

Selection on Observables as an IV Thought Experiment

Cooper Desmond

1 Introduction

Selection on observables and instrumental variables are usually taught as two separate causal arguments. This note runs a thought experiment that puts them in conversation. Take the linear model with a regressor of interest x_1 and exogenous controls X_2 , and let x_1 serve as its own instrument. After the controls are partialled out, the instrument is $\tilde{x}_1$, the part of x_1 the controls cannot explain. Effects are homogeneous throughout, so OLS and IV target the same constant coefficient β_1 , and the only question is what each design must assume to recover it. Treating x_1 as its own instrument lets that question be answered assumption by assumption, in the language of instruments.

Three interpretations follow, and together they show that nothing in the estimator, the asymptotics, or the form of the identifying conditions separates the two designs. First, by Frisch-Waugh-Lovell, OLS of y on (x_1, X_2) and just-identified IV with instrument z and controls X_2 are the same ratio on residualized data, so setting $z = x_1$ turns OLS into IV with no change in machinery. Second, the IV conditions map onto the selection on observables conditions: instrument validity, $\text{Cov}(\tilde{x}_{1i}, u_i) = 0$, is exactly the no-omitted-variable-bias condition; relevance is the requirement that some variance in x_1 survive the controls, which is the OLS full-rank condition; and exclusion is vacuous because the instrument and the regressor are the same variable. Third, the homoskedastic asymptotic variances are equivalent. The OLS variance $\sigma^2 / \text{Var}(\tilde{x}_{1i})$ is the IV variance $\sigma^2 / (R_{x_1 \sim \tilde{z}}^2 \text{Var}(x_{1i}))$ evaluated at $\tilde{z} = \tilde{x}_1$, where the first-stage $R_{x_1 \sim \tilde{x}_1}^2$ is $1 - R_{x_1 \sim X_2}^2$. The familiar variance inflation factor is therefore the inverse of a first-stage R^2 , and multicollinearity is a weak first stage for the instrument OLS uses implicitly.

What does separate the designs is the source of the identifying variation and what can be said in its defense. An OLS regression with controls claims that all of the variation in x_1 left over after the controls is as good as random, so the controls define the instrument and the assumption does the identifying, while an IV regression claims this only for the variation induced by an external z , a single source that can usually be described institutionally. The thought experiment makes the first claim explicit by writing it as the validity condition on an instrument, which is the form in which it can be questioned. The rest of the note works through this in order: Section 2 sets notation, Section 3 states the thought experiment and does a standard derivation of the selection on observables conditions, Section 4 reads those conditions as IV conditions, Section 5 shows the variance equivalence, and Section 6 draws out what the exercise says about how to read a regression with controls.

2 Setup and notation

Let there be n i.i.d. observations. Write $y \in \mathbb{R}^n$ for the outcome, $x_1 \in \mathbb{R}^n$ for the regressor of interest, $X_2 \in \mathbb{R}^{n \times k}$ for a full column rank matrix of exogenous controls (with the constant absorbed into X_2), and $z \in \mathbb{R}^n$ for a single instrument, so that the number of instruments equals the number of endogenous variables. The structural equation is

$$y = x_1\beta_1 + X_2\gamma + u, \quad (1)$$

where β_1 is the object of interest.

Define the projection and annihilator matrices associated with X_2 ,

$$P_2 = X_2(X_2'X_2)^{-1}X_2', \quad M_2 = I_n - P_2, \quad (2)$$

which are symmetric and idempotent. Residualization with respect to the controls is denoted by a tilde:

$$\tilde{y} = M_2y, \quad \tilde{x}_1 = M_2x_1, \quad \tilde{z} = M_2z. \quad (3)$$

For population statements, let $\lambda = E(X_{2i}X_{2i}')^{-1}E(X_{2i}x_{1i})$ be the coefficients of the population projection of x_{1i} on X_{2i} , and let $\tilde{x}_{1i} = x_{1i} - X_{2i}'\lambda$ be the population residual, with $\tilde{z}_i$ and λ_z defined in the same way for z . Because the constant is a column of X_2 , $E(\tilde{x}_{1i}) = 0$ and $E(X_{2i}\tilde{x}_{1i}) = 0$. Sample moments of the sample residuals converge to the corresponding population moments, so for example $n^{-1}\tilde{x}_1'\tilde{x}_1 \xrightarrow{p} E(\tilde{x}_{1i}^2) = \text{Var}(\tilde{x}_{1i})$.

3 The thought experiment and the selection on observables conditions

Set $z = x_1$, so that $\tilde{z} = \tilde{x}_1$, the regressor of interest residualized on the controls, is its own instrument. By Frisch-Waugh-Lovell, OLS of y on (x_1, X_2) and IV with instrument z and controls X_2 are the same ratio on residualized data,

$$\hat{\beta}_1 = (\tilde{z}'\tilde{x}_1)^{-1}\tilde{z}'\tilde{y}, \quad \begin{cases} \tilde{z} = \tilde{x}_1 & (\text{OLS: the instrument is } x_1 \text{ itself}) \\ \tilde{z} = M_2z & (\text{IV: } z \text{ external}). \end{cases}$$

While they are the same ratio, they rest on different assumptions. Selection on observables requires that, once X_2 is controlled for, there is no omitted variable bias. To make this explicit, write the structural error as $u = w\delta + \varepsilon$, where $w \in \mathbb{R}^n$ is an index of the determinants of y omitted from X_2 and ε is mean-independent of (x_1, X_2, w) . Residualizing (1) gives $\tilde{y} = \tilde{x}_1\beta_1 + \tilde{u}$ since $M_2X_2 = 0$, and because M_2 is symmetric and idempotent, $\tilde{x}_1'\tilde{u} = \tilde{x}_1'u = \delta\tilde{x}_1'w + \tilde{x}_1'\varepsilon$. Setting $\tilde{z} = \tilde{x}_1$,

$$\hat{\beta}_1 - \beta_1 = (\tilde{x}_1'\tilde{x}_1)^{-1}\tilde{x}_1'\tilde{u} = \underbrace{\delta(\tilde{x}_1'\tilde{x}_1)^{-1}\tilde{x}_1'w}_{\text{OVB}} + (\tilde{x}_1'\tilde{x}_1)^{-1}\tilde{x}_1'\varepsilon, \quad (4)$$

so that

$$\hat{\beta}_1 - \beta_1 = \delta \frac{n^{-1}\tilde{x}_1'w}{n^{-1}\tilde{x}_1'\tilde{x}_1} + \frac{n^{-1}\tilde{x}_1'\varepsilon}{n^{-1}\tilde{x}_1'\tilde{x}_1} \xrightarrow{p} \delta \frac{\text{Cov}(\tilde{x}_{1i}, w_i)}{\text{Var}(\tilde{x}_{1i})}. \quad (5)$$

The second term vanishes in the limit because ε is mean-independent of the regressors. The selection on observables assumption is that the limit in (5) is zero: either nothing relevant is omitted ($\delta = 0$), or, once X_2 is partialled out, x_1 is uncorrelated with the omitted determinants of y ($\text{Cov}(\tilde{x}_{1i}, w_i) = 0$). In either case the OVB term is zero and $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

In addition, we need some leftover variance in x_1 after partialling out the controls,

$$\frac{1}{n} \tilde{x}_1' \tilde{x}_1 \xrightarrow{p} \text{Var}(\tilde{x}_{1i}) > 0, \quad (6)$$

so that the denominator of the limit in (5) is nonzero. Equivalently, x_1 cannot be a linear combination of the columns of X_2 , so $[x_1, X_2]$ must have full column rank. The sample analogue, $\tilde{x}_1' \tilde{x}_1 > 0$, is what guarantees that $\hat{\beta}_1$ exists.

4 The selection on observables conditions as IV conditions

Now read the same ratio as an IV estimator. Since $\tilde{y} = \tilde{x}_1 \beta_1 + \tilde{u}$ and $\tilde{z}' \tilde{u} = \tilde{z}' u$,

$$\hat{\beta}_1 - \beta_1 = (\tilde{z}' \tilde{x}_1)^{-1} \tilde{z}' u \xrightarrow{p} \frac{\text{Cov}(\tilde{z}_i, u_i)}{\text{Cov}(\tilde{z}_i, \tilde{x}_{1i})} = \frac{\text{Cov}(\tilde{x}_{1i}, u_i)}{\text{Var}(\tilde{x}_{1i})}, \quad (7)$$

where the last equality sets $\tilde{z} = \tilde{x}_1$. The general IV probability limit collapses to the OLS one, with the instrument's covariance with the regressor replaced by the regressor's own residual variance. The IV conditions for consistency now map onto the two selection on observables conditions of Section 3, with exclusion holding trivially.

Validity. Validity requires the numerator of (7) to vanish,

$$\frac{1}{n} \tilde{x}_1' u \xrightarrow{p} 0, \quad (8)$$

so the residualized instrument must be uncorrelated with the structural error. Substituting $u = w\delta + \varepsilon$,

$$\text{Cov}(\tilde{x}_{1i}, u_i) = \delta \text{Cov}(\tilde{x}_{1i}, w_i) + \text{Cov}(\tilde{x}_{1i}, \varepsilon_i) = \delta \text{Cov}(\tilde{x}_{1i}, w_i), \quad (9)$$

since $\tilde{x}_{1i}$ is a function of (x_{1i}, X_{2i}) and ε is mean-independent of both. Instrument validity is therefore exactly the no-OVB condition of Section 3, either $\delta = 0$ or $\text{Cov}(\tilde{x}_{1i}, w_i) = 0$. As with any just-identified instrument, it cannot be tested directly. What can be done in its defense is the same in both designs, with a difference in scope. With an external instrument, the researcher makes an institutional argument that z is randomly or quasi-randomly assigned given X_2 , as in a lottery, and checks balance. Balance entails regressing pre-determined characteristics that were held out of X_2 on (z, X_2) and seeing if there is evidence against a zero coefficient on z . As-good-as-random assignment implies balance, so imbalance is evidence against validity, though the absence of imbalance does not necessarily establish it. In the thought experiment the same two moves are theoretically available for $\tilde{x}_1$.

In this case, an institutional argument would be about the selection process into x_1 given X_2 , and the balance check regresses pre-determined characteristics X_3 held out of the controls on (x_1, X_2) and asks whether there is evidence against a zero coefficient on x_1 , which by FWL is $\text{Cov}(\tilde{x}_{1i}, X_{3i}) = 0$. The main difference is in the scope, where the IV researcher defends only the variation in x_1 that z induces, while the OLS researcher must defend all of the variation left after the controls, and the check is informative only for characteristics deliberately left out of X_2 , since anything moved into the controls is balanced by construction.

Relevance. Relevance requires the denominator of (7) to be nonzero. Write the first stage as $x_1 = z\pi_1 + X_2\pi_2 + v$. With $z = x_1$ it fits perfectly, with $\pi_1 = 1$, $\pi_2 = 0$, and $v = 0$. Relevance therefore reduces to $\text{Var}(\tilde{x}_{1i}) > 0$, which is condition (6). In matrix form, let $X = [x_1, X_2]$ and $Z = [z, X_2] = X$. The rank condition, that $n^{-1}Z'X \xrightarrow{p} E(Z_i X_i')$ with $E(Z_i X_i')$ of full column rank, becomes the usual OLS requirement that $n^{-1}X'X \xrightarrow{p} E(X_i X_i')$ with $E(X_i X_i')$ nonsingular. The reduced form of y on (z, X_2) is the OLS regression itself, so the IV identity $\beta_1 = \text{reduced form/first stage}$ is the OLS coefficient divided by one.

Exclusion. Exclusion requires that z affect y only through x_1 . When $z = x_1$ there is no separate channel to exclude, so the condition is vacuous and adds nothing beyond validity.

5 Asymptotic variances

The two estimators are the same ratio and rest on matching assumptions. This section shows that the textbook homoskedastic asymptotic variances match as well.

5.1 OLS

Stack the regressors as $X = [x_1, X_2] \in \mathbb{R}^{n \times (1+k)}$ with i th row $X_i' = (x_{1i}, X_{2i}')$, and let $\theta = (\beta_1, \gamma')'$, so that (1) is $y = X\theta + u$. Assume selection on observables in its conditional-mean form, together with homoskedasticity,

$$E(u_i | X_i) = 0, \quad E(u_i^2 | X_i) = \sigma^2, \quad (10)$$

and that $Q_{XX} = E(X_i X_i')$ is finite and nonsingular, which, given that X_2 has full column rank, is equivalent to (6). The conditional-mean condition strengthens the zero-covariance condition of Section 3 to the form in which the textbook variance is derived. Since $\hat{\theta} - \theta = (X'X)^{-1}X'u$,

$$\sqrt{n}(\hat{\theta} - \theta) = \left(\frac{1}{n}X'X\right)^{-1} \frac{1}{\sqrt{n}}X'u. \quad (11)$$

By the LLN, $n^{-1}X'X \xrightarrow{p} Q_{XX}$. Each term $X_i u_i$ has mean zero by (10) and variance

$$E(u_i^2 X_i X_i') = E[E(u_i^2 | X_i) X_i X_i'] = \sigma^2 Q_{XX},$$

so by the CLT, $n^{-1/2}X'u \xrightarrow{d} N(0, \sigma^2 Q_{XX})$. Slutsky's theorem then gives

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2 Q_{XX}^{-1}). \quad (12)$$

The asymptotic variance of $\hat{\beta}_1$ is the (1, 1) element, $\sigma^2[Q_{XX}^{-1}]_{11}$. With λ and $\tilde{x}_{1i}$ as defined in Section 2,

$$[Q_{XX}^{-1}]_{11} = \left[E(x_{1i}^2) - E(x_{1i}X'_{2i})\lambda \right]^{-1} = \left[E(\tilde{x}_{1i}x_{1i}) \right]^{-1} = \left[E(\tilde{x}_{1i}^2) \right]^{-1} = \frac{1}{\text{Var}(\tilde{x}_{1i})}.$$

The first equality is the partitioned inverse formula. The second substitutes the definition of $\tilde{x}_{1i}$. The third uses $E(\tilde{x}_{1i}X'_{2i}) = 0$, and the last uses $E(\tilde{x}_{1i}) = 0$. Therefore

$$\text{AVar}(\sqrt{n}(\hat{\beta}_1^{\text{OLS}} - \beta_1)) = \sigma^2[Q_{XX}^{-1}]_{11} = \frac{\sigma^2}{\text{Var}(\tilde{x}_{1i})}. \quad (13)$$

5.2 IV

For IV, assume validity with homoskedasticity,

$$E(u_i | z_i, X_{2i}) = 0, \quad E(u_i^2 | z_i, X_{2i}) = \sigma^2, \quad (14)$$

and relevance, $\text{Cov}(\tilde{z}_i, \tilde{x}_{1i}) \neq 0$. By the partitioned FWL theorem, $\hat{\beta}_1^{\text{IV}}$ is the first element of $(Z'X)^{-1}Z'y$ and equals $(\tilde{z}'\tilde{x}_1)^{-1}\tilde{z}'\tilde{y}$. Since $\tilde{y} = \tilde{x}_1\beta_1 + \tilde{u}$ and $\tilde{z}'\tilde{u} = \tilde{z}'u$,

$$\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1) = \left(\frac{1}{n}\tilde{z}'\tilde{x}_1 \right)^{-1} \frac{1}{\sqrt{n}}\tilde{z}'u. \quad (15)$$

By the LLN, $n^{-1}\tilde{z}'\tilde{x}_1 \xrightarrow{p} E(\tilde{z}_i\tilde{x}_{1i}) = \text{Cov}(\tilde{z}_i, \tilde{x}_{1i})$. Each term $\tilde{z}_i u_i$ has mean zero by (14). Since $\tilde{z}_i$ is a function of (z_i, X_{2i}) , its variance is

$$E(u_i^2 \tilde{z}_i^2) = E[E(u_i^2 | z_i, X_{2i}) \tilde{z}_i^2] = \sigma^2 \text{Var}(\tilde{z}_i),$$

so by the CLT, $n^{-1/2}\tilde{z}'u \xrightarrow{d} N(0, \sigma^2 \text{Var}(\tilde{z}_i))$. Slutsky's theorem then gives

$$\text{AVar}(\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1)) = \frac{\sigma^2 \text{Var}(\tilde{z}_i)}{\text{Cov}(\tilde{z}_i, \tilde{x}_{1i})^2} = \frac{\sigma^2 \text{Var}(\tilde{z}_i)}{\text{Cov}(\tilde{z}_i, x_{1i})^2} = \frac{\sigma^2}{R_{x_1 \sim \tilde{z}}^2 \text{Var}(x_{1i})}. \quad (16)$$

The second equality uses $x_{1i} = X'_{2i}\lambda + \tilde{x}_{1i}$ and $\text{Cov}(\tilde{z}_i, X'_{2i}\lambda) = 0$, since $\tilde{z}_i$ is uncorrelated with the controls. The third divides the numerator and denominator by $\text{Var}(\tilde{z}_i)\text{Var}(x_{1i})$, where $R_{x_1 \sim \tilde{z}}^2 = \text{Corr}(x_{1i}, \tilde{z}_i)^2$ is the R^2 from regressing x_1 on the residualized instrument.

5.3 Equivalence under the thought experiment

To put the OLS variance in the IV form, decompose $x_{1i} = X'_{2i}\lambda + \tilde{x}_{1i}$. The two pieces are uncorrelated, since $\text{Cov}(X'_{2i}\lambda, \tilde{x}_{1i}) = \lambda'E(X_{2i}\tilde{x}_{1i}) - E(X'_{2i}\lambda)E(\tilde{x}_{1i}) = 0$. Hence

$$\text{Var}(x_{1i}) = \text{Var}(X'_{2i}\lambda) + \text{Var}(\tilde{x}_{1i}) \implies \text{Var}(\tilde{x}_{1i}) = (1 - R_{x_1 \sim X_2}^2) \text{Var}(x_{1i}), \quad (17)$$

where $R_{x_1 \sim X_2}^2 = \text{Var}(X'_{2i}\lambda)/\text{Var}(x_{1i})$ is the population R^2 from projecting x_1 on X_2 . The same decomposition gives $\text{Cov}(x_{1i}, \tilde{x}_{1i}) = \text{Var}(\tilde{x}_{1i})$, so the squared correlation between x_1 and its own residual is

$$R_{x_1 \sim \tilde{x}_1}^2 = \frac{\text{Var}(\tilde{x}_{1i})^2}{\text{Var}(x_{1i})\text{Var}(\tilde{x}_{1i})} = \frac{\text{Var}(\tilde{x}_{1i})}{\text{Var}(x_{1i})} = 1 - R_{x_1 \sim X_2}^2. \quad (18)$$

Substituting (17) and (18) into (13),

$$\begin{aligned} \text{AVar}(\sqrt{n}(\hat{\beta}_1^{\text{OLS}} - \beta_1)) &= \frac{\sigma^2}{\text{Var}(\tilde{x}_{1i})} = \frac{\sigma^2}{(1 - R_{x_1 \sim X_2}^2) \text{Var}(x_{1i})} \\ &= \frac{\sigma^2}{R_{x_1 \sim \tilde{x}_1}^2 \text{Var}(x_{1i})} = \text{AVar}(\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1)) \Big|_{\tilde{z}=\tilde{x}_1}. \end{aligned} \quad (19)$$

The variance equivalence in (19) is exact because OLS with controls is, mechanically, a just-identified IV estimator with instrument $\tilde{x}_1$. Since $\tilde{x}_1'x_1 = \tilde{x}_1'\tilde{x}_1$ and $\tilde{x}_1'X_2 = 0$, the OLS coefficient is $(\tilde{x}_1'x_1)^{-1}\tilde{x}_1'y$, the IV estimator of y on x_1 alone with $\tilde{x}_1$ as the instrument, and the controls drop out of the sampling error exactly. The content is in how the variance is written. The OLS form $\sigma^2/\text{Var}(\tilde{x}_{1i})$ puts all of the identifying variation in one number. The IV form $\sigma^2/(R_{x_1 \sim \tilde{x}_1}^2 \text{Var}(x_{1i}))$ factors that number into the total variation in the treatment and the share of it the instrument captures, which by (18) is $1 - R_{x_1 \sim X_2}^2$, the share the controls leave behind. The variance inflation factor $1/(1 - R_{x_1 \sim X_2}^2)$ is therefore the inverse of a first-stage R^2 , and it says that OLS can be imprecise not because x_1 varies little but because most of its variation lies along the controls, and only the orthogonal share, the useful variation in IV terms, identifies β_1 . The IV variance (16) has the same form with $R_{x_1 \sim \tilde{z}}^2$ in place of $R_{x_1 \sim \tilde{x}_1}^2$, so OLS is the case in which the instrument captures all of the leftover variation, and multicollinearity can then be considered a weak first stage for this implicit instrument under this exercise.

6 Conclusion

Once the estimator, the asymptotics, and the identifying conditions all line up, what remains to distinguish a selection on observables design from an IV design is the source of the identifying variation and what can be said in its defense. An OLS regression with controls asserts that all of the variation in x_1 left over after the controls is as good as random, which is the validity condition of Section 4 applied to $\tilde{x}_1$. An IV regression asserts the same thing only for the variation induced by z , a single source the researcher can usually describe institutionally and probe with balance checks. The selection on observables assumption is therefore an exogeneity claim about an instrument, and the instrument is all of the residual variation in the treatment. On this reading the controls do not identify anything. Instead, they define the instrument by fixing which variation in x_1 is left over, and the identifying work is done entirely by the assumption that this leftover variation is uncorrelated with u .

A possible practical lesson from this exercise is to read every OLS regression with controls as a just-identified IV regression and to ask the IV questions of it. Where does the variation in $\tilde{x}_1$ come from? Why should it be as good as random with respect to everything left in u ? How much of $\text{Var}(x_{1i})$ survives the controls, that is, how strong is the first stage? These are the questions one would ask of an instrument, and the thought experiment shows that they are the right questions for a control strategy as well. Framed this way, using many controls is not a substitute for an argument about where the remaining variation comes from. Instead, it only narrows the variation that the argument has to cover.