MATH6703: Applied Regression Analysis [1]

Dr. Md. Rezaul Karim
PhD(KU Leuven & UHasselt), MS(Biostatistics, UHasselt), MS(Statistics, JU)
Professor, Department of Statistics and Data Science
Jahangirnagar University (JU), Savar, Dhaka - 1342, Bangladesh
Mobile: 01912605556, Email: rezaul@juniv.edu

MS in Mathematics - 2024

[1] These course slides should not be reproduced nor used by others (without permission).

Introduction

1 Introduction

1.1 Learning Outcomes of the Course

1.2 Text and Reference List

Learning Outcomes of the Course

1. Understand fundamental concepts of correlation and regression analysis.
2. Learn to build simple and multiple regression models.
3. Understand assumptions and diagnostics of correlation and regression.
4. Explore practical applications in various fields.
5. Learn advanced regression models such as logistic regression, Poisson regression, Polynomial regression, Generalized Linear Model (GLM), Generalized Additive Models (GAM), Mixed effect model, Random effect models, etc.
6. Develop critical thinking skills in result interpretation.
7. Gain proficiency in statistical software for analysis.

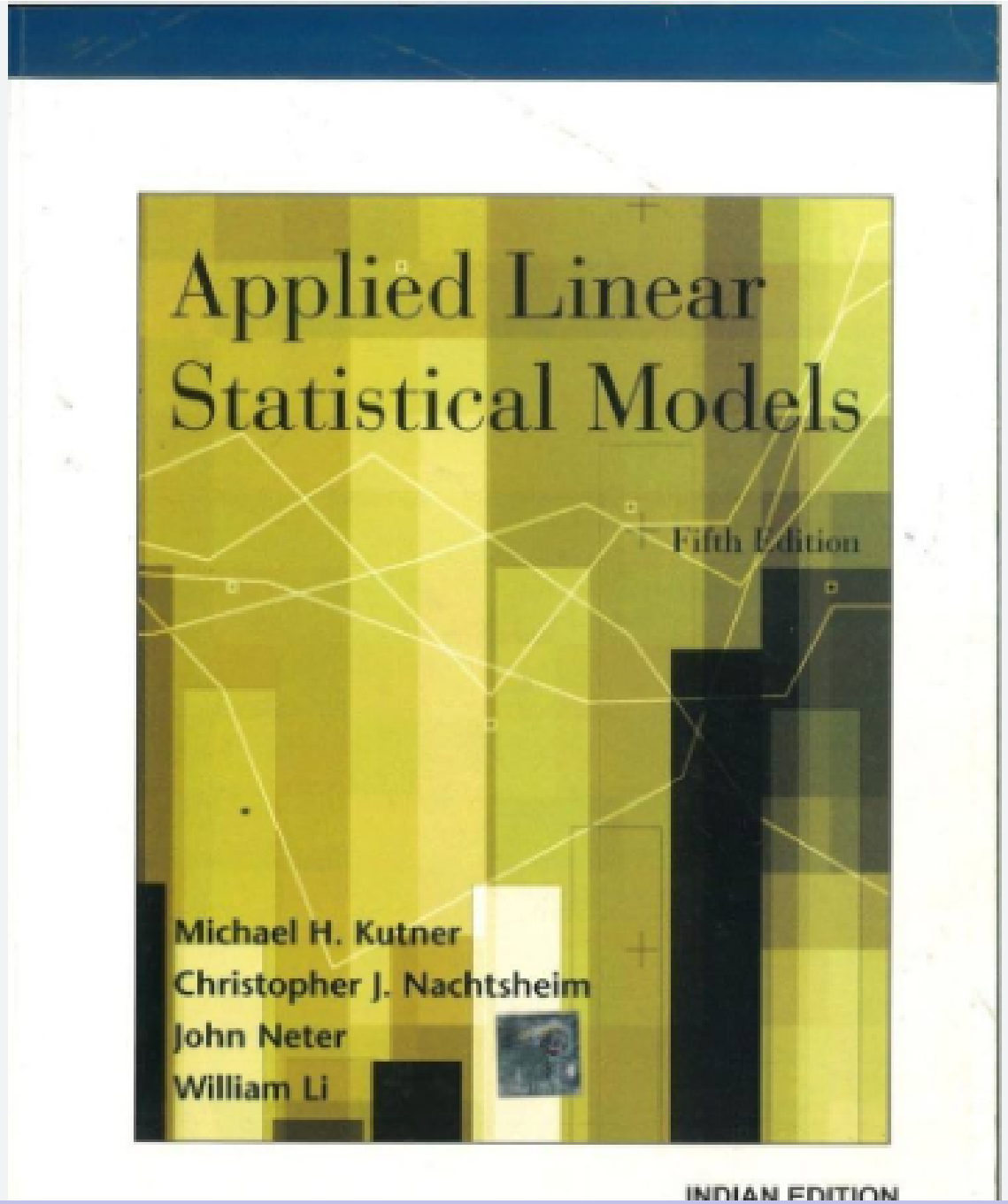
Text and Reference List

Text Book

1. Michael Kutner, Christopher Nachtsheim, John Neter, William Li (2004): *Applied Linear Statistical Models*. 5th edition. New York: McGraw-Hill/Irwin.
2. Gujarati, Damodar N. and Dawn C. Porter (2012): *Basic Econometrics*, 5th Edition. McGraw-Hill.

Reference list

1. Michael H. Kutner, Chris Nachtsheim, John Neter (2004): *Applied Linear Regression Models*. 5th edition. New York: McGraw-Hill/Irwin.
2. Greene, W. H. (2003): *Econometric Analysis*. Pearson Education.


Applied Linear
Statistical Models
Fifth Edition
Michael H. Kutner
Christopher J. Nachtsheim
John Neter
William Li
INDIAN EDITION

Lecture Outline I

Lecture Outline II

Lecture Outline III

Lecture Outline IV

Lecture Outline V

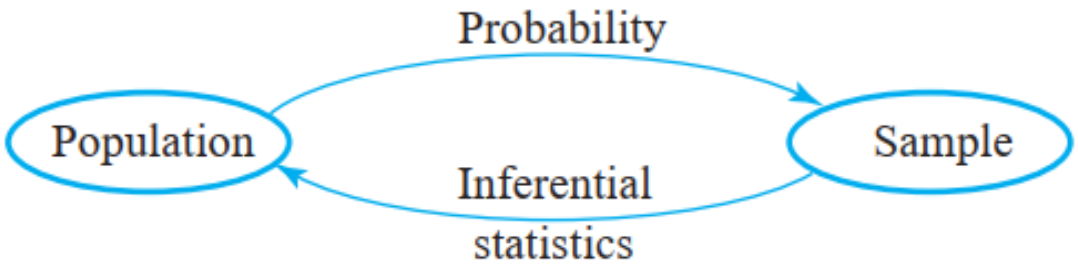
Lecture Outline VI

Chapter 1: Correlation and Association Analysis

2 Chapter 1: Correlation and Association Analysis

Statistical methods

- can be used to summarize or describe a collection of statistical methods statistics data; this is called *descriptive statistics*
- allows to make predictions ("inferences") from that data; this is called *inferential statistics*



The relationship between probability and inferential statistics

Parameter & Statistic

- measurable of a population characteristics ⇒ parameter
- measurable of a sample characteristics ⇒ statistic

Variable & it's types

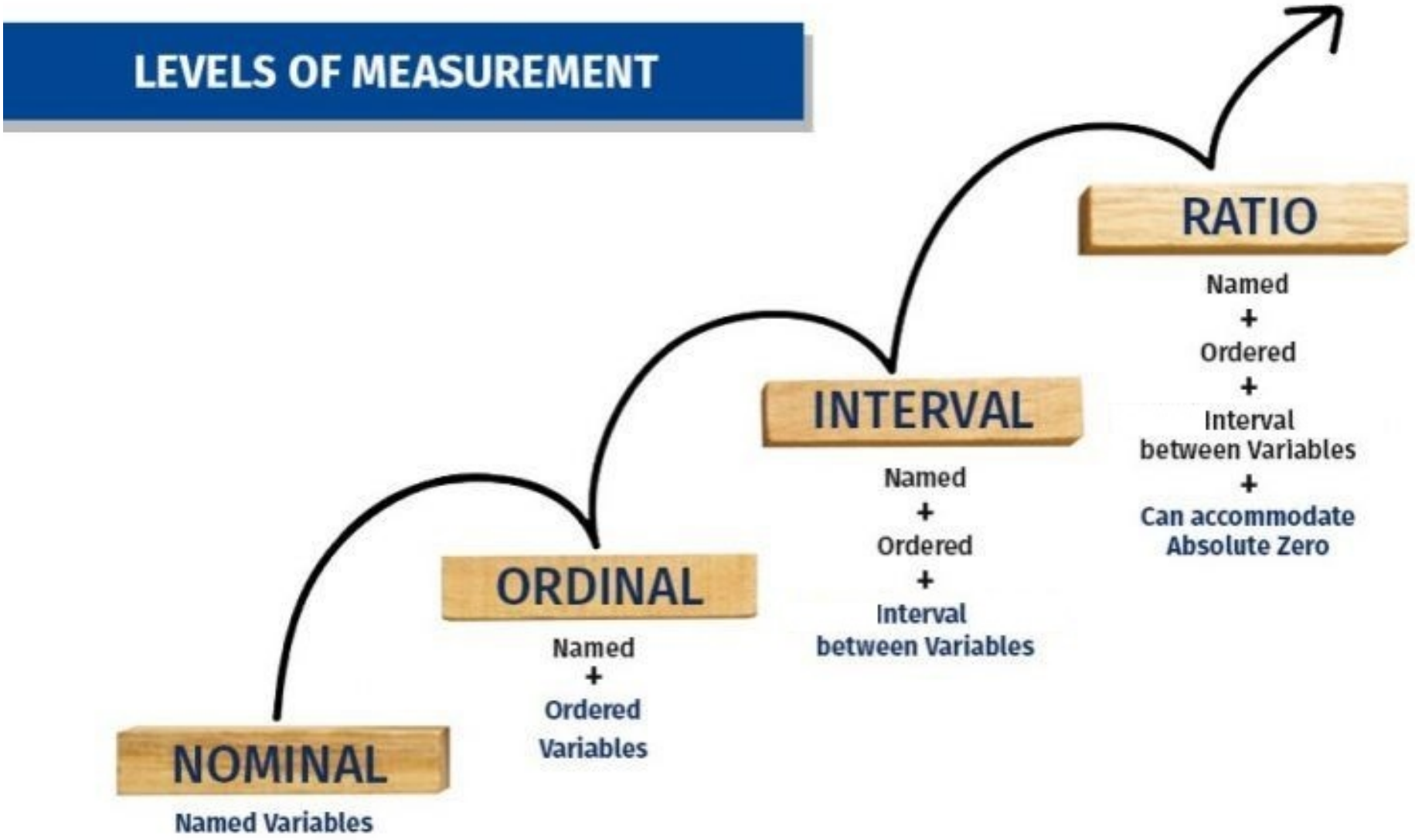
Variable

- is an attribute or characteristics of interest that vary or change respondent to respondent
- it is also called feature, or factor. In the context of statistics and data analysis, variables can be divided into two main types:
  - independent variables
  - dependent variables.
- Example: Suppose we are conducting a study to investigate the effect of studying time on exam scores. In this study, "studying time" is the independent variable because it is manipulated by the students themselves, and "exam scores" are the dependent variable because they are measured in response to changes in studying time. Both studying time and exam scores are examples of variables.

Types of Variable

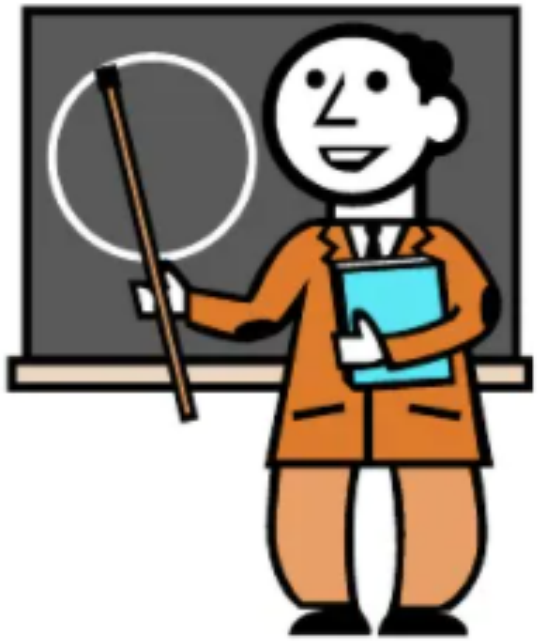
- qualitative variable (also known as categorical variable)
- quantitative variable (also known as numerical variable)
  - discrete variable
  - continuous variable

Level or Scales of Measurement



Nominal Scale

- Assign responses to different categories
- No numerical difference between categories
- Examples
  - Gender
  - Marital status
  - State of residence
  - College major
  - SSN
  - Zip code
  - Student ID



Ordinal Scale

- Set of categories that are ordered from least to most
- Don't know numerical distance from each category to the next
- Examples
  - Miss America results – first place, runner-up, second, third
  - Military rank
  - Letter grade in class
  - Degrees held
  - Medical condition (satisfactory, serious, guarded, critical)
  - Rank order of your preference from 1 to 4 of Ruffles, Doritos, Cheetos, Fritos

Interval Scale

- Scale with values, and there is the same numerical distance between each value
- This scale has an arbitrary zero point (no true meaningful zero point)
- Examples
  - How appealing is this cereal box to children?

    Not at all very

    -3 -2 -1 0 1 2 3

    
  - Current temperature
  - Many behavioral science questionnaires
  - IQ

Ratio Scale

- Scale with scores where there is the same numerical distance between each score
- The scale has a true, meaningful zero point that anchors the scale
- Only scale that allows you to make ratio comparisons, such as "Maribel's income is 35% more than Susan's"
- Examples
  - Weight of a package of candy
  - Number of times you return to a restaurant after visiting it the first time
  - Amount of money in your checking account
  - Number of questions correct on a quiz
  - Distance from San Antonio to Laredo

Summarizing bivariate data

- Tabular Method-Crosstabulation
  - A tabular summary of data for two variables. The classes for one variable are represented by the rows; the classes for the other variable are represented by the columns.
- Graphical method- scatter diagram

Example: Test Performances of MAT 101

ID	Gender	Test Performance	Study Hour	Score of STAT
1	Male	Good	10	71
2	Female	Good	11	75
3	Male	Excellent	14	85
4	Female	Excellent	10	90
5	Male	Poor	8	50
6	Female	Excellent	13	88
7	Male	Poor	6	45
8	Female	Excellent	15	80
9	Male	Good	14	65
10	Female	Good	10	82
11	Male	Excellent	14	92
12	Female	Poor	8	55
13	Male	Poor	5	40
14	Male	Good	10	68
15	Female	Good	9	62

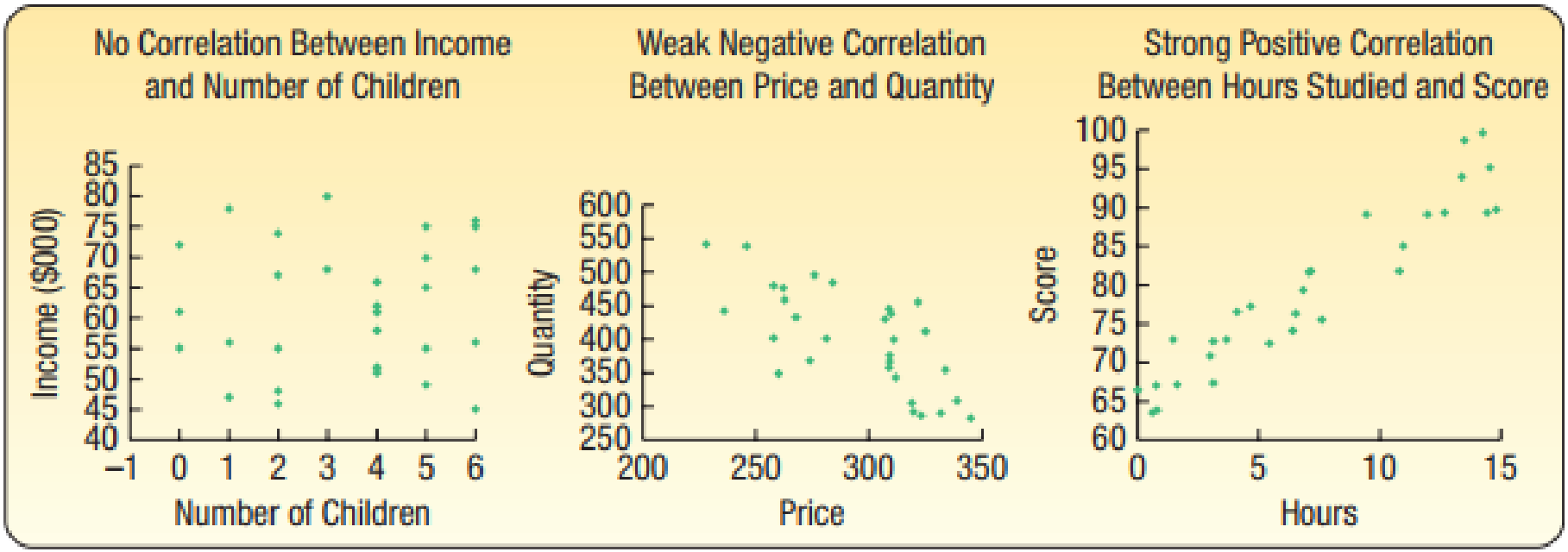
Table 1: The Crosstabulation of Gender and Test Performance.

Gender	Test Performance			Total
	Poor	Good	Excellent	
Male	3	3	2	8
Female	1	3	3	7
Total	4	6	5	15

Scatter Diagram

a scatter diagram is a graphic tool used to portray the relationship between two variables

- the dependent variable is scaled on the Y-axis and is the variable being estimated
- the independent variable is scaled on the X-axis and is the variable used as the predictor



Covariance is a statistical measure that quantifies the degree to which two random variables vary together. In other words, it assesses the relationship between two variables, indicating whether they tend to move in the same direction (positive covariance) or opposite directions (negative covariance). A positive covariance suggests that when one variable increases, the other variable tends to increase as well, while a negative covariance indicates that when one variable increases, the other tends to decrease. Sample covariance is denoted by s_{xy} and calculated by the following formula

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Example: Sample Covariance

Using a dataset containing the number of commercial advertisements aired and the corresponding sales volume for a product or service over a specific time period:

Table 2: Sample Data for the San Francisco Electronics Store

Week	Number of commercials (X)	Sales Volume ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

(i). Draw a scatter diagram of the data points representing the relationship between the number of commercial advertisements and the sales volume.

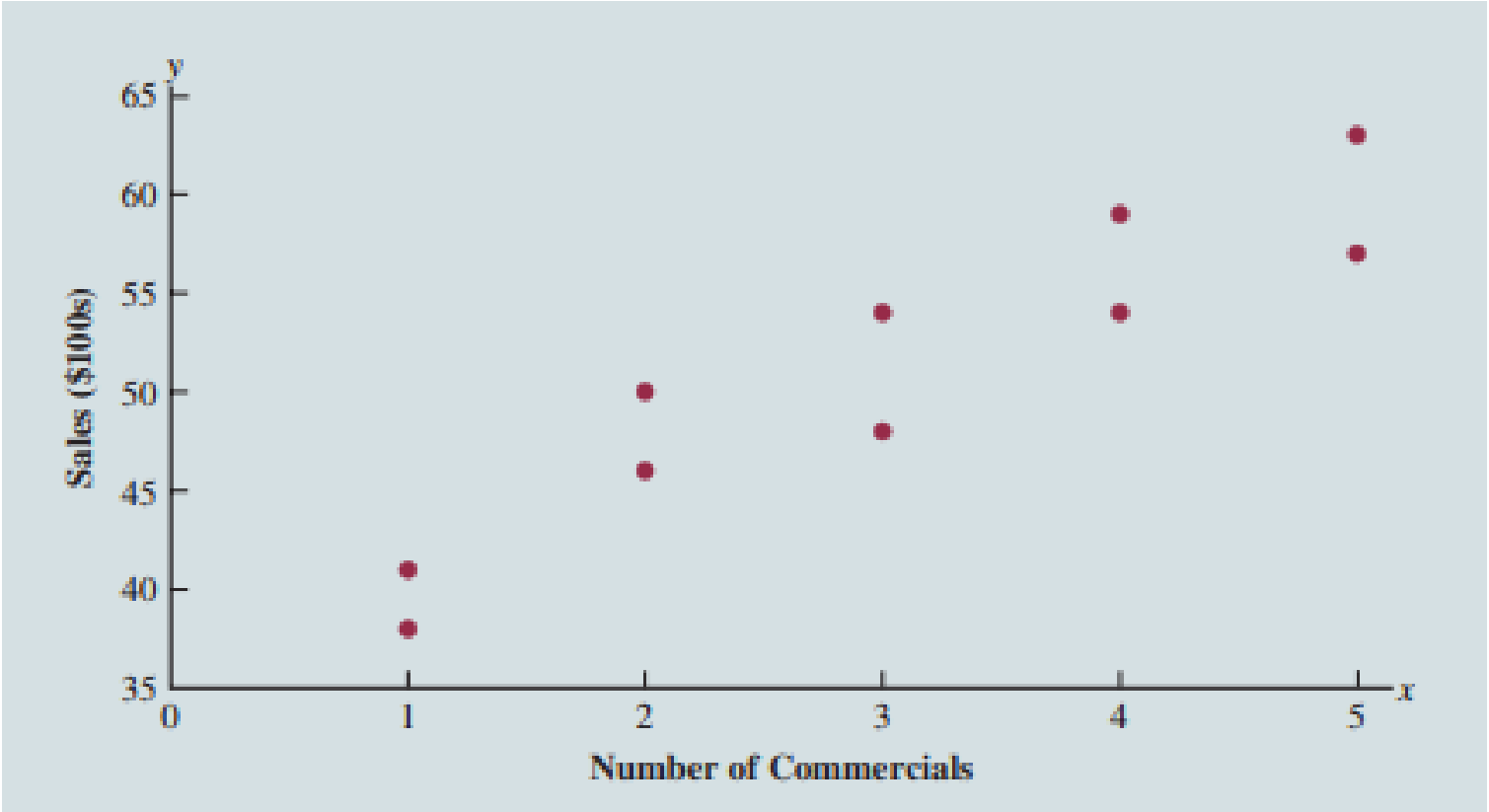
(ii). Based on the scatter diagram, describe the observed trend or pattern in the data.

(iii). Calculate the sample covariance between the number of commercial advertisements and the sales volume to quantitatively assess the degree of association between these two variables.

(iv). Interpret the sample covariance value in terms of the strength and direction of the relationship between the number of commercial advertisements and the sales volume.

Sample Covariance

- what is the nature of the relationship between x and y?



- sample covariance:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

TABLE 3.7 Calculations for the Sample Covariance

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

Correlation Analysis

Correlation analysis is a group of techniques to measure *the strength and direction of the relationship* between two variables. There are different types of correlation coefficients that can be used depending on the nature of the variables being analyzed and the assumptions of the data. Some of the common types of correlation coefficients include:

1. **Pearson correlation coefficient**: Measures linear relationship between two continuous variables.
2. **Spearman rank correlation coefficient**: Measures association between ranked variables.
3. **Kendall tau correlation coefficient**: Measures similarity of orderings of data pairs.
4. **Point-biserial correlation coefficient**: Measures association between continuous and binary variables.
5. **Phi coefficient**: Measures association between two binary variables.
6. **Cramér's V**: Measures association between two nominal variables.

Pearson's Correlation coefficient

- is a measure of the strength of the linear relationship between two variables
- is denoted by r and defined as

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

where $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

Working Formula

this correlation coefficient is called the Pearson's correlation coefficient whicgh is also defined as

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

where $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

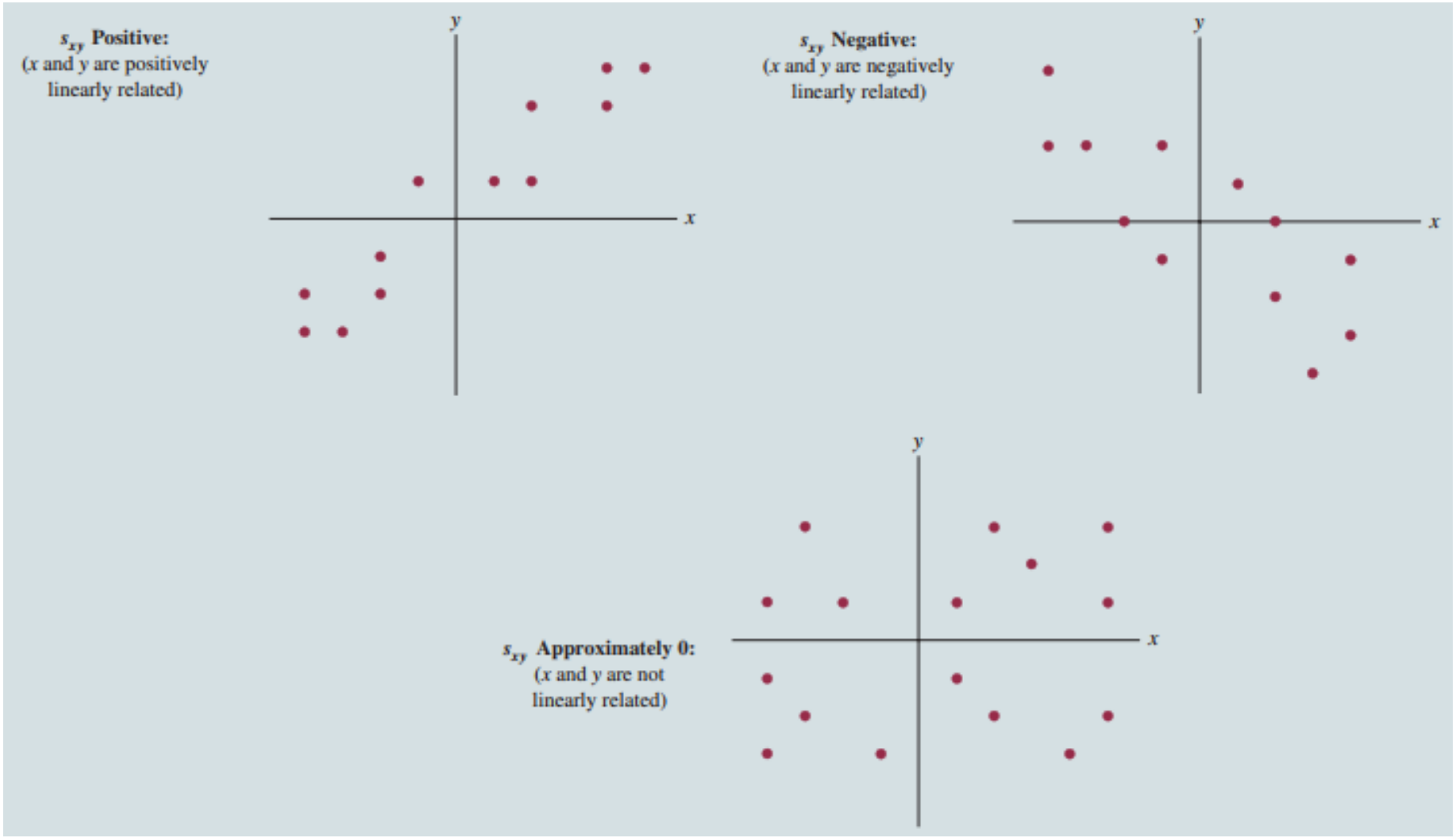
Let us now compute the sample correlation coefficient for the San Francisco electronics store. Using the data in Table 3.6, we can compute the sample standard deviations for the two variables:

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49$$

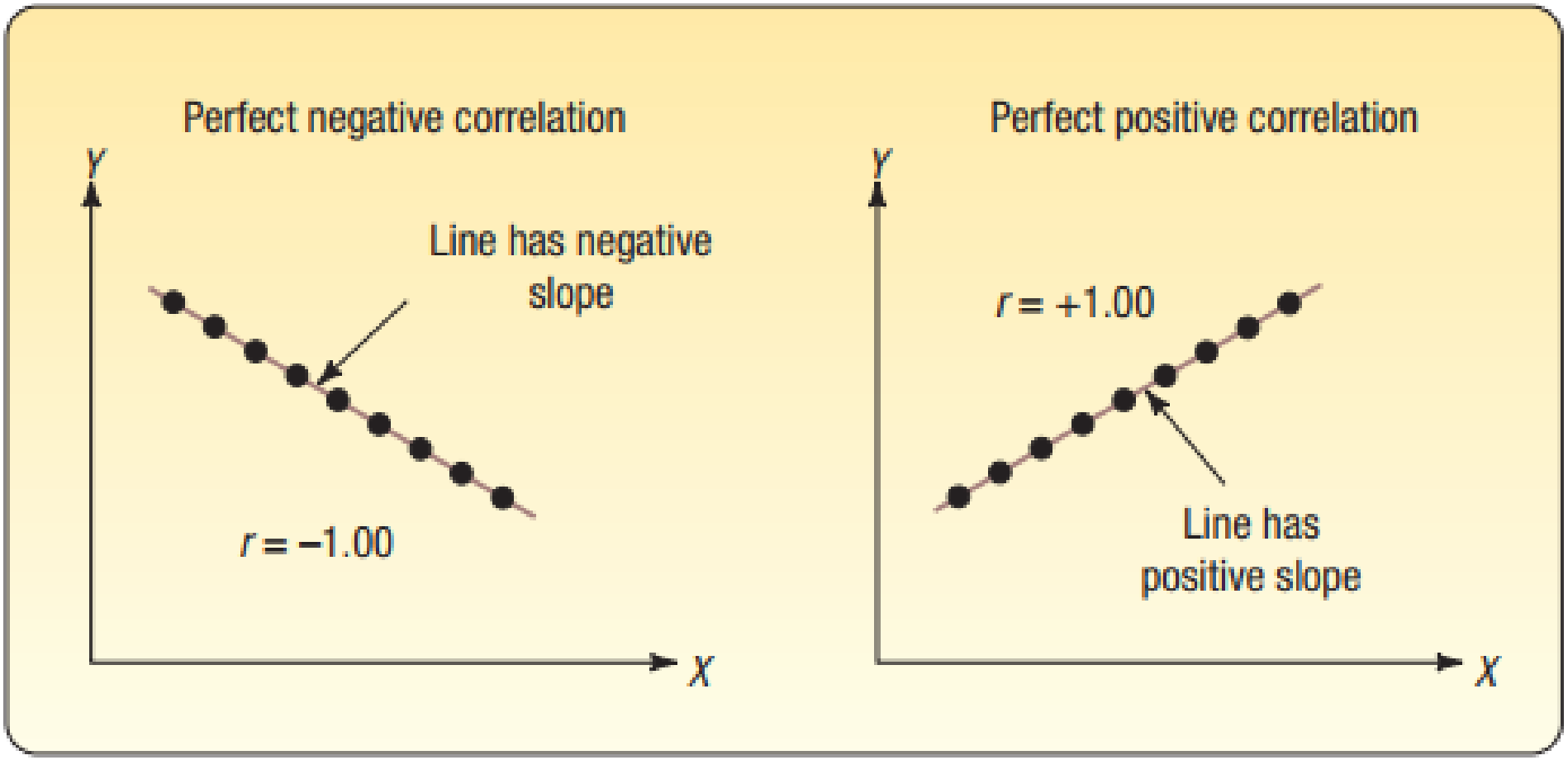
$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

Now, because $s_{xy} = 11$, the sample correlation coefficient equals

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = .93$$

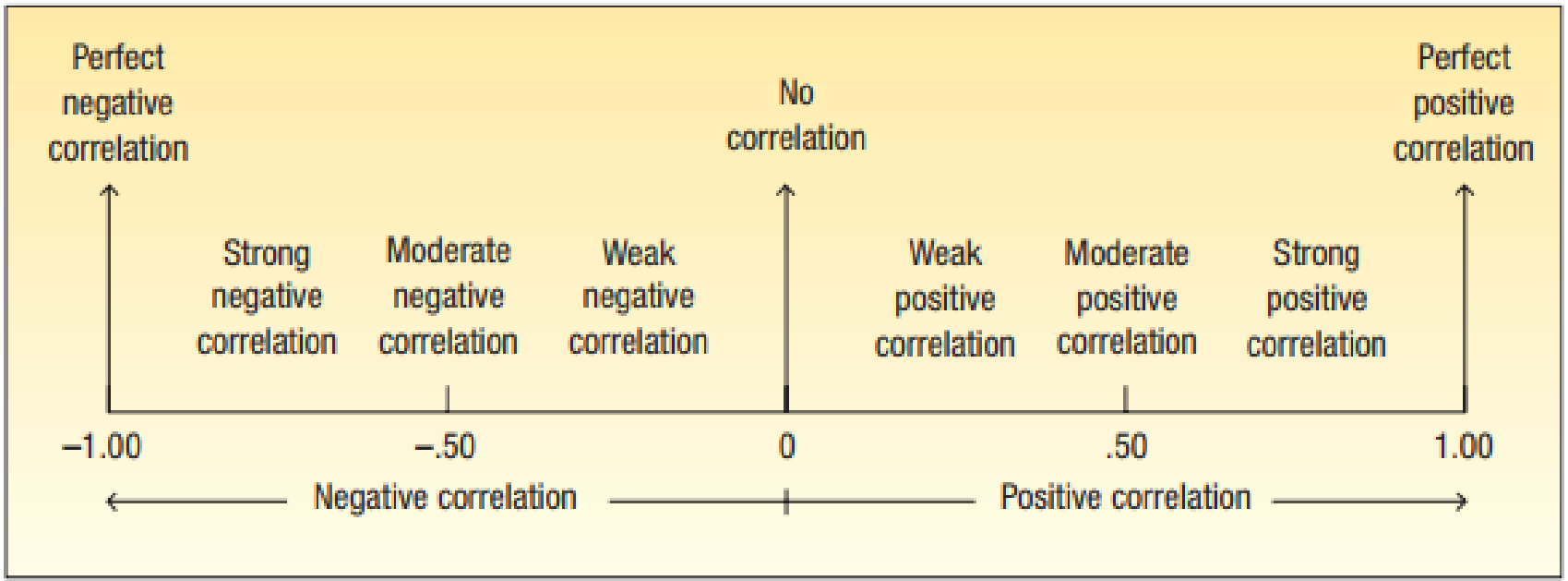

s_{xy} Positive:
(x and y are positively linearly related)
y
x
s_{xy} Negative:
(x and y are negatively linearly related)
y
x
s_{xy} Approximately 0:
(x and y are not linearly related)
y
x

Correlation Coefficient



Correlation Coefficient

the following drawing summarizes the strength and direction of the correlation coefficient



Characteristics of correlation coefficient

1. the sample correlation coefficient is identified by the lowercase letter r
2. it shows the direction and strength of the linear relationship between two interval or ratio-scale variables
3. it ranges from -1 up to and including $+1$
4. a value near 0 indicates there is little linear relationship between the variables
5. a value near 1 indicates a direct or positive linear relationship between the variables
6. a value near -1 indicates an inverse or negative linear relationship between the variables

Exercises

1. The following sample of observations were randomly selected.

x	4	5	3	6	10
y	4	6	5	7	7

(a). Draw a scatter diagram.
(b). Determine the correlation coefficient and interpret the relationship between x and y.
(c). Interpret these statistical measures.

Solution (b)

x	x	x^2	y^2	xy
4	4	16	16	16
5	6	25	36	30
3	5	9	25	15
6	7	36	49	42
10	7	100	49	70
$\sum x_i = 28$	$\sum y_i = 29$	$\sum x_i^2 = 186$	$\sum y_i^2 = 175$	$\sum x_i y_i = 173$
$\bar{x} = 5.6$	$\bar{y} = 5.8$			

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} = \frac{173 \times -5 \times 5.6 \times 5.8}{\sqrt{186 - 5 \times (5.6)^2}\sqrt{175 - 5 \times (5.8)^2}} = 0.7522$$

Exercises

2. The owner of Maumee Ford-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Car	Age (years)	Selling Price ($000)	Car	Age (years)	Selling Price ($000)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

(a). Draw a scatter diagram.
(b). Determine the correlation coefficient.
(c). Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

Testing the Significance of the Correlation Coefficient

1. hypothesis

$$H_0 : \rho = 0 \text{ (the correlation in the population is 0)}$$
$$H_1 : \rho \neq 0 \text{ (the correlation in the population is not 0)}$$

2. level of significance α
3. reject H_0 if

$$T > t_{\alpha/2,n-2} \quad \text{or} \quad T < -t_{\alpha/2,n-2}$$

4. test statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t \text{ distribution with } n-2 \text{ degrees of freedom}$$

5. decision

in general, if the null hypothesis is

$$H_0 : \rho = 0$$

if the null hypothesis is true, the test statistic T follows the student's-t distribution with $(n-2)$ degrees of freedom, i.e., $T \sim t(n-2)$

Table 3: decision rule for the test of hypothesis $H_0 : \rho = 0$

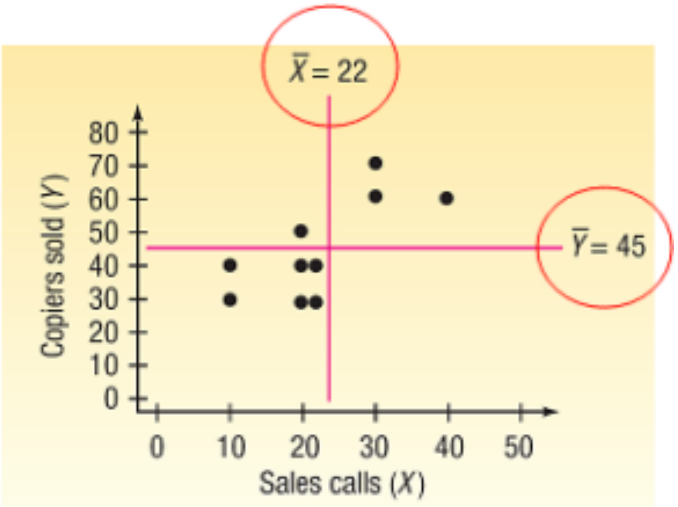
Alternative hypothesis	Reject H_0 if
$H_1 : \rho < 0$	$T < -t_{\alpha,n-2}$
$H_1 : \rho > 0$	$T > t_{\alpha,n-2}$
$H_1 : \rho \neq 0$	$T > t_{\alpha/2,n-2}$ or $T < -t_{\alpha/2,n-2}$

Correlation Coefficient - Example

Using the formula:

CORRELATION COEFFICIENT

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n-1)s_x s_y}$$



Sales Representative	Calls, Y	Sales, X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramirez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

Correlation Coefficient - Example

	A	B	C	D	E	F	G	H
1		Sales Representative	Sales Calls (x)	Copiers Sold (y)			Sales Calls (x)	Copiers Sold (y)
2		Brian Virost	96	41		Mean	96.00	45.00
3		Carlos Ramirez	40	41		Standard Error	11.04	3.33
4		Carol Saia	104	51		Median	84.00	43.00
5		Greg Fish	128	60		Mode	84.00	41.00
6		Jeff Hall	164	61		Standard Deviation	42.76	12.89
7		Mark Reynolds	76	29		Sample Variance	1828.57	166.14
8		Meryl Rumsey	72	39		Kurtosis	-0.32	-0.73
9		Mike Kiel	80	50		Skewness	0.46	0.36
10		Ray Snarsky	36	28		Range	144.00	42.00
11		Rich Niles	84	43		Minimum	36.00	28.00
12		Ron Broderick	180	70		Maximum	180.00	70.00
13		Sal Spina	132	56		Sum	1440.00	675.00
14		Soni Jones	120	45		Count	15.00	15.00
15		Susan Welch	44	31				
16		Tom Keller	84	30				
17		Total	1440	675				

We now insert these values into formula (13–1) to determine the correlation coefficient:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y} = \frac{6672}{(15-1)(42.76)(12.89)} = 0.865$$

Correlation Coefficient - Example

H_0: $\rho = 0$ (the correlation in the population is 0)

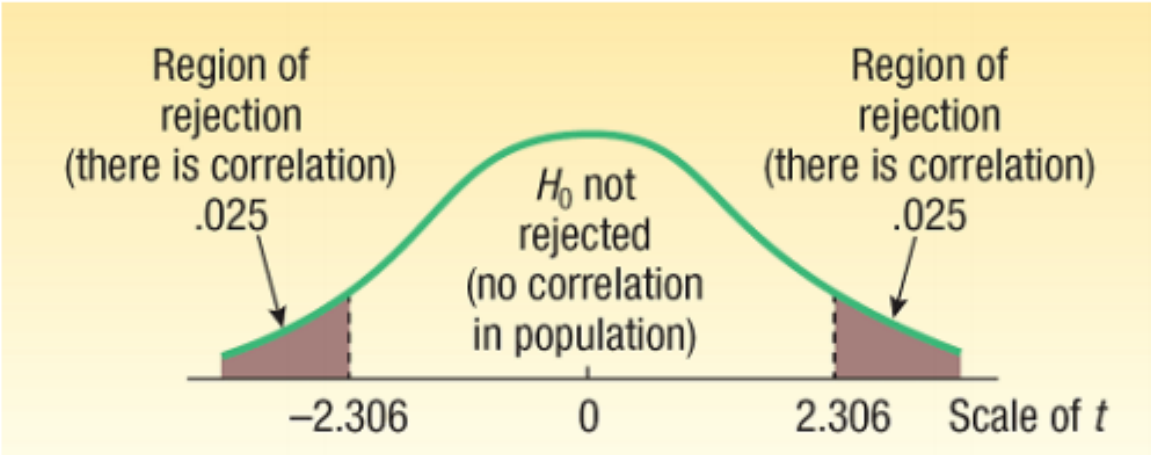
H_1: $\rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

$t > t_{0.025,8}$ or $t < -t_{0.025,8}$

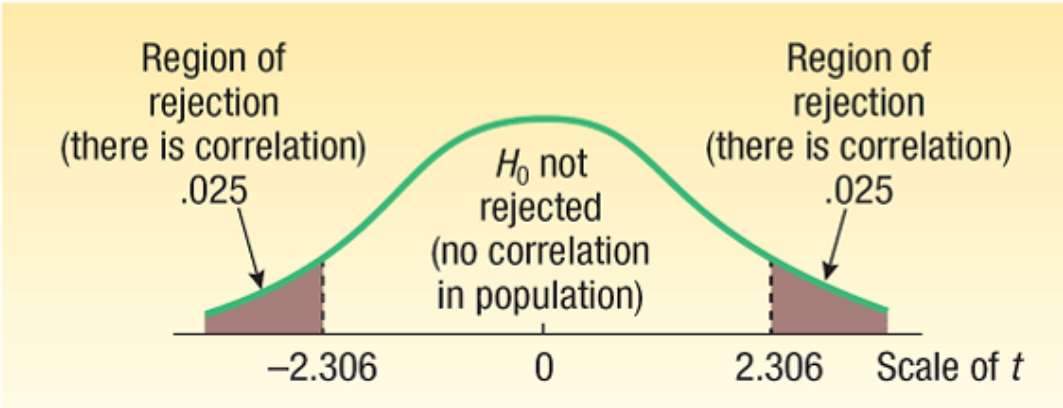
$t > 2.306$ or $t < -2.306$



Correlation Coefficient - Example

Computing t, we get

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$



The computed $t(3.297)$ is within the rejection region, therefore, we will reject H_0. This means the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

Exercises

7. The following hypotheses are given.

$$H_0: \rho \leq 0$$
$$H_1: \rho > 0$$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

$$H_0: \rho \geq 0$$
$$H_1: \rho < 0$$

A random sample of 15 paired observations has a correlation of −.46. Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?

10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?

11. The Airline Passenger Association studied the relationship between the number of passengers on a particular flight and the cost of the flight. It seems logical that more passengers on the flight will result in more weight and more luggage, which in turn will result in higher fuel costs. For a sample of 15 flights, the correlation between the number of passengers and total fuel cost was .667. Is it reasonable to conclude that there is positive association in the population between the two variables? Use the .01 significance level.

R Code

```
# Example data (you should replace this with your actual data)
data <- data.frame(
x1 = c(1, 2, 3, 4, 5),
x2 = c(2, 3, 4, 5, 6),
x3 = c(3, 4, 5, 6, 7)
)

# Compute correlation matrix
correlation_matrix <- cor(data)
print(correlation_matrix)
# Compute correlation matrix with 2 decimal places
correlation_matrix <- round(cor(data), 2)

# View the correlation matrix
print(correlation_matrix)
```

Python Code: Correlation Matrix

```
import pandas as pd

# Example data (you should replace this with your actual data)
data = pd.DataFrame({
  'x1': [1, 2, 3, 4, 5],
  'x2': [2, 3, 4, 5, 6],
  'x3': [3, 4, 5, 6, 7]
})

# Compute correlation matrix
correlation_matrix = data.corr()

print(correlation_matrix)
# Compute correlation matrix with 2 decimal places
correlation_matrix = data.corr().round(2)

# View the correlation matrix
print(correlation_matrix)
```

Spearman Rank Correlation

- is measured the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable
- If r_i, s_i are the ranks of the i-member according to the x and the y-quality respectively, then the rank correlation coefficient is

$$r_R = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of data points of the two variables and di is the difference in the ranks of the ith element of each random variable considered, i.e. $d_i = r_i - s_i$, is the difference between ranks

- the Spearman correlation coefficient, r_R, can take values from +1 to −1

Question (Q1)

Based on the following data, find the rank correlation between marks of English and Mathematics courses.

English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

Solution: The procedure for ranking these scores is as follows:

English (mark)	Maths (mark)	Rank (Engilsh)	Rank (Maths)	d_i	d_i^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	-3	9
62	65	6	5	1	1
64	56	5	9	-4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	-1	1
61	63	7	6	1	1

- the realized value of the Spearman Rank Correlation is

$$r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 54}{10(10^2-1)} = 0.6727$$

This indicates a strong positive relationship between the ranks individuals obtained in the Maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Equal Ranks or Tie in Ranks

- for tie observations, the rank correlation can be computed by adjusting $m(m^2-1)/12$ to the value of $\sum d_i^2$, where m stands for the number of items whose ranks are equal
- if there are more than one such group of items with common rank, this value is added as many times as the number of such groups
- then the formula for the rank correlation is

$$r_R = 1 - \frac{6\{\sum d_i^2 + \frac{m_1}{12}(m_1^2-1) + \frac{m_2}{12}(m_2^2-1) + \cdots\}}{n(n^2-1)}$$

Question (Q2)

Based on the following data, Find the rank correlation between marks of English and Mathematics courses.

English	56	75	45	71	61	64	58	80	76	61
Maths	70	70	40	60	65	56	59	70	67	80

Solution: The procedure for ranking these scores is as follows:

English (mark)	Maths (mark)	Rank (Engilsh)	Rank (Maths)	d_i	d_i^2
56	70	9	3	6	36
75	70	3	3	0	0
45	40	10	10	0	0
71	60	4	7	-3	9
61	65	6.5	6	0.5	0.25
64	56	5	9	-4	16
58	59	8	8	0	0
80	70	1	3	-2	4
76	67	2	5	-3	9
61	80	6.5	1	5.5	30.25

- the mark 61 is repeated 2 times in series X (in English) and hence $m_1 = 2$
- in series Y (in Math), the marks 70 occurs 3 times and hence $m_2 = 3$
- so the rank correlation is

$$r_R = 1 - \frac{6\{104.5 + \frac{2}{12}(2^2 - 1) + \frac{3}{12}(3^2 - 1) + \cdots\}}{10(10^2 - 1)}$$

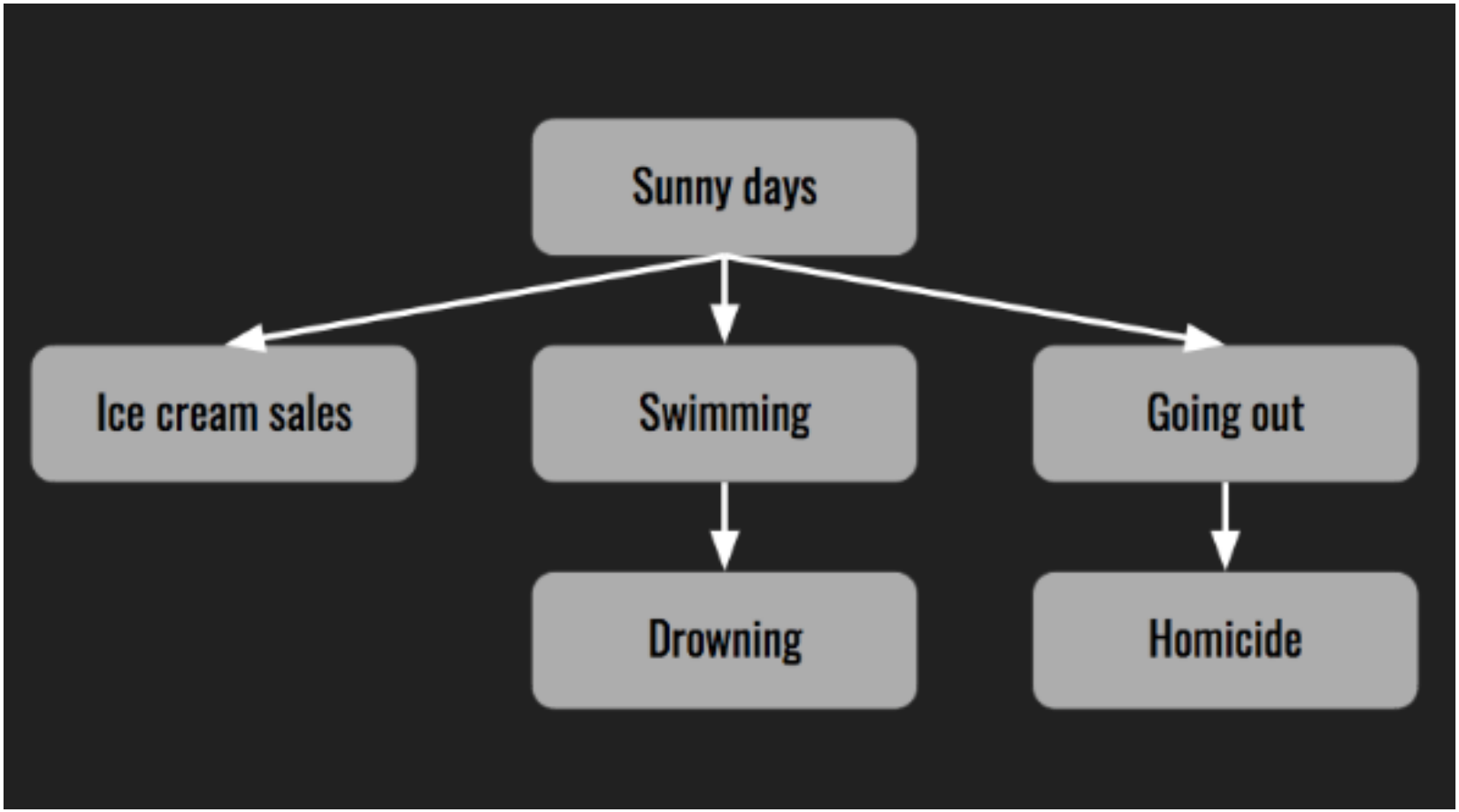
$$= 0.3515$$

Correlation is not causation!

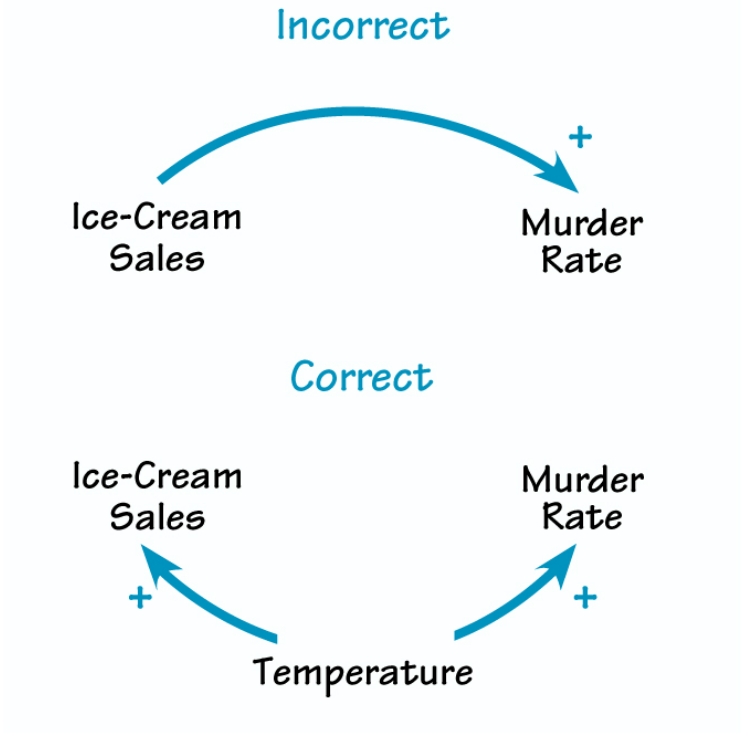
Correlation is not causation!

- a study showed that the ice cream sales is correlated with homicides in New York
  - as the sales of ice cream rise and fall, so do the number of homicides. Does the consumption of ice cream causing the death of the people?
  - No – two things are correlated doesn't mean one causes other

Consider underlying factors before conclusion



Don't conclude too fast!



- there is no causal relationship between the ice cream and rate of homicide, sunny weather is bringing both the factors together
- and yes, ice cream sales and homicide has a causal relationship with weather

R Code: Rank Correlation

```r
# Example data (you should replace this with your actual
    data)
x <- c(1, 2, 3, 4, 5)
y <- c(2, 3, 1, 5, 4)

# Calculate Spearman rank correlation coefficient
correlation <- cor(x, y, method = "spearman")
print(correlation)

# Alternatively, if you have data in a dataframe, you can
    use the cor() function directly on the dataframe:
# Example dataframe (you should replace this with your
    actual data)
data <- data.frame(
x = c(1, 2, 3, 4, 5),
y = c(2, 3, 1, 5, 4)
)

# Calculate Spearman rank correlation coefficient
correlation <- cor(data, method = "spearman")
print(correlation)
```

Python Code: Rank Correlation

In Python, you can find rank correlation using the scipy.stats module, which provides a function called spearmanr() to calculate the Spearman rank correlation coefficient. Here's how you can do it:

```
from scipy.stats import spearmanr
# Example data (you should replace this with your actual
    data)
x = [1, 2, 3, 4, 5]
y = [2, 3, 1, 5, 4]

# Calculate Spearman rank correlation coefficient
rho, p_value = spearmanr(x, y)
print("Spearman rank correlation coefficient:", rho)
print("p-value:", p_value)
```

Python Code: Rank Correlation

If you have two columns of data in a pandas DataFrame, you can calculate the rank correlation directly from the DataFrame. Here's an example:

```
# From Data file
import pandas as pd
# Example DataFrame (you should replace this with your
    actual data)
data = pd.DataFrame({
  'x': [1, 2, 3, 4, 5],
  'y': [2, 3, 1, 5, 4]
})

# Calculate Spearman rank correlation coefficient
rho, p_value = data.corr(method='spearman').iloc[0, 1]

print("Spearman rank correlation coefficient:", rho)
print("p-value:", p_value)
```

Kendall Tau Correlation Coefficient

Definition

The Kendall tau correlation coefficient, denoted by τ, measures the similarity of the orderings of data pairs.

Formula:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs.

- Suitable for ordinal data.
- Useful when dealing with tied ranks.

Example: Tau correlation coefficient

Suppose we have the following data on two variables, X and Y, with their corresponding ranks:

Observation	Rank
X	Y
10	15
15	10
20	20
25	25
30	40

Using this data, we can calculate the Kendall Tau correlation coefficient as follows:

- **Concordant Pairs:**
  In the context of correlation coefficients such as Kendall Tau, concordant pairs refer to pairs of observations where the ranks for both variables follow the same order. In other words, if (X_i, Y_i) and (X_j, Y_j) are two pairs of observations, they are considered concordant if both $X_i < X_j$ and $Y_i < Y_j$ or if both $X_i > X_j$ and $Y_i > Y_j$.
- **Discordant Pairs:** Discordant pairs, on the other hand, refer to pairs of observations where the ranks for the variables have opposite orders. In other words, if (X_i, Y_i) and (X_j, Y_j) are two pairs of observations, they are considered discordant if $X_i < X_j$ and $Y_i > Y_j$ or if $X_i > X_j$ and $Y_i < Y_j$.
  In the context of calculating correlation coefficients like Kendall Tau, understanding concordant and discordant pairs is crucial as they form the basis for determining the strength and direction of association between two variables based on their ranks.

Calculation of Kendall Tau

To calculate the Kendall Tau correlation coefficient:

1. Compare each pair of observations in terms of their ranks.
2. Determine whether they have the same order (concordant) or opposite order (discordant) for both variables.
3. Count the total number of concordant pairs (n_c) and discordant pairs (n_d).
4. Calculate the Kendall Tau coefficient using the formula:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Example Calculation

Suppose we have the following data on two variables, X and Y, with their corresponding ranks:

Observation	Rank
X	Y
10	15
15	10
20	20
25	25
30	40

Let's calculate the Kendall Tau correlation coefficient.

Concordant and Discordant Pairs

To find the number of concordant and discordant pairs in the given data, we need to compare each pair of observations in terms of their ranks and determine whether they are concordant or discordant.
Let's denote the observations as (X_i, Y_i) and (X_j, Y_j), where $i < j$.
A pair (X_i, Y_i) and (X_j, Y_j) is:

- **Concordant** if $X_i < X_j$ and $Y_i < Y_j$, or if $X_i > X_j$ and $Y_i > Y_j$.
- **Discordant** if $X_i < X_j$ and $Y_i > Y_j$, or if $X_i > X_j$ and $Y_i < Y_j$.

Let's analyze each pair:

1. $(10, 15)$ and $(15, 10)$: Discordant
2. $(10, 15)$ and $(20, 20)$: Concordant
3. $(10, 15)$ and $(25, 25)$: Concordant
4. $(10, 15)$ and $(30, 40)$: Concordant
5. $(15, 10)$ and $(20, 20)$: Concordant
6. $(15, 10)$ and $(25, 25)$: Concordant
7. $(15, 10)$ and $(30, 40)$: Concordant
8. $(20, 20)$ and $(25, 25)$: Concordant

9. $(20, 20)$ and $(30, 40)$: Concordant
10. $(25, 25)$ and $(30, 40)$: Concordant

So, out of the 10 pairs of observations, there are 9 concordant pairs and 1 discordant pair.

- Number of concordant pairs (n_c): 9
- Number of discordant pairs (n_d): 1
- Total number of pairs (n): $\frac{5(5-1)}{2} = 10$
- Kendall Tau coefficient (τ):

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} = \frac{9-1}{\frac{1}{2}(10)(9)} = \frac{8}{45} \approx 0.1778$$

The Kendall Tau correlation coefficient for the given data is $\tau = \frac{8}{45} \approx 0.1778$.

Point-Biserial Correlation Coefficient

Definition

The point-biserial correlation coefficient, denoted by r_{pb}, measures the strength and direction of association between a continuous variable and a binary variable. It is calculated by using

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_0}{s_X \sqrt{\frac{p(1-p)}{n}}}$$

where:

- $\overline{X}_1$ is the mean exam score for one group (e.g., males).
- $\overline{X}_0$ is the mean exam score for the other group (e.g., females).
- s_X is the standard deviation of the exam scores.
- p is the proportion of one group (e.g., proportion of males).
- n is the total sample size.
- Applicable when one variable is dichotomous.

Question

Suppose we are interested in examining the relationship between students' exam scores and their gender (male/female). We have the following data:

Student	Exam Score	Gender
1	85	Male
2	70	Female
3	90	Male
4	75	Female
5	80	Male
6	65	Female
7	95	Male
8	85	Female
9	88	Male
10	82	Female

Calculate the point-biserial correlation coefficient to assess the relationship between exam scores and gender.

Solution

Using the given data:

- $\overline{X}_1 = 87.6$ (mean exam score for males)
- $\overline{X}_0 = 75.4$ (mean exam score for females)
- $s_X = 8.67$ (standard deviation of exam scores)
- $p = 0.5$ (proportion of males in the sample)
- $n = 10$ (total sample size)

Substituting these values into the formula, we find:

$$r_{pb} = \frac{87.6 - 75.4}{8.67\sqrt{\frac{0.5\times(1-0.5)}{10}}} \approx 0.76$$

Therefore, the point-biserial correlation coefficient is approximately 0.76, indicating a strong positive relationship between exam scores and gender.

Phi Coefficient

Definition

The phi coefficient, denoted by ϕ, measures the association between between two dichotomous variables. It is essentially a special case of the Pearson correlation coefficient, specifically applicable when both variables are binary (i.e., they have only two possible values).

We have the following contingency table for two categorical variables:

	Variable 2		
Variable 1	Variable 2 = 0	Variable 2 = 1	Total
Variable 1 = 0	n_{00}	n_{01}	n_{0+}
Variable 1 = 1	n_{10}	n_{11}	n_{1+}
Total	n_{+0}	n_{+1}	n_{++}

Here, n_{00} represents the frequency of observations with Variable 1 being in category 0 and Variable 2 being in category 0, and so on.

Phi Coefficient with an example

We can then calculate the phi coefficient using the formula:

$$\phi = \frac{n_{00}n_{11} - n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$$

- Similar to Pearson correlation.
- Both variables are dichotomous.

Suppose we have data on two binary variables, Variable 1 and Variable 2, and we want to calculate the phi coefficient to measure the association between them. We have the following contingency table:

	Variable 2 = 0	Variable 2 = 1
Variable 1 = 0	20	30
Variable 1 = 1	40	10

9. Calculate the marginal frequencies:

$$n_{1+} = 40 + 10 = 50$$
$$n_{0+} = 20 + 30 = 50$$
$$n_{+1} = 30 + 10 = 40$$
$$n_{+0} = 20 + 40 = 60$$

10. Calculate the frequencies:

$$n_{00} = 20$$
$$n_{01} = 30$$
$$n_{10} = 40$$
$$n_{11} = 10$$

Using this contingency table, we can calculate the phi coefficient as follows:

1. Substitute the frequencies into the formula:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$$

2. Calculate the phi coefficient:

$$\phi = \frac{(10)(20) - (40)(30)}{\sqrt{(50)(50)(40)(60)}} = \frac{200 - 1200}{\sqrt{50000 \times 2400}}$$
$$= \frac{-1000}{\sqrt{120000000}} \approx -0.3162$$

So, the phi coefficient for this example is approximately -0.3162, indicating a moderate negative association between Variable 1 and Variable 2.

Cramér's V

Definition

Cramér's V is a measure of the strength of association between two nominal variables. It is defined as

$$V = \sqrt{\frac{\chi^2}{n(min(r,c)-1)}} = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$$

where

$$\chi^2 = \sum_i^r \sum_j^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

is the chi-square statistic, O_{ij} is the observed frequency in cell (i,j) of the contingency table. E_{ij} is the expected frequency in cell (i,j) of the contingency table. n is the total number of observations, and r and c are the number of rows and columns in the contingency table respectively.

- Ranges from 0 to 1.
- 0 indicates no association, 1 indicates perfect association.

Computation of Cramér's V

Suppose we have the following contingency table:

	Variable 1		
Variable 2	A	B	Total
C	20	30	50
D	10	40	50
Total	30	70	100

Step 1: Calculate the chi-square statistic

First, compute the expected frequencies for each cell:

$$E_{11} = \frac{50 \times 30}{100} = 15$$

$$E_{12} = \frac{50 \times 70}{100} = 35$$

$$E_{21} = \frac{50 \times 30}{100} = 15$$

$$E_{22} = \frac{50 \times 70}{100} = 35$$

Then, compute the chi-square statistic:

$$\begin{aligned}\chi^2 &= \frac{(20-15)^2}{15} + \frac{(30-35)^2}{35} + \frac{(10-15)^2}{15} + \frac{(40-35)^2}{35} \\ &= \frac{25}{15} + \frac{25}{35} + \frac{25}{15} + \frac{25}{35} \\ &= \frac{5}{3} + \frac{5}{7} + \frac{5}{3} + \frac{5}{7} \\ &= \frac{70}{21} \approx 3.333\end{aligned}$$

Step 2: Calculate Cramér's V

$$V = \sqrt{\frac{3.333}{100 \times (\min(2,2)-1)}} = \sqrt{\frac{3.333}{100}} \approx 0.577$$

So, in this example, Cramér's V is approximately 0.577. This value indicates a moderate to strong association between the two categorical variables. Therefore, we can interpret that there is a notable relationship between Variable 1 and Variable 2 in the given contingency table.

The Kappa Statistic

- The Kappa Statistic or Cohen's Kappa statistic is a statistical measure used to quantify the level of agreement between two raters (or judges, observers, surveys, etc.) who each classify items into categories.
- It assesses the level of agreement between two raters beyond what would be expected by chance.
- Kappa takes into account both observed agreement and agreement expected by chance.
- It is widely used in various fields to assess the reliability of ratings or classifications made by multiple raters.

The formula for Cohen's Kappa is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where:

- P_o is the proportion of observed agreement.
- P_e is the proportion of agreement expected by chance.

	Judgment 2		
Judgment 1	Yes	No	Total
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	n

where

$$P_o = \frac{a+d}{a+b+c+d}$$

$$P_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d}$$

Similarly:

$$P_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

Finally,

$$P_e = P_{\text{Yes}} + P_{\text{No}}$$

Interpretation of Cohen's Kappa

- $\kappa = 1$: Perfect agreement between raters.
- $\kappa = 0$: Agreement no better than chance.
- $\kappa < 0$: Agreement worse than chance.
- $0 < \kappa < 0.2$: Slight agreement.
- $0.2 \leq \kappa < 0.4$: Fair agreement.
- $0.4 \leq \kappa < 0.6$: Moderate agreement.
- $0.6 \leq \kappa < 0.8$: Substantial agreement.
- $0.8 \leq \kappa$: Almost perfect agreement.

Example of Cohen's Kappa

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the disagreement count data were as follows

	Readers 1		
Readers 2	Present	Absent	Total
Present	20	5	25
Absent	10	15	25
Total	30	20	50

Calculate Cohen's Kappa to assess the agreement between the two doctors.

The observed proportionate agreement is:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

To calculate pe (the probability of random agreement) we note that: Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time. Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time. So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

Similarly:

$$P_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$P_e = P_{\text{Yes}} + P_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

So, with a kappa value of 0.4, there is a moderate level of agreement beyond what would be expected by chance between the two readers. While it's not perfect agreement, it still suggests a meaningful level of consensus between them.

Python Code: The Kappa Statistic

```
# Observed agreement
num_agreements = 20 + 15
total_obs = 50
P_o = num_agreements / total_obs

# Expected agreement
P_present_reader1 = 25 / total_obs
P_absent_reader1 = 25 / total_obs
P_present_reader2 = 30 / total_obs
P_absent_reader2 = 20 / total_obs

P_e_present = P_present_reader1 * P_present_reader2
P_e_absent = P_absent_reader1 * P_absent_reader2
P_e = P_e_present + P_e_absent

# Cohen's Kappa
kappa = (P_o - P_e) / (1 - P_e)
print("Cohen's Kappa:", kappa)
```

Another Example

```
from sklearn.metrics import cohen_kappa_score

# Example data
doctor1_ratings = [3, 2, 1, 5, 4, 2, 3, 4, 1, 5]
doctor2_ratings = [4, 2, 1, 5, 3, 2, 3, 4, 2, 5]

# Calculate Kappa Statistic
kappa = cohen_kappa_score(doctor1_ratings, doctor2_ratings)
print("Kappa Statistic:", kappa)

# Alternatively,
import statsmodels.api as sm

# Example data
doctor1_ratings = [3, 2, 1, 5, 4, 2, 3, 4, 1, 5]
doctor2_ratings = [4, 2, 1, 5, 3, 2, 3, 4, 2, 5]

# Calculate Kappa Statistic
kappa = sm.stats.inter_rater.cohens_kappa(doctor1_ratings,
    doctor2_ratings)
print("Kappa Statistic:", kappa)
```

Partial Correlation

Partial correlation is a statistical technique used to measure the strength and direction of the relationship between two variables while controlling for the influence of one or more additional variables. The formula for calculating partial correlation coefficient (denoted as $r_{xy.z}$) between variables X and Y while controlling for variable Z is given by:

$$r_{xy.z} = \frac{r_{xy} - (r_{xz} \times r_{zy})}{\sqrt{(1 - r_{xz}^2)(1 - r_{zy}^2)}}$$

Where:

- r_{xy}: Correlation coefficient between variables X and Y
- r_{xz}: Correlation coefficient between variables X and Z
- r_{zy}: Correlation coefficient between variables Z and Y

Partial correlation helps in understanding the unique association between variables after accounting for the effects of other variables.

Example: Partial Correlation

Suppose we have three variables: X, Y, and Z. We want to calculate the partial correlation coefficient $r_{xy.z}$ between X and Y while controlling for Z.

Given correlation coefficients:

$$r_{xy} = 0.6, \quad r_{xz} = 0.4, \quad r_{zy} = 0.3$$

We can use the formula:

$$r_{xy.z} = \frac{r_{xy} - (r_{xz} \times r_{zy})}{\sqrt{(1 - r_{xz}^2)(1 - r_{zy}^2)}}$$

Substituting the given values:

$$r_{xy.z} = \frac{0.6 - (0.4 \times 0.3)}{\sqrt{(1 - 0.4^2)(1 - 0.3^2)}} = \frac{0.48}{\sqrt{0.84 \times 0.91}} \approx 0.549$$

So, the partial correlation coefficient between X and Y while controlling for Z is approximately 0.549.

Suppose we are interested in the association between two variables X and Y but want to control for other covariates $Z_1, Z_2, \ldots, Z_k$. The partial correlation is defined to be the Pearson correlation between two derived variables ϵ_x and ϵ_y, where

- ϵ_x = the residual from the linear regression of X on $Z_1, Z_2, \ldots, Z_k$
- ϵ_y = the residual from the linear regression of Y on $Z_1, Z_2, \ldots, Z_k$.

However, we are also often interested in the association between Y and all the predictors when considered as a group. This measure of association is given by the multiple correlation.

Python Code: The Partial Correlation

```
pip install pingouin
import pandas as pd
import pingouin as pg

# Sample data
data = {
  'X': [1, 2, 3, 4, 5],
  'Y': [2, 3, 5, 4, 6],
  'Z': [5, 4, 3, 2, 1]
}

# Create a DataFrame
df = pd.DataFrame(data)

# Calculate partial correlation between X and Y controlling for Z
partial_corr_result = pg.partial_corr(data=df, x='X', y='Y',
    covar='Z')

# Print the partial correlation coefficient and p-value
print("Partial Correlation Coefficient:",
    partial_corr_result['r'].values[0])
print("p-value:", partial_corr_result['p-val'].values[0])
```

Multiple Correlation

Suppose we have an outcome variable y and a set of predictors $x_1, \ldots, x_k$. The maximum possible correlation between y and a linear combination of the predictors $c_1x_1 + \ldots + c_kx_k$ is given by the correlation between y and the regression function $\beta_1x_1 + \ldots + \beta_kx_k$ and is called the multiple correlation between y and $\{x_1, \ldots, x_k\}$. It is estimated by the Pearson correlation between y and $b_1x_1 + \ldots + b_kx_k$, where $b_1, \ldots, b_k$ are the least-squares estimates of $\beta_1, \ldots, \beta_k$. The multiple correlation can also be shown to be equivalent to $\sqrt{\frac{\text{Reg SS}}{\text{Total SS}}} = \sqrt{R^2}$. So multiple correlation is defined as

$$R = \sqrt{R^2}$$

where R^2 is the coefficient of determination for the number of independent variables in the model.

- R ranges from 0 to 1. A higher R value suggests a stronger linear relationship between the variables and $R = 0$ indicates no linear relationship.

Python Code: The Multiple Correlation

```
import numpy as np

# Sample data: three variables X1, X2, X3
X1 = [1, 2, 3, 4, 5]
X2 = [2, 3, 5, 4, 6]
X3 = [5, 4, 3, 2, 1]

# Combine variables into a 2D array
data = np.array([X1, X2, X3])

# Calculate the correlation matrix
correlation_matrix = np.corrcoef(data)

# Extract the multiple correlation coefficient (square root of
    determinant of the correlation matrix)
multiple_correlation = np.sqrt(np.linalg.det(correlation_matrix))

print("Multiple correlation coefficient:", multiple_correlation)
```

Chapter Exercises

1. Consider the following dataset representing the scores of two students in two exams:

Student	Quiz 1	Quiz 2
A	80	75
B	90	85
C	70	65
D	85	80

Calculate the Spearman's rank correlation coefficient between Quiz 1 and Quiz 2 scores.

2. Suppose you have collected data on the height and weight of ten individuals:

Height (cm)	Weight (kg)
160	55
165	60
170	65
175	70
180	75
185	80
190	85
195	90
200	95
205	100

Compute the Pearson correlation coefficient between height and weight and interpret your result.

3. Given the following contingency table representing the relationship between two categorical variables:

	Variable 1	Total
Variable 2	A	B
C	20	30
D	10	40
Total	30	70

Calculate the phi correlation coefficient and interpret your result.

4. Consider the following contingency table representing the relationship between two categorical variables:

	Myocardial Infraction (MI)	Total	
Smoking Status	*Yes*	*No*	Total
Yes	40	30	70
No	10	40	50
Total	50	70	120

Calculate Cramér's V for the given contingency table and interpret your result.

5. Two raters independently evaluate the performance of 10 students in a quiz competition. Each rater scores the students as either "Pass" or "Fail." The ratings are compared to assess the agreement between the raters using Cohen's Kappa statistic.

Student	Rater 1	Rater 2
1	Pass	Pass
2	Pass	Fail
3	Fail	Fail
4	Pass	Pass
5	Pass	Pass
6	Fail	Pass
7	Fail	Fail
8	Pass	Pass
9	Pass	Pass
10	Fail	Fail

(i) Convert the ratings into binary labels. For example, you might assign 1 for "Pass" and 0 for "Fail."

(ii) Calculate the observed agreement (P_o) between Rater 1 and Rater 2's ratings.

(iii) Calculate the expected agreement (P_e) by chance.

(iv) Use the formula for Cohen's Kappa (κ) to calculate the statistic.

(v) Interpret the value of Cohen's Kappa in terms of the level of agreement between the two raters.

Python Code: The Kappa Statistic

```
from sklearn.metrics import cohen_kappa_score

# Data representing the ratings by two raters
rater1 = ["Pass", "Pass", "Fail", "Pass", "Pass", "Fail",
    "Fail", "Pass", "Pass", "Fail"]
rater2 = ["Pass", "Fail", "Fail", "Pass", "Pass", "Pass",
    "Fail", "Pass", "Pass", "Fail"]

# Define the possible categories
categories = ["Pass", "Fail"]

# Calculate Cohen's Kappa
kappa = cohen_kappa_score(rater1, rater2, labels=categories)

print("Cohen's Kappa:", kappa)
```

Chapter 2: Regression Analysis: Simple Linear Regression

3 Chapter 2: Regression Analysis: Simple Linear Regression

Problem & Motivation

Research Problem 1

Obstetricians sometimes order tests to measure estriol levels from 24-hour urine specimens taken from pregnant women who are near term because level of estriol has been found to be related to infant birthweight. The test can provide indirect evidence of an abnormally small fetus. Greene and Touchstone conducted a study to relate birthweight and estriol level in pregnant women[a]. The Sample data are presented in following Table 4. They want to find any relationship between the estriol level and birthright How can this relationship be quantified? What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

[a] Greene, J., & Touchstone, J. (1963). Urinary tract estriol: An index of placental function. American Journal of Obstetrics and Gynecology, 85(1), 1-9.

Table 4: Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Motivation for Regression Analysis

- The drawback of correlation analysis is that it only measures the strength and direction of the linear relationship between two variables. It does not provide information about causality or the nature of the relationship beyond linearity. Additionally, correlation coefficients can be affected by outliers and may not capture complex relationships that exist between variables.
- The motivation for regression analysis stems from the need to understand and model the relationship between variables more comprehensively. Regression analysis allows us to not only measure the strength and direction of the relationship but also to make predictions and infer causality, provided certain assumptions are met. By fitting a regression model, we can examine how changes in one variable are associated with changes in another variable while controlling for potential confounding factors. This enables deeper insights into the underlying mechanisms driving the relationship between variables.

Functional Relation

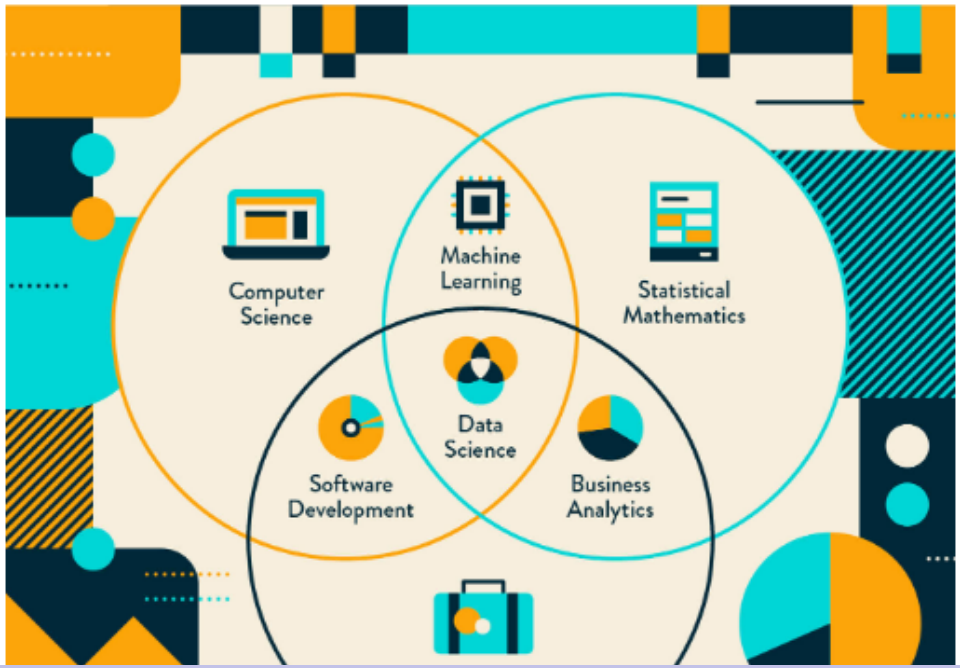
- Suppose f is a known function.

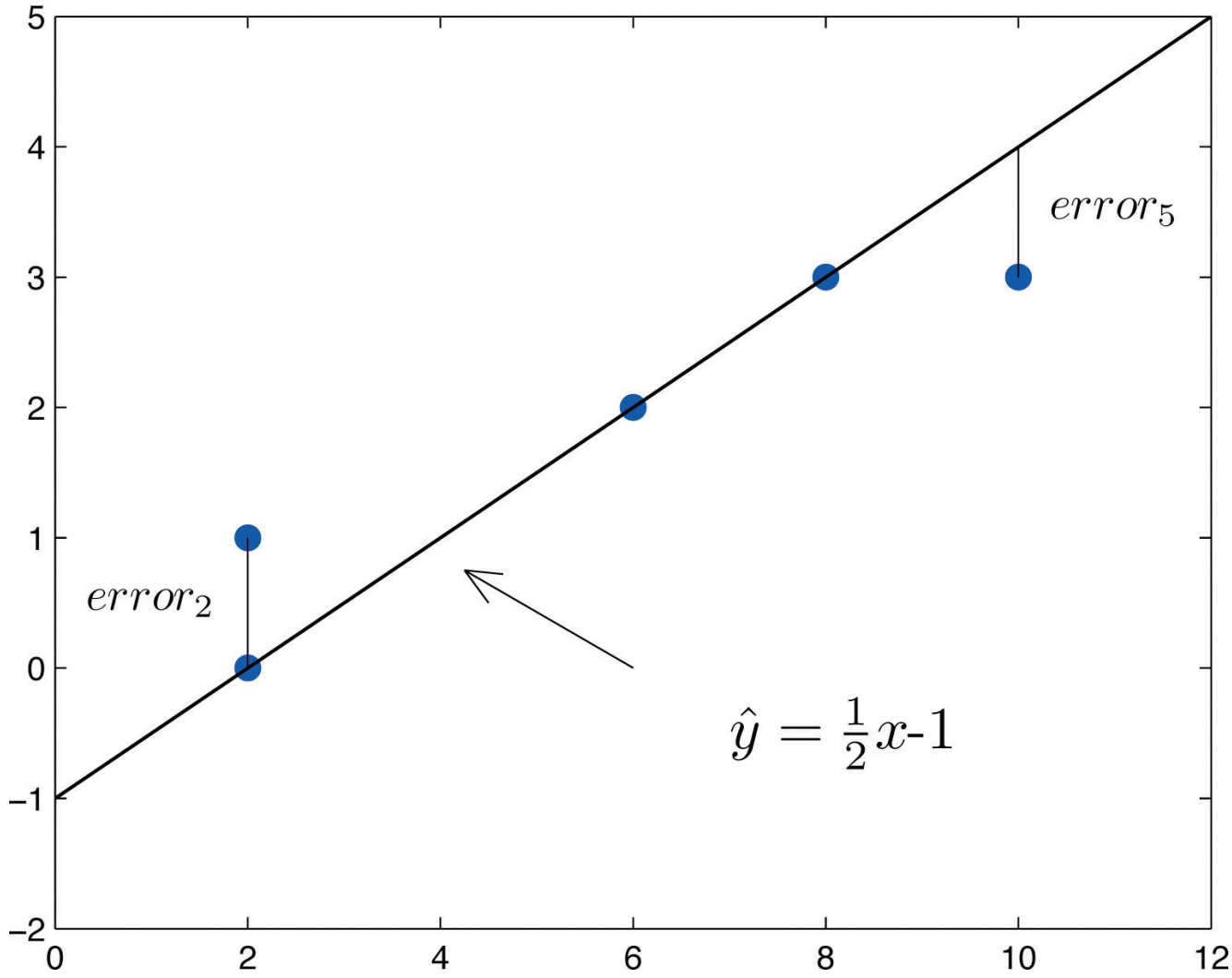
$$Y = g(X)$$

  Whenever X is known, Y is completely known.

- Examples:

(1) $y = 2x$ (2) $y = \frac{1}{2}x - 1$ (3) $y = 5 + \frac{1}{2}x^2$




5
4
3
2
1
0
−1
−2
0
2
4
6
8
10
12
$error_2$
$error_5$
$\hat{y} = \frac{1}{2}x\text{-}1$

Regression Analysis

- correlation analysis does not tell us why and how behind the relationship but it just says the relationship exists
- regression is to build a function of independent variables (also known as predictors) to predict a dependent variable (also called response) for example,

$$y = g(x_1, x_2, \ldots, x_p) + e$$

where y is a dependent variable, $x_1, x_2, \ldots, x_p$ independent variables, e is an error term.

- for example, banks assess the risk of home-loan applicants based on their
  - age
  - income
  - expenses
  - occupation
  - number of dependents,
  - total credit, etc.

- regression analysis: to find out the relationship and measure the dependency between a response variable Y and a set of covariates $\mathbf{X}$
- standard regression methods have focused on the estimation of conditional mean function $\mathbb{E}(Y|\mathbf{X}) = g(\mathbf{X})$ of a conditional response Y given a set of p covariates $\mathbf{X} = (X_1, ..., X_p)$

Historical Origin of the Term Regression

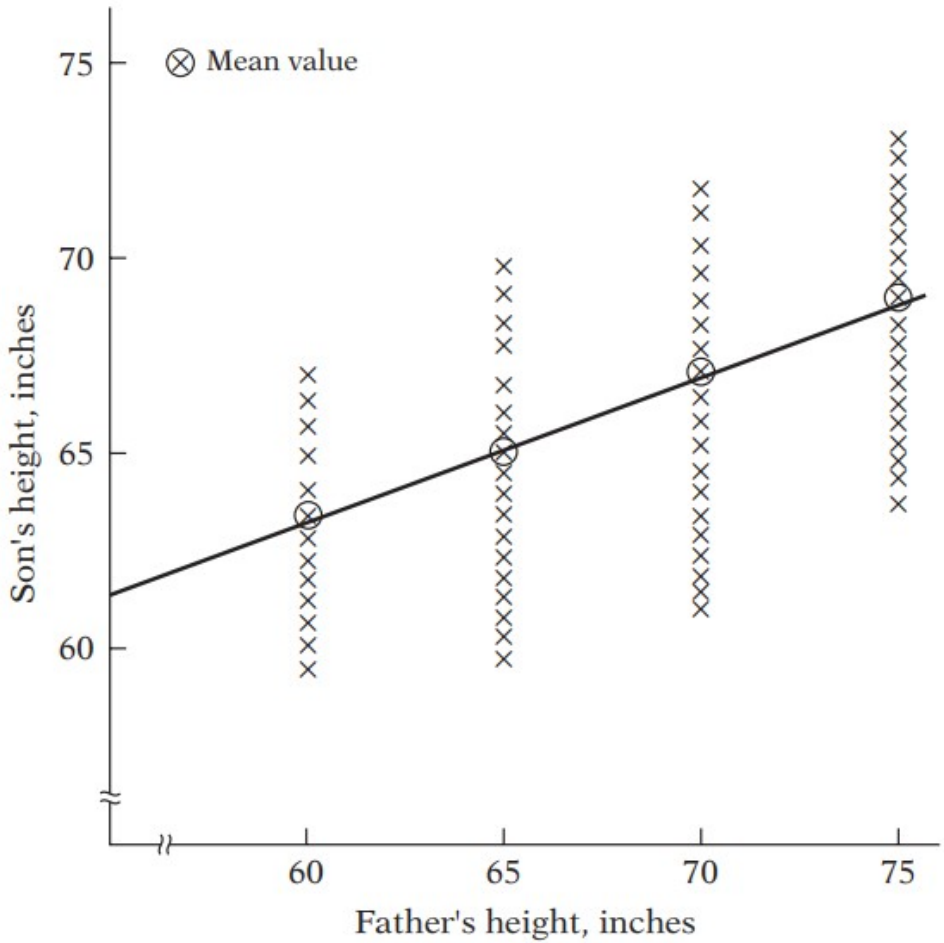


Figure 2: Hypothetical distribution of sons' heights corresponding to given heights of fathers

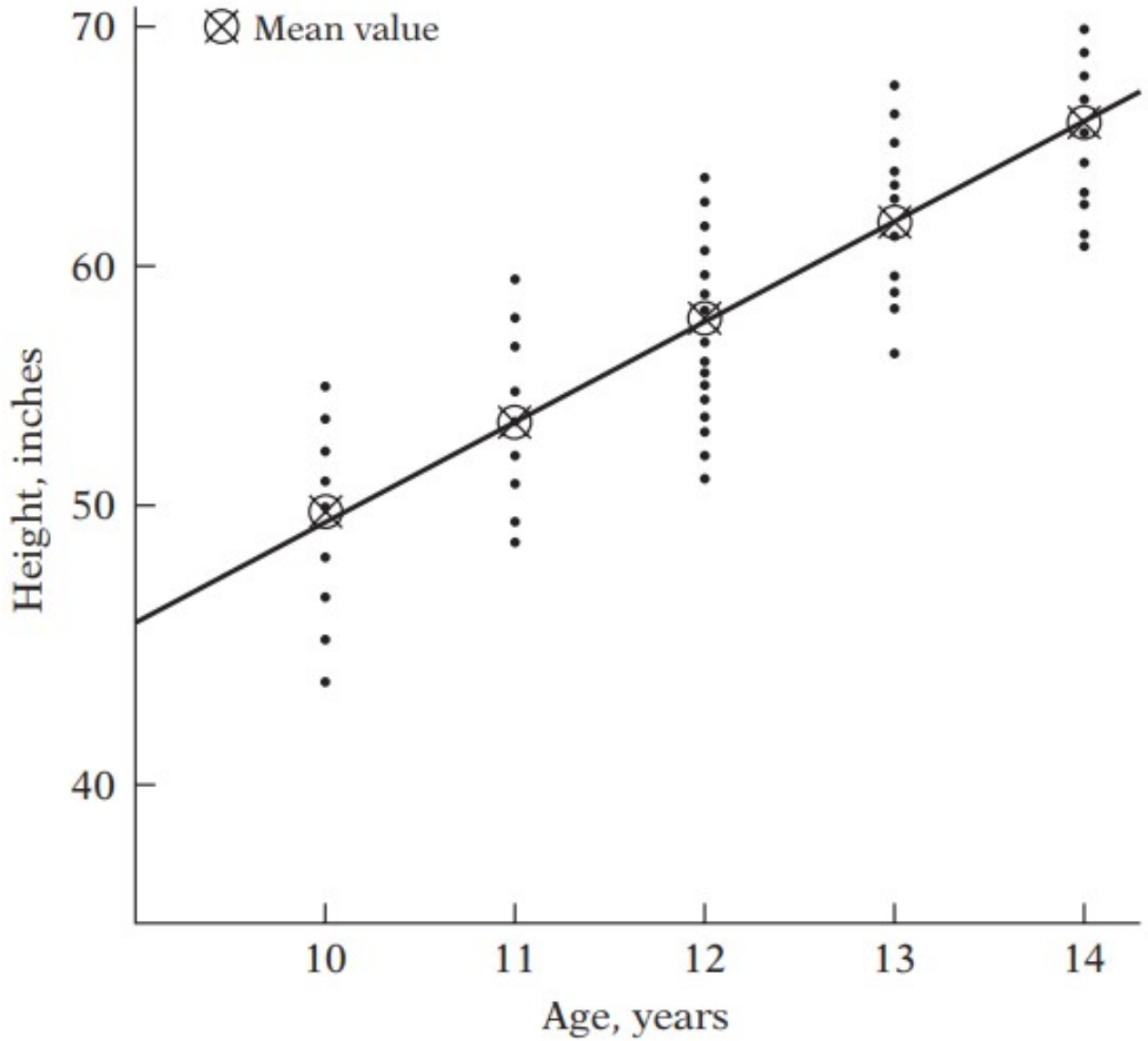


Figure 3: Hypothetical distribution of heights corresponding to given age

Dependent variable	**Explanatory variable**
⇕	⇕
Explained variable	Independent variable
⇕	⇕
Predictand	Predictor
⇕	⇕
Regressand	**Regressor**
⇕	⇕
Response	Stimulus
⇕	⇕
Endogenous	Exogenous
⇕	⇕
Outcome	Covariate
⇕	⇕
Controlled variable	Control variable

Main Use of Regression Analysis

The main use of regression analysis is to:

- Understand and quantify relationships between variables.
- Identify significant factors influencing a dependent variable.
- Assess the strength and direction of associations between variables.
- Examine how changes in one variable are associated with changes in another variable while controlling for potential confounding factors.
- Predict future outcomes or values based on historical data.
- Make informed decisions and recommendations based on statistical evidence.

Comparison of correlation analysis and regression analysis

Aspect	**Correlation Analysis**	**Regression Analysis**
Purpose	Measure the strength and direction of association between two (continuous) variables.	Model and quantify the relationship between a dependent variable and one or more independent variables.
Direction of Analysis	Examines relationship between variables without assuming causation.	Explores relationship between variables with assumption of causation.
Output	Produces correlation coefficient (e.g., Pearson's r, Spearman's ρ).	Provides regression coefficients (slope and intercept), along with measures of model fit (e.g., R-squared).

Table 5: Comparison between Correlation Analysis and Regression Analysis

Aspect	**Correlation Analysis**	**Regression Analysis**
Model Complexity	Relatively simple analysis; provides single summary statistic.	Can range from simple linear regression to complex models with multiple predictors.
Inference	Descriptive analysis; does not imply causation.	Allows for inference and hypothesis testing; enables predictions and causal conclusions.

Types of parametric regression analysis

- for continuous response variable
  - Simple Linear Regression
  - Multiple Regression
  - Polynomial Regression
  - Multivariate Regression
- for categorical response variable
  - Logistic Regression
    - binomial (also called ordinary) logistic regression
    - multinomial logistic regression
    - ordinal logistic regression
    - alternating logistic regressions
- for discrete response variable
  - Poisson Regression
- for survival-time (time-to-event) outcomes
  - Cox Regression (or proportional hazards regression)

Regression Models

another types of regression analysis:

- Ridge Regression
- Lasso Regression
- Bayesian regression
- Nonparametric regression

types of regression model in machine learning

- Generalized linear model (GLM)
- Additive model
- Generalized additive model (GAM)
- Random effect model
- Linear mixed model (LMM)
- Generalized linear mixed model (GLMM)
- Generalized estimating equations (GEE)

Simple Regression Model

To explore the relationship between independent variable x and dependent variable y we consider the following population regression equation:

$$E(y|x) = \beta_0 + \beta_1 x$$

where $E(y|x)$ = expected or average value of y for the give value of x.

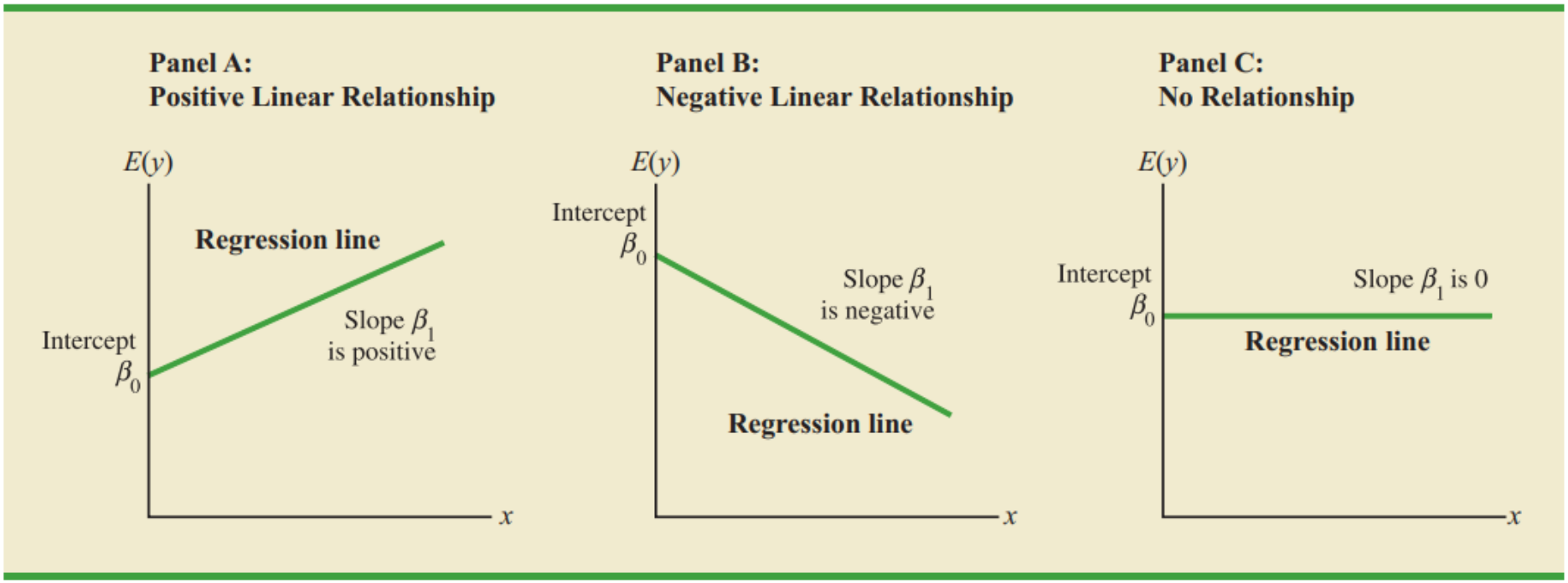
- Simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, n \tag{1}$$

where we assume, e_i is normally distributed with mean 0 and constant variance. That ie, $\mathbb{E}(e_i) = 0$ and $\text{Var}(e_i) = \sigma^2$ (say).

Remarks: Normality of the errors (or residuals) is not strictly required. However, the normality assumption in Equation (1) is necessary to perform hypothesis tests concerning regression parameters, as discussed in the next.

Graphical Presentation



- For $x = x_i$, we have one (sample) observation, $y = y_i$, the population regression function (PRF), it can be expressed as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i \\ &= \mathbb{E}(y|x_i) + e_i \end{aligned}$$

- In terms of the sample regression function (SRF), the observed y_i can be expressed as

$$y_i = \widehat{y}_i + \widehat{e}_i$$

- The estimated model (or estimated line) of (1) can be written as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where

  - $\widehat{y}_i$ = estimator of $\mathbb{E}(y|x_i)$ under the assumption $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - $\widehat{\beta}_0$ = estimator of β_0
  - $\widehat{\beta}_1$ = estimator of β_1

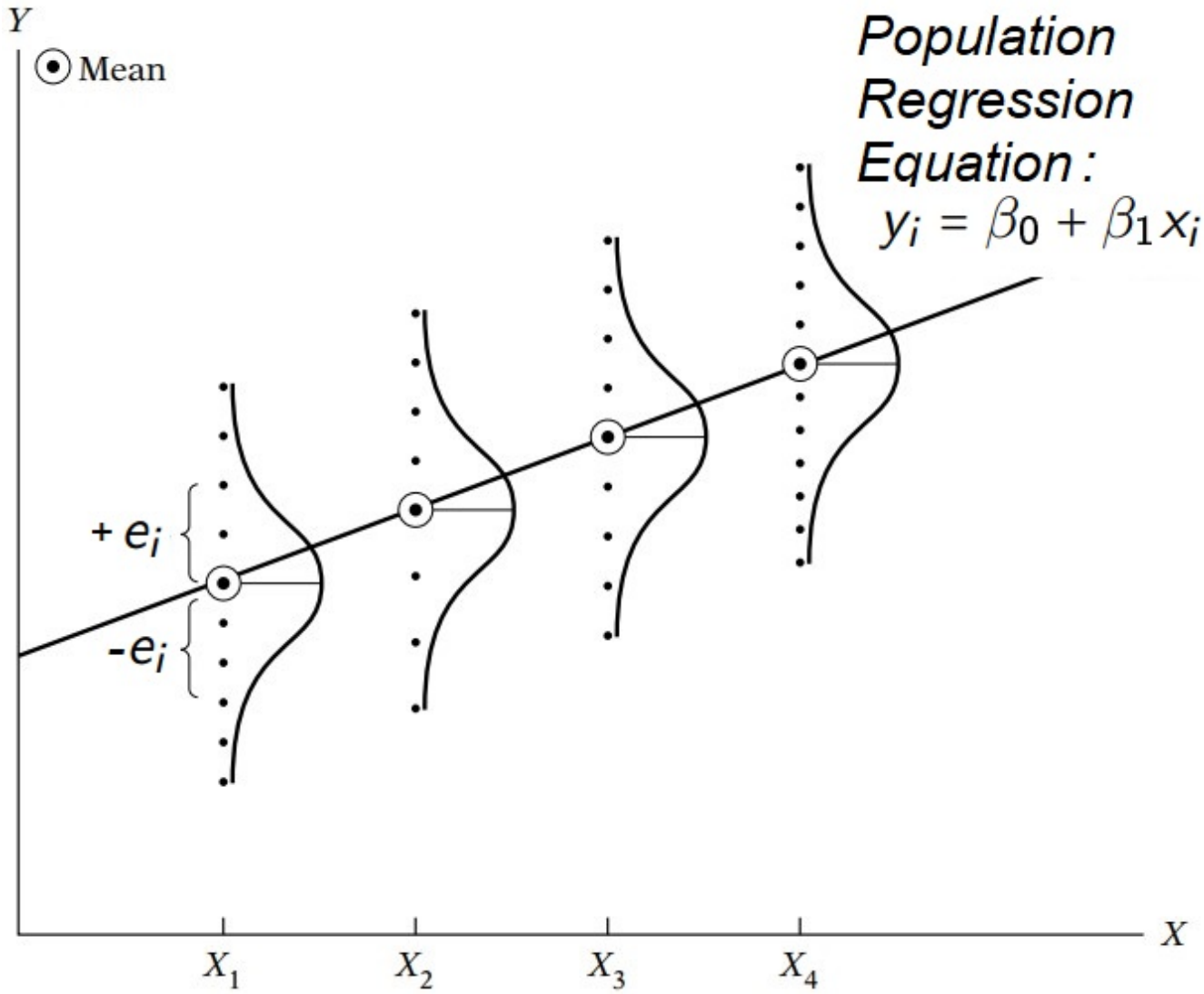
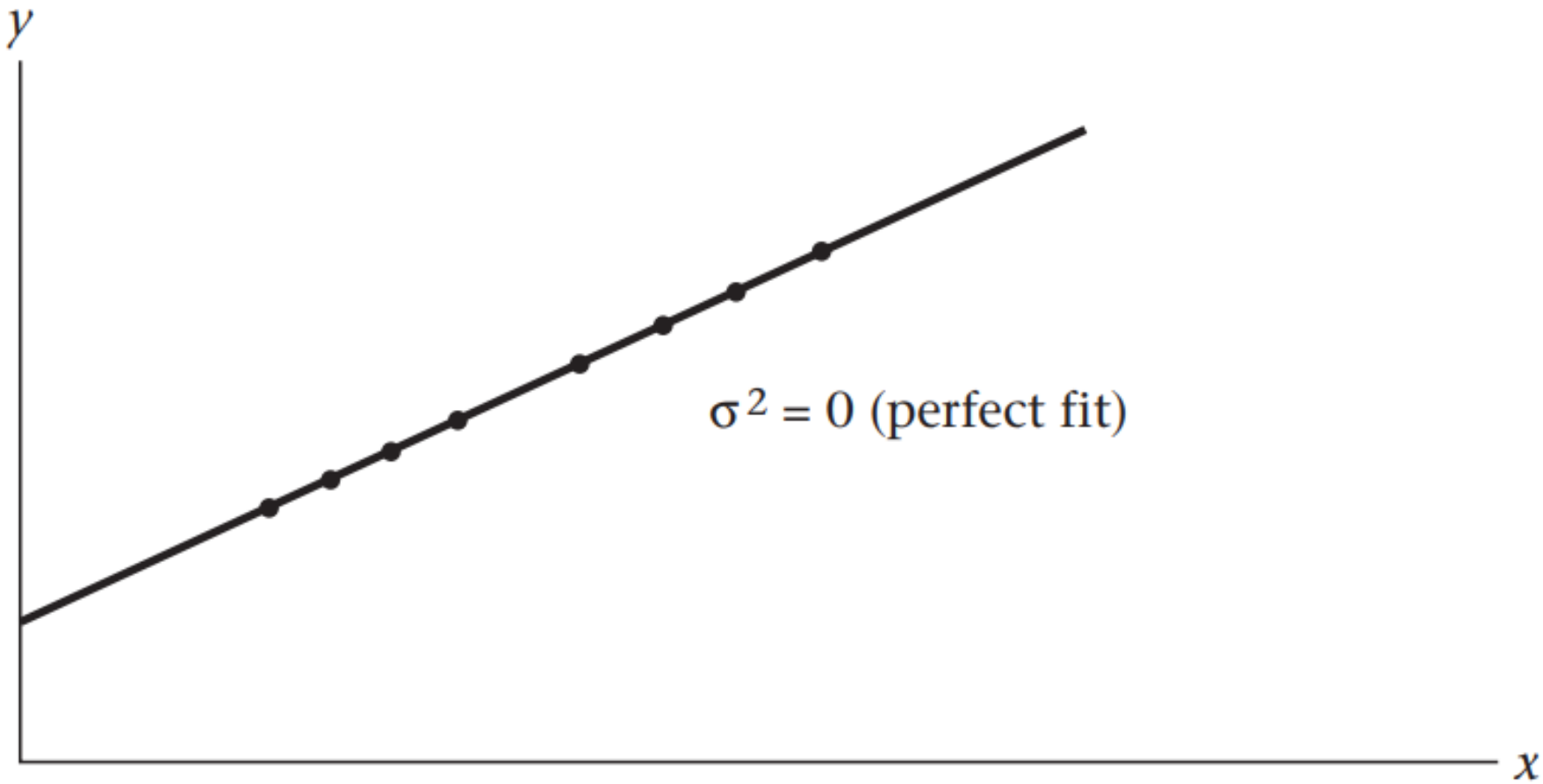
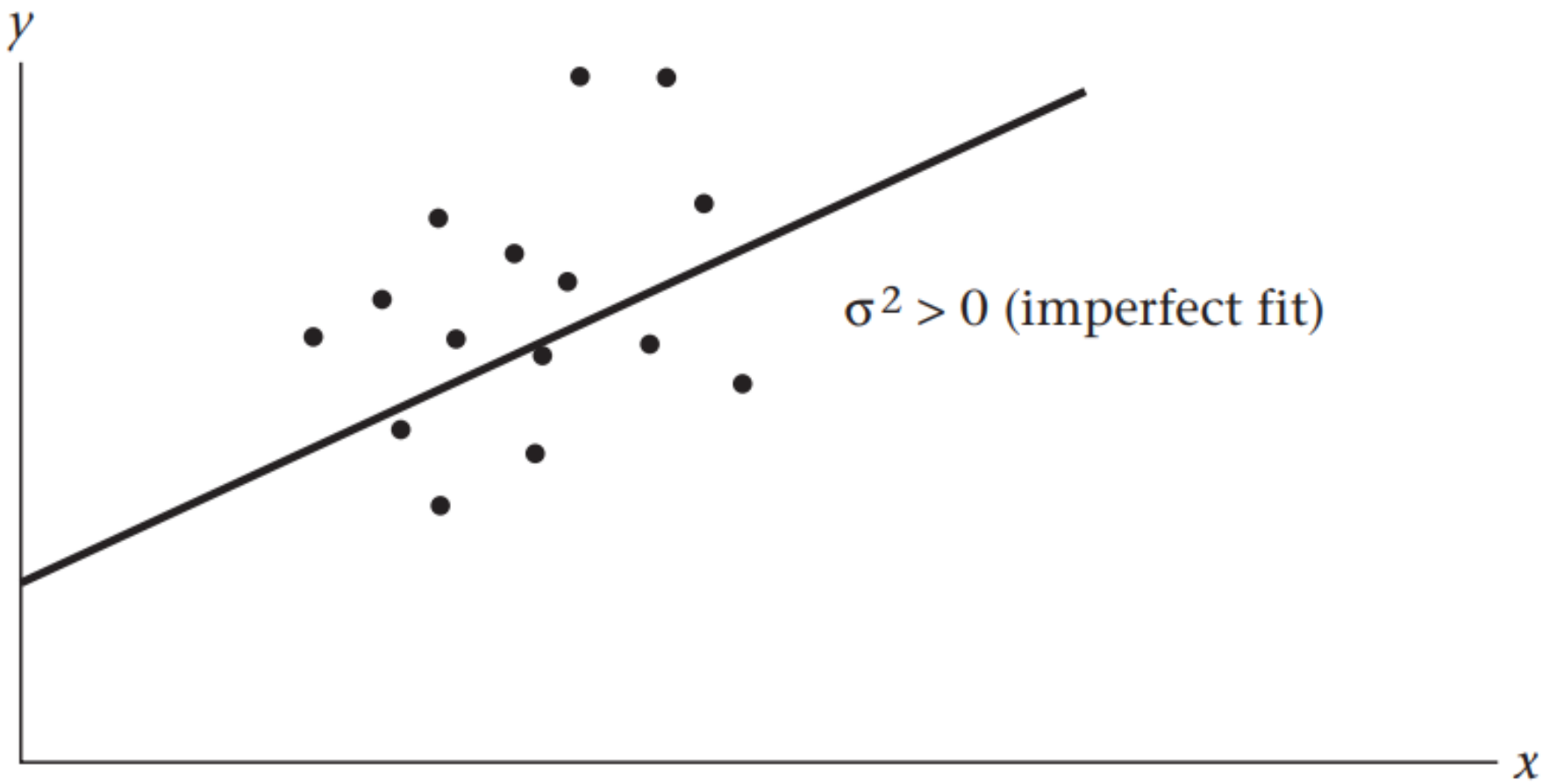
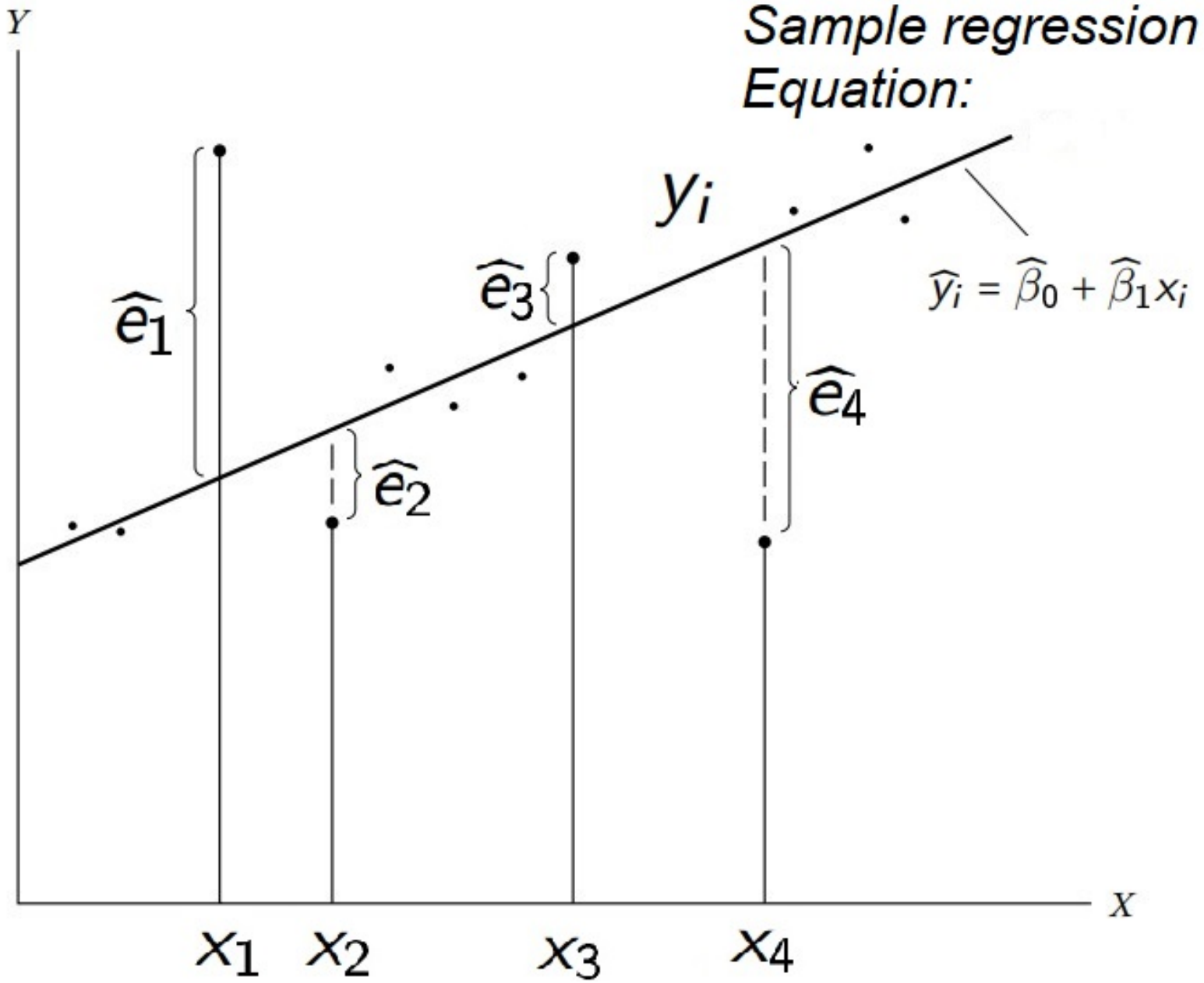


Figure 4: Conditional distribution of the disturbances e_i

The effect of σ^2 on the goodness of fit of a regression line




y
x
$\sigma^2 > 0$ (imperfect fit)


Y
Sample regression Equation:
y_i
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
$\hat{e}_1$
$\hat{e}_2$
$\hat{e}_3$
$\hat{e}_4$
X
x_1
x_2
x_3
x_4

A Probabilistic View of Regression Analysis

- regression analysis: to find out the relationship between a response variable Y and a set of covariates $\mathbf{X}$.
- response variable: Y and a set of covariates: $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$
- the conditional density:

$$\underbrace{f_{Y|\mathbf{X}}(y|\mathbf{x})}_{\text{parametric}} \leftarrow \underbrace{\begin{cases} \mu(\mathbf{x}) \\ \sigma^2(\mathbf{x}) \end{cases}}_{\text{parametric/nonparametric}}$$

- conditional Mean of Y given $\mathbf{X} = \mathbf{x}$:

$$\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \mu(\mathbf{X})$$

aims:

- estimation of unknown function $\mathbb{E}(Y|\mathbf{X}) = \mu(\mathbf{X})$

Distributional Assumptions Underlying Classical Regression

- if
  - $f_{y|\mathbf{x}}(y|\mathbf{x})$ is a normal density (Remarks: y must be a normal random variate)
  - $\mu(\mathbf{x})$ is a linear function (i.e., $\mu(x) = \beta_0 + \beta_1 x$) and
  - $\sigma^2(\mathbf{x})$ is a constant i.e., $\sigma^2(\mathbf{x}) = \sigma^2$ (say)

  then we use a *classical linear regression model* to estimate $\mathbb{E}(y|\mathbf{x})$
- in this case, we assume

$$\mathbb{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x + \mathbb{E}(e|\mathbf{x})$$

  where β_1 measures the marginal change in the mean of y due to a marginal change in $\mathbf{x}$.

We can now reformulate the assumptions of the classical regression model in the next step

Assumption of Classical Linear Model

The *classical (simple) linear regression model* is

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, n \tag{2}$$

Assumptions

1. **Linearity:** There exists a linear relationship between the independent variable X and the dependent variable Y.
2. **Independence:** The observations are independent of each other.
3. **Homoscedasticity:** The variance of the errors (residuals) is constant across all levels of the independent variable X. That is , $\text{Var}(e_i) = \sigma^2$.
4. **Normality:** The errors (residuals) follow a normal distribution with mean 0. That is , $\mathbb{E}(e_i) = 0$.
5. **No perfect multicollinearity:** There is no perfect linear relationship between the independent variable X and any other variable.

That is, $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, that is $y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Therefore, it is also called normal regression

How do you estimate parameters of the simple linear regression model?

- Suppose our simple regression model is

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, n$$

  where we assume, e_i is normally distributed with mean 0 and variance σ^2.
- usually, β_0, β_1 and σ^2 are unknown
- we can estimate β_0, β_1 and σ^2 by following methods
  - ordinary least square (OLS) method
  - maximum likelihood estimation
  - method-of-moments (or GMM)
  - ...

Interpretation of Slope Coefficient

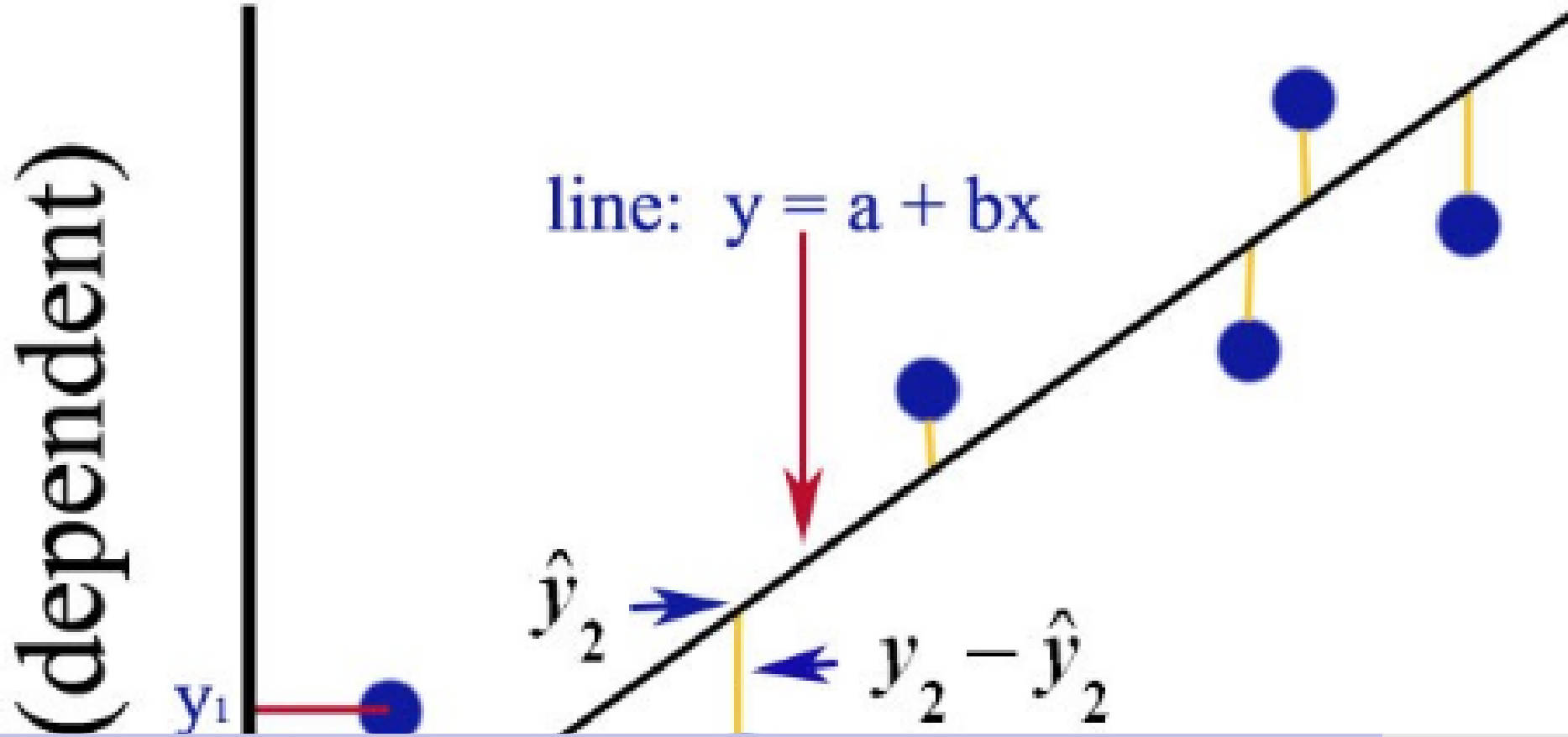
- the estimated regression model (or equation) is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

  where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are respectively, the estimator of β_0 and β_1
- $\widehat{\beta}_1$ is the slope of the fitted (estimated) line
  - it shows the amount of change in $\widehat{y}$ for a change of one unit in x
  - a positive value for $\widehat{\beta}_1$ indicates a direct relationship between the two variables and a negative value indicates an inverse relationship
  - the sign of $\widehat{\beta}_1$ and the sign of r, the correlation coefficient, are always the same

Least Square Estimation

A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual y values and the estimated (or predicted) values of y



Ordinary Least Square (OLS) Estimation

1. minimizing function:

$$Q = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$
$$= \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

2. the estimators for β_0 and β_1 can be found by solving the following normalized equation

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{3}$$

and

$$\frac{\partial Q}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{4}$$

OLS estimatior

from (3), we have

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

or

$$\beta_0 = \bar{y} - \beta_1 \bar{X}$$

from (4), we have

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad (5)$$

putting the value of β_0 in Equation (5), we have

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} - \beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = 0$$

$$\Rightarrow \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

- hence, the ordinary least square (OLS) estimators of β_0 and β_1 are

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

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

- the expression of $\widehat{\beta}_1$ can also be written as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r\left(\frac{s_y}{s_x}\right)$$

because

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

Remarks: Note that the method of least squares is appropriate whenever the average residual for each given value of x is 0— that is, when $\mathbb{E}(e|X = x) = 0$ in Equation (2). Normality of the errors (or residuals) is not strictly required. However, the normality assumption in Equation (2) is necessary to perform hypothesis tests concerning regression parameters, as discussed in the next.

The Estimated Error Variance or Standard Error

The error variance or standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression. The formulas that are used to compute the error variance or the standard error:

$$\widehat{\sigma}^2 = \frac{\sum_i^n (y_i - \widehat{y}_i)^2}{n-2} = \frac{\sum_i^n y_i^2 - \widehat{\beta}_0 \sum_i^n y_i - \widehat{\beta}_1 \sum_i^n x_i y_i}{n-2}$$

where

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Hence, the standard error of estimate is

$$\widehat{\sigma} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i^n y_i^2 - \widehat{\beta}_0 \sum_i^n y_i - \widehat{\beta}_1 \sum_i^n x_i y_i}{n-2}}$$

Example 3.1 (Sultan's Dine restaurants Sales Dataset)

To illustrate the least squares method, suppose data were collected from a sample of 10 Sultan's Dine restaurants located near to the university campuses. For the ith observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars).

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

(i). Show the relationship between the size of student population and the quarterly sales. Make a comment on the diagram.

(ii). Write down the regression model for this example, and mention the assumptions of the model.

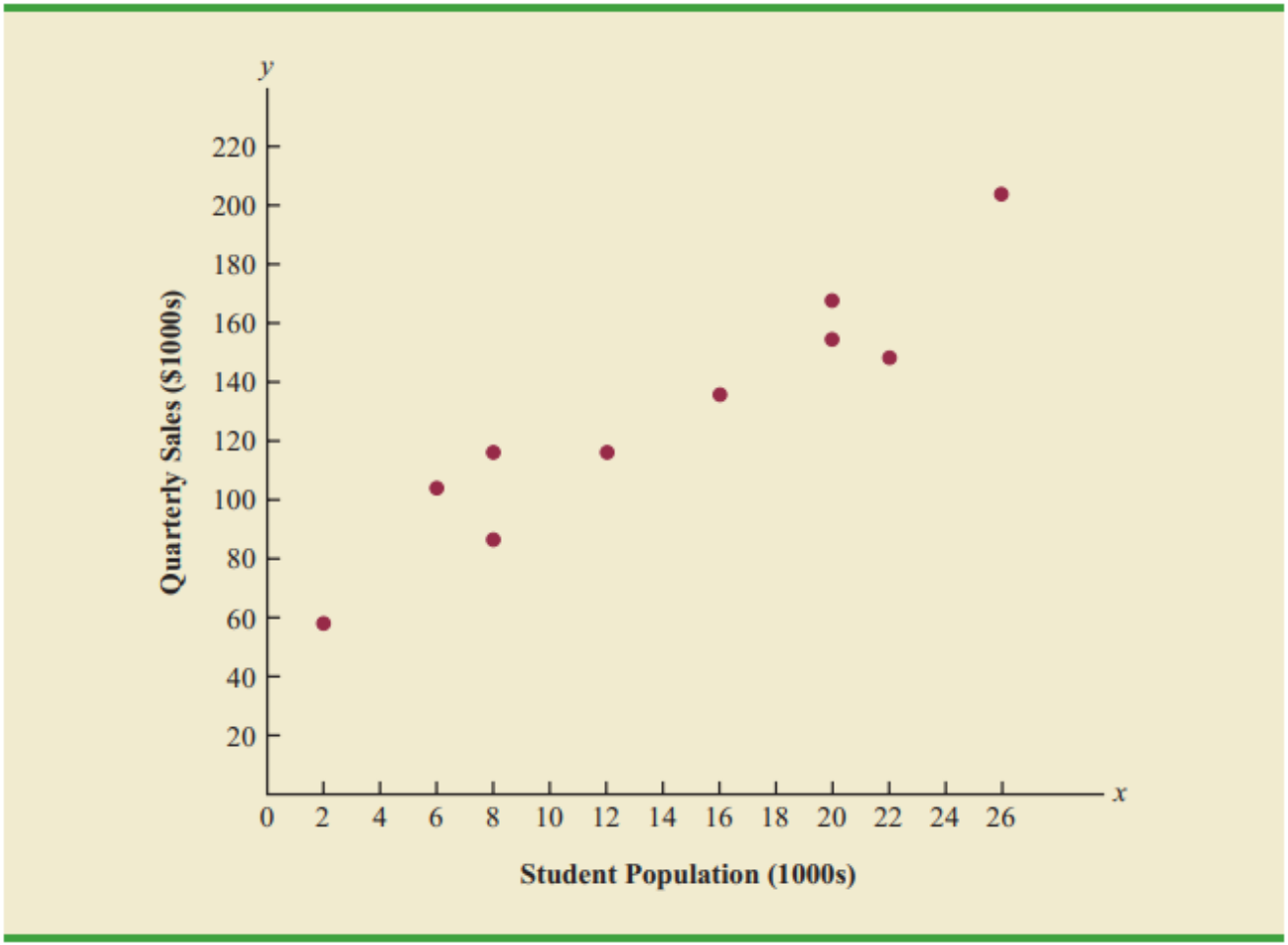
(iii). Write the estimated regression equation and find the least square estimates. Interpret the results.

(iv). Draw the regression line.

(v). Predict quarterly sales for a restaurant to be located near a campus with 16,000 students.

(vi). Find the value of the standard error of the estimates.

Student Population And Quarterly Sales Data: for 10 pair observations



- Simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, 10 \tag{6}$$

where

  - y_i = Quarterly Sales ($1000s)
  - x_i = bStudent Population (1000s)
  - β_0 is intercept
  - β_1 is slope coefficient
  - e_i error term and we assume, e_i is normally distributed with mean 0 and variance σ^2.

- The estimated regression model (or equation) of (6) is

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

where $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are respectively, the estimator of β_0 and β_1.

- The ordinary least square (OLS) estimators of β_0 and β_1 are

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \quad ; \quad \widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

The Least Squares Estimation

x_i	y_i	x_i^2	$x_i y_i$
2	58	4	116
6	105	36	630
8	88	64	704
8	118	64	944
12	117	144	1404
16	137	256	2192
20	157	400	3140
20	169	400	3380
22	149	484	3278
26	202	676	5252
Total= 140	1300	2528	21040

The ordinary least square (OLS) estimators of β_0 and β_1 are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{2840}{568} = 5$$

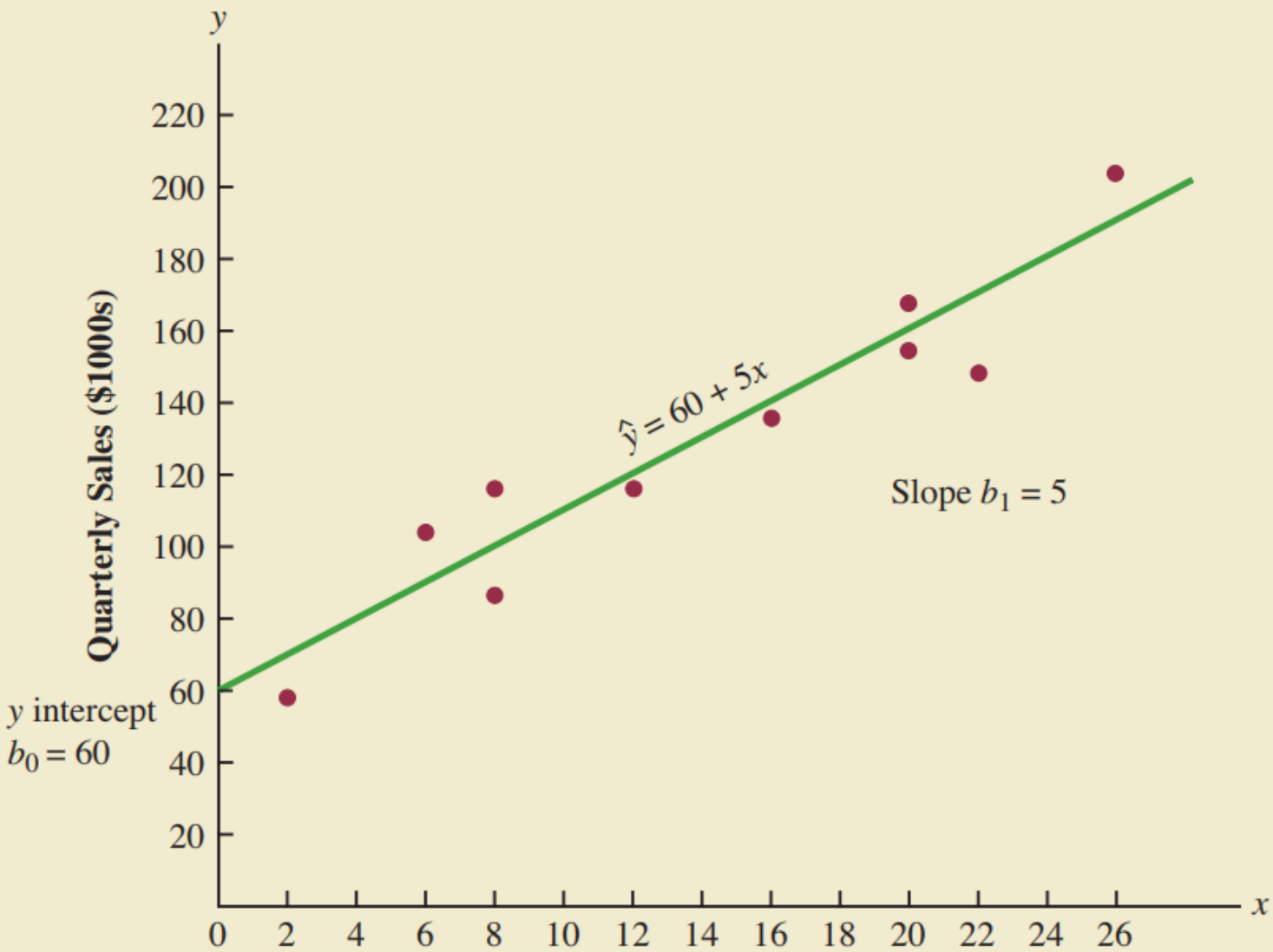
$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 130 - 5(14) = 60$$

Thus, the estimated regression equation is

$$\widehat{y}_i = 60 + 5x_i$$

The slope of the estimated regression equation ($\widehat{\beta}_1 = 5$) is positive, implying that as student population increases, sales increase. In fact, we can conclude that an increase in the student population of 1000 is associated with an increase of \$5000 in expected sales; that is, quarterly sales are expected to increase by \$5 per student.

FIGURE 14.4 GRAPH OF THE ESTIMATED REGRESSION EQUATION FOR ARMAND'S PIZZA PARLORS: $\hat{y} = 60 + 5x$



To predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\widehat{y} = 60 + 5 \times 16 = 140$$

Hence, we would predict quarterly sales of \$140,000 for this restaurant.

The Standard Error of the Estimates

x_i	y_i	$\widehat{y}_i = 60 + 5x_i$	$(y_i - \widehat{y}_i)^2$
2	58	70	144
6	105	90	225
8	88	100	144
8	118	100	324
12	117	120	9
16	137	140	9
20	157	160	9
20	169	160	81
22	149	170	441
26	202	190	144
140	**1300**	**760**	**1530**

$$\widehat{\sigma}^2 = \frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2} = \frac{1530}{10-2} = 765 \text{ and } \widehat{\sigma} = \sqrt{765} = 27.66$$

Hence the standard error of the estimates is $\widehat{\sigma} = 27.66$.

Coefficient of Determination

- the coefficient of determination is the proportion of the total variation in the dependent variable Y that is explained by the independent variable(s) in a regression model. It is denoted by R^2 and defined by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

  where
  - the total sum of squares (proportional to the variance of the data): $SST = \sum_i (y_i - \bar{y})^2$
  - The sum of squares of residuals, also called the Residual or Error Sum of Squares (SSE): $SSE = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i e_i^2$
- For example, suppose $R^2 = 0.9027$. This implies that 90.27% of the variability of the dependent variable is explained and the remaining 9.73% of the variability is still unexplained by the regression model.

For the Sultan's Dine restaurants Sales Dataset give in 3.1, we have

x_i	y_i	$(y_i - \bar{y})^2$	$\widehat{y_i} = 60 + 5x_i$	$(y_i - \widehat{y_i})^2$
2	58	5184	70	144
6	105	625	90	225
8	88	1764	100	144
8	118	144	100	324
12	117	169	120	9
16	137	49	140	9
20	157	729	160	9
20	169	1521	160	81
22	149	361	170	441
26	202	5184	190	144
140	**1300**	**15730**	**1300**	**1530**

$$\text{hence, } R^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{1530}{15730} = 0.9027$$

Classroom Practice

Example 3.2 (Classroom Practice)

The following sample of observations were randomly selected.

x	4	5	3	6	10
y	4	6	5	7	7

(a). Determine the regression equation.

(b). Write the estimated regression equation (or line).

(c). Determine the estimated value of y when x is 7.

(d). Find the value of the coefficient of determination and interpret your results.

Solution of the Exercises

Solution of the Exercise given in 3.2

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
4	4	16	16	16
5	6	25	36	30
3	5	9	25	15
6	7	36	49	42
10	7	100	49	70
$\sum x_i = 28$	$\sum y_i = 29$	$\sum x_i^2 = 186$	$\sum y_i^2 = 175$	$\sum x_i y_i = 173$
$\bar{x} = 5.6$	$\bar{y} = 5.8$			

(a). The least square estimator of β_0 and β_1 are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 0.3630 \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 3.7671$$

(b). The estimated equation (or line) is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = 3.7671 + 0.3630 x_i$$

(c). The estimated value of y is

$$\widehat{y}_i = 3.7671 + 0.3630 \times 7 = 6.3082$$

when x_i is 7.

Example: Coefficient of Determination

x_i	y_i	$(y_i - \bar{y})^2$	$\widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
4	4	3.24	5.2191	1.486205
5	6	0.04	5.5821	0.17464
3	5	0.64	4.8561	0.020707
6	7	1.44	5.9451	1.112814
10	7	1.44	7.3971	0.157688
$\sum_i x_i = \mathbf{28}$	$\sum_i y_i = \mathbf{29}$	SST=**6.8**		SSE=**2.9521**

where

$$\begin{aligned}\widehat{y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i \\ &= 3.7671 + 0.3630 x_i\end{aligned}$$

Hence

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{2.9521}{6.8} = 0.566$$

Relationship between R^2 and r_{xy}

- The coefficient of correlation measures the strength and direction of a linear relationship between two variables. The coefficient of correlation is denoted by r_{xy}.
- **Coefficient of Determination (R^2):**
  The relationship between R^2 and r is that the square of the coefficient of correlation (r_{xy}) is equal to the coefficient of determination (R^2) for the simple regression model. Mathematically,

$$R^2 = (r_{xy})^2$$

- Hence, for the previous example,

$$\begin{aligned} R^2 &= (r_{xy})^2 \\ &= (0.7522)^2 \\ &= 0.566 \end{aligned}$$

Remarks: Note that the relationship $R^2 = (r_{xy})^2$ only holds for the simple linear regression model.

- the estimated regression equation for simple linear regression model provides

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

- the sample correlation coefficient is

$$\begin{aligned} r_{xy} &= (\text{sign of } \widehat{\beta}_1)\sqrt{\text{coefficient of determination}} \\ &= (\text{sign of } \widehat{\beta}_1)\sqrt{R^2} \end{aligned} \tag{7}$$

where R^2 is the coefficient of determination for simple regression model $y_i = \beta_0 + \beta_1 x_i + e_i; \quad i = 1, 2, \ldots, n$

Remarks: Note that the equation (7) is true only for simple regression model.

Advantages and Disadvantages of R^2

- R^2 is a statistic that will give some information about the goodness-of-fit of a model
- in regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points
- an R^2 of 1 indicates that the regression predictions perfectly fit the data
- R^2 increases as we increase the number of variables in the model (R^2 is monotone increasing with the number of variables included i.e., it will never decrease)
- an adjusted R^2 is a modification of R^2 that adjusts for the number of explanatory terms in a model (p) relative to the number of data points (n)

Adjusted R^2

- The adjusted R^2 (denoted by $\bar{R}^2$) is defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the total number of explanatory variables in the model (not including the constant term), and n is the sample size

- it can also be written as:

$$\bar{R}^2 = 1 - \frac{SSE/\text{df}_e}{SST/\text{df}_t}$$

where df_t is the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance

Adjusted R^2

- the explanation of this statistic is almost the same as R^2 but it penalizes the statistic as extra variables are included in the model
- the term $\frac{n-1}{n-p-1}$ is called the penalty of using the more regressors in a model
- when the number of regressors p, increases, $(1 - R^2)$ will decrease, but $\frac{n-1}{n-p-1}$ will increase
- whether more regressors improve the explanatory power of a model depends on the trade-of between R^2 and the penalty
- For the previous example, $p = 1$ and $R^2 = 0.566$ and hence,

$$\bar{R}^2 = 1 - (1 - 0.566)\left(\frac{5-1}{5-1-1}\right) = 0.4212$$

Figure 5: Excel output for the classroom practice example.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.752246166
R Square	**0.565874295**
Adjusted R Square	**0.421165727**
Standard Error	0.991976948
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.847945205	3.847945	3.910441	0.142405612
Residual	3	2.952054795	0.984018		
Total	4	6.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.767123288	1.119648297	3.364559	0.043586	0.2039027	7.33034388
x	0.363013699	0.18357357	1.977483	0.142406	-0.221199331	0.94722673

Confidence Interval for β_0

The confidence interval for β_0 can be computed using the standard formula for linear regression parameter estimates. The formula for the confidence interval for β_0 is:

$$\widehat{\beta}_0 \pm t_{\frac{\alpha}{2},(n-2)} \cdot se(\widehat{\beta}_0)$$

Where:

- $\widehat{\beta}_0$ is the estimated coefficient for x,
- $t_{\alpha/2}$ is the critical value of the t-distribution with $n-2$ degrees of freedom at a significance level of $\alpha/2$ (where α is typically 0.05 for a 95% confidence interval),
- the standard error of the estimator $\widehat{\beta}_0$ is

$$se(\widehat{\beta}_0) = \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)}.$$

Confidence Interval for β_1

The confidence interval for β_1 can be computed using the standard formula for linear regression parameter estimates. The formula for the confidence interval for β_1 is:

$$\widehat{\beta}_1 \pm t_{\frac{\alpha}{2},(n-2)} \cdot se(\widehat{\beta}_1)$$

Where:

- $\widehat{\beta}_1$ is the estimated coefficient for x,
- $t_{\alpha/2}$ is the critical value of the t-distribution with $n-2$ degrees of freedom at a significance level of $\alpha/2$ (where α is typically 0.05 for a 95% confidence interval),
- the standard error of the estimator $\widehat{\beta}_1$ is

$$se(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Where:

- $$\widehat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(y_i - \widehat{y}_i)^2}{n-2}}$$

  is the estimated standard error of the residuals (or the square root of the mean squared error, often obtained from the regression output),
- n is the number of observations,
- $\bar{x}$ is the mean of the independent variable x.

Once you have these values, you can compute the confidence interval.

Test overall significance of the regression model

In a simple linear regression model, the F-test is used to assess the overall significance of the regression model. The hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The null hypothesis H_0 implying that the independent variable(s) do not have any effect on the dependent variable. The alternative hypothesis H_1 indicating that the independent variable(s) do have a significant effect on the dependent variable. The formula for the F-statistic in a simple linear regression model is:

$$F = \frac{SSR/1}{SSE/(n-2)}$$

Where:

- $SSR = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$ is the sum of squared regression (explained),
- SSE is the sum of squared error (residual) terms,

- Under the null hypothesis, the F-statistic follows an F-distribution with p and $n-p-1$ degrees of freedom, where p is the number of regressors (excluding the intercept) and n is the number of observations.
- In a simple linear regression model, $p = 1$ because you only have one independent variable (excluding the intercept).
- So, the degrees of freedom for the F-distribution in a simple linear regression model are 1 and $n-2$.
- Once you compute the F-statistic, you can compare it to the critical value from the F-distribution at a chosen significance level (e.g., $\alpha = 0.05$) to determine whether to reject the null hypothesis.
- If the F-statistic is greater than the critical value, you reject the null hypothesis and conclude that the model is significant. Otherwise, you fail to reject the null hypothesis.

Decision Rule for ANOVA F-test

Classical Approach

- Set the significance level α.
- Calculate the critical value $F_{\text{critical}} = F_{\alpha}(p, n-p-1)$ from the F-distribution with appropriate degrees of freedom.
- Decision Rule:
  - If calculated $F > F_{\text{critical}}$, reject H_0 and conclude that the regression model is statistically significant.
  - If calculated $F \leq F_{\text{critical}}$, fail to reject H_0 and conclude that the regression model is not statistically significant.

Decision Rule for ANOVA F-test

p-value Approach

- Calculate the p-value associated with the calculated F-statistic.
- Decision Rule:
  - If p-value $< \alpha$, reject H_0 and conclude that the regression model is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the regression model is not statistically significant.

Hypothesis Test: $H_0 : \beta_1 = 0$

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad 1, 2, \ldots, n$$

We want to test the null hypothesis:

$$H_0 : \beta_1 = 0$$

Under H_0, the test statistic is given by:

$$t = \frac{\widehat{\beta}_1 - 0}{\text{se}(\widehat{\beta}_1)}$$

$$\text{where, } se(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ and } \widehat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(y_i - \widehat{y}_i)^2}{n-2}}$$

We reject H_0 if $|t|$ exceeds the critical value $t_{\text{critical}} = t_{\frac{\alpha}{2},(n-2)}$ from the t-distribution with $n-2$ degrees of freedom, where n is the sample size.

Decision Rule for t-test

Classical Approach

- Set the significance level α.
- Calculate the critical value t_{critical} from the t-distribution with appropriate degrees of freedom.
- Decision Rule for each $\widehat{\beta}_1$:
  - If $|t| > t_{\text{critical}} = t_{\frac{\alpha}{2},(n-2)}$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is statistically significant.
  - If $|t| \leq t_{\text{critical}}$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is not statistically significant.

p-value Approach

- Calculate the p-value associated with each calculated t-statistic.
- Decision Rule for each $\widehat{\beta}_1$:
  - If p-value $< \alpha$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is not statistically significant.

Note that the p-value can be obtained by using the formula:

$$p\text{-value} = 2 \times \min(P(T < -|t|), P(T > |t|))$$

Confidence Interval for $E(Y|X = x)$

- The Confidence Interval for $E(Y|X = x^*)$ provides a range of values where we expect the mean response y to lie for a given value of x^* with a certain level of confidence.
- It is computed as:

$$\widehat{y}^* \pm t_{\frac{\alpha}{2},(n-2)} \cdot se(\widehat{y}^*)$$

where $\widehat{y}^*$ is the predicted value of y for a given x^*, $t_{\alpha/2}$ is the critical value of the t-distribution, n is the number of observations, and

$$se(\widehat{y}^*) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

is the standard error of the predicted value.

- x^* is the specific value of the independent variable for which you're predicting the response,
- The confidence level is typically chosen to be 95% ($\alpha = 0.05$).

For Sultan's Dine restaurants Sales Dataset given in Example 3.1, we have $\widehat{\sigma} = 13.829$. With $x^* = 10, \bar{x} = 14$, and $\sum(x_i - \bar{x})^2 = 568$, we have

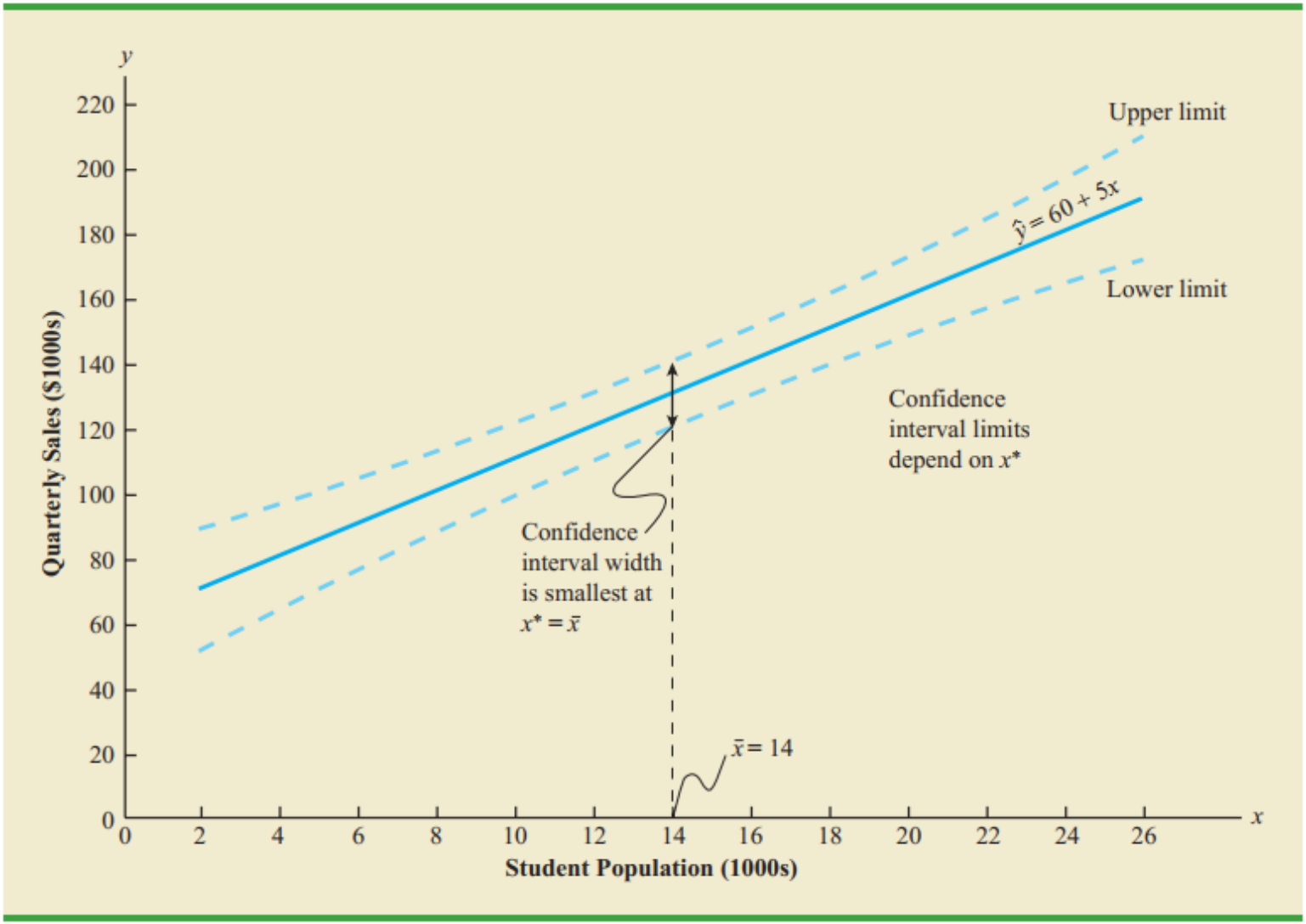
$$\begin{aligned} se(\widehat{y}^*) &= 13.829\sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} \\ &= 13.829\sqrt{0.1282} \\ &= 4.95 \end{aligned}$$

With $\widehat{y}^* = 110$ and a margin of error of $t_{0.025,8} \times se(\widehat{y}^*) = 2.306 \times 4.95 = 11.4147$, the 95% confidence interval for an average quarterly sales for the Sultan's Dine restaurants located near campus for fixed $x^* = 10$ is

$$110 \pm 11.4147$$

where $t_{\frac{\alpha}{2},(n-1)} = t_{0.025,8} = 2.306$.

Confidence and prediction intervals for sales Y at given values of student population X



Prediction Interval for an Individual Value of Y

- The Prediction Interval for an Individual Value of Y provides a range of values where we expect a new observation of Y to lie with a certain level of confidence.
- It is wider than the Confidence Interval for the mean response $E(Y|X = x^*)$ because it accounts for the variability of individual observations around the regression line.
- It is computed as:

$$\widehat{y}^* \pm t_{\frac{\alpha}{2},(n-2)} \cdot s_{pred}$$

where

$$s^2_{pred} = \widehat{\sigma}^2 + se(\widehat{y}^*)^2 \text{ and hence } s_{pred} = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

is the standard error of the predicted value and $\widehat{y}^*$ is the predicted value of Y for a given x^*.

- The confidence level is typically chosen to be 95% ($\alpha = 0.05$).

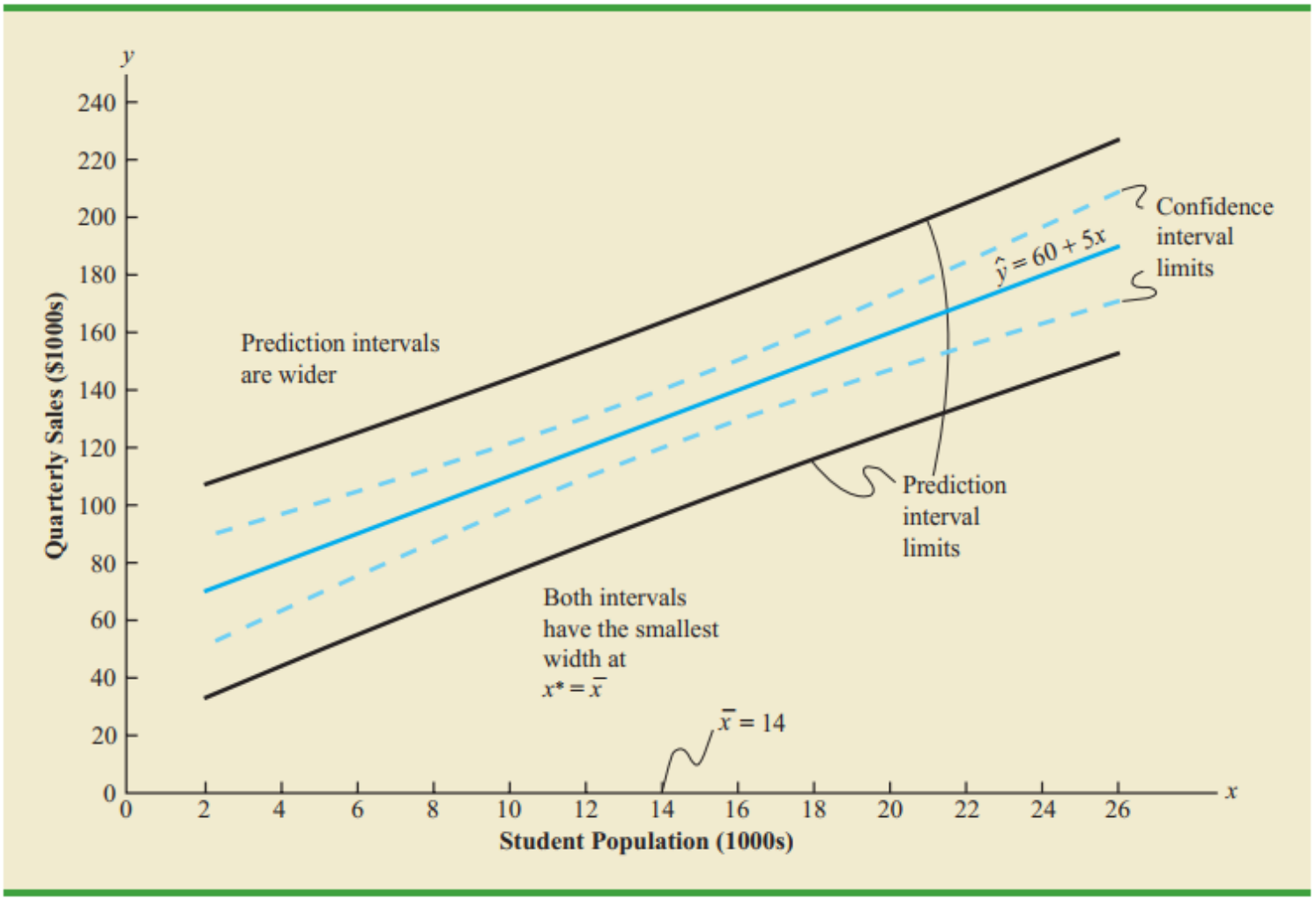
For Sultan's Dine restaurants Sales Dataset given in Example 3.1, the estimated standard deviation corresponding to the prediction of quarterly sales for a new restaurant located need to the campus with 10,000 students, is computed as follows

$$\begin{aligned} s_{pred} &= 13.829\sqrt{1+\frac{1}{10}+\frac{(10-14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

The 95% prediction interval for quarterly sales for the Sultan's Dine restaurants located near campus can be found $t_{\frac{\alpha}{2},(n-1)} = 2.306$. Thus, with $\widehat{y}^* = 110$ and a margin of error of $t_{0.025} \times s_{pred} = 2.306 \times 14.69 = 33.875$, the 95% prediction interval is

$$110 \pm 33.875$$

Confidence and prediction intervals for sales Y at given values of student population X



Example 3.3 (Obstetrics)

As discussed earlier in the Problem & Motivation section (Research Problem 1), obstetricians sometimes order tests to measure estriol levels from 24-hour urine specimens taken from pregnant women who are near term because level of estriol has been found to be related to infant birthweight. The test can provide indirect evidence of an abnormally small fetus. Greene and Touchstone conducted a study to relate birthweight and estriol level in pregnant women. The Sample data are presented in Table 6. They want to find any relationship between the estriol level and birthright How can this relationship be quantified? What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

For the Obstertrics dataset in the Example 3.3, we consider 'birthweight' is the dependent variable and 'estriol' is the independent variable for the problem because estriol levels are being used to try to predict birthweight. The relationship between estriol level and birthweight can be quantified by fitting a regression line that relates the two variables.

Table 6: Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

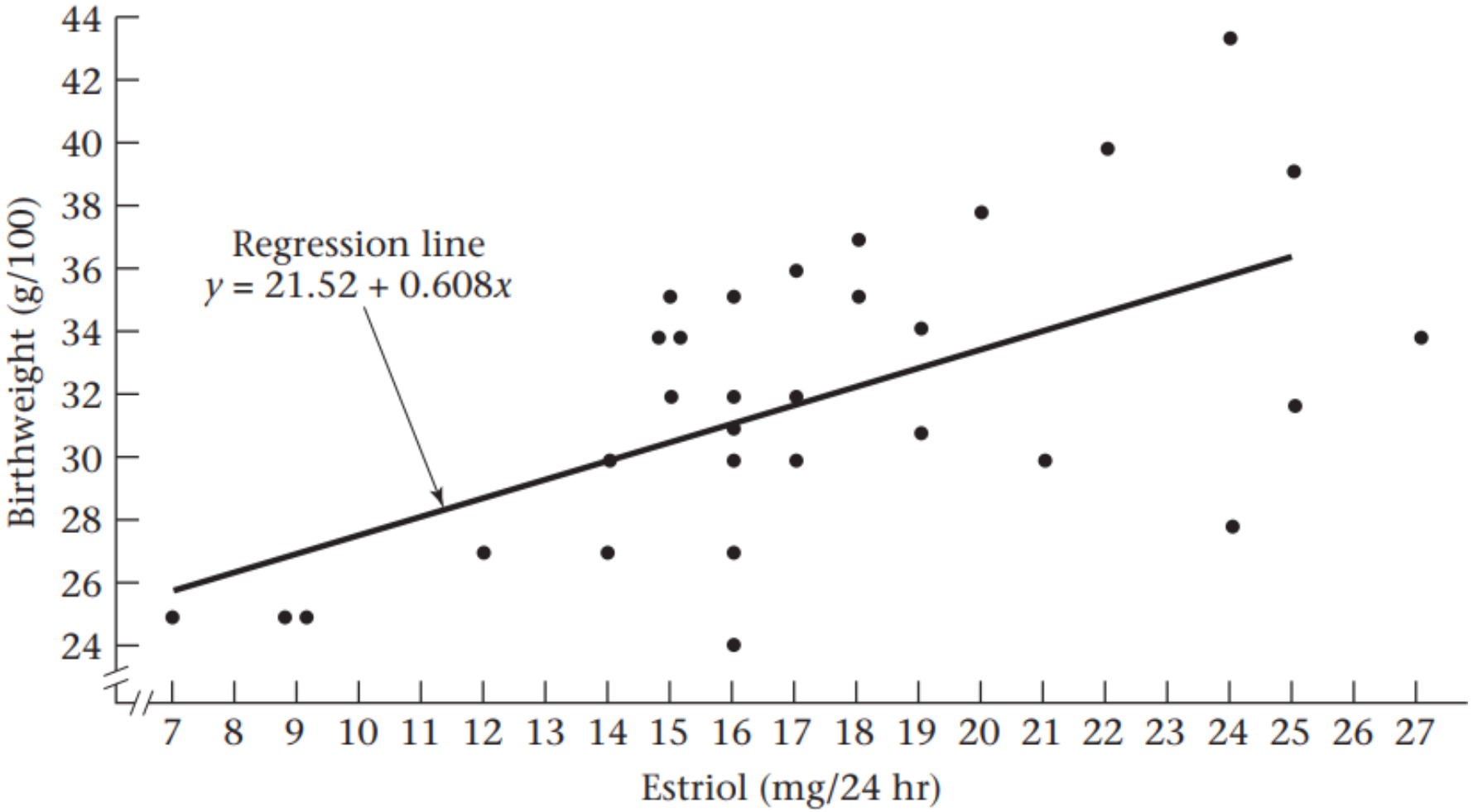


Figure 6: Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

To explore the relationship between estriol levels and birthweight, we created a scatter plot, as shown in Figure 6. If x = estriol level and y = birthweight, then we can postulate a linear relationship between y and x that is of the following form:

$$E(y|x) = \beta_0 + \beta_1 x$$

where $E(y|x)$ = expected or average birthweight (y) among women with a given estriol level (x). That is, for a given estriol-level x, the average birthweight $E(y|x) = \beta_0 + \beta_1 x$.

Let's assume e follows a normal distribution, with mean 0 and variance σ^2. The full linear-regression model then takes the following form:

$$y = \beta_0 + \beta_1 x + e$$

For the data given in Table 6, we have

$$\sum_{i=1}^{31} x_i = 534, \quad \sum_{i=1}^{31} x_i^2 = 9876, \quad \sum_{i=1}^{31} y_i = 992, \quad \sum_{i=1}^{31} x_i y_i = 17500$$

For computing the slope and intercept of the regression line, we consider the least square estimator of β_0 and β_1. These are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 0.608 \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 21.52$$

Thus, the regression line is given by

$$\widehat{y} = 21.52 + 0.608x$$

. This regression line is shown in Figure 6. The slope of 0.608 tells us that the predicted y increases by about 0.61 units per 1 mg/24 hr. Thus, the predicted birthweight increases by 61 g for every 1 mg/24 hr increase in estriol.

What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

- If $\widehat{y} = 2500/100 = 25$, then x can be obtained from the equation

$$25 = 21.52 + 0.608x$$

or

$$x = (25 - 21.52) \times 0.608 = 5.72$$

- Thus, if a woman has an estriol level of 5.72 mg/24 hr, then the predicted birthweight is 2500 g. Furthermore, the predicted infant birthweight for all women with estriol levels of ≤ 5 mg/24 hr is $< 2500g$ (assuming estriol can only be measured in increments of 1 mg/24 hr). This level could serve as a critical value for identifying high-risk women and trying to prolong their pregnancies.

Regression Parameter Confidence Intervals

- Standard errors and 95% confidence intervals for the regression parameters of the birthweight-estriol data given in the Example 3.3:
  - From the data, we have

$$se(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{14.60}{677.42}} = 0.147$$

  - a 95% confidence interval for β_1 is obtained from

$$0.608 \pm t_{29,0.025} \times 0.147 = 0.608 \pm 2.045(0.147)$$

$$= 0.608 \pm 0.300 = (0.308, 0.908)$$

- a 95% confidence interval for β_0 is
  - $\sum_{i=1}^{31} x_i = 534$, $x = 17.23$
  - Standard error of $\widehat{\beta}_0$: $se(\widehat{\beta}_0) = 2.62$
  - 95% confidence interval for β_0: $21.52 \pm t_{29,.975}(2.62) = (16.16, 26.88)$
- These intervals are rather wide due to the small sample size.

Assess the significant effect of birthweight on estriol level: that is, test the hypothesis is $H_0 : \beta_1 = 0$

- Thus,

$$t_{cal} = \widehat{\beta}_1 / se(\widehat{\beta}_1) = 0.608/0.147 = 4.14 \sim t_{29} \text{ under } H_0$$

- We find $t_{29,0.025} = 3.659$. Since $t_{cal} > t_{29,0.025}$, we reject H_0.
- We already have a 95% confidence interval for β_1 is $(0.308, 0.908)$.

What is your conclusion?

- To determine whether to reject the null hypothesis $H_0 : \beta_1 = 0$ based on the confidence interval $(0.308, 0.908)$, we need to check if the interval contains the value 0.
- Since the confidence interval $(0.308, 0.908)$ does not contain the value 0, we can reject the null hypothesis $H_0 : \beta_1 = 0$ at the 0.05 significance level. This means that there is evidence to suggest that the slope coefficient β_1 is not equal to zero in the regression model.

Residual Analysis

Residuals are the differences between the observed values of the dependent variable and the values predicted by the regression model. Residual analysis is a critical component of regression analysis as it helps to determine whether the assumptions made about the regression model appear to be valid. Key Aspects of Residual Analysis:

1. **Linearity**: Plot the residuals against the predicted values. A random scatter around zero suggests a linear relationship between the dependent and independent variables. Any systematic patterns (e.g., curves or clusters) indicate potential issues with linearity.
2. **Constant Variance (Homoscedasticity)**: Plot the residuals against the predicted values. The spread of residuals should remain roughly constant across all levels of the independent variable. Unequal spread or patterns in the residuals suggest heteroscedasticity, violating the assumption of constant variance.

3. **Normality**: Examine the distribution of residuals using a histogram or a Q-Q plot. Ideally, residuals should follow a normal distribution, allowing for reliable inference. Departures from normality may indicate potential issues, such as skewness or heavy tails, which could affect the validity of statistical tests.
4. **Influential Points**: residual analysis is also used to identify outliers and influential observations
   — high leverage points

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \quad ; \quad i = 1, \ldots, n$$

These points can significantly affect parameter estimates and may warrant further investigation.

5. Addressing Issues: If residual analysis reveals any discrepancies from the assumptions of the regression model, corrective actions may be necessary. This could involve transforming variables, removing outliers, using robust regression techniques, or exploring alternative model specifications.

Influential Observations, Outliers, and Cook's Distance

Influential Observations

Influential observations are data points that have a large impact on the estimated coefficients of the regression model. They can significantly alter the fit of the model if removed. Influential observations are identified using Cook's distance, where observations with Cook's distance greater than $4/n$ (where n is the number of observations) are considered influential.

Outliers

Outliers are data points that deviate significantly from the rest of the data. They can affect the regression model's accuracy and should be investigated to determine if they are genuine data points or errors. Outliers are identified by selecting observations with Cook's distance greater than a certain threshold (here, $4/n$).

Cook's Distance

Cook's Distance

Cook's distance measures the influence of each observation on the fitted values of the model. It combines the effect of leverage and residual to determine the influence of each data point.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot \text{MSE}}$$

where $\hat{y}_j$ is the jth fitted value, $\hat{y}_{j(i)}$ is the jth fitted value with the ith observation removed, p is the number of predictors in the model, and MSE is the mean squared error of the model.

Python Code: OLS for Linear Regression Model

```
# Simple Linear Regression Model
import statsmodels.api as sm
model =sm.OLS(boston.MEDV, sm.add_constant(boston.LSTAT))
result = model.fit()
print(result.summary())
```

```
import statsmodels.api as sm
model =sm.OLS(boston.MEDV, sm.add_constant(boston.LSTAT))
result = model.fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.544
Model:                            OLS   Adj. R-squared:                  0.543
Method:                 Least Squares   F-statistic:                     601.6
Date:                Fri, 27 Jan 2023   Prob (F-statistic):           5.08e-88
Time:                        03:26:52   Log-Likelihood:                -1641.5
No. Observations:                 506   AIC:                             3287.
Df Residuals:                     504   BIC:                             3295.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.5538      0.563     61.415      0.000      33.448      35.659
LSTAT         -0.9500      0.039    -24.528      0.000      -1.026      -0.874
==============================================================================
Omnibus:                      137.043   Durbin-Watson:                   0.892
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              291.373
Skew:                           1.453   Prob(JB):                     5.36e-64
Kurtosis:                       5.319   Cond. No.                         29.7
==============================================================================
```

R Code: Linear Regression Model

```
## Linear Regression Model
#data<-read.table(file.choose(),sep ="", header=T)
# Create the data frame
data <- data.frame(
x= c(2, 6, 8, 8, 12, 16, 20, 20, 22, 26),
y = c(58, 105, 88, 118, 117, 137, 157, 169, 149, 202)
)

# Perform simple linear regression
model <- lm(y ~ x, data = data)

# Print summary of the regression model
summary(model)
```

R Code: Residual Analysis

```r
# Get residuals
residuals <- resid(model)

# Residual plot
plot(model$fitted.values, residuals,
xlab = "Fitted values",
ylab = "Residuals",
main = "Residual Plot")
abline(h = 0, col = "red", lty = 2) # Add a horizontal line at
    y=0

# Histogram of residuals
hist(residuals,
main = "Histogram of Residuals",
xlab = "Residuals",
ylab = "Frequency",
col = "lightblue")

# QQ plot of residuals
qqnorm(residuals)
qqline(residuals)
title("QQ Plot of Residuals")
```

```r
# Residual plot
plot(model, which = 1)
attach(data)

# Cook's distance plot
plot(cook_dist, pch = 20, main = "Cook's Distance Plot", xlab =
    "Observation", ylab = "Cook's Distance")
abline(h = 4 / length(y), col = "red", lty = 2) # Add a
    horizontal line at the threshold

# Get influence measures
infl <- influence(model)

# Leverage points
leverage <- infl$hat

# Influential observations
cook_dist <- cooks.distance(model)

# Outliers
outliers <- which(cook_dist > 4 / length(y))
```

```r
# Print leverage points, influential observations, and outliers
print("Leverage points:")
print(which(leverage > mean(leverage) + 2 * sd(leverage)))

print("Influential observations:")
print(which(cook_dist > 4 / length(y)))

print("Outliers:")
print(outliers)
```

Python Code: Linear Regression Model

```
import numpy as np
import statsmodels.api as sm

# Define the data
x = np.array([2, 6, 8, 8, 12, 16, 20, 20, 22, 26])
y= np.array([58, 105, 88, 118, 117, 137, 157, 169, 149, 202])

# Add constant for intercept
x_with_const = sm.add_constant(x)

# Fit the model
model = sm.OLS(y, x_with_const).fit()

# Print summary of the regression model
print(model.summary())
```

Python Code: Residual Analysis

```
# Get residuals
residuals = model.resid

# Residual plot
plt.scatter(model.predict(), residuals)
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.axhline(y=0, color='r', linestyle='--') # Add a horizontal
    line at y=0
plt.show()

# Histogram of residuals
plt.hist(residuals, bins=10)
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Histogram of Residuals')
plt.show()

# QQ plot of residuals
sm.qqplot(residuals, line ='45')
plt.title('QQ Plot of Residuals')
plt.show()
```

```
import numpy as np
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import OLSInfluence

# Define the data
x = np.array([2, 6, 8, 8, 12, 16, 20, 20, 22, 26])
y = np.array([58, 105, 88, 118, 117, 137, 157, 169, 149, 202])

# Fit the linear regression model
X_with_const = sm.add_constant(x)
model = sm.OLS(y, X_with_const).fit()

# Get influence measures
influence = OLSInfluence(model)

# Leverage points
leverage = influence.hat_matrix_diag
```

```
# Influential observations
cook_dist = influence.cooks_distance[0]
print(cook_dist)
# Outliers
outliers = np.where(cook_dist > 4 / len(y))[0]

# Print leverage points, influential observations, and outliers
print("Leverage points:")
print(np.where(leverage > np.mean(leverage) + 2 *
    np.std(leverage))[0])

print("Influential observations:")
print(np.where(cook_dist > 4 / len(y))[0])

print("Outliers:")
print(outliers)

# Get influence measures
influence = OLSInfluence(model)

# Residual plot
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
plt.scatter(model.fittedvalues, model.resid)
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.axhline(y=0, color='red', linestyle='--')
```

```
# Cook's distance plot
plt.subplot(1, 2, 2)
plt.scatter(np.arange(len(cook_dist)), cook_dist, marker='o',
    color='blue')
plt.axhline(y=4 / len(y), color='red', linestyle='--')
plt.xlabel('Observation')
plt.ylabel("Cook's Distance")
plt.title("Cook's Distance Plot")

plt.tight_layout()
plt.show()
```

Chapter 3: Multiple Linear Regression Model

4 Chapter 3: Multiple Linear Regression Model

Problem & Motivation

Research Problem 2

Suppose age (days), birthweight (oz), and SBP are measured for 16 infants and the data are as shown in Table 7. What is the relationship between infant systolic blood pressure (SBP) and their age and birthweight? Can we predict SBP based on these factors?

Table 7: Sample data for infant blood pressure, age, and birthweight for 16 infants

i	Age (days) (x_1)	Birthweight (oz) (x_2)	SBP (mm Hg) (y)
1	3	135	89
2	4	120	90
3	3	100	83
4	2	105	77
5	4	130	92

Table 7 (continue)

i	Age (days) (x_1)	Birthweight (oz) (x_2)	SBP (mm Hg) (y)
6	5	125	98
7	2	125	82
8	3	105	85
9	5	120	96
10	4	90	95
11	2	120	80
12	3	95	79
13	3	120	86
14	4	150	97
15	3	160	92
16	3	125	88

Problem & Motivation

Research Problem 3

Let's delve into the exploration of the dataset. We'll utilize the **Boston House Prices Dataset**, consisting of 506 rows and 13 attributes, including a target column.

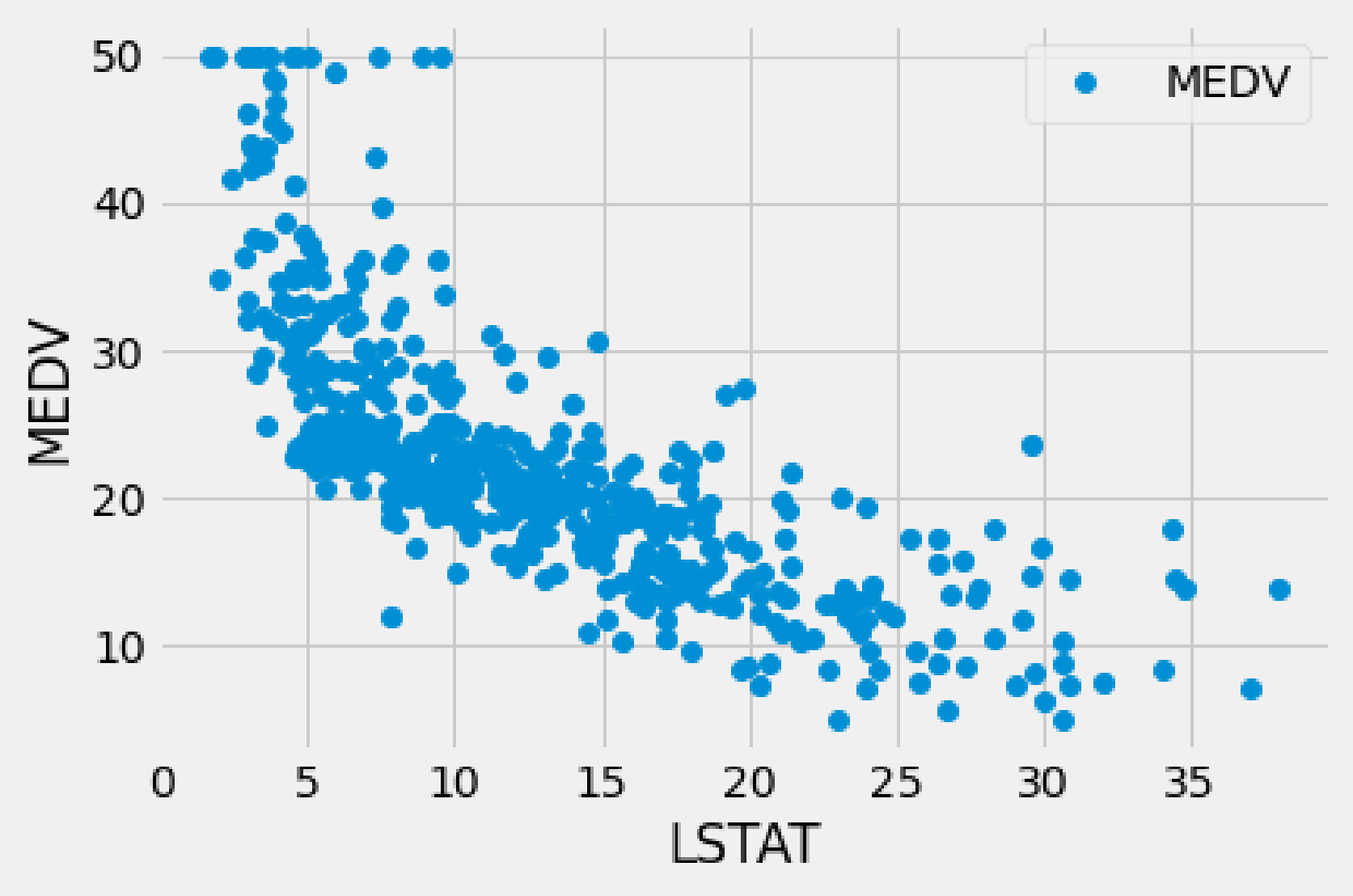
- Our objective is to forecast the median price value of owner-occupied homes. How do you predict the price?

How can we apply multiple regression to predict the price based on various attributes? Let's take a quick look at the dataset.

```
import io
boston = pd.read_csv(io.BytesIO(uploaded['boston.csv']))
# Dataset is now stored in a Pandas Dataframe
boston.head(10)
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
8	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9

- **Crim**: Per capita crime rate by town
- **Zn**: Proportion of residential land zoned for lots over 25,000 sq. ft.
- **Indus**: Proportion of non-retail business acres per town
- **Chas**: Charles River dummy variable (= 1 if tract bounds river; 0, otherwise)
- **Nox**: Nitrogen oxides concentration (parts per 10 million)
- **Rm**: Average number of rooms per dwelling
- **Age**: Proportion of owner-occupied units built before 1940
- **Dis**: Weighted mean of distances to five Boston employment centers
- **Rad**: Index of accessibility to radial highways
- **Tax**: Full-value property tax rate per \$10,000
- **Ptratio**: Pupil–Teacher ratio by town
- **B**: $1000(Bk - 0.63)^2$, where Bk is the proportion of Blacks by town
- **Lstat**: Lower status of the population (percent)
- **Medv**: Median Price value of owner-occupied homes in \$1000s


MEDV
50
40
30
20
10
MEDV
0
5
10
15
20
25
30
35
LSTAT

Python Code

```
from google.colab import files
uploaded = files.upload()
import io
boston = pd.read_csv(io.BytesIO(uploaded['boston.csv']))
# Dataset is now stored in a Pandas Dataframe
boston.head()
import matplotlib.pyplot as plt
boston.plot(x='LSTAT',y='MEDV',style='o')
plt.xlabel('LSTAT')
plt.ylabel('MEDV')
plt.show()
```

Multiple Regression Model

The multiple regression model is given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$

where:

- y_i is the dependent variable (response variable) for the ith observation.
- $x_{1i}, x_{2i}, \ldots, x_{pi}$ are the independent variables for the ith observation.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients corresponding to the independent variables $x_{1i}, x_{2i}, \ldots, x_{pi}$.
- e_i is the error term or residual for the ith observation.

Model Assumptions

Under the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$

the following assumptions are made:

1. **Linearity**: The relationship between the dependent variable y_i and the independent variables $x_{1i}, x_{2i}, \ldots, x_{pi}$ is linear.
2. **Independence of Errors**: The error terms e_i are independent of each other.
3. **Homoscedasticity**: The variance of the error terms e_i is constant for all values of the independent variables.
4. **Normality of Errors**: The error terms e_i are normally distributed with mean zero.

5. **No Perfect Multicollinearity**: There is no perfect linear relationship among the independent variables.
6. **No Autocorrelation**: The errors (e) are not correlated with each other over time or across observations.

These assumptions are essential for valid estimation and interpretation of the classical regression model.

Estimation Procedure by OLS in Matrix Notation

To estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ in the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$

using ordinary least squares (OLS) in matrix notation, we define:

$$\vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

In this case, the model can be written as

$$\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{e}$$

To find the estimated coefficients $\widehat{\beta}$ of β using ordinary least squares (OLS), we minimize the sum of squared residuals e:

$$\vec{e} = \vec{Y} - \boldsymbol{X}\widehat{\beta}$$

Calculation of Estimated Coefficients

The sum of squared residuals is given by:

$$\text{SSE} = \vec{e}^T \vec{e} = \left(\vec{Y} - \boldsymbol{X}\vec{\hat{\beta}}\right)^T \left(\vec{Y} - \boldsymbol{X}\vec{\hat{\beta}}\right)$$

To minimize SSE, we take the derivative with respect to $\vec{\hat{\beta}}$ and set it to zero:

$$\frac{\partial \text{SSE}}{\partial \vec{\hat{\beta}}} = -2\boldsymbol{X}^T(\vec{Y} - \boldsymbol{X}\vec{\hat{\beta}}) = 0$$

Solving for $\vec{\hat{\beta}}$, we get:

$$\vec{\hat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y}$$

This equation provides the estimated coefficients $\widehat{\beta}$ that minimize the sum of squared residuals.

Estimation Procedure of Error Variance

To estimate the error variance $\widehat{\sigma}^2$, we use the following formula:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\left(\vec{Y} - \boldsymbol{X}\vec{\widehat{\beta}}\right)^T\left(\vec{Y} - \boldsymbol{X}\vec{\widehat{\beta}}\right)$$

where:

- $\vec{\widehat{\beta}}$ is the vector of estimated coefficients obtained using ordinary least squares (OLS).
- n is the number of observations.
- p is the number of independent variables (excluding the intercept).

The term $\left(\vec{Y} - \boldsymbol{X}\vec{\widehat{\beta}}\right)^T\left(\vec{Y} - \boldsymbol{X}\vec{\widehat{\beta}}\right)$ represents the sum of squared residuals, which measures the unexplained variability in the dependent variable after accounting for the effects of the independent variables. Dividing by $n-p-1$, the degrees of freedom for the error term, provides an unbiased estimate of the error variance $\widehat{\sigma}^2$.

Coefficient of Determination (R^2)

The coefficient of determination (R^2) measures the proportion of the variance in the dependent variable (y) that is explained by the independent variables ($x_1, x_2, \ldots, x_p$) in the regression model.

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned}$$

where:

- SSE is the sum of squared errors (residuals), representing the unexplained variability in the dependent variable.
- SST is the total sum of squares, representing the total variability in the dependent variable.

Interpretation:

- R^2 ranges from 0 to 1, where a higher value indicates a better fit of the regression model to the data.
- R^2 represents the proportion of the variance in the dependent variable that is explained by the independent variables.
- For example, if $R^2 = 0.75$, it means that 75

Adjusted R^2

- the adjusted R^2 is denoted by $\bar{R}^2$ and is defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the total number of explanatory variables in the model (not including the constant term), and n is the sample size

- it can also be written as:

$$\bar{R}^2 = 1 - \frac{SSE/\text{df}_e}{SST/\text{df}_t}$$

where df_t is the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance

Hence,

$$\bar{R}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- the explanation of this statistic is almost the same as R^2 but it penalizes the statistic as extra variables are included in the model
- the term $\frac{n-1}{n-p-1}$ is called the penalty of using the more regressors in a model
- when the number of regressors p, increases, $(1-R^2)$ will decrease, but $\frac{n-1}{n-p-1}$ will increase
- whether more regressors improve the explanatory power of a model depends on the trade-of between R^2 and the penalty

Interpretation:

- $\bar{R}^2$ penalizes the addition of unnecessary predictors to the model, unlike R^2.
- It is always less than or equal to R^2, and it increases only if the new term improves the model more than would be expected by chance.
- Therefore, $\bar{R}^2$ is often preferred for comparing the goodness of fit of models with different numbers of predictors.

Example Dataset and Regression Calculations

Consider a dataset with 10 observations and two independent variables (x_1 and x_2).

i	x_{1i}	x_{2i}	y_i
1	9	16	10
1	13	14	12
1	11	10	14
1	11	8	16
1	14	11	18
1	15	17	20
1	16	9	22
1	20	16	24
1	15	12	26
1	15	12	28

To fit a multiple regression model, we use the least squares method to estimate the coefficients β_0, β_1, and β_2.

Calculation of $\vec{\hat{\beta}}$

For the given dataset, we have: $\boldsymbol{X} = \begin{bmatrix} 1 & 9 & 16 \\ 1 & 13 & 14 \\ 1 & 11 & 10 \\ 1 & 11 & 8 \\ 1 & 14 & 11 \\ 1 & 15 & 17 \\ 1 & 16 & 9 \\ 1 & 20 & 16 \\ 1 & 15 & 12 \\ 1 & 15 & 12 \end{bmatrix}$; $\vec{Y} = \begin{bmatrix} 10 \\ 12 \\ \vdots \\ 28 \end{bmatrix}$.

Now, let's calculate $\boldsymbol{X}^T\boldsymbol{X}$, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $\boldsymbol{X}^T\vec{Y}$, and $\vec{\hat{\beta}}$.

Calculation of $\boldsymbol{X}^T\boldsymbol{X}$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 9 & 13 & 11 & \dots & 15 \\ 16 & 14 & 10 & \dots & 12 \end{bmatrix} \begin{bmatrix} 1 & 9 & 16 \\ 1 & 13 & 14 \\ 1 & 11 & 10 \\ \vdots & \vdots & \vdots \\ 1 & 15 & 12 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 139 & 125 \\ 139 & 2019 & 1757 \\ 125 & 1757 & 1651 \end{bmatrix}$$

Calculation of $\vec{\hat{\beta}}$

Using matrix inversion, we find:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{\det(\boldsymbol{X}^T\boldsymbol{X})}\text{adj}(\boldsymbol{X}^T\boldsymbol{X}) = \begin{bmatrix} 3.369 & -0.135 & -0.112 \\ -0.135 & 0.012 & -0.003 \\ -0.112 & -0.003 & 0.012 \end{bmatrix}$$

Now, we calculate

$$\vec{\hat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y} = \begin{bmatrix} 2.821 \\ 1.591 \\ -0.475 \end{bmatrix}$$

Calculation of Error Variance Estimation

i	x_{1i}	x_{2i}	y_i	$\widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
1	9	16	10	9.542	0.209764
1	13	14	12	16.856	23.58074
1	11	10	14	15.573	2.474329
1	11	8	16	16.523	0.273529
1	14	11	18	19.871	3.500641
1	15	17	20	18.613	1.923769
1	16	9	22	24.003	4.012009
1	20	16	24	27.043	9.259849
1	15	12	26	20.988	25.12014
1	15	12	28	20.988	49.16814
Total	139	125	190	190	119.5229

- The error variance estimation, $\widehat{\sigma}^2$, is calculated as:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{119.5229}{10-2-1} = 17.0747$$

where n is the number of observations, p is the number of predictors, y_i is the observed value, and $\widehat{y}_i$ is the predicted value.

- Coefficient of determination:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{119.522}{330} = 0.6378$$

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} = 1 - \frac{(1-0.6378)(10-1)}{10-2-1} = 0.5343$$

Goodness-of-fit R^2 and Adjusted R^2

- The coefficient of determination $R^2 = 0.6378$ suggests that approximately 63.78% of the variance in the dependent variable (y) can be explained by the independent variables x_1 and x_2 included in the model.
- The Adjusted R^2 value takes into account the number of predictors and the sample size, providing a more conservative estimate of the model's goodness-of-fit. In this case, $\bar{R}^2 = 0.5343$ indicating that approximately 53.43% of the variance in the dependent variable(y) is explained by the independent variables x_1 and x_2 after adjusting for the number of predictors and the sample size.

F-test in Multiple Regression

- The F-test in multiple regression assesses the overall significance of the regression model.
- It tests whether at least one of the independent variables has a non-zero coefficient.
- The null hypothesis H_0 for the F-test is:

$$H_0 : \widehat{\beta}_1 = \widehat{\beta}_2 = \cdots = \widehat{\beta}_p = 0$$

where $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ are the coefficients of the independent variables.

ANOVA Table in Multiple Regression

- The ANOVA table in multiple regression assesses the overall significance of the regression model.
- It partitions the total variance in the dependent variable into explained variance and unexplained variance.
- The table includes sums of squares (SS), degrees of freedom (df), mean squares (MS), and the F-test statistic.

Source of Variation	**SS**	**df**	**MS**	**F**
Regression	SSR	p	$MSR = \frac{SSR}{p}$	
Residual (Error)	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	$F = \frac{MSR}{MSE}$
Total	SST	$n-1$		

- Reject the null hypothesis H_0 if the calculated F-statistic is greater than the critical value from the F-distribution.

Decision Rule for ANOVA F-test

Classical Approach

- Set the significance level α.
- Calculate the critical value F_{critical} from the F-distribution with appropriate degrees of freedom.
- Decision Rule:
  - If calculated $F > F_{\text{critical}}$, reject H_0 and conclude that the regression model is statistically significant.
  - If calculated $F \leq F_{\text{critical}}$, fail to reject H_0 and conclude that the regression model is not statistically significant.

Decision Rule for ANOVA F-test (Cont'd)

p-value Approach

- Calculate the p-value associated with the calculated F-statistic.
- Decision Rule:
  - If p-value $< \alpha$, reject H_0 and conclude that the regression model is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the regression model is not statistically significant.

Example: ANOVA Table

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	210.4651	105.2326	6.1625	0.0286
Residual	7	119.5349	17.0764		
Total	9	330			

The t-tests in Multiple Regression

- The t-tests in multiple regression assess the significance of individual coefficients (parameters) in the model.
- Each t-test tests the null hypothesis that the corresponding coefficient is zero.
- The t-test statistic for the coefficient $\widehat{\beta}_i$ is calculated as:

$$t = \frac{\widehat{\beta}_i}{\text{SE}(\widehat{\beta}_i)}$$

where $\widehat{\beta}_i$ is the estimated coefficient, and $\text{SE}(\widehat{\beta}_i)$ is its standard error.

Decision Rule for t-tests in Multiple Regression

Classical Approach

- Set the significance level α.
- Calculate the critical value t_{critical} from the t-distribution with appropriate degrees of freedom.
- Decision Rule for each $\widehat{\beta}_i$:
  - If $|t| > t_{\text{critical}}$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is statistically significant.
  - If $|t| \leq t_{\text{critical}}$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is not statistically significant.

Decision Rule for t-tests in Multiple Regression (Cont'd)

p-value Approach

- Calculate the p-value associated with each calculated t-statistic.
- Decision Rule for each $\widehat{\beta}_i$:
  - If p-value $< \alpha$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is not statistically significant.

ANOVA				330	
	df	*SS*	*MS*	*F*	*Significance F*
Regressior	2	210.4651	105.2326	6.1625	0.0286
Residual	7	119.5349	17.0764		
Total	9	330			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.821	7.585	0.372	0.721	-15.116	20.757
x1	1.591	0.455	3.499	0.010	0.516	2.666
x2	-0.475	0.451	-1.054	0.327	-1.540	0.591

Example 4.1

As discussed earlier in the Problem & Motivation section (Research Problem 2), age (days), birthweight (oz), and SBP are measured for 16 infants and the data are as shown in Table 7. What is the relationship between infant systolic blood pressure (SBP) and their age and birthweight? Can we predict SBP based on these factors?

Solution of the Research Problem 2

The multiple regression model for this problem is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad ; \quad i = 1, 2, \ldots, 16$$

where

- y is the SBP of infants
- x_1 is the Age of the infants
- x_2 is the birthweight (oz) of the infants
- e is the error term

According to the parameter-estimate column in python output, the regression equation is given by

$$y_i = 53.45 + 5.89 x_{1i} + 0.126 x_{2i} \quad ; \quad i = 1, 2, \ldots, 16$$

The regression equation tells us that for a newborn, the average blood pressure increases by an estimated 5.89 mm Hg per day of age and 0.126 mm Hg per ounce of birthweight.

Python Code: Hypertension Dataset

```
import pandas as pd
import statsmodels.api as sm

# Creating the dataframe from the provided data
data = {
  'Age': [3, 4, 3, 2, 4, 5, 2, 3, 5, 4, 2, 3, 3, 4, 3, 3],
  'Birthweight': [135, 120, 100, 105, 130, 125, 125, 105, 120,
      90, 120, 95, 120, 150, 160, 125],
  'SBP': [89, 90, 83, 77, 92, 98, 82, 85, 96, 95, 80, 79, 86,
      97, 92, 88]
}
df = pd.DataFrame(data)

# Adding a constant term for the intercept
X = sm.add_constant(df[['Age', 'Birthweight']])
y = df['SBP']

# Fitting the multiple regression model
model = sm.OLS(y, X).fit()

# Printing the summary of the regression model
print(model.summary())
```

OLS Regression Results

Dep. Variable:	SBP	R-squared:	0.881
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	48.08
Date:	Fri, 19 Apr 2024	Prob (F-statistic):	9.84e-07
Time:	02:49:38	Log-Likelihood:	-35.569
No. Observations:	16	AIC:	77.14
Df Residuals:	13	BIC:	79.46
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.4502	4.532	11.794	0.000	43.660	63.241
Age	5.8877	0.680	8.656	0.000	4.418	7.357
Birthweight	0.1256	0.034	3.657	0.003	0.051	0.200

Omnibus:	11.884	Durbin-Watson:	2.214
Prob(Omnibus):	0.003	Jarque-Bera (JB):	9.313
Skew:	1.222	Prob(JB):	0.00950
Kurtosis:	5.828	Cond. No.	892.

R Code: Hypertension Dataset

```
# Creating the dataframe from the provided data
data <- data.frame(
Age = c(3, 4, 3, 2, 4, 5, 2, 3, 5, 4, 2, 3, 3, 4, 3, 3),
Birthweight = c(135, 120, 100, 105, 130, 125, 125, 105, 120, 90,
    120, 95, 120, 150, 160, 125),
SBP = c(89, 90, 83, 77, 92, 98, 82, 85, 96, 95, 80, 79, 86, 97,
    92, 88)
)

# Fitting the multiple regression model
model <- lm(SBP ~ Age + Birthweight, data = data)

# Printing the summary of the regression model
summary(model)
```

```
> # Multiple Regression Model
> # Creating the dataframe from the provided data
> data <- data.frame(
+   Age = c(3, 4, 3, 2, 4, 5, 2, 3, 5, 4, 2, 3, 3, 4, 3, 3),
+   Birthweight = c(135, 120, 100, 105, 130, 125, 125, 105, 120, 90, 120, 95, 120, 150, 160, 125),
+   SBP = c(89, 90, 83, 77, 92, 98, 82, 85, 96, 95, 80, 79, 86, 97, 92, 88)
+ )
>
> # Fitting the multiple regression model
> model <- lm(SBP ~ Age + Birthweight, data = data)
>
> # Printing the summary of the regression model
> summary(model)

Call:
lm(formula = SBP ~ Age + Birthweight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0438 -1.3481 -0.2395  0.9688  6.6964

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.45019    4.53189  11.794 2.57e-08 ***
Age          5.88772    0.68021   8.656 9.34e-07 ***
Birthweight  0.12558    0.03434   3.657   0.0029 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.479 on 13 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.8626
F-statistic: 48.08 on 2 and 13 DF,  p-value: 9.844e-07
```

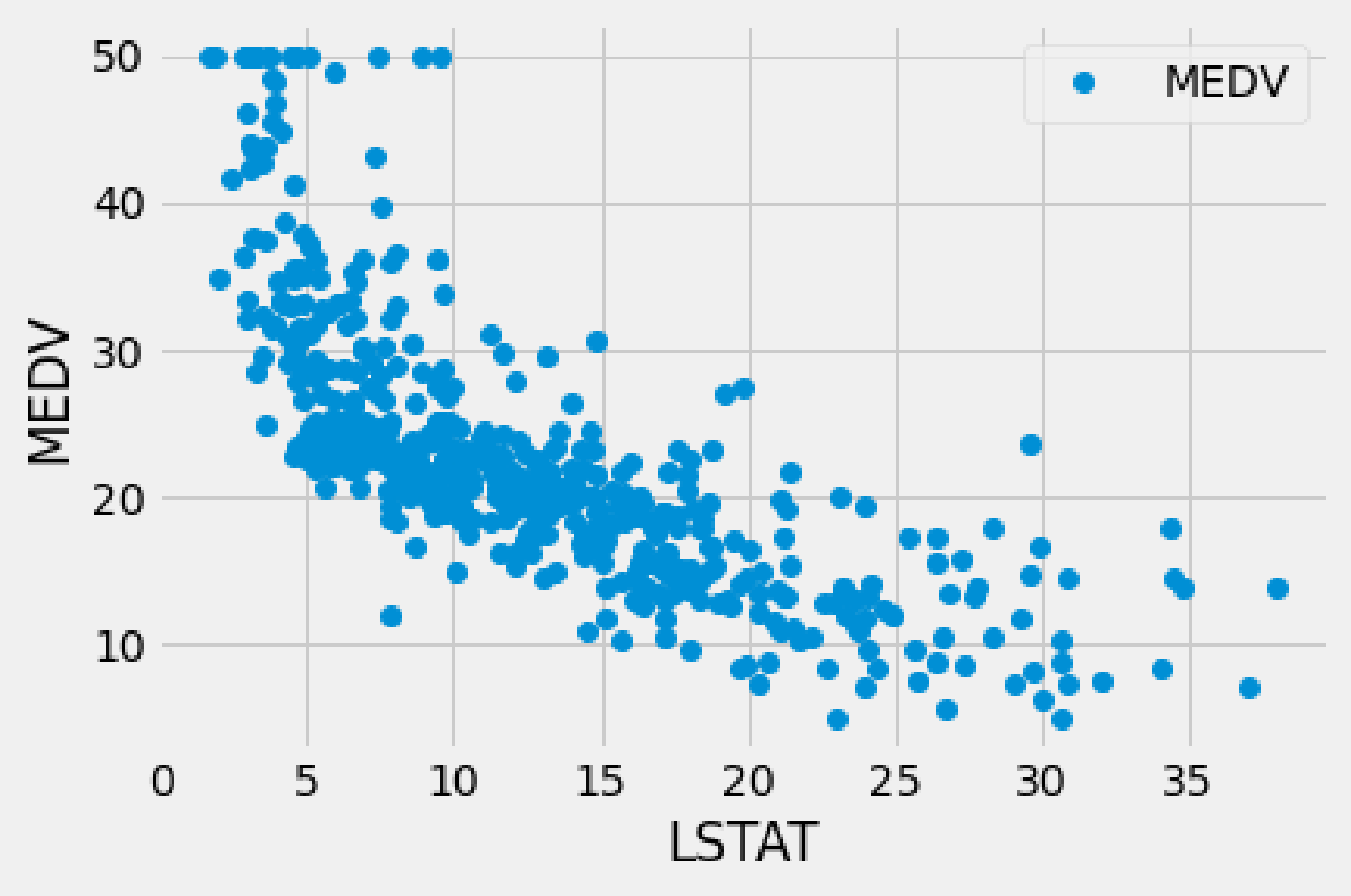
Example 4.2

As discussed earlier in the Problem & Motivation section (Research Problem 3), we want to forecast the median price value of owner-occupied homes. How do you predict the price? How can we apply multiple regression to predict the price based on various attributes? Let's take a quick look at the dataset.

```
import io
boston = pd.read_csv(io.BytesIO(uploaded['boston.csv']))
# Dataset is now stored in a Pandas Dataframe
boston.head(10)
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
8	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9

- **Crim**: Per capita crime rate by town
- **Zn**: Proportion of residential land zoned for lots over 25,000 sq. ft.
- **Indus**: Proportion of non-retail business acres per town
- **Chas**: Charles River dummy variable (= 1 if tract bounds river; 0, otherwise)
- **Nox**: Nitrogen oxides concentration (parts per 10 million)
- **Rm**: Average number of rooms per dwelling
- **Age**: Proportion of owner-occupied units built before 1940
- **Dis**: Weighted mean of distances to five Boston employment centers
- **Rad**: Index of accessibility to radial highways
- **Tax**: Full-value property tax rate per \$10,000
- **Ptratio**: Pupil–Teacher ratio by town
- **B**: $1000(Bk - 0.63)^2$, where Bk is the proportion of Blacks by town
- **Lstat**: Lower status of the population (percent)
- **Medv**: Median Price value of owner-occupied homes in \$1000s


MEDV
50
40
30
20
10
MEDV
0
5
10
15
20
25
30
35
LSTAT

Python Code: OLS for Multiple Linear Regression Model

```
# Multiple Linear Regression Model
import statsmodels.api as sm
X = boston[['LSTAT', 'CRIM']]
model =sm.OLS(boston.MEDV, sm.add_constant(X))
result = model.fit()
print(result.summary())

## Alternatively
import statsmodels.formula.api as smf
# formula: response ~ predictor + predictor
est = smf.ols(formula='MEDV ~ LSTAT + CRIM',
    data=boston).fit()
print(est.summary())
# In GLM framework
Gaussian_model = sm.GLM(boston.MEDV, sm.add_constant(X),
    family=sm.families.Gaussian()).fit()
print(Gaussian_model.summary())
```

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.546
Method:	Least Squares	F-statistic:	304.4
Date:	Fri, 27 Jan 2023	Prob (F-statistic):	2.33e-87
Time:	03:37:58	Log-Likelihood:	-1639.6
No. Observations:	506	AIC:	3285.
Df Residuals:	503	BIC:	3298.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	34.3192	0.574	59.816	0.000	33.192	35.446
LSTAT	-0.9114	0.043	-21.004	0.000	-0.997	-0.826
CRIM	-0.0704	0.036	-1.956	0.051	-0.141	0.000

Omnibus:	146.159	Durbin-Watson:	0.872
Prob(Omnibus):	0.000	Jarque-Bera (JB):	329.172
Skew:	1.517	Prob(JB):	3.32e-72
Kurtosis:	5.531	Cond. No.	32.7