MEMORANDUM

No 21/2014

Practical Correlation Bias Correction in Two-way Fixed Effects Linear Regression

Simen Gaure

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.sv.uio.no/econ
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 20/14	Rolf Aaberge, Tarjei Havnes and Magne Mogstad *A Theory for Ranking Distribution Functions*
No 19/14	Alice Ciccone *Is It All About CO2 Emissions? The Environmental Effects of Tax Reform for New Vehicles in Norway*
No 18/14	Mikolaj Czajkowski, Nick Hanley and Karine Nyborg *Social Norms, Morals and Self-interest as Determinants of Pro-environment Behaviours*
No 17/14	Karine Nyborg *Reciprocal Climate Negotiators: Balancing Anger against Even More Anger*
No 16/14	Karen Evelyn Hauge and Ole Røgeberg *Contribution to Public Goods as Individuals versus Group Representatives: Evidence of Gender Differences*
No 15/14	Moti Michael and Daniel Spiro *Skewed Norms under Peer Pressure: Formation and Collapse*
No 14/14	Daniel Spiro *Resource Prices and Planning Horizons*
No 13/14	Johan Gars and Daniel Spiro *Uninsurance through Trade*
No 12/14	Moti Michaeli and Daniel Spiro *Powerty in China as Seen from Outer Space*
No 11/14	Ingvild Almås, Åshild Auglænd Johnsen and Andreas Kotsadam *Powerty in China as Seen from Outer Space*

Previous issues of the memo-series are available in a PDF® format at:
http://www.sv.uio.no/econ/english/research/memorandum/

Practical correlation bias correction in two-way fixed effects linear regression

Simen Gaure

The Ragnar Frisch Centre for Economic Research, Oslo, Norway

Memo 21/2014-v1

This version August 2014

Abstract

When doing two-way fixed effects OLS estimations, both the variances and covariance of the fixed effects are biased. A formula for a bias correction is known, but in large datasets it involves inverses of impractically large matrices. We detail how to compute the bias correction in this case.

Keywords: Limited mobility bias, Two way fixed effects, Linear regression
JEL: C13, C33, C55, C87

1. Introduction

We consider a model of the type:

$$y = X\beta + D\theta + F\psi + \epsilon, \tag{1}$$

where $y \in \mathbb{R}^n$ is an outcome, X is a matrix of covariates, D is an $n \times k_\theta$ matrix resulting from dummy encoding a factor, F is an $n \times k_\psi$ matrix resulting from dummy encoding another factor, and ϵ is a normally distributed error term. This is a perfectly ordinary least squares system, but our assumption is that k_θ and k_ψ are large, e.g. of the order 10^5–10^7. This creates some computational challenges. The canonical example in the panel data econometrics literature of this kind of model can be found in [1], where the outcome y is the wage, D is a matrix of dummies for each individual, and F is a matrix of dummies for each firm. θ are time-constant individual fixed effects, ψ are time-constant firm fixed effects. They study the correlation $\operatorname{cor}(D\theta, F\psi)$ as a way to investigate whether "high wage" workers tend to work in "high wage" firms.

We assume that β, θ, ψ, and ϵ are estimated by OLS, e.g. with the methods in [7, 10, 17, 19]. It is shown in [3] that the variances $\tilde\sigma^2_\theta = \operatorname{var}(D\hat\theta)$ and $\tilde\sigma^2_\psi = \operatorname{var}(F\hat\psi)$ are positively biased, and that the covariance $\tilde\sigma_{\theta\psi} = \operatorname{cov}(D\hat\theta, F\hat\psi)$ is typically negatively biased, and they give explicit formulas for the magnitude of the biases. The biases can be substantial, and can even change the sign of the correlation estimate: $\tilde\rho_{\theta\psi} = \frac{\tilde\sigma_{\theta\psi}}{\sqrt{\tilde\sigma^2_\theta \tilde\sigma^2_\psi}}$. This particular type of bias is known as *limited mobility* bias.

A challenge with the bias correction formulas in [3] is that they involve the inverses of large square matrices, of size k_θ and k_ψ. Given that these quantities can be of the order 10^5–10^7, the method is impractical to use directly with commonly available computing contraptions. Some authors acknowledge the possible bias, but do not compute it, e.g. [6, 8, 9, 12, 20]. We therefore venture to evaluate the bias correction expressions in [3] without handling any large matrices. Our contributions are mainly in section 5, but for completeness and consistency of notation we include a derivation of the bias expressions from [3] in sections 3 and 4.

In applications there can be other sources of bias than the one corrected by the methods presented here. To solve such bias problem, other models should be used, as in [5, 16]. It is also pointed out in the literature, e.g. in [6], that the OLS assumption of independently identically distributed errors is dubious in some applications, though the bias expressions and the methods in section 5 can probably be adapted to at least some other error structures.

2. Preliminaries

We fix some notation and recall some standard facts about (orthogonal) projections. In general, we let I denote the identity matrix of appropriate size. We assume tacitly that our matrices are of the appropriate size. For a matrix A we denote by $\mathrm{R}(A)$ its column space, or range. We denote by M_A the projection onto the orthogonal complement of $\mathrm{R}(A)$. Note that in general, $M_A = M_A^t = M_A^2$ by the defining property of projections. For A of full column rank, we have

$$M_A = I - A(A^tA)^{-1}A^t, \tag{2}$$

but M_A is defined for any matrix A. For two matrices A and B we denote by $M_{A,B} = M_{B,A}$ the intersection $M_A \wedge M_B$, the projection onto the complement of the column space of the block matrix $\begin{bmatrix} A & B \end{bmatrix}$. In general, $M_{A,B}M_A = M_AM_{A,B} = M_{A,B}$, and $M_{A,B}A = 0$. A standard result in operator theory is that if $\mathrm{R}(A)$ is orthogonal to $\mathrm{R}(B)$, or if $\mathrm{R}(A) \subset \mathrm{R}(B)$, then $M_{A,B} = M_AM_B = M_BM_A$. We denote by $\mathbf{1} = (1, 1, \ldots, 1)$ a vector of the appropriate length where each coordinate equals 1. Thus, $M_{\mathbf{1}}$ is the projection which subtracts the mean.

We will now and then use the defining property of the trace, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, without mentioning.

With this notation, we may state some assumptions for our system in (1). There is no intercept in X. We have removed a reference group from ψ/F. There are no more collinearities in the system; in the language of [2], there is a single *connected group*, or *connected component*. These assumptions are necessary for identification of $\hat\theta$ and $\hat\psi$. In particular, $M_{F,X}D$ and $M_{D,X}F$ are assumed to be of full column rank, so that both $D^tM_{F,X}D$ and $F^tM_{D,X}F$ are invertible. We also have $M_DM_{\mathbf{1}} = M_{\mathbf{1}}M_D = M_D$. We do *not* assume that X is small, i.e. X may, among other covariates, contain one or more high dimensional dummy encoded factors, as in [7].

This lemma will come in handy later:

Lemma 2.1. If A and B are matrices, then $M_{A,B} = M_AM_{M_AB}$. If M_AB has full column rank, we have $M_{A,B} = M_A - M_AB(B^tM_AB)^{-1}B^tM_A$.

Proof. First, note that $M_A(I - M_{A,B})$ is a projection. Let $P = I - M_{M_AB}$. P is the projection onto the range of M_AB, i.e. $\mathrm{R}(P) = \mathrm{R}(M_AB)$. We show that $\mathrm{R}(M_AB) = R(M_A(I - M_{A,B}))$. Note that $\mathrm{R}(M_AB)$ is spanned by the columns of M_AB, i.e. $\mathrm{R}(M_AB) = M_A\,\mathrm{R}(B)$. On the right hand side the columns of $I - M_{A,B}$ span $R(A) + R(B)$. So that $\mathrm{R}(M_A(I-M_{A,B})) = M_A(\mathrm{R}(A)+\mathrm{R}(B)) = M_A\,\mathrm{R}(B)$. Two projections with the same range are equal, so $P = M_A(I - M_{A,B}) = M_A - M_{A,B}$. Substituting P, we obtain $I - M_{M_AB} = M_A - M_{A,B}$. Multiplying through with M_A yields $M_A - M_AM_{M_AB} = M_A - M_{A,B}$, which can be rewritten as $M_{A,B} = M_AM_{M_AB}$. In the case that M_AB has full column rank, we have from (2) that $M_{M_AB} = I - M_AB(B^tM_AB)^{-1}B^tM_A$. □

Remark 2.2. Given a vector v we note that [10, Algorithm 3.1] gives a procedure by which we can compute $M_{D,F}v$. It is not mentioned explicitly in [10] that the same method can be used to compute M_Av for an arbitrary $n \times k$ matrix A, not only for matrices arising from dummy-encoding. The theory and algorithm is the same, but the actual computation of each projection in [10, Algorithm 3.1(2) and (15)] corresponding to columns of A, is slightly more complicated. Such a procedure has been implemented in [11] through implementation of interactions between factors and continuous covariates; one may use a factor with a single level. In the present paper, there is also no intrinsic dependence on D and F being dummy encoded factors, most of the theory (except we don't need a reference group, and $M_{\mathbf{1}}M_D \neq M_D$) is the same if D and F are interactions between factors and covariates, or something else; but the author knows of no such application.

3. Variances and covariance bias

When deriving the bias correction formulas, we will follow the exposition in [3], but change the notation to reflect our emphasis on the projections of the type M_A, which we can compute.

As in [3, (8–10)], we have a biased sample estimate for the variance of $D\theta$:

$$\tilde{\sigma}_\theta^2 = \frac{\hat{\theta}^t D^t M_{\mathbf{1}} D\hat{\theta}}{n}, \tag{3}$$

for the variance of $F\psi$:

$$\tilde{\sigma}_\psi^2 = \frac{\hat{\psi}^t F^t M_{\mathbf{1}} F\hat{\psi}}{n}, \tag{4}$$

and for the covariance:

$$\tilde{\sigma}_{\theta\psi} = \frac{\hat{\theta}^t D^t M_{\mathbf{1}} F\hat{\psi}}{n}. \tag{5}$$

We take the expectation as in [3, (16)]:

$$\mathbb{E}(\tilde{\sigma}_\theta^2) = \mathbb{E}\left(\frac{\hat{\theta}^t D^t M_{\mathbf{1}} D\hat{\theta}}{n}\right) = \frac{\theta^t D^t M_{\mathbf{1}} D\theta}{n} + \frac{\operatorname{tr}(D^t M_{\mathbf{1}} D \operatorname{Var}(\hat{\theta}))}{n} \tag{6}$$

using the general formula for the expectation of a quadratic form

$$\mathbb{E}(x^t A x) = \mathbb{E}(x^t) A\, \mathbb{E}(x) + \operatorname{tr}(A \operatorname{Var}(x)) \tag{7}$$

with $A = D^t M_{\mathbf{1}} D$ and $x = \hat{\theta}$.

We are interested in the term $\sigma_\theta^2 = (\theta^t D^t M_{\mathbf{1}} D\theta)/n$. We can readily estimate the left hand side of (6) as $\tilde{\sigma}_\theta^2$ from the OLS estimate $\hat{\theta}$. To find the bias, i.e. the trace term in (6), we need an expression for $\operatorname{Var}(\hat{\theta})$. The problem is the same for $\tilde{\sigma}_\psi^2$ and $\tilde{\sigma}_{\theta\psi}$, but we detail it only for the θ case.

Remark 3.1. We note that the bias problem is symmetric in θ and ψ, even though not all our formulas will be syntactically symmetric. Also, $\tilde{\sigma}_\theta^2$, $\tilde{\sigma}_\psi^2$, and $\tilde{\sigma}_{\theta\psi}$ do not depend on which reference group we have picked, neither do they depend on whether the reference group is in θ or ψ. Indeed, $M_{\mathbf{1}} D\hat{\theta}$ and $M_{\mathbf{1}} F\hat{\psi}$ are independent of where the reference group is. To see this, a change of reference group has the same effect on $D\hat{\theta}$ and $F\hat{\psi}$ as a transformation of the type $D\hat{\theta} \mapsto D\hat{\theta} - \alpha\mathbf{1}$, $F\hat{\psi} \mapsto F\hat{\psi} + \alpha\mathbf{1}$. But we have $M_{\mathbf{1}}\mathbf{1} = 0$. That is, in e.g. (6), both $\tilde{\sigma}_\theta^2$ and σ_θ^2 are independent of the whereabouts of the reference group, so the trace term is independent of it as well. For simplicity, we do assume that the reference group is in ψ.

We may find a formula for $\operatorname{Var}(\hat{\theta})$ via the Frisch-Waugh-Lovell theorem. By multiplying through (1) with $M_{F,X}$ we have

$$M_{F,X} y = M_{F,X} D\theta + M_{F,X}\epsilon.$$

By standard OLS assumptions we have

$$\hat{\theta} = \theta + (D^t M_{F,X} D)^{-1} D^t M_{F,X}\epsilon,$$

and, by using the i.i.d. assumption $\operatorname{Var}(\epsilon) = \sigma_\epsilon^2 I$, we obtain

$$\begin{aligned}\operatorname{Var}(\hat{\theta}) &= (D^t M_{F,X} D)^{-1} D^t M_{F,X} \operatorname{Var}(\epsilon) M_{F,X} D (D^t M_{F,X} D)^{-1} \\ &= \sigma_\epsilon^2 (D^t M_{F,X} D)^{-1}. \end{aligned} \tag{8}$$

As usual, σ_ϵ^2 can be estimated from the residuals $\hat{\epsilon}$ when solving (1) by OLS.

That is, the bias term for $\tilde{\sigma}_\theta^2$ in (6) is

$$\delta_\theta = \hat{\sigma}_\epsilon^2 \operatorname{tr}((D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}} D)/n. \tag{9}$$

It is the $k_\theta \times k_\theta$ matrix inside the trace term which may be too large to be handled directly, as in [3, p. 687].

We may rewrite

$$\operatorname{tr}((D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}} D) = \operatorname{tr}(M_{\mathbf{1}} D (D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}}) = \operatorname{tr}(Q^t Q) \geq 0,$$

with $Q = (D^t M_{F,X} D)^{-1/2} D^t M_{\mathbf{1}}$, so the bias is non-negative. By symmetry between θ and ψ, the corresponding bias term for $\tilde{\sigma}^2_\psi$ is:

$$\delta_\psi = \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_{D,X} F)^{-1} F^t M_{\mathbf{1}} F)/n. \tag{10}$$

For the covariance in (5), note the general algebraic formula for a quadratic form with $A = A^t$, sometimes referred to as a polarization identity:

$$(x+y)^t A (x+y) = x^t A x + 2 x^t A y + y^t A y.$$

That is,

$$x^t A y = \frac{1}{2}\left((x+y)^t A (x+y) - x^t A x - y^t A y\right).$$

We use this and (7) on (5), with $x = D\hat{\theta}$, $y = F\hat{\psi}$, and $A = M_{\mathbf{1}}$. An algebraic excursion yields:

$$\begin{aligned}\mathbb{E}(\tilde{\sigma}_{\theta\psi}) = \frac{1}{n}(&\theta^t D^t M_{\mathbf{1}} F \psi + \\ &+ \sigma^2_\epsilon \operatorname{tr}(M_{\mathbf{1}} D (D^t M_{F,X} D)^{-1} D^t M_{F,X} M_{D,X} F (F^t M_{D,X} F)^{-1} F^t)).\end{aligned}$$

As in [3], we can use Lemma 2.1 to write: $M_{D,X} = M_X(I - D(D^t M_X D)^{-1} D^t M_X)$. We obtain $D^t M_{F,X} M_{D,X} F = -D^t M_{F,X} D (D^t M_X D)^{-1} D^t M_X F$ and rewrite the trace term as:

$$-\operatorname{tr}(M_{\mathbf{1}} D (D^t M_X D)^{-1} D^t M_X F (F^t M_{D,X} F)^{-1} F^t),$$

or, as in [3, (22)]

$$\delta_{\theta\psi} = -\sigma^2_\epsilon \operatorname{tr}(D^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_X D (D^t M_X D)^{-1})/n. \tag{11}$$

Remark 3.2. We see from the formulas that the magnitude of the bias increases with σ^2_ϵ, ceteris paribus. Although unobserved heterogeneity does not bias the OLS estimates, it does bias non-linear functions of them, such as $\tilde{\sigma}^2_\theta$. Other error structures than $\operatorname{Var}(\epsilon) = \sigma^2_\epsilon I$ can be accomodated by changing equation (8), provided an estimate for $\operatorname{Var}(\epsilon)$ can be found.

4. Independent covariates

In the special case when we have no X, or the columns of X are orthogonal to both D and F, we may simplify the bias correction formulas (9), (10), and (11) by observing that since $\mathrm{R}(X)$ is orthogonal to $\mathrm{R}(D)$ and $\mathrm{R}(F)$, we have $M_X D = D$, and $M_X F = F$. Also, $M_{D,X} = M_D M_X$, so that $M_{D,X} F = M_D M_X F = M_D F$, similarly $M_{F,X} D = M_F D$. From (2) we also have $D(D^t D)^{-1} D^t = I - M_D$.

We then obtain:

$$\begin{aligned}\delta'_\theta &= \hat{\sigma}^2_\epsilon \operatorname{tr}((D^t M_F D)^{-1} D^t M_{\mathbf{1}} D)/n,\\ \delta'_\psi &= \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t M_{\mathbf{1}} F)/n,\\ \delta'_{\theta\psi} &= -\hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t (I - M_D) M_{\mathbf{1}} F)/n.\end{aligned}$$

Using $(I - M_D) M_{\mathbf{1}} = M_{\mathbf{1}} - M_D$, and $(M_{\mathbf{1}} - M_D)^2 = M_{\mathbf{1}} - M_D$, we note that

$$\begin{aligned}\delta'_{\theta\psi} &= -\hat{\sigma}^2_\epsilon \operatorname{tr}((M_{\mathbf{1}} - M_D) F (F^t M_D F)^{-1} F^t (M_{\mathbf{1}} - M_D))/n\\ &= -\hat{\sigma}^2_\epsilon \operatorname{tr}(Q^t Q) \leq 0,\end{aligned}$$

with $Q = (F^t M_D F)^{-1/2} F^t (M_{\mathbf{1}} - M_D)$.

We may also rewrite:

$$\begin{aligned}
\delta'_{\theta\psi} &= \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t (M_D - M_{\mathbf{1}})F)/n \\
&= \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t M_D F)/n - \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t M_{\mathbf{1}} F)/n \\
&= \hat{\sigma}^2_\epsilon \operatorname{tr}(I)/n - \delta'_\psi \\
&= \hat{\sigma}^2_\epsilon \frac{k_\psi}{n} - \delta'_\psi.
\end{aligned}$$

In particular, we have $\delta'_\psi \geq \sigma^2_\epsilon \frac{k_\psi}{n}$.

By Remark 3.1, we also have

$$\delta'_{\theta\psi} = \hat{\sigma}^2_\epsilon \frac{k_\theta - 1}{n} - \delta'_\theta,$$

where we replace k_ψ with $k_\theta - 1$ in the numerator because of the reference group. This means we can write δ'_ψ in terms of δ'_θ:

$$\delta'_\psi = \delta'_\theta - \hat{\sigma}^2_\epsilon \frac{k_\theta - 1 - k_\psi}{n}.$$

That is, when the covariates X are uncorrelated with the factors D and F, the bias corrections are:

$$\begin{aligned}
\delta'_\theta &= \hat{\sigma}^2_\epsilon \operatorname{tr}((D^t M_F D)^{-1} D^t M_{\mathbf{1}} D)/n && \geq \hat{\sigma}^2_\epsilon \frac{k_\theta - 1}{n}, \\
\delta'_\psi &= \delta'_\theta - \hat{\sigma}^2_\epsilon \frac{k_\theta - 1 - k_\psi}{n} && \geq \hat{\sigma}^2_\epsilon \frac{k_\psi}{n}, \\
\delta'_{\theta\psi} &= -\delta'_\theta + \hat{\sigma}^2_\epsilon \frac{k_\theta - 1}{n} && \leq 0.
\end{aligned}$$

This is a computational advantage, since it suffices to compute a single trace.

5. Computing the trace

Computing the trace of a matrix is simple in theory, it is just to sum the diagonal elements. However, if the matrices in (9), (10), and (11) are too large to be handled by commonly available computers, we need some other method. Luckily, quantum physicists and others have studied such problems for quite some time. The following is one approach.

By using (7) with an x with $\mathbb{E}(x) = 0$ and $\operatorname{Var}(x) = I$, we obtain

$$\operatorname{tr}(A) = \mathbb{E}(x^t A x).$$

The right hand side can be estimated by sample means. It is shown in [13] that if we limit ourselves to real vectors, i.e. $x \in \mathbb{R}^m$, the least variance in $x^t A x$ with symmetric A is obtained by drawing x as *sign vectors*, i.e. uniformly in $\{-1, 1\}^m$. This method is also described in [4, Proposition 4.1].

That is, to compute the bias term δ_θ in (9), we estimate the expectation in:

$$\delta_\theta = \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} D (D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}} x)/n, \tag{12}$$

by sample means. This entails drawing an $x \in \{-1, 1\}^n$, then solve the equation

$$D^t M_{F,X} D v = D^t M_{\mathbf{1}} x, \tag{13}$$

for v by e.g. a conjugate gradient method (CG) like the one in [15, Algorithm 3], and compute $x^t M_{\mathbf{1}} D v$. This will be repeated a number of times and then averaged. The CG method has the advantage that it does not require a matrix representation of the linear operator $D^t M_{F,X} D$, it is sufficient with a procedure for computing the matrix-vector product, like the one in Remark 2.2.

The same method is used to compute the other bias-terms. We then obtain unbiased estimates for the variances and covariances, and may estimate the correlation.

The bias term for $\tilde{\sigma}^2_\psi$ is obtained from the bias term for $\tilde{\sigma}^2_\theta$ by interchanging F and D in (12):

$$\delta_\psi = \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_{\mathbf{1}} x)/n. \tag{14}$$

The bias term for $\tilde{\sigma}_{\theta\psi}$ becomes, from (11):

$$\delta_{\theta\psi} = -\hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_X D (D^t M_X D)^{-1} D^t M_{\mathbf{1}} x)/n. \tag{15}$$

Each sample requires two steps. We draw an $x \in \{-1, 1\}^n$, and solve

$$D^t M_X D v = D^t M_{\mathbf{1}} x,$$

for v. Then we solve

$$F^t M_{D,X} F w = F^t M_X D v,$$

for w. Finally, we compute $x^t M_{\mathbf{1}} F w$.

Remark 5.1. The operators M_D, M_F, $M_{F,X}$, $M_{D,X}$, and M_X are applied repeatedly in the CG iterations described above. The operators M_D and M_F are just centering on the means, i.e. subtraction of the group means. In general, by Remark 2.2, given a vector λ, $M_X\lambda$, $M_{F,X}\lambda$, and $M_{D,X}\lambda$ can be computed by the methods in [10], but unless X contains high dimensional dummy encoded factors, it is wise to use Lemma 2.1 to write $M_{F,X} = M_F M_{M_F X}$, and $M_{D,X} = M_D M_{M_D X}$, i.e. to apply two simpler operators in succession. We can precompute $M_F X$ and $M_D X$, and orthonormalize the columns; if the columns a_i of A are orthonormal, then $M_A\lambda$ is easy to compute: $M_A\lambda = \lambda - \sum_i \langle \lambda, a_i \rangle a_i$, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. That is, applying $M_{F,X}$, $M_{D,X}$ and M_X do not involve the possibly costly iterations of [10]. After orthonormalization we may anyway use that algorithm; with orthogonal columns it will terminate after one iteration. A fast, though numerically unstable, algorithm for orthonormalizing the columns of A, yielding a matrix Y with the same range as A, is $Y = A(L^t)^{-1}$ where L is the Cholesky decomposition of $A^t A = LL^t$. We clearly have $\mathrm{R}(A) = \mathrm{R}(Y)$, and it is readily seen that the columns of Y are orthonormal: $Y^t Y = L^{-1} A^t A (L^t)^{-1} = L^{-1} L L^t (L^t)^{-1} = I$. If $A^t A$ is close to singular a more stable algorithm should be used. However, this happens only if $M_D X$, $M_F X$ or X are close to being column rank deficient, which means that our original system in (1) is close to collinear. Respecifying the model is then probably a better option.

6. Summary

Given the model (1), and OLS estimates $\hat{\theta}$, $\hat{\psi}$, and $\hat{\sigma}^2_\epsilon$. To compute an estimate $\hat{\rho}_{\theta\psi}$ for $\rho_{\theta\psi} = \mathrm{cor}(D\theta, F\psi)$, we compute the biased estimates $\tilde{\sigma}^2_\theta$, $\tilde{\sigma}^2_\psi$, and $\tilde{\sigma}_{\theta\psi}$ as in eqs. (3), (4), and (5). We then estimate bias correction terms (12), (14), and (15) with sample means as in Section 5, with uniformly drawn $x \in \{-1, 1\}^n$:

$$\begin{aligned}
\delta_\theta &= \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} D (D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}} x)/n, \\
\delta_\psi &= \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_{\mathbf{1}} x)/n, \\
\delta_{\theta\psi} &= -\hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_X D (D^t M_X D)^{-1} D^t M_{\mathbf{1}} x)/n.
\end{aligned} \tag{16}$$

The unbiased variances and covariance are:

$$\begin{aligned}
\hat{\sigma}^2_\theta &= \tilde{\sigma}^2_\theta - \delta_\theta \\
\hat{\sigma}^2_\psi &= \tilde{\sigma}^2_\psi - \delta_\psi \\
\hat{\sigma}_{\theta\psi} &= \tilde{\sigma}_{\theta\psi} - \delta_{\theta\psi}
\end{aligned}$$

We then estimate the correlation $\rho_{\theta\psi} = \operatorname{cor}(D\theta, F\psi)$ between $D\theta$ and $F\psi$ as

$$\hat{\rho}_{\theta\psi} = \frac{\hat{\sigma}_{\theta\psi}}{\sqrt{\hat{\sigma}^2_\theta \hat{\sigma}^2_\psi}}.$$

If there are no covariates X, or they are uncorrelated with D and F, the bias corrections in (16) can be replaced with the ones from Section 4:

$$\begin{aligned}
\delta'_\theta &= \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} D (D^t M_F D)^{-1} D^t M_{\mathbf{1}} x)/n, \\
\delta'_\psi &= \delta'_\theta - \hat{\sigma}^2_\epsilon \frac{k_\theta - 1 - k_\psi}{n}, \\
\delta'_{\theta\psi} &= -\delta'_\theta + \hat{\sigma}^2_\epsilon \frac{k_\theta - 1}{n}.
\end{aligned}$$

Alternatively, the symmetric equations can be used:

$$\begin{aligned}
\delta'_\psi &= \hat{\sigma}^2_\epsilon \, \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_D F)^{-1} F^t M_{\mathbf{1}} x)/n, \\
\delta'_\theta &= \delta'_\psi - \hat{\sigma}^2_\epsilon \frac{k_\psi + 1 - k_\theta}{n}, \\
\delta'_{\theta\psi} &= -\delta'_\psi + \hat{\sigma}^2_\epsilon \frac{k_\psi}{n}.
\end{aligned}$$

Remark 6.1. In some cases k_θ or k_ψ are small enough that matrices of such size can be handled, as in [3, Section 5]. In that case δ'_θ or δ'_ψ can be computed directly, e.g. with

$$\delta'_\psi = \hat{\sigma}^2_\epsilon \operatorname{tr}((F^t M_D F)^{-1} F^t M_{\mathbf{1}} F)/n,$$

and we avoid sampling to estimate the trace.

7. Implementation suggestions

Application of the large matrices D, D^t, F, and F^t is simple in the software system R([18]). These sparse matrices are represented as *factors*, and are applied by subsetting for D and F, and by the function 'rowsum()' for their transposes. Similar mechanisms for handling sparse matrices are often available in other software systems. Efficient application of the matrix operators M_D, M_F, $M_{F,X}$, $M_{D,X}$ and M_X is described in Remark 5.1. Also, note that, due to the way we have formulated the computations in section 5, it is actually not necessary to remove a reference group from the factors at all. The CG-algorithm will converge to one of many possible solutions as in [14], but the ambiguity is annihilated by subsequent computations.

The CG method for solving an equation $Ax = b$ can easily be used for multiple vectors at the same time, i.e. with b a matrix of column vectors. The implementation in [11] of the computation $M_A v$ can also use a v which is a matrix, parallelizing over the columns of v, thus speeding up the computation. That is, when sampling the expectations in (12), (14), and (15), we can draw multiple vectors and use a matrix x of column vectors.

Since the expectations used to compute the bias corrections are estimated by sample means, they can be computed to arbitrary precision by taking enough samples. We can monitor the sample standard deviation and stop sampling when a desired relative accuracy in $\hat{\sigma}^2_\theta$ and $\hat{\sigma}^2_\psi$ has been reached. If we estimate $\hat{\sigma}^2_\theta$ and $\hat{\sigma}^2_\psi$ first, we can stop sampling for $\delta_{\theta\psi}$ when a desired absolute accuracy in $\hat{\rho}_{\theta\psi}$ has been reached. Also, the termination criterion for the CG algorithm can be set to finish when a solution just good enough for our expectation tolerance has been found. A suitable termination criterion can be found in [15]. However, computing a too imprecise solution can introduce bias, and also increase the variance in the expectation sampling, so that more samples may be needed.

Preliminary trials with $n = 3 \cdot 10^6, k_\theta = 3 \cdot 10^5, k_\psi = 3 \cdot 10^4$, bias and CG precision at 0.01, a true correlation of 0.3, and a biased correlation of 0.15, reveal that typically the required number of samples for the expectations are ≤ 10, which is 1–3 iterations on a 4-cpu computer, decreasing in n. The time requirements are 2–5 m. However, the number of required samples depends on $\hat{\sigma}^2_\epsilon$, the reason is that the larger $\hat{\sigma}^2_\epsilon$ is, the more accurately we have to estimate the traces to achieve the same precision in $\hat{\sigma}^2_\theta$, $\hat{\sigma}^2_\psi$, and $\hat{\sigma}_{\theta\psi}$. Fortunately, in practice there is typically not much heterogeneity left in (1) after controlling for β, θ and ψ, i.e. the more covariates in X, the smaller σ^2_ϵ. With large n, a case could be made for using only a single sample, this is what we do when we estimate $\theta = \mathbb{E}(\hat{\theta})$ with the "single sample" $\hat{\theta}$. By using 4 samples, relaxing the CG-tolerance so we only do 6 CG-iterations, and assuming the covariates X are independent of D and F, we obtain a correlation within 0.02 of the correct one in about 30 s.

A trial with a real dataset without covariates X, and a biased correlation of $\approx \; -0.18$, with $k_\theta = 2.3 \times 10^6$, $k_\psi = 62000$ and $n = 18 \times 10^6$ takes ≈ 35 m. With relaxed precision it finishes in 4 m, with the corrected correlation differing by 0.03 from the 0.01-precision estimate.

We should however keep in mind that our unbiased estimates $\hat{\sigma}^2_\theta$, $\hat{\sigma}^2_\psi$, and $\hat{\sigma}_{\theta\psi}$ still may be incorrect due to ordinary regression errors in the estimates $\hat{\theta}$, $\hat{\psi}$, and $\hat{\sigma}^2_\epsilon$.

An implementation of these bias corrections will be made available in a future version of [11].

8. References

[1] J. M. Abowd, F. Kramarz, and D. N. Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, March 1999. doi: 10.1111/1468-0262.00020. URL `http://www.jstor.org/stable/2999586`.

[2] J. M. Abowd, R. H. Creecy, and F. Kramarz. Computing person and firm effects using linked longitudinal employer-employee data. Longitudinal Employer-Household Dynamics Technical Papers 2002-06, Center for Economic Studies, U.S. Census Bureau, 2002. URL `http://ideas.repec.org/p/cen/tpaper/2002-06.html`.

[3] M. Andrews, L. Gill, T. Schank, and R. Upward. High wage workers and low wage firms: negative assortative matching or limited mobility bias? *Journal of the Royal Statistical Society(A)*, 171(3): 673–697, 2008. doi: 10.1111/j.1467-985X.2007.00533.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2007.00533.x/full`.

[4] Z. Bai, M. Fahey, and G. Golub. Some large-scale matrix computation problems. *Journal of Computation and Applied Mathematics*, (74):71–89, 1996. doi: 0.1016/0377-0427(96)00018-0. URL `http://www.sciencedirect.com/science/article/pii/0377042796000180`.

[5] C. Bartolucci and F. Devicienti. Better workers move to better firms: A simple test to identify sorting. IZA Discussion Paper 7601, Bonn, 2013. URL `http://hdl.handle.net/10419/90077`.

[6] D. Card, J. Heining, and P. Kline. Workplace heterogeneity and the rise of West German wage inequality*. *The Quarterly Journal of Economics*, 128(3):967–1015, 2013. doi: 10.1093/qje/qjt006. URL `http://qje.oxfordjournals.org/content/128/3/967.abstract`.

[7] A. Carneiro, P. Guimarães, and P. Portugal. Real wages and the business cycle: Accounting for worker, firm and job title heterogeneity. *American Economic Journal: Macroeconomics*, 4 (2):133–152, April 2012. doi: doi:10.1257/mac.4.2.133. URL `http://www.ingentaconnect.com/content/aea/aejma/2012/00000004/00000002/art00005`.

[8] T. Cornelißen and O. Hübler. Unobserved individual and firm heterogeneity in wage and job-duration functions: Evidence from german linked employer–employee data. *German Economic Review*, 12(4):469–489, 2011. ISSN 1468-0475. doi: 10.1111/j.1468-0475.2010.00528.x. URL `http://dx.doi.org/10.1111/j.1468-0475.2010.00528.x`.

[9] C. Davidson, F. Heyman, S. Matusz, F. Sjöholm, and S. Chun Zhu. Globalization and imperfect labor market sorting. IFN Working Paper 856, Stockholm, 2010. URL `http://hdl.handle.net/10419/81434`.

[10] S. Gaure. OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis*, 66:8–18, 2013. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2013.03.024. URL `http://www.sciencedirect.com/science/article/pii/S0167947313001266`.

[11] S. Gaure. *lfe: Linear group fixed effects*, 2013. URL `http://CRAN.R-project.org/package=lfe`. R package version 1.6.

[12] J.R. Graham, S. Li, and J. Qiu. Managerial attributes and executive compensation. *Review of Financial Studies*, 25(1):144–186, 2012. doi: 10.1093/rfs/hhr076. URL `http://rfs.oxfordjournals.org/content/25/1/144.abstract`.

[13] M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806. URL `http://dx.doi.org/10.1080/03610918908812806`.

[14] E.F. Kaasschieter. Preconditioned conjugate gradients for solving singular systems. *Journal of Computational and Applied Mathematics*, 24(12):265 – 275, 1988. ISSN 0377-0427. doi: http://dx.doi.org/10.1016/0377-0427(88)90358-5. URL `http://www.sciencedirect.com/science/article/pii/0377042788903585`.

[15] E.F. Kaasschieter. A practical termination criterion for the conjugate gradient method. *BIT Numerical Mathematics*, 28(2):308–322, 1988. ISSN 0006-3835. doi: 10.1007/BF01934094. URL `http://dx.doi.org/10.1007/BF01934094`.

[16] R. Mendes, G.J.v.d. Berg, and M. Lindeboom. An empirical assessment of assortative matching in the labor market. *Labour Economics*, 17(6):919 – 929, 2010. ISSN 0927-5371. doi: http://dx.doi.org/10.1016/j.labeco.2010.05.001. URL `http://www.sciencedirect.com/science/article/pii/S0927537110000618`.

[17] A. Ouazad. A2REG: Stata module to estimate models with two fixed effects. Statistical Software Components, Boston College Department of Economics, 2008. URL `http://econpapers.repec.org/RePEc:boc:bocode:s456942`.

[18] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[19] J.F. Schmieder. GPREG: Stata module to estimate regressions with two dimensional fixed effects. Statistical Software Components, Boston College Department of Economics, May 2009. URL `http://ideas.repec.org/c/boc/bocode/s457048.html`.

[20] T. Sørensen and R. Vejlin. The importance of worker, firm and match effects in the formation of wages. *Empirical Economics*, 45(1):435–464, 2013. ISSN 0377-7332. doi: 10.1007/s00181-012-0603-3. URL `http://dx.doi.org/10.1007/s00181-012-0603-3`.