Econ 1123: Section 1
Elena Llaudet

September 9, 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 219
- You can email me administrative questions to ellaudet@gmail.com. If you have substantive questions, please come to my office hours.

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.
- Problem sets will be returned in section.
- Problem sets that are not picked up in section will be left in a box outside of Betsy Stuppard’s office (Littauer Center 321)
- The solutions will be posted on the course website usually two days after they are due.
- Problem Set 1 is due in class on **Tuesday, September 14.**
**Bivariate Regression: An Example**

Population Linear Regression Model:
\[
\text{Test Scores}_i = \beta_0 + \beta_1 \text{SES}_i + u_i
\]

Where \(\text{SES}_i\) is a measure of the socio-economic status of the students. It goes from 1 to 3.

We have a data set of 35 observations.

**What is \(u_i\)?** They are the regression errors. They represent all the omitted factors that affect test scores.
Regression Results

Test Scores\(_i\) = \(\hat{\beta}_0 + \hat{\beta}_1\) SES\(_i\) + \(\hat{u}_i\)

| scores | Coef. | Robust Std. Err. | t | P>|t| | [95% Conf. Interval] |
|--------|-------|------------------|---|------|----------------------|
| ses    | 1.61858 | .347907         | 4.65 | 0.000 | .911158, 2.326882    |
| _cons  | 2.589235 | .7423807       | 3.49 | 0.001 | 1.07885, 4.09962     |

**Interpretation of Coefficients**

What is \(\hat{\beta}_0\) and what is its interpretation? \(\hat{\beta}_0 = 2.6\).

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? \(\hat{\beta}_1 = 1.6\).

It is the slope of the regression. It indicates how much change in test scores will be related to a one unit change in SES. In this case, for example, moving from an SES of 1 to an SES of 2 will be related with an increase in test scores of 1.6 points (in a scale of 1 to 10). **Is this big?** This is about 3/4 of a standard deviation of test scores. (Test score have a standard deviation of 2.16. If test scores were normally distributed, 95% of the values would be within two standard deviations of the mean.)

**Plot**

```
Code:
predict scoreshat
scatter scores scoreshat ses, msymbol(o i)
connect(. l)
```

**R\(^2\) Interpretation**

\[
R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{Explained sum of the squares (ESS)}}{\text{Total sum of the squares (TSS)}}
\]

\(0 \leq R^2 \leq 1\)

What is the \(R^2\) of our regression and what is its interpretation? \(R^2 = .33\).

It means that only 33% of the variation in test scores among the students is explained by their variation in their socio-economic status.

**Note:** A high \(R^2\) does not mean that the estimates are unbiased or statistically significant. The goal of an analysis is to determine the true coefficients, not to maximize \(R^2\).
Standard Error of the Regression Interpretation (SER)

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

It is a measure of the spread of the observations around the regression line. This is roughly the sample standard deviation of $\hat{u}$. So, it is in the same units as $Y$.

What is the SER of our regression and what is its interpretation? SER $\approx 1.8$ (SER $\approx$ Root Mean SE). This means that the sample standard deviation of $\hat{u}$ is roughly 1.8 points (in a scale of 1 to 10).

Hypothesis Testing - Terminology

The significance level of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

P-value: Probability of drawing a statistics (e.g. $\hat{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

Hypothesis Testing

Suppose we want to test:

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$.

How shall we interpret this test?

We are testing whether changes in SES are related in any way with changes in test scores. If we reject the null, it would mean that we have enough evidence to say that changes in SES are related with changes in test scores.

Hypothesis Testing: Steps

1. Construct the t-statistic: $t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of estimator}}$

   In our case, $t = \frac{1.6 - 0}{0.35} = 4.6$

   If $n$ is large, $t \approx N(0,1)$ (Central Limit Theorem)
2. Calculate the p-value: 
\[ p-value = P_{H_0}(|t| > |t_{act}|) \]
In our case,
\[ p-value = P_{H_0}(|t| > 4.6) \]
\[ p-value = 2\varphi(-4.6) = 0.0000 \]

3. Compare p-value to significance levels (\( \alpha \)) and decide whether we have enough evidence to reject the null hypothesis. In our case,
\[ p-value < \alpha ; \forall \alpha = \{0.001, 0.05, 0.01\} \]
We, therefore, reject the null hypothesis at all of the conventional levels of significance.

Confidence Intervals (CI): Example 95% CI of \( \beta_1 \)

\[
95\% \text{ Confidence Interval } = \{ \hat{\beta}_1 \pm 1.96 \ SE(\hat{\beta}_1) \} \\
= \{ 1.6 \pm 1.96 \times 0.35 \} \\
= (0.91, 2.29) 
\]

**What is the interpretation?** The true value of \( \beta_1 \) will be included in the 95% CI of \( \hat{\beta}_1 \) with .95 probability in repeated samples. In other words, if we draw 100 samples, only 95 of those will allow us to construct 95% CI where the true \( \beta_1 \) is included.

Does this mean that we know for sure that the true \( \beta_1 \) falls within (0.91, 2.29)? No. This sample might be one of the 5 samples out of 100 that has a 95% confidence interval that does not include \( \beta_1 \).

A Special Case: Binary Independent Variable

Population Linear Regression Model:
\[ \text{Test Scores}_i = \beta_0 + \beta_1 \text{ Private School}_i + u_i \]

Where Private School\(_i\) is a dummy variable that indicates whether the student attended a private school. The baseline category is public schools.

We have school data for our 35 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>35</td>
<td>5.457143</td>
<td>2.108357</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>private</td>
<td>35</td>
<td>0.7857143</td>
<td>0.4563492</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
What is $\hat{\beta}_0$ and what is its interpretation? $\hat{\beta}_0 = 4.9$
It is the average test score of the students attending public schools. Generally: $\hat{\beta}_0 = \mathbb{E}[Y \mid D = 0]$

When Private$_i = 0$: Test Scores$_i = 4.9 + 2 \times 0 = 4.9$

What is $\hat{\beta}_1$ and what is its interpretation? $\hat{\beta}_1 = 2.0$
It indicates that private school students have on average 2-point higher test scores than public school students. Generally: $\hat{\beta}_1 = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$

When Private$_i = 1$, Test Scores$_i = 4.9 + 2 \times 1 = 6.9$

Code:

```stata
predict scoreshat, xb
scatter scores scoreshat private, msymbol(o i) connect(. l)
```
The 3 OLS Assumptions

For our last example:
\[
\text{Test Scores}_i = \beta_0 + \beta_1 \text{ Private School}_i + u_i
\]
\begin{itemize}
  \item E \[ u_i | \text{ Private}_i \] = 0
  \item (Test Scores, Private) are i.i.d
  \item Large outliers are rare.
\end{itemize}

Only under these assumptions, the OLS estimator is unbiased and consistent.

Are they all reasonable assumptions in our example?
Let’s look at each one of them separately.

Other two assumptions

2. (Test Scores, Private) are i.i.d
Example of a violation of this assumption: If some of our students come from the same schools, then those observations will not be independent. Those observations will be nested within schools and we will have to correct for this using clustered standard errors.

3. Large outliers are rare
This assumption seems reasonable given the data at hand.

A fourth assumption: Homoskedasticity

Some times we can safely assume \( \text{var} (u_i | X_i) \) is constant. When that is the case, we say that we assume \( u_i \) to be homoskedastic. This will change the formulas for the calculation of variances and standard errors. This is how STATA calculates regressions by default.

However, most of the time in economic analyses the variance of the error term depends on one or more of the regressors and so we have heteroskedastic errors. To run a regression in STATA allowing for heteroskedasticity, you must add “,robust” at the end of the “regress” command.

Can we think of any variables that would be related both with the probability of attending a private school and with the student 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 affect the performance of the student in that school.

\[
\text{SES}_i \quad \downarrow \quad \downarrow
\]

Private Schools \[ i \] \( \rightarrow \) Test Scores \[ i \]

If this is true, then, we say that \( \text{SES}_i \) is an omitted variable and the regression results would suffer from omitted variable bias.
Wrongly Assuming Homoskedasticity

What can happen if you wrongly assume homoskedasticity? You can get misleading results. The homoskedastic standard errors tend to be smaller than the heteroskedastic ones. Your results might appear to be statistically significant, when they are not.

Are the estimates biased? No.

STATA Help: Problem Set 1

- You should include your .do file at the end of the problem set
- To read the help file for any Stata function, use the command “help”. Example: “help scatter”.
- To compute a number use the function “scalar” and to see the value use the function “display”

For example,
scalar p=40+50
display p
90