Applied Economics

Instrumental Variables

Economics Department
Universidad Carlos III de Madrid

Material from Stock and Watson (ch.12), Wooldridge (ch.15), Angrist and Pischke (ch.4)

Simple Regression and Multiple Regression Models

- Which is the relationship between the simple and multiple regression models? Let's see an example:

Multiple Regression Model (long model)

- $wages = \beta_0 + \beta_1 educ + \beta_2 IQ + u$
- $C(educ, u) = C(IQ, u) = 0$

Simple Regression Model (short model)

- $wages = \gamma_0 + \gamma_1 educ + v$

- Is there any relationship between γ_1 and β_1?

γ_1 and β_1

Using the long model (assuming $C(educ, u) = 0$):

$$\begin{aligned} C(educ, w) &= C(educ, \beta_0 + \beta_1\, educ + \beta_2 IQ + u) \\ &= \beta_1\, V(educ) + \beta_2\, C(educ, IQ) \end{aligned}$$

Comparing both models:

$$\gamma_1 = \frac{C(educ, w)}{V(educ)} = \beta_1 + \beta_2 \frac{C(educ, IQ)}{V(educ)}$$

Omitted Variable Bias

- Then, assuming $C(educ, u) = 0$:

$$\gamma_1 = \beta_1 + \beta_2 \frac{C(educ, IQ)}{V(educ)}$$

- Note that $\frac{C(educ, IQ)}{V(educ)}$ is the slope in a regression of IQ on $educ$.
- This equation defines the Omitted Variable Bias: $\gamma_1 - \beta_1$

- There is no OVB $(\gamma_1 = \beta_1)$ if at least one of the two conditions is verified:
  - intelligence is not relevant: $\beta_2 = 0$
  - education is not correlated with intelligence: $C(educ, IQ) = 0$

Is $\hat{\gamma}_1$ a consistent estimator of the parameter of interest?

The parameter of interest is β_1

$$\begin{aligned}
\hat{\gamma}_1 = \frac{\hat{C}(educ, wages)}{\hat{V}(educ)} &= \frac{\hat{C}(educ, \beta_0 + \beta_1 educ + \beta_2 IQ + u)}{\hat{V}(educ)} \\
&= \beta_1 + \beta_2 \frac{\hat{C}(educ, IQ)}{\hat{V}(educ)} \\
&\Rightarrow plim(\hat{\gamma}_1) = \beta_1 + \beta_2 \frac{C(educ, IQ)}{V(educ)}
\end{aligned}$$

- $plim(\hat{\gamma}_1) = \beta_1$ ($\hat{\gamma}_1$ is consistent) if
  - intelligence is not relevant: $\beta_2 = 0$ or
  - education is not correlated with intelligence: $C(educ, IQ) = 0$
- We can show that $V(\hat{\gamma}_1) \leq V(\hat{\beta}_1)$

Uncorrelated Regressors

- **If *educ* and *IQ* are not correlated we get two simple FOC:**

$$\hat{\beta}_1 = \frac{\hat{C}(educ, wages)}{\hat{V}(educ)}$$
$$\hat{\beta}_2 = \frac{\hat{C}(IQ, wages)}{\hat{V}(IQ)}$$

- Then:

$$\hat{\beta}_1 = \frac{\hat{C}(educ, wages)}{\hat{V}(educ)}$$
$$\hat{\beta}_2 = \frac{\hat{C}(IQ, wages)}{\hat{V}(IQ)}$$

- the estimates are the same as the OLS estimates in simple linear regression models:

$$\hat{\beta}_1 = \frac{\hat{C}(educ, wages)}{\hat{V}(educ)} = \hat{\gamma}_1$$

Correlated Regressors

- With **correlated regressors**, in the long model, FOC are more complicated:

$$\hat{C}(educ, wages) = \hat{\beta}_1 \hat{V}(educ) + \hat{\beta}_2 \hat{C}(educ, IQ)$$
$$\hat{C}(IQ, wages) = \hat{\beta}_1 \hat{C}(IQ, educ) + \hat{\beta}_2 \hat{V}(IQ)$$

- Dividing the first condition by $\hat{V}(educ)$:

$$\frac{\hat{C}(educ,wages)}{\hat{V}(educ)} = \hat{\beta}_1 + \hat{\beta}_2 \frac{\hat{C}(educ,IQ)}{\hat{V}(educ)}$$

- The OLS estimate in the simple model is $\hat{\gamma}_1 = \frac{\hat{C}(educ,wages)}{\hat{V}(educ)}$:

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\hat{C}(educ,IQ)}{\hat{V}(educ)}$$

Correlated Regressors

- Omitted Variable Bias:

$$\hat{\gamma}_1 - \hat{\beta}_1 = \hat{\beta}_2 \frac{\hat{C}(educ,IQ)}{\hat{V}(educ)}$$

- the OLS estimate $(\hat{\gamma}_1)$ captures two effects on wages:

1. effects of independent changes in *educ*: $\hat{\beta}_1$
2. effects of changes in *IQ* associated to changes in *educ*:

$$\hat{\beta}_2 \frac{\hat{C}(educ,IQ)}{\hat{V}(educ)}$$

- where $\frac{\hat{C}(educ,IQ)}{\hat{V}(educ)}$ captures changes in *IQ* due to changes in *educ*

Conditional Mean Independence

- If we estimate the short model when the long model is the true one we are not identifying the effect we want (in this example, the impact of education on wages). Why?
- Because the Conditional Mean Independence assumption is not satisfied:

$$E(v|educ) \neq 0$$

- When the Conditional Mean Independence assumption is not satisfied, we say that there is an endogeneity problem.
- If for any reason, X_j is correlated with the error term, we say that X_j is an endogenous variable.

First Application

- The instrumental variables method is common in applied economics when there are endogeneity problems related to omitted variables, as the one we saw before.
- The first applications, however, are related to estimations of elasticities for supply and demand of agricultural goods.
- Philip Wright (1928) used the idea of what it will be later called instrumental variables to estimate the demand elasticity using a simple demand equation:

  $ln(Q_i) = \beta_0 + \beta_1 ln(P_i) + u_i$, where Q is quantity and P price.
- Problem: prices and quantities are jointly determined by the intersection of supply and demand curves.

First Application (cont.)

- Then, an OLS estimation of quantities on prices cannot identify nor the supply neither the demand curve.
- The solution proposed by Wright was two find two type of factors: “(A) affecting demand conditions without affecting costs conditions or which (B) affecting costs conditions without affecting demand conditions”.
- Type (A) factors help to identify the supply curve, type (B) factors help to identify the demand curve.
- Wright proposed several factors: the price of substitutes as a factor affecting demand but not supply, and weather-related variables as factors affecting supply but not demand.

Introduction 1/2

Let's assume we want to estimate the following model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ where } C(X, u) \neq 0$$

- If $C(X, u) \neq 0$, X is an endogenous variable, and OLS yields inconsistent estimators.
- Estimations using Instrumental Variables (IV) use an additional variable (Z) to isolate the part of X not correlated with u.
- We ask Z to verify two conditions.

Introduction 2/2

- The two conditions:
  - Z is not correlated with the error: $C(Z,u)=0$. Z does not affect directly the variable of interest. **Exogeneity**
  - Z is correlated (partial correlation) with X (the endogenous variable): $C(Z,X)\neq 0$. **Relevance**
- If Z is relevant, its variation is related to the variation in X. If Z is exogenous, the part of the variation in X captured by Z is exogenous.
- The only reason for finding a relationship between Y and Z is due to the relevance of Z.
- Under these conditions, using Z as IV allows us to obtain consistent estimators even under endogeneity.

Example 1/3

Example: wage equation

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

- Is it reasonable to assume that $C(educ_i, u_i) = 0$?
- We can argue that ability is an omitted variable in the model. If *educ* is correlated with ability, the Conditional Mean Independence assumption will not be valid.
- A good instrument needs to be correlated with *educ* but not with ability, or any other factor in the error term. Any ideas?

Example 2/3

- Some of the instruments for education used in the literature: parental education, number of siblings, distance to the university, date of birth.
- For instance, Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He estimated a standard wage equation including other standard controls: experience, race, region.
- He used a dummy variable for whether someone grew up near a four year college as an instrumental variable for education.

Example 3/3

- Relevance: those students who grew up near a four year college are more likely to attend college (any argument against it?).
- Exogeneity: distance should not be related to the ability of individuals or to any other factor in the error term (any argument against it?).
- Card finds that the IV estimate of the return to education is almost twice as large as the OLS estimate (13.2% vs. 7.5%), but the standard error of the IV estimate is over 18 times larger than the OLS standard error.
- The 95% confidence interval for the IV estimate is from .024 and .239, which is a very wide range. The price we pay to get a consistent estimator.

Valid IV: Exogeneity

- In the case of one endogenous variable and one instrument is not possible to test if the instrument is exogenous: $C(Z_i, u_i) \neq 0$.
- In the example of Card (1995), we argue that the distance does not affect the wage through another mechanism. What if the distance is correlated with family income and family income is an omitted variable in the wage equation?
- In the example of Wright(1928) we need to argue that weather conditions do not affect the demand of the good.

Valid IV: Relevance

- The second condition ($C(X_i, Z_i) \neq 0$) is verifiable since we observe both variables:
- Regress X on Z (actually on all the exogenous variables):

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

- Test the hypothesis: $H_0 : \pi_1 = 0$
- If we reject H_0, we have evidence that X and Z are correlated, and then Z is relevant.
- If we do not reject H_0, we say that Z is a weak instrument, a problem that we will discuss later.

Two Stage Least Squares

If the instrument Z verifies both conditions, it is possible to get consistent estimators using the Two Stage Least Squares estimator (TSLS). As it sounds, TSLS has two stages -two regressions:

- In the first stage we isolate the part of X that is uncorrelated with u by regressing X on Z using OLS : $X_i = \pi_0 + \pi_1 Z_i + v_i$
  The idea is to use the part of X that can be predicted using Z: $\pi_0 + \pi_1 Z_i$. In this first stage, we obtain OLS estimates for π_0 and π_1 and we compute $\hat{X}_i$.
- The second stage is the OLS regression of Y on $\hat{X}$. Because Z is exogenous, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ is not correlated with u_i. The estimator in this second stage is called the TSLS estimator.

Two Stage Least Squares

- In applied work, using a specialized command, both stages are estimated at the same time (as always, we use robust standard errors). If we do it separately, we need to adjust standard errors in the second stage since we are using $\hat{X}_i$, an estimated variable.
- Formula: very simple in the case of one endogenous regressor and one instrument:

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

  where s represents the sample covariance between two variables.
- We can show that TSLS is a consistent estimator and normally distributed in large samples.

TSLS: Consistency

Let's start with the simple model: $Y_i = \beta_0 + \beta_1 X_i + u_i$ and apply covariance properties:

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

Under exogeneity of the instrument: $C(Z,u) = 0$ and $\beta_1 = C(Z,Y)/C(Z,X)$. Since the sample covariance is a consistent estimator of the covariance we can show that:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{C(Z,Y)}{C(Z,X)} = \beta_1$$

TSLS vs OLS

- TSLS:

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

The bias without imposing exogeneity of Z:

$$\hat{\beta}_1^{TSLS} \xrightarrow{p} \frac{C(Z,Y)}{C(Z,X)} = \beta_1 + \frac{C(Z,u)}{C(Z,X)}.$$

The bias then depends on two conditions: exogeneity and relevance.

- OLS:

$$\hat{\beta}_1^{OLS} = \frac{s_{XY}}{s_X^2}$$

We obtain the bias similarly:

$$\hat{\beta}_1^{OLS} \xrightarrow{p} \frac{C(X,Y)}{V(X)} = \beta_1 + \frac{C(X,u)}{V(X)}.$$

The bias depends on the exogeneity of X.

General Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$
$$+ \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{ri} + u_i$$

- We may have more controls: some endogenous ($X_1, ..., X_k$, potentially correlated with u) and some exogenous ($W_1, ..., W_r$, not correlated with u).
- To apply TSLS we need at least as many instruments (denoted as $Z_1, Z_2, ..., Z_m$) as endogenous variables ($m \geq k$).
- The coefficients are **exactly identified** if there are just enough instruments to estimate the parameters of the model ($m = k$). The coefficients are **overidentified** if there are more instruments than endogenous regressors ($m > k$).

TSLS: several instruments

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

- With more than one instrument for X_1 we would have more than one possible IV estimator, but none of them is efficient: the best instrument is a linear combination of all possible instruments.
- First stage (X_1 on the m instruments and the r exogenous controls):

$$X_{1i} = \pi_0 + \pi_1 Z_{1i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+r} W_{ri} + v_i$$

- Second stage: Y_i on $\hat{X}_i$ and the exogenous controls in the original equation $(W_{1i}, ..., W_{ri})$ using OLS.
- Relevance condition: at least one Z useful to predict X_1, given the $W's$.
- Exogeneity condition: each Z needs to be uncorrelated with u.

TSLS: several endogenous regressors

- Similar TSLS procedure as before, only that each endogenous regressor needs its own first stage regression. Each one of these regressions include the same controls: all the instruments and all the exogenous controls from the original equation.
- Second stage: Y_i on all the $\hat{X}_j$ and the exogenous controls in the original equation $(W_{1i}, ..., W_{ri})$ using OLS.
- Again, in our applications, we estimate both stages automatically using `gretl`. In this way we get the correct standard errors.

Testing for Endogeneity: Hausman Test

- If there is no endogeneity in the original model, both estimators, OLS and TSLS are consistent, but OLS is more efficient. Remember the Card example.
- Under endogeneity only TSLS is consistent.
- Therefore, it is important to have a test for endogeneity. We use the Hausman test for endogeneity (H_0 : Exogeneity).

Testing for Endogeneity 1/2

Given the following simplified model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + u_i$$

- If we have an additional exogenous variable (Z_1), we can apply a two-step procedure to test if X_1 is an endogenous variable:
- **First Step:** regress X_1 on all the exogenous variables (in our example W_1 and Z_1) and compute the residuals: $\hat{v}$.

$$X_1 = \pi_0 + \pi_1 Z_1 + \pi_2 W_1 + v$$

- Under exogeneity of X_1, because Z_1 and W_2 are not correlated with u (by assumption), the residuals $\hat{v}$ should neither be.

Testing for Endogeneity 2/2

- **Second Step:** estimate the original model adding $\hat{v}$ to the equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \alpha \hat{v}_i + \varepsilon_i$$

- Test the null hypothesis that X_1 is exogenous. Under this null, the coefficient of $\hat{v}$ should be not significant: $H_0)\ \alpha = 0$.
- If we reject H_0, we have evidence against X_1 being exogenous, then against using OLS.
- Note that we need an exogenous instrument to carry out this test.

Instruments validity: relevance

- With one endogenous regressor and several instruments: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + u_i$, with m additional exogenous variables: $Z_1, ..., Z_m$.
- Relevance is checked in the first stage regression: $X_{1i} = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + v_i$
- We test the null hypothesis that the coefficients of the instruments are jointly zero: $H0)\pi_1 = ... = \pi_m = 0$. The F-statistic is a measure of how much information is included in the instruments.
- “Weak” instruments explain very little of the variation in X_1, beyond that explained by the $W's$ (simple rule: F below 10).

Weak instruments

- If instruments are weak, the sampling distribution of TSLS and its t-statistic are not (at all) normal, even with n large. Statistical inference will not be correct.
- What to do? Get better instruments (not easy...)
- If you have many instruments, some are probably weaker than others and you can try dropping the weaker ones until you find a set that can be considered relevant.

Exogeneity

- If the coefficients are **exactly identified** we cannot test if the instruments are exogenous.
- If we have more instruments than endogenous variables, we can test the overidentifying restrictions (we use a Sargan test).
- The test allows us to know if the additional instruments are exogenous.

Overidentification Test

1. First Step: Estimate the original model by TSLS and obtain the TSLS residuals ($\hat{u}^{TSLS}$).
2. Second Step: Regress the residuals on all the exogenous variables (using OLS):
   $\hat{u}^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \ldots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \ldots + \delta_{m+r} W_{ri} + v_i$

- Compute nR^2. Under the null hypothesis that the additional instruments are exogenous:

$$LM = nR^2 \rightarrow \chi^2_q$$

where q is the number of additional instruments (the degree of overidentification).

Application

- We are interested in knowing the elasticity of demand for cigarettes. We use annual data on cigarette consumption and average prices paid by end consumer for the US. We use the following equation where Q_i is the number of packs sold per capita and P_i the real price in state i.

$$ln(Q_i) = \beta_0 + \beta_1 ln(P_i) + u_i$$

- Is it correct to use OLS to estimate β_1?
- If we want to apply TSLS we need at least one instrument. One candidate is the general sales tax per pack in each state: Tax_i.

Application (cont.)

- Conditions for a valid instrument:
  1. Relevance: correlated with P

     Using 1995 data the results from the first stage regression are (file cig_ch10.gdt):

     $$\widehat{ln(P_i)} = \underset{(0.0289)}{4.6165} + \underset{(0.0048)}{0.0307}\ Tax \quad T = 48 \quad R^2 = 0.4710$$

     The estimated coefficient of *Tax* is positive and significantly different than 0: more taxes higher after-tax prices. Variation in taxes explains 47% of the variation in prices among states.
  2. Exogeneity: it is not possible to check it formally. Argument: taxes affect the demand of cigarettes only through the price.

Application (cont.)

- TSLS estimation using *Tax* as an instrument for *P*, with robust standard errors:

  $$\widehat{ln(Q_i)} = \underset{(1.5283)}{9.7199} - \underset{(0.3189)}{1.0836}\, ln(P_i)$$

  A 1% increase in price decreases consumption on average by 1.08%.

- Potential problem: omitted variables correlated with taxes: if that's the case *Tax* will not be exogenous.
- For instance, states with higher income levels could also have lower taxes, and higher consumption levels.

Application (cont.)

- To try to solve this problem we include income in the regression (and assume it is exogenous):

$$ln(Q_i) = \beta_0 + \beta_1 ln(P_i) + \beta_2 ln(Ing_i) + u_i$$

- TSLS with *Tax* as an instrument for *P*, with robust standard errors:

$$\widehat{ln(Q_i)} = \underset{(1.2594)}{9.4307} - \underset{(0.3723)}{1.1434}\, ln(P_i) + \underset{(0.3117)}{0.2145}\, ln(Inc_i)$$

- We used only one instrument: demand elasticity exactly identified.
- We could try to add another instrument: one candidate is the cigarette-specific tax ($CigTax$). With two instruments demand elasticity is overidentified.

Application (cont.)

- TSLS with *Tax* and *CigTax* as instruments for P, with robust standard errors:

$$\widehat{ln(Q_i)} = \underset{(0.9592)}{9.8950} - \underset{(0.2496)}{1.2774}\, ln(P_i) + \underset{(0.2539)}{0.2804}\, ln(Inc_i)$$

- Compare the standard errors.
- Are these estimations reliable? Depends on the validity of the instruments.

Application (cont.)

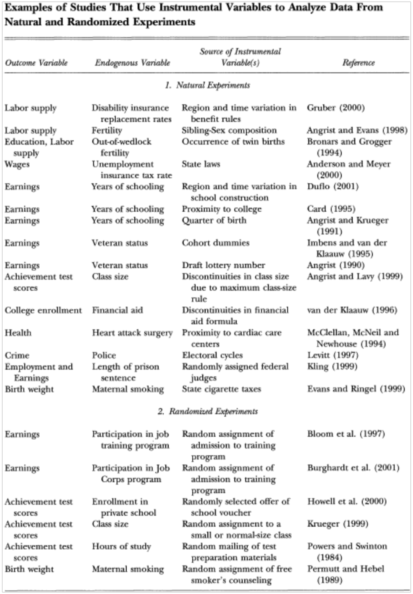
- Relevance: first stage regression:

$$ln(P_i) = \pi_0 + \pi_1 Tax_i + \pi_2 CigTax_i + \pi_3 ln(Inc_i) + u_i$$

  We test $H0)\pi_1 = \pi_2 = 0$, and the corresponding F-statistic is 209.676. We reject the null hypothesis that the instruments are weak.

- Exogeneity: with two instruments and one endogenous variable it is possible to run an overidentification test. The F-statistic from the Sargan test is 0.33, and given the χ^2_1 distribution of this statistic, the p-value is 0.5641. Then, we do not reject the null that both instruments are exogenous.

IV papers - Angrist and Krueger



Examples of Studies That Use Instrumental Variables to Analyze Data From Natural and Randomized Experiments

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*	*Reference*
		1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling-Sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)
		2. Randomized Experiments	
Earnings	Participation in job training program	Random assignment of admission to training program	Bloom et al. (1997)
Earnings	Participation in Job Corps program	Random assignment of admission to training program	Burghardt et al. (2001)
Achievement test scores	Enrollment in private school	Randomly selected offer of school voucher	Howell et al. (2000)
Achievement test scores	Class size	Random assignment to a small or normal-size class	Krueger (1999)
Achievement test scores	Hours of study	Random mailing of test preparation materials	Powers and Swinton (1984)
Birth weight	Maternal smoking	Random assignment of free smoker's counseling	Permutt and Hebel (1989)

Example

Using the file mroz.gdt, which has information on the participation of women in the labor market, we estimate the following wage equation:

$lwage = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp2 + \varepsilon$

1. Analyze if *educ* is an exogenous variable (use husband and parents' education as exogenous variables).
2. Discuss if husband and parents' education are good instruments for *educ*.
3. Estimate the effect of education on wages using the more appropriate method: OLS or TSLS.