On Simple Linear Regression

// FLATIRON SCHOOL

Why Linear Regression ?

- LR is a fundamental tool in the data scientist's kit.
- Practically speaking, using it is one or two lines of code.
- But it's crucial for us to understand the theory underlying it:
  - Building block for more complex tools
  - We are better data scientists if we understand both HOW and WHY

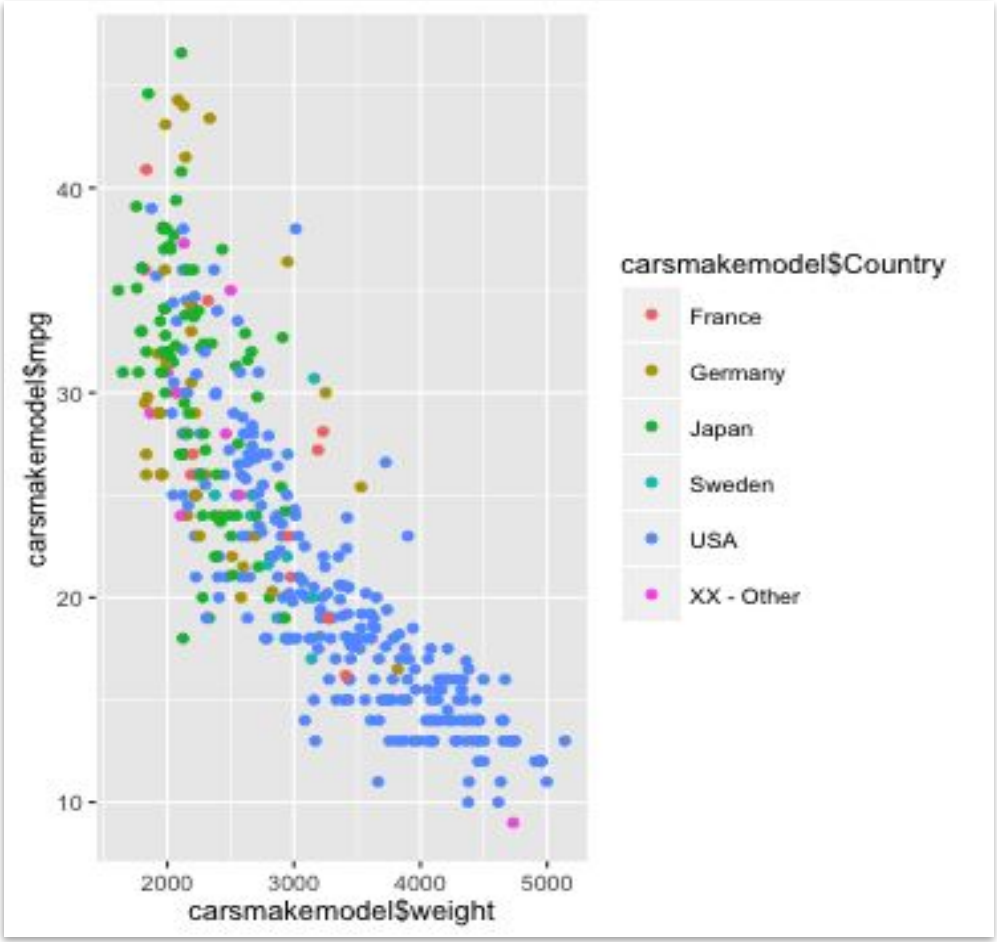
Simple Linear Regression

- Simple: functions of a single variable: $Y = f(X)$
- Linear: models are lines
- Regression: dependent variable is continuously-valued

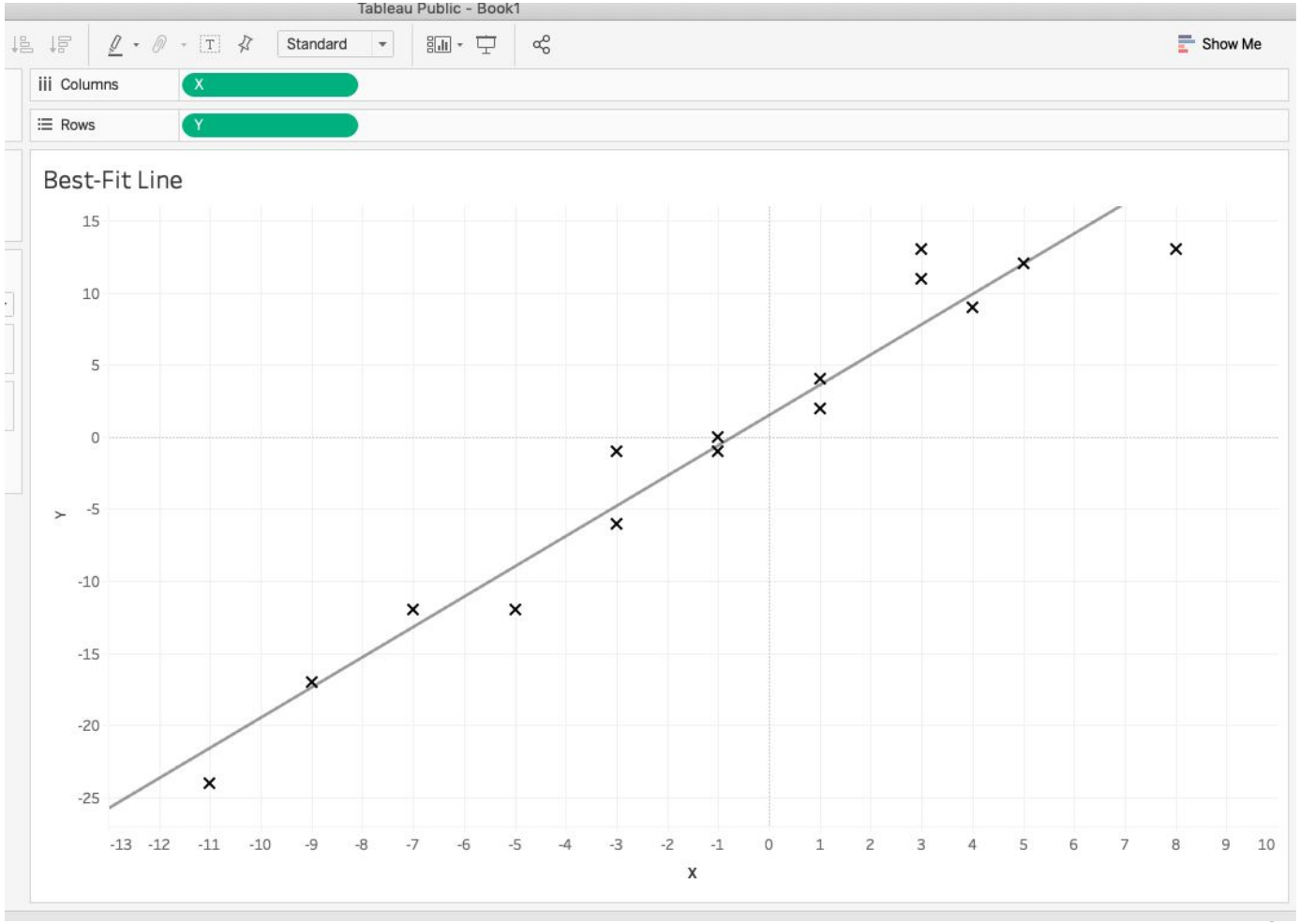
Inference and Prediction

- As population density increases, so do housing prices.
- As the number of trees decreases, the concentration of CO_2 goes up.

Car Weight and MPG



Best-Fit Line



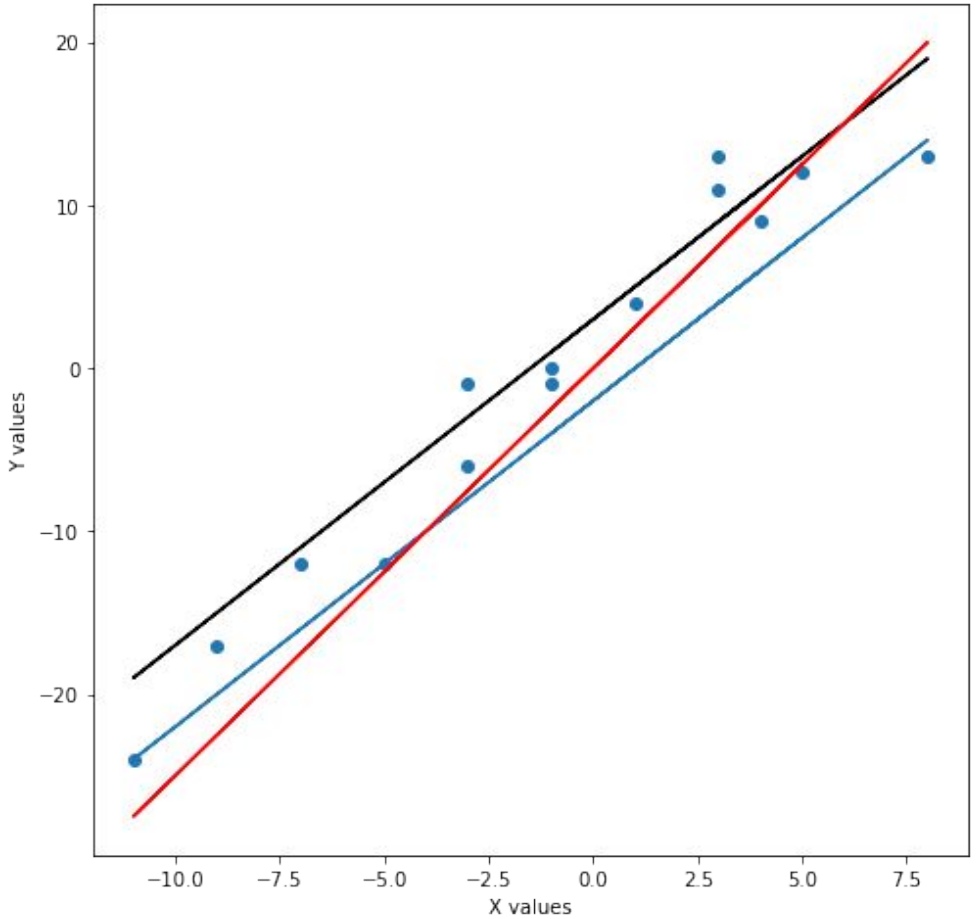
A Line as a Model

- Predictions for *all* values of the X variable
  - Model shape: $\hat{y} = \beta_1 x + \beta_0$
- Error as the distance between real and predicted values $E = y - \hat{y}$

$$E^2 = (y - \hat{y})^2$$

Goal: Minimize Error

- Which of these lines fits the data best?



How to Construct the Best-Fit Line

$$r_P = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

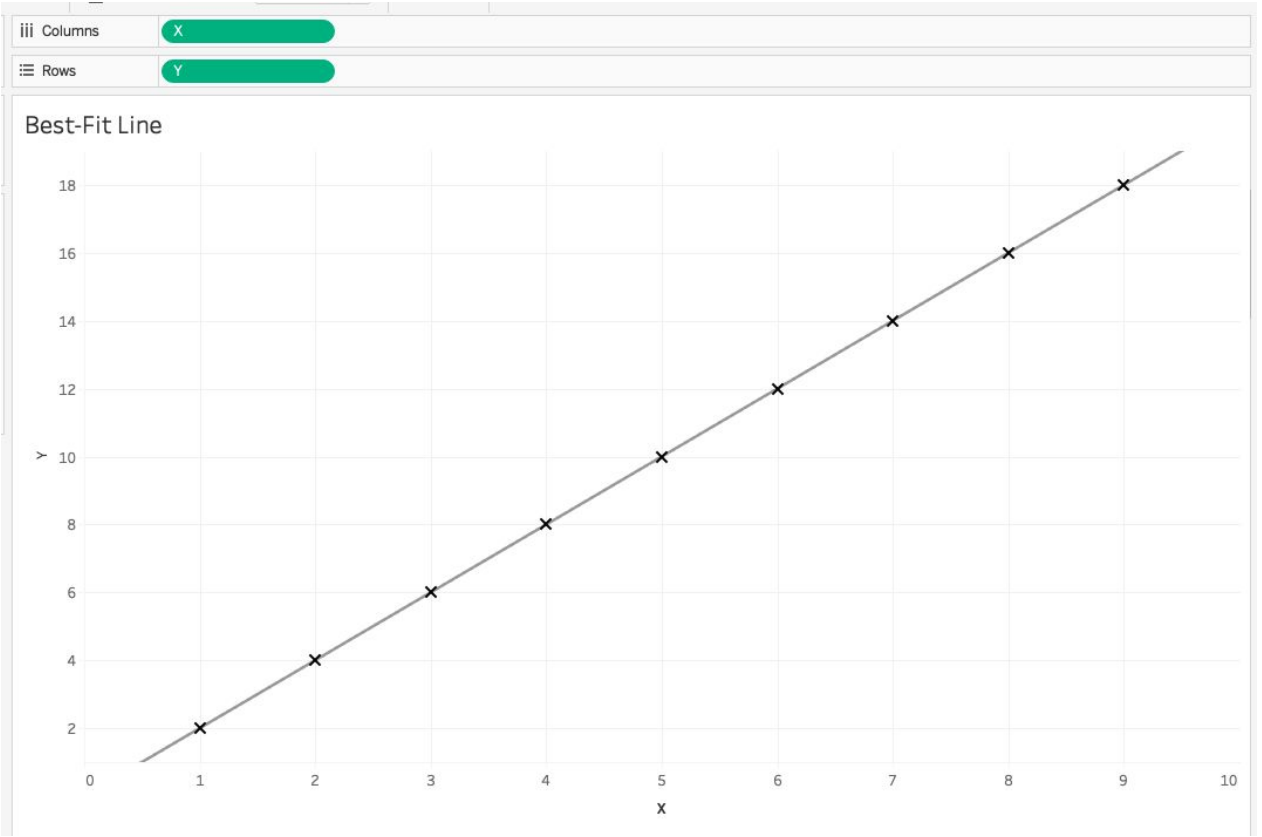
$$\beta_1 = r_P \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

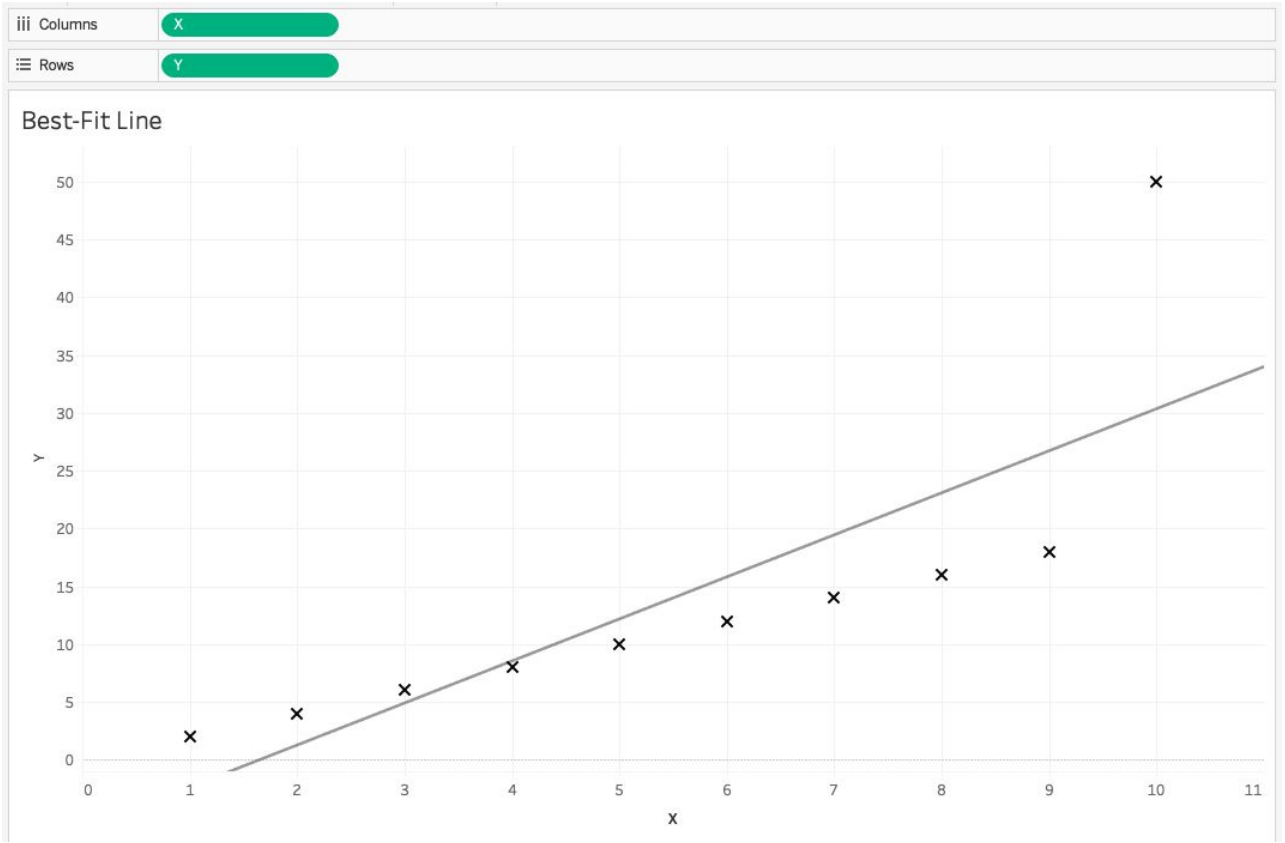
Outliers

# Sheet1 X	# Sheet1 Y
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16
9	18
10	50

Dropping Outliers



Keeping Outliers



Example

- Construct the best-fit line for the points:

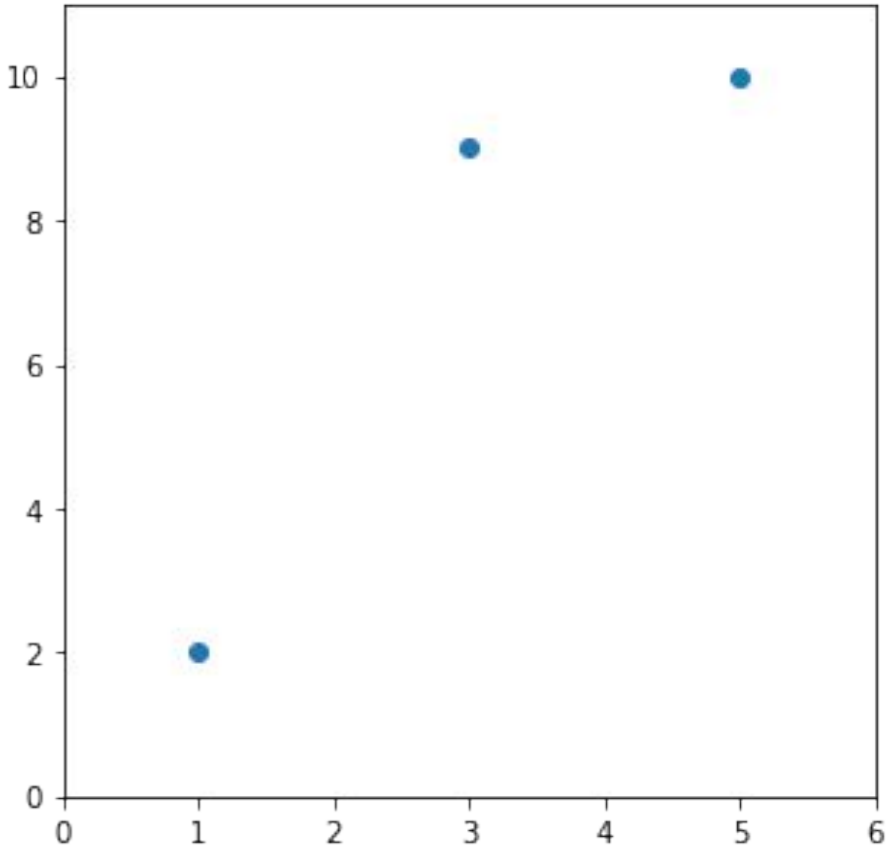
  (1, 2), (3, 9), and (5, 10).

- Remember:

$$r_P = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2}\sqrt{\Sigma_i (y_i - \bar{y})^2}}$$

$$\beta_1 = r_P \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$



$x_i : [1, 3, 5]$

$y_i : [2, 9, 10]$

Step 1

- Calculate x_bar and y_bar:

$$\bar{x} = \frac{1+3+5}{3} = 3$$

$$\bar{y} = \frac{2+9+10}{3} = 7$$

$x_i : [1, 3, 5]$

$y_i : [2, 9, 10]$

Step 2

- Calculate these products:

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = (1-3)(2-7) + (3-3)(9-7) + (5-3)(10-7) = 16$$

$$\Sigma(x_i - \bar{x})^2 = (1-3)^2 + (3-3)^2 + (5-3)^2 = 8$$

$$\Sigma(y_i - \bar{y})^2 = (2-7)^2 + (9-7)^2 + (10-7)^2 = 38$$

$x_i : [1, 3, 5]$

$y_i : [2, 9, 10]$

Step 3

- Calculate Pearson correlation:

$$r_P = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2}\sqrt{\Sigma_i (y_i - \bar{y})^2}} = \frac{16}{\sqrt{(8)(38)}} = \frac{4}{\sqrt{19}}$$

$x_i : [1, 3, 5]$

$y_i : [2, 9, 10]$

Step 4

- Calculate standard deviations:

$$\sigma_x = \sqrt{\frac{8}{3}}$$

$$\sigma_y = \sqrt{\frac{38}{3}}$$

$x_i : [1, 3, 5]$

$y_i : [2, 9, 10]$

Step 5

- Calculate the slope:

$$\beta_1 = r_P \frac{\sigma_y}{\sigma_x} = \frac{4}{\sqrt{19}} \left(\frac{\sqrt{\frac{38}{3}}}{\sqrt{\frac{8}{3}}} \right) = \frac{4}{\sqrt{19}} \left(\frac{\sqrt{38}}{\sqrt{8}} \right) = \frac{4\sqrt{2}}{2\sqrt{2}} = 2$$

$x_i : [1, 3, 5]$
$y_i : [2, 9, 10]$

Step 6

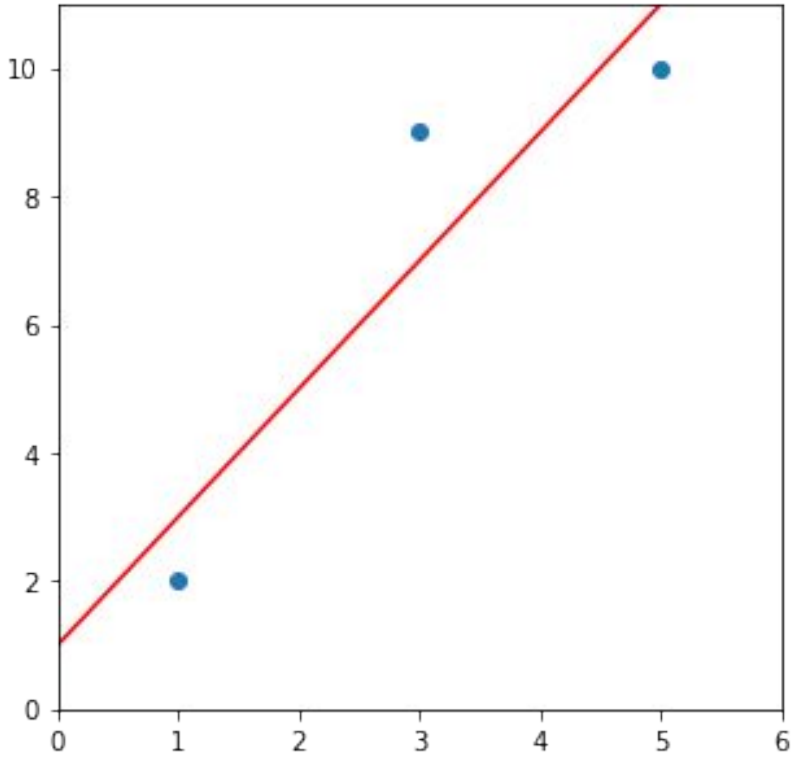
- Calculate the y-intercept:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 7 - (2)(3) = 1$$

Line

- We have our line!

$$\hat{y} = \beta_1 x + \beta_0 = 2x + 1$$



Surface and Contour Plots of SSE(m, b)