

Chapter 10

Instrumental
Variables
Regression

Introduction to
Econometrics

JAMES H. STOCK
MARK W. WATSON

Scientific Conception of Causality

X causes Y means that (holding all other relevant factors constant)

every time X happens Y happens (e.g. holding pressure constant at a standard level if I heat water to 212 °F).

Statistical conception of causality

(more useful in social science)

X causes Y means that (holding all other *observable* relevant factors constant)

almost every time X happens Y happens (e.g. holding income constant if the price of cars increase people will buy fewer of them)

In public policy analysis causality is (nearly) always the most important issue!

Typical public policy issue is to solve problem Y by doing something about issue X.

For example:

Problem: Children are performing poorly in school

Issue: Would reducing TV watching raise school performance?

Does TV watching cause children to do poorly in school?

Simply investigating *the statistical correlation* between TV watching and school performance is likely to tell you little about whether school performance could be improved by limiting TV watching

Suppose that you find that kids who watch a lot of TV also do poorly in school. This is not very convincing evidence that limiting TV watching would cause them to perform better in school.

1. For example: Poor school performance may make the kids feel badly about themselves, this may result in them zoning out and watching lots of TV
2. More likely. The kids who do poorly in school have uninvolved parents. Uninvolved parents are associated with more TV watching.

Ideal experiment to determine whether TV watching lowers school performance

Orwellian world

1. Take 10,000 kids at birth
2. Randomly assign half of them (5,000) to have TVs in their house and prohibit the other half from ever having access to TV.
3. Wait 10 years and compare school performance

Clearly,

- A. We'd never get IRB approval and
- B. Even if we did we couldn't keep the 2nd group from watching TV.

Instrumental variables is something we can do in lieu of this “pure” study.

Basic idea behind instrumental variables is to find some variable or event that influences the hypothesis variable (TV viewing) but not the dependent variable (school performance). If we can find a good one we can do a “quasi-experiment”.

Tough to think of a good variable in this case but imagine:

1. Some areas of the city have access to cable TV and some do not.
2. Having access to cable to TV can be used to predict how much you watch TV (more access results in more TV watching)
3. Access to Cable TV does not *directly* effect school performance.

If that were true we could use access to cable TV to help predict school performance.

By far

The hardest thing about doing instrumental variable (or simultaneous) equations is finding an event (or variable) that has a big influence on the hypothesis variable but has little or no direct influence on the dependent variable.

IN FACT, THIS IS THE HARDEST PART OF EMPIRICAL WORK AND (TO ECONOMISTS) THE MOST IMPORTANT ISSUE (BY FAR) IN DOING EMPIRICAL SOCIAL SCIENCE.

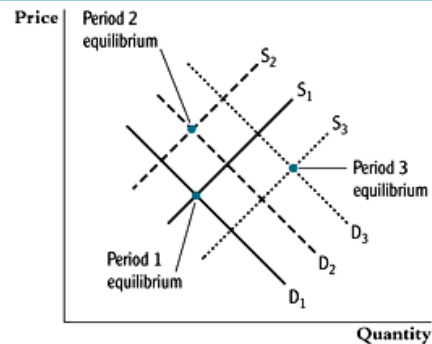
Instrumental Variables and the Search for Identification: From
Supply and Demand to Natural Experiments

Joshua D. Angrist, Alan B. Krueger

The Journal of Economic Perspectives, Vol. 15, No. 4
(Autumn, 2001), pp. 69-85

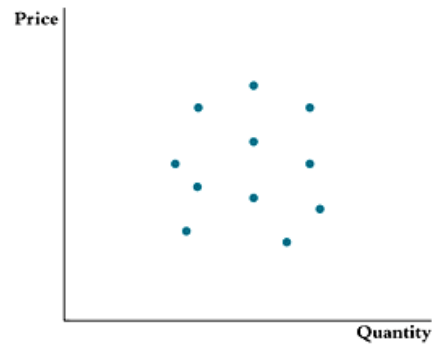
FIGURE 10.1

(a) 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 .



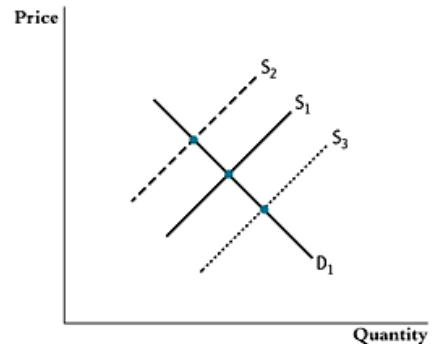
(a) Demand and Supply in Three Time Periods

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



(b) Equilibrium Price and Quantity for Eleven Time Periods

(c) 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.



(c) Equilibrium Price and Quantity When Only the Supply Curve Shifts

The next three slides each present one third of Figure 10.1.



Key Concept 10.2

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (10.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s):** Regress X_{1i} on the instrumental variables (Z_{1i}, \dots, Z_{mi}) and the included exogenous variables (W_{1i}, \dots, W_{ri}) using OLS. Compute the predicted values from this regression; call these \hat{X}_{1i} . Repeat this for all the endogenous regressors X_{2i}, \dots, X_{ki} , thereby computing the predicted values $\hat{X}_{1i}, \dots, \hat{X}_{ki}$.
2. **Second-stage regression:** Regress Y_i on the predicted values of the endogenous variables $(\hat{X}_{1i}, \dots, \hat{X}_{ki})$ and the included exogenous variables (W_{1i}, \dots, W_{ri}) using OLS. The TSLS estimators $\hat{\beta}_0^{TSLS}, \dots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in modern econometric software.

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad (10.12)$$

$i = 1, \dots, n$, where:

- Y_i is the dependent variable;
- u_i is the error term, which represents measurement error and/or omitted factors;
- X_{1i}, \dots, X_{ki} are k endogenous regressors, which are potentially correlated with u_i ;
- W_{1i}, \dots, W_{ri} are r included exogenous regressors, which are uncorrelated with u_i ;
- $\beta_0, \beta_1, \dots, \beta_{k+r}$ are unknown regression coefficients;
- Z_{1i}, \dots, Z_{mi} are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$); they are underidentified if $m < k$; and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.



Key Concept 10.1

TSLS estimator (p.328) is

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

OLS estimator (p.100) is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2} \Rightarrow s_{XY} = \hat{\beta}_1 s_X^2 = \hat{\beta}_1 [X, Y] s_X^2$$

where the notation after the last equality means the OLS coefficient obtained from a regression of X on Y.

Using this notation we can rewrite the formula for the TSLS estimator as

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{\hat{\beta}_1 [Z, Y] * s_{ZZ}}{\hat{\beta}_1 [Z, X] * s_{ZZ}} = \frac{\hat{\beta}_1 [Z, Y]}{\hat{\beta}_1 [Z, X]}$$



Key Concept 10.3

The Two Conditions for Valid Instruments

A set of m instruments Z_{1i}, \dots, Z_{mi} must satisfy the following two conditions to be valid:

1. Instrument Relevance

- *In general*, let \hat{X}_{1i}^* be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z 's) and the included exogenous regressors (W 's), and let "1" denote a regressor that takes on the value "1" for all observations (its coefficient is the intercept). Then $(\hat{X}_{1i}^*, \dots, \hat{X}_{ki}^*, W_{1i}, \dots, W_{ri}, 1)$ are not perfectly multicollinear.
- *If there is only one X* , then at least one Z must enter the population regression of X on the Z 's and the W 's.

2. Instrument Exogeneity

The instruments are uncorrelated with the error term, that is,
 $\text{corr}(Z_{1i}, u_i) = 0, \dots, \text{corr}(Z_{mi}, u_i) = 0.$



Key Concept 10.4

The IV Regression Assumptions

The variables and errors in the IV regression model in Key Concept 10.1 satisfy

1. $E(u_i | W_{1i}, \dots, W_{ri}) = 0$;
2. $(X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}, Z_{1i}, \dots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;
3. The X 's, W 's, Z 's, and u all have nonzero, finite fourth moments;
4. The W 's are not perfectly multicollinear; and
5. The two conditions for a valid instrument in Key Concept 10.3 hold.



Key Concept 10.5

A Rule of Thumb for Checking for Weak Instruments

The first-stage F -statistic is the F -statistic testing the hypothesis that the coefficients on the instruments Z_{1i}, \dots, Z_{mi} equal zero in the first stage of two stage least squares. When there is a single endogenous regressor, a first-stage F less than 10 indicates that the instruments are weak, in which case the TSLS estimator is biased (even in large samples), and TSLS t -statistics and confidence intervals are unreliable.



Key Concept 10.6

The Overidentifying Restrictions Test (the J -Statistic)

Let \hat{u}_i^{TSLs} be the residuals from TSLs estimation of Equation (10.12). Use OLS to estimate the regression coefficients in

$$\hat{u}_i^{TSLs} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i, \quad (10.17)$$

where e_i is the regression error term. Let F denote the homoskedasticity-only F -statistic testing the hypothesis that $\delta_1 = \cdots = \delta_m = 0$. The overidentifying restrictions test statistic is $J = mF$. Under the null hypothesis that all the instruments are exogenous, then in large samples J is distributed χ_{m-k}^2 , where $m - k$ is the “degree of overidentification,” that is, the number of instruments minus the number of endogenous regressors.

TABLE 10.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$			
Regressor	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	-0.94** (0.21)	-1.34** (0.23)	-1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
Intercept	0.21 (0.13)	0.45** (0.14)	0.37** (0.12)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage <i>F</i> -statistic	33.70	107.20	88.60
Overidentifying restrictions <i>J</i> -test and <i>p</i> -value	-	-	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the ten-year differences). The data are described in Appendix 10.1. The *J*-test of overidentifying restrictions is described in Key Concept 10.6 (its *p*-value is given in parentheses), and the first-stage *F*-statistic is described in Key Concept 10.5. Individual coefficients are statistically significant at the *5% level or **1% significance level.