Influence Analysis with Panel Data

Annalivia Polselli*

February 23, 2023

1 Introduction

Short panel data are widely used by applied researchers to conduct analyzes of interest. The nature of the research question or design itself may limit the number of observed units (e.g., number of countries, regions, or states; participants in an experiment; patients receiving a treatment, etc.). This data structure is common in many economic fields – for example, macroeconomic country-level analyzes, lab experiments, health studies, etc.

Short panel data may contain units that possess extreme values in the dependent variable and/or independent variables – respectively, *vertical outliers* (VO), *bad leverage* (BL) and *good leverage* (GL) points. These anomalies exert a disproportionate influence on least squares (LS) estimates leading to biases in the regression estimates, specifically in the estimated coefficients if BL and VO, or in the standard errors if GL (Donald and Maddala, 1993; Bramati and Croux, 2007; Verardi and Croux, 2009; Polselli, 2022). This is why it is important to identify the existence and the type of anomalous units, and how they influence the results when working with short panels.

In this paper, I develop a method to: (i) visually detect anomalous units in a panel data set, and identify their type; (ii) investigate how these units affect the LS estimates, and other units' influence on the LS estimates. I created two packages in the statistical software Stata: `xtlvr2plot` produces a leverage-versus-residual plot for panel data, and `xtinfluence` conducts the influence analysis with panel data[1].

2 Econometric Framework

Consider a linear panel regression model with fixed effects, where a dependent variable is regressed on a set of covariates and individual fixed effects. The model can be consistently estimated using OLS on the time-demeaned variables to obtain the well-known *within-group* estimator. However, it is shown that the presence of VO, BL and GL units can bias the LS coefficients and/or their standard errors.

Diagnostic plots (i.e., leverage-versus-squared residual plots) and measures of overall influence (e.g., Cook (1979)'s distance) are usually used to detect such anomalies. There are two problems arising with the available tools. First, diagnostic plots are built for cross-sectional data. Although these tools can be used even with panel data, they are complicated to implement due to the required data manipulation, and the graphical output is difficult to interpret due to the large number of displayed data points (see left plot in Figure 1). This can be overcome by using the average individual leverage of unit i at time t – a measure of the distance of the x-values of a unit from the values of other units – and the average normalised residual squared – a measure of the degree of outlyingness of a unit i. Plotting the average individual leverage over the average normalised residual squared shows the influence of each unit in the sample, and its type.

*Email: annalivia.polselli@essex.ac.uk. Post-doctoral research fellow at the Institute of Analytics and Data Science (IADS) and Centre for Micro-Social Change (MiSoc), University of Essex.

[1] The package is available at https://github.com/POLSEAN/Influence-Analysis.

Second, the popular Cook (1979)'s distance may fail to flag multiple anomalous cases in the data set because, by construction, it does not consider the mutual influence exerted by pairs of observations (Atkinson and Mulira, 1993; Chatterjee and Hadi, 1988; Rousseeuw and Van Zomeren, 1990; Rousseeuw, 1991). Pair-deletion measures can overcome this limit (Lawrance, 1995). I build on Lawrance (1995)'s approach by proposing measures for joint and conditional influence suitable for panel data models with fixed effects. Specifically,

- **Joint influence** is the influence exerted by a pair (i, j) on the LS estimates *jointly*, by comparing of the estimates *with* and *without* the pair in the sample. When i coincides with j, the measure informs on the individual influence of unit i – i.e., the Cook's distance for panel data (Banerjee and Frees, 1997; Belotti and Peracchi, 2020).

- **Joint effect** informs on unit i's influence within the (i, j) pair. For large values of the measure, i is the least influential unit of the pair, whose effect is *swamped* by the other unit, j. Thus, unit j contributes to drive most of the effect on LS estimates.

- **Conditional influence** is the influence exerted by unit i on the LS estimates *conditional on* removing unit j from the sample. In other words, it shows how the absence of j alters the influence i on the LS estimates.

- **Conditional effect** is unit i's influence *before* and *after* the deletion of unit j from the sample. For values greater than one, unit j is said to *mask* the influence of unit i. In this circumstance, the influence of i increases without j in the sample. Therefore, j contributes to drive most of the effect on the LS estimates and hides the influence of unit i.

The joint and conditional measures are constructed in a way that resembles a weighted and directed adjacency list from network analysis, displaying the existence and the strength of the links between pairs of units (i, j). Then, I can mobilize its graphical tools to analyze the relationship between unit i and unit j. This consists of plotting unit j's influence on i's.

3 The Method

The method consists of three steps that can be summarized in the points below.

1. Identify anomalous units and their type with `xtlvr2plot`. The right plot of Figure 1 is the graphical output of the command, showing the location of unit based on its leverage and normalised residual squared. GL units are located in the top-left quadrant, VO in the bottom-right quadrant, and BL units in the top-right quadrant; non-influential units are grouped in the cloud of points in the bottom-left quadrant. For example, from the right plot of Figure 1 units 10 and 40 are correctly classified as BL, units 20 and 50 as GL, and units 30 and 60 as VO, as expected from the data generating process.

2. Conduct the influence analysis with `xtinfluence`. The command generates a graph with four plots, as displayed in Figure 2. The color scale from dark blue to red shows the degree of influence/effect from the smallest to the largest value. The graphs should be read as follows:

   2.1. **Joint Influence Plot:** Identify the units with high individual influence (on the main diagonal) based on the the Cook's distance for panel data; identify pairs with high joint influence (on the off-diagonal). For example, BL and GL units (i.e., 10, 20, 40, and 50) have both high individual influences (in pink and red) and joint influences (light blue). Highly individually influential units also exert high joint influence with the rest of the units and, specifically, in correspondence of another anomaly of the same type (e.g., BL with BL units; unit 10 with 40).

2.2. Joint Effect Plot: Identify a pair with large joint effect (i.e., unit $j = \{10, 20, 40, 50\}$ with unit $i = 68$). Unit j is said to *swamp* the individual effect of unit i for large values of this measure. This occurs when the joint influence is high but the individual influence of i is small. Unit j should have already been detected at stages (1) and (2.1).

2.3. Conditional Influence Plot: Identify unit i that is influential *conditional on* removing j from the sample. The conditional influence is high for already highly individual influential units (i.e., 10, 20, 40, and 50) that, as expected, remain influential even after removing the rest of the units from the sample. These units should have already been detected in the previous steps.

2.4. Conditional Effect Plot: Identify a pair with large conditional effect (i.e., unit $j = \{10, 20, 40, 50\}$ with unit $i = 68$). Unit j is said to *mask* the effect of unit i on the LS estimates for values of the measure greater than one. In this case, the individual influence of unit i increases when j is removed from the sample. Compare the identified pairs with step (2.2). In this synthetic case, they coincide.

3. Units 10, 20, 40, and 50 that are detected in stages (1), (2.1) and (2.3) are anomalous and, specifically, BL and GL units. Stages (2.2) and (2.4) explain how their presence is altering the influence of other units in the sample (here, the influence of unit 68) and, hence, how their presence is driving the average effect on the LS estimates (i.e., by ignoring the effect of unit 68).

4 Main Findings

1. The leverage-residual plot for panel data (to the right of Figure 1) takes into account the full history of a unit, and is more informative about the existence and type of anomalous units than the plot (to the left) with each individual realisation over time.

Figure 1. *lvr2plot vs xtlvr2plot*

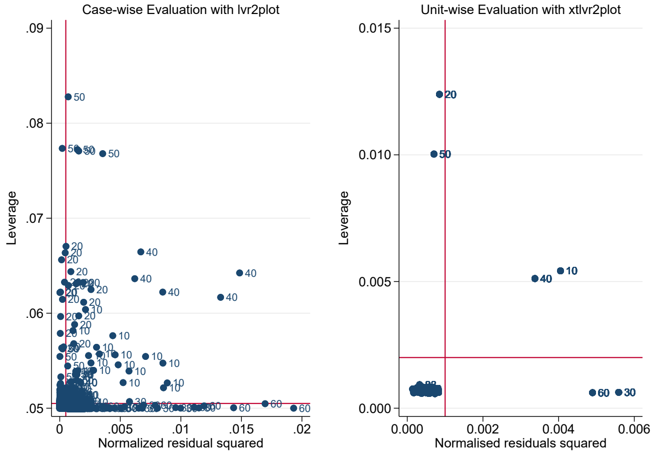
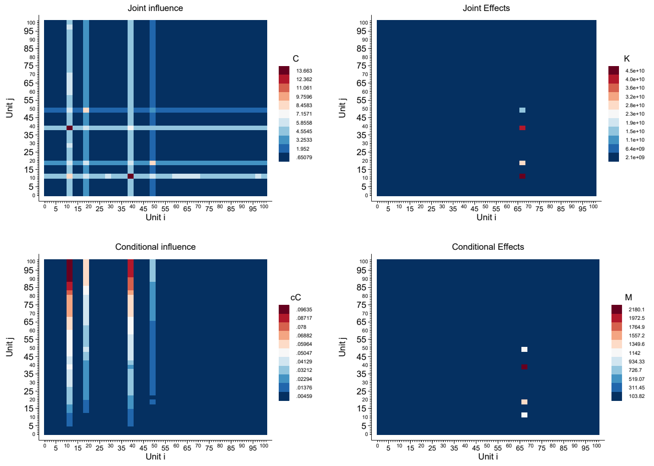


Figure 2. *xtinfluence: heat plot*



2. Joint and conditional measures are helpful in detecting GL and BL units, and showing how their presence alters the influence of other units in the sample. The strength of this method is that a unit, which is not individually influential according to Cook's distance, will always be detected if is influential jointly with, or in the absence of another highly influential unit. This will help the researcher understand how anomalous units drive their LS estimates.

5 Author's main message

In short panel data sets, the presence of anomalous units (i.e., vertical outliers, good and bad leverage units) has the potential to severely bias the least squares estimates (i.e., regression coefficients and standard errors). Available measures for their detection and classification may fail to accomplish this objective because: (i) leverage-versus-residual plots are designed for cross-sectional data, and (ii) Cook-like measures may fail to detect anomalous units in the presence of more influential ones in the data set. The method I propose overcomes these limits by taking into account the panel structure of the data, and the links between pairs of units. Once anomalous units are properly detected and identified, the researcher can deal with their presence according to the econometric literature.

References

Atkinson, A. and Mulira, H.-M. (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing*, 3(1):27–35.

Banerjee, M. and Frees, E. W. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, 92(439):999–1005.

Belotti, F. and Peracchi, F. (2020). Fast leave-one-out methods for inference, model selection, and diagnostic checking. *The Stata Journal*, 20(4):785–804.

Bramati, M. C. and Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The econometrics journal*, 10(3):521–540.

Chatterjee, S. and Hadi, A. S. (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Computational Statistics & Data Analysis*, 6(2):129–144.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Donald, S. G. and Maddala, G. (1993). 24 identifying outliers and influential observations in econometric models. In *Econometrics*, volume 11 of *Handbook of Statistics*, pages 663 – 701. Elsevier.

Lawrance, A. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):181–189.

Polselli, A. (2022). *Essays on Econometric Methods*. PhD thesis, University of Essex.

Rousseeuw, P. J. (1991). A diagnostic plot for regression outliers and leverage points. *Computational Statistics & Data Analysis*, 11(1):127–129.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639.

Verardi, V. and Croux, C. (2009). Robust regression in stata. *The Stata Journal*, 9(3):439–453.