Causality and Omitted Variable Bias

Panel Data Econometrics

Prof. Alexandre Gori Maia

State University of Campinas

Topics

- Omitted Variable Bias
- 2 Stage Least Squares
- Propensity Score Matching

Reference

Angrist, J.; Pischke, J. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Caps. 1-4, 2009.

Endogeneity

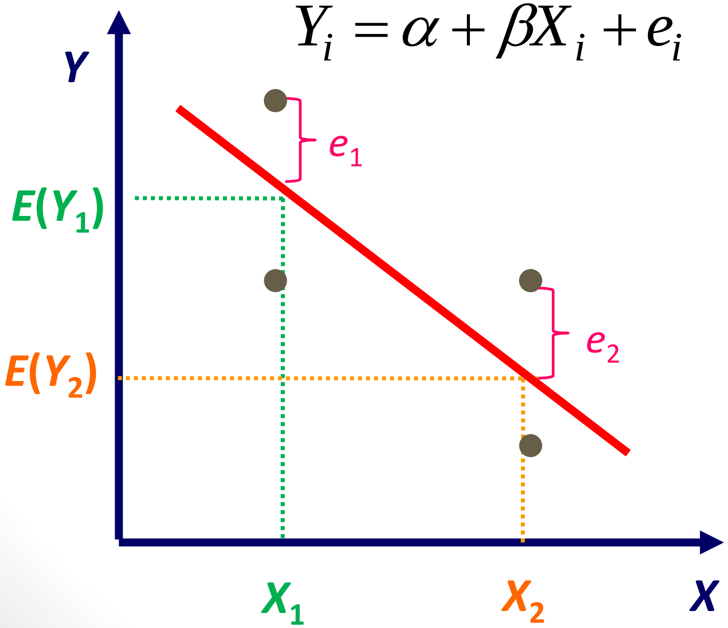
An important assumption of the OLS estimates is that the values of *X* are not related to the errors *e,* i.e.:

$$E(e|X) = 0$$

We say that the regressor *X* is endogenous when it is related to the errors *e*:

$$E(e|X) \neq 0$$

$$Y_i = \alpha + \beta X_i + e_i$$



We assume that, once we hold *X* constant, we can observe random variations of *Y* or *e*.

The problem is that, for example, when a positive effect of *e* on *Y* may also generate an impact on *X*. In this case, X can not be assumed to be constant, and we are not able to obtain unbiased estimates using OLS.

Sources – Omitted Variables

- Suppose 6 farms with 3 distinct land sizes (*A* in hectares);
- Suppose that, the larger the land size (*A*), the larger the agricultural production (*Y*);
- Imagine now that the total volume of credit accessed by each farm (*X*, in thousands) **has no** impact on agricultural production (*Y*). But those larger farms accessed more credit;

$A=2$	$A=2$	$A=4$	$A=4$	$A=6$	$A=6$
$Y=2000$	$Y=2200$	$Y=4200$	$Y=4000$	$Y=6200$	$Y=6000$
$X=2$	$X=4$	$X=6$	$X=8$	$X=10$	$X=12$

- If we relate the total volume of credit (*X*) with production (*Y*), without controls for land size, we can erroneously assume a positive relation between credit and production:

$Y=2000$	$Y=2200$	$Y=4200$	$Y=4000$	$Y=6200$	$Y=6000$
$X=2$	$X=4$	$X=6$	$X=8$	$X=10$	$X=12$

High values of *Y* are associated with high values of *X*, but *X* dos not determine *Y*.

Omitted Variables Bias

- Suppose that the population regression model is:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + e_i$$

- But we mistakenly consider the model:

$$Y_i = \tilde{\alpha} + \tilde{\beta}_1 X_1 + e_i$$

- The undue omission of X_2 in our model will bias the estimate of β_1.

- The bias in β_1 depends on both the value of β_2 and the correlation between X_1 and X_2. In general:

	Corr (X_1, X_2) > 0	**Corr (X_1, X_2) < 0**
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Exercise

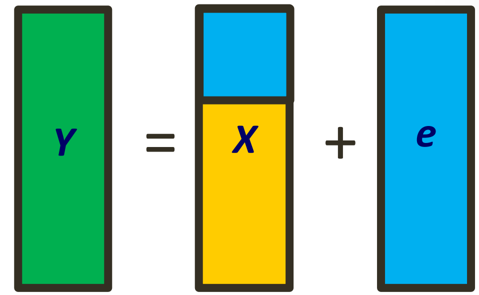
1) The dataset *Data_RelativeIncome.csv* contains a household smaple with information on relative income (average in the neighborhood) and income sufficiency (GORI MAIA, A. Relative Income, Inequality and Subjective Wellbeing: Evidence for Brazil. Social Indicators Research, v. 113, p. 1193-1204, n. 2013) :
   a) Analyze the relation between income sufficiency and log of relative income, without controls;
   b) Analyze the relation between income sufficiency and log of relative income, controlling for per capita income and other variables;

Intrumental Variables

We want to analyze: $Y_i = \alpha + \beta X_i + e_i$

But we have: $\text{Cov}(X_i, e_i) \neq 0$

The OLS estimators are biased even for large samples

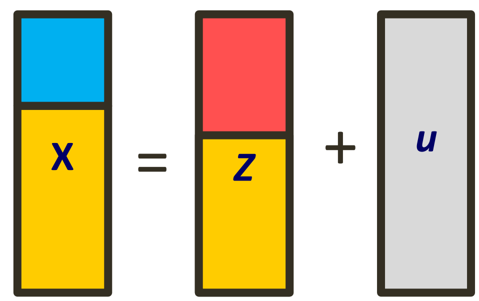


We need a instrument *Z* in such a way that:

$\text{Cov}(Z_i, e_i) = 0$ and $Cov(X_i, Z_i) \neq 0$

The portion of *Z* associated with *X* is:

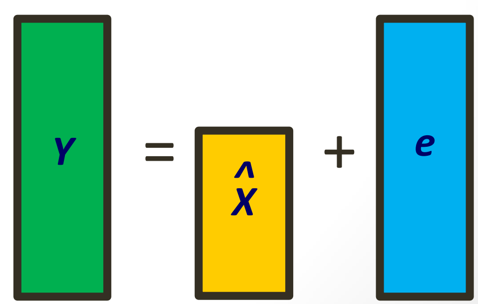
$$\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$$



The IV estimator is given by:

$$Y_i = \alpha + \beta \hat{X}_i + e_i$$

The IV estimator is consistent (unbiased for large samples) but can be biased for small samples



Two Stage Least Squares

Steps for the 2SLS:

1) **Identification:** we need at least one instrument for each endogenous regressor in the structural form;
2) **Reduced form**: algebraic transformation that defines each endogenous variable as a function of all exogenous variables (including instruments);
3) **Instrumental variable**: the predicted value of the reduced form for the endogenous variables;
4) **Structural form**: apply OLS after we replace the endogenous regressor by the instrumental variable predicted in the step 3;

The structural form is 1

$$Y_1 = \alpha + \beta_1 Y_2 + \beta_2 X + e$$

Z is the instrument for Y_2

Important

The 2SLS estimators are consistent but tend to be biased for small samples

The reduced form is: 2

$$Y_2 = \pi_0 + \pi_1 X + \pi_2 Z + u \xrightarrow{\text{OLS}} \hat{Y}_2 = \hat{\pi}_0 + \hat{\pi}_1 X + \hat{\pi}_2 Z \quad (3)$$

$$Y_1 = \alpha + \beta_1 \hat{Y}_2 + \beta_2 X + e \quad (4)$$

Example – Stata & R

- Suppose we have a model for *y*1 as a function of an endogenous regressor (*y*2), three exogenous controls (*x*1, *x*2 and *x*3) and two instruments for *y*2 (*z*1 and z2):

```
* ordinary least squares
regress y1 y2 x1 x2 x3

* two stage least squares
ivregress 2sls y1 (y2=z1 z2) x1 x2 x3
```

- The equivalent in R:

```
# ordinary least squares
ols <- lm(y1 ~ y2 + x1 + x2 + x3, data=dataname)
summary(ols)

# two stages least squares
tsls <- ivreg(y1 ~ y2 + x1 + x2 + x3 | .-y2 + z1 + z2, data=dataname)
summary(tsls)
```

Example – Python

- The equivalent in Python:

```
# ordinary least squares
x = pnad[['y2','x1','x2','x3']]
y = pnad['y1']
x = sm.add_constant(x)
ols = sm.OLS(y, x).fit()
print(ols.summary())

# package for 2sls
from linearmodels.iv import IV2SLS

# two stage least squares
pnad['const'] = 1
tsls = IV2SLS(dependent=pnad['y1'], \
               exog=pnad[['const','x1','x2','x3']],\
               endog=pnad['y2'], \
               instruments=pnad[['z1','z2']]).fit()
print(tsls.summary)
```

Exercise

1) The datase *Data_HealthIncome.csv* contains a household sample with information on health status and wage (MAIA, A. G. , RODRIGUES, C. G. . Saúde e mercado de trabalho no Brasil: diferenciais entre ocupados agrícolas e não agrícolas. Revista de Economia e Sociologia Rural (Impresso), v. 48, p. 737-765, n. 2010) :
    a) Analyze the relation between health status and wages using OLS;
    b) Analyze the relation between health status and wages using 2SLS;

Selection Bias

- We want to evaluate the impact of a program participation (*T*=0 or 1) on the outcome Y, controlling by **x** (vector of characteristics):

$$Y = \alpha + \boldsymbol{\beta}\mathbf{x} + \rho T + e$$

- But the selection of participants (*T*=1) and non-participants (*T*=0) is not random. This participation is defined by unobservable factors that are also related to the outcome *Y*, i.e.;

$$E(e|T) \neq 0$$

- Ideally, we wanted to estimate the *Average Treatment Effect* (ATE) by comparing the outcomes before the participation (Y_0) and after the participation (Y_1) for the same individuals.

$$ATE = E(Y_{1i} - Y_{0i})$$

- If we had a random selection:

$$ATE = E(Y_{1i} - Y_{0i}) = E(Y_i|T = 1) - E(Y_i|T = 0)$$

Matching

- Suppose a regression model with a treatment (*T*=1) and a control group (*T*=0) :

$$Y = \alpha + \boldsymbol{\beta}\mathbf{x} + \rho T + e$$

- Where *T* is not random and depends on non-observable factors :

$$E(e|T) \neq 0$$

- The *Propensity Score Matching* reduces the selection bias that is related to observable factors (**z**, which is a vector with characteristics determining both *Y* and *T*) by comparing treated and control individuals with similar characteristics (*propensity score* – p(**z**)):

$$p(\mathbf{z}) = prob(T = 1) = \boldsymbol{\pi}\mathbf{z} + u$$

- The treatment effect will be given by the *Average Effect of Treatment on the Treated* (ATT):

$$ATT = E[Y_{1i} - Y_{0i}|T_i = 1, p(\mathbf{z})] = E[Y_{1i}|T_i = 1, p(\mathbf{z}_i)] - E[Y_{0i}|T_i = 0, p(\mathbf{z}_i)]$$

Example – Stata & R

- Suppose we have a binary variable *T* designating a treatment that impacts the outcome *y*, and we also have three exogenous controls (*x*1, *x*2 and *x*3). The comparison between the OLS and PSM estimates in Stata can be given by:

```
* ols estimates for the impact of T on y
regress y T x1 x2 x3

* psm estimates for the impact of T on y
psmatch2 T x1 x2 x3, outcome(y)
```

- The equivalent in R:

```
# package for matching
library("MatchIt")

# ordinary least squares
ols <- lm(y ~ T + x1 + x2 + x3, data=mydata)
summary(ols)

# match mfa and non-mfa observations
matchobj <- matchit(T ~ x1 + x2 + x3, data=mydata)
summary(matchobj)                          # matchobj is an object
matchdata =match.data(matchobj)            # matchdata is a data frame with matched individuals
att <- lm(y ~ T, data=matchdata)     # mean comparison between treated and non-treated
summary(att)
```

Exercise

1) The datase *Data_MFA.xls* contains a household sample with information on the participation in the program *Mas Famílian en Accion* (MFA) in Colombia and poverty perception (MORALES MARTINEZ, D.; GORI MAIA, A. The impacts of cash transfers on subjective wellbeing and poverty: The case of Colombia. International Journal of Family and Economic Issues, 39(4), pp 616–633,2018) :
   a) Analyze the impact of the program MAF on poverty perception using OLS;
   b) Analyze the impact of the program MAF on the poverty perception using *propensity score matching*;