Calculating Advanced Statistics: Part II

Collection Editor:
National Council of Professors of Educational Administration

Calculating Advanced Statistics: Part II

Collection Editor:
National Council of Professors of Educational Administration

Authors:
Ana Rojas-LeBouef
John R. Slate

Online:
< http://cnx.org/content/col11348/1.2/ >

C O N N E X I O N S

Rice University, Houston, Texas

Table of Contents

Chapter 1

Multivariate Analysis of Variance: Part I[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

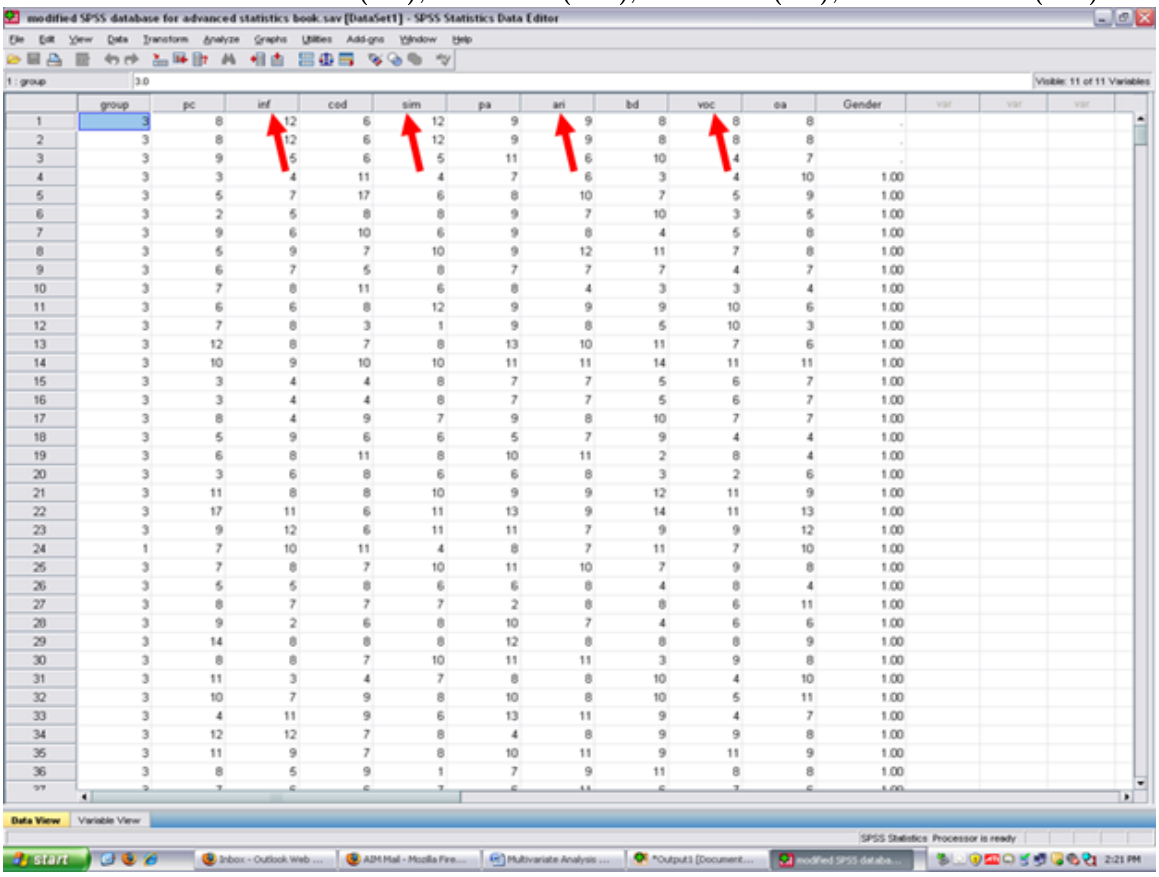
In this set of steps, readers will calculate a multivariate analysis of variance procedure, following the determination of the extent to which data for the dependent variables reflect normal distributions. Although a parametric statistical procedure requires that its data be reflective of a normal curve, the multivariate analysis of variance procedure is

[1] This content is available online at <http://cnx.org/content/m40728/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40728/latest/www.writingandstatisticalhelp

regarded as being sufficiently robust that it can withstand most violations. For detailed information regarding the assumptions underlying the multivariate analysis of variance (MANOVA) procedure, readers are referred to the Hyperstats Online Statistics Textbook at http://davidmlane.com/hyperstat/[4] ; to the *Electronic Statistics Textbook* (2011) at http://www.statsoft.com/textbook/[5] ; or to Andy Field's (2009) Discovering Statistics Using SPSS at http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967862&sr=1-1[6]

Research questions for which a MANOVA procedure is appropriate involve asking for differences in multiple dependent variables by group membership (i.e., more than two groups may be present). In addition to multiple dependent variables being present, multiple independent variables can be present as well. That is, differences in several achievement variables could be analyzed by student gender, student ethnicity, student socioeconomic status, and the like. A specific research question that could be addressed is, "What is the difference in academic achievement among elementary school students as a function of ethnic membership, gender, and grade level?" Academic achievement in this example could be reading, math, science, and social studies scores. The independent variables are ethnicity, gender, and grade level.

For this particular chapter, the research question on which we will conduct a MANOVA will be: What is the difference in verbal aptitude among elementary reading groups? Verbal aptitude consists of Verbal 1 (inf), Verbal 2 (sim), Verbal 3 (ari), and Verbal 4 (voc).



[7]

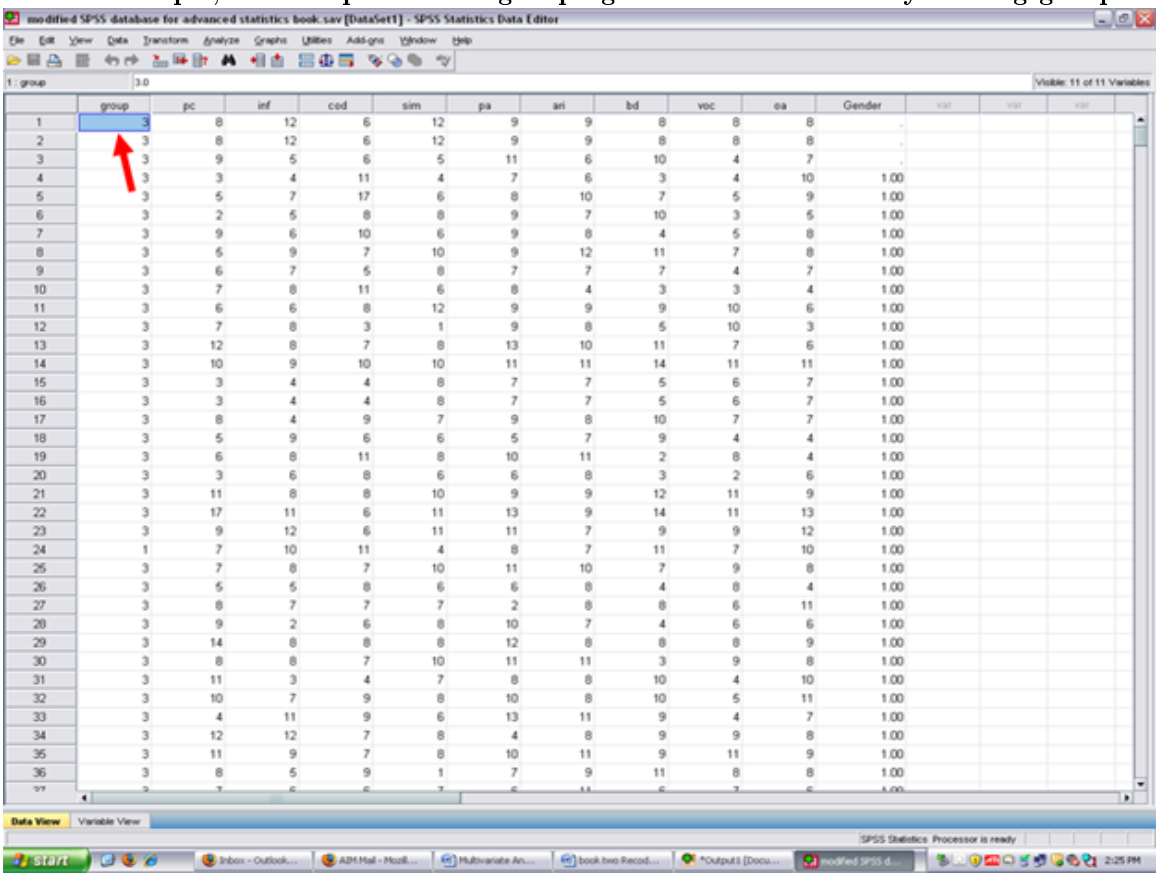
[4] http://davidmlane.com/hyperstat/

[5] http://www.statsoft.com/textbook/

[6] http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967862&sr=1-1

[7] http://cnx.org/content/m40728/latest/10.1.png/image

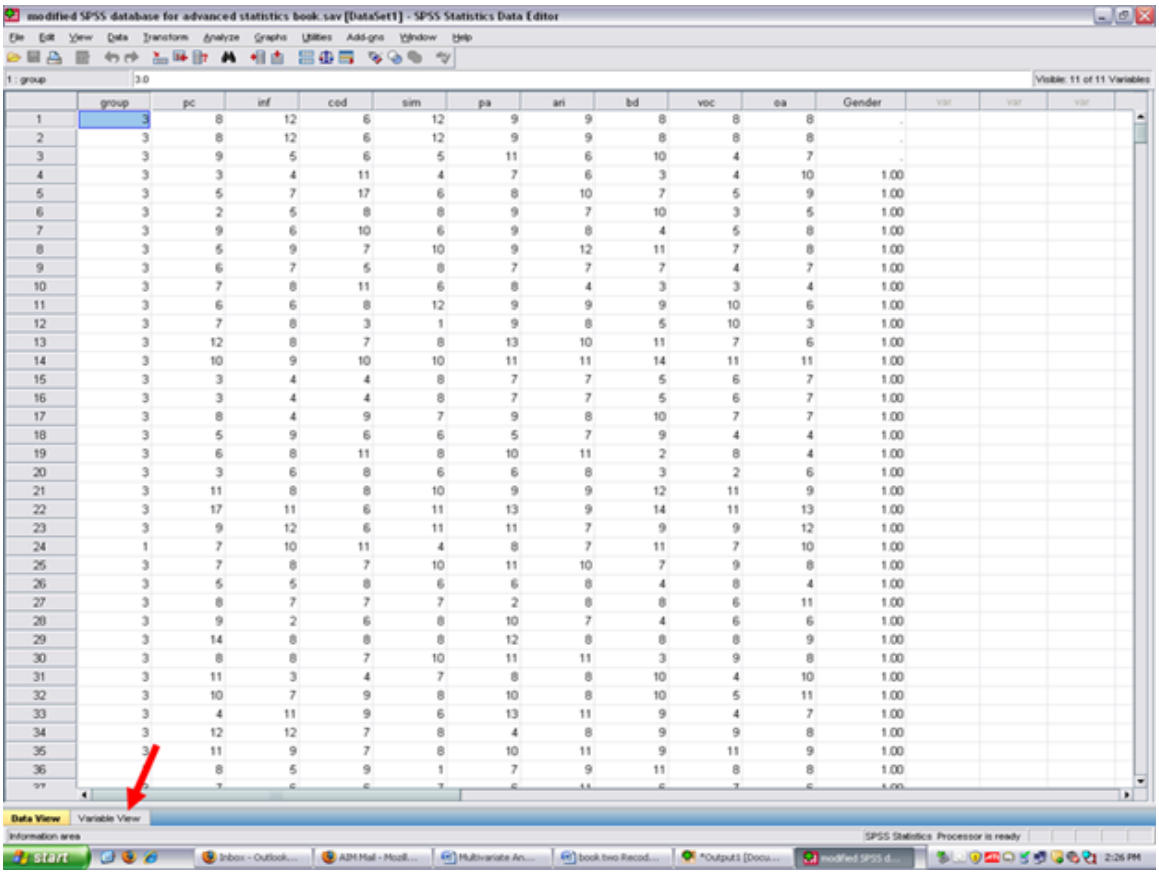
In this example, our independent or grouping variable is elementary reading group.



[8]

By clicking on variable view, you can see how many groups of students are present in the reading group variable; the names of the groups; and the numbers which have been assigned to each group.

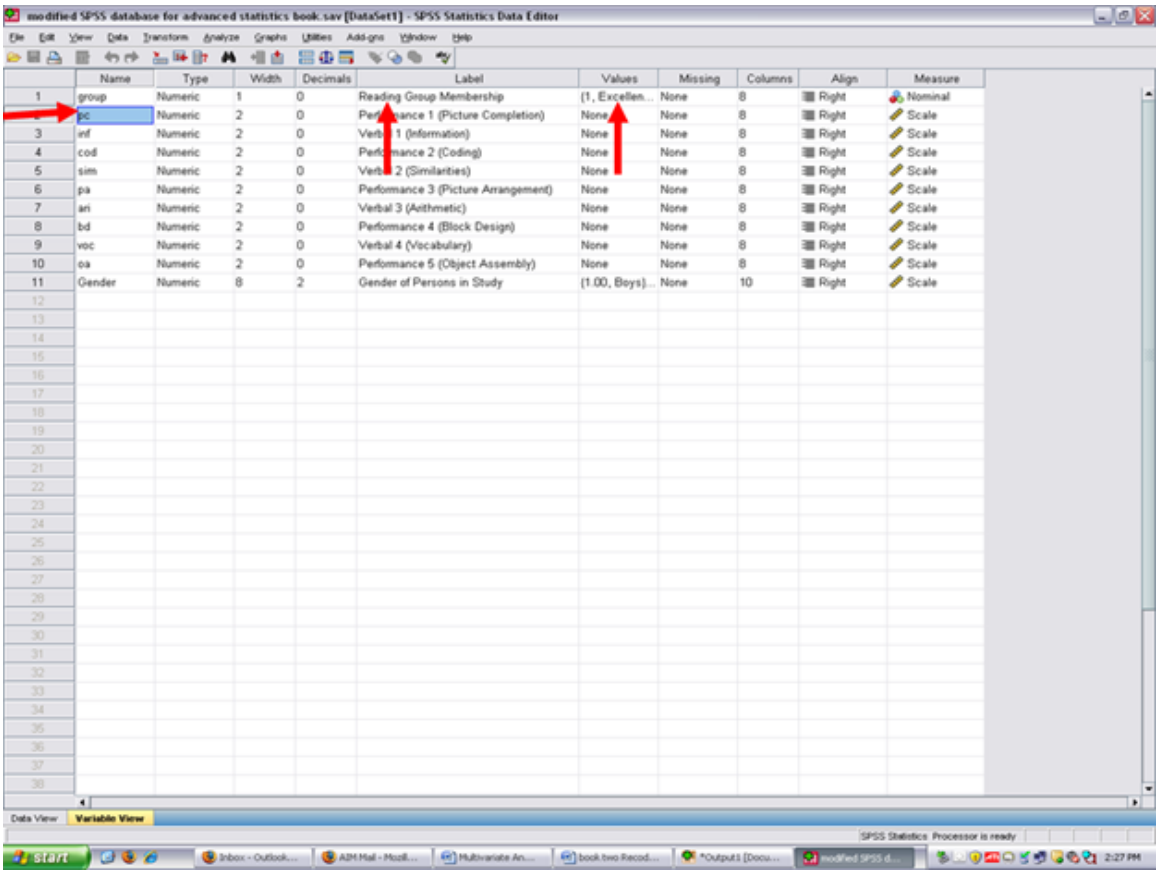
[8] http://cnx.org/content/m40728/latest/10.2.png/image



[9]

After clicking on Variable View, the following screen will appear: Arrows have been placed toward the name of the independent or grouping variable; toward the label assigned to this variable; and then toward the values of each group.

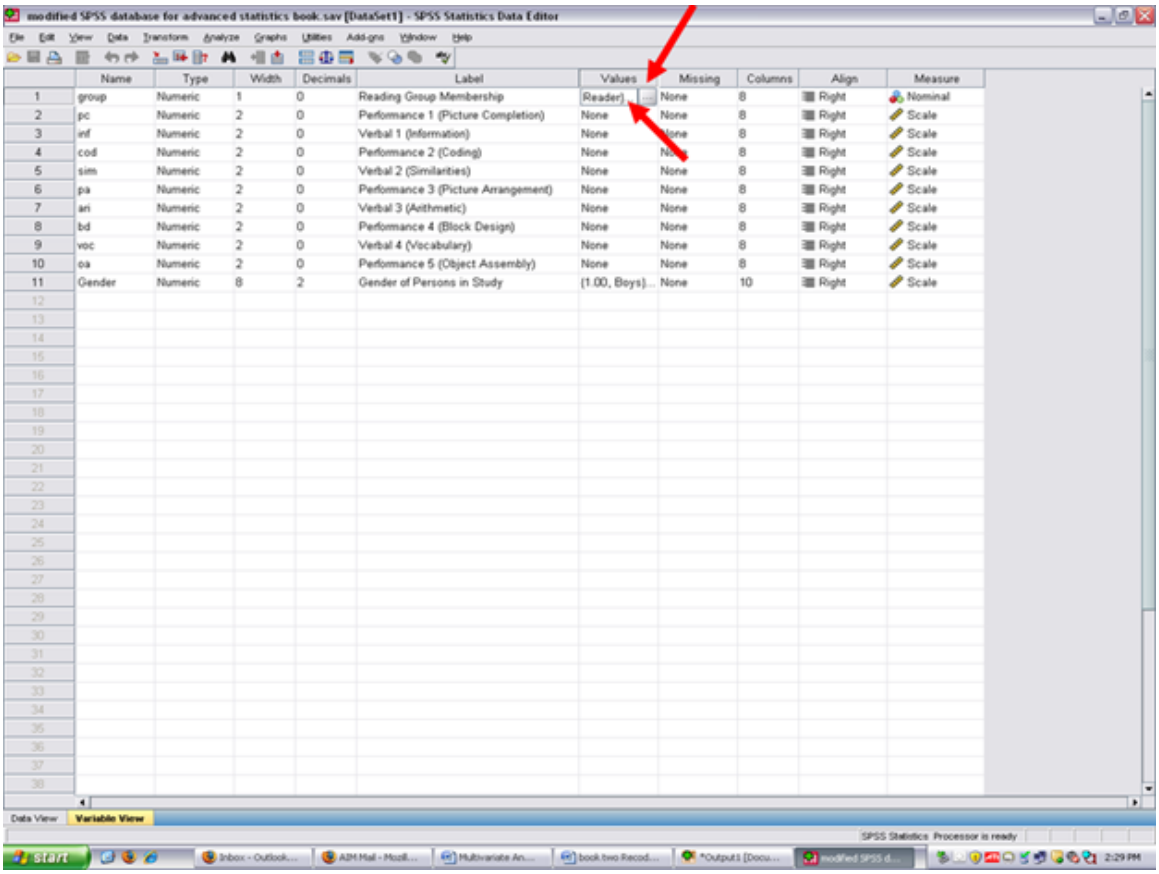
[9] http://cnx.org/content/m40728/latest/10.3.png/image



[10]

Now click on the cell for Reading Group Membership. Three dots will appear, indicating that another screen is beneath this one.

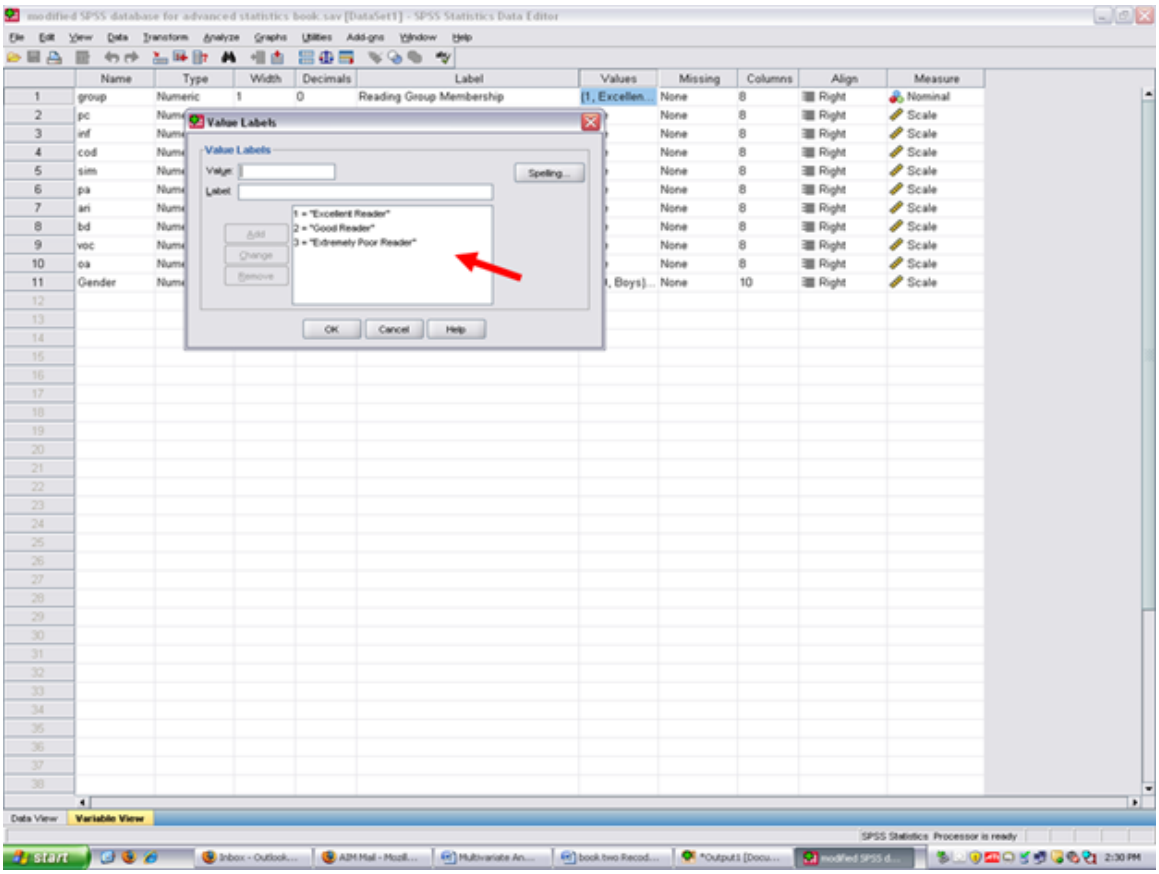
[10] http://cnx.org/content/m40728/latest/10.4.png/image



[11]

Click on the three dots shown in the screen above and the value and name of each of the reading groups will be shown. Our three groups are Excellent Reader (1), Good Reader (2), and Extremely Poor Reader (3).

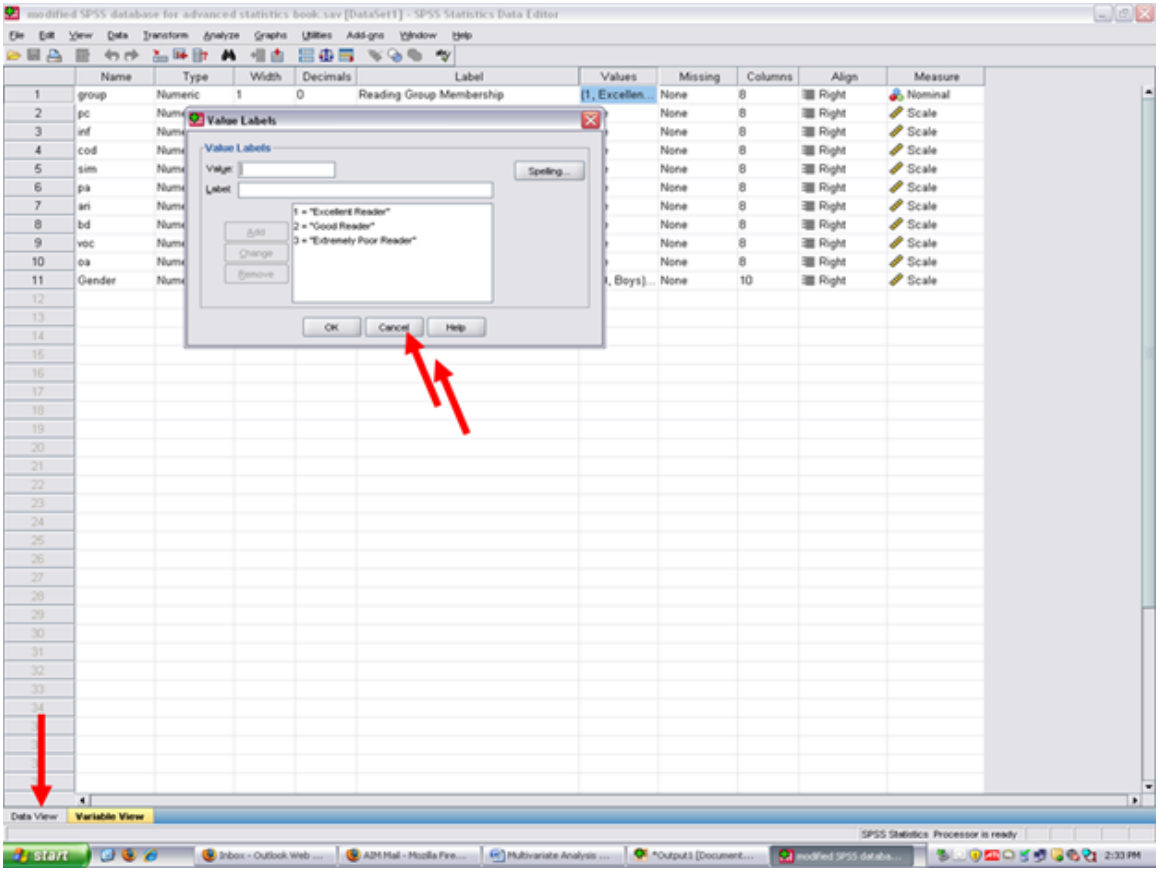
[11] http://cnx.org/content/m40728/latest/10.5.png/image



[12]

Now that you know what your independent (grouping) variable is and your dependent variables are, it is time to determine the extent to which the dependent variable data are normally distributed. Cancel out of the screen above. Then click on data view.

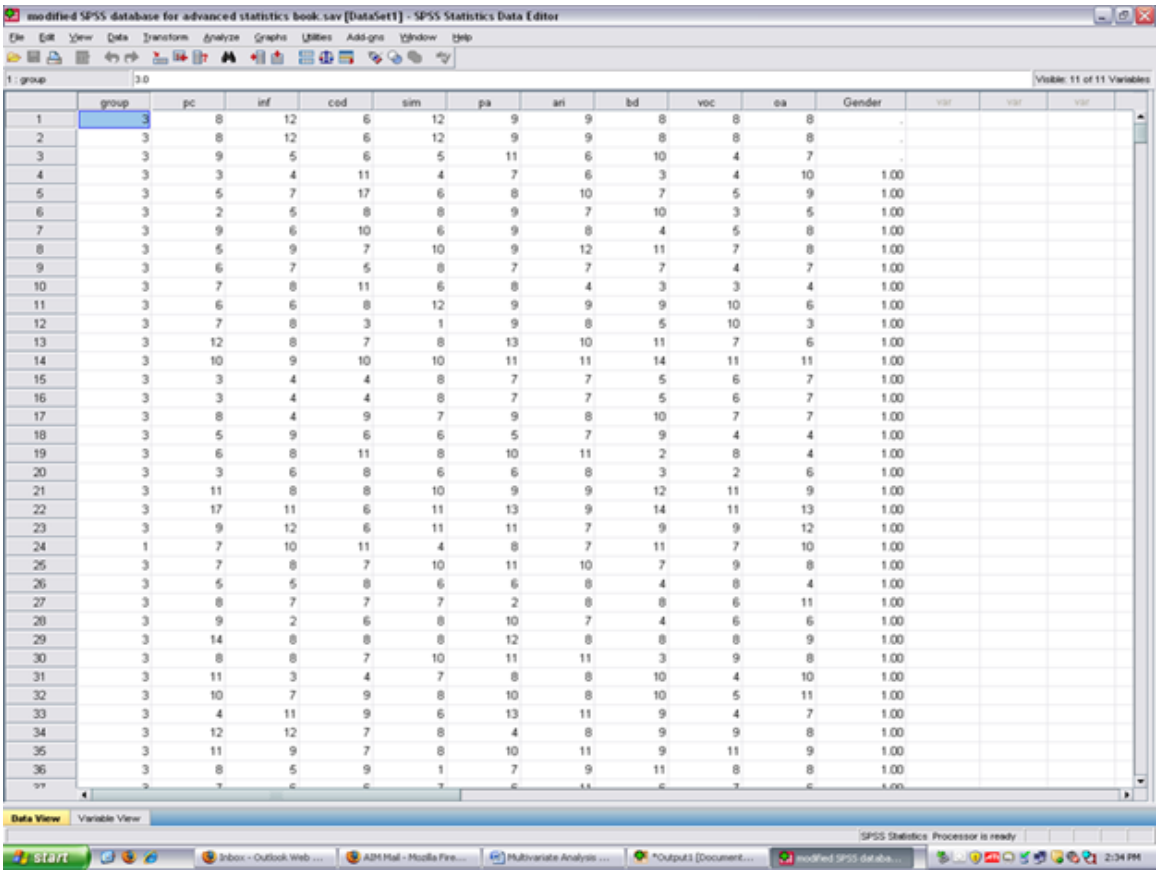
[12] http://cnx.org/content/m40728/latest/10.6.png/image



[13]

Your screen should now look like the one below.

[13] http://cnx.org/content/m40728/latest/10.7.png/image

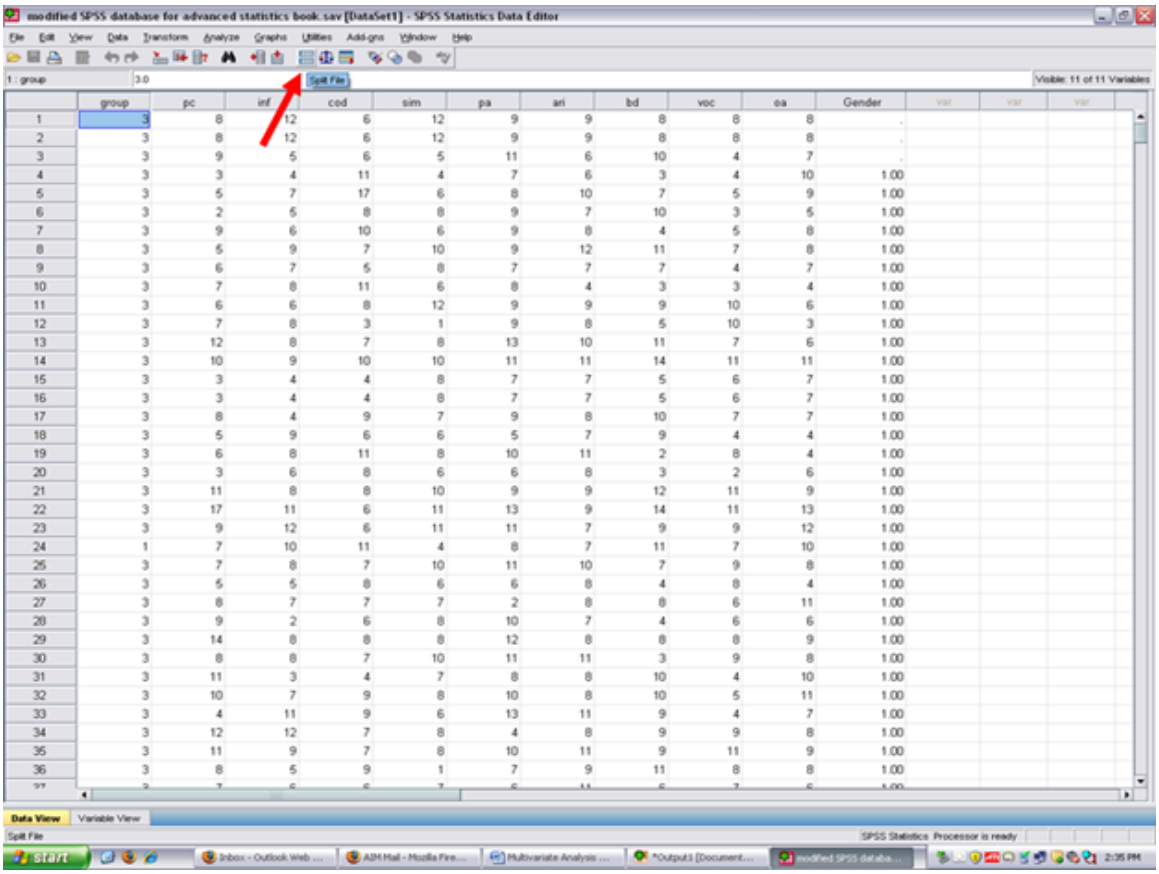


[14]

Check for Skewness and Kurtosis values falling within/without the parameters of normality (-3 to +3). To do so, you need to obtain these values for each of your three groups for each of your four dependent variables.

√ Split your file on the basis on your independent variable/fixed factor/grouping variable. To do so, click on the icon next to the scales. Holding your cursor on it will reveal Split File, as shown below.

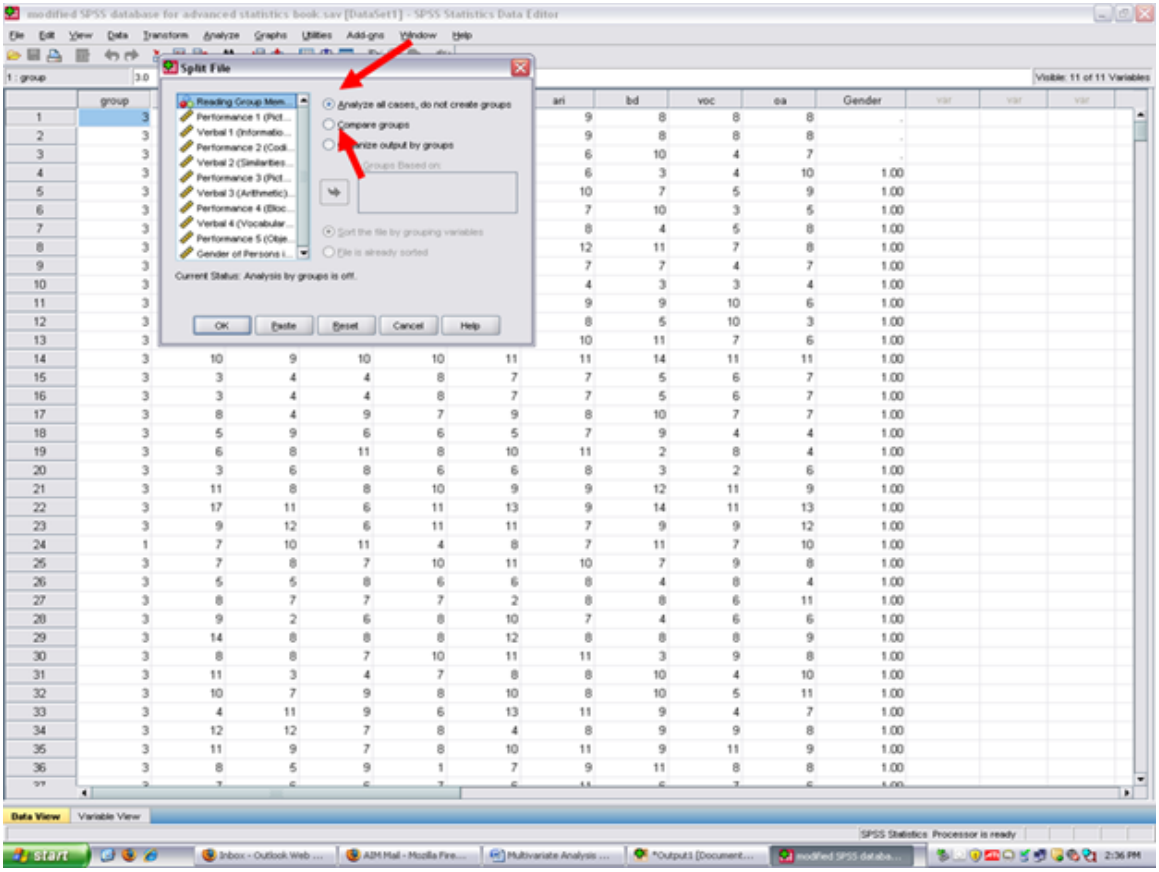
[14] http://cnx.org/content/m40728/latest/10.8.png/image



[15]

After clicking on the split file icon, the following screen will appear: You will note that the Analyze all cases, do not create groups is checked. This value is the default for SPSS as all cases are analyzed, unless otherwise specified. To obtain measures of normality for each of the three reading groups, the Compare groups button will need to be clicked.

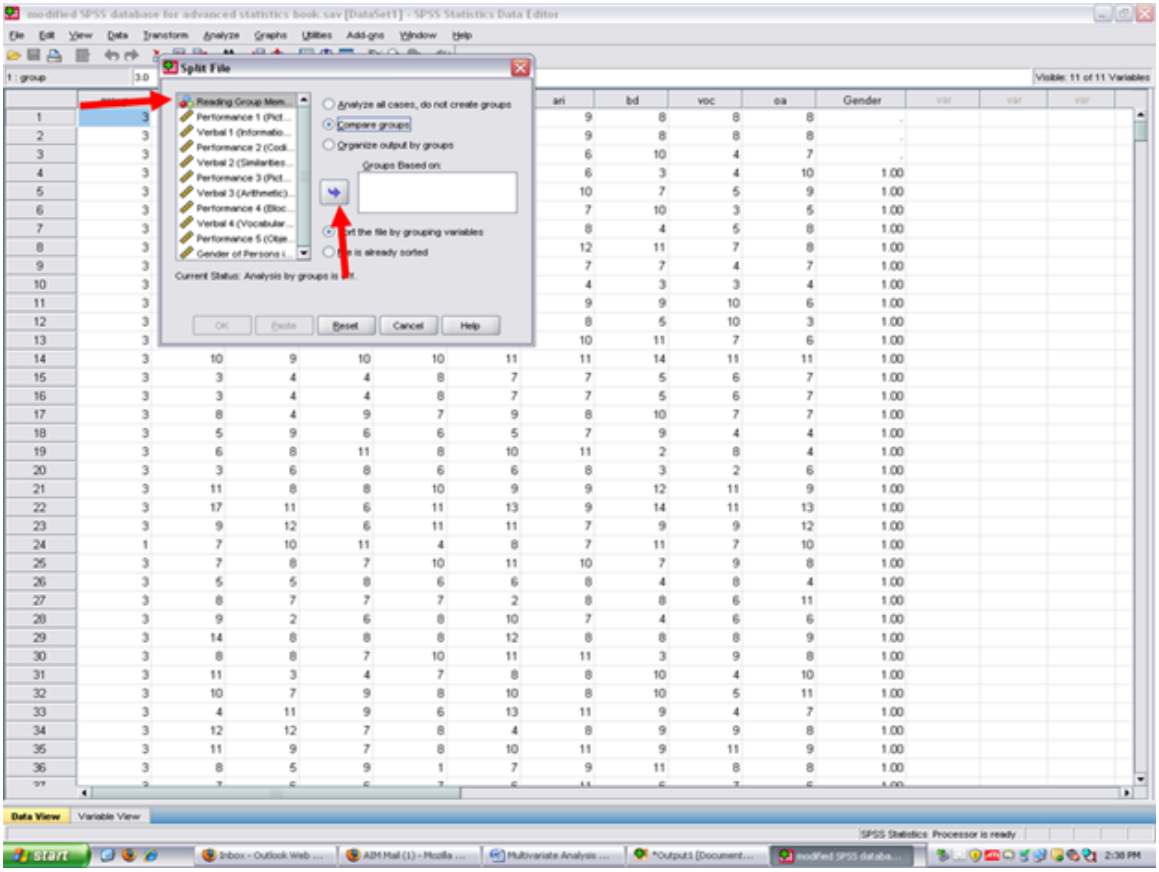
[15] http://cnx.org/content/m40728/latest/10.9.png/image



16

The screen below will appear after the Compare groups button has been clicked. Note that the Groups Based on rectangle has now become active. The independent (grouping) variable should be highlighted, as it already is, and then moved to the Groups Based on cell. After highlighting the Reading Group Membership variable, then click on the arrow below.

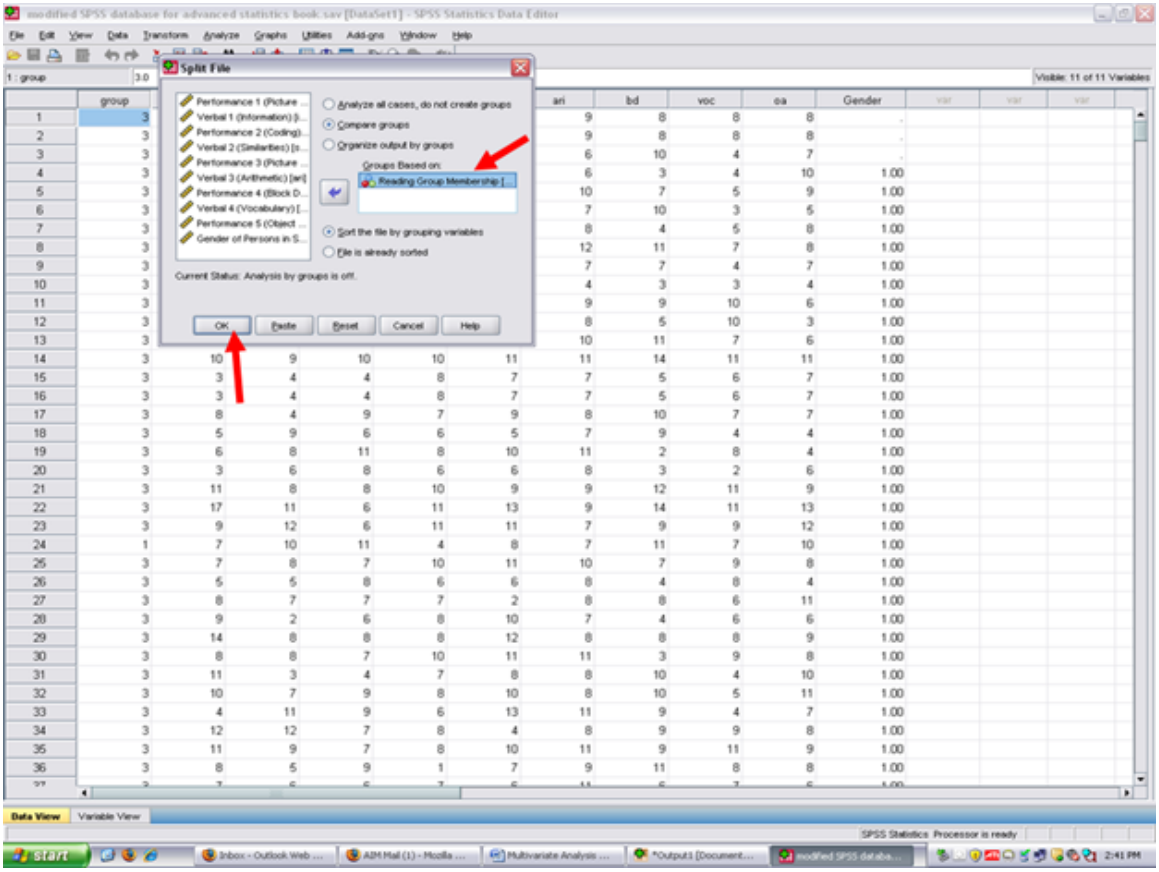
[16] http://cnx.org/content/m40728/latest/10.10.png/image



17

After clicking on the arrow, your independent variable of Reading Group Membership is now in the Groups Based on cell. Now click on OK.

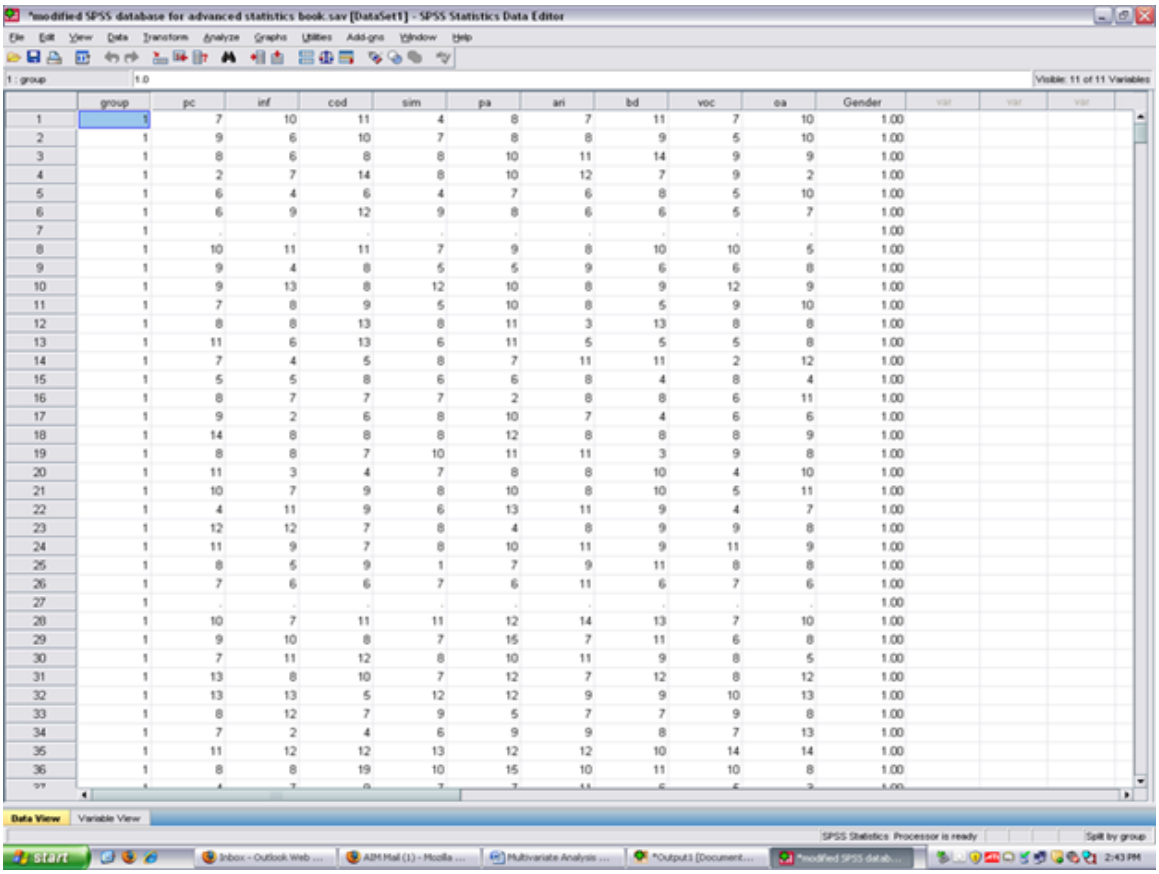
[17] http://cnx.org/content/m40728/latest/10.11.png/image



[18]

Now all SPSS analyses will be calculated separately for each of the three reading groups.

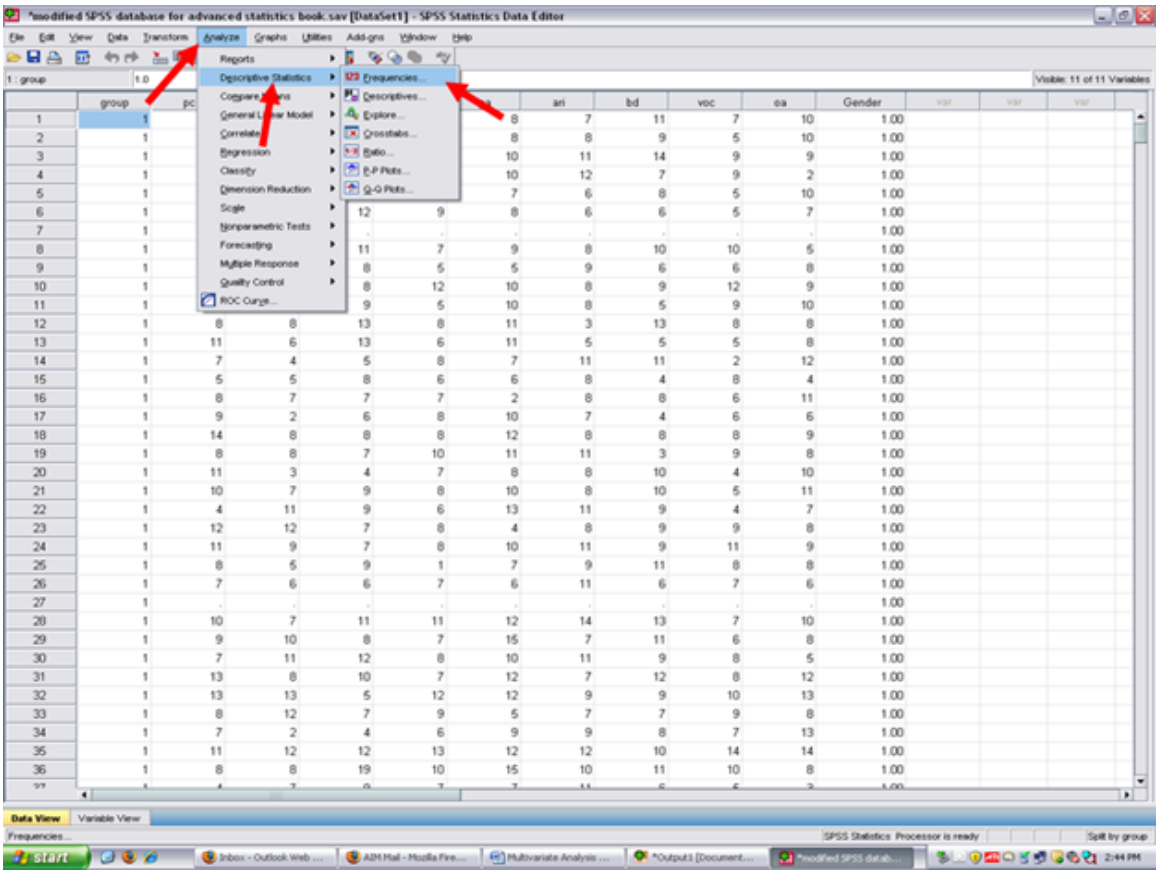
[18] http://cnx.org/content/m40728/latest/10.12.png/image



[19]

√ **Analyze**
*** Descriptive Statistics**
*** Frequencies**

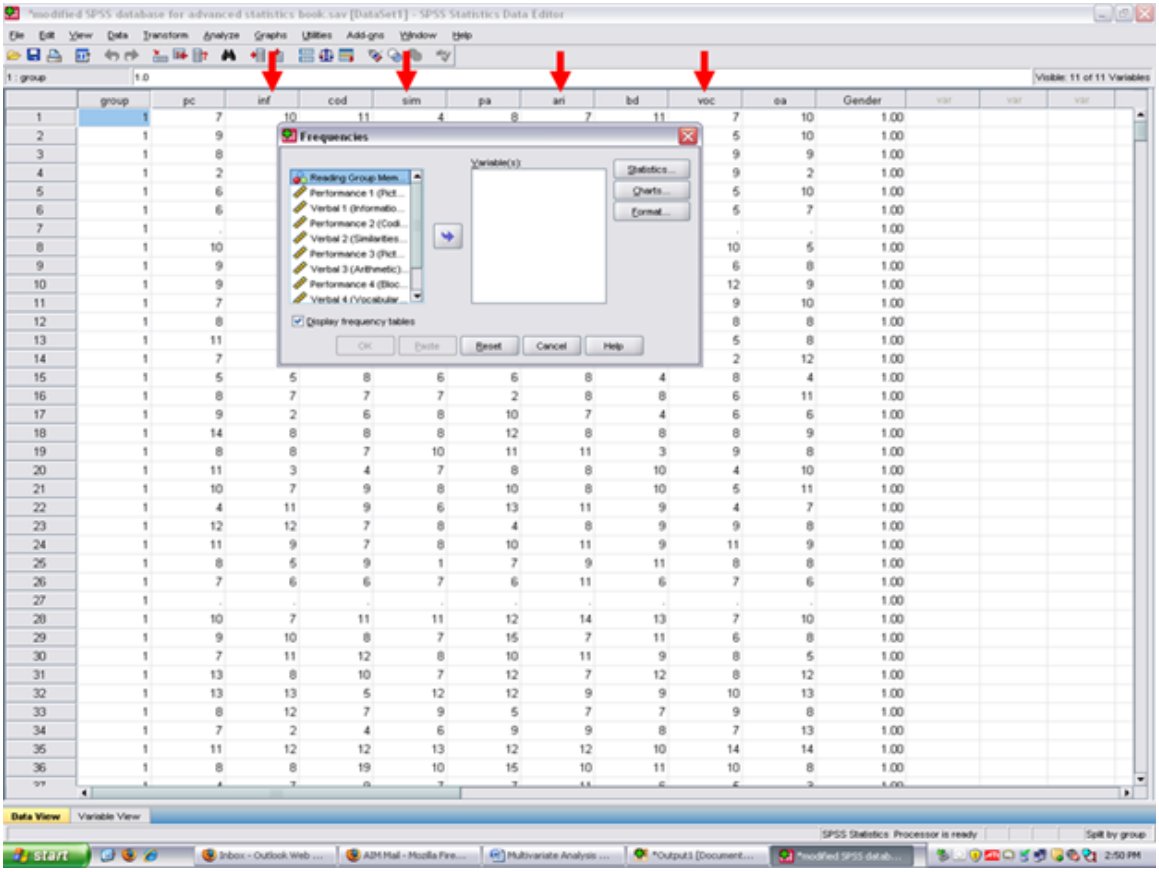
[19] http://cnx.org/content/m40728/latest/10.13.png/image



[20]

After clicking on frequencies, the following screen should now be present. Remember that the dependent variables in this example are Verbal 1 through Verbal 4 (i.e., inf, ari, voc)

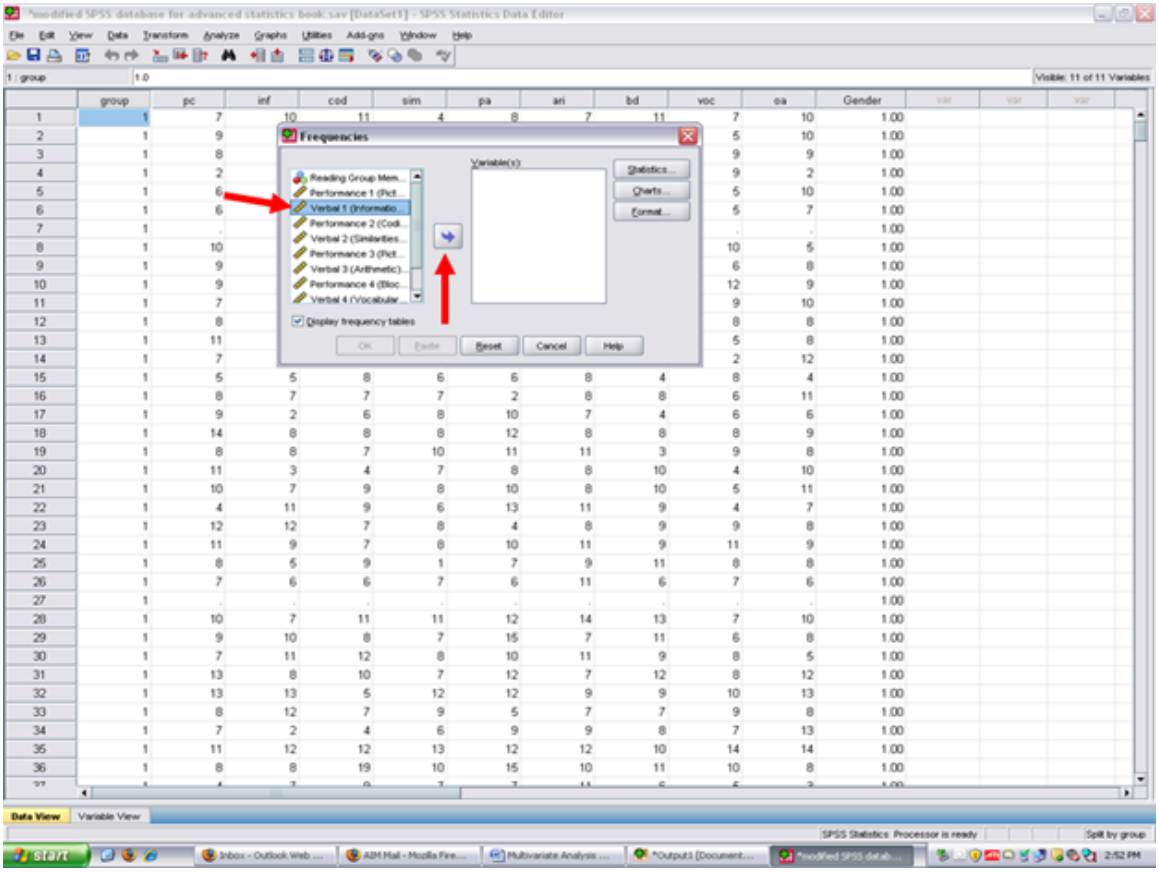
[20] http://cnx.org/content/m40728/latest/10.14.png/image



[21]

√ **Move over the dependent variables. In this example, highlight each of the dependent variables one at a time and click on the right arrow.**

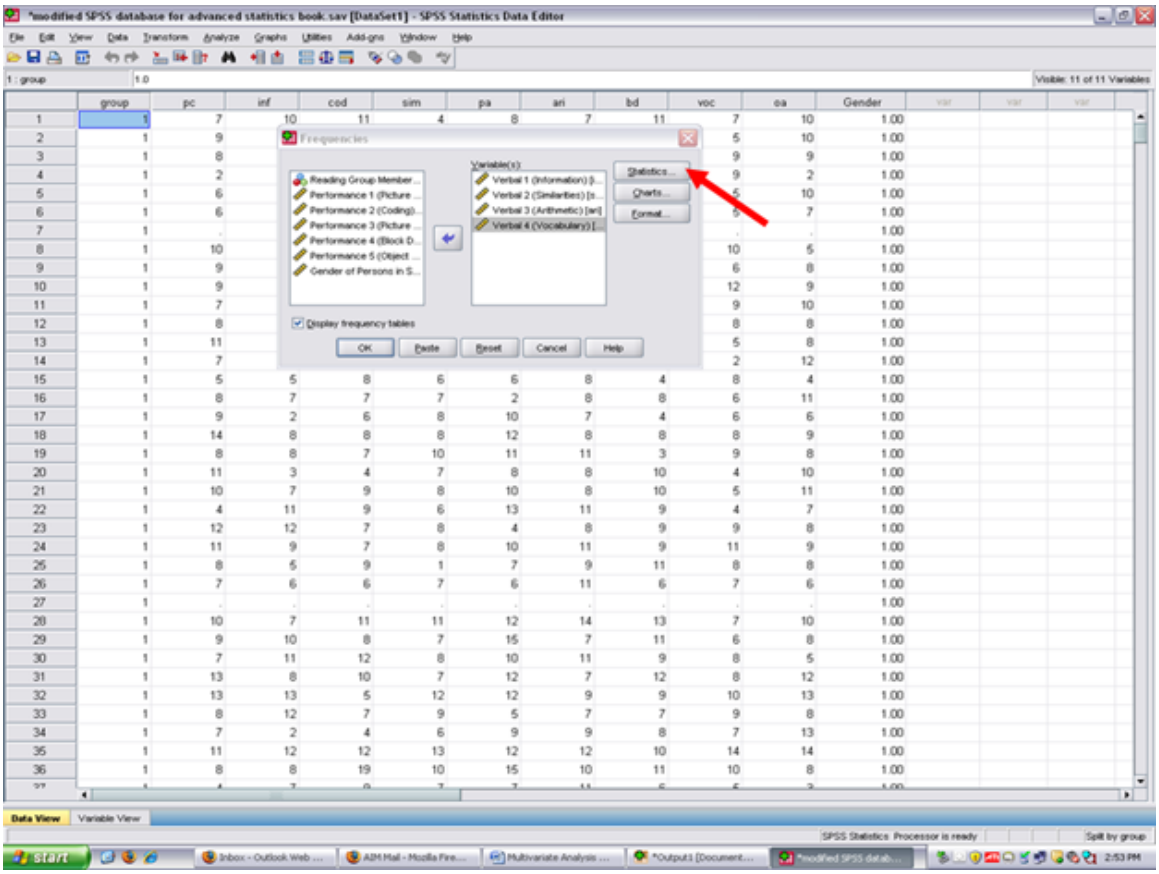
[21] http://cnx.org/content/m40728/latest/10.15.png/image



[22]

When all four dependent variables are in the Variable cell, then click on Statistics.

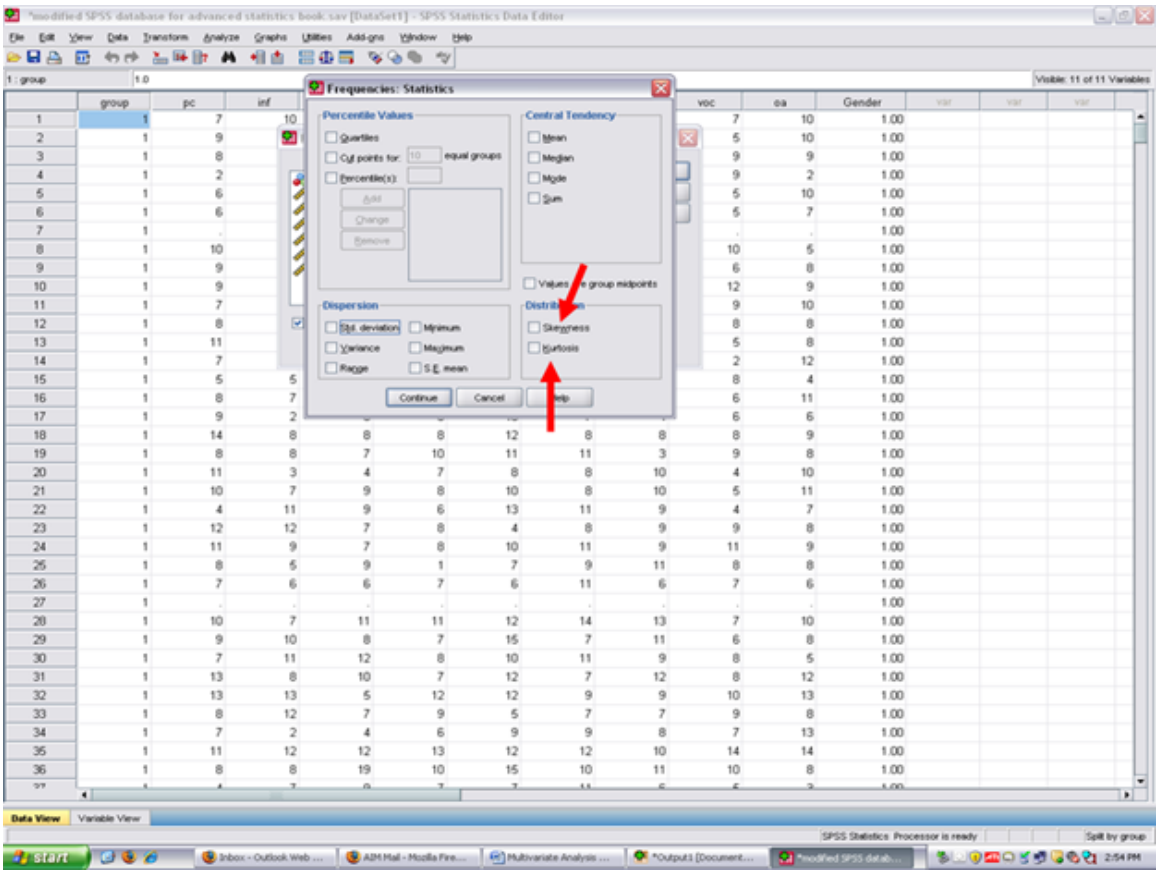
[22] http://cnx.org/content/m40728/latest/10.16.png/image



[23]

After clicking on statistics, the screen below will appear. Note that no statistics items are checked. For purposes of this example, only the skewness and kurtosis items will be checked. Although readers may obtain descriptive statistics at this screen, the MANOVA procedure itself can be used to obtain that information.

[23] http://cnx.org/content/m40728/latest/10.17.png/image



[24]

Though previously discussed in the steps and screenshots chapters for the basic statistical procedures, readers may find the following information helpful in understanding skewness and kurtosis and their importance for conducting statistical procedures.

* Skewness [Note. Skewness refers to the extent to which the data are normally distributed around the mean. Skewed data involve having either mostly high scores with a few low ones or having mostly low scores with a few high ones.] Readers are referred to the following sources for a more detailed definition of skewness: http://www.statistics.com/index.php?page=glossary&term_id=356[25] and http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[26]

To standardize the skewness value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the skewness value from the SPSS output and divide it by the Std. error of skewness. If the resulting calculation is within -3 to +3, then the skewness of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

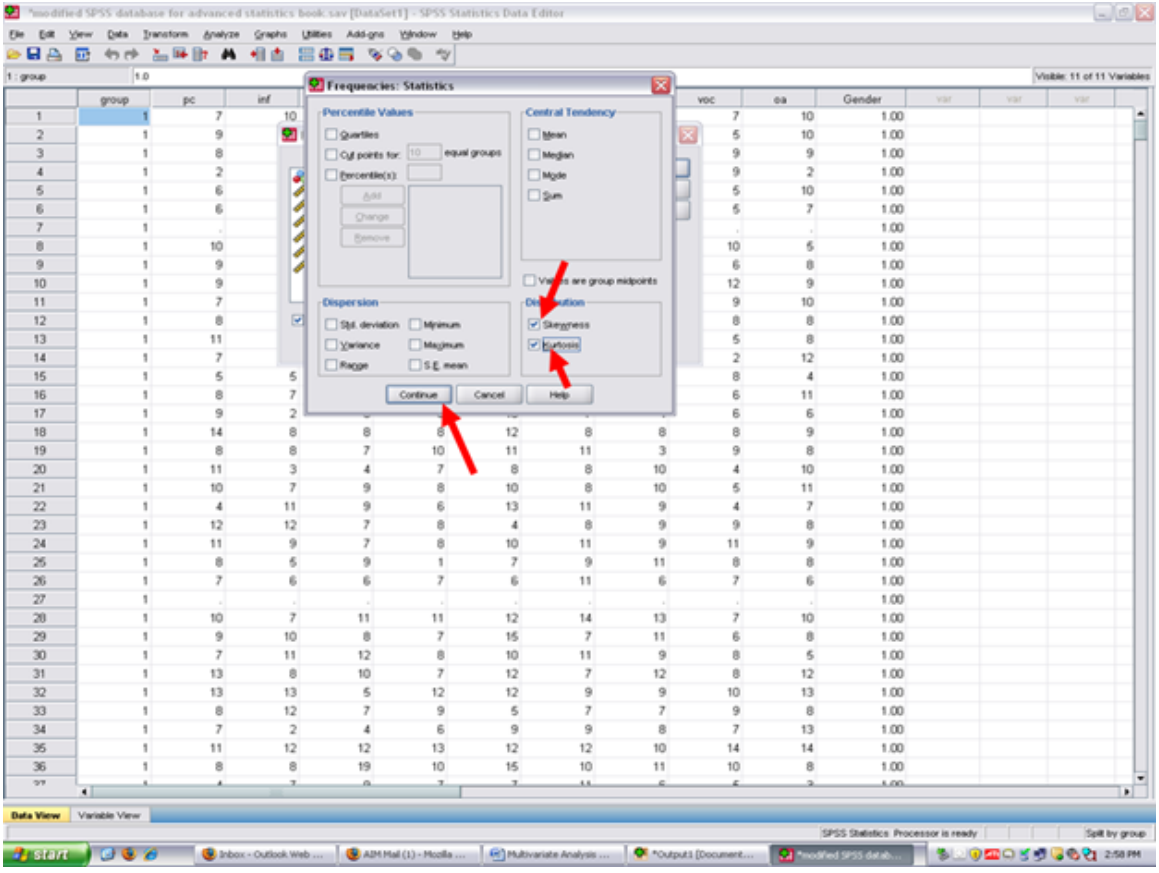
* Kurtosis [Note. Kurtosis also refers to the extent to which the data are normally distributed around the mean. This time, the data are piled up higher than normal around the mean or piled up higher than normal at the ends of the distribution.] Readers are referred to the following sources for a more detailed definition of kurtosis: http://www.statistics.com/index.php?page=glossary&term_id=326[27] and

[24] http://cnx.org/content/m40728/latest/10.18.png/image
[25] http://www.statistics.com/index.php?page=glossary&term_id=356
[26] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb
[27] http://www.statistics.com/index.php?page=glossary&term_id=326

http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[28]

To standardize the kurtosis value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the kurtosis value from the SPSS output and divide it by the Std. error of kurtosis. If the resulting calculation is within -3 to +3, then the kurtosis of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed. Then the kurtosis of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

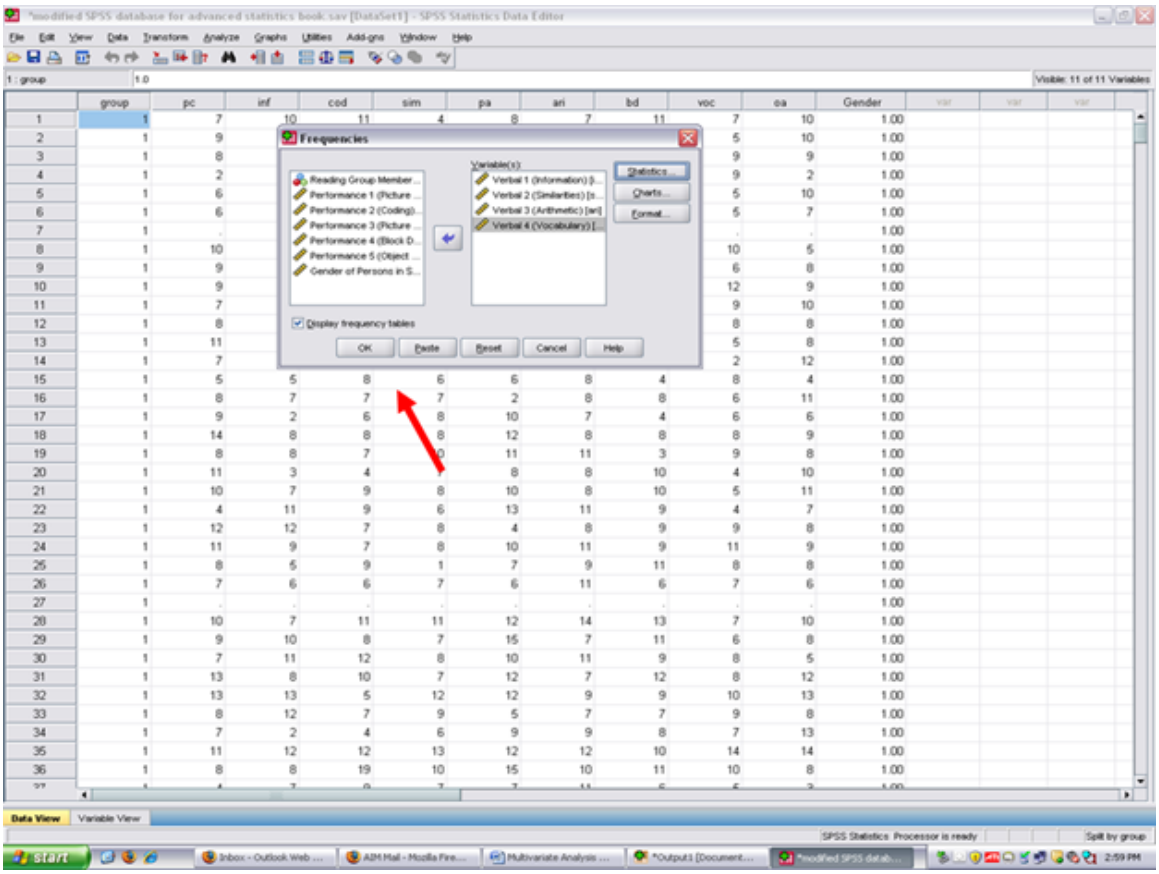
Click on Skewness, Kurtosis, and Continue



29

Now click on OK.

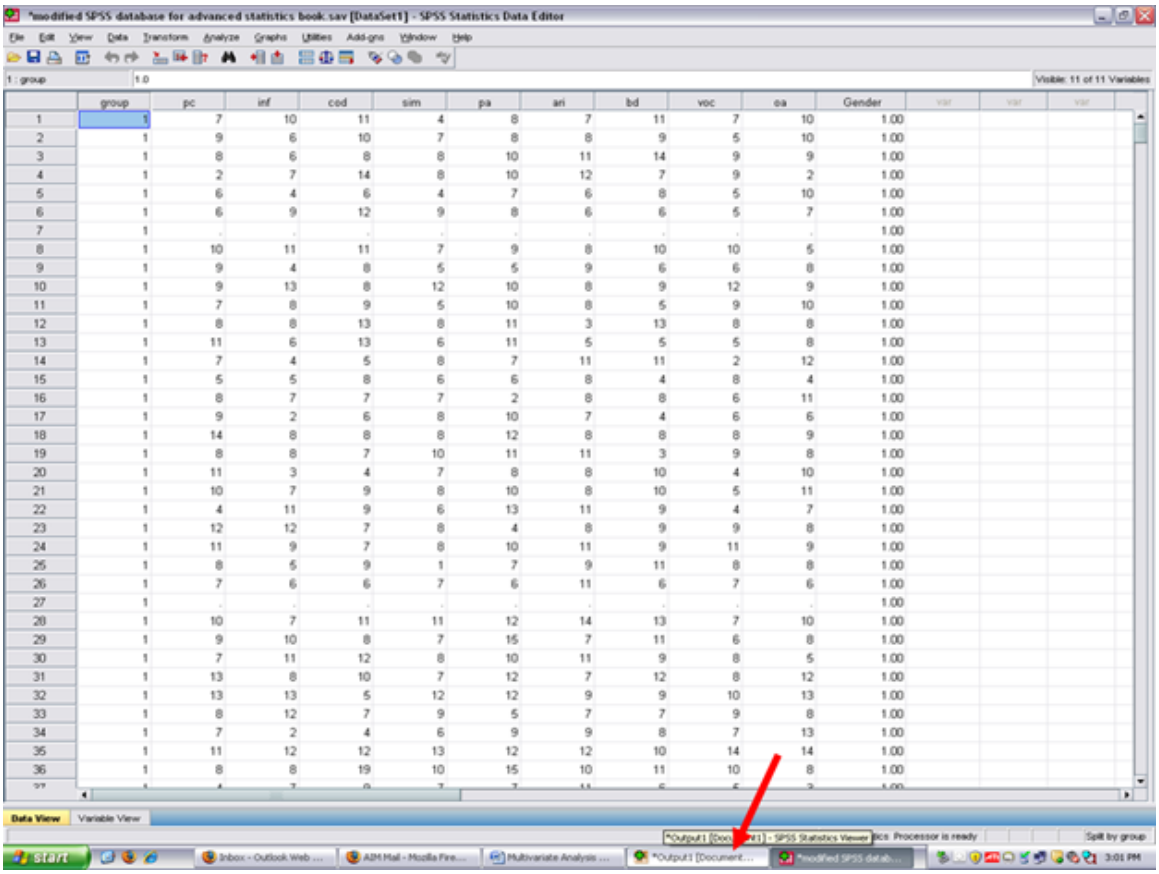
[28] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb
[29] http://cnx.org/content/m40728/latest/10.19.png/image



[30]

If SPSS does not send you to the Output screen, click on the Output button at the bottom of your screen so that you can view the results of the statistics you just had SPSS calculate for you.

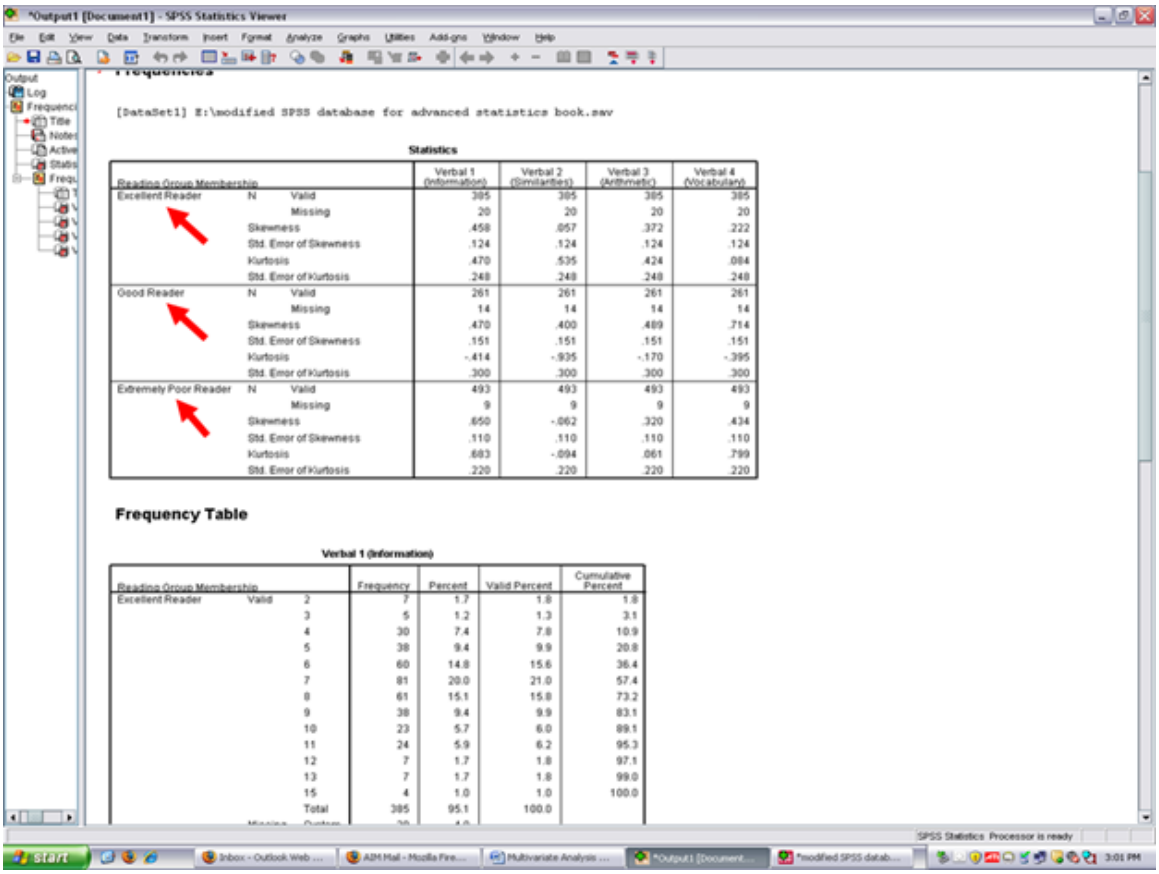
[30] http://cnx.org/content/m40728/latest/10.20.png/image



[31]

Clicking on the Output icon will result in the screen below being present. You will note that the Statistics table below contains statistics for the Excellent Reader group; for the Good Reader group; and for the Extremely Poor Reader group.

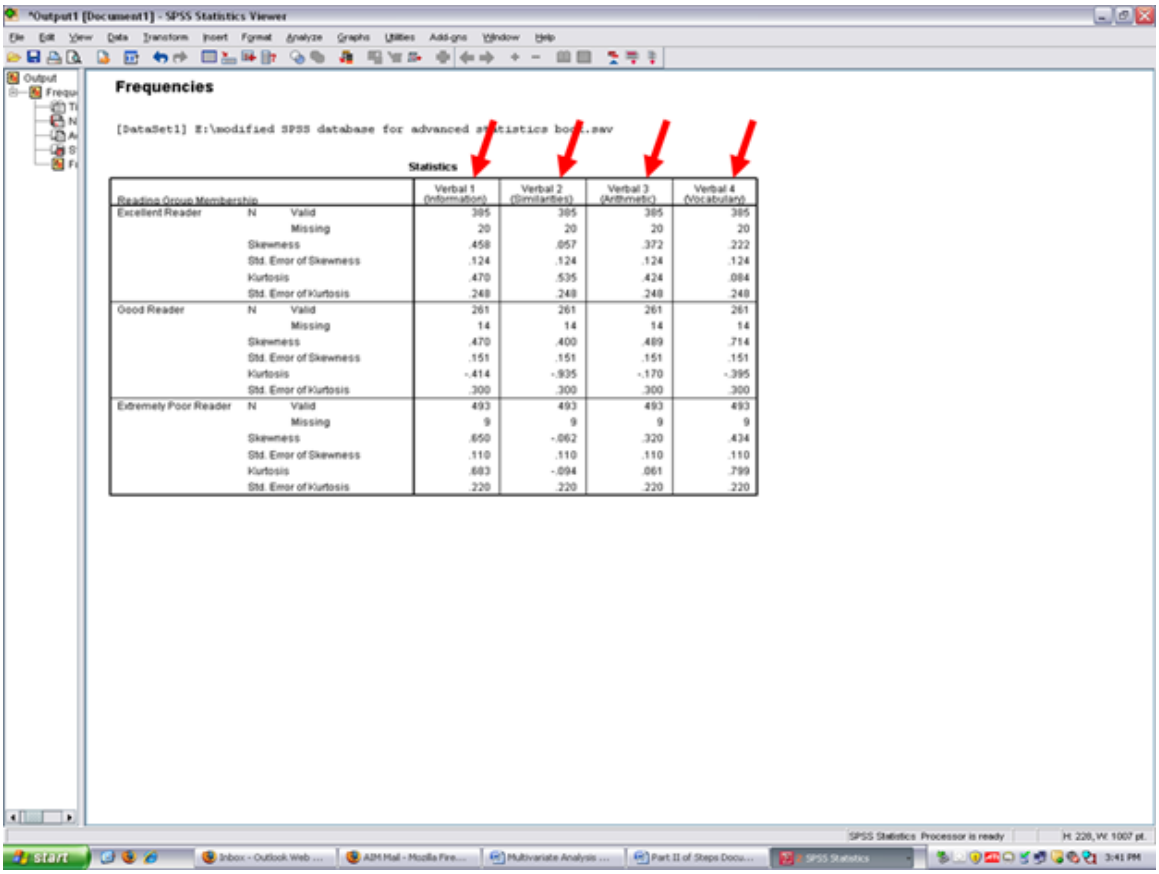
[31] http://cnx.org/content/m40728/latest/10.21.png/image



[32]

In the columns to the right of the three reading groups are the skewness and kurtosis values for the four dependent variables.

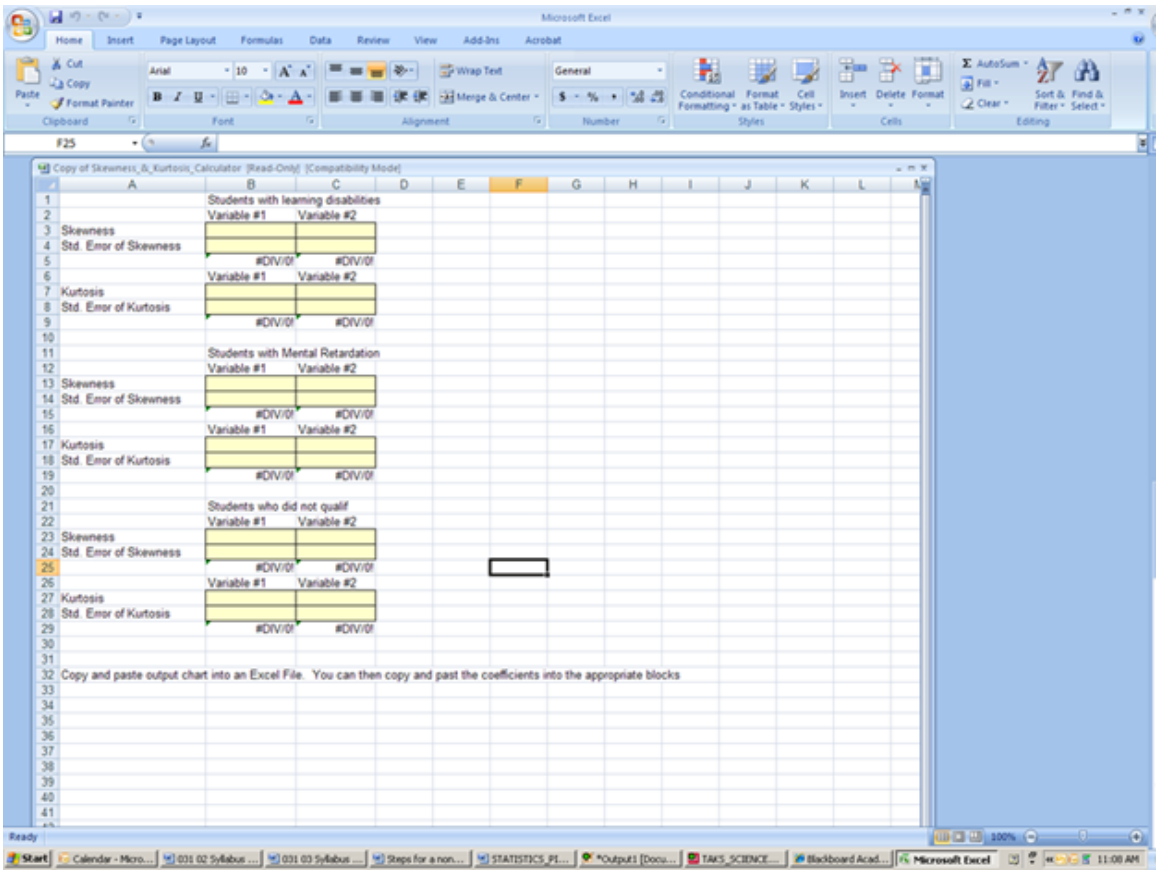
[32] http://cnx.org/content/m40728/latest/10.22.png/image



[33]

Using the skewness and kurtosis value for each dependent variable above for each of the three groups above, type them in one at a time into the standardized coefficients calculator. To be regarded as being normally distributed, the coefficient should be within -3 to +3 (Onwuegbuzie & Daniel, 2002).

[33] http://cnx.org/content/m40728/latest/10.23.png/image



[34]

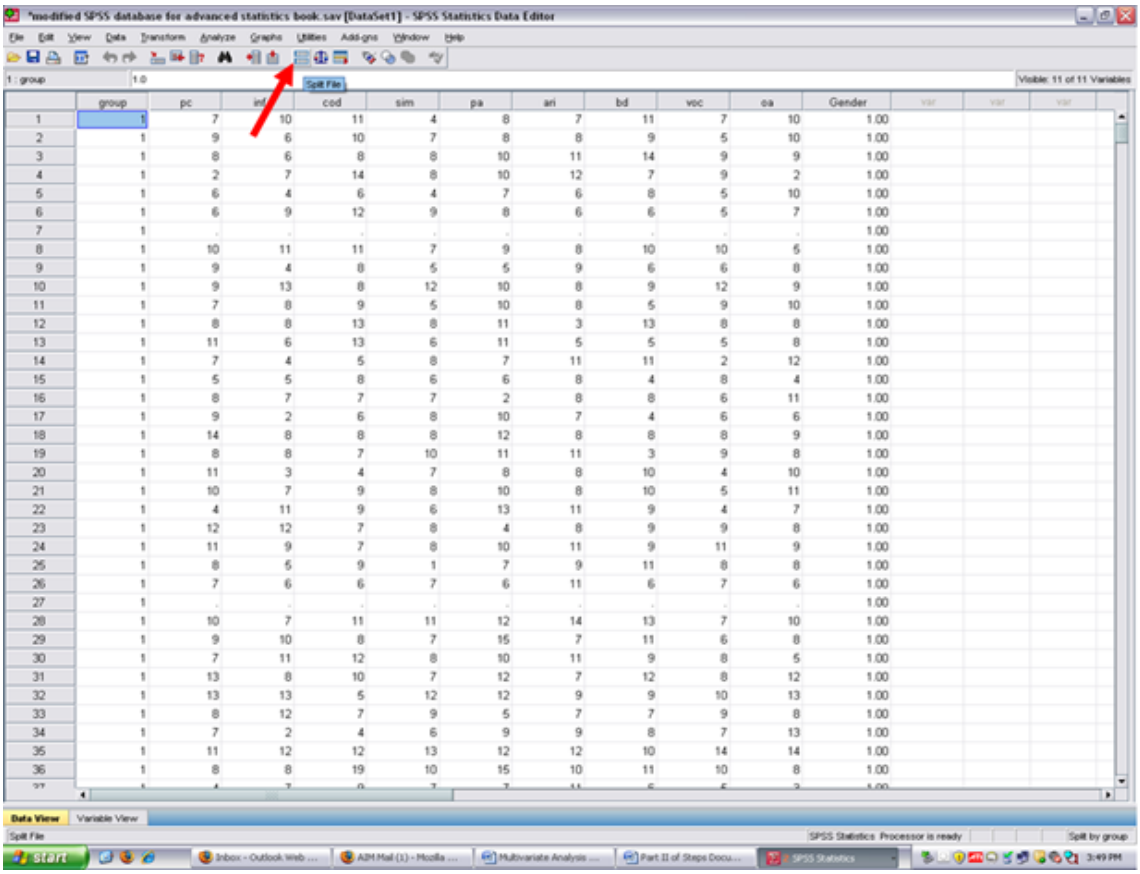
After calculating the standardized coefficients, record the number that are within +/-3 and the number that are outside of these boundaries. This information needs to be discussed in your results section where you document your checks of the assumptions underlying this particular statistical technique.

Now, before conducting the MANOVA to answer our research question, this dataset must be put back together. Remember that the dataset is currently split. As you were in the SPSS Output screen, make sure that you go back to the Data screen before continuing.

To unsplit the file,

√ Split Files (the icon next to the scales)

[34] http://cnx.org/content/m40728/latest/10.24.png/image



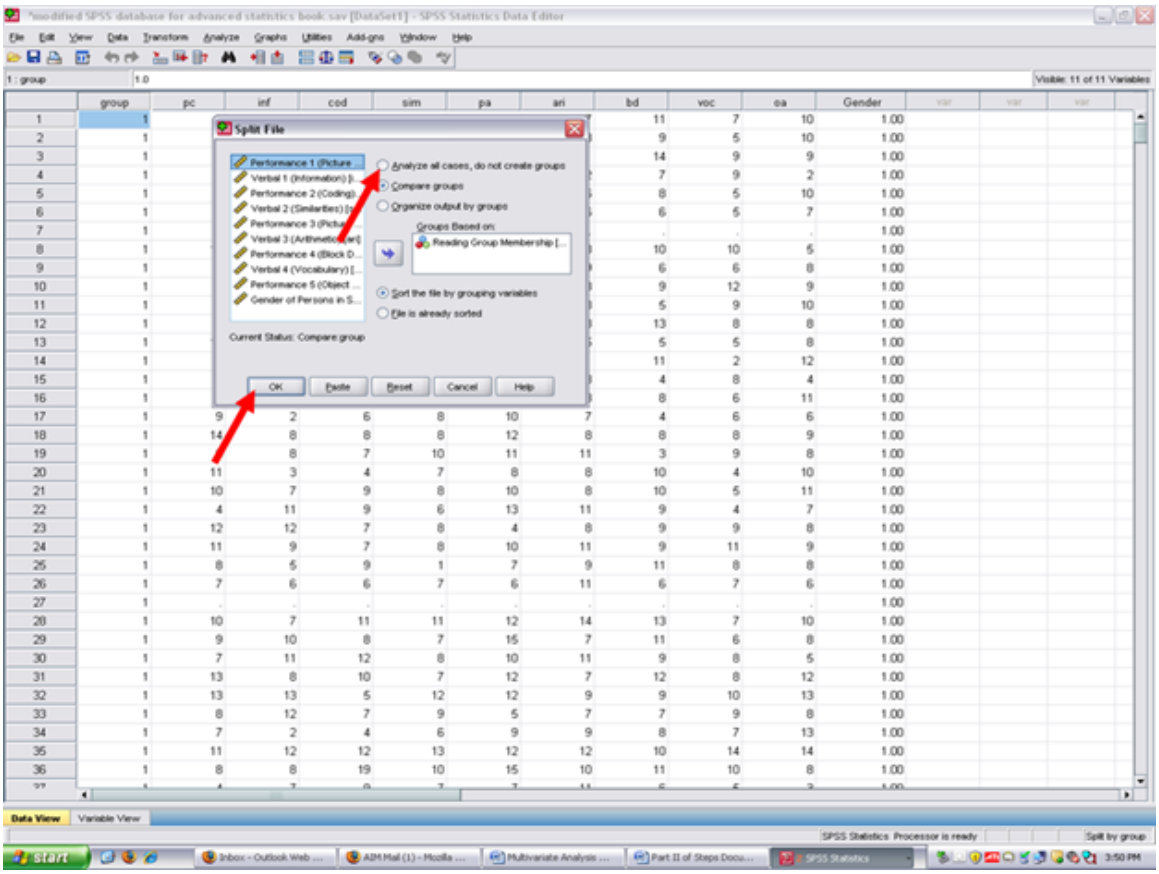
[35]

Clicking on split files will reveal this screen. Although two ways exist in which to have SPSS analyze all cases, the easiest is simply to click on:

√ Analyze all cases, do not create groups

Then, √ OK

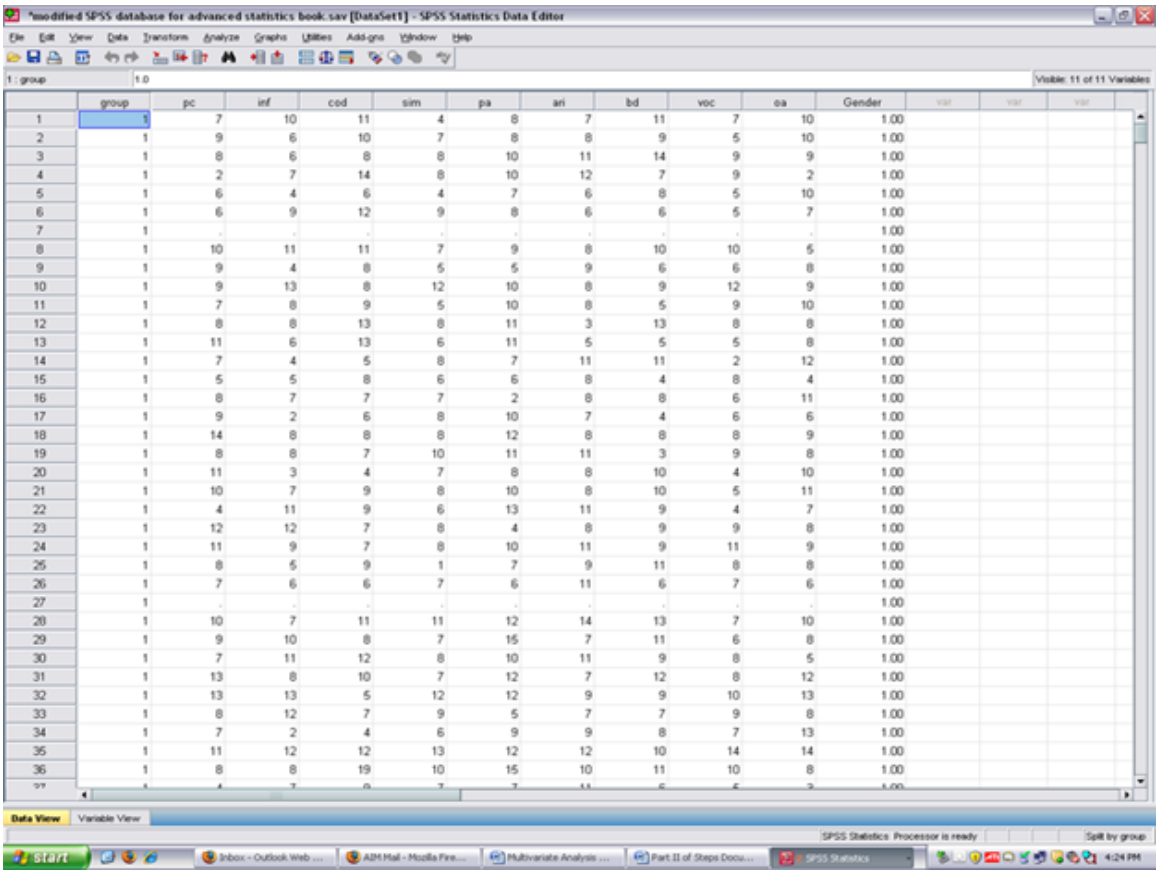
[35] http://cnx.org/content/m40728/latest/10.25.png/image



[36]

After clicking on OK, your screen should now look like the one below.

[36] http://cnx.org/content/m40728/latest/10.26.png/image



[37]

[37] http://cnx.org/content/m40728/latest/10.27.png/image

Chapter 2

Multivariate Analysis of Variance: Part II[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

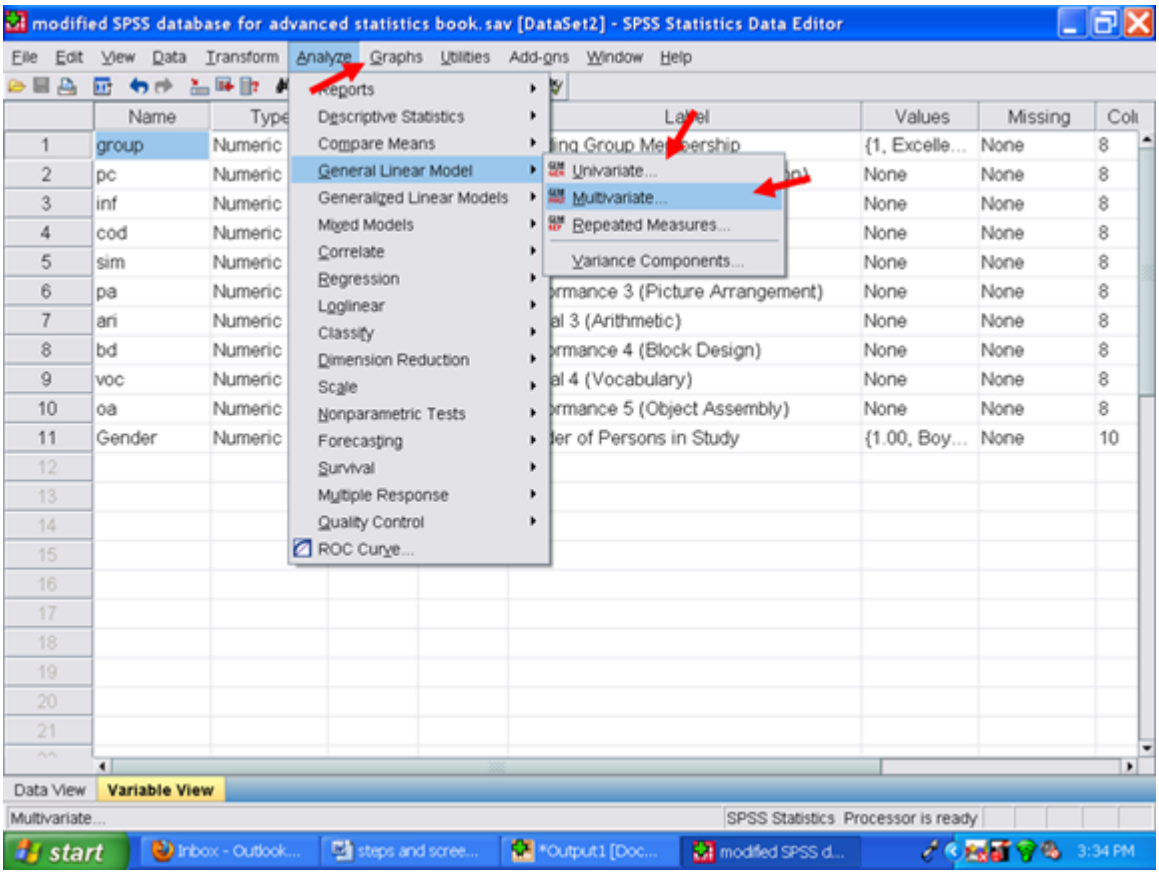
John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

We are now ready to conduct the MANOVA procedure.

√ **Analyze**
√ **General Linear Model**
√ **Multivariate**

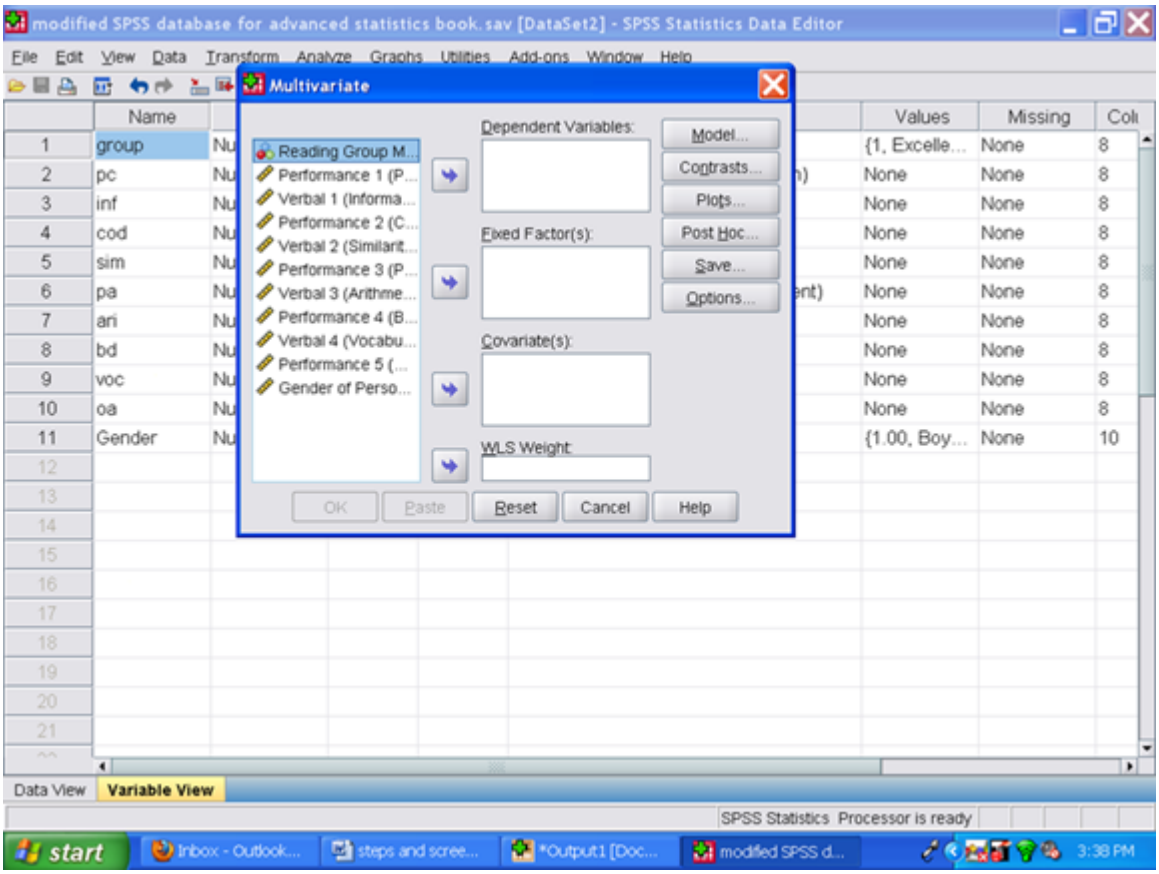
[1] This content is available online at <http://cnx.org/content/m40729/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40729/latest/www.writingandstatisticalhelp



[4]

After clicking on Multivariate, the following screen will appear. You will send your dependent variables to the Dependent Variables box and your independent variable to the Fixed Factor box. Remember that in this example, Verbal 1 through Verbal 4 (i.e., Information, Similarities, Arithmetic, and Vocabulary) constitute the dependent variables and Reading Group Membership is the independent variable.

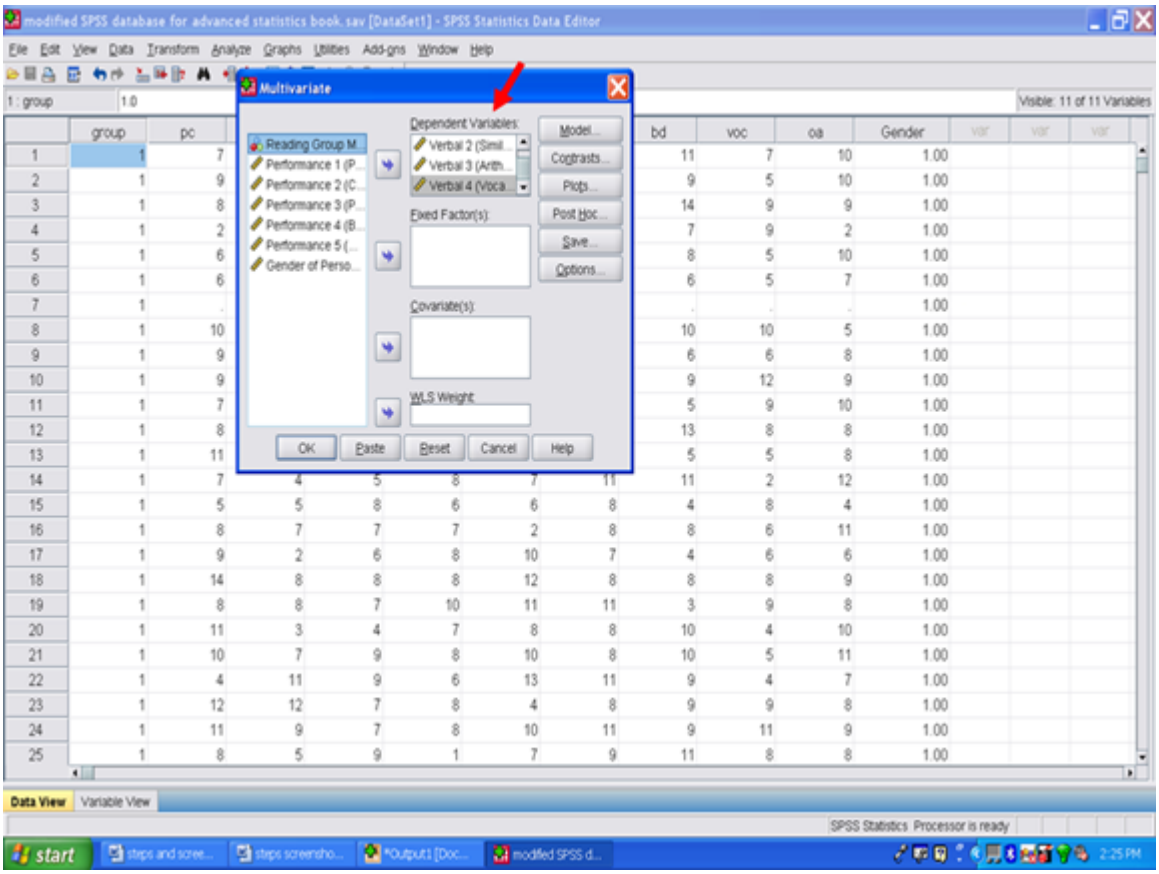
[4] http://cnx.org/content/m40729/latest/11.1.png/image



[5]

Now send the four dependent variables (i.e., Verbal 1 through Verbal 4) over one at a time to the Dependent Variables screen.

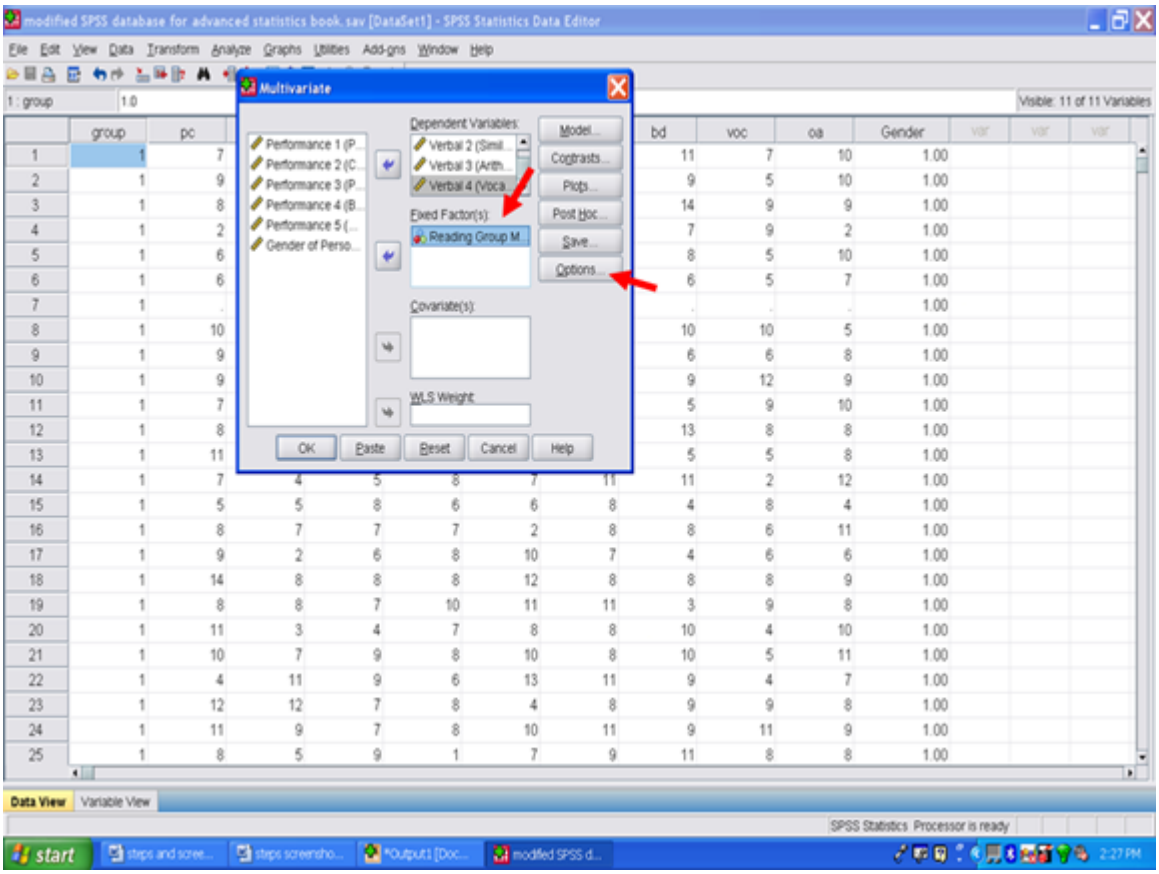
[5] http://cnx.org/content/m40729/latest/11.2.png/image



[6]

Then send over the independent variable, Reading Group Membership, to the Fixed Factor box. Then click on Options.

[6] http://cnx.org/content/m40729/latest/113.png/image



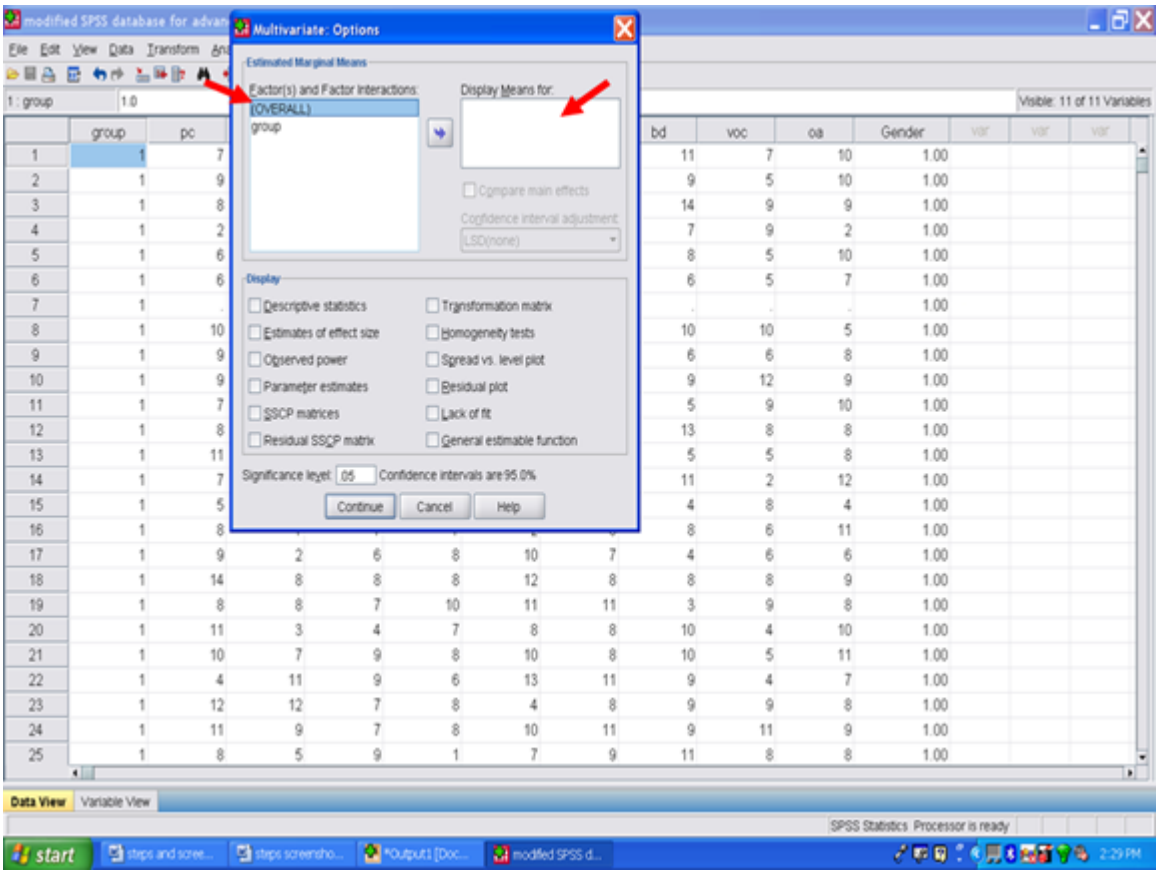
[7]

Clicking on options will then give you this screen.

We will use this screen to obtain descriptive statistics of our four dependent variables for each of our three reading groups; to obtain effect size estimates; and to determine the extent to which the assumptions underlying use of the MANOVA procedure are met.

Click on (Overall) and on group and send them to the Display Means for: box.

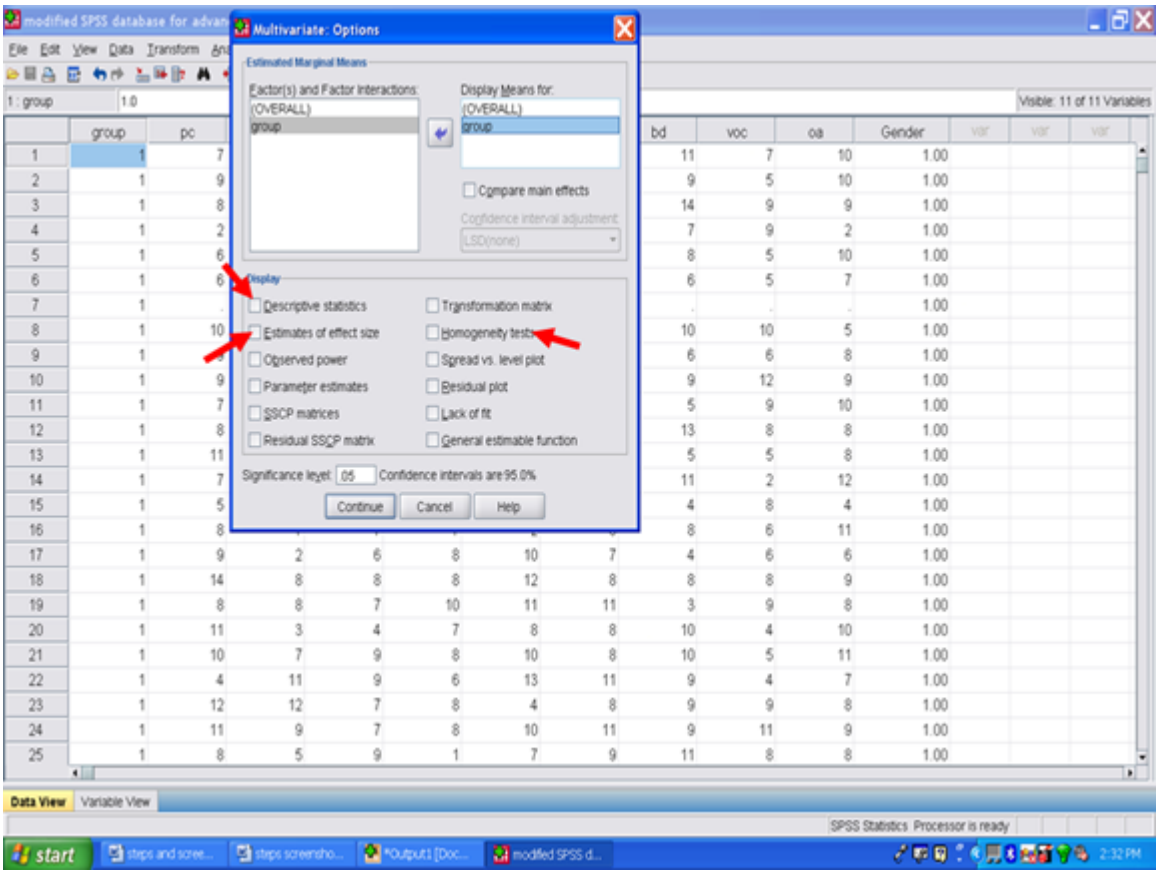
[7] http://cnx.org/content/m40729/latest/11.4.png/image



[8]

To obtain the information just mentioned, you will need to click on:
Descriptive Statistics
Estimates of Effect Size
Homogeneity tests

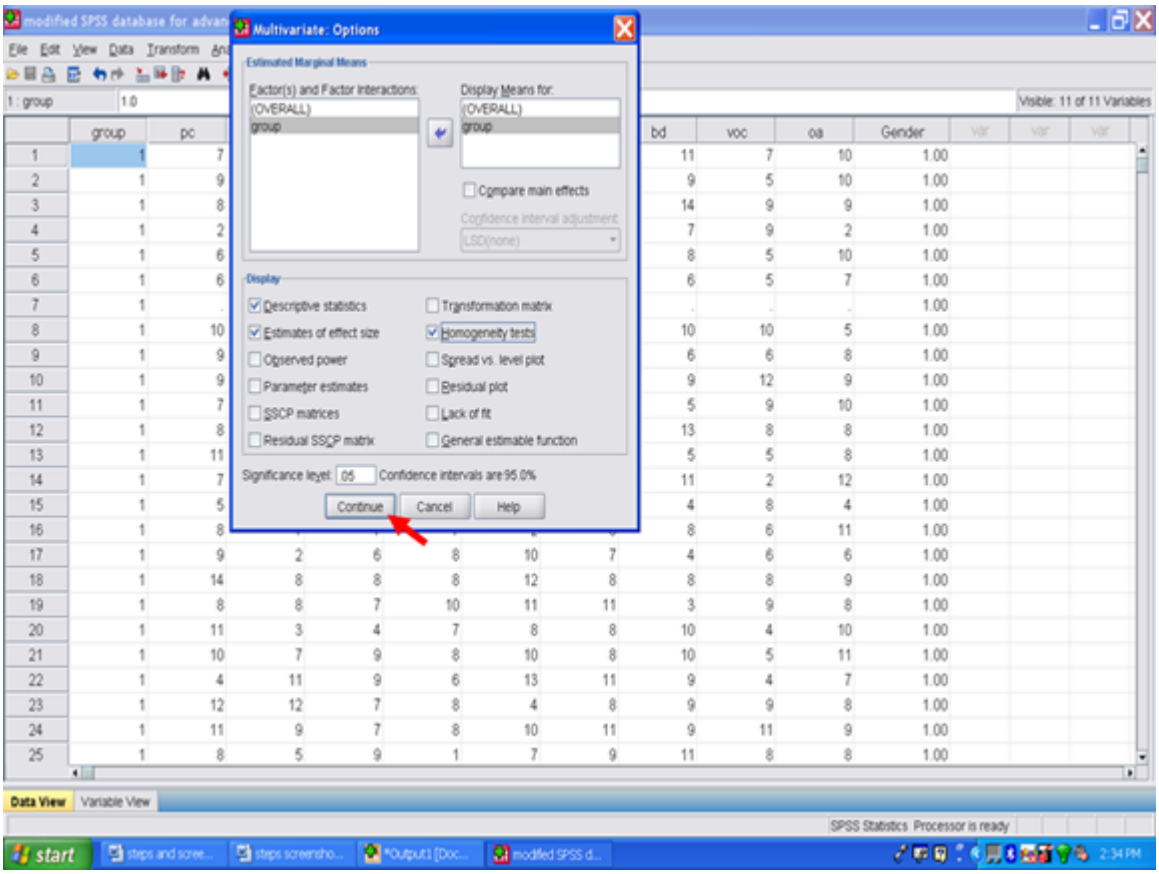
[8] http://cnx.org/content/m40729/latest/11.5.png/image



[9]

After checking the three boxes mentioned, then click on Continue

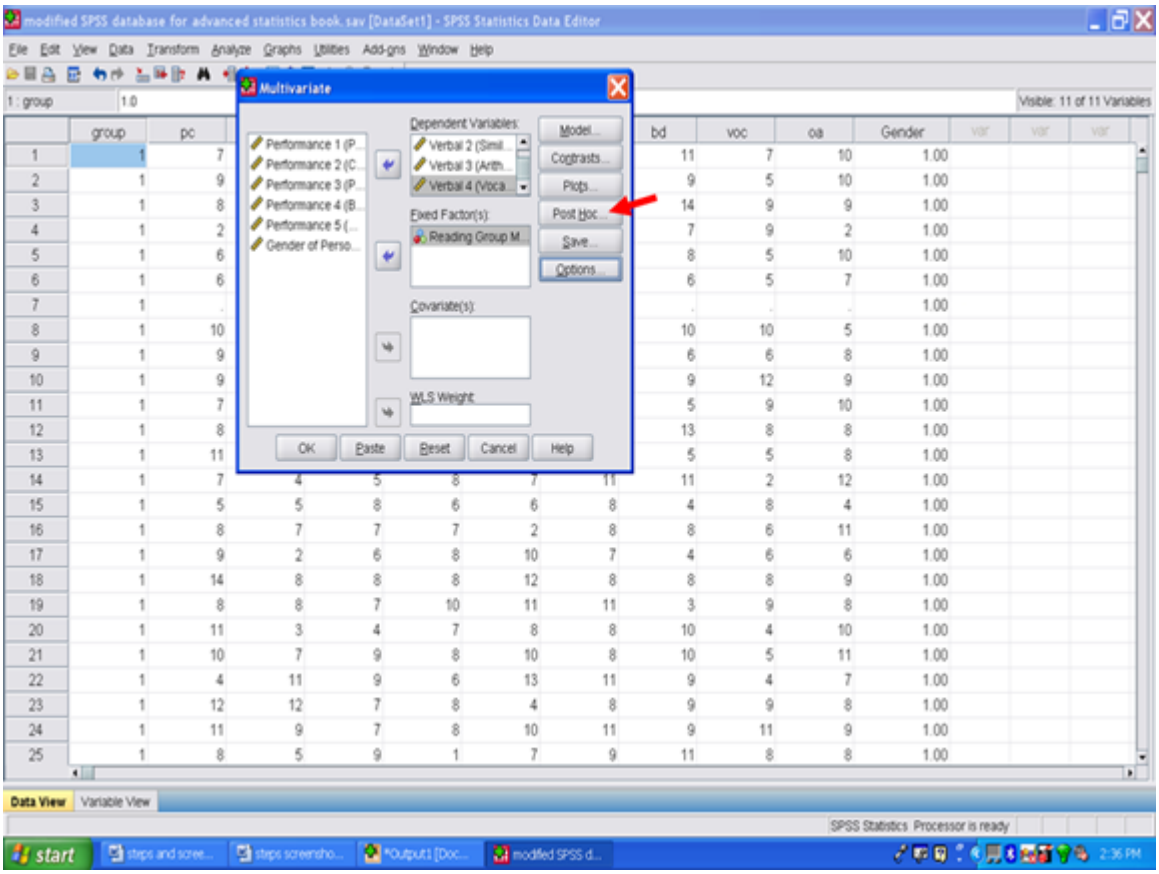
[9] http://cnx.org/content/m40729/latest/11.6.png/image



[10]

After clicking on Continue, the following screen will appear. Now click on Post Hoc so that pairwise analyses can be conducted.

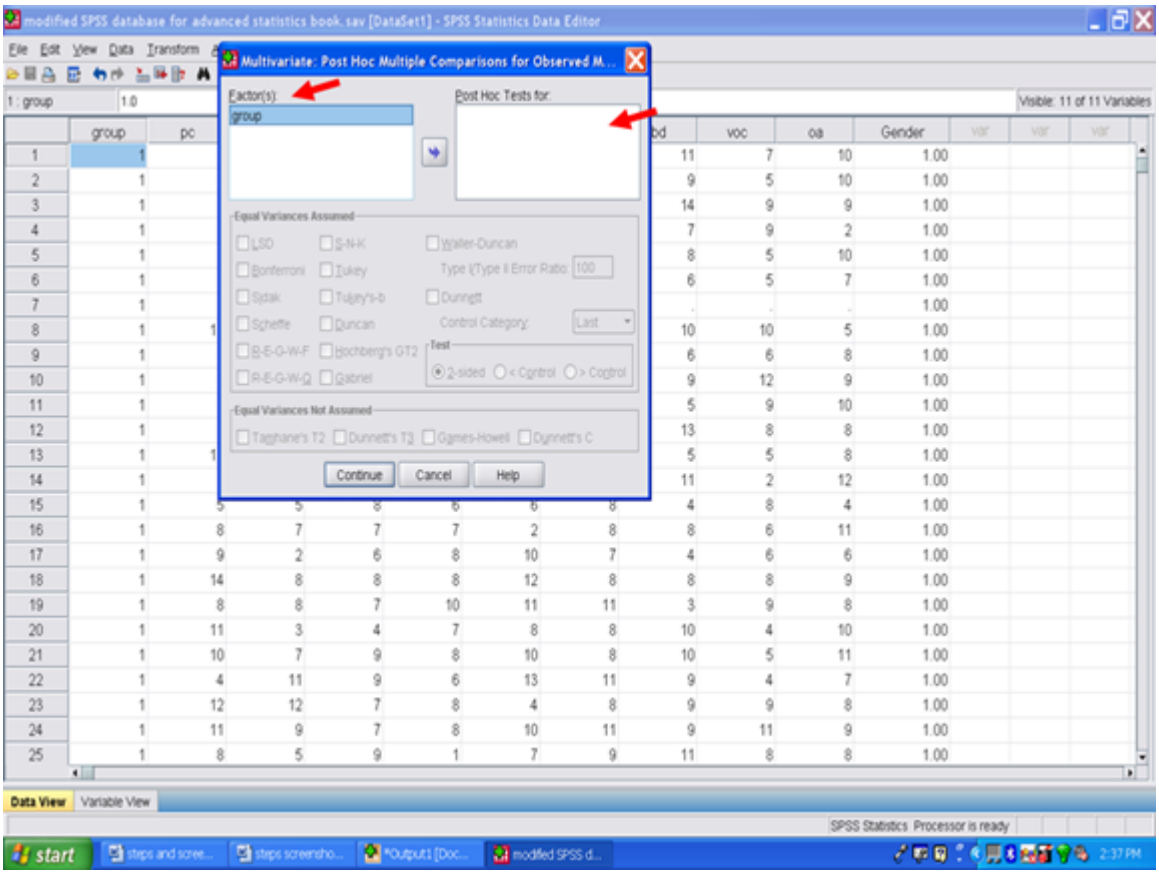
[10] http://cnx.org/content/m40729/latest/11.7.png/image



[11]

Clicking on Post Hoc will then give you the screen below. Click on group and send it to the box labeled Post Hoc Tests for:

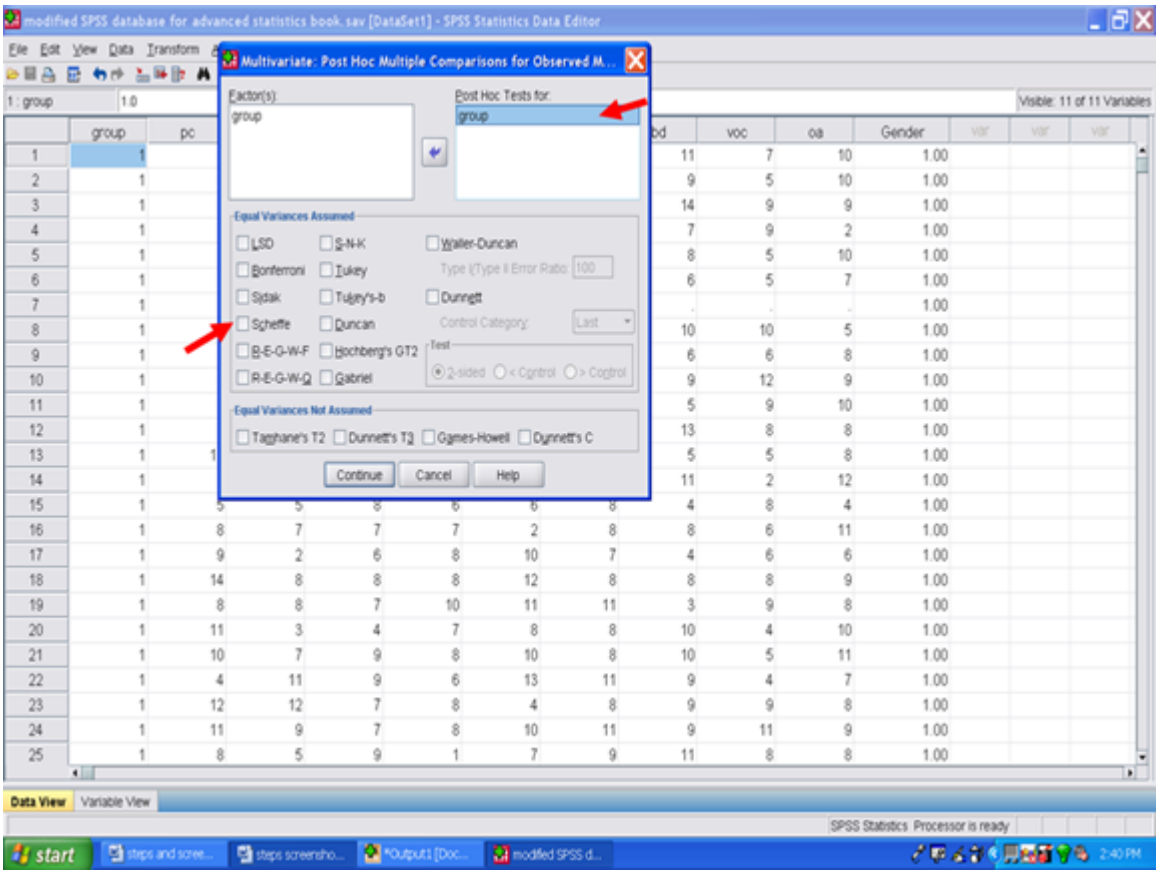
[11] http://cnx.org/content/m40729/latest/11.8.png/image



12

Once you send the independent variable over to the box labeled Post Hoc Tests for:, the different types of post hoc procedures become activated. For our purposes, we will click on the Scheffe' post hoc procedure. If you are analyzing data for your dissertation or theses, you might check with your chair to see if s/he has a preference for a different type. Scheffe'is a conservative post hoc procedure, thus the reason that we recommend its use. Note that the way that Scheffe is spelled in SPSS is not written correctly in compliance with APA. The way that we have typed it is correct.

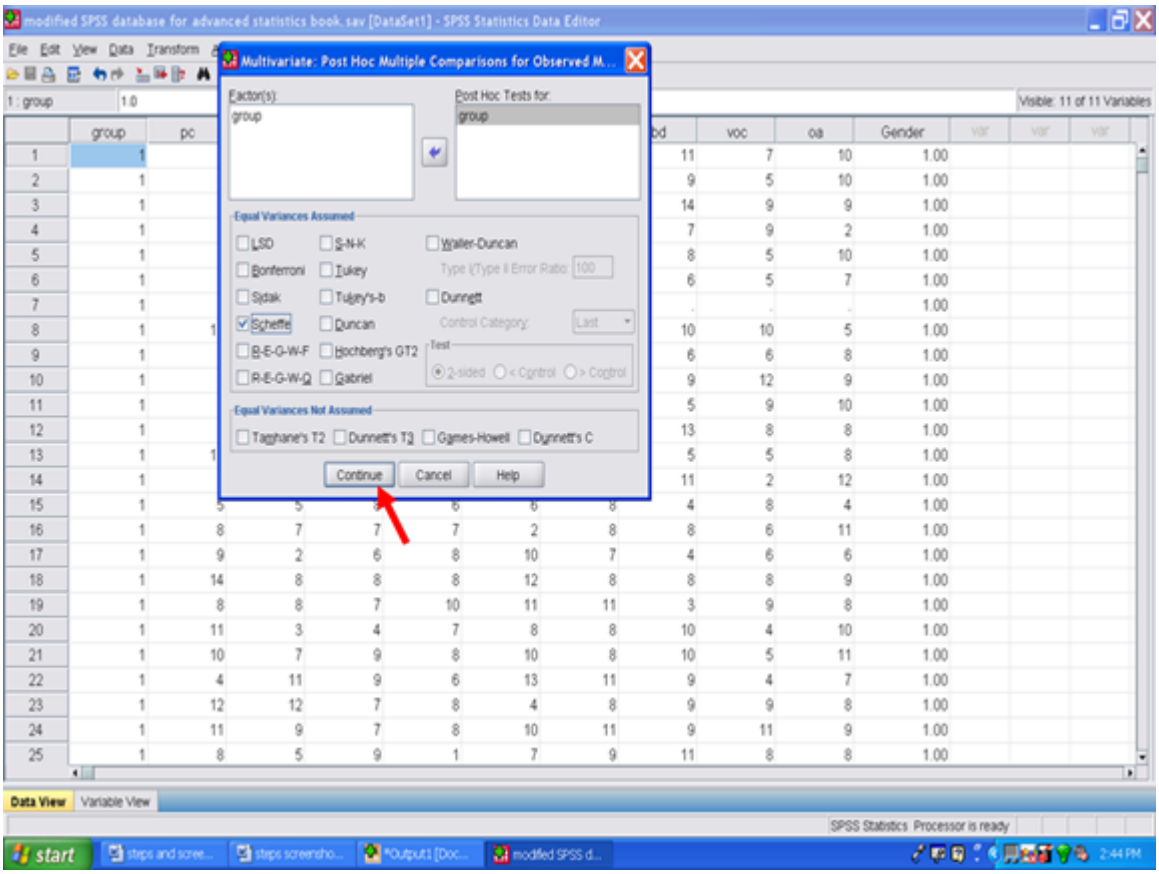
[12] http://cnx.org/content/m40729/latest/11.9.png/image



13

After clicking on the Scheffe' post hoc procedure, then click on Continue.

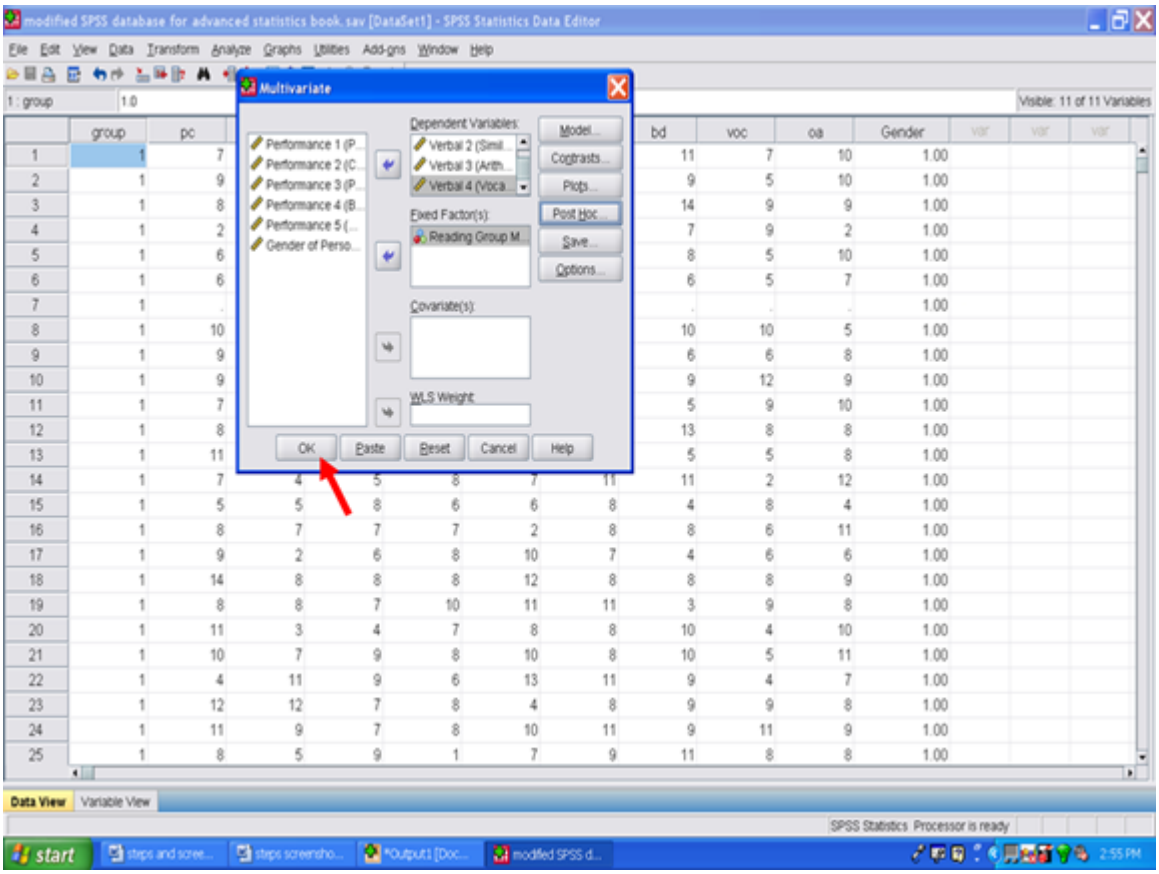
[13] http://cnx.org/content/m40729/latest/11.10.png/image



14

Then click on OK.

[14] http://cnx.org/content/m40729/latest/11.11.png/image



[15]

Your MANOVA has now been calculated. If you are not sent to the SPSS Output file, click on Output at the bottom of your screen so that you may view your MANOVA results.

[15] http://cnx.org/content/m40729/latest/11.12.png/image

Chapter 3

Multiple Analysis of Variance: Part III[1]

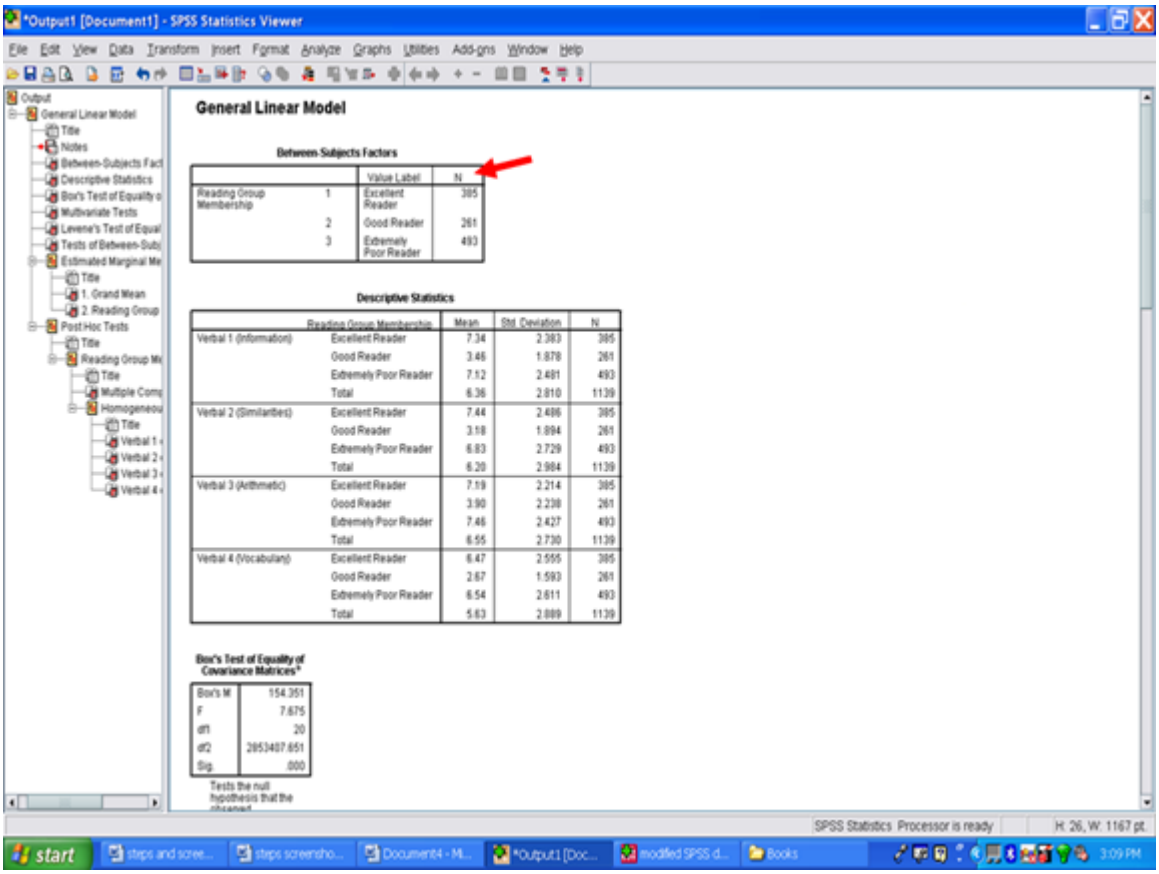
NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

- **John R. Slate** is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.
- **Ana Rojas-LeBouef** is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

Now that you are in the SPSS output screen, your screen should look like the following: Your first table is titled Between-Subjects Factors and it is entirely redundant with the table immediately below it. The column of n in the Between-Subjects Factors table is duplicated in the Descriptive Statistics table. Therefore, we will not use the Between-Subjects Factors table.

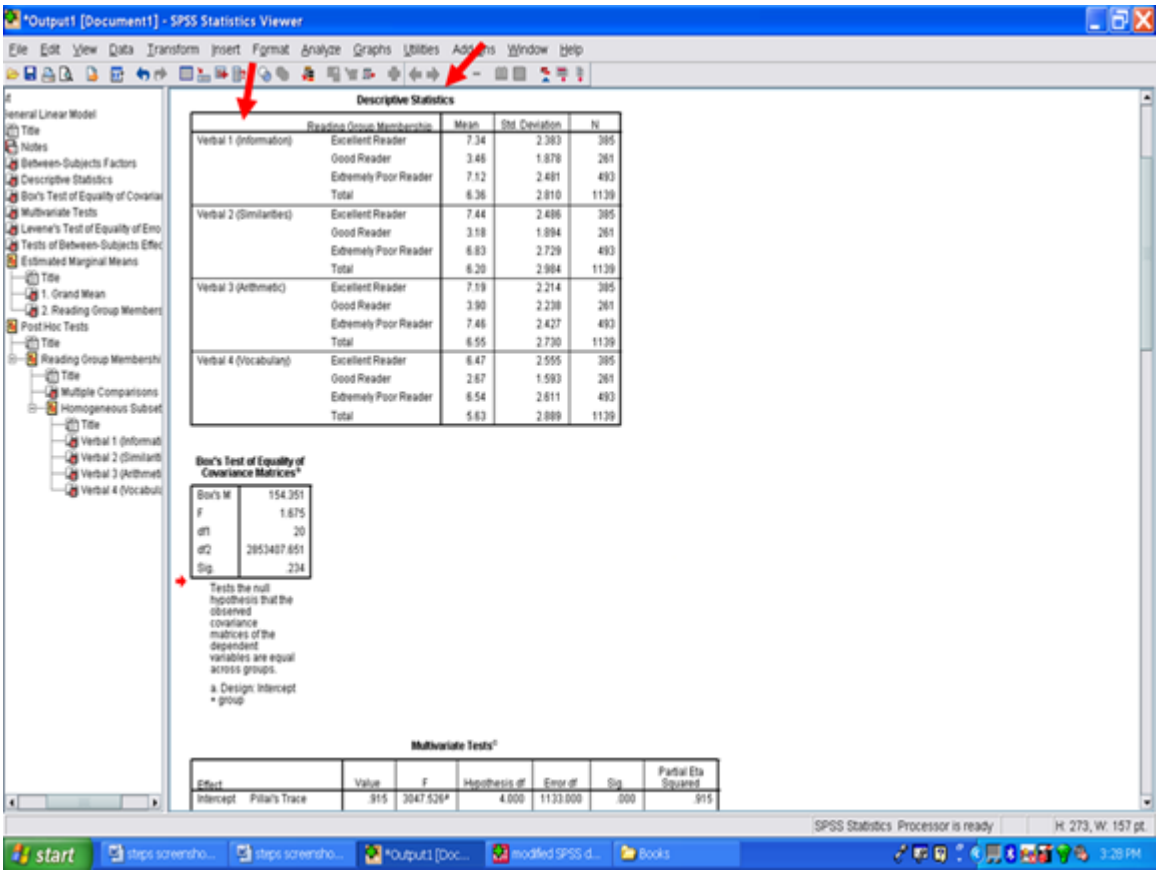
[1] This content is available online at <http://cnx.org/content/m40731/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40731/latest/www.writingandstatisticalhelp



[4]

Focusing on the Descriptive Statistics table will show you a column for the *M*, for the *SD*, and for the sample size, *n*. This information has been provided to you for each of the four dependent variables for each of the three groups. This information will be used in your Results section and should be kept.

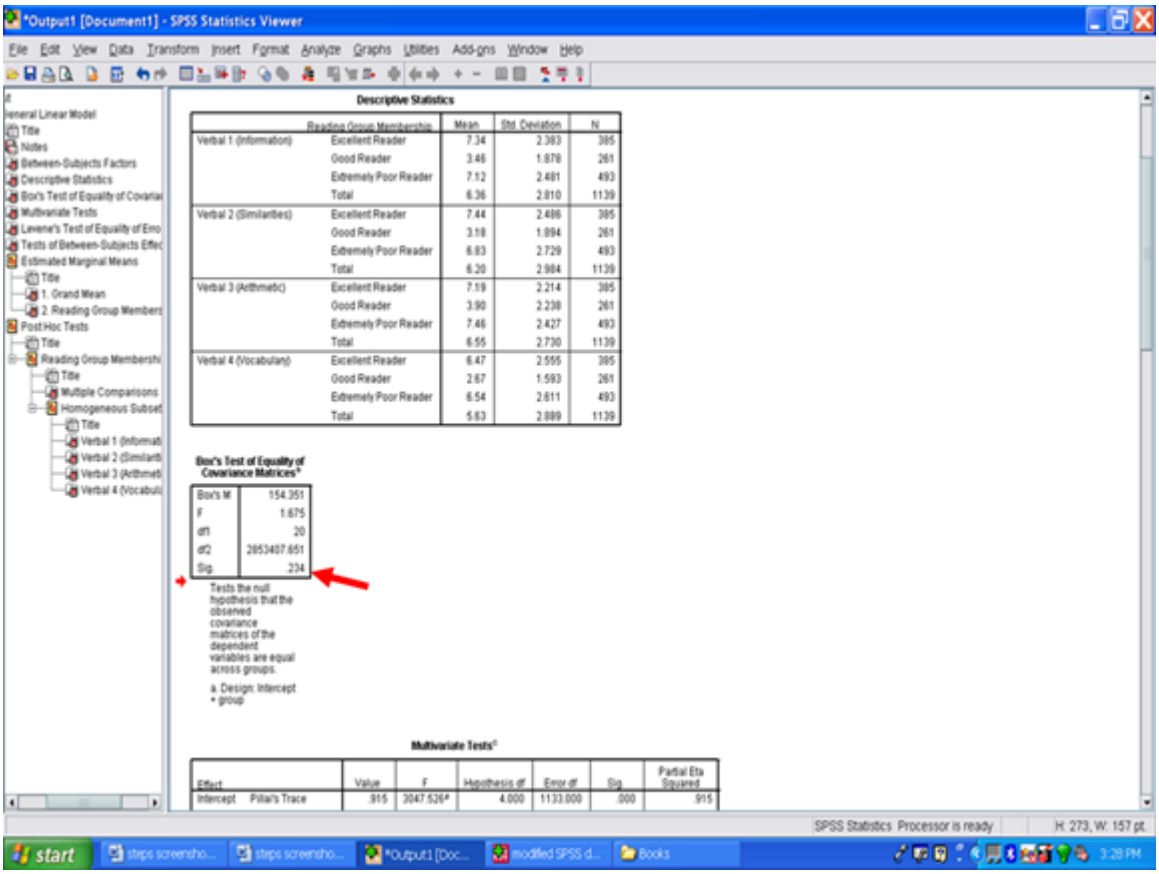
[4] http://cnx.org/content/m40731/latest/12.1.png/image



[5]

Underneath the Descriptive Statistics table is the Box's Test of Equality of Covariance Matrices. The information in this box involves checking one of the assumptions underlying use of a MANOVA procedure. The Sig. value in this box should be greater than .05 if the assumption is met. In the example below, the Sig. value is .234 which means that this particular assumption has been met. Readers should be informed whether the Box's *M* assumption was met or was violated. See Field (2010) for a detailed explanation of Box's *M* and the use of a MANOVA, even when this assumption is violated.

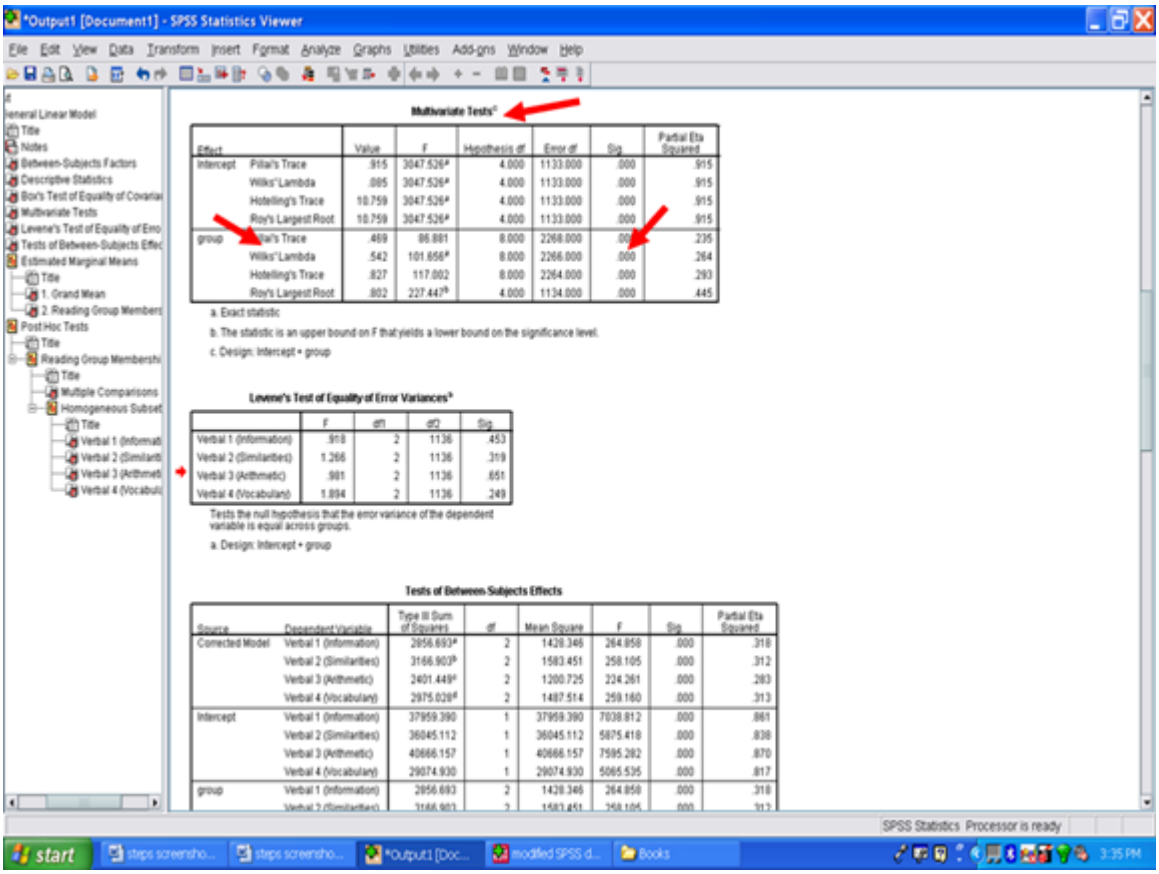
[5] http://cnx.org/content/m40731/latest/12.2.png/image



[6]

Underneath the Box's Test of Equality of Covariance Matrices is the Multivariate Tests table. This table is important because it indicates whether or not a statistically significant difference is present among the reading groups (i.e., our independent variable) in the aggregated dependent variable (i.e., aggregated Verbal 1 through Verbal 4). For our purposes, we will use Wilks' Lambda to determine whether or not a difference is present. In this example, a statistically significant difference is present.

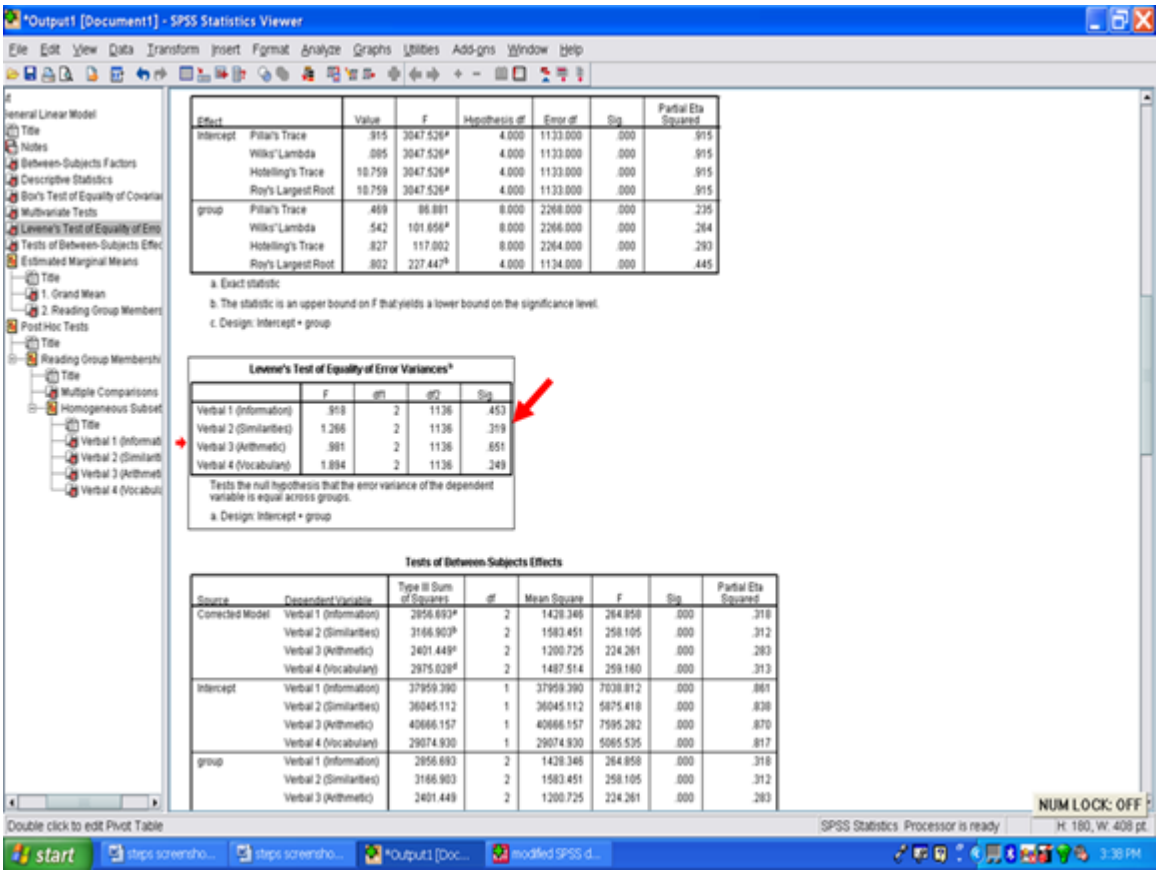
[6] http://cnx.org/content/m40731/latest/12.3.png/image



[7]

Next, the Levene's Test of Equality of Error Variances table is examined. This table involves checking the assumption that the variability is consistent within each dependent variable for each cell in the research design. The Sig. value in each row must be greater than .05 for each assumption to be met. In the table below, the assumption of equality of error variances is met for all four dependent variables as the Sig. values range from .249 to .651. Readers should be informed that this assumption for each dependent variable was met or was not met. See Field (2010) for a detailed discussion regarding Levene's Test and the robustness of MANOVA, even when this assumption is violated.

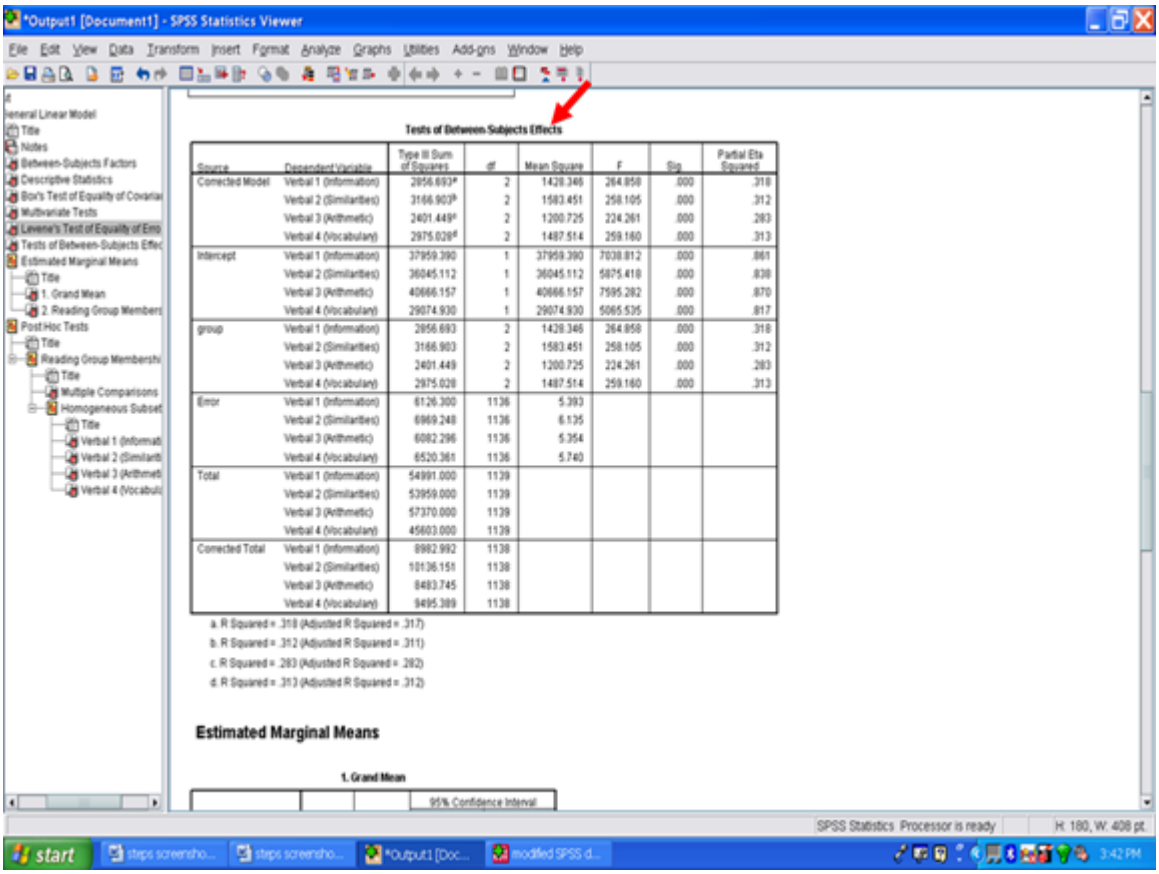
[7] http://cnx.org/content/m40731/latest/12.4.png/image



[8]

Underneath the Levene's Test of Equality table is the Tests of Between-Subjects Effects table. Contained in this table are the univariate ANOVAs regarding whether the reading groups differ on each of the four dependent variables. The Wilks' Lambda simply indicated whether an overall difference across the aggregated dependent variables was present. It did not provide information on whether the difference was for one, two, three, or all four of the dependent variables.

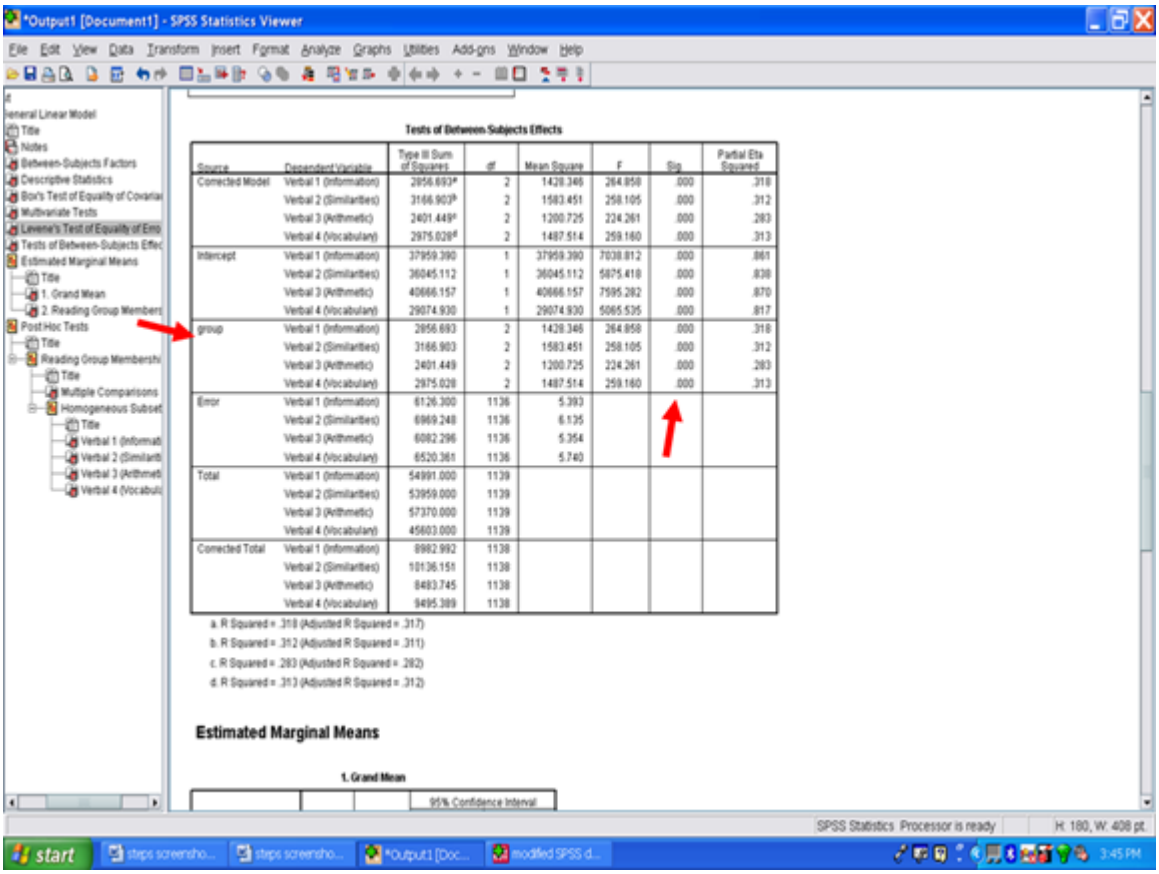
[8] http://cnx.org/content/m40731/latest/12.5.png/image



[9]

Find the row section that begins with the name of the independent variable. In our case, the independent variable is labeled group. Each row in this group section is a separate analysis of variance result. Examining the Sig. column for our four rows shows that each univariate ANOVA yielded a statistically significant result.

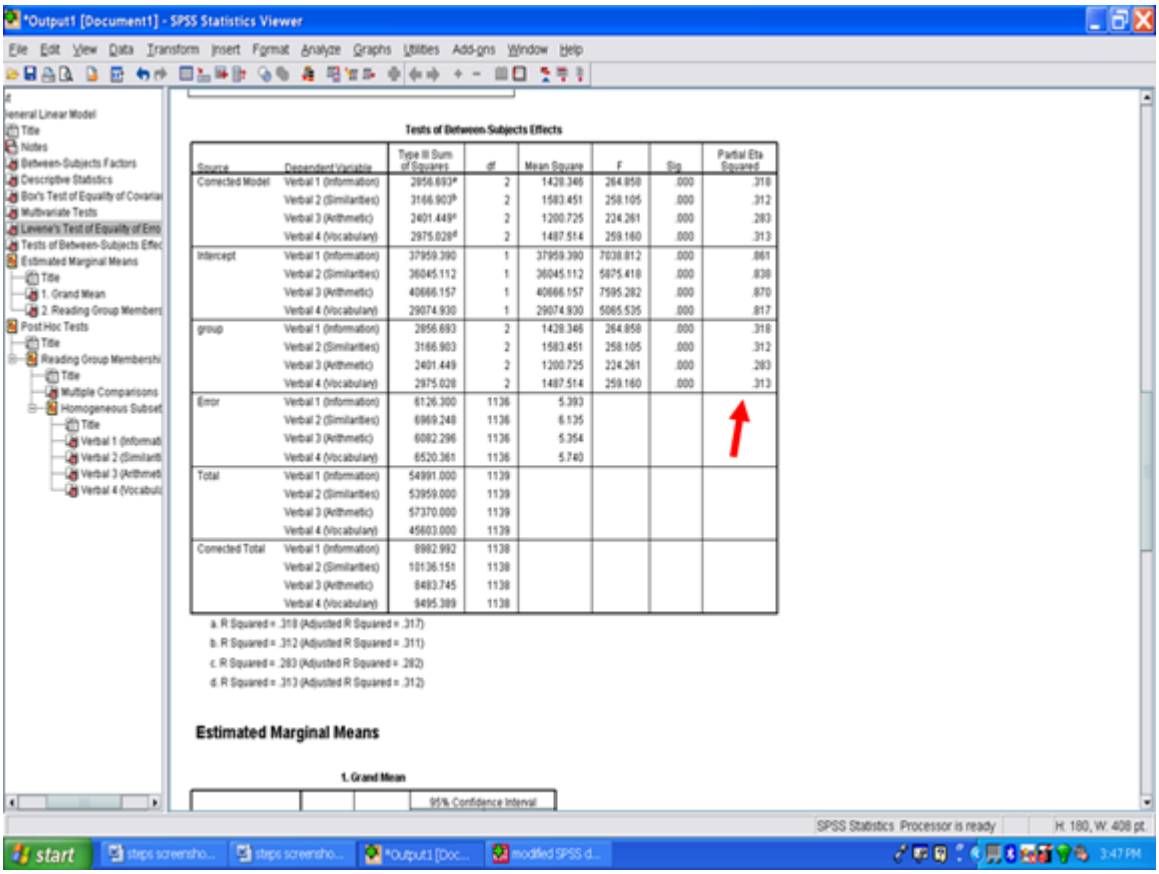
[9] http://cnx.org/content/m40731/latest/12.6.png/image



[10]

These results may be interpreted to mean that a statistically significant difference was present for Verbal 1, for Verbal 1, for Verbal 3, and for Verbal 4 as a function of the reading group. The effect size information for each statistically significant difference is present in the last column, the one labeled Partial Eta Squared. Partial eta squared, or n^2, is the effect size metric provided to you in the SPSS output. To interpret whether it is small, moderate, or large, see Cohen (1988).

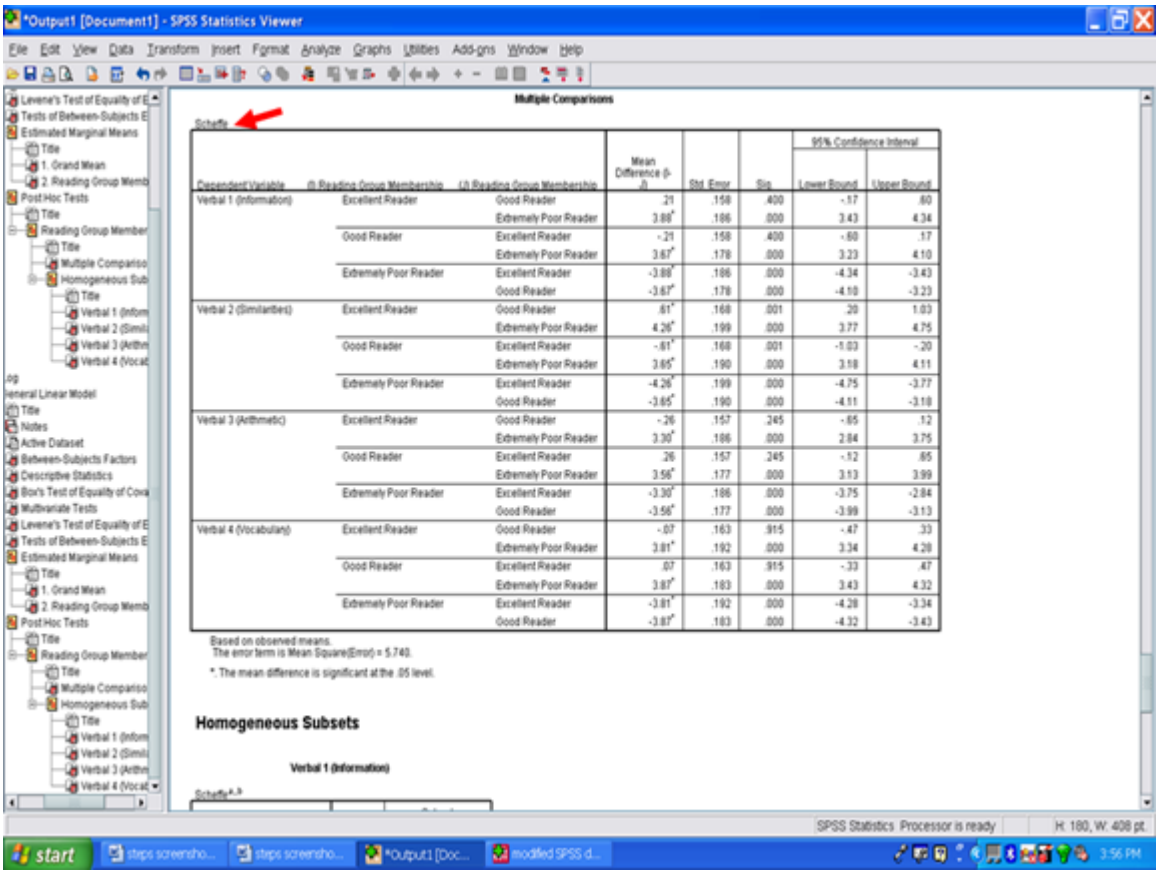
[10] http://cnx.org/content/m40731/latest/12.7.png/image



[11]

Because a statistically significant difference is present for each dependent variable, post hoc procedures must be examined to determine if all groups differ from each other or if only some of the groups differ. Post hoc procedure information is provided in the Multiple Comparisons table. The arrow below is pointed toward the specific type of post hoc we requested.

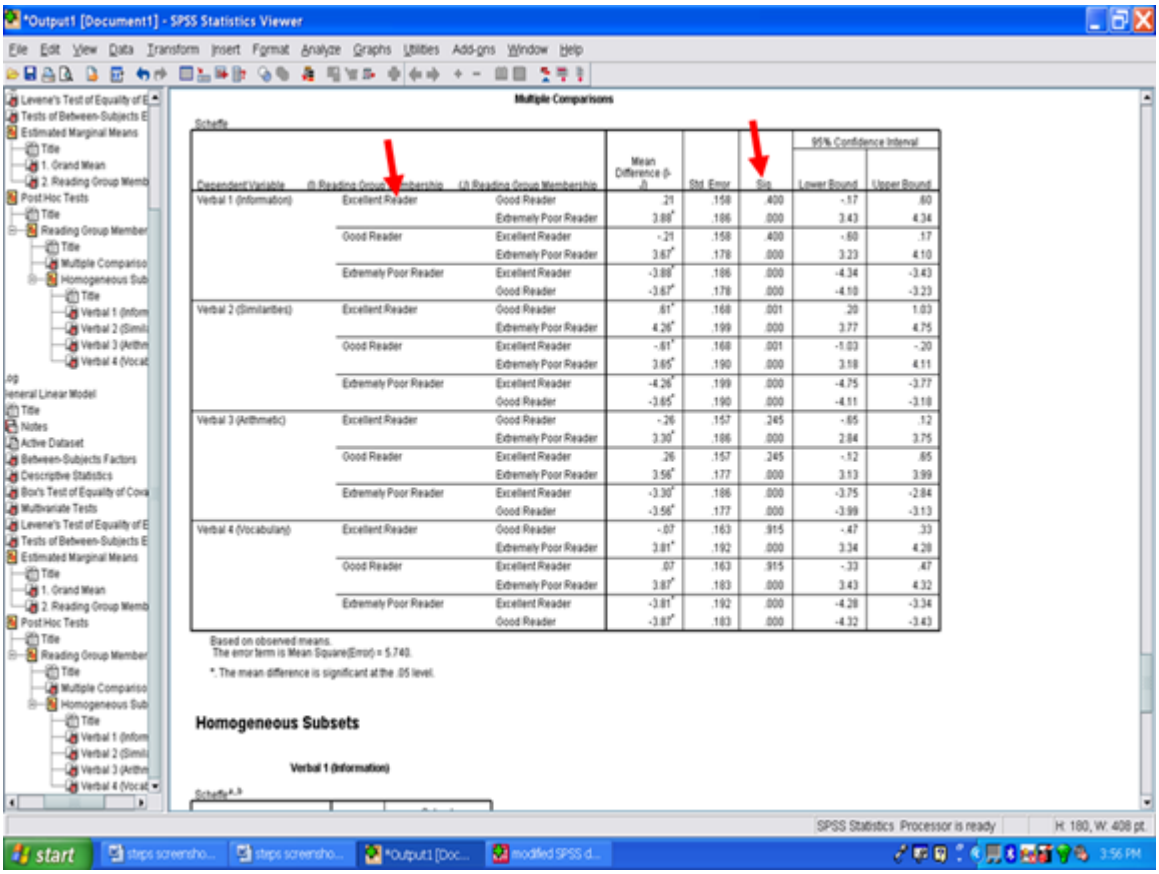
[11] http://cnx.org/content/m40731/latest/12.8.png/image



[12]

The most important column in this post hoc table is the Sig. column. For a pairwise comparison to be statistically significant, the Sig. value must be .05 or below. In the example below for the Verbal 1 set of comparisons, Excellent Readers did not differ from Good Readers, however, Excellent readers did differ from Extremely Poor Readers.

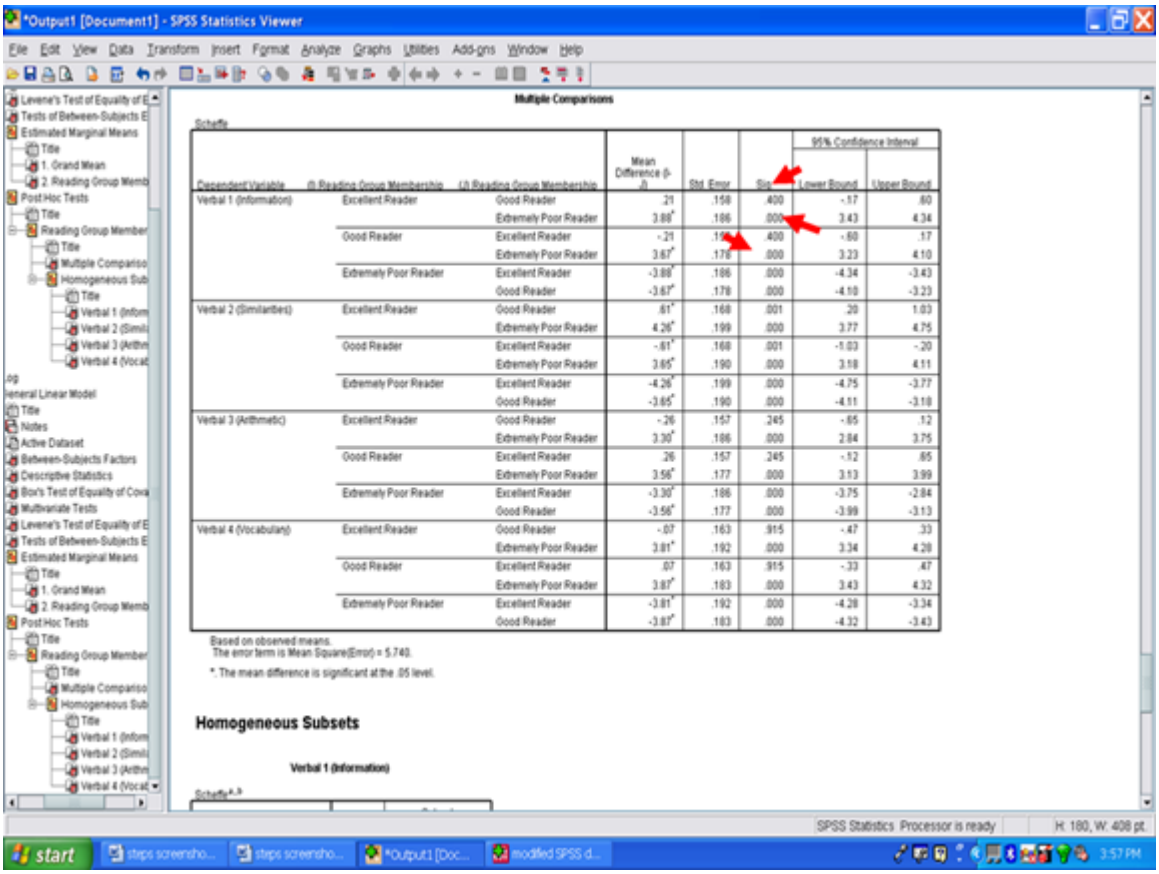
[12] http://cnx.org/content/m40731/latest/12.9.png/image



[13]

Readers should note that only three pairwise comparisons are unique in the table for each dependent variable. SPSS, however, provides information on six pairwise comparisons. When three groups are present, as they are in this example, row 1, row 2, and row 4 are unique. Row 1 is Excellent Reader compared to Good Reader. Row 2 is Excellent Reader compared to Extremely Poor Reader. Row 4 is Good Reader compared to Extremely Poor Reader.

[13] http://cnx.org/content/m40731/latest/12.10.png/image



[14]

Each of the four dependent variables in this example have post hoc procedures calculated for them and are present in the table above.

All of the useful information present in the SPSS output for the MANOVA procedure has now been covered.

[14] http://cnx.org/content/m40731/latest/12.11.png/image

Chapter 4

Discriminant Analysis: Assumptions[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

In this set of steps, readers will learn how to conduct a canonical discriminant analysis procedure. For detailed information regarding the assumptions underlying use of a discriminant analysis, readers are referred to the Hyperstats Online Statistics Textbook at http://davidmlane.com/hyperstat/[4] ; to the *Electronic Statistics Textbook* (2011)

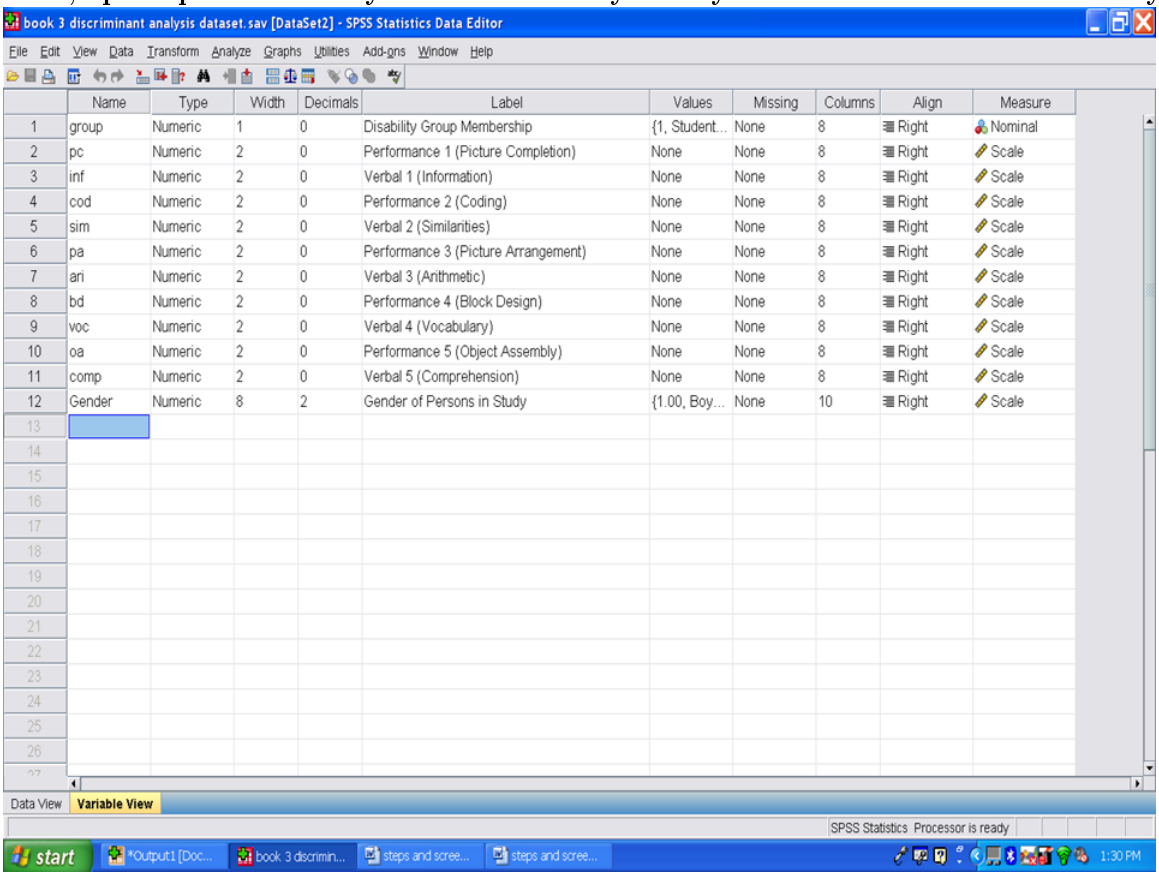
[1] This content is available online at <http://cnx.org/content/m40733/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40733/latest/www.writingandstatisticalhelp
[4] http://davidmlane.com/hyperstat/

at http://www.statsoft.com/textbook/[5] ; or to Andy Field's (2009) Discovering Statistics Using SPSS at http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967862&sr=1-1[6]

Research questions for which a discriminant analysis procedure is appropriate involve determining variables that predict group membership. For example, if two groups of persons are present such as completers and non-completers and archival data are available, then a discriminant analysis procedure could be utilized. Such a procedure could identify specific variables that differentiate group membership. As such, interventions could be developed and targeted toward the variables that predicted group membership. Other sample research questions for which a discriminant analysis might be appropriate: (a) What factors differentiates successful from unsuccessful students?; (b) What factors differentiate delinquents from nondelinquents?; (c) What set of test scores best differentiates students with LD, students who are failing, and students with MR?; and (d) What set of factors differentiates drop-outs from persisters?

For purposes of this chapter, our research question is: "What scholastic variables differentiate boys from girls?"

First, open up the dataset you intend to analyze for your canonical discriminant analysis.



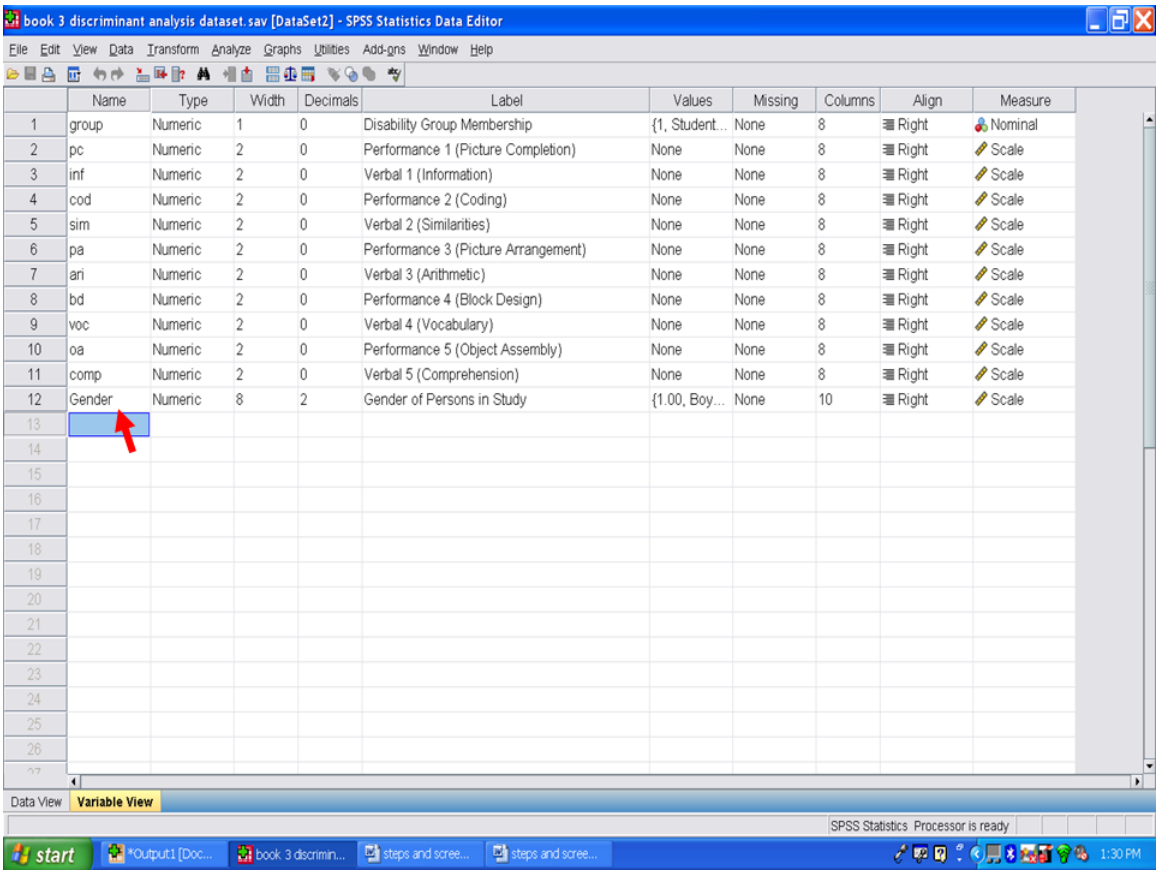
[7]

Our independent variable is gender. Boys are labeled as group 1 and girls are labeled as group 2.

[5] http://www.statsoft.com/textbook/

[6] http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967862&sr=1-1

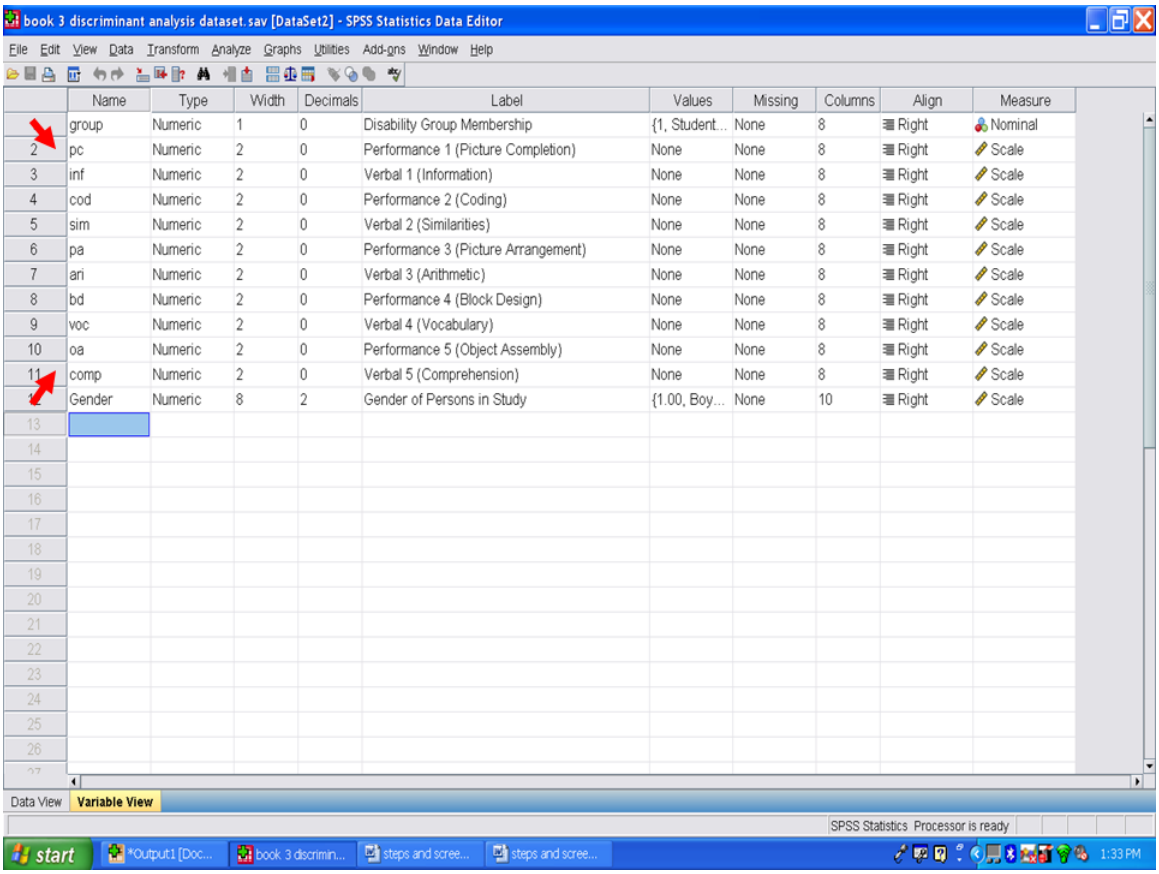
[7] http://cnx.org/content/m40733/latest/13.1.png/image



[8]

Our dependent variables, the ones we will use to differentiate boys from girls are 10 subscales from the Wechsler Intelligence Scale for Children-Third Edition: Picture Completion (pc), Information (inf), Coding (cod), Similarities (sim), Picture Arrangement (pa), Arithmetic (ari), Block Design (bd), Vocabulary (voc), Object Assembly (oa), and Comprehension (comp).

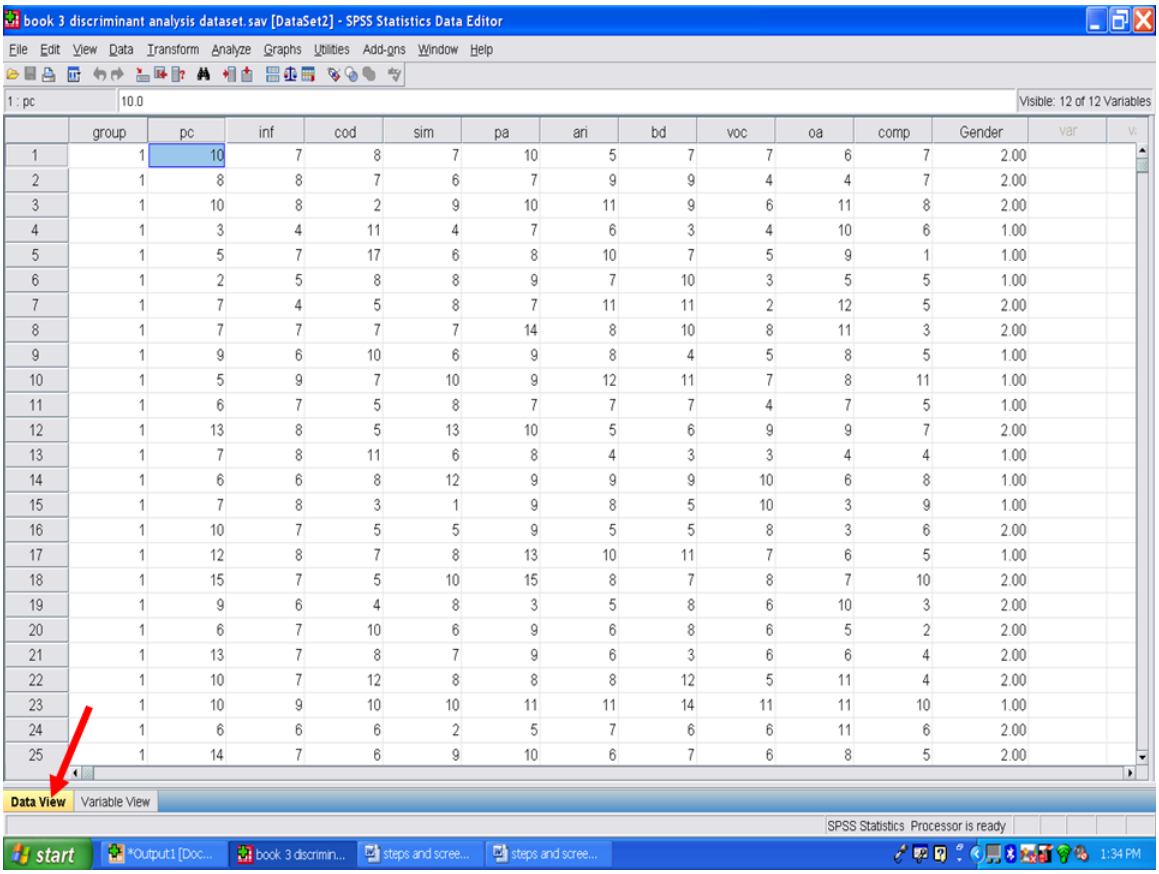
[8] http://cnx.org/content/m40733/latest/13.2.png/image



[9]

In the previous screenshots, we were in the variable view screen. Click on data view, shown below, so that your screen looks like the one below.

[9] http://cnx.org/content/m40733/latest/13.3.png/image



10

Prior to conducting a canonical discriminant function, we need to check the assumptions that underlie its use.

4.1 Normal Distribution

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution. You can examine whether or not variables are normally distributed with histograms of frequency distributions. However, note that violations of the normality assumption are usually not "fatal," meaning, that the resultant significance tests etc. are still "trustworthy." You may use specific tests for normality in addition to graphs. http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions[11]

We recommend that you calculate the standardized skewness coefficients and the standardized kurtosis coefficients, as discussed in other chapters.

*** Skewness [Note. Skewness refers to the extent to which the data are normally distributed around the mean. Skewed data involve having either mostly high scores with a few low ones or having mostly low scores with a few high ones.] Readers are referred to the following sources for a more detailed definition of skewness: http://www.statistics.com/index.php?page=glossary&term_id=356[12] and http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[13]**

[10] http://cnx.org/content/m40733/latest/13.4.png/image
[11] http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions
[12] http://www.statistics.com/index.php?page=glossary&term_id=356
[13] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb

To standardize the skewness value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the skewness value from the SPSS output and divide it by the Std. error of skewness. If the resulting calculation is within -3 to +3, then the skewness of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

*** Kurtosis [Note. Kurtosis also refers to the extent to which the data are normally distributed around the mean. This time, the data are piled up higher than normal around the mean or piled up higher than normal at the ends of the distribution.] Readers are referred to the following sources for a more detailed definition of kurtosis: http://www.statistics.com/index.php?page=glossary&term_id=326[14] and http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[15]**

To standardize the kurtosis value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the kurtosis value from the SPSS output and divide it by the Std. error of kurtosis. If the resulting calculation is within -3 to +3, then the kurtosis of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

4.2 Homogeneity of Variances/Covariances

It is assumed that the variance/covariance matrices of variables are homogeneous across groups. Again, minor deviations are not that important. http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions[16]

4.3 Correlations between Means and Variances

The major "real" threat to the validity of significance tests occurs when the means for variables across groups are correlated with the variances (or standard deviations). Intuitively, if there is large variability in a group with particularly high means on some variables, then those high means are not reliable. However, the overall significance tests are based on pooled variances, that is, the average variance across all groups. Thus, the significance tests of the relatively larger means (with the large variances) would be based on the relatively smaller pooled variances, resulting erroneously in statistical significance. In practice, this pattern may occur if one group in the study contains a few extreme outliers, who have a large impact on the means, and also increase the variability. To guard against this problem, inspect the descriptive statistics, that is, the means and standard deviations or variances for such a correlation. http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions[17]

After calculating the means and standard deviations for your variables for each of your groups, check them to determine if large variability is present in the means for one of your groups compared to the means for the other group.

4.4 The Matrix Ill-Conditioning Problem

Another assumption of discriminant function analysis is that the variables that are used to discriminate between groups are not completely redundant. As part of the computations in-

[14] http://www.statistics.com/index.php?page=glossary&term_id=326
[15] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb
[16] http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions
[17] http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions

volved in discriminant analysis, you will invert the variance/covariance matrix of the variables in the model. If any one of the variables is completely redundant with the other variables then the matrix is said to be *ill-conditioned*, and it cannot be inverted. For example, if a variable is the sum of three other variables that are also in the model, then the matrix is ill-conditioned. http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions[18]

What this assumption means is that each variable should be unique from any other variable in the analysis. Having one variable that includes another variable would be a violation of this assumption. An example of this would be using a total score that contains several subscale scores, all of which are used in the discriminant analysis.

4.5 Tolerance Values.

In order to guard against matrix ill-conditioning, constantly check the so-called tolerance value for each variable. This tolerance value is computed as *1 minus R-square* of the respective variable with all other variables included in the current model. Thus, it is the proportion of variance that is unique to the respective variable. In general, when a variable is almost completely redundant (and, therefore, the matrix ill-conditioning problem is likely to occur), the tolerance value for that variable will approach 0. http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions[19]

We will check this assumption, the tolerance values, when we examine the SPSS output.

[18] http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions
[19] http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions

Chapter 5

Discriminant Analysis: Part I[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

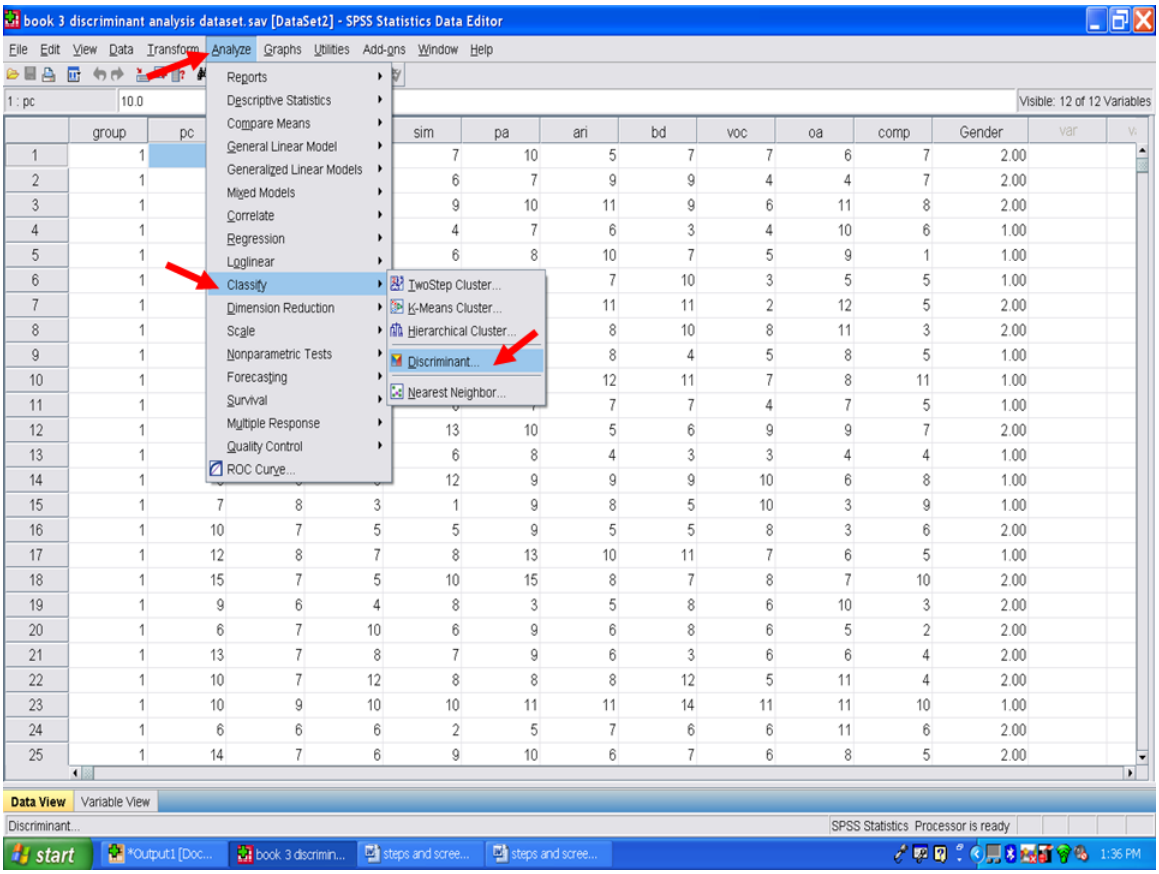
John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

Now that we have ascertained the assumptions underlying use of a discriminant analysis procedure, we will begin the steps.

Click on Analyze
Click on Classify
Click on Discriminant

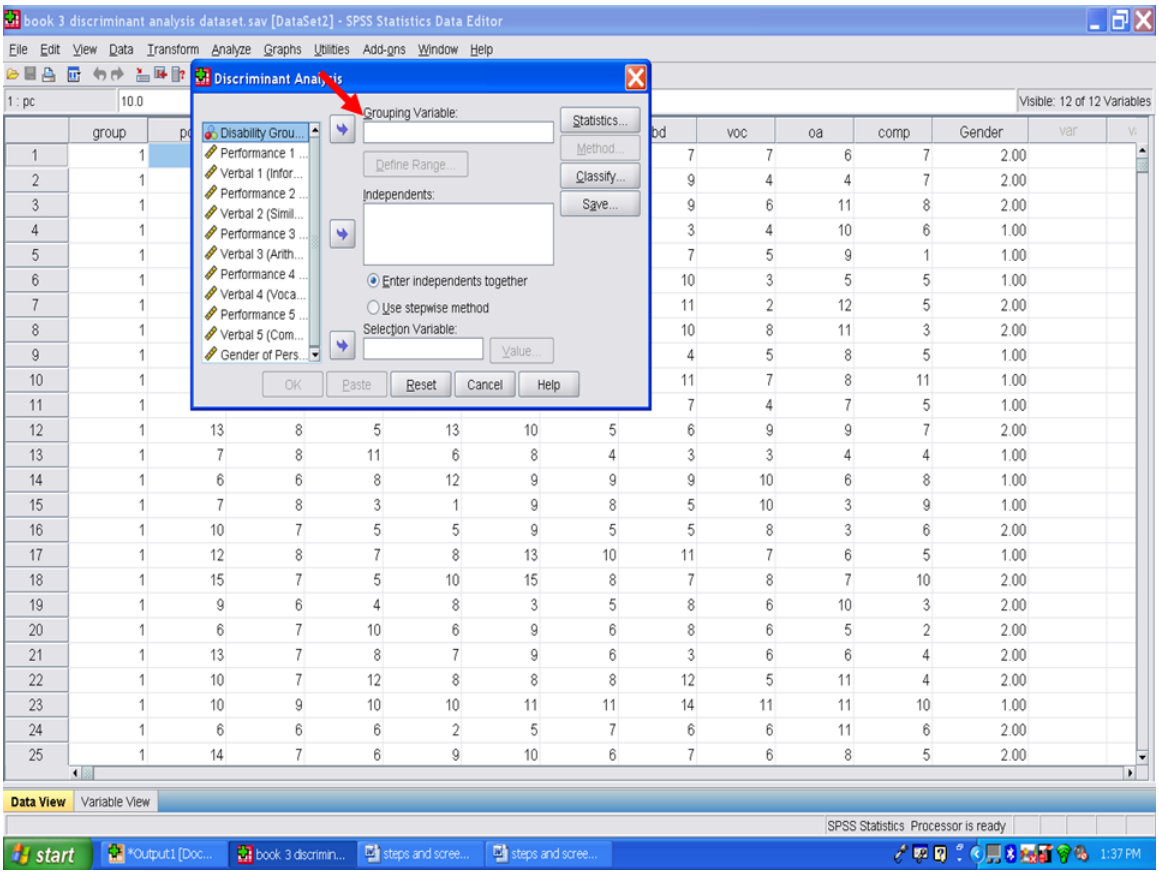
[1] This content is available online at <http://cnx.org/content/m40735/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40735/latest/www.writingandstatisticalhelp



[4]

The following screen will then appear. Grouping variable is your independent variable. Remember that in our example that gender is our independent or grouping variable.

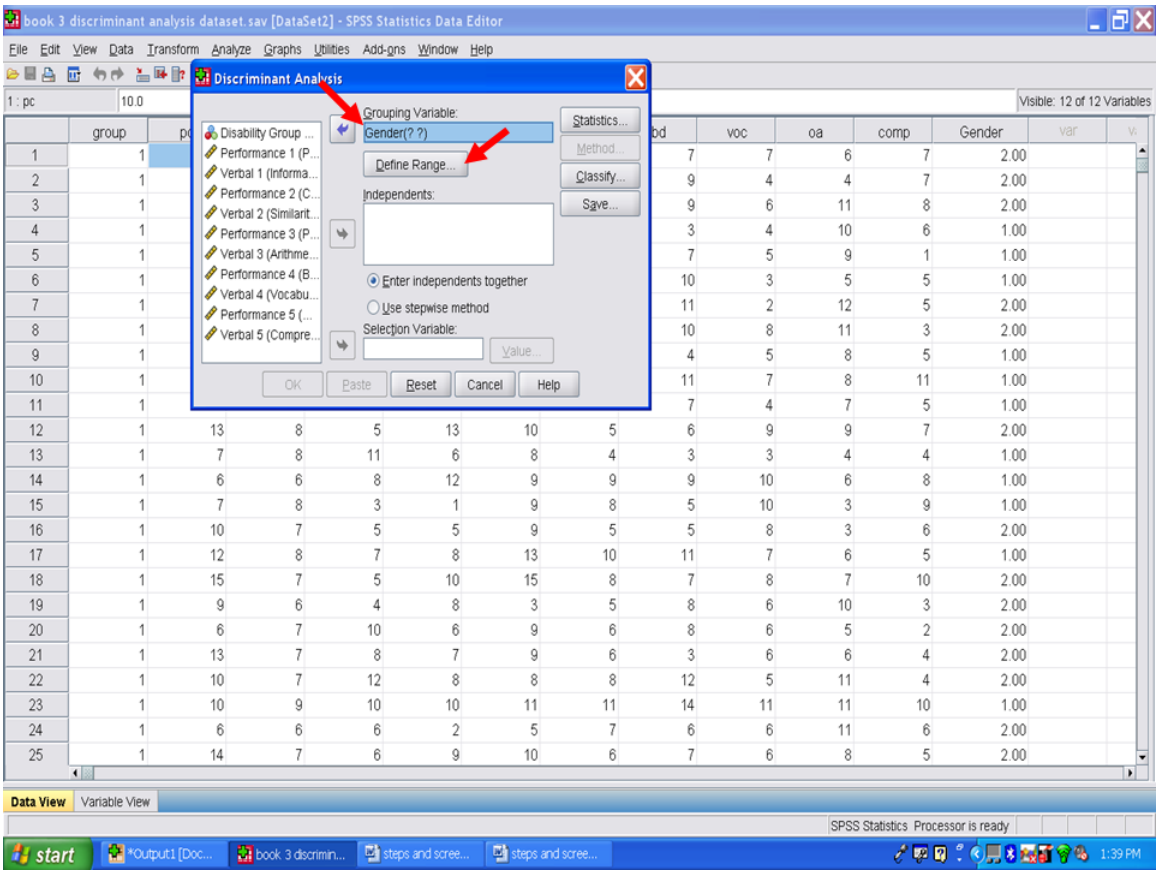
[4] http://cnx.org/content/m40735/latest/14.1.png/image



[5]

Highlight gender and send it into the Grouping variable cell. When you do so, you will note that after Gender is a set of () with question marks. Click on Define Range.

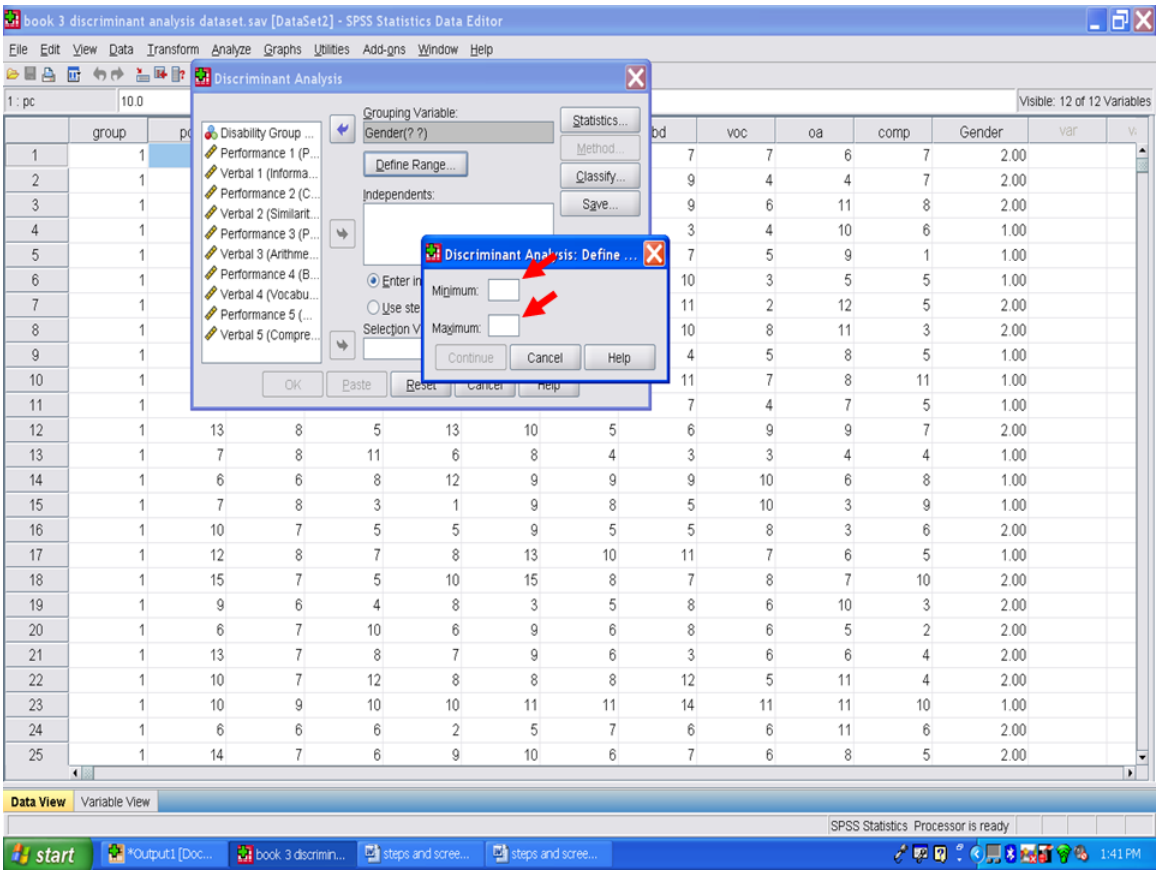
[5] http://cnx.org/content/m40735/latest/14.2.png/image



6

After clicking on Define Range, the following screen will appear: For the Minimum box, we will type in the number 1 because it represents boys. We will type in the number 2 in the Maximum box because it represents girls.

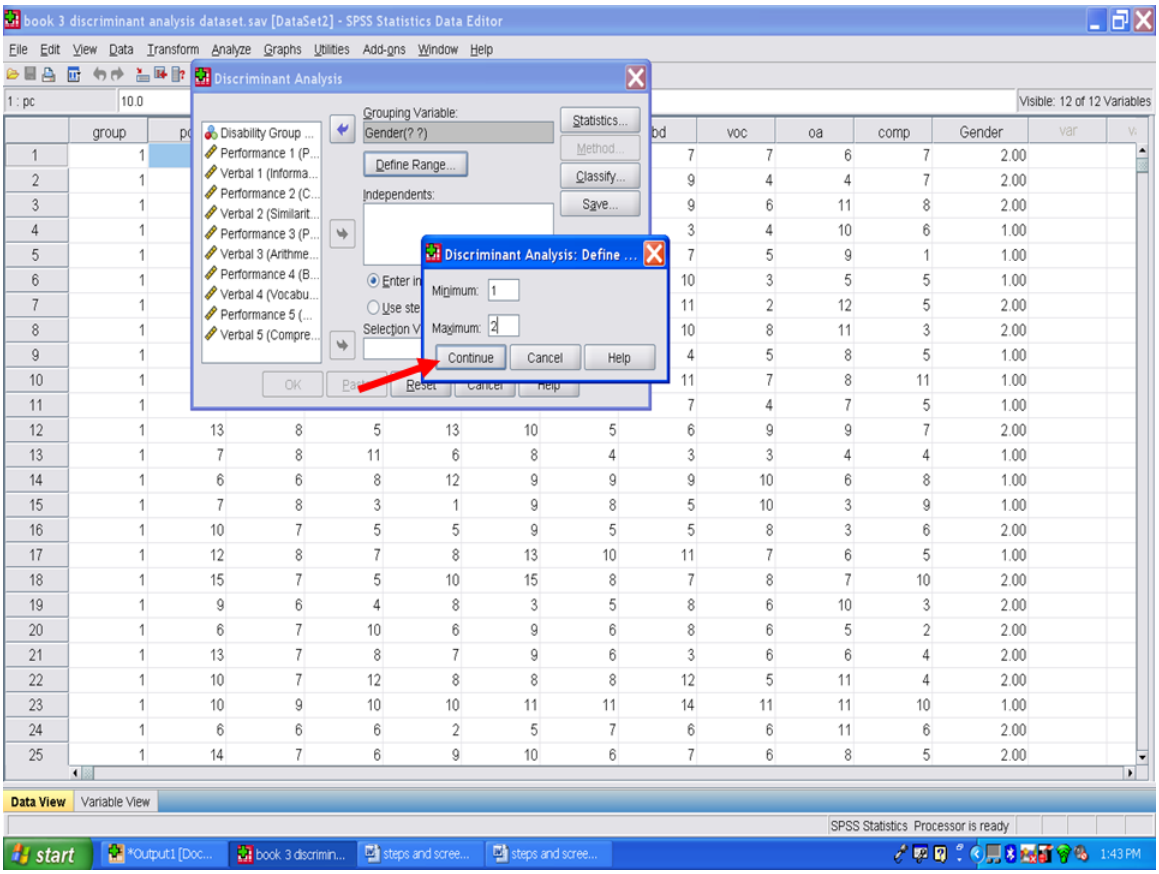
[6] http://cnx.org/content/m40735/latest/14.3.png/image



[7]

After typing in 1 and 2, then click on Continue.

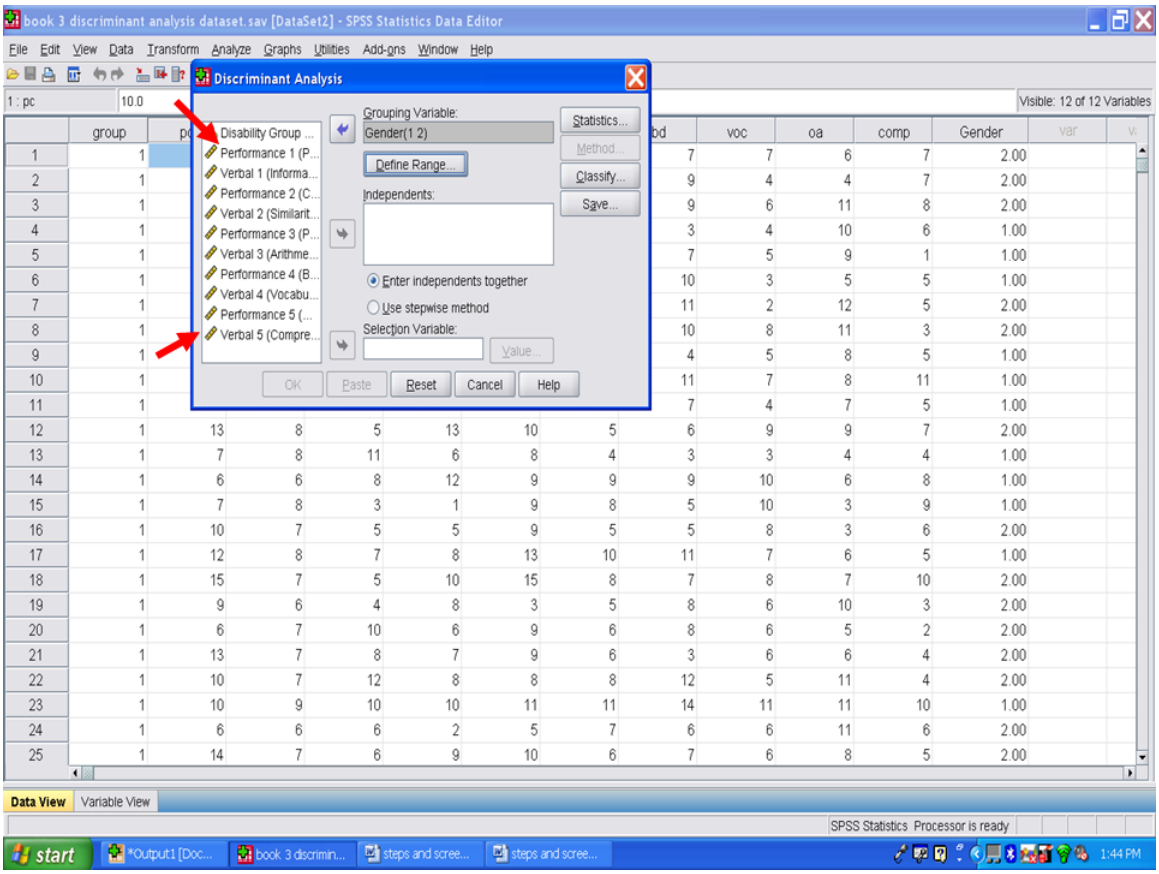
[7] http://cnx.org/content/m40735/latest/14.4.png/image



[8]

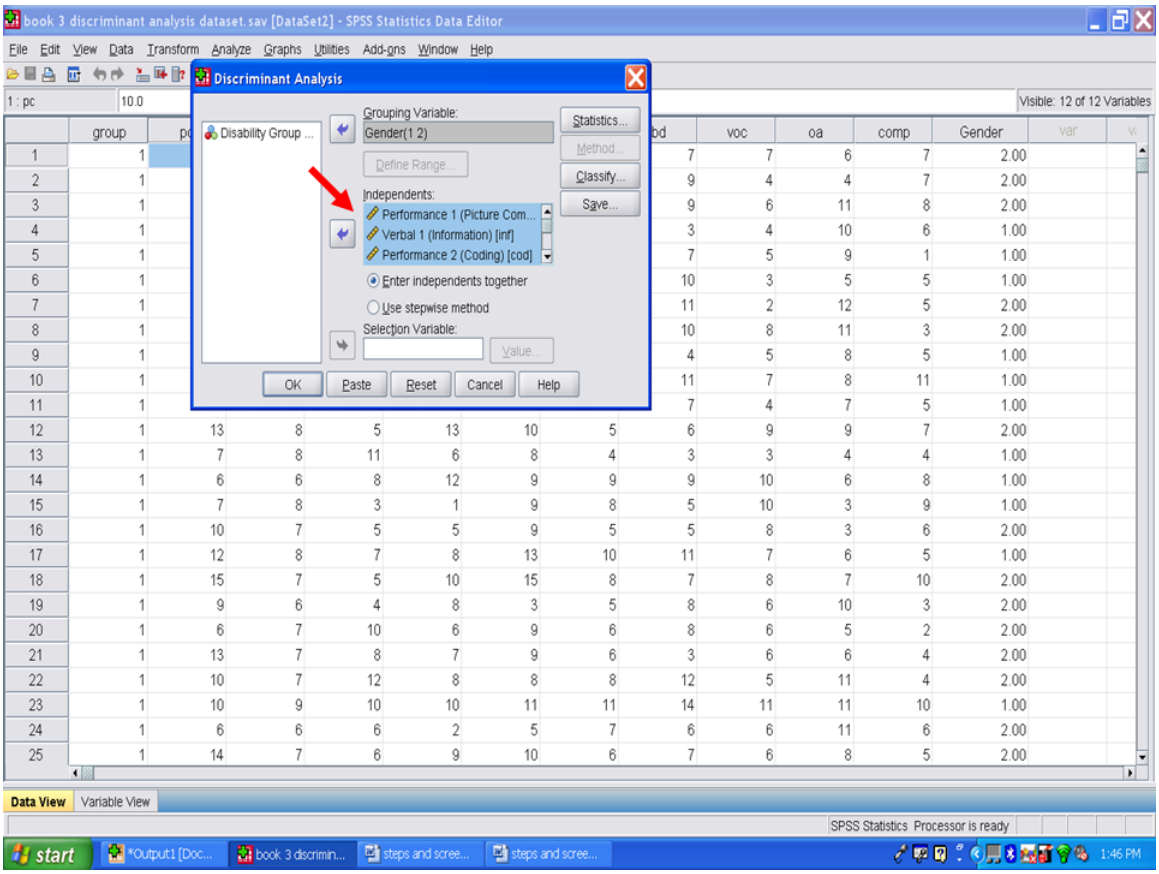
Next we will send over our 10 dependent variables, the 10 subscales, to the Independents box. Highlight Performance 1 through Verbal 5; click on the middle arrow; send them to the Independents box.

[8] http://cnx.org/content/m40735/latest/14.5.png/image



The screen should now look like the one below:

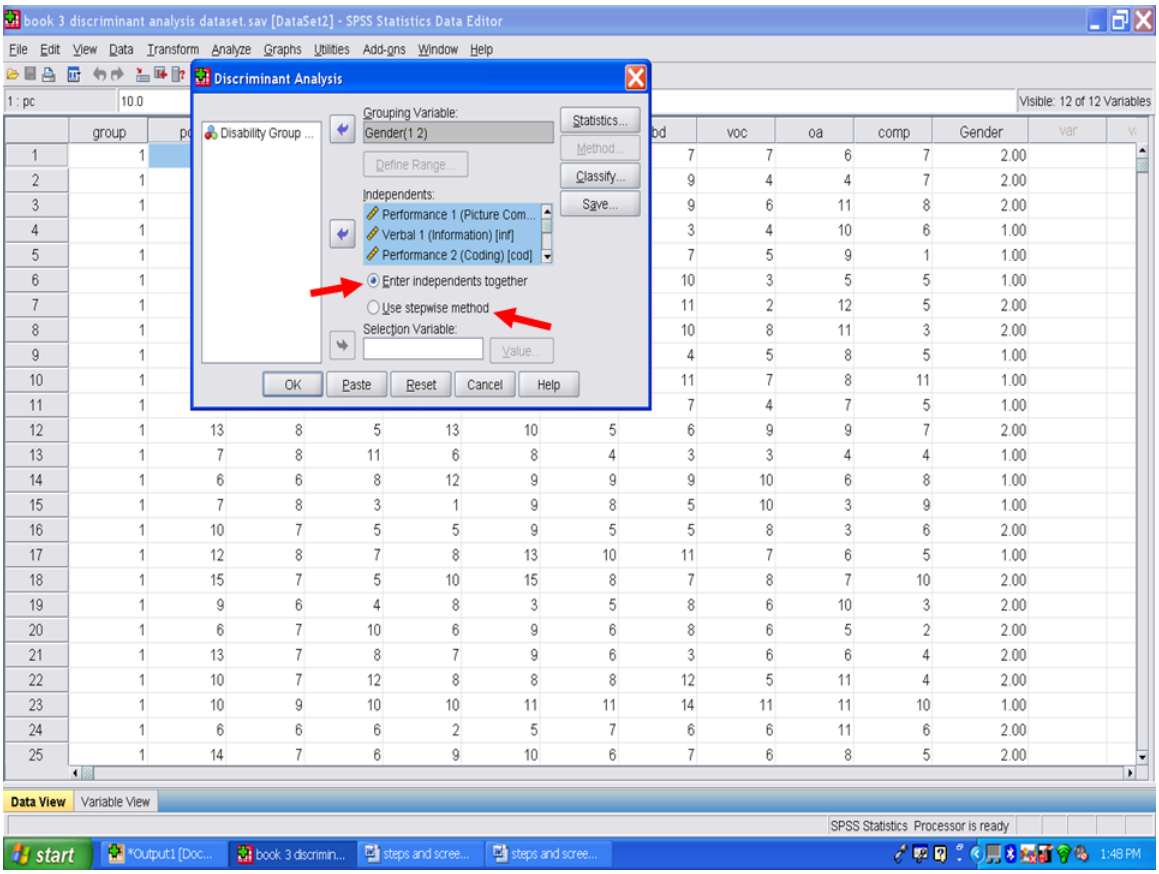
[9] http://cnx.org/content/m40735/latest/14.6.png/image



[10]

Next we will change the specific type of discriminant analysis from the default of Enter Independents Together to Use Stepwise Method.

[10] http://cnx.org/content/m40735/latest/14.7.png/image



[11]

[11] http://cnx.org/content/m40735/latest/14.8.png/image

Chapter 6

Discriminant Analysis: Part II[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

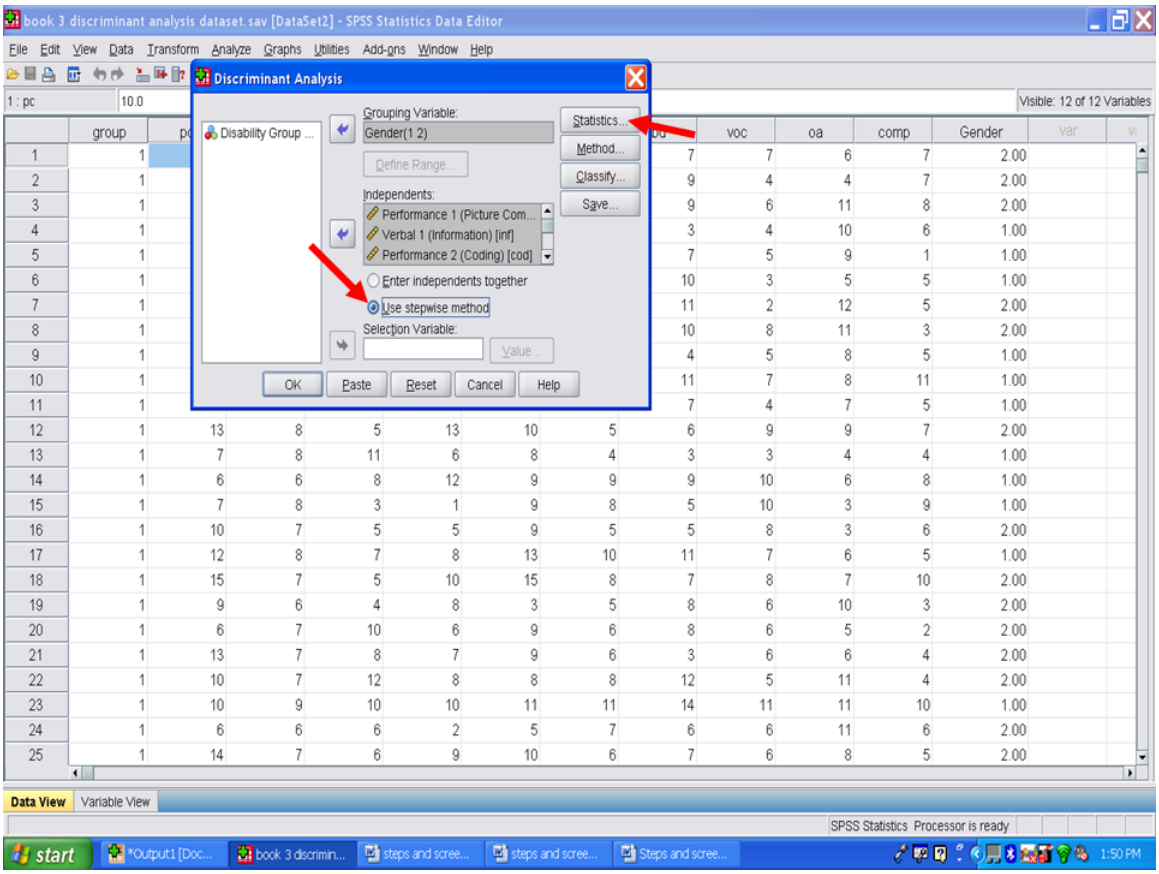
About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

Your screen should now look like the one below: Now click on Statistics.

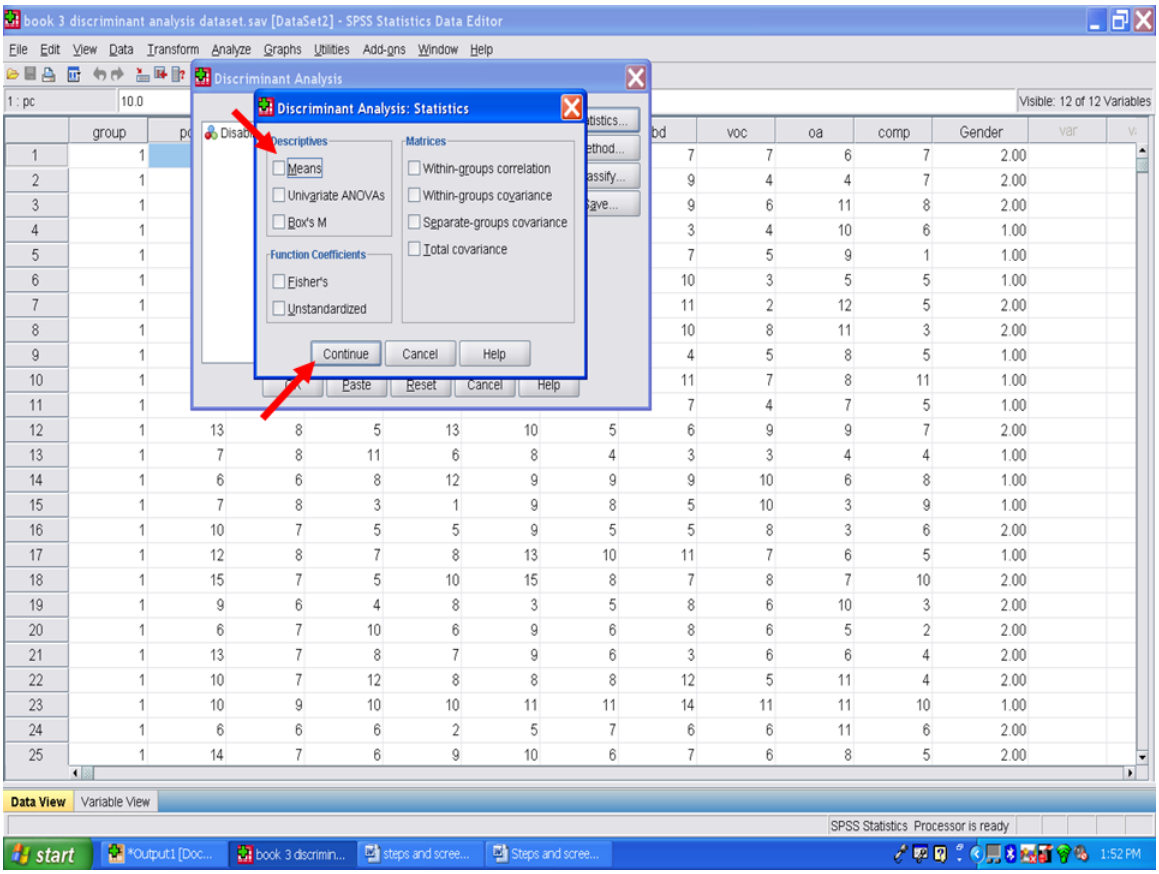
[1] This content is available online at <http://cnx.org/content/m40736/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40736/latest/www.writingandstatisticalhelp



[4]

The following screen then appears.
Click on Means.
Click on Continue.

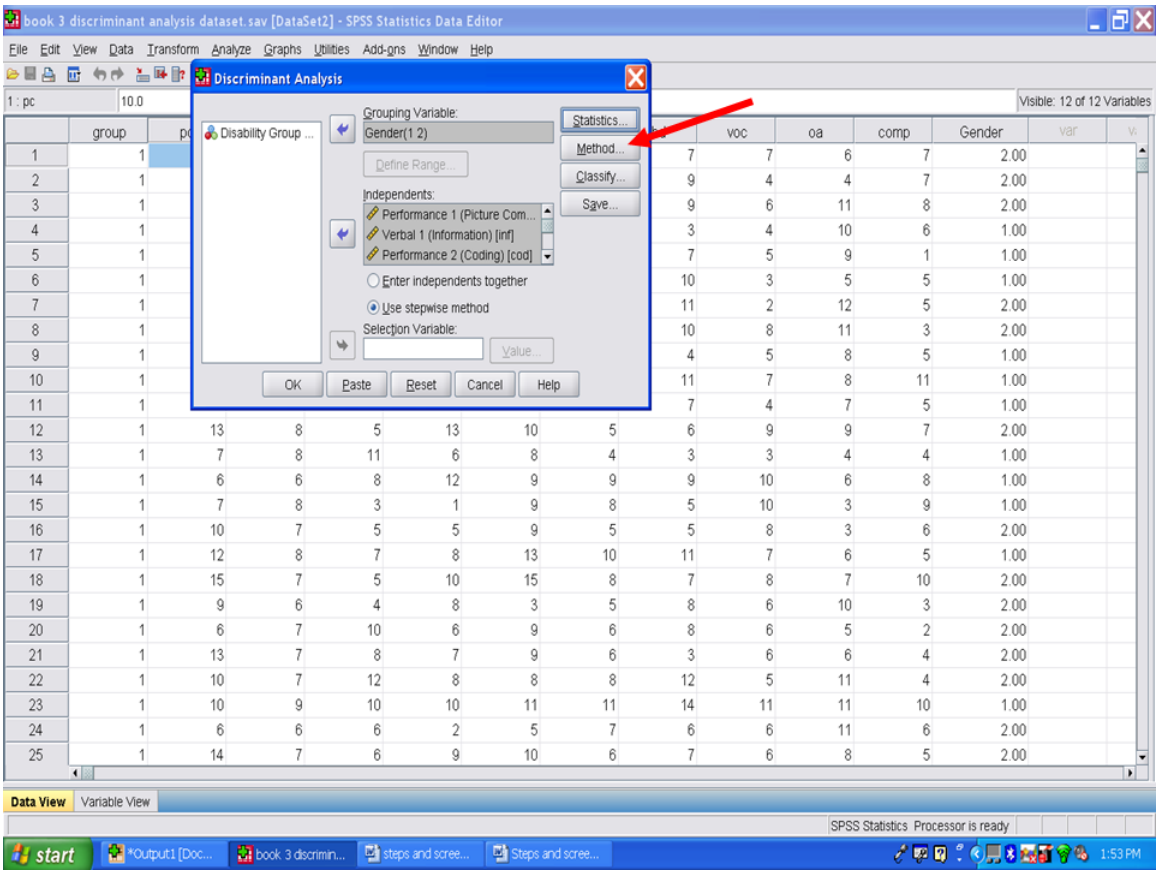
[4] http://cnx.org/content/m40736/latest/15.1.png/image



[5]

When the screen below appears, click on Method

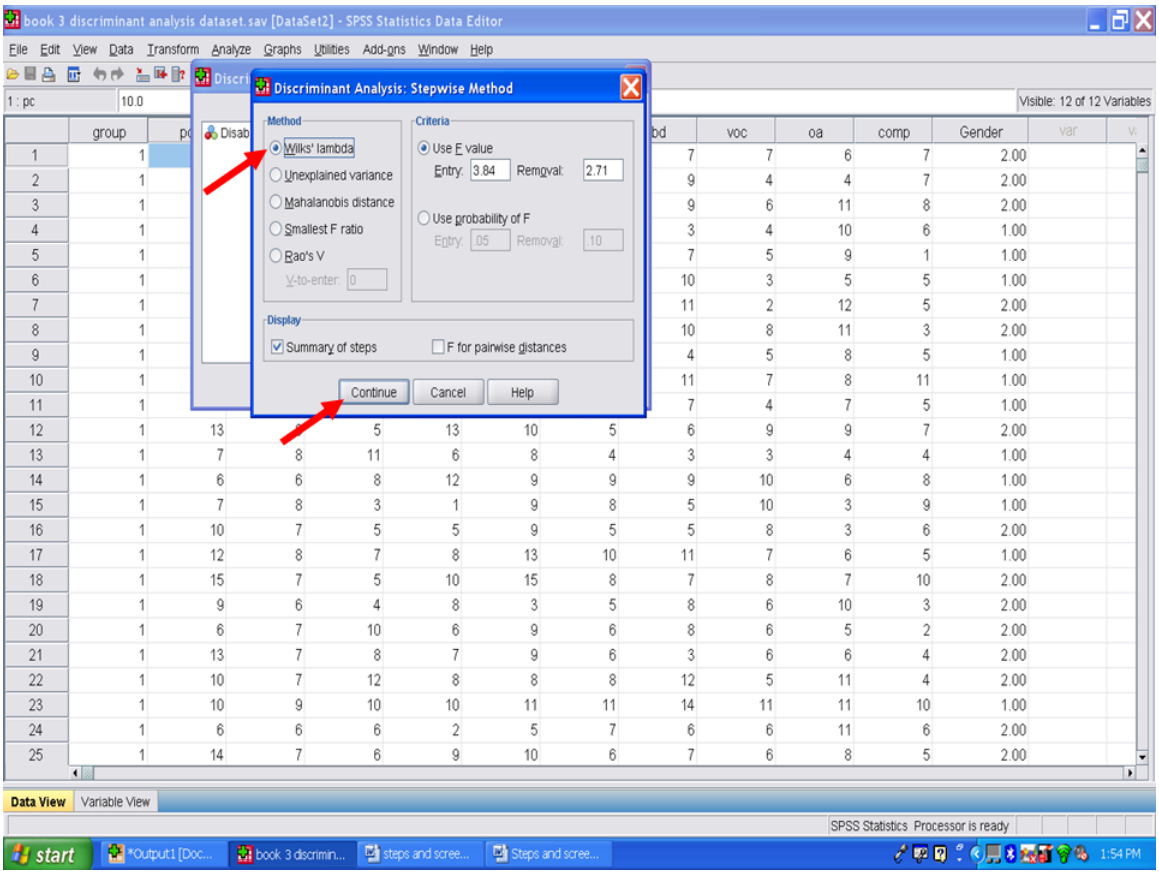
[5] http://cnx.org/content/m40736/latest/15.2.png/image



[6]

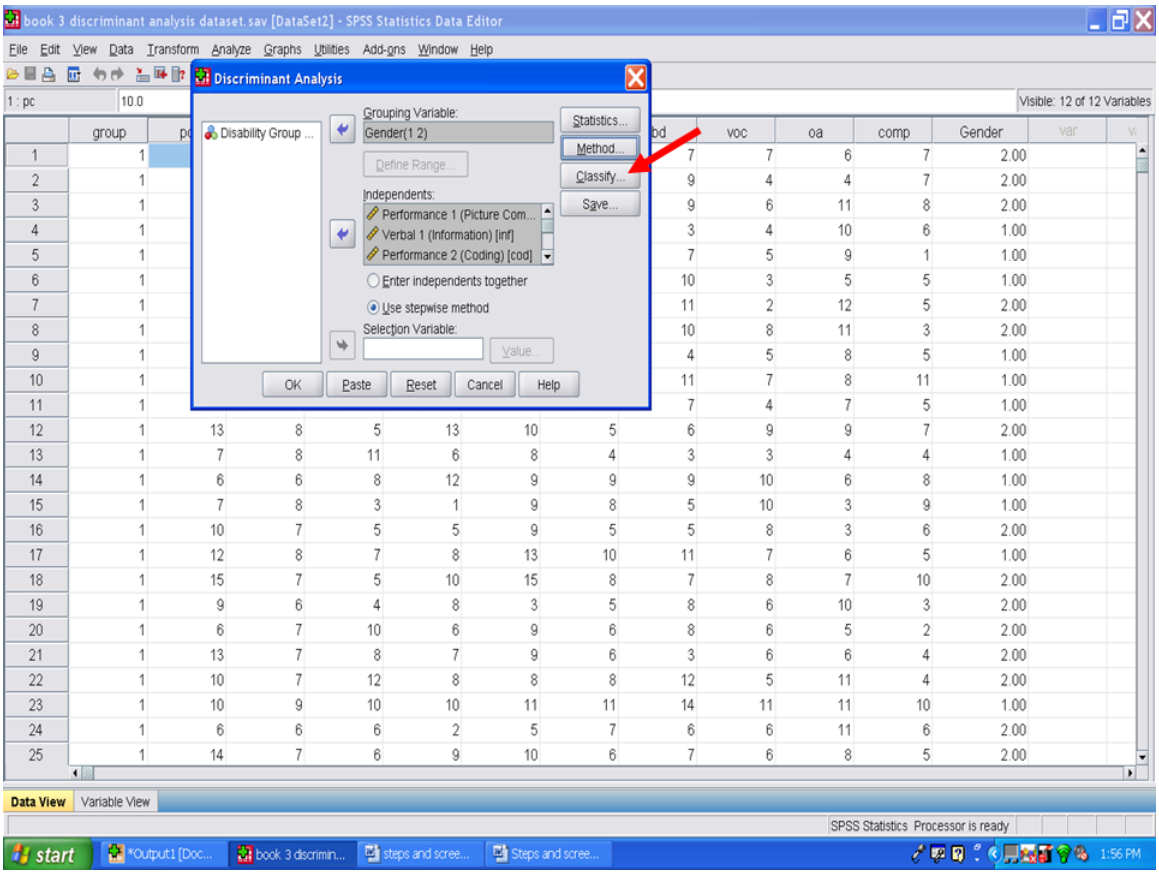
In the Method screen, we will use the default of Wilks' lambda. Click on Continue.

[6] http://cnx.org/content/m40736/latest/15.3.png/image



Then click on Classify.

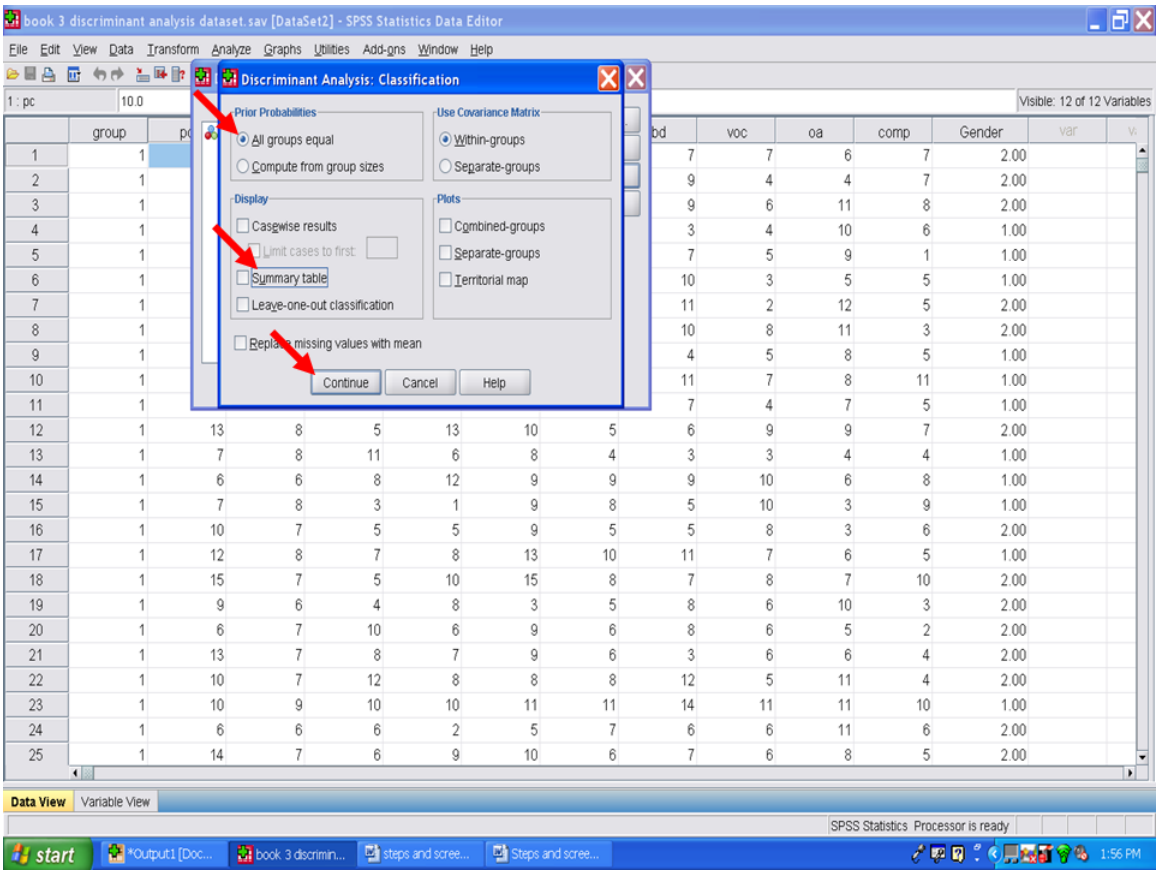
[7] http://cnx.org/content/m40736/latest/15.4.png/image



[8]

In the screen that appears, the All groups equal button is already clicked.
Click on Summary table under the Display.
Then click on Continue.

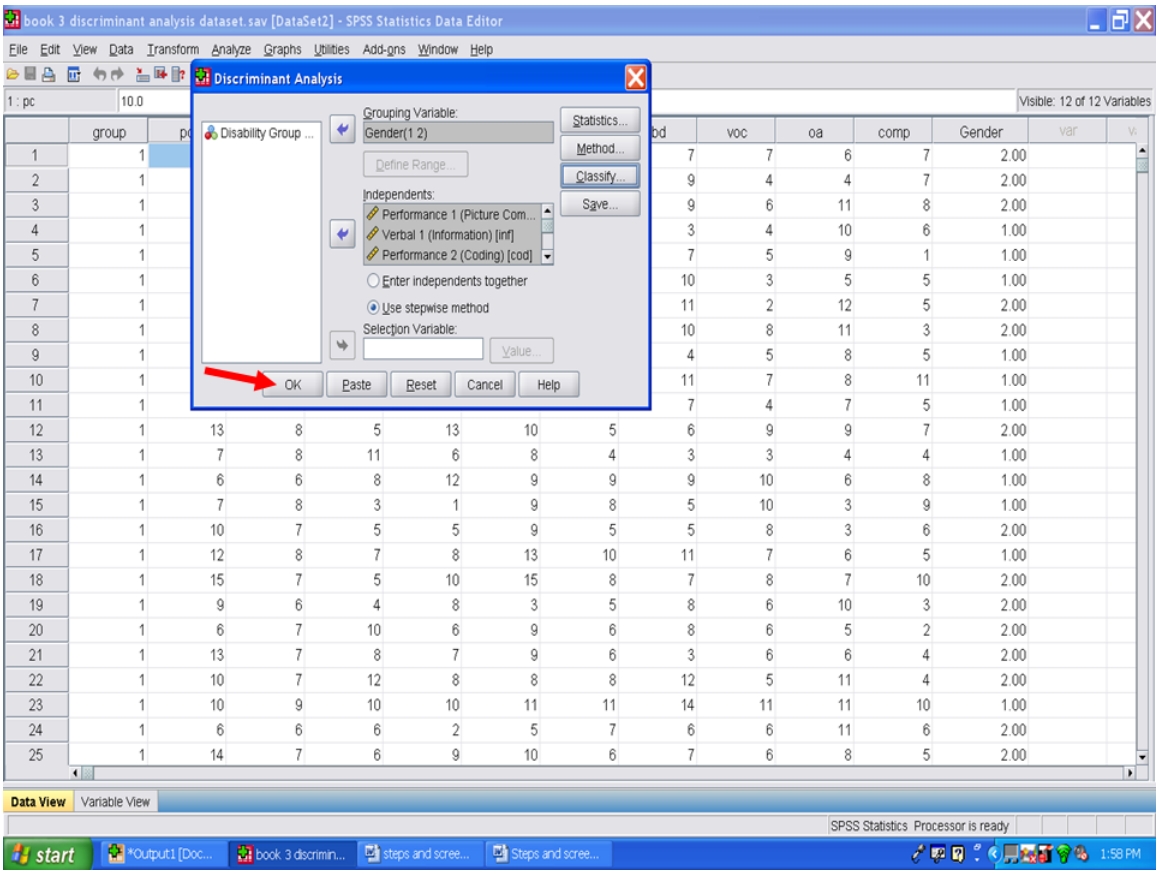
[8] http://cnx.org/content/m40736/latest/15.5.png/image



[9]

We are now ready to have SPSS calculate this procedure. Click on OK.

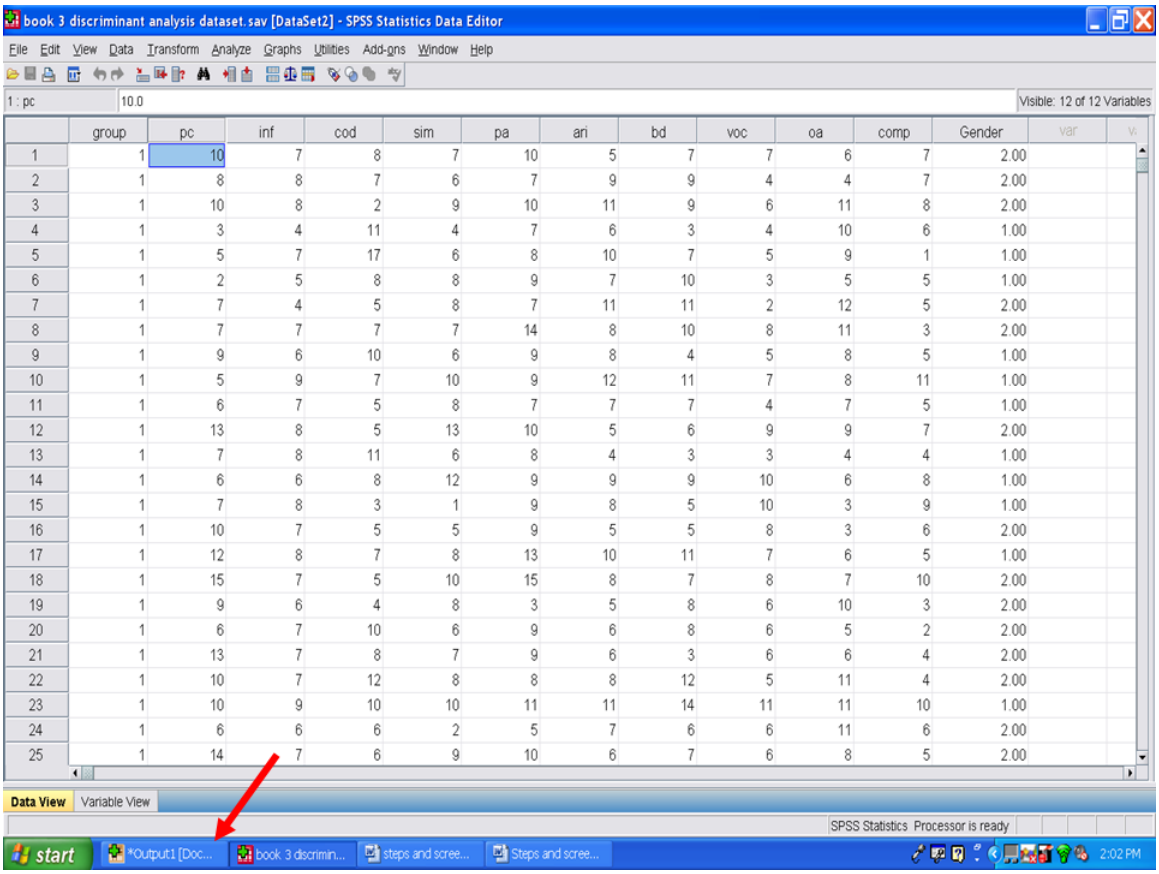
[9] http://cnx.org/content/m40736/latest/15.6.png/image



[10]

After clicking on OK, SPSS should send you to your output file. If not, click on the Output icon at the bottom of your screen.

[10] http://cnx.org/content/m40736/latest/15.7.png/image

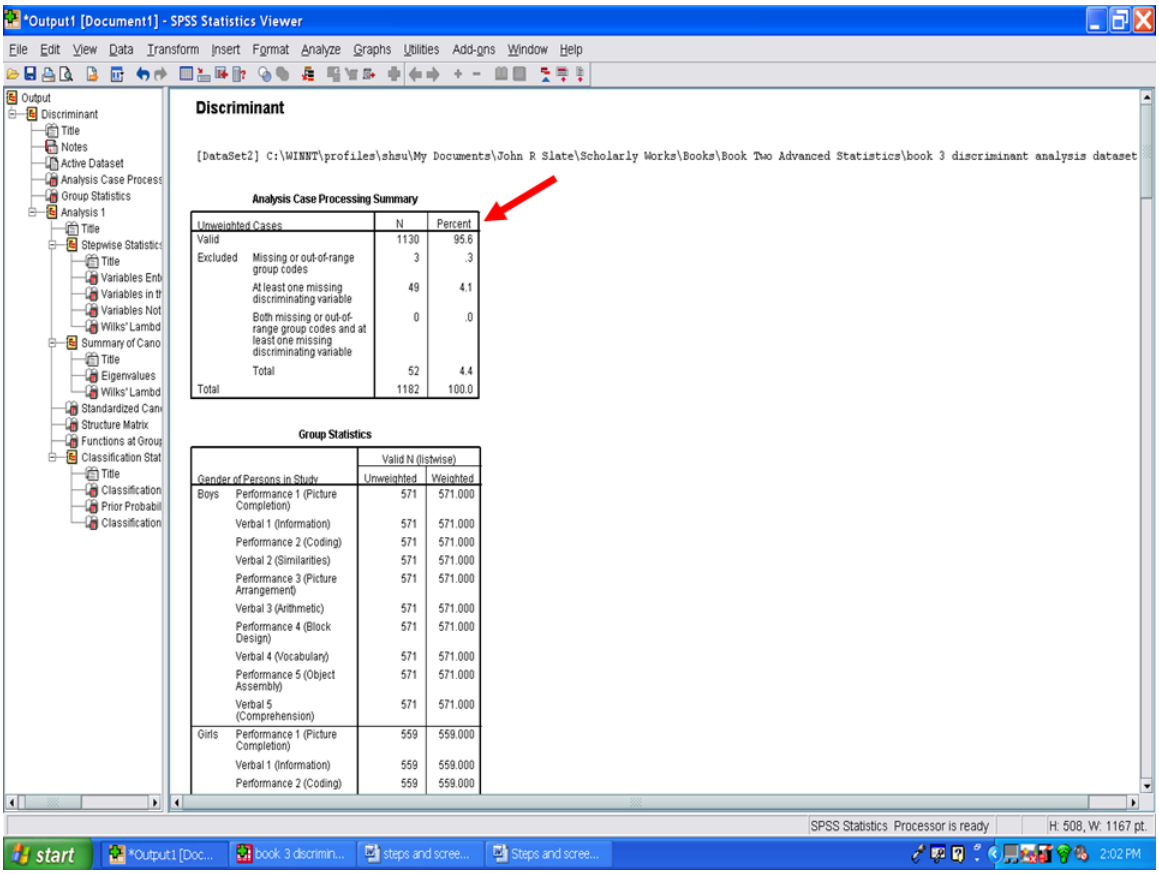


[11]

In the output screen, you should have the following output.

The first table shows you how many of your participants' data were used in the analysis. In this example, we had a total of 52 participants whose data were not analyzed.

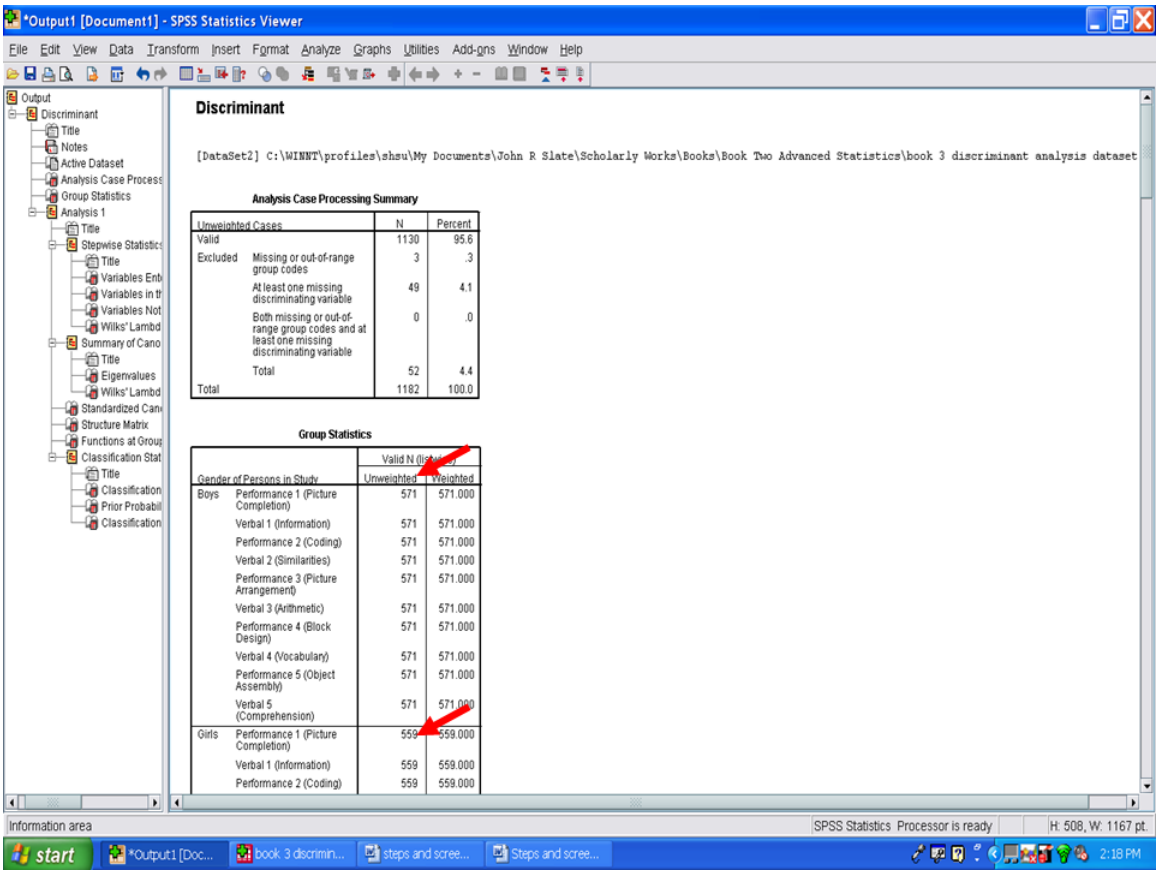
[11] http://cnx.org/content/m40736/latest/15.8.png/image



12

The second table in the output tells us we have 571 boys in group one and 559 girls in group two.

[12] http://cnx.org/content/m40736/latest/15.9.png/image



[13]

[13] http://cnx.org/content/m40736/latest/15.10.png/image

Chapter 7

Discriminant Analysis: Part III[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

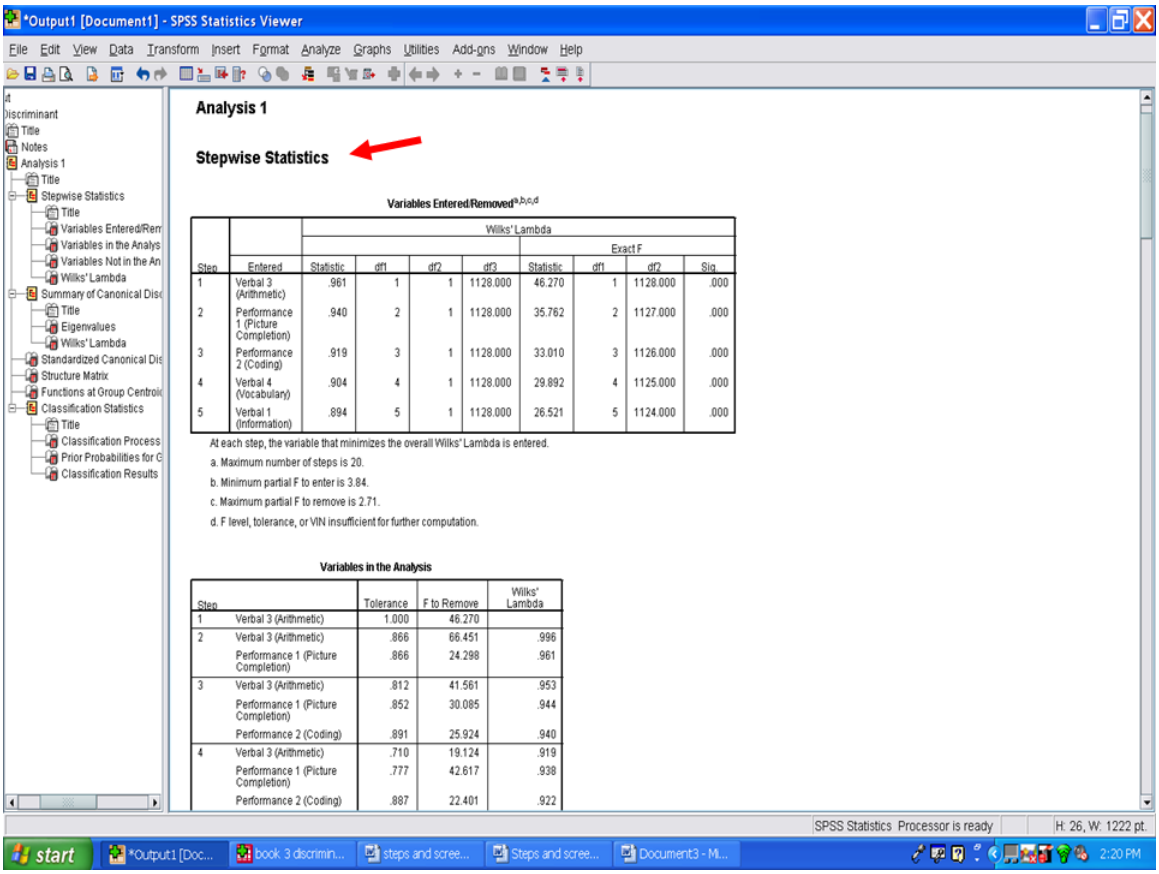
About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

The third table in the SPSS output for this canonical discriminant analysis is the Stepwise Statistics table. In this example, five of the 10 variables differentiated boys from girls.

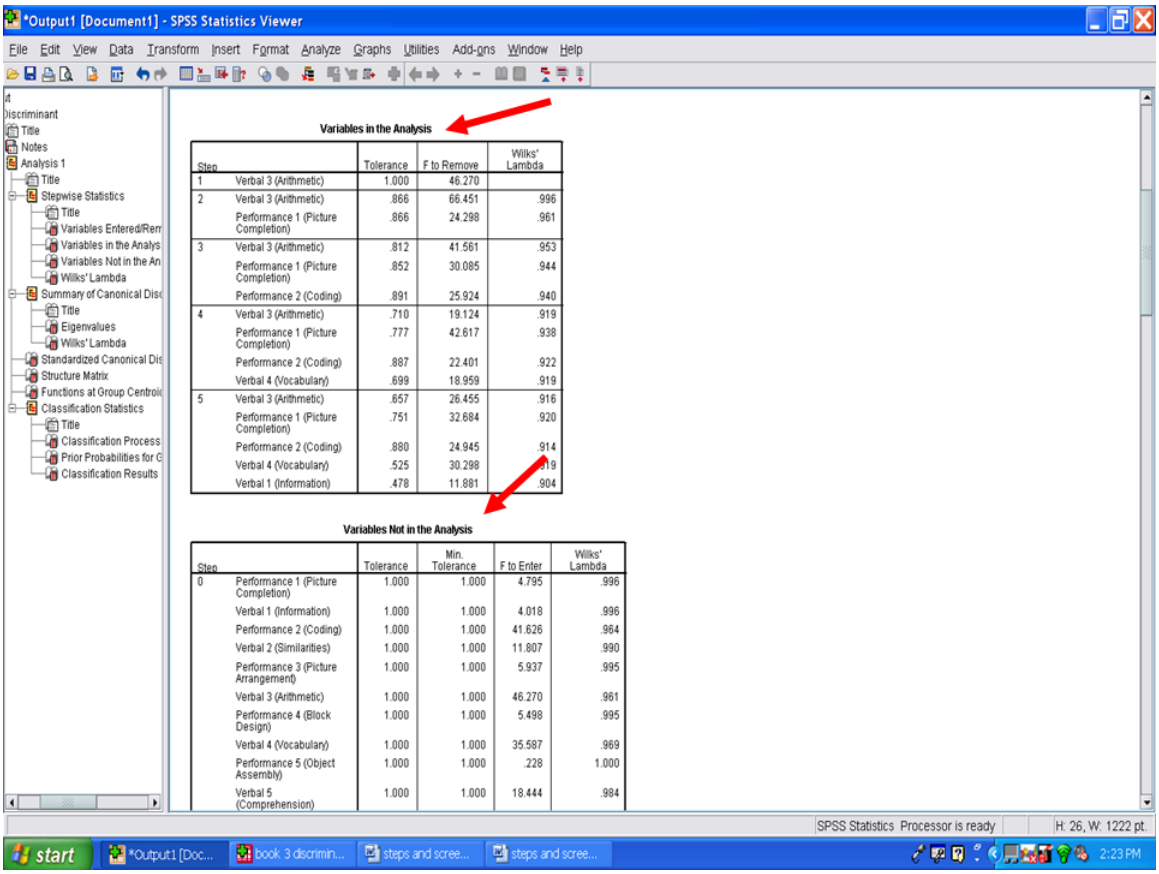
[1] This content is available online at <http://cnx.org/content/m40737/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40737/latest/www.writingandstatisticalhelp



[4]

The next two tables indicate which variables remained in the discriminant analysis and which variables were not utilized. The variables that were not utilized did not contribute statistically significantly to differentiating boys from girls. The process of identifying variables in the analysis and variables not in the analysis is depicted for each of the five steps.

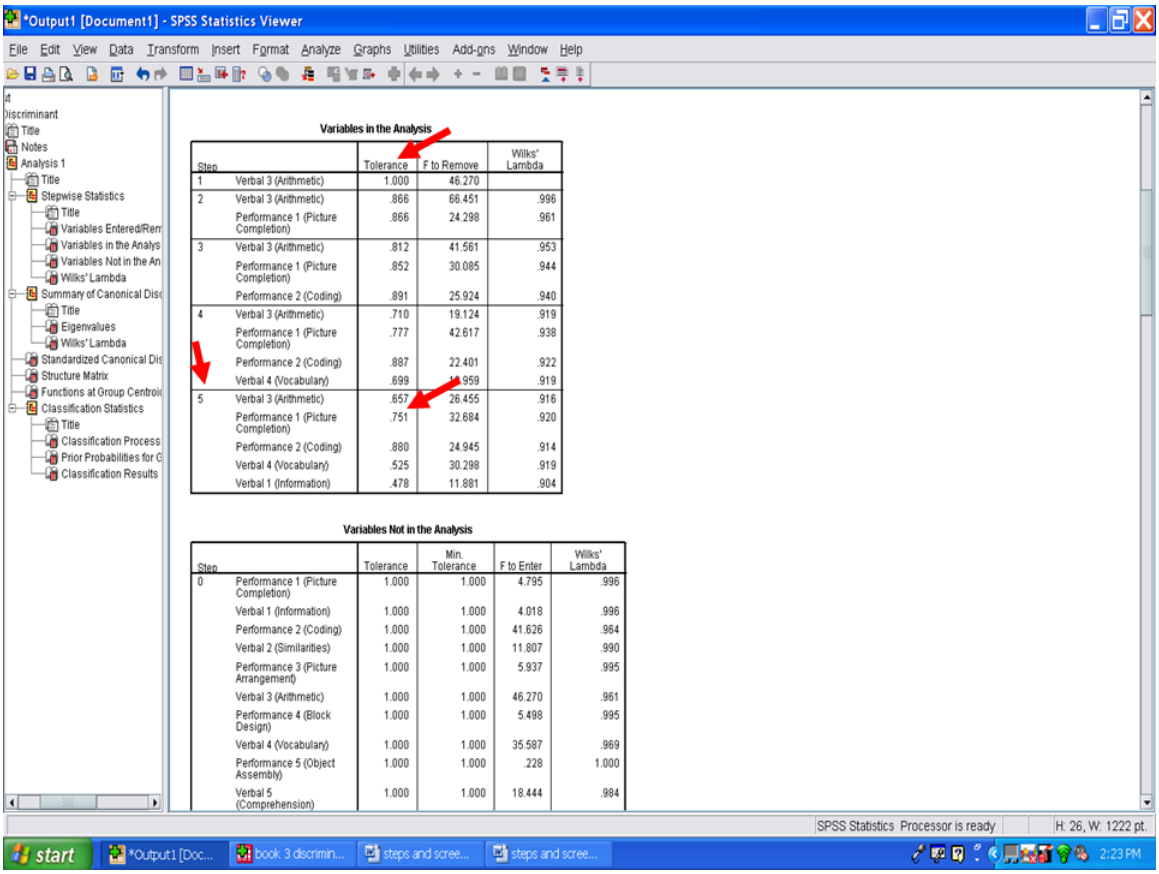
[4] http://cnx.org/content/m40737/latest/16.1.png/image



[5]

If you recall from the beginning of this chapter, one of the assumptions underlying use of a discriminant analysis procedure is tolerance values. The closer to 0 the tolerance values are, the more likely it is that the matrix is ill-conditioned and that overlap exists among the variables in the equation. In Step 5 below, you will note that all of the tolerance values are above 0, ranging from a low of .478 for Verbal 1 (Information) to a high of .880 for Performance 2 (Coding). Therefore, this assumption has not been violated.

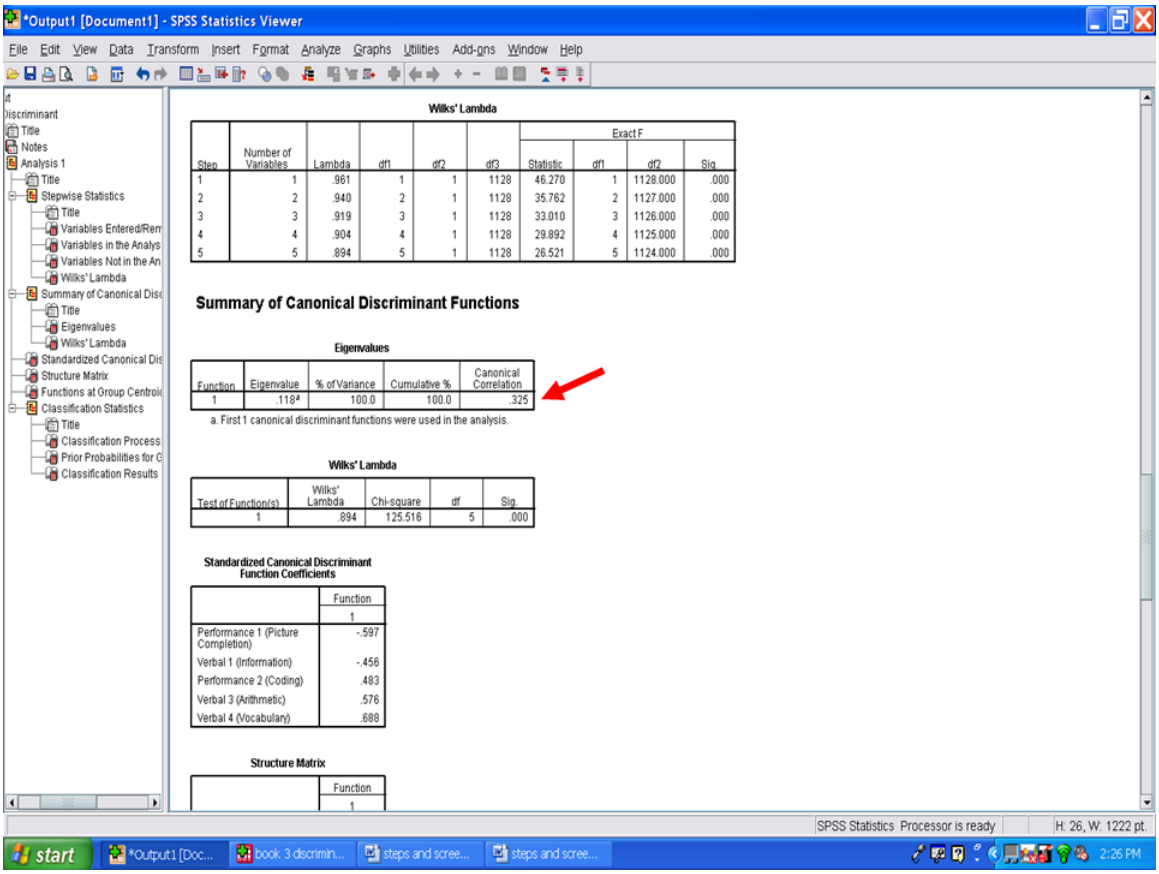
[5] http://cnx.org/content/m40737/latest/16.2.png/image



[6]

Under the Wilks' lambda table is a Summary of Canonical Discriminant Functions table labeled Eigenvalues. In this table is the canonical correlation, or *Rc*, which in this example is .325. This *Rc* value reflects the correlation between the groups and the discriminant function formed by the five variables that were statistically significant. The *Rc* may be interpreted as a Pearson correlation of the discriminant scores with the grouping variable. This information will be reported in your Results section.

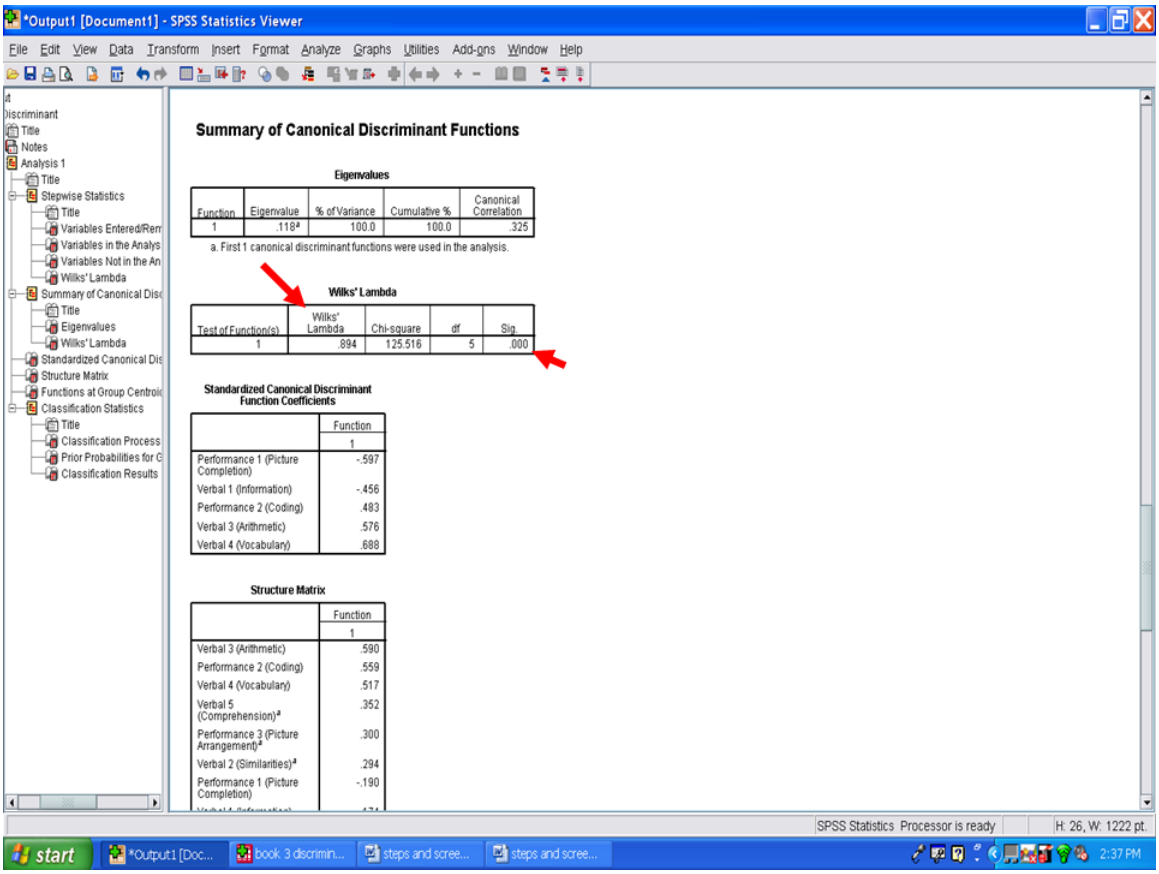
[6] http://cnx.org/content/m40737/latest/16.3.png/image



[7]

In the table below the Eigenvalues one is a table labeled Wilks' Lambda. This measure is what is used to determine whether the discriminant function is statistically significant in differentiating group membership. In our example, the Wilks' Lambda of .894 is statistically significant at the .001 level. This result means that the discriminant function, consisting of five variables, statistically significantly differentiated the group of boys from the group of girls. This information will be reported in the Results section.

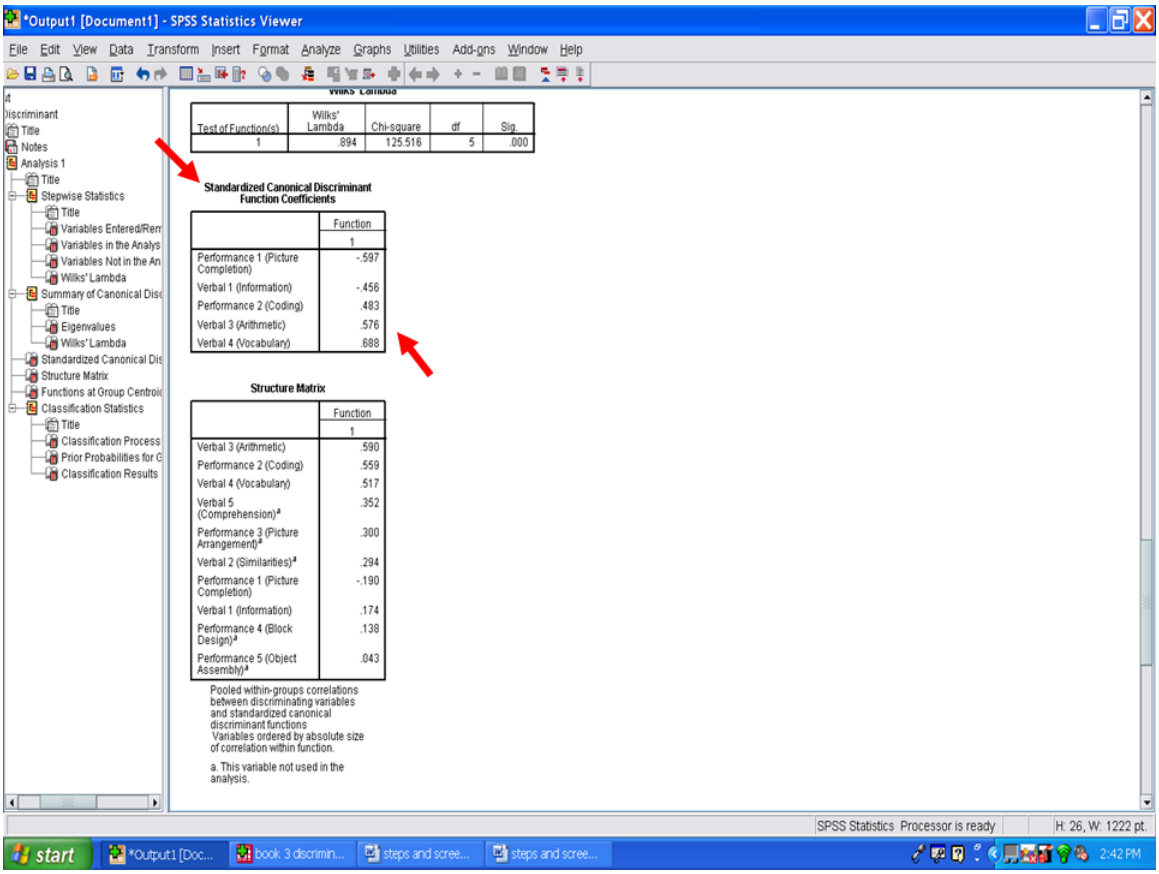
[7] http://cnx.org/content/m40737/latest/16.4.png/image



[8]

The table immediately below the Wilks' Lambda table is labeled Standardized Canonical Discriminant Function Coefficients. These values are also referred to as Standardized Discriminant Coefficients. Contained in this table are the values that depict the relative importance of each variable in differentiating group membership. These coefficients are comparable to Beta Weights in multiple regression. In this example, Verbal 4 (Vocabulary) is more important, with a coefficient of .688, than is Performance 2 (Coding), with a coefficient of .483, in differentiating boys from girls. These coefficients should be provided in your Results section.

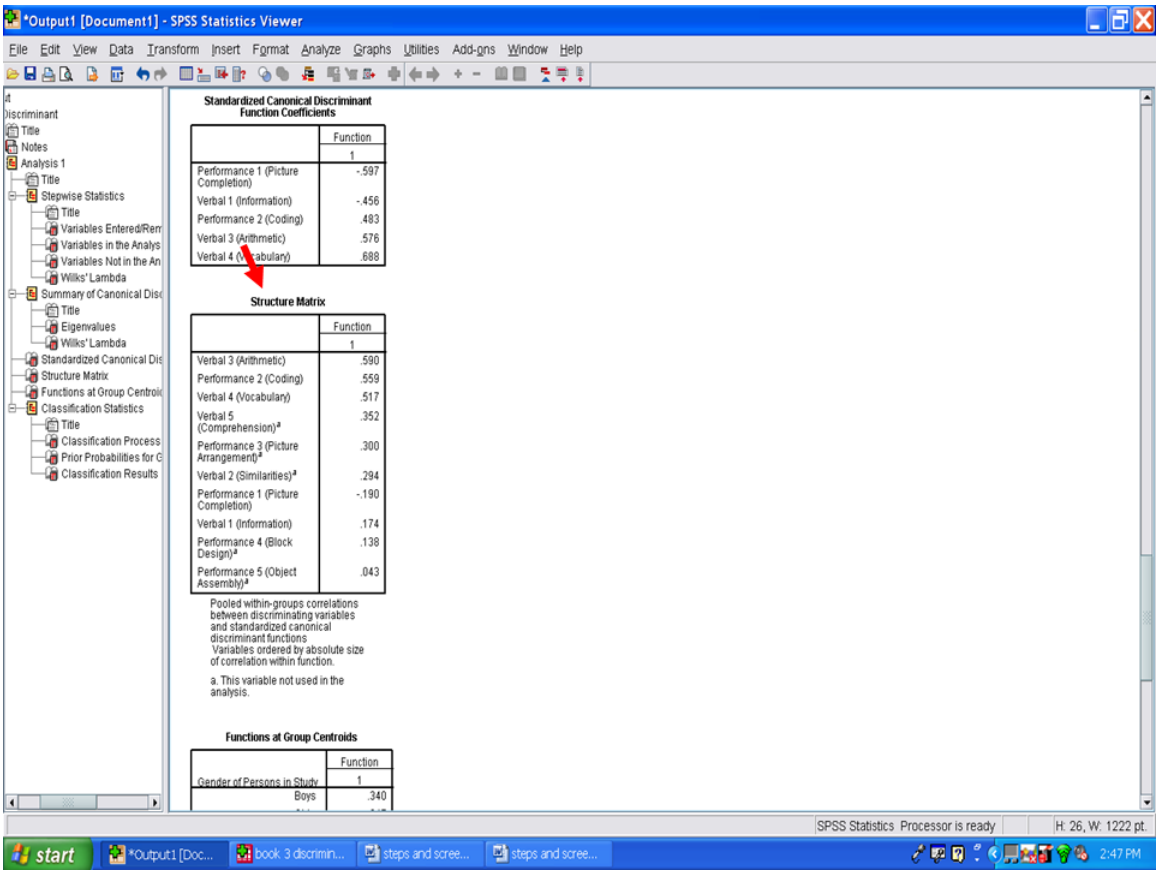
[8] http://cnx.org/content/m40737/latest/16.5.png/image



[9]

Next, we examine the table underneath the Standardized Canonical Discriminant Function Coefficients, the one labeled as Structure Matrix. Each of these values depicts the relationship of that variable with the discriminant function scores. Each value is analogous to a correlation coefficient. In the example below, Verbal 3 (Arithmetic) has a correlation of .59 with the discriminant function scores. Similarly, Performance 2 (Coding) and Verbal 4 (Vocabulary) have correlations of .559 and .517, respectively, with the discriminant function scores. These values may be interpreted as factor loadings in assigning a name to the discriminant function. As such, the first three variables just mentioned contribute more to the discriminant function than does the variable of Performance 5 (Object Assembly).

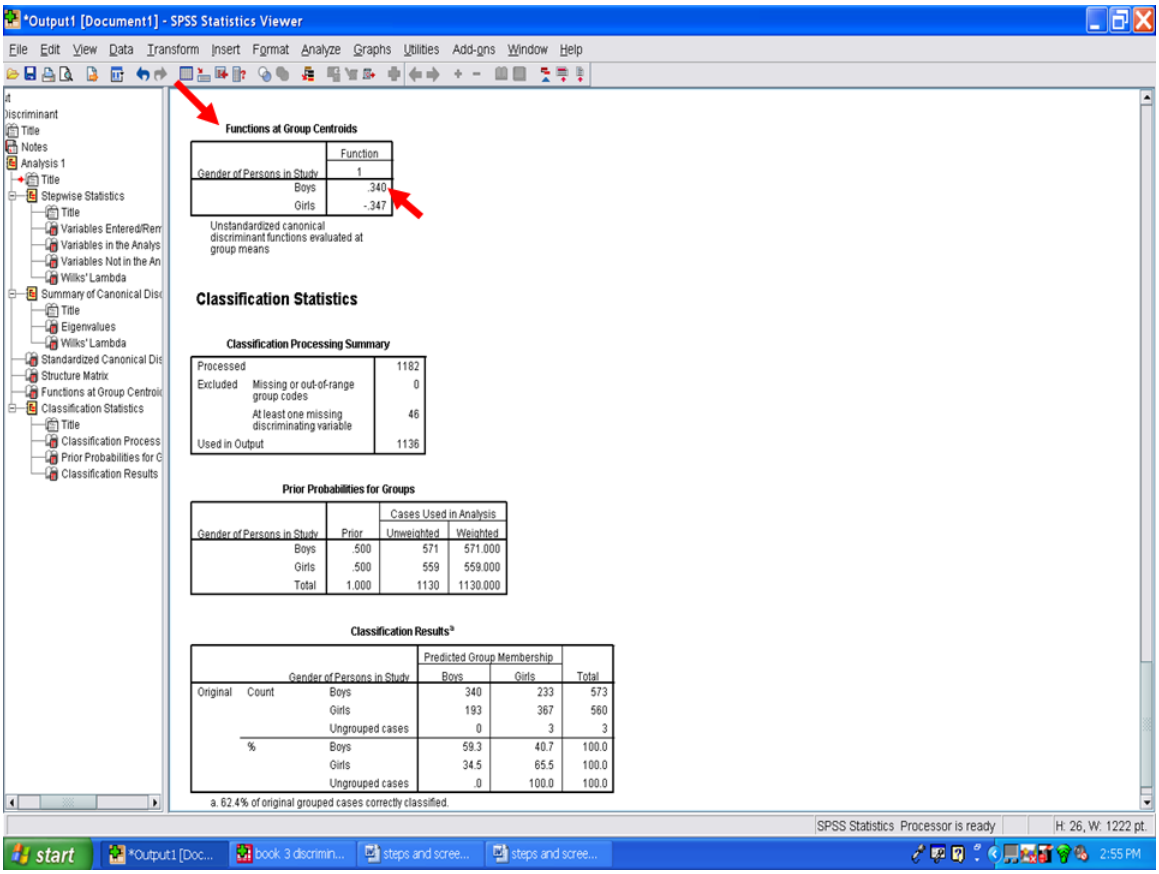
[9] http://cnx.org/content/m40737/latest/16.6.png/image



[10]

Under this table is a table labeled Functions at Group Centroids. These values are z-scores, with a *M* of 0 and a *SD* of 1. The group centroid indicates the most typical location of any person from a particular group. For this example, the group centroid for boys is .34 and the group centroid for girls is -.347. Using these values we can compare how far apart boys and girls are differentiated by the statistically significant discriminant function in our example. The further apart these two means or centroids are, the better the separation or differentiation between the two groups. The centroids should be reported in the Results section.

[10] http://cnx.org/content/m40737/latest/16.7.png/image



[11]

You have now successfully conducted a canonical discriminant analysis.

[11] http://cnx.org/content/m40737/latest/16.8.png/image

Chapter 8

Multiple Regression: Assumptions[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

In this set of steps, readers will learn how to conduct a multiple regression procedure. For detailed information regarding the assumptions underlying use of a multiple regression analysis, readers are referred to the Hyperstats Online Statistics Textbook at http://davidmlane.com/hyperstat/[4] ; to the *Electronic Statistics Textbook* (2011)

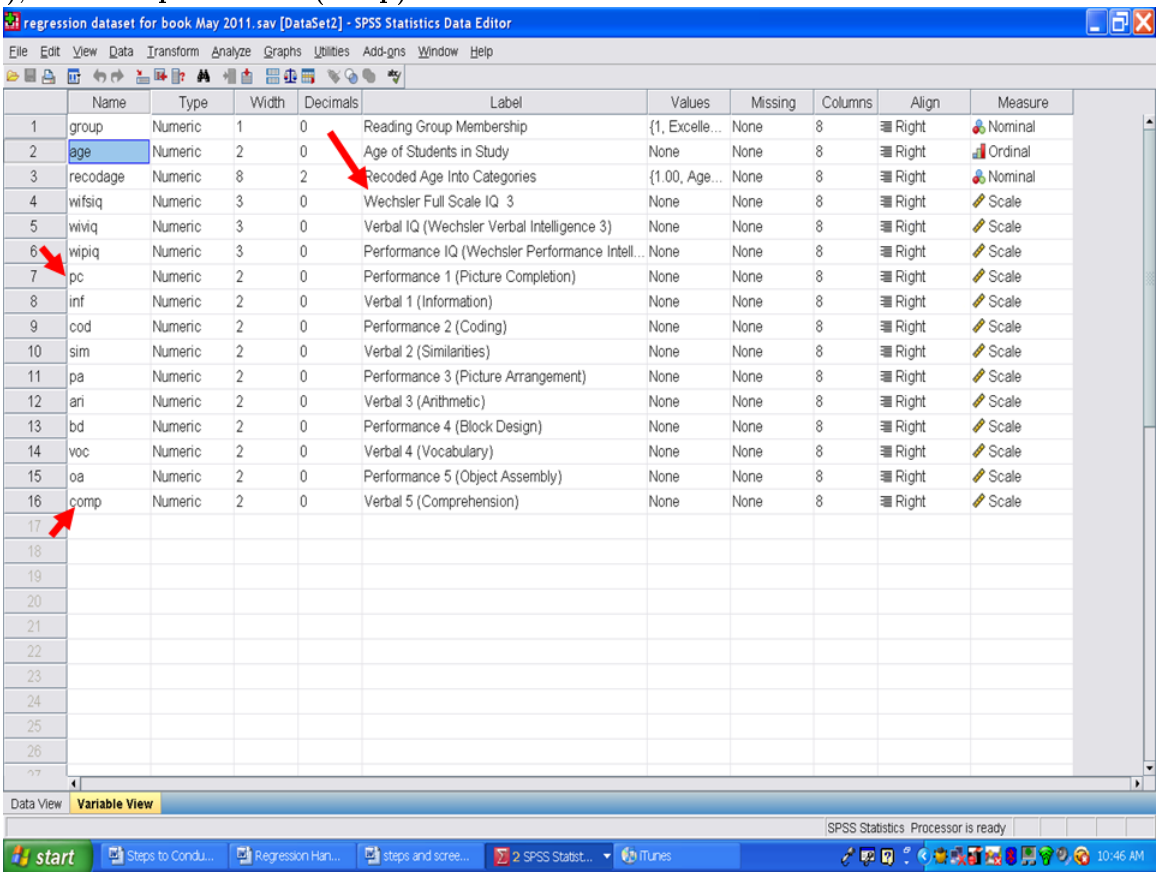
[1] This content is available online at <http://cnx.org/content/m40738/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40738/latest/www.writingandstatisticalhelp
[4] http://davidmlane.com/hyperstat/

at http://www.statsoft.com/textbook/[5] ; or to Andy Field's (2009) Discovering Statistics Using SPSS at http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967862&sr=1-1[6]

Research questions for which a multiple regression analysis is appropriate involve determining variables that predict a continuous variable. For example, if you want to predict the life expectancy of individuals and you have archival data available (e.g., health history, gender), then a multiple regression analysis procedure could be utilized. Such a procedure could identify specific variables that are predictive of a long or of a short life expectancy. As such, interventions could be developed and targeted toward the variables that were statistically significant predictors. Other sample research questions for which a multiple regression analysis might be appropriate: (a) What factors predict high scores on a scholastic aptitude measure?; and (b) What factors are predictive of high scores on a life satisfaction scale?

For the purposes of this chapter, our research question is: "What scholastic variables predict students' Full Scale IQ?"

Have your data set open in SPSS. In this dataset, we will determine what variables, if any, are predictive of students' Wechsler Full Scale IQ 3 (wifsiq). The 10 variables that we will use are: Picture Completion (pc), Information (inf), Coding (cod), Similarities (sim), Picture Arrangement (pa), Arithmetic (ari), Block Design (bd), Vocabulary (voc), Object Assembly (oa), and Comprehension (comp).



[7]

As with every statistical procedure, we need to check the underlying assumptions. One

[5] http://www.statsoft.com/textbook/

[6] http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304961

[7] http://cnx.org/content/m40738/latest/17.1.png/image

assumption involves the data being normally distributed. To check this assumption, we recommend that you calculate the standardized skewness coefficients and the standardized kurtosis coefficients, as discussed in other chapters.

*** Skewness [Note. Skewness refers to the extent to which the data are normally distributed around the mean. Skewed data involve having either mostly high scores with a few low ones or having mostly low scores with a few high ones.] Readers are referred to the following sources for a more detailed definition of skewness: http://www.statistics.com/index.php?page=glossary&term_id=356[8] and http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[9]**

To standardize the skewness value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the skewness value from the SPSS output and divide it by the Std. error of skewness. If the resulting calculation is within -3 to +3, then the skewness of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

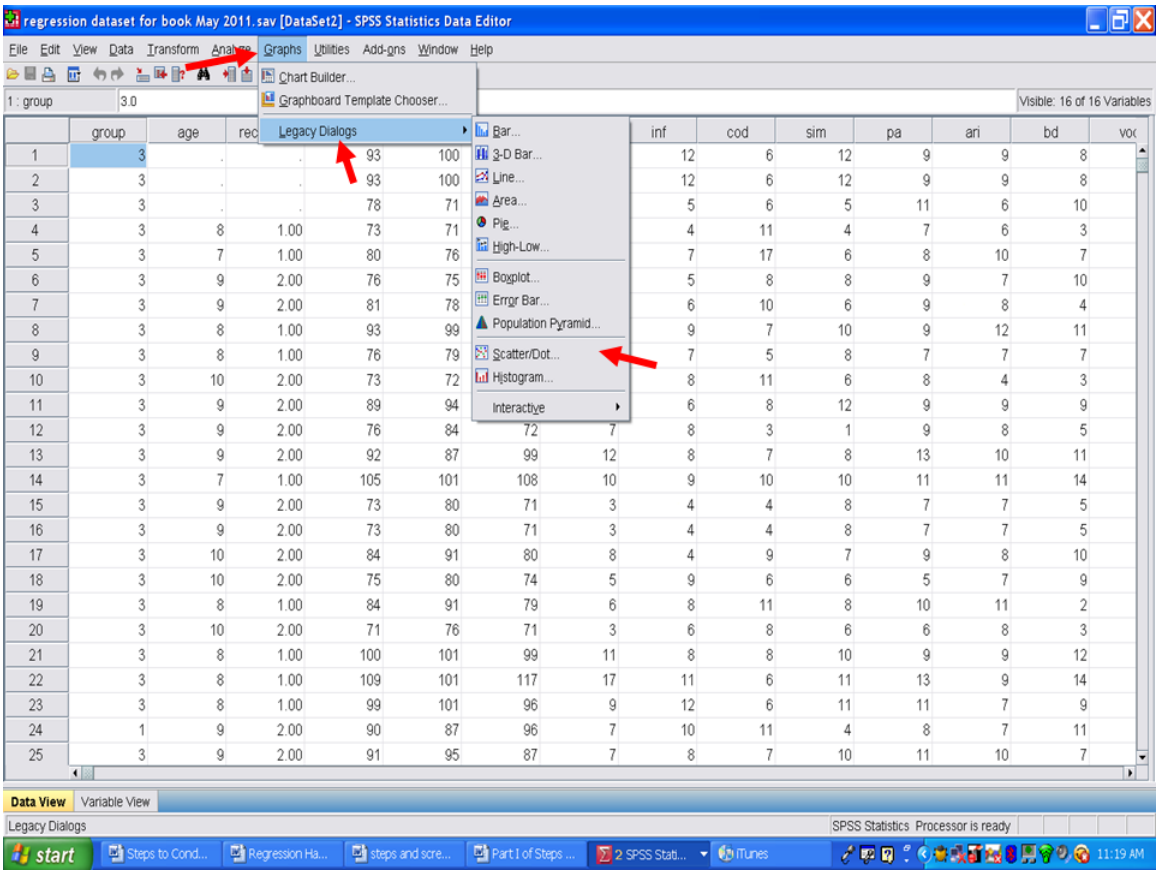
*** Kurtosis [Note. Kurtosis also refers to the extent to which the data are normally distributed around the mean. This time, the data are piled up higher than normal around the mean or piled up higher than normal at the ends of the distribution.] Readers are referred to the following sources for a more detailed definition of kurtosis: http://www.statistics.com/index.php?page=glossary&term_id=326[10] and http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb[11]**

To standardize the kurtosis value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the kurtosis value from the SPSS output and divide it by the Std. error of kurtosis. If the resulting calculation is within -3 to +3, then the kurtosis of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this +/-3 range, the dataset is not normally distributed.

Now that you have verified that your data are normally distributed, the extent to which linearity is present between each of the 10 independent variables listed above and the dependent variable of Full Scale IQ must be determined. For linearity, we will have SPSS conduct scatterplots for each IV and DV pair.

√ Graphs
√ Legacy Dialogs
√ Scatter/Dot

[8] http://www.statistics.com/index.php?page=glossary&term_id=356
[9] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb
[10] http://www.statistics.com/index.php?page=glossary&term_id=326
[11] http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb

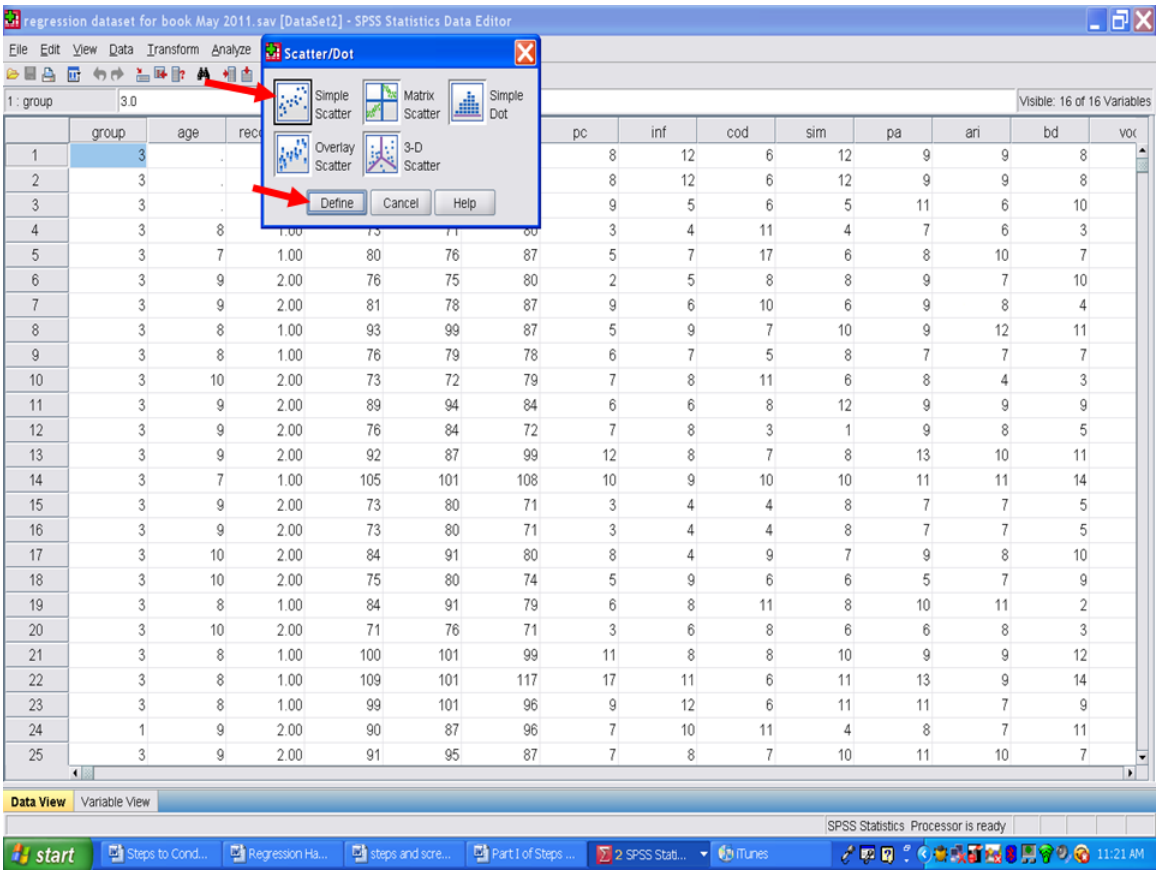


12

After clicking on Scatter/Dot, the following screen will appear. The Simple Scatter icon should be highlighted. If not, click on it.

√ Click on Define

[12] http://cnx.org/content/m40738/latest/17.2.png/image

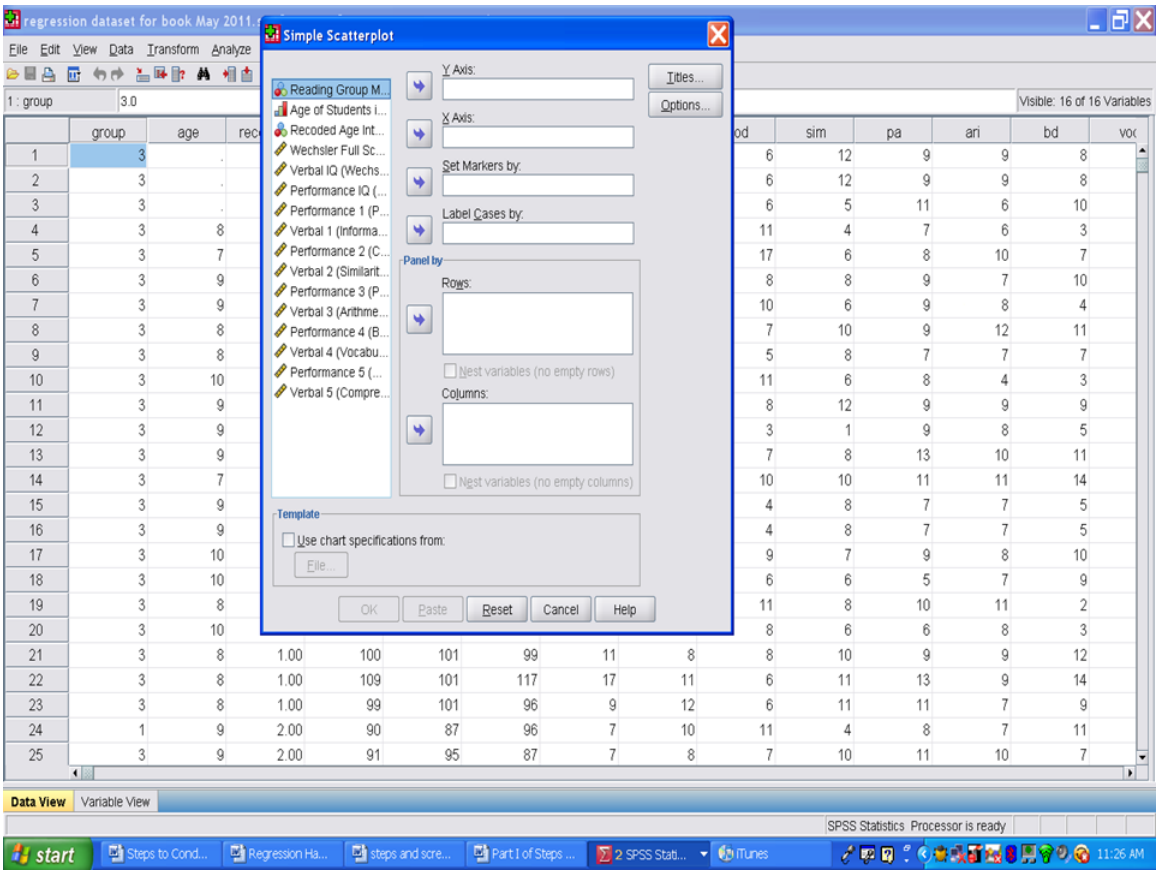


	group	age	rec				pc	inf	cod	sim	pa	ari	bd	voc
1	3	.					8	12	6	12	9	9	8	
2	3	.					8	12	6	12	9	9	8	
3	3	.					9	5	6	5	11	6	10	
4	3	8	1.00	73	71	80	3	4	11	4	7	6	3	
5	3	7	1.00	80	76	87	5	7	17	6	8	10	7	
6	3	9	2.00	76	75	80	2	5	8	8	9	7	10	
7	3	9	2.00	81	78	87	9	6	10	6	9	8	4	
8	3	8	1.00	93	99	87	5	9	7	10	9	12	11	
9	3	8	1.00	76	79	78	6	7	5	8	7	7	7	
10	3	10	2.00	73	72	79	7	8	11	6	8	4	3	
11	3	9	2.00	89	94	84	6	6	8	12	9	9	9	
12	3	9	2.00	76	84	72	7	8	3	1	9	8	5	
13	3	9	2.00	92	87	99	12	8	7	8	13	10	11	
14	3	7	1.00	105	101	108	10	9	10	10	11	11	14	
15	3	9	2.00	73	80	71	3	4	4	8	7	7	5	
16	3	9	2.00	73	80	71	3	4	4	8	7	7	5	
17	3	10	2.00	84	91	80	8	4	9	7	9	8	10	
18	3	10	2.00	75	80	74	5	9	6	6	5	7	9	
19	3	8	1.00	84	91	79	6	8	11	8	10	11	2	
20	3	10	2.00	71	76	71	3	6	8	6	6	8	3	
21	3	8	1.00	100	101	99	11	8	8	10	9	9	12	
22	3	8	1.00	109	101	117	17	11	6	11	13	9	14	
23	3	8	1.00	99	101	96	9	12	6	11	11	7	9	
24	1	9	2.00	90	87	96	7	10	11	4	8	7	11	
25	3	9	2.00	91	95	87	7	8	7	10	11	10	7	

[13]

The following screen should now be present.

[13] http://cnx.org/content/m40738/latest/17.3.png/image

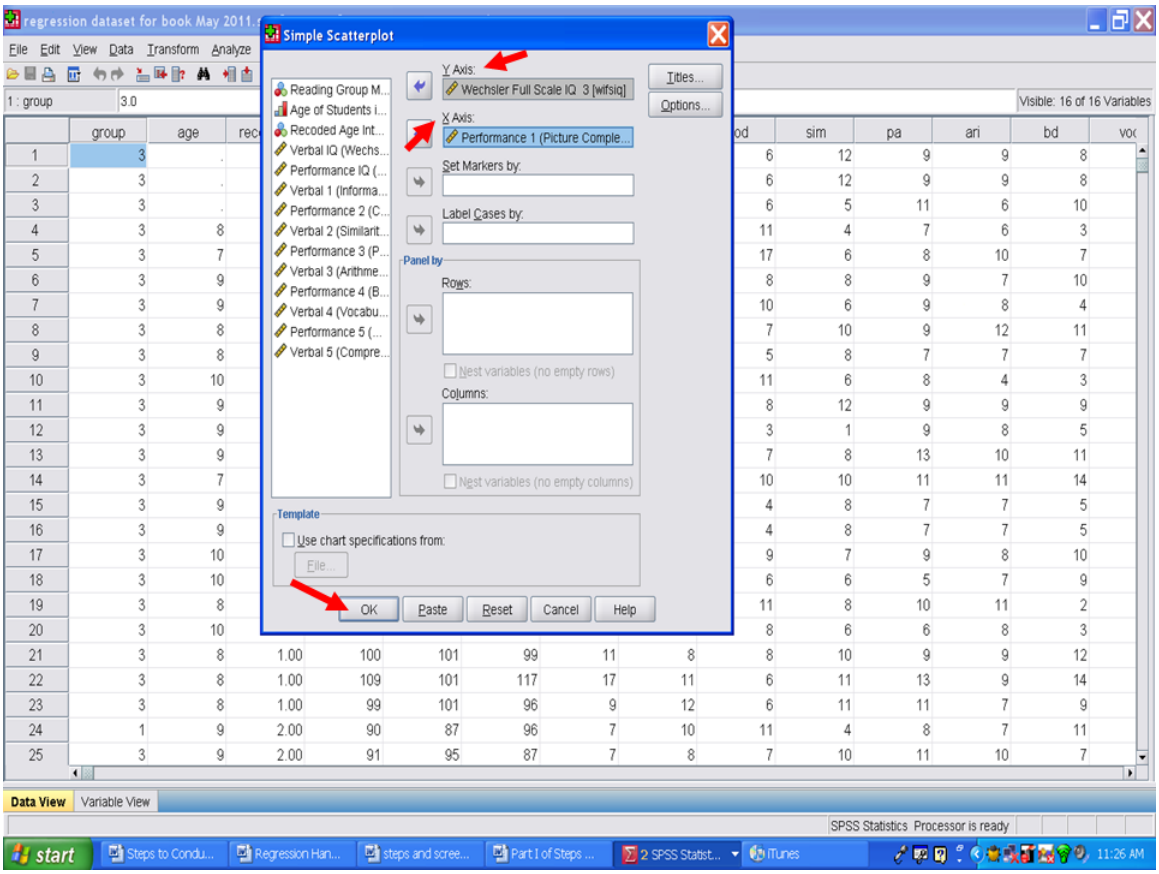


14

√ **Drag one of the two variables of interest to the first box (Y axis) on the right hand side and the other variable of interest to the second box (X axis) on the right hand side. It does not matter which variable goes in the X or Y axis because your scatterplot results will be the same. For our purposes, we will place the variable we are trying to predict, Wechsler Full Scale IQ 3, in the Y Axis box and one of the variables (i.e., Performance 1) we will use to try to predict it.**

√ **Once you have a variable in each of the first two boxes, click on the OK tab on the bottom left hand corner of the screen.**

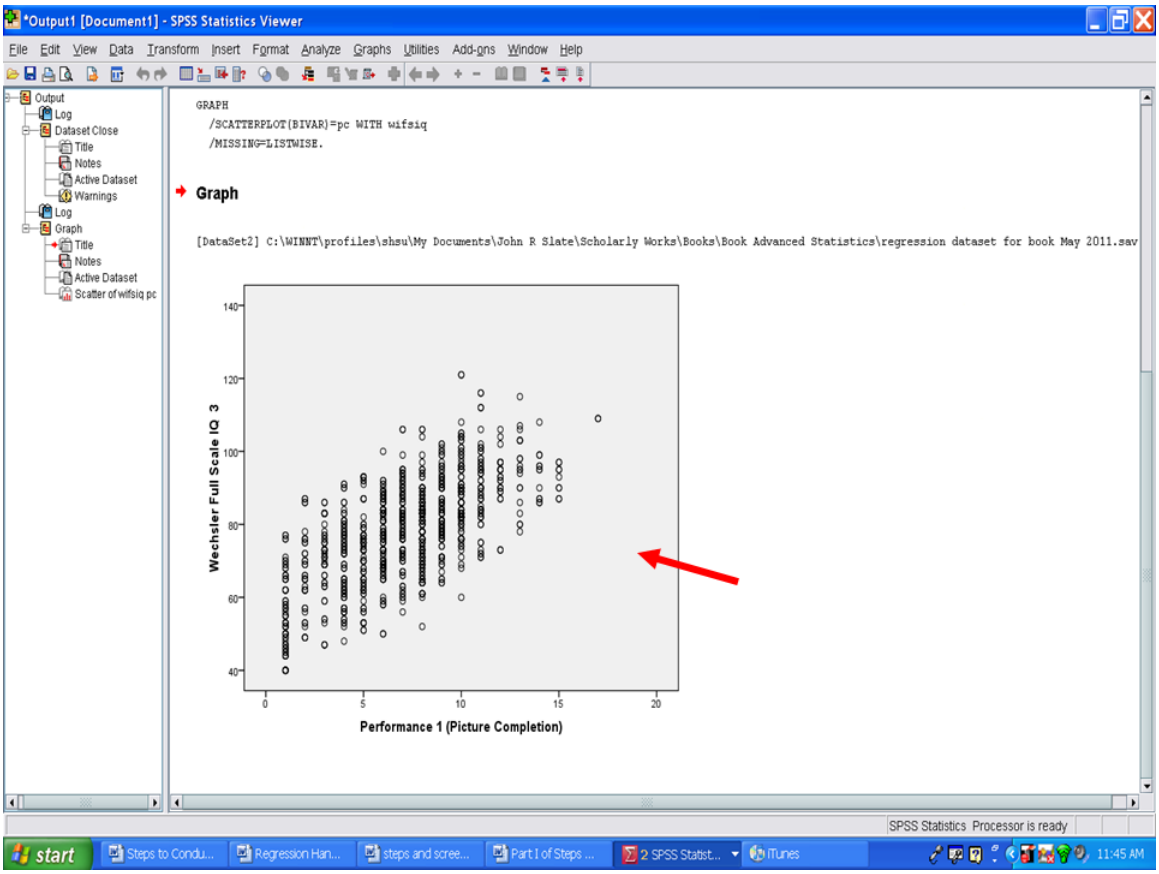
[14] http://cnx.org/content/m40738/latest/17.4.png/image



15

√ **Look at the scatterplot to determine whether a linear relationship is present. In the screenshot below, the relationship is very clearly linear.**

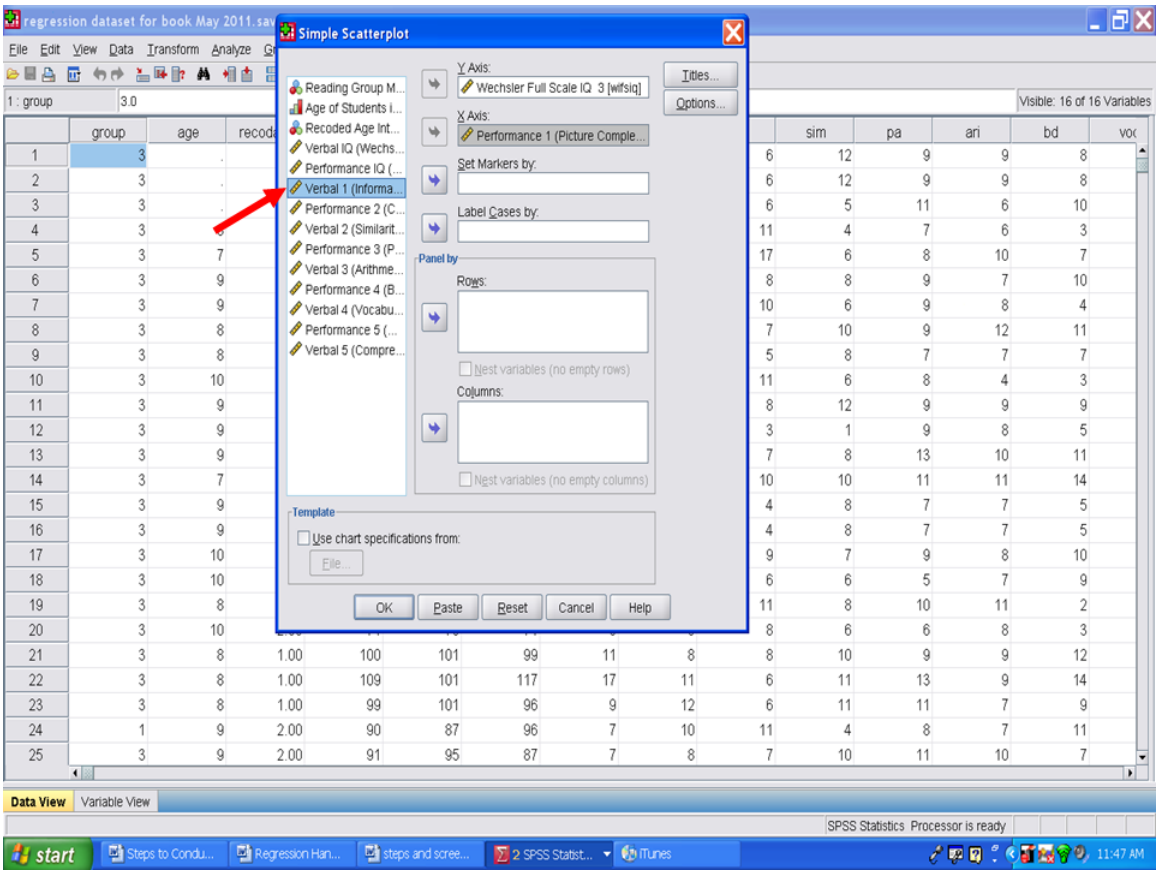
[15] http://cnx.org/content/m40738/latest/17.5.png/image



16

You will need to repeat this process, for this example, nine more times. Leave the dependent variable of Wechsler Full Scale IQ 3 in the Y Axis box and replace the variable in the X Axis box with the next variable (i.e., Verbal 1). Then click on OK.

[16] http://cnx.org/content/m40738/latest/17.6.png/image



17

After you have verified that linearity is present for each independent variable with the dependent variable, we will examine the extent to which multicollinearity is not present. Multicollinearity refers to having variables that are highly correlated with each other. When variables are highly correlated in a multiple regression analysis, the unique contribution of each variable in predicting the dependent variable is difficult to determine. The reason for this difficulty is that highly interrelated variables are being used to predict the same variance in the dependent variable. Researchers/statisticians disagree on the specific correlation value that must be present for multicollinearity to exist. Some persons contend that correlations above .70 are necessary whereas other persons contend that the correlations must be above .90 for multicollinearity to exist.

If multicollinearity is present, you can leave it as it is, and have SPSS calculate the multiple regression. Multicollinearity influences the results regarding each predictor's unique contribution. If your interest is in the overall or combined effect of the statistically significant predictors, then multicollinearity is not an issue. Other choices would be to remove one or more of the highly correlated variables from the regression analysis or to create an aggregate or composite of the highly correlated variables.

The choice that we recommend is to have SPSS calculate multicollinearity when the multiple regression analysis is calculated. More on this later.

[17] http://cnx.org/content/m40738/latest/17.7.png/image

Chapter 9

Multiple Regression: Part I[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

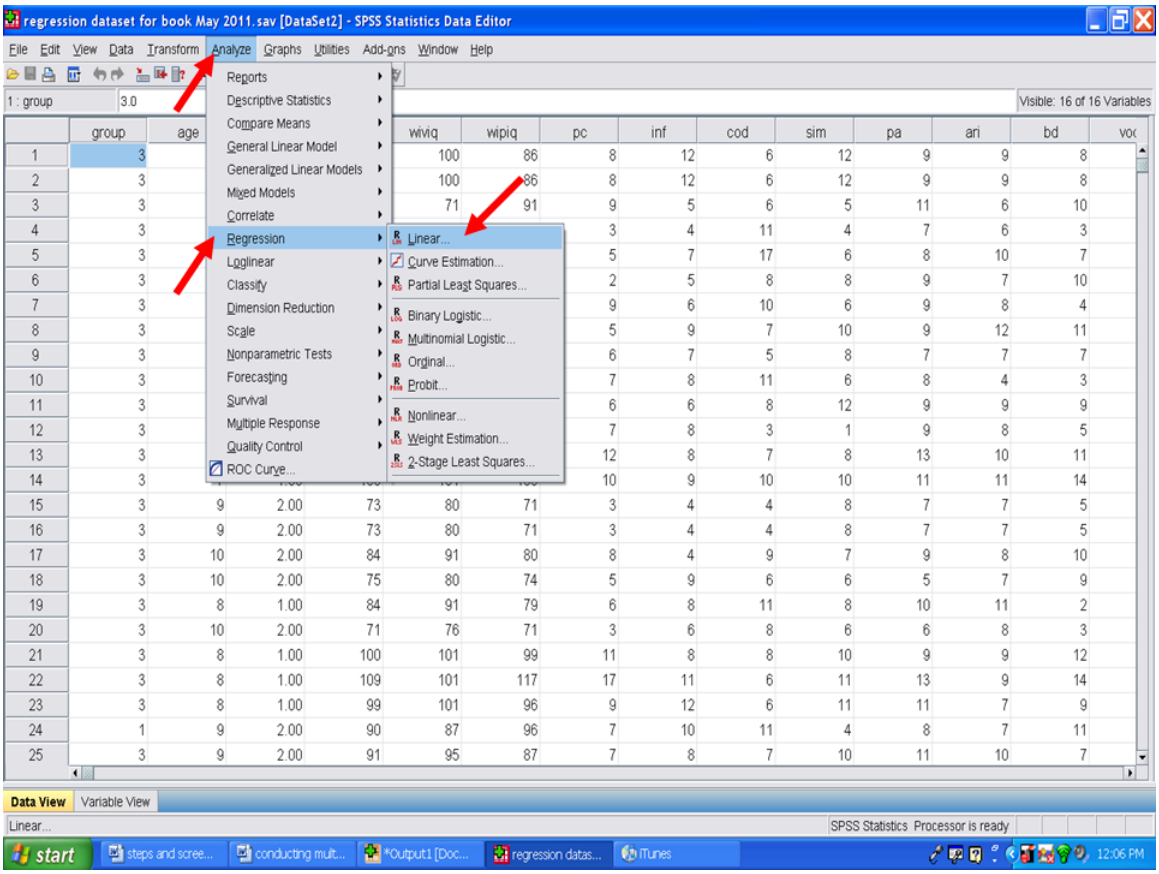
John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

From the data screen, click on

√ **Analyze**
√ **Regression**
√ **Linear**

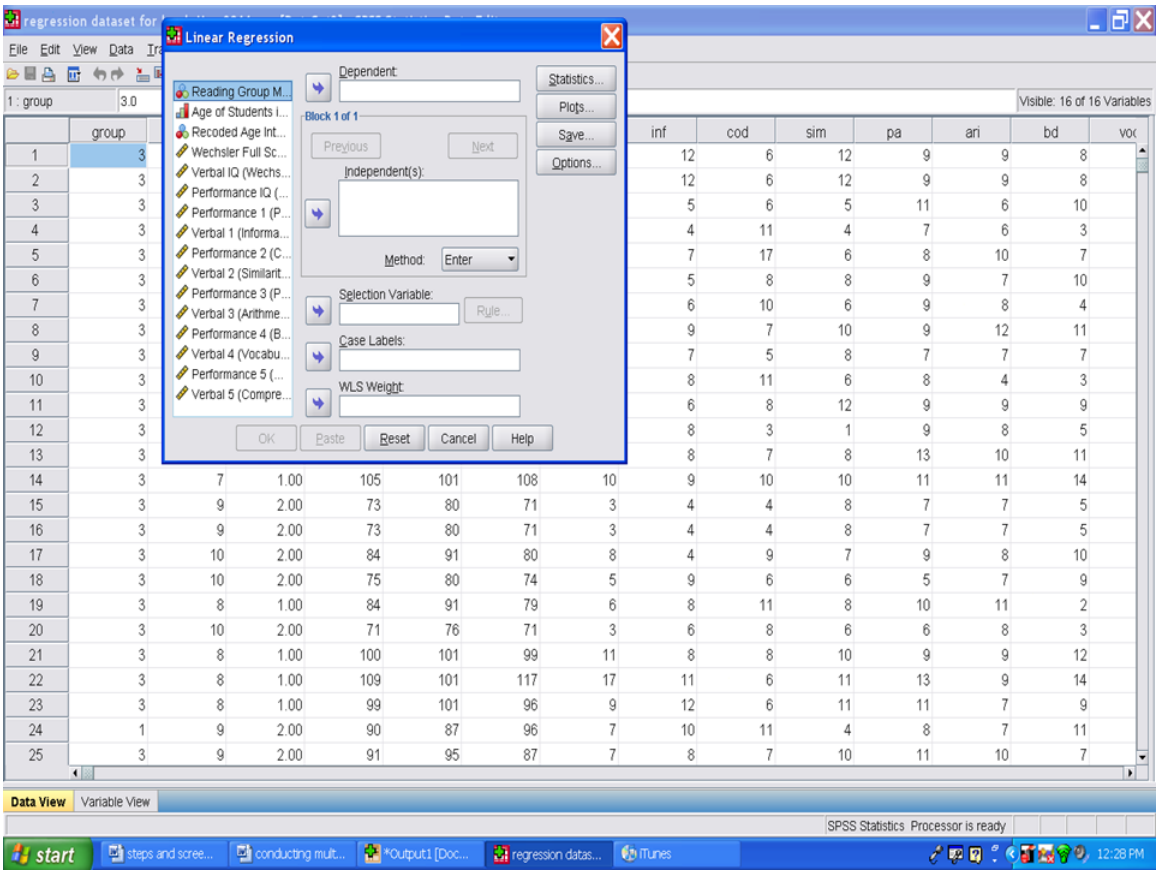
[1] This content is available online at <http://cnx.org/content/m40739/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40739/latest/www.writingandstatisticalhelp



[4]

Your screen should now look like the one below. We will send the dependent variable, the one we are trying to predict, to the Dependent box. Click on Wechsler Full Scale IQ 3 and send it to the Dependent box.

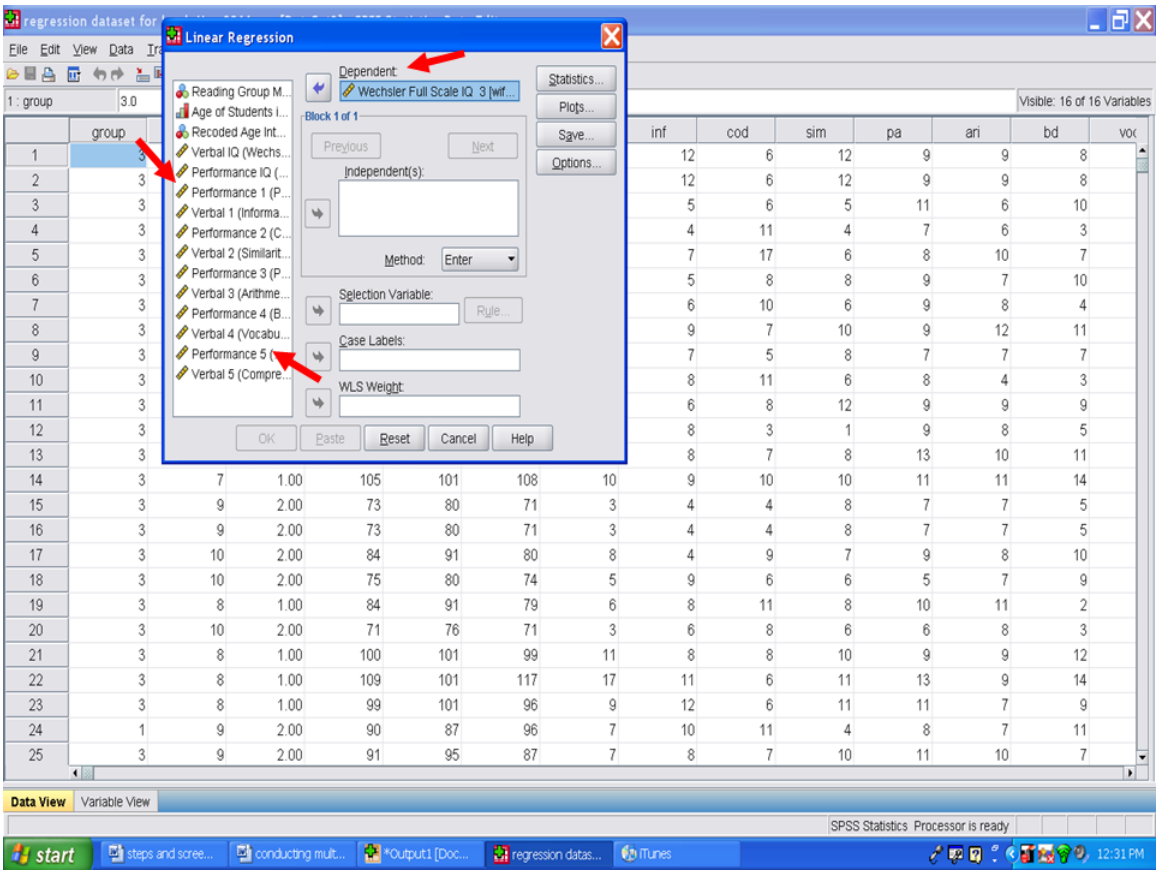
[4] http://cnx.org/content/m40739/latest/18.1.png/image



[5]

Your screen should now look like the following one. Next we will place all of the independent variables in the Independent(s) box. I this example, remember that we are using Performance 1 through Verbal 5.

[5] http://cnx.org/content/m40739/latest/18.2.png/image

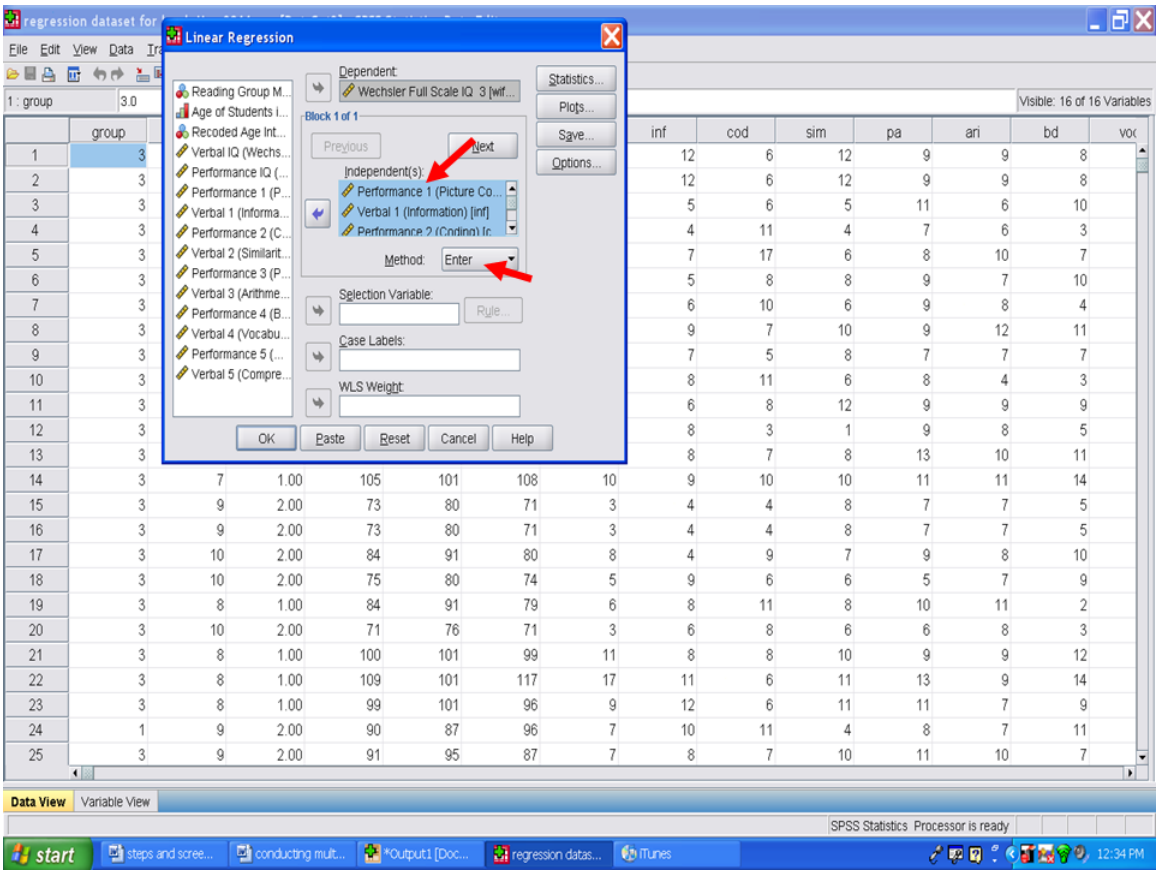


[6]

After you have placed all 10 of these variables in the Independent(s) box, your screen will look like the following one.

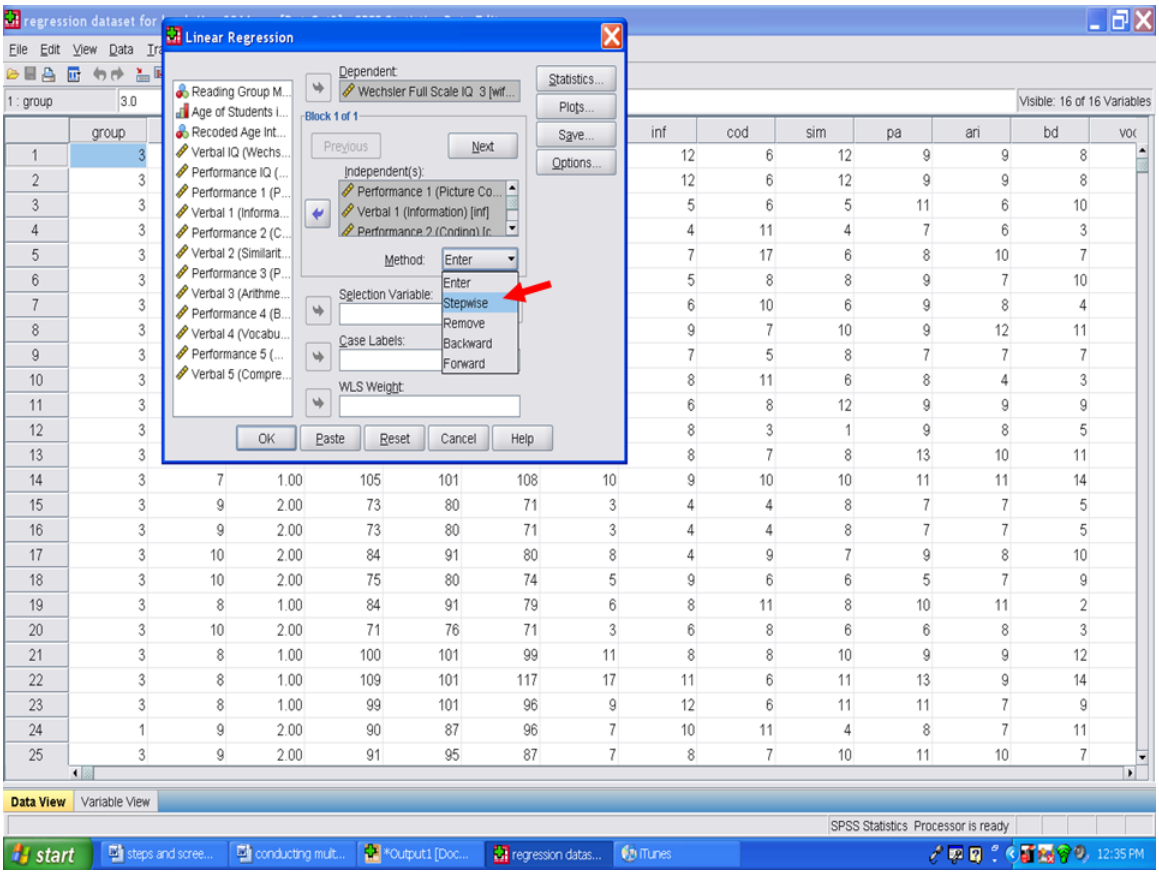
Next we will click on Method and change it from the default of Enter to Stepwise.

[6] http://cnx.org/content/m40739/latest/18.3.png/image



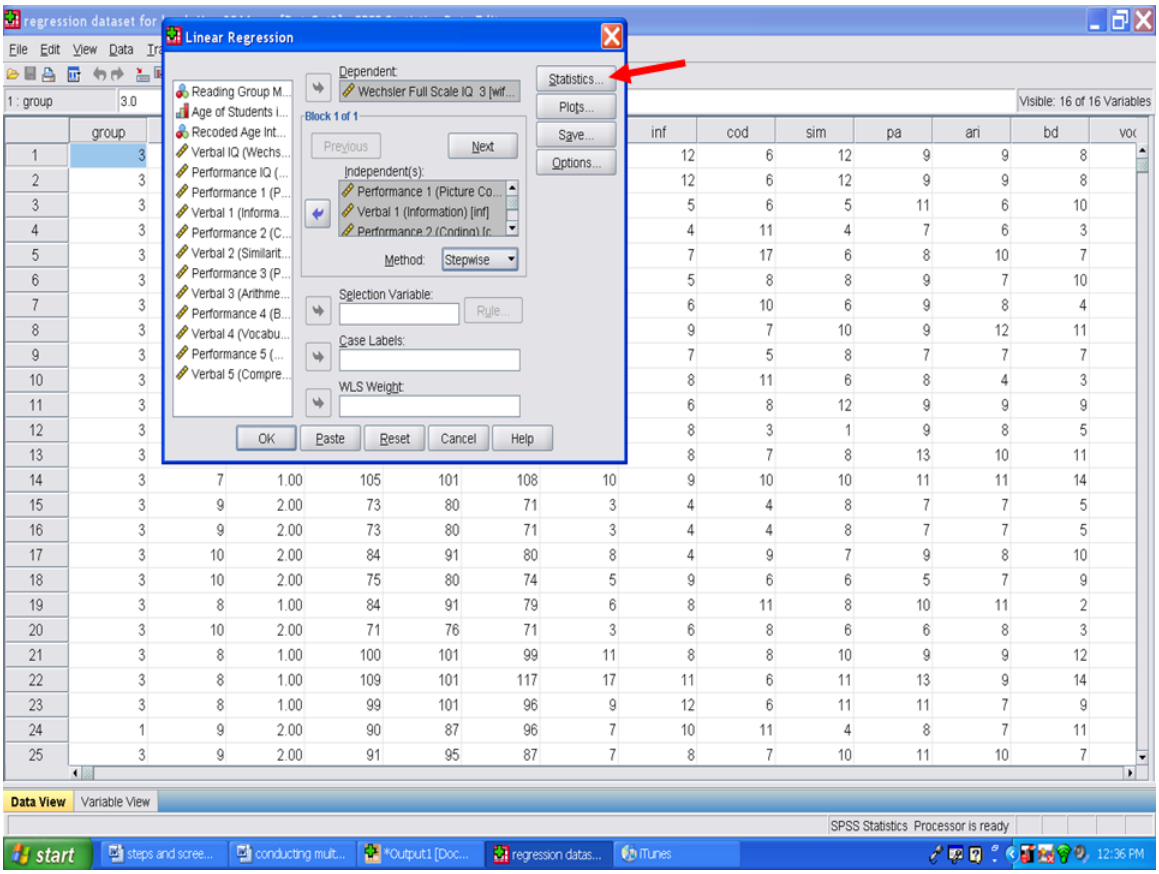
Click on Stepwise.

[7] http://cnx.org/content/m40739/latest/18.4.png/image



After clicking on Stepwise, your screen will look like the one below.
Click on Statistics

[8] http://cnx.org/content/m40739/latest/18.5.png/image



9

After clicking on Statistics, you will now

√ Click on Estimates (default already checked)

√ Model fit (default already checked)

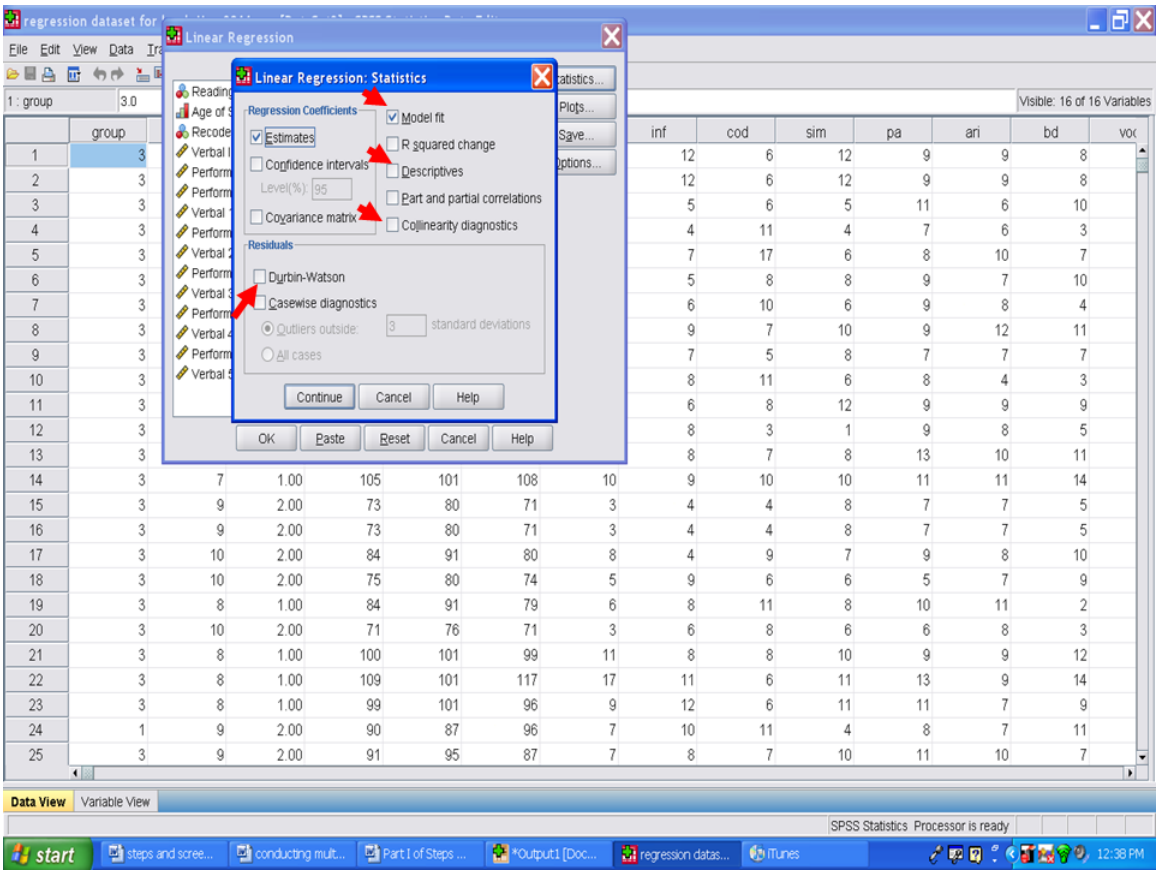
√ R squared change

√ Descriptives

√ Collinearity Diagnostics

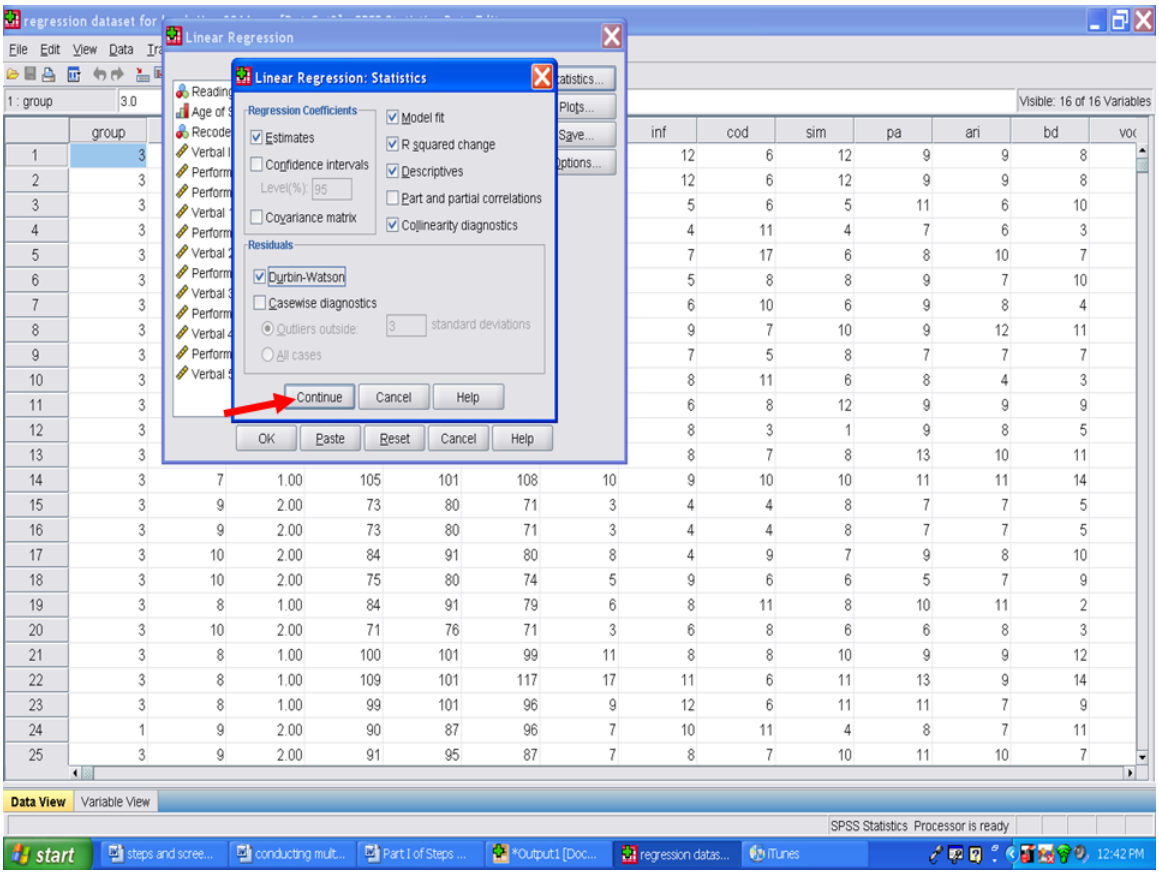
√ Durbin-Watson

[9] http://cnx.org/content/m40739/latest/18.6.png/image



Your screen should like the one below. Now click on Continue.

[10] http://cnx.org/content/m40739/latest/18.7.png/image



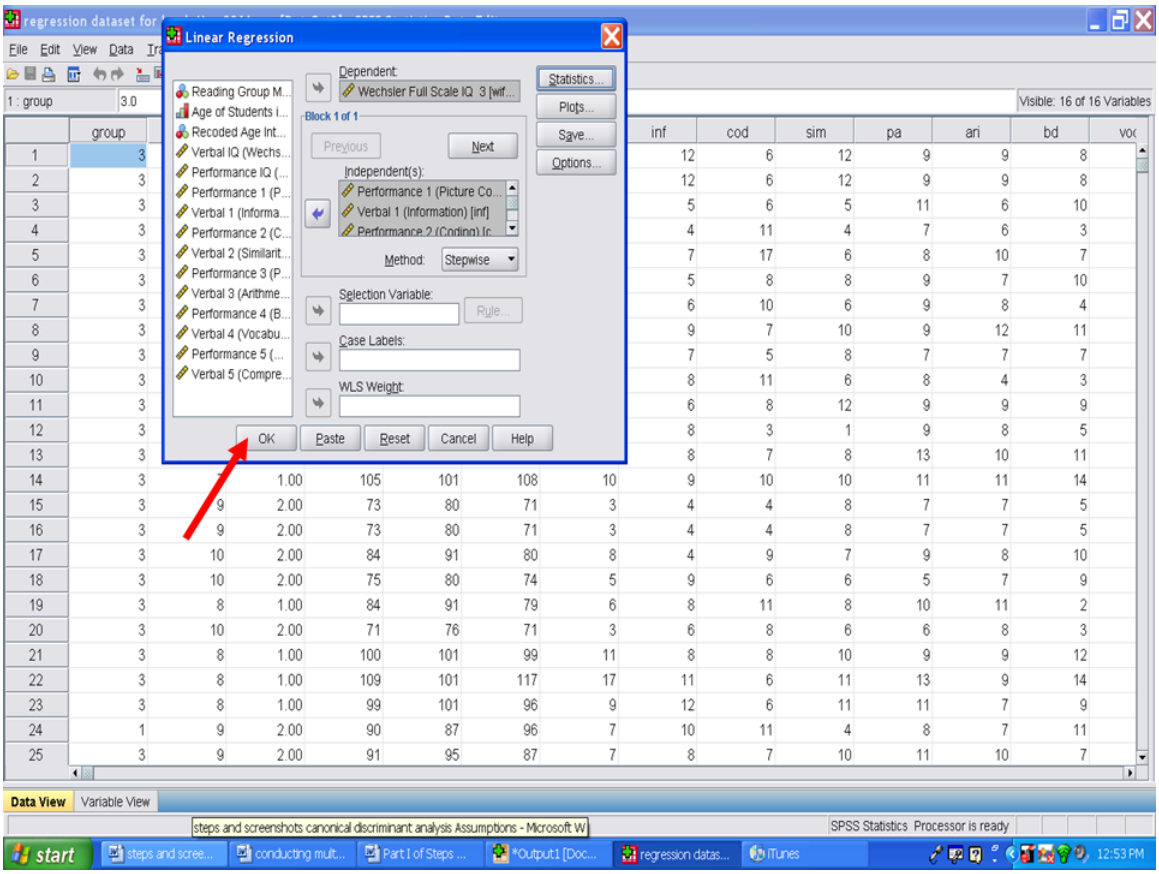
[11]

Next click on

√ OK

Go to the Output file if SPSS does not auomatically send you there.

[11] http://cnx.org/content/m40739/latest/18.8.png/image



[12]

[12] http://cnx.org/content/m40739/latest/18.89.png/image

Chapter 10

Multiple Regression: Part II[1]

NOTE: This chapter is published by NCPEA Press[2] and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

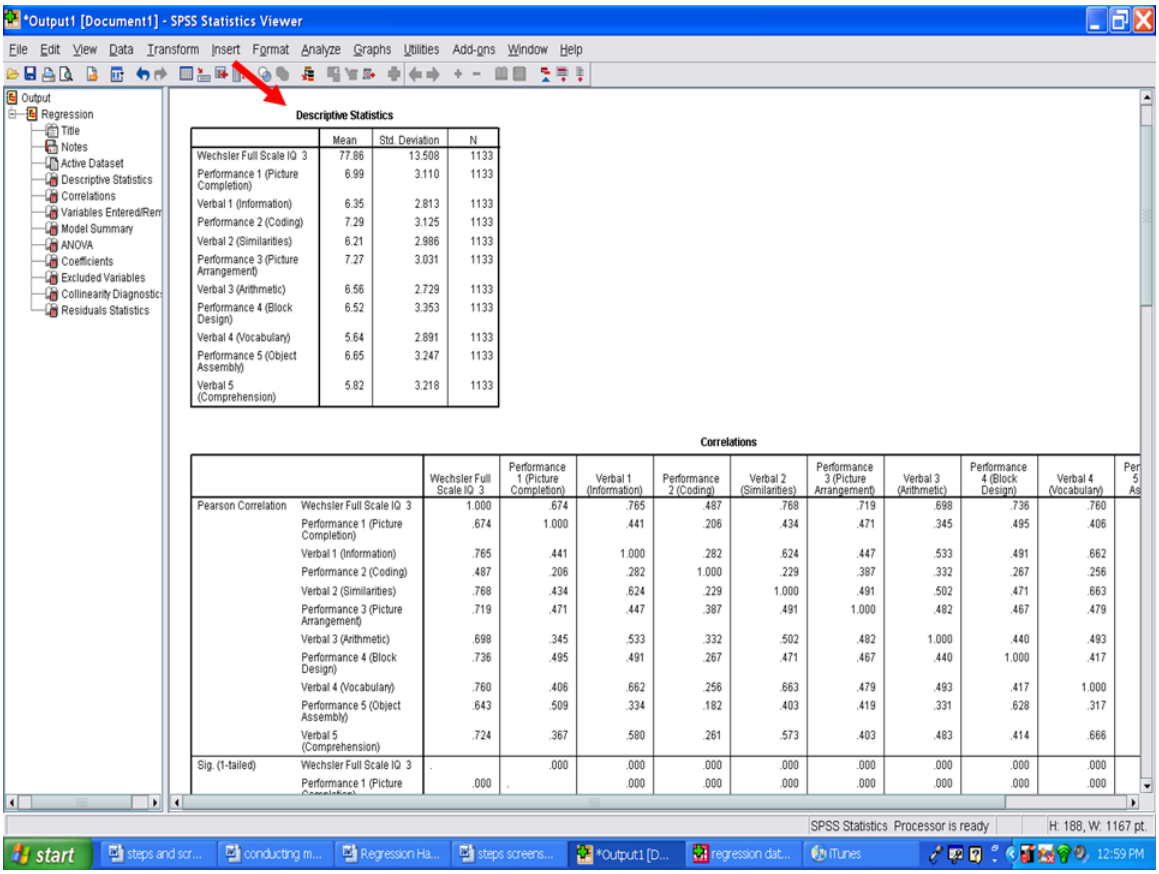
About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help[3]) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

The first table in your SPSS output file should be a Descriptive Statistics table. You will see a column for *M*, *SD*, and *n*. This information should be used in your Results section.

[1] This content is available online at <http://cnx.org/content/m40742/1.2/>.
[2] http://www.ncpeapublications.org/books.html
[3] http://cnx.org/content/m40742/latest/www.writingandstatisticalhelp

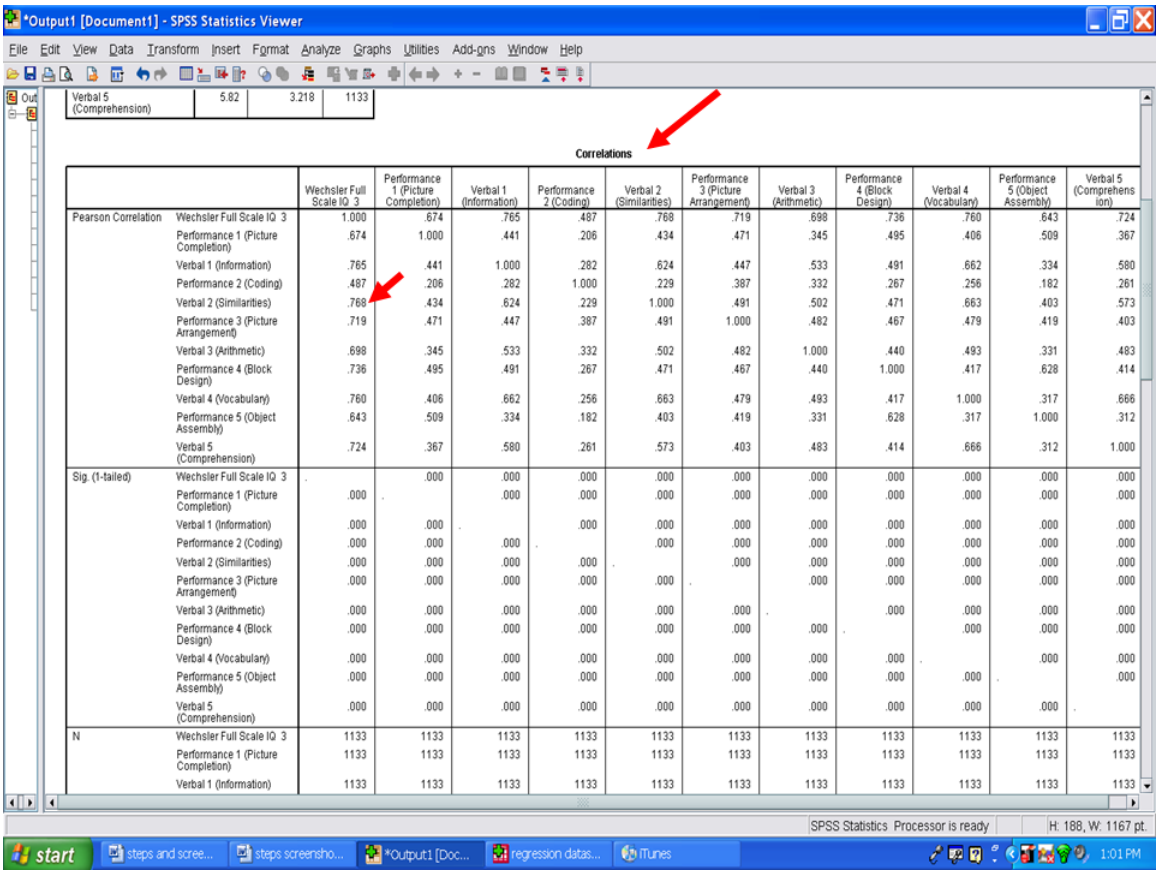


[4]

Underneath the Descriptive Statistics table is a table labeled Correlations. This table reflects the Pearson rs for each independent variable with the dependent variable, as well as the interrelationships among all of the variables.

Of the relationships of the independent variable with the Wechsler Full Scale IQ 3 (the dependent variable in this example), Verbal 2 (Similarities) has the highest correlation, .768. In a stepwise regression procedure, this variable should be the first statistically significant predictor.

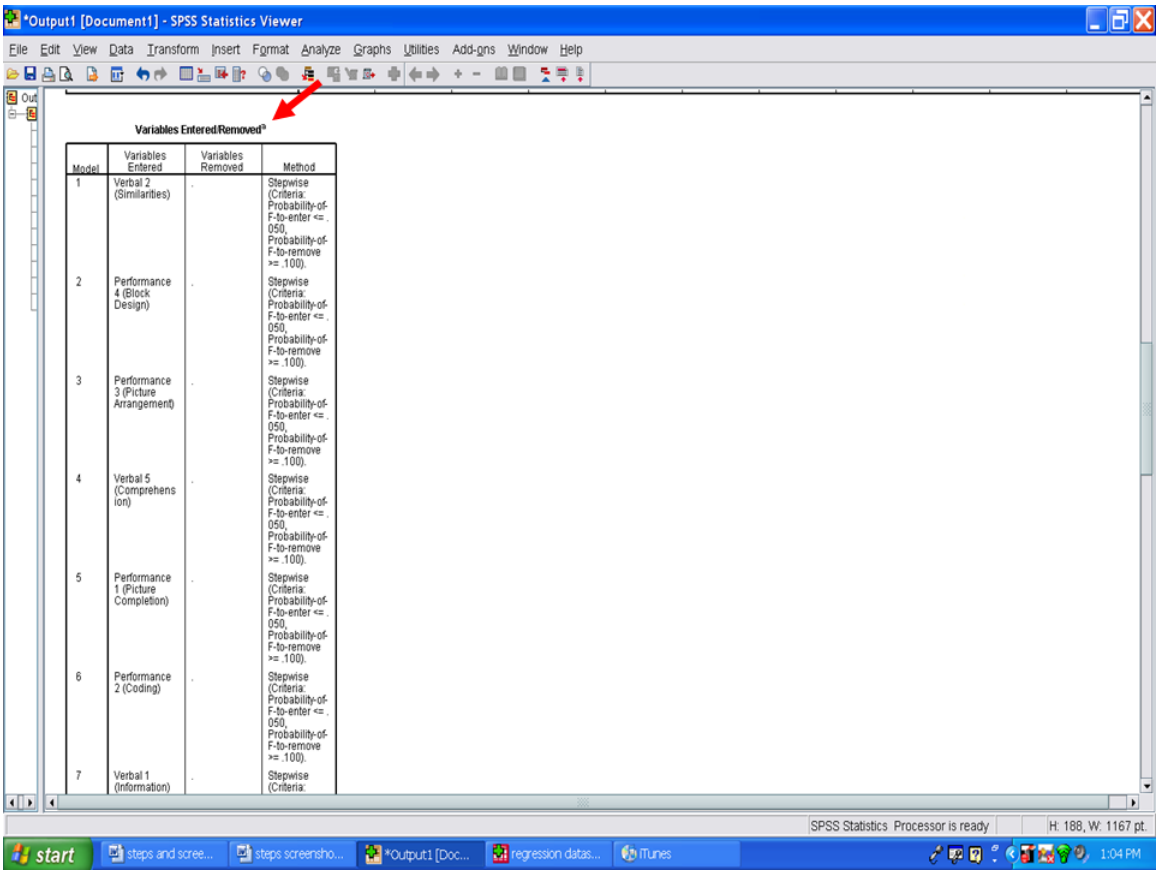
[4] http://cnx.org/content/m40742/latest/19.1.png/image



[5]

Next, you will see the Variables Entered/Removed table. We will not use the information in this table.

[5] http://cnx.org/content/m40742/latest/19.2.png/image

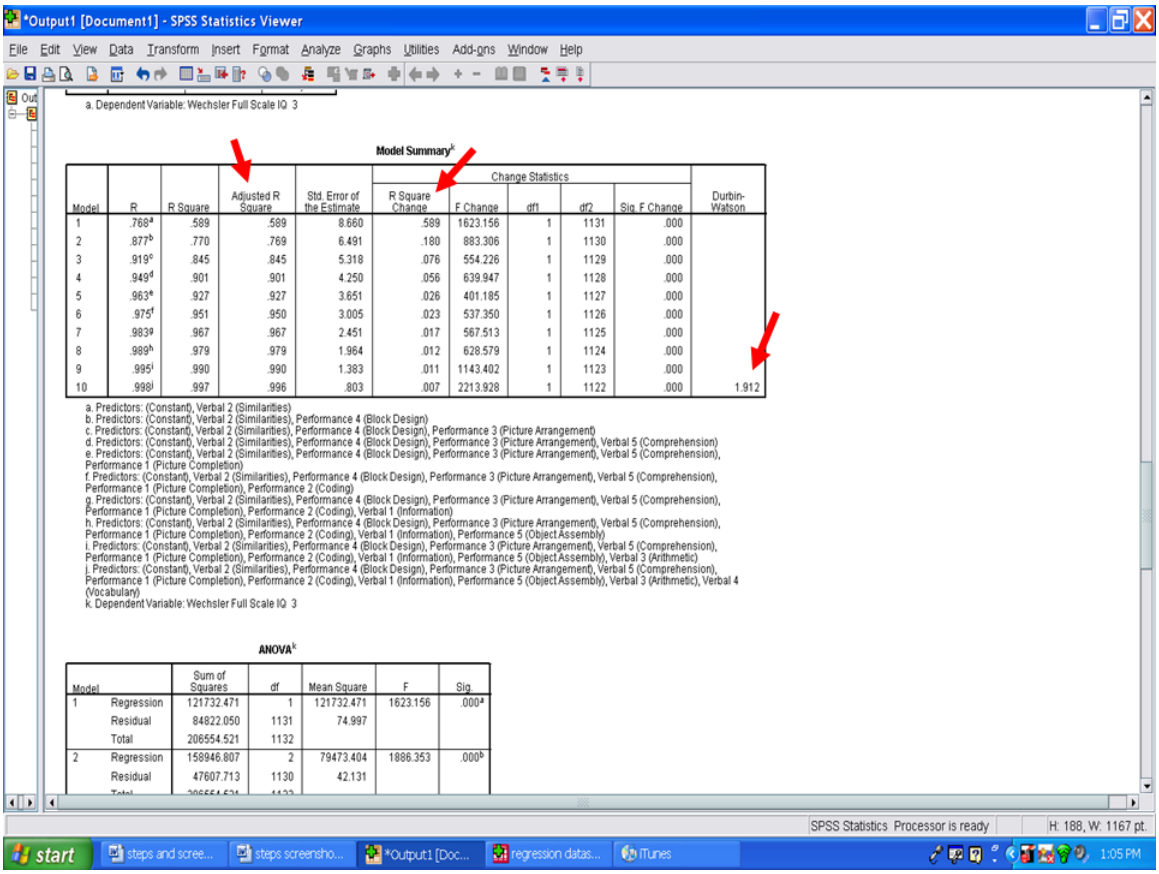


[6]

The table labeled Model Summary is an important table. This table is a summary of all of the steps entered/removed. Important columns in this table are:

Adjusted R Square
R Square Change
Durbin-Watson

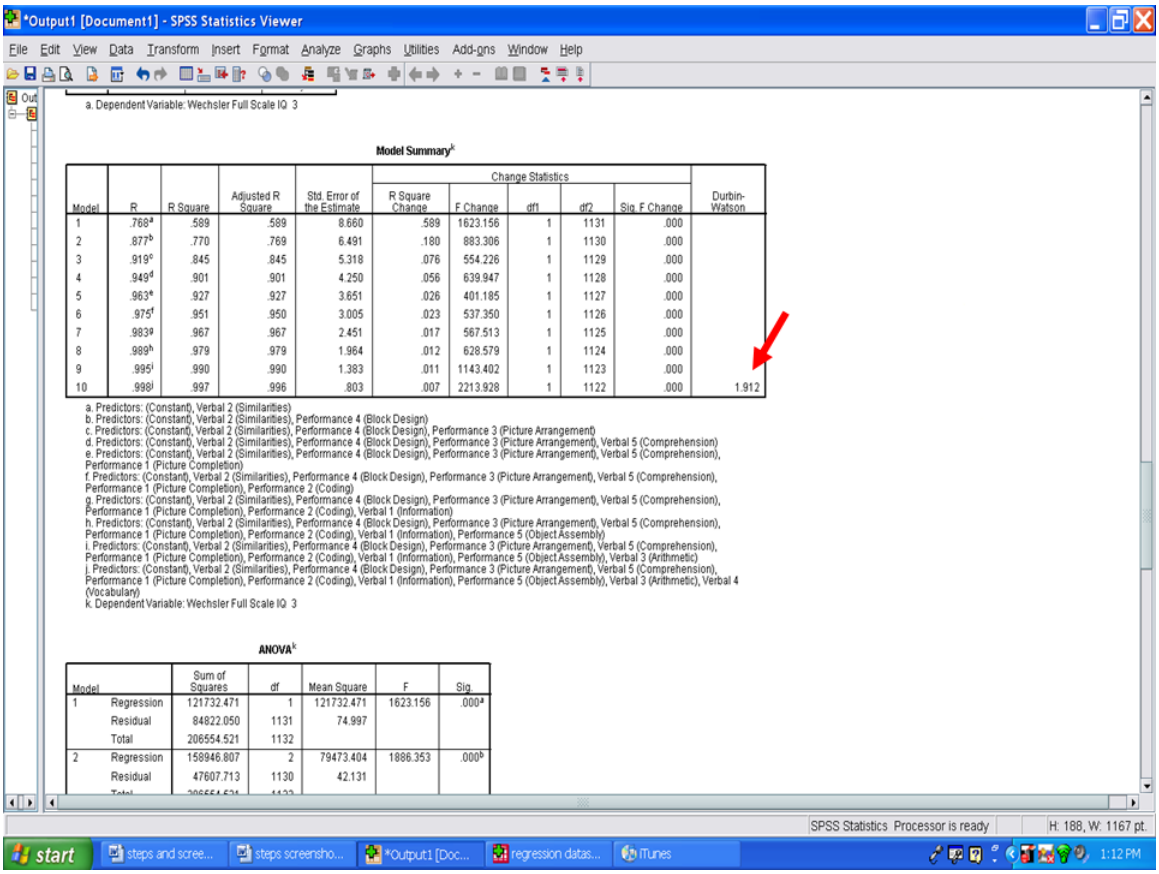
[6] http://cnx.org/content/m40742/latest/19.3.png/image



[7]

The Durbin-Watson statistic is a measure for significant residual autocorrelation. Ideally it should be close to 2. In our example, the Durbin-Watson statistic is 1.912. Therefore, this assumption has not been violated.

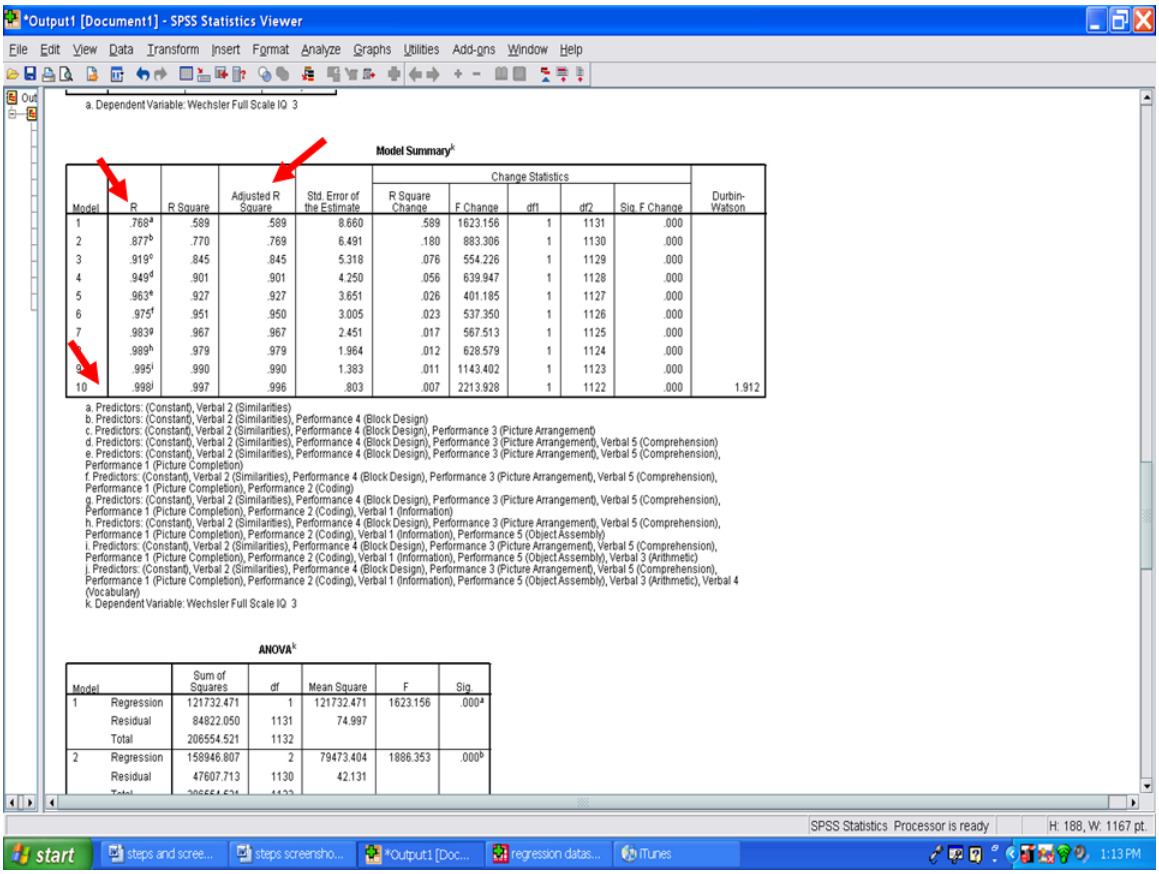
[7] http://cnx.org/content/m40742/latest/19.4.png/image



[8]

The Adjusted R Square column indicates the amount of variance that each model explains in the dependent variable (i.e., Wechsler Fll Scale IQ 3). The first model has the letter a next to it. This model contains a single independent variable, Verbal 2 (Similarities). If you recall, this variable has the highest Pearson r with the Wechsler Full Scale IQ 3. In this example, Verbal 2 (Similarities) accounts for 58.9% of the variance in the Wechsler Full Scale IQ 3.

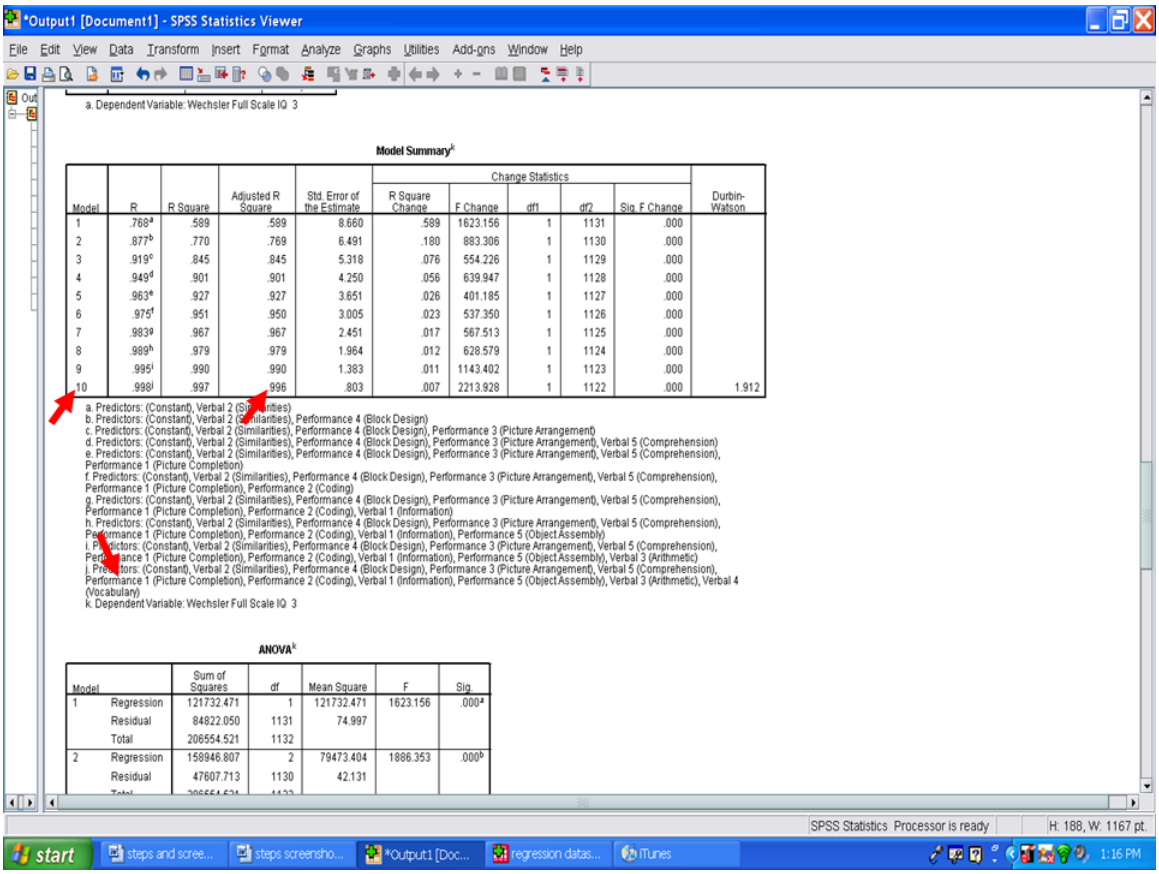
[8] http://cnx.org/content/m40742/latest/19.5.png/image



[9]

Each model, from 1 to 10, depicts the addition of another statistically significant variable in predicting the Wechsler Full Scale IQ 3. The final model, 10, indicates that all 10 independent variables accounted for 99.6% of the variance in the Wechsler Full Scale IQ 3.

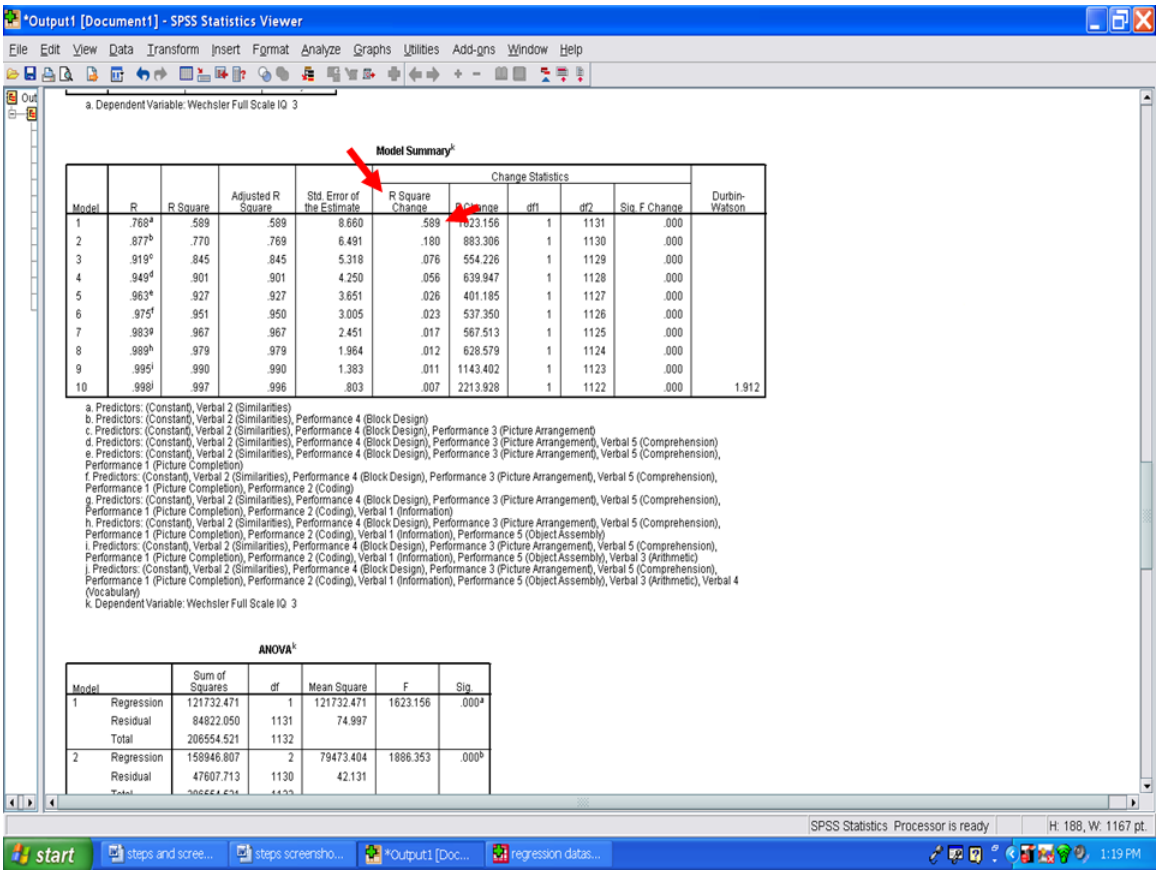
[9] http://cnx.org/content/m40742/latest/19.6.png/image



10

Next, examine the R Square Change column. Each value reflects the unique variance in the Wechsler Full Scale IQ 3 explained by each statistically significant predictor variable. For Model 1, Verbal 2 (Similarities) explained the most variance, 58.9%. In Model 2, Performance 4 (Block Design) added 18.0% of unique variance that it explained.

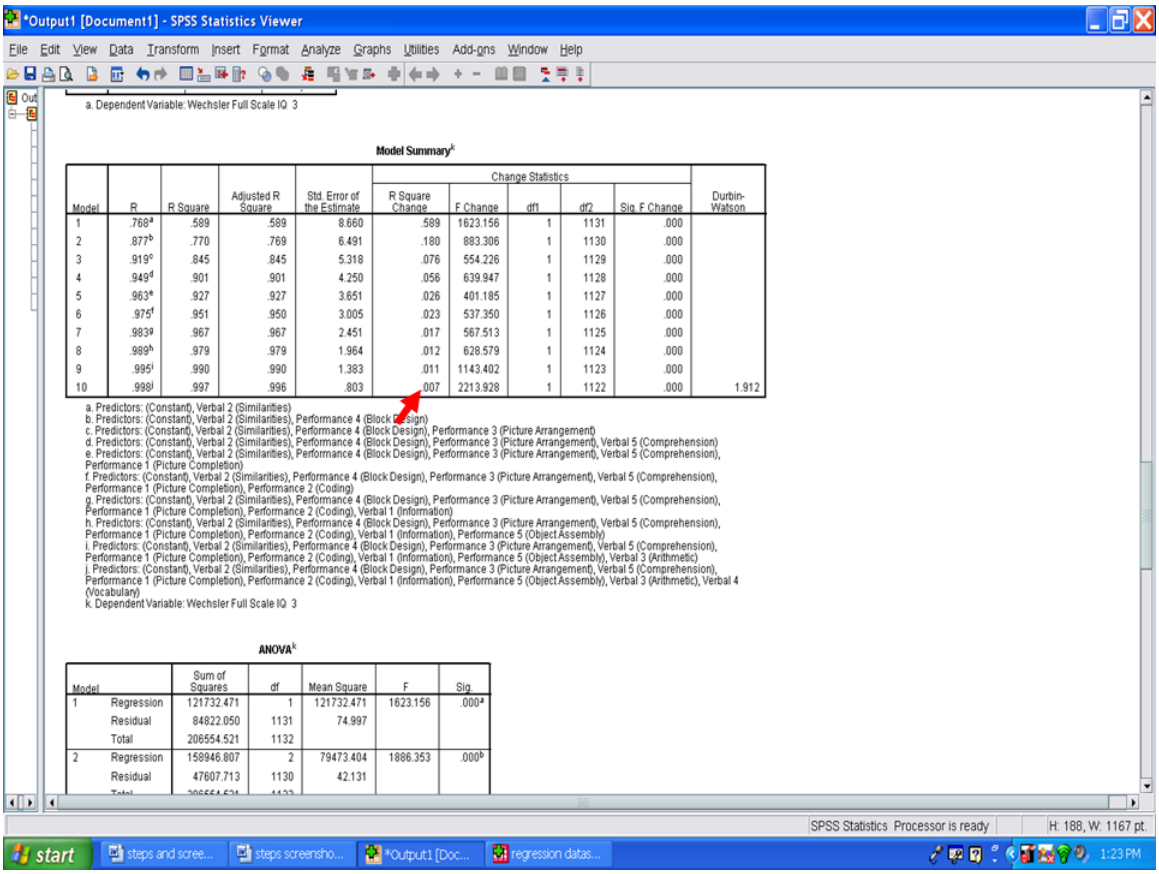
[10] http://cnx.org/content/m40742/latest/19.7.png/image



[11]

The final model, 10, only added 0.7% of additional variance explained.

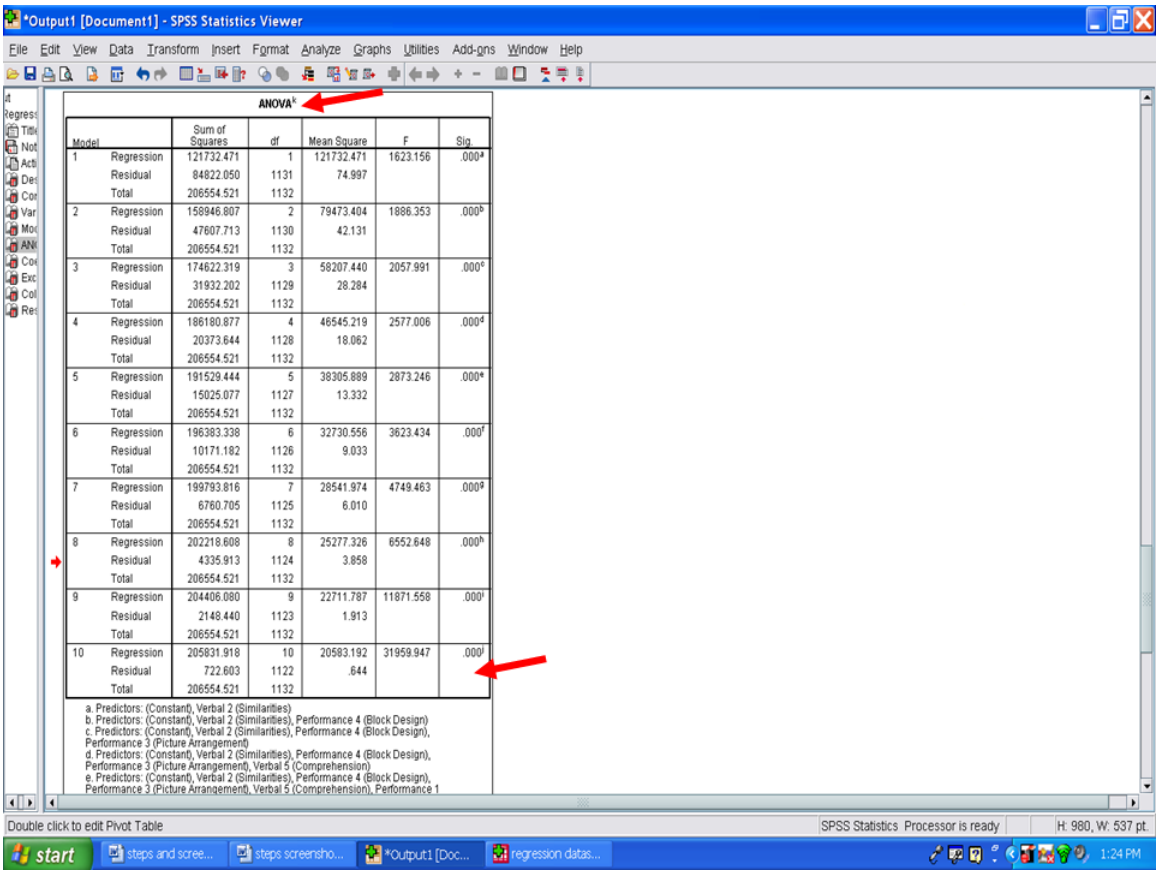
[11] http://cnx.org/content/m40742/latest/19.8.png/image



[12]

The next table of importance is the ANOVA table. Each model is examined to determine the extent to which it explains a statistically significant amount of the variance in the dependent variable. Of interest to us is the very last model, 10, which shows a statistically significant result, $F(10, 1122) = 31959.947$, $p < .001$.

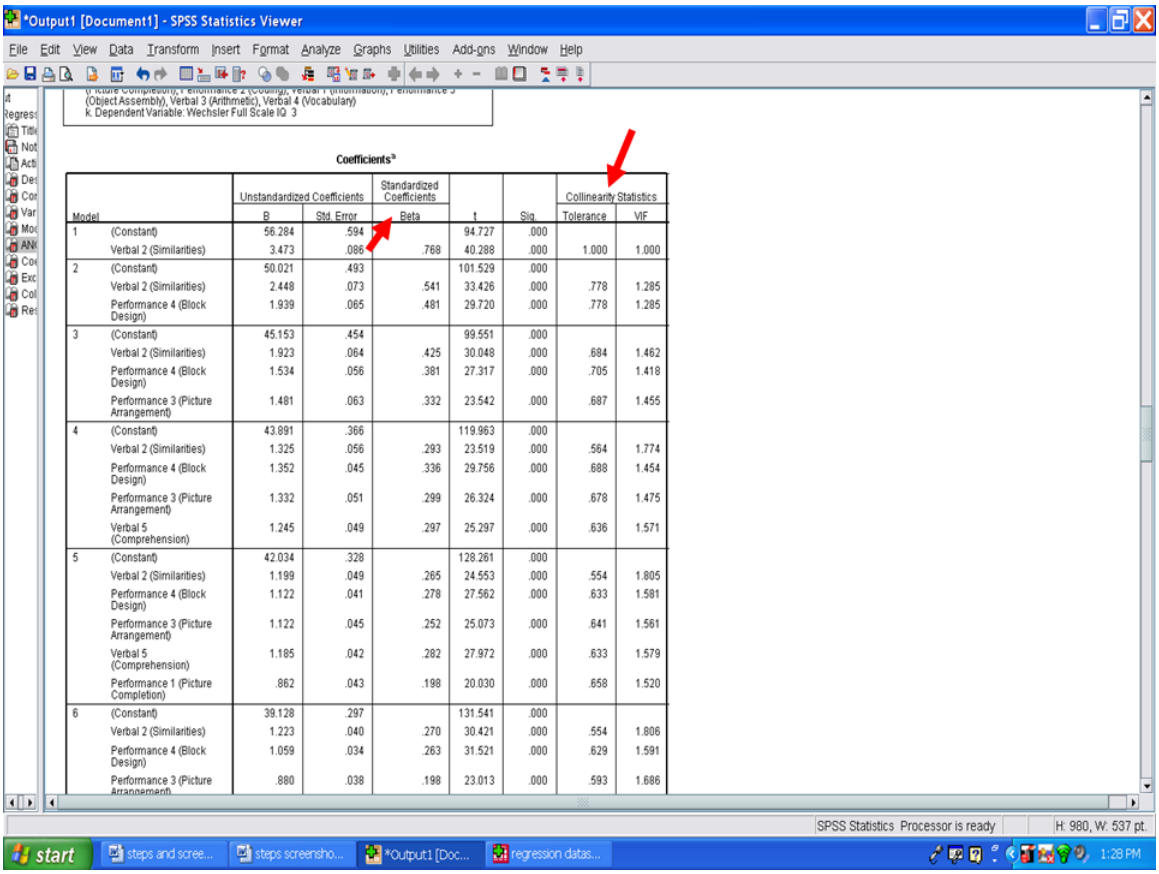
[12] http://cnx.org/content/m40742/latest/19.9.png/image



13

Underneath the ANOVA table is the Coefficients table. The important columns in this table are the Standardized Coefficients Beta and the Collinearity Statistics Tolerance and VIF ones. We will scroll down this table until we get to the final model, 10, information.

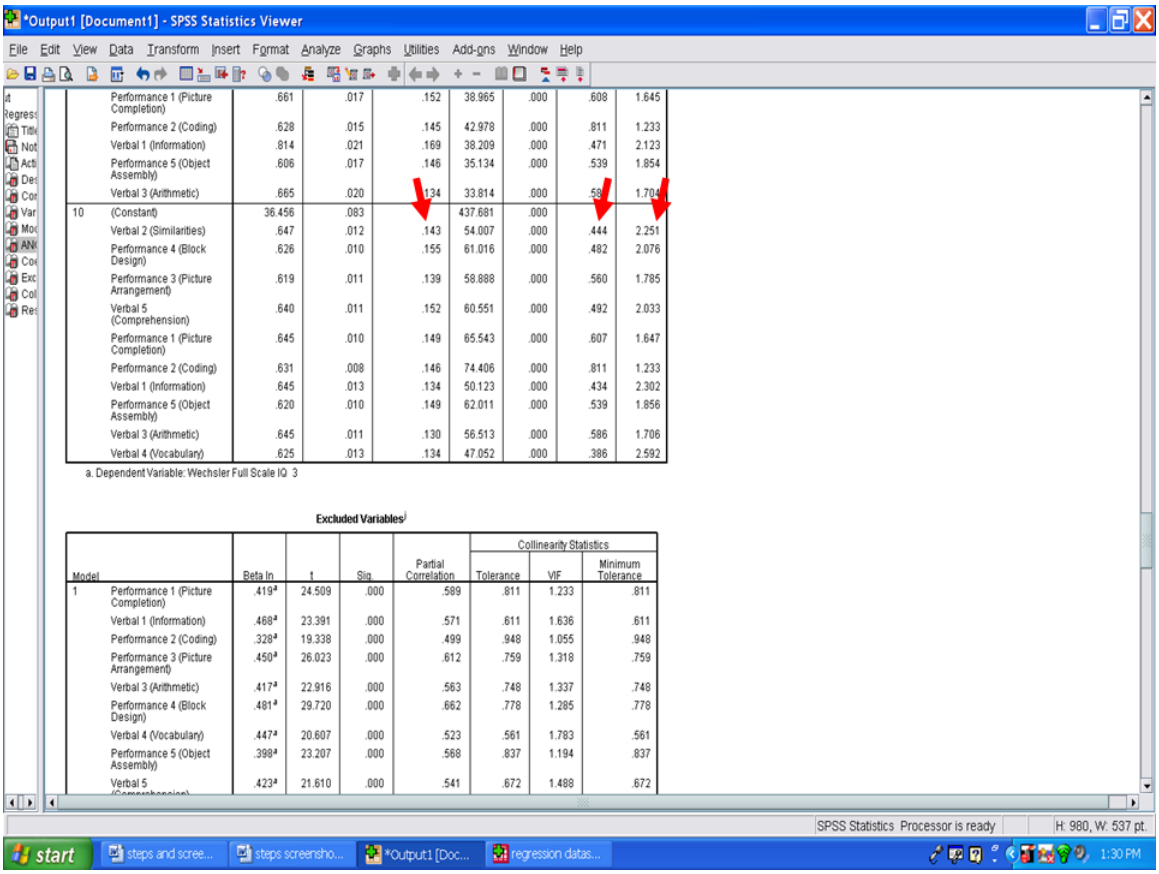
[13] http://cnx.org/content/m40742/latest/19.10.png/image



[14]

Scrolling down to the last model information shows us the following information. If you recall from the assumption checks, that multicollinearity was to be examined in the SPSS output. Multicollinearity is present when Tolerance values are below .1. As you can tell from the output below, Tolerance values range from a low of .386 to a high of .811. A second check for multicollinearity is the VIF column. Multicollinearity is present in the VIF column when the VIF values are greater than 10. In the example below, the VIF values range from a low of 1.233 to a high of 2.592. Therefore, multicollinearity is not present in this example.

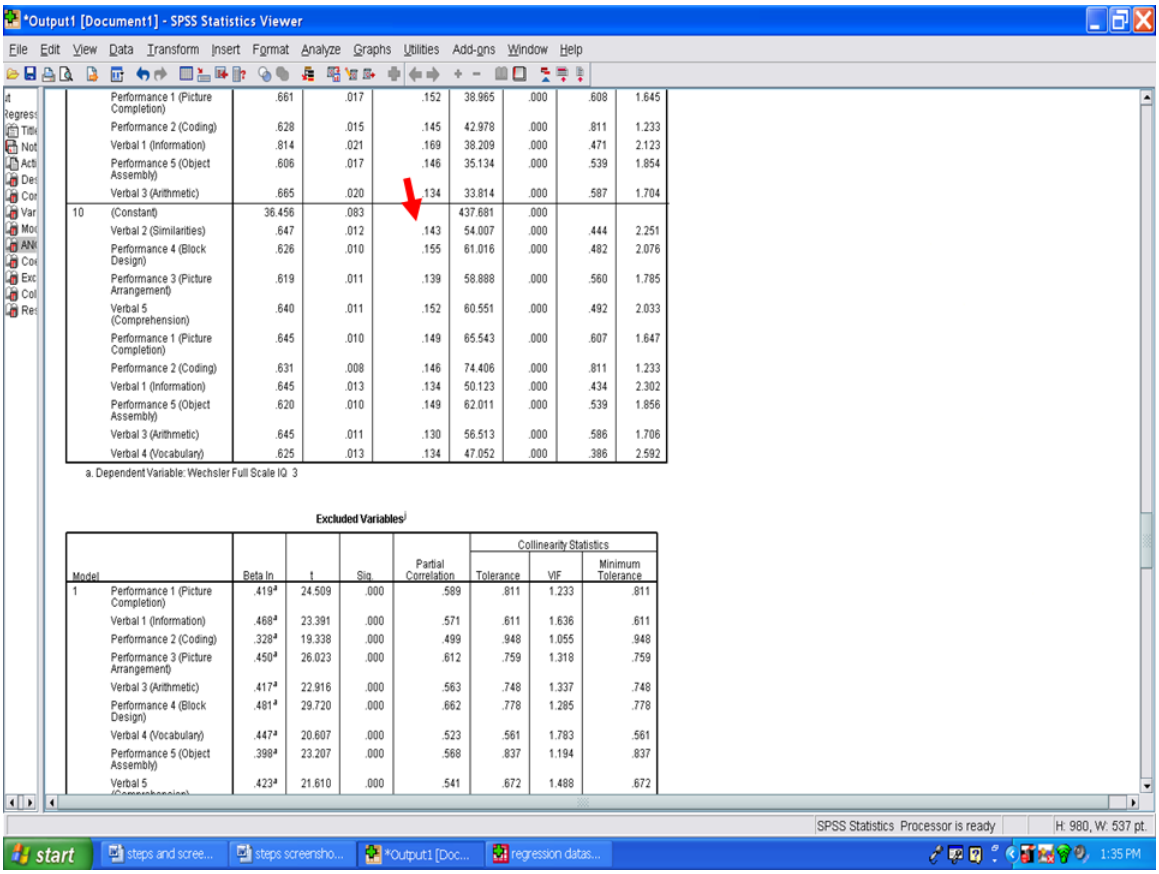
[14] http://cnx.org/content/m40742/latest/19.11.png/image



[15]

The Standardized Coefficients Beta column is to be examined next. Each of these values reflects the relative importance of each of these statistically significant predictors. In this column, you will see that the Betas range from a low of .130 to a high of .155, indicating that each has about the same degree of relative importance.

[15] http://cnx.org/content/m40742/latest/19.12.png/image



16

In your Results section, you should discuss the assumptions that you checked; the extent to which each assumption was met or not met; the descriptive statistics; and the information in the columns that was discussed at each step of the regression process.

You have now successfully gone through the calculation of a multiple regression analysis.

[16] http://cnx.org/content/m40742/latest/19.13.png/image

Attributions

Collection: *Calculating Advanced Statistics: Part II*
Edited by: National Council of Professors of Educational Administration
URL: http://cnx.org/content/col11348/1.2/
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multivariate Analysis of Variance: Part I"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40728/1.2/
Pages: 1-28
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multivariate Analysis of Variance: Part II"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40729/1.2/
Pages: 29-41
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multiple Analysis of Variance: Part III"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40731/1.2/
Pages: 43-54
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discriminant Analysis: Assumptions"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40733/1.2/
Pages: 55-61
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discriminant Analysis: Part I"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40735/1.2/
Pages: 63-71
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discriminant Analysis: Part II"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40736/1.2/
Pages: 73-83
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discriminant Analysis: Part III"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40737/1.2/
Pages: 85-93
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multiple Regression: Assumptions"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40738/1.2/
Pages: 95-103
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multiple Regression: Part I"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40739/1.2/
Pages: 105-114
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

Module: "Multiple Regression: Part II"
By: John R. Slate, Ana Rojas-LeBouef
URL: http://cnx.org/content/m40742/1.2/
Pages: 115-128
Copyright: John R. Slate, Ana Rojas-LeBouef
License: http://creativecommons.org/licenses/by/3.0/

About Connexions

Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities.

Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai. Connexions is part of an exciting new information distribution system that allows for **Print on Demand Books**. Connexions has partnered with innovative on-demand publisher QOOP to accelerate the delivery of printed course materials and textbooks into classrooms worldwide at lower prices than traditional academic publishers.