Applied Statistics and Econometrics
Lecture 7

Giuseppe Ragusa
gragusa@luiss.it

March 6, 2017

Outline

- Omitted variable bias
- Multiple regression
- OLS Measures of fit
- Sampling distribution of the OLS estimator

Omitted Variable Bias

- The error u arises because of factors, or variables, that influence Y but are not included in the regression function. There are always omitted variables.
- Sometimes, the omission of those variables can lead to bias in the OLS estimator.

Omitted variable bias, ctd.

- The **bias** in the OLS estimator that occurs as a result of an omitted factor, or variable, is called omitted variable bias.
- For omitted variable bias to occur, the omitted variable Z must satisfy two conditions:

The two conditions for omitted variable bias

1. Z is a determinant of Y (i.e. Z is part of u)
2. Z is correlated with the regressor X (i.e. $\text{corr}(Z, X) \neq 0$)

Both conditions must hold for the omission of Z to result in omitted variable bias.

Omitted variable bias, ctd.

Let's go back to the class size example:

$$testscore = \beta_0 + \beta_1 str + \underbrace{u}_{\beta_2 Z + \varepsilon}$$

- English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
- Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.

Accordingly, $\hat{\beta}_1$ is biased, that is, $\hat{\beta}_1 \xrightarrow{p} \beta_1 + bias$

- What is the direction of this bias? (That is, what is the sign of *bias*?)
- What does common sense suggest? If common sense fails you, there is a formula. . .

California Schools Data

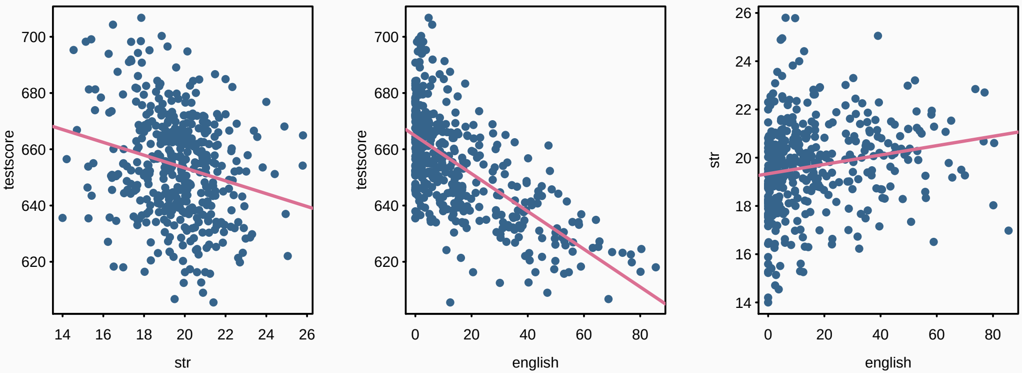
- The California School Dataset (`CASchool`) has data on the fraction of english learning in a district
- The variable is *english*

```
summary(CASchools$english)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       2       9      16      23      86
```

- Is this variable correlated, at least in the sample, with *str* and *testscore*?

California Schools Data



- Districts with fewer English Learners have higher test scores
- Districts with lower percent EL (PctEL) have smaller classes

Omitted variable bias, ctd.

A formula for omitted variable bias: recall the equation,

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{n}{n-1}s_X^2}$$

Under the assumption that $E[u_i|X_i] = 0$,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i\right] = E\left[(X_i - \bar{X})u_i\right] = \text{cov}(X_i, u_i) = 0$$

and, thus, there is not bias because:

$$E\left[\hat{\beta}_1\right] = \beta_1$$

But what if if $cov(X_i, u_i) \neq 0$?

Omitted variable bias, ctd.

Under assumptions LSA#2 and LSA#3 (even if LSA #1 is not true)

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{n}{n-1}s_X^2} \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

- If LSA #1 is correct, $\sigma_{Xu} = 0$ and
$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$
- If LSA #1 is incorrect, $\sigma_{Xu} \neq 0$ and
$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_{Xu}}{\sigma_X^2}$$

Omitted variable bias, ctd.

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\frac{\sigma_{Xu}}{\sigma_X^2}}_{\text{bias}}$$

- *Since*

$$u_i = \beta_2 Z_i + \varepsilon_i$$

the formula simplifies to

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\beta_2 \frac{\sigma_{XZ}}{\sigma_X^2}}_{\text{bias}}$$

Omitted variable bias, ctd.

Consider the class size example. It is very likely that:

- $\beta_2 < 0$ - as it is reasonable to assume that districts with more english learners have lower testscore (the sample analysis also seems to suggest so)
- $\sigma_{XZ} > 0$ - the covariance between *str* and *english* is probably positive (the sample analysis also seems to suggest so)
- Thus, the bias is probably negative

$$\underset{(-)}{\beta_2} \frac{\sigma_{XZ}}{\underset{(+)}{\sigma_X^2}} < 0$$

in which case we say that $\hat{\beta}_1$ is *downward biased*, that is, it is smaller than the “true” β_1.

Omitted variable bias, ctd.

The two conditions for the omitted variable bias can be expressed in terms of β_2 and σ_{XZ}

1. Z is a determinant of Y
   - $\implies \beta_2 \neq 0$
2. Z is correlated with the regressor X
   - $\implies \sigma_{XZ} \neq 0$

Three ways to overcome omitted variable bias

- Run a randomized controlled experiment in which (*str*) is randomly assigned: then *english* is still a determinant of TestScore, but *english* is uncorrelated with *str*. (This solution to OV bias is rarely feasible.)
- Adopt the "cross tabulation" approach, with finer gradations of *str* and *english* – within each group, all classes have the same *english*, so we control for *english* (But soon you will run out of data, and what about other determinants like family income and parental education?)
- Use a regression in which the omitted variable (*english*) is no longer omitted: include *english* as an additional regressor in a multiple regression. (That is what we will do next..)

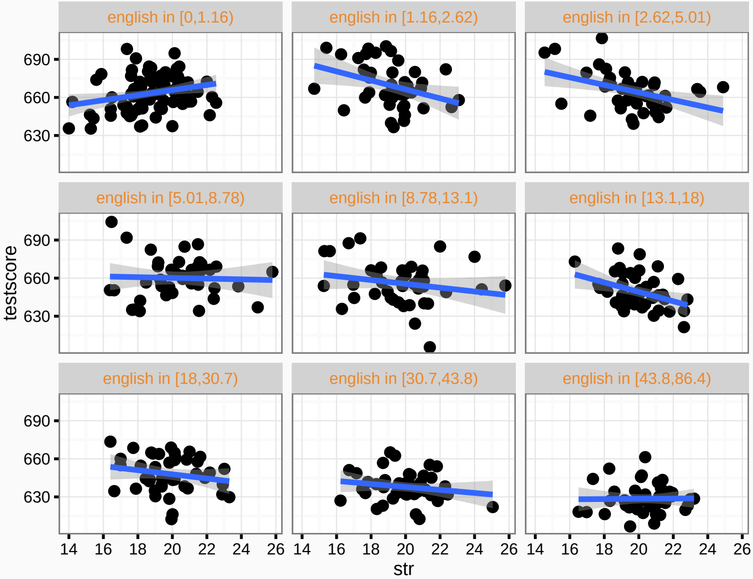
“Cross-tabulation”

- We “cut” *english* into intervals

```
##                              count
## english in [0,1.16)            84
## english in [1.16,2.62)         42
## english in [2.62,5.01)         42
## english in [5.01,8.78)         42
## english in [8.78,13.1)         42
## english in [13.1,18)           42
## english in [18,30.7)           42
## english in [30.7,43.8)         42
## english in [43.8,86.4)         42
```

- We estimate the linear model using data on each subset

Scatterplot



The Population Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots n$$

- Y is the dependent variable
    - $X_1, X_2, \ldots, X_n$ are the k independent variables (regressors)
- $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$ denote the ith observation on $Y_i, X_1, X_2, \ldots, X_k$
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding $X_2 \ldots . X_k$ constant
- β_2 = effect on Y of a change in X_2, holding $X_1, X_3, \ldots, X_k$ constant
- $\vdots$
- β_k = effect on Y of a change in X_k, holding $X_1 \ldots . X_{k-1}$ constant
- u_i= the regression error (omitted factors)
    - Satisfies $E[u_i | X_1, \ldots, X_k] = 0$

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots n$$

Consider changing X1 by ΔX1 while holding X2 constant:

- Population regression line before the change:

$$E[Y_i | X_{1i} = x_1, \ldots, X_{ki} = x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- Population regression line, after the change:

$$E[Y_i | X_{1i} = x_1 + \Delta x, \ldots, X_{ki} = x_k] = \beta_0 + \beta_1 (x_1 + \Delta x) + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\underbrace{E[Y_i | X_{1i} = x_1 + \Delta x, \ldots, X_{ki} = x_k] - E[Y_i | X_{1i} = x_1, \ldots, X_{ki} = x_k]}_{\text{effect of increasing } X_1 \text{by } \Delta x \text{unit, holding } X_2, \ldots, X_k \text{constant}} = \beta_1 \Delta x$$

The OLS Estimator in Multiple Regression

Assume for the moment that $k = 2$, that is, there are two regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- In this case the OLS estimator solves

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}))^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i the prediction (predicted value) based on the estimated line.
- Generalization of the case with one regressors

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i}))^2$$

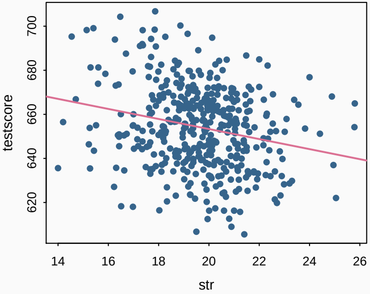
Graphic intuition

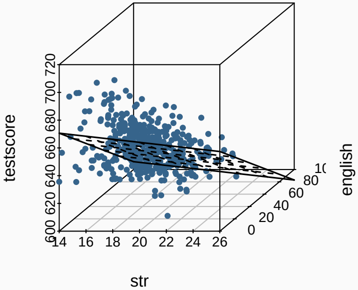
$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i})^2$$

Fits a line through points in $\mathbb{R}^2$

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

Fits a plane through points in $\mathbb{R}^3$





Matrix notation

The multivariate linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots n$$

can also be written in matrix form

$$Y = X\beta + u$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

OLS in Matrix Form

- Using matrix notation, the minimization of the residuals sum of squares can be compactly rewritten as

$$\min_{\beta}(Y - X\beta)'(Y - X\beta)$$

- The first order conditions are

$$X'(Y - X\beta) = 0 \implies \underbrace{X'X}_{(k+1)\times(k+1)} \underbrace{\beta}_{(k+1)\times 1} = \underbrace{\underbrace{X'Y}_{(k+1)\times 1}}$$

which is a system of linear equations (recall $Ax = b$, where $A = X'X$, $x = \beta$, and $b = X'Y$)

- From which we obtain the OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

provided $(X'X)^{-1}$ is invertible (determinant $\neq 0$)

Example: the California test score data

```
lm(testscore ~ str, data = CASchools)
```

```
##
## Call:
## lm(formula = testscore ~ str, data = CASchools)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

```
lm(testscore ~ str + english, data = CASchools)
```

```
##
## Call:
## lm(formula = testscore ~ str + english, data = CASchools)
##
## Coefficients:
## (Intercept)          str      english
##      686.03        -1.10        -0.65
```

- What happens to the coefficient on *str*?

Measures of Fit for Multiple Regression

- SER $=$ std. deviation of $\hat{u}_i$ (with d.f. correction)
- RMSE $=$ std. deviation of $\hat{u}_i$ (without d.f. correction)
- R^2= fraction of variance of Y explained by $X_1, \ldots, X_k$
- $\bar{R}^2$= "adjusted R2" $= R^2$ with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the SER and the RMSE are measures of the spread of the Ys around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

R^2 and $\bar{R}^2$ (adjusted R^2)

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}}) = \sum_{i=1}^{n} \hat{u}_i^2$$

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

The R^2 always increases when you add another regressor (why?) – a bit of a problem for a measure of “fit”

R^2 and $\bar{R}^2$ (adjusted R^2), ctd.

The $\bar{R}^2$ (the "adjusted R2") corrects this problem by "penalizing" you for including another regressor – the does not necessarily increase when you add another regressor.

$$\begin{aligned}\bar{R}^2 &= 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS} \\ &= R^2 - \left(\frac{k}{n-k-1}\right)\frac{SSR}{TSS}\end{aligned}$$

Note that $\bar{R}^2 < R^2$, however if n is large

$$\left(\frac{k}{n-k-1}\right) \to 0,$$

and the two will be very close.

California Example

- Regression of *testscore* against *str*

$$testscore = 698.93 - 2.28\, str, \quad R^2 = 0.0512$$

- Regression of *testscore* against *str* and *english*

$$testscore = 686.032 - 1.101\, str - 0.65\, english, \quad R^2 = 0.426$$

What – precisely – does this tell you about the fit of univariate regression compared with the bivariate regression?

The Least Squares Assumptions for Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots n$$

1. The conditional distribution of u given the X's has mean zero, that is,

   $$E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0,$$

   for all $(x_1, \ldots, x_k)$
2. $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, are i.i.d.
3. Large outliers are unlikely, $X_1, \ldots, X_k$ and Y have four moments

   $$E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y^4) < \infty$$

4. There is no perfect multicollinearity

Assumption #1: the conditional mean of u given the included Xs is zero.

$$E(u_i|X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to omitted variable bias, specifically, if an omitted variable *(a)* belongs in the equation (so is in u) and *(b)* is correlated with an included X then this condition fails and there is OV bias.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (we will see this later)

Assumption #2: $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, are i.i.d.

- This is the same assumption as we had before for a single regressor.
- This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: Large outliers are rare (finite fourth moments)

- This is the same assumption as we had before for a single regressor.
- As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

Assumption #4: Perfect multicollinearity

- **Perfect multicollinearity** is when one of the regressors is an exact linear function of the other regressors.
- If there is perfect multicollinearity, the OLS problem is not defined
- Modern computer software has ways to handle this problem. For instance, R drops one of the collinear variable

```
CASchools[["var0"]] <- CASchools[["str"]] * 2 + 5
## var0 is a linear transformation of str
lm(testscore ~ str + var0, data = CASchools)
```

```
##
## Call:
## lm(formula = testscore ~ str + var0, data = CASchools)
##
## Coefficients:
## (Intercept)          str         var0
##      698.93        -2.28           NA
```

Assumption #4: Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by “dropping” one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity, ctd.

- Imperfect and perfect multicollinearity are quite different despite the similarity of the names.
- **Imperfect multicollinearity** occurs when two or more regressors are very highly correlated.
    - ***Why the term "multicollinearity***"? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are "co-linear" – but unless the correlation is exactly ± 1, that collinearity is imperfect.
- One of the consequences of imperfect multicollinearity is that the standard errors of the coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanatory. (*Why?*)

Dummy Variable Trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Think of region of residence: Sicily, Lazio, Tuscany, etc.). If you include all these dummy variables and a constant, you will have perfect multicollinearity – this is sometimes called the dummy variable trap.

- Why is there perfect multicollinearity here?
- Solutions to the dummy variable trap:
  1. Omit one of the groups (e.g. Lazio), **or**
  2. Omit the intercept
- What are the implications of (1) or (2) for the interpretation of the coefficients?

We will see this later on with an example.

Hypothesis Tests and Confidence Intervals for a Single Coefficient

- Hypothesis tests and confidence intervals for a single coefficient in multiple regression follow the same logic and recipe as for the slope coefficient in a single-regressor model.
- Since

  $$\frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{SE(\hat{\beta}_1)}, \frac{\hat{\beta}_2 - E[\hat{\beta}_2]}{SE(\hat{\beta}_2)}, \dots, \frac{\hat{\beta}_k - E[\hat{\beta}_k]}{SE(\hat{\beta}_k)}$$

  are approximately distributed $N(0,1)$ (because, under the assumptions just given, the Central Limit Theorem applies)
- The hypothesis on β_1 (or any other coefficient) can be tested using the usual t-statistic

  $$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

  , and confidence intervals are constructed as

  $$(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1))$$
- The standard errors of the estimated coefficients are calculated as in the single regression case (*mutatis mutandis).*

Example: The California class size data

1. Single coefficient regression

```
summary_rob(lm(testscore ~ str, data = CASchools))

##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  698.933     10.364   67.44  < 2e-16
## str           -2.280      0.519   -4.39  1.1e-05
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 19 on 418 degrees of freedom
## Multiple R-squared:  0.0512,Adjusted R-squared:  0.049
## F-statistic: 19.3 on 1 and Inf DF,  p-value: 1.14e-05
```

Example: The California class size data, ctd.

1. Multiple regression

```
summary_rob(lm(testscore ~ str + english, data = CASchools))
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  686.032      8.728   78.60   <2e-16
## str           -1.101      0.433   -2.54    0.011
## english       -0.650      0.031  -20.94   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 14 on 417 degrees of freedom
## Multiple R-squared:  0.426,Adjusted R-squared:  0.424
## F-statistic:  448 on 2 and Inf DF,  p-value: <2e-16
```

Example: The California class size data, ctd.

- The coefficient on *str* in (2) is the effect on *testscore* of a unit change in *str*, **holding constant** the percentage of English Learners in the district
- The coefficient on *str* falls by one-half (why?)
- The 95% confidence interval for coefficient on STR in (2) is

$$(-1.10 \pm 1.96 \times 0.43) = (-1.95, -0.26)$$

- The t-statistic testing $H_0 : \beta_{str} = 0$ is

$$t = \frac{-1.10}{0.43} = -2.54,$$

  so we **reject** the hypothesis at the **5%** significance level
- We use heteroskedasticity-robust standard errors – for exactly the same reason as in the case of a single regressor.