Course notes

- Good work on good-use of statistics
- Some thoughts about bad use
- Distribute “Faith-Based Fudging” (Kleiman, 2003)
- Final exam
  - Take-home exam with Stata component (similar to problem sets)
  - Distributed Monday 17 May
  - Due Friday 21 May
Geometry of the Multiple Regression Model

- Multiple regressors (a plane) and one outcome (a height)
- Linear regression with one regressor can be visualized as a line through a scatterplot in the $(X, Y)$-plane.
- Linear regression with two regressors can be visualized as a plane through a cloud in the $(X_1, X_2, Y)$-space. The slope of $Y$ against $X_1$ is $\beta_1$, and the slope of $Y$ against $X_2$ is $\beta_2$.

OLS Estimator in Multiple Regression

With one regressor (and an intercept), we choose $b_0$ and $b_1$ to minimize
\[ \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \]

With two or more regressors (and an intercept), we choose $b_0$ and $b_1, \ldots, b_k$ to minimize
\[ \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \cdots - b_k X_{ki})^2 \]

- The OLS estimators that minimize the sum of squared prediction mistakes are $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$. (Complicated formula.)
- Predicted value for observation $i$ given the value of the explanatory variables for observation $i$:
  \[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} \]
- OLS residual: $\hat{u}_i = \hat{Y}_i - Y_i$
Test Scores and the Student-Teacher Ratio

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

The coefficient on STR falls from $-2.28$ to $-1.10$ when we hold constant the effect of percent English learners.

The first estimate suffered from omitted variable bias because it reflected both the effect of a change in the Student-Teacher Ratio and the omitted effect of more English learners.

The Percent English learners tends to lower the district test score (holding constant classroom size).
Using multiple regression results

Expected test score for a class room that has 22 students, of whom 5 percent are English learners?

\[
\text{TestScore} = 686.0 - 1.10 \times \text{STR} - 0.65 \times \text{PctEL}
\]
\[
= 686.0 - 1.10 \times 22 - 0.65 \times 5
\]
\[
= 658.6
\]

Expected test score for a class room that has 24 students, of whom 5 percent are English learners (or, the “same” classroom with two more students).

\[
\text{TestScore} = 686.0 - 1.10 \times \text{STR} - 0.65 \times \text{PctEL}
\]
\[
= 686.0 - 1.10 \times 24 - 0.65 \times 5
\]
\[
= 656.4
\]

(Adding +2 students reduces test scores by \(2 \times (-1.1) = -2.2\) points)

Using multiple regression results

Expected test score for a class room that has 22 students, of whom 10 percent are English learners?

\[
\text{TestScore} = 686.0 - 1.10 \times \text{STR} - 0.65 \times \text{PctEL}
\]
\[
= 686.0 - 1.10 \times 22 - 0.65 \times 10
\]
\[
= 655.3
\]

Expected test score for a class room that has 24 students, of whom 10 percent are English learners (or, the “same” classroom with two more students).

\[
\text{TestScore} = 686.0 - 1.10 \times \text{STR} - 0.65 \times \text{PctEL}
\]
\[
= 686.0 - 1.10 \times 24 - 0.65 \times 10
\]
\[
= 653.1
\]

(Adding +2 students reduces test scores by \(2 \times (-1.1) = -2.2\) points)
Aside: The simplest regressor

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \]
\[ = \beta_0 \cdot 1 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \]
\[ = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \text{ where } X_{0i} = 1 \text{ for all } i \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>testscr</th>
<th>constant</th>
<th>str</th>
<th>el_pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>690.8</td>
<td>1</td>
<td>17.88991</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>661.2</td>
<td>1</td>
<td>21.52466</td>
<td>4.583333</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>643.6</td>
<td>1</td>
<td>18.69723</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>647.7</td>
<td>1</td>
<td>17.35714</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>640.85</td>
<td>1</td>
<td>18.67133</td>
<td>13.85768</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>605.55</td>
<td>1</td>
<td>21.40625</td>
<td>12.40876</td>
</tr>
</tbody>
</table>

Every observation has the same value for “variable” \( X_{0i} \).
A regression with only regressor \( X_{0i} = 1 \) is equivalent to computing \( \bar{Y} \).

The Least Squares Assumptions

The point of the assumptions

- The OLS estimators are unbiased and consistent.
- The OLS estimators are jointly normal in large samples.
- Each OLS estimator is normally distributed, \( \hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}) \),
  \[ j = 1, \ldots, k. \]
\[ E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0 \]

Assumption #1: The Conditional Distribution of \( u_i \) Given \( X_{1i}, X_{2i}, \ldots, X_{ki} \)
Has a Mean of Zero

- \( Y_i \) is sometimes above the regression line and sometimes below the regression line; and
- Knowing all of the explanatory variables \( X_{1i}, X_{2i}, \ldots, X_{ki} \) does not help you guess whether \( Y_i \) is above or below the regression line (the regression line is the best guess).

Assumptions #2 and #3

- Random Sampling, (Quasi-)experimental design
  Assumption #2: \( (X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n \) Are i.i.d.
- Four Moments, No Severe Outliers
  Assumption #3: \( X_{1i}, X_{2i}, \ldots, X_{ki} \) and \( u_i \) Have Four Moments
Assumption #4: No Perfect Multicollinearity

Examples

• Fraction of English Learners
  Problem is $PctEL_i = 100 \times FracEL_i$ for every district.

• “Not very small classes” (Class size $\geq 12$)
  Problem is $NVS_i = 1$ for every observation

• Percentage of native English speakers
  Problem is $PctES_i = 100 - PctEL_i$ for every district.

Assumption #4: No Perfect Multicollinearity

Lessons

• Cannot have two (or more) regressors that contain exactly the same information.

• Stata, and other apps, will catch and “fix” perfect multicollinearity by dropping one or more variables from the regression command, but

• It’s better to control this process yourself.

Imperfect Multicollinearity

• Is the entire point of multiple regression: sorting out the influences of multiple, correlated influences, and

• Does not create the problem of Perfect Multicollinearity
Another multicollinearity example

Example: Are long work hours associated with higher wages?

\[ \text{Hourly wage}_i = \beta_0 + \beta_1 \text{Work hours}_i + \beta_2 \text{All other hours}_i + u_i \]

Problem: Work Hours + All Other Hours = 24 \times 365 for everyone (for all \( i \))

Potential Revisions:

\[ \text{Hourly wage}_i = \beta_0 + \beta_1 \text{Work hours}_i + u_i \]
\[ \text{Hourly wage}_i = \beta_0 + \beta_1 \text{Work hours}_i + \beta_2 \text{Leisure hours}_i + u_i \]

(The latter excludes other time, e.g., social reproduction.)

(Aside: also a good example of under-identification: is the wage high because the person works a lot of hours or does the person work a lot of hours because the wage is high?)

Distribution of OLS Estimators

- \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \) are unbiased (expectation) and consistent (large samples) estimators of \( \beta_0, \beta_1, \beta_2, \ldots, \beta_k \).
- \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \) are distributed multivariate normal (in large samples).
- \( \hat{\sigma}_{\hat{\beta}_j} \), the standard error of \( \hat{\beta}_j \), or \( SE(\hat{\beta}_j) \), or estimated standard deviation of the estimator \( \hat{\beta}_j \), can be estimated by long but straightforward formula. \( \hat{\sigma}_{\hat{\beta}_j} \) expresses how much sample estimates of \( \beta_j \) would vary in repeated sampling.
- Because the estimators are subject to sampling variation, we need to use statistical tests about the population parameters.
Hypothesis Tests and CI’s for a Single Coefficient

- Methods are identical to regression with one regressor.
  1. Form null and alternative hypotheses
  2. Estimate $\hat{\beta}_j$ and compute its standard error, $SE(\hat{\beta}_j)$
  3. Compute the $t$-statistic

\[
t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}
\]

4. Compute the $p$-value

\[
p\text{-value} = 2\Phi(-|t|)
\]

and test the null hypothesis.

- The 95 percent confidence interval for $\beta_j$ is

$$(\hat{\beta}_j - 1.96SE(\hat{\beta}_j), \hat{\beta}_j + 1.96SE(\hat{\beta}_j))$$

Test Scores and the Student-Teacher Ratio

$$\text{TestScore} = 686.0 - 1.10 \text{ STR} - 0.650 \text{ PctEL}$$

$$\text{ (8.7) \quad (0.43) \quad (0.031) }$$

1. $H_0 : \beta_{\text{STR}} = 0$ vs. $H_1 : \beta_{\text{STR}} \neq 0$

\[
t = \frac{-1.10 - 0}{0.43} = -2.54
\]

$p$-value $= 2\Phi(-2.54) = 1.1\%$

2. $H_0 : \beta_{\text{STR}} = -2.28$ vs. $H_1 : \beta_{\text{STR}} \neq -2.28$

\[
t = \frac{-1.10 - (-2.28)}{0.43} = 2.74
\]

$p$-value $= 2\Phi(-2.74) = 0.6\%$
Test Scores and the Student-Teacher Ratio

- 95 percent CI

\[ (-1.10 - 1.96 \times 0.43 \ , \ -1.10 + 1.96 \times 0.43) \]

\[ \beta_{\text{STR}} \in (-1.95 \ , \ -0.26) \]

- 95 percent CI for a two-student reduction

\[ (-1.95 \times 2 \ , \ -0.26 \times 2) \]

\[ (-3.90 \ , \ -0.52) \]

Effect of STR Holding $ Constant

Expn is expenditure per pupil.

\[
\text{TestScore} = 649.6 - 0.29 \ \text{STR} + 3.87 \ \text{Expn} - 0.656 \ \text{PctEL} \\
(15.5) \quad (0.48) \quad (1.59) \quad (0.032)
\]

Nifty use of multiple regression: by holding constant expenditure per pupil, we can learn the effect of changing class size without additional resource use (something else must be reduced to effect a decrease in the STR). N.B. the effect of STR is small and not significant in this specification.
Aside: normalizing variables

(Normalizing has nothing to do with the normal distribution.)

Intensive versus Extensive Measures and Models

1. Why use Percent English Learners instead of Number of English Learners?
2. Why use District Expenditure per Pupil instead of Total District Expenditure?

Joint Hypotheses

Why not simply test one hypothesis at a time?

Two reasons

1. If each test has a 5 percent chance of being wrong (the size that we have chosen by testing $p$-value \( \leq 0.05 \)), the chance of at least one test being wrong is substantially greater than 5 percent even if the tests are independent.

2. Because regressor variables are correlated, the test statistics are correlated and the tests are often not independent.

Next time: Tests of Joint Hypotheses