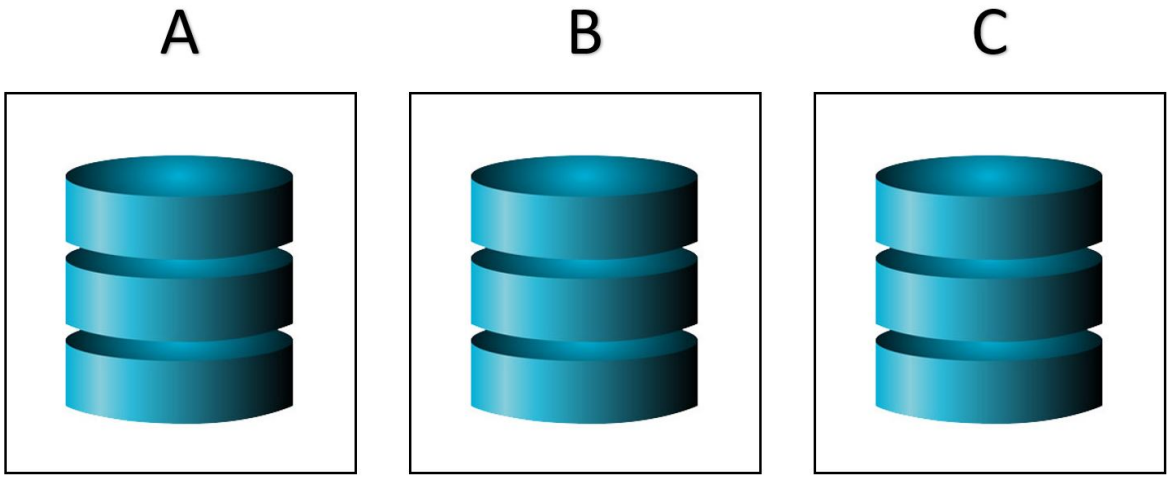
GMM across Unmerged Data Sets

Eric Bartelsman[1 2] & Richard Bräuer[2 3]

[1]Tinbergen Institute [2]VU Amsterdam [3]Halle Institute of Economic Research

October 8, 2019

Problem



- Researchers often face confidential data they are not allowed to access without anonymization
  - Patient Data
  - Employer-employee Data
  - Individual Firm Data
  - ...

Current Solutions

Micro Moments Databases

- Collect nonconfidential group info (Bartelsman & Barnes 2004; Bartelsman, Hagsten & Polder 2018)
- Allow analysis at level between micro and macro
- Observations correspond to groups of individual firms

Two sample IV

- Estimation with two different samples describing the same process (Angrist & Krueger 1992)
- Instruments $\overrightarrow{Z}$ and outcomes $\overrightarrow{Y}$ in sample 1
- End. variables $\overrightarrow{X}$ and instruments $\overrightarrow{Z}$ in sample 2

New estimator:

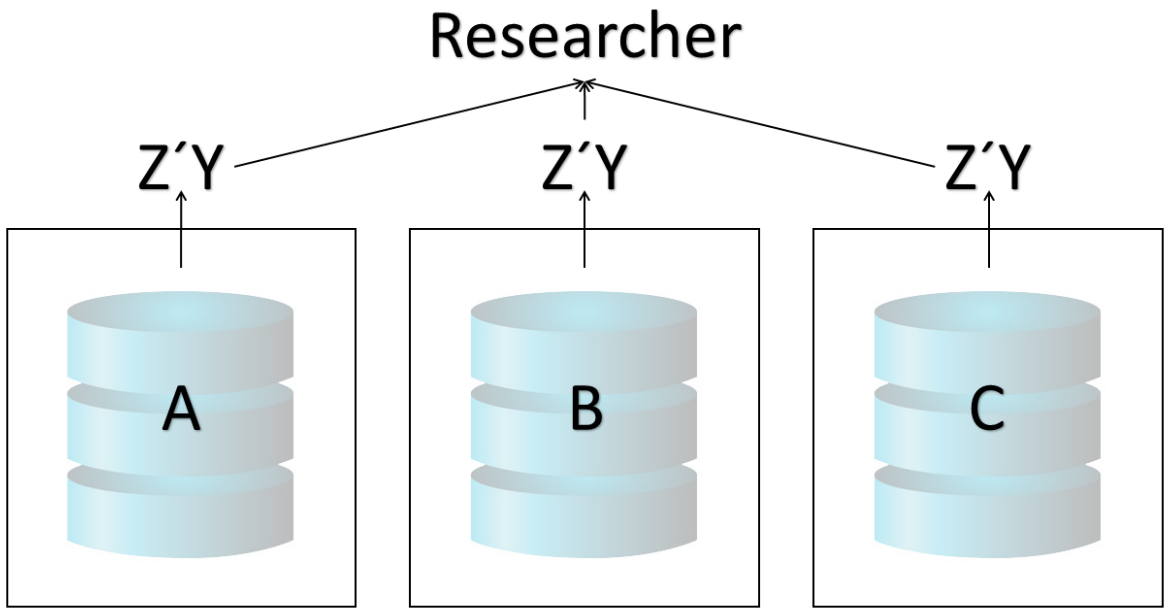
Arbitrary linear GMM spanning multiple unmerged data silos

Example: CompNet

- CompNet collects micro moments from representative firm data
- **However:** Some questions are hard to answer with moments
  - What effects do EU subsidies have on the targeted firms?
  - What is the return of investment on R&D for the firm?
  - What is the cross-country production function in a sector?
  - ...
- Firm level regression would be most straightforward

An Estimator for Cross-Data-Set Regressions

- Data split in several sets with $\boldsymbol{Z}$,$\boldsymbol{X}$ & $\boldsymbol{Y}$ (CompNet)
- $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$



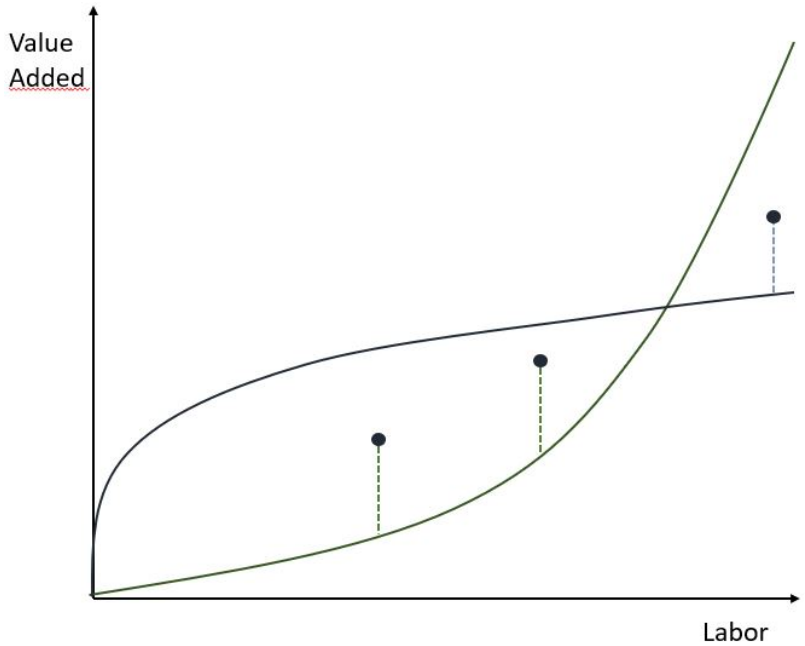
An Estimator for Cross-Data-Set Regressions

- Data split in several sets with $\boldsymbol{Z}$,$\boldsymbol{X}$ & $\boldsymbol{Y}$ (CompNet)
- $\hat{\beta} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$

$$(\boldsymbol{Z}'\boldsymbol{Y}) = \begin{bmatrix} \sum_i^N (1 * y_i) \\ \sum_i^N (z_i^1 * y_i) \\ \sum_i^N (z_i^2 * y_i) \\ \dots \end{bmatrix} = \underbrace{\begin{bmatrix} \sum_i^K (1 * y_i) \\ \sum_i^K (z_i^1 * y_i) \\ \sum_i^K (z_i^2 * y_i) \\ \dots \end{bmatrix}}_{\text{Data Source A}} + \underbrace{\begin{bmatrix} \sum_K^N (1 * y_i) \\ \sum_K^N (z_i^1 * y_i) \\ \sum_K^N (z_i^2 * y_i) \\ \dots \end{bmatrix}}_{\text{Data Source B}} \quad (1)$$

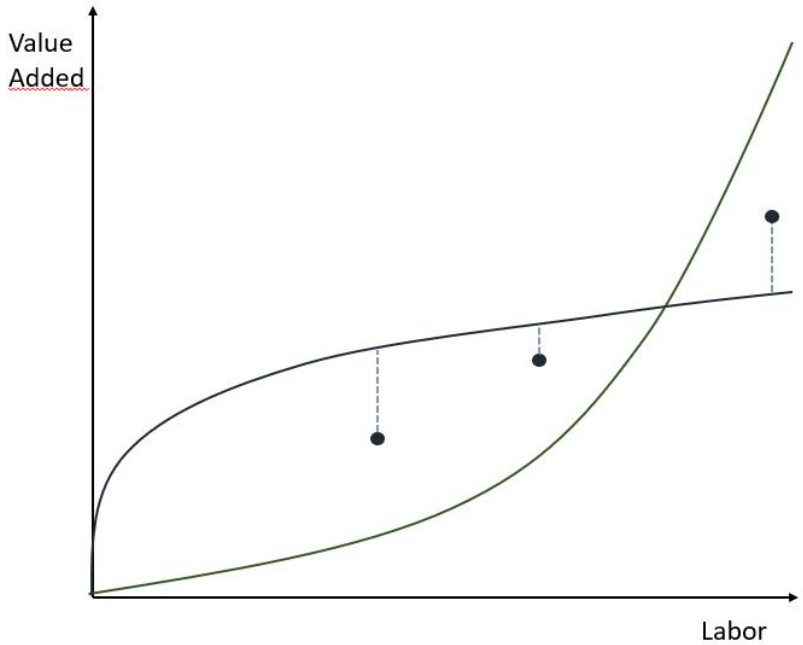
Cross-country-TFP-estimation

- TFP only comparable if derived from the same production function



Cross-country-TFP-estimation

- TFP only comparable if derived from the same production function



- Current round contains first cross-country comparable TFP from an estimated production function (experimental)

An Estimator for Cross-Data-Set Regressions

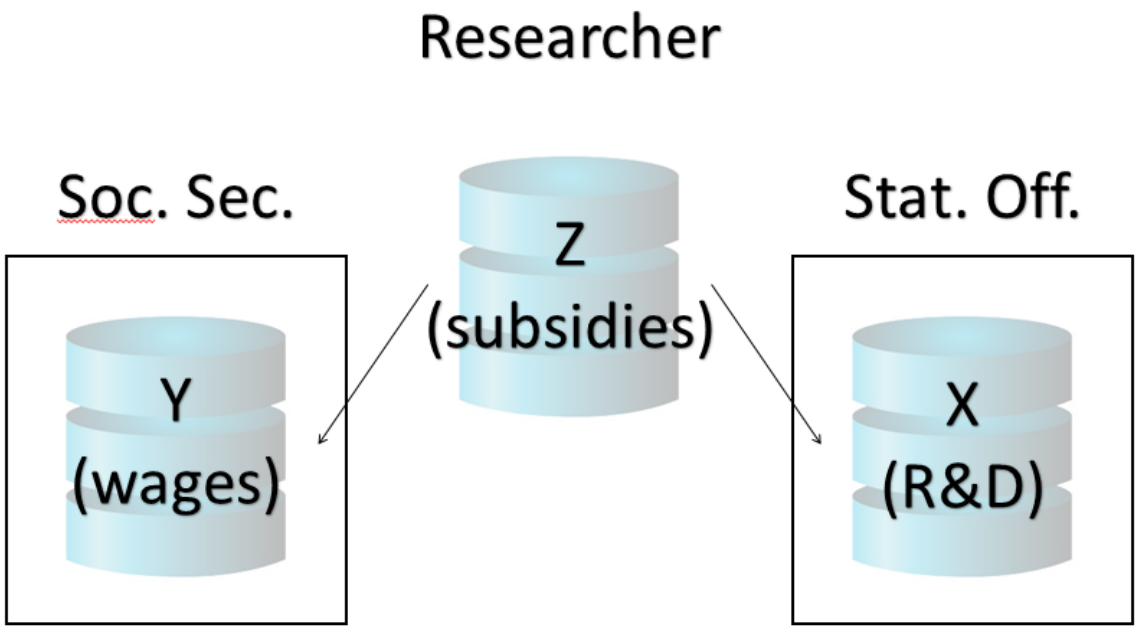
- Data split with $\boldsymbol{Y}$ & $\boldsymbol{X}$ in different data sets
- $\hat{\beta} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$

$$(\boldsymbol{Z}'\boldsymbol{Y}) = \begin{bmatrix} \sum_i^N (1 * y_i) \\ \sum_i^N (z_i^1 * y_i) \\ \sum_i^N (z_i^2 * y_i) \\ \cdots \end{bmatrix}$$

$$(\boldsymbol{Z}'\boldsymbol{X}) = \begin{bmatrix} \sum_i^N (1 * x_i) \\ \sum_i^N (z_i^1 * x_i) \\ \sum_i^N (z_i^2 * x_i) \\ \cdots \end{bmatrix}$$

An Estimator for Cross-Data-Set Regressions

- Example: Do R&D subsidies affect workers'/inventors' wages?



Conclusion

- Problem: Unmergeable data sets prevent regressions
- Contribution: New estimator that bridges separate data silos
- Application: Estimation of Cross-Country Production Function
- Code for the estimation: (soon)

Thank you for your attention

Cross-country-TFP-estimation

- TFP only comparable if derived from the same production function

Table: inputs, outputs and TFP for the example firms

Firm	labor	capital	output	TFP	coeff(l)	coeff(c)	country
Firm 1	25	16	20	1	$\frac{1}{2}$	$\frac{1}{2}$	POR
Firm 2	25	16	20	0.94	$\frac{2}{3}$	$\frac{1}{3}$	ROM
Firm 3	25	4	64	0.78	$\frac{2}{3}$	$\frac{1}{3}$	ROM

- Current round contains first cross-country comparable TFP from an estimated production function (Experimental)