

Probability and Statistics

Prepared by Dr. Israa Lewaa

Lecture 9

جامعة الجميع الذكية

تسعى لتكون الكيان الأول الذي يتبنى مبدأ المجتمع التعليمي
التعاوني ويقدم العلم مقابل العمل لتنمية وبناء المجتمعات.

STA 201

Correlation and Regression



جامعة الجميع الذكية
EVERYONE'S SMART UNIVERSITY

9.1 Correlation

9.2 Linear Regression

9.3 Measures of Regression and Prediction Intervals

9.4 Multiple Regression

Section 9.1: Correlation

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

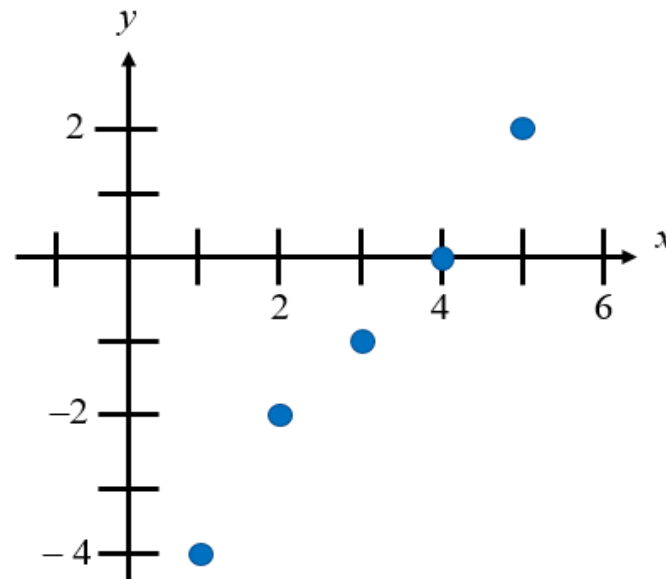
x is the **independent** (or **explanatory**) **variable**

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

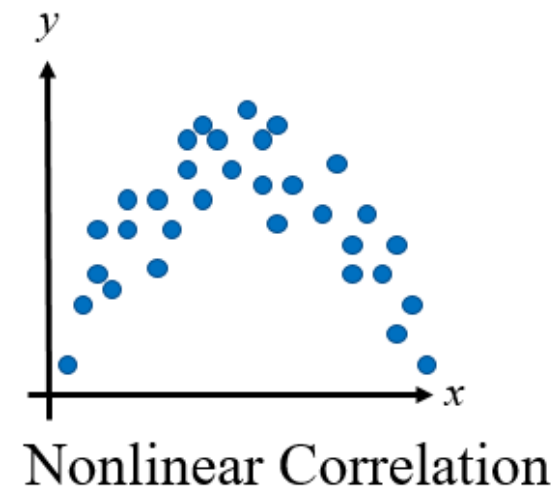
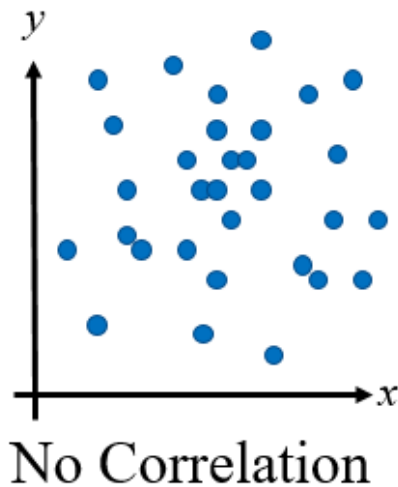
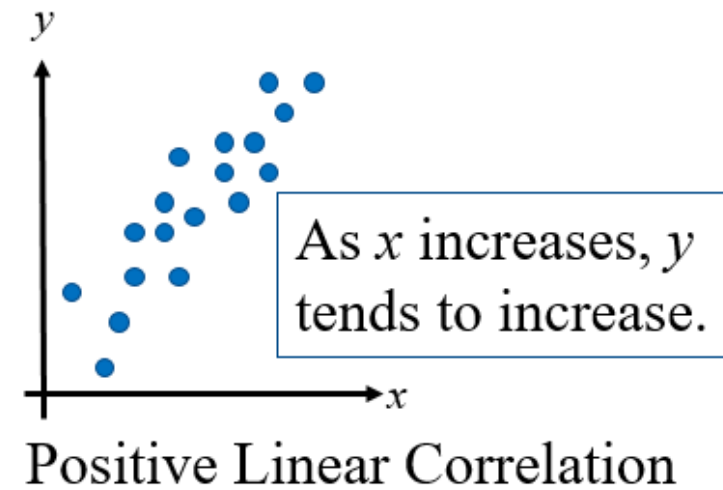
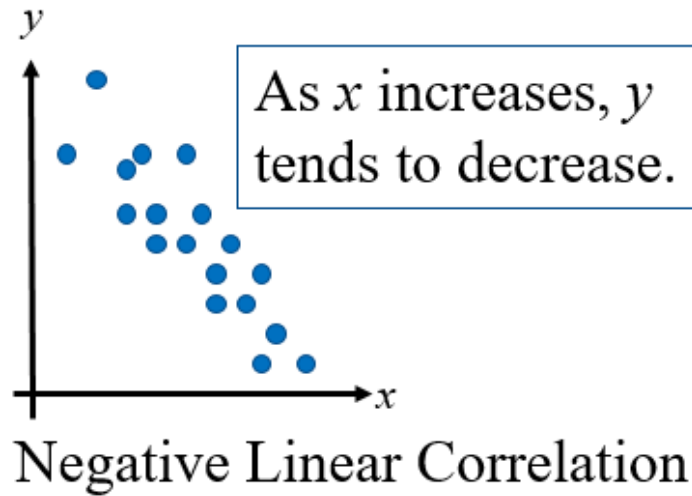
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

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



Types of Correlation



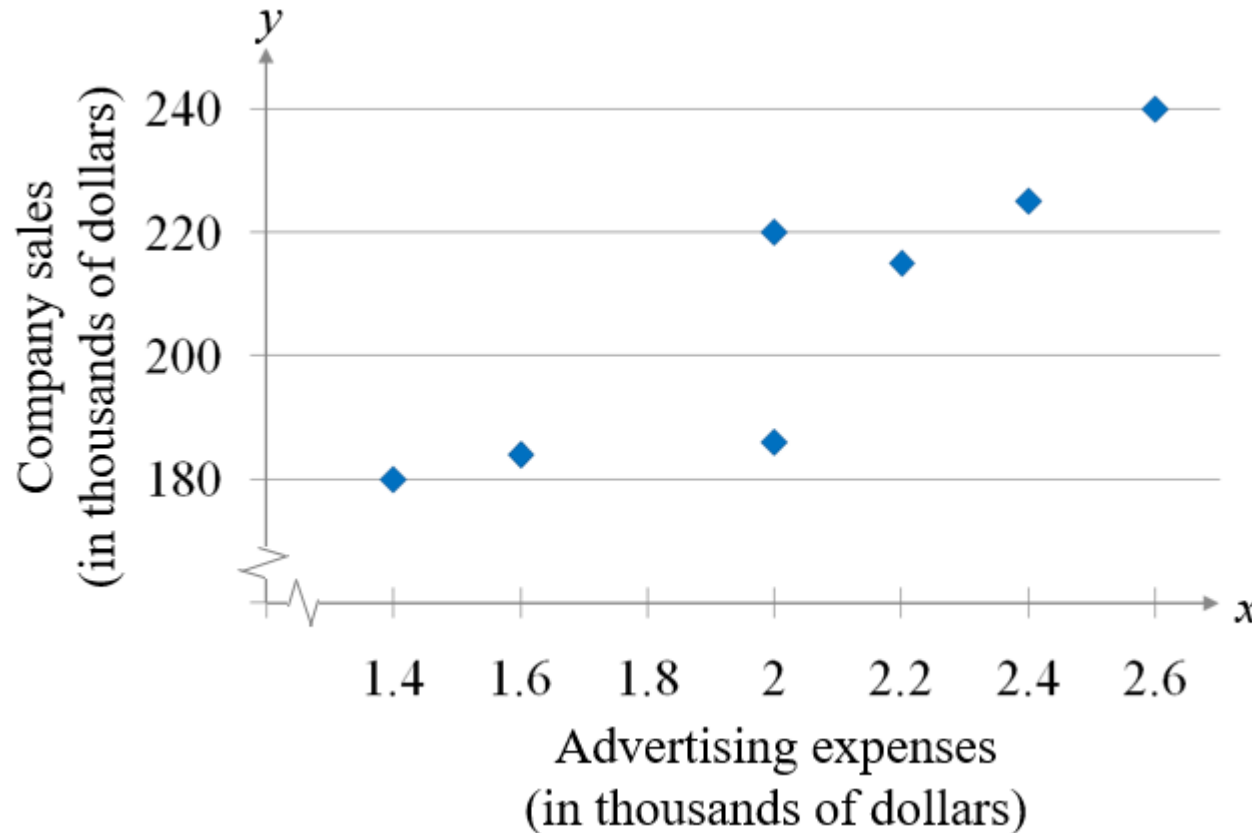
Example: Constructing a Scatter Plot

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

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

The symbol r represents the sample correlation coefficient.

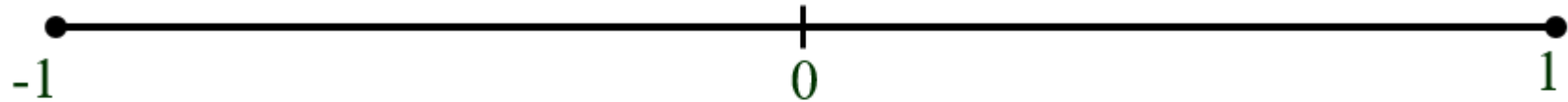
A formula for r is

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

n is the number
of data pairs

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

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



If $r = -1$ there is a perfect negative correlation



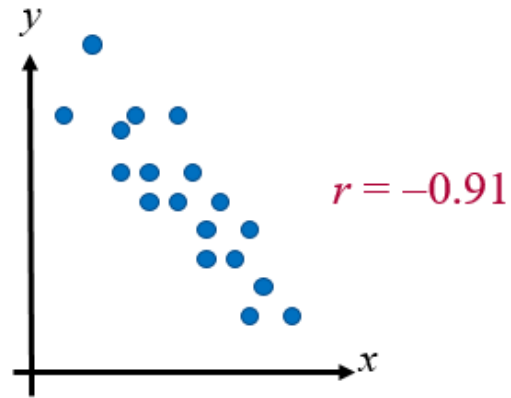
If r is close to 0 there is no linear correlation



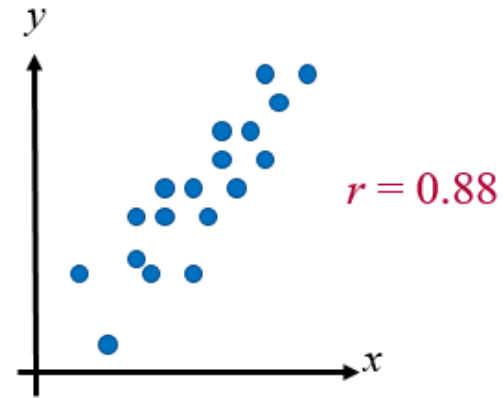
If $r = 1$ there is a perfect positive correlation



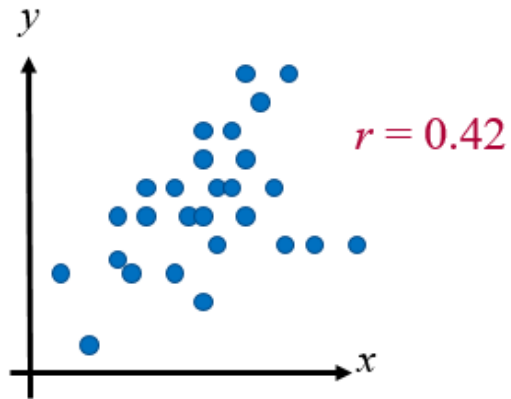
Linear Correlation



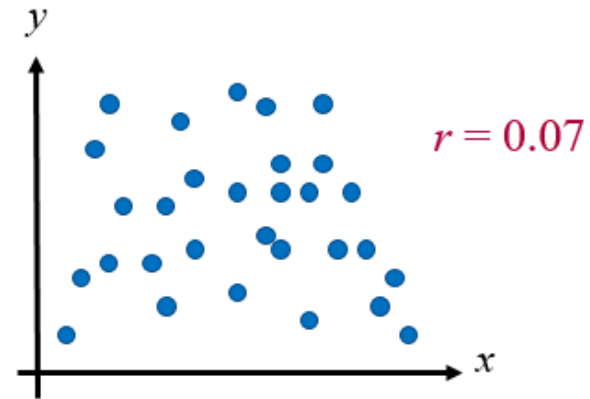
Strong negative correlation



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

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

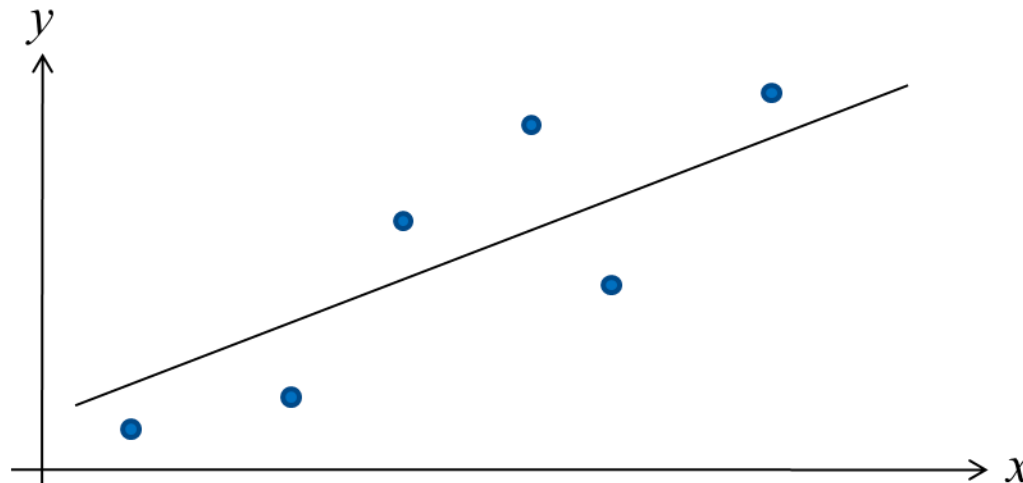
$$\begin{aligned} r &= \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} \\ &= \frac{8(3289.8) - (15.8)(1634)}{\sqrt{8(32.44) - 15.8^2} \sqrt{8(337,558) - 1634^2}} \\ &= \frac{501.2}{\sqrt{9.88} \sqrt{30,508}} \approx 0.9129 \end{aligned}$$

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

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

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 (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

- $\hat{y} = mx + b$ 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 $(\bar{x}, \bar{y})$

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

Recall from section 9.1:

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 Equation of a Regression Line

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

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

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

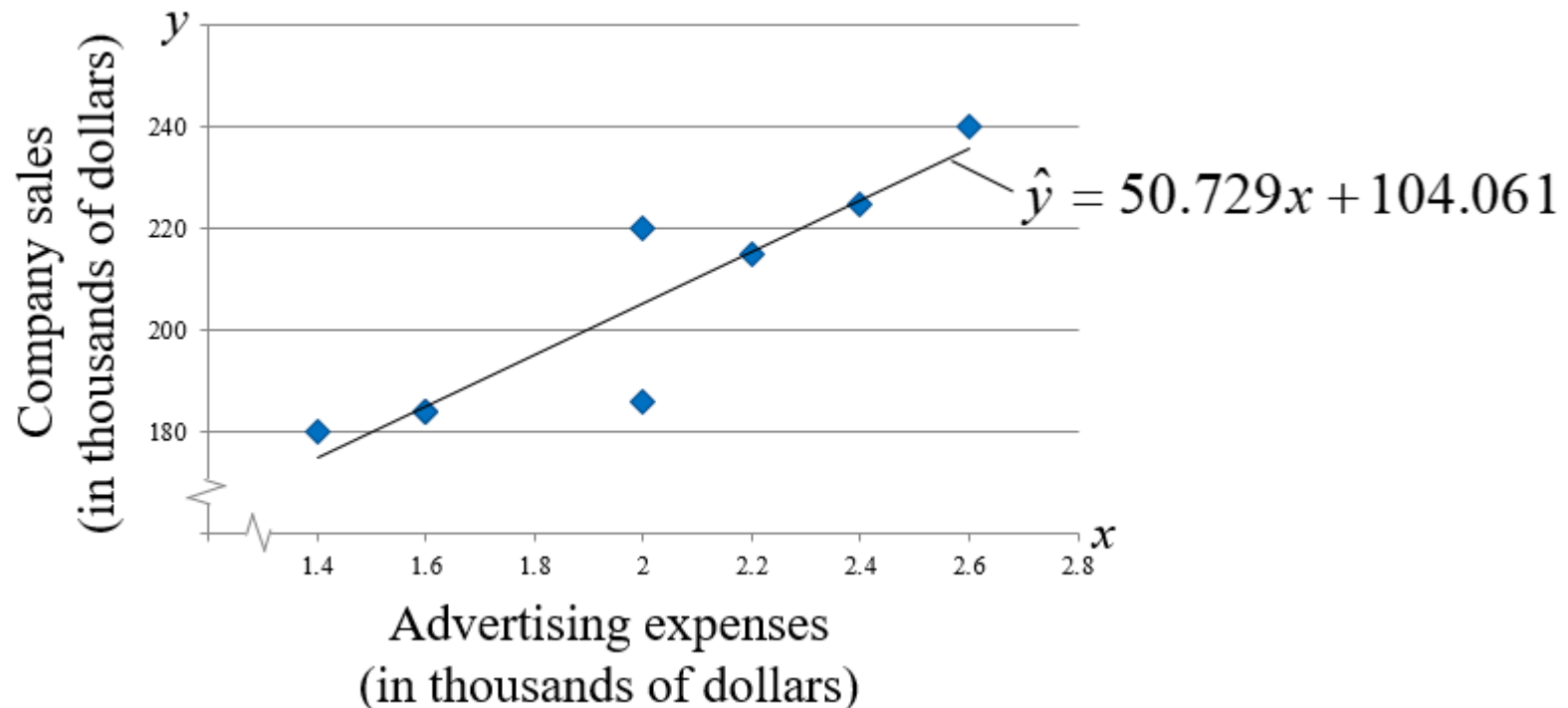
$$b = \bar{y} - m\bar{x} = \frac{1634}{8} - (50.72874)\frac{15.8}{8}$$

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

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

Solution: Finding the Equation of a Regression Line

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.



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

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

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.

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).

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**

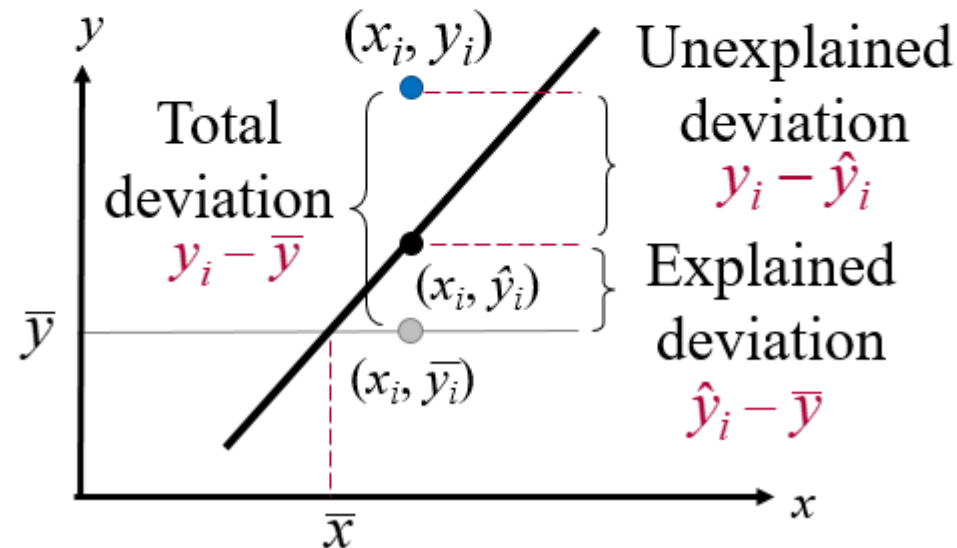
- The **explained deviation**

- The **unexplained deviation**

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

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

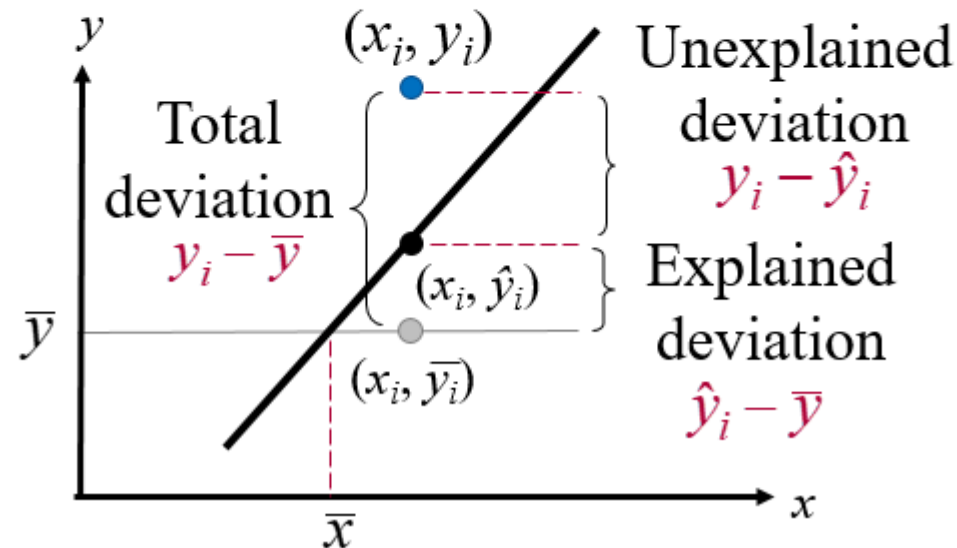
Unexplained Deviation = $y_i - \hat{y}_i$



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

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

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



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 .

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

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

- 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.

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.

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: 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: Standard Error of Estimate

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

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

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

Multiple regression equation

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \dots + m_kx_k$$

$x_1, x_2, x_3, \dots, x_k$ are independent variables

b is the y-intercept

y is the dependent variable

- * Because the mathematics associated with this concept is complicated, technology is generally used to calculate the multiple regression equation.

Example: Finding a Multiple Regression Equation

A researcher wants to determine how employee salaries at a certain company are related to the length of employment, previous experience, and education. The researcher selects eight employees from the company and obtains the data shown on the next slide. Use Minitab to find a multiple regression equation that models the data.

Employee	Salary, y	Employment (yrs), x_1	Experience (yrs), x_2	Education (yrs), x_3
A	57,310	10	2	16
B	57,380	5	6	16
C	54,135	3	1	12
D	56,985	6	5	14
E	58,715	8	8	16
F	60,620	20	0	12
G	59,200	8	4	18
H	60,320	14	6	17

Solution: Finding a Multiple Regression Equation

MINITAB

Regression Analysis: Salary, y versus x1, x2, x3

The regression equation is

$$\text{Salary, } y = 49764 + 364 x_1 + 228 x_2 + 267 x_3$$

Predictor	Coef	SE Coef	T	P
Constant	49764	1981	25.12	0.000
x1	364.41	48.32	7.54	0.002
x2	227.6	123.8	1.84	0.140
x3	266.9	147.4	1.81	0.144

S = 659.490 R-Sq = 94.4% R-Sq(adj) = 90.2%

The regression equation is

$$\hat{y} = 49,764 + 364x_1 + 228x_2 + 267x_3$$

Example: Predicting y-Values

Use the regression equation:

$$\hat{y} = 49,764 + 364x_1 + 228x_2 + 267x_3$$

to predict an employee's salary given 12 years of current employment, 5 years of experience, and 16 years of education.

Solution:

$$\begin{aligned}\hat{y} &= 49,764 + 364(12) + 228(5) + 267(16) \\ &= 59,544\end{aligned}$$

The employee's predicted salary is \$59,544.

9.1 Correlation

9.2 Linear Regression

9.3 Measures of Regression and Prediction Intervals

9.4 Multiple Regression