

Simple correlation and regression analysis

In this lecture, we are interested in whether two variables are associated. The concept of association is not only very important in statistics but also in life. For example, as teachers we are always interested in factors that exist among students with good grades. A medical doctor may want to know what factors that are present in people who are resistant to certain infections. A manager may want to know what makes some people more motivated in their work. In essence, we are always linking one variable to other variables.

Example (to start with)

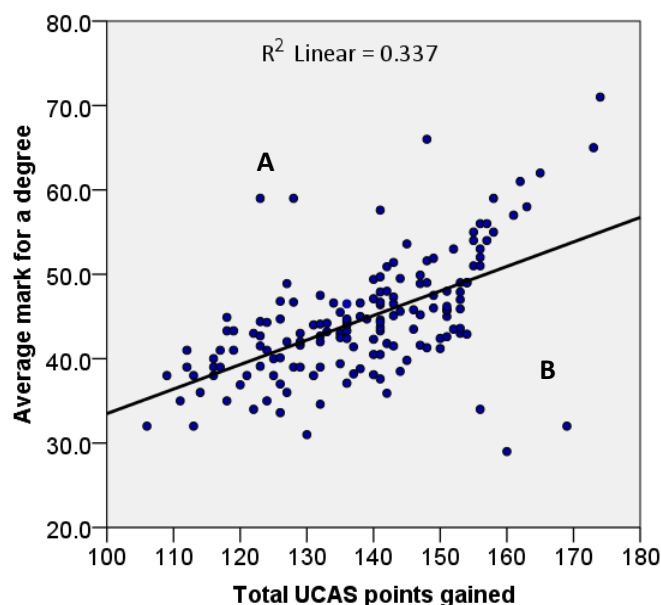
Universities recruit college students to their courses based on their UCAS points. It is said that students with higher UCAS points are more likely to be good students and therefore more likely to be successful in completing their degrees in three years.

Is this relationship between “Success in degree” and “UCAS points” generally correct? We hope so, because for years, universities have been enrolling student to their courses on that basis. Does this mean that a candidate with high UCAS points more likely to succeed? Does this mean that a candidate with low UCAS points more likely to fail? Are the UCAS points the cause for success or failure? Are there exceptions to this relationship? If so, are the exceptions to this rule so many that the relationship is simply not true.

Let us measure success at a degree by the average degree grade at the end of the three years, that is adding all the grades of the module passed at first attempt (i.e. before resit) and correlate this average to the UCAS point gained while at college. The data are for 158 students is displayed below.

File Edit View Data Transform Analyze Graphs Utilities			
20 :			
	Count	UCAS	Grade
1	1	106	32.0
2	2	109	38.0
3	3	111	35.0
4	4	112	39.0
5	5	112	41.0
6	6	113	38.0
7	7	113	32.0
8	8	114	36.0
9	9	116	39.0

1. Scatter diagram



The scatter plots show that there is a general pattern, a positive correlation between the two variables but the value of the **r-square** of **0.337** suggest that this relationship is weak. There are too many exceptions in this scatter. We call these outliers. They do not fit the norm. Point A is one of them. This student enrolled to the university with about 120 UCAS points but ended with a very high degree mark of around 60. This student has over-achieved compared to the norm (if there is such thing). Point B relates to a student who was enrolled with a large UCAS score but the average degree grade was fairly low of 30. These two points A and B are few and far between. The majority belong to a cloud of data somewhere in the middle.

Next, we look at the correlation matrix.

Correlations

		Total UCAS points gained	Average mark for a degree
Total UCAS points gained	Pearson Correlation	1	.580**
	Sig. (2-tailed)		.00000
	N	158	158
Average mark for a degree	Pearson Correlation	.580**	1
	Sig. (2-tailed)	.00000	
	N	158	158

**. Correlation is significant at the 0.01 level (2-tailed).

The **Pearson correlation coefficient** (R) is 0.580. It measures the strength of the association between the two variables. More importantly is the p-value (of 0.0000), which means that if the population correlation coefficient is "0", then there is no chance of getting a sample correlation of 0.580. In simple terms, when the p-value is less than 0.05, then we can conclude that the correlation is different from "0" (zero).

If there is a linear (straight line) relationship between the two variables, this is obtained by the SPSS output as follows:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.580 ^a	.337	.332	5.7845

a. Predictors: (Constant), Total UCAS points gained

b. Dependent Variable: Average mark for a degree

The next table shows the "Model summary". The corresponding coefficient of variation R-squared is 0.337. It represents the proportion of variation in the dependent variable that is explained by the changes independent variable. So, we can say that 33.7% of the variation in the average grade is explained by the UCAS points. The "Adjusted R-square" is only useful when we have more than one independent variable. So, we should leave it for now.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2649.526	1	2649.53	79.184	.000 ^b
	Residual	5219.827	156	33.460		
	Total	7869.353	157			

a. Dependent Variable: Average mark for a degree

b. Predictors: (Constant), Total UCAS points gained

The next output is the "Anova" table. Here SPSS has found one model, purely because there is only independent variable. Specifically, the F-ratio of 79.784, obtained by dividing the "Mean square" for regression by that of the residual, represents the amount of improvement in predicting the dependent variable (here, the average grade) using the model in relation to the weakness in the results. The sig (or p-value) of 0.000, suggests that the improvement brought about by the model is very unlikely to have happened by chance. We interpret this result as meaning that this model (Model 1) significantly improved our ability to predict the average degree grade. The bigger the F ratio the better the model.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	4.433	4.543		.976	.331
Total UCAS points gained	.291	.033	.580	8.899	.000

The next output “Coefficients” describes the model (Model 1) in a more precise manner. The loose equation of the line of best fit can be read from the column “B” under the “Unstandardized Coefficients” as:

$$\text{Average grade} = 4.433 + 0.291 \times \text{UCAS}$$

It says that one can predict the average degree grade from the UCAS points by multiplying the latter by 0.291 and then add 4.433 to the result.

So if a student was enrolled to the university with **160** UCAS points, then the student’s predicted average degree grade is:

$$\text{Average grade} = 4.433 + 0.291 \times 160 = 4.433 + 46.56 = 50.993 \approx 51$$

The sample gradient here is 0.291, which is the figure that determines if the amount by which the average degree grade changes if the UCAS points increases. Does this sample gradient come from a population of gradients with an average of “0” (zero) or not? The **sig** value of 0.000 represents the p-value for a t-test for the following hypotheses:

Null (H_0): Population gradient is 0

Null (H_1): Population gradient is different from 0

In other words, we are testing if the sample gradient of 0.291 is a very rare gradient among all sample gradients obtained from a population whose average gradient is equal to 0. The **p-value** of 0.000 is the probability of making an error if the null hypothesis is rejected. We conclude therefore that the population gradient has an average that is significantly different from 0.

Casewise Diagnostics^a

Case Number	Std. Residual	Average mark for a degree	Predicted Value	Residual
26	3.254	59.0	40.178	18.8219
42	3.003	59.0	41.631	17.3688
116	3.208	66.0	47.443	18.5566
151	-3.791	29.0	50.931	-21.9308
156	-3.725	32.0	53.546	-21.5463

a. Dependent Variable: Average mark for a degree

The next output is the “Casewise Diagnostics”. This table lists the individual cases where the difference between the “observed” and “predicted” values in the dependent variable (Average degree grades) is more than three standard deviations (as set in the run). We can choose to set it at 2 SDs instead. The table shows that there were 5 cases with a large deviation from the expected values. These cases have plots that do not fit the general trend in the scatter diagram. They can be detected as very far from the line of best fit. They are called outliers.

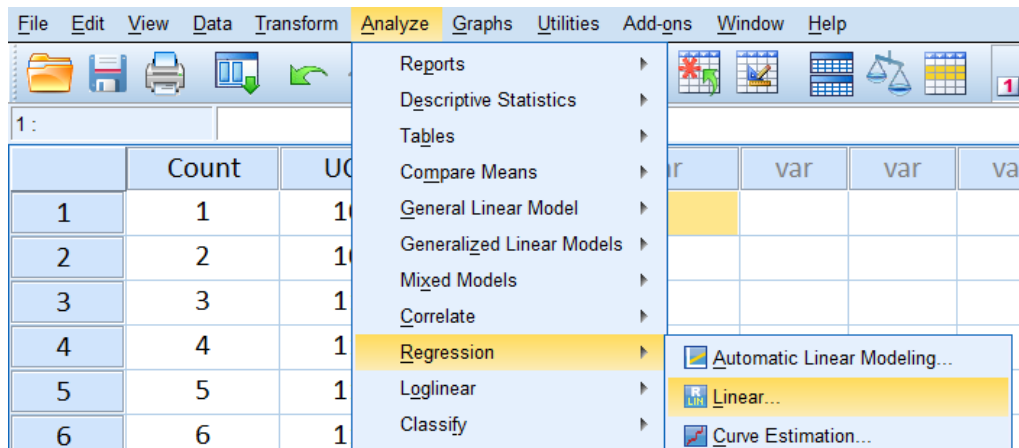
We can see, for example, that case **26** has an observed average degree grade of **59.0** but was predicted (by the model) to gain an average grade of **40.178**. This student has over-achieved in the degree because the predicted average grade is lower than the actual/observed average grade.

However, case **151** has under-achieved. The predicted average grade was 50.931 but this student only gained 29.0, which is much lower than expected.

The over- or under-achievement in the average grade can be seen from the sign (+ or -) of the “Residual” column.

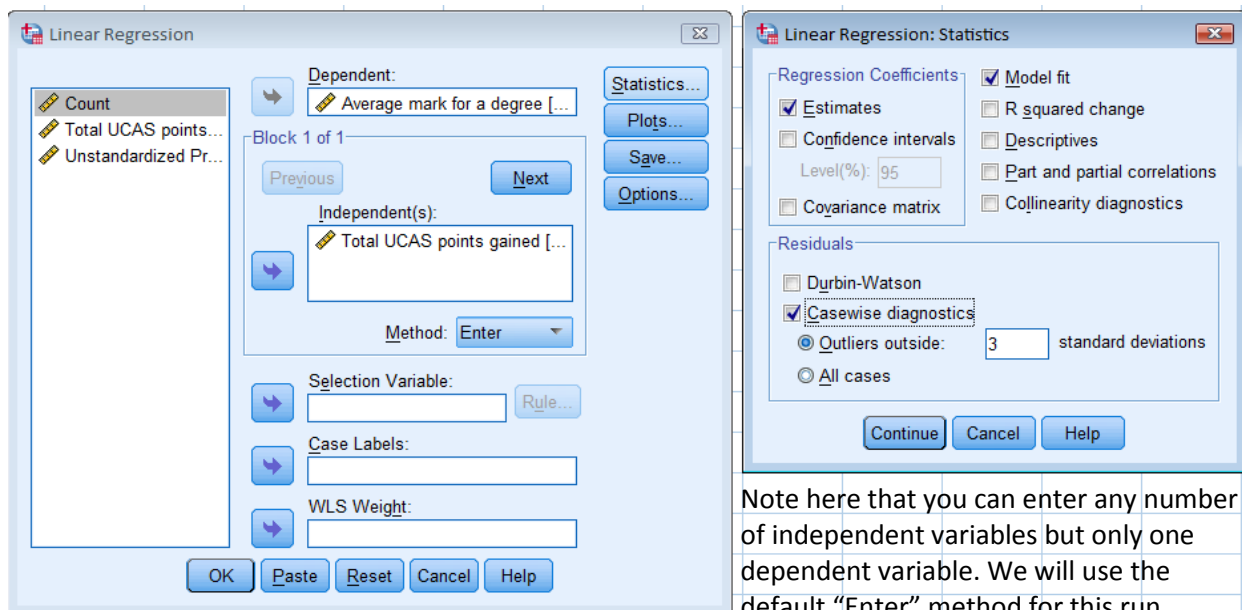
Using SPSS in correlation and regression

The above SPSS outputs were obtained by going to Analyze as follows:



- To obtain the correlation matrix, we go to “Correlate”, then “Bivariate”. Choose Pearson.
- To obtain the rest of the output, we go to “Regression”, then “Linear”.

In this example, the dependent variable is the average degree grade and the independent variable is the UCAS points.



Note here that you can enter any number of independent variables but only one dependent variable. We will use the default “Enter” method for this run.

Correlation versus Causation

It is important at this stage to remind you that a statistical association does not automatically mean that there is a causal relationship between the two variables. The number of UCAS points does not, by itself, condemn a student into a particular average grade for their degree. The analysis of correlation and by default, statistical association, only tells that the two variables vary together. It does not tell you why. It is simply an observation, albeit a significant observation that did not happen by chance.

It could be a **spurious** association, in the sense that the variables are not related logically, and that the association is due to something else. For instance, the number of UCAS points is due to the student's ability to work on their coursework and is so happens that most university modules are assessed by coursework method of assessment and hence, this explains the correlation between the two variables. If the choice of assessment method, at university, changes from coursework to “time-constrained exam”, then we may find that the two variables are not associated or that there is a weak correlation.

Rank Correlation

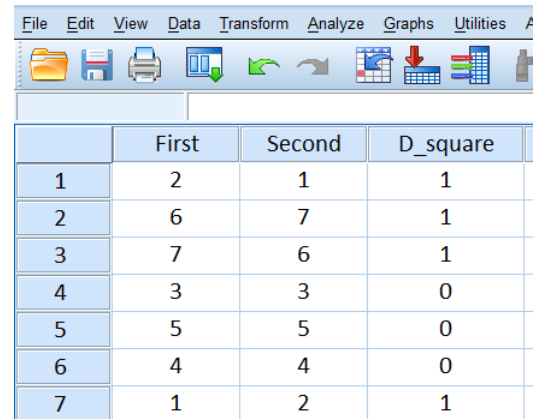
The previous example dealt with variables that are numerical and of continuous nature. If however, the data are ordinal in at least one of the variables, we need to use a non-parametric Spearman's rank correlation coefficient to measure the strength of association.

Example:

In a test of reliability, 7 repeat business customers to a bank were asked to rate the quality of service in relation to the advice given by their in-house agent. The rating is measure on a Likert scale 1 to 5, where 1=Very bad, and 7=Very good. The data are as follows:

The column "First" represents the first time the ratings were collected. The column "Second" represents the second time the ratings were collected.

The last column "D_square" represents the difference in the rating for each case.



	First	Second	D_square
1	2	1	1
2	6	7	1
3	7	6	1
4	3	3	0
5	5	5	0
6	4	4	0
7	1	2	1

The formula for the Spearman's rank correlation coefficient is:

$$R_s = 1 - \frac{6 \times \sum D^2}{n \times (n^2 - 1)}$$

- D is the difference between the scores of each case.
- n is the number of cases.

In this particular case, the value of $D^2 = 4$ and the value of $n = 7$. Substituted in the formula, the Spearman's rank correlation coefficient will be:

$$R_s = 1 - \frac{6 \times 4}{7 \times (7^2 - 1)} = 1 - \frac{24}{7 \times 48} = 1 - \frac{24}{336} = 1 - 0.0714285 = 0.9285715 \approx 0.929$$

Using SPSS

We go, "Analyze", "Correlate", then "Bivariate". Choose Spearman's.

The output obtained is:

Correlations				
			First	Second
Spearman's rho	First	Correlation Coefficient	1.000	.929**
		Sig. (2-tailed)	.	.003
		N	7	7
	Second	Correlation Coefficient	.929**	1.000
		Sig. (2-tailed)	.003	.
		N	7	7

** . Correlation is significant at the 0.01 level (2-tailed).

The interpretation is similar to the one give for Pearson's correlation coefficient on page 65.

The only problem with this formula is that it does not work well when you have tied scores. In this case SPSS tend to use other more complicated formulae.

Seminar exercises:

1. It is anticipated that the annual earnings of workers is associated to the number of years they spent in education. The following are SPSS output of this:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.733 ^a	.538	.535	1410.552

a. Predictors: (Constant), Number of years in education

b. Dependent Variable: Annual earning (in £)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	356454084.736	1	356454084.74	179.154	.000 ^b
	Residual	306407082.258	154	1989656.378		
	Total	662861166.994	155			

a. Dependent Variable: Annual earning (in £)

b. Predictors: (Constant), Number of years in education

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11395.52	540.560		21.081	.000
	Number of years in education	441.467	32.983	.733	13.385	.000

a. Dependent Variable: Annual earning (in £)

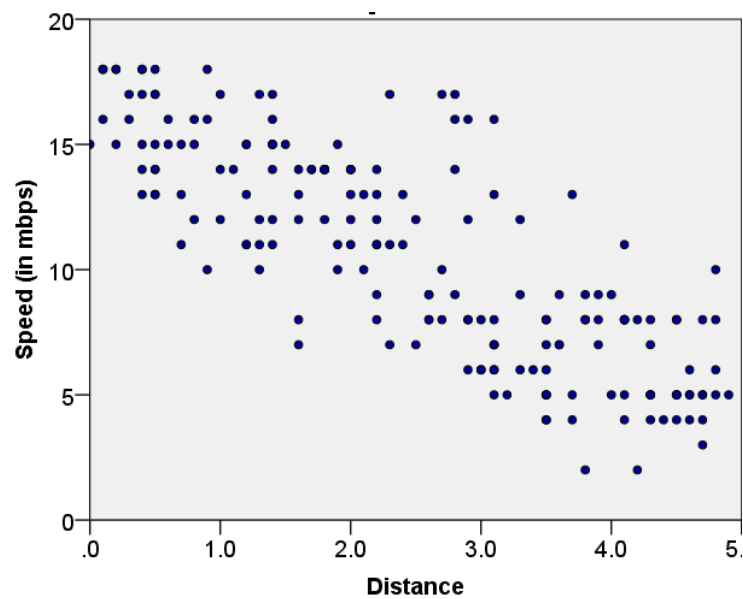
Casewise Diagnostics^a

Case Number	Std. Residual	Annual earning (in £)	Predicted Value	Residual
31	3.280	21320	16693.12	4626.877
115	-3.467	15334	20224.86	-4890.86

Required:

- a) What is the Pearson correlation coefficient? Is there a strong association between the variables?
- b) Does the model show significant improvement on the prediction of earnings?
- c) State the equation of the line of best fit.
- d) Is the gradient of this line significantly different to zero?
- e) How many outliers cases have been detected by SPSS?
- f) Which of these outlier cases has been doing better in terms of earnings in relation to length in education?

2. All broadband providing companies claim that the speed of your home broadband depends on the distance to the nearest telephone exchange. A survey conducted on



Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.813 ^a	.661	.660	2.516

a. Predictors: (Constant), Distance

b. Dependent Variable: Speed

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2287.85	1	2287.85	361.495	.000 ^b
	Residual	1170.84	185	6.329		
	Total	3458.68	186			

a. Dependent Variable: Speed

b. Predictors: (Constant), Distance

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16.737	.375		44.602	.000
	Distance	-2.492	.131	-.813	-19.013	.000

Required:

- What is the Pearson correlation coefficient? Is this a positive or negative association between the variables?
- Does the model show significant improvement on the prediction of earnings?
- State the equation of the line of best fit.
- Is the gradient of this line significantly different to zero?

Computing session

Download the data file "Broadband speed .sav"



Model answers

1.
 - a) 0.733 fairly strong
 - b) Yes it does. The p-value of 0.000 for the predictive model is fairly significant.
 - c) **Annual earning** = $11395.52 + 441.467 \times (\text{year in education})$
 - d) The gradient is significantly different to 0. The p-value (of 0.000) means that if the population gradient is "0", then there is no chance of getting a sample gradient of 0.733 (Can't reject the null hypothesis $\beta = 0$).
 - e) There are two outliers (case 31 and case 115).
 - f) Case 31. The predicted earning is lower than actual one.
2.
 - a) 0.813 Negative
 - b) Yes it does. The p-value of 0.000 for the predictive model is fairly significant.
 - c) **Boradband speed** = $16.737 - 2.492 \times \text{Distance}$
 - d) The gradient is significantly different to 0. The p-value (of 0.000) means that if the population gradient is "0", then there is practically no chance of getting a sample gradient of 0.813 (Can't reject the null hypothesis $\beta = 0$).