“Something’s just not right—our air is clean, our water is pure, we all get plenty of exercise, everything we eat is organic and free-range, and yet nobody lives past thirty.”
Outline of Lecture 3

- Omitted Variables Bias and Multiple Regression
- Sampling Distribution of OLS Estimator in Multiple Regression
- Homoskedasticity vs. Heteroskedasticity
- Hypothesis Tests (covered in supplementary notes & homework)
Omitted Variables Bias

- Consider the simple model with two regressors:

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

- Suppose that the variable \( X_{2i} \) is omitted from the regression (either because of model specification error, or maybe because you don't have data on the variable \( X_{2i} \))

- Then the regression model becomes:

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

- Q: LSA #1 is now \( E[v_i|X_{1i}] = 0 \). Is it satisfied here?
Key Result:

- Let $\text{Corr}(X_{1i}, X_{2i}) = \rho_{12} \neq 0$

  (Note: LSA #1 not satisfied, i.e., $\text{Corr}(X_{1i}, v_i) \neq 0$ even if $\text{Corr}(X_{1i}, u_i) = 0$)

- Then, we can prove that the OLS estimator has the following probability limit:

  $\hat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \rho_{12} \frac{\sigma_{X2}}{\sigma_{X1}}$

- This says that as the sample size increases, $\hat{\beta}_1$ does not get close to the true $\beta_1$ with high probability

Olivier Deschenes, UCSB, ESM 296, Winter 2018
Implications:

- If the regressor $X_{1i}$ is correlated with a variable that:
  
  (i) is also a predictor of the dependent variable $Y_i$,

  (ii) has been omitted from the model, then the OLS estimator will suffer from omitted variable bias (i.e. OLS is not consistent)

- In the house value example, omitted variable bias will arise if NOx concentrations are correlated with other predictors of house values (for example: house size, noise levels, other pollutants, etc) and if these factors are not controlled for in the regression
Conclusion on Omitted Variables Bias

• Omitted variable bias is a problem whether the sample size is small or large. Even in the limit experiment when \( n \rightarrow \infty \), the OLS estimator remains inconsistent.

• Whether this “bias” is large or small depends on:
  
  (i) the magnitude of the correlation between \( X_{1i} \) and the omitted variable (\( X_{2i} \) in the example). The larger \( |\rho_{12}| \), the larger is the bias.
  
  (ii) the magnitude of the regression coefficient on the omitted variable (\( \beta_2 \) in the example).

• The direction of the bias depends on the sign of \( \rho_{12} \) and \( \beta_2 \). If \( \rho_{12} > 0 \) and \( \beta_2 > 0 \), then the OLS estimator overstates \( \beta_1 \).
Solutions to Problem of Omitted Variables Bias:

1. Add more variables to the regression model

- Effectively, this improves the credibility of the assumption $E[u_i|X_i]=0$ (LSA#1)

- Why: the more variables you include, the more potential relevant predictors of $Y_i$ you include

- However: there is a bias/variance tradeoff in finite samples (including more regressors reduces the risk of OV bias but it also reduces the precision of OLS estimator)
  - Moreover, some important factors may be unobservable so it impossible to directly include controls for them

2. Later: Matching, Instrumental variables regression, Panel data models, and also controlled random experiments
The Population Multiple Regression Model

- Consider the case of two regressors:
  \[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad i = 1, \ldots, n \]

- \( \beta_0 \) = unknown population intercept

- \( \beta_1 \) = effect of a change in \( X_1 \) on \( Y \), holding \( X_2 \) constant

- \( \beta_2 \) = effect of a change in \( X_2 \) on \( Y \), 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 function \textit{before} the change in \( X_1 \):
  \[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]

- Population regression function, \textit{after} the change in \( X_1 \):
  \[ Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \]

- Difference: \( \Delta Y = \beta_1 \Delta X_1 \)
**Implications:**

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

**Similarly,**

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

**\( \beta_0 \) = \text{predicted value of } Y \text{ when } X_1 = X_2 = 0**

- Rarely a useful parameter
OLS Estimator in Multivariate Regression

- Recall the formula from bivariate regression
  \( Y_i = \beta_0 + \beta_1 X_{1i} + u_i: \)
  \[
  \hat{\beta}_1 = \frac{S_{X_1Y}}{S^2_{X_1}}
  \]

- Equivalent formula in multivariate setting
  \( Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i: \)
  \[
  \hat{\beta}_1 = \frac{S_{\tilde{X}_1Y}}{S^2_{\tilde{X}_1}}
  \]

- Where \( \tilde{X}_{1i} \) is the fitted residual from a regression of \( X_{1i} \) on a constant term and all the other regressors (here only \( X_{2i} ) \)

Olivier Deschenes, UCSB, ESM 296, Winter 2018
Multiple Regression in STATA

`regress price nox rooms, robust;`

Linear regression

|                | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|----------------|--------|-----------|-------|-------|----------------------|
| price          |        |           |       |       |                      |
| nox            | -1062.208 | 357.8614  | -2.97 | 0.003 | -1767.811 to -356.6063 |
| rooms          | 9836.748  | 924.3718  | 10.64 | 0.000 | 8014.146 to 11659.35  |
| _cons          | -33216.07 | 6655.565 | -4.99 | 0.000 | -46338.97 to -20093.17 |

\[\hat{\text{Price}} = -33216 - 1062 \times NOX + 9837 \times ROOMS\]
**Interpretation:**

- Recall the regression of *Price* on *NOx* (Lecture 2):
  \[
  \hat{\text{Price}} = 38068 - 2776 \times NOX
  \]

- Now include number of rooms (*Rooms*) as well:
  \[
  \hat{\text{Price}} = -33216 - 1062 \times NOX + 9837 \times ROOMS
  \]

- What happens to the coefficient on *NOx*?
- Why? (*Note*: \(\text{corr}(NOx, Rooms) = -0.29\))

- \(\Rightarrow\) In the model that omits *Rooms*, the regression attributes some of the effect of *Rooms* to *NOx*
Application of the Partial OLS Estimator Formula

- Recall \( \hat{\beta}_1 = \frac{S_{X_1} Y}{S_{X_1}^2} \)

- Step 1: Regress \( \text{NOx} \) on \( \text{Rooms} \), get fitted residuals:

\[
\text{. regress nox rooms, robust;}
\]

|                | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|----------------|--------|-----------|-------|-------|---------------------|
| nox            |        |           |       |       |                     |
| rooms          | -.4806996 | .1127166 | -4.26 | 0.000 | -.7029385 -.2584607 |
| _cons          | 8.549037  | .715984   | 11.94 | 0.000 | 7.137359 9.960715  |

\[
\text{. predict nox_resid, residual;}
\]
Recall $\hat{\beta}_1 = \frac{S_{\tilde{X}_1 Y}}{S_{\tilde{X}_1}^2}$

Step 2: Compute $\hat{\beta}_1$

```plaintext
.summ price nox_resid;

Variable | Obs  | Mean     | Std. Dev. | Min  | Max  |
---------|------|----------|-----------|------|------|
price    | 206  | 22723.11 | 9381.108  | 5000 | 50001|
nox_resid| 206  | -9.20e-10| 1.095036  | -1.802523 | 3.107652 |

.correlate price nox_resid, cov;
(obs=206)
```

\[
\hat{\beta}_1 = \frac{-1273.7}{1.095^2} = -1062
\]

Olivier Deschenes, UCSB, ESM 296, Winter 2018
**The Least Squares Assumptions in the Multiple Regression Model**

<table>
<thead>
<tr>
<th>6.4</th>
</tr>
</thead>
</table>

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

1. \( u_i \) has conditional mean zero given \( X_{1i}, X_{2i}, \ldots, X_{ki} \); that is,
   \[ E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0. \]

2. \( (X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n \) are independently and identically distributed (i.i.d.) draws from their joint distribution.

3. Large outliers are unlikely: \( X_{1i}, \ldots, X_{ki} \) and \( Y_i \) have nonzero finite fourth moments.

4. There is no perfect multicollinearity.

**LSA#1 is key:** An implication is that each regressor (X) is uncorrelated with the regression error \( u_i \)

**LSA#1 implies that there is no omitted variable in the model**
Discussion of the LSA’s for Multivariate Model

- LSA1: $E[u_i|X_{1i}, X_{2i}, ..., X_{Ki}] = 0$

- Key assumption: implies that the OLS estimator is consistent (i.e. no omitted variables bias)

- Remember that it is not testable without more information

- LSA2 and LSA3: technical assumptions, always maintained in this class

- LSA4: The regressors are perfectly multi-collinear if one of the regressors is a perfect linear function of another
  - We rule this out
Discussion of Perfect Multi-Collinearity

- LSA4: The regressors are perfectly multi-collinear if one of the regressors is a perfect linear function of some of the others.

- Example: \( X_{1i} = (=1 \text{ if observation } i \text{ is male}) \)
  \( X_{2i} = (=1 \text{ if observation } i \text{ is female}) \)
  So: \( X_{1i} + X_{2i} = 1 \), perfectly collinear with intercept.

- LSA4 is “testable”. If two (or more) regressors are perfectly collinear, Stata will throw one out of the regression model.

- It simply means that you cannot separately identify the effect of the multi-collinear regressors on \( Y \).
**Example**: Suppose you accidentally include $NOX$ twice in the regression:

```
regress price nox nox, robust
note: nox omitted because of collinearity
```

| Coef. | Std. Err. | t    | P>|t|   | [95% Conf. Interval] |
|-------|-----------|------|-------|-----------------------|
| nox   | -2775.674 | 414.4046 | -6.70  | 0.000     | -3592.739   | -1958.608 |
| nox   | (omitted) |       |       |           |           |            |
| _cons | 38068.27  | 2222.545 | 17.13  | 0.000     | 33686.17   | 42450.38  |

Linear regression

Number of obs = 206
F(  1,   204) = 44.86
Prob > F      = 0.0000
R-squared     = 0.1146
Root MSE      = 8849
"Imperfect" Multi-Collinearity

- Two variables that are highly correlated with each other, although not perfectly (i.e. correlation coefficient close to 1 or -1)

- The more multi-collinear $X_1$ and $X_2$ are, the more "unstable" the OLS estimates of $\beta_1$ and $\beta_2$ become, and also the larger their standard errors become

- Detectable by examining data and regression results
### Key Concept

**Large Sample Distribution of \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \)**

If the least squares assumptions (Key Concept 6.4) hold, then in large samples the OLS estimators \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) are jointly normally distributed and each \( \hat{\beta}_j \) is distributed \( N(\beta_j, \sigma^2_{\beta_j}), j = 0, \ldots, k \).

---

**Same results as in the bivariate regression model**

**OLS estimator is distributed with a Normal distribution (when n is large) due to the Central Limit Theorem (CLT)**

**Implication 1.** Can use Normal distribution for hypothesis tests

**Implication 2.** Formula for covariance matrix of OLS estimator depends on assumption of **homoskedasticity or heteroskedasticity** of the regression errors

*** Here we always proceed with the assumption of heteroskedasticity

Olivier Deschenes, UCSB, ESM 296, Winter 2018
Heteroskedasticity and Homoskedasticity

- What do these two terms mean?

- If \( \text{Var}(u_i | X_i = x) \) is constant – that is, if the variance of the conditional distribution of \( u_i \) given \( X_i \) does not depend on \( X_i \) – then \( u_i \) is said to be \textit{homoskedastic}

- Otherwise, \( u_i \) is \textit{heteroskedastic}

- Since it involves the unobserved regression error term, it is difficult to directly assess heteroskedasticity by looking at the data, especially in multivariate models

- *** We assume heteroskedastic errors and adjust our methods of inference to account for it in a general way

Olivier Deschenes, UCSB, ESM 296, Winter 2018
Implications of Homoskedasticity:

- Homoskedasticity of the error term $\text{Var}(u_i | X_i) = \sigma^2$ implies that the **conditional variance of $Y$ given $X$ is also constant**: 

- Consider simple bivariate model $Y_i = \beta_0 + \beta_1 X_i + u_i$

\[
\text{Var}(Y_i | X_i) = \text{Var}(\beta_0 | X_i) + \text{Var}(\beta_1 X_i | X_i) + \text{Var}(u_i | X_i) \\
= \beta_0 \text{Var}(1 | X_i) + \beta_1 \text{Var}(X_i | X_i) + \sigma^2 \\
= 0 + 0 + \sigma^2
\]

- [Note that all covariance terms are equal to 0 (by LSA#1)]
Looking at data scatter plot to assess homoskedasticity

Scatter plot and regression line for hourly wages vs. years of education (data source: Current Population Survey)
Sampling Variance of OLS Estimator Without Homoskedasticity in Bivariate Model

- Recall the earlier result

- When the sample size $n$ grows large, under assumptions LSA#1, LSA#2, and LSA#3, and without assuming homoskedasticity you can prove that:

$$
\hat{\beta}_1 \overset{A}{\approx} N \left( \beta_1, \frac{\text{Var}[(X_i - \mu_X)u_i]}{n \text{Var}(X_i)^2} \right)
$$

The standard errors reported by STATA under the “regress y x, robust” command is an estimate of the square root of the sampling variance of the OLS estimator.
Sampling Variance of OLS Estimator in Multivariate Regression

- The same logic applies here, but the formulas for the variance of the sampling distribution is more complicated (come to office hours if you want to know...)

- ** The OLS estimator has an approximately normal sampling distribution:

\[
\hat{\beta}_j \approx N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)
\]

- You should assume (at least in ESM 296) that \( \sigma^2_{\hat{\beta}_j} \) is derived under heteroskedasticity.
Conclusion on Heteroskedasticity:

1. Whether the errors are homoskedastic or heteroskedastic does not change how we estimate the slope coefficients in all of our regression models.

2. The sampling covariance matrix (i.e. $\text{Var}(\hat{\beta})$) derived under the assumption of heteroskedasticity simplifies (when $n$ is large) to the theoretically-correct covariance matrix in the special case of homoskedasticity.

So, we always use heteroskedasticity-robust standard errors and inference.
. regress price rooms;

(...)

|            | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|--------|-----------|-------|-----|----------------------|
| rooms      | 10347.35 | 620.3613  | 16.68 | 0.000 | 9124.209 - 11570.49  |
| _cons      | -42296.93 | 3921.414  | -10.79 | 0.000 | -50028.63 - 34565.23 |

. regress price rooms, robust;

Linear regression

Number of obs = 206
F(  1,   204) = 148.50
Prob > F      = 0.0000
R-squared     = 0.5769
Root MSE      = 6116.7

|            | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|--------|-----------|-------|-----|----------------------|
| rooms      | 10347.35 | 849.1011  | 12.19 | 0.000 | 8673.211 - 12021.49  |
| _cons      | -42296.93 | 5426.635 | -7.79 | 0.000 | -52996.41 - 31597.44 |