

A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters

BY TIEMEN WOUTERSEN*

ABSTRACT. This note proposes a new way to calculate confidence intervals of partially identified parameters. It extends earlier work by letting the point identified case be a special case, by using the regular bootstrap to construct the confidence interval, and by allowing for an unknown correlation between the estimators of the upper and the lower bounds.

KEYWORDS: Bounds, Partial Identification, Confidence Intervals

1. INTRODUCTION

CONFIDENCE INTERVALS FOR POINT IDENTIFIED PARAMETERS have been well studied in econometrics, see for example Newey and McFadden (1994) for an overview. More recently, confidence intervals have been developed for partially identified parameters. Horowitz and Manski (1998) develop interval estimates that asymptotically cover the entire identified region with fixed probability. Chernozhukov, Hong and Tamer (2007) extend this approach through formulating the problem of covering the entire identified region as a minimization problem. Imbens and Manski (2004) derive the confidence interval that covers each element in the identified region with fixed probability. This confidence interval is in general shorter than the earlier derived confidence intervals. In this note, we develop an easier and perhaps more intuitive way to calculate confidence the interval that covers each element in the identified region with fixed probability bootstrap. We also extend earlier work by letting the point identified case be a special case, by using the regular bootstrap to construct the confidence interval, and by allowing that different estimators

*Comments are welcome, woutersen@jhu.edu, or Johns Hopkins University, Department of Economics, 3400 North Charles street, Baltimore, MD 21218.

A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters 2

(e.g. a least squares and an instrumental variable estimator) or different datasets can be used to estimate the upper and the lower bound.

An easy way to calculate the confidence interval of a point identified parameter is by estimating the (asymptotic) distribution function of the parameter and then calculating its τ^{th} and the $(1-\tau)^{\text{th}}$ percentile. Similarly, one may be to approximate the (asymptotic) distribution of both the lower and upper bound of a scalar parameter. In general, the confidence interval of the lower bound contains smaller values of the parameter than the confidence interval of the upper bound. In this note, we propose to ‘average’ the confidence intervals by averaging the distribution function of the lower and the upper bound. After averaging the distribution functions, the confidence interval is constructed as if we construct the confidence interval of a point, i.e. by calculating the τ^{th} and the $(100-\tau)^{\text{th}}$ percentile. The resulting confidence interval has the same asymptotic coverage properties as the confidence interval of Imbens and Manski (2004) but is much easier to calculate. In particular, this simple procedure can be implemented by applying the bootstrap method to the lower and upper bound while no correction term is needed to ensure uniform convergence. Imbens and Manski (2004) need a correction term to ensure uniform convergence in case the parameter is point identified and also assume that the estimator of the lower bound coincides with the estimator of the upper bound in case of point identification of the parameter. The simple procedure of averaging the distribution functions does not need a correction term if the estimators of the lower and upper bound coincide: Averaging the identical distribution function still yields the desired distribution function and, in this point identified case, the usual confidence intervals for point identification. Another advantage of the simple procedure is that estimators of the lower and upper bound can be different (e.g. the ordinary least squares estimator for the upper bound and an instrumental variable estimator for the lower bound, each estimator using a different dataset). Moreover, parameter by parameter confidence intervals can be constructed for vector valued parameters.

2. CONFIDENCE INTERVAL

For now, let the parameter of interest θ and its upper and lower bound, θ_l and θ_u , be scalars. The following assumption is identical to Imbens and Manski (2004, assumption 1 (i):

Assumption 1: There are estimators for the lower and the upper bound, $\hat{\theta}_l$ and $\hat{\theta}_u$, that satisfy:

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_l - \theta_l \\ \hat{\theta}_u - \theta_u \end{pmatrix} \xrightarrow{d} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{pmatrix} \right),$$

uniformly in $P \in \mathcal{P}$ and there are estimators $\widehat{\sigma}_l^2, \widehat{\sigma}_u^2$, and $\widehat{\rho}$ for σ_l^2, σ_u^2 , and ρ that converge to their population values uniformly in $P \in \mathcal{P}$ (ρ may be equal to one in absolute value, as in the case where the width of the identification region is known).

Imbens and Manski (2004) need their assumptions 1 (ii)-(iii) so that the estimate of the difference between the lower and upper bound is well behaved. Here, we do not estimate that difference and do not need to make those assumptions. Moreover, we also do not need the assumption that $\theta_l = f(P, \lambda_l)$ where λ_l is a quantity which is known only to belong to a specialized set. For example, the estimators of the lower and upper bound could be different or could be based on different datasets without reference to some λ_l . Let CI_α^θ denote the symmetric confidence interval with coverage probability α . Let $\Phi(\cdot)$ denote the distribution function of the standard normal distribution. Let $\underline{\theta}_N$ and $\bar{\theta}_N$ be chosen such that

$$\Phi\left(\sqrt{N}\frac{\underline{\theta}_N - \hat{\theta}_l}{\widehat{\sigma}_l}\right) + \Phi\left(\sqrt{N}\frac{\underline{\theta}_N - \hat{\theta}_u}{\widehat{\sigma}_u}\right) = 1 - \alpha \tag{1}$$

$$\Phi\left(\sqrt{N}\frac{\bar{\theta}_N - \hat{\theta}_l}{\widehat{\sigma}_l}\right) + \Phi\left(\sqrt{N}\frac{\bar{\theta}_N - \hat{\theta}_u}{\widehat{\sigma}_u}\right) = 1 + \alpha \tag{2}$$

where $\widehat{\sigma}_l = \sqrt{\widehat{\sigma}_l^2}$ and $\widehat{\sigma}_u = \sqrt{\widehat{\sigma}_u^2}$. The confidence interval can then be constructed as

$$CI_\alpha^\theta = [\underline{\theta}_N, \bar{\theta}_N]. \tag{3}$$

The following theorem gives the uniform coverage result.

Theorem 1

Let assumption 1 hold and let $\alpha \geq \frac{1}{2}$. Then

$$\lim_{N \rightarrow \infty} \inf_{P \in \mathcal{P}} \Pr(\theta \in CI_\alpha^\theta) \geq \alpha.$$

A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters 4

Proof: See appendix.

The following lemma is central to this note.

Lemma 1

Let $Z \sim N(0, 1)$, let $\alpha, c \in \mathbb{R}$. and let $0 \leq \alpha \leq 1$. Let $\Phi(\cdot)$ denote the cumulative distribution function of Z . Then

$$P(1 - \alpha \leq \Phi(Z) + \Phi(Z + c) \leq 1 + \alpha) = \alpha.$$

It is easy to see that the lemma is true for $c = 0$ and $c \rightarrow \pm\infty$. Consider $c = 0$,

$$P(1 - \alpha \leq 2 \cdot \Phi(Z) \leq 1 + \alpha) = P(1 - \alpha \leq 2 \cdot U \leq 1 + \alpha)$$

where U is uniformly distributed on the interval $[0, 1]$ so that

$$P(1 - \alpha \leq 2 \cdot \Phi(Z) \leq 1 + \alpha) = \frac{1 + \alpha}{2} - \frac{1 - \alpha}{2} = \alpha.$$

For $c \rightarrow \infty$, we have

$$\begin{aligned} \lim_{c \rightarrow \infty} P(1 - \alpha \leq \Phi(Z) + \Phi(Z + c) \leq 1 + \alpha) \\ &= P(1 - \alpha \leq \Phi(Z) + 1 \leq 1 + \alpha) \\ &= P(0 \leq \Phi(Z) \leq \alpha) = \alpha. \end{aligned}$$

For a fixed data generating process, one can often use the bootstrap to calculate CI_{α}^{θ} . For example, the bootstrap can be used if the estimators of the lower and upper bound can be written as the sum of an ‘influence function’ in the terminology of Newey and McFadden (1994) and a term that converges in probability to zero, i.e. $\hat{\theta}_l = \theta_l + \frac{1}{\sqrt{N}} \sum_i g_l(X_i) + o_p(1)$, $\hat{\theta}_u = \theta_u + \frac{1}{\sqrt{N}} \sum_i g_u(X_i) + o_p(1)$ for some $g_l(\cdot)$ and $g_u(\cdot)$ and X being independent identically distributed (Horowitz, 2001, theorem 2.2). The bootstrap estimates the tail probabilities consistently in this case and can be used to approximate the tail probabilities of $\hat{\theta}_l$ and $\hat{\theta}_u$ as well as the average of the two. If, in addition, the assumptions of theorem 3.1 of Horowitz (2001) hold, then the bootstrap yields an asymptotic refinement over the regular first order asymptotics.

Imbens and Manski (2004) assume that $\theta_l = f(P, \lambda_l)$ and $\theta_u = f(P, \lambda_u)$ where are the (smallest and largest) elements of a set. For example, their section 3 discusses an example in which the value of the dependent variable is observed with probability p while the mean value of the unobserved observations cannot be estimated and is assumed to lie between zero and one, $\lambda \in [0, 1]$. If $p = 1$, i.e. we observe all outcomes, then the lower and upper bound coincide with the sample average.

Assumption 2: There are estimators for the lower and the upper bound, $\hat{\theta}_l$ and $\hat{\theta}_u$, that satisfy:

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_l - \theta_l \\ \hat{\theta}_u - \theta_u \end{pmatrix} \xrightarrow{d} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{pmatrix} \right),$$

and there are estimators $\hat{\sigma}_l^2$, and $\hat{\sigma}_u^2$, for σ_l^2 , and σ_u^2 , that converge in probability to their population values.

Theorem 2

Let assumption 1 hold and let $\alpha \geq \frac{1}{2}$. Then

$$\lim_{N \rightarrow \infty} \Pr(\theta \in CI_\alpha^\theta) \geq \alpha.$$

Proof: See appendix.

Theorem 2 allows us to use the bootstrap. In particular, the bootstrap can be used under the same conditions as before and parameter by parameter confidence intervals can be constructed for vector valued parameters.

3. APPENDIX

Lemma 1

Let $Z \sim N(0, 1)$, let $\alpha, c \in \mathbb{R}$. and let $0 \leq \alpha \leq 1$. Let $\Phi(\cdot)$ denote the cumulative distribution function of Z . Then

$$P(1 - \alpha \leq \Phi(Z) + \Phi(Z + c) \leq 1 + \alpha) = \alpha.$$

Theorem 1

Let assumption 1 hold and let $\alpha \geq \frac{1}{2}$. Then

$$\lim_{N \rightarrow \infty} \inf_{P \in \mathcal{P}} \Pr(\theta \in CI_\alpha^\theta) \geq \alpha.$$

A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters 6

Proof: Note that, by construction, $\underline{\theta}_N \leq \bar{\theta}_N$. We first consider $\theta_0 = \underline{\theta}$. For simplicity, we first prove the theorem under the stronger assumption that

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_l - \theta_l \\ \hat{\theta}_u - \theta_u \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{pmatrix} \right)$$

for known σ_l^2 , σ_u^2 and $\rho = 1$. Note that $\underline{\theta}$ would not be an element of the confidence interval if $H(\underline{\theta}) < 1 - \alpha$ or $H(\underline{\theta}) > 1 + \alpha$. In particular,

$$\begin{aligned} \Pr(\underline{\theta} \in CI_\alpha^\theta) &= P(1 - \alpha \leq H(\underline{\theta}) \leq 1 + \alpha) \\ &= P(1 - \alpha \leq \Phi(\sqrt{N} \frac{\underline{\theta} - \hat{\theta}_l}{\hat{\sigma}_l}) + \Phi(\sqrt{N} \frac{\underline{\theta} - \hat{\theta}_u}{\hat{\sigma}_u}) \leq 1 + \alpha) \\ &= P(1 - \alpha \leq \Phi(\sqrt{N} \frac{\underline{\theta} - \hat{\theta}_l}{\sigma_l}) + \Phi(\sqrt{N} \frac{\underline{\theta} - \bar{\theta}}{\sigma_u} + \sqrt{N} \frac{\bar{\theta} - \hat{\theta}_u}{\sigma_u}) \leq 1 + \alpha) \end{aligned}$$

For any N , $\frac{\underline{\theta} - \hat{\theta}_l}{\sigma_l} = \frac{\bar{\theta} - \hat{\theta}_u}{\sigma_u}$ has a standard normal distribution. Therefore,

$$\Pr(\underline{\theta} \in CI_\alpha^\theta) = P(1 - \alpha \leq Z + \Phi(\sqrt{N} \frac{\underline{\theta} - \bar{\theta}}{\sigma_u} + Z) \leq 1 + \alpha).$$

Lemma 1 applies so that

$$\Pr(\underline{\theta} \in CI_\alpha^\theta) = \alpha.$$

Similarly,

$$\Pr(\bar{\theta} \in CI_\alpha^\theta) = \alpha.$$

Since $\Pr(\theta_0 \in CI_\alpha^\theta)$ is minimized at $\theta_0 = \underline{\theta}$ or $\theta_0 = \bar{\theta}$, we have

$$\Pr(\underline{\theta} \in CI_\alpha^\theta) \geq \alpha.$$

Let assumption 1 hold and let $\alpha \geq \frac{1}{2}$. Let We first consider $\theta_0 = \underline{\theta}$. Then

$$\lim_{N \rightarrow \infty} \inf_{P \in \mathcal{P}} \Pr(\underline{\theta} \in CI_\alpha^\theta) = \lim_{N \rightarrow \infty} \inf_{P \in \mathcal{P}} .P(1 - \alpha \leq H(\underline{\theta}) \leq 1 + \alpha) \geq \alpha.$$

As above, $\lim_{N \rightarrow \infty} \Pr(\theta_0 \in CI_\alpha^\theta)$ is minimized at $\theta_0 = \underline{\theta}$ or $\theta_0 = \bar{\theta}$, so that

$$\lim_{N \rightarrow \infty} \inf_{P \in \mathcal{P}} \Pr(\underline{\theta} \in CI_\alpha^\theta) \geq \alpha.$$

REFERENCES

- [1] Andrews, D. W. K., and P. Guggenberger (2005): “Validity of Subsampling and “Plug-in Asymptotic” Inference for Parameters Defined by Moment Inequalities”; *Econometric Theory* (2009, forthcoming).
- [2] Bugni, F. A. (2008): “Bootstrap Inference in Partially Identified Models”, unpublished manuscript.
- [3] Chernozhukov, V., H. Hong and E. Tamer (2007): “Parameter Set Inference in a Class of Econometric Models,” *Econometrica*, 75(5), page 1243–1284.
- [4] Horowitz, J. L. (2001): “The Bootstrap” in *Handbook of Econometrics*, Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.
- [5] Horowitz, J. L., and C. Manski (1998): “Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations,” *Journal of Econometrics*, 84, 37-58.
- [6] Imbens, G. W. and C. F. Manski (2004): “Confidence Intervals for Partially Identified Parameters”, *Econometrica*, 72, 1845-1857.
- [7] Manski, C. (1990): “Nonparametric Bounds on Treatment Effects,” *American Economic Review Papers and Proceedings*, 80, 319-323.
- [8] Manski, C. (2003): “*Partial Identification of Probability Distributions*”, New York: Springer-Verlag.
- [9] Newey, W. K. (1991): “Uniform Convergence in Probability and Stochastic Equicontinuity”, *Econometrica*, 59, 1161-1167.
- [10] Newey, W. K., and D. McFadden (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

Endogeneity and Imperfect Instruments: Estimating Bounds for the Effect of Early Childbearing on High School Completion

Steffen Reinhold* Tiemen Woutersen[†]

March 3, 2009

Abstract

This paper derives informative bounds of the effect of early childbearing on high school completion. We allow the exclusion restriction of the instrument to be violated. In particular, we assume that the correlation between the instrument and the structural error is smaller than the correlation between the structural error and the endogenous regressor. We derive a confidence interval using the regular bootstrap and find that the least squares estimate is outside this confidence interval. That is, the bias of the least squares estimator is both substantial and statistically significant.

Keywords: Instrumental Variables, Validity, Identification, Bounds, Teenage Childbearing, Educational Attainment.

JEL Classification: C310, J130

*Mannheim Research Institute for the Economics of Aging, University of Mannheim; L13,17; 68131 Mannheim; Germany; email: reinhold@mea.uni-mannheim.de

[†]Johns Hopkins University, Department of Economics; 3400 N. Charles Street; MD 21218; email: woutersen@jhu.edu. We would like to thank Colleen Carey for valuable comments.

1 Introduction

There is a close association between early childbearing and adverse economic outcomes for mothers and their children such as poverty risk, lower educational achievement, and depressed earnings. While it is plausible to describe these outcomes as consequences of early fertility decisions, this correlation can also arise from a correlation of unobserved factors with early childbearing. Instrumental variables are one approach to deal with the problems of endogenous regressors. However, depending on the exact choice of the instruments, a wide range of estimates emerges.

An instrument needs to be correlated with the endogenous regressor teenage childbearing (relevance) but it must not be correlated with the structural error term in the outcome equation (validity). One problem with this approach is that often doubts linger as to whether the instruments satisfy this second assumption. In this paper, we employ a novel set identification approach with imperfect instruments based on the assumption that the correlation between the instrument and the structural error is smaller than the correlation between the structural error and the endogenous regressor, childbearing. Nevo and Rosen (2008) were the first to use this inequality. This approach relaxes the validity assumption, allowing limited correlation with the error term and thus widens the set of potential instruments. Nevo and Rosen's work builds on a recent literature on partial identification. Horowitz and Manski (1998) develop interval estimates that asymptotically cover the entire identified region with fixed probability. Chernozhukov, Hong and Tamer (2007) extend this approach through formulating the problem of covering the entire identified region as a minimization problem. Imbens and Manski (2004) derive the confidence interval that covers each element in the identified region with fixed probability. This confidence interval is in general shorter than the earlier derived confidence intervals. Woutersen (2008) develops an easier and perhaps more intuitive way to calculate the confidence interval that covers each element in the identified region with fixed probability bootstrap; the identified case is a special case.

We investigate whether we can identify informative bounds for the causal effect of teenage pregnancy on high school completion under the relaxed assumptions. A researcher may also

be interested in the question of how much her results are driven by the identifying assumptions by employing a weaker assumptions on the instruments. In addition, this paper helps to assess whether set estimators, in general, provide informative bounds on the parameters of interest in an application that is typical for the literature in labor and demographic economics. It is an empirical application of Woutersen's (2008) approach to estimating bounds for partially identified parameters.

In the literature on the effects of teenage childbearing on high school completion, the problem of potentially invalid instruments is particularly urgent: Hotz et al. (1997, 1999, 2005) use the occurrence of a miscarriage among women who have become pregnant as teenagers as an instrument for teenage childbearing. They use the sample of teenagers who have become pregnant because they argue that miscarriages are a random phenomenon for pregnant women. But since pregnancy itself is not a random event, their inference does not extend to the general population. This is because the occurrence of a miscarriage does not satisfy the strong assumption of validity when applied to the full sample of women. But even in the restricted sample there is the possibility that the instrument does not satisfy the validity assumption if miscarriages are correlated with unobserved family background or the woman's health behavior. However with the identifying assumptions of Nevo and Rosen (2008), the occurrence of a miscarriage becomes a legitimate instrument even in the full sample if one believes that the correlation between it and the unobserved factors in the outcome equation is smaller than the correlation between teenage childbearing and the unobserved factors.

Another potential instrument is age at menarche (Ribar, 1994). Age at menarche is correlated with teenage pregnancy and via this pathway also with teenage childbearing. Ribar therefore uses this instrument without a sample restriction to women who have become pregnant as teenagers. Using only age at menarche, Ribar finds a detrimental effect of teenage childbearing on educational achievement. Although Ribar's empirical setup is different, the same fundamental problem emerges. Age at menarche may be correlated with other unobserved factors like nutrition during childhood (Freedman, Khan, Serdula, Dietz, Srinivasan,

and Berenson 2002) which may itself influence education. This would render this instrument invalid.

Using different sets of instrument often leads to differing qualitative conclusions about the causal effect of teenage childbearing on high school completion as in this particular case. Whereas the results using the occurrence of a miscarriage indicate no adverse effect of teenage childbearing one finds a rather strong adverse effect on high school completion using age at menarche as an instrument.

The paper is organized as follows. After discussing the data, we present some details about estimating the bounds of the causal effect of teenage childbearing. We discuss in detail our identifying assumptions about the unobserved correlation between the instruments and the error term. In a further subsection, we also discuss how one can derive 95% confidence intervals for the parameter of interest using both Imbens and Manski (2004) and Woutersen (2008). In a next section we discuss the results.

2 Description of Data

We use the 2002 cycle of the National Survey of Family Growth (NSFG) a representative sample of 7643 women from the ages 15-44. We restrict our sample to women older than 20 years. At this age, high school should normally be completed. With this sample restriction we have 6443 observations in our complete sample. In addition, we also construct subsamples of women having become pregnant before age 18 ($N=1284$), ever pregnant women ($N=4810$), and women having children in their household ($N=3910$).

Thus, we use the instruments in less restricted samples than Hotz et al. who only estimated the effect of teenage childbearing in a sample of woman who experienced teenage pregnancy. Under the stronger validity assumptions this restriction is clearly reasonable. However, under our relaxed assumptions we can use the occurrence of a miscarriage even in the full sample of women as long as the correlation between teenage childbearing and the structural error term is smaller as the correlation between the occurrence of a miscarriage

and the structural error term.

As additional controls we also use dummies for religious affiliation (Catholic, Protestant, no religion, and other), race, age and age squared, a dummy for migrant status, intact family background, and dummies for parental education.

3 Empirical Model

3.1 Estimation of Bounds

Consider the following model:

$$Y = X\beta + W\delta + U \tag{1}$$

where Y is a dummy which takes the value of 1 for not completing high school, X is a dummy which takes the value 1 if the individual gives birth as a teenager. For those who miscarry this dummy can still take the value of 1 if a second teenage pregnancy results in a live-birth, but these cases are rare. W is a vector of other covariates including age at interview and its square, race, religion, parents' educational background, intact family background, and migration background. There is also a set of imperfect instruments Z including age at menarche and the occurrence of a miscarriage.

Controlling for other covariates in IV regressions is often important because the assumption of zero correlation between the error term and the instrument only needs to hold after conditioning on all exogenous variables. Since we do not require zero correlation between the instrument and the error term we can proceed with estimating the model without covariates. Also, for estimating we will partial out all other covariates, so that we essentially are estimating a simple bivariate model with one endogenous regressor. For this reason, we also discuss the following bivariate model:

$$Y = X\beta + U \tag{2}$$

There are also some imperfect instruments Z . Nevo and Rosen (2008) assume that the

correlations between the endogenous regressor and the error term and between the instrument and the error term have the same sign:

$$\rho_{XU}\rho_{Z_jU} \geq 0 \tag{3}$$

where j indexes the instrument in the case of multiple instruments. This is not restrictive because one can just multiply the instrument by negative 1 for convenience without loss of generality. But one needs to make an assumption about these correlations because one never can observe the true error term. In their fourth assumption, Nevo and Rosen assume that the correlation between the instruments and the error term is weaker in absolute terms than the correlation between the endogenous regressor and the error term, that is:

$$|\rho_{XU}| \geq |\rho_{ZU}| \tag{4}$$

This assumption considerably weakens the usual assumptions for instrumental variables which would require that $\rho_{ZU} = 0$. We believe that in our case these assumptions are satisfied.

Hotz et al. argue that pregnancies are not a random event in the population, and hence the occurrence of a miscarriage cannot be a valid instrument in the full sample because it is correlated with pregnancies. Becoming pregnant as a teenager may well be correlated with the error term, for example because of unobserved preferences for education or future earnings potential. Conditional on being pregnant, however, miscarriages are largely a random event with possible some behavioral risk factors such as illicit drug use. If miscarriages are largely random then the correlation between miscarriages and the error term must be smaller than the correlation between teenage childbearing and the error term. Similarly, there seems to be some variation in age at menarche which is not correlated with is also not strongly influenced by either family background or behavioral factors under the individual's control.

In addition, they impose the usual rank conditions ensuring that the probability limits of OLS and 2SLS estimators are well defined. We will rely for estimation on their second

proposition, and here it is important to also know the correlation between the instruments and the endogenous regressor. Our assumptions on these correlations are detailed in the next section. If all four assumptions are satisfied, then the identified set is given by:

$$\text{If } \sigma_{XZ} < 0 \text{ then:} \tag{5}$$

$$B^* = [\beta_Z^{IV}, \beta_{Z^*}^{IV}] \text{ if } \sigma_{XU} > 0 \tag{6}$$

$$B^* = [\beta_{Z^*}^{IV}, \beta_Z^{IV}] \text{ if } \sigma_{XU} < 0 \tag{7}$$

$\sigma_{XZ} < 0$ is the covariance between the endogenous regressor and the instrument, and σ_X , σ_Z are the standard deviations of the respective variables, and β_Z^{IV} is a 2SLS estimator using the instrument Z . Z^* is a new instrument which can be easily constructed from the data as $Z^* = \sigma_Z X - \sigma_X Z$.

$\sigma_{XZ} < 0$ is the covariance between the endogenous regressor and the instrument, and σ_X , σ_Z are the standard deviations of the respective variables. If $\sigma_{XZ} > 0$ then:

$$B^* = (-\infty, \min(\beta_Z^{IV}, \beta_{Z^*}^{IV})] \text{ if } \sigma_{XU} > 0 \tag{8}$$

$$B^* = [\max(\beta_Z^{IV}, \beta_{Z^*}^{IV}), +\infty) \text{ if } \sigma_{XU} < 0 \tag{9}$$

Nevo and Rosen (2008) show that one can combine the bounds implied by different instrumental variable estimates by finding their intersection which we also do. These bounds are sharp. Because the correlation between teenage childbearing and the occurrence of a miscarriage is negative we already have two-sided bounds using the occurrence of a miscarriage as an instrument. The question becomes whether age at menarche helps in tightening these bounds. This depends on whether the upper bound using age at menarche as an instrument is smaller than the upper bound using the occurrence of a miscarriage as an instrument.

3.2 Assumptions on the Correlation between the Instruments and the Unobserved Heterogeneity

Nevo and Rosen (2008) replace the assumption of no correlation between the instrument and the error term by the weaker assumption that this correlation has to be smaller in absolute terms than the correlation between the endogenous regressor and the error term. However, one needs to make an assumption about the sign of this correlation. In the following section we discuss these assumptions, and in addition, we provide key summary statistics for the observed variances and covariances of the endogenous regressor and the instruments which are needed to calculate the bounds.

We summarize the information about our instruments and our assumptions about key correlations in Tables 1 and 2. Given our assumptions on the correlation between the unobserved factors and our instruments, one can derive two-sided bounds using the occurrence of a miscarriage as an instrument, whereas one can estimate upper bounds using age at menarche (or its negative) as an instruments.

We assert that there is a positive correlation between the error term and early childbearing. In our example, a dummy for not completing high school is the outcome variable and teenage childbearing is the endogenous regressor. Unobserved factors such as a disadvantaged family background may make a teenager both more likely to have children very early and to drop out of high school. The covariances between the instruments and the dummy for early childbearing have the sign one would expect. Women who have experienced a miscarriage are less likely to have experienced a live-birth as teenagers, and teenagers with an early onset of their menarche are more likely to become pregnant and bear children as teenagers.

Furthermore, we assume that the occurrence of a miscarriage, henceforth Z_1 , is also positively correlated with the error term. In the full sample of all women, miscarriages are positively correlated with pregnancies and to the extent that teenagers from disadvantaged families have earlier pregnancies one would expect a positive correlation. In addition, the

occurrence of miscarriages may still be correlated with certain risk factors such as smoking or drinking (Hotz, McElroy, and Sanders 1999) or a very young age at conception (see Reinhold 2007, and cites therein) which also correlate positively with dropping out of high school.

For the second instrument, age at menarche (Z_2), the direction of the correlation is more open for argument. Weil (2007) uses age at menarche as a health indicator in his work and reports that in poorer countries the age at menarche is later. If poor living conditions are associated with a late age at menarche one could expect that $\rho_{Z_2U} > 0$. Notice however, that he uses this indicator to compare developed with developing countries. In our sample, we only have a comparison within the United States, and we assume that $\rho_{Z_2U} < 0$. Freedman et al. (2002) present evidence that obesity and an early age at menarche are positively correlated. Obesity is a marker of a disadvantaged family background which would indicate that women from disadvantaged family background may have a lower age at menarche.

The assumptions on the correlation structure are the same for the more general version of the model where there are additional covariates W . In this case, one regresses both the outcome and the endogenous regressor on these covariates and obtains residuals from these regressions. Let \tilde{X}, \tilde{Y} denote these residuals. Using these residuals one can estimate a bivariate model. Alternatively, one can also just use 2SLS estimation with additional covariates to obtain the same estimates.

3.3 Estimating Confidence Intervals Covering the True Parameter

Woutersen (2008) allows for the regular bootstrap to be used for confidence intervals. That is, we sample individuals with replacement and get an estimate for the lower and upper bound for each subsample. We generate 5000 subsamples. We then put all the estimates of the upper and lower bound in a ordered vector (with length 10,000) and calculate the 2.5% and 97.5% percentile. This confidence interval has the correct coverage and the asymptotic refinement of the bootstrap

4 Results

Table 3 presents the estimates of the bounds of the effect on early childbearing both with and without additional covariates for the complete sample and for subsamples of women who have ever been pregnant, women who have become pregnant as teenagers, and women who currently have children in their household. In addition, we present the OLS estimates of the coefficient on teenage childbearing in these samples. The bounds estimates use the occurrence of a miscarriage, negative of age at menarche, and the transformed Z^* as instruments. Using these regression estimates, we can derive bounds for the effect of teenage childbearing on high school completion.

We use Woutersen's (2008) method to calculate the 95% confidence intervals for the set estimates employing the bootstrap. In addition, we also estimated 95% confidence intervals using the method of Imbens and Manski (2004). The resulting confidence intervals only differed marginally and can be found in the appendix.

The OLS results in the last column confirm previous findings of the literature showing a clear association of teenage childbearing with lower educational attainment. Notably, the association is weakest in the teenage pregnancy sample indicating maybe self-selection of individuals with low educational prospects into early pregnancy.

Turning to the negative of age at menarche as an instrument, we can only derive upper bounds for the effect of teenage childbearing. Panel A shows the results for the complete sample, where the upper bound is estimated using 2SLS with the negative of age at menarche as an instrument. If one believes that this instrument is valid, then this is just the conventional point estimator of the causal effect of early childbearing. It is 0.218 when using no covariates and 0.146 when using covariates. These estimates of the upper bound of the effect are well below the OLS estimates as expected if the correlation between teenage childbearing and the structural error term is positive. However, they are very imprecisely estimated resulting in 95% confidence intervals for the true parameter that are not very informative. In all the other samples, a similar picture emerges.

Turning to miscarriage, we present estimates of the upper and lower bound using the

occurrence of a miscarriage or the transformed Z^* as instruments. Using the occurrence of a miscarriage as an instrument, it is possible to derive the lower bound for the causal effect of early childbearing. The upper bound is estimated using the transformed Z^* which is a linear combination of teenage childbirth and the occurrence of a miscarriage.

In panel A we again present the results for the complete sample where concerns about the validity of the instrument are most serious. In this sample, the estimated upper bound for the causal effect of early childbearing is well below the OLS estimate. Even considering the uncertainty associated with these estimates, the 95% confidence intervals using the occurrence of a miscarriage as instruments do not cover the corresponding OLS estimates.

Even if one uses the upper end of the confidence interval which is closest to OLS, one would find a smaller adverse effect of early childbearing on educational attainment. Notice that the estimates of the bounds are consistent with positive effects of early childbearing on educational attainment, and they are certainly consistent with a zero effect.

Turning to the sample of women who have ever been pregnant in panel B, we estimate a lower bounds for the effect of early childbearing of around zero. But the upper bound indicates in this sample a greatly increased risk of dropping out of high school for teenage mothers of comparable size to the OLS estimates. Turning to the sample of women who experienced a pregnancy as teenagers in panel C using the instruments leads to qualitatively different conclusions than the corresponding OLS estimates. Judging by the estimates of the lower bounds one cannot exclude the possibility that early childbearing is beneficial for educational achievement. And even when one is more conservative and looks only at the upper bounds one finds estimates of the upper bound that are well below the OLS estimates. In the sample of women who have children, a similar picture emerges as for the sample of ever pregnant women. Again, the upper bound of the estimate of the effect of early childbearing is quantitatively close to the OLS estimate.

In this particular case, age at menarche does not add much new information to sharpen the estimates because the associated standard errors are large. But somewhat similar to an overidentifying assumptions test, the presence of a second instrument is useful for a

specification test. Nevo and Rosen (2008) suggest inspecting the overlap of the set estimates using the different instruments. If there is no overlap the model would be misspecified. In all our cases, the set estimates overlap, so that we cannot reject the specification using this test.

5 Conclusion

In this paper we estimated bounds for the causal effect of early childbearing on educational attainment. We have found informative bounds for this parameter. The upper bound of the set estimate is well below the OLS estimate for all the instruments, therefore leading to different qualitative conclusions. The estimated bounds would even be consistent with no or even positive effects of early childbearing on educational attainment. Being conservative and only considering the upper bound of the effect one finds an adverse effect of early childbearing which is somewhat smaller than the OLS estimates suggest. However, especially when using age at menarche as an instrument, there is some estimation uncertainty associated with it. Combining both instruments one can tighten the set estimates even further because the upper bound estimated with age at menarche is smaller than the upper bound using the occurrence of a miscarriage as an instrument. However, one would still prefer the 95% confidence intervals only using the occurrence of a miscarriage as instrument because these bounds are more precisely estimated. But this overlap can also be seen as a specification test. If there was no overlap, then one would question whether the model is correctly specified. Constructing the overlap can also be seen as a way to reconcile different estimates when using alternative instruments. In our case, it is easy to reconcile the estimates because the confidence intervals for the parameters using each instrument individually overlap. Substantively, we find that the adverse effects of early childbearing are smaller than one would expect based on the usual OLS estimates confirming earlier empirical results. This lends more credibility to those results because one obtains them here under much weaker assumptions about the validity of instruments.

References

- Chernozhukov, Victor, Han Hong, and Elie Tamer (2007): “Parameter Set Inference in a Class of Econometric Models,” *Econometrica*, 75(5), 1243–1284.
- Freedman, David S., Laura Kettel Khan, Mary K. Serdula, William H. Dietz, Sathanur R. Srinivasan, and Gerald S. Berenson (2002): “Relation of Age at Menarche to Race, Time Period, and Anthropometric Dimensions: The Bogalusa Heart Study,” *Pediatrics*, 110(4), 1–7.
- Horowitz, Joel L. (2001): “The Bootstrap,” in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer. Amsterdam, North-Holland: Elsevier.
- Horowitz, Joel L., and Charles F. Manski (1998): “Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations,” *Journal of Econometrics*, 84, 37–58.
- Hotz, V. Joseph, Susan Williams McElroy, and Seth G. Sanders (1997): “The Costs and Consequences of Teenage Childbearing for Mothers and Government,” in *Kids Having Kids*, ed. by R. A. Maynard. Washington D.C.: Urban Institute Press.
- (1999): “Teenage Childbearing and its Life Cycle Consequences. Exploiting a Natural Experiment,” NBER Working Paper 7397.
- (2005): “Teenage Childbearing and its Life Cycle Consequences,” *Journal of Human Resources*, XL(3), 683–715.
- Imbens, Guido W., and Charles F. Manski (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72(6), 1845–1857.
- Nevo, Aviv, and Adam Rosen (2008): “Identification with Imperfect Instruments,” cemmap Working Paper 16/08.

Reinhold, Steffen (2007): “Essays in Demographic Economics,” Ph.D. thesis, Johns Hopkins University.

Ribar, David C. (1994): “Teenage Fertility and High School Completion,” *Review of Economics and Statistics*, 76(3), 413–424.

Weil, David N. (2007): “Accounting for the Effect of Health on Economic Growth,” *Quarterly Journal of Economics*, 122(3), 1265–1306.

Woutersen, Tiemen (2008): “A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters,” Mimeo. Johns Hopkins University.

Table 1: Summary Statistics

	Mean	Variance	$\sigma_{X,Z}$
Teenage childbearing	0.14	0.12	
-Age at menarche	-12.57	2.82	0.05
Occurrence of miscarriage	0.11	0.10	-0.01

Table 2: Identifying Assumptions

	ρ_{XU}	$\sigma_{X,Z}$	ρ_{ZU}
Teenage childbearing	+		
-Age at menarche	+	0.05	+
Occurrence of miscarriage	+	-0.01	+

Note: $\sigma_{X,Z}$ is estimated from data.

Table 3: Set Estimates and Bootstrapped 95% Confidence Intervals using Each Instrument Individually.

	Age at menarche		Miscarriage Bootstrapped CI		OLS
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
Panel A: Complete Sample. N=6443					
No covariates	$-\infty$	0.218	-0.501	0.310	0.368***
95% CI		0.464	-0.973	0.345	(0.017)
With covariates	$-\infty$	0.146	-0.428	0.239	0.291***
95% CI		0.439	-0.827	0.274	(0.017)
Panel B: Ever pregnant sample. N=4810					
No covariates	$-\infty$	0.108	0.052	0.306	0.338***
95% CI		0.347	-0.158	0.344	(0.018)
With covariates	$-\infty$	0.089	-0.039	0.229	0.263***
95% CI		0.366	-0.251	0.267	(0.018)
Panel C: Teenage pregnancy sample. N=1284					
No covariates	$-\infty$	-0.048	-0.117	0.156	0.253***
95% CI		2.172	-0.267	0.215	(0.028)
With covariates	$-\infty$	0.114	-0.115	0.094	0.174***
95% CI		2.244	-0.261	0.155	(0.029)
Panel D: Have kids sample. N=3910					
No covariates	$-\infty$	0.050	-0.036	0.286	0.320***
95% CI		0.307	-0.317	0.330	(0.020)
With covariates	$-\infty$	-0.096	-0.165	0.206	0.243***
95% CI		0.251	-0.487	0.249	(0.020)

Note: 5000 bootstrap replications were used.

6 Appendix

6.1 Construction of the upper bound when miscarriage is used as an instrument

Here we derive the upper bound as we do in the case of the occurrence of a miscarriage. Suppose we observe $\{X_i, Y_i, Z_i\}$ where $i = 1, \dots, N$. Let the realizations be identically and independently distributed. Also, let

$$Y_i = X_i\beta + \varepsilon_i, \text{ and} \quad (10)$$

$$\rho_{X\varepsilon} \geq \rho_{Z\varepsilon}$$

where β is a scalar. This yields

$$\frac{\text{Cov}(X, \varepsilon)}{\sigma_X \sigma_\varepsilon} \geq \frac{\text{Cov}(Z, \varepsilon)}{\sigma_Z \sigma_\varepsilon}$$

$$\frac{\text{Cov}(X, Y - X\beta)}{\sigma_X \sigma_\varepsilon} \geq \frac{\text{Cov}(Z, Y - X\beta)}{\sigma_Z \sigma_\varepsilon}$$

so that

$$\frac{\text{Cov}(X, Y) - \beta \cdot \text{Var}(X)}{\sigma_X} \geq \frac{\text{Cov}(Z, Y) - \beta \cdot \text{Cov}(Z, X)}{\sigma_Z}$$

and

$$\beta \cdot \left\{ \text{Cov}(Z, X) \frac{\sigma_X}{\sigma_Z} - \text{Var}(X) \right\} \geq \left\{ \text{Cov}(Z, Y) \frac{\sigma_X}{\sigma_Z} - \text{Cov}(X, Y) \right\}.$$

Thus,

$$\beta \leq \frac{\text{Cov}(Z, Y) \frac{\sigma_X}{\sigma_Z} - \text{Cov}(X, Y)}{\text{Cov}(Z, X) \frac{\sigma_X}{\sigma_Z} - \text{Var}(X)}.$$

since $\text{Var}(X) - \text{Cov}(Z, X) \frac{\sigma_X}{\sigma_Z} = \sigma_X (\sigma_X - \text{Cov}(Z, X)/\sigma_Z) > 0$.

Replacing

$$\frac{\text{Cov}(Z, Y) \frac{\sigma_X}{\sigma_Z} - \text{Cov}(X, Y)}{\text{Cov}(Z, X) \frac{\sigma_X}{\sigma_Z} - \text{Var}(X)}$$

with its empirical counterparts one obtains an upper bound. See Adam and Rosen (2008) for a more general version of this derivation.

6.2 Construction of the lower bound when miscarriage is used as an instrument

Here we derive the lower bound when using miscarriage as an instrument. The probability limit of the standard IV estimator is given by

$$\beta_Z^{IV} = \beta + \frac{\sigma_{ZU}}{\sigma_{XZ}}$$

The expression $\frac{\sigma_{ZU}}{\sigma_{XZ}}$ is negative because $\sigma_{XZ} < 0$ and $\sigma_{ZU} > 0$. Hence, β_Z^{IV} provides a lower bound when using the occurrence of a miscarriage as an instrument.

6.3 Construction of the upper bound when age at menarche is used as an instrument

Here we derive the upper bound when using the negative of age at menarche as an instrument.

The probability limit of the standard IV estimator is given by

$$\beta_Z^{IV} = \beta + \frac{\sigma_{ZU}}{\sigma_{XZ}}$$

The expression $\frac{\sigma_{ZU}}{\sigma_{XZ}}$ is positive because both the denominator and the numerator are positive.

Hence the conventional IV estimator provides an upper bound.

6.4 Imbens and Manski CI, not intended for publication

Table 4: Set Estimates and 95% Confidence Intervals using Each Instrument Individually.

	Age at menarche		Miscarriage		OLS
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
	Imbens & Manski CI				
Panel A: Complete Sample. N=6443					
No covariates	$-\infty$	0.218	-0.501	0.310	0.368***
95% CI		0.466	-0.896	0.346	(0.017)
With covariates	$-\infty$	0.146	-0.428	0.239	0.291***
95% CI		0.434	-0.775	0.275	(0.017)
Panel B: Ever pregnant sample. N=4810					
No covariates	$-\infty$	0.108	0.052	0.306	0.338***
95% CI		0.344	-0.151	0.343	(0.018)
With covariates	$-\infty$	0.089	-0.039	0.229	0.263***
95% CI		0.359	-0.240	0.266	(0.018)
Panel C: Teenage pregnancy sample. N=1284					
No covariates	$-\infty$	-0.048	-0.117	0.156	0.253***
95% CI		1.209	-0.261	0.216	(0.028)
With covariates	$-\infty$	0.114	-0.115	0.094	0.174***
95% CI		1.204	-0.255	0.155	(0.029)
Panel D: Have kids sample. N=3910					
No covariates	$-\infty$	0.050	-0.036	0.286	0.320***
95% CI		0.315	-0.312	0.327	(0.020)
With covariates	$-\infty$	-0.096	-0.165	0.206	0.243***
95% CI		0.258	-0.467	0.248	(0.020)

Note: Confidence intervals based on robust standard errors.

6.5 Woutersen CI, not intended for publication

Table 5: Set Estimates and 95% Confidence Intervals using Each Instrument Individually.

	Age at menarche		Miscarriage Woutersen CI		OLS
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
Panel A: Complete Sample. N=6443					
No covariates	$-\infty$	0.218	-0.501	0.310	0.368***
95% CI		0.466	-0.896	0.346	(0.017)
With covariates	$-\infty$	0.146	-0.428	0.239	0.291***
95% CI		0.434	-0.775	0.275	(0.017)
Panel B: Ever pregnant sample. N=4810					
No covariates	$-\infty$	0.108	0.052	0.306	0.338***
95% CI		0.344	-0.151	0.345	(0.018)
With covariates	$-\infty$	0.089	-0.039	0.229	0.263***
95% CI		0.359	-0.240	0.267	(0.018)
Panel C: Teenage pregnancy sample. N=1284					
No covariates	$-\infty$	-0.048	-0.117	0.156	0.253***
95% CI		1.209	-0.261	0.216	(0.028)
With covariates	$-\infty$	0.114	-0.115	0.094	0.174***
95% CI		1.204	-0.255	0.155	(0.029)
Panel D: Have kids sample. N=3910					
No covariates	$-\infty$	0.050	-0.036	0.286	0.320***
95% CI		0.315	-0.311	0.331	(0.020)
With covariates	$-\infty$	-0.096	-0.165	0.206	0.243***
95% CI		0.258	-0.466	0.251	(0.020)

Note: Confidence intervals based on robust standard errors.