Regression Estimation – Least Squares and Maximum Likelihood

Dr. Frank Wood

Least Squares Max(min)imization

- Function to minimize w.r.t. β_0, β_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

$$\begin{aligned} \frac{dQ}{d\beta_0} &= 0 \\ \frac{dQ}{d\beta_1} &= 0 \end{aligned}$$

go to board

Normal Equations

- The result of this maximization step are called the normal equations. b_0 and b_1 are called point estimators of β_0 and β_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…

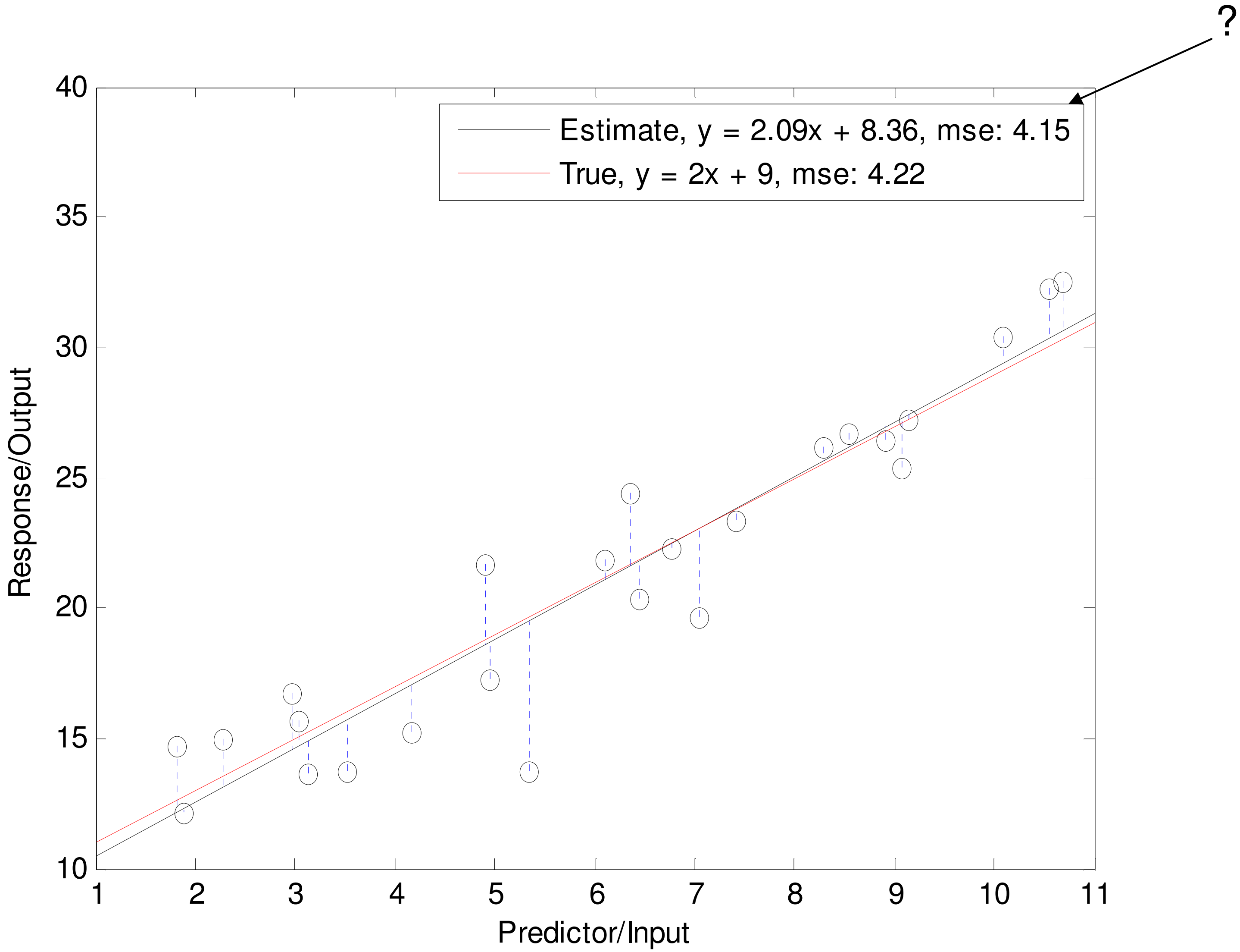
Write these on board

Solution to Normal Equations

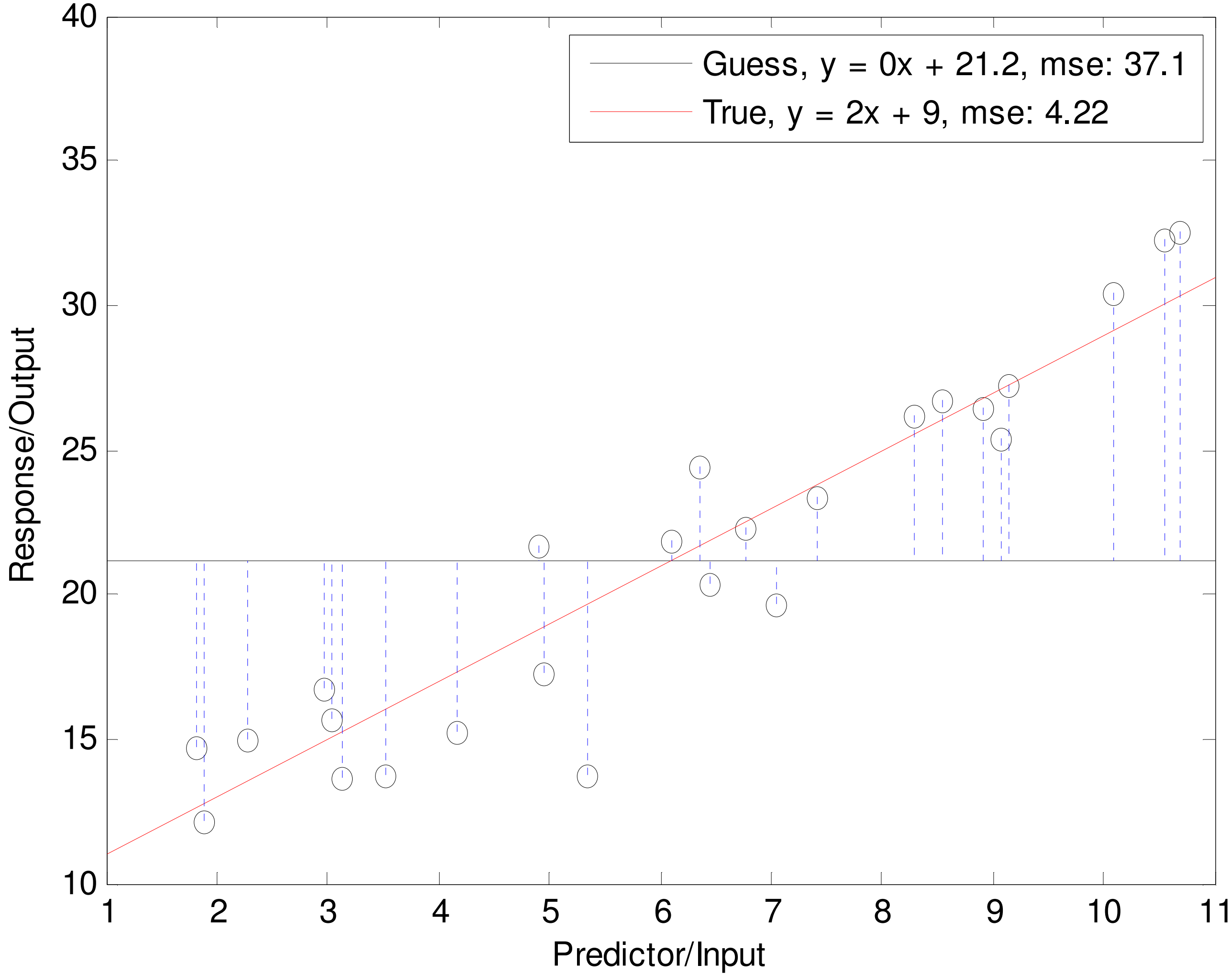
- After a lot of algebra one arrives at

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

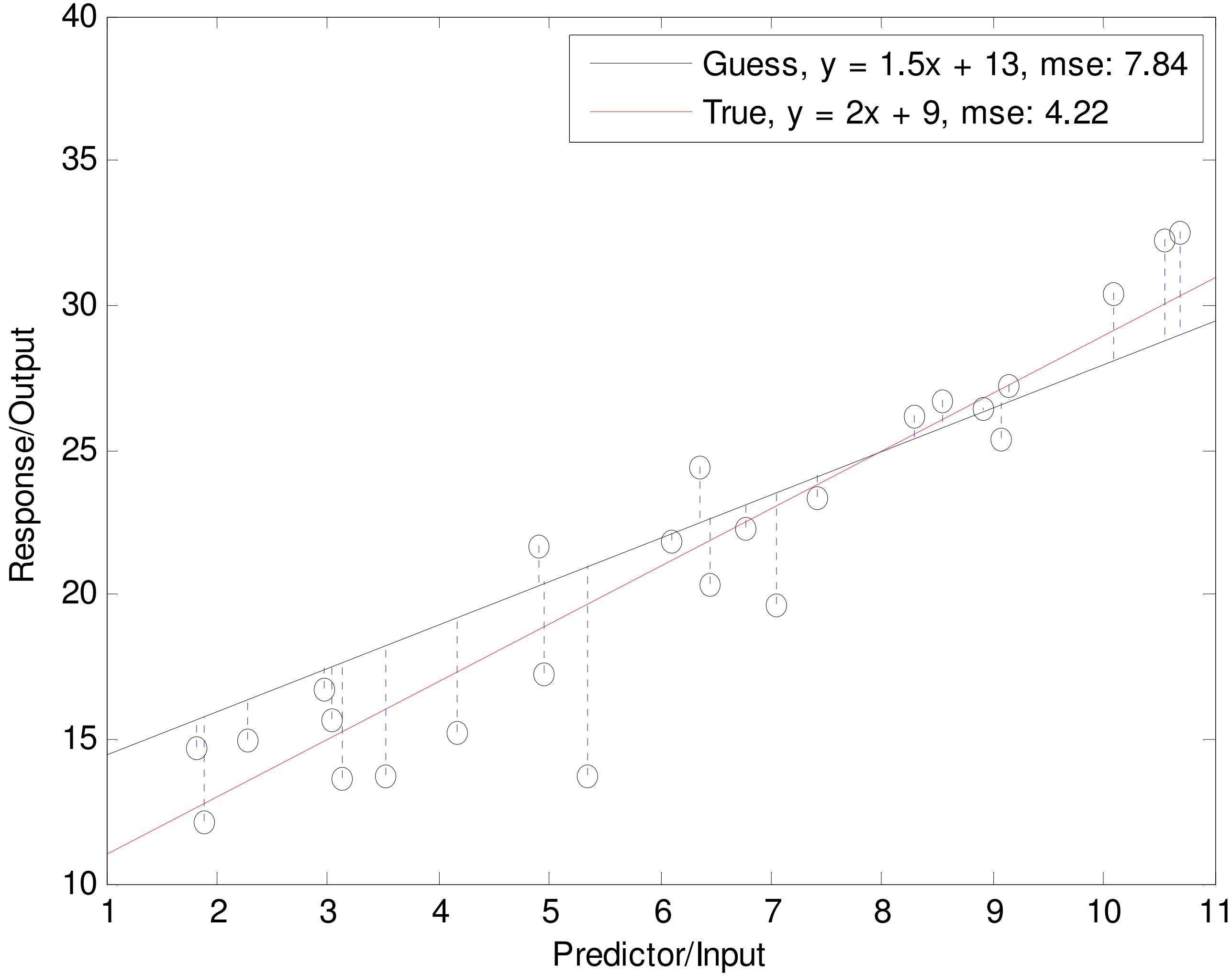
Least Squares Fit



Guess #1


Guess, y = 0x + 21.2, mse: 37.1
True, y = 2x + 9, mse: 4.22
40
35
30
25
20
15
10
1
2
3
4
5
6
7
8
9
10
11
Response/Output
Predictor/Input

Guess #2



Guess, y = 1.5x + 13, mse: 7.84
True, y = 2x + 9, mse: 4.22
Response/Output
Predictor/Input
40
35
30
25
20
15
10
1
2
3
4
5
6
7
8
9
10
11

Looking Ahead: Matrix Least Squares

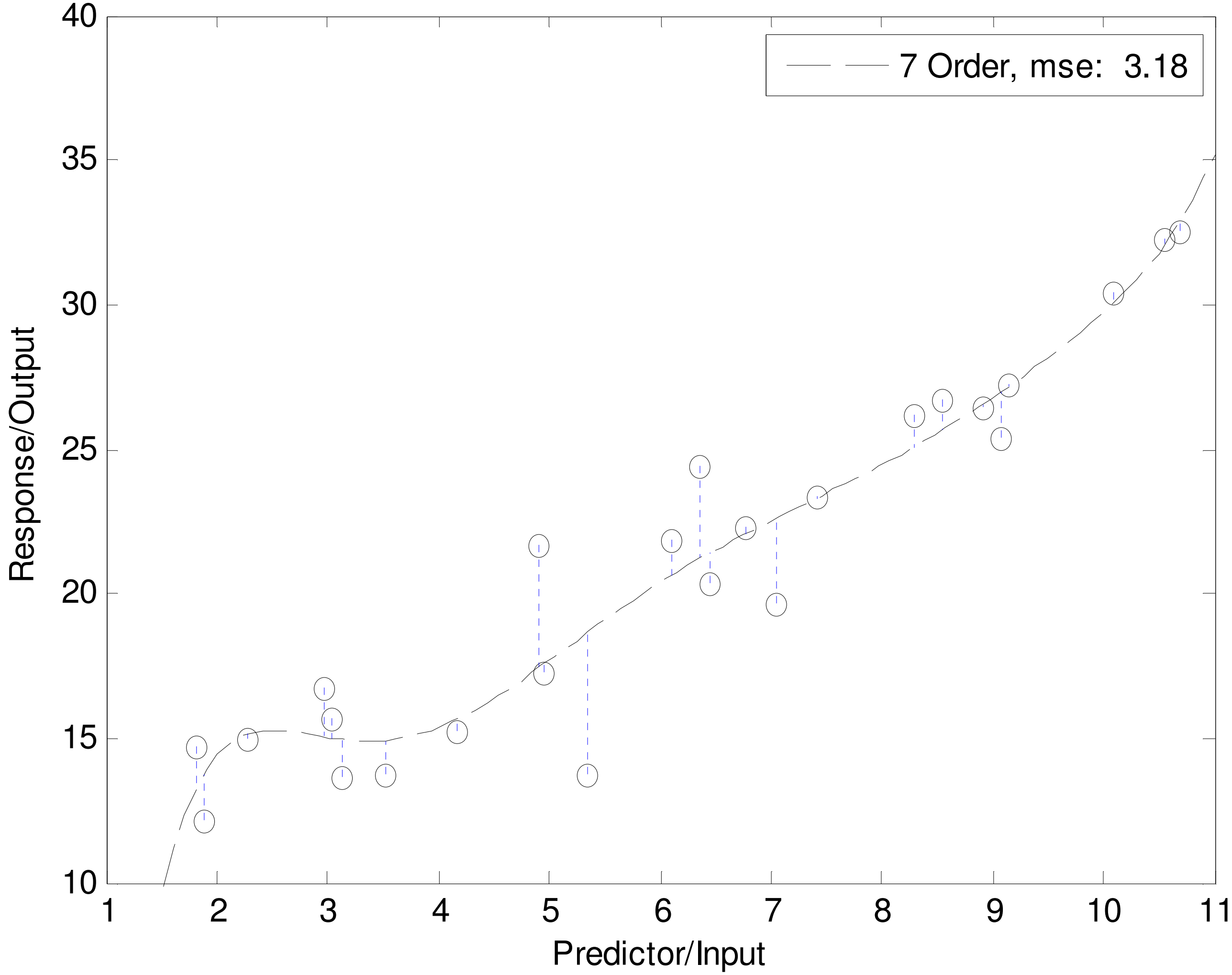
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 & 1 \\ X_2 & 1 \\ \vdots & \\ X_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix}$$

- 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

Properties of Solution

- The ith residual is defined to be

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

- The sum of the residuals is zero:

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

By first normal equation.

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

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

By second normal equation.

Properties of Solution

- The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the fitted value of the response variable for the i^{th} trial

$$
\begin{aligned}
\sum_i \hat{Y}_i e_i &= \sum_i (b_0 + b_1 X_i) e_i \\
&= b_0 \sum_i e_i + b_1 \sum_i e_i X_i \\
&= 0
\end{aligned}
$$

By previous properties.

Properties of Solution

- The regression line always goes through the point

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

Estimating Error Term Variance σ^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

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

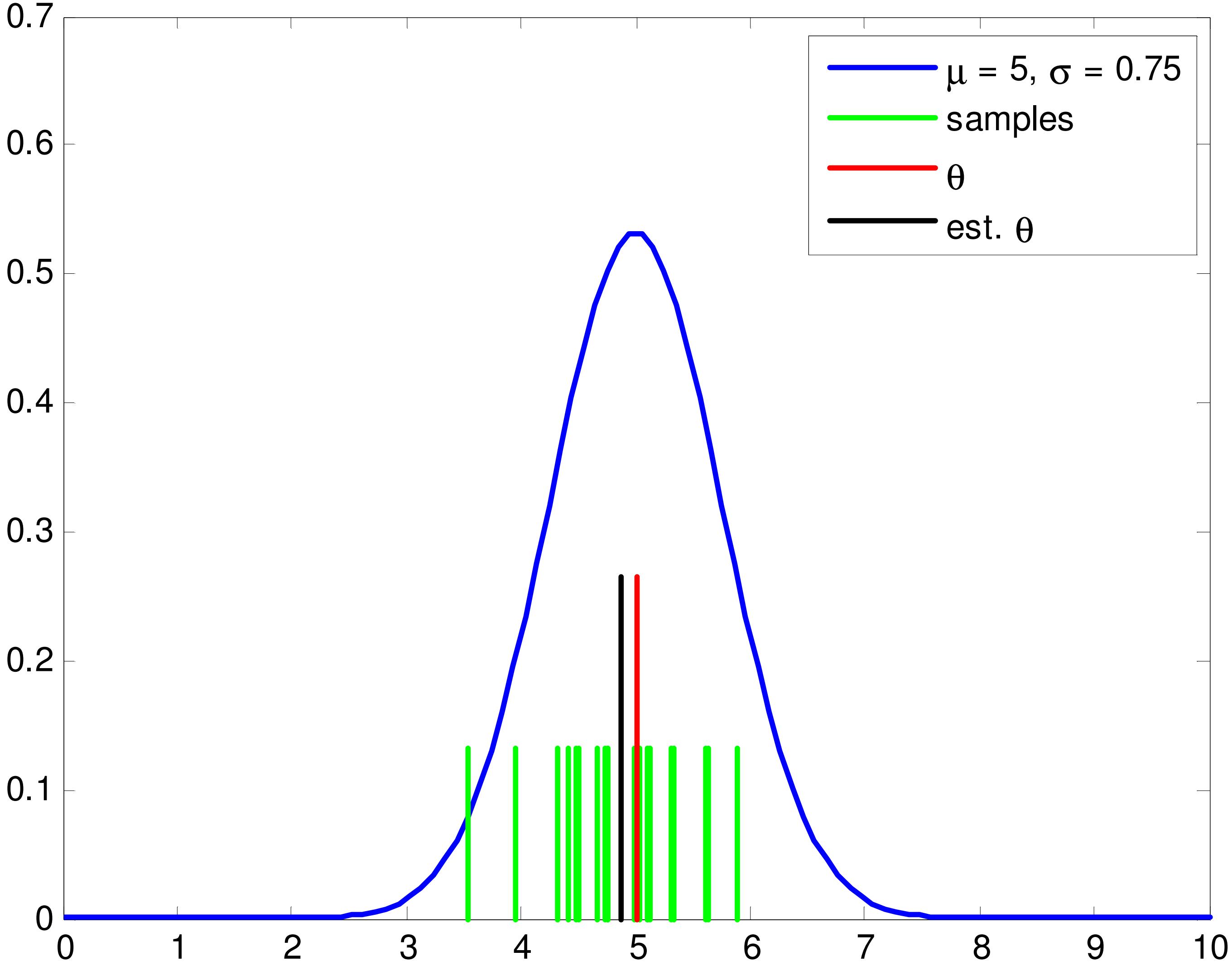
- Unknown quantity / parameter

$$\theta$$

- Definition: Bias of estimator

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

One Sample Example



run bias_example_plot.m

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(μ,σ^2) distribution

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

- 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

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

- This is an example of an unbiased estimator

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

Variance of Estimator

- Definition: Variance of estimator

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

- Remember:

$$V(cY) = c^2 V(Y)$$

$$V(\textstyle\sum_{i=1}^n Y_i) = \sum_{i=1}^n V(Y_i)$$

Only if the Y_i are *independent* with *finite variance*

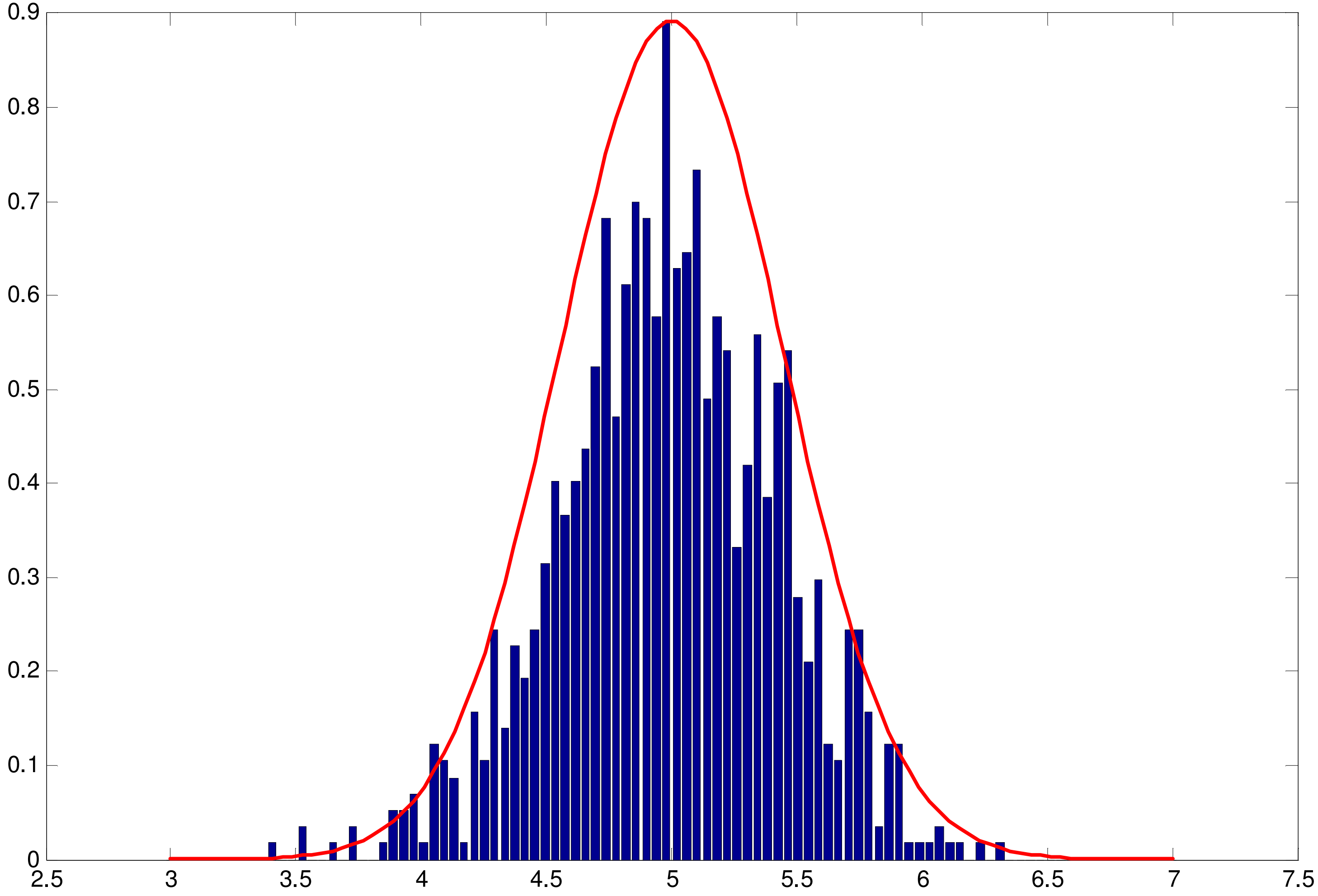
Example Estimator Variance

- For N(0,1) mean estimator

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

- Note assumptions

Distribution of sample mean estimator



1000 samples

Bias Variance Trade-off

- The mean squared error of an estimator

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

- Can be re-expressed

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

MSE = VAR + BIAS^2

- Proof

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

Trade-off

- Think of variance as confidence and bias as correctness.
  - Intuitions (largely) apply
- Sometimes a biased estimator can produce lower MSE if it lowers the variance.

Estimating Error Term Variance σ^2

- Regression model
- Variance of each observation Y_i is σ^2 (the same as for the error term ϵ_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 σ^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 σ^2

$$E(MSE) = \sigma^2$$

- The sum of squares SSE has n-2 degrees of freedom associated with it.

Normal Error Regression Model

- No matter how the error terms ϵ_i are distributed, the least squares method provides unbiased point estimators of β_0 and β_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 ϵ_i
- We will assume that the ϵ_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
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- $\epsilon_i \sim_{iid} N(0,\sigma^2)$
- $i = 1,\ldots,n$

Notational Convention

- When you see $\epsilon_i \sim_{\text{iid}} N(0,\sigma^2)$

- It is read as ϵ_i is distributed identically and independently according to a normal distribution with mean 0 and variance σ^2

- Examples
  - $\theta \sim$ Poisson(λ)
  - $z \sim G(\theta)$

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

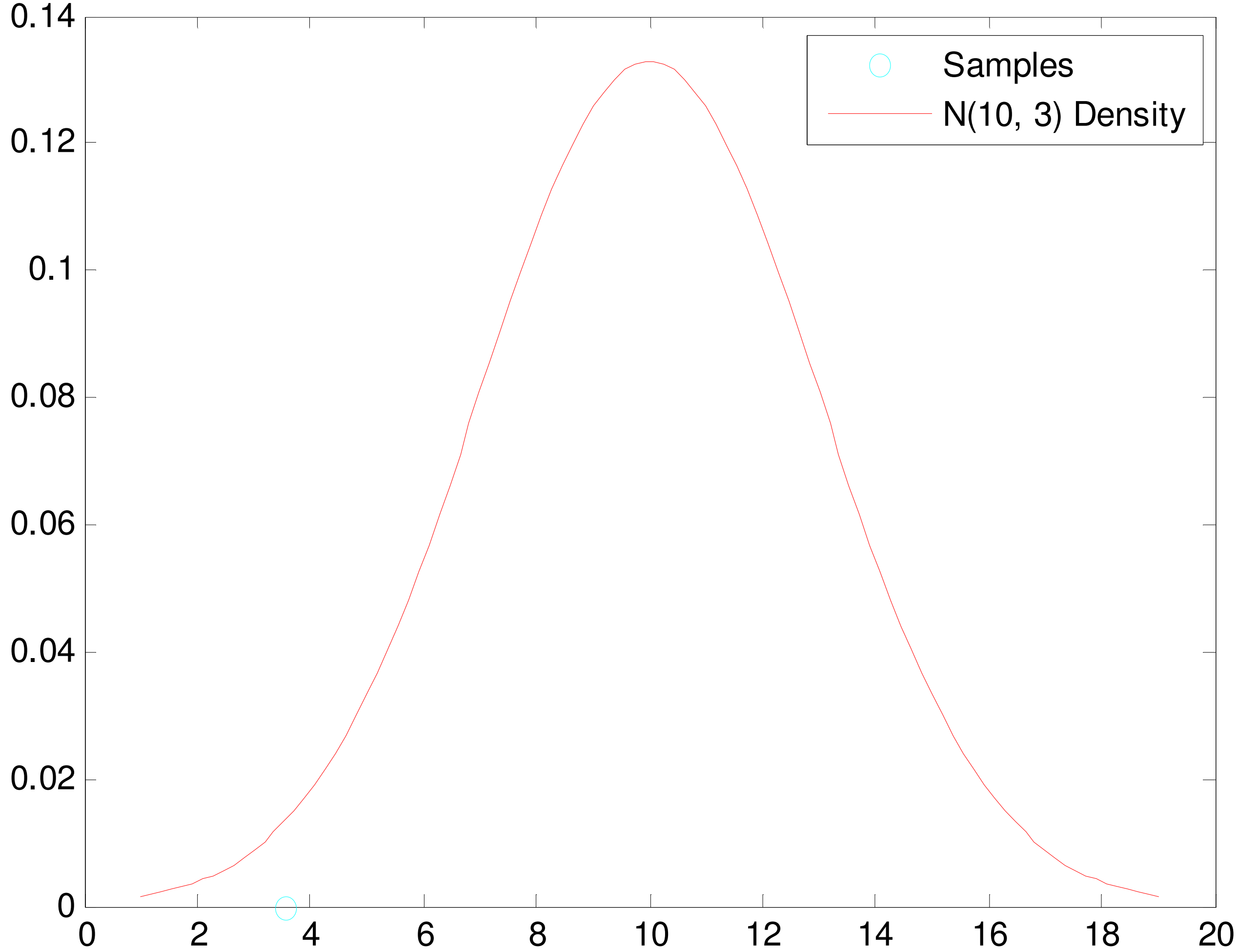
- If

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

then the likelihood function is

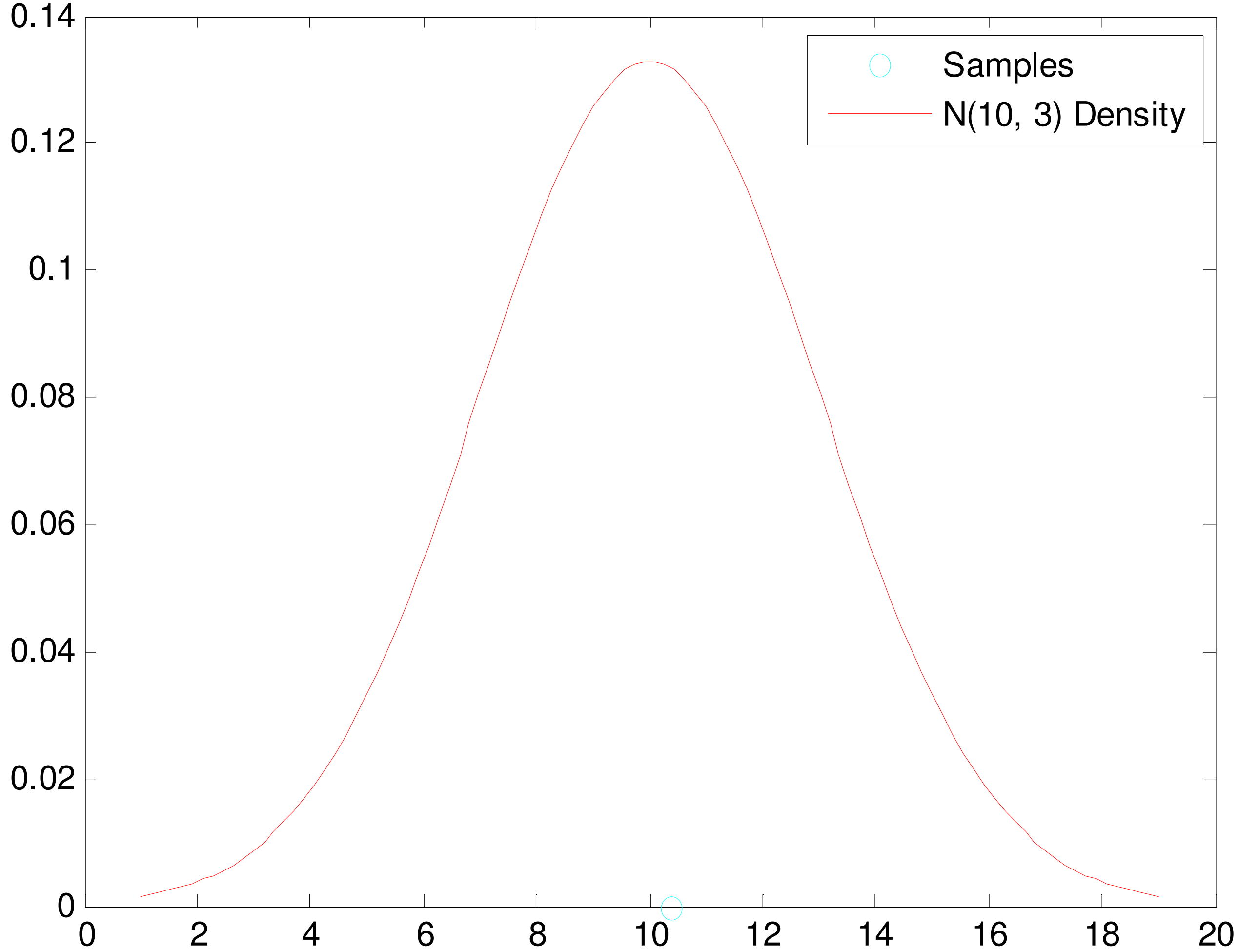
$$\mathcal{L}(\{X_i\}_{i=1}^n, \Theta) = \prod_{i=1}^n F(X_i; \Theta)$$

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



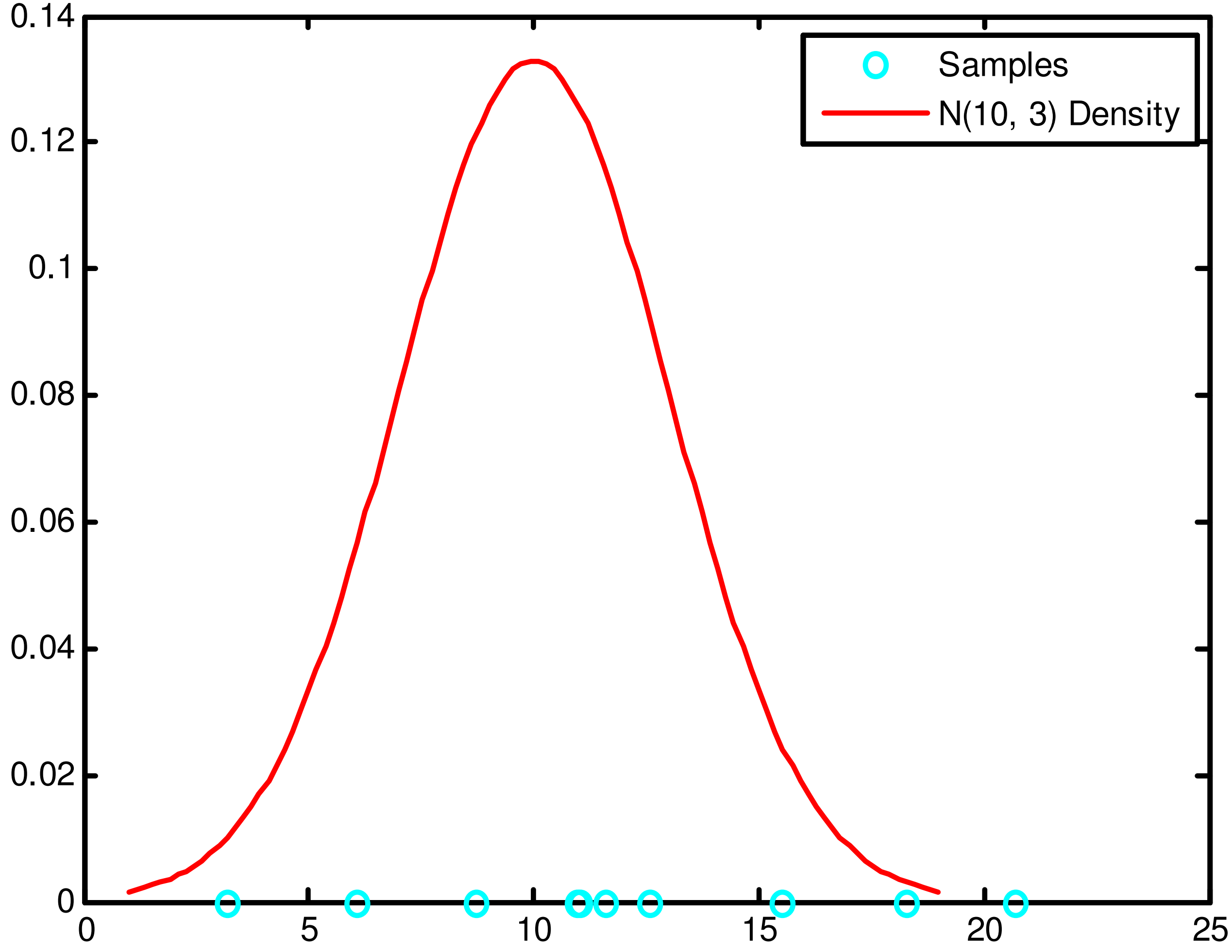
N=10, - log likelihood = 4.3038

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



N=10, - log likelihood = 4.3038

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



N=10, - log likelihood = 36.2204

Maximum Likelihood Estimation

- The likelihood function can be maximized w.r.t. the parameter(s) Θ, 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.

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

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

ML Normal Regression

- Likelihood function

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

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}$$

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

Maximum Likelihood Estimator(s)

- β_0
  - b_0 same as in least squares case
- β_1
  - b_1 same as in least squares case
- σ_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?