Influence Analysis with Panel Data using Stata

Annalivia Polselli

Institute for Analytics and Data Science (IADS)
University of Essex

Oceania Stata Conference 2023

February 9, 2023

Contex

- Small panel data sets with small N but larger than T
  - e.g., 50 US States, 38 OECD countries, 20 Italian regions, etc.
- Observational data may contain "anomalous" observations (Rousseeuw and Van Zomeren, 1990; Silva, 2001)
  - Exerting a disproportionate influence on the Least Squares (LS) estimates
  - Leading to biases in regression coefficients or standard errors (Donald and Maddala, 1993; Bramati and Croux, 2007; Verardi and Croux, 2009)

In this Presentation

- I present my method to
  - Visually detect and identify the type of anomalous unit
  - Understand how these affect the LS estimates
- I develop a *unit-wise* approach for the detection of anomalous units
  - As opposed to a *case-wise* (observational) approach
- The method can be conducted before or after the regression analysis

The Commands

- I propose two commands for a visual detection of anomalous units
  - `xtlvr2plot` – Leverage versus residual plot for panel data
  - `xtinfluence` – Influence analysis with panel data
- These commands can detect units that exhibit large values
  - in the outcome variable – *vertical outliers* ▸ VO
  - in the covariate space – *good leverage* points ▸ GL
  - in both directions – *bad leverage* points ▸ BL

  ▸ Plots ▸ DGP
- These commands are designed to be used with short panel data
  - e.g., cross-country macro panels, experimental panel data, health data with repeated units, etc.

Contribution

Diagnostic plots

- Leverage vs squared residual plots → `lvr2plot` and `lvr2plot2`
  - Only for cross-sectional data
  - Less handy for panel data (time-demeaned variables, case-wise visualization etc.)

Measures of overall influence

- Cook-like distances to detect anomalies
  - in cross-sectional data → `predict c, cooksd`
  - in panel data → `jackknife2, cooksd(`*newvar*`) bpd(`*newvar*`):`*command*

  - These metrics may fail to flag multiple atypical cases (Atkinson and Mulira, 1993; Chatterjee and Hadi, 1988; Rousseeuw and Van Zomeren, 1990)
    - A local approach can overcome this limit (Lawrance, 1995)

Econometric Framework

- A static linear panel regression model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}$$

- After the *within-group* (WG) transformation

$$\widetilde{y}_{it} = \widetilde{\mathbf{x}}_{it}'\boldsymbol{\beta} + \widetilde{u}_{it}$$

where $\widetilde{y}_{it} = y_{it} - T^{-1}\sum_t y_{it}$, etc.

- WG Estimator: $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{it}\widetilde{\mathbf{x}}_{it}'\right)^{-1}\sum_{i=1}^{N}\widetilde{\mathbf{x}}_{it}\widetilde{y}_{it}$
- LS Residuals: $\widehat{u}_{it} = \widetilde{y}_{it} - \widetilde{\mathbf{x}}_{it}'\widehat{\boldsymbol{\beta}}$
- Average normalised residual squared

$$\widehat{u}_i^* = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\widehat{u}_{it}}{\sqrt{\sum_i \widehat{u}_{it}^2}}\right)^2$$

Leverage

The **leverage of a unit** is a measure of the distance of the x-values of a unit from other units.

In panel data models, the individual leverage matrix

$$\underset{(T\times T)}{\mathbf{H}_{ii}} = \widetilde{\mathbf{X}}_i\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}_i' = \begin{pmatrix} h_{ii,11} & h_{ii,12} & \dots & h_{ii,1T} \\ h_{ii,21} & h_{ii,22} & \dots & h_{ii,2T} \\ \vdots & \vdots & \ddots & \vdots \\ h_{ii,T1} & h_{ii,T2} & \dots & h_{ii,TT} \end{pmatrix}$$

where $\widetilde{\mathbf{X}}_i$ is $T\times k$, and $\widetilde{\mathbf{X}}$ is $NT\times k$, with diagonal element $h_{ii,tt} = \widetilde{\mathbf{x}}_{it}'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{x}}_{it}$ and off-diagonal element $h_{ii,ts} = \widetilde{\mathbf{x}}_{it}'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{x}}_{is}$ for $t, s = 1, \dots, T$.

The **average individual leverage** of unit i at time t is

$$\overline{h}_i = \frac{1}{T}\sum_{t=1}^{T} h_{ii,tt}$$

xtlvr2plot: Syntax

xtlvr2plot – Leverage versus normalised residual squared plot for panel data.

xtlvr2plot *depvar* [*indepvar*] [*if*] [*in*] [, *options*]

options

graph_opts	graph options available for twoway scatter

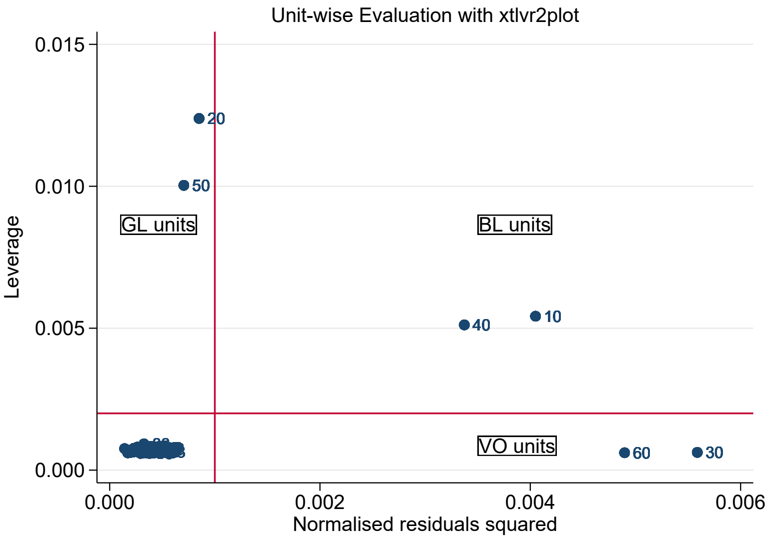
Generated variables

_lev	average individual leverage
_normres2	average individual residual squared

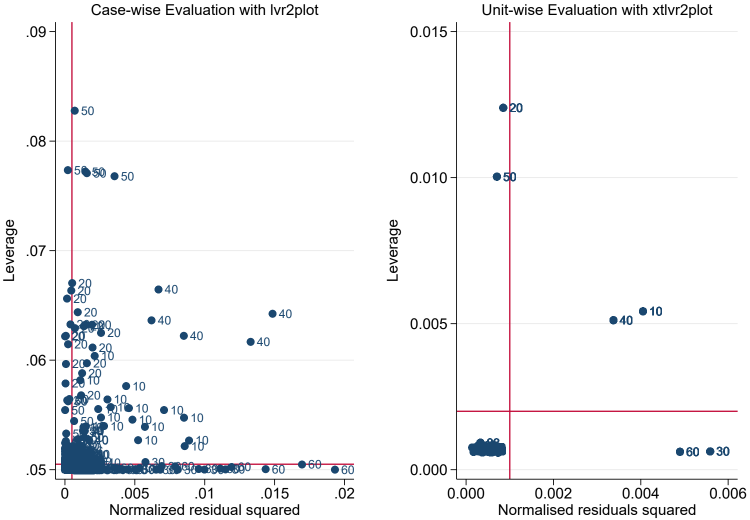
xtlvr2plot: Example

```
** Use of the 'xtlvr2plot' command
xtset id t

xtlvr2plot y x,                                    ///
    mlabel(id)                                     ///
    xlabel(, format(%9.3fc))                       ///
    ylabel(, angle(h) format(%9.3fc))              ///
    title("Unit-wise Evaluation", size(medsmall))  ///
    saving("xtlvr2plot_example.gph", replace)
```

xtlvr2plot: Plot



lvr2plot vs xtlvr2plot



xtlvr2plot: Summary Table

```
** Summary table w/detected anomalous units
** generated by 'xtlvr2plot'
```

```
            Anomalous units
-----------------------------------------
x-cutoff =    0.001
y-cutoff =    0.002
-----------------------------------------
Good leverage units
 - Count : 2
 - List  : 20 50
Bad leverage units
 - Count : 2
 - List  : 10 40
Vertical outliers
 - Count : 2
 - List  : 30 60
-----------------------------------------
```

Influence Analysis: Measures

- **Joint influence:** $\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})$
  - Influence exerted by a pair (i,j) on LS estimates *jointly*
  - Comparison of LS estimates *with* and *without* the *pair*
  - With $i = j$, $\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$ measures the individual influence of i

  Formula

- **Conditional influence:** $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})$
  - Influence exerted by i on LS estimates *conditional* on removing j from the sample
  - How the absence of j affects the influence i on LS estimates

  Formula

Influence Analysis: Measures

- **Joint influence:** $\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})
  - Influence exerted by a pair (i,j) on LS estimates *jointly*
  - Comparison of LS estimates *with* and *without* the *pair*
  - With $i = j$, $\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$ measures the individual influence of i

  Formula

- **Conditional influence:** $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})$
  - Influence exerted by i on LS estimates *conditional* on removing j from the sample
  - How the absence of j affects the influence i on LS estimates

  Formula

Influence Analysis: Measures

- **Joint influence:** $\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})$
  - Influence exerted by a pair (i,j) on LS estimates *jointly*
  - Comparison of LS estimates *with* and *without* the *pair*
  - With $i = j$, $\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$ measures the individual influence of i

  ▸ Formula

- **Conditional influence:** $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})$
  - Influence exerted by i on LS estimates *conditional* on removing j from the sample
  - How the absence of j affects the influence i on LS estimates

  ▸ Formula

Influence Analysis: Effects

- **Joint Effect**
  - $\mathrm{K}_{j|i} = \mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$
    - j *swamps* i
    - the most influential unit *swamps* the least
    - j drives the LS estimates *swamping* the effect of i
- **Conditional Effect**
  - $\mathrm{M}_{i(j)} = \mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$
    - j *masks* i
    - influence of i increases without j in the sample
    - j drives the LS estimates *masking* the effect of i

Influence Analysis: Effects

- **Joint Effect**
  - $\mathrm{K}_{j|i} = \mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$
    - j *swamps* i
    - the most influential unit *swamps* the least
    - j drives the LS estimates *swamping* the effect of i
- **Conditional Effect**
  - $\mathrm{M}_{i(j)} = \mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$
    - j *masks* i
    - influence of i increases without j in the sample
    - j drives the LS estimates *masking* the effect of i

Influence Analysis: Effects

- **Joint Effect**
  - $\mathrm{K}_{j|i} = \mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$
    - j *swamps* i
    - the most influential unit *swamps* the least
    - j drives the LS estimates *swamping* the effect of i
- **Conditional Effect**
  - $\mathrm{M}_{i(j)} = \mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$
    - j *masks* i
    - influence of i increases without j in the sample
    - j drives the LS estimates *masking* the effect of i

Influence Analysis: Effects

- **Joint Effect**
  - $\mathrm{K}_{j|i} = \mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$
    - j *swamps* i
    - the most influential unit *swamps* the least
    - j drives the LS estimates *swamping* the effect of i
- **Conditional Effect**
  - $\mathrm{M}_{i(j)} = \mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})/\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})$
    - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$
    - j *masks* i
    - influence of i increases without j in the sample
    - j drives the LS estimates *masking* the effect of i

xtinfluence: Syntax

`xtinfluence` – Influence analysis for panel data displaying the measures and effects of unit j against unit i. The size of the symbols is proportional to the magnitude of the calculated measures.

`xtinfluence` *depvar* [*indepvar*] [*if*] [*in*] [, *options*]

options

`figure(`*graphtype*`)`	display diagnostic plots like *graphtype* allows for the choice between `scatter` plot or `heat` plot; default is `scatter`
graph_opts	graph options available for `scatter` and `heatplot`
`saving(`*filename*`)`	save .dta and .pdf file with the specified name and location

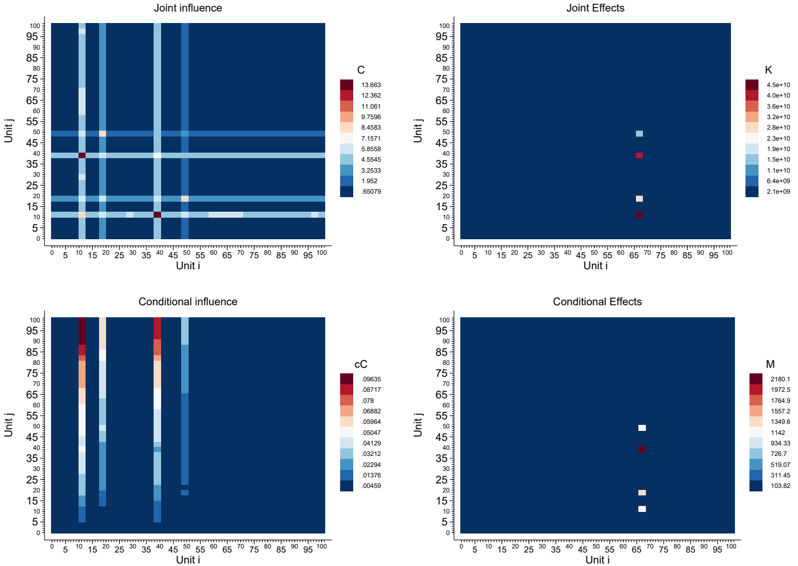
Saved data sets

filename`_adj_mtx.dta`	Data sets with the adjacency list for the influence measures and effects ▸ browse

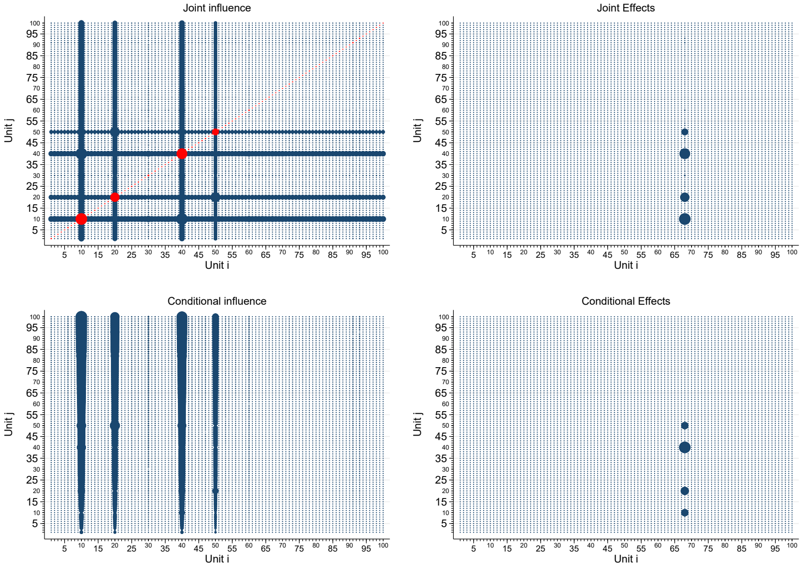
xtinfluence: Example

```
**Use of the 'xtinfluence' command
xtset id t

** Heat plot
xtinfluence y x, figure(heat)                                 ///
        keylabels(all) color(RdBu, reverse)                   ///
        xlabel(5(10)100, angle(h) labsize(small))             ///
        xmtick(##10) xmlabel(##2, angle(h))                   ///
        ylabel(5(10)100, angle(h))                            ///
        ymtick(##10) ymlabel(##2, angle(h))                   ///
        saving("xtinfluence_heat")

** Scatter plot
xtinfluence y x, figure(scatter)                              ///
        xlabel(5(10)100, angle(h) labsize(small))             ///
        xmtick(##10) xmlabel(##2, angle(h))                   ///
        ylabel(5(10)100, angle(h))                            ///
        ymtick(##10) ymlabel(##2, angle(h))                   ///
        saving("xtinfluence_scatter")
```

xtinfluence: heat plot



xtinfluence: scatter plot



`xtinfluence`: Summary Table

```
** Output table generated by 'xtinfluence'
```

```
              Influence analysis
---------------------------------------------
v1 = k+1 =  2
v2 = NT-N-k-1 = 1898
c1 = 4/N = .04
c2 = F(v1,v2,.5) =   0.6934
---------------------------------------------
Cii >= c1
 - Count : 5
 - List  : 10 20 30 40 50
Cii >= c2
 - Count : 4
 - List  : 10 20 40 50
i with K >= p99
 - Count : 6
 - List  : 7 33 44 63 68 88
i with M >= 1
 - Count : 3
 - List  : 44 63 68
---------------------------------------------
```

Summary of Method

1. Identify anomalous units and their type with `xtlvr2plot`
2. Conduct the influence analysis with `xtinfluence`
   2.1 **Joint Influence Plot**
   - Identify units with high individual influence (main diagonal)
   - Identify pairs with high joint influence (off-diagonal)
   - Highly influential units swamp all other units

   2.2 **Joint Effect Plot**
   - Identify pairs with largest effect
   - j swamps the effect of i
   - j must be detected in (1) and (2.1)

   2.3 **Conditional Influence Plot**
   - Identify influential i conditional to removing j
   - Check if same units as (1) and (2.1)

   2.4 **Conditional Effect Plot**
   - Identify pairs with largest effect
   - j masks the effect of i
   - Compare identified pairs with (2.2)
3. Units detected in (1), (2.1) and (2.3) are anomalous; (2.2) and (2.4) explain how they affect the influence of other units and, hence, LS estimates

How to treat anomalous units?

Once identified the type of anomaly in the sample,

1. Is it an actual error in the entry of the data?
   - Deal with measurement error
2. Is it a genuine extreme value in the entry of the data?
   - Robust estimation techniques if VO and BL units (Bramati and Croux, 2007; Verardi and Croux, 2009; Aquaro and Čížek, 2013, 2014; Jiao, 2022)
   - Jackknife-type standard errors if GL units (MacKinnon and White, 1985; Davidson et al., 1993; MacKinnon, 2013; Belotti and Peracchi, 2020; Polselli, 2022)

Thank you for your attention!

✉ annalivia.polselli[at]essex.ac.uk

https://github.com/POLSEAN

References I

Aquaro, M. and Čížek, P. (2013). One-step robust estimation of fixed-effects panel data models. *Computational Statistics & Data Analysis*, 57(1):536–548.

Aquaro, M. and Čížek, P. (2014). Robust estimation of dynamic fixed-effects panel data models. *Statistical Papers*, 55(1):169–186.

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.

Davidson, R., MacKinnon, J. G., et al. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

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. (2022). A simple robust procedure in instrumental variables regression.

References II

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. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325.

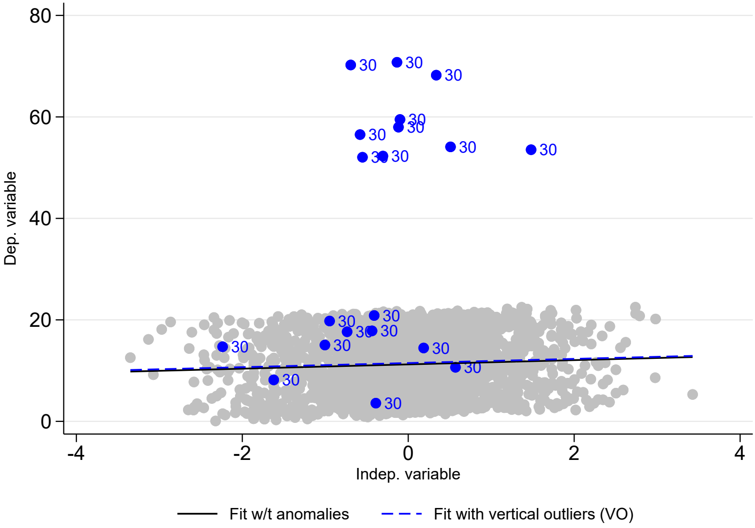
Polselli, A. (2022). *Essays on Econometric Methods*. PhD thesis, University of Essex.

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.

Silva, J. S. (2001). Influence diagnostics and estimation algorithms for powell's scls. *Journal of Business & Economic Statistics*, 19(1):55–62.

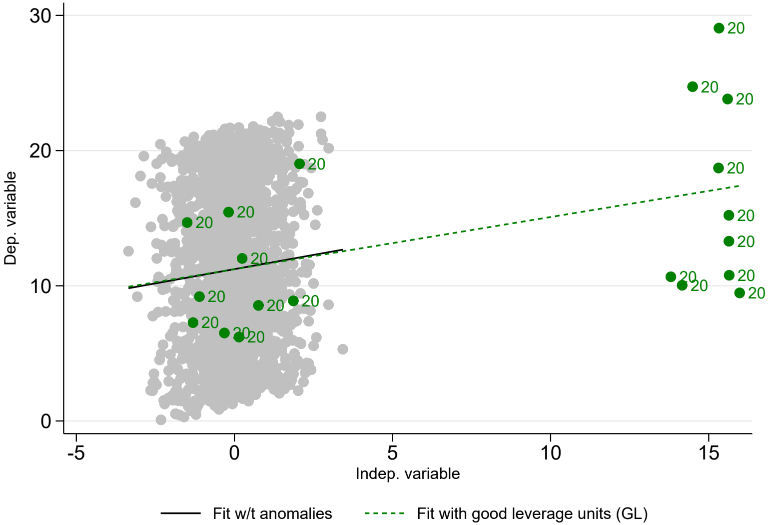
Verardi, V. and Croux, C. (2009). Robust regression in stata. *The Stata Journal*, 9(3):439–453.

Vertical outliers ▸ Back

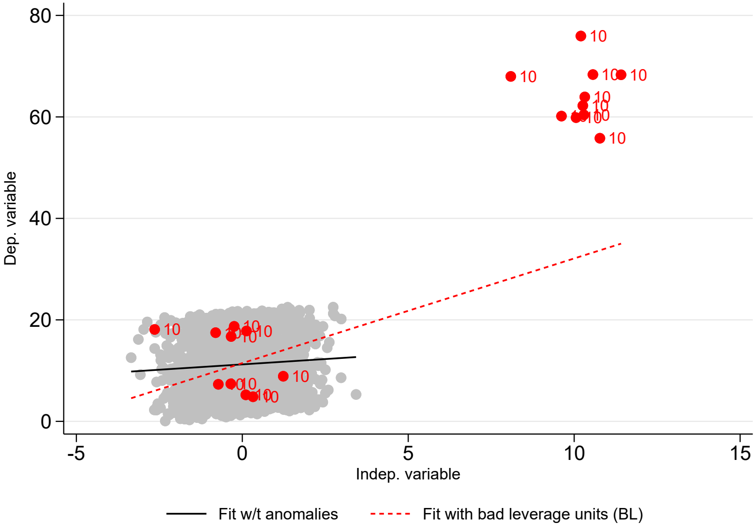


Good leverage units

Back

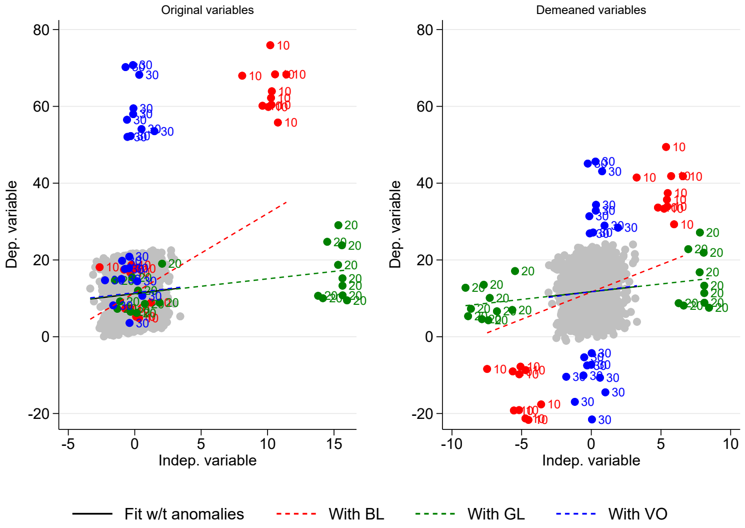


Bad leverage units ▸ Back



Anomalous units after time-demeaning

▸ Back



Directed Weighted Adjacency List

Back

	i	j	C	K	cC	M
1	1	1	.0318985	1	0	0
2	1	2	.0779802	2.444638	8.05e-06	.0002523
3	1	3	.0379366	1.189292	.000065	.0020391
4	1	4	.0812006	2.545595	.0000804	.0025191
5	1	5	.0384888	1.206603	.0000916	.0028703
6	1	6	.0619195	1.941144	.000091	.0028528
7	1	7	.0802803	2.516744	.0001116	.0034988
8	1	8	.0322271	1.010302	.0001236	.003874
9	1	9	.0102966	.3227937	.0001144	.0035852
10	1	10	34.86443	1092.981	.0001167	.0036569
11	1	11	.0380862	1.193983	.0001264	.0039615
12	1	12	.0524164	1.643225	.0001519	.0047621
13	1	13	.0510088	1.599099	.0001667	.005226
14	1	14	.0550416	1.725525	.0001834	.0057488
15	1	15	.0617752	1.936618	.0001679	.0052648
16	1	16	.0591808	1.855285	.000202	.0063336
17	1	17	.0512263	1.605917	.0001969	.0061739
18	1	18	.067513	2.116496	.0002049	.006424
19	1	19	.0904264	2.834818	.000237	.0074296
20	1	20	11.59427	363.474	.0005592	.0175295
21	1	21	.0564583	1.769938	.0002562	.0080332
22	1	22	.0020566	.0644732	.0002375	.0074454
23	1	23	.091529	2.869384	.0002585	.0081049
24	1	24	.026083	.8176892	.0002669	.0083674
25	1	25	.0945991	2.965631	.0003046	.0095503

Joint Influence Back

If $i \neq j$,

$$\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}}) = \big(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i,j)}\big)'\big(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\big)\big(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i,j)}\big)(s^2K)^{-1}$$

where

$$\widehat{\boldsymbol{\beta}}_{(i,j)} = \widehat{\boldsymbol{\beta}}_{(i)} - \big(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\big)^{-1}\big(\widetilde{\mathbf{X}}_i'\mathbf{M}_i^{-1}\mathbf{H}_{ij} + \widetilde{\mathbf{X}}_j'\big)\big(\mathbf{M}_j - \mathbf{H}_{ij}'\mathbf{M}_i^{-1}\mathbf{H}_{ij}\big)^{-1}\big(\mathbf{H}_{ij}'\mathbf{M}_i^{-1}\widehat{\mathbf{u}}_i + \widehat{\mathbf{u}}_j\big)$$

with $\mathbf{M}_j = \mathbf{I}_j - \mathbf{H}_j$ with $\mathbf{H}_{ij} = \widetilde{\mathbf{X}}_i(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}_j'$, and $\mathbf{H}_j = \widetilde{\mathbf{X}}_j(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}_j'$.
Note that $\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}}) = \mathrm{C}_{ji}(\widehat{\boldsymbol{\beta}})$.

If $i = j$,

$$\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}}) = \big(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\big)'\big(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\big)\big(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\big)(s^2K)^{-1}$$

where $\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \big(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\big)^{-1}\widetilde{\mathbf{X}}_i'\mathbf{M}_i^{-1}\widehat{\mathbf{u}}_i$.

This is Banerjee and Frees (1997) metrics as defined by Belotti and Peracchi (2020) for linear panel data models with fixed effects.

Both measures are distributed as $\mathrm{F}(\nu_1, \nu_2)$; a distributional cutoff can be chosen.

Conditional Influence ▸ Back

$$\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}}) = \left(\widehat{\boldsymbol{\beta}}_{(i,j)} - \widehat{\boldsymbol{\beta}}_{(j)}\right)' \left(\sum_{\substack{i=1 \\ i \neq j}}^{N} \widetilde{\mathbf{X}}'_{i(j)} \widetilde{\mathbf{X}}_{i(j)} \right) \left(\widehat{\boldsymbol{\beta}}_{(i,j)} - \widehat{\boldsymbol{\beta}}_{(j)}\right) (s^2 K)^{-1}$$

- $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}}) = 0$ for $i = j$
- $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}}) \neq \mathrm{C}_{j(i)}(\widehat{\boldsymbol{\beta}})$
- $\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}}) \approx F(\nu_1, \nu_2)$ from which a distributional cutoff can be chosen

Data generating process

Back

```
set seed 1408
set obs 100
gen id = _n
expand 20

bys id: generate t = _n
bys id: gen x = rnormal()

bys id: replace x = rnormal(10,1) if id==10 & t<=10  //BL unit
bys id: replace x = rnormal(10,1) if id==40 & t<=5   //BL unit

bys id: replace x = rnormal(15,1) if id==20 & t<=10  //GL unit
bys id: replace x = rnormal(15,1) if id==50 & t<=5   //GL unit

bys id: gen a = runiform(0,20)
bys id: gen y = 1 + 1*x + a + runiform()

bys id: replace y = y + rnormal(50,1) if id==10 & t<=10  //BL unit
bys id: replace y = y + rnormal(50,1) if id==40 & t<=5   //BL unit

bys id: replace y = y + rnormal(50,1) if id==30 & t<=10  //VO
bys id: replace y = y + rnormal(50,1) if id==60 & t<=5   //VO
```