Linear Regression with Multiple Regressors
(SW Chapter 6)

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  
(SW Section 6.1)

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); and
2. Z is correlated with the regressor X (i.e. corr(Z,X) ≠ 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,

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

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.

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

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

\[
p \rightarrow \frac{\sigma_{Xu}}{\sigma_X^2}
\]

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

where \( \rho_{Xu} = \text{corr}(X,u) \). If assumption #1 is correct, then \( \rho_{Xu} = 0 \), but if not we have....
The omitted variable bias formula:

\[
\hat{\beta}_1 \rightarrow \beta_1 + \left( \frac{\sigma_u}{\sigma_X} \right) \rho_{Xu}
\]

- If an omitted variable \( Z \) is \textit{both}:
  1. a determinant of \( Y \) (that is, it is contained in \( u \)); \textbf{and}
  2. correlated with \( X \),

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

- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. \textit{Is this actually going on in the CA data?}
- 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)

**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

<table>
<thead>
<tr>
<th>Percentage of English learners</th>
<th>Student–Teacher Ratio &lt; 20</th>
<th>Student–Teacher Ratio ≥ 20</th>
<th>Difference in Test Scores, Low vs. High STR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Test Score</td>
<td>n</td>
<td>Average Test Score</td>
</tr>
<tr>
<td>All districts</td>
<td>657.4</td>
<td>238</td>
<td>650.0</td>
</tr>
<tr>
<td>&lt; 1.9%</td>
<td>664.5</td>
<td>76</td>
<td>665.4</td>
</tr>
<tr>
<td>1.9–8.8%</td>
<td>665.2</td>
<td>64</td>
<td>661.8</td>
</tr>
<tr>
<td>8.8–23.0%</td>
<td>654.9</td>
<td>54</td>
<td>649.7</td>
</tr>
<tr>
<td>&gt; 23.0%</td>
<td>636.7</td>
<td>44</td>
<td>634.8</td>
</tr>
</tbody>
</table>
Causality and regression analysis

The test score/STR/fraction English Learners example shows that, if an omitted variable satisfies the two conditions for omitted variable bias, then the OLS estimator in the regression omitting that variable is biased and inconsistent. So, even if \( n \) is large, \( \hat{\beta}_1 \) will not be close to \( \beta_1 \).

This raises a deeper question: how do we define \( \beta_1 \)? That is, what precisely do we want to estimate when we run a regression?
What precisely do we want to estimate when we run a regression?

There are (at least) three possible answers to this question:

1. We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning.

   This can be useful at times, but isn’t very interesting intellectually and isn’t what this course is about.
2. We want to make forecasts, or predictions, of the value of $Y$ for an entity not in the data set, for which we know the value of $X$.

*Forecasting is an important job for economists, and excellent forecasts are possible using regression methods without needing to know causal effects. We will return to forecasting later in the course.*
3. We want to estimate the causal effect on \( Y \) of a change in \( X \).

This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect. Except when we discuss forecasting, the aim of this course is the estimation of causal effects using regression methods.
What, precisely, is a causal effect?

- “Causality” is a complex concept!

- In this course, we take a practical approach to defining causality:

  A causal effect is defined to be the effect 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:

Imagine an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR…

- In that experiment, students would be randomly assigned to classes, which would have different sizes.

- Because they are randomly assigned, all student characteristics (and thus $u_i$) would be distributed independently of $STR_i$.

- Thus, $E(u_i|STR_i) = 0$ – that is, LSA #1 holds in a randomized controlled experiment.
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 \( X \).

- Thus, the “control” and “treatment” groups differ in a systematic way, so \( \text{corr}(STR, PctEL) \neq 0 \)
• 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 $PctEL$.
  
  o 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.
  
  o 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. (This solution to OV bias is rarely feasible.)

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 you 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  
(SW Section 6.2)

Consider the case of two regressors:

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

- \( Y \) is the \textit{dependent variable}
- \( X_1, X_2 \) are the two \textit{independent variables} (regressors)
- \((Y_i, X_{1i}, X_{2i})\) denote the \( i^{\text{th}} \) observation on \( Y, X_1, \) and \( X_2 \).
- \( \beta_0 \) = unknown population intercept
- \( \beta_1 \) = effect on \( Y \) of a change in \( X_1 \), holding \( X_2 \) constant
- \( \beta_2 \) = 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

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

Consider changing \( X_1 \) by \( \Delta X_1 \) while holding \( X_2 \) constant:

Population regression line \textit{before} the change:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]

Population regression line, \textit{after} the change:

\[ Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \]
Before: \[ Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2X_2 \]

After: \[ Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2X_2 \]

Difference: \[ \Delta Y = \beta_1\Delta X_1 \]

So:

\[ \beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{ holding } X_2 \text{ constant} \]

\[ \beta_2 = \frac{\Delta Y}{\Delta X_2}, \text{ holding } X_1 \text{ constant} \]

\[ \beta_0 = \text{predicted value of } Y \text{ when } X_1 = X_2 = 0. \]
The OLS Estimator in Multiple Regression  
(SW Section 6.3)

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 \( \beta_0 \) and \( \beta_1 \).
Example: the California test score data

Regression of $TestScore$ against $STR$:

$$TestScore = 698.9 - 2.28 \times STR$$

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

$$TestScore = 686.0 - 1.10 \times STR - 0.65 PctEL$$

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

Model 2: OLS, using observations 1–420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const 686.032</td>
<td>8.72822</td>
<td>78.5993</td>
<td>0.0000</td>
</tr>
<tr>
<td>str</td>
<td>−1.10130</td>
<td>0.432847</td>
<td>−2.5443</td>
</tr>
<tr>
<td>el_pct</td>
<td>−0.649777</td>
<td>0.0310318</td>
<td>−20.9391</td>
</tr>
</tbody>
</table>

Mean dependent var 654.1565  S.D. dependent var 19.05335
Sum squared resid 87245.29  S.E. of regression 14.46448
$R^2$ 0.426431  Adjusted $R^2$ 0.423680
$F(2, 417)$ 223.8229  P-value($F$) 9.28e–67
Log-likelihood −1716.561  Akaike criterion 3439.123
Schwarz criterion 3451.243  Hannan–Quinn 3443.913

TestScore = 686.0 − 1.10\times STR − 0.65PctEL

More on this printout later...
Measures of Fit for Multiple Regression
(SW Section 6.4)

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

- The \( R^2 \) always increases when you add another regressor (\textit{why?}) – a bit of a problem for a measure of “fit”
$R^2$ and $\overline{R}^2$, ctd.

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

\[
Adjusted \ R^2: \ \overline{R}^2 = 1 - \left( \frac{n - 1}{n - k - 1} \right) \frac{SSR}{TSS}
\]

Note that $\overline{R}^2 < R^2$, however if $n$ is large the two will be very close.
Measures of fit, ctd.

Test score example:

(1) \[ \text{TestScore} = 698.9 - 2.28 \times \text{STR}, \]
\[ R^2 = .05, \ SER = 18.6 \]

(2) \[ \text{TestScore} = 686.0 - 1.10 \times \text{STR} - 0.65 \text{PctEL}, \]
\[ R^2 = .426, \ \overline{R}^2 = .424, \ SER = 14.5 \]

• What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?
• Why are the \( R^2 \) and the \( \overline{R}^2 \) so close in (2)?
The Least Squares Assumptions for Multiple Regression
(SW Section 6.5)

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

1. The conditional distribution of \( u \) given the \( X \)'s has mean zero, that is, \( E(u_i|X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0 \).
2. \((X_{1i}, \ldots, X_{ki}, Y_i), \ i =1, \ldots, n\), are i.i.d.
3. Large outliers are unlikely: \( X_1, \ldots, X_k \), and \( Y \) have four moments: \( E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty \).
4. There is no perfect multicollinearity.
Assumption #1: the conditional mean of \( u \) given the included \( Xs \) is zero.

\[ E(u|X_1 = x_1, \ldots, X_k = x_k) = 0 \]

- 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
  \( (1) \) belongs in the equation (so is in \( u \)) and
  \( (2) \) 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 (discussed in Ch. 7)
Assumption #2: \((X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n,\) 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.

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

- In the previous regression, $\beta_1$ 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 distribution of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.***
The Sampling Distribution of the OLS Estimator
(SW Section 6.6)

Under the four Least Squares Assumptions,

- The sampling distribution of $\hat{\beta}_1$ has mean $\beta_1$
- $\text{var}(\hat{\beta}_1)$ is inversely proportional to $n$.
- 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)
  
  \[ \hat{\beta}_1 - E(\hat{\beta}_1) \]

  - $\frac{\sqrt{\text{var}(\hat{\beta}_1)}}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $\mathcal{N}(0,1)$ (CLT)

  - These statements hold for $\hat{\beta}_1, \ldots, \hat{\beta}_k$

  Conceptually, there is nothing new here!
Multicollinearity, Perfect and Imperfect
(SW Section 6.7)

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

Some more examples of perfect multicollinearity

1. The example from before: you include \( STR \) twice,
2. 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.
3. Would there be perfect multicollinearity if the intercept (constant) were excluded from 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 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.
Imperfect multicollinearity, ctd.

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

- The idea: 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 don’t contain much information about what happens when $X_1$ changes but $X_2$ doesn’t. If 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…