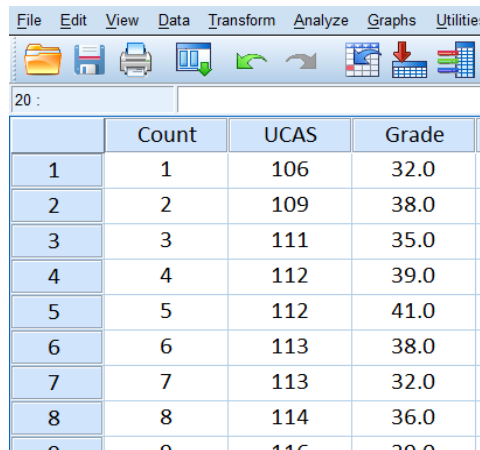


Multiple linear regression

Summary of the simple linear regression

The simple linear regression model is a model for a dependent (or outcome) variable that is explained by one single independent (or explanatory) variable in a straight line fashion.

When we encountered this topic, for the example of the average degree grade in relation to the UCAS points, we found that the best fit model for the two variables was represented by the equation:



	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

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

The outcome (or dependent) variable is the “average grade” and the explanatory (independent) variable is the “number of UCAS point”.

In this case, one variable “number of UCAS point” explains the variable “average grade”.

In general terms, this equation can be written as:

$$y = b_0 + (b_1 \times x_1)$$

Where $b_0 = 4.433$ and $b_1 = 0.291$

In the above equation, there is one explanatory variable x_1 . The most important number in this equation is the coefficient linked to this explanatory variable, namely, b_1 . Graphically, it represents the gradient (or slope) of the straight line but here, we call it simply, the coefficient of x_1 . If this coefficient is close to zero, then we can safely say that the explanatory x_1 does not explain the outcome variable. So, we are always on a look-out of whether this coefficient is equal to 0 or not.

In the example, we tested that this coefficient was significantly different from 0 and therefore concluded that the explanatory variable “UCAS points” seems to move in the same direction as the outcome variable “Average grade”.

For simple linear regression, it was important to look at the **Pearson correlation coefficient (R)** between the outcome and explanatory variables to evaluate the strength of association between them. In our example, the value of R was found to be 0.580 and was tested to be significantly different from 0.

Equally, it was important to look at the **coefficient of determination (R-square)** to ascertain how much of the variation in the outcome variable (Average grade) is explained by the explanatory variable (UCAS points). Our example shows that R-square was 0.337, which means that 33.7% of the variation in the average grades was explained by the number of UCAS points.

Linear regression with two variables

We all know, you and me, that students' average grade for their degree cannot solely rely on the number of UCAS points they bring with them from secondary education. I should hope so, because if that is the case, then studying at university does not add any value whatsoever to the degree. That is simply not the case. We also know that resit marks, in most cases, are capped to 40%. This generally affects the overall mark of the module and ultimately the overall grade. Let us introduce a new variable called “Resit” which represents the number of resits a student has taken during their degree.

	Count	UCAS	Resit	Grade
1	1	106	12	32.0
2	2	109	8	38.0
3	3	111	11	35.0
4	4	112	9	39.0
5	5	112	9	41.0
6	6	113	11	38.0

Does this additional variable improve the model for predicting the average grade of the student?

When creating a model with more than one explanatory variable, a couple of complications arise.

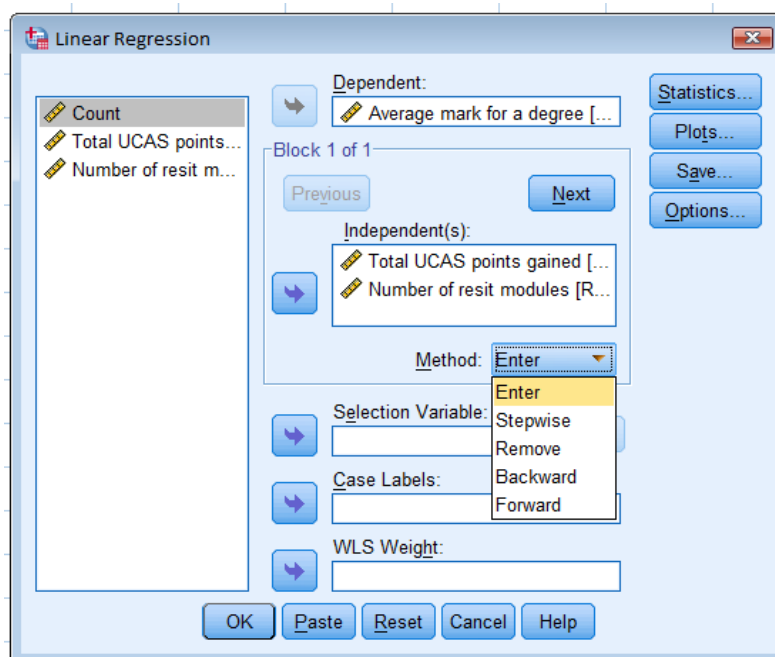
Firstly, we want to find a model which accounts for as much of the variance in the outcome variable as possible and therefore we can't just throw any new variable into it.

Secondly, it is to do with the correlation between the explanatory variables. If there is a significant correlation between two explanatory variables, it is unclear how much of the variance in the outcome variable is explained by each of them. In our example, the variance in average grade is explained by the number of UCAS points. But if the number of UCAS points correlate highly with the number of resit attempts, then the amount of variance in the average grade can be shared by both of explanatory variables. This problem is referred to as **Multicollinearity**.

It is possible to adjust a multiple regression model to account for this problem. If the model is created in steps, we can better estimate which of the explanatory variables accounts for the largest variance in the outcome variable. The changes in the values of R-square can be observed after each step. This means a new explanatory variable is added to the model only if it explains some unique variance in the outcome that is not accounted for by the previous explanatory variables in the model.

Stepwise Versus Enter methods

There are 5 methods available on SPSS as seen below.



The **Forward** method starts from the beginning. SPSS searches from the specified list of possible explanatory variables for the single one with the strongest correlation with the outcome and enters that

first. It continues to add explanatory variables in order of how much additional unique variance they explain. It only stops when there are no further variables that can explain additional variance.

The **Backward** method does the opposite. It begins with all of the explanatory variables available and then removes those which are not making a significant contribution to the variance of the outcome.

The **Stepwise** option is similar but uses both forward and backwards criteria for deciding when to add or remove an explanatory variable.

The **Enter** method allows you to control how the variables are entered in the model. At the simplest level all the variables could be entered together in a single group called a 'block'. However, variables can be entered in separate blocks of explanatory variables in order of their theoretical importance. Knowing the precise order is very difficult. **Enter** will include all variables in the specified block while **Remove** removes all variables in the specified block.

Although this may sound confusing, it will be clearer when we run the multiple regression analysis.

Using SPSS for multiple regression

Let us use our example with two explanatory variables. We start by looking at the correlation matrix.

Correlations

		Total UCAS points gained	Number of resit modules	Average mark for a degree
Total UCAS points gained	Pearson Correlation	1	-.507**	.580**
	Sig. (2-tailed)		.000	.000
Number of resit modules	Pearson Correlation	-.507**	1	-.904**
	Sig. (2-tailed)	.000		.000
Average mark for a degree	Pearson Correlation	.580**	-.904**	1
	Sig. (2-tailed)	.000	.000	

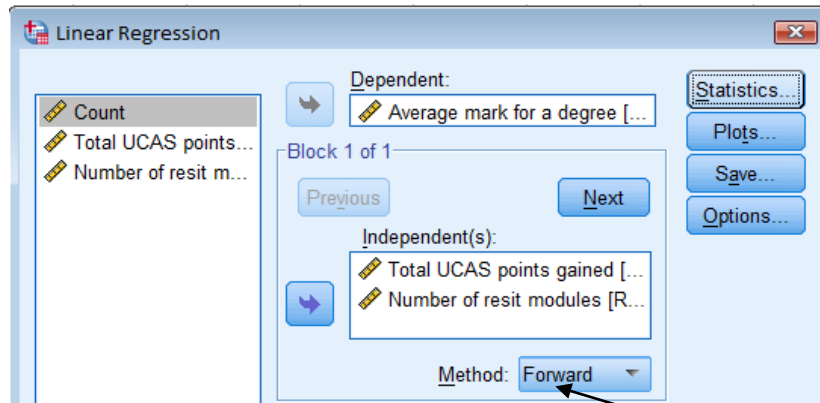
**. Correlation is significant at the 0.01 level (2-tailed).

One can see that there is moderate negative correlation between the two explanatory variables – correlation coefficient of -0.507 and this is statistically significant. This means that the variable "UCAS points" and "Resits" share quite a bit of the variance on the average grade.

Both explanatory variables are highly correlated with the outcome variable "Average grade". Indeed, the explanatory variable "Resit" is very strongly correlated with the outcome variable, albeit negatively, with a correlation coefficient of -0.904 . Very likely, therefore, the explanatory variable "Resit" will be the first variable chosen in the **Forward** method.

The two-tailed p-values show that all these correlations coefficients are statistically and significantly different from 0.

Let us run the various linear regression on SPSS. We start with the **Forward** method.



The output related to the model is produced below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.904 ^a	.817	.816	3.0395	.81686	695.803	1	156	.000000
2	.915 ^b	.837	.835	2.8786	.01993	18.927	1	155	.000025

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
1	(Constant)	54.502	.445			122.522	.000
	Number of resit modules	-1.822	.069	-.904		-26.378	.000
2	(Constant)	42.240	2.850			14.822	.000
	Number of resit modules	-1.655	.076	-.821		-21.793	.000
	Total UCAS points gained	.082	.019	.164		4.350	.000

Model 1: Includes on the “Resit” variable.

Model 2: Includes both the “Resit” and the “UCAS points” variables.

Model 1 (Adjusted R-square of 0.816), which means that the variable “Resit” explains 81.7% of variance of the average grades.

Model 2 (Adjusted R-square of 0.835), which means that the variable “UCAS points” only explains an additional of 1.99% of the variance of the average grades. There is a large amount of variance that is shared by the two explanatory variables.

The final model – Model 2, is given by the equation:

$$\text{Average grade} = 42.240 - 1.655 \times \text{Resit} + 0.082 \times \text{UCAS}$$

Note that there are two coefficients, one for each explanatory variable. Both these coefficients are statistically significant in the sense that they belong to a population of coefficients that are different from zero. The coefficient for the “Resit” variable is -1.655 . The coefficient for the “UCAS points” variable is $+0.082$

The reason, the coefficient for the “Resit” variable is negative is because we expect a negative correlation between “Average grades” and “Resit”. The more resits a student takes, the lower his/her average grade. Let us run the various linear regression on SPSS. We start with the **Backward** method.

The output using this method is produced below:

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.915 ^a	.837	.835	2.8786	.837	397.343	2	155	.000

a. Predictors: (Constant), Number of resit modules, Total UCAS points gained

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	
1	(Constant)	42.240	2.850		14.822
	Total UCAS points gained	.082	.019	.164	4.350
	Number of resit modules	-1.655	.076	-.821	-21.793

With this method, there is only one model - **Model 1**, which includes on the “Resit” and the “UCAS points” variables. This model accounts for 83.7% of the variance of average grades.

This model – Model 1, is given by the same equation, except the coefficients are in a different order.

$$\text{Average grade} = 42.240 + 0.082 \times \text{UCAS} - 1.655 \times \text{Resit}$$

So, when we consider only one explanatory variable, the general equation is of the form:

$$y = b_0 + (b_1 \times x_1)$$

If we consider two explanatory variables, the general equation is of the form:

$$y = b_0 + (b_1 \times x_1) + (b_2 \times x_2)$$

If we consider three explanatory variables, the general equation is of the form:

$$y = b_0 + (b_1 \times x_1) + (b_2 \times x_2) + (b_3 \times x_3)$$

And so on. More importantly, the coefficients change value each time we add another explanatory variable. This is because the final model needs to re-adjust to the additional variable.

Let us continue with this interesting example and add yet another controversial explanatory variable, that of “attendance”.

File	Edit	View	Data	Transform	Analyze	Graphs	Utilities	Add-ons	Window	Help
										
119 :										
	Count	UCAS	Resit	Attendance	Grade					
114	114	148	2	53	51.6					
115	115	148	2	91	49.0					
116	116	148	0	63	66.0					
117	117	149	2	58	47.5					

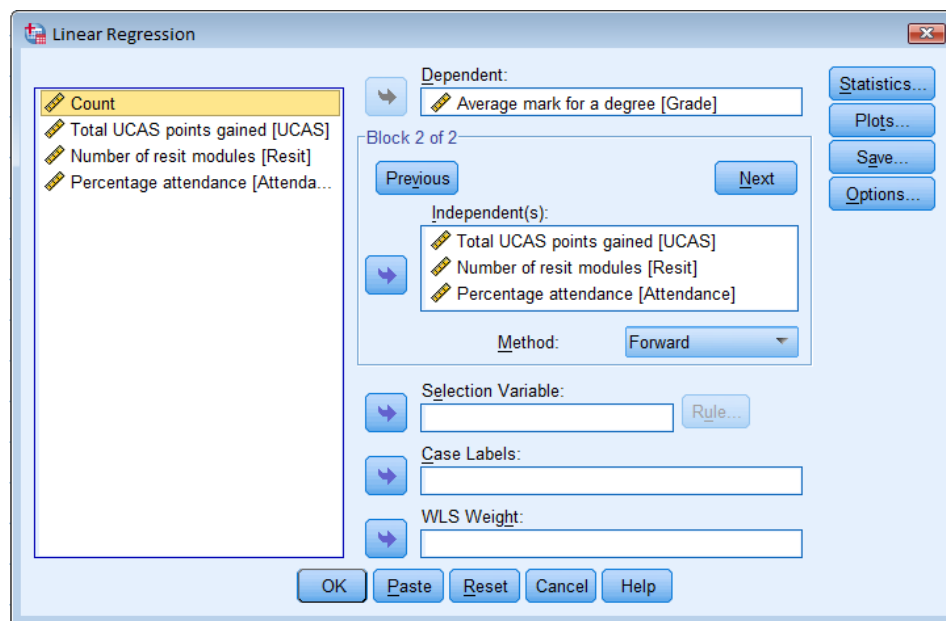
The correlation matrix.

Correlations

		Total UCAS points gained	Number of resit modules	Percentage attendance	Average mark for a degree
Total UCAS points gained	Pearson Correlation Sig. (2-tailed)	1	-.507** .000	.329** .000	.580** .000
Number of resit modules	Pearson Correlation Sig. (2-tailed)	-.507** .000	1	-.306** .000	-.904** .000
Percentage attendance	Pearson Correlation Sig. (2-tailed)	.329** .000	-.306** .000	1	.276** .000
Average mark for a degree	Pearson Correlation Sig. (2-tailed)	.580** .000	-.904** .000	.276** .000	1

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

We note here that “Attendance” shows a relatively weak correlation with the outcome variable “Average grade”. It is therefore not likely to contribute to the variance of the outcome in a major way. Let us run the **Forward** method of regression on SPSS.



The output is as follows:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	54.502	.445		122.522	.000
	Number of resit modules	-1.822	.069	-.904	-26.378	.000
2	(Constant)	42.240	2.850		14.822	.000
	Number of resit modules	-1.655	.076	-.821	-21.793	.000
	Total UCAS points gained	.082	.019	.164	4.350	.000

a. Dependent Variable: Average mark for a degree

We can see that SPSS started with “Resit” variable, then added the “UCAS points” variable and stop at that. It did not find significant to add the “Attendance” variable.

We can see much more clearly if we run the **Backward** method. The regression output follows:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.915 ^a	.83780	.83464	2.8790	.838	265.144	3	154	.000
2	.915 ^b	.83679	.83468	2.8786	-.001	.958	1	154	.329

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42.743	2.896		14.759	.000
	Total UCAS points gained	.086	.019	.172	4.459	.000
	Number of resit modules	-1.667	.077	-.827	-21.636	.000
	Percentage attendance	-.017	.017	-.034	-.979	.329
2	(Constant)	42.240	2.850		14.822	.000
	Total UCAS points gained	.082	.019	.164	4.350	.000
	Number of resit modules	-1.655	.076	-.821	-21.793	.000

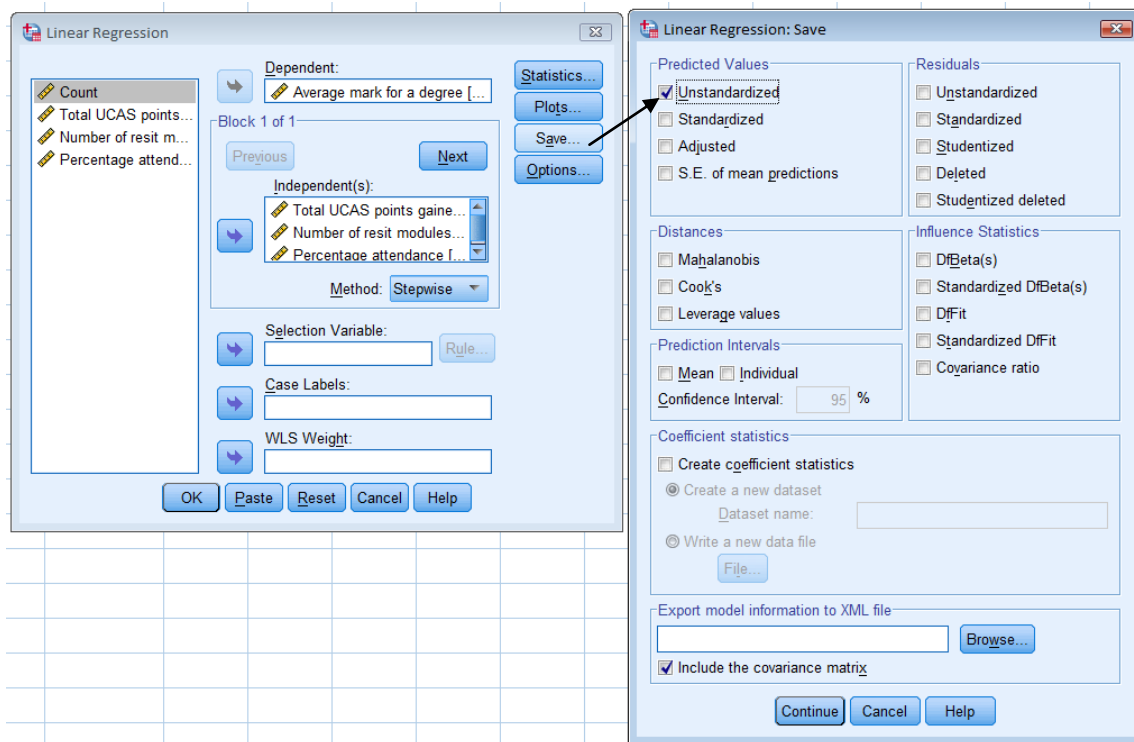
This method shows clearly what was done. SPSS included all the variables (including the Attendance), then removed the “Attendance “ variable from the Model 1 to obtain Model 2. There are two obvious reasons for this.

1. Although **model 2** has a slightly lower R-square value, it has a slightly higher “Adjusted R square” than **Model 1**, a value of the coefficient of determination that is much better for multiple regression because it adjust for the sample size.
2. The coefficient for “Attendance”, -0.017 is not statistically and significantly different from 0, judging by the large p-value of 0.329. This means that if we accept that the population gradient is different from zero, we are making an error of 32.9%. Therefore, this population “Attendance” gradient is zero. This, in turn, means that the average grade is not associated to the attendance.

The model (**Model 2**) for predicting the average grade at university remains the same as with two explanatory variables.

Predictive power of the model

Once the best and most appropriate model has been found, we can find the predicted values for the outcome variable in SPSS, by making sure you tick the “Unstandardized” box under “Predicted values” in the “Save” button as below.



The predictions for the available observed values of the independent (or explanatory) variables will show up on the “Data View” tab of SPSS as such.

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help						
1: PRE_1 31.08316496755881						
	Count	UCAS	Resit	Attendance	Grade	PRE_1
1	1	106	12	42	32.0	31.08316
2	2	109	8	35	38.0	37.94751
3	3	111	11	67	35.0	33.14797
4	4	112	9	50	39.0	36.53911
5	5	112	9	61	41.0	36.53911
6	6	113	11	51	38.0	33.31207
7	7	113	11	51	38.0	33.31207

By default, these predictions are labelled “Pre_1” but you can give another label to them.

Let us check one prediction. For case 1, the number of UCAS points were 106 and the students had 12 resits.

$$\text{Average grade} = 42.240 + 0.082 \times 106 - 1.655 \times 12 = 42.240 + 8.692 - 19.86 = 31.072$$

SPSS output gives a slightly different prediction of **31.08316**. This is because it uses the full version of the coefficient values. Ours are shown correct to 3 decimal places only.

Seminar Exercises

1. A double-glazed window maker advertised his prices. They are dependent to their size given by the width and the height.

A stepwise linear regression method was run and the SPSS output is given below:

Correlations

		Width of window in cm	Height of window in cm	Price of window in £
Width of window in cm	Pearson Correlation	1	.061	.697**
	Sig. (2-tailed)		.835	.006
Height of window in cm	Pearson Correlation	.061	1	.728**
	Sig. (2-tailed)	.835		.003
Price of window in £	Pearson Correlation	.697**	.728**	1
	Sig. (2-tailed)	.006	.003	

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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.728	.529	.490	10.142
2	.978	.957	.949	3.209

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	46.586	6.747		6.905	.000
	Height of window in cm	.284	.077	.728	3.674	.003
2	(Constant)	25.603	2.933		8.731	.000
	Height of window in cm	.269	.025	.687	10.950	.000
	Width of window in cm	.136	.013	.655	10.435	.000

Required:

- a) Name the explanatory variables and the outcome variable in this example.
- b) Is there a severe case of multicollinearity in this example?
- c) Name the best model for predicting the outcome variable. Give your reason.
- d) How much of the variance of the outcome variable is accounted for by this model?
- e) Write down the linear regression equation for this model.
- f) For the best model, are the coefficients statistically significant? Give your reasons.

Computing session

Activity one:

Download the data file "Average degree grade . sav"

- a) Use SPSS linear regression facility to produce the multiple regression output using:
 - I. The Forward method
 - II. The Backward method
 - III. The Stepwise method
- b) Compare the output for these different methods and discuss the results.

