Week 5: Simple Linear Regression

Brandon Stewart

Princeton

October 10, 12, 2016

\footnote{These slides are heavily influenced by Matt Blackwell, Adam Glynn and Jens Hainmueller. Illustrations by Shay O’Brien.}
Where We’ve Been and Where We’re Going…

Last Week

▶ hypothesis testing
▶ what is regression

This Week

▶ Monday:
  ⋆ mechanics of OLS
  ⋆ properties of OLS

▶ Wednesday:
  ⋆ hypothesis tests for regression
  ⋆ confidence intervals for regression
  ⋆ goodness of fit

Next Week

▶ mechanics with two regressors
▶ omitted variables, multicollinearity

Long Run

▶ probability
  → inference
  → regression

Questions?

Stewart (Princeton)
Week 5: Simple Linear Regression
October 10, 12, 2016
2 / 103
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

- Long Run
  - probability
  - → inference
  - → regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:

- **Next Week**
  - mechanics with two regressors
  - omitted variables, multicollinearity

- **Long Run**
  - probability → inference → regression

Questions?
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

Questions?

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

Next Week
- mechanics with two regressors
- omitted variables, multicollinearity

Long Run
- probability
  - → inference
  - → regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- **Next Week**
  - mechanics with two regressors
  - omitted variables, multicollinearity
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?
Macrostructure

The next few weeks,

- Linear Regression with Two Regressors
- Multiple Linear Regression
- Break Week
Macrostructure

The next few weeks,

- Linear Regression with Two Regressors
- Multiple Linear Regression
- Break Week
- Regression in the Social Science
- What Can Go Wrong and How to Fix It Week 1
- What Can Go Wrong and How to Fix It Week 2 / Thanksgiving
Macrostructure

The next few weeks,

- Linear Regression with Two Regressors
- Multiple Linear Regression
- Break Week
- Regression in the Social Science
- What Can Go Wrong and How to Fix It Week 1
- What Can Go Wrong and How to Fix It Week 2 / Thanksgiving
- Causality with Measured Confounding
- Unmeasured Confounding and Instrumental Variables
- Repeated Observations and Panel Data
Macrostructure

The next few weeks,

- Linear Regression with Two Regressors
- Multiple Linear Regression
- Break Week
- Regression in the Social Science
- What Can Go Wrong and How to Fix It Week 1
- What Can Go Wrong and How to Fix It Week 2 / Thanksgiving
- Causality with Measured Confounding
- Unmeasured Confounding and Instrumental Variables
- Repeated Observations and Panel Data

A brief comment on exams, midterm week etc.
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
1. Mechanics of OLS

2. Properties of the OLS estimator

3. Example and Review

4. Properties Continued

5. Hypothesis tests for regression

6. Confidence intervals for regression

7. Goodness of fit

8. Wrap Up of Univariate Regression

9. Fun with Non-Linearities
The population linear regression function

The population simple linear regression model can be stated as the following:

\[ r(x) = E[Y|X=x] = \beta_0 + \beta_1 x \]

This (partially) describes the data generating process in the population:

- \( Y \): dependent variable
- \( X \): independent variable
- \( \beta_0, \beta_1 \): population intercept and population slope (what we want to estimate)
The (population) simple linear regression model can be stated as the following:

\[ r(x) = E[Y|X = x] = \beta_0 + \beta_1 x \]
The population linear regression function

- The (population) simple linear regression model can be stated as the following:

\[ r(x) = E[Y|X = x] = \beta_0 + \beta_1 x \]

- This (partially) describes the data generating process in the population
The population linear regression function

- The (population) simple linear regression model can be stated as the following:

\[ r(x) = E[Y|X = x] = \beta_0 + \beta_1 x \]

- This (partially) describes the data generating process in the population
- \( Y \) = dependent variable
The population linear regression function

- The (population) simple linear regression model can be stated as the following:
  \[ r(x) = E[Y|X = x] = \beta_0 + \beta_1 x \]
- This (partially) describes the data generating process in the population
- \( Y = \) dependent variable
- \( X = \) independent variable
The population linear regression function

The (population) simple linear regression model can be stated as the following:

\[ r(x) = E[Y|X = x] = \beta_0 + \beta_1 x \]

This (partially) describes the data generating process in the population

- \( Y \) = dependent variable
- \( X \) = independent variable
- \( \beta_0, \beta_1 \) = population intercept and population slope (what we want to estimate)
The sample linear regression function

The estimated or sample regression function is:

\[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \]

\( \hat{\beta}_0, \hat{\beta}_1 \) are the estimated intercept and slope

\( \hat{Y}_i \) is the fitted/predicted value

We also have the residuals, \( \hat{u}_i \) which are the differences between the true values of \( Y \) and the predicted value:

\[ \hat{u}_i = Y_i - \hat{Y}_i \]

You can think of the residuals as the prediction errors of our estimates.
The sample linear regression function

- The estimated or sample regression function is:

\[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]
The sample linear regression function

- The estimated or sample regression function is:

\[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

- \( \hat{\beta}_0, \hat{\beta}_1 \) are the estimated intercept and slope

\( \hat{Y}_i \) is the fitted/predicted value.
The sample linear regression function

- The estimated or sample regression function is:

\[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

- \( \hat{\beta}_0, \hat{\beta}_1 \) are the estimated intercept and slope
- \( \hat{Y}_i \) is the fitted/predicted value

We also have the residuals, \( \hat{u}_i \) which are the differences between the true values of \( Y \) and the predicted value:

\[ \hat{u}_i = Y_i - \hat{Y}_i \]

You can think of the residuals as the prediction errors of our estimates.
The sample linear regression function

- The estimated or sample regression function is:

\[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

- \( \hat{\beta}_0, \hat{\beta}_1 \) are the estimated intercept and slope
- \( \hat{Y}_i \) is the fitted/predicted value
- We also have the residuals, \( \hat{u}_i \) which are the differences between the true values of \( Y \) and the predicted value:

\[ \hat{u}_i = Y_i - \hat{Y}_i \]
The sample linear regression function

- The estimated or sample regression function is:
  \[ \hat{r}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

- \( \hat{\beta}_0, \hat{\beta}_1 \) are the estimated intercept and slope
- \( \hat{Y}_i \) is the fitted/predicted value
- We also have the residuals, \( \hat{u}_i \) which are the differences between the true values of Y and the predicted value:
  \[ \hat{u}_i = Y_i - \hat{Y}_i \]

- You can think of the residuals as the prediction errors of our estimates.
Overall Goals for the Week

- Learn how to run and read regression
- Mechanics: how to estimate the intercept and slope?
- Properties: when are these good estimates?
- Uncertainty: how will the OLS estimator behave in repeated samples?
- Testing: can we assess the plausibility of no relationship ($\beta_1 = 0$)?
- Interpretation: how do we interpret our estimates?
Overall Goals for the Week

- Learn how to run and read regression
Overall Goals for the Week

- Learn how to run and read regression
- **Mechanics**: how to estimate the intercept and slope?
Overall Goals for the Week

- Learn how to run and read regression
- **Mechanics**: how to estimate the intercept and slope?
- **Properties**: when are these good estimates?
Overall Goals for the Week

- Learn how to run and read regression
- **Mechanics**: how to estimate the intercept and slope?
- **Properties**: when are these good estimates?
- **Uncertainty**: how will the OLS estimator behave in repeated samples?
Overall Goals for the Week

- Learn how to run and read regression
- **Mechanics**: how to estimate the intercept and slope?
- **Properties**: when are these good estimates?
- **Uncertainty**: how will the OLS estimator behave in repeated samples?
- **Testing**: can we assess the plausibility of no relationship ($\beta_1 = 0$)?
Overall Goals for the Week

- Learn how to run and read regression
- **Mechanics**: how to estimate the intercept and slope?
- **Properties**: when are these good estimates?
- **Uncertainty**: how will the OLS estimator behave in repeated samples?
- **Testing**: can we assess the plausibility of no relationship ($\beta_1 = 0$)?
- **Interpretation**: how do we interpret our estimates?
What is OLS?

- An estimator for the slope and the intercept of the regression line

Ordinary Least Squares (OLS):

\[
(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
\]

In words, the OLS estimates are the intercept and slope that minimize the sum of the squared residuals.
What is OLS?

- An estimator for the slope and the intercept of the regression line
- We talked last week about ways to derive this estimator and we settled on deriving it by minimizing the squared prediction errors of the regression, or in other words, minimizing the sum of the squared residuals:

  \[
  \hat{\beta}_0, \hat{\beta}_1 = \arg \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
  \]

In words, the OLS estimates are the intercept and slope that minimize the sum of the squared residuals.
What is OLS?

- An estimator for the slope and the intercept of the regression line
- We talked last week about ways to derive this estimator and we settled on deriving it by minimizing the squared prediction errors of the regression, or in other words, minimizing the sum of the squared residuals:
- **Ordinary Least Squares (OLS):**

\[
(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
\]
What is OLS?

- An estimator for the slope and the intercept of the regression line
- We talked last week about ways to derive this estimator and we settled on deriving it by minimizing the squared prediction errors of the regression, or in other words, minimizing the sum of the squared residuals:

  **Ordinary Least Squares (OLS):**

  \[
  (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
  \]

- In words, the OLS estimates are the intercept and slope that minimize the sum of the squared residuals.
How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?

Answer: We will minimize the squared sum of residuals.
Graphical Example

How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?

Answer: We will minimize the squared sum of residuals.
Graphical Example

How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?
Graphical Example

How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?

Answer: We will minimize the squared sum of residuals

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} u_i^2$$

Residual $u_i$ is “part” of $Y_i$ not predicted

$$u_i = Y_i - \hat{Y}_i$$

$Y$: % African Americans in State Legislature

$X$: % African American Population

$Y = -1.3 + .36 X$
Deriving the OLS estimator

Let's think about \( n \) pairs of sample observations:
\[
(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)
\]
Let \( \{ b_0, b_1 \} \) be possible values for \( \{ \beta_0, \beta_1 \} \)

Define the least squares objective function:
\[
S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
\]

How do we derive the LS estimators for \( \beta_0 \) and \( \beta_1 \)? We want to minimize this function, which is actually a very well-defined calculus problem.

1. Take partial derivatives of \( S \) with respect to \( b_0 \) and \( b_1 \).
2. Set each of the partial derivatives to 0.
3. Solve for \( \{ b_0, b_1 \} \) and replace them with the solutions.

To the board we go!
Deriving the OLS estimator

Let’s think about \( n \) pairs of sample observations: 
\((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\)
Deriving the OLS estimator

- Let’s think about $n$ pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
Deriving the OLS estimator

- Let’s think about $n$ pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$
Deriving the OLS estimator

- Let’s think about $n$ pairs of sample observations: 
  $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

  $$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$ 

- How do we derive the LS estimators for $\beta_0$ and $\beta_1$? We want to minimize this function, which is actually a very well-defined calculus problem.
Deriving the OLS estimator

- Let’s think about \( n \) pairs of sample observations: \((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\)
- Let \( \{b_0, b_1\} \) be possible values for \( \{\beta_0, \beta_1\} \)
- Define the least squares objective function:

\[
S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.
\]

- How do we derive the LS estimators for \( \beta_0 \) and \( \beta_1 \)? We want to minimize this function, which is actually a very well-defined calculus problem.

1. Take partial derivatives of \( S \) with respect to \( b_0 \) and \( b_1 \).
Deriving the OLS estimator

- Let’s think about \( n \) pairs of sample observations: 
  \((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\)
- Let \( \{b_0, b_1\} \) be possible values for \( \{\beta_0, \beta_1\} \)
- Define the least squares objective function:
  \[
  S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.
  \]

How do we derive the LS estimators for \( \beta_0 \) and \( \beta_1 \)? We want to minimize this function, which is actually a very well-defined calculus problem.

1. Take partial derivatives of \( S \) with respect to \( b_0 \) and \( b_1 \).
2. Set each of the partial derivatives to 0
Deriving the OLS estimator

- Let’s think about \( n \) pairs of sample observations: \((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\)
- Let \( \{b_0, b_1\} \) be possible values for \( \{\beta_0, \beta_1\} \)
- Define the least squares objective function:

\[
S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.
\]

- How do we derive the LS estimators for \( \beta_0 \) and \( \beta_1 \)? We want to minimize this function, which is actually a very well-defined calculus problem.
  1. Take partial derivatives of \( S \) with respect to \( b_0 \) and \( b_1 \).
  2. Set each of the partial derivatives to 0
  3. Solve for \( \{b_0, b_1\} \) and replace them with the solutions
Deriving the OLS estimator

- Let’s think about \( n \) pairs of sample observations: 
  \((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\)
- Let \( \{b_0, b_1\} \) be possible values for \( \{\beta_0, \beta_1\} \)
- Define the least squares objective function:
  \[
  S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.
  \]

- How do we derive the LS estimators for \( \beta_0 \) and \( \beta_1 \)? We want to minimize this function, which is actually a very well-defined calculus problem.
  1. Take partial derivatives of \( S \) with respect to \( b_0 \) and \( b_1 \).
  2. Set each of the partial derivatives to 0
  3. Solve for \( \{b_0, b_1\} \) and replace them with the solutions

- To the board we go!
The OLS estimator

Now we’re done! Here are the OLS estimators:

\[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \]

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]
Intuition of the OLS estimator

The intercept equation tells us that the regression line goes through the point \((Y, X)\):

\[ Y = \hat{\beta}_0 + \hat{\beta}_1 X \]

The slope for the regression line can be written as the following:

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \]

Sample Covariance between \(X\) and \(Y\)
Sample Variance of \(X\)

The higher the covariance between \(X\) and \(Y\), the higher the slope will be.

Negative covariances \(\rightarrow\) negative slopes; positive covariances \(\rightarrow\) positive slopes

What happens when \(X_i\) doesn’t vary?
What happens when \(Y_i\) doesn’t vary?
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\bar{Y}, \bar{X})\):

\[
\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}
\]
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\bar{Y}, \bar{X})\):
  \[
  \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}
  \]

- The slope for the regression line can be written as the following:
  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
  \]
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\bar{Y}, \bar{X})\):

  \[
  \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}
  \]

- The slope for the regression line can be written as the following:

  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
  \]

- The higher the covariance between \(X\) and \(Y\), the higher the slope will be.

Negative covariances \(\rightarrow\) negative slopes;
Positive covariances \(\rightarrow\) positive slopes
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\overline{Y}, \overline{X})\):
  \[
  \overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}
  \]

- The slope for the regression line can be written as the following:
  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
  \]

- The higher the covariance between \(X\) and \(Y\), the higher the slope will be.
- Negative covariances \(\rightarrow\) negative slopes; positive covariances \(\rightarrow\) positive slopes
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\bar{Y}, \bar{X})\):
  \[
  \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}
  \]

- The slope for the regression line can be written as the following:
  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
  \]

- The higher the covariance between \(X\) and \(Y\), the higher the slope will be.
- Negative covariances \(\rightarrow\) negative slopes;
  positive covariances \(\rightarrow\) positive slopes
- What happens when \(X_i\) doesn’t vary?
Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point \((\overline{Y}, \overline{X})\):
  \[
  \overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}
  \]

- The slope for the regression line can be written as the following:
  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
  \]

- The higher the covariance between \(X\) and \(Y\), the higher the slope will be.
- Negative covariances \(\rightarrow\) negative slopes; positive covariances \(\rightarrow\) positive slopes
- What happens when \(X_i\) doesn’t vary?
- What happens when \(Y_i\) doesn’t vary?
A Visual Intuition for the OLS Estimator
A Visual Intuition for the OLS Estimator
A Visual Intuition for the OLS Estimator
A Visual Intuition for the OLS Estimator
A Visual Intuition for the OLS Estimator
Mechanical properties of OLS

Later we'll see that under certain assumptions, OLS will have nice statistical properties. But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The residuals will be 0 on average:
   \[ \sum_{i=1}^{n} \hat{u}_i = 0 \]

2. The residuals will be uncorrelated with the predictor (\( \hat{\text{cov}} \) is the sample covariance):
   \[ \hat{\text{cov}}(X_i, \hat{u}_i) = 0 \]

3. The residuals will be uncorrelated with the fitted values:
   \[ \hat{\text{cov}}(\hat{Y}_i, \hat{u}_i) = 0 \]
Mechanical properties of OLS

- Later we’ll see that under certain assumptions, OLS will have nice statistical properties.

1. The residuals will be 0 on average:
   \[ \sum_{i=1}^{n} \hat{u}_i = 0 \]

2. The residuals will be uncorrelated with the predictor (\( \hat{\text{cov}} \) is the sample covariance):
   \[ \hat{\text{cov}}(X_i, \hat{u}_i) = 0 \]

3. The residuals will be uncorrelated with the fitted values:
   \[ \hat{\text{cov}}(\hat{Y}_i, \hat{u}_i) = 0 \]
Mechanical properties of OLS

- Later we’ll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The residuals will be 0 on average:
   \[ \sum_{i=1}^{n} \hat{u}_i = 0 \]

2. The residuals will be uncorrelated with the predictor (\(\hat{\text{cov}}\) is the sample covariance):
   \[ \hat{\text{cov}}(\hat{X}_i, \hat{u}_i) = 0 \]

3. The residuals will be uncorrelated with the fitted values:
   \[ \hat{\text{cov}}(\hat{Y}_i, \hat{u}_i) = 0 \]
Mechanical properties of OLS

- Later we’ll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The residuals will be 0 on average:

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0
\]
Mechanical properties of OLS

- Later we’ll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The residuals will be 0 on average:

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0
\]

2. The residuals will be uncorrelated with the predictor (\(\hat{\text{cov}}\) is the sample covariance):

\[
\hat{\text{cov}}(X_i, \hat{u}_i) = 0
\]
Mechanical properties of OLS

- Later we’ll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The residuals will be 0 on average:

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0
\]

2. The residuals will be uncorrelated with the predictor (\(\hat{\text{cov}}\) is the sample covariance):

\[
\hat{\text{cov}}(X_i, \hat{u}_i) = 0
\]

3. The residuals will be uncorrelated with the fitted values:

\[
\hat{\text{cov}}(\hat{Y}_i, \hat{u}_i) = 0
\]
OLS slope as a weighted sum of the outcomes

\[ \hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} W_i Y_i \]

Where here we have the weights, \( W_i \) as:

\[ W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

This is important for two reasons. First, it'll make derivations later much easier. And second, it shows that it is just the sum of a random variable. Therefore it is also a random variable.
One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

\[ \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i \]
OLS slope as a weighted sum of the outcomes

- One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

\[ \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i \]

Where here we have the weights, \( W_i \) as:

\[ W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

- This is important for two reasons. First, it'll make derivations later much easier. And second, it shows that \( \hat{\beta}_1 \) is just the sum of a random variable. Therefore it is also a random variable.
OLS slope as a weighted sum of the outcomes

One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i$$

Where here we have the weights, $W_i$ as:

$$W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

This is important for two reasons. First, it’ll make derivations later much easier. And second, it shows that is just the sum of a random variable. Therefore it is also a random variable.
One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

\[ \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i \]

Where here we have the weights, \( W_i \) as:

\[ W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

This is important for two reasons. First, it’ll make derivations later much easier. And second, it shows that is just the sum of a random variable. Therefore it is also a random variable.

To the board!
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Sampling distribution of the OLS estimator

Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.

\[
\text{Sample 1: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \Rightarrow (\hat{\beta}_0, \hat{\beta}_1)
\]

\[
\text{Sample 2: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \Rightarrow (\hat{\beta}_0, \hat{\beta}_1)
\]

\[\ldots\]

\[
\text{Sample } k-1: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \Rightarrow (\hat{\beta}_0, \hat{\beta}_1)
\]

\[
\text{Sample } k: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \Rightarrow (\hat{\beta}_0, \hat{\beta}_1)
\]

Just like the sample mean, sample difference in means, or the sample variance, it has a sampling distribution, with a sampling variance/standard error, etc.

Let's take a simulation approach to demonstrate:

▶ Pretend that the AJR data represents the population of interest
▶ See how the line varies from sample to sample
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

Sample 1: \( \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \) \( \Rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1 \)
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

Sample 1: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1

Sample 2: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

```
Sample 1: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1
Sample 2: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2
\vdots
Sample \( k-1 \): \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_{k-1}
Sample \( k \): \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_k
```

Just like the sample mean, sample difference in means, or the sample variance, it has a sampling distribution, with a sampling variance/standard error, etc.

Let’s take a simulation approach to demonstrate:

- Pretend that the AJR data represents the population of interest
- See how the line varies from sample to sample
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

\[
\text{Sample 1: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \quad \rightarrow \quad (\hat{\beta}_0, \hat{\beta}_1)_1
\]
\[
\text{Sample 2: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \quad \rightarrow \quad (\hat{\beta}_0, \hat{\beta}_1)_2
\]
\[\vdots\]
\[
\text{Sample } k-1: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \quad \rightarrow \quad (\hat{\beta}_0, \hat{\beta}_1)_{k-1}
\]
\[
\text{Sample } k: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \quad \rightarrow \quad (\hat{\beta}_0, \hat{\beta}_1)_k
\]

- Just like the sample mean, sample difference in means, or the sample variance
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

\[
\text{Sample 1: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1 \\
\text{Sample 2: } \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2 \\
\vdots \\
\text{Sample } k - 1: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_{k-1} \\
\text{Sample } k: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_k
\]

- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get estimates.

Sample 1: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1
Sample 2: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2
\vdots
Sample k - 1: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_{k-1}
Sample k: \{ (Y_1, X_1), \ldots, (Y_n, X_n) \} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_k

- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let’s take a simulation approach to demonstrate:
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

\[ \text{Sample } 1: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1 \]
\[ \text{Sample } 2: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2 \]
\[ \vdots \]
\[ \text{Sample } k - 1: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_{k-1} \]
\[ \text{Sample } k: \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_k \]

- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let’s take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it’s a machine that we plug data into and we get out estimates.

\[
\begin{align*}
\text{Sample 1: } & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1 \\
\text{Sample 2: } & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_2 \\
& \vdots \\
\text{Sample } k-1: & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_{k-1} \\
\text{Sample } k: & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_k
\end{align*}
\]

- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let’s take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`.
2. Use `lm()` to calculate the OLS estimates of the slope and intercept.
3. Plot the estimated regression line.
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
3. Plot the estimated regression line
Population Regression

![Graph showing the relationship between Log GDP per capita growth and Log Settler Mortality]
Randomly sample from AJR

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016 20 / 103
Randomly sample from AJR

![Graph showing the relationship between Log GDP per capita growth and Log Settler Mortality. The graph includes two regression lines and data points.](image-url)
Randomly sample from AJR

![Graph showing the relationship between Log GDP per capita growth and Log Settler Mortality.](image)

- **Log GDP per capita growth** vs. **Log Settler Mortality**
  - Data points scattered across the graph, illustrating a negative correlation.
  - Red line indicates the trend for theoretical data, while blue points show observed data.

---

**Stewart (Princeton)**

**Week 5: Simple Linear Regression**

**October 10, 12, 2016**
Randomly sample from AJR

[Graph showing the relationship between Log GDP per capita growth and Log Settler Mortality.]

Stewart (Princeton)
Log GDP per capita growth vs. Log Settler Mortality

Week 5: Simple Linear Regression
October 10, 12, 2016
Stewart (Princeton)
Randomly sample from AJR

Log GDP per capita growth vs. Log Settler Mortality

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016  20 / 103
Sampling distribution of OLS

You can see that the estimated slopes and intercepts vary from sample to sample, but that the "average" of the lines looks about right.

Sampling distribution of intercepts

\[ \beta^0 \]

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>300</td>
</tr>
</tbody>
</table>

Sampling distribution of slopes

\[ \beta^1 \]

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.5</td>
</tr>
<tr>
<td>-1.0</td>
</tr>
<tr>
<td>-0.5</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>300</td>
</tr>
</tbody>
</table>
Sampling distribution of OLS

- You can see that the estimated slopes and intercepts vary from sample to sample, but that the “average” of the lines looks about right.

Sampling distribution of intercepts

Sampling distribution of slopes

Is this unique?

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016
Sampling distribution of OLS

- You can see that the estimated slopes and intercepts vary from sample to sample, but that the “average” of the lines looks about right.

Sampling distribution of intercepts

Sampling distribution of slopes

- Is this unique?
Assumptions for unbiasedness of the sample mean

What assumptions did we make to prove that the sample mean was unbiased?

$E[\bar{X}] = \mu$

Just one: random sample

We'll need more than this for the regression case
What assumptions did we make to prove that the sample mean was unbiased?

$$\mathbb{E}[\bar{X}] = \mu$$
Assumptions for unbiasedness of the sample mean

- What assumptions did we make to prove that the sample mean was unbiased?
  \[ \mathbb{E}[\bar{X}] = \mu \]
- Just one: random sample
Assumptions for unbiasedness of the sample mean

What assumptions did we make to prove that the sample mean was unbiased?

\[ \mathbb{E}[ar{X}] = \mu \]

- Just one: random sample
- We’ll need more than this for the regression case
Our goal

What is the sampling distribution of the OLS slope $\hat{\beta}_1 \sim ?(?, ?)$? We need to fill in those ?s.

We'll start with the mean of the sampling distribution. Is the estimator centered at the true value, $\beta_1$? Most of our derivations will be in terms of the slope but they apply to the intercept as well.
Our goal

What is the sampling distribution of the OLS slope?

\[ \hat{\beta}_1 \sim ?(?, ?) \]
Our goal

What is the sampling distribution of the OLS slope?

\[ \hat{\beta}_1 \sim (?) , (?) \]

We need fill in those ?s.
Our goal

- What is the sampling distribution of the OLS slope?
  \[ \hat{\beta}_1 \sim (?) , (?) \]

- We need fill in those ?s.
- We’ll start with the mean of the sampling distribution. Is the estimator centered at the true value, \( \beta_1 \)?
Our goal

- What is the sampling distribution of the OLS slope?
  \[ \hat{\beta}_1 \sim ?(?, ?) \]

- We need fill in those ?s.
- We’ll start with the mean of the sampling distribution. Is the estimator centered at the true value, \( \beta_1 \)?
- Most of our derivations will be in terms of the slope but they apply to the intercept as well.
OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified.
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in $X$: There is variation in the explanatory variable.
4. Zero conditional mean: Expected value of the error term is zero conditional on all values of the explanatory variable.
5. Homoskedasticity: The error term has the same variance conditional on all values of the explanatory variable.
6. Normality: The error term is independent of the explanatory variables and normally distributed.
1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
OLS Assumptions Preview

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.

2. **Random Sampling**: The observed data represent a random sample from the population described by the model.
OLS Assumptions Preview

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.

2. **Random Sampling**: The observed data represent a random sample from the population described by the model.

3. **Variation in $X$**: There is variation in the explanatory variable.
OLS Assumptions Preview

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.

2. **Random Sampling**: The observed data represent a random sample from the population described by the model.

3. **Variation in $X$**: There is variation in the explanatory variable.

4. **Zero conditional mean**: Expected value of the error term is zero conditional on all values of the explanatory variable.
OLS Assumptions Preview

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.

2. **Random Sampling**: The observed data represent a random sample from the population described by the model.

3. **Variation in X**: There is variation in the explanatory variable.

4. **Zero conditional mean**: Expected value of the error term is zero conditional on all values of the explanatory variable.

5. **Homoskedasticity**: The error term has the same variance conditional on all values of the explanatory variable.
OLS Assumptions Preview

1. **Linearity in Parameters:** The population model is linear in its parameters and correctly specified.

2. **Random Sampling:** The observed data represent a random sample from the population described by the model.

3. **Variation in $X$:** There is variation in the explanatory variable.

4. **Zero conditional mean:** Expected value of the error term is zero conditional on all values of the explanatory variable.

5. **Homoskedasticity:** The error term has the same variance conditional on all values of the explanatory variable.

6. **Normality:** The error term is independent of the explanatory variables and normally distributed.
Hierarchy of OLS Assumptions
Hierarchy of OLS Assumptions

- **Identification**
  - Data Description
    - Variation in X

- **Unbiasedness**
  - Consistency
    - Variation in X
    - Random Sampling
    - Linearity in Parameters
    - Zero Conditional Mean

- **Gauss-Markov (BLUE)**
  - Asymptotic Inference (z and $\chi^2$)
    - Variation in X
    - Random Sampling
    - Linearity in Parameters
    - Zero Conditional Mean
    - Homoskedasticity

- **Classical LM (BUE)**
  - Small-Sample Inference (t and F)
    - Variation in X
    - Random Sampling
    - Linearity in Parameters
    - Zero Conditional Mean
    - Homoskedasticity
    - Normality of Errors
OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X + u \]

Note that it can be nonlinear in variables

\[ Y = \beta_0 + \beta_1 X + u \] or
\[ Y = \beta_0 + \beta_1 X^2 + u \] or
\[ Y = \beta_0 + \beta_1 \log(X) + u \]

\[ \beta_0, \beta_1: \text{Population parameters — fixed and unknown} \]
\[ u: \text{Unobserved random variable with } E[u] = 0 — \text{captures all other factors influencing } Y \text{ other than } X \]

We assume this to be the structural model, i.e., the model describing the true process generating \( Y \).
Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]
Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

- Note that it can be nonlinear in variables
Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

- Note that it can be nonlinear in variables
  - OK: \( Y = \beta_0 + \beta_1 X + u \) or
    \[ Y = \beta_0 + \beta_1 X^2 + u \] or
    \[ Y = \beta_0 + \beta_1 \log(X) + u \]

\( \beta_0, \beta_1 \): Population parameters — fixed and unknown

\( u \): Unobserved random variable with \( E[u] = 0 \) — captures all other factors influencing \( Y \) other than \( X \)

We assume this to be the structural model, i.e., the model describing the true process generating \( Y \).
Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

- Note that it can be nonlinear in variables
  - OK: \( Y = \beta_0 + \beta_1 X + u \) or \( Y = \beta_0 + \beta_1 X^2 + u \) or \( Y = \beta_0 + \beta_1 \log(X) + u \)
  - Not OK: \( Y = \beta_0 + \beta_1^2 X + u \) or \( Y = \beta_0 + \exp(\beta_1)X + u \)
OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

- Note that it can be nonlinear in variables
  - OK: \( Y = \beta_0 + \beta_1 X + u \) or
    \[ Y = \beta_0 + \beta_1 X^2 + u \] or
    \[ Y = \beta_0 + \beta_1 \log(X) + u \]
  - Not OK: \( Y = \beta_0 + \beta_1^2 X + u \) or
    \[ Y = \beta_0 + \exp(\beta_1) X + u \]

- \( \beta_0, \beta_1 \): Population parameters — fixed and unknown
OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

- Note that it can be nonlinear in variables
  - OK: \( Y = \beta_0 + \beta_1 X + u \) or \( Y = \beta_0 + \beta_1 X^2 + u \) or \( Y = \beta_0 + \beta_1 \log(X) + u \)
  - Not OK: \( Y = \beta_0 + \beta_1^2 X + u \) or \( Y = \beta_0 + \exp(\beta_1) X + u \)

- \( \beta_0, \beta_1 \): Population parameters — fixed and unknown
- \( u \): Unobserved random variable with \( E[u] = 0 \) — captures all other factors influencing \( Y \) other than \( X \)
OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

\[ Y = \beta_0 + \beta_1 X_1 + u \]

Note that it can be nonlinear in variables

- OK: \[ Y = \beta_0 + \beta_1 X + u \] or \[ Y = \beta_0 + \beta_1 X^2 + u \] or \[ Y = \beta_0 + \beta_1 \log(X) + u \]
- Not OK: \[ Y = \beta_0 + \beta_1^2 X + u \] or \[ Y = \beta_0 + \exp(\beta_1) X + u \]

- \( \beta_0, \beta_1 \): Population parameters — fixed and unknown
- \( u \): Unobserved random variable with \( E[u] = 0 \) — captures all other factors influencing \( Y \) other than \( X \)

We assume this to be the structural model, i.e., the model describing the true process generating \( Y \)
OLS Assumption II

Assumption (II. Random Sampling)

The observed data: \((y_i, x_i)\) for \(i = 1, \ldots, n\) represent an i.i.d. random sample of size \(n\) following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)
Assumption (II. Random Sampling)

The observed data:

\[(y_i, \ x_i) \text{ for } i = 1, \ldots, n\]

represent an i.i.d. random sample of size n following the population model.
Assumption (II. Random Sampling)

*The observed data:*

\[(y_i, x_i) \text{ for } i = 1, \ldots, n\]

*represent an i.i.d. random sample of size n following the population model.*

Data examples consistent with this assumption:
OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

\[(y_i, x_i) \text{ for } i = 1, ..., n\]

represent an i.i.d. random sample of size \(n\) following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly
OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

\[(y_i, \ x_i) \text{ for } i = 1, \ldots, n\]

represent an i.i.d. random sample of size \(n\) following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)
Assumption (II. Random Sampling)

The observed data:

\[(y_i, \ x_i) \text{ for } i = 1, \ldots, n\]

represent an i.i.d. random sample of size \(n\) following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
Assumption (II. Random Sampling)

The observed data:

\[(y_i, x_i) \text{ for } i = 1, \ldots, n\]

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)
OLS Assumption III

Assumption (III. Variation in $X$; a.k.a. No Perfect Collinearity)

The observed data: $x_i$ for $i = 1,...,n$ are not all the same value.

Satisfied as long as there is some variation in the regressor $X$ in the sample.

Why do we need this?

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

This assumption is needed just to calculate $\hat{\beta}$, i.e. identifying $\hat{\beta}$.

In fact, this is the only assumption needed for using OLS as a pure data summary.
**OLS Assumption III**

**Assumption (III. Variation in \( X \); a.k.a. No Perfect Collinearity)**

The observed data:

\[ x_i \quad \text{for} \quad i = 1, \ldots, n \]

are not all the same value.
### OLS Assumption III

**Assumption (III. Variation in \( X \); a.k.a. No Perfect Collinearity)**

*The observed data:*

\[ x_i \text{ for } i = 1, \ldots, n \]

*are not all the same value.*

Satisfied as long as there is some variation in the regressor \( X \) in the sample.

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

This assumption is needed just to calculate \( \hat{\beta} \), i.e., identifying \( \hat{\beta} \).

In fact, this is the only assumption needed for using OLS as a pure data summary.
OLS Assumption III

Assumption (III. Variation in $X$; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor $X$ in the sample.

Why do we need this?

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
$$
OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

\[ x_i \text{ for } i = 1, \ldots, n \]

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

Why do we need this?

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

This assumption is needed just to calculate \( \hat{\beta} \), i.e. identifying \( \hat{\beta} \).
OLS Assumption III

Assumption (III. Variation in $X$; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor $X$ in the sample.

Why do we need this?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

This assumption is needed just to calculate $\hat{\beta}$, i.e. identifying $\hat{\beta}$.

In fact, this is the only assumption needed for using OLS as a pure data summary.
Stuck in a moment

- Why does this matter?
Stuck in a moment

- Why does this matter?
Why does this matter? How would you draw the line of best fit through this scatterplot, which is a violation of this assumption?
OLS Assumption IV

Assumption (IV. Zero Conditional Mean)
The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u|X] = 0$$

Cov($X, u) = 0$ implies a slightly weaker condition

Given random sampling, $E[u|X] = 0$ also implies $E[u_i|x_i] = 0$ for all $i$

Violations:
Recall that $u$ represents all unobserved factors that influence $Y$. If such unobserved factors are also correlated with $X$, $Cov(X, u) \neq 0$

Example:
Wage = $\beta_0 + \beta_1$education + $u$. What is likely to be in $u$?

$\rightarrow$ It must be assumed $E[ability|educ=low] = E[ability|educ=high]$.
Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u|X] = 0$$
Assumption (IV. Zero Conditional Mean)

*The expected value of the error term is zero conditional on any value of the explanatory variable:*

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)

Examples:

Wage = \( \beta_0 + \beta_1 \text{education} + u \). What is likely to be in \( u \)?

- It must be assumed \( E[\text{ability}|\text{educ} = \text{low}] = E[\text{ability}|\text{educ} = \text{high}] \)
OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)
- Given random sampling, \( E[u|X] = 0 \) also implies \( E[u_i|x_i] = 0 \) for all \( i \)

Violations:

Recall that \( u \) represents all unobserved factors that influence \( Y \).

If such unobserved factors are also correlated with \( X \), \( \text{Cov}(X, u) \neq 0 \).

Example: Wage = \( \beta_0 + \beta_1 \text{education} + u \).

It must be assumed \( E[\text{ability}|\text{education}=\text{low}] = E[\text{ability}|\text{education}=\text{high}] \).
Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u|X] = 0$$

- $E[u|X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$
- Given random sampling, $E[u|X] = 0$ also implies $E[u_i|x_i] = 0$ for all $i$
OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

*The expected value of the error term is zero conditional on any value of the explanatory variable:*

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)
- Given random sampling, \( E[u|X] = 0 \) also implies \( E[u_i|x_i] = 0 \) for all \( i \)

Violations:

Recall that \( u \) represents all unobserved factors that influence \( Y \). If such unobserved factors are also correlated with \( X \), \( \text{Cov}(X, u) \neq 0 \).

Example: Wage = \( \beta_0 + \beta_1 \text{education} + u \). What is likely to be in \( u \)?

→ It must be assumed \( E[\text{ability}|\text{educ}=\text{low}] = E[\text{ability}|\text{educ}=\text{high}] \).
OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)
- Given random sampling, \( E[u|X] = 0 \) also implies \( E[u_i|x_i] = 0 \) for all \( i \)

Violations:
- Recall that \( u \) represents all unobserved factors that influence \( Y \)
Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u|X] = 0$$

- $E[u|X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$
- Given random sampling, $E[u|X] = 0$ also implies $E[u_i|x_i] = 0$ for all $i$

Violations:

- Recall that $u$ represents all unobserved factors that influence $Y$
- If such unobserved factors are also correlated with $X$, $\text{Cov}(X, u) \neq 0$
Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)
- Given random sampling, \( E[u|X] = 0 \) also implies \( E[u_i|x_i] = 0 \) for all \( i \)

Violations:

- Recall that \( u \) represents all unobserved factors that influence \( Y \)
- If such unobserved factors are also correlated with \( X \), \( \text{Cov}(X, u) \neq 0 \)
- Example: \( \text{Wage} = \beta_0 + \beta_1 \text{education} + u \). What is likely to be in \( u \)?
OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

\[ E[u|X] = 0 \]

- \( E[u|X] = 0 \) implies a slightly weaker condition \( \text{Cov}(X, u) = 0 \)
- Given random sampling, \( E[u|X] = 0 \) also implies \( E[u_i|x_i] = 0 \) for all \( i \)

Violations:

- Recall that \( u \) represents all unobserved factors that influence \( Y \)
- If such unobserved factors are also correlated with \( X \), \( \text{Cov}(X, u) \neq 0 \)

Example: \( \text{Wage} = \beta_0 + \beta_1 \text{education} + u \). What is likely to be in \( u \)?

\[ \rightarrow \text{It must be assumed } E[\text{ability} | \text{educ} = \text{low}] = E[\text{ability} | \text{educ} = \text{high}] \]
Violating the zero conditional mean assumption

How does this assumption get violated? Let's generate data from the following model:

\[ Y_i = 1 + 0.5X_i + u_i. \]

But let's compare two situations:

1. Where the mean of \( u_i \) depends on \( X_i \) (they are correlated)
2. No relationship between them (satisfies the assumption)
How does this assumption get violated? Let’s generate data from the following model:

$$Y_i = 1 + 0.5X_i + u_i$$
Violating the zero conditional mean assumption

How does this assumption get violated? Let’s generate data from the following model:

\[ Y_i = 1 + 0.5X_i + u_i \]

But let’s compare two situations:

1. Where the mean of \( u_i \) depends on \( X_i \) (they are correlated)
Violating the zero conditional mean assumption

How does this assumption get violated? Let’s generate data from the following model:

\[ Y_i = 1 + 0.5X_i + u_i \]

But let’s compare two situations:

1. Where the mean of \( u_i \) depends on \( X_i \) (they are correlated)
2. No relationship between them (satisfies the assumption)
Violating the zero conditional mean assumption

Assumption 4 violated

Assumption 4 not violated
Unbiasedness (to the blackboard)

With Assumptions 1-4, we can show that the OLS estimator for the slope is unbiased, that is $E[\hat{\beta}_1] = \beta_1$. 
Unbiasedness (to the blackboard)

With Assumptions 1-4, we can show that the OLS estimator for the slope is unbiased, that is $E[\hat{\beta}_1] = \beta_1$. 

TO THE BLACKBOARD!
Unbiasedness of OLS

Theorem (Unbiasedness of OLS)

Given OLS Assumptions I–IV:

\[ E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1 \]

The sampling distributions of the estimators \( \hat{\beta}_1 \) and \( \hat{\beta}_0 \) are centered about the true population parameter values \( \beta_1 \) and \( \beta_0 \).
Where are we?
Now we know that, under Assumptions 1-4, we know that

\[ \hat{\beta}_1 \sim ?(\beta_1, ?) \]
Now we know that, under Assumptions 1-4, we know that

\[ \hat{\beta}_1 \sim \? (\beta_1, \?) \]

That is we know that the sampling distribution is centered on the true population slope, but we don’t know the population variance.
Sampling variance of estimated slope
Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,
Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,
  1. Linearity
Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
Sampling variance of estimated slope

In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
4. Zero conditional mean of the errors
In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
4. Zero conditional mean of the errors
5. Homoskedasticity
Variance of OLS Estimators

How can we derive \( \text{Var}[ \hat{\beta}_0 ] \) and \( \text{Var}[ \hat{\beta}_1 ] \)? Let's make the following additional assumption:

**Assumption (V. Homoskedasticity)**

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

\[
\text{Var}[u | X] = \sigma_u^2
\]

This implies \( \text{Var}[u] = \sigma_u^2 \rightarrow \) all errors have an identical error variance (\( \sigma_u^2 \) for all \( i \))

Taken together, Assumptions I–V imply:

\[
E[Y | X] = \beta_0 + \beta_1 X
\]

\[
\text{Var}[Y | X] = \sigma_u^2
\]

Violation: \( \text{Var}[u | X = x_1] \neq \text{Var}[u | X = x_2] \) called heteroskedasticity.

Assumptions I–V are collectively known as the Gauss-Markov assumptions.
Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let’s make the following additional assumption:

Assumption (V. Homoskedasticity)
The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$\text{Var}[u|X] = \sigma^2$$

This implies $\text{Var}[u] = \sigma^2$ → all errors have an identical error variance ($\sigma^2_u = \sigma^2$ for all $i$)

Taken together, Assumptions I–V imply:

$$\text{E}[Y|X] = \beta_0 + \beta_1 X$$

$$\text{Var}[Y|X] = \sigma^2_u$$

Violation: $\text{Var}[u|X=x_1] \neq \text{Var}[u|X=x_2]$ called heteroskedasticity.

Assumptions I–V are collectively known as the Gauss-Markov assumptions.
Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let’s make the following additional assumption:

**Assumption (V. Homoskedasticity)**

*The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:*

$$\text{Var}[u|X] = \sigma_u^2$$
Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let’s make the following additional assumption:

**Assumption (V. Homoskedasticity)**

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$\text{Var}[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  - all errors have an identical error variance ($\sigma_u^2 = \sigma_u^2$ for all $i$)
Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let’s make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$\text{Var}[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  $\rightarrow$ all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all $i$)

- Taken together, Assumptions I–V imply:

$$E[Y|X] = \beta_0 + \beta_1 X$$

$$\text{Var}[Y|X] = \sigma_u^2$$
Variance of OLS Estimators

How can we derive \( \text{Var}[\hat{\beta}_0] \) and \( \text{Var}[\hat{\beta}_1] \)? Let’s make the following additional assumption:

### Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

\[
\text{Var}[u|X] = \sigma_u^2
\]

- This implies \( \text{Var}[u] = \sigma_u^2 \)
  
  \( \rightarrow \) all errors have an identical error variance \( (\sigma_{u_i}^2 = \sigma_u^2 \text{ for all } i) \)

- Taken together, Assumptions I–V imply:

  \[
  E[Y|X] = \beta_0 + \beta_1 X \\
  \text{Var}[Y|X] = \sigma_u^2
  \]

- Violation: \( \text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2] \) called heteroskedasticity.
Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let’s make the following additional assumption:

**Assumption (V. Homoskedasticity)**

*The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:*

$$\text{Var}[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  - all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all $i$)
- Taken together, Assumptions I–V imply:
  $$E[Y|X] = \beta_0 + \beta_1X$$
  $$\text{Var}[Y|X] = \sigma_u^2$$
- Violation: $\text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2]$ called heteroskedasticity.
- Assumptions I–V are collectively known as the Gauss-Markov assumptions
Deriving the sampling variance

\[
\text{var}[\hat{\beta}_1 | X_1, \ldots, X_n] = ??
\]
Deriving the sampling variance

$$\text{var}[\hat{\beta}_1 | X_1, \ldots, X_n] = ??$$

TO THE BLACKBOARD!
Theorem (Variance of OLS Estimators)

Given OLS Assumptions I–V (Gauss-Markov Assumptions):

\[
\text{Var}[\hat{\beta}_1 \mid X] = \frac{\sigma_u^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma_u^2}{\text{SST}_x}
\]

\[
\text{Var}[\hat{\beta}_0 \mid X] = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}
\]

where \(\text{Var}[u \mid X] = \sigma_u^2\) (the error variance).
Understanding the sampling variance

\[
\text{var}[\hat{\beta}_1|X_1, \ldots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\]

- What drives the sampling variability of the OLS estimator?
Understanding the sampling variance

\[ \text{var}[\hat{\beta}_1 | X_1, \ldots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of \( Y_i \), the higher the sampling variance
  - The lower the variance of \( X_i \), the higher the sampling variance
  - As we increase \( n \), the denominator gets large, while the numerator is fixed and so the sampling variance shrinks to 0.
Understanding the sampling variance

\[ \text{var}[\hat{\beta}_1|X_1, \ldots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of \( Y_i \), the higher the sampling variance
  - The lower the variance of \( X_i \), the higher the sampling variance
Understanding the sampling variance

\[ \text{var}[\hat{\beta}_1|X_1, \ldots, X_n] = \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of \( Y_i \), the higher the sampling variance
  - The lower the variance of \( X_i \), the higher the sampling variance
  - As we increase \( n \), the denominator gets large, while the numerator is fixed and so the sampling variance shrinks to 0.
Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $\text{Var}[u] = \sigma_u^2$?
Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $\text{Var}[u] = \sigma_u^2$? We can derive an estimator based on the residuals:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors $u_i$ are NOT the same as the residuals $\hat{u}_i$. Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line. We can measure scatter with the mean squared deviation:

$$\text{MSD}(\hat{u}) \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$$

Intuitively, which line is likely to be closer to the observed sample values on $X$ and $Y$, the true line $y_i = \beta_0 + \beta_1 x_i$ or the fitted regression line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$?
Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance \( \text{Var}[u] = \sigma_u^2 \)? We can derive an estimator based on the residuals:

\[
\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i
\]

Recall: The errors \( u_i \) are NOT the same as the residuals \( \hat{u}_i \).

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

\[
\text{MSD}(\hat{u}) \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - \bar{\hat{u}})^2 =
\]
Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $\text{Var}[u] = \sigma^2_u$? We can derive an estimator based on the residuals:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors $u_i$ are NOT the same as the residuals $\hat{u}_i$.

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

$$\text{MSD}(\hat{u}) \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}^2_i$$
Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $\text{Var}[u] = \sigma_u^2$? We can derive an estimator based on the residuals:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors $u_i$ are NOT the same as the residuals $\hat{u}_i$.

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

$$\text{MSD}(\hat{u}) \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$$

Intuitively, which line is likely to be closer to the observed sample values on $X$ and $Y$, the true line $y_i = \beta_0 + \beta_1 x_i$ or the fitted regression line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$?
Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the actual sample we have
Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the actual sample we have
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations

\[
\text{MSD}(\hat{u}) \text{ will slightly underestimate the error variance } \text{Var}[u] = \sigma^2_u \text{ on average.}
\]

In fact, we can show that with a single regressor \( X \) we have:

\[
E[\text{MSD}(\hat{u})] = \frac{n - 2}{n} \sigma^2_u \quad \text{(degrees of freedom adjustment)}
\]

Thus, an unbiased estimator for the error variance is:

\[
\hat{\sigma}^2_u = \frac{n}{n - 2} \text{MSD}(\hat{u}) = \frac{n}{n - 2} \sum_{i=1}^{n} \hat{u}_i^2
\]

We plug this estimate into the variance estimators for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \).
By construction, the regression line is closer since it is drawn to fit the actual sample we have.

Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations.

So the spread of the residuals $MSD(\hat{u})$ will slightly underestimate the error variance $\text{Var}[u] = \sigma^2_u$ on average.
Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the actual sample we have.

- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations.

- So the spread of the residuals $MSD(\hat{u})$ will slightly underestimate the error variance $\text{Var}[u] = \sigma_u^2$ on average.

- In fact, we can show that with a single regressor $X$ we have:

$$E[MSD(\hat{u})] = \frac{n - 2}{n} \sigma_u^2$$ (degrees of freedom adjustment)
Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the actual sample we have
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations
- So the spread of the residuals $MSD(\hat{u})$ will slightly underestimate the error variance $\text{Var}[u] = \sigma_u^2$ on average
- In fact, we can show that with a single regressor $X$ we have:

$$E[MSD(\hat{u})] = \frac{n-2}{n} \sigma_u^2 \text{ (degrees of freedom adjustment)}$$

- Thus, an unbiased estimator for the error variance is:

$$\hat{\sigma}_u^2 = \frac{n}{n-2} MSD(\hat{u}) = \frac{n}{n-2} \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$$

We plug this estimate into the variance estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$. 
Under Assumptions 1-5, we know that

\[ \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \]
Where are we?

- Under Assumptions 1-5, we know that
  \[ \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \]

- Now we know the mean and sampling variance of the sampling distribution.
Under Assumptions 1-5, we know that

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

Now we know the mean and sampling variance of the sampling distribution.

Next Time: how does this compare to other estimators for the population slope?
Where We’ve Been and Where We’re Going...

Last Week
▶ hypothesis testing
▶ what is regression

This Week
▶ Monday:
   ⋆ mechanics of OLS
   ⋆ properties of OLS
▶ Wednesday:
   ⋆ hypothesis tests for regression
   ⋆ confidence intervals for regression
   ⋆ goodness of fit

Next Week
▶ mechanics with two regressors
▶ omitted variables, multicollinearity

Long Run
▶ probability
→ inference
→ regression

Questions?
Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

- Long Run
  - probability
  - → inference
  - → regression

Questions?

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

- Long Run
  - probability → inference → regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS

- Wednesday:
  - hypothesis tests for regression
  - confidence intervals for regression
  - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

Long Run
- probability
  → inference
  → regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- **Next Week**
  - mechanics with two regressors
  - omitted variables, multicollinearity

- **Long Run**
  - probability → inference → regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression

Next Week
- mechanics with two regressors
- omitted variables, multicollinearity

Long Run
- probability
  - inference
  - regression

Questions?

Stewart (Princeton)
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
Where We’ve Been and Where We’re Going...

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

Next Week
- mechanics with two regressors
- omitted variables, multicollinearity

Long Run
- probability
  → inference
  → regression

Questions?

Stewart (Princeton)

Week 5: Simple Linear Regression

October 10, 12, 2016
Where We’ve Been and Where We’re Going…

- **Last Week**
  - hypothesis testing
  - what is regression

- **This Week**
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- **Next Week**
  - mechanics with two regressors
  - omitted variables, multicollinearity
Where We’ve Been and Where We’re Going...

- Last Week
  - hypothesis testing
  - what is regression

- This Week
  - Monday:
    - mechanics of OLS
    - properties of OLS
  - Wednesday:
    - hypothesis tests for regression
    - confidence intervals for regression
    - goodness of fit

- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

- Long Run
  - probability → inference → regression

Questions?
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Example: Epstein and Mershon SCOTUS data
Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
- Percentage of votes in liberal direction for each justice in a number of issue areas
Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
- Percentage of votes in liberal direction for each justice in a number of issue areas
- Segal-Cover scores for each justice
Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
- Percentage of votes in liberal direction for each justice in a number of issue areas
- Segal-Cover scores for each justice
- Party of appointing president
$y = \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 x$
\[ y = 27.6 + 41.2x \]
$y = 27.6 + 41.2x + u$
How to get $\beta_0$ and $\beta_1$

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \]

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \]
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Week 5: Simple Linear Regression

1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Where are we?
Where are we?

- **Identification**
  - Data Description
  - Variation in X

- **Unbiasedness**
  - Consistency
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean

- **Gauss-Markov (BLUE)**
  - Asymptotic Inference (z and $\chi^2$)
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity

- **Classical LM (BUE)**
  - Small-Sample Inference (t and F)
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
  - Normality of Errors
Under Assumptions 1-5, we know that

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

Now we know the mean and sampling variance of the sampling distribution.

How does this compare to other estimators for the population slope?
Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the

1. Best: Lowest variance in class
2. Linear: Among Linear estimators
3. Unbiased: Among Linear Unbiased estimators
4. Estimator.

Assumptions 1-5: the “Gauss Markov Assumptions”

The proof is detailed and doesn’t yield insight, so we skip it. (We will explore the intuition some more in a few slides)

Fails to hold when the assumptions are violated!
### Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is **BLUE**, i.e. the

1. **Best**: Lowest variance in class
2. **Linear**: Among Linear estimators
3. **Unbiased**: Among Linear Unbiased estimators
4. **Estimator**.

Fails to hold when the assumptions are violated!
OLS is BLUE :

**Theorem (Gauss-Markov)**

*Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the*

1. **Best:** Lowest variance in class
2. **Linear:** Among Linear estimators
3. **Unbiased:** Among Linear Unbiased estimators
4. **Estimator.**

- Assumptions 1-5: the “Gauss Markov Assumptions”
Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is **BLUE**, i.e. the

1. **Best**: Lowest variance in class
2. **Linear**: Among Linear estimators
3. **Unbiased**: Among Linear Unbiased estimators
4. **Estimator**.

- Assumptions 1-5: the “Gauss Markov Assumptions”
- The proof is detailed and doesn’t yield insight, so we skip it. (We will explore the intuition some more in a few slides)
Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is **BLUE**, i.e. the

1. **Best**: Lowest variance in class
2. **Linear**: Among Linear estimators
3. **Unbiased**: Among Linear Unbiased estimators
4. **Estimator**.

- Assumptions 1-5: the “Gauss Markov Assumptions”
- The proof is detailed and doesn’t yield insight, so we skip it. (We will explore the intuition some more in a few slides)
- Fails to hold when the assumptions are violated!
Gauss-Markov Theorem

All estimators

unbiased

linear

OLS is efficient in the class of unbiased, linear estimators.

OLS is BLUE—best linear unbiased estimator.
Where are we?

- Under Assumptions 1-5, we know that

\[ \hat{\beta}_1 \sim \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \]
Where are we?

- Under Assumptions 1-5, we know that

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

- And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \)
Where are we?

- Under Assumptions 1-5, we know that
  \[ \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \]

- And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \)

- What about the last question mark? What’s the form of the distribution?
Where are we?

- Under Assumptions 1-5, we know that
  \[ \hat{\beta}_1 \sim \text{?} \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \]

- And we know that \[ \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \] is the lowest variance of any linear estimator of \( \beta_1 \)

- What about the last question mark? What’s the form of the distribution?
Under Assumptions 1-5, we know that

\[ \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \]

And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \).

What about the last question mark? What’s the form of the distribution? Uniform?
Where are we?

- Under Assumptions 1-5, we know that

\[ \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \]

- And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \)

- What about the last question mark? What’s the form of the distribution? Uniform? \( t \)?
Where are we?

Under Assumptions 1-5, we know that

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

And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \)

What about the last question mark? What’s the form of the distribution? Uniform? t? Normal?
Where are we?

- Under Assumptions 1-5, we know that
  \[
  \hat{\beta}_1 \sim ? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)
  \]

- And we know that \(\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\) is the lowest variance of any linear estimator of \(\beta_1\)

- What about the last question mark? What’s the form of the distribution? Uniform? \(t\)? Normal? Exponential?
Where are we?

- Under Assumptions 1-5, we know that

  \[ \hat{\beta}_1 \sim? \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \]

- And we know that \( \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \) is the lowest variance of any linear estimator of \( \beta_1 \)

- What about the last question mark? What’s the form of the distribution? Uniform? \( t \)? Normal? Exponential? Hypergeometric?
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.’s:

\[
\hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i
\]
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.’s:
  \[
  \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i
  \]

- Mantra of the Central Limit Theorem:
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

\[ \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i \]

- Mantra of the Central Limit Theorem:
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.’s:

\[
\hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i
\]

- Mantra of the Central Limit Theorem:

"the sums and means of r.v.’s tend to be Normally distributed in large samples."
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.’s:
  \[ \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i \]

- Mantra of the Central Limit Theorem:
  “the sums and means of r.v.’s tend to be Normally distributed in large samples.”

- True here as well, so we know that in large samples:
  \[ \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1) \]
Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.’s:
  \[
  \hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i
  \]

- Mantra of the Central Limit Theorem:
  "the sums and means of r.v.’s tend to be Normally distributed in large samples."

- True here as well, so we know that in large samples:
  \[
  \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)
  \]

- Can also replace \( SE \) with an estimate:
  \[
  \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}[\hat{\beta}_1]} \sim N(0, 1)
  \]
Where are we?

Under Assumptions 1-5 and in large samples, we know that

\[ \hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \]
Where are we?

Under Assumptions 1-5 and in large samples, we know that

$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right)$$
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:
  1. Linearity
  2. Random (iid) sample
  3. Variation in $X_i$
  4. Zero conditional mean of the errors
  5. Homoskedasticity
  6. Errors are conditionally Normal
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:

1. Linearity
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:

  1. Linearity
  2. Random (iid) sample
  3. Variation in $X_i$
  4. Zero conditional mean of the errors
Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can’t get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
4. Zero conditional mean of the errors
5. Homoskedasticity
Sampling distribution in small samples

What if we have a small sample? What can we do then?  
Can’t get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in $X_i$
4. Zero conditional mean of the errors
5. Homoskedasticity
6. Errors are conditionally Normal
OLS Assumptions VI

Assumption (VI. Normality)
The population error term is independent of the explanatory variable, \( u \perp X \), and is normally distributed with mean zero and variance \( \sigma^2_u \):

\[ u \sim N(0, \sigma^2_u) \]

which implies

\[ Y | X \sim N(\beta_0 + \beta_1 X, \sigma^2_u) \]

Note: This implies homoskedasticity and zero conditional mean.

Together Assumptions I–VI are the classical linear model (CLM) assumptions.

The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)

Non-normality of the errors is a serious concern in small samples. We can partially check this assumption by looking at the residuals

Variable transformations can help to come closer to normality

We don't need normality assumption in large samples
OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp X$, and is normally distributed with mean zero and variance $\sigma_u^2$:

$$ u \sim N(0, \sigma_u^2), \quad \text{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2) $$

Note: This implies homoskedasticity and zero conditional mean.
OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, \( u \perp X \), and is normally distributed with mean zero and variance \( \sigma_u^2 \):

\[
u \sim N(0, \sigma_u^2), \quad \text{which implies} \quad Y|X \sim N(\beta_0 + \beta_1X, \sigma_u^2)
\]

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
Assumption (VI. Normality)

The population error term is independent of the explanatory variable, \( u \perp X \), and is normally distributed with mean zero and variance \( \sigma^2_u \):

\[
u \sim N(0, \sigma^2_u), \quad \text{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2_u)
\]

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators).
Assumption (VI. Normality)

The population error term is independent of the explanatory variable, \( u \perp X \), and is normally distributed with mean zero and variance \( \sigma_u^2 \):

\[
u \sim N(0, \sigma_u^2), \quad \text{which implies} \quad Y | X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)
\]

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can partially check this assumption by looking at the residuals

Variable transformations can help to come closer to normality.
OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp X$, and is normally distributed with mean zero and variance $\sigma^2_u$:

$$u \sim N(0, \sigma^2_u), \text{ which implies } Y\mid X \sim N(\beta_0 + \beta_1X, \sigma^2_u)$$

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can partially check this assumption by looking at the residuals
- Variable transformations can help to come closer to normality
Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp X$, and is normally distributed with mean zero and variance $\sigma^2_u$:

$$u \sim N(0, \sigma^2_u), \quad \text{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2_u)$$

Note: This implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can partially check this assumption by looking at the residuals
- Variable transformations can help to come closer to normality
- We don’t need normality assumption in large samples
Sampling Distribution for $\hat{\beta}_1$

Theorem (Sampling Distribution of $\hat{\beta}_1$)

Under Assumptions I–VI, $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \text{Var}[\hat{\beta}_1|X])$

where $\text{Var}[\hat{\beta}_1|X] = \sigma^2_u \sum_{i=1}^{n} (x_i - \bar{x})^2$

which implies

$\hat{\beta}_1 - \beta_1 \sqrt{\text{Var}[\hat{\beta}_1|X]} = \hat{\beta}_1 - \beta_1 \text{SE}(\hat{\beta}_1) \sim \mathcal{N}(0, 1)$

Proof.

Given Assumptions I–VI, $\hat{\beta}_1$ is a linear combination of the i.i.d. normal random variables:

$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} (x_i - \bar{x}) S_{StT} x_i u_i$

where $u_i \sim \mathcal{N}(0, \sigma^2_u)$.

Any linear combination of independent normals is normal, and we can transform/standarize any normal random variable into a standard normal by subtracting off its mean and dividing by its standard deviation.
Sampling Distribution for $ \hat{\beta}_1 $

**Theorem (Sampling Distribution of $ \hat{\beta}_1 $)**

**Under Assumptions I–VI,**

$$ \hat{\beta}_1 \sim N \left( \beta_1, \text{Var}[\hat{\beta}_1 | X] \right) $$

**where**

$$ \text{Var}[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} $$

**which implies**

$$ \hat{\beta}_1 - \beta_1 \quad \sqrt{\text{Var}[\hat{\beta}_1 | X]} = \hat{\beta}_1 - \beta_1 \quad \text{SE}(\hat{\beta}_1) \sim N(0,1) $$
Theorem (Sampling Distribution of $\hat{\beta}_1$)

Under Assumptions I–VI,

$$\hat{\beta}_1 \sim N(\beta_1, \text{Var}[\hat{\beta}_1 | X])$$

where

$$\text{Var}[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

which implies

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1 | X]}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta})} \sim N(0, 1)$$
Sampling Distribution for \( \hat{\beta}_1 \)

**Theorem (Sampling Distribution of \( \hat{\beta}_1 \))**

Under Assumptions I–VI,

\[
\hat{\beta}_1 \sim N \left( \beta_1, \text{Var}[\hat{\beta}_1 \mid X] \right)
\]

where

\[
\text{Var}[\hat{\beta}_1 \mid X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\]

which implies

\[
\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1 \mid X]}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta})} \sim N(0, 1)
\]

**Proof.**
Sampling Distribution for $\hat{\beta}_1$

**Theorem (Sampling Distribution of $\hat{\beta}_1$)**

Under Assumptions I–VI,

$$\hat{\beta}_1 \sim N\left(\beta_1, \text{Var}[\hat{\beta}_1 | X]\right)$$

where

$$\text{Var}[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

which implies

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1 | X]}} = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta})} \sim N(0, 1)$$

**Proof.**

Given Assumptions I–VI, $\hat{\beta}_1$ is a linear combination of the i.i.d. normal random variables:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\text{SST}_x} u_i \quad \text{where} \quad u_i \sim N(0, \sigma_u^2).$$

Any linear combination of independent normals is...
Sampling Distribution for $\hat{\beta}_1$

**Theorem (Sampling Distribution of $\hat{\beta}_1$)**

Under Assumptions I–VI,

$$\hat{\beta}_1 \sim N\left(\beta_1, \text{Var}[\hat{\beta}_1 | X]\right)$$

where

$$\text{Var}[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

which implies

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1 | X]}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta})} \sim N(0, 1)$$

**Proof.**

Given Assumptions I–VI, $\hat{\beta}_1$ is a linear combination of the i.i.d. normal random variables:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{\text{SST}_x} u_i \quad \text{where} \quad u_i \sim N(0, \sigma_u^2).$$

Any linear combination of independent normals is normal,
Sampling Distribution for $\hat{\beta}_1$

**Theorem (Sampling Distribution of $\hat{\beta}_1$)**

*Under Assumptions I–VI,*

$$\hat{\beta}_1 \sim N \left( \beta_1, \text{Var}[\hat{\beta}_1 | X] \right)$$

where

$$\text{Var}[\hat{\beta}_1 | X] = \frac{\sigma^2_u}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

which implies

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1 | X]}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta})} \sim N(0, 1)$$

**Proof.**

Given Assumptions I–VI, $\hat{\beta}_1$ is a linear combination of the i.i.d. normal random variables:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{SST_x} u_i \quad \text{where} \quad u_i \sim N(0, \sigma^2_u).$$

Any linear combination of independent normals is normal, and we can transform/standarize any normal random variable into a standard normal by subtracting off its mean and dividing by its standard deviation.
Sampling distribution of OLS slope

- If we have $Y_i$ given $X_i$ is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}[\hat{\beta}_1]} \sim t_{n-2}$$

The standardized coefficient follows a $t$ distribution $n-2$ degrees of freedom. We take off an extra degree of freedom because we had to one more parameter than just the sample mean.

All of this depends on Normal errors! We can check to see if the error do look Normal.
Sampling distribution of OLS slope

- If we have $Y_i$ given $X_i$ is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}$$
Sampling distribution of OLS slope

- If we have $Y_i$ given $X_i$ is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a $t$ distribution $n - 2$ degrees of freedom. We take off an extra degree of freedom because we had to one more parameter than just the sample mean.
Sampling distribution of OLS slope

- If we have $Y_i$ given $X_i$ is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a $t$ distribution $n - 2$ degrees of freedom. We take off an extra degree of freedom because we had to one more parameter than just the sample mean.

- All of this depends on Normal errors! We can check to see if the error do look Normal.
The t-Test for Single Population Parameters

- \( SE[\hat{\beta}_1] = \frac{\sigma_u}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \) involves the unknown population error variance \( \sigma_u^2 \)

- Replace \( \sigma_u^2 \) with its unbiased estimator \( \hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \), and we obtain:

**Theorem (Sampling Distribution of t-value)**

*Under Assumptions I–VI, the t-value for \( \beta_1 \) has a t-distribution with \( n - 2 \) degrees of freedom:

\[
T \equiv \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim \tau_{n-2}
\]
The t-Test for Single Population Parameters

- \( SE[\hat{\beta}_1] = \frac{\sigma_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \) involves the unknown population error variance \( \sigma_u^2 \)

- Replace \( \sigma_u^2 \) with its unbiased estimator \( \hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \), and we obtain:

### Theorem (Sampling Distribution of t-value)

Under Assumptions I–VI, the t-value for \( \beta_1 \) has a t-distribution with \( n - 2 \) degrees of freedom:

\[
T \equiv \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim \tau_{n-2}
\]

### Proof.

The logic is perfectly analogous to the t-value for the population mean — because we are estimating the denominator, we need a distribution that has fatter tails than \( N(0,1) \) to take into account the additional uncertainty.

This time, \( \hat{\sigma}_u^2 \) contains two estimated parameters (\( \hat{\beta}_0 \) and \( \hat{\beta}_1 \)) instead of one, hence the degrees of freedom \( = n - 2 \).
Where are we?

- Under Assumptions 1-5 and in large samples, we know that

\[ \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \]
Where are we?

- Under Assumptions 1-5 and in large samples, we know that
  \[
  \hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)
  \]

- Under Assumptions 1-6 and in any sample, we know that
  \[
  \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}
  \]
Where are we?

- Under Assumptions 1-5 and in large samples, we know that
  \[ \hat{\beta}_1 \sim N \left( \beta_1, \sigma_u^2 \sum_{i=1}^n (X_i - \bar{X})^2 \right) \]

- Under Assumptions 1-6 and in any sample, we know that
  \[ \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2} \]
Where are we?

- Under Assumptions 1-5 and in large samples, we know that
  \[
  \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)
  \]

- Under Assumptions 1-6 and in any sample, we know that
  \[
  \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}
  \]

Now let’s briefly return to some of the large sample properties.
Large Sample Properties: Consistency

We just looked formally at the small sample properties of the OLS estimator, i.e., how \( \hat{\beta}_0, \hat{\beta}_1 \) behaves in repeated samples of a given \( n \).

Now let's take a more rigorous look at the large sample properties, i.e., how \( \hat{\beta}_0, \hat{\beta}_1 \) behaves when \( n \to \infty \).

Theorem (Consistency of OLS Estimator)

Given Assumptions I–IV, the OLS estimator \( \hat{\beta}_1 \) is consistent for \( \beta_1 \) as \( n \to \infty \):

\[
\operatorname{plim}_{n \to \infty} \hat{\beta}_1 = \beta_1
\]

Technical note: We can slightly relax Assumption IV:

\[
E[ u | X ] = 0 \quad \text{(any function of } X \text{ is uncorrelated with } u)
\]

to its implication:

\[
\operatorname{Cov}[ u, X ] = 0 \quad \text{(} X \text{ is uncorrelated with } u \text{)}
\]

for consistency to hold (but not unbiasedness).
Large Sample Properties: Consistency

- We just looked formally at the small sample properties of the OLS estimator, i.e., how $(\hat{\beta}_0, \hat{\beta}_1)$ behaves in repeated samples of a given $n$. 

Theorem (Consistency of OLS Estimator)

Given Assumptions I–IV, the OLS estimator $\hat{\beta}_1$ is consistent for $\beta_1$ as $n \to \infty$:

$$\lim_{n \to \infty} \hat{\beta}_1 = \beta_1$$

Technical note: We can slightly relax Assumption IV:

$$E[u|X] = 0 \text{ (any function of } X \text{ is uncorrelated with } u)$$

to its implication:

$$\text{Cov}[u, X] = 0 \text{ (} X \text{ is uncorrelated with } u)$$

for consistency to hold (but not unbiasedness).
Large Sample Properties: Consistency

- We just looked formally at the small sample properties of the OLS estimator, i.e., how \( (\hat{\beta}_0, \hat{\beta}_1) \) behaves in repeated samples of a given \( n \).
- Now let’s take a more rigorous look at the large sample properties, i.e., how \( (\hat{\beta}_0, \hat{\beta}_1) \) behaves when \( n \to \infty \).

**Theorem (Consistency of OLS Estimator)**

Given Assumptions I–IV, the OLS estimator \( \hat{\beta}_1 \) is consistent for \( \beta_1 \) as \( n \to \infty \):

\[
\lim_{n \to \infty} \hat{\beta}_1 = \beta_1
\]

Technical note: We can slightly relax Assumption IV:

\[
E[u|X] = 0 \quad \text{(any function of } X \text{ is uncorrelated with } u)
\]

This implies:

\[
\text{Cov}[u, X] = 0 \quad \text{ (} X \text{ is uncorrelated with } u)
\]

for consistency to hold (but not unbiasedness).
Large Sample Properties: Consistency

- We just looked formally at the small sample properties of the OLS estimator, i.e., how \((\hat{\beta}_0, \hat{\beta}_1)\) behaves in repeated samples of a given \(n\).
- Now let’s take a more rigorous look at the large sample properties, i.e., how \((\hat{\beta}_0, \hat{\beta}_1)\) behaves when \(n \to \infty\).

**Theorem (Consistency of OLS Estimator)**

*Given Assumptions I–IV, the OLS estimator \(\hat{\beta}_1\) is consistent for \(\beta_1\) as \(n \to \infty\):*

\[
\operatorname{plim}_{n \to \infty} \hat{\beta}_1 = \beta_1
\]

- Technical note: We can slightly relax Assumption IV:
  
  \[E[u|X] = 0\] (any function of \(X\) is uncorrelated with \(u\))

  to its implication:

  \[\operatorname{Cov}[u, X] = 0\] (\(X\) is uncorrelated with \(u\))

  for consistency to hold (but not unbiasedness).
Large Sample Properties: Consistency

Proof.

Similar to the unbiasedness proof:

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

\[ \text{plim } \hat{\beta}_1 = \text{plim } \beta_1 + \text{plim } \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]  

(Wooldridge C.3 Property i)

\[ = \beta_1 + \frac{\text{plim } \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\text{plim } \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]  

(Wooldridge C.3 Property iii)

\[ = \beta_1 + \frac{\text{Cov}[X, u]}{\text{Var}[X]} \]  

(by the law of large numbers)

\[ = \beta_1 \quad \text{(Cov}[X, u] = 0 \text{ and } \text{Var}[X] > 0) \]

- OLS is inconsistent (and biased) unless \( \text{Cov}[X, u] = 0 \)
- If \( \text{Cov}[u, X] > 0 \) then asymptotic bias is upward; if \( \text{Cov}[u, X] < 0 \) asymptotic bias is downwards
Large Sample Properties: Consistency

Regression Analysis: Large Sample Properties

We covered what are called finite sample OLS estimators in the population model:

\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \epsilon. \]

Unbiasedness of OLS (derived in Chapter 11) is a finite sample property because it holds restriction that \( n \) must be at least as large as \( k + 1 \). Similarly, the fact that under the full set of Gauss-Markov assumptions, we added the classical linear model assumptions, the error is normally distributed and independent of the variables in the sample. In particular, we obtain exact sampling distributions, which we can then use to inference. If the error is not normally distributed, then an F statistic does not have an exact distribution. If the estimator of a parameter is not consistent, then you are wasting your time. There are a few different ways to describe consistency. Formal definitions and results are given in Appendix C; here, we focus on an intuitive understanding.

Let \( \hat{\beta}_j \) be the OLS estimator of \( \beta_j \) for some \( j \). For each \( n \), \( \hat{\beta}_j \) has a probability distribution (representing its possible values in different random samples of size \( n \)). Because \( \hat{\beta}_j \) is unbiased under Assumptions MLR.1 through MLR.4, this distribution has mean value \( \beta_j \). If this estimator is consistent, then the distribution of \( \hat{\beta}_j \) becomes more and more tightly distributed around \( \beta_j \) as the sample size grows. As \( n \) tends to infinity, the distribution of \( \hat{\beta}_j \) collapses to the single point \( \beta_j \). In effect, this means that we can make our estimator arbitrarily close to \( \beta_j \) if we can collect as much data as we want. This convergence is illustrated in Figure 5.1.
Large Sample Properties: Asymptotic Normality

- For statistical inference, we need to know the sampling distribution of $\hat{\beta}$ when $n \to \infty$.

**Theorem (Asymptotic Normality of OLS Estimator)**

Given Assumptions I–V, the OLS estimator $\hat{\beta}_1$ is asymptotically normally distributed:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \approx N(0, 1)$$

where

$$\widehat{SE}[\hat{\beta}_1] = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

with the consistent estimator for the error variance:

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma_u^2$$
Proof.

Proof is similar to the small-sample normality proof:

\[ \hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{SST_x} \right) u_i \]

\[ \sqrt{n} \left( \hat{\beta}_1 - \beta_1 \right) = \frac{\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

where the numerator converges in distribution to a normal random variable by CLT. Then, rearranging the terms, etc. gives you the right formula given in the theorem. For a more formal and detailed proof, see Wooldridge Appendix 5A.
Large Sample Inference

Proof.

Proof is similar to the small-sample normality proof:

\[
\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{SST_x} u_i
\]

\[
\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})u_i}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

where the numerator converges in distribution to a normal random variable by CLT. Then, rearranging the terms, etc. gives you the right formula given in the theorem. For a more formal and detailed proof, see Wooldridge Appendix 5A.

- We need homoskedasticity (Assumption V) for this result, but we do not need normality (Assumption VI).
- Result implies that asymptotically our usual standard errors, t-values, p-values, and CIs remain valid even without the normality assumption! We just proceed as in the small sample case where we assume normality.
- It turns out that, given Assumptions I–V, the OLS asymptotic variance is also the lowest in class (asymptotic Gauss-Markov).
Testing and Confidence Intervals

Three ways of making statistical inference out of regression:
Testing and Confidence Intervals

Three ways of making statistical inference out of regression:

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer $\beta_1$.
Testing and Confidence Intervals

Three ways of making statistical inference out of regression:

1. **Point Estimation**: Consider the sampling distribution of our point estimator \( \hat{\beta}_1 \) to infer \( \beta_1 \)

2. **Hypothesis Testing**: Consider the sampling distribution of a test statistic to test hypothesis about \( \beta_1 \) at the \( \alpha \) level
Testing and Confidence Intervals

Three ways of making statistical inference out of regression:

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer $\beta_1$

2. **Hypothesis Testing**: Consider the sampling distribution of a test statistic to test hypothesis about $\beta_1$ at the $\alpha$ level

3. **Interval Estimation**: Consider the sampling distribution of an interval estimator to construct intervals that will contain $\beta_1$ at least $100(1 - \alpha)\%$ of the time.
Testing and Confidence Intervals

Three ways of making statistical inference out of regression:

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer $\beta_1$.

2. **Hypothesis Testing**: Consider the sampling distribution of a test statistic to test hypothesis about $\beta_1$ at the $\alpha$ level.

3. **Interval Estimation**: Consider the sampling distribution of an interval estimator to construct intervals that will contain $\beta_1$ at least $100(1 - \alpha)$% of the time.

For 2 and 3, we need to know more than just the mean and the variance of the sampling distribution of $\hat{\beta}_1$. We need to know the full shape of the sampling distribution of our estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. 
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Null and alternative hypotheses review

- Null: \( H_0 : \beta_1 = 0 \)
Null and alternative hypotheses review

- **Null:** $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.

- **Alternative:** $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always “some effect”
  - Could do one-sided test, but you shouldn’t

Notice these are statements about the population parameters, not the OLS estimates.
Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship.

- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always “some effect”
  - Could do one-sided test, but you shouldn’t

Notice these are statements about the population parameters, not the OLS estimates.
Null and alternative hypotheses review

- **Null:** $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship.
- **Alternative:** $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always "some effect"
  - Could do one-sided test, but you shouldn't

Notice these are statements about the population parameters, not the OLS estimates.
Null and alternative hypotheses review

- **Null:** $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship

- **Alternative:** $H_a : \beta_1 \neq 0$
  - Claim we want to test
Null and alternative hypotheses review

- **Null**: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship

- **Alternative**: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always “some effect”
Null and alternative hypotheses review

- **Null:** $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship

- **Alternative:** $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always “some effect”
  - Could do one-sided test, but you shouldn’t
Null and alternative hypotheses review

- **Null**: \( H_0 : \beta_1 = 0 \)
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship

- **Alternative**: \( H_a : \beta_1 \neq 0 \)
  - Claim we want to test
  - Almost always “some effect”
  - Could do one-sided test, but you shouldn’t

- Notice these are statements about the population parameters, not the OLS estimates.
Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\hat{\beta}_1 - c}{\hat{SE}[\hat{\beta}_1]}$$
Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\hat{\beta}_1 - c}{SE[\hat{\beta}_1]}$$

- As we saw in the last section, if the errors are conditionally Normal, then under the null hypothesis we have:

$$T \sim t_{n-2}$$
Test statistic

- Under the null of $H_0: \beta_1 = c$, we can use the following familiar test statistic:
  \[ T = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)} \]

- As we saw in the last section, if the errors are conditionally Normal, then under the null hypothesis we have:
  \[ T \sim t_{n-2} \]

- In large samples, we know that $T$ is approximately (standard) Normal, but we also know that $t_{n-2}$ is approximately (standard) Normal in large samples too, so this statement works there too, even if Normality of the errors fails.
Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\hat{\beta}_1 - c}{\hat{SE}[\hat{\beta}_1]}$$

- As we saw in the last section, if the errors are conditionally Normal, then under the null hypothesis we have:

$$T \sim t_{n-2}$$

- In large samples, we know that $T$ is approximately (standard) Normal, but we also know that $t_{n-2}$ is approximately (standard) Normal in large samples too, so this statement works there too, even if Normality of the errors fails.

- Thus, under the null, we know the distribution of $T$ and can use that to formulate a rejection region and calculate p-values.
Rejection region

- Choose a level of the test, $\alpha$, and find rejection regions that correspond to that value under the null distribution:

$$
\mathbb{P} \left( - t_{\alpha/2, n-2} < T < t_{\alpha/2, n-2} \right) = 1 - \alpha
$$
Rejection region

- Choose a level of the test, $\alpha$, and find rejection regions that correspond to that value under the null distribution:

$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

- This is exactly the same as with sample means and sample differences in means, except that the degrees of freedom on the $t$ distribution have changed.
The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true.
p-value

The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true.

Mathematically:

$$P \left( \left| \frac{\hat{\beta}_1 - c}{SE[\hat{\beta}_1]} \right| \geq |T_{obs}| \right)$$
The interpretation of the p-value is the same: **the probability of seeing a test statistic at least this extreme if the null hypothesis were true**

Mathematically:

\[
P \left( \left| \frac{\hat{\beta}_1 - c}{SE[\hat{\beta}_1]} \right| \geq |T_{obs}| \right)
\]

If the p-value is less than \(\alpha\) we would reject the null at the \(\alpha\) level.
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find $t$-values such that:

$$P \left( - t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \leq t_{\alpha/2, n-2} \right) = 1 - \alpha$$
Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find \( t \)-values such that:

\[
P \left( -t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \leq t_{\alpha/2, n-2} \right) = 1 - \alpha
\]

- If we rearrange this as before, we can get an expression for confidence intervals:

\[
P \left( \hat{\beta}_1 - t_{\alpha/2, n-2} SE[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} SE[\hat{\beta}_1] \right) = 1 - \alpha
\]
Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find $t$-values such that:

$$
\mathbb{P}\left(- \frac{t_{\alpha/2,n-2}}{SE[\hat{\beta}_1]} \leq \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \leq \frac{t_{\alpha/2,n-2}}{SE[\hat{\beta}_1]}\right) = 1 - \alpha
$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$
\mathbb{P}\left(\hat{\beta}_1 - t_{\alpha/2,n-2}SE[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}SE[\hat{\beta}_1]\right) = 1 - \alpha
$$

- Thus, we can write the confidence intervals as:

$$
\hat{\beta}_1 \pm t_{\alpha/2,n-2}SE[\hat{\beta}_1]
$$
Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find $t$-values such that:

$$
P\left(- t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \leq t_{\alpha/2, n-2}\right) = 1 - \alpha
$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$
P\left(\hat{\beta}_1 - t_{\alpha/2, n-2}SE[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2}SE[\hat{\beta}_1]\right) = 1 - \alpha
$$

- Thus, we can write the confidence intervals as:

$$
\hat{\beta}_1 \pm t_{\alpha/2, n-2}SE[\hat{\beta}_1]
$$

- We can derive these for the intercept as well:

$$
\hat{\beta}_0 \pm t_{\alpha/2, n-2}SE[\hat{\beta}_0]
$$
Returning to our simulation example we can simulate the sampling distributions of the 95% confidence interval estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$. 

![Simulation Example Diagram]
Returning to our simulation example we can simulate the sampling distributions of the 95% confidence interval estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$. 
When we repeat the process over and over, we expect 95% of the confidence intervals to contain the true parameters. Note that, in a given sample, one CI may cover its true value and the other may not.
When we repeat the process over and over, we expect 95% of the confidence intervals to contain the true parameters. Note that, in a given sample, one CI may cover its true value and the other may not.
Prediction error

How do we judge how well a line fits the data?
Prediction error

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting $Y$ once we include $X$ into the regression model.

\[
SS_{\text{tot}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2
\]

Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or $SS_{\text{res}}$:

\[
SS_{\text{res}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]
Prediction error

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting $Y$ once we include $X$ into the regression model.
- Prediction errors without $X$: best prediction is the mean, so our squared errors, or the total sum of squares ($SS_{tot}$) would be:

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$
Prediction error

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting $Y$ once we include $X$ into the regression model.
- Prediction errors without $X$: best prediction is the mean, so our squared errors, or the total sum of squares ($SS_{tot}$) would be:

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or $SS_{res}$:

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$
Sum of Squares

Total Prediction Errors

Log GDP per capita growth vs Log Settler Mortality
Sum of Squares

Residuals

Log GDP per capita growth vs. Log Settler Mortality

Stewart (Princeton)
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

This is the fraction of the total prediction error eliminated by providing information on $X$.

Alternatively, this is the fraction of the variation in $Y$ is "explained by" $X$.

$R^2 = 0$ means no relationship

$R^2 = 1$ implies perfect linear fit
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?
- We quantify this question with the coefficient of determination or $R^2$. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

This is the fraction of the total prediction error eliminated by providing information on $X$. Alternatively, this is the fraction of the variation in $Y$ is "explained by" $X$.

$R^2 = 0$ means no relationship

$R^2 = 1$ implies perfect linear fit
By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?

We quantify this question with the coefficient of determination or $R^2$. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

This is the fraction of the total prediction error eliminated by providing information on $X$.
By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?

We quantify this question with the coefficient of determination or $R^2$. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

This is the fraction of the total prediction error eliminated by providing information on $X$.

Alternatively, this is the fraction of the variation in $Y$ is “explained by” $X$. 

**R-square**
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?
- We quantify this question with the coefficient of determination or $R^2$. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on $X$.
- Alternatively, this is the fraction of the variation in $Y$ is “explained by” $X$.
- $R^2 = 0$ means no relationship
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the $SS_{res}$ compared to the $SS_{tot}$?
- We quantify this question with the coefficient of determination or $R^2$. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on $X$.
- Alternatively, this is the fraction of the variation in $Y$ is “explained by” $X$.
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit
Is R-squared useful?

\[ R^2 = 0.66 \]
Is R-squared useful?

\[ R^2 = 0.96 \]
Is R-squared useful?

\[ R^2 = 0.27 \]
Is R-squared useful?
Why $r^2$?

To calculate $r^2$, we need to think about the following two quantities:

1. **TSS**: Total sum of squares
2. **SSE**: Sum of squared errors

\[
TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2.
\]

\[
SSE = \sum_{i=1}^{n} u_i^2.
\]

\[
r^2 = 1 - \frac{SSE}{TSS}.
\]
1−SSE/TSS = 0.45
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \{\hat{u}_i + (\hat{y}_i - \bar{y})\}^2
\]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left\{ \hat{u}_i + (\hat{y}_i - \bar{y}) \right\}^2 \\
= \sum_{i=1}^{n} \{ \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \} \]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \{\hat{u}_i + (\hat{y}_i - \bar{y})\}^2 \\
= \sum_{i=1}^{n} \{\hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\} \\
= \sum_{i=1}^{n} \hat{u}_i^2 + 2 \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left\{ \hat{u}_i + (\hat{y}_i - \bar{y}) \right\}^2
\]

\[
= \sum_{i=1}^{n} \left\{ \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \right\}
\]

\[
= \sum_{i=1}^{n} \hat{u}_i^2 + 2 \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

\[
= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]
Derivation

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \{ \hat{u}_i + (\hat{y}_i - \bar{y}) \}^2 \]

\[ = \sum_{i=1}^{n} \{ \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \} \]

\[ = \sum_{i=1}^{n} \hat{u}_i^2 + 2 \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \]

\[ = \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \]

\[ TSS = SSE + RegSS \]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \{\hat{u}_i + (\hat{y}_i - \bar{y})\}^2 \\
= \sum_{i=1}^{n} \{\hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\} \\
= \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

\[TSS = SSE + RegSS\]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left\{ \hat{u}_i + (\hat{y}_i - \bar{y}) \right\}^2 \\
= \sum_{i=1}^{n} \left\{ \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \right\} \\
= \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
TSS = SSE + RegSS
\]
Derivation

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \{ \hat{u}_i + (\hat{y}_i - \bar{y}) \}^2 \\
= \sum_{i=1}^{n} \{ \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \} \\
= \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
TSS = SSE + RegSS
\]
We can divide each side by the TSS:

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}
\]

\(r^2\) is a measure of how much of the variation in \(Y\) is accounted for by \(X\).
Coefficient of Determination

We can divide each side by the TSS:

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}
\]

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = 1
\]

\(r^2\) is a measure of how much of the variation in \(Y\) is accounted for by \(X\).
Coefficient of Determination

We can divide each side by the TSS:

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}
\]

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = 1
\]

\[
\frac{RegSS}{TSS} = 1 - \frac{SSE}{TSS} = r^2
\]
Coefficient of Determination

We can divide each side by the TSS:

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}
\]

\[
\frac{SSE}{TSS} + \frac{RegSS}{TSS} = 1
\]

\[
\frac{RegSS}{TSS} = 1 - \frac{SSE}{TSS} = r^2
\]

\(r^2\) is a measure of how much of the variation in \(Y\) is accounted for by \(X\).
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Mechanics of OLS

Properties of the OLS estimator

Example and Review

Properties Continued

Hypothesis tests for regression

Confidence intervals for regression

Goodness of fit

Wrap Up of Univariate Regression

Fun with Non-Linearities
OLS Assumptions Summary

- **Identification**
  - Data Description
  - Variation in X

- **Unbiasedness**
  - Consistency
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean

- **Gauss-Markov (BLUE)**
  - Asymptotic Inference (z and $\chi^2$)
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity

- **Classical LM (BUE)**
  - Small-Sample Inference (t and F)
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
  - Normality of Errors
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the "best fitting" line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$)

2. Predictive inference:
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$): "What's my best guess about the wage of a new worker who only has high school education?"
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).

Stewart (Princeton)
Week 5: Simple Linear Regression
October 10, 12, 2016 88 / 103
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the "best fitting" line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$ only).

2. Predictive inference:
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$): "What's my best guess about the wage of a new worker who only has high school education?"
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the “best fitting” line.

   Assumption required: III (variation in $X$) only.

2. Predictive inference:
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$): “What's my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).

Stewart (Princeton)
Week 5: Simple Linear Regression
October 10, 12, 2016  88 / 103
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.

2. **Predictive inference:**
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$): “What’s my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the **statistical properties** of the OLS estimator.
- However, these properties do not tell us what **types of inference** we can draw from the estimates.

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. **Predictive inference:**
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$): "What's my best guess about the wage of a new worker who only has high school education?"
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the **statistical properties** of the OLS estimator
- However, these properties do not tell us what **types of inference** we can draw from the estimates

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line
   - No inference about the underlying population intended
   - Assumption required: III (variation in $X$) only

2. **Predictive inference:**
   - Inference about a new observation coming from the same population
   - Example: Wage ($Y$) and education ($X$): “What’s my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling)
   - Assumptions desired: I (linearity)
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. **Predictive inference:**
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. **Predictive inference:**
   - Inference about a new observation coming from the same population.
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. Predictive inference:
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$):
     “What’s my best guess about the wage of a new worker who only has high school education?”
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. Descriptive inference:
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. Predictive inference:
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$):
     “What’s my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling).
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator
- However, these properties do not tell us what types of inference we can draw from the estimates

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line
   - No inference about the underlying population intended
   - Assumption required: III (variation in $X$) only

2. **Predictive inference:**
   - Inference about a new observation coming from the same population
   - Example: Wage ($Y$) and education ($X$):
     “What’s my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling)
   - Assumptions desired: I (linearity)
What Do the Regression Coefficients Mean Substantively?

- So far, we have learned the statistical properties of the OLS estimator.
- However, these properties do not tell us what types of inference we can draw from the estimates.

Three types of inference:

1. **Descriptive inference:**
   - Summarizing sample data by drawing the “best fitting” line.
   - No inference about the underlying population intended.
   - Assumption required: III (variation in $X$) only.

2. **Predictive inference:**
   - Inference about a new observation coming from the same population.
   - Example: Wage ($Y$) and education ($X$):
     “What’s my best guess about the wage of a new worker who only has high school education?”
   - Assumptions required: III and II (random sampling).
   - Assumptions desired: I (linearity).
What Do the Regression Coefficients Mean Substantively?

Causal inference:
▶ Inference about counterfactuals, i.e. hypothetical interventions to the same units
▶ Example: Wage ($Y$) and education ($X$):
   "What would my current wage be if I only had high school education?"
▶ Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
▶ In this sequence we will continue to discuss causal identification assumptions

Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference

Implications:
▶ When Assumptions I–IV are all satisfied, we can estimate the structural parameters $\beta$ without bias and thus make causal inference.
▶ However, we can make predictive inference even if some assumptions are violated.
What Do the Regression Coefficients Mean Substantively?

3 Causal inference:

Inference about counterfactuals, i.e. hypothetical interventions to the same units
Example: Wage ($Y$) and education ($X$):
"What would my current wage be if I only had high school education?"
Assumptions required (under the current framework): I, II, III and IV
In this sequence we will continue to discuss causal identification assumptions
Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference
Implications:
When Assumptions I–IV are all satisfied, we can estimate the structural parameters $\beta$ without bias and thus make causal inference.
However, we can make predictive inference even if some assumptions are violated.
What Do the Regression Coefficients Mean Substantively?

3 Causal inference:
   - Inference about counterfactuals, i.e. hypothetical interventions to the same units

Example: Wage ($Y$) and education ($X$):
"What would my current wage be if I only had high school education?"

Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)

In this sequence we will continue to discuss causal identification assumptions.

Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference.

Implications:

- When Assumptions I–IV are all satisfied, we can estimate the structural parameters $\beta$ without bias and thus make causal inference.
- However, we can make predictive inference even if some assumptions are violated.
What Do the Regression Coefficients Mean Substantively?

3. Causal inference:
   - Inference about counterfactuals, i.e. hypothetical interventions to the same units
   - Example: Wage ($Y$) and education ($X$):
     “What would my current wage be if I only had high school education?”
What Do the Regression Coefficients Mean Substantively?

Causal inference:

- Inference about counterfactuals, i.e. hypothetical interventions to the same units
- Example: Wage ($Y$) and education ($X$): “What would my current wage be if I only had high school education?”
- Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
What Do the Regression Coefficients Mean Substantively?

Causal inference:

- Inference about counterfactuals, i.e. hypothetical interventions to the same units
- Example: Wage ($Y$) and education ($X$): “What would my current wage be if I only had high school education?”
- Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
- In this sequence we will continue to discuss causal identification assumptions
What Do the Regression Coefficients Mean Substantively?

Causal inference:

- Inference about counterfactuals, i.e. hypothetical interventions to the same units
- Example: Wage ($Y$) and education ($X$):
  “What would my current wage be if I only had high school education?”
- Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
- In this sequence we will continue to discuss causal identification assumptions

Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference
What Do the Regression Coefficients Mean Substantively?

Causal inference:

- Inference about counterfactuals, i.e. hypothetical interventions to the same units
- Example: Wage \( (Y) \) and education \( (X) \):
  “What would my current wage be if I only had high school education?”
- Assumptions required (under the current framework): I, II, III and IV (zero conditional mean)
- In this sequence we will continue to discuss causal identification assumptions

- Notice in the wage example, how the omission of unobserved ability from the equation does or does not affect each type of inference

Implications:

- When Assumptions I–IV are all satisfied, we can estimate the structural parameters \( \beta \) without bias and thus make causal inference.
- However, we can make predictive inference even if some assumptions are violated.
OLS as a Best Linear Predictor (Review of BLUE)

Suppose that we want to predict the values of $Y$ given observed $X$ values. Suppose further that we've decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$\text{MSE} = E \left[ (Y_i - \hat{Y}_i)^2 \right] = E \left[ (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = E \left[ \hat{u}_i^2 \right]$$

The smaller a predictor makes $\text{MSE}$, the better.

Now, note that the sample version of $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$.

Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}_i^2$!

This implies that OLS is the best linear predictor in terms of MSE.

Which assumptions did we use to get this result?

$\begin{align*}
\text{Needed: Assumptions II (random sampling) and III (variation in } X) \\
\text{Not needed: Assumptions I (linearity) and IV (zero cond. mean)}
\end{align*}$

Note that Assumption I would make OLS the best, not just best linear, predictor, so it is certainly desired.
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.

Let's use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$MSE = \mathbb{E}[(Y_i - \hat{Y}_i)^2] = \mathbb{E}[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2] = \mathbb{E}[^2]$$

The smaller a predictor makes $MSE$, the better.

Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$.

Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}_i^2$.

This implies that OLS is the best linear predictor in terms of MSE.

Which assumptions did we use to get this result?

- Needed: Assumptions II (random sampling) and III (variation in $X$).
- Not needed: Assumptions I (linearity) and IV (zero cond. mean).

Note that Assumption I would make OLS the best, not just best linear, predictor, so it is certainly desired.
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.
- Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$\text{MSE} = E \left[ (Y_i - \hat{Y}_i)^2 \right] = E \left[ (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = E \left[ \hat{u}^2_i \right]$$

The smaller a predictor makes $\text{MSE}$, the better.

Now, note that the sample version of $\text{MSE}$ is

$$\frac{1}{n} \sum_{i=1}^{n} \hat{u}^2_i$$

Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}^2_i$!

This implies that OLS is the best linear predictor in terms of $\text{MSE}$.

Which assumptions did we use to get this result?

- Needed: Assumptions II (random sampling) and III (variation in $X$)
- Not needed: Assumptions I (linearity) and IV (zero cond. mean)

Note that Assumption I would make OLS the best, not just best linear, predictor, so it is certainly desired.
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.
- Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$MSE = E\left[(Y_i - \hat{Y}_i)^2\right] = E\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i)^2\right] = E\left[\hat{u}_i^2\right]$$

The smaller a predictor makes $MSE$, the better.
Suppose that we want to predict the values of $Y$ given observed $X$ values

Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population)

How to choose a good predictor? A popular criterion is mean squared error:

$$MSE = E \left[ (Y_i - \hat{Y}_i)^2 \right] = E \left[ (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = E \left[ \hat{u}_i^2 \right]$$

The smaller a predictor makes $MSE$, the better.

Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.

- Suppose further that we've decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

- How to choose a good predictor? A popular criterion is mean squared error:

\[
MSE = E \left[ (Y_i - \hat{Y}_i)^2 \right] = E \left[ (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = E \left[ \hat{u}_i^2 \right]
\]

The smaller a predictor makes $MSE$, the better.

- Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$.

- Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}^2$!
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values
- Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population)

How to choose a good predictor? A popular criterion is mean squared error:

$$
MSE = E \left[ (Y_i - \hat{Y}_i)^2 \right] = E \left[ (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = E \left[ \hat{u}_i^2 \right]
$$

The smaller a predictor makes $MSE$, the better.

- Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$
- Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}_i^2$!
- This implies that OLS is the best linear predictor in terms of MSE
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.
- Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$MSE = E[(Y_i - \hat{Y}_i)^2] = E[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2] = E[\hat{u}_i^2]$$

The smaller a predictor makes $MSE$, the better.

- Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$.
- Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}^2$!
- This implies that OLS is the best linear predictor in terms of MSE.
- Which assumptions did we use to get this result?

- Needed: Assumptions II (random sampling) and III (variation in $X$).
- Not needed: Assumptions I (linearity) and IV (zero cond. mean).

Note that Assumption I would make OLS the best, not just best linear, predictor, so it is certainly desired.
OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of $Y$ given observed $X$ values.
- Suppose further that we’ve decided to use a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily assume a true linear relationship in the population).

How to choose a good predictor? A popular criterion is mean squared error:

$$MSE = E [(Y_i - \hat{Y}_i)^2] = E [(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2] = E [\hat{u}_i^2]$$

The smaller a predictor makes $MSE$, the better.

Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$

Recall how we got the OLS estimator; we minimized $\sum_{i=1}^{n} \hat{u}^2$!

This implies that OLS is the best linear predictor in terms of MSE.

Which assumptions did we use to get this result?
- Needed: Assumptions II (random sampling) and III (variation in $X$)
- Not needed: Assumptions I (linearity) and IV (zero cond. mean)
Suppose that we want to predict the values of $Y$ given observed $X$ values

Suppose further that we’ve decided to *use* a linear predictor $\hat{\beta}_0 + \hat{\beta}_1 X$ (but not necessarily *assume* a true linear relationship in the population)

How to choose a good predictor? A popular criterion is *mean squared error*:

$$MSE = E[(Y_i - \hat{Y}_i)^2] = E[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2] = E[\hat{u}_i^2]$$

The smaller a predictor makes $MSE$, the better.

Now, note that the sample version of $MSE = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2$

Recall how we got the OLS estimator; we *minimized* $\sum_{i=1}^{n} \hat{u}^2$!

This implies that OLS is the **best linear predictor** in terms of MSE

Which assumptions did we use to get this result?

- Needed: Assumptions II (random sampling) and III (variation in $X$)
- *Not* needed: Assumptions I (linearity) and IV (zero cond. mean)

Note that Assumption I would make OLS the **best**, not just best linear, predictor, so it is certainly desired
State Legislators and African American Population

Interpretations of increasing quality:

> summary(lm(beo ~ bpop, data = D))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -1.31489 | 0.32775    | -4.012  | 0.000264 *** |
| bpop           | 0.35848  | 0.02519    | 14.232  | < 2e-16 *** |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385, Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF, p-value: < 2.2e-16

“African American population is statistically significant ($p < 0.001$)”
State Legislators and African American Population

Interpretations of increasing quality:

> summary(lm(beo ~ bpop, data = D))

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -1.31489 | 0.32775 | -4.012 | 0.000264 *** |
| bpop        | 0.35848  | 0.02519 | 14.232 | < 2e-16 *** |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385, Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF, p-value: < 2.2e-16

“Percent African American legislators increases with African American population (p < 0.001)”
State Legislators and African American Population

Interpretations of increasing quality:

> summary(lm(beo ~ bpop, data = D))

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| -1.31489 | 0.32775    | -4.012  | 0.000264 *** |
| bpop       | 0.35848  | 0.02519    | 14.232  | < 2e-16 *** |

---

Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385, Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16

“A one percentage point increase in the African American population causes a 0.35 percentage point increase in the fraction of African American state legislators (p < 0.001).”
State Legislators and African American Population

Interpretations of increasing quality:

> summary(lm(beo ~ bpop, data = D))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -1.31489 | 0.32775    | -4.012  | 0.000264 *** |
| bpop           | 0.35848  | 0.02519    | 14.232  | < 2e-16 *** |

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385, Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16

“A one percentage point increase in the African American population is associated with a 0.35 percentage point increase in the fraction of African American state legislators (p < 0.001).”
Ground Rules: Interpretation of the Slope

1. Give a short, but precise interpretation of the exact meaning of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.

2. Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.

3. Do not resort to unwarranted causal language: Say "predicts", "associated with", "expected difference" or "correlated with" instead of "causes", "leads" or "affects".

4. Give a short, but precise interpretation of statistical significance.

5. Give a short, but precise interpretation of practical significance. You want to discuss the magnitude of the slope in your particular application.
Give a short, but precise interpretation of the exact meaning of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.

- Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.
Ground Rules: Interpretation of the Slope

1. Give a short, but precise interpretation of the exact meaning of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.
   - Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.

2. Do not resort to unwarranted causal language: Say “predicts”, “associated with”, “expected difference” or “correlated with” instead of “causes”, “leads” or “affects”
Ground Rules: Interpretation of the Slope

1. Give a short, but precise interpretation of the exact meaning of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.
   - Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.

2. Do not resort to unwarranted causal language: Say “predicts”, “associated with”, “expected difference” or “correlated with” instead of “causes”, “leads” or “affects”

3. Give a short, but precise interpretation of statistical significance
Ground Rules: Interpretation of the Slope

1. Give a short, but precise interpretation of the **exact meaning** of the value of the slope coefficient referring to the concepts, units, direction, and magnitude.
   - Estimate suggests that one additional hour of reading the textbook is associated with 10 additional points on the exam.

2. Do not resort to unwarranted causal language: Say “predicts”, “associated with”, “expected difference” or “correlated with” instead of “causes”, “leads” or “affects”

3. Give a short, but precise interpretation of **statistical significance**

4. Give a short, but precise interpretation of **practical significance**. You want to discuss the **magnitude** of the slope in your particular application.
Reporting Statistical Significance

A reasonable way to think about statistical significance is to think about the precision of the estimates. If the slope is large substantively but just barely fails to reach conventional levels of significance, it may still be interesting.

Examples:

- We reject the null hypothesis that the slope is zero at the .05 level.
- The slope coefficient suggests that a one unit change in $X$ is associated with a 10 unit change in $Y$ ($p < .02$).
- The slope coefficient is fairly precisely estimated, the 95% confidence interval ranging from 8 to 10.
A reasonable way to think about statistical significance is to think about the precision of the estimates.
A reasonable way to think about statistical significance is to think about the **precision** of the estimates.

If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.
Reporting Statistical Significance

- A reasonable way to think about statistical significance is to think about the precision of the estimates.

- If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.

- Examples:
A reasonable way to think about statistical significance is to think about the precision of the estimates.

If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.

Examples:

- We reject the null hypothesis that the slope is zero at the .05 level.
- The slope coefficient suggests that a one unit change in X is associated with a 10 unit change in Y (p. < .02).
- The slope coefficient is fairly precisely estimated, the 95% confidence interval ranging from 8 to 10.
Reporting Statistical Significance

- A reasonable way to think about statistical significance is to think about the precision of the estimates.

- If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.

- Examples:
  - We reject the null hypothesis that the slope is zero at the .05 level.
  - The slope coefficient suggests that a one unit change in $X$ is associated with a 10 unit change in $Y$ ($p.< .02$).
Reporting Statistical Significance

- A reasonable way to think about statistical significance is to think about the precision of the estimates.

- If the slope is large substantively but just barely fails to reach conventional levels of significance it may still be interesting.

Examples:

- We reject the null hypothesis that the slope is zero at the .05 level.

- The slope coefficient suggests that a one unit change in $X$ is associated with a 10 unit change in $Y$ ($p < .02$).

- The slope coefficient is fairly precisely estimated, the 95% confidence interval ranging from 8 to 10.
Reporting Substantive Significance

Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!

To comment on substantive magnitude you should set up a "plausible" contrast keeping in mind (1) the distributions of variables and the (2) the substantive context.

Examples:

Stewart (Princeton)
Reporting Substantive Significance

- Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!
Reporting Substantive Significance

- Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!

- To comment on substantive magnitude you should set up a “plausible” contrast keeping in mind (1) the distributions of variables and the (2) the substantive context
Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!

To comment on substantive magnitude you should set up a “plausible” contrast keeping in mind (1) the distributions of variables and the (2) the substantive context

Examples:

Turnout on education: The slope estimate suggests that a high school dropout (10th percentile of the education distribution) on average has a .3 lower probability of voting compared to a college graduate (75th percentile of schooling).

The average probability of voting among HSDs is .2, so this corresponds to a 150 % increase for an average HSD.
Reporting Substantive Significance

- Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!

- To comment on substantive magnitude you should set up a “plausible” contrast keeping in mind (1) the distributions of variables and the (2) the substantive context

- Examples:

  Democracy on GDP: Going from the 25th to the 75th percentile of the GDP distribution (e.g. comparing Ghana and Spain) is associated with a 10 point increase in the average polity index, which corresponds to an increase from the 25th to the 52nd percentile of the democracy distribution
Statistical significance and substantive significance are not the same: with a large enough sample size even truly microscopic differences can be statistically significant!

To comment on substantive magnitude you should set up a “plausible” contrast keeping in mind (1) the distributions of variables and the (2) the substantive context

Examples:

Earnings on Schooling: The standard deviation is 2.5 years for schooling and $50,000 for annual earnings. Thus, the slope estimates suggest that a one standard deviation increase in schooling is associated with a .8 standard deviation increase in earnings.
Next Week
Next Week

- OLS with two regressors
Next Week

- OLS with two regressors
- Omitted Variables and Multicolinearity
Next Week

- OLS with two regressors
- Omitted Variables and Multicolinearity
- Dummy variables, interactions, polynomials
Next Week

- OLS with two regressors
- Omitted Variables and Multicolinearity
- Dummy variables, interactions, polynomials

Reading:
- Fox Chapter 5.2.1 (Least Squares with Two Variables)
- Fox Chapter 7.1-7.3 (Dummy-Variable Regression, Interactions)
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
1. Mechanics of OLS
2. Properties of the OLS estimator
3. Example and Review
4. Properties Continued
5. Hypothesis tests for regression
6. Confidence intervals for regression
7. Goodness of fit
8. Wrap Up of Univariate Regression
9. Fun with Non-Linearities
Fun with Non-Linearities

The linear regression model can accommodate non-linearity in $X$ (but not in $\beta$). We do this by first transforming $X$ appropriately. A useful transformation when variables are positive and right-skewed is the (natural) logarithm. The log transformation changes the interpretation of $\beta_1$:

$\rightarrow$ Regress log($Y$) on $X$: $\beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$.

$\rightarrow$ Regress $Y$ on log($X$): $\beta_1$ approximates increase in $Y$ associated with a percent increase in $X$.

Note that these approximations work only for small increments. In particular, they do not work when $X$ is a discrete random variable.
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first transforming $X$ appropriately

A useful transformation when variables are positive and right-skewed is the (natural) logarithm.

The log transformation changes the interpretation of $\beta_1$:

- Regress $\log(Y)$ on $X$ $\rightarrow$ $\beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$
- Regress $Y$ on $\log(X)$ $\rightarrow$ $\beta_1$ approximates increase in $Y$ associated with a percent increase in $X$

Note that these approximations work only for small increments.

In particular, they do not work when $X$ is a discrete random variable.
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first transforming $X$ appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$).
- We do this by first transforming $X$ appropriately.

A useful transformation when variables are positive and right-skewed is the (natural) logarithm.

- The log transformation changes the interpretation of $\beta_1$:

\[ \text{Regress } \log(Y) \text{ on } X \rightarrow \beta_1 \text{ approximates percent increase in } Y \text{ associated with one unit increase in } X \]

\[ \text{Regress } Y \text{ on } \log(X) \rightarrow \beta_1 \text{ approximates increase in } Y \text{ associated with a percent increase in } X \]

Note that these approximations work only for small increments.

In particular, they do not work when $X$ is a discrete random variable.
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first transforming $X$ appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of $\beta_1$:
  - Regress $\log(Y)$ on $X \rightarrow \beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$
Fun with Non-Linearities

- The linear regression model can accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first transforming $X$ appropriately

A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of $\beta_1$:  
  - Regress $\log(Y)$ on $X \rightarrow \beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$
  - Regress $Y$ on $\log(X) \rightarrow \beta_1$ approximates increase in $Y$ associated with a percent increase in $X$

Note that these approximations work only for small increments in particular, they do not work when $X$ is a discrete random variable
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first transforming $X$ appropriately

A useful transformation when variables are positive and right-skewed is the (natural) logarithm

The log transformation changes the interpretation of $\beta_1$:

- Regress $\log(Y)$ on $X$ $\rightarrow$ $\beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$  
- Regress $Y$ on $\log(X)$ $\rightarrow$ $\beta_1$ approximates increase in $Y$ associated with a percent increase in $X$
- Note that these approximations work only for small increments
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in $X$ (but not in $\beta$)
- We do this by first *transforming* $X$ appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of $\beta_1$:
  - Regress $\log(Y)$ on $X \rightarrow \beta_1$ approximates percent increase in $Y$ associated with one unit increase in $X$
  - Regress $Y$ on $\log(X) \rightarrow \beta_1$ approximates increase in $Y$ associated with a percent increase in $X$
  - Note that these approximations work only for small increments
  - In particular, they do not work when $X$ is a discrete random variable
\( \hat{\beta}_1 = 1.23 \)
\[ \hat{\beta}_1 = 1.23 \rightarrow \text{One additional soldier killed predicts } 1.23 \text{ additional soldiers wounded on average} \]
Wounded (Scale in Levels)

World War II
Civil War, North
World War I
Vietnam War
Civil War, South
Korean War
Okinawa
Operation Iraqi Freedom, Iraq
Iwo Jima
Revolutionary War
War of 1812
Aleutian Campaign
D-Day
Philippines War
Indian Wars
Spanish American War
Terrorism, World Trade Center
Yemen, USS Cole
Terrorism Khobar Towers, Saudi Arabia
Persian Gulf
Terrorism Oklahoma City
Persian Gulf, Op Desert Shield/Storm
Russia North Expedition
Moro Campaigns
China Boxer Rebellion
Panama
Dominican Republic
Israel Attack/USS Liberty
Lebanon
Texas War Of Independence
South Korea
Grenada
China Yangtze Service
Mexico
Nicaragua
Barbary Wars
Russia Siberia Expedition
Dominican Republic
China Civil War
Terrorism Riyadh, Saudi Arabia
North Atlantic Naval War
Franco-American Naval War
Operation Enduring Freedom, Afghanistan
Mexican War
Operation Enduring Freedom, Afghanistan Theater
Haiti
Texas Border Cortina War
Nicaragua
Italy Trieste
Japan

Number of Wounded
Wounded (Logarithmic Scale)

Log(Number of Wounded)

Number of Wounded

World War II
Civil War, North
World War I
Vietnam War
Civil War, South
Korean War
Okina
Operation Iraqi Freedom, Iraq
Iwo Jima
Revolutionary War
War of 1812
Aleutian Campaign
D-Day
Philippines War
Indian Wars
Spanish American War
Terrorism, World Trade Center
Yemen, USS Cole
Terrorism Khobar Towers, Saudi Arabia
Persian Gulf
Terrorism Oklahoma City
Persian Gulf, Op Desert Shield/Storm
Russia North Expedition
Moro Campaigns
China Boxer Rebellion
Panama
Dominican Republic
Israel Attack/USS Liberty
Lebanon
Texas War Of Independence
South Korea
Grenada
China Yangtze Service
Mexico
Nicaragua
Barbary Wars
Russia Siberia Expedition
Dominican Republic
China Civil War
Terrorism Riyadh, Saudi Arabia
North Atlantic Naval War
Franco–Amer Naval War
Operation Enduring Freedom, Afghanistan
Mexican War
Operation Enduring Freedom, Afghanistan Theater
Haiti
Texas Border Cortina War
Nicaragua
Italy Trieste
Japan

Stewart (Princeton)  Week 5: Simple Linear Regression  October 10, 12, 2016 100 / 103
Regression: Log-Level

\[ \hat{\beta}_1 = 0.0000237 \rightarrow \]
\[ \hat{\beta}_1 = 0.0000237 \quad \text{One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded on average} \]
\( \hat{\beta}_1 = 0.797 \)
Regression: Log-Log

\[ \hat{\beta}_1 = 0.797 \quad \rightarrow \quad \text{A percent increase in deaths predicts 0.797 percent increase in the wounded on average} \]