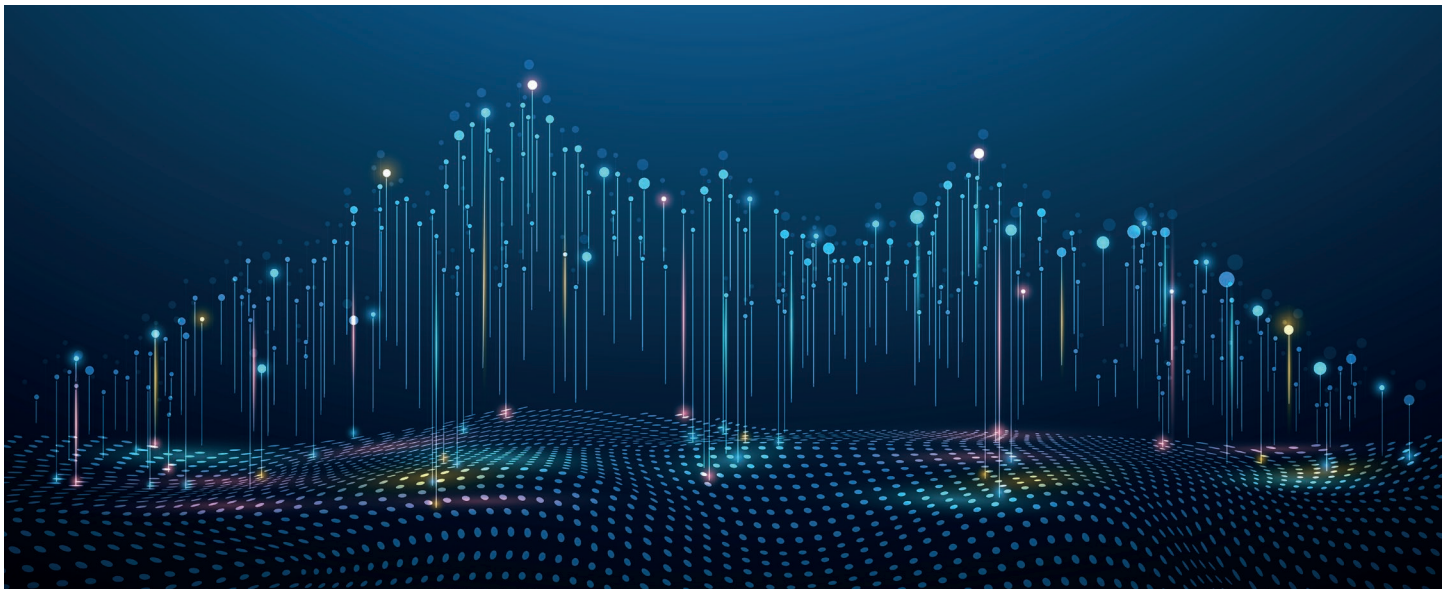
MiSoC ESRC Research Centre on Micro-Social Change

EXPLAINER



A new method to detect anomalous units with panel data and quantify their influence on the results

Annalivia Polselli

What we have developed

We have developed a new method for panel data to: (a) visually detect anomalous units in a longitudinal panel dataset and identify their type; and (b) investigate how these units affect the final results, and other units' influence.

Our new approach can help empirical researchers in the social sciences to better understand their datasets and estimation results.

Our new method is designed for panel data models with fixed effects which are commonly used by empirical researchers in the social sciences when data is collected over multiple time periods.

The method can be used *before* or *after* conducting the main regression analysis to explore the dataset and identify potential anomalies, and understand how possibly anomalous units drive their estimates.

Background

Short panel data, where the same unit is observed over multiple but small number of time periods, are widely used by applied researchers to conduct their analyses of interest. The nature of the research question or design may limit the number of observed units (e.g., number of countries, regions, or states; participants in an experiment; patients receiving a treatment; households/firms in a

Economic and Social Research Council

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

University of Essex

DOI: to follow

survey). This data structure is common in many economic fields, for example, macroeconomic country-level analyses, lab experiments, health studies, and has wide applicability.

This type of data may contain units that have extreme values in the dependent variable and/or independent variables; these are labelled *vertical outliers* (VO), *bad leverage* (BL) and *good leverage* (GL) points, respectively.

These anomalies exert a disproportionate influence on widely used estimation techniques such as the Ordinary Least Squares (OLS) estimates, leading to biased regression estimates (Donald and Maddala, 1993). For instance, BL and VO bias the estimated coefficients as shown in Figure 1, while GL bias the standard errors – even when choosing the conventional robust versions of the variance. This is why it is important to identify the presence and type of anomalous units, and how they influence the results when working with short panel data.

Existing statistical tools often used to detect such anomalies are: (a) diagnostic plots, such as leverage-vs-squared residual plots, and (b) measures of overall influence, like the Cook (1979)'s distance. There are two problems arising with these available tools.

First, most diagnostic plots are designed for cross-sectional data, where the information is collected at a single point in time. As a result, they do not account for the entire history of the units in the sample, but only assess the influence of each single realisation per unit. Consequently, it becomes challenging to evaluate the effective influence of a unit over the entire time series.

Second, the popular Cook (1979)'s distance may fail to detect 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). Lawrance (1995) shows that pair-wise deletion measures can overcome this limit for cross-sectional data, but no extension has been ever attempted for panel data.

These two limitations motivate our interested in developing of tools that address the challenges introduced by the time dimension of these type of data. This study hence proposes statistical measures that overcome the limitations of the aforementioned tools for the detection and classification of anomalous units, and introduces a method that takes into account the panel structure of the data and the links between pairs of units.

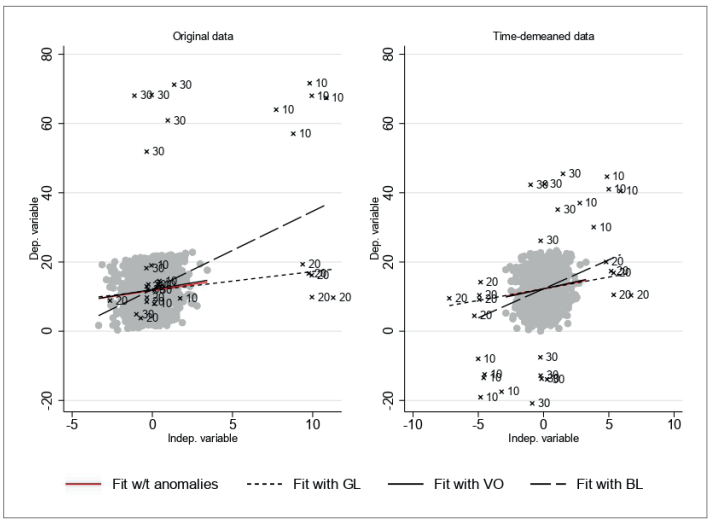
Our new method

Our new method derives suitable statistical measures for panel data that detect anomalous units and quantify their influence on other units, showing how the final estimates might be contaminated.

This new approach includes calculating these measures and using the graphical tools to investigate potential anomalies in the panel dataset and understand what and how units are driving the final results.

First, we propose two statistical measures to quantify: (a) the distance of the values of the

Figure 1 Example of anomalous units with panel data



Notes The graphs show the relationship between the dependent and independent variables, where some data points have been contaminated with good and bad leverage cases, and vertical outliers. Unit 10 is an example of bad leverage, unit 20 of a good leverage, and unit 30 of a vertical outlier. On the left is the scatter plot of the original data, while on the right is the scatter plot after the within-group transformation. The red line is fitted using uncontaminated units only; the dashed line using uncontaminated and good leverage points; the long-dashed line using uncontaminated and bad leverage points; the solid line uncontaminated units and vertical outliers.

covariates of a unit from the values of other units; and (b) the degree of 'outlyingness' of a unit (i.e., if the observations of one unit come from a different data generating process than the rest of the sample units). Plotting these two measures together on a graph (the leverage-vs-residual plot for panel data) is informative in identifying the type of anomaly (BL, GL, or VO) and inferring their potential influence on the OLS estimates.

Second, we build on Lawrance (1995)'s pair-wise deletion approach by proposing measures for quantifying the *joint and conditional influence* of units. Ploting the influence measures for the pair of units i and j informs on the existence and strength of the links between that pair of units. This graphical tool resembles a *weighted and directed adjacency* matrix from network analysis to analyse the relationship between units *i* and *j*, and it is easily interpretable.

How will this help future researchers?

The leverage-residual plot for panel data takes into account the full history of a unit, and is more informative about the existence and type of anomalous units than a plot obtained with cross-sectional measures where each individual realisation is displayed over time.

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 approach 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.

Our new method is designed for panel data models with fixed effects as the underlying algorithm uses the conventional within-group (time-demeaning) transformation to remove the effect of the fixed effects. It can be used before or after conducting the main regression analysis to explore the dataset and identify potential anomalies, and understand how possibly anomalous units drive their estimates.

The method is implemented in the statistical software Stata. The command **xtlvr2plot** produces leverage-versus-residual plots with panel data, and **xtinfluence** conducts the influence analysis with panel data. The Stata commands are currently available at our Github repository: https://github.com/POLSEAN/Influence-Analysis.

Once anomalous units are properly detected and identified, the researcher can deal with their presence according to the econometric literature. In fact, the researcher might be tempted to delete the anomalous units from the sample, but this is not always the best option because relevant information may be incorrectly discarded. For instance, the literature suggests to use of robust standard errors based on jackknife methods with GL units (MacKinnon and White, 1985; Chesher and Jewitt, 1987; MacKinnon, 2013; Belotti and Peracchi, 2020; MacKinnon et al., 2023), and robust estimation techniques, based on the estimation of the median instead of the mean, with BL and VO units (Bramati and Croux, 2007; Verardi and Croux, 2009; Aquaro and Čížek, 2013; Jiao et al., 2024, Klooster and Zhelonkin, 2024).

Read the working paper
https://arxiv.org/abs/2312.05700

Cite
Polselli, A. (2025). *A new method to detect anomalous units with panel data and quantify their influence on the results*. MiSoC Explainer (insert series reference number), University of Essex. DOI: to follow

DOI
DOI: to follow

References

- Aquaro, M. and Čížek, P. (2013). 'One-step robust estimation of fixed-effects panel data models' *Computational Statistics & Data Analysis*, 57(1):536-548
- 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.
- Chesher, A. and Jewitt, I. (1987). 'The bias of a heteroskedasticity consistent covariance matrix estimator'. E*conometrica: Journal of the Econometric Society*, pages 1217-1222.
- 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.
- Jiao, X., Pretis, F., and Schwarz, M. (2024). 'Testing for coefficient distortion due to outliers with an application to the economic impacts of climate change'. *Journal of Econometrics*, 239(1):105547.
- Klooster, J. and Zhelonkin, M. (2024). 'Outlier robust inference in the instrumental variable model with applications to causal effects'. *Journal of Applied Econometrics*, 39(1):86-106.
- Lawrance, A. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):181-189.
- MacKinnon, J. G. (2013). 'Thirty years of Heteroskedasticity-Robust Inference'. In *Recent advances and future directions in causality, prediction, and specification analysis*, pages 437-461. Springer.
- MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023a). 'Cluster-robust inference: A guide to empirical practice'. *Journal of Econometrics*, 232(2):272-299.
- MacKinnon, J. G. and White, H. (1985). 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties'. *Journal of Econometrics*, 29(3):305-325.
- Polselli, A. (2023). *Influence Analysis with Panel Data*. arXiv preprint arXiv:2312.05700.
- 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). 'Un-masking 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.