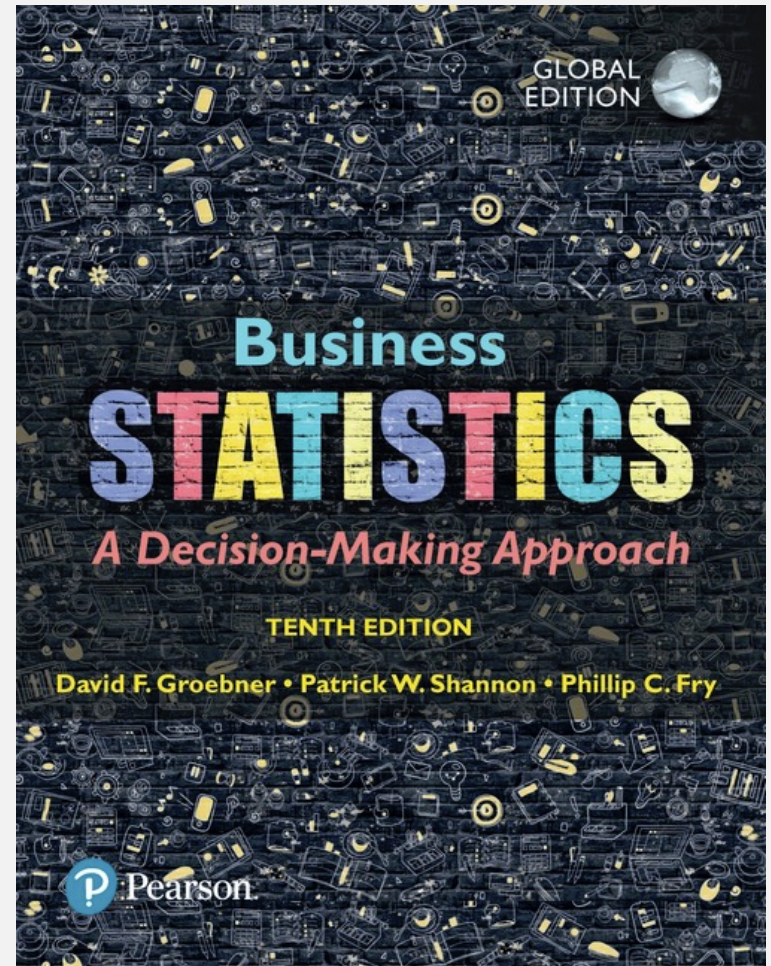


Chapter 14

Introduction to Linear Regression and Correlation Analysis



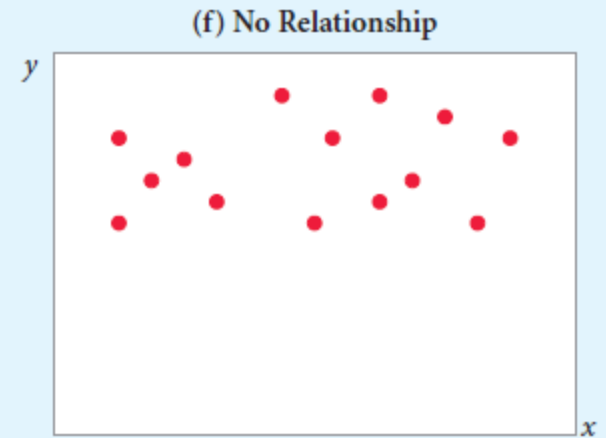
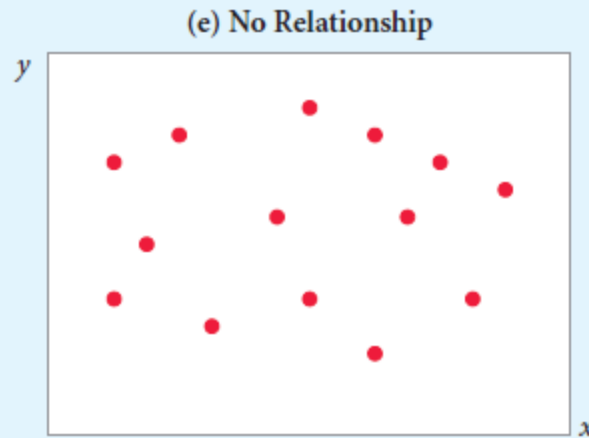
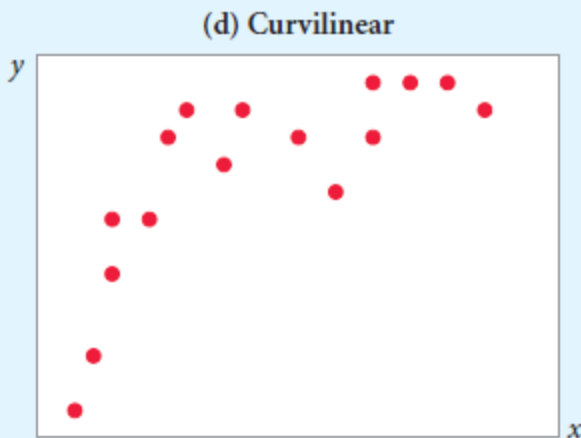
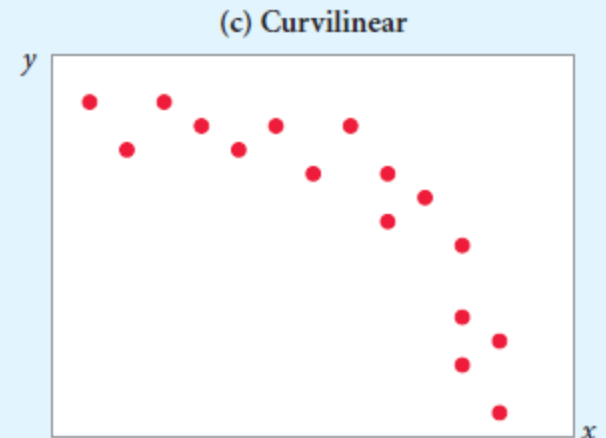
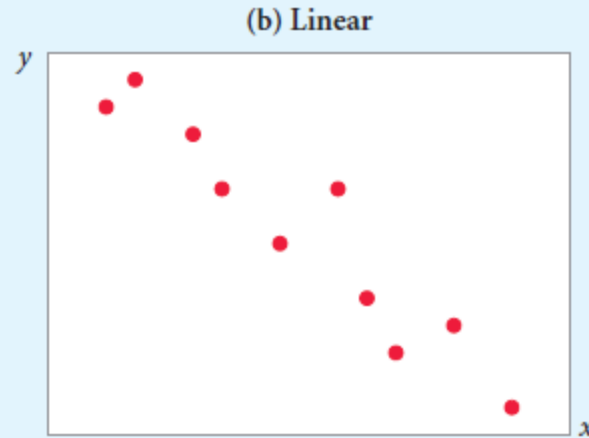
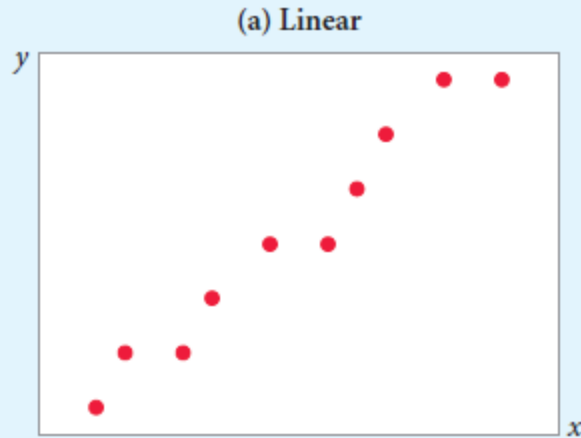
Learning Outcomes

- Outcome 1.** Calculate and interpret the correlation between two variables.
- Outcome 2.** Determine whether the correlation is significant.
- Outcome 3.** Calculate the simple linear regression equation for a set of data and know the basic assumptions behind regression analysis
- Outcome 4.** Determine whether a regression model is significant.
- Outcome 5.** Recognize regression analysis applications for purposes of description and prediction.
- Outcome 6.** Calculate and interpret confidence intervals for the regression analysis.
- Outcome 7.** Recognize some potential problems if regression analysis is used incorrectly.

14.1 Scatter Plots and Correlation

- Scatter Plot
 - A two-dimensional plot showing the values for the joint occurrence of two quantitative variables. The scatter plot may be used to graphically represent the relationship between two variables. It is also known as a scatter diagram.
- Correlation Coefficient
 - A quantitative measure of the strength of the linear relationship between two variables. The correlation ranges from -1.0 to + 1.0. A correlation of ± 1.0 indicates a perfect linear relationship, whereas a correlation of 0 indicates no linear relationship.

Two-Variable Relationships



Scatter Plot – Example Using Excel

2016

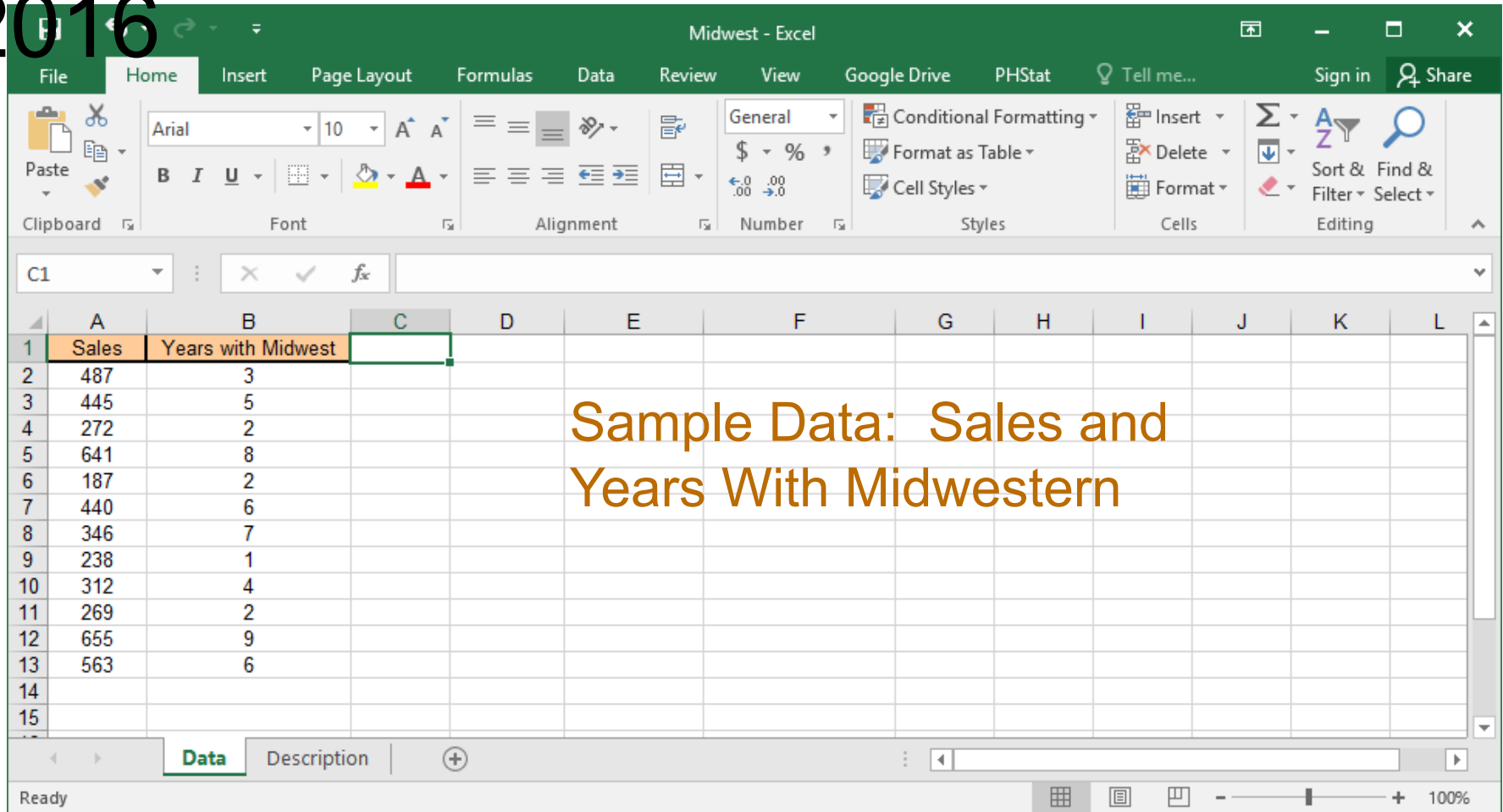
The director of marketing for Midwest Distribution Company is concerned about the rapid turnover in her sales force. In the course of exit interviews, she discovered a major concern with the compensation structure. At issue is the relationship between sales and number of years with the company. The data for a random sample of 12 sales representatives was used for analysis.

Objective: Use Excel 2016 to first create a scatter plot using the data file **Midwest.xlsx**.



Scatter Plot – Example Using Excel

2016

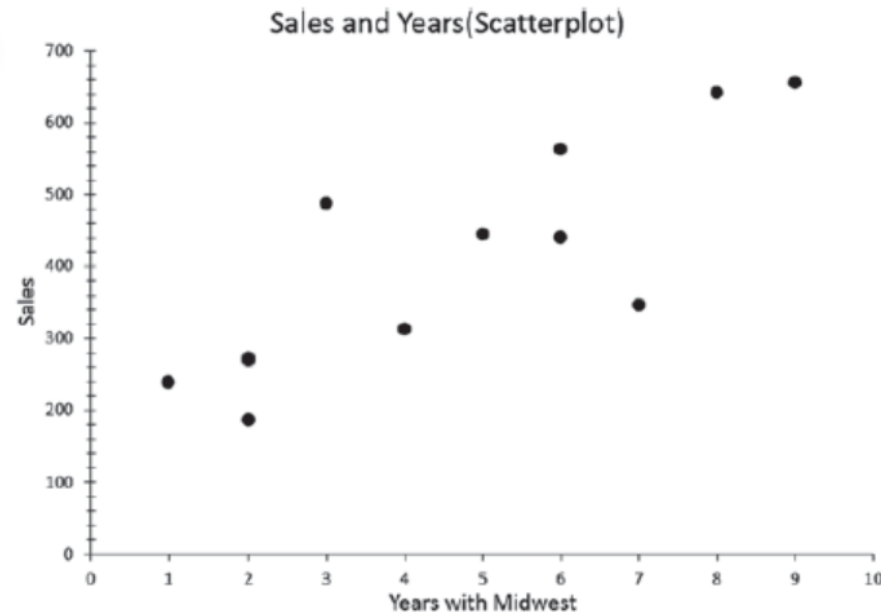


Scatter Plot – Example Using Excel 2016

Excel 2016 Instructions

1. Open file: **Midwest.xlsx**.
2. Move the *Sales* column to the right of the *Years* column.
3. Select data to be used in the chart.
4. On the **Insert** tab, click **Scatter (X, Y)** or **Bubble Chart**, and then click the **Scatter** option.
5. Use the **Design** tab of **Chart Tools** to add titles and remove the grid lines.
6. Use the **Design** tab of **Chart Tools** to move the chart to a new worksheet.

FIGURE 14.3 Excel 2016 Scatter Plot of Sales vs. Years with Midwest Distribution



The relationship between Sales and Years With Midwestern appears to be positive and linear.

The Correlation Coefficient

- Sample Correlation Coefficient:

- Algebraic

$$\frac{\sum \sum (x - \bar{x})(y - \bar{y})}{\sum \sum (x - \bar{x})^2 \sum \sum (y - \bar{y})^2}$$

$$\frac{\sum \sum (x - \bar{x})(y - \bar{y})}{\sum \sum (x - \bar{x})^2 \sum \sum (y - \bar{y})^2}$$

r - Sample correlation coefficient

n - Sample size

x - Value of the independent variable

y - Value of the dependent variable

The Correlation Coefficient

The Correlation Coefficient measures the strength of the linear relationship between two variables.

$$-1.0 \leq r \leq +1.0$$

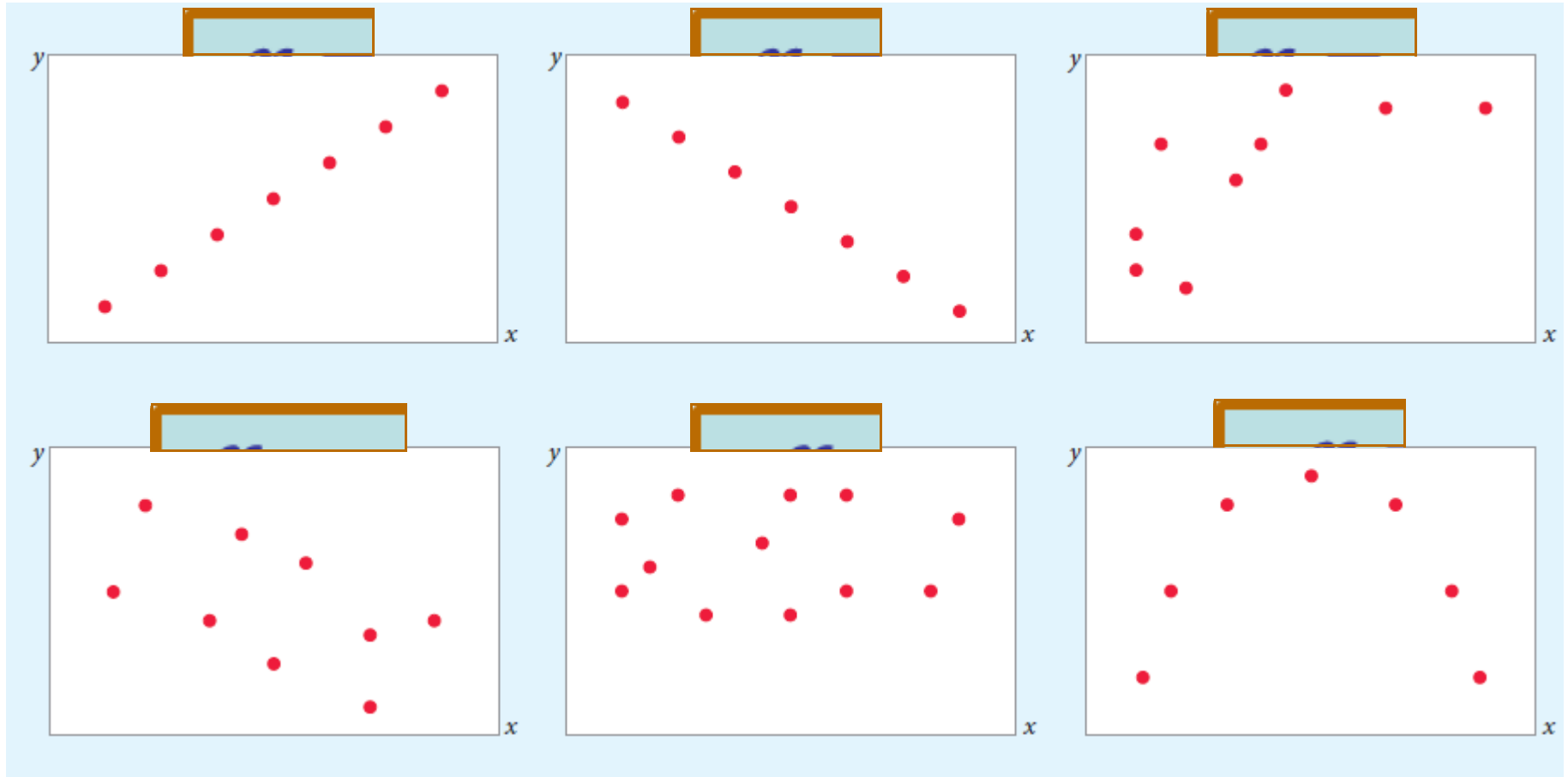
r close to 1.0 implies a strong positive linear relationship

r close to -1.0 implies a strong negative linear relationship

r close to 0.0 implies a weak linear relationship

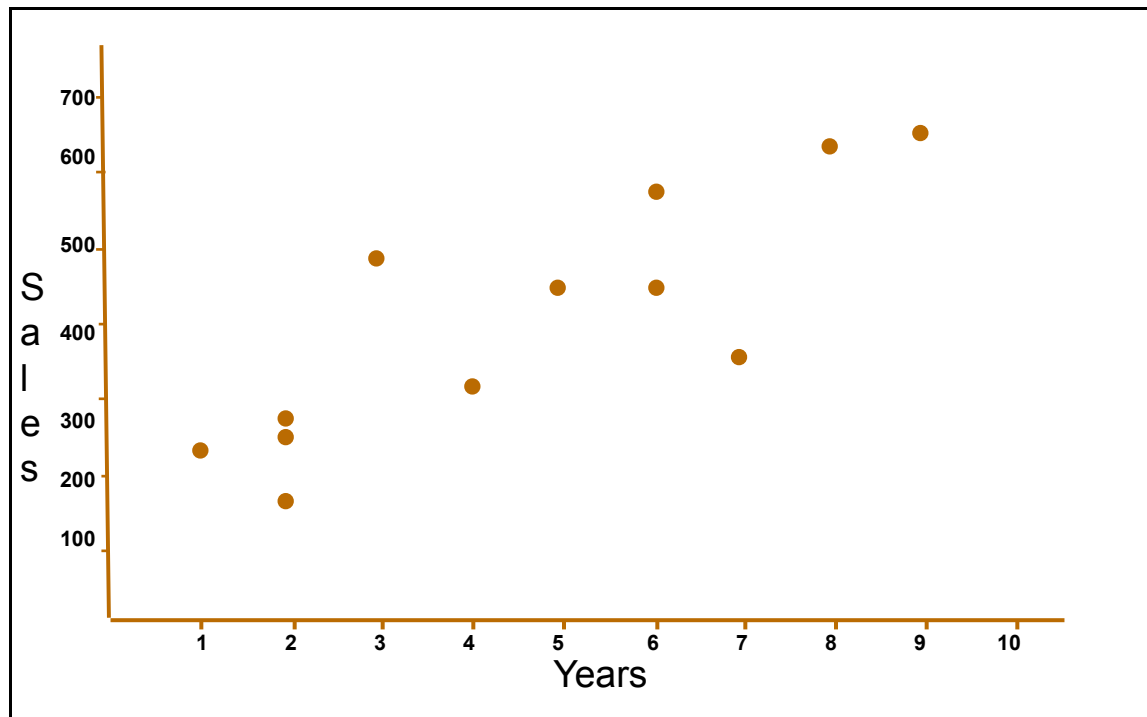


Correlation between Two Variables



The Correlation Coefficient - Example

The company is studying the relationship between sales (on which commissions are paid) and number of years a sales person is with the company. A random sample of 12 sales representatives is collected. Compute the correlation coefficient.



The Correlation Coefficient – Manual Calculation Example

Sales	Years					
y	x	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
487	3	-1.58	82.42	-130.22	2.50	6,793.06
445	5	0.42	40.42	16.98	0.18	1,633.78
272	2	-2.58	-132.58	342.06	6.66	17,577.46
641	8	3.42	236.42	808.56	11.70	55,894.42
187	2	-2.58	-217.58	561.36	6.66	47,341.06
440	6	1.42	35.42	50.30	2.02	1,254.58
346	7	2.42	-58.58	-141.76	5.86	3,431.62
238	1	-3.58	-166.58	596.36	12.82	27,748.90
312	4	-0.58	-92.58	53.70	0.34	8,571.06
269	2	-2.58	-135.58	349.80	6.66	18,381.94
655	9	4.42	250.42	1,106.86	19.54	62,710.18
563	6	1.42	158.42	224.96	2.02	25,096.90
$\Sigma = 4,855$	$\Sigma = 55$			$\Sigma = 3,838.92$	$\Sigma = 76.92$	$\Sigma = 276,434.92$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{4,855}{12} = 404.58 \quad \bar{x} = \frac{\Sigma x}{n} = \frac{55}{12} = 4.58$$

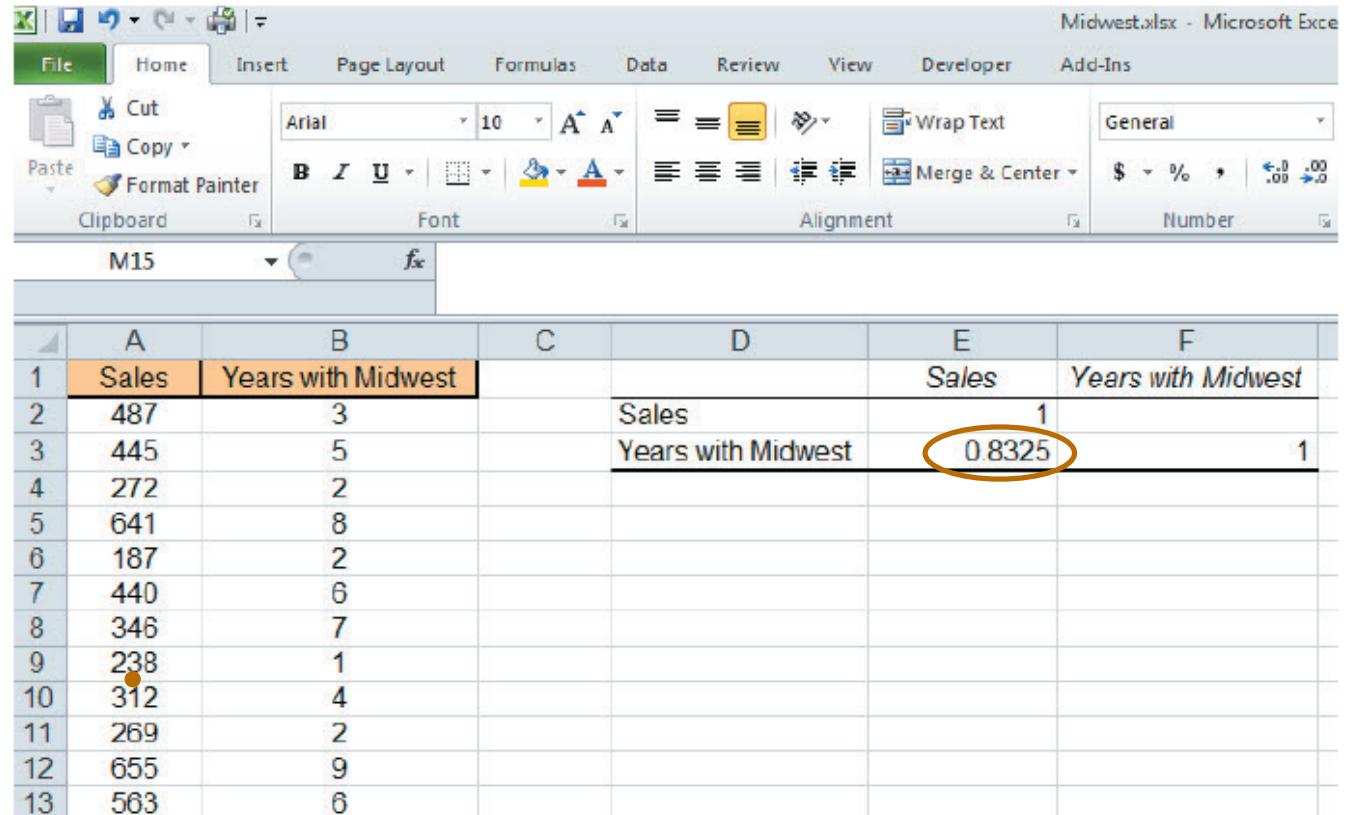
Using Equation 14.1,

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} = \frac{3,838.92}{\sqrt{(76.92)(276,434.92)}} = 0.8325$$



The Correlation Coefficient – Example Using Excel 2016

1. Open file.
2. Select **Data > Data Analysis.**
3. Select **Correlation.**
4. Define the data range.
5. Click on **Labels in First Row.**
6. Specify output choice.
7. Click **OK.**



Midwest.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Developer Add-Ins

Cut Copy Paste Format Painter Clipboard Font Alignment Number

M15

	A	B	C	D	E	F
1	Sales	Years with Midwest			Sales	Years with Midwest
2	487	3		Sales	1	
3	445	5		Years with Midwest	0.8325	1
4	272	2				
5	641	8				
6	187	2				
7	440	6				
8	346	7				
9	238	1				
10	312	4				
11	269	2				
12	655	9				
13	563	6				

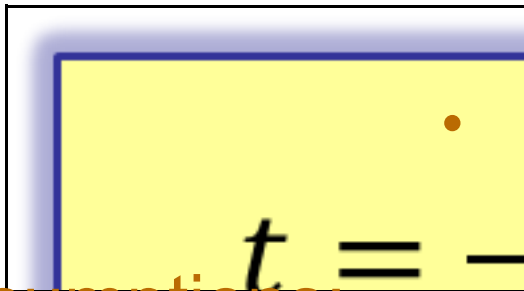
Note: Data are taken from previous example.

Significance Test for the Correlation

- The Null and Alternative Hypotheses:

$$H_0: \rho = 0 \text{ (no)}$$

- Test Statistic for Correlation:


$$t = -$$

ρ - Population corre
 t - Number of stanc

- Assumptions:

- The data are interval or ratio-level.
- The two variables (y and x) are distributed as a *bivariate normal* distribution.

Significance Test for the Correlation - Example

Midwestern Example

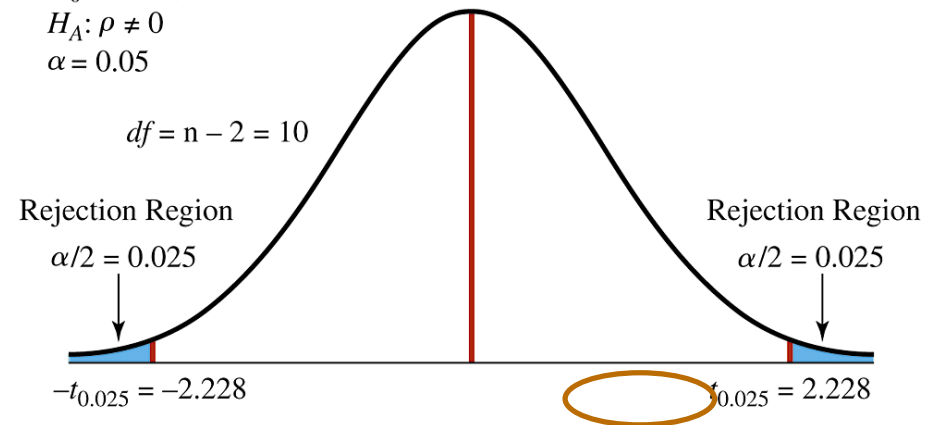
	A	B	C
1		Sales	Years with Midwest
2	Sales	1	
3	Years with Midwest	0.8325	1

Hypotheses:

$H_0: \rho = 0$ (no correlation)

$H_A: \rho \neq 0$

$\alpha = 0.05$



The calculated t -value is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.8325}{\sqrt{\frac{1-0.6931}{10}}} = 4.752$$

Decision Rule:

If $t > t_{0.025} = 2.228$, reject H_0 .

If $t < -t_{0.025} = -2.228$, reject H_0 .

Otherwise, do not reject H_0 .

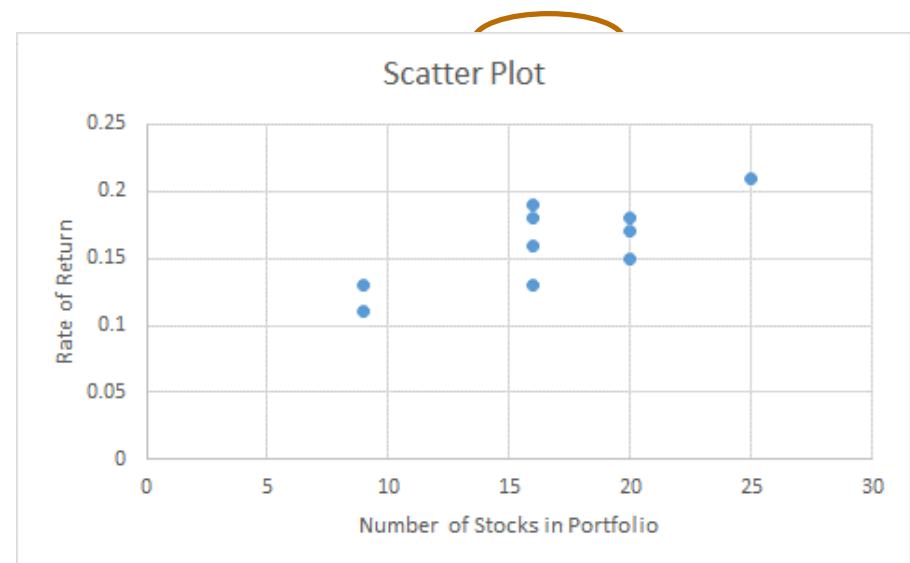
Because $4.752 > 2.228$, reject H_0 .

Based on the sample evidence, we conclude there is a significant positive linear relationship between years with the company and sales volume.

The Correlation Coefficient – Example

A money management company is interested in determining whether there is a positive linear relationship between the number of stocks in a client's portfolio and the portfolio annual rate of return. A sample of $n=10$ clients has been selected. The sample data are:

Number of Stocks	Rate of Return
9	0.13
16	0.16
25	0.21
16	0.18
20	0.18
16	0.19
20	0.15
20	0.17
16	0.13
9	0.11



The Correlation Coefficient – Example

	<i>Number of Stocks</i>	<i>Rate of Return</i>
Number of Stocks	1	
Rate of Return	0.780	1

$$r = 0.780$$

$$\begin{aligned} H_o : \rho &= 0.0 \\ H_A : \rho &\geq 0.0 \\ \alpha &= 0.05 \end{aligned} \quad t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.780}{\sqrt{\frac{1-0.780^2}{10-2}}} = 3.53$$

Since $t = 3.53 > t_{0.05, df=8} = 1.8595$ reject the null hypothesis.

Scatter Plot and Correlation Coefficient – Example Using Excel

ONLINE [Compatibility Mode] - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

PivotTable Recommended Table
PivotTables

Pictures Online Pictures
Illustrations

Shapes SmartArt Screenshot

Store My Apps Bing Maps People Graph
Apps

Recommended Charts

PivotChart Power View

Chart 3

	A	B	C	D	E	F	G	H	I	J	K
	Customer Account Number	Time (Minutes)	Purchases (\$)								
1											
2	62638	54.69	259								
3	58499	13.42	24								
4	79902	15.78	177								
5	85784	75.70	207								
6	99619	2.28	37								
7	88286	14.13	20								
8	60330	232.36	336								
9	10702	285.93	364								
10	8368	5.97	281								

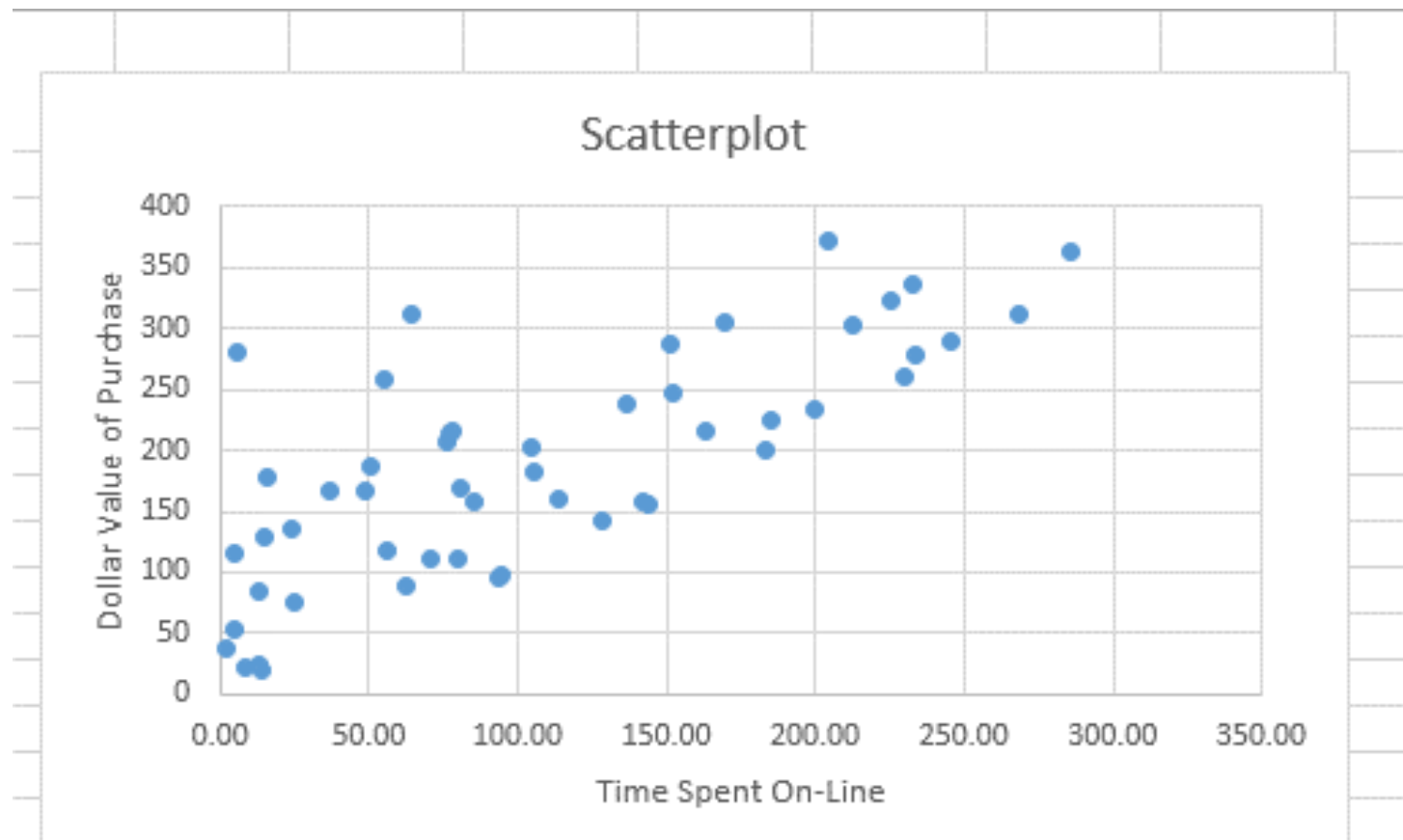
A random sample of 51 customers who made on-line purchases last quarter from an Internet retailer. The quarterly purchases (rounded to the nearest \$) for each customer and the amount of time spent viewing the retailer's catalog (in minutes) last quarter are recorded.

Scatter

Bubble

More Scatter Charts...

Scatter Plot and Correlation Coefficient – Example Using Excel



Scatter Plot and Correlation Coefficient – Example Using Excel

The screenshot shows the Microsoft Excel interface with the 'DATA' ribbon selected. The ribbon includes tabs for FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, and VIEW. The DATA tab is active, showing options like 'From Access', 'From Web', 'From Text', 'From Other Sources', 'Existing Connections', 'Refresh All', 'Connections', 'Sort', 'Filter', 'Clear', 'Reapply', 'Advanced', and 'Text to Columns'. Below the ribbon, the formula bar shows 'H18' and the function 'fx'. The worksheet displays a table of data with columns A, B, and C. A 'Correlation' dialog box is open, showing the following settings:

- Input**
 - Input Range:
 - Grouped By: ☒ Columns ☐ Rows
 - ☒ Labels in First Row
- Output options**
 - ☐ Output Range:
 - ☒ New Worksheet Ply:
 - ☐ New Workbook

Buttons for OK, Cancel, and Help are visible on the right side of the dialog box.

Using the Data Analysis Tool for calculating the correlation coefficient.

Scatter Plot and Correlation Coefficient – Example Using Excel

	A	B	C	D
1		Time (Minutes)	Purchases (\$)	
2	Time (Minutes)	1		
3	Purchases (\$)	0.756	1	
4				

$$H_o : \rho = 0.0$$

$$H_A : \rho \neq 0.0$$

$$\alpha = 0.05$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.756}{\sqrt{\frac{1-0.756^2}{51-2}}} = 8.08$$

Since $t = 8.08 > t_{0.05, df=49} = 2.0096$ we reject the null hypothesis

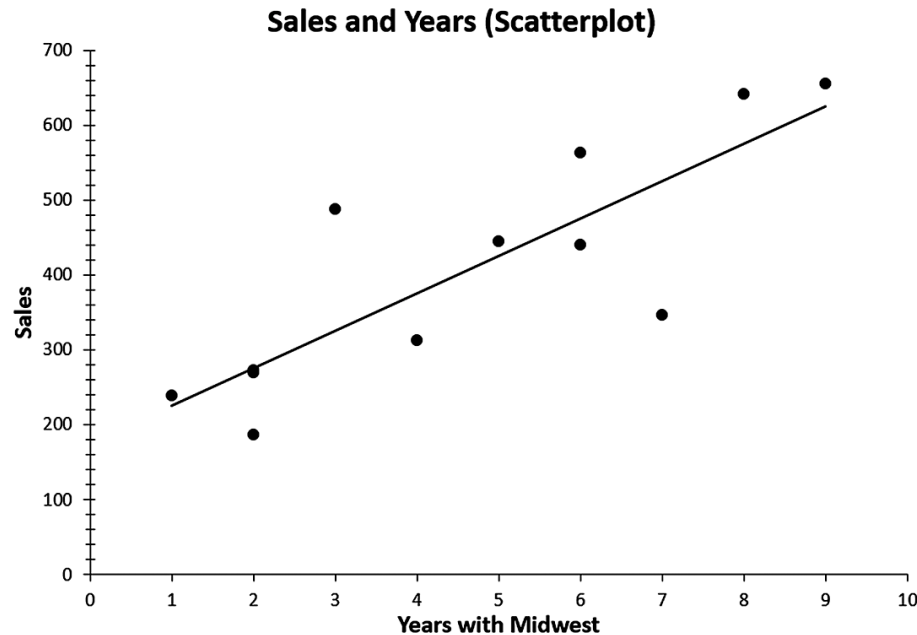
Correlation Analysis - Summary

- **Step 1:** Specify the population parameter of interest
- **Step 2:** Formulate the appropriate null and alternative hypotheses
- **Step 3:** Specify the level of significance
- **Step 4:** Compute the correlation coefficient and the test statistic
- **Step 5:** Construct the rejection region and decision rule.
- **Step 6:** Reach a decision
- **Step 7:** Draw a conclusion



14.2 Simple Linear Regression Analysis

A statistical method that is used to describe the linear relationship between two variables in the form of a straight that passes through the points on a scatterplot



Simple Linear Regression Analysis

- When there are only two variables - a dependent variable, and an independent variable, the technique is referred to as simple regression analysis
- When the relationship between the dependent variable and the independent variable is linear, the technique is simple linear regression

Dependent and Independent Variables

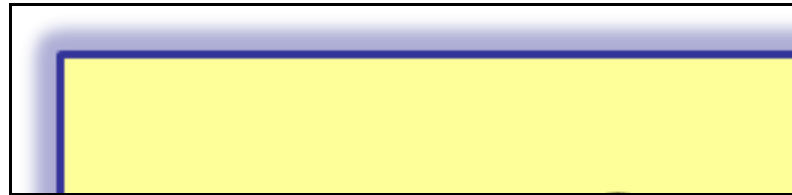
Dependent Variable – A variable whose values are thought to be a function of, or dependent on, the values of more or more other variables. This dependent variable is referred to as the y variable and is placed on the vertical axis of a scatterplot.

The Independent Variable – A variable whose values are thought to influence the values of the dependent variable. Independent variables are also called explanatory variables. The dependent variable is referred to as the x variable and is placed on the horizontal axis of a scatterplot.

The Regression Model

-

Regression Model



y - Value of the dependent variable

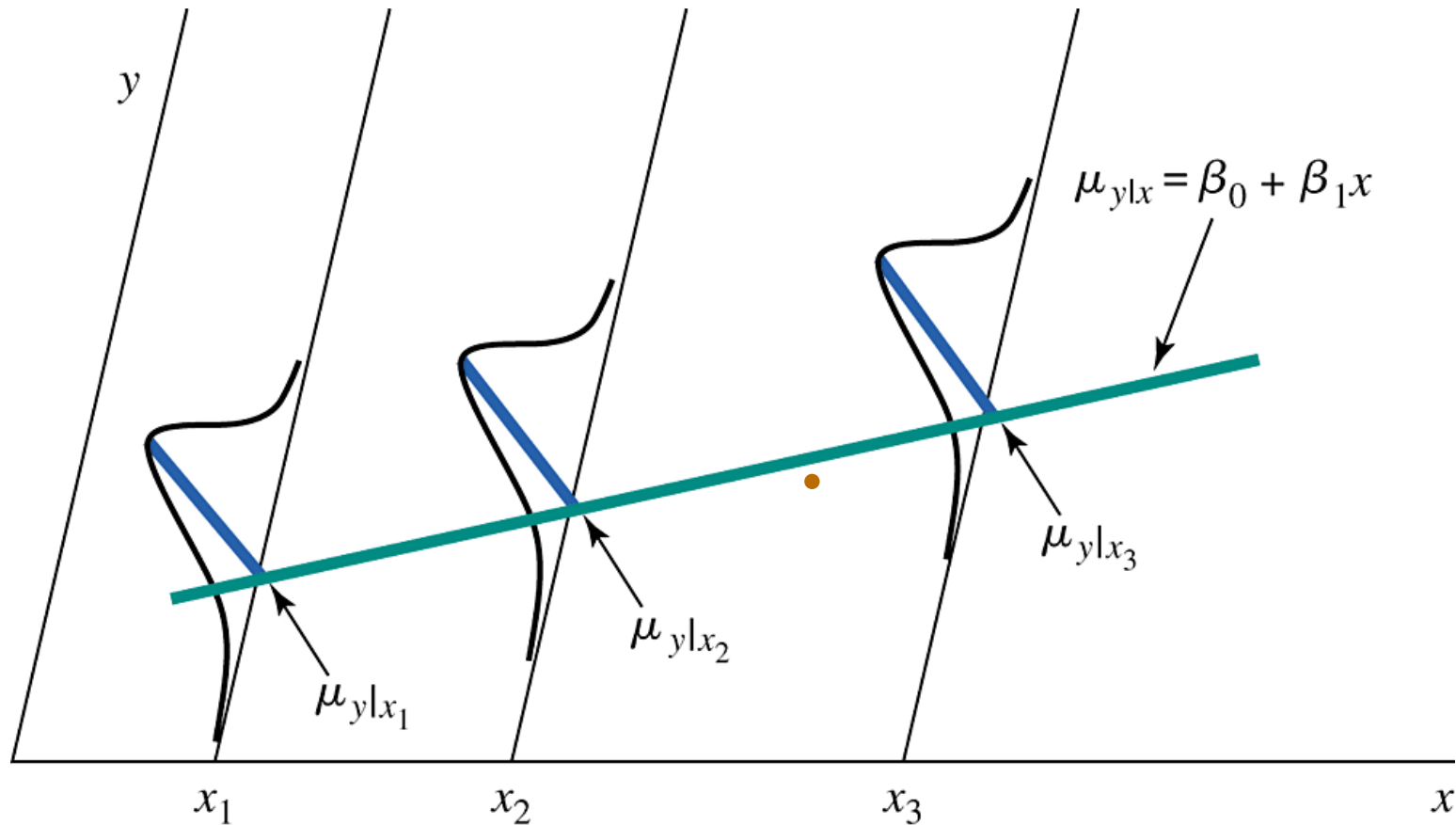
x - Value of the independent variable

$\hat{y}$ - Predicted value of the dependent variable

Linear Regression Assumptions

1. The random errors ϵ , are statistically independent
2. For each value of x there can exist many possible values of y and the distribution of y values is normally distributed.
3. The distributions of errors have equal variances for all possible levels of x
4. A straight line, called the population regression model (equation) will pass through the mean of the possible y values for all levels of x

Linear Regression Assumptions – Visual Representation



Meaning of the Regression Coefficients

- **Regression Slope (β_1)**
 - Measures the average change in the dependent variable for a one-unit change in x

Estimates of the Regression Coefficients

$$\hat{y} = b_o + b_1x$$

where :

$\hat{y}$ = estimated value of the dependent variable for a given value of x

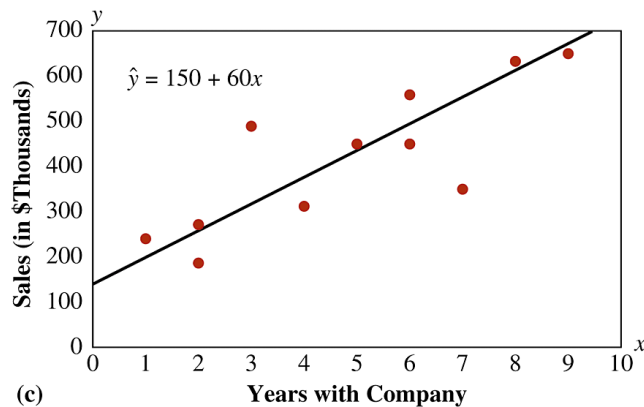
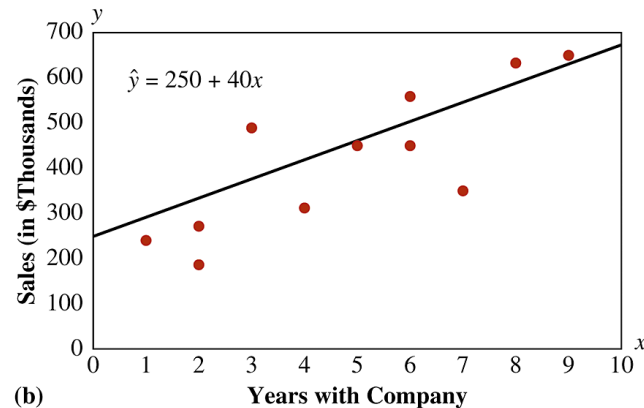
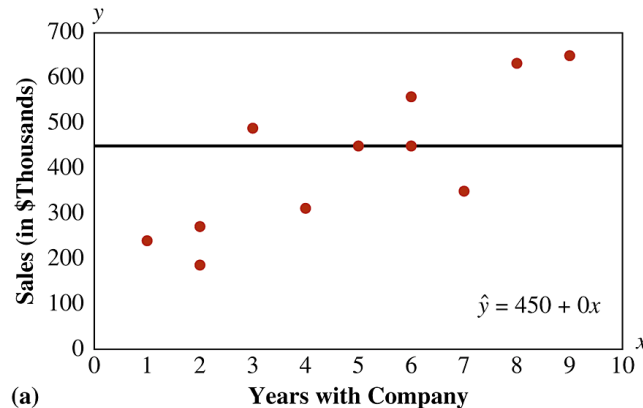
b_1 = estimate of the true population regression slope coefficient

b_o = estimate of the true population y intercept

How do we determine the values for b_o and b_1 ?

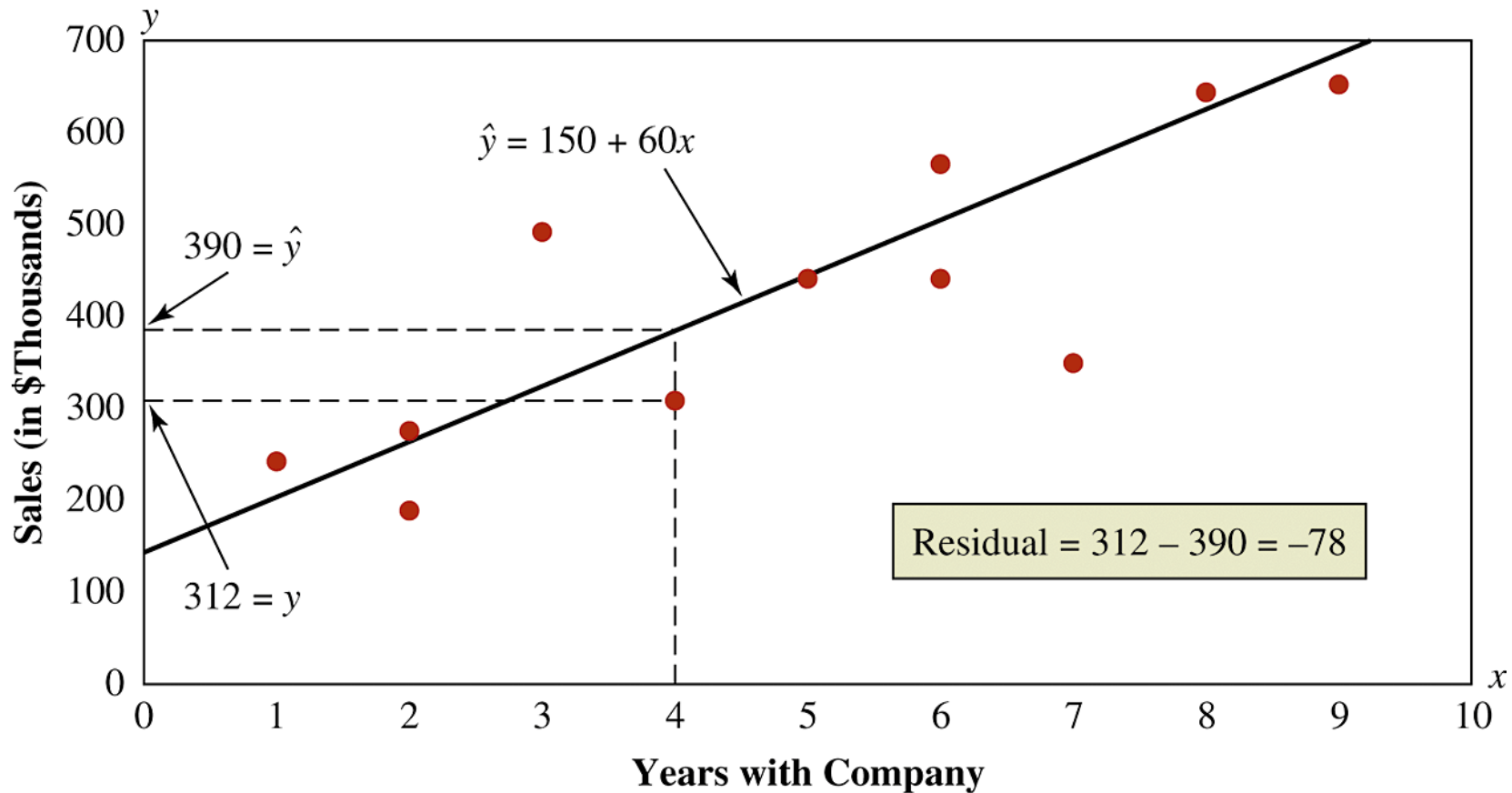


Regression Line Examples



Which Regression Line is Best? Examine Regression Errors

Computation of Regression Error - Example

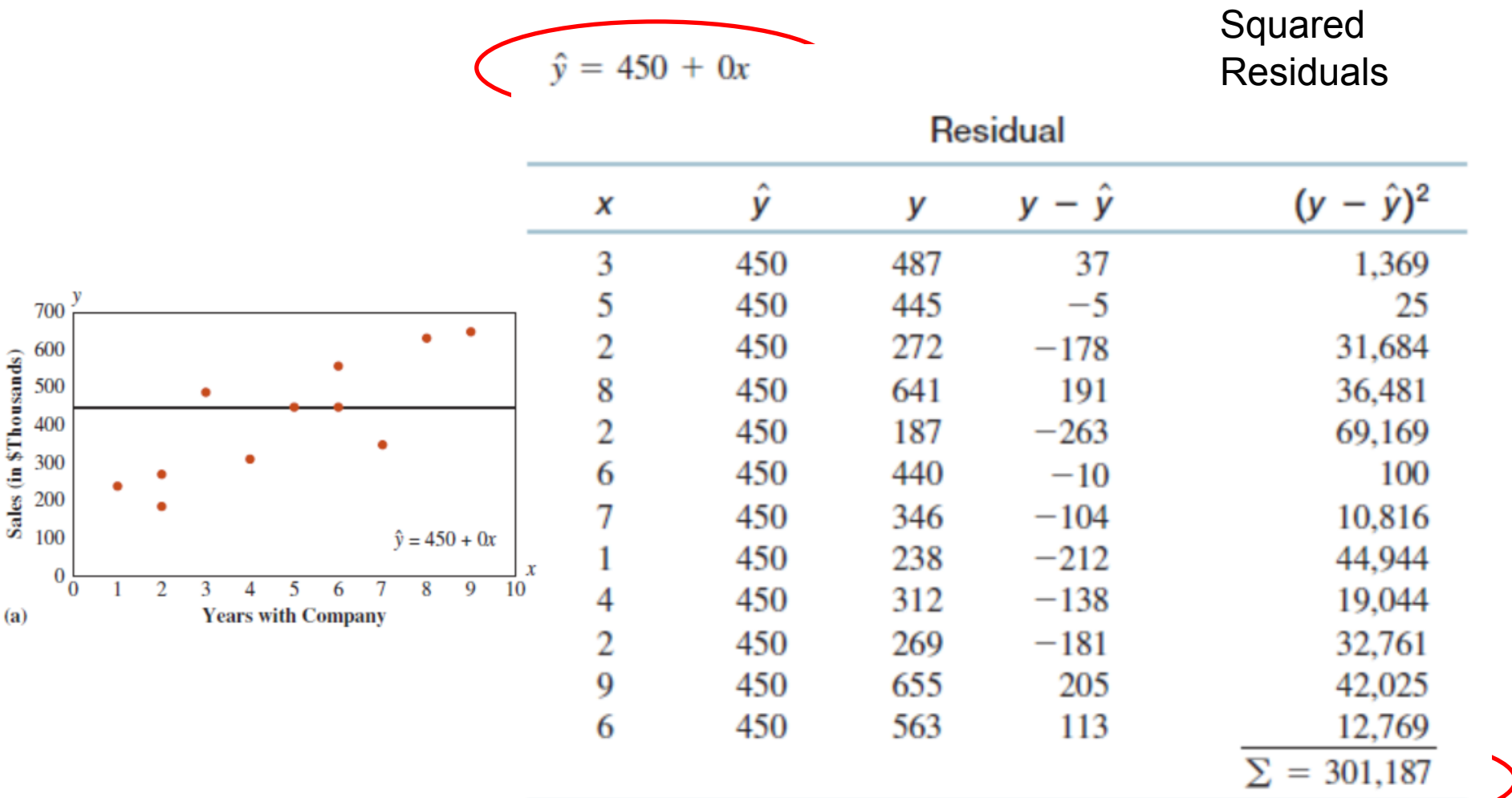


Least Squares Criterion

- The criterion for determining a regression line that minimizes the sum of squared prediction errors (residuals)
- **Residual:** The difference between the actual value of the dependent variable and the value predicted by the regression model.

Computation of Regression Residuals

– Trial-and-Error Example

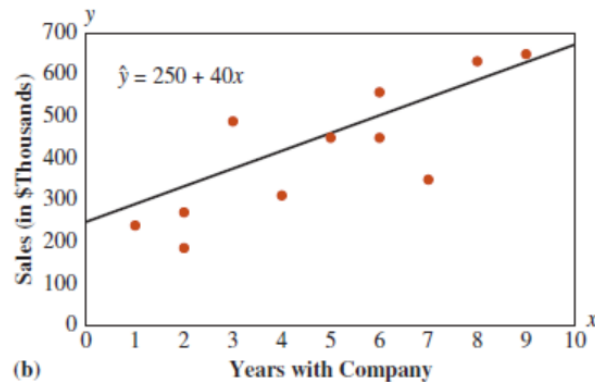


Computation of Regression Residuals

– Trial-and-Error Example

$$\hat{y} = 250 + 40x$$

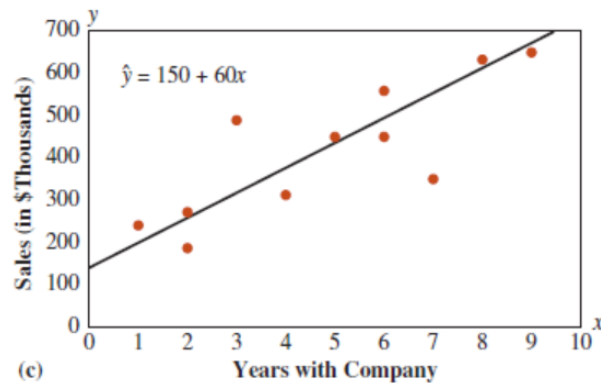
Squared
Residuals



Residual				
x	$\hat{y}$	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	370	487	117	13,689
5	450	445	-5	25
2	330	272	-58	3,364
8	570	641	71	5,041
2	330	187	-143	20,449
6	490	440	-50	2,500
7	530	346	-184	33,856
1	290	238	-52	2,704
4	410	312	-98	9,604
2	330	269	-61	3,721
9	610	655	45	2,025
6	490	563	73	5,329
				$\Sigma = 102,307$

Computation of Regression Residuals

– Trial-and-Error Example



$$\hat{y} = 150 + 60x$$

Squared
Residuals

Residual				
x	$\hat{y}$	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	330	487	157	24,649
5	450	445	-5	25
2	270	272	2	4
8	630	641	11	121
2	270	187	-83	6,889
6	510	440	-70	4,900
7	570	346	-224	50,176
1	210	238	28	784
4	390	312	-78	6,084
2	270	269	-1	1
9	690	655	-35	1,225
6	510	563	53	2,809
				$\Sigma = 97,667$

Least Squares Criterion

We need a more direct approach than trial-and-error! The answer lies in finding the slope and intercept such that the sum of squared residuals is minimized for the sample data.

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

The least squares equations:

$$b_1 = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2} \quad \text{and} \quad b_o = \bar{y} - b_1 \bar{x}$$

Least Squares Equations – Manual Calculations Example

y	x	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
487	3	-1.6	82.4	-131.84	2.56
445	5	0.4	40.4	16.16	0.16
272	2	-2.6	-132.6	344.76	6.76
641	8	3.4	236.4	803.76	11.56
187	2	-2.6	-217.6	565.76	6.76
440	6	1.4	35.4	49.56	1.96
346	7	2.4	-58.6	-140.64	5.76
238	1	-3.6	-166.6	599.76	12.96
312	4	-0.6	-92.6	55.56	0.36
269	2	-2.6	-135.6	352.56	6.76
655	9	4.4	250.4	1101.76	19.36
563	6	1.4	158.4	221.76	1.96
$\bar{y} = 404.6$	$\bar{x} = 4.6$			$\Sigma = 3838.92$	$\Sigma = 76.92$

$$b_1 = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2} = \frac{3838.92}{76.92} = 49.91$$

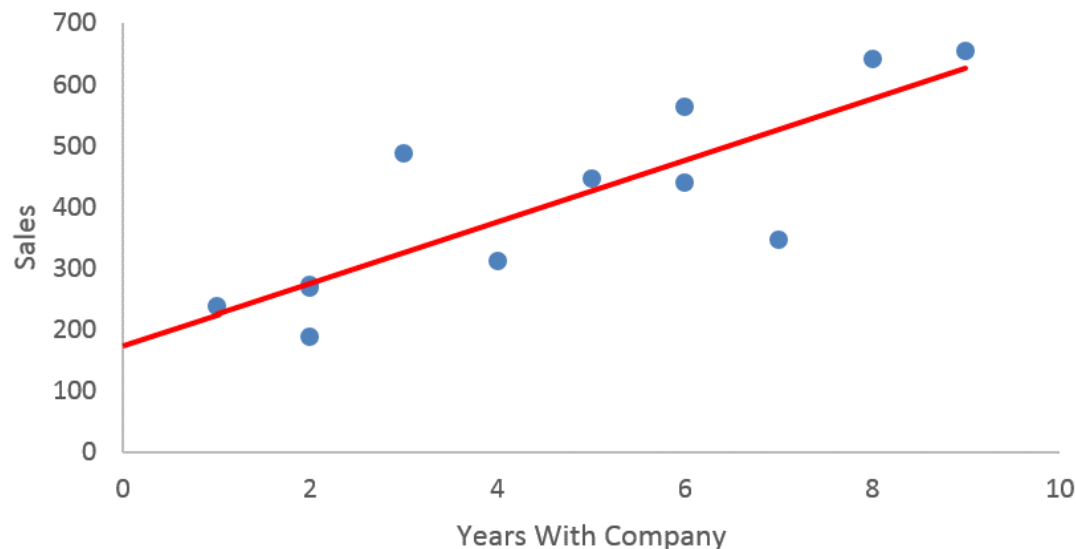
$$b_o = \bar{y} - b_1 \bar{x} = 404.6 - (49.91)(4.6) = 175.01$$

Estimated Regression Equation -Example

$$b_1 = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2} = \frac{3838.92}{76.92} = 49.91 \quad b_o = \bar{y} - b_1 \bar{x} = 404.6 - (49.91)(4.6) = 175.01$$

$$\hat{y} = 175.01 + 49.91(x)$$

Midwestern Company



Minimum Sum of Squares Residuals- Example

y	x	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
487	3	324.74	162.26	26328.31
445	5	424.56	20.44	417.79
272	2	274.83	-2.83	8.01
641	8	574.29	66.71	4450.22
187	2	274.83	-87.83	7714.11
440	6	474.47	-34.47	1188.18
346	7	524.38	-178.38	31819.42
238	1	224.92	13.08	171.09
312	4	374.65	-62.65	3925.02
269	2	274.83	-5.83	33.99
655	9	624.2	30.8	948.64
563	6	474.47	88.53	7837.56
			$\Sigma = 84,842.35$	

The Least
Squares
equations
minimize
SSE



Excel 2016 Regression Results

1. Open file
2. Select Data Analysis
3. Select Regression
4. Define y

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	Regression Statistics						
4	Multiple R	0.8325	$\min \sum_{i=1}^n (y - \hat{y})^2 = 84,834.29$				
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	Observation	Predicted	Residuals				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

$$\hat{y} = 175.83 + 49.91(x)$$

(Regression results differ slightly from manual calculations due to rounding.)

Test for Significance of the Regression Slope Coefficient

- Hypotheses:

- A slope of 0 would

-- -- -- -- --

Test Statistic for Test of the Significance of the Slope Coefficient

- Hypotheses:

- Test Statistic:

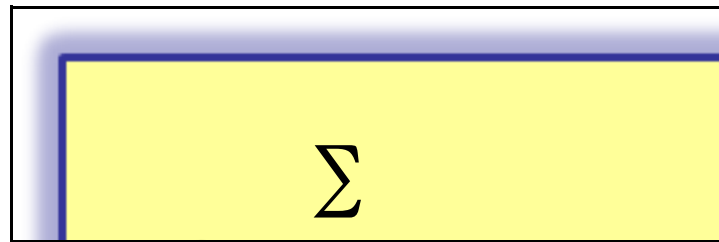
$$H_0: \beta$$

$$b_1 - \beta$$

b_1 - Sample regression
" " " " " "

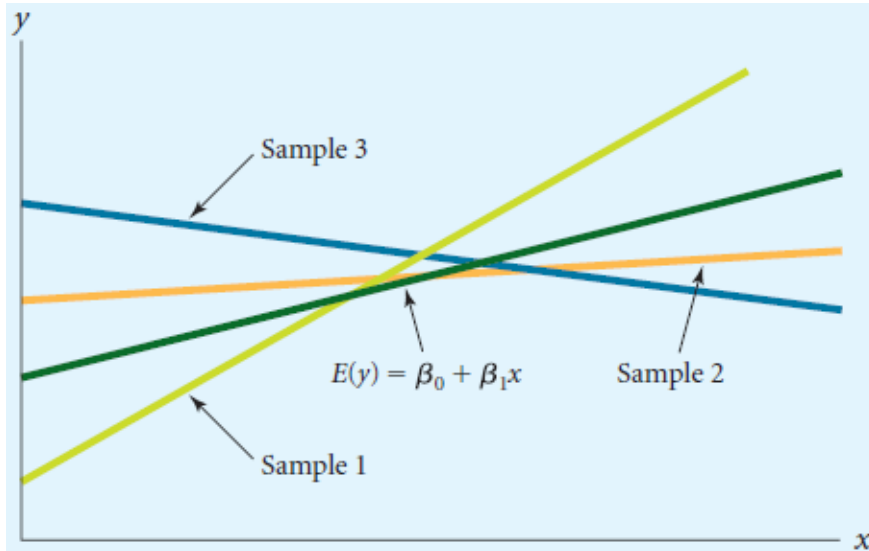
Standard Error of the Slope

- Simple Regression Estimator for the Standard Error of the Slope:

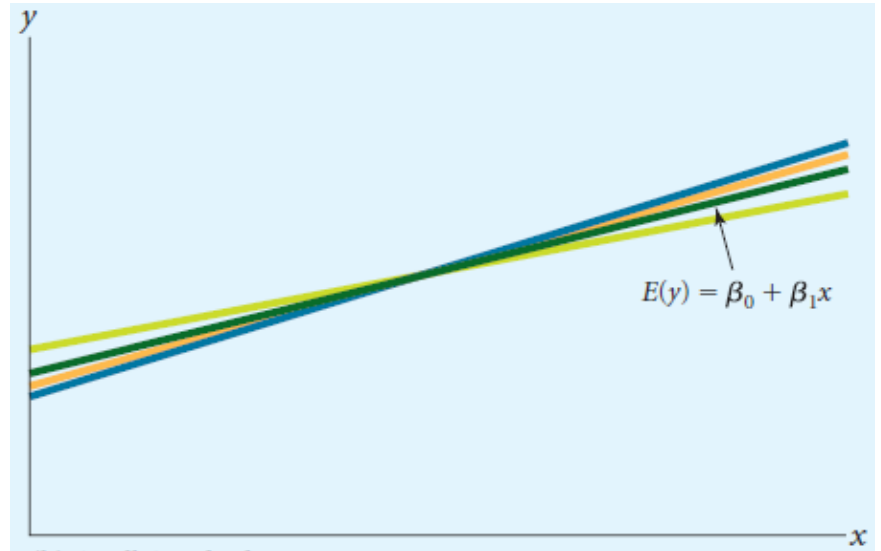

$$\Sigma$$

s_{b_1} - Standard deviation
slope (*standard error*)

Standard Error of the Slope



Large Standard Error



Small Standard Error

Standard Error of the Slope- Example

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325					
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

$$S_e = \sqrt{SSE} = \text{Standard Error of the Estimate}$$

$$S_{b_1} = \sqrt{\frac{S_e}{\sum (x - \bar{x})^2}} = \text{Standard Error of Slope Coefficient}$$



Test Statistic for Test of the Significance of the Slope Coefficient

$$H_o : B_1 = 0.0$$

$$H_1 : B_1 \neq 0.0$$

$$\alpha = 0.05$$

Test Statistic

$$t = \frac{b_1 - B_1}{S_{b_1}} = \frac{49.91 - 0.0}{10.5021} = 4.752$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325					
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

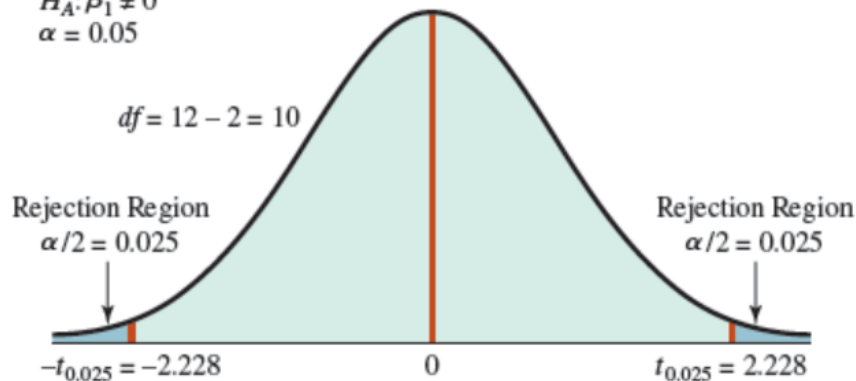
Test Statistic for Test of the Significance of the Slope Coefficient

Hypotheses:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$$\alpha = 0.05$$



The calculated t is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{49.91 - 0}{10.50} = 4.752$$

Decision Rule:

If $t > t_{0.025} = 2.228$, reject H_0 .

If $t < -t_{0.025} = -2.228$, reject H_0 .

Otherwise, do not reject H_0 .

Because $4.752 > 2.228$, we reject the null hypothesis and conclude that the true slope is not 0. Thus, the simple linear relationship that utilizes the independent variable, years with the company, is useful in explaining the variation in the dependent variable, sales volume.

p-value for Test of the Significance of the Slope Coefficient

$$H_o : B_1 = 0.0$$

$$H_1 : B_1 \neq 0.0$$

$$\alpha = 0.05$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325					
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

p-value

Because p-value = 0.0008 < alpha/2 = 0.025, reject Ho

Sums of Squares

- Total Sum of Squares:

$$\sum_{i=1}^n$$

- Sum of Squares Regression:

$$\sum_{i=1}^n$$

n - Sample size

y_i - i^{th} value of the dependent variable

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

Sums of Squares - Example

SSR

SSE

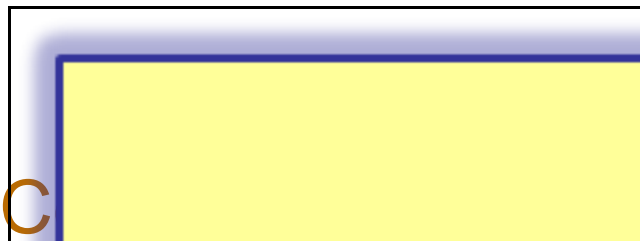
SST

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325					
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

$$SST = SSR + SSE$$

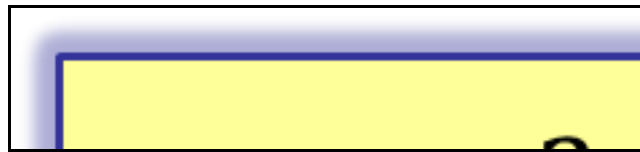
The Coefficient of Determination R^2

- The portion of the total variation in the dependent variable that is explained by its relationship with the independent variable



SSR - Sum of squares

- Coefficient of Determination for the Single Independent Variable Case



Sum of squares

The Coefficient of Determination R^2

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325	$R^2 = \frac{SSR}{SST} = \frac{191,600.62}{276,434.92} = 0.6931$				
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

Test Statistic for Significance of the Coefficient of Determination

$$H_o : \rho^2 = 0.0$$

$$H_A : \rho^2 > 0.0$$

The independent variable does not explain a significant proportion of the total variation in the dependent variable

Test Statistic

$$F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE}$$

$$df, D_1 = 1 \text{ and } D_2 = n - 2$$



Test Statistic for Significance of the Coefficient of Determination

$$H_o : \rho^2 = 0.0$$

$$H_A : \rho^2 > 0.0$$

$$\alpha = 0.05$$

Test Statistic

$$F = \frac{MSR}{MSE}$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.8325					
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
19							
20	RESIDUAL OUTPUT						
21	<i>Observation</i>	<i>Predicted</i>	<i>Residuals</i>				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

Because $F = 22.59 > F_{\text{critical}, 0.05} = 4.965$, reject the null hypothesis



p-value for Significance of the Coefficient of Determination

$$H_o : \rho^2 = 0.0$$

$$H_A : \rho^2 > 0.0$$

$$\alpha = 0.05$$

Because p-value = 0.0008 < alpha = 0.05, reject the null hypothesis

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.8325		p-value = 0.0008			
5	R Square	0.6931					
6	Adjusted R Square	0.6624					
7	Standard Error	92.1055					
8	Observations	12					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	191,600.62	191,600.62	22.59	0.0008	
13	Residual	10	84,834.29	8,483.43			
14	Total	11	276,434.92				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
18	Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102
19							
20	RESIDUAL OUTPUT						
21	Observation	Predicted	Residuals				
22	1	325.56	161.44				
23	2	425.38	19.62				
24	3	275.65	-3.65				

14.3 Uses for Regression Analysis

- Description – When we are primarily interested in analyzing the relationship between the x and y variables as measured by the regression slope coefficient
- Prediction – When we are primarily interested in predicting what the value of the y variable will be when we know a value of the x variable.

Regression Analysis for Description - Example

The Environmental Protection Agency (EPA) is interested in the relationship between vehicle mileage and the CO2 emitted by the vehicle. To analyze the relationship, staff members have collected sample data from 58 vehicles and used Excel to compute the following regression output.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2	<i>Regression Statistics</i>						
3	Multiple R	0.9589					
4	R Square	0.9194					
5	Adjusted R Square	0.9180					
6	Standard Error	16.2705					
7	Observations	58					
8							
9	ANOVA						
10		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
11	Regression	1	169115.67	169115.67	638.83	0.0000	
12	Residual	56	14824.81	264.73			
13	Total	57	183940.48				
14							
15		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
16	Intercept	703.39	14.20	49.53	0.0000	674.94	731.84
17	Combined MPG	-13.64	0.54	-25.28	0.0000	-14.72	-12.56

$$\hat{y} = 703.39 - 13.64(\text{mpg})$$

$$H_o : B_1 = 0.0$$

$$H_1 : B_1 \neq 0.0$$

$$\alpha = 0.05$$

Because p-value = 0.0000 < 0.05/2 we reject the null hypothesis



Regression Analysis for Description – Regression Slope Analysis

- Confidence Interval Estimate for the Regression Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

s_{b_1} - Standard deviation
slope coefficient

Regression Analysis for Description

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2	<i>Regression Statistics</i>						
3	Multiple R	0.9589					
4	R Square	0.9194					
5	Adjusted R Square	0.9180					
6	Standard Error	16.2705					
7	Observations	58					
8							
9	ANOVA						
10		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
11	Regression	1	169115.67	169115.67	638.83	0.0000	
12	Residual	56	14824.81	264.73			
13	Total	57	183940.48				
14							
15		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
16	Intercept	703.39	14.20	49.53	0.0000	674.94	731.84
17	Combined MPG	-13.64	0.54	-25.28	0.0000	-14.72	-12.56

$$b_1 \pm t_{0.05, df=56} S_{b_1}$$

Based on the sample data, with 95% confidence, we believe that for each increase on one mpg, the mean change in CO2 is between -14.72 and -12.56 grams with a point estimate of -13.64 grams



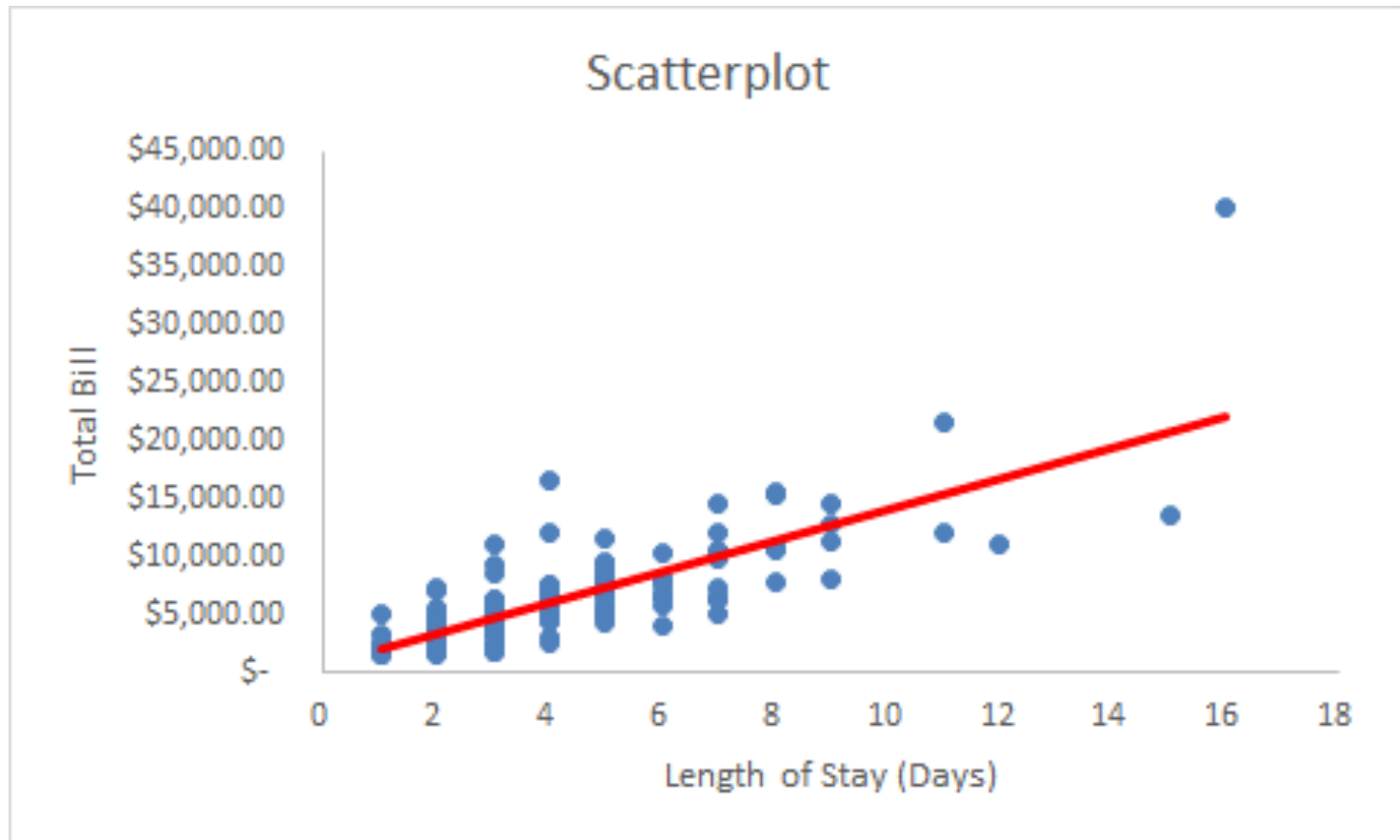
Regression Analysis for Prediction

Hospital administrators wish to predict the total hospital bill based on knowing the patient's length of stay in the hospital. Data were collected for 138 patients and the following regression output was produced by Excel 2016

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.77					
R Square	0.60					
Adjusted R Square	0.59					
Standard Error	2894.78					
Observations	138.00					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	1683440143.46	1683440143.46	200.89	0.00	
Residual	136	1139647630.81	8379761.99			
Total	137	2823087774.28				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	527.61	483.81	1.09	0.28	-429.17	1484.38
Length of Stay	1352.80	95.44	14.17	0.00	1164.05	1541.54

$$\hat{y} = 527.61 + 1352.80(days)$$

Regression Analysis for Prediction – Scatterplot Example



$$\hat{y} = 527.61 + 1352.80(days)$$

Regression Analysis for Prediction – Point Estimate

Relevant Range for the x variable = 1 to 16 days

$$\hat{y} = 527.61 + 1352.80(days)$$

Point Prediction Value for x = 5 days

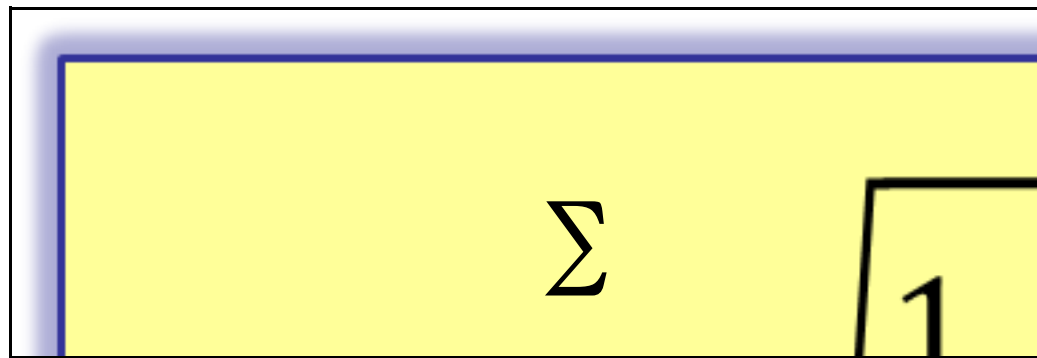
$$\hat{y} = 527.61 + 1352.80(5) = \$7,291.59$$

Point Prediction Value for x = 9 days

$$\hat{y} = 527.61 + 1352.80(9) = \$12,702.81$$



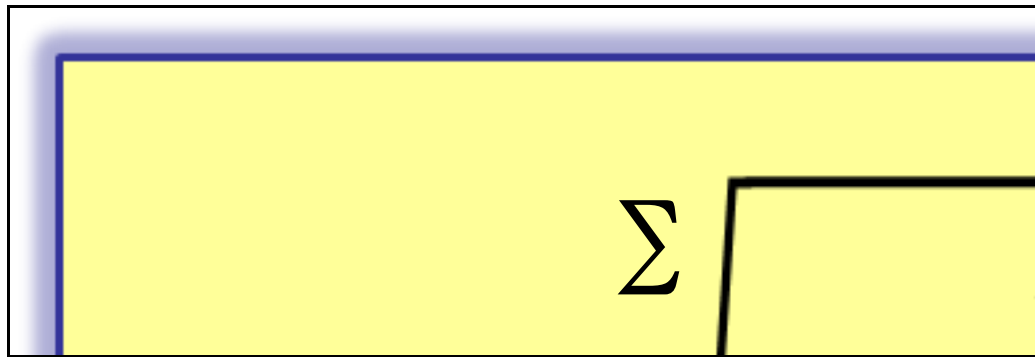
Confidence Inter


$$\Sigma \sqrt{1}$$

$\hat{y}$ - Point estimate of the depen
 t - Critical value with $n - 2$ deg
 n - Sample size

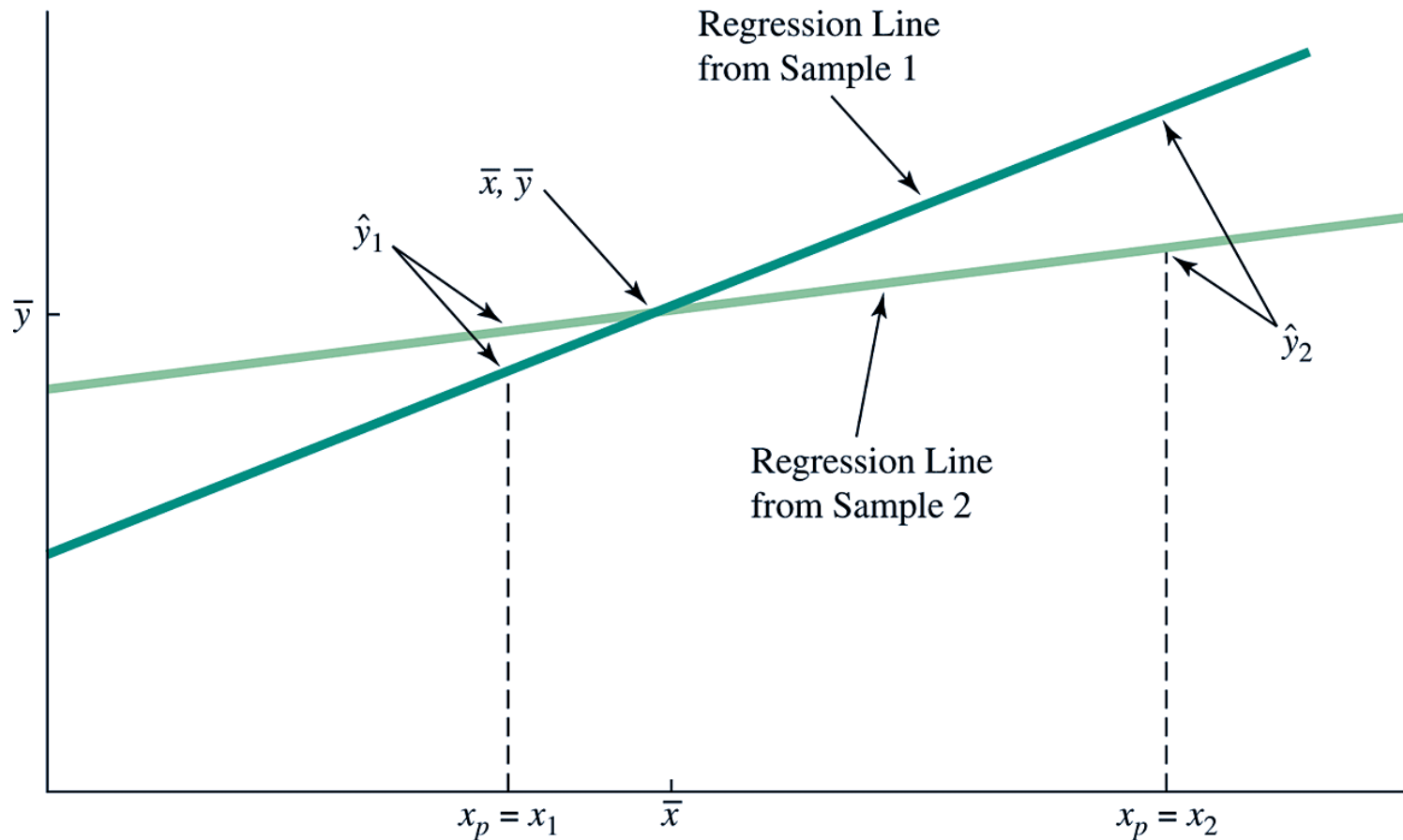
Prediction Interval

• • • • •

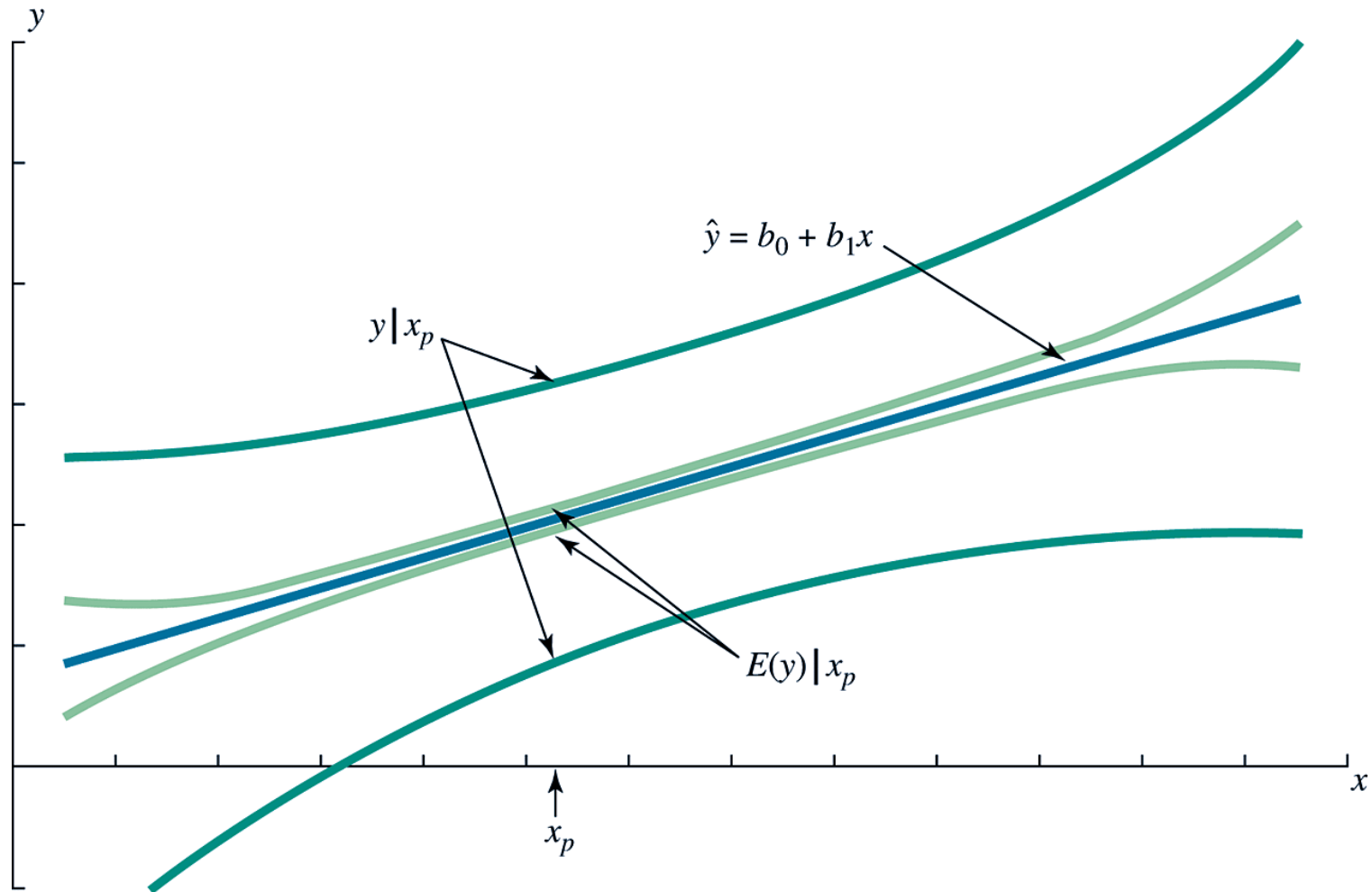


The term $(x_p - \bar{x})^2$ has a
confidence and prediction

Potential Variation in y as x_p Moves Farther from $\bar{x}$



Confidence and Prediction Intervals



Confidence and Prediction Intervals Using Excel 2016 and XLSTAT – Hospital Example

$$x_p = 5 \text{ days}$$

Predictions for the new observations:							
Observation	Pred(Total Chges)	Std. dev. on pred. (Mean)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Std. dev. on pred. (Observation)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
PredObs1	7,291.59	253.83	6,789.63	7,793.54	2,905.89	1,545.01	13,038.16

Point Estimate

Condence Interval
6,789.63 ——— 7,793.54

Prediction Interval
1,545.01—13,038.16