DISCUSSION PAPER SERIES

IZA DP No. 4689

Instrumental Variables Estimation with Partially Missing Instruments

Magne Mogstad
Matthew Wiswall

January 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Instrumental Variables Estimation with Partially Missing Instruments

Magne Mogstad
Statistics Norway
and IZA

Matthew Wiswall
New York University

Discussion Paper No. 4689
January 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4689
January 2010

ABSTRACT

Instrumental Variables Estimation with Partially Missing Instruments*

We examine instrumental variables estimation in situations where the instrument is only observed for a sub-sample, which is fairly common in empirical research. Typically, researchers simply limit the analysis to the sub-sample where the instrument is non-missing. We show that when the instrument is non-randomly missing, standard IV estimators require strong, auxiliary assumptions to be consistent. In many (quasi)natural experiments, the auxiliary assumptions are unlikely to hold. We therefore introduce alternative IV estimators that are robust to non-randomly missing instruments without auxiliary assumptions. A Monte-Carlo study illustrates our results.

JEL Classification: C31, C34

Keywords: instrumental variables, partially missing instruments, sample selection, sub-sample estimation

Corresponding author:

Magne Mogstad
Research Dept. of Statistics Norway
Pb 8131 Dep
0033 Oslo
Norway
E-mail: magne.mogstad@ssb.no

* Thanks to Rolf Aaberge, Erik Biørn, Stephane Bonhomme, Christian Brinch, Bryan Graham, Konrad Menzel, and Arvid Raknerud for helpful comments. The Norwegian Research Council has provided financial support for this project.

1 Introduction

Instrumental variables (IV) estimation is a major workhorse in applied economics and is gaining ground in other sciences as well. In some applications, the instrument is only observed for some sub-sample and missing for the remaining sample, and there are reasons to suspect that the instrument is not missing at random. To fix ideas, consider estimating the return to schooling using receipt of a college tuition subsidy as an instrument for schooling. Suppose the tuition subsidy is allocated randomly among all individuals who apply, conditional on some observable characteristics. However, given some cost of applying for the subsidy, only a subsample of potential college students apply. In this case, the instrument (receipt of tuition subsidy) satisfies the standard IV assumption of conditional independence within the sub-sample of applicants, since it is randomly allocated conditional on observed characteristics. But the instrument is missing or undefined for the subsample of individuals who did not apply. The instrument is missing non-randomly if applicants for tuition subsidy have a different level of unobservables affecting earnings, such as innate ability, compared to non-applicants. More generally, instruments available only for choice-based sub-samples have this type of missing instrument structure.

This paper examines IV estimation with partially missing instruments. Our contributions are to characterize the sources to inconsistency in standard IV estimators when the instruments are missing non-randomly, and propose alternative IV estimators that are robust to non-randomly missing instruments.[2] Commonly, researchers simply limit the analysis to the sub-sample where the instrument is non-missing. We show that when the instrument is non-randomly missing, standard IV estimators applied to the sub-sample where the instrument is non-missing require strong, auxiliary assumptions to yield consistent estimates. We also show that this is true for the pooled IV approach proposed by Angrist et al. (2009). They construct an instrument for the full sample from a linear projection of the partially observed instrument on the covariates in the sub-sample where the instrument is non-missing. That auxiliary assumptions are required even under otherwise ideal conditions for IV estimation – strong instruments and homogeneous treatment effects – makes partially missing instruments particularly worrisome.

In this paper, we make clear the auxiliary assumptions required for IV estima-

[2] Our focus is on missing *instruments*; not missing outcomes, nor missing covariates – on which there is a large literature in statistics and econometrics (see e.g. Little and Rubin, 2002). In addition, we are not discussing situations where the model is under-identified, i.e. insufficient instruments to identify parameters of interest.

tors to yield consistent estimates when the instrument is non-randomly missing. There are two main auxiliary assumptions, either of which is sufficient for consistent estimation: (i) the regression error is mean-independent of the covariates in the sub-sample where the instrument is non-missing, or (ii) the conditional expectation of the instrument is linear in the covariates.

As discussed in detail below, in many (quasi)natural experiments, neither of these auxiliary assumptions are likely to hold. The standard assumption of the regression error being mean-independent of the covariates in the full-sample, does not imply that auxiliary assumption (i) holds. In fact, not even full-independence in the full-sample implies mean-independence in the sub-sample. On the contrary, in the general case where selection into the sub-sample with non-missing instruments depends on unobservables, such as if high ability students are more likely to apply for aid, auxiliary assumption (i) is violated.

In addition, auxiliary assumption (ii) is too strong in many IV applications, where the relationship between the instrument and the covariates is unlikely to be linear. A leading example is with a binary instrument (e.g. receipt of a tuition subsidy), where the relationship is generally non-linear. It is important to emphasize that for the issue of partially missing instruments, the relationship of primary interest is that between the instrument and the included covariates; not the relationship between the endogenous regressor and the instrument, nor the relationship between the outcome of interest and the endogenous regressor. As is well known, consistency of IV estimators does not require the correct specification of the relationship between the dependent variable and the endogenous regressor in the first stage (Kelejian 1971). Furthermore, using a linear model to describe the relationship between the outcome and the endogenous regressor can be justified even when the outcome is binary or the endogenous regressor takes on multiple values.[3] By contrast, when the instrument is non-randomly missing, the consistency of the IV estimator may depend on the specification of the relationship between the instrument and the included covariates.

To avoid having to impose any auxiliary assumptions, we introduce two alternative IV estimators that are robust to non-randomly missing instruments. The first estimator uses only the sub-sample where the instrument is non-missing. The second estimator generalizes the pooled IV approach to construct a full-sample instrument using non-parametric regression. This second estimator preserves the

[3]See Angrist (2001) for a discussion linear IV estimation with a binary outcome variable. Angrist and Imbens (1995) discuss linear IV estimation with an endogenous regressor that takes on multiple values, so-called variable treatment intensity.

efficiency gain from pooling samples (both the sample with and without missing instruments), as in Angrist et al (2009), but does not require any auxiliary assumptions to yield consistent estimates. In a Monte Carlo study, we demonstrate that both of these IV estimators are robust to non-randomly missing instruments without auxiliary assumptions, in stark contrast to the standard IV estimators.

The next section describes the main model and assumptions, before discussing a number of well-known IV applications where the instruments are partially missing. Section 3 examines standard IV estimation using only the sub-sample where the instrument is non-missing. Section 4 examines consistency of the full-sample instrument proposed by Angrist et al. (2009). Section 5 describes the proposed IV estimators that are robust to non-randomly missing instruments without auxiliary assumptions. Section 6 illustrates our results using a Monte-Carlo study, before Section 7 concludes.

2 Model

Consider a simple, constant-effect regression model in which an individual's outcome Y depends linearly on a scalar regressor of interest S and a vector of covariates $X = [1, x_2, \ldots, x_K]'$:

$$Y = \beta S + X'\delta + \epsilon, \tag{1}$$

where ϵ is the regression error, which is normalized $E[\epsilon] = 0$, without loss of generality. To focus on the missing instrument issue, we assume that the marginal effects are constant and homogeneous. This implies that β is the same in any subsample.[4] To make the model non-trivially different from a bivariate model, we assume that at least one of the included covariates are correlated with the regressor of interest conditional on the other covariates, that is, $Cov(x_k, S|X_{-k}) \neq 0$ for at least one $k = 1, 2, ..., K$.

Assumption 1 states that the the regression error in equation (1) is mean-independent of the covariates, so that S is the only potentially endogenous variable in the full sample. Throughout this paper, we will assume that Assumption 1 holds.

Assumption 1 *Full-sample exogeneity:* $E[\epsilon|X] = E[\epsilon]$

[4] See e.g. Imbens and Angrist (1994), Angrist and Imbens (1995), Heckman et al. (2006), and Mogstad and Wiswall (2009) for a discussion of IV estimation in the presence of heterogeneous treatment effects and variable treatment intensity, when the instrument is fully observed.

Given Assumption 1, OLS on equation (1) produces consistent estimates of the parameters $[\delta, \beta]$ whenever $Cov(\epsilon, S|X) = 0$. The motivation for IV estimation is that $Cov(\epsilon, S|X) \neq 0$.

Our point of departure is that the instrument is only observed for some sub-sample, but missing for the remaining sample. For simplicity, suppose that there are two sub-samples, denoted by $R = \{0, 1\}$, and that there exists a scalar instrument Z, which is undefined or missing for the sub-sample with $R = 0$ but observed for the sub-sample with $R = 1$. We consider estimation of the parameter β of the endogenous variable S, under standard IV assumptions imposed on the sub-sample where the instrument is non-missing, $R = 1$:

Assumption 2 *First stage exists:* $E[SZ|X, R = 1] \neq 0$

Assumption 3 *Cond. mean indep.:* $E[\epsilon|Z, X, R = 1] = E[\epsilon|X, R = 1]$

These assumptions imply that conditional on the covariates, the instrument is correlated with the endogenous regressor and, further, that the regression error is mean-independent of the instrument, within the sub-sample where the instrument is non-missing. In the special case where the instrument is observed for the full sample, $pr(R = 1) = 1$, Assumptions 2 and 3 become the standard IV assumptions. Given that the instrument is missing for the sub-sample with $R = 0$, it is not meaningful to impose Assumptions 2 and 3 for the full-sample. Instead, we follow previous studies performing IV estimation with partially observed instruments in imposing Assumptions 2 and 3 on the sub-sample where the instrument is non-missing.

As stated in Definition 1, the instrument is *missing at random* if the sub-sample where the instrument is non-missing has the same average level of unobservables affecting the outcome of interest, compared to the sub-sample with missing instruments.

Definition 1 *Instrument missing at random:* $E[\epsilon|R = 1, X] = E[\epsilon|X]$

Note that the degree of selection into the sub-sample where the instrument is non-missing is given by $E[\epsilon|R = 1, X]$, since $E[\epsilon|X] = E[\epsilon]$ under Assumption 1 and we normalize $E[\epsilon] = 0$.

2.1 Partially Missing Instruments: Some Examples

Below, we show that partially missing instruments are actually fairly common in empirical research, in part because of the nature of the instruments used but also because of data availability. And further, it is quite likely that the partially missing instruments are missing non-randomly.

To make clear the distinction between fully and partially observed instruments, consider the vast literature using instruments to estimate the returns to schooling (for a review, see e.g. Card, 2001). In this literature, interest is in the relationship between wages (Y) and years of schooling (S), conditional on some covariates (X). A major concern is that years of schooling is endogenous to potential wages. Instruments derived from compulsory schooling laws (see e.g. Angrist and Krueger, 1991) and college proximity (see e.g. Card, 1993) are full-sample instruments, since an individual's quarter of birth and location of residence are in principle available for the full sample. However, other instruments for schooling are partially missing.

An common partially missing instrument for schooling is tuition subsidy, which was first used by Kane and Rouse (1993) in a study of the return to college. This instrument is partially missing since receipt of college subsidy is only observed for applicants, and missing for non-applicants. In general, the take-up rate of tuition subsidies targeted toward students from low income families is quite low. Many people eligible for support do not apply for it. Carneiro and Heckman (2002) discuss two possible explanations. First, the non-monetary costs of applying for financial aid are high, especially for low income people because the application process is complex. Second, many eligible persons perceive that even with a substantial tuition subsidy, the returns to college education for them are too low to pay for the foregone earnings required to attend school. Both explanations suggest that the tuition subsidy instrument is likely to be missing for a sizable subgroup and, further, that it is missing non-randomly since applicants for college subsidy can be expected to have substantially different potential wages compared to non-applicants.

Even instruments that are in theory available for everyone may in practice be partially missing because of data availability. For example, the study of returns to education by Aakvik et al. (2009) use the staged implementation of a major reform in the comprehensive school system Norway as an instrument for schooling. This instrument is partially missing, since there is no information about the implementation of the reform for about a quarter of all municipalities in Norway. The instrument is missing non-randomly if potential wages differ systematically

depending on municipality of residency.[5]

Another example of partially missing instruments is from the large and growing literature investigating the effects of family size (S) on child outcome (Y). In this literature, the commonly used instrument – twin on second (or higher) birth– is missing for the sub-sample of children whose parents choose to only have one child. The motivation for using the twin instrument is that it is thought of as an exogenous shock to family size, conditional on some covariates like the mother's age.[6]

In a similar vein, the literature investigating the relationship between family size (S) and maternal labor supply (Y), faces the problem of partially missing instruments. For example, Angrist and Evans (1998) uses mixed gender sibship as an instrument for family size. This instrument is missing for the sub-sample of mothers who choose to only have one child. The twin and mixed gender sibship instruments are missing non-randomly, insofar family size is endogenous; if exogenous, there would be no need to search for an instrument in the first place.[7]

The problem of partially missing instruments also exists in the literature looking at the relationship between institutions on economic performance. For example, Acemoglu et al. (2001) investigate the effect of property rights (S), as a proxy for institutions, on per capita income (Y). To address the concern for selection bias, they use mortality rates among early European settlers in different colonies to instrument for property rights. Their theory is that property rights are affected by settler mortality through the institutions brought by Europeans to their colonies. However, the instrument is only observed for countries that were colonized, and missing for Non-European countries never colonized and, of course, for European countries themselves. The instrument is missing non-randomly if colonized countries have a different level of unobservables affecting income per

[5]The same partially missing instrument is used to study the intergenerational transmission of human capital in Black et al. (2005b).

[6]Rosenzweig and Wolpin (1980) are the first to use twin birth as an instrument to estimate the effect of family size on child outcome. More recently, Black et al. (2005), Caceres-Delpiano (2006), and Mogstad and Wiswall (2009) have used twin birth to examine the effects of family size on child outcome. To avoid including the outcomes of twins themselves, these studies restrict the sample to children with at least one sibling.

[7]Partially missing instruments are also used in the literature looking at the effect of marital dissolution on the economic wellbeing of women. Both Bedard and Deschenes (2005) and Ananat and Michaels (2008) use the gender of the first born as an instrument for divorce. The idea is that families have preferences for sons over daughters. This instrument is missing for couples without children, and missing non-randomly if their potential divorce probability is different from couples with children.

capita than the rest of the world.

In such applications, where instruments are partially missing, there are two options. In all the above examples, researchers simply limit the analysis to the sub-sample where the instruments are non-missing. For example, Black et al. (2005) estimate the effect of family size on the educational attainment of first born from families with two or more children, using twin at second birth as instrument for family size. And Acemoglu et al. (2001) estimate the impact of property rights on per capita income, restricting the sample to colonized countries for which settler mortality is observed. The second option is to form an instrument that is defined for the full sample. In a study of family size and child outcome, Angrist et al. (2009) propose a full-sample instrument based on a linear projection of the instrument on the covariates in the sub-sample where it is non-missing. Their motivation for constructing such full-sample instruments is that pooling samples in this way may generate efficiency gains. Below we consider both ways of dealing with missing instruments.

3 Sub-Sample IV Estimation

In this section, we apply standard IV methods to the sub-sample with non-missing instrument, and do not use any of the data from the sub-sample with missing instruments. As a benchmark, we first briefly summarize IV estimation when the instrument is fully observed, before turning our attention to the case where the instrument is partially missing.

3.1 Fully Observed Instrument

When the instrument is fully observed, $pr(R = 1) = 1$, the standard moment based IV estimator is given by the sample analog of:

$$[\beta(IV), \delta(IV)] = E[Q'W]^{-1}E[Q'Y], \tag{2}$$

where $Q = [Z, X]$ and $W = [S, X]$.[8] As is well known, a numerically equivalent IV estimator for β can be formed by first projecting Z on X and forming the residual. The residual for Z is $Z^\star = Z - X'\psi$, where the vector of linear projection

[8]The moment based estimator is implemented in commonly used statistical packages, such as STATA's IVREGRESS routine. Since our model is exactly identified, the moment based IV estimator is equivalent to the two-stage least-squares estimator.

coefficients in the full-sample is $\psi = E[X'X]^{-1}E[X'Z]$. The IV estimator for β can then be expressed as the sample analog of:

$$\beta(IV) = E[Z^\star S]^{-1}E[Z^\star Y]. \tag{3}$$

Substituting (1), we have

$$\beta(IV) = \beta + E[Z^\star S]^{-1}E[Z^\star \epsilon]$$

Under Assumptions 1-3, the sample analog of $\beta(IV)$ yields a consistent estimate of β since $E[Z^\star\epsilon] = E_X\{E[Z^\star\epsilon|X]\}$ and for all X,

$$E[Z^\star\epsilon|X] = E[Z\epsilon|X] - X'\psi E[\epsilon|X] = E_{Z|X}ZE[\epsilon|Z,X] = 0.$$

3.2 Partially Missing Instrument

Next, consider the case where the instrument is partially missing, $pr(R = 1) \in (0, 1)$. The standard moment based IV estimator applied to the sub-sample with non-missing instruments, $R = 1$, is given by the sample analog to:

$$[\beta(IV, 1), \delta(IV, 1)] = E[Q'W|R = 1]^{-1}E[Q'Y|R = 1]. \tag{4}$$

As above, a numerically equivalent IV estimator for β can be formed by first projecting Z on X in the sub-sample with $R = 1$. The vector of coefficient from this linear projection for Z is $\psi_1 = E[X'X|R = 1]^{-1}E[X'Z|R = 1]$, from which we can form the residual $Z^* = Z - X'\psi_1$ in the sub-sample with $R = 1$. The IV estimator for β can then be expressed as the sample analog of:

$$\beta(IV, 1) = E[Z^*S|R = 1]^{-1}E[Z^*Y|R = 1]. \tag{5}$$

Proposition 1 shows that the inconsistency in the standard IV estimator applied to the sub-sample where the instrument is non-missing is the product of two terms. The first term capture the degree of selection into the sub-sample where the instrument is non-missing, given by $E[\epsilon|R = 1, X]$. And the second term reflects the difference between the linear projection of Z on X, $X'\psi_1$, and the conditional expectation function between the instrument and the included covariates, $E[Z|X, R = 1]$. Note that $E[Z|X, R = 1]$ is in general an unknown and possibly non-linear function of the X variables.

Proposition 1 *Under Assumptions 1- 3,* $\beta(IV,1) = \beta$ *if and only if* $E_{X|R=1}\{E[\epsilon|R=1,X](E[Z|R=1,X] - X'\psi_1)\} = 0$.

Proof Substituting (1) into (5) we get

$$\beta(IV,1) = \beta + E[Z^*S|R=1]^{-1}E[Z^*\epsilon|R=1].$$

Under Assumptions 1 and 2, $\beta(IV,1) = \beta$ if and only if $E[Z^*\epsilon|R=1] = 0$. By iterating expectations, we can write

$$E[Z^*\epsilon|R=1] = E_{X|R=1}\{E[Z\epsilon|R=1,X] - X'\psi_1 E[\epsilon|R=1,X]\}$$

Assumption 3 implies $Cov(Z,\epsilon|R=1,X) = 0$ and therefore

$$E[Z\epsilon|R=1,X] = E[Z|R=1,X]E[\epsilon|R=1,X].$$

Substituting this expression, we have

$$E[Z^*\epsilon|R=1] = E_{X|R=1}\{E[\epsilon|R=1,X](E[Z|R=1,X] - X'\psi_1)\},$$

which gives us

$$\beta(IV,1) = \beta + E[Z^*S|R=1]^{-1}E_X\{E[\epsilon|R=1,X](E[Z|R=1,X] - X'\psi_1)\},$$

where Assumption 2 implies that $E[Z^*S|R=1] \neq 0$.

QED

Proposition 1 raises the question of which auxiliary assumptions ensure consistent estimates from the standard IV estimator applied to the sub-group where the instrument is non-missing? Corollary 1 states that the sample analog of $\beta(IV,1)$ consistently estimates β if at least one of the following conditions holds: (i) the instrument is missing at random conditional on covariates, (ii) the covariates are mean-independent of regression error in the non-missing sub-sample, or (iii) the conditional expectation of the instrument is linear in the covariates.

Corollary 1 *$\beta(IV,1)=\beta$ if at least one of the following auxiliary assumptions are imposed:*
(i) $E[\epsilon|R=1,X]=E[\epsilon|X]$ *for all* $X=x$
(ii) $E[\epsilon|R=1,X]=E[\epsilon|R=1]$ *for all* $X=x$
(iii) $E[Z|R=1,X]=X'\psi_1$

Proof With auxiliary assumptions (i) and (iii), $\beta(IV,1)=\beta$ follows immediately from Proposition 1. Under auxiliary assumption (ii), we have

$$\begin{aligned}&E_{X|R=1}\{E[\epsilon|R=1,X](E[Z|R=1,X]-X'\psi_1)\}\\&=E[\epsilon|R=1]E_{X|R=1}\{E[Z|R=1,X]-X'\psi_1\}=0,\end{aligned}$$

since

$$E_{X|R=1}\{E[Z|R=1,X]\}=E[Z|R=1]=E_{X|R=1}\{X'\psi_1\}.$$

QED

Assuming that the instrument is randomly missing, that is, imposing auxiliary assumption (i) in Corollary 1, implies that there is no selection on unobservables into a particular sub-sample.[9] In the family size application, for instance, this assumption assumes away the very reason for instrumenting, namely that fertility is endogenous: If only children have the same potential outcome as children with siblings, there is no need to instrument for family size.

Auxiliary assumption (ii) is violated when the regression error is not mean-independent of the covariates within the sub-sample where the instrument is non-missing. It should be emphasized that the standard assumption of the regression error being mean-independent of the covariates in the full-sample, Assumption 1, does not imply that the covariates are mean-independent in either of the two sub-samples. In fact, not even full-independence in the full-sample implies mean-independence in either of the two sub-samples.[10]

On the contrary, if the instrument is non-randomly missing then it is likely that the distribution of covariates differ between the full-sample and the sub-sample

[9] Assuming that the instrument is missing at random is analog to the much used assumption of outcomes or regressors missing at random (Little and Rubin, 2002), which has received criticism for being far too strong (see e.g. Frangakis and Rubin, 1999).

[10] And conversely, full-independence in a particular sub-sample does not imply mean-independence in the full sample.

where the instrument is non-missing. This occurs in situations where the selection into the sub-sample depends on covariates, such as when $R = 1\{\gamma_0\epsilon + X'\gamma_1 > 0\}$. For example, the probability of having another child is known to decrease with mother's age, and the probability of applying for tuition subsidy may fall with parental wealth. If the instrument is missing non-randomly, $E[\epsilon|X, R = 1]$ will generally not be equal to $E[\epsilon|R = 1]$, even if ϵ and X are fully independent in the full-sample. Another possibility is with conditional heteroscedasticity in the full-sample, that is, $\epsilon \sim N(0, \sigma(X)^2)$ so that $E[\epsilon|X] = 0$ and $Var[\epsilon|X] = \sigma(X)^2$. Then, $E[\epsilon|R = 1, X] \neq 0$ whenever the instrument is non-randomly missing, implying that $E[\epsilon|R = 1, X] \neq E[\epsilon|X]$. Consequently, the regression error is mean-independent of the covariates in the full-sample, but not in sub-samples.

Auxiliary assumption (iii) from Corollary 1 is violated when the the conditional expectation of the instrument $E[Z|X, R = 1]$ is non-linear in X. The result that consistency of the standard IV estimator when the instrument is non-randomly missing may depend on the specification of the relationship between the instrument and the covariates is worrisome, especially since researchers typically pay little attention to this relationship. For instance, we know that $E[Z|X, R = 1]$ is generally non-linear if the instrument is binary, such as with twin births or a dummy variable for receipt of a tuition subsidy.[11] Furthermore, in the family size application, it is well known that the probability of a women having twin birth increases with her age, at an increasing rate. And in the returns to schooling example, the receipt of means-tested tuition subsidies may very well be related non-linearly to parental wealth.

An interesting special case where auxiliary assumption (iii) in Corollary 1 holds, namely when the instrument is mean-independent of the covariates, $E[Z|X, R = 1] = E[Z|R = 1]$. This implies that $E[Z|X, R = 1] = X'\psi_1 = \mu_z$ where μ_z is the mean of the instrument. Hence, the conditional expectation function between the instrument and the included covariates is constant, and $\beta(IV, 1) = \beta$. However, the condition of the instrument being mean-independent of the covariates is generally quite strong, except in the case where Z is truly randomly assigned as in a controlled laboratory experiment. In many (quasi)natural experiments, like the examples discussed above, the instrument is not mean-independent of the covariates.

Finally, it should be noted that the auxiliary assumptions in Corollary 1 are sufficient, but not necessary. From Proposition 1, it is clear that $\beta(IV, 1) = \beta$

[11] If X is a dummy variable, then $E[Z|X]$ is necessarily linear in X. However, in general $E[Z|X]$ is not linear in X.

when non-linearities cancel each other out, such that the weighted average of the difference between $E[Z|X, R = 1]$ and $X'\psi_1$ becomes zero. Although this may happen by chance, it seems to be a too strong auxiliary assumption, and it is not testable since the weights $E[\epsilon|R = 1, X]$ are unobserved.

4 Full-Sample Instrument

An alternative to limiting the analysis to the sub-sample where the instrument is non-missing, is to form an instrument that is defined for the full sample. A motivation for constructing such full-sample instruments is that pooling samples in this way may generate efficiency gains. Angrist et al. (2009) propose such a full-sample instrument using a linear projection of the partially observed instrument on the covariates in the sub-sample with non-missing instrument. Their linear projection instrument can be expressed as

$$Z_{LP} = \begin{cases} 0 & \text{if } R = 0 \\ Z - X'\psi_1 & \text{if } R = 1 \end{cases} \tag{6}$$

where ψ_1 is the vector of coefficients from the regression of Z on X in sub-sample with $R = 1$, as defined above. Using the instrument Z_{LP}, the moment based IV estimator applied to the full-sample is given by the the sample analog of:

$$[\beta(IV - LP), \delta(IV - LP)] = E[Q'_{LP}Y]E[Q'_{LP}W]^{-1} \tag{7}$$

where $Q_{LP} = [Z_{LP}, X]$ and $W = [S, X]$.[12] As above, a numerically equivalent IV estimator for β can be formed by first projecting Z_{LP} on X. The vector of coefficient from this linear projection is $\phi = E[X'X]^{-1}E[X'Z_{LP}]$, from which we can form the residual $Z^*_{LP} = Z_{LP} - X'\phi$. The moment based IV estimator applied to the full-sample can then be expressed as the sample analog of:

[12] Another way to view the linear projection instrument is as the following linear imputation method:

$$Z_{LI} = \begin{cases} X'\psi_1 & \text{if } R = 0 \\ Z & \text{if } R = 1, \end{cases}$$

where the instrument Z_{LI} is constructed by imputing the missing instrument for the $R = 0$ sub-sample using the observed covariates X. The imputation uses the linear predictor $X'\psi_1$, constructed using the covariates X observed in both sub-samples, and the ψ_1 coefficients from the regression of the instrument on the covariates in the $R = 1$ sub-sample. It is straightforward to show that using Z_{LI} is equivalent to using Z_{LP} as the instrument in the moment based IV estimator applied to the full-sample.

$$\beta(IV - LP) = E[Z^*_{LP}S]^{-1}E[Z^*_{LP}Y],$$

For a full sample with N observations, the sample analog estimators are defined by

$$\hat{\beta}(IV - LP) = \sum_{i=1}^{N}(Z^*_{LP,i}S_i)^{-1}\sum_{i=1}^{N}(Z^*_{LP,i}Y_i)$$

$$\hat{\beta}(IV, 1) = \sum_{i=1}^{N_1}(Z^*_i S_i)^{-1}\sum_{i=1}^{N_1}(Z^*_i Y_i)$$

where, without loss of generality, we have ordered the observations for the $R = 1$ sub-sample from $i = 1, \ldots, N_1$ and the remaining observations from $N_1 + 1, \ldots, N$. Lemma 1 states that applying the moment based IV estimator to the full-sample based on the linear projection instrument is equivalent to using the moment based IV estimator based on the partially observed instrument in the sub-sample where the instrument is non-missing. Hence, both estimators provide consistent estimates of β under the auxiliary assumptions stated in Corollary 1.[13]

Lemma 1 $\hat{\beta}(IV - LP) = \hat{\beta}(IV, 1)$.

Proof Because the linear projection coefficient $\phi = E[X'X]^{-1}E[X'Z_{LP}] = 0$, $Z^*_{LP} = Z_{LP}$. Since $Z_{LP} = R(Z - X'\psi_1) = RZ^*$, it follows that:

$$\hat{\beta}(IV - LP) = \{\sum_{i=1}^{N_1} Z^*_i S_i + \sum_{i=N_1+1}^{N} 0S_i\}^{-1}\{\sum_{i=1}^{N_1} Z^*_i Y_i + \sum_{i=N_1+1}^{N} 0Y_i\}$$

$$= \hat{\beta}(IV, 1)$$

QED

[13]It should be noted that Angrist et al. (2009) in their Lemma providing the econometric justification for their IV strategy, only assume that conditional on the covariates, the instrument is correlated with the endogenous regressor and that the regression error is independent of the instrument, within the sub-sample where the instrument is non-missing (our Assumptions 1-3).

5 Robust IV Estimators

This section introduces two alternative IV estimators that are robust to non-randomly missing instruments without auxiliary assumptions.

The first robust IV estimator uses only the sub-sample where the instrument is non-missing. Let the support of X be denoted as $\mathbf{X}$. Conditioning on some realization of $X = x$, we can form a consistent estimator of β for each $x \in \mathbf{X}$ from the sample analog of:

$$\beta_x = \frac{Cov(Y, Z|X = x, R = 1)}{Cov(S, Z|X = x, R = 1)},$$

Given a vector of weights over $\mathbf{X}$, $W = \{W(x)\}_{x \in \mathbf{X}}$, we can then form a weighted average of each of β_x as

$$\beta(W, 1) = \sum_{x \in \mathbf{X}} \beta_x W(x), \tag{8}$$

where the weights $W(x)$ are positive for all $x \in \mathbf{X}$ and sum to 1. The weights could be chosen optimally to minimize the variance of the estimator. As the sample analog of $\beta(W, 1)$ is a non-parametric estimator of β, it is clear that $\beta(W, 1) = \beta$ under Assumptions 1-3. The simulation exercises below provides an example of how to estimate $\beta(W, 1)$.

The second robust IV estimator generalizes the linear projection method to construct a full-sample instrument using non-parametric regression. This full-sample instrument is defined as:

$$Z_{NP} = \begin{cases} 0 & \text{if } R = 0 \\ Z - E[Z|X, R = 1] & \text{if } R = 1 \end{cases} \tag{9}$$

where $E[Z|X, R = 1]$ is estimated using non-parametric methods in the sub-sample with non-missing instrument. The advantage of this instrument over the linear projection instrument, is that Z_{NP} does not restrict the relationship between Z and X to be linear. Using the instrument Z_{NP}, the moment based IV estimator applied to the full-sample is given by the the sample analog of:

$$\beta(IV - NP), \delta(IV - NP)] = E[Q'_{NP}Y]E[Q'_{NP}W]^{-1} \tag{10}$$

where $Q_{NP} = [Z_{NP}, X]$ and $W = [S, X]$. As the sample analog of $\beta(IV - NP)$ is a non-parametric estimator of β, it is clear that $\beta(IV - NP) = \beta$ under Assumptions 1-3. The simulation exercises below provides an example of how to

estimate $\beta(IV - NP)$.

6 Simulation

This section uses a Monte-Carlo simulation to illustrate our results. To fix ideas, let us return to the example where college tuition subsidy (Z) is allocated randomly among applicants conditional on parent's wealth (X). But, as argued above, the instrument is only observed for applicants, $R = 1$, and missing for individuals who did not apply, $R = 0$. An individual's wage Y is specified as a linear function of a binary indicator for college attendance S and a scalar covariate X.

Specifically, we specify the following data generating process:

$$Y = \alpha + \delta X + \beta S + \epsilon, \qquad X \sim N(0,1), \epsilon \sim N(0,1), \alpha = 1, \delta = 1, \beta = 1$$

$$S = 1\{\kappa_1 X + \kappa_2 Z + \kappa_3 \epsilon + \eta > 0\}, \qquad \eta \sim N(0,1), \kappa_k = 1, k = 1,2,3$$

$$Z = \chi_2 X^2 + \phi, \qquad \phi \sim N(0,1),$$

$$R = 1\{\gamma_1 X + \gamma_2 \epsilon + \omega > 0\}, \qquad \omega \sim N(0,1), \gamma_1 = 1, \gamma_2 = 1$$

In this data generating process, the instrument is missing non-randomly, $E[\epsilon|R = 1, X] \neq E[\epsilon|X]$. It is also clear that Assumptions 1-3 are satisfied. Further, the instrument Z is non-linearly related to X whenever $\chi_2 \neq 0$.

Table 1 provides simulation results when changing the degree of non-linearity between Z and X. In particular, the first column sets $\chi_2 = 0$, whereas columns 2-4 specify $\chi_2 > 0$. Each column computes 6 different estimators for β, using 500 replications from the data generating process with sample sizes of 50,000. All of the estimators are averages over the replications.

The first row computes the OLS estimator. The next row estimates the moment based IV estimator, defined in (2), in the benchmark case where the instrument is fully observed, $pr(R = 1) = 1$. The final four rows focus attention on

the case with partially observed instrument, $pr(R = 1) \in (0, 1)$. In particular, the third row estimates the moment based IV estimator applied to the sub-sample with non-missing instrument, defined in (5). And the fourth row computes the IV estimator proposed by Angrist et al. (2009), defined in (7), based on a full-sample instrument constructed from a linear projection of the partially observed instrument on the covariates in the sub-sample with non-missing instrument. Finally, the two last rows estimate the proposed IV estimators, $\beta(W, 1)$ defined in (8) and $\beta(IV - NP)$ defined in (10), which are robust to non-randomly missing instruments without auxiliary assumptions. To estimate $\beta(W, 1)$ we partition the scalar covariate X defined over the range $(-1, 1)$ into 41 equally spaced cells. Next, we compute β_X within each of these cells, before forming $\beta(W, 1)$ as the simple average of these β_X. To estimate $\beta(IV - NP)$, we construct Z_{NP} by using a fourth order polynomial in X to approximate $E[Z|X, R = 1]$.

As is clear from the results provided in the first row, the OLS estimates are severely biased upward because of the endogeneity in S, by about 78-95 percent depending on the specification of χ_2. In comparison, it is evident from the second row that the standard moment based IV estimator perform well in the benchmark case where the instrument is fully observed; this is true for all specifications of χ_2. The results in the first column confirm that all the IV estimators provide consistent estimates of β also in the special case where there is a linear relationship between Z and X, that is, $\chi_2 = 0$.

However, the performance of the alternative IV estimators differ substantially in columns 2-4, in which there is a non-linear relationship between Z and X. The third row demonstrates that the IV estimator applied to the sub-sample where the instrument is non-missing exhibit exhibits substantial bias when $\chi_2 \neq 0$: the estimate of β is 1.14 when $\chi_2 = 0.5$ and 1.44 when $\chi_2 = 2$. This finding confirms the result in Proposition 1, stating that Assumptions 1-3 do not ensure that $\beta(IV, 1)$ yields consistent estimate of $\beta(IV)$ – auxiliary assumptions are required. The fourth row shows that also the IV estimates based on the full-sample instrument from the linear projection method are substantially biased when $\chi_2 \neq 0$. Indeed, as stated in Lemma 1, it is clear that using $\beta(IV - LP)$ is equivalent to using $\beta(IV, 1)$. Finally, the last two rows confirm that the two alternative IV estimators are robust to non-randomly missing instruments without auxiliary assumptions, with estimates of β very close to the true value of 1.

7 Conclusion

This paper examines IV estimation in situations where the instrument is observed from some sub-sample and missing for the remaining sample. In such cases, which are quite common in empirical research, there are two options. Typically, researchers simply limit the analysis to the sub-sample where the instrument is non-missing. The second option is to form an instrument that is defined for the full sample and pool samples, as suggested by Angrist et al. (2009). We show that when the instrument is non-randomly missing, the ordinary IV assumptions are inadequate for standard IV estimators to be consistent – in both cases, strong, auxiliary assumptions are necessary. One sufficient auxiliary assumption is that the conditional expectation function between the instrument and the included exogenous covariates is truly linear. However, in many (quasi)natural experiments the auxiliary assumptions are unlikely to hold. For example, the relationship between a binary instrument and the covariates is necessarily non-linear. We therefore introduce alternative IV estimators that are consistent under the ordinary IV assumptions. A Monte-Carlo study illustrates the inconsistency in standard IV estimators when the relationship between the instrument and the covariates is non-linear. Moreover, it shows that the proposed IV estimators are robust to non-randomly missing instruments without auxiliary assumptions.

The result that non-linearities in the relationship between a non-randomly missing instrument and the covariates may matter for the consistency of standard IV estimators is worrisome, given that linear specifications are generally viewed as innocent in IV estimation. As discussed above, there are several reasons for this view. First, consistency of the IV estimator does not require the correct specification of the relationship between the endogenous regressor and the instrument in the first stage. Also, a linear specification of the relationship between the outcome and the endogenous regressor can be justified even when the outcome is binary or the endogenous regressor takes on multiple values. However, in situations where the instrument is only observed for some sub-sample, allowing for a non-linear relationship between the instrument and the covariates is preferable.

References

ANGRIST, J. (2001): "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19(1), 2–16.

ANGRIST, J. D., AND W. N. EVANS (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88(3), 450–77.

ANGRIST, J. D., AND G. IMBENS (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of American Statistical Association*, 90(430), 431–442.

ANGRIST, J. D., AND A. B. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

ANGRIST, J. D., V. LAVY, AND A. SCHLOSSER (2009): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *MIT Working Paper, Revised Version*.

ANGRIST, J. D., AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist Companion*. Princeton University Press, Princeton, New Jersey.

BLACK, S. E., P. J. DEVEREUX, AND K. G. SALVANES (2005): "The More the Merrier? The Effects of Family Size and Birth Order on Children's Education," *Quarterly Journal of Economics*, 120, 669–700.

CACERES-DELPIANO, J. (2006): "The Impacts of Family Size On Investment in Child Quality," *Journal of Human Resources*, 41(4), 738–754.

CARD, D. (1993): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," 4483.

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5), 1127–60.

CARNEIRO, P., AND J. J. HECKMAN (2002): "The Evidence on Credit Constraints in Post–secondary Schooling," *Economic Journal*, 112(482), 705–734.

CHAMBERS, R. L., AND R. DUNSTAN (1986): "Estimating Distribution Functions from Survey Data," *Biometrika*, 73(3), 597–604.

EISENHAUER, J. G. (2003): "Regression through the Origin," *Teaching Statistics*, 25(3), 76–80.

FRANGAKIS, C., AND D. RUBIN (1999): "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86(2), 365–379.

HECKMAN, J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

KANE, T. J., AND C. E. ROUSE (1993): "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?," NBER Working Papers 4268, National Bureau of Economic Research, Inc.

KELEJIAN, H. H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*, 66(334), 373–4.

LITTLE, R. J. A., AND D. B. RUBIN (2002): *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, 2 edn.

MOGSTAD, M., AND M. WISWALL (2009): "How Much Should We Trust Linear Instrumental Variables Estimators? An Application to Family Size and Children's Education," *Statistics Norway Discussion Paper*, (586).

ROSENZWEIG, M. R., AND K. I. WOLPIN (1980): "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48(1), 227–240.

YUAN, J., AND R. SAVICKAS (2009): "Cross-Sectional Estimation Biases in Risk Premia and Zero-Beta Excess Returns," *Working Paper, George Washington University*.

Table 1: Simulation Results

Degree of Non-Linearity (χ_2)		0	0.5	1	2
Sample, Estimator	IV Missing		Slope Est. (β)		
			(True value: 1)		
1) Full Sample, OLS	–	1.9490	1.9011	1.8379	1.7762
2) Full Sample, IV	NO	0.9992	0.9997	1.0000	1.0003
3) Non-Missing Sample, IV	YES	0.9993	1.1420	1.2648	1.4389
4) Full Sample, Linear Projection IV	YES	0.9993	1.1420	1.2648	1.4389
5) Non-Missing Sample, Robust IV	YES	0.9941	0.9936	0.9925	0.9838
6) Full Sample, Robust IV	YES	0.9993	0.9992	0.9991	0.9989

Notes: Simulation from 500 replications of the data generating process described in Section 6. See also Section 6 for definitions of the estimators. Estimates are averages of each estimator over the 500 replications.