Influence Analysis with Panel Data using Stata

Annalivia Polselli

Institute for Analytics and Data Science
University of Essex

UK Stata Conference

September 7, 2023

Motivation

- Short panel data sets (small N but $N \gg T$) are common in many fields of Economics
  - Macro-level panel data (e.g., 50 US States)
  - Cell-group data (e.g., gender-age-occupation)
  - Experimental panel data (e.g., limited no. participants)
- Observational data may contain "anomalous" observations (Rousseeuw and Van Zomeren, 1990; Silva, 2001)
  - Vertical outliers (VO), good leverage (GL) points, bad leverage (BL) points Example DGP
- Large influence on the Least Squares (LS) estimates $\Longrightarrow$ Biased regression coefficients or standard errors (Donald and Maddala, 1993; Bramati and Croux, 2007; Verardi and Croux, 2009)

Motivation

- **Diagnostic plots** (leverage-vs-residual plots)
  - for cross-sectional data: `lvr2plot`/`lvr2plot2`
  - Less handy for panel data
- **Measures of influence** (Cook (1979)'s distance)
  - for cross-sectional data:
    `predict c, cooksd`
  - for 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) unlike *pair-wise measures* (Lawrance, 1995)

In this talk

- I present a method to
  1. **Detect** anomalous units and **identify** their type
  2. **Show** how these affect the LS estimates
- I follow a *unit-wise* approach (full history of a unit)
- I develop two commands in Stata
  - `xtlvr2plot` – Leverage-vs-residual plot for panel data
  - `xtinfluence` – Pair-wise influence measures with panel data
- I apply the method to a cross-country study
  - Berka et al. (2018, AER)

Model and estimator

Static linear panel regression model with fixed effects

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

Model 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., and $\boldsymbol{\beta}$ is a vector of parameters.

The WG Estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{it}\widetilde{\mathbf{x}}_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{it}\widetilde{y}_{it}$$

Overview

1. Identify anomalous units with `xtlvr2plot`

2. Understand how anomalous units may affect the LS estimates with `xtinfluence`

    2.1 Joint influence and joint effect

    2.2 Conditional influence and conditional effect

xtlvr2plot: Syntax

xtlvr2plot – Leverage-versus-normalised residual squared plot for panel data.
xtset 'panelvar' 'timevar' is required.

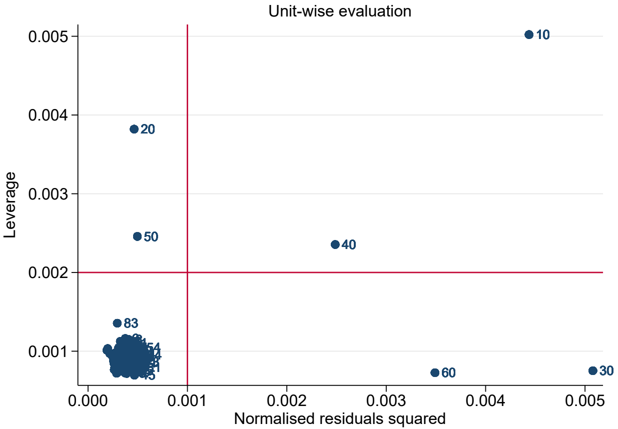
xtlvr2plot *depvar* [*indepvar*] [*if*] [*in*] [, *options*]

options	
graph_opts	graph options allowed for twoway scatter
Generated variables	
_lev	average individual leverage
_normres2	average individual residual squared

xtlvr2plot: Example

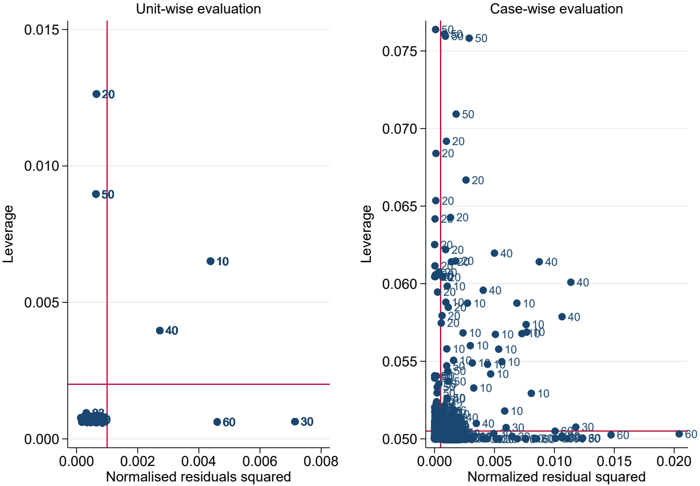
```
** Use of the 'xtlvr2plot' command
xtset id time

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`: Leverage-vs-residual plot



Note: Units 10 and 40 are set to be bad leverage units; units 20 and 50 good leverage units; units 30 and 60 vertical outliers. Formulae

xtlvr2plot vs lvr2plot



Note: Units 10 and 40 are set to be bad leverage units; units 20 and 50 good leverage units; units 30 and 60 vertical outliers.

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
--------------------------------------------------
```

Note: Units 10 and 40 are set to be bad leverage units; units 20 and 50 good leverage units; units 30 and 60 vertical outliers.

Overview

1. Identify anomalous units with `xtlvr2plot`

2. Understand how anomalous units may affect the LS estimates with `xtinfluence`

    2.1 Joint influence and joint effect

    2.2 Conditional influence and conditional effect

Joint measures

- For $i \neq j$, joint influence is

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

where $\widehat{\boldsymbol{\beta}}_{(i,j)}$ is WG estimator w/t units i and j, s is RMSE, K is no. covariates

  - Influence exerted by a pair (i,j) on LS estimates *jointly*
  - Comparison of LS estimates *with* and *without* the *pair*
  - For $i = j$, i's individual influence (as in Belotti and Peracchi (2020))

- The joint effect is

$$\mathrm{K}_{j|i} = \frac{\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})}{\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})}$$

  - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$, j *alters* the effect of i
    - j either enhances or reduces the effect of i on the LS estimates, based on the conditional effect

Joint measures

- For $i \neq j$, joint influence is

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

where $\widehat{\boldsymbol{\beta}}_{(i,j)}$ is WG estimator w/t units i and j, s is RMSE, K is no. covariates

  - Influence exerted by a pair (i,j) on LS estimates *jointly*
  - Comparison of LS estimates *with* and *without* the *pair*
  - For $i = j$, i's individual influence (as in Belotti and Peracchi (2020))

- The joint effect is

$$\mathrm{K}_{j|i} = \frac{\mathrm{C}_{ij}(\widehat{\boldsymbol{\beta}})}{\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})}$$

  - How much the pair is influential wrt i
  - For large values of $\mathrm{K}_{j|i}$, j *alters* the effect of i
    - j either enhances or reduces the effect of i on the LS estimates, based on the conditional effect

Conditional measures

- For $i \neq j$, conditional influence is

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

  - Influence exerted by i on LS estimates without j in the sample
  - How the absence of j affects the influence i on LS estimates

- The conditional effect is

$$\mathrm{M}_{i(j)} = \frac{\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})}{\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})}$$

  - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$, influence of i increases without j in the sample $\implies j$ *masks* i.
  - If $\mathrm{M}_{i(j)} < 1$, influence of i decreases without j in the sample $\implies j$ *boosts* i

Conditional measures

- For $i \neq j$, conditional influence is

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

  - Influence exerted by i on LS estimates without j in the sample
  - How the absence of j affects the influence i on LS estimates

- The conditional effect is

$$\mathrm{M}_{i(j)} = \frac{\mathrm{C}_{i(j)}(\widehat{\boldsymbol{\beta}})}{\mathrm{C}_{ii}(\widehat{\boldsymbol{\beta}})}$$

  - How influence of i changes before and after the deletion of j
  - If $\mathrm{M}_{i(j)} \geq 1$, influence of i increases without j in the sample $\Longrightarrow$ j *masks* i.
  - If $\mathrm{M}_{i(j)} < 1$, influence of i decreases without j in the sample $\Longrightarrow$ j *boosts* i

xtinfluence: Syntax

xtinfluence – Calculates and displays joint and conditional measures/effects of pairs of units i and j. The size of the symbols is proportional to the magnitude of the calculated measures. xtset 'panelvar' 'timevar' is required.

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

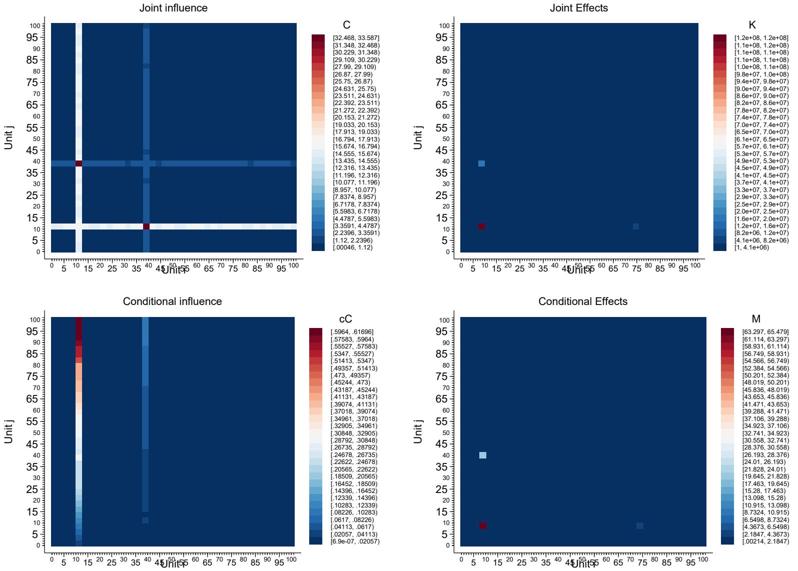
options

figure(*graphtype*)	displays diagnostic plots as *graphtype*. Allowed *graphtype* are scatter plot or heat plot; default is scatter
graph_opts	graph options allowed for scatter and heatplot
saving(*filename*)	saves .dta and .pdf file with the specified name and location
Generated variables	
_newid	assigns a new numeric identifier to sorted 'panelvar'
Saved data sets	
*filename*_adj_mtx.dta	Automatically saves a data set with the influence measures and effects generated by the command

xtinfluence: Example

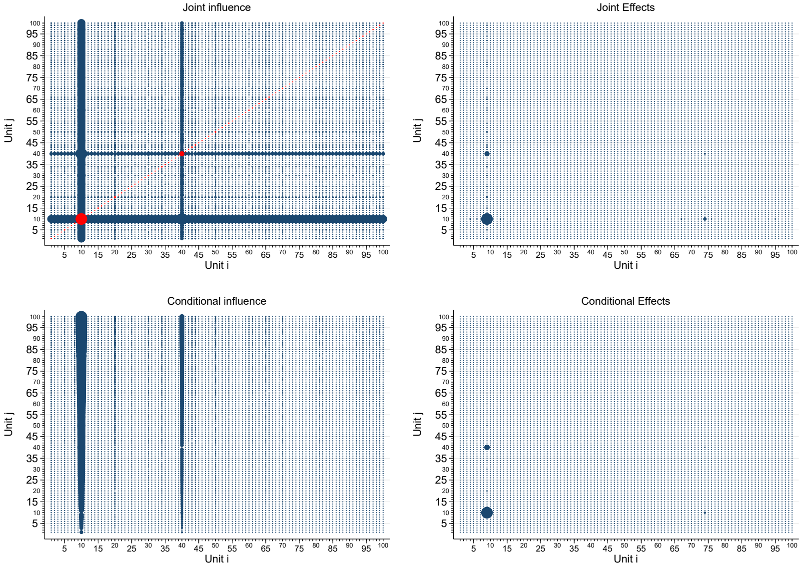
```
**Use of the 'xtinfluence' command
xtset id t

** Heat plot
xtinfluence y x, figure(heat)                                  ///
        keylabels(all, interval) color(RdBu, reverse)          ///
        lev(30) statistic(max)                                 ///
        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: Plot



Note: Units 10 and 40 are bad leverage units; units 20 and 50 good leverage units; units 30 and 60 vertical outliers. ▸ Adj-mtx

xtinfluence: Plot



Note: Units 10 and 40 are bad leverage units; units 20 and 50 good leverage units; units 30 and 60 vertical outliers. ▸ Adj-mtx

xtinfluence: Summary table

Variable	Obs	Mean	Std. dev.	Min	Max
C	10,000	.3811386	2.200585	2.35e-11	33.58732
K	10,000	16156.08	1242556	4.42e-08	1.23e+08
cC	10,000	.0038312	.0353837	0	.6169614
M	9,900	.0305928	.6922132	4.39e-06	65.47916

```
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 : 8
 - List  : 8 10 20 34 40 43 50 65
Cii >= c2
 - Count : 2
 - List  : 10 40
i with K >= p99
 - Count : 30
 - List  : 3 4 6 9 11 13 14 19 24 27 47 49 55 57 62 64 67 68 69 71 72 74 76 77 79 84 86 89 93 95
j with K >= p99
 - Count :
 - List  :
i with M >= 1
 - Count : 2
 - List  : 9 74
j with M >= 1
 - Count : 2
 - List  : 10 40
```

Empirical example

- Berka et al. (2018, AER) – Macro-panel data
  - Objective: Study relationship between real exchange rate and sectoral productivity in the Eurozone
  - Units of observation: 9 EU countries
  - Time period: 1995–2007

▸ Overview ▸ Scatter ▸ xtlvr2plot ▸ xtinfluence ▸ Summary

Conclusion

- The proposed STATA commands allow to systematically
  1. Identify anomalous units and their type (unit-wise leverage-vs-residual plot)
  2. Investigate how anomalous units may affect the LS estimates (joint and conditional influence and effects)

- Once the type of anomaly is identified, the literature suggests, e.g.,
  1. Methods for measurement error if error in the data entry
  2. Robust estimation techniques if VO and BL (Bramati and Croux, 2007; Verardi and Croux, 2009; Aquaro and Čížek, 2013, 2014; Jiao, 2022)
  3. Jackknife-type standard errors if GL (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/Influence-Analysis

@AnnalivPolselli

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.

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.

Berka, M., Devereux, M. B., and Engel, C. (2018). Real exchange rates and sectoral productivity in the eurozone. *American Economic Review*, 108(6):1543–81.

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.

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

References II

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. Unpublished, Last accessed: 07/02/2023.

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.

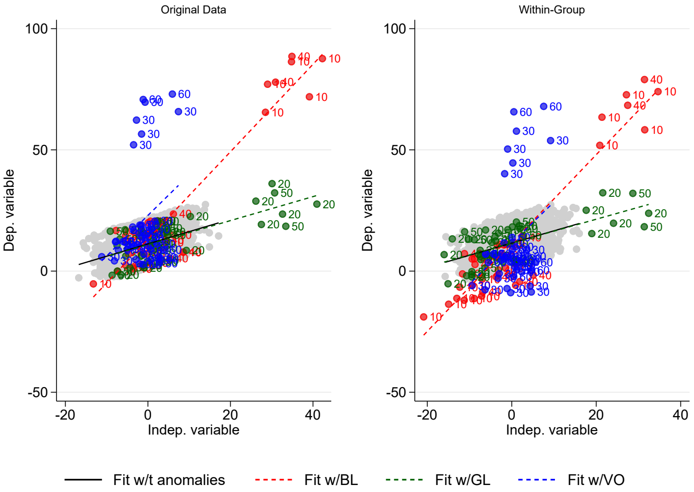
Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

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.

Scatter Plot DGP

▸ Back



Note: Units 10 and 40 are bad leverage units; units 20 and 50 are good leverage units; units 30 and 60 are vertical outliers.

DGP ▸ Back

```
loc numobs 100
set obs 100
gen id = _n
expand 20

bys id: generate t = _n
bys id: gen z = rnormal(0,5)
**GL
bys id: replace z = z + rnormal(30,1) if id==20 & t<=5
bys id: replace z = z + rnormal(30,1) if id==50 & t<=2
**for BL
bys id: replace z = z + rnormal(30,1) if id==10 & t<=5
bys id: replace z = z + rnormal(30,1) if id==40 & t<=2
**line
bys id: gen a = runiform(0,20)
bys id: gen y = 1 + .5*z + a + runiform()
**BL
bys id: replace y = y + rnormal(50,1) if id==10 & t<=5
bys id: replace y = y + rnormal(50,1) if id==40 & t<=2
*VO
bys id: replace y = y + rnormal(50,1) if id==30 & t<=5
bys id: replace y = y + rnormal(50,1) if id==60 & t<=2
```

Example: Berka et al. (2018) Back

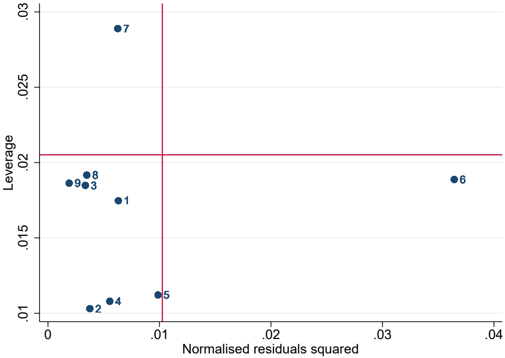
- They study relationship between real exchange rate and sectoral productivity in the Eurozone
- Regression model:

$$RER_{it} = \beta TFP_{it} + \mathbf{x}'_{it}\boldsymbol{\gamma} + \alpha_i + u_{it}$$

RER_{it}: real exchange rate in log
TFP_{it}: total factor productivity in log
$\mathbf{x}_{it}$: other controls
α_i: country fixed effects

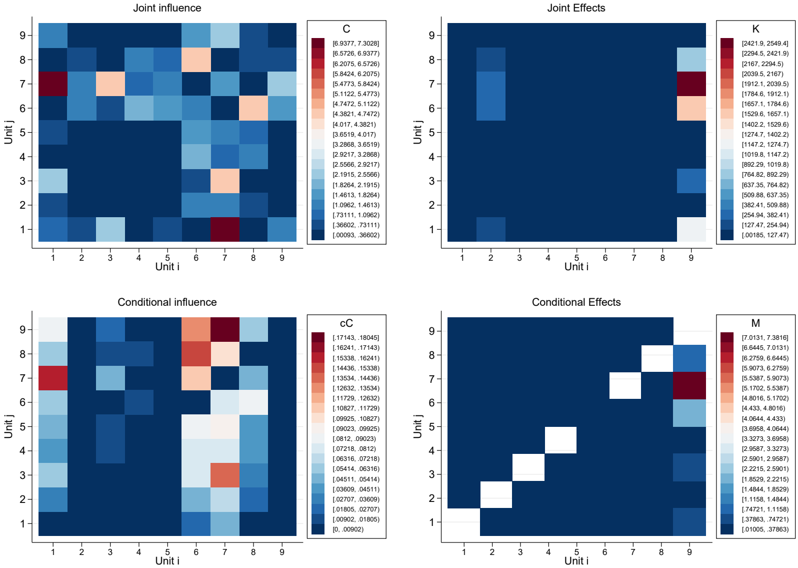
- Finding strong correlation between TFP and RER among high-income countries with floating nominal exchange rates
- Sample: 9 EU countries
- Time Period: 1995–2007
- Table 4, specification (2a)

Example: Leverage-vs-residual plot ▸ Back



Note: 1-Austria, 2-Belgium, 3-Finland, 4-France, 5-Germany, 6-Ireland, 7-Italy, 8-Netherlands, 9-Spain.

Example: Network-like plots ▸ Back



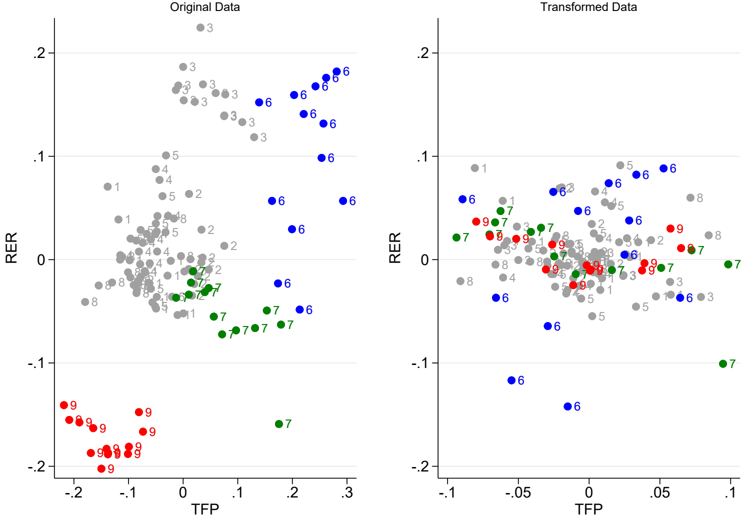
Example: Summary

Back

Variable	Obs	Mean	Std. dev.	Min	Max
C	81	1.0233	1.472976	.0009253	7.30281
K	81	97.87085	368.2484	.0018538	2549.404
cC	81	.032125	.0439157	0	.1804506
M	72	.2303033	.8915019	.0046645	7.381636

```
                Influence analysis

v1 = k+1 =  2
v2 = NT-N-k-1 = 184
c1 = 4/N = .4444444444444444
c2 = F(v1,v2,.5) =   0.6958

Cii >= c1
 - Count : 4
 - List  : 1 6 7 8
Cii >= c2
 - Count : 3
 - List  : 1 6 7
i with K >= p99
 - Count : 1
 - List  : 9
j with K >= p99
 - Count :
 - List  :
i with M >= 1
 - Count : 1
 - List  : 9
j with M >= 1
 - Count : 2
 - List  : 6 7
```

Example: Scatter ‣ Back



Note: 1-Austria, 2-Belgium, 3-Finland, 4-France, 5-Germany, 6-Ireland, 7-Italy, 8-Netherlands, 9-Spain.

*filename*_adj_mtx.dta ▸ Back

The saved data set resembles a directed and weighted adjacency list

	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

Residual and Leverage

Back

- The **average normalised residual** squared

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

where $\widehat{u}_{it} = \widetilde{y}_{it} - \widetilde{\mathbf{x}}_{it}'\widehat{\boldsymbol{\beta}}$ are LS Residuals.

Cut-off value: $c_{\widehat{u}_i^*} = \frac{2}{NT}$

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

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

where $h_{ii,tt} = \widetilde{\mathbf{x}}_{it}'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{x}}_{it}$, and $h_{ii,ts} = \widetilde{\mathbf{x}}_{it}'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{x}}_{is}$ for $t, s = 1, \ldots, T$.

Cut-off value: $c_{\overline{h}_i} = \frac{2(K+1)}{NT}$

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