

CORRELATION AND REGRESSION

© Nick Lee and Mike Peters 2016.

QUESTION 1.

Complete the following sentence:

The purpose of the correlation coefficient is to measure the strength of the linear relationship between two random variables.

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 0.804.

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

An increase in weight could possibly cause an increase in blood pressure.

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. False.

QUESTION 3.

Complete the following sentence:

The coefficient of determination measures the strength of the linear relationship between the dependent variable and the predictor (independent) variable.

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 0.69.

The coefficient of determination is 0.48.

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

It gives a measure of how strong the relationship is between the predictor variable and the dependent variable.

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

There is a correlation between the heights.

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 $\hat{y} = b_0 + b_1x$ to model the data, answer the following questions:

The value of b_0 is 61.23.

The value of b_1 is 1.64.

The value b_1 tells us that for every additional hour of study the exam result increased by 1.64.

The y intercept is the result a student could expect if they spend zero hours studying. This prediction is unreliable since it is outside 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 $t_\alpha = 0.95$ for 10 degrees of freedom, the 95% confidence limits for the mean heights of sons are father's height minus 0.12 and plus 0.8.

This tells us we can be about 95% confident the mean height of all sons whose father's heights are 1.65 will be between 1.53 and 2.45.

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 £ 1000 increase in inventory, sales could rise by £ 10,000.

The sales resulting from a £ 15,000 investment in inventory and a £ 10,000 investment in advertising are £ 255,000.

QUESTION 7.

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

1. The random errors are independent.
2. The random error has a normal distribution.
3. No autocorrelation.

State the null and alternative hypotheses for the F test.

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

H_1 : at least one of the parameters does not equal 0.

QUESTION 8.

The residual for the i th observation is the difference between the observed value of the independent variable and the predicted 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

The influence of an observed outcome value for a particular data point.

Cook's distance measures

The effect on the model of deleting an observation and recalculating the model.

Mahalanobis distance measures

Measures the distance between the data point and the mean of the predictors.

QUESTION 9.

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

A good starting point is to plot the standardised residuals.

In a normal P-P plot (normal probability plot) the diagonal line represents a perfect normal distribution and the points are the standardised residuals. The closer the observed residuals follow the straight line, the closer to normality they are.

The Kolmogorov-Smirnov test tests whether a variable is normally distributed.

Autocorrelation is measured by the Durbin-Watson test and the output ranges from 0 to 4. A value of 2 represents no correlation.

In the box below, explain the following equation:

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

The equation measures the collinearity of the predictors. The closer they are to 1 the more serious the problem becomes. A low value of the coefficient of determination will cause this to occur.

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?