	21	23, 49	23, 49	21, 39	49		49	7, 46	
	Points above	23, 49	10, 21	21	None	23, 32	23, 46, 49	23, 46	44, 49	44, 49

Table 2. Low and high average cut-offs for known influential case (point 1) and other cases – all simulations.

DFBETAS (β_0)		DFBETAS (β_1)		DFBETAS (β_2)		DFBETAS (β_3)		DFBETAS (β_4)		Cook's D		DFFITS		COVRATIO		Leverage	
Point 1 cut-off	Other cut-offs	Point 1 cut-off	Other cut-offs	Point 1 cut-off	Other cut-offs	Point 1 cut-off	Other cut-offs	Point 1 cut-off	Other cut-offs	Point 1 cut-off	Others cut-offs	Point 1 cut-off	Others cut-offs	Point 1 cut-off	Others cut-offs	Point 1 cut-off	Others cut-offs
Normal errors, $n = 20$, $p = 2$																	
−0.547	−0.571	−0.559	−1.430									−0.711	−1.599	0.680	0.576		
0.561	0.982	0.549	0.596							0.224	0.453	0.717	0.644	1.469	1.362	0.233	0.255
$t(3)$ Errors, $n = 20$, $p = 2$																	
−0.572	−0.586	−0.578	−1.092									−0.767	−1.290	0.587	0.582		
0.576	0.783	0.576	0.580							0.249	0.372	0.770	0.675	1.469	1.413	0.231	0.254
Log-normal errors, $n = 20$, $p = 2$																	
−0.491	−0.550	−0.576	−1.250									−0.553	−1.379	0.545	0.511		
0.663	0.928	0.574	0.589							0.252	0.421	0.969	0.774	1.474	1.389	0.232	0.255
Normal errors, $n = 50$, $p = 5$																	
−0.343	−0.334	−0.341	−0.334	−0.342	−0.424	−0.343	−0.334	−0.343	−0.333			−0.702	−0.899	0.689	0.683		
0.342	0.375	0.341	0.335	0.343	0.355	0.344	0.332	0.342	0.334	0.093	0.102	0.700	0.647	1.385	1.365	0.200	0.208
$t(3)$ Errors, $n = 50$, $p = 5$																	
−0.340	−0.335	−0.345	−0.339	−0.340	−0.434	−0.342	−0.336	−0.346	−0.338			−0.735	−0.932	0.577	0.626		
0.344	0.378	0.342	0.337	0.337	0.342	0.341	0.338	0.344	0.337	0.100	0.108	0.740	0.662	1.390	1.368	0.199	0.207
Log-normal errors, $n = 50$, $p = 5$																	
−0.314	−0.326	−0.334	−0.333	−0.336	−0.437	−0.340	−0.335	−0.333	−0.332			−0.535	−0.807	0.537	0.620		
0.363	0.393	0.335	0.332	0.334	0.345	0.336	0.332	0.339	0.334	0.099	0.107	0.927	0.726	1.392	1.368	0.200	0.208

Table 3. Simulation results, $t(3)$ errors, $n = 20,\ p = 2$.

Method		DFBETAS		Cook's D	DFFITS	COVRATIO	Leverage
Influential point not present							
Traditional	Low cut-off	−0.447	−0.447		−0.633	0.700	
	High cut-off	0.447	0.447	0.721	0.633	1.300	0.200
		1.711	1.670	0.178	1.743	3.170	1.707
		(0.029)	(0.029)	(0.013)	(0.032)	(0.036)	(0.025)
Bootstrap	Low cut-off	−0.586	−0.581		−0.778	0.587	
		(0.006)	(0.006)		(0.007)	(0.003)	
	High cut-off	0.577	0.583	0.251	0.768	1.466	0.232
		(0.006)	(0.006)	(0.002)	(0.007)	(0.003)	(0.001)
	Average no. of points (SE)	1.347	1.337	1.214	1.489	1.446	1.141
		(0.022)	(0.022)	(0.019)	(0.023)	(0.021)	(0.016)
Influential point present							
Traditional	Average no. of points (SE)	1.971	2.131	1.001	2.035	2.128	1.694
		(0.023)	(0.026)	(0.011)	(0.026)	(0.032)	(0.020)
	Percent of times point identified	0.985	0.993	0.934	0.989	0.640	0.998
		(0.004)	(0.003)	(0.008)	(0.003)	(0.015)	(0.001)
Bootstrap	Low cut-off	−0.585	−1.067		−1.264	0.582	
		(0.005)	(0.005)		(0.007)	(0.003)	
	High cut-off	0.773	0.579	0.366	0.680	1.416	0.253
		(0.005)	(0.005)	(0.003)	(0.005)	(0.002)	(0.001)
	Average number of points (SE)	1.607	1.694	1.353	1.691	1.185	1.233
		(0.016)	(0.017)	(0.015)	(0.018)	(0.022)	(0.014)
	Percent of times point identified	0.948	0.978	0.982	0.977	0.494	0.970
		(0.007)	(0.005)	(0.004)	(0.005)	(0.016)	(0.005)

Table 4. Simulation results, log-normal errors, $n = 20,\ p = 2$.

Method		DFBETAS		Cook's D	DFFITS	COVRATIO	Leverage
Influential point not present							
Traditional	Low cut-off	−0.447	−0.447		−0.632	0.700	
	High cut-off	0.447	0.447	0.721	0.632	1.300	0.200
	Average number of points (SE)	1.335	1.715	0.202	1.651	3.149	1.670
		(0.028)	(0.029)	(0.014)	(0.025)	(0.034)	(0.025)
Bootstrap	Low cut-off	−0.495	−0.585		−0.553	0.552	
		(0.005)	(0.006)		(0.004)	(0.004)	
	High cut-off	0.666	0.584	0.257	0.983	1.470	0.232
		(0.008)	(0.006)	(0.002)	(0.008)	(0.003)	(0.001)
	Average number of points (SE)	1.135	1.324	1.145	1.277	1.475	1.142
		(0.022)	(0.021)	(0.017)	(0.022)	(0.020)	(0.016)
Influential point present							
Traditional	Average number of points (SE)	1.997	2.293	1.064	2.072	2.059	1.699
		(0.024)	(0.027)	(0.008)	(0.025)	(0.030)	(0.021)
	Percent of times pointt identified	0.996	0.998	0.988	0.997	0.830	0.994
		1.997	2.293	1.064	2.072	2.059	1.699
Bootstrap	Low cut-off	−0.547	−1.216		−1.337	0.513	
		(0.004)	(0.004)		(0.006)	(0.003)	
	High cut-off	0.915	0.589	0.413	0.784	1.394	0.254
		(0.006)	(0.005)	(0.003)	(0.006)	(0.002)	(0.001)
	Average number of points (SE)	1.593	1.685	1.280	1.639	1.221	1.213
		(0.016)	(0.017)	(0.014)	(0.017)	(0.021)	(0.014)
	Percent of times point identified	0.963	0.991	0.995	0.998	0.683	0.967
		(0.006)	(0.003)	(0.002)	(0.001)	(0.015)	(0.006)

Table 5. Simulation results, log-normal errors, $n = 50$, $p = 5$.

Method		DFBETAS					Cook's D	DFFITS	COVRATIO	Leverage
Influential point not present										
Traditional	Low cut-off	−0.283	−0.283	−0.283	−0.283	−0.283		−0.633	0.700	
	High cut-off	0.283	0.283	0.283	0.283	0.283	0.884	0.633	1.300	0.200
	Average number of points (SE)	2.811	3.000	3.228	3.065	3.000	0.048	3.072	7.106	2.673
		(0.040)	(0.040)	(0.042)	(0.041)	(0.041)	(0.007)	(0.038)	(0.061)	(0.041)
Bootstrap	Low cut-off	−0.311	−0.337	−0.334	−0.334	−0.337		−0.530	0.531	
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	
	High cut-off	0.366	0.335	0.332	0.331	0.332	0.098	0.934	1.392	0.200
		(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.005)	(0.001)	(0.000)
	Average number of points (SE)	2.384	2.449	2.511	2.455	2.431	2.377	2.342	2.989	2.789
		(0.027)	(0.027)	(0.026)	(0.027)	(0.027)	(0.024)	(0.028)	(0.028)	(0.028)
Influential point present										
Traditional	Average number of points (SE)	3.373	2.666	3.223	4.052	3.283	1.001	2.958	4.528	2.614
		(0.041)	(0.029)	(0.036)	(0.045)	(0.041)	(0.001)	(0.036)	(0.045)	(0.033)
	% of times point identified	0.996	0.900	1.000	0.929	0.920	1.000	1.000	1.000	1.000
		(0.002)	(0.010)	(0.000)	(0.009)	(0.009)	(0.000)	(0.000)	(0.000)	(0.000)
Bootstrap	Low cut-off	−0.326	−0.333	−0.435	−0.335	−0.332		−0.802	0.619	
		(0.001)	(0.002)	(0.003)	(0.002)	(0.002)		(0.003)	(0.002)	
	High cut-off	0.393	0.332	0.345	0.332	0.334	0.107	0.730	1.368	0.207
		(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)	(0.001)	(0.000)
	Average number of points (SE)	2.614	2.330	2.649	2.931	2.681	2.083	2.129	2.461	2.401
		(0.029)	(0.022)	(0.027)	(0.032)	(0.032)	(0.024)	(0.025)	(0.028)	(0.023)
	% of times point identified	0.995	0.872	1.000	0.900	0.909	1.000	1.000	1.000	1.000
		(0.002)	(0.011)	(0.000)	(0.009)	(0.009)	(0.000)	(0.000)	(0.000)	(0.000)

cut-offs, the disparity in size falling as n increases, though still noticeable at $n = 50$. This pattern was repeated for the cases of $t(3)$ errors and log-normal errors, with cut-offs larger than their traditional counterparts. However, for the heavy-tailed and skewed error distributions, there was also a noticeable tendency for the cut-offs to become asymmetric, particularly so for small n in the log-normal case. In a sense, the bootstrap distribution of the influence measure is responding to particular features of the underlying error distribution. For the heavy-tailed error case, the bootstrap distributions of the influence measures tend to be heavier tailed than the traditional size-adjusted cut-offs suggest, particularly in small samples, while for the skewed error case, the influence measures have skewed bootstrap distributions, more so in small samples. As a result, the bootstrap cut-offs being larger and asymmetric compared with their traditional counterparts mean that the bootstrap method may not flag some points flagged by traditional techniques in small samples. This result is sensible in that the points not flagged in this circumstance may *not* be all that unusual in a small sample context or in a situation where a skewed underlying error distribution exists. For the DFFITS and DFBETAS measures, the traditional method routinely flags a higher percentage of points than the bootstrap method, even when there is no deliberately inserted influential point (so that such points occur only at random). Also, the percentage of points flagged by the bootstrap method is relatively unaffected by the deliberate insertion of an influential point in the simulated data – this behaviour is appropriate in that other, randomly occurring, potentially influential points are likely to be *relatively less* influential than the deliberately inserted point, and the bootstrap distribution takes this internal scaling into account automatically.

Our methods produce particularly promising results for the Cook's D measure, with cut-offs becoming steadily more stringent than the traditional method as n increases. For $n = 20$, when no influential point is deliberately inserted into the data, the traditional Cook's D measure flags only a very small percentage of data points on average as suspicious compared with the DFBETAS measure, while the bootstrap method flags roughly the same number of points as noted using DFBETAS. For $n = 50$ with no inserted influential point, the traditional method routinely flags *no* points, while the bootstrap Cook's D measure flags roughly the same number of points as flagged for the DFBETAS measures. In the presence of an inserted influential point, the traditional cut-off for Cook's D routinely flagged only that point, despite more points appearing influential with respect to multiple betas. The bootstrap method for Cook's D identified about the same number of points as DFBETAS, detecting the deliberately inserted point reliably.

Our results suggest that, apart from the case of Cook's D, the traditional cut-offs are too stringent when samples are small and model assumptions are not satisfied, while for Cook's D, the traditional cut-off is too liberal. For each of the non-normal error distributions, when there was a deliberately inserted influential point, the bootstrap method performed well, identifying the point in most cases, although marginally less often than the traditional method, except in the case of Cook's D, where our method typically identified the point more frequently. This phenomenon is perhaps unsurprising since while the inserted point may appear unusual in large sample, normal-errors settings (that is, those for which the traditional cut-offs are designed), it may not be so unusual in a small-sample, non-normal error situation. The bootstrap automatically provides a kind of 'internal scaling' that takes account of features of the original data, such as heavy tails or skewness in the error distribution.

References

Atkinson, A.C. (1986), 'Masking Unmasked', *Biometrika*, 73, 533–541.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.

Brown, G.C., and Lawrance, A.J. (2000), 'Theory and Illustration of Regression Influence Diagnostics', *Communications in Statistics: Theory and Methods*, 29, 2079–2107.

Chatterjee, S., and Hadi, A.S. (1986), ‘Influential Observations, High Leverage Points, and Outliers in Linear Regression’, *Statistical Science*, 1, 379–393.

Cook, R.D. (1977), ‘Detection of Influential Observation in Linear Regression’, *Technometrics*, 19, 15–18.

Cook, R.D. (1979), ‘Influential Observations in Linear Regression’, *Journal of the American Statistical Association*, 74, 169–174.

Cook, R.D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

Davison, A.C., and Hinkley, D.V. (1997), *Bootstrap Methods and Their Applications*, Cambridge: Cambridge University Press.

Davison, A.C., and Tsai, C.-L. (1992), ‘Regression Model Diagnostics’, *International Statistical Review*, 60, 337–353.

Efron, B. (1979), ‘Bootstrap Methods: Another Look at the Jackknife’, *The Annals of Statistics*, 7, 1–26.

Efron, B. (1992), ‘Jackknife-After-Bootstrap Standard Errors and Influence Functions (with discussion)’, *Journal of the Royal Statistical Society, Series B*, 54, 83–127.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Belmont: Chapman and Hall.

Fox, J. (1991), *Regression Diagnostics*, Newbury Park, CA: Sage Publications.

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C., and He, X. (2002), ‘Influence Diagnostics and Other Tests for Semiparametric Mixed Models’, *Journal of the Royal Statistical Society, Series B*, 64, 565–579.

Kim, C. (1996), ‘Cook’s Distance in Spline Smoothing’, *Statistics and Probability Letters*, 31, 139–144.

Lawrance, A.J. (1995), ‘Deletion Influence and Masking in Regression’, *Journal of the Royal Statistical Society, Series B*, 57, 181–189.

Martin, M.A., and Roberts, S. (2006), ‘An Evaluation of Bootstrap Methods for Outlier Detection in Least Squares Regression’, *Journal of Applied Statistics*, 33, 705–722.

Martin, M.A., Roberts, S., and Zheng, L. (in press), ‘Delete-2 and Delete-3 Procedures for Unmasking in Regression’, *Australian and New Zealand Journal of Statistics*.

Muller, E.K., and Mok, M.C. (1997), ‘The Distribution of Cook’s D Statistics’, *Communications in Statistics: Theory and Methods*, 26, 525–546.

Neter, J., Wasserman, W., and Kutner, M.H. (1990), *Applied Linear Statistical Models* (3rd ed.), Boston: Irwin.

Pena, D. (2005), ‘A New Statistic for Influence in Linear Regression’, *Technometrics*, 47, 1–12.

Pena, D., and Yohai, V.J. (1995), ‘The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix’, *Journal of the Royal Statistical Society, Series B*, 57, 145–156.

Ramsey, F., and Schafer, D. (2002), *The Statistical Sleuth – A Course in Methods of Data Analysis*, Belmont, CA: Duxbury Press.

Weisberg, S. (1983), ‘Some Principles for Regression Diagnostics and Influence Analysis’ (discussion of a paper by Hocking), *Technometrics*, 25, 240–244.