UMP OPEN COURSEWARE

Universiti Malaysia PAHANG
Engineering • Technology • Creativity

CHAPTER 5

LINEAR REGRESSION AND CORRELATION

Expected Outcomes

- ✓ Able to use simple and multiple linear regression analysis, and correlation.
- ✓ Able to conduct hypothesis testing for simple and multiple linear regression models.
- ✓ Able to identify the best fit model.

PREPARED BY: DR SITI ZANARIAH SATARI & DR ROSLINAZAIRIMAH ZAKARIA

CONTENT

5.1: CORRELATION

Introductory Concepts

- ✓ Suppose you wish to investigate the relationship between a dependent variable (*y*) and independent variable (*x*)

  - **Independent variable** (*x*) **–** explanatory/ predictor/regressor/exogeneous/controlled **variable.**
  - **Dependent variable** (*y*) **–** the response/endogeneous variable.

- ✓ In other word, the value of *y* depends on the value of *x*.

Example 1

Suppose you wish to investigate the relationship between the numbers of hours students spent studying for an examination and the mark they achieved.

Students	A	B	C	D	E	F	G	H
numbers of hours (*x*)	5	8	9	10	10	12	13	15
Final marks (*y*)	49	60	55	72	65	80	82	85

Numbers of hours students spent studying for an examination
(*x* – Independent variable)

will effect →

the mark (*y*) they achieved.
(*y* – Dependent variable)

- ✓ Students will obtain a better final marks if they spent more time to study.
- ✓ Therefore, the final marks depend on the number of hours students spent studying for an examination.

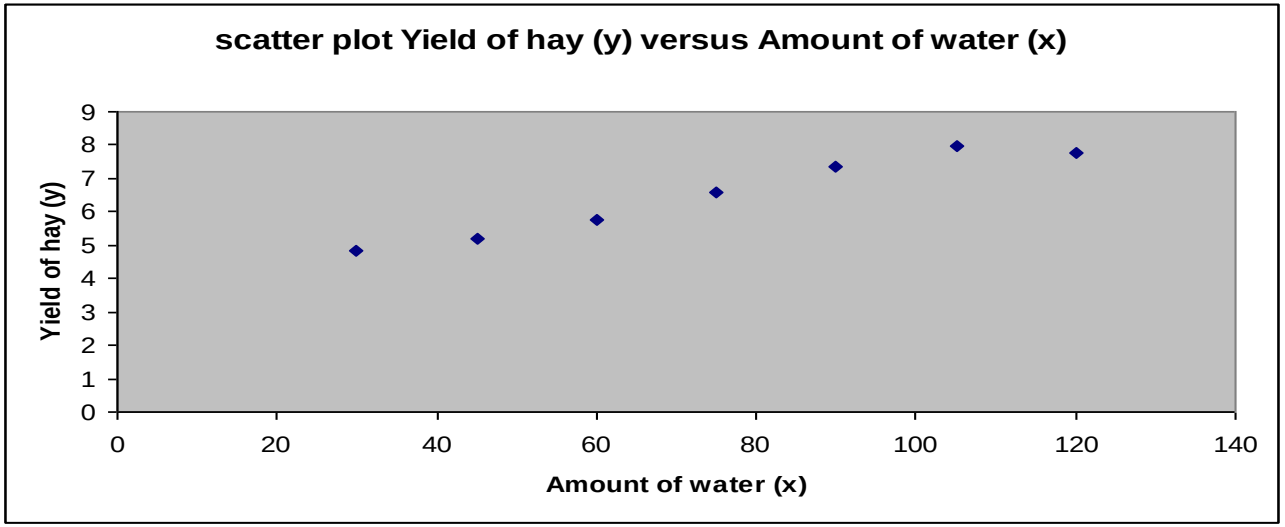
Other Examples

1. The weight at the end of the spring (x) and the length of the spring (y)
2. Temperature (x) and pressure (y)
3. Child's age (x) and height(y)
4. A student's mark in a computer test (x) and the mark in a mathematics test (y)
5. The diameter of the stem of a plant (x) and the average length of the leaf of the plant (y)

Scatter Plot

When pairs of values are plotted (*x* vs *y*), a scatter plot is produced

- ✓ To see how the data looks like and relate with each other
- ✓ To investigate the relationship (association) between x and y

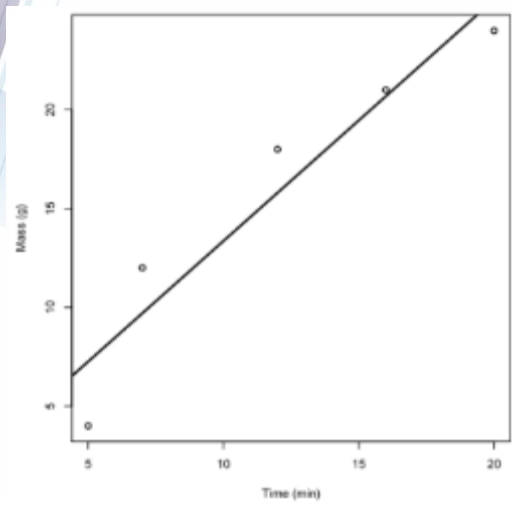


Linear Relationship/Correlation

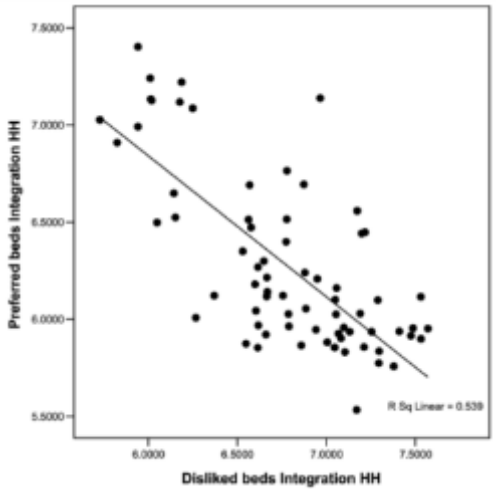
- ✓ We can observed some pattern/trend/relationship from the scatter plot
- ✓ **Linear relationship/correlation**
  - If the points on the scatter plot appear to lie near a straight line **(Simple regression line/line of best fit/trend line)**
- ✓ Or you would say that there is a **linear correlation/linear relationship** between *x* and *y*
- ✓ Sometimes a scatter plot shows a curvilinear (polynomial curve) or nonlinear (log, exp, etc…) relationship between data.

Exercise: Plot a scatter diagram for Example 1. Is there any correlation between *x* and *y*?

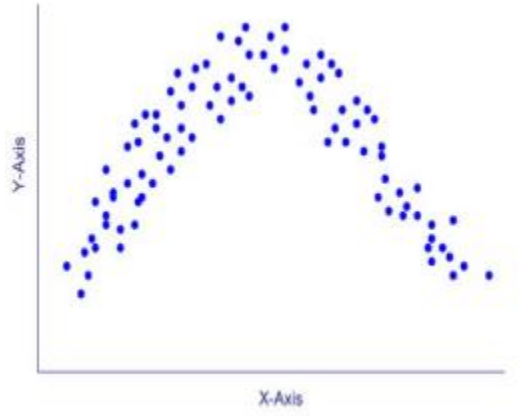
Examples of scatter plot



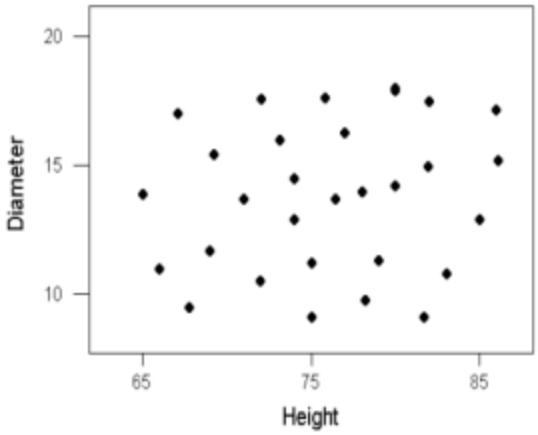
a) Positive linear trend



b) Negative linear trend



c) Curvilinear trend



d) No trend

Correlation Coefficient

- The Pearson product-moment correlation coefficient (Karl Pearson 1900), *r*, is a numerical value between -1 and 1 inclusive.
- Used to indicate the **linear degree** of scatter plot.
- The value of correlation coefficient shows the **strength of the association** or the **linear direction** between the variables.
- It is independent of the units of scale of the variables (dimensionless).
- The correlation coefficient is not robust since it is **easily affected by outliers**.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$-1 \le r \le 1$$

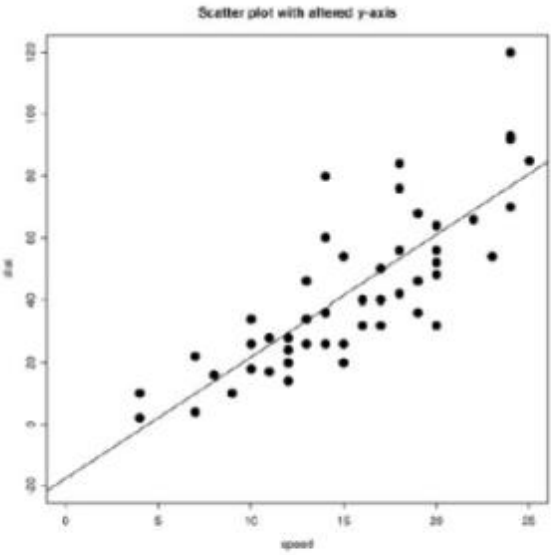
$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

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

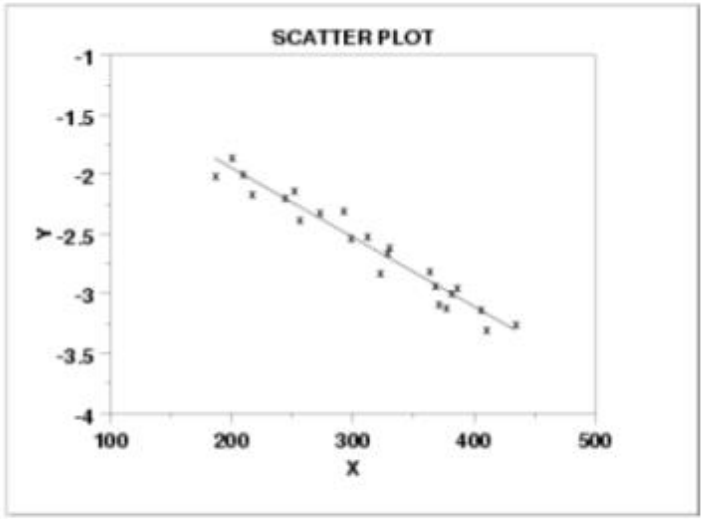
$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Inferences in Correlation Coefficient

- ✓ The nearer the value of r to 1 or -1, the closer the points on the scatter plot are to regression line.
  - Nearer to 1 is strong positive linear correlation/relationship
  - Nearer to -1 is strong negative linear correlation/relationship



(a) Weak positive correlation ($r = 0.5$)



(b) Strong negative correlation ($r = -0.95$)

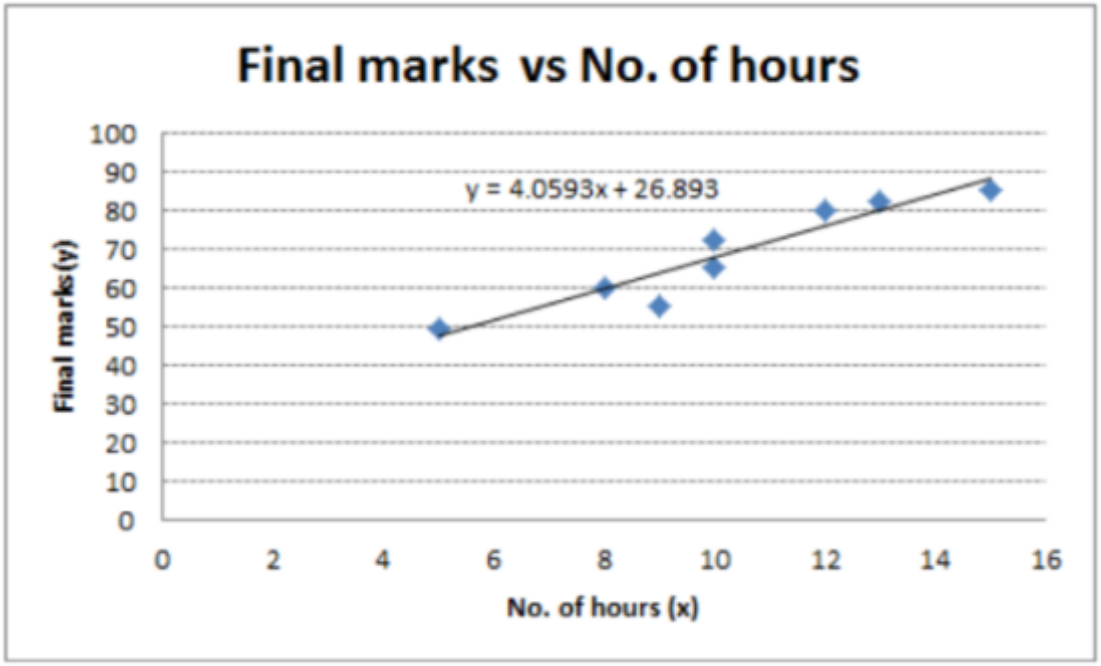
Scatter plots of positive and negative correlations

Strength of correlation coefficient

Scale	Correlation sign
$r = -1.0$	perfect negatively correlated where all the data fall on the line of negative slope
$-1 < r \leq -0.7$	Strong negative
$-0.7 < r \leq -0.5$	Moderate negative
$-0.5 < r < 0$	Weak negative
0	No correlation
$0 < r < 0.5$	Weak positive
$0.5 \leq r < 0.7$	Moderate positive
$0.7 \leq r < 1$	Strong positive
$r = 1.0$	perfect positively correlated where all the data fall on the line of positive slope

Example 2

Using data from Example 1, draw a scatter plot and calculate the correlation coefficient value and interpret its value.



Scatter plot of Final Marks and No. of Hours

No. i	Hours x_i	Marks y_i	Calculation		
			x_i^2	$x_i y_i$	y_i^2
1	5	49	25	245	2401
2	8	60	64	480	3600
3	9	55	81	495	3025
4	10	72	100	720	5184
5	10	65	100	650	4225
6	12	80	144	960	6400
7	13	82	169	1066	6724
8	15	85	225	1275	7225
$n=8$	82	548	908	5891	38784

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{n} = 5891 - \frac{82(548)}{8} = 274$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 908 - \frac{82^2}{8} = 67.5$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 38784 - \frac{548^2}{8} = 1246$$

The correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{274}{\sqrt{67.5}\sqrt{1246}} = 0.9448$$

Comment:

r = 0.9448 shows a strong positive linear correlation. There is a strong positive linear relationship between the number of hours students spent to study and their final marks. Those who spent more time will obtain better result.

Using Calculator (model: fx–570MS)

CLEAR all the previous data (memory): $\boxed{shift}$ $\boxed{mode}$ All $\boxed{3}$

Step 1 – Set the calculator to "regression" and "linear" modes

$\boxed{mode}$ $\boxed{mode}$ Reg $\boxed{2}$ Lin $\boxed{1}$

Step 2 – Key-in the pairs of data x $\boxed{,}$ y $\boxed{,}$

5 $\boxed{,}$ 49 $\boxed{M+}$

8 $\boxed{,}$ 60 $\boxed{M+}$

$\vdots$

15 $\boxed{,}$ 85 $\boxed{M+}$

Step 3 – Find the values of $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$ using

$\boxed{shift}$ S-Sum $\boxed{1}$ then use the arrow keys.

Step 4 – We may check the value of r, $\overline{x}$ and $\overline{y}$ using

$\boxed{shift}$ S-Var $\boxed{2}$ then use the arrow keys.

5.2 THE COEFFICIENT OF DETERMINATION

- Denoted by r^2
- Interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- The percentage of the variation in dependent variable that can be explained by the independent variable.
- The value of r^2 is between 0 and 1 inclusive ($0 \leq r^2 \leq 1$).

Interpretation of coefficient of determination value, R^2

Value of R^2	Interpretation
$r^2 = 0$	the dependent variable cannot be predicted from the independent variable
$r^2 = 1$	the dependent variable can be predicted without error from the independent variable
$(0 < r^2 < 1)$	indicate the extent to which the dependent variable is predicted from the independent variable.
$r^2 = 0.1$	10 % of the variation in *y* can be explained by *x* or 10% of the variation in *y* is predictable from *x*
$r^2 = 0.8$	80 % of the variation in *y* can be explained by *x* or 80% of the variation in *y* is predictable from *x*

Example 3

Using the results from Example 2, find the coefficient of determination and interpret its value.

Solution

The value of coefficient of determination:

$r^2 = (0.94482)^2 = 0.892647$

Interpretation:

89.26 % of the variation in the final marks (y) can be explained by the number of hours students spent to study (x). 10.7% of the variation in y is due to other factors such as study environment.

5.3 SIMPLE LINEAR REGRESSION

- ✓ Regression method describes how one variable depends on other variables.
- ✓ Simple linear regression model is a model with a single independent variable *x* that has a relationship with a response variable *y* and it can be represented by **an equation of a straight line** (line of best fit).
- ✓ If we have more than one independent variables then it becomes multiple linear regression.

Simple Linear Regression Model

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

Where

β_0 = unknown constant intercept (regression coefficient)
β_1 = unknown constant slope (regression coefficient)
ε = random error component
(zero mean and unknown variance)

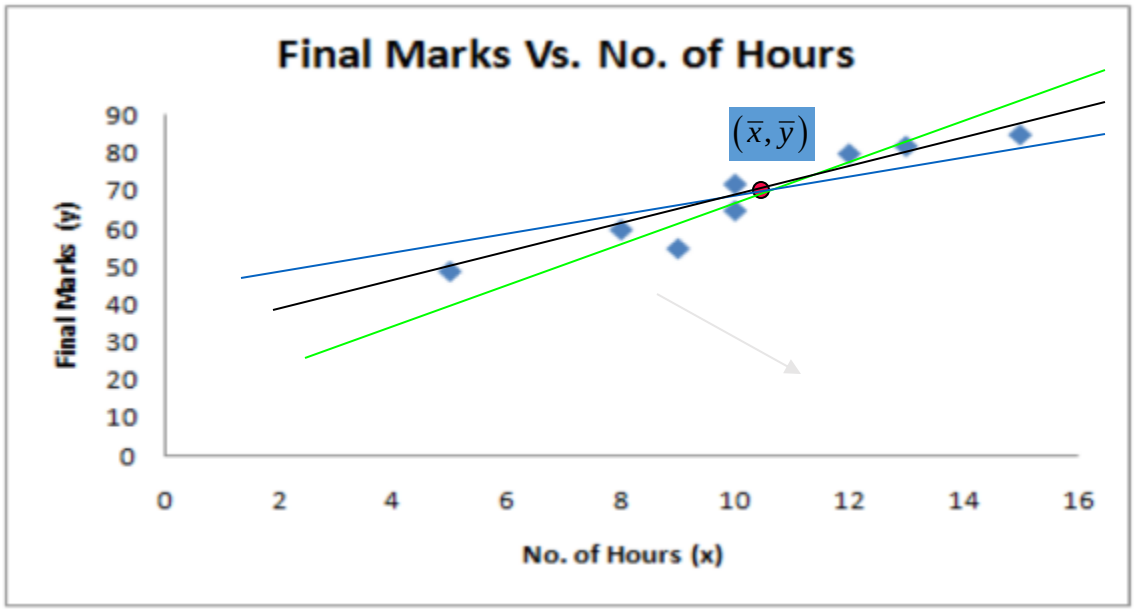
Line of Best Fit

- ✓ A mathematical way of fitting the regression line (least of square method).
- ✓ The line of best fit must pass through the means of both sets of data, i.e. the point $(\bar{x}, \bar{y})$.

Example 1

$\bar{x} = 10.25$

$\bar{y} = 68.5$



Properties of Simple Linear Regression Line

- A line that best represents the data on a scatter plot that passes through $(\bar{x}, \bar{y})$.
- A line that minimizes the sum of squared differences between observed values (y) and predicted values $(\hat{y})$.
- Considered a good approximation of the true relationship between two variables.
- A linear line with regression constant $(\hat{\beta}_0)$ or the y-intercept.
- A linear line with regression coefficient $(\hat{\beta}_1)$ or the slope of the regression line. It describes the average change in the dependent variable (y) for each unit change in the independent variable (x).
- Used to estimate value of a variable given a value of the other (interpolation/extrapolation).

Guideline for Using Regression Equation

1. If there is no linear correlation, do not use the regression equation to make prediction.

2. When using the regression equation for predictions, stay within the scope of the available sample data. Be careful with extrapolation method **(estimating beyond the original observation range)**.

3. A regression equation based on old data is not necessarily suitable for current data.

4. Do not make predictions about a population that is different from the population from which the sample data were drawn.

Estimation of Model Parameters

- ✓ The least squares method is used to estimate the parameters β_0 and β_1.
- ✓ The estimated (fitted) equation of the simple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where

$\hat{y}$: estimated dependent (or response) variable

x : independent (or predictor/ regressor /explanatory) variable

$\hat{\beta}_0$: estimate of *y*–intercept, the point at which the line intersects the y-axis (regression constant)

$\hat{\beta}_1$: estimate of slope, the amount of increase/decrease of *y* for each unit increase (or decrease) in *x* (regression coefficient).

Least square regression line of *y* on *x*

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

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

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

Where:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \quad , \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{n} \quad , \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Example 4

Using data from Example 1,

(a) calculate the least square estimates of the slope and *y*-intercept of the linear regression line, then write its equation.

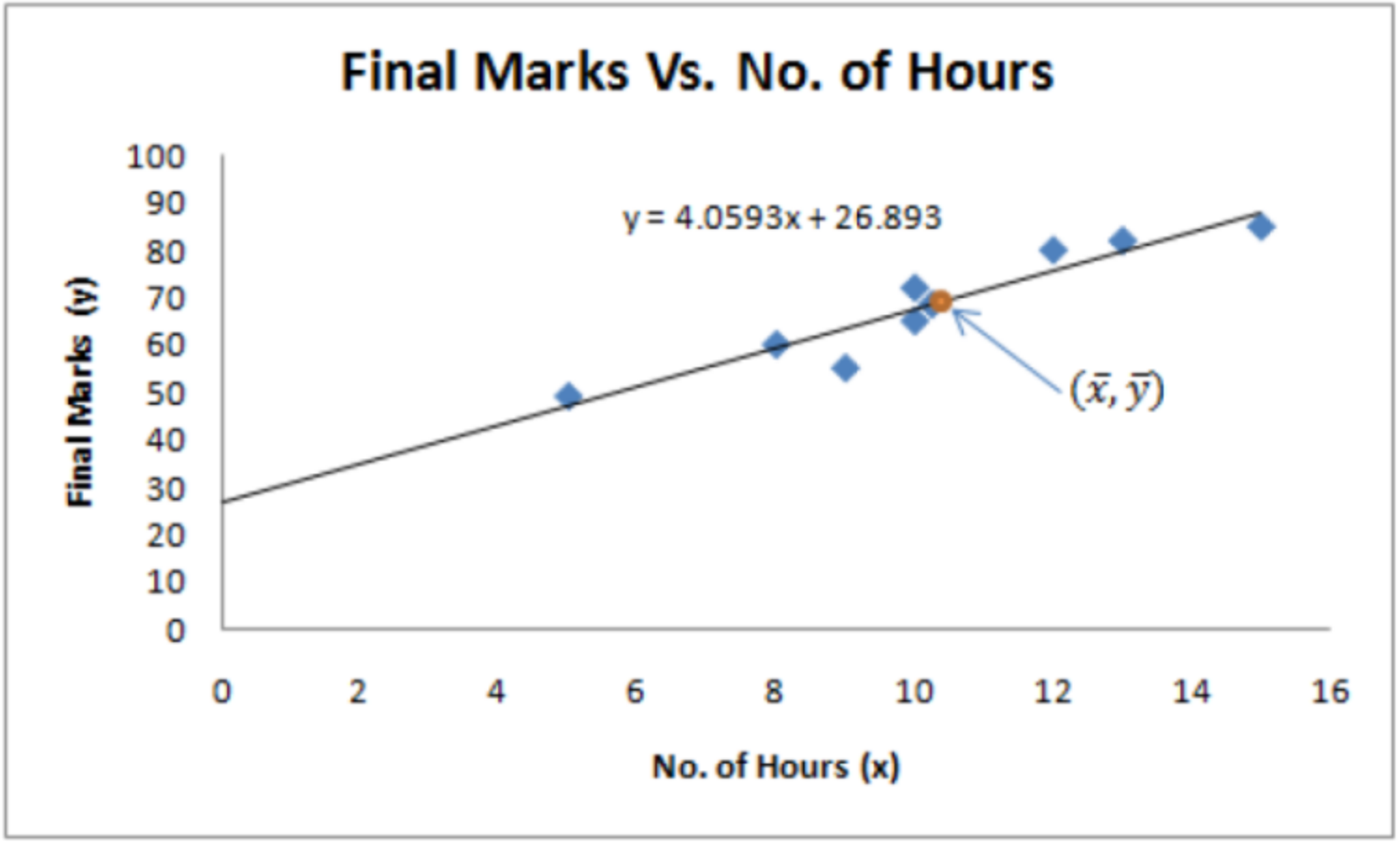
$S_{xy} = 274 \qquad S_{xx} = 67.5 \qquad S_{yy} = 1246 \qquad \bar{x} = 10.25 \qquad \bar{y} = 68.5 \qquad n = 8$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{274}{67.5} = 4.0593$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 68.5 - (4.0593)(10.25) = 26.8922$$

$$\Rightarrow \quad \hat{y} = 4.0593 + 26.8922x$$

Example 4


Final Marks Vs. No. of Hours
y = 4.0593x + 26.893
$(\bar{x}, \bar{y})$
Final Marks (y)
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
No. of Hours (x)

Example 4

(b) interpret the values of intercept and slope.

$\hat{\beta}_0$: If the hour spent to study is zero (student do not spent any hour for study), student will achieve 26.89 marks.

$\hat{\beta}_1$: Marks will increase by 4.06 for every hour student's spent to study.

(b) use the equation of the fitted line to predict what marks would be observed when the number of hours are 16 and 20 hours. Do you think the marks are logical?

When $x = 16$, $\hat{y} = 26.8922 + 4.0593(16) = 91.841$

When $x = 20$, $\hat{y} = 26.8922 + 4.0593(20) = 108.0782$

Not logical since marks can't go beyond 100 marks or negative value.

5.4 HYPOTHESIS TESTING FOR SIMPLE REGRESSION MODEL

RECALL THAT:

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

There are two parameters in simple linear regression model:

β_0 = Intercept

β_1 = Slope

5.4.1 HYPOTHESIS TESTING FOR INTERCEPT

The regression line is passing through the origin if the intercept is zero, $\beta_0 = 0$.

Hypothesis (two-sided test):

H_0 : The intercept of regression line is zero or $\beta_0 = 0$

H_1 : The intercept of regression line is not zero or $\beta_0 \neq 0$

The Test Statistic:

$$t = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0}{\sqrt{\left(\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}\right)\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

Standard error of the intercept

Residual mean square

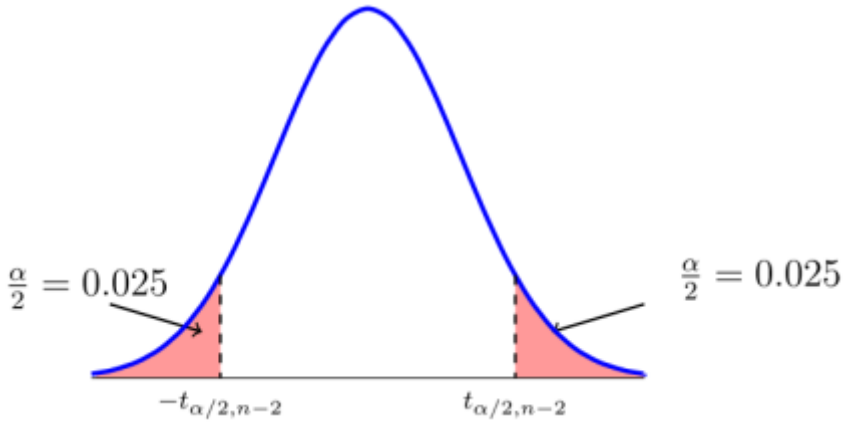
How to reject *H*0?

1. Using critical value

If $t_{test} > \left(critical\ value = t_{\frac{\alpha}{2},n-2}\right)$ or $t_{test} < \left(critical\ value = -t_{\frac{\alpha}{2},n-2}\right)$, then reject H_0

2. Using P-value

If $P-\text{value} < \alpha$, then reject H_0



CONCLUSION: (example)

We can conclude that the intercept of regression line is not zero at α = 0.05 significance level.

That is, the regression line does not pass through the origin.

Example 5

Using the data and results from Examples 3 and 4, test the hypothesis that $\beta_0 = 0$ at $\alpha=0.05$ significance level.

SOLUTION:

$S_{xy} = 274$ $S_{xx} = 67.5$ $S_{yy} = 1246$ $\hat{y} = 4.0593 + 26.8922x$

$\bar{x} = 10.25$ $\bar{y} = 68.5$ $n = 8$

$$t = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0}{\sqrt{\left(\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}\right)\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

Example 5: solution

Step 1 – State the hypothesis

H_0 : the intercept of regression line is zero or $\beta_0 = 0$

H_1 : the intercept of regression line is not zero or $\beta_0 \neq 0$

Step 2 – Test Statistics

$$t_{test} = \frac{\hat{\beta}_0 - \beta_0}{se\left(\hat{\beta}_0\right)} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left[\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}\right]\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{26.8922 - 0}{\sqrt{\frac{1246 - 4.0593(274)}{6}\left(\frac{1}{8} + \frac{10.25^2}{67.5}\right)}} = 4.3924$$

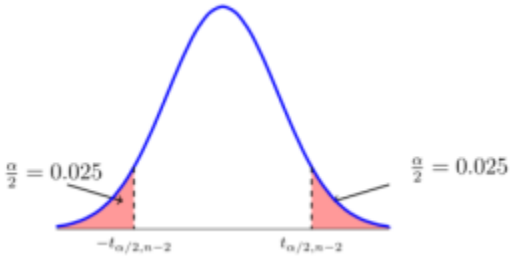
Step 3 - Critical value

$$\pm t_{\alpha/2, n-2} = \pm t_{0.025,6} = \pm 2.4469$$

Example 5: solution

Step 4 - Compare test statistic and critical value

$$4.3924 > 2.4469$$



Decision: Reject H_0, therefore $\beta_0 \neq 0$. Therefore, we can conclude that the intercept of regression line is not zero at α=0.05 significance level. Graphically, t_{test}=4.3924 is in the rejection region.

Step 5- Conclusion

There is sufficient evidence that the regression line does not pass through the origin at α = 0.05.

5.4.2 HYPOTHESIS TESTING FOR SLOPE

Two variables have a linear relationship if the slope of regression line is not zero, $\beta_1 \neq 0$.

Linearity Hypothesis (two-sided test):

H_0 : The slope of regression line is zero or $\beta_1 = 0$

H_1 : The slope of regression line is not zero or $\beta_1 \neq 0$

The Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}\left(\frac{1}{S_{xx}}\right)}} = \frac{\hat{\beta}_1}{\sqrt{\left(\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}\right)\left(\frac{1}{S_{xx}}\right)}}$$

Standard error of the slope

Residual mean square

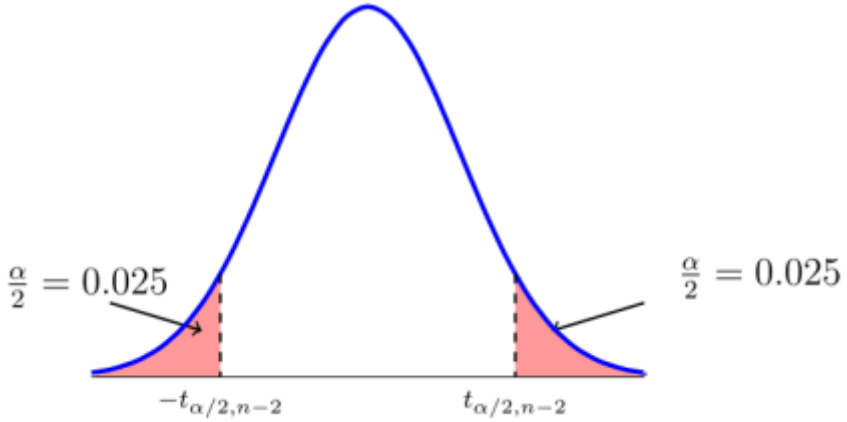
How to reject *H*0?

1. Using critical value

If $t_{test} > \left(critical\ value = t_{\frac{\alpha}{2},n-2}\right)$ or $t_{test} < \left(critical\ value = -t_{\frac{\alpha}{2},n-2}\right)$, then reject H_0

2. Using P-value

If $P-\text{value} < \alpha$, then reject H_0



CONCLUSION: (example)

We can conclude that the slope of regression line is not zero at α = 0.05 significance level.

That is, there is sufficient evidence that the variables *x* and *y* have a linear relationship at α = 0.05 .

Example 6

Again, using the data and results from Examples 3 and 4, test the linearity between *x* and *y* at α=0.05 significance level.

Solution:

Step 1 – State the hypothesis

H_0 : the slope of regression line is zero or $\beta_1 = 0$
H_1 : the slope of regression line is not zero or $\beta_1 \neq 0$

$S_{xy} = 274$	$S_{xx} = 67.5$	$S_{yy} = 1246$	$\hat{y} = 4.0593 + 26.8922x$
$\bar{x} = 10.25$	$\bar{y} = 68.5$	$n = 8$	

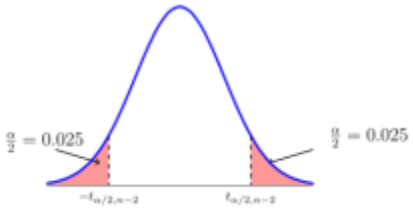
Example 6: solution

Step 2 – Test Statistics

$$t_{test} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\left[\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}\right]\left(\frac{1}{S_{xx}}\right)}} = \frac{4.0593 - 0}{\sqrt{\frac{1246 - 4.0593(274)}{6}\left(\frac{1}{67.5}\right)}} = 7.0636$$

Step 3 - Critical value

$$\pm t_{\alpha/2,n-2} = \pm t_{0.025,6} = \pm 2.4469$$



Step 4 - Compare test statistic and critical value

$$7.0636 > 2.4469$$

Decision: Reject H_0, therefore $\beta_1 \neq 0$. Therefore, we can conclude that the slope of regression line is not zero at α=0.05 significance level. Graphically, t_{test}=7.0636 is in the rejection region.

Step 5- Conclusion

There is sufficient evidence that the variables *x* and *y* have a linear relationship at α = 0.05.

5.4.3 HYPOTHESIS TESTING FOR SLOPE USING ANOVA

Hypothesis:

H_0 : The slope of regression line is zero or $\beta_1 = 0$

H_1 : The slope of regression line is not zero or $\beta_1 \neq 0$

Source of variations	Sum of squares	Degrees of Freedom	Mean of Squares	f_{test}
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	$MS_R = SS_R/1$	MS_R/MS_{Res}
Residual	$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}$	$n-2$	$MS_{Res} = SS_{Res}/(n-2)$	
Total	$SS_T = S_{yy}$	$n-1$		

ANOVA table for hypothesis testing of slope of regression line

How to reject *H*0?

1. Find critical value

$$f_{\alpha,1,n-2} \ (right\ tailed\ test)$$

2. Compare and decide

$$\text{If } f_{test} > f_{\alpha,1,n-2}, \text{ then Reject } H_0: \beta_1 = 0,$$

that is the slope of regression line is not zero.

3. CONCLUSION:

There is a linear relationship between the dependent and independent variables at α.

Example 7

Given a fitted model,

$$\hat{y} = 2627.82 - 37.15x, \text{where } n = 20,\ S_{yy} = 1693737.60 \text{ and } S_{xy} = -41112.65.$$

Complete the ANOVA table and test the hypothesis that $\beta_1 = 0$ at significance level $\alpha = 0.01$.

Solution:

Step 1 – State the hypothesis

H_0 : the slope of regression line is zero or $\beta_1 = 0$
H_1 : the slope of regression line is not zero or $\beta_1 \neq 0$

Example 7: solution

Step 2 – Complete the ANOVA table:

Source of variation	Sum of squares	Degrees of freedom	Mean Square	f_{test}
Regression	1527334.95	1	1527334.95/1 = 1527334.95	1527334.95/9244.59 = 165.21
Residual	166402.65	18	166402.65/18 = 9244.59	
Total	1693737.60	19		

Step 3 – Critical value: $f_{0.01,1,8} = 8.29 \ \ (right\ tailed\ test)$

Step 4 – Compare and decide: $(f_{test} = 165.21) > (f_{0.01,1,8} = 8.29)$, Reject H_0

Step 5 – Conclusion: There is a linear relationship between the dependent and independent variables at 0.01 significance level.

5.5 REGRESSION ANALYSIS USING MICROSOFT EXCEL

Step 1 : Key-in the data in Excel. Excel requires that the variables are in adjoining columns.

Students	numbers of hours (x)	Final marks (y)
A	5	49
B	8	60
C	9	55
D	10	72
E	10	65
F	12	80
G	13	82
H	15	85

Step 2 : Tools – Data Analysis – Regression – enter the data range (y & x) – ok

Computer Output - Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9447995
R Square	0.8926461
Adjusted R Square	0.87475378
Standard Error	4.72163395
Observations	8

r = 0.9448, Strong Linear positive correlation

R^2 = 0.8926, 89.26% of the variation in y can be explained by x.

Use adjusted R square if more than one independent variable

$\sqrt{MS_{Res}} = \sqrt{22.29383} = 4.7216$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1112.237037	1112.237	49.88991	0.000403286
Residual	6	133.762963	22.29383		
Total	7	1246			

Alternative: Use ANOVA table to test the linear relationship between variables.

The intercept of regression line is not zero (P-value < 0.05)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	26.8925926	6.122634851	4.392323	0.004606	11.91103386	41.87415	11.91103	41.87415
X Variable 1	4.05925926	0.574698983	7.063279	0.000403	2.653020479	5.465498	2.65302	5.465498

$\hat{y} = 26.8926 + 4.0593x$

x and y have linear relationship (P-value < 0.05)

5.6 MULTIPLE LINEAR REGRESSION ANALYSIS

- ✓ A multiple regression equation is use to describe linear relationships involving more than two variables.

- ✓ A multiple linear regression equation expresses a linear relationship between a response variable *y* and two or more predictor/regressor variable ($x_1, x_2, \ldots, x_k$).

- ✓ The general form of a multiple regression equation is

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

- ✓ A multiple linear regression equation identify the plane that gives the best fit to the data.

5.6.1 MULTIPLE LINEAR REGRESSION EQUATION

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where

n : sample size

k : number of regressor (independent) variables

y : response (dependent) variable

$x_1, x_2 \ldots, x_k$: regressor (independent) variables

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$: unknown parameters (regression coefficients)

β_0 : y-intercept of regression line

β_1 : shows the mean change in x_1 when $x_2 \ldots, x_k$ are held constants

ε : the error term

Estimated Form of the Multiple Linear Regression Equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

where

$\hat{y}$: predicted value of y

$x_1, x_2 \ldots, x_k$: regressor (independent) variables

k : number of regressor (independent) variables

$\hat{\beta}_0$: estimated value of y-intercept

$\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$: estimated value of regression coefficients

Examples of real situation

1) A manufacturer of jams wants to know where to direct its marketing efforts when introducing a new flavour. Regression analysis can be used to help determine the profile of heavy users of jams. For instance, a company might predict the number of flavours of jam a household might have at any one time on the basis of a number of independent variables such as, number of children living at home, age of children, gender of children, income and time spent on shopping.

2) Many companies use regression to study an impact of market share, purchase frequency, product ownerships, and product & brand loyalty on markets segments.

Examples of real situation

3) Company directors explore the relationships of employee salary levels to geographic location, unemployment rates, industry growth, union membership, industry type, or competitive salaries.

4) Financial analysts look for causes of high stock prices by analysing dividend yields, earning per share, stock splits, consumer expectation of interest rates, savings levels and inflation rates.

5) Medical researchers use regression analysis to seek links between blood pressure and independent variables such as age, social class, weight, smoking habits and race.

6) Doctors explore the impact of communications, number of contacts, and age of patient on patient satisfaction with service.

5.6.2 COMPUTING THE MULTIPLE LINEAR REGRESSION MODEL

Using the least square method, the multiple linear regression equation is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

where the estimated regression coefficients are

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

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Example 8

A sales manager of computer company needs to predict sales of monopod in selected market area. He believes that advertising expenditures and the population in each market area can be used to predict monopod sales.

He gathered sample of monopod sales, advertising expenditures and the population as shown below.

Find the estimated multiple linear regression model which gives the best fit to the data.

EXAMPLE 8, cont...

Market Area	Advertising Expenditures (x_1) RM (in Thousands)	Population (x_2) (Thousands)	Monopod sales (y) RM (in Thousands)
A	1.0	200	100
B	5.0	700	300
C	8.0	800	400
D	6.0	400	200
E	3.0	100	100
F	10.0	600	400

Example 8: Solution

Since we have 2 independent variables, so the multiple regression equation is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

where

$n = 6$ and $k = 2$

$\hat{y}$ = predicted value of y

x_1 = the value of the first independent variable

x_2 = the value of the second independent variable

$\hat{\beta}_0$ = estimate value of y-intercept

$\hat{\beta}_1$ = slope associated with x_1 (the change in $\hat{y}$ if x_2 is held constant and x_1 varies by 1 unit)

$\hat{\beta}_2$ = slope associated with x_2 (the change in $\hat{y}$ if x_1 is held constant and x_2 varies by 1 unit)

Solve Multiple Linear Regression Using Microsoft Excel

STEP 1: Excel – key in data in column form and adjacent to each other.

Market Area	Advertising Expenditures (x 1) RM (in Thousands)	Population (x 2) (Thousands)	Monopod sales (y) RM (in Thousands)
A	1	200	100
B	5	700	300
C	8	800	400
D	6	400	200
E	3	100	100
F	10	600	400

STEP 2: Tools – Data Analysis – Regression – enter the data range (y & x) – ok

STEP 3: Analyse the Excel Output.

Example 8: solution using Microsoft Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98674032
R Square	0.97365646
Adjusted R Square	0.9560941
Standard Error	28.8827296
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	92497.3638	46248.68	55.43996	0.004275739
Residual	3	2502.636204	834.2121		
Total	5	95000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.39718805	25.98627725	0.246176	0.821429	-76.30282156	89.097198	-76.302822	89.0971977
X Variable 1	20.4920914	5.882177186	3.48376	0.039947	1.772360776	39.211822	1.77236078	39.211822
X Variable 2	0.28049209	0.068601659	4.088707	0.026442	0.062170791	0.4988134	0.06217079	0.49881339

$$\hat{y} = 6.3972 + 20.4921x_1 + 0.2805x_2$$

Interpreting the Values in the Equation

$$\hat{y} = 6.3972 + 20.4921x_1 + 0.2805x_2$$

Coefficient	Explanation
$\hat{\beta}_0 = 6.3972$	The estimated value of *y* when x_1 and x_2 are both zeros.
	The estimated value of monopod sales is RM6.3972 thousand when the advertising expenditures and population are zeros.
$\hat{\beta}_1 = 20.4921$	When the value of x_2 is held constant, the estimated value of *y* increases by RM20.4921 (in thousands) for every RM1000 of x_1.
	When the population is constant, the estimated value of monopod sales is increases by RM20.4921 (in thousands) for every RM1000 of advertising expenditures.
$\hat{\beta}_0 = 0.2805$	When the value of x_1 is held constant, the estimated value of *y* increases by RM0.2805 (in thousands) for every 1000 of x_2.
	When the advertising expenditures is constant, the estimated value of monopod sales is increases by RM0.2805 (in thousands) for every 1000 of population.

What is Regression Model Objectives?

Multiple linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

is used to:

1. Predict the value of dependent variable
   - ✓ Example: predicting sales and output

2. Estimate the marginal effects of each independent variable
   - ✓ Example: want to know how changes of independent variables x_j, $j = 1, \ldots, k$ change the dependent variable.

Example 9: Making predictions with the multiple regression model

Assume that the computer company had recently spent RM4,000 for advertisement in a particular market area which has a population of 500,000 people. Using the multiple linear regression model obtained previously, the sales manager can predict the total monopod sales as follows:

When x_1=4 and x_2 = 500 (in thousands), then

$$\begin{aligned}\hat{y} &= 6.3972 + 20.4921x_1 + 0.2805x_2 \\ &= 6.3972 + 20.4921(4) + 0.2805(500) \\ &= 228.613 \quad \rightarrow \quad RM\,228613\end{aligned}$$

5.6.3 INTERPRETATION OF REGRESSION STATISTICAL OUPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98674032
R Square	0.97365646
Adjusted R Square	0.9560941
Standard Error	28.8827296
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	92497.3638	46248.68	55.43996	0.004275739
Residual	3	2502.636204	834.2121		
Total	5	95000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.39718805	25.98627725	0.246176	0.821429	-76.30282156	89.097198	-76.302822	89.0971977
X Variable 1	20.4920914	5.882177186	3.48376	0.039947	1.772360776	39.211822	1.77236078	39.211822
X Variable 2	0.28049209	0.068601659	4.088707	0.026442	0.062170791	0.4988134	0.06217079	0.49881339

Multiple *R*

The coefficient of multiple correlations *R* is the positive square root of R^2.

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

The value of *R* can range from 0 to +1. The closer to +1, the stronger the relationship. The closer to 0, the weaker the relationship.

The Coefficient of Multiple Determinations r^2

- ✓ Measure the percentage of variation in the *y* variable associated with the use of the set *x* variables
- ✓ A percentage that shows the variation in the *y* variable that's explain by its relation to the combination of x_1 and x_2.

$$R^2 = \frac{SSR}{SST}$$

where

$$SSR = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y} - \left(\frac{1}{n}\right) \mathbf{Y}^T \mathbf{1}\mathbf{1}^T \mathbf{Y}$$

$$SST = \mathbf{Y}^T \mathbf{Y} - \left(\frac{1}{n}\right) \mathbf{Y}^T \mathbf{1}\mathbf{1}^T \mathbf{Y}$$

and $0 \le R^2 \le 1$

$R^2 = 0$ → **when all** $\hat{\beta}_i = 0 (i = 1, \ldots, k)$

$R^2 = 1$ → **when all observations fall directly on the fitted response surface, i.e. when** $Y_t = \hat{Y}_t$ **for all *i*. (the regression equation is good)**

Adjusted R^2

The adjusted coefficient of determination is the multiple coefficient of determination R^2 modified to account for the number of variables and the sample size.

$$adjusted\ R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

PROPERTIES:

1. When comparing a multiple regression equation to others, it is better to use the adjusted R^2.
2. The adjusted R^2 can take any values less than or equal to 1. (can be negative)
3. If the adjusted R^2 is close to 1, it indicates a better fit.
4. If the adjusted R^2 is negative, it indicates that the model contains terms that do not help to predict the response.

Interpretation of Multiple *R*, R^2, and adjusted R^2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98674032
R Square	0.97365646
Adjusted R Square	0.9560941
Standard Error	28.8827296
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	92497.3638	46248.68	55.43996	0.004275739
Residual	3	2502.636204	834.2121		
Total	5	95000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.39718805	25.98627725	0.246176	0.821429	-76.30282156	89.097198	-76.302822	89.0971977
X Variable 1	20.4920914	5.882177186	3.48376	0.039947	1.772360776	39.211822	1.77236078	39.211822
X Variable 2	0.28049209	0.068601659	4.088707	0.026442	0.062170791	0.4988134	0.06217079	0.49881339

$R = 0.9867$. Strong relationship exist between sales and advertising expenditures and population size

97.4% of monopod sales in the market area is explained by advertising expenditures and population size

Used this values if we want to compare which model is the best to fit the data

$$\hat{y} = 6.3972 + 20.4921x_1 + 0.2805x_2$$

Standard error

- ✓ Measure the extent of the scatter, or dispersion, of the sample data points about the multiple regression planes.
- ✓ Compare the standard error between two or more regression equation.
- ✓ Smallest standard error:
  - Data is less dispersed
  - Data is closed to regression line.

ANOVA test

The Hypothesis testing is:

H_0: Neither of the independent variables is related to the dependent variables ($\beta_1 = \beta_2 = \ldots = \beta_k = 0$)

H_1: At least one of the independent variables is related to the dependent variables ($\beta_j \neq 0$ for at least one j)

- ✓ Reject H_0 if *significance F* $< \alpha$
- ✓ *significance F = P value*

Interpretation of ANOVA table

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98674032
R Square	0.97365646
Adjusted R Square	0.9560941
Standard Error	28.8827296
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	92497.3638	46248.68	55.43996	0.004275739
Residual	3	2502.636204	834.2121		
Total	5	95000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.39718805	25.98627725	0.246176	0.821429	-76.30282156	89.097198	-76.302822	89.0971977
X Variable 1	20.4920914	5.882177186	3.48376	0.039947	1.772360776	39.211822	1.77236078	39.211822
X Variable 2	0.28049209	0.068601659	4.088707	0.026442	0.062170791	0.4988134	0.06217079	0.49881339

H_0: Neither of the independent variables is related to the dependent variables ($\beta_1 = \beta_2 = 0$)

H_1: At least one of the independent variables is related to the dependent variables ($\beta_j \neq 0$ for at least one j, j =1,2)

0.0043 < 0.05 So reject H_0

Hence, at least one of the independent variables is related to dependent variable.

$$\hat{y} = 6.3972 + 20.4921x_1 + 0.2805x_2$$

5.7: MODEL SELECTION

TIPS: Model Selection in simple way

- ✓ Use common sense and practical considerations to include or exclude variables.
- ✓ Consider the *P*-value from ANOVA table (the measure of the overall significance of multiple regression equation -significance F value) displayed by computer output. The smaller the better.
- ✓ Consider equation with high values of r^2 for simple linear regression or high value of adjusted r^2 and try include only a few significant variables.
- ✓ Find the linear correlation coefficient *r* for each pair of variables being considered. If 2 predictor values have a very high *r*, there is no need to include them both. Exclude the variable with the lower value of *r*.

Model Selection for multiple regression using Microsoft Excel

- If there are k independent variables (x_1, x_2, …,x_k), there will be a total of $2^k - 1$ possible models.
- **Example**: If there are 3 variables (A, B, C)
  - Single variable: A, B, C
  - Two variables: AB, AC, BC
  - Three variables: ABC
    - Total = 7 possible models
    - Do the regression analysis using Microsoft Excel separately for each model.

Model Selection: PROCEDURE

1. Perform the regression analysis for all $2^k - 1$ possible models using Microsoft Excel.

2. Summarise all the Excel Output in a table. The summary table should contain
   - P-value, r^2, adjusted r^2, and regression equation.

3. Chose the best model that fit the data.

Example 10

The following table summarize the multiple regression analysis for the response variable (y) which is weight (in pounds), and the predictor (x) variables are H (height in inches), W (waist circumference in cm), and C (cholesterol in mg).

x	P-value	r^2	Adjusted r^2	Regression equation
H	0.0001	0.273	0.254	$-139 + 4.55H$
W	0.0000	0.790	0.785	$-44.1 + 2.37W$
C	0.874	0.001	0.000	$173 - 0.003C$
H, W	0.0000	0.877	0.870	$-206 + 2.66H + 2.15W$
H, C	0.0002	0.277	0.238	$-148 + 4.65H + 0.006C$
W, C	0.0000	0.804	0.793	$-42.8 + 2.41W + 0.01C$
H, W, C	0.0000	0.880	0.870	$-199 + 2.55H + 2.18W - 0.005C$

Example 10: solution

a) If only one predictor variable is used to predict weight, which single variable is best? Why?

x	P-value	r^2	Adjusted r^2	Regression equation
H	0.0001	0.273	0.254	-139 + 4.55H
W	0.0000	0.790	0.785	-44.1 + 2.37W
C	0.874	0.001	0.000	173 - 0.003C

Answer:

Choose Waist (W): $\hat{y} = -44.1 + 2.37W$

since $(\text{P-value} = 0.0000) < (\alpha = 0.05)$, has significant effect with high $r^2 = 0.790$.

Height (H) is not chosen because the correlation is considered very weak with $r^2 = 0.273$.

Example 10: solution

b) If exactly two predictor variables are used to predict weight, which two variables should be chosen? Why?

x	P-value	r^2	Adjusted r^2	Regression equation
H, W	0.0000	0.877	0.870	$-206 + 2.66H + 2.15W$
H, C	0.0002	0.277	0.238	$-148 + 4.65H + 0.006C$
W, C	0.0000	0.804	0.793	$-42.8 + 2.41W + 0.01C$

Answer:

Choose Height, Waist (H, W): $\hat{y} = -206 + 2.66H + 2.15W$ with (P-value $= 0.0000$) $<$ $(\alpha = 0.05)$ (significant effect) and high adjusted $r^2 = 0.870$.

Since C has very low (P-value $= 0.0000$) and very weak correlation ($r^2 = 0.001$), so any combination with C is not preferred. Hence choose H, W.

Example 10: solution

c) Which regression equation is best for predicting weight? Why?

x	P-value	r^2	Adjusted r^2	Regression equation
W	0.0000	0.790	0.785	$-44.1 + 2.37W$
H, W	0.0000	0.877	0.870	$-206 + 2.66H + 2.15W$
H, W, C	0.0000	0.880	0.870	$-199 + 2.55H + 2.18W - 0.005C$

Answer:

Hence, $\hat{y} = -206 + 2.66H + 2.15W$ is the best for predicting weight based on the above reasons.

REFERENCES

1. A.G. Bluman. 2007. *Elementary Statistics: A Step by Step Approach.* Sixth Edition. McGraw-Hill.
2. D.C. Montgomery, E.A. Peck and G.G. Vining. 2012. *Introduction to Linear Regression Analysis*. 5th Edition, Wiley Series in Probability and Statistics, John Wiley and Sons, Inc. New Jersey.
3. D.C. Montgomery and G.C. Runger. 2011. *Applied Statistics and Probability for Engineers (SI version)*. Fifth Edition, John Wiley and Sons, Inc. New Jersey.
4. W. Navidi. 2011. *Statistics for Engineers and Scientists*. Third Edition. McGraw-Hill
5. Triola, M.F. 2006. *Elementary Statistics.* 10^{th} Edition. UK: Pearson Education.
6. Satari S. Z. et al. Applied Statistics Module New Version. 2015. Penerbit UMP. Internal used.

Thank You.

NEXT: CHAPTER 6 Goodness of Fit and Contingency Table.