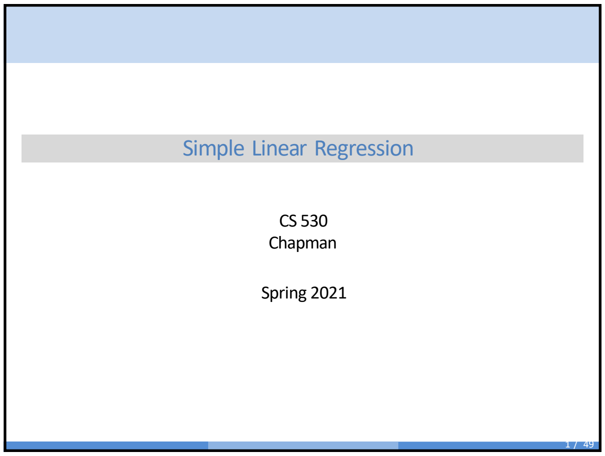


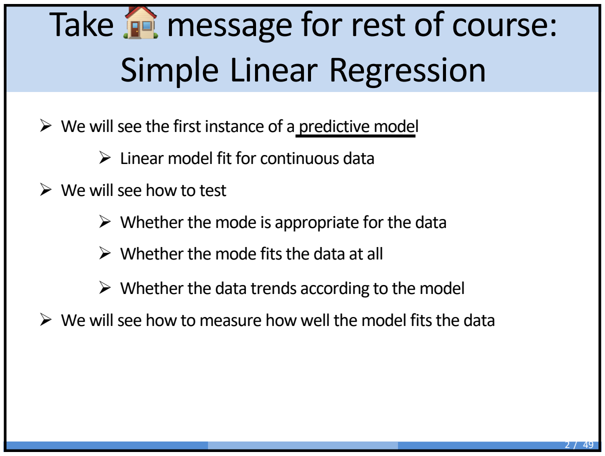
Simple Linear Regression

CS 530
Chapman

Spring 2021

1 / 49

1

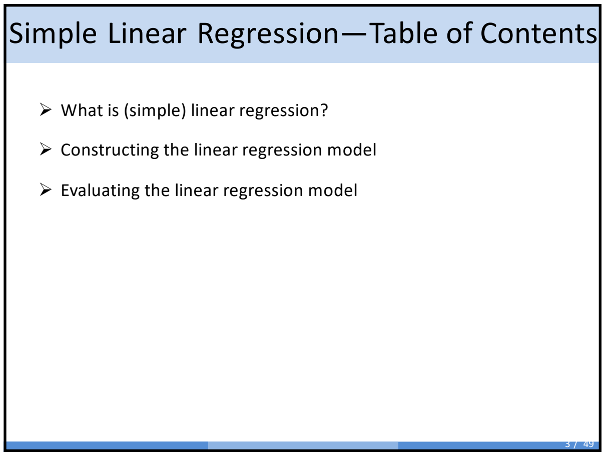


Take 🏠 message for rest of course: Simple Linear Regression

- We will see the first instance of a predictive model
  - Linear model fit for continuous data
- We will see how to test
  - Whether the mode is appropriate for the data
  - Whether the mode fits the data at all
  - Whether the data trends according to the model
- We will see how to measure how well the model fits the data

2 / 49

2

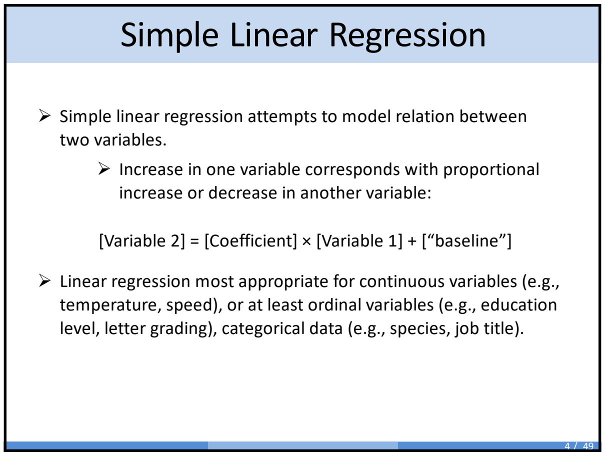


Simple Linear Regression—Table of Contents

- What is (simple) linear regression?
- Constructing the linear regression model
- Evaluating the linear regression model

3 / 49

3



Simple Linear Regression

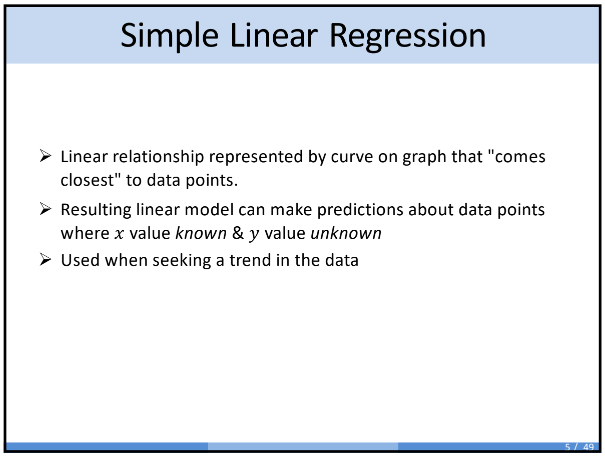
- Simple linear regression attempts to model relation between two variables.
  - Increase in one variable corresponds with proportional increase or decrease in another variable:

[Variable 2] = [Coefficient] × [Variable 1] + [“baseline”]

- Linear regression most appropriate for continuous variables (e.g., temperature, speed), or at least ordinal variables (e.g., education level, letter grading), categorical data (e.g., species, job title).

4 / 49

4

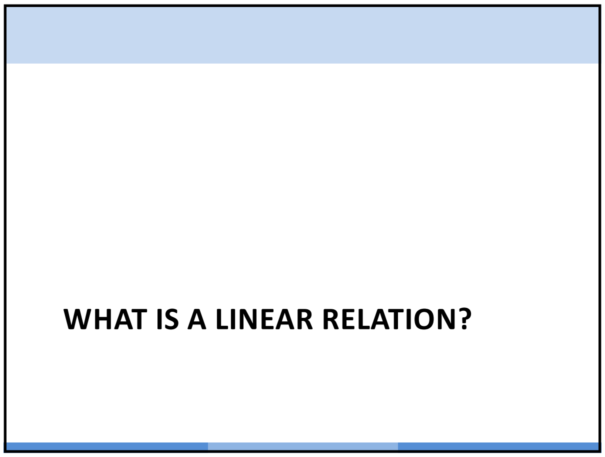


Simple Linear Regression

- Linear relationship represented by curve on graph that "comes closest" to data points.
- Resulting linear model can make predictions about data points where x value *known* & y value *unknown*
- Used when seeking a trend in the data

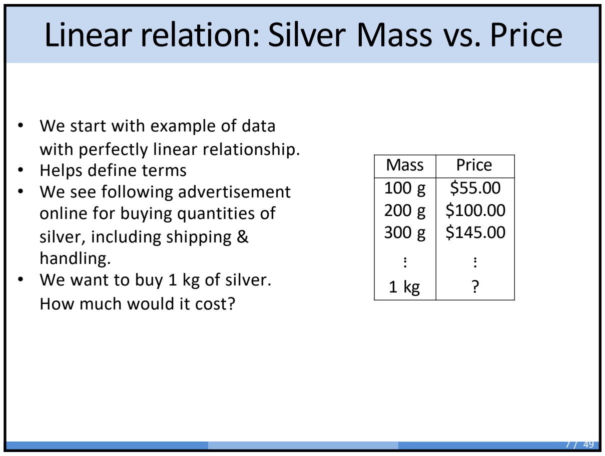
5 / 49

5



WHAT IS A LINEAR RELATION?

6



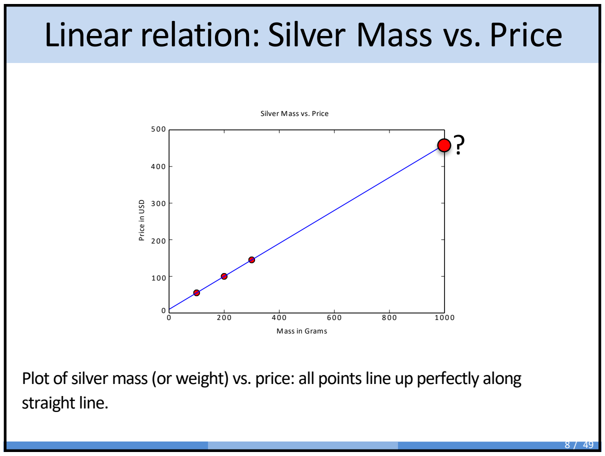
Linear relation: Silver Mass vs. Price

- We start with example of data with perfectly linear relationship.
- Helps define terms
- We see following advertisement online for buying quantities of silver, including shipping & handling.
- We want to buy 1 kg of silver. How much would it cost?

Mass	Price
100 g	$55.00
200 g	$100.00
300 g	$145.00
⋮	⋮
1 kg	?

7 / 49

7

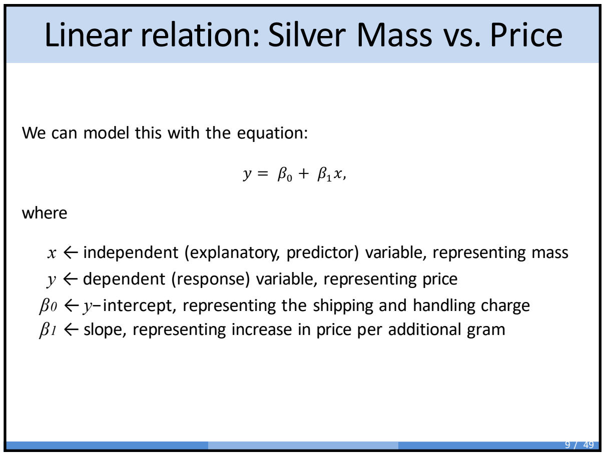


Linear relation: Silver Mass vs. Price

Plot of silver mass (or weight) vs. price: all points line up perfectly along straight line.

8 / 49

8



Linear relation: Silver Mass vs. Price

We can model this with the equation:

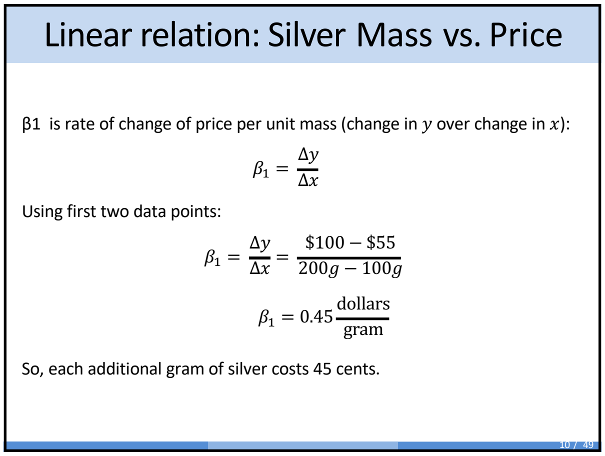
$$y = \beta_0 + \beta_1 x,$$

where

$x \leftarrow$ independent (explanatory, predictor) variable, representing mass
$y \leftarrow$ dependent (response) variable, representing price
$\beta_0 \leftarrow y-$intercept, representing the shipping and handling charge
$\beta_1 \leftarrow$ slope, representing increase in price per additional gram

9 / 49

9



Linear relation: Silver Mass vs. Price

β1 is rate of change of price per unit mass (change in y over change in x):

$$\beta_1 = \frac{\Delta y}{\Delta x}$$

Using first two data points:

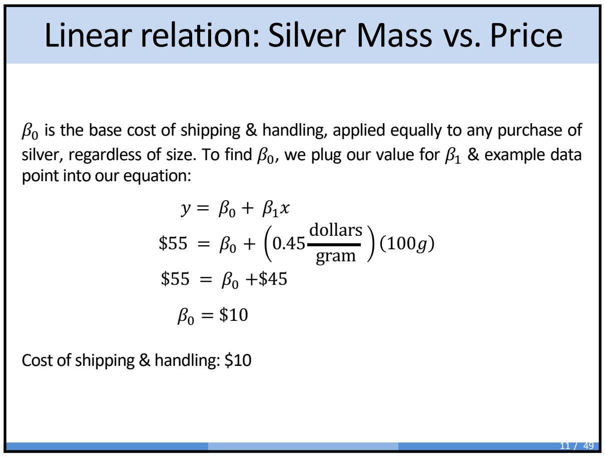
$$\beta_1 = \frac{\Delta y}{\Delta x} = \frac{\$100 - \$55}{200g - 100g}$$

$$\beta_1 = 0.45\frac{\text{dollars}}{\text{gram}}$$

So, each additional gram of silver costs 45 cents.

10 / 49

10



Linear relation: Silver Mass vs. Price

β_0 is the base cost of shipping & handling, applied equally to any purchase of silver, regardless of size. To find β_0, we plug our value for β_1 & example data point into our equation:

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

$$\$55 = \beta_0 + \left(0.45\frac{\text{dollars}}{\text{gram}}\right)(100g)$$

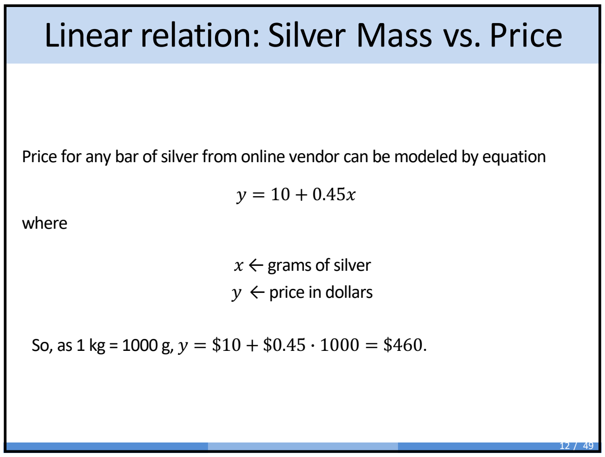
$$\$55 = \beta_0 + \$45$$

$$\beta_0 = \$10$$

Cost of shipping & handling: $10

11 / 49

11



Linear relation: Silver Mass vs. Price

Price for any bar of silver from online vendor can be modeled by equation

$$y = 10 + 0.45x$$

where

$x \leftarrow$ grams of silver
$y \leftarrow$ price in dollars

So, as 1 kg = 1000 g, $y = \$10 + \$0.45 \cdot 1000 = \$460.$

12 / 49

12

LINEAR REGRESSION: INTRODUCTION VIA SIMPLE EXAMPLE

13

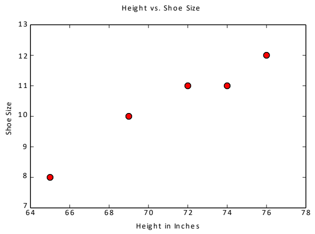
Linear Regression: Height vs. Shoe Size

Simple example: group of men with following heights & shoe sizes. What shoe size would we predict for man of average American height (70 inches)?

Height (inches)	Shoe Size
72	11
74	11
76	12
69	10
65	8
⋮	⋮
70	?

14

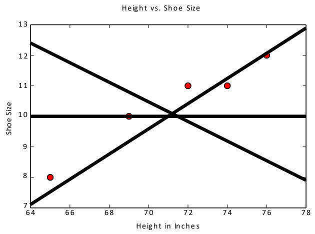
Linear Regression: Height vs. Shoe Size



Scatter plot of height vs. shoe size.
This time it is not a perfect linear relation.

15

Linear Regression: Height vs. Shoe Size



There is infinite number of lines—infinite models—to potentially fit to this data. What is "best fit" line—the "best" model?

16

Linear Regression: Height vs. Shoe Size

Data points don't fit perfectly onto line, but relationship still appears "line-like" in nature. We can model this with equation:

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

where

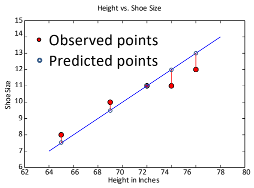
x ← man's height in inches
y ← man's shoe size
β_0 ← theoretical shoe size for a man who is 0" tall
β_1 ← increase in shoe size for each additional 1" in height
ϵ ← error term (a.k.a. noise)

17

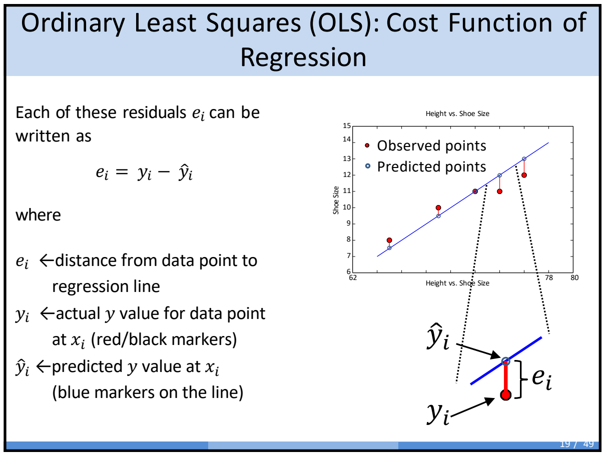
Ordinary Least Squares (OLS): Cost Function of Regression

No line can pass through every point in this data set.
Vertical distances between the observed data points & our model line termed residuals—represented by red vertical lines in the graph.

Model (line) with best fit defined as one with "least overall distance" between observed points & predicted points—*best-fit model minimizes sum of residuals.*



18



Ordinary Least Squares (OLS): Cost Function of Regression

Each of these residuals e_i can be written as

$$e_i = y_i - \hat{y}_i$$

where

e_i ←distance from data point to regression line

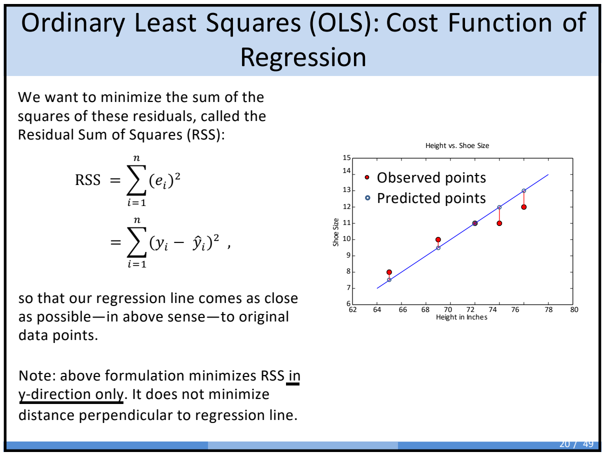
y_i ←actual y value for data point at x_i (red/black markers)

$\hat{y}_i$ ←predicted y value at x_i (blue markers on the line)

Height vs. Shoe Size
Observed points
Predicted points
Shoe Size
Height vs. Shoe Size
$\hat{y}_i$
e_i
y_i

19 / 49

19



Ordinary Least Squares (OLS): Cost Function of Regression

We want to minimize the sum of the squares of these residuals, called the Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (e_i)^2$$

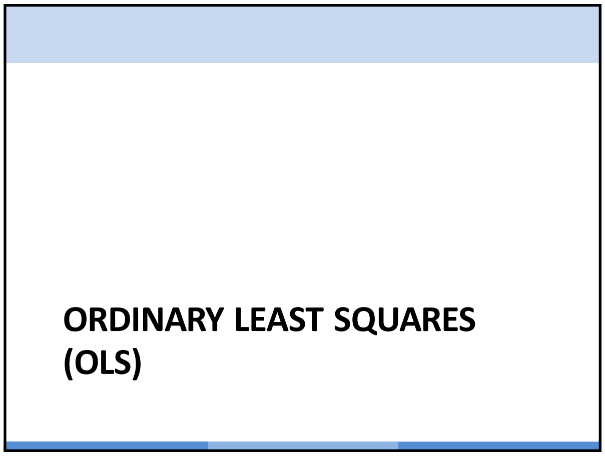
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ ,$$

so that our regression line comes as close as possible—in above sense—to original data points.

Note: above formulation minimizes RSS in y-direction only. It does not minimize distance perpendicular to regression line.

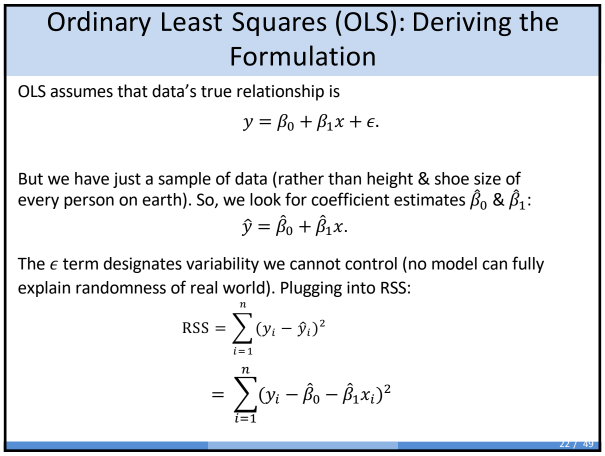
20 / 49

20



ORDINARY LEAST SQUARES (OLS)

21



Ordinary Least Squares (OLS): Deriving the Formulation

OLS assumes that data's true relationship is

$$y = \beta_0 + \beta_1 x + \epsilon.$$

But we have just a sample of data (rather than height & shoe size of every person on earth). So, we look for coefficient estimates $\hat{\beta}_0$ & $\hat{\beta}_1$:

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

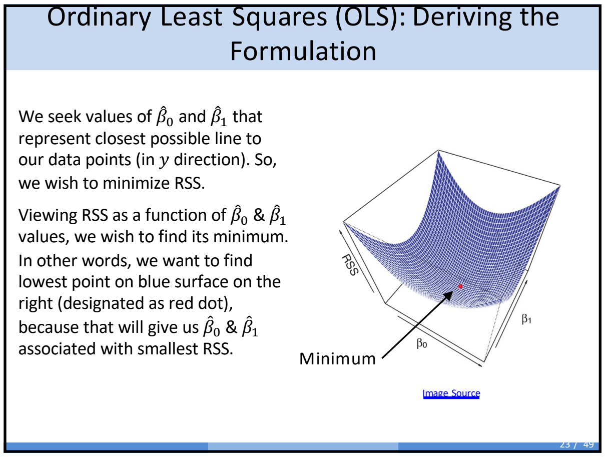
The ϵ term designates variability we cannot control (no model can fully explain randomness of real world). Plugging into RSS:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

22 / 49

22



Ordinary Least Squares (OLS): Deriving the Formulation

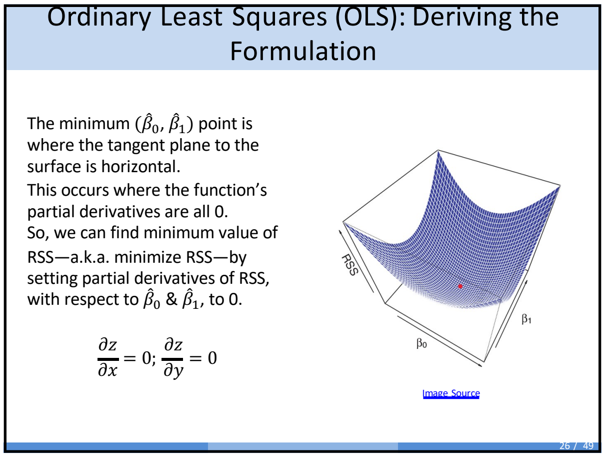
We seek values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that represent closest possible line to our data points (in y direction). So, we wish to minimize RSS.

Viewing RSS as a function of $\hat{\beta}_0$ & $\hat{\beta}_1$ values, we wish to find its minimum. In other words, we want to find lowest point on blue surface on the right (designated as red dot), because that will give us $\hat{\beta}_0$ & $\hat{\beta}_1$ associated with smallest RSS.

Image Source

23 / 49

23



Ordinary Least Squares (OLS): Deriving the Formulation

The minimum $(\hat{\beta}_0, \hat{\beta}_1)$ point is where the tangent plane to the surface is horizontal.

This occurs where the function's partial derivatives are all 0.

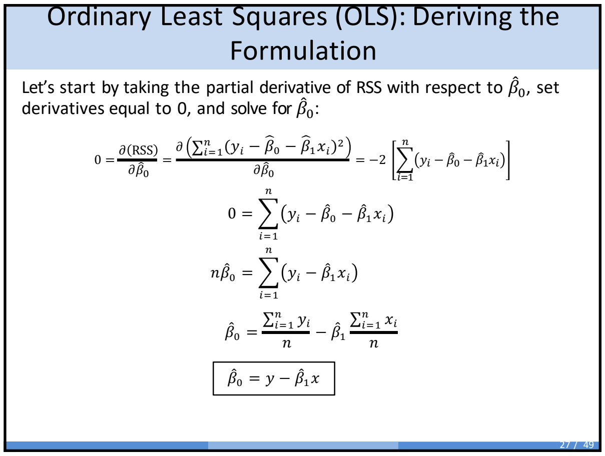
So, we can find minimum value of RSS—a.k.a. minimize RSS—by setting partial derivatives of RSS, with respect to $\hat{\beta}_0$ & $\hat{\beta}_1$, to 0.

$$\frac{\partial z}{\partial x} = 0; \frac{\partial z}{\partial y} = 0$$

Image Source

26 / 49

26



Ordinary Least Squares (OLS): Deriving the Formulation

Let's start by taking the partial derivative of RSS with respect to $\hat{\beta}_0$, set derivatives equal to 0, and solve for $\hat{\beta}_0$:

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

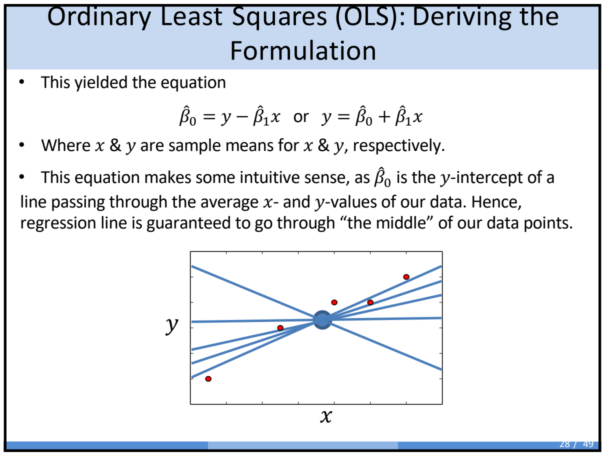
$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

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

27



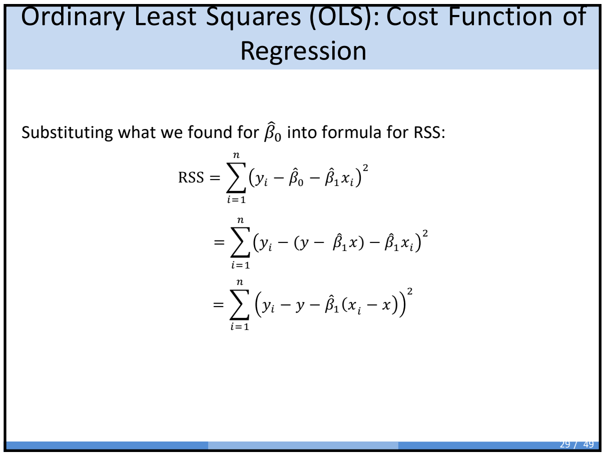
Ordinary Least Squares (OLS): Deriving the Formulation

- This yielded the equation

$$\hat{\beta}_0 = y - \hat{\beta}_1 x \quad \text{or} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Where x & y are sample means for x & y, respectively.
- This equation makes some intuitive sense, as $\hat{\beta}_0$ is the y-intercept of a line passing through the average x- and y-values of our data. Hence, regression line is guaranteed to go through "the middle" of our data points.

28



Ordinary Least Squares (OLS): Cost Function of Regression

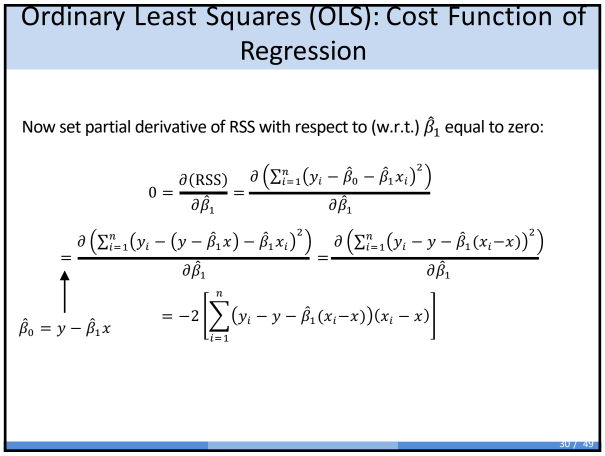
Substituting what we found for $\hat{\beta}_0$ into formula for RSS:

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

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

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

29



Ordinary Least Squares (OLS): Cost Function of Regression

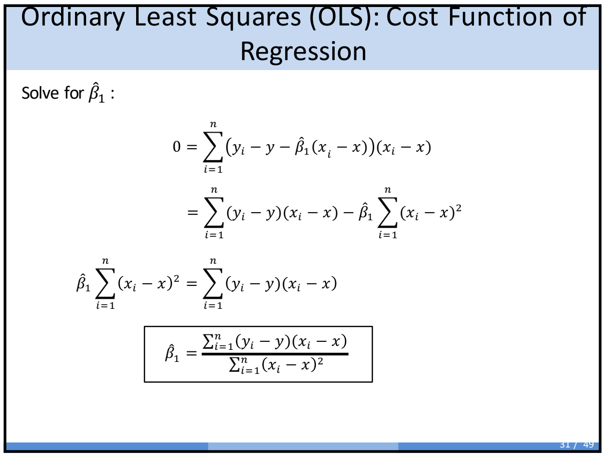
Now set partial derivative of RSS with respect to (w.r.t.) $\hat{\beta}_1$ equal to zero:

$$0 = \frac{\partial(\mathrm{RSS})}{\partial\hat{\beta}_1} = \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\right)}{\partial\hat{\beta}_1}$$

$$\underset{\uparrow \; \hat{\beta}_0 = y - \hat{\beta}_1 x}{=} \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - (y - \hat{\beta}_1 x) - \hat{\beta}_1 x_i\right)^2\right)}{\partial\hat{\beta}_1} = \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - y - \hat{\beta}_1 (x_i - x)\right)^2\right)}{\partial\hat{\beta}_1}$$

$$= -2\left[\sum_{i=1}^{n}\left(y_i - y - \hat{\beta}_1 (x_i - x)\right)(x_i - x)\right]$$

30



Ordinary Least Squares (OLS): Cost Function of Regression

Solve for $\hat{\beta}_1$:

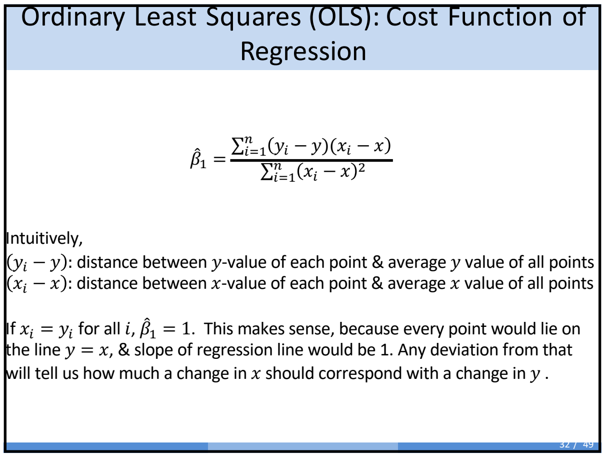
$$0 = \sum_{i=1}^{n}\left(y_i - y - \hat{\beta}_1 (x_i - x)\right)(x_i - x)$$

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

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

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

31



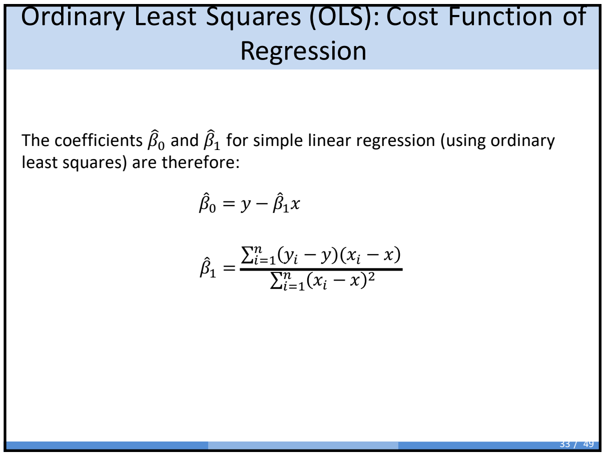
Ordinary Least Squares (OLS): Cost Function of Regression

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

Intuitively,
$(y_i - y)$: distance between y-value of each point & average y value of all points
$(x_i - x)$: distance between x-value of each point & average x value of all points

If $x_i = y_i$ for all i, $\hat{\beta}_1 = 1$. This makes sense, because every point would lie on the line $y = x$, & slope of regression line would be 1. Any deviation from that will tell us how much a change in x should correspond with a change in y .

32



Ordinary Least Squares (OLS): Cost Function of Regression

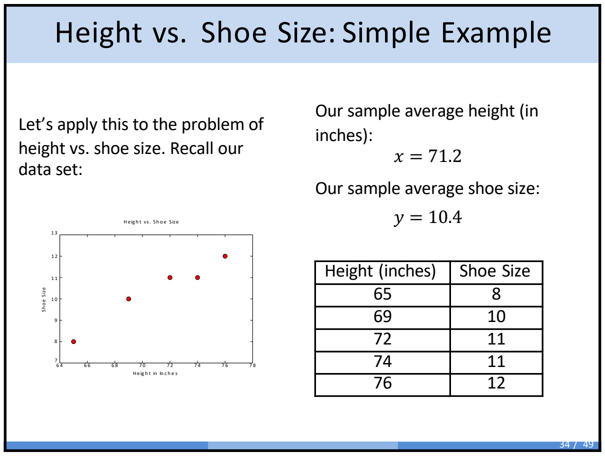
The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ for simple linear regression (using ordinary least squares) are therefore:

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

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

33 / 49

33



Height vs. Shoe Size: Simple Example

Let's apply this to the problem of height vs. shoe size. Recall our data set:

Height vs. Shoe Size
Shoe Size
Height in Inches
7
8
9
10
11
12
13
64
66
68
70
72
74
76
78

Our sample average height (in inches):

$$x = 71.2$$

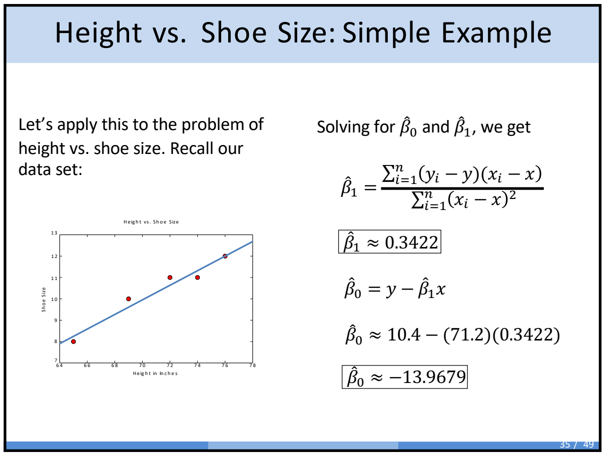
Our sample average shoe size:

$$y = 10.4$$

Height (inches)	Shoe Size
65	8
69	10
72	11
74	11
76	12

34 / 49

34



Height vs. Shoe Size: Simple Example

Let's apply this to the problem of height vs. shoe size. Recall our data set:

Height vs. Shoe Size
Shoe Size
Height in Inches
7
8
9
10
11
12
13
64
66
68
70
72
74
76
78

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$, we get

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

$$\boxed{\hat{\beta}_1 \approx 0.3422}$$

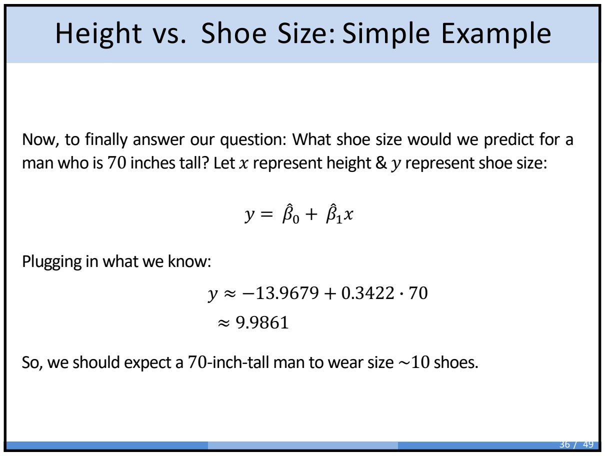
$$\hat{\beta}_0 = y - \hat{\beta}_1 x$$

$$\hat{\beta}_0 \approx 10.4 - (71.2)(0.3422)$$

$$\boxed{\hat{\beta}_0 \approx -13.9679}$$

35 / 49

35



Height vs. Shoe Size: Simple Example

Now, to finally answer our question: What shoe size would we predict for a man who is 70 inches tall? Let x represent height & y represent shoe size:

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

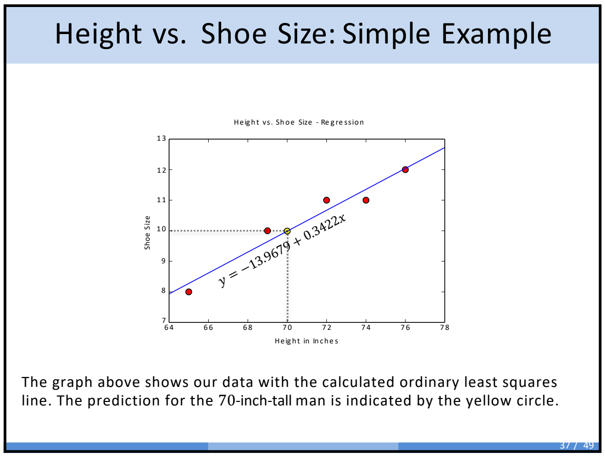
Plugging in what we know:

$$y \approx -13.9679 + 0.3422 \cdot 70$$
$$\approx 9.9861$$

So, we should expect a 70-inch-tall man to wear size ~10 shoes.

36 / 49

36

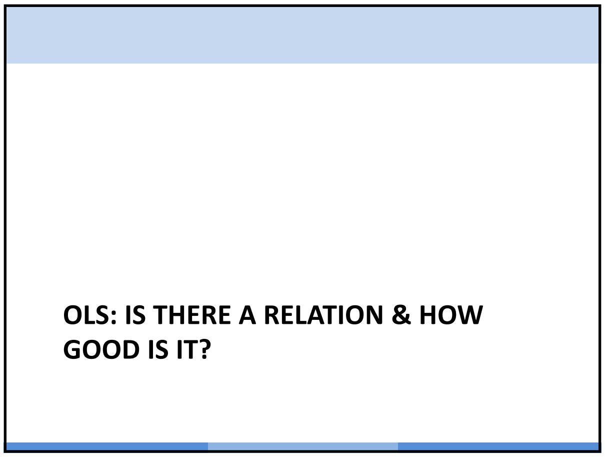


Height vs. Shoe Size: Simple Example

The graph above shows our data with the calculated ordinary least squares line. The prediction for the 70-inch-tall man is indicated by the yellow circle.

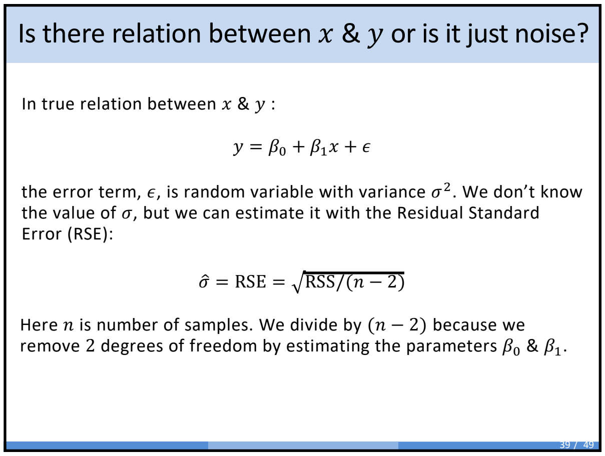
37 / 49

37



OLS: IS THERE A RELATION & HOW GOOD IS IT?

38



Is there relation between x & y or is it just noise?

In true relation between x & y :

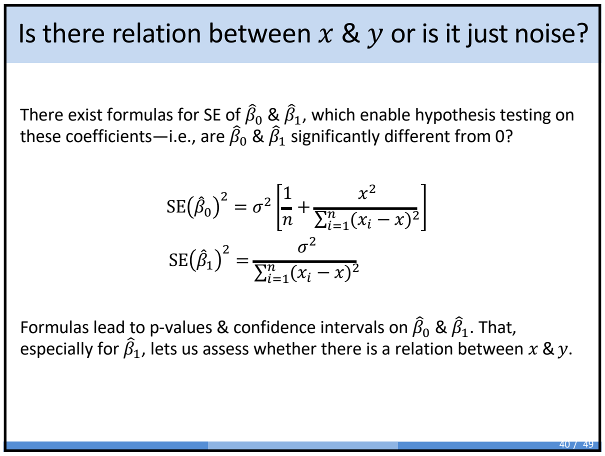
$$y = \beta_0 + \beta_1 x + \epsilon$$

the error term, ϵ, is random variable with variance σ^2. We don't know the value of σ, but we can estimate it with the Residual Standard Error (RSE):

$$\hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

Here n is number of samples. We divide by $(n-2)$ because we remove 2 degrees of freedom by estimating the parameters β_0 & β_1.

39



Is there relation between x & y or is it just noise?

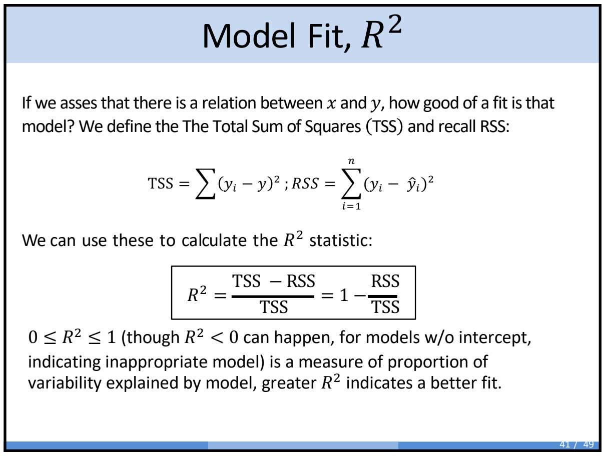
There exist formulas for SE of $\hat{\beta}_0$ & $\hat{\beta}_1$, which enable hypothesis testing on these coefficients—i.e., are $\hat{\beta}_0$ & $\hat{\beta}_1$ significantly different from 0?

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{x^2}{\sum_{i=1}^{n}(x_i - x)^2}\right]$$

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

Formulas lead to p-values & confidence intervals on $\hat{\beta}_0$ & $\hat{\beta}_1$. That, especially for $\hat{\beta}_1$, lets us assess whether there is a relation between x & y.

40



Model Fit, R^2

If we asses that there is a relation between x and y, how good of a fit is that model? We define the The Total Sum of Squares (TSS) and recall RSS:

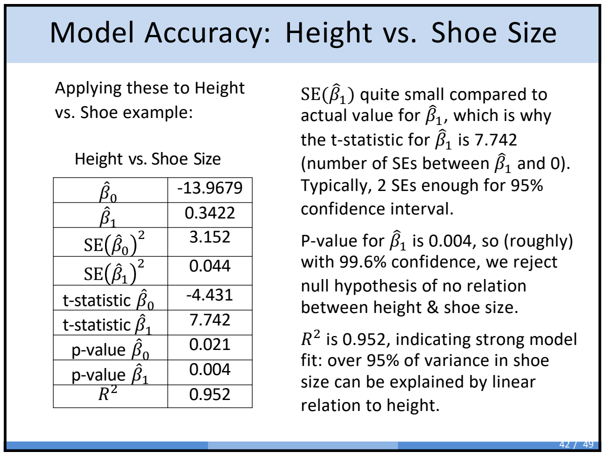
$$\mathrm{TSS} = \sum (y_i - y)^2\,;RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We can use these to calculate the R^2 statistic:

$$\boxed{R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}}$$

$0 \le R^2 \le 1$ (though $R^2 < 0$ can happen, for models w/o intercept, indicating inappropriate model) is a measure of proportion of variability explained by model, greater R^2 indicates a better fit.

41



Model Accuracy: Height vs. Shoe Size

Applying these to Height vs. Shoe example:

Height vs. Shoe Size

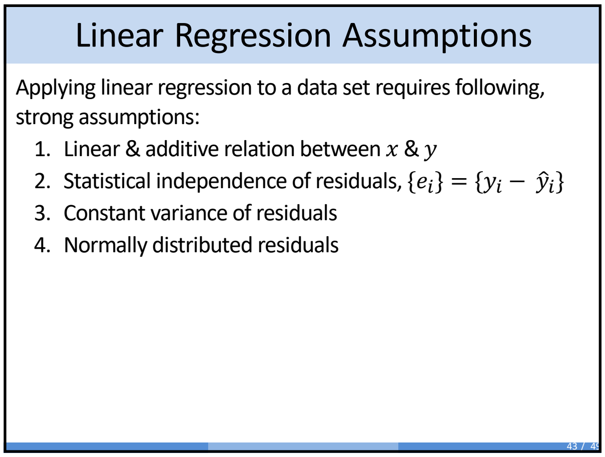
$\hat{\beta}_0$	-13.9679
$\hat{\beta}_1$	0.3422
$\mathrm{SE}(\hat{\beta}_0)^2$	3.152
$\mathrm{SE}(\hat{\beta}_1)^2$	0.044
t-statistic $\hat{\beta}_0$	-4.431
t-statistic $\hat{\beta}_1$	7.742
p-value $\hat{\beta}_0$	0.021
p-value $\hat{\beta}_1$	0.004
R^2	0.952

$\mathrm{SE}(\hat{\beta}_1)$ quite small compared to actual value for $\hat{\beta}_1$, which is why the t-statistic for $\hat{\beta}_1$ is 7.742 (number of SEs between $\hat{\beta}_1$ and 0). Typically, 2 SEs enough for 95% confidence interval.

P-value for $\hat{\beta}_1$ is 0.004, so (roughly) with 99.6% confidence, we reject null hypothesis of no relation between height & shoe size.

R^2 is 0.952, indicating strong model fit: over 95% of variance in shoe size can be explained by linear relation to height.

42

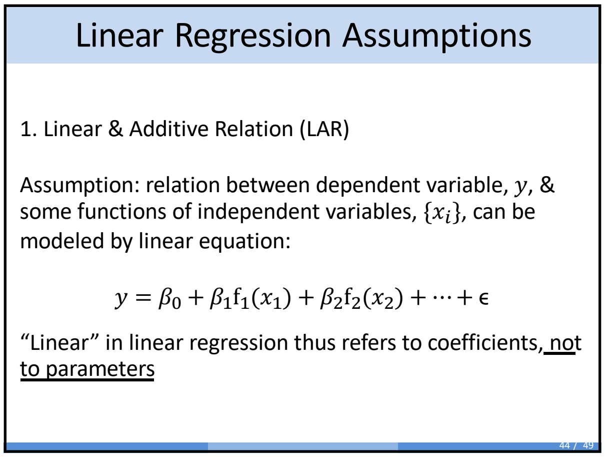


Linear Regression Assumptions

Applying linear regression to a data set requires following, strong assumptions:

1. Linear & additive relation between x & y
2. Statistical independence of residuals, $\{e_i\} = \{y_i - \hat{y}_i\}$
3. Constant variance of residuals
4. Normally distributed residuals

43



Linear Regression Assumptions

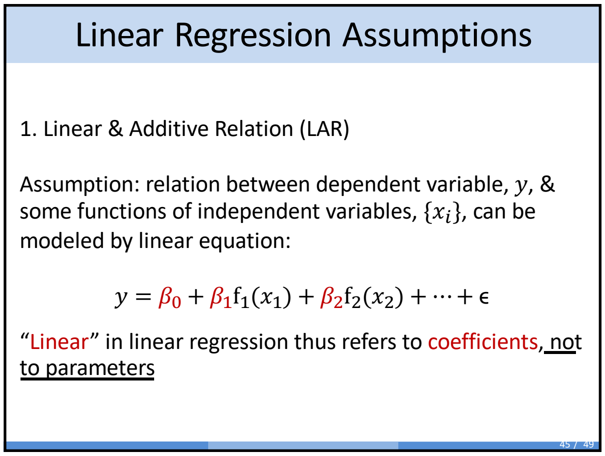
1. Linear & Additive Relation (LAR)

Assumption: relation between dependent variable, y, & some functions of independent variables, $\{x_i\}$, can be modeled by linear equation:

$$y = \beta_0 + \beta_1 \mathrm{f}_1(x_1) + \beta_2 \mathrm{f}_2(x_2) + \cdots + \epsilon$$

"Linear" in linear regression thus refers to coefficients, not to parameters

44



Linear Regression Assumptions

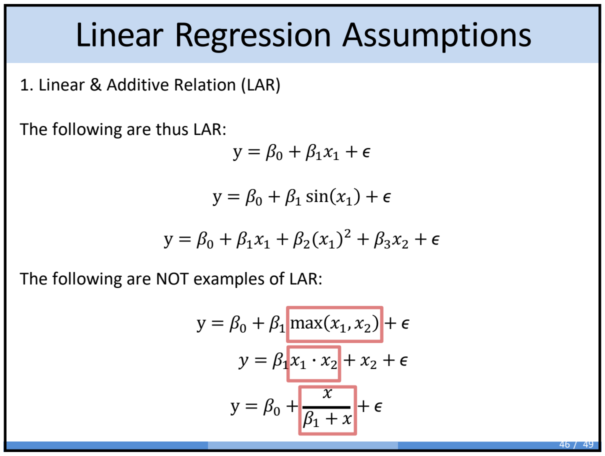
1. Linear & Additive Relation (LAR)

Assumption: relation between dependent variable, y, & some functions of independent variables, $\{x_i\}$, can be modeled by linear equation:

$$y = \beta_0 + \beta_1 \mathrm{f}_1(x_1) + \beta_2 \mathrm{f}_2(x_2) + \cdots + \epsilon$$

"Linear" in linear regression thus refers to coefficients, not to parameters

45



Linear Regression Assumptions

1. Linear & Additive Relation (LAR)

The following are thus LAR:

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

$$y = \beta_0 + \beta_1 \sin(x_1) + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (x_1)^2 + \beta_3 x_2 + \epsilon$$

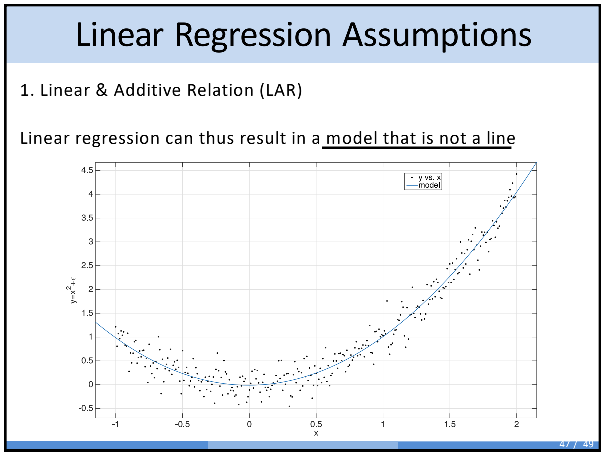
The following are NOT examples of LAR:

$$y = \beta_0 + \beta_1 \boxed{\max(x_1, x_2)} + \epsilon$$

$$y = \beta_1 \boxed{x_1 \cdot x_2} + x_2 + \epsilon$$

$$y = \beta_0 + \boxed{\frac{x}{\beta_1 + x}} + \epsilon$$

46

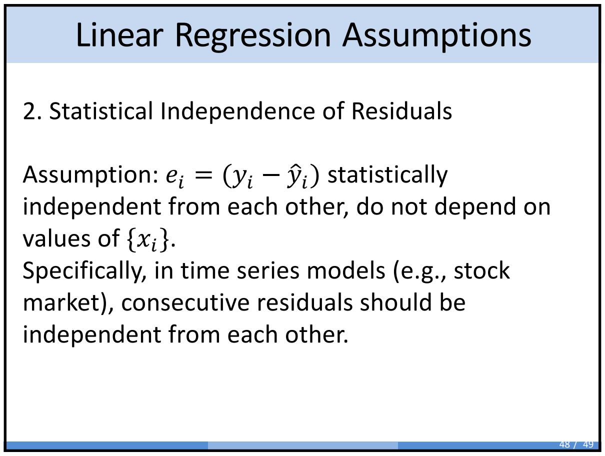


Linear Regression Assumptions

1. Linear & Additive Relation (LAR)

Linear regression can thus result in a model that is not a line

47

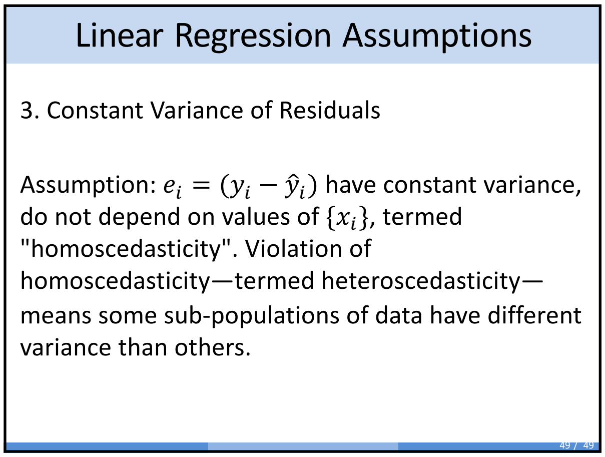


Linear Regression Assumptions

2. Statistical Independence of Residuals

Assumption: $e_i = (y_i - \hat{y}_i)$ statistically independent from each other, do not depend on values of $\{x_i\}$.
Specifically, in time series models (e.g., stock market), consecutive residuals should be independent from each other.

48

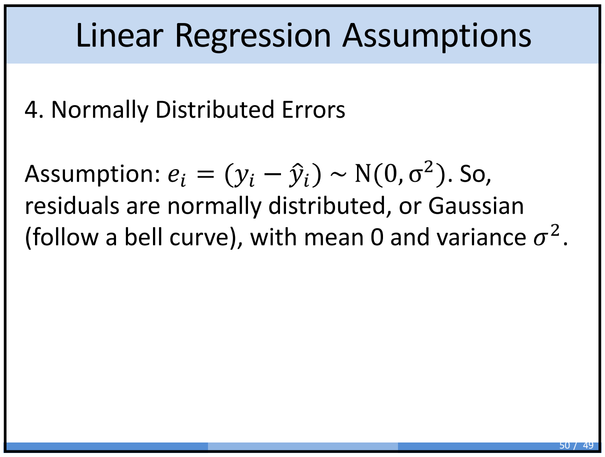


Linear Regression Assumptions

3. Constant Variance of Residuals

Assumption: $e_i = (y_i - \hat{y}_i)$ have constant variance, do not depend on values of $\{x_i\}$, termed "homoscedasticity". Violation of homoscedasticity—termed heteroscedasticity—means some sub-populations of data have different variance than others.

49

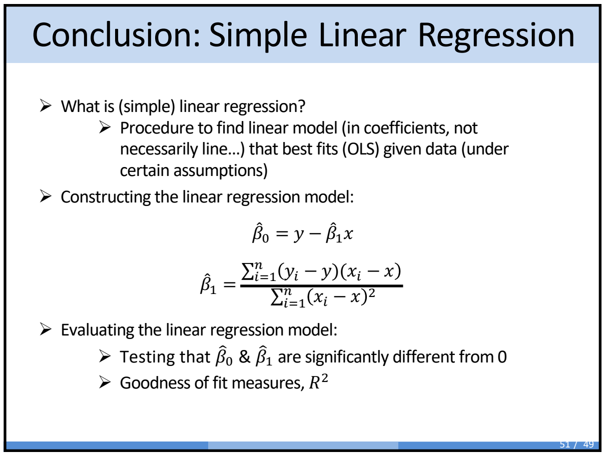


Linear Regression Assumptions

4. Normally Distributed Errors

Assumption: $e_i = (y_i - \hat{y}_i) \sim \mathrm{N}(0, \sigma^2)$. So, residuals are normally distributed, or Gaussian (follow a bell curve), with mean 0 and variance σ^2.

50



Conclusion: Simple Linear Regression

- What is (simple) linear regression?
  - Procedure to find linear model (in coefficients, not necessarily line…) that best fits (OLS) given data (under certain assumptions)
- Constructing the linear regression model:

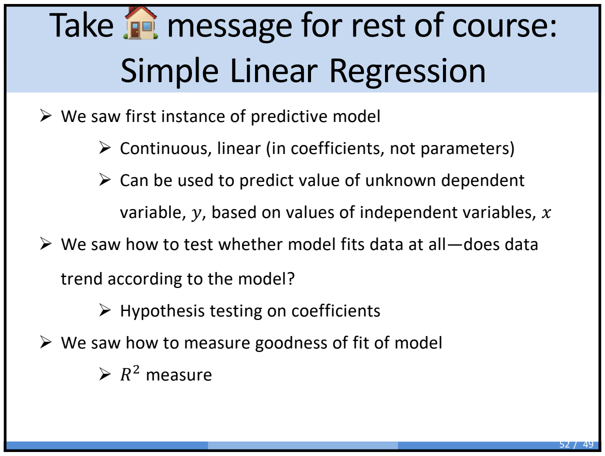
$$\hat{\beta}_0 = y - \hat{\beta}_1 x$$

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

- Evaluating the linear regression model:
  - Testing that $\hat{\beta}_0$ & $\hat{\beta}_1$ are significantly different from 0
  - Goodness of fit measures, R^2

51 / 49

51



Take 🏠 message for rest of course: Simple Linear Regression

- We saw first instance of predictive model
  - Continuous, linear (in coefficients, not parameters)
  - Can be used to predict value of unknown dependent variable, y, based on values of independent variables, x
- We saw how to test whether model fits data at all—does data trend according to the model?
  - Hypothesis testing on coefficients
- We saw how to measure goodness of fit of model
  - R^2 measure

52 / 49

52