Pelatihan ABCD
Modul 4-2: Linear Regression

Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung

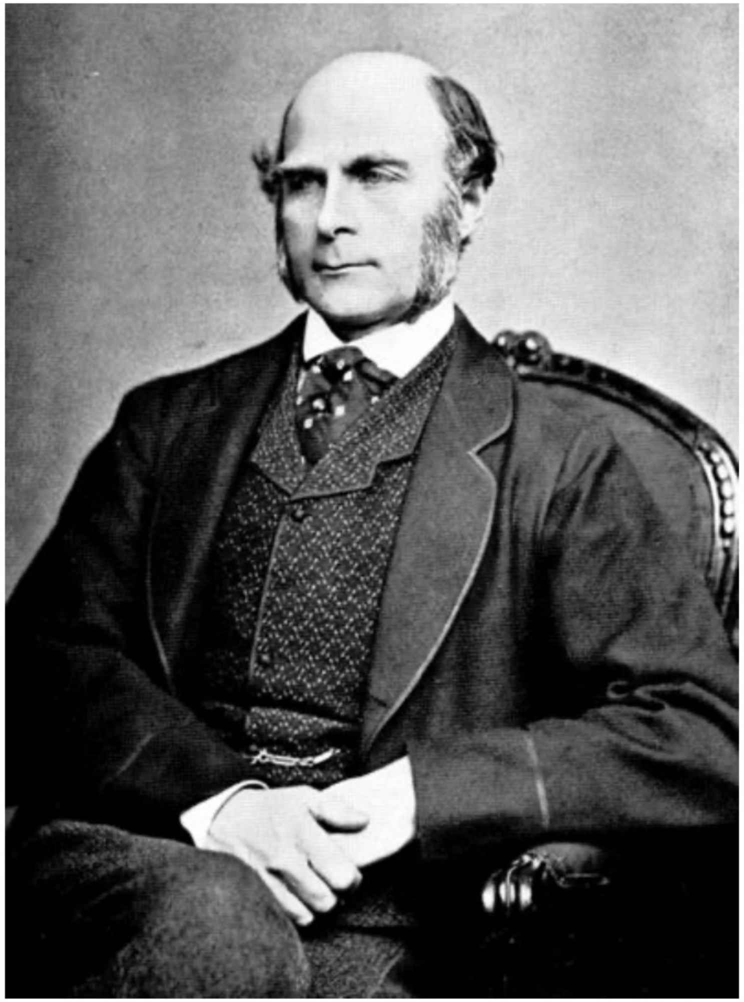
Unviersitas Singaperbangsa Karawang

Contents

- Concept of Linear Regression
- Evaluating Regression

History

- This all started in the 1800s with a guy named Francis Galton. Galton was studying the relationship between parents and their children. In particular, he investigated the relationship between the heights of fathers and their sons.
- What he discovered was that a man's son tended to be roughly as tall as his father.
- However, Galton's breakthrough was that the son's height tended to be **closer to the overall average height of all people**

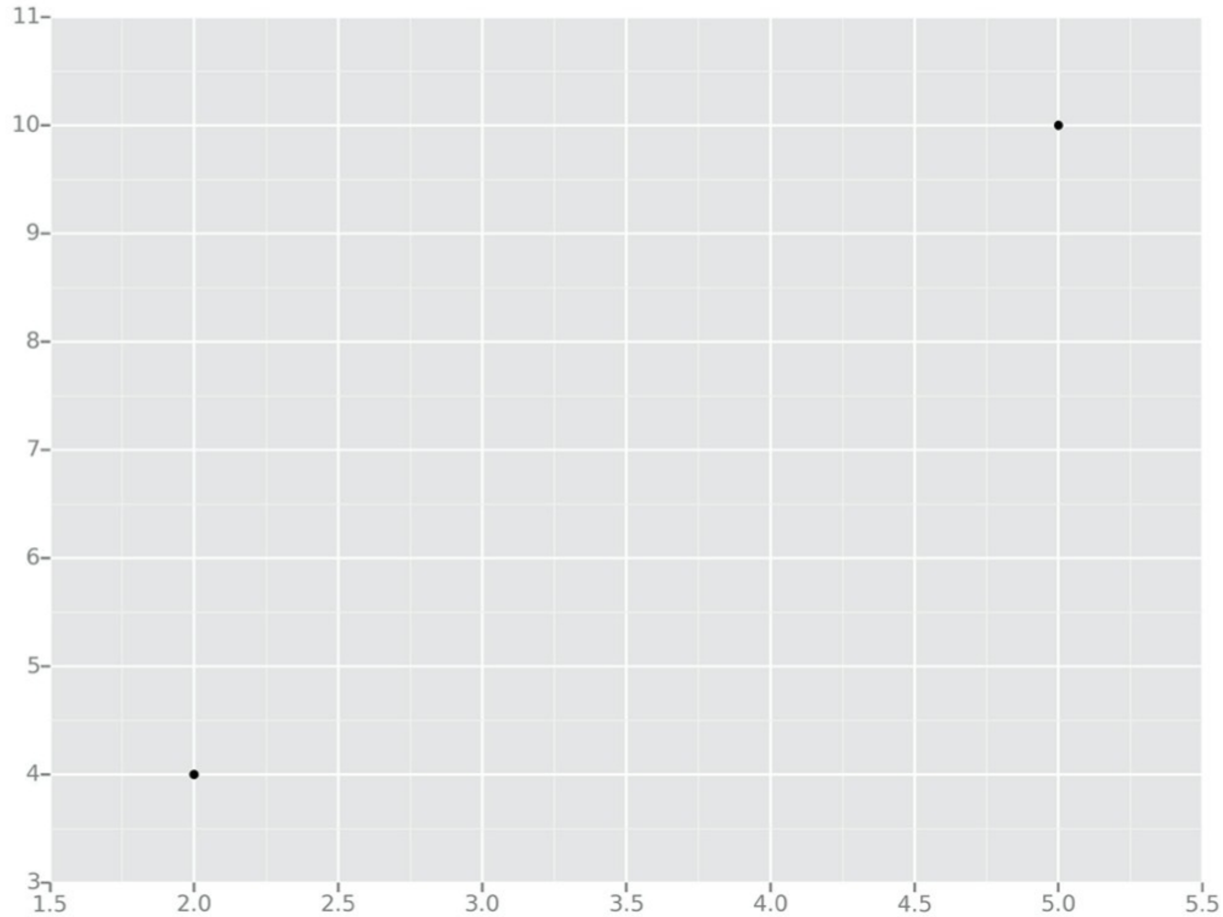


Example

- Let's take Shaquille O'Neal as an example. Shaq is really tall: 2.2 meters.
- If Shaq has a son, chances are he'll be pretty tall to. However, Shaq is such an anomaly that there is also a very good chance that his son will be **not be as tall as Shaq**.
- Turns out this is the case: Shaq's son is pretty tall (2 meters), but not nearly as tall as his dad.
- Galton called this phenomenon regression, as in "A father's son's height tends to regress (or drift towards) the mean (average) height."

Example

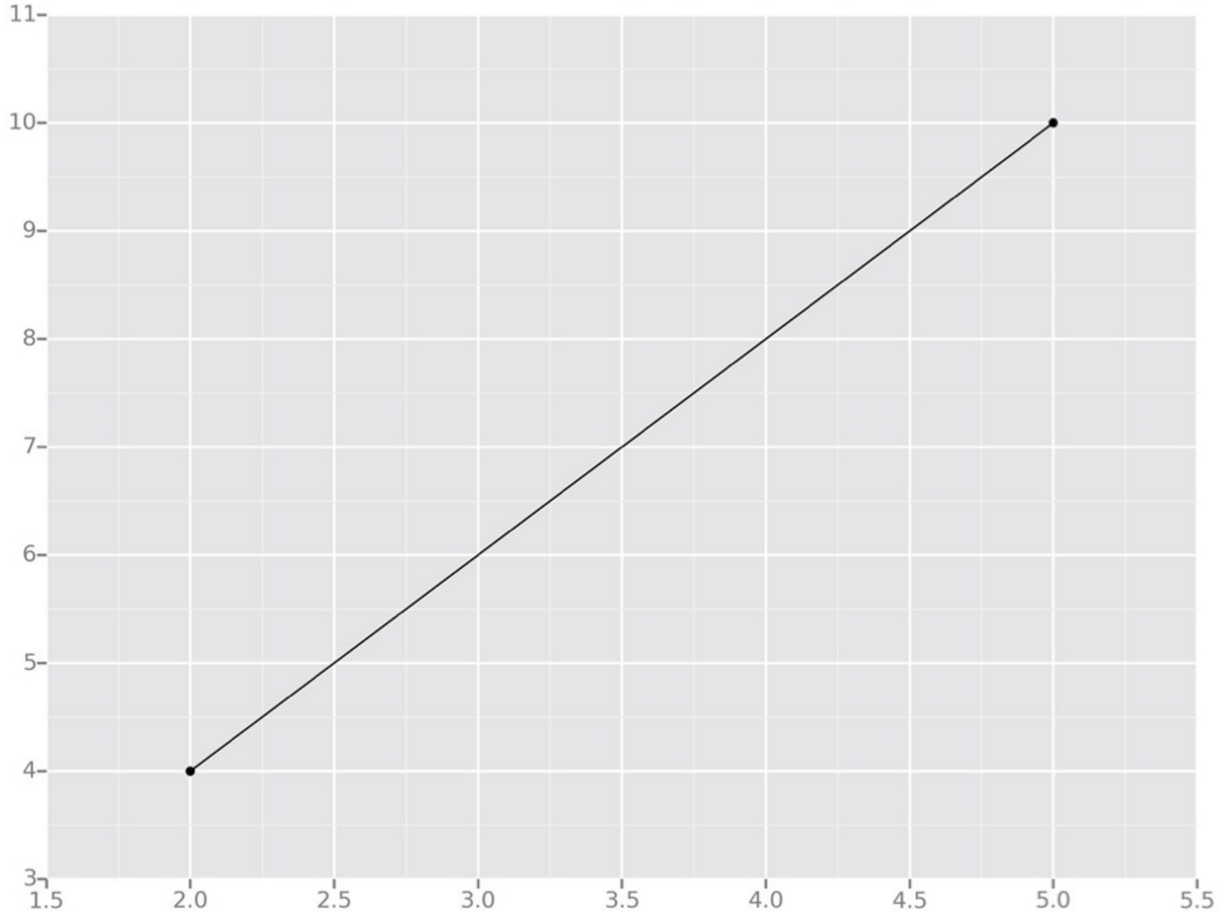
- Let's take the simplest example: calculating a regression with only 2 data points.



A = (2,4) and B = (5,10)

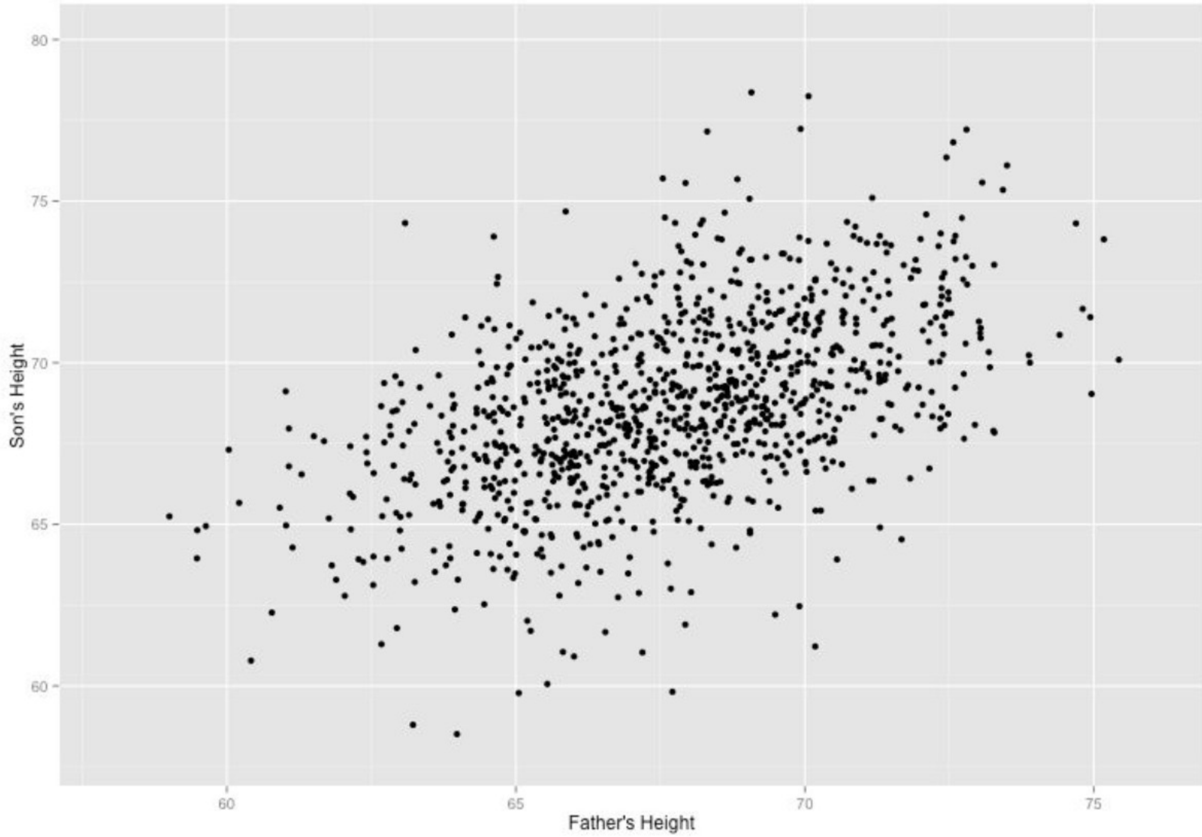
Example

- All we’re trying to do when we calculate our regression line is draw a line that’s as close to every dot as possible.
- For classic linear regression, or “Least Square Method”, you only measure the closeness in the “up and down” direction.



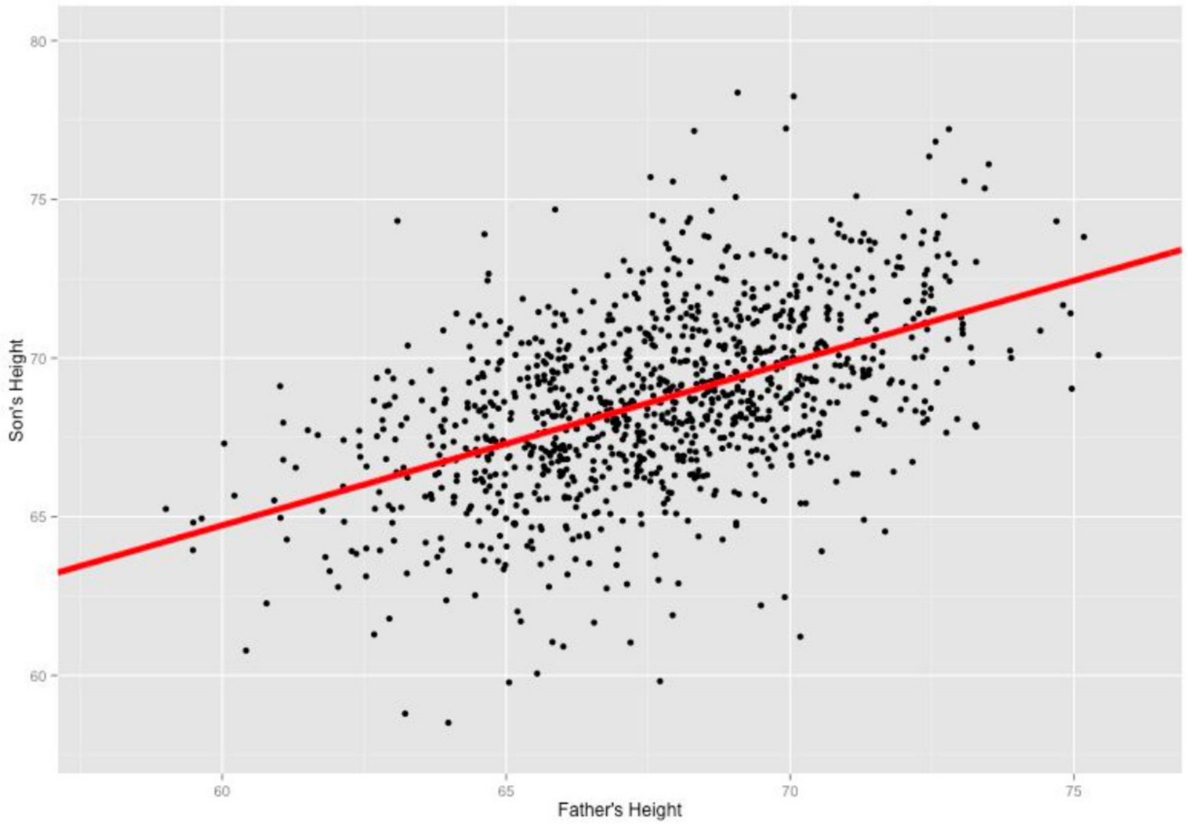
Example

- Now wouldn't it be great if we could apply this same concept to a graph with more than just two data points?
- By doing this, we could take multiple men and their son's heights and do things like tell a man how tall we expect his son to be. Before he even has a son!



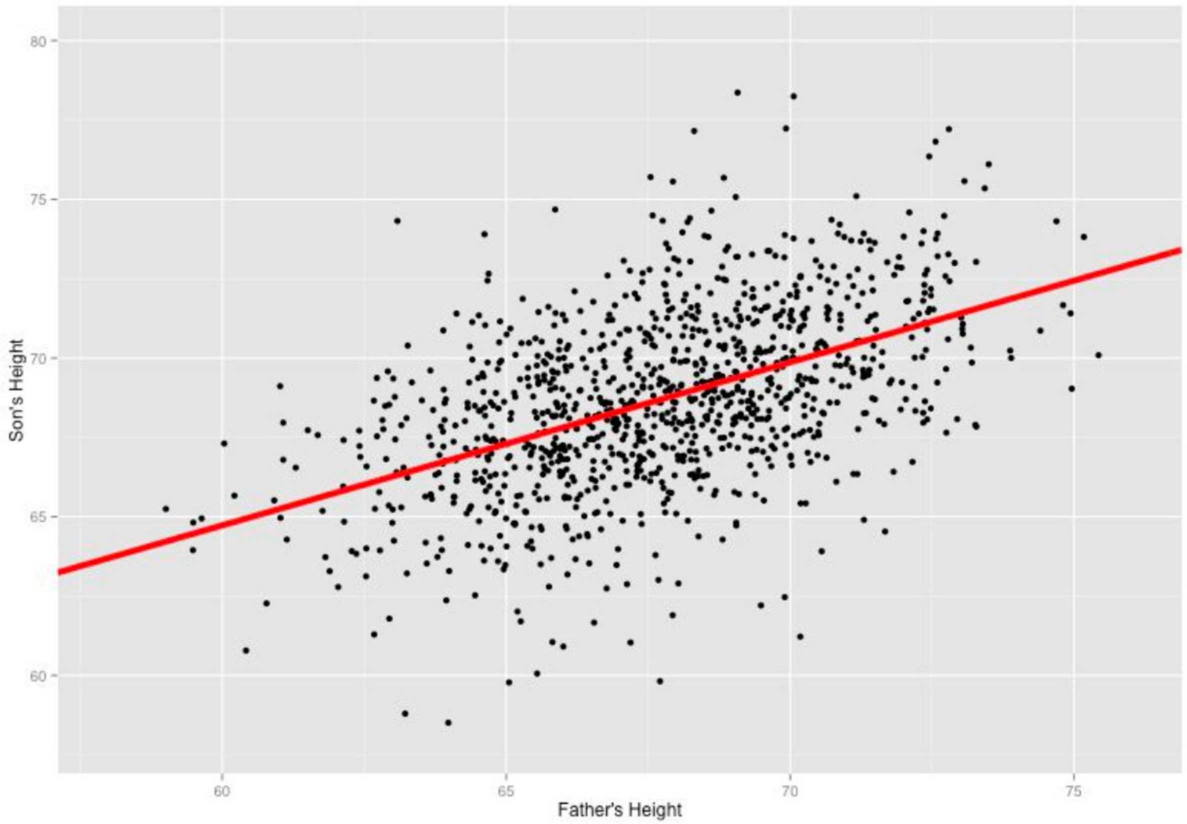
Example

- Our goal with linear regression is to **minimize the vertical distance** between all the data points and our line.
- So, in determining the **best line**, we are attempting to minimize the distance **between all the points** and their distance to our line.



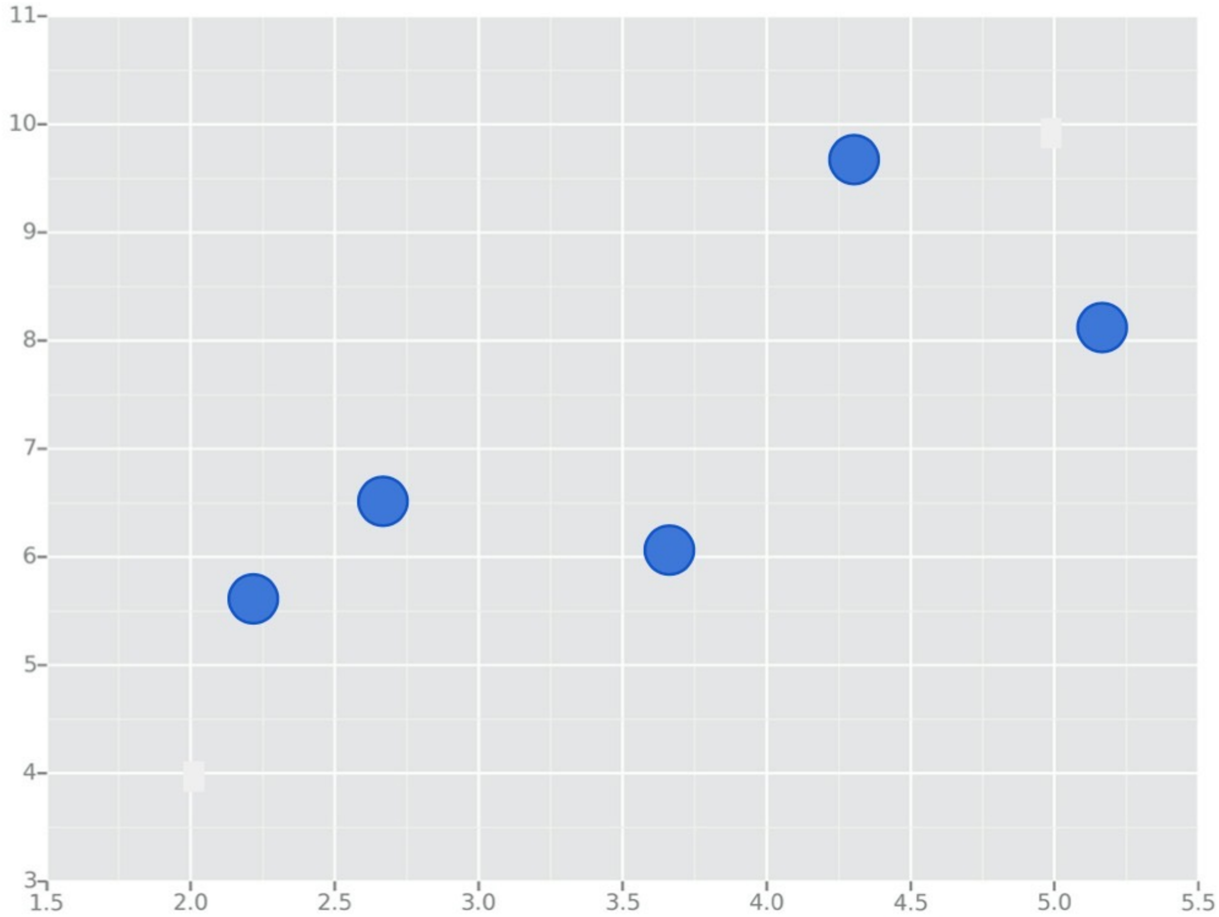
Example

- There are lots of different ways to minimize this (sum of squared errors, sum of absolute errors, etc), but all these methods have a general goal of minimizing this distance.



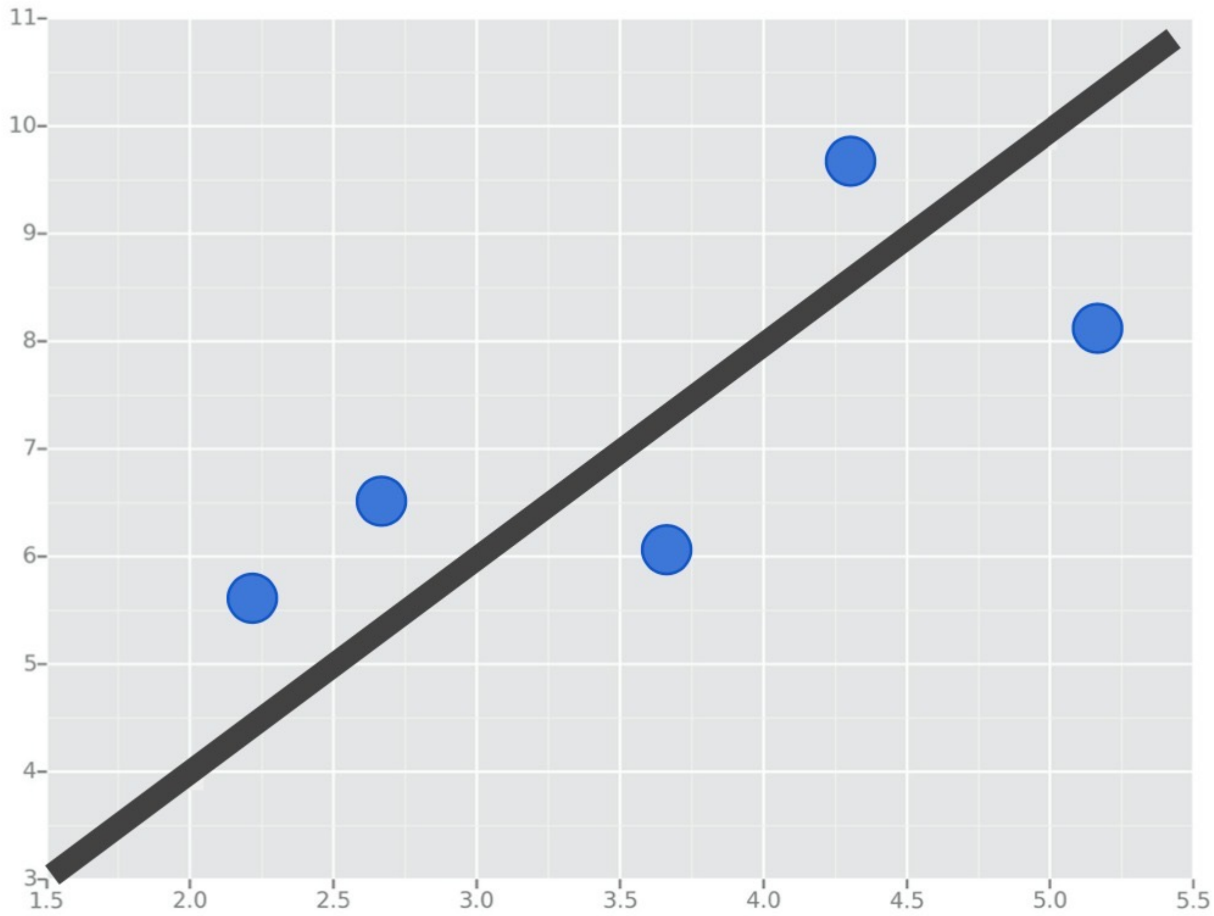
Example

- For example, one of the most popular methods is the least squares method.
- Here we have blue data points along an x and y axis



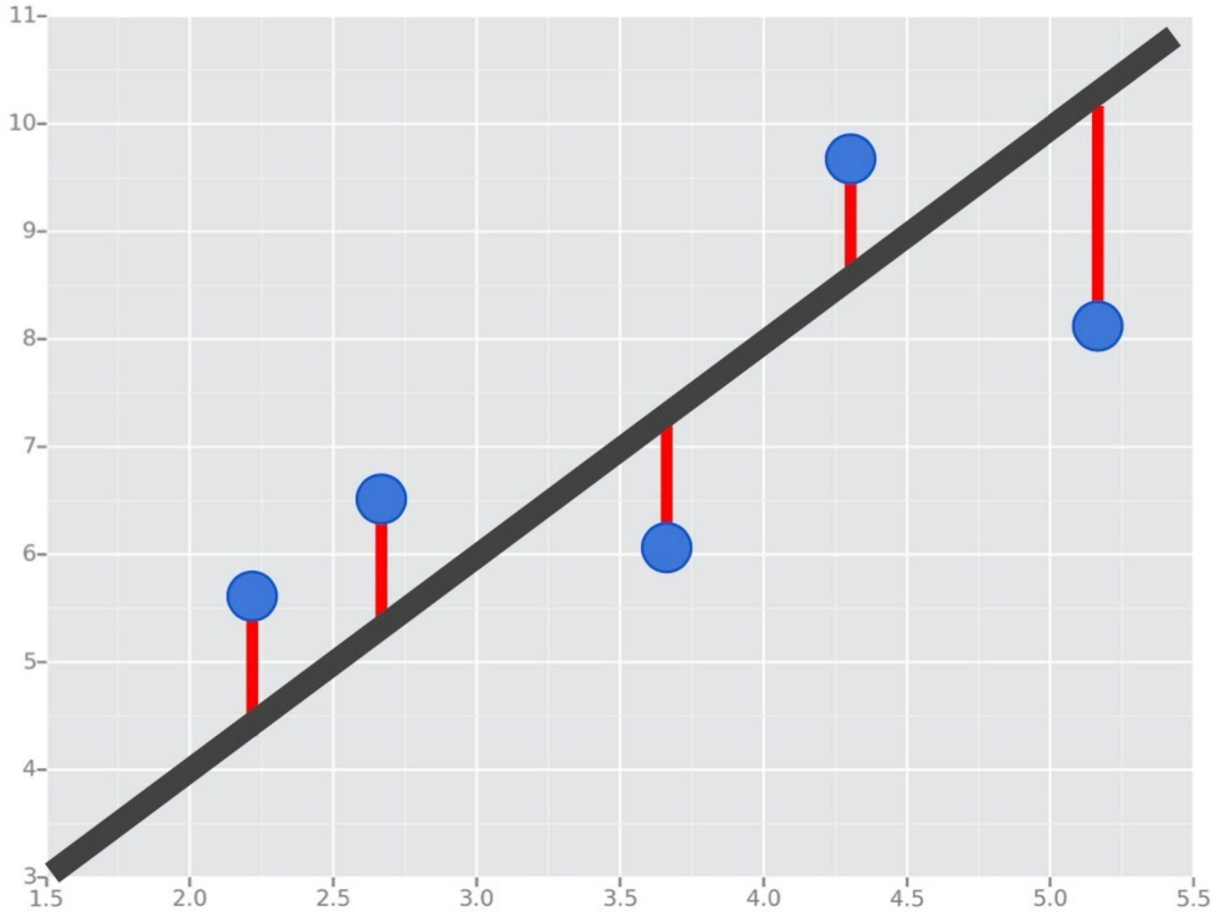
Example

- Now we want to fit a linear regression line.
- The question is, how do we decide which line is the best fitting one?



Example

- We'll use the **Least Squares Method**, which is fitted by **minimizing the sum of squares of the residuals**.
- The residuals for an observation is the difference between the observation (the y-value) and the fitted line.



Evaluating Regression

- Regression is a task when a model attempts to **predict continuous values** (unlike categorical values, which is classification)
  - For example, attempting to predict the price of a house given its features is a **regression task.**
  - Attempting to predict the country a house is in given its features would be a **classification task**.
- The most common evaluation metrics for regression:
  1. **Mean Absolute Error**
  2. **Mean Squared Error**
  3. **Root Mean Square Error**

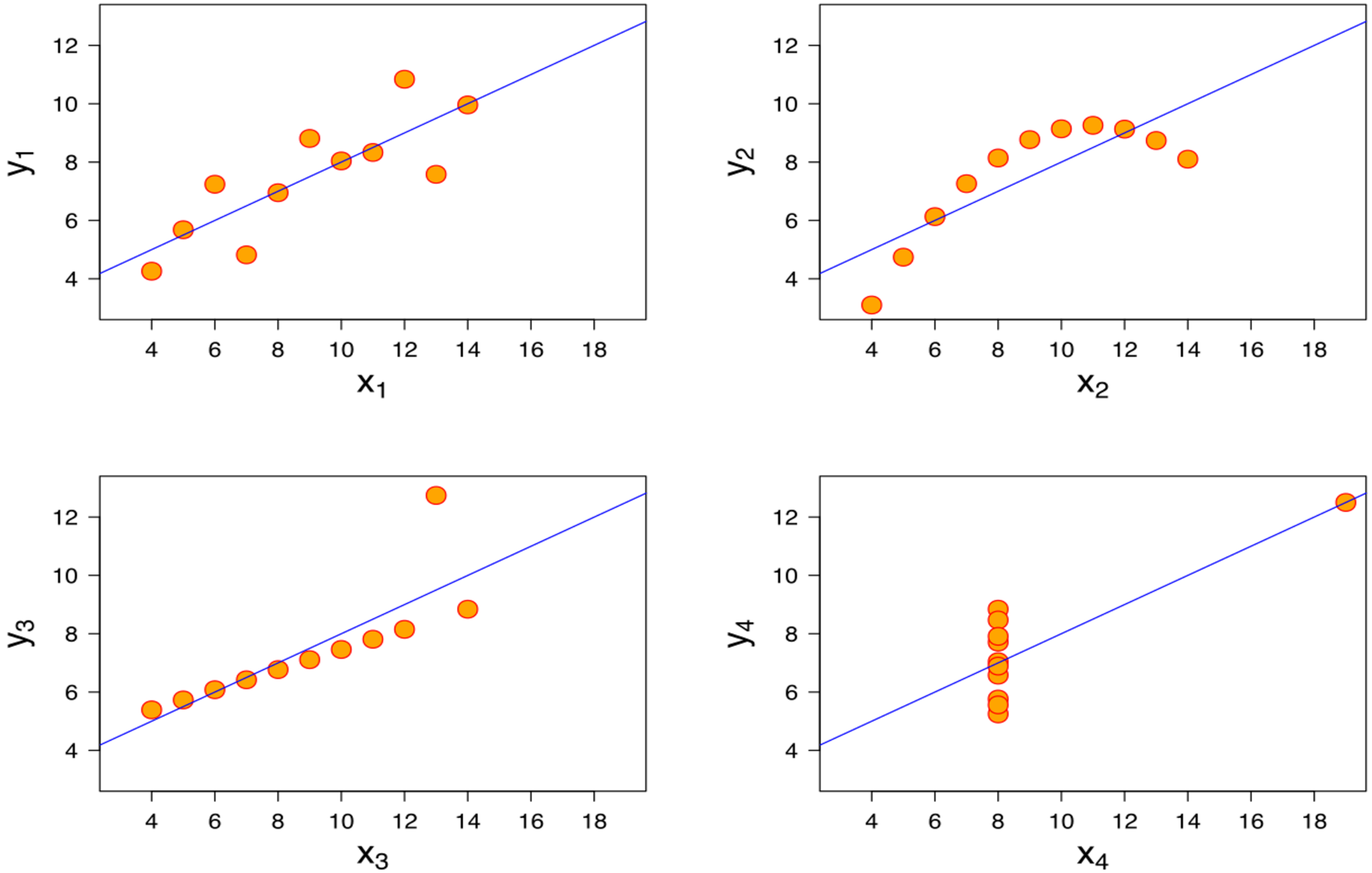
1. Mean Absolute Error (MAE)

- This is the mean of the absolute value of errors

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

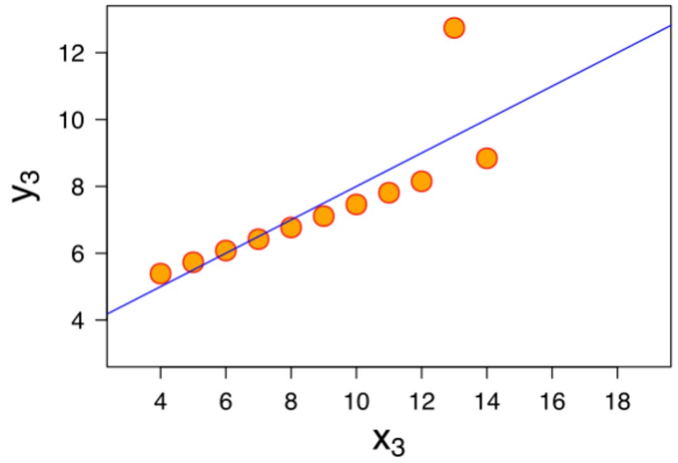
1. Mean Absolute Error (MAE)

- However, MAE won't punish large errors

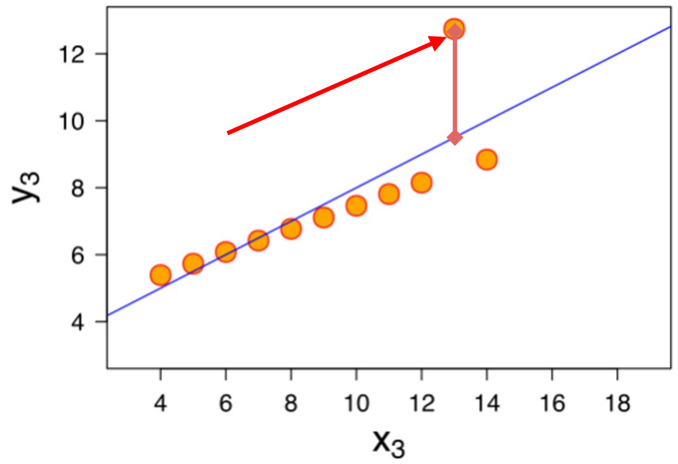


1. Mean Absolute Error (MAE)

- However, MAE won't punish large errors



- We want our error metrics to account for these



2. Mean Squared Error (MSE)

- Mean Squared Error (MSE) is the mean of the squared errors.
- Larger errors are noted more than with MAE, making MSE more popular.

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

Root Mean Square Error (RMSE)

- Root Mean Square Error (RMSE) is the root of the mean of the squared errors.
- Most popular (has same units as y)

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

Linear Regression Lab Project

- Open your Jupyter Notebook
- Use Scikit-Learn and Python to create a linear regression model
- Solve the project exercise