2024 ■ Vol. 20 ■ no. 2

Identifying Influential Observations in Multiple Regression

Carmel Camilleri [a], Udi Alter [a] & Robert A. Cribbie [a]

[a]Department of Psychology, York University

Abstract ■ Linear models are particularly vulnerable to influential observations which disproportionately affect the model's parameter estimates. Multiple statistics and numerous cut-off values have been proposed to detect highly influential observations including Cook's Distance (CD), Standardized Difference of Fits (DFFITS) and Standardized Difference of Beta (DFBETAS). This paper reports on a Monte Carlo simulation study that assesses the effectiveness of these methods and recommended cut-off values under various conditions, including different sample sizes, numbers of predictors, strengths of variable associations, and non-sequential versus sequential analysis approaches within a multiple linear regression framework. The findings suggest that the proportion of observations identified as highly influential varies significantly based on the chosen diagnostic method and the thresholds used for detection. Consequently, researchers should consider the implications of their methodological choices and the thresholds they apply when identifying influential data points.

Keywords ■ Influential Cases, Monte Carlo Simulation, Outliers, Cook's Distance, DFFITS, DFBETAS, Regression.

✉ carmel01@yorku.ca

10.20982/tqmp.20.2.p096

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

Introduction

When analyzing data using linear models, it is common for particular observations to be inconsistent with the others in the data (Barnett & Lewis, 1984). Data points which are inconsistent with other observations in the dataset are referred to as outliers. Outliers often affect the results of statistical analyses in a substantial way. Therefore, it is valuable to identify potential outliers and assess their impact on the results and conclusions of a study. In some instances, it is easy to identify outliers via a simple graph (e.g., scatter plot), whereas in others, outliers cannot be detected as easily but nevertheless have an important effect on the analysis results.

There are three primary ways to quantify outlying cases: leverage, discrepancy, and influence. A high-leverage data point is one with "extreme" values on the predictor variable(s) (Faraway, 2004; Wei et al., 1998). For instance, an individual could have an extreme value on one variable (e.g., income) or a combination of variables (e.g., income and depression), and in both situations, these observations would have high leverage. A high discrepancy data point has an unusual outcome value given its predictor value (Faraway, 2004). For example, when utilizing the least squares method to plot a regression line on a set of observations, an observation far away from the line of best fit can be identified as having a high discrepancy. Although leverage and discrepancy are crucial considerations in outlier detection, this paper focuses on data points considered "influential." Influence is a function of discrepancy and leverage. A highly influential case has an unusually large effect on the estimated parameters of the linear model (e.g., intercept, slope), whereby removing this case will substantially change the coefficient estimates (Salkind, 2010). The identification and appropriate treatment of influential cases are vital because a highly influential case may alter the parameter estimates (and significance tests) such that the coefficients do not adequately represent the relationships in question for the bulk of the data.

That said, researchers should not instinctively remove data deemed influential. Instead, a researcher should flag this value for further investigation. Anscombe (1960)

places outliers into two distinct categories: 1) observations arising from errors in the data (e.g., data entry errors, such as reporting in centimeters instead of inches or misplacing a decimal point); and 2) observations arising from the inherent variability of the data. Identifying to which category a highly influential observation belongs can sometimes be straightforward and other times extremely complicated. Obviously, errors in the dataset need to be addressed, but what to do about genuine cases that strongly influence the results is often unclear, with no consensus in the literature (Dhakal, 2017). These decisions are often highly subjective and context-dependent, but still might have strong implications. The subsequent section provides an overview of three popular model-based methods for detecting influential cases: Cook's distance (CD), Difference in Fits (DFFITS), and Difference in Betas (DFBETAS).

Commonly Employed Methods for Detecting Influential Cases

Using figures is an initial step that one should take when approaching outlier analysis (Felt et al., 2017; Hebbali, 2020; Tukey, 1977). For example, a scatter plot can be a vital tool to detect outliers when there is a single outcome and a single predictor variable. Including a regression line with a scatter plot can also greatly assist researchers in visually identifying influential cases. Researchers are encouraged to utilize visualization strategies combined with outlier detection statistics to better understand the influential data points. Next, there are various statistical techniques that researchers can use to identify highly influential cases.

Cook's Distance (CD)

CD is one of the most common methods used to locate influential observations in a dataset (Zhu et al., 2012). Cases with higher leverage and higher discrepancy have increased CD scores (Cook, 1977). CD reflects the change in the fitted response (predicted) values when the ith data point is removed. CD is calculated as:

$$CD_i = \frac{\sum_{j=1}^{N}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)}{(p+1)\ MSE}$$

where $\hat{Y}_j$ is the predicted response value for case j, $\hat{Y}_{j(i)}$is the predicted response value for case j, where the fit does not include observation i, p is the number of predictors (plus 1 to account for the intercept term), N is the sample size, and MSE is the model mean squared error based on all observations. Note that the numerator assesses the difference in the predicted response when the ith case is included versus not included in the data. The denominator accounts for the expected variability between observed and predicted response values.

CD is not a statistical significance test that researchers should solely use to accept or reject particular observations in a given dataset (Cook, 2011); instead, it is best used to indicate the extent to which each observation is influential or outlying. Deciding on an appropriate cut-off value for identifying when an observation is highly influential is thus an imperative, but often complicated exercise. Cook (1977) stated that any observation with a CD value greater than 1 should be considered an influential case when deciding upon a particular cut-off value for CD. A cut-off of 1 is a popular cut-off value for interpreting CD among researchers and textbook authors (e.g., Cohen et al., 2014; Tabachnick & Fidell, 2019). However, this approach may be considered too conservative (i.e., not able to detect highly influential outliers) as it rarely "catches" any cases due to its high threshold (McDonald, 2002). For this reason, researchers who want a more conservative approach sometimes use 0.5 as a cut-off point, with CD values larger than 0.5 considered highly influential (Cook & Weisberg, 1982). Another popular cut-off for CD is $4/(N-p-1)$ (Cook, 1977). It is obvious that with even a small sample size, this cut-off is much smaller than CD = 1 or CD = 0.5. For example, with $N = 20$ and two predictor variables, $4/(N-p-1) = 4/(20-2-1) = 0.24$.

Standardized Difference in Fits (DFFITS)

An alternative approach to identifying influential cases is DFFITS. CD and DFFITS follow similar logic: delete one observation at a time, then refit the regression model on $N-1$ observations and explore the difference in the model parameters. The equation for DFFITS is as follows:

$$DFFITTS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sigma_{(i)}\sqrt{h_{ii}}}$$

The numerator describes the difference in the predicted values for Y_i with data point i included and without data point i included in the regression model. $\sigma_{(i)}$ represents the standard error estimated without the ith point included and h_{ii} is the leverage value for the point. DFFIT is an unstandardized version of DFFITS. The difference between these two tests is that DFFIT only computes the numerator values of DFFITS, excluding the denominator. Numerous cut-offs have been proposed to identify cases that are influential when using DFFITS. The initial cut-off employed by Welsch and Kuh (1977) was $2\frac{\sqrt{(p+1)}}{(N-p-1)}$, whereas Belsley et al. (1980) proposed cut-offs of 2 and $2\sqrt{p/N}$. The cut-offs 2 and $2\sqrt{p/N}$ are used in the current study.

Standardized Difference of Betas (DFBETAS)

Following the same intuitive logic as the CD and DFFITS procedures, DFBETAS is an influential case detection method that investigates the standardized difference in re-

Table 1 ■ Overview of Outlier Cut-Off Values

Influential Case Detection Method	Cut-Off
CD	
Cook (1977)	1
Cook and Weisberg (1982)	0.5
Cook (1977)	$4/(N-p-1)$
DFFITS	
Belsley et al. (1980)	2
Belsley et al. (1980)	$2\sqrt{p/N}$
DFBETAS	
Belsley et al. (1980)	2
Belsley et al. (1980)	$2/\sqrt{N}$
Bollen and Jackman (1985)	1

gression coefficients when a particular (ith) case is present, versus not present, in the dataset. The difference between CD or DFFITS and DFBETAS is that DFBETAS assesses how each coefficient in the model changes when deleting the ith observation, rather than how the predicted values change (as with CD and DFFITS). DFBETAS can be calculated as:

$$DFBETAS_{ij} = \frac{\hat{b}_k - \hat{b}_{(k)i}}{s_{(i)}\sqrt{(X'X)_{jj}}}$$

where $\hat{b}_k$ is the kth coefficient estimate from the regression model calculated using all the data, $\hat{b}_{(k)i}$ is the kth coefficient estimate from the regression model calculated without the ith observation, and $(X'X)_{jj}$ is the (j,j)th (diagonal) element of $(\mathbf{X}'\mathbf{X})^{-1}$ for all observations (where $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the variance-covariance matrix of the predictor variables). As per the previous numeric cut-off strategies, there are various thresholds that one can utilize to determine if a case should be deemed influential. Belsley et al. (1980) proposed a general cut off of 2 and a sample size-adjusted cut-off of $2/\sqrt{N}$. Values that exceed this threshold are to be considered influential. Bollen and Jackman (1985) proposed an alternative cut-off of DFBETAS value at 1.

Table 1 presents an overview of the numeric cut-offs proposed for each of the methods for quantifying high influence.

Sequential vs. Non-Sequential

When considering the various methods and cut-off values available to researchers, two main approaches are available to detect influential cases: sequential and non-sequential. For the sake of this discussion, we will assume that the researcher has decided to remove cases deemed influential; however, this does not need to be the default. Non-sequential entails researchers removing all values that exceed a particular cut-off value in one step. The sequential strategy (Aggarwal & Sathe, 2017) consists of researchers flagging the most extreme value that exceeds the cut-off value, removing it, and, if necessary, identifying further cases by rerunning the model using the updated data and cut-off value. This process is repeated iteratively until no other values exceed the selected cut-off. Since outlying cases often inflate the standard errors of models (and influential case detection methods), sequential methods 're-run' the models and influential case detection methods after removing the most influential case. We are unaware of any previous literature exploring the advantages or disadvantages of one approach over the other.

Current Study

Despite their accessibility in the literature and availability in statistical software, recommendations regarding implementing the CD, DFFITS, and DFBETAS influential case detection methods are limited and sometimes confusing, with no details regarding the consequences of choosing one method over another. Furthermore, there is little research comparing the methods or the cut-offs proposed for each method. Given the impact that influential observations can have on the results of linear model analyses, we sought to compare available strategies. This study uses a Monte Carlo simulation approach to compare different outlier detection strategies and cut-offs within a multiple linear regression framework. The current study compares the procedures (along with the associated cut-offs) in terms of the proportion of influential cases identified when the data are sampled from a multivariate normal distribution. Given the nature of the study, there are no formal hypotheses regarding the effects; instead, our interest is in highlighting relevant differences among the procedures and cut-offs in terms of the proportion of cases identified as highly influential. To the best of our knowledge, this is the first study to compare the CD, DFFITS and DFBETAS procedures, using all of the recommended cut-offs outlined in Table 1, with respect to

Table 2 ■ Summary of Simulation Parameters and Parameter Values

Parameter	Value
N	25, 50, 100, 250, 500, 1000
COV	Equal: COV(VAR1, VAR2) = COV(VAR1, VAR3) = COV(VAR2, VAR3) = 0.1, 0.3, or 0.5;
	Unequal: COV(VAR1, VAR2) = 0.1, COV(VAR1, VAR3) = 0.3, COV(VAR2, VAR3) = 0.5
es	Equal: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0.1, 0.3$, or 0.5;
	Unequal: $\beta_1 = 0.1, \beta_2 = 0.3, \beta_3 = 0.5, \beta_4 = 0.1, \beta_5 = 0.3, \beta_6 = 0.5$
p	1, 3, 6
s/ns	Sequential, Non-Sequential

Note. Sample size (N), covariance between predictor variables (COV), effect size of regression coefficients between predictor and outcome variables (es), note, unequal conditions are only for 6 predictor models, number of predictor variables (p), sequential and non-sequential outlier detection method (s/ns). Note that the covariance between predictors in the table only shows the conditions with three predictors. However, with six predictors, the pattern of association follows the same logic only with 15 unique variable pairings.

identifying influential cases in a multiple regression framework.

Monte Carlo Simulation Study

A Monte Carlo study was used to compare the proportion of influential cases identified using three influential case methods, CD, DFFITS and DFBETAS. In addition, the cut-offs were also varied for each method (See Table 1). The study design is a 3 (number of predictors in the model) $\times$ 6 (total sample size) $\times$ 5 (coefficient effect sizes) $\times$ 6 (relationship between predictor variables) $\times$ 2 (sequential vs. non-sequential), and removing any redundant conditions (e.g., 0.1, 0.3, and 0.5 between-predictor correlations with a single predictor), resulting in a total of 210 unique conditions. The simulation parameters are summarized in Table 2.

Procedure

Multivariate normal data was simulated using the `SimDesign` package (Chalmers & Adkins, 2020) in R (R Core Team, 2021). The multiple regression model was conducted via the `lm` function within the `stats` package in R. Once the model was run, each of the influential case methods was employed (i.e., CD, DFFITS, DFBETAS) along with each of their accompanying cut-off values to identify any influential cases. CD, DFFITS, and DFBETAS were computed using the `cooks.distance`, `DFFITS`, and `DFBETAS` functions, respectively (all within the stats package in R). For each condition, 5000 simulations were conducted. The proportion of outliers detected for each method, and the accompanying cut-off value, was determined by taking the number of cases deemed as highly influential and dividing that number by N.

Simulation Conditions.

The conditions used within the Monte Carlo study are presented in Table 2.

Sample Size. The selected sample-size values were chosen to reflect typical sample sizes employed in psychology research. The selected sample sizes included in the simulation are $N = 25, 50, 100, 250, 500, and 1000$.

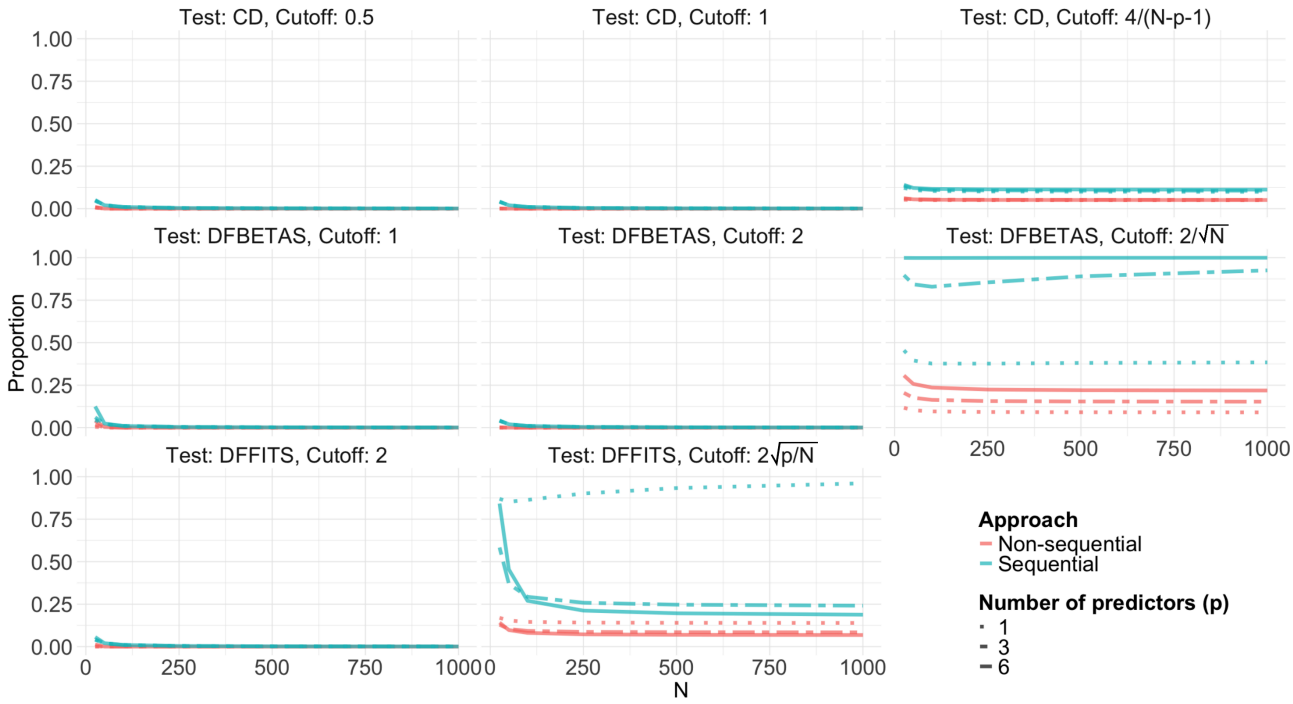
Number of Predictors. The number of predictors used in each model was $p = 1, 3$, and 6. These predictors were selected to represent typical designs used in social science research (e.g., Tabachnick & Fidell, 2019; Yarkoni & Westfall, 2017).

Relationship Between Predictors. The relationship between predictors (covariance) in the model varied across numerous conditions. In particular, the covariance among predictors was manipulated to assess the degree to which relationships among predictors affect the proportion of outliers. The population covariances among predictors were set at 0.1, 0.3, 0.5. Each pair of predictors either had the same relationship (e.g., in the 3-predictors conditions, the relationship was either [0.1, 0.1, 0.1], [0.3, 0.3, 0.3], or [0.5, 0.5, 0.5]) or "heterogeneous" correlations with varying or "mixed" magnitudes (e.g., COV(VAR1, VAR2) = 0.1, COV(VAR1, VAR3) = 0.3, COV(VAR2, VAR3) = 0.5, etc.).

Effect Size of Regression Coefficients. The effect size of the regression coefficients between the predictor and outcome variable varied across conditions. More specifically, various standardized regression coefficients were utilized to examine the unique relationships among each predictor and the outcome variable. Effect size regression coefficients were set at 0.1, 0.3, 0.5. In some instances, all regression coefficients had the same effect size value (i.e., [0.1, 0.1, 0.1], [0.3, 0.3, 0.3], and [0.5, 0.5, 0.5]), whereas others included a "mix" of effect sizes in the three (i.e., [0.1, 0.3, 0.5]) and six predictor model (i.e., [0.1, 0.3, 0.5, 0.1, 0.3, 0.5]). These values reflect the partial relationship between predictor variables and the outcome.

Approach: Sequential and Non-Sequential. The influential case method (and associated cut-off value) was

Figure 1 ■ Simulation Results: Total Proportion of Influential Cases Identified



conducted using either a sequential or non-sequential approach. As described in detail above, the sequential approach removes the most influential case (that exceeds the cut-off), and then repeats the procedure with the new data. This procedure continues until no case exceeds the cut-off. The non-sequential approach removes all cases that exceed the cut-off all at once (with no further influential case detection).

Results

Across all manipulated conditions in the simulation study, the magnitude of the effects and covariance (i.e, the strength of the relationship between the predictors and the outcome, or the strength of the relationships among the predictors themselves) had very little influence on the proportion of identified influential cases as indicated by the arrow and red square.

Therefore, to avoid redundancy, both Figure 1 and Figure 2 illustrate the marginal results, collapsing across all effect sizes and covariances.“ The results are summarized in Figure 1. Additionally, please refer to Figure 2, which highlights the smaller proportions (< .30) of outlier detection. The tabulated raw results and simulation code are available on OSF: osf.io/ehvym/.

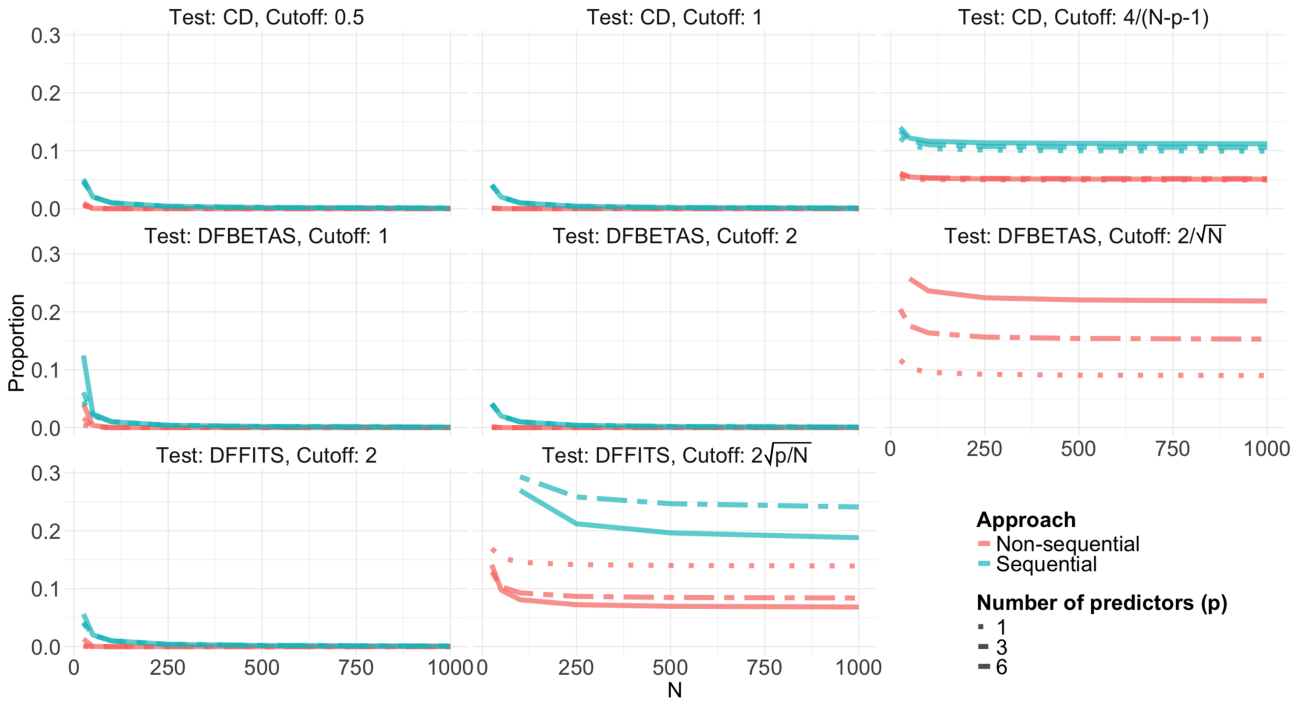
CD

CD cut-offs can be separated into two groups: non-sample-size-dependent cut-offs (1 and 0.5) and a sample-size-dependent cut-off of $4/(N-p-1)$. A non-sequential approach with a cut-off of 1 and 0.5 identified less than 1% of cases as influential, while a cut-off of $4/(N-p-1)$ identified approximately 5%-7% of cases as influential. When taking a sequential approach, non-sample-size-dependent cut-offs still identified less than 1% of cases as influential, while a cut-off of $4/(N-p-1)$ identified approximately 12%-15% of cases as influential. Across all conditions, CD was relatively consistent with respect to the proportion of influential cases observed across different sample sizes, number of predictors, and effect sizes.

DFFITS

Using a non-sequential approach, a cut-off of 2 identified less than 1% of cases as influential across all conditions. The $2\sqrt{p/N}$ cut-off identified approximately 7%-17% of cases as influential across all conditions, with rates lower as N increased. Alternatively, when employing a sequential approach, a cut-off of 2 identified less than 1% of cases as influential, while a cut-off of $2\sqrt{p/N}$ identified 20%-90% of cases across all conditions. Within a simple linear

Figure 2 ■ Simulation Results: Proportion of Influential Cases with 30% or Fewer Identified. These results replicate the same findings as Figure 1, with an emphasis on proportions of outliers identified less than 30% of the time.



regression framework (i.e., one predictor), for $N > 100$, the proportion of influential cases identified increased with N. In a three and six-predictor model, the proportion of influential cases observed rapidly decreased as N increased.

DFBETAS

Starting with the non-sequential approach, DFBETAS with non-sample-size dependent cut-offs (1, 2) were conservative, identifying less than 1% of cases as influential, while DFBETAS with a sample-size-dependent cut-off of $2\sqrt{N}$ identified 12%-30% of cases as influential. When employing a sequential approach, non-sample-size-dependent cut-offs identified less than 1% of cases as influential, while a cut-off of $2\sqrt{N}$ identified approximately 37%-100% of cases as influential across all conditions. As demonstrated by the limited variability in the proportion of observed influential cases, sequential and nonsequential approaches with non-sample size-dependent cut-offs remained relatively unaffected by simulation conditions (e.g., sample size, number of predictors). When discussing a sequential approach with a cut-off of $2\sqrt{N}$, there was a substantially greater proportion of observed influential cases compared to the non-sequential method with a wide interval of observed cases. For a cut-off of $2\sqrt{N}$, the rates also increased with the number of predictors. As an example, approximately 38% of cases were identified as influential with one predictor, compared to approximately 99% of cases being identified as influential in a six-predictor model.

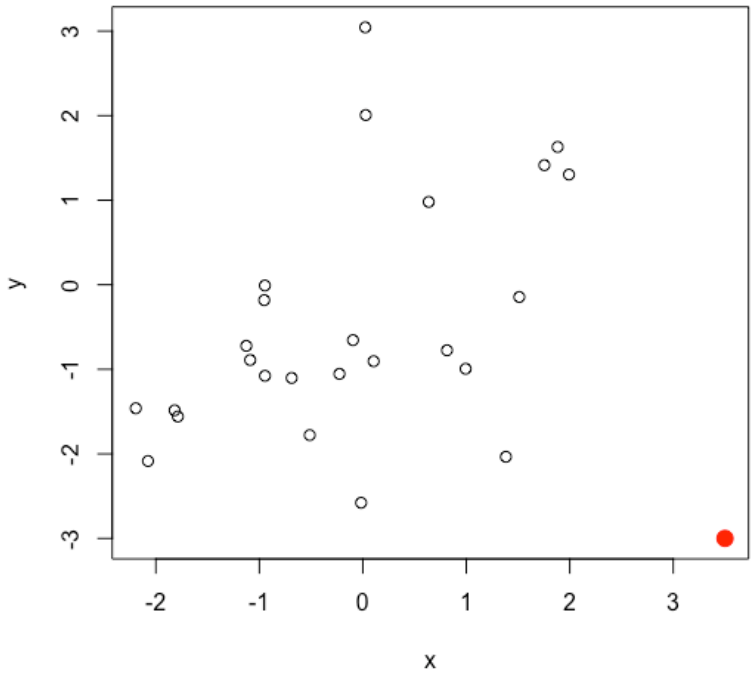
Monte Carlo Simulation Summary

When using DFBETAS or DFFITS with a sequential approach and a sample-size- dependent cut-off, there was a large amount of variability in the proportion of outliers observed across simulation conditions, with rates approaching 100% of cases being deemed influential. CD was relatively robust across simulation conditions (i.e., the most consistent in detecting outliers regardless of condition) in terms of the mean proportion of identified influential cases, with rates elevated, as expected, in sequential conditions. When comparing CD, DFFITS and DFBETAS, non-sample-size-dependent cut-offs always identified less than 2% of cases as influential in the non-sequential cases, with slightly higher rates in the sequential conditions.

Simulating Outlier-Contaminated Data

Given that the CD statistic with a $4/(N-p-1)$ cut-off and non-sequential approach was consistent across the number of predictors and sample size, detecting, on average,

Figure 3 ■ Contaminated Data Point Example. The red dot represents a contaminated observation of (2.5, - 2.5) to a positive, moderate bivariate association between two normally distributed variables, each with a mean of 0 and a standard deviation of 1.



5% of cases as outliers, we find this approach to be the best of all options investigated. Thus, an additional simulation was performed to understand how well this method-threshold combination correctly detects data artificially contaminated with outliers. We added outliers into the data by manually simulating points that were clearly inconsistent with the observations along the regression line. For example, inserting a contaminating observation of (2.5, - 2.5) to a positive, moderate bivariate association between two normally distributed variables, each with a mean of 0 and a standard deviation of 1. Please see Figure 3 to view a simple example that illustrates a set of 25 observations with a single contaminated data point. The simulation is a 2 (number of predictors in the model; $p = 1, 3$) $\times$ 2 (total sample size; $N = 50, 100$) $\times$ 2 (number of outliers added to the data; outliers = 1, 3) design, resulting in a total of 8 unique conditions. For each condition, 500 simulations were conducted. In all simulation iterations, our recommended outlier detection method achieved a 100% detection rate in identifying outliers added to the dataset.

Discussion

Outlier detection is an important linear model diagnostic that is commonly used by researchers prior to analyzing and interpreting results. However, implementing influential case methods can be confusing for researchers, especially since there is little information available regarding the implications of selecting particular methods and their accompanying cut-off values. The current study sought to inform researchers on how the particular influential case method and cut-off they select can affect the proportion of detected influential cases across various linear models, even with “perfectly” (i.e., artificially generated using a simulation) multivariate normal data and no anticipated outliers. This study investigated CD, DFFITS and DFBETAS across varying parameters and with various proposed cut-offs to better understand the differences when identifying influential cases in a model. Simulation results supported the contention that there is high variability across influential case methods/cut-offs. Once researchers identify a particular influential case method and a preferred cut-off, they are then faced with selecting either a sequential or nonsequential approach for their analysis. Given the lack of resources in the literature, this study sought to examine how each approach can affect the proportion of influential cases observed.

When comparing the number of cases identified as influential for non-sample size dependent cut-offs (CD = .5 or

1; DFITTS = 2; DFBETA = 1 or 2) across all methods, near zero percent of cases were flagged as influential in most conditions. The proportions of influential cases for non-sample-size-dependent cut-offs were obtained under perfectly normal multivariate data with no outliers inserted into the data, and hence, these methods are performing as expected. However, even when multivariate normal data are simulated, there is an expectation that some of the cases will be extreme/influential.

CD was the most consistent method across conditions in the simulation, having higher rates with sequential methods and with the sample-size-dependent cut-off. On the other hand, the rates for the sequential versions of the DFFITS with a cut-off of $2\sqrt{p/N}$ and DFBETAS with a cut-off of $2\sqrt{N}$ varied substantially by sample size and the number of predictors included in the model. Even with a non-sequential approach, the rates consistently identified more cases as influential than the sample-size-dependent cut-offs for CD. More specifically, these methods consistently identified more than 10% of cases as influential within the model. This level of outliers represents a higher level than what we would expect from distributions simulated from the multivariate normal distribution (Aguinis et al., 2013); as such, we do not recommend employing these strategies to detect influential cases. As expected, selecting a sequential approach yielded a greater proportion of cases identified as influential than a non-sequential approach. This finding was consistent across all influential case methods/cut-offs and simulation parameters. From a theoretical perspective, this finding is logical: as outliers are removed from the dataset, the standard deviation of the residuals generally gets smaller as the line of best fit gets pulled toward the other data points in the distribution. This process results in higher influence statistics. More so, influential cases can inflate the standard error of linear models, implying a need to re-run the models and re-identify influential cases after removing the most extreme influential case. However, the results of the current study suggest that the sequential approach generally identifies an unacceptably high number of cases as influential.

Recommendations

To blindly remove outliers that exceed a cut-off is a practice that we neither endorse nor encourage. We advise researchers to conduct a sensitivity analysis, to construct a model with outliers removed and compare its performance to that of the original model. If the model's fit and estimated parameters undergo meaningful alterations after outlier removal, researchers should investigate these influential data points further. If no meaningful alterations to the model occur, we recommend leaving those values in the model. This systematic approach facilitates a more comprehensive evaluation of the model's validity and sheds light on the impact of specific cases.

Influential case statistics should be used in conjunction with visualizations (e.g., scatter plots, residuals vs. leverage plots). Visualizations provide an additional tool for researchers to investigate cases identified by the selected influential case method as extreme.

After taking the necessary steps needed to better understand each case identified as influential, researchers may decide to remove an outlier (if they have legitimate reasons to do so) or simply leave the case in the model and report the results with and without the influential case (i.e., sensitivity analyses). Most importantly, transparency and justification are necessary when a researcher makes these decisions. Informing the reader of the process taken to identify and deal with influential cases increases the transparency and reliability of the current study. These open science practices are one of many ways researchers can improve social science's validity and reliability (Hales et al., 2018).

The results of the present Monte Carlo study inform researchers regarding the consequences of their selection of a particular influential case detection method and cut-off with respect to the proportion of "flagged" highly influential cases. This highlights the importance of researchers' degrees of freedom in detecting influential cases. As previously stated, we recommend that researchers use the non-sequential CD approach with a sample size dependent cut-off of $4/(N-p-1)$. This method provided an excellent balance between consistency (across N and p) and level of conservativeness (5% of cases detected as influential). Further, this approach was also able to correctly flag contaminating observations with 100% accuracy. That said, there is no "right" rate of rejected cases for an outlier detection approach/cut-off. However, some rates are clearly unacceptable (e.g., > 30%). In our simulation, we expected a reasonable or acceptable rate to be low but non-zero, and this study helped identify procedures that fall into this broad category.

Outside of the simulation study, however, deciding on an acceptable rate of outlier detection can be a challenging but necessary task. The decision of an acceptable rate can vary significantly depending on the research context, question, study design, target population, and the researcher's judgement. Questions that researchers may ask themselves in guiding their decision are whether it is reasonable that a certain proportion (e.g., 5%) of the data is identified as outliers, how removing these cases (if necessary) would impact the sample size and statistical power of the model, and are there alternative statistical techniques or robust methods that could be used to mitigate the influence of outliers without excluding them entirely? Because these decisions are subjective and at the discretion of the research

team, we invite researchers to consider them carefully in advance, provide clear justifications for their choices, and ensure transparency when reporting their results.

Limitations and Future Research

Given the nature of the current study, some limitations deserve notice. One limitation is that the current study was not able to investigate all possible Monte Carlo simulation conditions (e.g., sample size, number of predictors, cut-offs, and relationships among variables). Therefore, it is possible that the results would not generalize to other conditions. As well, we also recognize that the cut-offs employed in this manuscript are not an exhaustive list of all the cut-offs proposed for our influential case detection methods of interest. As the current study sought to include the most popular and recommended cut-offs in the literature, future research should expand the cut-offs used in the current study to understand how different cut-offs affect the proportion of outliers identified.

The current study also did not investigate situations in which the assumptions of linear models were violated (e.g., nonlinearity, heteroscedasticity). Future research may benefit from violating particular model assumptions and by including a more diverse array of simulation parameters, to see the effect of these conditions on the proportion of cases deemed influential.

Finally, as identifying influential cases with DFBETAS depends directly on the number of predictors in the model, future research is encouraged to evaluate the DFBETAS approach with familywise error control (e.g., Bonferroni-Holm; Holm, 1979) to see how this change affects the mean proportion of influential cases observed.

Conclusion

This study revealed substantial heterogeneity in the various influential case methods and cut-offs used for identifying outliers in a multiple linear regression framework. More specifically, this research highlights the need for methodological and statistical awareness among researchers before approaching outlier analysis. Understanding the variability among mainstream influential case methods and their associated cut-offs accentuates how decisions may affect the validity and reproducibility of results. Based on the results of this study, we recommend that researchers utilize the CD approach with a cut-off of $4/(N-p-1)$ for identifying influential cases in multiple regression scenarios. It is hoped that this paper will help researchers make informed decisions regarding the approach they select for identifying highly influential cases in multiple regression models.

Authors' note

We would like to thank Linda Farmus, Nataly Beribisky, and Naomi Martinez Gutierrez for their suggestions and discussions that added to the quality of the manuscript. This project was partially funded by a Social Sciences and Humanities Research Council of Canada Discovery Grant to Robert A. Cribbie (Grant #: 435-2016-1057).

References

Aggarwal, C. C., & Sathe, S. (2017). *Outlier ensembles*. Springer International Publishing. doi: 10.1007/978-3-319-54765-7.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. doi: 10.1177/1094428112470848.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, *2*(2), 123–146. doi: 10.1080/00401706.1960.10489888.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data* (2nd). John Wiley & Sons.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*.

Bollen, K. A., & Jackman, R. W. (1985). Regression diagnostics. *Sociological Methods & Research*, *13*(4), 510–542. doi: 10.1177/0049124185013004004.

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable monte carlo simulations with the simdesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. doi: 10.20982/tqmp.16.4.p248.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioural sciences*. Psychology Press.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18. doi: 10.2307/1268249.

Cook, R. D. (2011). Cook's distance. *International Encyclopedia of Statistical Science*, 301–302. doi: 10.1007/978-3-642-04898-2_189.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman; Hall. https://hdl.handle.net/11299/37076

Dhakal, C. P. (2017). Dealing with outliers and influential points while fitting regression. *Journal of Institute of Science and Technology*, *22*, 61–65. doi: 10.3126/jist.v22i1.17741.

Faraway, J. J. (2004). *Linear models with R*. CRC Press.

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, *8*(863), 1–9. doi: 10.3389/fpsyg.2017.00863.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2018). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*(1), 13–31. doi: 10.1007/s40614-018-00186-8.

Hebbali, A. (2020, February 10). *olsrr: Tools for building OLS regression models* [R-Packages]. https://cran.r-project.org/web/packages/olsrr/index.html

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

McDonald, B. (2002). *A teaching note on Cook's distance - a guideline* [Mro.massey.ac.nz]. http://hdl.handle.net/10179/4352

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Salkind, N. (2010). *Encyclopedia of research design*. Sage. doi: 10.4135/9781412961288.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th). Pearson.

Tukey, J. W. (1977). *Exploratory data analysis, vol. 2* (Vol. 2). Springer.

Wei, B., Hu, Y., & Fung, W. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, *25*(1), 25–37. doi: 10.1111/1467-9469.00086.

Welsch, R. E., & Kuh, E. (1977, March 1). *Linear regression diagnostics* [Papers.ssrn.com]. https://ssrn.com/abstract=260362

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393.

Zhu, H., Ibrahim, J. G., & Cho, H. (2012). Perturbation and scaled Cook's distance. *The Annals of Statistics*, *40*(2), 785–811. doi: 10.1214/12-aos978.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/ehvym

Citation

Camilleri, C., Alter, U., & Cribbie, R. A. (2024). Identifying influential observations in multiple regression. *The Quantitative Methods for Psychology*, *20*(2), 96–105. doi: 10.20982/tqmp.20.2.p096.

Received: 04/01/2024 ~ Accepted: 14/06/2024