



The
University
Of
Sheffield.

**Going
Further
with
SPSS 16.**

CiCS
Jean Russell
Bob Booth
May 2010
AP-SPSS6

Contents

1. INTRODUCTION	3
1.1 MORE ON VARIABLES AND ANALYSIS	3
2. STARTING SPSS.....	5
2.1 SAVING AND LOADING SPSS DATA	6
3. ADVANCED ANALYSES	7
3.1 FACTORIAL ANOVA (AND ANCOVA).....	7
3.2 CHECKING THAT THE RESIDUALS ARE NORMALLY DISTRIBUTED.....	10
3.3 WHY STATISTICIANS PREFER GRAPHS TO TEST FOR CHECKING NORMALITY OF RESIDUALS	12
3.4 EXERCISE	12
3.5 LINEAR REGRESSION.....	13
3.6 EXERCISE	16
3.7 CATEGORICAL AND CONTINUOUS EXPLANATORY VARIABLES: ANOVA OR LINEAR REGRESSION.....	17
3.8 EXERCISE	19
3.9 LOGISTIC REGRESSION	19
3.10 EXERCISE	22
4. INTERACTING WITH A WORD DOCUMENT	24
4.1 PUTTING DATA TABLES IN WORD DOCUMENTS.....	24
4.2 PUTTING GRAPHS IN WORD DOCUMENTS	24
4.3 EXPORTING TO WEB PAGE AND GRAPHICS FILES.....	24
5. SYNTAX	25
5.1 EXERCISES	25
6. FURTHER READING	26

1. Introduction

This document is for people who have used SPSS and have decided to go further that simple univariate and bivariate statistics. SPSS 16 is available on the Managed XP Service, and can be installed on personal Windows Vista XP computers and older Macs.

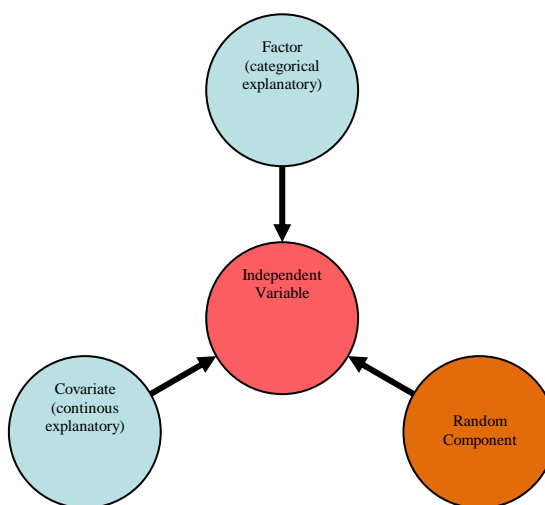
If you have Snow Leopard or Windows 7 then you will need SPSS 18. This is on the whole similar to SPSS 16.

1.1 More on variables and analysis

As we are moving to more complex analyses we need a slightly more complex model of what sort of variables we have. Firstly we need to distinguish between **dependent variables** and **explanatory (independent) variables**.

First let's deal with the **dependent variable**. There is normally only one of these in an analysis (ignoring MANOVA and Repeated Measures Analysis of Variance for now). The dependent variable is the outcome variable. It is seen as the variable that you are primarily interested in. A dependent variable can be either categorical or continuous. It also is the variable that contains all the randomness in the study. This is the one to worry if it is distributed correctly but make sure you know which distribution it should be distributed like, because that distribution may depend on the **explanatory variables**.

An **explanatory variable** is any that may explain some of the differences in values of the **dependent variable**. I don't like calling these variables **independent** variables because independence means no correlation and often there are correlations between different **explanatory variables**. Admittedly ideally they should not be correlated, but that almost never happens. You may have one, two or more of these in an analysis and just like a **dependent variable** they can be both categorical and continuous. However, sometimes continuous explanatory variables are known as **factors** and sometimes continuous explanatory variables are known as **covariates**.



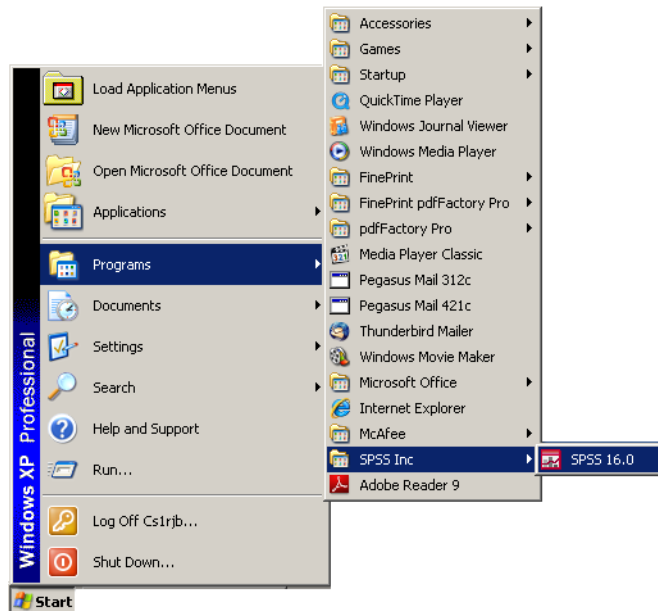
You therefore have now two ways of classifying all the variables in your analysis:

		Dependent	Explanatory	
			Factor	Covariate
Continuous	Integer			
	Rational			
Categorical	Binary			
	Ordinal			
	Multinomial			

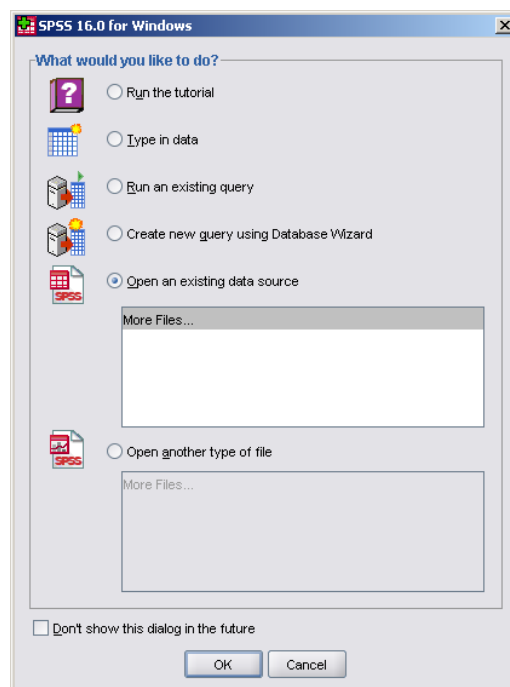
Therefore for most analyses done here you should be able to fit every variable in your table into one of the clear cells above. Remember you can have more than one factor and more than one covariate.

2. Starting SPSS

SPSS can be installed on Windows XP, Vista and 7 computers. After installation the software will be available from the **Start** menu, under **Programs**. On the Managed XP Service, SPSS is also available from the **Programs** menu.



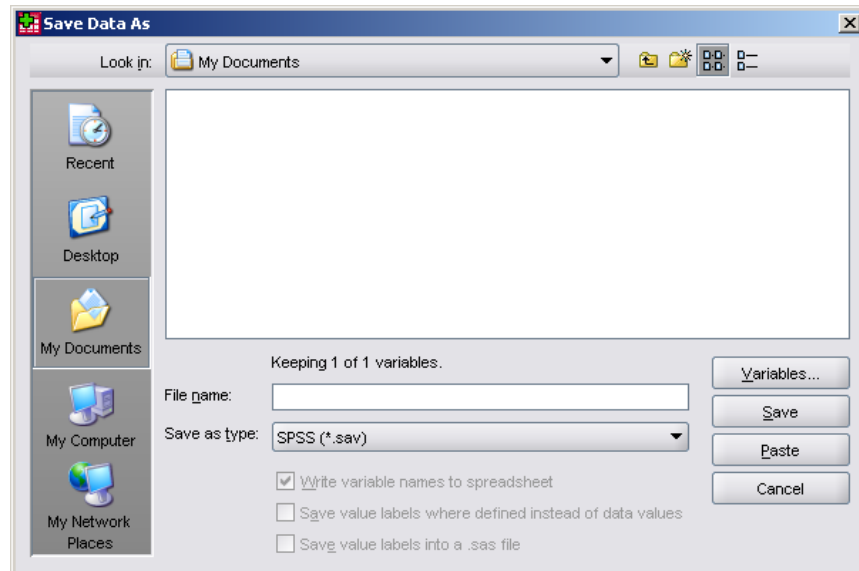
When SPSS starts you will see the following dialogue box.




2.1 Saving and Loading SPSS Data

Once you have entered all your data into SPSS you should save it into a data file. Ideally you should save your data many times before this, say after defining your variables, then after inputting data for several units, and so on until your data is complete.

To save your data file either click the **File** menu and select **Save**, or click the usual Save button. You will see a typical Windows Save dialogue.



From here you can specify the name for your data file, and use the **Save in** field to specify the drive and folder in which to save the file. You can even use the New Folder button to create a new folder in which to save your data file. 

SPSS data files are usually saved with the extension **.sav**, but in the **Save as type** field you can see some other formats for SPSS data. Data can be saved in text files with the extension **.txt** or **.dat**, or as Excel data with the extension **.xls**. You can also save data in other, less common, formats.

When you return to SPSS you can load existing SPSS data files using **Open** from the **File** menu, or by using the usual Open button from the Toolbar. All the data files ending in **.sav** will be available, and you can load any data file by selecting it, then clicking the **Open** button. In addition you can load data from other applications, as described in the following sections.

3. Advanced Analyses

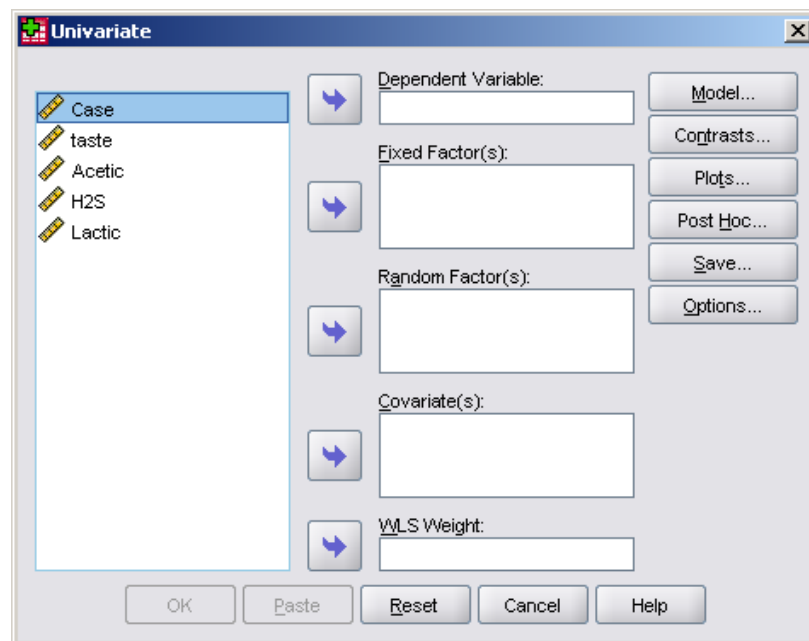
3.1 Factorial ANOVA (and ANCOVA)

Use for: Telling the differences between mean values of single dependent variables when there is more than one grouping variable and or when there are both factors and covariates.

Limitations: It is assumed that the populations are normally distributed and have equal variance. It also assumes that the samples are independent of each other.

Factorial ANOVA, ANCOVA and various other forms of ANOVA are all gathered together as a group of techniques know as General Linear Models. Technically Linear Regression should also be in this group (if there are no factors and you do analysis of covariance that is identical with Regression) but SPSS has decided to ignore this.

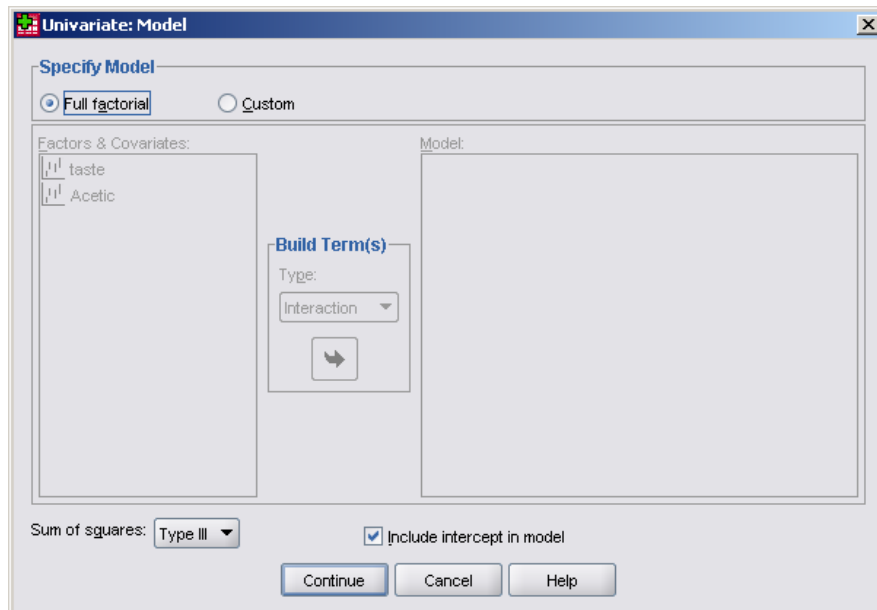
From the **Analyze** menu, select **General Linear Models** and then **Univariate**. This will bring up the following dialogue box:



In the **Dependent Variable** section, you should put the continuous outcome variable that you are interested in. In the **Fixed Factors(s)** section, specify the explanatory factors which have had fixed or determined levels, these would include things like gender and college. **Random Factor(s)** are where the selection of levels is a feature of the sampling that has taken place. Example of a random factor may be country of origin of students entering a University. Some levels like the UK will come up year after year, but it is unlikely that every year there would be students from Tonga. So in a sense the countries that turn up are a result of a random process. Normally, although we are interested

in removing the variance due to a Random Factor, we are not interested in estimating its effect too precisely. In the **Covariate(s)** section, we would place continuous explanatory variables.

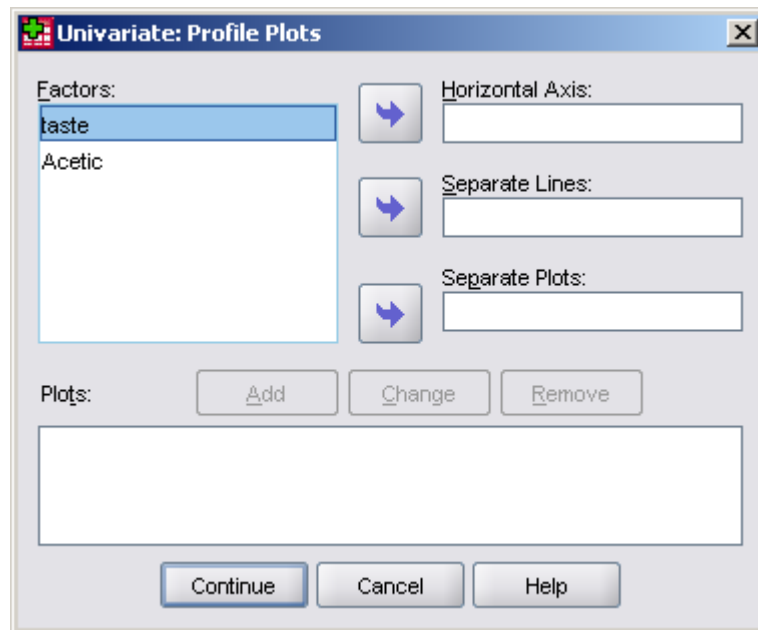
Click the **Model** button to see the following dialogue box:



With simple ANOVAs like the present one, you can normally go for the **Full factorial** model. With complex ones (i.e. with three or more factors and covariates) it is a good idea to do a custom model, such as the main effects (factors and covariates) and their two-way interactions. Three-way interactions and above are notoriously difficult to handle. It is normally a good idea to keep **Sum of Squares** set as **Type III**. On a few occasions a change to **Type IV** may be sensible. Types **I** and **II** are the same as **Type III** where you are carrying out a fully designed study, otherwise they cause problems with interpretation. On the whole they are not recommended.

The **Contrasts** button is beyond the scope of this text. If you want to access the full power of ANOVA then you really do need to get to grips with contrasts. I suggest that you read section 8.2.10 Planned Contrasts from Field p 325.

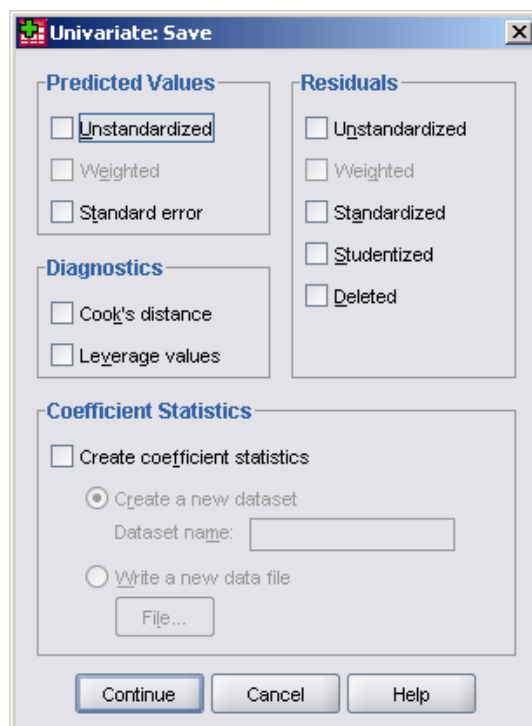
Click the **Plots** button of the original dialogue box to see the following:



I find that plots tend to be the first place I go once I have a significant result, so I suggest you put all the main effects and two-way interactions into the plot. One rule of thumb is for the two way interaction to put the factor with the most categories along the **Horizontal Axis**. You will need to click the **Add** button to add each graph to the list.

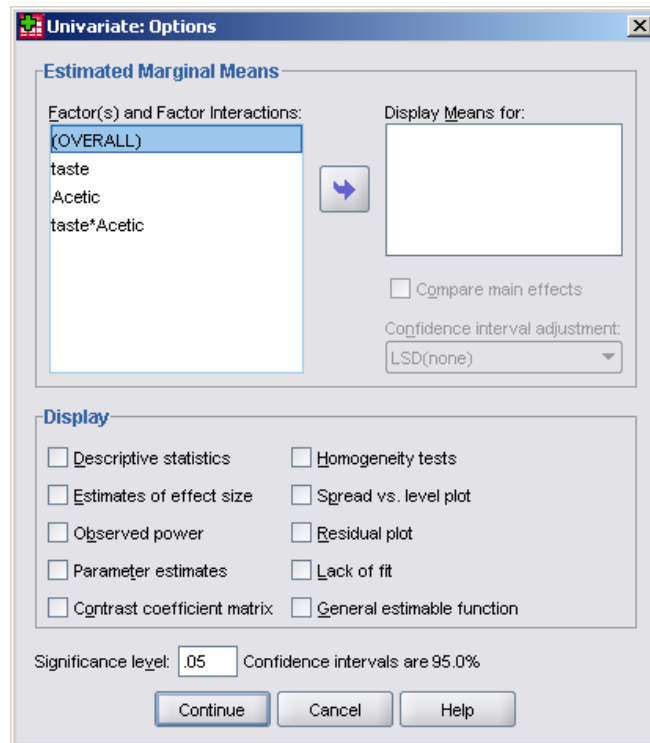
The **Post Hoc** button of the original dialogue box offers you exactly the same choice as it does in One-way ANOVA.

Click the **Save** button of the original dialogue box to see the following:



If you are interested in testing normality assumption it is essential that you save some sort of **Residuals**. I suggest that you save the **Studentized** residuals. This will add a new variable to your data set which will be called something like **SRE_1**. It is this new variable that needs to have normal distribution, not the original data!

The final button in this set is the **Options** button which will bring up the following dialogue box:



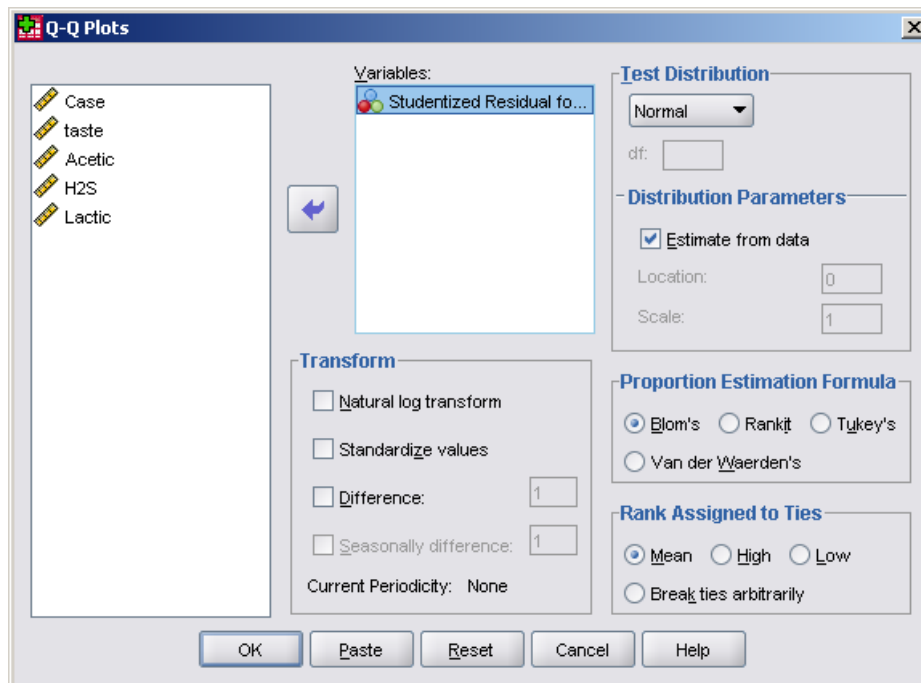
I suggest you specify all the factors and factor interactions in the **Display Means for:** box as you do not know which you will be interested in, until you have the ANOVA table in the output. It is also a good idea to tick the **Homogeneity tests** and **Spread vs. level plot**.

When you have worked through all these dialogue boxes, click the **OK** button.

For Further information: Field chapters 10 and 9 (yes in that order). Chapter 10 is pp 389-426, chapter 9 pp 363-388

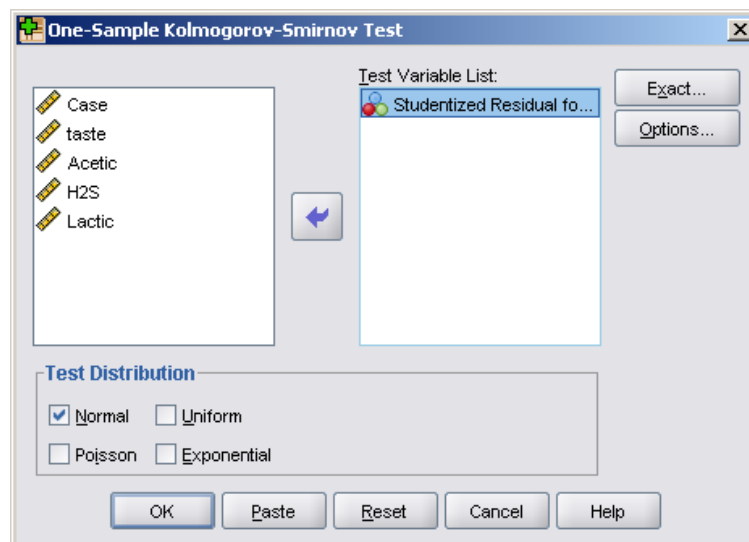
3.2 Checking that the residuals are normally distributed

If you have followed the previous instructions you will have a new variable called something like **SRE_1** added to your data set. As I have said earlier: it is this variable that needs to be normally distributed. The preferred method by most statisticians is to use a normal probability plot. To do this, click the **Analyze** menu and select **Descriptive Statistics** then **Q-Q plot**. You will see the following dialogue box:



You can ignore most of this dialogue box. You need to put the calculated residuals (**SRE_1**) in the **Variables** section and, because these are **Studentized residuals**, they should automatically have a **Location** of zero and a **Standard deviation** of 1. You can remove the tick from the box labelled **Estimate from data**.

However some people are not satisfied with plots! In order to do a test, click the **Analyze** menu, select **Non-parametric Tests** then select **1 sample K-S** to bring up the following dialogue box:



You need to specify the **Test Variable** that contains the **Studentized Residuals** (e.g. **SRE_1**) and click the **OK** button.

3.3 Why Statisticians Prefer Graphs to Test for Checking Normality of Residuals

You rarely see statisticians actually using a Kolmogorov-Smirnov test or quoting it in published papers. This is because the test is sensitive precisely when it does not need to be. So far I have implied that it is the residuals that need to be normally distributed; in fact what needs to be normally distributed is the estimates with the fixed effects removed. These have a variance related in size to the residuals, but there is a theorem in statistics that says approximately that the more cases you have per fixed effect, the closer the distribution of the estimate is to normal. This implies that the more cases you have the less you need to bother about the residuals being normally distributed. The Kolmogorov-Smirnov test also becomes more sensitive (i.e. you can detect smaller departures from the normal) as the number of cases increases. Therefore we tend to prefer to eyeball the graphs and make our own decision.

Here are some rules of thumb based on the ratio of degrees of freedom used in the model to residual degrees.

- If the ratio is > 0.2 then you should cite other studies to establish normality of the data. It is going to be rare for either Kolmogorov-Smirnov or a Q-Q plot to be informative.
- If the ratio is < 0.2 and is > 0.05 then do both the Kolmogorov-Smirnov and the Q-Q plot.
- If the ratio is < 0.05 then do a Q-Q plot and only consider transformations when there is a very clear deviation from a straight line.

3.4 Exercise

Demonstration

Yeast Percentage.sav is the data collected from a designed experiment, ANOVAs often are the analysis of experiments. The idea is to determine how the two factors affect the amount of yeast grown. The factors are the amount of glycerine at 10 mg, 20 mg or 30 mg and how the speed the solution is agitated 10, 20 or 30 revs per minute. There are five replications of each combination.

1. Perform the full Factorial ANOVA on the data and decide whether the interaction term is important. What is the f-value, degrees of freedom and p-value of the interaction term?

F-value	DF (top)	DF (Bottom)	P-Value

2. Repeat the ANOVA and this time get out the differences in the mean value for the important factors and carry out an appropriate and save the studentized residuals. What are the average values (and confidence interval) for each speed?

Speed (revs per min)	Mean	Lower bound CI	Upper bound CI
10			
20			
30			

3. Carry out a Q-Q normality plot on the studentized residuals and a Kolmogorov-Sminoff (K-S) test to look at the normality of the residuals. What is the approximate size of the largest deviation from the normal assumption?

What is the p-Value of K-S test?

Would you conclude from this evidence that the data was a good enough approximation to normally distributed?

Yes

no

Continuing exercises

Take one of the following data sets and carry out an ANOVA for your self

- Oil Viscosity Experiment** is data from a study looking at the viscosity of a liquid produced when different proportions of oil and filler are added. Oil is added at 0, 10, 20, and 30 ml while the filler is added at 0, 12, 24 36 48 and 60 mg into 100 ml of water. You need to answer whether the interaction is important, what is the effect of oil and filler and whether the data is normally distributed.
- University Graduate Salaries** are the salaries graduates obtain in their first employment following graduation. The questions is how the salaries vary with time. However salaries may also be affected by what their degree was in. Also do males differ from females and if so how? This data though all the explanatory variable are categorical, the data is unbalanced (this is not an experiment) and there is a good chance that the salary might need a log transformation.

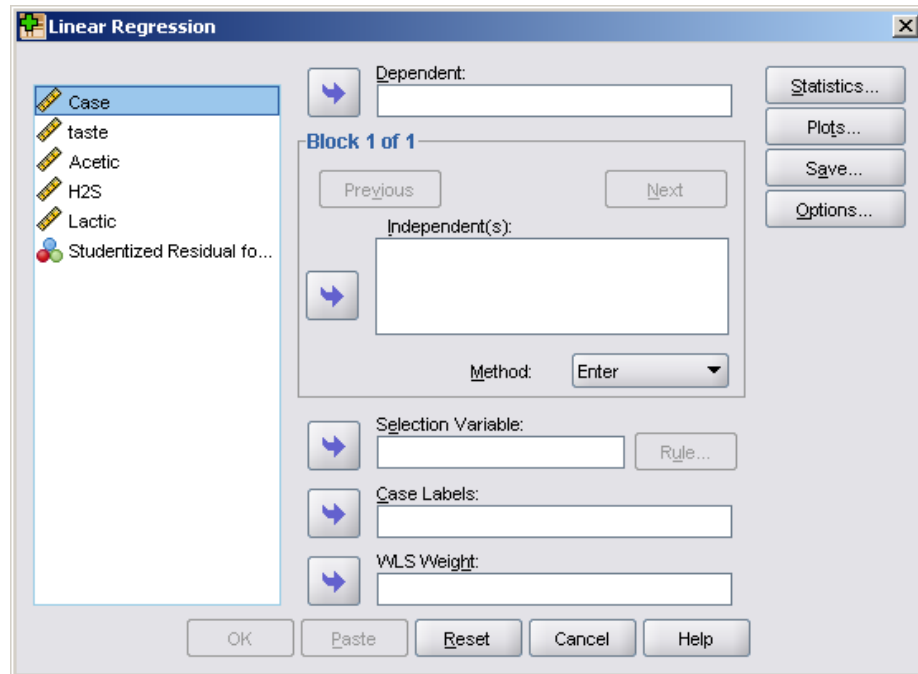
3.5 Linear Regression

Use for: Sorting out relationships between a continuous dependent variable and continuous (or binary) explanatory variables.

Limitations: It is assumed that once the model is fitted, the residuals are normally distributed from a single normal distribution. It does not handle categorical variables that have more than two categories at all well; indeed the only way to deal correctly with non-binary categorical variables is to create dummy variables and use those in the analysis. It is also

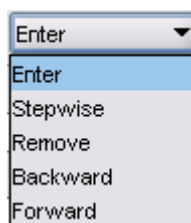
assumed that none of the explanatory variables (independent) are highly correlated with each other. Another rule of thumb is you need at least five cases for each explanatory variable and preferably twenty.

To do a Regression click the **Analyze** menu, select **Regression** and then select **Linear**. This will bring up the following dialogue box:



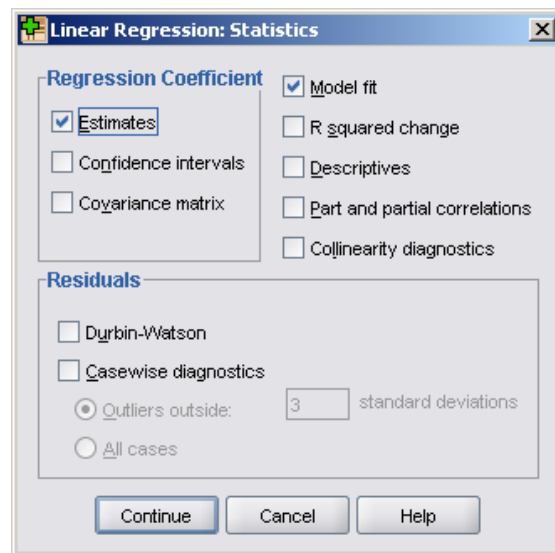
The dependent variable is the variable that you want to explain, and should be continuous. It is the residuals of this that need to be normally distributed. The independent variables or explanatory variables

The **Method** drop down menu gives you the following options:



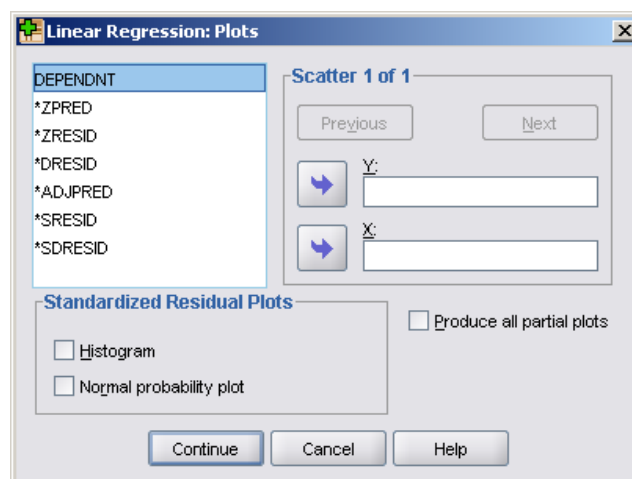
If you know the model you want to fit then select **Enter**. There are three model selection techniques: **Stepwise**, **Backward** and **Forward**. **Forward** method starts with no terms (explanatory variables) in the model and at each step adds the most significant term until there are no more significant terms. **Backward** method starts with all terms in the model then deletes the least significant term until there are only significant terms in the model. **Stepwise** checks both the possibility of putting a term in or removing a term and chooses the best at each step. That leaves **Remove** which is used on the second or third step to remove already entered terms to calculate the significance between the model with them in or with them removed.

If you click the **Statistics** button you will see the following dialogue box:



I suggest that you at least tick the **R squared change** and the **Confidence intervals** as these are often terms that you wish to report. If you are suspicious that some of your explanatory (independent) variables are strongly correlated you may also wish to have the **Collinearity diagnostics** and the **Part and partial correlations**. When you have made your selection click the **Continue** button.

Click the **Plots** button or the original dialogue box to see the following:



I suggest that you tick the box labelled **Normal probability plot** at least. For a scatterplot a useful plot would be the Studentized Deleted Residual **SDRESID** against the Adjusted Predictor **ADJPRED**. When you have selected your plots then click the **Continue** button. The **Save** and **Options** buttons are for the specialist use and a basic user is unlikely to want to change any of their current settings. So you can now click the **OK** button.

For further reading: Field Chapter 5 pp143-217

3.6 Exercise

Demonstration

American wages.sav looks at the hourly rate of pay in specific areas. There are questions on how the other income affects the rate of pay people will accept. The dependent variable is rate.

1. Carry out a regression looking at the average yearly earnings of spouse (ERSP), average yearly earnings of other family members (ERNO), average yearly non-earned income and average family assets holdings. How many of the terms entered are significant?

2. Carry out a forward stepwise regression to put in only the significant explanatory variables into the model. How many models does the analysis take?

Are the terms the same as those significant in the other model?

Yes

No

Compare the estimates from the forced entry and the forward model selection process

Term	Fixed Entry	Forward selection final model
constant		
ERSP		
ERNO		
NEIN		
ASSET		

3. It is suggested that the hours worked, the racial makeup, number of dependents, age of respondent and the level of education all had marked effect on the accepted hourly rate. To do this put these in by forward regression and then use backward to eliminate non significant crucial terms. Also save the studentized residual.

Term	Fixed Entry	Forward selection final model	Final model allowing for demographics
constant			
ERSP			
ERNO			
NEIN			
ASSET			

What do you notice about average age?

Carry out a Q-Q plot on the residuals. Are you happy with these as normally distributed?

Yes No **Have a go yourself**

- 3. Cheese.sav** is data on the taste of cheese. The researchers are interested in what levels of Acetic Acid, Hydrogen Sulphide and Lactic acid make for the best taste. A starting point for the regression could be a straight forward enter of all variables, but consider plotting taste against each of the levels!
- 4. Car Time.sav** simple question, do men spend more time in their cars than women? Gender is binary which means it does not need recoding to be a dummy variable. Does age or extroversion affect the result at all
- 5. Clean Air.sav** is based on Boston. The question is does clean air have any effect on the median price of the houses in an area. The following may also affect the price of houses in the area
- per capita crime rate
 - proportion of residential land zoned for lots over 25000 sq ft
 - proportion of non-retail business acres per town
 - Charles River dummy variable (=1 if tracts bound by river)
 - number of rooms per dwelling
 - Proportion of owner occupied units built prior to 1940
 - Weighted distance from 5 Boston employment centres
 - Index of accessibility to radial high ways
 - full value property tax rate per £10000
 - Pupil-Teacher ratio by town
 - $1000(BK-0.63)^2$ where BK is the proportion of blacks by town
 - % lower status of the population

Choose a sensible model from these prices and then look at the affect of clean air on the median price of houses in the district

3.7 But I have both categorical and continuous explanatory variables should I use Anova or Linear regression

There is not a right answer here. Indeed SPSS is following tradition and hiding the fact that, when it comes to basic formulae, ANOVA and Linear Regression are identical. So at this stage there become two approaches that will cross the divide. If the dominant explanatory variables are categorical then go with ANOVA and use the Covariate box to fit the continuous variables.

However if your dominant variables are categorical then use the Linear Regression but you will have to recode your categorical variables to a set of binary dummy variables. You always create one less dummy variable than the number of categories. There are several ways of doing this coding. If you have a nominal variable then simply choosing a reference category and then setting up variable for every other category works. So if you have the variable of hair colour and the options blonde, auburn, brown, and black then you can code as follows:

Step 1: take brown as reference category (default)

Step 2:

Hair colour	Has Blonde hair	Has Auburn hair	Has Black Hair
Blonde	Yes (1)	No(0)	No(0)
Auburn	No(0)	Yes (1)	No(0)
Brown	No(0)	No(0)	No(0)
Black	No(0)	No(0)	Yes (1)

If they have not got blonde, auburn or black hair then they must have brown!

However some categorical data is ordinal, so you get something like exam grades of A, B, C, D, and E. Then as there is an order to them you might want to reflect this in the grading. In which case something like the following works well:

Grade	AvsBCDEF	ABvsCDEF	ABCvsDEF	ABCDvsEF	ABCDEvsF
A	1	0.5	.333	.25	.2
B	-.2	0.5	.333	.25	.2
C	-.2	-.25	.333	.25	.2
D	-.2	-.25	-.333	.25	.2
E	-.2	-.25	-.333	-.5	.2
F	-.2	-.25	-.333	-.5	-1

If you go into ANOVA and look under contrast the first coding is simple and the second example coding is difference.

You can create such variables using the compute dialogue or using the recode dialogue as it is only a matter of assigning values to specific categories. Both these codings exist as contrasts in Anova. The categorical one is simple and the ordinal one is called the Helmert.

3.8 Exercise

Demonstration

This is using a dataset that comes with SPSS. It is probably in the file called:

C:\Program Files\SPSSInc\SPSS16\Samples\cars.sav

It is data which gives the fuel consumption with an number of measures about the car. One of these measures is the country that the car is made in.

1. Open Car.sav
2. Calculate the dummy variables. As the data set is American in origin most of the cars are American. Therefore it is sensible to make American the default category. To do this using the compute dialog
3. Run a backwards regression and see how big a difference having a Japanese or European car makes

How many extra miles do you get if you choose a European car rather than an American?	
How many extra miles do you get if you choose a Japanese car rather than an American?	

- 4.

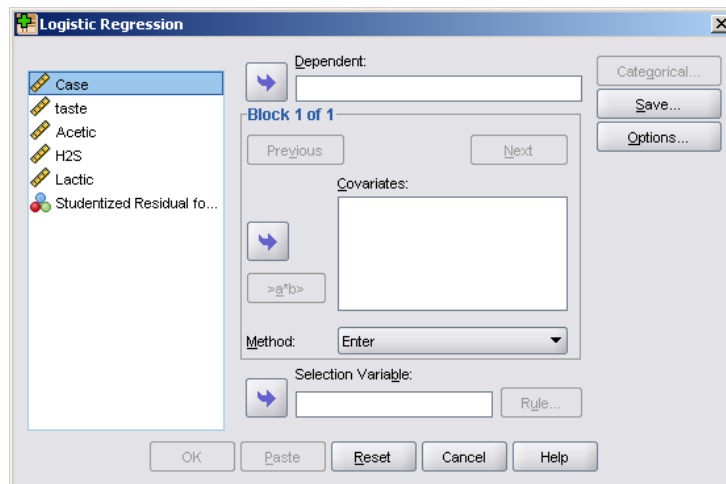
3.9 Logistic Regression

Use for: Sorting out the relationship between a binary dependent and categorical or continuous explanatory (independent) variables.

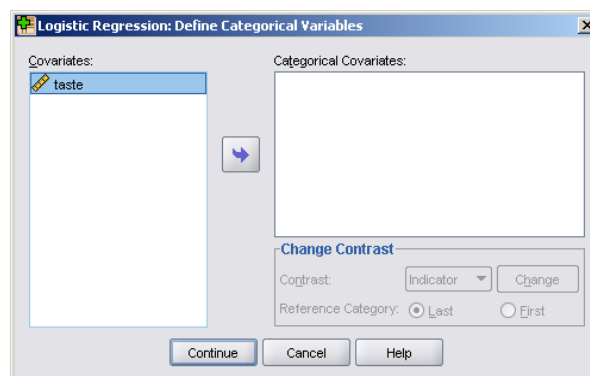
Limitations: Logistic Regression does not work well when any of the explanatory variables (independent) are highly correlated with each other. For logistic regression to give good estimates you need at least 10 cases of your less frequent outcome per variable. Also the proportion of cases should be between 5% and 95% in order to get reliable estimates. If your data does not conform to these limits then you need to use a package like Cytel's LogXact.

Please note, Logistic Regression was a much later addition to SPSS than any other procedure covered up to now. Although many things will seem very similar some things will differ markedly.

To do a Logistic Regression, click the **Analyze** menu, select **Regression** then select **Binary Logistic**. You will see the following dialogue box:



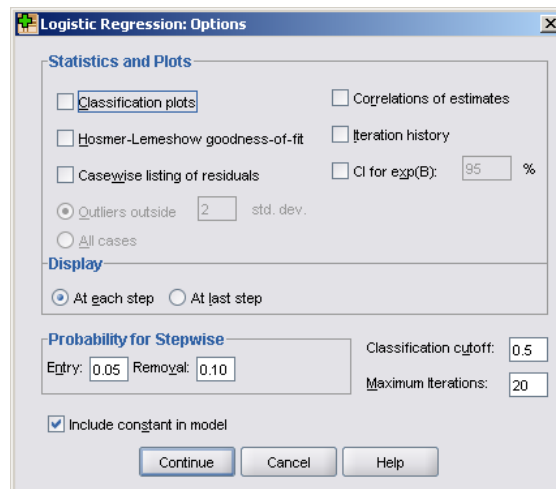
Your **Dependent** variable should be a binary variable. That means that you can rephrase it as a simple yes/no question and get all the data. In this case it is whether somebody has passed or not when sitting an exam. In the block you need to put the appropriate covariates for your analysis, in this example score on a preliminary test, which group they were assigned to and how many weeks of experience they had. Note that these are both continuous and categorical. Binary Logistic regression handles both sort of variable but you have to tell the program which variables are categorical. To do this you press the **Categorical** button to bring up the following dialogue box.



You need to take the categorical explanatory variables into the **Categorical Covariates** box. When you have done this, click the **Continue** button.

The change contrast in this dialogue box is useful when you want to test specific things. I advise you to either look them up in the manual (an extra CD which you can get from CiCS for £2) and also read about **dummy variables** in **Field pg 208** as contrasts are already-calculated dummy variables. A good introduction to contrasts is found the first chapter of **Multivariate Analysis of Variance and Repeated Measures** by D. J. Hand and C.C Taylor, available through Google books.

Click the **Options** button to see the following dialogue box:



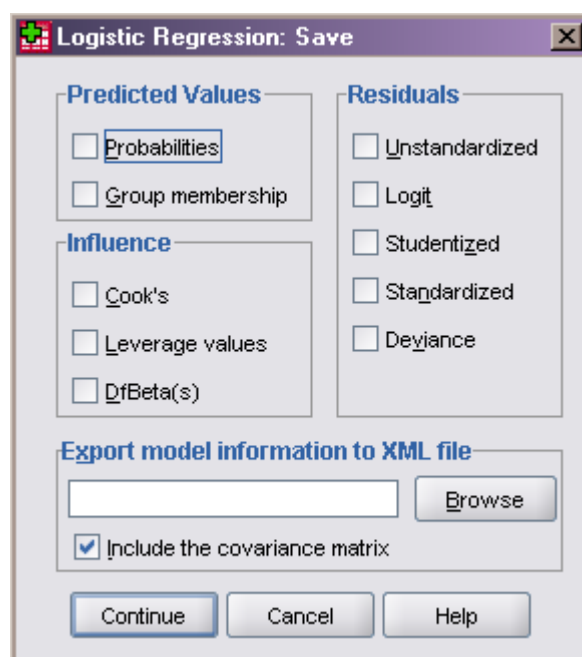
I suggest that you tick the **Hosmer-Lemeshow goodness of fit** box and **CI for exp(B)**. The first gives an idea of how good a fit the model is, the second is the confidence interval of the odds ratio! For some reason SPSS has stuck with the mathematical formula instead of the commonly used name.

For further reading: see Field Chapter 6 pp 218-268

Residuals in Logistic Regression

The first thing to realise is that logistic regression does not use assume a normal distribution, rather it assumes a binomial distribution. Thus the simplistic approach of using the residuals by calculating them the same way as for anova and regression does not work.

So ignore the top four in the residuals box in the save file



Deviance appear to be very similar to the Anscombe residual which is defined as “as normal as possible”. Therefore if you want to take check distributional assumptions then you need to select these. Then you can check for normality as with ANOVA and linear regression.

Warning: if your dependents are purely categorical then it is highly unlikely that the residuals will look at all normal. They will instead have quite a strong step pattern as the values they can take is restricted.

3.10 Exercise

Demonstration

Chronic Heart Disease is a study into indicators of the likely of someone having chronic heart disease. The main question is whether chronic heart disease is related to kidney disease. Kidney disease would be indicated by high levels of serum creatine. However before we get there we need to allow for other known indicators of heart disease. Raised Catecholamine is an indicator of stress levels and would tend to raise blood pressure.

1. Carry out a Logistic Regression to see if creatine is related to chronic heart disease. What is the p-value for Serum Creatine

2. Carry out an analysis to see if adjusted for other explanatory variables Serum Creatine is significant. What now is the p-value for serum creatine?

3. Finally use forward LR to see if diastolic or systolic blood pressure with raised level of catecholamine are significant indicatory variables.

Interaction term	Yes	No
Catecholamine and SBP	<input type="checkbox"/>	<input type="checkbox"/>
Catecholamine and DBP	<input type="checkbox"/>	<input type="checkbox"/>

What now is the p-value now for serum creatine?

4. Finally fit the model that seems most useful using enter, and then check whether serum creatine is significant also get the confidence intervals for the Odds ratio (ExpB). What now is the p-value for serum creatine?

What are the odds ratios with their confidence intervals

	Odds Ratio	Confidence interval
Age		
Raised Catecholamine (Cat)		
Cholesterol level		
Has smoked		
Diastolic Blood Pressure (DBP)		
CAT by DBP		
Serum Creatine		
Constant		

Having a go yourself:

1. Pass Teams.sav looks at the pass rate of job trainees. Each trainee has set a test prior to training and is assigned to a team (red, green, blue), when the team thinks a trainee is ready they put him/her in for the test which is a simple pass or fail. Therefore the trainees have different amounts of experience. Create a model and find out what effects the pass rate of a candidate.
2. Obesity.sav is looking at risk factors that are associated with being overweight. Obesity is a y/n variable and the risk factors are:
 - Sex
 - Systolic Blood Pressure
 - Diastolic Blood Pressure
 - Serum Cholesterol
 - Age in Years

Please note that BMI is used in the calculation of whether a person is obese. So it will predict obesity rather too well to be used in the model.

4. Interacting with a Word Document

4.1 Putting Data Tables in Word Documents

To copy an SPSS table into a Word table, select the table in the Output window and choose **Copy** from the **Edit** menu. Then go to Word and use **Paste** from the **Edit** menu.

To copy an SPSS table into Word as simple text, select the table in the Output window and choose **Copy** from the **Edit** menu. Then go to Word and from the **Edit** menu use **Paste Special** then select **Unformatted Text**.

To copy an SPSS table into Word as an SPSS object, select the table in the Output window and choose **Copy Object** from the **Edit** menu. Then go to Word and use **Paste** from the **Edit** menu.

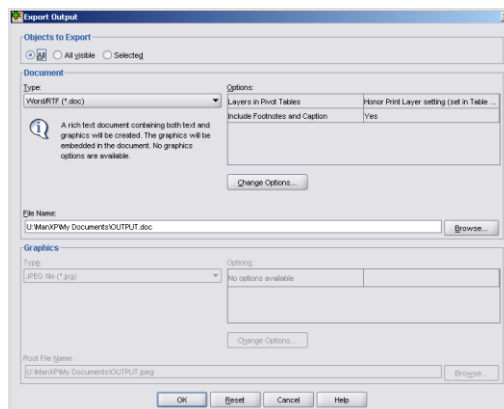
4.2 Putting Graphs in Word Documents

To paste as a simple picture, select the graph and choose **Copy** from the **Edit** menu. Then go to Word and use **Paste** from the **Edit** menu.

To paste as an object select the graph and choose **Copy Object** from the **Edit** menu. Then go to Word and use **Paste** from the **Edit** menu. This is not editable within Word. Interactive graphs can only be pasted as objects.

4.3 Exporting to Web Page and Graphics Files

To do this select **Export** from the **File** menu of the **Output** window. Use the **Export** list to specify objects to export. For table only files you can export to HTML or to plain text.



Use the **File Type** list to choose HTML for data, or a graphics format for charts. Charts can be exported as Windows metafile (WMF), Windows bitmap (BMP), encapsulated PostScript (EPS), JPEG, TIFF, CGM, PNG, or Macintosh PICT.

5. Syntax

You can display the line entry commands required to carry out an analysis or create a graph. Select **Options** from the **Edit** menu, then choose the table labelled **Viewer** and check the **Display commands in the log** box. Once the log is displayed on screen, any analysis you perform will write the equivalent line entry commands into the on-screen log.

A text file that contains the commands to carry out an analysis in SPSS is called a Syntax File. To open one of these go to the **File** menu, select **New**, then select **Syntax**. This is a straight text file and you type in the relevant commands. It can be saved to run again later, which you cannot do with a series of keystrokes.

5.1 Exercises

Demonstration

Open the data set C:\Program Files\SPSSInc\SPSS16\Samples\bankloan.sav.

Go to the compute dialog and feed in the syntax for computing the total debt. Then press “paste” to see the syntax

Go to the recode dialog and put in the coding for those who have graduated from highschool or higher. Press paste button. Copy and paste this syntax so you have it four times. Alter the syntax so we have variables that cover “graduated from highschool” “gone to college” “graduated from college” and “post graduate qualification”. Run this syntax.

Create a scatterplot matrix for the variables in the regression

Get descriptive statistics in a table and hit paste to save this to the syntax file

Do a correlation of “Years with current employer” “Years at current address” “Household income in thousands” the four education variables and the Total debt (you have calculated the last five variables).

Go back to the correlation dialog and this time hit paste. Put the word “with” in the correlation syntax that has been put into your syntax file before the total debt variable. Run the altered syntax.

Create the syntax for a backward regression and do paste.

Save the syntax file

Leave SPSS

Open SPSS

Open the data set C:\Program Files\SPSSInc\SPSS16\Samples\bankloan.sav.

Open the syntax file and re run the analysis

Take the descriptive table and put it into a Word document

Take the scatter graph and put that into a Word document.

6. Further Reading

SPSS Books

Discovering Statistics Using SPSS, Andy Field, Sage Publications, ISBN 0-7619-4452-4

This is a book that is both a statistics text and an SPSS primer. It covers a large number of techniques (all introduced in this course and more) along with the background theory of how they work. For those who want to go further, I have referenced the relevant pages in Field for each statistical technique covered in this userguide.

SPSS Made Simple, Kinnear & Grey, ISBN 0-8637-7350-8, £12.43 for v 15

A good introduction to SPSS, costing about half the price of a single manual. It is written by academics in Aberdeen.

SPSS Guide to Data Analysis, Marya Norusis, ISBN 0-13-020399-8. £44.99

More than a manual, it details why and how to use SPSS for analysing your data. It has become a classic, but costs about twice as much as Kinnear & Grey. If you want a book from SPSS, this is preferable to a manual.

General Statistics Books

Using Multivariate Statistics 5th Edition, isbn 78-0205465255 £41.11

This is a good intermediate statistical text book, it includes syntax for most analyses and consideration on when tests are valid or not. It is more advanced than Field but really does give you a depth of understanding.

The Cartoon Guide to Statistics, ISBN 0-5062731025, £9.99

A basic introduction to statistical thinking.

How to Lie with Statistics, ISBN 0-393310728, £5.99

This book is a good read, even if the closest you get to statistics is reading what someone else has done. It goes through the basic ways that research may be reported to mislead.