Linear Regression

Machine Learning Context

- Given a set of data with associated measurements
- Predict the results of future measurements given a set of known results

Data could be a collection of images, measurements say "this is a duck".

Data could be numerical (such as time intervals) and measurements could be numerical (such as speed of an object or a stock price).

Simplest case is finding a linear relationship.

Basic Problem

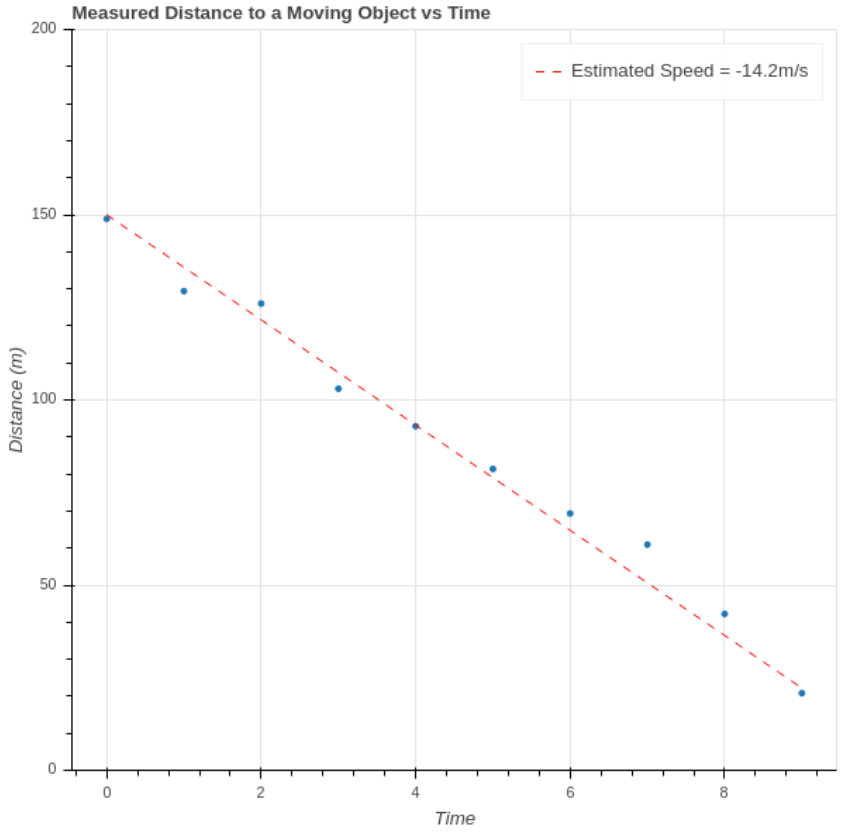

Figure 1: Physics Experiment

Engine size and MPG

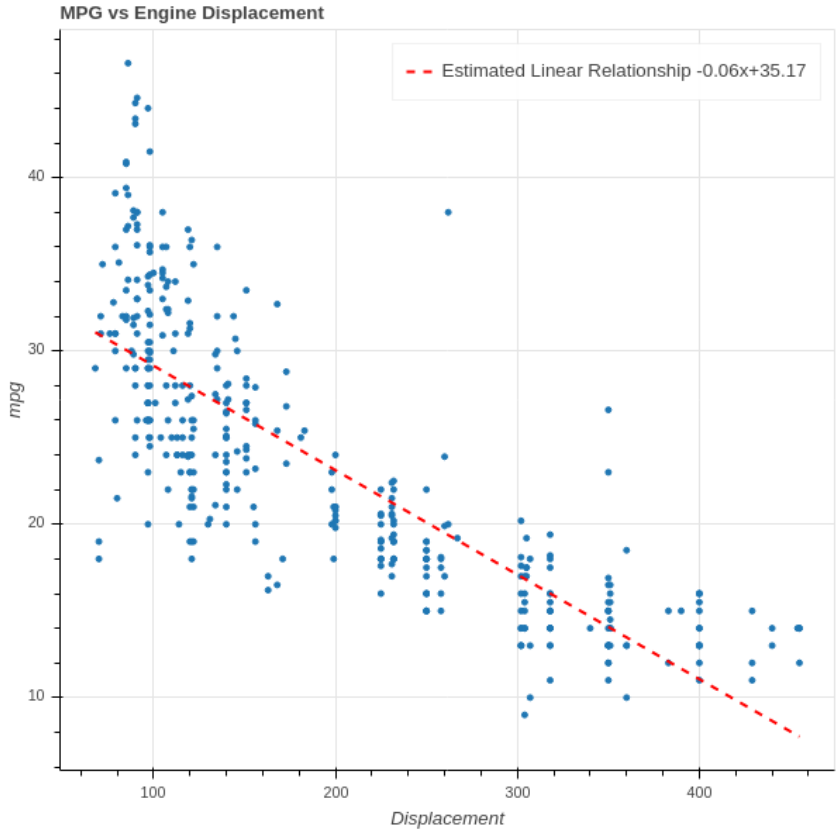

Figure 2: MPG vs Displacement

Mean Squared Error

Data consists of pairs $\{(x_i, y_i)\}$.

$$MSE(m, b) = \frac{1}{N} \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

Minimize MSE

Write E instead of MSE for simplicity.

$$\frac{\partial E}{\partial m} = \frac{1}{N}\sum_{1}^{N} -2x_i(y_i - mx_i - b)$$

$$\frac{\partial E}{\partial b} = \frac{1}{N}\sum_{1}^{N} -2(y_i - mx_i - b)$$

Compute the derivatives

$$\begin{aligned} \frac{1}{N}(\sum_{i=1}^{N} x_i^2)m + & \quad \frac{1}{N}(\sum_{i=1}^{N} x_i)b = & \quad \frac{1}{N}\sum_{i=1}^{N} x_i y_i \\ \frac{1}{N}(\sum_{i=1}^{N} x_i)m + & & b = & \quad \frac{1}{N}\sum_{i=1}^{N} y_i \end{aligned}$$

- $\overline{x} = \frac{1}{N}\sum x_i$
- $\overline{y} = \frac{1}{N}\sum y_i$
- S_{xx}, S_{xy}, and S_{yy} are $\frac{1}{N}\sum x_i^2$, $\frac{1}{N}\sum x_i y_i$, $\frac{1}{N}\sum y_i^2$respectively.

Solve to find the minima

$$\begin{aligned} S_{xx}m+ \quad & \overline{x}b = S_{xy} \\ \overline{x}m+ \quad & b = \overline{y} \end{aligned}$$

Solution

$$m = \frac{S_{xy} - \overline{xy}}{S_{xx} - \overline{x}^2}$$

$$b = \frac{S_{xx}\overline{y} - S_{xy}\overline{x}}{S_{xx} - \overline{x}^2}$$