Applied Analytics and Predictive Modeling

Spring 2021

Lecture-8

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Some of the slides adapted from Intro to Data Mining Tan et al. 2nd edition

Today’s agenda

- Linear regression
- Case Study-2

Linear Regression

Linear Regression

The technique is used to **predict** the value of one variable (the dependent variable - y) **based on** the value of other variables (independent variables x_1, x_2,…x_k)

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

Modeling

- The first order linear model

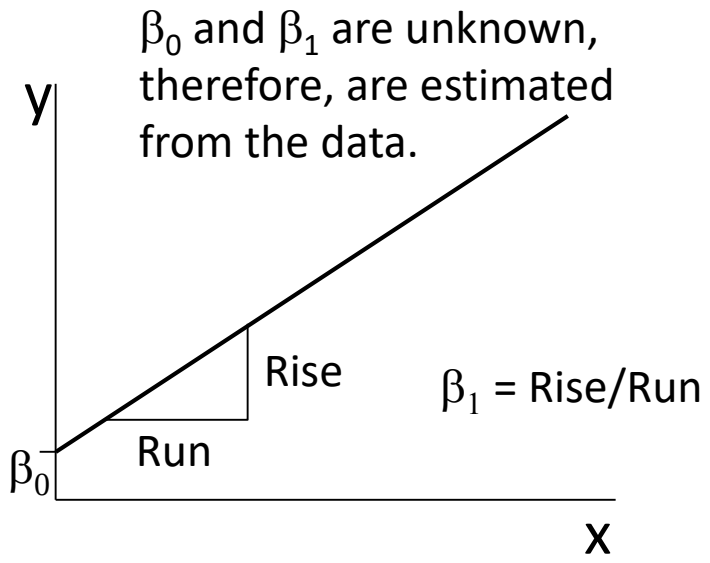
y = dependent variable

x = independent variable

β_0 = y-intercept

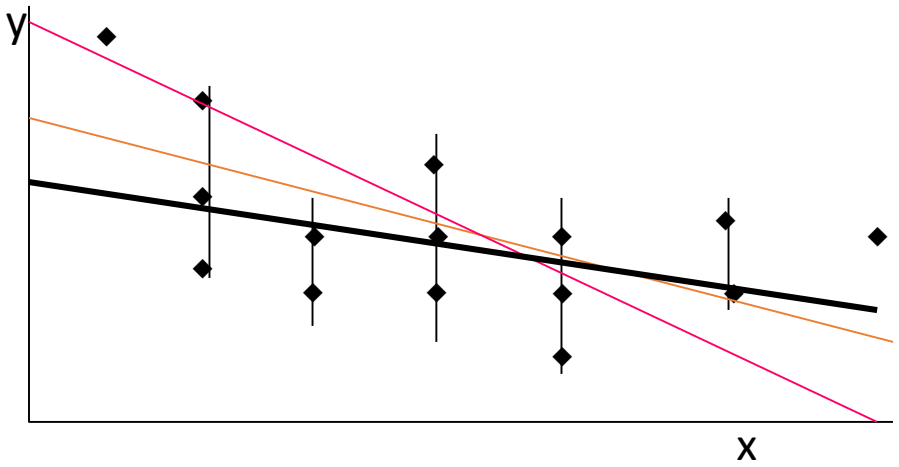
β_1 = slope of the line

$\mathcal{E}$ = error variable



Estimating the coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.

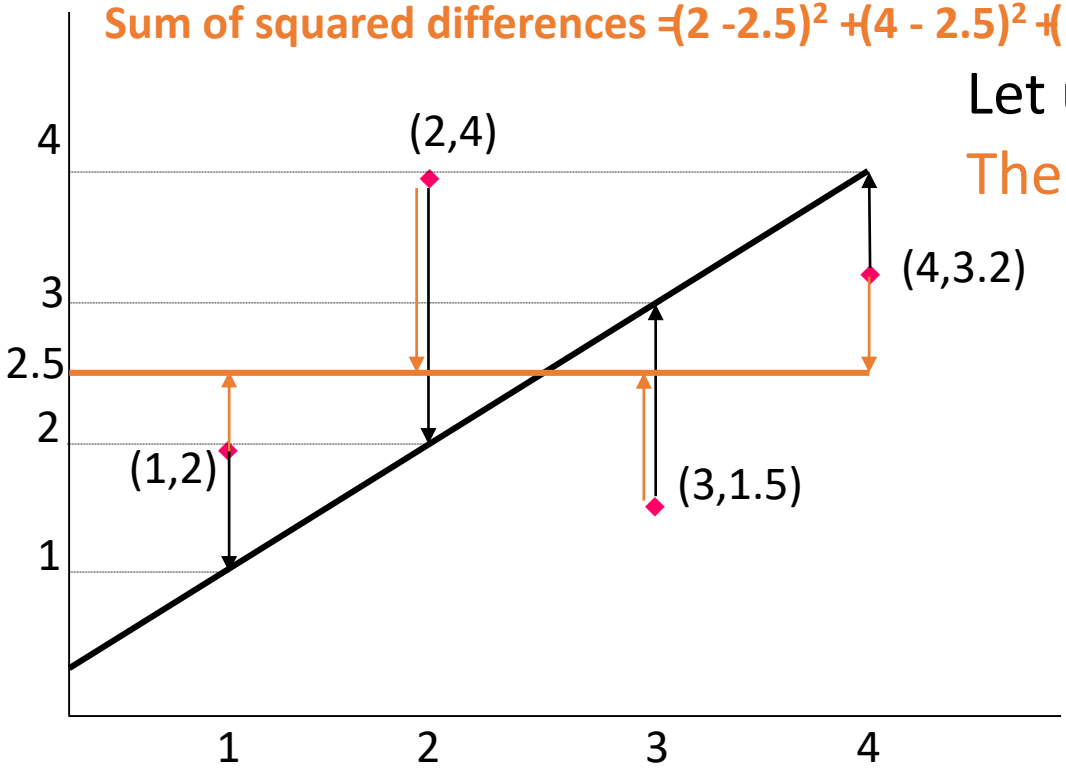


The question is:
Which straight line fits best?

The best line is the one that minimizes the sum of squared vertical differences between the points and the line.

Sum of squared differences $= (2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences $= (2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



Let us compare two lines

The second line is horizontal

The smaller the sum of squared differences the better the fit of the line to the data.

To calculate the estimates of the coefficients that minimize the differences between the data points and the line, use the formulas:

$$b_1 = \frac{\text{cov}(X,Y)}{s_x^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

The regression equation that estimates the equation of the first order linear model is:

$$\hat{y} = b_0 + b_1 x$$

Relationship between odometer reading and a used car's selling price.

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data recorded.
- Find the regression line.

Car	Odometer	Price
1	37388	5318
2	44758	5061
3	45833	5008
4	30862	5795
5	31705	5784
6	34010	5359
.	.	.
.	.	.
.	.	.

Independent variable x

Dependent variable y

Solution to calculate b_0 and b_1 we need to calculate several statistics first;

$$\bar{x} = 36{,}009.45; \qquad s_x^2 = \frac{\sum (x_i - x)^2}{n-1} = 43{,}528{,}688$$

$$\bar{y} = 5{,}411.41; \qquad \text{cov}(X,Y) = \frac{\sum (x_i - x)(y_i - y)}{n-1} = -1{,}356{,}256$$

where n = 100.

$$b_1 = \frac{\text{cov}(X,Y)}{s_x^2} = \frac{-1{,}356{,}256}{43{,}528{,}688} = -.0312$$

$$b_0 = \bar{y} - b_1\bar{x} = 5411.41 - (-.0312)(36{,}009.45) = 6{,}533$$

$$\boxed{\hat{y} = b_0 + b_1 x = 6{,}533 - .0312x}$$

Example-1 - Demo

- Go to the excel sheet to plot a linear regression line for this data

Subject	Age	beats per minute
1	43	91
2	31	72
3	25	65
4	42	87
5	57	110
6	59	120

Exercise

- Build a linear regression model for the diamonds_small.csv uploaded on Piazza

- How do we do this in Python – demo

Case Study-2