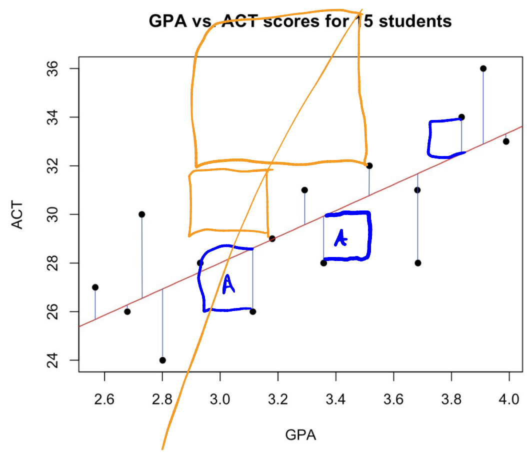
MATH 134A Review: Least-squares regression

The least-squares regression line is the line that makes the sum of the squared residuals as small as possible.



Let $(x_1, y_1), \ldots, (x_n, y_n)$ be points on a scatterplot. Find the equation of a line $\hat{y} = a + bx$ so that $(y_1 - \hat{y}(x_1))^2 + \cdots + (y_n - \hat{y}(x_n))^2$ is minimized.

Derivation

Define

$$Q(a, b) = (y_1 - \hat{y}(x_1))^2 + \cdots + (y_n - \hat{y}(x_n))^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

We will find a and b so that $\frac{\partial}{\partial a} Q = 0$ and $\frac{\partial}{\partial b} Q = 0$. Observe

$$\frac{\partial}{\partial a} Q = 2(na + b\sum_{i=1}^n x_i - \sum_{i=1}^n y_i) = 0.$$

A·B = 0

A = 0 or B = 0

Define $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$ and $\bar{y} = \frac{y_1 + \cdots + y_n}{n}$. Then

$$a = \bar{y} - b\bar{x}.$$

Y-intercept

It remains to solve for b. Observe

$$\frac{\partial}{\partial b} Q = -2\sum_{i=1}^n (x_i y_i - a x_i - b x_i^2) = -2\sum_{i=1}^n (x_i y_i - x_i \bar{y} + b x_i \bar{x} - b x_i^2) = 0.$$

A · B = 0

Then

$$b = \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y})}{\sum_{i=1}^n (x_i^2 - x_i \bar{x})}.$$

Slope