IV Regression

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.
Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.

The textbook comes with online resources and study guides. Other references will be given from time to time.

In this lesson you will learn ...

- The problem of identification.
- The IV model, assumptions, and estimator.
- The two-stage least-squares estimator and its sampling distribution.
- The general IV regression model and assumptions for causal inference.
- Instrument relevance and exogeneity in the general IV model.
- Checking instrument validity.

Identification in a Supply-Demand Diagram

Philip Wright's Problem

- Philip Wright was concerned with an important economic problem of his day: how to set an import tariff — a tax on imported goods. The key to understanding the economic effect of a tariff was having quantitative estimates of the demand and supply curves of the goods.
- Consider the problem of estimating the elasticity of demand for butter from the demand equation

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

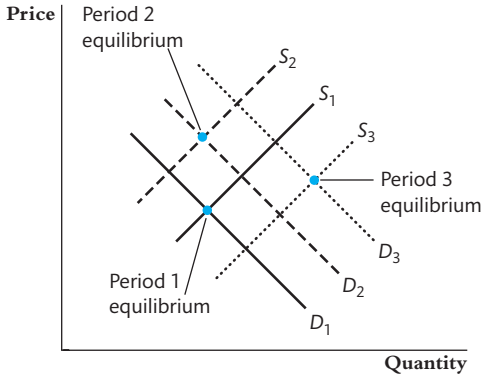
  where Q_i^{butter} is the ith observation on the quantity of butter consumed, P_i^{butter} is its price, and u_i represents other factors that affect demand, such as income and consumer tastes.
- The coefficient β_1 has the interpretation of the elasticity of Y with respect to X.
- Because of the interactions between supply and demand, the regressor $\ln(P_i^{\text{butter}})$ is likely correlated with the error term u_i.

Identification in a Supply-Demand Diagram

$\ln(P_i^{\text{butter}})$ is correlated with u_i.

- In year 1, the demand and supply curves for the first period are denoted D_1 and S_1.
- The first period's equilibrium price and quantity are determined by their intersection.
- In year 2, demand increases from D_1 to D_2 and supply decreases from S_1 to S_2.
- The second period's equilibrium price and quantity are determined by the new intersection.
- In year 3, demand increases again to D_3, supply increases to S_3.
- A new equilibrium price and quantity are determined.
- Because the points have been determined by changes in both demand and supply, they cannot be used to identify the demand or supply curve.
- Consider a scatterplot of the equilibrium price/quantity points. Fitting a line to these points will estimate neither a demand curve nor a supply curve!
- To identify the demand curve, we would need to fix the supply curve. And vice versa: To identify the supply curve, we would need to fix the demand curve.

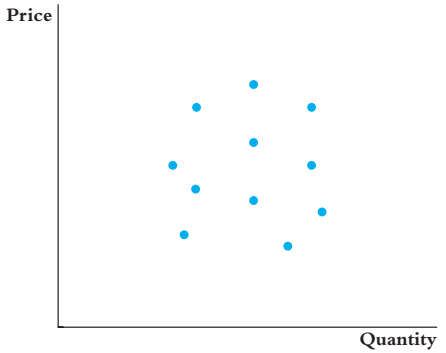
Identification in a Supply-Demand Diagram



(a) Demand and supply in three time periods

Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.

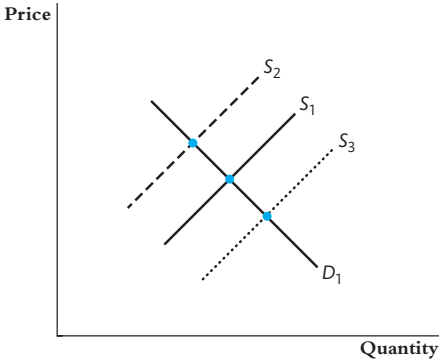
Identification in a Supply-Demand Diagram



(b) Equilibrium price and quantity for 11 time periods

This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?

Identification in a Supply-Demand Diagram



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

When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.

Identification in a Supply-Demand Diagram

Philip Wright's Problem

- Wright discovered that we need some third variable — the instrumental variable — that shifts supply but does not shift demand.
- The instrument must be correlated with the price.
  - The instrument shifts the supply curve, which leads to a change in the price.
- The instrument must be uncorrelated with the error term.
  - The instrument does not shift the demand curve.
- Wright considered weather as an instrument.
  - below-average rainfall in a dairy region could impair grazing and thus reduce butter production at a given price — it would shift the supply curve downwards (i.e. to the left) and increase the equilibrium price.
- **Instrument relevance:**
  dairy-region rainfall has a direct influence on the supply of butter and therefore on price.
- **Instrument exogeneity:**
  dairy-region rainfall has no direct influence on the demand for butter.

Instrumental Variables Regression

- **IV regression:** A method to obtain a consistent estimator of the unknown causal coefficients when a regressor X is correlated with the error term u.
- **Understand IV regression:** Separate the variation in X into two parts: one part that is correlated with u and another part that is not. If you could isolate the uncorrelated part, you could disregard the variations in X that bias the OLS estimates. Information about the movements in X that are uncorrelated with u is obtained from "instrumental variables."
- Let β_1 be the causal effect of X on Y.

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

  where u_i is the error term representing omitted factors that determine Y_i.
- If X_i and u_i are correlated, the OLS estimator is inconsistent.
- Instrumental variables estimation uses a variable Z to isolate that part of X that is uncorrelated with u.

Instrumental Variables Regression

Dealing with cor(X_i, u_i)

- **Valid Instruments**
  1. Instrument Relevance: $\text{cor}(Z_i, X_i) \neq 0$.
  2. Instrument Exogeneity: $\text{cor}(Z_i, u_i) = 0$.
- **Relevance:** The instrument's variation is related to the variation in the explanatory variable.
- **Exogeneity:** The part of the variation of the explanatory variable captured by the instrumental variable is exogenous.
- **Definitions:**
  - Endogenous variable: Variables correlated with the error term.
  - Exogenous variable: Variables uncorrelated with the error term.

Two-Stage Least Squares Estimator

- **Stage 1**
  The first stage decomposes X into a component that may be correlated with the regression error and a component that is uncorrelated with the error term.

$$X_i = \underbrace{\pi_0 + \pi_1 Z_i}_{\text{correlated}} + \underbrace{v_i}_{\text{uncorrelated}}$$

  where π_0 is the intercept, π_1 is the slope, and v_i is the error term.
- **Stage 2**
  The second stage uses the uncorrelated component to estimate β_1. Regress Y_i on the predicted value $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$:

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

  The estimators from the second-stage regression are the TSLS estimators $\hat{\beta}_0^{\text{TSLS}}$ and $\hat{\beta}_1^{\text{TSLS}}$.

Sampling Distribution of the TSLS Estimator

Large Sample Distribution

- For a single regressor X and a single instrument Z, the TSLS estimator has a simple formula.
- The TSLS estimator of β_1 is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X:

$$\hat{\beta}_1^{\text{TSLS}} = \frac{s_{ZY}}{s_{ZX}}$$

- In large samples, $\hat{\beta}_1^{\text{TSLS}}$ is consistent and normally distributed.

$$\hat{\beta}_1^{\text{TSLS}} \xrightarrow{p} \beta_1$$
$$\hat{\beta}_1^{\text{TSLS}} \sim \mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1^{\text{TSLS}}})$$
$$\sigma^2_{\hat{\beta}_1^{\text{TSLS}}} = \frac{1}{n} \frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}$$

- Because $\hat{\beta}_1^{\text{TSLS}}$ is normally distributed in large samples, hypothesis tests about β_1 can be performed by computing the t-statistic.

The Demand for Cigarettes

- Use TSLS to estimate the elasticity of demand for cigarettes using annual data for the 48 contiguous U.S. states for 1985 through 1995.
- **Regressand:** $Q^{\text{cigarettes}}$
  The number of packs of cigarettes sold per capita in the state.
- **Regressor:** $P^{\text{cigarettes}}$
  The average real price per pack of cigarettes, including all taxes.
- **Instrument:** $SalesTax$
  The portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack (in real dollars, deflated by the Consumer Price Index).
- **Instrument relevance:** A high sales tax increases the after-tax sales price.
- **Instrument exogeneity:** The sales tax must affect the demand for cigarettes only indirectly through the price. General sales tax rates vary from state to state, because of political considerations, not factors related to the demand for cigarettes.

The Demand for Cigarettes

- **First Stage:**

$$\widehat{\ln(P^{\text{cigarettes}})} = \underset{(0.03)}{4.62} + \underset{(0.005)}{0.031}\ SalesTax \quad \bar{R}^2 = 0.47$$

- As expected, higher sales taxes mean higher after-tax prices.
- The variation in the sales tax on cigarettes explains 47% of the variance of cigarette prices across states.
- **Second Stage:**

$$\widehat{\ln(Q^{\text{cigarettes}})} = \underset{(1.53)}{9.72} - \underset{(0.32)}{1.08}\ \widehat{\ln(P^{\text{cigarettes}})}$$

- An increase in the price of 1% reduces consumption by 1.08%. This suggests that the demand for cigarettes is surprisingly elastic. However, there clearly are omitted variables — At the very least we must control for income.
- To include control variables in the regression, we extend the simple TSLS to allow multiple regressors.

The General IV Regression Model

- The general IV regression model is

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

- Y_i: The dependent variable.
- $X_{1i}, \ldots, X_{ki}$: k endogenous regressors potentially correlated with u_i.
- $W_{1i}, \ldots, W_{ri}$: r exogenous regressors, uncorrelated with u_i.
- u_i: The error term, which captures measurement errors and/or omitted factors.
- $Z_{1i}, \ldots, Z_{mi}$: m instrumental variables.
- The coefficients are
  - **overidentified** if there are more instruments than endogenous regressors ($m > k$).
  - **exactly identified** if $m = k$.
  - **underidentified** if $m < k$.
- **Estimating the IV regression model requires exact identification or overidentification.**

The General IV Regression Model

- **First Stage:**

  Regress each of the k (endogenous) regressors $X_{1i}, \ldots, X_{ki}$ on:

  the m instrumental variables $Z_{1i}, \ldots, Z_{mi}$,
  on the r (exogenous) control variables $W_{1i}, \ldots, W_{ri}$.

  Obtain the k estimated values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.

- **Second Stage:**

  Regress the explained variable Y_i on:

  the k estimated values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$, and
  on the r (exogenous) control variables $W_{1i}, \ldots, W_{ri}$.

  Obtain the $k + r + 1$ estimated coefficients $\hat{\beta}_0^{\text{TSLS}}, \ldots, \hat{\beta}_{k+r}^{\text{TSLS}}$.

- Each endogenous regressor requires its own first-stage regression.
- Both regressions are typically estimated by OLS, including an intercept.

Two Conditions for Valid Instruments

The TSLS estimator is consistent and has a normal sampling distribution in large samples IF:

1. **Instrument Relevance**
   - Let $\hat{X}_{1i}$ be the predicted value from the regression of X_{1i} on the instruments (Z) and the included exogenous regressors (W):

     $$\hat{X}_{1i}, \ldots, \hat{X}_{ki}, W_{1i}, \ldots, W_{ri}, 1 \text{ are not perfectly multicollinear.}$$

     (1 is the intercept).
2. **Instrument Exogeneity**
   - The instruments are uncorrelated with the error term:

     $$\text{cor}(Z_{1i}, u_i) = \ldots = \text{cor}(Z_{mi}, u_i) = 0.$$

IV Regression Assumptions

- **The variables satisfy:**
  1. $\text{E}[u_i | W_{1i}, \ldots, W_{ri}] = 0$
  2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution
  3. X, W, Z, and Y have non-zero finite fourth moments (outliers unlikely)
  4. The instruments are valid, in that they satisfy:
     - Instrument relevance.
     - Instrument exogeneity.
- **Estimator:**
  The TSLS estimator is consistent and normally distributed in large samples.
- **Inference:**
  Valid, including hypothesis tests and confidence intervals.
- **Standard errors:**
  Second-stage regression standard errors are incorrect! Correct standard errors are provided by specialized packages — always use heteroskedasticity-robust standard errors.

Dealing with Weak Instruments

Checking for Weak Instruments

- **Weak instruments Rule of Thumb:** for a single endogenous regressor, $F_{\text{first-stage}} < 10$.
- The first-stage F-statistic tests the hypothesis that the coefficients on the instruments are all equal to zero in the first stage of two stage least squares.
- If the instruments are weak, the TSLS estimator is biased even in large samples.
  t-statistics and confidence intervals are unreliable.
- **What to do?**
  - If you have many instruments, discard the weakest instruments.
    With weak instruments, the standard errors are invalid, so if these standard errors become smaller after dropping instruments, that is meaningless.
  - J-test for overidentifying restrictions
  - find stronger instruments!
  - Go beyond TSLS.

Dealing with Weak Instruments

Overidentifying Restrictions Test: J-Statistics

- Use OLS to estimate a regression of the residuals from TSLS estimation ($\hat{u}_i^{\text{TSLS}}$) on the instruments (Z) and exogenous regressors (W):

  $$\hat{u}_i^{\text{TSLS}} = \delta_0 + \delta_1 Z_{1i} + \ldots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \ldots + \delta_{m+r} W_{ri} + e_i$$

  where e_i is the regression error term.
- Use the estimated regression coefficients $\hat{\delta}_0, \ldots, \hat{\delta}_m$ to test the hypothesis:

  $$\delta_1 = \ldots = \delta_m = 0$$

- Let F denote the homoskedasticity-only F-statistic. The overidentifying restrictions test statistic is $J = mF$.
- Under the null hypothesis that all the instruments are exogenous, if e_i is homoskedastic, in large samples J is distributed χ^2_{m-k}, where $m - k$ is the degree of overidentification.
- The degree of overidentification, $m - k$, is the number of instruments (Z) minus the number of endogenous regressors (W).

The Demand for Cigarettes

- The IV regression suffers from omitted variable bias: Include income as a control variable.
- One instrument, $SalesTax$:

$$\widehat{\ln(Q^{\text{cigarettes}})} = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\ \ln(P^{\text{cigarettes}}) + \underset{(0.31)}{0.21}\ \ln(Inc)$$

- Two instruments, $SalesTax$ and $CigTax$:

$$\widehat{\ln(Q^{\text{cigarettes}})} = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\ \ln(P^{\text{cigarettes}}) + \underset{(0.25)}{0.28}\ \ln(Inc)$$

- The standard errors on the estimated demand elasticity is smaller with two instruments: The regression explains more of the variation in cigarette prices.

The Demand for Cigarettes: Long-Run Elasticity

- To estimate the long-run price elasticity, consider quantity and price changes that occur over 10-year periods and two instruments, $SalesTax$ and $CigTax$. Three models are estimated and compared. The first model uses a single instrument, $SalesTax$.
- The first stage regression is:

$$\widehat{\ln(P^{\text{cigarettes}}_{1995}) - \ln(P^{\text{cigarettes}}_{1985})} = \underset{(0.03)}{0.53} - \underset{(0.22)}{0.22}\ \left[\ln(Inc_{1995}) - \ln(Inc_{1985})\right] + \underset{(0.0044)}{0.0255}\ \left[\ln(SalesTax_{1995}) - \ln(SalesTax_{1985})\right]$$

- The first-stage F-statistic for the null hypothesis

$$H_0:\ SalesTax_{1995} - SalesTax_{1985} = 0$$
$$F = t^2 = (0.0255/0.0044)^2 = 33.7.$$

- Since $F > 10$, the instrument $SalesTax$ is not weak.

The Demand for Cigarettes: Long-Run Elasticity

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)

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

The Demand for Cigarettes: Long-Run Elasticity

- **The instruments are not weak:**
  - The first-stage F-statistics are 33.7, 107.2, and 88.6.
  - In all three cases the first-stage F-statistics exceed 10.
- The regressions in columns (1) and (2) are exactly identified — We cannot use the J-test. (each have a single instrument and a single endogenous regressor)
- The regression in column (3) is overidentified (there is one overidentifying restriction, $m - k = 2 - 1 = 1$).
- **The null hypothesis that both the instruments are exogenous is rejected:**
  - The J-statistic for column (3) is 4.93.
  - The critical value from the χ^2_1 distribution with significance level 5% is 3.84.

Summary

- Instrumental variables regression is a way to estimate causal coefficients when one or more regressors are correlated with the error term.
- An endogenous variable is one that is correlated with the error term in the equation. Exogenous variables are uncorrelated with this error term.
- A valid instrument is (1) correlated with the included endogenous variable and (2) exogenous.
- IV regression requires at least as many instruments as included endogenous variables.
- The TSLS estimator has two stages. First, the included endogenous variables are regressed against the included exogenous variables and the instruments. Second, the dependent variable is regressed against the included exogenous variables and the predicted values of the included endogenous variables from the first-stage regression.
- Weak instruments are not strongly correlated with the included endogenous variables. Weak instruments make the TSLS estimator biased and TSLS confidence intervals and hypothesis tests unreliable.
- If an instrument is not exogenous, the TSLS estimator is inconsistent.

Keywords

instrumental variables (IV) regression instrument endogenous variable exogenous variable instrument relevance instrument exogeneity two-stage least squares exactly identified overidentified underidentified reduced form first-stage regression second-stage regression weak instruments test of overidentifying restrictions

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 12, Exercise 5.

Consider the IV regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

where X_i is correlated with u_i and Z_i is an instrument. Suppose that the first three assumptions are satisfied.

1. $\mathrm{E}[u_i | W_{1i}, \ldots, W_{ri}] = 0$;
2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;
3. X, W, Z, and Y have non-zero finite fourth moments (outliers unlikely);

Which IV assumption is not satisfied when:

1. Z_i is independent of (Y_i, X_i, W_i)?
2. $Z_i = W_i$?
3. $W_i = 1$ for all i?
4. $Z_i = X_i$?