Econ 1123: Section 2
Elena Laulet

September 16, 2010

Grade

Grade:
- Problem sets: 30%
- Midterm (October 14): 25%
- Final (TBD): 45%

Problem sets:
- graded from 0-10
- -5 points if handed in before solutions are posted
- -10 points if handed in after solutions are posted
- The lowest grade will be dropped
- Problem set 8 cannot be dropped and counts double

In both the midterm and the final, you are allowed a "cheat-sheet"

Contact Information

- Sections are voluntary.
- My office hours are Thursdays 5pm-7pm in Littauer Mezzanine 34-36 (Note room change)
- You can email me administrative questions to ellaudet@gmail.com. If you have substantive questions, please come to my office hours.
- Any substantial STATA related questions should be sent to the STATA TF: Shai (shaiber@gmail.com)

Problem Sets

- To complete the problem sets, we encourage you to work in small groups (maximum of 3). Each person, however, must submit their own write-up.
- You must submit a hard copy of each problem set in class (usually on Tuesdays).
- Make sure to indicate the name of your TF and the name of your collaborators at the beginning of the problem set.
- Also make sure to attach your .do file at the end of the problem set.
General Information

- Problem set 1 has been graded (on a 0-10 scale). Problem sets that are not picked up in section will be left in a box in Littauer 200A.
- The solutions should already be posted on the course website. I highly encourage you for you to take a look at them.
- I post the handouts that I use in section on the website (under Section Pages/Elena’s Sections).
- The handouts on the web contain the answers to the questions raised in section.

Outline

1. Review
2. Binary Regressors
3. Bivariate Regression
   - Omitted Variable Bias
4. Multivariate Regression
   - Interpretation of Coefficients
   - Joint Hypothesis Testing
5. STATA Help for Problem Set 2

Review of Last Week
Based on Common Mistakes in Problem Set 1

A few general comments that apply to problem sets as well as exams
- Make sure that you are answering the question. Try to say no more and no less.
- In your answers, round numbers to two or three decimals.
- Pay attention to the signs. An effect of -3 is not the same as an effect of +3.
- If at all possible, make it easy for us to grade your answers (e.g., good handwriting if not typing the answers, don’t cram your answers, ...).

Do not forget to mention the units in the interpretation of coefficients
- The real-world magnitude of effects
- SE (ΔT \( \hat{\beta} \)) = ΔT SE (\( \hat{\beta} \))
- Meaning of Heteroskedasticity/Homoskedasticity
  - What do we learn about the assumption of homoskedasticity if the standard errors of the coefficients change significantly depending on whether we run the regression with robust standard errors or not? Nothing.
  - Meaning of RMSE
  - Remember that with one binary/dummy regressor the \( \hat{\beta} \) has a special interpretation.
Interpretation of $\hat{\beta}$s of Binary Regressors

Example from last week’s section slides:

Test Scores$_i = \beta_0 + \beta_1$ Private School$_i + u_i$

$\hat{\beta}_0 = [\text{Test Scores}_i | \text{Private Schools}_i = 0]$

$\hat{\beta}_1 = [\text{Test Scores}_i | \text{Private Schools}_i = 1] - [\text{Test Scores}_i | \text{Private Schools}_i = 0]$

Generally: $Y_i = \beta_0 + \beta_1 D_i + u_i$ where $D_i$ is a dummy var.

$\hat{\beta}_0 = [Y | D = 0]$

$\hat{\beta}_1 = [Y | D = 1] - [Y | D = 0]$

What if you try to run:

$TS_i = \beta_0 + \beta_1 \text{Public School}_i + \beta_2 \text{Private School}_i + u_i$?

You will face a problem of perfect multicollinearity (also known as the dummy variable trap).

For all observations: $\text{Public School}_i - \text{Private School}_i = 1 = \text{constant term}$

So, one of two things will happen. Either you won’t be able to run the regression or the software will automatically drop one of the variables of the regression.

Bivariate Regression: An Example

What would be the interpretation of coefficients then if you run:

$TS_i = \beta_0 + \beta_1 \text{Public School}_i + \beta_2 \text{Private School}_i + u_i$?

$\hat{\beta}_0 = [\text{Test Scores}_i | \text{Public Schools}_i = 1]$

$\hat{\beta}_1 = [\text{Test Scores}_i | \text{Private Schools}_i = 1]$

Generally: $Y_i = \beta_0 D_1 i + \beta_1 D_2 i + u_i$ where $D_1 i$ and $D_2 i$ are dummy variables such as $D_1 i + D_2 i = 1$

$\hat{\beta}_0 = [Y | D_1 = 1]$

$\hat{\beta}_1 = [Y | D_2 = 1]$

Test Scores$_i = 4.9 + 2 \text{ Private School}_i + u_i$

\[0.4 \quad 0.8\]
Omitted Variable Bias

(A violation of the first OLS assumption)

Omitted Variable Bias occurs when one (or more) of the omitted variables in the regression:

(a) is correlated with the independent variable of interest, AND
(b) the omitted variable is a determinant of our dependent variable.

Consequences:
The OLS estimator is biased \((E(\hat{\beta}) \neq \beta)\) and inconsistent \((\hat{\beta} \text{ does not converge to } \beta)\).
\[
\hat{\beta} \rightarrow P \beta + \rho_{X,u} \frac{\sigma_u}{\sigma_X},
\]
where \(\rho_{X,u}\) is the correlation between \(X\) & \(u\).

How can we identify the sign or direction of the bias?

Supposing \(Z\) to be the omitted variable determinant of \(Y\) and correlated with \(X\), then:
\[
\text{Sign of the OVB} = \text{Sign of } \text{corr}(Z,X) \times \text{Sign of } \text{corr}(Z,Y)
\]

Omitted Variable Bias in our Example

Going back to our example, we can think of many variables that are both related with the probability of attending a private school and that are determinants of students’ performance. The socio-economic status of a student, for example, will affect the likelihood that he or she attend a private school and it will also directly affect the performance of the student in that school.

\[
\begin{align*}
\text{SES}_i & \quad \downarrow & \quad \downarrow \\
\text{Private Schools}_i & \quad \rightarrow & \quad \text{Test Scores}_i
\end{align*}
\]

If this is true, then, we say that \(\text{SES}_i\) is an omitted variable and the coefficient on private school would suffer from omitted variable bias.

What would be the expected sign of the bias in this case?

We shall expect \(\text{SES}_i\) and Private Schools\(_i\) to be positively correlated. And, we shall expect \(\text{SES}_i\) and Test Scores\(_i\) to be also positively correlated. So \((+ \times +) = (+)\)

Does this mean that omitting \(\text{SES}_i\) from our regression is making \(\beta\) too small or too large? Too large.

What can we do to solve the problem?: If the omitted variable is measurable, we can eliminate the bias by including the variable in the regression.

How can we interpret the distance between the data and the line?

There are factors that affect students’ performance other than the type of school attended by the students. The errors of the regression represent all of the omitted factors that affect test scores.

Does the existence of these error terms mean that the regression suffers from omitted variable bias? No.

Omitted Variable Bias

Going back to our example, we can think of many variables

that are both related with the probability of attending a private school and that are determinants of students’ performance. The socio-economic status of a student, for example, will affect the likelihood that he or she attend a private school and it will also directly affect the performance of the student in that school.

\[
\begin{align*}
\text{SES}_i & \quad \downarrow & \quad \downarrow \\
\text{Private Schools}_i & \quad \rightarrow & \quad \text{Test Scores}_i
\end{align*}
\]

If this is true, then, we say that \(\text{SES}_i\) is an omitted variable and the coefficient on private school would suffer from omitted variable bias.
Econ 1123: Section 2
Review
Binary Regressors
Bivariate Regression
Omitted Variable Bias
Multivariate Regression
Interpretation of Coefficients
Joint Hypothesis Testing
STATA Help for Problem Set 2

Multivariate Regression: An Example

Population Linear Regression Model:
Test Scores_i = \beta_0 + \beta_1 \text{ Private Schools}_i + \beta_2 \text{ SES}_i + u_i

We have a data set of 61 observations.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>scores</td>
<td>61</td>
<td>5.459016</td>
<td>2.894228</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>private</td>
<td>61</td>
<td>.360557</td>
<td>.494758</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ses</td>
<td>61</td>
<td>1.819672</td>
<td>.785452</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Regression Results

Test Scores_i = \hat{\beta}_0 + \hat{\beta}_1 \text{ Private Schools}_i + \hat{\beta}_2 \text{ SES}_i + \hat{u}_i

<table>
<thead>
<tr>
<th>scores</th>
<th>Coef.</th>
<th>Robust Std. Err.</th>
<th>t</th>
<th>Prob</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>.7395488</td>
<td>.4807247</td>
<td>1.51</td>
<td>.137</td>
<td>-1.417425</td>
</tr>
<tr>
<td>ses</td>
<td>1.327975</td>
<td>.2552757</td>
<td>4.80</td>
<td>.000</td>
<td>.985017</td>
</tr>
<tr>
<td>_cons</td>
<td>2.786094</td>
<td>.5349354</td>
<td>5.21</td>
<td>.000</td>
<td>1.715764</td>
</tr>
</tbody>
</table>

Interpretation of Coefficients

What is \hat{\beta}_0 and what is its interpretation? \hat{\beta}_0 = 2.8
It is the intercept of the regression. In this case it does not have substantive interpretation. (SES never takes values of 0).

What is \hat{\beta}_1 and what is its interpretation? \hat{\beta}_1 = .7
It indicates the change in test scores associated with attending a private school (instead of a public school), while holding SES constant (or, ceteris paribus).

Is this big? This is about 1/3 of a standard deviation of test scores. (Test scores have a standard deviation of 2.09 points).

Is it statistically significant at the 5% significance level? No. \hat{\beta}_1/SE(\hat{\beta}_1) = .7/ .49 < 1.96, so we cannot reject the null hypothesis that Private Schools, is = 0 at the 5% significance level.

How does \hat{\beta}_1 compare to the \hat{\beta}_1 that we got with the regression that did not control for SES_i (e.g., is this smaller or larger; is it more or less statistically significant)?

The \hat{\beta}_1 estimated in the model that controls for SES_i is smaller, which is what we would expect since we expected that omitting SES_i from the regression was positively biasing the coefficient on Private School;
While in the regression without SES Private Schools_i was statistically significant at the 5% level, here it is not.
What is $\hat{\beta}_2$ and what is its interpretation? $\hat{\beta}_2 = .3$

It indicates the change in test scores associated with a one unit change in the socio-economic status of the student, while holding the type of school attended constant (or, ceteris paribus).

Is this big? This is more than 1/2 of a standard deviation of test scores. (Test scores have a standard deviation of 2.09 points).

Is it statistically significant at the 5% significance level? Yes. $\hat{\beta}_2 / SE(\hat{\beta}_2) = 1.3/.3 > 1.96$, so we reject the null hypothesis that $ses_i$ is $= 0$ at the 5% significance level.

Suppose we want to test:

$H_0 : \beta_1 = \beta_2 = 0$

$H_A : \beta_1 \neq 0$ or $\beta_2 \neq 0$.

How shall we interpret this test?

We are testing whether neither of our two independent variables matter (i.e., whether neither of them are associated with changes in test scores).

To test joint hypothesis, we use the F-test, which in large samples is distributed as $F_{q,\infty}$ and can be approximated with $\chi^2 / q$; where $q$ is the number of restrictions under the null.

What is $q$ in this case? $q = 2$.

Shall we reject the null? Yes. Because the probability of getting a value greater than the value actually obtained is around 0.0000, then, we reject the null hypothesis at all of the conventional levels of significance.
Is it ever possible for two variables to be jointly significant but individually insignificant? Yes. If the variables are correlated (if there is imperfect multicollinearity between the two variables).

Is it ever possible for two variables to be individually significant but jointly insignificant? No, by definition.

- To sort the data, use the command "sort [NAME OF VARIABLE TO USE FOR THE SORTING]"
- To list the data, use the command "list [NAME OF VARIABLES TO BE LISTED]"
- To drop an observation, use the command "drop if [CONDITION TO BE USED FOR THE DROPPING: EXAMPLE: X > 3]"