Chapter 24

Advanced Quantitative Data Analysis

Daniel Muijs

Doing Regression Analysis in SPSS

When we want to do regression analysis in SPSS, we have to go through the following steps:

1 As usual, we choose *analyse*.
2 In the pop-down menu that appears, we go to *regression*.
3 A new pop-down menu appears. We choose *linear.*
4 A new screen appears. From the list of variables on the left, we will first choose our dependent variable, in this case 'pupil attainment', and then click on the button next to dependent. The variable name now appears in the 'dependent' box:
5 Next, we have to choose our predictors. Then we click on 'ok'.

The following is the output from our analysis:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
dimension0[a]	School climate, teaching quality, student ability, school leadership	.	Enter

a. Dependent variable: school grades in English.

Model Summary

Model	R	R Square	Adjusted R Square	Std Error of the Estimate
dimension0	.538	.31	.29	9.78882

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7026. 942	3	2342.314	24.445	.000
	Residual	54522.170	569	95.821		
	Total	61549.112	572			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std Error	Beta	t	Sig.
1	(Constant)	142.617	9.581		14.886	.000
	School Climate	.156	.104	.261	2.544	.005
	Teaching Quality	.207	.292	.233	2.260	.006
	School Leadership	.353	.075	.293	2.610	.003
	Student Ability	.467	.054	.418	3.975	.000

The first box of the output is labelled **'Variables entered/removed'**. This is simply a list of all the predictors we have entered into the equation. It also gives the method we have used (more about that later).

The second box is labelled **'Model summary'**. This is an important one, as it gives us the measures of how well our overall model, i.e. our three predictors together, is able to predict English grades. The first measure in the table is called R. This is a measure of how well our predictors predict the outcome, but we need to take the squared root of R to get a more accurate measure. This is R squared, which SPSS shows us in the next column. This gives us the amount of variance in pupil attainment explained by the four predictor variables together. R squared varies between 0 and 1. The next column is labelled adjusted R squared. This is, as the name implies, a correction to R squared, which takes into account that we are looking at a sample rather than at the population. As the model is likely to fit the population less well than the sample, R squared is adjusted downwards to give us a measure of how well our model is likely to fit in the population. Adjusted R squared also lies between 0 and 1. In this case it is 0.29, which suggests fit of our model to the data is modest. As a rough guide, the following rule of thumb can be used to see how well our model fits the data:

<0.1: poor fit
0.11 – 0.3: modest fit
0.31 – 0.5: moderate fit
> 0.5: strong fit

The final column gives us the standard error of the estimate. This is a measure of how much R is predicted to vary from one sample to the next.

The third output box is labelled **ANOVA**. We will not worry about that one here.

The fourth box, **Coefficients**, gives us some important information, as that is where we will be able to look at the b, beta and significance of our three predictors separately.

The first column gives us the names of our predictor variables. The variable labelled 'constant' is the intercept, or a. The second column gives us our b coefficients, the value that Y will change by if X changes by 1 unit. If we look at school climate, that value is .156. So, if scores on the school climate scale go up by 1, pupil attainment is predicted to go up by 0.156. In the next column, the standard error for each of these 'b's is given.

The following column contains the Beta parameters. What is Beta? One problem with b is that because variables are often measured using different scales, you can't use b values to see which of your variables has the strongest influence on the dependent variable. That is why if we want to look at the *effect size* of each of our variables, we need to standardise the variables so that they are all measured on the same scale. The Betas give us these standardised coefficients. Betas vary between 0 and 1, with, as usual, 1 being the strongest effect, and are interpreted similarly to correlation coefficients.

The final column in this box gives us the statistical significance of the relationship between each predictor and the dependent variable. In other words, how likely is it that we would have found a relationship this strong in our sample if there wasn't one in the population? Again, interpretation is the same as for correlation coefficients.

Logistic Regression in SPSS

To calculate logistic regression in SPSS, using pass/fail as our dependent variable, we need to do the following:

1. As usual, we choose *analyse*.
2. In the pop-down menu that appears, we go to *regression*.
3. A new pop-down menu appears. We choose *binary logistic*.
4. A new screen appears. From the list of variables on the left, we will first choose our dependent variable, in this case 'passfail', and then click on the button next to dependent. The variable name now appears in the 'dependent' box.
5. Next, we have to choose our predictors. We choose family gender and click on the arrow next to the 'covariates' box. We do the same for 'I often don't understand things in school' and 'I'm one of the best in my class at all subjects'. Then we click on 'ok'.

The output then appears, and it is the most confusing one yet! Luckily, we don't require all the information provided.

Logistic regression

Notes

Output Created		26-Mar-2010 15:13:58
Comments		
Input	Data	F:\
	Active Dataset	DataSet1
	File Label	Written by SPSS for Windows
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	889
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing
Syntax		LOGISTIC REGRESSION VARIABLES passfail /METHOD=ENTER gender schsc2 schsc3 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
Resources	Processor Time	00:00:00.016
	Elapsed Time	00:00:00.024

Case Processing Summary

UnweightedCases[a]		N	Percent
Selected Cases	Included in Analysis	884	99.4
	Missing Cases	5	.6
	Total	889	100.0
Unselected Cases		0	.0
Total		889	100.0

[a]. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value		Internal Value
dimension0	fail	0
	pass	1

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			passfail		Percentage correct
			fail	pass	
Step 0	passfail	fail	0	414	.0
		pass	0	470	100.0
	Overall Percentage				53.2

[a.] Constant is included in the model.
[b.] The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.127	.067	3.543	1	.060	1.135

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Gender	.654	1	.419
		School climate	129.014	1	.000
		School leadership	156.486	1	.000
	Overall Statistics		186.409	3	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi square	df	Sig.
Step 1	Step	202.407	3	.000
	Block	202.407	3	.000
	Model	202.407	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1019.527[a]	.205	.273

[a.] Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			passfail		Percentage correct
			fail	pass	
Step 1	passfail	fail	292	122	70.5
		pass	135	335	71.3
Overall Percentage					70.9

[a.] The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender	.095	.153	.387	1	.534	1.100
	School climate	.537	.094	32.312	1	.000	1.710
	School leadership	.777	.098	62.333	1	.000	2.175
	Constant	-3.425	.363	89.230	1	.000	.033

[a.] Variable(s) entered on step 1: gender, schsc2, schsc3.

The first set of boxes provides us with some technical details and information on variables, which are not very important.

The useful information starts with the section labelled **'Block 0'.** This section gives us the statistics for the baseline model, which doesn't contain any of the independent variables.

The first box in this set gives us the comparison between the predicted and actual scores. As we have no independent variables in this baseline model, the prediction is simply that everyone will pass, which turns out to be correct in 53.2% of cases (you can find this number in the bottom-right corner of the table, in the row labelled 'Overall percentage').

The second box gives us the regression coefficients for the variables in the model. The only variable in the model at present is constant (the intercept), which has no substantive interpretation. The third box contains the variables not in the baseline model.

The next section of the output is labelled **'Block 1: Method = Enter'**. This section contains the statistics for the model with the independent variables included.

The first box gives us the so-called 'omnibus test of model coefficients'. This gives us an indication of whether or not the model with our independent variables fits the data better (i.e. gives us a better prediction of individual scores) than the baseline model. We can find significance in the final column on this table, and we can see (significance is less than .05) that the model is significant, which means that our model with the three predictors fits better than a model with no predictors.

The second box provides us with the Pseudo R square statistics. There are two measures, Cox & Snell and Nagelkerke. Both use a somewhat different formula, but both are equally valid. In this case, Cox & Snell is .20, and Nagelkerke is .27. These numbers indicate modest improvement in fit over the baseline model (0–.1 would indicate poor improvement in fit, .1–.3 modest improvement, .3–.5 moderate improvement and more than .5 strong improvement).

The third box is called the classification table, and gives us the comparison between predicted scores and the actual scores. We can see, for example, that 292 pupils who were predicted by our model (with the three predictors) were predicted to fail and did indeed fail, while 122 were predicted to fail and in fact passed. In total, 70.9% of our predictions were accurate, which though far from perfect is a clear improvement over the baseline model, where 53.2% of predictions were accurate.

The final box provides the data for the independent variables. In the box labelled 'sig', we can see that the variables school climate and school leadership are significant, while gender isn't. The regression coefficients are given under B, and show that an increase of one on the scale for schsc3 increases the probability of a pass on the outcome variable by .777.