Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

Correlation bias correction in two-way fixed-effects linear regression

Simen Gaure*

Received 23 October 2014; Accepted 12 November 2014

When doing two-way fixed-effects ordinary least squares 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. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: limited mobility bias; linear regression; two-way fixed effects

1 Introduction

We consider a linear 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, β is a vector of parameters. D is an $n \times k_\theta$ matrix resulting from dummy encoding a factor, with parameter vector θ. F is an $n \times k_\psi$ matrix resulting from dummy encoding another factor, with parameter vector ψ, and ϵ is a normally distributed error term. This is a perfectly ordinary least squares (OLS) system, but our assumption is that k_θ and k_ψ are large, for example, of the order 10^5–10^7. This creates some computational challenges.

Remark 1.1
Note that the phrase 'fixed effects' are used slightly differently in statistics and in econometrics. When using OLS, every parameter is of course a fixed effect, as opposed to a random effect, but it is customary in some econometric circles to refer to the θ and ψ in (1) as the fixed effects, that is, time-constant individual effects.

The canonical example in the panel data econometrics literature of this kind of model can be found in Abowd et al. (1999), 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 $\mathrm{cor}(D\theta, F\psi)$ as a way to investigate whether 'high-wage' workers tend to work in 'high-wage' firms.

Some authors (e.g. Card et al., 2013) do variance decompositions of the form

$$\mathrm{var}(y) = \mathrm{var}(X\beta) + \mathrm{var}(D\theta) + \mathrm{var}(F\psi) + 2\mathrm{cov}(D\theta, F\psi) + 2\mathrm{cov}(\cdots) + \cdots$$

to use changes in the decomposition over time to study wage inequality trends.

The Ragnar Frisch Centre for Economic Research, 0349 Oslo, Norway
***Email: Simen.Gaure@frisch.uio.no**

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

The ISI's Journal for the Rapid Dissemination of Statistics Research

We assume that β, θ, ψ, and ϵ are estimated by OLS, for example, with the methods of Carneiro et al. (2012), Gaure (2013b), Ouazad (2008), and Schmieder (2009). Andrews et al. (2008) show that the variances $\tilde{\sigma}_\theta^2 = \mathrm{var}(D\hat{\theta})$ and $\tilde{\sigma}_\psi^2 = \mathrm{var}(F\hat{\psi})$ are positively biased, and that the covariance $\tilde{\sigma}_{\theta\psi} = \mathrm{cov}\left(D\hat{\theta}, F\hat{\psi}\right)$ 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: $\tilde{\rho}_{\theta\psi} = \tilde{\sigma}_{\theta\psi}(\tilde{\sigma}_\theta^2\tilde{\sigma}_\psi^2)^{-1/2}$. This particular type of bias is known as *limited mobility* bias.

A challenge with the bias correction formulas of Andrews et al. (2008) is that they involve the inverses of large square matrices, of sizes 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. Card et al., 2013; Cornelißen & Hübler, 2011; Davidson et al., 2010; Graham et al., 2012; Sørensen & Vejlin, 2013). We therefore venture to evaluate the bias correction expressions 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 in Section 3.

In applications, there can be other sources of bias than the one corrected by the methods presented here, for example, related to endogenous selection. To solve such bias problems, other models could be used, as in Bartolucci & Devicienti (2013) and Mendes et al. (2010). It is also pointed out in the literature, for example by Card et al. (2013), that the OLS assumption of independently identically distributed (i.i.d.) errors is dubious in some applications. Adaptation of the bias expressions to heteroscedastic and clustered residuals is sketched in Section 4. The results of some trial runs are reported in the Appendix.

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 and vectors are of the appropriate size. The letters A and B are used to denote general matrices and have no fixed meaning. The letter Q is used to temporarily name particular matrices for clarity. The letter m is used for an arbitrary natural number.

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 $[\ A\ B\ ]$. 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 that subtracts the mean. For a vector x, we denote by $\mathrm{d}(x)$ the diagonal matrix with x on the diagonal. We will now and then use the defining property of the trace, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, without mentioning. This includes 'doubling' of projections: $\mathrm{tr}(AM_B) = \mathrm{tr}(M_BAM_B)$. We use $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot,\cdot)$ to denote the numerical sample variance and covariance of sequences of numbers. The capitalized versions $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot,\cdot)$ are used for variance and covariance matrices of random variables.

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 collinearities in the system; in the language of Abowd et al. (2002), there is a single *connected group*, or *connected component*. These assumptions are necessary for identification of $\hat{\theta}$ and $\hat{\psi}$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

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 do *not* assume that X is small; that is, X may, among other covariates, contain one or more high-dimensional dummy-encoded factors, as in Carneiro et al. (2012).

Remark 2.1

Given a vector v, we note that Gaure (2013b, Algorithm 3.1) gives a procedure by which we can compute $M_{D,F}v$. It is not mentioned explicitly by Gaure (2013b) 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 are the same, but the actual computation of each projection of Gaure (2013b, Algorithm 3.1(2) and (15)) corresponding to columns of A, is slightly more complicated. Such a procedure has been implemented by Gaure (2013a) through implementation of interactions between factors and continuous covariates; one may use a factor with a single level. In the present paper, there is no intrinsic dependence on D and F being dummy-encoded factors; most of the theory is the same if D and F are interactions between factors and covariates, or something else; but the author knows of no such application.

The following lemma will come in handy later.

Lemma 2.2

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

Proof

First, we note that $M_A(I - M_{A,B})$ is a projection. Now, let $P = I - M_{M_AB}$. P is the projection onto $\mathrm{R}(M_AB)$, that is, $\mathrm{R}(P) = \mathrm{R}(M_AB)$. Note that $\mathrm{R}(M_AB)$ is spanned by the columns of M_AB, that is, $\mathrm{R}(M_AB) = M_A\mathrm{R}(B)$. We also have that the columns of $I - M_{A,B}$ span $R(A) + R(B)$. So that $\mathrm{R}(M_A(I - M_{A,B})) = M_A(\mathrm{R}(A) + \mathrm{R}(B)) = M_A\mathrm{R}(B) = \mathrm{R}(P)$. Two projections with the same range are equal, so $P = M_A(I - M_{A,B})$. That is, $I - M_{M_AB} = M_A - M_{A,B}$. 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 a full-column rank, we have from (2) that $M_{M_AB} = I - M_AB(B^tM_AB)^{-1}B^tM_A$. □

3 Variance and covariance bias

When deriving the bias correction formulas, we will follow the exposition by Andrews et al. (2008) but change the notation to reflect our emphasis on the projections of the type M_A, which we can compute. For the asymptotic statistical properties, we refer the reader to their paper. As in Andrews et al. (2008, (8–10)), we have biased sample estimates for the variance of $D\theta$ and $F\psi$ and their covariance:

$$\tilde{\sigma}_\theta^2 = \mathrm{var}\left(D\hat{\theta}\right) = \frac{\hat{\theta}^tD^tM_1D\hat{\theta}}{n}, \tag{3}$$

$$\tilde{\sigma}_\psi^2 = \mathrm{var}\left(F\hat{\psi}\right) = \frac{\hat{\psi}^tF^tM_1F\hat{\psi}}{n}, \tag{4}$$

$$\tilde{\sigma}_{\theta\psi} = \mathrm{cov}\left(D\hat{\theta}, F\hat{\psi}\right) = \frac{\hat{\theta}^tD^tM_1F\hat{\psi}}{n}. \tag{5}$$

We take the expectation of (3) as in Andrews et al. (2008, (16)):

$$\mathbb{E}\left(\frac{\hat{\theta}^t D^t M_{\mathbf{1}} D\hat{\theta}}{n}\right) = \frac{\theta^t D^t M_{\mathbf{1}} D\theta}{n} + \delta_\theta, \tag{6}$$

where the bias term

$$\delta_\theta = \frac{\mathrm{tr}\left(D^t M_{\mathbf{1}} D \mathrm{Var}\left(\hat{\theta}\right)\right)}{n},$$

is found by using the general formula for the expectation of a quadratic form

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

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

Our interest is in the term

$$\sigma_\theta^2 = \frac{\theta^t D^t M_{\mathbf{1}} D\theta}{n}.$$

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 δ_θ, we need an expression for $\mathrm{Var}(\hat{\theta})$. 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 $D\hat{\theta} \mapsto D\hat{\theta} - \alpha\mathbf{1}, F\hat{\psi} \mapsto F\hat{\psi} + \alpha\mathbf{1}$, for some $\alpha \in \mathbb{R}$. But we have $M_{\mathbf{1}}\mathbf{1} = 0$. That is, in for example, (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 $\mathrm{Var}(\hat{\theta})$ via the Frisch–Waugh–Lovell theorem. By multiplying through (1) with $M_{F,X}$ and using standard OLS assumptions, including the i.i.d. assumption $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2 I$, we obtain

$$\mathrm{Var}\left(\hat{\theta}\right) = \sigma_\epsilon^2 (D^t M_{F,X} D)^{-1}. \tag{8}$$

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

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

It is the $k_\theta \times k_\theta$ matrix inside the trace term that may be too large to be handled directly, as in Andrews et al. (2008, p. 687). By symmetry between θ and ψ, the corresponding bias term for $\tilde{\sigma}_\psi^2$ is

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

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, $(x+y)^t A(x+y) = x^t A x + 2x^t A y + y^t A y$. By taking expectations and using (7), we obtain

$$\mathbb{E}(x^t A y) = \mathbb{E}(x^t) A \mathbb{E}(y) + \mathrm{tr}(A\mathrm{Cov}(x,y)). \tag{11}$$

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

We use (11) on (5), with $x = D\hat{\theta}, y = F\hat{\psi}$, and $A = M_1$. An algebraic excursion yields

$$\mathbb{E}\left(\tilde{\sigma}_{\theta\psi}\right) = \frac{\theta^t D^t M_1 F \psi}{n} + \frac{\sigma_\epsilon^2}{n}\mathrm{tr}(M_1 D(D^t M_{F,X} D)^{-1} D^t M_{F,X} M_{D,X} F (F^t M_{D,X} F)^{-1} F^t).$$

As in Andrews et al. (2008), we can use Lemma 2.2 to write $M_{D,X} = M_X(I - D(D^t M_X D)^{-1} D^t M_X)$. We then obtain $M_{F,X} M_{D,X} F = -M_{F,X} D(D^t M_X D)^{-1} D^t M_X F$ and rewrite the trace term as $-\mathrm{tr}(M_1 D(D^t M_X D)^{-1} D^t M_X F (F^t M_{D,X} F)^{-1} F^t)$, or use the transposed version as in Andrews et al. (2008, (22)), so that the covariance bias can be written as

$$\delta_{\theta\psi} = -\frac{\sigma_\epsilon^2}{n}\mathrm{tr}(D^t M_1 F (F^t M_{D,X} F)^{-1} F^t M_X D (D^t M_X D)^{-1}). \tag{12}$$

4 Heteroscedastic and clustered residuals

If the residuals ϵ do not satisfy $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2 I$, our formulas will be different. The simplification leading to (8) is no longer valid. Instead, we obtain

$$\mathrm{Var}\left(\hat{\theta}\right) = R V_\epsilon R^t,$$

where $V_\epsilon = \mathrm{Var}(\epsilon)$ and $R = (D^t M_{F,X} D)^{-1} D^t M_{F,X}$. This yields the following bias correction formulas:

$$\delta'_\theta = \mathrm{tr}(M_1 D R V_\epsilon R^t D^t M_1)/n, \tag{13}$$

$$\delta'_\psi = \mathrm{tr}(M_1 F S V_\epsilon S^t F^t M_1)/n, \tag{14}$$

$$\delta'_{\theta\psi} = \mathrm{tr}(M_1 D R V_\epsilon S^t F^t M_1)/n, \tag{15}$$

where $S = (F^t M_{D,X} F)^{-1} F^t M_{D,X}$.

The simplest version of clustering is when there are n groups, one for each observation. This is the heteroscedasticity assumption. That is, the residuals are independent, but with different variances. An estimate for V_ϵ is the diagonal matrix with the squared residuals on the diagonal (White, 1980):

$$\hat{V}_\epsilon^h = \mathrm{d}\,(\hat{\epsilon})^2\,.$$

With fewer groups in the clustering, we describe them with a matrix C, which is the dummy encoding of the clustering categorical variable. Following Cameron et al. (2011, (2.4)), we can use the estimate

$$\hat{V}_\epsilon^c = \mathrm{d}\,(\hat{\epsilon})\, C C^t \mathrm{d}\,(\hat{\epsilon})\,.$$

We see that the heteroscedastic case corresponds to $C = I$. With multiway clustering, the alternating sum over one-way clusters in the work of Cameron et al. (2011, (2.13)) can be used.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

The ISI's Journal for the Rapid Dissemination of Statistics Research

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 (12) 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 $\mathrm{Var}(x) = I$, we obtain

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

The right-hand side can be estimated by sample means. It was shown by Hutchinson (1989) that if we limit ourselves to real vectors, that is, $x \in \mathbb{R}^m$, the least variance in $x^t A x$ with symmetric A is obtained by drawing x as *sign vectors*, that is, uniformly in $\{-1, 1\}^m$. This method is also described by Bai et al. (1996, Proposition 4.1), and, for complex $x \in \mathbb{C}^m$ by Iitaka & Ebisuzaki (2004).

That is, to compute the bias term δ_θ in (9), we estimate σ_ϵ^2, and the expectation in

$$\delta_\theta = \sigma_\epsilon^2 \mathbb{E}(x^t M_{\mathbf{1}} D (D^t M_{F,X} D)^{-1} D^t M_{\mathbf{1}} x)/n, \tag{16}$$

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

$$D^t M_{F,X} D v = D^t M_{\mathbf{1}} x, \tag{17}$$

for v, and computing $x^t M_{\mathbf{1}} D v$. Solving (17) can be performed, for example, with a conjugate gradient method (CG) like the one described by Kaasschieter (1988, Algorithm 3). 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 2.1.

The same method is used to compute the other bias terms. The bias term for $\tilde{\sigma}_\psi^2$ is obtained from the bias term for $\tilde{\sigma}_\theta^2$ by interchanging F and D in (16):

$$\delta_\psi = \sigma_\epsilon^2 \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_{\mathbf{1}} x)/n. \tag{18}$$

The bias term for $\tilde{\sigma}_{\theta\psi}$ becomes, from (12),

$$\delta_{\theta\psi} = -\sigma_\epsilon^2 \mathbb{E}(x^t M_{\mathbf{1}} F (F^t M_{D,X} F)^{-1} F^t M_X D (D^t M_X D)^{-1} D^t M_{\mathbf{1}} x)/n. \tag{19}$$

Each sample in (19) 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

$$F^t M_{D,X} F w = F^t M_{\mathbf{1}} x,$$

for w. Finally, we compute $w^t F^t M_X D v$.

As usual, σ_ϵ^2 can be estimated from the residuals $\hat{\epsilon}$ when solving (1) by OLS. With the typically large number of observations in these kind of models, the estimate $\hat{\sigma}_\epsilon^2$ will be a very good estimate of σ_ϵ^2. For the heteroscedastic and cluster robust corrections in Section 4, the computations are similar, but note that the formulas for δ'_θ and δ'_ψ are of

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

the form $\mathbb{E}(x^tQ^tQx)$, so rather than solving two large equations for each, we can get away with one, that is, compute $z = Qx$ and z^tz. Indeed, if sampling of the traces in Section 4 is carried out simultaneously (with the same x) for the three bias corrections, the vectors $R^tD^tM_{\mathbf{1}}x$ and $S^tF^tM_{\mathbf{1}}x$ from (13) and (14) can be reused in (15), so that no separate CG iterations need to be run for $\delta'_{\theta\psi}$. This, of course, also holds for the i.i.d. expression $\delta_{\theta\psi}$ in Section 3 provided those expressions are written symmetrically. However, such simultaneous sampling may introduce correlation between the bias estimates.

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 earlier. The operators M_D and M_F are just centring on the means, that is, subtraction of the group means. In general, by Remark 2.1, given a vector $\lambda, M_X\lambda, M_{F,X}\lambda$, and $M_{D,X}\lambda$ can be computed by the methods of Gaure (2013b), but unless X contains high-dimensional dummy-encoded factors or is otherwise too large, we can use Lemma 2.2 to write $M_{F,X} = M_FM_{M_FX}$ and $M_{D,X} = M_DM_{M_DX}$, that is, apply two simpler operators in succession. We can precompute M_FX and M_DX and orthonormalize the columns; if the columns a_i of a matrix 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 Gaure (2013b). After orthonormalization, we may anyway use that algorithm; with orthogonal columns, it will terminate after one iteration. A fast, but numerically unstable algorithm for orthonormalizing the columns of A, yielding a matrix B with the same range as A, is $B = A(L^t)^{-1}$ where L is the Cholesky decomposition of $A^tA = LL^t$. We clearly have $\mathrm{R}(A) = \mathrm{R}(B)$, so that $M_A = M_B$, and it is readily seen that the columns of B are orthonormal: $B^tB = L^{-1}A^tA(L^t)^{-1} = L^{-1}LL^t(L^t)^{-1} = I$. If A^tA is close to singular, a more stable algorithm should be used. However, in our setting, this happens only if M_DX, M_FX, or X is 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 $\hat\sigma^2_\epsilon$, an estimate $\hat\rho_{\theta\psi}$ for the correlation $\rho_{\theta\psi} = \mathrm{cor}(D\theta, F\psi)$ can be found by computing the biased estimates $\tilde\sigma^2_\theta, \tilde\sigma^2_\psi$, and $\tilde\sigma_{\theta\psi}$ as in (3), (4), and (5). We then estimate bias correction terms (16), (18), and (19) with sample means as in Section 5, with uniformly drawn $x \in \{-1, 1\}^n$:

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

$$\hat\delta_\psi = \hat\sigma^2_\epsilon\hat{\mathbb{E}}(x^tM_{\mathbf{1}}F(F^tM_{D,X}F)^{-1}F^tM_{\mathbf{1}}x)/n, \tag{20}$$

$$\hat\delta_{\theta\psi} = -\hat\sigma^2_\epsilon\hat{\mathbb{E}}(x^tM_{\mathbf{1}}F(F^tM_{D,X}F)^{-1}F^tM_XD(D^tM_XD)^{-1}D^tM_{\mathbf{1}}x)/n,$$

where we use $\hat{\mathbb{E}}$ to denote a sample mean. The bias-corrected variances and covariance are

$$\hat\sigma^2_\theta = \tilde\sigma^2_\theta - \hat\delta_\theta,$$

$$\hat\sigma^2_\psi = \tilde\sigma^2_\psi - \hat\delta_\psi,$$

$$\hat\sigma_{\theta\psi} = \tilde\sigma_{\theta\psi} - \hat\delta_{\theta\psi}.$$

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

The ISI's Journal for the Rapid Dissemination of Statistics Research

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

$$\hat{\rho}_{\theta\psi} = \hat{\sigma}_{\theta\psi}\left(\hat{\sigma}_\theta^2\hat{\sigma}_\psi^2\right)^{-1/2}.$$

For heteroscedastic and clustered residuals, estimated bias corrections $\hat{\delta}'_\theta, \hat{\delta}'_\psi$, and $\hat{\delta}'_{\theta\psi}$ can be found by using the expressions in Section 4.

Appendix A: Some trial runs

The method for bias correction described earlier has been implemented with the tools available in the work of Gaure (2013a). To illustrate that the method works in practice, in reasonable time, we present the result of some trial runs of the algorithm described in Section 5.

A.1. Implementation

Application of the large matrices C, C^t, D, D^t, F, and F^t is simple in the software system R (R Core Team, 2014). These sparse matrices are represented as *factors* and are applied by subsetting for C, 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.

Because 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 monitor the standard deviation of the sample mean and stop sampling when a desired relative accuracy in $\hat{\sigma}_\theta^2$ and $\hat{\sigma}_\psi^2$ has been reached. We estimate $\hat{\sigma}_\theta^2$ and $\hat{\sigma}_\psi^2$ first, with a relative tolerance. We stop sampling for $\delta_{\theta\psi}$ when a desired absolute accuracy in $\hat{\rho}_{\theta\psi}$ has been reached.

The termination criterion for the CG algorithm is set so that the iterations finish when a solution good enough for our expectation tolerance has been found. We use the termination criterion of Kaasschieter (1988). Computing a too imprecise solution can introduce bias and also increase the variance in the expectation sampling, so that more samples may be needed. The number of required CG iterations depends on the spectrum of A and is therefore data dependent. In short, the tolerances we choose for the sample means and the CG iterations can have a substantial impact on the timing, and the choice depends on what we intend to use the estimates for.

A.2. Datasets

Because datasets of such a large size typically are non-public for either commercial or data protection reasons, we have created some datasets by randomly drawing observations. We have used the model

$$y_{ijt} = x_1 + x_2 + \theta_i + \psi_j + \epsilon_{ijt}, \tag{A.1}$$

where ψ_j plays the role of time-constant firm effects and θ_i plays the role of time-constant individual effects. The number of observations for each individual ranges uniformly from five to seven, and the observation period is exogenously drawn. The initial size of the firms is drawn from a $\chi^2(k_\theta/k_\psi)$ distribution to obtain a variation in firm size. The ψ_j's and θ_i's are initially drawn from normal distributions with zero expectation and unit variance. Now and then, individual i 'changes job', in such a way that the probability of picking a firm j depends monotonically on $|\theta_i - \psi_j|$. Thus, we create a correlation between $D\theta$ and $F\psi$, where D and F are as before, dummy-encoded individuals and firms. When we have drawn all job changes, that is, D and F have been constructed, we select the largest connected component,

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

which in all cases consists of more than 99% of the data. Then, θ and ψ are linearly scaled so that $\text{var}(D\theta) = 8$ and $\text{var}(F\psi) = 2$. The covariate x_1 is drawn from a normal distribution, but with some correlation with θ and ψ. Similarly with x_2,

$$x_1 \sim \mathcal{N}(0,1) + 0.1\theta_i + 0.9\psi_j,$$

$$x_2 \sim \mathcal{N}(0,1) + 0.2x_1 - 0.9\theta_i + 0.2\psi_j.$$

The residual ϵ is drawn from a normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$.

Loosely speaking, the biases we are studying come from two sources. Andrews et al. (2008) suggest that the important source for the correlation bias is *limited mobility*, which they operationalize as *number of movers per firm*. That is, given a firm, how many of its employees have also worked for another firm in the observation period. Their specific trials may suggest that the correlation bias is below 0.02 when the number of movers per firm exceeds ≈15. However, as they point out, there is also another source of bias, namely σ_ϵ^2. This makes it hard to give general guidelines for the size of the biases; they depend on both the mobility and the residual variance. We therefore make some datasets by varying both the mobility and σ_ϵ^2. Our mobility parameter is a hazard, that is, the probability per observation period of changing a job. Our low-mobility datasets are described in Table A.1, where we report the sizes and the biased variances and covariances together with the correlation bias. We have created four small datasets and two large ones. The large ones are of the size studied by Card et al. (2013). The low-mobility datasets have $\sigma_\theta^2 = 8, \sigma_\psi^2 = 2$, and $\sigma_{\theta\psi} \approx 0.8$; thus, $\rho_{\theta\psi} \approx 0.2$. The mobility hazard is 0.0623 for the small datasets and 0.0447 for the two large ones. The integers n, k_θ, and k_ψ are in units of 10^5. The column $\tilde{\delta}_\rho$ is the correlation error $\rho_{\theta\psi} - \tilde{\rho}_{\theta\psi}$, due to the bias and finite sample errors in the estimates $\tilde{\sigma}_\theta^2, \tilde{\sigma}_\psi^2$, and $\tilde{\sigma}_{\theta\psi}$. Because of the data generation process, we do not manage to keep the covariance $\sigma_{\theta\psi}$ entirely constant over the datasets. The actual $\sigma_{\theta\psi}$ is therefore also tabulated.

Table A.1. Description of low-mobility datasets.

Name	n	k_θ	k_ψ	σ_ϵ^2	$\tilde{\sigma}_\theta^2$	$\tilde{\sigma}_\psi^2$	$\tilde{\sigma}_{\theta\psi}$	$\tilde{\rho}_{\theta\psi}$	$\tilde{\delta}_\rho$	$\sigma_{\theta\psi}$
l1	60	10	1	0.1	8.034	2.020	0.782	0.194	0.006	0.800
l2	60	10	1	1	8.352	2.207	0.615	0.143	0.059	0.806
l3	60	10	1	8	10.811	3.595	−0.669	−0.107	0.308	0.802
l4	60	10	1	20	15.030	6.004	−2.887	−0.304	0.505	0.804
l5	1200	200	15	1	8.366	2.213	0.600	−0.139	0.061	0.800
l6	1200	200	15	20	15.341	6.254	−3.216	−0.328	0.528	0.800

Table A.2. Description of high-mobility datasets.

Name	n	k_θ	k_ψ	σ_ϵ^2	$\tilde{\sigma}_\theta^2$	$\tilde{\sigma}_\psi^2$	$\tilde{\sigma}_{\theta\psi}$	$\tilde{\rho}_{\theta\psi}$	$\tilde{\delta}_\rho$	$\sigma_{\theta\psi}$
h1	60	10	1	0.1	8.018	2.007	0.831	0.207	0.002	0.836
h2	60	10	1	1	8.217	2.067	0.789	0.191	0.018	0.835
h3	60	10	1	8	9.677	2.513	0.459	0.093	0.116	0.834
h4	60	10	1	20	12.276	3.289	−0.101	−0.016	0.224	0.832
h5	1200	200	15	1	8.225	2.073	0.771	0.187	0.021	0.831
h6	1200	200	15	20	12.530	3.435	−0.360	−0.055	0.263	0.832

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

The ISI's Journal for the Rapid Dissemination of Statistics Research

Table A.3. Timing of low-mobility datasets.

Name	$\hat{\sigma}^2_\theta$	$\hat{\sigma}^2_\psi$	$\hat{\sigma}_{\theta\psi}$	$\hat{\rho}_{\theta\psi}$	$\hat{\delta}_\rho$	K	Time
l1	8.016	2.003	0.792	0.198	0.002	4	117
l2	8.000	2.007	0.799	0.199	0.003	4	451
l3	7.995	1.990	0.812	0.204	−0.004	4	794
l4	8.005	2.004	0.793	0.198	0.003	4	923
l5	7.999	2.002	0.800	0.200	0.000	1	4321
l6	7.988	2.005	0.798	0.199	0.001	1	7940

Table A.4. Timing of high-mobility datasets.

Name	$\hat{\sigma}^2_\theta$	$\hat{\sigma}^2_\psi$	$\hat{\sigma}_{\theta\psi}$	$\hat{\rho}_{\theta\psi}$	$\hat{\delta}_\rho$	K	Time
h1	8.000	2.003	0.834	0.208	0.001	4	74
h2	8.003	2.005	0.833	0.208	0.001	4	200
h3	7.968	2.004	0.837	0.209	−0.001	4	554
h4	8.002	2.027	0.845	0.210	−0.002	4	662
h5	7.999	2.002	0.829	0.207	0.001	1	2369
h6	8.005	1.995	0.832	0.208	0.000	1	4758

Our high-mobility datasets can be found in Table A.2. They have $\sigma^2_\theta = 8, \sigma^2_\psi = 2$, and $\sigma_{\theta\psi} \approx 0.83$; thus, $\rho_{\theta\psi} \approx 0.208$. The mobility hazard is 0.175 for the small datasets and 0.113 for the two large ones.

A.3. Timing

The trials have been run on a Dell M520 sporting two octacore Intel Xeon E5-2470 CPUs running at 2.3 GHz, with 192 GiB of 1333 MHz DDR3 memory. Installed OS was Ubuntu Linux 14.04.1 LTS. R version 3.1.1 was used. The computer was otherwise idle during the trial runs. We used a single CPU for our trials, that is, no parallelization. Independent sampling is otherwise quite well suited for parallelization. For the bias corrections, we tried to ensure a maximum relative error of 1% for the variances, and a maximum absolute error of 0.01 for the correlation, that is, an absolute covariance tolerance of ≈0.04. It turned out that very few samples needed to be taken for the trace estimations. For the large datasets, l5, l6, h5, and h6, a single sample yielded more than enough precision. All our estimates are well within the tolerance we have chosen. In Table A.3, we report the bias-corrected estimates together with the number of samples (K) and elapsed time in seconds. The column $\hat{\delta}_\rho$ is the correlation error $\rho_{\theta\psi} - \hat{\rho}_{\theta\psi}$ after bias correction. Note that this is not the definitive word on how fast the method is, as implementation details also matter, but it is an illustration that it runs in reasonable time for quite large datasets.

The timing of the high-mobility datasets can be found in Table A.4. We see that the bias corrections are generally faster to perform for these datasets. This is due to fewer required CG iterations. The required number of CG iterations depends on the spectrum of the operator. It is not surprising that the eigenvalues have a more amenable structure in the better-identified high-mobility datasets.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

References

Abowd, JM, Creecy, RH & Kramarz, F (2002), '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.

Abowd, JM, Kramarz, F & Margolis, DN (1999), 'High wage workers and high wage firms', *Econometrica*, **67**(2), 251–333.

Andrews, M, Gill, L, Schank, T & Upward, R (2008), 'High wage workers and low wage firms: negative assortative matching or limited mobility bias?' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**(3), 673–697.

Bai, Z, Fahey, M & Golub, G (1996), 'Some large-scale matrix computation problems', *Journal of Computation and Applied Mathematics*, **74**, 71–89.

Bartolucci, C & Devicienti, F (2013), 'Better workers move to better firms: a simple test to identify sorting', IZA Discussion Paper 7601, Bonn.

Cameron, AC, Gelbach, JB & Miller, DL (2011), 'Robust inference with multiway clustering', *Journal of Business & Economic Statistics*, **29**(2), 238–249.

Card, D, Heining, J & Kline, P (2013), 'Workplace heterogeneity and the rise of West German wage inequality', *The Quarterly Journal of Economics*, **128**(3), 967–1015.

Carneiro, A, Guimarães, P & Portugal, P (2012), 'Real wages and the business cycle: accounting for worker, firm and job title heterogeneity', *American Economic Journal: Macroeconomics*, **4**(2), 133–152.

Cornelißen, T & Hübler, O (2011), '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.

Davidson, C, Heyman, F, Matusz, S, Sjöholm, F & Zhu, SC (2010), 'Globalization and imperfect labor market sorting', IFN Working Paper 856, Stockholm.

Gaure, S (2013a), lfe: linear group fixed effects. R package version 1.6.

Gaure, S (2013b), 'OLS with multiple high dimensional category variables', *Computational Statistics & Data Analysis*, **66**, 8–18.

Graham, J, Li, S & Qiu, J (2012), 'Managerial attributes and executive compensation', *Review of Financial Studies*, **25**(1), 144–186.

Hutchinson, M (1989), 'A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines', *Communications in Statistics—Simulation and Computation*, **18**(3), 1059–1076.

Iitaka, T & Ebisuzaki, T (2004), 'Random phase vector for calculating the trace of a large matrix', *Physical Review E*, **69**, 057701.

Kaasschieter, E (1988), 'A practical termination criterion for the conjugate gradient method', *BIT Numerical Mathematics*, **28**(2), 308–322.

Mendes, R, Berg, G & Lindeboom, M (2010), 'An empirical assessment of assortative matching in the labor market', *Labour Economics*, **17**(6), 919–929.

Ouazad, A (2008), A2REG: stata module to estimate models with two fixed effects, Statistical Software Components, Boston College Department of Economics.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.68

The ISI's Journal for the Rapid Dissemination of Statistics Research

R Core Team (2014), R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Schmieder, J (2009), GPREG: stata module to estimate regressions with two dimensional fixed effects, Statistical Software Components, Boston College Department of Economics.

Sørensen, T & Vejlin, R (2013), 'The importance of worker, firm and match effects in the formation of wages', *Empirical Economics*, **45**(1), 435–464.

White, H (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica*, **48**(4), 817–838.