

Introduction to SPSS

SPSS is a statistical package commonly used in the social sciences, particularly in marketing, psychology and sociology. It is also used frequently in market research. SPSS has a user-friendly graphical interface, but also allows programming. This document is a tutorial on doing basic tasks in SPSS using the menu-driven interface. It covers material seen in decision science statistics courses (DECS 433, 434), and necessary for the research methods in marketing course (MKTG-450). More advanced topics are covered in class by the faculty member teaching research methods in marketing.

0. Before you begin

To follow the examples in the document, you need the student version of SPSS installed on your laptop. This software is bundled with the textbook for MKTG 450. All the screenshots in this document are based on SPSS 12.0 for Windows Student Version. Alternatively, you may work through the examples using the full-fledged version of SPSS available in the “special software” workstations in the Kellogg computing labs or lease a copy of SPSS (see section 9 of this document). Before you begin, download “**tek_spss.zip**” from the following Web page:

www.kellogg.northwestern.edu/kis/tek/ongoing/spss.htm

This Web page will also include the most recent version of this document. Please extract the files from the ZIP file to a directory of your choice in your laptop. These files contain the data used in the different examples below.

Note that the SPSS student version seems to be a bit unstable – as we were writing this document, the program quit unexpectedly several times. Save your data and output often to avoid losing work if the program crashes. The full-fledged version of SPSS does not present this problem.

1. Reading text (ASCII) data

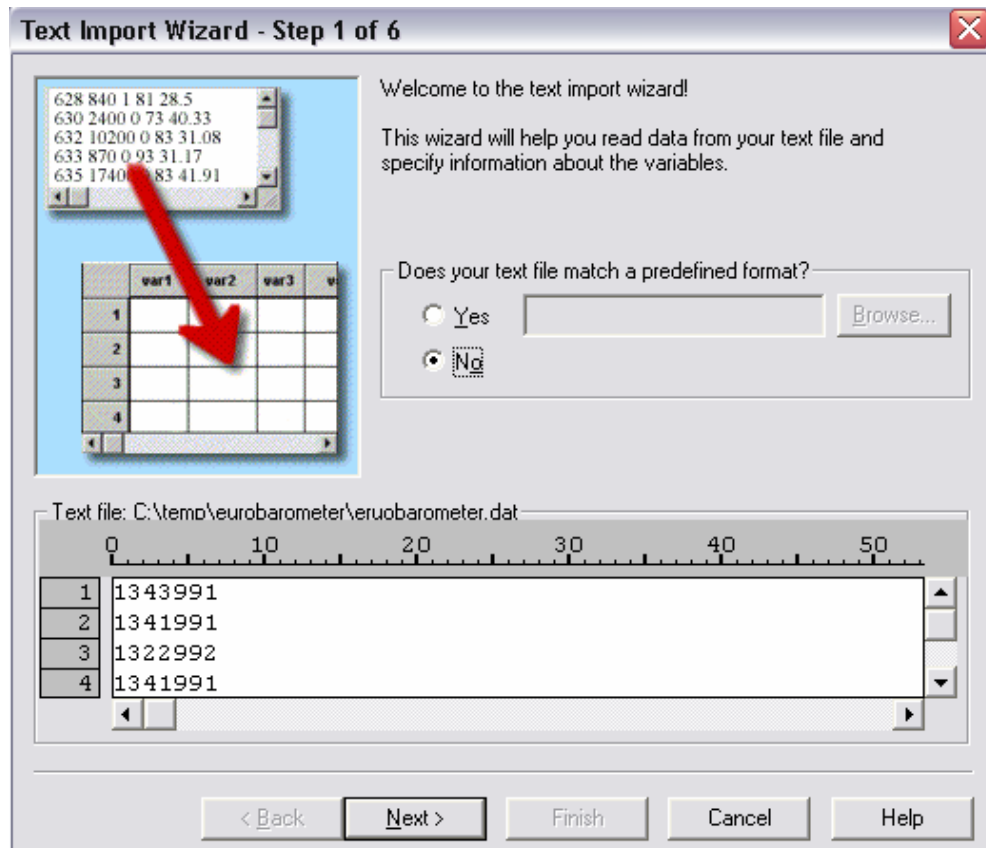
Often times, data is made available in text format, also called “ASCII”, which stands for American Standard Code for Information Interchange, which defines the character coding that computers understand. The researcher needs to read the file into a package such as SPSS, making sure the variables included in the file are defined according to specifications usually provided with the data file. The file specifications are typically in a “codebook”. Text files are convenient for two reasons: they can be read by different programs and they are small. Data files in binary

format created by a given program (such as SPSS or Excel) tend to be larger due to the “overhead” in space created by the definition of the data: an SPSS data file will include not just the values of each variable, but also the names of the variables, the type (numeric or alphanumeric), and any associated variable and value labels.

The two most frequent types of text data files are files in “fixed” and “delimited” formats. The typical layout for both is to have variables in columns, with one observation per row. In a fixed format, each variable takes a certain number of spaces in the file: for example, a variable that takes values between 0 and 1,000 would take 4 spaces or columns in the file. In delimited files, by contrast, the values of different variables are separated by a specific delimiter, such as a comma (e.g., in comma-separated values or CSV files) or even a pipe (“|”).

1.1. Reading a fixed ASCII format file

The next few examples are going to use a subset from a Eurobarometer survey collected between October and November 2003. Open SPSS and cancel the dialog box that opens when you launch the program. From the File menu, select “Read Text Data.” Locate the “eurobarometer.dat” file in the directory in which you extracted the files from “tek_spss.zip.” Once you select the file, SPSS will open the “Text import wizard”, which will walk you through the steps of reading the data into SPSS format. As you see in the screenshot below, you need additional information to parse the different variables included in the file:



The following table summarizes information from the original dataset’s codebook:

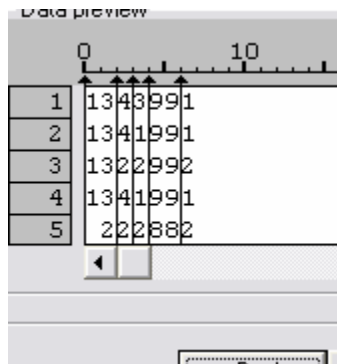
Exhibit 1: Codebook of Eurobarometer sample file

Variable name	Question	Column location	Possible responses and codes for missing values
country	What is your country of citizenship?	1-2	2=Cyprus 13=Turkey
q1	On the whole, how satisfied are you with your life in general? Would you say you are...?	3	1= "Very satisfied" 2="Fairly satisfied" 3="Not very satisfied" 4="Not at all satisfied" 8="dk / no opinion" 9="refusal/na" Missing values: 8, 9
q12	Would you say you are very proud, fairly proud, not very proud or not at all proud to be [NATIONALITY - refer to citizenship]?	4	1 "Very proud" 2 "Fairly proud" 3 "Not very proud" 4 "not at all proud" 7 "Does not feel to be [Cypriot/Turkish]" 8 "dk / no opinion" 9 "refusal/na" Missing values: 7, 8, 9
d4	What is the year of your birth? (Code last two digits of the year)	5-6	Missing values: 99
d14	Sex of respondent (Do not ask – mark appropriate)	7	1="Male" 2="Female"

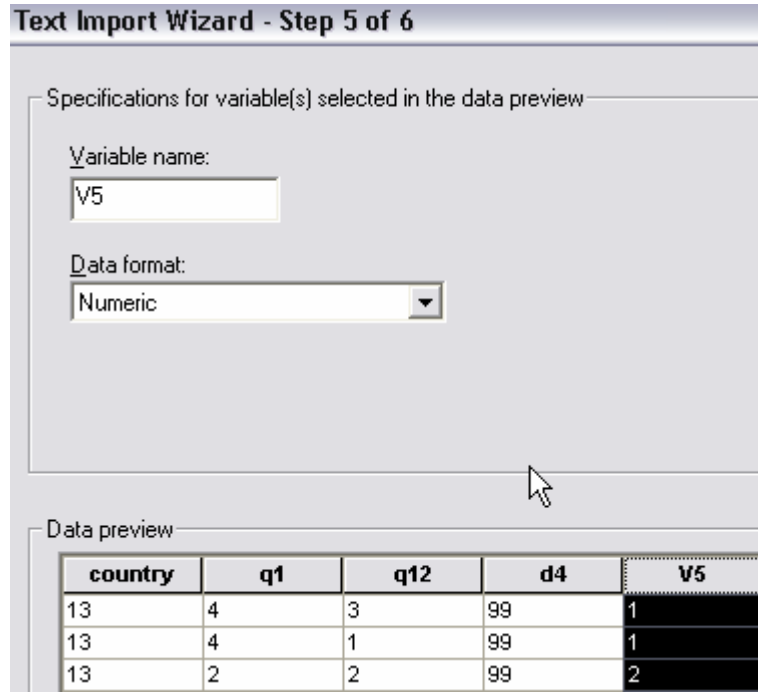
In the SPSS Text Import Wizard, click on “Next” to get to **Step 2**. Under “How are your variables arranged?”, select “Fixed”. Under “Are variable names included at the top of your file?” select “No” and click on “Next”.

In **step 3** of the Text Import Wizard, the default options should be to start at row 1 as the first case (observation), with each line representing one case. We are going to read all the cases. Click on “Next.”

In **step 4** of the Text Import Wizard, you will need to define the variables by clicking between them – a click between columns 2 and 3 will insert a line. If you make a mistake, point to the arrow in the preview, drag it beyond the ruler, and release the mouse button:



Click on “Next” to get to **step 5**. In this step, you will click on each of the columns to name the variables: country, q1, q12, d4 and d14. Note that variable names can be up to eight characters long.



Click on “Next” – you do not need to save the format for future use. Click on “Finish” to import the data. Check to make sure you have 1500 cases. The data will be displayed as a spreadsheet. From the File menu, select “Save as” and save the file as an SPSS data file (call it “eurobarometer.sav”). Note that the “Save as” dialog box includes a “Variables” button. This allows you to select the variables that you want to save.

1.2. Reading delimited ASCII data

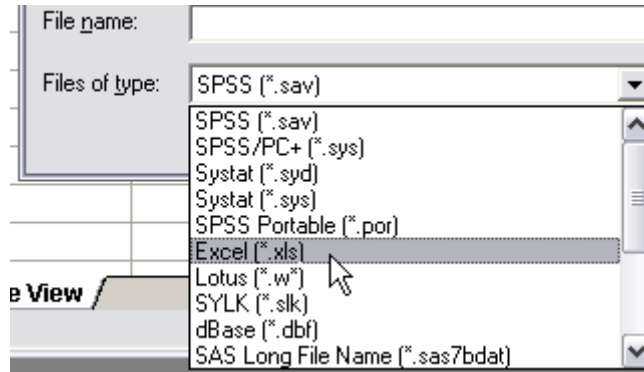
The procedure to read delimited ASCII data is the same as above (from the File menu, select “Read Text Data”) except that in **step 2** of the Text Import Wizard you need to specify that the file is delimited and whether the first row in it contains variable names. **In step 3**, the first line of the file in which there is data will adjust according to your answer regarding the availability of variable names in the first row. Finally, in **step 4**, you are able to specify the delimiter. Make sure to have only “Comma” selected as the delimiter.

To try reading this type of file, import “compustat_sample.csv” into SPSS. This file is comma-delimited and its format is equivalent to those created as a result of web surveys using SurveyZ or ViewsFlash. For more information about web surveys, point your browser to:

www.kellogg.northwestern.edu/kis/websurveys/

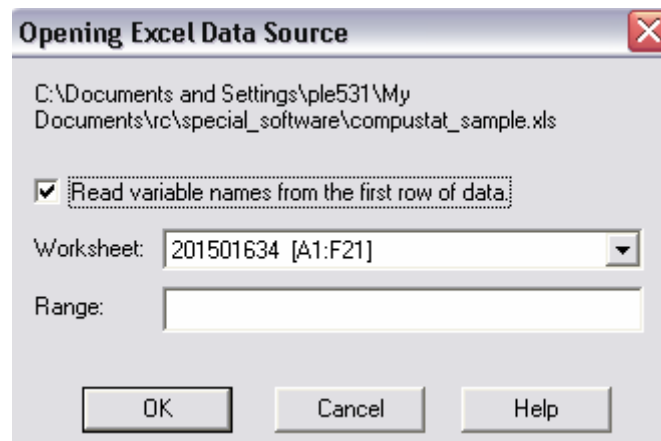
2. Importing data from Excel

To import data from an Excel spreadsheet, select “Open Data” from the File menu. In the Open File dialog box, select “Excel (*.xls)” under “Files of Type.”



Among the sample files for this session, point SPSS to “**compustat_sample.xls**”. Once you click on the “Open” button, SPSS will prompt you to select the worksheet within the file, as well as whether to read the variables in the first row of data. You may also specify a particular range of cells within the worksheet, in the same way you would specify a range within Excel. For example, you could retrieve only the data included in “A12:F21”.

Note that SPSS will read variables in the first row relative to the range you selected. When you specify range A12:F21, SPSS will use the contents of row 11 to name the variables. Hence, you need to be careful if want to use this feature and unselect the “Read variable names from the first row of data” box if the variables names are not in the first row of the range you specified.



It is important to keep in mind that SPSS expects cases or observations to be represented in rows, while the columns are variables.

3. Computing and recoding data

3.1. Defining missing values, variable and value labels

Going back to the SPSS data file we created in section 1.1, open “eurobarometer.sav” by choosing “Open Data” from the File menu. Once you have opened this dataset, notice that in the bottom of the SPSS window there are two tabs, labeled “Data view” (the default, where you see the values that each variable takes for each case), and “Variable View”. The latter shows the attributes for each one of the variables in the file, and it allows you to create variable and value labels. Variable and value labels are useful when you are working with the data, as well as when you produce output from analysis with SPSS. In addition, this view allows you to define “user missing values”.

User missing values (as opposed to a “system” missing value, which is simply the lack of data in a particular variable for a given case/observation) are values that have been defined in a survey for individuals who did not respond¹. For example, in **Exhibit 1** (page 3 of this document), for question 1, values of 8 and 9 would be defined as user-missing. Missing values are typically excluded from various statistical procedures by the software (KStat, in contrast, does not deal with missing values easily).

	Name	Type	Width	Decimals	Label	Values	Missing
1	country	Numeric	2	1		None	None
2	q1	Numeric	1	0		None	None
3	q12	Numeric	1	0		None	None
4	d4	Numeric	2	1		None	None
5	d14	Numeric	1	0		None	None

Variable labels can be up to 256 characters long, providing a description of the variable (versus the variable name, which is a mnemonic of up to 8 characters).

Value labels can be up to 60 characters long, and their purpose is to provide information about the meaning of each category in an ordinal variable. For example, variable d14 takes values of 1 (“Male”) and 2 (“Female”). You can relate the values (1 and 2) to labels (Male and Female) using the Variable View in SPSS.

¹ System missing values show up as a dot (“.”) in SPSS’ data view. User-missing values show up as the number that originally represented it. Different user-missing values allow inclusion of categories outside the normal range of values (for example, respondents who marked two answers or perhaps wrote down an additional answer) and allow an interested researcher to assess the quality of the survey instrument. In other words, the user-missing values convey some information as to why the respondent did not fit in the allowed responses.

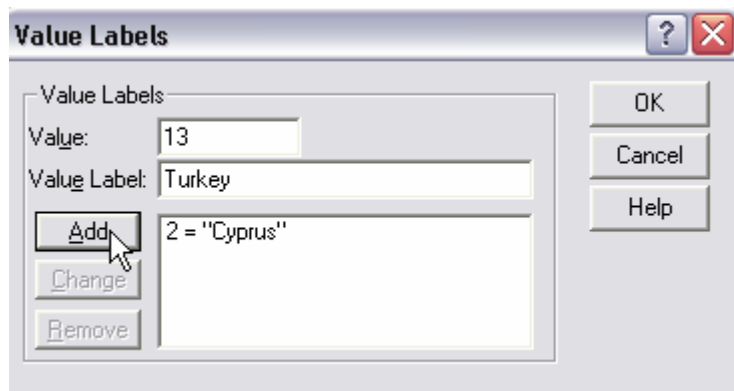
Click on the Variable View tab to define missing values, variable and value labels for our sample file, using the information in **Exhibit 1**. To create the label for the country variable, double-click on the “Label” field for each of the variables and type the corresponding label.

Name	Type	Wid	Deci	Label
country	Numeric	2	1	Respondent's country of citizenship
q1	Numeric	1	0	Respondent's level of satisfaction with life in general
q12	Numeric	1	0	Respondent's degree of pride of being Cypriot/Turkish
d4	Numeric	2	1	Respondent's year of birth
d14	Numeric	1		Respondent's gender

Next, to create value labels, click on the “Values” field for each of the variables in our sample file. Notice that a grey square will be displayed (see screenshot below). Click on it to open the “Value Labels” dialog box.

	Name	Type	Wi	Dec	Label	Values
1	country	Numeric	2	1	Respondent's	None
2	q1	Numeric	1	0	Respondent's	None
3	q12	Numeric	1	0	Respondent's	None

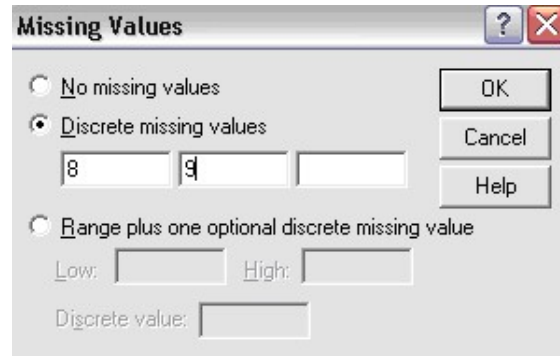
In the Value Labels dialog box, you can link the values of one of your variables with labels that you type in. For example, for the “country” variable, a value of 2 corresponds to Cyprus. In the “Value” field, type “2”; press the tab key to get to the “Value Label” field and type “Cyprus”. Click on the “Add” button on the left to register that value and add the label for 13. Click on the “Add” button again and then on “OK” to close the Value Labels dialog box for the country variable.



Based on the information provided in the forth column of **Exhibit 1**, complete the creation of value labels for the remaining variables in the sample file.

Finally, the process of defining missing values is similar to that of value labels. In the “Variable View” of the data, click on the “Missing” field of a variable. As in the case of value labels, a grey box will appear. In our sample dataset, only variables q1, q12 and d4 have user missing values.

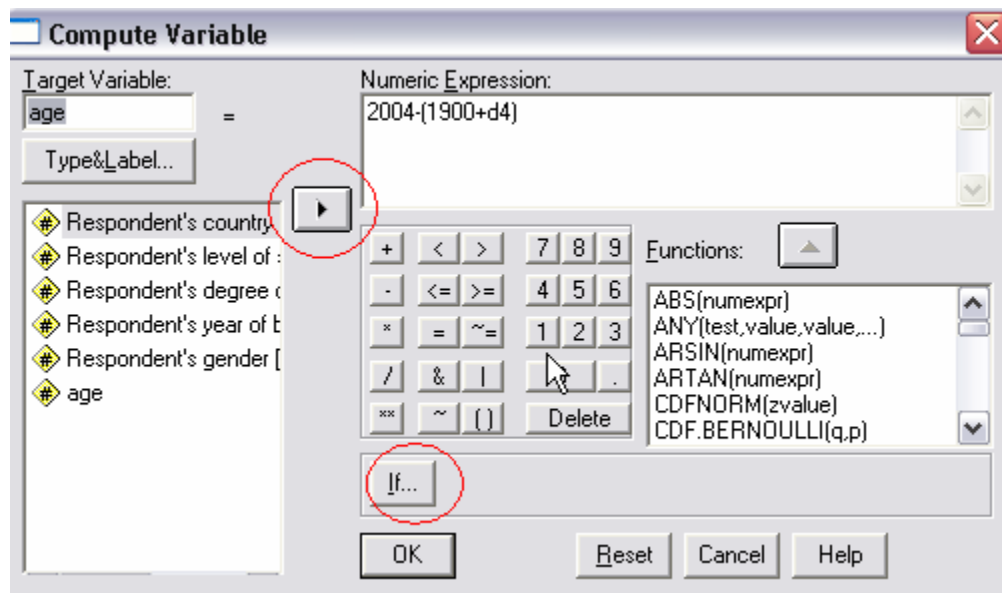
For our sample dataset, it is sufficient to choose “discrete missing values”. For example, for variable q1, the missing values are 8 and 9. Other datasets might have a range of values that can be considered missing (for example, 80 through 90), plus some extra value which is out of the specified range (99). Using the information from **Exhibit 1**, define the missing values for variables q1, q12 and d4. Once you are done, save the dataset to preserve your changes.



3.2. Computing new variables

Suppose that you would like to calculate the age of your respondents as the current year (2004) less their year of birth (variable d4). To do this, select “Compute” from the “Transform” menu. In the Compute Variable dialog box, you need to specify the name for the new variable (“Target variable”) and the expression that will generate the values for it. Note that you can also specify the type and label.

To generate the age variable, type “age” under target variable. In “numeric expression”, type: 2004-(1900+d4). Instead of typing d4, you may also click on the variable name and then on the arrow button (circled below) next to the list of existing variables.



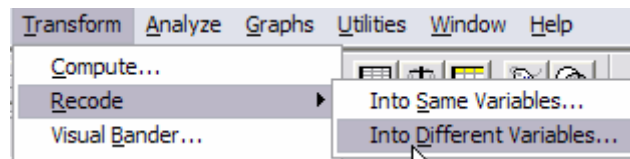
SPSS also allows you to use functions (listed on the right side; use the help files to find out about specific functions); you may also perform the computation for a certain subset of the cases by using a conditional statement (for example, you could calculate the age only for respondents from Turkey). The latter is accomplished by specifying a condition - click on the “If” button (circled in the picture above) to define one. Keep in mind that most of the functions work across variables. For example, the “mean” function works for creating the mean, for each case, of the variables included as argument to the function. To get the mean of a variable, either for all cases or by groups as defined by a categorical variable, use the features available under the “Analyze” menu.

Typically, conditional computations are performed to deal with branching (sometimes called “skip pattern”) in a survey: sets of questions that apply only to respondents that answered a given question in a specific way. Another reason for conditional statements is performing computations depending on the range of values. For example, the level of discount in the price of a product might depend on the volume purchased.

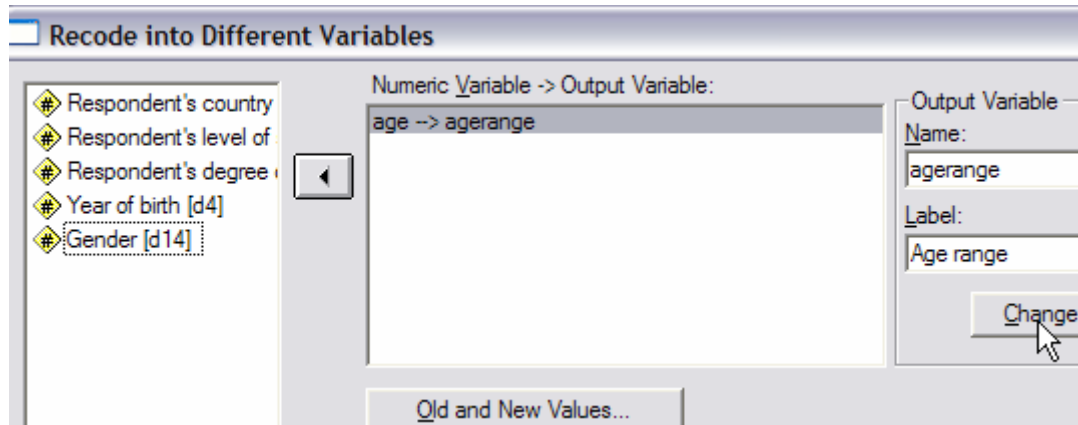
3.3. Recoding variables

There are different reasons for “recoding” a variable. For example, you may want to collapse a variable with 10 categories into just three categories, or, as you analyze the data, you might have found cases that were data entry mistakes. You may recode into a new variable, which is equivalent to creating a new variable, or you may recode into the same variable. The latter is equivalent to overwriting the values of the variable. As a general rule, you should avoid recoding into the same variable unless you are a proficient SPSS user and you have very good reasons for the recode.

Suppose we want to recode the age variable into four categories: teenagers (up to 19 years old), young adults (20 to 44 years old), middle aged (45 to 64 years old) and older adults (65 years and older). To do this, select “Recode into different variable” from the “Transform” menu.



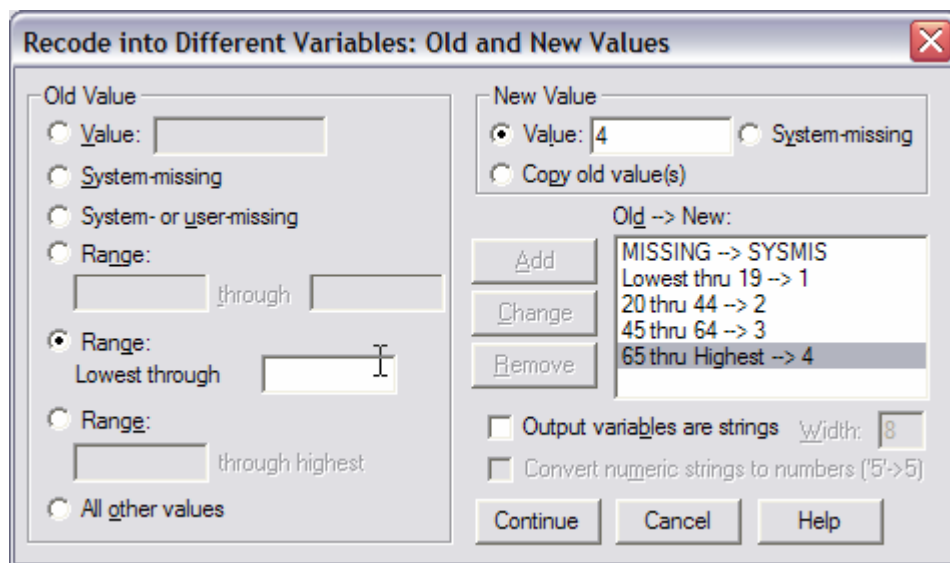
In the recode dialog window, first click on the variable name (age), and then click on the arrow button to move it to the right box. Then click on the “Name” field to give the new variable a name (call it “agerange”) and type a label for it. Click on the “Change” button so that this information is registered. Next, click on the “Old and new values...” button.



In the “Recode into Different Variables: Old and New Values” dialog box we will specify how the values of age map into values of the new variable. First, we will map all system- or user-missing values into system-missing in the new variable. On the left panel, select “System- or user-missing”. On the right panel, click on “System-missing” and then click on the “Add” button.

Next, we will map the ages below 20 into a value of “1”. Click on the second “Range” radius button (lowest through...) and type “19”. On the right panel, click on the “Value” radius button and type “1”. Again, click “Add” to update the “Old -> New” list.

Use the first “Range” radius button to map the 20 through 44 range into a value of 2, and the 45 through 64 range into a value of 3. Finally, use the third “Range” radius button to map ages 65 and up into a value of 4. If you make a mistake, you can click on the value you have added in the “Old -> New” list to remove it. Once you are done, click on the “Continue” button to return to the previous dialog box; click on “OK” to generate the new variable. The next screenshot displays how your Old and New Values should look like when you are done.



Since we only have 1500 cases, it is relatively easy to eyeball the data to check that the recode worked as you expected it to. However, it is best to run descriptive statistics to check this.

With large datasets, this is often times the only way of checking that a transformation worked properly.

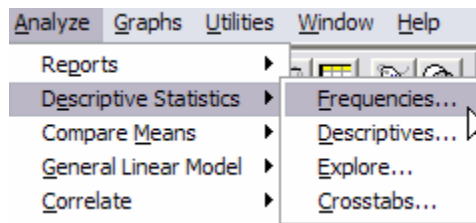
In this particular example, you may want to go back to the “Variable View” and define the appropriate value labels for the newly created variable (section 3.1 above).

4. Frequencies, descriptive statistics and crosstabs

To check whether the transformation worked, the easiest way to verify would be, first, to get a count of the non-missing values for both “age” and “agerange”, and then, get the descriptive statistics of “age” for each of the groups defined by “agerange”.

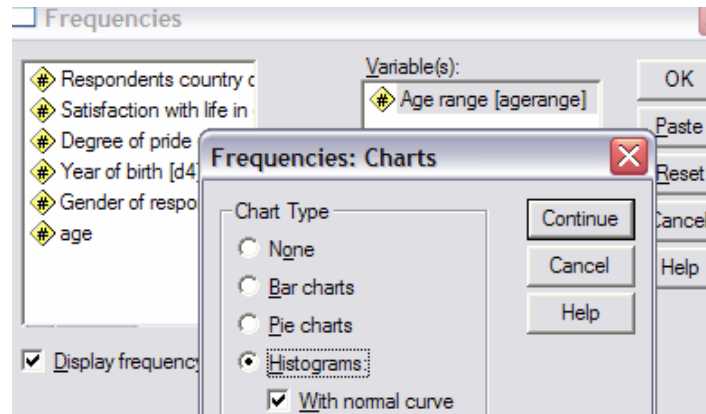
4.1. Frequencies

To get the frequency for each of the values in the “agerange” variable, select “Frequencies from “Descriptive Statistics”, under the “Analyze” menu.



As in previous examples, to select a variable, click on it in the list shown in the dialog box (in this case, the Frequencies dialog box), and then click on the arrow button to select it for analysis. At this stage, you may also select additional variables for analysis (for example, country, q1 or q12). Be careful if you select a variable such as the year of birth or age, since the resulting table will be very long. Make sure that the “Display frequency tables” option is selected.

Click on the “Charts” button and select “Histograms”; also, check the “with normal curves” option. Click on the “Continue” button to return to the previous window, and then click on the OK button to compute the frequencies. SPSS will open a second window, its SPSS Output Viewer. It is in this window that any output will be displayed. The output of the frequencies of “agerange” (except the histogram) has been pasted below. Each of the tables and chart is an object. You can click on it to copy and paste it into a document (the tables can go as formatted text, “rich text format”, into Word, while the chart can be copied as a bitmap). Note that SPSS uses the variable and value labels you have defined.



Hint: If you have a dialog box open in the main SPSS window (Data Viewer), you will not be able to select or edit any object in the SPSS output viewer. The output viewer allows you to do some editing on the objects, as well as add headings. You can also save the entire output as an SPSS output file (a file with extension “spo”).

Statistics

Age range

N	Valid	1496
	Missing	4

Age range

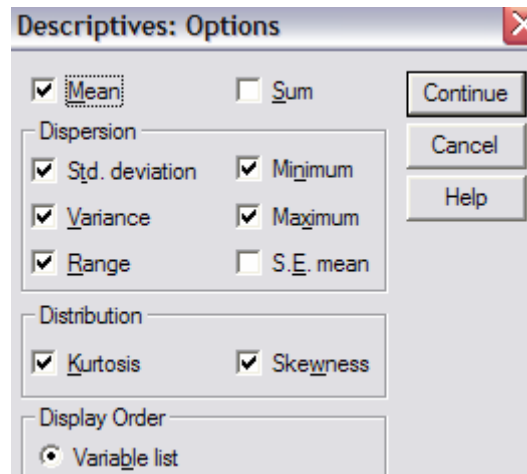
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Teenager	119	7.9	8.0	8.0
	Young adult	774	51.6	51.7	59.7
	Middle aged adult	424	28.3	28.3	88.0
	Older adult	179	11.9	12.0	100.0
	Total	1496	99.7	100.0	
Missing	System	4	.3		
Total		1500	100.0		

Note that the Frequencies dialog window also includes a “Statistics” button. This allows you to select some descriptive statistics that are calculated across all the categories or values. Finally, the “Format” button allows you to specify the order in which the categories are displayed in the output; there is also an option to suppress the table of frequencies if there are more categories than a specified number (the default is 10 categories).

If you examine the results above, you see there were only four missing values out of 1500. This should be the same number of missing values in the “age” variable. Otherwise, it would mean that we missed some range of values when we recoded “age”. While it may seem trivial in this particular case, it is easy to make mistakes when you deal with continuous variables or with categorical variables with many categories.

4.2 Descriptive statistics

To get the descriptive statistics for the “age” variable, select the “Descriptives” option from “Descriptive Statistics” under the “Analyze” menu. Select “age” for analysis and click in the “Options” button to select the desired descriptive statistics. Select the mean, standard deviation, variance, range, minimum, maximum, kurtosis and skewness. Then click on the “Continue” button to return to the previous window.



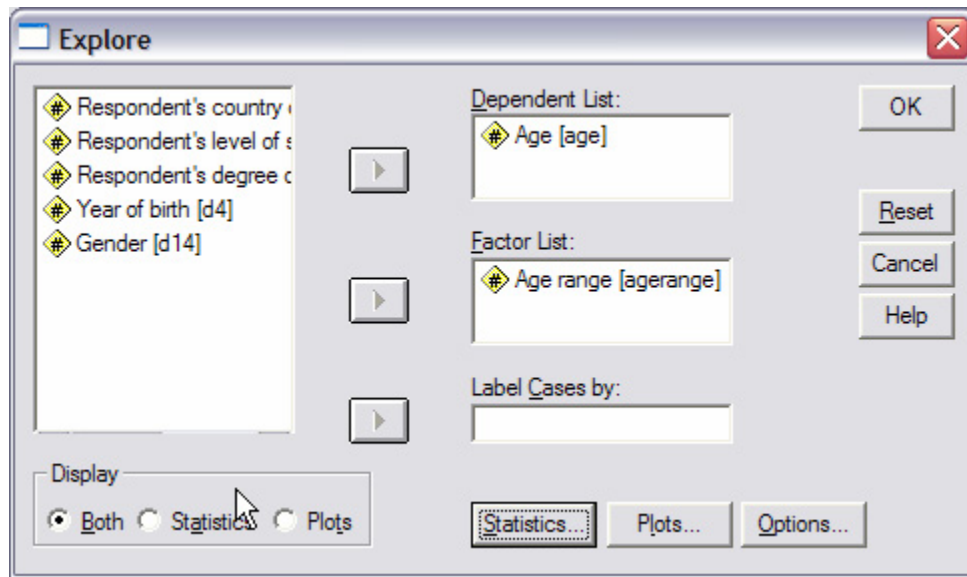
Note that the descriptives dialog box allows you to create and save the standardized values or “z scores” (SPSS subtracts the mean for the variable from each one of the values and divides by the standard deviation) for the variables analyzed.

To produce the descriptive statistics for “age”, click on the OK button. The output will be displayed in the output viewer, following the output of the frequencies analysis. Looking at the output shows you that there are only four missing values for “age”, which means that our recode is probably correct.

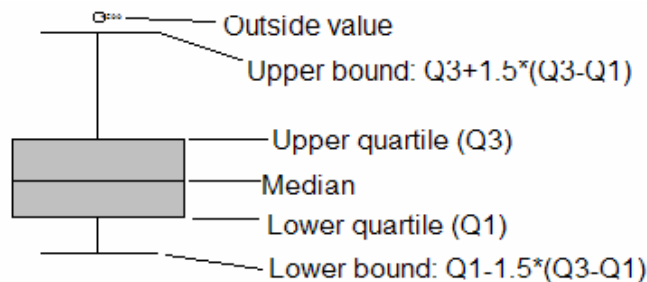
4.3. “Exploring” the data

Finally, you can use the “Explore” feature of “Descriptive Statistics” to get descriptive statistics of age for each of the groups defined by age range. In the “Explore” dialog box, select age as the dependent variable (you can add several variables to this list) and age range as the factor. You can also produce results by more than one factor. SPSS will produce the descriptive statistics and plots for each of the break ups you have specified.

By default, SPSS will produce both descriptive statistics and stem-and-leaf and box-and-whisker plots. Both of these plots (invented by statistician John Tukey in the 1960s and 1970s) are useful tools to visualize the frequency distribution of a variable. *Stem-and-leaf* plots are easiest to understand if you examine the plot corresponding to “young adults” in the output. The stems will be the decades (2, 3 and 4), while the leaves are the years (0 through 9). You can recover the data by pairing each decade with one of the leaves. Looking at the plot, you see that the values are relatively evenly distributed. This is not the case if you look at the stem-and-leaf plot for older adults, where the distribution looks skewed towards the lower age values.



Box-and-whisker plots are another way of looking at the same information. The whiskers (the horizontal lines at the end of the vertical line) are defined in relation to the “inner fences².” The lower inner fence is defined as the lower quartile (“Q1,” the 25th percentile) less 1.5 times the interquartile range (Q3-Q1); thus, the lower whisker is the first value in the distribution above the lower inner fence. The upper whisker is the first value below the upper inner fence, defined as the upper quartile (“Q3,” the 75th percentile) plus 1.5 the interquartile range. Values outside the whiskers are suspect outliers (“outside values”); SPSS displays them with a circle and labels them with a case ID. The box depicts the interquartile range (the range between the 25% percentile and the 75% percentile) of the distribution. These plots are useful to check whether your data is skewed and to see whether there are potential outliers.



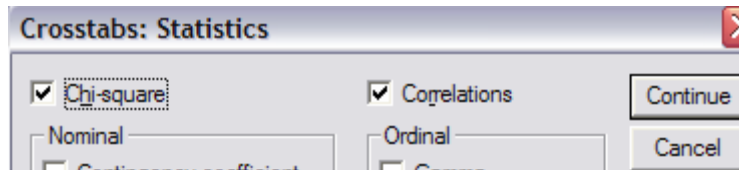
4.4. Cross tabulations

Finally, you may want to create a contingency table to visualize the relationship between two categorical variables such as country and the level of satisfaction with life in general (variable q1). Select “Crosstabs” from “Descriptive Statistics” under the Analyze menu.

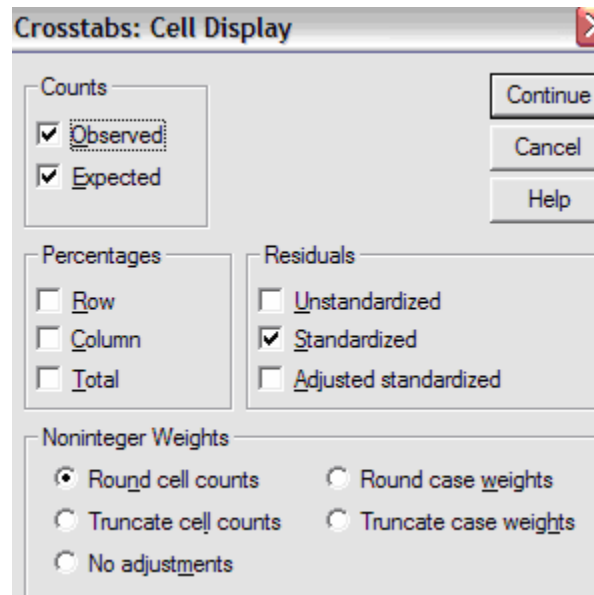
² Tukey also defined “outside fences.” The definition is like that of the inner fences, except that the interquartile range was multiplied by 2 instead of 1.5. Values beyond the outside fences were called “far out” values.

In the Crosstabs dialog box, select variable q1 (level of satisfaction with life in general) as the variable of analysis in rows, and select country as the columns. SPSS also allows you to add an additional layer, such as gender. What this does is to generate a cross tabulation of q1 by country for women and one for men. Make sure the “Suppress tables” option is not checked.

The “Statistics” button allows you to select a number of measures of association between categorical variables. Select “Chi-square” and “Correlation”; the latter will be a Spearman or rank correlation coefficient³.



Click on the Continue button to return to the previous window. Next, click on the “Cells” button to select what will be displayed in the cross tabulation output.



For this example, select “Observed” and “Expected” under Counts, as well as “Standardized” under Residuals. The first is the actual count in each cell, while the second is the expected count if the variables (q1 and country) were independent. Click on the “Next” button to close the Cell Display dialog window and then the “OK” button to produce the output (part of it is shown in the next page).

The expected count for the cell will be equal to the total for the column times the total for the row, divided by the number of observations. For example, for the first row (very satisfied) there

³ The computation is similar to that of the (Pearson) correlation coefficient, using the ranking of each value instead of the actual value. It is suitable for cases in which the location of a response on a scale is arbitrary. For example, it is not necessary to assume that the “distance” between “very satisfied” and “fairly satisfied” is the same as the distance between “fairly satisfied” and “not very satisfied.” Check the chapter on non-parametric statistics in the DECS-433/434 textbook.

are a total of 270 cases, from a total of 1499 cases (500 from Cyprus and 999 from Turkey). The expected count of very satisfied respondents from Cyprus would be $E_1 = (270 \cdot 500 / 1499) = 90.1$, compared to an actual count $O_1 = 123$.

The standardized residuals will be equal to difference between the observed count (O_1) count and the expected (E_1) count, divided by the square root of the expected. Thus, for the first cell, the standard residual is:

$$(O_1 - E_1) / \sqrt{E_1} = (123 - 90.1) / \sqrt{90.1} = 3.5.$$

Standardized residuals below 1 are not significant, while those above 2 are significant. If you sum the squared standardized residuals for all the cells in the table, the result is the Pearson χ^2 statistic. In this case, it has a value of 50.4 and the degrees of freedom are determined by (number of rows - 1) times (number of columns - 1). In this case, the test has 3 degrees of freedom (4-1 times 2-1).

The significance of the Pearson χ^2 will give you an indication of whether the variables are related in some way; the null hypothesis is that there is no association. In this case, we can reject the null and conclude that there is some association between the level of satisfaction with life in general and the respondent's country of birth – the test does not tell us anything about the direction of the relation or about causation (refer to the chapter on hypothesis testing in your marketing research book for more details).

Satisfaction with life in general * Country of citizenship Crosstabulation

			Country of citizenship		Total
			Cyprus	Turkey	
Satisfaction with life in general	Very satisfied	Count	123	147	270
		Expected Count	90.1	179.9	270.0
		Std. Residual	3.5	-2.5	
	Fairly satisfied	Count	279	544	823
		Expected Count	274.5	548.5	823.0
		Std. Residual	.3	-.2	
	Not very satisfied	Count	77	163	240
		Expected Count	80.1	159.9	240.0
		Std. Residual	-.3	.2	
	Not at all satisfied	Count	21	145	166
		Expected Count	55.4	110.6	166.0
		Std. Residual	-4.6	3.3	
Total	Count	500	999	1499	
	Expected Count	500.0	999.0	1499.0	

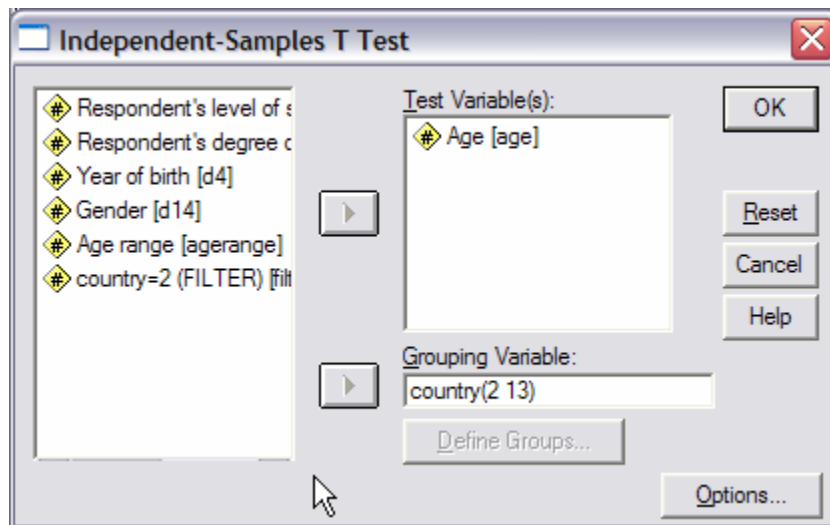
5. Comparing means and proportions

In the last example, we used the Pearson χ^2 to test whether the proportions of respondents for each level of satisfaction were the same across countries, i.e., that the proportion of Cypriot respondents feeling very satisfied is the same as the proportion of Turkish respondents feeling very satisfied, etc. In the next examples, we use *t*-tests to compare groups.

5.1. Independent samples *t*-test

Next, we could ask whether the average age of the respondents diverges depending on the country of origin or not. To do this, we can use a *t*-test, where the null hypothesis is that the average age of Turkish respondents is equal to that of Cypriot respondents ($H_0: \mu_{\text{age, Turkey}} = \mu_{\text{age, Cyprus}}$). Since the cases in our sample are the results of separate surveys in Cyprus and in Turkey, the most appropriate test would be one for independent samples, where the variances are not assumed to be equal.

From the Analyze menu, select “Compare Means > Independent samples T test”. This will open the dialog box shown below. Click on the age variable and click on the upper arrow button to make it the test variable. The grouping variable will be country. Click on the lower arrow button to move the country variable to the “Grouping Variable” field. Notice that its name is followed by two question marks (“country(?,?)”). You need to specify the two groups to be compared. While we only have two values in this sample (2 and 13), you could define the groups based on a continuous variable (by declaring a cut point) or select two values from a categorical variable that has more than two.



Click on the “Define Groups” button and type 2 to define Group 1 and 13 to define Group 2. Click on the “Continue” button to return to the previous window. The “OK” button should be clickable now (it was previously “greyed out”, as the “Define Groups” appears in the screenshot above). Click on OK to perform the test.

In addition to printing the means broken by the groups defined, SPSS will produce the *t*-tests for both assumed equal variances and unequal variances. We do not have any information that would suggest that the variances can be assumed to be equal, and assuming that the variances are not equal is a more conservative approach. Note in the output that there is Levene *F*-test for the equality of variances (the null hypothesis is that the variances are equal). In this case, the null can be rejected, thus the better *t*-statistic is the one computed without assuming equal variances.

The table below is a slightly edited version of the actual output. The *t*-statistic for the equality of means is 12.09, which is significant at the .05 level for a two tailed test; SPSS prints the *p*

value associated with the t -statistic under the “Sig. (2 tailed)” heading. Thus, we would reject the null hypothesis.

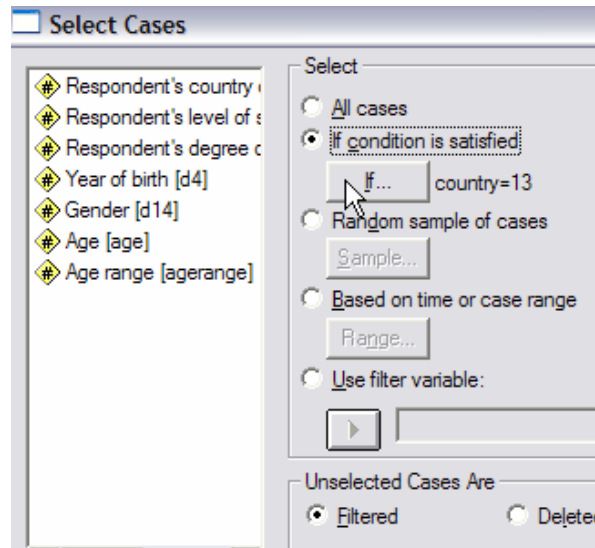
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Age	Equal variances assumed	51.377	.000	13.033	1494	.000
	Equal variances not assumed			12.094	824.733	.000

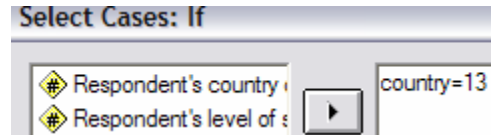
5.2. Intermediate step: Selecting a sub-set of cases for analysis

One of the strengths of statistical packages is to allow easy manipulation of the data. Suppose we would like to perform a one sample t -test to determine whether the average age (variable age) for Turkish respondents (value of country is 13) is significantly different than 39. As you saw in the results of the example in section 5.1, the average age for Turkish respondents was 38.04.

To select the cases of Turkish respondents, select “Select cases” under the “Data” menu. The default is “All cases.” Click on the “If condition is satisfied” radius button and then on the “IF” button to specify the condition. This will open a dialog box similar to the one you saw in section 3.2 (computing a new variable). In the screenshot below, notice that SPSS offers two options to deal with the unselected cases: filter or delete. Filtering will preserve the unselected data in your dataset, but will not use it subsequent computations until you go back to the Select cases dialog box and select “All cases.”

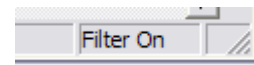


Either type “country=13” in the field reserved for writing the condition, or, as before, click on the variable name on the left panel and on the arrow button, typing the rest of the condition. Click on continue to return to the previous window.



Once you click OK on the Select cases window, the “filter” is on until you disable it. SPSS provides two visual cues. The first is to strike through the observation numbers; the second cue is in the program’s status bar, where you can read “Filter On”.

	country	q1
40	2.0	
41	2.0	
42	13.0	

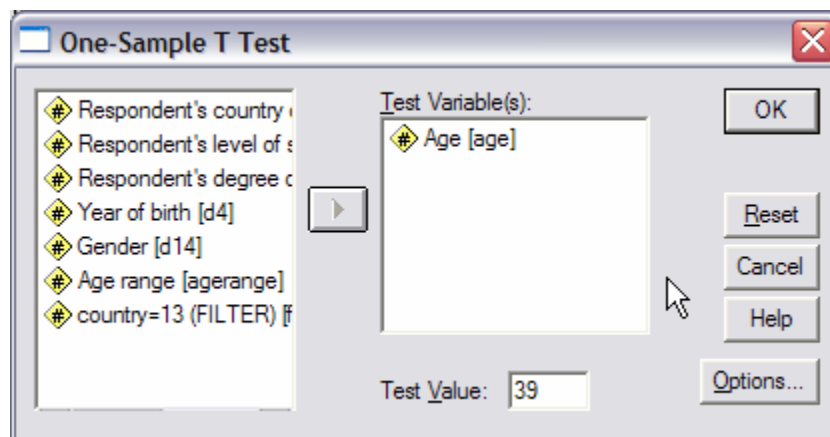


Notice also that SPSS created an indicator variable called “filter_\$” that takes a value of 1 if country=13 and it is equal to zero otherwise. If you filter the data in a different manner, this “flag” variable will be overwritten.

5.3. A one sample t-test

Having selected the Turkish respondents, you can now perform the one-sample *t*-test easily (see screenshot below). In this case, the null hypothesis is $H_0: \mu_{\text{age, Turkey}} = 39$.

Select “Compare means > One sample T test” from the “Analyze” menu. Next, select variable age as the test variable, and, finally, type “39” in the “Test Value” field. Click OK to perform the analysis.



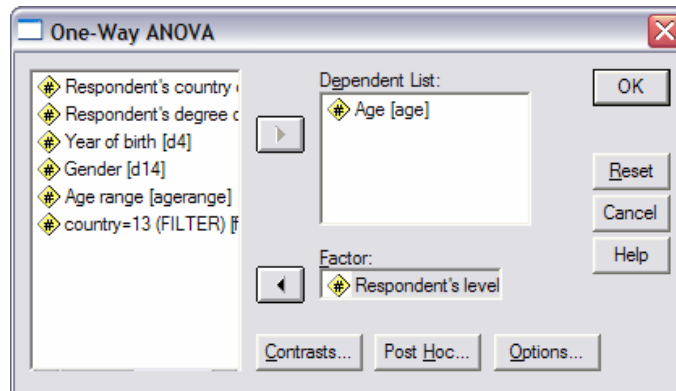
In the output, you will see that the null can still be rejected, albeit at a lower level of significance (0.042) than in the previous example:

One-Sample Test						
Test Value = 39						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Age	-2.034	995	.042	-.95582	-1.8780	-.0337

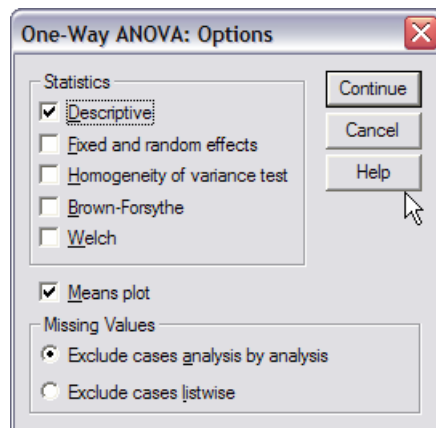
6. ANOVA

T-tests allow comparisons of only two groups at a time. To investigate differences among more than two groups (where the null hypothesis is the equality of all means across groups), Analysis of variance is more suitable.

In this example, we are going to run an ANOVA analysis of the age variable and compare two of the means. With the filter for Turkish respondents still in effect, select “Compare Means > One way ANOVA” from the “Analyze” menu. Select age as the dependent list and q1, the respondent’s level of satisfaction with life in general, as the factor or treatment effect.



Next, click on the “Options” button and select “Descriptive” statistics and a “Means” plot. Click on continue to get the one-way ANOVA window.

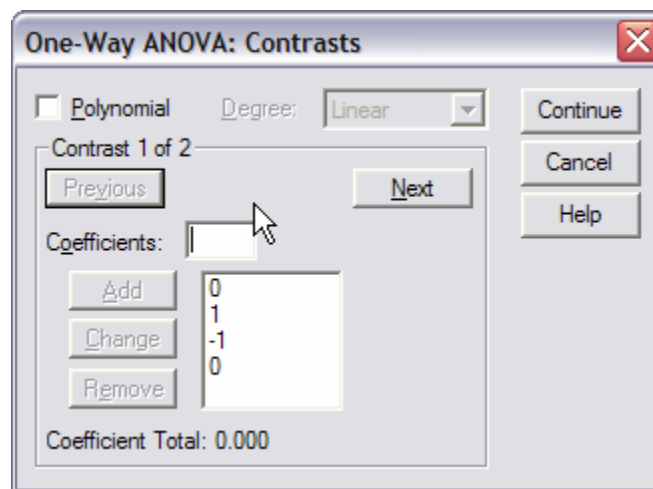


Now, click on the “Contrasts” button. In this window, you can specify a test on the different means (in this case, q1 has four categories, thus there are four means: 1=very satisfied to 4=not at all satisfied) as a linear combination for the means. For example, if you wanted to compare whether the average ages for respondents that declared being fairly satisfied (2) and not very satisfied (3) are equal, you could specify H_0 such that:

$$H_0: (0 \times \mu_1) + (1 \times \mu_2) + (-1 \times \mu_3) + (0 \times \mu_4) = 0$$

In other words, you are testing: $H_0: \mu_2 = \mu_3$. To do this in SPSS, in the Contrasts window you type in the coefficients that each mean gets in the null hypothesis. In this case, these coefficients are 0, 1, -1, and 0. Type each of them in the “Coefficients” field in the Contrasts window and click on the “Add” button.

You can specify additional hypothesis tests or “contrasts” by clicking on the “Next” button. For example, you could test whether the average age for very satisfied Turkish respondents is the same as the average age for fairly satisfied respondents: $H_0: \mu_1 = \mu_2$. The corresponding coefficients for this hypothesis would be 1, -1, 0 and 0. When you are done entering hypothesis, click on the “Continue” button to go back to the one-way ANOVA window.



In the one-way ANOVA window, click on the “OK” button to produce the results of this analysis.

7. Regression and regression diagnostics⁴

For the remainder of this document, we will use an SPSS data file called “**regression_example_BBD.sav**”. Please load these data into SPSS.

⁴ If you need good reference texts beyond what is covered in the DECS-433/434 textbook, you may find the following books useful: Peter Kennedy’s *A Guide to Econometrics* (MIT Press, 2003); Damodar Gujarati’s *Basic Econometrics* (Mc-Graw Hill, 2003); or William Greene’s *Econometric Analysis* (Prentice Hall, 2003). All of them are available in the NU library.

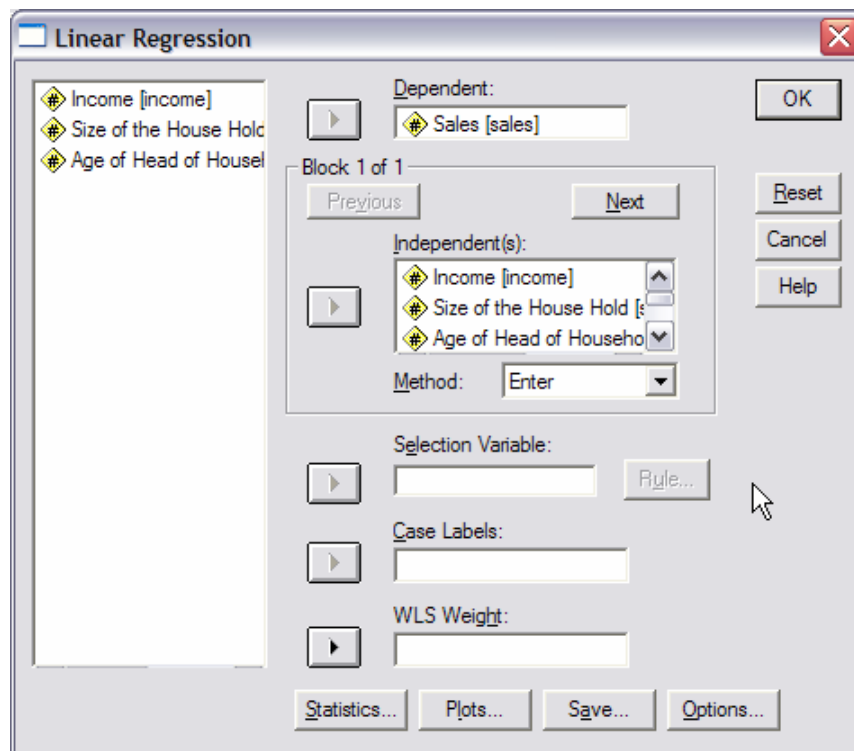
Suppose you have received data on a random sample of BBD (mail-order apparel company) customers. The data file contains last year’s BBD sales (in thousands of dollars), household income (in thousands of dollars), size of the household and age of the head of household.

7.1. Regression

To run a regression of sales against the three other remaining variables, select “Regression > Linear” from the “Analyze” menu. Select “sales” as the dependent variable; select income, size of household and age of head of household as the independent or explanatory variables. When you are done, click on OK to perform the analysis.

SPSS will print a summary of the model, followed by the analysis of variance. The model summary includes both the R-squared, as well as the adjusted R-squared, correcting for the number of regressors included; in terms of the ANOVA, recall that the R-squared is the proportion of the total variance of the dependent variable explained by variation in the explanatory variables (regression sum of squares). In this case, the R-squared indicates that 33.1% is explained by the independent variables.

The ANOVA includes the *F* test for the null hypothesis of all the coefficients being equal to zero ($H_0: \beta_1 = \beta_2 = \beta_3 = 0$).



In this case, the null hypothesis of all coefficients being equal to zero can be rejected.

ANOVA(b)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	770565.077	3	256855.026	32.325	.000(a)
	Residual	1557415.409	196	7945.997		
	Total	2327980.486	199			

Finally, the results of the regression follow. SPSS shows the regression coefficients, together with their corresponding standard deviation, as unstandardized coefficients. The *t*-statistics are also printed. For this particular regression, your results will show that the age of the head of household is not significantly different than zero.

Coefficients(a)

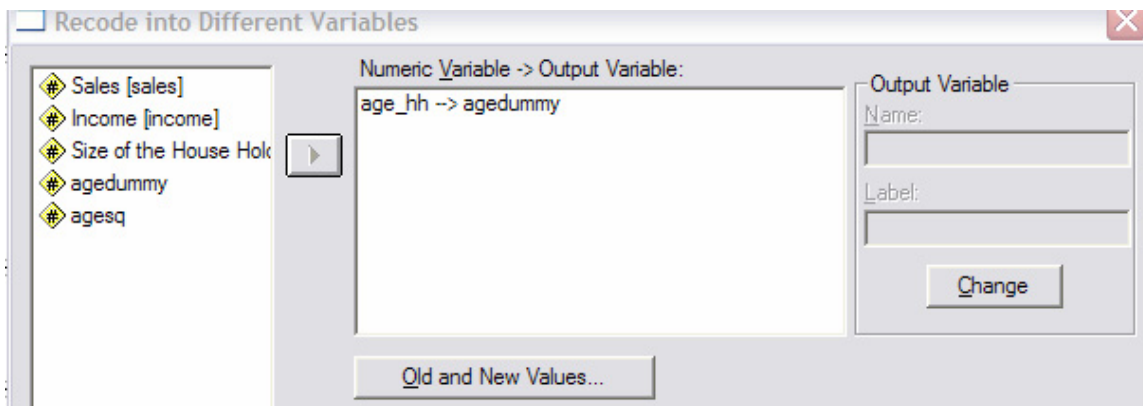
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45.359	43.089		1.053	.294
	Income	1.775	.217	.479	8.181	.000
	Size of the House Hold	22.122	4.190	.309	5.279	.000
	Age of Head of Household	.449	.768	.034	.585	.559

a Dependent Variable: Sales

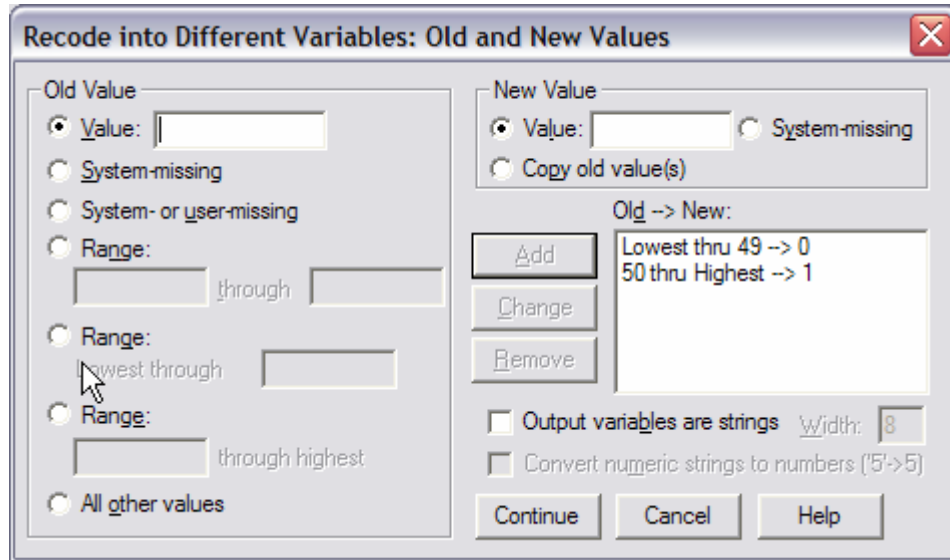
7.2. Intermediate step: creating a dummy variable

Suppose (just for illustration purposes) that you suspect that age matters only if the head of household is more or less than 50 years old. Thus, you would like to run a similar regression, except that instead of age of the head of household, you would include a dummy with a value of 0 if the head of household is less than 50 years old, and a value of one if the head of household is 50 years or older.

To create the dummy, you will follow the same steps shown in section 3.3. Creating the dummy is like recoding a variable. To do this, select “Recode > Into different variable” from the “Transform” menu. To create the dummy, click on the “age_hh” variable, and then on the arrow to move the variable over to the “Numeric variable -> Output variable” field. In the “Output variable” field, type “agedummy” and click on the “Change” button.



Next, click on the “Old and New Values” button to recode ages lowest through 49 into a value of 0, and 50 through the highest into a value of 1. Click on the “Continue” button to return to the previous window. Finally, click on OK to create the dummy variable.



To run the regression with the newly created dummy variable, go back to the regression window, remove “age_hh” from the regressors and add “agedummy.”

7.3. Regression diagnostics

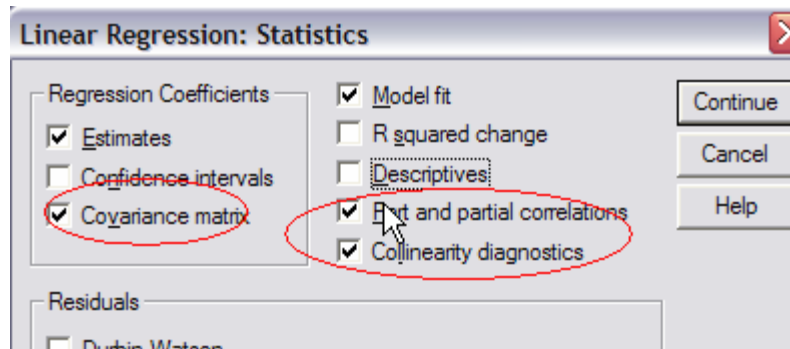
In this sub-section we go through some tests that are easily done with the student version of SPSS. The full fledged version of SPSS offers additional tests and options⁵.

7.3.1. COLLINEARITY

Multicollinearity in a regression causes the standard errors of the parameter estimates to be large; the parameters are still best linear unbiased estimators or “BLUE”, as you may remember from your statistics class. In addition, estimated parameters are unstable: a small change in the data (such as excluding a couple of observations) will result in dramatic changes in the estimated parameters.

Given that regressors are typically correlated, multicollinearity is a matter of degree. To detect collinearity, you inspect the partial correlations between the variables in your regression, the variance inflation factors (VIF) or the condition numbers. You can request these statistics using the “Statistics” button in the regression dialog box.

⁵ Alternative packages such as SAS or Stata offer more powerful options for regression diagnostics. SAS, while popular in corporate environments and academia, has a steep learning curve. Stata is easier to learn, but used mostly in academic settings; it is taught in the Empirical Methods in Strategy (MGMT 469) course. Both packages are available in the special software workstations in the Kellogg computer labs.



If you see a large partial correlations⁶ (e.g., 0.8), you should deal with the collinearity. However, the problem with bivariate correlations is that they do not give you information about one of your independent variables being a linear combination of two or more independent variables. Hence, inspecting the VIFs is possibly a better solution, as is looking at the condition numbers.

A rule of thumb with VIFs is that a VIF higher than 10 is a source of concern – this is simply a rule of thumb. Recall that the square root of the VIF tells you how much larger the standard error is relative to a case where you do not have collinearity. For example, a VIF of 4 indicates that the standard errors are twice as large as if there was no collinearity.

Condition numbers are statistics based on the eigenvalues of the matrix of independent variables. The rule of thumb is to consider condition numbers greater than 15 or 20 as a source of concern.

Yet another way of detecting collinearity is inspecting the correlations between the coefficients (select “Covariance matrix” among the statistics in the regression – see screenshot above). High correlations between pairs of coefficients could indicate collinearity.

7.3.2. HETEROSKEDASTICITY

Heteroskedasticity makes your parameter estimates no longer BLUE – they are still unbiased, but no longer have the minimum variance.

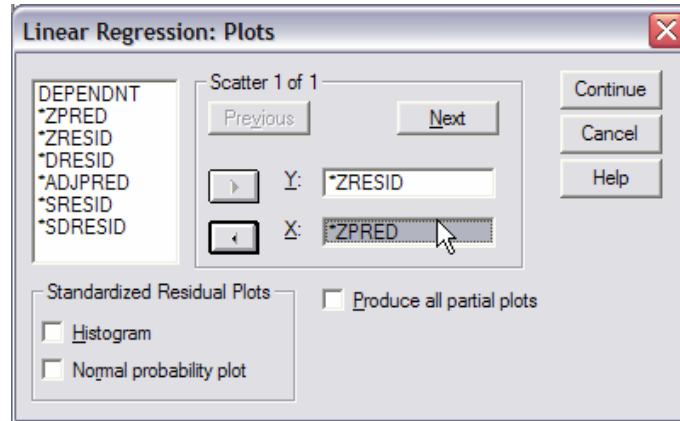
Unfortunately, SPSS does not have built in procedures to test for heteroskedasticity⁷. The tests can be performed by writing some code or by performing a series of analysis in sequence to obtain the test statistic. See section 7.5 below for the step by step instructions to perform this test.

Despite not having built in procedures to test for heteroskedasticity, you can plot the standardized residuals (ZRESID) against the standardized predicted values (ZPRED). If there is no heteroskedasticity, the plot should look random. If you see a pattern, such as a funnel shape or a curve, this indicates heteroskedasticity. A curved shape, in particular, could indicate some non-linearity in the relation that you failed to take into account. You can find the “culprit” by dropping one of the independent variables and checking whether the plot becomes random again.

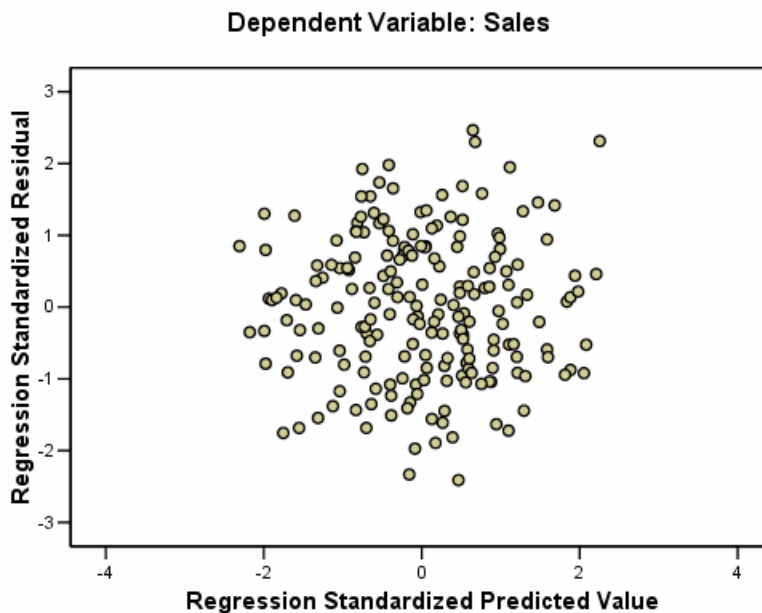
⁶ The “partial correlations” are Pearson correlation coefficients.

⁷ This procedure is built into other packages, such as SAS or Stata.

To obtain these plots, in the regression dialog box, click on the “Plots” button. Select ZRESID as the Y variable and ZPRED as the X variable. Click on the “Continue” button to return to the previous dialog box.



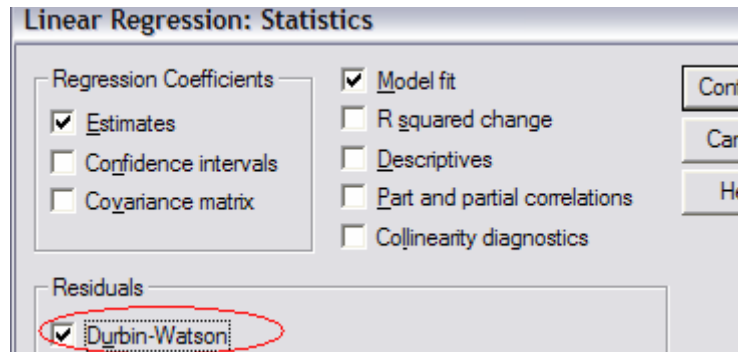
For the regression of sales against income, size of household and the age dummy variable, the plot of the standardized residuals against the standardized predicted values looks as follows:



7.3.3. SERIAL CORRELATION

If you are working with time series data, you may have serial correlation: the error terms from different time periods might be correlated. If the correlation is between one time period and the period immediately before it (the error term in period t is correlated with the error term in period $t-1$) then you have first order serial correlation. Serial correlation will underestimate the standard errors associated to the parameter estimates, making them look more significant than they really are. The parameter estimates are still unbiased.

The most popular test for first order serial correlation is the Durbin-Watson test. You can produce the Durbin-Watson statistic by requesting it among the statistics produced by the regression (even if this dataset is not a time series). In the regression dialog box, click on the “Statistics” button and check the box next to “Durbin-Watson” to select the test.



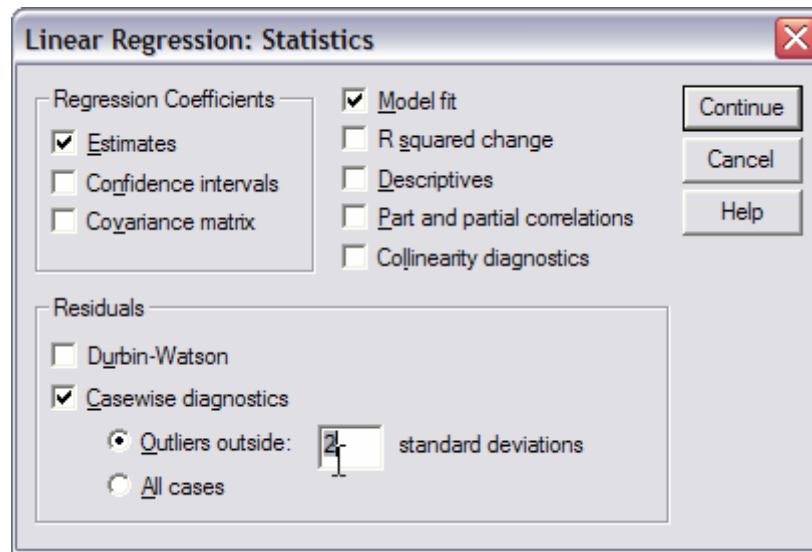
The rule of thumb is that a Durbin-Watson close to 2 indicates no serial correlation, a D-W greater than two indicates negative serial correlation, and a D-W below 2 indicates positive serial correlation. Since there is an area of indeterminacy and the exact distribution of the Durbin-Watson depends on the data matrix, to rule out serial correlation you need to compare the D-W from your regression against the critical values for the D-W in the back of an econometrics textbook⁸.

7.3.4. OUTLIERS

The scatter produced above for heteroskedasticity (standardized residuals against standardized predicted values) also allows you a visual check for the presence of outliers.

Another check for outliers comes from the resulting distribution of standardized residuals. A standardized residual greater than 3 (or smaller than -3) is a reasonably good indicator that the observation could be considered an outlier. To obtain descriptive statistics and a list of potential suspect observations, click on the “Statistics” button in the regression dialog box. Select “Casewise diagnostics”. The default is to print outliers outside 3 standard deviations. In our regression, no observation will be listed with this criterion. You could type “2” standard deviations instead, to see a list of cases for which the absolute value of the standardized residual exceeds 2.

⁸ The tables are usually reproduced from N.E. Savin and K.J. White, “The Durbin-Watson test for serial correlation with extreme sample sizes of many regressors”, *Econometrica*, 45(8), 1977, pp. 1992-1995.



You can also request other measures of influence and leverage by using the “Save” button in the regression dialog window. See the following section.

7.4. Saving predicted values and residuals

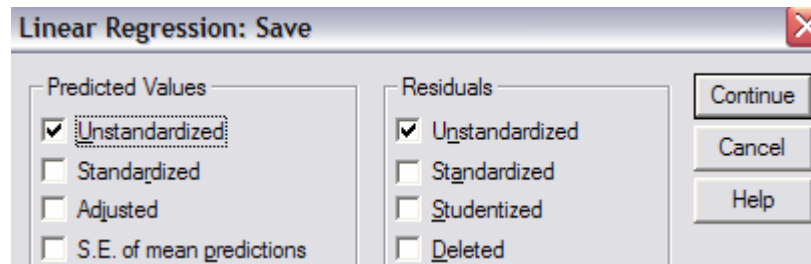
For certain tests and graphics, you may want to save the predicted values or residuals from a regression. The regression feature of SPSS also allows you to do this. In the regression dialog box, click on the “Save” button. For this example, select the unstandardized predicted values (i.e., this predicted values will be in the same units as your dependent variable) and the unstandardized residuals. Note that this dialog window allows you to generate studentized residuals, leverage values, and Cook’s distance. Click on the “OK” button to return to the regression dialog box.

If you request any of the available statistics for each residual, rather than eyeballing the data (you could be dealing with thousands of observations), your next step should be to explore the data as shown in section 4.3 and request “Outliers” among the statistics to be printed (click on the Statistics button).

Once you run the regression, two new variables will appear on your dataset: PRE_1 (the predicted values) and RES_1 (the residuals from the regression). You can manipulate these variables as you would any other in the dataset. If you run further regressions, the predicted values will be named with different numbers: PRE_2, PRE_3, etc.

Note: Having the residuals, you can easily compute the *Jarque-Bera test* for normality. Simply run some descriptive statistics on the residuals (res_1), requesting the skewness (S) and the kurtosis (K). The Jarque-Bera test statistic is $JB = n \cdot (S^2/6 + K^2/24) \sim \chi^2_{(2)}$, where n is the number of observations; the null hypothesis is that the residuals are not normally distributed. You can use the SPSS (compute “1-cdf.chisq($JB,2$)”) or Excel (“=chidist($JB,2$)”) to get the p -value.

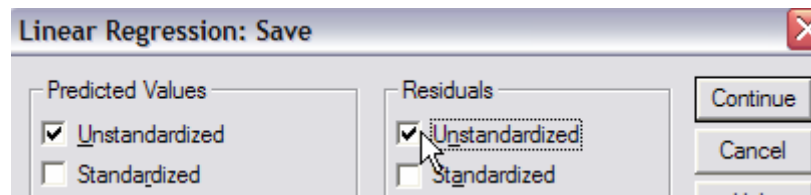
Refer to section 8.4 of this document if you only have two variables and want to plot the predicted values for a bivariate regression.



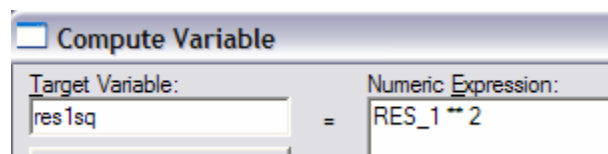
7.5. The Breusch-Pagan test for heteroskedasticity, step by step

This section walks you through the steps you need to follow to compute the Breusch-Pagan statistic in the same version as in KStat. This computation uses the predicted values of the regression. In reality, the computation of the Breusch-Pagan test can be more nuanced, since you can include other variables in the computation of the score⁹.

Step 1: If you followed the example in section 7.4, you have already done this step. Run the regression (sales against income, size of household and the dummy variable for the age of the head of household) and save the unstandardized residuals and the unstandardized predicted values. The residuals can be saved by selecting the relevant option in the “Save” button of the regression dialog window. If this is the first set of residuals you save, the new variables will be named “res_1” and “pre_1”, respectively.

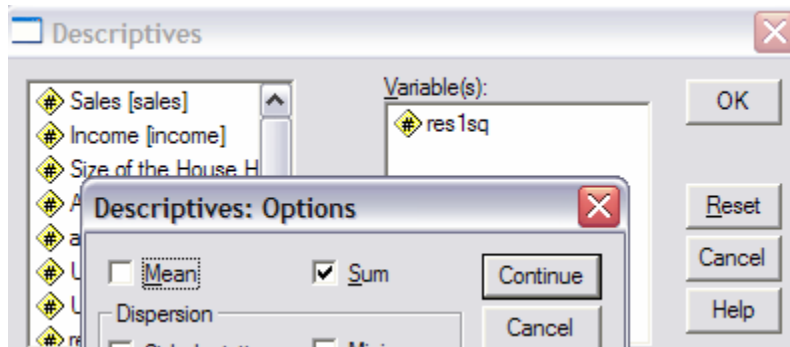


Step 2: Create a new variable (call it “res1sq”) that will be the square of the residuals. To do this, select “Compute” from the “Transform” menu. If there is a previous expression in the window, click on the “Reset” button to clear it. The exponent operator is the double asterisk (**): $\text{res1sq} = \text{res}_1^{**2}$. Click on OK to perform the computation.



Step 3: Calculate the sum of the squared residuals. From the “Analyze” menu, select “Descriptive Statistics > Descriptives...” and select “Sum” in the “Options” button.

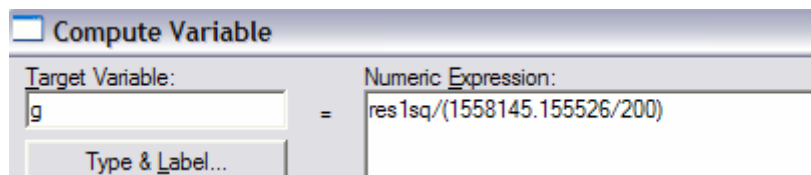
⁹ Refer to Gwilym Pryce (2002). “Heteroscedasticity Testing and correcting in SPSS.” This document is available online at the following URL: pages.infinit.net/rlevesqu/Tutorials/HeteroscedasticityTestingAndCorrectingInSPSS.zip. It also reviews the steps necessary to check for the normality of the residuals, since the Breusch-Pagan statistic is misleading if the residuals are non-normal. Finally, it also gives examples of several solutions to heteroskedasticity, including the widely used White standard errors.



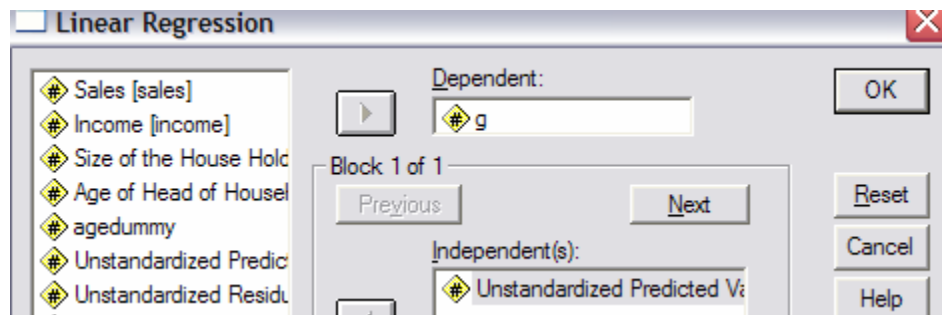
Step 4: Go to the Output Viewer and click and double-click on the descriptives output table to select it. Double-click on the sum to select it. Copy this value (either press CTRL-C or select “Copy” from the “Edit” menu) into the cache.

Descriptive Statistics		
	N	Sum
res1sq	200	1558145.15
Valid N (listwise)	200	

Compute a new variable (call it “g”) that is equal to the squared residual divided by the sum of the squared residuals, divided by the number of observations: $g = \text{res1sq}/(\text{RSS}/n)$. In this case, the residuals sum of squares is “1558145.15526”, and the number of observations is 200



Step 5: Run a regression of g against the predicted values (pre_1). Again, if there are selections from the previous regression, click on the “Reset” button to clear them.



Step 6: This step can be done in Excel or in SPSS (using the Compute dialog window). Calculate the Breusch-Pagan test score, $B = 1/2 \cdot (\text{REGSS}) \sim \chi^2_{(1)}$. In this case, you can see from

the ANOVA table of the resulting output is “0.881” (again, double-click a couple of times to be able to copy this number with all its digits). Thus, in this case, $B = 0.4403$:

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.88061726	1	.881	.598	.440 ^a
	Residual	291.801	198	1.474		
	Total	292.687	199			

Step 7: Get the p -value associated to the χ^2 you just computed (with one degree of freedom). Again, you can perform this calculation in SPSS or in Excel. In SPSS, you would use the CDF.CHISQ function in the Compute dialog window: $B_pval = 1 - \text{cdf.chisq}(B, 1)$. Evidently, you could have performed steps 6 and 7 in one calculation.

In Excel, you can use the CHIDIST function: $=\text{chidist}(0.4403, 1)$.

In both cases, you will obtain a p -value of 0.5070, so we cannot reject the null hypothesis. In the Breusch-Pagan test the null is homoskedasticity.

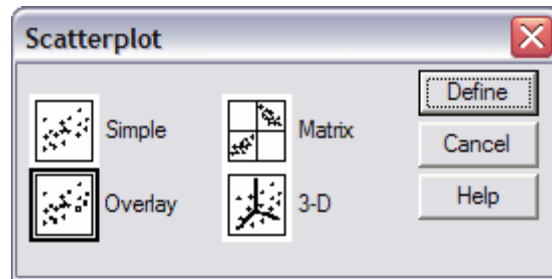
8. Graphics

The SPSS “Graphs” menu includes different kinds of graphics, some of which we have used in the examples above. Here, we will work on four examples to illustrate the features you can find in SPSS; the examples will use the dataset named “**regression_example_BBD.sav**”. As with any procedure in SPSS, you can use the “Select cases” from the “Data” menu to restrict your graphics to a subset of the cases in your dataset. This feature was shown in section 5.2 of this document.

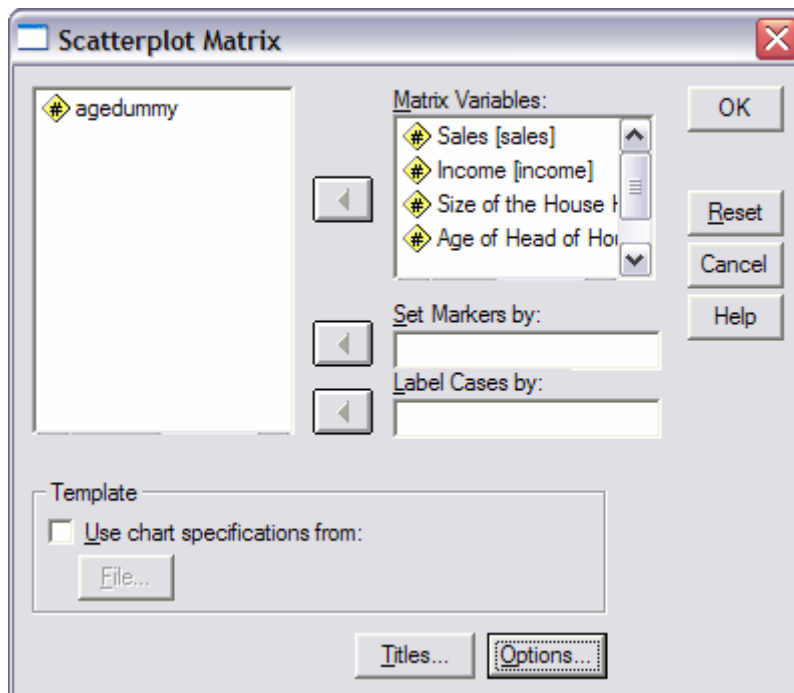
For any of the graphs, there are further formatting options in the SPSS Output Viewer. You can change the color of the chart elements, include value labels, etc. You may also adjust the size of the output so that it is more convenient to paste in a Word document, for example.

8.1. Matrix of scatterplots

A good way of visualizing relations in your data is creating a matrix of scatter plots. In other words, use SPSS to generate a scatter plot of each of the variables in your dataset against the rest. To do this, select “Scatter...” from the “Graph” menu. Click on “Matrix” and then click on the “Define” button to open the scatterplot dialog window.



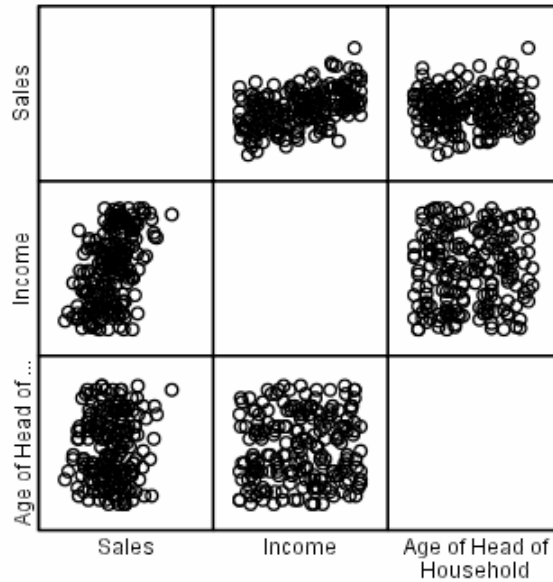
In the “Scatterplot Matrix” dialog window, select sales, income, size and age of household as the “Matrix variables”.



Note that you can also specify titles for the resulting chart by clicking on the “Titles” button; the “Options” button allows you to specify how missing values are treated. The default is “listwise”. This means that if you are analyzing four variables (as an example), if the value of one of the variables is missing for any of the cases, the case is dropped from the analysis. Another option is to exclude cases variable by variable. In this case, cases are dropped depending on the variables necessary to compute a specific statistic. For example, if you are computing a correlation matrix and “sales” is missing in one case, the case will be excluded from the correlations that involve sales. The downside to this is that the correlations are based on different samples of cases.

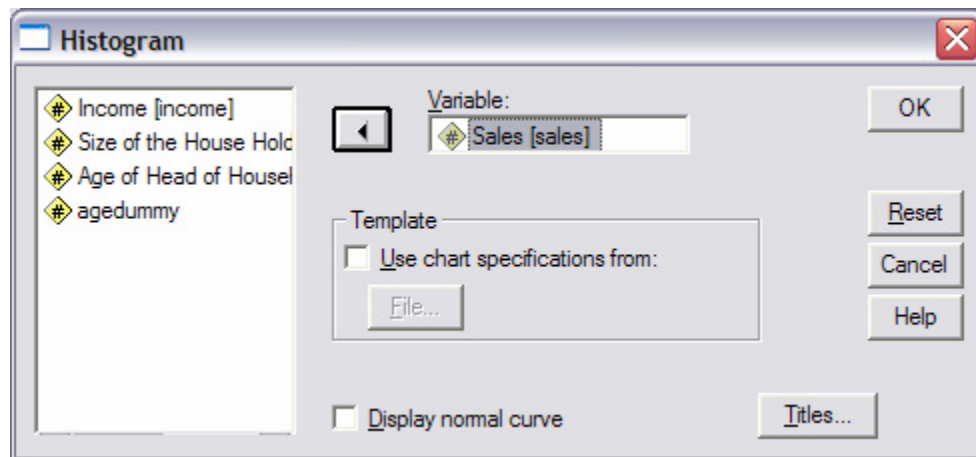
Click on the “OK” button to produce the matrix of scatter plots. The scatter plot matrix inserted below was created with only three of the variables: sales, income and size of household. Notice that the diagonals must be empty (a scatter of a variable against itself). The scatter plot of

sales against income shows you a positive relation, as you would expect. The plots of categorical variables are typically harder to read since the values overlap¹⁰.



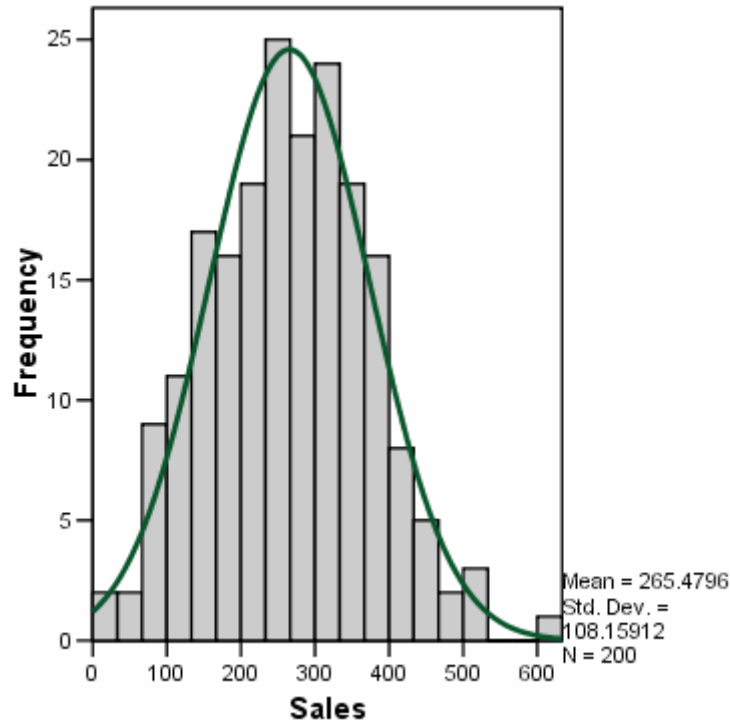
8.2. Histograms

To create a histogram, select “Histogram” from the “Graph” menu to open the Histogram dialog box. Click on “Sales” and move it to the “Variable” field. Also, click on the “Display normal curve” box. SPSS will overlay a normal curve on the histogram. The normal curve overlay allows you to visualize whether the variable under analysis (sales, in this example) is close or far from resembling a normal distribution.



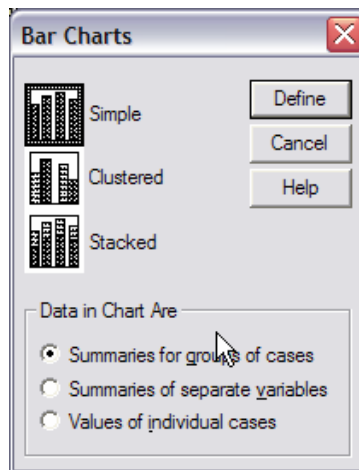
¹⁰ The full version of SPSS allows “jittering” or “dithering.” This process offsets identical (x,y) pairs by adding a very small random value (uniformly distributed) to each observation so that you can visualize a relation between the variables, if it exists. Other packages also allow changing the size of the marker symbol in the chart to reflect the relative frequency of the (x,y) pairs.

Click on the “OK” button to produce the histogram. The histogram inserted below is an edited version of the default output:

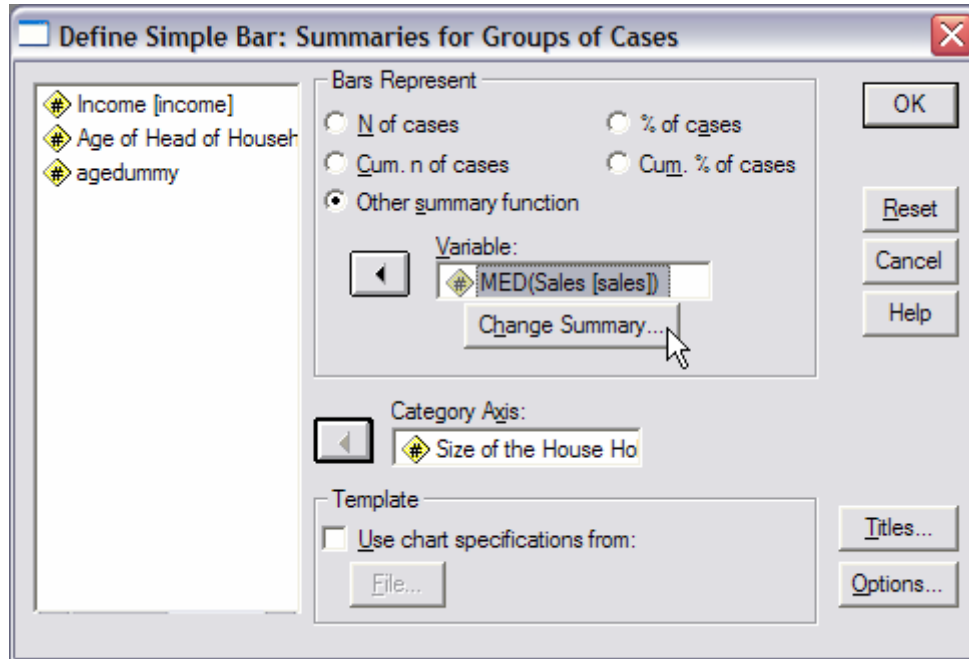


8.3. Bar charts

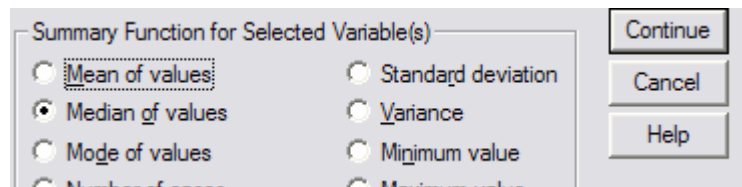
Finally, this last example shows you how to produce a bar chart that depicts the median sales level for each different household size. Select “Bar...” from the “Graph” menu. Make sure that the “Summaries of groups of cases” radius button is selected. Click on “Simple” and then on the “Define” button”.



In the “Define Simple Bar” window, click on the “Other summary function” radius button. Then, click on “Sales.” Click on the arrow button next to the “Variable” field to select sales as the analysis variable. By default, SPSS will propose to present the mean as the summary function.

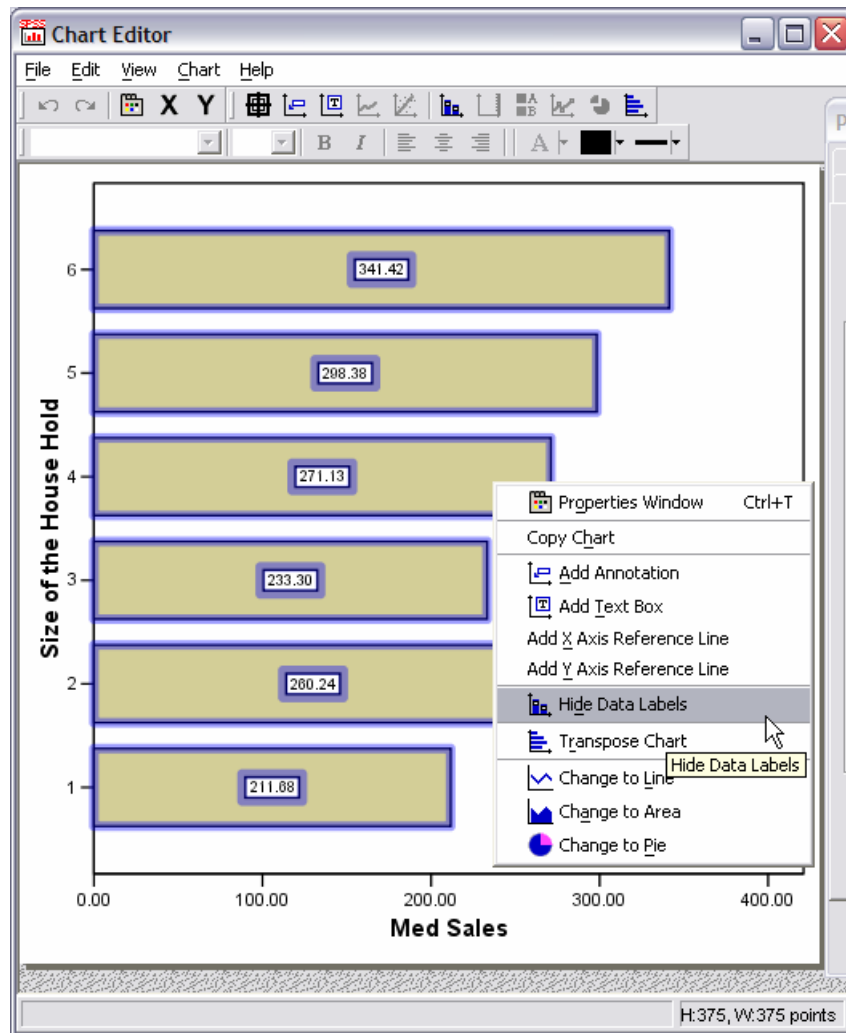


Click on the “Change Summary...” button to select the “Median of values” instead of the mean. This will open a dialog box that offers a list of different summary functions (see screenshot below).



Click on the “Continue” button to return to the previous dialog box (“Define simple bar”). Click on the “size of household” variable and on the arrow button next to the “Category axis” field, to make size of household the horizontal axis in your bar chart. Finally, click on the “OK” button to produce the bar chart.

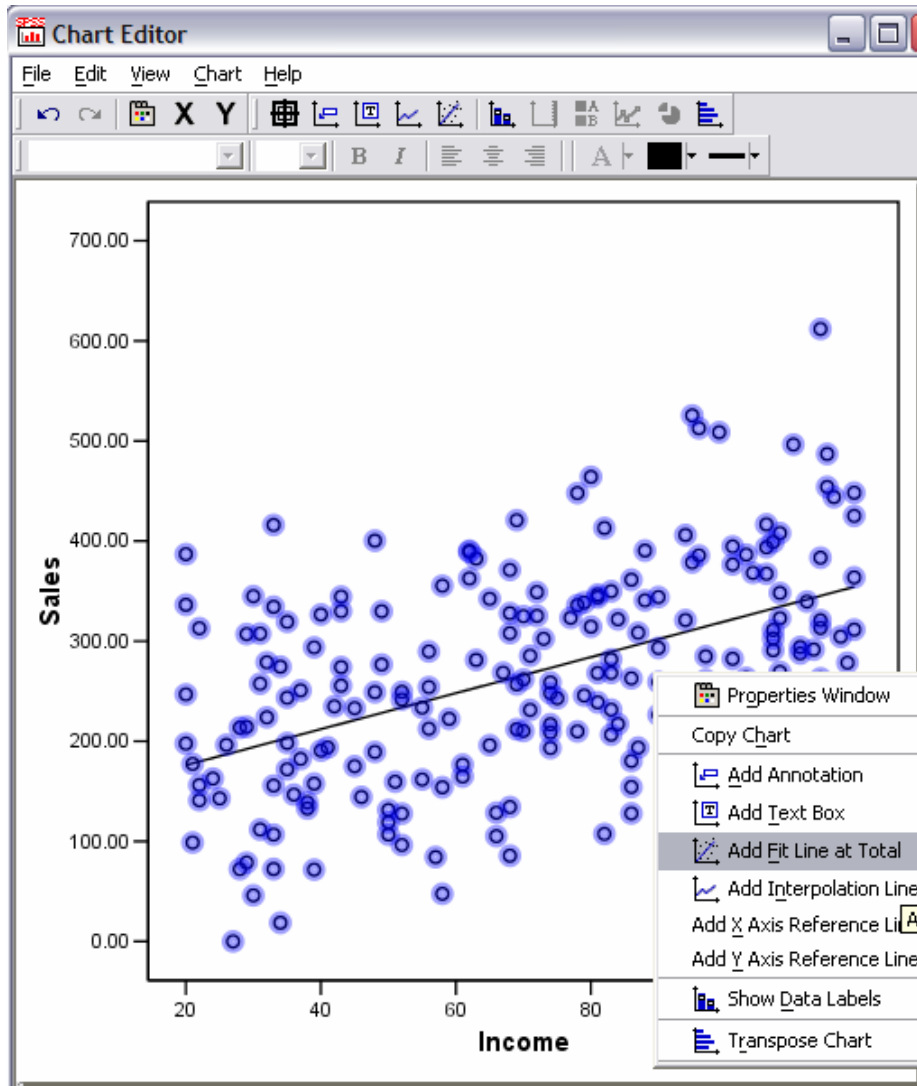
As mentioned at the beginning of section 8, you can use the SPSS Output Viewer to further edit your chart. Just double-click on the chart in the Output Viewer to open it in the “Chart Editor”, shown in the screenshot below.



8.4. A scatterplot with a regression line

In some cases, you may simply want a scatter plot of two variables and the corresponding regression line (for the bivariate regression). You can add the regression line to the scatter plot by editing the chart.

To do so, produce a scatter plot with sales on the vertical (Y) axis and income on the horizontal (X) axis. Double-click on the chart to open it in the Chart Editor. Double-click on the “cloud” area of the chart (on the markers that represent your data) to select it. Finally, right click to see the context menu (which you can see on the screen shot below). The “Add fit line at total” will place the regression line and the R-squared for it on the chart.



Keep in mind that this fit line corresponds to the predicted values of a regression of sales against income. If you want to plot the predicted values of a multivariate regression, refer to section 7.4 on how to save the predicted values from a regression.

9. Limitations of the student version

- The full-fledged version consists of SPSS Base, usually accompanied by additional modules. At Kellogg, the “special software” workstations include SPSS Base, the Advanced Models, Regression Models, and Tables modules; Northwestern does not have a license for the Categories (includes perceptual maps) and Conjoint (conjoint analysis) modules. There are more than 10 additional modules available from SPSS for special applications; for the most part, SPSS Base is sufficient. A full list of SPSS modules is available at www.spss.com/spss/family.cfm.

- The student version allows a maximum of 50 variables and 1500 cases, which is sufficient for most applications within a course, but typically not enough when you deal with actual data.
- You cannot aggregate the dataset (for example, create a file with the average level of satisfaction of respondents by country of origin) or merge the dataset with an additional data file.
- You cannot write SPSS programs, “syntax”. While you may think that you will not ever write SPSS programs, doing so saves you considerable time in carrying out your analysis and being able to replicate them. For example, you may update or correct the data and then need to update all the analyses you had done before. Alternatively, you may want to replicate a number of tables for a sub-sample of your data only. Finally, another good reason for it is dealing with large datasets (over 1GB): syntax allows you to script the analysis and test it on a sub-sample of the data, and then run all the analysis on the full sample, which can take a long time depending on the number of observations and the nature of the analysis.
- The student version does not have all the analytical features available in the full-fledged version of SPSS. Logistic regression or multinomial logit models, for example, are not available in the student version.
- The full fledged version of SPSS allows users to read data by querying SQL servers, which is one reason for its wide use in the corporate sector.

Northwestern University has an agreement with SPSS that allows six or 12 month “rentals” (via e-Academy, elms.e-academy.com/northwestern; prices range from \$40 to \$125, depending on the specific rental chosen) of either the SPSS Base software or the SPSS Graduate Pack. The latter includes SPSS Base, SPSS Regression Models (which includes binary and multinomial logistic regression and non-linear regression) and SPSS Advanced Models (which includes procedures such as general linear models, hierarchical loglinear models, survival models, etc).

10. Using SPSS syntax

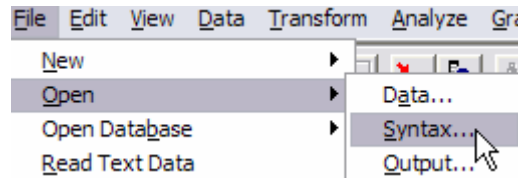
Writing programs in SPSS (using SPSS “syntax”) allows you to use additional features not reflected in the menus and dialog windows. In addition, it also allows you to replicate all the steps in your analysis quickly, which is extremely convenient if your dataset has been corrected or updated.

If you look at the SPSS output viewer carefully, you will notice that, under “Notes”, it produces a subsection titled “syntax”. For example, if you go back to the frequency tables for “agerange”, the output includes the syntax to generate the tables and histogram.

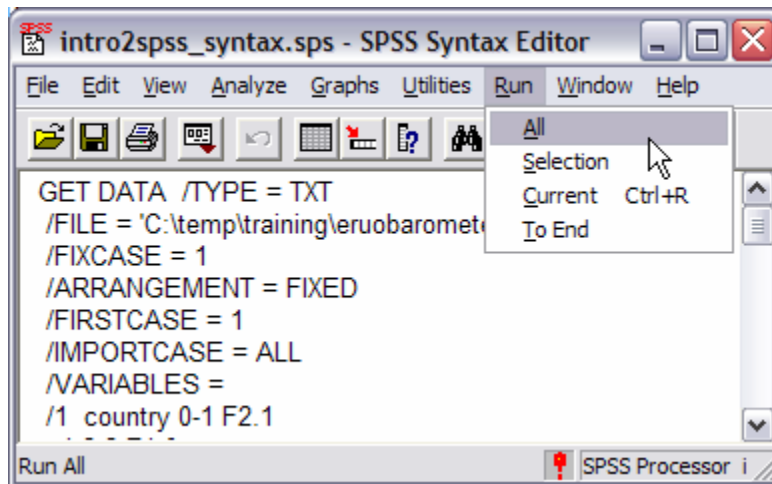


If you use the full fledged version of SPSS, most menus include a “Paste syntax” button that will paste the commands into the SPSS Syntax Editor (see screenshot below). This is a convenient method of keeping track of your work, as well as a way of learning syntax. There are commands and options in SPSS that are not reflected in any of the menus; all commands are explained in the SPSS Syntax Reference manual.

The ZIP file with sample data files also includes a text file called “intro2spss_syntax.sps” that contains the syntax necessary for all of the examples from this handout¹¹. This file assumes that you have placed all the files from “tek_spss.zip” in a subdirectory of your H drive (your network home directory) called “spstraining”. To run it in the full fledged version of SPSS, select “Open > Syntax...” from the “File” menu:



Select the file and click OK. This will open the SPSS Syntax Editor, which is a window not available in the student version. To run all the commands in the syntax file, select “All” under the “Run” menu. Else, you can execute the instructions piece by piece by selecting a block of syntax and running a “Selection” instead.



¹¹ To view this file in your laptop, you can open it in Word or in an editor. If you modify it, be sure to save it as a text file.

The basic rules for writing SPSS syntax are:

- Commands are not case sensitive. In the sample syntax file that accompanies this document, SPSS commands are in upper case, but this is not necessary.
- Commands have to begin in column 1 of a line. You can use plus (+) or minus (-) signs in the first column if you want to indent the command specification to make the command file more readable.
- Command options start with a slash (“/”).
- If multiple lines are used for a command, column 1 of each continuation line must be blank (in the sample below, this is why the options are indented with spaces).
- Commands are terminated with a period (this is optional).
- Comments are indicated by an asterisk (“*”) in the first column.

When you create a syntax file and execute it, SPSS does not carry out certain commands until required to do so. For example, if your syntax file ends with a “GET DATA” command, the command will be read but not executed until the data is necessary (for example, if you run a frequency). To force SPSS to execute the command, add “EXECUTE.” after the command.

Some of the editing done using the SPSS Chart Editor (e.g., making the bar chart horizontal and including the value labels for each bar) does not have a syntax equivalent. Instead, SPSS has resorted to creating “chart template files”, which are binary files with extension “sct”. These files are created by editing a chart in the Chart Editor and selecting “Save chart template” from the “File” menu can then be called in a command file by adding a / TEMPLATE=”path” option to the GRAPH command.

11. Additional resources for learning SPSS

For a detailed explanation of any of SPSS commands, check the SPSS manuals. A set of them is available from the student technical support center (TSC in room 163 of the Jacobs Center). Their PDF versions are also part of the SPSS installation in the “special software” workstations in the Kellogg computer labs. Some useful online resources are:

- SPSS tutorials, a list of links compiled by Raynald Levesque:
pages.infinit.net/rlevesqu/spss.htm
- A list of resources to learn SPSS compiled by UCLA:
www.ats.ucla.edu/stat/spss/default.htm
- SPSS online training workshop from Central Michigan University:
<http://calcnet.mth.cmich.edu/org/spss/>



NORTHWESTERN
UNIVERSITY