

Module 4 - Multiple Logistic Regression

Objectives

- Understand the principles and theory underlying logistic regression
- Understand proportions, probabilities, odds, odds ratios, logits and exponents
- Be able to implement multiple logistic regression analyses using SPSS and accurately interpret the output
- Understand the assumptions underlying logistic regression analyses and how to test them
- Appreciate the applications of logistic regression in educational research, and think about how it may be useful in your own research

Start Module 4: Multiple Logistic Regression

Using multiple variables to predict dichotomous outcomes.

You can jump to specific pages using the contents list below. If you are new to this module start at the overview and work through section by section using the 'Next' and 'Previous' buttons at the top and bottom of each page. Be sure to tackle the exercise and the quiz to get a good understanding.

Contents

4.1 Overview

4.2 An introduction to Odds and Odds Ratios

Quiz A

4.3 A general model for binary outcomes

4.4 The logistic regression model

4.5 Interpreting logistic equations

4.6 How good is the model?

4.7 Multiple Explanatory Variables

4.8 Methods of Logistic Regression

4.9 Assumptions

4.10 An example from LSYPE

4.11 Running a logistic regression model on SPSS

4.12 The SPSS Logistic Regression Output

4.13 Evaluating interaction effects

4.14 Model diagnostics

4.15 Reporting the results of logistic regression

Quiz B

Exercise

4.1 Overview

What is Multiple Logistic Regression?


In the last two modules we have been concerned with analysis where the outcome variable (sometimes called the dependent variable) is measured on a continuous scale. However many of the variables we meet in education and social science more generally have just a few, maybe only two categories. Frequently we have only a dichotomous or binary outcome. For example this might be whether a student plans to continue in full-time education after age 16 or not, whether they have identified Special Educational Needs (SEN), whether they achieve a certain threshold of educational achievement, whether they do or do not go to university, etc. Note that here the two outcomes are *mutually exclusive* and one must occur. We usually code such outcomes as 0 if the event does not occur (e.g. the student does not plan to continue in FTE after the age of 16) and 1 if the event does occur (e.g. the student does plan to continue in FTE after age 16).

This module first covers some basic descriptive methods for the analysis of binary outcomes. This introduces some key concepts about percentages, proportions, probabilities, odds and odds-ratios. We then show how variation in a binary response can be modeled using regression methods to link the outcome to explanatory variables. In analyzing such binary outcomes we are concerned with modeling the *probability* of the event occurring given the level/s of one or more explanatory (independent/predictor) variables. This module is quite difficult because there are many new concepts in it. However if you persevere through the rationale below, you will find (hopefully!) that the examples make sense of it all. Also, like all things, the concepts and applications will grow familiar with use, so work through the examples and take the quizzes.

Running through the example of SPSS

As with previous modules you can follow us through the real-world examples that we use. You may now be familiar with the main adapted version of the LSYPE dataset but we also have a more specialized one for use with this module – the one we used in

the previous module. We recommend that you retrieve them from the ESDS website – playing around with them will really help you to understand this stuff!

LSYPE 15,000 

MLR LSYPE 15,000 

4.2 An introduction to Odds, Odds Ratios and Exponents

Let's start by considering a simple association between two dichotomous variables (a 2 x 2 crosstabulation) drawing on the LSYPE dataset. The outcome we are interested in is whether students aspire to continue in Full-time education (FTE) after the age of 16 (the current age at which students in England can choose to leave FTE). We are interested in whether this outcome varies between boys and girls. We can present this as a simple crosstabulation (**Figure 4.2.1**).

Figure 4.2.1: Aspiration to continue in full time education (FTE) after the age of 16 by gender: Cell counts and percentages

			Pupil wants to continue in FTE after age 16		Total
			0 No	1 Yes	
Gender	0 Male	Count	1837	6015	7852
		%	23.4%	76.6%	100.0%
	1 Female	Count	1003	6576	7579
		%	13.2%	86.8%	100.0%
Total		Count	2840	12591	15431
		%	18.4%	81.6%	100.0%

We have coded not aspiring to continue in FTE after age 16 as 0 and aspiring to do so as 1. Although it is possible to code the variable with any values, employing the values 0 and 1 has advantages. The mean of the variable will equal the proportion of cases with the value 1 and can therefore be interpreted as a probability. Thus we can see that the percentage of all students who aspire to continue in FTE after age 16 is 81.6%. This is equivalent to saying that the *probability* of aspiring to continue in FTE in our sample is 0.816.

Odds and odds ratios

However another way of thinking of this is in terms of the *odds*. Odds express the likelihood of an event occurring relative to the likelihood of an event not occurring. In our sample of 15,431 students, 12,591 aspire to continue in FTE while 2,840 do not aspire, so the odds of aspiring are $12591/2840 = 4.43:1$ (this means the ratio is 4.43 to 1, but we conventionally do not explicitly include the :1 as this is implied by the odds). The odds tell us that if we choose a student at random from the sample they

are 4.43 times more likely to aspire to continue in FTE than not to aspire to continue in FTE.

We don't actually have to calculate the odds directly from the numbers of students if we know the proportion for whom the event occurs, since the odds of the event occurring can be gained directly from this proportion by the formula (Where p is the probability of the event occurring.):

$$odds = \frac{P}{1 - P}$$

Thus the odds in our example are:

$$Odds = [p/(1-p)] = .816 / (1-.816) = .816 / .184 = 4.43.$$

The above are the *unconditional odds*, i.e. the odds in the sample as a whole. However odds become really useful when we are interested in how some other variable might affect our outcome. We consider here what the odds of aspiring to remain in FTE are separately for boys and girls, i.e. *conditional* on gender. We have seen the odds of the event can be gained directly from the proportion by the formula $odds = p/(1-p)$.

So for boys the odds of aspiring to continue in FTE = .766/(1-.766) = 3.27

While for girls the odds of aspiring to continue in FTE = .868/(1-.868) = 6.56.

These are the *conditional odds*, i.e. the odds depending on the condition of gender, either boy or girl.

We can see the odds of girls aspiring to continue in FTE are higher than for boys. We can in fact directly compare the odds for boys and the odds for girls by dividing one by the other to give the *Odds Ratio* (OR). If the odds were the same for boys and for girls then we would have an odds ratio of 1. If however the odds differ then the OR will depart from 1. In our example the odds for girls are 6.53 and the odds for boys are 3.27 so the $OR = 6.56 / 3.27 = 2.002$, or roughly 2:1. This says that girls are *twice as likely* as boys to aspire to continue in FTE.

Note that the way odd-ratios are expressed depends on the baseline or comparison category. For gender we have coded boys=0 and girls =1, so the boys are our natural base group. However if we had taken girls as the base category, then the odds ratio would be $3.27 / 6.56 = 0.50:1$. This implies that boys are *half as likely* to aspire to continue in FTE as girls. You will note that saying “Girls are twice as likely to aspire as boys” is actually identical to saying “boys are half as likely to aspire as girls”. Both figures say the same thing but just differ in terms of the base.

Odds Ratios from 0 to just below 1 indicate the event is *less likely* to happen in the comparison than in the base group, odds ratios of 1 indicate the event is *exactly as likely* to occur in the two groups, while odds ratios from just above 1 to infinity indicate the event is *more likely* to happen in the comparator than in the base group.

Extension D provides a table that shows the equivalence between ORs in the range 0 to 1 with those in the range 1 to infinity.

Seeing the relationship as a model

An interesting fact can be observed if we look at the odds for boys and the odds for girls in relation to the odds ratio (OR).

$$\text{For boys (our base group) the odds} = 3.27 * 1 = 3.27$$

$$\text{For girls the odds} = 3.27 * 2.002 = 6.56.$$

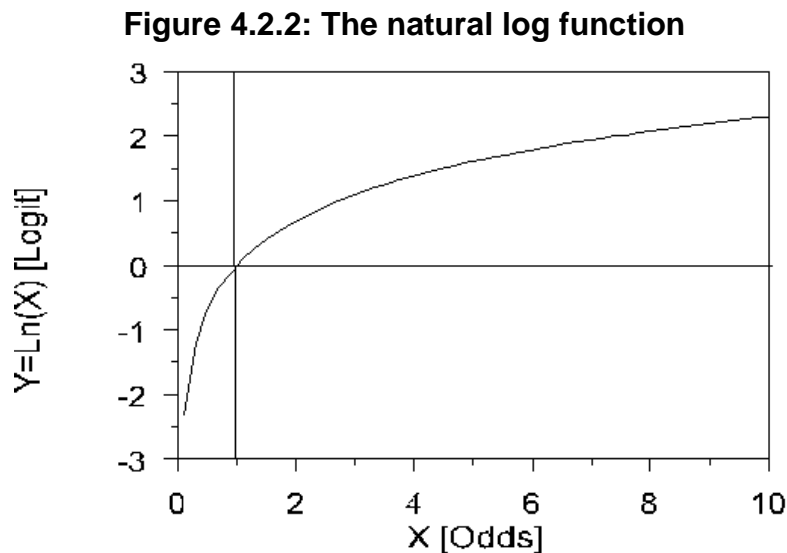
So another way of looking at this is that the odds for each gender can be expressed as a constant multiplied by a gender specific multiplicative factor (namely the OR).

$$p/(1-p) = \text{constant} * \text{OR}.$$

However there are problems in using ORs directly in any modeling because they are asymmetric. As we saw in our example above, an OR of 2.0 indicates the same relative ratio as an OR of 0.50, an OR of 3.0 indicates the same relative ratio as an OR of 0.33, an OR of 4.0 indicates the same relative ratio as an OR of 0.25 and so on. This asymmetry is unappealing because ideally the odds for males would be the opposite of the odds for females.

Odds, Log odds and exponents

This asymmetry problem disappears if we take the 'log' of the OR. 'Log' doesn't refer to some sort of statistical deforestation... rather a mathematical transformation of the odds which will help in creating a regression model. Taking the log of an OR of 2 gives the value $\text{Log}(2) = +0.302$ and taking the log of an OR of 0.50 gives the value $\text{Log}(0.5) = -0.302$. See what's happened? The Log of the OR, sometimes called the logit (pronounced 'LOH-jit', word fans!) makes the relationships symmetric around zero (the OR's become plus and minus .302). Logits and ORs contain the same information, but this difference in mathematical properties makes logits better building blocks for logistic regression. But what is a log function? How does it transform the ORs? Well, the natural log function looks like this (**Figure 4.2.2**):



So if we take the log of each side of the equation we can then express the log odds as:

$$\text{Log} [p/(1-p)] = \text{constant} + \text{log} (OR)$$

If the constant is labelled a , the log of the OR is labelled b , and the variable gender (x) takes the value 0 for boys and 1 for girls, then:

$$\text{Log} [p/(1-p)] = a + bx$$

Note that taking the log of the odds has converted this from a multiplicative to an additive relationship with the *same form as the linear regression equations we have*

discussed in the previous two modules (it is not essential, but if you want to understand how logarithms do this it is explained in **Extension E**). So the log of the odds can be expressed as an additive function of $a + bx$. This equation can be generalised to include any number of explanatory variables:

$$\text{Log} [p/(1-p)] = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n.$$

Output from a logistic regression of gender on educational aspiration

If we use SPSS to complete a logistic regression (more on this later) using the student level data from which the summary **Figure 4.2.1** was constructed, we get the logistic regression output shown below (**Figure 4.2.3**).

Figure 4.2.3: Output from a logistic regression of gender on aspiration to continue in FTE post 16

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	gender (1)	.694	.043	259.217	1	.000	2.002
	Constant	1.186	.027	1979.810	1	.000	3.274

Let’s explain what this output means. The B weights give the linear combination of the explanatory variables that best predict the log odds. So we can determine that the log odds for:

Male: $\text{Log} [p/(1-p)] = 1.186 + (0.694 * 0) = 1.186$

Female: $\text{Log} [p/(1-p)] = 1.186 + (0.694 * 1) = 1.880$

The inverse of the log function is the *exponential* function, sometimes conveniently also called the *anti-logarithm* (nice and logical!). So if we want to convert the log odds back to odds we take the *exponent* of the log odds. So the odds for our example are:

Male: $\text{Exp} (1.186) = 3.27$

Female: $\text{Exp} (1.880) = 6.55$

The odds ratio is given in the SPSS output for the gender variable [indicated as Exp(B)] showing that girls are twice as likely as boys to aspire to continue in FTE.

By simple algebra we can rearrange the formula $\text{odds} = [p/(1-p)]$ to solve for probabilities, i.e. $p = [\text{odds}/(1+\text{odds})]$:

Males: $p = 3.27 / (1+3.27) = .766$

$$\text{Females: } p = 6.55 / (1 + 6.55) = .868.$$

These probabilities, odds and odds ratios - derived from the logistic regression model - are identical to those calculated directly from **Figure 4.2.1**. This is because we have just one explanatory variable (gender) and it has only two levels (girls and boys). This is called a *saturated* model for which the expected counts and the observed counts are identical. The logistic regression model will come into its own when we have an explanatory variable with more than two values, or where we have multiple explanatory variables. However what we hope this section has done is show you how probabilities, odds, and odds ratios are all related, how we can model the proportions in a binary outcome through a linear prediction of the log odds (or logits), and how these can be converted back into odds ratios for easier interpretation.

Take the quiz to check you are comfortable with what you have learnt so far. If you are not perturbed by maths and formulae why not check out **Extension E** for more about logs and exponents.

4.3 A general model for binary outcomes

The example we have been using until now is very simple because there is only one explanatory variable (gender) and it has only two levels (0=boys, 1=girls). With this in mind, why should we calculate the logistic regression equation when we could find out exactly the same information directly from the cross-tabulation? The value of the logistic regression equation becomes apparent when we have multiple levels in an explanatory variable or indeed multiple explanatory variables. In logistic regression we are not limited to simple dichotomous independent variables (like gender) we can also include ordinal variables (like socio-economic class) and continuous variables (like age 11 test score). So the logistic regression model lets us extend our analysis to include multiple explanatory variables of different types.

We can think of the data in **Figure 4.2.1 (Page 4.2)** in two ways. One is to think of them as two proportions, the proportion of students who aspire to continue in FTE in two independent samples, a sample of boys and a sample of girls. The other way is to think of the data as 13,825 observations, with the response always either 0 (does not wish to continue in FTE) or 1 (wishes to continue in FTE). Thus our response or outcome distribution is actually what is known as a *binomial distribution* since it is made up of only two values, students either do not aspire (0) or do aspire (1) to continue in FTE.

As another way to consider the logic of *logistic regression*, consistent with what we have already described but coming at it from a different perspective, let's consider first why we cannot model a binary outcome using the *linear regression* methods we covered in modules 2 and 3. We will see that significant problems arise in trying to use linear regression with binary outcomes, which is why logistic regression is needed.

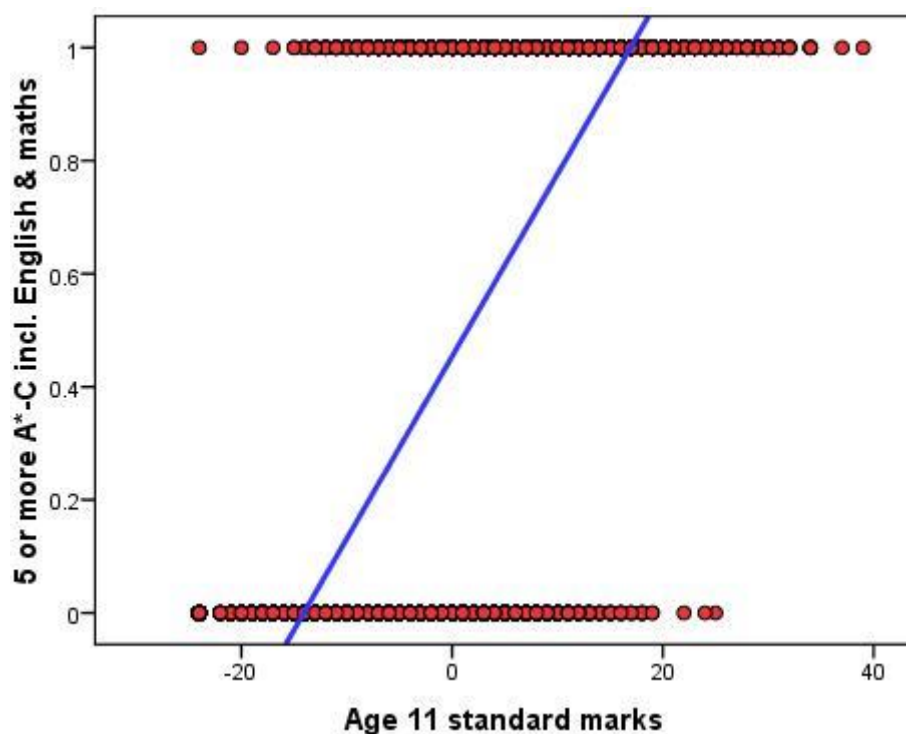
The problem with linear regression for binary outcomes

A new example: Five or more GCSE passes at A*-C including English and maths

For this example let us take as our outcome as whether a student achieves the conventional measure of exam success in England, which is achieving five or more GCSE passes at grades A*-C, including English and maths (*fiveem*). This is a

frequently used measure of a student's 'success' in educational achievement at age 16. It is also used at an institutional level in school performance tables in England by reporting the proportion of students in each secondary school achieving this threshold, attracting considerable media attention. Our variable *fiveem* is coded '0' if the student did not achieve this threshold and '1' if the student did achieve it. We want to predict the probability of achieving this outcome depending on test score at age 11. We can fit a *linear regression* to this binary outcome as shown in **Figure 4.3.1** below.

Figure 4.3.1: A linear regression of age 11 test score against achieving five or more GCSE grades A*-C including English and maths (*fiveem*)



The linear regression of age 11 score on *fiveem* give the following regression equation:

$$\hat{Y} = .454 + .032 * X \text{ (where } X = \text{age 11 score which can range from } -24 \text{ to } 39\text{).}$$

The predicted values take the form of proportions or probabilities. Thus at the average age 11 score (which was set to 0, see **Extension A**) the predicted probability is simply the intercept or .454 (i.e. 45.4%). At an age 11 score 1 SD below the mean ($X = -10$) the predicted probability = $.454 + .032 * -10 = .134$, or 13.4%. At an age 11 score 1 SD above the mean ($X = 10$) the predicted probability = $.454 + .032 * 10 = .774$, or 77.4%.

However there are two problems with linear regression that make it inappropriate to use with binary outcomes. One problem is conceptual and the other statistical.

Lets deal with the statistical problem first. The problem is that a binary outcome violates the assumption of *normality* and *homoscedasticity* inherent in linear regression. Remember that linear regression assumes that most of the observed values of the outcome variable will fall close to those predicted by the linear regression equation and will have an approximately *normal distribution* (See **Page 2.6**). Yet with a binary outcome only two Y values exist so there can only be two residuals for any value of X, either 1, predicted value (when Y=1) or 0, predicted value (when Y=0). The assumption of normality clearly becomes nonsensical with a binary response, the distribution of residuals cannot be normal when the distribution only has two values. It also violates the assumption of *homoscedasticity*, namely that the variance of errors is constant at all levels of X. If you look at **Figure 4.3.1** it is apparent that near the lower and upper extremes of X, where the line comes close to the floor of 0 and the ceiling of 1, the residuals will be relatively small, but near the middle values of X the residuals will be relatively large. Thus there are good statistical reasons for rejecting a linear regression model for binary outcomes.

The second problem is conceptual. Probabilities and proportions are different from continuous outcomes because they are bounded by a minimum of 0 and a maximum of 1, and by definition probabilities and proportions cannot exceed these limits. Yet the linear regression line can extend upwards beyond 1 for large values of X and downwards below 0 for small values of X. Look again at **Figure 4.3.1**. We see that students with an age 11 score below -15 are actually predicted to have a less than zero (<0) probability of achieving *fiveem*. Equally problematic are those with an age 11 score above 18 who are predicted to have a probability of achieving *fiveem* greater than one (>1). These values simply make no sense.

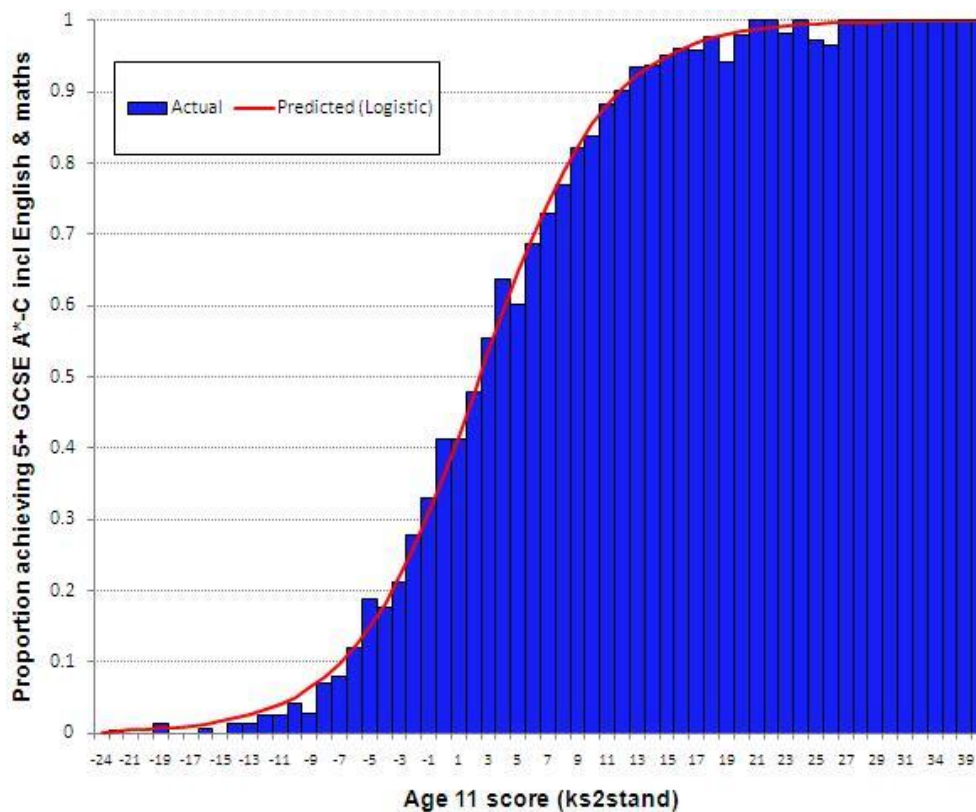
We could attempt a solution to the boundary problem by assuming that any predicted values of $Y > 1$ should be truncated to the maximum of 1. The regression line would be straight to the value of 1, but any increase in X above this point would have no influence on the predicted outcome. Similarly any predicted values of $Y < 0$ could be

truncated to 0 so any decrease in X below this point would have no influence on the predicted outcome. However there is another functional form of the relationship between X and a binary Y that makes more theoretical sense than this 'truncated' linearity. Stay tuned...

4.4 The logistic regression model

To see this alternative form for the relationship between age 11 score and *fiveem*, let us plot the actual data from LSYPE. In **Figure 4.4.1** the blue bars show the actual proportion of students with each age 11 test score that achieved *fiveem*.

Figure 4.4.1: Probability of 5+ GCSE A*-C including English & maths by age 11 test score



We can see that the relationship between age 11 score and *fiveem* actually takes the form of an S shaped curve (a 'sigmoid'). In fact whenever we have a binary outcome (and are thus interested in modeling proportions) the sigmoid or S shaped curve is a better function than a linear relationship. Remember that in linear regression a one unit increase in X is assumed to have the same impact on Y wherever it occurs in the distribution of X. However the S shape curve represents a nonlinear relationship between X and Y. While the relationship is approximately linear between probabilities of 0.2 and 0.8, the curve levels off as it approaches the ceiling of 1 and the floor of 0. The effect of a unit change in age 11 score on the predicted probability is relatively

small near the floor and near the ceiling compared to the middle. Thus a change of 2 or 3 points in age 11 score has quite a substantial impact on the probability of achieving *fiveem* around the middle of the distribution, but much larger changes in age 11 score are needed to effect the same change in predicted probabilities at the extremes. Conceptually the S-shaped curve makes better sense than the straight line and is far better at dealing with probabilities.

The logistic function

There are many ways to generate an S shaped curve mathematically, but the logistic function is the most popular and easiest to interpret. A function is simply a process which transforms data in a systematic way – in this example it transforms log odds into a proportion. We described on **Page 4.2** that there can be a linear and additive relationship between our explanatory variables and the log odds of the event occurring, so we can predict the log odds from our explanatory variables.

$$\text{Log } [p/(1-p)] = a + bx.$$

The logistic function transforms the log odds to express them as predicted probabilities. First, it applies the reverse of the log (called the exponential or anti-logarithm) to both sides of the equation, eliminating the log on the left hand side, so the odds can be expressed as:

$$p/(1-p) = \text{Exp}^{(a+bx)}.$$

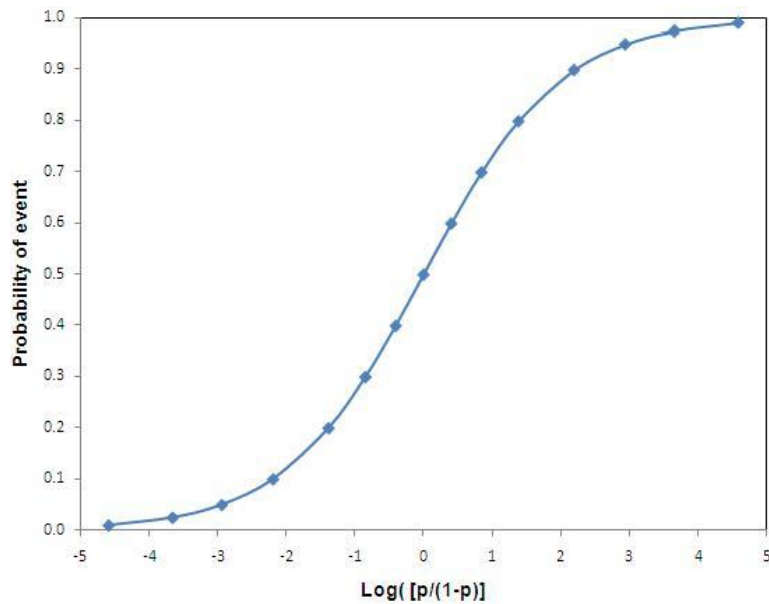
Second, the formula can be rearranged by simple algebra¹ to solve for the value p.

$$p = \text{Exp}^{(a+bx)} / [1 + \text{Exp}^{(a+bx)}]$$

So the logistic function transforms the log odds into predicted probabilities. **Figure 4.4.2** shows the relationship between the log odds (or logit) of an event occurring and the probabilities of the event as created by the logistic function. This function gives the distinct S shaped curve.

¹ You will have to take my word on this, but a formula $p / (1-p) = x$ can be rearranged to $p = x / (1+x)$.

Figure 4.4.2: The logistic function



Look back to **Figure 4.4.1** where the blue bars shows the actual proportion of students achieving *fiveem* for each age 11 test score. We have superimposed over the actual figures a red line that shows the predicted probabilities of achieving *fiveem* as modeled from a logistic regression using age 11 test score as the explanatory variable. Comparing these predicted probabilities (red line) to the actual probability of achieving *fiveem* (blue bars) we can see that the modeled probabilities fit the actual data extremely well.

Why do we model the log odds rather than probabilities or proportions?

The log odds are more appropriate to model than probabilities because log odds do not have the floor of 0 and the ceiling of 1 inherent in probabilities. Remember the probability of an event occurring cannot be <0 or >1 . What the log odds does is to 'stretch' the proportion scale to eliminate these floor and ceiling effects. They do this by (i) transforming the probabilities to odds, and (ii) taking the log of the odds.

Odds remove the ceiling

We saw in **Figure 4.4.1** that there is a non-linear relationship between X and Y - for example, we need larger changes in X to effect the same proportionate increase in Y at the ceiling compared to near the middle. Odds can model this property because larger changes in odds are needed to effect the same change in the probabilities

when we are at the ceiling than at the middle of the curve. Let's look at specific figures using **Figure 4.4.3** which shows the relationship between probabilities, odds and log odds.

Figure 4.4.3: Probabilities, odds and log odds

p	odds[p/(1-p)]	Log (Odds) or logit
0.01	0.01	-4.60
0.05	0.05	-2.94
0.1	0.11	-2.20
0.2	0.25	-1.39
0.3	0.43	-0.85
0.4	0.67	-0.41
0.5	1.00	0.00
0.6	1.50	0.41
0.7	2.33	0.85
0.8	4.00	1.39
0.9	9.00	2.20
0.95	19.00	2.94
0.99	99.00	4.60

A change in the probability of an event occurring from .5 to .6 is associated with a change in odds from 1.0 to 1.5 (an increase of 0.5 in the odds). However a similar change in probability from .8 to .9 reflects a much larger change in the odds from 4.0 to 9.0 (an increase of 5 in the odds). Thus modeling the odds reflects the fact that we need larger changes in X to effect increases in proportions near the ceiling of the curve than we do at the middle.

Log odds remove the floor (as well as the ceiling)

Transforming probabilities into odds has eliminated the ceiling of 1 inherent in probabilities, but we are still left with the floor effect at 0, since odds, just like proportions, can never be less than zero. However taking the log of the odds also removes this floor effect.

- The log of odds below 1 produce negative numbers
- The log of odds equal to 1 produce 0
- The log of odds above 1 produce positive numbers

The log odds still has the non-linear relationship with probability at the ceiling, since a change in p from .5 to .6 is associated with an increase in log odds from 0 to 0.4, while a change in probability from .8 to .9 is associated with an increase in log odds from 1.39 to 2.20 (or 0.8). However they also reflect the non-linear relationship at the floor, since a decrease in probability from .5 to .4 is associated with a decrease in log odds from 0 to -0.4, while a decrease in probability from .2 to .1 is associated with a decrease in log odds from -1.39 to -2.20 (or -0.8) (see **Figure 4.4.3**). Also, as we discussed on **Page 4.2** log odds have the advantage that they are symmetrical around 0. A probability of 0.8 that an event will occur has log odds of 1.39, and the probability of 0.2 that the event will not occur has log odds of -1.39. So the log odds are symmetrical around 0.

4.5 Interpreting logistic equations

We have seen that if we try to predict the probabilities directly we have the problem of non-linearity, specifically the floor at 0 and the ceiling at 1 inherent in probabilities. But if we use our explanatory variables to predict the log odds we do not have this problem. However while we can apply a linear regression equation to predict the log odds of the event, people have a hard time understanding log odds (or logits).

Remember that a logit is just a log of the odds, and odds are just a function of p (the probability of a 1). We can convert the log odds back to odds by applying the reverse of the log which is called the *exponential* (sometimes called the *anti-logarithm*) to both sides. Taking the exponent eliminates the log on the left handside so the odds can be expressed as:

$$p/(1-p) = \text{Exp}(^{a+bx}).$$

We can also rearrange this equation to find the probabilities as:

$$p = \text{Exp}(^{a+bx}) / [1 + \text{Exp}(^{a+bx})]$$

which is the logistic function, which converts the log odds to probabilities.

So now rather than log odds or logits, which people are not happy talking about, we can talk about odds and probabilities, which people are happier talking about (at least relatively!).

Interpreting the logistic regression for our example *Fiveem*

So let's return to our example of modelling the probability of achieving five or more GCSE A*-C grades including English & maths (*fiveem*) from age 11 test score.

The SPSS logistic regression output is shown in the table below.

Figure 4.5.1: Logistic regression for *Fiveem* by age 11 score

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	ks2stand	.235	.004	3513.724	1	.000	1.265
	Constant	-.337	.023	210.536	1	.000	.714

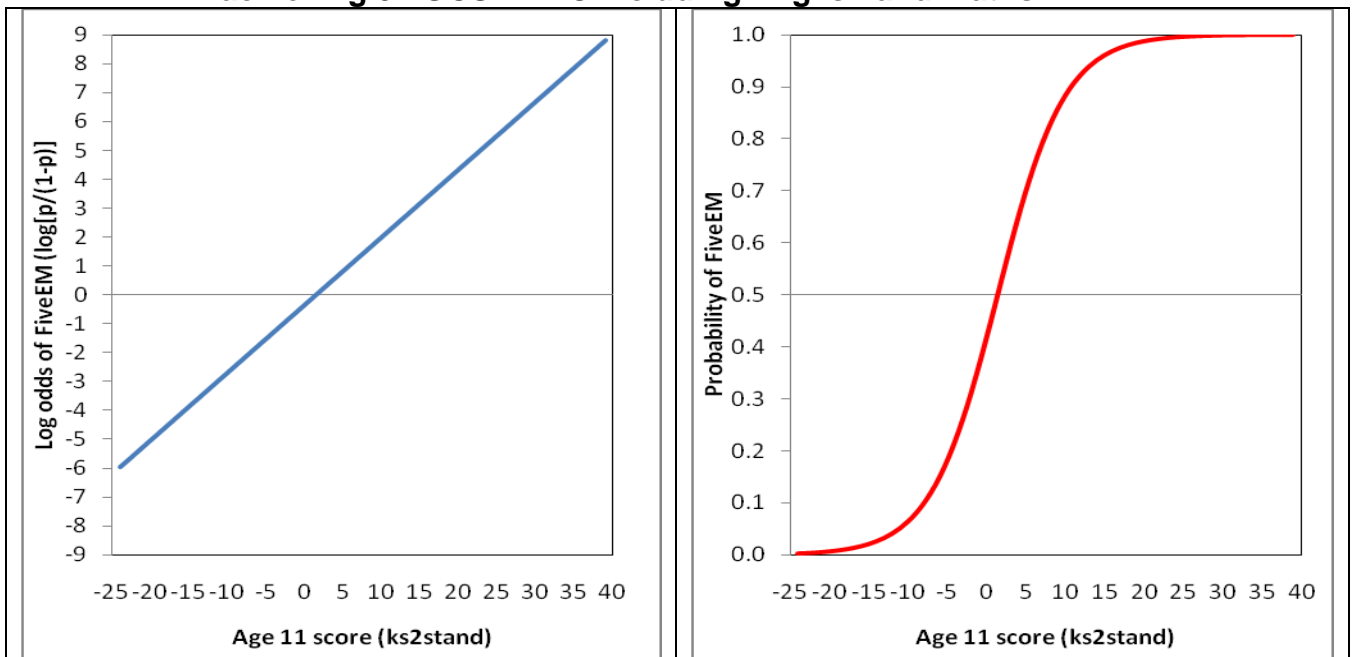
a. Variable(s) entered on step 1: ks2stand.

The B coefficients describe the logistic regression equation using age 11 score to predict the log odds of achieving *fiveem*, thus the logistic equation is:

$$\log [p/(1-p)] = -.337 + .235 * \text{age 11 score}.$$

Figure 4.5.2 lets us visualize the equation. The left hand chart shows the linear relationship between age 11 score and the log odds of achieving *fiveem*. This line has an intercept of $-.337$ and a slope of $.235$ and is clearly linear. However we can use the logistic function to transform the log odds to predicted probabilities, which are shown in the right hand chart. Looking back to **Figure 4.4.1** on **Page 4.4** we see how well these predicted probabilities match the actual data on the proportion of pupils achieving *fiveem* at each age 11 score.

Figure 4.5.2: Relationship between age 11 score and (a) the log odds of achieving 5+ GCSE A*-C including English & maths (b) the probability of achieving 5+ GCSE A*-C including English and maths.



The logistic regression equation indicates that a one unit increase in age 11 test score is associated with a $.235$ increase in the log odds of achieving *fiveem*. Taking the exponent of the log odds, indicated in the output as $\text{Exp}(B)$, gives the Odds Ratio, which shows that a one unit increase in age 11 test score increases the odds of achieving *fiveem* by a multiplicative factor of 1.27 . Various procedures also exist to calculate the effects of a unit change in the b on the probability of Y occurring.

However the effect on probabilities depends on the point of the logistic curve at which the effect is calculated (e.g. a one unit change in age 11 score from -10 to -9 would give a different change in the predicted probability than a one unit change in age 11 score from 0 to 1). This is why we typically stick to ORs as the main way of interpreting the logistic regression results. (For more detail on interpreting the age 11 coefficient see **Pages 4.10 and 4.12**).

Summary

So in summary we have seen that when attempting to predict probabilities (which we are doing when we model binary outcomes) linear regression is inappropriate, both for statistical and conceptual reasons. With binary outcomes the form of the relationship between an explanatory variable X and the probability of Y is better modeled by an S-shaped curve. While the relationship between X and the probability of Y is non-linear (it is in fact curvilinear), and therefore cannot be modeled directly as a linear function of our explanatory variables, there can be a linear and additive combination of our explanatory variables that predict the log odds, which are not restricted by the floor of 0 and ceiling of 1 inherent in probabilities. These predicted log odds can be converted back to odds (by taking the exponential) and to predicted probabilities using the logistic function.

How are the logistic regression coefficients computed?

In logistic regression, the regression coefficients deal in probabilities so they cannot be calculated in the same way as they are in linear regression. While in theory we could do a linear regression with logits as our outcome, we don't actually have logits for each individual observation we just have 0's or 1's. The regression coefficients have to be estimated from the pattern of observations (0's and 1's) in relation to the explanatory variables in the data. We don't need to be concerned with exactly how this works but the process of *maximum likelihood estimation* (MLE) starts with an initial arbitrary "guesstimate" of what the logit coefficients should be. The MLE seeks to manipulate the b 's to maximize the log likelihood (LL) which reflects how likely it is (i.e. the log odds) that the observed values of the outcome may be predicted from the explanatory variables. After this initial function is estimated the residuals are tested and a re-estimate is made with an improved function and the process is repeated

(usually about half a dozen times) until convergence is reached (that is until the improvement in the LL does not differ significantly from zero).

4.6 How good is the model?

The Deviance (-2LL) statistic

We will need to ascertain how good our regression model is once we have fitted it to the data – does it accurately explain the data, or does it incorrectly classify cases as often as it correctly classifies them? The deviance, or -2 log-likelihood (-2LL) statistic, can help us here. The deviance is basically a measure of how much unexplained variation there is in our logistic regression model – the higher the value the less accurate the model. It compares the difference in probability between the predicted outcome and the actual outcome for each case and sums these differences together to provide a measure of the total error in the model. This is similar in purpose to looking at the total of the residuals (the sum of squares) in linear regression analysis in that it provides us with an indication of how good our model is at predicting the outcome. The -2LL statistic (often called the deviance) is an indicator of how much unexplained information there is after the model has been fitted, with large values of -2LL indicating poorly fitting models. Don't worry about the technicalities of this – as long as you understand the basic premise you'll be okay!

The deviance has little intuitive meaning because it depends on the sample size and the number of parameters in the model as well as on the goodness of fit. We therefore need a standard to help us evaluate its relative size. One way to interpret the size of the deviance is to compare the value for our model against a 'baseline' model. In linear regression we have seen how SPSS performs an ANOVA to test whether or not the model is better at predicting the outcome than simply using the mean of the outcome. The change in the -2LL statistic can be used to do something similar: to test whether the model is significantly more accurate than simply always guessing that the outcome will be the more common of the two categories. We use this as the baseline because in the absence of any explanatory variables the 'best guess' will be the category with the largest number of cases.

Let's clarify this with our *fiveem* example. In our sample 46.3% of student achieve *fiveem* while 53.7% do not. The probability of picking at random a student who does not achieve the *fiveem* threshold is therefore slightly higher than the probability of

picking a student who does. If you had to pick one student at random and guess whether they would achieve *fiveem* or not, what would you guess? Assuming you have no other information about them, it would be most logical to guess that they would not achieve *fiveem* – simply because a slight majority do not. This is the baseline model which we can test our later models against. This is also the logistic model when only the constant is included. If we then add explanatory variables to the model we can compute the improvement as follows:

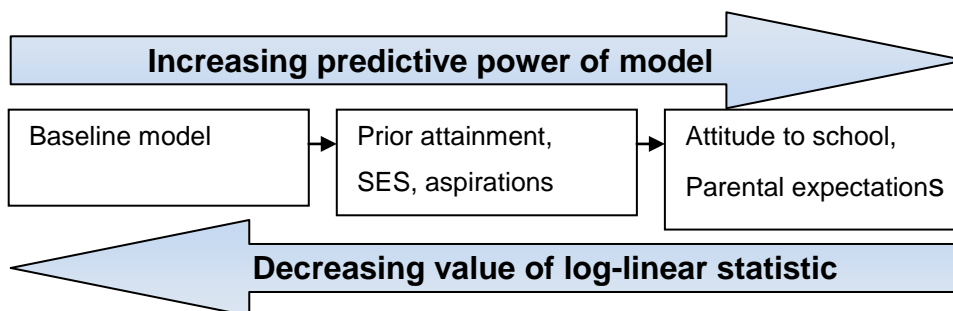
$$X^2 = [-2LL (baseline)] - [-2LL (new)]$$

with *degrees of freedom* = $k_{baseline} - k_{new}$, where k is the number of parameters in each model.

If our new model explains the data better than the baseline model there should be a significant reduction in the deviance (-2LL) which can be tested against the chi-square distribution to give a p value. Don't worry - SPSS will do this for you! However if you would like to learn more about the process you can go to **Extension F**.

The deviance statistic is useful for more than just comparing the model to the baseline - you can also compare different variations of your model to see if adding or removing certain explanatory variables will improve its predictive power (**Figure 4.6.1**)! If the deviance (-2LL) is decreasing to a statistically significant degree with each set of explanatory variables added to the model then it is improving at accurately predicting the outcome for each case.

Figure 4.6.1 Predictors of whether or not student goes to university



R^2 equivalents for logistic regression

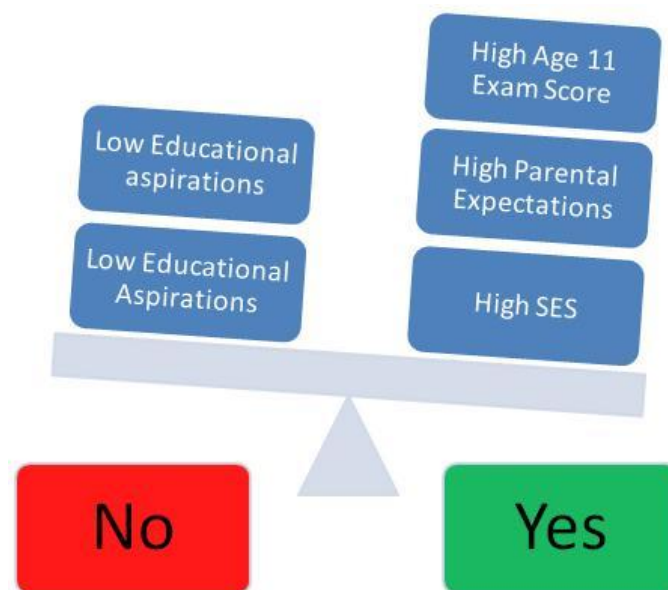
Another way of evaluating the effectiveness of a regression model is to calculate how strong the relationship between the explanatory variable(s) and the outcome is. This was represented by the R^2 statistic in linear regression analysis. R^2 , or rather a form of it, can also be calculated for logistic regression. However, somewhat confusingly there is more than one version! This is because the different versions are pseudo- R^2 statistics that approximate the amount of variance explained rather than calculate it precisely. Remember we are dealing with probabilities here! Despite this it can still sometimes be useful to examine them as a way of ascertaining the substantive value of your model.

The two versions most commonly used are *Hosmer & Lemeshow's R^2* and *Nagelkerke's R^2* . Both describe the proportion of variance in the outcome that the model successfully explains. Like R^2 in multiple regression these values range between '0' and '1' with a value of '1' suggesting that the model accounts for 100% of variance in the outcome and '0' that it accounts for none of the variance. Be warned: they are calculated differently and may provide conflicting estimates! These statistics are readily available through SPSS and we'll show you how to interpret them when we run through our examples over the next few pages.

4.7 Multiple explanatory variables

As with linear regression, the more information we have to predict the outcome the more likely it is that we will be able to develop good models. Like multiple linear regression, multiple logistic regression allows the researcher to add many *explanatory variables* to the model. For example, if we know about the student's prior attainment, their gender, their attitude to school, their socio-economic status, their parent's expectations for them and so on, we can use all the explanatory variables together to better estimate which category of the outcome variable they will most likely fall into (see for example **Figure 4.7.1** below).

Figure 4.7.1 Multiple explanatory variables used to make classifications for binary variables



Of course this will be true only if our additional explanatory variables actually add significantly to the prediction of the outcome! As in linear regression, we want to know not only how well the model overall fits the data, but also the individual contributions of the explanatory variables. SPSS will calculate standard errors and significance values for all variables added to our model, so we can judge how much they have added to the prediction of the outcome.

Statistical significance of explanatory variables

As in linear regression we want to know not only how the model overall fits the data but also the individual contribution of the explanatory variables. The use of the Wald statistic is analogous to the t-test performed on the regression coefficients in linear regression to test whether the variable is making a significant contribution to the prediction of the outcome, specifically whether the explanatory variable's coefficient is significantly different from zero. SPSS calculates and reports the Wald statistic and importantly the associated probability (p-value). Some caution is necessary however when interpreting the Wald statistic. If the coefficient (B) is large then this can result in the standard error becoming disproportionately large which can lead to an inaccurately small Wald statistic. This can result in false conclusions being drawn about the contribution of the explanatory variable to the model (you are more likely to reject the significance of an explanatory variable that is actually important). The Wald statistic is a useful indicator of whether or not an explanatory variable is important but should be interpreted with care! If in doubt you should compare the deviance (-2LL) of a model including the explanatory variable to a previous model without the variable to see whether the reduction in -2LL is statistically significant. We will show you an example of this later.

Effect size of explanatory variables

The above tells us whether an explanatory variable makes a statistically significant contribution to predicting the outcome, but we also want to know the size or magnitude of the association. In linear regression the regression coefficients (b) are the increase in Y for a one unit increase in X. In logistic regression we are not predicting a continuous dependent variable but the log odds of the outcome occurring. Thus in logistic regression the b coefficient indicates the increase in the log odds of the outcome for a one unit increase in X. However as we have seen these log odds (or logits) do not provide an intuitively meaningful scale to interpret the change in the dependent variable. Taking the exponent of the log odds allows interpretation of the coefficients in terms of Odds Ratios which are substantive to interpret. Helpfully, SPSS gives this OR for the explanatory variable labeled as $\text{Exp}(B)$.

Dichotomous (or dummy) explanatory variables

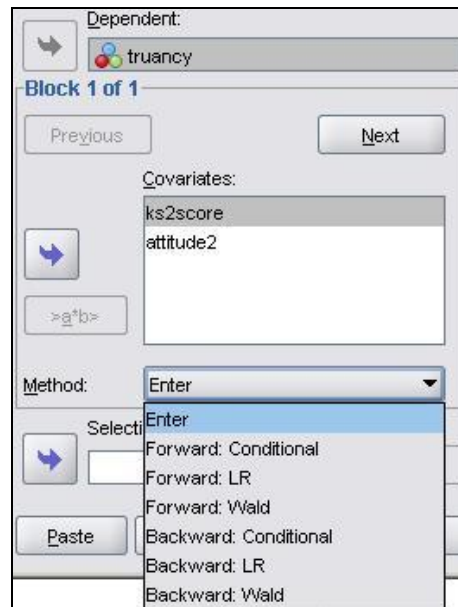
For a dichotomous explanatory variable the OR is simply the difference between the odds for the base category ($x=0$) and the other category ($x=1$). Thus in our earlier example for gender and aspirations the OR was 2.0 indicating girls ($x=1$) were twice as likely as boys ($x=0$) to aspire to continue in FTE. While the OR is sufficient for meaningful interpretation, some researchers also like to express the OR in percentage terms. Subtracting 1 from the OR and multiplying by 100 gives the percentage change. Thus $(2-1) * 100 =$ a 100% increase in the odds.

Continuous explanatory variables

What about a continuous predictor, like age 11 score? **Figure 4.5.1** (on **Page 4.5**) indicates the OR $[\text{Exp}(B)]$ associated with age 11 score is 1.27, thus a one unit increase in age 11 standard score increases the odds of achieving *fiveem* by a factor of 1.27 or 27%. Given that age 11 score is a continuous variable that has a standard deviation (SD) of 10, it would be more meaningful to compare the increase in odds associated with a 1SD change, or a 10 unit increase in age 11 score. If we multiply the SD by the b coefficient of .235, there is a 2.35 increase in the log odds for a 1 SD change in age 11 score. Remembering that to convert this into a statement about the odds of the outcome - rather than the log odds - we have to take the exponential, so $\text{Exp}(2.35)$ or $e^{2.35} = 10.5$. So a 1 SD increase in age 11 score increases the odds of achieving *fiveem* by a factor of 10.5. If we wanted to calculate this as a percentage change then $(10.5-1)*100=$ a 950% change in the odds. Wow!

4.8 Methods of Logistic Regression

As with linear regression we need to think about how we enter explanatory variables into the model. The process is very similar to that for multiple linear regression so if you're unsure about what we're referring to please check the section entitled 'methods of regression' on **Page 3.2**. The control panel for the method of logistic regression in SPSS is shown below.



As you can see it is still possible to group the explanatory variables in blocks and to enter these blocks in to the model in order of importance. Thus the above screen shot show we are at 'Block 1 of 1', but we can use the 'Next' button to set up a second block if we want to. The 'Enter' option should also be familiar - when selected, all explanatory variables (here labeled 'covariates' by SPSS – just to add an extra little challenge!) in the specific block are forced into the model simultaneously.

The main difference for logistic regression is that the automated 'stepwise' entry methods are different. Once again the forward and backward methods are present. They differ in how they construct the regression model, with the forward method adding explanatory variables to a basic model (which includes only the constant, B_0) and the backwards method removing explanatory variables from the full model (one including all the specified explanatory variables). SPSS makes these decisions based on whether the explanatory variables meet certain criteria. You can choose three different types of criteria for both forward and backward stepwise entry methods:

'Conditional', 'LR' and 'Wald'. 'LR' stands for Likelihood Ratio which is considered the criterion least prone to error.

We haven't gone into too much detail here partly because stepwise methods confuse us but mainly because they are not generally recommended. They take important decisions away from the researcher and base them on mathematical criteria rather than sound theoretical logic. Stepwise methods are only really recommended if you are developing a theory from scratch and have no empirical evidence or sensible theories about which explanatory variables are most important. Most of the time we have some idea about which predictors are important and the relative importance of each one, which allows us to specify the entry method for the regression analysis ourselves.

4.9 Assumptions

You will find that the assumptions for logistic regression are very similar to the assumptions for linear regression. If you need a recap, rather than boring you by repeating ourselves like statistically obsessed parrots (the worst kind of parrot) we direct you to our multiple regression assumptions on **Page 3.3**. However, there are still three key assumptions which you should be aware of:

Linearity (sort of...): For linear regression the assumption is that the outcome variable has a linear relationship with the explanatory variables, but for logistic regression this is not possible because the outcome is binary. The assumption of linearity in logistic regression is that any explanatory variables have a linear relationship with the *logit* of the outcome variable. ‘What are they on about now?’ we imagine you’re sighing. If the relationship between the log odds of the outcome occurring and each of the explanatory variables is not linear then our model will not be accurate. We’ll discuss how to evaluate this in the context of SPSS over the coming pages, but the best way to check that the model you are creating is sensible is by looking at the model fit statistics and pseudo R^2 . If you are struggling with the concept of logits and log odds you can revise **Pages 4.2 and 4.4** of this module.

Independent errors: Identical to linear regression, the assumption of *independent errors* states that errors should not be correlated for two observations. As we said before in the simple linear regression module, this assumption can often be violated in educational research where pupils are *clustered* together in a hierarchical structure. For example, pupils are clustered within classes and classes are clustered within schools. This means students within the same school often have a tendency to be more similar to each other than students drawn from different schools. Pupils learn in schools and characteristics of their schools, such as the school ethos, the quality of teachers and the ability of other pupils in the school, may affect their attainment. In large scale studies like the LSYPE such clustering can to some extent be taken care of by using design weights which indicate the probability with which an individual case was likely to be selected within the sample. Thus published analyses of LSYPE (see **Resources Page**) specify school as a cluster variable and apply published design weights using the SPSS *complex samples* module. More generally researchers can

control for clustering through the use of multilevel regression models (also called hierarchical linear models, mixed models, random effects or variance component models) which explicitly recognize the hierarchical structure that may be present in your data. Sounds complicated, right? It definitely can be and these issues are more complicated than we need here where we are focusing on understanding the essentials of logistic regression. However if you feel you want to develop these skills we have an excellent sister website provided by another NCRM supported node called **LEMMA** which explicitly provides training on using multilevel modeling including for logistic regression. We also know of some good introductory texts on multilevel modelling and you can find all of this among our **Resources**.

Multicollinearity: This is also identical to multiple regression. The assumption requires that predictor variables should not be highly correlated with each other. Of course predictors are often correlated with each other to some degree. As an example, below is a correlation matrix (**Figure 4.9.1**) that shows the relationships between several *LSYPE* variables.

Figure 4.9.1: Correlation Matrix – searching for multicollinearity issues

		Correlations			
		SEC of head of household	IDACI score	KS2 score	Evening homework per week
SEC of head of household	Pearson Corr	1	.421**	-.332**	-.134**
	Sig. (2-tailed)		.000	.000	.000
	N	12829	12814	11681	11565
IDACI score	Pearson Corr	.421**	1	-.271**	-.132**
	Sig. (2-tailed)	.000		.000	.000
	N	12814	15754	14290	14057
KS2 score	Pearson Corr	-.332**	-.271**	1	.210**
	Sig. (2-tailed)	.000	.000		.000
	N	11681	14290	14301	12742
Evening homework per week	Pearson Corr	-.134**	-.132**	.210**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	11565	14057	12742	14073

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

As you might expect, IDACI (Income Deprivation Affecting Children Index) is significantly related to the SEC (socio-economic classification) of the head of the

household but the relationship does not appear strong enough (Pearson's $r = .42$) to be considered a problem. Usually values of $r = .8$ or more are cause for concern. As before the Variance Inflation Factor (VIF) and tolerance statistics can be used to help you verify that multicollinearity is not a problem (see **Page 3.3**).

4.10 An example from LSYPE

Now that we have discussed the theory underlying logistic regression let's put it into practice with an example from the *LSYPE* dataset. As we have said, in England, the success of education is often judged by whether a student achieve five or more GCSE at grades A* - C', including the subjects of English and maths, when they take these exams at age 16 (our humble *fiveem* variable). This criterion is considered a benchmark of academic success - it can dictate whether or not a student is advised to continue their studies after the age of 16 and it plays a role in deciding which courses of study are available to them. You will know by now that we have a strong research interest in equity issues in relation to educational achievement. So for this example we will explore the extent to which ethnicity, social class and gender are associated with the probability of a student achieving 5 or more GCSE A*-C grades including English and maths (*fiveem*).

Descriptive statistics

As with all research we should not run straight to the statistical analyses, even though this might be the sexier bit because it is so clever! The starting point should always be simple descriptive statistics so we better understand our data before we engage with the more complex stuff. So what is the pattern of association between our key variables of ethnicity, SEC and gender and *fiveem*?

Remember, as we said on **Page 4.2**, that the advantage of coding our binary response as 0 and 1 is that the mean will then indicate the proportion achieving our outcome (the value coded 1). We can therefore just ask for simple means of *fiveem* by our independent variables (**Figure 4.10.1**).

Figure 4.10.1: Mean, N and SD of fiveem by student background variables

fiveem 5 or more A*-C incl. English & maths * ethnic Ethnic					
ethnic Ethnic group	Mean	N	SD	Odds	OR
0 White British	.475	9896	.499	0.91	base
1 Mixed heritage	.440	766	.497	0.79	0.87
2 Indian	.589	1004	.492	1.43	1.58
3 Pakistani	.367	922	.482	0.58	0.64
4 Bangladeshi	.420	709	.494	0.73	0.80
5 Black Caribbean	.326	555	.469	0.48	0.53
6 Black African	.428	584	.495	0.75	0.83
7 Any other group	.522	630	.500	1.09	1.21
Total	.466	15066	.499	0.87	-

fiveem 5 or more A*-C incl. English & maths * SECshort					
SECshort Socio-Economic Class	Mean	N	SD	Odds	OR
0 Missing	.400	2847	0.49	0.67	1.60
1 Managerial & professional	.659	4581	.474	1.93	4.62
2 Intermediate	.451	3976	.498	0.82	1.97
3 Routine, semi-routine or	.295	4000	.456	0.42	base
Total	.460	15404	.499	0.85	-

fiveem 5 or more A*-C incl. English & maths * gender					
gender Gender	Mean	N	SD	Odds	OR
0 Male	.428	7655	.495	0.75	base
1 Female	.506	7426	.500	1.02	1.37
Total	.466	15081	.499	0.87	-

Note: You will remember from module 2 that LSYPE was unable to code the SEC of the head of household for quite a large proportion (18%) of LSYPE students. These cases are coded 0 which has been defined as the missing value for *SECshort*. However, we do not want to lose this many cases from the analysis. To include these cases we have redefined the missing values property for *SECshort* to 'no missing values'. This means that those cases where SEC is not known (0) are included in the above table. We will explain how we to do this on **Page 4.13**.

We see that the proportion of White British students achieving *fiveem* is 48%. The proportion is substantially higher among Indian students (59%) and substantially lower among Black Caribbean students (33%), with the other ethnic groups falling in between these two extremes. There is also a substantial association between SEC and *fiveem*. While 29% of students from low SEC home achieve *fiveem*, this rises to 45% for students from middle SEC homes and 66% for students from high SEC homes. Finally there is also a difference related to gender, with 43% of boys achieving *fiveem* compared to 51% of girls.

In the table you will see that we have also calculated and included the odds and the odds ratios (OR) as described on **Page 4.2**. You will remember the odds are calculated as $[p/(1-p)]$ and represent the odds of achieving *fiveem* relative to the proportion not achieving *fiveem*. We did this in Microsoft EXCEL, but you could equally easily use a calculator. The OR compares the odds of success for a particular group to a base category for that variable. For the variable ethnicity we have selected White British as the base category. From the OR we can therefore say that Indian students are 1.58 times more likely than White British students to achieve *fiveem*. From **Page 4.7** you will also remember that I can express this in percentage terms ($1-OR * 100$), so Indian students are 58% more likely to achieve *fiveem* than White British students. Conversely for Black Caribbean students the OR is 0.53, so Black Caribbean students are about half as likely as White British students to achieve *fiveem*. In percentage terms ($1-OR * 100$) they are 47% less likely to achieve *fiveem*. The ORs for SEC and for gender were calculated in the same way.


4.11 Running a logistic regression model on SPSS

So we can see the associations between ethnic group, social class (SEC), gender and achievement quite clearly without the need for any fancy statistical analysis. Why would we want to get involved in logistic regression modeling? There are three rather good reasons:

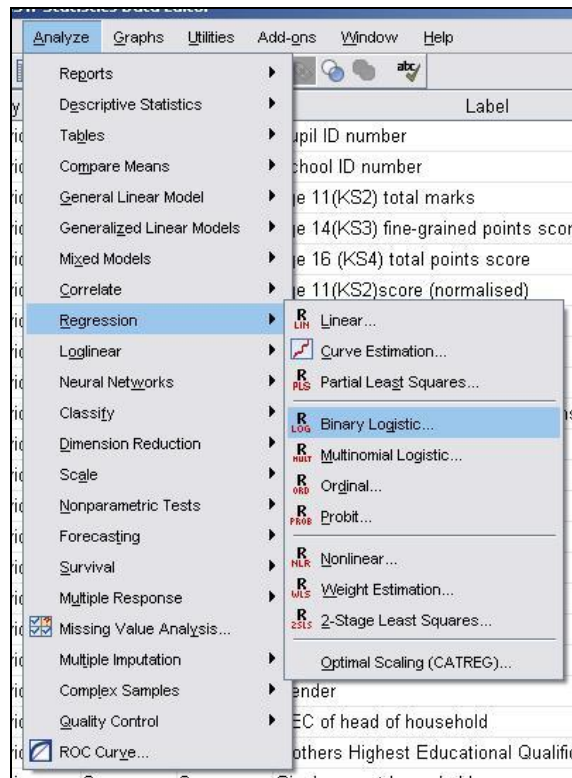
- To evaluate the statistical significance of the above associations. Remember that this data represents only a sample (although a very large sample) from the population of all students in England (approximately 600,000 students in any one year group). Are the effects in the sample sufficiently large relative to their standard errors that they are likely to be true in the population?
- To gauge the effect of one explanatory variable (e.g. ethnicity) on the outcome when controlling for other variables also associated with the outcome (e.g. SEC and gender). Specifically we are interested here in what the OR for ethnicity looks like after we have controlled for differences in exam achievement associated with SEC and gender.
- To gauge the extent and significance of any *interactions* between the explanatory variables in their effects on the outcome.

To do this we will need to run a logistic regression which will attempt to predict the outcome *fiveem* based on a student's ethnic group, SEC and gender.

Setting up the logistic regression model

Let's get started by setting up the logistic regression analysis. We will create a logistic regression model with three explanatory variables (ethnic, SEC and gender) and one outcome (*fiveem*) – this should help us get used to things! You can open up the LSYPE 15,000 Dataset  to work through this example with us.

Take the following route through SPSS: **Analyse > Regression > Binary Logistic**

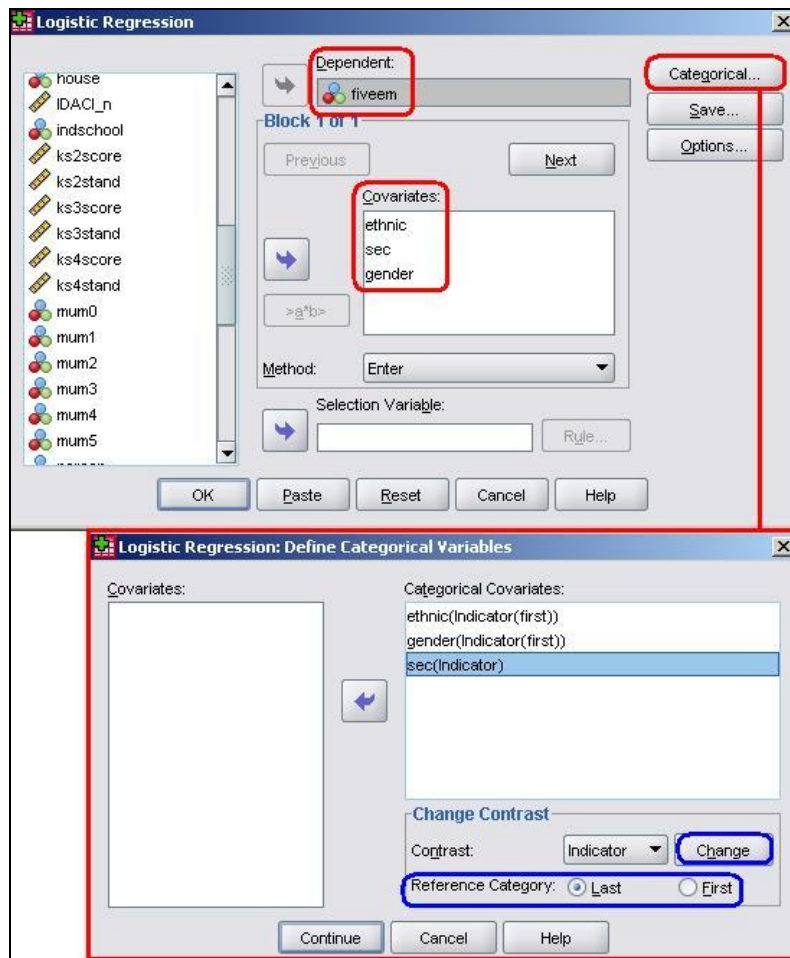


The logistic regression pop-up box will appear and allow you to input the variables as you see fit and also to activate certain optional features. First of all we should tell SPSS which variables we want to examine. Our outcome measure is whether or not the student achieves five or more A*-Cs (including Maths and English) and is coded as '0' for no and '1' for yes. This variable is labelled **fiveem** and should be moved in to the *Dependent* box.

Any explanatory variables need to be placed in what is named the *covariates* box. If the explanatory variable is continuous it can be dropped in to this box as normal and SPSS can be trusted to add it to the model, However, the process is slightly more demanding for categorical variables such as the three we wish to add because we need to tell SPSS to set up dummy variables based on a specific baseline category (we do not need to create the dummies ourselves this time... hooray!).

Let's run through this process. To start with, move *ethnic*, *SEC* and *Gender* into the covariates box. Now they are there we now need to define them as categorical variables. To do this we need to click the button marked 'Categorical' (a rare moment of simplicity from our dear friend SPSS) to open a submenu. You need to move all of

the explanatory variables that are categorical from the left hand list (Covariates) to the right hand window... in this case we need to move all of them!



The next step is to tell SPSS which category is the reference (or baseline) category for each variable. To do this we must click on each in turn and use the controls on the bottom right of the menu which are marked 'Change Contrast'. The first thing to note is the little drop down menu which is set to 'Indicator' as a default. This allows you to alter how categories within variables are compared in a number of ways (that you may or may not be pleased to hear are beyond the scope of this module). For our purposes we can stick with the default of 'indicator', which essentially creates dummy variables for each category to compare against a specified reference category – a process which you are probably getting familiar with now (if not, head to **Page 3.6**).

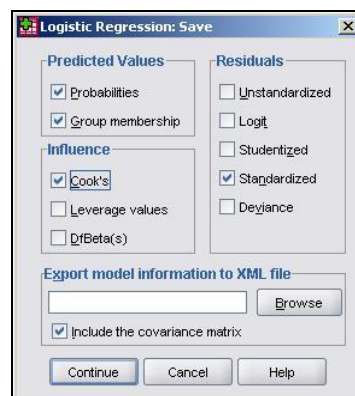
All we need to do then is tell SPSS whether the first or last category should be used as the reference and then click 'Change' to finalize the setting. For our *Ethnic* variable

the first category is '0' White-British (the category with the highest number of participants) so, as before, we will use this as the reference category. Change the selection to 'First' and click 'Change'. For the *Gender* variable we only have two categories and could use either male ('0') or female ('1') as the reference. Previously we have used male as the reference so we will stick with this (once again, change the selection to 'First' and click 'Change'). Finally, for Socio Economic Class (*sec*) we will use the least affluent class as the reference ('Never worked/long term unemployed - 8'). This time we will use the 'Last' option given that the SEC categories are coded such that the least affluent one is assigned the highest value code. Remember to click 'Change'! You will see that your selections have appeared in brackets next to each variable and you can click 'Continue' to close the submenu.

Notice that on the main Logistic Regression menu you can change the option for which *method* you use with a drop down menu below the *covariates* box. As we are entering all three explanatory variables together as one block you can leave this as 'Enter'. You will also notice that our explanatory variables (Covariates) now have 'Cat' printed next to them in brackets. This simply means that they have been defined as categorical variables, not that they have suddenly become feline (that would just be silly).

The Logistic Regression Sub-menus

Now that our variables have been defined we can start playing with the option menus. Beware SPSS's desire to dazzle you with a myriad of different tick boxes, options and settings - some people just like to show off! We'll guide you through the useful options. The save sub-menu is very useful and can be seen below.



If you recall, the save options actually create new variables for your data set. We can ask SPSS to calculate four additional variables for us:

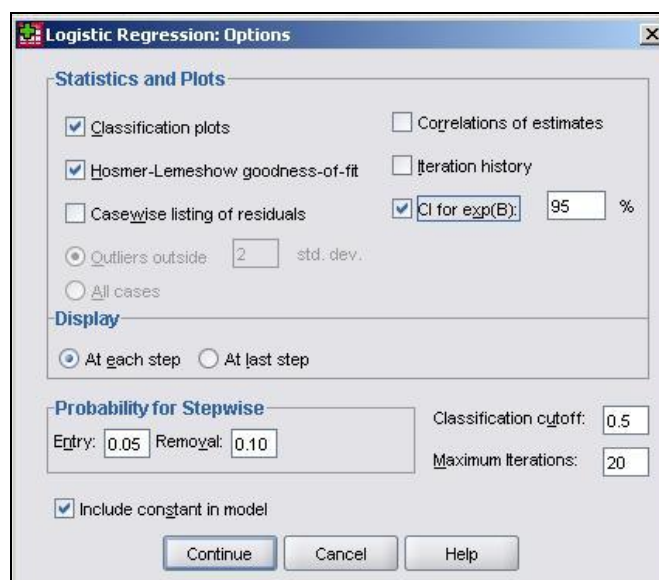
Predicted probabilities – This creates a new variable that tells us for each case the predicted probability that the outcome will occur (that *fiveem* will be achieved) based on the model.

Predicted Group Membership – This new variable estimates the outcome for each participant based on their predicted probability. If the predicted probability is >0.5 then they are predicted to achieve the outcome, if it is $<.5$ they are predicted not to achieve the outcome. This .5 cut-point can be changed, but it is sensible to leave it at the default. The predicted classification is useful for comparison with the actual outcome!

Residuals (standardized) – This provides the *residual* for each participant (in terms of standard deviations for ease of interpretation). This shows us the difference between the actual outcome (0 or 1) and the probability of the predicted outcome and is therefore a useful measure of error.

Cook's – We've come across this in our travels before. This generates a statistic called *Cook's distance* for each participant which is useful for spotting cases which unduly influence the model (a value greater than '1' usually warrants further investigation).

The other options can be useful for the statistically-minded but for the purposes of our analysis the options above should suffice (we think we are fairly thorough!). Click on *Continue* to shut the sub-menu. The next sub-menu to consider is called *options*:



Again we have highlighted a few of the options here:

Classification plots – Checking this option requests a chart which shows the distribution of outcomes over the probability range. This is useful for visually identifying where the model makes most incorrect categorizations. This will make more sense when we look at one on the **Page 4.12!**

Hosmer-Lameshow Goodness of fit – This option provides a X^2 (Chi-square) test of whether or not the model is an adequate fit to the data. The null hypothesis is that the model is a 'good enough' fit to the data and we will only reject this null hypothesis (i.e. decide it is a 'poor' fit) if there are sufficiently strong grounds to do so (conventionally if $p < .05$). We will see that with very large samples as we have here there can be problems with this level of significance, but more on that later.

CI for exp(B) – CI stands for confidence interval and this option requests the range of values that we are confident that each odds ratio lies within. The setting of 95% means that there is only a $p < .05$ that the value for the odds ratio, **exp(B)**, lies outside the calculated range (you can change the 95% confidence level if you are a control freak!).

Click on *continue* to close the sub-menu. Once you are happy with all the settings take a deep breath... and click **OK** to run the analysis.

4.12 The SPSS Logistic Regression Output

SPSS will present you with a number of tables of statistics. Let's work through and interpret them together. First of all we get these two tables (**Figure 4.12.1**):

Figure 4.12.1: Case Processing Summary and Variable Encoding for Model

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	12347	78.3
	Missing Cases	3423	21.7
	Total	15770	100.0
Unselected Cases		0	.0
Total		15770	100.0

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

Original Value	Internal Value
no	0
yes	1

The *Case Processing Summary* simply tells us about how many cases are included in our analysis. The second row tells us that 3423 participants are missing data on some of the variables included in our analysis (they are missing either ethnicity, gender or *fiveem*, remember we have included all cases with missing SEC), but this still leaves us with 12347 cases to analyse. The *Dependent Variable Encoding* reminds us how our outcome variable is encoded – '0' for 'no' (Not getting 5 or more A*-C grades including Maths and English) and '1' for 'yes' (making the grade!).

Next up is the *Categorical Variables Encoding Table* (**Figure 4.12.2**). It acts as an important reminder of which categories were coded as the reference (baseline) for each of your categorical explanatory variables. You might be thinking 'I can remember what I coded as the reference category!' but it's easy to get lost in the output because SPSS has a delightful tendency to rename things just as you are becoming familiar with them... In this case 'parameter coding' is used in the SPSS logistic regression output rather than the value labels so you will need to refer to this table later on. Let's consider the example of ethnicity. White British is the reference category because it does not have a parameter coding. Mixed heritage students will be labeled "ethnic(1)" in the SPSS logistic regression output, Indian students will be labeled "ethnic(2)",

Pakistani students “ethnic(3)” and so on. You will also see that ‘Never worked/long term unemployed’ is the base category for SEC, and that each of the other SEC categories has a ‘parameter coding’ of 1-7 reflecting each of the seven dummy SEC variables that SPSS has created. This is only important in terms of how the output is labeled, nothing else, but you will need to refer to it later to make sense of the output.

Figure 4.12.2: Categorical Variables Coding Table

		Categorical Variables Codings							
		Frequency	Parameter coding						
			(1)	(2)	(3)	(4)	(5)	(6)	(7)
ethnic Ethnic group	0 White British	8319	.000	.000	.000	.000	.000	.000	.000
	1 Mixed heritage	626	1.000	.000	.000	.000	.000	.000	.000
	2 Indian	800	.000	1.000	.000	.000	.000	.000	.000
	3 Pakistani	707	.000	.000	1.000	.000	.000	.000	.000
	4 Bangladeshi	489	.000	.000	.000	1.000	.000	.000	.000
	5 Black Caribbean	452	.000	.000	.000	.000	1.000	.000	.000
	6 Black African	463	.000	.000	.000	.000	.000	1.000	.000
	7 Any other group	491	.000	.000	.000	.000	.000	.000	1.000
sec MP social class	1 Higher Managerial and professional occupations	1528	1.000	.000	.000	.000	.000	.000	.000
	2 Lower managerial and professional occupations	2995	.000	1.000	.000	.000	.000	.000	.000
	3 Intermediate occupations	896	.000	.000	1.000	.000	.000	.000	.000
	4 Small employers and own account workers	1609	.000	.000	.000	1.000	.000	.000	.000
	5 Lower supervisory and technical occupations	1397	.000	.000	.000	.000	1.000	.000	.000
	6 Semi-routine occupations	1566	.000	.000	.000	.000	.000	1.000	.000
	7 Routine occupations	1341	.000	.000	.000	.000	.000	.000	1.000
	8 Never worked/long term unemployed	1015	.000	.000	.000	.000	.000	.000	.000
gender Gender	0 Male	6329	.000						
	1 Female	6018	1.000						

The next set of output is under the heading of *Block 0: Beginning Block* (Figure 4.12.3):

Figure 4.12.3: Classification Table and Variables in the Equation

Classification Table^{a,b}

		Observed		Predicted		Percentage Correct
				5 or more A*-C incl. English & maths		
		no	yes			
Step 0	5 or more A*-C incl. English & maths	no	6422	0	100.0	
		yes	5925	0	.0	
Overall Percentage						52.0

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	-.081	.018	19.995	1	.000	.923

This set of tables describes the baseline model – that is a model that does not include our explanatory variables! As we mentioned previously, the predictions of this baseline model are made purely on whichever category occurred most often in our dataset. In this example the model always guesses ‘no’ because more participants did not achieve 5 or more A*-C grades than did (6422 compared to 5925 according to our first column). The *overall percentage* row tells us that this approach to prediction is correct 52.0% of the time – so it is only a little better than tossing a coin!

The *Variables in the Equation* table shows us the coefficient for the constant (B_0). This table is not particularly important but we’ve highlighted the significance level to illustrate a cautionary tale! According to this table the model with just the constant is a statistically significant predictor of the outcome ($p < .001$). However it is only accurate 52% of the time! The reason we can be so confident that our baseline model has some predictive power (better than just guessing) is that we have a very large sample size – even though it only marginally improves the prediction (the effect size) we have enough cases to provide strong evidence that this improvement is unlikely to be due to sampling. You will see that our large sample size will lead to high levels of statistical significance for relatively small effects in a number of cases.

We have not printed the next table *Variables not Included in the Model* because all it really does is tell us that none of our explanatory variables were actually included in this baseline model (Block 0)... which we know anyway! It is however worth noting the number in brackets next to each variable – this is the ‘parameter coding’ we mentioned earlier. As you can see, you will need to refer to the *Categorical Variables Encoding Table* to make sense of these!

Now we move to the regression model that includes our explanatory variables. The next set of tables begins with the heading of *Block 1: Method = Enter* (**Figure 4.12.4**):

Figure 4.12.4: Omnibus Tests of Coefficients and Model Summary

		Chi-square	df	Sig.
Step 1	Step	1566.727	15	.000
	Block	1566.727	15	.000
	Model	1566.727	15	.000

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	15529.838 ^a	.119	.159

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

The *Omnibus Tests of Model Coefficients* is used to check that the new model (with explanatory variables included) is an improvement over the baseline model. It uses chi-square tests to see if there is a significant difference between the Log-likelihoods (specifically the -2LLs) of the baseline model and the new model. If the new model has a significantly reduced -2LL compared to the baseline then it suggests that the new model is explaining more of the variance in the outcome and is an improvement! Here the chi-square is highly significant ($chi-square=1566.7$, $df=15$, $p<.000$) so our new model is significantly better.

To confuse matters there are three different versions; *Step*, *Block* and *Model*. The *Model* row always compares the new model to the baseline. The *Step* and *Block* rows are only important if you are adding the predictors to the model in a stepwise or hierarchical manner. If we were building the model up in stages then these rows would compare the -2LLs of the newest model with the previous version to ascertain whether or not each new set of explanatory variables were causing improvements. In this case we have added all three explanatory variables in one block and therefore have only one step. This means that the chi-square values are the same for step, block and model. The *Sig.* values are $p < .001$, which indicates the accuracy of the model improves when we add our explanatory variables.

The *Model Summary* (also in **Figure 4.12.4**) provides the -2LL and pseudo-R² values for the full model. The -2LL value for this model (15529.8) is what was compared to

the -2LL for the previous null model in the ‘omnibus test of model coefficients’ which told us there was a significant decrease in the -2LL, i.e. that our new model (with explanatory variables) is significantly better fit than the null model. The R^2 values tell us approximately how much variation in the outcome is explained by the model (like in linear regression analysis). We prefer to use the Nagelkerke’s R^2 (circled) which suggests that the model explains roughly 16% of the variation in the outcome. Notice how the two versions (Cox & Snell and Nagelkerke) do vary! This just goes to show that these R^2 values are approximations and should not be overly emphasized.

Moving on, the Hosmer & Lemeshow test (**Figure 4.12.5**) of the goodness of fit suggests the model is a good fit to the data as $p=0.792 (>.05)$. However the *chi-squared* statistic on which it is based is very dependent on sample size so the value cannot be interpreted in isolation from the size of the sample. As it happens, this *p value* may change when we allow for interactions in our data, but that will be explained in a subsequent model on **Page 4.13**. You will notice that the output also includes a *contingency table*, but we do not study this in any detail so we have not included it here.

Figure 4.12.5: Hosmer and Lameshow Test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	4.674	8	.792

More useful is the *Classification Table* (**Figure 4.12.6**). This table is the equivalent to that in *Block 0* (**Figure 4.12.3**) but is now based on the model that includes our explanatory variables. As you can see our model is now correctly classifying the outcome for 64.5% of the cases compared to 52.0% in the null model. A marked improvement!

Figure 4.12.6: Classification Table for Block 1

Observed			Predicted		Percentage Correct
			5 or more A*-C incl. English & maths		
			no	yes	
Step 1	5 or more A*-C incl. English & maths	no	4234	2188	65.9
		yes	2193	3732	63.0
Overall Percentage					64.5

a. The cut value is .500

However the most important of all output is the *Variables in the Equation* table (Figure 4.12.7). We need to study this table extremely closely because it is at the heart of answering our questions about the joint association of ethnicity, SEC and gender with exam achievement.

Figure 4.12.7: Variables in the Equation Table Block 1

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a			133.180	7	.000			
ethnic								
ethnic(1)	-.129	.089	2.116	1	.146	.879	.739	1.046
ethnic(2)	.679	.080	72.405	1	.000	1.972	1.686	2.305
ethnic(3)	-.080	.088	.839	1	.360	.923	.777	1.096
ethnic(4)	.387	.104	13.981	1	.000	1.473	1.202	1.804
ethnic(5)	-.564	.107	27.909	1	.000	.569	.462	.701
ethnic(6)	-.051	.106	.228	1	.633	.950	.771	1.171
ethnic(7)	.335	.100	11.148	1	.001	1.399	1.149	1.703
sec			1182.976	7	.000			
sec(1)	2.431	.101	577.399	1	.000	11.371	9.325	13.864
sec(2)	1.726	.089	374.634	1	.000	5.618	4.717	6.691
sec(3)	1.223	.105	134.974	1	.000	3.398	2.764	4.177
sec(4)	1.200	.094	164.172	1	.000	3.320	2.763	3.989
sec(5)	.783	.097	65.001	1	.000	2.187	1.808	2.646
sec(6)	.563	.096	34.711	1	.000	1.757	1.456	2.119
sec(7)	.318	.099	10.213	1	.001	1.374	1.131	1.669
gender(1)	.393	.039	104.052	1	.000	1.482	1.374	1.598
Constant	-1.476	.086	293.797	1	.000	.229		

a. Variable(s) entered on step 1: ethnic, sec, gender.

This table provides the regression coefficient (**B**), the **Wald statistic** (to test the statistical significance) and the all important Odds Ratio (**Exp (B)**) for each variable category.

Looking first at the results for SEC, there is a highly significant overall effect ($Wald=1283, df=7, p<.000$). The b coefficients for all SECs (1-7) are significant and positive, indicating that increasing affluence is associated with increased odds of achieving *fiveem*. The Exp(B) column (the Odds Ratio) tells us that students from the highest SEC homes are eleven (11.37) times more likely than those from lowest SEC homes (our reference category) to achieve *fiveem*. Comparatively those from the SEC group just above the poorest homes are about 1.37 times (or 37%) more likely to achieve *fiveem* than those from the lowest SEC group. The effect of gender is also significant and positive, indicating that girls are more likely to achieve *fiveem* than boys. The OR tells us they are 1.48 times (or 48%) more likely to achieve *fiveem*, even after controlling for ethnicity and SEC (refer back to **Page 4.7** 'effect size of explanatory variables' to remind yourself how these percentages are calculated).

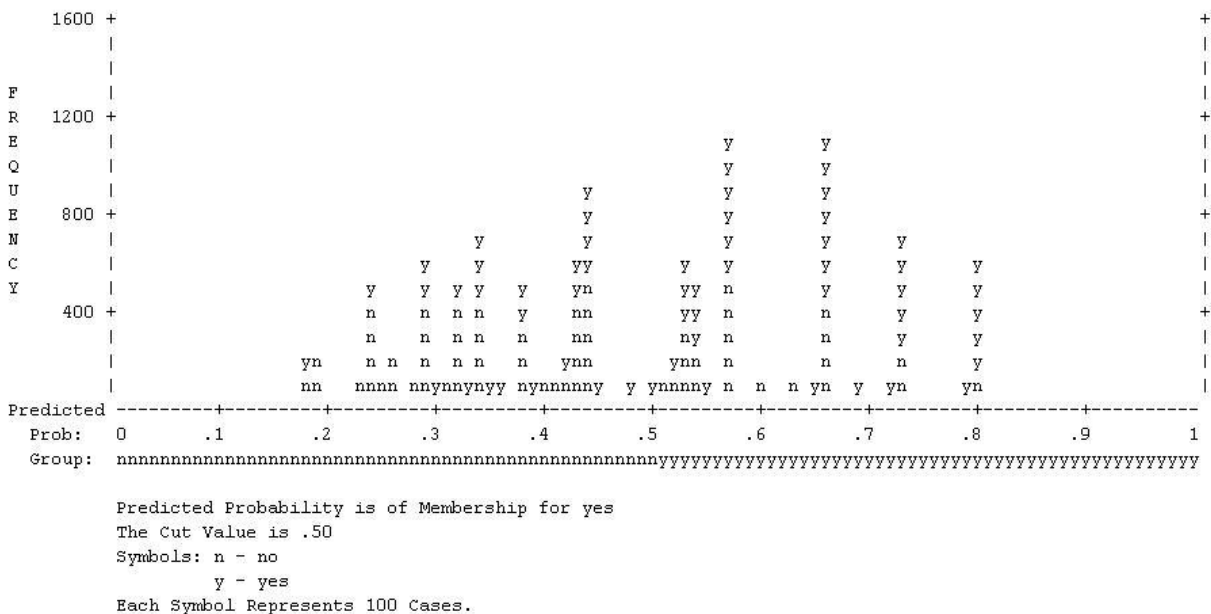
Most importantly, controlling for SEC and gender has changed the associations between ethnicity and *fiveem*. The overall association between *fiveem* and ethnicity remains highly significant, as indicated by the overall Wald statistic, but the size of the b coefficients² and the associated ORs for most of the ethnic groups has changed substantially. This is because the SEC profile for most ethnic minority groups is lower than for White British, so controlling for SEC has significantly changed the odds ratios for these ethnic groups (as it did in our multiple linear regression example). We saw in **Figure 4.10.1** that Indian students (Ethnic(2)) were significantly more likely than White British students to achieve *fiveem* (OR=1.58), and now we see that this increases even further after controlling for SEC and gender (OR=1.97). Bangladeshi students (Ethnic(4)) were previously significantly less likely than White British students to achieve *fiveem* (OR=.80) but are now significantly more likely (OR=1.47). Pakistani (Ethnic(3)) students were also previously significantly less likely than White British students to achieve *fiveem* (OR=.64) but now do not differ significantly after controlling for SEC (OR=.92). The same is true for Black African (Ethnic(6)) students (OR change from .83 to .95). However the OR for Black Caribbean (Ethnic(5)) students has not changed much at all (OR change .53 to .57) and they are still significantly less likely to

². Before running this model we ran a model that just included ethnic group to estimate the b coefficients and to test the statistical significance of the ethnic gaps for *fiveem*. We haven't reported it here because the Odds Ratios from the model are identical to those shown in **Figure 4.10.1**. However the b coefficients and their statistical significance are shown as Model 1 in **Figure 4.15.1** where we show how to present the results of a logistic regression.

achieve *fiveem* than White British students, even after accounting for the influence of social class and gender.

The final piece of output is the classification plot (**Figure 4.12.8**). Conceptually it answers a similar question as the classification table (see **Figure 4.12.6**) which is ‘how accurate is our model in classifying individual cases’? However the classification plot gives some finer detail. This plot shows you the frequency of categorizations for different predicted probabilities and whether they were ‘yes’ or ‘no’ categorizations. This provides a useful visual guide to how accurate our model is by displaying how many times the model would predict a ‘yes’ outcome based on the calculated predicted probability when in fact the outcome for the participant was ‘no’.

Figure 4.12.8: Observed groups and Predicted Probabilities



If the model is good at predicting the outcome for individual cases we should see a bunching of the observations towards the left and right ends of the graph. Such a plot would show that where the event did occur (*fiveem* was achieved, as indicated by a ‘y’ in the graph) the predicted probability was also high, and that where the event did not occur (*fiveem* was not achieved, indicated by a ‘n’ in the graph) the predicted probability was also low. The above graph shows that quite a lot of cases are actually in the middle area of the plot, i.e. the model is predicting a probability of around .5 (or


a 50:50 chance) that *fiveem* will be achieved. So while our model identifies that SEC, ethnicity and gender are significantly associated with the *fiveem* outcome, and indeed can explain 15.9% of the variance in outcome (quoting the Nagelkerke pseudo- R^2), they do not predict the outcome for *individual students* very well. This is important because it indicates that social class, ethnicity and gender do not *determine* students' outcomes (although they are significantly associated with it). There is substantial individual variability that cannot be explained by social class, ethnicity or gender, and we might expect this reflects individual factors like prior attainment, student effort, teaching quality, etc.

Let's move on to discuss interaction terms for now – we will save explaining how to test the assumptions of the model for a little later. Something to look forward to!

4.13 Evaluating interaction effects

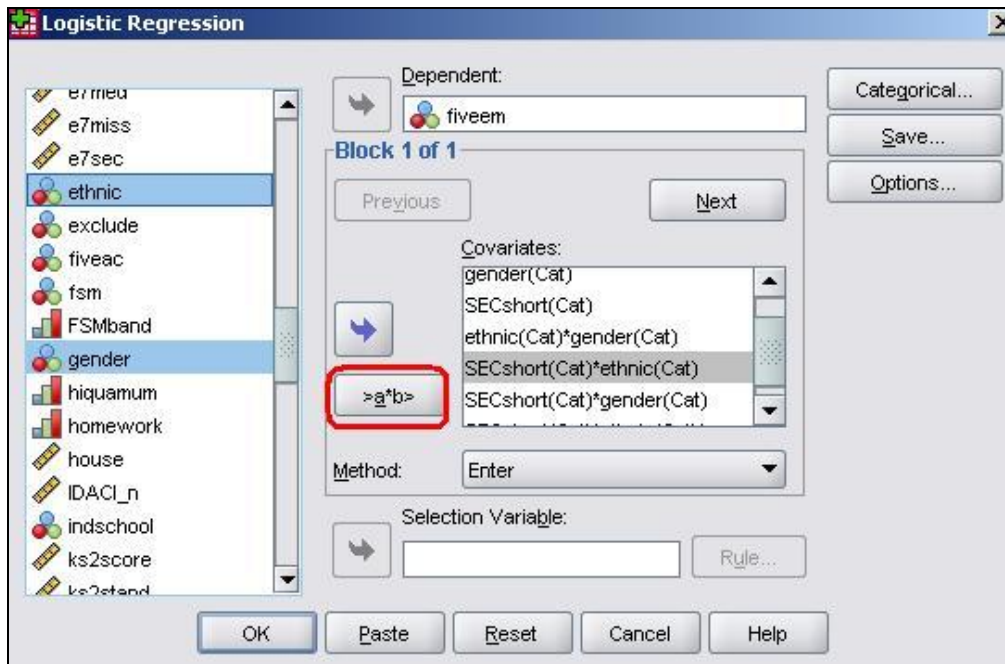
We saw in **Module 3** when modeling a continuous measure of exam achievement (the age 14 average test score) that there were significant interactions between ethnic group and SEC (if you want to remind yourself about interaction effects head to **Page 3.11**). There are therefore strong grounds to explore whether there are interaction effects for our measure of exam achievement at age 16.

The first step is to add all the interaction terms, starting with the highest. With three explanatory variables there is the possibility of a 3-way interaction (ethnic * gender * SEC). If we include a higher order (3 way) interaction we must also include all the possible 2-way interactions that underlie it (and of course the main effects). There are three 2-way interactions: ethnic*gender, ethnic*SEC and Gender*SEC. Our strategy here is to start with the most complex 3-way interaction to see if it is significant. If it is not then we can eliminate it and just test the 2-way interactions. If any of these are not significant then we can eliminate them. In this way we can see if any interaction terms make a statistically significant contribution to the interpretation of the model.

In this example we will use the MLR LSYPE 15,000  dataset because it contains some useful extra variables which we created for the last module. The process for creating a model with interaction terms is very similar to doing it without them so we won't repeat the whole process in detail (see the previous page, **Page 4.12**, if you require a recap). However, there is a key extra step which we describe below...

Entering interaction terms to a logistic model

The masters of SPSS smile upon us, for adding interaction terms to a logistic regression model is remarkably easy in comparison to adding them to a multiple linear regression one! Circled in the image below is a button which is essentially the 'interaction' button and is marked as '>a*b>'. How very helpful! All you have to do is highlight the two (or more) variables you wish to create an interaction term for in the left hand window (hold down 'control' on your keyboard while selecting your variables to highlight more than one) and then use the '>a*b>' button to move them across to the right hand window as an interaction term.



The two variables will appear next to each other separated by a '*'. In this way you can add all the interaction terms to your model.

Reducing the complexity of the model

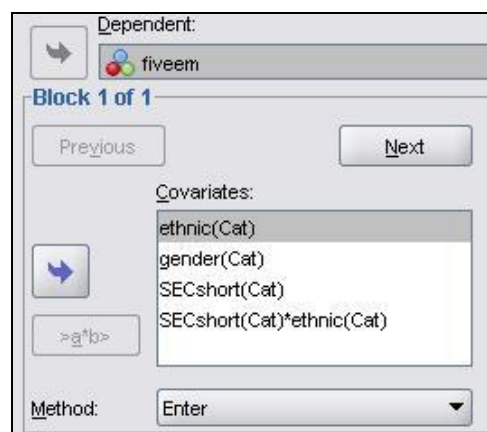
If we were to create interaction terms involving all levels of *SEC* we would probably become overwhelmed by the sheer number of variables in our model. For the two-way interaction between ethnicity and *SEC* alone we would have seven *ethnic* dummy variables multiplied by seven *SEC* dummy variables giving us a total of 49 interaction terms! Of course, we could simplify the model if we treated *SEC* as a continuous variable, we would then have only seven terms for the interaction between *ethnic* * *SEC*. While it would be a more parsimonious model (because it has fewer parameters to model the interaction), treating *SEC* as a continuous variable would mean omitting the nearly 3,000 cases where *SEC* was missing. The solution we have taken to this problem, as described before on **Page 3.12**, is to use the shortened version of the *SEC* variable called *SECshort* which has only three (rather than eight) *SEC* categories (plus a code for missing values). That should make our lives a little less confusing!

Even though we have chosen to use the three category *SEC* measure, the output is very extensive when we include all possible interaction terms. We have a total of 55 interaction terms (three for *gender***SECshort*, seven for *ethnic***gender*, 21 for

ethnic*SECshort and a further 21 for ethnic*SECshort*gender). You will forgive us then if we do not ask you to run the analysis with all the interactions! Instead we will give you a brief summary of the preliminary analyses, before asking you to run a slightly less complex model. Our first model included all the three-way and two way interactions as well as the main effects. It established that three-way interaction was not significant ($p=0.91$) and so could be eliminated. Our second model then included just all the two-way interactions (and main effects). This showed that the gender*SECshort and the ethnic*gender interactions were also not significant but the ethnic*SECshort interaction was significant. The final model therefore eliminated all but the ethnic*SECshort interaction which needs to be included along with the main effects.

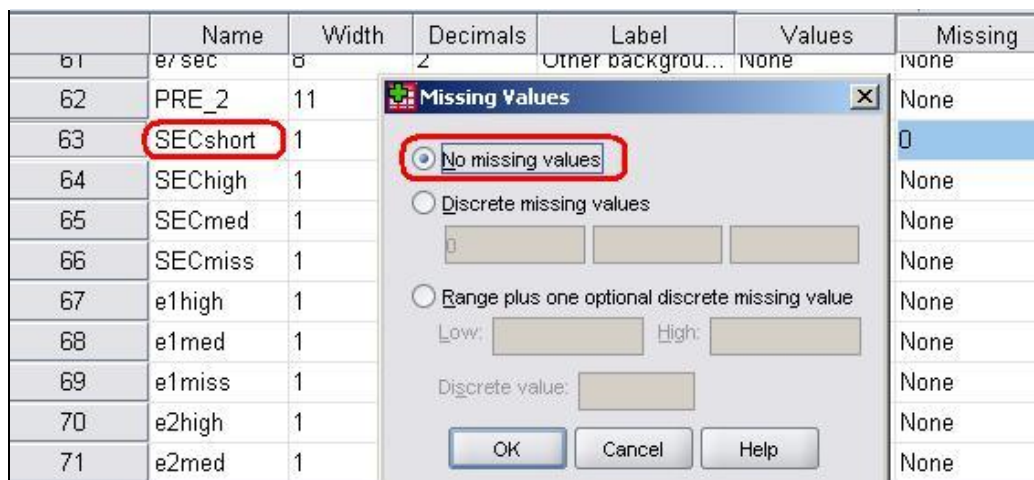
Running the logistic model with an interaction term

So let's run this final model including the ethnic*SECshort interaction. Maybe you want to run through this example with us. In this model the 'dependent' variable is *fiveem* (our Outcome Variable) and the 'covariates' (our explanatory variables) are *ethnic*, *gender*, *SECshort*, and *ethnic*SECshort* (the interaction term, which is entered in the way that we showed you earlier on this page). Your final list of variables should look like the one below.



Remember to tell SPSS which variables are categorical and set the options as we showed you on **Page 4.11!**

Before running this model you will need to do one more thing. Wherever it was not possible to estimate the SEC of the household in which the student lived *SECshort* was coded 0. To exclude these cases from any analysis the ‘missing value’ indicator for *SECshort* is currently set to the value ‘0’. As discussed on **Page 3.9**, it is actually very useful to include a dummy variable for missing data where possible. If we want to include these cases we will need to tell SPSS. Go to the ‘Variable view’ and find the row of options for *SECshort*. Click on the box for *Missing* and change the option to ‘No missing values’ (see below) and click OK to confirm the change.



This will ensure that SPSS makes us a dummy variable for SEC missing. You can now click **OK** on the main menu screen to run the model!

Interpreting the output

The results of this final model are shown below. Rather than show you all of the output as on the previous page (**Page 4.12**), this time we will only show you the ‘Variables in the Equation’ table (**Figure 4.13.1**) as it is most relevant to interpreting interaction effects.

Figure 4.13.1: Variables in the Equation Table with Interaction Terms

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp (B)
ethnic			65.003	7	.000	
ethnic(1)	-.079	.171	.212	1	.645	.924
ethnic(2)	.784	.128	37.592	1	.000	2.191
ethnic(3)	.152	.125	1.477	1	.224	1.165
ethnic(4)	.544	.123	19.702	1	.000	1.723
ethnic(5)	-.246	.210	1.377	1	.241	.782
ethnic(6)	.068	.156	.188	1	.665	1.070
ethnic(7)	.605	.163	13.798	1	.000	1.831
gender(1)	.340	.035	96.900	1	.000	1.405
SECshort			915.572	3	.000	
SECshort(1)	.551	.071	60.374	1	.000	1.736
SECshort(2)	1.735	.061	817.996	1	.000	5.669
SECshort(3)	.818	.062	173.754	1	.000	2.266
SECshort * ethnic			43.758	21	.003	
SECshort(1) by ethnic(1)	-.005	.251	.000	1	.984	.995
SECshort(1) by ethnic(2)	-.065	.198	.108	1	.743	.937
SECshort(1) by ethnic(3)	-.328	.198	2.760	1	.097	.720
SECshort(1) by ethnic(4)	-.320	.190	2.831	1	.092	.726
SECshort(1) by ethnic(5)	-.492	.318	2.386	1	.122	.612
SECshort(1) by ethnic(6)	-.288	.246	1.373	1	.241	1.334
SECshort(1) by ethnic(7)	-.177	.241	.536	1	.464	.838
SECshort(2) by ethnic(1)	-.129	.217	.352	1	.553	.879
SECshort(2) by ethnic(2)	-.159	.215	.553	1	.457	.853
SECshort(2) by ethnic(3)	-.519	.237	4.823	1	.028	.595
SECshort(2) by ethnic(4)	-1.009	.366	7.609	1	.006	.365
SECshort(2) by ethnic(5)	-.782	.263	8.812	1	.003	.457
SECshort(2) by ethnic(6)	-.447	.231	3.750	1	.053	.640
SECshort(2) by ethnic(7)	-.418	.239	3.057	1	.080	.658
SECshort(3) by ethnic(1)	-.019	.235	.007	1	.935	.981
SECshort(3) by ethnic(2)	-.145	.179	.659	1	.417	.865
SECshort(3) by ethnic(3)	-.418	.187	5.003	1	.025	.659
SECshort(3) by ethnic(4)	-.711	.226	9.906	1	.002	.491
SECshort(3) by ethnic(5)	-.065	.276	.055	1	.815	.937
SECshort(3) by ethnic(6)	-.001	.273	.000	1	.997	1.001
SECshort(3) by ethnic(7)	-.453	.234	3.750	1	.053	.636
Constant	-1.184	.051	529.007	1	.000	.306

a. Variable(s) entered on step 1: ethnic, gender, SECshort, SECshort * ethnic .

The overall Wald for the SECshort*ethnic interaction is significant (WALD=43.8, df=21, $p<.005$) so we proceed to look at the individual regression coefficients. You will need to refer to the *Categorical Variables Encoding* Table to remind yourself what each of these coefficients represents (for example SECshort(1)=missing; SECshort(2)= high SEC and SECshort(3)= middle SEC, with low SEC as the reference category). There are statistically significant coefficients for the interaction between Pakistani and middle and high SEC, between Bangladeshi and middle and high SEC and between Black Caribbean and high SEC.

Working out the ORs with interaction effects is somewhat tricky (remember we encountered a similar issue for multiple linear regression modules on **Page 3.11**). As we have discussed, each B coefficient represents the change in the logit of our outcome predicted by a one unit change in our explanatory variable, but this is more complicated when we are also have interactions between our explanatory variables.

- Each of the ethnic coefficients represents the difference between that ethnic group and 'White British' students, but crucially only for students in the baseline category for *SECshort* (i.e. low SEC students).
- For *SECshort* the coefficients represent the difference between each of the medium and high SEC categories and the baseline category of low SEC, but only for "White British" students.
- The coefficients for each ethnic * SECshort interaction term represent how much the SEC contrasts vary for each ethnic group, relative to the size of the SEC effect among White British students.

Interpreting the SEC gap for different ethnic groups

We can see that SEC has a substantial association with achievement among White British students. White British students from high SEC homes are 5.67 times more likely to achieve *fiveem* than White British students from low SEC homes. How do we determine the size of the SEC effect among other ethnic groups? Well, when interaction terms are included in the model we need to calculate the predicted probabilities by adding the B coefficients together, as we did in multiple linear regression. So to estimate the SEC gap for Black Caribbean students we add the coefficient for high SEC [SECshort(2)] and the B for the interaction between Black

Caribbean and high SEC [SECshort(2) by ethnic(5)]. This gives = 1.735 + -.784 = 0.953. What does this mean? Not much yet – we need to take the exponent to turn it into a trusty odds ratio: $\text{Exp}(0.953)=2.59$ (we used the =EXP() function in EXCEL or you could use a calculator). This means that among Black Caribbean students High SEC students are only 2.6 times more likely to achieve *fiveem* than low SEC students - the SEC gap is much smaller than among White British students. The important point to remember is that you cannot simply add up the Exp(B)s to arrive here – it only works if you add the B coefficients in their original form and then take the exponent of this sum!

Interpreting the ethnic gaps at different levels of SEC

The model has told us what the ethnic gaps are among low SEC students (the reference category). Suppose I wanted to know what the estimated size of the ethnic gaps was among high SEC students, how would I do this? To find out you would rerun the model but *set high SEC as the base or reference category*. The coefficients for each ethnic group would then represent the differences between the average for that ethnic group and White British students among those from high SEC homes.

Currently SECshort is coded as follows with the last category used as the reference group.

SECshort value	Label
0	Missing SEC
1	High SEC
2	Middle SEC
3	Low SEC (Reference category LAST)

We can simply recode the value for missing cases from 0 to 9 and set the reference category to the first value, so High SEC becomes the reference category, as shown below:

SECshort value	Label
1	High SEC (Reference category FIRST)
2	Middle SEC
3	Low SEC
4	Missing SEC

You can do this through the *Transformation-Recode into new variable* windows menu (see **Foundation Module**) or simply through the following syntax:

```
RECODE SECshort (0=4) (ELSE=Copy) INTO SECshortnew.  
EXECUTE
```

We then rerun the model simply adding SECshortnew as our SEC measure. It is important to note that computationally this is an exactly equivalent model to when low SEC was the reference category. The coefficients for other variables (for example, gender) are identical, the contrast between low and high SEC homes is the same (you can check the B value in the output below), and the R^2 and log-likelihood are exactly the same. All that has varied is that the coefficients printed for ethnicity are now the contrasts among high SEC rather than low SEC homes.

The output is shown below (**Figure 4.13.2**). For convenience we have added labels to the values so you can identify the groups. As you know, this is not done by SPSS so it is vital that you refer to the *Categorical variables encoding* table when interpreting your output. It is apparent that the ethnic gaps are substantially different among high SEC than among low SEC students. Among low SEC students the only significant contrasts were that Indian, Bangladeshi and Any other ethnic group had higher performance than White British (see **Figure 4.13.1**). However among students from high SEC homes while Indian students again achieve significantly better outcomes than White British students, both Black Caribbean (OR=.36, $p < .005$) and Black African (OR=.685, $p < .025$) are significantly less likely to achieve *fiveem* than White British students by a considerable margin. Black Caribbean students are only about one third as likely to achieve *fiveem* as White British high SEC students. In percentage terms we can say they are 64% ($0.358 - 1 * 100$) less likely to achieve *fiveem* than White British students of the same SEC group.

Figure 4.13.2: Variables in the Equation Table with high SEC as the reference category

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
ethnic			69.014	7	.000	
ethnic(1) [Mixed]	-.208	.134	2.399	1	.121	.812
ethnic(2) [Indian]	.625	.172	13.176	1	.000	1.868
ethnic(3) [Pakistani]	-.367	.201	3.349	1	.067	.693
ethnic(4) [Bangladeshi]	-.465	.345	1.820	1	.177	.628
ethnic(5) [Black Caribbean]	-1.028	.159	41.593	1	.000	.358
ethnic(6) [Black African]	-.379	.170	4.990	1	.025	.685
ethnic(7) [Any other ethnic group]	.186	.175	1.130	1	.288	1.205
gender(1) [Girls]	.340	.035	96.900	1	.000	1.405
SECshortnew			915.572	3	.000	
SECshortnew(1) [Middle SEC]	-.917	.053	295.757	1	.000	.400
SECshortnew(2) [Low SEC]	-1.735	.061	817.996	1	.000	.176
SECshortnew(3) [Missing SEC]	-1.184	.063	347.706	1	.000	.306
SECshortnew * ethnic			43.758	21	.003	
SECshortnew(1) by ethnic(1)	.110	.209	.275	1	.600	1.116
SECshortnew(1) by ethnic(2)	.014	.213	.004	1	.947	1.014
SECshortnew(1) by ethnic(3)	.102	.244	.174	1	.676	1.107
SECshortnew(1) by ethnic(4)	.298	.394	.572	1	.449	1.347
SECshortnew(1) by ethnic(5)	.717	.239	8.973	1	.003	2.049
SECshortnew(1) by ethnic(6)	.448	.281	2.536	1	.111	1.565
SECshortnew(1) by ethnic(7)	-.035	.243	.021	1	.886	.966
SECshortnew(2) by ethnic(1)	.129	.217	.352	1	.553	1.138
SECshortnew(2) by ethnic(2)	.159	.215	.553	1	.457	1.173
SECshortnew(2) by ethnic(3)	.519	.237	4.823	1	.028	1.681
SECshortnew(2) by ethnic(4)	1.009	.366	7.609	1	.006	2.743
SECshortnew(2) by ethnic(5)	.782	.263	8.812	1	.003	2.186
SECshortnew(2) by ethnic(6)	.447	.231	3.750	1	.053	1.563
SECshortnew(2) by ethnic(7)	.418	.239	3.057	1	.080	1.519
SECshortnew(3) by ethnic(1)	.124	.227	.299	1	.585	1.132
SECshortnew(3) by ethnic(2)	.095	.229	.170	1	.680	1.099
SECshortnew(3) by ethnic(3)	.191	.252	.576	1	.448	1.211
SECshortnew(3) by ethnic(4)	.689	.374	3.388	1	.066	1.991
SECshortnew(3) by ethnic(5)	.290	.288	1.020	1	.313	1.337
SECshortnew(3) by ethnic(6)	.735	.255	8.330	1	.004	2.085
SECshortnew(3) by ethnic(7)	.242	.250	.936	1	.333	1.273
Constant	.551	.040	190.940	1	.000	1.734

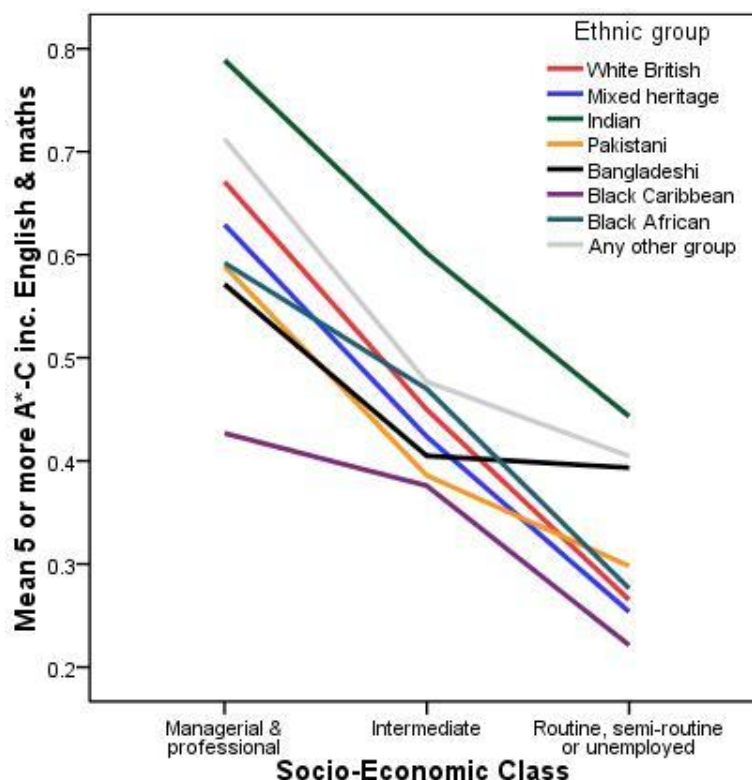
If we wanted to evaluate the ethnic gaps among students from middle SEC homes we would simply follow the same logic as above, recoding our values so that middle SEC was the first (or last) category and setting the reference category appropriately.

Putting it all together - viewing the interactions graphically

What is most helpful in understanding these interactions is to plot the data graphically. This gives us an easy visual indicator to help in interpreting the regression output and the nature of any interaction effects. We have done this in **Figure 4.13.3** below. You can use **Graphs > Legacy Dialogs > Line** to create this graph or alternatively you can use the syntax below (see the **Foundation Module** if you require further guidance). Here we have plotted the actual means for *fiveem*, but you could equally plot the predicted probabilities if you saved them from the model (see **Page 4.11**). Note that in the graph we have omitted cases where SEC is missing by returning the missing value for SECshort to '0' before requesting the graph.

`GRAPH /LINE(MULTIPLE)=MEAN(fiveem) BY SECshort BY ethnic.`


Figure 4.13.3: Mean Number of Students with Five or More A*-C grades (inc. English and Maths) by SEC and Ethnicity



The line graph shows a clear interaction between SEC and ethnicity. If the two explanatory variables did not interact we would expect all of the lines to have approximately the same slope (for example, the lines on the graph would be parallel when there is no interaction effect) but it seems that the effect of SEC on *fiveem* is different for different ethnic groups. For example the relationship appears to be very linear for White British students (blue line) – as the socio-economic group becomes more affluent the probability of *fiveem* increases. This not the case for all of the ethnic groups. For example, with regard to Black Caribbean students there is a big increase in *fiveem* as we move from low SEC to intermediate SEC, but a much smaller increase as we move to high SEC. As you (hopefully) can see, the line graph is a good way of visualizing an interaction between two explanatory variables.

Now that we have seen how to create and interpret out logistic regression models both with and without interaction terms we must again turn our attention to the important business of checking that the assumptions underlying our model are met and that the results are not misleading due to any extreme cases.

4.14 Model diagnostics

On **Page 4.9** we discussed the assumptions and issues involved with logistic regression and were relieved to find that they were largely familiar to us from when we tackled multiple linear regression! Despite this, testing them can be rather tricky. We will now show you how to perform these diagnostics using SPSS based on the model we used as an example on **Page 4.11** (using the MLR LSYPE 15,000  dataset).

Linearity of Logit

This assumption is confusing but it is not usually an issue. Problems with the linearity of the logit can usually be identified by looking at the model fit and pseudo R^2 statistics (Nagelkerke R^2 , see **Page 4.12**, **Figure 4.12.4**). The Hosmer and Lemeshow test, which as you may recall was discussed on **Page 4.12** and shown as SPSS output in **Figure 4.12.5** (reprinted below) is a good test of how well your model fits the data. If the test is not statistically significant (as is the case with our model here!) you can be fairly confident that you have fitted a good model.

Figure 4.14.1: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	4.674	8	.792

With regard to the Nagelkerke R^2 you are really just checking that your model is explaining a reasonable amount of the variance in the data. Though in this case the value of .159 (about 16%) is not high in absolute terms it is highly statistically significant.

Of course this approach is not perfect. Field (2009, p.296, see **Resources**) suggests an altogether more technical approach to testing the linearity of the logit if you wish to have more confidence that your model is not violating its assumptions.

Independent Errors

As we mentioned on **Page 4.9** checking for this assumption is only really necessary when data is clustered hierarchically and this is beyond the scope of this website. We thoroughly recommend our sister site LEMMA (see **Resources**) if you want to learn more about this.

Multicollinearity

It is important to know how to perform the diagnostics if you believe there might be a problem. The first thing to do is simply create a correlation matrix and look for high coefficients (those above .8 may be worthy of closer scrutiny). You can do this very simply on SPSS: **Analyse > Correlate > Bivariate** will open up a menu with a single window and all you have to do is add all of the relevant explanatory variables into it and click **OK** to produce a correlation matrix.

If you are after more detailed colinearity diagnostics it is unfortunate that SPSS does not make it easy to perform them when creating a logistic regression model (such a shame, it was doing so well after including the 'interaction' button). However, if you recall, it is possible to collect such diagnostics using the menus for multiple linear regression (see **Page 3.14**)... because the tests of multicollinearity are actually independent of the type of regression model you are making (they examine only the explanatory variables) you can get them from running a multiple linear regression using the exact same variables as you used for your logistic regression. Most of the output will be meaningless because the outcome variable is not continuous (which violates a key assumption of linear regression methods) but the multicollinearity diagnostics will be fine! Of course we have discussed this whole issue in the previous module (**Page 3.14**).

Influential Cases

On **page 4.11** we showed you how to request the model's residuals and the *Cook's distances* as new variables for analysis. As you may recall from the previous module (**Page 3.14**), if a case has a Cook's distance greater than one it may be unduly influencing your model. Requesting the Cook's distance will have created a new variable in your dataset called *COO_1* (note that this might be different if you have created other variables in previous exercises – this is why clearly labeling variables is so useful!). To check that we have no cases where Cook's distance is greater than one we can simply look at the frequencies: **Analyse > Descriptive Statistics > Frequencies**, add *Coo_1* into the window, and click **OK**. Your output, terrifyingly, will look something like **Figure 4.14.2** (only much, much longer!):

Figure 4.14.2: Frequency of Cook's distance for model

Cook's influence statistics

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00014	24	.2	.2	.2
	.00018	466	3.0	3.8	4.0
	.00023	734	4.7	5.9	9.9
	.00023	20	.1	.2	10.1
	.00025	346	2.2	2.8	12.9
	.00026	130	.8	1.1	13.9
	.00026	40	.3	.3	14.3
	.00029	414	2.6	3.4	17.6
	.00031	444	2.8	3.6	21.2
	.00033	1	.0	.0	21.2
	.00033	31	.2	.3	21.5
	.00034	53	.3	.4	21.9
	.00035	13	.1	.1	22.0
	.00035	77	.5	.6	22.6
	.00036	650	4.1	5.3	27.9
	.00039	23	.1	.2	28.1
	.00040	349	2.2	2.8	30.9

This is not so bad though – remember we are looking for values greater than one and these values are in order. If you scroll all the way to the bottom of the table you will see that the highest value Cook's distance is less than .014... nowhere near the level of 1 at which we need to be concerned.

Finally we have reached the end of our journey through the world of Logistic Regression. Let us now take stock and discuss how you might go about pulling all of this together and reporting it.

4.15 Reporting the results of logistic regression

Our interest here has been not only in the association between ethnic group, social class, gender and exam achievement, but also how the relationship between ethnic group and exam achievement changes as we account for other explanatory variables (like SEC) and interaction effects. It is therefore appropriate to present the results not just for the last model but also for the preceding models. In a report we would present the results as shown in the table below.

Model 1 shows the simple association between ethnic group and the *fiveem* outcome. Model 2 shows what happens when we add SECshort and gender to the model. Model 3 shows the significant interaction that exists between ethnic group and SECshort which needs to be taken into account. Summarising the results of the three models alongside each other in this way lets you tell the story of your analysis and show how your modeling developed.

The elements of this table (**Figure 4.15.1**) that you choose to discuss in more detail in your text will depend on the precise nature of your research question, but as you can see it provides a fairly concise presentation of nearly all of the key relevant statistics.

Figure 4.15.1: reporting the results of logistic regression

Variable	Model 1			Model 2			Model 3		
	B	SE	OR	B	SE	OR	B	SE	OR
Constant	-1.100	.020	.91	-1.088	.043	.34	-1.184	.051	.31
Ethnic group									
Mixed heritage	-.142	.076	.87	-.133	.079	.88	-.079	.171	.92
Indian	.458 ***	.067	1.58	.678 ***	.070	1.97	.784 ***	.128	2.19
Pakistani	-.447 ***	.071	.64	-.138	.074	.87	.152	.125	1.16
Bangladeshi	-.222 **	.079	.80	.223 **	.082	1.25	.544 ***	.123	1.72
Black Caribbean	-.626 ***	.093	.53	-.632 ***	.097	.53	-.246	.210	.78
Black African	-.190 *	.086	.83	-.010	.091	.99	.068	.156	1.07
Any other group (base = White British)	.188 *	.082	1.21	.335 ***	.086	1.40	.605 ***	.163	1.83
Gender									
Female (base= male)				.342 ***	.034	1.41	.340 ***	.035	1.40
Socio-Economic Class (SEC)									
Missing				.476	.053	1.61	.551 ***	.071	1.74
High				1.585	.049	4.88	1.735 ***	.061	5.67
Medium (base= low)				.705	.048	2.02	.818 ***	.062	2.27
Interaction ethnic * SEC									
Mixed heritage * missing SEC							-.005	.251	1.00
Mixed heritage * high SEC							-.129	.217	.88
Mixed heritage * medium SEC							-.019	.235	.98
Indian * missing SEC							-.065	.198	.94
Indian * high SEC							-.159	.215	.85
Indian * medium SEC							-.145	.179	.86
Pakistani * missing SEC							-.328	.198	.72
Pakistani * high SEC							-.519 *	.237	.59
Pakistani * medium SEC							-.418 *	.187	.66
Bangladeshi * missing SEC							-.320	.190	.73
Bangladeshi * high SEC							-1.009 **	.366	.36
Bangladeshi * medium SEC							-.711 **	.226	.49
Black Caribbean * missing SEC							-.492	.318	.61
Black Caribbean * high SEC							-.782 **	.263	.46
Black Caribbean * medium SEC							-.065	.276	.94
Black African * missing SEC							.288	.246	1.33
Black African * high SEC							-.447	.231	.64
Black African * medium SEC							.001	.273	1.00
Any other * missing SEC							-.177	.241	.84
Any other * high SEC							-.418	.239	.66
Any other * medium SEC							-.453	.234	.64
-2LL	20652			19335			19291		
	$\chi^2 = 165, df=7, p < .001$			$\chi^2 = 1317, df=4, p < .001$			$\chi^2 = 44, df=21, p < .001$		
Nagelkerke R ²	1.5%			12.5%			12.9%		
Hosmer & Lemeshow test	$p=1.00$			$p=0.026$			$p=0.535$		
Classification accuracy	54.7%			63.8%			64.0%		

If you want to see an example of a published paper presenting the results of a logistic regression see:


- Strand, S. & Winston, J. (2008). Educational aspirations in inner city schools. *Educational Studies*, 34, (4), 249-267.

Conclusion

We hope that now you have braved this module you are confident in your knowledge about what logistic regression is and how it works. We hope that you are confident about creating and interpreting your own logistic regression models using SPSS. Most of all we hope that all of the formula has not frightened you away... Logistic regression can be an extremely useful tool for educational research, as we hope our LSYPE example has demonstrated, and so getting to grips with it can be a very useful experience! Whew... why not have a little lie down (and perhaps a stiff drink) and then return to test your knowledge with our quiz and exercise?

Exercise

We have seen that prior attainment, specifically age 11 test score, is a strong predictor of later achievement. Maybe some of the ethnic, social class and gender differences in achievement at age 16 reflect differences that were already apparent in attainment at age 11? This would have substantive implications for education policy, because it would indicate that attention would need to focus as much on what has happened during the primary school years up to age 11 as on the effects of education during the secondary school years up to age 16.

Use the LSYPE 15,000 dataset  to work through each of the following questions. Answer them in full sentences with supporting tables or graphs where appropriate as this will help when you to better understand how you may apply these techniques to your own research. The answers are on the next page.

Note: *The variable names as they appear in the SPSS dataset are listed in brackets. We have also included some hints in italics.*

Question 1

Exam score at age 11 is included in the LSYPE dataset as *ks2stand*. Before we include it in our model we need to know how it is related to our outcome variable, *fiveem*. Graphically show the relationship between *ks2stand* and *fiveem*.

Use a bar chart.

Question 2

In this module we have established that ethnicity, Socio-economic class and gender can all be used to predict whether or not students pass 5 exams with grades A-C at age 16 (our trusty *fiveem*). Does including age 11 exam score (*ks2stand*) to this main effects model as an explanatory variable make a statistically significant contribution to predicting *fiveem*?

Run a logistic regression model with these variables:

Outcome: fiveem

Explanatory: ethnic, SECshort, gender, ks2stand

Question 3

Does adding age 11 score as an explanatory variable substantially improve the fit of the model? That is to say, does it improve how accurately the model predicts whether or not students achieve *fiveem*?

Run the analysis again but this time with two blocks, including ethnic, SECshort and gender in the first block and ks2stand in the second. Examine the -2LL and pseudo-R² statistics.

Question 4

Following on from question 3, what difference (if any) does adding age 11 score to the model make to the ethnic, gender and SEC coefficients? What is your interpretation of this result?

No need to carry out further analysis – just examine the ‘Variables in the equation’ tables for each block.

Question 5

Is there an interaction effect between SEC and age 11 score?

*Run the model again but this time including a SECshort*ks2stand interaction. You may also wish to graph the relationship between age 11 exam score and SEC in the context of fiveem.*

Question 6

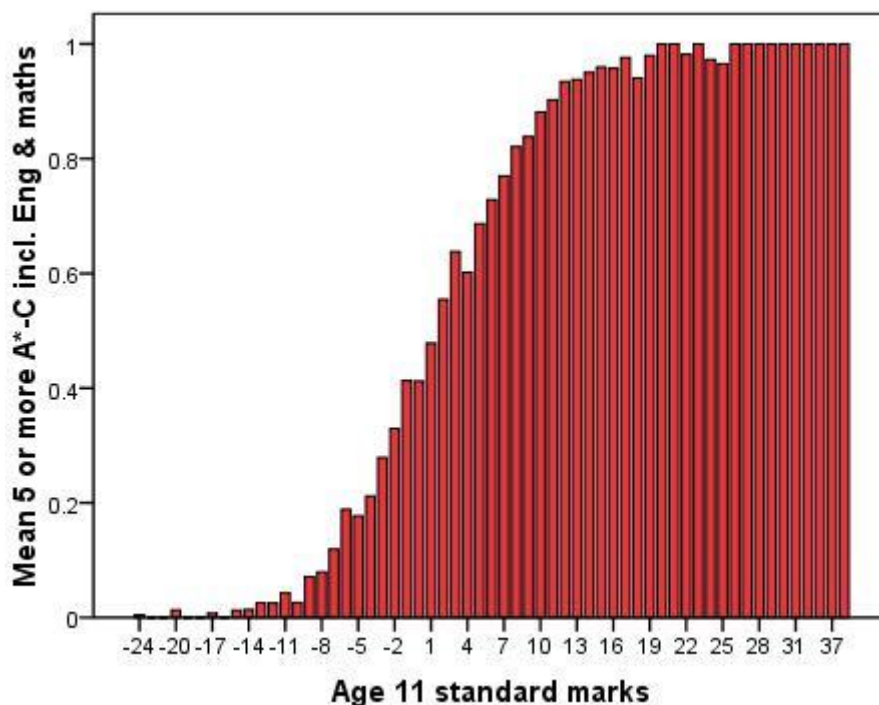
Are there any overly influential cases in this model?

You will need to get the Cook’s distances for each case. This may require you to re-run the model with different option selections.

Answers

Question 1

There is arguably more than one way to achieve this but we have gone for a bar graph which uses *ks2stand* as the category (X axis) and the mean *fiveem* score as the main variable (Y axis). If you have forgotten how to do this on SPSS we run through it as part of the **Foundation Module**.



The mean *fiveem* score (which varies between 0 and 1) provides an indication of how many students passed five or more GCSEs (including maths and English) at each level of age 11 standard score. As you can see, the mean score is far lower for those with lower Age 11 scores but right up at the maximum of 1 for those with the highest scores. Note that the shape of the graph matches the 'sigmoid' shape for binary outcomes which we have seen throughout this module. Based on this graph there appears to be strong evidence to suggest that age 11 is a good predictor of *fiveem*.

Question 2

The table below shows that the explanatory variable for age 11 score is indeed a statistically significant predictor of *fiveem* (this can be ascertained from the 'sig.' column). The 'Exp(B)' column shows that the odds ratio is 1.273; meaning that a one unit change in age 11 score (an increase of 1 point) changes the odds of achieving *fiveem* increase by a multiplicative factor of 1.273. This is very substantial when you consider how large the range of possible age 11 scores is!

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	ethnic			238.396	7	.000	
	ethnic(1)	-.073	.120	.376	1	.540	.929
	ethnic(2)	1.295	.107	145.134	1	.000	3.650
	ethnic(3)	.820	.118	48.675	1	.000	2.270
	ethnic(4)	.812	.136	35.610	1	.000	2.253
	ethnic(5)	-.030	.143	.045	1	.832	.970
	ethnic(6)	.986	.158	38.725	1	.000	2.680
	ethnic(7)	.771	.147	27.579	1	.000	2.161
	SECshort			200.476	2	.000	
	SECshort(1)	.953	.067	199.694	1	.000	2.592
	SECshort(2)	.471	.066	50.606	1	.000	1.602
	gender(1)	.521	.053	97.306	1	.000	1.684
	ks2stand	.241	.005	2610.550	1	.000	1.273
	Constant	-1.276	.062	418.673	1	.000	.279

a. Variable(s) entered on step 1: ks2stand.

See **Page 4.5** if you want to review how to interpret logistic regression coefficients for continuous explanatory variables.

Question 3

In order to explore just how much impact adding age 11 exam score as an explanatory variable had on the model we re-ran the model but entered *ks2stand* as a second block (block 2). The classification Tables, Nagelkerke R² and omnibus tests can then be compared across these blocks to assess the impact of accounting for prior achievement.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step	1196.751	10	.000
Block	1196.751	10	.000
Model	1196.751	10	.000

Model Summary

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
	14447.731 ^a	.100	.134

Classification Table^a

Observed	Fiveem	Predicted		% Correct
		Fiveem		
		no	yes	
Fiveem	no	4415	1509	74.5
	yes	2554	2826	52.5
Overall %				64.1

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step	5391.043	1	.000
Block	5391.043	1	.000
Model	6587.794	11	.000

Model Summary

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
	9056.688 ^a	.442	.589

Classification Table^a

Observed	Fiveem	Predicted		% Correct
		Fiveem		
		no	yes	
Fiveem	no	4902	1022	82.7
	yes	1133	4247	78.9
Overall %				80.9

As we have discussed on **Page 4.12**, the omnibus test tells us whether or not our model is better at predicting the outcome than the 'baseline' model (which always predicts whichever of the two outcomes was more frequent in the data). The 'Block' row tells us whether the new block significantly improves the number of correct predictions compared to the previous block. As you can see, the omnibus test table for Block 2 indicates that the addition of *ks2stand* does indeed improve the accuracy of predictions to a statistically significant degree (highlighted - 'Sig' is <.05). The 'Model' row test is also significant for both blocks, suggesting both are better at predicting the outcome than the baseline model.

The classification table (third table down) shows us just how much more accurately block 2 describes the data compared to block 1. The model defined as block 1 correctly classifies 64.1% of cases – an improvement over the baseline model but still not great. The inclusion of ks2stand in block 2 increases the number of correct classifications substantially, to 80.9%. Finally, the Model Summary table helps to confirm this. The deviance (-2LL) is substantially lower for block 2 than for block 1 and the Nagelkerke pseudo R^2 is .589 (59% of variance explained) for block 2 compared to .134 (13% of variance explained) for block 1. Overall, age 11 score is clearly very important for explaining age 16 exam success!

Question 4

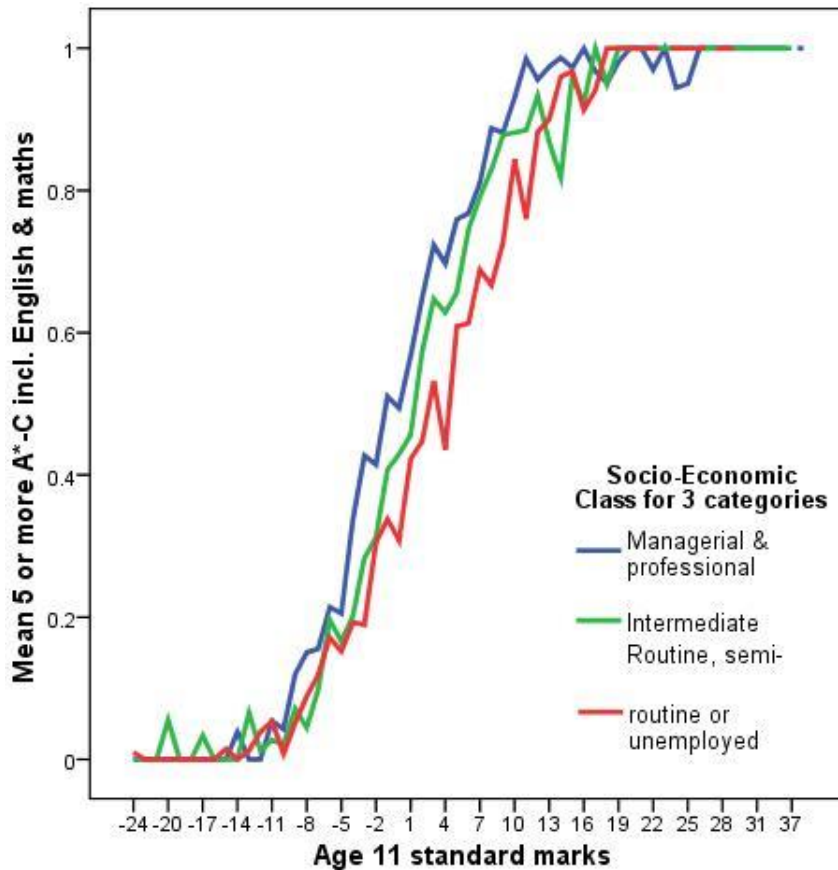
Below you will see the 'Variables in the Equation' table for block 2 with the 'sig' and Exp(B) columns from the same table for block 1. These are taken from the same SPSS output that we generated for question 3. Notice how in most cases the odds ratios [Exp(B)] are less in block 2 than they were in block 1. In addition, note that three explanatory variables are not statistically significant in both versions of the model.

For example, as highlighted, students from a managerial or professional family background [SECshort(1)] are 4.7 times more likely to achieve *fiveem* than those from routine, semi-routine or unemployed family backgrounds for our model excluding age 11 exam score. However, when we include age 11 score in block 2 the odds ratio is much lower: those from the wealthiest group are only 2.6 times more likely to achieve *fiveem* than those from the least wealthy group. This suggests that some the difference in success rate that appears to be due to social-class differences between students may in fact be explained by prior attainment. There may also be an interaction between prior attainment and social class.

	Block 2						Block 1	
	B	S.E.	Wald	df	Sig.	Exp(B)	Sig.	Exp(B)
ethnic			238.396	7	.000		.000	
ethnic(1)	-.073	.120	.376	1	.540	.929	.304	.911
ethnic(2)	1.295	.107	145.134	1	.000	3.650	.000	2.043
ethnic(3)	.820	.118	48.675	1	.000	2.270	.590	.954
ethnic(4)	.812	.136	35.610	1	.000	2.253	.002	1.371
ethnic(5)	-.030	.143	.045	1	.832	.970	.000	.557
ethnic(6)	.986	.158	38.725	1	.000	2.680	.186	1.173
ethnic(7)	.771	.147	27.579	1	.000	2.161	.000	1.465
SECshort			200.476	2	.000		.000	
SECshort(1)	.953	.067	199.694	1	.000	2.592	.000	4.696
SECshort(2)	.471	.066	50.606	1	.000	1.602	.000	1.973
gender(1)	.521	.053	97.306	1	.000	1.684	.000	1.466
ks2stand	.241	.005	2610.550	1	.000	1.273		
Constant	-1.276	.062	418.673	1	.000	.279	.000	.329

Question 5

Before exploring the interaction statistically it is worth first examining the relationship by looking at a line graph (though remembering that this graph does not account for the influence of other explanatory variables such as gender and ethnicity):



As you may recall from **Page 4.13**, if the lines in this graph were approximately parallel we would expect there to be *no* interaction. Though this does not appear to be the case here the interaction is certainly not clear cut... you could argue that in the middle range of age 11 scores there is a clear ordering by affluence whereby the wealthiest group has the highest pass rate and the least wealthy the lowest. There are considerable fluctuations, particularly at the upper and lower ends of the age 11 score range, but these may not be as important as they appear given the large range of possible age 11 score values.

Let us check the 'Variables in the equation' for the logistic regression model when we include a $ks2stand*SECshort$ interaction term. As you can see it doesn't actually help

that much! Though the significance level is greater than the commonly used 5% ($p < .05$) it is still less than 10%. The comparison between the lower SEC group and the intermediate SEC group (for White-British males only) is statistically significant at the .05 level. Deciding whether or not to include this interaction in your final model would be a judgment call. Given that the interaction effect does not appear to be particularly pronounced (the odds ratios are relatively small for the interactions and the odds ratios for the other variables have changed very little) we would probably not include it for the sake of parsimony.

	B	S.E.	Wald	df	Sig.	Exp(B)
ethnic			237.343	7	.000	
ethnic(1) [Mixed Heritage]	-.070	.120	.341	1	.559	.932
ethnic(2) [Indian]	1.292	.107	145.240	1	.000	3.639
ethnic(3) [Pakistani]	.815	.117	48.496	1	.000	2.259
ethnic(4) [Bangladeshi]	.804	.135	35.413	1	.000	2.235
ethnic(5) [Black Caribbean]	-.028	.143	.038	1	.846	.973
ethnic(6) [Black African]	.978	.157	38.563	1	.000	2.659
ethnic(7) [Any other]	.772	.146	27.823	1	.000	2.164
SECshort			189.510	2	.000	
SECshort(1) [Managerial & Pro]	.936	.068	189.160	1	.000	2.550
SECshort(2) [Intermediate]	.454	.067	46.554	1	.000	1.575
gender(1)	.520	.053	96.920	1	.000	1.683
ks2stand	.227	.008	823.286	1	.000	1.255
SECshort * ks2stand			4.724	2	.094	
SECshort(1) by ks2stand	.019	.011	2.918	1	.088	1.019
SECshort(2) by ks2stand	.023	.011	4.031	1	.045	1.023
Constant	-1.2E0	.062	419.773	1	.000	.282

Question 6

As we saw on **Page 4.14**, cases that are having a particularly large influence on a model can be identified by requesting a statistic called Cook's distance. If this statistic is greater than 1 for a given case than that case may be a an outlier that is powerful enough to unduly influence a model (this is usually a more significant issue when you have a smaller sample. The table below shows the cooks distances produced by the interaction model we created in Question 5. We have completely removed the middle section because it is a horrendously long table. The Cook's distances are in the left-most column and listed in numerical order. As you can see largest vale (at the bottom of the table) is .05158. This is much lower than a value of 1, which would usually be cause for concern. It seems that our model has no cases that are overly influential.

Analog of Cook's influence statistics

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00000	1	.0	.0	.0
	.00000	1	.0	.0	.0
	.00000	1	.0	.0	.0
	.04017	1	.0	.0	100.0
	.04183	1	.0	.0	100.0
	.05158	1	.0	.0	100.0
	Total	11304	71.7	100.0	
Missing	System	4466	28.3		
Total		15770	100.0		