GE2262 Business Statistics
Topic 8: Simple Linear Regression Exercises

Q1
Suppose that Y = 67+27X for all values of X and Y. What can you say about the sample correlation coefficient r between X and Y?

a) r = -1
b) r = 0
c) r = 1
d) The magnitude of r is unknown until it is estimated using sample data

Q2
Suppose that r = 0, where r is the sample correlation coefficient between X and Y. Then

a) X and Y are not related at all
b) X and Y are very closely related
c) Neither of the above is necessarily true

Q3
You want to predict expenditure on hamburger purchases Y using one of the income measures X or Z. If $r_{XY} = 0.4$, $r_{ZY} = 0.3$ and $r_{XZ} = 0.6$, then we should use

a) X
b) Z
c) Not use this information

Q4
You want to predict the consumption sauce W using one of the income measures X or Z. Given the information $r_{WX} = 0.6$, $r_{XZ} = 0.9$ and $r_{ZW} = -0.8$, you should use

a) X
b) Z
c) Not use this information

Q5
Suppose we find that r_{XY} is very close to -1 for a sample. That means

a) X and Y have very weak correlation
b) The X and Y values lie very close to a straight line with slope that equals -1
c) Neither of the above

Q6

A fuel-oil distribution company has collected data over a number of years in order to determine the statistical relationship between the daily temperature and the consumption of fuel-oil in single family dwellings. Given the temperature, the company would like to be able to predict the consumption of fuel-oil in order to service their customers better. The company draws sample observations which yield the following information:

Consumption of fuel-oil (litres)	Daily temperature (in Centigrade)
7.0	-6
6.3	-3
5.1	0
4.6	3
3.4	6
2.9	9
1.3	12
1.0	15
0.6	18

a) Plot the data on a scatter diagram.

b) Estimate the linear regression equation of daily fuel consumption on daily temperature. Superimpose the estimated regression line on the graph constructed in a).

c) Relate the values of the estimated regression coefficients to your diagram.

d) Predict the level of consumption of fuel when the daily temperature is 7.5c.

e) Measure the strength of the linear relationship between the daily temperature and daily fuel consumption in terms of the sample correlation coefficient. Interpret this value of the coefficient which you obtain.

Q7

You want to develop a model to predict assessed value of homes based on gross area. A sample of 15 single-family apartments is selected in the Midwestern district. The assessed value (Y, in hundred thousands of dollars) and the gross area of the apartments (X, in thousands of square feet) are recorded, with the following results of statistical analysis:

$$\hat{Y} = 51.915 + 16.633X$$

$$r = 0.812$$

a) Interpret the meaning of the Y-intercept and the slope of the regression model.
b) Predict the assessed value for an apartment whose gross area is 1750 square feet.
c) Interpret the meaning of the coefficient of determination in this problem.

Q8

Circulation is the lifeblood of the publishing business. The larger the sales of magazine, the more it can charge advertisers. Recently, a circulation gap has appeared between the publishers' reports of magazines' newsstand sales and subsequent audits by the Audit Bureau of Circulations. The data in the file Circulation represent the reported and audited newsstand yearly sales (in thousands) for the following 10 magazines (i.e. n = 10)

Magazine	Reported (X) ←independent variable	Audited (Y) ←dependent variable
YM	621.0	299.6
CosmoGirl	359.7	207.7
Rosie	530.0	325.0
Playboy	492.1	336.3
Esquire	70.5	48.6
TeenPeople	567.0	400.3
More	125.5	91.2
Spin	50.6	39.1
Vogue	353.3	268.6
Elle	263.6	214.3

Source: *Data extracted from M. Rose, "In Fight for Ads, Publishers Often Overstate Their Sales,"* The Wall Street Journal, *August 6, 2003, pp. A1, A10.*

For these data, $b_0 = 26.724$, $b_1 = 0.5719$ and $S_{b1} = 0.0668$

a) Interpret the meaning of the slope, b_1, in this problem.
b) Predict the audited newsstand sales for a magazine that reports newsstand sales of 300,000
c) At the 0.05 level of significance, is there evidence of a linear relationship between reported sales and audited sales?
d) Construct a 95% confidence interval estimate of the population slope β_1.

Q9

An agent for a residential real estate company in a large city would like to be able to predict the monthly rental cost for apartments, based on the size of the apartment, as defined by square footage. A sample of 25 apartments in a particular residential neighborhood was selected, and the information gathered revealed the following:

Rent	Size	Rent	Size
950	850	1800	1369
1600	1450	1400	1175
1200	1085	1450	1225
1500	1232	1100	1245
950	718	1700	1259
1700	1485	1200	1150
1650	1136	1150	896
935	726	1600	1361
875	700	1650	1040
1150	956	1200	755
1400	1100	800	1000
1650	1285	1750	1200
2300	1985		

a) Use the least-squares method to find the regression coefficients b_0 and b_1.
b) Interpret the meaning of b_0 and b_1 in this problem.
c) Predict the monthly rent for an apartment that has 1,000 square feet.
d) Why would it not be appropriate to use the model to predict the monthly rent for apartments that have 500 square feet?
e) At the 0.05 level of significance, is there evidence of a linear relationship between the size of the apartment and the monthly rent?
f) Construct a 95% confidence interval estimate of the population slope, β_1.

Q10

A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below:

Company	Sales	Earnings
C1	40.2	5.3
C2	10.4	3.7
C3	18.6	4.4
C4	71.7	8.0
C5	58.6	6.6
C6	46.8	5.1
C7	17.5	2.6
C8	11.9	1.7
C9	19.6	3.5
C10	51.2	8.2
C11	28.6	6.0
C12	69.2	10.8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.887504192**
R Square	**0.787663691**
Adjusted R Square	**0.76643006**
Standard Error	**1.258490697**
Observations	**12**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**58.75117833**	**58.7511783**	**37.0951014**	**0.000117143**
Residual	**10**	**15.83798834**	**1.58379883**		
Total	**11**	**74.58916667**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**1.64574594**	**0.728503942**	**2.259076233**	**0.047442872**	**0.022538011**	**3.26895387**	**0.022538011**	**3.268953869**
X Variable 1	**0.103873619**	**0.017054814**	**6.090574801**	**0.000117143**	**0.065873126**	**0.14187411**	**0.065873126**	**0.141874111**

a) Determine a regression equation
$\hat{Y} = a + bx$
by the least-squares principle so that we can predict the value of earnings based on the value of sales. Interpret the meaning of a and b in this problem.

b) Find the sample coefficient of correlation between sales and earnings. Is there evidence that there exists a positive linear relationship between the two variables? Use the 0.05 level of significance.

c) What is the coefficient of determination? Interpret the meaning of this coefficient.

d) For a company with $35.0 million is sales, predict the earnings.

e) Corresponding to the regression equation in part (a), the regression model is given by
$Y = \alpha + \beta x + \varepsilon$
where $\varepsilon \sim N(0, \sigma^2)$.
Test H_0: $\beta \geq 0.15$ against H_1: $\beta < 0.15$ at the 0.05 level of significance.

Q11

The following table shows, for eight brands of tea, the average purchase quantity per buyer (Y) and the purchasing frequency (X) in a year.

Y (kg)	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
X	24	21	22	22	18	13	9	6

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.647613**
R Square	**0.419402**
Adjusted R Square	**0.322636**
Standard Error	**0.396999**
Observations	**8**

ANOVA

	df	SS	MS	F	Significance F
Regression	**1**	**0.683101**	**0.683101**	**4.334174**	**0.082521**
Residual	**6**	**0.945649**	**0.157608**		
Total	**7**	**1.62875**			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	**2.028994**	**0.40167**	**5.0514**	**0.00233**	**1.046144**	**3.011844**	**1.046144**	**3.011844**
X Variable 1	**0.04643**	**0.022302**	**2.081868**	**0.082521**	**-0.008141**	**0.101001**	**-0.00814**	**0.101001**

a) Estimate the regression of average purchase quantity per buyer on purchase frequency.
b) Interpret the slope of the estimated regression line.
c) Estimate the average purchase per buyer for the purchasing frequency 20.
d) Test the hypothesis that the slope of the population regression line is zero against that the slope is not zero. Use a 5% level of significance.
e) Find and interpret the coefficient of determination. Comment on the result.

Q12

In a simple linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, the sample information are obtained as follows:

X_i	2	2	4	4	6	6
Y_i	10	12	6	8	1	5

After fitting the above regression model, the following output is obtained from Excel:

	Coefficients	**Standard Error**	**t Stat**	**P-value**
Intercept	**15**	**1.870828693**	**8.0178373**	**0.0013127**
X Variable 1	**-2**	**0.433012702**	**-4.618802**	**0.00989**

a) Write down the prediction equation for Y. Find the predicted values ($\hat{Y}_i$) when X_i = 2, 4 and 6.

b) Use the observed values (Y_i) and predicted values ($\hat{Y}_i$) compute the SSE for this regression model.

c) Find $\overline{Y}$ and hence calculate the SST for this regression model.

d) Make use of the results obtained in parts (b) and (c), find and interpret the coefficient of determination (R^2).

e) Based on your result obtained in part (d), or otherwise, find the coefficient of correlation.

f) According to the p-value(s) from the Excel output, is there any evidence showing that variable X is an important factor affecting Y?

Q13

A retails company conducted a study on the cost of opening stores in 2004. the information about number of stores (NUMSTORE), total store area (in square feet) (STORESIZE), and corresponding total cost in (USD) to set up these stores (COST) is collected from 14 areas. The administrator of the company wants to develop an equation that is helpful in pricing the opening of new stores. With the data available, two regression models are constructed with COST as the dependent variable. The Excel regression results are given below. Use the outputs to answer the following questions ($\alpha = 0.05$).

Regression output using COST and NUMSTORES

Regression Statistics	
Multiple R	0.9419
R Square	0.8872
Adjusted R Square	0.8778
Standard Error	4307
Observations	14

	Coefficients	Standard Error	t Stat	P-value
Intercept	16593.65	2687.05	6.18	0.0000476
NUMSTORE	2600.68	267.66	9.72	0.0000005

Regression output using COST and STORESIZE

Regression Statistics	
Multiple R	0.6488
R Square	0.4209
Adjusted R Square	0.3727
Standard Error	9760
Observations	14

	Coefficients	Standard Error	t Stat	P-value
Intercept	10961.14	10232.92	1.07	0.31
STORESIZE	21.76	7.37	2.95	0.01

a) What is the sample linear regression equation relating COST and NUMSTORE?

b) Suppose the population linear regression equation relating COST and STORESIZE is $COST = \beta_0 + \beta_1 STORESIZE + \varepsilon$, find the point estimate and 95% interval estimate for β_1.

c) A claim is made that each new store adds at least 3000USD of cost. Can you find any evidence to support this claim?

d) Which variable, NUMSTOTE or STORESIZE, is more useful to explain the variations in store setup cost? Why?

e) Based on the regression results, describe the relationship between number of stores in an area and the corresponding setup cost.

f) Based on the regression results, predict the total setup cost for a business area that needs 1000 square feet of store area.

Q14

In planning for an orientation gathering with new Management Science (MS) major students, the Head of the Department wants to emphasize the importance of doing well in the major courses in order to get better-paying jobs after graduation. To support this point, the Head plans to show that there is a strong positive correlation between starting salaries (SALARY) for recent MS graduates and their grade-point averages (GPA) in the major courses. Records for seven of last year's MS graduates are selected at random and given in the table. The Excel regression results are given below.

GPA in Major Courses	Starting Salary (thousands of dollars)
2.58	11.5
3.27	13.8
3.85	14.5
3.50	14.2
3.33	13.5
2.89	11.6
2.23	10.6

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.966913851
R Square	0.934922395
Adjusted R Square	0.921906874
Standard Error	0.431547809
Observations	7

	Coefficients	Standard Error	t Stat	P-value
Intercept	4.553802733	0.988203581	4.608162549	0.005797239
GPA in Major Courses	2.670825906	0.315129149	8.475337535	0.000375688

a) What is the sample linear regression equation relating SALARY and GPA? Interpret the intercept and slope of the sample linear regression equation.

b) What is the estimated starting salary for MS graduates with GPA 4.0?

c) Test whether there is a positive linear relationship between SALARY and GPA with level of significance 0.01.

d) The personal secretary of the Head of Department has conducted another study on the same group of graduated to investigate the relationship between SALARY and their IELTS results (IELTS) with the coefficient of correlation 0.92. Which variable, GPA or IELTS, is more useful to explain the variations in starting salary of MS graduates? Explain.

Q15

In a manufacturing process, the assembly line speed, X_i (feet per minute) was thought to affect the number of defective parts found, Y_i during the inspection process. To test this theory, managers devised a situation in which the same batch of parts was inspected visually at a variety of line speeds. The following tables list the collected data and the Excel output.

Line Speed, X_i	Number of Defective Parts Found, Y_i
20	220
20	200
40	160
60	130
40	180

SUMMARY OUTPUT

	Coefficients	Standard Error	t Stat	P-value
Intercept	250	13.4519	18.5848	0.0004
Line Speed	-2	0.3450	-5.7966	0.0103

a) Develop the estimated regression equation that relates line speed to the number of defective parts found. Interpret the intercept and slope of the estimated regression equation?

b) What is the estimated number of defective parts found with line speed 55 feet per minute?

c) At a 0.05 level of significance, determine whether line speed and number of defective parts found are negatively related.

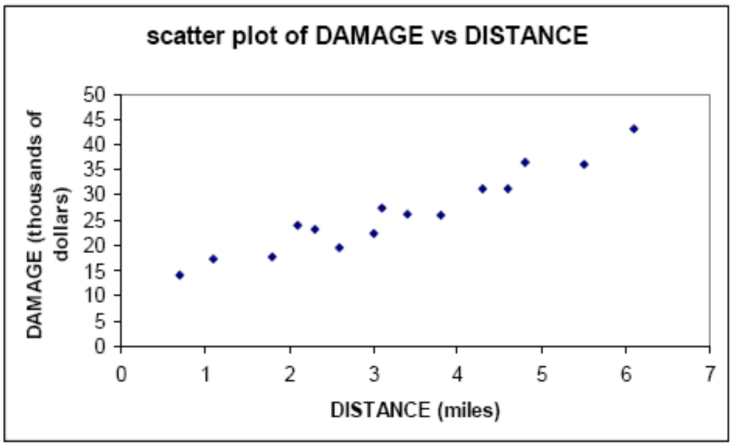
d) Find the predicted values, $\hat{Y}_i$ when X_i = 20, 40 and 60. Compute the SSE for this regression model using the observed values, Y_i and predicted values, $\hat{Y}_i$. Find $\overline{Y}$ and hence calculate the SST for this regression model.

e) From the results in (d), find the coefficient of determination. Compute the percentage of the total variation unexplained by the estimated regression equation.

f) Determine and interpret the correlation coefficient.

Q16

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage, y, and the distance between the fire and the nearest fire station, x, are recorded for each fire. The results are given as below ($\alpha = 0.05$).



SUMMARY OUTPUT (part)

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1		841.766358	156.88616	1.2478E-08
Residual	13	69.75097535	5.365459643		
Total	14	911.5173333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	10.27792855	1.420277811	7.236562082	6.5856E-06	7.20960489	13.34625221
X Variable 1		0.392747749	12.52542054	1.2478E-08	4.070850801	5.767810653

a) According to the scatter diagram, describe the relationship between X and Y.
b) Find the regression equation.
c) Interpret the slope of the regression equation.
d) Is there sufficient evidence to show that X and Y are linearly correlated? Test this hypothesis using p-value approach and use $\alpha = 0.05$.
e) Find coefficient of determination and interpret the result.
f) Can we use the above regression equation to do prediction? Why?
g) Find correlation coefficient.
h) Find the 95% confidence interval of β_1.