The Simple Linear Regression Model

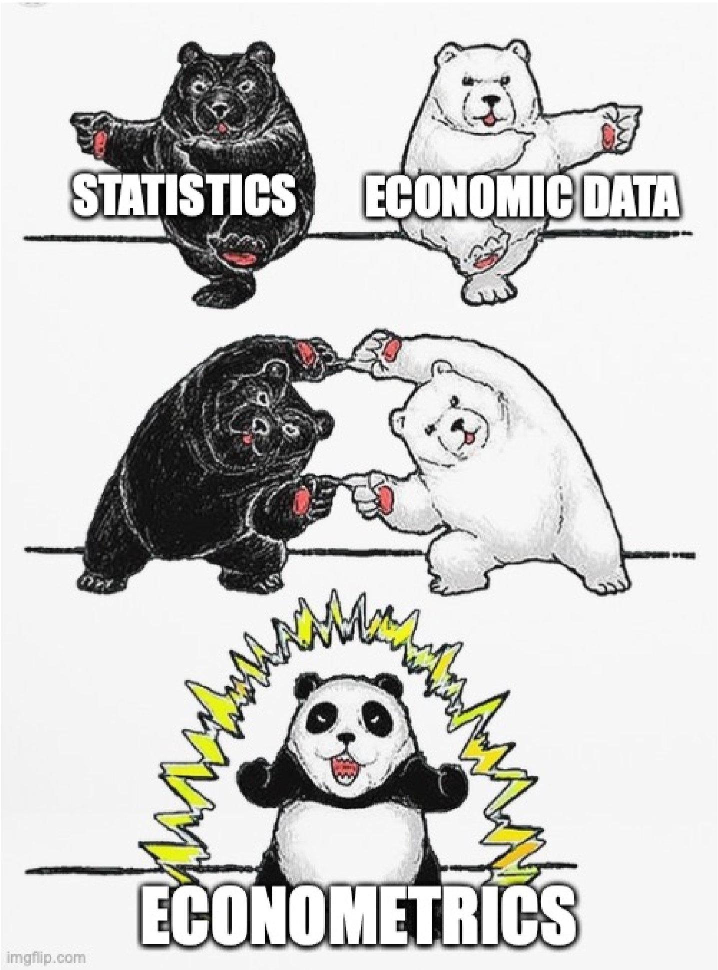
EC295

Justin Smith

Wilfrid Laurier University

Fall 2022

What is Econometrics?



What is Econometrics

- Defining characteristics of econometrics
  - Observational data
  - Use of regression analysis
- Motivating statistical models with economic models
  - Focus on causality
- This class introduces you to linear regression
  - Building block for many future economics classes
  - You will use this technique in EC481

Introduction to Linear Regression

- Economic analysis often involves relating two or more variables
  - Does age of school entry affect test scores?
  - Does childhood health insurance affect adult health?
  - Does foreign competition affect domestic innovation?
- These relationships are typically used for
  - **Causal Inference**: the independent effect of one variable on another
  - **Prediction**: estimating value of one variable given values of another
- Which one you use depends on goals of your analysis
  - Causal inference is important in policy analysis
  - Prediction is useful for guessing unknown values of a variable
- We will develop a model to use for these goals

Context

- A big issue in education is the size of school classes
- Parents often in favour of smaller classes
  - More attention paid to individual students
  - Classes easier to control
  - Can do more interactive work
- But, smaller classes are more expensive
  - More teaching resources per student
- Important to measure benefit of smaller classes
  - Compare against cost to see if worthwhile
- Book repeatedly discusses models in context of class size and student performance

What Are We Trying to Model?

- We want to relate test scores to class size
- Hard to do this for specific individuals
    - Many reasons why test scores differ between people
    - Even people in same class sizes have very different scores
- Instead focus on the systematic relationship
- We do this by focusing on average test scores
    - How do average test scores change with class size?
- Several reasons to use the average
    - Highlights systematic patterns between variables
    - It is mathematically optimal way to predict a variable given another
    - Intuitively appealing

What Are We Trying to Model?

- Mathematically we focus on the **Conditional Expectation**
- In the context of test scores, the conditional expectation is

$$E[TestScore|STR]$$

- This is the average test score for each class size
- STR is Student Teacher Ratio, a measure of class size

Reminder about Expected Values

The **Expected Value** $E[Y]$ of a random variable Y is its weighted average

The **Conditional Expectation** $E[Y|X]$ is the weighted average of a variable Y at specific values of another variable X

What Are We Trying to Model?

- Big problem: we do not know how average test scores relate to class size
    - Could be linear
    - Could be non-linear
    - Could some other weird function
- Unfortunately, we will **never know** exactly how they relate 😭
- Instead we approximate this relationship
- In EC295 our we use linear models for the approximation
    - Often a good guess at true relationship
    - But unknown true model is probably more complicated

The Linear Regression Model

- A linear model relating test scores to each class size is

$$TestScore = \beta_0 + \beta_{STR} STR + u$$

- Several important components of this model
  - $TestScore$ are individual test scores
  - STR are individual class sizes
  - β_{STR} is the slope
    - Effect of one-unit change in class size on test scores
  - β_0 is the intercept parameter
    - Test scores when class size is zero
  - u is everything except class size that determines test scores

The Linear Regression Model

This model breaks test scores in to two pieces

1. Population Regression Function

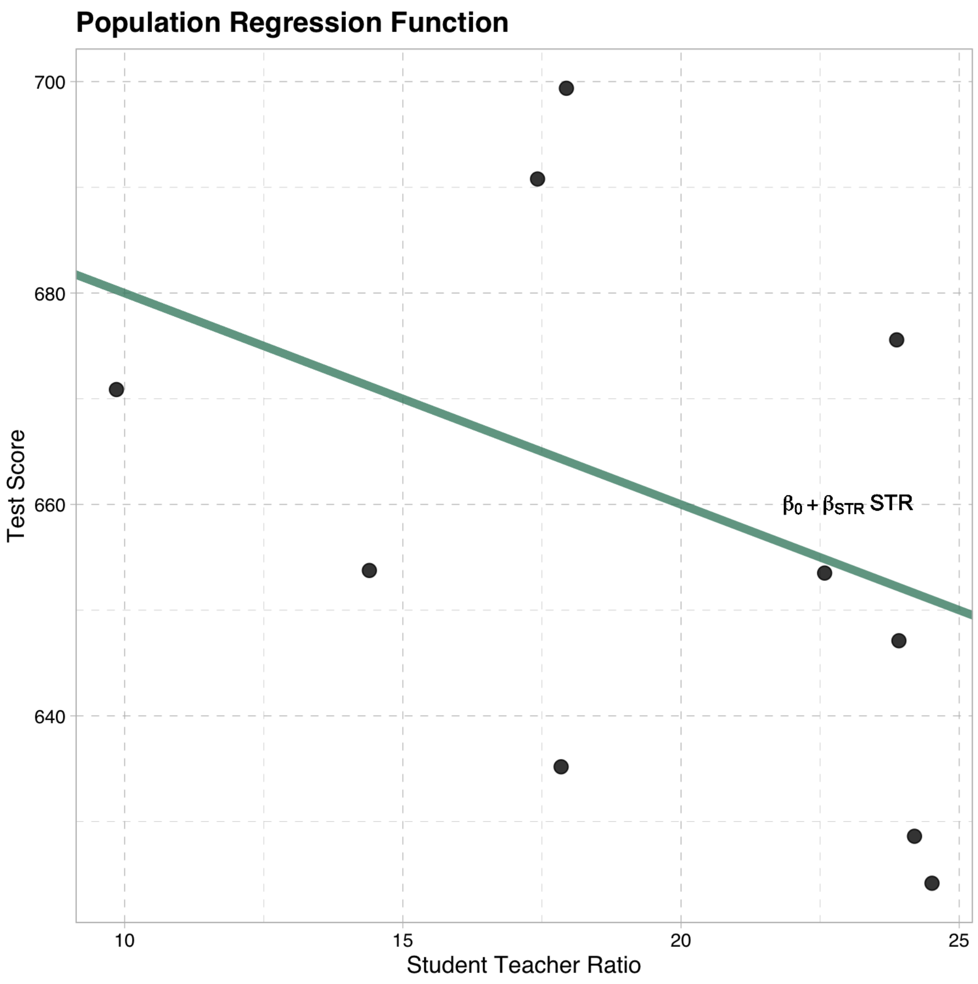
$$\beta_0 + \beta_{STR}STR$$

   The predictable part of test scores

2. Error Term

$$u = TestScore - \beta_0 - \beta_{STR}STR$$

   The unobserved and unpredictable part of test scores



The Linear Regression Model

- Another big problem: We do not know the values of β_0 and β_{STR}
  - They are parameters that we do not observe
- We also do not observe u
  - The unobserved error term
- Suppose we need to know these parameters
- How do we proceed from here?
- Answer: we **estimate** β_0 and β_{STR} with a sample of data
  - There are several estimation methods
  - We will focus on **Ordinary Least Squares (OLS)**

Drawing a Sample from the Population

- To estimate our model, we need to collect data on test scores and class sizes
- Imagine collecting a sample of size n
  - e.g. test scores and class sizes from 50 classes in different schools
  - $n = 50$ in this case
- The population regression model holds for each member of the sample

$$TestScore_i = \beta_0 + \beta_{STR} STR_i + u_i$$

  - The subscript i identifies a specific member of the sample
- Test scores are assumed to be linearly related to class size for each member of the sample

Ordinary Least Squares

Ordinary Least Squares

A method that estimates regression parameters by choosing the ones that minimize the sum of the squared distance between the estimated regression line and each data point

- To implement OLS, replace the unknowns of the population model with estimates

$$TestScore_i = \hat{\beta}_0 + \hat{\beta}_{STR} STR_i + \hat{u}_i$$

  - $\hat{\beta}_0$ estimates β_0
  - $\hat{\beta}_{STR}$ estimates β_{STR}
  - $\hat{u}_i$ is the residual (estimates the error)

- OLS chooses $\hat{\beta}_0$ and $\hat{\beta}_{STR}$ to minimize the sum of the squared residual

Ordinary Least Squares

- The sum of the squared residual is

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (TestScore_i - \hat{\beta}_0 - \hat{\beta}_1 STR_i)^2$$

- To solve, take derivative[1] above with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set to zero

$$\sum_{i=1}^{n} (TestScore_i - \hat{\beta}_0 - \hat{\beta}_{STR} STR_i) = 0$$

$$\sum_{i=1}^{n} (TestScore_i - \hat{\beta}_0 - \hat{\beta}_{STR} STR_i) STR_i = 0$$

- These are the OLS Normal Equations

1. If you don't know calculus, don't worry about it. I will not ask you to take a derivative in this class.

Ordinary Least Squares

- Use these equations to solve for $\hat{\beta}_0$ and $\hat{\beta}_{STR}$

Ordinary Least Squares Estimators (for our example)

$$\hat{\beta}_0 = \overline{TestScore} - \hat{\beta}_1 \overline{STR}$$

$$\hat{\beta}_{STR} = \frac{\sum_{i=1}^n (STR_i - \overline{STR})(TestScore_i - \overline{TestScore})}{\sum_{i=1}^n (STR_i - \overline{STR})^2} = \frac{\widehat{cov}(STR_i, TestScore_i)}{\widehat{var}(STR_i)}$$

- The estimates of the intercept and slope based on our sample
- 💥**Important**💥: these will differ from one sample to another
  - We will return to sampling variation later

Ordinary Least Squares

The estimated model has its own terminology

1. Sample Regression Function

$$\hat{\beta}_0 + \hat{\beta}_{STR} STR$$

   The line constructed with the OLS estimators

2. Predicted Value

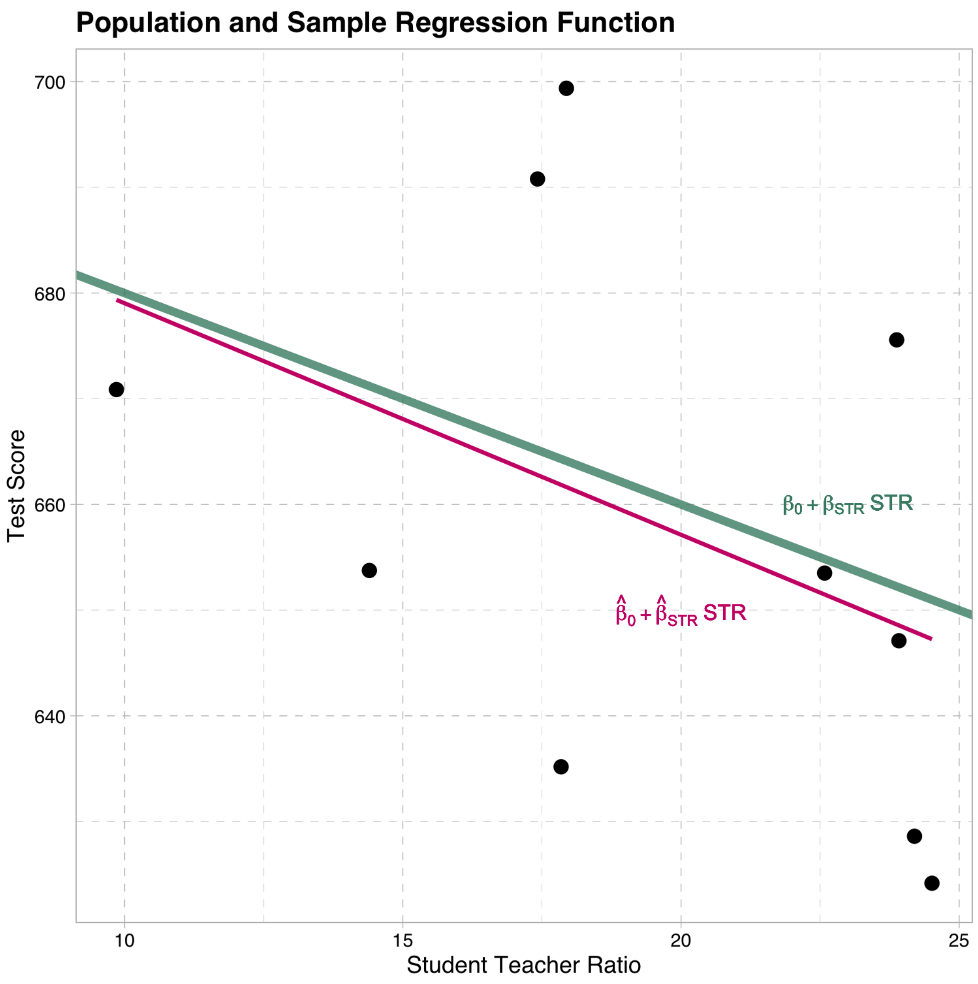
$$\widehat{TestScore}_i = \hat{\beta}_0 + \hat{\beta}_{STR} STR_i$$

   The value of $TestScore_i$ implied by the sample regression function

3. Residual

$$\hat{u}_i = TestScore_i - \hat{\beta}_0 - \hat{\beta}_{STR} STR_i$$

   The difference between the actual value of $TestScore_i$ and its prediction



General Model

- So far we have used a specific example
- A population regression function for any outcome and any independent variable is

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

Ordinary Least Squares Estimators

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

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

Sample Regression Function

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

Predicted Value

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_{STR} X_i$$

Residual

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

Example: The Effect of Class Size on Test Scores

- **Question:** Are class size and student achievement related?
- We will create simulated data to explore the relationship
  - We set the process generating the data
  - Lets us control the true values of the parameters
  - We set these values to create realistic data
- The simulated data will mimic actual data we see on test scores
- We will use this dataset to explore linear regression
  - We will see mechanics of estimation
  - Also how sampling variation affects estimates

Example: The Effect of Class Size on Test Scores

- Suppose the population regression function is

$$TestScore_i = \beta_0 + \beta_1 STR_i + u_i$$

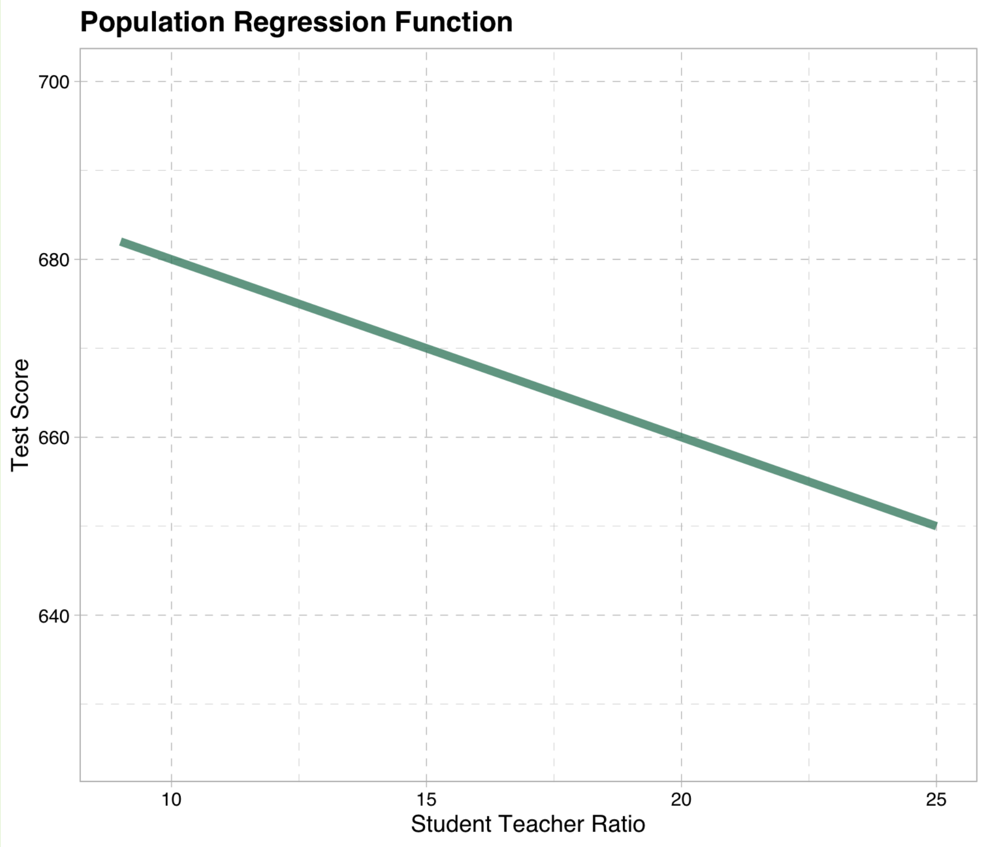
    - β_1 is effect of one more student per teacher
    - β_0 is test score when class size is zero
        - Does not have a useful interpretation in this example
- u are determinants of test scores other than student-teacher ratio
    - Natural ability
    - Student background
    - School/teacher quality
    - etc

Example: Effect of Class Size on Test Scores

- Set the population regression equation as

$$TestScore = 700 - 2 * STR + u$$

  - Says that $\beta_0 = 700$, $\beta_1 = -2$
  - These are fictional population values
    - In reality we would never know these
    - We are pretending we know them for instructional reasons



Example: Effect of Class Size on Test Scores

- Next step is to estimate β_0 and β_1
  - As though we did not know their values
- First take sample of data from population
- We will draw 420 observations with a simple random sample
- Stata code on right

Stata Code

```
clear
set obs 420
set seed 12345

gen str = rnormal(20,2)
gen u = rnormal(0,20)

gen testscr = 700 -2 * str + u
```

Example: Effect of Class Size on Test Scores

- Before estimating parameters, summarize the data

Stata Code and Output

```
sum testscr str
```

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     testscr |        420    659.1345    20.67156    593.118   713.0748
         str |        420    20.13071    2.103167    14.2861   27.30753
```

- Note scale of test scores
  - Simulate scores from a standardized test
  - Standardized tests often scaled to have mean 650, standard deviation 20
- Roughly 20 students per teacher in these fictional districts

Example: Effect of Class Size on Test Scores

- Estimate intercept and slope by OLS

Stata Code and Output

```
regress testscr str
```

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     15.45
       Model |  6383.10498         1  6383.10498   Prob > F        =    0.0001
    Residual |  172661.265       418  413.065226   R-squared       =    0.0357
-------------+----------------------------------   Adj R-squared   =    0.0333
       Total |  179044.369       419  427.313531   Root MSE        =    20.324

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -1.855817    .472094    -3.93   0.000    -2.783791   -.9278429
       _cons |   696.4934    9.55519    72.89   0.000     677.7112    715.2756
------------------------------------------------------------------------------
```

Example: Effect of Class Size on Test Scores

- The OLS estimates are

$$\hat{\beta}_1 = -1.86$$

$$\hat{\beta}_0 = 696.49$$

- The sample regression function is

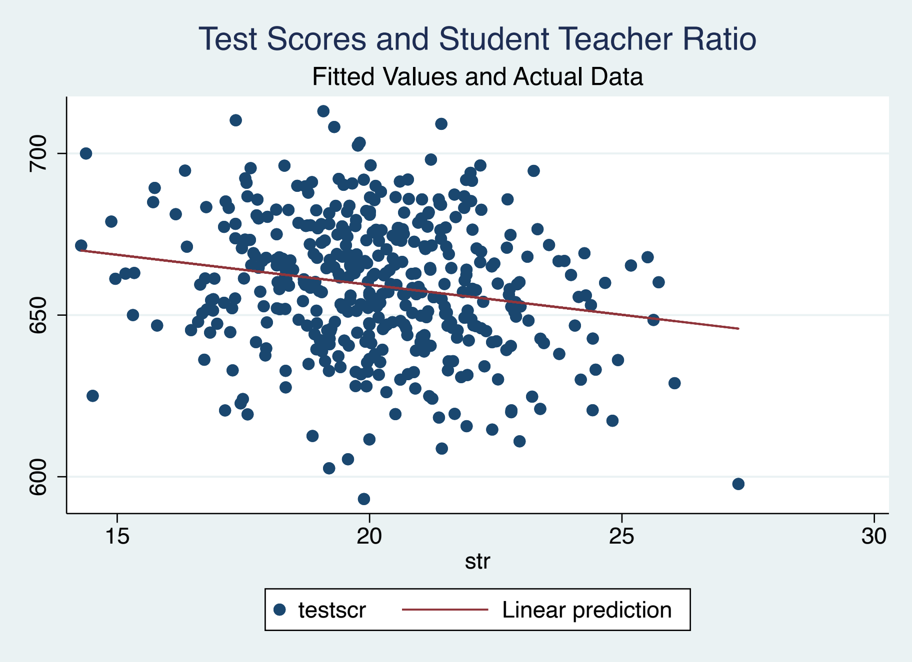
$$\widehat{TestScore} = 696.49 - 1.86 STR$$

  - Use to generate predictions of test scores
  - Simply plug in a value for STR, and compute $\widehat{TestScore}$

Example: Effect of Class Size on Test Scores

Stata Code

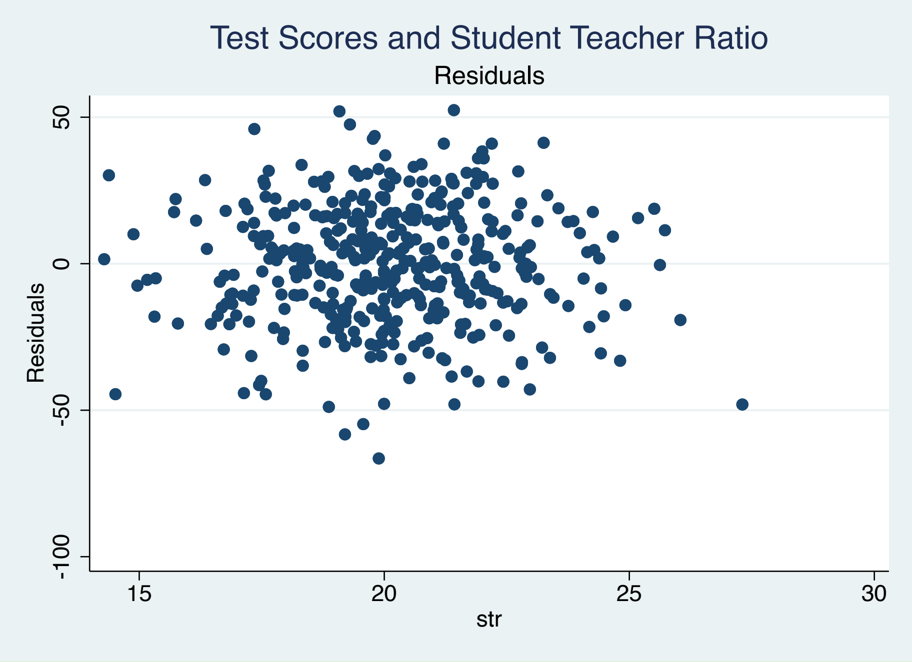
```
predict fitted, xb
twoway (scatter testscr str)(line fitted str), title(Test Scores and Student Teacher Ratio)
subtitle(Fitted Values and Actual Data)
```



Example: Effect of Class Size on Test Scores

Stata Code

```
predict resid, residual
twoway (scatter resid str), title(Test Scores and Student Teacher Ratio) subtitle(Residuals)
```



Measures of Fit

Introduction

- OLS is one way to estimate a linear regression model
- It is important to know how well the method works
- One way is to examine the fit of our regression line
  - How close to the line are the datapoints?
  - Does X explain a large fraction of variation in Y?
- These are the algebraic properties of our estimator
  - Mathematical relationships hold true **in each sample**
- Different from the statistical properties
  - The behaviour of estimators **across repeated samples**
  - Necessarily hypothetical because we only have one sample

Measures of Fit

R-Squared

- The Coefficient of Determination R^2 measures the fraction of the variation in y that is explained by the independent variables

$$R^2 = \frac{ESS}{TSS}$$

- TSS is the Total Sum of Squares

$$TSS = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

  - A measure of the spread in the Y_i

Measures of Fit

- ESS is the Explained Sum of Squares

$$ESS = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$$

- And the Residual Sum of Squares (SSR) is

$$SSR = \sum_{i=1}^{N} (\hat{u}_i)^2$$

- R^2 ranges between 0 and 1
  - $R^2 = 0$ means that X explains none of the variation in Y
    - Scatterplot between Y and X is a cloud with no obvious linear relationship
  - $R^2 = 1$ means that X explains all of the variation in Y
    - Data in scatterplot between Y and X fall along a straight line

Measures of Fit

- R^2 is also equal to the square of correlation coefficient between y_i and $\hat{y}_i$
    - $R^2 = 1$ is perfect correlation between prediction and actual values
- An important relationship between sums of squares is

$$TSS = ESS + SSR$$

    - Part of any movement of y_i away from its average is explainable by factors in the regression
    - Other part is related to unobserved factors
- As a result, you can reexpress

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

Measures of Fit

- Important to be cautious when using R^2
- In real applications, R^2 is often very low
  - Does not mean regression is bad
  - Just means we have not captured all factors that explain Y
- A low R^2 does not imply a poor estimate of β_1
  - β_1 measures effect on Y from changing X, all else equal
  - R^2 measures fraction of total variation in Y is explained by X
  - Concepts are independent of each other
- In class size example $R^2 = 0.036$
  - Many other factors besides student-teacher ratio explain test scores

Measures of Fit

Standard Error of Regression (SER)

- Can also measure fit with spread of data around regression line
- The residual $\hat{u}_i$ is deviation of Y_i from prediction

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

- The standard error of regression (SER) is the standard deviation of $\hat{u}_i$
  - The average distance of Y_i from its prediction $\hat{Y}_i$

$$SER = s_{\hat{u}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}$$

Example

- Recall the regression output from earlier

```
regress testscr str
```

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     15.45
       Model |  6383.10498         1  6383.10498   Prob > F        =    0.0001
    Residual |  172661.265       418  413.065226   R-squared       =    0.0357
-------------+----------------------------------   Adj R-squared   =    0.0333
       Total |  179044.369       419  427.313531   Root MSE        =    20.324

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -1.855817    .472094    -3.93   0.000    -2.783791   -.9278429
       _cons |   696.4934    9.55519    72.89   0.000     677.7112    715.2756
------------------------------------------------------------------------------
```

Example

- The sums of squares are
  - $ESS = 6383.10$
  - $SSR = 172661.27$
  - $TSS = 179044.37$
- $R^2 = 0.056$ is in the top right corner
- You can verify that
  - $SST = SSE + SSR$
  - $R^2 = \frac{SSE}{SST}$
- The SER is called the Root MSE (Mean Square Error) in the output
  - From the output $SER = 20.32$

Least Squares Assumptions for Causal Inference

- So far we have defined β_1 only as the **slope**
- The slope could be two things
  1. The (standardized) correlation between X and Y
     - What happens to Y when we change X?
  2. The causal effect of X on Y
     - What happens to Y when we change X and **nothing else that affects Y changes**
- In many applications we want the causal effect
  - What happens to my income if I get a university degree?
  - How does getting a COVID shot affect the likelihood of infection?
- In this section we establish what needs to be true for OLS to estimate a causal effect

Least Squares Assumptions for Causal Inference

Correlation Example

- Regression of Income on Schooling with observational data

$$Inc = \beta_0 + \beta_1 Schl + u$$

- β_1 shows how income changes with schooling
- Probably represents only a correlation
  - People with more schooling were already smarter
  - Would have earned more even without schooling
- Slope reflects partly effect of schooling, partly effect of intelligence

Causation Example

- Regression of test scores on class size when students randomly assigned to classes

$$TestScore = \beta_0 + \beta_1 ClassSize + u$$

- β_1 shows how bigger classes affect scores
- Probably a causal effect because
  - Randomization of class size means it is unrelated to other factors
  - Students in big classes are no different from those in small ones
- Slope reflects only independent effect of class size on scores

Least Squares Assumptions for Causal Inference

- For OLS to estimate the causal effect the following things need to be true

Assumptions for Causal Inference

The model relating Y to X is

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

where β_1 is explicitly defined as the causal effect, **and**:

1. The error u is not systematically related to X on average:

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

2. (X_i, Y_i) are independent and identically distributed (iid)

3. Large outliers are unlikely

Least Squares Assumptions for Causal Inference

Assumption 1: Zero Conditional Mean of the Error

- The average error term u_i, conditional on X_i, is zero

$$E[u_i|X_i] = 0$$

- Means that unobserved factors are unrelated to the independent variable
    - No linear or non-linear relationship between the two
    - Zero correlation and covariance between u_i and X_i
- Intuitively, at each X_i positive and negative errors tend to average out to zero
- Assumption implies the population regression function accurately describes the conditional mean of Y_i
    - Average Y_i is linearly related to X_i

Least Squares Assumptions for Causal Inference

- Why do we need to assume $E[u_i|X_i] = 0$?
- It allows us to claim $\hat{\beta}_1$ is unbiased
    - Average of $\hat{\beta}_1$ over repeated samples equals β_1
- When β_1 is the causal effect and $\hat{\beta}_1$ is an unbiased estimate of it, we can infer causality
    - $E[u_i|X_i] = 0$ means no unobserved factors change systematically with X_i
    - When this is true, $\hat{\beta}_1$ estimates the causal effect of X_i on Y_i
- This is an **assumption**
- We will never know for sure if it is true
    - Best we can do is assess whether we think it is reasonable
    - Most of the time, it is probably not (we will discuss later in the course)

Least Squares Assumptions for Causal Inference

OLS Estimates Unbiased Causal Effect

OLS Estimates Biased Effect

Least Squares Assumptions for Causal Inference

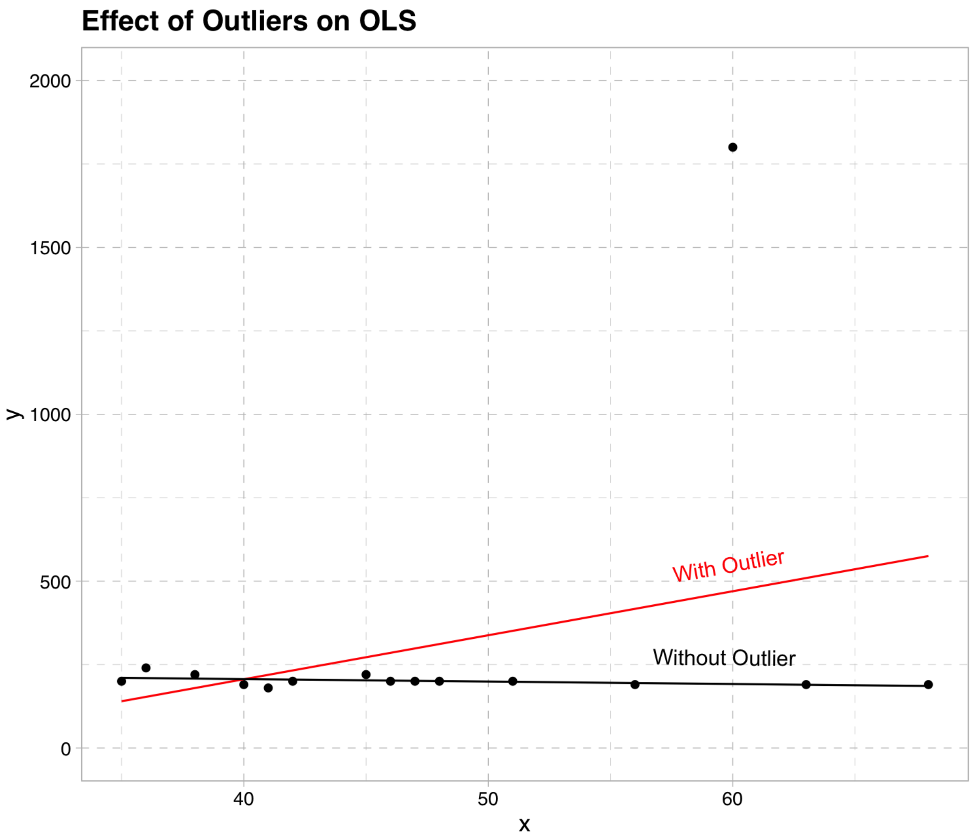
Assumption 2: (X_i, Y_i) are iid

- When sampling, we draw both X_i and Y_i for each person
- Assumption is they are independent, and have the same distribution across people
- If we have a simple random sample, this will be true
  - Observations come from same population
  - Chosen so that everyone has same chance of being in sample
  - Then one pair (X_i, Y_i) gives no info about other (X_i, Y_i)
  - Each (X_i, Y_i) has same distribution
- Assumption sometimes fails with different sampling schemes
  - Ex: time series and panel data

Least Squares Assumptions for Causal Inference

Assumption 3: Large Outliers Unlikely

- Outlier: an observation on X or Y far outside usual range of data
- OLS estimators are sensitive to outliers
    - Regression line on right is flat without outlier
    - Regression line tilts up significantly with one outlier



Least Squares Assumptions for Causal Inference

- Outliers happen for several reasons
  - Data entry error
    - Recording height in cm instead of inches for 1 observation
    - Accidentally shifting decimal place
    - Entering a totally wrong value
  - Naturally occurring issues that are not errors
    - One large country in sample of small countries
    - One big donor in sample of charitable giving
- Important to check data for outliers
  - Examine summary statistics before doing regression
  - E.g. mean, standard deviation, max, min, iqr, etc.

Sampling Distribution of OLS Estimators

Introduction

- The estimator $\hat{\beta}_1$ is a quantity computed from a sample
- Its value therefore varies from sample to sample
    - It is a .red[random variable]
- The sampling distribution of $\hat{\beta}_1$ describes the likelihood of values it can take across random samples
- The sampling distribution helps us test claims about β_1 through hypothesis tests
- For hypothesis tests, we need to know the sampling distribution
- In this section we derive it using our assumptions

Sampling Distribution of OLS Estimators

The Mean of $\hat{\beta}_1$

- Like all random variables, $\hat{\beta}_1$ has a mean and variance
- We compute these values as part of the description of the sampling distribution
- To compute the mean, start with the formula for $\hat{\beta}_1$

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

- First step is to rearrange the formula
- Rewrite numerator as

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})(\beta_1(X_i - \bar{X}) + u_i - \bar{u}))$$

Sampling Distribution of OLS Estimators

- Multiplying out the brackets

$$= \sum_{i=1}^{n} (\beta_1 (X_i - \bar{X})^2 + (X_i - \bar{X})(u_i - \bar{u}))$$

$$= \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u})$$

- The last term can be simplified

$$\sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n} (X_i - \bar{X}) u_i - \sum_{i=1}^{n} (X_i - \bar{X}) \bar{u}$$

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

Sampling Distribution of OLS Estimators

- The estimator $\hat{\beta}_1$ is the sum of two things
  - The parameter it is estimating
  - A weighted sum of the (unknown) errors
- The expected value of $\hat{\beta}_1$ is then

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

$$= E[\beta_1|X_i] + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}|X_i\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})E[u_i|X_i]}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Sampling Distribution of OLS Estimators

- Our first assumption is that $E[u_i|X_i] = 0$, so

$$E[\hat{\beta}_1|X_i] = \beta_1$$

- For a given value of X_i, the average of $\hat{\beta}_1$ is β_1
- To find the **overall** average, use the law of iterated expectations

$$E[\hat{\beta}_1] = E[E[\hat{\beta}_1|X_i]]$$

Sampling Distribution of OLS Estimators

- Substituting in $E[\hat{\beta}_1|X_i] = \beta_1$

$$E[\hat{\beta}_1] = E[\beta_1] = \beta_1$$

  - Intuition: Since the average at each X_i is zero, the overall average is also zero

-The resulting mean of the OLS estimator is

Mean of the OLS Estimator

$$E[\hat{\beta}_1] = \beta_1$$

Sampling Distribution of OLS Estimators

- $E[\hat{\beta}_1] = \beta_1$ means that $\hat{\beta}_1$ is unbiased
- Why is this important?
  - **If we could repeatedly sample** the average of $\hat{\beta}_1$ would be β_1
  - The only reason $\hat{\beta}_1$ differs from β_1 **in any one sample** is sampling error
    - A sample does not always match the population
  - Unbiased estimators are preferable to biased estimators
    - Biased estimators differ from parameter it is estimating because of sampling error **and** because it is systematically wrong
  - Statisticians will generally prefer an unbiased estimator
- If β_1 is the causal effect and $\hat{\beta}_1$ is an unbiased estimate of it, we can attribute causality to the estimated relationship between X_i and Y_i

Sampling Distribution of OLS Estimators

Variance of $\hat{\beta}_1$

- The expected value tells us the middle of the distribution
- We also need to know how spread out the values of $\hat{\beta}_1$ are from the mean across samples
- The key measure of this is the variance
- Start with the alternate formula for $\hat{\beta}_1$ we derived above

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

Sampling Distribution of OLS Estimators

- Rewrite the denominator using the sample variance of X_i

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

- where $s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
- Multiply numerator and denominator by $\frac{1}{n}$

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

- From this point forward, we assume that we have a large sample
  - With large samples, estimators are very close to parameters
  - So $\bar{X} \approx \mu_X$ and $s_X^2 \approx \sigma_X^2$
  - Also, $\frac{n-1}{n} \approx 1$

Sampling Distribution of OLS Estimators

- Substitute these values into the formula

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)u_i}{\sigma_X^2}$$

- Now use the variance operator

$$VAR(\hat{\beta}_1) = VAR\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)u_i}{\sigma_X^2}\right)$$

- Since β_1 is a fixed parameter,

$$VAR(\hat{\beta}_1) = VAR\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)u_i}{\sigma_X^2}\right)$$

Sampling Distribution of OLS Estimators

- We will now make heavy use of the properties of variance
- Because σ_X^2 is a fixed constant

$$VAR(\hat{\beta}_1) = \frac{1}{(\sigma_X^2)^2} VAR\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)u_i\right)$$

- Because $\frac{1}{n}$ is a fixed constant

$$VAR(\hat{\beta}_1) = \frac{1}{(\sigma_X^2)^2 n^2} VAR\left(\sum_{i=1}^{n}(X_i - \mu_X)u_i\right)$$

- Finally, because X_i and u_i are unrelated

$$VAR(\hat{\beta}_1) = \frac{n}{(\sigma_X^2)^2 n^2} VAR\left((X_i - \mu_X)u_i\right)$$

Sampling Distribution of OLS Estimators

- Simplifying, we have the final variance formula

Variance of OLS Estimator

$$VAR(\hat{\beta}_1) = \frac{VAR\left((X_i - \mu_X)u_i\right)}{n(\sigma_X^2)^2}$$

- Important things to note about the spread of $\hat{\beta}_1$
  - The larger is n, the smaller is the variance
    - More data reduces sampling variation
  - The larger is σ_X^2, the smaller is the variance
    - When X_i is more spread out, it is easier to estimate the linear relationship
  - A larger spread in u_i increases the variance

Sampling Distribution of OLS Estimators

The Distribution of $\hat{\beta}_1$

- We know the mean and variance of the distribution of $\hat{\beta}_1$
- What about the shape?
- If we assume a big sample we can apply the Central Limit Theorem (CLT)
    - The sum of independent random variables from the same population is approximately Normally distributed
- $\hat{\beta}_1$ is an average

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)u_i}{\sigma_X^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

Sampling Distribution of OLS Estimators

- Central Limit Theorem says $\hat{\beta}_1$ has a Normal distribution
- We previously derived the mean and variance
- This gives us the distribution of the OLS estimator

Distribution of OLS Estimator

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{VAR\left((X_i - \mu_X)u_i\right)}{n(\sigma_X^2)^2}\right)$$

Example

- Simulate the sampling distribution of $\hat{\beta}_1$
- Code to the right:
  - Assumes model is

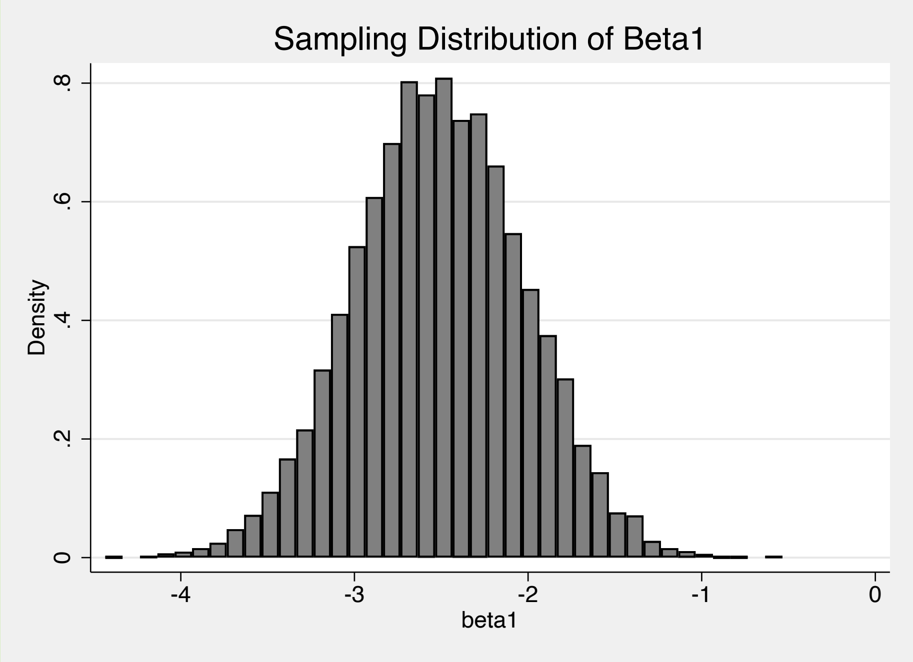
$$TestScore = 700 - 2 * STR + u$$

  - Draws 420 observation on Y and X
  - Computes $\hat{\beta}_1$ based on sample
  - Repeats this 9999 times
  - Plots distribution of 9999 $\hat{\beta}_1$ values

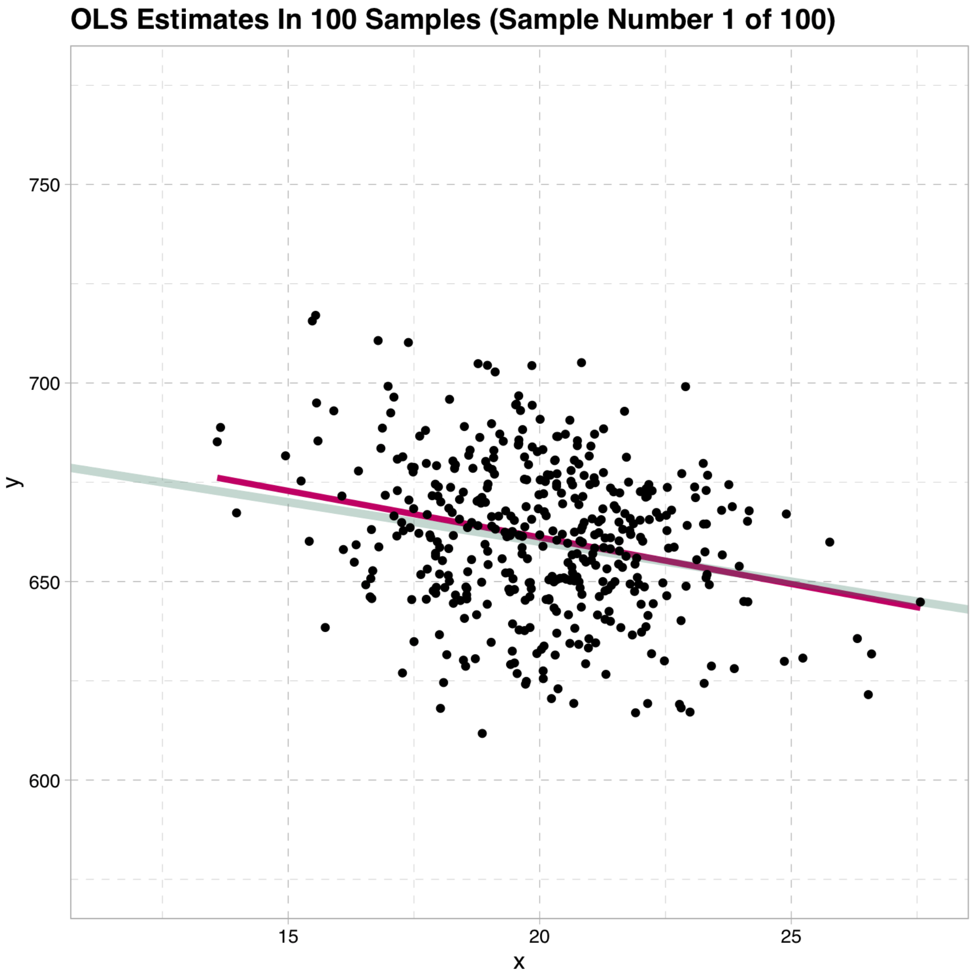
```
clear all
local sims = 9999
set obs `sims'
set more off
gen beta1 = .

forvalues x = 1/`sims' {
    preserve
    clear
    qui set obs 420
    gen str = rnormal(20,2)
    gen u = rnormal(0,20)
    gen testscr = 700 -2 * str + u
    qui regress testscr str
    restore
    qui replace beta1 = _b[str] in `x'
}
```

Example

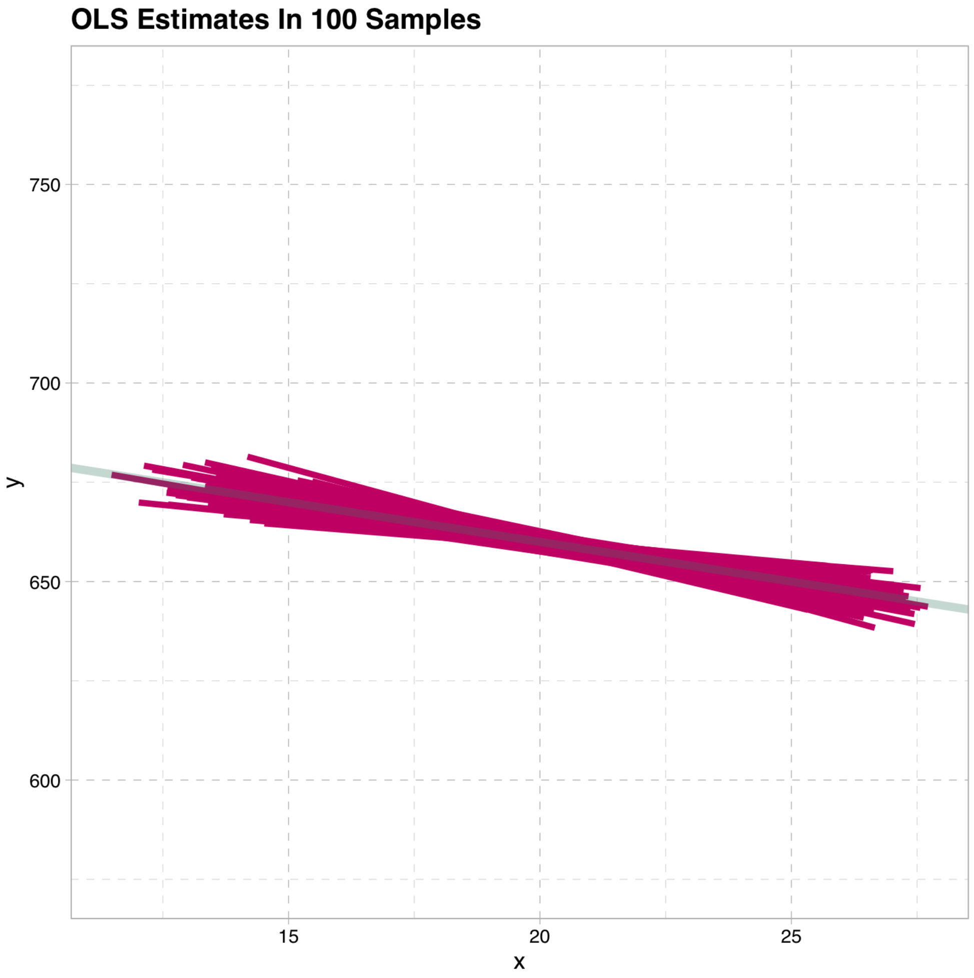
```
twoway hist beta1, title(Sampling Distribution of Beta1) scheme(s2mono)
```



Example



OLS Estimates In 100 Samples (Sample Number 1 of 100)
750
700
650
600
y
15
20
25
x



OLS Estimates In 100 Samples
750
700
650
600
y
15
20
25
x