Correlation Analysis and Linear Regression

Learning Outcomes

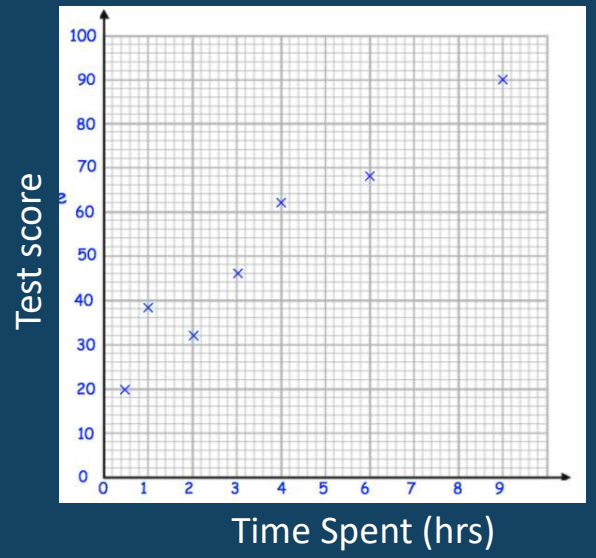
Learning Outcomes

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

1. Identify the types of relationship between independent and dependent variables using a scatter diagram.
2. Describe the relationship between two sets of variables using Correlation Coefficient
3. Interpret the regression equation with reference to the slope and the intercept of the regression equation.
4. Evaluate regression equation's ability to predict using standard error of estimate and coefficient of determination
5. Solve real-life business problems by applying regression and correlation analysis.

Introduction to Correlation Analysis

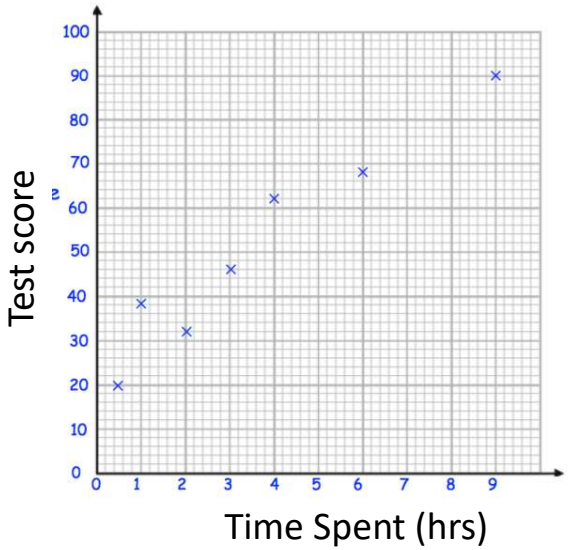
- You had always wondered "Will more time spent on revision results in higher score in a test?"
- So, you collected the test scores of 10 of your classmates on their spent on revising for Business Statistics test and their corresponding test scores.
- You plot them on a graph as shown and noticed that the more time spent on revision, the higher the test score.
- Based on your observation, you think the relationship between the time spent on revision and test score is quite strong.



- In this lesson, you will be apply a group of techniques called **Correlation Analysis** to measure the relationship between two variables, such as the Time spent on revision and Test score in the example above.

Scatter Diagram

- The basic idea of correlation analysis is to measure the relationship between 2 variables.
- The usual first step is to plot the data in a **Scatter Diagram**, such as the one shown on the right.
- Scatter diagrams serve two purposes:
  - ✓ Provides visual information whether variables X and Y share any special relationship.
  - ✓ Helps to determine the type of equation to use to describe the relationship.

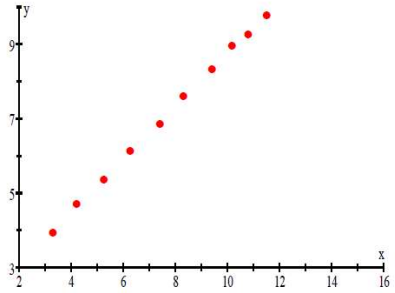


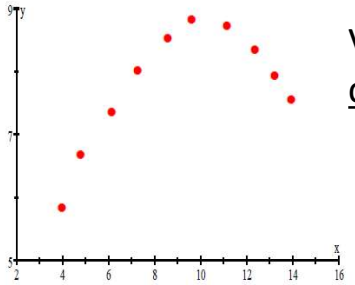
- On the Scatter Diagram shown,
  - Time Spent is referred to as **independent variable**. It provides the basis for estimating the dependent variable.
  - Test Score is referred to as the **dependent variable.** It is the variable that is being estimated.

Scatter Diagram

- Shown on the right are typical relationships which Scatter diagrams can provide visually:

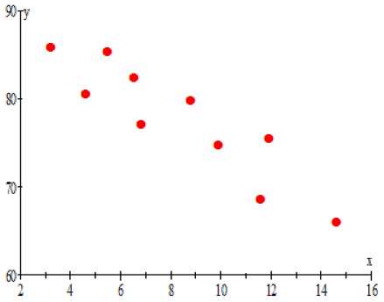
Variables X and Y have a positive relationship if X increases, Y increases (i.e. there is a upward trend).

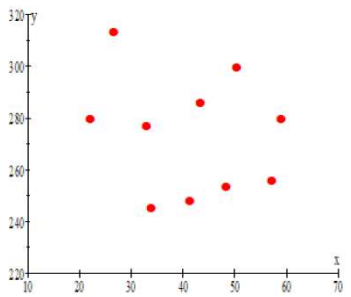




Variables X and Y have a curvilinear relationship

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





Variables X and Y no obvious relationship

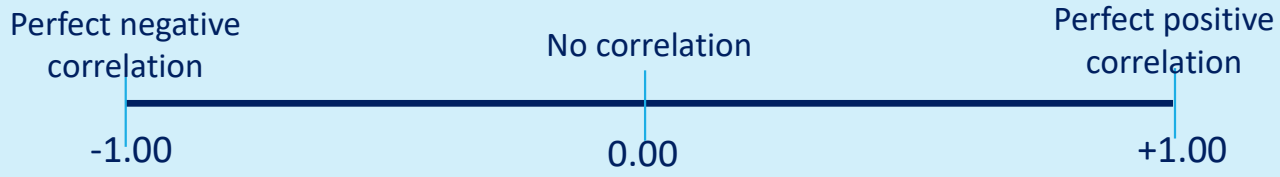
Correlation Coefficient

- It is difficult to use our eyes to determine through the scatter diagram whether there is indeed a linear relationship between the variables.
- The Correlation Coefficient describes the strength of relationship between 2 sets of variables.

Correlation Coefficient

A measure of the strength of the linear relationship between two variables

- ✓ Designated by letter r
- ✓ Shows the direction and strength of the linear relationship between two interval variables



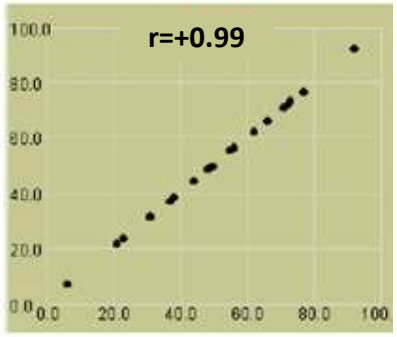
Correlation Coefficient

- The table below shows how the correlation coefficient, r indicates the linear relationship between two variables

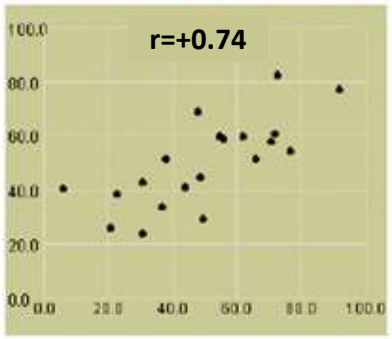
$-1 \le r \le 1$		
Strength of linear relationship	**Positive**	**Negative**
Perfect	$r = 1$	$r = -1$
Very strong	$0.8 \le r < 1$	$-1 < r \le -0.8$
Strong	$0.4 \le r < 0.8$	$-0.8 < r \le -0.4$
Weak	$0.2 \le r < 0.4$	$-0.4 < r \le -0.2$
Little / no relationship	$0 \le r < 0.2$	$-0.2 < r \le 0$

Correlation Coefficient

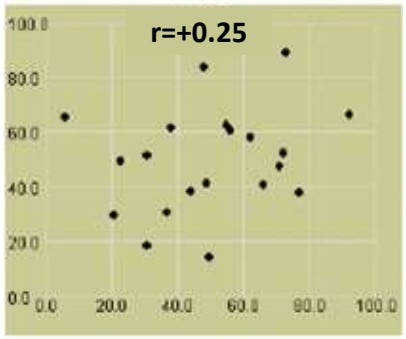
Example 1: Shape of Scatter Diagrams and Correlation Coefficient



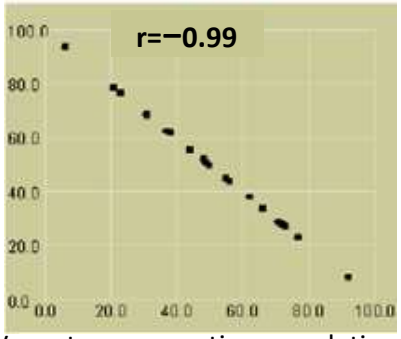
Very strong, positive correlation



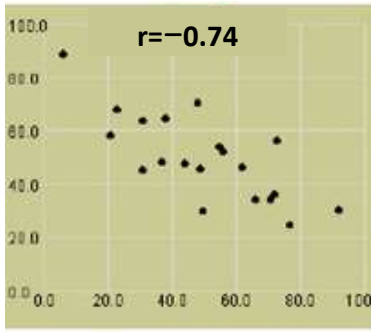
Strong, positive correlation



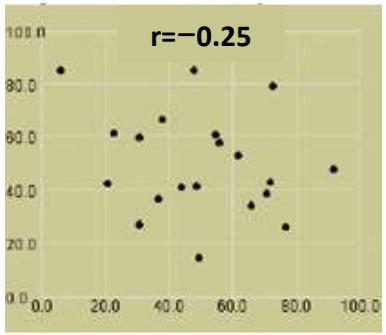
Weak, positive correlation



Very strong, negative correlation



Strong, negative correlation



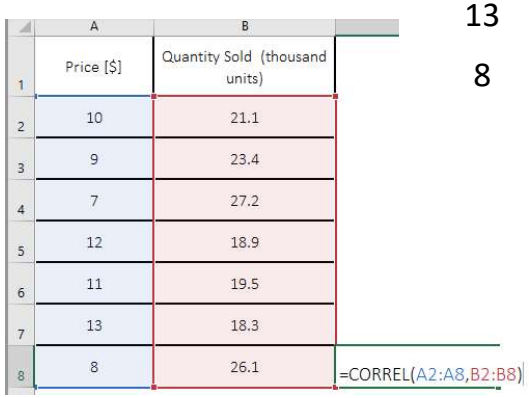
Weak negative correlation

Correlation Coefficient

Example 2: John is an entrepreneur who owns a business selling an electronic gadget. He plans to increase the price of his gadget but is worried about the falling demand when price increases. Based on the data collected, compute and interpret the Correlation Coefficient, r

Price [$]	Quantity Sold (thousand units)
10	21.1
9	23.4
7	27.2
12	18.9
11	19.5
13	18.3
8	26.1

Method 1: In Excel, type in = CORREL(array1, array2) where array1 and array2 are the 2 columns of data on price and quantity sold.



The computed Correlation Coefficient, r = −0.975. Therefore, there is a **very strong negative** linear relationship between monthly quantity sold and price of electronic gadgets.

Correlation Coefficient

Method 2: Use **Regression** function in Excel Toolpak Add-in (See setup here*).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97463804
R Square	0.94991932
Adjusted R Square	0.93990318
Standard Error	0.87325336
Observations	7

Multiple R refers to Correlation Coefficient

Positive value means Positive Relationship
Negative value means Negative Relationship

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.1428571	1.68298	22.6639	3.1E-06	33.8166	42.4691	33.8166	42.4691
X Variable 1	-1.60714286	0.16503	-9.7385	0.00019	-2.0314	-1.1829	-2.0314	-1.1829

From the summary output, Correlation Coefficient , r = -0.975. Therefore, there is a **very strong negative linear relationship** between monthly quantity sold and price of electronic gadgets.

*https://www.excel-easy.com/examples/regression.html

Correlation Coefficient

Example 3: State the value of the correlation coefficient and describe the relationship between X and Y based on the following Summary Output.

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 95.0*
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

From the summary output, Correlation Coefficient , r = +0.897. Therefore, there is a **very strong positive linear relationship** between X and Y.

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 of the form $y = mx + c$ where
  - x is independent variable
  - y is dependent variable.
  - c coefficient of Intercept
  - m is coefficient(slope) of X variable 1

	Coefficients	*Standard Error*
Intercept	81.04809	13.88088
X Variable 1	0.964381	0.238061

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

Least Squares Regression Line

- Principle of Least Square: A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual y values and the predicted values of y.
- Illustration:

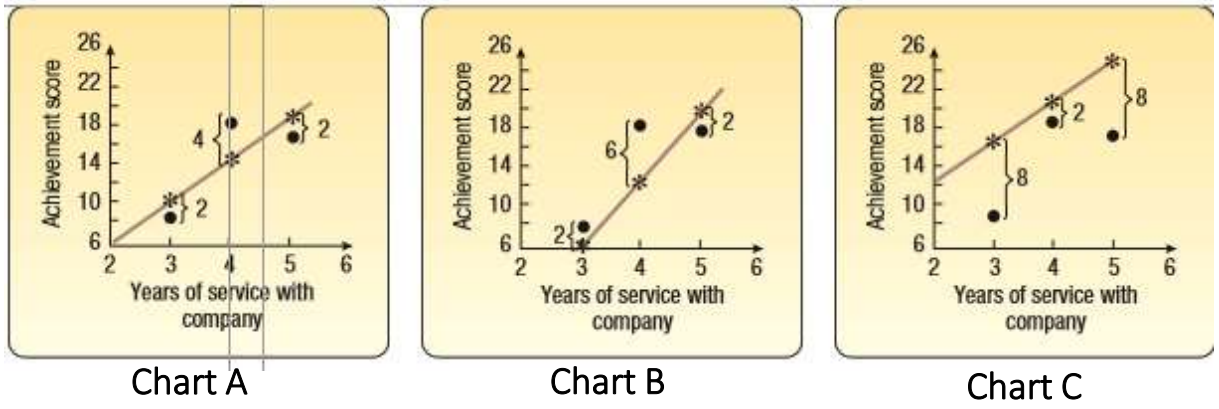


Chart A Chart B Chart C

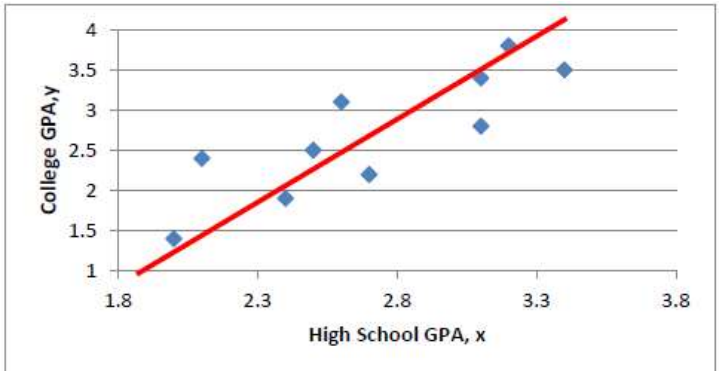
✓Chart A, drawn using the least squares method, is the best fitting line. Its sum of the squares of the vertical deviations is at a minimum with sum of the squares = 24

✓Chart B & C are drawn differently and their sum of the squares is 44 and 132 respectively.

Simple Linear Regression

Example 4: The Graph on right shows the high school GPA and the college GPA at the end of the 1st year for 10 different students.

a) Using the summary output, state the correlation coefficient and the relationship between the two variables.
b) Write the equation of the regression line Y on X.
c) Find the college GPA if the High School GPA is 3.6.
d) Find the High School GPA if the College GPA is 2.3



SUMMARY OUTPUT

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

	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

Simple Linear Regression

a) Using the summary output, state the correlation coefficient and the relationship between the two variables.

The correlation coefficient is +0.844. The two variables has a **very strong**, **positive** and **linear** relationship.

b) Write the equation of the regression line Y on X.

The linear regression line is $y = 1.347x - 0.9504$

c) Find the college GPA if the High School GPA is 3.6.

x= 3.6, $y = 1.347(3.6) - 0.9504$ = 3.8988

College GPA = 3.899

d) Find the High School GPA if the College GPA is 2.3

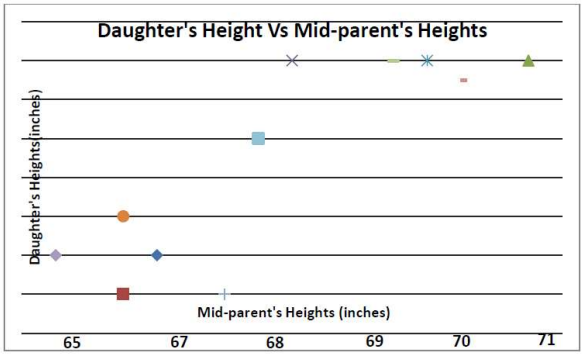
y= 2.3, $2.3 = 1.347x - 0.9504$

$x = 2.413.$

High school GPA = 2.413

Simple Linear Regression

Example 5: In a study, heights of eleven female students and their mid-parent's heights in inches were collected. The mid-parent's height is the average of father's and mother's heights.



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, $y = mx + c$.

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

d) Briefly state the physical significance of the coefficient, m.

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

Simple Linear Regression

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

The correlation coefficient is +0.85. The relationship between mid-parent's height (x) and daughter's height (y) has a **very strong**, **positive** and **linear** relationship.

b) Find the equation of the line of best fit, y=mx+c.

The line of best fit is $y = 0.956x + 1.65$

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

x= 69, $y = 0.956(69) + 1.65$ = 67.6 inches

Estimated height of daughter = 67.6 inches

d) Briefly state the physical significance of the coefficient, m.

When the mid parent's height increases by 1 inch, the daughter's height is estimated to increase by 0.956 inch.

Evaluating a Regression Equation's ability to predict

We can use 2 statistics to evaluate a Regression Equation's ability to predict:

- Standard Error of Estimate
- Coefficient of Determination

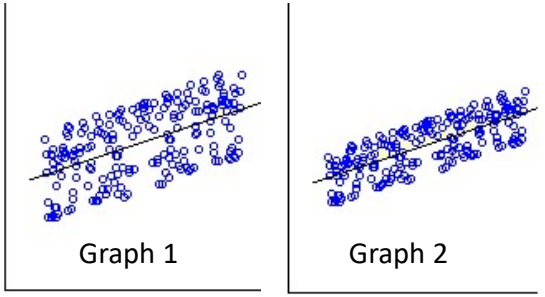
Evaluating a Regression Equation's ability to predict

- Standard Error of Estimate
  - Perfect prediction is almost impossible
  - Hence, we need a way to measure how inaccurate an estimate might be

Standard Error of Estimate, S_e

A measure of the dispersion, or scatter, of the observed value around the regression line for a given x

If S_e is large, (Graph 1)	If S_e is small, (Graph 2)
Greater will be the dispersion of points from the regression line.	Data points will be closely spaced from the regression line.
Average error of estimate from regression line will be larger.	Average error of estimate from regression line will be smaller.
Estimate based on that regression line will be less accurate.	Estimate based on that regression line will be more accurate.



Evaluating a Regression Equation's ability to predict

- Coefficient of Determination

Coefficient of Determination, r^2

A measure of the amount of variation in the dependent variable Y that is explained by the variation in the independent variable X

- r^2 must fall between two limits zero and one $0 \leq r^2 \leq 1$
- The higher the value of r^2, the higher the explanatory power of the regression line.
  - ✓ If $r^2 = 1$, variations in Y is fully explained by the variations in X, i.e. perfect correlation between X and Y.
  - ✓ If $r^2 = 0$, variations in Y is not explained by the variations in X at all, i.e. no relationship between X and Y.

Evaluating a Regression Equation's ability to predict

Example 6: With the summary output taken from Example 5,

a) State and interpret the standard error of estimate.

b) State and interpret the coefficient of determination.

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) The Standard Error of Estimate = 1.451. The average dispersion of data points from the regression line is 1.451 inches of the daughter's height.

b) The Coefficient of Determination= 0.723. 72.3% of the variation in predicted daughter's height is explained by the variation in mid height of parents.