Accounting for Sample Overlap in Metaanalysis

Heiko Rachinger
University of Vienna
(joint with Pedro Bom, Deusto Business School, Bilbao)

MAER-NET 2015
IES, Prague
September 2015

Introduction

Motivation: Meta-Analysis in Economics

- Macroeconomic data is observational and aggregated by nature
  $\Longrightarrow$ same/similar samples typically used by multiple studies
- Empirical studies usually report multiple estimates based on the same/similar data
  $\Longrightarrow$ empirical results tend to be positively correlated

Question: How should meta-analysis account for this correlation?

Introduction

Contributions

- Identify sample overlap as a source of correlation, especially relevant with economic data
- Illustrate the implications for statistical properties when not taking into account this correlation
- Propose a weighted least squares estimator to circumvent these statistical difficulties
- Illustrate how this WLS can be made operational using information available from the individual studies

Introduction

Relationship to Multilevel estimation

- Multilevel estimators taking into account correlation structures due to
  - dependency at some level (typically at study level: estimates within a study might be correlated because of similar data sets, similar modelling strategies, similar mistakes)
- In macroeconomics:
  - dependence over studies since different studies use similar data sets.
  - independence within studies: different data sets or split data sets in subsamples

Introduction

A few examples in the meta-analysis literature in macroeconomics with sample overlap

Study	Subject
Stanley (1998)	Ricardian equivalence
de Mooij & Ederveen (2003)	FDI effects of taxation
Doucouliagos & Paldam (2010)	Growth effects of aid
Havranek & Irsova (2011)	Vertical productivity spillovers from FDI
Alptekin & Levine (2012)	Growth effects of military expenditures
Bom & Ligthart (2013)	Output elasticity of public infrastructure

Simulations

Setup

- Interest in estimating the mean, $\mu = 0$, of a normal population:

$$Y \sim N(\mu, \sigma^2) \tag{1}$$

- meta-analyst retrieves M individual estimates of μ, each of them using a representative sample S_i of size N_i (for simplicity we assume $N_i = N$)
  - $M_1 = (1 - \lambda) M$ independent samples, $\{S_1, ...S_{M_1}\}$,
  - $M_2 = \lambda M$ samples, $\{S_{M_1}, ..., S_M\}$ with ρN overlap (repeated observations) in each sample

Simulations

Setup

- Types of overlap
  - (Case A) overlap only with S_1 (first of the independent samples) (extreme case of correlation)
    - extreme and somehow unrealistic scenario.
    - S_1 corresponds to e.g. U.S. 1980-2000.
  - (Case B) overlap with all independent samples (S_1 overlaps with S_{M_1+1},S_2 with S_{M_1+2}, etc) (for $\lambda > 1/2$, start again from beginning)
    - Other extreme scenario since typically there are few samples that are repeated much more heavily
  - two extreme cases and truth lies somewhere in between.

Simulations

Setup

Estimators

- Infeasible efficient estimator: average of all independent observations underlying the estimates
  - infeasible because we only have the estimates but not the underlying observations
- Simple average of all estimates
- inverse variance weighted average of all estimates
  - commonly used one
  - in our setup ($N_i = N$) very similar to the simple average
- inverse variance weighted average of the first M_1 estimates (samples without any overlap with any other sample)
  - data underlying the estimates is independent $\Longrightarrow$ good estimator
  - in practice, very few samples don´t have any overlap $\Longrightarrow$ somehow arbitrary and dropping many estimates which is against the philosophy of metaanalysis

Simulations

Setup

We look at the following statistics:

- size of a t-test for $H_0 : \mu = \mu_0 = 0$ (should be close to nominal level)
- relative efficiency measured as mean square error relative to the one of the infeasible inefficient estimator

$$\frac{MSE_{efficient}}{MSE_{estimator}}$$

Simulations

Size of the inverse variance weighted estimator (at 5% level)

a) Case A)

	i) M=30			ii) M=100			iii) M=500		
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	8.4	23.3	39.3	15.4	43.9	61.2	37.3	72.0	79.4
0.5	14.5	40.0	56.8	27.0	60.5	74.2	56.2	81.9	88.7
0.8	18.5	47.1	63.0	34.5	66.9	78.5	62.3	84.2	89.2
1	21.4	51.8	67.1	39.1	70.3	80.5	66.4	88.0	92.0

b) Case B)

	i) M=30			ii) M=100			iii) M=500		
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	6.7	6.1	14.2	5.9	7.7	15.1	6.0	7.7	14.6
0.5	7.6	11.5	27.6	7.2	13.3	27.0	8.0	11.5	25.9
0.8	9.1	15.9	36.1	9.8	15.5	36.0	9.5	13.6	33.7
1	10.0	17.1	36.2	11.9	19.4	39.3	10.2	17.0	41.1

Simulations

Efficiency: MSE relative to MSE of infeasible efficient (Case A)

a) relative efficiency of the inverse variance weighted estimator

	i) M=30			ii) M=100			iii) M=500		
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	0.79	0.45	0.23	0.58	0.18	0.086	0.21	0.043	0.01
0.5	0.65	0.26	0.16	0.36	0.097	0.052	0.09	0.021	0.01
0.8	0.52	0.22	0.16	0.27	0.080	0.053	0.07	0.015	0.01
1	0.51	0.22	0.24	0.24	0.074	0.077	0.06	0.015	0.01

b) relative efficiency of IVW of M_1 samples

	i) M=30			ii) M=100			iii) M=500		
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	0.82	0.54	0.24	0.85	0.52	0.26	0.77	0.56	0.21
0.5	0.86	0.71	0.35	0.87	0.64	0.35	0.87	0.62	0.29
0.8	0.93	0.83	0.56	0.93	0.79	0.55	0.96	0.78	0.60
1	0.97	0.99	0.97	0.97	0.98	0.99	0.97	0.97	0.96

Simulations

Efficiency: MSE relative to MSE of infeasible efficient (Case B)

a) relative efficiency of the inverse variance weighted estimator

	i) M=30			ii) M=100					
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	0.95	0.91	0.66	0.94	0.90	0.69	0.93	0.91	0.65
0.5	0.92	0.92	0.54	0.93	0.86	0.64	0.89	0.87	0.60
0.8	0.88	0.91	0.64	0.90	0.92	0.61	0.86	0.92	0.62
1	0.86	0.98	0.98	0.87	0.97	0.98	0.85	0.97	0.98

b) relative efficiency of IVW of M_1 independent samples

	i) M=30			ii) M=100					
$\rho \backslash \lambda$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
0.2	0.84	0.51	0.23	0.82	0.54	0.26	0.84	0.53	0.23
0.5	0.84	0.69	0.34	0.86	0.62	0.37	0.86	0.67	0.38
0.8	0.92	0.80	0.52	0.93	0.85	0.50	0.92	0.81	0.51
1	0.98	0.97	0.98	0.97	0.97	0.98	0.98	0.97	0.98

Simulations

Comments

t-tests on the mean are clearly oversized

$\Longrightarrow$ Two effects

- Numerator (in Case A): converges to a weighted average of the limit of the first sample estimate and $\mu_0 = 0$.

$$
\begin{aligned}
\hat{\mu}_{Meta-EW} &= \frac{1}{M}\sum_{j=1}^{M}\hat{\mu}_j = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N} y_{ji}\right) \\
&= \frac{1}{M}\left[\frac{1}{N}\sum_{i}^{N} y_{1i} + \sum_{j=2}^{(1-\lambda)M}\frac{1}{N}\sum_{i}^{N} y_{ji} + \sum_{j=(1-\lambda)M}^{M}\frac{1}{N}\sum_{i}^{N} y_{ji}\right] \\
&= \lambda\frac{1}{N}\sum_{i}^{\rho N} y_{1i} + \frac{1}{M}\left[\frac{1}{N}\sum_{\rho N+1}^{N} y_{1i} + \sum_{j=(1-\lambda)M}^{M}\frac{1}{N}\sum_{\rho N+1}^{N} y_{ji} + \sum_{j=2}^{(1-\lambda)M}\frac{1}{N}\sum_{i}^{N} y_{ji}\right] \\
&\underset{M\to\infty}{\to} \lambda\ \tilde{\mu}_{1,\rho N} + 0
\end{aligned}
$$

- This effect does not vanish for M$\longrightarrow \infty$ as long as proportion of samples with overlap, λ, doesn´t vanish!

Simulations

Comments

- Denominator: estimate of standard error of meta-estimate too small (we erronously assume we have more info than we do have)
  $\Longrightarrow$ t-stat too large $\Longrightarrow$ too many rejections
  $\Longrightarrow$ **clearly oversized tests: large Type I error**
  - effect gets worse with M (for Case A)
  - effect is much worse in Case A

Simulations

Comments

- relative MSE
  - IVW can become very inefficient
  - For Case A: effect clearly gets worse with M
  - IVW only using independent samples gets worse with increasing λ

$\Longrightarrow$ Adding new estimates can potentially harm (to a potentially large degree) the inference and therefore the meta-study

Our Objective

- Correct for the correlation structure in a Generalized LS fashion
- Using information on the overlap (provided in the primary studies) in order to construct the covariance matrix of the primary estimates

The general case

In economics, meta-analysis is more commonly applied on a slope parameter of a linear regression model, such as θ in

$$y = \alpha + \theta x + u. \tag{2}$$

Assuming the classical linear model assumptions, a meta-analysis model of M collected estimates of θ can be written as

$$\hat{\boldsymbol{\theta}} = \theta \boldsymbol{\iota} + \boldsymbol{\epsilon}, \qquad \mathrm{E}(\boldsymbol{\epsilon}) = 0, \quad \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \equiv \boldsymbol{\Omega}, \tag{3}$$

where $\boldsymbol{\Omega}$ is the variance-covariance matrix of $\boldsymbol{\epsilon}$.

- If primary samples don't overlap, then $\boldsymbol{\Omega}$ is diagonal
- If samples overlap, then some off-diagonal elements will be nonzero
- The 'generalized weights' meta-estimator is then given by

$$\tilde{\theta}_G = (\boldsymbol{\iota}'\boldsymbol{\Omega}^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\theta}}. \tag{4}$$

The general case

What are the elements of Ω?

In the case of OLS estimates:

- Diagonal elements (i.e., variances):

$$\text{var}(\epsilon_i) \approx \frac{\sigma_u^2}{N_i \sigma_x^2}, \tag{5}$$

  which are obtained from the primary studies.

- Off-diagonal elements (i.e., covariances):

$$\text{cov}(\epsilon_p, \epsilon_q) \approx \frac{C}{N_p}\text{var}(\epsilon_p) \approx \frac{C}{N_q}\text{var}(\epsilon_q) \tag{6}$$

  which can be computed from information in reported in primary studies

Special cases:

- 'Inverse-variance' weights: off diagonal elements of Ω are 0
- 'Simple average' (equal weights): off diagonal elements of Ω are 0, diagonal elements $= 1$

Simulations revisited

Size (Case a)

a) Size of the inverse variance weighted estimator

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	8.4	23.3	39.3
0.5	14.5	40.0	56.8
0.8	18.5	47.1	63.0
1	21.4	51.8	67.1

ii) M=100

0.2	0.5	0.8
15.4	43.9	61.2
27.0	60.5	74.2
34.5	66.9	78.5
39.1	70.3	80.5

b) Size of the proposed WLS

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	5.1	5.5	6.7
0.5	5.9	6.1	7.1
0.8	5.1	5.7	7.6
1	5.5	5.3	5.4

ii) M=100

0.2	0.5	0.8
5.8	6.3	8.2
6.5	5.2	7.3
6.1	6.0	9.8
5.8	4.1	5.7

Simulations revisited

Size (Case b)

a) Size of the inverse variance weighted estimator

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	6.7	6.1	14.2
0.5	7.6	11.5	27.6
0.8	9.1	15.9	36.1
1	10.0	17.1	36.2

ii) M=100

0.2	0.5	0.8
5.9	7.7	15.1
7.2	13.3	27.0
9.8	15.5	36.0
11.9	19.4	39.3

b) Size of the proposed WLS

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	5.8	4.4	5.8
0.5	5.6	5.5	4.9
0.8	5.6	6.7	7.3
1	4.6	5.8	5.5

ii) M=100

0.2	0.5	0.8
4.9	5.7	5.4
5.0	6.7	6.4
5.5	6.1	5.8
6.9	5.9	6.4

Simulations revisited

relative MSE (Case a)

a) relative efficiency of the inverse variance weighted estimator

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	0.79	0.45	0.23
0.5	0.65	0.26	0.16
0.8	0.52	0.22	0.16
1	0.51	0.22	0.24

ii) M=100

0.2	0.5	0.8
0.58	0.18	0.086
0.36	0.097	0.052
0.27	0.080	0.053
0.24	0.074	0.077

b) relative efficiency of the proposed WLS

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	0.89	0.68	0.34
0.5	0.89	0.71	0.37
0.8	0.92	0.80	0.50
1	0.97	0.99	0.97

ii) M=100

0.2	0.5	0.8
0.86	0.56	0.26
0.87	0.62	0.31
0.92	0.75	0.40
0.97	0.98	0.99

Simulations revisited

relative MSE (Case b)

a) relative efficiency of the inverse variance weighted estimator

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	0.95	0.91	0.66
0.5	0.92	0.92	0.54
0.8	0.88	0.91	0.64
1	0.86	0.98	0.98

ii) M=100

0.2	0.5	0.8
0.94	0.90	0.69
0.93	0.86	0.64
0.90	0.92	0.61
0.87	0.97	0.98

b) relative efficiency of the proposed WLS

i) M=30

$\rho \backslash \lambda$	0.2	0.5	0.8
0.2	0.96	0.91	0.65
0.5	0.93	0.91	0.52
0.8	0.95	0.90	0.60
1	0.98	0.97	0.98

ii) M=100

0.2	0.5	0.8
0.95	0.90	0.67
0.95	0.85	0.60
0.96	0.90	0.57
0.97	0.97	0.98

Simulations

Some comments

- proposed WLS controls size well
- in Case A, proposed WLS has potentially much higher relative efficiency
- efficiency of proposed WLS relative to infeasible efficient one improves with increasing ρ and decreasing λ
- efficiency of proposed WLS relative to IVW improves with ρ and λ
- IVW deterioriates (in terms of size and relative efficiency) with number of estimates M

Technical issues

Technical complications that are likely to arise in practice:

- Data aggregation issues (e.g., quarterly vs. yearly time-series)
- Different estimation methods (e.g., OLS vs. IV)

Under some assumptions, it can be shown that the overlapping covariance between...

- ...aggregated and disaggregated estimates is determined by the overlapping variance of the 'disaggregated' estimate
- ...OLS and IV estimates is determined by the overlapping variance of the OLS estimate

Extensions to the basic model

The meta-analysis model can be extended to address:

- observed heterogeneity, by including moderator variables
- publication bias, by including the standard errors

Conclusion

Contributions

- Identify sample overlap as a source of correlation, especially relevant with economic data
- Illustrate the implications for statistical properties when not taking into account this correlation
- Propose a weighted least squares estimator to circumvent these statistical difficulties
- Illustrate how this WLS can be made operational using information available from the individual studies
- Up to now: clear homogenous setup. More work is necessary for dealing with heterogenous setup.

Thank you very much for your attention!

Technical issues

How to find the overlapping variances? Run the auxiliary regression:

$$\ln \text{var}(\epsilon_i) = \alpha_0 + \alpha_1 \ln N_i + \sum_{l=1}^{L} \phi_l D_{li} + \varepsilon_i, \tag{7}$$

where:

- $\hat{\alpha}_1$ should be approximately -1
- D_{li} is a model design characteristic that may effect the variance of the estimator (e.g., IV estimation)

Once estimated, we can use this question to infer the sampling variance for any sample size and model design characteristic.