PLAUSIBLY EXOGENOUS

Author(s): Timothy G. Conley, Christian B. Hansen and Peter E. Rossi

Source: *The Review of Economics and Statistics*, February 2012, Vol. 94, No. 1 (February 2012), pp. 260-272

Published by: The MIT Press

Stable URL: http://www.jstor.com/stable/41349174

PLAUSIBLY EXOGENOUS

Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi*

Abstract—Instrumental variable (IV) methods are widely used to identify causal effects in models with endogenous explanatory variables. Often the instrument exclusion restriction that underlies the validity of the usual IV inference is suspect; that is, instruments are only plausibly exogenous. We present practical methods for performing inference while relaxing the exclusion restriction. We illustrate the approaches with empirical examples that examine the effect of 401(k) participation on asset accumulation, price elasticity of demand for margarine, and returns to schooling. We find that inference is informative even with a substantial relaxation of the exclusion restriction in two of the three cases.

I. Introduction

INSTRUMENTAL variable (IV) techniques are among the most widely used empirical tools in economics. Identification of a treatment parameter of interest typically comes from an exclusion restriction: some IV has a correlation with the endogenous regressor but no correlation with the unobservables influencing the outcome of interest. Such exclusion restrictions are often debatable. Authors routinely devote a great deal of effort to convincing readers that their assumed exclusion restriction is a good approximation; that is, they argue that their instruments are plausibly exogenous. Inference about the treatment parameter is then typically conducted under the assumption that the restriction holds exactly. This paper presents an alternative approach to inference for IV models with instruments whose validity is debatable. We provide an operational definition of *plausibly* (or *approximately*) *exogenous instruments* and present simple, tractable methods of conducting inference that are consistent with instruments being only plausibly exogenous.[1]

Our definition of plausibly exogenous instruments comes from relaxing the IV exclusion restriction. We define a parameter γ that reflects how close the exclusion restriction is to being satisfied in the following model:[2]

$$Y = X\beta + Z\gamma + \varepsilon. \tag{1}$$

In this regression, Y is an outcome vector, X is a matrix of endogenous treatment variables, ε are unobservables, and Z is a matrix of instruments assumed uncorrelated with ε. When X is endogenous, the parameters β and γ are not jointly identified, so prior information or assumptions about γ are used to obtain estimates of the parameters of interest: β. The IV exclusion restriction is equivalent to the dogmatic prior belief that γ is identically 0. Our definition of *plausible exogeneity* corresponds to having prior information that implies γ is near 0 but perhaps not exactly 0. This assumption relaxes the IV exclusion assumption but still provides sufficient structure to allow inference to proceed.

We present four complementary inference strategies that use prior information about γ to differing extents. The first approach specifies only the set of possible γ values—the support of γ. Interval estimates for β, the treatment parameter of interest, can be obtained conditional on any potential value of γ. Taking the union of these interval estimates across different γ values provides a conservative (in terms of coverage) interval estimate for β. A virtue of this method is that it requires only specification of a range of plausible values for γ without requiring complete specification of a prior distribution. Its chief drawback is that the resulting interval estimates may be wide.

Our second and third strategies use prior information about the distribution of potential values of γ while stopping short of a full specification of error distributions. We view prior probabilities for γ as analogous to objective probabilities in a two-step data-generating process (DGP) where first nature draws γ according to the prior distribution, and then the data are drawn from the specified DGP given this value of γ. Interval estimates are interpreted as having a particular confidence level from an ex ante point of view for this two-step DGP.

Prior beliefs about γ are routinely held by researchers. Usual arguments employed by researchers to justify their instruments as being "plausibly exogenous" are analogous to statements of beliefs that there is a high probability that γ is near 0 and that the probability of more extreme values is diminishing. Such beliefs define a prior distribution for γ. We consider two ways to use this prior information. One is a straightforward modification of the union of confidence intervals approach mentioned above. The second uses a large-sample approximation and is very convenient.

Our fourth strategy is to undertake a full Bayesian analysis that requires priors over all model parameters (not just γ) and assumptions about the error distributions. We outline two specific ways to form priors for γ: one takes γ to be independent of the rest of the model and the other allows beliefs about γ to depend on β. Priors for γ that depend on other model parameters are much easier to handle in this Bayesian framework versus our other methods.

One key feature of our approaches is that they provide valid inference statements for any beliefs about the validity

Received for publication October 8, 2008. Revision accepted for publication August 26, 2010.

* Graduate School of Business, University of Chicago.

We thank seminar participants at the University of Chicago, Brown University, Brigham Young University, University of Wisconsin, University of Cincinnati, and Syracuse University for helpful comments. We also appreciate the comments of the referees and the editor. Funding was generously provided by the William S. Fishman Faculty Research Fund, the IBM Corporation Faculty Research Fund, and the Kilts Center for Marketing at the University of Chicago Graduate School of Business.

[1] Stata code for the methods of sections IIIA and IIIB is on Christian Hansen's Web site, currently http://faculty.chicagogsb.edu/christian.hansen/research/. R code for Bayesian inference is available in the contributed package *bayesm*.

[2] The same basic ideas we develop in this paper can also be easily applied in nonlinear structural models with potentially unidentified parameters; see Conley, Hansen, and Rossi (2007).

The Review of Economics and Statistics, February 2012, 94(1): 260–272

of the instruments.[3] It is well known that the sensitivity of the 2SLS estimator of β to violations of the exclusion restriction depends on the strength of the instruments.[4] For example, relatively minor deviations from the exclusion restriction may greatly decrease precision relative to the case where $\gamma = 0$ when instruments are weak, whereas large deviations may have only small influences on precision when the instruments are strong. The desire to use instruments that are strong but may violate the exclusion restriction provides a direct motivation for the methods of this paper. In many applications, instruments can yield informative results even under appreciable deviations from an exact exclusion restriction. We illustrate this through empirical examples.

The remainder of this paper is organized as follows. In section II, we revisit model (1) and discuss the importance of instrument strength in determining the influence of violations of the exclusion restriction. We present the inference methods formally in section III. In section IV, we illustrate our methods and sensitivity in three example applications: estimating the effect of 401(k) participation on asset accumulation motivated by Abadie (2003) and Poterba, Venti, and Wise (1995); estimating the price elasticity of demand for margarine following Chintagunta, Dubé, and Goh (2005); and estimating the returns to schooling as in Angrist and Krueger (1991). In section V, we discuss related literature, and section VI concludes.

II. Model

We are interested in estimating the parameter β in a simultaneous equation model represented in limited information form as

$$Y = X\beta + Z\gamma + \varepsilon, \tag{2}$$

$$X = Z\Pi + V, \tag{3}$$

where Y is an $N \times 1$ vector of outcomes; X is an $N \times s$ matrix of endogenous variables, $E[X\varepsilon] \neq 0$, with treatment parameter of interest β; Z is an $N \times r$ matrix of instruments where $r \geq s$ with $E[Z'\varepsilon] = 0$; Π is a matrix of first-stage coefficients; and γ is our parameter measuring the plausibility of the exclusion restriction. This model generalizes easily to allow additional predetermined or exogenous regressors.[5] The difference between the model defined above and the usual IV model is the presence of the term $Z\gamma$ in the structural equation. As discussed above, the usual IV assumption corresponds to the exclusion restriction that $\gamma \equiv 0$, which may be viewed as a dogmatic prior on γ. Our formalization of the notion of plausible exogeneity of Z corresponds to allowing deviations from this dogmatic prior on γ.

While we present the model with constant coefficients, our approach allows heterogeneous treatment effects. For example, in the model with $y_i = x_i'\beta_i + z_i'\gamma_i + u_i$ where $E[z_i u_i] = 0$, $E[x_i u_i] \neq 0$, and γ_i and β_i are jointly independent of x_i and z_i, we have that $y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$ satisfies $E[x_i\varepsilon_i] \neq 0$ and $E[z_i\varepsilon_i] = 0$ where $\varepsilon_i = u_i + x_i'(\beta_i - \beta) + z_i'(\gamma_i - \gamma)$, $\beta = E[\beta_i]$, and $\gamma = E[\gamma_i]$. Thus, the only difference between this model and the model in equation (2) is that β should be interpreted as the average treatment effect of X on Y and γ should be interpreted as the average partial effect of Z on Y.[6]

Finally, it is worth noting that the strength of the relationship between Z and X, captured by Π in equation (3), plays an important role in determining what can be learned about β in equation (2) just as it does in any IV model. The intuition for this can be seen easily in the special case where β and γ are both scalars. In this case, $\widehat{\beta} = (Z'X)^{-1}Z'Y \xrightarrow{p} \beta + \gamma/\Pi$, from which it follows that $\widehat{\beta}$ is far more sensitive to γ when Π is small. This basic intuition holds for all of the inferential approaches we consider in the following section. In particular, small ranges for plausible values of γ will lead to large decreases in the precision of inference relative to the case when $\gamma \equiv 0$ when the first-stage relationship is weak (Π is small) but may lead to only minor losses in precision when Π is large. This behavior is a manifestation of the point from Bound, Jaeger, and Baker (1995) and others that there is typically a trade-off between instrument strength and degree of violation of the exclusion restriction.

III. Inference Procedures

In this section, we consider four methods for inference about β without assuming γ is exactly 0. In the first, we assume only that the support of γ is known and consider construction of confidence regions for β by essentially taking a union of γ-specific confidence intervals. In the second and third, we view γ as a random parameter and assume beliefs about γ can be described by a proper prior distribution. We view the data-generating process as a two-stage process where a value for γ is drawn and then data are generated from equations (2) and (3) given this value of γ. We obtain frequentist confidence regions that have correct coverage from an ex ante point of view under the assumed distribution for γ. The second approach constructs a confidence region as a union of prior-weighted γ-specific confidence intervals, and the third approach employs a large sample approximation in which prior uncertainty about the exclusion restriction is modeled as being of the same order of magnitude as sampling uncertainty to obtain an approximate distribution for the treatment

[3] Note that this allows for beliefs for γ that are not centered at 0.

[4] See, for example, Angrist and Krueger (1994) and Bound, Jaeger, and Baker (1995).

[5] It is straightfoward to allow such additional regressors $\tilde{W}$ into the model by setting $\tilde{Y} = \tilde{W}B_1 + \tilde{X}\beta + \tilde{Z}\gamma + \varepsilon$ and $\tilde{X} = \tilde{W}B_2 + \tilde{Z}\Pi + V$. This model reduces to models (1) and (2) by defining Y, X, and Z as residuals from a projection on the space spanned by $\tilde{W}$, that is, as $Y = (I - P_{\tilde{W}})\tilde{Y}$, $X = (I - P_{\tilde{W}})\tilde{X}, Z = (I - P_{\tilde{W}})\tilde{Z}$.

[6] We also note that this framework applies to the case where $Y = X\beta + g(Z) + u$ and u satisifies the usual conditions. In this case, we have $\gamma = E[z_i z_i']^{-1}E[z_i g(z_i)]$, the projection coefficient of $g(Z)$ onto Z, and $\varepsilon = u + (g(Z) - Z\gamma)$, which is uncorrelated with Z. We provide an example within a simple LATE model in a working paper version of this paper: Conley et al. (2007).

effect estimator. In the fourth and final approach, we again adopt a prior distribution over γ and couple this with a prior over all the other model parameters and additional assumptions about the distribution of the unobserved errors ε and V, which allow us to pursue inference in a fully Bayesian manner.

A. *Union of Confidence Intervals with γ Support Assumption*

Our first inference method uses only a support assumption about γ. Specifically, suppose prior information consists of knowledge of the support for γ, $\mathcal{G}$, which is bounded. If the true value of γ was the value $\gamma_0 \in \mathcal{G}$, then we could subtract $Z\gamma_0$ from both sides of the equation in model (2) and estimate

$$(Y - Z\gamma_0) = X\beta + \varepsilon$$

using any estimation method based on the orthogonality of the instruments Z and errors ε. The usual asymptotic approximations could be employed to obtain a $(1-\alpha)$ confidence interval for β under the assumption that the true value of γ equals γ_0. In theory, a set of such confidence intervals could be constructed for all points in the support $\mathcal{G}$, and the union of these γ-specific confidence regions for β will have coverage of at least $(1-\alpha)$. Our approach is simply to approximate this union of confidence intervals.[7]

For ease of exposition, we present details for the two-stage least squares (2SLS) estimator of β. Under the maintained assumption that $\gamma = \gamma_0$,

$$\widehat{\beta}_N(\gamma_0) \equiv (X'P_ZX)^{-1}X'P_Z(Y - Z\gamma_0),$$

where the projection matrix $P_Z \equiv Z(Z'Z)^{-1}Z'$. Simplifying this expression yields

$$\widehat{\beta}_N(\gamma_0) = \beta + (X'P_ZX)^{-1}X'P_Z\varepsilon,$$

from which it will follow under conventional regularity conditions that

$$\sqrt{N}(\widehat{\beta}_N(\gamma_0) - \beta) \xrightarrow{d} N(0, V(\gamma_0)), \tag{4}$$

where $V(\gamma_0)$ is the usual asymptotic covariance matrix for 2SLS with $Y - Z\gamma_0$ as the dependent variable.

For simplicity, we suppose that s, the dimension of β, equals 1 in the following. The discussion generalizes immediately to $s > 1$ at the cost of complicating the notation.[8] Using equation (4), we could estimate a symmetric $(1-\alpha)$ confidence interval for β under the maintained assumption that $\gamma = \gamma_0$ in the usual way:

$$CI_N(1-\alpha, \gamma_0) = \left[\widehat{\beta}_N(\gamma_0) \pm c_{1-\alpha/2}\sqrt{\widehat{V}_N(\gamma_0)/N}\right], \tag{5}$$

where $\widehat{V}_N(\gamma_0)$ is a consistent estimator of $V(\gamma_0)$ and the critical value $c_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. Of course, the quantity in equation (5) is simply the $(1-\alpha)$ confidence interval for β constructed in the usual fashion from the output of any statistical package from the 2SLS regression of $Y - Z\gamma_0$ on X using Z as instruments. For each element of $\mathcal{G}$, we could construct such an interval and define a $(1-\alpha)$ confidence interval for β as the union of this set of confidence intervals:

$$CI_N(1-\alpha) = \cup_{\gamma_0 \in \mathcal{G}} CI_N(1-\alpha, \gamma_0). \tag{6}$$

Since we know that $\gamma \in \mathcal{G}$ and that the intervals $CI_N(1-\alpha, \gamma_0)$ were all constructed such that $Pr\{\beta \in CI_N(1-\alpha, \gamma_0)\} \to 1-\alpha$ when $\gamma = \gamma_0$, it follows immediately that asymptotically $Pr\{\beta \in CI_N(1-\alpha)\} \geq 1-\alpha$. That is, $CI_N(1-\alpha)$ will cover the true parameter value with at least probability $(1-\alpha)$ asymptotically. The interval $CI_N(1-\alpha)$ is easily approximated in practice by gridding up the support $\mathcal{G}$ and taking the union of $CI_N(1-\alpha, \gamma_0)$ over the grid points for γ_0.

A (weakly) shorter version of $CI_N(1-\alpha)$ is available if we allow the γ_0-specific intervals to be asymmetric. For each γ_0, we can define a potentially asymmetric confidence interval for β using an additional parameter $\delta(\gamma_0) \in [0, \alpha]$, which describes the degree of asymmetry in the interval, which may depend on γ_0. Under the maintained assumption that $\gamma = \gamma_0$, this confidence interval is

$$\begin{aligned} &CI_N(1-\alpha, \gamma_0, \delta(\gamma_0)) \\ &= [\widehat{\beta}_N(\gamma_0) + c_{\alpha-\delta(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}, \\ &\qquad \widehat{\beta}_N(\gamma_0) + c_{1-\delta(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}]. \end{aligned} \tag{7}$$

Again under conventional regularity conditions, it follows that $Pr\{\beta \in CI_N(1-\alpha, \gamma_0, \delta(\gamma_0))\} \to 1-\alpha$ as $N \to \infty$ if $\gamma = \gamma_0$. Likewise, we can define a $(1-\alpha)$ confidence interval for β as the union of this set of confidence intervals,

$$CI_N(1-\alpha, \delta(\cdot)) = \cup_{\gamma_0 \in \mathcal{G}} CI_N(1-\alpha, \gamma_0, \delta(\gamma_0)), \tag{8}$$

where the expression $\delta(\cdot)$ is used to denote the function mapping $\mathcal{G}$ into our asymmetry parameter. The interval $CI_N(1-\alpha, \delta(\cdot))$ has at least $(1-\alpha)$ coverage for any function $\delta(\cdot)$, and the minimum length interval can be found as the solution to the problem of minimizing the length of $CI_N(1-\alpha, \delta(\cdot))$ by choice of $\delta(\cdot)$. The shortest possible interval length is given as the solution to

$$\begin{aligned} \min_{\delta(\cdot)} &\int_{-\infty}^{\infty} 1\{b \in CI_N(1-\alpha, \delta(\cdot))\} db \\ &\text{s.t.} \quad \delta(\gamma_0) \in [0, \alpha] \text{ for all } \gamma_0 \in \mathcal{G}, \end{aligned} \tag{9}$$

[7] This approach is clearly related to the literature on bounds; see, for example, Manski (2003).

[8] An interesting issue that arises in overidentified models ($r > s$) is that one can in principle learn about subspaces of γ. One approach would be to combine one of our inference procedures with the approach of Small (2007). A fully Bayesian procedure also naturally accounts for this, and one can in principle look at the posteriors for γ.

where $1\{\cdot\}$ is the indicator function, which is 1 when the event in the braces is true.[9]

In practice, we anticipate that often there will be only modest gains from calculating the shortest interval by solving the minimization problem in equation (9) compared with the easy-to-compute union of symmetric intervals, equation (6). Such small gains are illustrated in the empirical examples.

The chief drawback of the union of confidence intervals approach is that the resulting confidence regions may be large. In a sense, this approach produces valid intervals by requiring correct coverage in every possible case, including cases that a researcher may believe unlikely. Alternatively, one may be willing to use more prior information than just the support of γ. In particular, if one is willing to assign a prior distribution over potential values for γ, intervals that use this additional information are feasible. These intervals will generally be much narrower than those produced using the bounds given above.

B. *Unions of Prior-Weighted Confidence Intervals*

A natural way to use prior information beyond a support restriction is to construct a union of confidence intervals as in the previous section, allowing oneself to weight the intervals for different values of γ depending on prior beliefs about likely values of γ. A way to achieve this weighting is by allowing the levels of the confidence intervals that go into forming the union to differ depending on the likelihood of the corresponding values of γ. In particular, we can choose low levels of confidence for unlikely values of γ and higher levels of confidence for more likely values. Under the specified distribution for γ, the union of these regions will have correct ex ante coverage and, because an additional choice variable has been added relative to the previous section, may be shorter than the bounds in section IIIA.

We begin by defining another γ_0-specific confidence interval with an additional degree of freedom, allowing the confidence level α to also depend on γ_0. Thus, we define a $(1 - a(\gamma_0))$ confidence interval for β conditional on $\gamma = \gamma_0$ as

$$\begin{aligned} &CI_N(1 - a(\gamma_0), \gamma_0, \delta(\gamma_0)) \\ &= [\widehat{\beta}_N(\gamma_0) + c_{a(\gamma_0)-\delta(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}, \\ &\qquad \widehat{\beta}_N(\gamma_0) + c_{1-\delta(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}]. \end{aligned} \tag{10}$$

Without any information about the distribution of γ beyond a support condition, the only way to ensure correct ex ante coverage of $(1 - \alpha)$ is to set $a(\cdot) \equiv \alpha$ and take the union of confidence sets as was done above. This union of confidence intervals is a natural place to start and will certainly produce a confidence region for β that has correct coverage. However, the additional information available in a specified prior over possible values for γ opens the possibility of achieving correct coverage with a shorter interval by weighting the confidence intervals according to the prior.

We define a union of prior-weighted confidence intervals as

$$CI_{F,N}(1 - \alpha, a(\cdot), \delta(\cdot)) = \cup_{\gamma_0 \in \mathcal{G}} CI_N(1 - a(\gamma_0), \gamma_0, \delta(\gamma_0)) \tag{11}$$

$$\text{subject to}$$

$$\delta(\gamma_0) \in [0, \alpha(\gamma_0)] \text{ and } \int_{\mathcal{G}} \alpha(\gamma_0) dF(\gamma_0) = \alpha,$$

where $CI_N(1 - a(\gamma_0), \gamma_0, \delta(\gamma_0))$ is a $(1 - a(\gamma_0))$(possibly asymmetric) confidence interval given $\gamma = \gamma_0$ defined by equation (10). The constraint $\int_{\mathcal{G}} \alpha(\gamma_0) dF(\gamma_0) = \alpha$ ensures ex ante coverage of $(1-\alpha)$, under the prior distribution F. Under regularity conditions, $CI_{F,N}(1 - \alpha, a(\cdot), \delta(\cdot))$ has coverage at least $(1 - \alpha)$, as $N \to \infty$ (see the appendix).

The ability to choose the level of the confidence intervals $(1 - a(\gamma_0))$ to vary according to the hypothesized value of γ_0 allows us to implicitly weight the confidence intervals obtained for different values of γ_0 according to prior beliefs about the likelihood of particular values of γ. We are able to choose low levels of confidence for unlikely values of γ and higher levels of confidence for more likely values. Under the specified distribution for γ, the union of these regions will have correct coverage and, because an additional choice variable has been added, will be (weakly) shorter than the bounds in section IIA if the functions $a(\cdot)$ and $\delta(\cdot)$ are chosen well. The choices of $a(\cdot)$ and $\delta(\cdot)$ that minimize the size of the interval solve the following problem:

$$\min_{a(\cdot),\delta(\cdot)} \int_{-\infty}^{\infty} 1\{b \in CI_{F,N}(1 - \alpha, a(\cdot), \delta(\cdot))\} db. \tag{12}$$

Note that this problem corresponds to a modification of the choice problem for the minimum-length interval given only support information. The modification is to introduce an additional free parameter $a(\cdot)$, so the interval corresponding to the solution of equation (12) will always be weakly smaller than the confidence region obtained without using the distributional information about γ.

Table 1 illustrates the difference between intervals constructed using only support information as in section IIIA and intervals that make use of a fully specified prior. We consider a highly stylized example where γ may take on one of two values: γ_1 or γ_2. With $\gamma = \gamma_1$, the estimator of β takes on a value of 1 and has a standard error of 1, and when $\gamma = \gamma_2$, the estimator of β is 4 with a standard error of 2. The columns labeled "Support Restriction" present intervals constructed using the approach of section IIIA, and the columns labeled

[9] We note that in some ways, this is similar to the problem considered in Imbens and Manski (2004). We are implicitly maintaining the assumption throughout that $E[x_i z_i]$ is full rank with minimum singular value bounded away from 0 and that the support of γ is of positive length bounded away from 0 under which there is no uniformity problem. We also note that the minimization problem (9) would ensure uniformity under much weaker conditions as it essentially adopts the solution presented in Imbens and Manski (2004).

TABLE 1.—INTERVAL ESTIMATES FOR TWO-POINT EXAMPLE

	Support Restriction		Fully Specified Prior	
	Symmetric	Asymmetric	$P(\gamma = \gamma_1) = .5$	$P(\gamma = \gamma_1) = .9$
90% Confidence Interval	$(-0.645, 7.289)$	$(-0.282, 6.759)$	$(-0.645, 6.162)$	$(-1.007, 3.179)$

Ninety percent confidence intervals for example where γ may take on two values: γ_1 and γ_2. When $\gamma = \gamma_1$, the estimate of β is 1 with a standard error of 1, and when $\gamma = \gamma_2$, the estimate of β is 4 with a standard error of 2. Intervals in the "Support Restriction" columns are obtained imposing only that γ takes on one its two possible values. "Symmetric" is the union of the two symmetric intervals, and "Asymmetric" is the minimum length union. The "Fully Specified Prior" columns report "prior weighted" unions where the column heading indicates the prior specification.

"Fully Specified Prior" present intervals that make use of prior beliefs over the probability of each potential value of γ.

We first consider the "Support Restriction" results. The union of symmetric intervals is trivially constructed by taking the usual 90% confidence interval for γ_1, $(-0.645, 2.645)$, and γ_2, (.710, 7.289), and forming the interval as the minimum of the lower end points and maximum of the upper end points. To get the length-minimizing union of intervals imposing only the support retriction, we then note that we can increase the lower end point of the γ_1 interval $(-0.645, 2.645)$ while simultaneously increasing its upper end point. Retaining 90% coverage requires that the increase in the upper end point of the γ_1 interval be larger than the increase in the lower end point of the γ_1 interval due to the shape of the normal distribution, but this is irrelevant from the standpoint of the union of intervals as long as the upper end point of the γ_1 interval remains smaller than the upper end point of the γ_2 interval. A similar argument holds for decreasing the lower and upper end points of the γ_2 interval. The length-minimizing interval occurs where the two lower end points and the two upper end points coincide. In this example, this occurs when the γ_1 interval has lower and upper tail probabilities of .099999996 and .000000004, respectively, and when the γ_2 interval has lower and upper tail probabilities of .016 and .084, respectively.

For the prior-weighted intervals, we consider two different prior specifications. In the first, we assume each potential value of γ is equally likely, and we assume γ_1 occurs with 90% probability in the second. In both cases, we see gains over the minimum length union that uses only the support restriction, with the gains being much larger with the more asymmetric prior. Specifically, the interval under equal prior probabilities is $(-0.645, 6.162)$, and the interval when γ_1 is 90% likely is $(-1.007, 3.179)$. The narrowing of the intervals is due to two factors. When prior probabilities are unequal, the length of the interval may be reduced by downweighting the unlikely γ event by substantially reducing the level and length of the associated confidence interval. This reduction in the level of the interval associated with the unlikely event can be done while maintaining ex ante coverage at the desired level with only a slight increase in the length of the other interval because the unlikely γ event has low prior probability. The second factor is that one can also play favorites in the equal probability case by downweighting the interval associated with the value of γ that produces the larger variance estimator of β. We have approximately a 95% interval for $\gamma = \gamma_1$ and an 85% interval for $\gamma = \gamma_2$ in this example. It is important to note that any prior, including the uniform, is imposing additional prior information beyond what is provided by simply specifying the support. This fact is also illustrated in the empirical examples.

C. γ *Local-to-Zero Approximation*

Our third approach uses a large-sample approximation that models uncertainty about γ as being the same order of magnitude as sampling uncertainty. The econometric jargon for this strategy is that γ is treated as being local-to-zero.[10] This treatment produces the following approximation to the distribution of $\widehat{\beta}$:

$$\begin{aligned} &\widehat{\beta} \overset{approx}{\sim} N(\beta, V_{2SLS}) + A\gamma, \\ &A = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z), \\ &\gamma \sim F. \end{aligned} \tag{13}$$

The first term in this expression, $N(\beta, V_{2SLS})$, is the usual 2SLS asymptotic distribution. V_{2SLS} is the typical variance-covariance matrix estimator for 2SLS, returned by any standard software package.[11] The second term, which is assumed independent of the first, reflects the influence of exogeneity error. The distribution of the exogeneity error term depends on sample moments in the matrix A and the specified prior distribution F for γ.

This approximation is easy to use. The approximate distribution for $\widehat{\beta}$ takes its most convenient form when one uses a Gaussian prior for γ, say, $N(\mu_\gamma, \Omega_\gamma)$. With such priors, the distribution for $\widehat{\beta}$ is, of course, Gaussian:

$$\widehat{\beta} \overset{approx}{\sim} N(\beta + A\mu_\gamma, V_{2SLS} + A\Omega_\gamma A').$$

This approximation is easily implemented with any conventional software package and a researcher-specified μ_γ and Ω_γ.

In situations that dictate a non-Gaussian prior F, confidence intervals for β are easily constructed by simulating from the distribution of deviations of $\widehat{\beta}$ from β: $\Delta = \beta - \widehat{\beta}$ where

$$\Delta \sim N(0, V_{2SLS}) + A\gamma, \quad \gamma \sim F.$$

[10] A formal statement of the asymptotic sequence and derivation of the result is provided in the appendix. The key component of the derivation is treating γ as being of the same order of magnitude as the sampling error: that is, $\gamma = \eta/\sqrt{N}$, where η follows a distribution.

[11] Standard covariance matrix estimators used in estimating V_{2SLS} remain consistent under the definition of plausible exogeneity where specification error is of the same order as sampling error. Examples include the Huber-Eicker-White estimator for independent data or a heteroskedasticity, autocorrelation consistent estimator as in Andrews (1991) for time series or Conley (1999) for spatial data.

Draws from the Δ distribution can be constructed as follows:

1. Use any standard software package to compute A and the 2SLS covariance matrix V_{2SLS}.
2. Generate one draw, Δ_1, from the desired distribution by generating a $N(0, V_{2SLS})$ draw and adding it to A times a draw from F.
3. Repeat step 2 B times for some large number B to generate a set of Δ draws: $\Delta_1, \Delta_2, \ldots, \Delta_B$.
4. Compute percentiles of the B draws to use for confidence intervals. For example, find the $\alpha/2$ and $1 - \alpha/2$ percentiles, and label them $c_{\alpha/2}$ and $c_{1-\alpha/2}$, respectively.
5. Construct a $(1 - \alpha)$, confidence interval for β as $[\widehat{\beta} - c_{1-\alpha/2}, \widehat{\beta} - c_{\alpha/2}]$.

One nice aspect of the approximate distribution in equation (13) is that the relationship between strength of instruments and the impact of exogeneity errors is transparent. A given exogeneity error γ is multiplied by A. Thus, the size of A determines how strongly exogeneity errors influence inference about β. The strength of instruments is relevant as it determines the $Z'X$ term. Weak instruments by definition have low magnitudes of $Z'X$. As $Z'X$ occurs twice in the denominator of A versus only once in its numerator, the influence of small $Z'X$ will be akin to that of dividing by a small number. Weak instruments with small $Z'X$ will therefore amplify exogeneity errors compared to strong instruments with large $Z'X$.

This approach to performing inference with plausibly exogenous instruments is appealing in that it is extremely simple to implement. In the case of a mean 0 normal prior, it requires only an adjustment to the asymptotic variance. The simplicity of the approach with this prior lends itself to examining how results vary with changes in prior beliefs, as discussed in section V. It will produce valid frequentist inference under the assumption that the prior is correct and will provide robustness relative to the conventional approach (which assumes $\gamma \equiv 0$) even when incorrect.

D. *Full Bayesian Analysis*

In the previous subsections, we have considered two types of prior information: knowledge of the support of γ and explicit prior distributions over γ. A Bayesian approach to inference is a natural complement to these methods that incorporate prior information about part of the model defined by equations (1) and (2). Of course, Bayesian inference will require priors over the other model parameters, as well as assumptions regarding the distribution of the error terms to complete the model likelihood. We let $p(Data|\beta, \gamma, \Pi, \theta)$ be the likelihood of the data conditional on the treatment and reduced-form parameters, (β, γ, Π), and the parameters characterizing the distribution of the error terms, θ. Our inference will be based on the posterior distribution for β, Π, and θ given the data, integrating out γ:

$$p(\beta, \Pi, \theta|Data) \propto \int p(Data|\beta, \gamma, \Pi, \theta) p_\gamma \times (\gamma|\beta, \Pi, \theta) p_{\{\beta,\Pi,\theta\}}(\beta, \Pi, \theta) d\gamma, \quad (14)$$

where $p_{\{\beta,\Pi,\theta\}}(\beta, \Pi, \theta)$ is the prior distribution over the model parameters and $p_\gamma(\gamma|\beta, \Pi, \theta)$ is the prior distribution over γ, which in principle is allowed to depend on all other model parameters. We note that allowing this dependence is straightforward in the Bayesian setting and allows a great deal of flexibility in the way prior information regarding the exogeneity error γ is incorporated. For example, it is simple to allow prior beliefs that γ is likely to be a small proportion of the value of β or beliefs about the exclusion restriction in terms of the unidentified population R^2 of the regression of Z on the structural error, in which case the prior would depend on the distributional parameters θ. Either of these approaches to prior information is cumbersome in the frequentist frameworks outlined previously.

In the Bayesian analyses reported in our empirical examples, we consider possible priors for γ that do and do not depend on β:

$$\text{Prior 1}: \quad \gamma \sim N(\mu, \delta^2 I). \quad (15)$$

$$\text{Prior 2}: \quad \gamma|\beta \sim N(0, \delta^2 \beta^2 I). \quad (16)$$

With prior 1, we need some idea of the size of the direct effect γ without reference to the treatment effect β. This information may come from other data sources, or we may have some intuition about potential benchmark values for γ. We anticipate using prior 2 with δ small, based on the idea that the effects of Z on Y should be smaller than the effect of X on Y and that the treatment effect could be used to benchmark γ were it available. This prior is one representation of the core idea that the exclusion restriction need not hold exactly but that deviations should be small. Prior 2 is a nonstandard prior in the Bayesian simultaneous equations literature where independent priors are typically used for model coefficients. Prior 2 assesses the conditional prior distribution of γ given β and can be coupled with a standard diffuse normal prior on β. To complete the model, we use a Gaussian distribution for (ε, V) that is independent of Z. We use this model for simplicity, and because the focus of this paper is on the prior for γ, we could of course employ Bayesian methods with any parametric likelihood or a nonparametric approach as in, for example, Conley, Rossi, and McCullogh (2006). In our examples, we use a Gaussian likelihood and priors. Computation in this case is quite similar to standard approaches in the Bayesian literature, so we leave computational details to the appendix.

We anticipate that in many applications, the Bayesian posterior intervals and the local-to-0 confidence intervals under the same prior for γ will be close. This similarity occurred in each of our empirical examples below. We suspect that

this is because the Bayesian posterior for β is largely influenced by two components: the likelihood, which by standard arguments will lead the posterior to behave similar to the asymptotic distribution for $\widehat{\beta}$ when identification is strong, and the prior over γ. Since we use the same priors over γ and identification is fairly strong based on standard criteria in at least two of our examples, the close correspondence between the two is not surprising.

IV. Examples Illustrating Inference with Alternate Priors and Instruments

We use a set of empirical examples to illustrate our methods in practice. Because of the importance of prior beliefs and because researchers will likely differ on their exact prior beliefs, it will be useful to apply our procedures under more than one assumption regarding γ and compare the resulting inference for β. For example, in the support-restriction-only approach in section IIIA, with one instrument, one could take the support of γ to be an interval $[-\delta, \delta]$ and plot a confidence interval of interest versus many different values of δ. For the other methods presented in sections IIIB through IIID, a fully specified prior distribution is required, and thus the analysis requires choosing different distributions for γ. In these cases, one can proceed by selecting a parametric family for γ and then varying the distributional parameters. For example, with one instrument, one could take γ to be normally distributed with mean 0 and variance δ^2.

We present three illustrative example applications of our methods: the effect of participating in a 401(k) plan upon accumulated assets, the demand for margarine, and the returns to schooling. We have chosen these three examples to illustrate both the breadth of potential applications for our methods and some of the variety of specifications for γ priors that may prove useful. The 401(k) application provides an example where some researchers may anticipate a violation of the exclusion restriction. Our methods can readily accommodate this by using priors for γ that are not centered at 0. In-demand estimation priors for a wholesale price direct effect, γ, might be usefully specified as depending on the price elasticity of interest. This is easily captured by using priors for γ that depend on β. Finally, the returns-to-schooling application provides an example scenario where a prior for γ can be grounded in existing research. In all applications, we assume independence across observations and estimate covariance matrices using the Huber-Eicker-White heteroskedasticity-consistent covariance matrix estimator.

A. *Effect of 401(k) Participation upon Asset Accumulation*

Our first example application examines the effect of 401(k) plan participation upon asset accumulation.[12] The outcome of interest is net financial assets (1991 dollars), and the treatment of interest is an indicator for 401(k) participation in the following regression:

$$\text{Net financial assets} = \beta \times 401(k)\ \text{participation} + X\lambda + Z\gamma + u.$$

X is a matrix of covariates with five age category indicators, seven income category indicators, family size, four education category indicators, a marital status indicator, a two-earner status indicator, a defined benefit pension indicator, an IRA participation indicator, and a home ownership indicator. The instrument Z is an indicator for 401(k) plan eligibility: whether an individual in the household works for a company with a 401(k) plan.

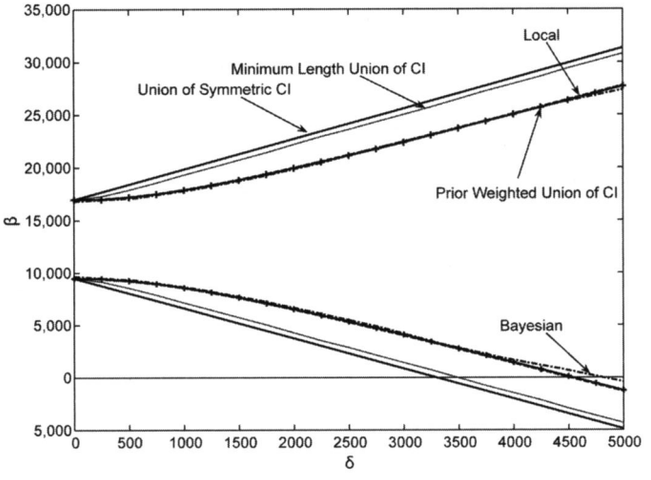
An argument for 401(k) eligibility being a valid instrument is put forth in a series of articles by Poterba, Venti, and Wise (1994, 1995, 1996). These authors argue that eligibility for a 401(k) can be taken as exogenous given income, and 401(k) eligibility and participation are of course correlated. They (1994, 1995, 1996) use this argument to estimate the effect of 401(k) eligibility on assets. For our example, we follow Abadie (2003), who uses this same basic set of arguments to motivate the use of 401(k) eligibility as an instrument for 401(k) participation. The main claim for 401(k) eligibility's being exogenous is that eligibility is determined by employers and so is plausibly taken as an exogenous conditional on covariates. Of course, the argument for the exogeneity of 401(k) eligibility is hardly watertight. For example, one might conjecture that firms that introduced 401(k)s did so due to pressure from their employees, implying that firms with plans are those with employees who really like saving.[13]

Figure 1 displays results for the full array of our methods with γ priors centered at 0. This figure plots five sets of confidence intervals for an array of assumptions about prior information indexed by the parameter δ. The widest set of solid lines presents 95% confidence intervals using the method in section IIIA with a union of symmetric γ_0-specific intervals. Their corresponding support restrictions are of the form $\gamma \in [-2\delta, +2\delta]$. The lines that lie just inside them are the minimum-length bound from section IIIA with the same $[-2\delta, +2\delta]$ support condition. The remaining three intervals are very close to each other, well within the support-restriction-only intervals. Among this set of lines, the + lines correspond to the union of prior-weighted intervals approach from section IIIB with γ prior of $N(0, \delta^2)$, the dashed lines correspond to the local-to-0 method of section IIIC with γ prior of $N(0, \delta^2)$, and the dot-dash lines are 95% Bayesian credibility intervals, .025 and .975 quantiles of the posterior for β, from section IIID, again obtained with a γ prior of $N(0, \delta^2)$.

[12] See Poterba et al. (1995), Abadie (2003), Benjamin (2003), and Chernozhukov and Hansen (2004) for further details about sample construction and descriptive statistics.

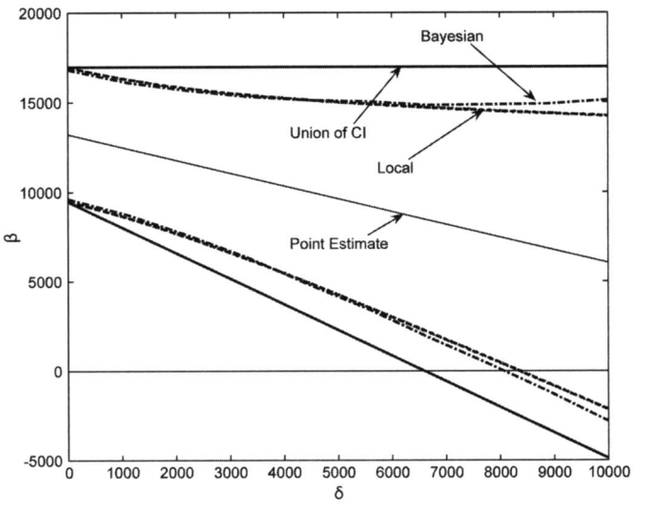
[13] See Engen, Gale, and Sholz (1996) for arguments as to why 401(k) eligibility is not a valid instrument conditional on available observables.

FIGURE 1.—95% INTERVAL ESTIMATES FROM 401(K) EXAMPLE



This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets using each of our proposed methods and across various prior settings. The definition of δ differs between the support only intervals; "Union of Symmetric CI" (dark solid line) and "Minimum Length Union of CI" (light solid line) and the intervals that use the full prior; "Prior Weighted Union of CI" (+ line), "Local" (dark dashed line), and "Bayesian" (dark dash-dot line). The intervals given by the curves "Union of Symmetric CI" and "Minimum Length Union of CI" impose only the prior information that the support of γ is $[-2\delta, 2\delta]$. For the remaining intervals, we impose the prior that $\gamma \sim N(0, \delta^2)$. The "Local" (dark-dashed line) and "Prior Weighted Union of CI" (+ line) are almost coincident and indistinguishable in the figure.

FIGURE 2.—95% INTERVAL ESTIMATES FOR 401(K) EXAMPLE WITH POSITIVE PRIOR



This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets across various prior settings using priors that the direct effect of 401(k) eligibility on net financial assets is nonnegative. The definition of δ differs between the different intervals. The "Union of CI" intervals impose only the prior information that the support of γ is $[0, \delta]$. The "Local" interval imposes the prior that $\gamma \sim U(0, \delta)$, and the "Bayesian" interval imposes γ is normally distributed with mean and variance corresponding to the mean and variance of a $U(0, \delta)$ random variable.

The dominant feature of figure 1 is that there are basically two sets of intervals, those with support conditions only and those with a prior distribution for γ. Since the intervals constructed with support conditions necessarily require bounded support, they are not strictly comparable with the Gaussian prior distributions. However, we are confident that differences in support are not the cause of this discrepancy; it is due to the introduction of the distributional information. Similar qualitative results to those with Gaussian priors obtain using uniform priors with support $[-2\delta, +2\delta]$. The coincidence of the other three intervals is a combination of their common priors and the large amount of information in the data. The two support-restriction-only intervals are close in all our example applications, so henceforth we plot only one of them to minimize clutter. Likewise, we omit plotting intervals for the prior-weighted union of confidence intervals as they are very close to the local-to-0 intervals in our applications.

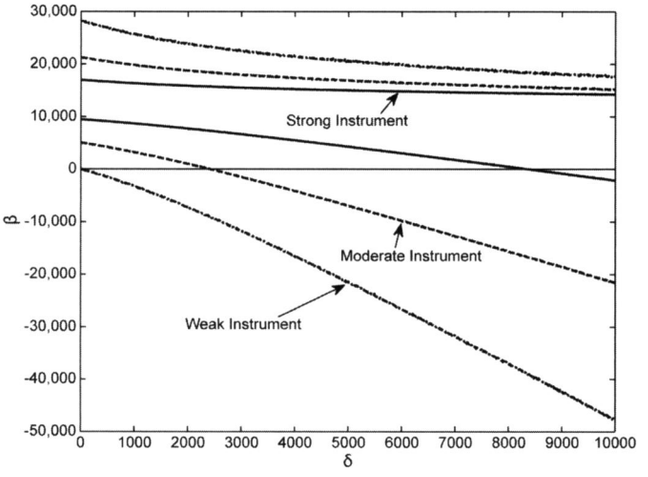
Of course, priors centered at 0 may be inappropriate in this example since many stories about violations of the exclusion restriction imply a positive direct effect of 401(k) eligibility on saving. Such beliefs are easily dealt with by using priors for γ with a positive mean. Figure 2 displays results for three methods, using priors consistent with beliefs that γ is positive. Priors are again indexed by the parameter δ. The solid lines present 95% confidence intervals using the method in section IIIA with a union of symmetric γ_0-specific intervals. Their corresponding support restrictions are of the form $\gamma \in [0, +\delta]$. The dashed lines are our local-to-0 95% confidence interval estimates from section IIIC using the prior that γ is uniformly distributed on $[0, +\delta]$. The dot-dash lines present Bayesian 95% credibility intervals from section IIID with a prior for γ that is normally distributed with the same mean and variance used for the local-to-0 estimates.[14] Since priors with positive means result in intervals shifting location, we also plot a solid line corresponding to the center point of our local-to-0 95% confidence intervals as a point estimate. The point and interval estimates of β of course shift downward as the prior mean for γ increases.

The results displayed in figure 2 suggest that there is still a significant effect of 401(k) participation on net assets, even with substantial departures from perfect instruments. For example, take the widest intervals with the support restriction of $\gamma \in [0, 4,000]$, clearly distinct from $\gamma \equiv 0$. The corresponding confidence set for β is approximately [\$3,700, \$17,000]. While certainly different from the [\$9,500, \$17,000] interval under perfect ($\gamma \equiv 0$) instruments, many would still consider the [\$3,700, \$17,000] interval evidence that β is of an economically important size.

We also use this example to illustrate how the strength of the relationship between the instruments and endogenous variables affects the analysis and the trade-offs between strength of instruments and plausibility of the exclusion restriction. In figure 3, we plot 95% interval estimates under Uniform $[0, \delta]$ priors using the local-to-0 approach for differing strengths of instruments. The instruments in this figure were generated by taking the 401(k) eligibility instrument in the data and adding noise to it in such a way that it continues to take on values of only 0 and 1. We consider "strong," "moderate," and "weak" instruments which, respectively, correspond

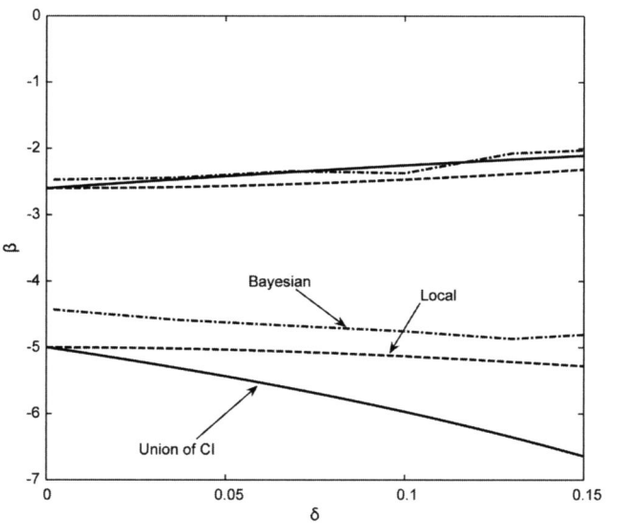
[14] $\gamma \sim N(\delta/2, \delta^2/12)$.

FIGURE 3.—95% INTERVAL ESTIMATES FOR 401(K) EXAMPLE WITH POSITIVE PRIOR AND DIFFERING STRENGTHS OF INSTRUMENTS



This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets for different strengths of instruments across various prior settings using priors that the direct effect of 401(k) eligibility on net financial assets is likely nonnegative. The figure shows "Local" interval estimates using a prior that $\gamma \sim U(0, \delta)$. The strong instrument is 401(k) eligibility from the data. The moderate and weak instruments are formed by adding noise to 401(k) eligibility.

FIGURE 4.—95% INTERVAL ESTIMATES FROM MARGARINE EXAMPLE WITH $\gamma|\beta$ PRIOR



This figure presents 95% confidence intervals for the price elasticity of the demand for margarine across various prior settings. The definition of δ differs between the different intervals. The "Union of CI" intervals impose only the prior information that the support of γ is $[-2\delta|\beta|, 2\delta|\beta|]$. The "Local" and "Bayesian" intervals impose the prior that $\gamma \sim N(0, \delta^2\beta^2)$.

to adding no noise, a moderate amount of noise, and a large amount of noise to the original instrument.[15]

Looking at figure 3, we can clearly see the influence of the strength of the first-stage relationship on inference for the structural parameter of interest. As indicated earlier, the width of the confidence set increases more rapidly for weaker instruments. It is interesting that the confidence set for β with the weak instrument under the assumption that the instrument is perfect ($\delta = 0$) is wider than the confidence set for β with the strong instrument even when allowing for the direct effect of 401(k) eligibility to be as large as \$10,000 with uniform beliefs over [0, 10,000]. These gains to using the stronger instrument become even more pronounced once one starts allowing even modest violations of the exclusion restriction with the weak instrument. While this example is obviously artificial, it clearly illustrates the tradeoffs between the strength and plausibility of instruments and certainly illustrates that in some scenarios, it will be preferable, in terms of learning about β, to use a strong instrument that may not satisfy the exclusion restriction rather than a weak one that does.

B. *Price Elasticity of Demand for Margarine*

Our second example application concerns price endogeneity in demand estimation, a canonical econometric problem. We use as our example the problem of estimating demand for margarine using the data of Chintagunta et al. (2005). The sample consists of weekly purchases and prices for the four most popular brands of margarine in the Denver area for 117 weeks from January 1993 to March 1995. Pooling across brands, we estimate the following model,

$$\log(Share) = \beta \log(retail\ price) + X\lambda + Z\gamma + u,$$

where X includes brand indicators, feature and display indicators, and their interactions with brand indicators. Following Chintagunta et al. (2005), we use log wholesale prices as an instrument, Z, for retail prices.

The argument for plausible exogeneity of wholesale prices is that they should primarily vary in response to cost shocks and should be much less sensitive to retail demand shocks than retail prices. In this example application, we illustrate the use of priors for γ that depend on β. It seems quite possible that researchers would be comfortable assuming that the direct effect of a wholesale price could be benchmarked relative to the elasticity with respect to the corresponding retail price.

Figure 4 displays results for three methods, using priors for γ that depend on β. Priors are again indexed by the parameter δ. The solid lines present 95% confidence intervals using the method in section IIIA with a union of symmetric γ_0-specific intervals with support restriction $\gamma \in [-2\delta\beta, +2\delta\beta]$. The dashed lines are our local-to-0 95% confidence interval estimates from section IIIC using a prior that γ given β is distributed $N(0, \delta^2\beta^2)$.[16] The dot-dash lines present Bayesian

[15] We generate new instruments z^* as $z^* = \phi z + (1 - \phi)w$ where z is 401(k) eligibility, ϕ is a Bernoulli(p) random variable, w is Bernoulli($\bar{z}$) random variable, $\bar{z}$ is the sample mean of z, and ϕ and w are independent. We consider three strengths of instruments: "strong" instruments with $p = 1$, "moderate" instruments with $p = .5$, and "weak" instruments with $p = .3$. By usual measures of instrument strength, none of the settings corresponds to weak instruments. For example, the first-stage F-statistics for the strong, moderate, and weak settings are, respectively, 7,767.9, 1,094.9, and 351.8. We use the terminology to simply denote the relative strength of the instruments.

[16] This prior for the local-to-0 approach is implemented using the consistent 2SLS estimator $\hat{\beta}_{2SLS}$. In other words, our prior distribution is specified to be $N(0, \delta^2\hat{\beta}^2_{2SLS})$.

95% credibility intervals from section IIID with a prior that the distribution of γ given β is $N(0, \delta^2\beta^2)$.

Unlike the 401(k) example considered above, there is a notable difference between all the intervals in figure 4. For most values of δ, the Bayesian intervals are smaller than both of the others. The two factors of support differences and information in a full prior distribution of course drive some of the discrepancy between the Bayesian and support-restriction-only intervals. An additional source of discrepancy that is likely more important here than in the previous example application is the relatively small amount of information in the data. This leads us to believe that much of the discrepancy between the Bayesian intervals and those of the local-to-0 approximation is due to small sample effects.

A qualitative conclusion from figure 4 that is common across methods is that there can be a substantial violation of the exclusion restriction without a major change in the demand elasticity estimates. Inferences change little for a range of direct wholesale price effects up to 10% of the size of the retail price effect. Take, for example, the local-to-0 estimates, at $\delta = 0$ the 95% confidence interval is $(-5, -2.5)$ and at $\delta = 10\%$ the 95% confidence interval is $(-5.5, -2.3)$. Put on a standard markup basis using the inverse elasticity rule, the corresponding mark-up intervals are [20% to 40%] and [18% to 44%]. For many, if not all, purposes, this is a small change in the implied mark-ups.

C. *Returns to Schooling*

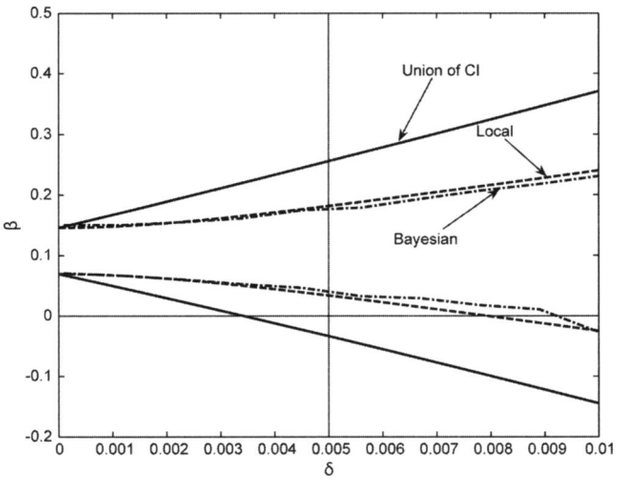
Our final example application is estimating the returns to schooling using quarter of birth as instruments, as in Angrist and Krueger (1991). The sample consists of 329,509 males from the 1980 U.S. Census who were born between 1930 and 1939. For this illustration, we estimate the following model determining log wages:

$$\log(Wage) = \beta School + X\lambda + Z\gamma + u,$$

where the dependent variable is the log of the weekly wage, *School* is reported years of schooling, and X is a vector of covariates consisting of state and year of birth fixed effects. To sidestep weak and many instrument issues, we use only the three-quarter of birth indicators, with being born in the first quarter of the year as the excluded category, as instruments Z, and do not report results using interactions between quarter of birth and other regressors.[17]

Angrist and Krueger (1991) argue that quarter of birth is a valid instrument, correlated with schooling attainment and uncorrelated with unobserved taste or ability factors that influence earnings. Angrist and Krueger (1991) examine data from three decennial censuses and find that people born in the first quarter of the year have less schooling on average than those born later in the year. This correlation of quarter of birth and schooling is uncontroversial. However, there is considerable debate about these instruments' validity due to correlation between birth quarter and other determinants of wages (Bound & Jaeger, 1996; Bound, Jaeger, & Baker, 1995). Bound et al. (1995) go beyond this in providing well-motivated back-of-the-envelope calculations of a plausible range for direct effects of quarter of birth on wages. They come up with an approximate magnitude of a direct effect of quarter of birth on wages of about 1%. Such calculations are directly useful in our framework, informing our choice of prior for γ.

FIGURE 5.—95% INTERVAL ESTIMATES FROM RETURNS-TO-SCHOOLING EXAMPLE



This figure presents 95% confidence intervals for the returns to schooling across various prior settings. The definition of δ differs between the different intervals. The "Union of CI" intervals impose only the prior information that the γ takes on values within the cube $[-2\delta, 2\delta]^3$. The "Local" and "Bayesian" intervals impose the prior that $\gamma \sim N(0, \delta^2 I_3)$ where I_3 is a 3×3 identity matrix.

Results are displayed in figure 5. This figure plots three sets of confidence intervals for an array of assumptions about prior information indexed by the parameter δ. The solid lines represent 95% confidence intervals using the method in section IIIA with a union of symmetric γ_0-specific intervals. Their corresponding support restrictions are of the form $\gamma \in [-2\delta, +2\delta]^3$. The dashed lines present 95% confidence intervals for our local-to-0 method in section IIIC using priors for γ that are $N(0, \delta^2 I)$. Finally, the dot-dash lines present Bayesian 95% credibility intervals using the model in section IIID with $N(0, \delta^2 I)$ priors for γ. The vertical line at $\delta = .005$ provides a reference point for priors motivated by the Bound et al. approximate magnitude for the direct effect of quarter of birth of 1%.

The intervals in figure 5 suggest that the data are essentially uninformative about the returns to schooling under priors consistent with the evidence in Bound et al. (1995). Using the Bound et al. (1995) calculations as an upper bound on the magnitude of γ would require us to focus attention in a δ range near .005. At $\delta = .005$, the local-to-0 95% confidence interval for β is [3.4% to 18.3%], which we consider uninformative about the returns to years of school. In order for these

[17] The first-stage F-statistic from the specification with three instruments is 36.07, which is well within the range where one might expect the usual asymptotic approximation to perform adequately.

confidence intervals to be informative in our judgment, prior beliefs regarding γ must be much more concentrated near 0. For example, using the support-restriction-only intervals, one would need to be sure that the magnitude of γ was less than .002 to obtain a confidence interval for β that excluded 5%.

V. Other Approaches

There are, of course, many approaches to performing sensitivity analysis and for dealing with unidentified parameters in the statistics and econometrics literatures. The methods we discuss in this paper complement these approaches and provide economists tractable ways to proceed when using less-than-perfect instruments and to summarize changes to inference across a variety of prior beliefs about the unidentified parameter γ.

Many classical approaches to sensitivity analysis focus on the bias of a point estimator of a parameter of interest and investigate how the bias of the estimator relates to unidentified nuisance parameters. (See, for example, Rosenbaum, 2002, for a textbook discussion of this approach, as well as Rosenbaum, 1987, Gastwirth, Krieger, & Rosenbaum, 1998, and Imbens, 2003, for specific examples. This basic approach has also been considered in the IV framework by Angrist, Imbens, & Rubin, 1996, and Hahn & Hausman, 2003.

Relative to the approaches noted, the methods we advocate not only address the changes in location of the parameter estimate but also provide simple ways to combine the information in the data with beliefs about γ to obtain valid inferential statements, such as confidence intervals, about the parameter of interest, β. As such, the methods we advocate are clearly related to the broader literature regarding obtaining bounds for partially identified parameters. Manski (2003) provides an overview of such approaches, and Horowitz and Manski (1995), Hotz, Mullin, and Sanders (1997), and Manski and Pepper (2000) provide interesting examples. More recently, Chernozhukov, Hong, and Tamer (2007) provide a very general framework for obtaining inferential statements in models in which parameters are set identified rather than point identified. In contemporaneous research,[18] Small (2007) provides an approach to obtaining bounds in overidentified linear IV models. Nevo and Rosen (2008) also consider bounds in linear IV models assuming the signs of correlations among unobservables and instruments are known. As with Small (2007) and Nevo and Rosen (2008), our bounds approach in section IIIA is clearly a special case of the general bounds framework, which is easy to implement and involves placing support restrictions over an easily interpretable parameter.

A drawback of the bounds approaches is that they do not make use of any beliefs about the likelihood of possible violations of the exclusion restriction that a researcher might have, and using these beliefs may help refine inference. Incorporating completely specified prior distributions for unidentified parameters is easy in a Bayesian framework. Priors over unidentified parameters have been used in, for example, Frangakis and Rubin (2002), Hirano et al. (2000), and Imbens and Rubin (1997) in treatment effects contexts. Our approaches in sections IIIB to IIID also make use of completely specified priors for γ in an IV model and provide a useful complement to the existing literature.

[18] See also Berkowitz, Caner, and Fang (2006).

VI. Conclusion

When using IV methods, researchers routinely provide informal arguments that their instruments satisfy the instrument exclusion restriction but recognize that this may only be approximately true. However, inference in these settings then typically proceeds under the assumption that the IV exclusion restriction holds exactly. We have presented alternative approaches to inference that do not impose the assumption that instruments exactly satisfy an exclusion restriction; they need only be plausibly exogenous. Our methods provide an improved match between researchers' assumptions of plausible exogeneity and their methods of inference.

All of our approaches involve using some sort of prior information regarding the extent of deviations from the exact exclusion restriction. Many of the usual arguments that researchers use to justify exclusion restrictions are naturally viewed as providing information about prior beliefs about violation of these restrictions. Our contribution is to provide a practical method of incorporating this information. We provide a tool set for applied researchers to conduct inference about parameters of interest even when the the set of available instruments is imperfect. We demonstrate the utility of our approach through three empirical applications. While decreasing inference precision, inference regarding the parameters of interest remains economically informative under a priori moderate violations of the exclusion restriction in two of the three applications. Useful inference is clearly feasible with instruments that are only plausibly exogenous.

Overall, our methods provide tools to applied researchers that allow them to expand the set of instruments they consider and the set of problems they tackle. Since our methods account for both sampling uncertainty and uncertainty about the validity of the instruments, they allow researchers to do inference about treatment effects even when instruments may not be perfect. Viewing research in this way shifts the focus from finding instruments that are perfect to finding instruments that allow economically informative inference after accounting for their possible imperfection.

REFERENCES

Abadie, A., "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics* 113 (2003), 231–263.

Andrews, D. W. K., "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica* 59 (1991), 817–858.

Angrist, J. D., G. W. Imbens, and D. B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91 (1996), 444–455.

Angrist, J. D., and A. Krueger, "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (1991), 979–1014.
Angrist, J. D., and A. Krueger, "Why Do World War II Veterans Earn More Than Nonveterans?" *Journal of Labor Economics* 12 (1994), 74–97.
Benjamin, Daniel J., "Do 401(k)s Increase Saving? Evidence from Propensity Score Subclassification," *Journal of Public Economics* 87 (2003), 1259–1290.
Berkowitz, D., M. Caner, and Y. Fang, "Are Nearly 'Exogenous Instruments' Reliable?" Mimeograph (2006).
Bound, J., and D. A. Jaeger, "On the Validity of Season of Birth as an Instrument in Wage Equations: A Comment on Angrist and Krueger's 'Does Compulsory Attendance Affect Schooling and Earnings?"' NBER working paper no. 5835 (1996).
Bound, J., D. A. Jaeger, and R. M. Baker, "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 (1995), 443–450.
Chernozhukov V., and C. Hansen, "The Impact of 401K Participation on Savings: An IV-QR Analysis," this REVIEW 86 (2004), 735–751.
Chernozhukov, V., H. Hong, and E. Tamer, "Estimation and Inference on Identified Parameter Sets," *Econometrica* 75 (2007), 1243–1284.
Chintagunta, P. K., J. P. Dubé, and K. Y. Goh, "Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household-Level Brand Choice Models," *Management Science* 51 (2005), 832–849.
Conley, T. G., "GMM with Cross Sectional Dependence," *Journal of Econometrics* 92 (1999), 1–45.
Conley, T., C. Hansen, and P. Rossi, "Plausibly Exogenous," SSRN working paper (2007).
Conley, T., C. Hansen, P. Rossi, and R. E. McCullogh, "A Non-Parametric Bayesian Approach to the Instrumental Variable Problem," University of Chicago, working paper (2006).
Engen, E. M., W. G. Gale, and J. K. Scholz, "The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives* 10 (1996), 113–138.
Frangakis, C., and D. Rubin, "Principal Stratification in Causal Inference," *Biometrics* 58 (2002), 21–29.
Gastwirth, J. L., A. M. Krieger, and P. R. Rosenbaum, "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika* 85 (1998), 907–920.
Hahn, J., and J. A. Hausman, "IV Estimation with Valid and Invalid Instruments," MIT Department of Economics working paper no. 03-26 (2003).
Hirano, K., G. Imbens, D. Rubin, and X. Zhou, "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics* 1 (2000), 69–88.
Horowitz J., and C. F. Manski, "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica* 63 (1995), 281–302.
Hotz, V. J., C. Mullin, and S. Sanders, "Bounding Causal Effects from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing," *Review of Economic Studies* 64 (1997), 575–603.
Imbens, G. W., "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review, Papers and Proceedings* 93 (2003), 126–132.
Imbens, G. W., and C. F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica* 72 (2004), 1845–1857.
Imbens, G. W., and D. Rubin, "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *Annals of Statistics* 25 (1997), 305–327.
Manski, C. F., *Partial Identification of Probability Distributions* (Berlin: Springer-Verlag, 2003).
Manski, C. F., and J. V. Pepper, "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica* 68 (2000), 997–1010.
Nevo, A., and A. Rosen, "Inference with Imperfect Instrumental Variables," mimeograph (2008).
Poterba, J. M., S. Venti, and D. Wise, "Targeted Retirement Saving and the Net Worth of Elderly Americans," *American Economic Review* 84 (1994), 180–185.
——— "Do 401(k) Contributions Crowd Out Other Private Saving?" *Journal of Public Economics* 58 (1995), 1–32.
——— "How Retirement Saving Programs Increase Saving," *Journal of Economic Perspectives* 10 (1996), 91–112.
Rosenbaum, P. R., "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika* 74 (1987), 13–26.
——— *Observational Studies*, 2nd ed. (Berlin: Springer-Verlag, 2002).
Rossi, P. G. Allenby, and R. E. McCulloch, *Bayesian Statistics and Marketing* (Hoboken, NJ: Wiley, 2005).
Small, D., "Sensitivity Analysis for Instrumental Variables Regression with Overidentifying Restrictions," *Journal of the American Statistical Association* 102 (2007), 1049–1058.
White, H., *Asymptotic Theory for Econometricians*, rev. ed. (San Diego: Academic Press, 2001).

APPENDIX

A.1. Regularity Conditions for Convergence of Union of Prior-Weighted Confidence Intervals

Standard regularity conditions[19] and continuity of $a(\cdot), a(\cdot)$ are sufficient for $CI_{F,N}(1-\alpha, a(\cdot), a(\cdot))$ defined by equation (11) to have proper limiting coverage. To see this, note,

$$\begin{aligned}&\Pr\{\beta \in CI_N(\gamma_0) \mid \gamma = \gamma_0\}\\ &\quad = \Pr\{z_{\alpha(\gamma_0)-a(\gamma_0)} \le -\widehat{V}(\gamma_0)^{-1/2}\sqrt{N}(\widehat{\beta}(\gamma_0) - \beta) \le z_{1-a(\gamma_0)} \mid \gamma = \gamma_0\}\\ &\quad = G_N(z_{1-a(\gamma_0)}) - G_N(z_{\alpha(\gamma_0)-a(\gamma_0)}),\end{aligned}$$

where G_n does not depend on γ since

$$\begin{aligned}&\widehat{\beta}(\gamma_0) - \beta \mid \gamma = \gamma_0 \text{ is } (X'P_ZX)^{-1}X'P_z\varepsilon\\ &\text{and } \widehat{V}(\gamma_0) = h(X, Z, e(\gamma_0))\\ &\text{where } e(\gamma_0) \mid \gamma = \gamma_0 \text{ is } Y - Z\gamma_0 - X\widehat{\beta}(\gamma_0) = \varepsilon - X(X'P_zX)^{-1}X'P_z\varepsilon.\end{aligned}$$

Also, under standard regularity conditions,

$$\begin{aligned}&-\widehat{V}(\gamma_0)^{-1/2}\sqrt{N}(\widehat{\beta}(\gamma_0) - \beta)\\ &\quad = h(X, Z, e(X, Z, \varepsilon))^{-1/2}\left(\frac{X'P_zX}{N}\right)\frac{1}{\sqrt{N}}X'P_z\varepsilon \xrightarrow{d} N(0,1)\\ &\quad \Longrightarrow G_N(w) \longrightarrow \Phi(w) \text{ pointwise for all } w.\end{aligned}$$

Now

$$\begin{aligned}\Pr\{\beta \in CI_{F,N}\} &= \int \Pr\{\beta \in CI_{F,N} \mid \gamma = \gamma_0\}dF(\gamma_0)\\ &\ge \int \Pr\{\beta \in CI_N(\gamma_0) \mid \gamma = \gamma_0\}dF(\gamma_0)\\ &= \int [G_N(z_{1-a(\gamma_0)}) - G_N(z_{\alpha(\gamma_0)-a(\gamma_0)})]dF(\gamma_0)\\ &\longrightarrow \int [1 - a(\gamma_0)]dF(\gamma_0) - \int [\alpha(\gamma_0) - a(\gamma_0)]dF(\gamma_0)\\ &= 1 - \alpha,\end{aligned}$$

where the first inequality follows because $\{\beta \in CI_N(\gamma_0) \mid \gamma = \gamma_0\}$ implies $\{\beta \in CI_{F,N} \mid \gamma = \gamma_0\}$; the interchange of the limit and integral follows from $|G_N(h(\gamma))dF| \le dF$, which is integrable, $\alpha(\cdot), a(\cdot)$ continuous, and convergence a.e.; and the last equality from $\int \alpha(\gamma_0)dF(\gamma_0) = \alpha$ by construction.

A.2. Behavior of 2SLS Estimator under γ Local-to-0 Approximation

To obtain the approximation in section IIIC, we model γ as being local to 0.[20] Explicitly referencing the dependence of γ on the sample size using the subscript N, we represent γ in structural equation (2) as

[19] For example, one could assume the following: As $N \to \infty$, the following convergence results hold jointly: (a) $Z'Z/N \xrightarrow{p} M_{ZZ}$, for $M_{ZZ} = \lim E\{Z'Z/N\}$, a positive-definite matrix; (b) $Z'X/N \xrightarrow{p} M_{ZX}$, for $M_{ZX} = \lim E\{Z'X/N\}$, a full-rank matrix; (c) $Z'\varepsilon/N \xrightarrow{p} 0$, and $Z'\varepsilon/\sqrt{N} \xrightarrow{d} N(0, V)$ for $V = \lim E\{Z'\varepsilon\varepsilon'Z/N\}$.

[20] It is a straightforward extension to model γ as being local to any known value.

$$\gamma_N = \eta/\sqrt{N} \text{ where } \eta \sim G. \tag{A1}$$

We assume η is independent of X, Z, and ε. In our approach, we equate prior information about plausible values of γ with knowledge of the distribution G. This approach differs from other local approaches in that we do not treat η as a constant but rather as a random variable. This produces limiting behavior in which not just the location but also the shape of the asymptotic distribution is influenced by the uncertainty about the value of γ.

The normalization by $\sqrt{N}$ in the definition of γ_N is designed to produce asymptotics in which the uncertainty about exogeneity and usual sampling error is of the same order, and so both factor into the asymptotic distribution. If instead γ_N were equal to η/N^b for $b < 1/2$, the asymptotic behavior would be determined completely by the exogeneity error η/N^b, and if b were greater than $1/2$, the limiting distribution would be determined completely by the usual sampling behavior. The modeling device we use may be regarded as a thought experiment designed to produce an approximation in which both exogeneity error and sampling error play a role and not as the actual DGP.

The 2SLS estimator can be written as

$$\hat{\beta}_N = (X'P_Z X)^{-1} X'P_Z Y.$$

Substitution of our model for Y yields

$$\hat{\beta}_N = (X'P_Z X)^{-1} X'P_Z X'\beta + (X'P_Z X)^{-1} X'P_Z Z\eta/\sqrt{N} + (X'P_Z X)^{-1} X'P_Z \varepsilon.$$

Then, rearranging and scaling by $\sqrt{N}$ yields

$$\sqrt{N}(\hat{\beta}_N - \beta) = [\sqrt{N}(X'P_Z X)^{-1} X'P_Z \varepsilon] + (X'P_Z X)^{-1} X'P_Z Z\eta.$$

Standard regulatory conditions imply that the term in brackets converges in distribution to

$$\left(M'_{ZX} M_{ZZ}^{-1} M_{ZX}\right)^{-1} M'_{ZX} M_{ZZ}^{-1} \nu,$$

which has the usual 2SLS limiting distribution, and the second term converges in distribution to $(M'_{ZX} M_{ZZ}^{-1} M_{ZX})^{-1} M'_{ZX} \eta$.

To use this approximation, it necessary to be able to consistently estimate the asymptotic variance of ν, V. To see that standard estimators of V remain consistent, note that $\widehat{\beta}_N$ is consistent under A1 and that $\widehat{\beta}_N - \beta = O_p(N^{-1/2})$. Therefore, we can form residuals $\widehat{\varepsilon} = Y - X\widehat{\beta}_N = \varepsilon + Z\frac{\eta}{\sqrt{N}} - X(\widehat{\beta}_N - \beta)$ where $\widehat{\beta}_N - \beta = O_p(N^{-1/2})$ and apply standard arguments to demonstrate consistency (see White, 2001).

Before concluding, we note that as with all other asymptotics, this local asymptotic sequence is a way to form an approximation for the behavior of a statistic and should not be viewed as a literal description of reality. By using this sequence, we obtain an approximation in which both sampling error and uncertainty about the exclusion restriction play a role. In practice, we believe the right way to use this approximation is to decide on what prior beliefs one has about γ and then plug them into expression (13).

A.3. MCMC Details

We outline our general strategy for full Bayesian inference for the model in equations (1) and (2). The only difference between our sampler and the sampler of Rossi, Allenby, and McCulloch (2005) is the inclusion of the term involving γ in the structural equation. For the empirical examples, the data were scaled by the standard deviation of Y, and then the priors used were $\Sigma \equiv \text{Cov}(\varepsilon, V) \sim$ Inverse Wishart(5,5I), and $\beta \sim N(0, 100)$. R code to implement these samplers is available on request from the authors.

Let Θ denote the parameters of the error term distribution—the joint distribution of (ε_i, ν_i). In the normal case, $\Theta = \Sigma$ is a covariance matrix. In an MCMC scheme, we alternate between drawing the regression coefficients (β, γ, Π) and the error term parameters. A basic Gibbs sampler structure is given by

$$\Theta|\beta, \gamma, \Pi, (X, Y, Z) \tag{GS.1}$$
$$\beta, \gamma, \Pi|\Theta, (X, Y, Z) \tag{GS.2}$$

In the normal case, equation (GS.1) may be done as in Rossi et al. (2005). The draw in (GS.2) is accomplished by a set of two draws:

$$\beta, \gamma|\Pi, \Theta, (X, Y, Z) \tag{GS.2a}$$
$$\Pi|\Theta, (X, Y, Z) \tag{GS.2b}$$

Given Θ, we can standardize appropriately (by subtracting the mean vector and premultiplying by the inverse of the Cholesky root of Σ). The draws in equation (GS2.a) are then done by realizing that, given Π, we "observe" ν_i and can compute the conditional distribution of ε_i. Given (β, γ), the draw of Π is done by a restricted regression model. Rossi et al. (2005) provide details of these draws.

For the prior on gamma in equation (A1), we cannot draw (β, γ) in one draw but must draw from the appropriate conditionals:

$$\gamma|\beta, \Pi, \Theta, (X, Y, Z) \tag{GS.2a.1}$$
$$\beta|\gamma, \Pi, \Theta, (X, Y, Z) \tag{GS.2a.2}$$