Questions and answers for Chapter 9

1. You want to find out what factors predict achievement in mathematics. Develop a model that you think can explain this.

   There are of course many possible models that might predict mathematics achievement. As an example we will use a model based on a number of plausible predictors in our sample: we suggest that 'I'm among the best in my class at maths', 'I like going to school' and 'gender' will predict mathematics achievement.

2. Calculate your model SPSS. What is R squared, and what does it mean?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.455[a]	.207	.202	10.89574

a. Predictors: (Constant), gender, among th best in my class at all subjects, like going to school

Our output shows us that R Squared is .207. This means that our three variables together explain about 21% of the variance in maths achievement. This suggests a modest fit of our model to the data.

3. Calculate your model in SPSS. What is your b and what does it mean?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	59.387	1.995		29.773	.000
	like going to school	.545	.425	.050	1.283	.200
	among th best in my class at all subjects	5.722	.498	.441	11.489	.000
	gender	.940	.925	.039	1.016	.310

a. Dependent Variable: school grades maths

The regression coefficient, b, represents the amount the dependent variable will change by if the independent variable changes by one unit. The b coefficient for 'I like going to school' is .54. This means that on average, if pupils go up 1 point on the 'I like going to school' scale (i.e. from disagree to agree) maths achievement will improve by 0.545 points (which in this case is equivalent to 0.545%). The b for 'I'm among the best in my class at maths' is 5.722. This means that on average, if pupils go up 1 point on the 'I'm among the best in my class at maths' scale (i.e. from disagree to agree) maths achievement will improve by 5.722 points. Finally, the b for gender is .940. This means that on average if a pupil is a girl rather than a boy maths achievement will improve by 0.94 points.

4. Calculate your model in SPSS. What is Beta, and what does it mean?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	59.387	1.995		29.773	.000
	like going to school	.545	.425	.050	1.283	.200
	among th best in my class at all subjects	5.722	.498	.441	11.489	.000
	gender	.940	.925	.039	1.016	.310

a. Dependent Variable: school grades maths

Beta is the standardised regression coefficient, which allows us to compare the effect of variables measured on different scales. In this case we can see that 'I'm among the best in my class at school' is the strongest predictor of maths achievement, with a Beta of .441, a moderate effect size. 'I like going to school', with a Beta of .5, and gender, with a Beta of .39, are both weak predictors of maths achievement.

5. Calculate your model in SPSS. What is the p-value, and what does it mean?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	59.387	1.995		29.773	.000
	like going to school	.545	.425	.050	1.283	.200
	among th best in my class at all subjects	5.722	.498	.441	11.489	.000
	gender	.940	.925	.039	1.016	.310

a. Dependent Variable: school grades maths

The p-value is our measure of statistical significance and will tell us whether it is likely that we would have found a relationship of this size in the sample if there was no relationship in the population. Looking at our p-values the only variable that is significant is 'I'm among the best in my class at maths', with a p-value of <.001. The other two variables both have p-values well over .05. This means that while we can be quite confident that there will be a relationship between thinking 'I'm among the best in my class at maths' and maths achievement in the population, we can't say this about either gender or 'I like going to school'.

6. If you find a model that fits well, does that mean your predictors cause your dependent variable? Why (not)?

   Not necessarily. Think again of our three conditions. Regression models clearly demonstrate whether or not the predictors are related to the dependent variable, so that condition can be fulfilled. As for the condition of the relationship not being caused by a third, underlying variable, regression does a better job than correlation in that you can include other possible causes in the model. However, it is unlikely that we will have included all possible variables. Finally, the time condition is not demonstrated by regression analysis.

7. What is a dummy variable, and when do you use it?

   A dummy variable is created by turning a nominal variable into one or more two-category variables. To do this, what we need to do is make one category into our reference category, to which the others are going to be compared. Lets take state schools as our reference category in this example. We are first going to compare children in Catholic schools with children in state schools, and then children in local authority schools with children in state schools. How do we do this? We will have to make two new variables, one for Catholic and one for local authority schools. We will have to recode our variable school type so that all Catholic schools are coded as 1, and all other schools as 0.

   We have to use dummy variables when we have nominal predictors in regression models.

8. When would you use regression rather than correlation?

Regression analysis has a number of advantages over correlation. Firstly, it allows us to develop more accurate models by allowing us to include a number of different predictors of an outcome we are interested in. As the relationship of each individual predictor to the dependent variable is controlled for the relationship with any other variables in the model, regression also gives a more accurate picture of the strength of those relationships. Regression analysis also provides us with a number of useful diagnostics to test the validity of our models.

9. You want to find out which factors predict responses to the question 'I get good marks in English'. Develop a model that you think can explain this.

'I get good marks in English' is an ordinal variable, therefore ordinal regression would be the most appropriate statistical method to use. Thinking of possible predictors, we could look at gender, 'the teachers think I'm good at English' and 'I like going to school'. Again, many different models are of course possible.

10. Calculate your model in SPSS. Does your model fit the data?

The main sections of the output we need to answer this question are given below.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	901.609			
Final	241.019	660.590	7	.000

Link function: Logit.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	95.709	86	.222
Deviance	68.051	86	.923

Link function: Logit.

Pseudo R-Square

Cox and Snell	.527
Nagelkerke	.578
McFadden	.309

Link function: Logit.

The first box, labelled 'model fitting information' provides us with a comparison between the baseline model with no independent variables, called 'intercept only', with the model with the three predictors which is called 'final'. A Chi Square test was conducted to look at improvement in prediction compared to the baseline model. If the test is significant this

indicates that our model fits better than the baseline model with no predictors. As we can see, this is clearly the case here (sig is less than .05).

More measures of model fit are provided in the next box (Pearson and Deviance). These two measures compare the actual results for each respondent (i.e. do they agree strongly, agree somewhat, disagree somewhat or disagree strongly that school is fun) with the outcome predicted by our model. Again we need to look at the significance level. Unlike many of the measures we have looked at, in this case we want the difference between the expected and actual results to be non-significant. This is because if our model fits well, the observed and expected cell counts should be similar (i.e. respondents have given the answer we predicted based on our model). This is the case here, which suggests that our model fits the data.

The Pseudo R Squared statistics are given in the following box. When we look at Cox and Snell and Nagelkerke's measures (as in logistic regression), we find a good fit, with Pseudo R squareds of over .5.

11. Calculate your model in SPSS. What does it tell you about the coefficients?

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[engsc1 = 1]	-8.414	.378	495.260	1	.000	-9.155	-7.673
	[engsc1 = 2]	-5.363	.325	272.752	1	.000	-5.999	-4.726
	[engsc1 = 3]	-2.021	.270	56.032	1	.000	-2.551	-1.492
Location	[gender=1]	-.337	.143	5.521	1	.019	-.618	-.056
	[gender=2]	0[a]	.	.	0	.	.	.
	[attsc7=1]	-.269	.205	1.729	1	.189	-.670	.132
	[attsc7=2]	-.026	.213	.015	1	.904	-.443	.391
	[attsc7=3]	-.036	.181	.040	1	.841	-.392	.319
	[attsc7=4]	0[a]	.	.	0	.	.	.
	[engsc4=1]	-6.960	.408	290.760	1	.000	-7.760	-6.160
	[engsc4=2]	-5.742	.323	316.771	1	.000	-6.374	-5.110
	[engsc4=3]	-2.756	.269	104.914	1	.000	-3.284	-2.229
	[engsc4=4]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

In this box we can see the parameters for the individual variables. The 'threshold' statistics refer to the dependent variable, and are less important for us at this stage. The statistics for the independent variables are given under 'location'. What we can see here is that we have statistics not for the variables as a whole, but for each category of the variable. We can see this most clearly when we look at the variable 'The teachers think I'm good at English' (engsc4). Firstly, we can see that there is one category, 4, which

corresponds to an 'agree strongly' response, that does not have a significance level calculated (see column labelled 'sig') and for which the coefficient is 0 (see column labelled 'estimates'). This is because this is the reference category, to which we compare all the others, like we did when we used nominal predictors in multiple linear regression. We can see here that responding 1 (disagree strongly) is significantly (column labelled 'sig') related to responses on the dependent variable. When we look at the estimates, we can see that the coefficient is -6.96, which means that respondents who disagree strongly that teachers think they are good at maths are less likely to agree they get good marks in English than respondents who agree strongly that teachers think they are good at maths (our reference category 4). Category 2 (this corresponds to a response of disagree somewhat) is also related significantly to responses on the independent variable. The coefficient is -5.74, so respondents who disagree somewhat that teachers think they are good at maths are less likely to agree that they get good marks in maths than respondents who agree strongly that teachers think they are good at maths, though this is less the case than it was for respondents who disagreed strongly (as the coefficient here was -6.96). The third category, which corresponds to agree somewhat is also significant, though the coefficient, at -2.756 is lower still. Overall we would say that there is therefore a relationship between 'the teacher thinks I'm good at maths' and 'I get good marks in maths', and that the relationship is neatly ordered.

When we look at the other two variables, gender and 'I like going to school' (attsc7), we can again see that one of the categories is the reference category. For gender this is category 2 (girl), for 'I like going to school' this is category 4 (agree strongly). When we

look at the estimates we can see that boys (coefficient -.337) are somewhat less likely to agree that they get good marks in English. None of the categories for 'I like going to school' are significant.

12. You want to find out which factors predict a pass (over 75%) or fail (75% or below) in mathematics. Develop a model that you think can explain this.

    To do this we need to choose passfail as our dependent variable. This variable is categorical, so it makes most sense for us to use logistic regression as our analysis method.

    Again, many models are possible, but we will use the same three predictors as we did for questions 1 to 5, I'm among the best in my class at maths', 'I like going to school' and 'gender'.

13. Calculate your model in SPSS. Does your model fit the data?

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	147.847	3	.000
	Block	147.847	3	.000
	Model	147.847	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1072.569[a]	.154	.206

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

Classification Table[a]

Observed			Predicted		
			passfail		
			fail	pass	Percentage Correct
Step 1	passfail	fail	274	139	66.3
		pass	138	332	70.6
	Overall Percentage				68.6

a. The cut value is .500

The first box here gives us the '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 next 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 .15, and Nagelkerke is .21. These numbers indicate modest improvement in fit over the baseline model (0-.1 would indicate poor improvement in fir, .1-.3 modest improvement, .3-.5 moderate improvement and more than .5 strong improvement).

The next box is called the classification table, and gives us the comparison between predicted scores and the actual scores. We can see, for example, that 274 pupils who were predicted to fail by our model (with the three predictors) did indeed fail, while 138 were predicted to fail and in fact passed. In total, 68.6% 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.

14. Calculate your model in SPSS. What does it tell you about the coefficients?

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender	.345	.151	5.215	1	.022	1.412
	attsc7	.071	.068	1.101	1	.294	1.074
	mathsc1	.867	.081	114.165	1	.000	2.380
	Constant	-2.830	.359	62.212	1	.000	.059

a. Variable(s) entered on step 1: gender, attsc7, mathsc1.

In the box labelled 'sig' we can see that the variables gender and 'I'm one of the best in my class at maths' (mathsc1) are significant, while 'I like going to school' (attsc7), isn't. The regression coefficients are given under B, and show that an increase of one on the scale for mathsc1, for example, increases the probability of a pass on the outcome variable by .867.