**HOW TO PREPARE YOUR DATA**

by Simon Moss

|  |
| --- |
| **Introduction** |

To analyse quantitative data, researchers need to

* choose which techniques they should apply to analyze their data, such as ANOVAs, linear regression analysis, neural networks, and so forth
* prepare their data—such as recode variables, manage missing data, and identify outliers
* test the assumption of these techniques they chose to conduct
* implement the techniques they chose to conduct

Surprisingly, the last phase—implement the techniques—is the simplest. In contrast, researchers often dedicate hours, days, or even week to the preparation of data and the evaluation of assumptions. This document will help you prepare your data in SPSS. Another document will help you test the assumptions. In particular

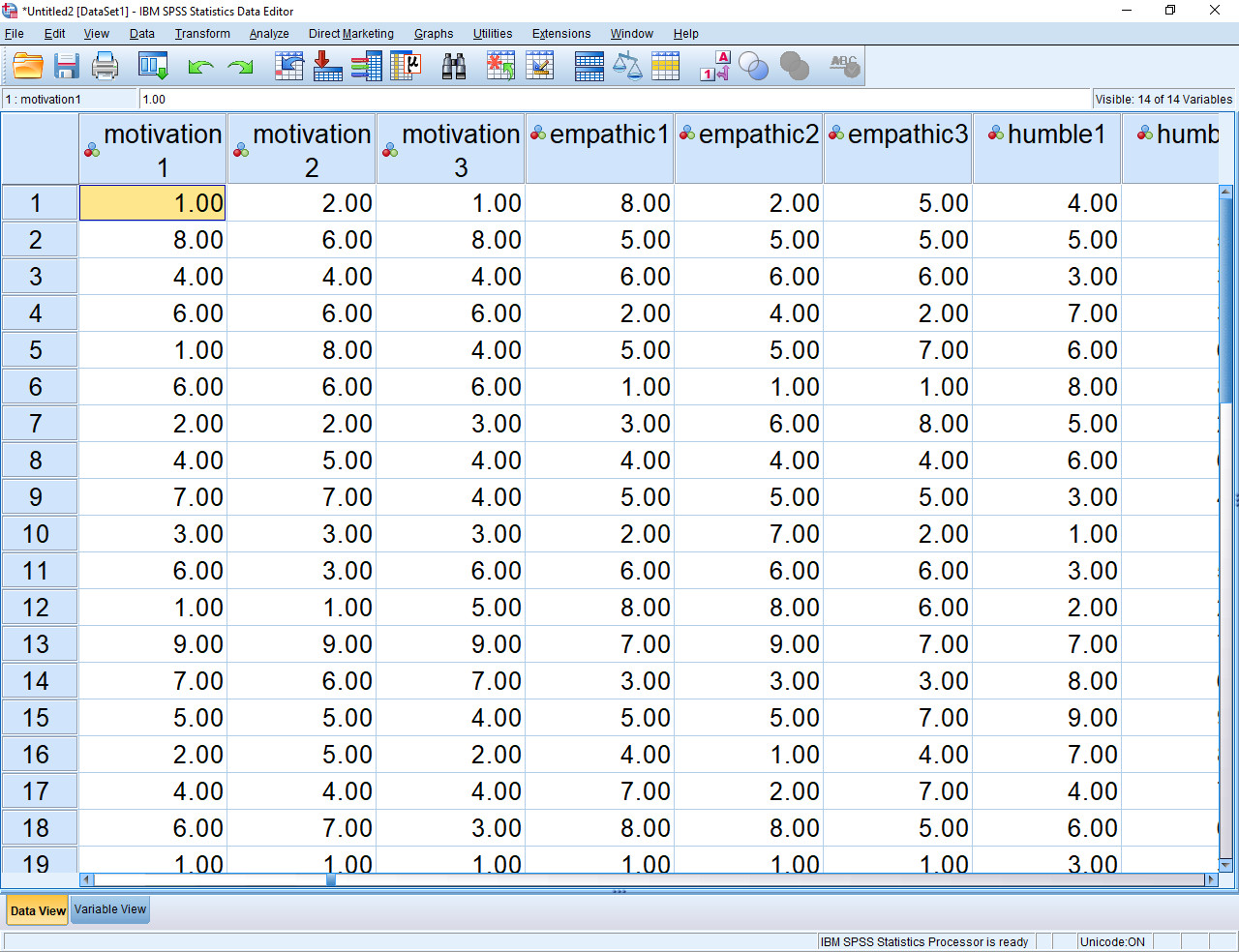
* this document will describe a series of activities you need to complete
* you should complete these activities in the order they appear in this document
* in practice, you might not need to complete all these activities, however

**Illustration**

To learn how to prepare the data, this document will refer to a simple example. Suppose you want to ascertain which supervisory practices enhance the motivation of research candidates. To explore this question, research candidates might complete a survey that includes a range of questions and measures, as outlined in the following table

|  |  |
| --- | --- |
| Topic | Questions |
| Motivation | On a scale from 1 to 10, please indicate the extent to which you feel  1 Absorbed in your work at university  2 Excited by your research  3 Motivated during the morning |
| Empathic supervisors | On a scale from 1 to 10, please indicate the extent to which your supervisor  4 Is understanding of your concerns  5 Shows empathy when you are distressed  6 Ignores your emotions |
| Humble supervisors | On a scale from 1 to 10, please indicate the extent to which your supervisor  7 Admits their faults  8 Admits their mistakes  9 Conceals their limitations |
| Demographics | 10 What is your gender?  11 Are you married, de facto, divorced, separated, widowed, or single? |

An extract of the data appears in the following table. To practice these activities, you could enter data that resembles this spreadsheet.



|  |
| --- |
| **SPSS syntax** |

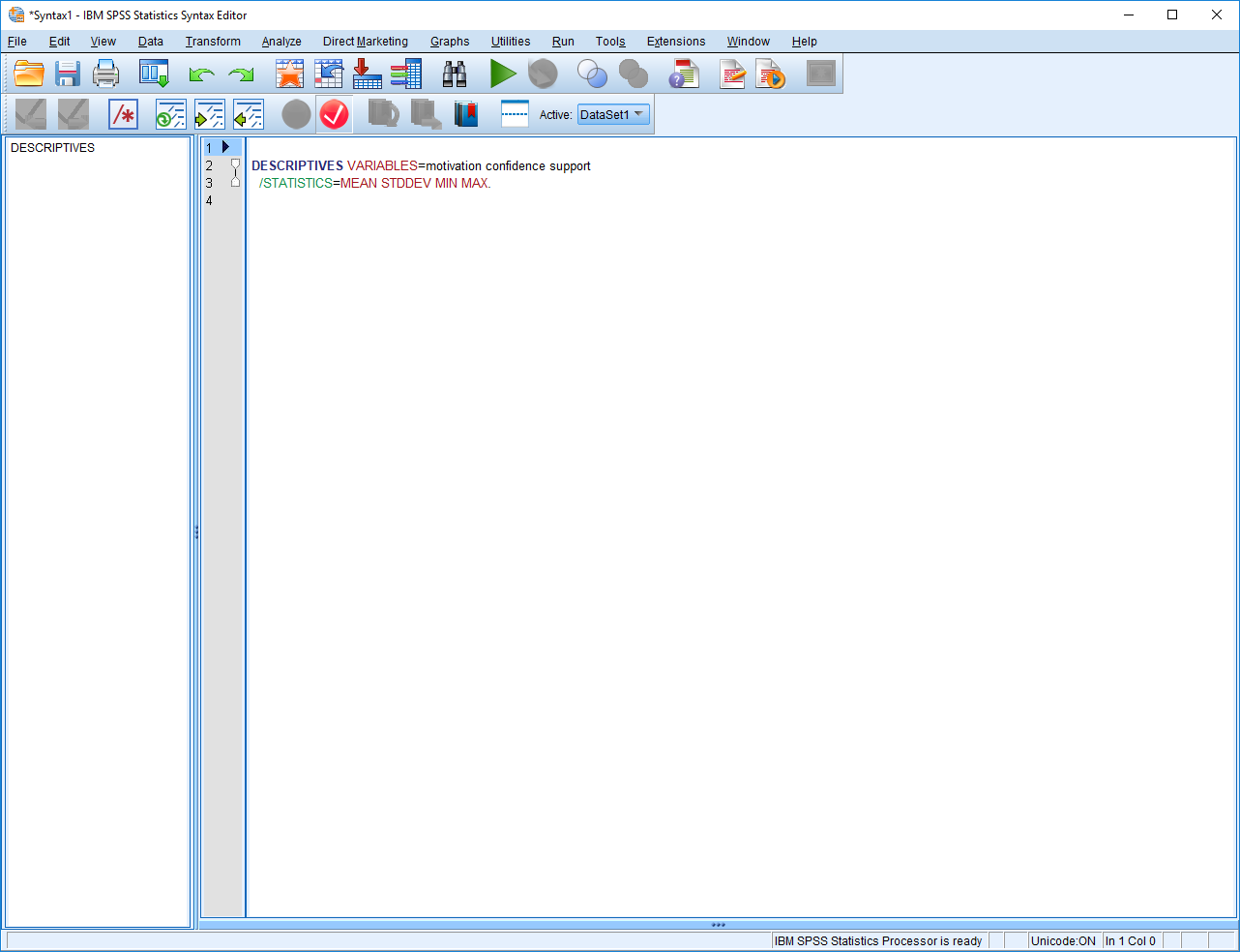
To complete these activities, researchers can either choose the relevant menus and options or they can construct a syntax file—similar to a computer code. Initially, some research candidates assume the syntax files will be confusing. But, after a few minutes, most researchers discover the syntax files are straightforward and efficient. For example, when researchers use these syntax files, they can

* modify and improve their analysis efficiently, without needing to select the various menus and options again
* repeat their analysis with another data set

This article will demonstrate how you can prepare your data using either the menus or the syntax files.

**How to use syntax files**

To learn how to use syntax files, enter some data into a data file. And then, begin some analysis. For example, you might choose “Analyse”, “Descriptive Statistics”, and “Descriptives” before entering a variable into the box called “Variables”. However, rather than press OK, press Paste instead, to generate the following code.



* Whenever you press “Paste”, SPSS generates the code—called syntax—that corresponds to this procedure
* Initially, this code might look illegible. But, you can probably roughly guess what the code indicates.
* For example, if you wanted to explore another variable, you would merely enter this variable alongside the other variables you chose in the top row
* You could use the “File” menu to save this file now or to open this file in the future
* To execute this code, simply highlight the code with your mouse and press the green arrow or the run menu.

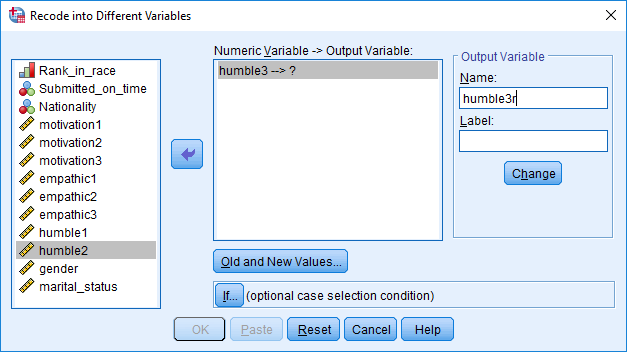
|  |
| --- |
| **1 Recode your data if necessary** |

Sometimes you need to modify some of your data—called recoding. The following table outlines some instances in which data might need to be recoded. After you scan this table, decide whether you might need to recode some of your variables.

|  |  |
| --- | --- |
| Reason to recode | Example |
| To blend specific categories into broader categories | The researcher might want to reduce married, de facto divorced, separated, widowed, or single to two categories: “living with a partner” versus “not living with a partner” |
| To create consistency across similar questions | To measure the humility of supervisors, participants indicate, on a scale of 1 to 10, the extent to which their supervisor   * admits faults * admits mistakes * conceals limitations   One participant might indicate 7, 8, and 3 on these three questions. In this instance,   * high scores on the first two questions, but low scores on the third question, indicates elevated humility. * therefore, the researcher should not merely sum these three responses to estimate the overall humility of the supervisor—because a high score might indicate the supervisors often admits faults and mistakes or often conceals limitations * to override this problem, the researcher could recode the responses to conceals limitations * in particular, on this item, the researcher could substract the score of participants from 11—one higher than is the maximum * a 9 would become a 2, a 2 would become a 9, and so forth * this procedure is called reverse coding, because high scores become low scores and vice versa   In contrast, if the responses spanned from 1 to 5, you would subtract each number from 6 to reverse code. |

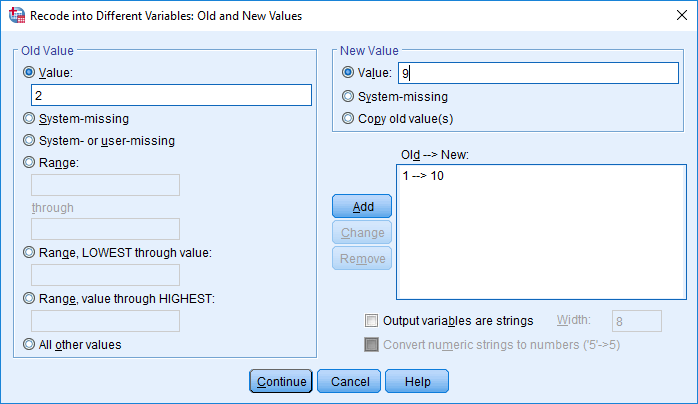
**How to use the menu and options to recode data**

To use the menus and options to recode data, choose the “Transform” menu and then “Recode into Different Variables”, to generate this box.



Then

* into the box called “Numerical Variable --> Output Variable”, specify the variable you want to recode, as shown in the previous example
* in the box called “Name”, enter a label for the new variable, such as “humble3r”—many researchers simply append an “r” to the end of previous label to indicate the data has been revised
* press “Change”
* click “Old and New values” to generate the following screen



Your task is now to convert the original values to the new values, such as 1 to 10. For example, you could

* enter “2” in the top box on the left, and enter “9” in the top box on the right
* press “Add” to transfer this translation from 2 to 9 into the box called “Old-->New
* apply to all the other changes, such as 3 to 8.
* then press continue and OK

**How to use syntax to recode data**

If you pressed “Paste” instead of “OK”, you would receive syntax that resembles the following:

|  |
| --- |
| RECODE humble3 (1=10) (2=9) (3=8) (4=7) (5=6) (6=5) (7=4) (8=3) (9=2) (10=1) INTO humble3r.  EXECUTE. |

You could then copy and paste the top row of this code and then modify the code slightly to modify other variables as well, as the following syntax shows.

|  |
| --- |
| RECODE humble3 (1=10) (2=9) (3=8) (4=7) (5=6) (6=5) (7=4) (8=3) (9=2) (10=1) INTO humble3r.  RECODE empathic3 (1=10) (2=9) (3=8) (4=7) (5=6) (6=5) (7=4) (8=3) (9=2) (10=1) INTO empathic3r.  RECODE marital\_status (1=1) (2=1) (3=0) (4=0) (5=0) (6=0) INTO marital\_statusr  EXECUTE. |

To implement this code, you would need to

* highlight all this code with your mouse
* press the green arrow or choose the run menu

In the future, to apply this procedure to other variables, you could merely copy, paste, and then modify this code accordingly. For example, you would update the name of these variables and the numbers you want to change. Once you implement this code

* return to your data file
* scroll to the right; the new columns, called humble3r, empathic3r, and marital\_statusr should appear.

|  |
| --- |
| **2 Assess internal consistency** |

Consider the following subset of data. Each row corresponds to one participant. The first three columns present answers to the three questions that assess the humility of supervisors, after recoding the third item. The final column presents the average of the other columns. In subsequent analyses, researchers will often utilize this final column—the average of several items— instead of the previous columns because

* trivial events, such as misconstruing one word, can appreciably affect the response to a specific question
* but, these events are not as likely to affect the average of several responses to the same extent
* that is, these averages tend to be more reliable or consistent over time

Consequently, researchers often compute the average, or sum, of a set of items or columns. This average or sum is sometimes called a composite scale or simply a scale.

|  |  |  |  |
| --- | --- | --- | --- |
| Humility 1  Admits their faults | Humility 2  Admits their mistakes | Humility 3r  Conceals their limitations after recoding | Average |
| 4 | 8 | 6 | 6 |
| 1 | 3 | 5 | 3 |
| 2 | 6 | 4 | 4 |

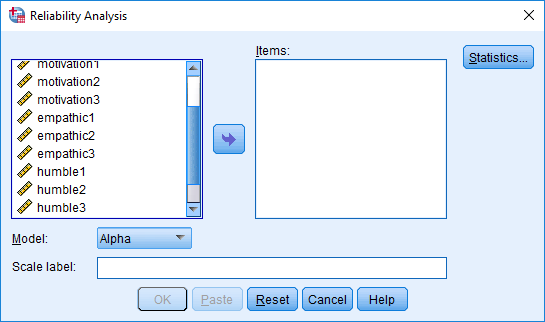
So, when should researchers construct these composite scales? That is, when should researchers integrate several distinct questions or items into one measure. Researchers tend to construct these composite scales when

* past research—such as factor analyses or similar techniques—indicates these individual questions or items correspond to the same measure or scale
* these questions or items are highly correlated with each other. That is, high scores on one item, such as “Admit their faults”, tend to coincide with high scores on the other items, such as “Admit their mistakes”

To determine whether these questions or items are highly correlated with each other, many researchers compute an index called a Cronbach’s alpha. Values above 0.7 on this index tend to indicate the questions or items are adequately related to each other.

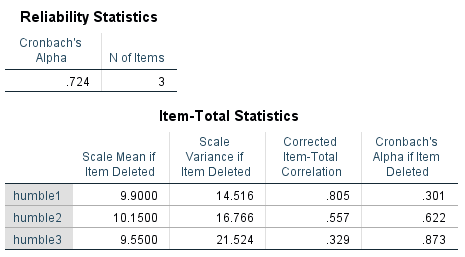
**How to use the menus and options to calculate Cronbach’s alpha**

To use the menus and options to calculate Cronbach’s alpha, choose the “Analyse” menu, then “Scale” and finally “Reliability analysis” to generate the following screen



To compute Cronbach’s alpha

* transfer the individual questions or items, such as humble1, humble2, and humble3r, into the box called “Items”
* click the “Statistics” button and choose “Scale if item deleted”.
* then press “Continue” and “OK” to generate the following output



As this output shows

* Cronbach’s alpha for this humility scale is .724
* according to Nunnally (1978), values above .7 indicate that Cronbach’s alpha is adequate; in other words, the three items correlate with each other to an adequate extent
* the researcher could thus combine these items to generate a composite scale

Yet, the final column in the second table uncovers a vital insight. According to this column, if the researcher had excluded the third item, Cronbach’s alpha would be appreciably higher at .83. And, when Cronbach’s alpha is appreciably higher, the results are more likely to be significant: power increases. So, should the researcher exclude this item from the composite?

* If the scale has been utilized and validated extensively, researchers are reluctant to exclude items
* If the scale has not been utilized and validated extensively, the researcher may exclude this item from subsequent analyses
* However, scales that comprise fewer than 3 items are often not particularly reliable or easy to interpret.
* Therefore, in this instance, the researcher would probably retain all the items.

**How to use syntax to calculate Cronbach’s alpha**

If you pressed “Paste” instead of “OK”, you would receive syntax that resembles the following:

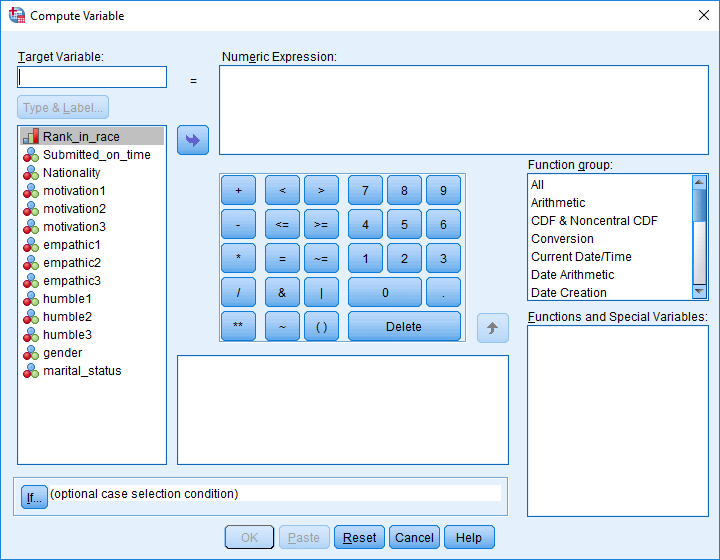
|  |
| --- |
| RELIABILITY  /VARIABLES=humble1 humble2 humble3r  /SCALE('ALL VARIABLES') ALL  /MODEL=ALPHA  /SUMMARY=TOTAL. |

To use this syntax

* change “humble1 humble2 humble3r” to the names of your items or questions
* if applicable, remember to include the reverse coded item instead of the original item
* repeat for each scale or subscale; that is, copy and then paste these five lines and then specify the names of your items or questions in the second line
* you can also replace ALL VARIABLES with the name of your composite scale, such as HUMILITY COMPOSITE

|  |
| --- |
| **3 Construct the scales** |

If the Cronbach’s alpha is sufficiently high, you can then compute the average or sum of these items. For example, you would choose the “Transform” menu and then “Compute variables” to generate this screen. This screen can be utilized to construct new columns, including composite scales, from previous columns.



To construct the scale

* In the box called “Target Variable”, enter a name for this new column or scale, such as “Humility”
* In the box called “Numeric expression”, enter the expression MEAN(humility1, humility2, humility3r), except specify your items instead of humility1, humility2, humility3r.
* Do not include the items or questions you had previously decided to exclude
* Press OK

This procedure will then generate an additional column in your datafile, called “Humility” in this example. To locate this column, scroll towards the right of this screen.

**How to use syntax to construct the composite scales**

The syntax will resemble the following

|  |
| --- |
| COMPUTE humble = (humble1, humble2, humble3r).  COMPUTE empathic = (empathic1, empathic2, empathic3r).  COMPUTE motivation = (motivation1, motivation2, motivation3)  EXECUTE. |

If you exclude the “EXECUTE” command, SPSS will calculate these composite scales but will not implement these changes to the data file. However, you need only include this command once to implement these changes to the data file.

**Mean versus sum or total**

When researchers construct these composite files, they might calculate the sum or total of specific items instead of the mean or average. For example, they could include the expression

* SUM(humility1, humility2, humility3r) or
* humility1 + humility2 + humility3r.

However, if possible, researchers should utilize the mean instead of the total for two reasons. First, the mean scores are easier to interpret:

* to illustrate, if the responses can range from 1 to 10, the mean of these items also ranges from 1 to 10.
* therefore, a researcher will immediately realize that a mean score of 1.5 is low, but cannot as readily interpret a total of 24

Second, the mean scores are accurate even if the participants had not answered all the questions. To demonstrate,

* if the participant had specified 3 and 5 on the first two items, but overlooked the third item, SPSS will derive the mean from the answered questions
* in this example, the mean will be 4
* but, in this example, SPSS will not derive a total but instead generate a missing cell.

**How to construct composites when the response options differ across the items**

In the previous examples, the responses to each item could range from 1 to 10. However, suppose you want to combine these two items

* what is your height in cm?
* what is your shoe size?

If you constructed the mean of these two items, the final composite would primarily depend on height rather than shoe size. Instead, whenever the range of responses differs between the items you want to combine, you should first convert these data to z scores and then average these z scores. To achieve this goal

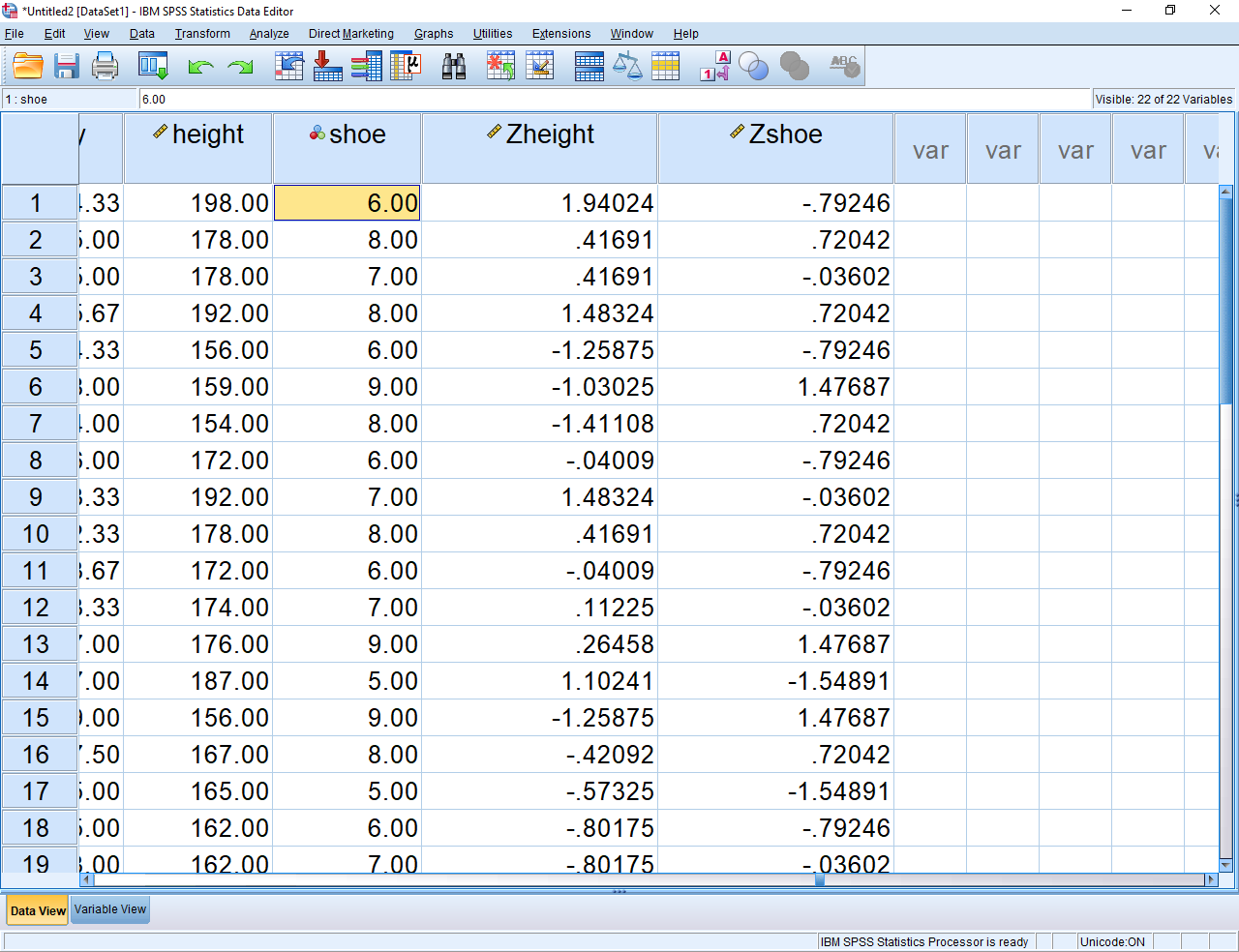
* choose “Analyse”, “Descriptive Statistics”, and “Descriptives” before entering these items into the box called “Variables”.
* tick “Save standardized values as variables” and press OK
* Or modify the following syntax.

|  |
| --- |
| DESCRIPTIVES VARIABLES=height shoe  /SAVE  /STATISTICS=MEAN STDDEV MIN MAX. |

This procedure will generate additional columns in your data file, each beginning with the letter Z, as the following example shows. Specifically

* To generate these columns, SPSS calculates the mean and standard deviation of the original column, such as height or show
* Then, SPSS deducts this mean from the original scores and divides by the standard deviation
* The new columns are called z scores
* By definition, the mean of these columns is 0 and the standard deviation is 1.

These two added columns comprise the same standard deviation and, therefore, can be blended into a composite—with the syntax COMPUTE size = mean(Zheight, Zshoe).



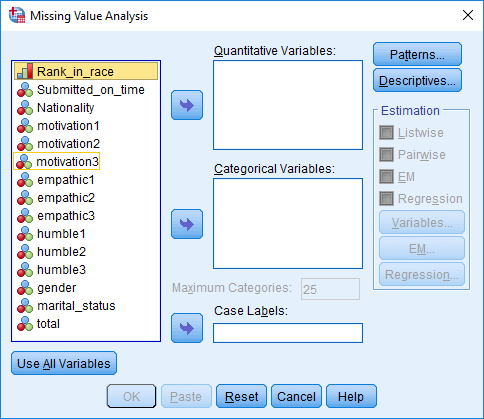
|  |
| --- |
| **4 Manage missing data** |

In many data sets, some of the data are missing. Participants might overlook some questions, for example. However

* if participants have overlooked some, but not all, the items or questions on a composite scale, you do not need to be too concerned; SPSS will derive the mean from the items or questions that have been answered
* if participants had overlooked all the items or questions on a composite scale—or overlooked a measure that is not a composite scale—you need to manage these missing data somehow

**How to use the menus and options to manage missing data**

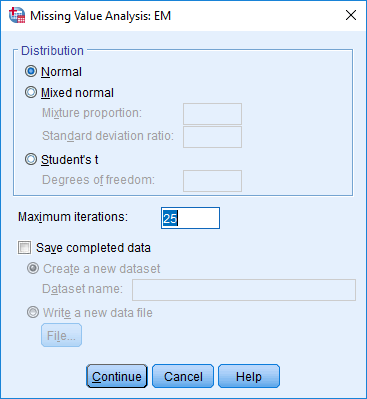
Researchers have developed a variety of techniques to manage missing data. One technique that is suitable in many circumstances is called expectation maximization. To conduct this technique, choose the “Analyze” menu and then “Missing Value Analysis” to generate the following screen.



Then, in the box called “Quantitative Variables”,

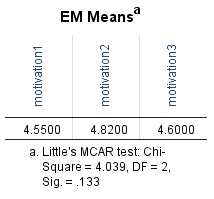
* specify your numerical variables—that is, variables in which everyone is assigned a real number, such as weight, height, or most composite scales
* however, if you want to include composite scales, exclude the specific items or questions these scales entailed
* exclude variable that are not relevant to your analyses.

Next, in the box called “Categorical Variables”, specify your categorical variable—that is, variables in which everyone is assigned a category or name; to illustrate, for the variable gender, participants might be assigned the categories male, female, or other. Finally, tick the “EM” option on the right side, and select the “EM” button to generate the following screen



This screen enables you to construct a new data file—a data file that is identical to the original data except, somehow, the missing data has been replaced with actual data. To generate this data file

* tick “Save completed data” and then choose “Write a new data file”.
* press “file” and then type the name of a new data file—a file that you store in a particular directory on a hard drive, USB, or some other location.
* after you press Continue and then OK, you should be able to locate this data file in this directory.
* in addition, this procedure will generate a series of tables in the output, including the following



This table offers some insight as to whether expectation maximization is valid in these circumstances, to utilize this table

* if this p value is not significant, the data are called “completing missing at random”.
* expectation maximization is thus applicable; for the remainder of your analyses, use the new data file you constructed in which missing data was replaced with numbers
* if this p value is not significant, expectation maximization may not be suitable.
* Instead, perhaps you need to analyze a data file that comprises missing data

**Syntax to manage missing data**

If you press paste, instead of OK, the syntax will resemble the following

|  |
| --- |
| MVA VARIABLES=motivation1 motivation2 motivation3 empathic2 empathic3 humble1 gender haircolor  /CATEGORICAL=gender haircolor  /EM(TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25 TDF=20 OUTFILE='C:\Users\smoss\Newfile.sav'). |

When using this syntax, note that

* you must include all relevant variables after “MVA VARIABLES=” but only the categorical variables after “CATEGORICAL=”
* you can change C:\Users\smoss\Newfile.sav to a directory and filename that is more relevant to you research.

**Rationale behind expectation maximization**

Expectation maximization, although simple to apply in practice, is underpinned by a complex rationale. In essence

* expectation maximization utilizes the responses on other measures to guess what the missing data would have been had the participant not overlooked the item or question
* expectation maximization overcomes some of the limitations of other techniques; for example, this approach introduces errors to the variances and covariances to generate more accurate estimates of the missing data

**Listwise versus pairwise**

If you cannot utilize expectation maximization but your data file contains missing data, most techniques are still valid, despite some exceptions. To illustrate

* if you conduct repeated-measures ANOVAs, SPSS will simply disregard the rows that contain missing data on relevant variables. A more sophisticated technique, called multi-level analysis, circumvents this problem
* in other instances, the researcher can choose between several options.

For example, suppose you plan to conduct multiple regression, factor analysis, or several other techniques. While conducting these techniques, you will usually be granted the option to choose “Exclude cases listwise”, “Exclude cases pairwise”, or perhaps one or more other alternatives. Roughly speaking

* “Exclude cases listwise” disregards rows of data that yield one or more instances of missing data on a relevant variable—a variable you have included in the analysis
* “Exclude cases pairwise” utilizes the data from these participants wherever possible.

To clarify this pairwise alternative, suppose that 50 individuals completed Questions 1 and 2 but only 45 individuals completed Question 3. When conducting multiple regression, factor analysis, or several other techniques, SPSS initially calculates the correlation between all relevant variables. SPSS with thus

* derive the correlation between Questions 1 and 2 from all 50 participants
* derive the correlation between Questions 1 and 3 and between Questions 2 and 3 from 45 participants only
* SPSS will then utilize these correlations to complete the subsequent steps in the analysis.

So, which of these alternatives should you apply? In general, use “Pairwise” whenever possible, because this option is more likely to generate significant results. However, if “Pairwise” generates an error, use “Listwise” instead.

* To choose pairwise while using syntax, insert the row “ /MISSING PAIRWISE” somewhere within the command.
* Otherwise, insert the row “ /MISSING LISTWISE”

|  |
| --- |
| **5 Examine redundancies or multi-collinearity** |

When researchers conduct analyses, one or more variables may be somewhat redundant. For example, suppose a researcher wants to assess an interesting theory. According to this theory, if supervisors are tall, research candidates might feel more supported by an influential person, enhancing their motivation. To test this possibility, 100 research candidates complete questions in which they indicate

* their level of motivation, on a scale from 1 to 10
* the height of their supervisor
* the shoe size of their supervisor

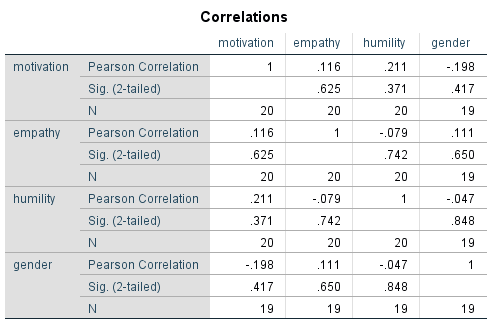
The problem, however, is that height and shoe size are highly correlated with each other. If someone is tall, their feet tend to be long. If someone is short, their feet tend to be small. Two variables that are highly related to each other are called multi-colinear. In these circumstances

* including both height and shoe size will diminish the likelihood that either variable is significantly associated with candidate motivation
* in other words, multi-collinearity reduces statistical power
* instead, researchers should either discard one of these variables, such as shoe size, or somehow combine these variables into one composite, as shown previously.

**How to use the menus and options to compute correlations**

To identify multi-collinearity, one simple method is to calculate the correlation between all the variables you plan to include in your analyses. To achieve this goal, you could

* Choose the “Analyze” menu and then select “Correlate” and finally “Bivariate”
* Transfer all the variables you plan to subject to the analyses—such as the composite scales and other important variables—to the box called “Variables”
* Include all numerical variables and dichotomous variables—variables in which participants can be assigned one of two categories
* Press OK to generate an output that resembles the following table, called a correlation matrix



* In this instance, none of the correlations are especially high. For example, the correlation between motivation and humility is .211
* Correlations about 0.8 might indicate multi-collinearity and could reduce power, especially if these variables are all predictors or independent variables
* Correlations above 0.7 could also be high enough to reduce power, particularly if the sample size is quite small, such as less than 100.

**How to use syntax to compute correlations and manage large correlation table**

If you press paste, instead of OK, the syntax will resemble the following

|  |
| --- |
| CORRELATIONS  /VARIABLES=motivation empathy humility gender  /PRINT=TWOTAIL NOSIG  /MISSING=PAIRWISE. |

Sometimes, the correlation matrix comprises too many variables and is thus unwieldy. You could potentially divide the matrix into three tables:

* the first table examines the correlations between one set of variables--perhaps the outcome measures.
* the second table examines the correlations between the remaining variables.
* the third table demonstrates how the first set correlates with the second set. To construct this third table, insert the term “WITH” between the two sets, as the following syntax indicates.
* note, in this syntax, SPSS will disregard rows of text that begin with an asterisk. This text is usually comments, recorded by the researcher.

|  |
| --- |
| \*Set 1  CORRELATIONS  /VARIABLES=motivation empathy  /PRINT=TWOTAIL NOSIG  /MISSING=PAIRWISE.  \*Set 2  CORRELATIONS  /VARIABLES= humility gender  /PRINT=TWOTAIL NOSIG  /MISSING=PAIRWISE.  \*Set 3  CORRELATIONS  /VARIABLES=motivation empathy WITH humility gender  /PRINT=TWOTAIL NOSIG  /MISSING=PAIRWISE. |

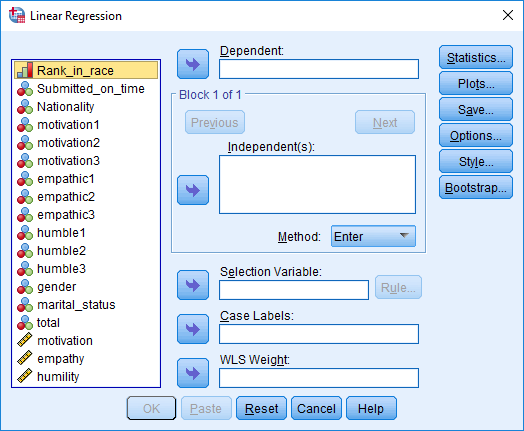
**Other measures of multi-collinearity: Variable inflation factor**

Unfortunately, these correlations do not uncover all instances of multi-collinearity. To illustrate, suppose that

* the researcher wants to construct a new variable, called compassion, equal to empathy + humility—as the following table shows
* surprisingly, compassion might only be moderately correlated with empathy and humility
* thus, a variable might be only moderately correlated with other variables—but highly correlated with a combination of other variables
* yet, even this pattern represents multi-collinearity and diminishes power
* indeed, if one variable is derived from of other variables in the analysis, SPSS will generate an error message. This pattern is called singularity and is tantamount to extreme multicollinearity

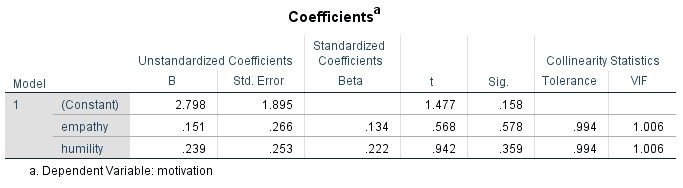
|  |  |  |
| --- | --- | --- |
| Empathy | Humility | Compassion |
| 8 | 6 | 14 |
| 3 | 5 | 8 |
| 6 | 4 | 10 |

Because you might not be able to extract these patterns from the correlations, you might need to calculate other indices instead. Typically, researchers calculate these indices while, rather than before, they conduct the main analyses. To illustrate, if conducting a linear or multiple regression analysis, you would first choose the “Analyse” menu and then select “Regression” and finally “Linear” to generate this screen



Then

* transfer your outcome measure to the box labelled “Dependent”
* transfer your predictors to the box labelled “Independents
* select “Statistics” and then tick “Colinearity diagnostics” before pressing “Continue” and “OK” to generate output that resembles the following table



The final column, VIF, represents “Variable inflation factor”. To interpret these figures

* a VIF that exceeds 5 indicates multicollinearity—and suggests one or more predictors need to be omitted or combined; ; a VIF that exceeds 10 is especially concerning
* strictly speaking, VIF is the variance of a regression coefficient divided by what the variance of this coefficient would have been had all other predictors been omitted
* if the other predictors are uncorrelated, VIF will equal 1
* if the other predictors are correlated, VIF exceeds 1

If you had pressed paste, the syntax would resemble the following

|  |
| --- |
| REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN  /DEPENDENT motivation  /METHOD=ENTER empathy humility. |

**How to manage instances of multi-collinearity**

If you do uncover multi-collinearity, you could exclude one of the variables from subsequent analyses or combine items or scales that are highly related to each other. To combine these items or scales, apply the procedures that were discussed in the previous section called “Construct the scales”.

|  |
| --- |
| **6 Identify outliers** |

**Classes of outliers**

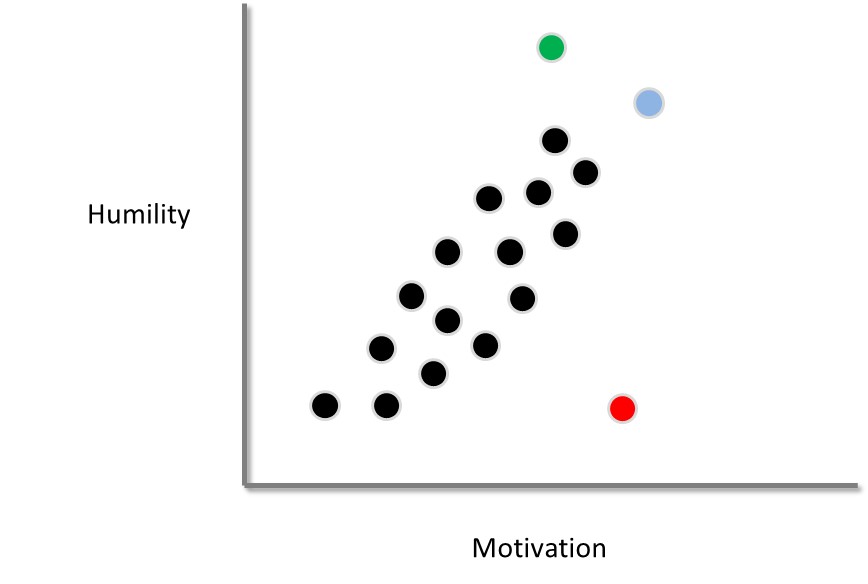
Finally, you need to identify and address the issue of outliers. An outlier is a score, or set of scores, that departs markedly from other scores. Researchers sometimes differentiate univariate outliers, multivariate outliers, and influential cases. The following table defines these three kinds of outliers.

|  |  |
| --- | --- |
| Kind of outlier | Definition |
| Univariate outlier | * A univariate outlier is an extreme score on one variable—a score that is appreciably higher or lower than all the other scores on that variable |
| Multivariate outlier | * A multivariate outlier is a combination of scores in one row—such as one person—that differs appreciably from similar combinations in other rows |
| Influential cases | * An influential case is a person, animal, or other row in the data file that greatly affects the outcome of a statistical test |

To differentiate these three kinds of outliers, consider the following graph. In this graph, each dot represents a different research candidate. The green dot for example, is probably a univariate outlier—humility is very high in this candidate relative to other candidates. However,

* the blue dot may be a multivariate outlier;
* this dot is not excessively high on humility and motivation; yet, the combination of humility and motivation seems quite high relative to everyone else
* nevertheless, the blue dot is consistent with the overall pattern and, therefore, might not change the results greatly.

The red dot, however, seems to diverge from the overall pattern and, therefore, might shift the results significantly. This red dot might thus be a multivariate outlier and an influential case.



**Causes or sources of outliers**

Outliers can be ascribed to one of three causes:

* Outliers might represent errors—such as mistakes in data entry
* Outliers might indicate the person or unit does not belong to the population of interest. For example, the red dot might correspond to a school candidate, instead of a research candidate, who received this survey in error
* Outliers could be legitimate; in the population, some people are just quite distinct.

**Effects of outliers**

Outliers, even if legitimate rather than mistakes, can generate complications and should perhaps be omitted. In particular

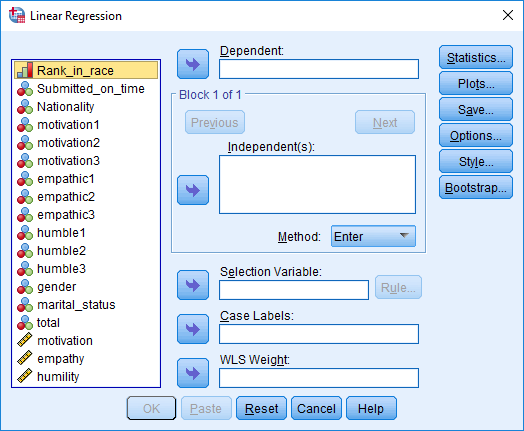
* influential cases in particular reduce the reliability of findings; if this outlier had not been included, the results might have been very different
* when the data comprises outliers, the assumption of normality is typically violated; hence, the p values tend to be inaccurate
* outliers can increase the variability within group and, therefore, can sometimes diminish the likelihood of significant results

**How to identify outliers**

To help identify univariate outliers, you should first

* select the “Analyze” menu, “Descriptive statistics” and “Frequencies” to assess the frequency of each variable. To illustrate, if the responses on some variable are supposed to range from 1 to 10, an 11 would indicate an error

To identify multivariate outliers, you could calculate a statistic called the Mahalanobis distance. To achieve this goal, choose the “Analyse” menu and then select “Regression” and finally “Linear” to generate this screen—regardless of which technique you plan to conduct later.



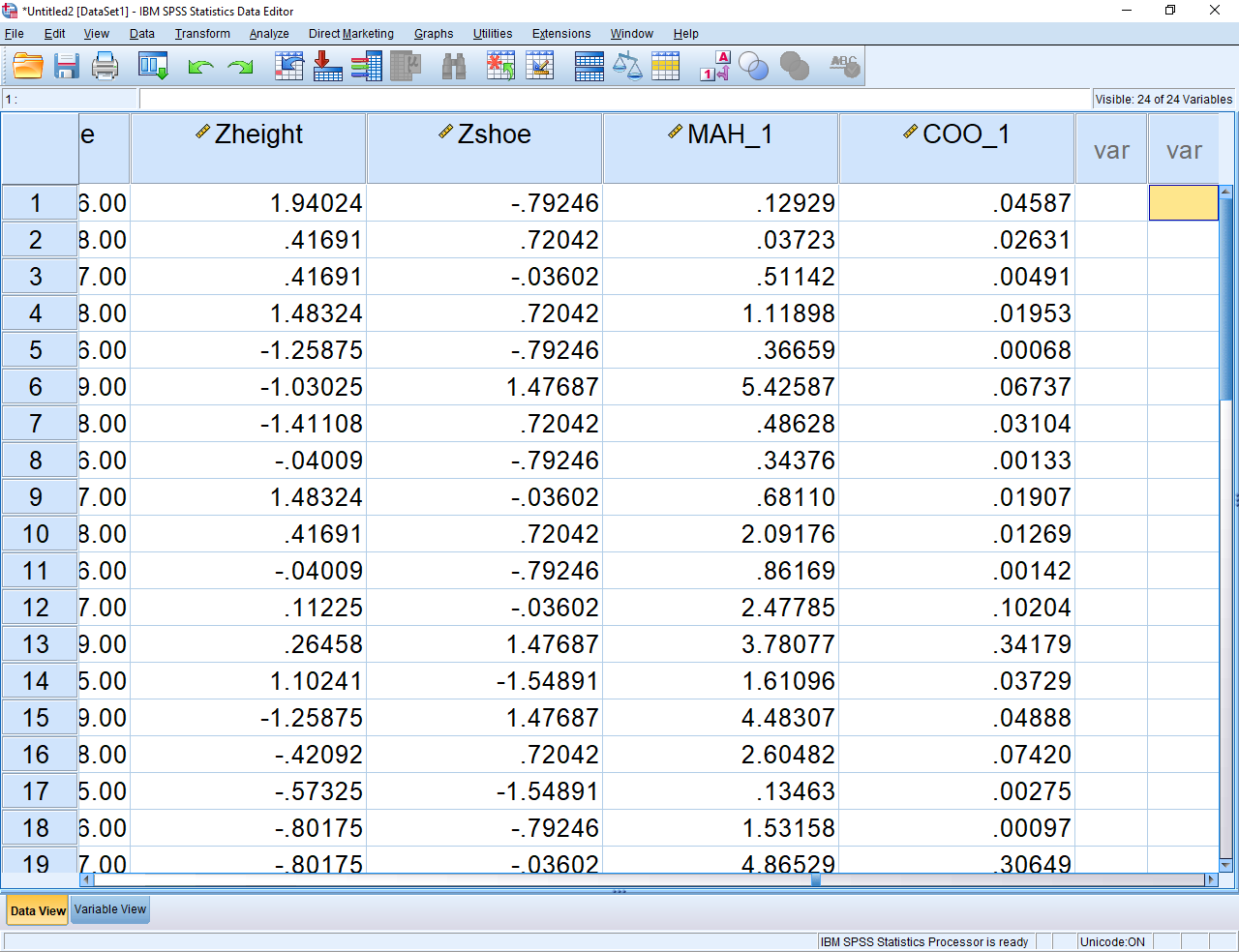
Then

* transfer all your numerical variables and dichotomous variables of interest into the box labelled “Independents
* transfer some irrelevant variable, such as ID, into the box labelled “Dependent”. You may need to create a new, artificial column if all the variables in the data file are indeed relevant.
* select “Save” and tick “Mahalanobis” and “Cooks” before clicking “Continue” and “OK”

If you had pressed paste, the syntax would resemble the following

|  |
| --- |
| REGRESSION  /MISSING PAIRWISE  /STATISTICS COEFF OUTS R ANOVA  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN  /DEPENDENT motivation  /METHOD=ENTER empathy humility  /SAVE MAHAL COOK. |

Besides a series of tables, this procedure will also generate additional columns in your data file, labelled MAH\_1 and COO\_1 respectively, as shown below



MAH\_1, or the Mahalanobis distance, represents the extent to which each row or participant differs from the other rows or participants. To identify which of these rows or participants are outliers

* open Microsoft Excel. Type "=CHIINV(0.01, 50)" in one of the cells--that is, type everything that appears within these quotation marks
* change 50 to the number of Independent Variables in the analysis—or variables that appeared after 'ENTER' in the previous syntax. This number corresponds to the degrees of freedom
* A value will then appear in the cell.
* Mahalanobis values that appreciably exceed this value are outliers at the p < .01 level.

These outliers should be excluded from subsequent analysis. That is, you could delete the row, and save the data file with another name.

**Influential cases**

The Mahalanobis distances will signify multivariate outliers but not necessarily all influential cases. The method you should use to generate influential cases varies across techniques. That is

* for some techniques, influential cases are hard to identify
* for linear or multiple regression, influential cases are easy to identify
* simply repeat the previous regression—except specify the outcome measure in the box called “Dependent Variable” and the predictors in the box called “Independent Variables”
* the column called COO\_1 represents an index called Cooks Distance
* perhaps use the “Data” menu and “Sort cases” to arrange the rows from the highest to the lowest Cook’s distance.

If a Cook’s distance exceeds 1, or is substantially higher than almost all the other Cook’s distances in the data file, the corresponding row or participant is an influential case. You should repeat the analysis but after excluding this participant.

|  |
| --- |
| **References** |

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98-104.

Nunnally, J. C. (1978). Psychometric theory (2nd edition). New York: McGraw-Hill.