Week 5: Simple Linear Regression

Brandon Stewart[1]

Princeton

October 10, 12, 2016

[1] 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...

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression
- This Week
  - Monday:

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression
- This Week
  - Monday:
    - mechanics 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

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

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

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.

The population linear regression function

The population linear regression function

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

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

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
- β_0, β_1 = population intercept and population slope (what we want to estimate)

The sample linear regression function

The sample linear regression function

- The estimated or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

The sample linear regression function

- The estimated or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope

The sample linear regression function

- The estimated or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope
- $\widehat{Y}_i$ 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:

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

The sample linear regression function

- The estimated or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

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

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

- You can think of the residuals as the prediction errors of our estimates.

Overall Goals for the Week

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

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:

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):

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

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):

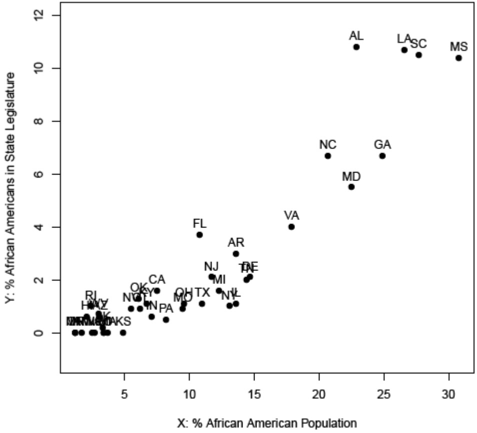
$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

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

Graphical Example

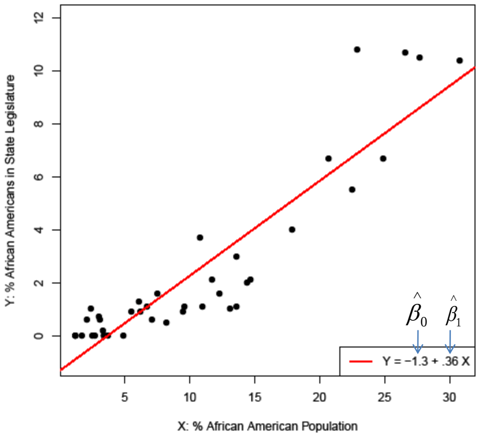
Graphical Example

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



Graphical Example

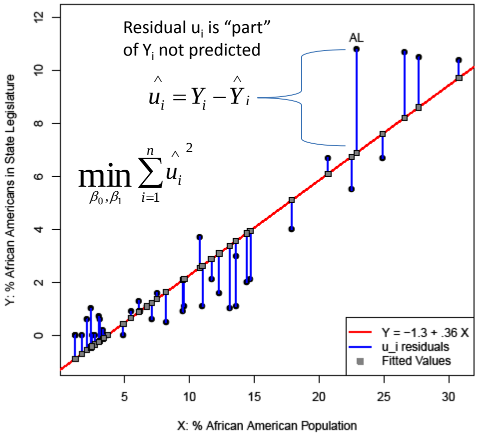
How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?



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



Deriving the OLS estimator

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$

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: $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$
- 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 β_0 and β_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 β_0 and β_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 β_0 and β_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 β_0 and β_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 β_0 and β_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**:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

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

Intuition of the OLS estimator

Intuition of the OLS estimator

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

Intuition of the OLS estimator

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

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

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

Intuition of the OLS estimator

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

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

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

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

Intuition of the OLS estimator

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

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

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- 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 $(\overline{Y}, \overline{X})$:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$
- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- 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?

Intuition of the OLS estimator

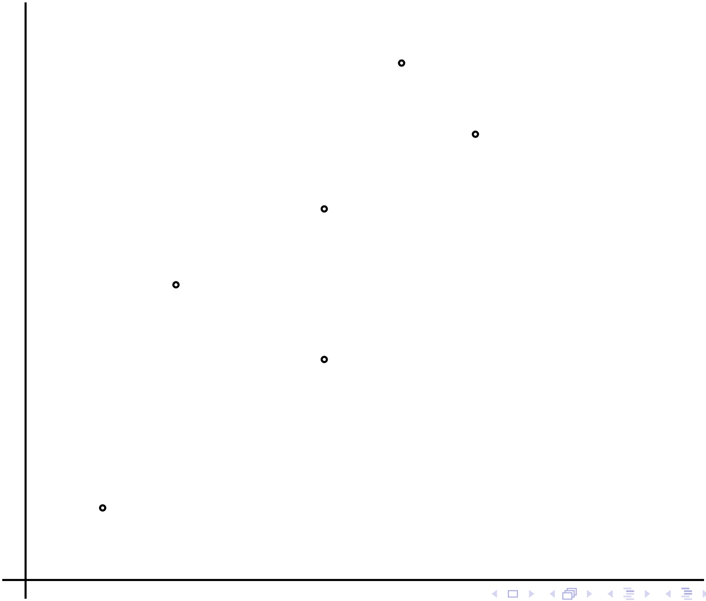
- The intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$

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

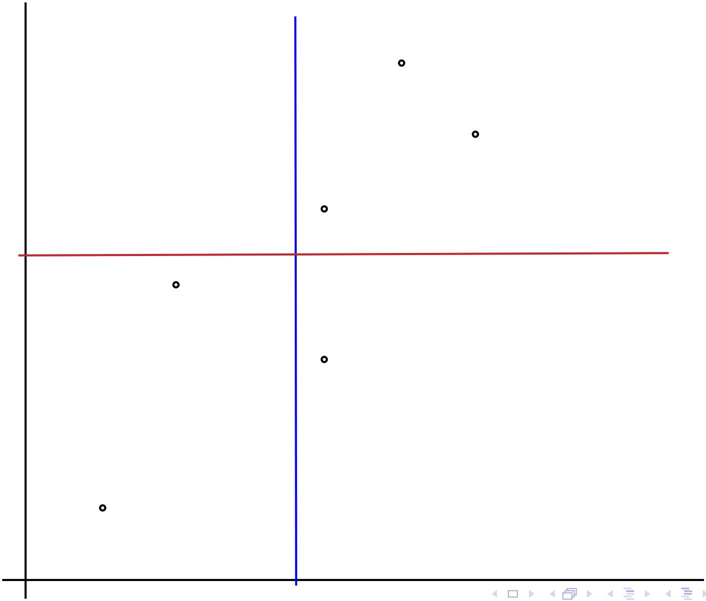
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- 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?

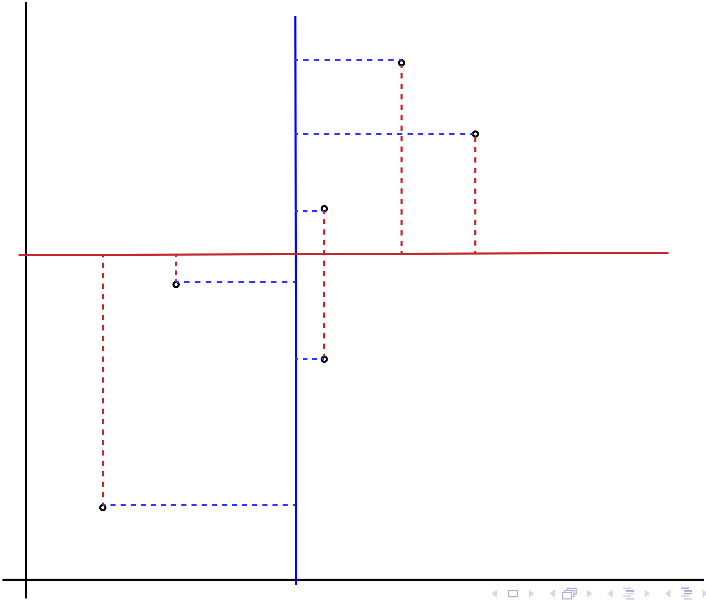
A Visual Intuition for the OLS Estimator



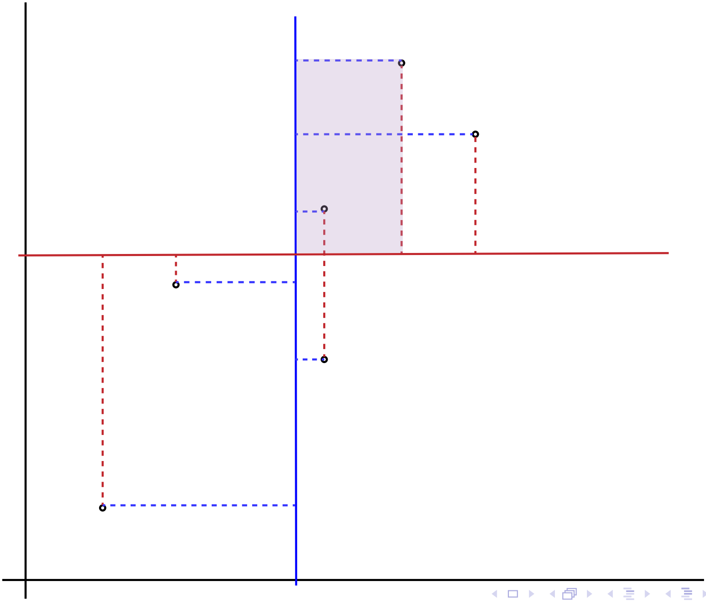
A Visual Intuition for the OLS Estimator



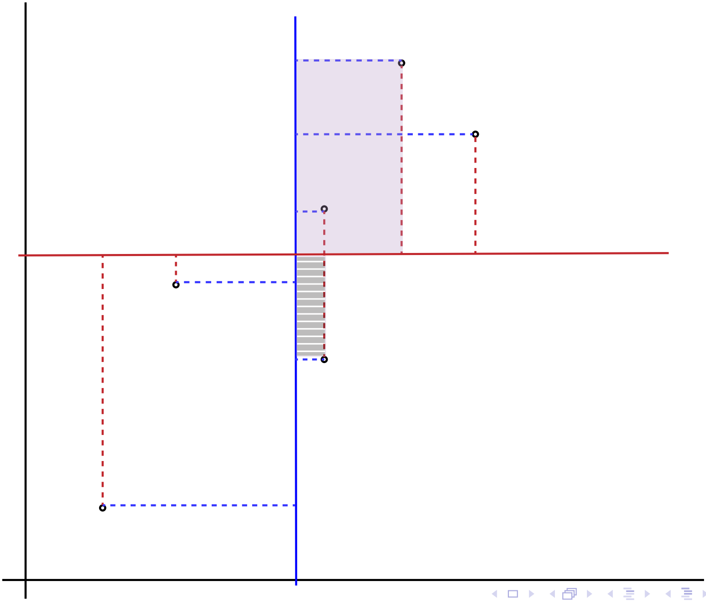
A Visual Intuition for the OLS Estimator



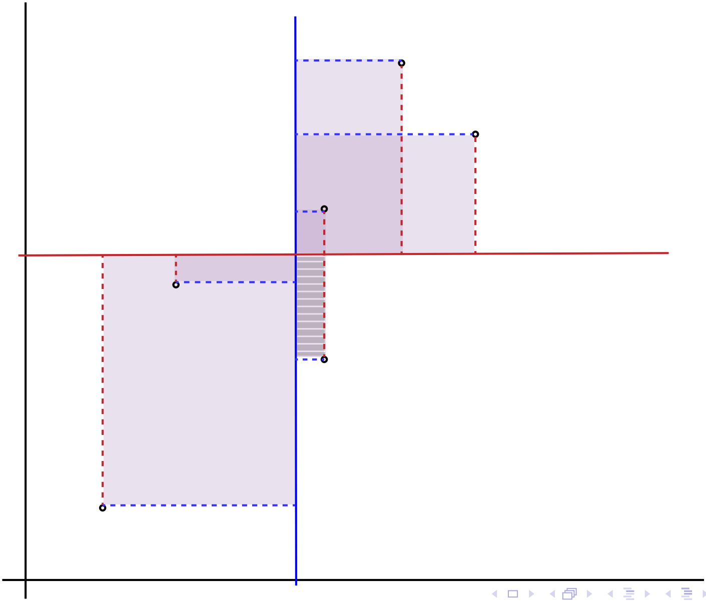
A Visual Intuition for the OLS Estimator



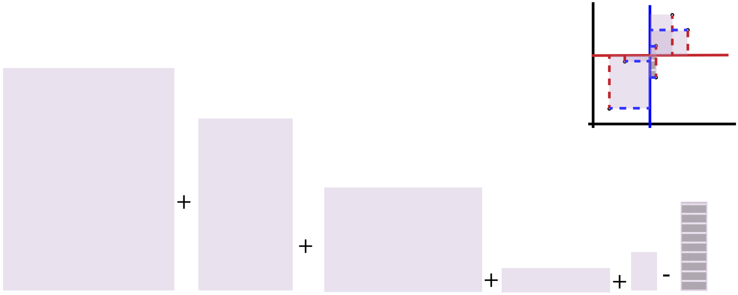
A Visual Intuition for the OLS Estimator



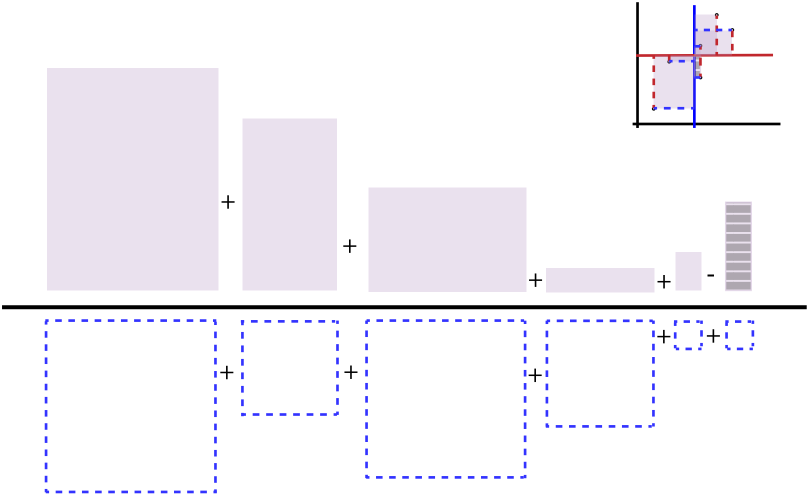
A Visual Intuition for the OLS Estimator



A Visual Intuition for the OLS Estimator



A Visual Intuition for the OLS Estimator



Mechanical properties of OLS

Mechanical properties of OLS

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

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.

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:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

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:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

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

$$\widehat{\text{cov}}(X_i, \widehat{u}_i) = 0$$

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:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

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

$$\widehat{\text{cov}}(X_i, \widehat{u}_i) = 0$$

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

$$\widehat{\text{cov}}(\widehat{Y}_i, \widehat{u}_i) = 0$$

OLS slope as a weighted sum of the outcomes

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.

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

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.

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

- Where here we have the weights, W_i as:

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

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.

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

- Where here we have the weights, W_i as:

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

- 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.

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.

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

- Where here we have the weights, W_i as:

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

- 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!

Sampling distribution of the OLS estimator

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

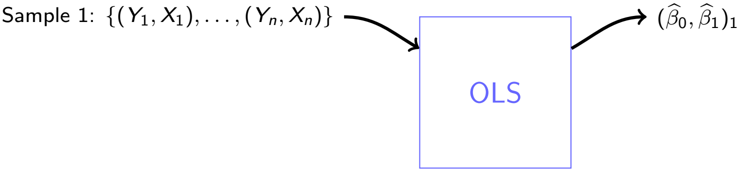
- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.

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.

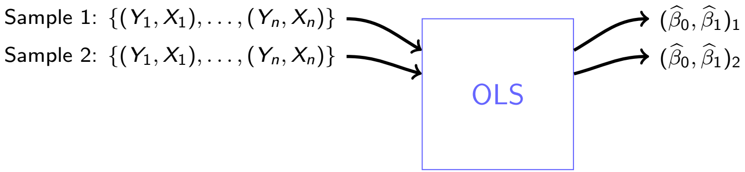
Sampling distribution of the OLS estimator

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



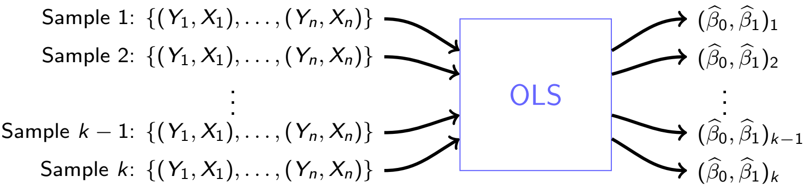
Sampling distribution of the OLS estimator

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



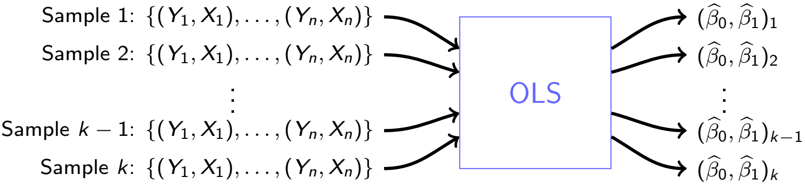
Sampling distribution of the OLS estimator

- 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

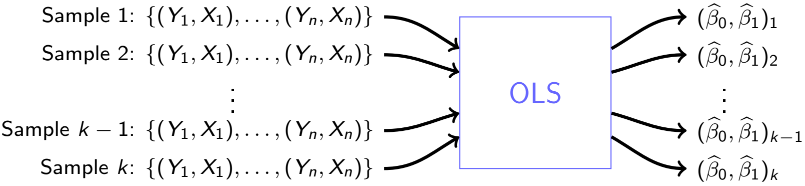
- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance

Sampling distribution of the OLS estimator

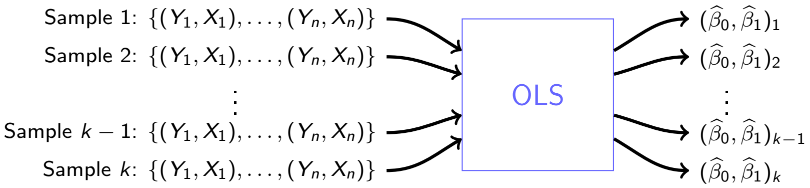
- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.



- 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

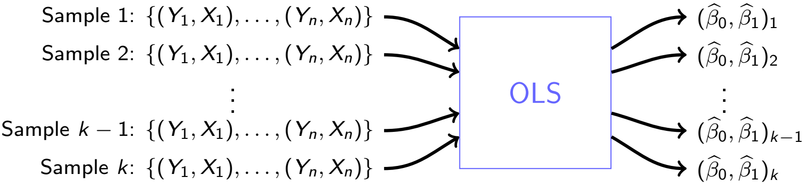
- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.



- 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

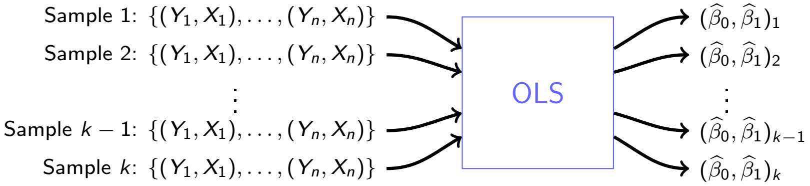
- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates.



- 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.



- 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

Simulation procedure

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

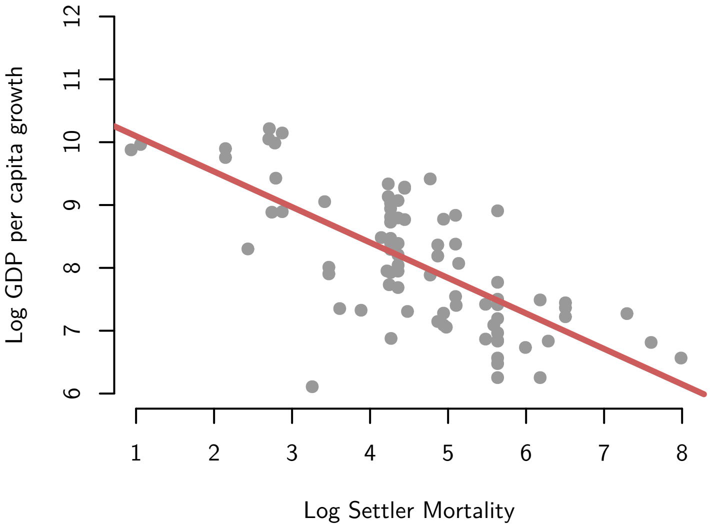
Simulation procedure

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

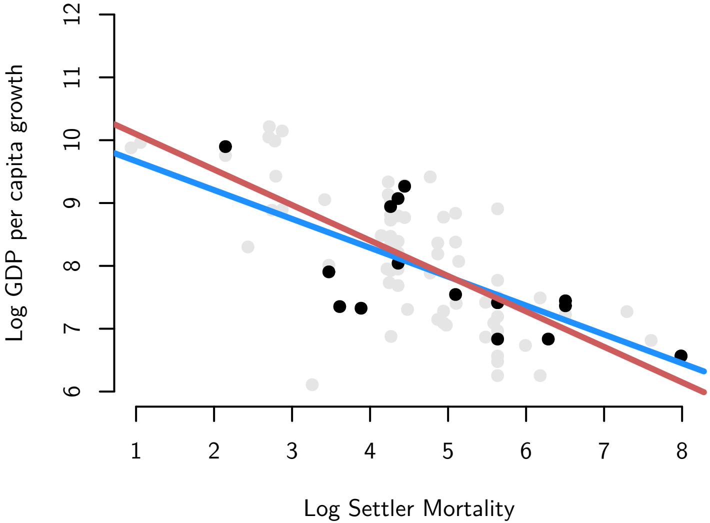
Simulation procedure

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

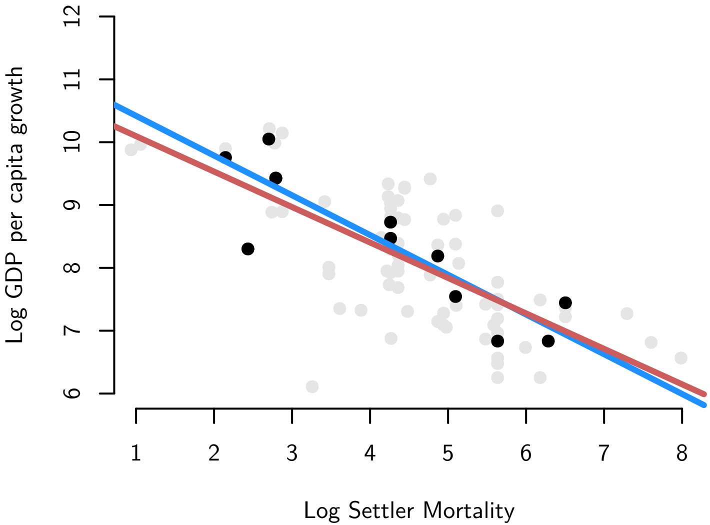
Population Regression



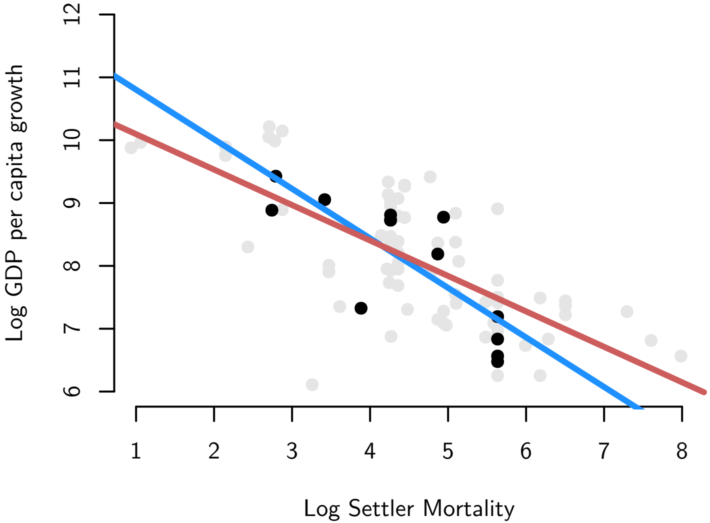
Randomly sample from AJR



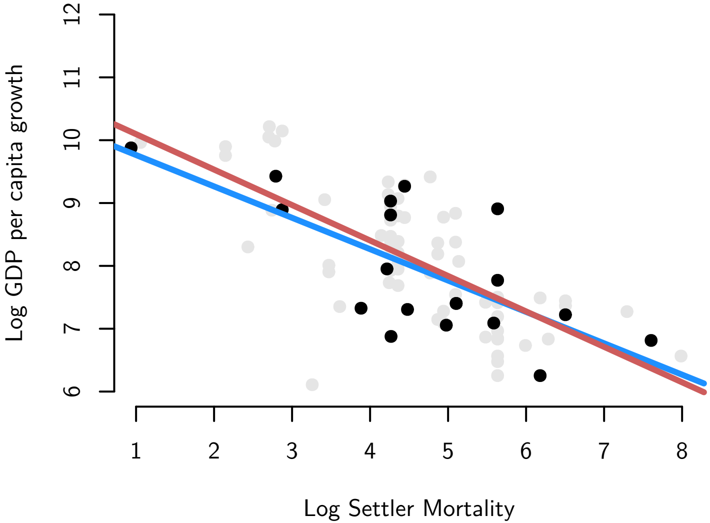
Randomly sample from AJR



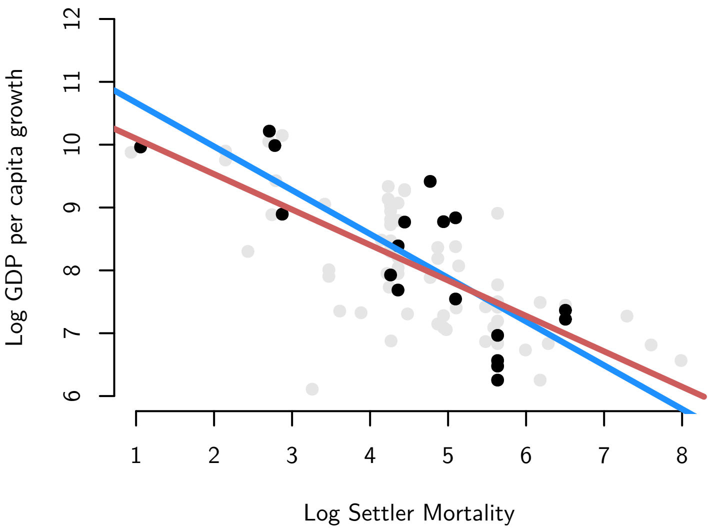
Randomly sample from AJR



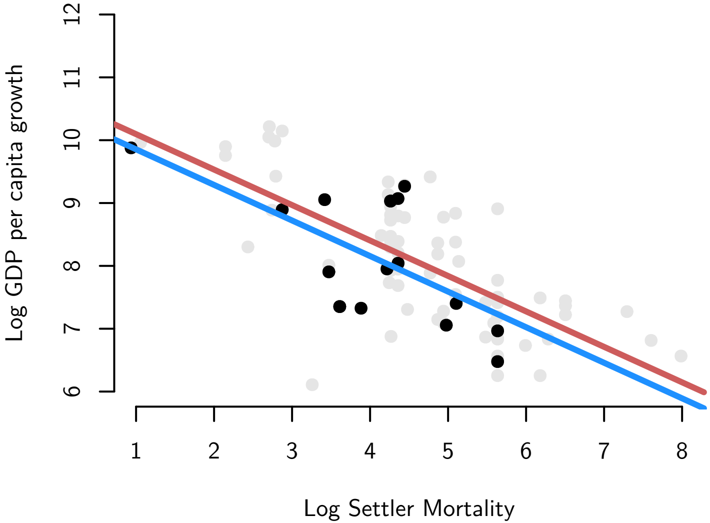
Randomly sample from AJR



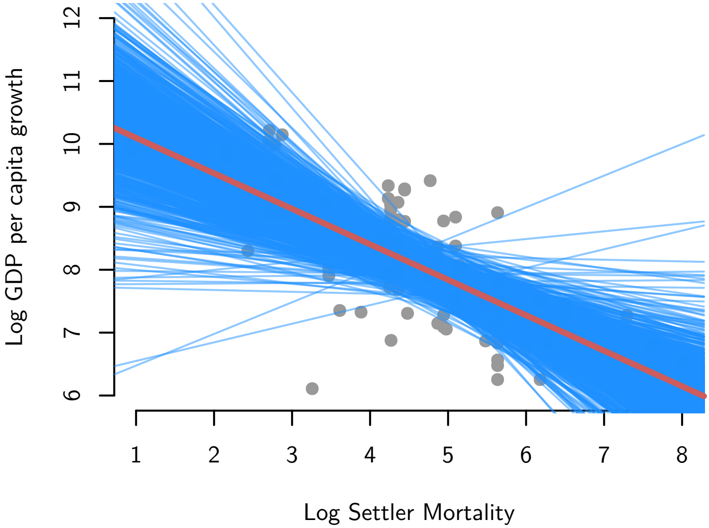
Randomly sample from AJR



Randomly sample from AJR



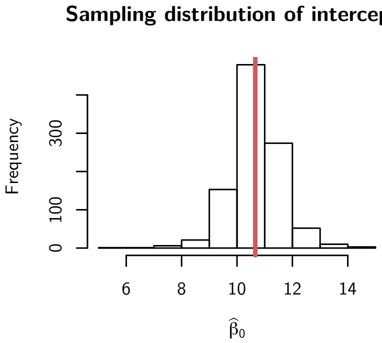
Randomly sample from AJR

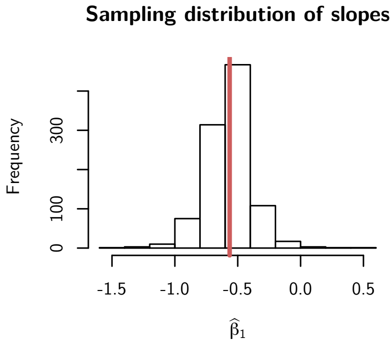


Sampling distribution of OLS

Sampling distribution of OLS

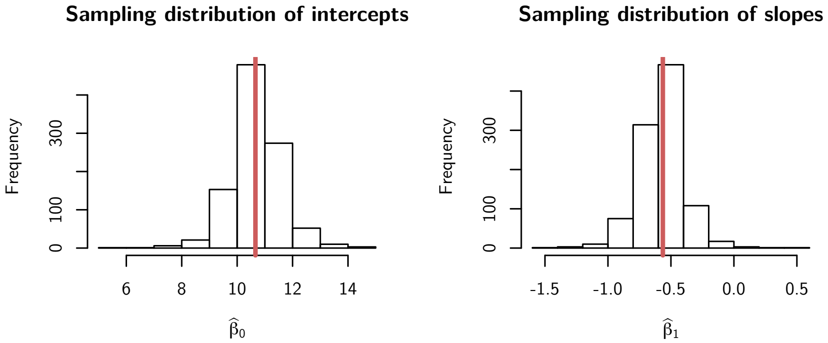
- You can see that the estimated slopes and intercepts vary from sample to sample, but that the "average" of the lines looks about right.





Sampling distribution of 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.



- Is this unique?

Assumptions for unbiasedness of the sample mean

Assumptions for unbiasedness of the sample mean

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

$$\mathbb{E}[\overline{X}] = \mu$$

Assumptions for unbiasedness of the sample mean

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

$$\mathbb{E}[\overline{X}] = \mu$$

- Just one: random sample

Assumptions for unbiasedness of the sample mean

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

$$\mathbb{E}[\overline{X}] = \mu$$

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

Our goal

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- We'll start with the mean of the sampling distribution. Is the estimator centered at the true value, β_1?

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- We'll start with the mean of the sampling distribution. Is the estimator centered at the true value, β_1?
- Most of our derivations will be in terms of the slope but they apply to the intercept as well.

OLS Assumptions Preview

OLS Assumptions Preview

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

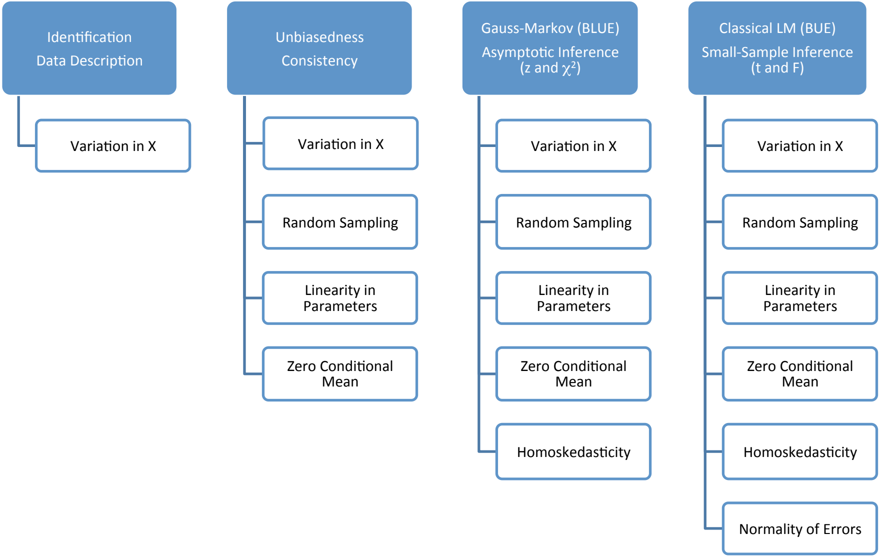
1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.
4. Zero conditional mean: 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



OLS Assumption I

OLS Assumption I

Assumption (I. Linearity in Parameters)

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

$$Y = \beta_0 + \beta_1 X_1 + u$$

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*

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
    $Y = \beta_0 + \beta_1 log(X) + u$

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
    $Y = \beta_0 + exp(\beta_1) X + u$

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
    $Y = \beta_0 + exp(\beta_1)X + u$
- β_0, β_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
    $Y = \beta_0 + exp(\beta_1) X + u$

- β_0, β_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
    $Y = \beta_0 + exp(\beta_1) X + u$
- β_0, β_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

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.

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:

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:

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

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:

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i, \ x_i) \textit{ for } i = 1, ..., n$$

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

Potential Violations:

- Time series data (regressor values may exhibit persistence)

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

Potential Violations:

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

OLS Assumption III

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.

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.

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:

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

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

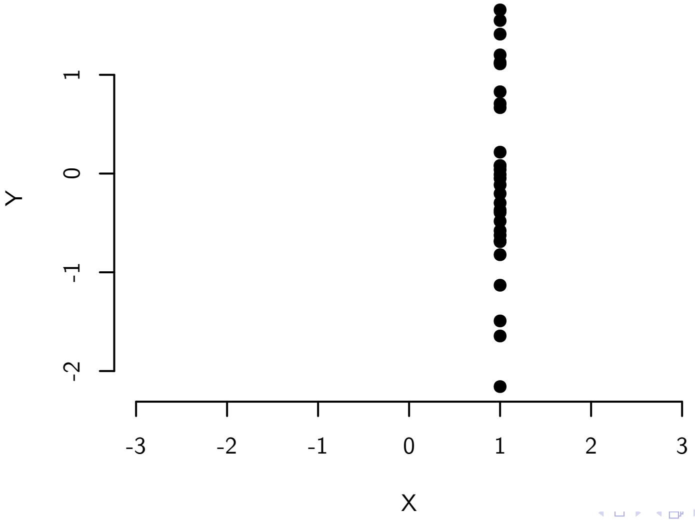
- Why does this matter?

Stuck in a moment

- Why does this matter?

Stuck in a moment

- 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

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:

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

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 $\text{Cov}(X, u) = 0$

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

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

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:

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

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, $\text{Cov}(X, u) \neq 0$

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

  $\longrightarrow$ It must be assumed $E[ability|educ = low] = E[ability|educ = high]$

Violating the zero conditional mean assumption

Violating the zero conditional mean assumption

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

$$Y_i = 1 + 0.5X_i + u_i$$

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

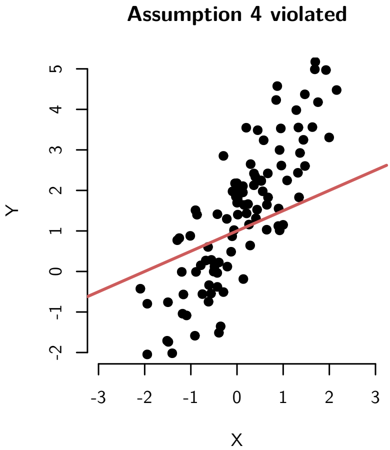
How does this assumption get violated? Let's generate data from the following model:

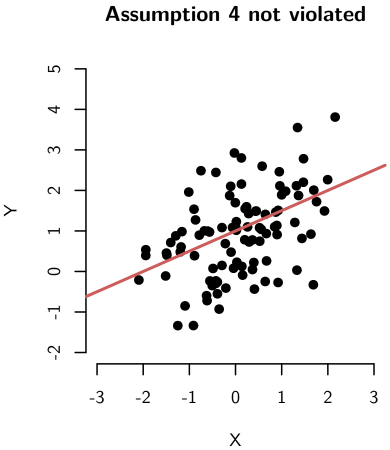
$$Y_i = 1 + 0.5X_i + u_i$$

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



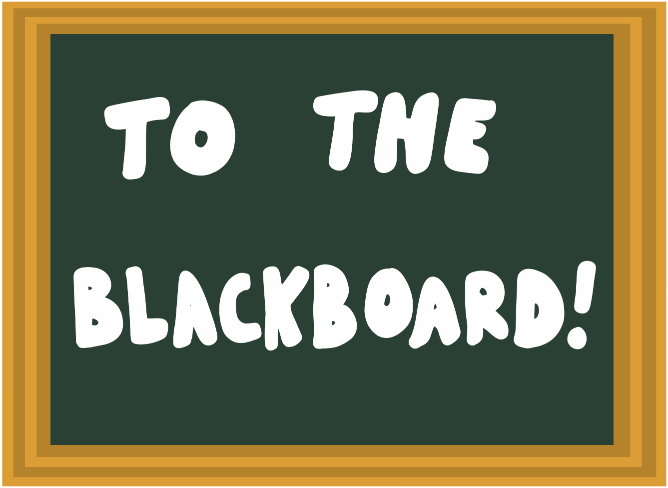


Unbiasedness (to the blackboard)

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

Unbiasedness (to the blackboard)

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



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 β_1 and β_0.

Where are we?

Where are we?

- Now we know that, under Assumptions 1-4, we know that

$$\widehat{\beta}_1 \sim ?(\beta_1, ?)$$

Where are we?

- Now we know that, under Assumptions 1-4, we know that

$$\widehat{\beta}_1 \sim ?(\beta_1, ?)$$

- 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

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
4. Zero conditional mean of the errors

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
4. Zero conditional mean of the errors
5. Homoskedasticity

Variance of OLS Estimators

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:

Variance of OLS Estimators

How can we derive Var$[\hat{\beta}_0]$ and 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:

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

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:

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

- 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)

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:

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

- 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:

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

- 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:

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

- 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:

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

- 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:

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

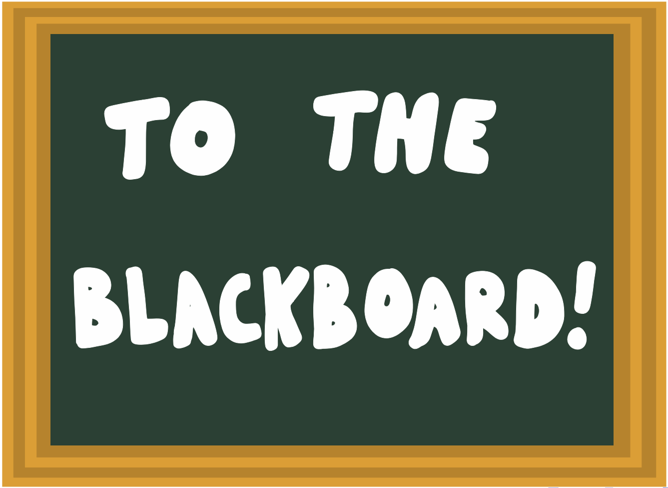
- Violation: $\text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2]$ called heteroskedasticity.
- Assumptions I–V are collectively known as the Gauss-Markov assumptions

Deriving the sampling variance

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

Deriving the sampling variance

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



Variance of OLS Estimators

Theorem (Variance of OLS Estimators)

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

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

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

where $Var[u \mid X] = \sigma_u^2$ *(the error variance).*

Understanding the sampling variance

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

- What drives the sampling variability of the OLS estimator?

Understanding the sampling variance

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

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of Y_i, the higher the sampling variance

Understanding the sampling variance

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

- 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

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

- 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 $Var[u] = \sigma_u^2$?

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
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 $\hat{u}_i$.

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
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:

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

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
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:

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

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
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:

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

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

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

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:

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

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 $\hat{\beta}_1$.

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ? \left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)$$

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

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ? \left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)$$

- 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...

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression
- This Week
  - Monday:

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - what is regression
- This Week
  - Monday:
    - mechanics 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

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

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?

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

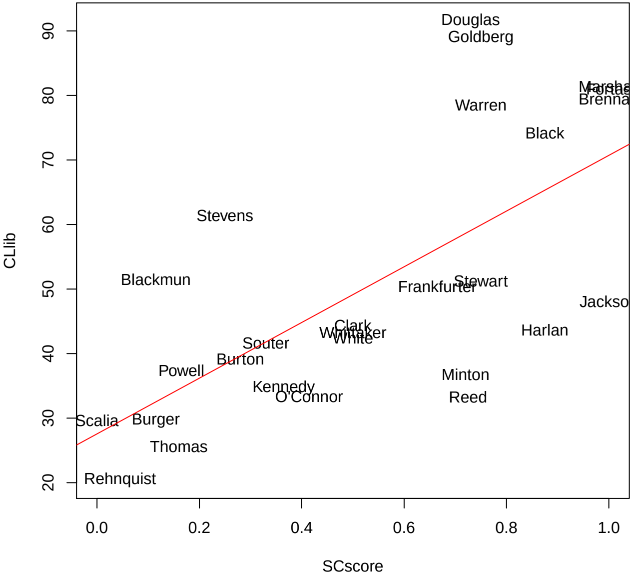
- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
- Percentage of votes in liberal direction for each justice in a number of issue areas

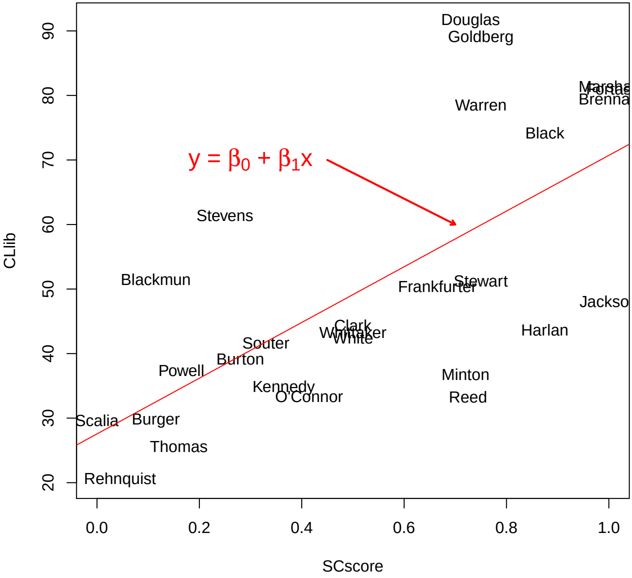
Example: Epstein and Mershon SCOTUS data

- 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

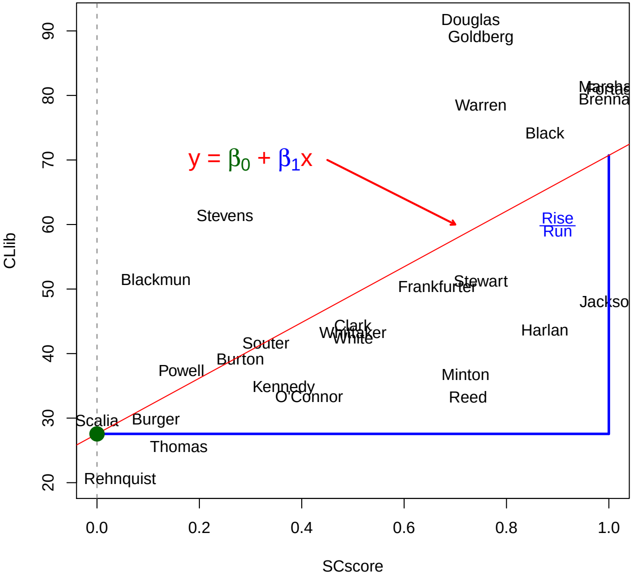
Example: Epstein and Mershon SCOTUS data

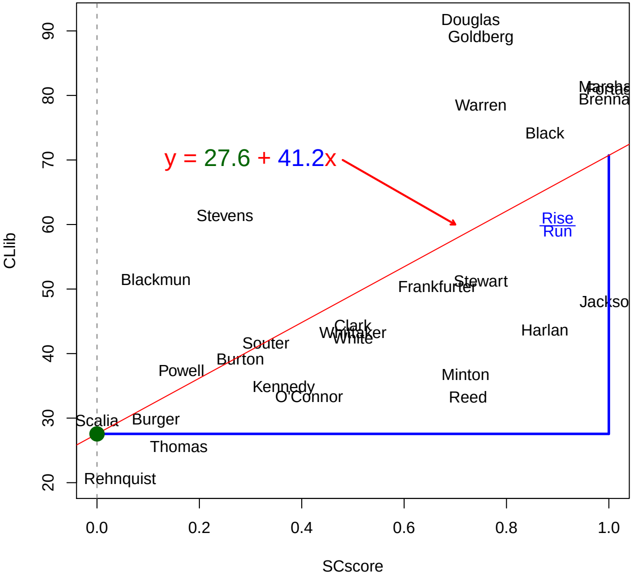
- Data on 27 justices from the Warren, Burger, and Rehnquist courts (can be interpreted as a census)
- Percentage of votes in liberal direction for each justice in a number of issue areas
- Segal-Cover scores for each justice
- Party of appointing president


Douglas
Goldberg
Marsha
Fortas
Brenna
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
CLib
SCscore
20
30
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
1.0

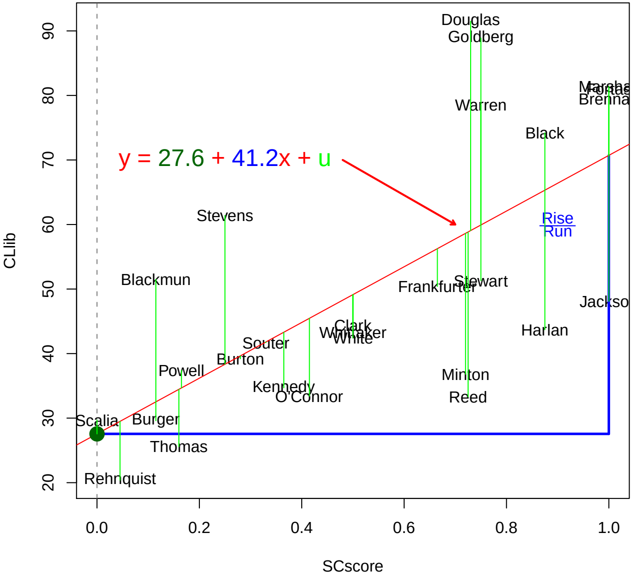


$y = \beta_0 + \beta_1 x$
Douglas
Goldberg
Marsha
Fortas
Brenna
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
CLib
SCscore
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90


$y = \beta_0 + \beta_1 x$
Rise
Run
Douglas
Goldberg
Brenna
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
CLib
SCscore
20
30
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
1.0



y = 27.6 + 41.2x
Rise
Run
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Scalia
Burger
Thomas
Rehnquist
CLib
SCscore
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90


y = 27.6 + 41.2x + u
Rise
Run
CLib
SCscore
Douglas
Goldberg
Warren
Black
Brenna
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90

How to get β_0 and β_1

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

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

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 χ^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

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

- Now we know the mean and sampling variance of the sampling distribution.
- How does this compare to other estimators for the population slope?

OLS is BLUE :(

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.*

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”

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

OLS is BLUE :(

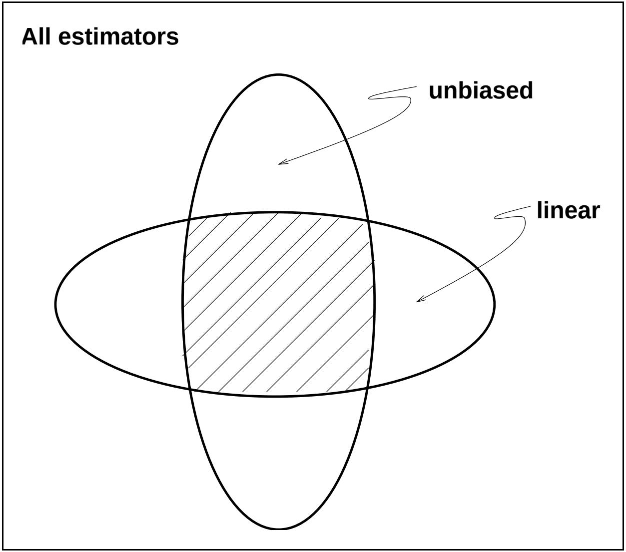
Theorem (Gauss-Markov)

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"
- 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



Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ is the lowest variance of any linear estimator of β_1

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ is the lowest variance of any linear estimator of β_1
- What about the last question mark? What's the form of the distribution?

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ is the lowest variance of any linear estimator of β_1
- What about the last question mark? What's the form of the distribution?

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

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

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

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

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

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

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

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

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

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

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

Large-sample distribution of OLS estimators

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

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

- 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:

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

- 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:

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

- 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:

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

- 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:

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

Large-sample distribution of OLS estimators

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

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

- 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:

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

- Can also replace SE with an estimate:

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

Where are we?

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

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:

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

OLS Assumptions VI

Assumption (VI. Normality)

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

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

Note: This implies homoskedasticity and zero conditional mean.

OLS Assumptions VI

Assumption (VI. Normality)

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

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

Note: This implies homoskedasticity and zero conditional mean.

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

OLS Assumptions VI

Assumption (VI. Normality)

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

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

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)

OLS Assumptions VI

Assumption (VI. Normality)

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

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

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

OLS Assumptions VI

Assumption (VI. Normality)

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

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

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
- We don't need normality assumption in large samples

Sampling Distribution for $\widehat{\beta}_1$

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

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

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

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

Proof.

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

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

Proof.

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

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

Any linear combination of independent normals is

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

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

Proof.

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

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

Any linear combination of independent normals is normal,

Sampling Distribution for $\widehat{\beta}_1$

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

Under Assumptions I–VI,

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

where

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

which implies

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

Proof.

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

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

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:

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

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:

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

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

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

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:

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

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

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

- 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:

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

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

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

- 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 σ_u^2
- Replace σ_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 β_1 has a t-distribution with $n-2$ degrees of freedom:

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

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 σ_u^2
- Replace σ_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 β_1 has a t-distribution with $n-2$ degrees of freedom:

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

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

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

Where are we?

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

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

Where are we?

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

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

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

Now let's briefly return to some of the large sample properties.

Large Sample Properties: Consistency

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.

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* $n \to \infty$.

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 $\widehat{\beta}_1$ is consistent for β_1 as $n \to \infty$:

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1$$

- Technical note: We can slightly relax Assumption IV:

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

to its implication:

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

for consistency to hold (but not unbiasedness).

Large Sample Properties: Consistency

Proof.

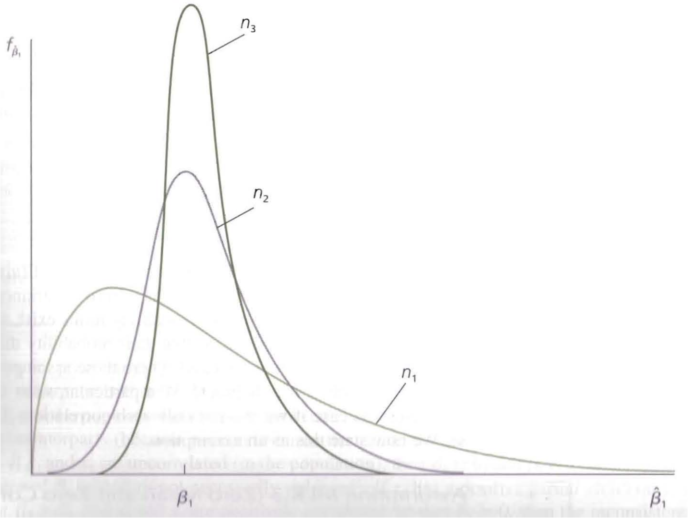
Similar to the unbiasedness proof:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_i^n (x_i - \bar{x}) u_i}{\sum_i^n (x_i - \bar{x})^2} \\
\text{plim } \widehat{\beta}_1 &= \text{plim } \beta_1 + \text{plim } \frac{\sum_i^n (x_i - \bar{x}) u_i}{\sum_i^n (x_i - \bar{x})^2} \quad \text{(Wooldridge C.3 Property i)} \\
&= \beta_1 + \frac{\text{plim } \frac{1}{n} \sum_i^n (x_i - \bar{x}) u_i}{\text{plim } \frac{1}{n} \sum_i^n (x_i - \bar{x})^2} \quad \text{(Wooldridge C.3 Property iii)} \\
&= \beta_1 + \frac{\text{Cov}[X, u]}{\text{Var}[X]} \quad \text{(by the law of large numbers)} \\
&= \beta_1 \quad (\text{Cov}[X, u] = 0 \text{ and } \text{Var}[X] > 0)
\end{aligned}$$

□

- 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



Sampling distributions of $\hat{\beta}_1$, for sample sizes $n_1 < n_2 < n_3$

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 $\widehat{\beta}_1$ is asymptotically normally distributed:

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

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:

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma_u^2$$

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. □

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 β_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 β_1
2. Hypothesis Testing: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α 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 β_1
2. Hypothesis Testing: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α level
3. Interval Estimation: Consider the sampling distribution of an interval estimator to construct intervals that will contain β_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 β_1
2. Hypothesis Testing: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α level
3. Interval Estimation: Consider the sampling distribution of an interval estimator to construct intervals that will contain β_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 $\hat{\beta}_1$.

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.

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

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: $H_a : \beta_1 \neq 0$

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:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

- 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:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- 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:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- 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, α, and find rejection regions that correspond to that value under the null distribution:

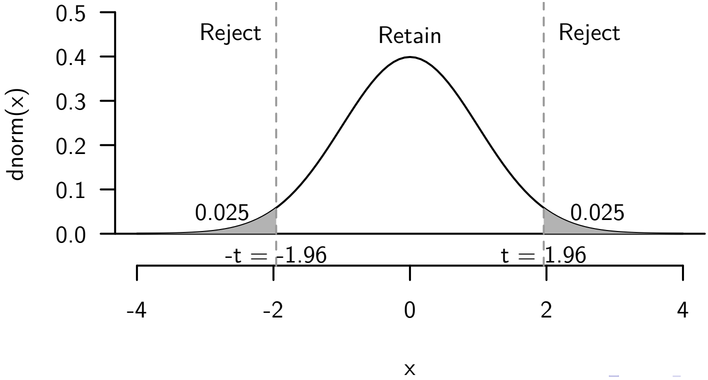
$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

Rejection region

- Choose a level of the test, α, 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.



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

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:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

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:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

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:

$$\mathbb{P}\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2} \right) = 1 - \alpha$$

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:

$$\mathbb{P}\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2} \right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \right) = 1 - \alpha$$

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:

$$\mathbb{P}\left(- t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]\right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]$$

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:

$$\mathbb{P}\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]\right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

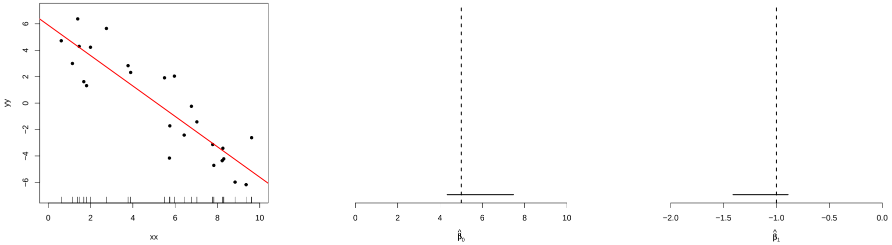
$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]$$

- We can derive these for the intercept as well:

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_0]$$

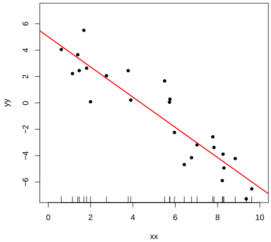
CIs Simulation Example

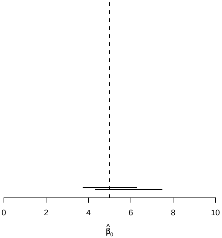
Returning to our simulation example we can simulate the sampling distributions of the 95 % confidence interval estimates for $\widehat{\beta}_1$ and $\widehat{\beta}_0$

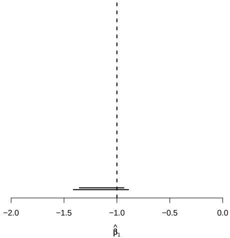


CIs Simulation Example

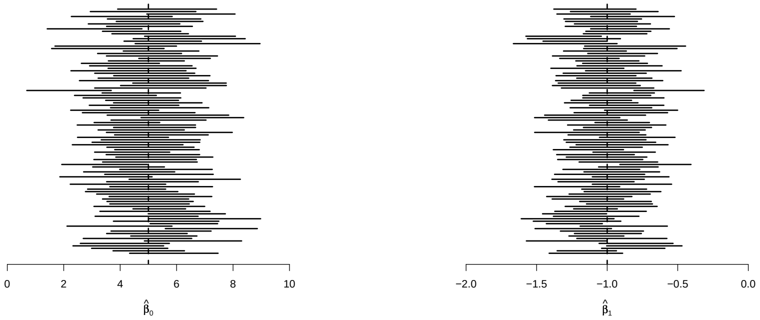
Returning to our simulation example we can simulate the sampling distributions of the 95 % confidence interval estimates for $\widehat{\beta}_1$ and $\widehat{\beta}_0$



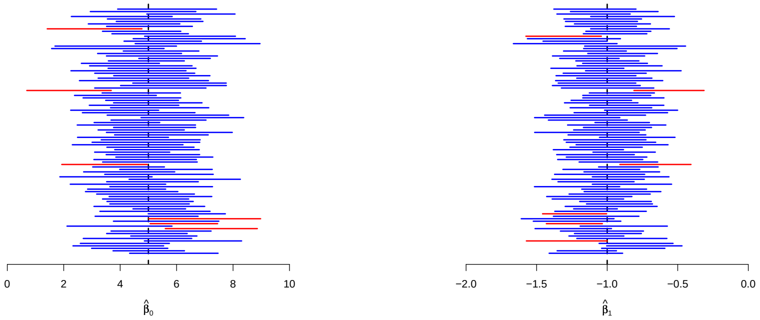




CIs Simulation Example



CIs Simulation Example



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.

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:

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Prediction error

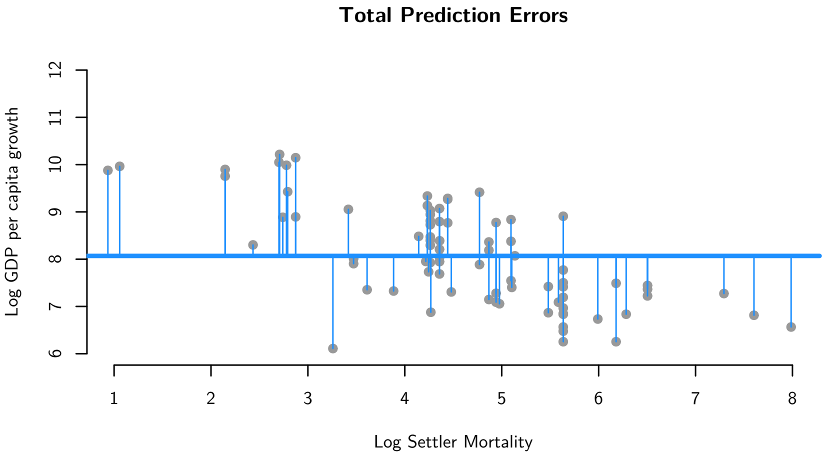
- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting Y once we include X into the regression model.
- Prediction errors without X: best prediction is the mean, so our squared errors, or the total sum of squares (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

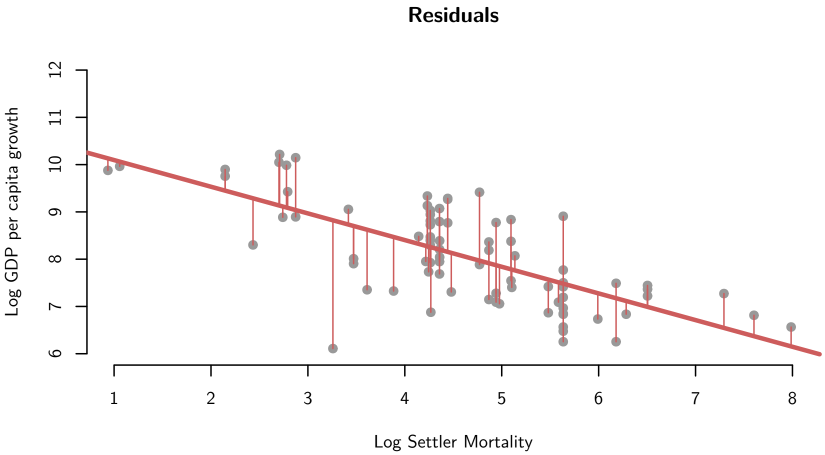
- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Sum of Squares



Sum of Squares



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-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:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

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.

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-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

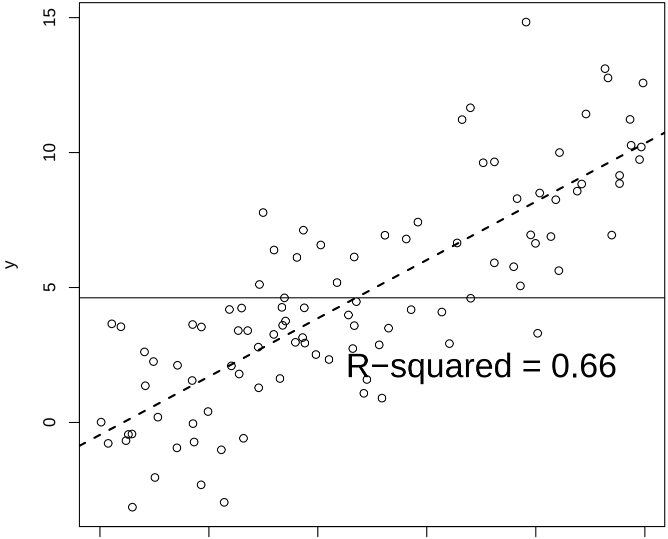
R-square

- 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:

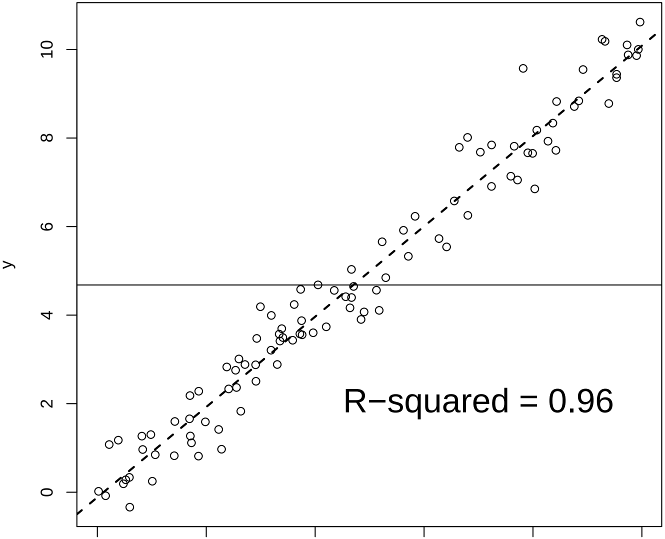
$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- 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

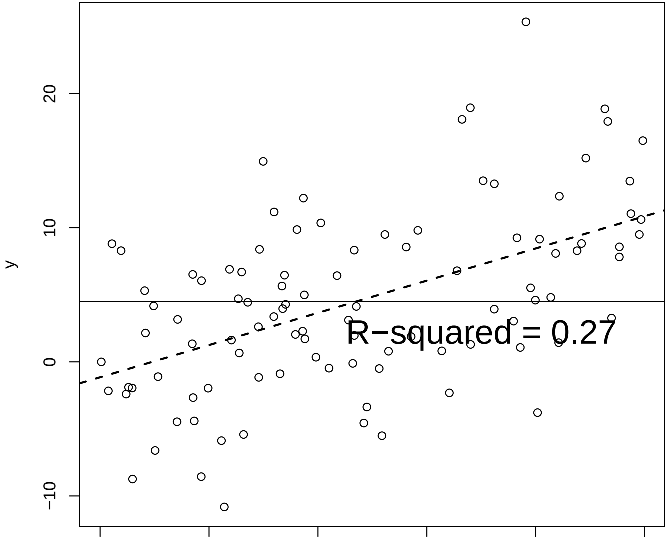
Is R-squared useful?



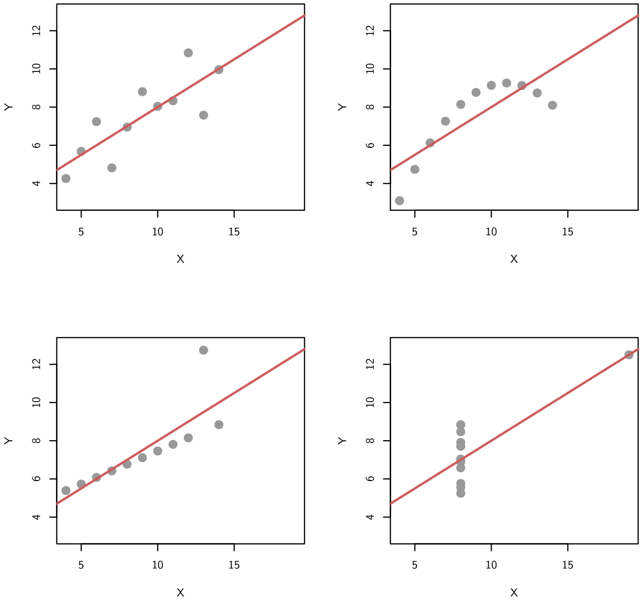
Is R-squared useful?



Is R-squared useful?



Is R-squared useful?



Why r^2?

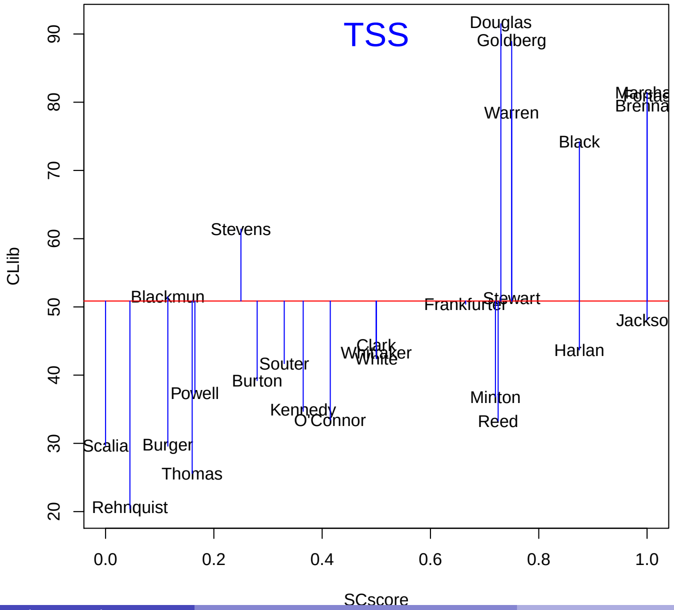
To calculate r^2, we need to think about the following two quantities:

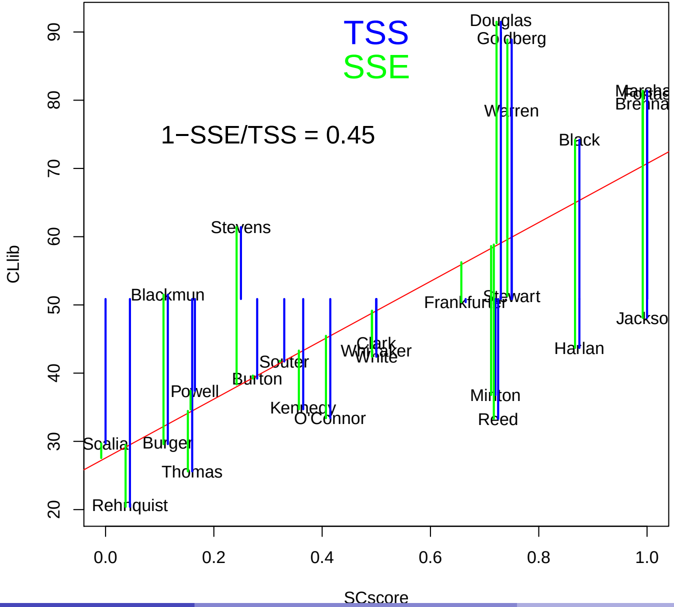
1. TSS: Total sum of squares
2. SSE: Sum of squared errors

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

$$SSE = \sum_{i=1}^{n} u_i^2.$$

$$r^2 = 1 - \frac{SSE}{TSS}.$$


TSS
CLlib
SCscore
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90
Douglas
Goldberg
Warren
Black
Brenna
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Harlan
Clark
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist


TSS
SSE
1−SSE/TSS = 0.45
CLlib
SCscore
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Powell
Burger
Scalia
Thomas
Rehnquist
Souter
Burton
Kennedy
O'Connor
Clark
Whittaker
White
Frankfurter
Stewart
Minton
Reed
Harlan
Jackso
Brenna
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90

Derivation

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \quad \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2$$

Derivation

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\ &= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\}\end{aligned}$$

Derivation

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\ &= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\ &= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2\end{aligned}$$

Derivation

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2
\end{aligned}$$

Derivation

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSE + RegSS
\end{aligned}$$

Derivation

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSE + RegSS
\end{aligned}
$$

Derivation

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSE + RegSS
\end{aligned}$$

Derivation

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\{\hat{u}_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{n}\{\hat{u}_i^2 + 2\hat{u}_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSE + RegSS
\end{aligned}
$$

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSE}{TSS} + \frac{RegSS}{TSS} = 1$$

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:

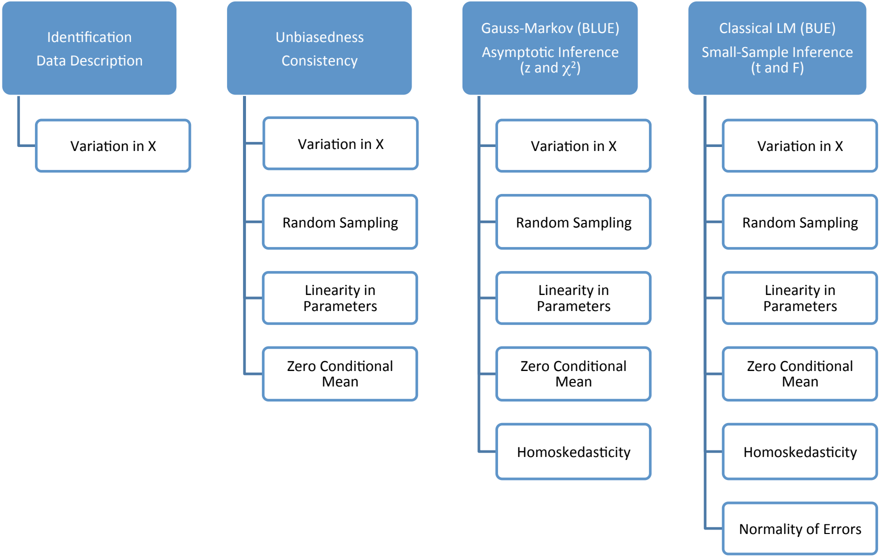
$$\frac{SSE}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSE}{TSS} + \frac{RegSS}{TSS} = 1$$

$$\frac{RegSS}{TSS} = 1 - \frac{SSE}{TSS} = r^2$$

r^2 is a measure of how much of the variation in Y is accounted for by X.

OLS Assumptions Summary



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:

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:

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

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

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

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

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?

What Do the Regression Coefficients Mean Substantively?

3. Causal inference:

What Do the Regression Coefficients Mean Substantively?

3. Causal inference:
   - Inference about counterfactuals, i.e. hypothetical interventions to the same units

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?

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)

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

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

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 β 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)

OLS as a Best Linear Predictor (Review of BLUE)

- Suppose that we want to predict the values of Y given observed X values

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)

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.

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 $MSE = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$

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* $\sum_{i=1}^{n} \hat{u}^2$!
- 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:

$$MSE = E\left[(Y_i - \hat{Y}_i)^2\right] = E\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2\right] = E\left[\hat{u}_i^2\right]$$

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* $\sum_{i=1}^{n} \hat{u}^2$!
- This implies that OLS is the best linear predictor in terms of MSE
- Which assumptions did we use to get this result?

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* $\sum_{i=1}^{n}\hat{u}^2$!
- 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)

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* $\sum_{i=1}^{n} \hat{u}^2$!
- 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

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.

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

Reporting Statistical Significance

- A reasonable way to think about statistical significance is to think about the precision of the estimates

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.

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:

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

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

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

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:

  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

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:

  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)

Fun with Non-Linearities

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first transforming X appropriately

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- 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 β)
- 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 β_1:

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- 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 β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ 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 β)
- 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 β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in Y associated with one unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- 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 β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in Y associated with one unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X
  - Note that these approximations work only for small increments

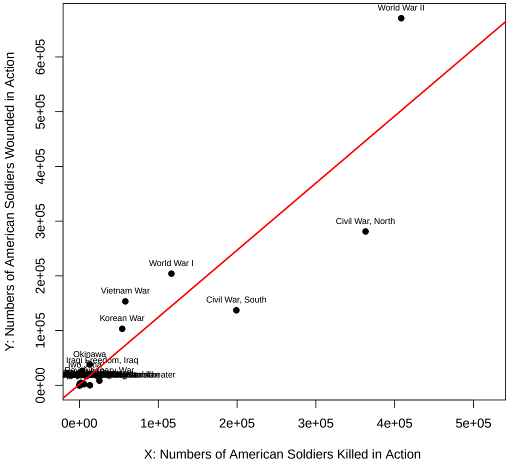
Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first transforming X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

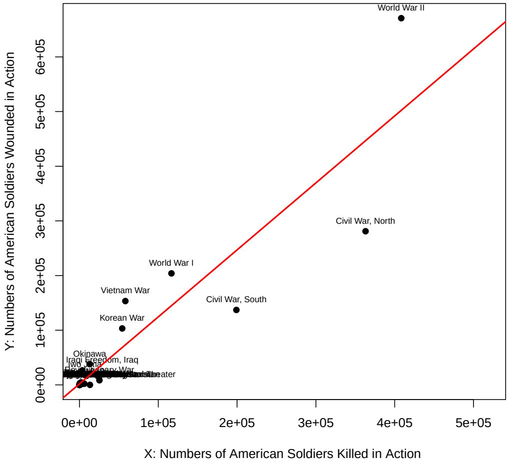
- The log transformation changes the interpretation of β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in Y associated with one unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X
  - Note that these approximations work only for small increments
  - In particular, they do not work when X is a discrete random variable

Example from the American War Library



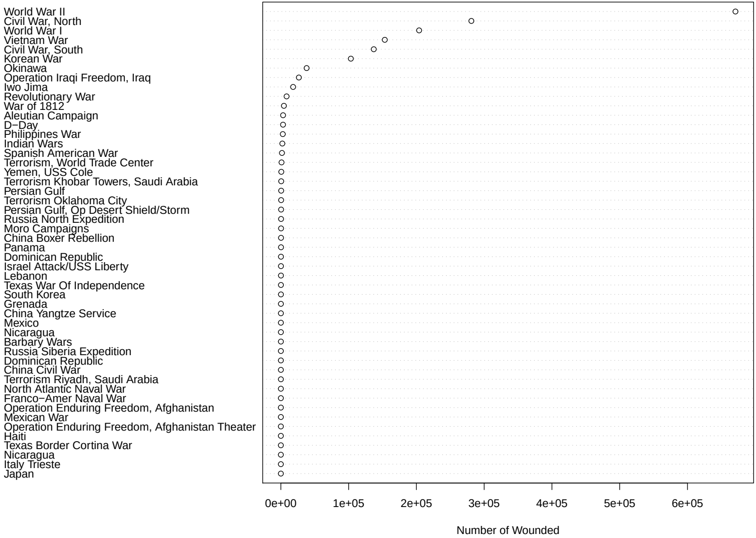
$\hat{\beta}_1 = 1.23 \longrightarrow$

Example from the American War Library

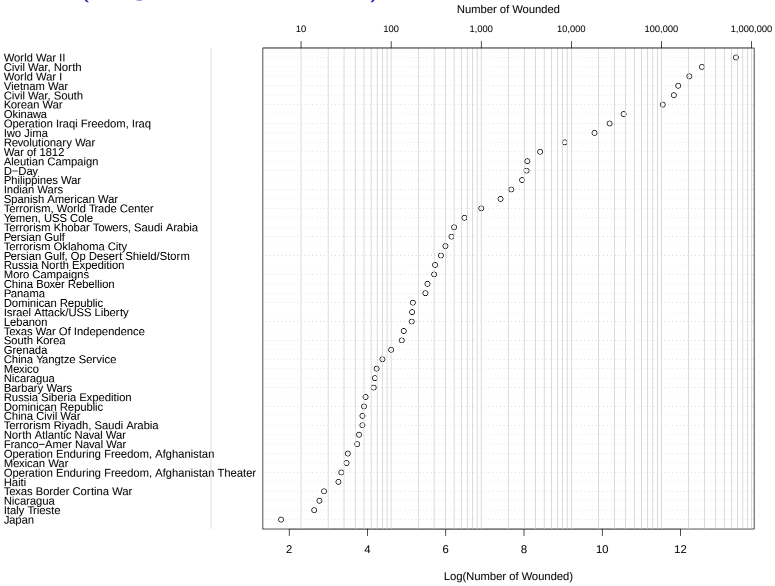


$\hat{\beta}_1 = 1.23 \longrightarrow$ One additional soldier killed predicts 1.23 additional soldiers wounded on average

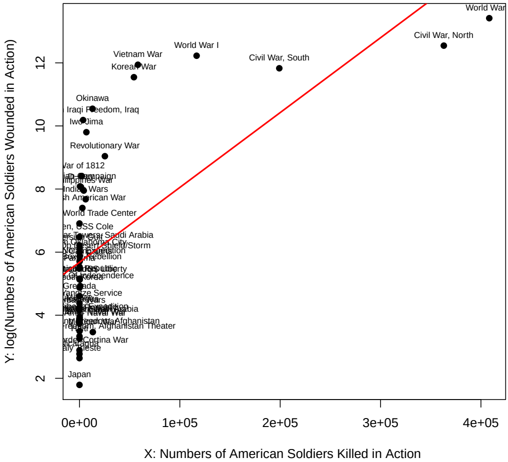
Wounded (Scale in Levels)



Wounded (Logarithmic Scale)

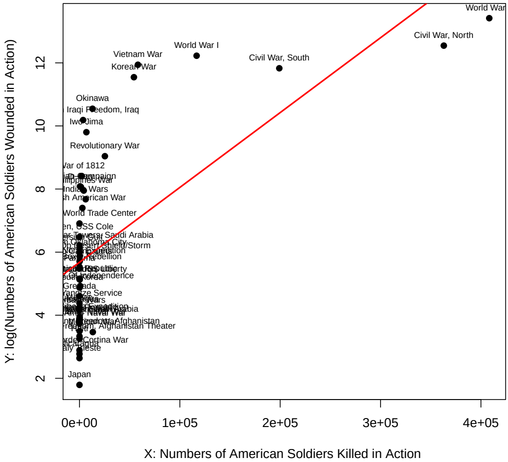


Regression: Log-Level



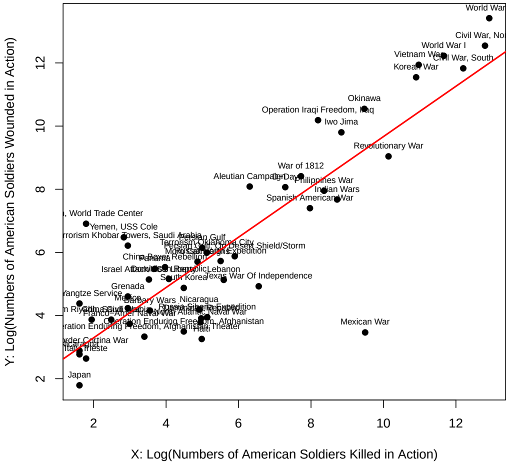
$\hat{\beta}_1 = 0.0000237 \longrightarrow$

Regression: Log-Level



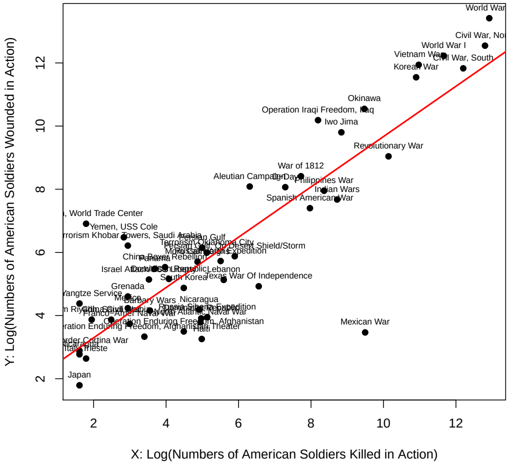
$\hat{\beta}_1 = 0.0000237 \longrightarrow$ One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded on average

Regression: Log-Log



$\hat{\beta}_1 = 0.797 \longrightarrow$

Regression: Log-Log



$\hat{\beta}_1 = 0.797 \longrightarrow$ A percent increase in deaths predicts 0.797 percent increase in the wounded on average

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." 2000.

Wooldridge, Jeffrey. 2000. *Introductory Econometrics.* New York: South-Western.