7. Instrumental variables

LPO 8852: Regression II

Sean P. Corcoran

IV - introduction

An **instrumental variable** can be used to "carve out" exogenous variation in a explanatory variable that would otherwise be endogenous. Some useful applications:

- Addressing OVB when adequate controls or panel data are unavailable
- Certain "natural experiments"
- Partial or incomplete random assignment
- Fuzzy regression discontinuity (RD)
- Correcting for measurement error

Preliminary: some rules of covariance

Recall that for random variables X, Y, and Z and constant a:

1. $Cov(aX, Y) = aCov(X, Y)$
2. $Cov(X + c, Y) = Cov(X, Y)$
3. $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
4. $Cov(X, X) = Var(X)$

All of these follow from the definition of covariance:

$$Cov(X, Y) = \sigma_{XY} = E\left[(X - E(X))(Y - E(Y))\right]$$

Covariance algebra applied to simple regression

Suppose we are interested in the population relationship between Y and X, which we believe can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

The covariance between Y and X can be written as:

$$Cov(Y, X) = Cov(\beta_0 + \beta_1 X + u, X)$$

$$Cov(Y, X) = \beta_1 Cov(X, X) + Cov(u, X)$$

$$\sigma_{YX} = \beta_1 \sigma_X^2 + \sigma_{Xu}$$

$$\frac{\sigma_{YX}}{\sigma_X^2} = \beta_1 + \frac{\sigma_{Xu}}{\sigma_X^2}$$

Covariance algebra applied to simple regression

The slope estimator you learned in Reg 1 is the sample analog of the term on the LHS:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

In large samples, this converges to the population quantity σ_{YX}/σ_X^2. However, this quantity is <u>not</u> generally β_1, but β_1 plus a bias term, as the above formula shows.

You will recognize the previous slide's formula as the omitted variables bias formula for a simple regression.

Covariance algebra applied to simple regression

If the population covariance between X and u is 0 ($\sigma_{Xu} = 0$) then:

$$\frac{\sigma_{YX}}{\sigma_X^2} = \beta_1$$

This is called the **method of moments** derivation of β_1. As long as $\sigma_{Xu} = 0$, the population covariance between X and Y divided by the variance of X gives you β. In practice, we use the sample covariance and sample variance (s_{XY} and s_X^2) to estimate β:

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

IV introduction

Now consider the same population regression function for Y, but where $Cov(X_i, u_i) \neq 0$ (X_i is **endogenous**).

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Suppose a third variable Z_i is available, where:

1. $Cov(Z_i, X_i) \neq 0$
2. $Cov(Z_i, u_i) = 0$

Z_i is an **instrumental variable** or **instrument**. A *valid instrument* satisfies the above two properties. These are the key identification assumptions for IV.

IV introduction

What can a (valid) instrument do for us? Begin again with:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Applying covariance algebra:

$$Cov(Z, Y) = Cov(Z, \beta_0 + \beta_1 X + u)$$

$$Cov(Z, Y) = \beta_1 Cov(Z, X) + Cov(Z, u)$$

$$\sigma_{ZY} = \beta_1 \sigma_{ZX} + \sigma_{Zu}$$

$$\frac{\sigma_{ZY}}{\sigma_{ZX}} = \beta_1 + \frac{\sigma_{Zu}}{\sigma_{ZX}}$$

IV introduction

If the two identification assumptions hold ($\sigma_{Zu} = 0$ and $\sigma_{ZX} \neq 0$) then:

$$\beta_{1,IV} = \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{Cov(Z,Y)}{Cov(Z,X)}$$

This is the method of moments derivation of β_1 using the instrument Z.

IV introduction

$\beta_{1,IV}$ can be estimated using sample covariance between Z and Y (s_{ZY}) and the sample covariance between Z and X (s_{ZX}):

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

Compare this to the OLS simple regression slope estimator below. If X itself satisfies the two identifying assumptions (i.e., it is exogenous), it can be used as an "instrument for itself" and the above simplifies to the traditional OLS estimator:

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

IV identifying assumptions

1. $Cov(Z_i, X_i) \neq 0$
2. $Cov(Z_i, u_i) = 0$

We cannot test the second of these conditions, since u is unobserved. We must rely on theory or introspection to rationalize this.

We *can* offer evidence for the first condition, by estimating this regression and testing for significance of $\hat{\pi}_1$:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

IV identifying assumptions

When might the second condition be violated? Suppose we are interested in the relationship between class attendance and final grade among college students:

$$\text{final grade} = \beta_0 + \beta_1(\text{days absent}) + u$$

We can hypothesize that $\beta_1 < 0$, but one might worry that the coefficient in the above regression is not causal. There may be factors associated with class attendance that are also associated with the final grade (e.g., general motivation, subject interest, family resources).

Average *potential* outcomes for students with more/less days absent are not the same, on average.

IV identifying assumptions

Consider as a potential instrument distance from home to school (Z). It is plausible that students who live further from school miss more class due to unanticipated weather, traffic, etc. This would imply $Cov(Z, X) \neq 0$. One could also argue that weather and traffic are uncorrelated with potential outcomes.

However, what if less engaged students choose to live further (on average) from school? Or if housing prices are high near school such that distance to school is related to ability to pay? Then distance to school may be correlated with potential outcomes, or $Cov(Z, u) \neq 0$.

More on this later.

Application to charter school lottery

We are interested in the following *structural equation*, where for student i, Y_i is a test score and D_i = attendance at a charter school:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

Presumably, attendance at charter schools is non-random and correlated with potential outcomes. OLS estimation of the above may suffer from omitted variables bias: $Cov(u_i, D_i) \neq 0$.

Fortunately, at over-subscribed charters, schools hold random lotteries which determine offers of admission. Let $Z_i = 1$ if a student receives a random offer to attend a charter school and $Z_i = 0$ otherwise.

Application to charter school lottery

Z_i is a valid instrumental variable for D_i since $Cov(Z_i, D_i) > 0$ and $Cov(Z_i, u_i) = 0$. We can use instrumental variables to estimate the population slope coefficient β_1:

$$\beta_{1,IV} = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)}$$

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(D_i - \bar{D})}$$

Application to charter school lottery

Alternatively, consider the following two simple regressions:

$$Y_i = \alpha + \rho Z_i + w_i$$

$$D_i = \gamma + \phi Z_i + v_i$$

The former equation is the **reduced form**, the latter is the **first stage**. Z_i is randomly assigned, so there is no OVB in either case. The population slope coefficients can be written as:

$$\rho = \frac{Cov(Z_i, Y_i)}{Var(Z_i)}$$

$$\phi = \frac{Cov(Z_i, D_i)}{Var(Z_i)}$$

Application to charter school lottery

Note that:

$$\frac{\rho}{\phi} = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} \times \frac{Var(Z_i)}{Var(Z_i)} = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)}$$

That is, the reduced form slope coefficient divided by the first stage coefficient is $\beta_{1,IV}$—the causal effect of charter school attendance.

One can use the sample analogs $\hat{\rho}$ and $\hat{\phi}$ to estimate $\beta_{1,IV}$.

Application to charter school lottery

Both Z_i and D_i are binary variables in this example. Another way to write the slope coefficients in the reduced form and first stage are:

$$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$
$$\phi = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$

ρ is the difference in the mean of Y for those randomly assigned an offer and those not—the **ITT**. ϕ is the difference in the proportion treated for those randomly assigned an offer and those not. The ratio is $\beta_{1,IV}$:

$$\frac{\rho}{\phi} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

KIPP example (Angrist & Pischke, ch. 3)

KIPP randomizes the *offer* of a seat at its schools. Angrist et al. (2010, 2012) examined the impact attending KIPP Lynn (MA).

- 629 applicants between 2005 and 2008
- 446 assigned via the lottery and had complete data
- 303 (68%) were offered a seat
- 221 of the 303 enrolled in KIPP (73%)
- 3.5% of lottery losers enrolled in KIPP

The lottery should yield treatment and control groups that are on average equivalent at baseline, including unobserved heterogeneity and pre-treatment outcomes. See Table 3.1, column 3 panel A.

KIPP example (Angrist & Pischke, ch. 3)

TABLE 3.1
Analysis of KIPP lotteries

		KIPP applicants			
	Lynn public fifth graders (1)	KIPP Lynn lottery winners (2)	Winners vs. losers (3)	Attended KIPP (4)	Attended KIPP vs. others (5)
Panel A. Baseline characteristics					
Hispanic	.418	.510	−.058 (.058)	.539	.012 (.054)
Black	.173	.257	.026 (.047)	.240	−.001 (.043)
Female	.480	.494	−.008 (.059)	.495	−.009 (.055)
Free/Reduced price lunch	.770	.814	−.032 (.046)	.828	.011 (.042)
Baseline math score	−.307	−.290	.102 (.120)	−.289	.069 (.109)
Baseline verbal score	−.356	−.386	.063 (.125)	−.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1.000	1.000 —
Math score	−.363	−.003	.355 (.115)	.095	.467 (.103)
Verbal score	−.417	−.262	.113 (.122)	−.211	.211 (.109)
Sample size	3,964	253	371	204	371

Notes: This table describes baseline characteristics of Lynn fifth graders and reports estimated offer effects for Knowledge Is Power Program (KIPP) Lynn applicants. Means appear in columns (1), (2), and (4). Column (3) shows differences between lottery winners and losers. These are coefficients from regressions that control for risk sets, namely, dummies for year and grade of application and the presence of a sibling applicant. Column (5) shows differences between KIPP students and applicants who did not attend KIPP. Standard errors are reported in parentheses.

KIPP example (Angrist & Pischke, ch. 3)

Using *math score* estimates from Panel B of Table 3.1:

Reduced form:

$$\hat{\rho} = Avg[Y_i|Z_i = 1] - Avg[Y_i|Z_i = 0] = 0.355$$

First stage:

$$\hat{\phi} = Avg[D_i|Z_i = 1] - Avg[D_i|Z_i = 0] = 0.741$$

Instrumental variables estimate:

$$\frac{\hat{\rho}}{\hat{\phi}} = \frac{0.355}{0.741} = 0.479$$

KIPP example (Angrist & Pischke, ch. 3)

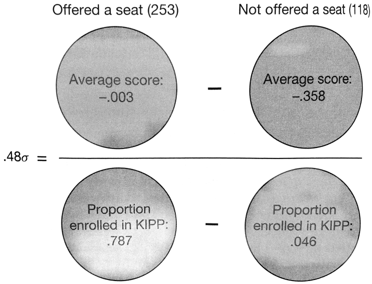
"The IV estimator converts KIPP *offer* estimates into KIPP *attendance* effects." The logic:

- The instrument (Z) has a causal effect on KIPP enrollment (D)—the *first stage*
- The instrument (Z) is randomly assigned (or is as good as randomly assigned) and is thus unrelated to omitted variables in the main equation: the *independence assumption*
- The instrument (Z) affects Y only through D: the *exclusion restriction*

The exclusion restriction implies that the 0.355 difference in Y is attributable only to the 0.741 increase in KIPP attendance.

KIPP example (Angrist & Pischke, ch. 3)

FIGURE 3.2
IV in school: the effect of KIPP attendance on math scores



Note: The effect of Knowledge Is Power Program (KIPP) enrollment described by this figure is $.48\sigma = .355\sigma/.741$.

KIPP example (Angrist & Pischke, ch. 3)

The quantity estimated by IV here is a **local average treatment effect (LATE)**. Why?

- The ratio tells us the ATE for those induced into treatment by the instrument (the lottery offer): the *compliers*
- The first stage is driven by compliers.
- We don't learn anything about *always-takers* or *never-takers*.
- We assume there are no *defiers*. This is called a **monotonicity** assumption—the instrument pushes treatment in one direction only.

If the treated population includes always-takers, the LATE and TOT are not generally the same. (Always-takers are among the treated and may have different potential outcomes from the compliers).

Local average treatment effect (LATE)

In this example it is important to note the difference between *assignment to treatment* ($Z_i = 1$ or $Z_i = 0$) and *actual treatment*, or the receipt of treatment ($D_i = 1$ or $D_i = 0$). These can differ in "broken experiments."

In the following example, suppose there are 100 individuals randomized (50/50) to treatment. In the population, 50% are compliers, 30% are always-takers, and 20% are never-takers. We will assume *no defiers*.

Local average treatment effect (LATE)

	Instrument Z_i (assignment to treatment)					
	$Z_i = 1$	$Z_i = 0$	$Z_i = 1$ Pr(D=1)	$Z_i = 0$ Pr(D=1)	$Z_i = 1$ Share	$Z_i = 0$ Share
Compliers	$Z_i = 1$ $D_i = 1$	$Z_i = 0$ $D_i = 0$	100	0	50	50
Always takers	$Z_i = 1$ $D_i = 1$	$Z_i = 0$ $\mathbf{D_i = 1}$	100	100	30	30
Never takers	$Z_i = 1$ $\mathbf{D_i = 0}$	$Z_i = 0$ $D_i = 0$	0	0	20	20
Defiers	$Z_i = 1$ $\mathbf{D_i = 0}$	$Z_i = 0$ $\mathbf{D_i = 1}$	0	100	0	0

By randomization, the $Z_i = 1$ and $Z_i = 0$ groups should include the same proportions of compliers, always-takers, and never-takers, in expectation.

Local average treatment effect (LATE)

Now consider potential outcomes for these groups, $Y(1)$ and $Y(0)$. Notice the heterogeneous treatment effect.

			Potential outcomes		
	$Z_i = 1$	$Z_i = 0$	Y(1)	Y(0)	ATE
Compliers	$Z_i = 1$	$Z_i = 0$			
	$D_i = 1$	$D_i = 0$	500	250	250
Always takers	$Z_i = 1$	$Z_i = 0$			
	$D_i = 1$	$\mathbf{D_i = 1}$	600	400	200
Never takers	$Z_i = 1$	$Z_i = 0$			
	$\mathbf{D_i = 0}$	$D_i = 0$	300	250	50

Local average treatment effect (LATE)

$$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= 500 * 0.50 + 600 * 0.30 + 250 * 0.20 \\ &- (\underbrace{250 * 0.50}_{\text{compliers}} - \underbrace{600 * 0.30}_{\text{always-takers}} - \underbrace{250 * 0.20}_{\text{never-takers}}) \\ &= 125 \end{aligned}$$

The reduced form (ITT) $\rho = 125$.

Local average treatment effect (LATE)

$$\phi = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$

$$\begin{aligned} E[D_i|Z_i = 1] - E[D_i|Z_i = 0] &= 1 * 0.50 + 1 * 0.30 + 0 * 0.20 \\ &\quad - (\underbrace{0 * 0.50}_{\text{compliers}} - \underbrace{1 * 0.30}_{\text{always-takers}} - \underbrace{0 * 0.20}_{\text{never-takers}}) \\ &= 0.5 \end{aligned}$$

The first stage $\phi = 0.5$. So $\rho/\phi = 125/0.5 = 250$—the LATE or **complier average treatment effect**.

Local average treatment effect (LATE)

More intuition:

- The ITT is a weighted average of the *compliers'* treatment effect and zero. The ITT is not affected by always-takers or never-takers, since their actual treatment receipt is the same whether $Z_i = 1$ or $Z_i = 0$.
- Treatment assignment (Z) is used to estimate the proportion of compliers in the treatment group (which includes both compliers and always-takers). Call this p_c below.
- The estimated proportion of compliers is used to adjust ("scale up") the ITT to obtain and estimate of LATE.

$$(LATE \times p_c) + (0 \times (1 - p_c)) = ITT$$

$$LATE = \frac{ITT}{p_c}$$

Example 2

N=125 students where 40% were assigned randomly to a school tutoring intervention.

	Actual treatment (D)			Pr of total	
Assigned treatment (Z)	Tutoring	None	Total	Tutoring	No tutoring
Tutoring	38	12	50	0.760	0.240
No tutoring	10	65	75	0.133	0.867
Total	48	77	125		

	Mean outcomes (Y)			
	Tutoring	None	All	
Tutoring	72.3	61.4	69.7	
No tutoring	75.6	54.8	57.6	
			12.1	<–Reduced form (ITT)

Example 2

Reduced form:

$$\hat{\rho} = Avg[Y_i|Z_i = 1] - Avg[Y_i|Z_i = 0] = 69.7 - 57.6 = 12.1$$

First stage:

$$\hat{\phi} = Avg[D_i|Z_i = 1] - Avg[D_i|Z_i = 0] = 0.760 - 0.133 = 0.627$$

$$\hat{\beta}_{1,IV} = \hat{\rho}/\hat{\phi} = 12.1/0.627 = 19.3$$

Example 2

The point estimate of 19.3 is a LATE: the impact of tutoring on the population induced into this intervention by the random assignment.

- It is <u>not</u> informative about never-takers: those who would not receive the tutoring intervention in any case.
- It is <u>not</u> informative about always-takers: those who would receive the tutoring intervention in any case.
- The table suggests about 13% always-takers. Since the treated group includes always-takers, LATE differs from ATT. (If this were small, we might be able to say that the LATE approximates the ATT).

For another example see Angrist & Pischke chapter 3 on the Minneapolis Domestic Violence Experiment in the 1980s (Table 3.3).

IV bivariate regression—estimator properties

- The IV estimator is *consistent* as long as the instrument is valid and the identification assumptions hold:
  - $Cov(Z_i, X_i) \neq 0$
  - $Cov(Z_i, u_i) = 0$
- The IV estimator is *biased* in finite samples (as long as X and u are in fact correlated). More on this later.
- The IV estimator is *less efficient* than OLS.

Note: see Wooldridge chapter 15 for good coverage of this.

IV bivariate regression—estimator properties

Assuming homoskedasticity, the asymptotic variance of $\hat{\beta}_{1,IV}$ is:

$$Var(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

- σ_x^2 is the variance of x (can estimate with s_x^2)
- $\rho_{x,z}^2$ is the squared correlation between X and Z (can estimate with R^2 from a regression of X on Z)
- σ^2 is the variance of the residuals, conditional on Z (can estimate with residuals from the IV model)

IV bivariate regression—estimator properties

Compare to the asymptotic variance of $\hat{\beta}_{1,OLS}$ from Reg 1:

$$Var(\hat{\beta}_{1,OLS}) = \frac{\sigma^2}{n\sigma_x^2}$$

The presence of $\rho_{x,z}^2 \leq 1$ in the denominator of $Var(\hat{\beta}_{1,IV})$ implies that the variance of the IV estimator will be *greater* than that of the OLS estimator. This is intuitive: we are using a portion of the variation in X to estimate β.

Instruments for continuous variables

The charter lottery and tutoring examples might be called "broken experiments," where assignment to treatment was random, but treatment delivery (or participation) was not. Both Z and D were binary variables.

Instrumental variables are also used in *natural experiments* and other designs where the natural experiment creates an instrument. The instrument allows us to isolate exogenous variation in X, where X is a continuous variable.

Instruments for continuous variables

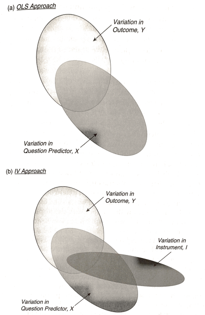


Figure 10.1 Graphical analog for the population variation and covariation among outcome Y, potentially endogenous question predictor X, and instrument, I, used for distinguishing the OLS and IV approaches (a) OLS Approach: bivariate relationship between Y and X, (b) IV Approach: trivariate relationship among Y, X, and I.

Source: Murnane & Willett chapter 11

Instrument for family size

Consider the following regression intended to estimate the effect of family size on a child's years of schooling (Becker's "quantity-quality tradeoff"):

$$educ = \beta_0 + \beta_1 famsize + u$$

Family size is likely endogenous, correlated with omitted variables related to both family size and educational attainment. Can we find an instrumental variable Z that satisfies the identification assumptions?

- $Cov(Z, famsize) \neq 0$
- $Cov(Z, u) = 0$

Instrument for family size

Some potential candidates:

- Twins: an "exogenous shock" to family size (Angrist, Lavy, & Schlosser, 2010)
- Sibling sex composition: families with two boys or two girls are more likely to have a third child (Angrist & Evans, 1998)
- Female child as first child, especially in countries where having a male child is culturally important

Instrument for family size

Issues to consider:

- First stage: both have a positive association with family size. See Table 3.4 (later).
- Exclusion restriction: can we assume the presence of twins or two siblings of the same sex has no effect on educational attainment, except through family size? Consider effects of gender mix on educational inputs.
- Independence: can we assume that the presence of twins or two siblings of the same sex is uncorrelated with omitted variables in u? Consider in-vitro fertilization.
- Monotonicity: instrument only affects compliers in the same direction. Consider case where some parents respond to having same-sex siblings by having *fewer* kids, not more.

Instrument for family size

If the identification assumptions hold, we again have two regressions to estimate:

$$educ_i = \alpha + \rho Z_i + w_i$$

$$famsize_i = \gamma + \phi Z_i + v_i$$

These are the reduced form and first stage equations. The IV estimate of β_1 in the structural model can be obtained as $\hat{\rho}/\hat{\phi}$.

Is $\hat{\rho}$ insignificant? If there is no reduced form effect, there isn't a LATE!

Precision and IV

Again consider the asymptotic variance of $\hat{\beta}_{1,IV}$:

$$Var(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

- A weak first stage (small $\rho_{x,z}^2$ above) can yield large standard errors
- A large sample size (n) can help counteract this.

In the family size example, twins have a larger first stage, but there are fewer cases of twins. The gender mix first stage is smaller, but the number of treated cases is much larger. Again, see Table 3.4 (later).

Weak instruments

We cannot verify the exclusion restriction $Cov(Z, u) = 0$ in a simple regression, so there always remains some possibility this is violated. For large sample bias, we *hope* this correlation is small. In large samples:

$$\text{plim}\hat{\beta}_{1,IV} = \beta_1 + \frac{Cov(Z, u)}{Cov(Z, X)}$$

or

$$\text{plim}\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(Z, u)}{Corr(Z, X)} \times \frac{\sigma_u}{\sigma_x}$$

Even when $Corr(Z, u)$ is small, the bias can be large if $Corr(Z, X)$ is small. A large sample does *not* help in this case. More on this later.

Two stage least squares

In the family size example, we considered two excluded instruments: twin births and the gender mix of the first two children. If both instruments are valid, can we use them together to estimate the causal effect of *famsize*?

Yes: using **two-stage least squares** (2SLS)

- Combines multiple instrumental variables efficiently.
- Can include covariates when the instrument is imperfect (i.e., where the assumption $Cov(Z, u) = 0$ does not hold without conditioning on the covariates).

Two stage least squares—single instrument

Stage 1: get predicted values from first stage regression with single instrument Z_i

$$\widehat{X_i} = \widehat{\gamma} + \widehat{\phi} Z_i$$

Stage 2: use predicted values in the structural equation

$$Y_i = \beta_0 + \beta_1 \widehat{X_i} + u_i$$

In this example, the 2SLS estimator of β_1 is equivalent to the IV estimator (the ratio of the reduced form slope to the first stage slope, or $\hat{\rho}/\hat{\phi}$)

Two stage least squares—single instrument

Can also include covariates (here, X_2):

Stage 1: get predicted values

$$\widehat{X_{1i}} = \widehat{\gamma} + \widehat{\phi} Z_i + \widehat{\alpha_1} X_{2i}$$

Stage 2: use predicted values in the structural equation

$$Y_i = \beta_0 + \beta_1 \widehat{X_{1i}} + \alpha_2 X_{2i} + u_i$$

X_{1i} is the endogenous explanatory variable X_{2i} is an exogenous covariate. Note the covariate(s) need to be included in <u>both</u> stages!!

Two stage least squares—multiple instruments

Stage 1: get predicted values

$$\widehat{X_{1i}} = \widehat{\gamma} + \widehat{\phi_1} Z_{1i} + \widehat{\phi_2} Z_{2i} + \widehat{\alpha_1} X_{2i}$$

Stage 2: use predicted values in the structural equation

$$Y_i = \beta_0 + \beta_1 \widehat{X_{1i}} + \alpha_2 X_{2i} + u_i$$

Note the covariate(s) needs to be included in <u>both</u> stages!! The two instruments Z_{1i} and Z_{2i} are only in the first stage.

With multiple instruments, the estimator of β_1 in Stage 2 is a weighted average of estimators that use each instrument individually.

Multiple instruments and multiple endogenous variables

Stage 1: get predicted values *for each endogenous variable*

$$\widehat{X_{1i}} = \widehat{\gamma_1} + \widehat{\phi_{11}} Z_{1i} + \widehat{\phi_{12}} Z_{2i} + \widehat{\alpha_{11}} X_{3i}$$

$$\widehat{X_{2i}} = \widehat{\gamma_2} + \widehat{\phi_{21}} Z_{1i} + \widehat{\phi_{22}} Z_{2i} + \widehat{\alpha_{21}} X_{3i}$$

Stage 2: use predicted values in the structural equation

$$Y_i = \beta_0 + \beta_1 \widehat{X_{1i}} + \beta_2 \widehat{X_{2i}} + \alpha X_{3i} + u_i$$

Here X_{1i} and X_{2i} are endogenous explanatory variables. X_{3i} is an exogenous covariate.

Two stage least squares: family size example

First stages:

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320 (.052)	.437 (.050)			.449 (.050)
Same-sex sibships			.079 (.012)	.073 (.010)	.076 (.010)
Male		−.018 (.010)		−.020 (.010)	−.020 (.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

Two stage least squares: family size example

OLS vs. 2SLS

Table 3.5
OLS and 2SLS estimates of the quantity-quality trade-off

Dependent variable	OLS estimates (1)	2SLS estimates: Twins instruments (2)	2SLS estimates: Same-sex instruments (3)	2SLS estimates: Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

Two stage least squares: family size example

Some things to note:

- Change in sign from OLS to IV estimates
- Lack of precision in IV estimates
- Some improvement in precision when multiple instruments are used

Stata example 1

Implementing 2SLS in Stata:

```
ivregress 2sls y1 x1 x2 x3 (y2 = z1), first
```

- 2sls is the estimation method (`ivregress` has others)
- *y1* is the dependent variable
- *x1-x3* are the exogenous explanatory variables
- *y2* is the endogenous explanatory variable
- *z1* is the excluded instrument
- This example has only one endogenous explanatory variable and one excluded instrument
- `first` option displays first-stage results (recommended)

See in-class example 1 using *Card.dta*

Some notes on terminology in IV

- A *structural equation* typically has endogenous explanatory variables. It is usually the causal relationship of interest.
- *Reduced form equations* only have exogenous explanatory variables on the RHS. We referred to the equation with Y on the LHS and Z on the RHS as the "reduced form," but the first-stage equation is technically also a "reduced form" equation.
- The *Wald estimator* is the name given to estimation with one endogenous variable and one excluded instrument, in which the reduced form slope coefficient is divided by the first stage slope coefficient. (You most often see this term used when Z and X are binary).

Some notes on terminology in IV

An IV model is:

- *just identified* if the number of excluded instruments in the vector Z_i equals the number of endogenous explanatory variables.
- *under-identified* if the number of excluded instruments in the vector Z_i is less than the number of endogenous explanatory variables.
- *over-identified* if the number of excluded instruments in the vector Z_i is more than the number of endogenous explanatory variables.

If the model is under-identified, no consistent IV estimator exists. In other words, you need at least as many excluded instruments as endogenous explanatory variables.

Testing IV assumptions

"Anyone using an IV estimator should conduct and report tests of the following:" (Nichols, 2007)

- Strength of first stage (test for weak instruments)
- Test for endogeneity
- Overidentification test—where possible (requires > 1 instrument)
- Misspecification of functional form

See Cameron & Trivedi chapter 6 for more details and examples.

Strength of first stage

Remember, weak instruments mean less precision *and* bias (even in large samples). How strong should the relationship be?

- After `ivregress` can use `estat firststage` command to get a measure of the strength of association between instruments Z and endogenous explanatory variables X
- Use the F statistic for joint significance of the instruments in the first stage. $\mathbf{F > 10}$ is historically used as the rule-of-thumb for rejecting a weak instrument.
- aka "under-identification test"
- See Cameron & Trivedi (chapter 6) for other related statistics

Strength of first stage

The `estat firststage` output includes a table with "rule-of-thumb" critical values from Stock and Yogo (2005) that can be used. These values (measures of the predictive power of the excluded instruments) imply some limit of the bias to a percent of OLS. See in-class example using *Card.dta*

- Lee et al. (2022) find bias in confidence intervals with F statistics as high as 104.7
- 10 is not the magic number, and neither is 104.7—in practice, need to worry more the smaller F is

Strength of first stage

What to do in the case of weak instruments?

- Find better instruments, duh! A transformation of your existing instrument(s) can help in some cases.
- Can use LIML estimation (vs 2SLS) which has more desirable finite sample properties, especially if instruments are weak.
- Also see `condivreg` command for robust inference with weak instruments, or `jive` (Jackknife IV estimator). I haven't used these.
- Anderson-Rubin (1949) confidence intervals that adjust standard errors for weak instruments (see `weakiv` package).

Endogeneity test

We are typically using IV because we believe an explanatory variable in the structural model is endogenous. If that variable is in fact *exogenous*, IV is *consistent* but *less efficient*. Would prefer OLS in this case.

- After `ivregress` can use `estat endogenous` command
- If this follows `ivreg 2sls`, the Durbin-Wu-Hausman test statistic is reported.
- H_0 : explanatory variable(s) is exogenous
- Rejection suggests explanatory variable(s) is endogenous and IV is appropriate. Failure to reject would suggest OLS is preferable.

See in-class example using *Card.dta*

Over-identification test

It is impossible to test the identification assumption $Cov(Z, u) = 0$ in the just-identified case. But one can test for the validity of instruments in the *over-identified* case (more instruments than endogenous variables).

- After `ivregress` can use `estat overid` command
- If this follows `ivreg 2sls`, the Sargan (1958) chi-squared test statistic is reported ($df =$ number of over-identified restrictions)
- H_0 : all instruments are valid
- Rejection can be interpreted as one or more instruments are not valid (or, model was mis-specified to begin with). Note a lack of rejection should not be interpreted as confirmation of validity.

Over-identification test

If you have instruments that are truly exogenous, then it is also true that the squares and cross-products of these instruments are exogenous. Nichols (2007) recommends an overid test in which these are added to the list of instruments.

Fixed effects panel model with IV

Implementing 2SLS in Stata with panel data and fixed effects:

```
xtivreg y1 x1 x2 x3 (y2 = z1), fe small
```

- 2sls is the estimation method
- The panel variables should already be set using `xtset`
- Note `small` option is for small-sample statistics (i.e., report t and F statistics)

Nonlinear IV

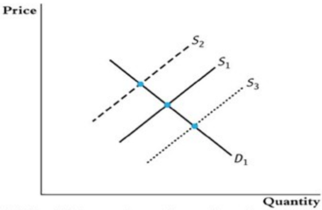
When endogeneous variable is binary:

- It is important not to use a nonlinear model (probit/logit) in the first stage. Z will be related to the error term in the second stage even when Z is valid. (Re: the predicted values $\hat{X}$ will depend on the level of other covariates).
- Can use linear probability model (this is what `ivreg` does).
- There are alternatives: see Huntington-Klein ch. 17, Wooldridge, `etregress` command.
- For binary outcome variables, see `ivprobit` with two step option
- For binary outcome and treatment, see `biprobit` as an alternative.

Applications of IV

Early applications: identifying parameters of the demand curve. Observed prices reflect simultaneous effects of supply and demand. Need instrument Z that affects price that is unrelated to other demand-side factors in u.

$$D = \alpha_0 + \alpha_1 P + u$$



(c) Equilibrium price and quantity when only the supply curve shifts

Applications of IV

Angrist and Krueger (1991) - compulsory schooling

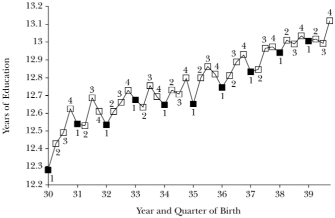
- Let X be years of schooling and Y be (log) earnings. Can we find an instrument Z that affects years of schooling but not earnings (except through years of schooling)?
- Consider a child born in Q4: starts school before turning age 6. At age 16, has completed 10+ years of school
- Consider a child born in Q1: starts school the following year. At age 16, as completed 9+ years of school

Prediction: for kids who drop out at 16, those born in Q1 have less completed schooling.

Applications of IV

Angrist and Krueger (1991) - compulsory schooling

Figure 1
Mean Years of Completed Education, by Quarter of Birth

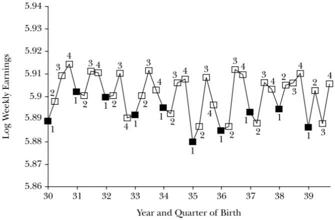


Source: Authors' calculations from the 1980 Census.

Applications of IV

Angrist and Krueger (1991) - compulsory schooling

Figure 2
Mean Log Weekly Earnings, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

Applications of IV

- Vietnam draft lottery number (Angrist & Krueger 1992)
- Proximity to colleges (Card 1993)
- Weather: rainfall, snow, temperature (see Mellon 2020 on rainfall)
- Election cycles and effect on policing
- Terrorist attacks
- Dorm room assignments
- Random judge or case worker assignment
- Immigration (e.g., Mariel Boat Lift)
- Topological features (Hoxby 2000)
- "Shift-share" designs: applying aggregate shocks to local baseline measures

COVID-19 as an instrument?

Instrumental variables often emerge from "natural experiments" in which an exogenous force changes behavior or exposure to "treatment" in unanticipated ways. Might COVID-19 be used as an instrument for, say, the use of online instruction or some other policy change?

Most likely not (Bacher-Hicks & Goodman, 2020).

COVID-19 as an instrument?

Consider the following regression:

$$Y_i = \beta_0 + \beta_1 \text{online}_i + u_i$$

where $\text{online}_i = 1$ if student i is receiving online instruction. Generally we would be concerned about self-selection into remote instruction and OVB.

However, COVID-19 forced many learners into remote instruction. One might even argue there was some idiosyncratic variation in this change, given variation in district decisions. Perhaps local variation in COVID severity could be used as an IV for remote instruction?

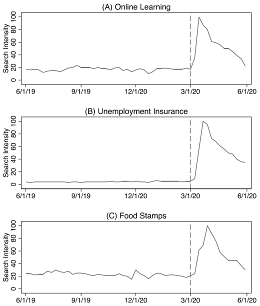
COVID-19 as an instrument?

Problems:

- Exclusion restriction: the instrument should be unrelated to other contemporaneous changes that might affect Y. Unlikely for lots of reasons! Negative effect of the pandemic, compensating behavior by parents.
- External validity: even if we could identify the causal effect of remote instruction via the pandemic, the LATE estimates are unlikely to translate to other times and places.

COVID-19 as an instrument?

Figure 2. Search Intensity for Online Learning and Economic Indicators



Notes: This figure presents the weekly popularity for the keywords "Online Learning", "Unemployment Insurance", and "Food Stamps" in the United States from June 1, 2019 through June 1, 2020. Keyword trend data come from Google Trends. Popularity is indexed relative to the week of March 29, 2020, the week in which "Online Learning" was most popular across the weeks in this sample.

Measurement error

IV is often used as a solution to the "errors-in-variables" problem: when the variables we are using in a regression are measured with error. Types of measurement error:

1. Measurement error in the dependent variable: observe y instead of y^*
2. Measurement error in an explanatory variable: observe x instead of x^*

See Wooldridge chapter 9.

Measurement error in the dependent variable

We observe y instead of y^*:

$$y^* = \beta_0 + \beta_1 x_1 + u$$
$$y = y^* + e_0$$

The regression we are forced to estimate:

$$y = \beta_0 + \beta_1 x_1 + \underbrace{(u + e_0)}_{v}$$

As long as measurement error e_0 is uncorrelated with the explanatory variable x_1, the OLS esimator of β_1 is unbiased and consistent. (The same logic applies to regressions with multiple explanatory variables).

Measurement error in the dependent variable

If u and e_0 are uncorrelated, then $Var(v) = Var(u + e_0) = \sigma_u^2 + \sigma_{e0}^2$ which is greater than σ_u^2. This means greater variance in the OLS estimator (larger standard errors).

It is often assumed that e_0 is uncorrelated with the explanatory variables, but is this a reasonable assumption?

Measurement error in an explanatory variable

We observe x_1 instead of x_1^*:

$$y = \beta_0 + \beta_1 x_1^* + u$$
$$x_1 = x_1^* + e_1$$

The regression we are forced to estimate:

$$y = \beta_0 + \beta_1(x_1 - e_1) + u$$
$$y = \beta_0 + \beta_1 x_1 + \underbrace{(u - \beta_1 e_1)}_{v}$$

If measurement error e_1 is uncorrelated with the *observed* measure x_1, then OLS is unbiased and consistent. (The variance of the estimator will again be larger). This is an unusual assumption, however.

Classical measurement error

It is more reasonable to think e_1 is uncorrelated with the *unobserved* measure x_1^*. This would arise when the observed x_1 is the sum of the true explanatory variable and random noise: "classical measurement error".

Then there is necessarily correlation between x_1 and e_1:

$$Cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = \sigma_{e1}^2$$

This means there is correlation between x_1 and v in the regression we estimate:

$$Cov(x_1, v) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

NOTE: these use the rule $Cov(X, Y) = E(XY) - E(X)E(Y)$.

Classical measurement error

The large sample bias of the OLS estimator $\hat{\beta}_1$ is shown here:

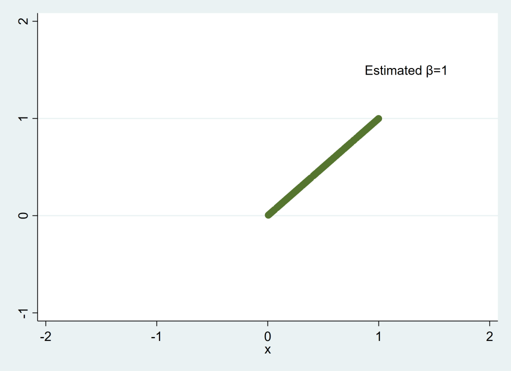
$$\begin{aligned}
\text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{Cov(x_1, v)}{Var(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \\
&= \beta_1 \left(1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}\right) \\
&= \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}\right)
\end{aligned}$$

β_1 is being multiplied by a number < 1. This is called **attenuation bias**. The estimator is biased toward zero. The amount of bias depends on the variability in x_1^* (signal) versus the variability in e_1 (noise).

Example

Let x be a random draw from the uniform (0, 1) distribution, and let the true model be $y = x$. Create a simulated dataset with N=300.

Example - no measurement error

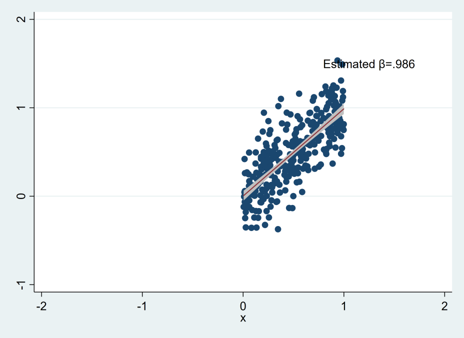


Example

Introduce some noise into the *dependent* variable.

- $y = y^* + e_0$ where $e_0 \sim N(0, 0.25^2)$

Example - measurement error in the dependent variable



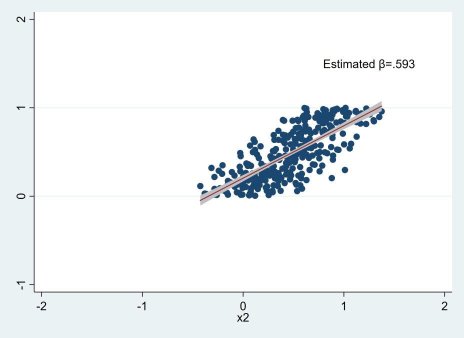
Confidence interval for slope of (.889, 1.084).

Example

Introduce some noise into the *explanatory* variable (classical measurement error), and regress the clean y on the x with measurement error.

- $x = x^* + e_1$ where $e_1 \sim N(0, 0.25^2)$

Example - measurement error in the explanatory variable



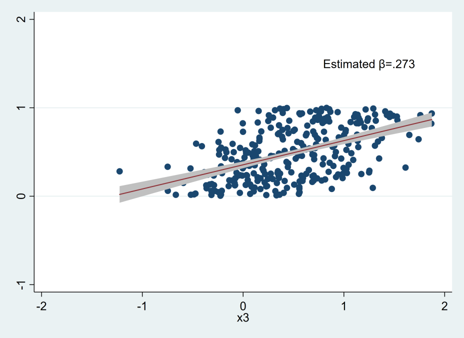
Confidence interval for slope of (.535, .651).

Example

Increase the amount of noise in the explanatory variable, and regress the clean y on the x with measurement error.

- $x = x^* + e_2$ where $e_2 \sim N(0, 0.50^2)$

Example - measurement error in the explanatory variable



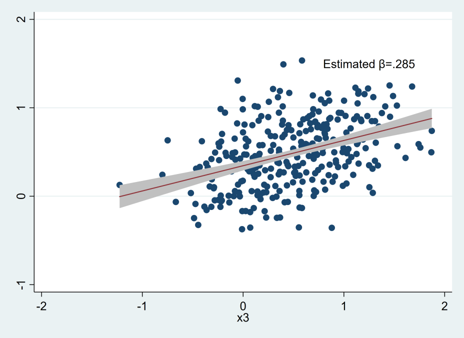
Confidence interval for slope of (.221, .326).

Example

Repeat, but regress the y with measurement error on the x with measurement error.

- $y = y^* + e_0$ where $e_0 \sim N(0, 0.25^2)$
- $x = x^* + e_2$ where $e_2 \sim N(0, 0.50^2)$

Example - measurement error in both



Confidence interval for slope of (.213, .358).

Using an instrumental variable with measurement error

The regression we would like to estimate (versus what we actually estimate) is:

$$\begin{aligned} y^* &= \beta_0 + \beta_1 x^* + u \\ y - e_0 &= \beta_0 + \beta_1 (x - e_2) + u \\ y &= \beta_0 + \beta_1 x + \underbrace{(u + e_0 - \beta_1 e_2)}_{v} \end{aligned}$$

As shown earlier, $Cov(v, x) \neq 0$, so OLS will be biased and inconsistent.

Using an instrumental variable with measurement error

Suppose we have a third measure z that tells us whether or not the underlying x^* is below or above 0.5. This is a much coarser measure than x^*, but it is known for sure—not measured with error.

- $y = y^* + e_0$ where $e_0 \sim N(0, 0.25^2)$
- $x = x^* + e_2$ where $e_2 \sim N(0, 0.50^2)$
- $z = 0$ if $x^* < 0.5$ and $z = 1$ if $x^* > 0.5$

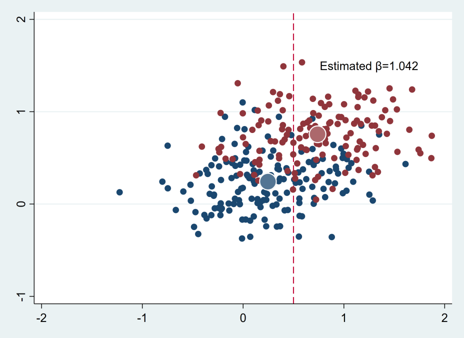
Using an instrumental variable with measurement error

Does z satisfy the key assumptions of a valid instrumental variable?

1. $Cov(z, x) \neq 0$: YES
2. $Cov(z, v) = 0$: YES (conditional on x). z is not correlated with y, except through x

Try 2SLS/IV using z as an instrument for x. That is, regress x on z, get fitted values $\hat{x}$, and then regress y on $\hat{x}$.

Example - using IV



Confidence interval for slope of (.789, 1.295).

Example - using IV

Reduced form:

$$Avg[Y_i|Z_i = 1] - Avg[Y_i|Z_i = 0] = 0.755 - 0.243 = 0.512$$

First stage:

$$Avg[X_i|Z_i = 1] - Avg[X_i|Z_i = 0] = 0.739 - 0.247 = 0.492$$

Instrumental variables (Wald) estimate:

$$\frac{0.512}{0.492} = 1.042$$

Using an instrumental variable with measurement error

Why does this work?

- $\hat{x}$ is "purged" of noise since it only represents variation in x that is explained by z
- The Wald estimate is the change in y associated with a change in $\hat{x}$

$$Avg[X_i|Z_i = 1] - Avg[X_i|Z_i = 0] =$$
$$Avg[X_i^*|Z_i = 1] + Avg[e_2|Z_i = 1] - Avg[X_i^*|Z_i = 0] - Avg[e_2|Z_i = 0]$$

The mean of e_2 does not vary with Z_i, so those terms drop out.

$$Avg[X_i^*|Z_i = 1] - Avg[X_i^*|Z_i = 0]$$

IV solutions in practice

Examples where a second measure can potentially be used as a instrument for the x measured with error:

- Test scores: instrumenting one test measure with another
- Salary reports: obtaining independent reports from the employer and employee
- Family measures: obtaining independent reports from spouses
- Education: two independent reports on educational attainment (Ashenfelter & Krueger, 1994)