Regression Introduction and Estimation Review

Dr. Yang Feng

Formal Statement of Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- ϵ_i is a random error term with mean $\mathbb{E}(\epsilon_i)$ and variance $\text{Var}(\epsilon_i) = \sigma^2$
- $i = 1, \ldots, n$

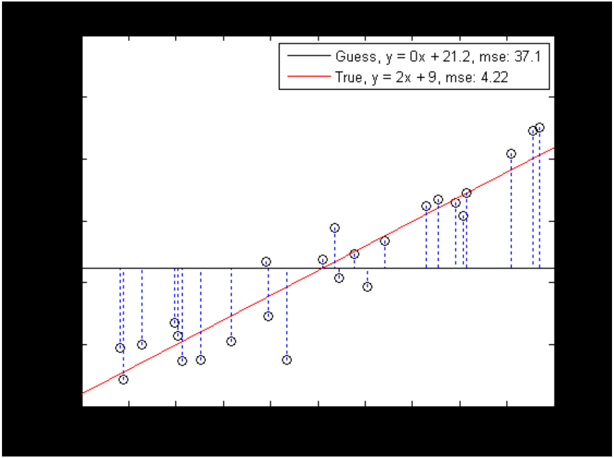
Least Squares Linear Regression

- Seek to minimize

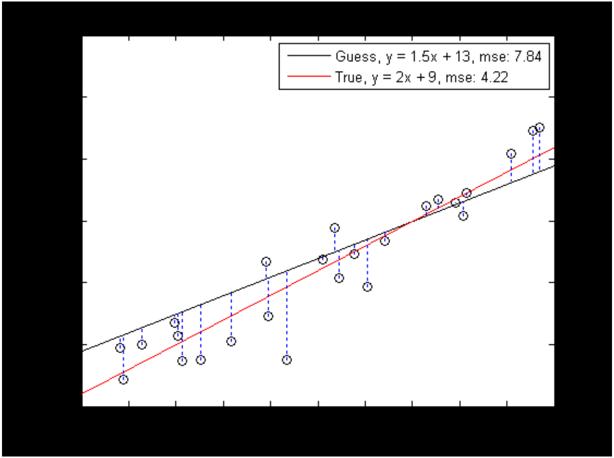
$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- By careful choice of b_0 and b_1 where b_0 is a point estimator for β_0 and b_1 is the same for β_1
  How?

Guess #1



Guess #2



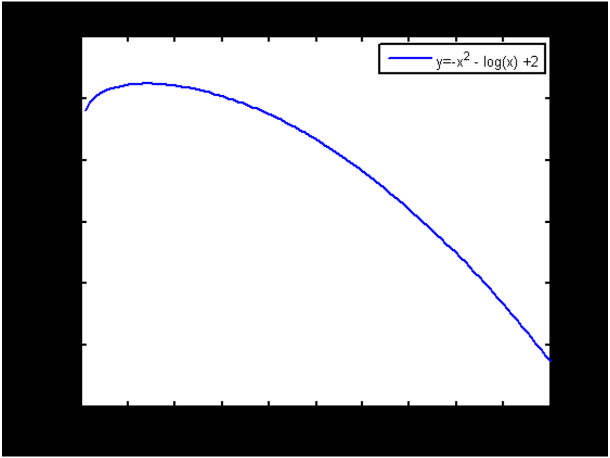
Function maximization

- Important technique to remember!
  - Take derivative
  - Set result equal to zero and solve
  - Test second derivative at that point
- Question: does this always give you the maximum?
- Going further: multiple variables, convex optimization

Function Maximization

Find the maximum value of x that satisfies the function

$$-x^2 + ln(x) = a, x > 0$$



Least Squares Max(min)imization

- Function to minimize w.r.t. b_0 and b_1 – b_0 and b_1 are called point estimators of β_0 and β_1 respectively

$$Q = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

- Minimize this by maximizing -Q
- Either way, find partials and set both equal to zero

$$\begin{aligned} \frac{dQ}{db_0} &= 0 \\ \frac{dQ}{db_1} &= 0 \end{aligned}$$

Normal Equations

- The result of this maximization step are called the normal equations.

$$\begin{aligned} \sum Y_i &= nb_0 + b_1 \sum X_i \\ \sum X_i Y_i &= b_0 \sum X_i + b_1 \sum X_i^2 \end{aligned}$$

- This is a system of two equations and two unknowns. The solution is given by...

Solution to Normal Equations

After a lot of algebra one arrives at

$$\begin{aligned} b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ b_0 &= \bar{Y} - b_1 \bar{X} \\ \bar{X} &= \frac{\sum X_i}{n} \\ \bar{Y} &= \frac{\sum Y_i}{n} \end{aligned}$$