School of Information Technology

Module	:	Business Statistics
Topic	:	Correlation Analysis and Linear Regression

Learning Outcomes:

By 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.

Topic: Correlation Analysis and Linear Regression

QUESTION 1

A survey is conducted to find the demand for a product in thousands of units, Y and its price in cents, X in six different market areas. The scatter diagram and the Excel summary output of the data are given below.

a) State the correlation coefficient and comment on the relationship between the demand for the product and price

b) Write down the equation of line of best fit which we can predict the demand for the product in terms of its price

c) Estimate the demand for the product when it is priced at 20 cents.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9721118
R Square	0.9450013
Adjusted R Square	0.9312517
Standard Error	12.074828
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10020.79	10020.79	68.72903	0.001155783
Residual	4	583.2059	145.8015		
Total	5	10604			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	340.52941	33.83741	10.0637	0.000548	246.5816981	434.4771	246.5817	434.4771
X Variable 1	-14.86765	1.793379	-8.2903	0.001156	-19.8468659	-9.88843	-19.8469	-9.88843

QUESTION 2

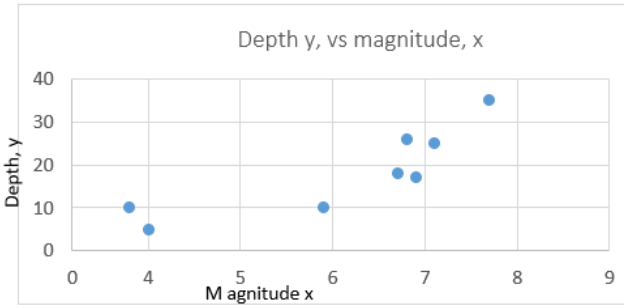
A researcher wants to determine if there is a linear relationship between the magnitudes of earthquakes and their depths below the surface at the epicentre. The scatter diagram and the Excel summary output of the data are given below.

a) State the correlation coefficient r and comment on the relationship between the magnitude, X and depth, Y.

b) Find the equation of the line of best fit.

c) Estimate the depth below the surface at the epicentre when the magnitude is 5.5.

d) Comment on the reliability of the predicted depth when the magnitude of earthquake is 11.5.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.848252943
R Square	0.719533055
Adjusted R Square	0.672788564
Standard Error	5.718196516
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	503.3133717	503.313372	15.392895	0.007771717
Residual	6	196.1866283	32.6977714		
Total	7	699.5			

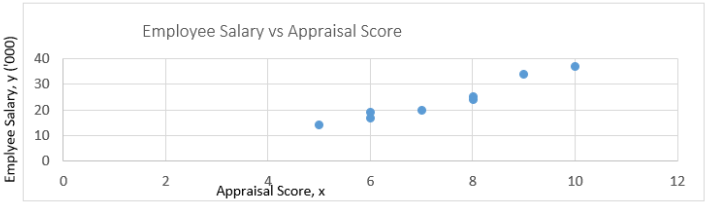
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-17.4092638	9.311050593	-1.86974215	0.1107184	-40.19258384	5.3740562	-40.192584	5.3740562
X Variable 1	5.833826388	1.486939652	3.92337805	0.0077717	2.195416132	9.4722366	2.1954161	9.4722366

QUESTION 3

A human resource management researcher is interested to know the relationship between performance appraisal scores, X and employee's salary Y (in thousands of dollars). A sample of eight employees was taken of their salary and performance appraisal scores. The scatter diagram and the Excel summary output of the data are given below.

a) State the correlation coefficient r and comment on the relationship between the performance appraisal scores, X and salary, Y.

b) Find the equation of the line of best fit.



c) State and interpret the coefficient of determination and standard error of estimate.

d) Explain if you would advise an employee to perform well in his/her job.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97054928
R Square	0.941965905
Adjusted R Square	0.932293556
Standard Error	2.108185107
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	432.8333333	432.8333333	97.3875	6.24576E-05
Residual	6	26.66666667	4.444444444		
Total	7	459.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	-10.66666667	3.566283606	-2.990975437	0.02428959	-19.3930483	-1.940285047	-19.393048
X Variable 1	4.666666667	0.472884601	9.868510526	6.2458E-05	3.509559731	5.823773602	3.5095597

SUPPLEMENTARY QUESTIONS

QUESTION 4

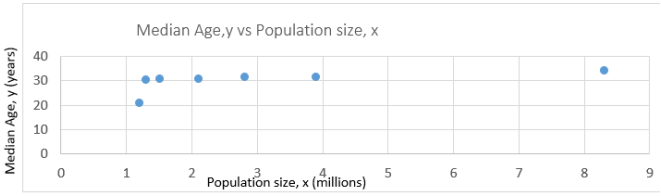
City planners collected data to examine if there is a relationship between the population size, X and the median age of resident, Y for a random sample of cities. Excel summary output of the data are given below.

a) State the correlation coefficient r and comment on the relationship between population size, X and the median age of resident, Y.

b) Find the equation of the line of best fit.

c) Estimate the median age of resident with the population size is 0.9 million.

d) State and interpret the coefficient of determination and standard error of estimate.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.979625974
R Square	0.959667048
Adjusted R Squa	0.951600458
Standard Error	0.297802618
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10.55085372	10.55085372	118.9681145	0.00011256
Residual	5	0.443431997	0.088686399		
Total	6	10.99428571			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95*
Intercept	29.94293743	0.183856398	162.8604595	1.65596E-10	29.4703195	30.41555535	29.4703
X Variable 1	0.526039715	0.048228443	10.90725055	0.000112561	0.40206455	0.650014876	0.40206

Answers:

Q1 a) −0.972

b) y = −14.868x + 340.529

c) 43.169 ('000)

Q2 a) +0.848

b) y = 5.834x – 17.409

c) 14.678

Q3 a) +0.971

b) y = 4.667x – 10.667

c) Standard error of estimate = 2.108

Coefficient of determination = 0.942

Q4 a) +0.9796

b) y = 0.526x +29.943

c) Standard error of estimate = 0.298

Coefficient of determination = 0.95966