TECHNICAL WORKING PAPER SERIES

HIERARCHICAL BAYES MODELS WITH MANY INSTRUMENTAL VARIABLES

Gary Chamberlain
Guido W. Imbens

Technical Working Paper 204

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 1996

The authors thank Joshua Angrist, James Powell, and Peter Rossi for helpful comments, and thank Alan Krueger for making his data available to us. Financial support was provided by the National Science Foundation. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper 204
September 1996

HIERARCHICAL BAYES MODELS WITH MANY INSTRUMENTAL VARIABLES

ABSTRACT

In this paper, we explore Bayesian inference in models with many instrumental variables that are potentially weakly correlated with the endogenous regressor. The prior distribution has a hierarchical (nested) structure. We apply the methods to the Angrist-Krueger (AK, 1991) analysis of returns to schooling using instrumental variables formed by interacting quarter of birth with state/year dummy variables. Bound, Jaeger, and Baker (1995) show that randomly generated instrumental variables, designed to match the AK data set, give two-stage least squares results that look similar to the results based on the actual instrumental variables. Using a hierarchical model with the AK data, we find a posterior distribution for the parameter of interest that is tight and plausible. Using data with randomly generated instruments, the posterior distribution is diffuse. Most of the information in the AK data can in fact be extracted with quarter of birth as the single instrumental variable. Using artificial data patterned on the AK data, we find that if all the information had been in the interactions between quarter of birth and state/year dummies, then the hierarchical model would still have led to precise inferences, whereas the single instrument model would have suggested that there was no information in the data. We conclude that hierarchical modeling is a conceptually straightforward way of efficiently combining many weak instrumental variables.

Gary Chamberlain
Department of Economics
Littauer Center 123
Harvard University
Cambridge, MA 02138
and NBER
CHAMBER@ECGC.HARVARD.EDU

Guido W. Imbens
Department of Economics
Littauer Center 117
Harvard University
Cambridge, MA 02138
and NBER
guido_imbens@harvard.edu

HIERARCHICAL BAYES MODELS WITH MANY INSTRUMENTAL VARIABLES[1]

1. INTRODUCTION

Recently a literature has emerged focusing on models with instrumental variables that are only weakly correlated with the endogenous regressor. Part of the literature (Nelson and Startz (1990), Maddala and Jeong (1992)) has focused on sampling distributions of standard estimators such as the two-stage least squares (TSLS) estimator, and the poor approximation of those sampling distributions by a normal distribution. A second strand (Bound, Jaeger, and Baker (1995)), focusing on the case with multiple weak instrumental variables, concluded that the small sample results from an older literature (Nagar (1959), Sawa (1969)) on the bias of TSLS towards ordinary least squares (OLS) are very relevant for this case, even with many observations. To improve inference in the case with weak instruments, Bekker (1994) suggests an alternative asymptotic approximation to the sampling distribution, based on increasing the number of instrumental variables along with the sample size. Staiger and Stock (1994) suggest a third asymptotic approximation based on a vanishing correlation between the instrumental variables and the endogenous regressor. Angrist and Krueger (1995) and Angrist, Imbens, and Krueger (1995) use sample splitting ideas to propose alternative estimators that do not have the bias towards OLS that plagues TSLS.

In this paper, we suggest how Bayesian inference might proceed in models with many instruments that are potentially weakly correlated with the endogenous regressor. We develop a hierarchical model that enables us to combine many instruments. Our model is similar in spirit to the hierarchical model used by Rossi, McCulloch, and Allenby (1995) in an analysis of consumer behavior, although our focus is different. Our concern is with the effect the hierarchical structure, and in particular the choice of prior distribution, has on (1) the bias associated with the use of many instruments; and (2) the misleading appearance of

high precision in standard large-sample approximations with many instruments. We apply this model to an example that has motivated much of the theoretical work in this area, the Angrist and Krueger (1991), henceforth AK, analysis of returns to schooling using quarter of birth to form instrumental variables.

A key insight of AK was that quarter of birth might be a valid instrument for the effect of years of schooling on earnings because of the link through compulsory schooling laws. In addition to estimating linear regressions of earnings on schooling using only three quarter-of-birth dummies as instruments, they investigate models with instrumental variables formed with interactions between quarter of birth and year of birth, and between quarter of birth and state of birth, giving 180 instrumental variables. Bound, Jaeger, and Baker (1995) show that randomly generated instrumental variables, designed to match the AK data set, give results that look remarkably similar to the results based on the actual instrumental variables. In particular, inference based on a normal approximation to the sampling distribution of the TSLS estimator misleadingly suggests that the randomly generated instruments are powerful enough to reveal a precisely estimated relationship between earnings and schooling. In our investigation we use even more instruments than AK, by interacting all year of birth and state of birth pairs with quarter of birth to generate instruments. Using a hierarchical model with the AK data, we find a posterior distribution for the parameter of interest that is tight and plausible. Using data with randomly generated instrumental variables, the posterior distribution is diffuse. We find that with the AK data, most of the information can in fact be extracted by using a single instrument. Using artificial data, patterned on the AK data, we find that if all the information had been in the interactions between quarter of birth and state/year dummies, then the hierarchical model would still have led to precise inferences, whereas the single instrument model would have suggested there was no information in the data. We conclude that hierarchical modeling is a conceptually straightforward way of efficiently combining many weak instrumental variables.

2. THE MODEL

There is a family of probability distributions $\{P_\theta : \theta \in \Theta\}$, and we observe $\{Z_i\}_{i=1}^n$, where the random variables Z_i are independently and identically distributed (i.i.d.) according to P_θ for some value of θ in the parameter space Θ. To simplify notation, let Z denote a random variable that is distributed according to P_θ. Our observed variables consist of $Z = (S, Y, R, W)$, and they are the the subject of the following model:

$$S = \gamma_1' R + \gamma_2' W + V_1 \tag{1}$$

$$Y = \gamma_3' R + \beta \gamma_2' W + V_2,$$

where the disturbances (V_1, V_2) are independent of (R, W) with a bivariate normal distribution:

$$\begin{pmatrix} V_{1i} \\ V_{2i} \end{pmatrix} \mid R_i, W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma) \qquad (i = 1, \ldots, n) \tag{2}$$

under P_θ. This gives the P_θ distribution for (S, Y) conditional on (R, W), and is the basis for our likelihood function. There are proportionality restrictions in this bivariate regression model: the coefficients $(\beta\gamma_2)$ on the vector W in the regression function for Y are proportional to the coefficients (γ_2) in the regression function for S. The proportionality factor is the scalar β, a key parameter of interest. We shall assume that the P_θ distribution for (R, W) is not informative for the parameters of interest. (Let $\theta = (\theta_1, \theta_2)$, where $\theta_1 = (\gamma_1, \gamma_2, \gamma_3, \beta, \Sigma, \alpha, \Omega)$ and the hyperparameters α and Ω will be introduced below; the P_θ distribution of (R, W) only depends on θ_2, and θ_1 and θ_2 are independent under the prior distribution on Θ.)

One simple motivation for this model is based on potential outcomes. There is a potential outcome Y^t corresponding to treatment level t. The potential outcome varies linearly with the treatment:

$$Y^t = Y^0 + \beta t,$$

where Y^0 is the potential outcome with treatment level 0, and β is the effect per unit of treatment, which is the "return" to a year of schooling in our case. The potential outcome

is only observed for one of the treatment levels. This observed treatment level is S, which gives an observed outcome Y of

$$Y = Y^S = Y^0 + \beta S. \tag{3}$$

Define $X' = (R', W')$. The potential outcome Y^0 has a linear predictor $\zeta' X$ (with $\zeta := \arg\min_a E_\theta(Y^0 - a'X)^2$); then defining the disturbance $U = Y^0 - \zeta' X$ gives the orthogonal decomposition

$$Y^0 = \zeta' X + U = \zeta_1' R + \zeta_2' W + U, \quad E_\theta(XU) = 0. \tag{4}$$

The vector W of instrumental variables satisfies the exclusion restriction $\zeta_2 = 0$. R contains a constant identically equal to one, so that $E_\theta(U) = 0$.

Let $\xi' X$ denote the linear predictor of S given X. Defining the disturbance $V_1 = S - \xi' X$ gives the orthogonal decomposition

$$S = \xi' X + V_1 = \gamma_1' R + \gamma_2' W + V_1, \quad E_\theta(XV_1) = 0. \tag{5}$$

Substituting (4) and (5) into (3) (with $\zeta_2 = 0$) gives

$$Y = (\zeta_1 + \beta\gamma_1)' R + \beta\gamma_2' W + (U + \beta V_1).$$

Define $\gamma_3 = \zeta_1 + \beta\gamma_1$, $V_2 = U + \beta V_1$, $\pi_1' = (\gamma_1', \gamma_2')$, and $\pi_2' = (\gamma_3', \beta\gamma_2')$. Then we have the following reduced form:

$$\begin{aligned} S &= \pi_1' X + V_1, \quad E_\theta(XV_1) = 0 \\ Y &= \pi_2' X + V_2, \quad E_\theta(XV_2) = 0. \end{aligned} \tag{6}$$

Adding the assumption that the distribution of (V_1, V_2) conditional on X is $\mathcal{N}(0, \Sigma)$ gives our model.

Our data is a subset of the data used by AK containing males born in either the first or fourth quarters between 1930 and 1939. The outcome variable Y is the log of weekly earnings in 1979. The treatment variable S is years of school completed. The

predictor variables R consist of indicator variables $(R_j)_{j=1}^m$ based on the individual's state of birth and year of birth. There are 510 state/year cells (fifty states plus the District of Columbia and ten years); $R_{ji} = 1$ if individual i is from state/year cell j, and $R_{ji} = 0$ otherwise. We discard state/year cells with ten or fewer observations, leaving $m = 496$ cells and $n = 163,456$ observations. (This resulted in dropping 59 observations.) The vector of instrumental variables is formed by interacting quarter of birth with these state/year indicators. Let $Q_i = 1$ if individual i was born in the fourth quarter, and $Q_i = 0$ otherwise. Then $W_i' = (R_{1i}Q_i, \ldots, R_{mi}Q_i)$. This generalizes the models used by AK who only interacted quarter of birth with state *or* year dummies. This modification increases the effective number of instrumental variables from 180 to 496.

Given a prior distribution on the regression coefficients γ_1, γ_2, γ_3, β, and covariance matrix Σ, one can evaluate the posterior distribution for β. With m, the dimension of γ_2, large, however, it will be seen that a conventional choice for a "diffuse" prior distribution is in fact very informative. If the prior density for γ_2 is constant, then the prior distribution dogmatically asserts that the instrumental variables are *collectively* very powerful predictors of S. In this case with large m it is important to restrict the variability of the parameters, and the importance of the choice of prior distribution reflects this. We therefore impose a structure on the prior distribution in the form of a hierarchical (nested) model linking the reduced-form parameters for state/year cells. Let $\gamma_1 = (\gamma_{1j})_{j=1}^m$, $\gamma_2 = (\gamma_{2j})_{j=1}^m$, $\gamma_3 = (\gamma_{3j})_{j=1}^m$, and let $\gamma^j = (\gamma_{1j}, \gamma_{2j}, \gamma_{3j})'$, $\gamma = (\gamma^j)_{j=1}^m$. We assume that

$$\gamma^j \,|\, \alpha, \Omega \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\alpha, \Omega) \qquad (j = 1 \ldots, m).$$

We employ improper priors for β, Σ^{-1}, and the hyperparameter α:

$$p(\gamma, \beta, \Sigma^{-1}, \alpha, \Omega) \propto p(\gamma \,|\, \alpha, \Omega) p(\Omega) |\Sigma|^{3/2}.$$

We obtain very similar results using proper, suitably diffuse, priors for β, Σ^{-1}, α. We complete the specification of the prior distribution with a Wishart distribution for Ω^{-1}:

$$\Omega^{-1} \sim \mathcal{W}(k, H).$$

A conventional diffuse, but improper, prior for Ω^{-1} would correspond to $k = 0$, $H^{-1} = 0$. We show in the Appendix that, for a simplified version of our model, an improper prior for Ω^{-1} results in an improper posterior, corresponding to a distribution that puts all its mass on $\Omega^{-1} = \infty$ and implying that $\Omega = 0$ with posterior probability one. So we shall use a proper prior; choices for k and H are discussed below.

The Appendix describes how to sample from the posterior distribution. We use the structure of the model to set up a Gibbs sampling algorithm that will converge to the posterior distribution. See Tanner and Wong (1987), Gelfand and Smith (1990), Gelman and Rubin (1992), Chib and Greenberg (1996), and Geweke (1995) for general discussions of Gibbs sampling, and Jacquier, Polson, and Rossi (1994), Rossi, McCulloch, and Allenby (1995), and Geweke (1994) for recent empirical applications in economics.[2]

3. THE BASIC RESULTS

We shall compare the posterior distributions with inferences based on the TSLS estimator. First some notation: let $\hat{\pi}_1$ and $\hat{\pi}_2$ denote the least-squares estimates of π_1 and π_2 in (6), and set $\hat{S}_i = \hat{\pi}_1' X_i$. The TSLS estimator $\hat{\beta}_{\text{TSLS}}$ is the coefficient on $\hat{S}$ in the least-squares regression of $(Y_i)_{i=1}^n$ on $(R_i, \hat{S}_i)_{i=1}^n$. (The OLS estimator of β is the coefficient on S in the least-squares regression of $(Y_i)_{i=1}^n$ on $(R_i, S_i)_{i=1}^n$.)

Let $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{2j}$ denote the coefficients on R_j and W_j in $\hat{\pi}_1$, and let $\hat{\gamma}_{3j}$ denote the coefficient on R_j in $\hat{\pi}_2$. Let $\hat{\gamma}^j = (\hat{\gamma}_{1j}, \hat{\gamma}_{2j}, \hat{\gamma}_{3j})'$ and let $\hat{D}_\gamma$ be the sample covariance matrix of $\{\hat{\gamma}^j\}_{j=1}^m$: $\hat{D}_\gamma = \sum_{j=1}^m (\hat{\gamma}^j - \bar{\hat{\gamma}})(\hat{\gamma}^j - \bar{\hat{\gamma}})'/m$, where $\bar{\hat{\gamma}} = \sum_{j=1}^m \hat{\gamma}^j/m$. Then we specify the Wishart $\mathcal{W}(k, H)$ prior for Ω^{-1} to have $k = 3$, $H^{-1} = C \cdot k \cdot \hat{D}_\gamma$, and $C = .001$. The value $k = 3$ is the smallest value such that there is probability one that Ω is nonsingular. The the data dependent $\hat{D}_\gamma$ provides a convenient normalization; its use is not essential. We shall discuss the implications of the choice of C after presenting the main results.

We shall also compare our results with those from a nonhierarchical model that specifies an improper flat prior distribution for γ. In that case the full prior distribution has

density

$$p(\gamma, \beta, \Sigma^{-1}) \propto |\Sigma|^{3/2}.$$

We shall see that a very similar posterior distribution for β can be obtained within the hierarchical model by setting the parameter C to a large value such as $C = 1000$.

In Tables 1 and 2 we report the results for the posterior distributions. Table 1 contains the results for our subset of the AK data. The instrumental variables are based on the indicator for birth in the fourth quarter: $W_{ji} = R_{ji}Q_i$, where $Q_i = 1$ if individual i was born in the fourth quarter, $Q_i = 0$ otherwise. Table 2 contains the results for the same data, but with the actual quarter of birth replaced by randomly generated indicators, with probability .5 on the first quarter and .5 on the fourth quarter.

The first row in both tables gives summary measures of the posterior distribution for β corresponding to an improper (flat) prior on γ without the hierarchical structure. We report the mean, standard deviation, median, .025 and .975 quantiles. The real data and the random instrument data give very similar results. With the real data, there is posterior .95 probability that β lies between .058 and .089. With the random instrument data, the .95 interval is from .047 to .081. The TSLS estimator leads to a similar inference: the point estimate (standard error) is .073 (.008) for the real data and .063 (.009) for the random instrument data. This is the focus of the Bound, Jaeger, and Baker (1995) study, who argue that sampling-based inference with the random instrument data can be very misleading. The first row shows that this issue also arises when using Bayesian methods.

The second row shows that the same phenomenon of misleading inference with many weak instruments can be observed strictly within the hierarchical model for certain choices of a prior distribution. When we set the parameter C for the prior distribution on Ω to 1000, the posterior distribution for β is very similar to the flat prior, nonhierarchical results. The similarity arises because a large C implies an a priori large Ω, corresponding to an essentially flat prior distribution on γ. Conversely, a flat prior distribution on γ corresponds to an a priori large Ω.

The third row reports on the posterior distribution for β in the hierarchical model with C, the parameter of the prior distribution for Ω, equal to .001. In this case, substantial differences between the real and random instrument data sets emerge. With the real data, the posterior distribution suggests that the parameter of interest is quite precisely estimated. With random instruments, however, the posterior distribution is diffuse, with a standard deviation about ten times larger than with the real data.

These results suggest that the problem of misleading inference with many instrumental variables can be viewed in terms of the choice of prior distribution in a hierarchical model. We can consider the information content of various choices for the prior distribution and use that to decide on a suitable choice for a specific application. Since C^{-1} is a scale parameter of the prior distribution of Ω^{-1}, the quantiles of the prior distribution of $\sqrt{\Omega_{22}}$ are scaled by $\sqrt{C}$ relative to a prior with $C = 1$. In particular, with $C = 1000$, the .025 and .975 quantiles of the prior distribution for $\sqrt{\Omega_{22}}$ are approximately 14 and 1000 for both the real data and the random instrument data. With $C = .001$, these prior quantiles are approximately .014 and 1.00. The $C = 1000$ specification is a very informative prior relative to our data set. In the random instrument case, where the true value of γ_{2j} is zero for all j and so the true value of Ω_{22} is zero, the $C = 1000$ prior implies that the .025 and .975 quantiles of the posterior distribution for $\sqrt{\Omega_{22}}$ are 1.47 and 1.67, which results in a very misleading inference for β. The choice of $C = .001$ appears to give a suitably diffuse prior that avoids this problem. From the comparison of the AK data and the random QOB data, however, we cannot judge whether $C = .001$ is too small. We therefore shall set up four artificial data sets that will help us to explore the implications of various prior distributions.

Alternatively, one can argue for small values of C on the basis of the substantive problem. AK argue that quarter of birth affects years of schooling due to compulsory schooling laws. Suppose that eligibility for September enrollment in the first grade requires that the individual be born before a cutoff date. If the cutoff date is the first of January,

then individuals born in the fourth quarter ($Q = 1$) would on average be three quarters of a year younger when they start school than the individuals born in the first quarter ($Q = 0$). The other extreme is a cutoff date between April first and October first, in which case the $Q = 1$ individuals would be one quarter of a year older than the $Q = 0$ individuals. If all individuals start school in September of the first year they are eligible, and leave school at the minimum legal age, then the coefficient (γ_{2j}) on Q in the schooling regression would vary from .75 in the state/year cells with a January cutoff to -.25 in the state/year cells with an April–October cutoff. So the maximum variance of γ_{2j} would be .25, or a standard deviation of .5. This is likely to be a severe overestimate of the variance of γ_{2j}, since most people do not leave school as soon as they are legally allowed to.

A value of $C = 1$, however, corresponds to a prior .95 probability interval for $\sqrt{\Omega}_{22}$ of $(.44, 32)$, which is almost entirely to the right-hand side of the plausible upper limit of .5. The prior .95 probability interval corresponding to $C = .001$, equal to $(.014, 1.00)$, appears much more appropriate.

4. FOUR ARTIFICIAL DATA SETS

In order to investigate the properties of the posterior distribution when population values of the parameters are close to or on the boundary of the parameter space, as well as to replicate the results in the previous section, we generated four artificial data sets. In each case we fixed the sample size at $n = 162,000$ and the number of state/year cells at $m = 500$, giving 324 individuals in each cell. Within each cell, 162 individuals are born in the fourth quarter ($Q = 1$) and an equal number are born in the first quarter ($Q = 0$). The symmetry across the cells is for computational reasons, speeding up the Gibbs sampler.

The reduced form disturbances (V_1, V_2) have covariance matrix

$$\Sigma^0 = \begin{pmatrix} 10.72 & -.75 \\ -.75 & .46 \end{pmatrix}.$$

This corresponds to the posterior mean for Σ based on the AK data, with the sign of

the correlation between V_1 and V_2 reversed, so that the plim of the OLS estimate (when $\gamma_2 = 0$) is $-.75/10.72 = -.070$, in contrast to the population value of β, which we set at $\beta^0 = .098$. This allows us to see the effect of various prior distributions on the bias of the TSLS estimator as well as the effect on the standard error, when there are many instrumental variables. In all cases the proportionality restrictions in (1) are imposed, so that the coefficient on W_j in the Y equation ($\beta^0\gamma_{2j}$) is β^0 times the coefficient on W_j in the S equation (γ_{2j}).

The four data sets are generated from populations which differ in the choice of hyper-parameters α_2 and Ω_{22} corresponding to the mean and variance of γ_{2j}:

1. Nonzero Mean, Nonzero Variance

$$\alpha_2^0 = .151, \quad \sqrt{\Omega_{22}^0} = .123;$$

2. Zero Mean, Nonzero Variance

$$\alpha_2^0 = 0, \quad \sqrt{\Omega_{22}^0} = .123;$$

3. Nonzero Mean, Zero Variance

$$\alpha_2^0 = .151, \quad \sqrt{\Omega_{22}^0} = 0;$$

4. Zero Mean, Zero Variance

$$\alpha_2^0 = 0, \quad \sqrt{\Omega_{22}^0} = 0.$$

The first data set is drawn from a population that mimics the actual data. The other data sets are created by fixing the population mean and/or variance of the slope coefficients γ_{2j} to zero.[3] In data set 2, it is essential to use many instruments because the average value of γ_{2j} is equal to zero. The information on β^0 in this data set is concentrated in the interactions between quarter of birth and state/year dummy variables, and would be lost if we used Q as the only instrumental variable. The last data set, with both the mean and variance of γ_{2j} equal to zero, contains no information about β^0. It corresponds to the random instrument case. Data sets 2 and 4 represent the two dangers that are to

be avoided by choice of prior distribution, and specifically by choice of C. If the prior distribution of Ω_{22} puts too much mass close to zero, the posterior distribution for β in data set 2 will be diffuse although there is information in the data. If the prior distribution of Ω_{22} puts too much mass on large values, the posterior distribution for β in data set 4 will be misleadingly tight around the OLS estimate.

The TSLS estimates (and standard errors) for the four data sets are as follows. Data set 1: -.025 (.008); Data set 2: -.052 (.009); Data set 3: -.037 (.008); Data set 4: -.064 (.009). In all cases the TSLS estimator appears to be badly biased towards the plim of the OLS estimator (-.070) and away from the population value of $\beta^0 = .098$. The standard errors are similar in all cases, about .01. For data set 4, corresponding to the random instrument case, the TSLS inference provides a fairly tight, and again very misleading, .95 confidence interval of -.082 to -.047. We obtain very similar inferences from the posterior distribution of β if we set the parameter C of the prior distribution for Ω equal to 1000.

Table 3 reports the mean, standard deviation, median, .025 and .975 quantiles of the posterior distribution for β when the parameter C of the prior for Ω is set at .001. For data sets one and three, the posterior distribution for β is fairly concentrated around the true value. The posterior is less concentrated for data set 2, but there is still strong evidence in favor of a positive coefficient. This is interesting because here the population mean of γ_{2j} is zero, so that the information for β^0 is coming solely from the interactions between quarter of birth and state/year dummy variables. In data set 4, which contains no information about β^0, the posterior distribution is quite diffuse, with a .95 interval of -.539 to .143.

5. GAINS FROM MANY INSTRUMENTAL VARIABLES

We shall investigate whether there is a gain from using 500 instrumental variables relative to a single instrument. Is it better to ignore the potential information in the instrumental variables created by interacting quarter of birth with state and year of birth dummy variables? The answer depends upon the amount of variation in γ_{2j} across state/year cells,

where $\gamma_{1j} + \gamma_{2j}Q_i$ is the conditional mean of education for individual i in state/year cell j. The variance of γ_{2j} is Ω_{22}.

Table 4 reports the mean, standard deviation, median, .025 and .975 quantiles of the posterior distribution of $\sqrt{\Omega_{22}}$, as well as quantiles of the prior distribution.[4] The parameter C of the prior for Ω is set at .001. With the real data (row 1), the posterior .95 interval extends from .087 to .152. When the actual quarter of birth is replaced by randomly generated indicators (row 2), the posterior distribution for $\sqrt{\Omega_{22}}$ is concentrated close to zero, with the .95 interval extending from .013 to .056. The next four rows in the table correspond to the four artificial data sets from the previous section. The posterior distribution clearly distinguishes between the data sets where the true value of $\sqrt{\Omega_{22}}$ is .123 (data sets 1 and 2) and where the true value is zero (data sets 3 and 4). In data sets 2 and 4, the population mean of γ_{2j} is zero (i.e., $\alpha_2^0 = 0$), and so there is information on β only when $\Omega_{22}^0 > 0$. We saw in Table 3 that the posterior distribution for β is informative in the case of data set 2 but diffuse for data set 4. This corresponds to the posterior distribution for Ω_{22} being concentrated away from 0 for data set 2 and concentrated close to 0 for data set 4.

We can examine the gain from using many instrumental variables by using a restricted prior in which we constrain $\Omega_{22} = 0$, so that $\gamma_{2j} \equiv \alpha_2$. Then $\gamma_2'W$ in (1) (which equals $\sum_{j=1}^{m} \gamma_{2j}R_jQ$) reduces to $\alpha_2 Q$, leaving us with a single instrumental variable. Define

$$\tilde{\gamma}^j = \begin{pmatrix} \gamma_{1j} \\ \gamma_{3j} \end{pmatrix}, \quad \tilde{\Omega} = \begin{pmatrix} \Omega_{11} & \Omega_{13} \\ \Omega_{31} & \Omega_{33} \end{pmatrix}.$$

The prior distribution for $\tilde{\gamma} \equiv (\tilde{\gamma}^j)_{j=1}^m$ specifies that

$$\tilde{\gamma}^j \mid \alpha, \tilde{\Omega} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{pmatrix} \alpha_1 \\ \alpha_3 \end{pmatrix}, \tilde{\Omega}\right) \qquad (j = 1, \ldots, m).$$

We employ improper priors for Σ^{-1}, α:

$$p(\tilde{\gamma}, \beta, \Sigma^{-1}, \alpha, \tilde{\Omega}) \propto p(\tilde{\gamma} \mid \alpha, \tilde{\Omega}) p(\tilde{\Omega}) |\Sigma|^{3/2} p(\beta).$$

The prior for $\tilde{\Omega}^{-1}$ is $\mathcal{W}(k, H)$, with $k = 2$, $H^{-1} = C \cdot k \cdot \tilde{\hat{D}}_\gamma$, where $C = .001$ and $\tilde{\hat{D}}_\gamma$ is the submatrix of $\hat{D}_\gamma$ formed from the first and third rows and columns. The prior for β is improper ($p(\beta) \propto 1$) except when this gives an improper posterior for β, with the .025 and .975 quantiles tending to $-\infty$ and $+\infty$. We found that this occurs in the real data with quarter of birth replaced by randomly generated indicators (random QOB), and in artificial data sets 2 and 4, where $\alpha_2^0 = 0$. In these cases we used a proper but diffuse prior for β: $\mathcal{N}(0, 10^6)$.

Table 5 reports on the posterior distribution for β that results from the restricted prior specification. With the real data (row 1), the standard deviation of the posterior distribution is slightly larger for the restricted prior than for the general hierarchical prior (with $C = .001$) in Table 1: .022 versus .017. With artificial data sets 1 and 3, in which α_2^0 = .151, comparison with Table 3 shows that the posterior standard deviations are similar whether or not the prior imposes $\Omega_{22} = 0$. So in these cases, the tightness of the posterior distribution is not much affected by whether we use a single instrumental variable or five hundred.

The comparison is very different when $\alpha_2^0 = 0$, for then the population mean of γ_{2j} is zero and the information for β comes from the interactions of quarter of birth with the state/year dummy variables. With data set 2, the restricted prior leads to an extremely diffuse posterior distribution for β, whereas the posterior distribution in Table 3 (row 2) is very informative on the sign of β. There is a tradeoff, however, because when the population value of Ω_{22} is zero ($\Omega_{22}^0 = 0$), as with the random quarter of birth data and with data set 4, the restricted prior leads to a much more diffuse posterior than is obtained with our general hierarchical prior. The restricted prior does a better job of revealing that there is no information for β in this case. More generally, we can decrease C from its value of .001 in Table 3 to smaller, but still positive values. This will lead to better inference in data set 4, but to worse inference in data set 2. Increasing C will have the opposite effects.

6. CONCLUSION

In this paper we have shown how an analysis from a Bayesian perspective might deal with models with many, weak instruments. We develop a hierarchical model and argue that the choice of prior distribution of the variance should reflect concern with two cases. First is the case with no information in the instrumental variables. The prior can guard against misleading inference in this case by putting mass close to zero. Second, the prior distribution should reflect concern over loss of information from not using all the instruments. This suggests that the prior distribution should put weight on large values of the variance. In specific applications, the relative concern with both cases should govern the choice of prior distribution. The type of calculations performed on artificial data sets in this paper can prove useful in making such decisions.

APPENDIX

1. EVALUATING THE POSTERIOR DISTRIBUTION

Hierarchical Prior

Let $d = \{z_i\}_{i=1}^n$ denote the data, that is, the observed values of the Z_i, with $z_i' = (s_i, y_i, r_i', w_i')$. Let $x_i' = (r_i', w_i')$ and let q_i denote the observed value of the quarter-of-birth indicator Q_i. Define $f_i' = (1, q_i)$, $M_j = \sum_{i:r_{ji}=1} f_i f_i'$, and $\Lambda_j = \Sigma \otimes M_j^{-1}$. Let $\hat{\pi}_1$ and $\hat{\pi}_2$ denote the least-squares estimates of π_1 and π_2 in (6), and let $\hat{\pi}^{j\prime} = (\hat{\pi}_{1j}, \hat{\pi}_{1,m+j}, \hat{\pi}_{2j}, \hat{\pi}_{2,m+j})$. Note that

$$\hat{\pi}^j \mid \{x_i\}_{i=1}^n \overset{\text{ind}}{\sim} \mathcal{N}(T\gamma^j, \Lambda_j) \quad \text{under } P_\theta \qquad (j = 1, \ldots, m),$$

where

$$T := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \beta & 0 \end{pmatrix}. \tag{A.1}$$

In addition, since $\gamma^j \mid \{x_i\}_{i=1}^n, \beta, \Sigma, \alpha, \Omega \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\alpha, \Omega)$,

$$\hat{\pi}^j \mid \{x_i\}_{i=1}^n, \beta, \Sigma, \alpha, \Omega \overset{\text{ind}}{\sim} \mathcal{N}(T\alpha, T\Omega T' + \Lambda_j).$$

We choose starting values for β, Σ, and Ω as follows: $\beta = 0$, $\Sigma = \hat{\Sigma}$ (based on the least-squares residuals: $\hat{e}_i' = (s_i - \hat{\pi}_1' x_i, y_i - \hat{\pi}_2' x_i)$, $\hat{\Sigma} = \sum_{i=1}^n \hat{e}_i \hat{e}_i'/n$), and $\Omega = \hat{D}_\gamma$. Given these starting values, we cycle through the following four steps of the Gibbs sampler.

1. Sample from the conditional distribution of (α, γ) given d, β, Σ, Ω. First sample from the marginal distribution of α. This distribution is normal with

$$E(\alpha \mid d, \beta, \Sigma, \Omega) = \left(\sum_{j=1}^m T'(T\Omega T' + \Lambda_j)^{-1} T \right)^{-1} \sum_{j=1}^m T'(T\Omega T' + \Lambda_j)^{-1} \hat{\pi}^j,$$

$$V(\alpha \mid d, \beta, \Sigma, \Omega) = \left(\sum_{j=1}^m T'(T\Omega T' + \Lambda_j)^{-1} T \right)^{-1}.$$

Then sample from the conditional distribution of γ given α. This distribution is normal with

$$E(\gamma^j \mid \alpha, d, \beta, \Sigma, \Omega) = F_{1j}\hat{\gamma}^j + F_{2j}\alpha,$$

$$V(\gamma^j \mid \alpha, d, \beta, \Sigma, \Omega) = (T'\Lambda_j^{-1}T + \Omega^{-1})^{-1},$$

$$\mathrm{Cov}(\gamma^j, \gamma^l \mid \alpha, d, \beta, \Sigma, \Omega) = 0 \qquad (j, l = 1, \ldots, m;\ j \neq l),$$

where

$$F_{1j} = (T'\Lambda_j^{-1}T + \Omega^{-1})^{-1}T'\Lambda_j^{-1}T$$

$$F_{2j} = (T'\Lambda_j^{-1}T + \Omega^{-1})^{-1}\Omega^{-1}$$

$$\hat{\gamma}^j = (T'\Lambda_j^{-1}T)^{-1}T'\Lambda_j^{-1}\hat{\pi}^j.$$

2. Sample from the conditional distribution of Ω^{-1} given $d, \gamma, \beta, \Sigma, \alpha$.

$$\Omega^{-1} \mid d, \gamma, \beta, \Sigma, \alpha \sim \mathcal{W}(k + m, (G + H^{-1})^{-1})$$

where

$$G = \sum_{j=1}^{m} (\gamma^j - \alpha)(\gamma^j - \alpha)'.$$

3. Sample from the conditional distribution of β given $d, \gamma, \Sigma, \alpha, \Omega$. Note that

$$\hat{\pi}^j \mid \{x_i\}_{i=1}^n \overset{\text{ind}}{\sim} \mathcal{N}(a_j + g_j\beta, \Lambda_j) \quad \text{under } P_\theta,$$

where

$$a_j = \begin{pmatrix} \gamma^j \\ 0 \end{pmatrix}, \quad g_j = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \gamma_{2j} \end{pmatrix}. \tag{A.2}$$

Hence the conditional distribution of β is normal with

$$E(\beta \mid d, \gamma, \Sigma, \alpha, \Omega) = \left(\sum_{j=1}^{m} g_j'\Lambda_j^{-1}g_j\right)^{-1} \sum_{j=1}^{m} g_j'\Lambda_j^{-1}(\hat{\pi}^j - a_j)$$

$$V(\beta \mid d, \gamma, \Sigma, \alpha, \Omega) = \left(\sum_{j=1}^{m} g_j'\Lambda_j^{-1}g_j\right)^{-1}.$$

4. Sample from the conditional distribution of Σ^{-1} given $d, \gamma, \beta, \alpha, \Omega$.

$$\Sigma^{-1} \mid d, \gamma, \beta, \alpha, \Omega \sim \mathcal{W}(n, \left(\sum_{i=1}^{n} e_i e_i'\right)^{-1}) \quad \text{where} \quad e_i = \begin{pmatrix} s_i - \gamma_1' r_i - \gamma_2' w_i \\ y_i - \gamma_3' r_i - \beta \gamma_2' w_i \end{pmatrix}.$$

We use a normal approximation to this Wishart distribution, with the same first and second moments. The approximation is a very good one (n = 162,000), and it reduces the computation time.

Cycling through these four steps of the Gibbs sampler will eventually lead to draws that can be considered draws from the posterior distribution of $(\gamma, \beta, \Sigma^{-1}, \alpha, \Omega^{-1})$ given $\{Z_i\}_{i=1}^n = d$.

Restricted Prior

The results in Table 5 are obtained by modifying the previous algorithm to impose $\Omega_{22} = 0$. Given the starting values $\beta = 0$, $\Sigma = \hat{\Sigma}$, $\tilde{\Omega} = \tilde{\hat{D}}_\gamma$, we cycle through the following four steps of the Gibbs sampler.

1. Sample from the conditional distribution of $(\alpha, \tilde{\gamma})$ given $d, \beta, \Sigma, \tilde{\Omega}$. First sample from the marginal distribution of α. This distribution is normal with

$$E(\alpha \mid d, \beta, \Sigma, \tilde{\Omega}) = \left(\sum_{j=1}^{m} T'(T\Omega^* T' + \Lambda_j)^{-1} T\right)^{-1} \sum_{j=1}^{m} T'(T\Omega^* T' + \Lambda_j)^{-1} \hat{\pi}^j,$$

$$V(\alpha \mid d, \beta, \Sigma, \tilde{\Omega}) = \left(\sum_{j=1}^{m} T'(T\Omega^* T' + \Lambda_j)^{-1} T\right)^{-1},$$

where Ω^* is defined to be the 3×3 matrix with $\Omega^*_{jk} = \Omega_{jk}$ for $j, k \in \{1, 3\}$ and $\Omega^*_{jk} = 0$ if j or $k = 2$. Then sample from the conditional distribution of $\tilde{\gamma}$ given α. Note that

$$\hat{\pi}^j \mid \{x_i\}_{i=1}^n \overset{\text{ind}}{\sim} \mathcal{N}(\lambda + \tilde{T}\tilde{\gamma}^j, \Lambda_j) \quad \text{under } P_\theta,$$

where

$$\lambda = \begin{pmatrix} 0 \\ \alpha_2 \\ 0 \\ \beta\alpha_2 \end{pmatrix}, \quad \tilde{T} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

So the conditional distribution of $\tilde{\gamma}$ is normal with

$$E(\tilde{\gamma}^j \mid \alpha, d, \beta, \Sigma, \tilde{\Omega}) = F_{1j}\hat{\tilde{\gamma}}^j + F_{2j}\begin{pmatrix}\alpha_1\\ \alpha_3\end{pmatrix},$$

$$V(\tilde{\gamma}^j \mid \alpha, d, \beta, \Sigma, \tilde{\Omega}) = (\tilde{T}'\Lambda_j^{-1}\tilde{T} + \tilde{\Omega}^{-1})^{-1},$$

$$\mathrm{Cov}(\tilde{\gamma}^j, \tilde{\gamma}^l \mid \alpha, d, \beta, \Sigma, \tilde{\Omega}) = 0 \qquad (j, l = 1, \ldots, m;\ j \neq l),$$

where

$$F_{1j} = (\tilde{T}'\Lambda_j^{-1}\tilde{T} + \tilde{\Omega}^{-1})^{-1}\tilde{T}'\Lambda_j^{-1}\tilde{T}$$

$$F_{2j} = (\tilde{T}'\Lambda_j^{-1}\tilde{T} + \tilde{\Omega}^{-1})^{-1}\tilde{\Omega}^{-1}$$

$$\hat{\tilde{\gamma}}^j = (\tilde{T}'\Lambda_j^{-1}\tilde{T})^{-1}\tilde{T}'\Lambda_j^{-1}(\hat{\pi}^j - \lambda).$$

2. Sample from the conditional distribution of $\tilde{\Omega}^{-1}$ given $d, \tilde{\gamma}, \beta, \Sigma, \alpha$.

$$\tilde{\Omega}^{-1} \mid d, \tilde{\gamma}, \beta, \Sigma, \alpha \sim \mathcal{W}(k+m, (G + H^{-1})^{-1})$$

where

$$G = \sum_{j=1}^{m}(\tilde{\gamma}^j - \begin{pmatrix}\alpha_1\\ \alpha_3\end{pmatrix})(\tilde{\gamma}^j - \begin{pmatrix}\alpha_1\\ \alpha_3\end{pmatrix})'.$$

3. Sample from the conditional distribution of β given $d, \tilde{\gamma}, \Sigma, \alpha, \tilde{\Omega}$. This distribution is normal with

$$E(\beta \mid d, \tilde{\gamma}, \Sigma, \alpha, \tilde{\Omega}) = \left(\sum_{j=1}^{m} g'\Lambda_j^{-1}g + \psi\right)^{-1}\sum_{j=1}^{m} g'\Lambda_j^{-1}(\hat{\pi}^j - a_j)$$

$$V(\beta \mid d, \tilde{\gamma}, \Sigma, \alpha, \tilde{\Omega}) = \left(\sum_{j=1}^{m} g'\Lambda_j^{-1}g + \psi\right)^{-1},$$

where

$$a_j = \begin{pmatrix}\gamma_{1j}\\ \alpha_2\\ \gamma_{3j}\\ 0\end{pmatrix}, \quad g = \begin{pmatrix}0\\ 0\\ 0\\ \alpha_2\end{pmatrix},$$

$\psi = 0$ for the improper prior on β, and $\psi = 10^{-6}$ for the $\mathcal{N}(0, 10^6)$ prior.

4. Sample from the conditional distribution of Σ^{-1} given $d, \tilde{\gamma}, \beta, \alpha, \tilde{\Omega}$.

$$\Sigma^{-1} \mid d, \tilde{\gamma}, \beta, \alpha, \tilde{\Omega} \sim \mathcal{W}(n, \left(\sum_{i=1}^{n} e_i e_i'\right)^{-1}) \quad \text{where} \quad e_i = \begin{pmatrix} s_i - \gamma_1' r_i - \alpha_2 q_i \\ y_i - \gamma_3' r_i - \beta\alpha_2 q_i \end{pmatrix}.$$

Nonhierarchical Prior

Tables 1 and 2 report results for a nonhierarchical prior with $p(\gamma, \beta, \Sigma^{-1}) \propto |\Sigma|^{3/2}$. Given the starting values $\beta = 0$ and $\Sigma = \hat{\Sigma}$, we cycle through the following three steps of the Gibbs sampler.

1. Sample from the conditional distribution of γ given d, β, Σ. This distribution is normal with

$$\begin{aligned} E(\gamma^j \mid d, \beta, \Sigma) &= (T'\Lambda_j^{-1}T)^{-1}T'\Lambda_j^{-1}\hat{\pi}^j, \\ V(\gamma^j \mid d, \beta, \Sigma) &= (T'\Lambda_j^{-1}T)^{-1}, \\ \mathrm{Cov}(\gamma^j, \gamma^l \mid d, \beta, \Sigma) &= 0 \qquad (j, l = 1, \ldots, m;\ j \neq l), \end{aligned}$$

where T is defined in (A.1).

2. Sample from the conditional distribution of β given d, γ, Σ. This distribution is normal with

$$\begin{aligned} E(\beta \mid d, \gamma, \Sigma) &= \left(\sum_{j=1}^{m} g_j'\Lambda_j^{-1}g_j\right)^{-1} \sum_{j=1}^{m} g_j'\Lambda_j^{-1}(\hat{\pi}^j - a_j) \\ V(\beta \mid d, \gamma, \Sigma) &= \left(\sum_{j=1}^{m} g_j'\Lambda_j^{-1}g_j\right)^{-1}, \end{aligned}$$

where a_j and g_j are defined in (A.2).

3. Sample from the conditional distribution of Σ^{-1} given d, γ, β.

$$\Sigma^{-1} \mid d, \gamma, \beta \sim \mathcal{W}(n, \left(\sum_{i=1}^{n} e_i e_i'\right)^{-1}) \quad \text{where} \quad e_i = \begin{pmatrix} s_i - \gamma_1' r_i - \gamma_2' w_i \\ y_i - \gamma_3' r_i - \beta\gamma_2' w_i \end{pmatrix}.$$

2. IMPROPER PRIOR FOR Ω^{-1}

Consider the following special case of our model:

$$Y_i = \gamma' W_i + V_i, \quad V_i \,|\, W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \text{under } P_\theta$$
$$\gamma_j \,|\, \Omega \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Omega) \qquad (i = 1,\ldots,n;\ j = 1,\ldots,m),$$

where $\gamma = (\gamma_j)_{j=1}^m$ is an $m \times 1$ vector and Y_i is scalar. Define $\Psi = \Omega^{-1}$ and specify $\Psi \sim \mathcal{W}(k,H)$. An improper prior corresponds to $k = 0$, $H^{-1} = 0$: $p(\Psi) \propto \Psi^{-1}$. Suppose that $\sum_{i:r_{ji}=1} q_i^2 = 1$ for $j = 1,\ldots,m$ and define $\hat{\gamma}_j = \sum_{i:r_{ji}=1} q_i y_i$, $h = \sum_{j=1}^m \hat{\gamma}_j^2$.

The improper prior distribution results in an improper posterior distribution:

$$p(\Psi \,|\, d) \propto g(\Psi) := (\Psi^{-1}+1)^{-m/2} \exp[-\tfrac{1}{2}(\Psi^{-1}+1)^{-1}h]\Psi^{-1};$$
$$\int_1^\infty g(\Psi)d\Psi > 2^{-m/2}\exp(-h/2)\int_1^\infty \Psi^{-1}d\Psi = \infty.$$

Since $\int_0^1 g(\Psi)d\Psi < \infty$, we can interpret the improper posterior distribution as a distribution that puts all its mass on $\Psi = \infty$, implying that $\Omega = 0$ with posterior probability one.

FOOTNOTES

[1] The authors thank Joshua Angrist, James Powell, and Peter Rossi for helpful comments, and thank Alan Krueger for making his data available to us. Financial support was provided by the National Science Foundation.

[2] Geweke (1994) applies Gibbs sampling to a reduced-rank regression model; reduced rank is a feature of our model.

[3] For all four data sets, the other values of α^0 are $\alpha_1^0 = 12.672$ and $\alpha_3^0 = 5.879$. For data sets 1 and 2, the other values of Ω^0 are $\Omega_{11}^0 = .677$, $\Omega_{12}^0 = -.098$, $\Omega_{13}^0 = .080$, $\Omega_{23}^0 = -.011$, $\Omega_{33}^0 = .013$. For data sets 3 and 4, Ω^0 is the same as for data sets 1 and 2 except that $\Omega_{12}^0 = \Omega_{22}^0 = \Omega_{23}^0 = 0$.

[4] The prior distributions for $\sqrt{\Omega_{22}}$ differ because the data sets differ in their values of $\hat{D}_\gamma$.

REFERENCES

Angrist, J. and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Angrist, J. and A. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13, 225–235.

Angrist, J., G. Imbens, and A. Krueger (1995): "Jackknife Instrumental Variables Estimation," National Bureau of Economic Research, Technical Working Paper No. 172.

Bekker, P. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

Bound, J., D. Jaeger, and R. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Chib, S. and E. Greenberg (1996): "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 12, 409–431.

Gelfand, A. and A. Smith (1990): "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. and D. Rubin (1992): "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Geweke, J. (1994): "Bayesian Comparison of Econometric Models," Federal Reserve Bank of Minneapolis, Research Department, Working Paper 532.

Geweke, J. (1995): "Monte Carlo Simulation and Numerical Integration," Federal Reserve Bank of Minneapolis, Research Department, Staff Report 192; forthcoming in *Handbook of Computational Economics*, ed. by H. Amman, D. Kendrick, and J. Rust, Amsterdam: North-Holland.

Jacquier, E., N. Polson, and P. Rossi (1994): "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business & Economic Statistics*, 12, 371–389.

Maddala, G. S. and J. Jeong (1992): "On the Exact Small Sample Distribution of the Instrumental Variable Estimator," *Econometrica*, 60, 181–183.

Nagar, A. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Nelson, C. and R. Startz (1990): "Some Further Results on the Exact Small Sample Prop-

erties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

Rossi, P., R. McCulloch, and G. Allenby (1995): "Hierarchical Modelling of Consumer Heterogeneity: An Application to Target Marketing," in *Case Studies in Bayesian Statistics*, Volume II, *Lecture Notes in Statistics*, 105, eds. C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla, New York: Springer-Verlag, pp. 323–349.

Sawa, T. (1969): "The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," *Journal of the American Statistical Association*, 64, 923–937.

Staiger, D. and J. Stock (1994): "Instrumental Variables Regression with Weak Instruments," National Bureau of Economic Research, Technical Working Paper No. 151.

Tanner, M. and W. Wong (with discussion) (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–550.

TABLE 1

RETURNS TO SCHOOLING USING QUARTER OF BIRTH AS INSTRUMENTAL VARIABLES (REAL QOB DATA)

Hierarchical Model	C	Mean (sd)	Median	Quantile .025	Quantile .975
NO	–	.073 (.008)	.073	.058	.089
YES	1000	.074 (.008)	.074	.057	.090
YES	.001	.080 (.017)	.080	.046	.115

TABLE 2

RETURNS TO SCHOOLING USING QUARTER OF BIRTH AS INSTRUMENTAL VARIABLES (RANDOM QOB DATA)

Hierarchical Model	C	Mean (sd)	Median	Quantile .025	Quantile .975
NO	–	.064 (.009)	.064	.047	.081
YES	1000	.063 (.010)	.063	.045	.082
YES	.001	.060 (.156)	.062	-.253	.377

TABLE 3

POSTERIOR FOR β IN FOUR ARTIFICIAL DATA SETS ($\beta^0 = .098$)

Data Set	Mean (s.d.)	Median	Quantile .025	Quantile .975
1 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = .123$)	.110 (.023)	.109	.069	.158
2 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = .123$)	.191 (.071)	.181	.084	.362
3 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = 0$)	.104 (.028)	.102	.055	.164
4 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = 0$)	-.166 (.199)	-.157	-.539	.143

TABLE 4

POSTERIOR AND PRIOR FOR $\sqrt{\Omega_{22}}$

Data Set	Posterior: Mean (s.d.)	Median	Quantile .025	Quantile .975	Prior: Quantile .025	Quantile .500	.975
Real QOB	.119 (.017)	.119	.087	.152	.014	.047	.997
Random QOB	.029 (.011)	.027	.013	.056	.014	.046	.993
1 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = .123$)	.100 (.014)	.100	.074	.128	.009	.029	.636
2 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = .123$)	.085 (.017)	.084	.053	.117	.009	.029	.636
3 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = 0$)	.023 (.009)	.022	.010	.044	.009	.029	.630
4 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = 0$)	.036 (.014)	.035	.012	.065	.009	.029	.630

TABLE 5

POSTERIOR FOR β WITH RESTRICTED PRIOR ($\Omega_{22} = 0$)

Data Set	Mean (s.d.)	Median	Quantile .025	Quantile .975
Real QOB	.096 (.022)	.095	.054	.139
Random QOB	1.49 (311)	.076	-693	714
1 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = .123$)	.090 (.024)	.089	.048	.141
2 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = .123$)	-1.22 (295)	.055	-670	656
3 ($\alpha_2^0 = .151, \sqrt{\Omega_{22}^0} = 0$)	.111 (.029)	.109	.060	.173
4 ($\alpha_2^0 = 0, \sqrt{\Omega_{22}^0} = 0$)	-1.19 (332)	-.002	-785	754