Instrumental Variables Regression

Ming-Ching Luoh

Three important threats to internal validity are:

- omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
- simultaneous causality bias (X causes Y, Y causes X);
- errors-in-variables bias (X is measured with error).

Instrumental variables regression can eliminate bias when $\mathrm{E}(u|X) \neq 0$— using an *instrumental variable*, Z.

- Instrumental variable (IV) regression is a general way to obtain a consistent estimator of the unknown coefficients of the population regression function when the regressor, X, is correlated with the error term, u.
- The variation in X has two parts: one part that is correlated with u (the part that causes the problems), and a second part that is uncorrelated with u.
- If you had information that allowed you to isolate the second part, then you could focus on those variations in X that are uncorrelated with u.

- The information about the movements in X that are uncorrelated with u is gleaned from one or more additional variables, called instrumental variables or simply instruments.
- Instrumental variables regression uses these additional variables as tools or “instruments" to isolate the movements in X that are uncorrelated with u, which in turn permit consistent estimation of the regression coefficients.

The IV Estimator with a Single Regressor and a Single Instrument

The IV Model and Assumptions

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

- If X_i and u_i are correlated, the OLS estimator is inconsistent.
- Instrumental variables estimation uses an additional , "instrumental" variable Z to isolate that part of X_i that is uncorrelated with u_i.

Terminology: endogeneity and exogeneity

An *endogenous* variable is one that is correlated with u.
An *exogenous* variable is one that is uncorrelated with u.

Historical note:

- "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, or, a variable subject to simultaneous causality.
- However, this definition is narrow and IV regression can be used to address omitted variable bias and errors-in-variable bias, not just to simultaneous causality bias.

Two conditions for a valid instrument

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

For an instrumental variable (an "instrument") Z to be valid, it must satisfy two conditions:

1. Instrument relevance: $\text{Cov}(Z_i, X_i) \neq 0$
2. Instrument exogeneity: $\text{Cov}(Z_i, u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later), How can you use Z_i to estimate β_1?

The Two Stage Least Squares (TSLS) Estimator

As it sounds, TSLS has two stages— two regressions:

(1) First isolates the part of X that is uncorrelated with u: regress X on Z using OLS.

$$X_i = \pi_0 + \pi_1 Z_i + v_i \tag{1}$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them.
- Compute the predicted values of X_i, $\hat{X}_i$, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1, \cdots, n$.

(2) Replace X_i by $\hat{X}_i$ in the regression of interest:
regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \tag{2}$$

- Because $\hat{X}_i$ is uncorrelated with u_i in large samples, so the first least squares assumption holds.
- Thus β_1 can be estimated by OLS using regression (2).
- This argument relies on large samples (so π_0 and π_1 are well estimated using regression (1)).
- This resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Summary:

Suppose you have a valid instrument, Z_i.

- Stage 1:

  Regress X_i on Z_i, obtain the predicted values $\hat{X}_i$.

- Stage 2:

  Regress Y_i on $\hat{X}_i$, the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

Then $\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

Another approach:

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

Thus,

$$\begin{aligned} & \mathrm{Cov}(Y_i, Z_i) \\ = \; & \mathrm{Cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\ = \; & \mathrm{Cov}(\beta_0, Z_i) + \mathrm{Cov}(\beta_1 X_i, Z_i) + \mathrm{Cov}(u_i, Z_i) \\ = \; & \beta_1 \mathrm{Cov}(X_i, Z_i) \end{aligned}$$

where $\mathrm{Cov}(u_i, Z_i) = 0$ (instrument exogeneity).

Thus.

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(X_i, Z_i)}$$

$$\beta_1 = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(X_i, Z_i)}$$

The IV estimator replaces these population covariances with sample covariances.

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

s_{YZ} and s_{XZ} are the sample covariances.
This is the TSLS estimator - just a different derivation.

Why $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$?

$$
\begin{aligned}
\hat{\beta}_1^{TSLS} &= \frac{s_{\hat{X}Y}}{s^2_{\hat{X}}} \quad (\text{ from 2nd stage}) \\
&= \frac{\hat{\pi}_1 s_{ZY}}{\hat{\pi}_1^2 s_Z^2} \quad (\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i) \\
&= \frac{s_{ZY}}{\hat{\pi}_1 s_Z^2} \quad \left(\hat{\pi}_1 = \frac{s_{ZX}}{s_Z^2}\right) \\
&= \frac{s_{YZ}}{s_{XZ}}
\end{aligned}
$$

Third explanation: Derivation from the "reduced form"

The "reduced form" relates Y to Z and X to Z:

$$\begin{aligned} X_i &= \pi_0 + \pi_1 Z_i + v_i \\ Y_i &= \gamma_0 + \gamma_1 Z_i + w_i \end{aligned}$$

where w_i is an error term. Because Z is exogenous, Z is uncorrelated with both v_i and w_i.

- **The idea**: A unit change in Z_i results in a change in X_i of π_1 and a change in Y_i of γ_1. Because that change in X_i arises from the exogenous change in Z_i, that change in X_i is exogenous. Thus an exogenous change in X_i of π_1 units is associated with a change in Y_i of γ_1 units— so the effect on Y of an exogenous change in X is $\beta_1 = \frac{\gamma_1}{\pi_1}$ units.

The math:

$$\begin{aligned} X_i &= \pi_0 + \pi_1 Z_i + v_i \\ Y_i &= \gamma_0 + \gamma_1 Z_i + w_i \end{aligned}$$

Solve the X equation for Z:

$$Z_i = -\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1} X_i - \frac{1}{\pi_1} v_i$$

Substitute this into the Y equation and collect terms:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 Z_i + w_i \\ &= \gamma_0 + \gamma_1 \left(-\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1} X_i - \frac{1}{\pi_1} v_i \right) + w_i \\ &= \left(\gamma_0 - \frac{\pi_0 \gamma_1}{\pi_1} \right) + \frac{\gamma_1}{\pi_1} X_i + \left(w_i - \frac{\gamma_1}{\pi_1} v_i \right) \\ &\equiv \beta_0 + \beta_1 X_i + u_i \end{aligned}$$

where $\beta_1 = \frac{\gamma_1}{\pi_1} \equiv \frac{\frac{\text{Cov}(Y,Z)}{\text{Var}(Z)}}{\frac{\text{Cov}(X,Z)}{\text{Var}(Z)}} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(X,Z)}$.

Consistency of the TSLS estimator

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

The sample covariances are consistent: $s_{YZ} \xrightarrow{p} \text{Cov}(Y, Z)$ and $s_{XZ} \xrightarrow{p} \text{Cov}(X, Z)$. Thus,

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta_1$$

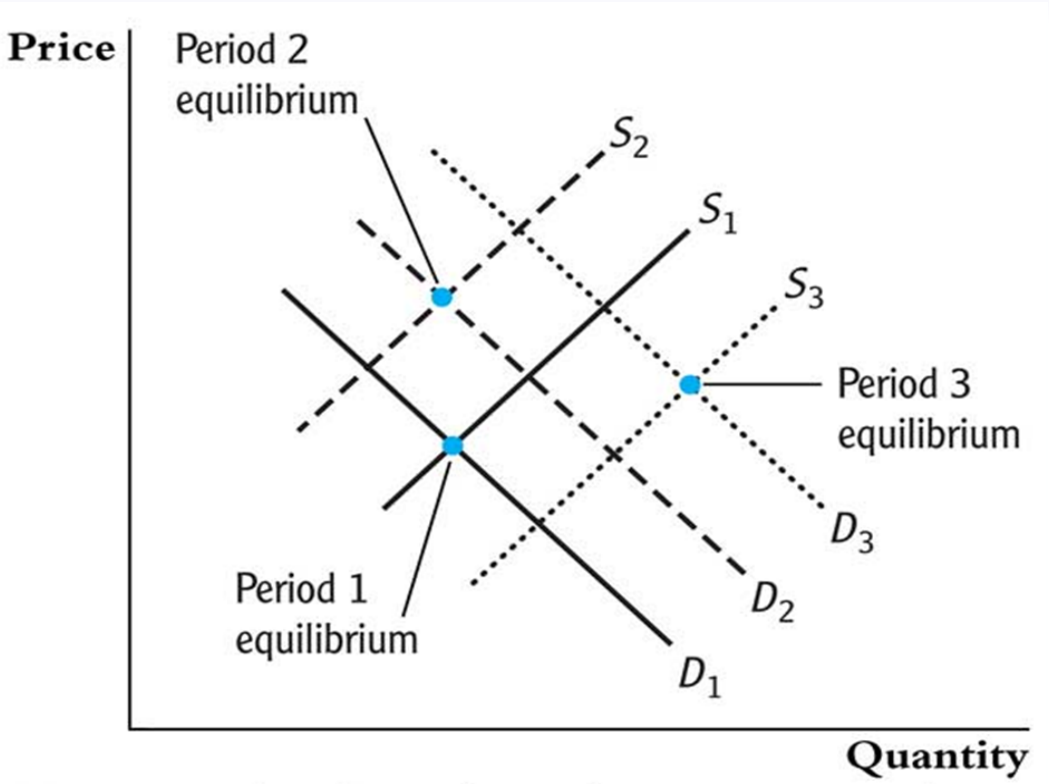
- The instrument relevance condition, $\text{Cov}(X, Z) \neq 0$, ensures that you don't divide by zero.

Example #1: Supply and demand for butter

IV regression was originally developed to estimate demand elasticities for agricultural goods, for example butter:

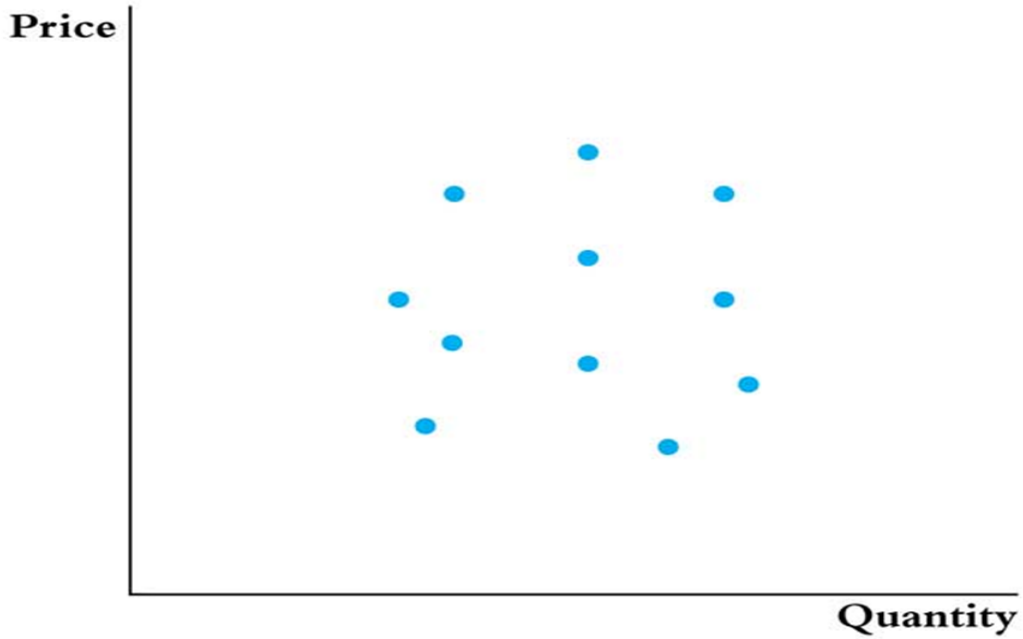
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- β_1 = price elasticity of butter = percent change in quantity for a 1% change in price.
- Data: observations on price and quantity of butter for different years.
- The OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ suffers from simultaneous causality bias (why?)



- Simultaneous causality bias in the OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ arises because price and quantity are determined by the interaction of demand *and* supply.

This interaction of demand and supply produces

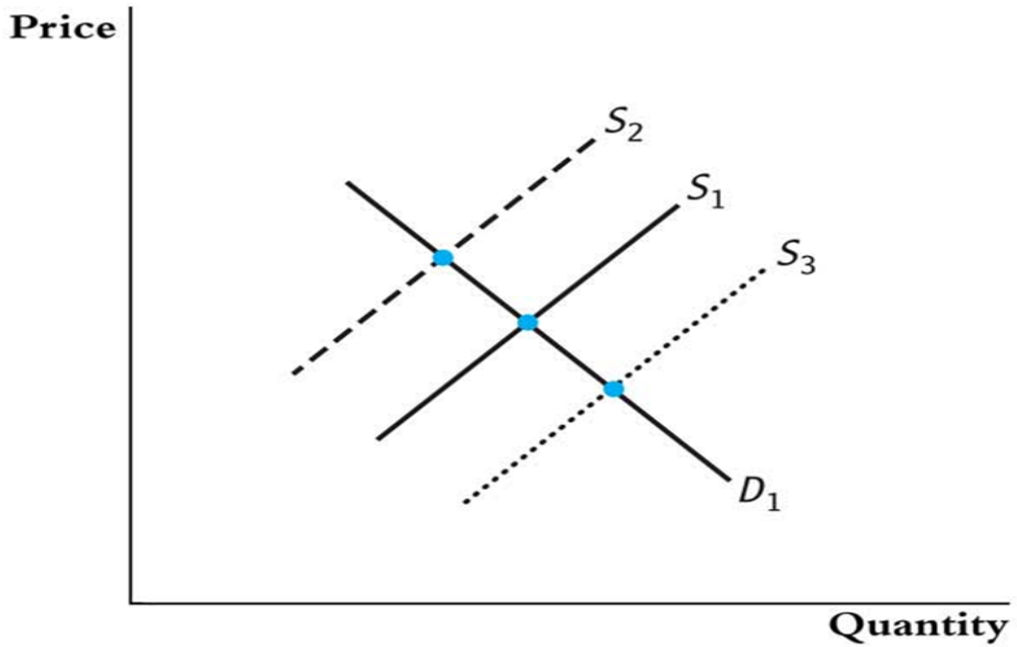


(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?

No!

What would you get if only supply shifted?



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

- TSLS estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply.
- Z is a variable that shifts supply but not demand.

TSLS in the supply-demand example:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Let Z = rainfall in dairy-producing regions. Is Z a valid instrument?

(1) Exogenous? $\text{Cov}(\text{rain}_i, u_i) = 0$?
Plausibly: whether it rains in dairy-producing regions shouldn't affect demand.

(2) Relevant? $\text{Cov}(\text{rain}_i, \ln(P_i^{butter})) \neq 0$?
Plausibly: insufficient rainfall means less grazing means less butter.

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

$Z_i = \text{rain}_i$ = rainfall in dairy-producing regions.

- Stage 1: regress $\ln(P_i^{butter})$ on rain_i, get $\widehat{\ln(P_i^{butter})}$.

  $\widehat{\ln(P_i^{butter})}$ isolates changes in log price that arise from supply (part of supply, at least).

- Stage 2: regress $\ln(Q_i^{butter})$ on $\widehat{\ln(P_i^{butter})}$.

  The regression counterpart of using shifts in the supply curve to trace out the demand curve.

Example #2: Test scores and class size

- The California regressions still could have omitted variable bias (e.g. parental involvement).
- This bias could be eliminated by using IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:
  (1) relevant: $\text{Cov}(Z_i, STR_i) \neq 0$.
  (2) exogenous: $\text{Cov}(Z_i, u_i) = 0$.

Here is a (hypothetical) instrument:

- some districts, randomly hit by an earthquake, “double up" classrooms:

  $Z_i = Quake_i = 1$ if hit by quake, $= 0$ otherwise.
- Do the two conditions for a valid instrument hold?
- The earthquake makes it as if the districts were in a random assignment experiment. Thus the variation in STR arising from the earthquake is exogenous.
- The first stage of TSLS regresses STR against $Quake$, thereby isolating the part of STR that is exogenous (the part that is “as if" randomly assigned).

Inference using TSLS

- In large samples, the sampling distribution of the TSLS estimator is normal.
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. $\pm 1.96SE$.
- The idea behind the large-sample normal distribution of the TSLS estimator is that - like all the other estimators we have considered— it involves an average of mean zero $i.i.d.$ random variables, to which we can apply the CLT.
- See SW App. 12.3 for the details.

Sampling Distribution of the TSLS Estimator

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

Substitute in

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

and simplify,

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z}) \\
= \quad &\frac{1}{n-1}\sum_{i=1}^{n}\left(\beta_1(X_i - \bar{X}) + (u_i - \bar{u})\right)(Z_i - \bar{Z}) \\
= \quad &\beta_1\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \\
&+\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})
\end{aligned}$$

Thus

$$\begin{aligned}\hat{\beta}_1^{TSLS} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Z_i-\bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})} \\ &= \beta_1 + \frac{\frac{1}{n-1}\sum_{i=1}^{n}(u_i-\bar{u})(Z_i-\bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})}\end{aligned}$$

Subtract β_1 from each side and we get,

$$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(u_i-\bar{u})(Z_i-\bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})}$$

Multiplying through by $\sqrt{n-1}$ and making the approximation that $\sqrt{n-1} \simeq \sqrt{n}$ yields:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) \simeq \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i-\bar{u})(Z_i-\bar{Z})}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})}$$

- First consider the numerator, in large samples,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z}) \xrightarrow{d} N\left(0, \mathrm{Var}\left[(Z - \mu_z)u\right]\right)$$

- Next consider the denominator:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \mathrm{Cov}(X, Z) \text{ by the LLN}$$

where $\mathrm{Cov}(X, Z) \neq 0$ because the instrument is relevant by assumption.

Put these together:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) \simeq \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z}) \xrightarrow{d} N(0, \mathrm{Var}[(Z - \mu_z)u])$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \mathrm{Cov}(X, Z)$$

$$\hat{\beta}_1^{TSLS} \xrightarrow{d} N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$$

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\mathrm{Var}[(Z - \mu_z)u]}{[\mathrm{Cov}(X, Z)]^2}$$

$$\hat{\beta}_1^{TSLS} \xrightarrow{d} N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$$

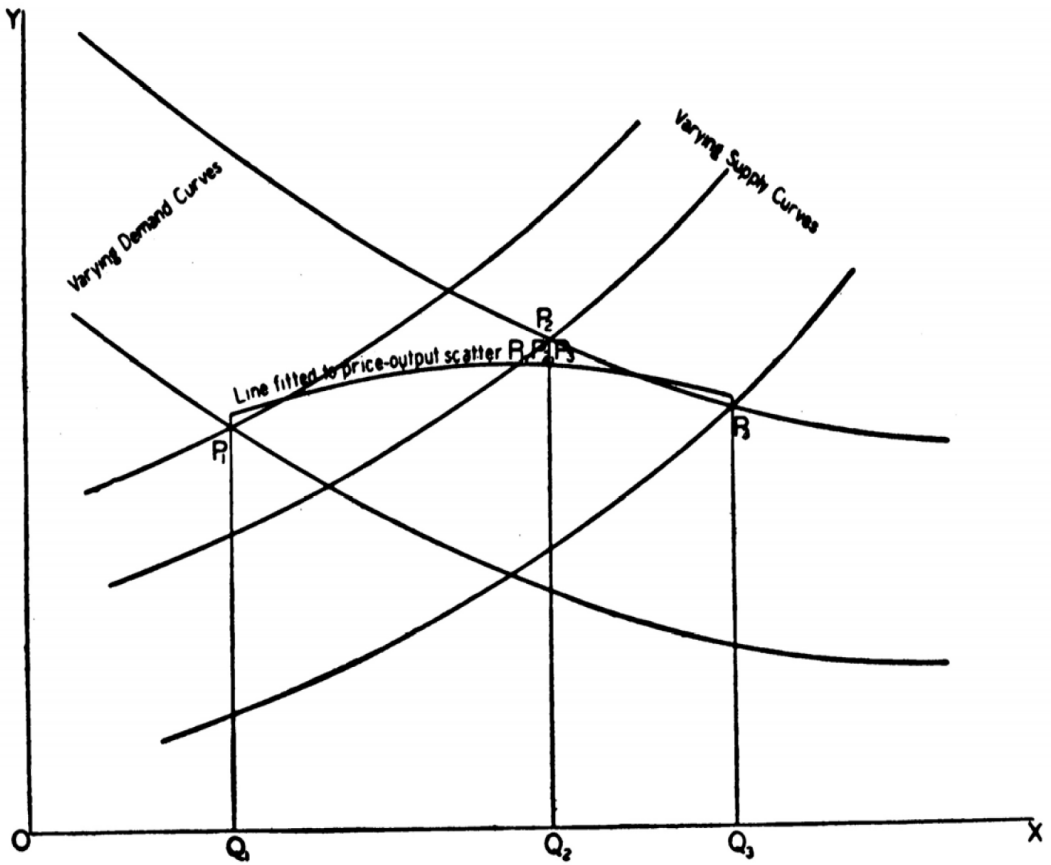
- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples.
- This all assumes that the instruments are valid - we'll discuss what happens if they aren't valid later.
- Important note on standard errors:
  - The OLS standard errors from the second stage regression are not correct— they don't take into account the estimation in the first stage ($\hat{X}_i$ is estimated).
  - Instead, use a single specialized command that computes the TSLS estimator and the correct SEs.
  - As usual, use heteroskedasticity-robust SEs.

A complete digression: The early history of IV regression

- How much money would be raised by an import tariff on animal and vegetable oils (butter, flaxseed oil, soy oil, etc.)?
- To do this calculation you need to know the elasticities of supply and demand, both domestic and foreign.
- This problem was first solved in Appendix B of Wright (1928), "The Tariff on Animal and Vegetable Oils."

Figure 4, p. 296, from Appendix B (1928):

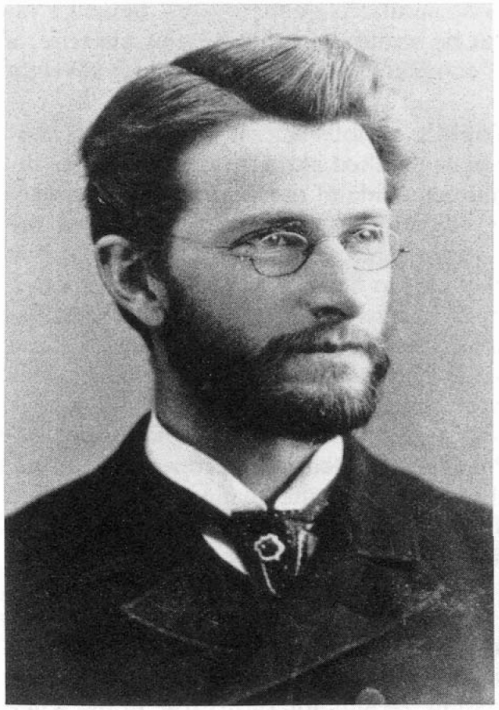
FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Who wrote Appendix B of Philip Wright (1928)?

··· this appendix is thought to have been written with or by his son, Sewall Wright, an important statistician. (SW, p. 425)

Who were these guys and what's their story?



Philip Wright (1861-1934)

itinerant economist and bad poet

MA Harvard, Econ, 1887

Sewall Wright (1889-1988)

pathbreaking genetic statistician

ScD Harvard, Biology, 1915

Derivation of the IV estimator in Wright (1928, p. 314)

Now multiply each term in this equation by A (the corresponding deviation in the price of a substitute) and we shall have:

$$eA \times P = A \times O - A \times S_1.$$

Suppose this multiplication to be performed for every pair of price-output deviations and the results added, then:

$$e \sum A \times P = \sum A \times O - \sum A \times S_1 \text{ or } e = \frac{\sum A \times O - \sum A \times S_1}{\sum A \times P}.$$

But A was a factor which did not affect supply conditions; hence it is uncorrelated with S_1; hence $\sum A \times S_1 = 0$; and hence $e = \frac{\sum A \times O}{\sum A \times P}$.

Who wrote this?

Summary Statistics: selected words, constructions

	Philip		Sewall			Appendix B	
	mean	standard deviation	mean	standard deviation	*t*	mean	standard deviation
noun followed by coordinating conjunction	26.8	7.0	17.3	4.6	5.55	27.0	5.0
to	29.5	5.8	20.9	6.1	4.79	28.0	8.6
now	1.6	1.5	0.1	0.3	4.74	1.1	1.0
when	2.4	2.1	0.3	0.7	4.72	1.8	1.2
in	22.7	5.3	29.8	5.5	-4.34	18.5	5.8
so	2.1	1.6	0.7	0.8	3.82	2.0	1.7
n	25		20			6	

Notes: The entries in columns 2 and 3 are the mean and standard deviations of the counts, per 1000 words, of the stylometric indicator in column 1 in the 25 blocks undisputedly written by Philip Wright. Columns 4 and 5 contain this information for the 20 blocks undisputedly written by Sewall Wright. The next column contains the two-sample *t*-statistic testing the hypothesis that the mean counts are the same for the two authors. The final two columns contain means and standard deviations for the 6 blocks from Appendix B. Shaded indicators occur in the excerpt in Exhibit 2. Source: J.H. Stock and F. Trebbi, "Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17 (2003), 177 – 194.

Application to the Demand for Cigarettes

- How much will a hypothetical cigarette tax reduce cigarette consumption?
- To answer this, we need the elasticity of demand for cigarettes, that is, β_1 in the regression,

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

- Will the OLS estimator plausibly be unbiased? *Why or why not?*

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Panel data:

- Annual cigarette consumption and average prices paid (including tax).
- 48 continental US states, 1985-1995.

Proposed instrumental variable:

- Z_i = general sales tax per pack in the state = $SalesTax_i$.
- Is this a valid instrument?
  (1) Relevant? $corr(SalesTax_i, \ln(P_i^{cigarettes})) \neq 0$?
  (2) Exogenous? $corr(SalesTax_i, u_i) = 0$?

For now, use data for 1995 only.

- First stage OLS regression:

$$\ln(\widehat{P_i^{cigarettes}}) = 4.63 + .031\, SalesTax_i, n = 48$$

- Second stage OLS regression with correct, heteroskedasticity-robust standard errors.

$$\begin{array}{rcl} \ln(\widehat{Q_i^{cigarettes}}) & = & 9.72 - 1.08 \ln(\widehat{P_i^{cigarettes}}) \\ & & (1.53) \quad (0.32) \end{array}$$

STATA ***Example*****: Cigarette demand, First stage**

Instrument = Z = *rtaxso* = general sales tax (real $/pack)

X *Z*

```
. reg lravgprs rtaxso if year==1995, r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   40.39
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4710
                                                      Root MSE      =  .09394

------------------------------------------------------------------------------
             |               Robust
    lravgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtaxso |   .0307289   .0048354     6.35   0.000     .0209956    .0404621
       _cons |   4.616546   .0289177   159.64   0.000     4.558338    4.674755
------------------------------------------------------------------------------
```

X-hat

```
. predict lravphat;
```

Now we have the predicted values from the 1st stage

Second stage

Y *X-hat*

```
. reg lpackpc lravphat if year==1995, r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   10.54
                                                      Prob > F      =  0.0022
                                                      R-squared     =  0.1525
                                                      Root MSE      =  .22645

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravphat |  -1.083586   .3336949    -3.25   0.002    -1.755279   -.4118932
       _cons |   9.719875   1.597119     6.09   0.000     6.505042    12.93471
------------------------------------------------------------------------------
```

- These coefficients are the TSLS estimates
- The standard errors are wrong because they ignore the fact that the first stage was estimated

Combined into a single command:

```
                        Y          X          Z
. ivregress 2sls lpackpc (lravgprs = rtaxso) if year==1995, vce(robust);

Instrumental variables (2SLS) regression                Number of obs =      48
                                                        Wald chi2(1)  =   12.05
                                                        Prob > chi2   =  0.0005
                                                        R-squared     =  0.4011
                                                        Root MSE      =  .18635

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3122035    -3.47   0.001    -1.695494    -.471679
       _cons |   9.719876   1.496143     6.50   0.000     6.78749     12.65226
------------------------------------------------------------------------------
Instrumented:  lravgprs          This is the endogenous regressor
Instruments:   rtaxso            This is the instrumental varible
------------------------------------------------------------------------------
```

Estimated cigarette demand equation:

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08 \widehat{\ln(P_i^{cigarettes})}$$
$$(1.53) \quad (0.31)$$

Summary of IV Regression with a Single X and Z

- A valid instrument Z must satisfy two conditions:

  (1) relevance: $corr(Z_i, X_i) \neq 0$
  (2) exogeneity: $corr(Z_i, u_i) = 0$

- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$.

- The key idea is that the first stage isolates part of the variation in X that is uncorrelated with u.

- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual.

The General IV Regression Model

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors ($X_1, \cdots, X_k$).
  - multiple included exogenous variables ($W_1, \cdots, W_r$). These need to be included for the usual omitted variables reason.
  - multiple instrumental variables ($Z_1, \cdots, Z_m$). More (relevant) instruments can produce a smaller variance of TSLS: the R^2 of the first stage increases, so you have more variation in $\hat{X}$.

Example: Demand for Cigarettes

- Another determinant of cigarette demand is income; omitting income could result in omitted variable bias.
- Cigarette demand with one X, one W, and 2 instruments (2 Z's):

$$\ln(Q_i^{ciga}) = \beta_0 + \beta_1 \ln(P_i^{ciga}) + \beta_2 \ln(Income_i) + u_i$$

Z_{1i} = general sales tax component only

Z_{2i} = cigarette-specific tax component only

- Other W's might be state effects and/or year effects (in panel data).

The general IV regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

- Y_i is the dependent variable.
- $X_{1i}, \cdots, X_{ki}$ are the endogenous regressors (potentially correlated with u_i).
- $W_{1i}, \cdots, W_{ri}$ are the included exogenous variables or included exogenous regressors (uncorrelated with u_i).
- $\beta_0, \beta_1, \cdots, \beta_{k+r}$ are the unknown regression coefficients.
- $Z_{1i}, \cdots, Z_{mi}$ are the m instrumental variables (the excluded exogenous variables).

Identification:

- In general, a parameter is said to be identified if different values of the parameter would produce different distributions of the data.
- In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (m) and the number of endogenous regressors (k).
- Intuitively, if there are fewer instruments than endogenous regressors, we can't estimate $\beta_1, \cdots, \beta_k$.
- For example, suppose $k = 1$ but $m = 0$ (no instruments)!

The coefficients $\beta_1, \cdots, \beta_k$ are said to be:

- exactly identified if $m = k$.

  There are just enough instruments to estimate $\beta_1, \cdots, \beta_k$.
- overidentified if $m > k$.

  There are more than enough instruments to estimate $\beta_1, \cdots, \beta_k$. If so, you can test whether the instruments are valid (a test of the "overidentifying restrictions") - we'll return to this later.
- underidentified if $m < k$.

  There are too few enough instruments to estimate $\beta_1, \cdots, \beta_k$
  If so, you need to get more instruments!

General IV regression: TSLS, 1 endogenous regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

- Instruments: $Z_{1i}, \cdots, Z_{mi}$.
- First stage:
  - Regress X_1 on all the exogenous regressors: regress X_1 on $W_1, \cdots, W_r, Z_1, \cdots, Z_m$ by OLS.
  - Compute predicted values $\hat{X}_{1i}$, $i = 1, \cdots, n$.
- Second stage:
  - Regress Y on $\hat{X}_1, W_1, \cdots, W_r$ by OLS.
  - The coefficients from this second stage regression are the TSLS estimators, but SEs are wrong.
- To get correct SEs, do this in a single step.

Example: Demand for cigarettes

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

$$\begin{aligned} Z_{1i} &= \text{general sales tax}_i \\ Z_{2i} &= \text{cigarette specific tax}_i \end{aligned}$$

- Endogenous variable: $\ln(P_i^{cigarettes})$ ("one X").
- Included exogenous variable: $\ln(Income_i)$ ("one W").
- Instruments (excluded endogenous variables): general sales tax, cigarette specific tax ("two Zs").
- Is the demand elasticity β_1 overidentified, exactly identified, or underidentified?

Example: Cigarette demand, one instrument

```
              Y       W          X          Z
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =       48
                                                       F(  2,    45) =     8.19
                                                       Prob > F      =   0.0009
                                                       R-squared     =   0.4189
                                                       Root MSE      =   .18957

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495     -.413375    .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112    11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso        STATA lists ALL the exogenous regressors
                                     as instruments - slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

- Running IV as a single command yields correct *SE*s
- Use , r for heteroskedasticity-robust *SE*s

Example: Cigarette demand, two instruments

Y W X Z_1 Z_2

```
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  2,    45) =   16.17
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4294
                                                       Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax    STATA lists ALL the exogenous regressors
                                      as "instruments" - slightly different
                                      terminology than we have been using
------------------------------------------------------------------------------
```

TSLS estimates, Z = sales tax (m = 1)

$$\widehat{\ln(Q_i^{ciga})} = 9.43 - 1.14\widehat{\ln(P_i^{ciga})} + 0.21\ln(Income_i)$$
$$(1.26) \quad (0.37) \qquad (0.31)$$

TSLS estimates, Z = sales tax, cig-only tax (m = 2)

$$\widehat{\ln(Q_i^{ciga})} = 9.89 - 1.28\widehat{\ln(P_i^{ciga})} + 0.28\ln(Income_i)$$
$$(0.96) \quad (0.25) \qquad (0.25)$$

- Smaller *SE*s for $m = 2$. Using 2 instruments gives more information, more variation.
- Low income elasticity (not a luxury good); income elasticity not statistically significantly different from 0.
- Surprisingly high price elasticity.

Implications: Sampling distribution of TSLS

- If the IV regression assumptions hold, then the TSLS estimator is normally distributed in large samples.
- Inference (hypothesis testing, confidence intervals) proceeds as usual.
- Two notes about standard errors.
  - The second stage *SE*s are incorrect because they don't take into account estimation in the first stage; to get correct *SE*s, run TSLS in a single command.
  - Use heteroskedasticity-robust *SE*s, for the usual reason.
- *All this hinges on having valid instruments*.

Checking Instrument Validity

Recall the two requirements for valid instruments:

1. *Relevance*: At least one of the instruments is correlated with X.

2. *Exogeneity*: All the instruments must be uncorrelated with the error term:
   $\text{Cov}(Z_{1i}, u_i) = 0, \cdots, \text{Cov}(Z_{mi}, u_i) = 0$.

What happens if one of these requirements isn't satisfied? How can we check? And what do we do?

Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

First stage regression:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+k} W_{ki} + u_i$$

- The instruments are relevant if at least one of $\pi_1, \cdots, \pi_m$ are nonzero.
- The instruments are said to be **weak** if all the $\pi_1, \cdots, \pi_m$ are either zero or nearly zero.
- **Weak instruments** explain very little of the variation in X, beyond that explained by the W's.

What are the consequences of weak instruments?

Consider the simplest case:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ X_i &= \pi_0 + \pi_1 Z_i + v_i \end{aligned}$$

- The IV estimator is $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$.
- If $\text{Cov}(X, Z)$ is zero or small, then s_{XZ} will be small. With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of $\hat{\beta}_1^{TSLS}$ (and its t-statistic) is not well approximated by its large-n normal approximation.

Why does normal approximation fail?

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- If $\text{Cov}(X, Z)$ is small, small changes in s_{XZ} can induce big changes in $\hat{\beta}_1^{TSLS}$.
- Suppose in one sample you calculate $s_{XZ} = .00001$!
- Thus the large-n normal approximation is a poor approximation to the sampling distribution of $\hat{\beta}_1^{TSLS}$.
- A better approximation is that $\hat{\beta}_1^{TSLS}$ is distributed as the ratio of two correlated normal random variables (see SW App. 12.4).
- If instruments are weak, the usual methods of inference are unreliable - potentially very unreliable.

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

When the instrument is irrelavant, $\text{Cov}(Z_i, X_i) = 0$, the denomenator is approximately

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X}) &\simeq \frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu_Z)(X_i - \mu_X) \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} r_i = \bar{r}\end{aligned}$$

Let $\sigma_r^2 = \text{Var}\left[(Z_i - \mu_Z)(X_i - \mu_X)\right]$, $\sigma_{\bar{r}}^2 = \frac{\sigma_r^2}{n}$.

Let $q_i = (Z_i - \mu_Z)u_i$, $\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i$, $\sigma_q^2 = \mathrm{Var}\left[(Z_i - \mu_Z)u_i\right]$, $\sigma_{\bar{q}}^2 = \frac{\sigma_q^2}{n}$, then in large samples,

$$\hat{\beta}_1^{TSLS} \cong \beta_1 + \frac{\bar{q}}{\bar{r}} = \beta_1 + \left(\frac{\sigma_{\bar{q}}}{\sigma_{\bar{r}}}\right)\left(\frac{\frac{\bar{q}}{\sigma_{\bar{q}}}}{\frac{\bar{r}}{\sigma_{\bar{r}}}}\right)$$

$$= \beta_1 + \left(\frac{\sigma_q}{\sigma_r}\right)\left(\frac{\frac{\bar{q}}{\sigma_{\bar{q}}}}{\frac{\bar{r}}{\sigma_{\bar{r}}}}\right)$$

- If the instrument is irrelevant, $\mathrm{E}(r_i) = \mathrm{Cov}(Zi, X_i) = 0$, then $\bar{r}$ is the sample average of the random variable r_i, $i = 1, \cdots, n$, which are i.i.d, have variance σ_r^2, and have a mean of zero.
- It follows that the central limit theorem applies to $\bar{r}$. $\frac{\bar{r}}{\sigma_{\bar{r}}}$ is approximately distributed $N(0, 1)$.

- Therefore, in large samples, the distribution of $\hat{\beta}_1^{TSLS} - \beta_1$ is the distribution of aS, where $a = \frac{\sigma_q}{\sigma_r}$ and S is the ratio of two random variables, each of which has a standard normal distribution. And because X_i and u_i are correlated, these two normal random variables are correlated.
- The large-sample distribution of the TSLS estimator when the instrument is irrelevant is complicated. In fact, it is centered on the probability limit of the OLS estimator.
- Thus, when the instrument is irrelevant, TSLS does not eliminate the bias in OLS, and has a non-normal distribution even in large samples.

Measuring the strength of instruments in practice:
The first-stage F-statistic

- The first stage regression (one X):
  Regress X on $Z_1, \cdots, Z_m, W_1, \cdots, W_k$.
- Totally irrelevant instruments $\Leftrightarrow$ all the coefficients on $Z_1, \cdots, Z_m$ are zero.
- The first-stage F-statistic tests the hypothesis that $Z_1, \cdots, Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

Checking for weak instruments with a single X

- Compute the first-stage F-statistic.
  Rule-of-thumb: If the first stage F-statistic is less than 10, then the set of instruments is weak.
- If so, the TSLS estimator will be biased, and statistical inferences can be misleading.
- Note that simply rejecting the null hypothesis of that the coefficients on the Z's are zero is not enough— we actually need substantial predictive content for the normal approximation to be a good one. (see SW App. 12.5)

- Let β_1^{OLS} denote the probability limit of the OLS estimator $\hat{\beta}_{OLS}$, and let $\beta_1^{OLS} - \beta_1$ denote the asympototic bias of the OLS estimator.

- It is possible to show that the bias of the TSLS is approximately

$$\mathrm{E}(\hat{\beta}_1^{TSLS}) - \beta_1 \approx \frac{\beta_1^{OLS} - \beta_1}{\mathrm{E}(F) - 1}$$

  where $\mathrm{E}(F)$ is the expectation of the first-stage F-statistic.

- If $\mathrm{E}(F) = 10$, then the bias of TSLS, relative to the bias of OLS, is approximately 1/9, or just over 10%, which is small enough to be acceptable in many applications.

What to do if you have weak instruments?

- Get better instruments (!)
- If you have many instruments, some are probably weaker than others, then it's a good idea to drop the weaker ones (dropping an irrelevant instrument will increase the first-stage F).

Checking Assumption #2: Instrument Exogeneity

- Instrument exogeneity: All the instruments are uncorrelated with the error term:
  $\text{Cov}(Z_{1i}, u_i) = 0, \cdots, \text{Cov}(Z_{mi}, u_i) = 0.$
- If the instruments are not uncorrelated with the error term, the first stage of TSLS doesn't successfully isolate a component of X that is uncorrelated with the error term, so $\hat{X}$ is correlated with u and TSLS is inconsistent.
- If there are more instruments than endogenous regressors, it is possible to test - *partially* - for instrument exogeneity.

Testing overidentifying restrictions

Consider the simplest case:

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

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}.
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #*Z*'s > #*X*'s (overidentified).

Suppose # instruments = $m > \#X's = k$ (overidentified).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

The J-test of overidentifying restrictions

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$, using the actual X's (not the $\hat{X}$'s used to estimate the second stage)
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$.
3. Regress $\hat{u}_i$ against $Z_{1i}, \cdots, Z_{mi}, W_{1i}, \cdots, W_{ri}$.
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i}, \cdots, Z_{mi}$ are all zero.
5. The J-statistic is $J = mF$

$J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i}, \cdots, Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i}, \cdots, Z_{mi}, W_{1i}, \cdots, W_{ri}$.

Distribution of the J-statistic

- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m - k$ degrees of freedom.
- If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.

Checking Instrument Validity:

Summary

The two requirements for valid instruments:

1. **Relevance** (special case of one X)
    - At least one instrument must enter the population counterpart of the first stage regression.
    - If instruments are weak, then the TSLS estimator is biased and the and t-statistic has a non-normal distribution
    - To check for weak instruments with a single included endogenous regressor, check the first-stage F.
        - If $F > 10$, instruments are not weak - use TSLS
        - If $F < 10$, weak instruments - take some action

2. Exogeneity

- All the instruments must be uncorrelated with the error term: $corr(Z_{1i}, u_i) = 0, \cdots, corr(Z_{mi}, u_i) = 0$.
- We can partially test for exogeneity: if $m > k$, we can test the hypothesis that all are exogenous, against the alternative that as many as $m - 1$ are endogenous (correlated with u).
- The test is the J-test, constructed using the TSLS residuals.

Application to the Demand for Cigarettes

Why are we interested in knowing the elasticity of demand for cigarettes?

- Theory of optimal taxation: optimal tax is inverse to elasticity: smaller deadweight loss if quantity is affected less.
- Externalities of smoking - role for government intervention to discourage smoking.
  - second-hand smoke (non-monetary).
  - monetary externalities.

Panel data set

- Annual cigarette consumption, average prices paid by end consumer (including tax), personal income.
- 48 continental US states, 1985-1995.

Estimation strategy

- Having panel data allows us to control for unobserved state-level characteristics that enter the demand for cigarettes, as long as they don't vary over time.
- But we still need to use IV estimation methods to handle the simultaneous causality bias that arises from the interaction of supply and demand.

Fixed-effects model of cigarette demand

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

- $i = 1, \cdots, 48$, $t = 1985, 1986, \cdots, 1995$.
- α_i reflects unobserved omitted factors that vary across states but not over time, e.g. attitude towards smoking.
- Still, $corr(\ln(P_{it}^{cigarettes}), u_{it})$ is plausibly nonzero because of supply/demand interactions.
- Estimation strategy:
  - Use panel data regression methods to eliminate α_i.
  - Use TSLS to handle simultaneous causality bias.

Panel data IV regression: Two approaches

(a) The "n-1 binary indicators" method.

(b) The "changes" method (when T=2).

(a) The "n-1 binary indicators" method

Rewrite

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

as

$$\begin{aligned} \ln(Q_{it}^{cigarettes}) &= \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) \\ &\quad + \gamma_2 D2_{it} + \cdots + \gamma_{48} D48_{it} + u_{it} \end{aligned}$$

$$\begin{aligned} \text{Instruments} : Z_{1it} &= \text{general sales tax}_{it} \\ Z_{2it} &= \text{cigarette} - \text{specific tax}_{it} \end{aligned}$$

This now fits in the general IV regression model:

$$\begin{aligned}\ln(Q_{it}^{cigarettes}) &= \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) \\ &\quad + \gamma_2 D2_{it} + \cdots + \gamma_{48} D48_{it} + u_{it}\end{aligned}$$

- X(endogenous regressor) $= \ln(P_{it}^{cigarettes})$.
- 48 W's (included exogenous regressors) $= \ln(Income_{it})$, $D2_{it}, \cdots, D48_{it}$.
- Two instruments $= Z_{1it}, Z_{2it}$.
- Now estimate this full model using TSLS!

(b) The "changes" method (when T=2)

- One way to model long-term effects is to consider 10-year changes, 1985-1995.
- Rewrite the regression in "changes" form:

$$\begin{aligned} & \ln(Q_{i1995}^{cigarettes}) - \ln(Q_{i1985}^{cigarettes}) \\ = \; & \beta_1 \left(\ln(P_{i1995}^{cigarettes}) - \ln(P_{i1985}^{cigarettes}) \right) \\ + \; & \beta_2 \left(\ln(Income_{i1995}) - \ln(Income_{i1985}) \right) \\ + \; & (u_{i1995} - u_{i1985}) \end{aligned}$$

- Must create "10-year change" variables, for example: 10-year change in log price = $\ln(P_{i1995}) - \ln(P_{i1985})$.
- Then estimate the demand elasticity by TSLS using 10-year changes in the instrumental variables.

STATA: Cigarette demand

First create "10-year change" variables

10-year change in log price

$$= \ln(P_{it}) - \ln(P_{it-10}) = \ln(P_{it}/P_{it-10})$$

```
. gen dlpackpc = log(packpc/packpc[_n-10]);
. gen dlavgprs = log(avgprs/avgprs[_n-10]);
. gen dlperinc = log(perinc/perinc[_n-10]);
. gen drtaxs  = rtaxs-rtaxs[_n-10];
. gen drtax   = rtax-rtax[_n-10];
. gen drtaxso = rtaxso-rtaxso[_n-10];
```

_n-10 is the 10-yr lagged value

Use TSLS to estimate the demand elasticity by using the "10-year changes" specification

Y W X Z

```
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso) , r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   12.31
                                                      Prob > F      =  0.0001
                                                      R-squared     =  0.5499
                                                      Root MSE      =  .09092

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -.9380143   .2075022    -4.52   0.000    -1.355945   -.5200834
    dlperinc |   .5259693   .3394942     1.55   0.128    -.1578071    1.209746
       _cons |   .2085492   .1302294     1.60   0.116    -.0537463    .4708446
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso
------------------------------------------------------------------------------
```

NOTE:

- *All the variables – Y, X, W, and Z's – are in 10-year changes*
- *Estimated elasticity = –.94 (SE = .21) – surprisingly elastic!*
- *Income elasticity small, not statistically different from zero*
- *Must check whether the instrument is relevant…*

Check instrument relevance: compute first-stage F

```
.  reg dlavgprs drtaxso dlperinc , r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  2,    45) =   16.84
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5146
                                                      Root MSE      =  .06334

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0254611   .0043876     5.80   0.000      .016624    .0342982
    dlperinc |  -.2241037   .2188815    -1.02   0.311    -.6649536    .2167463
       _cons |   .5321948   .0295315    18.02   0.000     .4727153    .5916742
------------------------------------------------------------------------------

.  test drtaxso;

 ( 1)  drtaxso = 0

       F(  1,    45) =   33.67
                 Prob > F =    0.0000
```

*We didn't need to run "test" here because with m=1 instrument, the F-statistic is the square of the t-statistic, that is, 5.80*5.80 = 33.67*

First stage F = 33.7 > 10 so instrument is not weak

Can we check instrument exogeneity? *No…m = k*

Cigarette demand, 10 year changes – 2 IVs

```
                      Y          W      X         Z1     Z2
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso drtax) , vce(r);

Instrumental variables (2SLS) regression              Number of obs =      48
                                                      Wald chi2(2)  =   45.44
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.5466
                                                      Root MSE      =  .08836

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -1.202403   .1906896    -6.31   0.000    -1.576148   -.8286588
    dlperinc |   .4620299   .2995177     1.54   0.123    -.1250139    1.049074
       _cons |   .3665388   .1180414     3.11   0.002     .1351819    .5978957
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso drtax
------------------------------------------------------------------------------
```

drtaxso = general sales tax only
drtax = cigarette-specific tax only
Estimated elasticity is -1.2, even more elastic than using general sales tax only!

Test the overidentifying restrictions

```
. predict e, resid;        Computes predicted values for most recently
                           estimated regression (the previous TSLS regression)
. reg e drtaxso drtax dlperinc;     Regress e on Z's and W's

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =    1.64
       Model |  .037769176     3  .012589725            Prob > F      =  0.1929
    Residual |  .336952289    44  .007658007            R-squared     =  0.1008
-------------+------------------------------           Adj R-squared =  0.0395
       Total |  .374721465    47  .007972797            Root MSE      =  .08751

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0127669   .0061587     2.07   0.044      .000355    .0251789
       drtax |  -.0038077   .0021179    -1.80   0.079     -.008076    .0004607
    dlperinc |  -.0934062   .2978459    -0.31   0.755    -.6936752    .5068627
       _cons |    .002939   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------
. test drtaxso drtax;

 ( 1)  drtaxso = 0                  Compute J-statistic, which is m*F,
 ( 2)  drtax = 0                    where F tests whether coefficients on
                                    the instruments are zero
       F(  2,    44) =    2.47           so J = 2 × 2.47 = 4.93
            Prob > F =    0.0966     ** WARNING - this uses the wrong d.f. **
```

The correct degrees of freedom for the J-statistic is $m–k$:

- $J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\ldots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{mi}$.
- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m–k$ degrees of freedom
- Here, $J = 4.93$, distributed chi-squared with d.f. = 1; the 5% critical value is 3.84, so reject at 5% sig. level.
- In STATA:

```
.  dis "J-stat = " r(df)*r(F) "  p-value = "  chiprob(r(df)-1,r(df)*r(F));
J-stat = 4.9319853  p-value = .02636401
```

J = 2 × 2.47 = 4.93 p-value from chi-squared(1) distribution

Check instrument relevance: compute first-stage *F*

```
          X       Z1       Z2     W
.  reg dlavgprs drtaxso drtax dlperinc , r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  3,    44) =   66.68
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.7779
                                                      Root MSE      =  .04333

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |    .013457   .0031405     4.28   0.000     .0071277    .0197863
       drtax |   .0075734   .0008859     8.55   0.000     .0057879    .0093588
    dlperinc |  -.0289943   .1242309    -0.23   0.817    -.2793654    .2213767
       _cons |   .4919733   .0183233    26.85   0.000     .4550451    .5289015
------------------------------------------------------------------------------

.  test drtaxso drtax;

 ( 1)  drtaxso = 0
 ( 2)  drtax = 0

       F(  2,    44) =   88.62          88.62 > 10 so instruments aren't weak
            Prob > F =    0.0000
```

Summary of these results:

TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q^{cigarettes}_{i,1995}) - \ln(Q^{cigarettes}_{i,1985})$

Regressor	(1)	(2)	(3)
$\ln(P^{cigarettes}_{i,1995}) - \ln(P^{cigarettes}_{i,1985})$	−0.94 (0.21) [−1.36, −0.52]	−1.34 (0.23) [−1.80, −0.88]	−1.20 (0.20) [−1.60, −0.81]
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34) [−0.16, 1.21]	0.43 (0.30) [−0.16, 1.02]	0.46 (0.31) [−0.16, 1.09]
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage F-statistic	33.7	107.2	88.6
Overidentifying restrictions J-test and p-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The J-test of overidentifying restrictions is described in Key Concept 12.6 (its p-value is given in parentheses), and the first-stage F-statistic is described in Key Concept 12.5. Heteroskedasticity-robust standard errors are given in parentheses beneath coefficients, and 95% confidence intervals are given in brackets.

How should we interpret the J-test rejection?

- J-test rejects the null hypothesis that both the instruments are exogenous.
- This means that either $rtaxso$ is endogenous, or $rtax$ is endogenous, or both.
- The J-test doesn't tell us which!! You must think!
- Why might $rtax$ (cig-only tax) be endogenous?
  - Political forces: history of smoking or lots of smokers $\Rightarrow$ political pressure for low cigarette taxes.
  - If so, cig-only tax is endogenous.
- This reasoning doesn't apply to general sales tax, $\Rightarrow$ use just one instrument, the general sales tax.

The Demand for Cigarettes: Summary of Empirical Results

- Use the estimated elasticity based on TSLS with the general sales tax as the only instrument:

  Elasticity = -.94, SE = .21.

- This elasticity is surprisingly large (not inelastic) - a 1% increase in prices reduces cigarette sales by nearly 1%. This is much more elastic than conventional wisdom in the health economics literature.

- This is a long-run (ten-year change) elasticity. *What would you expect a short-run (one-year change) elasticity to be - more or less elastic?*

Remaining threats to internal validity?

- Omitted variable bias?
  - Panel data estimator; probably OK.
- Functional form mis-specification
  - A related question is the interpretation of the elasticity: using 10-year differences, the elasticity interpretation is long-term. Different estimates would be obtained using shorter differences.

Remaining threats to internal validity, ctd.

- Remaining simultaneous causality bias?
  - Not if the general sales tax a valid instrument:
    - relevance?
    - exogeneity?
- Errors-in-variables bias? *Interesting question: are we accurately measuring the price actually paid?*
- Selection bias? (*no, we have all the states*)

Overall, this is a credible estimate of the long-term elasticity of demand although some problems might remain.

Where Do Valid Instruments Come From?

- Valid instruments are (1) relevant and (2) exogenous.
- One general way to find instruments is to look for exogenous variation - variation that is "as if" randomly assigned in a randomized experiment - that affects X.
  - Rainfall shifts the supply curve for butter but not the demand curve, rainfall is "as if" randomly assigned.
  - Sales tax shifts the supply curve for cigarettes but not the demand curve, sales taxes are "as if" randomly assigned.

Example: Cardiac Catheterization

Does cardiac catheterization (心導管) improve longevity of heart attack patients?

Y_i= survival time (in days) of heart attack patient

X_i= 1 if patient receives cardiac catheterization,

= 0 otherwise.

- Clinical trials show that $CardCath$ affects $SurvivalDays$.
- But is the treatment effective “in the field"?

$$SurvivalDays_i = \beta_0 + \beta_1 CardCath_i + u_i$$

- Is OLS unbiased? The decision to treat a patient by cardiac catheterization is endogenous - it is made in the field by EMT technician depends on u_i (unobserved patient health characteristics).
- If healthier patients are catheterized, then OLS has simultaneous causality bias and OLS overstates overestimates the CC effect.
- Propose instrument: distance to the nearest CC hospital - distance to the nearest “regular" hospital.

- Z = differential distance to CC hospital.
  - Relevant? If a CC hospital is far away, patient won't bet taken there and won't get CC.
  - Exogenous? If distance to CC hospital doesn't affect survival, then $\mathrm{Cov}(distance, u_i) = 0$. So exogenous.
  - If patients location is random, then differential distance is "as if" randomly assigned.
  - The 1st stage is a linear probability model: distance affects the probability of receiving treatment.
- Results (McClellan, McNeil, Newhous, *JAMA*, 1994):
  - OLS estimates significant and large effect of CC.
  - TSLS estimates a small, often insignificant effect.

Example: Peer Behavior Effects in Elementary School

Figlio, David N. (2007), "Boys Named Sue: Disruptive Children and Their Peers," *Education Finance and Policy* 2:4, 376-94.

- What is the effect on student performance of having disruptive children in the classroom?
- Y = Math test score
  X = measure of how disruptive your classmate's are
- What is the motivation for using instrumental variables?
- Proposed instrument:
  Z = fraction of male classmates with female names

Figure 1: Percentage of children suspended 5+ days on at least one occasion

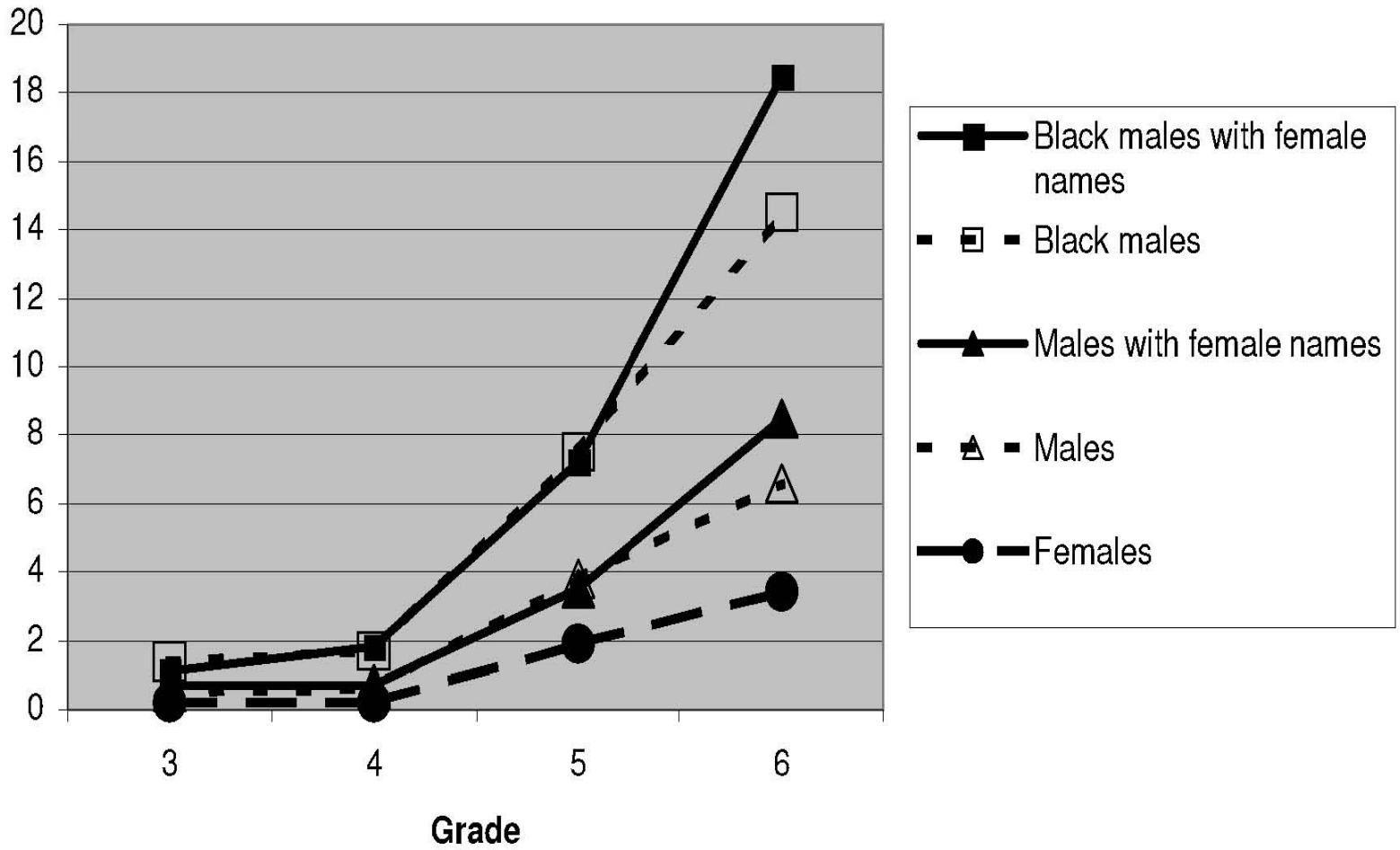


Table 2. First-stage estimates of the relationship between boys with female names and the rate of classroom disruption faced by students
Dependent variable: Fraction of classmates suspended for 5+ days

	(1)	(2)
Child fixed effects	YES	YES
Grade dummies	YES	YES
Fraction of male classmates with female names	0.003 (0.013)	0.000 (0.011)
Fraction of African-American male classmates with female names		0.033 (0.026)
Fraction of male classmates with female names x grade 6	0.226 (0.040)	-0.270 (0.026)
Fraction of African-American male classmates with female names x grade 6		1.056 (0.066)
Average third grade national percentile ranking of classmates (coefficient x 10)	-0.003 (0.000)	-0.003 (0.000)
Fraction of classmates who are African-American	0.036 (0.002)	0.034 (0.002)
Fraction of classmates who are male	0.042 (0.004)	0.038 (0.004)
Fraction of classmates who are free-lunch eligible	0.013 (0.002)	0.010 (0.002)
Fraction of classmates who are immigrants	-0.017 (0.002)	-0.016 (0.002)
Partial r-squared of female names variables	0.03	0.08

Notes: Standard errors adjusted for clustering are in parentheses beneath coefficient estimates. Data are for students in grades three through six

Table 3. Instrumental variables estimates of
the effect of disruptive classmates on student outcomes

	(3)	(4)	(5)	(6)
Child fixed effects	YES	YES	YES	YES
Grade dummies	YES	YES	YES	YES
Controls for fraction Black, third grade scores, fraction males, low income, immigrants among peers	YES	YES	YES	YES
Controls for fraction of male classmates with female names	NO	YES	NO	YES
Controls for fraction of Black male classmates with female names	NO	NO	NO	YES

Table 3 - IV results, ctd.

	(3)	(4)	(5)	(6)
Instruments employed	Fraction male classmates with female names (F), F x grade 6	Fraction male classmates with female names x grade 6	Fraction male classmates with female names (F), Fraction Black male classmates with female names (BF), F x grade 6, BF x grade 6	Fraction male classmates with female names x grade 6, Fraction Black male classmates with female names x grade 6
DEPENDENT VARIABLE	IV COEFFICIENT ESTIMATE ON FRACTION OF CLASSMATES SUSPENDED AT LEAST ONCE FOR 5+ DAYS			
Mathematics test score (national percentile ranking)	-65.76 (**16.13**)	-57.55 (**30.04**)	-114.39 (**8.47**)	-124.19 (**9.71**)
Child suspended at least once for 5+ days	0.94 (0.15)	0.84 (0.27)	0.98 (0.08)	1.01 (0.09)

Notes: Standard errors adjusted for clustering are in parentheses beneath coefficient estimates. Data are for students in grades three through six

Example: Effects of Empire

Feyrer, James, and Bruce Sacerdote (2009), "Colonialism and Modern Income— Islands as Natural Experiments," *Review of Economics and Statistics*, 91:2, 245-262..

- Does having been colonized historically affect modern economic well being?
- Data: n = 80 island economies (Atlantic, Pacific, Indian Oceans)
  Y = log GDP per capita or infant mortality
  X = number of years under colonial rule
  Z = 12-month average of east-west wind speed,
  12-month std. deviation of east-west wind speed
  W = geographic dummies, area, latitude

Figure 1
GDP Per Capita versus Years of Colonialism

Circles represent islands in the Atlantic, triangles are islands in the Pacific and squares are islands in the Indian Ocean.

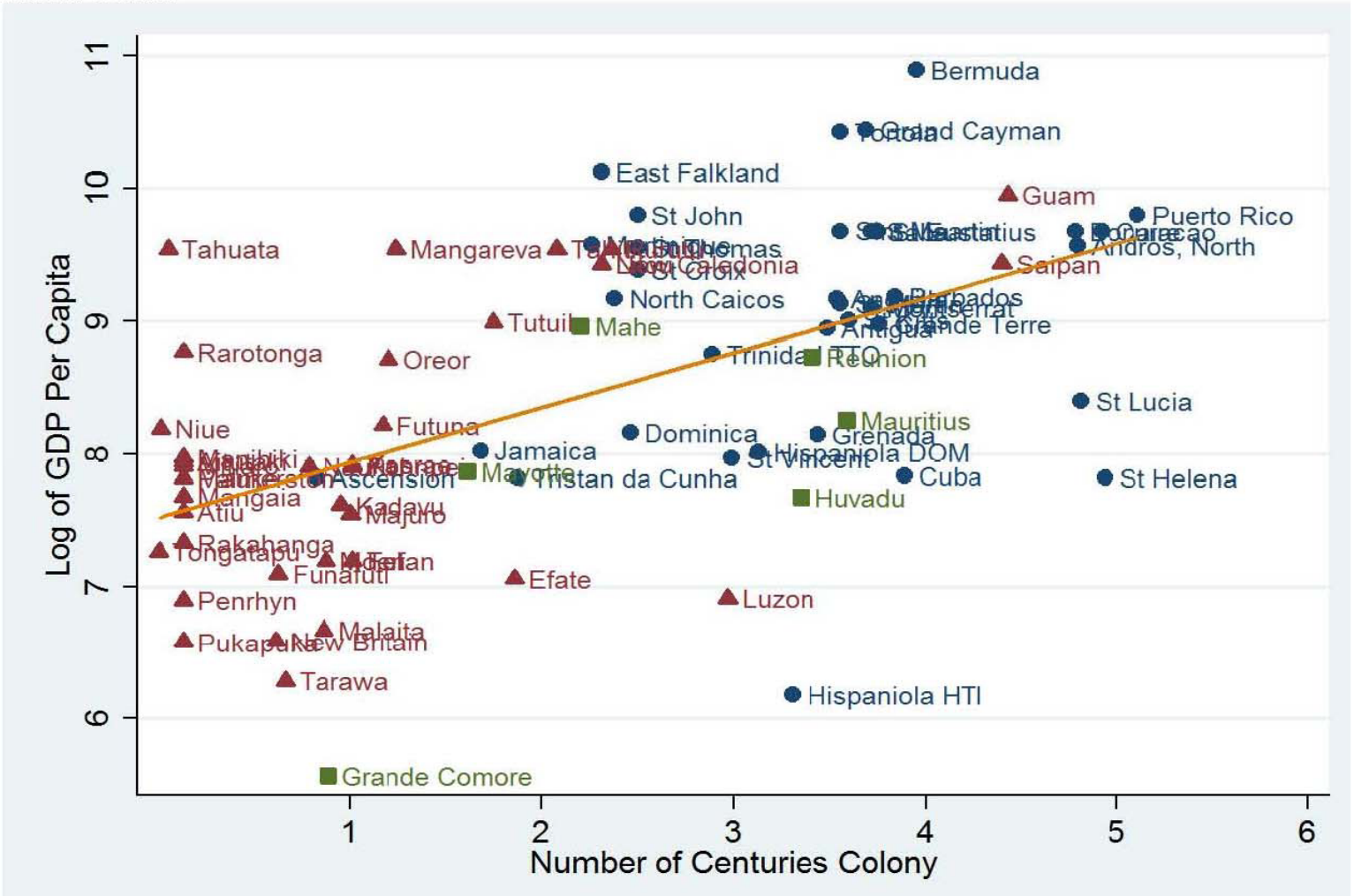


Figure 2
Years of Colonialism Versus Easterly Vector of Wind

Circles represent islands in the Atlantic, triangles are islands in the Pacific and squares are islands in the Indian Ocean.

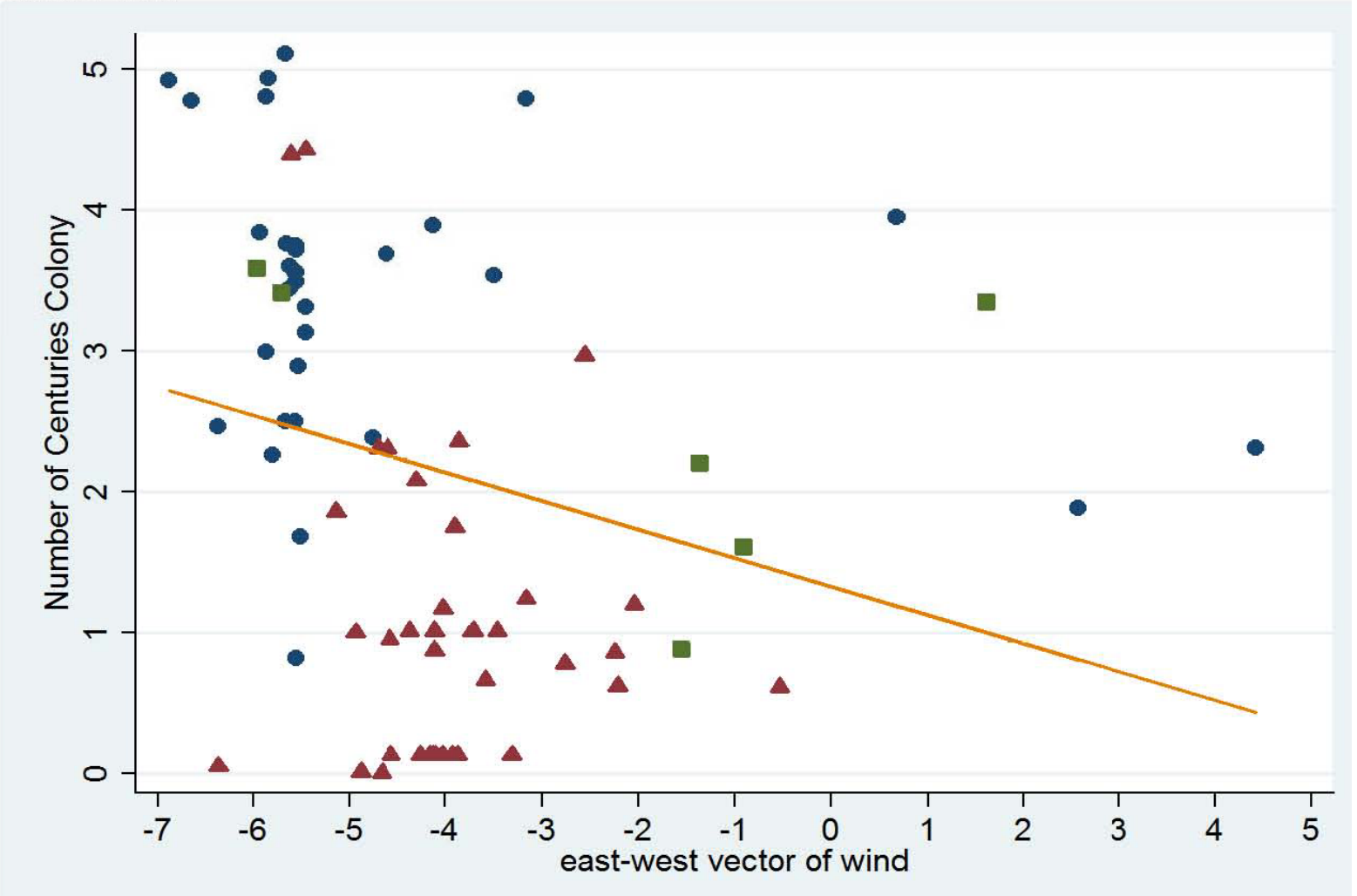


Table 11
Outcomes Regressed on Years of Colonization

We regress Log GDP per capita and infant mortality on the number of years the island spent as a colony of a European power. Columns (1), (2), (4), (6) and (7) are OLS. Columns (3), (5) and (8) are two stage least squares where we instrument for centuries of colonial rule or the first year as a colony using the 12 month average and standard deviation of the east-west wind speed for each island.

	(1) Log GDP Capita	(2) Log GDP Capita	(3) Log GDP Capita - IV	(4) Log GDP Capita	(5) Log GDP Capita- IV	(6) Infant Mortality Per 1000	(7) Infant Mortality Per 1000	(8) Infant Mortality Per 1000 - IV
Number of Centuries a Colony	0.413	0.450	0.441			-2.801	-2.611	-10.244
	(0.065)**	(0.083)**	(0.157)**			(1.156)*	(1.259)*	(4.344)*
First Year a Colony				-0.396	-0.545			
				(0.101)**	(0.232)*			
Final Year A Colony				0.014	0.007			
				(0.014)	(0.017)			
Remained A Colony in 2000				0.800	0.732			
				(0.149)**	(0.206)**			
Abs(Latitude)		0.048	0.048	0.039	0.042		-0.763	-0.771
		(0.011)**	(0.011)**	(0.011)**	(0.013)**		(0.211)**	(0.221)**
Area in millions of sq km		-21.046	-20.984	-20.429	-23.791		263.524	321.185
		(3.937)**	(3.961)**	(4.707)**	(6.169)**		(149.986)+	(143.722)*
Island is in Pacific		0.779	0.767	0.747	0.944		-7.427	-18.724
		(0.457)+	(0.522)	(0.470)	(0.569)		(9.498)	(13.608)
Island is in Atlantic		0.615	0.622	0.427	0.298		-7.349	-1.117
		(0.400)	(0.410)	(0.367)	(0.403)		(8.581)	(8.555)
Constant	7.524	6.172	6.192	13.673	16.356	24.771	41.579	60.751
	(0.166)**	(0.526)**	(0.659)**	(1.942)**	(4.173)**	(3.677)**	(10.898)**	(18.551)**
Observations	80	80	80	80	80	80	80	80
R-squared	0.320	0.578	0.578	0.642	0.630	0.080	0.353	0.082

Robust standard errors in parentheses. We cluster at the island group level since several of the islands (e.g. the Cook Islands and the Federated States of Micronesia) are used as separate observations from a cluster of politically related yet geographically distinct islands.

+ significant at 10%; * significant at 5%; ** significant at 1%

Summary: IV Regression

- A valid instrument let us isolate a part of X that is uncorrelated with u, and that part can be used to estimate the effect of a change in X on Y.
- IV regression hinges on having valid instruments:
  (1) Relevance: check via first-stage F.
  (2) Exogeneity: Test overidentifying restrictions via the J-statistic.
- A valid instrument isolates variation in X that is “as if" randomly assigned.