

# Regression Estimation - Least Squares and Maximum Likelihood

Dr. Frank Wood

# Least Squares Max(min)imization

- ▶ Function to minimize w.r.t.  $\beta_0, \beta_1$

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

- ▶ Minimize this by maximizing  $-Q$
- ▶ Find partials and set both equal to zero

$$\frac{dQ}{d\beta_0} = 0$$

$$\frac{dQ}{d\beta_1} = 0$$

# Normal Equations

- ▶ The result of this maximization step are called the normal equations.  $b_0$  and  $b_1$  are called point estimators of  $\beta_0$  and  $\beta_1$  respectively.

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

- ▶ 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

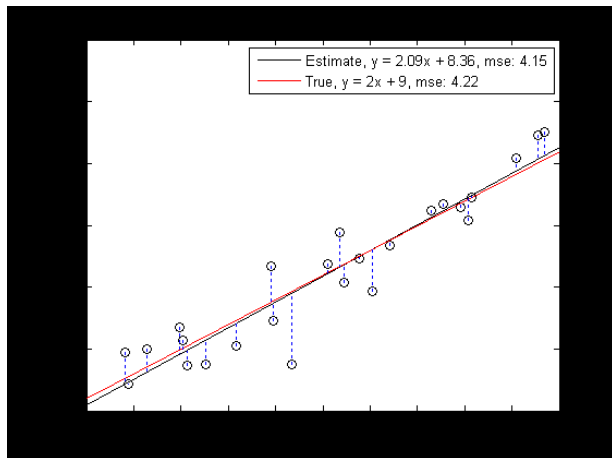
$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

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

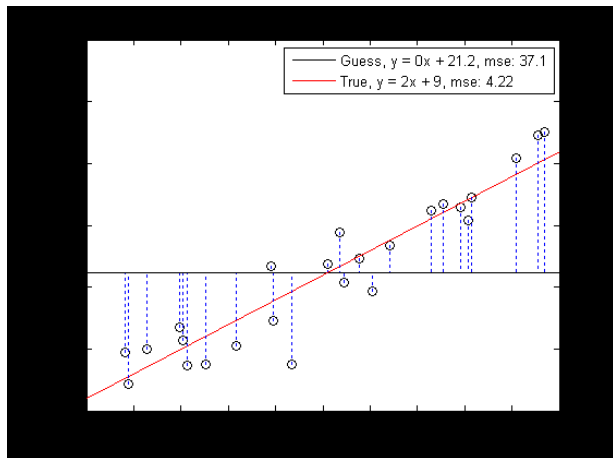
$$\bar{X} = \frac{\sum X_i}{n}$$

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

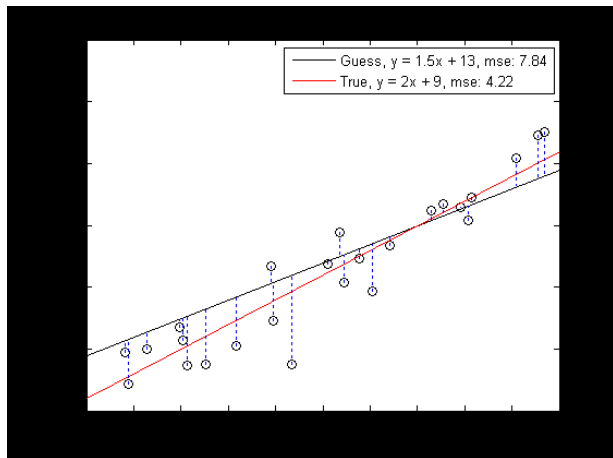
# Least Squares Fit



# Guess #1



## Guess #2



## Looking Ahead: Matrix Least Squares

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 & 1 \\ X_2 & 1 \\ \vdots & \\ X_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix}$$

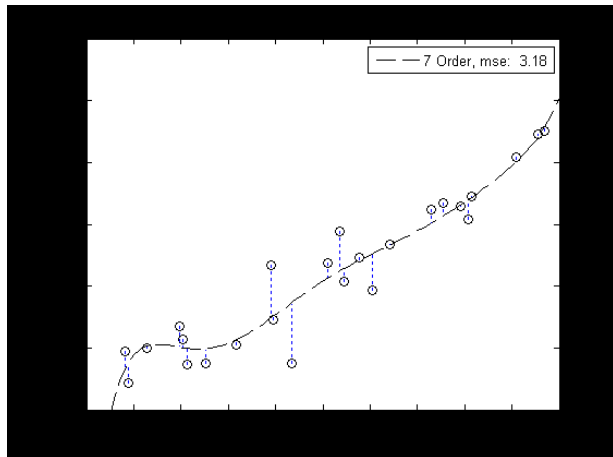
Solution to this equation is solution to least squares linear regression (and maximum likelihood under normal error distribution assumption)



# Questions to Ask

- ▶ Is the relationship really linear?
- ▶ What is the distribution of the of “errors”?
- ▶ Is the fit good?
- ▶ How much of the variability of the response is accounted for by including the predictor variable?
- ▶ Is the chosen predictor variable the best one?

# Is This Better?



# Goals for First Half of Course

- ▶ How to do linear regression
  - ▶ Self familiarization with software tools
- ▶ How to interpret standard linear regression results
- ▶ How to derive tests
- ▶ How to assess and address deficiencies in regression models

## Estimators for $\beta_0, \beta_1, \sigma^2$

- ▶ We want to establish properties of estimators for  $\beta_0, \beta_1$ , and  $\sigma^2$  so that we can construct hypothesis tests and so forth
- ▶ We will start by establishing some properties of the regression solution.

# Properties of Solution

- ▶ The  $i^{th}$  residual is defined to be

$$e_i = Y_i - \hat{Y}_i$$

- ▶ The sum of the residuals is zero:

$$\begin{aligned}\sum_i e_i &= \sum (Y_i - b_0 - b_1 X_i) \\ &= \sum Y_i - nb_0 - b_1 \sum X_i \\ &= 0\end{aligned}$$

# Properties of Solution

The sum of the observed values  $Y_i$  equals the sum of the fitted values  $\hat{Y}_i$

$$\begin{aligned}\sum_i Y_i &= \sum_i \hat{Y}_i \\&= \sum_i (b_1 X_i + b_0) \\&= \sum_i (b_1 X_i + \bar{Y} - b_1 \bar{X}) \\&= b_1 \sum_i X_i + n\bar{Y} - b_1 n\bar{X} \\&= b_1 n\bar{X} + \sum_i Y_i - b_1 n\bar{X}\end{aligned}$$

# Properties of Solution

The sum of the weighted residuals is zero when the residual in the  $i^{th}$  trial is weighted by the level of the predictor variable in the  $i^{th}$  trial

$$\begin{aligned}\sum_i X_i e_i &= \sum_i (X_i (Y_i - b_0 - b_1 X_i)) \\ &= \sum_i X_i Y_i - b_0 \sum_i X_i - b_1 \sum_i (X_i^2) \\ &= 0\end{aligned}$$

## Properties of Solution

The regression line always goes through the point

$$\bar{X}, \bar{Y}$$



# Estimating Error Term Variance $\sigma^2$

- ▶ Review estimation in non-regression setting.
- ▶ Show estimation results for regression setting.

# Estimation Review

- ▶ An estimator is a rule that tells how to calculate the value of an estimate based on the measurements contained in a sample
- ▶ i.e. the sample mean

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

# Point Estimators and Bias

- ▶ Point estimator

$$\hat{\theta} = f(\{Y_1, \dots, Y_n\})$$

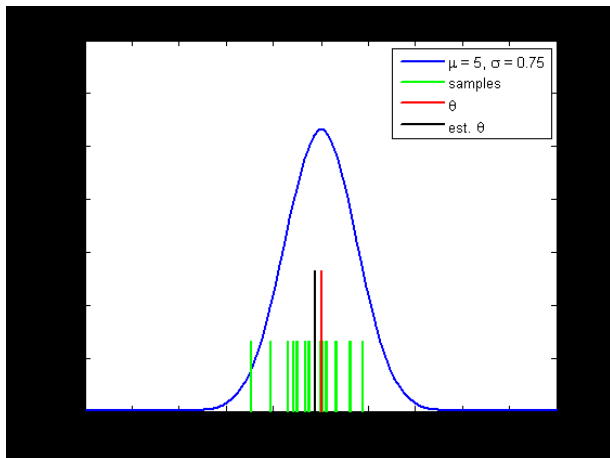
- ▶ Unknown quantity / parameter

$$\theta$$

- ▶ Definition: Bias of estimator

$$B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

# One Sample Example



# Distribution of Estimator

- ▶ If the estimator is a function of the samples and the distribution of the samples is known then the distribution of the estimator can (often) be determined
  - ▶ Methods
    - ▶ Distribution (CDF) functions
    - ▶ Transformations
    - ▶ Moment generating functions
    - ▶ Jacobians (change of variable)

## Example

- ▶ Samples from a  $Normal(\mu, \sigma^2)$  distribution

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

- ▶ Estimate the population mean

$$\theta = \mu, \quad \hat{\theta} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

# Sampling Distribution of the Estimator

- First moment

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(Y_i) = \frac{n\mu}{n} = \theta\end{aligned}$$

- This is an example of an unbiased estimator

$$B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = 0$$

# Variance of Estimator

- Definition: Variance of estimator

$$\text{Var}(\hat{\theta}) = \mathbb{E}([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2)$$

- Remember:

$$\begin{aligned}\text{Var}(cY) &= c^2 \text{Var}(Y) \\ \text{Var}\left(\sum_{i=1}^n Y_i\right) &= \sum_{i=1}^n \text{Var}(Y_i)\end{aligned}$$

Only if the  $Y_i$  are independent with finite variance



## Example Estimator Variance

- For  $N(0, 1)$  mean estimator

$$\begin{aligned}\text{Var}(\hat{\theta}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

- Note assumptions

# Central Limit Theorem Review

## Central Limit Theorem

Let  $Y_1, Y_2, \dots, Y_n$  be iid random variables with  $\mathbb{E}(Y_i) = \mu$  and  $\text{Var}(Y_i) = \sigma^2 < \infty$ . Define.

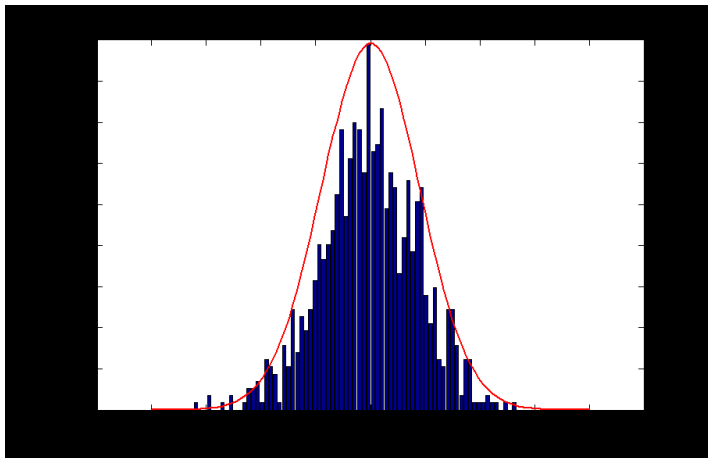
$$U_n = \sqrt{n} \left( \frac{\bar{Y} - \mu}{\sigma} \right) \quad \text{where} \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (1)$$

Then the distribution function of  $U_n$  converges to a standard normal distribution function as  $n \rightarrow \infty$ .

## Alternately

$$P(a \leq U_n \leq b) \rightarrow \int_a^b \left( \frac{1}{\sqrt{2\pi}} \right) e^{-\frac{u^2}{2}} du \quad (2)$$

# Distribution of sample mean estimator



# Bias Variance Trade-off

- ▶ The mean squared error of an estimator

$$MSE(\hat{\theta}) = \mathbb{E}([\hat{\theta} - \theta]^2)$$

- ▶ Can be re-expressed

$$MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) + (B(\hat{\theta}))^2$$

$$MSE = VAR + BIAS^2$$

Proof

$$\begin{aligned}MSE(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) \\&= \mathbb{E}([\hat{\theta} - \mathbb{E}(\hat{\theta})] + [\mathbb{E}(\hat{\theta}) - \theta])^2 \\&= \mathbb{E}([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2) + 2\mathbb{E}([\mathbb{E}(\hat{\theta}) - \theta][\hat{\theta} - \mathbb{E}(\hat{\theta})]) + \mathbb{E}([\mathbb{E}(\hat{\theta}) - \theta]^2) \\&= \text{Var}(\hat{\theta}) + 2\mathbb{E}([\mathbb{E}(\hat{\theta})][\hat{\theta} - \mathbb{E}(\hat{\theta})] - \theta[\hat{\theta} - \mathbb{E}(\hat{\theta})]) + (B(\hat{\theta}))^2 \\&= \text{Var}(\hat{\theta}) + 2(0 + 0) + (B(\hat{\theta}))^2 \\&= \text{Var}(\hat{\theta}) + (B(\hat{\theta}))^2\end{aligned}$$

# Trade-off

- ▶ Think of variance as confidence and bias as correctness.
  - ▶ Intuitions (largely) apply
- ▶ Sometimes choosing a biased estimator can result in an overall lower MSE if it exhibits lower variance.
- ▶ Bayesian methods (later in the course) specifically introduce bias.

# Estimating Error Term Variance $\sigma^2$

- ▶ Regression model
- ▶ Variance of each observation  $Y_i$  is  $\sigma^2$  (the same as for the error term  $\epsilon_i$ )
- ▶ Each  $Y_i$  comes from a different probability distribution with different means that depend on the level  $X_i$
- ▶ The deviation of an observation  $Y_i$  must be calculated around its own estimated mean.

## $s^2$ estimator for $\sigma^2$

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

- ▶ MSE is an unbiased estimator of  $\sigma^2$

$$\mathbb{E}(MSE) = \sigma^2$$

- ▶ The sum of squares SSE has  $n-2$  “degrees of freedom” associated with it.
- ▶ Cochran's theorem (later in the course) tells us where degree's of freedom come from and how to calculate them.



# Normal Error Regression Model

- ▶ No matter how the error terms  $\epsilon_i$  are distributed, the least squares method provides unbiased point estimators of  $\beta_0$  and  $\beta_1$ 
  - ▶ that also have minimum variance among all unbiased linear estimators
- ▶ To set up interval estimates and make tests we need to specify the distribution of the  $\epsilon_i$
- ▶ We will assume that the  $\epsilon_i$  are normally distributed.

# Normal Error Regression 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
- ▶  $\epsilon_i \sim_{iid} N(0, \sigma^2)$   
*note this is different, now we know the distribution*
- ▶  $i = 1, \dots, n$

# Notational Convention

- ▶ When you see  $\epsilon_i \sim_{iid} N(0, \sigma^2)$
- ▶ It is read as  $\epsilon_i$  is distributed identically and independently according to a normal distribution with mean 0 and variance  $\sigma^2$
- ▶ Examples
  - ▶  $\theta \sim \text{Poisson}(\lambda)$
  - ▶  $z \sim G(\theta)$

# Maximum Likelihood Principle

The method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data.

# Likelihood Function

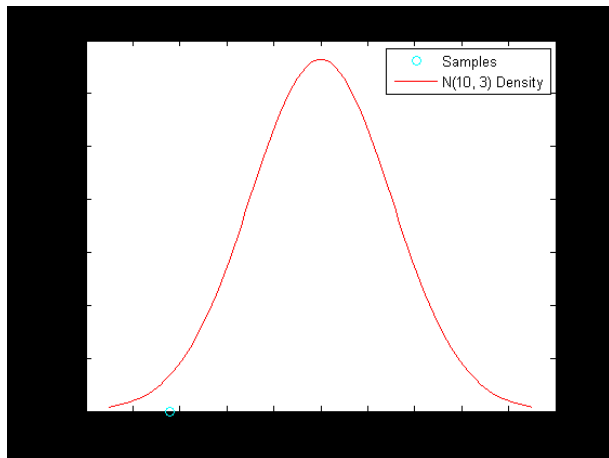
If

$$X_i \sim F(\Theta), i = 1 \dots n$$

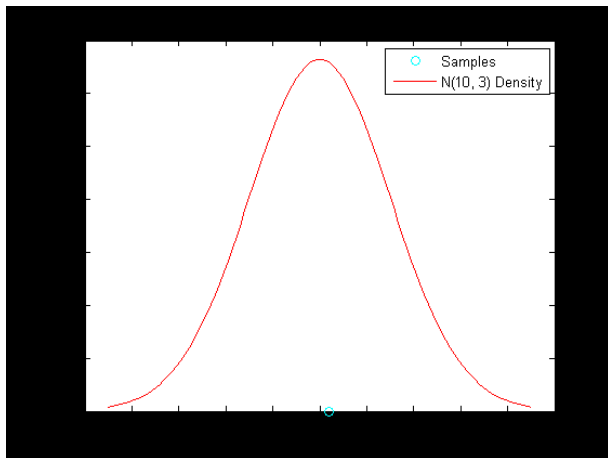
then the likelihood function is

$$\mathcal{L}(\{X_i\}_{i=1}^n, \Theta) = \prod_{i=1}^n F(X_i; \Theta)$$

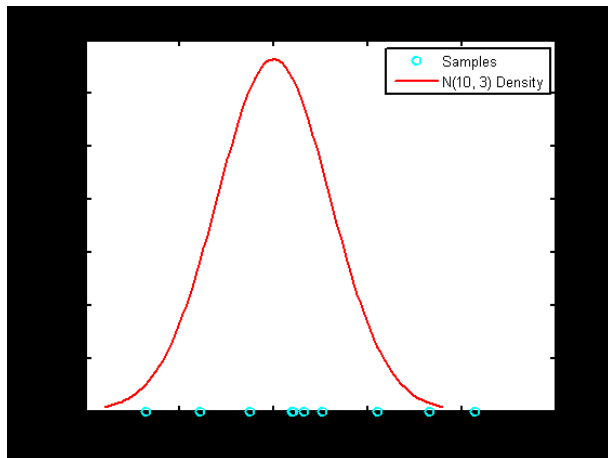
## Example, $N(10, 3)$ Density, Single Obs.



## Example, $N(10, 3)$ Density, Single Obs. Again



## Example, $N(10, 3)$ Density, Multiple Obs.





# Maximum Likelihood Estimation

- ▶ The likelihood function can be maximized w.r.t. the parameter(s)  $\Theta$ , doing this one can arrive at estimators for parameters as well.

$$\mathcal{L}(\{X_i\}_{i=1}^n, \Theta) = \prod_{i=1}^n F(X_i; \Theta)$$

- ▶ To do this, find solutions to (analytically or by following gradient)

$$\frac{d\mathcal{L}(\{X_i\}_{i=1}^n, \Theta)}{d\Theta} = 0$$

## Important Trick

Never (almost) maximize the likelihood function, maximize the log likelihood function instead.

$$\begin{aligned}\log(\mathcal{L}(\{X_i\}_{i=1}^n, \Theta)) &= \log\left(\prod_{i=1}^n F(X_i; \Theta)\right) \\ &= \sum_{i=1}^n \log(F(X_i; \Theta))\end{aligned}$$

Quite often the log of the density is easier to work with mathematically.

# ML Normal Regression

Likelihood function

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}\end{aligned}$$

which if you maximize (how?) w.r.t. to the parameters you get...

# Maximum Likelihood Estimator(s)

- ▶  $\beta_0$   
 $b_0$  same as in least squares case
- ▶  $\beta_1$   
 $b_1$  same as in least squares case
- ▶  $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n}$$

- ▶ Note that ML estimator is biased as  $s^2$  is unbiased and

$$s^2 = MSE = \frac{n}{n-2} \hat{\sigma}^2$$

# Comments

- ▶ Least squares minimizes the squared error between the prediction and the true output
- ▶ The normal distribution is fully characterized by its first two central moments (mean and variance)
- ▶ Food for thought:
  - ▶ What does the bias in the ML estimator of the error variance mean? And where does it come from?