Online Appendix

10 October 2014
Revised, 11 January 2016

FIXED EFFECTS, INVARIANCE, AND SPATIAL VARIATION IN INTERGENERATIONAL MOBILITY

Gary Chamberlain*
Harvard University

ABSTRACT

Chetty, Hendren, Kline, and Saez (2014) and Chetty and Hendren (2015) document variation across commuting zones in intergenerational mobility. With over 700 commuting zones, the task of estimating place effects involves a high-dimension parameter space. I consider a class of invariant estimators in a multivariate model. An infeasible oracle estimator provides a lower bound on risk. The oracle corresponds to an invariant prior distribution, which I use to construct a feasible random-effects estimator. In the univariate (scalar outcome) case, this estimator is in the James and Stein (1961) family of estimators. Its risk function is evaluated using the parameter space for the original, fixed-effects model. In the univariate case, this risk function depends only on a scalar noncentrality parameter. In the empirical application, the random-effects estimator almost achieves the oracle bound on risk over the relevant part of the parameter space, and there is a substantial improvement over the risk of the least-squares estimator.

Keywords: fixed effects, place effects, optimal invariant estimator, minimax, random effects, James and Stein

* I thank Raj Chetty, Nathan Hendren, and Maximilian Kasy for helpful discussions.

FIXED EFFECTS, INVARIANCE, AND SPATIAL VARIATION IN INTERGENERATIONAL MOBILITY

1. INTRODUCTION

I shall work with a multivariate normal linear model:

$$Y \mid x \sim \mathcal{N}(x\beta, I_N \otimes \Sigma),$$

where Y is $N \times M$, x is $N \times K$, β is $K \times M$, and Σ is a $M \times M$ positive-definite matrix. I am interested in applications with large K. For example, Y_{ij} could be the score for student i on test j. The students can be grouped in various ways, and x could include indicator variables for school, classroom, teacher, and other groups. Coefficients on indicator variables are sometimes called "fixed effects." It will be useful to partition $x = (\, x_1 \quad x_2 \,)$, with $x\beta = x_1\beta_1 + x_2\beta_2$, and β_1 is $K_1 \times M$, β_2 is $K_2 \times M$. The indicator variables are in x_2, and x_1 includes a constant (and other variables) so that coefficients on indicator variables can be interpreted as deviations from an average coefficient. The relative dimensions could have a large number N of students, a large number of indicator variables making K_2 large with a large number of fixed effects in β_2, with a small number K_1 of variables in x_1 and a small number M of tests. (Could have $M = 1$.)

We may want to impose restrictions on β_2, such as having the same fixed effect in the predictors for multiple test scores. This could take the form

$$\beta_2 = \tau\gamma',$$

where τ is the $K_2 \times 1$ vector of fixed effects and γ is $M \times 1$. More generally, there could be l fixed effects corresponding to each group, with $1 \leq l \leq M$, so that τ is $K_2 \times l$ and γ is $M \times l$. We can impose normalizations, because τ and γ are not separately identified. If $l = M$, then β_2 is unrestricted.

This is related to the model in Chamberlain and Moreira (2009). Their focus is on the estimation of γ, treating τ as a nuisance parameter. Because τ has high dimension, there is an incidental

parameters problem caused by the nonlinear term $\tau\gamma'$. Their solution is to use invariance arguments to construct a statistic whose distribution does not depend upon the incidental parameters. Then a marginal likelihood function can be based on that statistic, with a low dimension parameter space. I shall use their arguments to deal with γ, and my focus is on estimating $\tau\gamma'$.

Simplify notation by temporarily dropping the term $x_1\beta_1$. There is a convenient linear transformation (given x) of $x_2\tau\gamma'$ that I'll call μ. Imposing normalizations, we can express μ as

$$\mu = \omega\lambda',$$

where ω is a $K_2 \times l$ matrix with orthonormal columns: $\omega'\omega = I_l$, and the $M \times l$ matrix λ has an $l \times l$ submatrix that is lower triangular with positive diagonal elements. The positive-definite matrix Σ has the factorization $\Sigma = \sigma\sigma'$, where σ is a lower triangular matrix with positive diagonal elements. We can index the family of distributions for Y conditional on x using the parameter $\theta = (\omega, \lambda, \sigma)$.

Using a quadratic loss function based on forecasting Y, there is a class of invariant estimators for μ. The risk function for an invariant estimator depends on the parameter θ only through $\sigma^{-1}\lambda$. The risk function does not depend on the high dimension parameter ω.

I construct an oracle estimator for μ that is allowed to depend on $\alpha(\theta) = (\lambda, \sigma)$. The oracle is an optimal invariant estimator, and I can compare feasible invariant estimators to it. The oracle is based on the posterior mean for ω using an invariant prior distribution.

I use a random-effects model to motivate a particular (feasible) invariant estimator. The key feature of this random-effects model is that the distribution it implies for ω is the invariant prior distribution. The risk function for the random-effects estimator is evaluated using the parameter space for our original, fixed-effects model. This risk function depends only upon $\sigma^{-1}\lambda$, and can be compared to the oracle risk function. This comparison is particularly simple when $M = 1$; for example, the scalar outcome Y_i could be the score for student i on a single test. Then the risk function depends only on the scalar noncentrality parameter $\delta = \mu'\mu/\sigma^2 = \lambda^2/\sigma^2$. When $M = 1$, this random-effects estimator is in the James and Stein (1961) family of estimators.

My estimator requires a specification for which variables are in x_2. In the random-effects model, this corresponds to assigning a mean of zero to β_2. But my approach corresponds to a fixed-effects model in that it does not require any additional specification for the distribution of β_2 conditional on x.

Section 2 sets up a canonical form for the model, which simplifies the subsequent analysis. Section 3 applies standard invariance arguments, and Section 4 constructs an optimal invariant estimator when $\alpha(\theta) = (\lambda, \sigma)$ is given. Section 5 uses a random-effects model based on the invariant prior to obtain a feasible invariant estimator.

The application in Section 6 draws on work by Chetty, Hendren, Kline, and Saez (2014) and Chetty and Hendren (2015). There is a sample of parents moving across commuting zones with children of age less than 23. There are measures of parent income and child outcome measures such as an indicator for college attendance and the child's income rank at age 26. Based on the age of the child at the move, a variable is constructed that measures the exposure of the child to the new neighborhood. For each origin-destination pair of commuting zones (o, d), a separate least-squares fit provides regression coefficients that are used to form a statistic S_{od}. In comparing two children for whom exposure to d differs by one year, with both children having parents at quantile p of the income distribution, the predicted difference in the outcomes is S_{od}. Each row of my Y matrix corresponds to an (o, d) pair, with N between 3000 and 5000 in the samples that I use. The elements in the row are constructed from the S_{od} statistics. Weights are used, based on the number of families moving from origin o to destination d, or based on an estimated variance for S_{od}.

For example, using college attendance for the child outcome and $p = .25$ for parent income rank, we would have a single column in Y ($M = 1$). Jointly using college attendance and income rank for the child outcomes and $p = .25$ and .75 for the income rank of the parents would give $M = 4$ columns in the Y matrix.

The x matrix is constructed using indicator variables for commuting zones. There is a column for each commuting zone. The (o, d) row of x has $+1$ in the column for d and -1 in the column for

o, and the row is multiplied by the weight for the (o, d) pair. Corresponding to each column of Y, there is a column of the β matrix, with a coefficient for each commuting zone. The columns of x sum to zero, and we can normalize the coefficients in β to sum to zero across commuting zones. So the coefficients can be interpreted as deviations from an average coefficient, and we have $\beta = \beta_2$, $x = x_2$, and $K = K_2$ is greater than 500. I interpret β as providing place effects that summarize differences across commuting zones in intergenerational mobility.

A one-factor model has $\beta = \tau\gamma'$, where τ is a single column with an element for each commuting zone, and the $M = 4$ elements in γ provide (relative) effects on the four variables in Y. Restricting β to have rank $= 2$ could correspond to separate factors for the child's college attendance and income rank outcomes.

2. CANONICAL FORM

The observation is the realized value of an $N \times M$ matrix Y. We shall be conditioning on the value of an $N \times K$ matrix x, which is observed. Partition $x = (\, x_1 \quad x_2\,)$, where x_1 is $N \times K_1$ and x_2 is $N \times K_2$. Our model specifies a conditional distribution for Y given x, as a function of the parameter $(\beta_1, \tau, \gamma, \sigma)$:

$$Y \,|\, x \stackrel{d}{=} x_1\beta_1 + x_2\tau\gamma' + W\sigma',$$

where τ is $K_2 \times l$ with $l \leq M$, γ is $M \times l$, and the $M \times M$ matrix σ is lower-triangular with positive diagonal elements. The components of the $N \times M$ matrix W are, conditional on x, independent and identically distributed $\mathcal{N}(0, 1)$, which we shall denote by

$$\mathcal{L}(W) = \mathcal{N}(0, I_N \otimes I_M).$$

(For a random matrix V, the notation $\mathcal{L}(V) = \mathcal{N}(\mu, \Lambda)$ indicates that the vector formed by joining the rows of V has a multivariate normal distribution with covariance matrix Λ and mean vector formed by joining the rows of μ.) We cannot, without further restrictions, distinguish between (τ, γ) and $(\tau c', \gamma c^{-1})$, where c is any nonsingular $l \times l$ matrix. It will be convenient to restrict the submatrix formed from the first l rows of γ to be lower triangular with positive diagonal elements.

Let $\tilde{x}_2$ denote the residual from the least-squares projection of x_2 on x_1:

$$\tilde{x}_2 = x_2 - x_1 a, \quad x_1' \tilde{x}_2 = 0.$$

Let

$$\pi = \beta_1 + a\tau\gamma'$$

so that

$$x_1\beta_1 + x_2\tau\gamma' = x_1\pi + \tilde{x}_2\tau\gamma'.$$

Assume that $\tilde{x}_2\tau$ has full column rank ($= l$). Note that β_1 unrestricted implies that π is unrestricted.

Let h denote the rank of x_1. The singular value decomposition (SVD) of x_1 gives

$$x_1 = q_1 d_1 s_1',$$

where q_1 is $N \times h$ with $q_1'q_1 = I_h$, d_1 is a $h \times h$ diagonal matrix with positive diagonal elements, and s_1 is $K_1 \times h$ with $s_1's_1 = I_h$. Let r denote the rank of $\tilde{x}_2$ (with $r \geq l$). The SVD of $\tilde{x}_2$ gives

$$\tilde{x}_2 = q_2 d_2 s_2',$$

where q_2 is $N \times r$ with $q_2'q_2 = I_r$, d_2 is a $r \times r$ diagonal matrix with positive diagonal elements, and s_2 is $K_2 \times r$ with $s_2's_2 = I_r$. Note that $x_1'\tilde{x}_2 = 0$ implies that $q_1'q_2 = 0$. The rank of x is $h + r$, which must be $\leq N$. If $h + r < N$, let the columns of q_3 be an orthonormal basis for $\{q_1, q_2\}^\perp$ (the orthogonal complement of the linear space spanned by the columns of q_1 and q_2). Then the $N \times N$ matrix $q = (\, q_1 \quad q_2 \quad q_3\,)$ has orthonormal columns: $q'q = I_N$. If $h + r = N$, then simply set $q = (\, q_1 \quad q_2\,)$, and again q is an $N \times N$ matrix with $q'q = I_N$.

We can use the one-to-one function $q'Y$ as the observation, with

$$\begin{pmatrix} q_1' \\ q_2' \\ q_3' \end{pmatrix} Y \,|\, x \stackrel{d}{=} \begin{pmatrix} d_1 s_1' \\ 0 \\ 0 \end{pmatrix} \pi + \begin{pmatrix} 0 \\ d_2 s_2' \\ 0 \end{pmatrix} \tau\gamma' + \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \sigma'$$

(where we have used $q'W \stackrel{d}{=} W$). Define

$$\tilde{\pi} = d_1 s_1' \pi, \quad \tilde{\tau} = d_2 s_2' \tau.$$

If x_1 has full column rank, then $\pi = s_1 d_1^{-1}\tilde{\pi}$; if $\tilde{x}_2$ has full column rank, then $\tau = s_2 d_2^{-1}\tilde{\tau}$. Let $\mathcal{F}_{l,r}$ denote the set of $r \times l$ matrices whose columns are orthogonal and have unit length:

$$\mathcal{F}_{l,r} = \{a \in \mathcal{R}^{r\times l} : a'a = I_l\}.$$

The matrix $\tilde{\tau}$ has the QR factorization

$$\tilde{\tau} = \omega\rho', \quad \omega \in \mathcal{F}_{l,r}, \quad \rho\rho' = \tilde{\tau}'\tilde{\tau},$$

where ρ is lower triangular with positive diagonal elements. Let $\lambda = \gamma\rho$ and note that the submatrix formed from the first l rows of λ is lower triangular with positive diagonal elements. Then we have

$$\begin{pmatrix} q_1' \\ q_2' \\ q_3' \end{pmatrix} Y \mid x \stackrel{d}{=} \begin{pmatrix} \tilde{\pi} \\ \omega\lambda' \\ 0 \end{pmatrix} + \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \sigma'.$$

To simplify notation, I shall focus on estimating $(\omega, \lambda, \sigma)$ using $Z_1 = q_2'Y$ and $Z_2 = q_3'Y$. So, conditional on x, the canonical form is

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} \omega\lambda' \\ 0 \end{pmatrix} + \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \sigma',$$

where $V_1 \sim \mathcal{N}(0, I_r \otimes I_M)$, $V_2 \sim \mathcal{N}(0, I_{n-r} \otimes I_M)$ (with $n = N - h$), and V_1 and V_2 are independent. (I shall use $q_1'Y$ as the estimate of $\tilde{\pi}$; this implies that the estimate of $x_1\pi$ is the least-squares projection of Y on x_1.)

3. INVARIANCE

We shall be using standard invariance arguments; see, for example, Eaton (1983, 1989). The sample space for the observation $Z' = (\, Z_1' \quad Z_2' \,)$ is

$$\mathcal{Z} = \{z = (z_1, z_2) : z_1 \in \mathcal{R}^{r\times M},\ z_2 \in \mathcal{R}^{(n-r)\times M}\}.$$

Let $\theta = (\omega, \lambda, \sigma)$ denote the parameter. The parameter space is

$$\Theta = \mathcal{F}_{l,r} \times \mathcal{D}_l \times \mathcal{G}_T^+,$$

where $\mathcal{D}_l$ is the set of $M \times l$ matrices with the $l \times l$ submatrix formed from the first l rows a lower triangular matrix with positive diagonal elements, and $\mathcal{G}_T^+$ is the group of $M \times M$ lower-triangular matrices with positive diagonal elements.

Consider the group

$$G = O(r) \times O(n-r) \times {\mathcal{G}_{\mathcal{T}}}^+,$$

where $O(k)$ denotes the group of $k \times k$ orthogonal matrices. Let $g = (g_1, g_2, g_3)$ denote an element of G. Define an action of the group on the sample space:

$$m_1 \colon G \times \mathcal{Z} \to \mathcal{Z}, \quad m_1(g, z) = (g_1 z_1 g_3', g_2 z_2 g_3'),$$

and abbreviate $m_1(g, z) = g \cdot z$. Define an action of the group on the parameter space:

$$m_2 \colon G \times \Theta \to \Theta, \quad m_2(g, \theta) = (g_1 \omega, g_3 \lambda, g_3 \sigma),$$

and abbreviate $m_2(g, \theta) = g \cdot \theta$.

Let P_θ denote the distribution of Z (conditional on x) when the parameter takes on the value θ.

$$\begin{pmatrix} g_1 Z_1 g_3' \\ g_2 Z_2 g_3' \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} g_1 \omega \lambda' g_3' \\ 0 \end{pmatrix} + \begin{pmatrix} g_1 V_1 \\ g_2 V_2 \end{pmatrix} \sigma' g_3' \stackrel{d}{=} \begin{pmatrix} (g_1 \omega)(g_3 \lambda)' \\ 0 \end{pmatrix} + \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} (g_3 \sigma)',$$

so

$$\mathcal{L}(Z) = P_\theta \quad \Rightarrow \quad \mathcal{L}(g \cdot Z) = P_{g \cdot \theta},$$

and the model is invariant under the actions of G on the sample space and the parameter space.

Consider estimation of $\mu = E_\theta(Z_1) = \omega \lambda'$. The loss function is based on the following prediction problem. We observe Y, whose distribution is in our model, and we want to predict the value of an independent draw Y^* from the same distribution (conditional on x): $\mathcal{L}(Y^* \,|\, x) = \mathcal{L}(Y \,|\, x)$. Given estimates $\tilde{\pi}_e$ and μ_e, our prediction for Y^* is

$$\hat{Y}^* = q \begin{pmatrix} \tilde{\pi}_e \\ \mu_e \\ 0 \end{pmatrix}.$$

The forecast loss function is

$$E_{(\beta_1,\tau,\gamma,\sigma)}\left[\text{trace}[\Sigma^{-1}(Y^* - \hat{Y}^*)'(Y^* - \hat{Y}^*)]\right] = \text{trace}[\Sigma^{-1}(\tilde{\pi} - \tilde{\pi}_e)'(\tilde{\pi} - \tilde{\pi}_e)]$$
$$+ \text{trace}[\Sigma^{-1}(\mu - \mu_e)'(\mu - \mu_e)] + NM$$

(with $\Sigma = \sigma\sigma'$). So for estimation of μ, the action space is $\mathcal{A} = \mathcal{R}^{r\times M}$ and the loss function is

$$L\colon \Theta \times \mathcal{A} \to \mathcal{R}, \quad L(\theta, a) = \text{trace}[\Sigma^{-1}(\mu - a)'(\mu - a)].$$

The action of G on the action space is given by $g \cdot a = g_1 a g_3'$. Then $L(g \cdot \theta, g \cdot a) = L(\theta, a)$ for all $g \in G$, $a \in \mathcal{A}$, and $\theta \in \Theta$. So the loss function is invariant.

An estimator (decision rule) $\hat{\mu}\colon \mathcal{Z} \to \mathcal{A}$ maps the sample space into the action space. An estimator $\hat{\mu}$ is invariant if for all $g \in G$, $z \in \mathcal{Z}$ we have

$$\hat{\mu}(g \cdot z) = g \cdot \hat{\mu}(z).$$

The risk function for an estimator $\hat{\mu}$ expresses expected loss as a function of the parameter:

$$R(\theta, \hat{\mu}) = E_\theta[L(\theta, \hat{\mu}(Z))].$$

If $\hat{\mu}$ is an invariant estimator, then the risk function depends on $\theta = (\omega, \lambda, \sigma)$ only through $\sigma^{-1}\lambda$; it does not depend upon ω. A key step in the argument is that for any $\theta = (\omega, \lambda, \sigma) \in \Theta$, we can choose $g \in G$ such that $g \cdot \theta = (e_l, \sigma^{-1}\lambda, I_M)$, where e_l is the matrix formed from the first l columns of I_r. To see this, note that for any $\omega \in \mathcal{F}_{l,r}$, we can choose $g_1 \in O(r)$ so that $g_1\omega = e_l$. Choose $g_3 = \sigma^{-1}$. Then $g \cdot \theta = (e_l, \sigma^{-1}\lambda, I_M)$. Let $\theta_* = (e_l, \sigma^{-1}\lambda, I_M)$ denote this function of θ. Now use the invariance of the loss function, the invariance of the estimator, and the invariance of the model:

$$\begin{aligned} R(\theta, \hat{\mu}) &= E_\theta[L(\theta, \hat{\mu}(Z))] \\ &= E_\theta[L(g \cdot \theta, g \cdot \hat{\mu}(Z))] \\ &= E_\theta[L(\theta_*, \hat{\mu}(g \cdot Z))] \\ &= E_{g\cdot\theta}[L(\theta_*, \hat{\mu}(Z))] \\ &= E_{\theta_*}[L(\theta_*, \hat{\mu}(Z))] \\ &= R(\theta_*, \hat{\mu}). \end{aligned} \tag{1}$$

4. OPTIMALITY

I shall construct an oracle estimator for $\mu = E_\theta(Z_1)$ that is allowed to depend on the following function of θ: $\alpha(\theta) = (\lambda, \sigma)$. The oracle will be an optimal invariant estimator, and we can compare feasible invariant estimators to it.

The oracle is constructed by minimizing posterior loss with respect to a uniform prior distribution on $\mathcal{F}_{l,r}$. The construction of the uniform distribution on $\mathcal{F}_{l,r}$ follows Eaton (1989, Proposition 7.1, p. 100). Start with an $r \times l$ matrix T of independent standard normal random variables: $\mathcal{L}(T) = \mathcal{N}(0, I_r \otimes I_l)$. With probability one the rank of T is l, in which case the polar decomposition of T gives $T = US$, where $U \in \mathcal{F}_{l,r}$ and $S = (T'T)^{1/2}$ is the unique positive-definite matrix satisfying $S^2 = T'T$. The distribution of $U = T(T'T)^{-1/2}$ is uniform in that it has the following invariance property: for any $g \in O(r)$, $\mathcal{L}(gU) = \mathcal{L}(U)$. To see this, note that $gT \stackrel{d}{=} T$, and so

$$gU \stackrel{d}{=} gT(T'T)^{-1/2} \stackrel{d}{=} gT[(gT)'(gT)]^{-1/2} \stackrel{d}{=} T(T'T)^{-1/2} \stackrel{d}{=} U.$$

Let $\mathcal{L}(U) = \eta$ denote the uniform distribution on $\mathcal{F}_{l,r}$. Let $f(z \,|\, \theta)$ denote the density of P_θ:

$$f(z \,|\, (\omega, \lambda, \sigma)) = (2\pi)^{-n/2} |\Sigma|^{-n/2} \exp\Big[-\frac{1}{2} \mathrm{trace}\big[\Sigma^{-1}[(z_1 - \omega\lambda')'(z_1 - \omega\lambda') + z_2'z_2]\big]\Big],$$

with $\Sigma = \sigma\sigma'$. Consider the average risk, averaging over the uniform distribution on $\mathcal{F}_{l,r}$ with $\alpha = (\lambda, \sigma)$ given:

$$\begin{aligned} A(\hat{\mu}; \alpha) &= \int_{\mathcal{F}_{l,r}} R((\omega, \alpha), \hat{\mu})\, \eta(d\omega) \\ &= \int_{\mathcal{F}_{l,r}} \int_{\mathcal{Z}} L((\omega, \alpha), \hat{\mu}(z)) f(z \,|\, (\omega, \alpha))\, dz\, \eta(d\omega). \end{aligned}$$

We can minimize the average risk, without constraining $\hat{\mu}$ to be invariant, by reversing the order of the double integral and minimizing the inner integral separately at each value for z:

$$A(\hat{\mu}; \alpha) \geq \int_{\mathcal{Z}} \left[\min_{t \in \mathcal{A}} \int_{\mathcal{F}_{l,r}} L((\omega, \alpha), t) f(z \,|\, (\omega, \alpha))\, \eta(d\omega) \right] dz.$$

So

$$\hat{\mu}^*(\,\cdot\,;\alpha) = \arg\min_{\hat{\mu}} A(\hat{\mu};\alpha)$$

is obtained by setting $\hat{\mu}^*(z;\alpha)$ equal to the minimizing value for t in the inner integral, which is equivalent to minimizing posterior expected loss:

$$\hat{\mu}(z;\alpha) = \arg\min_{t\in\mathcal{A}} \int_{\mathcal{F}_{l,r}} L((\omega,\alpha),t) f(z\,|\,(\omega,\alpha))\,\eta(d\omega) \Big/ \int_{\mathcal{F}_{l,r}} f(z\,|\,(\omega,\alpha))\,\eta(d\omega).$$

With our quadratic loss function, posterior expected loss is minimized by the posterior mean:

$$\hat{\mu}^*(z;\alpha) = \int_{\mathcal{F}_{l,r}} \omega\lambda' f(z\,|\,(\omega,\alpha))\,\eta(d\omega) \Big/ \int_{\mathcal{F}_{l,r}} f(z\,|\,(\omega,\alpha))\,\eta(d\omega).$$

Because

$$f(z|(\omega,\lambda,\sigma)) = f_1(z_1\,|\,(\omega,\lambda,\sigma)) \cdot f_2(z_2\,|\,\sigma),$$

we can replace $f(z\,|\,(\omega,\alpha))$ by $f_1(z_1\,|\,(\omega,\alpha))$ in the formula for the posterior mean. For calculation, take independent draws $U^{(j)}$ from the invariant distribution η, and approximate $\hat{\mu}^*(z;\alpha)$ by

$$\frac{1}{J}\sum_{j=1}^{J} U^{(j)}\lambda' f(z\,|\,(U^{(j)},\alpha)) \Big/ \frac{1}{J}\sum_{j=1}^{J} f(z\,|\,(U^{(j)},\alpha)).$$

The invariance condition for the oracle is

$$\hat{\mu}^*(g\cdot z;\alpha(g\cdot\theta)) = g\cdot\hat{\mu}^*(z;\alpha(\theta)).$$

This condition implies that for all $\theta\in\Theta$,

$$R(\theta,\hat{\mu}^*(\,\cdot\,;\alpha(\theta))) = R(\theta_*,\hat{\mu}^*(\,\cdot\,;\alpha(\theta_*))),$$

where, as above, $\theta_* = (e_l,\sigma^{-1}\lambda,I_M)$. The argument is similar to that used in equation (1). To see that $\hat{\mu}^*(\,\cdot\,;\alpha)$ satisfies the invariance condition, note that for any $g=(g_1,g_2,g_3)\in G$:

$$f(g\cdot z\,|\,\theta) = |g_3|^{-n} f(z\,|\,g^{-1}\cdot\theta),$$

and so the invariance of η implies that

$$\begin{aligned}
\hat{\mu}^*(g\cdot z;\alpha(g\cdot\theta)) &= \int_{\mathcal{F}_{l,r}} \omega(g_3\lambda)' f(g\cdot z\,|\,(\omega,g_3\lambda,g_3\sigma)\,\eta(d\omega) \Big/ \int_{\mathcal{F}_{l,r}} f(g\cdot z\,|\,(\omega,g_3\lambda,g_3\sigma))\,\eta(d\omega) \\
&= g_1 \int_{\mathcal{F}_{l,r}} g_1^{-1}\omega\lambda' g_3' f(z\,|\,(g_1^{-1}\omega,\lambda,\sigma))\,\eta(d\omega) \Big/ \int_{\mathcal{F}_{l,r}} f(z\,|\,(g_1^{-1}\omega,\lambda,\sigma))\,\eta(d\omega) \\
&= g_1 \int_{\mathcal{F}_{l,r}} \omega\lambda' f(z\,|\,(\omega,\lambda,\sigma))\,\eta(d\omega) g_3' \Big/ \int_{\mathcal{F}_{l,r}} f(z\,|\,(\omega,\lambda,\sigma))\,\eta(d\omega) \\
&= g\cdot\hat{\mu}(z;\alpha(\theta)).
\end{aligned}$$

For an invariant estimator $\hat{\mu}$, the risk $R((\omega,\alpha),\hat{\mu})$ at any $\omega\in\mathcal{F}_{l,r}$ equals the average risk $A(\hat{\mu};\alpha)$, and so the oracle provides a lower bound on risk for invariant estimators:

$$R((\omega,\alpha),\hat{\mu}) = A(\hat{\mu};\alpha) \geq A(\hat{\mu}^*(\,\cdot\,;\alpha);\alpha) = R((\omega,\alpha),\hat{\mu}^*(\,\cdot\,;\alpha)).$$

Furthermore, $\hat{\mu}^*(\,\cdot\,;\alpha)$ is minimax when α is given: for any estimator $\hat{\mu}$ (which need not be invariant),

$$\sup_{\omega\in\mathcal{F}_{l,r}} R((\omega,\alpha),\hat{\mu}^*(\,\cdot\,;\alpha)) = A(\hat{\mu}^*(\,\cdot\,;\alpha);\alpha) \leq A(\hat{\mu};\alpha) \leq \sup_{\omega\in\mathcal{F}_{l,r}} R((\omega,\alpha),\hat{\mu}).$$

5. RANDOM EFFECTS MODEL

I shall use a random-effects model to motivate a particular (feasible) invariant estimator. It is important to stress that the risk function for this estimator will be evaluated using the parameter space for the original, fixed-effects model. In particular, the risk will be compared to the lower bound provided by the oracle.

The random-effects model specifies a distribution for $\tilde{\tau}$:

$$\tilde{\tau} \sim \mathcal{N}(0, I_r \otimes I_l).$$

The key feature of this distribution is that $\tilde{\tau}(\tilde{\tau}'\tilde{\tau})^{-1/2}$ has the uniform distribution η on $\mathcal{F}_{l,r}$. There is a family of distributions $\{Q_\xi : \xi\in\Theta_{\mathrm{RE}}\}$ for Z, indexed by the parameter $\xi=(\gamma,\sigma)$ with parameter space

$$\Theta_{re} = \mathcal{D}_l \times \mathcal{G}_T^+.$$

Q_ξ specifies that Z_1 and Z_2 are independent with

$$Z_1 \sim \mathcal{N}(0, I_r \otimes (\gamma\gamma' + \Sigma)), \quad Z_2 \sim \mathcal{N}(0, I_{n-r} \otimes \Sigma)$$

(with $\Sigma = \sigma\sigma'$). The covariance matrix I_l for a row of $\tilde{\tau}$ is a normalization. If the covariance matrix were $\Phi = \phi\phi'$, then the covariance matrix for a row of Z_1 would be $\gamma\Phi\gamma' + \Sigma$, and we could replace $\tilde{\tau}$ by $\tilde{\tau}\phi'^{-1}$ and replace γ by $\gamma\phi$.

Conditional on $(\tilde{\tau}, \gamma, \sigma)$, the distribution of Z_1 is $\mathcal{N}(\tilde{\tau}\gamma', I_r \otimes \Sigma)$. So given $\xi = (\gamma, \sigma)$, the posterior mean of $\tilde{\tau}$ is

$$E_\xi(\tilde{\tau} \mid Z_1 = z_1) = z_1\Sigma^{-1}\gamma(\gamma'\Sigma^{-1}\gamma + I_l)^{-1}.$$

For an invariant estimator of $\mu = \tilde{\tau}\gamma'$, I shall use

$$\hat{\mu}_{re}(z) = z_1\hat{\Sigma}^{-1}\hat{\gamma}(\hat{\gamma}'\hat{\Sigma}^{-1}\hat{\gamma} + I_l)^{-1}\hat{\gamma}',$$

with $\hat{\Sigma} = \hat{\sigma}\hat{\sigma}'$. The invariance condition that $\hat{\mu}_{re}(g \cdot z) = g_1\hat{\mu}_{re}(z)g_3'$ is satisfied provided that

$$\hat{\gamma}(g \cdot z) = g_3\hat{\gamma}(z), \quad \hat{\sigma}(g \cdot z) = g_3\hat{\sigma}(z).$$

To motivate a particular choice for $\hat{\sigma}$, note that $E_\xi(Z_2'Z_2) = (n-r)\Sigma$, which suggests using $\hat{\Sigma}(z) = z_2'z_2/(n-r)$. Provided that $\hat{\Sigma}(z)$ is positive definite, there will be a unique $\hat{\sigma}(z) \in \mathcal{G}_T^+$ with $\hat{\sigma}(z)\hat{\sigma}(z)' = \hat{\Sigma}(z)$, and this $\hat{\sigma}$ satisfies the invariance condition $\hat{\sigma}(g \cdot z) = g_3\hat{\sigma}(z)$.

To motivate a particular choice for $\hat{\gamma}$, note that

$$E_\xi(\sigma^{-1}Z_1'Z_1\sigma'^{-1}) = r(\sigma^{-1}\gamma\gamma'\sigma'^{-1} + I_M).$$

A spectral decomposition gives

$$\hat{\sigma}^{-1}(z)z_1'z_1\hat{\sigma}(z)'^{-1}/r = \sum_{j=1}^{M} \kappa_j\nu_j\nu_j' \quad \text{with} \quad \kappa_1 \geq \kappa_2 \cdots \geq \kappa_M \geq 0$$

and orthonormal eigenvectors $\nu_j'\nu_j = 1$, $\nu_j'\nu_k = 0$ $(j \neq k)$. Let

$$A(z) = \hat{\sigma}(z)\left(\sum_{j=1}^{l}(\kappa_j - 1)^+\nu_j\nu_j'\right)\hat{\sigma}(z)'$$

(with $t^+ = \max\{t, 0\}$ for $t \in \mathcal{R}$). Let $\bar{l} \leq l$ denote the rank of $A(z)$. If the submatrix formed from the first $\bar{l}$ rows and columns of $A(z)$ has rank $\bar{l}$, then there is a unique $M \times l$ matrix $\hat{\gamma}(z)$ such that the submatrix formed from the first $\bar{l}$ columns is in $\mathcal{D}_{\bar{l}}$, the remaining columns (if any) are zero, and

$$\hat{\gamma}(z)\hat{\gamma}(z)' = A(z).$$

This $\hat{\gamma}$ satisfies the invariance condition $\hat{\gamma}(g \cdot z) = g_3'\hat{\gamma}(z)$.

If $M = 1$, these choices for $\hat{\sigma}$ and $\hat{\gamma}$ give

$$\hat{\mu}_{re}(z) = (1 - 1/F_{stat})^+ \hat{\mu}_{ls}(z),$$

where the least-squares estimate of μ is $\hat{\mu}_{ls}(z) = z_1$, and the F-statistic for testing $\mu = 0$ is

$$F_{stat} = \frac{z_1'z_1/r}{z_2'z_2/(n-r)}.$$

This estimator is in the James and Stein (1961) family of (positive-part) estimators:

$$\hat{\mu}_{JS+}(z) = (1 - c\frac{z_2'z_2}{z_1'z_1})^+ z_1,$$

which dominate $\hat{\mu}_{ls}$ if $r \geq 3$ and c is any number in the interval

$$0 < c < \frac{2(r-2)}{n-r+2}.$$

See Sclove (1968) for a discussion of this result. Our estimator $\hat{\mu}_{re}$ has $c = r/(n-r)$ and satisfies the dominance condition if $r \geq 5$ and $n - r > 10$.

If $\tilde{x}_2$ has full column rank, then $\beta_2 = s_2 d_2^{-1} \mu$ and, if $M = 1$,

$$\hat{\beta}_{2,re}(z) = (1 - 1/F_{stat})^+ \hat{\beta}_{2,ls}(z).$$

Each element of the least-squares estimate of β_2 is multiplied by the same factor to obtain the random-effects estimate. This is a consequence of using the invariant prior and would not generally

hold in other random-effects models, which would imply a different specification for the distribution of β_2 conditional on x. The invariant prior distribution for $\tilde{\tau}$ is not meant to be a subjective choice, motivated, for example, by exchangeability. The invariant prior implies that the covariance matrix for β_2 is proportional to $s_2 d_2^{-2} s_2' = (\tilde{x}_2' \tilde{x}_2)^{-1}$. If one wanted to model the covariance matrix of β_2 conditional on x, and, perhaps motivated by exchangeability, chose a covariance matrix proportional to I_{K_2}, then the implied covariance matrix for $\tilde{\tau}$ would be proportional to d_2^2. But I am only using a random-effects model to motivate a feasible approximation to the optimal invariant estimator in the fixed-effects model. That calls for an invariant prior for $\tilde{\tau}$.

If $M = 1$, the risk function $R(\theta, \hat{\mu})$ for an invariant estimator depends upon $\theta = (\omega, \lambda, \sigma)$ only through the scalar noncentrality parameter $\delta = \mu'\mu/\sigma^2 = \lambda^2/\sigma^2$. I use simulation to evaluate the risk functions for the oracle $\hat{\mu}^*(\,\cdot\,; \alpha(\theta))$ and for $\hat{\mu}_{re}$ at $\theta = \theta_* = (e_1, \sqrt{\delta}, 1)$. In order to have a relevant range of values for δ in the empirical application, I shall use a .95 confidence interval. It is based on the distribution of F_{stat}, which is noncentral F with noncentrality parameter δ and degrees of freedom r and $(n - r)$. I take the observed value of F_{stat} and find the value for δ such that F_{stat} is at the .025 quantile of this noncentral F distribution. This gives the upper bound for the confidence interval. The value for δ that puts the observed F_{stat} at the .975 quantile gives the lower bound for the confidence interval.

If $M = 1$, the integrals in the formula for the oracle estimator can be evaluated to give

$$\hat{\mu}^*(z; \alpha) = H(\frac{r}{2} - 1, \frac{\sqrt{\delta}}{\sigma}||z_1||)\sqrt{\delta}\sigma z_1/||z_1||,$$

where $||z_1|| = (z_1' z_1)^{1/2}$, $H(v, u) = I_{v+1}(u)/I_v(u)$, and $I_v(u)$ is a modified Bessel function (a special function for which there are standard computational algorithms). See the Appendix.

6. APPLICATION

I shall draw on work by Chetty, Hendren, Kline, and Saez (2014) and Chetty and Hendren (2015). Chetty and Hendren use data constructed in Chetty et al. (2014) to form a sample of parents moving from commuting zone o to commuting zone d with children of age less than 23.

As in the earlier paper, there is a measure of parent income rank (p) and there are child outcome measures (c) such as an indicator for college attendance and the child's income rank at age 26. A variable is constructed that measures the exposure (ex) of the child to the new neighborhood. The number of families moving from origin o to destination d is n_{od}, and I shall use the (o, d) pairs with $n_{od} \geq 100$. For each of these (o, d) pairs, a least-squares projection of c on a constant, ex, $ex \cdot p$, and additional variables in m gives

$$\hat{c}_i = b_{1,od} \cdot ex_i + b_{2,od} \cdot ex_i \cdot p_i + b'_{3,od} \cdot m_i \qquad (i = 1, \ldots, n_{od}).$$

Here ex_i is the amount of time that child i spent growing up in the destination neighborhood: ex_i = (23 − child$_i$'s age at move), and p_i is the parent income rank in the national distribution. The additional variables in the vector m_i are a constant, p_i, s_i, s_i^2, $s_i \cdot p_i$, $s_i^2 \cdot p_i$, where s_i is the child's cohort. Let S_{od} denote the statistic $b_{1,od} + .25 \cdot b_{2,od}$. In comparing two children (from the same cohort) for whom ex differs by one year, with both children having parents at the .25 quantile of the income distribution ($p = .25$), the predicted difference in the outcomes is S_{od}. The vector Y and the matrix x are formed using the weights $w_{od} = \sqrt{n}_{od}$. Each element of Y corresponds to an (o, d) pair, and the (o, d) element of Y is $w_{od}S_{od}$. The x matrix has a column for each commuting zone. Row (o, d) of x has w_{od} in the column for commuting zone d, with $-w_{od}$ in the column for commuting zone o, and zeros in the other columns. In the notation of our general model, we have $M = 1$, $x = x_2 = \tilde{x}_2$, x_1 is null, $K = K_2$, and $\beta = \beta_2$, which is unrestricted. The regression function is

$$E(Y_{od} \,|\, x) = w_{od}(\beta_d - \beta_o).$$

I interpret β as providing place effects that summarize differences across commuting zones in intergenerational mobility.

To go from our estimates of μ to estimates of β, we need a normalization, because the columns of x sum to zero and so $r = \text{rank}(x) \leq K - 1$. Let $\mathbf{1}$ denote a $K \times 1$ vector of ones. Then $0 = q'_2 x\mathbf{1} = d_2 s'_2 \mathbf{1}$ implies that $s'_2\mathbf{1} = 0$, so that $\mathbf{1}$ is orthogonal to the r columns of s_2. We can normalize $\mathbf{1}'\beta = 0$, so that the place effects sum to 0. If $r = K - 1$, then $\mu = d_2 s'_2 \beta$ together with

the normalization $0 = \mathbf{1}'\beta$ gives the unique solution $\beta = s_2 d_2^{-1}\mu$. The least-squares estimate of μ is $\hat{\mu}_{ls} = Z_1$. Let $\hat{\beta}_{ls} = s_2 d_2^{-1}\hat{\mu}_{ls}$ and $\hat{\beta}_{re} = s_2 d_2^{-1}\hat{\mu}_{re}$.

For a simple summary measure, I shall use the standard deviation (SD) of the estimated place effects, weighting by the population in the 2000 Census. The data are from the Chetty and Hendren (2015) Online Data Tables 3 and 5 (http://www.equality-of-opportunity.org). With college attendance (*col*) as the child outcome c, there are $N = 4931$ commuting zone (o, d) pairs that satisfy the $n_{od} \geq 100$ requirement. The rank of x is $r = K - 1 = 586$. Multiplying by 100 to convert the probability of college attendance to percentage points, we have

$$\mathrm{SD}(\hat{\beta}_{ls}^{col,.25}) = .48, \quad \mathrm{SD}(\hat{\beta}_{re}^{col,.25}) = .24.$$

With the least-squares estimate, a one standard deviation increase in a place effect corresponds to a predicted increase of .48 percentage points in the probability of college attendance (per year of exposure). With the random effects estimate, the predicted increase is .24 percentage points. With the latter estimate, 20 years of exposure imply a predicted increase of 4.7 percentage points.

The value for the F statistic is $F_{stat} = 1.96$ with $r = 586$ and $n - r = 4345$. The .95 interval for the noncentrality parameter δ is $[.76 \cdot r, 1.18 \cdot r]$. Over this interval, the ratio of $\hat{\mu}_{re}$ risk to oracle risk varies from 1.006 to 1.004. So the feasible invariant estimator is almost achieving the oracle bound on risk. The ratio of least-squares risk to oracle risk varies from 2.32 to 1.84. So the oracle and its feasible approximation provide substantial risk improvements over the least-squares estimator.

The unweighted standard deviations of the estimated place effects are $\mathrm{SD}(\hat{\beta}_{ls}^{col,.25}) = .95$, $\mathrm{SD}(\hat{\beta}_{re}^{col,.25}) = .47$. The population weights matter because the ratio of largest to smallest is over two thousand. The least-squares estimates of the individual place effects are all reduced by the same factor: $1 - 1/F_{stat} = .49$, and so the ratio of the standard deviations is the same as before: $\mathrm{SD}(\hat{\beta}_{re}^{col,.25})/\mathrm{SD}(\hat{\beta}_{ls}^{col,.25}) = .49$. This is a consequence of using an invariant prior in the random-effects model, in order to match the optimal invariant estimator in the fixed-effects model. An alternative would be to require a separate specification for the distribution of β conditional on

x, so we would no longer be using a fixed-effects approach. With survey sampling of individuals, random sampling can motivate an independent and identically distributed (i.i.d.) specification for individual effects. The random sampling motivation is less obvious here, because the commuting zones form a partition of the country. Even if an i.i.d. specification is adopted for the place effects, this need not hold conditional on x. If we did assume that the covariance matrix of β conditional on x is proportional to an identity matrix, then the implied covariance matrix for μ would be proportional to d_2^2 instead of the invariant prior specification of I_r. This can make a difference, because the ratio of the largest to smallest diagonal elements of d_2^2 is over ten thousand. (The diagonal matrix d_2^2 contains the nonzero eigenvalues of $x'x$.)

Now use the .75 quantile of the income distribution for parents and set $S_{od} = b_{1,od} + .75 \cdot b_{2,od}$. In comparing two children from the same cohort whose families move from o to d, with parents at the .75 quantile of the income distribution, the predictive effect of an additional year of exposure to d is S_{od}. Using S_{od} and n_{od} to construct Y and x as before, and using population-weighted standard deviations, gives

$$\mathrm{SD}(\hat{\beta}_{ls}^{col,.75}) = .40, \quad \mathrm{SD}(\hat{\beta}_{re}^{col,.75}) = .19.$$

With the least-squares estimate, a one standard deviation increase in a place effect corresponds to a predicted increase of .40 percentage points in the probability of college attendance (per year of exposure), for a child with parents at the .75 quantile of the income distribution. With the random effects estimate, the predicted increase is .19 percentage points, and 20 years of exposure gives a predicted increase of 3.8 percentage points. The value for the F statistic is $F_{stat} = 1.90$ with $r = 586$ and $n - r = 4345$. The .95 interval for δ is $[.70 \cdot r, 1.12 \cdot r]$. Over this interval, the ratio of $\hat{\mu}_{re}$ risk to oracle risk varies from 1.006 to 1.005. The ratio of least-squares risk to oracle risk varies from 2.42 to 1.90.

The above results use the weights $w_{od} = \sqrt{n}_{od}$ in constructing Y and x. An alternative is to use the estimated covariance matrix of $(b_{1,od}, b_{2,od})$ to provide an estimated variance for $S_{od} = b_{1,od} + p \cdot b_{2,od}$ (with $p = .25$ or .75). Let var_{od} denote this estimated variance and use $w_{od} =$

$var_{od}^{-1/2}$ for the weights. This gives similar results:

$$\text{SD}(\hat{\beta}_{ls}^{col,.25}) = .48, \quad \text{SD}(\hat{\beta}_{re}^{col,.25}) = .24, \quad \text{SD}(\hat{\beta}_{ls}^{col,.75}) = .38, \quad \text{SD}(\hat{\beta}_{re}^{col,.75}) = .16.$$

The F statistics are 2.01 for $p = .25$ and 1.71 for $p = .75$.

With income rank at age 26 (kr) as the child outcome c, there are $N = 3094$ commuting zone (o, d) pairs that satisfy the $n_{od} \geq 100$ requirement. The rank of x is $r = K - 1 = 508$. Multiply by 100 to convert the income rank from quantiles to percentiles. Using the weights $w_{od} = \sqrt{n}_{od}$ gives

$$\text{SD}(\hat{\beta}_{ls}^{kr,.25}) = .33, \quad \text{SD}(\hat{\beta}_{re}^{kr,.25}) = .038, \quad \text{SD}(\hat{\beta}_{ls}^{kr,.75}) = .40, \quad \text{SD}(\hat{\beta}_{re}^{kr,.75}) = .105.$$

Using the weights $w_{od} = var_{od}^{-1/2}$ gives

$$\text{SD}(\hat{\beta}_{ls}^{kr,.25}) = .33, \quad \text{SD}(\hat{\beta}_{re}^{kr,.25}) = .078, \quad \text{SD}(\hat{\beta}_{ls}^{kr,.75}) = .39, \quad \text{SD}(\hat{\beta}_{re}^{kr,.75}) = .052.$$

Using the variance weights makes more of a difference here than it did with the college outcome. With the variance weights, the F statistics are 1.31 with $p = .25$ and 1.15 with $p = .75$. The random effects estimate with $p = .25$ implies that a one standard deviation increase in a place effect corresponds to a predicted increase of .078 percentiles of income rank per year of exposure. At the .75 quantile of the parent income distribution, the predicted increase is .052 percentiles. With 20 years of exposure, the predicted increases in income rank are 1.6 and 1.0 percentiles. With $p = .25$, the .95 interval for the noncentrality parameter δ is $[.15 \cdot r, .50 \cdot r]$. Over this interval, the ratio of $\hat{\mu}_{re}$ risk to oracle risk varies from 1.03 to 1.01. The ratio of least-squares risk to oracle risk varies from 7.60 to 3.02. So the oracle and its feasible approximation provide substantial risk improvements over the least-squares estimator.

In the multivariate model with $M = 4$, let

$$Y = \left(Y^{col,.25} \quad Y^{col,.75} \quad Y^{kr,.25} \quad Y^{kr,.75} \right),$$

where $Y^{col,p}$ is the vector constructed above using the college outcome and with parents at quantile p of the income distribution; $Y^{kr,p}$ is constructed in the same way, using the child's income rank at

age 26. There are $N = 3094$ commuting zone (o, d) pairs that satisfy the $n_{od} \geq 100$ requirement for both the college and income outcomes. I shall use $w = \sqrt{n}_{od}$ (with n_{od} from the college outcome). With β unrestricted, the standard deviations for the four columns of $\hat{\beta}_{re}$ are

$$\text{SD}(\hat{\beta}_{re}^{col,.25}) = .27, \quad \text{SD}(\hat{\beta}_{re}^{col,.75}) = .19, \quad \text{SD}(\hat{\beta}_{re}^{kr,.25}) = .052, \quad \text{SD}(\hat{\beta}_{re}^{kr,.75}) = .101.$$

Imposing the rank 2 restriction that $\beta = \tau\gamma'$, where τ is $K \times 2$ and γ' is 2×4, gives

$$\text{SD}(\hat{\beta}_{re}^{col,.25}) = .26, \quad \text{SD}(\hat{\beta}_{re}^{col,.75}) = .18, \quad \text{SD}(\hat{\beta}_{re}^{kr,.25}) = .050, \quad \text{SD}(\hat{\beta}_{re}^{kr,.75}) = .101.$$

This rank restriction corresponds to separate factors for the college and income rank outcomes. The two-factor model provides a good summary of the unrestricted estimates. Restricting to a single factor so that $\beta = \tau\gamma'$,where τ is $K \times 1$ and γ' is 1×4, gives

$$\text{SD}(\hat{\beta}_{re}^{col,.25}) = .26, \quad \text{SD}(\hat{\beta}_{re}^{col,.75}) = .18, \quad \text{SD}(\hat{\beta}_{re}^{kr,.25}) = .020, \quad \text{SD}(\hat{\beta}_{re}^{kr,.75}) = .009.$$

The standard deviations of the place effects for the college outcomes are not affected, but there is a sharp drop for the income rank outcomes. The one-factor model does not provide a good summary of the unrestricted estimates.

7. CONCLUSION

I have developed a fixed-effects model along with an oracle bound on the risk of invariant estimators. The oracle estimator uses an invariant prior, which I have incorporated into a random-effects model to obtain a feasible estimator. This estimator almost achieves the oracle bound over the relevant part of the (fixed-effects) parameter space in the empirical application. There is a substantial reduction in risk compared with the least-squares estimator. The random-effects estimator requires a specification for which variables are in x_2 (with $x\beta = x_1\beta_1 + x_2\beta_2$). This corresponds to assigning a mean of zero to β_2. The estimator does not require a separate specification for the covariance matrix of β_2 conditional on x, because this is chosen to mimic the oracle in the fixed-effects model. An alternative is a random-effects approach that separately develops a model for

the distribution of β_2 conditional on x, perhaps using exchangeability arguments. The form of my estimator suggests entertaining specifications in which the covariance matrix of β_2 depends upon x. This may matter in the application, because there are large differences across the commuting zones in their scale, which is reflected in substantial differences between population-weighted and unweighted standard deviations.

APPENDIX

Formula for the Oracle

With $M = l = 1$ and (λ, σ) given, the distribution of Z_1 is $P_{1,\omega} = \mathcal{N}(\lambda\omega, \sigma^2 I_r)$. The risk function is

$$\tilde{R}(\omega, \hat{\mu}) = \sigma^{-2} \int_{\mathcal{R}^r} ||\hat{\mu}(z_1) - \lambda\omega||^2 P_{1,\omega}(dz_1).$$

Let ζ denote surface measure on the unit sphere $\mathcal{S}^{r-1}$ of dimension $r-1$ in $\mathcal{R}^r$, and consider the average risk, averaging over the uniform distribution on $\mathcal{S}^{r-1}$:

$$A(\hat{\mu}) = \sigma^{-2} \int_{\mathcal{S}^{r-1}} \int_{\mathcal{R}^r} ||\hat{\mu}(z_1) - \lambda\omega||^2 P_{1,\omega}(dz_1)\zeta(d\omega)/\zeta(\mathcal{S}^{r-1}).$$

We can minimize the average risk, without constraining $\hat{\mu}$ to be invariant, by reversing the order of the double integral and minimizing the inner integral separately at each value for z_1:

$$A(\hat{\mu}) \geq c\sigma^{-r-2} \int_{\mathcal{R}^r} \left[\min_{t \in \mathcal{R}^r} \int_{\mathcal{S}^{r-1}} ||t - \lambda\omega||^2 \exp[-\frac{1}{2\sigma^2}||z_1 - \lambda\omega||^2]\zeta(d\omega) \right] dz_1,$$

with the constant $c = (2\pi)^{-r/2}/\zeta(\mathcal{S}^{r-1})$. The minimizing value for t in the inner integral gives the minimizing value for $\hat{\mu}(z_1)$. So

$$\hat{\mu}^* = \arg\min_{\hat{\mu}} A(\hat{\mu})$$

is given by

$$\hat{\mu}^*(z_1) = \arg\min_{t \in \mathcal{R}^r} \int_{\mathcal{S}^{r-1}} [||t||^2 - 2\lambda t'\omega] \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\zeta(d\omega). \tag{2}$$

($\hat{\mu}^*(z_1)$ corresponds to $\hat{\mu}^*(z; \alpha)$ in the text.)

Let $e_1 \in S^{r-1}$ denote the first column of the identity matrix I_r. To find the solution in (2), we can let $b = t/||t||$ (with the arbitrary value e_1 if $t = 0$) and consider

$$\arg \min_{||t|| \geq 0, b \in \mathcal{S}^{r-1}} [||t||^2 - 2\lambda||t|| \frac{1}{D} \int_{\mathcal{S}^{r-1}} b'\omega \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\zeta(d\omega)] \tag{3}$$

with

$$D = \int_{\mathcal{S}^{r-1}} \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\zeta(d\omega).$$

Choose $h \in O(r)$ such that $hz_1 = ||z_1||e_1$. Note that

$$\begin{aligned}\int_{\mathcal{S}^{r-1}} b'\omega \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\zeta(d\omega) &= \int_{\mathcal{S}^{r-1}} (hb)'h\omega \exp(\frac{\lambda}{\sigma^2}(hz_1)'h\omega)\zeta(d\omega) \\ &= b'h' \int_{\mathcal{S}^{r-1}} \omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\zeta(d\omega)\end{aligned}$$

because the measure ζ is invariant under the action of $O(r)$ on $\mathcal{S}^{r-1}$: $\zeta(h^{-1}B) = \zeta(B)$ for any measurable subset B of $\mathcal{S}^{r-1}$.

Define

$$q = \int_{\mathcal{S}^{r-1}} \omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\zeta(d\omega).$$

Let $c \in O(r)$ have the form

$$c = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{c} \end{pmatrix}, \quad \tilde{c} \in O(r-1),$$

so that $ce_1 = e_1$. Note that

$$\begin{aligned} cq &= \int_{\mathcal{S}^{r-1}} c\omega \exp(\frac{\lambda}{\sigma^2}||z_1||(ce_1)'c\omega)\zeta(d\omega) \\ &= \int_{\mathcal{S}^{r-1}} \omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\zeta(d\omega) \\ &= q. \end{aligned}$$

Partition q as

$$q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix},$$

where q_1 is scalar and q_2 is $(r-1) \times 1$. Then $cq = q$ implies that

$$cq = \begin{pmatrix} q_1 \\ \tilde{c}q_2 \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix},$$

and so $\tilde{c}q_2 = q_2$ for all $\tilde{c} \in O(r-1)$, which implies that $q_2 = 0$.

In addition, we can show that $q_1 > 0$. Define

$$A = \{\omega \in \mathcal{S}^{r-1} : e_1'\omega > 0\}, \quad A^c = \mathcal{S}^{r-1} - A.$$

Note that

$$\int_A e_1'\omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\,\zeta(d\omega) > \int_A e_1'\omega\,\zeta(d\omega),$$
$$\int_{A^c} e_1'\omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\,\zeta(d\omega) > \int_{A^c} e_1'\omega\,\zeta(d\omega),$$

and so

$$q_1 = \int_{\mathcal{S}^{r-1}} e_1'\omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\,\zeta(d\omega) > \int_{\mathcal{S}^{r-1}} e_1'\omega\,\zeta(d\omega) = 0.$$

Because $q_1 > 0$ and $q_2 = 0$, we have $q/||q|| = e_1$, and

$$\arg\max_{b\in\mathcal{S}^{r-1}} b'h'q = h'q/||q|| = h'e_1 = z_1/||z_1||.$$

So

$$\arg\max_{b\in\mathcal{S}^{r-1}} \int_{\mathcal{S}^{r-1}} b'\omega \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\,\zeta(d\omega) = z_1/||z_1||,$$

and

$$\max_{b\in\mathcal{S}^{r-1}} \int_{\mathcal{S}^{r-1}} b'\omega \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\,\zeta(d\omega) = q_1.$$

We can complete the solution to (3) with

$$\arg\min_{||t||\geq 0} [||t||^2 - 2\lambda||t||q_1/D] = \lambda q_1/D.$$

So the solution in (2) is

$$\hat{\mu}^*(z_1) = \lambda(q_1/D)z_1/||z_1||. \tag{4}$$

With $h \in O(r)$ such that $hz_1 = ||z_1||e_1$, we have

$$D = \int_{\mathcal{S}^{r-1}} \exp(\frac{\lambda}{\sigma^2} z_1'\omega)\,\zeta(d\omega) = \int_{\mathcal{S}^{r-1}} \exp(\frac{\lambda}{\sigma^2}(hz_1)'h\omega)\,\zeta(d\omega)$$
$$= \int_{\mathcal{S}^{r-1}} \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\,\zeta(d\omega),$$

and

$$q_1/D = e_1'q/D = \int_{\mathcal{S}^{r-1}} e_1'\omega \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\zeta(d\omega) \Big/ \int_{\mathcal{S}^{r-1}} \exp(\frac{\lambda}{\sigma^2}||z_1||e_1'\omega)\zeta(d\omega).$$

The integrals can be simplified using the following result:

$$\int_{\mathcal{S}^N} f(\alpha'\omega)\,\zeta_N(d\omega) = \zeta_{N-1}(\mathcal{S}^{N-1}) \int_{[-1,1]} f(s)(1-s^2)^{N/2-1}\,ds$$

for all $\alpha \in \mathcal{S}^N$, $N \geq 1$, and all measurable f on $([-1,1], \mathcal{B}_{[-1,1]})$ that are either bounded or nonnegative. (The term $\mathcal{B}_E$ denotes the Borel σ-algebra over the topological space E; ζ_N denotes surface measure on $\mathcal{S}^N$.) See Stroock (1999, pp. 88, 89, 213–215). Applying this result (with $r \geq 2$) gives

$$q_1/D = \int_{[-1,1]} s\exp(\frac{\lambda}{\sigma^2}||z_1||s)(1-s^2)^{(r-3)/2}\,ds \Big/ \int_{[-1,1]} \exp(\frac{\lambda}{\sigma^2}||z_1||s)(1-s^2)^{(r-3)/2}\,ds, \quad (5)$$

which shows that $q_1/D < 1$. Because q_1 and D are positive, we have $0 < q_1/D < 1$.

The integrals in (5) can be related to the modified Bessel function $I_v(u)$, a special function for which there are standard computational algorithms. There is an integral representation and a derivative formula:

$$I_v(u) = \frac{(\frac{1}{2}u)^v}{\pi^{\frac{1}{2}}\Gamma(v+\frac{1}{2})} \int_{[-1,1]} (1-s^2)^{v-\frac{1}{2}} \exp(us)\,ds \qquad (v > -\frac{1}{2}), \quad (6)$$

$$\frac{1}{u}\frac{d}{du}[u^{-v}I_v(u)] = u^{-v-1}I_{v+1}(u); \quad (7)$$

see Abramowitz and Stegun (1972, formulas 9.6.18 and 9.6.28, p. 376). Note that

$$\frac{d}{du}I_v(u) = \frac{(\frac{1}{2})^v v u^{v-1}}{\pi^{\frac{1}{2}}\Gamma(v+\frac{1}{2})} \int_{[-1,1]} (1-s^2)^{v-\frac{1}{2}} \exp(us)\,ds \quad (8)$$

$$+ \frac{(\frac{1}{2}u)^v}{\pi^{\frac{1}{2}}\Gamma(v+\frac{1}{2})} \int_{[-1,1]} s(1-s^2)^{v-\frac{1}{2}} \exp(us)\,ds,$$

$$\frac{1}{u}\frac{d}{du}[u^{-v}I_v(u)] = \frac{1}{u}[-vu^{-v-1}I_v(u) + u^{-v}\frac{d}{du}I_v(u)]. \quad (9)$$

Equations (8) and (6) imply that

$$\frac{\frac{d}{du}I_v(u)}{I_v(u)} = \frac{v}{u} + \frac{\int_{[-1,1]} s(1-s^2)^{v-\frac{1}{2}} \exp(us)\,ds}{\int_{[-1,1]} (1-s^2)^{v-\frac{1}{2}} \exp(us)\,ds}.$$

Equations (7) and (9) imply that

$$\frac{I_{v+1}(u)}{I_v(u)} = -\frac{v}{u} + \frac{\frac{d}{du}I_v(u)}{I_v(u)}.$$

So we have

$$\frac{\int_{[-1,1]} s(1-s^2)^{v-\frac{1}{2}} \exp(us)\, ds}{\int_{[-1,1]} (1-s^2)^{v-\frac{1}{2}} \exp(us)\, ds} = \frac{I_{v+1}(u)}{I_v(u)}. \tag{10}$$

Applying (10) to (5), we can express q_1/D in terms of modified Bessel functions:

$$q_1/D = I_{\frac{r}{2}}(\frac{\lambda}{\sigma^2}||z_1||)/I_{\frac{r}{2}-1}(\frac{\lambda}{\sigma^2}||z_1||).$$

Using this result in (4), when $r \geq 2$ we can write the solution in (2) as

$$\hat{\mu}^*(z_1) = H(\frac{r}{2} - 1, \frac{\lambda}{\sigma^2}||z_1||)\lambda z_1/||z_1||,$$

with

$$H(v,u) = I_{v+1}(u)/I_v(u).$$

REFERENCES

Abramowitz, M. and I. Stegun (1972): *Handbook of Mathematical Functions*. New York: Dover Publications.

Chamberlain, G. and M. Moreira (2009): "Decision Theory Applied to a Linear Panel Data Model," *Econometrica*, 77, 107–133.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014): "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 129(4), 1553–1623.

Chetty, R. and N. Hendren (2015): "The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates," Unpublished manuscript, Harvard University.

Eaton, M. (1983): *Multivariate Statistics: A Vector Space Approach*. New York: John Wiley & Sons.

Eaton, M. (1989): *Group Invariance Applications in Statistics*. Regional Conference Series in Probability and Statistics, Vol. 1. Haywood, CA: Institute of Mathematical Statistics.

James, W. and C. Stein (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press.

Sclove, S. (1968): "Improved Estimators for Coefficients in Linear Regression," *Journal of the American Statistical Association*, 63, 596–606.

Stroock, D. (1999): *A Concise Introduction to the Theory of Integration*, Third Edition. Boston: Birkhäuser.