NANYANG POLYTECHNIC

Course : Diploma in Electronic Systems
Diploma in Telematics & Media Technology
Diploma in Aerospace Systems & Management
Diploma in Electrical Engineering with Eco-Design
Diploma in Mechatronics Engineering
Diploma in Digital & Precision Engineering
Diploma in Aeronautical & Aerospace Technology
Diploma in Biomedical Engineering
Diploma in Nanotechnology & Materials Science
Diploma in Engineering with Business
Diploma in Information Technology
Diploma in Financial Informatics
Diploma in Cybersecurity & Forensics
Diploma in Infocomm & Security
Diploma in Chemical & Pharmaceutical Technology
Diploma in Biologics & Process Technology
Diploma in Chemical & Green Technology

Module : Engineering Mathematics 2B / Mathematics 2B/ Computing Mathematics 2 – EG1761/2008/2681/2916/2961
EGB/D/F/H/J/M207
IT1201/1531/1631/1761
CLB/C/G201

Topic 2 : Linear Regression and Correlation

Objectives :

At the end of this lesson, the student should be able to:

1 explain the line of best fit and linear correlation between two variables
2 find the correlation coefficient and the equation of regression line
3 interpret the output from simple linear regression analysis and find the predicted values

Topic 2: Linear Regression and Correlation

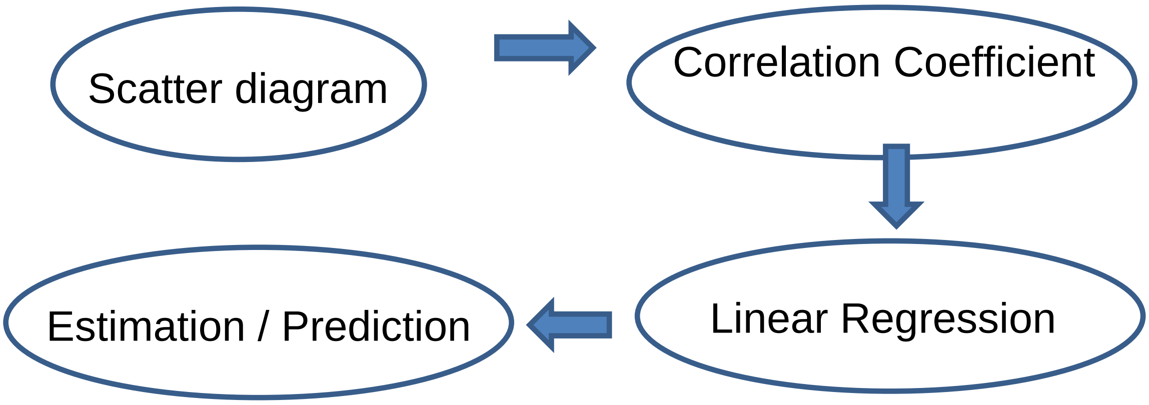
2.1 Introduction

- In the previous chapter we have been dealing with data of one variable. In this chapter, we will study data with two variables and the relationship between them.

- Examples of data with two variables:

  (a) Class test results vs final exam results,
  (b) Blood pressure vs age of a person,
  (c) Price of a car vs price of a 3 room HDB flat.

- An overview of this chapter is shown below:



Scatter diagram

Correlation Coefficient

Linear Regression

Estimation / Prediction

2.2 Scatter Diagram

- Given pairs of observed data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, we can plot them on the $x - y$ axes to obtain a scatter diagram.

- Scatter diagrams are useful as they provide visual information whether variables X and Y share any special relationship. Some examples of scatter diagrams are shown below:

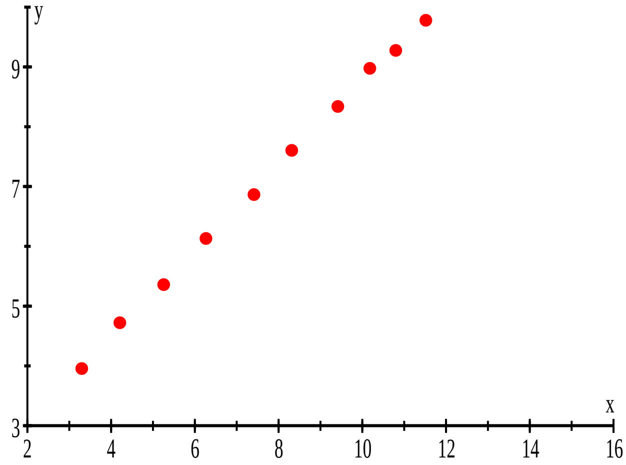


Diagram 2.2.1

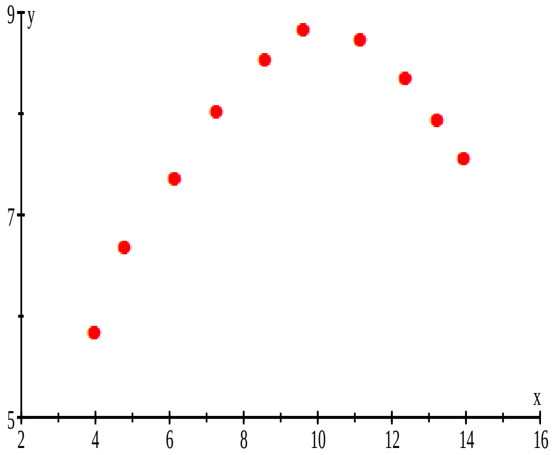


Diagram 2.2.2

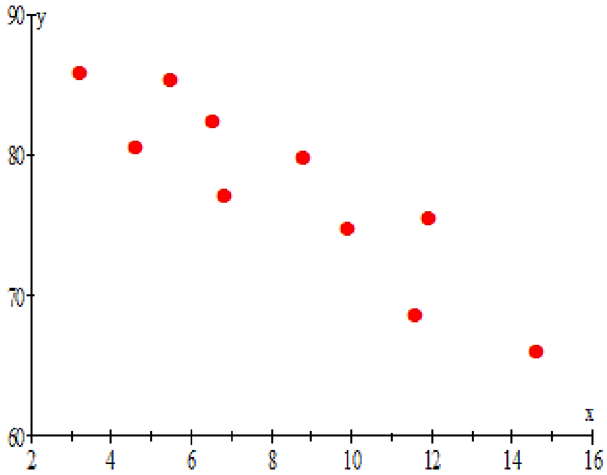


Diagram 2.2.3

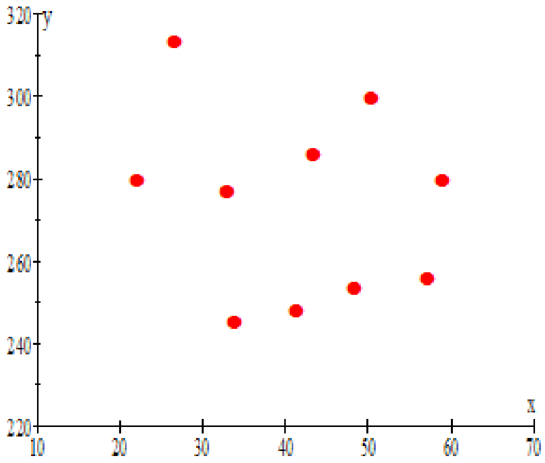


Diagram 2.2.4

- Variables X and Y have a **positive** relationship if X increases, Y increases (i.e. there is a **upward** trend). Variables X and Y have a **negative** relationship if X increases, Y decreases (i.e. there is a **downward** trend).

- Variables X and Y have a **linear** relationship if the observed values of X and Y can be described using a straight line equation $y = mx + c$.

- Diagrams 2.2.1 and 2.2.3 illustrate a positive and negative **linear** relationship between X and Y respectively. Diagram 2.2.2 shows that X and Y have a **curvilinear** relationship while Diagram 2.2.4 shows that X and Y have no obvious relationship.

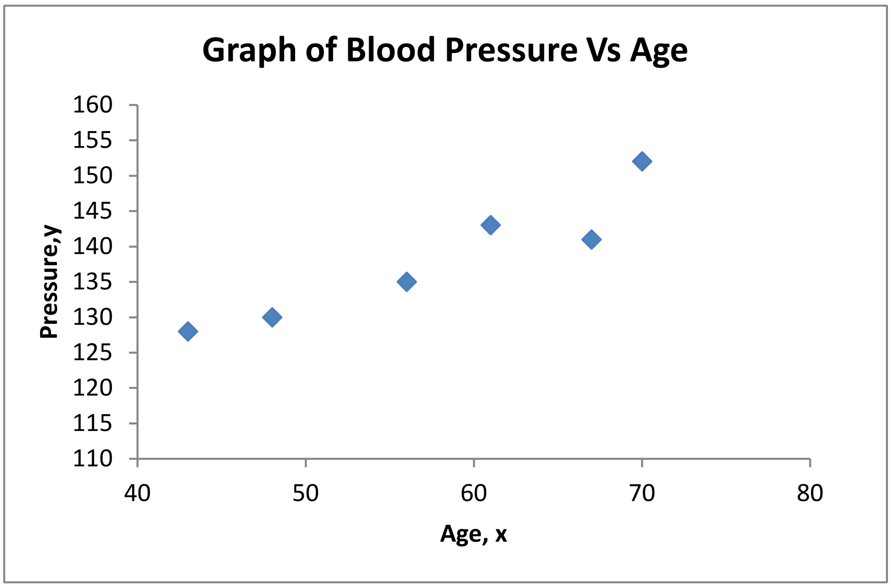
Example 2.2-1

For each of the scatter plot shown below, describe whether

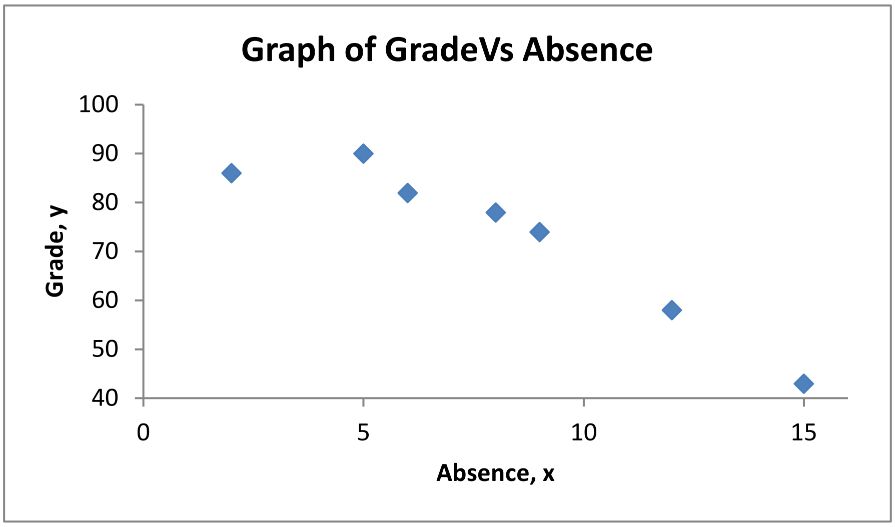
(i) the relationship between X and Y is positive or negative,

(ii) the relationship between X and Y is linear, curvilinear or not obvious.

(a)



(b)



Solution:

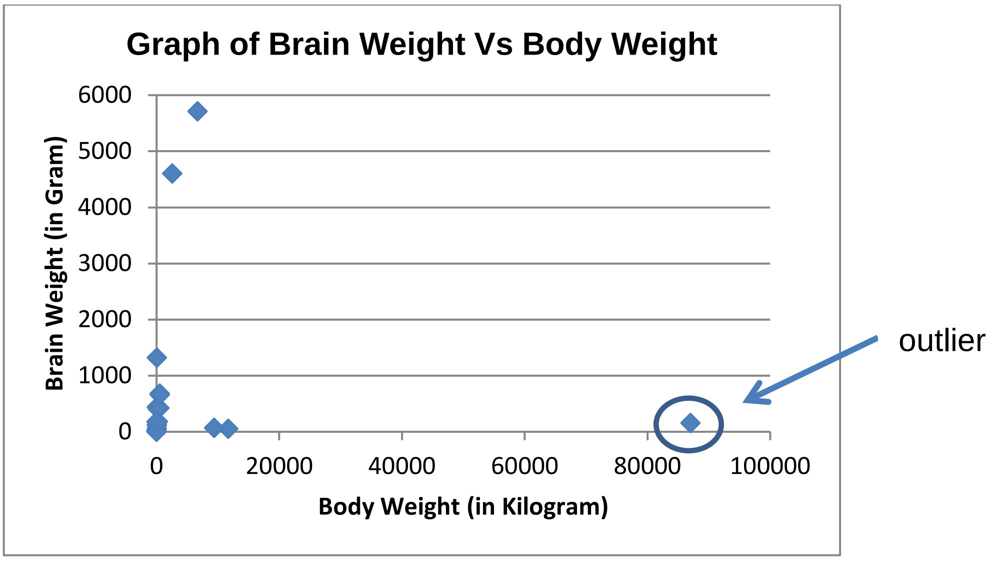
a) The relationship between X and Y is ______________ and _________________.

b) The relationship between X and Y is ______________ and _________________.

- In some situations we may have **outliers** (observation points that is distant from the rest) that will distort the shape of the scatter plot. We may need to apply **transformation** of the data values so that the relationship between the variables is more visible.

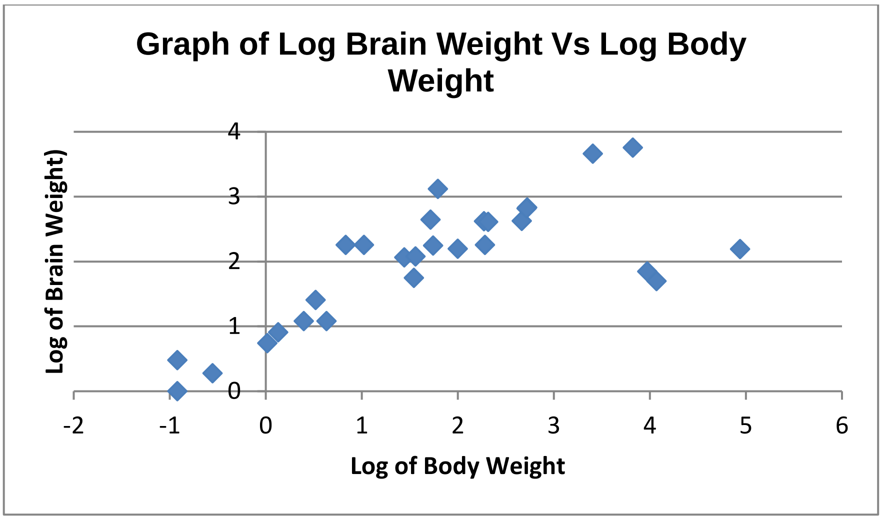
Example 2.2-2

We want to investigate the relationship between body and brain weights of different animals. The scatter diagram of the brain weights (in grams) and the body weights (in kilograms) of 28 animals is shown below:



We observed at least one point that is distant from the rest. Hence it distorts the shape of the scatter diagram and we may not see any clear relationship between the variables.

We can apply transformation by taking **logarithm** to both the observed values of body and brain weight and plot the scatter diagram shown below:



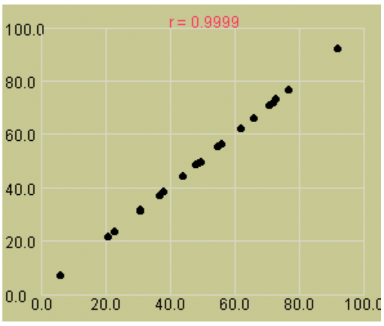
Now we observe that the new variables exhibit a clearer linear relationship.

2.3 Correlation Coefficient

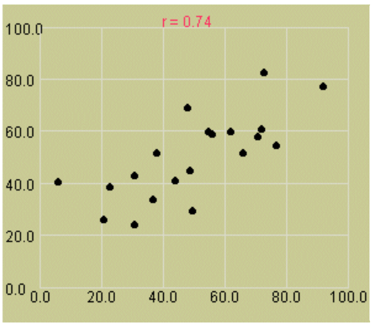
- Sometimes it is difficult to use our eyes to determine through the scatter diagram whether there is indeed a linear relationship between the variables. (Refer to Example 2.2.1).
- Therefore we will need precise mathematical calculation to help us determine the **degree of linearity** in the relationship between two variables. This is done through the **Pearson product moment correlation coefficient, *r*.**
- The table below shows how the correlation coefficient, *r* indicates the **linear** relationship between two variables.

$-1 \le r \le 1$		
Strength of linear relationship	**Positive**	**Negative**
Perfect	$r = 1$	$r = -1$
Very strong	$0.8 \le r < 1$	$-1 < r \le -0.8$
Strong	$0.4 \le r < 0.8$	$-0.8 < r \le -0.4$
Weak	$0.2 \le r < 0.4$	$-0.4 < r \le -0.2$
Little / no relationship	$0 \le r < 0.2$	$-0.2 < r \le 0$

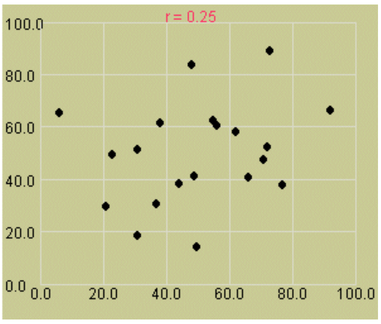
- The examples below show the shape of various scatter diagrams and r values.



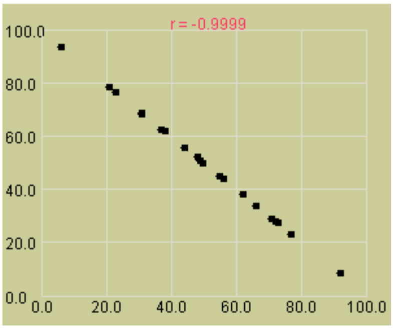
Very strong, positive correlation



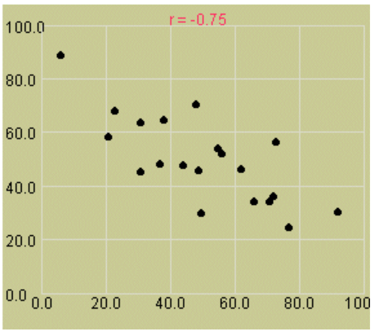
Strong positive correlation



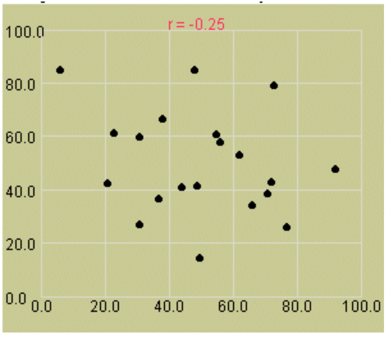
Weak positive correlation



Very strong, negative correlation



Strong negative correlation



Weak negative correlation

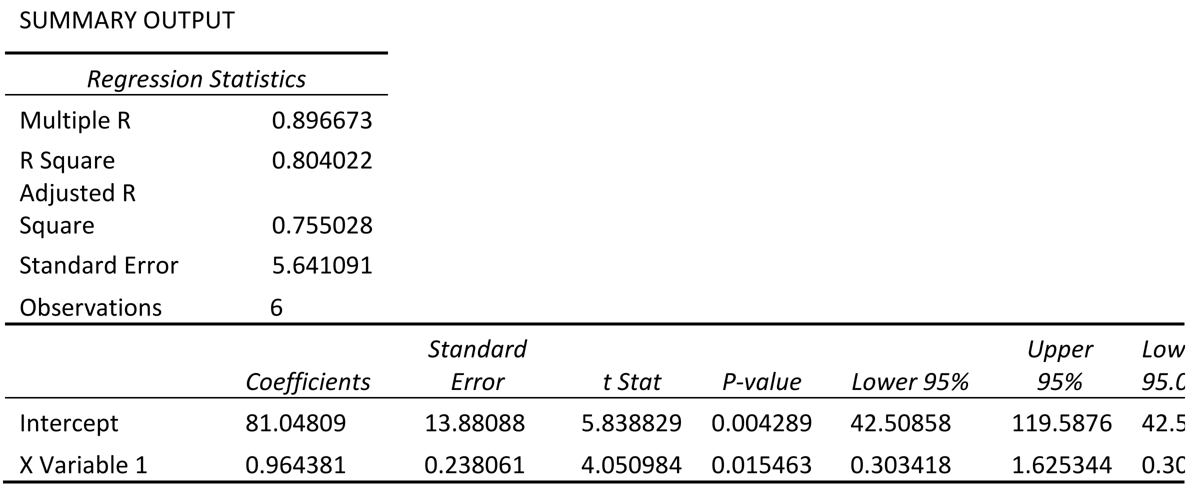
Source of data: http://www.seeingstatistics.com/seeing1999/resources/opening.html

- We can obtain the correlation coefficient of a data set through EXCEL. The correlation coefficient is obtained from “Multiple R” and the sign of “X variable coefficient”. The two examples below illustrate how to interpret the summary output from EXCEL.

Example 2.3-1

For each of the summary output below, state the value of the correlation coefficient and describe the relationship between X and Y

(a)

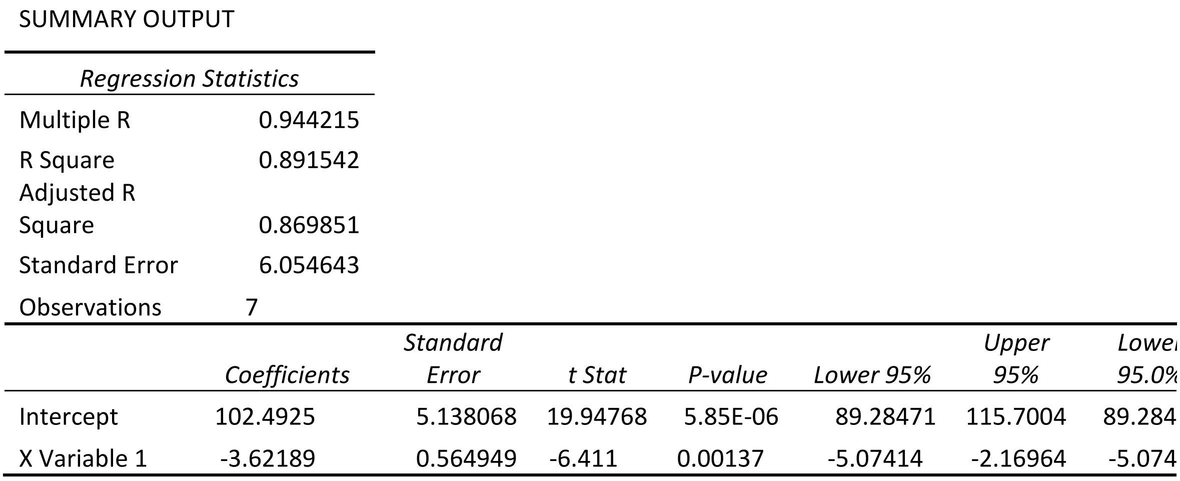


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.896673
R Square	0.804022
Adjusted R Square	0.755028
Standard Error	5.641091
Observations	6

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Low 95.C*
Intercept	81.04809	13.88088	5.838829	0.004289	42.50858	119.5876	42.5
X Variable 1	0.964381	0.238061	4.050984	0.015463	0.303418	1.625344	0.30

(b)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.944215
R Square	0.891542
Adjusted R Square	0.869851
Standard Error	6.054643
Observations	7

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lowe 95.0%*
Intercept	102.4925	5.138068	19.94768	5.85E-06	89.28471	115.7004	89.284
X Variable 1	-3.62189	0.564949	-6.411	0.00137	-5.07414	-2.16964	-5.074

Solution:

a) The correlation coefficient, *r* = ______________. X and Y has a _____________, ____________ linear relationship.

b) The correlation coefficient, *r* = ______________. X and Y has a _____________, ____________ linear relationship.

2.4 Simple Linear Regression

- If scatter diagram and correlation coefficient indicate that two variables share a linear relationship, we will model them using a **straight line** equation and see how one variable (**dependent variable**) changes its value according to another variable (**independent variable**).

- Some examples of dependent and independent variables

Dependent variables Y	Independent variables X
Blood pressure	Age
Sales of cold drinks	Climate temperature
Price of house bought	Monthly salary
Final exam score	Number of lessons missed

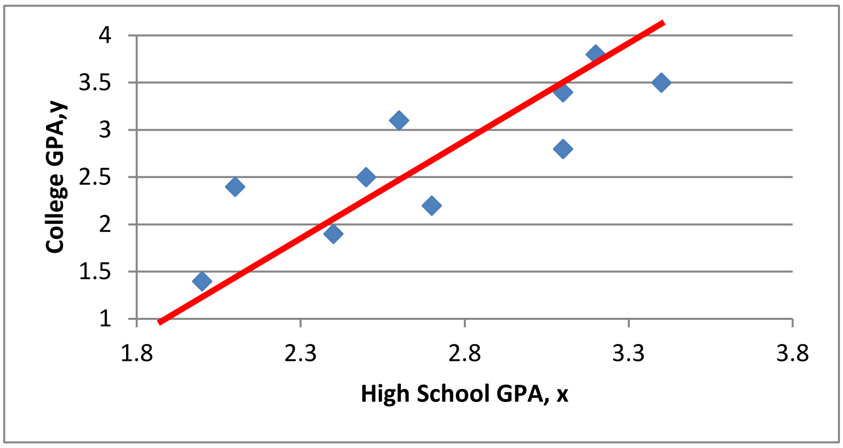
- Usually we will denote the independent variable as X and dependent variable as Y.
- A linear regression line Y on X is of the form $y = mx + c$, where the values of m and c can be obtained from EXCEL summary output (refer to Example 2.3-1):
  - m = coefficient of X variable,
  - c = coefficient of Intercept.

- The equation of the linear regression line (best fit line) is obtained using the **principle of least squared error** (refer to Appendix 2.2).

Example 2.4-1

The table below shows the high school GPA and the college GPA at the end of the 1st year for 10 different students:

Student	High School GPA, x	College GPA, y
1	2.7	2.2
2	3.1	2.8
3	2.1	2.4
4	3.2	3.8
5	2.4	1.9
6	3.4	3.5
7	2.6	3.1
8	2.0	1.4
9	3.1	3.4
10	2.5	2.5



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.843923
R Square	0.712206
Adjusted R Square	0.676232
Standard Error	0.433342
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.717716	3.717716	19.79767	0.002141
Residual	8	1.502284	0.187786		
Total	9	5.22			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	−0.95037	0.831773	-1.14258	0.286254	-2.86844	0.967706
X Variable 1	1.346999	0.302733	4.449458	0.002141	0.648895	2.045103

(a) Using the summary output

(i) state the correlation coefficient and the relationship between the two variables.

(i) write the equation of the regression line Y on X.

(b) Using the equation of the line in part (aii), find the college GPA if the High School GPA is 3.6.

(c) Using the line of best fit in part (aii), find the High School GPA if the College GPA is 2.3.

Solution:

(ai) The correlation coefficient is ________________. The two variables has a ____________ , _______________ and ______________ relationship.

(aii) The linear regression line Y on X is _________________________.

(b) when $x = 3.6$,

(c) when $y = 2.3$,

Example 2.4-2

The Financial World magazine uses its own complex formula to estimate how much the following brand names would be worth in cash. The table gives the brand name, its value in billions of dollars, Y and the company's revenue in billions, X:

Brand Name	Revenue	Value
Marlboro	15.4	31.2
Coca-Cola	0.4	4.4
Budweiser	6.2	10.1
Pepsi-Cola	5.5	9.6
Nescafe	4.3	8.5
Kellogg	4.7	8.4
Winston	3.6	6.1
Pampers	4	6.1
Camel	2.3	4.4
Campbell	2.4	3.9
Nestle	6	3.7
Hennessy	0.9	3
Heineken	3.5	2.7
Johnnie Walker	1.5	2.6
Louis Vuitton	0.9	2.6
Hershey	2.6	2.3
Guinness	1.8	2.3
Barbie	0.8	2.2
Kraft	2.8	2.2
Smirnoff	1	2.2
Del Monte	2.3	1.6
Wrigley's	1	1.5
Schweppes	1.3	1.4
Tampax	0.6	1.4
Heinz	0.8	1.3
Quaker	1.1	1.2
Colgate	1.1	1.2
Gordon's	0.6	1.1
Hermes	0.5	1
Kleenex	0.7	0.8
Carlsberg	0.8	0.7
Haagen-Dazs	0.5	0.6
Fisher-Price	0.6	0.6
Nivea	0.9	0.6
Sara Lee	0.8	0.5
Oil of Olay	0.6	0.5
Planters	0.7	0.5
Green Giant	1	0.4
Jell-o	0.3	0.4
Band-Aid	0.2	0.2
Ivory	0.4	0.2
Birds Eye	0.3	0.2

Source of data: Financial World, August 12, 1992

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	968.6591	968.6591	337.9662	4.11E-21
Residual	40	114.6457	2.866142		
Total	41	1083.305			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-0.55226	0.333114	-1.65787	0.105167	-1.22551
X Variable 1	1.819783	0.098988	18.38386	4.11E-21	1.61972

(a) Using the summary output, write the equation of line of best fit.

(b) Using the equation in (a), find the value of the brand name if the company's revenue is $5 billion, $10 billion and $25 billion,

Solution:

(a) The regression line is: $y =$ ______ $+$ _______ $* x$

(b) Substituting the values 5, 10 and 25 in for x and computing the values of y yield the following predicted values of brand names when the revenues are \$$5$, \$$10$ and \$$25$ billion:

Revenue	Predicted value
\$5 billion	$-0.55226 + 1.819783$ (____) = \$__________ billion
\$10 billion	$-0.55226 + 1.819783$ (____) = \$__________ billion
\$25 billion	$-0.55226 + 1.819783$ (____) = \$__________ billion

2.5 Reliability of estimation / prediction

(a) Regression and correlation analysis only attempts to find a relationship between two variables. Even if there is a very strong linear relationship, we **cannot conclude** any **causation** between the variables.

(i.e. If there is a strong positive linear relationship between blood pressure and age, we cannot conclude that age **causes** high blood pressure).

(b) Given a value of the independent variable X, we say the estimation on Y is **reliable** if

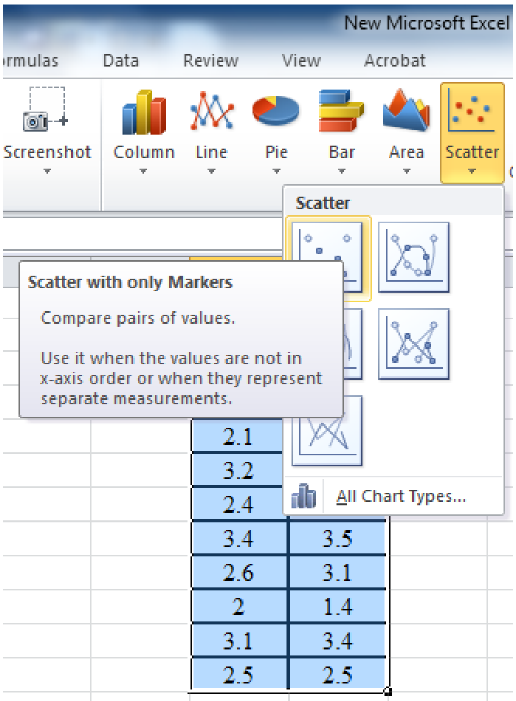
- the x value is within the data range of X provided,
- the correlation coefficient r is close to 1 or -1.

Appendix 2.1 Using EXCEL for Regression and Correlation

A2.1.1 Plot scatter diagram and regression line

- Refer to Example 2.4-1

Step 1: Highlight the data values for *x* and *y* (*x* on left *y* on right column) and go to "Insert" tab, select "Scatter". You may use the chart tools to add titles or to do adjustment of other settings.

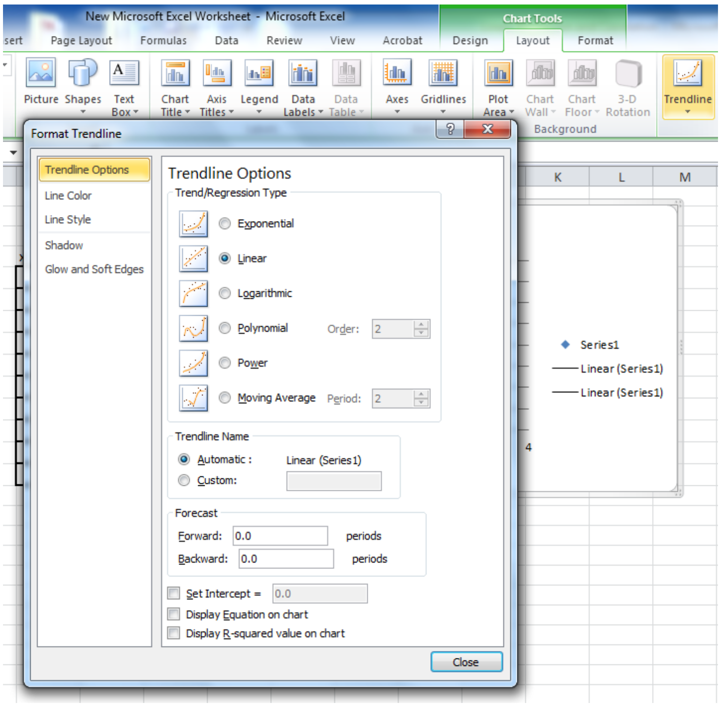


Scatter with only Markers

Compare pairs of values.

Use it when the values are not in x-axis order or when they represent separate measurements.

2.1	
3.2	
2.4	
3.4	3.5
2.6	3.1
2	1.4
3.1	3.4
2.5	2.5



Format Trendline

Trendline Options | Line Color | Line Style | Shadow | Glow and Soft Edges

Trendline Options

Trend/Regression Type

○ Exponential
● Linear
○ Logarithmic
○ Polynomial Order: 2
○ Power
○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Series1)
○ Custom:

Forecast

Forward: 0.0 periods
Backward: 0.0 periods

□ Set Intercept = 0.0
□ Display Equation on chart
□ Display R-squared value on chart

Close

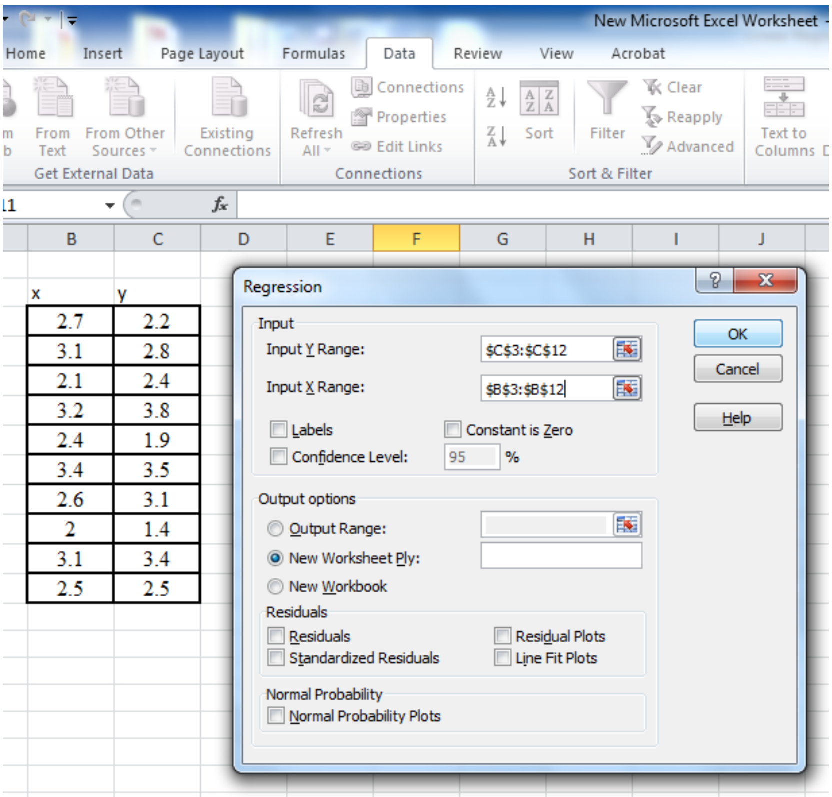
Step 2: You can highlight the scatter diagram under "Chart Tools", choose the tab "Layout" and select "Trendline" → "Linear Trendline". Under "Trendline Options" you may check the box "display equation on chart" to obtain the equation of the linear regression line.

A2.1.2 Generate Summary Output

- Using Example 2.4-1

  Step 1: Ensure you have the Analysis TookPak installed in Excel (Chapter 1). Under “Data” tab → “Data Analysis” → Regression.

  Step 2: Highlight the data cells for the X and Y variables respectively. Click “OK” and the summary output will be generated.



x	y
2.7	2.2
3.1	2.8
2.1	2.4
3.2	3.8
2.4	1.9
3.4	3.5
2.6	3.1
2	1.4
3.1	3.4
2.5	2.5

Input

Input Y Range: $C$3:$C$12

Input X Range: $B$3:$B$12

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

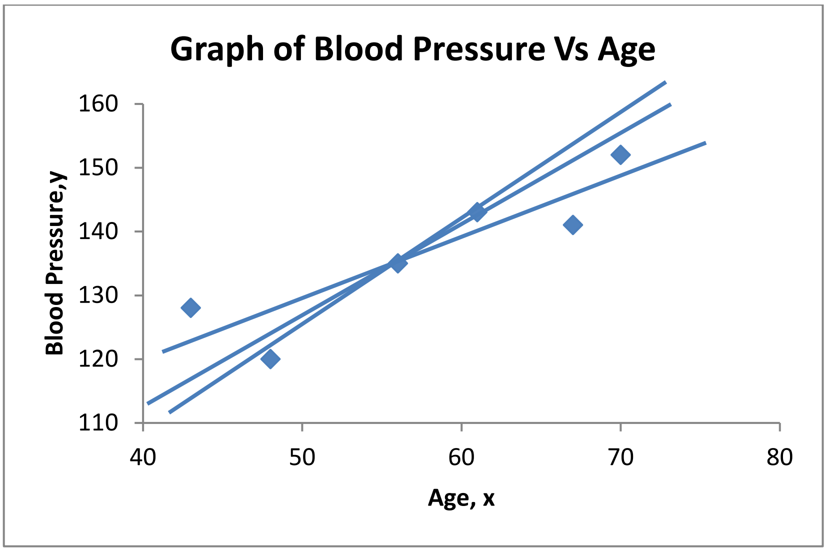
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

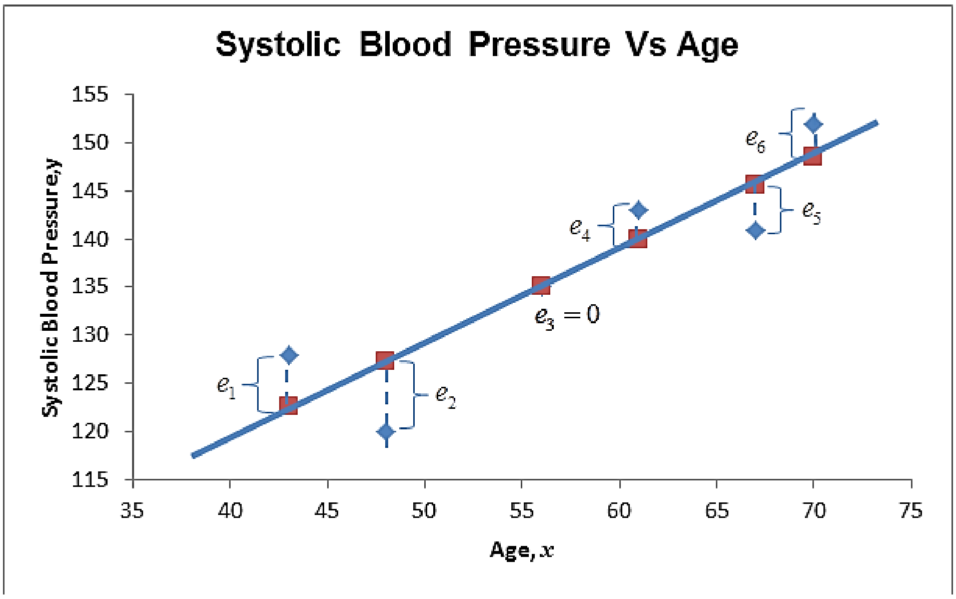
Appendix 2.2 Principle of Least Squared Error

- Given a scatter diagram, there can be a few possible straight lines to model the linear pattern. To avoid any ambiguity, statisticians adopt the "best fit" line that meets the criteria of having the least squared error value.



- The "dots" on the scatter diagram represent the **actual** observed values of the variables, while the values on the best fit line are just **estimations**.

- The errors "$e_1,\ldots,e_n$ " represent the difference between the actual and estimated values (error). The line that has the least value of $\sum_{r=1}^{n} e_r^2$ is the best fit line.

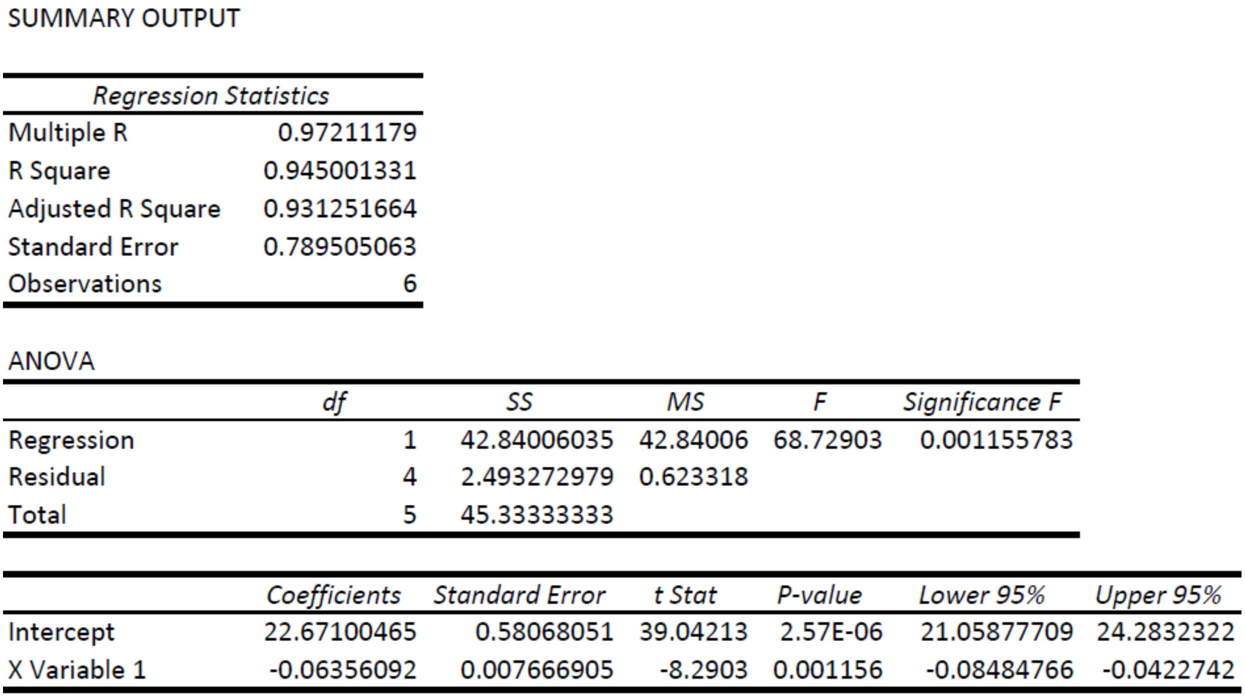


Tutorial 2: Linear Regression and Correlation

A Self Practice Questions

1 Each table gives the summary output of the linear regression analysis of y on x. Write down the correlation coefficient and comment on the relationship between x and y.

(a)



SUMMARY OUTPUT

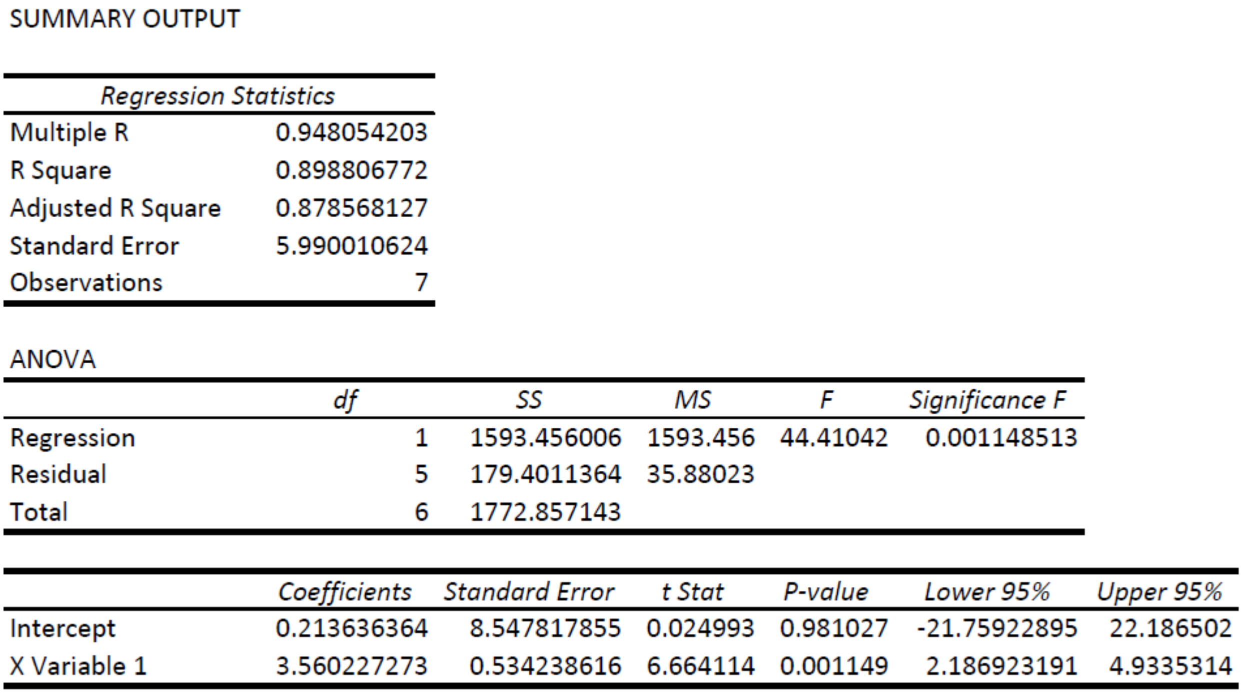
Regression Statistics	
Multiple R	0.97211179
R Square	0.945001331
Adjusted R Square	0.931251664
Standard Error	0.789505063
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	42.84006035	42.84006	68.72903	0.001155783
Residual	4	2.493272979	0.623318		
Total	5	45.33333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22.67100465	0.58068051	39.04213	2.57E-06	21.05877709	24.2832322
X Variable 1	-0.06356092	0.007666905	-8.2903	0.001156	-0.08484766	-0.0422742

(b)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.948054203
R Square	0.898806772
Adjusted R Square	0.878568127
Standard Error	5.990010624
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1593.456006	1593.456	44.41042	0.001148513
Residual	5	179.4011364	35.88023		
Total	6	1772.857143			

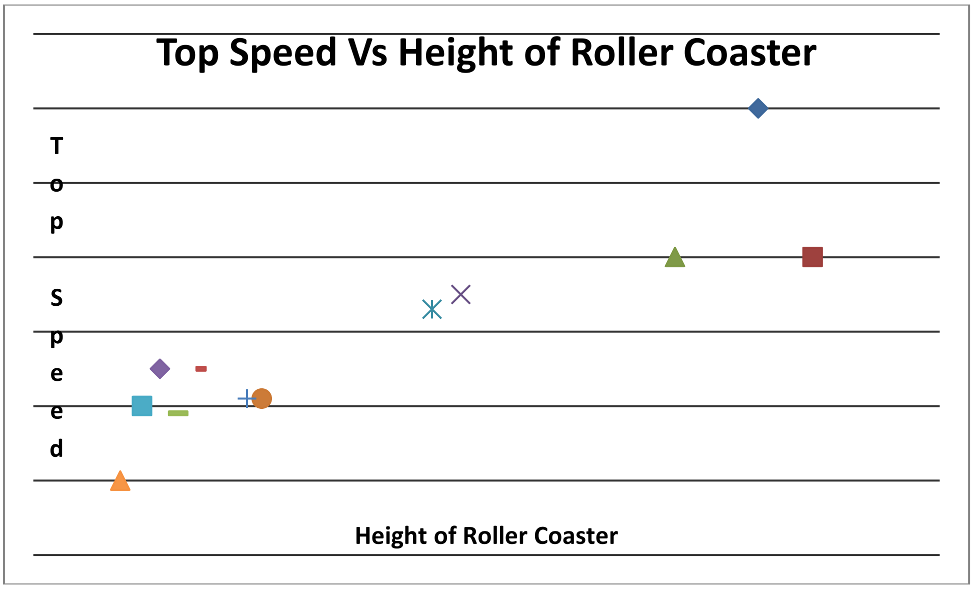
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.213636364	8.547817855	0.024993	0.981027	-21.75922895	22.186502
X Variable 1	3.560227273	0.534238616	6.664114	0.001149	2.186923191	4.9335314

2 Write down the equation of the regression line in the form $y = mx + c$ for Question 1 (a). Explain what do the values of m and c represent.

3 Using your answer to Question 2, estimate the value of y when $x = 20$.

B Discussion Questions

1 A survey is conducted on the relationship between the maximum height in feet of the roller coasters and their top speeds in miles per hour. The scatter diagram and the Excel summary output of the data are given below:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89321
R Square	0.79782
Adjusted R Square	0.77760
Standard Error	6.29004
Observations	12

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	39.06121	8.16739	4.7826	0.0007	20.8631	57.2593
X Variable 1	0.170724	0.02718	6.2818	9.1E05	0.11017	0.23128

(i) State the correlation coefficient and comment on the relationship between the height of roller coaster (x) and the top speed (y).

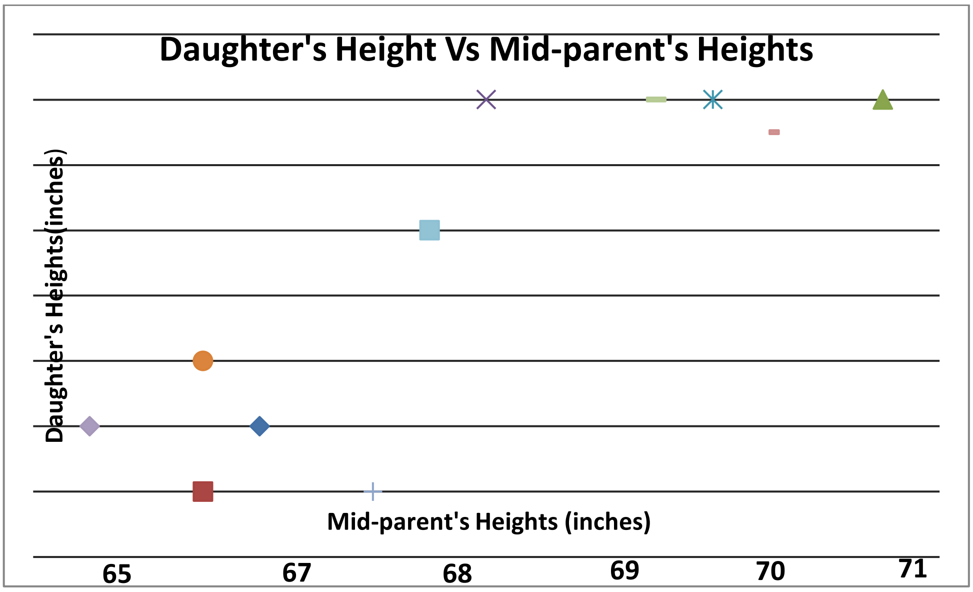
(ii) Find the equation of the line of best fit.

(iii) What is the predicted top speed for a new roller coaster of height 325 feet?

(iv) What must be the height of a new roller coaster if it is designed to go at a top speed of 90 miles per hour?

2 A study is carried out to investigate the relationship between the mid-parent's height and the daughter's height. Mid-parent's height is the average of father's and mother's heights. The heights of eleven female students and their mid-parent's heights in inches were collected.

The scatter diagram and the Excel summary output of the data are given below:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8504
R Square	0.7232
Adjusted R Square	0.6924
Standard Error	1.4506
Observations	11.000

	Coefficients	*Standard Error*	*t Stat*
Intercept	1.6497	13.363	0.1235
X Variable 1	0.9555	0.1971	4.8487

(a) State the correlation coefficient and comment on the relationship between mid-parent's height (x) and daughter's height (y).

(b) Find the equation of the line of best fit, $y = a + bx$.

(c) Predict the daughter's height if the mid-parent's height is 69 inches.

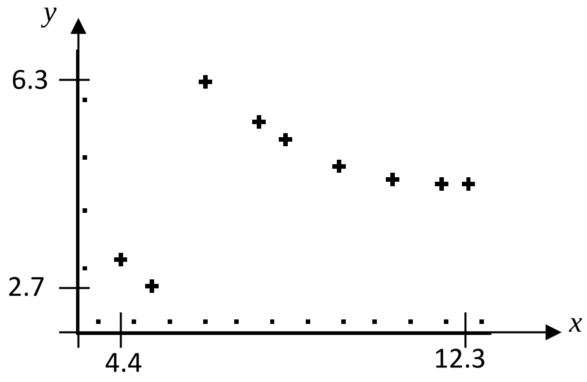
(d) Briefly state the physical significance of the coefficients a and b.

(e) Comment on the reliability of the estimate of daughter’s height when the mid – parent’s height is 73 inches.

3 In an experiment involving two chemicals x and y, a researcher recorded observations of values of y for controlled values of x and the summary output and scatter diagram are shown below:.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.26334686
R Square	0.06935157
Adjusted R Square	-0.06359821
Standard Error	2.92006982
Observations	9



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4.447901248	4.447901	0.521637	0.49355973
Residual	7	59.68765431	8.526808		
Total	8	64.13555556			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.30387433	4.365926621	1.214834	0.263813	-5.019901634	15.6276503
X Variable 1	0.66662966	0.922997026	0.722245	0.49356	-1.51591149	2.84917081

(a) Explain whether a linear model is appropriate.

(b) The researcher realised that some of the observations came from contaminated materials. He then considered only the seven pairs of observations for which the values of x exceeded 6 and discarded the other observations.

(i) On the scatter diagram, circle the data points that are to be removed.

(ii) The researcher proposed two models for the remaining seven pairs of data:

Model A: $y = ax^2 + b$, correlation coefficient, $r = -0.912961$

Model B: $y = a \ln x + b$, correlation coefficient, $r = -0.970794$

where a and b are constants

State which model is a better choice, giving a reason for your choice.

(iii) Hence using the better model with $a = -2.69, b = 11.0$, estimate the value of x when $y = 6.1$. For this model, comment on the validity of this estimated value.

<u>Answers</u>

A1 (a) $r = -0.972$ (b) $r = 0.948$

A2 $y = -0.0636x + 22.7$

A3 estimated value of $y = 21.4$

B1 (i) $r = 0.893$ (ii) $y = 0.171x + 39.1$ (iii) 94.5 mph

(iv) 298 feet

B2 (a) $r = 0.850$ (b) $y = 0.956x + 1.65$ (c) 67.6 inches

B3 (bi) $(4.4, 3.2)$, $(5.1, 2.7)$ (bii) Model B (biii) estimated $x = 6.18$