Accounting for Sample Overlap in Meta-Analysis*

Pedro R.D. Bom†

October 17, 2014

Abstract

A common feature of meta-analyses in economics, especially in macroeconomics and related subfields, is that the samples underlying the reported effect sizes overlap. The resulting positive correlation between effect sizes decreases the efficiency of standard meta-estimation methods. This paper argues that the variance-covariance matrix describing the structure of dependency between primary estimates can be feasibly specified as function of information that is typically reported in the primary studies. Meta-estimation efficiency can then be enhanced by using the resulting matrix in a Generalized Least Squares fashion.

JEL code: C13

Keywords: meta-analysis, meta-regression, sample overlap

*The author gratefully acknowledges financial support from Heinrich Graf Hardegg'sche Stiftung. The author is also thankful to Heiko Rachinger, Joshua Sherman, and participants of the MAER-Net 2014 Athens Colloquium, Greece, for their valuable comments and suggestions.

†Department of Economics, University of Vienna, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, Phone: +43-1-4277-37477, Fax: +43-1-4277-37498, E-mail: `pedro.bom@univie.ac.at`.

1 Introduction

Meta-analysis is a powerful statistical technique. By combining empirical results reported in multiple studies, it acts as to 'enlarge' the underlying sample from which an effect size is to be inferred. Not surprisingly, meta-analysis has been extensively used in fields such as biology, medical research, and psychology, among others.[1] In these areas, empirical research is typically conducted using randomized controlled trials, which employ a necessarily limited number of subjects. By mimicking a study with a larger number of subjects, meta-analysis allows to estimate an effect size more precisely.

But the precision gains of meta-analysis are not always so clear. In economics, empirical research is mostly based on observational (rather than experimental) data, which typically requires a higher level of statistical sophistication.[2] As a result, the disparity in empirical results in a given literature is often dominated by differences in study design characteristics—e.g., model specification, estimation method, and definition of variables—rather than sampling error. Pooling estimates from different studies is thus more likely to increase the variability of population parameters than to inform on the true size of a particular one. Moreover, in those fields of economics where data is by nature aggregated—as is the case in macroeconomics and related subfields—it is common to see the same (or nearly the same) data being repeatedly employed in several different studies.[3] Overlapping samples imply a positive correlation between the resulting estimates, which has implications for their optimal (efficiency-maximizing)

[1]Recent examples in these fields include research on biodiversity and ecological restoration (Benayas, Newton, Diaz, and Bullock, 2009), the study of H5N1 infections in humans (Wang, Parides, and Palese, 2012), and research on the neural bases of social cognition and story comprehension (Mar, 2011).

[2]There is a growing body of literature testing economic theories using experimental methods, however, which has motivated a number of meta-analyses of empirical research conducted in the lab. Recent examples include Weizsäcker (2010), Cooper and Dutcher (2011), and Engel (2011).

[3]A non-exhaustive list of meta-studies on macroeconomics-related topics where sample overlap may be an issue includes: Stanley (1998) on Ricardian equivalence; de Mooij and Ederveen (2003) on the FDI effects of taxation; de Dominicis, Florax, and de Groot (2008) on the relationship between income inequality and economic growth; Eickmeier and Ziegler (2008) on the forecast quality of dynamics factor models; Doucouliagos and Paldam (2010) on the growth effects of aid; Havranek (2010) on the trade effects of currency unions in gravity models of international trade; Efendic, Pugh, and Adnett (2011) on institutional quality and economic performance; Feld and Heckemeyer (2011) on FDI and taxation; Havranek and Irsova (2011) on vertical FDI productivity spillovers; Alptekin and Levine (2012) on the effect of military expenditures on economic growth; Adam, Kamas, and Lagou (2013) on the effects of globalization and capital market integration on capital taxes; Bom and Ligthart (2013) on the output elasticity of public capital; Celbis, Nijkamp, and Poot (2013) on the impact of infrastructure on exports and imports; Gechert (2013) on the output effects of fiscal policy shocks; and Melo, Graham, and Brage-Ardao (2013) on the output elasticity of transport infrastructure.

combination. Whereas the issue of study design heterogeneity has been successfully tackled within the context of a meta-regression model, the problem of overlapping samples has been largely ignored. The present paper addresses this issue.

To account for estimate dependency caused by sample overlap, I propose a 'generalized weights' meta-estimator. This method requires the full specification of the variance-covariance matrix of the primary estimates in terms of observables. I show how, under some assumptions, the elements of this matrix can be approximately written as function of quantities that are typically reported in the primary studies, such as samples sizes, sample overlap, and standard errors. This variance-covariance matrix can then be used to optimally weight the observations in the meta-sample. The generalized weights meta-estimator is thus a feasible application of the generalized least squares (GLS) principle. Intuitively, each observation in the meta-regression model is weighted according to how much independent sampling information it contains. Under no sample overlap, the generalized weights meta-estimator reduces to the 'inverse-variance' meta-estimator, which weights each primary estimate according to the reciprocal of its variance.[4]

To clarify ideas and build intuition, I provide in Section 2 a simple example of a three-study meta-analysis of the mean of a normal population. This example helps understand the difference between first-best and second-best efficiency and shows that, under sample overlap, no meta-estimator is first-best efficient. It also illustrates the efficiency gains of accounting for sample overlap by comparing the mean squared errors of the generalized weights meta-estimator, the inverse-variance meta-estimator, and the simple-average meta-estimator. Section 3 then describes the derivation of the generalized weights meta-estimator in the general case. It shows how the elements of the variance-covariance matrix should be computed, given the information collected from the primary studies. Section 4 provides an application of the generalized weights meta-estimator to a meta-sample of estimated output elasticities of public capital meta-analyzed by Bom and Ligthart (2013). The main result from this application is that accounting for sample overlap can significantly affect meta-estimates. Section 5 summarizes the findings and discusses limitations and directions for future research.

[4]See Stanley and Jarrell (1989), Stanley (2005), and Stanley and Doucouliagos (2010).

2 A Simple Example

Consider an economic variable of interest—say, household disposable income in a given year—whose population is normally distributed with mean μ and variance σ^2:

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

where, for simplicity, σ^2 is assumed to be known. The interest lies on the magnitude of the population mean, μ. Suppose that three studies are available, each of them computing and reporting one estimate of μ using the sample average estimator. Denote the sample average reported by the i-th study by $\bar{y}_i$ and the corresponding sample of size N_i by $S_i = \{y_{i1}, ..., y_{iN_i}\}$, for $i = 1, 2, 3$, so that $\bar{y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. Finally, assume that samples S_1 and S_2 overlap by $C \leq \min\{N_1, N_2\}$ observations—i.e., C independent realizations of Y are contained in both S_1 and S_2—but are totally independent of S_3. The question is then: how can we efficiently meta-estimate μ if only $\bar{y}_i$, N_i, and C (but not the individual observations in S_i) are reported in the primary studies?

Before tackling this question, let us consider the (unrealistic) case where the meta-analyst observes the primary samples S_1, S_2, and S_3. In this case, the efficient 'meta-estimator' of μ would simply average across the $N_1 + N_2 - C + N_3$ independent realizations of Y, excluding the C 'duplicated' observations in S_2:

$$\begin{aligned} \tilde{y}_F &= \frac{1}{N-C}\left(\sum_{j=1}^{N_1} y_{1j} + \sum_{k=1}^{N_2-C} y_{2k} + \sum_{l=1}^{N_3} y_{3l}\right) \\ &= \left(\frac{N_1}{N-C}\right)\bar{y}_1 + \left(\frac{N_2-C}{N-C}\right)\bar{y}_2^c + \left(\frac{N_3}{N-C}\right)\bar{y}_3, \end{aligned} \tag{2}$$

where $N \equiv N_1 + N_2 + N_3$ is the total number of observations in the three samples, and $\bar{y}_2^c = \frac{1}{N_2-C}\sum_{k=1}^{N_2-C} y_{2k}$ denotes the sample average of S_2 after excluding the C overlapping observations. The second line of (2) shows that the full information estimator can be written as a weighted average of the individual sample averages (after adjusting for sample overlap), using as weights the fraction of independent observations in each primary sample. Because all primary sampling information would be taken into account, this estimator would be first-best efficient. Below, I refer to $\bar{y}_F$ as the 'full information' meta-estimator.

What if, more realistically, the meta-analyst does not observe the primary samples? Then, a feasible meta-estimator must only depend on the observed $\bar{y}_i$'s:

$$\tilde{y}_S = \omega_1\bar{y}_1 + \omega_2\bar{y}_2 + \omega_3\bar{y}_3, \tag{3}$$

where $\omega_i \in (0,1)$ is the weight assigned to $\bar{y}_i$, $i = 1,2,3$. This implies that, because $\bar{y}_2^c$ is not observed, $\tilde{y}_S$ cannot replicate the full information estimator $\tilde{y}_F$. Unless samples do not overlap (i.e., $C = 0$), therefore, a feasible meta-estimator will not be first-best efficient. The question then is: how to choose the ω_i's so as to achieve second-best efficiency?

Assume first that samples do not overlap (i.e., $C = 0$) so that $\tilde{y}_S$ is also first-best efficient for the optimal choice of weights. If primary samples are equally sized (i.e., $N_1 = N_2 = N_3$), then according to (2) the weights should also be identical: $\omega_1 = \omega_2 = \omega_3 = 1/3$. If primary samples have different sizes, however, the weights should differ; in particular, $\omega_i = N_i/N$, for $i = 1,2,3$, so that primary estimates based on larger samples are given higher weights. This is equivalent to weighting using the inverse of the variance of $\bar{y}_i$—a common procedure in meta-analysis—because this variance is proportional to $1/N_i$.

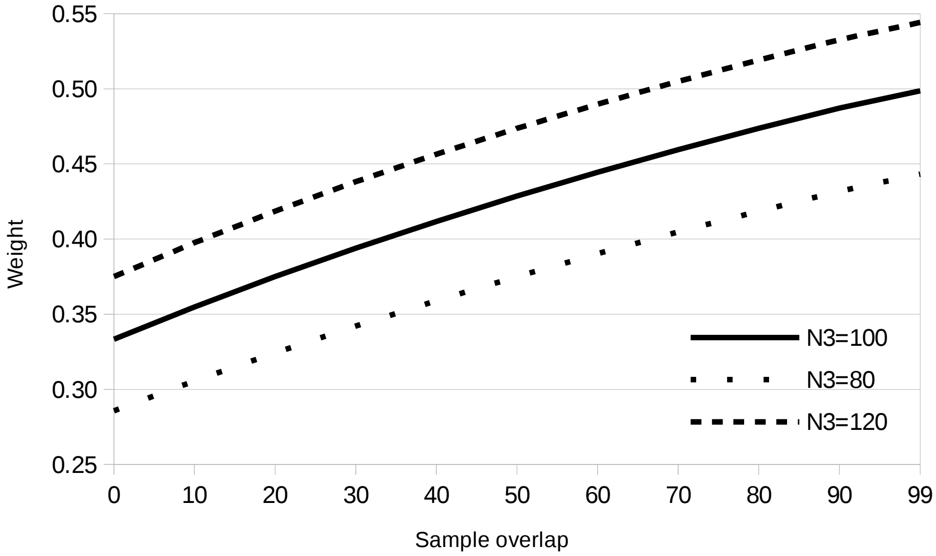
But if samples S_1 and S_2 overlap to some extent (i.e., $C > 0$), then $\bar{y}_1$ and $\bar{y}_2$ will be positively correlated. The optimal (second-best) weights must thus take this correlation into account. Using the procedure described below in Section 3.1, it can be shown analytically that the optimal weights in this case are:

$$\omega_1 = \frac{N_1 - C}{N - 2C - \frac{C^2 N_3}{N_1 N_2}}, \qquad \omega_2 = \frac{N_2 - C}{N - 2C - \frac{C^2 N_3}{N_1 N_2}}, \qquad \omega_3 = \frac{N_3 - \frac{C^2 N_3}{N_1 N_2}}{N - 2C - \frac{C^2 N_3}{N_1 N_2}},$$

which clearly depend on the degree of sample overlap, C. Below, I refer to the meta-estimator $\tilde{y}_S$ with these optimal (second-best) weights as the the 'generalized weights' meta-estimator. Note that these weights reduce to the 'inverse-variance' weights (i.e., $\omega_i = N_i/N$) for $C = 0$ and to 1/3 if, additionally, $N_1 = N_2 = N_3$.

To illustrate the relationship between estimation weights, sample size, and degree of sample overlap, Figure 2 plots the estimation weight of $\bar{y}_3$ as a function of C assuming $N_1 = N_2 = 100$ and various values of N_3. If $N_3 = 100$ (solid line), the weight of $\bar{y}_3$ is 1/3 for $C = 0$ and monotonically increases with C, approaching 1/2 as C approaches 100. In-

Figure 1: Optimal ω_3 as a function of C for $N_1 = N_2 = 100$ and various values of N_3

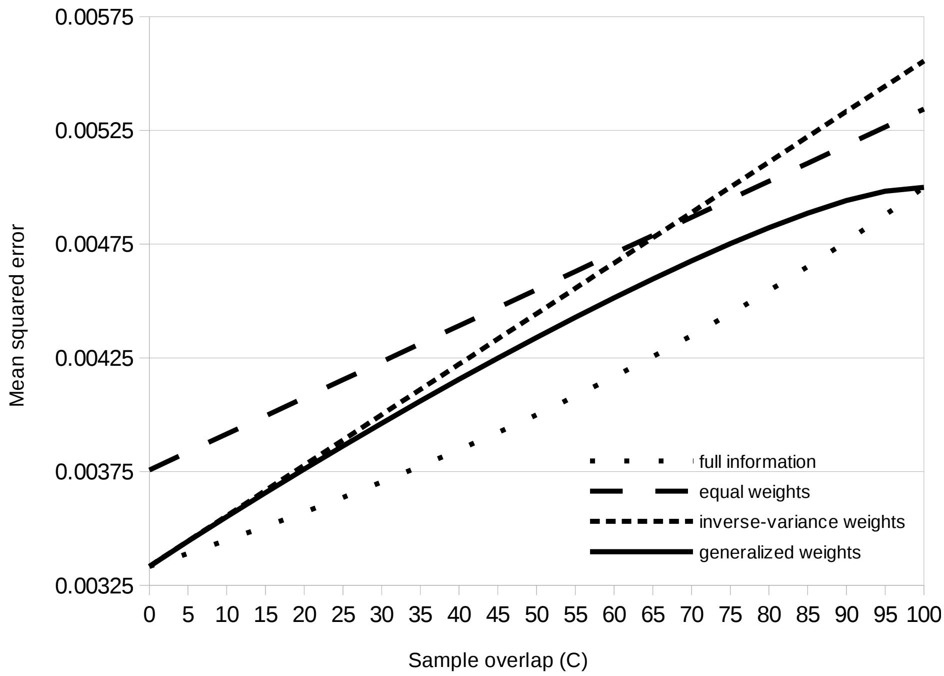


tuitively, because the primary samples are equally sized, each primary estimate receives the same weight if there is no sample overlap ($C = 0$). Full sample overlap, in turn, implies a weight of 1/2 for $\bar{y}_1 = \bar{y}_2$ (because S_1 and S_2 are in fact the same sample) and 1/2 for $\bar{y}_3$.[5] Clearly, a larger (smaller) size of S_3 implies a larger (smaller) weight to $\bar{y}_3$—as depicted by the dashed (dotted) lines—for any value of C.

How large are the efficiency gains from accounting for sample overlap? Figure 2 plots the mean squared error (MSE) of the meta-estimator $\tilde{y}_S$ as a function of sample overlap in the cases of equal weights, inverse-variance weights, and generalized weights. To compare first-best and second-best efficiency, Figure 2 also shows the MSE of the full information estimator (i.e., $\tilde{y}_F$). If there is no sample overlap (i.e., $C = 0$), the inverse-variance weights and generalized weights meta-estimators are identical to the full information estimator, so that first-best efficiency is achieved (the MSE of the simple average meta-estimator is higher because sample sizes differ). For positive values of C, however, the MSE of the generalized weights meta-estimator is lower than both the equal weights and the inverse-variance weights

[5]Rigorously speaking, the weights in (4) are not defined for $C = N_1 = N_2$—but only as C approaches $N_1 = N_2$—since the weight's denominator would in this case be zero. Intuitively, because $\bar{y}_1$ and $\bar{y}_2$ are identical, their weights cannot be disentangled.

Figure 2: MSE of the various meta-estimators as a function of sample overlap



Notes: The sizes of the primary samples are $N_1 = 140$, $N_2 = 100$, and $N_3 = 60$.

meta-estimators— which do not account for sample overlap—but larger than the first-best full information estimator. As C approaches its maximum value of 100 ($= \min\{N_1, N_2\}$), the generalized weights meta-estimator is again first-best efficient.[6] Note, finally, that a large degree of sample overlap may render the inverse-variance weights meta-estimator less efficient than a simple average.

3 The Generalized Weights Meta-Estimator

In economics, the object of meta-analysis is typically a slope parameter of a linear regression model. Assume, without loss of generality, that an economic variable of interest y (e.g., household disposable income in a given year) is linearly related to a single covariate x (e.g., average years of schooling of the household's income earners) in the population:

$$y = \alpha + \theta x + u, \tag{4}$$

[6]Note that $\tilde{y}_F$ can be computed in this case, because $\bar{y}_2^c = 0$.

where u is an independent and normally distributed error term with mean zero and variance σ_u^2. For simplicity, x is assumed to be non-stochastic (i.e., fixed for repeated samples). The interest lies in the magnitude of θ. A literature search reveals that M estimates of θ are available. Denote the i-th primary estimate by $\hat{\theta}_i$ and the respective primary sample of size N_i by $S_i = \{(y_{i1}, x_{i1}), (y_{i2}, x_{i2}), ..., (y_{iN_i}, x_{iN_i})\}$, for $i = \{1, 2, ..., M\}$. I allow for overlapping samples, so that samples are not necessarily independent from each other. In particular, I denote by C_{pq} the number of observations that are common to S_p and S_q. Up to sample overlap, however, samples are collections of independent realizations of y (given x).

3.1 The Baseline Case

Suppose that each primary estimate of θ is obtained by running an Ordinary Least Squares (OLS) regression of y on x using the N_i observations in S_i:

$$\hat{\theta}_i = \frac{\sum\limits_{j=1}^{N_i} \tilde{y}_{ij}\tilde{x}_{ij}}{\sum\limits_{j=1}^{N_i} \tilde{x}_{ij}^2} = \theta + \frac{\sum\limits_{j=1}^{N_i} \tilde{x}_{ij}u_{ij}}{\sum\limits_{j=1}^{N_i} \tilde{x}_{ij}^2}, \qquad \text{for } i = 1, 2, ..., M, \tag{5}$$

where $\tilde{y}_{ij} \equiv y_{ij} - N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ and $\tilde{x}_{ij} \equiv x_{ij} - N_i^{-1}\sum_{j=1}^{N_i} x_{ij}$ denote the demeaned dependent and independent variables, respectively. Because, under our assumptions, the OLS estimator is unbiased and consistent for θ, we can write (5) as

$$\hat{\theta}_i = \theta + \epsilon_i, \tag{6}$$

where ϵ_i is a sampling error component with zero mean.[7] Equation (6) provides the basis of a meta-analysis model of θ. Collecting the M primary estimates in vector $\hat{\boldsymbol{\theta}} \equiv [\hat{\theta}_1 \;\; \hat{\theta}_2 \;\; \ldots \;\; \hat{\theta}_M]'$, and the respective sampling errors in vector $\boldsymbol{\epsilon} \equiv [\epsilon_1 \;\; \epsilon_2 \;\; \ldots \;\; \epsilon_M]'$, we can write the meta-analysis model (6) as

$$\hat{\boldsymbol{\theta}} = \theta\boldsymbol{\iota} + \boldsymbol{\epsilon}, \qquad \mathrm{E}(\boldsymbol{\epsilon}) = 0, \quad \mathrm{var}(\boldsymbol{\epsilon}) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \equiv \boldsymbol{\Omega}, \tag{7}$$

where $\boldsymbol{\iota}$ denotes a $M \times 1$ vector of ones.

[7] For simplicity, we assume the homogeneity case where primary studies estimate a unique population effect, θ. Of course, differences in study design can lead the population parameters to be heterogeneous. This case case could be easily accommodated by means of a meta-regression model, which would include dummy variables capturing study design characteristics (see Stanley and Jarrell, 1989).

What are the elements of the variance-covariance matrix $\mathbf{\Omega}$? The main diagonal contains the variances of the ϵ_i's, which are given by

$$\mathrm{var}(\epsilon_i) = \frac{\sigma_u^2}{\sum_{j=1}^{N_i} \tilde{x}_{ij}^2} \approx \frac{\sigma_u^2}{N_i \sigma_x^2}, \tag{8}$$

where σ_x^2 is the population variance of x. Clearly, ϵ_i is heteroskedastic, since its variance is inversely proportional to N_i. These variances can be computed by squaring the standard errors of $\hat{\theta}_i$, which are typically reported in the primary studies. The off-diagonal elements of $\mathbf{\Omega}$ represent pairwise covariances between primary estimates and describe the dependency structure implied by sample overlap. Consider two primary estimates, $\hat{\theta}_p$ and $\hat{\theta}_q$, whose underlying samples of sizes N_p and N_q are denoted by S_p and S_q. Allow for sample overlap and denote by $C_{pq} \geq 0$ the number of observations of S_p overlapping with S_q, and by $C_{qp} \geq 0$ the number of observations of S_q overlapping with S_p; in this section, $C_{pq} = C_{qp}$.[8] Appendix A.1 then shows that the covariance between ϵ_p and ϵ_q is

$$\mathrm{cov}(\epsilon_p, \epsilon_q) \approx \frac{C_{pq} C_{qp}}{N_p N_q} \mathrm{cov}(\epsilon_{p,c}, \epsilon_{q,c}), \tag{9}$$

where $\epsilon_{i,c}$, for $i = \{p, q\}$, is the sampling error component corresponding to $\hat{\theta}_i$ if only the overlapping observations are used. But for the overlapping observations, $\hat{\theta}_p = \hat{\theta}_q$, so that $\epsilon_{p,c} = \epsilon_{q,c}$. Hence, $\mathrm{cov}(\epsilon_{p,c}, \epsilon_{q,c}) = \mathrm{var}(\epsilon_{p,c}) = \mathrm{var}(\epsilon_{q,c})$. Because, from (8), $\mathrm{var}(\epsilon_{p,c}) = \frac{N_p}{C_{pq}} \mathrm{var}(\epsilon_p)$ and $\mathrm{var}(\epsilon_{q,c}) = \frac{N_q}{C_{qp}} \mathrm{var}(\epsilon_q)$, we can finally write (9) as

$$\mathrm{cov}(\epsilon_p, \epsilon_q) \approx \frac{C_{qp}}{N_q} \mathrm{var}(\epsilon_p) = \frac{C_{pq}}{N_p} \mathrm{var}(\epsilon_q), \tag{10}$$

which can be computed from the information reported in the primary studies.[9]

Having specified the elements in matrix $\mathbf{\Omega}$, efficient meta-estimation of θ amounts to employing Generalized Least Squares (GLS)—by pre-multiplying both sides of (7) by $\mathbf{\Omega^{-1/2}}$—so that the meta-regression model actually estimated is

$$\boldsymbol{\hat{\theta}^*} = \theta \boldsymbol{\iota^*} + \boldsymbol{\epsilon^*}, \qquad \mathrm{E}(\boldsymbol{\epsilon}^*) = 0, \quad \mathrm{var}(\boldsymbol{\epsilon}^*) = \mathrm{E}(\boldsymbol{\epsilon}^* \boldsymbol{\epsilon}^{*\prime}) \equiv \boldsymbol{I}_M, \tag{11}$$

[8] The distinction between C_{pq} and C_{qp} will matter below, when discussing data aggregation.

[9] Using (7) and (8), it follows that the correlation coefficient between ϵ_p and ϵ_q is given by $\frac{C_{pq}}{\sqrt{N_p N_q}}$.

where $\hat{\boldsymbol{\theta}}^* \equiv \boldsymbol{\Omega}^{-1/2}\hat{\boldsymbol{\theta}}$, $\boldsymbol{\iota}^* \equiv \boldsymbol{\Omega}^{-1/2}\boldsymbol{\iota}$, $\boldsymbol{\epsilon}^* \equiv \boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon}$, and $\boldsymbol{I}_M$ is the $M \times M$ identity matrix. In essence, the meta-analysis model (11) is reweighted so that the residuals appear homoskedastic and uncorrelated. Intuitively, the larger the variance of a primary estimate or the more correlated it is with another primary estimate, the lower its estimation weight will be. This procedure gives rise to what we call the 'generalized weights' meta-estimator:

$$\tilde{\theta}_G = (\boldsymbol{\iota}'\boldsymbol{\Omega}^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\theta}}, \tag{12}$$

which I alluded to in Section 2. It follows from (11) that the sampling variance of $\tilde{\theta}_G$ is

$$\text{var}(\tilde{\theta}_G) = (\boldsymbol{\iota}'\boldsymbol{\Omega}^{-1}\boldsymbol{\iota})^{-1}. \tag{13}$$

It is important to note that, if the underlying samples of any two primary estimates, S_p and S_q, perfectly overlap (i.e., $C_{pq} = C_{qp} = N_p = N_q$), then $\text{cov}(\epsilon_p, \epsilon_q) = \text{var}(\epsilon_p) = \text{var}(\epsilon_q)$, which implies singularity (and thus non-invertibility) of $\boldsymbol{\Omega}$.

3.2 Data Aggregation Issues

In the previous section, I assumed that primary samples are drawn from the same population model, defined by (4), implying that the primary data are defined at the same level of aggregation. In practice, however, empirical studies often employ data defined at different levels of aggregation. Time series studies, for instance, may employ data at different frequencies (e.g., yearly, quarterly, or monthly data). Cross section or panel data studies, on the other hand, may use data at different layers of geographical aggregation (e.g., regional or national). This section discusses the particular pattern of sample overlap that may arise in such cases and how they affect the off-diagonal elements of $\boldsymbol{\Omega}$.

In this context, suppose that two primary estimates, $\hat{\theta}_p$ and $\hat{\theta}_q$, are obtained from overlapping samples containing data aggregated at different levels. Let S_p denote the sample of disaggregated data and S_q the sample of aggregated data. Each overlapping observation in S_q aggregates at least F overlapping observations of S_p. Hence, if S_q consists of yearly time series data and S_p contains quarterly data, then $F = 4$. Assuming that the population model is valid at both levels of aggregation, what are the variances and covariance between ϵ_p and ϵ_q? Clearly, the variances of ϵ_p and ϵ_q are still given by (8), after noting that σ_u^2 and σ_x^2 now depend on the level of data aggregation:

$$\mathrm{var}(\epsilon_i) \approx \frac{\sigma_{u_i}^2}{N_i \sigma_{x_i}^2}, \qquad \text{for } i = p, q, \tag{14}$$

which, again, can be obtained from the primary studies. In terms of the covariance between $\epsilon_{p,c}$ and $\epsilon_{q,c}$, Appendix A.2 shows that, for the overlapping observations (i.e., C_{pq} in S_p and C_{qp} in S_q), it is given by the variance of the least aggregated of the two primary estimates:

$$\mathrm{cov}(\epsilon_{p,c}, \epsilon_{q,c}) = \mathrm{var}(\epsilon_{p,c}). \tag{15}$$

Because this variance can itself be approximated by $\frac{N_p}{C_{pq}}\mathrm{var}(\epsilon_p)$, we can use it in (9) to find

$$\mathrm{cov}(\epsilon_p, \epsilon_q) \approx \frac{C_{qp}}{N_q}\mathrm{var}(\epsilon_p). \tag{16}$$

3.3 Other Estimation Methods

Section 3.1 assumed that θ is estimated by OLS. While this is mostly the case in practice, some studies may nevertheless apply different estimation techniques. The typical choice among these is the instrumental variables (IV) estimator, which is designed to address endogeneity concerns. In short, the IV estimator replaces x_i by an instrumental variable, z_i, in the expression of the OLS estimator (5):

$$\hat{\theta}_i = \frac{\sum_{j=1}^{N_i} \tilde{y}_{ij}\tilde{z}_{ij}}{\sum_{j=1}^{N_i} \tilde{x}_{ij}\tilde{z}_{ij}} = \theta + \frac{\sum_{j=1}^{N_i} \tilde{z}_{ij}u_{ij}}{\sum_{j=1}^{N_i} \tilde{x}_{ij}\tilde{z}_{ij}}, \tag{17}$$

where $\tilde{z}_{ij}$ denotes deviations from its average.

The question is, again, how does the presence of IV estimates affect the elements of $\mathbf{\Omega}$ as defined in Section 3.1? From (17) it follows that the variance of the error term, $\epsilon_i = \hat{\theta}_i - \theta$, is approximately given by

$$\mathrm{var}(\epsilon_i) \approx \frac{\sigma_u^2}{N_i \sigma_x^2 \rho_{xz}^2}, \tag{18}$$

where σ_x^2 is the population variance of the covariate, x, and ρ_{xz} is the correlation coefficient between x and its instrument, z. An estimate of this variance should be reported in the primary study. As for the covariance, Appendix A.1 shows that (9) holds irrespective of the primary estimates being OLS or IV. Moreover, if both primary estimates are IV, then the overlapping covariance equals both overlapping variances, so that (10) also holds. But what if one primary estimate is OLS and the other is IV? Denote the OLS estimator by $\hat{\theta}_p$ and the IV estimator by $\hat{\theta}_q$. Then, Appendix A.3 shows that the overlapping covariance equals the variance of the OLS estimator, so that

$$\mathrm{cov}(\epsilon_p, \epsilon_q) \approx \frac{C_{qp}}{N_q} \mathrm{var}(\epsilon_p). \tag{19}$$

4 An Application: the Output Elasticity of Public Capital

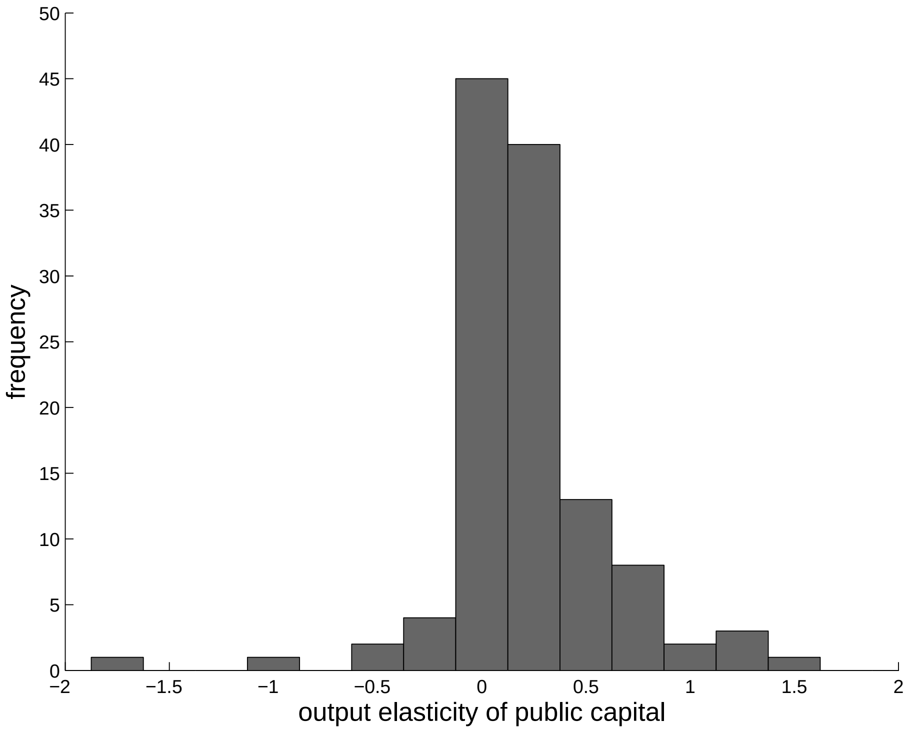
This section provides an application of the generalized weights meta-estimator to the output elasticity of public capital, using the meta-dataset of Bom and Ligthart (2013). Section 4.1 briefly describes the meta-dataset. Section 4.2 discusses the meta-estimation results.

4.1 Brief Description of the Meta-Sample

Bom and Ligthart (2013) conduct a meta-analysis of empirical studies that estimate the output elasticity of public capital using the production function approach. Their meta-sample contains 578 estimates from 68 studies conducted between 1983 and 2007. It exhibits severe sample overlap, both because of within-study sample overlap (i.e., multiple reported measurements in the same study using the same or similar underlying sample) and between-study sample overlap (i.e., the same or similar sample used by multiple studies). Indeed, almost half of the measurements (278) are based on data for the United States.

As discussed above, the generalized weights meta-estimator is not well-defined when two or more primary samples exactly overlap, since the variance-covariance matrix of the sampling error component would then be singular. In fact, the sample overlap issue is so severe in Bom and Ligthart's (2013) meta-dataset that the vast majority—458, more precisely—of their primary estimates are derived from samples that exactly overlap. This application therefore only uses the 120 estimates from 46 studies that contain some independent sampling information.

Figure 3: Histogram of Estimated Output Elasticities of Public Capital



Notes: The histogram is based on 120 estimates used by Bom and Ligthart (2013).

Figure 4.1 provides an histogram of the 120 output elasticities of public capital. The estimates lie between -1.726 and 1.601, with a mean of 0.212 and a median of 0.142. This mean-larger-median property, which implies that the empirical distribution is skewed toward large positive values, also characterizes the full meta-sample of 578 estimates. Bom and Ligthart (2013) observe that publication bias may cause this asymmetry. Here, however, we ignore publication bias issues and focus directly on meta-estimating the underlying output elasticity of public capital.

4.2 Meta-Estimation Results

Table 4.2 presents the results of meta-estimating the output elasticity of public capital using the 120 primary estimates discussed in the previous section. It compares the meta-estimates from the generalized weights meta-estimator with those from the simple average and inverse-variance meta-estimators. Along with the meta-estimates, Table 4.2 reports standard errors

Table 1: The Output Elasticity of Public Capital: Meta-Estimation Results

Meta-Estimator	Meta-Estimate	Std. Error	95% Conf. Interval
Simple Average	0.212	0.036	[0.141–0.283]
Inverse-Variance Weights	0.028	0.007	[0.014–0.041]
Generalized Weights	0.087	0.014	[0.059–0.115]

Notes: The results are based on 120 estimates used by Bom and Ligthart (2013).

and 95% confidence intervals.

As reported in the previous section, the (simple) average output elasticity of public capital in the meta-sample amounts to 0.212. Its relatively large standard error of 0.036, which partly reflects the lack of efficiency of this meta-estimator, implies a wide confidence interval of 0.141–0.283. The inverse-variance weights meta-estimate, in contrast, is only 0.028—roughly a tenth of the simple average—with a rather narrow confidence interval of 0.014–0.041. Despite the obvious efficiency gains from allowing for different sample sizes, the small standard error is nevertheless biased downwards, as it does not account for sample overlap. Finally, the generalized weights meta-estimate is 0.087, more than three times larger the inverse-variance weights meta-estimate. Its standard error is also about twice as large, reflecting both the the correction of the downward bias and the efficiency gains of correcting for sample overlap. Note that the confidence intervals of the generalized weights and inverse-variance weights meta-estimates do not overlap, which suggests that the changes for meta-analytical inference of accounting for sample overlap are non-negligible. The bottom line of these results is that adjusting the weights can affect meta-estimation results and inference to a large extent in the case of substantial sample overlap.

5 Concluding Remarks

Meta-analyzes in economics, especially in the field of macroeconomics and related subfields, often face the issue of overlapping primary samples. This problem arises when several studies report estimates of an effect size that are based on primary samples with common observations. Sample overlap gives rise to dependency between the primary estimates being meta-analyzed, thus decreasing the efficiency of standard meta-analytical estimation methods.

This paper argues that, although first-best meta-estimation efficiency is unattainable under sample overlap, second-best efficiency can be achieved by fully specifying the variance-covariance matrix of the model's error component. The paper shows that elements of this matrix can be approximated using information either readily available from the primary studies (such as the variances of the reported estimates and the corresponding samples sizes) or at least computable from the information reported in the primary studies (such as the number of overlapping observations).

I provide an application of the generalized weights meta-estimator to Bom and Ligthart's (2013) meta-analysis of the output elasticity of public capital. The results suggest that accounting for sample overlap can significantly change the meta-estimate of the true effect size. Whereas the inverse-variance weights meta-estimator (which accounts for different sample size but assumes no overlap) gives a meta-estimate of 0.028, the generalized weights meta-estimator (which accounts for both different sample sizes and sample overlap) gives a meta-estimate of 0.087. This increase is substantial both statistically and economically.

But the generalized weights meta-estimator suffers from one important drawback: although the efficiency gains from this method increase with the degree of sample overlap, it cannot deal with two or more estimates from exactly overlapping samples. This forces the meta-analyst to be selective, which inevitably requires subjective evaluation. Moreover, it cannot provide an answer to the long-standing question among meta-analysts of how many estimates per study to include in the meta-analysis (see, e.g., Bijmolt and Pieters, 2001).

One possible way to tackle this issue is to allow for a random effects component in the meta-analytical model. If strictly positive, the variance of this error component would show up in the main diagonal of the error term's variance-covariance matrix, breaking down its singularity. In fact, the reason why so many estimates are typically reported from the same primary sample is because they differ in some way (observed or not); this random component is thus likely to be empirically relevant in practice. I intend to pursue this line of research in the future.

Appendix

A.1 Deriving the Covariance Between Overlapping Estimates

As in Section 3, consider two primary estimates of θ, $\hat{\theta}_p$ and $\hat{\theta}_q$, obtained from samples S_p and S_q of sizes N_p and N_q. The primary samples S_p and S_q overlap to same extent. Denote by C_{pq} the number of elements of S_p overlapping with S_q, and by C_{qp} the number of elements of S_q overlapping with S_p. We can split sample S_p into a subset of $\bar{N}_p \equiv N_p - C_{pq}$ independent observations and a second subset of C_{pq} observations overlapping with sample S_q. Similarly, S_q is split into a subset of C_{qp} observations overlapping with S_p and a subset of $\bar{N}_q \equiv N_q - C_{qp}$ independent observations.

OLS primary estimates. Using tildes to denote deviations from averages—i.e., $\tilde{y}_{pi} \equiv y_{pi} - \bar{y}_p$ and $\tilde{x}_{pi} \equiv x_{pi} - \bar{x}_p$—the OLS estimator $\hat{\theta}_p$ can be written as

$$
\begin{aligned}
\hat{\theta}_p &= \frac{\sum_{j=1}^{\bar{N}_p} \tilde{y}_{pj}\tilde{x}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}^2} + \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{y}_{pj}\tilde{x}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}^2} \\
&= \frac{\sum_{j=1}^{\bar{N}_p} \tilde{x}_{pj}^2}{\sum_{j=1}^{N_p} \tilde{x}_{pj}^2} \frac{\sum_{j=1}^{\bar{N}_p} \tilde{y}_{pj}\tilde{x}_{pj}}{\sum_{j=1}^{\bar{N}_p} \tilde{x}_{pj}^2} + \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{x}_{pj}^2}{\sum_{j=1}^{N_p} \tilde{x}_{pj}^2} \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{y}_{pj}\tilde{x}_{pj}}{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{x}_{pj}^2} \\
&= \frac{\mathrm{SST}_x^{\bar{N}_p}}{\mathrm{SST}_x^{N_p}} \hat{\theta}_{p,p} + \frac{\mathrm{SST}_x^{C_{pq}}}{\mathrm{SST}_x^{N_p}} \hat{\theta}_{p,c},
\end{aligned}
\tag{A.1}
$$

where $\hat{\theta}_{p,p}$ and $\hat{\theta}_{p,c}$ denote the OLS estimators using only the $\bar{N}_p$ independent and the C_{pq} overlapping observations, respectively. The term $\mathrm{SST}_x^{N_p} \equiv \sum_{j=1}^{N_p} \tilde{x}_{pj}^2 = \mathrm{SST}_x^{\bar{N}_p} + \mathrm{SST}_x^{C_{pq}}$ is the total sum of squares of x in S_p, which equals the total sum of squares of x for the $\bar{N}_p$ independent observations ($\mathrm{SST}_x^{\bar{N}_p}$) and the total sum of squares of x for the C_{pq} overlapping observations ($\mathrm{SST}_x^{C_{pq}}$). Equation (A.1) simply writes the OLS estimator $\hat{\theta}_p$ as a convex combination of the subsample estimators $\hat{\theta}_{p,p}$ and $\hat{\theta}_{p,c}$. Noting that $\frac{\mathrm{SST}_x^{\bar{N}_p}}{\mathrm{SST}_x^{N_p}} \approx \frac{\bar{N}_p}{N_p}$ and $\frac{\mathrm{SST}_x^{C_{pq}}}{\mathrm{SST}_x^{N_p}} \approx$

$\frac{C_{pq}}{N_p}$, we can write (A.1) as

$$\hat{\theta}_p \approx \frac{\bar{N}_p}{N_p}\hat{\theta}_{p,p} + \frac{C_{pq}}{N_p}\hat{\theta}_{p,c}. \tag{A.2}$$

Following a similar procedure for $\hat{\theta}_q$, we find

$$\hat{\theta}_q \approx \frac{\bar{N}_q}{N_q}\hat{\theta}_{q,q} + \frac{C_{qp}}{N_q}\hat{\theta}_{q,c}. \tag{A.3}$$

IV primary estimates. The approximations (A.2) and (A.3) also hold in the case of IV primary estimates. Denoting the demeaned instrumental variable by $\tilde{z}_{pi} \equiv z_{pi} - \bar{z}_p$, the IV estimator reads

$$\begin{aligned}
\hat{\theta}_p &= \frac{\sum_{j=1}^{\bar{N}_p} \tilde{y}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} + \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{y}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} \\
&= \frac{\sum_{j=1}^{\bar{N}_p} \tilde{x}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} \frac{\sum_{j=1}^{\bar{N}_p} \tilde{y}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{\bar{N}_p} \tilde{x}_{pj}\tilde{z}_{pj}} + \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} \frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{y}_{pj}\tilde{z}_{pj}}{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}}.
\end{aligned} \tag{A.4}$$

Hence, by noting that $\frac{\sum_{j=1}^{\bar{N}_p} \tilde{x}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} \approx \frac{\bar{N}_p}{N_p}$ and $\frac{\sum_{j=\bar{N}_p+1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}}{\sum_{j=1}^{N_p} \tilde{x}_{pj}\tilde{z}_{pj}} \approx \frac{C_{pq}}{N_p}$, we obtain (A.2).

Derivation of Equation (9). The covariance between ϵ_p and ϵ_q is then

$$\begin{aligned}
\mathrm{cov}(\epsilon_p, \epsilon_q) &= \mathrm{E}(\epsilon_p\epsilon_q) \\
&= \mathrm{E}[(\hat{\theta}_p - \theta)(\hat{\theta}_q - \theta)] \\
&\approx \mathrm{E}\left[\left(\frac{\bar{N}_p}{N_p}(\hat{\theta}_{p,p} - \theta) + \frac{C_{pq}}{N_p}(\hat{\theta}_{p,c} - \theta)\right)\left(\frac{\bar{N}_q}{N_q}(\hat{\theta}_{q,q} - \theta) + \frac{C_{qp}}{N_q}(\hat{\theta}_{q,c} - \theta)\right)\right] \\
&= \mathrm{E}\left[\frac{C_{pq}C_{qp}}{N_pN_q}(\hat{\theta}_{p,c} - \theta)(\hat{\theta}_{q,c} - \theta)\right] \\
&= \frac{C_{pq}C_{qp}}{N_pN_q}\mathrm{E}(\epsilon_{p,c}\epsilon_{q,c}) \\
&= \frac{C_{pq}C_{qp}}{N_pN_q}\mathrm{cov}(\epsilon_{p,c}, \epsilon_{q,c}),
\end{aligned} \tag{A.5}$$

where I have used that $\mathrm{E}[(\hat{\theta}_{p,p} - \theta)(\hat{\theta}_{q,q} - \theta)] = \mathrm{E}[(\hat{\theta}_{p,p} - \theta)(\hat{\theta}_{q,c} - \theta)] = \mathrm{E}[(\hat{\theta}_{q,q} - \theta)(\hat{\theta}_{p,c} - \theta)] = 0$ in going from the third to the fourth line.

A.2 Overlapping Covariance for Data Aggregated at Different Levels

This sections derives equation (15). Using (5) and noting that $u_{qi} = u_{p(i-1)F+1} + \cdots + u_{piF}$, we can write for the overlapping observations C_{pq} and C_{qp}:

$$\begin{aligned}
\operatorname{cov}(\epsilon_{p,c}, \epsilon_{q,c}) &= \mathrm{E}\left[\frac{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi} u_{pi} \sum_{i=1}^{C_{qp}} \tilde{x}_{qi}(u_{p(i-1)F+1} + \cdots + u_{piF})}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi}^2}\right] \\
&= \frac{(\tilde{x}_{p1} + \cdots + \tilde{x}_{pF})\tilde{x}_{q1}\sigma_{u_p}^2 + \cdots + (\tilde{x}_{pC_{pq}-F+1} + \cdots + \tilde{x}_{pC_{pq}})\tilde{x}_{qC_{qp}}\sigma_{u_p}^2}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi}^2} \\
&= \frac{\sigma_{u_p}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi}^2}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi}^2} = \frac{\sigma_{u_p}^2}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2} = \operatorname{var}(\epsilon_{p,c}), \qquad \text{(A.6)}
\end{aligned}$$

where I have used that $\tilde{x}_{qi} = \tilde{x}_{p(i-1)F+1} + \cdots + \tilde{x}_{piF}$ in going from the second line to the first term in the last line.

A.3 Overlapping Covariance Between OLS and IV Estimators

Here, I derive the covariance between the OLS and IV estimators for the overlapping observations. Let ϵ_p and ϵ_q denote the sampling error terms corresponding to the OLS and IV estimators, respectively. I assume that the data is aggregated at the same level, so that $\tilde{x}_p = \tilde{x}_q$, $u_p = u_q$, and $C_{pq} = C_{qp}$. Using (5) and (17), we find

$$\begin{aligned}
\operatorname{cov}(\epsilon_{p,c}, \epsilon_{q,c}) &= \mathrm{E}\left[\frac{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi} u_{pi} \sum_{i=1}^{C_{qp}} \tilde{z}_{qi} u_{qi}}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi} \tilde{z}_{pi}}\right] \\
&= \frac{\tilde{x}_{p1}\tilde{z}_{q1}\sigma_{u_p}^2 + \tilde{x}_{p2}\tilde{z}_{q2}\sigma_{u_p}^2 + \cdots + \tilde{x}_{pC_p}\tilde{z}_{qC_q}\sigma_{u_p}^2}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi} \tilde{z}_{pi}} \\
&= \frac{\sigma_{u_p}^2 \sum_{i=1}^{C_{pq}} \tilde{x}_{pi} \tilde{z}_{qi}}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2 \sum_{i=1}^{C_{qp}} \tilde{x}_{qi} \tilde{z}_{pi}} = \frac{\sigma_{u_p}^2}{\sum_{i=1}^{C_{pq}} \tilde{x}_{pi}^2} = \operatorname{var}(\epsilon_{p,c}). \qquad \text{(A.7)}
\end{aligned}$$

References

Adam, A., P. Kammas, and A. Lagou (2013): "The Effect of Globalization on Capital Taxation: What Have We Learned After 20 Years of Empirical Studies?," *Journal of Macroeconomics*, 35, 199–209.

Alptekin, A., and P. Levine (2012): "Military Expenditure and Economic Growth: A Meta-Analysis," *European Journal of Political Economy*, 28, 636–650.

Benayas, J., A. Newton, A. Diaz, and J. Bullock (2009): "Enhancement of Biodiversity and Ecosystem Services by Ecological Restoration: A Meta-Analysis," *Science*, 325, 1121–1124.

Bijmolt, T. H. A., and R. G. M. Pieters (2001): "Meta-Analysis in Marketing when Studies Contain Multiple Measurements," *Marketing Letters*, 12, 157–169.

Bom, P. R., and J. E. Ligthart (2013): "What Have We Learned from Three Decades of Research on the Productivity of Public Capital?," *Journal of Economic Surveys*, doi: 10.1111/joes.12037.

Celbis, M. G., P. Nijkamp, and J. Poot (2013): "How Big Is the Impact of Infrastructure on Trade? Evidence from Meta-Analysis ," UNU-MERIT Working Papers, United Nations University, No. 2013-32.

Cooper, D., and E. Dutcher (2011): "The Dynamics of Responder Behavior in Ultimatum Games: a Meta-Study," *Experimental Economics*, 14, 519–546.

de Dominicis, L., R. J. G. M. Florax, and H. L. F. de Groot (2008): "A Meta-Analysis on the Relationship between Income Inequality and Growth," *Scottish Journal of Political Economy*, 55, 654–682.

Doucouliagos, H., and M. Paldam (2010): "Conditional Aid Effectiveness: A Meta-Study," *Journal of International Development*, 22, 391–410.

Efendic, A., G. Pugh, and N. Adnett (2011): "Institutions and Economic Performance: A Meta-Regression Analysis," *European Journal of Political Economy*, 27, 586–599.

EICKMEIER, S., AND C. ZIEGLER (2008): "How Successful are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach," *Journal of Forecasting*, 27, 237–265.

ENGEL, C. (2011): "Dictator Games: a Meta Study," *Experimental Economics*, 14, 583–610.

FELD, L. P., AND J. H. HECKEMEYER (2011): "FDI and Taxation: A Meta-Study," *Journal of Economic Surveys*, 25, 233–272.

GECHERT, S. (2013): "What Fiscal Policy is Most Effective? A Meta-Regression Analysis," Working Paper, Institut für Makroökonomie und Konjunkturforschung, No.117.

HAVRÁNEK, T. (2010): "Rose Effect and the Euro: Is the Magic Gone?," *Review of World Economics*, 146, 241–261.

HAVRÁNEK, T., AND Z. IRSOVA (2011): "Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect Is," *Journal of International Economics*, 85, 234–244.

MAR, R. (2011): "The Neural Bases of Social Cognition and Story Comprehension," *Annual Review of Psychology*, 62, 103–134.

MELO, P. C., D. J. GRAHAM, AND R. BRAGE-ARDAO (2013): "The Productivity of Transport Infrastructure Investment: A Meta-Analysis of Empirical Evidence," *Regional Science and Urban Economics*, 43, 695–706.

MOOIJ, R. D., AND S. EDERVEEN (2003): "Taxation and Foreign Direct Investment: A Synthesis of Empirical Research," *International Tax and Public Finance*, 10, 673–693.

STANLEY, T. D. (1998): "New Wine in Old Bottles: A Meta-Analysis of Ricardian Equivalence," *Southern Economic Journal*, 64, 713–727.

——— (2005): "Beyond Publication Bias," *Journal of Economic Surveys*, 19, 309–345.

STANLEY, T. D., AND S. B. JARRELL (1989): "Meta-Regression Analysis: A Quantitative Method of Literature Reviews," *Journal of Economic Surveys*, 3, 161–170.

——— (2010): "Picture This: A Simple Graph that Reveals Much Ado About Research," *Journal of Economic Surveys*, 24, 170–191.

Wang, T., M. Parides, and P. Palese (2012): "Seroevidence for H5N1 Influenza Infections in Humans: Meta-Analysis," *Science*, 335, 1463.

Weizsäcker, G. (2010): "Do We Follow Others when We Should? A Simple Test of Rational Expectations," *American Economic Review*, 100, 2340–2360.