Econ 422 – Lecture Notes
Part III

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

Introduction to Multiple Regression

Outline

1. Omitted variable bias
2. Causality and regression analysis
3. Multiple regression and OLS
4. Measures of fit
5. Sampling distribution of the OLS estimator

Omitted Variable Bias

The error u arises because of factors that influence Y but are not included in the regression function; so, 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 is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted factor "Z" must be:

1. A determinant of Y (i.e. Z is part of u); **and**

2. 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.

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
2. 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. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula…

Omitted variable bias, ctd.

A formula for omitted variable bias: recall the equation,

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

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$. Under Least Squares Assumption 1,

$$E[(X_i - \mu_X)u_i] = \text{cov}(X_i,u_i) = 0.$$

But what if $E[(X_i - \mu_X)u_i] = \text{cov}(X_i,u_i) = \sigma_{Xu} \neq 0$?

Omitted variable bias, ctd.

In general (that is, even if Assumption #1 is not true),

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

$$\xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu},$$

where ρ_{Xu} = corr(X,u). If assumption #1 is valid, then $\rho_{Xu} = 0$, but if not we have….

The omitted variable bias formula:

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}}$$

If an omitted factor Z is ***both***:

(1) a determinant of Y (that is, it is contained in u); ***and***

(2) correlated with X,

then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased (and is not consistent).

The math makes precise the idea that districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the ESL factor results in overstating the class size effect.

Is this is actually going on in the CA data?

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall "test score gap" = 7.4)

Digression on causality and regression analysis

What do we want to estimate?

- What is, precisely, a causal effect?
- The common-sense definition of causality isn't precise enough for our purposes.
- In this course, we define a causal effect as the effect that is measured in an ***ideal randomized controlled experiment***.

Ideal Randomized Controlled Experiment

- *Ideal*: subjects all follow the treatment protocol – perfect compliance, no errors in reporting, etc.!
- *Randomized*: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- *Controlled*: having a control group permits measuring the differential effect of the treatment
- *Experiment*: the treatment is assigned as part of the experiment: the subjects have no choice, so there is no "reverse causality" in which subjects choose the treatment they think will work best.

Back to class size:

- Conceive an ideal randomized controlled experiment for measuring the effect on *Test Score* of reducing *STR*…
- How does our observational data differ from this ideal?
  - The treatment is not randomly assigned
  - Consider *PctEL* – percent English learners – in the district. It plausibly satisfies the two criteria for omitted variable bias: $Z = PctEL$ is:
    1. a determinant of Y; ***and***
    2. correlated with the regressor X.
  - The "control" and "treatment" groups differ in a systematic way – $\text{corr}(STR, PctEL) \neq 0$

- ***Randomized controlled*** experiments:
  - Randomization + control group means that any differences between the treatment and control groups are random – not systematically related to the treatment
  - We can eliminate the difference in *PctEL* between the large (control) and small (treatment) groups by examining the effect of class size among districts with the same *PctEL*.
  - If the only systematic difference between the large and small class size groups is in *PctEL*, then we are back to the randomized controlled experiment – within each *PctEL* group.
  - This is one way to "control" for the effect of *PctEL* when estimating the effect of *STR*.

Return to omitted variable bias

Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment (*STR*) is randomly assigned: then *PctEL* is still a determinant of *TestScore*, but *PctEL* is uncorrelated with *STR*. (*But this is unrealistic in practice.*)
2. Adopt the "cross tabulation" approach, with finer gradations of *STR* and *PctEL* – within each group, all classes have the same *PctEL*, so we control for *PctEL* (*But soon we will run out of data, and what about other determinants like family income and parental education*?)
3. Use a regression in which the omitted variable (*PctEL*) is no longer omitted: include *PctEL* as an additional regressor in a multiple regression.

The Population Multiple Regression Model

Consider the case of two regressors:

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

- Y is the *dependent variable*
- X_1, X_2 are the two *independent variables* (*regressors*)
- (Y_i, X_{1i}, X_{2i}) denote the i^{th} observation on Y, X_1, and X_2.
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding X_2 constant
- β_2 = effect on Y of a change in X_2, holding X_1 constant
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

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

Consider changing X_1 by ΔX_1 while holding X_2 constant:

Population regression line ***before*** the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Population regression line, ***after*** the change:

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

Before: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

After: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Difference: $\Delta Y = \beta_1 \Delta X_1$

So:

$\beta_1 = \dfrac{\Delta Y}{\Delta X_1}$, **holding X_2 constant**

$\beta_2 = \dfrac{\Delta Y}{\Delta X_2}$, **holding X_1 constant**

β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression

With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of β_0, β_1, and β_2.**

Example: the California test score data

Regression of *TestScore* against *STR*:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

- What happens to the coefficient on *STR*?
- Why? (*Note*: corr(*STR*, *PctEL*) = 0.19)

Multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

More on this printout later…

Measures of Fit for Multiple Regression

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

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

$\bar{R}^2$ = "adjusted R^2" = 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 *Y*'s 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 $\bar{R}^2$

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

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

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

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

$$\textit{Adjusted } R^2\text{:} \quad \bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.

Measures of fit, ctd.

Test score example:

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR,$

$$R^2 = .05,\ SER = 18.6$$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65PctEL,$

$$R^2 = .426,\ \bar{R}^2 = .424,\ SER = 14.5$$

- *What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?*
- *Why are the R^2 and the $\bar{R}^2$ so close in (2)?*

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, $E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare: $X_1,\ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty,\ldots, E(X_{ki}^4) < \infty$, $E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

Assumption #1: the conditional mean of *u* given the included *X*'s is zero.

$$E(u|X_1 = x_1,\dots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in u) and (2) is correlated with an included X, then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution – ***if possible*** – is to include the omitted variable in the regression.

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

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: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
    str |  (dropped)
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression, β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples…

With these least squares assumptions in hand, we now can derive the sampling dist'n of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

The Sampling Distribution of the OLS Estimator

Under the four Least Squares Assumptions,

- The exact (finite sample) distribution of $\hat{\beta}_1$ has mean β_1, $\text{var}(\hat{\beta}_1)$ is inversely proportional to n; so too for $\hat{\beta}_2$.
- Other than its mean and variance, the exact (finite-n) distribution of $\hat{\beta}_1$ is very complicated; but for large n…
- $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (law of large numbers)
- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
- So too for $\hat{\beta}_2, \ldots, \hat{\beta}_k$

Conceptually, there is nothing new here!

Multicollinearity, Perfect and Imperfect

<u>Some more examples of perfect multicollinearity</u>

- The example from earlier: you include *STR* twice.
- Second example: regress *TestScore* on a constant, D, and B, where: $D_i = 1$ if $STR \leq 20$, $= 0$ otherwise; $B_i = 1$ if $STR > 20$, $= 0$ otherwise, so $B_i = 1 - D_i$ and there is perfect multicollinearity
- Would there be perfect multicollinearity if the intercept (constant) were somehow dropped (that is, omitted or suppressed) in this regression?
- This example is a special case of…

The 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 (Freshmen, Sophomores, Juniors, Seniors, Other). 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. Senior), or
  2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

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

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 this term? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are collinear – but unless the correlation is exactly ± 1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- Intuition: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data are pretty much uninformative about what happens when X_1 changes but X_2 doesn't, so the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.
- The math? See SW, App. 6.2

Next topic: hypothesis tests and confidence intervals…