

Probability and Statistics

Prepared by Dr. Alfaisal Hasan

STA 201
Probability and Statistics

Lecture NO 9





مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY

Probability and Statistics

Lecture NO 9

STA 201

Probability and Statistics

Prepared by
Dr. Alfaisal Hasan

www.midocean.ae



Chapter

Topics

- Chapter 1: Introduction To Statistics
- Chapter 2: Descriptive Statistics
- Chapter 3: Probability
- Chapter 4: Discrete Probability Distributions
- Chapter 5: Normal Probability Distributions
- Chapter 6: Confidence Intervals
- Chapter 7: Hypothesis Testing with One Sample
- Chapter 8: Hypothesis Testing with Two Samples
- **Chapter 9: Correlation And Regression**
- Chapter 10: Chi-Square Tests and The F -Distribution

Chapter Outline

9.1 Correlation

9.2 Linear Regression

9.3 Measures of Regression and Prediction Intervals

Section 9.1: Correlation

A relationship between two variables. The data can be represented by ordered pairs (x, y)

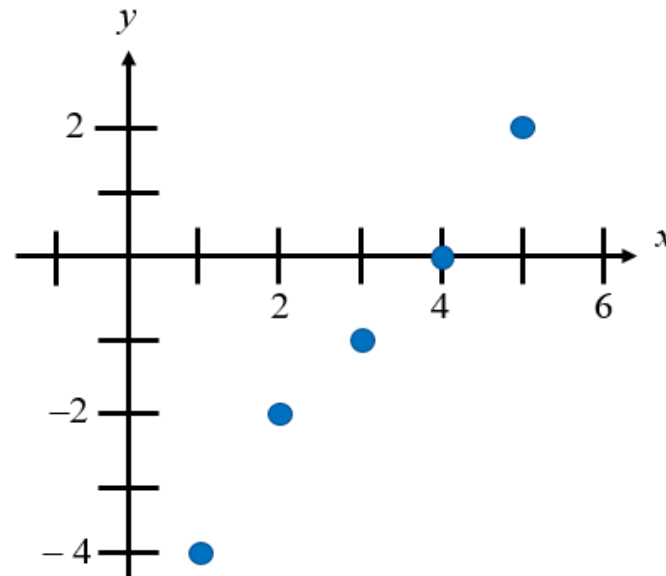
x is the **independent** (or **explanatory**) **variable**

y is the **dependent** (or **response**) **variable**

A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

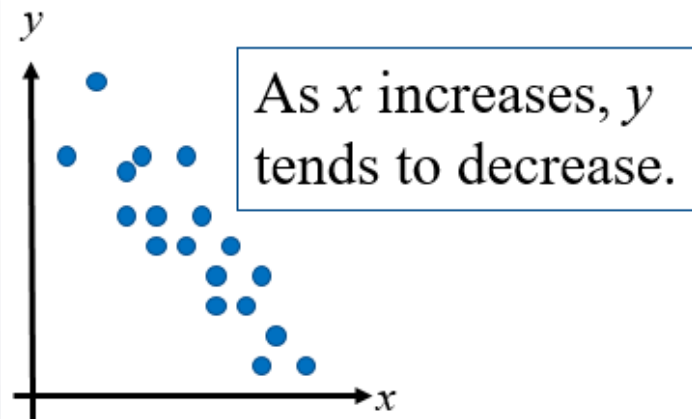
Example:

x	1	2	3	4	5
y	-4	-2	-1	0	2

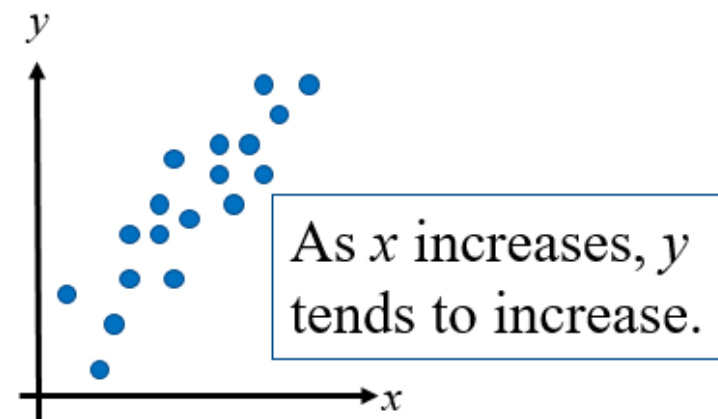


Types of Correlation

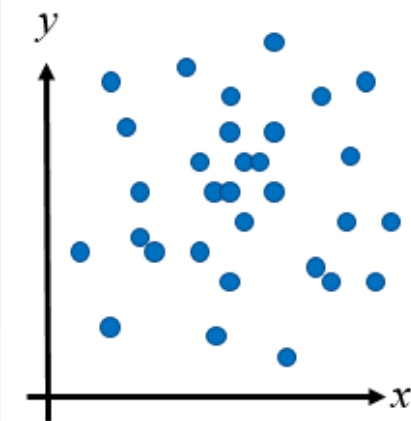
Chapter



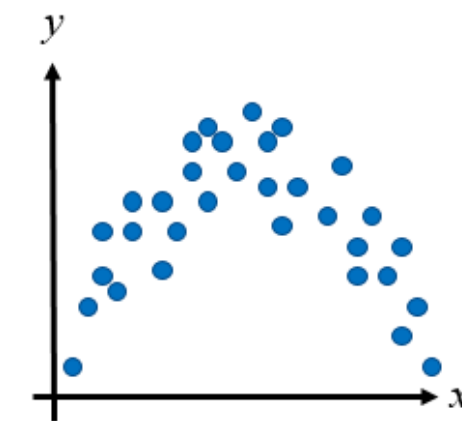
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

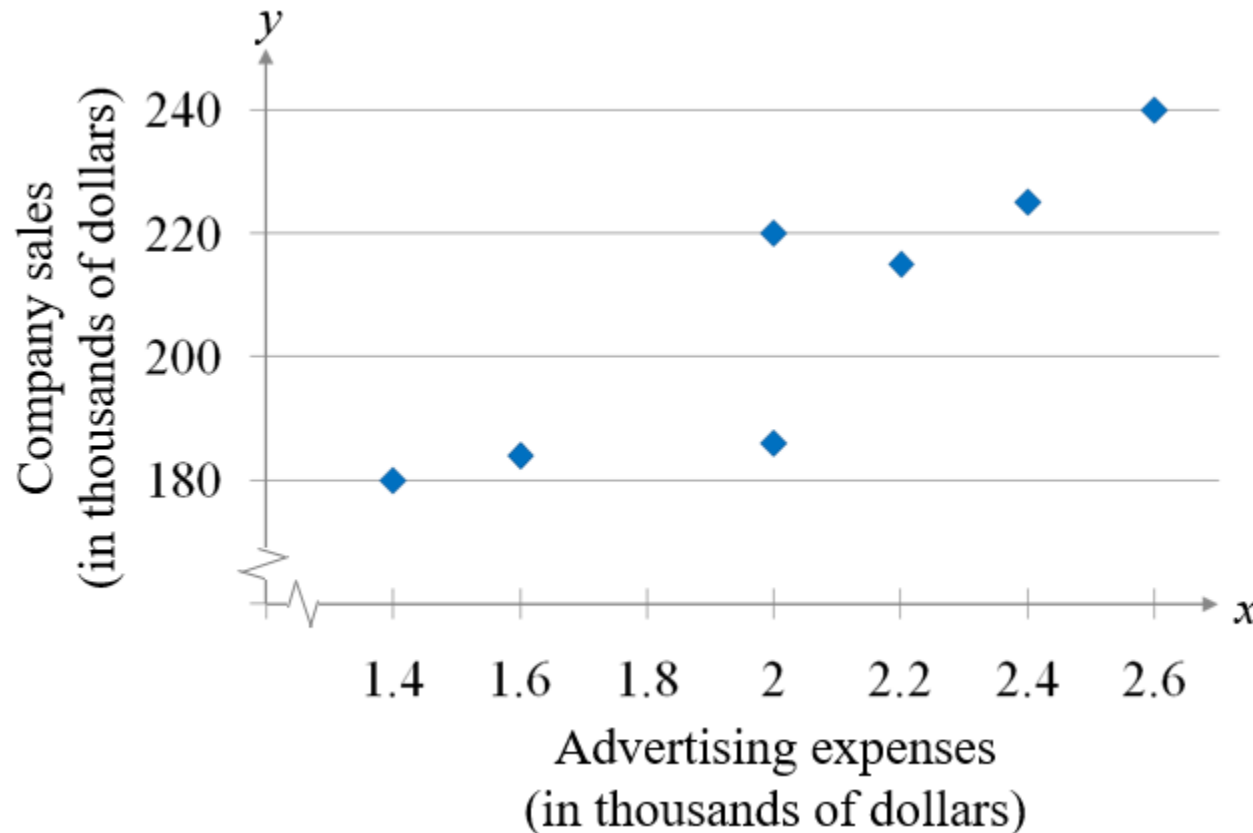
Example: Constructing a Scatter Plot

A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are shown in the table.

Display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation or no linear correlation.

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: Constructing a Scatter Plot



Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Appears to be a **positive linear correlation**. As the advertising expenses increase, the sales tend to increase.

Correlation coefficient

Correlation coefficient

A measure of the strength and the direction of a linear relationship between two variables.

The symbol r represents the sample correlation coefficient.

A formula for r is

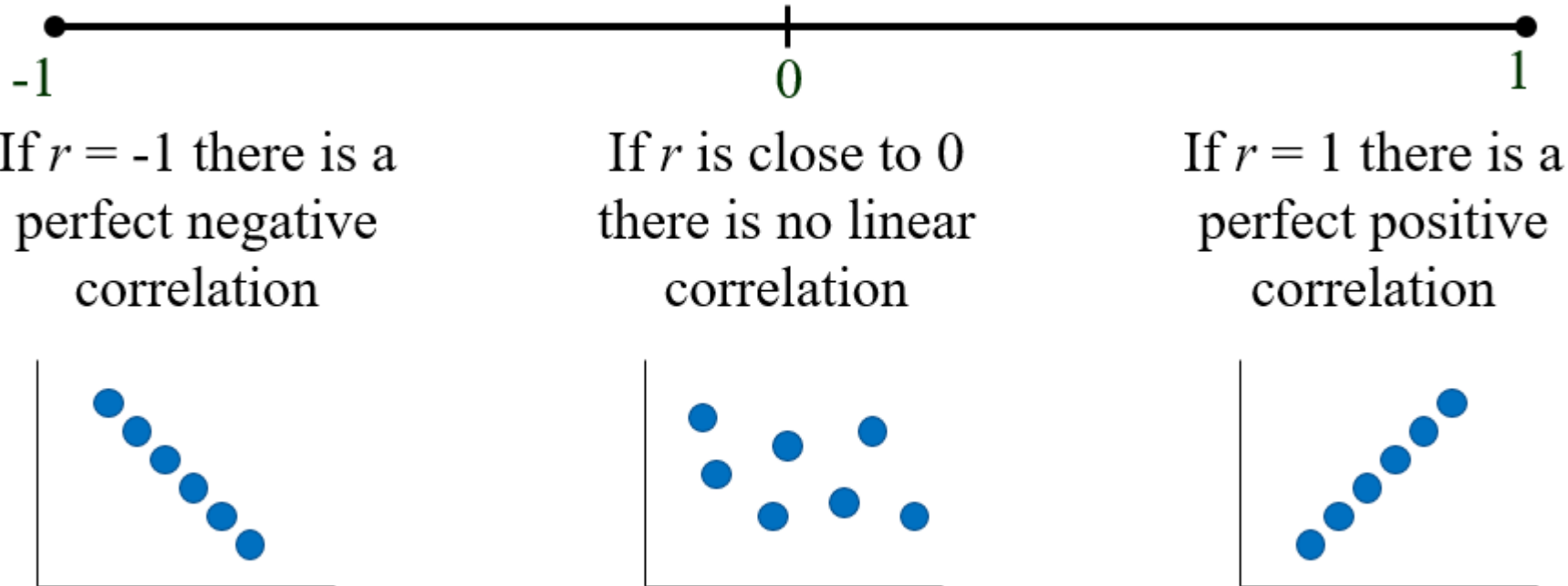
$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

n is the number
of data pairs

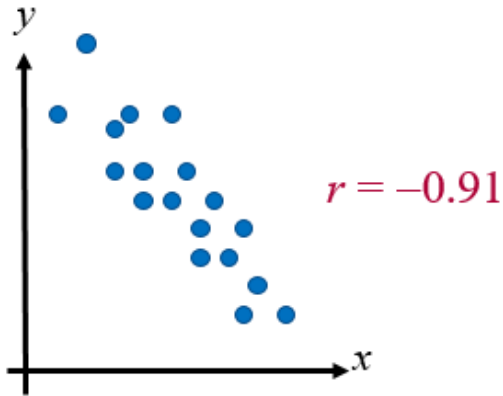
The population correlation coefficient is represented by ρ (rho).

Correlation coefficient

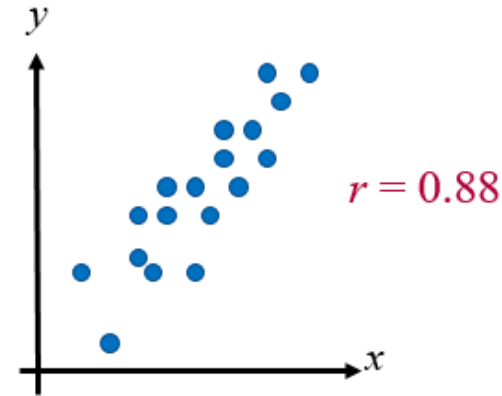
The range of the correlation coefficient is -1 to 1.



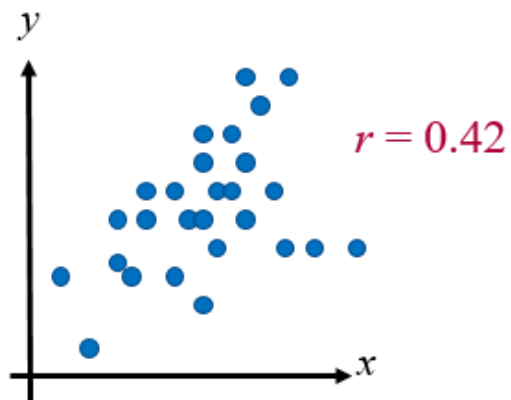
Linear Correlation



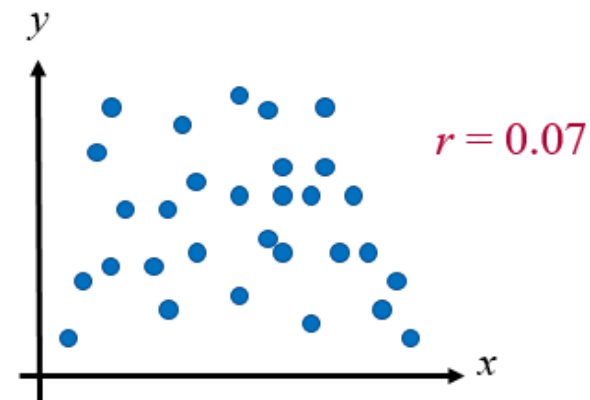
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Example: Finding the Correlation Coefficient

Calculate the correlation coefficient for the advertising expenditures and company sales data. What can you conclude?

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: Finding the Correlation Coefficient

x	y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$	$\Sigma y^2 = 337,558$

Solution: Finding the Correlation Coefficient

$$\Sigma x = 15.8 \quad \Sigma y = 1634 \quad \Sigma xy = 3289.8 \quad \Sigma x^2 = 32.44 \quad \Sigma y^2 = 337,558$$

$$\begin{aligned} r &= \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} \\ &= \frac{8(3289.8) - (15.8)(1634)}{\sqrt{8(32.44) - 15.8^2} \sqrt{8(337,558) - 1634^2}} \\ &= \frac{501.2}{\sqrt{9.88} \sqrt{30,508}} \approx 0.9129 \end{aligned}$$

$r \approx 0.913$ suggests a strong positive linear correlation. As the amount spent on advertising increases, the company sales also increase.

Example: Finding the Correlation Coefficient

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

Age (years)	Weight (Kg)
7	12
6	8
8	12
5	10
6	11
9	13

Example: Finding the Correlation Coefficient

Age (years) (x)	Weight (Kg) (y)	xy	X ²	Y ²
7	12	84	49	144
6	8	48	36	64
8	12	96	64	144
5	10	50	25	100
6	11	66	36	121
9	13	117	81	169
$\Sigma x = 41$	$\Sigma y = 66$	$\Sigma xy = 461$	$\Sigma x^2 = 291$	$\Sigma y^2 = 742$

Example: Finding the Correlation Coefficient

$$r = \frac{(6) \times (461) - (41) \times (66)}{\sqrt{[(6) \times (291) - (41)^2] \cdot [(6) \times (742) - (66)^2]}}$$

$$r = 0.759$$

strong direct correlation

Example: Finding the Correlation Coefficient

EXAMPLE: Relationship between
Anxiety and Test Scores

Test score (Y)	Anxiety (X)
2	10
3	8
9	2
7	1
6	5
5	6

Example: Finding the Correlation Coefficient

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X =$ 32	$\sum Y =$ 32	$\sum X^2 =$ 230	$\sum Y^2 =$ 204	$\sum XY =$ 1 29

Example: Finding the Correlation Coefficient

$$r = \frac{(6) \times (129) - (32) \times (32)}{\sqrt{((6) \times (230) - 32^2)((6) \times (204) - 32^2)}}$$

$$r = \frac{774 - 1024}{\sqrt{(356) \times (200)}} = -0.94$$

Indirect Strong correlation

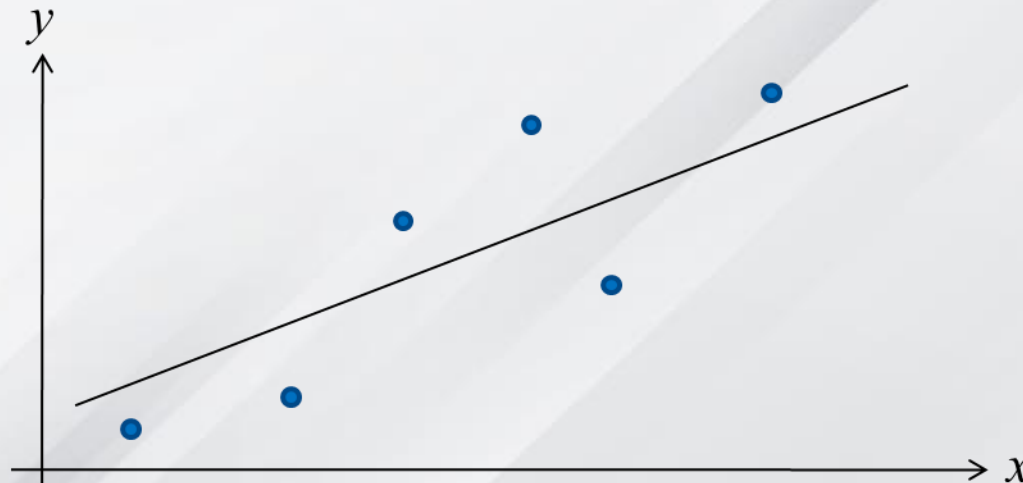
Section 9.2: Linear Regression

- Find the equation of a regression line
- Predict y -values using a regression equation

Regression Line

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that best models the data (**regression line**).

Can be used to predict the value of y for a given value of x .



Regression Line

Regression line (line of best fit)

The line for which the sum of the squares of the residuals is a minimum.

The equation of a regression line for an independent variable x and a dependent variable y is

$$\hat{y} = mx + b$$

Predicted
 y -value for
a given x -
value

Slope

y -intercept

The equation of Regression Line

- $\hat{y} = mx + b$ where

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \quad b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m \frac{\sum x}{n}$$

- $\bar{y}$ is the mean of the y -values in the data
- $\bar{x}$ is the mean of the x -values in the data
- The regression line always passes through the point $(\bar{x}, \bar{y})$

Example: Finding the Equation of a Regression Line

Find the equation of the regression line for the following data.

X	1	3	4	6	7	9
Y	6	4	4	3	2	1

Solution: Finding the Equation of a Regression Line

Recall from section 9.1:

	x	y	x^2	xy
	1	6	1	6
	3	4	9	12
	4	4	16	16
	6	3	36	18
	7	2	49	14
	9	1	81	9
Σ	30	20	192	75

Solution: Finding the Equation of a Regression Line

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$m = \frac{(6) \times (75) - (30) \times (20)}{(6) \times (192) - 30^2} = -0.595$$

$$b = \bar{y} - m\bar{x}$$

$$b = \frac{20}{6} - (-0.595) \frac{30}{6} = 6.31$$

Equation of regression line

$$\hat{y} = -0.595x + 6.31$$

Example: Finding the Equation of a Regression Line

Find the equation of the regression line for the advertising expenditures and company sales data.

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: Finding the Equation of a Regression Line

Recall from section 9.1:

x	y	xy	x^2
2.4	225	540	5.76
1.6	184	294.4	2.56
2.0	220	440	4
2.6	240	624	6.76
1.4	180	252	1.96
1.6	184	294.4	2.56
2.0	186	372	4
2.2	215	473	4.84
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$

Solution: Finding the Equation of a Regression Line

$$\Sigma x = 15.8 \quad \Sigma y = 1634 \quad \Sigma xy = 3289.8 \quad \Sigma x^2 = 32.44$$

$$m = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{8(3289.8) - (15.8)(1634)}{8(32.44) - 15.8^2}$$

$$= \frac{501.2}{9.88} \approx 50.72874$$

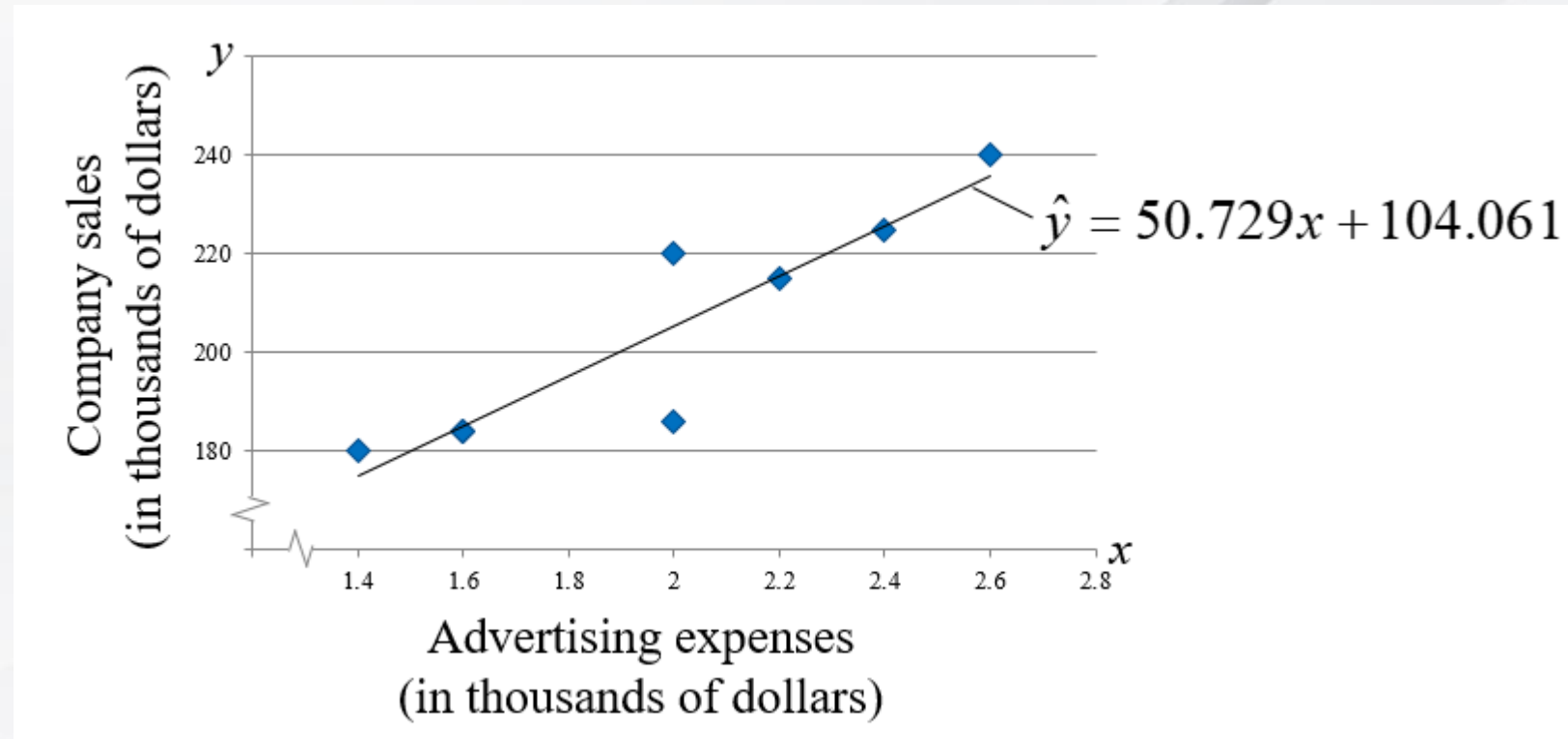
$$b = \bar{y} - m\bar{x} = \frac{1634}{8} - (50.72874)\frac{15.8}{8}$$

$$= 204.25 - (50.72874)(1.975) \approx 104.0607$$

Equation of the regression line $\hat{y} = 50.729x + 104.061$

Example: Predicting y-Values Using Regression Equations

To sketch the regression line, use any two x -values within the range of the data and calculate the corresponding y -values from the regression line.



Solution: Predicting y-Values Using Regression Equations

The regression equation for the advertising expenses (in thousands of dollars) and company sales (in thousands of dollars) data is $\hat{y} = 50.729x + 104.061$. Use this equation to predict the *expected* company sales for the following advertising expenses. (Recall from section 9.1 that x and y have a significant linear correlation.)

1. 1.5 thousand dollars
2. 1.8 thousand dollars
3. 2.5 thousand dollars

Solution: Predicting y-Values Using Regression Equations

$$\hat{y} = 50.729x + 104.061$$

1. 1.5 thousand dollars

$$\hat{y} = 50.729(1.5) + 104.061 \approx 180.155$$

When the advertising expenses are \$1500, the company sales are about \$180,155.

2. 1.8 thousand dollars

$$\hat{y} = 50.729(1.8) + 104.061 \approx 195.373$$

When the advertising expenses are \$1800, the company sales are about \$195,373.

Solution: Predicting y-Values Using Regression Equations

3. 2.5 thousand dollars

$$\hat{y} = 50.729(2.5) + 104.061 \approx 230.884$$

When the advertising expenses are \$2500, the company sales are about \$230,884.

Prediction values are meaningful only for x -values in (or close to) the range of the data. The x -values in the original data set range from 1.4 to 2.6. So, it would not be appropriate to use the regression line to predict company sales for advertising expenditures such as 0.5 (\$500) or 5.0 (\$5000).

Section 9.3: Measures of Regression and Prediction Intervals

Variation About a Regression Line

Three types of variation about a regression line

- Total variation

- Explained variation

- Unexplained variation

To find the total variation, you must first calculate

- The **total deviation**

- The **explained deviation**

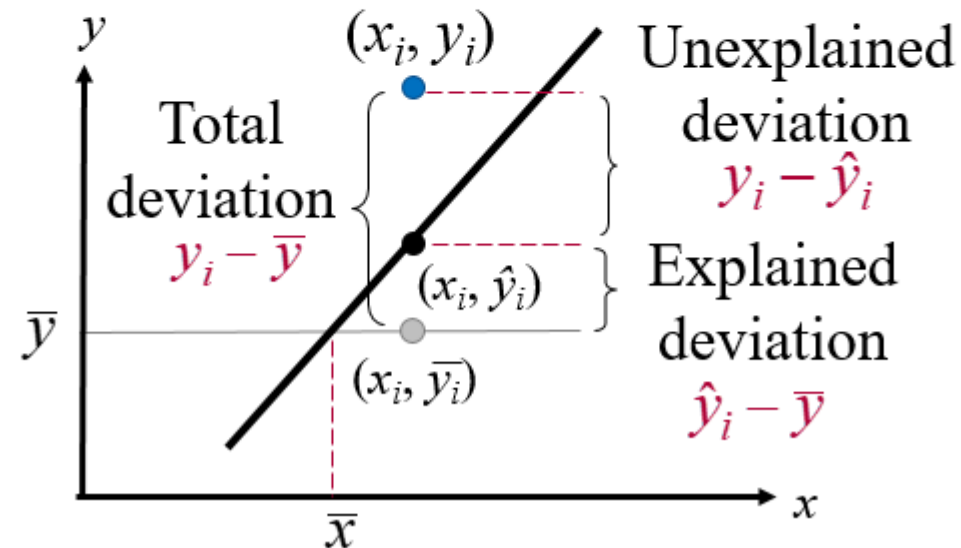
- The **unexplained deviation**

Variation About a Regression Line

Total Deviation = $y_i - \bar{y}$

Explained Deviation = $\hat{y}_i - \bar{y}$

Unexplained Deviation = $y_i - \hat{y}_i$



Variation About a Regression Line

Total variation

- The sum of the squares of the differences between the y -value of each ordered pair and the mean of y .

$$\text{Total variation} = \sum (y_i - \bar{y})^2$$

Explained variation

- The sum of the squares of the differences between each predicted y -value and the mean of y .

$$\text{Explained variation} = \sum (\hat{y}_i - \bar{y})^2$$

Variation About a Regression Line

Unexplained variation

- The sum of the squares of the differences between the y -value of each ordered pair and each corresponding predicted y -value.

$$\text{Unexplained variation} = \sum (y_i - \hat{y}_i)^2$$

The sum of the explained and unexplained variation is equal to the total variation.

$$\text{Total variation} = \text{Explained variation} + \text{Unexplained variation}$$

Coefficient of Determination

Coefficient of determination

- The ratio of the explained variation to the total variation.
- Denoted by r^2

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Example: Coefficient of Determination

The correlation coefficient for the advertising expenses and company sales data as calculated in Section 9.1 is $r \approx 0.913$. Find the coefficient of determination. What does this tell you about the explained variation of the data about the regression line? About the unexplained variation?

Solution: $r^2 \approx (0.913)^2$
 ≈ 0.834

About **83.4%** of the variation in the company sales can be explained by the variation in the advertising expenditures. About **16.9%** of the variation is unexplained.

The Standard Error of Estimate

Standard error of estimate

- The standard deviation of the observed y_i -values about the predicted $\hat{y}$ -value for a given x_i -value.
- Denoted by s_e .

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

n is the number of ordered pairs in the data set

- The closer the observed y -values are to the predicted y -values, the smaller the standard error of estimate will be.

Example: The Standard Error of Estimate

The regression equation for the advertising expenses and company sales data as calculated in section 9.2 is

$$\hat{y} = 50.729x + 104.061$$

Find the standard error of estimate.

Solution:

Use a table to calculate the sum of the squared differences of each observed y -value and the corresponding predicted y -value.

Solution: The Standard Error of Estimate

x	y	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
2.4	225	225.81	$(225 - 225.81)^2 = 0.6561$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	220	205.52	$(220 - 205.52)^2 = 209.6704$
2.6	240	235.96	$(240 - 235.96)^2 = 16.3216$
1.4	180	175.08	$(180 - 175.08)^2 = 24.2064$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	186	205.52	$(186 - 205.52)^2 = 381.0304$
2.2	215	215.66	$(215 - 215.66)^2 = 0.4356$
			$\Sigma = 635.3463$

↑
unexplained variation

Solution: The Standard Error of Estimate

- $n = 8, \Sigma(y_i - \hat{y}_i)^2 = 635.3463$

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{635.3463}{8 - 2}} \approx 10.290$$

The standard error of estimate of the company sales for a specific advertising expense is about \$10.29.

Chapter Summary

Chapter

9.1 Correlation

9.2 Linear Regression

9.3 Measures of Regression and Prediction Intervals

Un Solved Examples

Chapter

If the correlation coefficient $r = -0.02$, this means that there is strong negative (indirect) correlation.

إذا كان معامل الارتباط $r = -0.02$ فهذا يعني أن هناك علاقة قوية عكسية

Un Solved Examples

Chapter

If the equation of regression is given by $\hat{y} = 50x + 100$, then the predicting y-value at $x = 8$ is equal to

إذا كانت معادلة الانحدار $\hat{y} = 50x + 100$ فإن تنبؤ قيمة y عندما تكون قيمة $x = 8$ يساوي



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

