

# Module 3 - Multiple Linear Regressions

## Objectives

- Understand the strength of Multiple linear regression (MLR) in untangling cause and effect relationships
- Understand how MLR can answer substantive questions within the field of educational research, using the LSYPE dataset for examples
- Understand the assumptions underlying MLR and how to test they are met
- Understand how to explore interaction effects between variables
- Be able to implement and interpret MLR analyses using SPSS
- Appreciate the applications of MLR in educational research, and possibly in your own research

## **Start Module 3: Multiple Linear Regression**

Using multiple explanatory variables for more complex regression models.

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 exercises and the quiz to get a firm understanding.

## **Contents**

- 3.1 Overview
- 3.2 The Multiple Linear Regression Model
- 3.3 Assumptions of Multiple Linear Regression
- 3.4 Using SPSS to model the LSYPE data
- 3.5 A model with a continuous explanatory variable (Model 1)
- 3.6 Adding dichotomous nominal explanatory variables (Model 2)
- 3.7 Adding nominal variables with more than two categories (Model 3)
- 3.8 Predicting scores using the regression model
- 3.9 Refining the model: treating ordinal variables as dummy variables (Model 4)
- 3.10 Comparing coefficients across models
- 3.11 Exploring interactions between a dummy and a continuous variable (Model 5)
- 3.12 Exploring interactions between two nominal variables (Model 6)
- 3.13: A value added model (Model 7)
- 3.14 Model diagnostics and checking your assumptions
- 3.15 Reporting your results

Quiz

Exercise

## 3.1 Overview

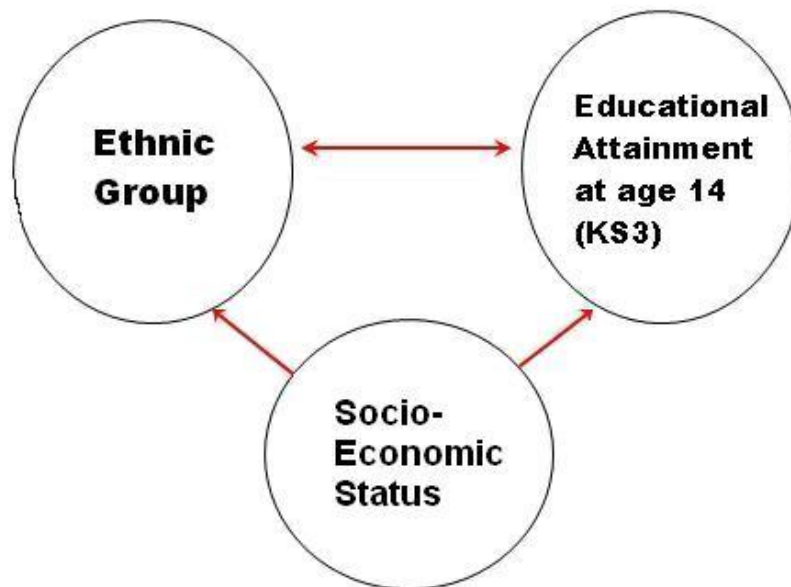
### What is multiple linear regression?

In the previous module we saw how simple linear regression could be used to predict the value of an outcome variable based on the value of a suitable explanatory variable. This is a useful technique but it is limited - usually a number of different variables will predict an outcome. For example, how good a student is at football is not just related to how many hours they practice a week. There is likely to be a relationship between ability and practice but discussing this in isolation from other important variables would most likely be a considerable over-simplification. The young player's spatial-awareness and physical fitness are also likely to contribute to their overall level of ability. Their ability may partly stem from personality traits that are related to confidence and teamwork.

Of course we can go even further than this and say that sometimes the explanatory variables can influence each other as well as the outcome itself! For example, the impact of training on ability is bound to be dependent on the level of motivation the student feels. Perhaps they are turning up to training but not putting any effort in because they don't really like football! The real world is very complicated but luckily, with regression analysis, we can at least partially model that complexity to gain a better understanding. Multiple linear regression (MLR) allows the user to account for multiple explanatory variables and therefore to create a model that predicts the specific outcome being researched. Multiple linear regression works in a very similar way to simple linear regression.

Consider the example of understanding educational attainment. It is well known that there is a strong and positive correlation between social class and educational attainment. There is evidence that pupils from some (though not all) minority ethnic groups do not achieve as well in the English education system as the majority White British group. However there is also a strong relationship between ethnic group and social class, with many minority ethnic groups experiencing higher socio-economic disadvantage than the White British group. It is therefore not possible to say from raw scores alone whether the lower attainment of some minority ethnic groups reflects something particular about belonging to that ethnic group or reflects the fact that some ethnic groups are particularly socially and economically disadvantaged. The relationship may look a little like the one presented below (**Figure 3.1.1**).


**Figure 3.1.1: The third variable problem**



Multiple regression offers a way to address these issues. If we put all the variables we have into one analysis, we can assess the impact of one factor when another is taken into account. Thus multiple regression can allow us to assess the association of ethnicity and attainment *after the variance in attainment associated with social class is taken into account*. A wide range of further variables can also be included to build up highly detailed and complex models, e.g. family composition, maternal educational qualifications, students' attitude to school, parents educational aspirations for the student, etc.

### **Running through the examples and exercises using SPSS**

As in the previous module, we provide worked examples from LSYPE which you can follow through using SPSS. You will find that we make a lot of transformations to the dataset as we perform the various analyses. It could get confusing! We recommend that you do not make the transformations yourself (one small error could dramatically alter your output) and instead use the pre-prepared variables in the MLR LSYPE 15,000 dataset. However if you really do want to perform all of the transformations yourself you can always use the original LSYPE 15,000 data file.

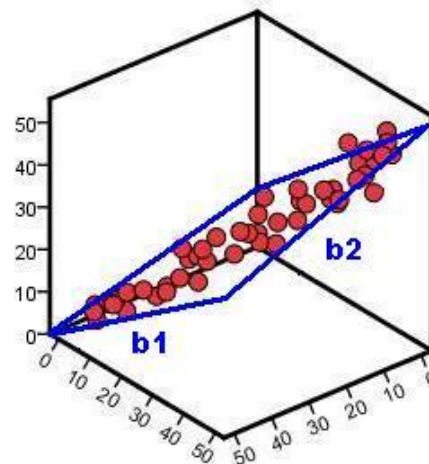
LSYPE 15,000 

MLR LSYPE 15,000 

## 3.2 The multiple regression model

The simple linear regression model is based on a straight line which has the formula  $\hat{Y} = a + bX$  (where **a** is the intercept and **b** is the gradient). You'll be relieved to hear that multiple linear regression also uses a linear model that can be formulated in a very similar way! Though it can be hard to visualize a linear model with two explanatory variables, we've had a go at showing you what it may look like by adding a 'plane' on the 3D scatterplot below (**Figure 3.2.1**). Roughly speaking, this plane models the relationship between the variables.

**Figure 2.2.1: A multiple regression plane**



The plane still has an intercept. This is the value of the outcome when both explanatory variables have values of zero. However there are now two gradients, one for each of the explanatory variables (**b<sub>1</sub>** on the x-axis and **b<sub>2</sub>** on the z-axis). Note that these gradients are the regression coefficients (B in the SPSS output) which tell you how much change in the outcome (Y) is predicted by a unit change in that explanatory variable. All we have to do to incorporate these extra explanatory variables in to our model is add them into the linear equation:

$$\hat{Y} = a + b_1x_1 + b_2x_2$$

As before, if you have calculated the value of the intercept and the two b-values you can use the model to predict the outcome  $\hat{Y}$  (pronounced “Y Hat” and used to identify the *predicted* value of Y for each case as distinct from the *actual* value of Y for the case) for any values of the explanatory variables ( $x_1$  and  $x_2$ ). Note that it is very difficult to visualize a scatterplot with more than two explanatory variables (it involves picturing four or more dimensions - something that sounds a bit 'Twilight Zone' to us and causes our poor brains to shut down...) but the same principle applies. You simply add a new **b** value (regression coefficient) for each additional explanatory variable:

$$\hat{Y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

Potentially you can include as many variables as you like in a multiple regression but realistically it depends on the design of your study and the characteristics of your sample.

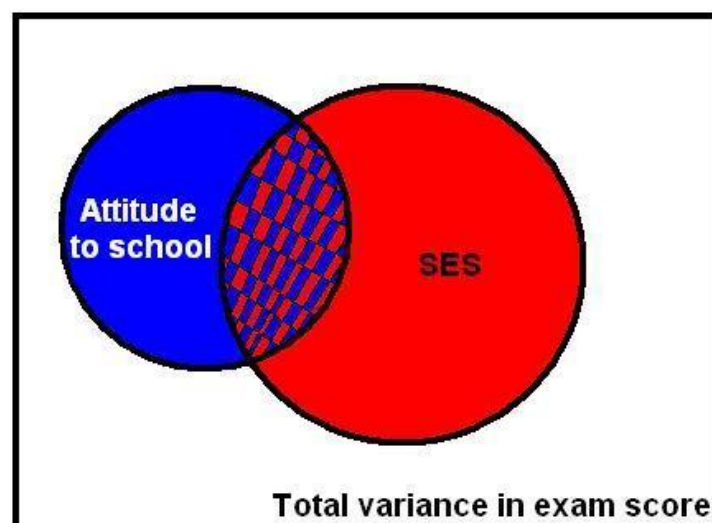
### **Multiple r and $r^2$**

For simple linear regression it was important to look at the correlation between the outcome and explanatory variable (Pearson's  $r$ ) and the  $r^2$  (the coefficient of determination) to ascertain how much of the variation in the outcome could be explained by the explanatory variable. Similar statistics can be calculated to describe multiple regression models.

Multiple  $r$  is the equivalent of Pearson's  $r$  though rather than representing the magnitude and direction of a relationship between two variables it shows the strength of the relationship between the outcome variable and the values predicted by the model as a whole. This tells us how well the model predicts the outcome (sometimes researchers say how well the model fits the data). A multiple  $r$  of 1 means a perfect fit while a multiple  $r$  of 0 means the model is very poor at predicting the outcome.

The  $r^2$  can be interpreted in the exact same way as for simple linear regression: it represents the amount of variation in the outcome that can be explained by the model, although now the model will include multiple explanatory variables rather than just one. The diagram below (**Figures 3.2.2** - lovingly prepared on 'MS Paint') might help you to visualize  $r^2$ . Imagine the variance in the outcome variable 'Exam Grade' is represented by the whole square and 'SES' (socio-economic status) and 'Attitude to School' are explanatory variables, with the circles representing the variance in exam grade that can be explained or accounted for by each.

**Figure 3.2.2: SES and Attitude to School predicting Exam Grade**



In **Figure 2.2.2** the square represents the total variation in exam score in our sample. The red circle represents the variance in exam score that can be predicted (we might say explained) by SES. Now we add a further variable, the blue circle - attitude to school. This variable also explains a large proportion of the variance in exam score. Because attitude to school and SES are themselves related, some of the variance in exam score that can be explained by attitude is already explained by SES (hatched red and blue area). However, attitude can also explain some unique variance in exam score that was not explained by SES. The red, blue and hatched areas combined represent  $r^2$ , the total variance in exam score explained by the model. This is greater than would be accounted for by using either SES or attitude to school on its own.

### **Methods of Variable Selection**

When creating a model with more than one explanatory variable a couple of complications arise. Firstly, we may be unsure about which variables to include in the model. We want to create a model which is detailed and accounts for as much of the variance in the outcome variable as possible but, for the sake of parsimony, we do not want to throw *everything* in to the model. We want our model to be elegant, including only the relevant variables. The best way to select which variables to include in a model is to refer to previous research. Relevant empirical and theoretical work will give you a good idea about which variables to include and which are irrelevant.

Another problem is correlation between explanatory variables. When there is correlation between two explanatory variables it can be unclear how much of the variance in the outcome is being explained by each. For example, the hatched area in **Figure 3.2.2** represents the variance in exam score which is shared by both SES and attitude to school. It is difficult to ascertain which variable is foremost in accounting for this shared variance because the two explanatory variables are themselves correlated. This becomes even more complicated as you add more explanatory variables to the model!

It is possible to adjust a multiple regression model to account for this issue. If the model is created in steps we can better estimate which of the variables predicts the largest change in the outcome. Changes in  $r^2$  can be observed after each step to find out how much the predictive power of the model improves after each new explanatory variable is added. This means a new explanatory variable is added to the model only if it explains some unique variance in the outcome that is not accounted for by variables already in the model (for example, the blue or red section in **Figure 2.2.2**).

**SPSS** allows you to alter how variables are entered and also provides options which allow the computer to sort out the entry process for you. The controls for this are shown below, but we'll go into the overall process of doing a multiple regression

analysis in more detail over the coming pages. For now it is worth examining what these different methods of variable selection are.

### **Stepwise/Forward/Backward**

We've grouped these methods of entry together because they use the same basic principle. Decisions about the explanatory variables added to the model are made by the computer based entirely on statistical criteria.

The **Forward** method starts from scratch - the computer searches from the specified list of possible explanatory variables for the one with the strongest correlation with the outcome and enters that first. It continues to add variables in order of how much additional (unique) variance they explain. It only stops when there are no further variables that can explain additional (unique) variance that is not already accounted for by the variables already entered.

The **Backward** method does the opposite - it begins with all of the specified potential explanatory variables included in the model and then removes those which are not making a significant contribution.

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

We don't generally recommend using stepwise methods! As **Field (2010)** observes, they take important decisions away from the researcher by making decisions based solely on mathematical criteria (related entirely to your specific dataset), rather than on broader theory from previous research! They can be useful if you are starting from scratch with no theory but such a scenario is rare.

### **Enter/Remove**

The 'Enter' method allows the researcher to control how variables are entered into the model. At the simplest level all the variables could be entered together in a single group called a 'block'. This makes no assumptions about the relative importance of each explanatory variable. However variables can be entered in separate blocks of explanatory variables. In this 'hierarchical' regression method the researcher enters explanatory variables into the model grouped in blocks in order of their theoretical relevance in relation to the outcome. Decisions about the blocks are made by the researcher based on previous research and theoretical reasoning. Generally knowing the *precise* order of importance is not possible, which is why variables that are considered of similar importance are entered as a single block. **Enter** will include all variables in the specified block while **Remove** removes all variables in the specified block.



Some of this may sound confusing. Don't worry too much if you don't get it straight away - it will become clearer when you start running your own multiple regression analyses. The main thing is that you have some understanding about what each entry method does.

### 3.3 Assumptions of multiple linear regression

The assumptions for multiple linear regression are largely the same as those for simple linear regression models, so we recommend that you revise them on **Page 2.6**. However there are a few new issues to think about and it is worth reiterating our assumptions for using multiple explanatory variables.

**Linear relationship:** The model is a roughly linear one. This is slightly different from simple linear regression as we have multiple explanatory variables. This time we want the outcome variable to have a roughly linear relationship with each of the explanatory variables, taking into account the other explanatory variables in the model.

**Homoscedasticity:** Ahhh, homoscedasticity - that word again (just rolls off the tongue doesn't it)! As for simple linear regression, this means that the variance of the residuals should be the same at each level of the explanatory variable/s. This can be tested for each separate explanatory variable, though it is more common just to check that the variance of the residuals is constant at all levels of the predicted outcome from the full model (i.e. the model including all the explanatory variables).

**Independent errors:** This means that residuals should be uncorrelated.

#### Other important things to consider

As with simple regression, the assumptions are the most important issues to consider but there are also other potential problems you should look out for:

**Outliers/influential cases:** As with simple linear regression, it is important to look out for cases which may have a disproportionate influence over your regression model.

**Variance in all explanatory variables:** It is important that your explanatory variables... well, vary! Explanatory variables may be continuous, ordinal or nominal but each must have at least a small range of values even if there are only two categorical possibilities.

**Multicollinearity:** Multicollinearity exists when two or more of the explanatory variables are highly correlated. This is a problem as it can be hard to disentangle which of them best explains any shared variance with the outcome. It also suggests that the two variables may actually represent the same underlying factor.

**Normally distributed residuals:** The residuals should be normally distributed.

We've moved through these issues quite quickly as we have tackled most of them before. You can review the simple linear regression assumptions on **page 2.6** if you feel a little bit rusty.

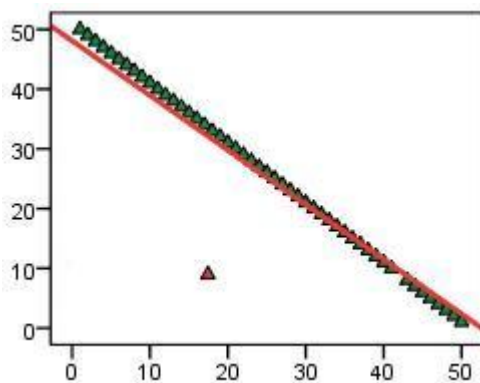
#### Checking the assumptions

Here is an assumptions checklist for multiple regression:

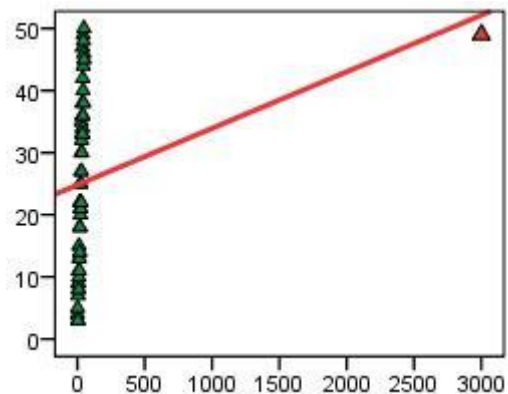
**1. Linear relationships, outliers/influential cases:** This set of assumptions can be examined to a fairly satisfactory extent simply by plotting scatterplots of the relationship between each explanatory variable and the outcome variable. It is important that you check that each scatterplot is exhibiting a linear relationship between variables (perhaps adding a regression line to help you with this). Alternatively you can just check the scatterplot of the actual outcome variable against the predicted outcome.

Now that you're a bit more comfortable with regression and the term residual you may want to consider the difference between outliers and influential cases a bit further. Have a look at the two scatterplots below (**Figures 3.3.1 & 3.3.2**):

**Figure 3.3.1: Scatterplot showing a simple outlier**



**Figure 3.3.2: Scatterplot showing an outlier that is an influential case**



Note how the two problematic data points influence the regression line in differing ways. The simple outlier influences the line to a far lesser degree but will have a very large residual (distance to the regression line). SPSS can help you spot outliers by identifying cases with particularly large residuals. The influential case outlier dramatically alters the regression line but might be harder to spot as the residual is small - smaller than most of the other more representative points in fact! A case this extreme is very rare! As well as examining the scatterplot you can also use *influence statistics* (such as the Cook's distance statistic) to identify points that may unduly influence the model. We will talk about these statistics and how to interpret them during our example.

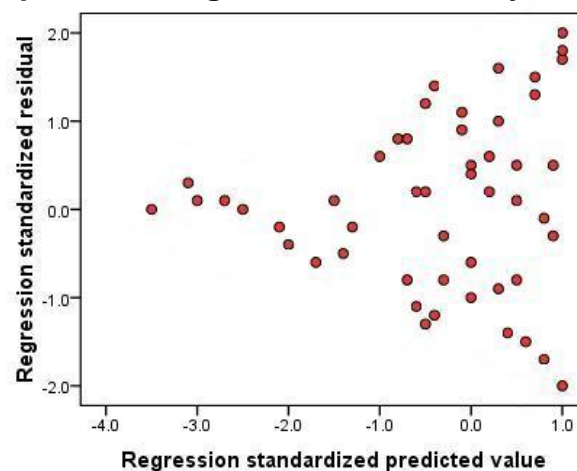
**2. Variance in all explanatory variables:** This one is fairly easy to check - just create a histogram for each variable to ensure that there is a range of values or that data is spread between multiple categories. This assumption is rarely violated if you have created good measures of the variables you are interested in.

**3. Multicollinearity:** The simplest way to ascertain whether or not your explanatory variables are highly correlated with each other is to examine a correlation matrix. If correlations are above .80 then you may have a problem. A more precise approach

is to use the collinearity statistics that SPSS can provide. The Variance inflation factor (VIF) and tolerance statistic can tell you whether or not a given explanatory variable has a strong relationship with the other explanatory variables. Again, we'll show you how to obtain these statistics when we run through the example!

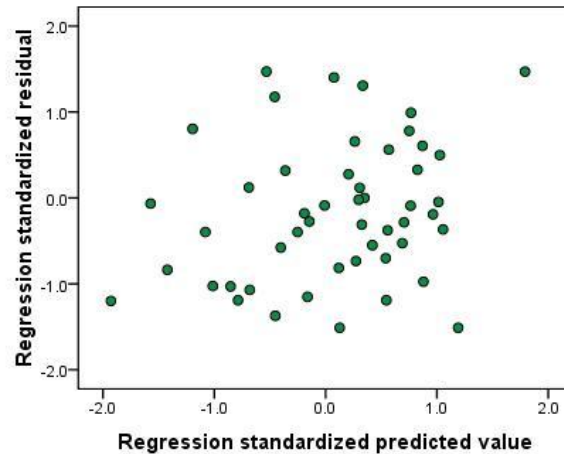
**4. Homoscedasticity:** We can check that residuals do not vary systematically with the predicted values by plotting the residuals against the values predicted by the regression model. Let's go into this in a little more depth than we did previously. We are looking for any evidence that residuals vary in a clear pattern. Let's look at the examples below (**Figure 3.3.3**).

**Figure 3.3.3: Scatterplot showing heteroscedasticity - assumption violated**



This scatterplot is an example of what a scatterplot might look like if the assumption of homoscedasticity is not met (this can be described as heteroscedasticity). The data points seem to funnel towards the negative end of the x-axis indicating that there is more variability in the residuals at higher predicted values than at lower predicted values. This is problematic as it suggests our model is more accurate when estimating lower values compared to higher values! In cases where the assumption of homoscedasticity is not met it may be possible to transform the outcome measure (see **Extension A**).

**Figure 3.3.4: Scatterplot showing homoscedasticity - assumption met**

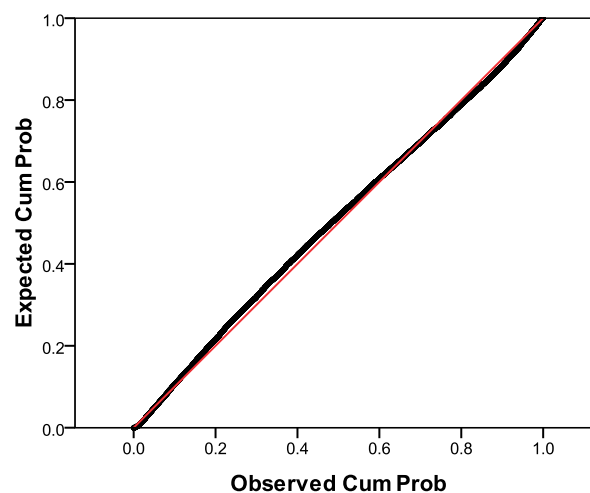


That's better! In **Figure 3.3.4** the data points seem fairly randomly distributed with a fairly even spread of residuals at all predicted values.

**5. Independent errors:** As we have stated before this assumption is rather tricky to test but luckily it only really applies to data where repeated measures have been taken at several time points. It should be noted that, as we said on **Page 2.6**, if there is a high degree of clustering then multi-level multiple regression may be appropriate. Using the SPSS complex samples module, mixed module or separate multi-levening modelling packages such as MLWin may be the only solution (see **page 2.6** for more detail).

**6. Normally distributed residuals:** A histogram of the residuals (errors) in our model can be used to check that they are normally distributed. However it is often hard to tell if the distribution is normal from just a histogram so additionally you should use a P-P plot as shown below (**Figure 3.3.5**):

**Figure 3.3.5: P-P plot of standardized regression residual**



As you can see the expected and observed cumulative probabilities, while not matching perfectly, are fairly similar. This suggests that the residuals are approximately normally distributed. In this example the assumption is not violated.

### **A Note on Sample Size**

The size of the data set that you're analyzing can be very important to the regression model that you can build and the conclusions that you can draw from it. In order for your regression model to be reliable and capable of detecting certain effects and relationships you will need an appropriate sample size. There is a general rule of thumb for this:

For each explanatory variable in the model 15 cases of data are required (see **Field, 2009, pp. 645-647**).

This is useful **BUT** it should be noted that it *is* an oversimplification! A good sample size depends on the strength of the effect or relationship that you're trying to detect - the smaller the effect you're looking for the larger the sample you will need to detect it! For example, you may need a relatively small sample to find a statistically significant relationship between age and reading ability. Reading ability usually develops with age and so you can expect a strong association will emerge even with a relatively small sample. However if you were looking for a relationship between reading ability and something more obscure, say time spent watching television, you would probably find a weaker correlation. To detect this weaker relationship and be confident that it exists in the actual population you will need a larger sample size.

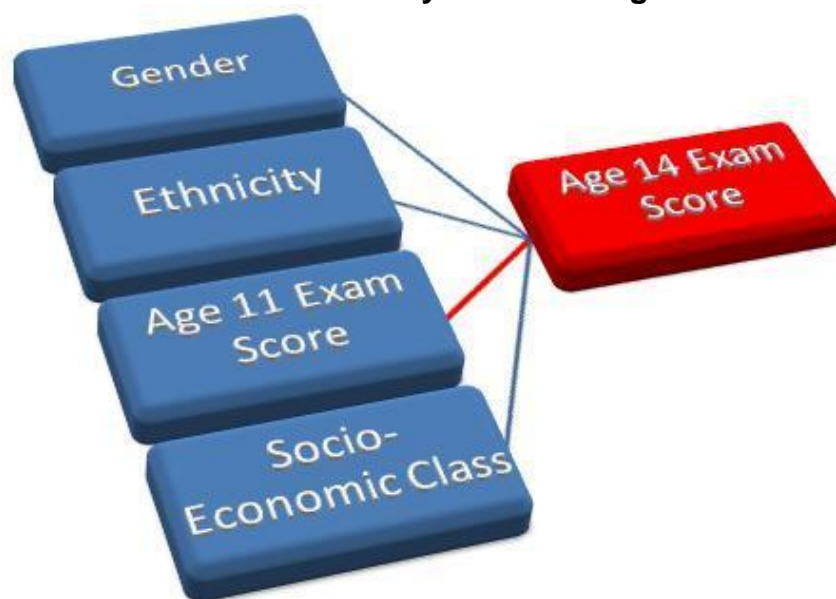
On this website we're mainly dealing with a very large dataset of over 15,000 individual participants. Though in general it can be argued that you want as big a sample as practically possible some caution is required when interpreting data from large samples. A dataset this large is very likely to produce results which are 'statistically significant'. This is because the sheer size of the sample overwhelms the random effects of sampling - the more of the population we have spoken to the more confident we can feel that we have adequately represented them. This is of course a good thing but a 'statistically significant' finding can have the effect of causing researchers to overemphasise their findings. A p-value does not tell the researcher how large an effect is and it may be that the effect is statistically significant but so small that it is not important. For this reason it is important to look at the effect size (the strength) of an effect or relationship as well as whether or not it is statistically likely to have occurred by chance in a sample. Of course it is also important to consider *who* is in your sample. Does it represent the population you want it to?

If you would like more information about sample size we recommend that you check out **Field (2009, p.645)**. There is also software that will allow you to estimate quite precisely the sample size you need to detect a difference of a given size with a given level of confidence. One example is the dramatically named 'GPower', which can be downloaded for free (the link is in our **resources**). With this in mind let us put our new knowledge on regression analysis into practice by running through an example!


### 3.4 Using SPSS to model the LSYPE data

In our example of simple linear regression in the previous module we found that prior academic achievement (age 11) is a good predictor of future academic achievement (age 14). This seems reasonable but surely there is more to it than that! The socio-economic class (SEC) of the parents is known to be related to students' academic achievement. The media talk about a gender gap in exam scores and inequalities between different ethnic groups. If we knew about these other variables could we improve the predictive power of our model? Multiple regression provides us with the tools we need to explore these questions!

**Figure 3.4.1: Factors which may influence Age 14 exam score**



The rest of this module is largely dedicated to an example which will build these variables into our multiple regression model and improve our understanding of the relationship between these factors and educational achievement. Over the next few pages we will be building up a model for predicting achievement during age 14 (KS3) exams. There will be seven different versions of the model as we build up your knowledge and refine the structure of our regression model with each new variable.

We will show you how to run this process on SPSS. Why not follow us through using the LSYPE MLR 15,000  dataset? The variables are already there for you so you will be able to run the analyses without creating the new variables (you'll see what we mean by this when we get started). Come on, it will be fun. Like a convoy.

Though it is not advisable to use anything other than a continuous variable for an outcome variable in multiple linear regression it is possible to use ordinal and nominal variables as explanatory variables. Before we can do this the data needs to

be set up in a specific way on SPSS. The ordinal and nominal variables must be coded as numbers so that SPSS can understand them.

### **Ordinal variables**

The process for ordinal variables is straight forward as it simply means ranking the categories. For example, for socio-economic class (SEC) data provided by parents on their occupation and employer have been coded using the Office for National Statistics socio-economic classification (for further details on this coding system see our **Resources**). Our measure of SEC has eight categories arranged in rank order with 'Higher managerial and professional occupations' coded as 1 through to 'Never worked or long-term unemployed' coded as 8.

While SEC is usually treated as an ordinal rather than a continuous variable, for the purpose of this example we will initially treat SEC as if it were a scale. However, ordinal variables should only be used in this way when there are at least five categories (levels) within the variable, a reasonable spread of cases across the levels, and a roughly linear relationship with the outcome. All these conditions are met for SEC (see **Figure 3.4.2** and later **Figure 3.5.2**). Regression is a very 'robust' procedure when it comes to treating ordinal variables as continuous explanatory variables so given these conditions are met this is permissible, although we will discuss some limitation to treating ordinal variables in this way later in the analysis.

**Figure 3.4.2: Frequency table for SEC of the home**

		Frequency	Percent	Valid Percent
Valid	1 Higher Managerial and professional occupations	1567	9.9	12.2
	2 Lower managerial and professional occupations	3083	19.5	24.0
	3 Intermediate occupations	932	5.9	7.3
	4 Small employers and own account workers	1672	10.6	13.0
	5 Lower supervisory and technical occupations	1454	9.2	11.3
	6 Semi-routine occupations	1637	10.4	12.8
	7 Routine occupations	1409	8.9	11.0
	8 Never worked/long term unemployed	1075	6.8	8.4
	Total	12829	81.4	100.0
Missing	0 missing	2941	18.6	
Total		15770	100.0	

### **Nominal variables - simple dichotomies**

Some nominal variables are simple dichotomies which mean they have only two mutually exclusive categories (e.g. you are either eligible for a free school meal or you are not – you can only belong to one of two categories). These are called dichotomous or binary variables because they have only two categories. Adding such variables to your regression module is fairly simple because we can simply give each category a numeric code (e.g. for gender code males as '0' and females as '1').



The output will represent a direct comparison between the two categories – this will become clearer when we run our example!

### **Nominal variables - multiple categories**

When you have nominal variables with multiple categories that cannot be ranked it requires a slightly more complicated approach. How do you numerically code a variable like school type? We can assign numbers to different school types, e.g. 0 for 'community schools', 1 for 'independent schools', 2 for 'foundation schools' 3, for 'voluntary-aided schools' and so on. However these numbers do not represent more or less of something as they do with SEC. In this case it is necessary to set up a number of comparisons such that a *reference category* (say 'community schools') is compared to each of the other categories. This means you have to create a series of new binary variables (for example 'independent school', 'foundation school' and 'Voluntary-aided school') where each case is coded '1' if it is from that particular school type and coded '0' otherwise. This procedure is often called setting up dummy variables. There should be one less dummy variable than the number of categories in the variable. So if we had *four* types of school we would need to choose one to be our base or reference category (e.g. community schools) and then create *three* dummy variables to compare each of the remaining categories (independent, foundation and voluntary-aided) to this reference category.

This is easiest to understand with examples. Over the next few pages we will introduce a variable of each of these types into our model.

### 3.5 A model with a continuous explanatory variable (Model 1)

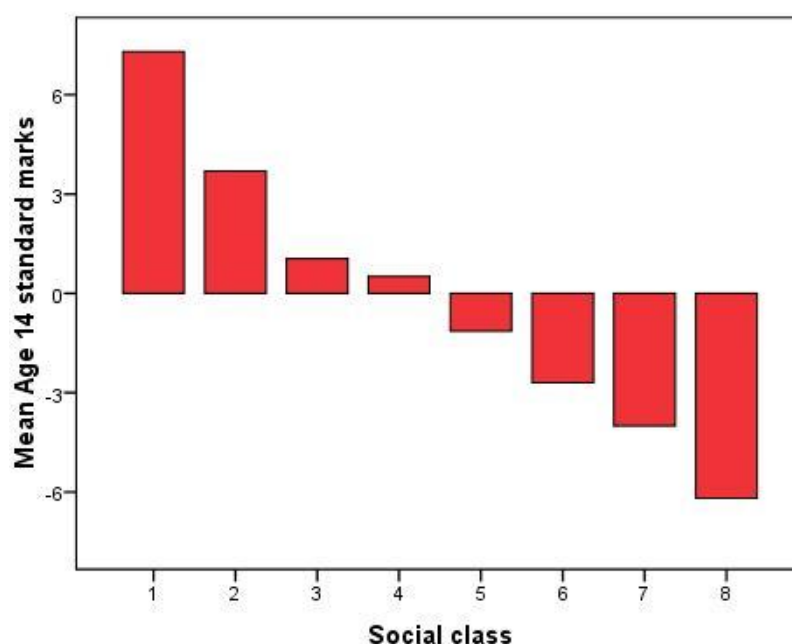
Rather than start by throwing all possible explanatory variables into the regression model let's build it up in stages. This way we can get a feel for how adding different variables affects our model.

Our Outcome variable is *ks3stand* - the standard score in national tests taken at the end of Key Stage 3 (KS3) at age 14, which we used as the outcome measure in the previous module. To begin our analysis we will start with Social Economic Class (SEC) as an explanatory variable. SEC represents the socio-economic class of the home on a scale of '1' (Higher Managerial and professional occupations) to '8' (Never worked/long term unemployed). There is a strong relationship between SEC and mean age 14 score, as shown in **Figures 3.5.1 and 3.5.2** below.

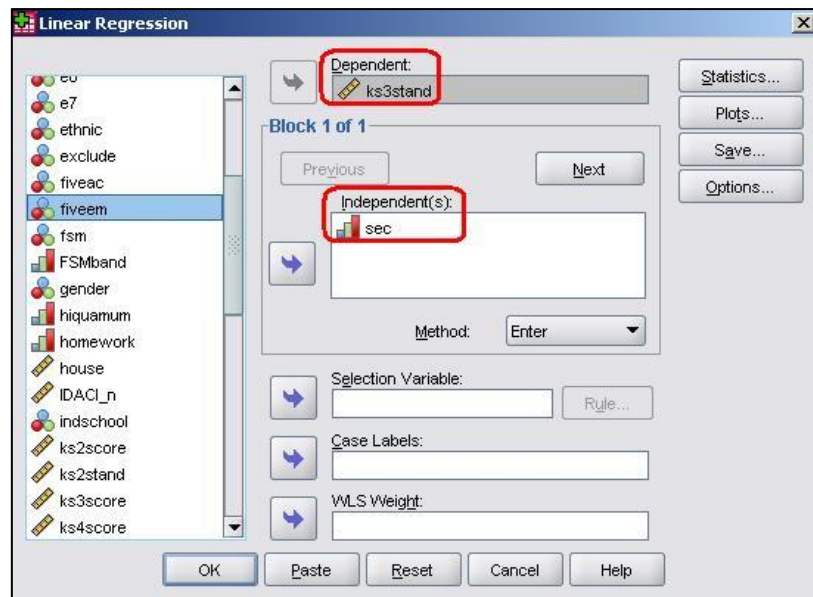
**Table 3.5.1: Mean age 14 score by SEC**

Social class (SEC)	Mean	N	SD
1 Higher Managerial & professional occupations	7.30	1378	9.761
2 Lower managerial & professional occupations	3.70	2851	9.179
3 Intermediate occupations	1.05	899	8.813
4 Small employers and own account workers	.51	1585	9.287
5 Lower supervisory and technical occupations	-1.14	1421	8.803
6 Semi-routine occupations	-2.70	1580	8.823
7 Routine occupations	-4.00	1367	9.063
8 Never worked/long term unemployed	-6.19	1019	9.162
Total	.39	12100	9.936

**Figure 3.5.2: Mean age 14 score by SEC**



We will start by entering SEC in our regression equation. Take the following route through SPSS: **Analyse > Regression > Linear**. This is the exact same route which we took for simple linear regression so you may well recognize the pop-up window that appears. The variable *ks3stand* goes in the *dependent* box and the variable *sec* is placed in the *independents* box. Note that we have selected 'Enter' as our *Method*.



We are not going to run through all of the diagnostic tests that we usually would this time – we will save that for when we add more variables over the coming pages! Let's just click **OK** as it is and see what SPSS gives us.

### **SPSS output for multiple linear regression**

In this basic analysis SPSS has only provided us with four tables. The first simply tells us which variables we have included in the model so we haven't reproduced that here. The other three provide more useful information about our model and the contribution of each of our explanatory variables. The process of interpreting most of these statistics is the same for multiple linear regression as we saw for simple linear regression in **Module 2**.

**Figure 3.5.3: Multiple  $r$  and  $r^2$  for model**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.389 <sup>a</sup>	.151	.151	9.154

The *Model Summary* (**Figure 3.5.3**) offers the multiple  $r$  and coefficient of determination ( $r^2$ ) for the regression model. As you can see  $r^2 = .151$  which indicates that 15.1% of the variance in age 14 standard score can be explained by our regression model. In other words the success of a student at age 14 is strongly related to the social economic class of the home in which they reside (as we saw in

**Figure 3.5.2).** However there is still a lot of variation in outcomes between students that is not related to SEC.

**Figure 3.5.4: ANOVA for model fit**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	180687.920	1	180687.920	2156.304	.000 <sup>a</sup>
	Residual	1013754.450	12098	83.795		
	Total	1194442.370	12099			

Whether or not our regression model explains a statistically significant proportion of the variance is ascertained from the *ANOVA* table (**Figure 3.5.4**), specifically the F-value (penultimate column) and the associated significance value. As before, our model predicts the outcome more accurately than if we were just using the mean to model the data ( $p < .000$ , or less than .0005, remembering that SPSS only rounds to 3 significant figures).

**Figure 3.5.5: Coefficients for model**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.579	.176		43.120	.000
	sec Socio-economic class of the home	-1.725	.037	-.389	-46.436	.000

The *Coefficients* table (**Figure 3.5.5**) gives the Constant or intercept term and the regression coefficients (**b**) for each explanatory variable. The constant value (7.579) represents the intercept, which is the predicted age 14 score when SEC=0 (note that SEC is never actually 0 in our data where the values of SEC range from 1-8, the constant is just important for the construction of the model). The other value here is the regression coefficients (**b**) for SEC. This indicates that for every unit increase in SEC the model predicts a *decrease* of -1.725 in age 14 standard score. This may sound counter-intuitive but it actually isn't – remember that SEC is coded such that lower values represent higher social class groups (e.g. 1 = 'Higher Managerial and professional', 8 = 'Never worked/long term unemployed').

We can use the regression parameters to calculate the predicted values from our model, so the predicted age 14 score when SEC=1 (higher managerial) is  $7.579 + (1 \times -1.725) = 5.85$ . By comparison the predicted age 14 score when SEC=8 (long term unemployed) is  $7.579 + (8 \times -1.725) = -6.22$ . There is therefore roughly a 12 score point gap between the highest and lowest SEC categories, which is a substantial difference. Finally the t-tests and 'sig.' values in the last two rows tell us that the variable is making a statistically significant contribution to the predictive power of the model – it appears that SEC is since the t-statistic is statistically significant ( $p < .000$ ).

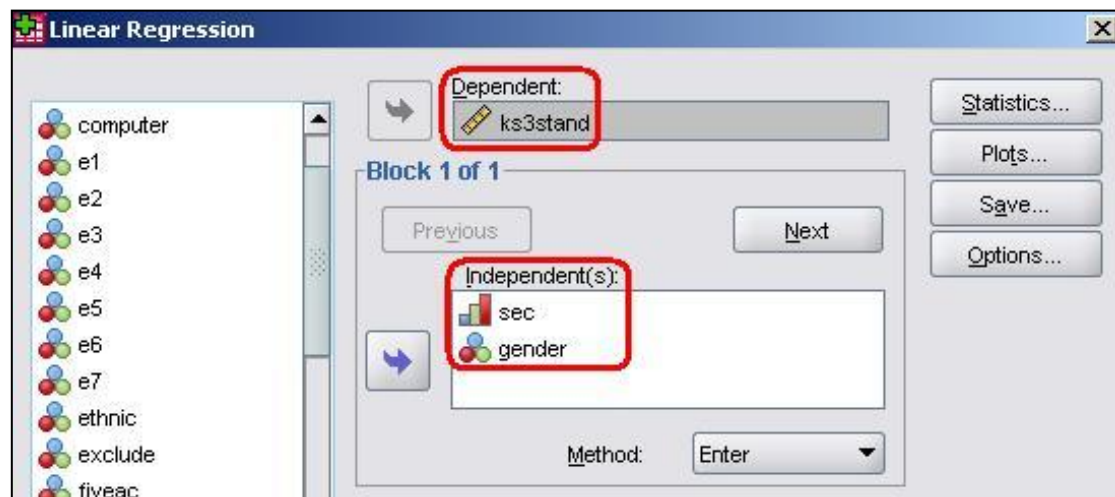
### 3.6 Adding dichotomous nominal explanatory variables (Model 2)

As discussed on the previous page, SEC can reasonably be treated as a scale, but what do we do with variables which are nominal? What if the categories cannot be placed in a rank order? Let us take the example of gender. Gender is usually a dichotomous variable – participants are either male or female. **Figure 3.6.1** displays the mean age 14 standard scores for males and females in the sample. There is a difference of a whole score point between the scores of males and females, which suggests a case for adding *gender* to our regression model.

**Figure 3.6.1: Mean age 14 score by gender**

	Mean KS3 score	N	Std. Deviation
0 Male	-.45	7378	10.174
1 Female	.62	7140	9.710
Total	.08	14518	9.963

Take the following route through SPSS: **Analyse > Regression > Linear**. Add *gender* to the *independents* box – we are now repeating the multiple regression we performed on the previous page but adding gender as an explanatory variable. Do not worry about all of the extra options and assumptions yet – we will come to that later! Just click **OK**.



You will be starting to get familiar with these three tables now.

**Figure 3.6.2: Multiple  $r$  and  $r^2$  for model**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.394 <sup>a</sup>	.155	.155	9.113

Our *Model Summary* (**Figure 3.6.2**) tells us that our new model now has  $r^2 = .155$  which suggests that 15.5% of the total variance in age 14 score can be explained. This is only very slightly more than the previous model (15.1%). But has the inclusion of gender made a significant contribution to explaining age 14 test score? We evaluate this through inspecting the coefficients table.

**Figure 3.6.3: Coefficients for model**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.032	.194		36.202	.000
	Socio-economic class of the home	-1.722	.037	-.389	-46.172	.000
	Gender	1.198	.167	.060	7.169	.000

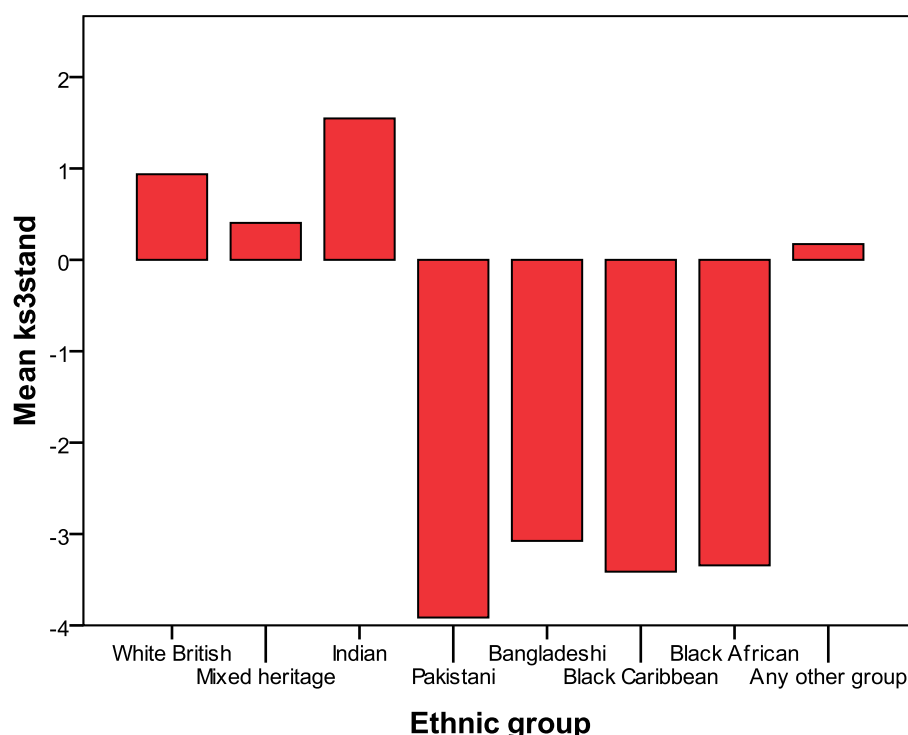
The *Coefficient* table (**Figure 3.6.3**) provides us with a fresh challenge: how do we interpret the b-coefficient for gender? Actually it is delightfully simple! If you recall the code '0' was used for males and '1' for females. The b-coefficient tells us how much higher or lower the category coded 1 (females) score in direct comparison to the category coded 0 (males) when the other variables in the model (currently SEC) are controlled (held fixed). The B coefficient for gender indicates that females score on average 1.2 standard marks higher than males, whatever the SEC of the home. The t-tests indicate that both explanatory variables are making a statistically significant contribution to the predictive power of the model.

What are the relative strengths of SEC and gender in predicting age 14 score? We cannot tell this directly from the coefficients since these are not expressed on a common scale. A one unit increase in SEC does not mean the same thing as a one unit increase in gender. We can get a rough estimate of their relative size by evaluating the difference across the full range for each explanatory variable, so the range for SEC is  $7 \times -1.72$  or 12.0 points, whereas the range for gender is just 1.2 points (girls versus boys). Another way of judging the relative importance of explanatory variables is through the *Beta* ( $\beta$ ) weights in the fourth column. These are a standardised form of b which range between 0 and 1 and give a common metric which can be compared across all explanatory variables. The effects of SEC is large relative to gender, as can be seen by the relative difference in beta values (-.389 versus .060). Note you ignore the sign since this only indicates the direction, whether the explanatory variable is associated with an increase or decrease in outcome scores, it is the absolute value of Beta which is used to gauge its importance. You will remember this from comparing the strength of correlation coefficients which we completed in the Simple Linear Regression module (**see page 2.4**). The results indicate that SEC is a more powerful predictor of age 14 score than gender, but both make a contribution to the explanation of variance in age 14 score.

### 3.7 Adding nominal variables with more than two categories (Model 3)

In the section above we added gender to our model and discussed how we interpret the  $b$  coefficient as a direct comparison of two categories. But what do we do when we have more than two categories to compare? Let's take a look at the example of ethnicity. **Figure 3.7.1** plots the mean age 14 score for each ethnic group. This shows that Pakistani, Bangladeshi, Black African and Black Caribbean students on average have a mean score around 3 points lower than White British students. Ethnic group is definitely a candidate to include as an explanatory variable in our regression model.

**Figure 3.7.1: Bar chart of mean age 14 score by ethnic group**



On **page 3.4** we mentioned the use of 'dummy variables' as a method for dealing with this issue. Where we have nominal variables with more than two categories we have to choose a reference (or comparison) category and then set up dummy variables which will contrast each remaining category against the reference category. See below for an explanation of how the ethnic group variable is coded into seven new dichotomous 'dummy' variables.

### ***Nominal variable with more than two categories: Ethnicity***

This requires the use of dummy variables. The variable is originally coded '0' to '7' with each code representing a different ethnic group. However we cannot treat this as an ordinal variable - we cannot say that 'Black Caribbean (coded '5') is 'more' of something than White British (coded '0'). What we need to do is compare each ethnic group against a reference category. The most sensible reference category is 'White British' (because it contains the largest number of participants), so we want to contrast each ethnic group against 'White British'. We do this by creating seven separate variables, one for each minority ethnic group (dummy variables). The group we leave out (White British) will be the reference group ('0' code).

This has already been done in the dataset so why not take a look. The new variable 'e1' takes the value of '1' if the participant is of 'Mixed Heritage' and '0' otherwise, while 'e2' takes the value of '1' if the pupil is Indian and '0' otherwise, and so on.

We have already coded the dummy variables for you but it is important to know how it is done on SPSS. You can use the **Transform > Recode into new variable** route to create each new variable individually. We have discussed the use of **Transform** for creating new variables briefly in our Foundation Module. There are hundreds of options through this menu and we think the best way to learn is to have a play with it! The alternative to generating each dummy variable individually is using syntax. We have included the syntax for the recoding of the ethnicity variable below! You don't need to use the syntax if you don't want to as we have already created the variables in our datasets but it is useful to know how to generate them.

### **SYNTAX ALERT!!!**

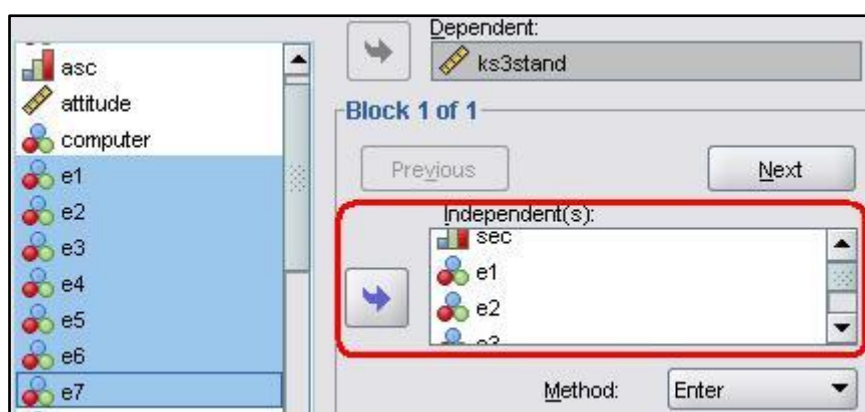
```
RECODE Ethnic (1=1)(else=0) into e1.  
RECODE Ethnic (2=1)(else=0) into e2.  
RECODE Ethnic (3=1)(else=0) into e3.  
RECODE Ethnic (4=1)(else=0) into e4.  
RECODE Ethnic (5=1)(else=0) into e5.  
RECODE Ethnic (6=1)(else=0) into e6.  
RECODE Ethnic (7=1)(else=0) into e7.  
VAR LABELS  
  e1 "Mixed heritage" e2 "Indian" e3 "Pakistani" e4 "Bangladeshi" e5 "Black  
  Caribbean" e6 "Black African" e7 "Any other ethnic group".
```

Coding variables through the SPSS menu options is relatively easy once you are used to the software, but can be very time-consuming. Using Syntax is a good way of saving time!

We can now include our dummy variables for ethnic group (e1 through to e7). Take the same familiar path through SPSS: **Analyse > Regression > Linear**. Add



*ks3stand* as the *Dependent* and move all of the relevant variables into the *Independents* box: *sec*, *gender*, *e1*, *e2*, *e3*, *e4*, *e5*, *e6* and *e7*.



Click **OK** when you're ready.

You will see that the new model has improved the amount of variance explained with  $r^2 = .170$ , or 17.0% of the variance (**Figure 3.7.2**), up from 15.5% in the previous model. We won't print the ANOVA table again but it does show that the new model once again explains more variance than the baseline (mean) model to a statistically significant level.

**Figure 3.7.2: Model 3 summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.412 <sup>a</sup>	.170	.169	9.037

More interesting for our understanding and interpretation is the *coefficients* table (**Figure 3.7.3**).

**Figure 3.7.3: Regression coefficients for model 3**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.180	.197		36.412	.000
	sec Socio-economic class	-1.674	.038	-.378	-43.595	.000
	gender Gender	1.234	.166	.062	7.439	.000
	e1 Mixed heritage	-.376	.375	-.008	-1.002	.316
	e2 Indian	1.400	.338	.035	4.142	.000
	e3 Pakistani	-2.343	.361	-.056	-6.494	.000
	e4 Bangladeshi	-.465	.432	-.009	-1.076	.282
	e5 Black Caribbean	-4.251	.436	-.082	-9.746	.000
	e6 Black African	-3.294	.437	-.064	-7.539	.000
	e7 Any other ethnic group	.199	.433	.004	.459	.646

Firstly a quick glance at the b coefficients (**Figure 3.7.3**) shows SEC and gender are still significant predictors, with a decrease of -1.674 score points for every unit increase in SEC and with girls scoring 1.234 points higher than boys. The b-coefficients for the ethnic dummy variables can be interpreted in a similar way to the

interpretation of gender. The coefficients represent the difference in age 14 test score between being in the specified category and being in the reference category (White British) when the other variables are all controlled. For example, the model indicates that Indian students achieve 1.40 *more* standard score points than White British students, while Black Caribbean student achieve -4.25 *less* standard score points than White British students. Remember these coefficient are after controlling for SEC and gender.

Though it is clear that SEC score is the most important explanatory variable, looking down the *t* and *sig* columns tells us that actually most of the ethnic dummy variables make a statistically significant contribution to predicting age 14 score ( $p < .05$ ). After we have controlled for SEC and gender, there is no statistically significant evidence that students of Mixed Heritage, Bangladeshi and Any Other ethnic group achieve different results to White British students. However on average Indian students score significantly higher than White British students while Pakistani, Black Caribbean and Black African pupils score significantly lower ( $p < .000$ ).

### 3.8 Predicting scores using the Regression Model

Let's look at how we can use model 3 to predict the score at age 14 that a given student from a specific background would be expected to achieve. Take this opportunity to look back at the coefficient table for model 3 (**Figure 3.7.3, Page 3.7**). The intercept is 7.18, this is the predicted age 14 exam score for our reference group, which is where SEC=0, gender=0 (boy) and ethnicity=0 (White British). The coefficient for SEC shows that each unit increase in SEC is associated with a decrease of about -1.67 score points in test score. The coefficients for gender shows the average difference between girls and boys, and the coefficients for each ethnic group shows the average difference between the relevant ethnic group and White British students. There is no interaction term (more on this on **Page 3.11**) so the model assumes the effect of gender and ethnicity are the same at all levels of SEC. For example, whatever the SEC of the home or whatever ethnic group, girls on average score 1.234 points higher than boys. Equally whatever the SEC of the home or gender of the student, Black Caribbean students score 4.25 points below White British students of the same gender.

So let's see how the predicted values are calculated. This may initially seem quite complicated but what we are doing is actually very straightforward. There are a total of 10 terms in our regression equation for Model 3. There is the intercept, which is constant for all cases, and there are nine regression coefficients: a coefficient for SEC, a coefficient for gender and seven coefficients for ethnicity, one for each ethnic group. As we described on **Page 3.2** in standard notation the calculation of the predicted age 14 score (labelled as  $\hat{Y}$ ) for any case would be written as:

$$\hat{Y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_9x_9$$

Where  $\hat{Y}$  = the predicted age 14 score;  $a$ = the intercept;  $b_1$ = the regression coefficient for variable 1;  $x_1$ = the value of variable 1,  $b_2$ = the regression coefficient for variable 2;  $x_2$ = the value of variable 2.... and so on through to  $b_9$  and  $x_9$  for variable 9. We can calculate the predicted value for any case simply by typing in the relevant quantities ( $a$ ,  $b_1$ ,  $x_1$ ,  $b_2$ ,  $x_2$ ... etc) from the regression equation. Four examples are shown below.

*For a White British, boy, from SEC=1 (higher managerial & professional home)*

The predicted value would be:

$$\hat{Y} = \text{intercept} + (1 \times \text{SEC coefficient})$$

$$\hat{Y} = 7.18 + (1 \times -1.674) = 5.51.$$

Because gender=0 (male) and ethnic group=0 (White British) there is no contribution from these terms.

*For a White British, girl, from SEC=1*

The predicted value would be:

$$\hat{Y} = \text{intercept} + (1 * \text{SEC coefficient}) + (\text{Gender coefficient})$$

$$\hat{Y} = 7.18 + (1 * -1.674) + (1.234) = 6.74.$$

Again, because ethnic group=0 there is no contribution from the ethnic terms.

*For a Black Caribbean, boy, from SEC=1*

The predicted value would be:

$$\hat{Y} = \text{intercept} + (1 * \text{SEC coefficient}) + (\text{Black Caribbean coefficient}) =$$

$$\hat{Y} = 7.180 + (1 * -1.674) + (-4.251) = 1.26.$$

Because gender=0 there is no contribution from this term.

*For a Black Caribbean, girl, from SEC=1*

The predicted value would be:

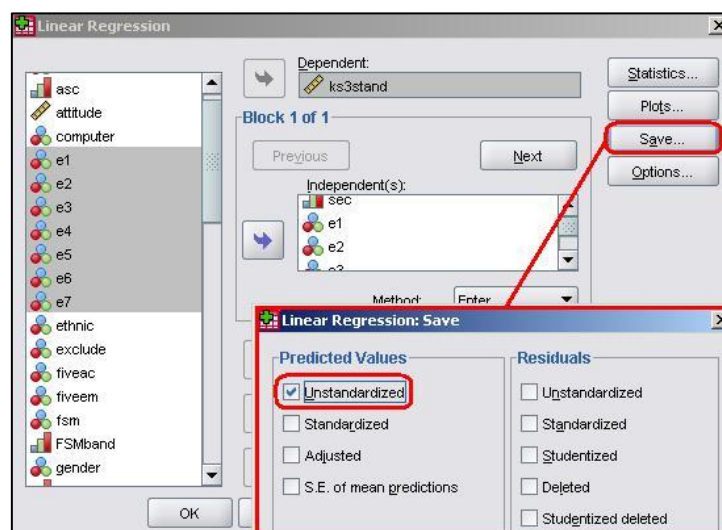
$$\hat{Y} = \text{intercept} + (1 * \text{SEC coefficient}) + (\text{Gender coefficient}) + (\text{Black Caribbean coefficient})$$

$$\hat{Y} = 7.180 + (1 * -1.674) + (1 * 1.234) + (-4.251) = 2.49.$$

Once you get your head around the numbers what we are doing is actually very straightforward.

The key point to notice is that whatever the value of SEC, girls are always predicted to score 1.234 points higher than boys. Equally whatever the SEC of the home, Black Caribbean students are always predicted to score 4.25 point below White British students of the same gender.

Rather than manually calculating the predicted values for all possible combinations of values, when specifying the multiple regression model we can ask SPSS to calculate and save the predicted values for every case. These predicted values are already saved in the LSYPE 15,000 MLR dataset (they are called PRE\_1). If you want to do this yourself can rerun the analysis for Model 3 as described on **Page 3.7** using the LSYPE 15,000 dataset but this time also click on the save button:



Add a tick in the '*Predicted values (unstandardized)*' option in the pop-up box. This will create a new variable in your SPSS file called PRE\_1 which will hold the predicted age 14 score for each student, as calculated by the model.

We can then plot these values to give us a visual display of the predicted variables. Let us look at the relationship between ethnic group and SEC. We will just plot the predicted values for boys since, as we have seen, the pattern of predicted values for girls will be identical except that every predicted value will be 1.234 points higher than the corresponding value for boys. We can plot the graph using the menu options as shown in **Module 1**, or we can do this more simply using syntax:

**SYNTAX ALERT!**

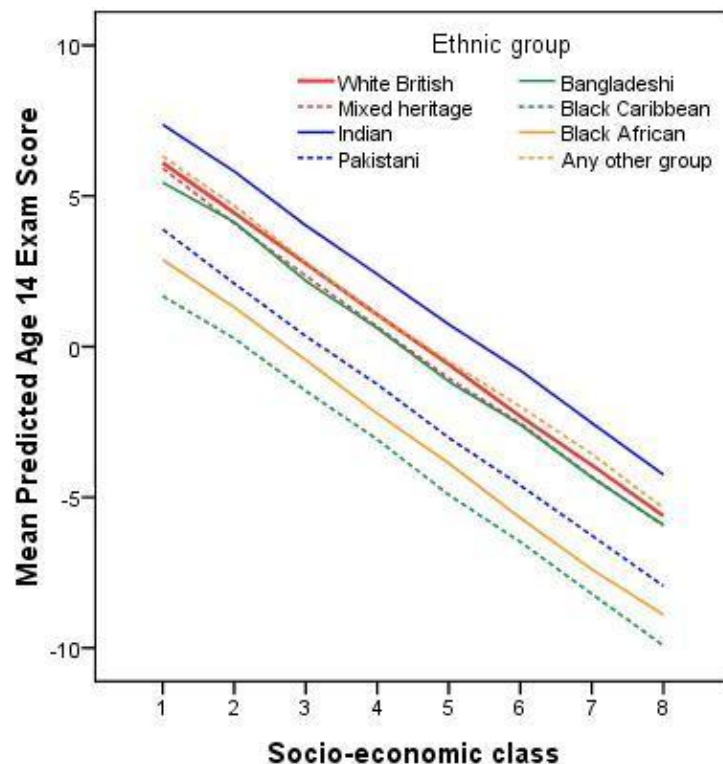
TEMPORARY.

SELECT IF gender=0.

GRAPH /LINE(MULTIPLE) MEAN (pre\_1) BY SEC by Ethnic.

The results are shown in **Figure 3.8.1**.

**Figure 3.8.1: Regression lines for ethnic group, SEC and age 14 attainment**



The important point you should notice is that the fitted regression lines for each ethnic group have different intercepts but the same slope, i.e. the regression lines are parallel. There are two equivalent ways of expressing the figure. We can say that the effect of SEC on attainment is the same for all ethnic groups, or we can say

the effect of ethnicity on attainment is the same for all social classes. It's the same thing (like two sides of a coin).

We will return to this type of line graph when we start to explore interaction effects on **Page 3.11** but for now let's discuss ways of refining our existing model further.


### 3.9 Refining the model – treating ordinal variables as a set of dummy variables (Model 4)

#### Minimising the effect of missing data

So far we have treated SEC as a continuous variable or scale. What are the implications of having done this? Treating a variable as a scale means that any case with a missing value on the variable is lost from the analysis. The frequency table for SEC was shown in **Figure 3.4.2**. A total of 2941 cases (18.6%) were missing a value for SEC, which is a high level of data loss.

One way of coping with this is to recode SEC into dummy variables, as we did with ethnic group (**Page 3.7**), and to explicitly include the ‘missing’ values as an extra category. This has several advantages, it:

- Prevents the loss of data that would come from omitting all cases with missing values, as happens when SEC is treated as a scale variable (excluding missing data in this way is known as ‘Listwise’ deletion)
- Allows for the direct modelling of missing data rather than imputing missing values, for example by mean substitution, which has its own interpretative problems
- Allows for non-linearity in the relationship between the ordinal categories and student attainment
- Can simplify the interpretation of the relationship between the explanatory variable and the outcome, since we can directly contrast against a reference category (e.g. compare all SEC categories against long term unemployed)
- Can ensure a consistent base in terms of the sample size across a range of hierarchical regression models by retaining all cases (including those with missing values) as new explanatory variables are added.

We used the following syntax to change SEC into a series of dummy variables. You don’t need to do this though as we have already prepared the dummy variables for you in the LSYPE 15000 MLR  file. Aren’t we nice?

### SYNTAX ALERT!!!

More syntax for your enjoyment! Are you beginning to see how it works?

```
RECODE sec (1=1)(else=0) into sc1.  
RECODE sec (2=1)(else=0) into sc2.  
RECODE sec (3=1)(else=0) into sc3.  
RECODE sec (4=1)(else=0) into sc4.  
RECODE sec (5=1)(else=0) into sc5.  
RECODE sec (6=1)(else=0) into sc6.  
RECODE sec (7=1)(else=0) into sc7.  
RECODE sec (missing=1)(else=0) into sc0.
```

### VARIABLE LABELS

```
sc1 ' Higher managerial & professional'  
sc2 ' Lower managerial & professional'  
sc3 ' Intermediate occupations'  
sc4 ' Small employers & own account workers'  
sc5 'Lower supervisory & technical occupations'  
sc6 ' Semi-routine occupations'  
sc7 ' Routine occupations'  
sc0 ' SEC missing'.
```

Note that the variable we have not created a dummy for (long-term unemployed) will be our reference category in the analysis. Let us repeat our last model but replace SEC with the eight terms sc0 to sc7. Repeat the regression we did on **Page 3.7** using *ks3stand* as the *Dependent* variable and *gender*, *e1 – e7* and *sc0 – sc7* as the independent variables. Rather than assessing the effect of SEC as a regression coefficient we get a direct measure of how each category (including our ‘missing cases’ category) contrasts with the base category (long term unemployed).

**Figure 3.9.1** presents the model summary and the ANOVA table. From the Model Summary we see that the model  $r^2$  is 15.1%. This is lower than for model 3 where the model accounted for 17.0% of the variance. However this reflects two factors: the change from treating SEC as a scale variable to modelling it as a set of dummy variables and the increase in sample size associated with including the previously omitted 2,900 or so cases. We can see from the ANOVA table that we are including 14,518 cases in this analysis (the total df shows the number of cases - 1), rather than the 12,100 cases included in model 3. We will not pursue the relative contribution of these two factors here, since the increase in sample size is reason enough for preferring the treatment of SEC as a set of dummy variables.



**Figure 3.9.1: model summary and the ANOVA table**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.388 <sup>a</sup>	.151	.150	9.186

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	217347.017	16	13584.189	160.998	.000 <sup>a</sup>
	Residual	1223522.937	14501	84.375		
	Total	1440869.954	14517			

Figure 3.9.2 shows the regression coefficients from the model.

**Figure 3.9.2: Regression coefficients for Model 4**

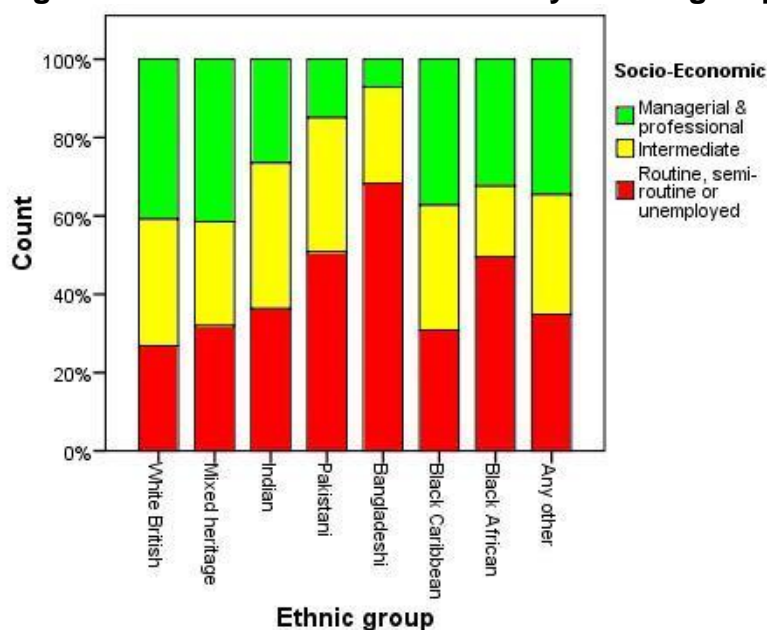
Model		Unstandardized Coefficients		Standard Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.507	.331		-16.636	.000
	sc1 Higher managerial & professional	12.467	.399	.365	31.255	.000
	sc2 Lower managerial & professional	9.036	.356	.359	25.402	.000
	sc3 Intermediate	6.459	.438	.155	14.735	.000
	sc4 small employers & own account	5.949	.384	.185	15.508	.000
	sc5 Lower supervisory & technical	4.130	.395	.122	10.454	.000
	sc6 semi-routine	2.751	.384	.085	7.157	.000
	sc7 routine	1.328	.396	.039	3.356	.001
	sc0 missing	3.813	.350	.147	10.901	.000
	gender Gender	1.109	.153	.056	7.262	.000
	e1 Mixed heritage	-.378	.348	-.008	-1.085	.278
	e2 Indian	1.518	.308	.038	4.927	.000
	e3 Pakistani	-2.864	.326	-.070	-8.789	.000
	e4 Bangladeshi	-1.056	.373	-.023	-2.835	.005
	e5 Black Caribbean	-3.939	.402	-.076	-9.808	.000
	e6 Black African	-2.956	.400	-.058	-7.398	.000
	e7 Any other ethnic group	.010	.390	.000	.025	.980

As we saw before, there is clearly a strong relationship between SEC and age 14 attainment, even after accounting for gender and ethnicity. Breaking the SEC variable down into its individual categories and comparing them to the base category of 'long term unemployed' makes interpretation of the coefficients more intuitive. For example, students from 'Higher managerial and professional' homes are predicted to obtain 12.5 more standard score marks than those from homes where the main parent is long term unemployed. Students from 'lower managerial and professional' homes achieve 9.0 more marks, those from intermediate homes 6.5 more marks and so on. You can see the ordinality in the data from the decreasing B coefficients: as the SEC of the home decreases there is a reduction in the extent of the 'boost' to age 14 standard score above the reference group of students from homes where the head of the household is long term unemployed. Being able to interpret the difference between categories in this way is very useful! We can see from the t statistic and associated 'sig' values that all SEC contrasts are highly statistically significant, including for those students with missing values for SEC.

### 3.10 Comparing coefficients across models

A significant objective in multiple regression modelling, as we observed in the introduction, is to assess the association of a variable with an outcome after controlling for the influence of other variables. **Figure 3.10.1** shows the relationship between ethnicity and SEC and it is apparent that students from minority ethnic backgrounds are less likely to be from the more affluent socio-economic classes than those from a White British background. Regression can be used to ascertain whether the ethnic gaps in attainment at age 14 result from these observed differences in SEC between ethnic groups. We can undertake this analysis by comparing the coefficients for our variable of interest (ethnic group) both before and after including the other 'control' variables in the multiple regression model (SEC and gender).

**Figure 3.10.1: % SEC breakdown by Ethnic group**



**Figure 3.10.3** shows the relationship between ethnic group and age 14 standard score both before and after controlling for the influence of gender and the SEC of the home. In both cases the reference category is White British students. The blue bars represent the unadjusted difference in mean age 14 scores between ethnic groups (the values for coefficients  $e_1$  to  $e_7$  when these are the only explanatory variables included in the regression model). The red bars display the ethnic coefficients  $e_1$  to  $e_7$  from model 4, after gender and SEC have been controlled (that is the variance in age 14 score that is accounted for by gender and SEC has been removed using the regression analysis). Indian students on average scored higher than White British students, and this difference was even more pronounced when their greater level of deprivation was taken into account, with an increase in the gap from 0.8 to 1.4 points. The gap for Bangladeshi students reduced from -3.9 points to around -1.0 point after adjustment, a reduction of 73%. There were smaller but still significant reductions in the size of the gap for Pakistani students (from -4.7 points to -2.9

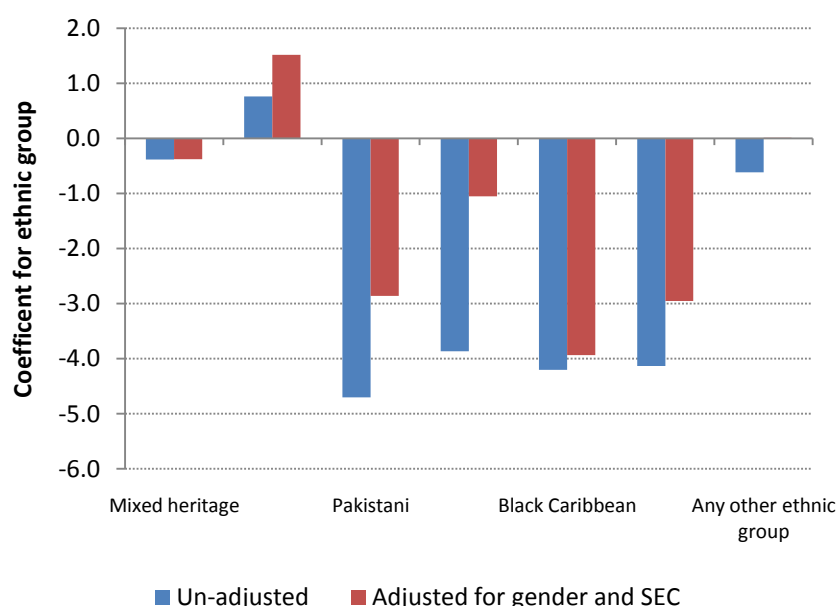
points) and for Black African students (from -4.1 to -3.0 points). However the average gap between Black Caribbean and White British students hardly changed at all, reducing only from -4.2 to -4.0 points.

**Table 3.10.2: Regression coefficients for ethnic groups before and after controlling for gender and SEC of the home**

	Unadjusted coefficients	Coefficients adjusted for gender and SEC
Intercept	.790	-5.507
Mixed heritage	-.39	-.38
Indian	.76	1.52
Pakistani	-4.70	-2.86
Bangladeshi	-3.87	-1.06
Black Caribbean	-4.20	-3.94
Black African	-4.13	-2.96
Any other ethnic group	-.62	.01
Sample size	14831	14517

*Note:* The sample size reduction in the adjusted model reflects the exclusion of 314 cases where gender was not known.

**Figure 3.10.3: Relationship between ethnic group and age 14 standard score before and after controlling for gender and SEC of the home**



There are many other variables in our LSYPE dataset than just SEC, gender and ethnicity, and it is likely that some of these may explain more of the ethnic gaps in attainment. However at this stage it is sufficient to show how coefficients can be compared across regression models to demonstrate the principle involved. One of the strengths of multiple regression is being able to ascertain the relative importance of an explanatory variable once others have been taken into account.

### 3.11 Exploring interactions between a dummy and a continuous variable (Model 5)

So far we have considered only the **main effects** of each of our explanatory variables (SEC, gender and ethnic group) on attainment at age 14. That is we have evaluated the association of each factor with educational attainment while holding all other factors constant. This involves a strong assumption that the effects of SEC, gender and ethnic group are **additive**, that is there are no interactions between the effect of these variables. For example we have assumed that the 'effect' of SEC is the same for all ethnic groups, or equivalently that the 'effect' of ethnicity is the same at all levels of SEC. However there is the possibility that SEC and ethnic group may **interact** in terms of their effect on attainment, that the relationship between ethnicity and attainment may be different at different levels of SEC (or put the other way around that the relationship between SEC and attainment may vary for different ethnic groups, it's the same thing). Is this assumption of additive effects valid, and how can we test it?

In any multiple regression model there exists the possibility of interaction effects. With only two explanatory variables there can of course be only one interaction, between explanatory variable 1 and explanatory variable 2, but the greater the number of variables in your model the higher the number of possible interactions. For example if we have 10 explanatory variables then there are 45 possible pairs of explanatory variables that may interact. It is unwieldy to test all possible combinations of explanatory variables, and indeed such 'blanket testing' may give rise to spurious effects, simply because at the 5% significance level some of the interactions might be significant by chance alone. Your search for possible interactions should be guided by knowledge of the existing literature and theory in your field of study. In relation to the literature of educational attainment, there is quite strong emerging literature suggesting interactions between ethnicity and SEC (**e.g. Strand, 1999; 2008**). Let's evaluate whether there is a statistically significant interaction between ethnicity and SEC in the current data, returning to the variables we used for model 3 (**Page 3.7**).

#### **Step 1: Creating the interaction terms**

Is it reasonable to assume that ethnic group differences in attainment are the same at all levels of SEC? We mentioned above that there is an emerging literature that suggests this may not be the case. One way to allow for different slopes in the relationship between SEC and attainment for different ethnic groups is to include extra variables in the model that represent the interactions between SEC and ethnic group.


For the purpose of this first example we treat SEC as a continuous variable, as we did in Models 1-3 (**Pages 3.4 to 3.8**). We want to create additional explanatory

variables that will represent the effect of SEC within each ethnic group. We do this simply by multiplying each of our ethnic group dummy variables by SEC. The table below (**Figure 3.11.1**) shows the computed values for some selected cases. *e1sec* will be a separate variable containing the SEC values only for cases where ethnicity =1 (Mixed heritage students); *e2sec* will be a separate variable contain the SEC values only for cases where ethnicity =2 (Indian students) and so on. Remember there has to be an omitted category against which these dummy variables are contrasted and this is White British students.

**Figure 3.11.1: Table showing examples of new interaction variables**

Ethnic group	SEC	<i>e1sec</i>	<i>e2sec</i>	<i>e3sec</i>
e1=1 (Mixed)	2	2	0	0
e1=1 (Mixed)	5	5	0	0
e1=1 (Mixed)	8	8	0	0
e2=1 (Indian)	1	0	1	0
e2=1 (Indian)	2	0	2	0
e2=1 (Indian)	5	0	5	0
e3=1 (Pakistani)	1	0	0	1
e3=1 (Pakistani)	5	0	0	5
e3=1 (Pakistani)	8	0	0	8
etc				

The inclusion of the terms *e1sec* to *e7sec*, called the interaction between ethnic group and SEC, allows for the relationship between SEC and attainment to vary for different ethnic groups. If these interaction terms are significant we say there is an **interaction effect**.

We have created these variables for you in the LSYPE 15000 MLR  but if you like you can do it yourself using the compute menu (**See Module 1**) or by using the following syntax:

**SYNTAX ALERT!**

```
COMPUTE e1sec= e1 * SEC.
COMPUTE e2sec= e2 * SEC.
COMPUTE e3sec= e3 * SEC.
COMPUTE e4sec= e4 * SEC.
COMPUTE e5sec= e5 * SEC.
COMPUTE e6sec= e6 * SEC.
COMPUTE e7sec= e7 * SEC.
```

**NOTE:** Unlike the SPSS multiple linear regression procedure, other SPSS statistical procedures which we will use later (such as multiple logistic regression) allow you to specify interactions between chosen explanatory variables without having to explicitly calculate the interaction terms yourself. This can save you some time. However it is no bad thing to calculate these terms yourself here because it should help you to understand exactly what SPSS is doing when evaluating interactions. Also whether you calculate these interactions terms yourself or the computer calculates these terms for you, you still have to be able to interpret the interaction coefficients in the regression output. So bear with it!

### **Step 2: Adding the interaction terms to the model**

Now we add the seven variables *e1sec* to *e7sec* to our model. Go to the main regression menu again and add *e1sec*, *e2sec*, *e3sec*, *e4sec*, *e5sec*, *e6sec*, *e7sec*, *sec*, *gender*, and *e1- e7*. As always, *ks3stand* is our *Dependent* variable. Before moving on select the *SAVE* submenu and place a tick in the *unstandardised residuals* box. SPSS will save the predicted values for each case and, as this is the second time we have requested predicted values, will name the new variable *PRE\_2*. These predicted values will be useful later in plotting the interaction effects. Click **OK** to run the regression.

The coefficients table from the SPSS regression output is shown below.

**Figure 3.11.2: Coefficients for Model 5**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	7.811	.226		34.526	.000
sec Socio-economic class of the home	-1.836	.048	-.415	-38.326	.000
gender Gender	1.220	.166	.062	7.366	.000
e1 Mixed heritage	-.772	.759	-.017	-1.016	.309
e2 Indian	-.159	.775	-.004	-.206	.837
e3 Pakistani	-5.528	.918	-.131	-6.024	.000
e4 Bangladeshi	-6.865	1.439	-.137	-4.771	.000
e5 Black Caribbean	-6.560	.914	-.127	-7.177	.000
e6 Black African	-5.332	.916	-.103	-5.819	.000
e7 Any other ethnic group	-.642	.894	-.013	-.718	.473
e1sec Mixed heritage * SEC	.104	.167	.011	.622	.534
e2sec Indian * SEC	.372	.159	.046	2.344	.019
e3sec Pakistani * SEC	.639	.161	.088	3.965	.000
e4sec Bangladeshi * SEC	1.079	.222	.142	4.865	.000
e5sec Black Caribbean * SEC	.581	.200	.052	2.899	.004
e6sec Black African * SEC	.452	.168	.049	2.688	.007
e7sec Any other * SEC	.211	.179	.021	1.180	.238

How do we interpret the output? As before the intercept term (Constant) refers to the predicted values for the reference or base category, which is where SEC=0,

gender=0 (boy) and ethnicity=0 (White British). However the coefficient for SEC now represents the effect of SEC *for the reference group only* (White British students). For White British students, attainment drops by 1.836 standard score points for every unit increase in the value of SEC. To evaluate the effect of SEC for each ethnic group, we adjust the overall SEC coefficient by combining it with the relevant ethnic\*sec interaction term. Thus the slope of SEC for Black Caribbean students is  $-1.836 + .581 = -1.26$ , significantly less steep than the slope for White British students. This is indicated by the significant p value for the Black Caribbean \* SEC interaction term ( $p < .000$ ).

A good way of interpreting this data is to calculate what the predicted age 14 standard scores are from the model:

*Predicted age 14 score for male White British students when SEC=5 (Lower supervisory):*

$$\hat{Y} = \text{intercept} + (5 * \text{SEC coefficient})$$

$$\hat{Y} = 7.81 + (5 * -1.836) = -1.37$$

As gender=0 (male) and ethnic group=0 (White British) there is no contribution from these terms.

*Predicted age 14 score for male Black Caribbean students when SEC=5 (lower supervisory).*

$$\hat{Y} = \text{intercept} + (\text{coeff. for Black Caribbean}) + (5 * \text{SEC coefficient}) + (5 * \text{Black Caribbean by SEC interaction})$$

$$\hat{Y} = 7.81 + -6.56 + (5 * -1.836) + (5 * .581) = -5.02$$

As before we can get SPSS to plot the full set of predicted values for all ethnic group and SEC combinations using the predicted values from the model that we created earlier (by default the variable was named *PRE\_2*). Again we will plot only the values for boys since the pattern for girls is identical, except that all predicted values are 1.22 score points higher.

The syntax below temporarily filters girls out of the analysis AND draws a line graph:

**SYNTAX ALERT!**

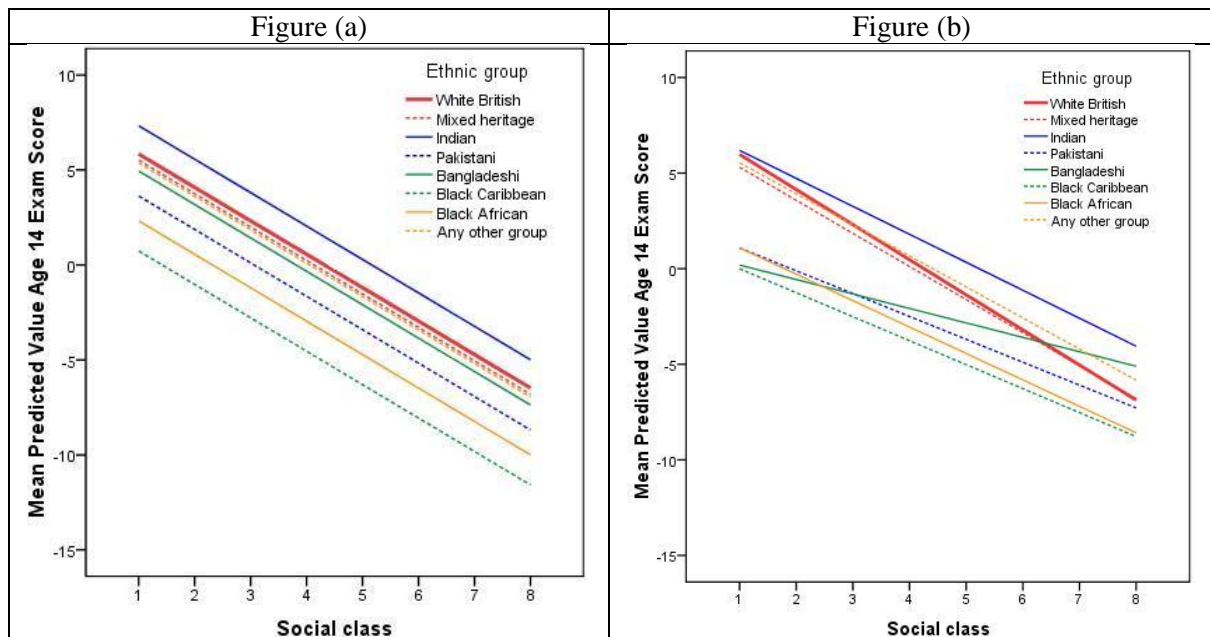
TEMPORARY.

SELECT IF GENDER=0.

GRAPH /LINE(MULTIPLE) MEAN(pre\_2) BY SEC by Ethnic.

The graph is shown in Panel (b) of **Figure 3.11.3**. For reference the regression lines from the model without interactions is shown in Panel (a).

**Figure 3.11.3: Regression lines between ethnic group, SEC and attainment  
(a) without and (b) with interactions between ethnic group and SEC.**



These interaction effects between ethnic groups and SEC are highly statistically significant, particularly for the Pakistani, Bangladeshi, Black Caribbean and Black African groups. We can see this from the sig values for the interaction terms which show  $p < .000$  for Pakistani and Bangladeshi and  $p < .01$  for the Black Caribbean and Black African groups. They are also quite large as can be seen in **Figure 3.11.3**. Note here that the lines are no longer parallel because we have allowed for different slopes in our regression model. Thus the slope for White British students is significantly steeper than for most ethnic minority groups, indicating the difference in attainment between students from high SEC and low SEC homes is particularly pronounced for White British students. Looking at the predicted values we see that the differences between ethnic groups from lower SEC homes are much smaller than the differences among high SEC homes. Rather than a constant difference between White British and Black Caribbean students of 4.25 score points at every SEC value, as indicated by model 3 without the interaction terms, the difference is actually 6.0 points at SEC=1 (Higher managerial and professional homes) , 3.7 points at SEC=5 (lower supervisory) and only 1.9 points at SEC=8 (long term unemployed). There are clear interaction effects.

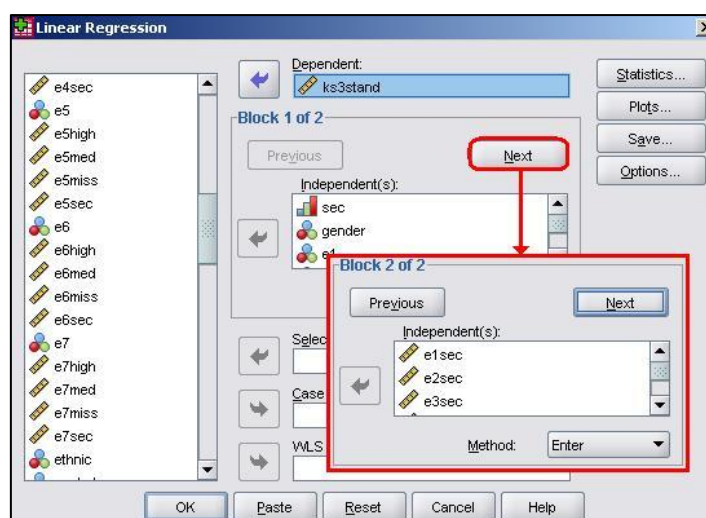
#### **Have we improved the fit of our model?**

The inclusion of the interaction terms does not at first glance appear to have substantially improved the overall fit of the model; the  $r^2$  has only risen from 17.0% to 17.3% (**Figure 3.11.4**). You might ask therefore whether the cost in added complexity caused by adding seven new interaction variables to the model was justified. While we do not appear to have explained a lot of additional variance in

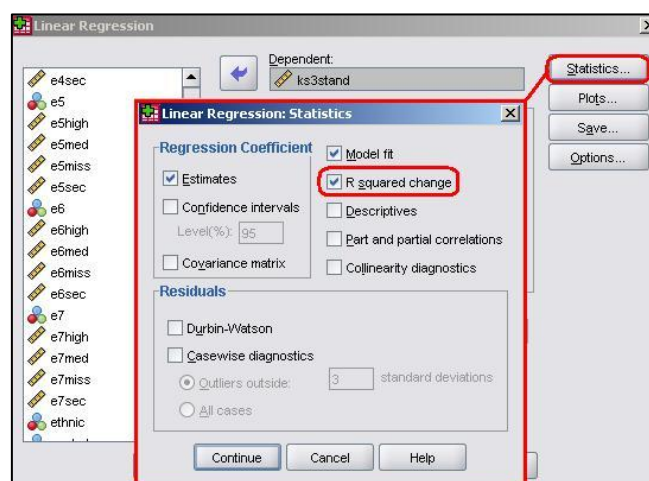


age 14 standard score, we can test the significance of the increase in  $r^2$  by re-running our regression analysis with a few changes.

- (a) On the main regression menu add the explanatory variables in two 'blocks'. The first (Block 1) is entered as normal and should include only SEC, gender and the ethnic dummy variables ( $e1 - e7$ ). This is the 'main effects' model. Click the **Next** button (shown below) to switch to a blank window and enter the variables for the second block. In this second window (Block 2) add the interaction terms which we created ( $e1sec - e7sec$ ). This is the 'interaction' model. Note that the variables you included in Block 1 are automatically included in the interaction model. Including a second block simply means adding new variables to the model specified in the first block.



- (b) Before moving on go to the **Statistics** sub-menu (one of the buttons on the right of the main regression menu) and check the box marked '*R squared change*'. This essentially asks SPSS to directly compare the predictive power ( $r^2$ ) of each model and to test if the difference between the two is statistically significant. This way we can directly test whether adding the interaction effect terms improves our model.



When you are happy with the setup, click **OK** to run the analysis.

For an alternative method for re-running the analysis use the syntax below.

**SYNTAX ALERT!**

```
REGRESSION /MISSING LISTWISE /STATISTICS COEFF R ANOVA CHANGE
/CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT ks3stand
/METHOD=ENTER sec gender e1 to e7 /ENTER= e1sec to e7sec.
```

We get the following output under **Model Summary**.

**Figure 3.11.4: Model summary and change statistics for Model 5**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.412 <sup>a</sup>	.170	.169	9.037	.170	270.135	9	11896	.000
2	.416 <sup>b</sup>	.173	.172	9.021	.003	7.000	7	11889	.000


We are interested here in the columns headed '*Change Statistics*' and specifically in the second row which compares Block 2 (the interaction model) against Block 1 (the main effects model). We can see that the increase in  $r^2$  for the interaction model, while small at 0.3%, is highly statistically significant ('Sig. F Change',  $p < .000$ ). So while the increase in overall  $r^2$  is small the model with interactions gives a significantly better fit to the data than we get if the interactions are not included. In short our interaction model is a much more precise and accurate summary of the pattern of mean scores across our different explanatory variables. However the relatively low  $r^2$  at 17.3% indicates that there is considerable variation in attainment between students within each category of the explanatory variables. Thus predictions of the attainment of *any individual student* based simply on knowledge of their SEC, ethnicity and gender, will have a large degree of imprecision. We will see later how adding further explanatory variables (such as prior attainment at age 11, **Page 3.13**) can substantially improve the  $r^2$  for the model.

### 3.12 Exploring interactions between two nominal variables (Model 6)

The above process is relatively easy to compute (yes, I'm afraid it will get a little harder below!) but has the same problem of data loss we identified earlier. The 2941 cases that have no valid value for SEC are excluded from the model. As before we can avoid this data loss if we transform SEC to a set of dummy variables and explicitly include missing values as an additional dummy category. This has the advantages outlined on **Page 3.8**. However if we keep the full eight SEC categories this would lead to a very large number of interaction terms: 7 ethnic group dummies \* 8 SEC dummies (including the dummy variable for missing cases) = 56 separate interaction terms! This is substantially higher than the seven new variables we included when we treated SEC as a continuous variable on **page 3.11**. Needless to say this analysing this would be a painful experience... In the interest of a parsimonious model it is unhelpful to add so many additional variables. One way to make the interactions more manageable is to 'collapse' the SEC variable into a smaller number of values by combining some categories.

#### **Step 1: Collapse SEC to three levels**

The Office for National Statistics socio-economic class (NS-SEC) coding system was used in this study. The system includes criteria for identifying 8, 5 or 3 class versions of SEC (see **Resources** page). To simplify the structure we recode SEC from the eight to the three class version, which combines higher and lower managerial and professional (class 1 and 2), intermediate, small employers and lower supervisory (classes 3 to 5) and semi-routine, routine and unemployed groups (classes 6 to 8). We will also create a new category for those with missing values for SEC.

All of the variables and interaction terms over the rest of this page have already been created for you in the LSYPE 15000 MLR , so you do not need to create them yourself. However, if you would like to follow the process exactly then feel free! You can either use the **Recode** menu (see **Module 1**) or you can run the syntax below (which will also produce a frequency table for the new variables) to create the required 3 item SEC variable:

#### **SYNTAX ALERT!**

```
RECODE SEC (0=0)(1 thru 2=1)(3 thru 5=2)(6 thru 8=3) INTO SECshort.  
VALUE labels SECshort 0'missing' 1'Managerial & professional' 2 'Intermediate'  
3'Routine, semi-routine or unemployed'.  
FORMATS secshort (F1.0).  
FREQ SECshort.
```

The new variable looks like this:

**Figure 3.12.1: Frequencies for collapsed SEC variable**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 missing	2941	18.6	18.6	18.6
	1 Managerial & professional	4650	29.5	29.5	48.1
	2 Intermediate	4058	25.7	25.7	73.9
	3 Routine, semi-routine or unemployed	4121	26.1	26.1	100.0
	Total	15770	100.0	100.0	

We can then calculate dummy variables for the collapsed variable, taking category 3 (low SEC) as the base or reference category.

**SYNTAX ALERT!**


RECODE secshort (1=1)(else=0) into SEChigh.  
 RECODE secshort (2=1)(else=0) into SECmed.  
 RECODE secshort (0=1)(else=0) into SECmiss.  
 FORMATS sechigh SECmed SECmiss (F1.0).

We should note that collapsing a variable is not only useful if we want to test interactions, it is most often necessary where the number of cases in a cell is particularly low. **Figure 3.12.2** shows the Crosstab of SEC and ethnic group. We can see there were only three Bangladeshi students in SEC class 1. By combining SEC classes 1 and 2 we increase the number of Bangladeshi students in the high SEC cell to 35. This is still relatively low compared to other ethnic groups, but will provide a more robust estimate than previously.

**Figure 3.12.2: Crosstabulation of SEC by ethnic group**

		sec MP social class								Total
		1 Higher Managerial and professional occupations	2 Lower managerial and professional occupations	3 Intermediate occupations	4 Small employers and own account workers	5 Lower supervisory and technical occupations	6 Semi-routine occupations	7 Routine occupations	8 Never worked/long term unemployed	
ethnic	0 White British	1211	2248	639	1060	1048	1066	917	289	8478
	1 Mixed heritage	75	195	54	64	55	92	62	54	651
	2 Indian	72	141	58	150	91	112	125	55	804
	3 Pakistani	32	75	23	189	34	74	101	189	717
	4 Bangladeshi	3	32	5	56	62	65	62	212	497
	5 Black Caribbean	29	146	71	27	52	70	36	39	470
	6 Black African	48	109	39	21	28	77	31	132	485
	7 Any other group	80	95	25	78	52	53	43	80	506
Total		1550	3041	914	1645	1422	1609	1377	1050	12608

**Step 2: Create the interaction terms**

As before to create the interaction terms we simply multiply each ethnic dummy variable by the relevant SEC dummy variable. Again, we have done this for you in the LSYPE 15000 MLR  dataset but if you want to do it yourself you can use the **Compute** option (see **Module 1**) or the syntax below.

**SYNTAX ALERT!**

```
COMPUTE e1high = e1*sechigh.  
COMPUTE e1med = e1*secmed.  
COMPUTE e1miss = e1*secmiss.  
COMPUTE e2high = e2*sechigh.  
COMPUTE e2med = e2*secmed.  
COMPUTE e2miss = e2*secmiss.  
COMPUTE e3high = e3*sechigh.  
COMPUTE e3med = e3*secmed.  
COMPUTE e3miss = e3*secmiss.  
COMPUTE e4high = e4*sechigh.  
COMPUTE e4med = e4*secmed.  
COMPUTE e4miss = e4*secmiss.  
COMPUTE e5high = e5*sechigh.  
COMPUTE e5med = e5*secmed.  
COMPUTE e5miss = e5*secmiss.  
COMPUTE e6high = e6*sechigh.  
COMPUTE e6med = e6*secmed.  
COMPUTE e6miss = e6*secmiss.  
COMPUTE e7high = e7*sechigh.  
COMPUTE e7med = e7*secmed.  
COMPUTE e7miss = e7*secmiss.  
FORMATS e1high to e7miss (F1.0).  
VAR LABELS  
  /e1high 'Mixed heritage * high' e1med 'Mixed heritage * medium' e1miss 'Mixed  
heritage * missing'  
  /e2high 'Indian * high' e2med 'Indian * medium' e2miss 'Indian * missing'  
  /e3high 'Pakistani * high' e3med 'Pakistani * medium' e3miss 'Pakistani * missing'  
  /e4high 'Bangladeshi * high' e4med 'Bangladeshi * medium' e4miss 'Bangladeshi *  
missing'  
  /e5high 'Black Caribbean*high' e5med 'Black Caribbean*med' e5miss 'Black  
Caribbean*missing'  
  /e6high 'Black African*high' e6med 'Black African * medium' e6miss 'Black African *  
missing'  
  /e7high 'Any other * high' e7med 'Any Other * medium' e7miss 'Any Other *  
missing'.
```

**Step 3: Add the interaction terms to the model**

Now we can see the relationship between ethnic group, SEC and attainment using the full sample of students. Run the regression analysis including all main and interaction variables: *ks3stand* (Dependent), *SEChigh*, *SECmed*, *SECmiss*, *gender*, *e1-e7* and all interaction terms (e.g. *e1high*, *e1med*, *e1miss*, *e2high*, *e2med*, *e2miss*,

etc. Twenty-one terms in total). Remember to request predicted values from the **SAVE** submenu (which will be saved by default to the variable *PRE\_3* because *this is the third time we have asked SPSS to save predicted values*). **Figure 3.12.3** shows the coefficient output.

**Figure 3.12.3: Regression coefficients output for Model 6**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.142	.211		-19.630	.000
	SEChigh	9.056	.258	.412	35.057	.000
	SECmed	4.147	.268	.184	15.480	.000
	SECmiss	2.500	.311	.096	8.040	.000
	gender Gender	1.073	.154	.054	6.973	.000
	e1 Mixed heritage	-.437	.687	-.010	-.637	.524
	e2 Indian	1.883	.580	.048	3.247	.001
	e3 Pakistani	-1.802	.529	-.044	-3.407	.001
	e4 Bangladeshi	-.452	.546	-.010	-.827	.408
	e5 Black Caribbean	-2.668	.804	-.051	-3.319	.001
	e6 Black African	-3.388	.643	-.066	-5.273	.000
	e7 Any other ethnic group	.361	.739	.007	.489	.625
	e1high Mixed heritage * high	-.353	.912	-.005	-.387	.699
	e1med Mixed heritage * medium	.494	1.004	.005	.492	.623
	e1miss Mixed heritage * missing	-.046	1.092	.000	-.042	.966
	e2high Indian * high	-1.650	.883	-.020	-1.870	.062
	e2med Indian * medium	-.054	.812	-.001	-.066	.947
	e2miss Indian * missing	.130	.905	.002	.144	.885
	e3high Pakistani * high	-1.259	1.070	-.011	-1.176	.240
	e3med Pakistani * medium	-1.979	.815	-.025	-2.428	.015
	e3miss Pakistani * missing	-2.278	.855	-.028	-2.664	.008
	e4high Bangladeshi * high	-1.861	1.685	-.009	-1.104	.270
	e4med Bangladeshi * medium	-2.733	1.016	-.025	-2.690	.007
	e4miss Bangladeshi * missing	-1.443	.867	-.018	-1.665	.096
	e5high Black Caribbean * high	-3.167	1.086	-.034	-2.916	.004
	e5med Black Caribbean * medium	-1.304	1.128	-.013	-1.156	.248
	e5miss Black Caribbean * missing	-.680	1.235	-.006	-.551	.582
	e6high Black African*high	-1.473	1.015	-.015	-1.451	.147
	e6med Black African * medium	1.053	1.220	.008	.863	.388
	e6miss Black African * missing	2.244	1.085	.020	2.069	.039
	e7high Any other * high	-.402	1.069	-.004	-.376	.707
	e7med Any Other * medium	-.857	1.083	-.009	-.792	.429
	e7miss Any Other * missing	-.441	1.114	-.004	-.396	.692

The output initially might look a little overwhelming as there are a considerable number of variables included, but this is still small compared to many models! We're not sure if that is reassuring or not... The good thing is that the interpretation of the output is substantially the same as we saw on **Page 3.11**.

- The constant coefficient gives the intercept for our reference group, which is White British, boys from low SEC homes.

- The SEChigh and SECmedium coefficients are directly interpretable as the 'boost' to attainment for *White British pupils* associated with residing in medium and high SEC homes. There is a strong boost for Medium SEC (4.2 score points) and particularly for high SEC homes (9.1 score points).
- The ethnic coefficients represent the difference in attainment between each ethnic group and White British students in the reference group (low SEC homes). We see for example that Black Caribbean low SEC students on average score about 2.7 points lower than their White British low SEC peers. In contrast Indian students score about 1.9 points higher than their low SEC White British peers. Both effects are highly statistically significant ( $p < .001$ ).
- The ethnic \* SEC interaction terms show how the boosts associated with medium and high SEC homes vary by ethnic group. Take for example Black Caribbean students from high SEC homes. The boost associated with high SEC homes for White British students is about 9.1 points, but for Black Caribbean students it is lower,  $9.056 + (-3.167) = 5.9$  score points.

As before a good way of interpreting this data is to calculate what the predicted age 14 standard scores are from the model.

*Predicted age 14 score for White British boys from high SEC homes:*

As White British boys are the reference group this value will be calculated from just two terms: intercept + high SEC coeff.

$$\hat{Y} = -4.142 + 9.056 = 4.914$$

*Predicted age 14 score for Black Caribbean boys from high SEC homes:*

This will be calculated from: intercept + Black Caribbean coeff. + high SEC coefficient + Black Caribbean\*High interaction coeff.

$$\hat{Y} = -4.142 + -2.668 + 9.056 + -3.167 = -.921$$

Note that gender is just modelled as a main effect (it has not been allowed to interact with SEC or ethnic group), so you would just add 1.073 to get the predicted values for girls from any ethnic or SEC group. Again we can plot the predicted value that we saved earlier when we specified the regression model (the values were saved as the variable *PRE\_3*).

**NOTE:** For the purpose of plotting this graph we have excluded the cases where SEC was missing by first setting the missing values code for SECshort to 0 (remember 0 indicated missing values). If you want to know how we did this, view the syntax below:



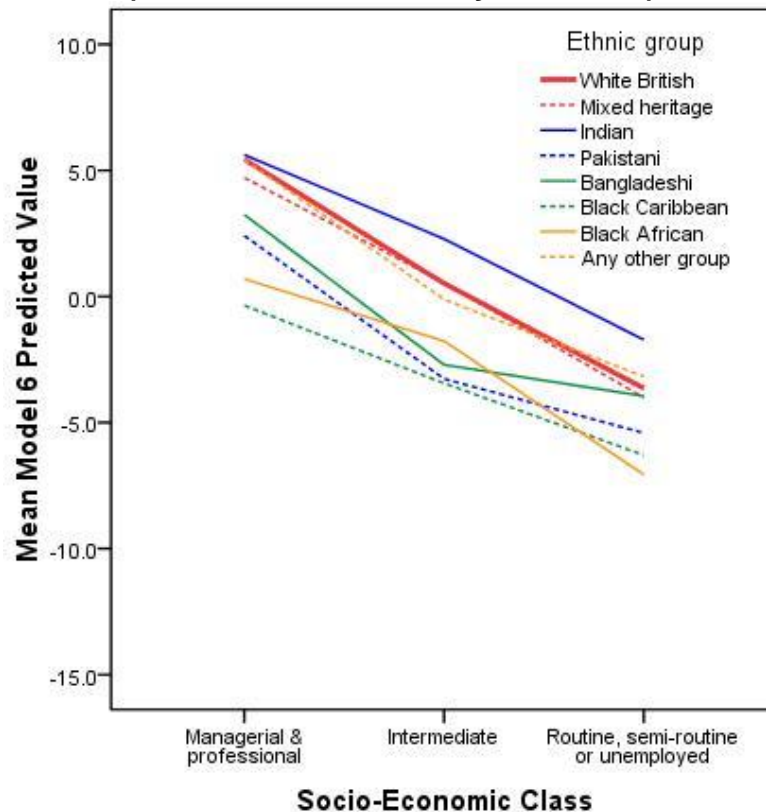
**SYNTAX ALERT!**

*MISSING VALUES secshort(0).*

*GRAPH /LINE(MULTIPLE)MEAN(pre\_3) BY secshort by Ethnic.*

If you prefer, you can use the **Select Cases** option.

**Figure 3.12.4: Predicted values for attainment at age 14 including interactions between ethnic group and SEC (Both coded as dummy variables)**



There are significant interactions between the Pakistani and Bangladeshi groups and medium SEC ( $p=.015$  and  $p=.007$  respectively) and between Black Caribbean and high SEC ( $p=.004$ ). The effects are not only statistically significant they are also quite large, as can be seen in **Figure 3.12.3**. Note here that the regression lines are no longer parallel because we have allowed for different slopes in our regression model. The slope for White British students is significantly steeper than for most ethnic minority groups indicating the differences between high SEC and low SEC homes is particularly pronounced for White British students. Looking at the coefficients in **Figure 3.12.3** we see that the differences between ethnic groups from lower SEC homes are much smaller than the differences among high SEC homes. Note that the significance tests for the ethnic group coefficients in the SPSS output are for ethnic differences in the reference group of low SEC homes. If we want to test the significance of the ethnic group differences in high SEC homes we can just change



our reference group to high SEC. Similarly if we want to test the significance of the ethnic difference at medium SEC we could change reference group to Medium SEC.

### **Have we improved the fit of our model?**

Again we can test whether we have significantly improved the  $r^2$  by entering the variables in two blocks and calculating the  $r^2$  change (see **Page 3.11**). Block 1 should include SEC, gender and ethnic group (the 'main effects' model), while the interaction terms should be added in Block 2 (the 'interaction' model).

**Figure 3.12.5: Model summary and change statistics for Model 6**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.365 <sup>a</sup>	.133	.132	9.303	.133	227.169	10	14821	.000
2	.368 <sup>b</sup>	.135	.134	9.296	.003	2.083	21	14800	.003

In **Figure 3.12.5**, which we received as part of our regression output, we are interested here in the columns headed 'Change Statistics' and specifically in the second row which compares Block 2 (the interaction model) against Block 1 (the main effects model). We can see that the increase in  $r^2$  for the interaction model, while small at 0.3%, is highly statistically significant ( $p<.003$ ). So while the increase in overall  $r^2$  is small, the model with interactions gives a significantly better fit to the data than we get if the interactions are not included.

### 3.13: A value added model (Model 7)

In all the models we have computed so far we have not included the variable which we saw in the SLR module had the biggest impact on age 14 score, namely test score at age 11. We now will include age 11 test score in our model. We should note that adding this variable changes the conceptual nature of our outcome. By evaluating attainment at age 14, after the variance associated with attainment at age 11 has been taken into account, we are effectively evaluating *progress* between age 11 and age 14. Positive residuals will indicate students making greater than average progress age 11 to age 14, while negative residuals will indicate pupil making less than the average progress between age 11 and 14.

So let's run Model 6 again, but this time also include age 11 score as a explanatory variable. Additionally, so that we can test whether our interaction terms give a statistically significant increase in  $r^2$  or not, we will enter the variables in two separate blocks and calculate the  $r^2$  change, as we did on **Page 3.11**. First, put *ks3stand* in the *Dependent* box. Next add *ks2stand*, *SEChigh*, *SECmed*, *SECmiss*, *gender*, and *e1-e7* as explanatory variables in Block 1. Finally add all the interaction terms (e.g. *e1high*, *e1med*, *e1miss*, *e2high*, *e2med*, *e2miss*, etc.) as explanatory variables in Block 2. If you want to save the predicted variables they will be called *PRE\_4* but you do not need them this time so you may prefer to uncheck the *unstandardized Predictors* box in the **SAVE** menu.

The figure below shows the model summary (**Figure 3.13.1**). The  $r^2$  increase massively from 17.0% in Model 6 up to 79.6% in this model. This testifies to the power of prior attainment as a predictive factor. By knowing how well a student was achieving at age 11, we can explain nearly 80% of the variance in how well they achieve at age 14. It is notable though that the ethnic by SEC interaction terms do **not** increase  $r^2$  significantly. Comparing the main effects model with the interaction model we see that there is no evidence that the inclusion of the interaction terms raises the  $r^2$  of the model ('Sig. F Change',  $p = .127$ ). The model explains 79.6% of the variance in age 14 scores with or without the interaction effects.

**Figure 3.13.1: Model summary when including prior attainment**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.892 <sup>a</sup>	.796	.796	4.408	.796	4.490E3	12	13832	.000
2	.892 <sup>b</sup>	.796	.796	4.407	.000	1.356	21	13811	.127

**Figure 3.13.2** presents the regression output. While age 11 score is the most powerful predictor in the model (look at the Beta values to get a relative idea) it is not the only significant explanatory variable in the model. SEC still has a significant association with student progress. White British students from medium SEC homes

make 1.2 points greater progress than students of the same prior attainment from low SEC homes, and White British students from high SEC homes make 2.5 points more progress than students with the same age 11 scores from low SEC homes. Gender is also still highly significant, with girls making 0.54 points more progress than boys after control for prior attainment and SEC. Remember that these are effects on student progress. They therefore indicate that social class and gender gaps increase between age 11 and age 14, that is the gaps get wider.

**Figure 3.13.2: Coefficients output (Model 7)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.272	.103		-12.377	.000
	SEChigh	2.408	.128	.112	18.748	.000
	SECmed	1.199	.130	.055	9.236	.000
	SECmiss	.722	.151	.028	4.795	.000
	gender Gender	.539	.075	.028	7.169	.000
	e1 Mixed heritage	-.616	.334	-.014	-1.846	.065
	e2 Indian	1.420	.281	.037	5.059	.000
	e3 Pakistani	.569	.257	.014	2.213	.027
	e4 Bangladeshi	.322	.269	.007	1.197	.231
	e5 Black Caribbean	-.698	.404	-.014	-1.727	.084
	e6 Black African	.610	.367	.011	1.663	.096
	e7 Any other ethnic group	1.748	.381	.034	4.590	.000
	e1high Mixed heritage * high	.549	.445	.007	1.234	.217
	e1med Mixed heritage * medium	.949	.487	.011	1.949	.051
	e1miss Mixed heritage * missing	.949	.533	.009	1.781	.075
	e2high Indian * high	-.339	.429	-.004	-.790	.429
	e2med Indian * medium	.069	.393	.001	.177	.860
	e2miss Indian * missing	.334	.440	.004	.760	.447
	e3high Pakistani * high	.844	.539	.007	1.566	.117
	e3med Pakistani * medium	-.358	.396	-.005	-.904	.366
	e3miss Pakistani * missing	-.655	.419	-.008	-1.562	.118
	e4high Bangladeshi * high	.558	.828	.003	.673	.501
	e4med Bangladeshi * medium	-1.078	.499	-.010	-2.161	.031
	e4miss Bangladeshi * missing	.023	.429	.000	.054	.957
	e5high Black Caribbean * high	-.454	.539	-.005	-.842	.400
	e5med Black Caribbean * medium	-.609	.560	-.006	-1.087	.277
	e5miss Black Caribbean * missing	-.044	.616	.000	-.072	.943
	e6high Black African * high	-.365	.558	-.003	-.654	.513
	e6med Black African * medium	.287	.667	.002	.431	.667
	e6miss Black African * missing	-.090	.603	-.001	-.149	.881
	e7high Any other * high	-.958	.554	-.009	-1.730	.084
	e7med Any Other * medium	-.846	.553	-.008	-1.529	.126
	e7miss Any Other * missing	-.768	.575	-.007	-1.335	.182
	ks2stand Age 11 standard marks	.845	.004	.858	211.393	.000

Why are the interaction terms no longer significant? The results indicate that some of the ethnic by SEC interactions at age 14 reflect differences in prior attainment at age 11 (not included in earlier models). For example Black Caribbean high SEC students

tend to have lower scores at age 11 than White British high SEC students, and this is one reason the age 14 attainment of Black Caribbean high SEC students may have been lower than their White British high SEC peers. Once we take this factor into account the interaction terms are no longer significant. The only exception is Bangladeshi \* Medium SEC ( $p=.031$ ). This may have emerged as statistically significant at the .05 confidence level due to the fact that multiple pairwise comparisons (t-tests) were made. At the .05 level you can expect 1 in 20 pairwise comparisons to emerge as statistically significance by chance alone!

We have arrived at the final model for our regression analysis. HOORAY! Given the interaction terms are not significant, we revert to the eight class SEC variable we used earlier, treated as a set of dummy variables. Our final model (Model 7) is: *ks3stand* (Dependent), *sc0-sc7*, *gender*, *e1-e7* and *ks2stand*.

The  $r^2$  value of this model is 79.6%, and the coefficients table is presented below.

**Figure 3.13.3: Final regression model**

Model	Unstandardized Coefficients		Standardized Coeffs	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1.621	.166		-9.752	.000
sc1 Higher managerial & professional	3.478	.203	.104	17.166	.000
sc2 Lower managerial & professional	2.388	.180	.097	13.266	.000
sc3 Intermediate	1.494	.218	.037	6.868	.000
sc4 small employers & own account	2.006	.191	.064	10.491	.000
sc5 Lower supervisory & technical	.954	.196	.029	4.864	.000
sc6 semi-routine	.593	.192	.019	3.093	.002
sc7 routine	.339	.196	.010	1.723	.085
sc0 missing	1.070	.175	.042	6.104	.000
gender Gender	.556	.075	.029	7.429	.000
e1 Mixed heritage	-.027	.170	-.001	-.160	.873
e2 Indian	1.413	.150	.037	9.420	.000
e3 Pakistani	.370	.161	.009	2.299	.021
e4 Bangladeshi	.265	.185	.006	1.430	.153
e5 Black Caribbean	-.926	.199	-.018	-4.647	.000
e6 Black African	.571	.221	.010	2.586	.010
e7 Any other ethnic group	1.117	.201	.022	5.556	.000
ks2stand Age 11 standard marks	.842	.004	.855	210.502	.000

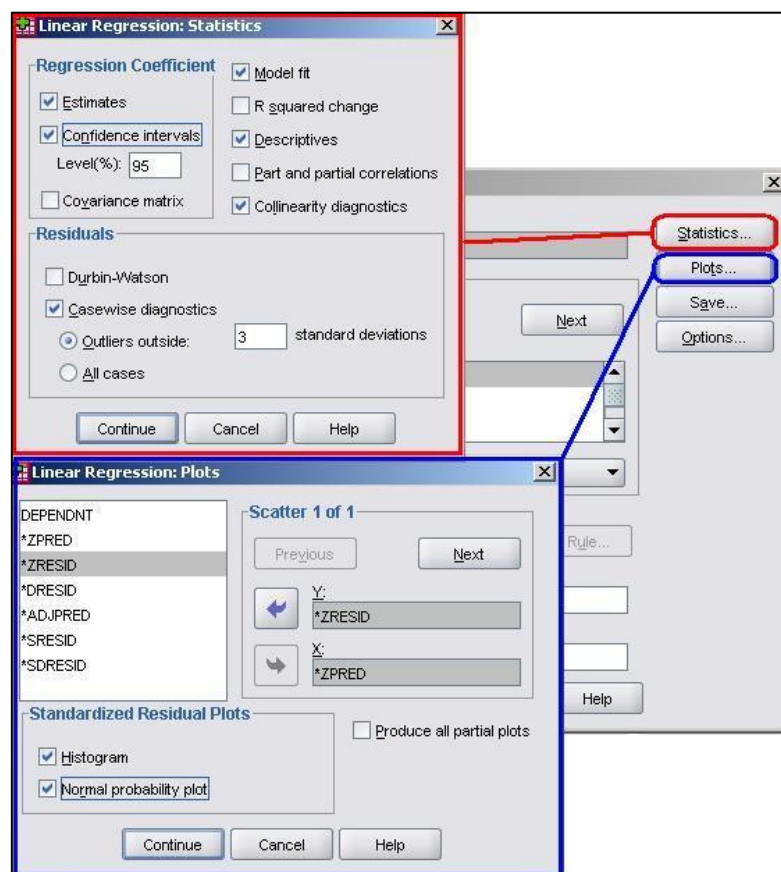
The next step is to test the adequacy of the model specification and whether the assumptions of multiple regression analysis, as outlined on **Page 3.3**, have been successfully met.

### 3.14 Model diagnostics and checking your assumptions

So far we have looked at building a multiple regression model in a very simple way. We have not yet engaged with the assumptions and issues which are so important to achieving valid and reliable results. In order to obtain the relevant diagnostic statistics you will need to run the analysis again, this time altering the various SPSS option menus along the way.

Let's use this opportunity to build model 7 from the beginning. Take the following route through SPSS: **Analyse > Regression > Linear** and set up the regression. We will use model 7 which is: *ks3stand* as the outcome variable, with the explanatory variables as *ks2stand*, *gender*, *e1-e7* (ethnicity) and *sc0-sc7* (Socio-economic class). Don't click ok yet!

We will need to make changes in the submenus in order to get access to the necessary information for checking the assumptions and issues. Let's start with the *Statistics* and *Plots* submenus.



Many of these options should be familiar to you from the previous module.

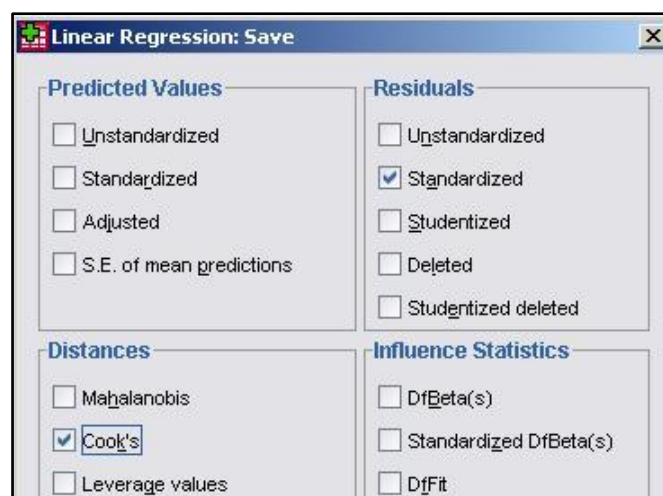
We will request the *Estimates*, *Descriptives* and *Model fit* from the **Statistics** submenu. We also recommend that you get the *Confidence Intervals* this time as

they provide the range of possible values for each of your explanatory variable's  $b$  coefficients within which you can be 95% sure that the true value lies. In addition we now have the potential issue of our explanatory variables being too highly correlated, so we should also get hold of the *Multicollinearity Diagnostics*.

It is worth also collecting the *Casewise Diagnostics*. These will tell us which cases have residuals that are three or more standard deviations away from the mean. These are the cases with the largest errors and may well be outliers (note that you can change the number of standard deviations from 3 if you wish to be more or less conservative).

You should exercise the same options as before in the **Plots** menu. Create a scatterplot which plots the standardized predicted value (ZPRED) on the x-axis and the standardized residual on the y-axis (ZRESID) so that you can check the assumption of homoscedasticity. As before we should also request the *Histogram* and *Normal Probability Plot* (P-P plot) in order to check that our residuals are normally distributed. Head back to **Page 2.7** of our previous module if you need to jog your memory about how to do all of this on SPSS.

We should also obtain some useful new variables from the **Save** menu.



From the Residuals section it is worth requesting the *Standardized* residuals as these can be useful for additional analysis. It is also worth getting the *Cook's* distance from the Distances section. The Cook's distance statistic is a good way of identifying cases which may be having an undue influence on the overall model. Cases where the Cook's distance is greater than 1 may be problematic. Once you have obtained them as a separate variable you can search for any cases which may be unduly influencing your model. We don't need the *Unstandardized Predicted values* for our purposes here.

Now that we have selected our outcome and explanatory variables and altered all of the relevant submenus it is time to run the analysis... click **OK**.

SPSS seems to have had a great time and has spat out a vast array of tables and plots, some of which are so alarmingly large that they do not even fit on the screen! We hope that, by now, you are getting used to SPSS being overenthusiastic and do not find this too disconcerting! Rather than reproduce all of that extraneous information here we will discuss only the important bits.

The **Descriptive Statistics** table is always worth glancing over as it allows you to understand the basic spread of your data. Note that the dummy variables for ethnicity and SEC can only vary between 0 and 1. Next we have a truly monstrous **Correlations** table. We have not included it because it would probably crash the internet... or at least make this page harder to read! However, it is very useful to know the correlations between the variables and whether they are statistically significant. The **Correlations** table is also useful for looking for multicollinearity. If any two explanatory variables have a Pearson's coefficient of 0.80 or greater there may be cause for concern – they may actually be measures of the same underlying factor. We have also ignored the **Variables Entered/Removed** table as it merely provides a summary of all of the variables we have included in our current model.

The **model summary** (**Figure 3.14.1**) provides us with a new value for  $r^2$  for our expanded model,  $r^2 = .797$ . The model explains about 80% of the variance in age 14 score. From the ANOVA table we can see that  $F = 3198.072$ ,  $df = 17$ ,  $p < .0005$ . This means that, as hoped, the regression model we have constructed is better at predicting the outcome variable than using the mean outcome (it generates a significantly smaller sum of residuals).

**Figure 3.14.1:  $r$  and  $r^2$  for expanded model**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.893 <sup>a</sup>	.797	.797	4.392

The **Coefficients** table (**Figure 3.14.2**) is frighteningly massive to account for the large number of variables it now encompasses. However, aside from a few small additions, it is interpreted in the exact same way as in the previous example so don't let it see your fear! We won't go through each variable in turn (we think you're probably ready to have a go at interpreting this yourself now) but let's look at the key points for diagnostics. Note we've had to shrink our table down to fit it on the screen!

**Figure 3.14.2: Coefficients table for full model**

Model	Unstandardized Coefficients		Stand. Coeff	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tol	VIF
(Constant)	-1.621	.166		-9.75	.000	-1.946	-1.295		
Age 11 standard marks	.842	.004	.855	210.50	.000	.834	.849	.888	1.13
Gender	.556	.075	.029	7.43	.000	.409	.703	.996	1.00
Mixed heritage	-.027	.170	-.001	-.16	.873	-.361	.307	.975	1.03
Indian	1.413	.150	.037	9.42	.000	1.119	1.707	.966	1.04
Pakistani	.370	.161	.009	2.30	.021	.055	.686	.916	1.09
Bangladeshi	.265	.185	.006	1.43	.153	-.098	.627	.905	1.10
Black Caribbean	-.926	.199	-.018	-4.65	.000	-1.317	-.536	.972	1.03
Black African	.571	.221	.010	2.59	.010	.138	1.003	.966	1.04
Any other ethnic group	1.117	.201	.022	5.57	.000	.723	1.511	.975	1.03
SEC missing	1.070	.175	.042	6.10	.000	.726	1.413	.311	3.22
Higher managerial and professional occupations	3.478	.203	.104	17.17	.000	3.081	3.875	.398	2.52
Lower managerial and professional occupations	2.388	.180	.097	13.27	.000	2.035	2.741	.273	3.66
Intermediate occupations	1.494	.218	.037	6.87	.000	1.068	1.921	.506	1.97
Small employers and own account workers	2.006	.191	.064	10.49	.000	1.631	2.380	.394	2.54
Lower supervisory and technical occupations	.954	.196	.029	4.86	.000	.569	1.338	.412	2.43
Semi-routine occupations	.593	.192	.019	3.09	.002	.217	.969	.399	2.51
Routine occupations	.339	.196	.010	1.72	.085	-.046	.724	.429	2.33

Because we requested multicollinearity statistics and confidence intervals from SPSS you will notice that we have four more columns at the end of the coefficients table. The *95% confidence interval* tells us the upper and lower bounds for which we can be confident that the true value of b coefficient lies. Examining this is a good way of ascertaining how much error there is in our model and therefore how confident we can be in the conclusions that we draw from it. Finally the *Collinearity Statistics* tell us the extent to which there is multicollinearity between our variables. If the value for the *Tolerance* is less than 10 and the value of the *VIF* is close to 1 for each explanatory variable then there is probably no cause for concern. The VIF for some of the SEC variables suggests we may have some issues with multicollinearity which require further investigation. However looking at the **correlations** table reveals that correlations between variables are weak despite often being statistically significant which allays our concerns about multicollinearity.



**Collinearity Diagnostics** emerge from our output next. We will not discuss this here because understanding the exact nature of this table is beyond the scope of this website. The table is part of the calculation of the collinearity statistics. The **Casewise Diagnostics** table is a list of all cases for which the residual's size exceeds 3. We haven't included it here because as you can see there are over 100 cases with residuals of this size! There are several ways of dealing with these outliers. If it looks as though they are the result of a mistake during data entry the case could be removed from analysis. Close to one hundred cases seems like a lot but is actually not too unexpected given the size of our sample – it is less than 1% of the total participants. The outliers will have a relatively small impact on the model but keeping them means our sample may better represent the diversity of the population.

We created a variable which provides us with the **Cook's Distance** for each case which is labelled as *COO\_1* in your dataset. If a case has a Cook's distance of greater than 1 it may be an overly influential case that warrants exclusion from the analysis. You can look at the descriptive statistics for Cook's distance to ascertain if any cases are overly influential. If you have forgotten how to calculate the descriptive statistics, all you need to do is take the following route through SPSS: **Analyze > Descriptive Statistics > Descriptives** (see **Module 1** if you require a reminder). **Figure 3.14.3** shows the output. As you can see the *maximum* value of Cook's distance in our sample is .00425 which far less than the value of 1 which may be a cause for concern. We do not appear to have any problematic cases in our sample.

**Figure 3.14.3: Descriptive statistics for Cook's distance Model 7**

	N	Minimum	Maximum	Mean	Std. Deviation
Cook's Distance	13845	.00000	.00425	.0000758	.00017036
Valid N (listwise)	13845				

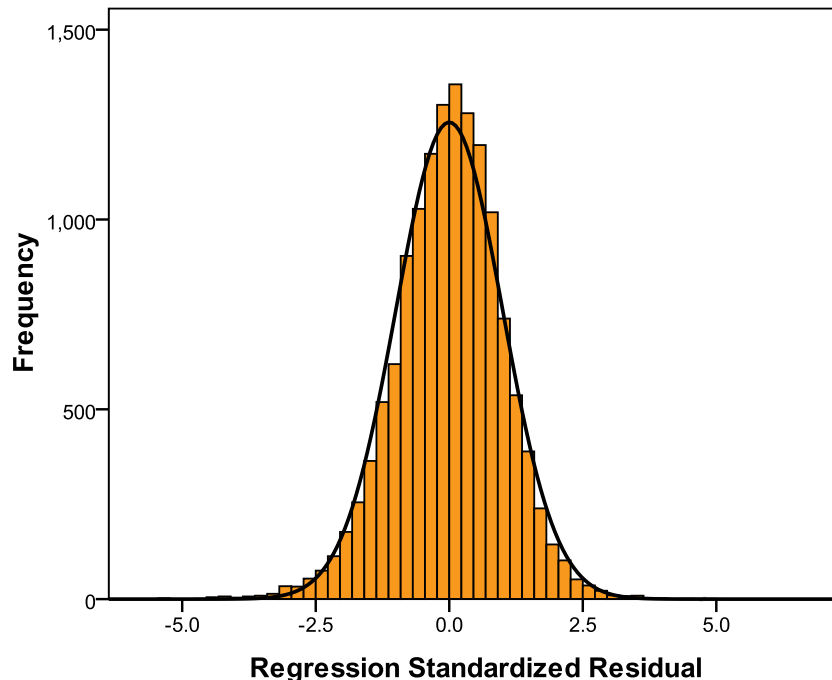
The **Residuals Statistics** (**Figure 3.14.4**) summarize the nature of the residuals and predicted values in the model (big surprise!). It is worth glancing at so you can get a better understanding of the spread of values that the model predicts and the range of error within the model.

**Figure 3.14.4: Residual statistics for model**

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-22.75	35.24	.41	8.704	13845
Residual	-29.355	20.986	.000	4.390	13845
Std. Predicted Value	-2.660	4.002	.000	1.000	13845
Std. Residual	-6.683	4.778	.000	.999	13845

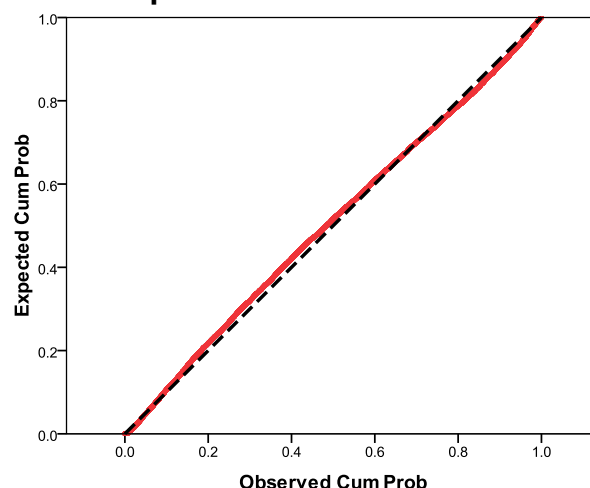
Next we have the plots and graphs that we requested. A **Histogram** of the residuals (**Figure 3.14.5**) suggests that they are close to being normally distributed but there are more residuals close to zero than perhaps you would expect.

**Figure 3.14.5: Histogram of standardized model residuals**



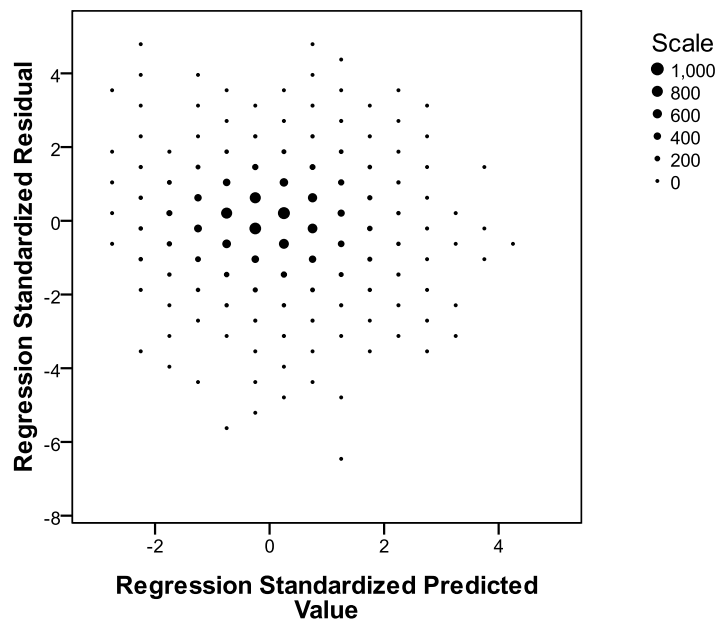
The **P-P plot** (**Figure 3.14.6**) is a little more reassuring. There does seem to be some deviation from normality between the observed cumulative probabilities of 0.2 and 0.6 but it appears to be minor. Overall there does not appear to be a severe problem with non-normality of residuals.

**Figure 3.14.6: P-P plot of standardized model residuals**



This **Scatterplot** (which we have altered with binning in **Figure 3.14.7** to clarify) shows that the residuals are not distributed in any pattern with the predicted values. This suggests that our model does not violate the assumption of homoscedasticity.

**Figure 3.14.7: Scatterplot of standardized residuals against standardized predicted values**



Finally, we created a variable for the **Standardized Residuals** of the model which has appeared in your data file labelled as *ZRE\_1*. If you wanted to perform certain analyses regarding which groups or cases the model is more accurate for (e.g. do certain ethnic groups have a smaller mean residual than others?) then creating this variable is very useful.

Now we have run our multiple regression with all of the explanatory variables let's have a look at how to report the results...

### 3.15 Reporting your results

As we said before, you should check the style guide for your university or target audience before writing up and avoid cutting and pasting SPSS output into your report! That said, we could report our multiple regression in the following way:

A multiple linear regression was carried out to ascertain the extent to which age 11 test scores, socio-economic class, gender and ethnicity can predict age 14 test scores. The regression model predicted 79.7% of the variance. The model was suitable for predicting the outcome ( $F = 3198.1$ ,  $df = 17$ ,  $p < .000$ ). The coefficients for the explanatory variables are tabulated below:

	B	SE	t	Sig.
Constant	-1.62	.166	-9.9	.000
Age 11 (KS2) score	.842	.004	210.5	.000
Gender (Girls versus boys)	.566	.075	7.4	.000
Mixed heritage	-.027	.170	-.16	.873
Indian	1.41	.150	9.4	.000
Pakistani	.37	.161	2.3	.021
Bangladeshi	.27	.185	1.4	.153
Black Caribbean	-.93	.199	-4.7	.000
Black African	.57	.221	2.6	.010
Any other ethnic group	1.12	.201	5.6	.000
<i>White British (reference)</i>	0			
Higher managerial & professional	3.48	.203	17.2	.000
Lower managerial & professional	2.39	.180	13.3	.000
Intermediate occupations	1.49	.218	6.9	.000
Small employers and own account workers	2.01	.191	10.5	.000
Lower supervisory & technical occupations	.95	.196	4.9	.000
Semi-routine occupations	.59	.192	3.1	.002
Routine occupations	.34	.196	1.7	.085
SEC missing data	1.07	.175	6.1	.000
<i>Never worked/Unemployed (reference)</i>	0			


Age 11 score was the strongest predictor of age 14 score. However gender, ethnicity and socio-economic class still accounted for a statistically significant amount of the

variance. Students from higher managerial and professional homes obtained on average a score 3.5 points higher than the lowest social class (long-term unemployed) even when prior attainment at age 11, gender and ethnicity were controlled. In relation to ethnicity, Black Caribbean students scored approximately 1.0 point lower at age 14, and Indian students scored 1.4 points higher, than White British students. Both results were highly statistically significant ( $p < .000$ ). Gender differences were also statistically significant with girls scoring 0.6 of a mark higher than boys, again after control for prior attainment, ethnicity and social class.

We think you should take our quiz and work through our exercises before moving on to the next module. Go on – it will be FUN! Well, sort of...

## Exercise

The following questions are slightly different in style to the ones you encountered in the previous module. We asked you to run a full analysis last time but we are now dealing with far more complex models and a single worked example would be way too big! Instead we have broken the process down into smaller questions which are easier to digest and more pleasant for your statistical palette. Don't worry; we will still be testing you! You will still need to perform a full multiple linear regression analysis. Note that you will also need to use the skills you learnt in previous modules.

Use the LSYPE 15,000 dataset  to work through each question. As before we recommend that you answer them in full sentences with supporting tables or graphs where appropriate as this will help when you come to report your own research. The answers are on the next page.

**Note:** *The variable names as they appear in SPSS dataset are listed in brackets.*

### Question 1

There is a variable in the LSYPE data (*singlepar*) which indicates whether the pupil lives in a single parent family (value=1) or not (value=0). What percentage of pupils in the sample live in single parent families (*singlepar*)?

*Use Frequencies to answer this question.*

### Question 2

Does the percentage of pupils with single parents (*singlepar*) vary across different ethnic groups (*ethnicity*) and is the association statistically significant?

*Use chi-square for this analysis.*

### Question 3

Is living in a single parent family (*singlepar*) related to educational attainment at age 14 (*ks3stand*), our outcome variable? Graphically display the relationship.

*Use a bar chart.*

#### Question 4

Is the relationship between age 14 attainment (*ks3stand*) and single parent family (*singlepar*) statistically significant if you add *singlepar* to model 7 as an explanatory variable? What is the importance of the single parent family variable relative to the other explanatory variables in the model?

*Run model 7 (Page 3.13) adding 'single parent family' (singlepar) as an explanatory variable.*

#### Question 5

Does adding the single parent family variable improve the predictive power of model 7 by a statistically significant increment?

*Calculate  $R^2$  change using two blocks for your regression analysis.*

#### Question 6

Does adding the single parent family (*singlepar*) variable cause any issues for the assumption of homoscedasticity of variance?

*Check the scatterplot of predicted score and standardized residual.*

## Answers

### Question 1

This question can be answered by creating a frequency table (head to the **Foundation Module** if you have forgotten how to do this).

Single parent household					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	11682	74.1	74.7	74.7
	yes	3950	25.0	25.3	100.0
	Total	15632	99.1	100.0	
Missing	missing	138	.9		
Total		15770	100.0		

As you can see just over 25% of the students in the sample come from a single parent home.

### Question 2

This question requires a crosstabulation with chi-square analysis. If you are rusty on this, head over to **Page 2.2**.

Ethnic group * Single parent household Crosstabulation					
			Single parent household		Total
			no	yes	
Ethnic group	White British	Count	7709	2329	10038
		% within Ethnic group	76.8%	23.2%	100.0%
	Mixed heritage	Count	437	353	790
		% within Ethnic group	55.3%	44.7%	100.0%
	Indian	Count	875	129	1004
		% within Ethnic group	87.2%	12.8%	100.0%
	Pakistani	Count	772	155	927
		% within Ethnic group	83.3%	16.7%	100.0%
	Bangladeshi	Count	607	106	713
		% within Ethnic group	85.1%	14.9%	100.0%
	Black Caribbean	Count	251	315	566
		% within Ethnic group	44.3%	55.7%	100.0%
	Black African	Count	329	279	608
		% within Ethnic group	54.1%	45.9%	100.0%
	Any other group	Count	477	166	643
		% within Ethnic group	74.2%	25.8%	100.0%
	Total	Count	11457	3832	15289
		% within Ethnic group	74.9%	25.1%	100.0%



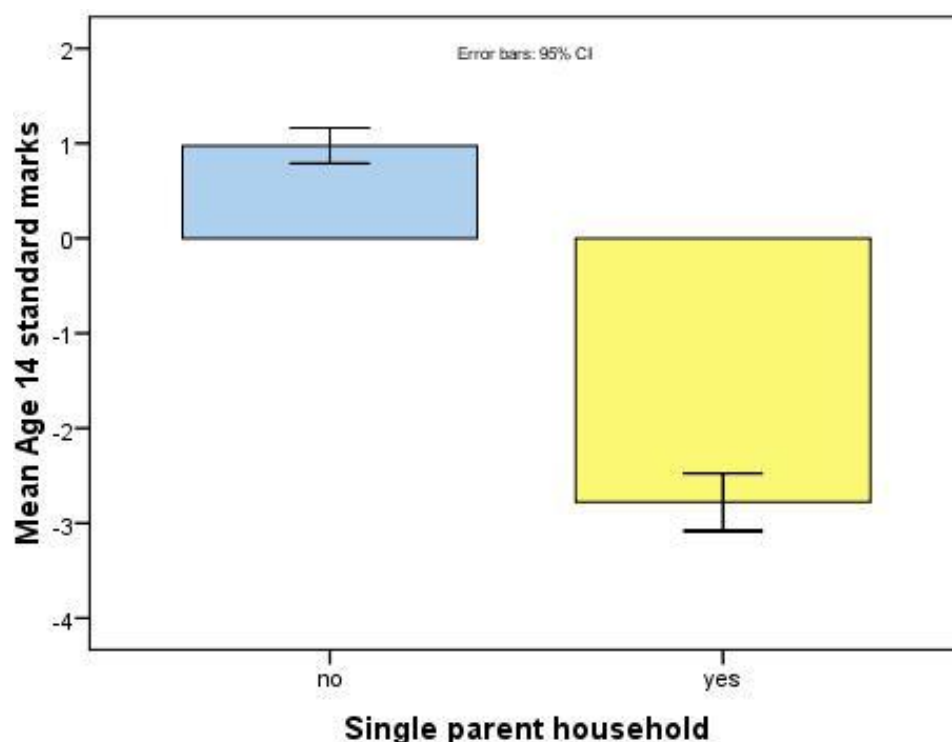
The table shows that the percentage of those from a single parent family does indeed vary between ethnic groups. For example, about 23% of the students from White British backgrounds are from single parent households compared to nearly 56% of those from Black Caribbean backgrounds. We can test the statistical significance of this association using chi-square.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	756.597 <sup>a</sup>	7	.000
Likelihood Ratio	697.905	7	.000
Linear-by-Linear Association	86.367	1	.000
N of Valid Cases	15289		

As the test shows, the chi-square value of 756.6 is statistically significant ( $p < .005$ ) so it is unlikely that an association of this strength could have occurred in our sample if there was no such association in the overall population.

### Question 3

A bar chart which uses the mean of the age 14 exam scores (*ks2stand*) on the y-axis is best for answering this question. If you can't quite recall how to do this the process is described in the **Foundation Module**.



As you can see, the mean age 14 score for students from a single parent families is substantially lower than average while those from backgrounds with two parents score slightly higher than average.

#### Question 4

You will need to re-run model 7 (on **Page 3.13**) but add the single parent family (*singlepar*) variable as a predictor. The basic procedures for running a regression module start on **Page 3.4**. The table required for answering this question is the *coefficients* table:

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1.126	.174		-6.461	.000
Gender	.555	.075	.028	7.407	.000
Mixed heritage	.125	.172	.003	.728	.467
Indian	1.289	.151	.033	8.558	.000
Pakistani	.239	.162	.006	1.475	.140
Bangladeshi	.079	.187	.002	.422	.673
Black Caribbean	-.695	.202	-.013	-3.432	.001
Black African	.656	.221	.012	2.970	.003
Any other ethnic group	1.137	.202	.022	5.630	.000
SEC missing	.817	.178	.032	4.586	.000
Higher Managerial and professional occupations	3.063	.207	.092	14.784	.000
Lower managerial and professional occupations	2.071	.183	.085	11.322	.000
Intermediate occupations	1.348	.218	.033	6.195	.000
Small employers and own account workers	1.624	.195	.052	8.324	.000
Lower supervisory and technical occupations	.594	.199	.018	2.981	.003
Semi-routine occupations	.409	.192	.013	2.125	.034
Routine occupations	.079	.198	.002	.401	.689
Age 11 standard marks	.839	.004	.852	2.08E2	.000
Single parent household	-.854	.093	-.038	-9.165	.000

We have highlighted the single parent family variable. The columns marked *t* and *sig* test tell us that the variable is contributing to the model to a statistically significant degree ( $p < .005$ ). The B-coefficient in the first column suggests that, even after all the other variables in the model are held constant, those students from single parent families score an average of -.854 less standard marks at age 14 than their peers from families with two parents. Though this is significant, the *Beta* column puts this in

perspective by providing a standardized coefficient for all variables. The Beta value of -.038 is much smaller than the one for age 11 exam score (.852) which shows that prior attainment is a more powerful predictor of exam score by far.

### Question 5

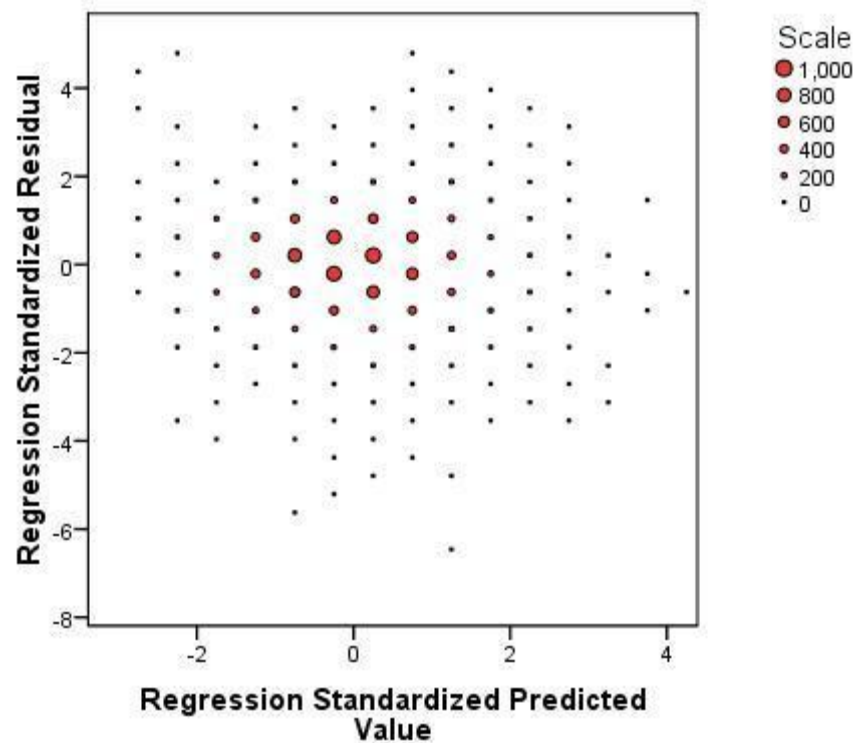
Answering this question requires you to break the model down in to two blocks, with the first block being the original model 7 and the second block being model 7 plus the single parent family variable. You will also need to request [glossary term='R Squared' page='/fac/soc/wie/research-new/srme/glossary']R-square[/glossary] change statistics from SPSS. The process for doing both of these things is explained at the end of **Page 3.11**. The *Model Summary* is shown below, complete with R-Square Change statistics.

Model Summary <sup>c</sup>								
Model	R	R Square	Adjusted R Square	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.893 <sup>a</sup>	.797	.797	.797	3166.007	17	13717	.000
2	.893 <sup>b</sup>	.798	.798	.001	83.992	1	13716	.000

The highlighted *R square Change* column for 'model 2' (where *singlepar* was added) shows that  $r^2$  only increases by .001 compared to the original model 7 (labelled '1' here – just to confuse you!). This means that only an additional 0.1% of the variance in age 14 exam score was explained by the new model. This is a small amount but that does not mean it is not a significant amount. The *Sig. F Change* indicates that the enhanced model ('2') is better at predicting the outcome to its predecessor ('1') to statistically significant level ( $p < .005$ ).

### Question 6

To check this assumption you will need to examine a scatterplot which has the standardized predicted values for each participant on the x-axis and the standardized residual for each participant on the y-axis. This is achieved using the *Plots* submenu on the right hand side of the main regression window. **Pages 3.3 and 3.14** discuss assumptions and how to test them if you are unsure how to do this.



The plot looks relatively unchanged from the one we saw when running diagnostics on model 7 (**Page 3.14**). The points are spread out in a fairly random manner which suggests that our assumption of homoscedasticity is likely to be safe!