ECOM20001
Econometrics 1

Lecture Note 4
Single Linear Regression - Estimation

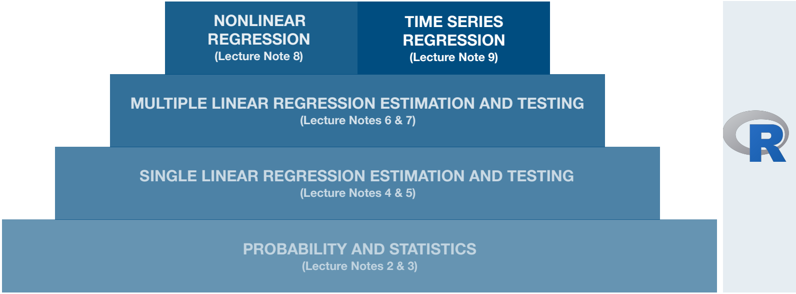
A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 4

Summary of Key Concepts

- Motivating example: class size and test scores
- Single linear regression model
- Population true values
  - population regression line, slope, intercept, errors
- Ordinary Least Squares (OLS) Estimator
  - sample regression line, slope estimate, intercept estimate, residuals, predicted values
- Model fit
  - R^2 and Standard Error of the Regression
- The 3 OLS Assumptions
  - independence, IID, no outliers
- Sampling distribution of the OLS estimator

Building our Econometric Toolkit



Are Expensive Programs to Reduce Class Size Worth It?

Class Size in New York City Schools Rises, but the Impact Is Debated

By JENNIFER MEDINA FEB. 21, 2009

Does Class Size Count?

BY SARA MOSLE MAY 4, 2013 2:42 PM 272

Hire More Teachers, Decrease the Class Size

Yvonne Mason is an AP language and British literature teacher at Mauldin High School in Mauldin, S.C.

UPDATED MARCH 26, 2015, 6:51 AM

Class Size and Test Scores

- Do students in smaller classes achieve higher grades?
- Major question in economics with very mixed evidence: some studies find smaller classes get better grades, others find no evidence of a relationship
- Critical issue for education policy that directly informs government budgeting processes
- Fields of economics: Labour Economics, Public Economics, Economics of Education

Class Size and Test Scores: Project STAR

Project STAR

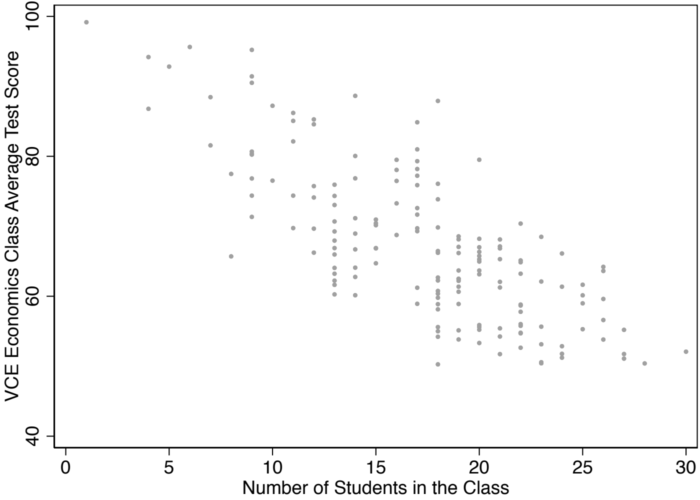
The first phase, termed Project STAR (Student-Teacher Achievement Ratio),[8] randomly assigned teachers and students to three groups, "small" (13 to 17), "regular" (22 to 25) classes with a paid aide, and "regular" (22 to 25) classes with no aide. In total some 6,500 students in about 330 classrooms at approximately 80 schools participated.

Using both standardized and curriculum based testing, the initial study concluded that small classes produced "substantial improvement in early learning and cognitive studies" with the effect about double for minority students. As this is considered the seminal study (in an area that has received much political attention) there have been many attempts to reinterpret the data.

Class Size and Test Scores: Classroom Data

classid	class_size	grade
1	24	35.66
2	16	77.95
3	23	48.80
4	19	53.91
5	19	50.04
6	18	74.52
7	14	72.09
8	24	55.37
9	26	40.04
10	23	51.40
11	16	42.51
12	13	58.11
13	24	48.53
14	17	69.84
15	23	53.15
16	17	57.38
17	23	49.01
18	7	91.16
19	16	68.13
20	28	29.86
21	15	72.60
22	20	61.74
23	24	43.07
24	16	55.50
25	22	46.72

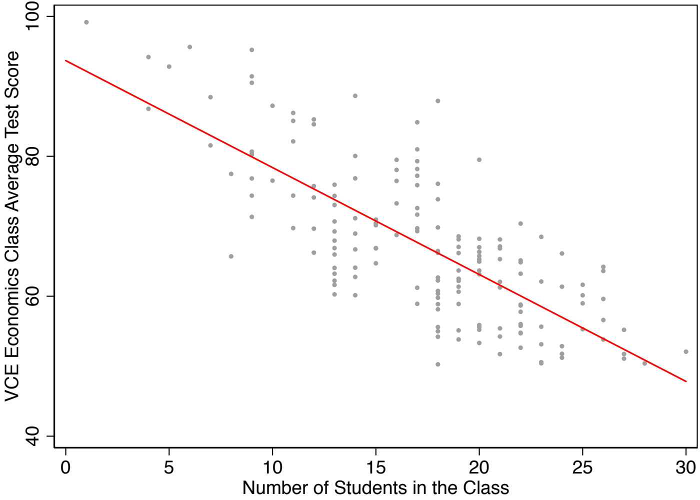
Class Size and Test Scores: Scatter Plot



What Does the Prime Minister Want to Know?

- How strong is the relationship between class size and test scores?
  - This question demands a quantitative answer
- Is the relationship statistically significant?
  - that is, is the relationship real or is it just a chance finding from random sampling?
- To what extent does class size explain or predict test scores?

Class Size and Test Scores: Line of Best Fit



Single Linear Regression

- Over the next 2 sets of lecture notes, we are going to build up a single linear regression model
- The model is used for empirically shedding light on relationships between pairs of economic variables, X (e.g., class size) and Y (test score)
  - strength of the relationship
  - statistical significance of the relationship
  - how much one variable explains/predicts the other variable

Single Linear Regression

- These tools are necessary for rigorously determining whether economic relationships actually exist in data
  - class size vs student performance
    $\rightarrow$ do students in smaller classes perform better?
  - household income vs health/obesity
    $\rightarrow$ are higher income people healthier?
  - market prices vs the number of competing firms
    $\rightarrow$ does competition reduce prices?
  - gender vs wages
    $\rightarrow$ is there a gender wage gap?
  - foreign aid and economic development vs wages
    $\rightarrow$ does providing foreign aid improve income, well-being, or mortality?

Single Linear Regression

- The linear regression model is:

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

where $i = 1, \ldots, N$ indexes observations and

Y_i:	dependent variable (or regressand or left-hand variable)
X_i:	independent variable (or regressor or right-hand variable)
β_0:	intercept parameter of the population regression line
β_1:	slope parameter of the population regression line
u_i:	error term
$\beta_0 + \beta_1 X_i$:	population regression line

Single Linear Regression Example

- From the example above, the linear regression model is:

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

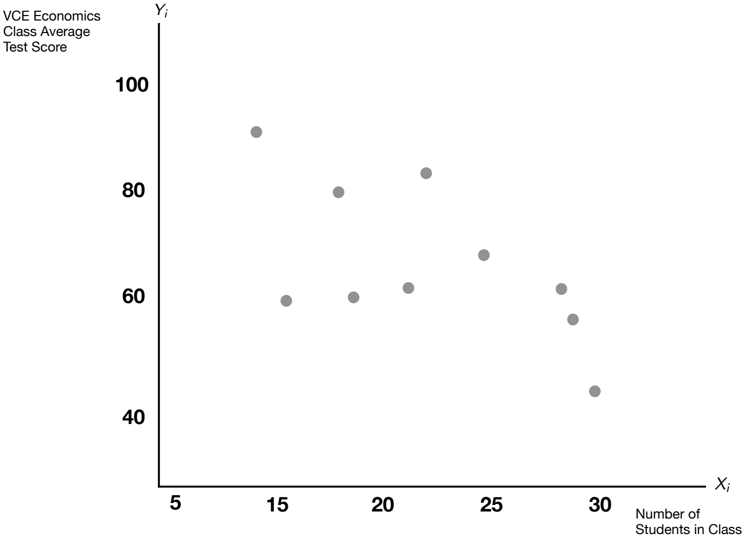
where $i = 1, \ldots, 1000$ classes, and

Y_i:	average test score of class i
X_i:	number of students in class i
β_0:	intercept
β_1:	slope measuring the test score – class size relationship
u_i:	error term: all other factors responsible for differences between test scores and the population regression line
$\beta_0 + \beta_1 ClassSize_i$:	population regression line for the test score and class size relationship

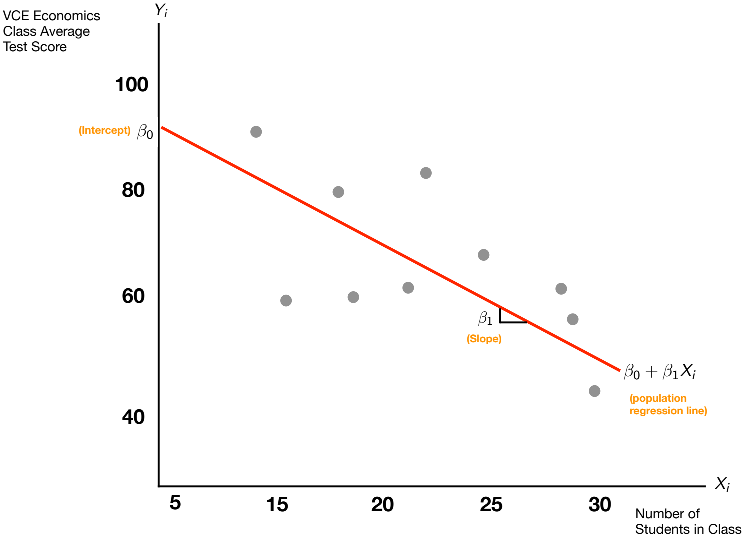
Parameters in a Single Linear Regression

- Parameters in a population regression line are good ways to describe the linear relation between Y_i and X_i
  - β_0 (intercept): the value that Y_i would take if X_i were zero
  - β_1 (slope): the change in Y_i if X_i were to increase by 1 unit
- In our test score and class size example, we expect β_1 to be negative
  - if we increase $ClassSize_i$ by 1 student, TestScore_i has an associated decrease by β_1 points (out of 100) on average
- In this example, how would we interpret β_0?
  - the average test score of a class with zero students

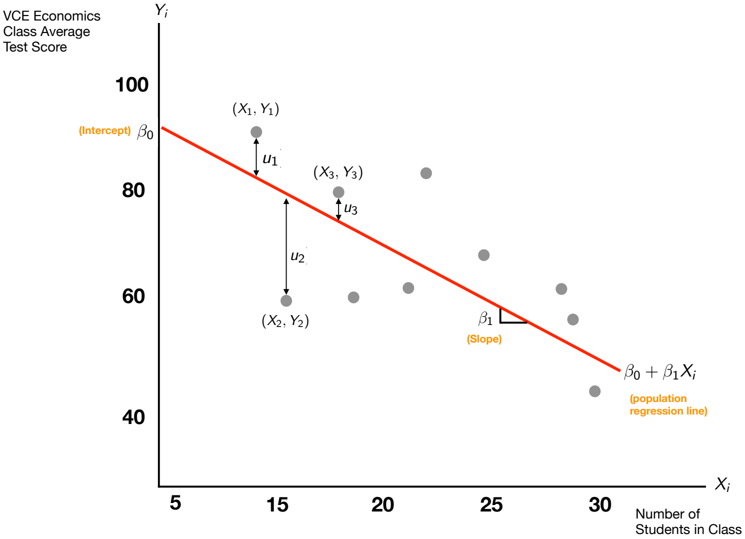
Single Linear Regression Graphically



Single Linear Regression Graphically



Single Linear Regression Graphically



How to Estimate the "Best" Population Regression Line?

- We want to use data to estimate the unknown parameters of the population regression line: β_0 and β_1
- If we have a random sample of (X_i, Y_i)'s $i = 1, 2, \ldots, n$ drawn from the population, it is possible to obtain estimates of β_0 and β_1, which we denote $\hat{\beta}_0$ and $\hat{\beta}_1$
- Similar to estimating the population mean of a single random variable μ_X using an average $\bar{X}$ from a random sample
  - population parameters: β_0 and β_1 is like μ_X
  - sample estimates: $\hat{\beta}_0$ and $\hat{\beta}_1$ is like $\bar{X}$
- We will use the Ordinary Least Squares (OLS) estimator to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$

Mistakes

- The OLS estimator for β_0 and β_1 tries to minimise mistakes
- For observation i, the model's mistake in predicting observation i given values of b_0 for β_0 and b_1 for β_1 is:

$$mistake_i = \underbrace{Y_i}_{data} - \underbrace{(b_0 + b_1 X_i)}_{model\ \ prediction}$$

- Adding up all of the squared mistakes from the model given candidate b_0 and b_1 parameters in predicting each of the $i = 1, \ldots, N$ observations in our sample of (X_i, Y_i)'s we obtain:

$$\sum_{i=1}^{N} \left(\underbrace{Y_i - b_0 - b_1 X_i}_{mistake_i} \right)^2$$

- We square the mistakes so that we add up the total amount of mistakes made by the model, both positive and negative, in predicting the Y_i values

The OLS Estimator Minimises the Mistakes

- Formally, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unique values of b_0 and b_1 that minimise the sum of squared mistakes

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

- The OLS regression line, or the sample regression line, is the straight line constructed using the OLS estimators:

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

- The predicted value $\hat{Y}_i$ for observation i is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The residual $\hat{u}_i$ for observation i is:

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

- Adding the predicted value and residual we get back to the original data:

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

OLS Estimators

- Using calculus, we can find the b_0 and b_1 values that minimise

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

to obtain formulas for the corresponding OLS estimators:

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

and

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

- $\hat{\beta}_1$ the sample covariance of X and Y (e.g., s_{XY}) divided by the sample variance of X (e.g., s_X^2)
- $\hat{\beta}_0$ is the sample mean of Y less the product of $\hat{\beta}_1$ and the sample mean of X

Summary of Population Values and OLS Estimators

Population Values		Sample Estimates	
Population Regression Line:	$\beta_0 + \beta_1 X$	OLS Regression Line	$\hat{\beta}_0 + \hat{\beta}_1 X$
Intercept:	β_0	Intercept Estimate:	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
Slope:	β_1	Slope Estimate:	$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$
Error Term:	u_i	Residual:	$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
		Predicted Value:	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

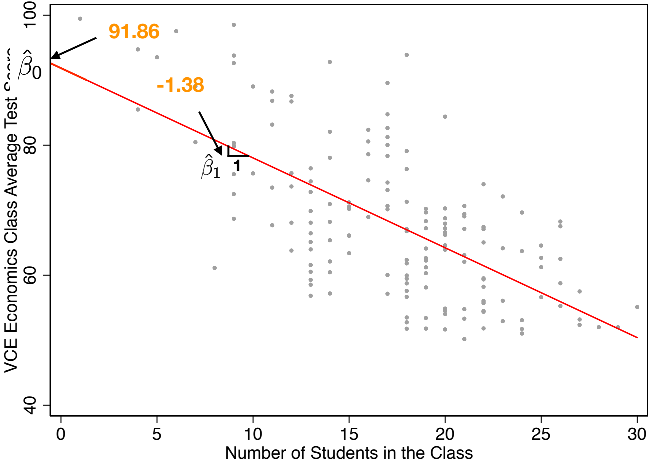
Test Score vs Class Size Example

- In our test scores and class size example, the OLS estimate of the sample regression line is:

$$\widehat{TestScore}_i = 91.86 - 1.38\, ClassSize_i$$

  so $\hat{\beta}_0 = 91.86$ and $\hat{\beta}_1 = -1.38$
- Example interpretations:
  - a 1 student $\uparrow$ in class size is associated with a 1.38 point (out of 100) $\downarrow$ in test scores on average
  - a 5 student $\uparrow$ in class size is associated with a 5 $\times$ 1.38=6.9 point (out of 100) $\downarrow$ in test scores on average

Test Score vs Class Size Example Graphically



Measures of Model Fit

- With OLS estimates of a linear regression in hand, some questions naturally arise
- How well does the regression line describe the data?
- Does the regressor X_i account for much or for little of the variation in the dependent variable Y_i?
- Are the observations tightly clustered around the regression line, or are they spread out?

Measures of Model Fit: R-Squared (R^2)

- The regression R^2 is the fraction of the sample variance of Y_i that is explained by (or predicted by) X_i
- To see this, remember that we can write the dependent variable in the data Y_i in terms of the predicted value and residual from the regression:

$$Y_i = \underbrace{\hat{Y}_i}_{explained\ by\ X_i} + \underbrace{\hat{u}_i}_{not\ explained\ by\ X_i}$$

- $\hat{Y}_i$ is the part of the Y_i explained by our regression; $\hat{u}_i$ is the part of the Y_i that is not explained by our regression
- The R^2 is computed as the ratio of the sample variance of $\hat{Y}_i$ to the sample variance of Y_i

Measures of Model Fit: R-Squared (R^2)

- Mathematically, the R^2 is computed using the explained sum of squares (ESS) and the total sum of squares (TSS):

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

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

- With ESS and TSS, we compute the R^2 as:

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

- More informally, the ratio can be interpreted as:

$$R^2 = \frac{Variation\ in\ Y\ Explained\ by\ X}{Total\ Variation\ in\ Y}$$

Measures of Model Fit: R-Squared (R^2)

- Alternatively, the R^2 can written in terms of the fraction of the variance of Y_i not explained by X_i
- For this definition of R^2, it is helpful to define the sum of squared residuals (SSR):

$$SSR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{u}_i{}^2$$

- Importantly, the TSS, ESS, and SSR are related as follows:

$$TSS = ESS + SSR$$

which in words can be interpreted as:

$$\begin{aligned} Total\ Variation\ in\ Y = &(Variation\ in\ Y\ Explained\ by\ X) \\ &+ (Variation\ in\ Y\ Not\ Explained\ by\ X) \end{aligned}$$

- And the R^2 can alternatively be computed as:

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

Measures of Model Fit: R-Squared (R^2)

- The R^2 of an OLS regression is bounded between 0 and 1. How we interpret it?
- Extreme Cases
  - $R^2 = 1$ implies that the variation in X_i explains <u>all</u> of the variation in Y_1
  - $R^2 = 0$ implies that the variation in X_i explains <u>none</u> of the variation in Y_1
- More realistically, the R^2 will be between 0 and 1
  - The larger the value of R^2, the large the share of the variation in Y_i is explained by X_i
- <u>Note</u>: as econometricians, our main goal is rarely (if ever) to try to maximise the R^2
  - We will come back to this later

Standard Error of the Regression

- The Standard Error of the Regression (SER) is another way of assessing model fit
- SER is computed using the residuals' standard deviation:

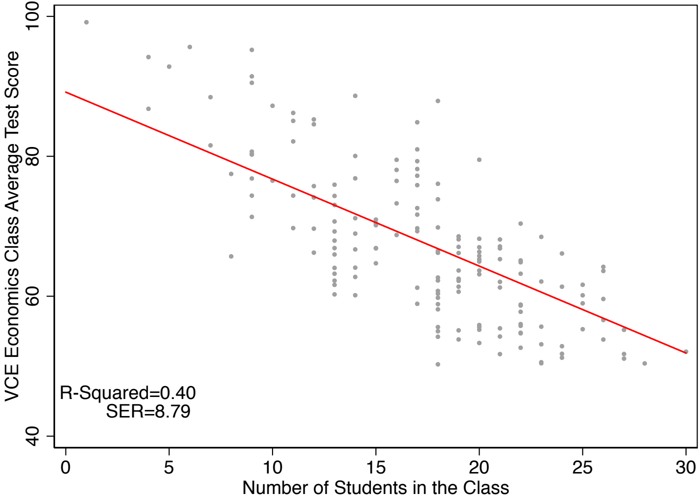
$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}$$

where

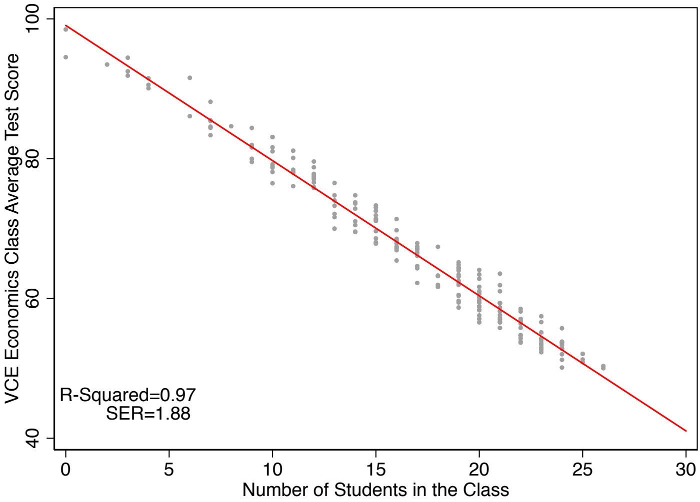
$$s_{\hat{u}}^2 = \frac{SSR}{n-2}$$

- We divided by $n-2$ and not n because we lose two "degrees of freedom" from estimating β_0 and β_1 in computing SSR
  - In practice if n is large dividing by n or $n-2$ is makes little difference
- Interpretting SER
  - the SER is the typical size of the regression error
  - the larger SER, the more spread the data points are around the regression line, implying a "noisier" regression model estimate

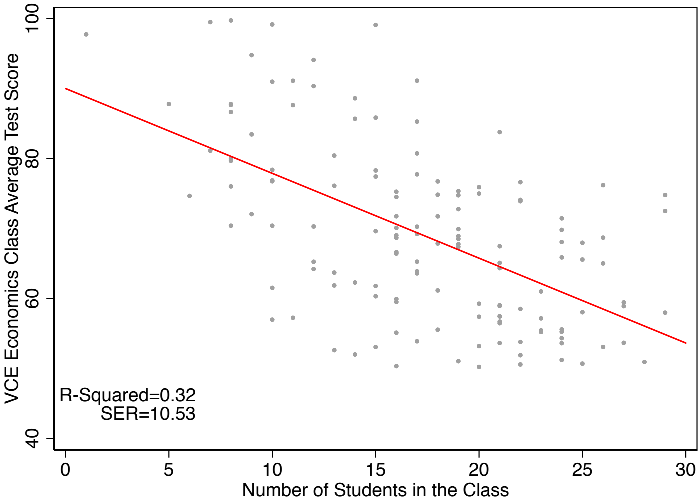
R^2 and SER Graphically



R^2 and SER Graphically



R^2 and SER Graphically



The Least Squares Assumptions

- Single Linear Regression Model:

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

- The validity of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1 rests on <u>three</u> least squares assumptions
- These assumptions are also critical for the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$
  $\rightarrow$ under the assumptions, we can conduct hypothesis tests and construct confidence intervals for β_0 and β_1

Least Squares Assumption #1: Independence

- Assumption 1: The conditional distribution of u_i given X_i has a mean of 0:
$$E(u_i|X_i) = 0$$
- From lecture note 2, recall $E(u_i|X_i) \implies corr(X_i, u_i) = 0$
- Common reinterpretation: other factors u_i in the error of the regression model are unrelated to the regressor X_i
- Therefore, assumption #1 is violated if the correlation between X_i and u_i is not zero

Least Squares Assumption #1: Independence

- To build a graphical representation of OLS Assumption 1, it is useful to consider the expectation of Y_i conditional on X_i:

$$\begin{aligned} E(Y_i|X_i) &= E(\beta_0 + \beta_1 X_i + u_i) \\ &= E(\beta_0|X_i) + E(\beta_1 X_i|X_i) + E(u_i|X_i) \\ &= \beta_0 + \beta_1 X_i + \underbrace{E(u_i|X_i)}_{=0} \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

- So the conditional independence assumption of $E(u_i|X_i) = 0$ implies that $E(Y_i|X_i)$ is just the population regression line:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

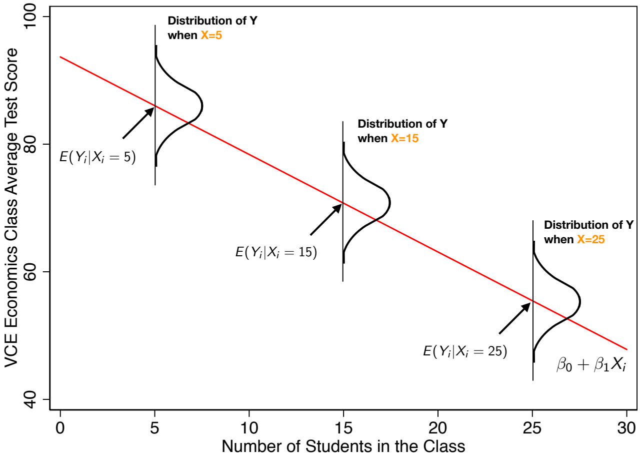
Least Squares Assumption #1: Independence

- Conditional on X_i (or for a given value of X_i), the expectation of Y_i is $\beta_0 + \beta_1 X_i$, the population regression line
- What about the shape of the distribution of Y_i around its conditional expectation given X_i?
- Returning to our regression model:

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

  if we are conditioning on X_i (meaning X_i is fixed/known/given), then the only source of randomness in Y_i comes from u_i
- That is, the shape of the Y_i distribution given X_i will be determined by the shape of the u_i distribution
  - for example, if u_i has a N(0,1) distribution, then the conditional distribution $Y_i|X_i$ will have a $N(\beta_0 + \beta_1 X_i, 1)$ distribution

Least Squares Assumption #1 Graphically



Understanding How Conditional Means Work Under Least Squares Assumption #1

- From 3 slides above, the expected value of Y_i given X_i is:

  $$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

  as long as Least Squares Assumption #1 holds $E(u_i|X_i) = 0$
- We also say that this equation describes the expected value of Y_i conditional on X_i, or simply the conditional mean of Y_i
- If we change the value of X_i, this conditional mean function tells how the expected value (or conditional mean) of Y_i changes as we change X_i

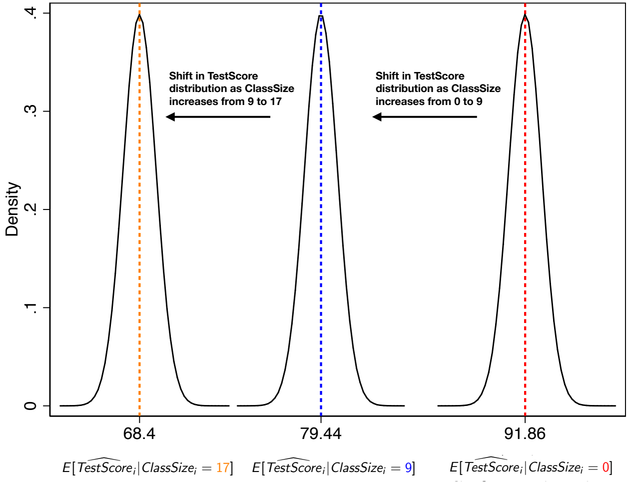
Understanding How Conditional Means Work Under Least Squares Assumption #1

- In our test score vs. class size example, recall that the estimated conditional mean function is:

$$\widehat{TestScore}_i = 91.86 - 1.38 ClassSize_i$$

- Consider some examples of the conditional mean of $\widehat{TestScore}_i$ as we choose different values of $ClassSize_i$:
  - $ClassSize_i = 0$, the conditional mean of $\widehat{TestScore}_i$ is: $E[\widehat{TestScore}_i | ClassSize_i = 0] = 91.86 - 1.38 \times 0 = 91.86$
  - $ClassSize_i = 9$, the conditional mean of $\widehat{TestScore}_i$ is: $E[\widehat{TestScore}_i | ClassSize_i = 9] = 91.86 - 1.38 \times 9 = 79.44$
  - $ClassSize_i = 17$, the conditional mean of $\widehat{TestScore}_i$ is: $E[\widehat{TestScore}_i | ClassSize_i = 17] = 91.86 - 1.38 \times 17 = 68.40$
- We can further graph how changing $ClassSize_i$ changes the expected value of $\widehat{TestScore}_i$

Understanding How Conditional Means Work Under Least Squares Assumption #1



Recall: given a value (or "conditional on") $ClassSize_i$ like 0, 9 or 17, all the variation in $\widehat{TestScore}_i$ is due to variation in $\hat{u}_i$

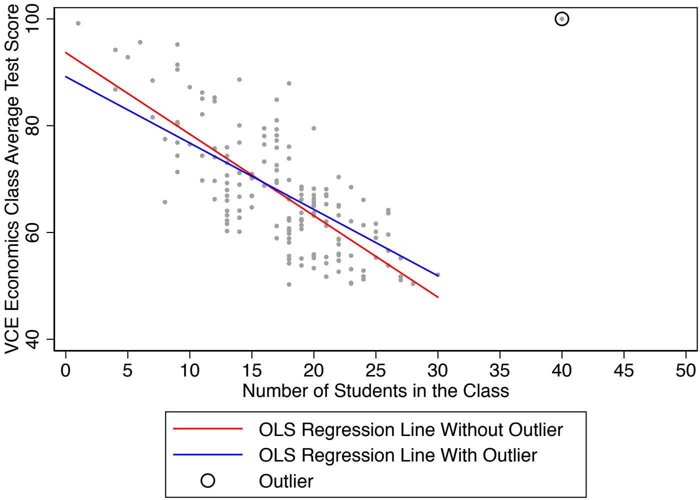
Least Squares Assumption #2: IID

- Assumption 2: The dataset $(X_i, Y_i), i = 1, \ldots, n$ are independent and identically distributed (iid) across observations
- Many sampling schemes in creating economic datasets such as population survey data satisfy the IID assumption
- For the bulk of the subject we will work with the IID assumption, but will introduce time series models at the end of the subject that relax it
- Note: much of ECOM 30002: Econometrics 2 is devoted to relaxing Assumptions #1 and #2 with a diverse array of econometric techniques for dealing with practical situations when using data where these assumptions fail to hold

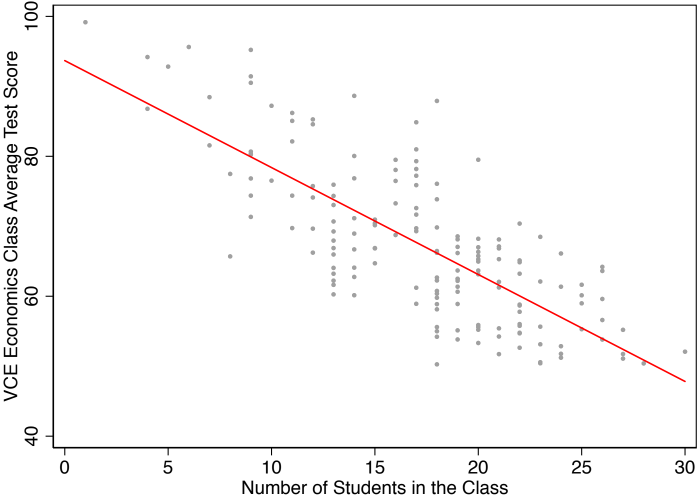
Least Squares Assumption #3: No Outliers

- Assumption 3: Large outliers are unlikely
- Outliers are values of (X_i, Y_i) pairs that are far outside the usual range of the data and can make OLS regression results misleading
- Formally, this assumption requires that the kurtosis of the X_i and Y_i be finite. That is, the tails of their distributions do not go off to infinity
- Often, outliers arise from data coding errors and can be found using scatter plots
- In doing econometric studies, always look at scatter plots of your data first to find and remove outliers (if legitimate)

Least Squares Assumption #3 Graphically



Least Squares Assumption #3 (Outlier Removed)



Reminder: Sampling Distribution of the Sample Average

- Recall a key insight from Lecture 2: because the sample average $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ is constructed using a random sample $Y_1, \ldots, Y_n$, the sample average $\bar{Y}$ <u>itself is a random variable</u>
- Law of Large Numbers (LLN): $\bar{Y}$ will become close to the true value of the population mean μ_Y as n grows
- Central Limit Theorem (CLT): the distribution of $\bar{Y}$ is converges to $N(\mu_Y, \sigma_{\bar{Y}})$ as n grows large, where the standard deviation of $\sigma_{\bar{Y}}$ can be estimated using the sample standard deviation of $\bar{Y}$
- The LLN and CLT are what allow us to use $\bar{Y}$ to estimate μ_Y, conduct hypothesis tests on μ_Y, and construct confidence intervals for μ_Y

Sampling Distribution of the OLS Estimator

- All of these insights regarding the sample average $\bar{Y}$ and population mean μ_Y carry over directly to the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ and the population values of β_0 and β_1
- Because the OLS estimators $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ are constructed using a random sample $(X_1, Y_1), (X_2, Y_2) \ldots, (X_n, Y_n)$, the OLS estimators themselves are random variables
- This implies that, like $\bar{Y}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ have sampling distributions

Sampling Distribution of the OLS Estimator

- Under the 3 Least Squares Assumptions, the mean of the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are their true values β_0 and β_1
- That is, the OLS estimators are unbiased:

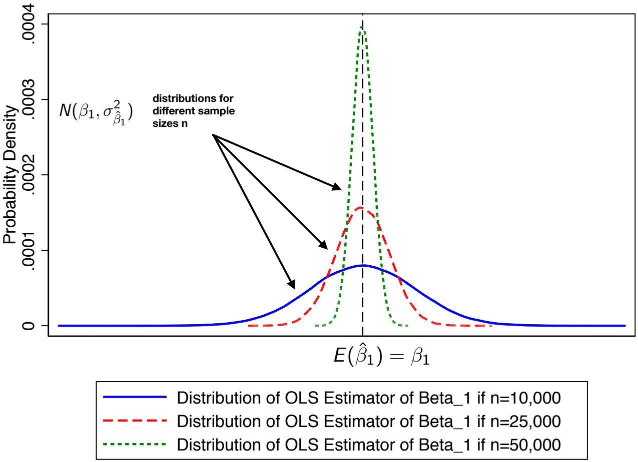
$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1$$

- If n is sufficiently large, the CLT implies that the marginal distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal distributions:

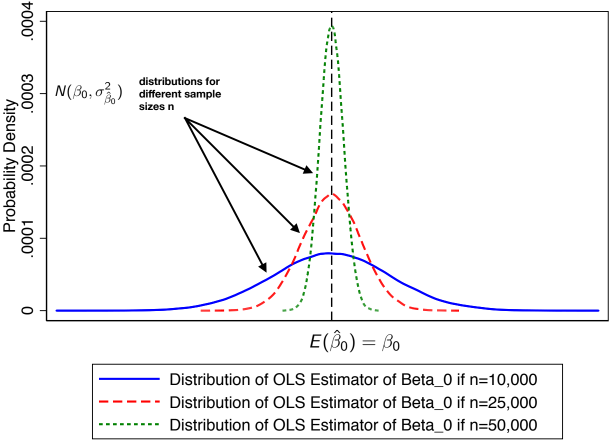
$$N(\beta_0, \sigma^2_{\hat{\beta}_0}) \text{ and } N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are also consistent estimators of β_0 and β_1: as the sample size n gets large, the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ will be close to their true values β_0 and β_1 with high probability

Sampling Distribution of $\hat{\beta}_1$ Graphically



Sampling Distribution of $\hat{\beta}_0$ Graphically



Sampling Distribution of the OLS Estimator

- The calculation for $\sigma^2_{\hat{\beta}_1}$ as n gets large yields:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var[(X_i - \mu_X)u_i)]}{[var(X_i)]^2}$$

- Key point #1: as n grows, $\sigma^2_{\hat{\beta}_1}$ shrinks (like with $\sigma^2_{\bar{Y}}$)
- Key point #2: as $var(X_i)$ grows, $\sigma^2_{\hat{\beta}_1}$ shrinks
  - more disperse X_i values yield more accurate $\hat{\beta}_1$ estimates
- Key point #3: as $var(u_i)$ grows, $\sigma^2_{\hat{\beta}_1}$ grows
  - noisier u_i error values yield less accurate $\hat{\beta}_1$ estimates
  - scatter plot of (X_i, Y_i) gets more disperse around sample regression line as $var(u_i)$ grows

Next Steps: Hypothesis Testing and Confidence Intervals

- We now know how to fit a linear regression line to a scatter plot from a random sample of (X_i, Y_i) $i = 1, \ldots, n$ using Ordinary Least Squares (OLS) to obtain $\hat{\beta}_1$ and $\hat{\beta}_0$ for estimating population true values β_1 and β_0
- We also now know from the LLN and CLT that the distribution of $\hat{\beta}_1$ and $\hat{\beta}_0$ is approximately distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1})$ and $N(\beta_0, \sigma^2_{\hat{\beta}_0})$ as sample size n gets large
- This allows us to test hypotheses about the populations true values for β_1 and β_0 and construct confidence intervals for the true values using a random sample of data drawn from the population