Chapter Four

Correlation and Regression

4-1 Introduction

- Inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.

- **Correlation** is a statistical method used to determine whether a relationship between variables exists.

- **Regression** is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

4-1 Introduction

- In a **simple relationship**, there are only two types of variables under study; an **independent variable**, and a **dependent variable**.

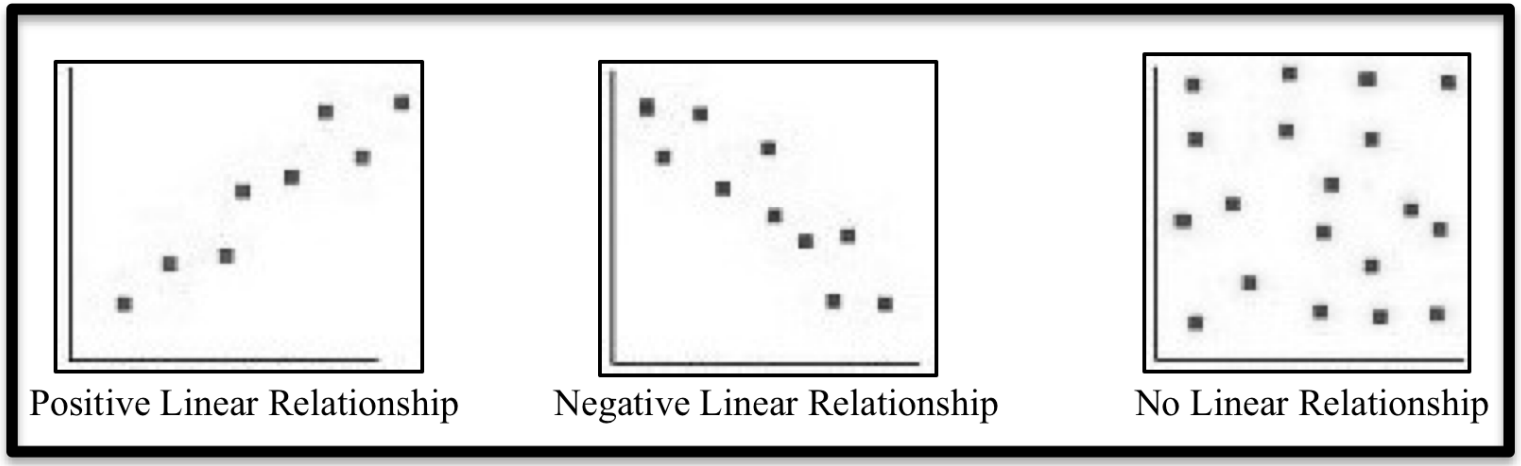
- Simple relationship can be positive or negative:

  - A **positive relationship** exists when both variables increase or decrease at the same time.

  - A **negative relationship** exists when one variable increases and the other variable decreases.

4-2 Scatter Plots

- A **scatter plot** is a visual way to describe the nature of the relationship between the independent and dependent variables.



4-3 Correlation Coefficient

- The **correlation coefficient** is measure of how variables are related, it measures the strength and direction of a linear relationship between two variables.
  - The symbol for the population correlation coefficient is ρ (rho).
  - The symbol for the sample correlation coefficient is r.
  - The range of the correlation coefficient is from -1 to $+1$.

4-3 Correlation Coefficient

- If there is a **strong positive linear relationship** between the variables, the value of ***r*** will be close to +1.
- If there is a **strong negative linear relationship** between the variables, the value of ***r*** will be close to –1.
- When there is **no linear relationship** between the variables or only a weak relationship, the value of *r* will be close to 0.

4-3 Correlation Coefficient

Correlation Coefficient Value	Meaning
+1	Complete Positive Linear Relationship
0.70 ⇔ 0.99	Strong Positive Linear Relationship
0.50 ⇔ 0.69	Moderate Positive Linear Relationship
0.01 ⇔ 0.49	Weak Positive Linear Relationship
0	No Linear Relationship
-0.01 ⇔ -0.49	Weak Negative Linear Relationship
-0.50 ⇔ -0.69	Moderate Negative Linear Relationship
-0.70 ⇔ -0.99	Strong Negative Linear Relationship
-1	Complete Negative Linear Relationship

4-4 Pearson linear correlation coefficient

- Formula for the **Pearson linear correlation coefficient** (r_P)

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

where n is the number of data pairs (sample size).

4-4 Pearson linear correlation coefficient

- Example:
  - A researcher wishes to determine if a person's age is related to the number of hours he or she exercises per week. The data for the sample are shown below.

Age x	18	26	32	38	52	59
Hours y	10	5	2	3	1.5	1

Compute the value of the correlation coefficient.

4-4 Pearson linear correlation coefficient

							$\sum$
Age x	18	26	32	38	52	59	225
Hours y	10	5	2	3	1.5	1	22.5
x^2	324	676	1024	1444	2704	3481	9653
y^2	100	25	4	9	2.25	1	141.25
xy	180	130	64	114	78	59	625

$$r_P = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} = \frac{(6)(625) - (225)(22.5)}{\sqrt{[(6)(9653) - (225)^2][(6)(141.25) - (22.5)^2]}}$$

$$= \frac{3750 - 5062.5}{\sqrt{[57918 - 50625][847.5 - 506.25]}} = \frac{-1312.5}{\sqrt{[7293][341.25]}} = \frac{-1312.5}{\sqrt{2488736.25}}$$

$$= \frac{-1312.5}{1577.57} = -0.83$$

There is a strong negative linear relationship, which means that older people tend to exercise less on average.

4-5 Spearman rank correlation coefficient

- Formula for the **Spearman rank correlation coefficient** (r_s)

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

- where ***d*** = difference in the ranks and ***n*** = number of data pairs

4-5 Spearman rank correlation coefficient

- Example:
  - The table below shows the total number of tornadoes that occurred at some states from 1962 to 1991 and the record high temperatures for the same states.
    Is there a relationship between the number of tornadoes and the record high temperatures?

State	Tornadoes	Record High Temp
AL	668	112
CO	781	118
FL	1590	109
IL	798	117
KS	1198	121
NY	169	108
PA	310	111
TN	360	113
VT	21	105
WI	625	114

4-5 Spearman rank correlation coefficient

	Tornado	R_1	Temp	R_2	$R_1 - R_2$	d^2
	668	6	112	5	1	1
	781	7	118	9	–2	4
	1590	10	109	3	7	49
	798	8	117	8	0	0
	1198	9	121	10	–1	1
	169	2	108	2	0	0
	310	3	111	4	–1	1
	360	4	113	6	–2	4
	21	1	105	1	0	0
	625	5	114	7	–2	4
$\sum$	-	-	-	-	0	64

4-5 Spearman rank correlation coefficient

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{(6)(64)}{(10)(10^2-1)} = 1 - \frac{384}{(10)(100-1)}$$

$$= 1 - \frac{384}{(10)(99)} = 1 - \frac{384}{990} = 1 - 0.39 = 0.61$$

- There is a moderate positive linear relationship between the number of tornados and the record high temperatures.

4-5 Spearman rank correlation coefficient

- Example:
  - To study the relationship between the students grade in Statistics and Mathematics, five students were chosen and their grades are as displayed. Is there a relationship between the grades of Statistics and Mathematics?

Student ID #	Statistics Grades	Mathematics Grades
1200003	F	D
1200004	A	C
1200005	C	B
1200006	D	F
1200007	B	A

4-5 Spearman rank correlation coefficient

	Statistics Grades	R_1	Mathematics Grades	R_2	$R_1 - R_2$	d^2
	F	1	D	2	-1	1
	A	5	C	3	2	4
	C	3	B	4	-1	1
	D	2	F	1	1	1
	B	4	A	5	-1	1
$\sum$	-	-	-	-	0	8

4-5 Spearman rank correlation coefficient

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{(6)(8)}{(5)(5^2-1)} = 1 - \frac{48}{(5)(25-1)}$$

$$= 1 - \frac{48}{(5)(24)} = 1 - \frac{48}{120} = 1 - 0.40 = 0.60$$

There is a moderate positive linear relationship between grades. So, high grades in Mathematics means high grades in Statistics.

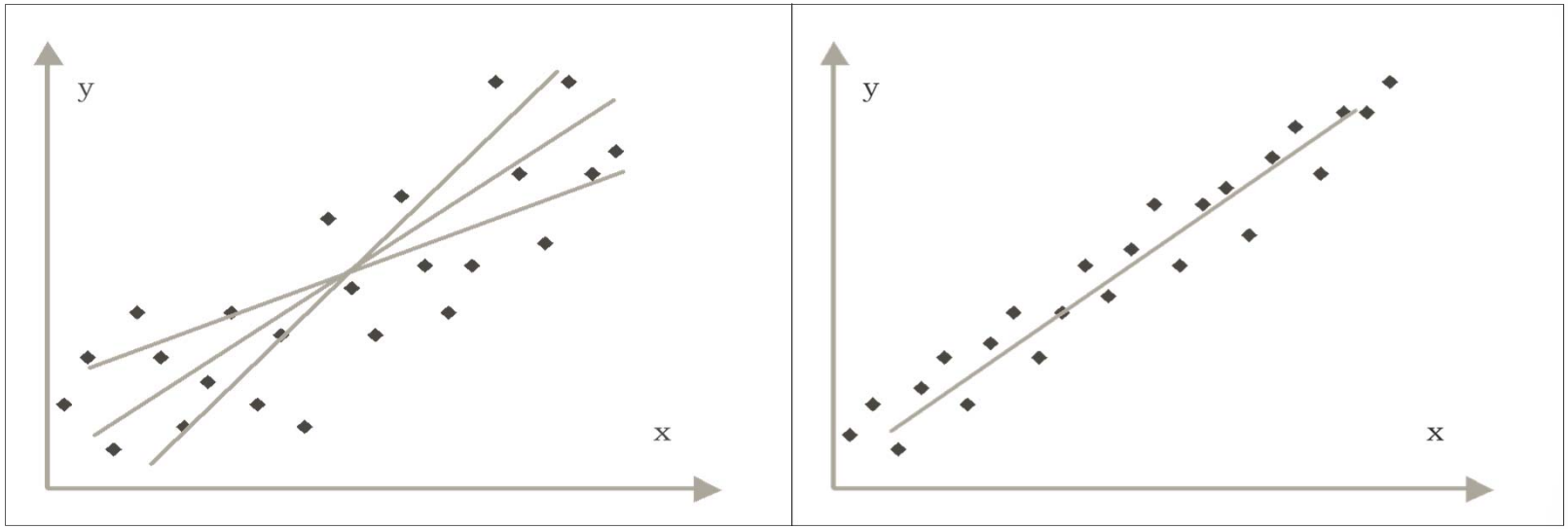
4-6 Regression Line

- If the value of the correlation coefficient is significant (will not be discussed here), the next step is to determine the equation of the **regression line**, which is the data's line of best fit.

- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.

4-6 Regression Line

- Scatter Plot and Regression Line



4-6 Regression Line

- Equation of the Regression Line
  - The equation of the regression line is written as $y'=a+bx$, where b is the slope of the line and a is the y' intercept.
  - The regression line can be used to predict a value for the **dependent** variable (y) for a given value of the **independent** variable (x).
  - **Caution**: Use x values within the experimental region when predicting y values.

4-6 Regression Line

- Formulas for the regression line ***y'=a+bx***:

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

$$a = \frac{\sum y - b\sum x}{n}$$

where ***a*** is the ***y'*** intercept and ***b*** is the slope of the line.

4-6 Regression Line

- Example:
  - Find the equation of the regression line and find ***y'*** when ***x = 35*** years. Remember that no regression should be done when ***r*** is not significant.

Age x	18	26	32	38	52	59
Hours y	10	5	2	3	1.5	1

4-6 Regression Line

							$\sum$
Age x	18	26	32	38	52	59	225
Hours y	10	5	2	3	1.5	1	22.5
x^2	324	676	1024	1444	2704	3481	9653
y^2	100	25	4	9	2.25	1	141.25
xy	180	130	64	114	78	59	625

4-6 Regression Line

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{(6)(625) - (225)(22.5)}{(6)(9653) - (225)^2}$$

$$= \frac{3750 - 5062.5}{57918 - 50625} = \frac{-1312.5}{7293} = -0.18$$

$$a = \frac{(22.5) - (0.18)(225)}{6} = \frac{(22.5) - (-40.5)}{6} = \frac{22.5 + 40.5}{6} = \frac{63}{6}$$

$$= 10.5$$

When ***x=35***, then

$$y' = a + bx = (10.5) + (-0.18)(35) = (10.5) + (-6.3)$$

$$= (10.5) - (6.3) = 4.2 \text{ hours}$$

So, a person who is 35 years old tends to exercise 4.2 hours per weak on average.