



# EIN 5332

## Simple Linear Regression

Chapter 23 Sections 23.5-11

Karen E. Schmahl Ph.D., P.E.

# Correlation Analysis

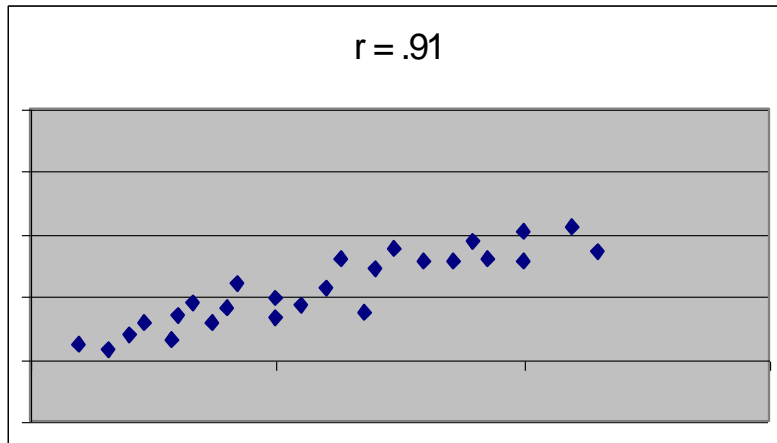
## Pearson Correlation Coefficient – r

- Reflects the degree of linear relationship between two variables.

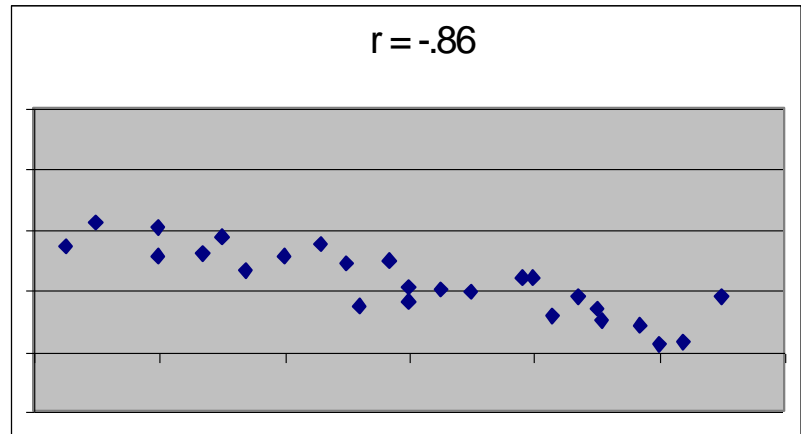
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

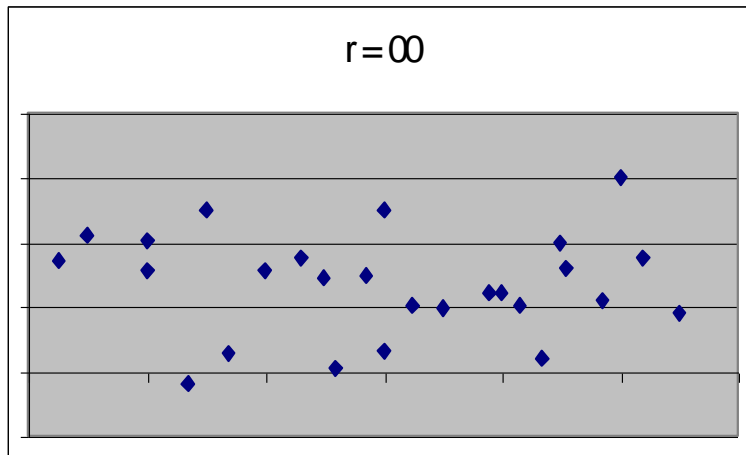
## Positive Correlation



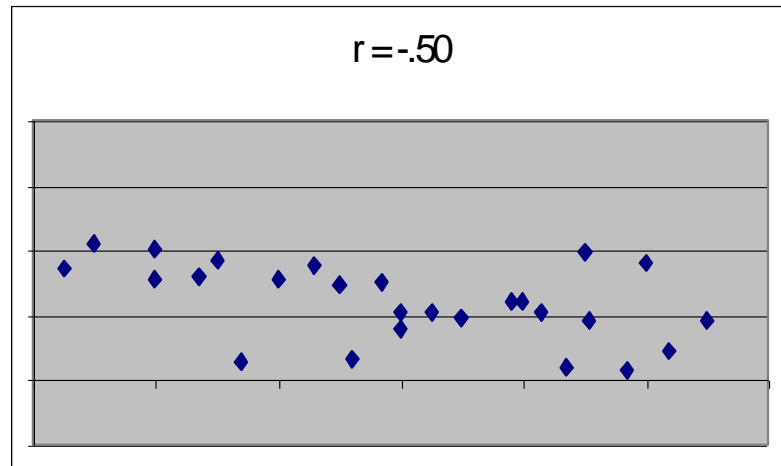
## Negative Correlation



## No Correlation



## ? Correlation



# Correlation Comprehension Question

Which of the following statements concerning the coefficient of linear correlation,  $r$  is not true?

- A.  $r = -1$  indicates a perfect relationship
- B.  $r = 0$  indicates the absence of a relationship
- C. The relation ship between the variables must be nonlinear.
- D.  $r = 0.81$  has the same predictive power as  $r = -.81$ .

# Correlation Comprehension Question

Given the following pairs of values of  $x$  and  $y$ ,

$X$  2, 3, 5, 6

$Y$  6, 8, 3, 7

What is the first step in determining if there is a linear association between  $x$  and  $y$ ?

- A. Draw a scatter plot
- B. Calculate the mean and standard deviation of  $x$  and  $y$
- C. Calculate the correlation coefficient
- D. Do nothing, get more data.

# Least Squares Line

Equation of a line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where

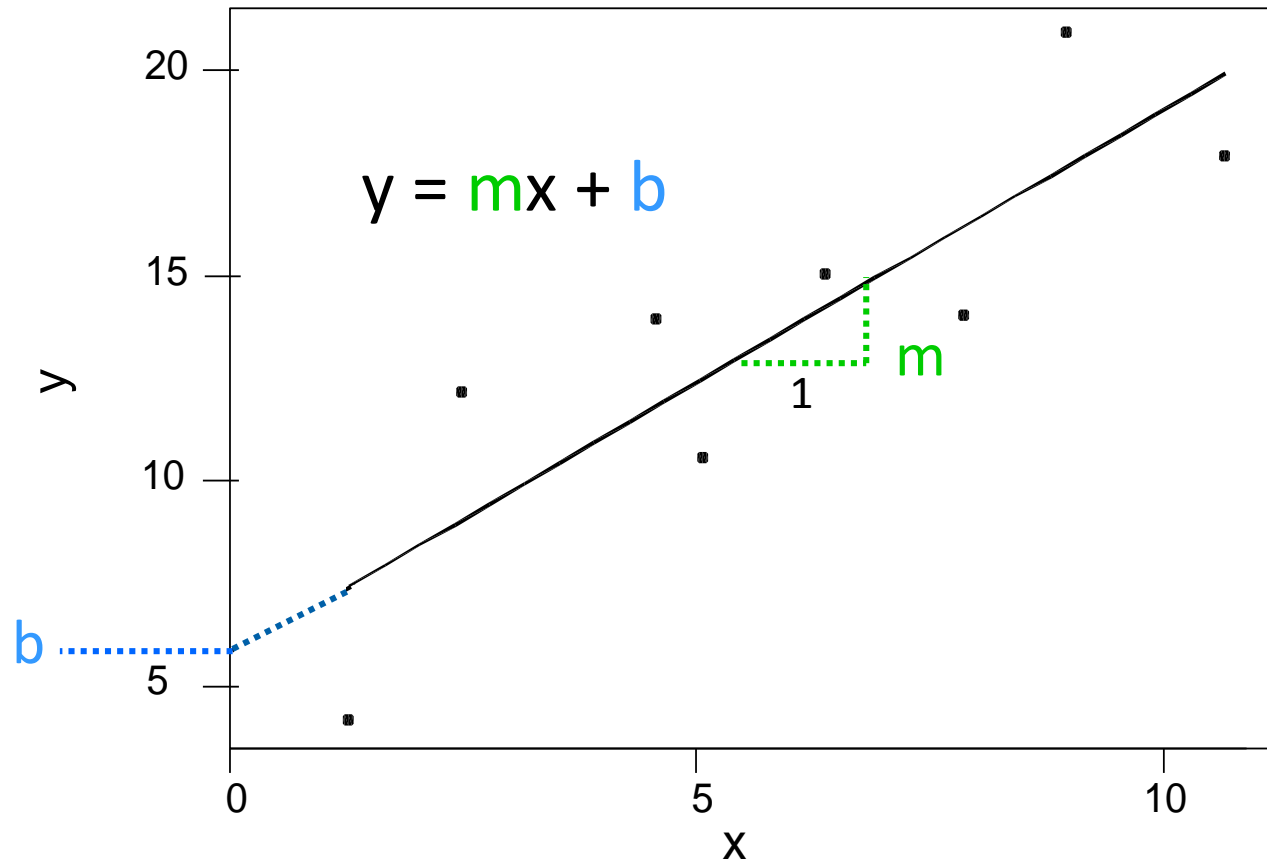
$\hat{\beta}_0$  = estimate of the y intercept

$\hat{\beta}_1$  = estimate of the slope

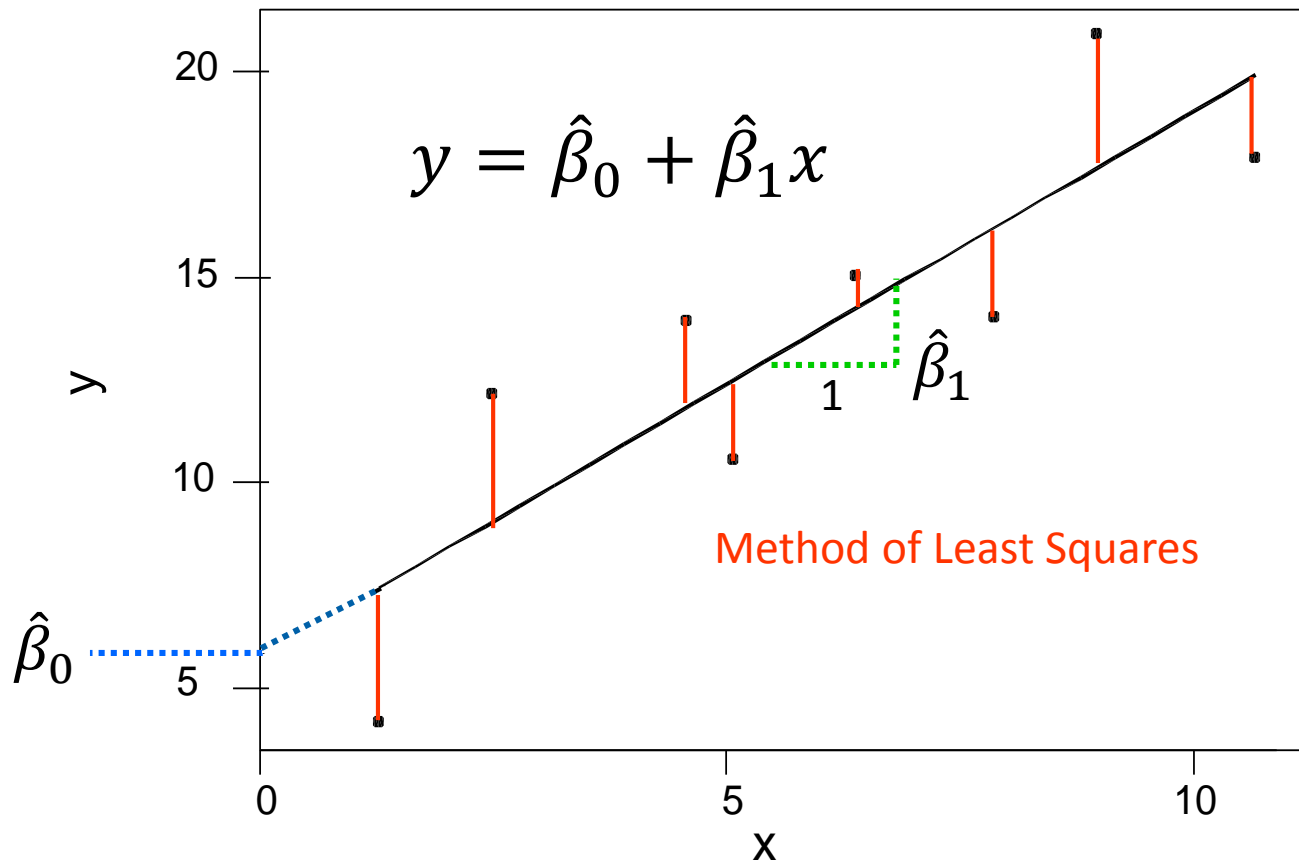
Note: anytime you see a “ ^ ” over a character, that means that you are going to be estimating the value.

# Equation of a Line

– other common notation



# Least Squares Line





# Least Squares Line

Equation of a line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where

$\hat{\beta}_1$  = estimate of the slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\hat{\beta}_0$  = estimate of the y intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Calculating the least squares line

Using equivalent formula for slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Set up data in corresponding pairs in table.

Calculate elements of formula:  $\sum x$ ,  $\sum x^2$ ,  $\sum xy$ ,  $\bar{x}$ ,  $\bar{y}$

Plug into formula to calculate slope  $\hat{\beta}_1$

Plug into formula to calculate intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured. The data collected is in the box below.

Time	Loss		
x	y	xy	x <sup>2</sup>
2	0.7	1.4	4
4	6.9	27.6	16
6	7.2	43.2	36
8	11.9	95.2	64
10	16.2	162	100
12	22.5	270	144
$\bar{x}$ , 7	$\bar{y}$ 10.9	$\sum xy$ 599.4	$\sum x^2$ 364

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

What is the equation of the least squares line?

The slope  $\hat{\beta}_1$  is

$$\hat{\beta}_1 = \frac{599.4 - 6(7)(10.9)}{364 - 6(7^2)} = 2.02$$

Time	Loss
x	y
2	0.7
4	6.9
6	7.2
8	11.9
10	16.2
12	22.5

xy	x <sup>2</sup>
1.4	4
27.6	16
43.2	36
95.2	64
162	100
270	144

$\bar{x}$ ,	$\bar{y}$	$\sum xy$	$\sum x^2$
7	10.9	599.4	364

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

What is the equation of the least squares line?

The slope  $\hat{\beta}_1$  is

$$\hat{\beta}_1 = \frac{599.4 - 6(7)(10.9)}{364 - 6(7^2)} = 2.02$$

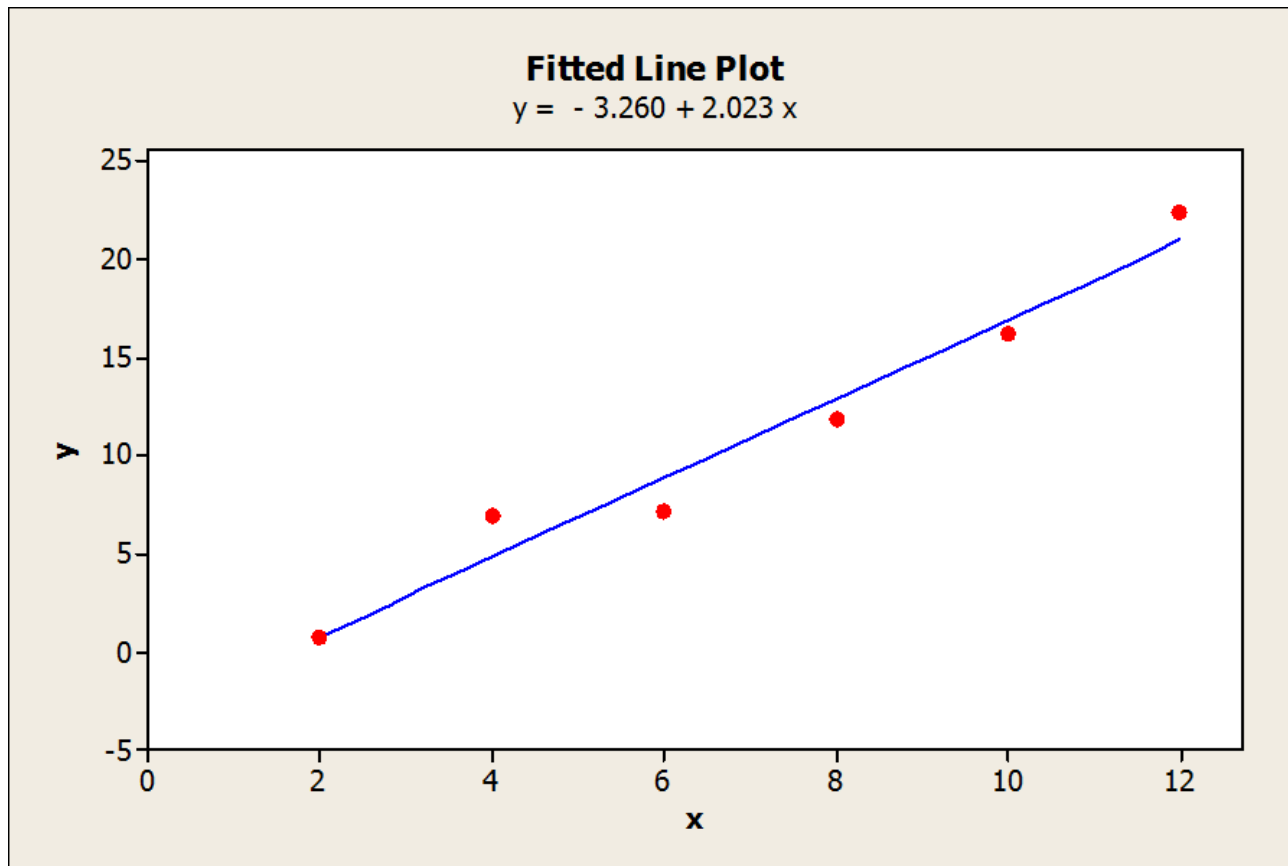
Time	Loss
x	y
2	0.7
4	6.9
6	7.2
8	11.9
10	16.2
12	22.5

xy	x^2
1.4	4
27.6	16
43.2	36
95.2	64
162	100
270	144

$\bar{x},$	$\bar{y}$	$\sum xy$	$\sum x^2$
7	10.9	599.4	364

The intercept  $\hat{\beta}_0$  is

- A. -.50      B. +1.85  
C. 0          D. -3.26



$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$y = -3.26 + 2.02x$$

# Predicting Response

Set up problem with

Independent (predictor) Variable –  $x$

Dependent (response) Variable –  $y$

Use least square equation to predict the value of  $y$  for given value of  $x$ :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

# Predicting Response

In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured.

Which of the following statements is NOT true?

- A. The amount of loss is dependent on the time so loss is the response variable.
- B. With the resulting least squares line, we can estimate the loss given the immersion time.
- C. Time is what we can control so it is the independent variable.
- D. The amount of hydrochloric acid will predict the loss so it is the predictor variable.
- E. All the above statements are true.



In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured. The data collected is in the box below.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = -3.26 + 2.02x$$

Time	Loss
x	y
2	0.7
4	6.9
6	7.2
8	11.9
10	16.2
12	22.5

$\bar{x},$        $\bar{y}$   
7      10.9

Based on the least squares equation, what would you predict the loss to be at 3 hours?

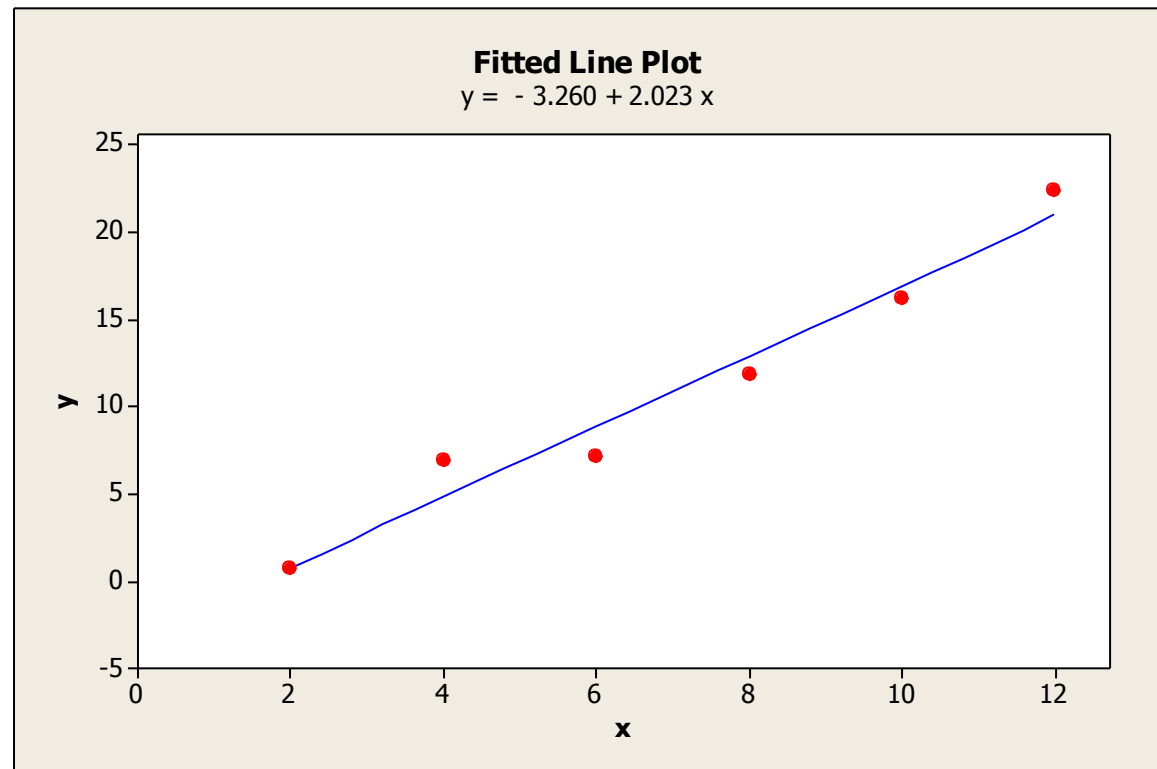
- a) 9.32      b) 4.83      c) 3.24      d) 2.81

What would you expect the loss to be at 6 hours?

- a) 8.86      b) 7.23      c) 9.64      d) 8.38

In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured. The data collected is in the box below.

Time	Loss
x	y
2	0.7
4	6.9
6	7.2
8	11.9
10	16.2
12	22.5



What is the estimate for y when x=1?

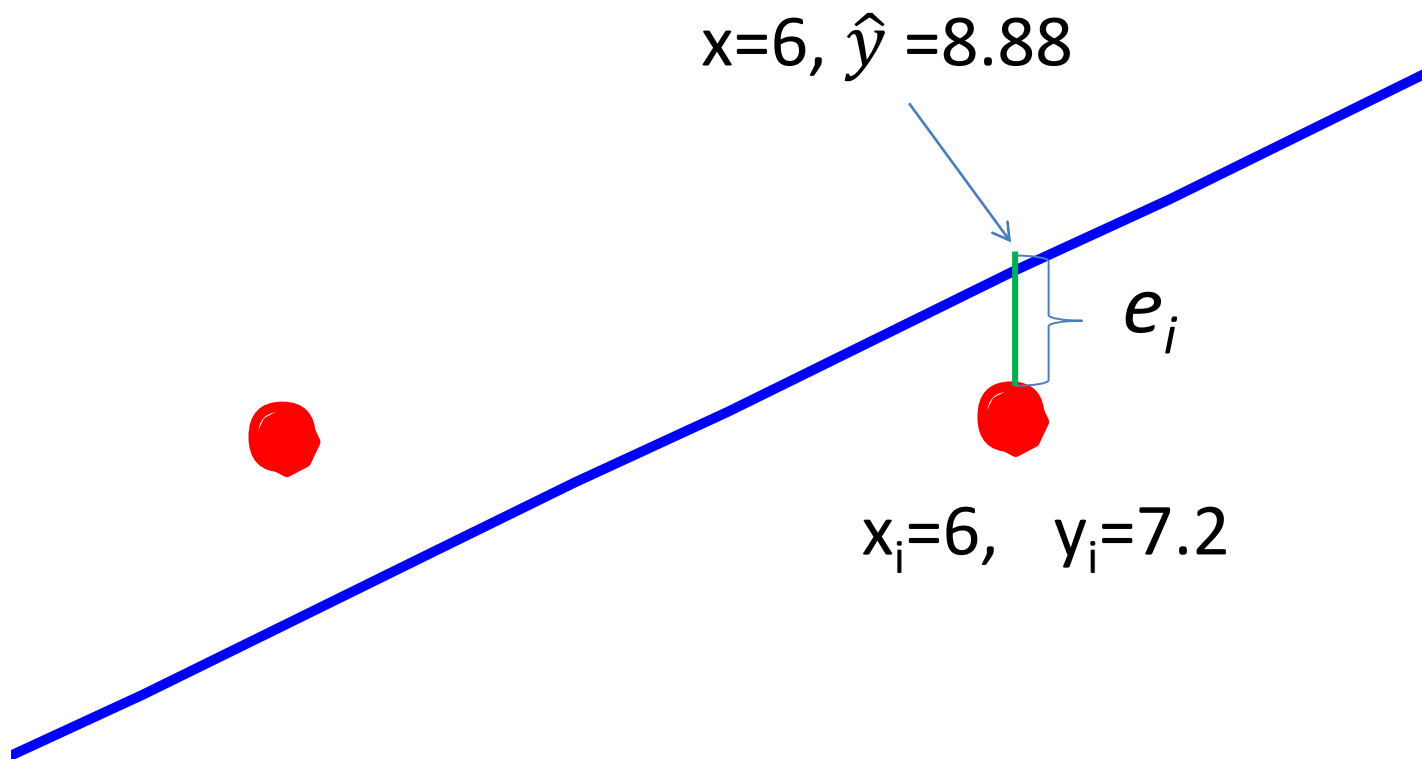
A. -1.237

B. 1.24

C. Cannot be determined.

# Least Squares Line Limitations

- Can not be considered valid outside the range of data
- Use for linear relationships only.
- Use scatterplot to look for outliers that may be an influential point.



$$\begin{aligned}\text{Residual} = \text{Error} &= e_i = y_i - \hat{y} \\ &= 7.2 - 8.88 = -1.68\end{aligned}$$

# Determination of residuals

	Observed Values	Fitted values	Residual (error values)
$x_i$	$y_i$	$\hat{y}$	$e_i = y_i - \hat{y}$
2	0.7	0.786	-0.086
4	6.9	4.831	2.069
6	7.2	8.877	-1.677
8	11.9	12.923	-1.023
10	16.2	16.969	-0.769
12	22.5	21.014	1.486

Which of the statements below is true of the sum of the residuals?

- A. The sum is positive because there is a positive correlation between x and y.
- B. The sum is negative because there is a negative correlation between x and y.
- C. The sum is zero because that the line that would fit best would minimize the sum of the residuals.

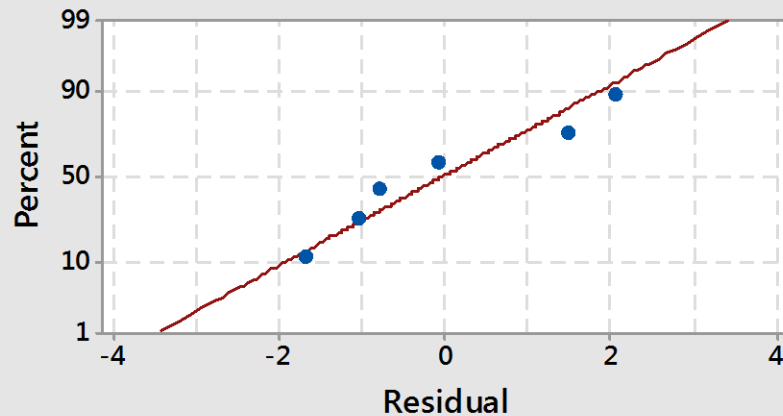
# Assumptions for Errors in Linear Models

In the simplest situation, the following assumptions are satisfied:

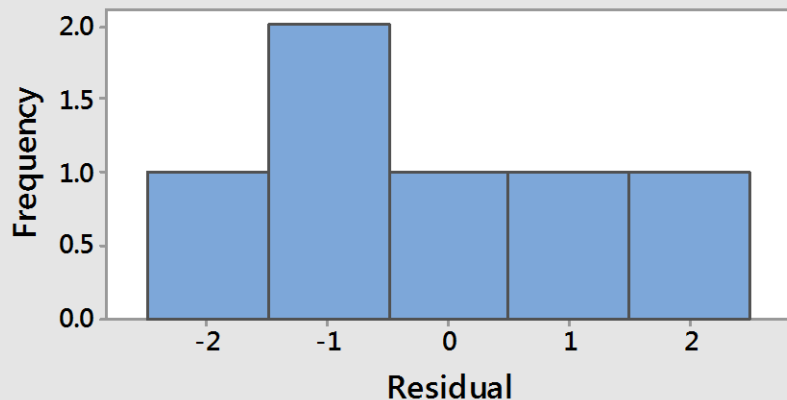
1. The errors  $\varepsilon_1, \dots, \varepsilon_n$  are random and independent.
2. The errors  $\varepsilon_1, \dots, \varepsilon_n$  all have mean 0.
3. The errors  $\varepsilon_1, \dots, \varepsilon_n$  are normally distributed.

# Residual Analysis Techniques

Normal Probability Plot



Histogram



## Check for Normality

- Normal probability Plot
- Histogram

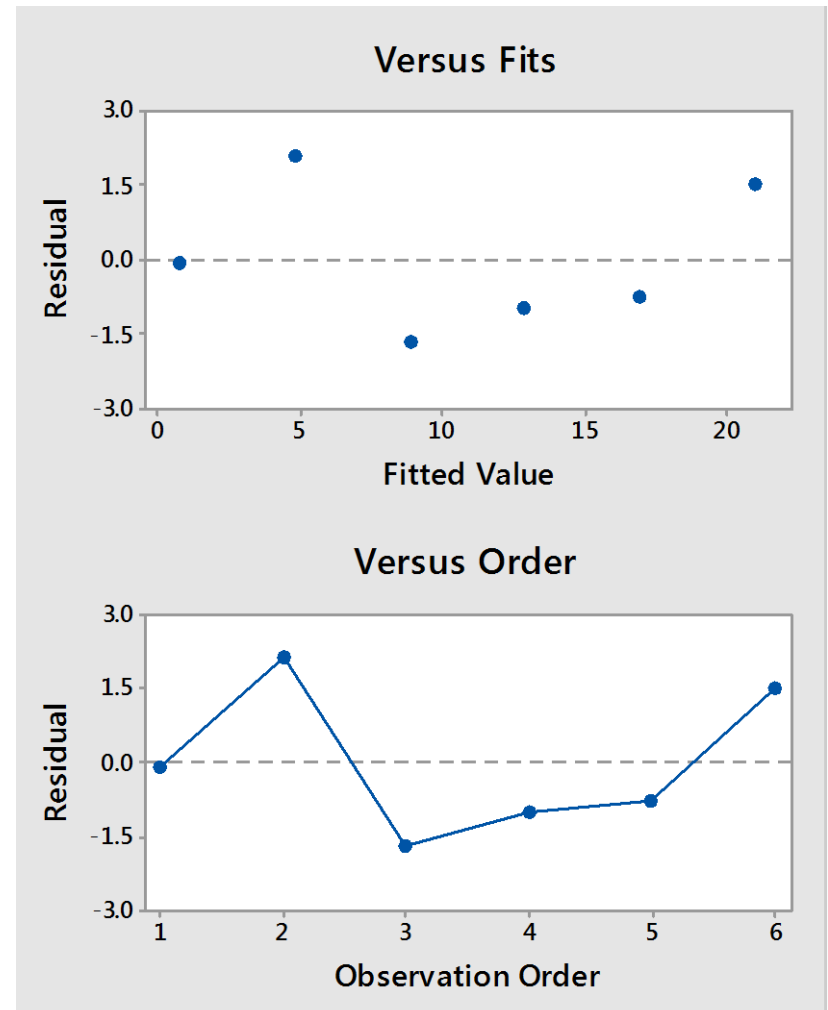
# Residual Analysis Techniques

## Residuals vs Fitted values

Verify correctness of model over the range of fitted values.

## Residuals vs Observation order

Verify “stability” of process over time.



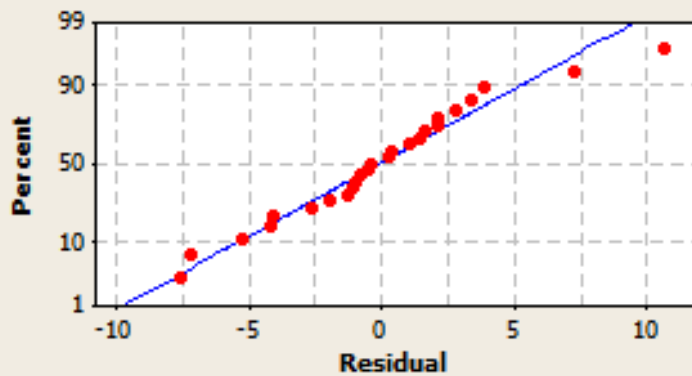




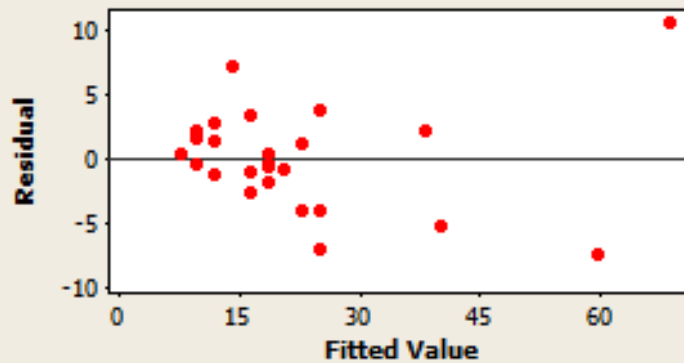
# 23.9 Analysis of Residuals: Fitted Values

Residual Plots for Delivery Time (y)

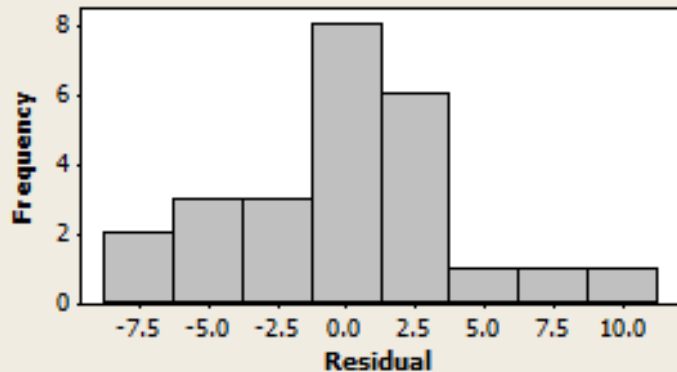
Normal Probability Plot



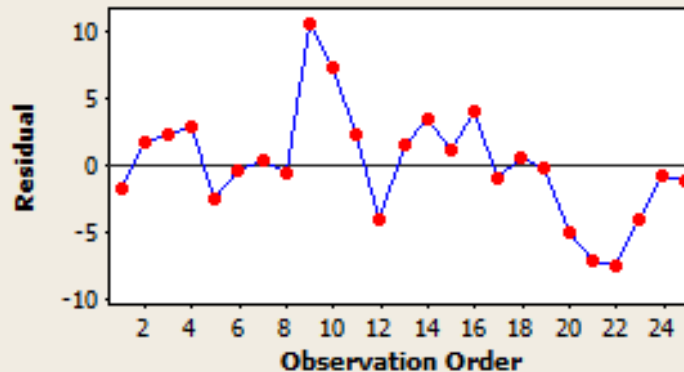
Versus Fits



Histogram



Versus Order



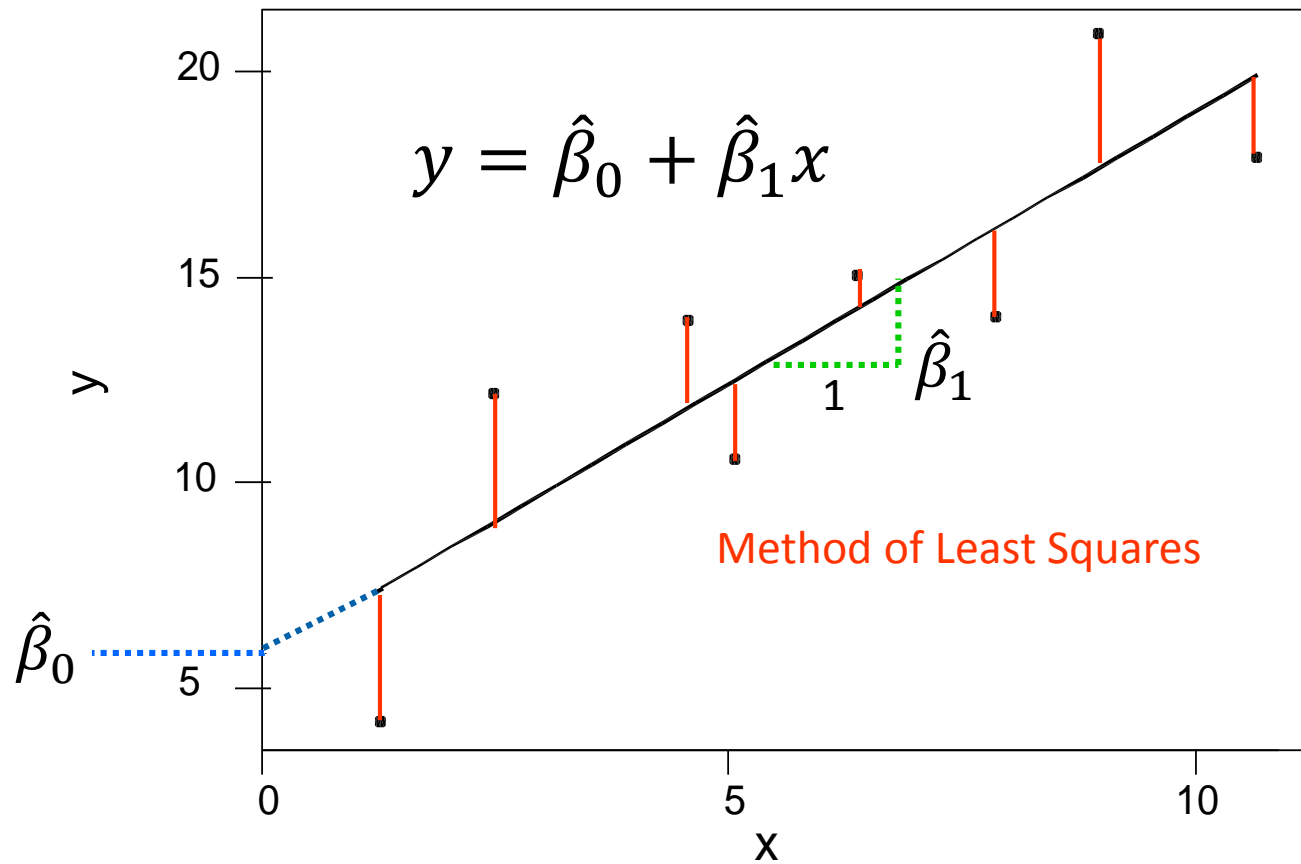
Minitab:  
Stat  
Regression  
Regression  
Graph  
Four in 1



# Vocabulary Review

- The **linear model** is  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- The **dependent variable** is  $y_i$
- The **independent variable** is  $x_i$
- The **regression coefficients** are  $\beta_0$  and  $\beta_1$
- The **error** is  $\varepsilon_i$
- The line  $y = \beta_0 + \beta_1 x$  is the **true regression line**.
- The quantities  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are called the **least squares coefficients** and can be computed.

# Least Squares Line



# Error Sum of Squares

*(How far were we off?)*

$x_i$	$y_i$	$\hat{y}$	$e_i = y_i - \hat{y}$	$e^2$
2	0.7	0.786	-0.086	0.007
4	6.9	4.831	2.069	4.279
6	7.2	8.877	-1.677	2.813
8	11.9	12.923	-1.023	1.046
10	16.2	16.969	-0.769	0.591
12	22.5	21.014	1.486	<u>2.207</u>

10.9

10.943 Error Sum of Squares

$$\sum_{i=1}^n (y_i - \hat{y})^2$$

# Total Sum of Squares

*How much variation do we have total for y?*

$x_i$	$y_i$	$\hat{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	0.7	0.786	-10.200	104.040
4	6.9	4.831	-4.000	16.000
6	7.2	8.877	-3.700	13.690
8	11.9	12.923	1.000	1.000
10	16.2	16.969	5.300	28.090
12	22.5	21.014	11.600	<u>134.560</u>
	$\bar{y}$			
	10.9			297.38

Total Sum of Squares

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

# Regression Sum of Squares

*How much variation was explained by the regression model (least squares line)?*

$$\frac{\text{Regression Sum of Squares} + \text{Error Sum of Squares}}{\text{Total Sum of Squares}}$$

Regression Sum of Squares

$$= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2$$

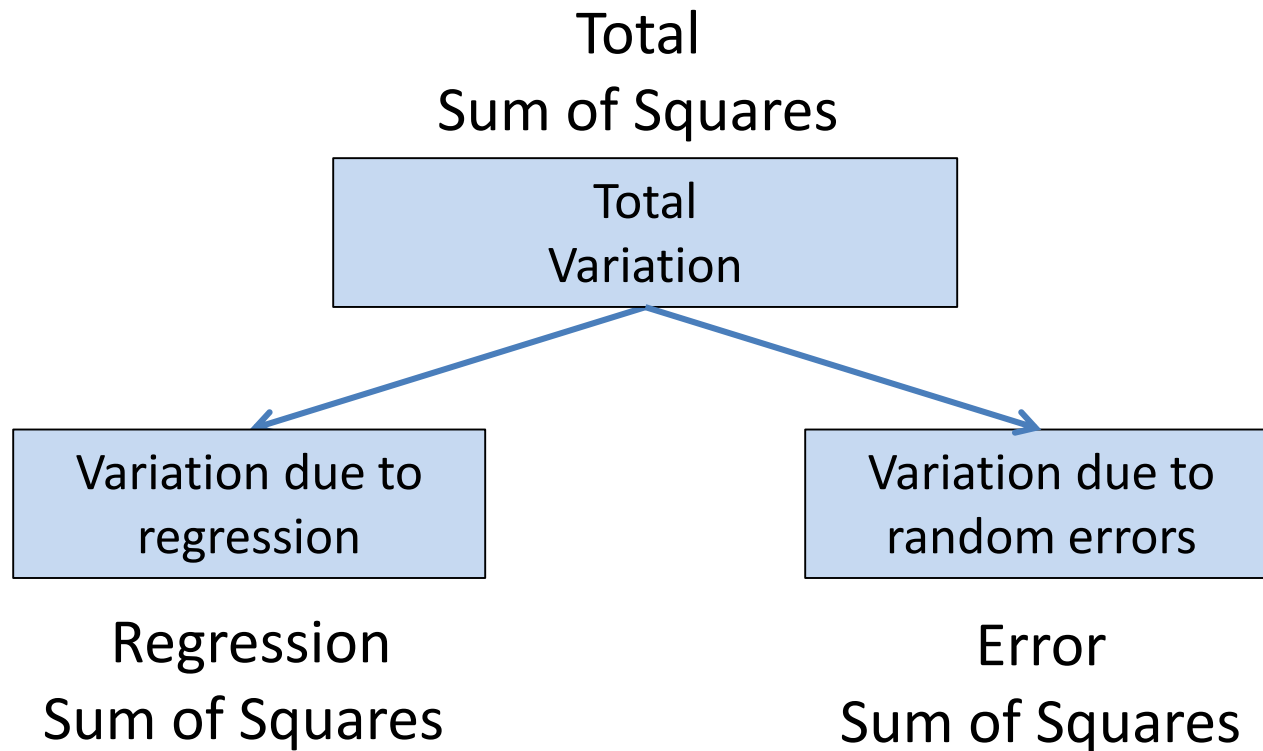
# Regression Sum of Squares

*How much variation was explained by the regression model (least squares line)?*

## Regression Sum of Squares

$$= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2$$

$$\begin{array}{ccccc} = & 297.380 & - & 10.943 & = 286.437 \\ & \text{Total} & & \text{Error} & \\ & \text{Sum of Squares} & & \text{Sum of Squares} & \text{Regression} \\ & & & & \text{Sum of Squares} \end{array}$$





# Goodness of Fit

Coefficient of Determination =  $r^2$

$$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

$$r^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$r^2$  = proportion variance in y explained by regression

# Goodness of Fit

## Correlation Coefficient

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

## Coefficient of Determination = $r^2$

$$r^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

In an experiment to determine corrosion rates, steel plates were placed in a solution of 10% hydrochloric acid for various periods of time (in hours) and the weight loss (in mg) was measured. The data collected is in the box below.

$$y = -3.26 + 2.02x$$

Time	Loss
x	y
2	0.7
4	6.9
6	7.2
8	11.9
10	16.2
12	22.5

Regression Sum of Squares = 286.437

Total Sum of Squares = 297.380

Coefficient Determination =  $r^2$

$$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}} = \frac{286.437}{297.380} = 0.963$$

Correlation Coefficient:  $r = 0.98$

# Degrees of Freedom

Degrees of Freedom (df or  $\nu$ ) =  $n-1$

- Is equal to the number of observation values minus the number of parameters estimated in calculating the statistic in question.
- For the sample variance, the 1 is subtracted because we are using  $\bar{x}$ , which was calculated from the sample data.

$$\begin{array}{l} \text{Sample} \\ \text{Variance} \end{array} \quad s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

# Total Sum of Squares

$x_i$	$y_i$	$\hat{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	0.7	0.786	-10.200	104.040
4	6.9	4.831	-4.000	16.000
6	7.2	8.877	-3.700	13.690
8	11.9	12.923	1.000	1.000
10	16.2	16.969	5.300	28.090
12	22.5	21.014	11.600	<u>134.560</u>
				297.38
				Total Sum of Squares

$\bar{y}$   
10.9

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

6 observation values  
 – 1 calculated value  
5 degrees of freedom

# Error Sum of Squares

$x_i$	$y_i$	$\hat{y}$	$e_i = y_i - \hat{y}$	$e^2$
2	0.7	0.786	-0.086	0.007
4	6.9	4.831	2.069	4.279
6	7.2	8.877	-1.677	2.813
8	11.9	12.923	-1.023	1.046
10	16.2	16.969	-0.769	0.591
12	22.5	21.014	1.486	<u>2.207</u>
				10.943 Error Sum of Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \hat{y} = -3.26 + 2.02x$$

$$\sum_{i=1}^n (y_i - \hat{y})^2$$

6 observation values  
– 2 calculated values  
 4 degrees of freedom

# Regression Sum of Squares

Regression Sum of Squares  
+ Error Sum of Squares

---

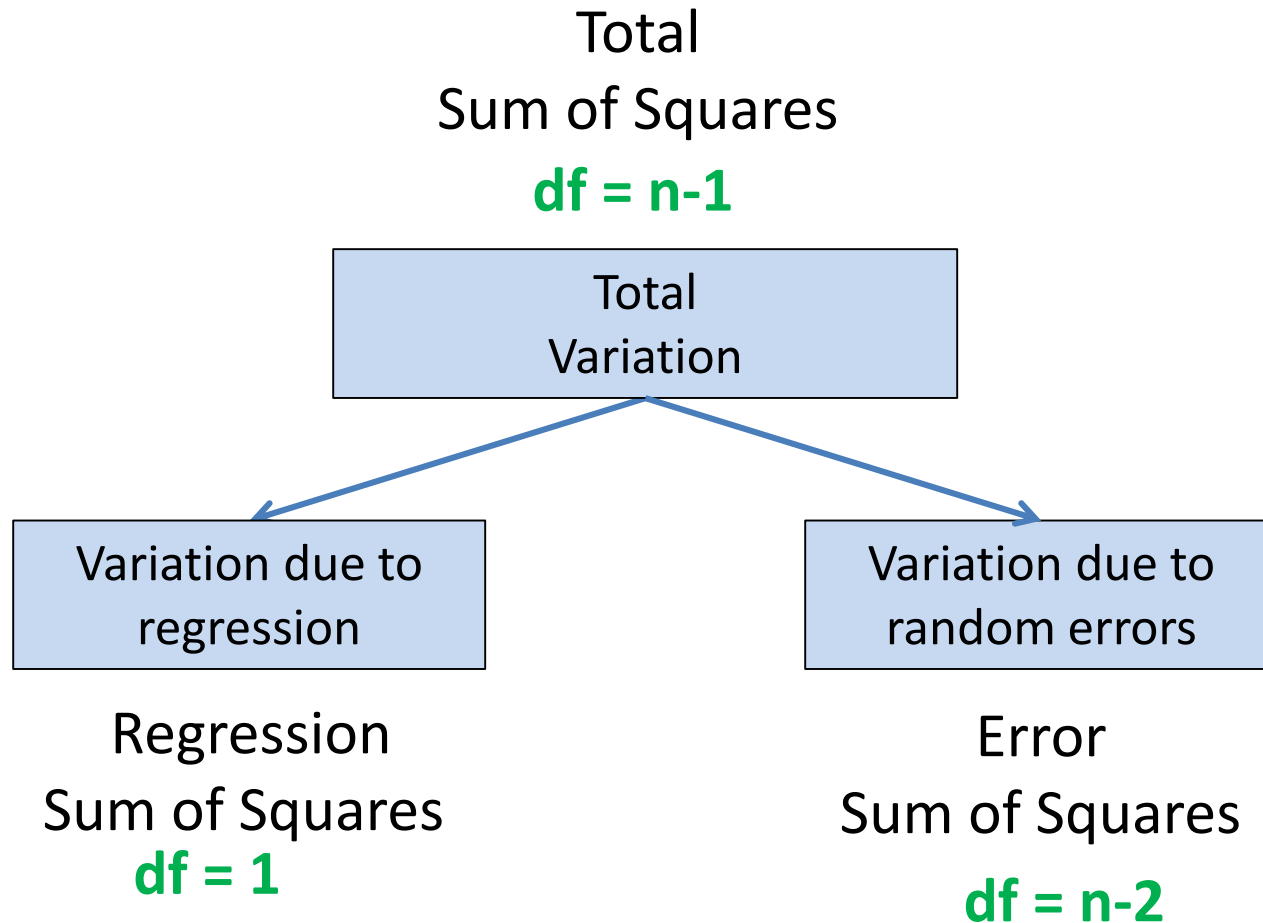
Total Sum of Squares

Regression Sum of Squares

$$= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2$$

$$DF_{\text{Regression}} = DF_{\text{Total}} - DF_{\text{Error}} = 1$$

# Degrees of Freedom for Simple Linear Regression





# How confident are we in the model?

## Analysis of Variance – simple linear model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0$
Regression	$SS_{regression}$	1	$MS_{regression}$	$F_0 = \frac{MS_{regression}}{MS_{error}}$
Error	$SS_{error}$	$n - 2$	$MS_{error}$	
Total	$SS_{total}$	$n - 1$		

How confident are we that the model explains a significant portion of the variation?

# Mean Square

The regression equation is


$$y = - 3.26 + 2.02 x$$

**S = 1.65404** R-Sq = 96.3% R-Sq(adj) = 95.4%

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	286.44	286.44	104.70	0.001
Residual Error	<b>4</b>	<b>10.94</b>	<b>2.74</b>		
Total	5	297.38			

Errors (residuals)


$$SS_{error} = \sum_{i=1}^n (y_i - \hat{y})^2 \quad df = n - 2$$

$$MSE = \frac{SS_{error}}{df} = \frac{SS_{error}}{n-2} = \frac{10.94}{4} = 2.74$$

$$2.74 = s^2 = 1.645^2$$

# Mean Square

The regression equation is

$$y = - 3.26 + 2.02 x$$

S = 1.65404 R-Sq = 96.3% R-Sq(adj) = 95.4%

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	286.44	286.44	104.70	0.001
Residual Error	4	10.94	2.74		
Total	5	297.38			

$$\begin{aligned} MSR &= \frac{SS_{\text{regression}}}{df} = \frac{SS_{\text{regression}}}{1} \\ &= \frac{286.44}{1} = 286.44 \end{aligned}$$

# The F Test

- The test statistic is  $F_{test} = \frac{MS_{regression}}{MS_{error}}$
- If there is no difference in the variation due to the regression and the variation due to the error, the two estimates are presumed to be similar and the ratio will be close to 1
- Calculating the  $F$ -test statistic tests the null hypothesis that there is no difference because of the regressor.

# The F Test

- The test statistic is  $F_{test} = \frac{MS_{regression}}{MS_{error}}$
- If there is a difference, we suspect that the regressor (x) causes the observed difference.
- Using an  $F$  table, we should reject the null hypothesis and conclude that the regressor causes a difference, at the significance level of  $\alpha$ , if

$$F_{test} > F_{\alpha,1,n-2}$$

The regression equation is

$$y = -3.26 + 2.02x$$

S = 1.65404 R-Sq = 96.3% R-Sq(adj) = 95.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	286.44	286.44	104.70	0.001
Residual Error	4	10.94	2.74		
Total	5	297.38			

$$F_{test} = \frac{MS_{regression}}{MS_{error}}$$

$$= \frac{286.44}{10.94} = 104.70$$

If acceptable risk,  $\alpha$ , equals 5%

$$F_{critical} = F_{\alpha,1,n-2} = F_{.05,1,4} = 7.71$$

$F_{test} > F_{\alpha,1,n-2}$  therefore reject null hypothesis  
that x does not cause a change in y.

Probability		v1	
Point	v2	1	2
0.100	1	39.86	49.50
0.050		161.45	199.50
0.010		4052.18	4999.50
0.100	2	8.53	9.00
0.050		18.51	19.00
0.010		98.50	99.00
0.100	3	5.54	5.46
0.050		10.13	9.55
0.010		34.12	30.82
0.100	4	4.54	4.32
0.050		7.71	6.94
0.010		21.20	18.00

# Practice

The regression equation is

$$Y = 69.87 + 0.6077 x$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	208.816	208.816		
Error	18	347.734			
Total	19	556.550			

Given the above Minitab output, if my x value was 75, I would predict y to be

- A. 117.3      B. 115.4      C. 122.5      D. 130.1

# Practice

The regression equation is

$$Y = 69.87 + 0.6077 x$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	208.816	208.816		
Error	18	347.734			
Total	19	556.550			

The total number of observations (pairs of x-y values) collected was

- A. 18                      B. 19                      C. 20                      D. 21



# Practice

The regression equation is

$$Y = 69.87 + 0.6077 x$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	208.816	208.816		
Error	18	347.734			
Total	19	556.550			

$MS_{\text{Error}}$  is equal to

- A. 208.82      B. 19.32      C. 17.39      D. 18.30

$F_{\text{test}}$  is equal to

- A. 10.81      B. 0.0925      C. 0.375      D. 2.67

# Practice

The regression equation is

$$Y = 69.87 + 0.6077 x$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	208.816	208.816	10.81	0.004
Error	18	347.734	19.319		
Total	19	556.550			

The accepted level of risk,  $\alpha$ , is equal to 5%. The  $F_{\text{critical}}$  for comparison will then be

- A.  $F_{.025,1,18}$       B.  $F_{.05,18,1}$       C.  $F_{.05,1,18}$       D.  $F_{.025,1,18}$

$F_{\text{critical}}$  is equal to

- A. 8.99      B. 3.01      C. 4.41      D. 246

# Practice

The regression equation is

$$Y = 69.87 + 0.6077 x$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	208.816	208.816	10.81	0.004
Error	18	347.734	19.319		
Total	19	556.550			

From the Analysis of Variance, it can be concluded that

- A. Since  $F_{test} > F_{critical}$  the null hypothesis that a change  $x$  does not cause a change in  $y$  is rejected.
- B. With 95% confidence, it can be said that the regression model explains a portion of the variation.
- C. A and B

# Practice

$$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

The regression equation is  
 $Y = 69.87 + 0.6077 x$

## Analysis of Variance

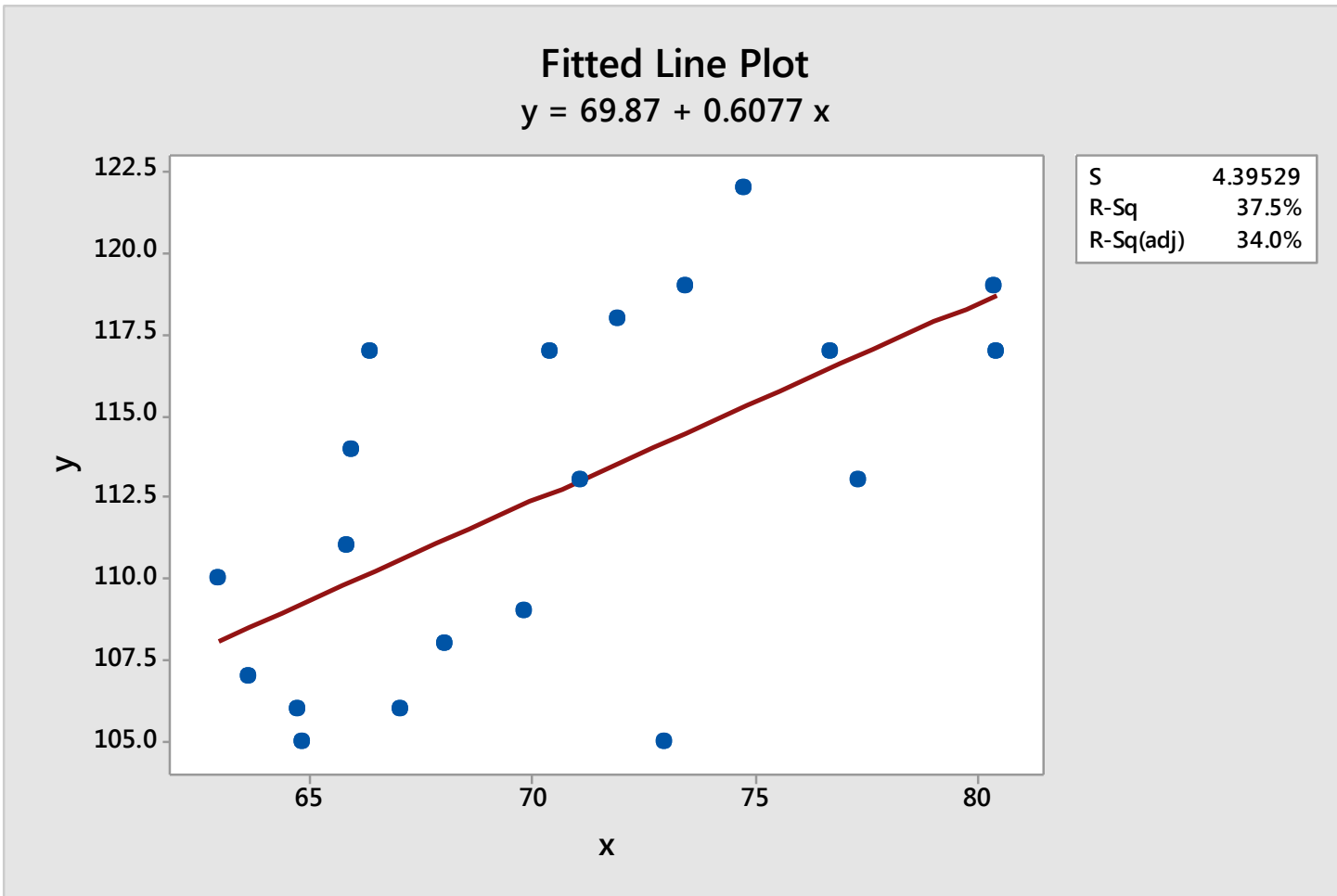
Source	DF	SS	MS	F	P
Regression	1	208.816	208.816	10.81	0.004
Error	18	347.734	19.319		
Total	19	556.550			

What is the coefficient of determination,  $r^2$ ?

- A. 0.375      B. 0.624      C. 2.80      D. .601

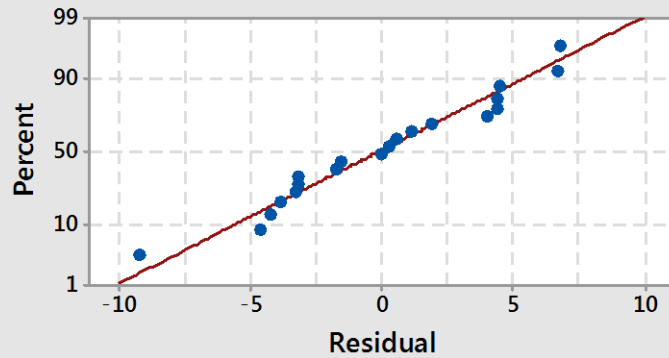
What percent of the variation in the data is explained by the regression model?

- A. 37.5%      B. 62.4%      C. 80.3%      D. 60.1%

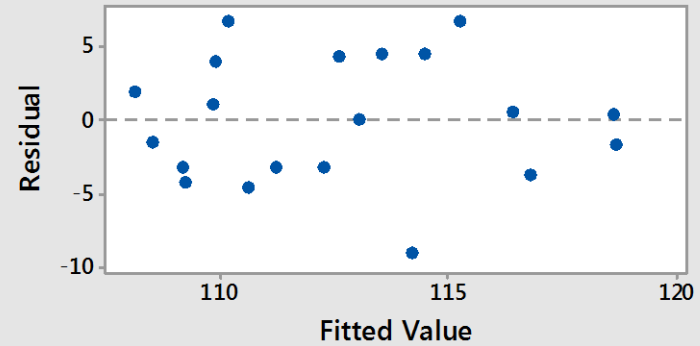


## Residual Plots

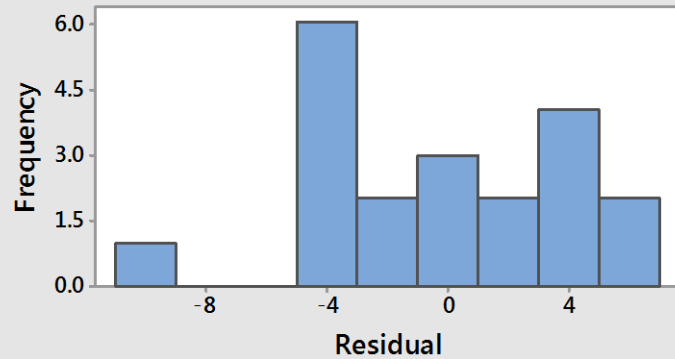
Normal Probability Plot



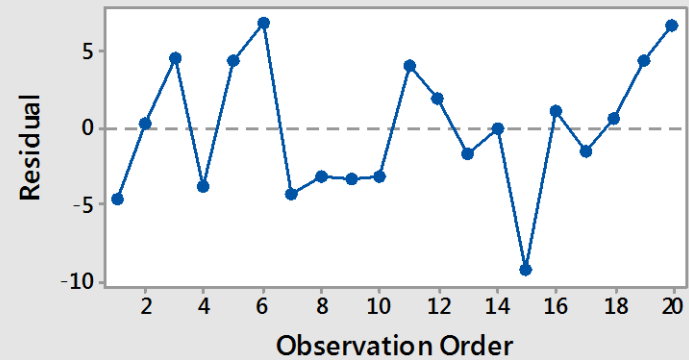
Versus Fits



Histogram



Versus Order



# Regression Reminders

- The regression model describes the region for which it models and should not be an accurate representation for extrapolated values..
- A true cause-and-effect relationship does not necessarily exist when two variables are correlated.
- Least-squares predictions are based on history data, which may not represent future relationships.
- This lecture only covered the simplest case – simple linear regression.



# Related Assignments

Please see Blackboard for related assignments