

Module 2 - Simple Linear Regression

OBJECTIVES

1. Know how to graph and explore associations in data
2. Understand the basis of statistical summaries of association (e.g. variance, covariance, Pearson's r)
3. Be able to calculate correlations and simple linear regressions using SPSS/PASW
4. Understand the formula for the regression line and why it is useful
5. See the application of these techniques in education research (and perhaps your own research!)

Start Module 2: Simple Linear Regression

Get started with the basics of regression analysis.

There are multiple pages to this module that you can access individually by 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, extension materials and the quiz to get a firm understanding.

CONTENTS

2.1 Overview

2.2 Association

2.3 Correlation

2.4 Correlation Coefficients

2.5 Simple Linear Regression

2.6 Assumptions

2.7 Using SPSS for Simple Linear Regression part 1 - running the analysis

2.8 Using SPSS for Simple Linear Regression part 2 - interpreting the output

Quiz (Online only)

Exercise

2.1 Overview

What is simple linear regression?

Regression analysis refers to a set of techniques for predicting an outcome variable using one or more explanatory variables. It is essentially about creating a model for estimating one variable based on the values of others. Simple linear regression is regression analysis in its most basic form - it is used to predict a continuous (scale) outcome variable from *one* continuous explanatory variable. Simple Linear regression can be conceived as the process of drawing a line to represent an association between two variables on a scatterplot and using that line as a linear model for predicting the value of one variable (outcome) from the value of the other (explanatory variable). Don't worry if this is somewhat baffling at this stage, it will become much clearer later when we start displaying bivariate data (data about two variables) using scatterplots!

Correlation is excellent for showing association between two variables. Simple Linear regression takes correlation's ability to show the strength and direction of an association a step further by allowing the researcher to use the pattern of previously collected data to build a predictive model. Here are some examples of how this can be applied:

- Does time spent revising influence the likelihood of obtaining a good exam score?
- Are some schools more effective than others?
- Does changing school have an impact on a pupil's educational progress?

It is important to point out that there are limitations to regression. We can't always use it to analyse association. We'll start this module by looking at association more generally.

Running through the examples and exercises using SPSS

We're keen on training you with real world data so that you may be able to apply regression analysis to your own research. For this reason all of the examples we use come from the LSYPE and we provide an adapted version of the LYSPE dataset for you to practice with and to test your new found skills from. We recommend that you run through the examples we provide so that you can get a feel for the techniques and for SPSS/PASW in preparation for tackling the exercises.

2.2 Association

How do we measure association? Correlation and Chi-Square

It is useful to explore the concepts of association and correlation at this stage as it will hold us in good stead when we start to tackle regression in greater detail. Correlation basically refers to statistically exploring whether the values of one variable increase or decrease systematically with the values of another. For example, you might find an association between IQ test score and exam grades such that if individuals have high IQ scores they also get good exam grades while individuals who get low scores on the IQ test do poorly in their exams.

This is very useful but association cannot always be ascertained using correlation. What if there are only a few values or categories that a variable can take? For example, can gender be correlated with school type? There are only a few categories in each of these variables (e.g. male, female). Variables that are sorted into discrete categories such as these are known in SPSS/PASW as nominal variables (see our page on types of data in the prologue). When researchers want to see if two nominal variables are associated with each other they can't use correlation but they can use a crosstabulation (crosstab). The general principle of crosstabs is that the proportion of *actual* observations in each category may differ from what would be *expected* to be observed by chance in the sample (if there were no association in the population as a whole). Let's look at an example using SPSS/PASW output based on LSYPE data (**Figure 2.2.1**):

Figure 2.2.1: Crosstabulation of gender and exclusion rate

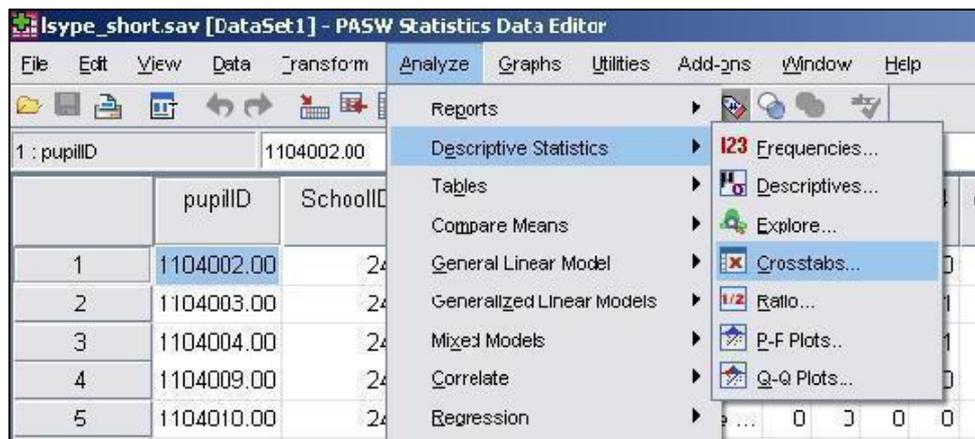
			exclude		Total
			No	Yes	
Gender	Male	Count	6037	1039	7076
		% within Gender	85.3%	14.7%	100.0%
	Female	Count	6352	397	6749
		% within Gender	94.1%	5.9%	100.0%
Total	Count	12389	1436	13825	
	% within Gender	89.6%	10.4%	100.0%	

This table shows how many males and females in the LSYPE sample were temporarily excluded in the three years before the data was collected. If there was no association between gender and exclusion you would expect the proportion of males excluded to be the same or very close to the proportion of females excluded. The *% within gender* for each row (each gender) displays the percentage of individuals within each cell. It appears that there is an association between gender and exclusion - 14.7% of males have been temporarily excluded compared to 5.9% of females. Though there appears to be an association we must be careful. There is bound to be some difference between males and females and we need to be sure that this difference is statistically improbable (if there was really no association) before we can say it reflects an underlying phenomenon. This is where chi-square comes in! Chi-square is a test that allows you to check whether or not an association is statistically significant.

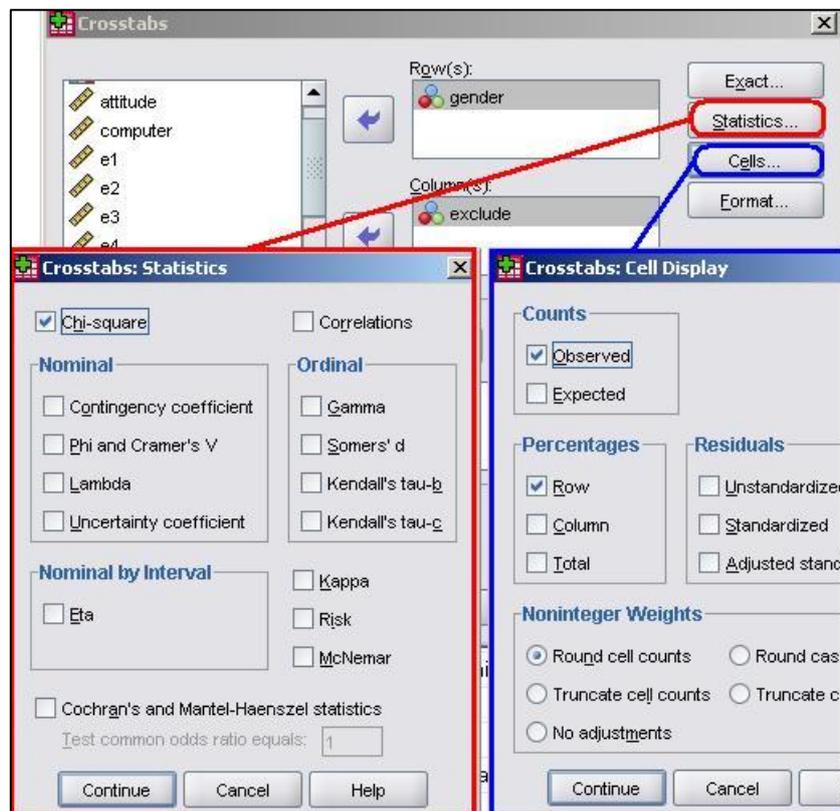
How to perform a Chi-square test using SPSS/PASW

Why not follow us through this process using the LSYPE 15,000 dataset. We also have a video demonstration.

To draw up a crosstabulation (sometimes called a contingency table) and perform a chi-square analysis, take the following route through the SPSS data editor: **Analyze > Descriptive Statistics > Crosstabs**



Now move the variables you wish to explore into the columns/rows boxes using the list on the left and the arrow keys. Alternatively you can just drag and drop! In order to perform the chi-square test click on the **Statistics** option (shown below) and check the *chi-square box*. There are many options here that provide alternative tests for association (including those for ordinal variables) but don't worry about those for now.



Click on continue when you're done to close the statistics pop-up. Next click on the **Cells** option and under *percentages* check the *rows* box. Click on *continue* to close the pop up and, when you're ready click **OK** to let SPSS weave its magic...

Interpreting the output

You will find that SPSS will give you an occasionally overwhelming amount of output but don't worry - we need only focus on the key bits of information.

Figure 2.2.2: SPSS output for chi-square analysis

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * exclude	13825	87.7%	1945	12.3%	15770	100.0%

Gender * exclude Crosstabulation

			exclude		Total
			No	Yes	
Gender	Male	Count	6037	1039	7076
		% within Gender	85.3%	14.7%	100.0%
	female	Count	6352	397	6749
		% within Gender	94.1%	5.9%	100.0%
Total		Count	12389	1436	13825
		% within Gender	89.6%	10.4%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	287.458 ^a	1	.000		
Continuity Correction ^b	286.513	1	.000		
Likelihood Ratio	297.717	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	287.437	1	.000		
N of Valid Cases	13825				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 701.02.

b. Computed only for a 2x2 table

The first table of **Figure 2.2.2** tells us how many cases were included in our analysis. Just below 88% of the participants were included. It is important to note that over a tenth of our participants have no data, but 88% is still a very large group - 13,825 young people! It is very common for there to be missing cases as often participants will miss out questions on the survey or give incomplete answers. Missing values are not included in the analysis. However it is important to question why certain cases may be missing when performing any statistical analysis - find out more in our missing values section.

The second table is basically a reproduction of the one at the top of this page (**Figure 2.2.1**). The part outlined in red tells us that males were disproportionately likely to have been excluded in the past three years compared to females (14.7% of males were excluded compared to 5.9% of females). Chi-square is less accurate if there are not at least five individuals expected in each cell. This is not a problem in this example but it is always worth checking, particularly when your research involves a smaller sample size, more variables or more categories within each variable (e.g. variables such as social class or school type).

The third table shows that the difference between males and females is statistically significant using the Pearson Chi-square (as marked in red). The Chi-square *value* of 287.5 is definitely significant at the $p < .05$ level (see *Asymp. Sig.* column). In fact the value is .000 which means that $p < .0005$. In other words the probability of getting a difference of this size between the observed and expected values purely by chance (if in fact there was no association between the variables) is less than a 0.05% or 1 in 2,000! We can therefore be confident that there is a real difference between the exclusion rates of males and females.

Choosing an approach to analysing association

Chi-square provides a good method for examining association in nominal (and sometimes ordinal) data but it cannot be used when data is continuous. For example, what if you were recording the number of pupils per school as a variable? Thousands of categories would be required! One option would be to create categories for such continuous data (e.g. 1-500, 501-1000, etc.) but this creates two difficult issues: How do you decide what constitutes a category and to what extent is the data oversimplified by such an approach? Generally where continuous data is being used a statistical correlation is a preferable approach for exploring association. Correlation is a good basis for learning regression and will be our next topic.

2.3 Correlation

Visually identifying association - Scatterplots

Scatterplots are the best way to visualize correlation between two continuous (scale) variables, so let us start by looking at a few basic examples. The graphs below illustrate a variety of different bivariate relationships, with the horizontal axis (x-axis) representing one variable and the vertical axis (y-axis) the other. Each point represents one individual and is dictated by their score on each variable. Note that these examples are fabricated for the purpose of our explanation - real life is rarely this neat! Let us look at each in turn:

Figure 2.3.1: Displays three scatterplots overlaid on one another and represents maximum or 'perfect' negative and positive correlations (blue and green respectively). The points displayed in red provide us with a cautionary tale! It is clear there is a relationship between the two as the points display a clear arching pattern. However they are not statistically correlated as the relationship is not consistent. For values of X up to about 75 the relationship with Y is positive but for values of 100 or more the relationship is negative!

Figure 2.3.1: Examples of 'perfect' relationships

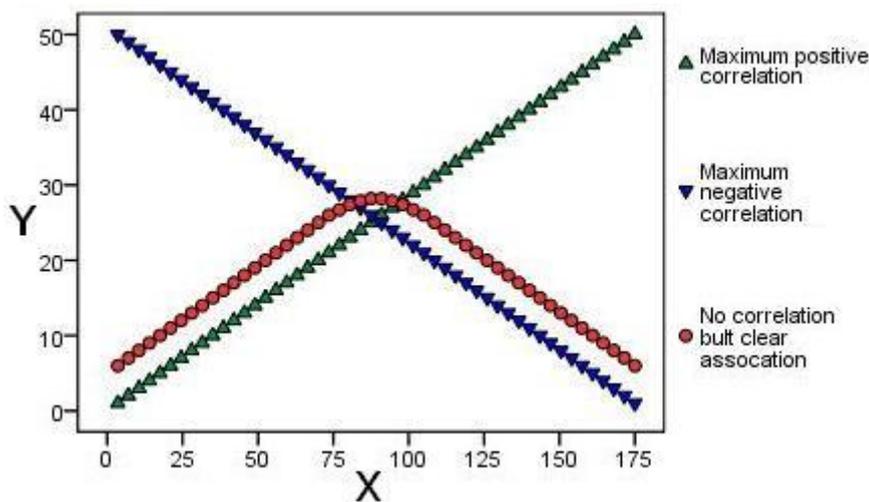


Figure 2.3.2: These two scatterplots look a little bit more like they may represent real life data. The green points show a strong positive correlation as there is a clear relationship between the variables - if a participant's score on one variable is high their score on the other is also high. The red points show a strong negative relationship. A participant with a high score on one variable has a low score on the other.

Figure 2.3.2: Examples of strong correlations

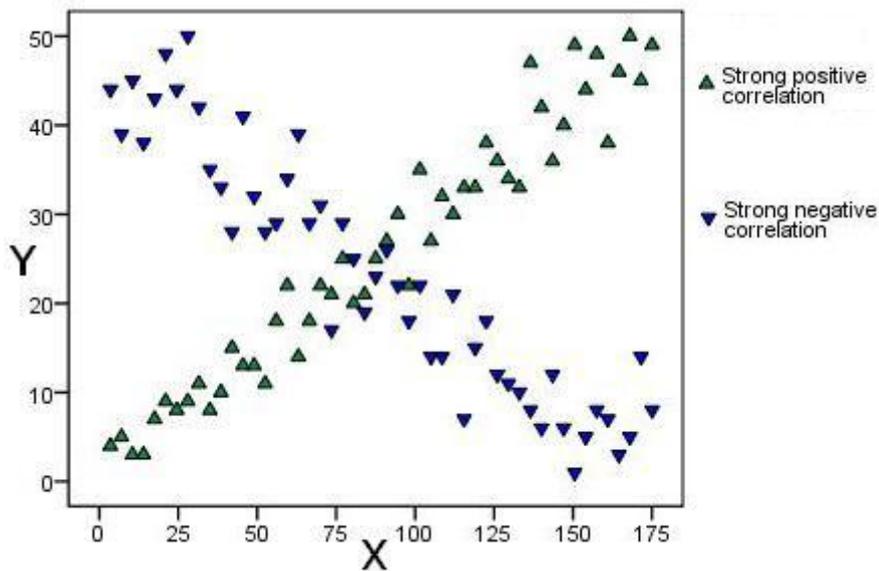
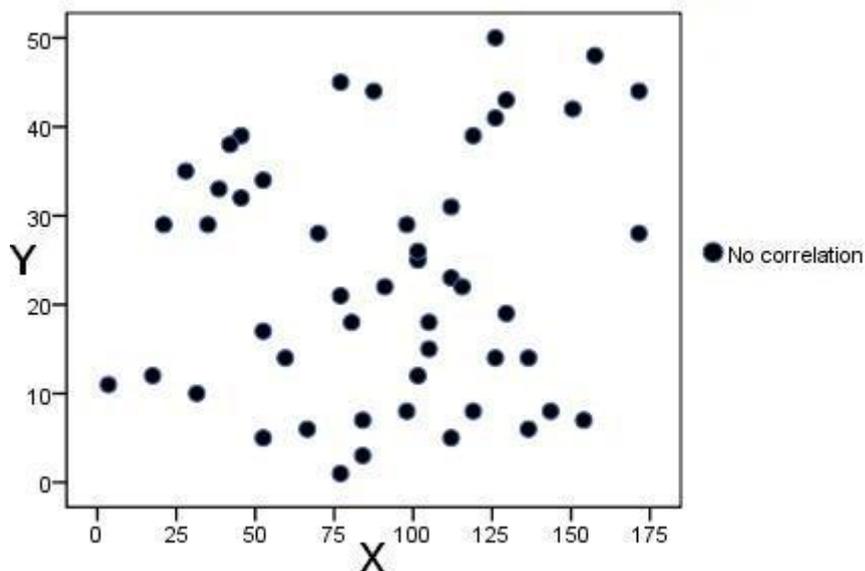


Figure 2.3.3: This final scatterplot demonstrates a situation where there is no apparent correlation. There is no discernible relationship between an individual's score on one variable and their score on another.

Figure 2.3.3: Example of no clear relationship



Several points are evident from these scatterplots.

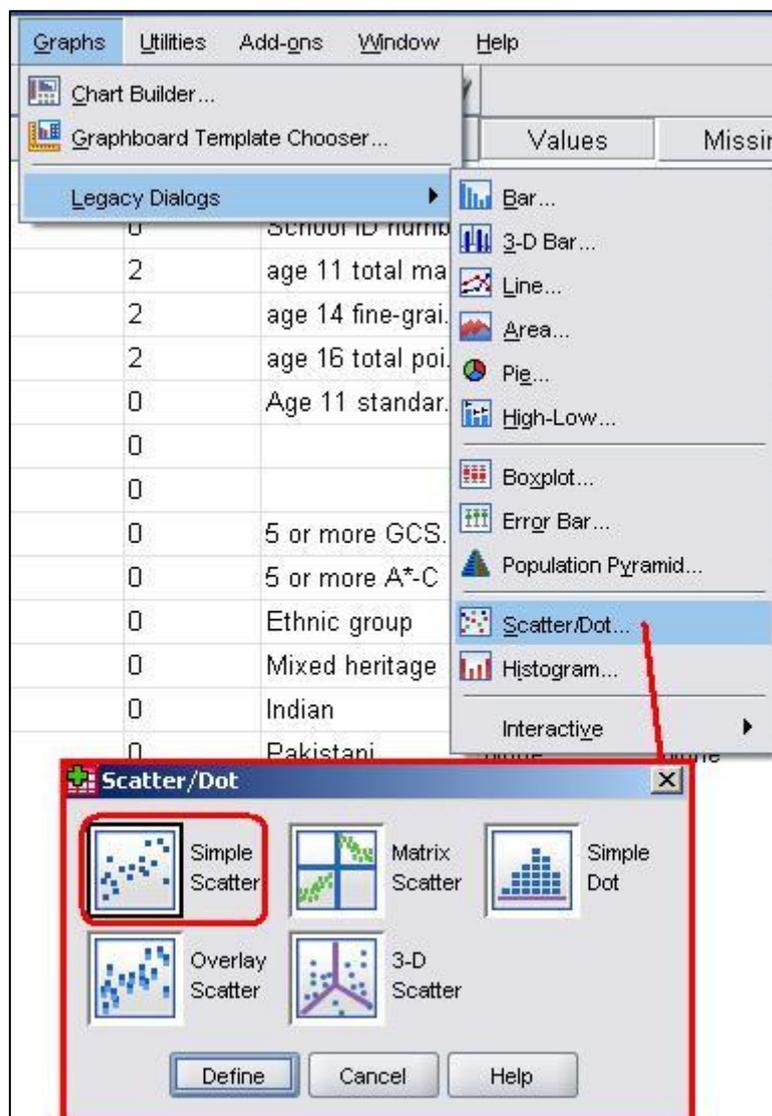
- When one variable increases when the other increases the correlation is positive; and when one variable increases when the other decreases the correlation is negative.
- The strongest correlations occur when data points fall exactly in a straight line (**Figure 2.3.1**).
- The correlation becomes weaker as the data points become more scattered. If the data points fall in a random pattern there is no correlation (**Figure 2.3.3**).

- Only relationships where one variable increases or decreases systematically with the other can be said to demonstrate correlation - at least for the purposes of regression analysis!

The next page will talk about how we can statistically represent the strength and direction of correlation but first we should run through how to produce scatterplots using SPSS/PASW.

How to produce a Scatterplot using SPSS

Let's use the LSYPE 15,000 dataset to explore the relationship between Key Stage 2 exam scores (exams students take at age 11) and Key Stage 3 exam scores (exams students take at age 14). Take the following route through SPSS: **Graphs > Legacy Dialogs > Scatter/Dot** (shown below). A small box will pop up giving you the option of a few different types of scatterplot. Feel free to explore these in more depth if you like (*overlay* was used to produce the scatterplots above) but in this case we're going to choose *Simple/Scatter*. Click **Define** when you're happy with your selection.



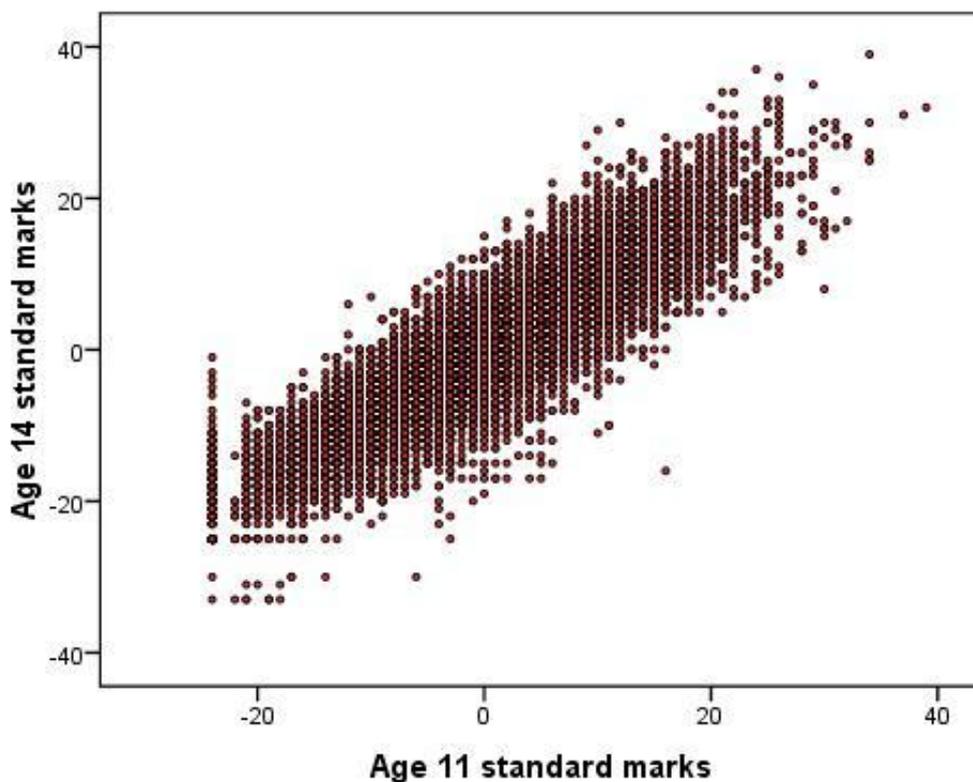
A pop-up will now appear asking you to define your scatterplot. All this means is you need to decide which variable will be represented on which axis by dragging the variables from the

list on the left into the relevant boxes (or transferring them using the arrows). We have put the age 14 assessment score (ks3stand) on the vertical y-axis and age 11 scores (ks2stand) on the horizontal x-axis. SPSS/PASW allows you to label or mark individual cases (data points) by a third variable which is a useful feature but we will not need it this time. When you are ready click on **OK**.



Et Voila! The scatterplot (**Figure 2.3.4**) shows that there is a relationship between the two variables. Higher scores at age 11(ks2) are related to higher scores at age 14 (ks3) while lower age 11 scores are related to lower age 14 scores - there is a positive correlation between the variables.

Figure 2.3.4: Scatterplot of ks2 and ks3 exam scores



Note that there is what is called a 'floor effect' where no age 11 scores are below approximately '-25'. They form a straight vertical line in the scatterplot. We discuss this in more detail on our extension page about transforming data. This is important but for now you may prefer to stay focussed on the general principles of simple linear regression.

Now that we know how to generate scatterplots for displaying relationships visually it is time to learn how to understand them statistically.

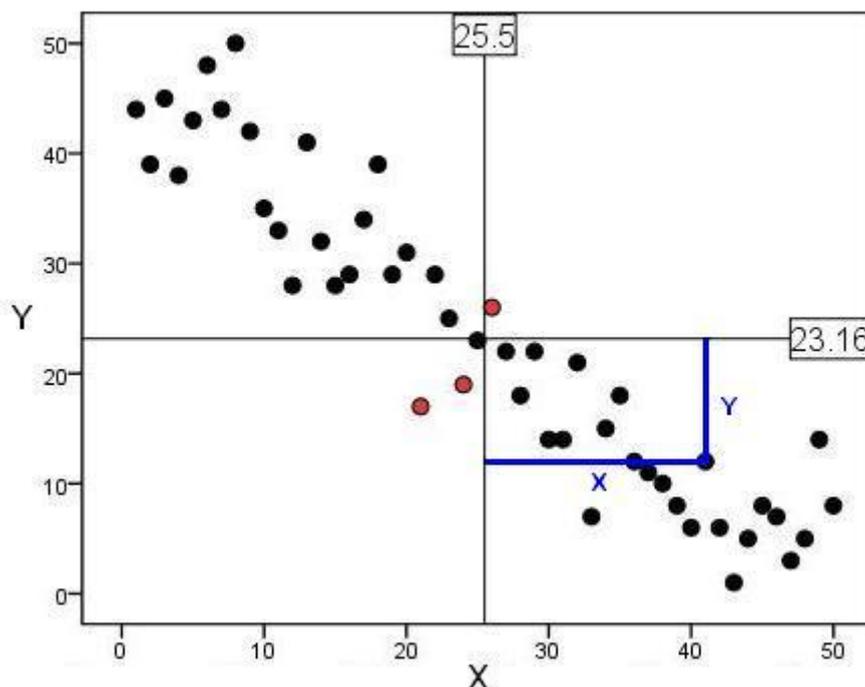
2.4 Correlation Coefficients

Pearson's r

Graphing your data is essential to understanding what it is telling you. Never rush the statistics, get to know your data first! You should always examine a scatterplot of your data. However it is useful to have a numeric measure of how strong the relationship is. The Pearson r correlation coefficient is a way of describing the strength of an association in a simple numeric form.

We're not going to blind you with formulae but it is helpful to have some grasp of how the stats work. The basic principle is to measure how strongly two variables relate to each other, that is to what extent do they *covary*. We can calculate the covariance for each participant (or case/observation) by multiplying how far they are above or below the mean for variable X by how far they are above or below the mean for variable Y. The blue lines coming from the case on the scatterplot below (**Figure 2.4.1**) will hopefully help you to visualize this.

Figure 2.4.1: Scatterplot to demonstrate the calculation of covariance



The black lines across the middle represent the mean value of X (25.50) and the mean value of Y (23.16). These lines are the reference for the calculation of covariance for all of the participants. Notice how the point highlighted by blue lines is above the mean for one variable but below the mean for the other. A score below the mean creates a negative difference (approximately $10 - 23.16 = -13.6$) while a score above the mean is positive (approximately $41 - 25.5 = 15.5$). If an observation is above the mean on X and also above the mean on Y than the product (multiplying the differences together) will be positive. The product will also be positive if the observation is below the mean for X and below the mean for Y. The product will be negative if the observation is above the mean for X and below the mean for Y or vice versa. Only the three points highlighted in red produce positive products

in this example. All of the individual products are then summed to get a total and this is divided by the product of the standard deviations of both variables in order to scale it (don't worry too much about this!). This is the correlation coefficient, Pearson's r .

What does Pearson's r tell us?

The correlation coefficient tells us two key things about the association:

- **Direction** - A positive correlation tells us that as one variable gets bigger the other tends to get bigger. A negative correlation means that if one variable gets bigger the other tends to get smaller (e.g. as a student's level of economic deprivation decreases their academic performance increases).
- **Strength** - The weakest linear relationship is indicated by a correlation coefficient equal to 0 (actually this represents no correlation!). The strongest linear correlation is indicated by a correlation of -1 or 1. The strength of the relationship is indicated by the magnitude of the value regardless of the sign (+ or -), so a correlation of -0.6 is equally as strong as a correlation of +0.6. Only the direction of the relationship differs.

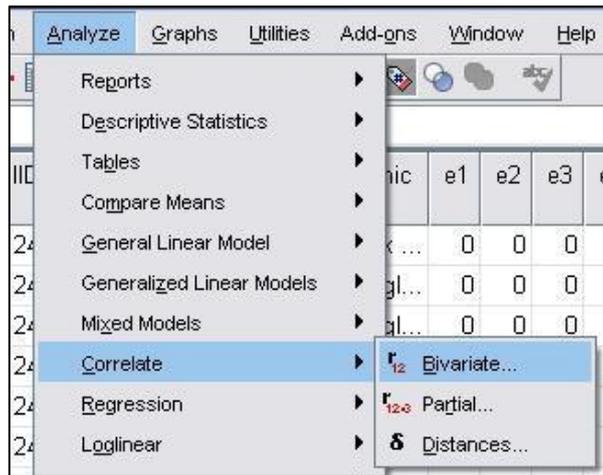
It is also important to use the data to find out:

- **Statistical significance** - We also want to know if the relationship is statistically significant. That is, what is the probability of finding a relationship like this in the sample, purely by chance, when there is no relationship in the population? If this probability is sufficiently low then the relationship is statistically significant.
- **How well the correlation describes the data** - This is best expressed by considering how much of the variance in the outcome can be explained by the explanatory variable. This is described as the proportion of variance explained, r^2 (sometimes called the coefficient of determination). Conveniently the r^2 can also be found just by squaring the Pearson correlation coefficient. The r^2 provides us with a good gauge of the substantive size of a relationship. For example, a correlation of 0.6 explains 36% ($0.6^2 = .036$) of the variance in the outcome variable.

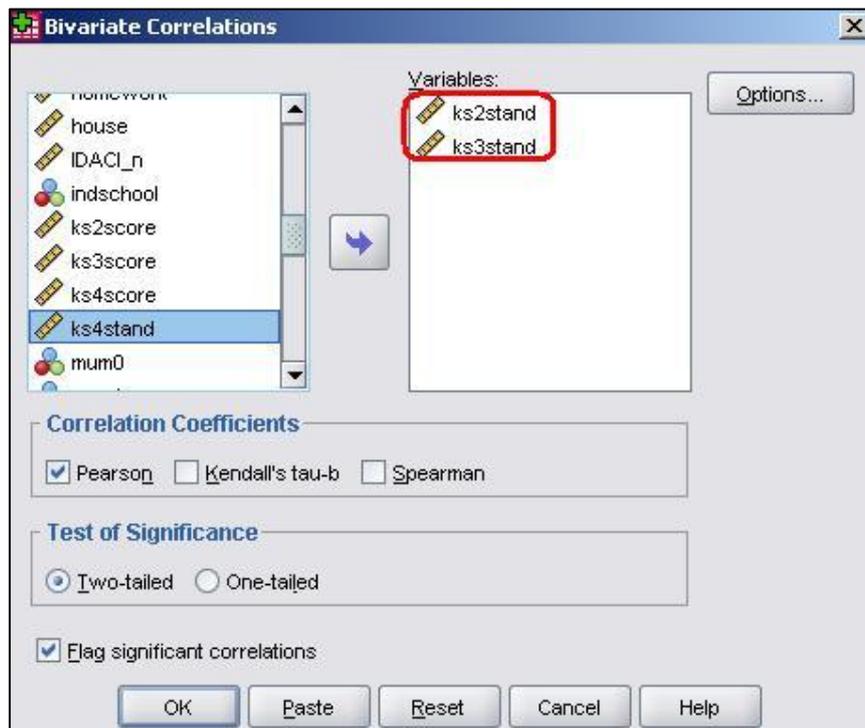
How to calculate Pearson's r using SPSS

Let us return to the example used on the previous page - the relationship between age 11 and age 14 exam scores. This time we will be able to produce a statistic which explains the strength and direction of the relationship we observed on our scatterplot. This example once again uses the LSYPE 15,000 dataset. Take the following route through SPSS:

Analyze > Correlate > Bivariate (shown below).



The pop-up menu (shown below) will appear. Move the two variables that you wish to examine (in this case ks2stand and ks3stand) across from the left hand list into the *variables* box. Note that continuous variables are preceded by a ruler icon - for Pearson's *r* all variables need to be continuous. The other options are fine so just click **OK**.



SPSS will provide you with the following output (**Figure 2.4.2**). This correlation table could contain multiple variables which is why it appears to give you the same information twice! We've circled in red the figure for Pearson's *r* and below that is the level of statistical significance. As inferred from the scatterplot on the previous page, there is a positive correlation between age 11 and age 14 exam performances such that a high score in one is associated with a high score in the other. The value of .886 is strong - it means that one variable accounts for about 79% of the variance in the other ($r^2 = .886 \times .886 = .785$). The significance level of .000 means that there is a less than .0005 possibility that this difference may have occurred purely due to sampling and so it is highly likely that the relationship exists in the population as a whole.

Figure 2.4.2: Correlation matrix for ks2 and ks3 exam scores

		Age 11 standard marks	Age 14 standard marks
Age 11 standard marks	Pearson Correlation	1	.886**
	Sig. (2-tailed)		.000
	N	14301	14138
Age 14 standard marks	Pearson Correlation	.886**	1
	Sig. (2-tailed)	.000	
	N	14138	14832

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

Correlation of ordinal variables - Spearman's rho

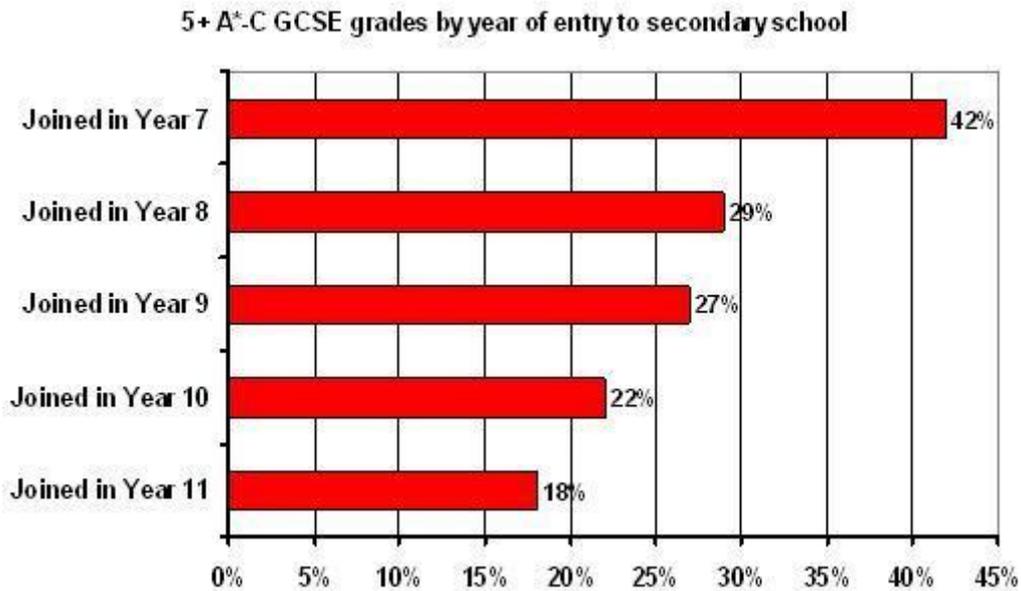
Pearson's r is a correlation coefficient for use when variables are continuous (scale). In cases where data is ordinal the Spearman Rho rank order correlation is more appropriate. Luckily this works in a similar way to Pearson's r when using SPSS/PASW and produces output which is interpreted in the same way. If your bivariate correlation involves ordinal variables perform the procedure exactly as you would for a Pearson's correlation (**Analyze > Correlate > Bivariate**) but be sure to check the *Spearman* box in the *correlation coefficients* section of the pop up box and deselect the *Pearson* option.

Correlation and causation

It is very important that correlation does not get confused with causation. All that a correlation shows us is that two variables are related, not that one necessarily causes the other. Consider the following example.

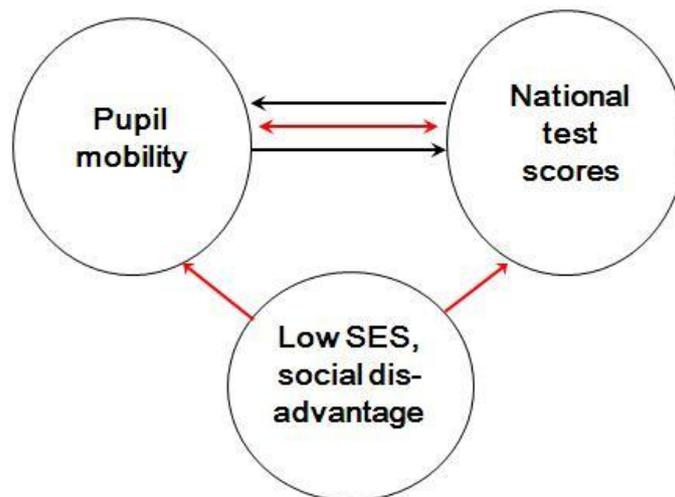
When looking at National test results, pupils who joined their school part of the way through a key stage tend to perform at the lower end of the attainment scale than those who attended school for the whole of the key stage. The figure below (**Figure 2.4.3**) shows the proportion of pupils who achieve 5 or more A*-C grades at GCSE (in one region of the UK). While 42% of those who were in the same school for the whole of the secondary phase achieved this only 18% of those who joined during year 11 did so.

Figure 2.4.3: Proportion of students with 5+ A*-C grades at GCSE by year they joined their school



It may appear that joining a school later leads to poorer exam attainment and the later you join the more attainment declines. However we cannot necessarily infer that there is a causal relationship here. It may be reverse causality, for example where pupils with low attainment and behaviour problems are excluded from school. Or it might be that the relationship between mobility and attainment arises because both are related to a third variable, such as socio-economic disadvantage.

Figure 2.4.4: Possible causal routes for the relationship between GCSE attainment and pupil mobility



This diagram (**Figure 2.4.4**) represents three variables but there may be hundreds involved in the relationship! We will begin to tackle issues like this more in the multiple linear regression module. For now let's turn our attention to simple linear regression.

2.5 Simple linear regression

Regression and correlation

If you are wrestling with all of this new terminology (which is very common if you're new to stats, so don't worry!) then here is some good news: Regression and correlation are actually fundamentally the same statistical procedure. The difference is essentially in what your purpose is - what you're trying to find out. In correlation we are generally looking at the strength of a relationship between two variables, X and Y, where in regression we are specifically concerned with how well we can *predict* Y from X.

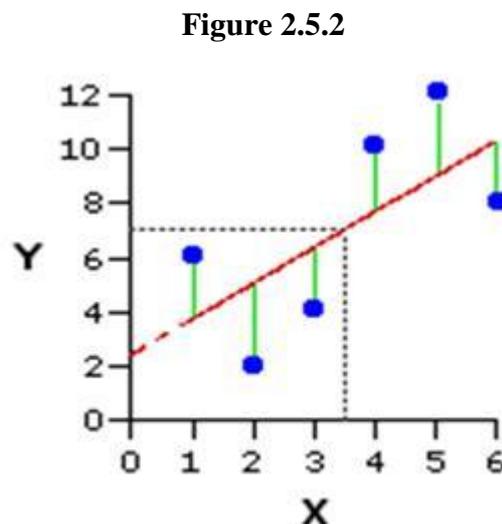
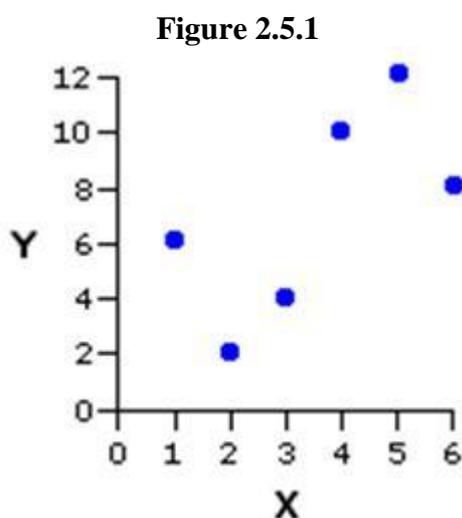
Examples of the use of regression in education research include defining and identifying under achievement or specific learning difficulties, for example by determining whether a pupil's reading attainment (Y) is at the level that would be predicted from an IQ test (X). Another example would be screening tests, perhaps to identify children 'at risk' of later educational failure so that they may receive additional support or be involved in 'early intervention' schemes.

Calculating the regression line

In regression it is convenient to define X as the explanatory variable (or independent) variable and Y as the outcome (or dependent) variable. We are concerned with determining how well X can predict Y.

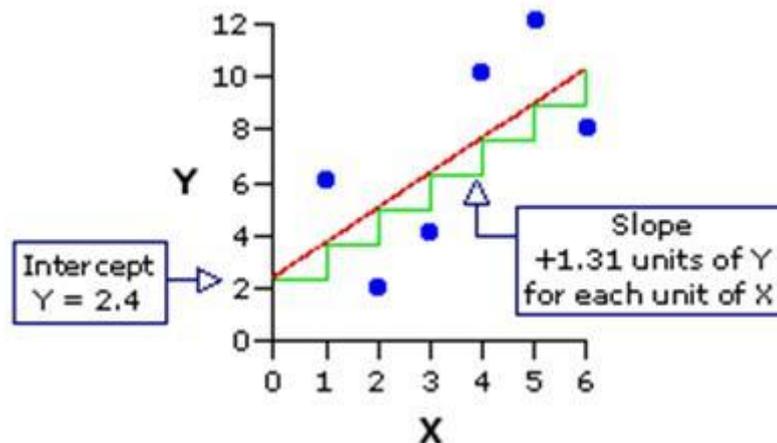
It is important to know which variable is the outcome (Y) and which is the explanatory (X)! This may sound obvious but in education research it is not always clear - for example does greater interest in reading predict better reading skills? Possibly. But it may be that having better reading skills encourages greater interest in reading. Education research is littered with such 'chicken and egg' arguments! Make sure that you know what *your* hypothesis about the relationship is when you perform a regression analysis as it is fundamental to your interpretation.

Let's try and visualize how we can make a prediction using a scatterplot:



- **Figure 2.5.1** plots five observations (XY pairs). We can summarise the linear relationship between X and Y best by drawing a line straight through the data points. This is called the regression line and is calculated so that it represents the relationship as accurately as possible.
- **Figure 2.5.2** shows this regression line. It is the line that minimises the differences between the actual Y values and the Y value that would be predicted from the line. These differences are squared so that negative signs are removed, hence the term 'sum of squares' which you may have come across before. You do not have to worry about how to calculate the regression line - SPSS/PASW does this for you!

Figure 2.5.3



- The line has the formula $Y = A + BX$, where **A** is the intercept (the point where the line meets the Y axis, where $X = 0$) and **B** is the slope (gradient) of the line (the amount Y increases for each unit increase in X), also called the regression coefficient. **Figure 2.5.3** shows that for this example the intercept (where the line meets the Y axis) is 2.4. The slope is 1.31, meaning for every unit increase in X (an increase of 1) the predicted value of Y increases by 1.31 (this value would have a negative sign if the correlation was negative).
- The regression line represents the predicted value of Y for each value of X. We can use it to generate a predicted value of Y for any given value of X using our formula, even if we don't have a specific data point that covers the value. From **Figure 2.5.3** we see that an X value of 3.5 predicts a Y value of about 7.
- Of course, the model is not perfect. The vertical distance from each data point to the regression line (see **Figure 2.5.2**) represents the error of prediction. These errors are called residuals. We can take the average of these errors to get a measure of the average amount that the regression equation over-predicts or under-predicts the Y values. The higher the correlation, the smaller these errors (residuals), and the more accurate the predictions are likely to be.

We will have a go at using SPSS/PASW to perform a linear regression soon but first we must consider some important assumptions that need to be met for simple linear regression to be performed.

2.6 Assumptions

The assumptions of linear regression

Simple linear regression is only appropriate when the following conditions are satisfied:

- **Linear relationship:** The outcome variable Y has a roughly linear relationship with the explanatory variable X.
- **Homoscedasticity:** For each value of X, the distribution of residuals has the same variance. This means that the level of error in the model is roughly the same regardless of the value of the explanatory variable (homoscedasticity - another disturbingly complicated word for something less confusing than it sounds).
- **Independent errors:** This means that residuals (errors) should be uncorrelated.

It may seem as if we're complicating matters but checking that the analysis you perform is meeting these assumptions is vital to ensuring that you draw valid conclusions.

Other important things to consider

The following issues are not as important as the assumptions because the regression analysis can still work even if there are problems in these areas. However it is still vital that you check for these potential issues as they can seriously mislead your analysis and conclusions.

- **Problems with outliers/influential cases:** It is important to look out for cases which may unduly influence your regression model by differing substantially to the rest of your data.
- **Normally distributed residuals:** The residuals (errors in prediction) should be normally distributed.

Let us look at these assumptions and related issues in more detail - they make more sense when viewed in the context of how you go about checking them.

Checking the assumptions

The below points form an important checklist:

1. First of all remember that for linear regression the outcome variable must be continuous. You'll need to think about Logistic and/or Ordinal regression if your data is categorical (fear not - we have modules dedicated to these).
2. **Linear relationship:** There must be a roughly linear relationship between the explanatory variable and the outcome. Inspect your scatterplot(s) to check that this assumption is met.
3. You may run into problems if there is a restriction of range in either the outcome variables or the explanatory variable. This is hard to understand at first so let us look at an example. **Figure 2.6.1** will be familiar to you as we've used it previously - it shows the relationship between age 11 and age 14 exam scores. **Figure 2.6.2** shows the same variables but includes only the most able pupils at age 11, specifically those who received one of the highest 10% of ks2 exam scores.

Figure 2.6.1: Scatterplot of age 11 and age 14 exam scores - full range

Correlation = .886 ($r^2 = .785$)

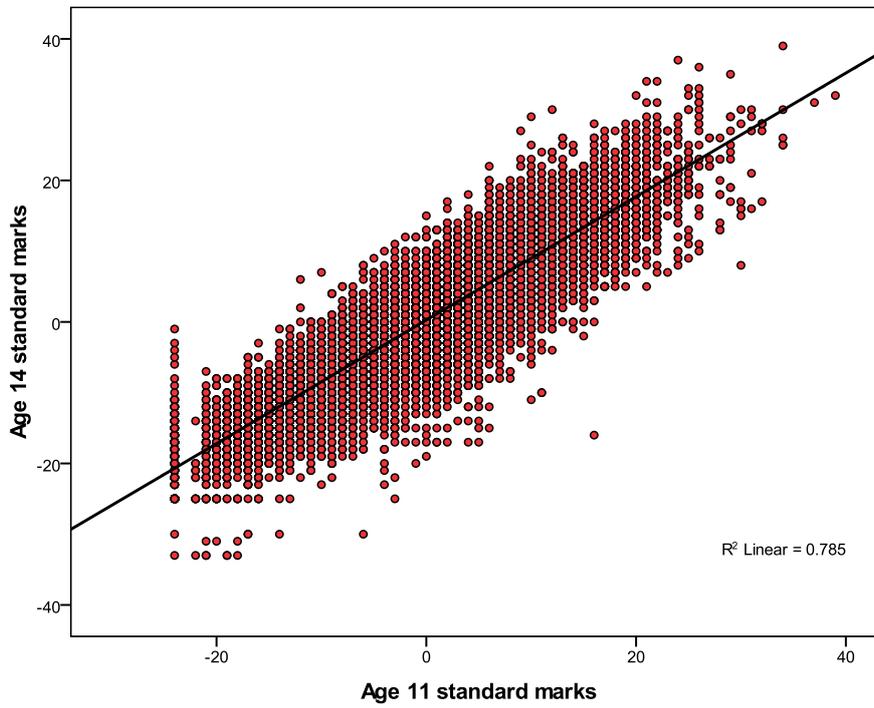
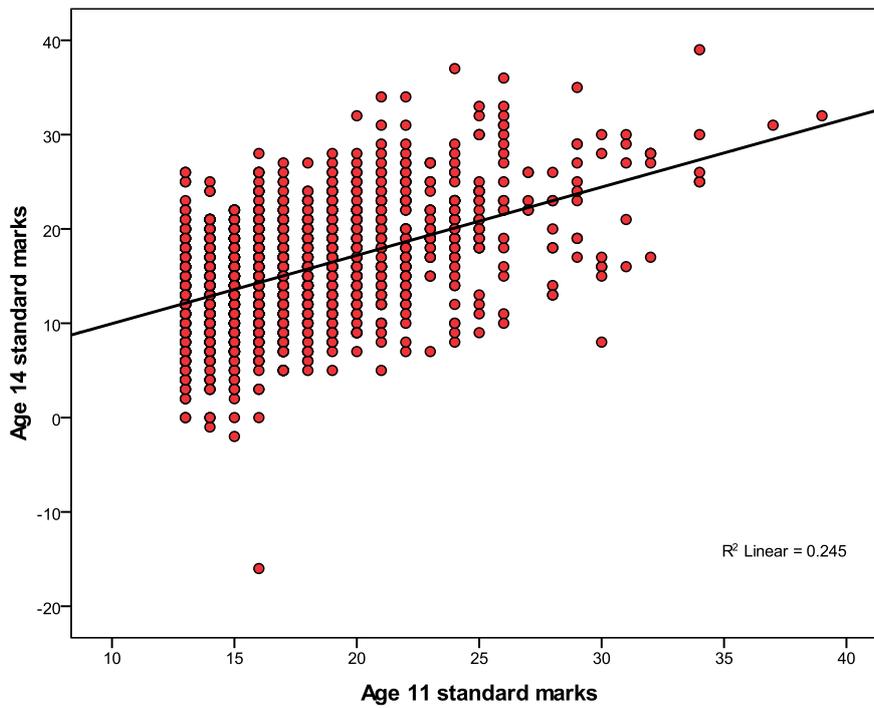


Figure 2.6.2: Scatterplot of age 11 and age 14 exam scores – top 10% age 11 scores only

Correlation = .495 ($r^2 = .245$)



Note how the restriction evident in **Figure 2.6.2** severely limits the correlation which drops from .886 to .495, explaining only 25% rather than 79% (approx) of the variance in age 14 scores. The moral of the story is that your sample must be representative of any dimensions relevant to your research question. If you wanted to know the extent to which exam score at age 11 predicted to exam score at age 14 you will not get accurate results if you sample only the high ability students! Interpret r^2 with caution - if you reduce the range of values of the variables in your analysis than you restrict your ability to detect relationships within the wider population.

4. Outliers: Look out for outliers as they can substantially reduce the correlation. Here is an example of this:

Figure 2.6.3: Perfect correlation

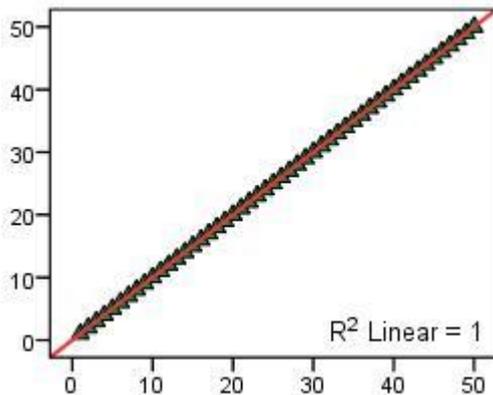
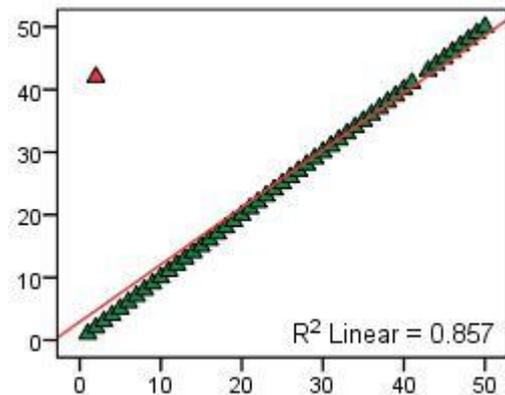


Figure 2.6.3: Same correlation with outlier



The single outlier in the plot on the right (**Figure 2.6.3**) reduces the correlation (from 1.00 to 0.86). This demonstrates how single unique cases can have an unduly large influence on your findings - it is important to check your scatterplot for potential outliers. If such cases seem to be problematic you may want to consider removing them from the analysis. However it is always worth exploring these rogue data points - outliers may make interesting case studies in themselves. For example if these data points represented schools it might be interesting to do a case study on the individual 'outlier' school to try to find out why it produces such strikingly different data. Findings about unique cases may sometimes have greater merit than findings from the analysis of the whole data set!

5. Influential cases: These are a type of outlier that greatly affects the slope of a regression line. The following charts compare regression statistics for a dataset with and without an influential point.

Figure 2.6.4: Without influential point

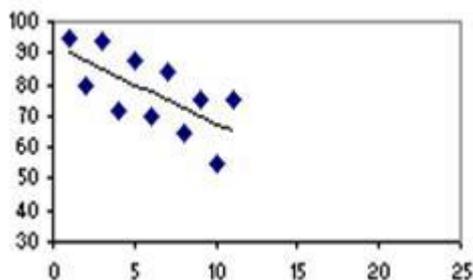
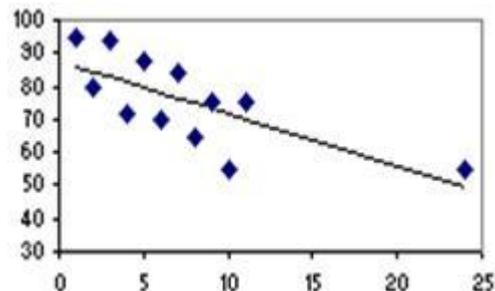


Figure 2.6.5: With influential point



Regression equation: $Y = 92.54 - 2.5X$

Regression equation: $Y = 87.59 - 16X$

Slope: $b_0 = -2.5$

Slope: $b_0 = -1.6$

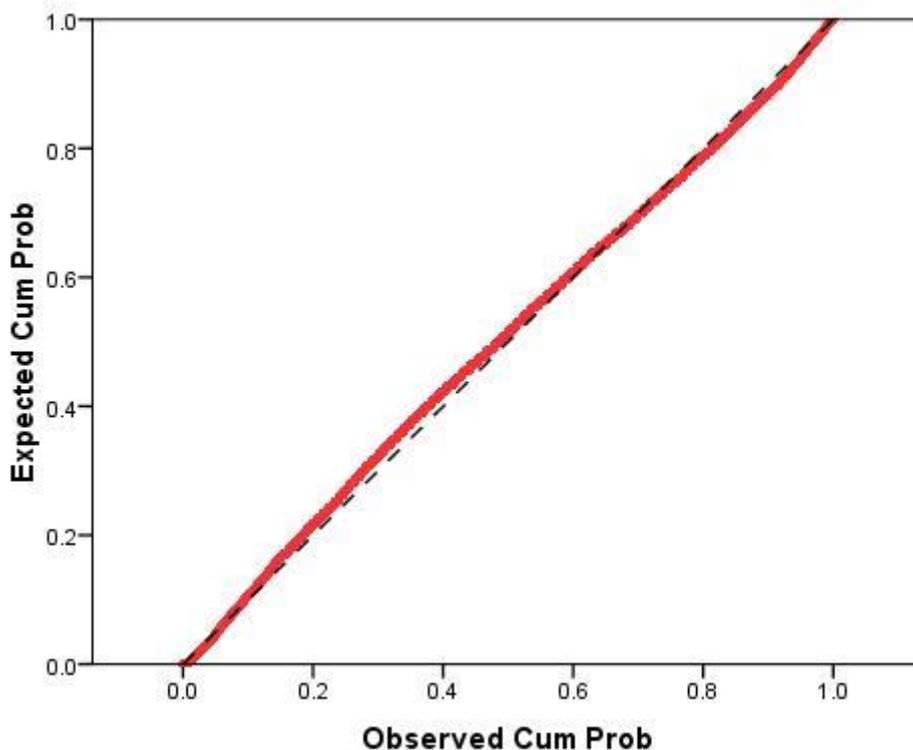
$r^2 = 0.46$

$r^2 = 0.52$

The chart on the right (**Figure 2.6.5**) has a single influential point, located at the high end of the X-axis (where $X = 24$). As a result of that single influential point the slope of the regression line decreases dramatically from -2.5 to -1.6. Note how this influential point, unlike the outliers discussed above, did not reduce the proportion of variance explained, r^2 (coefficient of determination). In fact the r^2 is slightly higher in the example with the influential case! Once again it is important to check your scatterplot in order to identify potential problems such as this one. Such cases can often be eliminated as they may just be errors in data entry (occasionally everyone makes typos...).

6. Normally distributed residuals: Residual plots can be used to ensure that there are no systematic biases in our model. A histogram of the residuals (errors) in our model can be used to loosely check that they are normally distributed but this is not very accurate. Something called a 'P-P plot' (**Figure 2.6.6**) is a more reliable way to check. The P-P plot (which stands for probability-probability plot) can be used to compare the distribution of the residuals against a normal distribution by displaying their respective cumulative probabilities. Don't worry; you do not need to know the inner workings of the P-P plot - only how to interpret one. Here is an example:

Figure 2.6.6: P-P plot of residuals for a simple linear regression



The dashed black line (which is hard to make out!) represents a normal distribution while the red line represents the distribution of the residuals (technically the lines represent the

cumulative probabilities). We are looking for the residual line to match the diagonal line of the normal distribution as closely as possible. This appears to be the case in this example – though there is some deviation the residuals appear to be essentially normally distributed.

7. Homoscedasticity: The residuals (errors) should not vary systematically across values of the explanatory variable. This can be checked by creating a scatterplot of the residuals against the explanatory variable. The distribution of residuals should not vary appreciably between different parts of the x-axis scale – meaning we are hoping for chaotic scatterplot with no discernable pattern! This may make more sense when we come to the example.

8. Independent errors:

We have shown you how you can test the appropriateness of the first two assumptions for your models, but the third assumption is rather more challenging. It can often be violated in educational research where pupils are clustered together in a hierarchical structure. For example, pupils are clustered within classes and classes are clustered within schools. This means students within the same school often have a tendency to be more similar to each other than students drawn from different schools. Pupils learn in schools and characteristics of their schools, such as the school ethos, the quality of teachers and the ability of other pupils in the school, may affect their attainment.

Even if students were randomly allocated to schools, social processes often act to create this dependence. Such clustering can be taken care of by the using of design weights which indicate the probability with which an individual case was likely to be selected within the sample. For example in published analyses of LSYPE clustering was controlled by specifying school as a cluster variable and applying published design weights using the SPSS/PASW complex samples module. More generally researchers can control for clustering through the use of multilevel regression models (also called hierarchical linear models, mixed models, random effects or variance component models) which explicitly recognised the hierarchical structure that may be present in your data.

Sounds complicated, right? It definitely can be and these issues are beyond the scope of this website. However if you feel you want to develop these skills we have an excellent sister website provided by another NCRM supported node called LEMMA which explicitly provides training on using multilevel modelling. We also know a good introductory text on multilevel modelling which you can find among our resources.

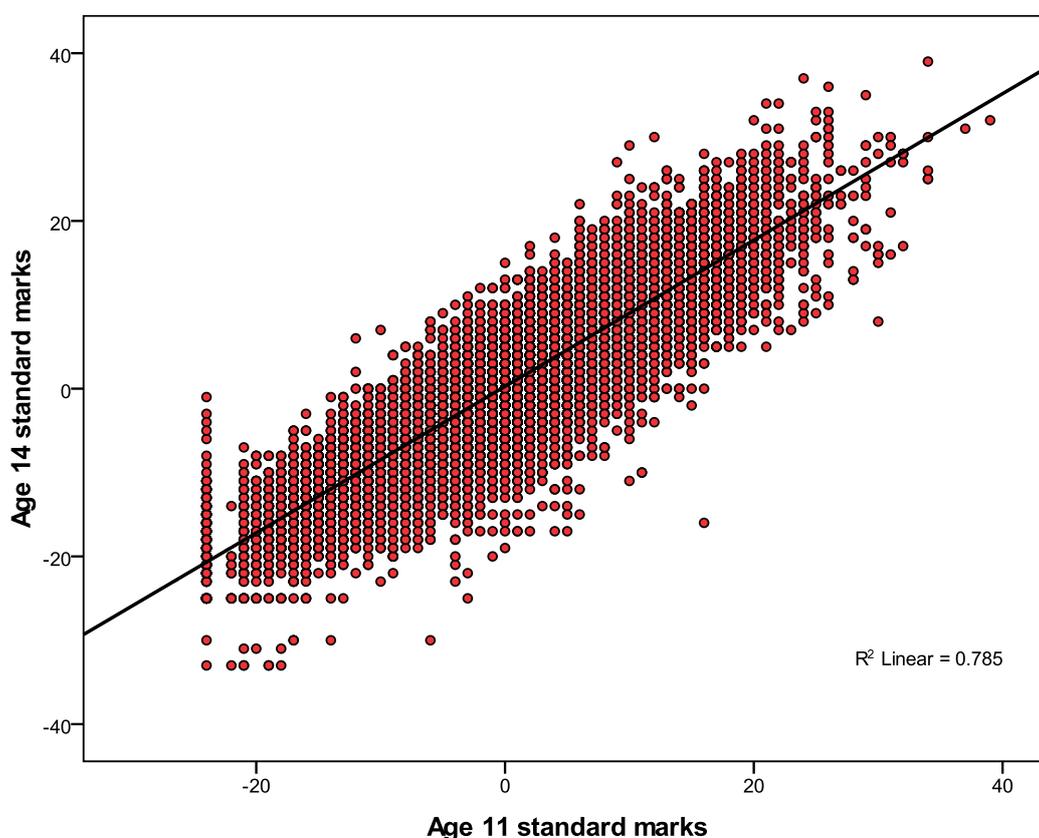
The next page will show you how to complete a simple linear regression and check the assumptions underlying it (well... most of them!) using SPSS/PASW.

2.7 SPSS and simple linear regression example

A Real Example

Let's work through an example of this using SPSS/PASW. The LSYPE dataset can be used to explore the relationship between pupils' Key Stage 2 (ks2) test score (age 11) and their Key Stage 3 (ks3) test score (age 14). Let's have another look at the scatterplot, complete with regression line, below (**Figure 2.7.1**). Note that the two exam scores are the standardized versions (mean = 0, standard deviation = 10), see Extension B on transforming data for more information if you're unsure about what we mean by this.

Figure 2.7.1: Scatterplot of KS2 and KS3 Exam scores



This regression analysis has several practical uses. By comparing the *actual* age 14 score achieved by each pupil with the age 14 score that was *predicted* from their age 11 score (the residual) we get an indication of whether the student has achieved more (or less) than would be expected given their prior score. Regression thus has the potential to be used diagnostically with pupils, allowing us to explore whether a particular student made more or less progress than expected? We can also look at the average residuals for different groups of students: Do pupils from some schools make more progress than those from others? Do boys and girls make the same amount of progress? Have pupils who missed a lot of lessons made less than expected progress? It is hopefully becoming clear just how useful a simple linear regression can be when used carefully with the right research questions.

Let's learn to run the basic analysis on SPSS/PASW. Why not follow us through the process using the LSYPE 15,000 dataset. You may also like to watch our video demonstration.

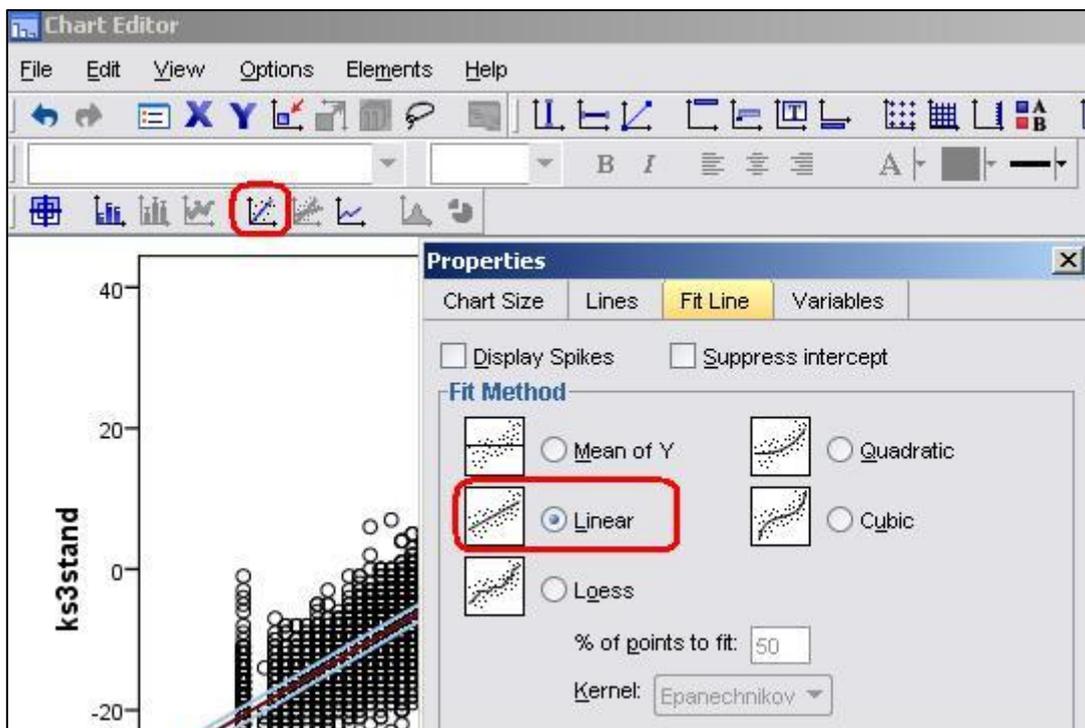
The scatterplot

As we've already shown you how to use SPSS/PASW to draw a scatterplot so we won't bore you with the details again. However you must be sure to put the standardized ks2 and ks3 score variables on to the relevant axes (shown below). The *ks3stand* variable is our outcome variable and so is placed on the Y-axis. Our explanatory variable variable is *ks2stand* and so is placed on the X-axis.



When you have got your scatterplot you can use the *chart editor* to add the regression line. Simply double click on the graph to open the editor and then click on this icon: 

This opens the *properties* pop-up. Select the *Linear* fit line, click **Apply** and then simply close the pop-up.



You can also customise the axes and colour scheme of your graph - you can play with the *chart editor* to work out how to do this if you like (we've given ours a makeover so don't worry if it looks a bit different to yours - the general pattern of the data should be the same).

It is important to use the scatterplot to check for outliers or any cases which may unduly influence the regression analysis. It is also important to check that the data points use the full scale for each variable and there is no restriction in range.

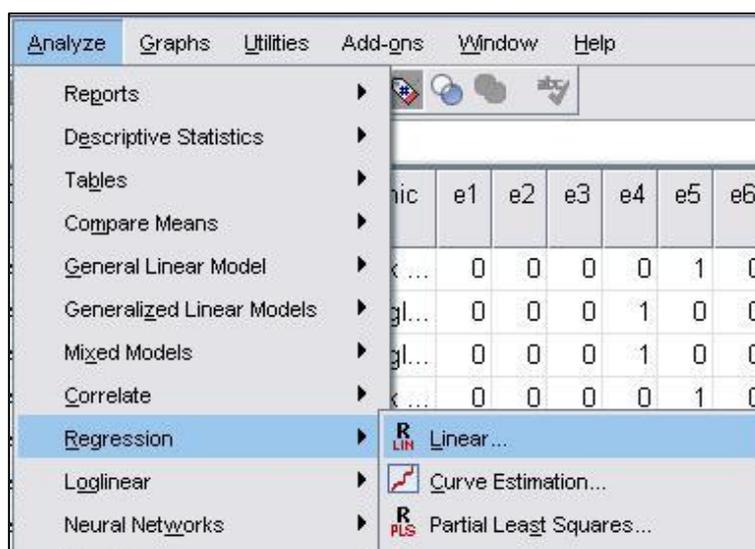
Looking at the scatterplot we've produced (**Figure 2.7.1**) there do seem to be a few outliers but given the size of the sample they are unlikely to influence the model to a large degree. It is helpful to imagine the regression line as a balance or seesaw - each data point has an influence and the further away it is from the middle (or the pivot to stretch our analogy!) the more influence it has. The influence of each case is relative to the total number of cases in the sample. In this example an outlier is one case in about 15,000 so an individual outlier, unless severely different from the expected value, will hold little influence on the model as a whole. In smaller data sets outliers can be much more influential.

Of course simply inspecting a scatterplot is not a very robust way of deciding whether or not an outlier is problematic - it is a rather subjective process! Luckily SPSS provides you with access to a number of diagnostic tools which we will discuss in more depth, particularly in coming modules.

More concerning are the apparent floor effects where a number of participants scored around -25 (standardized) at both age 11 and age 14. This corresponds to a real score of '0' in their exams! Scores of zero may have occurred for many reasons such as absence from exams or being placed in an inappropriate tier for their ability (e.g. taking an advanced version of a test when they should have taken a foundation one). This floor effect provides some extra challenges for our data analysis.

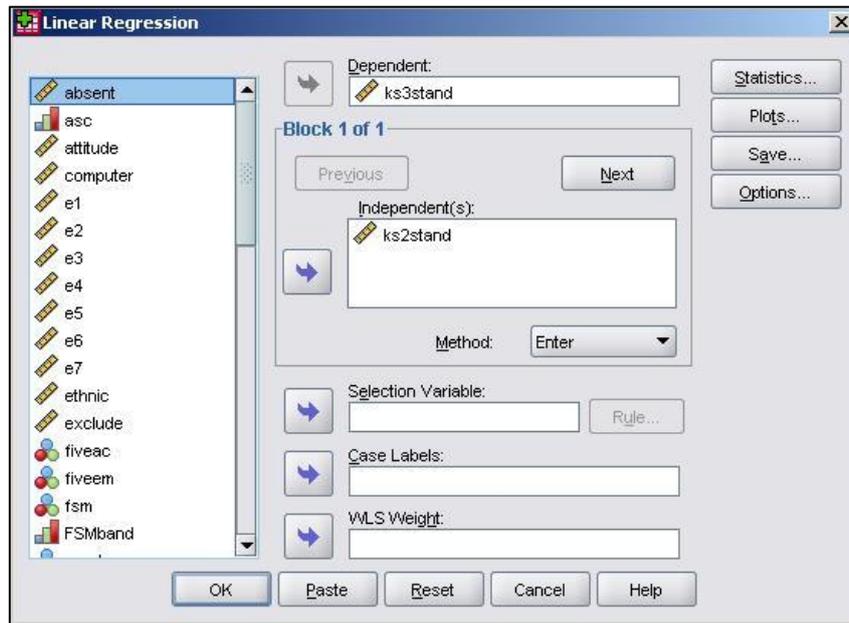
The Simple Linear Regression

Now that we've visualized the relationship between the ks2 and ks3 scores using the scatterplot we can start to explore it statistically. Take the following route through SPSS: **Analyse > Regression > Linear** (see below).



The menu shown below will pop into existence. As usual the full list of variables is listed in the window on the left. The ks3 exam score ('ks3stand') is our outcome variable so this goes

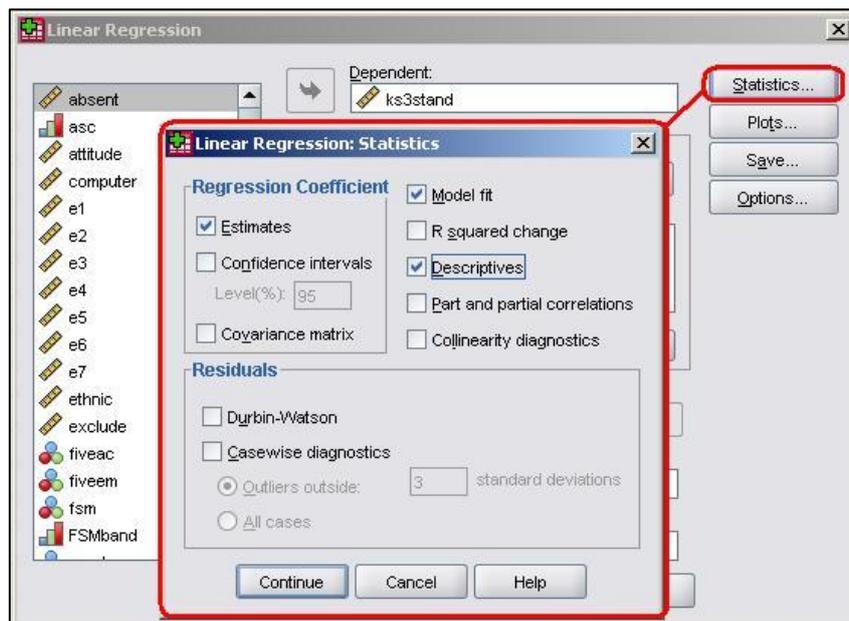
in the window marked *dependent*. The KS2 exam score ('ks2stand') is our explanatory variable and so goes into the window marked *independent(s)*.



Note that this window can take multiple variables and you can toggle a drop down menu called *method*. You will see the purpose of this when we move on to discuss multiple linear regression but leave 'Enter' selected for now. Don't click OK yet...

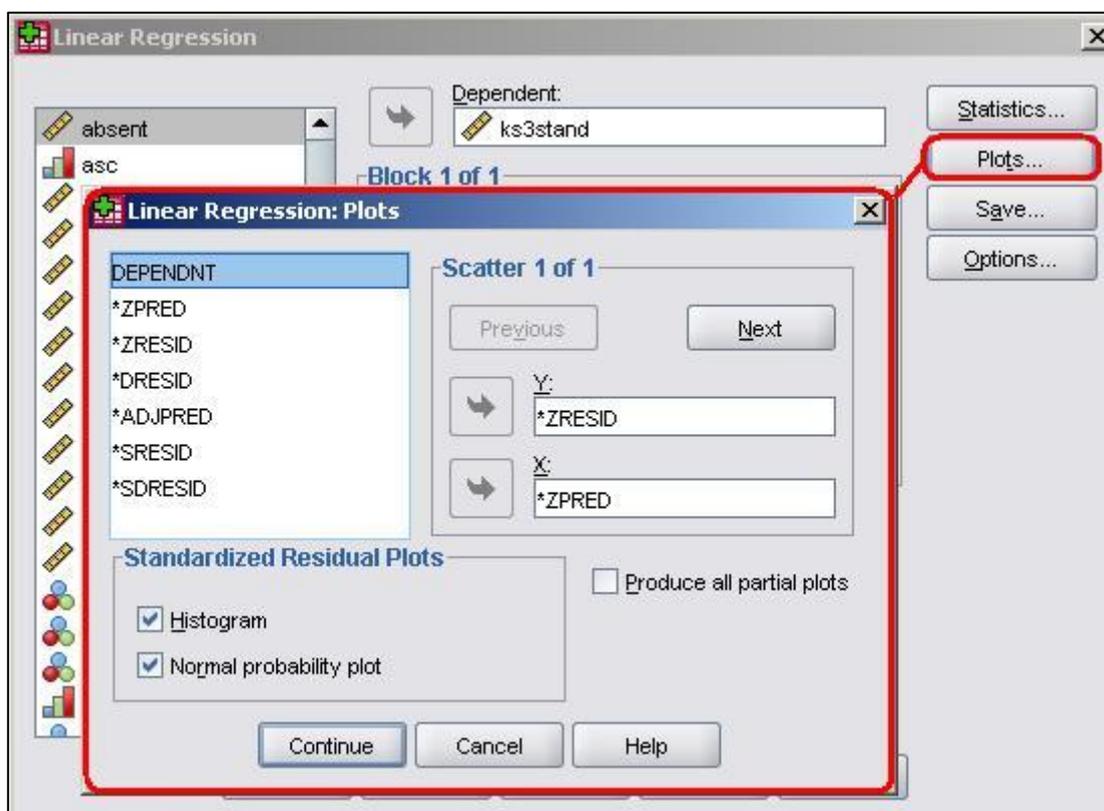
Adjusting the sub-menus

In the previous examples we have been able to leave many of the options set to their default selections but this time we are going to need to make some adjustments to the *statistics*, *plots* and *save* options. We need to tell SPSS that we need help checking that our assumptions about the data are correct! We also need some additional data in order to more thoroughly interrogate our research questions. Let's start with the *statistics* menu:



Many of these options are more important for multiple linear regression and so we will deal with them in the next module but it is worth ticking the *descriptives* box and getting the *model fit* statistics (this is ticked by default). Note also the *Residuals* options which allow you to identify outliers statistically. By checking *casewise diagnostics* and then 'outliers outside: 3 standard deviations' you can get SPSS to provide you a list of all cases where the residual is very large. We haven't done it for this dataset for one reason: we want to keep your first dabbling with simple linear regression... well, simple. This data set is so large that it will produce a long list of potential outliers that will clutter your output and cause confusion! However we would recommend doing it in most cases. To close the menu click **Continue**.

It is important that we examine a plot of the residuals to check that a) they are normally distributed and b) that they do not vary systematically with the predicted values. The *plots* option allows us to do this - click on it to access the menu as shown below.



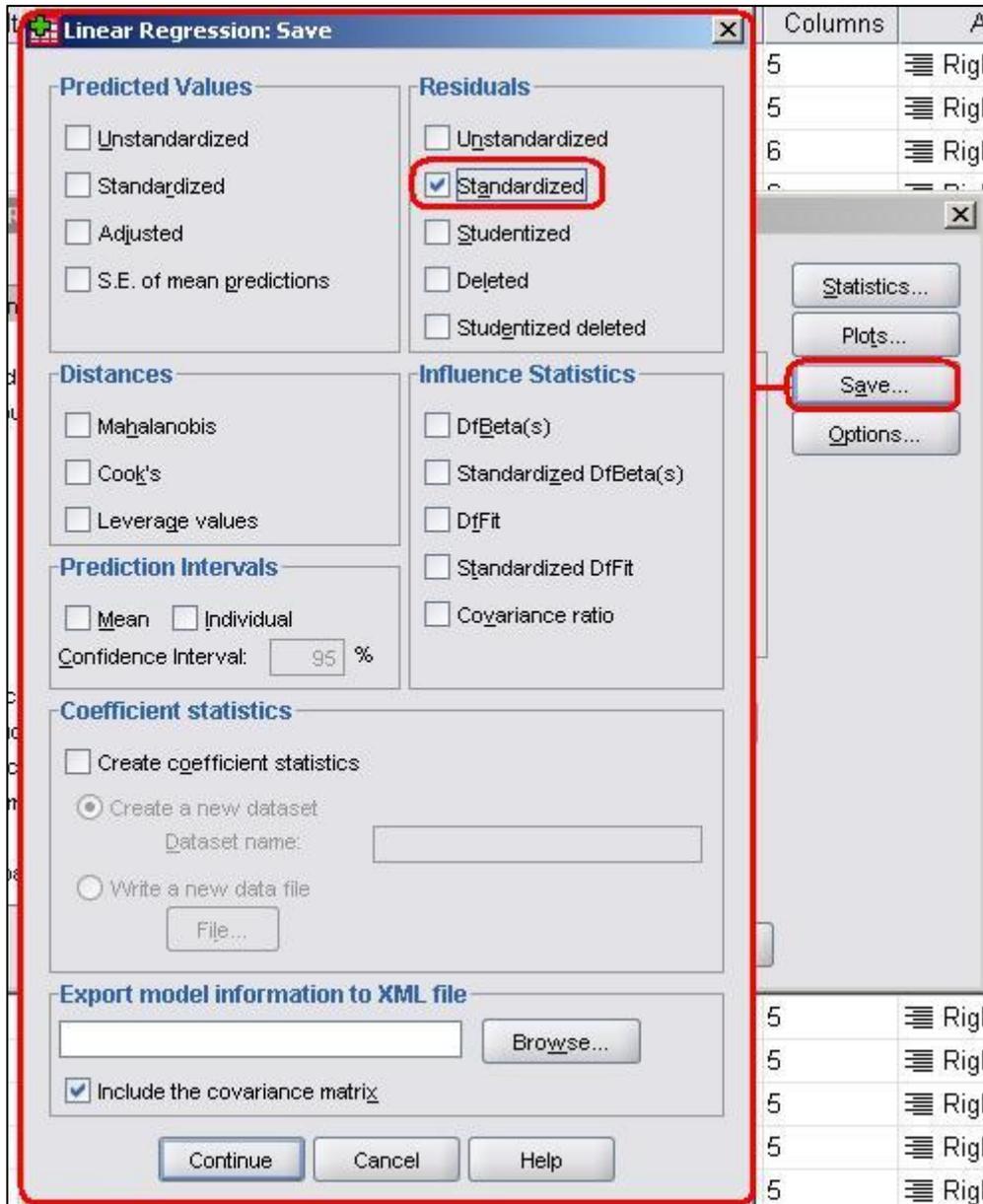
Note that rising feeling of irritation mingled with confusion that you feel when looking at the list of words in the left hand window. Horrible aren't they? They may look horrible but actually they are very useful - once you decode what they mean they are easy to use and you can draw a number of different scatterplots. For our purposes here we need to understand two of the words:

- ***ZRESID** - The standardized residuals for each case.
- ***ZPRED** - The standardized predicted values for each case.

As we have discussed, the term standardized simply means that the variable is adjusted such that it has a mean of zero and standard deviation of one - this makes comparisons between variables much easier to make because they are all in the same 'standard' units. By plotting ***ZRESID** on the Y-axis and ***ZPRED** on the X-axis you will be able to check the

assumption of homoscedasticity - residuals should not vary systematically with each predicted value and variance in residuals should be similar across all predicted values. You should also tick the boxes marked *Histogram* and *P-P plot*. This will allow you to check that the residuals are normally distributed. To close the menu click **Continue**.

Finally let's look at the *Save* options:



This menu allows you to generate new variables for each case/participant in your data set. The *Distances* and *Influence Statistics* allow you to interrogate outliers in more depth but we won't overburden you with them at this stage. In fact the only new variable we want is the standardized residuals (the same as our old nemesis *ZRESID) so check the relevant box as shown above. You could also get *Predicted* values for each case along with a host of other adjusted variables if you really wanted to! When ready, close the menu by clicking **Continue**.

Don't be too concerned if you're finding this hard to understand. It takes practice! We will be going over the assumptions of linear regression again when we tackle multiple linear

regression in the next module. Once you're happy that you have selected all of the relevant options click **OK** to run the analysis. SPSS/PASW will now get ridiculously overexcited and bombard you with output! Don't worry you really don't need all of it. Let's try and navigate this output together on the next page...

2.8 SPSS and simple linear regression output

Interpreting Simple Linear Regression SPSS/PASW Output

We've been given a quite a lot of output but don't feel overwhelmed: picking out the important statistics and interpreting their meaning is much easier than it may appear at first (you can follow this on our video demonstration). The first couple of tables (**Figure 2.8.1**) provide the basics:

Figure 2.8.1: Simple Linear regression descriptives and correlations output

Descriptive Statistics			
	Mean	Std. Deviation	N
Age 14 standard marks	.34	9.768	14138
Age 11 standard marks	.09	9.918	14138

Correlations			
		Age 14 standard marks	Age 11 standard marks
Pearson Correlation	Age 14 standard marks	1.000	.886
	Age 11 standard marks	.886	1.000
Sig. (1-tailed)	Age 14 standard marks	.	.000
	Age 11 standard marks	.000	.
N	Age 14 standard marks	14138	14138
	Age 11 standard marks	14138	14138

The *Descriptive Statistics* simply provide the mean and standard deviation for both your explanatory and outcome variables. Because we are using standardized values you will notice that the mean is close to zero. They are not *exactly* zero because certain participants were excluded from the analysis where they had data missing for either their age 11 or age 14 score. Extension B discusses issues of missing data. More useful is the *correlations* table which provides a correlation matrix along with probability values for all variables. As we only have two variables there is only one correlation coefficient. A correlation of .886 ($P < .0005$) suggests there is a strong positive relationship between age 11 and age 14 exam scores. There is also a table entitled *Variables Entered/Removed* which we've not included on this page because it is not relevant! It becomes important for multiple linear regression, so we'll discuss it then.

The next three tables (**Figure 2.8.2**) get to the heart of the matter, examining your regression model statistically:

Figure 2.8.2: SPSS Simple linear regression model output

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.886 ^a	.785	.785	4.525

a. Predictors: (Constant), Age 11 standard marks

b. Dependent Variable: Age 14 standard marks

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1059510.757	1	1059510.757	51750.500	.000 ^a
	Residual	289412.550	14136	20.473		
	Total	1348923.307	14137			

a. Predictors: (Constant), Age 11 standard marks

b. Dependent Variable: Age 14 standard marks

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.261	.038		6.848	.000
	Age 11 standard marks	.873	.004	.886	227.487	.000

a. Dependent Variable: Age 14 standard marks

The *Model Summary* provides the correlation coefficient and coefficient of determination (r^2) for the regression model. As we have already seen a coefficient of .886 suggests there is a strong positive relationship between age 11 and age 14 exam scores while $r^2 = .785$ suggests that 79% of the variance in age 14 score can be explained by the age 11 score. In other words the success of a student at age 14 is strongly predicted by how successful they were at age 11.

The *ANOVA* tells us whether our regression model explains a statistically significant proportion of the variance. Specifically it uses a ratio to compare how well our linear regression model predicts the outcome to how accurate simply using the mean of the outcome data as an estimate is. Hopefully our model predicts the outcome more accurately than if we were just guessing the mean every time! Given the strength of the correlation it is not surprising that our model is statistically significant ($p < .0005$).

The *Coefficients* table gives us the values for the regression line. This table can look a little confusing at first. Basically in the *(Constant)* row the column marked *B* provides us with our intercept - this is where $X = 0$ (where the age 11 score is zero – which is the mean). In the *Age 11 standard marks* row the *B* column provides the gradient of the regression line which is the regression coefficient (**B**). This means that for every one standard mark increase in age 11 score (one tenth of a standard deviation) the model predicts an increase of 0.873 standard marks in age 14 score. Notice how there is also a standardized version of this second *B*-value which is labelled as *Beta* (**β**). This becomes important when interpreting multiple explanatory variables so we'll come to this in the next module. Finally the t-test in the second row tells us whether the ks2 variable is making a statistically significant contribution to the predictive

power of the model - we can see that it is! Again this is more useful when performing a multiple linear regression.

Interpreting the Residuals Output

The table below (**Figure 2.8.3**) summarises the residuals and predicted values produced by the model.

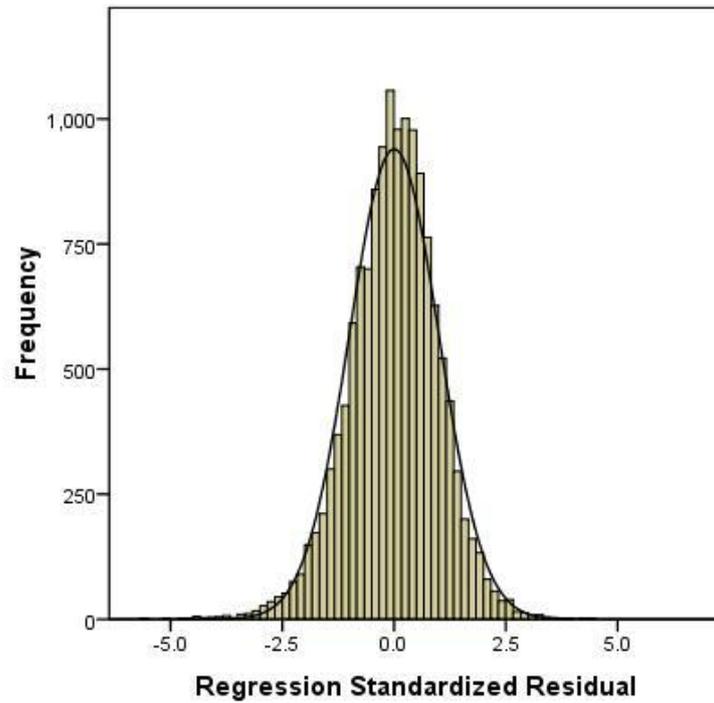
Figure 2.8.3: SPSS simple linear regression residuals output

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-20.69	34.30	.34	8.657	14138
Residual	-30.226	20.011	.000	4.525	14138
Std. Predicted Value	-2.429	3.923	.000	1.000	14138
Std. Residual	-6.680	4.422	.000	1.000	14138

It also provides standardized versions of both of these summaries. You will also note that you have a new variable in your data set: **ZRE_1** (you may want to re-label this so it is a bit more user friendly!). This provides the standardized residuals for each of your participants and can be analysed to answer certain research questions. Residuals are a measure of error in prediction so it may be worth using them to explore whether the model is more accurate for predicting the outcomes of some groups compared to others (e.g. Do boys and girls make the same amount of progress?).

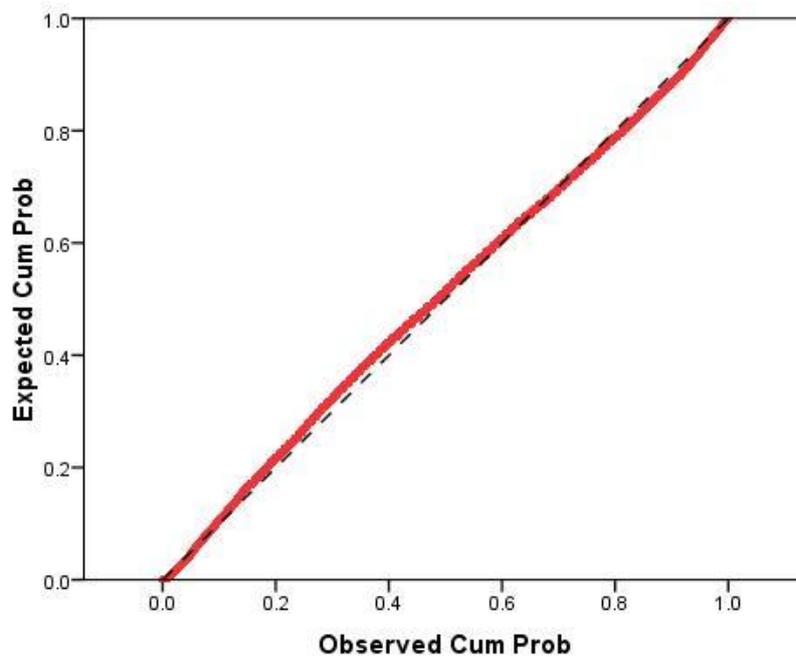
We requested some graphs from the *Plots* sub-menu when running the regression analysis so that we could test our assumptions. The histogram (**Figure 2.8.4**) shows us that we may have problems with our residuals as they are not quite normally distributed - though they roughly match the overlaid normal curve, the residuals are clearly clustering around and just above the mean more than is desirable.

Figure 2.8.4: Histogram of residuals for the simple linear regression



We have also generated a P-P plot to check that our residuals are normally distributed (**Figure 2.8.5**). We can use this plot to compare the observed residuals with what we'd expect if they were normally distributed (represented by the diagonal line). We can see that, aside from a minor departure at the observed cumulative probability of 0.4, the data is normally distributed.

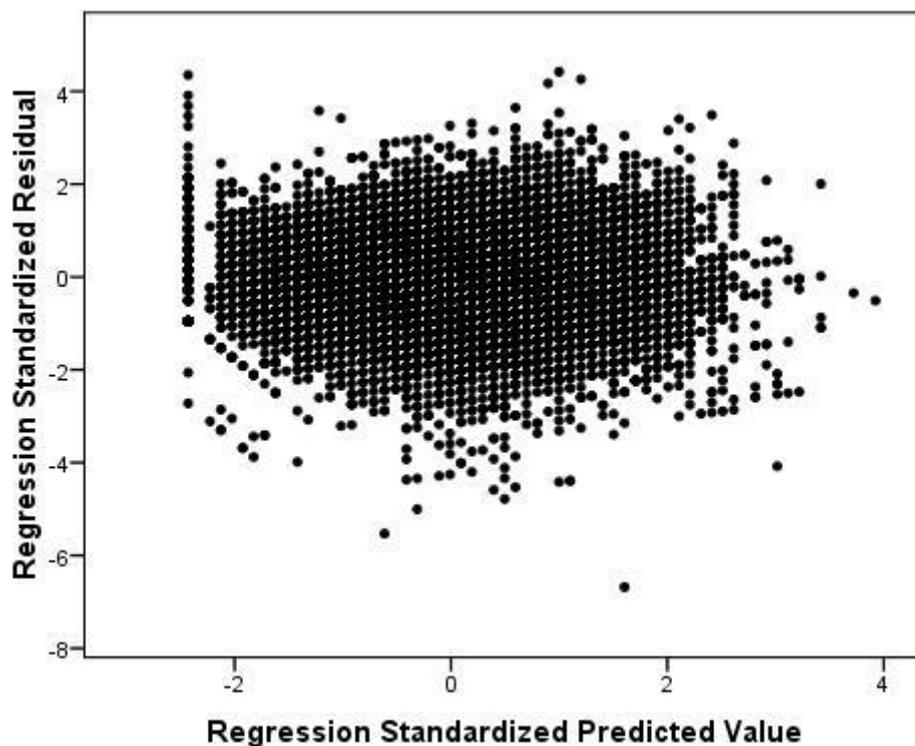
Figure 2.8.5: P-P plot for simple linear regression model residuals



As you will see, when using real world data such imperfections are to be expected! There are some issues with the distribution of the residuals and we may need to make a judgement call about whether to remove outliers or alter our model to fix this problem. In this case we decided that the benefits of keeping all cases (including the students with low grades!) in our sample outweighed the issues regarding the ambiguous interpretation of whether or not our residuals were normally distributed.

Finally, the scatterplot (**Figure 2.8.6**) shows us how large the standardized residual was for each case at each value of the predicted outcome. We are hoping that this will look pretty random as this would fulfil our assumption of homoscedasticity (that word again... must get you a good score in Scrabble).

Figure 2.8.6: Scatterplot of residuals and predicted values



Generally there is a problem with a large range of scores at the lowest end of the predicted value (X-axis). This is due to 'floor' effects in the data. It is often worth considering identifying outlying cases and maybe removing them from the analysis but in this example such cases are too interesting to remove! Despite our floor effect, overall the residuals at each predicted value do not appear to vary differently with the exception of a few outliers so it looks as if we have met the assumption.

Note that scatterplots can look like giant ink blobs when datasets are as large as this one and this can make interpreting them tricky. SPSS/PASW has a facility called 'binning' that is not as rubbish as it sounds (sorry) and can help us here. We discuss binning in **Extension C**.

Summary

Overall our regression model provides us with a good method of predicting age 14 exam scores by using age 11 scores.

We could report this in the following way:

A simple linear regression was carried out to ascertain the extent to which age 11 (ks2) assessment scores can predict age 14 (ks3) assessment scores. A strong positive correlation was found between ks2 and ks3 scores ($r = .89$) and the regression model predicted 79% of the variance. The model was a good fit for the data ($F = 51751, p < .0005$).

	B	SE B	β
Constant	.261	.038	
KS2 score	.873	.004	.886

We highly recommend you check the style guide for your university or target audience before writing up. Different institutions work under different criteria and often require very specific styles and formatting. Note that it is important not to simply cut and paste SPSS output into your report - it looks untidy and, as you know, it is full of unnecessary detail!

Perhaps you have been introduced to a few too many new ideas in this module and you need a little lie down. Rest assured that if you can grasp the basics of simple linear regression then you are off to a flying start. Our next module on multiple linear regression simply expands on the ideas you have already bravely survived here.

We recommend that you take our quiz and work through our exercises to consolidate your knowledge. Then move on to the next module!

Exercise

Simple Linear Regression Module Exercise

The following three questions are essentially research questions. You can work through them using the LSYPE 15,000 dataset and your new found statistical super powers! We recommend that you answer them in full sentences with supporting tables or graphs where appropriate – this will help when you come to report your own research. There is a link to the answers at the bottom of the page.

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

Question 1

Is there a statistically significant association between whether or not a student has been truant in the last twelve months (*truancy*) and whether or not they achieve 5 GCSE qualifications of A* - C, including maths and English (*fiveem*)?

Use crosstabulation and a chi-square analysis to answer this question and provide the supporting statistics.

Question 2

We have seen that age 11 exam scores have a strong positive correlation with age 14 exam scores. Is there similar association between age 11 (*ks2stand*) exam scores and exams scores at age 16 (*ks4stand*)?

Draw a scatterplot and perform a bivariate correlation to answer this question.

Question 3

The Income Deprivation Affecting Children Index (IDACI) provides a standardized measure of how relatively deprived the students in the LSYPE sample are. The mean score is 0 (the measure is standardized) while negative values represent students from relatively affluent backgrounds and positive values students from relatively poor backgrounds. Can IDACI score (*IDACI_n*) be used to predict students' exam scores at age 14 (*ks3stand*)?

*Perform a simple linear regression using *ks3stand* as the outcome variable and *IDACI_n* as the explanatory variable to answer this question. Be sure to check that the assumptions of the analysis are met.*

Answers

Simple Linear Regression Module Exercise

Question 1

In this case we are looking for association with two nominal variables. The best way to do this is to use a chi-square analysis. The expected crosstabulation is shown below. Page 1.2 runs you through the process of creating a crosstabulation and calculating a chi-square if you are stuck.

			5 or more A*-C incl. English & maths		Total
			no	yes	
truant in last 12 months	No	Count	5773	6206	11979
		% within truant in last 12 months	48.2%	51.8%	100.0%
	Yes	Count	1568	556	2124
		% within truant in last 12 months	73.8%	26.2%	100.0%
Total		Count	7341	6762	14103
		% within truant in last 12 months	52.1%	47.9%	100.0%

Of those students who had been truant in the last year 73.8% did not achieve 5 A*-C grades at GCSE (including Maths and English). This is compared to 48.2% of students who had not been truant. It seems that those who have played truant are less likely to achieve the GCSE target. We can test how probable it is to find this association by chance when in fact there isn't one in the population by using the Chi-square test table shown below.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	474.858 ^a	1	.000		
Continuity Correction ^b	473.832	1	.000		
Likelihood Ratio	494.190	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	474.824	1	.000		
N of Valid Cases	14103				

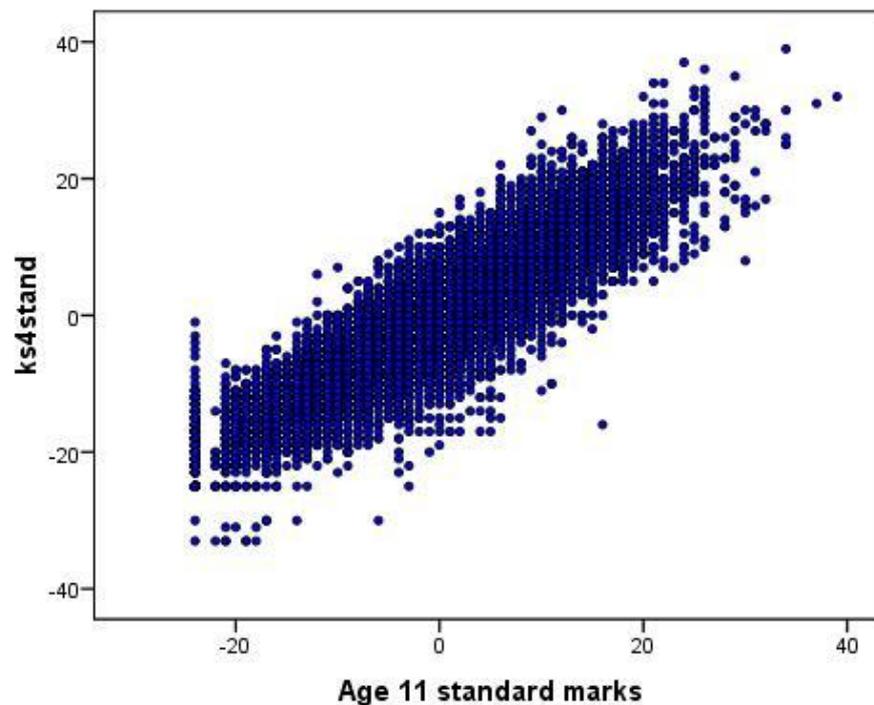
The *Pearson Chi-Square* confirms that the association we have spotted is statistically significant. The value of 474.858 (df = 1) has 'Asymp Sig' of .000, which means that $p < .0005$. SPSS/PASW has done its very best to befuddle us with a huge amount of confusingly labelled output but we have emerged victorious. Huzzah!

Overall your answer to this question could look a bit like this:

An association was found between whether a student had been truant (in the last 12 months) and whether they had achieved 5 or more A* - C grades at GCSE (including English and Maths), Chi-square = 474.86, $df = 1$, $p < .0005$. Examination of the cross-tabulation suggests that this association stems from those who have played truant in the last twelve months being more unlikely to achieve 5 A* - C grades at GCSE.

Question 2

In this case we are looking at an association between two continuous variables and so we can explore how they correlate. The first thing to do in such situations is create a scatterplot of the data! After that we can calculate a correlation coefficient for the relationship (usually a Pearson's coefficient). **Pages 2.3 and 2.4** explain how to generate a scatterplot and how to perform a bivariate correlation if you need to refresh. You should get a scatterplot a little like the one below:



You have probably noticed how similar this scatterplot is to the one for the age 11 and age 14 data! It looks as though past performance has strong positive correlation with more recent performance even over this greater timescale. Let us check this statistically with a correlation coefficient.

Correlations

		Age 16 standard marks	Age 11 standard marks
Age 16 standard marks	Pearson Correlation	1	.886**
	Sig. (2-tailed)		.000
	N	14832	14138
Age 11 standard marks	Pearson Correlation	.886**	1
	Sig. (2-tailed)	.000	
	N	14138	14301

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

According to the ‘*Pearson Correlation*’ row the coefficient of .886 is strong and positive just as we may have expected based on the scatterplot. The value of .000 in the ‘*Sig*’ row suggests that if there was no association between these two variables the probability of obtaining a coefficient of this strength with a sample of this size would be very small, $p < .0005$. Your answer to this question could look like this:

A strong positive correlation was found between age 11 and age 16 exam scores, $r = .886$, $p < .0005$.

Question 3

This question is the trickiest as it requires you to perform a full simple linear regression analysis! Pages 2.7 and 2.8 run through this if you feel you need to refresh your memory. You will get a colossal amount of output but we’ll only take you through the important bits.

Correlations

		Age 16 standard marks	IDACI normal score
Pearson Correlation	Age 16 standard marks	1.000	-.340
	IDACI normal score	-.340	1.000
Sig. (1-tailed)	Age 16 standard marks		.000
	IDACI normal score	.000	
N	Age 16 standard marks	14820	14820
	IDACI normal score	14820	14820

The correlation between Age 16 exam score and IDACI is negative, $r = -.340$ which means that as level of deprivation increases GCSE scores decrease.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.340 ^a	.116	.116	9.388

The model summary tells us how much of the variance in our outcome variable is accounted for by our explanatory variable (that is how much variance does the simple linear regression model account for). The R Square column gives us this value, which is known as the

coefficient of determination, r^2 , as .116. In other words IDACI score accounts for 12% of the variance in age 16 exam score.

ANOVA^b

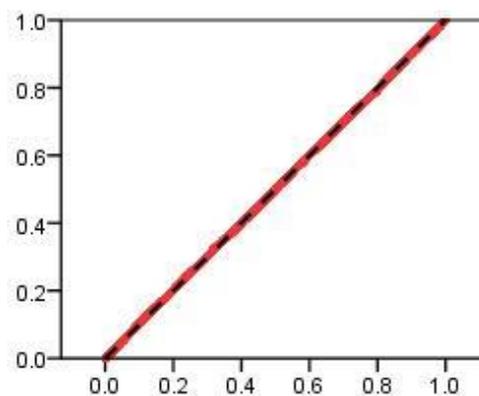
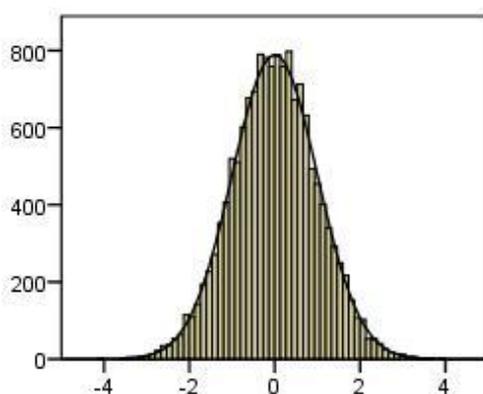
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	171099.780	1	171099.780	1941.436	.000 ^a
	Residual	1305918.216	14818	88.131		
	Total	1477017.996	14819			

The ANOVA table tells us whether or not our simple linear regression model is better at predicting the outcome variable than simply using the mean of the outcome variable. The F-ratio of 1941.4 (see column *F*) is statistically significant at $p < .0005$ (Column: *Sig* = .000), suggesting that our model does improve the prediction.

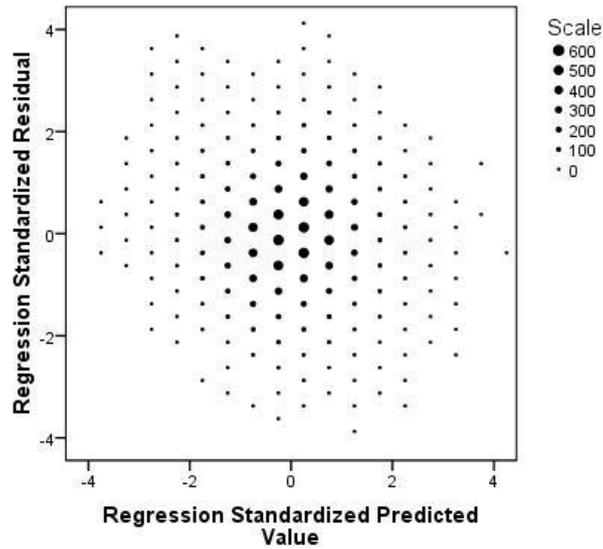
Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.077	.077		1.000	.317
	IDACI normal score	-3.440	.078	-.340	-44.062	.000

The coefficients table provides lots of information about the model. The intercept (where IDACI score = 0) is provided in the 'B' column for the '(Constant)' row as .077 while the gradient of the regression line (the coefficient) is in the 'B' column of the 'IDACI normal score' row: -3.440. This means that for standard deviation that the IDACI score increases the predicted age 16 score decreases by 3.4 standard units. The final two columns for this row ($t = -44.062$, $p < .0005$) tell us that IDACI is a statistically significant predictor of the outcome (age 16 exam score).



Above are miniaturized versions of the histogram and P-P plots that you should find in your output. As you can see the residuals show no major departures from the normal distribution to be overly concerned with.



The above scatterplot can be used to identify homoscedasticity. Note that we have used ‘binning’ on our version so that we can make out the pattern of the data more easily. The plot seems to show a fairly random spread of residuals (as we would hope for our assumptions). Homoscedasticity does not appear to be problematic.

Overall your answer to the question might look a bit like this:

A simple linear regression was carried out to ascertain the extent to which income deprivation (IDACI) can predict age 16 (ks4) assessment scores. A positive correlation was found between IDACI and ks4 score ($r = .340$) and the regression model predicted 12% of the variance. The model was appropriate for predicting the outcome variable ($F = 1941, p < .0005$).

	B	SE B	β
Constant	.077	.077	
IDACI (normalized)	-3.44	.078	-.340