Plausibly Exogenous

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

14 March 2006. This draft: 31 July 2008.

Abstract

Instrumental variable (IV) methods are widely used to identify causal effects in models with endogenous explanatory variables. In many cases, the instrument exclusion restriction that underlies the validity of the usual IV inference is suspect; that is, the instruments are 'plausibly exogenous.' We develop practical methods of performing inference while relaxing the exclusion restriction. These methods provide tools for applied researchers who want to proceed with less-than-perfectly valid instruments. In addition, our framework enables a concise description of the tradeoff between instrument strength and the degree of exclusion restriction violation. We illustrate the approach with empirical examples that examine the effect of 401(k) participation upon asset accumulation, price elasticity of demand for margarine, and returns-to-schooling. We find that inference is quite informative even with a substantial relaxation of the exclusion restriction in two of the three cases.

Keywords: Instrumental Variables, Sensitivity Analysis, Priors, Treatment Effects

JEL Codes:C3,C11

*Graduate School of Business, the University of Chicago 5807 South Woodlawn Ave., Chicago, IL 60637. Email: tconley1@chicagogsb.edu, chansen1@chicagogsb.edu, peter.rossi@chicagogsb.edu. We thank seminar participants at the University of Chicago, Brown University, and Brigham Young University for helpful comments. We also appreciate the comments of 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 Univeristy of Chicago Graduate School of Business.

1 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 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 towards convincing the reader that their assumed exclusion restriction is a good approximation; i.e. they argue 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:

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

In this regression, Y is an outcome vector, X is a matrix of endogenous treatment variables, ε are unobservables, Z is a matrix of instruments that are assumed uncorrelated with ε and this orthogonality condition is the basis for estimation. 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 zero. Our definition of plausible exogeneity corresponds to having prior information that implies γ is near zero but perhaps not exactly zero. This assumption relaxes the IV exclusion assumption but still provides sufficient structure to allow estimation and inference to proceed.

[1]Stata code for the methods of Sections 4.1 and 4.3 is on Christian Hansen's website, currently http://faculty.chicagogsb.edu/christian.hansen/research/ . R code for Bayesian inference is available in the contributed package *bayesm*. Matlab code for other methods is available from the authors upon request.

We present three complementary estimation strategies that utilize prior information about γ to differing extents. The first approach only specifies the set of possible γ values, i.e. 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 strategy is to use prior information about the distribution of potential values of γ, while stopping short of a full specification of the error terms' distribution. 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, 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 practically very convenient.

Our third strategy is to undertake a full Bayesian analysis which requires priors over all model parameters (not just γ) and assumptions about the error distributions. Distributional assumptions can be very flexible via recent non-parametric Bayesian IV methods (e.g. Conley et al (2006)). 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 upon other model parameters are much easier to handle in this Bayesian framework versus our other methods.

Each proposed method focuses on interval estimates (inference) for β which provide a measure of what can be learned about β given the information in the data and beliefs about

γ. This focus differs from many other approaches that consider only bias and thus do not provide a complete picture of what one can learn about β. Our approach adds substantial value to simple sensitivity analyses performed by careful researchers that report results for a handful of alternate values for γ; see Angrist and Krueger (1994). For example, a researcher with a handful of different confidence intervals for β (each corresponding to a different value for γ) will be able to assess the variability of results across these γ values. However, such analysis does not immediately provide inference for β which is the ultimate goal of the empirical exercise.[2] We show how to take such information and combine it with varying degrees of prior information/beliefs about γ to construct confidence sets for β, enabling the researcher to conduct valid inference about the treatment effect β.

Our methods may greatly expand the set of available instruments and change the way researchers think about the scope of IV techniques. Instead of being restricted to using instruments for which the exclusion restriction is nearly certain, researchers may entertain the use of any instrument for which they are able to define beliefs about its direct effect γ. In many applications, instruments can yield informative results even under appreciable deviations from an exact exclusion restriction. We illustrate this through empirical examples.

One of the key features of our approaches is that they provide valid inference statements for any beliefs about the validity of the instruments. Thus, we can use our framework to examine the tradeoff between instrument strength and the degree of exclusion restriction violation (plausibility). 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.[3] In our framework, we may readily see how interval estimates for β depend on both instrument strength and plausibility. For example, relatively minor deviations from the exclusion re-

[2]For example, usual sensitivity analysis proceeds by considering results for a handful of γ values and then draws one of two conclusions. If the results are not "too different", one goes ahead with the analysis conducting inference as if γ were identically 0. Otherwise, one concludes that the results are too variable to be useful. The first conclusion is problematic as it ignores the researcher's uncertainty about the true value of γ. Thus, it is at odds with the beliefs that led to the sensitivity analysis and will produce intervals that tend to be too narrow. The latter conclusion will tend to be too pessimistic even with weak instruments and mildly informative prior beliefs about γ.

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

striction may greatly decrease precision relative to the case where $\gamma = 0$ when instruments are weak, whereas large deviations may have only small influences upon precision when the instruments are strong. This phenomenon is illustrated in an empirical example where we see the gains to using strong, but less plausible, versus weak, but more plausible, instruments can be substantial. Our methods allow a researcher to construct valid inference under any such scenario and to explicitly compare these intervals to choose between instrument sets that vary in strength and plausibility. The desire to use strong but less plausible instruments provides a direct motivation for the methods of this paper.

In our main presentation, we restrict ourselves to the linear IV model with constant coefficients and the 2SLS estimator for ease of exposition.[4] We discuss the model in more detail and show that within the context of linear models treating a constant coefficient model is essentially without loss of generality; see Section 3. In particular, we show that commonly employed models with heterogeneous treatment effects, including the local average treatment effect (LATE) model of Angrist and Imbens (1995) and Angrist, Imbens, and Rubin (1996), fall within our framework by re-interpreting β and γ in equation (1). Our methods also extend immediately to any structural model setting with unidentified parameters. Any situation in which there is an unidentified parameter about which reasonable *a priori* information exists can be treated using the approach to inference in this paper.

The remainder of this paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, we revisit model (1), show how it relates to relevant models with heterogeneous treatment effects, and discuss the importance of instrument strength. We present the inference methods formally in Section 4; and in Section 5, we consider how results vary with prior beliefs. In Section 6, we illustrate our methods in three example applications: estimating the effect of 401(k) participation upon 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 (2003); and estimating the returns to schooling as in Angrist and Krueger (1991). Section 7 concludes.

[4]The analysis could easily be extended to other estimators or to weak instrument robust inference procedures.

2 Other Approaches

Our methods complement other approaches in the literature. As discussed below, a typical treatment of suspect instruments focuses on biases or bounds for β. In contrast, we focus on procedures that provide confidence interval estimates of treatment effects incorporating information from both the data and researchers' beliefs regarding potential violations of the exclusion restriction.

A typical treatment of suspect instruments proceeds by performing a sensitivity analysis for the bias of an estimator of β. Assumptions are made about possible values for unidentified parameters, and one investigates how these values are related to the bias of the estimator.[5] Examples include Angrist and Krueger (1994) who characterize violations of the exclusion restriction exactly as in equation (1) and note that the difference between β and the probability limit of its two-stage least squares (2SLS) estimator is a simple function of γ. They use this function of γ and external information to assign a value for the bias of the 2SLS estimator. Angrist, Imbens, and Rubin (1996) also provide an expression for the large-sample limit of the 2SLS estimator in the local average treament effect (LATE) model when the exclusion restriction does not hold. They then construct a measure of bias that depends on the unidentified direct effect of Z on Y. Hahn and Hausman (2003) provide an expression for the bias of 2SLS that depends on the population R^2 of the infeasible regression of Z on $Y - X\beta$. They combine this with variance expressions to compare MSE for OLS and 2SLS estimators when the exclusion restriction is violated, in order to provide conditions when one prefers 2SLS to OLS point estimators (in MSE) when both are inconsistent.[6]

While the above approach to sensitivity analysis is clearly useful, it does suffer from some drawbacks. The focus on biases in point estimation inherently ignores estimation precision. Of course, it is essentially impossible to draw inferences from the data without considering

[5] See Rosenbaum (2002) for a textbook discussion.

[6] Another popular device is to posit parameter values govering the distribution of a hypothetical unobserved variable influencing outcomes. Rosenbaum (1987) conducts sensitivity analysis in the case of matched pairs sampling with binary treatments and outcomes in the presence of unobservables that may affect the treatment state. Gastwirth et al. (1997) and Imbens (2003) extend this by explicitly considering an unobservable that may affect both the treatment and the response.

precision. Therefore, the typical approach provides no guidance about what to do to obtain valid inference about treatment parameters. This point remains even if the focus were shifted from bias to interval estimates. As we discuss below, it is straighforward to obtain interval estimates for β given any hypothetical value of γ by estimating a new equation with $Y - Z\gamma$ as the outcome variable. Thus, we can extend the usual sensitivity analysis for bias to consider a set of confidence intervals for β. However, this does not address the question of how to combine the information the intervals contain (potentially with prior information as well) to construct an interval estimate of β.

One way to address the drawbacks mentioned above is to not focus on point estimation and instead use the information in the data along with prior restrictions to estimate bounds for β. In general, such restrictions will not point identify β even with an infinite amount of data but will instead identify a set of β's that are consistent with the observed data and beliefs. Even though one can not point identify β, one can obtain valid confidence sets for β. For such sets to be informative, one must impose some form of prior restrictions. Many approaches proceed by restricting the support of unidentified parameters.[7] An excellent example contemporaneous with our paper is Small (2007) who presents an approach to obtaining bounds for overidentified instrumental variables models. His procedure is based on inverting the Anderson and Rubin (1949, AR) statistic. It uses the information available from standard overidentification tests and augments this with an assumption about the support of unidentified parameters that characterize directions against which overidentification tests have low power. A drawback of this approach is that it does not apply under some forms of treatment effect heterogeneity; for example, it will not apply in the LATE model. In this case, different instrument sets estimate different treatment parameters; and overidentifying tests

[7]See Manski (2003) for an overview of such approaches. Example uses of bounds are Manski and Pepper (2000) and Hotz, Mullin, and Sanders (1997). Manski and Pepper (2000) consider treatment effect bounds with instruments that are assumed to monotonically impact conditional expectations, which is roughly analogous to assuming $\gamma \in [0, \infty]$. Hotz, Mullin, and Sanders (1997) model their dataset as a contamination mixture of subpopulations with an IV exclusion restriction holding in only one subpopulation. They make prior assumptions about the nature of the contamination (informed by auxiliary data) and utilize Horowitz and Manski (1995) bounds. Bounds in this context are clearly related to the recent econometrics literature on set identification as in, for example, Chernozhukov, Hong, and Tamer (2007).

may reject not because of violation of the exclusion restriction but simply due to treatment effect heterogeneity.

Our first approach in Section 4.1 also falls within the bounds framework. We impose support restrictions on γ and couple these with the information from the data to form confidence sets for β. Our approach is very easy to implement, being based entirely on conventional 2SLS point estimates and standard errors, and involves placing support restrictions over an easily interpretable parameter. In addition, it applies immediately in important heterogeneous effects models such as the LATE model. The chief drawback of confidence sets that only use support restrictions is that they make no use of a researcher's beliefs about the plausibility of different values of γ and thus are pessimistic relative to approaches that use these beliefs.

An analytically different way to proceed is to use approximations obtained in a model in which specification or exogeneity error and sampling error are of the same order of magnitude. Using this approximation, both specification error and sampling error will influence the limiting distribution of an estimator of β. This idea was employed in the context of estimation with invalid instruments in Hahn and Hausman (2003) who examine asymptotic MSE under local deviations of the exclusion restriction where $\gamma = C/\sqrt{N}$ for some constant C where N is the sample size. In research contemporaneous to ours, Berkowitz, Caner, and Fang (2006) use this approximation to show the AR statistic converges to a noncentral χ^2 random variable with noncentrality that depends on the unknown nuisance parameter C. They show that a delete-d jackknife can produce a large sample distribution that is close to the limiting distribution of the AR statistic regardless of the value of C. Their results are obtained in an iid setting, and it is unclear whether they are readily generalized to settings with heterogeneity or dependence, especially cases with treatment effect heterogeneity where β is interpreted as the LATE. Our local approach in Section 4.3 also fits into this basic framework. However, it differs substantially from other approaches in that we explicitly incorporate a researcher's beliefs about plausible values of the specification error, as captured by γ, by allowing this parameter to be a nondegenerate random variable in any finite sample. Using this formulation allows researchers to incorporate prior beliefs over exclusion restriction violations without requiring that those beliefs be dogmatic and thus provides a

better match to the way most people think about instrumental variables and their inference methods. Our method is extremely easy to implement, requiring only a simple modification of standard 2SLS regression output, and applies immediately in the LATE model and allows for heterogeneity and dependence more generally.

None of these alternative approaches makes use of prior beliefs beyond support restrictions. It appears that prior beliefs are held by researchers in many cases. In particular, when researchers claim their instruments are plausibly exogenous, this is analogous to stating that they believe there is a high probability that γ is near 0 and that the probability of more extreme values is diminishing. That is, they have in mind a shape for the distribution of likely values of γ. There are also clearly examples where external information about γ is available; for example, see Angrist and Krueger (1994). When such information is available, one might wish to use the posterior of γ or asymptotic distribution of an estimator of γ obtained from this auxiliary information as a prior. There are, of course, many other ways that one may choose to characterize beliefs depending on the particular application.

The main contributions of this paper are in carefully considering a simple framework in which to discuss the notion of plausibly exogenous instruments and providing simple to implement approaches to inference that allow one to incorporate prior information to varying degrees. All of our discussion is done in terms of equation (1) where we can characterize beliefs about the exclusion restriction in terms of the readily interpretable parameter γ. γ may be viewed as the direct effect of Z on the outcome which corresponds to a quantity, a partial effect, about which most applied researchers have strong intuition and has natural units which facilitates prior construction.

3 Model

We discuss each of the inference procedures mentioned above in the context of a linear structural model. In this section, we present the model and briefly discuss how it encompasses standard models with heterogeneous treatment effects. Extensions to nonlinear models are discussed in the appendix.

We are interested in estimating the parameter β in a simultaneous equation model rep-

resented 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 excluded 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 obviously to allow for additional predetermined or exogenous regressors.[8] 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 encompasses the usual models that allow for 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 (2) is that β should be interpreted as the average treatment effect of X on Y and γ should be interpreted as the average effect of Z on Y. A similar though more complicated set of arguments also applies immediately in the LATE model of Angrist and Imbens (1995) and Angrist, Imbens, and Rubin (1996). In this case, β should be interpreted as the LATE and γ as the average direct effect of Z on Y, but nothing is affected outside of interpretation.[9]

Finally, it is worth noting that the strength of the relationship between Z and X, captured by Π in (3), plays an important role in determining what can be learned about β in (2) just

[8] See the appendix.

[9] We demonstrate this in the appendix. 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.

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 senstitive 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 (1996) and others that there is typically a tradeoff between instrument strength and plausibility. All of the inference approaches we present in the following section provide ways for one to think formally about this tradeoff.

4 Inference Procedures

In this section, we consider four methods for inference about β. Each provides a way to perform inference about β without assuming γ is exactly zero. 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 $\gamma-$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 (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' $\gamma-$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 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.

4.1 Union of Confidence Intervals with γ Support Assumption

Our first inference method utilizes only a support assumption about γ. Specifically, suppose prior information consists of knowledge of the support for γ, $\mathcal{G}$, which is bounded.[10] 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 $\gamma-$specific confidence regions for β will have coverage of at least $(1-\alpha)$. Our approach is simply to approximate this union of confidence intervals.

For ease of exposition, we present details for the two-stage least squares (2SLS) estimator of β.[11] Under the maintained assumption that $\gamma = \gamma_0$

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

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

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

from which it will follow under conventional regularity conditions[12] 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.

For simplicity, we suppose that s, the dimension of β, equals 1 in the following; the

[10] Of course, if $\mathcal{G}$ is unbounded and one is unwilling to place further restrictions on γ, confidence regions for β will also be unbounded.

[11] The same approach could be applied to any conventional estimator of β.

[12] See the appendix for an example set of regularity conditions.

discussion generalizes immediately to $s > 1$ at the cost of complicating the notation.[13] Using (4), we could estimate a symmetric (1-α) 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 (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. We give details of how to construct this interval in the appendix and illustrate its use in the empirical examples. We do not focus on it here as it is notationally cumbersome, and we expect there will generally be only modest gains to its use in practice. The 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

[13] 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. In the LATE model, it is less clear that any useful information may be extracted because different valid instruments will generally estimate different local average treatment effects.

requiring correct coverage in every possible case, including the worst. 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.

4.2 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 allowing oneself to "weight" the intervals for different values of γ differently depending on prior beliefs about how likely different values of γ are. 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 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 *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 4.1.

We note that formally setting up this problem is notationally complicated and so relegate the formal discussion to the appendix. Below, we illustrate the potential gains and discuss the problem heuristically in the context of an extremely simple stylized example. We also consider this approach in the empirical results presented below. While this approach offers the potential for substantial gains relative to the support-restriction-only interval when one is willing to place additional structure on the information about γ, numerical solution of the problem presents a greater computational challenge. Therefore, after presenting the approach heuristically below, we move on to explore a different approximate approach to using prior information about the distribution of γ which is extremely simple to implement in practice.

For our example, we consider a case where γ may take on only one of two values: γ_1 or γ_2. With $\gamma = \gamma_1$, the estimator of β takes on a value of one and has a standard error of one; and when $\gamma = \gamma_2$, the estimator of β is four with a standard error of two. We

present interval estimates using the bounds approach from the previous section and the prior-weighted approach of this section in Table 1. The columns labeled "Support Restriction" present intervals constructed using the approach of the previous section, and the columns labeled "Fully Specified Prior" present intervals which make use of prior beliefs over the probability of each potential value of γ.

Looking first at 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 endpoints and maximum of the upper endpoints. To get the length-minimizing union of intervals imposing only the support retriction, we then note that we can increase the lower endpoint of the γ_1-interval (-0.645,2.645) while simultaneously increasing its upper endpoint. Retaining 90% coverage requires that the increase in the upper endpoint of the γ_1-interval be larger than the increase in the lower endpoint 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 endpoint of the γ_1-interval remains smaller than the upper endpoint of the γ_2-interval. A similar argument holds for decreasing the lower and upper endpoints of the γ_2-interval. The length minimizing interval occurs where the two lower endpoints and the two upper endpoints 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 that there are 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.

4.3 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.'[14] This treatment produces the following approximation to the distribution of $\widehat{\beta}$:

$$\begin{aligned} &\widehat{\beta} \overset{approx}{\sim} N(\beta, V_{2SLS}) + A\gamma, \\ &A = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z), \\ &\gamma \sim F. \end{aligned} \tag{7}$$

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.[15] 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.

[14] 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.

[15] 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. We provide a simple illustration in the appendix.

This approximation is easy to use. As is evident in equation (7), 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: $\eta = \beta - \widehat{\beta}$ where

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

Draws from the η distribution can be constructed as follows:

(1) Use any standard software package to compute A as a function of sample moments and the 2SLS covariance matrix V_{2SLS}.

(2) Generate one draw, η_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 η draws: $\eta_1, \eta_2, ..., \eta_B$.

(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-α) confidence interval for β as $[\widehat{\beta} - c_{1-\alpha/2}, \widehat{\beta} - c_{\alpha/2}]$.

One nice aspect of the approximate distribution in (7) is that 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 $Z'X$.

We anticipate that in many applications this approach will return confidence intervals that are close to Bayesian posterior intervals under the same prior for γ, which occured 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 similarly 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.

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 zero 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 5 below. 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.

4.4 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 (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 :

$$p(\beta, \Pi, \theta|Data) \propto \int p(Data|\beta, \gamma, \Pi, \theta) p_\gamma(\gamma|\beta, \Pi, \theta) p_{\{\beta,\Pi,\theta\}}(\beta, \Pi, \theta) d\gamma \tag{8}$$

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 one 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 upon β :

$$\text{Prior 1:} \qquad \gamma \sim N(\mu, \delta^2 I) \tag{9}$$

$$\text{Prior 2:} \qquad \gamma|\beta \sim N(0, \delta^2\beta^2 I). \tag{10}$$

With prior 1, we will need to have 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 non-standard 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, but note that we could of course employ Bayesian methods with any parametric likelihood or a nonparametric approach via flexibly approximating an arbitrary likelihood as in, for example, Conley et al. (2006). Given the likelihood and the priors, the chief difficulty in conducting Bayesian inference is in evaluating the posterior distribution which will typically be done using MCMC methods. Given the Gaussian likelihood and

priors, computation is quite similar to standard approaches in the Bayesian literature; so we leave computational details to the appendix.

It is important to note that full Bayesian analysis allows exact small sample Bayesian inference as well as considerable flexibility in prior specification. There are certainly many applications where small sample considerations are paramount, e.g. due to weak or large numbers of instruments. If relatively diffuse priors are used on β and π and a sufficiently flexible distribution is specified for (ε, V), the requirements of a full Bayesian analysis are only modestly higher than the methods above using large sample approximations. All approaches hinge on a careful assessment of the prior on γ.

5 Illustrating Inference with Alternate Priors and Instruments

In the preceding section, we outlined basic approaches to inference that may be adopted in situations in which there is uncertainty about an IV exclusion restriction. In each, we model the specification error through an unidentified parameter γ and do inference utilizing assumptions about possible values for this parameter. The main benefit from our procedures is in providing researchers with a practical way to learn about treatment effects from the data with less-than-perfect instruments. Because of the importance of prior beliefs and because researchers will likely differ on their exact prior beliefs, we anticipate that it will be useful to apply our procedures under more than one assumption regarding γ and compare the resulting inference for β. This analysis gives the researcher substantial insight into ranges of beliefs that produce economically similar inference. In addition, when there are more instruments than endogenous regressors, a researcher may perform the exercise with different sets of instruments of differing strengths and plausibility to gain information on the tradeoff between instrument strength and plausibility.

The exact approach one takes to considering different priors will vary slightly under each of the different methods. For example, in the support-restriction-only approach in Section 4.1 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 4.2 through 4.4 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 zero and variance δ^2. Again one could compute confidence intervals (or credibility intervals in the case of Bayesian inference) as a function of δ. We provide illustrative examples of analogous exercises for each of our empirical examples in Figures 1-5 in Section 6.

The above methods are also well-suited to examining the tradeoff between strength and plausibility of instruments. In cases where the researcher has two sets of instruments and wants to choose one, a graph displaying a range of confidence sets for varying γ priors is straightforward to construct for each set of instruments. An example set of results for two instrument set/prior combinations is available in Figure 3, which is based on data from our first empirical example in Section 6.1. The solid lines in Figure 3, labeled strong instrument, trace out the upper and lower endpoints for a 95% confidence interval for β as a function of candidate γ prior distributions indexed by δ. These prior distributions are each Uniform [0,δ]. Likewise, the dashed lines, labeled moderate instrument, trace out upper and lower edges of confidence intervals with a relatively weaker set of instruments. Suppose a researcher believes that the moderate instrument is certain to satisfy the exclusion restriction $\delta = 0$, but her γ prior for the strong instrument is Uniform [0,1000] so $\delta = 1000$. The length of the 95% confidence interval for β corresponding to each of these sets of instruments and priors is given by the vertical distance between the solid lines at $\delta = 1000$ and the gap between the dashed lines at $\delta = 0$. Clearly in this example, more precise inference about β is available using the stronger set of instruments with a larger believed potential departure from the exclusion restriction.

6 Empirical Examples

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 as well as 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 zero. 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.

6.1 Effect of 401(k) Participation upon Asset Accumulation

Our first example application examines the effect of 401(k) plan participation upon asset accumulation. Our data are those of Poterba, Venti, and Wise (1995), Abadie (2003), Benjamin (2003), and Chernozhukov and Hansen (2004) from wave 4 of the 1990 Survey of Income and Program Participation which consists of observations on 9915 households with heads aged 25-64 with at least one adult employed (excluding self-employed). 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 is a vector of covariates that includes 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 firm with a 401(k) plan. Further details and descriptive statistics can be found in, for example, Benjamin (2003) or Chernozhukov and Hansen (2004).

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. Their main claim is the fact that eligibility is determined by employers and so is plausibly taken as exogenous conditional on covariates. In particular, if individuals made employment decisions based on income and within jobs classified by income categories it is random whether or not a firm offers a 401(k) plan, the exclusion restriction would be justified. Poterba, Venti, and Wise (1996) contains an overview of suggestive evidence based on pre-program savings used to substantiate this claim. Of course, the argument for the exogeneity of 401(k) eligibility is hardly watertight. For example, one might conjecture that firms which 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. People might also select employment on the basis of available retirement packages in addition to income. For these and other reasons, Engen, Gale, and Sholz (1996) argue that 401(k) eligibility is not a valid instrument and is positively related to the unobservables affecting financial asset accumulation.

First, we display 2SLS estimates for various values of γ in Panel A of Table 2. We chose the specific values of γ by deciding that we felt a \$10,000 direct effect of eligibility on financial assets would be quite large and then choosing what we felt were sensible intermediate values. Looking at the results, we see that the point estimate of β varies quite a lot as we change the value of γ and that the estimated standard errors are fairly stable across this range of γ values. These results are suggestive about the sensitivity of the 2SLS estimator of β to different values of γ but do not, by themselves, allow us to make inferences about β. Thus, we turn to the methods discussed in this paper.

Figure 1 displays results for the full array of our methods with γ priors centered at zero. 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 4.1 with a union of symmetric γ_0-specific intervals. Their corresponding support restrictions are of the form $\gamma \in [-2\delta, +2\delta]$. The dashed lines that lie just inside of them are the minimum-length bound from Section 4.1 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 dashed lines correspond to the union of prior-weighted intervals approach from Section 4.2 with γ prior of $N(0, \delta^2)$, the solid lines correspond to the local-to-zero method of Section 4.3 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 4.4 again obtained with a γ prior of $N(0, \delta^2)$.

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. It is important to note that 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 will omit plotting intervals for the prior-weighted union of confidence intervals as they are very close to the local-to-zero intervals in our applications.

While priors centered at zero may be adequate for many researchers, they would not be appropriate for those who agree with Engen, Gale, and Sholz (1996) and think there is likely a positive direct effect of 401(k) eligibility upon 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 4.1 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-zero 95% confidence interval estimates from Section 4.3 using the prior that γ is uniformly distributed on $[0, +\delta]$. The dot-dash lines present Bayesian 95% credibility intervals from Section 4.4 with a prior for γ that is normally distributed with the same mean and variance used for the local-to-zero estimates (a mean of $\frac{1}{2}\delta$, variance of $\frac{1}{12}\delta^2$). Since priors with positive means result in intervals shifting

location with δ we also plot, as point estimates, a solid line corresponding to the center point of our local-to-zero 95% confidence intervals. 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 upon net assets, even with substantial departures from perfect instruments. For example, take the widest intervals with the support restriction of $\gamma \in [0, 4000]$, clearly distinct from $\gamma \equiv 0$. The corresponding confidence set for β is approximately [\$3700, \$17,000]. While certainly different from the [\$9500, \$17,000] interval under perfect ($\gamma \equiv 0$) instruments, many would still consider the [\$3700, \$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 impacts 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-zero 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 only take on values of zero and one.[16] We consider "strong", "moderate", and "weak" instruments which respectively correspond to adding no noise, a moderate amount of noise, and a large amount of noise to the original instrument.

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. More interesting are the tradeoffs one can make when trying to choose between stronger and weaker instru-

[16]Specifically, 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 different 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 used in the econometrics literature regarding weak instruments, none of the setting correspond to weak instruments. For example, the first stage F-statistics for the strong, moderate, and weak settings are respectively 7767.9, 1094.9, and 351.8. We use the terminology to simply denote the relative strength of the instruments, but note that we could adapt our approach to a weak instrument robust procedure.

ments. 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 [0,10000]. Before entertaining using only the weak instrument, a researcher would need to believe that the direct effect of 401(k) eligibility could be as large as \$6000-\$7000, which seems quite large, and that the weak instrument satisfies the exclusion restriction almost perfectly. 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 trade-offs between the strength and plausibility of instruments and certainly illustrates that there are scenarios where it will be preferable, in terms of learning about β, to use a strong but less credible instrument rather than a weak but more credible one. Of course, this will depend on the relative strengths of the instruments and how firmly one believes the exclusion restriction is satisfied for each of them in any particular example. An appealing feature of the approaches considered in this paper is that they allow a researcher to assess these tradeoffs in any given application. Moreover, the approaches we develop are indispensible if one wishes to use strong instruments for which the exclusion restriction is believed unlikely to hold.

6.2 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, Dubé, and Goh (2003) . The sample consists of weekly purchases and prices for the 4 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, Dube, and Goh (2003) 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 direct effect of a wholesale price could be benchmarked relative to the elasticity with respect to the corresponding retail price.

We display 2SLS estimates for various values of γ in Panel B of Table 2. In this case, we decided that it seemed unlikely that the direct effect of wholesale prices could be more than 30% of the direct effect of retail prices and then chose a set of values consistent with these beliefs. Looking at the results, we see that the point estimate of β varies between -5 and -3 as we change the value of γ and that the estimated standard errors are quite variable. These results suggest that the estimated price elasticity is fairly sensitive to the value of γ, though in all cases we find that demand is quite elastic. As before, these results are suggestive but do not allow us to make inferences about β, so we turn to the methods discussed in this paper.

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 4.1 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-zero 95% confidence interval estimates from Section 4.3 using a prior that of γ given β is distributed $N(0, \delta^2\beta^2)$.[17] The dot-dash lines present Bayesian 95% credibility intervals from Section 4.4 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 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-restiction-only intervals. An additional source of discrepancy that is likely more important here than in the

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

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-zero 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 ten per cent of the size of the retail price effect. Take for example the local-to-zero 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 mark-up 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.

6.3 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.[18]

The use of quarter of birth as an instrument is motivated by the fact that quarter of birth is correlated with years of schooling due to compulsory schooling laws. The typical law requires students to attend first grade in the year in which they turn age 6 and continue school until age 16. This means that individuals born early in the year will usually be in the

[18]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.

middle of 10th grade when they turn 16 and can drop out while those born late in the year will have finished 10th grade before they reach age 16.

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 which influence earnings. Angrist and Krueger (1991) examine data from three decennial censuses and find that people born in the first quarter of the year do indeed 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 (e.g. Bound and Jaeger (1996) and Bound, Jaeger, and Baker (1995)). Bound, Jaeger, and Baker (1995) go beyond this in providing well motivated 'back of the envelope' calculations of a plausible range for direct effects of quarter of birth upon wages. They come up with an approximate magnitude of a direct effect of quarter of birth upon wages of about 1%. Such calculations are directly useful in our framework, informing our choice of prior for γ.

2SLS estimates for various values of γ are given in Panel C of Table 2. In this case, it is harder to determine a small set of sensible values due to the larger number of instruments. We used the benchmark of a 1% effect of quarter of birth and then chose a few different values for each of three coefficients where the effect of any one quarter relative to any other was no greater than 1%. Obviously, there are a large number of ways to do this, which suggests another advantage of our systematic way of viewing the problem relative to choosing a few candidate values for γ. Looking at the results, we see that the point estimate of β are again quite sensitive to the value of γ and that the estimated standard errors are fairly stable for this set of γ values. These results are again suggestive about the sensitivity of the 2SLS estimator of β to different values of γ but do not provide inference for β. Thus, we discuss results from using the methods of this paper below.

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 4.1 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-zero

method in Section 4.3 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 4.4 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, Jaeger, and Baker 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, Jaeger, and Baker (1995). Using the Bound, Jaeger, and Baker (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-zero 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 confidence intervals to be informative in our judgment, prior beliefs regarding γ must be much more concentrated near zero. For example, using the support-restriction-only intervals, one would need to be sure the magnitude of γ was less than .002 to obtain a confidence interval for β that excluded 5%.

7 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 toolset for the applied researcher 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.

Our methods also allow researchers to directly confront the issue of using stronger but less plausible instruments versus weaker but more plausible instruments. It is well-known that weaker instruments produce less precise inference than stronger instruments. Our results also allow a researcher to assess how strength of beliefs in the validity of the exclusion restriction affect the precision of inference about treatment effects. Thus, they allow researchers to assess whether the decrease in precision associated with more diffuse beliefs about exclusion restriction violations is offset by increases in precision due to using stronger instruments.

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 obtain useful 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 plausible instruments that allow for economically informative inference after accounting for their possible imperfection.

8 Appendix

8.1 Including Additional Regressors

For ease of exposition we have stated our model without any 'included' exogenous regressors. It is straightfoward to allow such additional regressors $\tilde{W}$ into the model :

$$\tilde{Y} = \tilde{W}B_1 + \tilde{X}\beta + \tilde{Z}\gamma + \varepsilon \tag{11}$$

$$\tilde{X} = \tilde{W}B_2 + \tilde{Z}\Pi + V \tag{12}$$

This model reduces to model (1) and (2) by defining Y, X, and Z as residuals from a projection upon the space spanned by $\tilde{W}$, *i.e.* as

$$Y = (I - P_{\tilde{W}})\tilde{Y}, \quad X = (I - P_{\tilde{W}})\tilde{X}, \quad Z = (I - P_{\tilde{W}})\tilde{Z}.$$

8.2 LATE Model

We consider a LATE model with binary treatment and instrument though extensions to multivalued treatments and instruments should follow similarly as in Angrist and Imbens (1994, 1995). We also switch to the potential outcomes framework and let $Y_i(x, z)$ be the outcome for individual i when treatment is set to x and the instrument is set to z. Similarly, let $X_i(z)$ be the treatment for individual i when the instrument is set to z. In the data we observe z_i, $x_i = X_i(1)z_i + X_i(0)(1 - z(i))$, and $y_i = Y_i(1,1)x_i z_i + Y_i(1,0)x_i(1 - z_i) + Y_i(0,1)(1 - x_i)z_i + Y_i(0,0)(1 - x_i)(1 - z_i)$. The LATE model typically makes the following assumptions:

A1. (Random Assignment) $Z_i \perp (Y_i(0,0), Y_i(1,0), Y_i(0,1), Y_i(1,1), X_i(0), X_i(1))$ and data are independent across individuals.

A2. (Exclusion) $Y_i(x, z) = Y_i(x, z')$ for all z, z', and x.

A3. (Monotonicity) $X_i(1) \geq X_i(0)$.

A4. (Instrument Relevance) $E[X_i(1)] \neq E[X_i(0)]$.

Under these assumptions, Angrist and Imbens (1994, 1995) and Angrist, Imbens, and Rubin (1996) show that the IV estimator estimates the LATE defined as $\beta_{LATE} = E[Y_i(1) - Y_i(0)|X_i(1) > X_i(0)]$ where $Y_i(1) = Y_i(1,1) = Y_i(1,0)$ under A2 and $Y_i(0)$ is defined similarly. In the following, we show that model (2) follows from a LATE model with A2 replaced with

A2'. (No Interaction) $Y_i(1,0) - Y_i(0,0) = Y_i(1,1) - Y_i(0,1)$.

A2' replaces the exclusion restriction with a condition that implies that there are no interactions between X and Z but allows for Z to have a direct effect on Y.

We start by noting that under A2', $y_i = Y_i(0,0) + [Y_i(1,0) - Y_i(0,0)]x_i + [Y_i(0,1) - Y_i(0,0)]z_i$ from which it follows that $y_i - \bar{y} = \beta_{LATE}(x_i - \bar{x}) + \gamma_{LATE}(z_i - \bar{z}) + \varepsilon_i - \bar{\varepsilon}$ where $\varepsilon_i = Y_i(0,0) + (\beta_i - \beta_{LATE})x_i + (\gamma_i - \gamma_{LATE})z_i$, $\beta_i = Y_i(1,0) - Y_i(0,0)$, $\gamma_i = Y_i(0,1) - Y_i(0,0)$, and $\bar{w} = \frac{1}{N}\sum_i w_i$ for any variable w_i. That this model is the same as (2) for β_{LATE} defined above and $\gamma_{LATE} = E[\gamma_i]$ follows by showing that $E[(z_i - \bar{z})(\varepsilon_i - \bar{\varepsilon})] = 0$.

$\varepsilon_i - \bar{\varepsilon}$ consists of three terms: $Y_i(0,0) - \bar{Y}(0,0)$, $(\beta_i - \beta_{LATE})x_i - \frac{1}{N}\sum_i(\beta_i - \beta_{LATE})x_i$, and $(\gamma_i - \gamma_{LATE})z_i - \frac{1}{N}\sum_i(\gamma_i - \gamma_{LATE})z_i$. For the first term, we have that $E[(z_i - \bar{z})(Y_i(0,0) - \bar{Y}(0,0))] = 0$ from A1. For the second term, we have $E[(z_i - \bar{z})((\beta_i - \beta_{LATE})x_i - \frac{1}{N}\sum_i(\beta_i - \beta_{LATE})x_i)] = E[(z_i - \bar{z})(\beta_i - \beta_{LATE})x_i] - E[(z_i - \bar{z})\frac{1}{N}\sum_i(\beta_i - \beta_{LATE})x_i]$. The first component can be written as $E[(z_i - \bar{z})\beta_i X_i(0)] - E[(z_i - \bar{z})\beta_{LATE}X_i(0)] + E[(z_i - \bar{z})\beta_i(X_i(1) - X_i(0))z_i] - E[(z_i - \bar{z})\beta_{LATE}(X_i(1) - X_i(0))z_i]$. The first two terms in this expression are 0 under A1. We also have that $E[(z_i - \bar{z})\beta_i(X_i(1) - X_i(0))z_i] = E[z_i(z_i - \bar{z})]E[\beta_i(X_i(1) -$

$X_i(0))] = E[z_i(z_i - \bar{z})]E[\beta_i|X_i(1) > X_i(0)]Pr(X_i(1) > X_i(0)) = E[z_i(z_i - \bar{z})]\beta_{LATE}Pr(X_i(1) > X_i(0))$ where the first equality is by A1, the second is by A3, and the last by the definition of β_{LATE} and that $E[(z_i - \bar{z})\beta_{LATE}(X_i(1) - X_i(0))z_i] = E[z_i(z_i - \bar{z})]E[\beta_{LATE}(X_i(1) - X_i(0))] = E[z_i(z_i - \bar{z})]\beta_{LATE}Pr(X_i(1) > X_i(0))$ where the first equality is by A1 and the second is by A3. Thus, $E[(z_i - \bar{z})\beta_i(X_i(1) - X_i(0))z_i] - E[(z_i - \bar{z})\beta_{LATE}(X_i(1) - X_i(0))z_i] = 0$. It follows immediately that $E[(z_i - \bar{z})(\beta_i - \beta_{LATE})x_i] = 0$, and $E[(z_i - \bar{z})\frac{1}{N}\sum_i(\beta_i - \beta_{LATE})x_i] = 0$ follows similarly. We can now choose γ_{LATE} to make the covariance between $z_i - \bar{z}$ and the final term $(\gamma_i - \gamma_{LATE})z_i - \frac{1}{N}\sum_i(\gamma_i - \gamma_{LATE})z_i$ zero. In particular, we have $E[(z_i - \bar{z})((\gamma_i - \gamma_{LATE})z_i - \frac{1}{N}\sum_i(\gamma_i - \gamma_{LATE})z_i)] = E[(z_i - \bar{z})(\gamma_i z_i - \frac{1}{N}\sum_i \gamma_i z_i)] - \gamma_{LATE}E[(z_i - \bar{z})^2]$ which we can set equal to zero and solve for γ_{LATE} to obtain $\gamma_{LATE} = E[(z_i - \bar{z})(\gamma_i z_i - \frac{1}{N}\sum_i \gamma_i z_i)]/E([(z_i - \bar{z})^2]$. Then note that $E([(z_i - \bar{z})^2] = (1 - 1/N)\sigma_z^2$ and that $E[(z_i - \bar{z})(\gamma_i z_i - \frac{1}{N}\sum_i \gamma_i z_i)] = (1 - 1/N)\sigma_z^2 E[\gamma_i]$ under A1 to conlude that $\gamma_{LATE} = E[\gamma_i]$. The conclusion then follows.

8.3 Regularity Conditions for 2SLS

We will make use of the following standard high-level assumption given below to derive the asymptotic properties of the inference procedure in Sections 4.1, 4.2 and 4.3. This assumpton imposes a standard set of regularity conditions which are implied by a variety of more primative stochastic assumptions; see e.g. White (2001). Existence of the limits in the assumption could be relaxed at the cost of more complicated notation. It should be noted that these assumptions correspond to those used for the typical asymptotic approximations for 2SLS which are known to provide poor approximations when the correlation between the instruments and endogenous regressors is weak, i.e. when condition (ii) is approximately violated, or when the degree of overidentification, $r - s$, is large. Extension of the basic approach to settings with weak or many instruments should be straightforward.

Assumption A1

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) $Z'X/N \xrightarrow{p} M_{ZZ}$, 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= $\lim E\{Z'\varepsilon\varepsilon'Z/N\}$

8.4 Length Minimizing and Prior-Weighted Intervals

In the following, we show how length minimizing intervals may be constructed. We start by considering the case where we impose the support restriction that $\gamma \in \mathcal{G}$ but do not make use of a fully specified prior. In

this case, the only gains relative to the simple union of symmetric intervals discussed in Section 4.1 is due to potentially allowing for asymmetric intervals. We then present the formal problem that is solved to obtain the prior-weighted union of intervals of Section 4.2. We note that the difference between these two problems is that the latter removes a constraint from the former and so will produce weakly shorter intervals.

For each γ_0 we can define a potentially asymmetric confidence interval for β using an additional parameter $a(\gamma_0) \in [0, \alpha]$ 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

$$\begin{aligned} & CI_N(1-\alpha, \gamma_0, a(\gamma_0)) \\ = \quad & \left[\widehat{\beta}_N(\gamma_0) + c_{\alpha - a(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}, \quad \widehat{\beta}_N(\gamma_0) + c_{1-a(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}\right]. \end{aligned} \tag{13}$$

Again under conventional regularity conditions, it follows that $Pr\{\beta \in CI_N(1-\alpha, \gamma_0, a(\gamma_0))\} \to 1-\alpha$ 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:

$$CI_N(1-\alpha, a(\cdot)) = \cup_{\gamma_0 \in \mathcal{G}} CI_N(1-\alpha, \gamma_0, a(\gamma_0)) \tag{14}$$

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

$$\min_{a(\cdot)} \int_{-\infty}^{\infty} 1\{b \in CI_N(1-\alpha, a(\cdot))\} db \quad \text{s.t.} \quad a(\gamma_0) \in [0, \alpha] \text{ for all } \gamma_0 \in \mathcal{G} \tag{15}$$

where $1\{\cdot\}$ is the indicator function which is one when the event in the braces is true.

In practice, we anticipate that often there will be only modest gains from calculating the shortest interval via solving the minimization problem in (15) compared with the easy-to-compute union of symmetric intervals (6). In situations where the variation in $\widehat{\beta}_N(\gamma_0)$ across γ_0 values is large relative to the estimated standard errors, even if all of the weight at the extreme values of $\widehat{\beta}_N(\gamma_0)$ is concentrated in one tail the changes to the overall interval length will be small. In addition, when the standard errors are much larger than the range of $\widehat{\beta}_N(\gamma_0)$ estimates there will also be little scope for moving away from equal tailed intervals.

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

$$\begin{aligned} & CI_N(1-a(\gamma_0), \gamma_0, a(\gamma_0)) \\ = \quad & \left[\widehat{\beta}_N(\gamma_0) + c_{a(\gamma_0)-a(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}, \quad \widehat{\beta}_N(\gamma_0) + c_{1-a(\gamma_0)}\sqrt{\widehat{V}_N(\gamma_0)/N}\right]. \end{aligned} \tag{16}$$

Without any information about the distribution of potential values of γ beyond a support condition, the only way to insure correct *ex ante* coverage of (1-α) 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 in the present context 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

$$\begin{aligned} CI_{F,N}(1-\alpha, a(\cdot), a(\cdot)) &= \cup_{\gamma_0 \in \mathcal{G}} CI_N(1-a(\gamma_0), \gamma_0, a(\gamma_0)) \\ &\text{subject to} \\ a(\gamma_0) &\in [0, \alpha(\gamma_0)] \text{ and } \int_{\mathcal{G}} \alpha(\gamma_0) dF(\gamma_0) = \alpha \end{aligned} \tag{17}$$

where $CI_N(1-a(\gamma_0), \gamma_0, a(\gamma_0))$ is a $(1-a(\gamma_0))$(possibly asymmetric) confidence interval given $\gamma = \gamma_0$ defined by (16). The constraint $\int_{\mathcal{G}} \alpha(\gamma_0) dF(\gamma_0) = \alpha$ ensures *ex ante* coverage of (1-α), under the prior distribution F. Under regularity conditions in the next section, $CI_{F,N}(1-\alpha, a(\cdot), a(\cdot))$ has coverage at least $(1-\alpha)$ as $N \to \infty$.

The choices of $a(\cdot)$ and $a(\cdot)$ that minimize the size of the interval solve the following problem:

$$\min_{a(\cdot), a(\cdot)} \int_{-\infty}^{\infty} 1\{b \in CI_{F,N}(1-\alpha, a(\cdot), a(\cdot))\} db. \tag{18}$$

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 (18) will always be weakly smaller than the confidence region obtained without using the distributional information about γ.

8.5 Regularity Conditions for Convergence of Union of Prior-Weighted Confidence Intervals

Assumption A1 and continuity of $a(\cdot), a(\cdot)$ are sufficient for $CI_{F,N}(1-\alpha, a(\cdot), a(\cdot))$ defined by (17) to have proper limiting coverage. To see this note,

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

where G_n does not depend on γ since

$$\widehat{\beta}(\gamma_0) - \beta \mid \gamma = \gamma_0 \text{ is } (X'P_Z X)^{-1} X' P_z \varepsilon$$

$$\text{and } \widehat{V}(\gamma_0) = h(X, Z, e(\gamma_0))$$

$$\text{where } e(\gamma_0) \mid \gamma = \gamma_0 \text{ is } Y - Z\gamma_0 - X\widehat{\beta}(\gamma_0) = \varepsilon - X(X'P_zX)^{-1}X'P_z\varepsilon.$$

Also, under standard regularity conditions,

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

Now

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

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 $|G_N(h(\gamma))dF| \leq dF$ 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.

8.6 Behavior of 2SLS Estimator under γ Local to Zero Approximation

To obtain the approximation in Section 4.3, we model γ as being local to zero.[19] Explicitly referencing the dependence of γ upon the sample size via a subscript N we represent γ in structural equation (2) as

$$\gamma_N = \eta/\sqrt{N} \quad \text{where} \quad \eta \sim G. \tag{19}$$

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 are of the same order and so both factor into the

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

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. To the extent that both sources of error do play a role, the approximation will tend to be more accurate than the approximation obtained when either source of error dominates.

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

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

Assumption A1 implies that the term in brackets converges in distribution to $\left(M'_{ZX} M_{ZZ}^{-1} M_{ZX}\right)^{-1} M'_{ZX} M_{ZZ}^{-1} v$, which has the usual 2SLS limiting distribution, and the second term converges in distribution to $\left(M'_{ZX} M_{ZZ}^{-1} M_{ZX}\right)^{-1} M'_{ZX}\eta$.

To use this approximation, it necessary to be able to consistently estimate the asymptotic variance of v, 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, e.g. White (2001). As a simple example, consider the case where $V = \sigma^2 M_{ZZ}$. In this case, M_{ZZ} can be estimated as usual by $Z'Z/N$. For our estimate of σ^2, we use

$$\begin{aligned}
\widehat{\sigma}^2 &= \widehat{\varepsilon}'\widehat{\varepsilon}/N \\
&= \varepsilon'\varepsilon/N + 2(\varepsilon'Z/N)\frac{\eta}{\sqrt{N}} - 2(\varepsilon'X/N)(\widehat{\beta}_N - \beta) \\
&\quad - 2\frac{\eta}{\sqrt{N}}(Z'X/N)(\widehat{\beta}_N - \beta) + (\eta'Z'Z\eta/N^2) \\
&\qquad + (\widehat{\beta}_N - \beta)'(X'X/N)(\widehat{\beta}_N - \beta) \\
&\xrightarrow{p} \sigma^2.
\end{aligned}$$

We could similarly show consistency for any standard robust covariance matrix estimator that allows for estimation of parameters by simply modifying the corresponding proof to account for γ_N as above.

Before concluding, we note that, as with all 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 (7).

8.7 Nonlinear Models

For ease of exposition, we considered each of the inference methods in the main text in the context of the linear IV model. Linear IV models are a leading case in which to apply our methods, but our basic results apply in any context in which there are unidentified parameters and in which one may reasonably claim to have some sort of prior information about their plausible values. The Bayesian approach discussed in Section 4.4 immediately extends to nonlinear models, here we briefly discuss the extenstion to nonlinear models for our other three approaches.

For concreteness, suppose that we are interested in performing inference about a parameter θ defined by as the optimizer of an objective function

$$h_N(W;\theta,\gamma)$$

where W are the observed data, γ is an additional parameter, and θ and γ are not jointly identified from the objective function but for a given value of γ, $\theta(\gamma)$ is identified. To implement the bounds approach, we suppose that there is a true value for γ which is unknown but is known to belong to a set $\mathcal{G}$. We define

$$\widehat{\theta}(\gamma_0) = \arg\max_{\theta} h_N(W;\theta,\gamma_0)$$

and suppose that $\sqrt{N}(\widehat{\theta}(\gamma)-\theta) \xrightarrow{d} N(0,V(\gamma))$. That is, we assume that if we knew γ, the estimator obtained by maximizing the objective function at that value of γ would be consistent and asymptotically normal.[20] Further suppose we have access to a consistent estimator, $\widehat{V}_N(\gamma_0)$, of the asymptotic variance of $\widehat{\theta}(\gamma_0)$ under the hypothesis that $\gamma_0 = \gamma$. Then for each value of $\gamma_0 \in \mathcal{G}$, we may obtain an estimator $\widehat{\theta}(\gamma_0)$ and, under the maintained hypothesis that $\gamma_0 = \gamma$, construct a confidence interval analogous to (16) whose level might depend on γ_0 as

$$\begin{aligned} CI(1-a(\gamma_0),\gamma_0,a(\gamma_0)) &= (\widehat{\theta}(\gamma_0) + (\widehat{V}_N(\gamma_0)/N)^{1/2} c_{\alpha-a(\gamma_0)}, \\ &\quad \widehat{\theta}(\gamma_0) + (\widehat{V}_N(\gamma_0)/N)^{1/2} c_{1-a(\gamma_0)}). \end{aligned} \tag{20}$$

Confidence intervals using the methods in Sections 4.1 and 4.2 can be directly constructed by plugging (20) into expression (14) to obtain a union of confidence intervals utilizing only the knowledge of the support of γ. Likewise, (20) can be plugged into expression (17) to obtain a prior-weighted confidence interval which enables computaion of minimum length prior-weighted interval via solution of (18).

For the local-to-zero approach in Section 4.3, we suppose that there are prior beliefs that $\gamma = \gamma^* + \eta/\sqrt{N}$ where $\eta \sim G$ and η is indepenent of all other variables. In other words, we assume that we know that γ is local to some value γ^*. As before, the normalization by $\sqrt{N}$ is best thought of as a thought experiment made to produce an asymptotic approximation in which both uncertainty about the true value of γ and sampling

[20] This condition will be satisfied under the usual conditions for consistency and asymptotic normality of M-estimators.

error play a part. If we then consider $\widehat{\theta}(\gamma^*)$ as our point estimator we can obtain an approximation for the distribution of $\widehat{\theta}(\gamma^*)$ via the usual approach of linearizing first order conditions. Under regularity conditions

$$\sqrt{N}(\widehat{\theta}(\gamma^*) - \theta) \xrightarrow{d} \left[-H(\theta,\gamma^*)^{-1}v\right] - H(\theta,\gamma^*)^{-1}J(\theta,\gamma^*)\eta$$

where v is a normally distributed random variable, $H(\theta,\gamma^*)$ is the limit of $\frac{\partial^2 h_N(W;\theta,\gamma^*)}{\partial\theta\partial\gamma'}$ and $J(\theta,\gamma^*)$ is the limit of $\frac{\partial^2 h_N(W;\theta,\gamma^*)}{\partial\theta\partial\gamma'}$. The term in brackets corresponds to the usual limit distribution for $\widehat{\theta}$, the effect of uncertainty about exogeneity is to add in the term $-H(\theta,\gamma^*)^{-1}J(\theta,\gamma^*)\eta$ whose distribution of course depends on the prior G.

8.8 MCMC Details

We outline our general strategy for full Bayesian inference for the model in (1) and (2). The only difference between our sampler and the sampler of Rossi et al (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 $\Sigma \equiv Cov(\varepsilon, V) \sim$ 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, i.e. the joint distribution of $(\varepsilon_i,\ v_i)$. 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) (GS.1)

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

In the normal case, (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) (GS.2a)

$\Pi|\Theta$, (X,Y,Z) (GS.2b)

Given Θ, we can standardize appropriately (by substracting the mean vector and pre-multiplying by the inverse of the Cholesky root of Σ). The draws in (GS2.a) are then done by realizing that, given Π, we "observe" v_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, chapter 7) provides details of these draws.

For the prior on gamma in (17), we cannot draw (β, γ) in one draw but must draw from the appropriate conditionals.

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

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

We should note that for datasets with very weak instruments, the MCMC sampler defined above can be highly autocorrelated, particularly for diffuse settings of the prior on γ. This is not a problem if sufficient draws can be completed.

We also note that one could use Bayesian methods based on other parametric likelihoods for (ε, V) or a nonparametric Bayesian approach. For example, Conley et al (2006) demonstrate how fully non-parametric Bayesian inference can be conducted by using a Dirichlet Process prior for the distribution of the error terms.

9 References

Abadie, A (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231-263, 2003.

Anderson, T. W. and Rubin, H. (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics,* 20, 46-63.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817-858.

Angrist, J. D. and Imbens, G. W. (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica,* 62, 467-475.

Angrist, J. D. and Imbens, G. W. (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association,* 90, 431-442.

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

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

Angrist, J. D. and A. Krueger (1994): "Why Do World War II Veterans Earn More than Nonveterans?" *Journal of Labor Economics,* 12, 74-97.

Benjamin, Daniel J. (2003). "Do 401(k)s Increase Saving? Evidence From Propensity Score Subclassification," *Journal of Public Economics*, 87(5-6), 1259-1290.

Berkowitz, D., Caner, M., and Fang, Y. (2006): "Are Nearly 'Exogenous Instruments' Reliable?" mimeo.

Bound, J., and D. A. Jaeger (1996): "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 5835.

Bound, J., D. A. Jaeger, and R. M. 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(430), 443-450.

Chernozhukov V. and C. Hansen (2004): "The Impact of 401K Participation on Savings: an IV-QR Analysis." *Review of Economics and Statistics*. 86(3), 735-751.

Chernozhukov, V., Hong, H., and Tamer, E. (2007): "Estimation and Inference on Identified Parameter Sets," *Econometrica,* 75(5), 1243-1284.

Chintagunta, P.K., Dubé, J.P., and Goh, K.Y. (2005): "Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household-Level Brand Choice Models", *Management Science*, 51(5), 832-849.

Conley T.G. (1999) "GMM with Cross Sectional Dependence" *Journal of Econometrics* 92:1-45.

Conley T., C. Hansen, P. Rossi, and R. E. McCullogh (2006): "A Non-parametric Bayesian Approach to the Instrumental Variable Problem." University of Chicago, GSB, Working Paper.

Engen, E. M., W. G. Gale and J. K. Scholz (1996): "The Illusory Effects of Saving Incentives on Saving." *Journal of Economic Perspectives*, 10(4), 113-138.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998) "Dual and Simultaneous Sensitivity Analysis for Matched Pairs" *Biometrik*a 85 907-920.

Hahn, J. and Hausman, J. A., (2003): "IV Estimation with Valid and Invalid Instruments." MIT Department of Economics Working Paper No. 03-26.

Horowitz J. and C. F. Manski (1995) Econometrica "Identification and Robustness with Contaminated and Corrupted Data" *Econometrica* 63: 281-302.

Hotz VJ, C. Mullins, S. Sanders (2007) "Bounding Causal Effects from a Contaminated Natural Experiment: Anaylsing the Effects of Teenage Childbearing" *Review of Economic Studies* 64: 575-603.

Imbens, G. W. (2003) "Sensitivity to Exogeneity Assumptions in Program Evaluation" *American Economic Review, Papers and Proceedings* 93 126-132.

Manski, C. F. (2003) *Partial Identification of Probability Distributions* Springer-Verlag.

Manski, C. F. and Pepper, J. V. (2000) "Monotone Instrumental Variables: With an Application to the Returns to Schooling" *Econometrica*, 68 997-1010.

Poterba, J. M., S. Venti and D. Wise (1994): "Targeted Retirement Saving and the Net Worth of Elderly Americans," *American Economic Review*, 84(2), 180-185

Poterba, J. M., S. Venti and D. Wise (1995): "Do 401(k) Contributions Crowd Out Other Private Saving?," *Journal of Public Economics*, 58(1), 1-32.

Poterba, J. M., S. Venti and D. Wise (1996): "How Retirement Saving Programs Increase Saving," *Journal of Economic Perspectives* 10(4), 91-112.

Rosenbaum, P. R. (1987): "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies" *Biometrika* 74, 13-26.

Rosenbaum, P. R. (2002): *Observational Studies* 2nd ed. Springer-Verlag.

Rossi, P. G. Allenby, and R.E. McCulloch (2005): *Bayesian Statistics and Marketing* John Wiley and Sons Ltd New York.

Small, D. (2007): "Sensitivity Analysis for Instrumental Variables Regression with Overidentifying Restrictions," *Journal of the American Statistical Association,* 102, 1049-1058.

White, H. (2001): *Asymptotic Theory for Econometricians.* San Diego: Academic Press, revised edn.

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)

90% 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 one with a standard error of one, and when $\gamma=\gamma_2$, the estimate of β is four with a standard error of two. Intervals in the "Support Restriction" columns are obtained imposing only that γ takes on one of 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.

Table 2. 2SLS Estimates for Different Values of γ in Empirical Examples

A. 401(k) Example

	γ=0	γ=2500	γ=5000	γ=7500	γ=10000
$\hat{\beta}(\gamma)$	13222	9636	6049	2463	-1124
s.e.	1913	1934	1915	1917	1920

B. Demand for Margarine Example

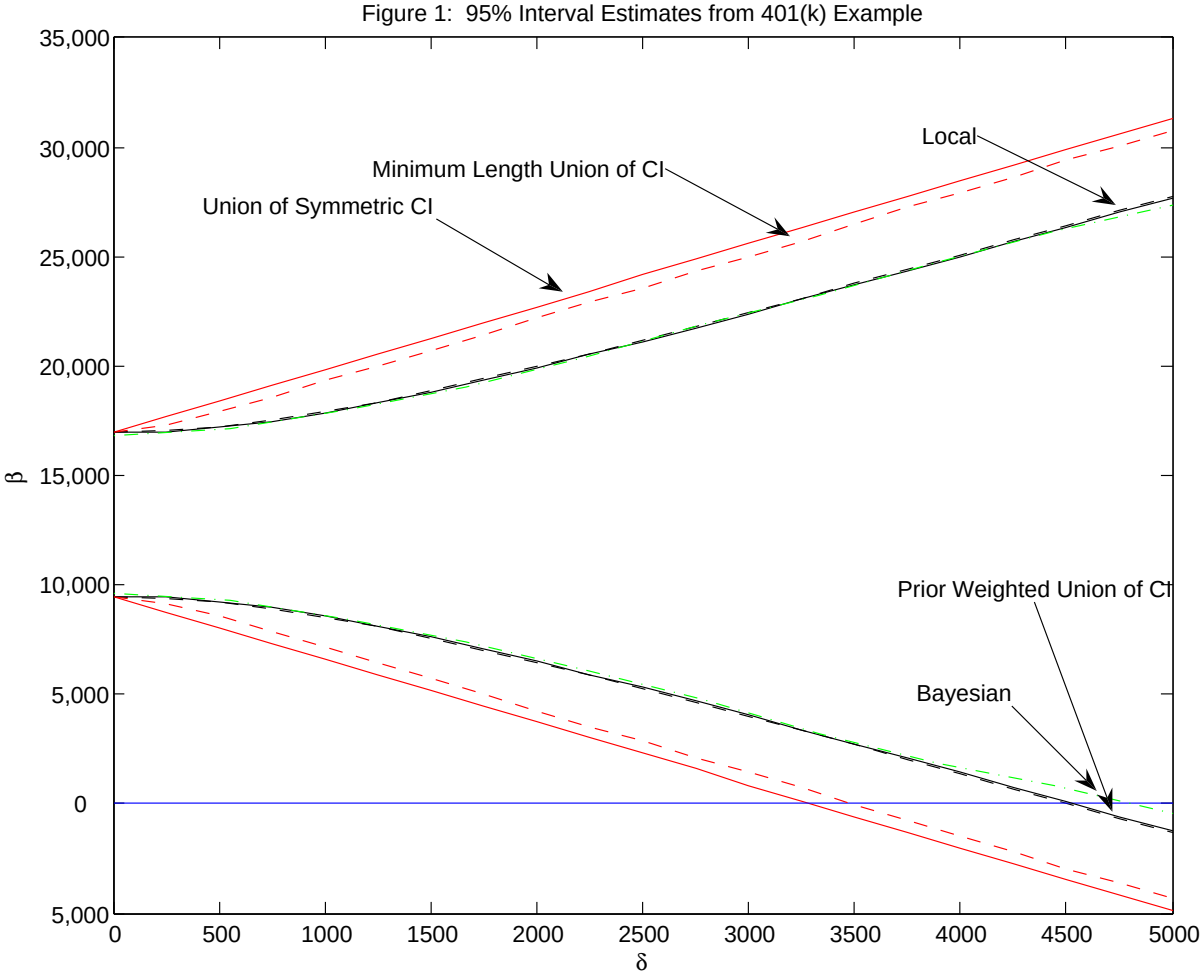
	γ=0	γ=-.075β	γ=-.15β	γ=.075β	γ=.15β
$\hat{\beta}(\gamma)$	-3.799	-4.302	-4.959	-3.401	-3.079
s.e.	0.611	0.707	0.858	0.544	0.495

C. Returns to Schooling Example

	γ_1=0, γ_2=0, γ_3=0	γ_1=0, γ_2=0, γ_3=.005	γ_1=0, γ_2=0, γ_3=.01	γ_1=.005, γ_2=.005, γ_3=.005	γ_1=.01, γ_2=.01, γ_3=.01
$\hat{\beta}(\gamma)$	0.1077	0.0797	0.0516	0.0792	0.0508
s.e.	0.0196	0.0192	0.0192	0.0192	0.0192

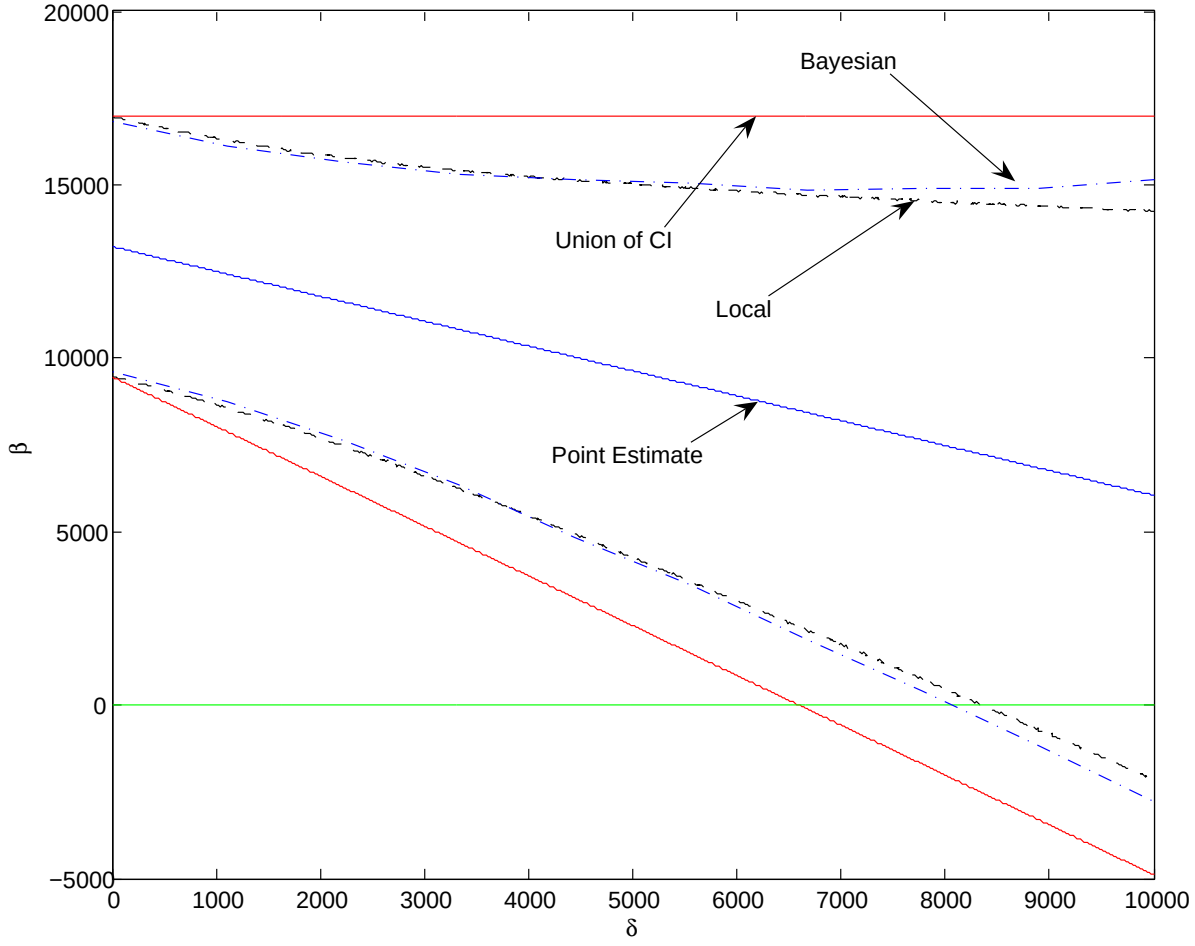
Point estimates, $\hat{\beta}(\gamma)$ and estimated standard errors, s.e., for the treatment effect of interest for various values of γ, the "direct effect" of Z on Y. In the 401(k) example, Z is whether an individual's firm offers a 401(k) plan, and Y is net financial assets. In the demand example, Z is the log wholesale price of margarine of a given brand, and Y is the log market share of that brand. In the schooling example, Z_1, Z_2, and Z_3 are respectively indicators for being born in the second through fourth quarters of the year.

Figure 1: 95% Interval Estimates from 401(k) Example



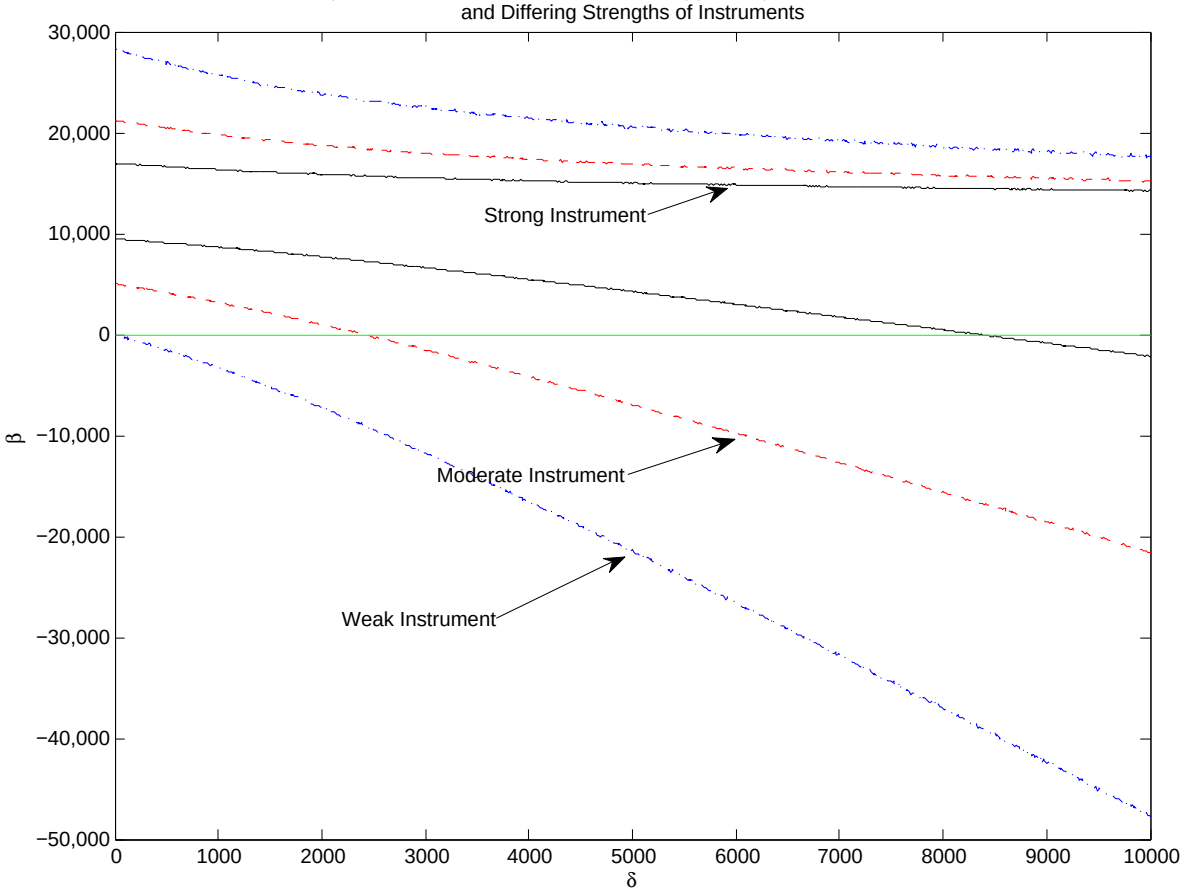
Note: This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets using each of our proposed methods and across various prior settings. The definition of δ differs between the support only intervals; "Union of Symmetric CI" and "Minimum Length Union of CI"; and the intervals that use the full prior; "Prior Weighted Union of CI", "Local", and "Bayesian." 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 $\gamma \sim N(0, \delta^2)$.

Figure 2: 95% Interval Estimates for 401(k) Example with Positive Prior



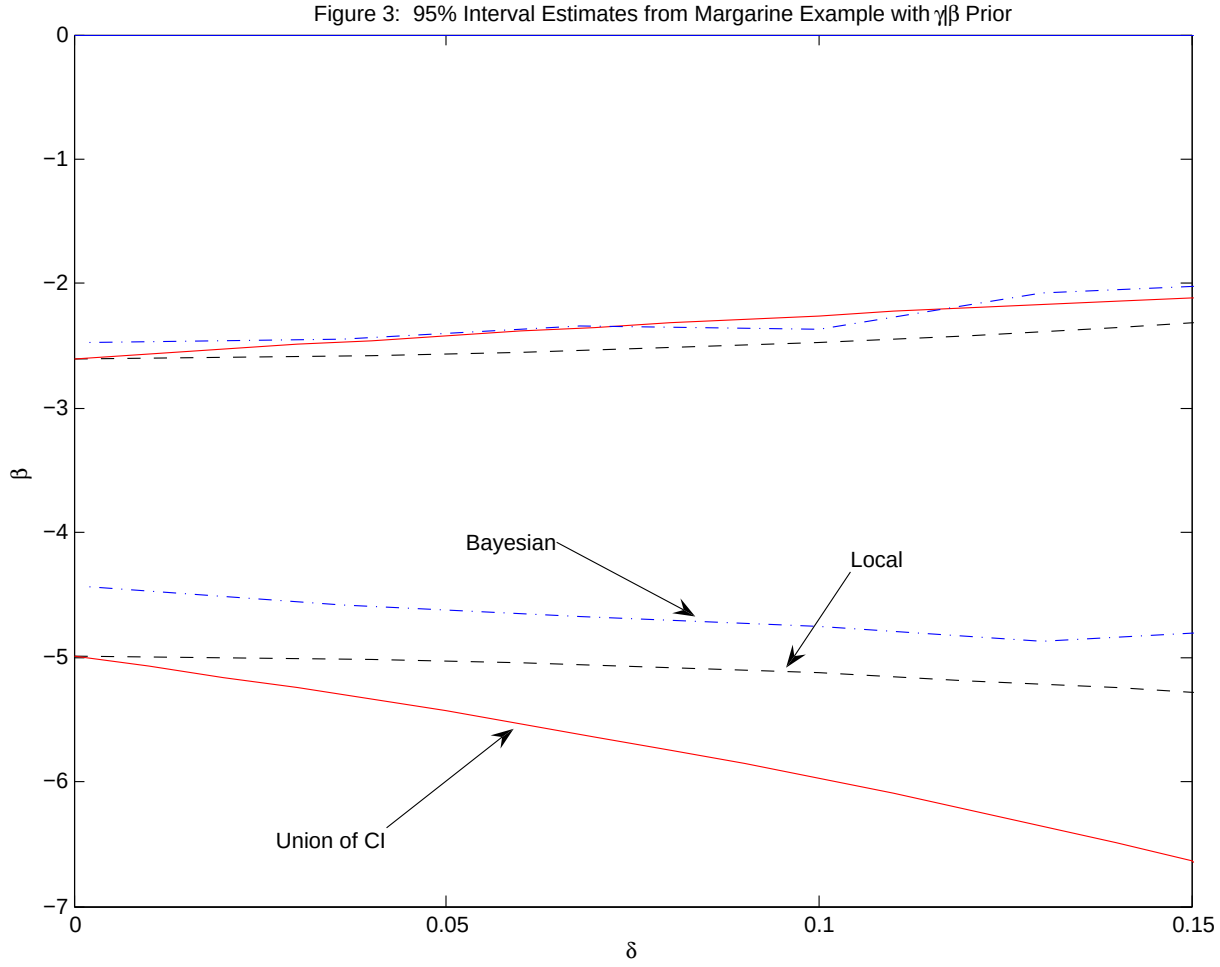
Note: This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets across various prior settings using priors that the direct effect of 401(k) eligibility on net financial assets is nonnegative. The definition of δ differs between the different intervals. 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.

Figure 3: 95% Interval Estimates for 401(k) Example with Positive Prior and Differing Strengths of Instruments



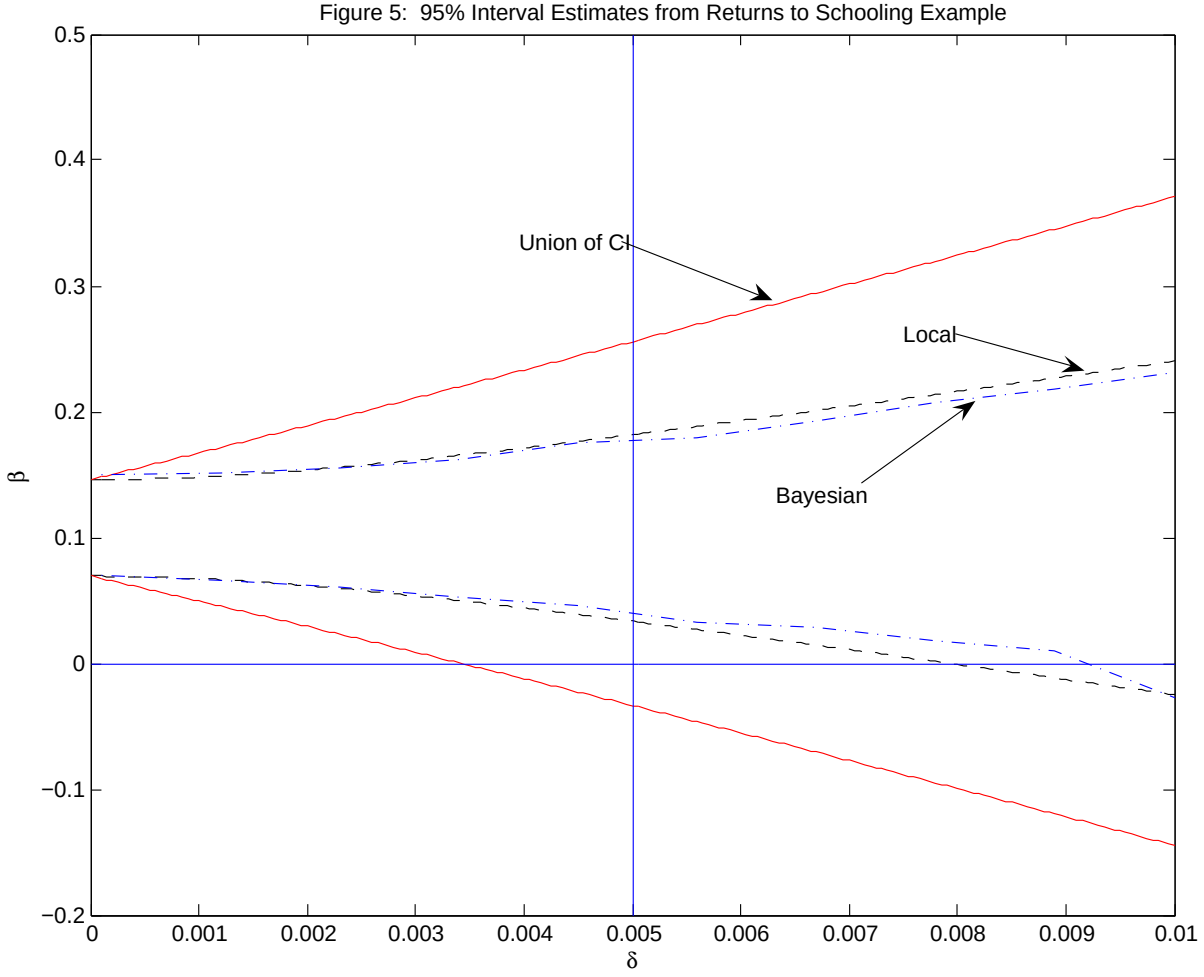
Note: This figure presents 95% confidence intervals for the effect of 401(k) participation on net financial assets for different strengths of instruments across various prior settings using priors that the direct effect of 401(k) eligibility on net financial assets is likely nonnegative. The definition 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 strong instrument is 401(k) eligibility from the data. The moderate and weak instrument are formed by adding noise to 401(k) eligility.

Figure 3: 95% Interval Estimates from Margarine Example with γ|β Prior



Note: 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 $\gamma \sim N(0, \delta^2\beta^2)$.

Figure 5: 95% Interval Estimates from Returns to Schooling Example



Note: 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.