Univariate Regression

AECN 896-002

Outline

Univariate Regression: Introduction

Plan

- simple univariate regression analysis for the next two weeks
- multivariate regression after that

Population and Sample

Population

A set of ALL individuals, items, phenomenon, that you are interested in learning about

Example

- Suppose you are interested in the impact of eduction on income across the U.S. Then, the population is all the individuals in U.S.
- Suppose you are interested in the impact of water pricing on irrigation water demand for farmers in NE. Then, your population is all the farmers in NE.

Population

Important

Population differs depending on the scope of your interest

- If you are interested in understanding the impact of COVID-19 on child education achievement at the global scale, then your population is every sinlge kid in the world
- If you are interested in understanding the impact of COVID-19 on child education achievement in U.S., then your population is every sinlge kid in U.S.

Sample

Sample

Sample is a subset of population that you observe

- data on education, income, and many other things for 300 individuals from each State
- data on water price, irrigation water use, and many other things for 500 farmers who farm in the Upper Republican Basin (southwest corner of NE)

Econometrics

Learn about the population using sample

Simple linear regression model

Consider a phenomenon in the population that is correctly represented by the following model (This is the model you want to learn about using sample),

A simple model in the population

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

- y: to be explained by x (dependent variable)
- x: explain y (independent variable , covariate , explanatory variable)
- u: parts of y that cannot be explained by x (error term)
- β_0 and β_1: real numbers that gives the model a quantitative meaning (parameters)

What does β_1 measure?

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

If you change x by 1 unit while holding u (everything else) constant,

$$\begin{aligned} y_{before} &= \beta_0 + \beta_1 x + u \\ y_{after} &= \beta_0 + \beta_1 (x + 1) + u \end{aligned}$$

The difference in y_{before} and y_{after},

$$\Delta y = \beta_1$$

That is, y changes by β_1.

We call β_1 the ceteris paribus (with everything else fixed) causal impact of x on y.

What does β_0 measure?

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

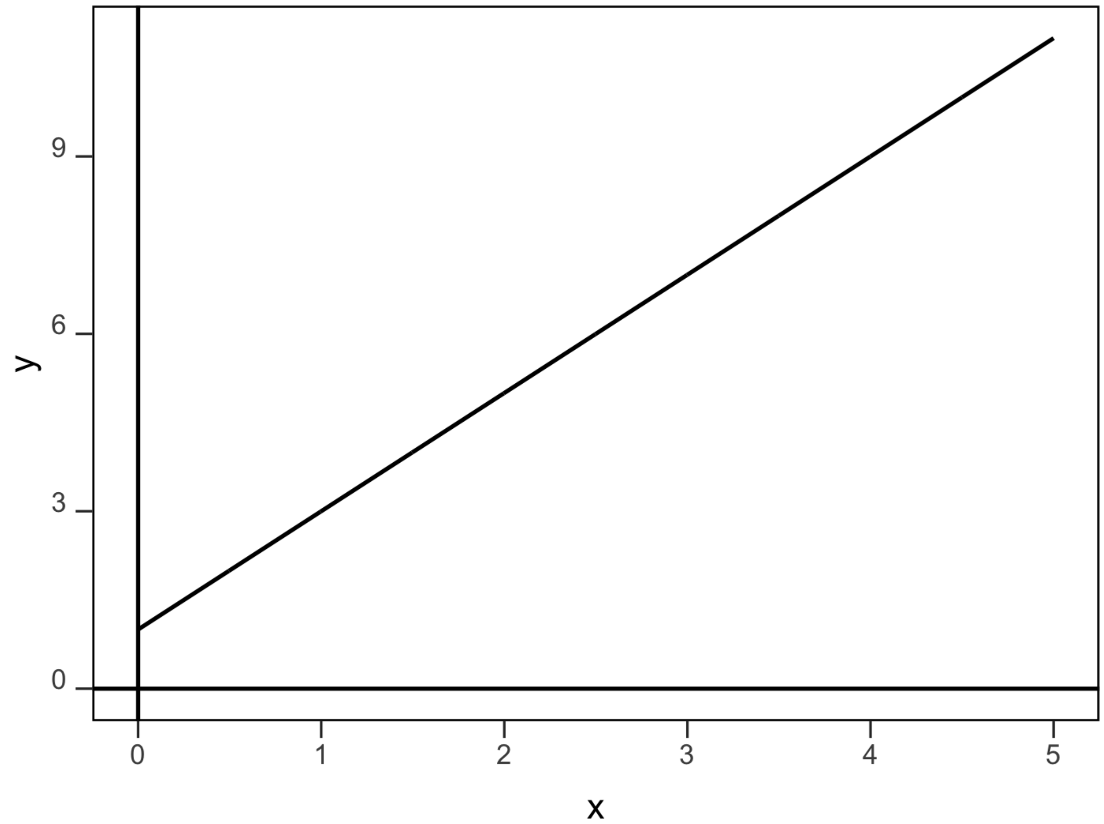
When $x = 0$ and $u = 0$,

$$y = \beta_0$$

So, β_0 represents the intercept (let's see this graphically).

Graphical representation

- β_0: intercept
- β_1: coefficient (slope)



Why do we want ceteris paribus causal impact?

Example: Quality of College

You

- have been admitted to University A (better, more expensive) and B (worse, less expensive)
- are trying to decide which school to attend
- are interested in knowing a boost in your future income to make a decision

You have found the following data

University	average income	sample size
A	130.13	500
B	90.13	500

Question

Should you assume the difference 40 is the expected boost you would get if you are to attend University A instead of B?

What would you be interested in?

Let's say your ability score is 6 out of 10 (the higher, the better),

$$\begin{aligned} &(1) \quad E[inc|A, ability = 9] - E[inc|B, ability = 6] \\ &(2) \quad E[inc|A, ability = 6] - E[inc|B, ability = 6] \end{aligned}$$

Which one would like you to know?

Aside: Contional Expectation

$E[Y|X]$ represents expected value of Y conditoinal on X (For a given value of X, the expected value of Y).

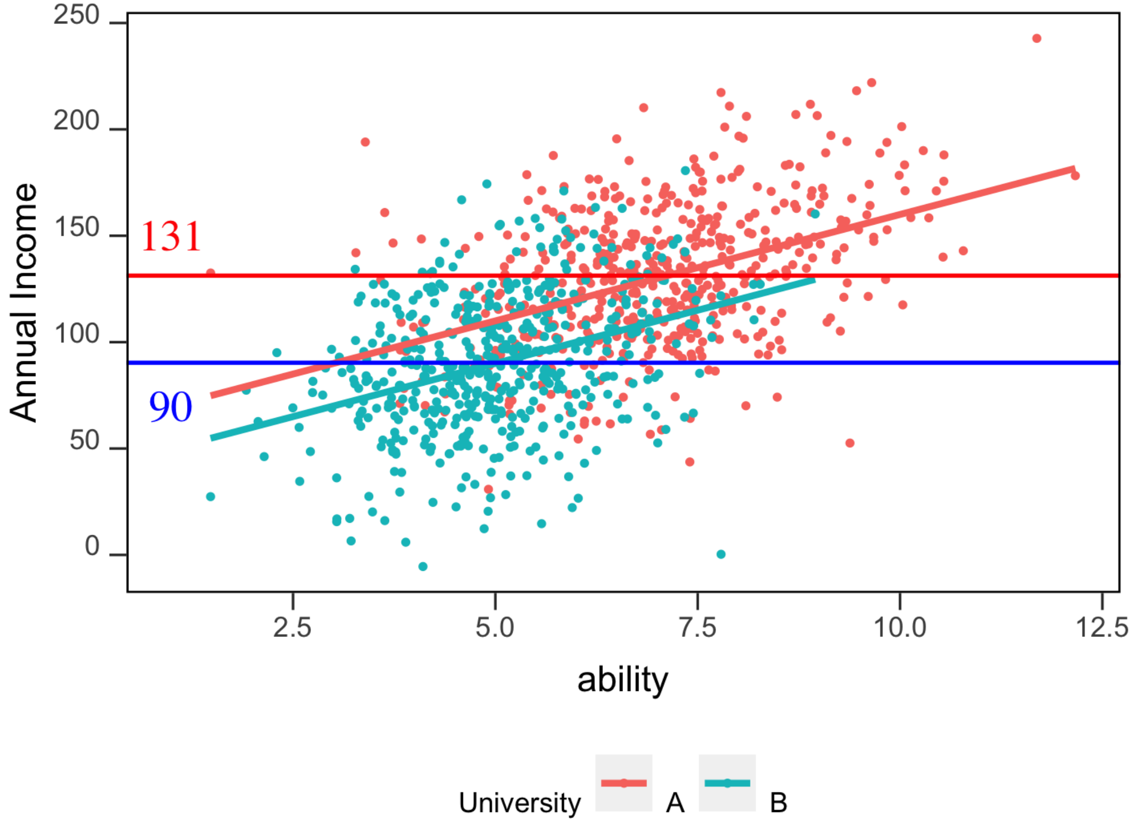
Ceteris Paribus Impact of School Quality

Why ceteris paribus impact?

- you want ability (an unobservable) to stay fixed when you change the quality of school because your innate ability is not going to miraculously increase by simply attending school A
- you don't want the impact of school quality to be confounded with something else

What do you observe?

- red sloped line: $E[income|A, ability]$
- blue sloped line: $E[income|B, ability]$



Example of a simple linear model

Corn yield and fertilizer

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Questions

- what is in the error term?
- are you comfortable with this model?

Estimating β_1 using sample

$$yield = \beta_0 + \beta_1 fertilizer + u$$

- you do not know β_0 and β_1, and would like to estimate them
- you observe a series of $\{yield_i, fertilizer_i\}$ combinations $(i = 1, \dots, n)$
- you would like to estiamte β_1, the impact of fertilizer on yield, **ceteris paribus** (with everything else fixed)

Question

How could we possibly find the **ceteris paribus** impact of fertilizer on yield when we do not observe whole bunch of other factors (error term)?

Crucial conditions to identify the ceteris paribus impact

It turns out, the following condition between x and u needs to be satisfied,

Mean independence

- mathematically:

$$E[u|x] = E[u]$$

- **verbally**: the average value of the error term (collection of all the unobservables) is the same at any value of x, and that the common average is equal to the average of u over the entire population
- **(almost) interchangeably**: the error term is not correlated with x

Correlation and Mean Independence

Note

Mean independence of u and x implies no correlation. But, no correlation does not imply mean independence.

Mean Independence Implies Correlation (proof)

$$
\begin{aligned}
Cov(u,x) =& E[(u-E[u])(x-E[x])] \\
=& E[ux]-E[u]E[x]-E[u]E[x]+E[u]E[x] \\
=& E[ux] \\
=& E_x[E_u[u|x]] \quad \text{(iterated law of expectation)}
\end{aligned}
$$

If zero conditional mean condition $(E(u|x)=0)$ is satisfied,

$$Cov(u,x) = E_x[0] = 0$$

Crucial conditions to identify the ceteris paribus impact

\$E(u)=0\$

This is always satisfied as long as an intercept is included in the model:

$$y = \beta_0 + \beta_1 x + u_1, \quad \text{where} \quad E(u_1) = \alpha$$

Rewriting the model,

$$y = \beta_0 + \alpha + \beta_1 x + u_1 - \alpha$$

$$= \gamma_0 + \beta_1 x + u_2$$

where, $\gamma_0 = \beta_0 + \alpha$ and $u_2 = u_1 - \alpha$.

Now, $E[u_2] = 0$.

Crucial conditions to identify the ceteris paribus impact

zero conditional mean

Combining mean independence and $E[u] = 0$,

$$\text{mean independence:} \qquad E(u|x) = E(u)$$

$$\Rightarrow\text{zero conditional mean:} \qquad E(u|x) = 0$$

Verbally

x and u are not correlated (not systematically related to one another)

Going back to the college-income example

The model

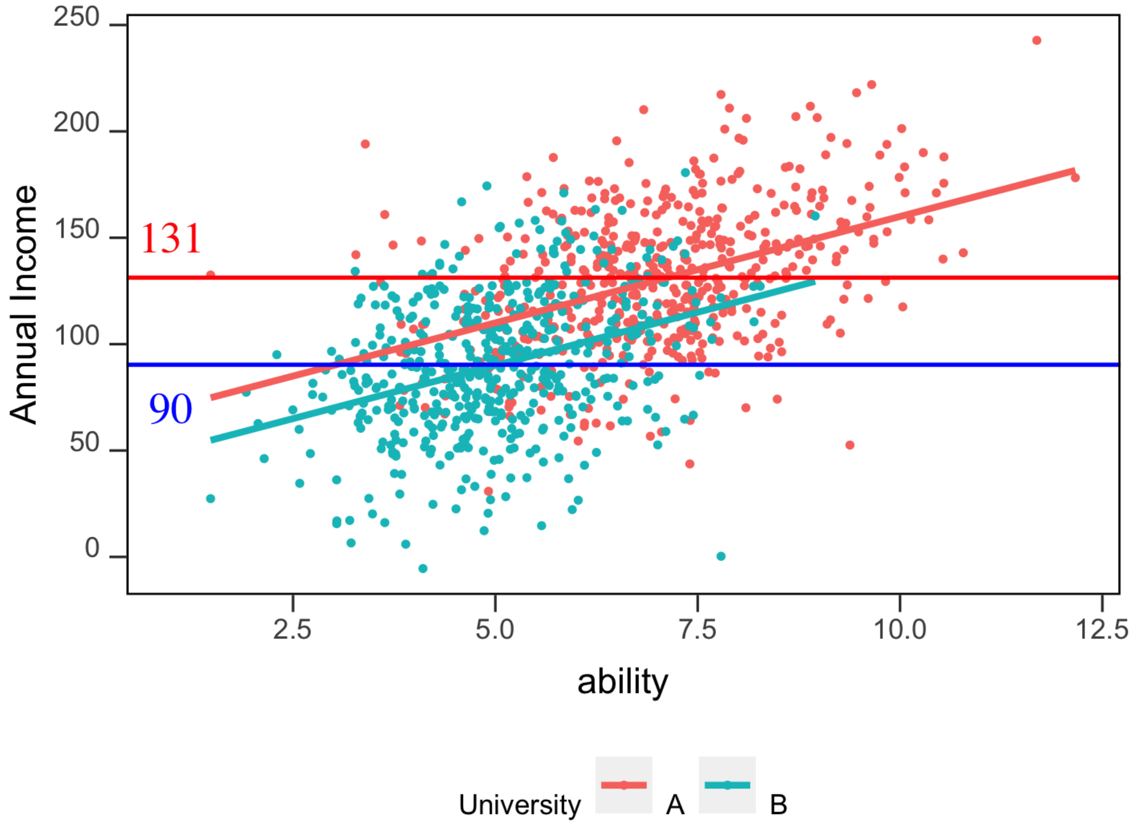
$$Income = \beta_0 + \beta_1 College\ \ A + u$$

where $College\ \ A$ is 1 if attending college A, 0 if attending college B, and u is the error term that includes ability.

Zero conditional mean satisfied?

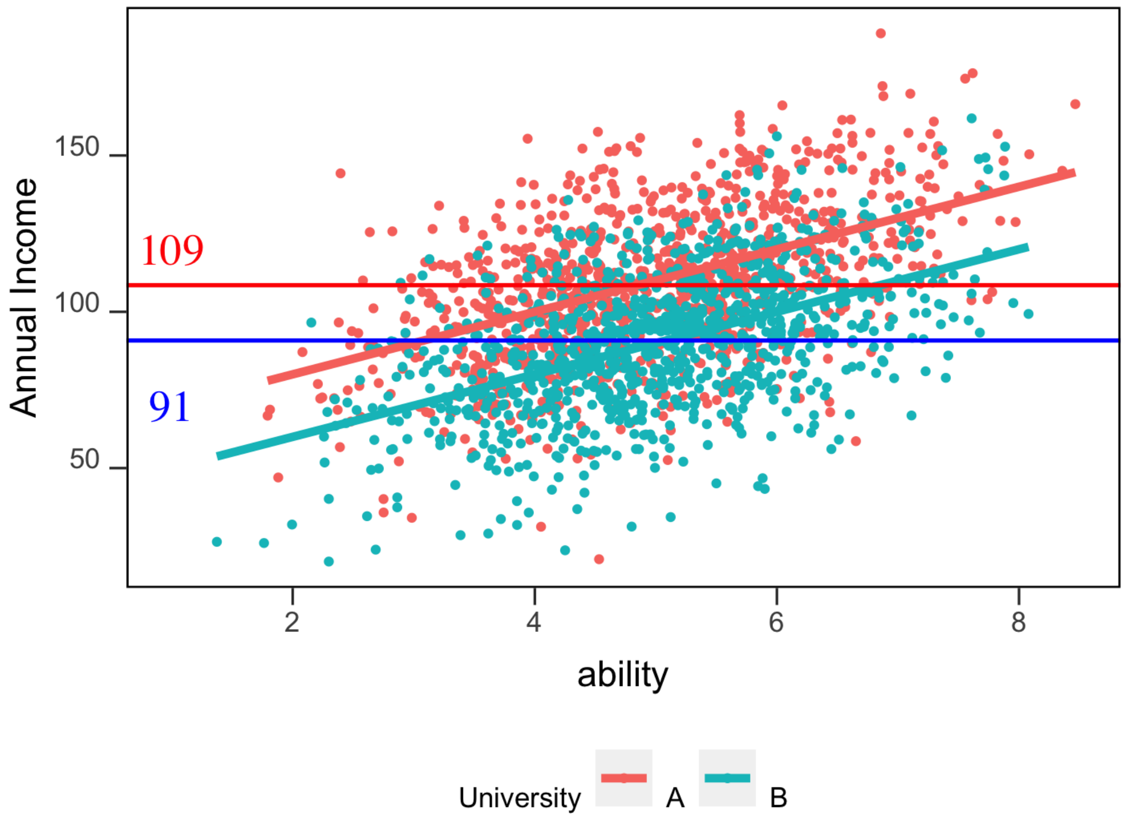
$$E[u(ability)|collegeA] = 0?$$

That is, are going to college A and ability (correlate) systematically related with each other? Or, is college choice correlated with ability?



Annual Income
ability
250
200
150
100
50
0
2.5
5.0
7.5
10.0
12.5
131
90
University
A
B

This is what it would like if college choice and ability are not correlated:



Another Example

yield-fertilizer relationship

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Questions

- What's in u? (note that factors that do not affect yield are not part of u)
- Is it correlated with fertilizer?

Exercise

- consider a phenomenon you are interested in understanding
  - dependent variable (variable to be explained)
  - explanatory variable (variable to explain)
- construct a simple linear modela
- identify what is in the error term
- check if they are correlated withe explanatory variable or not

Estimation of Parameters via OLS

So far

- You have collected data with n observations on y and x
- This random sample is denoted as $\{(y_i, x_i) : i = 1, \ldots, n\}$
- For each i, we can write:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The data set and model

Objective

Estimate the impact of lot size on house price

Model

$$price_i = \beta_0 + \beta_1 lotsize_i + u_i$$

- $price_i$: house price (\$) of house i
- $lotsize_i$: lot size of house i
- u_i: error term (everything else) of house i

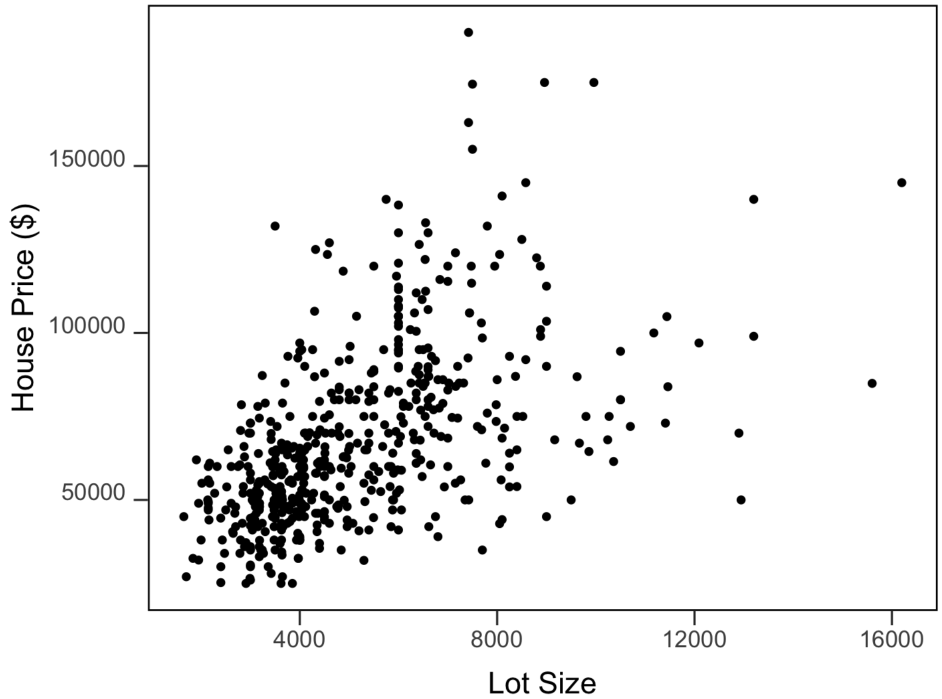
Data set we are going to use

R code: Loading a data set

```r
#--- load the AER package  ---#
library(AER) # load the AER package

#--- load the HousePrices data set ---#
data(HousePrices) # load

#--- take a look ---#
head(HousePrices[, 1:5])
```

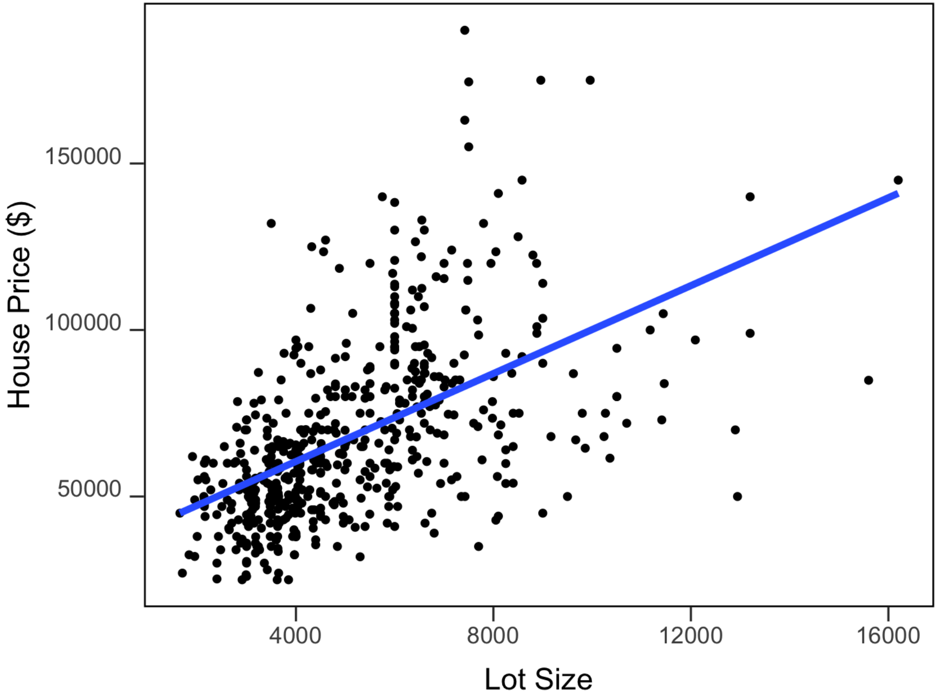
```
##   price lotsize bedrooms bathrooms
## 1 42000    5850        3         1
## 2 38500    4000        2         1
## 3 49500    3060        3         1
## 4 60500    6650        3         1
## 5 61000    6360        2         1
## 6 66000    4160        3         1
##   stories
## 1       2
## 2       1
## 3       1
## 4       2
## 5       1
## 6       1
```

Random sample and regression



Random sample and regression

- We want to draw a line like this, the slope of which is an estimate of β_1
- A way: Ordinary Least Squares (OLS)



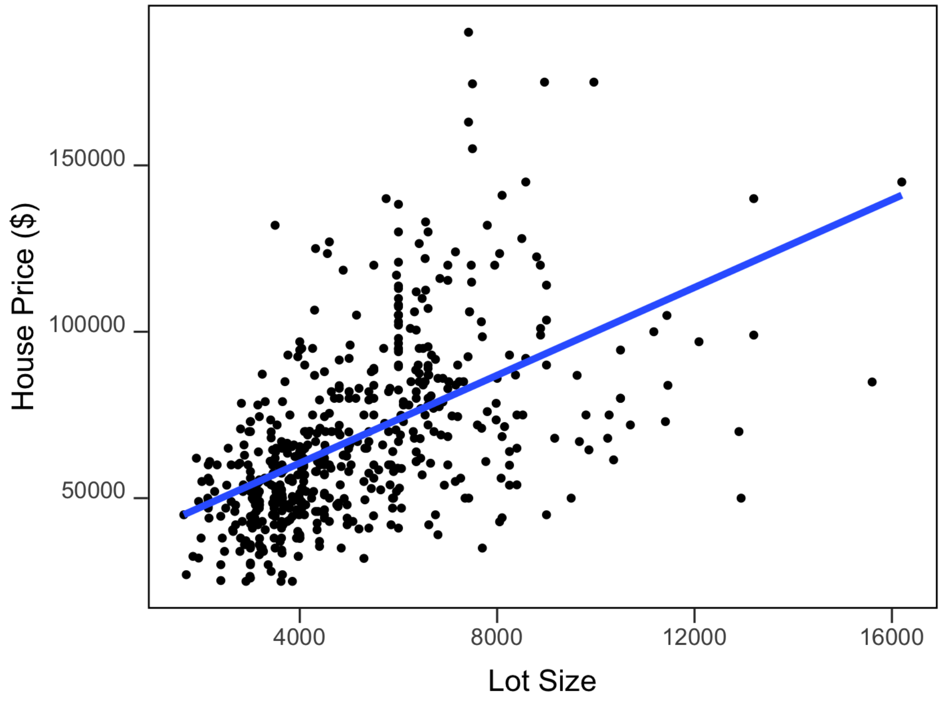
Residuals

For particular values of $\hat{\beta}_0$ and $\hat{\beta}_1$ you pick, the modeled value of y for individual i is $\hat{\beta}_0 + \hat{\beta}_1 x_i$.

Then, the residual for individual i is:

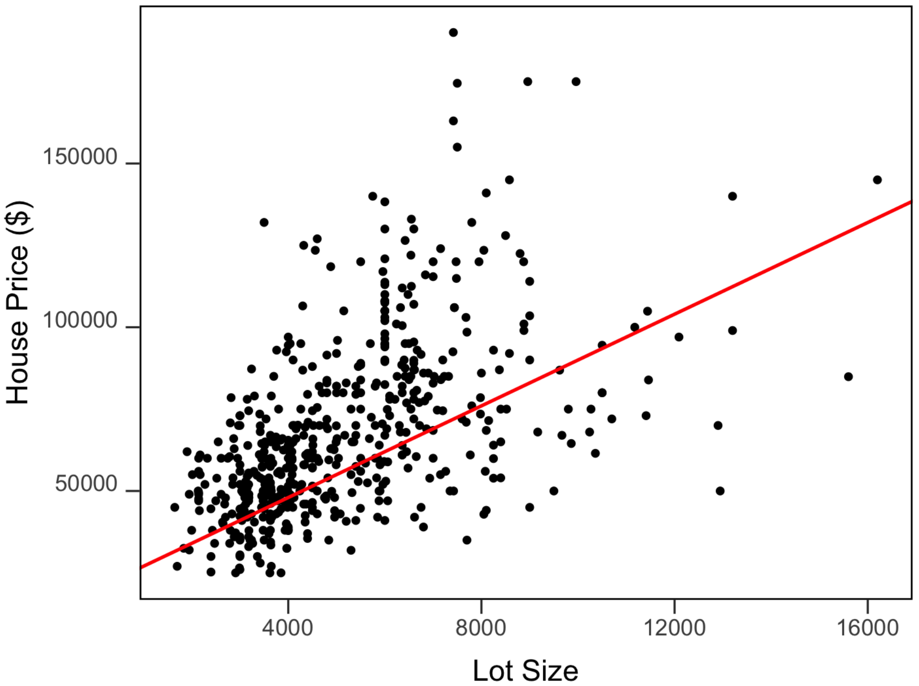
$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

That is, residual is the observed value of the dependent variable less the value of modeled value. For different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, you have a different value of residual.

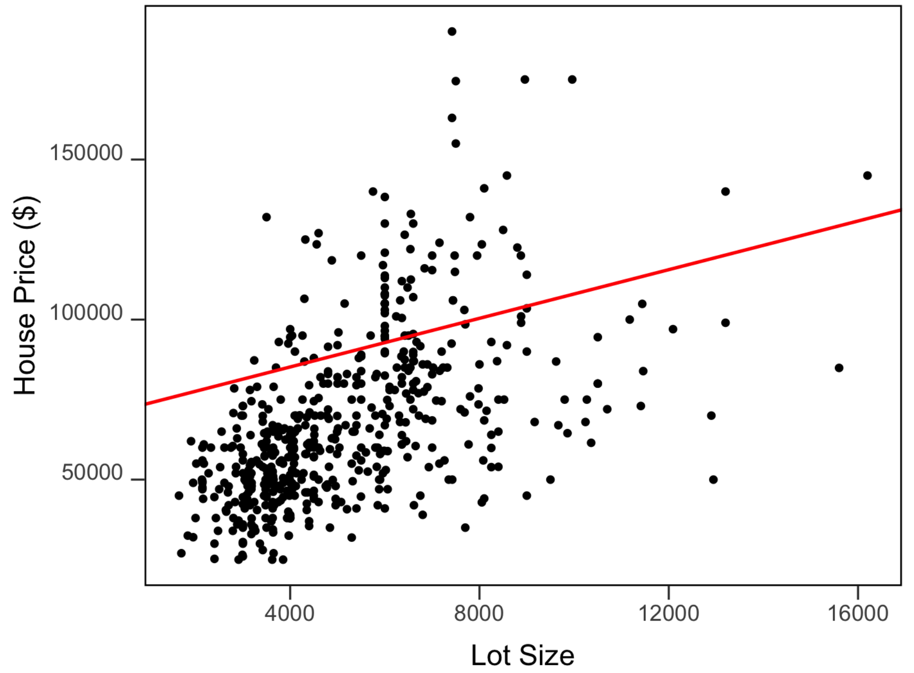

150000
100000
50000
House Price ($)
4000
8000
12000
16000
Lot Size

- Among all the possible values of β_0 and β_1, which one is the best?
- What criteria do we use (what does the best even mean?)

two example



- $\hat{\beta}_0 = 20000$
- $\hat{\beta}_1 = 7$



- $\hat{\beta}_0 = 70000$
- $\hat{\beta}_1 = 3.8$

Ordinary Least Squares (OLS) Methods

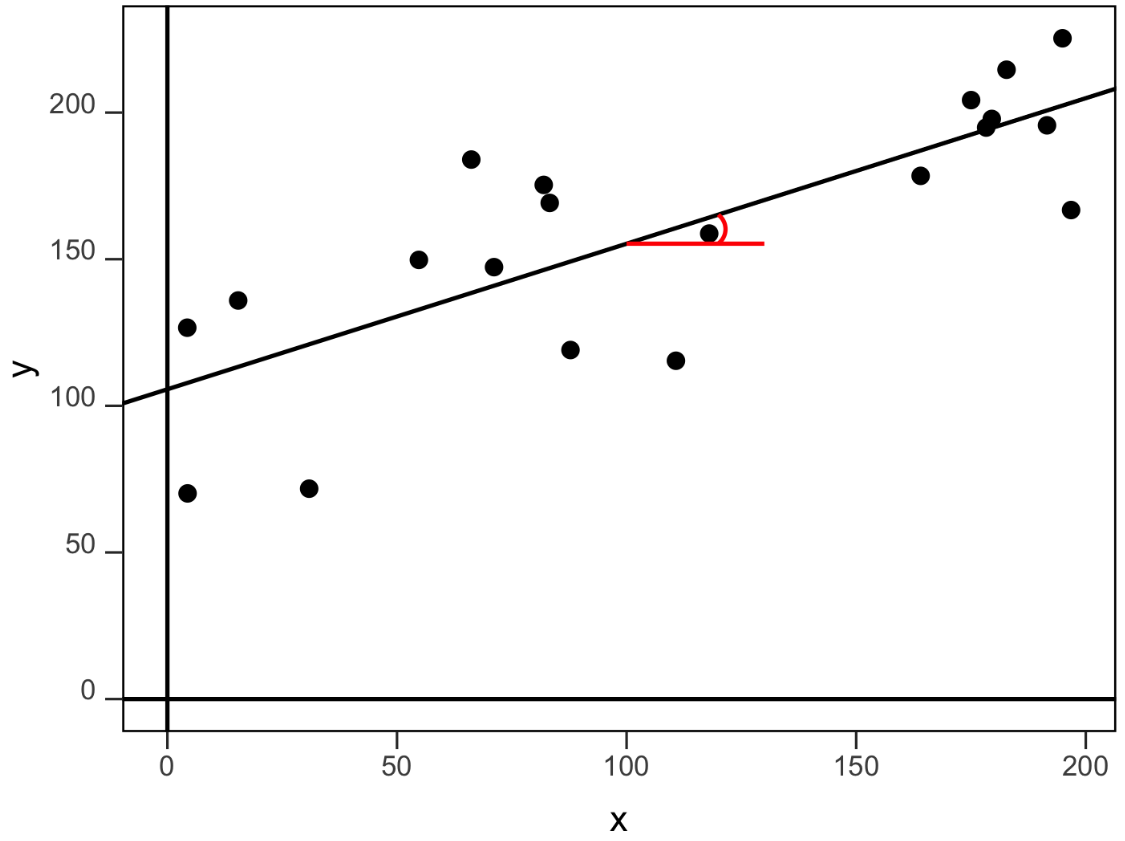
Idea

Let's find the value of β_0 and β_1 that minimizes the squared residuals!

Mathematically

$$Min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2, \text{where } \ \hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

OLS Visualization



Questions

- Why do we square the residuals, and then sum them up together? What's gonna happen if you just sum up residuals?
- How about taking the absolute value of residuals, and then sum them up?

Deriving OLS estimates

Mathematical problem to solve

$$Min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Steps

- partial differentiation of the objective function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$
- solve for $\hat{\beta}_0$ and $\hat{\beta}_1$

OLS derivation: FOC

$$Min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

FOC:

$$\frac{\partial}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$$

$$\frac{\partial}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} x_i \cdot [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = \sum_{i=1}^{n} x_i \cdot \hat{u}_i = 0$$

OLS estimators: analytical formula

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

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

$$\text{where } \; \bar{y} = \sum_{i=1}^{n} y_i/n \; \text{ and } \; \bar{x} = \sum_{i=1}^{n} x_i/n$$

Estimators vs Estimates

Estimators

Specific rules (formula) to use once you get the data

Estimates

Numbers you get once you plug values (your data) into the formula

OLS demonstration in R

Model

$$price = \beta_0 + \beta_1 lotsize + u$$

OLS Estimator Formula

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

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

R code: hard way

```
y <- HousePrices$price
x <- HousePrices$lotsize

#--- beta_1 ---#
b1_num <- sum((x - mean(x)) * (y - mean(y)))
b1_denom <- sum((x - mean(x))^2)
b1 <- b1_num / b1_denom
b1
```

```
## [1] 6.598768
```

OLS demonstration in R

Model

$$price = \beta_0 + \beta_1 lotsize + u$$

Estimation

We can use the `feols()` function from the `fixest` pacakge.

```
library(fixest)

#--- run OLS on the above model ---#
# lm(dep_var ~ indep_var,data=data_name)
uni_reg <- feols(price ~ lotsize, data = HousePrices)
uni_reg
```

```
## OLS estimation, Dep. Var.: price
## Observations: 546
## Standard-errors: Standard
##                Estimate   Std. Error
## (Intercept) 34136.2000 2491.100000
## lotsize         6.5988    0.445847
##              t value Pr(>|t|))
## (Intercept)  13.703 < 2.2e-16 ***
## lotsize      14.801 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 22,525.7   Adj. R2: 0.285766
```

Lots of information is stored in the regression results (`uni_reg`)

```
##  [1] "call"             "call_env"
##  [3] "coefficients"     "coeftable"
##  [5] "collin.min_norm" "cov.unscaled"
##  [7] "fitted.values"    "fml"
##  [9] "fml_all"          "hessian"
## [11] "ll_null"          "means"
## [13] "method"           "method_type"
## [15] "multicol"         "nobs"
## [17] "nobs_origin"      "nparams"
## [19] "obs_selection"    "residuals"
## [21] "scores"           "se"
## [23] "sigma2"           "sq.cor"
## [25] "ssr"              "ssr_null"
```

Estimated coefficients:

```
##  (Intercept)      lotsize
## 34136.191565     6.598768
```

Predicted values at the observation points:

```
## [1] 72738.98 60531.26 54328.42 78018.00
## [5] 76104.35
```

Residuals:

```
## [1] -30738.98 -22031.26  -4828.42
## [4] -17518.00 -15104.35
```

You can have a nice quick summary of the regression results with `summary()` function:

```
## OLS estimation, Dep. Var.: price
## Observations: 546
## Standard-errors: Standard
##               Estimate  Std. Error
## (Intercept) 34136.2000 2491.100000
## lotsize         6.5988    0.445847
##             t value Pr(>|t|))
## (Intercept)  13.703 < 2.2e-16 ***
## lotsize      14.801 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 22,525.7   Adj. R2: 0.285766
```

Model

$$price = \beta_0 + \beta_1 lotsize + u$$

Estimated Model

This is the estimated version of the expected value of y conditional on x.

$$price = 3.4136 \times 10^4 + 6.599 \times lotsize$$

This is called sample regression function (SRF) , and it is an estimation of $E[price|lotsize]$, the population regression function (PRF).

Model

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

Population Regression Function (PRF)

$$E[y|x] = \beta_0 + \beta_1 x$$

Sample Regression Function (SRF)

Estimated version of PRF, where estimates of β_0 and β_1 are plugged into the PRF:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Important:

- OLS Regression is about predicting the expected value of the dependent variable conditoinal on the explanatory variables.
- $\hat{\beta}_1$ is an estimate of how a change in x affects the expected value of y.

R code: Prediction

```r
#--- access fitted values for sample points ---#
uni_reg$fitted.values[1:5]
```

```
## [1] 72738.98 60531.26 54328.42 78018.00
## [5] 76104.35
```

```r
#--- for values of lotsize that are not in the sample ---#
newdata <- data.frame(lotsize = c(3000, 12000, 15000))
predict(uni_reg, newdata = newdata)
```

```
## [1]  53932.49 113321.40 133117.71
```

Exercise: The impact of lotsize

Your current lot size is 3000. You are thinking of expanding your lot by 1000 (with everything else fixed), which would cost you 5,000 USD. Should you do it? Use R to figure it out.

R code: impact of lotsize

```r
#--- access the coefficient values  ---#
uni_reg$coefficients
```

```
##  (Intercept)      lotsize
## 34136.191565     6.598768
```

```r
# class(uni_reg)

#--- assess the impact ---#
uni_reg$coefficients["lotsize"] * 1000 - 5000
```

```
##  lotsize
## 1598.768
```

R^2: Goodness of fit

R^2 is a measure of how good your model is in predicting the dependent variable (explaining variations in the dependent variable) compared to just using the average of the dependent variable as the predictor.

You can decompose observed value of y into two parts: fitted value and residual

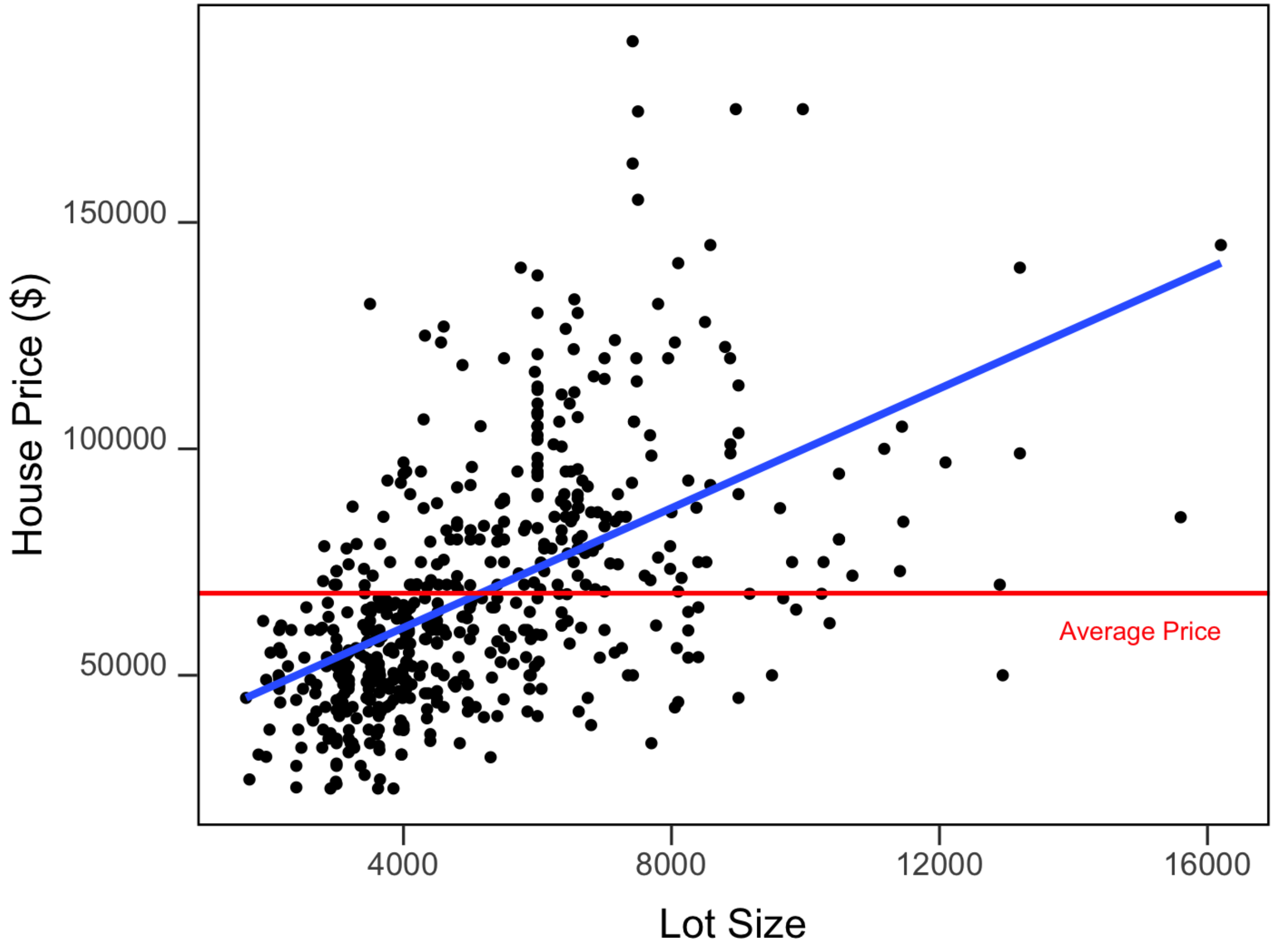
$$y_i = \hat{y}_i + \hat{u}_i, \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

now, subtracting $\bar{y}$ (sample average of y),

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{u}_i$$

- $y_i - \bar{y}$: how far away the actual value of y for ith observation from the sample average $\bar{y}$ is (actual deviation from the mean)
- $\hat{y}_i - \bar{y}$: how far away the predicted value of y for ith observation from the sample average $\bar{y}$ is (explained deviation from the mean)
- $\hat{u}_i$: the residual for ith observation

- $y_i - \bar{y}$
- $\hat{y}_i - \bar{y}$
- $\hat{u}_i$



total sum of squares (SST)

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

explained sum of squares (SSE)

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

residual sum of squares (SSR)

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

" Definition: R^2

$$R^2 = SSE/SST = 1 - SSR/SST$$

The value of R^2 always lies between 0 and 1 as long as an intercept is included in the econometric model.

What does it measure?

R^2 is a measure of how much improvement in predictin the depdent variable you've made by including independent variable(s) $(y = \beta_0 + \beta_1 x + u)$ compared to when simply using the mean of dependent variable as the predictor $(y = \beta_0 + u)$.

Important

- It tells nothing about how well you have estimated the causal ceteris paribus impact of x on y (β_1).
- As an economist, we typically do not care about how well we can prefict yield, rather we care about how well we have predicted β.

Problem

- While we observe the dependent variable (otherwise you cannot run regression), we cannot observe β_1.
- So, we get to check how good estimated models are in predicting the dependent variable (which we do not care), but we can never test whether they have estimated β_1 well.
- This means that we need to carefully examines whether the assumptions necessary for good estimation of β_1 is satisfied (next topic).

Small Sample Properties of OLS

Small sample property of OLS estimators

What is an estimator?

- A function of data that produces an estimate (actual number) of a parameter of interest once you plug in actual values of data
- OLS estimators: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

What is small sample property?

Properties that hold whatever the size of observation (small or large) is prior to obtaining actual estimates (before getting data)

- Put more simply: what can you expect from the estimators before you actually get data and obtain estimates?
- Difference between small sample property and the algebraic properties we looked at earlier?

OLS is just a way of using available information to obtain estimates. Does it have desirable properties?

- Unbiasedness
- Efficiency

As it turns out, OLS is a very good way of using available information!!

Unbiasedness

What does unbiased mean?

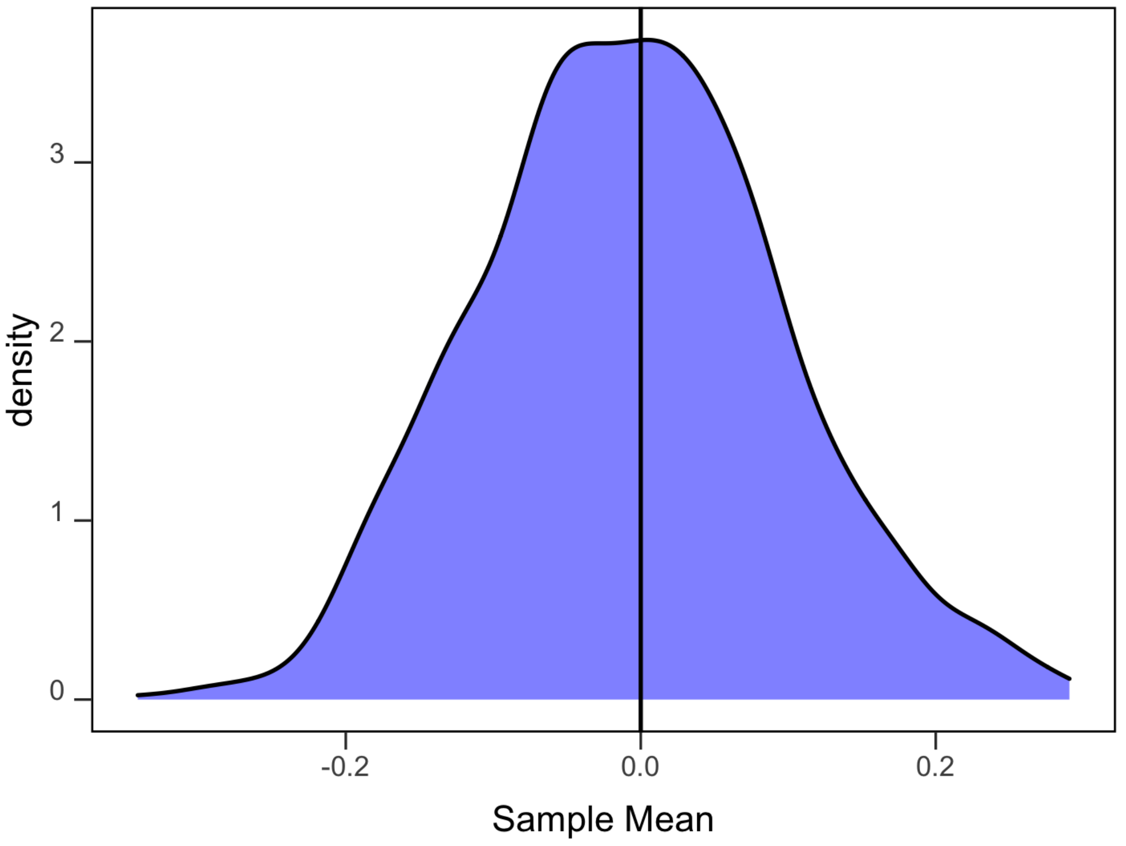
- Consider a problem of estimating the expected value of a single variable, x
- A good estimator is sample mean: $\frac{1}{n}\sum_{i}^{n} x_i$

R code: Sample Mean

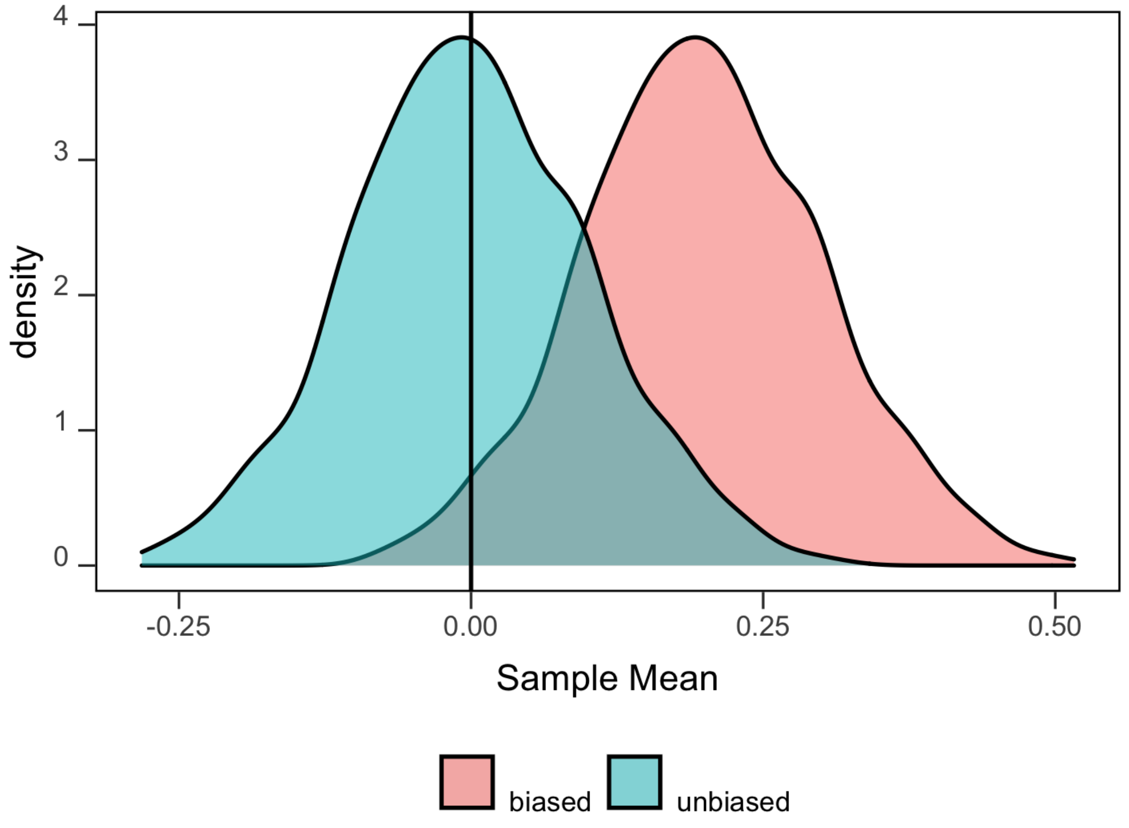
```
#--- set the number of observations ---#
n <- 100

#--- generate random values ---#
x_seq <- rnorm(n) # Normal(mean=0,sd=1)

#--- calcualte the mean ---#
mean(x_seq)
```

```
## [1] 0.03750092
```

This is what unbiased estimation looks like:



This is what biased estimation looks like:



Unbiasedness of OLS estimators

Unbiasedness of OLS estimators

Under certain conditions , OLS estimators are unbiased. That is,

$$E[\hat{\beta}_1] = E\Big[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\Big] = \beta_1$$

(We do not talk about unbiasedness of $\hat{\beta}_0$ because we are almost never interested in the intercept. Given the limited time we have, it is not worthwhile talking about it)

Certain Conditions

SLR.1: Linear in Parameters (Wooldridge, 2015)

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

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

(**Note** : this definition is from the textbook by Wooldridge)

SLR.2: Random sampling (Wooldridge, 2015)

We have a random sample of size n, $(x_i, y_i) : i = 1, 2, \ldots, n$, following the population model.

Non-random sampling

- Example: You observe income-education data only for those who have income higher than $\$25K$
- Benevolent and malevolent kinds:
  - exogenous sampling
  - endogenous sampling
- We discuss this in more detial later

SLR.3: Sample variation in covariates (Wooldridge, 2015)

The sample outcomes on x, namely, $x_i, i = 1, \ldots, n$, are not all the same value.

SLR.4: Zero conditional mean (Wooldridge, 2015)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

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

Along with random sampling condition, this implies that

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

Good and bad empiricists

Good Empiricists

- have ability to judge if the above conditions are satisfied for the particular context you are working on
- have ability to correct (if possible) for the problems associated with the violations of any of the above conditions
- knows the context well so you can make appropriate judgments

Unbiasedness of OLS estimators

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \left[\text{because } \sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = 0\right] \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SST_x} \quad \left[\text{where, } SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}\beta_1(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}
\end{aligned}
$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$$

$$\text{Since } \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \text{ and}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = SST_x,$$

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

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

Taking, expectation of $\hat{\beta}_1$ conditional on $\mathbf{x} = \{x_1, \ldots, x_n\}$,

$$\Rightarrow E[\hat{\beta}_1|\mathbf{x}] = E[\beta_1|\mathbf{x}] + E[(1/SST_x) \sum_{i=1}^{n} (x_i - \bar{x}) u_i|\mathbf{x}]$$

$$= \beta_1 + (1/SST_x) \sum_{i=1}^{n} (x_i - \bar{x}) E[u_i|\mathbf{x}]$$

So, if condition 4 $\left(E[u_i|\mathbf{x}] = 0\right)$ is satisfied,

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

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

Unbiasedness of OLS estimators

Reconsider the following example

$$price = \beta_0 + \beta_1 \times lotsize + u$$

- $price$: house price (USD)
- $lotsize$: lot size
- u: error term (everything else)

Questions

- What's in u?
- Do you think $E[u|x]$ is satisfied? In other words (roughly speaking), is u uncorrelated with x?

Important notes (again)

- Unbiasedness property of OLS estimators says nothing about the estimate that we obtain for a given sample
- It is always possible that we could obtain an unlucky sample that would give us a point estimate far from β_1, and we can never know for sure whether this is the case.

Variance of OLS estimator

- OLS estimators are random variables, which means that they have distributions
- OLS estimators have variance (how spread out OLS estimates can be)

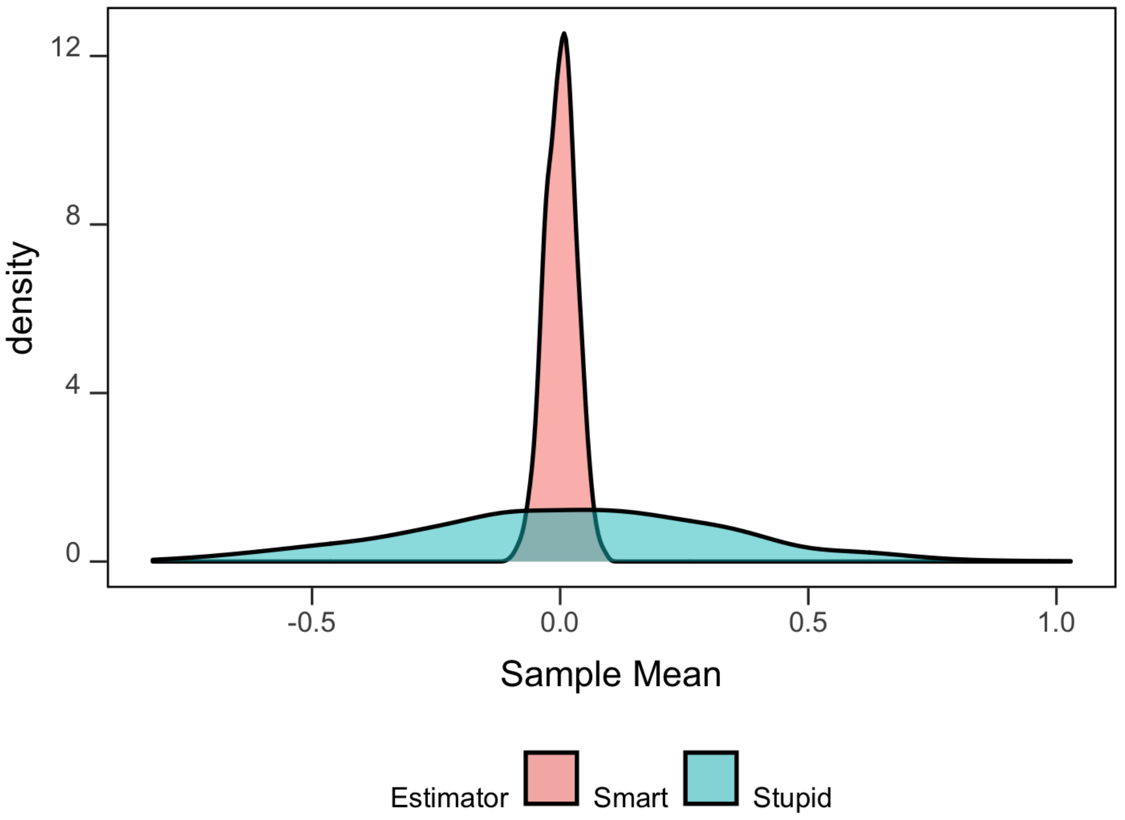
Example

Consider two estimators of $E[x]$:

$$\theta_{smart} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (n = 1000)$$

$$\theta_{stupid} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

Variance of estimators



Variance of OLS estimators

Variance of OLS estimators

If $Var(u|x) = \sigma^2$ and the four conditions (we used to prove unbiasedness of OLS estimators) are satisfied,

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

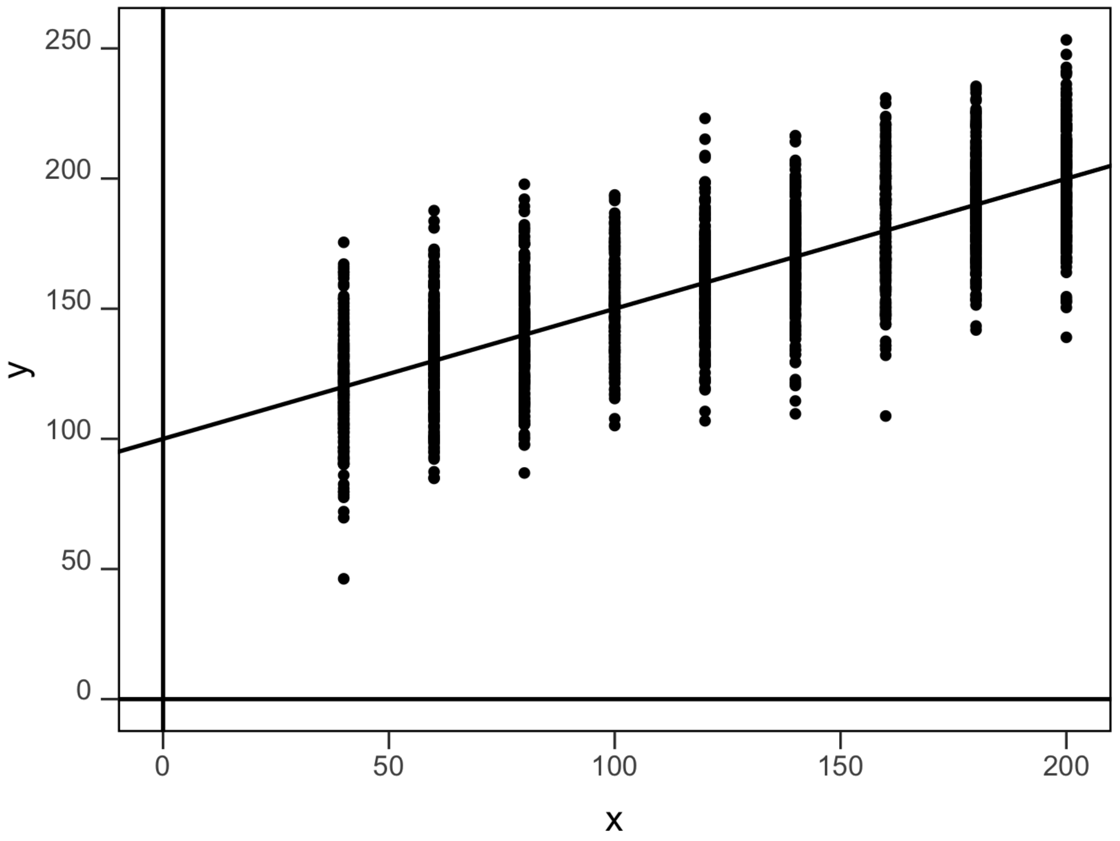
Homoskedasticity

The error u has the same variance give any value of the covariate x $(Var(u|x) = \sigma^2)$

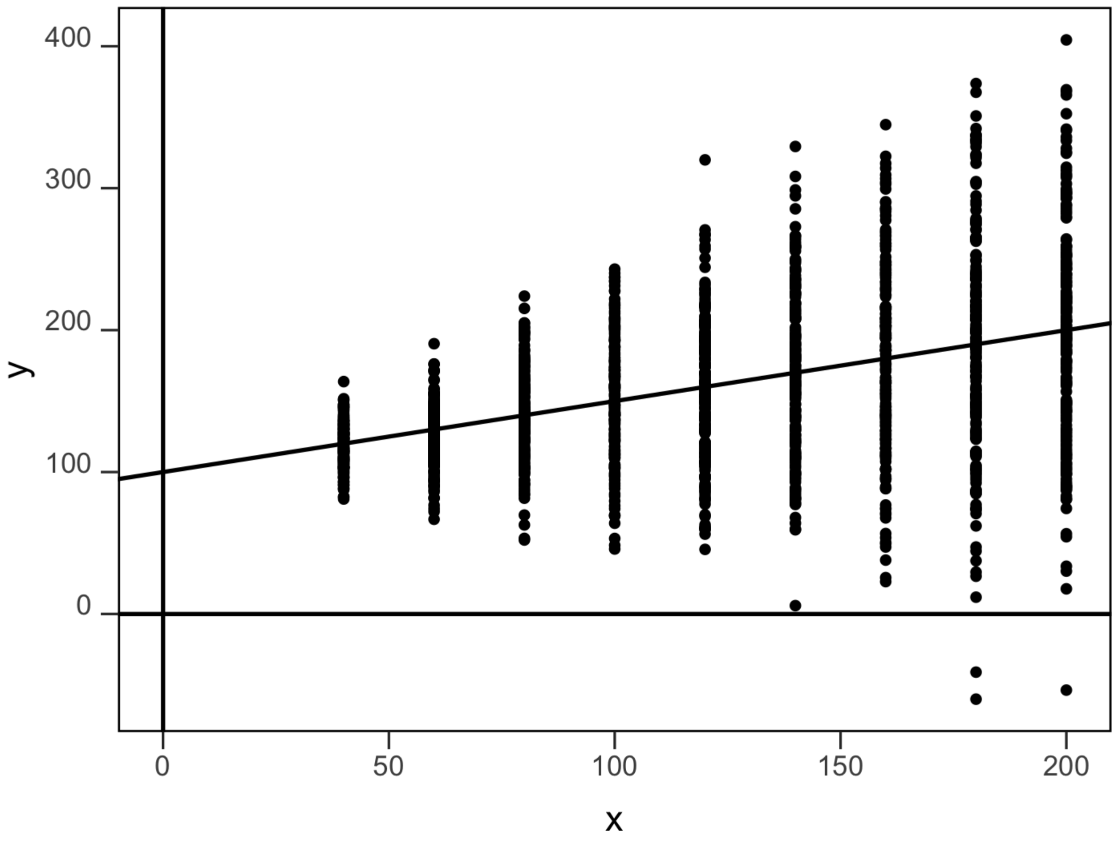
Heterokedasticity

The variance of the error u differs depending on the value of x $(Var(u|x) = f(x))$

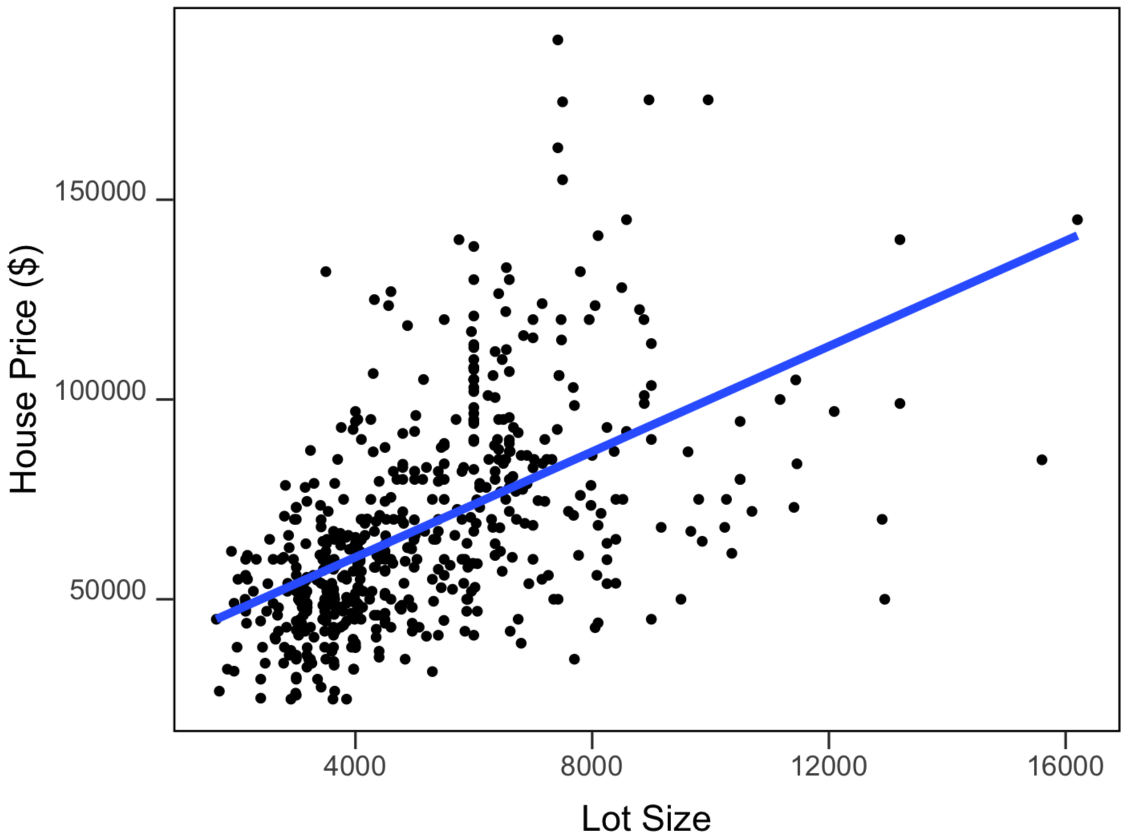
Homoskedastic Error



Heteroskedastic Error



House Price Example



Homoskedasticity Condition (Assumption)

- We did NOT use this condition to prove that OLS estimators are unbiased
- In most applications, homoskedasticity condition is not satisfied, which has important implications on:
  - estimation of variance (standard error) of OLS estimators
  - significance test

(A lot more on this issue later)

Variance of the OLS estimators

$$Var(\hat{\beta}_1|x) = \sigma^2/SST_x$$

What can you learn from this equation?

- the variance of OLS estimators is smaller (larger) if the variance of error term is smaller (larger)
- the greater (smaller) the variation in the covariate x, the smaller (larger) the variance of OLS estimators
  - if you are running experiments, spread the value of x as much as possible
  - you will rarely have this luxury

Gauss-Markov Theorem

Under conditions $SLR.1$ through $SLR.5$, OLS estimators are the best linear unbiased estimators (BLUEs)

In other words,

No other unbiased linear estimators have smaller variance than the OLS estimators (desirable efficiency property of OLS)

Estimating the error variance

- $Var(\hat{\beta}_1|x) = \sigma^2/SST_x$ will never be known. But, you can estimate it.
- Once you estimate $Var(\hat{\beta}_1|x)$, you can test the statistical significance of $\hat{\beta}_1$ (More on this later)

$$Var(u_i) = \sigma^2 = E[u_i^2] \ \left(Var(u_i) \equiv E[u_i^2] - E[u_i]^2\right)$$

- So, $\frac{1}{n}\sum_{i=1}^{n} u_i^2$ is an unbiased estimator of $Var(u_i)$
- What is the problem with this estimator?

We don't observe u_i (error), but we observe $\hat{u}_i$ (residuals)

Error and Residual

$$y_i = \beta_0 + \beta_1 x_i + u_i$$
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

Residuals as unbiased estimators of error

$$\begin{aligned} \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ \hat{u}_i &= \beta_0 + \beta_1 x_i + u_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ \Rightarrow \hat{u}_i - u_i &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i \\ \Rightarrow E[\hat{u}_i - u_i] &= E[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i] = 0 \end{aligned}$$

We know $E[\hat{u}_i - u_i] = 0$, so, why don't we use $\hat{u}_i$ (observable) in place of u_i (unobservable)?

How about $\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$ as an estimator of σ^2?

Unfortunately, $\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$ is a biased estimator of σ^2

Algebraic property of OLS

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ and } \sum_{i=1}^{n} x_i \hat{u}_i = 0$$

- this means that once you know the value of $n-2$ residuals, you can find the value of the other two by solving the above equations
- so, it's almost as if you have $n-2$ value of residuals instead of n

Unbiased estimator of the variance of the error term

We use $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$, which satisfies $E[\frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2] = \sigma^2$

Since $sd(\hat{\beta}_1) = \sigma / \sqrt{SST_x}$, the natural estimator of $sd(\hat{\beta}_1)$ is

$$se(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2} / \sqrt{SST_x},$$

which is called standard error of $\hat{\beta}_1$.

Later, we use $se(\hat{\beta}_1)$ for testing.

R code: Standard Error

```
## OLS estimation, Dep. Var.: price
## Observations: 546
## Standard-errors: Standard
##               Estimate  Std. Error
## (Intercept) 34136.2000 2491.100000
## lotsize         6.5988    0.445847
##              t value Pr(>|t|))
## (Intercept)  13.703 < 2.2e-16 ***
## lotsize      14.801 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 22,525.7   Adj. R2: 0.285766
```

Functional Form and Scale

Functional Form

Note

- transformation of variables is allowed without disturbing our analytical framework as long as the model is linear in parameter .
- transformation of variables change the interpretation of the coefficients estimates

Golas

- present popular functional forms
- use simple calculus to examine the interpretation of the coefficients

log-linear

$$log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

linear-log

$$y_i = \beta_0 + \beta_1 log(x_i) + u_i$$

log-log

$$log(y_i) = \beta_0 + \beta_1 log(x_i) + u_i$$

Log-linear functional form

Model

$$log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

Calculus

Differentiating the both sides wrt x_i,

$$\frac{1}{y_i} \cdot \frac{\partial y_i}{\partial x_i} = \beta_1 \Rightarrow \frac{\Delta y_i}{y_i} = \beta_1 \Delta x_i$$

Interpretation

β_1 measures a percentage change in y_i when x_i is increased by one unit

Log-linear model

Model

$$log(wage) = \beta_0 + \beta_1 educ + u$$

Calculus

Differentiating both sides with respect to $educ$,

$$\frac{1}{wage}\frac{\partial wage}{\partial educ} = \beta_1 \Rightarrow \frac{\Delta wage}{wage} = \beta_1 \Delta educ$$

Interpretation

If education increases by 1 year $(\Delta educ = 1)$, then wage increases by $\beta_1 * 100\%$ $(\frac{\Delta wage}{wage} = \beta_1)$

Log-linear model: Example

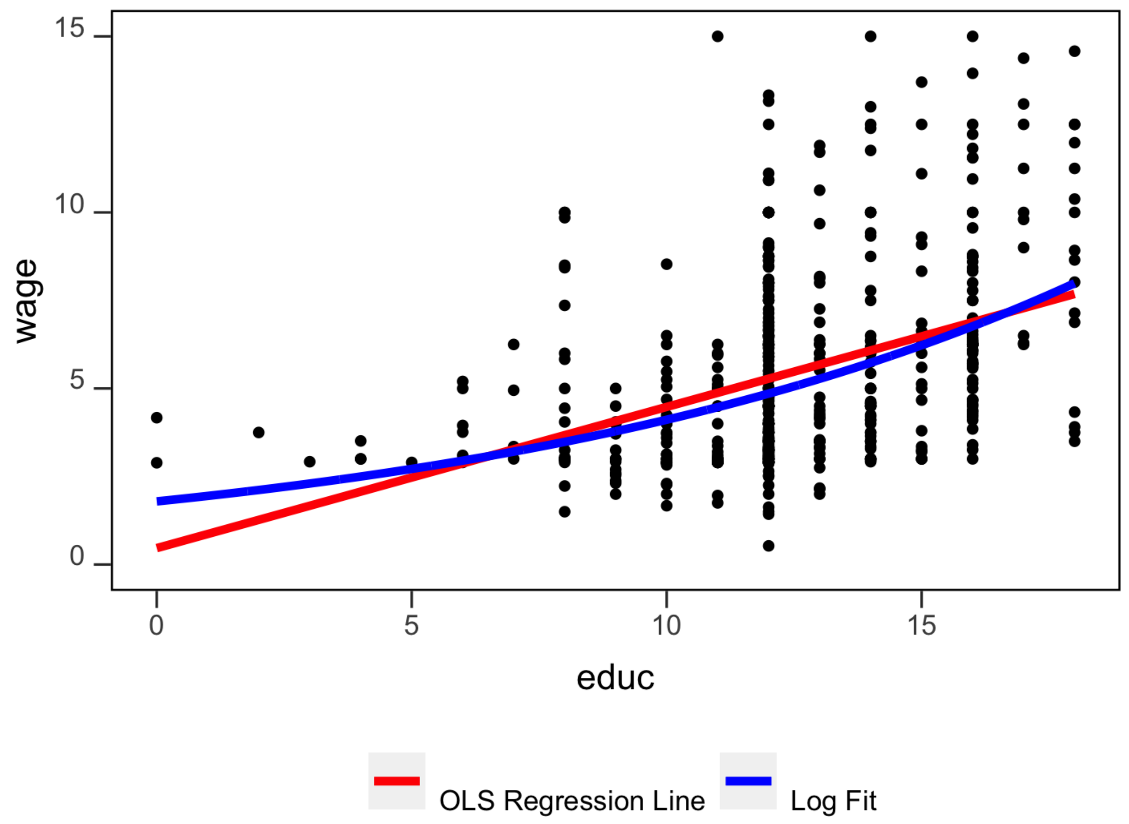
If you estimate the following model using the wage dataset:

$$log(wage) = \beta_0 + \beta_1 educ + u$$

Then, the estimated equation is the following:

$$log(\widehat{wage}) = 0.584 + 0.083educ$$

$$E[\widehat{wage}] = e^{0.584+0.083educ}$$



wage
educ
15
10
5
0
0
5
10
15
OLS Regression Line
Log Fit

Functional form: Linear-log

Model

$$y_i = \beta_0 + \beta_1 log(x_i) + u_i$$

Calculus

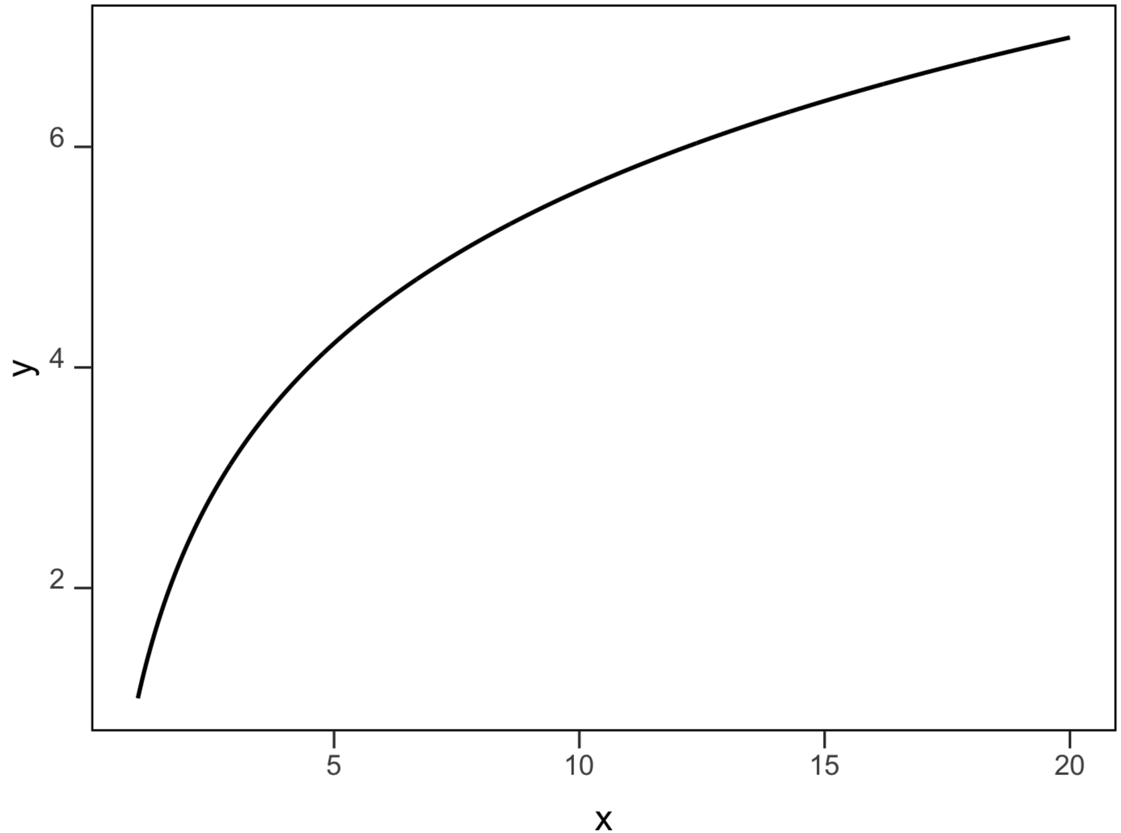
Differentiating the both sides wrt x_i,

$$\frac{\partial y_i}{\partial x_i} = \beta_1 / x_i \Rightarrow \Delta y_i = \beta_1 \frac{\Delta x_i}{x_i}$$

Interpretation

When x increases by 1%, y increases by β_1

$$y = \beta_0 + \beta_1 log(x) = \mathbf{1} + \mathbf{2} \times log(x)$$



Functional form: Log-log

Model

$$log(y_i) = \beta_0 + \beta_1 log(x_i) + u_i$$

Calculus

Differentiating the both sides wrt x_i,

$$\frac{\partial y_i}{y_i} / \frac{\partial x_i}{x_i} = \beta_1 \Rightarrow \frac{\Delta y_i}{y_i} = \beta_1 \frac{\Delta x_i}{x_i}$$

Interpretation

A percentage change in x would result in a β_1 percentage change in y_i (constant elasticity)

Simple Linear Regression

- In these models, the dependent variable and independent variable are non-linearly related, how come are these models called simple linear model?
- linear in simple linear model means that the model is linear in parameter , but not in variable

Non-linear (in parameter) Models

Example

$$y_i = \beta_0 + x_i^{\beta_1} + u_i$$
$$y_i = \frac{x_i}{\beta_0 + \beta_1 x_i} + u_i$$

Notes

Transformation of the dependent and independent variables would not affect the propertirs of the OLS estimator as long as the model is linear in parameter.