10

CORRELATION AND REGRESSION

QUESTION 1.

Complete the following sentence:

The purpose of the correlation coefficient is to measure the ________ of the ________ relationship between two ________.

QUESTION 2.

The following table shows the blood pressures and weights of 10 patients and you have been asked to find out if there is a correlation between weight and blood pressure. This information could be used by a doctor in order to offer dietary advice to the patients at risk of developing chronic illnesses. You are only expected to perform the statistical analysis and make the medical specialists aware of the limitations of the study.

Patient	Pressure	Weight
1	145	210
2	155	245
3	160	260
4	155	230

(Continued)

(Continued)

Patient	Pressure	Weight
5	130	175
6	140	185
7	135	230
8	165	249
9	150	200
10	130	190

The Pearson correlation coefficient is ________.

In the box below, explain your results using language that the medical specialists will be able to understand:

Is the following statement true or false?

The analysis shows that if your weight is within expected levels you will not get high blood pressure. True/False.

QUESTION 3.

Complete the following sentence:

The coefficient of determination measures the ________ of the ________ relationship between the ________ and the ________.

The table below shows the heights of 12 fathers and their sons. Your task is to find out if there is a correlation between the heights of the fathers and the heights of their respective sons.

Father's height	1.65	1.6	1.7	1.63	1.73	1.57	1.78	1.68	1.73	1.7	1.75	1.8
Son's height	1.73	1.68	1.73	1.65	1.75	1.68	1.73	1.65	1.8	1.7	1.73	1.78

The Pearson correlation coefficient is ________.

The coefficient of determination is ________.

In the box below explain what the coefficient of determination tells us:

In the box below, give an interpretation of your results.

QUESTION 4.

The table below shows the results of a mathematics exam.

Student	Hours studied	Exam result
1	10	78
2	15	83
3	8	75
4	7	77
5	13	80
6	15	85
7	20	95
8	10	83
9	5	85
10	5	68

Using the equation $\hat{y} = b_0 + b_1 x$ to model the data; answer the following questions:

The value of b_0 is ________.

The value of b_1 is ________.

The value b_1 tells us that for every ______________ of study the exam result increased by ________.

The y intercept is the result a student could expect if they spend ________ hours studying. This prediction is unreliable since it is ________ the range of exam results studied.

QUESTION 5.

The table below shows the heights of 12 fathers and their sons (this is the same table as in question 3).

Father's height	1.65	1.6	1.7	1.63	1.73	1.57	1.78	1.68	1.73	1.7	1.75	1.8
Son's height	1.73	1.68	1.73	1.65	1.75	1.68	1.73	1.65	1.8	1.7	1.73	1.78

Complete the following:
At $t_{0.975}$ = ________ for ________ degrees of freedom, the 95% confidence limits for the mean heights of sons whose fathers' heights are 1.65 are ________ ± ________.

This tells us we can be about ________ confident the ________ height of all sons whose father's heights are ________ will be between ________ and ________.

QUESTION 6.

You have been asked to predict the number of sales of walking boots for a company that specialises in outdoor hiking equipment. You have been given the following regression equation developed by your predecessor which relates sales to inventory investment and advertising expenditure:

$$\hat{y} = 25 + 10x_1 + 8x_2$$

where

x_1 = inventory investment (£1000s)

x_2 = advertising expenditure (£1000s)

$\hat{y}$ = sales (£1000s)

The coefficient b_1 means for every _________ increase in _________, sales could rise by _________.

The sales resulting from a £15,000 investment in inventory and a £10,000 investment in advertising are _________.

QUESTION 7.

The multiple regression model depends upon three assumptions. These assumptions are:

1. ____
2. ____
3. ____

State the null and alternative hypotheses for the F test.

$H_0: \beta_1 = \beta_2 \ldots = \beta_3 = 0$

H_1: at least one of the _______ does not equal _______.

QUESTION 8.

The residual for the ith observation is the difference between the _______ value of the _______ variable and the _______ value of the dependent variable.

There are three procedures to test whether a residual exerts a lot of influence on the regression equation.

Leverage measures
Cook's distance measures
Mahalanobis distance measures

QUESTION 9.

An important aspect of developing regression equations is the testing of assumptions.

A good starting point is to _______ the standardised residuals.

In a normal P-P plot (normal probability plot) the diagonal line represents a _______ and the points are the _______. The closer the _______ follow the , the closer to _______ they are.

The Kolmogorov-Smirnov test tests whether a _______ is _______ distributed.

Autocorrelation is measured by the _______ test and the output ranges from _______ to _______. A value of 2 represents _______.

In the box below, explain the following equation:

$$VIF(x_j) = \frac{1}{1 - R_j^2}$$

MINI PROJECT

You have been commissioned by a business to investigate the effectiveness of different types of advertising in the promotion of its products. You have chosen to look at radio advertising and magazine advertising. In a period of one month, you collected data from a sample of 22 cities of approximately equal populations. This data is shown in the table below.

City	Sales (£000)	Radio (£000)	Magazine (£000)
1	973	0	40
2	1119	0	40
3	875	25	25
4	625	25	25
5	910	30	30
6	971	30	30
7	931	35	35
8	1177	35	35
9	882	40	25
10	982	40	25
11	1628	45	45
12	1577	45	45
13	1044	50	0
14	914	50	0

(Continued)

(Continued)

City	Sales (£000)	Radio (£000)	Magazine (£000)
15	1329	55	25
16	1330	55	25
17	1405	60	30
18	1436	60	30
19	1521	65	35
20	1741	65	35
21	1866	70	40
22	1717	70	40

The Managing Director vaguely remembers some statistics he studied many years ago and has asked you to produce a report which gives an appropriate regression equation so that he can decide which advertising medium to invest in to boost sales. In order to save embarrassing him, you decide to include an interpretation of the gradients and the intercept value. You are also expected to inform him of the most effective advertising medium.

You know you will be in trouble if your recommendation proves to be wrong, so in order to provide evidence for your recommendation, you decide to construct a 95% confidence interval of the gradient between sales and radio advertising. Also, you decide to check at the 0.05 level of significance whether each independent variable makes a significant contribution to the regression model. You plot the data and being a good statistician (and wanting to keep your job!), you decide to check assumptions with appropriate tests and diagrams.

And finally...

If I revert to my childhood behaviours, am I simply regressing or is it a case of multiple regression?