DUM 2413 STATISTICS & PROBABILITY

CHAPTER 6
CORRELATION AND SIMPLE LINEAR REGRESSION

PREPARED BY:
DR. CHUAN ZUN LIANG; DR. NORATIKAH ABU; DR. SITI ZANARIAH SATARI
FACULTY OF INDUSTRIAL SCIENCES & TECHNOLOGY
chuanzl@ump.edu.my; atikahabu@ump.edu.my; zanariah@ump.edu.my

EXPECTED OUTCOMES

- Able to identify the linear relationship between a dependent and independent variables visually
- Able to determine the direction and magnitude between a dependent and independent variables
- Able to apply the simple linear regression model for forecasting in application problems
- Able to interpret a correlation coefficient, coefficient of determination and coefficient of regression

CONTENT

6.1 CORRELATION

6.2 COEFFICIENT OF DETERMINATION

6.3 SIMPLE LINEAR REGRESSION

REGRESSION MODEL
DISCUSS IN THIS COURSE

REGRESSION MODEL

LINEAR REGRESSION MODEL

Do not discussed here

SIMPLE LINEAR REGRESSION
*Involved one independent variable
MODEL: $y = \beta_0 + \beta_1 x + \varepsilon$

MULTIPLE LINEAR REGRESSION
*Involved ≥ 2 independent variables
MODEL: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

$$\left.\begin{array}{l}\text{Simple linear regression model} : \beta_0, \beta_1 \\ \text{Multiple linear regression model} : \beta_0, \beta_1, \beta_2, \ldots, \beta_k\end{array}\right\} \text{parameters}$$

6.1
SIMPLE LINEAR REGRESSION ANALYSIS AND CORRELATION

6.2
THE COEFFICIENT OF DETERMINATION

REGRESSION MODELS

- Used an EQUATION to describe the relationship between one dependent variable, y and the independent variable(s), x
  - Independent variable *(Likewise "Manipulated variable" in an experiment)*
    Factor whose effects are studied by the experimenter/researcher
  - Dependent variable *(Likewise "Responding variable" in an experiment)*
    Factor (QUANTITATIVE VARIABLE) whose value varies with the change of independent variable(s)
- Used mainly for prediction or estimation

FOR EXMAPLE

A physician wants to determine the relationship between the hemoglobin concentration and age.

SOLUTION:

Dependent variable : *The hemoglobin concentration*
Independent variable : *Age*

MEASUREMENT OF LINEAR RELATIONSHIP AND EVALUATION OF REGRESSION MODELS

1

- How to determine *the linear relationship* between dependent variable, y and independent variable, x by *graphically*?
- *Solution: Scatter plot*

2

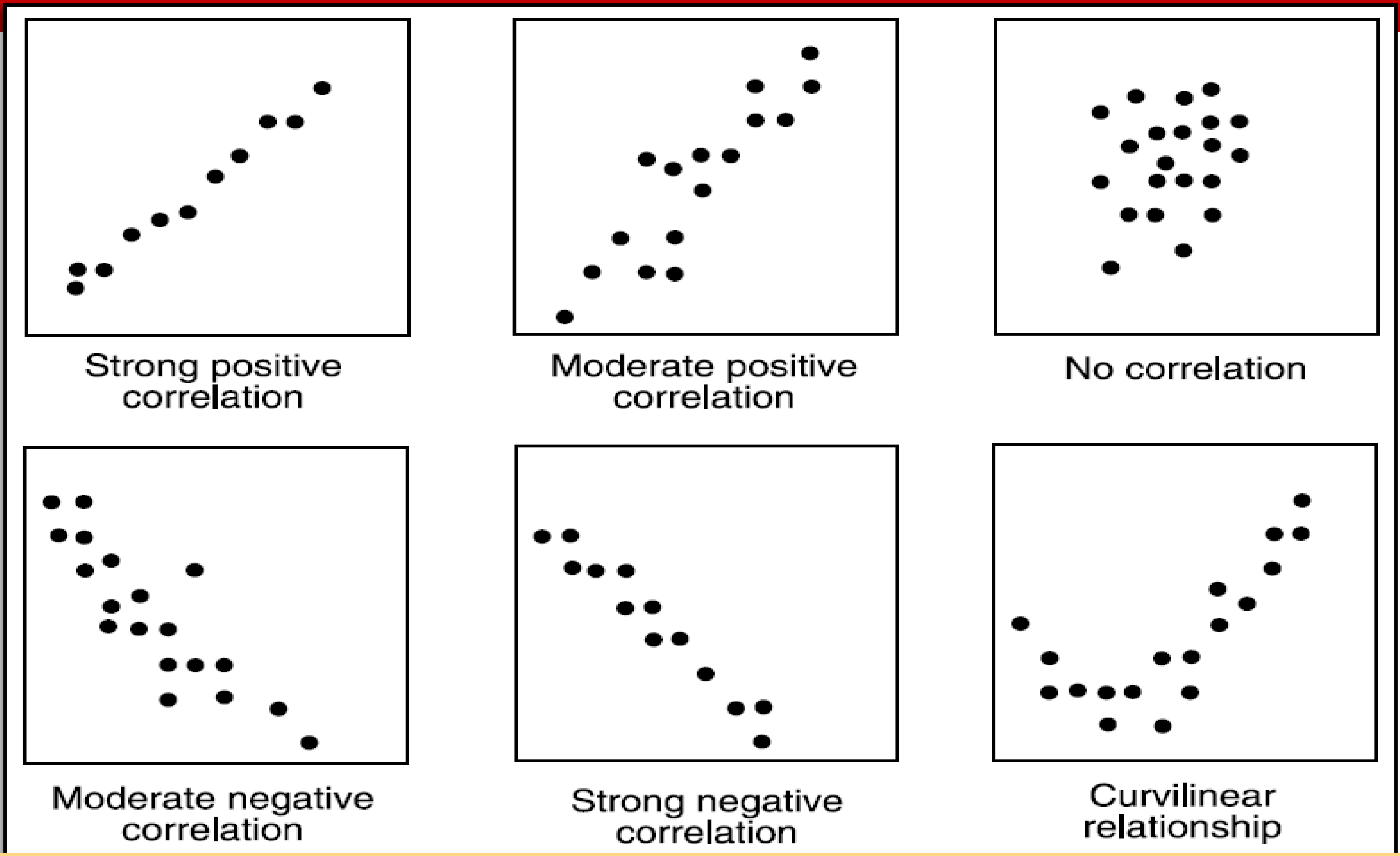
- How to measure the *direction* and the *strength* of a linear relationship (association) between dependent and independent variable(s)?
- *Solution: Correlation Coefficient (Pearson product-moment correlation),* r

3

- How well the sample data fit a statistical model?
- *Solution: Coefficients of determination,* r^2

1. SCATTER PLOT

A graph in which the value of two variables are plotted along two axes, the pattern of resulting points revealing any correlation present.



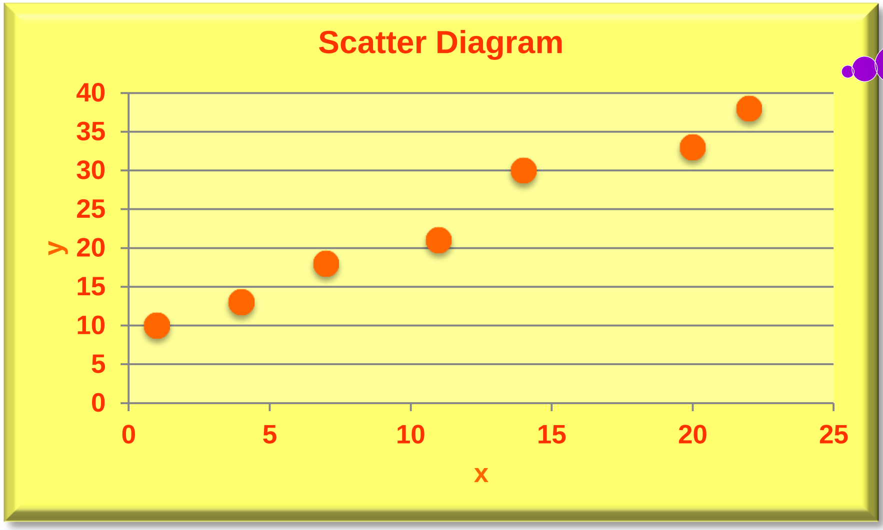
SOURCE: http://m3.zutyct.com/54672d7b5f2f2?geo=MY&aid=249&lp_domain=b1.zexwwx.com&clickid=1416047995

EXAMPLE 6.1

Given the following data collected from a research study. Based on these data, *construct a scatter diagram* and *described the linear relationship* between the variables x and y.

x	1	4	7	11	14	20	22
y	10	13	18	21	30	33	38

SOLUTION



Positive linear relationship

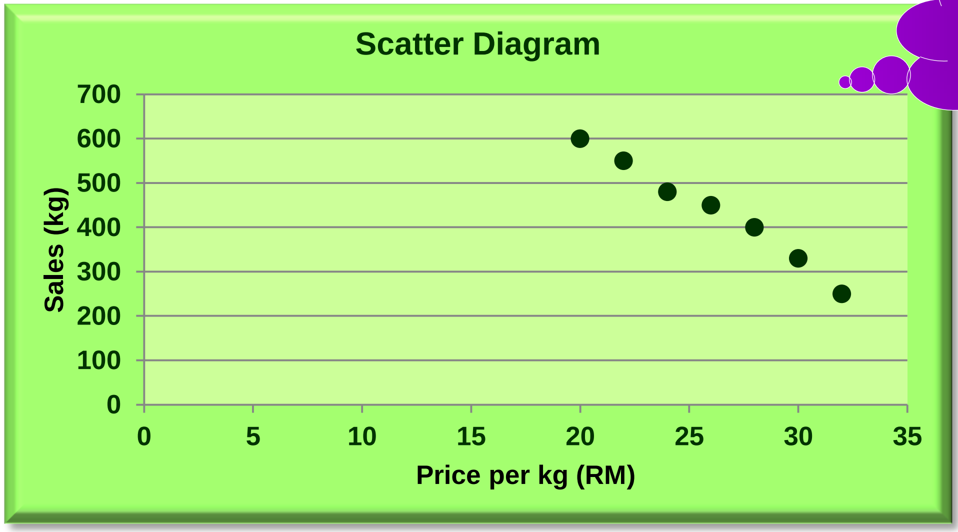
EXERCISE 6.1

Mr. Siew is a fisherman who company supplied prawns to restaurants. The demand for prawns depends on the price per kg. The collected data is illustrated in the table below.

Price per kg (RM)	20	22	24	26	28	30	32
Sales (kg)	600	550	480	450	400	330	250

Based on the information above, *construct a scatter diagram* and *described the linear relationship* between the prices and the demands of prawns.

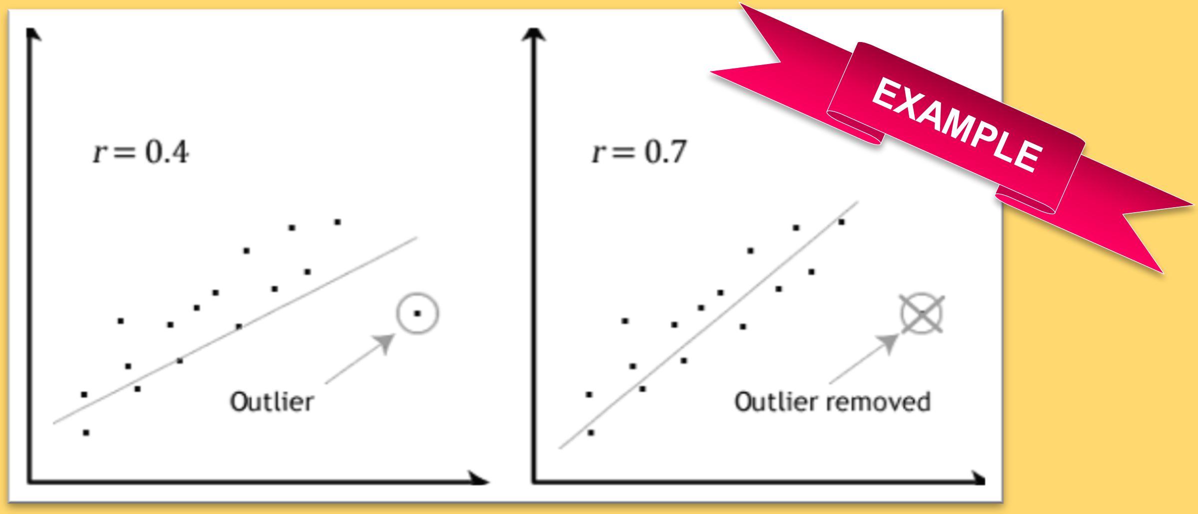
SOLUTION



Negative linear relationship

2. PRODUCT MOMENT CORRELATION

- The *Pearson product-moment correlation coefficient* [NOT SUITABLE FOR ORDINAL VARIABLE] is *a linear measure of the linear correlation between two variables*, x and y, giving a value *between -1 and +1 inclusive*, where *–1 is perfect negatively correlated*, *0 is no correlation*, and *+1 is perfect positively correlated*.
- The correlation coefficient is *not robust* since it is *easily affected by outliers*. Therefore, it should always check the scatter plot with the r value.



SOURCE: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide-2.php

2. PRODUCT MOMENT CORRELATION

The correlation coefficient can be shown to be equal to

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}; \quad -1 \le r \le +1$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}; S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}; S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

****Note: When calculating the correlation coefficient, both variables x and y *do not required same* in *units of measurements*. However, the *sample size, n should be equal.***

2. PRODUCT MOMENT CORRELATION

The Classification of the Strength for Correlation Coefficients

NEGATIVE Direction (-)	POSITIVE Direction (+)
NO CORRELATION between the two variables, x and y $r = 0$	
WEAK negatively correlated $-0.5 < r < 0$	WEAK positively correlated $0 < r < 0.5$
MODERATE negatively correlated $-0.7 < r \le -0.5$	MODERATE positively correlated $0.5 \le r < 0.7$
STRONG negatively correlated $-1.0 < r \le -0.7$	STRONG positively correlated $0.7 \le r < 1.0$
PERFECT negatively correlated $r = -1.0$	PERFECT positively correlated $r = 1.0$

EXAMPLE 6.2

A finance analyst interested *to study the linear relationship* between the *interest rates for housing loans* and *the number of applicants who applied for the loans* during economy grown down season. The collected data in a particular month of his study as illustrated in the table below.

Interest rate in %	6.0	6.2	6.5	6.8	7.0	7.2	7.5	7.8	8.0	8.2	8.4	8.7
Number of applicants	80	80	78	75	70	60	60	55	50	48	45	40

(a) Identify the *independent and dependent variables*.
(b) Find Pearson's product moment *correlation coefficient* and interpret it value.

EXAMPLE 6.2-CONTINUE

SOLUTION

(a) **Dependent variable**, $y =$ **Number of applicants; Independent variable**, $x =$ **Interest rates**

(b)

$$\sum_{i=1}^{12} x_i = 88.3,\ \sum_{i=1}^{12} x_i^2 = 658.35,\ \sum_{i=1}^{12} y_i = 741,\ \sum_{i=1}^{12} y_i^2 = 48063,\ \sum_{i=1}^{12} x_i y_i = 5313.6$$

$$S_{xx} = \sum_{i=1}^{12} x_i^2 - \frac{\left(\sum_{i=1}^{12} x_i\right)^2}{n} = 658.35 - \frac{88.3^2}{12} = 8.6092$$

$$S_{yy} = \sum_{i=1}^{12} y_i^2 - \frac{\left(\sum_{i=1}^{12} y_i\right)^2}{n} = 48063 - \frac{741^2}{12} = 2306.25$$

$$S_{xy} = \sum_{i=1}^{12} x_i y_i - \frac{\left(\sum_{i=1}^{12} x_i\right)\left(\sum_{i=1}^{12} y_i\right)}{n} = 5313.6 - \frac{(88.3)(741)}{12} = -138.925$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-138.925}{\sqrt{(8.6092)(2306.25)}} = -0.9859$$

Interpretation: Since $r = -0.9859$, **there is *strong negative linear relationship* between the interest rates and the number of applicants for the loans.**

EXERCISE 6.2

A real estate agent believes *that the monthly rent houses depend on the size of the houses*. A sample of eight houses in a residential area was selected and the information gathered is shown in the table below.

Monthly rent (RM'00)	Size ('00 square feet)
12	10
16	14
9	8
15	12
9	7
17	14
9	7
16	11

(a) Identify the *independent and dependent variables*.
(b) Find *product-moment correlation coefficient* and interpret it value.

EXERCISE 6.2-CONTINUE

(a) **Dependent variable,** $y =$ **Monthly rental; Independent variable,** $x =$ **Size of the house**

(b) $\sum_{i=1}^{8} x_i = 83, \ \sum_{i=1}^{8} x_i^2 = 919, \ \sum_{i=1}^{8} y_i = 103, \ \sum_{i=1}^{8} y_i^2 = 1413, \ \sum_{i=1}^{8} x_i y_i = 1136$

$$S_{xx} = \sum_{i=1}^{8} x_i^2 - \frac{\left(\sum_{i=1}^{8} x_i\right)^2}{n} = 919 - \frac{83^2}{8} = 57.875$$

$$S_{yy} = \sum_{i=1}^{8} y_i^2 - \frac{\left(\sum_{i=1}^{8} y_i\right)^2}{n} = 1413 - \frac{103^2}{8} = 86.875$$

$$S_{xy} = \sum_{i=1}^{8} x_i y_i - \frac{\left(\sum_{i=1}^{8} x_i\right)\left(\sum_{i=1}^{8} y_i\right)}{n} = 1136 - \frac{(83)(103)}{8} = 67.375$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{67.375}{\sqrt{(57.875)(86.875)}} = 0.9502$$

Interpretation: Since $r = 0.9502$, **there is *strong positive linear relationship* between the size of house and the monthly rental.**

3. THE COEFFICIENT OF DETERMINATION

The coefficient of determination, $r^2;\ 0 \le r^2 \le 1$ is a measure of the usefulness of linear regression model.

- **In statistics, r^2 is denoted as *the proportion* of the *total variation* in the n observed values of the *dependent variable* that is *explained/predictable* by the independent *variable(s)*.**
  - **The nearer $r^2 = 1$, *the larger is the utility of the model* in *predicting* y**
- **The coefficient of determination is given as r^2, where r is correlation coefficient**
- **The value of r^2 can be interpret as**

  i. When $r^2 = 0$,
  - ***The dependent variable cannot be explained/predicted from independent variable(s).***

  ii. When $0 < r^2 < 1$,
  - **$r^2 (100\%)$ *of the total variation in y can be explained/predicted by independent variable(s).***

  iii. When $r^2 = 1$,
  - ***The dependent variable can be predicted without error from the independent variable(s).***

EXAMPLE 6.3

Dr. Bazli wants to investigate ***the length of time taken to revise statistics lessons affect the final examination scores***. He randomly selected eight students from his class and collects the following data as illustrated in the table below.

Students	A	B	C	D	E	F	G	H
Time (Hours)	19	12	34	42	9	18	51	26
Examination score	55	47	70	98	37	71	96	85

(a) Identify the ***independent and dependent variables*** involved Dr. Bazli's study.
(b) Find the ***correlation coefficient*** and interpret its value.
(c) Find the ***coefficient of determination*** and interpret it value.

EXAMPLE 6.3-CONTINUE

(a) **Dependent variable,** $y =$ **Final examination scores;**
Independent variable, $x =$ **Time taken to revise statistics lessons before final examination**

(b) $\sum_{i=1}^{8} x_i = 211,\ \sum_{i=1}^{8} x_i^2 = 7107,\ \sum_{i=1}^{8} y_i = 559,\ \sum_{i=1}^{8} y_i^2 = 42589,\ \sum_{i=1}^{8} x_i y_i = 16822$

$$S_{xx} = \sum_{i=1}^{8} x_i^2 - \frac{\left(\sum_{i=1}^{8} x_i\right)^2}{n} = 7107 - \frac{211^2}{8} = 1541.8750;\ S_{yy} = \sum_{i=1}^{8} y_i^2 - \frac{\left(\sum_{i=1}^{8} y_i\right)^2}{n} = 42589 - \frac{559^2}{8} = 3528.8750$$

$$S_{xy} = \sum_{i=1}^{8} x_i y_i - \frac{\left(\sum_{i=1}^{8} x_i\right)\left(\sum_{i=1}^{8} y_i\right)}{n} = 16822 - \frac{(211)(559)}{8} = 2078.3750$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{2078.3750}{\sqrt{(1541.8750)(3528.8750)}} = 0.8910$$

Interpretation: Since $r = 0.8910,$ **there is *strong positive linear relationship* between the time taken to revise statistics lessons and examination marks.**

(c) Coefficient of determination, $r^2 = 0.8910^2 = 0.7939$

Interpretation: *79.39% of total variation* in examination marks *can be explained by* the time taken to revise statistics lessons before examination.

EXERCISE 6.3

The manager of Khairul trading company claimed that there is a relationship between *the amount of mileage claims made by salesman* and *their monthly sales*. In order to consolidate his claim's, he collects data on the amount of sales and the mileage claims made by 7 randomly selected salesmen.

Salesman	A	B	C	D	E	F	G
Mileage claims (RM'00)	8	5	8	11	9	12	7
Sales (RM'000)	12	10	14	16	15	19	11

(a) Identify the *independent and dependent variables*.
(b) Find the *Pearson's product moment correlation coefficient*.
(c) Find the *coefficient of determination* and interpret it value.

EXERCISE 6.3-CONTINUE

(a) **Dependent variable,** $y =$ **The monthly sales;**
Independent variable, $x =$ **The amount of mileage claims made by salesmen**

(b) $\sum_{i=1}^{7} x_i = 60,\ \sum_{i=1}^{7} x_i^2 = 548,\ \sum_{i=1}^{7} y_i = 97,\ \sum_{i=1}^{7} y_i^2 = 1403,\ \sum_{i=1}^{7} x_i y_i = 874$

$$S_{xx} = \sum_{i=1}^{7} x_i^2 - \frac{\left(\sum_{i=1}^{7} x_i\right)^2}{n} = 548 - \frac{60^2}{7} = 33.7143;\ S_{yy} = \sum_{i=1}^{7} y_i^2 - \frac{\left(\sum_{i=1}^{7} y_i\right)^2}{n} = 1403 - \frac{97^2}{7} = 58.8571$$

$$S_{xy} = \sum_{i=1}^{7} x_i y_i - \frac{\left(\sum_{i=1}^{7} x_i\right)\left(\sum_{i=1}^{7} y_i\right)}{n} = 874 - \frac{(60)(97)}{7} = 42.5714$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{42.5714}{\sqrt{(33.7143)(58.8571)}} = 0.9557$$

Interpretation: Since $r = 0.9557,$ **there is *strong positive linear relationship* between the amount of mileage claims made by salesmen and the monthly sales.**

(c) Coefficient of determination, $r^2 = 0.9557^2 = 0.9134$

Interpretation: *91.34% of the total variation* in the monthly sales *can be explained by* the amount of mileage claims made by salesmen.

EXERCISE 6.4

In University Malaysia Pahang, the final grade obtained by students is composed of 60% carry marks and 40% final exam scores. A study is conducted to investigate the relationship between ***the carry marks*** and ***final exam scores***. A sample of 10 students from the Faculty of Engineering Technology are selected and the data are illustrated in the table below.

Student	1	2	3	4	5	6	7	8	9	10
Carry Mark	40	31	27	42	36	53	53	32	49	38
Final Exam	24	25	21	36	15	34	38	14	29	23

(a) Identify the ***independent and dependent variables***.

(b) Draw a ***scatter diagram*** and give comment.

(c) Calculate $\sum y_i$, $\sum y_i^2$, $\sum x_i$, $\sum x_i^2$ and $\sum x_i y_i$.

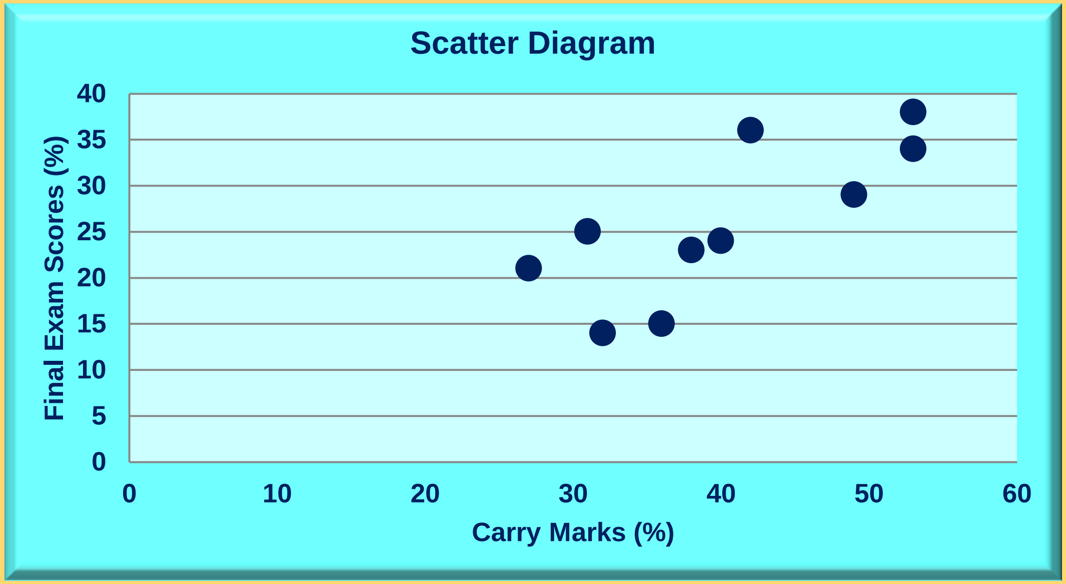
(d) Calculate ***correlation coefficient*** (r) and interpret its value.

(e) Give your general comment of the students' performance.

EXERCISE 6.4-CONTINUE

(a) Dependent variable, $y =$ Final Exam Scores; Independent variable, $x =$ Carry Marks

(b)



The scatter diagram above depicts a *positive linear relationship* between the carry marks and final exam scores.

EXERCISE 6.4-CONTINUE

(c) $\sum y_i = 259,\ \sum y_i^2 = 7329,\ \sum x_i = 401,\ \sum x_i^2 = 16837,\ \sum x_i y_i = 10913$

(d)

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 16837 - \frac{401^2}{10} = 756.9;\ S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 7329 - \frac{259^2}{10} = 620.9$$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 10913 - \frac{(401)(259)}{10} = 527.1$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{527.1}{\sqrt{(756.9)(620.9)}} = 0.7689$$

Interpretation: There is *strong positive linear relationship* between the carry marks and final exam scores.

(e) **Based on the table, we can observed that some of the students who obtained high carry marks, but did not performed very well in final exam.**

6.3
SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION

- ***The Simple Linear Regression Equation (Population)*** **is useful to describe the linear relationship between the mean of the responses, y, and independent variable, x, that can be expressed as**

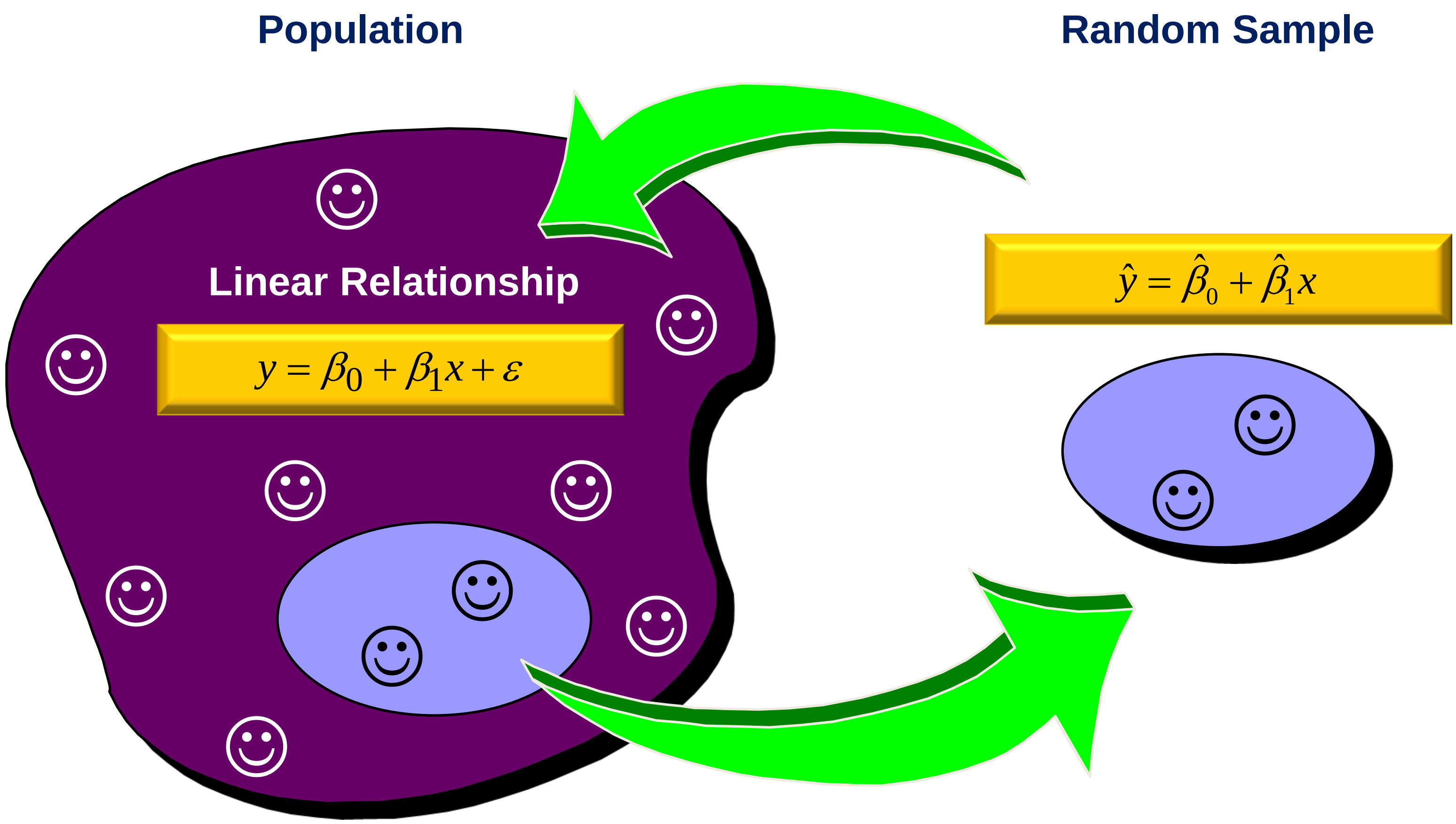
Population Y-Intercept | **Population Slope** | **Random Error**

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

Dependent variable | **Independent variable**

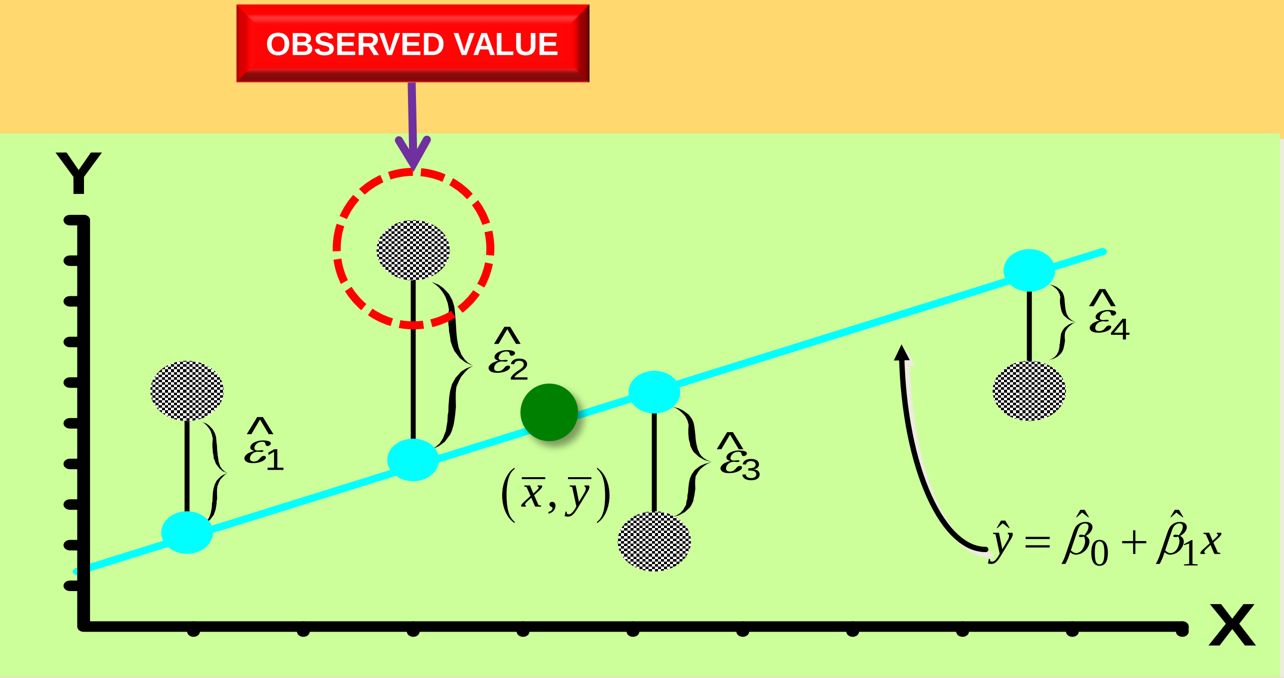
where parameters β_0 and β_1 are unknown

SIMPLE LINEAR REGRESSION



SIMPLE LINEAR REGRESSION

The parameters, β_0 and β_1, can be estimate by minimizing the sum of squared residuals, $\sum \varepsilon^2 = \sum (y - \hat{y})^2$ using least square estimation method, Therefore, the **best-fitting line** is obtained will be always pass through $(\bar{x}, \bar{y})$.



SOURCE: https://www.slideshare.net/vermaumeshverma/linear-regression-38653351

PARAMETERS ESTIMATION AND ITS INTERPRETATION

Point Estimate for *Y-Intercept*, $\hat{\beta}_0$

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

where

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{y} = \frac{\sum y}{n}$$

INTERPRETATION (EXAMPLE):

If $\hat{\beta}_0 = 4$, then the average y is expected to be 4 when $x = 0$.

Point Estimate for Slope, $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

INTERPRETATION (EXAMPLE):

If $\hat{\beta}_1 = 2$, then the average y is expected to increase by 2 for every 1 unit increase in x.

EXAMPLE 6.4

A study is conducted to investigate the relationship between *blood pressure rise* and *sound pressure level*. The data of the study is given in the table below.

Blood pressure rise (mmHg)	1	0	1	2	5	4	6	2
Sound pressure level (dB)	60	63	65	70	70	80	90	80

(a) Identify the *independent and dependent variables*.
(b) Calculate the value of *correlation coefficient* and interpret its value.
(c) Estimate the *regression coefficient* and hence write the equation of the estimated regression line.
(d) Find the *predicted mean rise in blood pressure level* associated with a sound pressure level of 100 decibels.

EXAMPLE 6.4-CONTINUE

(a) **Dependent variable, y =Blood pressure rise**;
Independent variable, x =Sound pressure level

(b)

$$\sum x = 578, \sum x^2 = 42494, \sum y = 21, \sum y^2 = 87, \sum xy = 1635, n = 8$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 42494 - \frac{578^2}{8} = 733.5000$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 87 - \frac{21^2}{8} = 31.8750$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 1635 - \frac{(578)(21)}{8} = 117.7500$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{117.7500}{\sqrt{(733.5000)(31.8750)}} = 0.7701$$

Interpretation: There is *strong positive linear relationship* between sound pressure level and blood pressure rise

EXAMPLE 6.4-CONTINUE

(c) $\bar{x} = 72.2500,\ \bar{y} = 2.6250$

Slope, $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{117.7500}{733.5000} = 0.1605$

Intercept, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 2.6250 - 0.1605(72.2500) = -8.9711$

Simple Linear Regression Equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{y} = -8.9711 + 0.1605x$$

(d) **The predicted mean rise in blood pressure level associated with a sound pressure level of 100 decibels:** $\bar{y} = 2.6250$

EXERCISE 6.5

A study is conducted to investigate the relationship between *the cost (RM million) of fire damage* and *distance (km) between the fire station and the location involves in the fire accident*. The regression method is used to analyse the data in the table below.

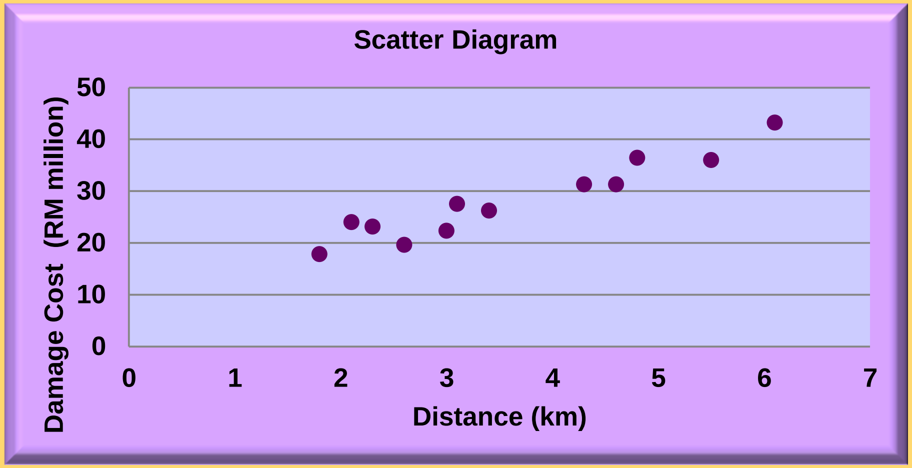
Cost	26.2	17.8	31.3	23.1	27.5	36.0	22.3	19.6	31.3	24.0	43.2	36.4
Distance	3.4	1.8	4.6	2.3	3.1	5.5	3.0	2.6	4.3	2.1	6.1	4.8

(a) Identify the *independent and dependent variables*.
(b) Draw a scatter plot and give comment.
(c) Calculate $\sum y_i,\ \sum y_i^2,\ \sum x_i,\ \sum x_i^2$ and $\sum x_i y_i$.
(d) Calculate *correlation coefficient* (r) and interpret its value.
(e) Calculate *coefficient of determination* (r^2) and interpret its value.
(f) Estimate the *regression parameters* and write the estimated *linear regression model*.
(g) Give the *interpretation* of *regression coefficient* $(\hat{\beta}_1)$.
(h) Predict the cost when the distance is 10km.
(i) Predict the distance when the cost is RM10 million.
(j) What is the *mean cost* when the distance is 20km?

EXERCISE 6.5-CONTINUE

(a) Dependent variable, $y =$ Cost of fire damage;
Independent variable, $x =$ Distance between the fire station and the location involves in the fire accident

(b)



The scatter diagram above illustrated a *positive linear relationship* between the distance and damage cost.

EXERCISE 6.5-CONTINUE

(c) $\sum y_i = 338.70, \sum y_i^2 = 10197.17, \sum x_i = 43.60, \sum x_i^2 = 180.02, \sum x_i y_i = 1342.57$

(d) $$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 180.02 - \frac{43.60^2}{12} = 21.6067$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 10197.17 - \frac{338.70^2}{12} = 637.3625$$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 1342.57 - \frac{(43.60)(338.70)}{12} = 111.96$$

$$\text{Correlation Coefficient, } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{111.96}{\sqrt{(21.6067)(637.3625)}} = 0.9541$$

Interpretation: There is *strong positive linear relationship* between the distance and damage cost.

EXERCISE 6.5-CONTINUE

(e) Coefficient of determination, $r^2 = 0.9541^2 = 0.9103$

Interpretation: ***91.03% total variation*** **in damage cost** ***can be explained by*** **the distance.**

(f) **Slope,** $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{111.9600}{21.6067} = 5.1817;$

Intercept, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 28.2250 - 5.1817(3.6333) = 9.3983$

The Estimated Linear Regression Model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{y} = 9.3983 + 5.1817x$$

(g) $\hat{\beta}_1 = 5.1817$

Interpretation: **The damage cost is expected** ***increase*** **by RM5.1817 million for** ***every 1km increase*** **in distance.**

EXERCISE 6.5-CONTINUE

(h) $\hat{y} = 9.3983 + 5.1817x$;

When $x = 10$ km,

$\hat{y} = 9.3983 + 5.1817(10) = \text{RM}\,61.2153$ millions

(i) $\hat{y} = 9.3983 + 5.1817x$;

When $\hat{y} = \text{RM}\,10$ million,

$$x = \frac{10 - 9.3983}{5.1817} = 0.1161\text{ km}$$

(j) **Mean cost,** $\bar{y} = 28.2250$

THANK YOU

END OF CHAPTER 6