Logistics

- **Exam Date:** *Wednesday, April 8 from 3-6pm in Sever Hall 113*
  - The exam covers material from 1010a, 1010b, and 1123 (not 1011a, 1011b, or 1126).
  - No calculators needed or permitted.

- **Econometrics Office Hours in Littauer 219:**
  - Thursday 4/2: 4 - 6pm
  - Friday 4/3: 3 - 5pm
  - Monday 4/6: 9:30 - 11:30am

- [http://economics.harvard.edu/pages/honors](http://economics.harvard.edu/pages/honors) has previous exams, review session videos, and slides.
<table>
<thead>
<tr>
<th>Section</th>
<th>Subsections</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ordinary Least Squares (OLS)</td>
<td>- The Error Term</td>
</tr>
<tr>
<td></td>
<td>- Heteroskedastic errors</td>
</tr>
<tr>
<td></td>
<td>- Hypothesis Testing</td>
</tr>
<tr>
<td></td>
<td>- Polynomials</td>
</tr>
<tr>
<td></td>
<td>- Logs</td>
</tr>
<tr>
<td></td>
<td>- Interaction Terms</td>
</tr>
<tr>
<td></td>
<td>- Conditional Mean Independence</td>
</tr>
<tr>
<td></td>
<td>- Omitted Variable Bias</td>
</tr>
<tr>
<td></td>
<td>- Internal &amp; External Validity</td>
</tr>
<tr>
<td>2. Binary Dependent Variables</td>
<td>- Problems with OLS</td>
</tr>
<tr>
<td></td>
<td>- Probit &amp; Logit</td>
</tr>
<tr>
<td>3. Panel Data</td>
<td></td>
</tr>
<tr>
<td>4. Instrumental Variables</td>
<td>- Advantages</td>
</tr>
<tr>
<td></td>
<td>- Fixed Effects</td>
</tr>
<tr>
<td></td>
<td>- Autocorrelation</td>
</tr>
<tr>
<td>5. Forecasting</td>
<td>- Testing Stationarity</td>
</tr>
<tr>
<td></td>
<td>- Models</td>
</tr>
<tr>
<td></td>
<td>- Testing Stationarity</td>
</tr>
</tbody>
</table>
Section 1

Ordinary Least Squares (OLS)
Ordinary Least Squares

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

**Assumptions:**
- Observations \((X_{1i}, X_{2i}, Y_i)\) are independent and identically distributed (\textit{iid})
- No perfect multicollinearity of \(X\)s
- Linear form is correctly-specified
- Conditional Mean Independence of \(X\)

If these assumptions hold, then OLS estimates \textit{unbiased}, \textit{consistent}, and asymptotically-normal coefficients

\[
\text{regress y x1 x2, robust}
\]
Perfect Multicollinearity

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \]

Regressors are **perfectly multicollinear** if one can be expressed as a linear function of the others:

**Ex:** if for all \( i \), any of these are true

- \( X_{1i} = X_{2i} \)
- \( X_{3i} = X_{2i} - 2X_{1i} \)
- \( \ldots \)

This is especially common if we include an intercept & fail to omit a dummy term (**dummy variable trap**)
The Error Term $u_i$

The error term $u_i$ is **unobserved** and typically the culprit behind our econometric woes.

The error term contains all the stuff that affects $Y_i$ but isn’t explicitly in the regression.

$$WAGE_i = \beta_0 + \beta_1 \text{EDUC}_i + u_i$$

In this case, $u_i$ includes the effect of:

- age
- past work experience
- health
- ... (all things that affect wage other than education)
The Error Term $u_i$

Suppose we control for age explicitly:

$$WAGE_i = \beta_0 + \beta_1 \text{EDUC}_i + \beta_2 \text{AGE}_i + u_i$$

Then $u_i$ includes the effect of...

- age
- past work experience
- health
- ... (all things that affect wage other than training & educ)
Homoskedastic v. Heteroskedastic Errors

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

Homoskedastic or Heteroskedastic is an assumption about the pattern of errors \( u_i \)

- **Homoskedastic**: \( \text{Var}(u|X) \) is constant for all \( X \)
- **Heteroskedastic**: \( \text{Var}(u|X) \) can vary with \( X \)

Homoskedasticity is a strong assumption that we basically never have enough evidence to make
Homoskedastic Errors: $\text{Var}(u|X)$ constant for all $X$

Heteroskedastic Errors: $\text{Var}(u|X)$ may change with $X$
Homoskedastic v. Heteroskedastic Errors

- The problem is that error $u$ is always unobserved

- Fortunately, if we allow for heteroskedasticity, standard error estimates will be right, even if the errors are homoskedastic

- **NEVER** assume homoskedasticity

- In STATA, use the “robust” command

  ```
  regress y x1 x2, robust
  ```
Hypothesis Testing – Single Variable

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \]

Suppose we wanted to test if \( \beta_1 \) is statistically different from a constant \( C \), where \( C \) is usually 0:

**Null** Hypothesis \( H_0 : \beta_1 = C \)

**Alternative** Hypothesis \( H_a : \beta_1 \neq C \)

We calculate a \( t \)-statistic using our estimate \( \hat{\beta}_1 \) and its standard error:

\[ t = \frac{\hat{\beta}_1 - C}{\text{se}(\hat{\beta}_1)} \]
Single Hypothesis Testing – One Variable

\[ t = \frac{\hat{\beta}_1 - C}{\text{se}(\hat{\beta}_1)} \]

For a 95% two-sided confidence test, we reject the null hypothesis when \( t \geq 1.96 \) or \( t \leq -1.96 \):

Make sure you also understand how to construct 95% confidence intervals!
Joint Hypothesis Testing – Multiple Variables

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \]

Suppose we wanted to test **multiple** coefficients are different from 0

\[ H_0 : \quad \beta_1 = \beta_2 = \beta_3 = 0 \]

\[ H_a : \quad \text{At least one of } \beta_1, \beta_2, \text{ or } \beta_3 \text{ is nonzero} \]

Now we have to use a **F-test**, which is like a multiple **t-test** that takes into account the correlation between \( \hat{\beta}_1, \hat{\beta}_2, \text{ and } \hat{\beta}_3 \)

**Note:** If we reject \( H_0 \), we cannot say which coefficient(s) is/are non-zero, only that at least one is non-zero
Single Hypothesis Testing

\[ H_0 : \beta_1 = 0 \]
\[ H_a : \beta_1 \neq 0 \]

Use a t-test

Joint Hypothesis Testing

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \]
\[ H_a : \textbf{At least} \text{ one of } \beta_1, \beta_2, \text{ or } \beta_3 \text{ is nonzero} \]

Use a F-test
Testing a Linear Combination of Coefficients

Suppose we wanted to test $\beta_1 = \beta_2$:

\[ H_0 : \beta_1 - \beta_2 = 0 \]
\[ H_a : \beta_1 - \beta_2 \neq 0 \]

Our t-test statistic is $t = \hat{\beta}_1 - \hat{\beta}_2$ and variance is:

\[ \text{Var}(\beta_1 - \beta_2) = \text{Var}(\beta_1) + \text{Var}(\beta_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \]

From the variance formula:

\[ \text{Var}(A \pm B) = \text{Var}(A) + \text{Var}(B) \pm 2\text{Cov}(A, B) \]
Polynomial regressions

**Quartic**

Figure: \( Y = 4 + 6X - X^2 \)

**Cubic**

Figure: \( Y = 12 + 4X - 6X^2 + X^3 \)
Examples of Regressions with Polynomials

Regressing with polynomials useful whenever $Y$ and $X$ do not have a linear relationship.

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i \]

- Diminishing marginal returns ($\beta_2 < 0$)
  - Kitchen output $\sim$ number of chefs
  - Total cost $\sim$ quantity

- Increasing marginal returns ($\beta_2 > 0$)
  - Cell-phone carrier demand $\sim$ number of antennas
  - Most natural monopolies
Testing a Regression with Polynomials

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \]

Suppose we wanted to conduct the following hypothesis test:

\[ H_0 : \ Y \text{ has a linear relationship with } X \]
\[ H_a : \ Y \text{ is non-linear with } X \]

Mathematically:

\[ H_0 : \ \beta_2 = \beta_3 = 0 \]
\[ H_a : \ \text{Either } \beta_2 \neq 0, \beta_3 \neq 0, \text{ or both} \]

Testing **multiple** coefficients, so have to use an **F-test**
Interpreting Coefficients

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i \]

What is the average effect of changing \(X\) from \(X = x\) to \(X = x + \Delta x\), holding all else fixed?

**BEFORE:** \(Y_{\text{before}} = \beta_0 + \beta_1 x + \beta_2 x^2 + u\)

**AFTER:** \(Y_{\text{after}} = \beta_0 + \beta_1 (x + \Delta x) + \beta_2 (x + \Delta x)^2 + u\)

On average, the effect of \(\Delta x\) is:

\[
E[Y_{\text{after}} - Y_{\text{before}}] = \beta_1 \Delta x + \beta_2 [(x + \Delta x)^2 - x^2]
\]

Notice that the effect of changing \(\Delta x\) depends on the original \(x\)!
log-Regressions

Cases of Regressions with $\ln$:

- **Linear-Log**
  \[
  Y = \beta_0 + \beta_1 \ln X + u
  \]
  A one percent increase in $X$ is associated with a $(0.01 \times \beta_1)$ increase in $Y$

- **Log-Linear**
  \[
  \ln Y = \beta_0 + \beta_1 X + u
  \]
  A one unit increase in $X$ is associated with a $(100 \times \beta_1)$ percent change in $Y$

- **Log-Log**
  \[
  \ln Y = \beta_0 + \beta_1 \ln X + u
  \]
  A one percent increase in $X$ is associated with a $\beta$ percent change in $Y$

Ex: price elasticities of demand and supply
Interaction Terms

Interaction terms are the product of two or more variables.

Ex:

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \]

Interaction terms allow for heterogeneous treatment effects by group.
Interaction Terms – Binary Case

Testing for wage discrimination?

\[ WAGE_i = \beta_0 + \beta_1 F_i + \beta_2 B_i + \beta_3 (F_i \times B_i) + u_i \]

- \( F_i = 1 \) if female; 0 otherwise
- \( B_i = 1 \) if black; 0 otherwise

How would we test for any wage discrimination?

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \]
\[ H_a : \text{At least one of } \beta_1, \beta_2, \beta_3 \text{ is non-zero} \]
Interaction Terms – Hybrid Case

\[ \text{WAGE}_i = \beta_0 + \beta_1 EDU_i + \beta_2 F_i + \beta_3 (EDU_i \times F_i) + u_i \]

The average wage when \( EDU = 0 \) is...

- \( \beta_0 \) for males
- \( \beta_0 + \beta_2 \) for females

Interaction terms allow for different intercepts across groups!

The effect of one additional year of education is...

- \( \beta_1 \) for males
- \( \beta_1 + \beta_3 \) for females

Interaction terms allow for different slopes across groups!
Combining It All

Polynomials, logs, interactions and control variables:

\[
\ln Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 D_i + \beta_4 (X_i \times D_i) + \cdots \\
\beta_5 (X_i^2 \times D_i) + \beta_6 W_{1i} + \beta_1 W_{2i} + u_i
\]
Interpreting Coefficients

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i \]

“All else equal, a unit increase in \( X \) is \textit{associated} with a \( \beta \) change in \( Y \) on average”

But economists care about \textit{causality}

When can we claim causality?

“All else equal, a unit increase in \( X \) \textit{causes} a \( \beta \) change in \( Y \) on average”

Causality in OLS requires \textit{conditional mean independence} of \( X \)
Conditional Mean Independence

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i \]

Conditional Mean Independence

Conditional Mean Independence of \( X \) requires:

\[
\mathbb{E}[u|X, W_1, W_2] = \mathbb{E}[u|W_1, W_2]
\]

Intuition:

- CMI implies that those with high \( X \) are not different from those with low \( X \)
- CMI implies \( X \) is **as-if randomly assigned**. This analogs an actual randomized experiment, so we can make statements about causality
Endogeneity and Exogeneity

Regressors that satisfy \textit{conditional mean independence} are called \textbf{exogeneous}

$\Rightarrow$ Exogenous $X$s are \textbf{as-if} randomly assigned

Regressors that violate \textit{conditional mean independence} are called \textbf{endogenous}

$\Rightarrow$ OLS with endogenous regressors yields biased coefficients that \textbf{cannot} be interpreted causally
Omitted Variable Bias

One of the most common violations of CMI is **omitted variable bias**

Omitted Variable Bias occurs when we fail to control for a variable in our regression

Suppose we ran:

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Instead of:

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \]
Omitted Variable Bias (OVB)

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Conditions for OVB?

Omitted Variable Bias

OVB arises if a variable \( W \) is omitted from the regression and

1. \( W \) is a determinant of \( Y \)
2. \( W \) is correlated with \( X \)
Example: Omitted Variable Bias

Suppose we were investigating the effect of **Student-Teacher Ratio** ($X_1$) on $Y$ = average district test scores

We ran the following 5 regressions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student–teacher ratio ($X_1$)</td>
<td>$-2.28^{**}$</td>
<td>$-1.10^*$</td>
<td>$-1.00^{**}$</td>
<td>$-1.31^{**}$</td>
<td>$-1.01^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.52)</td>
<td>(0.43)</td>
<td>(0.27)</td>
<td>(0.34)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Percent English learners ($X_2$)</td>
<td>$-0.650^{**}$</td>
<td>$-0.122^{**}$</td>
<td>$-0.488^{**}$</td>
<td>$-0.130^{**}$</td>
<td>$-0.130^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.033)</td>
<td>(0.030)</td>
<td>(0.036)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>Percent eligible for subsidized lunch ($X_3$)</td>
<td></td>
<td></td>
<td>$-0.547^{**}$</td>
<td></td>
<td>$-0.529^{**}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.024)</td>
<td></td>
<td>(0.038)</td>
</tr>
<tr>
<td>Percent on public income assistance ($X_4$)</td>
<td></td>
<td></td>
<td></td>
<td>$-0.790^{**}$</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.068)</td>
<td>(0.059)</td>
</tr>
<tr>
<td>Intercept</td>
<td>698.9**</td>
<td>686.0**</td>
<td>700.2**</td>
<td>698.0**</td>
<td>700.4**</td>
</tr>
<tr>
<td></td>
<td>(10.4)</td>
<td>(8.7)</td>
<td>(5.6)</td>
<td>(6.9)</td>
<td>(5.5)</td>
</tr>
</tbody>
</table>
Example: Omitted Variable Bias

Column (2) suggests that “Percent English learners ($X_2$)” causes omitted variable bias in Regression (1)

<table>
<thead>
<tr>
<th>Regressor</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student–teacher ratio ($X_1$)</td>
<td>$-2.28^{**}$</td>
<td>$-1.10^*$</td>
<td>$-1.00^{**}$</td>
<td>$-1.31^{**}$</td>
<td>$-1.01^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.52)</td>
<td>(0.43)</td>
<td>(0.27)</td>
<td>(0.34)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Percent English learners ($X_2$)</td>
<td>$-0.650^{**}$</td>
<td>$-0.122^{**}$</td>
<td>$-0.488^{**}$</td>
<td>$-0.130^{**}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.033)</td>
<td>(0.030)</td>
<td>(0.036)</td>
<td></td>
</tr>
<tr>
<td>Percent eligible for subsidized lunch ($X_3$)</td>
<td>$-0.547^{**}$</td>
<td></td>
<td></td>
<td>$-0.529^{**}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td></td>
<td></td>
<td>(0.038)</td>
<td></td>
</tr>
<tr>
<td>Percent on public income assistance ($X_4$)</td>
<td></td>
<td></td>
<td>$-0.790^{**}$</td>
<td></td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.068)</td>
<td></td>
<td>(0.059)</td>
</tr>
<tr>
<td>Intercept</td>
<td>698.9$^{**}$</td>
<td>686.0$^{**}$</td>
<td>700.2$^{**}$</td>
<td>698.0$^{**}$</td>
<td>700.4$^{**}$</td>
</tr>
<tr>
<td></td>
<td>(10.4)</td>
<td>(8.7)</td>
<td>(5.6)</td>
<td>(6.9)</td>
<td>(5.5)</td>
</tr>
</tbody>
</table>

Therefore, % English as a Second Language learners must affect average test scores and be correlated with Student-teacher ratio
Omitted Variable Bias

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

The direction of the bias depends on the direction of the two OVB conditions, i.e. how \( W \), \( X \), and \( Y \) are correlated.

<table>
<thead>
<tr>
<th>Corr(( W, Y )) &gt; 0</th>
<th>Corr(( W, Y )) &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corr(( W, X )) &gt; 0</td>
<td>+</td>
</tr>
<tr>
<td>Corr(( W, X )) &lt; 0</td>
<td>-</td>
</tr>
</tbody>
</table>

When the two correlations go in the same correlation, the bias is positive. When opposite, the bias is negative.
Omitted Variable Bias – Example 1

How does the student-teacher ratio affect test scores?

\[
\text{Avg Test Scores}_i = \beta_0 + \beta_1 \left( \frac{\# \text{ Students}}{\# \text{ Teachers}} \right)_i + u_i
\]

Omitted Variable Bias?

\( W = \% \) English as a Second Language learners

\[
\text{Corr}(W, \frac{\# \text{ Students}}{\# \text{ Teachers}}) > 0
\]

\[
\text{Corr}(W, \text{Avg Test Scores}) < 0
\]

Hence, there is **negative** OVB if we neglect to control for \% ESL.
Omitted Variable Bias - Example 1

Avg Test Scores\(_i\) = \(\beta_0 + \beta_1\left(\frac{\# \text{ Students}}{\# \text{ Teachers}}\right)\_i + u_i\)

Let \(W = \% \text{ English as a Second Language learners}\). Notice that \(W_i\) is in \(u_i\):

\[u_i = \gamma W_i + e_i\]

\[\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right) \rho X u\]

\[\rho X u = \text{corr}(X, u) = \text{corr}(X, \gamma W + e)\]

\[= \text{corr}(X, \gamma W) + \text{corr}(X, e)\]

\[= \gamma \text{corr}(X, W) + 0\]

\[\rho X u > 0\]

since \(\gamma < 0\) and \(\text{corr}(X, W) > 0\), and SEs are always positive.

Hence, \(\hat{\beta}_1 < \beta_1 \iff \hat{\beta}_1\) is negatively biased.
Omitted Variable Bias – Example 2

Did stimulus spending reduce local unemployment during the Great Recession?

\[
\begin{pmatrix}
\text{DISTRICT UNEMPLOYMENT}_i \\
\text{LOCAL STIMULUS SPENDING}_i
\end{pmatrix}
= \beta_0 + \beta_1 \times \begin{pmatrix}
\text{LOCAL STIMULUS SPENDING}_i \\
\text{LOCAL STIMULUS SPENDING}_i + u_i
\end{pmatrix}
\]

Omitted Variable Bias? \( W = \) previous unemployment in the area

\[
\text{Corr}(W, \text{District Unemployment}) > 0
\]
\[
\text{Corr}(W, \text{Stimulus Spending}) > 0
\]

Hence, there is positive OVB if we fail to control for initial unemployment
Deriving the OVB Bias Formula

Suppose we *naively* believe the model to be

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

But we’ve omitted \( W \), so the true model is:

\[ Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + e_i \]

\[ \hat{\beta}_1 = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \]

\[ = \frac{\text{Cov}(X_i, \beta_0 + \beta_1 X_i + \gamma W_i + e_i)}{\text{Var}(X_i)} \]

\[ = 0 + \frac{\text{Cov}(X_i, \beta_1 X_i) + \text{Cov}(X_i, \gamma W_i) + 0}{\text{Var}(X_i)} \]

\[ = \beta_1 + \gamma \frac{\text{Cov}(X_i, W_i)}{\text{Var}(X_i)} \]

Rearranging by \( \text{Corr}(X, u) = \frac{\text{Cov}(X, u)}{\sigma_X \sigma_u} \) yields the OVB equation.
Fixing OVB

Fixing some causes of OVB is straightforward – we just control for variables by including them in our regression.

However, usually implausible to control for all potential omitted variables.

Other common strategies for mitigating OVB include:

- Fixed effects and panel data
- Instrument Variable (IV) regression
Internal Validity

**Internal validity** is a measure of how well our estimates capture what we intended to study.

- Suppose we wanted to study the impact of student-teacher ratio on education outcomes?
  - **Internal Validity**: Do we have an unbiased estimate of the true causal effect?
Threats to Internal Validity

1. Omitted variable bias
2. Wrong functional form
   - Are we assuming linearity when really the relationship between $Y$ and $X$ is nonlinear?
3. Errors-in-variables bias
   - Measurement errors in $X$ biases $\hat{\beta}$ toward 0
4. Sample selection bias
   - Is the sample representative of the population?
5. Simultaneous causality bias
   - Is $Y$ “causing” $X$?
6. “Wrong” standard errors
   - Homoskedasticity v. heteroskedasticity
   - Are our observations iid or autocorrelated?
Assessing Internal Validity of a Regression

When assessing the internal validity of a regression:

- No real-world study is 100% internally valid
  - Do not write “Yes, it is internally valid.”
- Write intelligently
  - What two conditions would potential omitted variables have to satisfy?
  - Why might there be measurement error?
  - ...
- Lastly, assess whether you think these threats to internal validity are large or small
  - i.e. Is your $\hat{\beta}$ estimate very biased or only slightly biased?
  - Which direction is the bias? Why? There could be multiple OVBs acting in different directions
External Validity

External validity measures our ability to extrapolate conclusions from our study to outside its own setting.

- Does our study of California generalize to Massachusetts?
- Can we apply the results of a study from 1990 to today?
- Does our pilot experiment on 1,000 students scale to an entire country?
  - Ex: more spending on primary school in the Perry Pre-School Project

No study can be externally valid to all other settings.
Section 2

Binary Dependent Variables
Binary Dependent Variables

Previously, we discussed regression for continuous $Y$, but sometimes $Y$ is binary (0 or 1)

**Examples of binary $Y$:**

- Harvard admitted or denied
- Employed or unemployed
- War or no war
- Election victory or loss
- Mortgage application approved or rejected

When $Y$ is binary, predicted $Y$ is the probability that $Y = 1$:

$$\hat{Y}_i = \Pr(Y_i = 1)$$
Binary Dependent Variables

OLS is generally problematic when $Y$ is binary, because

- it can generate $\Pr(Y = 1)$ greater than 1 or less than 0
- it assumes changes in $X$ have a constant effect on $\Pr(Y = 1)$
Probit

Instead, we put a non-linear wrapper:

\[ Y_i = \Phi(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i) \]

Using Probit:

- \( \Phi(\cdot) \) is the **normal cumulative density function**
- \( z\text{-score} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \)
Probit and logit nearly identical (except very close to \( Y = 0 \) and \( Y = 1 \))
Both Probit and Logit are usually estimated using **Maximum Likelihood Estimation**

What are the coefficients $\beta_1, \beta_2, \ldots, \beta_j$ that produce expected probabilities $\hat{Y}_i = \Pr(Y_i = 1)$ most consistent with the data we observe $\{Y_1, Y_2, \ldots, Y_n\}$?
Section 3

Panel Data
Panel Data means we observe the same group of entities over time.

- $N =$ number of entities
- $T =$ number of time periods

Previously, we studied cross-section data, which was a snapshot of entities at just one period of time ($T = 1$).
Panel Data

Examples:
- Alcohol-related fatalities by state over time
  - What is the effect of a beer tax on alcohol-related fatalities?
- Terrorist fatalities by country over time
  - What is the effect of repressing political freedom on terrorism?
- GPA by student over 1st grade, 2nd grade, 3rd grade, ...
  - What is the effect of a good teacher on test scores?
- Crop output by farmer over time
  - Do irrigation and fertilization strategies help farmers in developing countries?
- Local unemployment by city by month
  - Did stimulus spending improve local labor market conditions during the Great Recession?
Advantages of Panel Data

Panel Data enables us to better control for entity-specific, time-invariant effects because we observe how the same entity responds to different $X$’s.

Example:
Suppose we could observe the same school with different Student-Teacher ratios over time.

Then we would be less worried about biases arising from comparing schools with different student compositions.

This can greatly allay our fears of omitted variable bias (but still have to worry about time-variant causes of OVB).
Entity Fixed Effects

\[ Y_{it} = \alpha_i + \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it} \]

Entity Fixed Effects \( \alpha_i \) allow each entity \( i \) to have a different intercept

- Using entity FE controls for any factors that vary across states but are constant over time
  - e.g. geography, environmental factors, anything relatively static across time

- Entity FE are like per-entity dummy variables

Entity FE means we are using only **within-entity** variation for identification
Time Fixed Effects

\[ Y_{it} = \alpha_i + \gamma_t + \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it} \]

**Time Fixed Effects** $\gamma_t$ control for factors constant across entities but not across time.

Time FE are basically dummy variables for time. Ex:

\[ Y_{it} = \alpha_i + \gamma_1 1\{t = 2013\} + \gamma_2 1\{t = 2012\} + \cdots + \gamma_{10} 1\{t = 2014\} + \ldots \]

\[ \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it} \]
Autocorrelation in Panel Data

Typically we assume that observations are independent, so $u_i$ are independent too.

With panel data, we observe the same entity over time, so $u_{it}$ and $u_{it+1}$ may be correlated. Shocks are likely to linger more than a single period.

Since it is the same entity over time, we say the errors are serially correlated or autocorrelated (i.e. correlated with itself).

Assuming independence will yield wrong standard error estimates!
Clustering Standard Errors by Entity

To account for this autocorrelation, we cluster standard errors by entity

```
xtreg y x1 x2, fe vce(cluster entity)
```

This assumes that observations across different entities are still independent, but observations within the same entity may be correlated.
Section 4

Instrumental Variables
Instrumental Variables

**Instrumental Variables** (IV) are useful for estimating models

- with simultaneous causality or

- with omitted variable bias
  - IV especially useful when we cannot plausibly control for all omitted variables

More generally, IV useful whenever conditional mean independence of $X$ fails
Instrumental Variables

What is the causal relationship between $X$ and $Y$?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Suppose our model suffers from omitted variable bias and simultaneous casuality, so CMI fails.

Hence, OLS produces a biased estimate of $\hat{\beta}_1$ that we cannot interpret causally.

Suppose we have another variable $Z$ that is a valid instrument, then we can recover a $\hat{\beta}_{IV}$ with a causal interpretation.
IV Conditions

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + u_i \]

Suppose OLS yields a biased estimate \( \hat{\beta}_1 \) because conditional mean independence fails:

\[ \mathbb{E}[u_i | X_i, W_{1i}] \neq \mathbb{E}[u_i | W_{1i}] \]

**Conditions for IV**

\( Z \) is a valid instrumental variable for \( X \) if:

1. **Relevance:** \( Z \) is related to \( X \):
   \[ \text{Corr}(Z, X) \neq 0 \]

2. **Exogeneity:** Controlling for \( W \)s, the only effect that \( Z \) has on \( Y \) goes through \( X \):
   \[ \text{Corr}(Z, u) = 0 \]
IV Conditions – Graphically

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Suppose we are investigating the effect of \( X \rightarrow Y \) but we suffer OVB and simultaneous causality:

- Omitted variables \( W_{1i} \) and \( W_{2i} \):
  - \( W_1 \rightarrow Y \) and \( W_1 \leftrightarrow X \)
  - \( W_2 \rightarrow Y \) and \( W_2 \leftrightarrow X \)

- Simultaneous causality
  - \( Y \rightarrow X \)
  - \( Y \rightarrow S_1 \rightarrow X \)

Suppose we have an instrument \( Z \)
Instrumental Variables

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Because of OVB and simultaneous causality: \( \mathbb{E}[u|X] \neq 0 \)
Instrumental Variables

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]
There can be multiple instruments $Z_1$ and $Z_2$ for the same $X$.

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]
IV

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

What is NOT allowed (by Condition 2: Exogeneity) –

- \( Z_i \leftrightarrow Y_i \)
- \( Z_i \leftrightarrow W_{1i} \) and \( Z \leftrightarrow W_{2i} \)
- \( Z_i \leftrightarrow S_{1i} \)
- \( \ldots \)

Suppose we also include control variable \( W_{1i} \)

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + u_i \]

What is NOT allowed (by Condition 2: Exogeneity) –

- \( Z_i \leftrightarrow Y_i \)
- \( Z_i \leftrightarrow W_{2i} \)
- \( Z_i \leftrightarrow S_{1i} \)
- \( \ldots \)
Conditions for IV: Intuition

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

\( Z \) is an instrumental variable for \( X \) in this model if:

**Condition 1: Relevance**

\( Z \) is related to \( X \) \( \text{Corr}(Z, X) \neq 0 \)

**Condition 2: Exogeneity of \( Z \)**

\( \text{Corr}(Z, u) = 0 \)

Two ways of saying the **Exogeneity** condition:

- \( Z \) is as-if randomly assigned
- The only effect of \( Z \) on \( Y \) goes through \( X \)

Both C1 and C2 must hold after controlling for \( Ws \)
### Examples of IVs

<table>
<thead>
<tr>
<th>$Z$</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How does prenatal health affect long-run development?</strong></td>
<td>Pregnancy during Ramadan</td>
<td>Prenatal health</td>
</tr>
<tr>
<td><strong>What effect does serving in the military have on wages?</strong></td>
<td>Military draft lottery #</td>
<td>Military service</td>
</tr>
<tr>
<td><strong>What is the effect of riots on community development?</strong></td>
<td>Rainfall during month of MLK assassination</td>
<td>Number of riots</td>
</tr>
</tbody>
</table>

Each of these examples requires some control variables $W$s for the exogeneity condition to hold. In general, arguing the exogeneity condition can be very difficult.
Testing the Validity of Instruments

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i \]

Conditions for IV

\( Z \) is an instrumental variable for \( X \) in this model if:

1. **Relevance**: \( \text{Corr}(Z, X) \neq 0 \)
2. **Exogeneity**: \( \text{Corr}(Z, u) = 0 \)

- Testing **Condition 1** is straightforward, since we have data on both \( Z \) and \( X \)
- Testing **Condition 2** is trickier, because we never observe \( u \). In fact, we can only test Condition 2 when we have more instruments \( Zs \) than endogenous \( Xs \)
Testing Condition 1: Relevance

**Condition 1: Relevance**

$Z$ must be related to $X$. i.e. $\text{Corr}(Z, X) \neq 0$

We need the relationship between $X$ and $Z$ to be meaningfully “large”

**How to check?**

Run **first-stage** regression with OLS (if we have multiple instruments, include all of them)

$$X_i = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 W_{1i} + \alpha_4 W_{2i} + \cdots + \nu_i$$

Check that the F-test on all the coefficients on the instruments $\alpha_1, \alpha_2$

- If $\hat{F} > 10$, we claim that $Z$ is a strong instrument
- If $\hat{F} \leq 10$, we have a weak instruments problem
Testing Condition 2: Exogeneity

J-test for overidentifying restrictions:

\[ H_0 : \text{Both } Z_1 \text{ and } Z_2 \text{ satisfy the exogeneity condition} \]
\[ H_a : \text{Either } Z_1, Z_2, \text{ or both are invalid instruments} \]

```
ivregress y w1 w2 (x = z1 z2), robust
estat overid
display "J-test = " r(score) " p-value = " r(p_score)
```

If the p-value < 0.05, then we reject the null hypothesis that all our instruments are valid.

But just like in the F-test case, rejecting the test alone does not reveal which instrument is invalid, only that at least one fails the exogeneity condition.
Testing Condition 2: Exogeneity

**Condition 2: Exogeneity of Z**

Z is as-if randomly assigned. i.e. $\text{Corr}(Z, u) = 0$

To check instrument exogeneity, we need more instruments Zs than endogenous Xs

Suppose there is one treatment variable of interest $X$, multiple Zs, potentially, multiple control variables $W$s.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

If $Z_1$ and $Z_2$ are both valid instruments, then

$\text{Corr}(Z_1, u) = 0$ and $\text{Corr}(Z_2, u) = 0$
Two-Stage Least Squares

IV regression is typically estimated using **two-stage least squares**

**First Stage:** Regress $X$ on $Z$ and $W$

$$X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 W_i + v_i$$

**Second Stage:** Regress $Y$ on $\hat{X}$ and $W$

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_i + u_i$$

```
ivregress 2sls y w (x = z), robust
```

**Intuition:** If the instrument satisfies the two IV conditions:

- The first stage removes the “endogenous” parts of $X$
- $\hat{X}$ satisfies conditional mean independence
Local Average Treatment Effect (LATE)

If OLS is unbiased, $\hat{\beta}_{OLS}$ identifies the **average treatment effect**

The IV estimate $\hat{\beta}_{IV}$ identifies the **local average treatment effect** (LATE)

**INTUITION:**
LATE is the weighted-average treatment effect for entities affected by the instrument, weighted more heavily toward those most affected by the instrument $Z$

The word *local* indicates the LATE is the average for this affected group known as **compliers**. Compliers are those affected by the instrument (i.e. they **complied** with $Z$)
ATE v. LATE

\[ \beta_{1,i} = \text{causal impact of } X \text{ on } Y \text{ for individual } i \]
\[ \Pi_{1,i} = \text{correlation between } X \text{ and } Z \text{ for individual } i \]

LATE = ATE if at least one of the following is true

- no heterogeneity in treatment effects of \( X \) on \( Y \)
  \[ \beta_{1,i} = \beta_1 \text{ for all } i \]

- no heterogeneity in first-stage relationship between \( Z \) and \( X \)
  \[ \Pi_{1,i} = \Pi_1 \text{ for all } i \]

- no correlation between the heterogeneity in response to instrument \( Z \) and heterogeneity in response to treatment \( X \)
  \[ \text{Cov}(\beta_{1,i}, \Pi_{1,i}) = 0 \]
More LATE

Suppose we are estimating the causal effect of $X$ on $Y$:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{2i} + \beta_3 W_{3i} + u_i$$

We have two valid instruments $Z_1$ and $Z_2$.

- We just use $Z_1$ and run 2SLS to estimate $\hat{\beta}_{2SLS}$
- We just use $Z_2$ and run 2SLS to estimate $\tilde{\beta}_{2SLS}$

Should we expect our estimates to equal?

$$\hat{\beta}_{2SLS} \stackrel{?}{=} \tilde{\beta}_{2SLS}$$

No.

Different groups may respond to the different instruments $Z_1$ and $Z_2$. So each instrument may have a different LATE
Internal Validity with IVs

If the IV is valid, then instrument variable regression takes care of:

- Omitted Variable Bias
- Simultaneous Causality Bias
- Errors-in-variables (or measurement error)

Thus, internal validity in an IV regression is mostly about assessing:

- Relevance of $Z$
- Exogeneity of $Z$
Section 5

Forecasting
With forecasting, forget causality. It’s all about **prediction**.

How can we use past $Y$ to predict future $Y$?

**i.e.** What can $(Y_{t-4}, Y_{t-3}, \ldots, Y_t)$ tell us about $Y_{t+1}$ or $Y_{t+n}$?

**Examples of $Y_t$**

- GDP
- Oil prices
- Stock market indices
- Exchanges rates
- …
Forecasting Vocabulary

- **Lag**: $Y_{t-p}$ is the $p^{th}$ lag of $Y_t$

- **Autocovariance** – covariance of a variable with a lag of itself
  
  \[
  \text{Cov}(Y_t, Y_{t-j}) \quad \text{“} j^{th} \text{ autocovariance”}
  \]

- **Autocorrelation** – correlation of a variable with a lag of itself
  
  \[
  \text{Corr}(Y_t, Y_{t-j}) \quad \text{“} j^{th} \text{ autocorrelation”}
  \]

- Both autocovariance and autocorrelation measure how $Y_{t-j}$ is related to $Y_t$
Stationarity

For the past to be useful for predicting the future, the process must be stationary

Let $Y_t$ be a process that evolves over time: $(Y_{t_0}, Y_{t_0+1}, \ldots, Y_T)$.

$Y_t$ is stationary if all three are true:

- $\mathbb{E}[Y_t]$ is constant over time
- $\text{Var}(Y_t)$ is constant over time
- Autocorrelation $\text{Corr}(Y_t, Y_{t-j})$ depends only on $j$ and not $t$

i.e. the behavior of $Y_t$ isn't fundamentally changing over time
Examples of Non-Stationarity

Non-stationary. \( \text{Var}(Y_t) \) is increasing over time

Non-stationary. \( \mathbb{E}[Y_t] \) is increasing over time
First Differences

Even for $Y_t$ is non-stationary, first-differences might be stationary!

$$\Delta Y_t = Y_t - Y_{t-1}$$
# Forecasting Models

Assuming $Y_t$ is stationary:

### AR($p$): Autoregressive Model of Order $p$:

Regression of $Y_t$ on $p$ lags of $Y_t$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$$

**Ex:** AR(1): $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$

**Ex:** AR(4): $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + u_t$

### ADL($p,q$): Autoregressive Distributed Lag Model

Regression of $Y_t$ on $p$ lags of $Y_t$ and $q$ lags of $X_t$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + u_t$$
Model Selection: Choosing number of lags $p$

How many lags $p$ should the model include?

$$\text{AR}(p) : \quad Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$$

We choose $p$ by minimizing the information criterion:

**Information Criterion**

- information criterion is a measure of how much information in our dataset is not captured by our model

Intuitive that we want to choose the model (i.e. choose $p$) with the smallest IC
Minimizing Information Criterion IC(p)

Choose \( p \) to minimize **Bayes’ information criterion**, \( \text{BIC}(p) \)

\[
\min_{0 \leq p \leq p_{\text{max}}} \text{BIC}(p) = \min_{0 \leq p \leq p_{\text{max}}} \ln \left( \frac{\text{SSR}(p)}{T} \right) + (p + 1) \left( \frac{\ln T}{T} \right)
\]

- \( \text{SSR}(p) \) is the sum of squared residuals when number of lags is \( p \)
  - \( \text{SSR}(p) \) is the variation in \( Y_t \) not captured by the model

- \( \left( \frac{\ln T}{T} \right) \) is a “penalty” factor associated with increasing \( p \) by 1
  - Need this penalty term because \( \text{SSR}(p) \) is always decreasing in \( p \)

- BIC trades off the *decrease* in bias from including important lags against the *increase* in variance from including irrelevant lags
Testing Stationarity

- Often difficult to confirm \( Y_t \) is stationary
- Breaks (or **structural breaks**) imply non-stationarity because the underlying relationships between \( Y_t \) and \( Y_{t-j} \) have changed

**Ex:** Was there a break? If so, when did the break occur?

Levels \( Y_t \)  
Differences \( Y_t - Y_{t-1} \)
Chow Test for Structural Breaks

Suppose you want to test if there was a break at specific time $\tau$:

\[ T_0 \quad \tau \quad T_1 \]

\[ D_{\tau,t} = \begin{cases} 
1 & \text{if } t \geq \tau \\
0 & \text{if } t < \tau 
\end{cases} \]

**AR(1):** \[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 D_{\tau,t-1} + \gamma_1 (Y_{t-1} \times D_{\tau,t-1}) + u_t \]

Chow test:

\[ H_0 : \delta_1 = \gamma_1 = 0 \]

If reject:

We have evidence that the relationship between $Y_t$ and lag $Y_{t-1}$ differs over time. Hence, we reject stationarity.
QLR Test

We can use the Chow Test when testing for a break at a specific time $\tau$

What if we want to test for any structural breaks across all time periods?

We calculate the **Quandt Likelihood Ratio Statistic**, which is the maximum Chow statistic across all $\tau$ in the central 70% of the time interval

$$QLR = \max_{\tau} F(\tau)$$

Calculating the Chow statistic for $\{\tau_0, \tau_0 + 1, \tau_0 + 2, \ldots, \tau_1 - 1, \tau_1\}$ to find the maximum manually can be very time-consuming

Fortunately, someone has written a `qlr` command in STATA
Good luck!