

Lecture 8 - Economics 113

Professor Spearot

▶ **Agenda**

1. OLS and unbiased estimates
2. Variance of the OLS estimates
3. Multivariate Regression

Simple Regression Model

Biased or unbiased

- ▶ When is $\hat{\beta}_1$ a good estimate, where "good" is defined as unbiased?
- ▶ By unbiased, $E[\hat{\beta}_1|x] = \beta_1$
 - ▶ $\hat{\beta}_1$'s are centered around β_1
- ▶ $\hat{\beta}_1$ Unbiased if the following assumptions hold!
 1. Linear in parameters: $y_i = \beta_0 + \beta_1 x_i$
 2. Random sample of size n . $\{(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots (x_n, y_n)\}$
 3. Zero conditional mean: $E(u|x) = 0$
 4. $\sigma_x^2 > 0$.

Simple Regression Model

Biased or unbiased

- ▶ Simple example
- ▶ Suppose that the population is characterized by:

$$y = 3 - 2x_1 + u$$

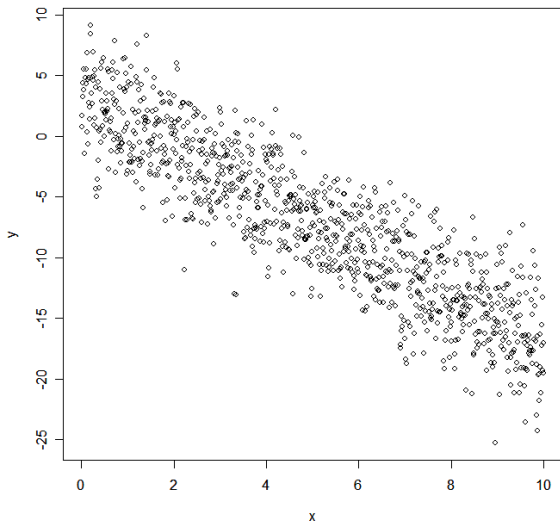
- $\beta_0 = 3$
 - $\beta_1 = -2$
 - u distributed normal, mean 0 and sd 3
 - x_1 's are between 0.01 and 10, spaced evenly
 - 1000 people
- ▶ Estimate using:

$$y = \beta_0 + \beta_1 x_1 + u$$

- ▶ Plot y on x

Simple Regression Model

Biased or unbiased



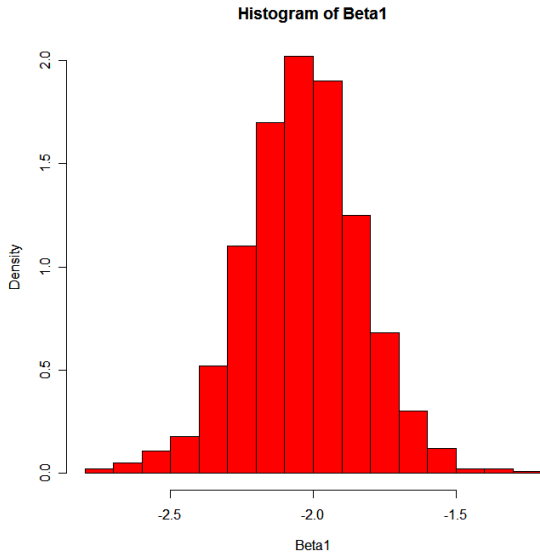
Simple Regression Model

Biased or unbiased

- ▶ Suppose that we sample 30 people from the population, and estimate β_1 via OLS
- ▶ First sample: $\hat{\beta}_1 = -1.951$
- ▶ Second sample: $\hat{\beta}_1 = -1.890$
- ▶ Third sample: $\hat{\beta}_1 = -1.559$
- ▶ They're all wrong. Is this a problem?
- ▶ Keep sampling!!
- ▶ Sample 1000 times
- ▶ Plot a histogram of the estimates of $\hat{\beta}_1$
- ▶ How does the distribution of estimates compare to -2 ?

Simple Regression Model

Biased or unbiased



OLS - Variance

Basics

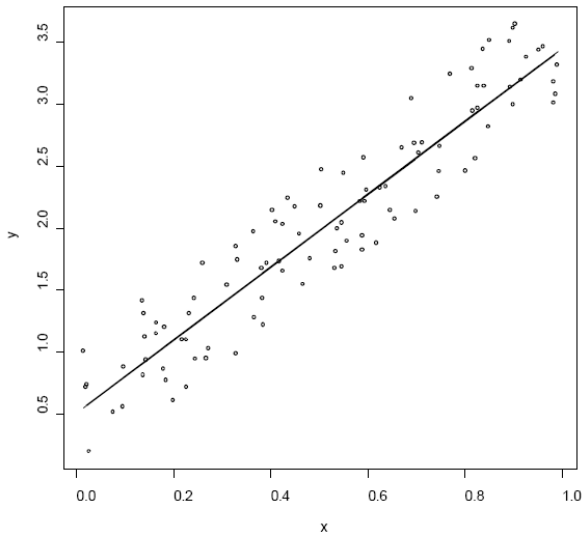
- ▶ If assumptions 1-4 hold, $\hat{\beta}_1$ is centered around β_1 .
 - ▶ *Central tendency says nothing about dispersion.*
- ▶ We are also interested in estimating $Var(\hat{\beta}_1)$
 - ▶ Clearly, from the previous histogram, there is variance in the estimate $\hat{\beta}_1$
 - ▶ Is the estimate of $\hat{\beta}_1$ precise/reliable?
- ▶ Assumption 5 - Homoskedastic Errors:

$$Var [u|x] = \sigma^2$$

- ▶ Variance of errors is common across x .
 - ▶ Assumptions 1-5 are called the "Gauss-Markov Assumptions"
- ▶ If $Var [u|x] \neq Var [u]$, errors are heteroskedastic.

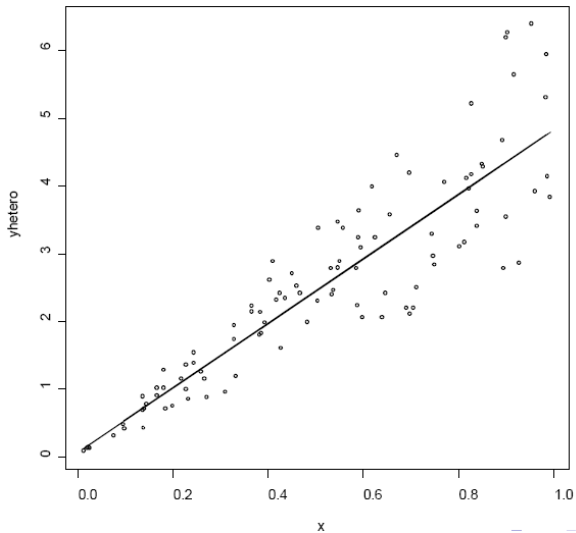
OLS - Variance

Homoskedastic Errors



OLS - Variance

Heteroskedastic Errors



OLS - Variance

Estimate Variance

- ▶ Variance of the slope parameter:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \hat{\mu}_x)^2}$$

- ▶ What do I need for these variance estimates?
 - ▶ An estimate of σ^2 :

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

- ▶ Why $n - 2$?
- ▶ $\hat{\sigma}^2$ requires estimating $\hat{\beta}_0$ and $\hat{\beta}_1$.

OLS

Handling non-linearity

- ▶ *Standard error* of $\hat{\beta}_1$:

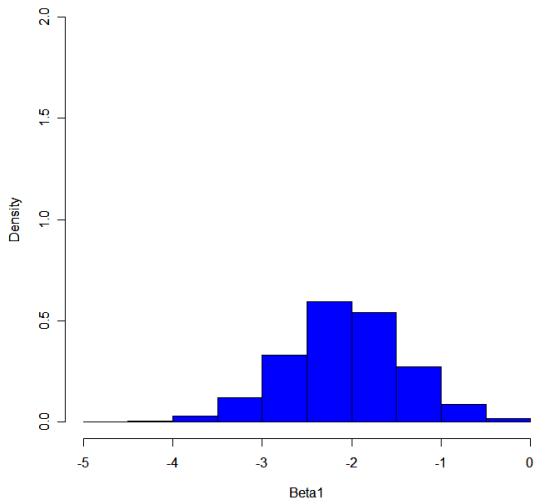
$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \hat{\mu}_x)^2}}$$

- ▶ Dispersion of $\hat{\beta}_1$ around β_1 , same scale as β_1
- ▶ How does $\hat{\sigma}$ effect the precision of our estimates? Why?
 - ▶ Higher $\hat{\sigma}$ yields more higher standard errors (lower precision).
 - ▶ With higher $\hat{\sigma}$, there is more noise, and thus it is harder to get a precise estimate of $\hat{\beta}_1$
 - ▶ Using the original example, compare the following two situations:
 - ▶ u distributed normal, mean 0 and sd **10**
 - ▶ u distributed normal, mean 0 and sd **3**

OLS - Variance

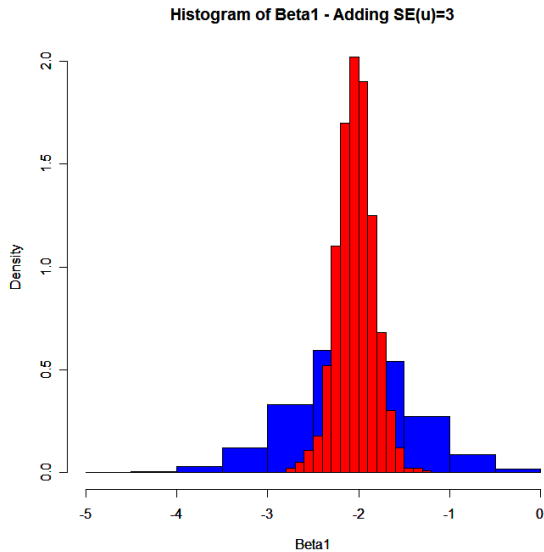
Estimate Variance

Histogram of Beta1 - SD(u)=10



OLS - Variance

Estimate Variance



Multivariate Regression

Introduction

- ▶ Example:

$$Grade = \beta_0 + \beta_1 hrs_study + u$$

- ▶ We estimate $\hat{\beta}_1 = 4.35$.
- ▶ Will a person who studies an extra hour per week get 4.35 more points?
 - ▶ Not if those who study more attend class more often.
 - ▶ *Attend* is an unobserved variable.
- ▶ Remember it must be that $E[u|x] = E[u|hrs_study] = 0$

Multivariate Regression

Introduction

- ▶ A new try:

$$Grade = \beta_0 + \beta_1 hrs_study + \beta_2 Attend + u$$

- *Attend* is number of classes attended.

- ▶ What is $E[u|x]$?
 - ▶ $E[u|hrs_study, Attend] = 0$
 - ▶ If this holds, β_1 and β_2 will be unbiased estimates.
- ▶ Do you think that $E[u|hrs_study, Attend] = 0$ is sensible?
- ▶ What else could be contained in u ?

Multivariate Regression

Introduction

- ▶ Partial solution: Put in as many variables as possible.
- ▶ New estimating equation with k variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- ▶ β_0 still an intercept
 - ▶ $\beta_1, \beta_2, \dots, \beta_k$ are slope parameters
 - ▶ Assume $E[u|x_1, x_2, \dots, x_k] = 0$
- ▶ How do we estimate these models?
- ▶ Least squares techniques (derivatives!)
- ▶ Generate $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$

Multivariate Regression

Using the estimates

- ▶ Predicted value \hat{y} is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

- ▶ How do we interpret the $\hat{\beta}$'s?
 - ▶ Holding everything constant (*ceteris paribus*), the effect of x_l is $\hat{\beta}_l$.
- ▶ Example: Predicting who to admit to college.
- ▶ What should we include if we want to predict freshman GPA?

Multivariate Regression

Using the estimates

- ▶ Suppose we estimated the following:

$$\widehat{FreshGPA} = 1.29 + 0.5HS_GPA + 0.0003SAT$$

- ▶ What does 1.29 mean?
- ▶ Suppose two students have identical GPA's but SAT's of 1150 and 950.
- ▶ What is the predicted difference in *FreshGPA*?
- ▶ Take a difference:

$$\begin{aligned}\Delta\widehat{FreshGPA} &= 0.0003\Delta SAT \\ &= 0.0003(1150 - 950) = 0.6\end{aligned}$$

- ▶ Student A has $HS_GPA = 3.2$, $SAT = 1250$.
- ▶ Student B has $HS_GPA = 3.4$, $SAT = 1180$.
- ▶ To maximize $\widehat{FreshGPA}$, who do we admit?