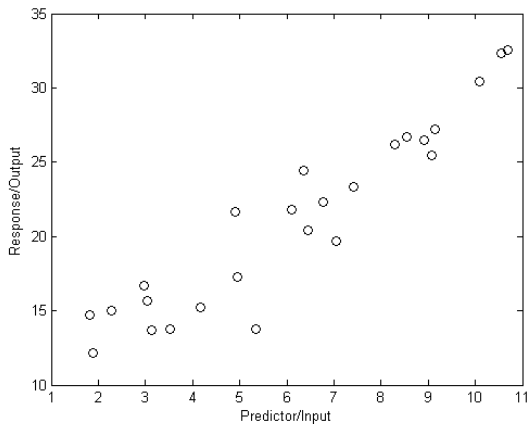


# Regression Introduction and Estimation Review

Dr. Frank Wood

## Quick Example - Scatter Plot



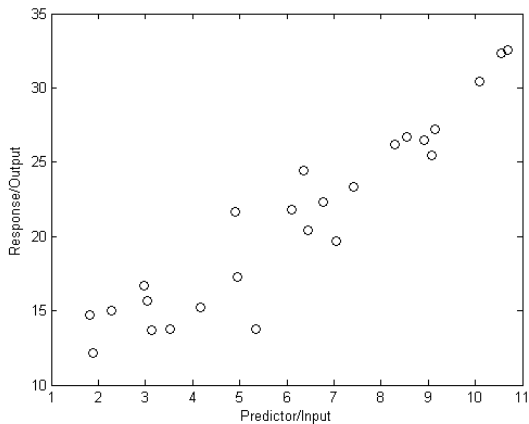
# Linear Regression

- ▶ Want to find parameters for a function of the form

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

- ▶ Distribution of error random variable not specified

## Quick Example - Scatter Plot



# 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
- ▶  $\beta_0$  and  $\beta_1$  are parameters
- ▶  $X_i$  is a known constant, the value of the predictor variable in the  $i^{th}$  trial
- ▶  $\epsilon_i$  is a random error term with mean  $\mathbb{E}(\epsilon_i)$  and variance  $\text{Var}(\epsilon_i) = \sigma^2$
- ▶  $i = 1, \dots, n$

# Properties

- ▶ The response  $Y_i$  is the sum of two components
  - ▶ Constant term  $\beta_0 + \beta_1 X_i$
  - ▶ Random term  $\epsilon_i$
- ▶ The expected response is

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \beta_0 + \beta_1 X_i + \mathbb{E}(\epsilon_i) \\ &= \beta_0 + \beta_1 X_i\end{aligned}$$

# Expectation Review

- ▶ Definition

$$\mathbb{E}(X) = \int X P(X) dX, X \in \mathcal{R}$$

- ▶ Linearity property

$$\begin{aligned}\mathbb{E}(aX) &= a \mathbb{E}(X) \\ \mathbb{E}(aX + bY) &= a \mathbb{E}(X) + b \mathbb{E}(Y)\end{aligned}$$

- ▶ Obvious from definition

## Example Expectation Derivation

$$P(X) = 2X, 0 \leq X \leq 1$$

Expectation

$$\begin{aligned}\mathbb{E}(X) &= \int_0^1 X P(X) dX \\ &= \int_0^1 2X^3 dX \\ &= \frac{2X^4}{4} \Big|_0^1 \\ &= \frac{1}{2}\end{aligned}$$



# Expectation of a Product of Random Variables

If  $X, Y$  are random variables with joint distribution  $P(X, Y)$  then the expectation of the product is given by

$$\mathbb{E}(XY) = \int_{XY} XY P(X, Y) dX dY.$$

## Expectation of a product of random variables

What if  $X$  and  $Y$  are independent? If  $X$  and  $Y$  are independent with density functions  $f$  and  $g$  respectively then

$$\begin{aligned}\mathbb{E}(XY) &= \int_{XY} XYf(X)g(Y)dXdY \\ &= \int_X \int_Y XYf(X)g(Y)dXdY \\ &= \int_X Xf(X) \left[ \int_Y Yg(Y)dY \right] dX \\ &= \int_X Xf(X) \mathbb{E}(Y) dX \\ &= \mathbb{E}(X) \mathbb{E}(Y)\end{aligned}$$

# Regression Function

- ▶ The response  $Y_i$  comes from a probability distribution with mean

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$$

- ▶ This means the regression function is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X$$

Since the regression function relates the means of the probability distributions of  $Y$  for a given  $X$  to the level of  $X$

# Error Terms

- ▶ The response  $Y_i$  in the  $i^{th}$  trial exceeds or falls short of the value of the regression function by the error term amount  $\epsilon_i$
- ▶ The error terms  $\epsilon_i$  are assumed to have constant variance  $\sigma^2$

# Response Variance

Responses  $Y_i$  have the same constant variance

$$\begin{aligned}\text{Var}(Y_i) &= \text{Var}(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \text{Var}(\epsilon_i) \\ &= \sigma^2\end{aligned}$$

## Variance ( $2^{nd}$ central moment) Review

- ▶ Continuous distribution

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int (X - \mathbb{E}(X))^2 P(X) dX, X \in \mathcal{R}$$

- ▶ Discrete distribution

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \sum_i (X_i - \mathbb{E}(X))^2 P(X_i), X \in \mathcal{Z}$$

## Alternative Form for Variance

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\&= \mathbb{E}((X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2)) \\&= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\&= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \\&= \mathbb{E}(X^2) - \mathbb{E}(X)^2.\end{aligned}$$

## Example Variance Derivation

$$P(X) = 2X, 0 \leq X \leq 1$$

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\&= \int_0^1 2X X^2 dX - \left(\frac{2}{3}\right)^2 \\&= \frac{2X^4}{4} \Big|_0^1 - \frac{4}{9} \\&= \frac{1}{2} - \frac{4}{9} = \frac{1}{18}\end{aligned}$$



# Variance Properties

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) \text{ if } X \perp\!\!\!\perp Y$$

$$\text{Var}(a + cX) = c^2 \text{Var}(X) \text{ if } a, c \text{ both constant}$$

More generally

$$\text{Var}\left(\sum a_i X_i\right) = \sum_i \sum_j a_i a_j \text{Cov}(X_i, X_j)$$

# Covariance

- ▶ The covariance between two real-valued random variables  $X$  and  $Y$ , with expected values  $\mathbb{E}(X) = \mu$  and  $\mathbb{E}(Y) = \nu$  is defined as

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu)(Y - \nu))$$

- ▶ Which can be rewritten as

$$\text{Cov}(X, Y) = \mathbb{E}(XY - \nu X - \mu Y + \mu\nu),$$

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \nu \mathbb{E}(X) - \mu \mathbb{E}(Y) + \mu\nu,$$

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mu\nu.$$

# Covariance of Independent Variables

If  $X$  and  $Y$  are independent, then their covariance is zero. This follows because under independence

$$\mathbb{E}(XY) = \mathbb{E}(X) \mathbb{E}(Y) = \mu\nu.$$

and then

$$\text{Cov}(XY) = \mu\nu - \mu\nu = 0.$$

# Least Squares Linear Regression

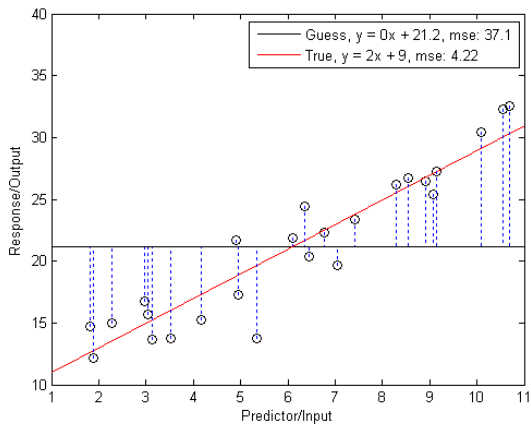
- Seek to minimize

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

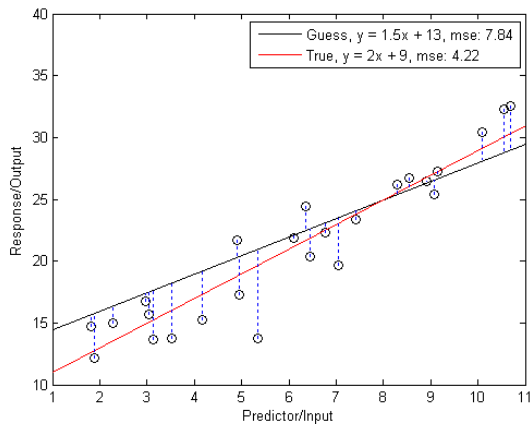
- By careful choice of  $b_0$  and  $b_1$  where  $b_0$  is a point estimator for  $\beta_0$  and  $b_1$  is the same for  $\beta_1$

How?

# Guess #1



## Guess #2



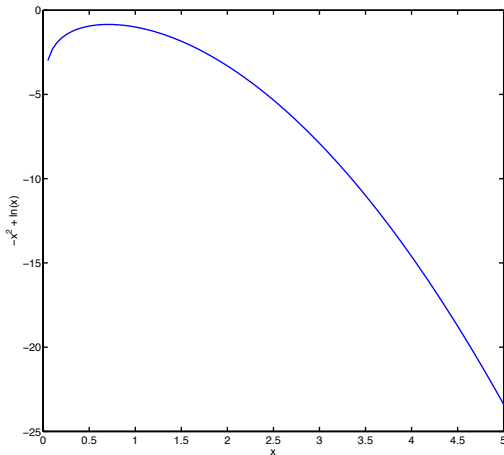
# Function maximization

- ▶ 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

$$\operatorname{argmax}_x -x^2 + \ln(x)$$





# 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  $\beta_0$  and  $\beta_1$  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

$$\frac{dQ}{db_0} = 0$$

$$\frac{dQ}{db_1} = 0$$

# 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

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$\bar{X} = \frac{\sum X_i}{n}$$

$$\bar{Y} = \frac{\sum Y_i}{n}$$