Computational Statistics and Data Analysis 108 (2017) 133–145

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A multi-row deletion diagnostic for influential observations in small-sample regressions

Daniel T. Kaffine[a,*], Graham A. Davis[b]

[a] Department of Economics, University of Colorado Boulder, Boulder CO 80309, USA
[b] Division of Economics and Business, Colorado School of Mines, Golden CO 80401, USA

HIGHLIGHTS

- The economic inferences from growth regressions are often sensitive to influence points in the sample.
- We introduce a multi-row deletion method for identifying influence points.
- The method is simple to use and is intuitively appealing.
- We illustrate our method using simulated and real data.
- Our method complements DFBETAS and robust regression.

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 12 October 2016
Accepted 12 October 2016
Available online 26 October 2016

Keywords:
Regression diagnostics
Multi-row deletion
Inference
Influence points
Outliers
Growth regressions
Institutions
Resource curse

ABSTRACT

The inference from ordinary least-squares regressions is often sensitive to the presence of one or more influential observations. A multi-row deletion method is presented as a simple diagnostic for influential observations in small-sample data sets. Multi-row deletion is shown to be complementary to two related diagnostic tests, DFBETAS and robust regression. As an illustration, the technique is applied both to simulated data and to a real data set from an influential study examining the role of institutions for economic growth in resource-rich economies. Multi-row deletion reveals that the key economic insight that institutions matter is sensitive to small variations in sample, indicating additional analysis may be warranted.

1. Introduction

Applied researchers sometimes perform regressions on small ($N \approx 100$) data sets. A particular concern when working with these data sets is the potential for the inference from the regression to be sensitive to a few influential data points. In this paper, we propose a simple multi-row deletion analysis (MRDA) for identifying influential observations in small data sets that are likely to have only a few (≈5%) influential observations. Our MRDA approach draws a large number of near-N random subsamples in a multi-row deletion exercise. We introduce simple graphical methods for identifying whether a few influential observations are causing wide swings in regression coefficients and/or t-statistics and then identifying these

[*] Corresponding author. Fax: +303 492 8960.
E-mail addresses: daniel.kaffine@colorado.edu (D.T. Kaffine), gdavis@mines.edu (G.A. Davis).

http://dx.doi.org/10.1016/j.csda.2016.10.007
0167-9473/© 2016 Elsevier B.V. All rights reserved.

observations where they are shown to exist.[1] As we will show, our approach is complementary to two other diagnostic tools, DFBETAS and robust regression.

For context, consider the case of growth regressions. Growth regressions have been criticized on many levels, one of which is the instability of ordinary least-squares (OLS) coefficient estimates as the sample is altered in small ways (Brock and Durlauf, 2001; Durlauf et al., 2005). In many cases, small changes in sample lead to a reversal of the economic inference from the modeling exercise. One example is the finding by Auerbach et al. (1994) that the high social returns to equipment investment found by DeLong and Summers (1991) hinge on the inclusion of Botswana in the sample. Recently, Herndon et al. (2014) examine the results in Reinhart and Rogoff (2010) and find their average growth estimates for countries with high public debt/GDP ratios are sensitive to exclusion or inclusion of country-year observations for New Zealand. Easterly (2005) argues more generally that extreme observations drive the results in growth regressions.

Growth regressions are typically subject to investigation of specification robustness—for small changes in the specification, does the inference from the coefficients of interest remain robust (Levine and Renelt, 1992; Sala-i-Martin et al., 2004; Hauk and Wacziarg, 2009; Ley and Steel, 2012)? The sensitivity of economic inference to sample has received less systematic attention and is rarely rigorously explored by growth researchers (Lorgelly and Owen, 1999; Sturm and de Haan, 2005). Where the sample size is less than 30, researchers may be able to use brute force techniques to identify influential observations (e.g. Knack, 2003), though many growth regressions have sample sizes of around 100 and brute force methods become unmanageable.

Our premise is that researchers are interested in knowing when the modeling inference from a regression analysis is driven by a single observation or small group of observations, regardless of what the formal regression statistics say for their baseline regression (e.g., Stokey, 1994).[2] For example, for small changes in sample, do the signs on the coefficients of interest change? Does the magnitude of a coefficient move from being meaningful to irrelevant? Does the t-statistic change from signaling statistical significance to signaling statistical insignificance? In addition, the presence of influential observations may indicate data errors or model misspecification. At a minimum, they signal further analysis may be warranted before drawing general conclusions. Their presence should also be communicated to readers in keeping with Kennedy's 10th Commandment of Applied Econometrics: "Thou shalt confess in the presence of sensitivity" (Kennedy, 2002).

We illustrate the MRDA approach using simulated and real data. In the simulated data, we take ten samples of 90 data points and perturb three in each. We lead the reader through a near-N multi-row deletion simulation of OLS regressions on the data followed by an analysis of histograms of coefficient estimates and t-statistics from these simulations. Ordered tails plots are then used to identify the data points most frequently missing in the relevant tails of the histograms. In most cases, they are determined to be the perturbed data points. We then show how robust regression and DFBETAS underperform in the same analysis. For the real data analysis, we use the cross-section OLS growth regressions in Mehlum et al.'s (2006) (MMT) influential paper on the Resource Curse. MRDA shows that the inclusion of a single country in the data set is shown to be necessary and sufficient for their key economic insight to hold; inclusion of Malaysia in a near-N subsample provides evidence that strong institutions can turn resource abundance into a blessing. By contrast, exclusion of Malaysia in a near-N subsample suggests that strong institutions may not overcome the curse. Additional row deletion exercises reveal that other results in their paper are sensitive to groups of countries voluntarily left out of the sample. Our exploration of the data in MMT, as well as DeLong and Summers (1991), highlights that MRDA can provide insight to applied researchers that may not be apparent from DFBETAS and robust regression. Finally, we explore the computational limitations of MRDA. We find that it is useful primarily for small data sets with no more than 5% of the data points likely to be influence points.

2. Diagnostic methods for identifying influential observations

2.1. Multi-row deletion analysis

Before we introduce MRDA in more detail, it is useful to categorize the various types of extreme observations, for the vernacular is far from uniform. Following the exposition in Birkes and Dodge (1993) for least squares analysis, let h_{ii} be the ith diagonal entry in the matrix $X(X'X)^{-1}X'$, where X is the $N \times (p+1)$ matrix of explanatory variables x_{ij} augmented by a column of 1's. h_{ii} is the "potential" of the ith data point to influence the regression. The fitted value $\hat{y}_i$ can be expressed as $\hat{y}_i = (1-h_{ii})\hat{y}_i^{\star} + h_{ii}y_i = \hat{y}_i^{\star} + h_{ii}(y_i - \hat{y}_i^{\star})$, where $\hat{y}_i^{\star}$ is the estimate obtained when observation i is removed from the data.[3] An observation that has an arbitrarily extreme value for one of the independent variables is called a *leverage* point, corresponding to a large h_{ii}. An observation is an *outlier* if it is sufficiently extreme in the dependent variable conditional on the independent variables, as measured by the residual $y_i - \hat{y}_i$. By contrast, an observation has an *outlier effect* if it is

[1] While we will generally discuss and apply our method in the context of ordinary least squares, in principle it can be applied more generally, with the caveat that estimating large numbers of resamples may be computationally intensive for more complex estimation techniques. We also focus on cross-section regressions. An extension of our approach to panel data may be a useful area of inquiry.

[2] As Temple (2000) points out, OLS regression diagnostics may not signal the presence of influence points since the tests are carried out on the estimated residuals rather than the true disturbances.

[3] Note that in the following discussion we are assuming the (linear) model is correctly specified. See Figure Appendix B.1 in Appendix B for a bivariate graphical representation of outliers, leverage and influence points related to the discussion below.