American Economic Review: Papers & Proceedings 2016, 106(5): 400–404
http://dx.doi.org/10.1257/aer.p20161082

Fixed Effects, Invariance, and Spatial Variation in Intergenerational Mobility†

By Gary Chamberlain*

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 (with K_2 large), and x_1 includes a constant (and other variables) so that coefficients on indicator variables can be interpreted as deviations from an average coefficient.

We can impose the restriction that the same fixed effect appears in the predictors for multiple test scores:

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

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 $\tau\gamma'$.

I. Canonical Form

The model implies the following canonical form:

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

where $V_1 \sim \mathcal{N}(0, I_r \otimes I_M)$, $V_2 \sim \mathcal{N}(0, I_{n-r} \otimes I_M)$, and V_1 and V_2 are independent. The sample space is

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

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 $\mathcal{F}_{l,r}$ is the set of $r \times l$ matrices with orthonormal columns (Stiefel manifold):

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

$\mathcal{D}_l$ 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.

*Department of Economics, Harvard University, Cambridge, MA 02138 (e-mail: gary_chamberlain@harvard.edu). I thank Raj Chetty, Nathan Hendren, and Maximilian Kasy for helpful discussions.

†Go to http://dx.doi.org/10.1257/aer.p20161082 to visit the article page for additional materials and author disclosure statement.

Let $\tilde{x}_2 = x_2 - x_1 a$ denote the residual from the least-squares projection of x_2 on x_1, with $x_1'\tilde{x}_2 = 0$. The singular value decomposition 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 $s_2's_2 = I_r$. Let $\tilde{\tau} = d_2 s_2'\tau$, with the QR factorization $\tilde{\tau} = \omega\rho'$, where $\omega \in \mathcal{F}_{l,r}$ and ρ is lower triangular with positive diagonal elements. Under the normalization that $\gamma \in \mathcal{D}_l$, we have $\lambda = \gamma\rho \in \mathcal{D}_l$. ($x_1$ has rank h and $n = N - h$.)

II. Invariance

I use standard invariance arguments; see Eaton (1989). Consider the group

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

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:

$$m_1(g, z) = (g_1 z_1 g_3', g_2 z_2 g_3'),$$

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

$$m_2(g, \theta) = (g_1\omega, g_3\lambda, g_3\sigma),$$

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:

$$Z \sim P_\theta \;\Rightarrow\; g \cdot Z \sim P_{g\cdot\theta},$$

and so 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 action space is $\mathcal{R}^{r\times M}$ and the loss function (with $\Sigma = \sigma\sigma'$) is

$$L(\theta, a) = \text{trace}[\Sigma^{-1}(\mu - a)'(\mu - a)].$$

The action of G on the action space is $g \cdot a = g_1 a g_3'$. Then $L(g \cdot \theta, g \cdot a) = L(\theta, a)$, and the loss function is invariant.

An estimator maps the sample space into the action space. An estimator $\hat{\mu}$ is invariant if $\hat{\mu}(g \cdot z) = g \cdot \hat{\mu}(z)$. 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. For all $\theta \in \Theta$,

$$R((\omega, \lambda, \sigma), \hat{\mu}) = R((e_l, \sigma^{-1}\lambda, I_M), \hat{\mu}),$$

where e_l is the matrix formed from the first l columns of I_r. If $M = 1$, the risk function depends only on the scalar noncentrality parameter $\delta = \mu'\mu/\sigma^2 = \lambda^2/\sigma^2$.

III. Optimality

I construct an oracle estimator $\hat{\mu}^*$ that is allowed to depend on the following function of $\theta\colon \alpha(\theta) = (\lambda, \sigma)$. The oracle provides a lower bound on risk for invariant estimators $\hat{\mu}$:

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

Furthermore, $\hat{\mu}^*$ provides a minimax bound: for any estimator $\hat{\mu}$ (which need not be invariant),

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

Let $f(z \mid \theta)$ denote the density of P_θ. The oracle is obtained as the posterior mean of μ using the invariant prior distribution η:

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

The distribution η on $\mathcal{F}_{l,r}$ is invariant in that $U \sim \eta$ implies $g_1 U \sim \eta$ for any $g_1 \in O(r)$. If $M = 1$, there is an explicit formula for $\hat{\mu}^*$ using a modified Bessel function.

IV. Random Effects Model

I use a random-effects model based on a normal prior distribution $\tilde{\tau} \sim \mathcal{N}(0, I_r \otimes I_l)$. The key feature of this distribution is that $\tilde{\tau}(\tilde{\tau}'\tilde{\tau})^{-1/2}$ has the invariant distribution η on $\mathcal{F}_{l,r}$. The estimator is

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

with $\hat{\sigma}$ and $\hat{\gamma}$ chosen so that the estimator is invariant. If $M = 1$,

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

where the least-squares estimate of μ is $\hat{\mu}_{ls}(z) = z_1$, and the F-statistic for testing $\mu = 0$ is $F_{stat} = (z_1'z_1/r)/(z_2'z_2/(n-r))$. ($t^{+} = \max\{t, 0\}$ for $t \in \mathcal{R}$.) This estimator is in the James and Stein (1961) family of (positive-part) estimators:

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

which dominate $\hat{\mu}_{ls}$ if $r \geq 3$ and c is any number in the interval $0 < c < 2(r-2)/(n-r+2)$. 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 $n - r > 10$.

V. Application

I draw on work by Chetty et al. (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 at ages 18–23 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

$$\hat{c}_i = b_{1,od} \cdot ex_i + b_{2,od} \cdot ex_i \cdot p_i + b_{3,od}' \cdot m_i$$

($i = 1, \ldots, n_{od}$). Here ex_i is the amount of time that child i spent growing up in the destination neighborhood: $ex_i = (23 - \text{child}_i\text{'s age at move})$, 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 $b_{1,od} + 0.25 \cdot b_{2,od}$. In comparing two children (from the same cohort) for whom ex differs by one year, with both children having parents at the 0.25 quantile of the income distribution ($p = 0.25$), 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, $M = 1$, $x = x_2 = \tilde{x}_2$, x_1 is null, $K = K_2$, and $\beta = \beta_2$, which is unrestricted. The regression function is

$$E(Y_{od} \mid x) = w_{od}(\beta_d - \beta_o);$$

β provides 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. We can normalize the place effects to sum to 0, with $\beta = s_2 d_2^{-1}\mu$. 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. With college attendance (col) as the child outcome c, there are $N = 4{,}931$ 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

$$\text{SD}(\hat{\beta}_{ls}^{col,0.25}) = 0.48, \quad \text{SD}(\hat{\beta}_{re}^{col,0.25}) = 0.24.$$

With the least-squares estimate, a one standard deviation increase in a place effect corresponds to a predicted increase of 0.48 percentage points in the probability of college attendance (per year of exposure). With the random effects estimate, the predicted increase is 0.24 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 $n - r = 4{,}345$. The 0.95 interval for the noncentrality parameter δ is

$[0.76 \cdot r, 1.18 \cdot r]$. 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 $\mathrm{SD}(\hat{\beta}_{ls}^{col,0.25}) = 0.95$, $\mathrm{SD}(\hat{\beta}_{re}^{col,0.25}) = 0.47$. The population weights matter because the ratio of largest to smallest is over 2,000. The least-squares estimates of the individual place effects are all reduced by the same factor: $1 - 1/F_{stat} = 0.49$, and so the ratio of the standard deviations is the same as before: $\mathrm{SD}(\hat{\beta}_{re}^{col,0.25})/\mathrm{SD}(\hat{\beta}_{ls}^{col,0.25}) = 0.49$. 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. The invariant prior implies that the covariance matrix for β is proportional to $s_2 d_2^{-2} s_2'$. The invariant prior is not meant to be a subjective choice, motivated, for example, by exchangeability. Even if an i.i.d. specification is adopted for the unconditional distribution of place effects, we need a distribution conditional on x, as in a correlated-random effects model. 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 $\tilde{\tau}$ 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 10,000. (The diagonal matrix d_2^2 contains the nonzero eigenvalues of $\tilde{x}_2' \tilde{x}_2$.)

Now use the 0.75 quantile of the income distribution for parents and set $S_{od} = b_{1,od} + 0.75 \cdot b_{2,od}$. Using population-weighted standard deviations gives

$$\mathrm{SD}(\hat{\beta}_{ls}^{col,0.75}) = 0.40, \quad \mathrm{SD}(\hat{\beta}_{re}^{col,0.75}) = 0.19.$$

With the random effects estimate, 20 years of exposure gives a predicted increase of 3.8 percentage points in the probability of college attendance. The value for the F statistic is $F_{stat} = 1.90$ with $r = 586$ and $n - r = 4{,}345$. The 0.95 interval for δ is $[0.70 \cdot r, 1.12 \cdot r]$. 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 $p = 0.25$ or 0.75). Let var_{od} denote this estimated variance and use $w_{od} = \mathrm{var}_{od}^{-1/2}$ for the weights. This gives similar results: $\mathrm{SD}(\hat{\beta}_{ls}^{col,0.25}) = 0.48, \mathrm{SD}(\hat{\beta}_{re}^{col,0.25}) = 0.24$, $\mathrm{SD}(\hat{\beta}_{ls}^{col,0.75}) = 0.38$, $\mathrm{SD}(\hat{\beta}_{re}^{col,0.75}) = 0.16$. The F statistics are 2.01 for $p = 0.25$ and 1.71 for $p = 0.75$.

With income rank at age 26 (kr) as the child outcome c, there are $N = 3{,}094$ 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 $\mathrm{SD}(\hat{\beta}_{ls}^{kr,0.25}) = 0.33$, $\mathrm{SD}(\hat{\beta}_{re}^{kr,0.25}) = 0.038$, $\mathrm{SD}(\hat{\beta}_{ls}^{kr,0.75}) = 0.40$, $\mathrm{SD}(\hat{\beta}_{re}^{kr,0.75}) = 0.105$. Using the weights $w_{od} = \mathrm{var}_{od}^{-1/2}$ gives $\mathrm{SD}(\hat{\beta}_{ls}^{kr,0.25}) = 0.33, \mathrm{SD}(\hat{\beta}_{re}^{kr,0.25}) = 0.078$, $\mathrm{SD}(\hat{\beta}_{ls}^{kr,0.75}) = 0.39$, $\mathrm{SD}(\hat{\beta}_{re}^{kr,0.75}) = 0.052$. 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 $p = 0.25$ and 1.15 with $p = 0.75$. The random-effects estimate with $p = 0.25$ implies that a one standard deviation increase in a place effect corresponds to a predicted increase of 0.078 percentiles of income rank per year of exposure. At the 0.75 quantile of the parent income distribution, the predicted increase is 0.052 percentiles. With 20 years of exposure, the predicted increases in income rank are 1.6 and 1.0 percentiles. With $p = 0.25$, the 0.95 interval for the noncentrality parameter δ is $[0.15 \cdot r, 0.50 \cdot r]$. 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

$$Y = (Y^{col,0.25} \;\; Y^{col,0.75} \;\; Y^{kr,0.25} \;\; Y^{kr,0.75}),$$

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. Imposing the rank 2 restriction that $\beta = \tau\gamma'$, where τ is $K \times 2$ and γ' is 2×4, corresponds to separate factors for the college and income rank outcomes. The results are very similar to the unrestricted estimates. Restricting to a single factor, so τ is $K \times 1$ and γ' is 1×4, 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.

VI. 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.

REFERENCES

Chamberlain, Gary, and Marcelo J. Moreira. 2009. "Decision Theory Applied to a Linear Panel Data Model." *Econometrica* 77 (1): 107–33.

Chetty, Raj, and Nathaniel Hendren. 2015. "The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates." http://www.equality-of-opportunity.org (accessed May 4, 2015).

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel 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.

Eaton, Morris L. 1989. "Group Invariance Applications in Statistics." *Regional Conference Series in Probability and Statistics* (1): 1–133.

James, W., and Charles Stein. 1961. "Estimation with Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, edited by Jerzy Neyman, 361–79. Berkeley, CA: University of California Press.

Sclove, Stanley L. 1968. "Improved Estimators for Coefficients in Linear Regression." *Journal of the American Statistical Association* 63 (322): 596–606.