Instructions for Conducting Multiple Linear Regression Analysis in SPSS

Multiple linear regression analysis is used to examine the relationship between two or more independent variables and one dependent variable. The independent variables can be measured at any level (i.e., nominal, ordinal, interval, or ratio). However, nominal or ordinal-level IVs that have more than two values or categories (e.g., race) must be recoded prior to conducting the analysis because linear regression procedures can only handle interval or ratio-level IVs, and nominal or ordinal-level IVs with a maximum of two values (i.e., dichotomous). The dependent variable MUST be measured at the interval- or ratio-level. In this demonstration we use base year standardized reading score (BY2XRSTD) as the dependent variable and socio-economic status (BYSES), family size (BYFAMSIZ), self-concept (BYCNCPT1), urban residence, (URBAN), rural residence (RURAL), and sex (GENDER) as independent variables. In order to conduct the analysis, we have to recode two variables from the original data set [i.e., urbanicity (G8URBAN) into urban residence (URBAN), and rural residence (RURAL), and sex (SEX) into sex (GENDER)]. We include instructions for this step as well. Although not used in the analysis, the data set also includes two identification variables [student ID (STU_ID) and school ID (SCH_ID)] and one weight variable (F2PNWLWT).

Copies of the data set and output are available on the companion website. The data set file is entitled, "REGRESSION.SAV". The output file is entitled, "Multiple Linear Regression results.spv".

The following instructions are divided into three sets of steps:

1. Recode G8URBAN and SEX into new dichotomous variables (i.e., URBAN, RURAL, & GENDER)
2. Conduct preliminary analyses

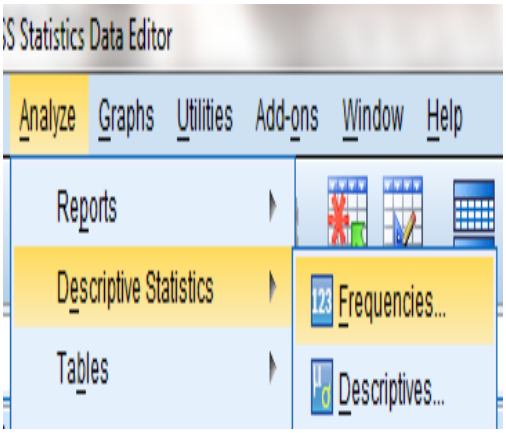
a. Examine descriptive statistics of the continuous variables
b. Check the normality assumption by examining histograms of the continuous variables
c. Check the linearity assumption by examining correlations between continuous variables and scatter diagrams of the dependent variable versus independent variables.

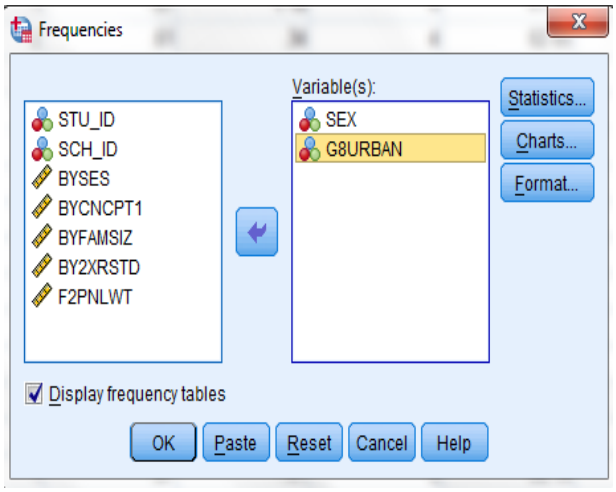
3. Conduct multiple linear regression analysis
    a. Run model with dependent and independent variables
    b. Model Check
        i. Examine collinearity diagnostics to check for multicollinearity
        ii. Examine residual plots to check error variance assumptions (i.e., normality and homogeneity of variance)
        iii. Examine influence diagnostics (residuals, dfbetas) to check for outliers
        iv. Examine significance of coefficient estimates to trim the model
    c. Revise the model and rerun the analyses based on the results of steps i-iv.
    d. Write the final regression equation and interpret the coefficient estimates.

To get started, open the SPSS data file entitled, REGRESSION.SAV.

STEP I: Recode SEX and G8URBAN into dichotomous variables

Run frequencies for SEX and G8URBAN. This will be helpful later for verifying that the dichotomous variables were created correctly. At the **Analyze** menu, select **Descriptive Statistics**. Click **Frequencies.** Click SEX and G8URBAN and move them to the **Variables** box. Click **OK.**





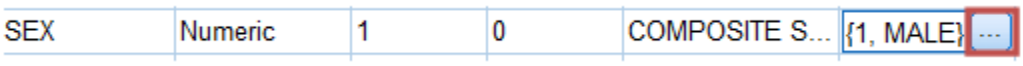
COMPOSITE SEX

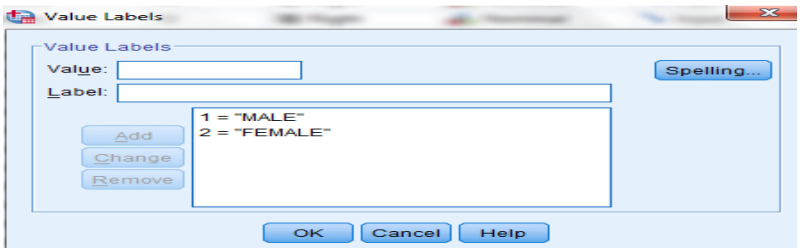
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	720	48.0	48.0	48.0
	FEMALE	780	52.0	52.0	100.0
	Total	1500	100.0	100.0	

URBANICITY COMPOSITE

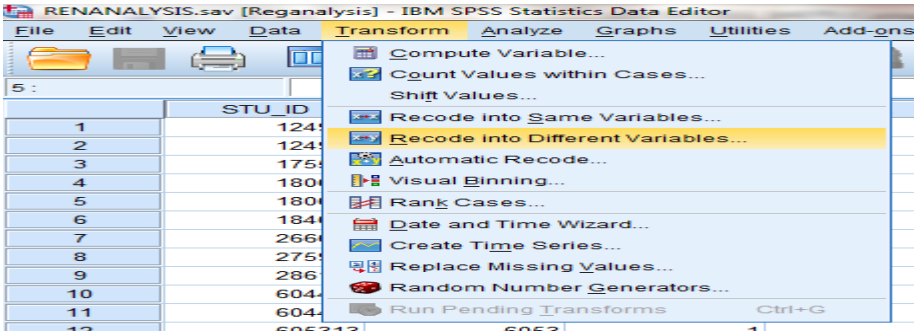
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	URBAN	394	26.3	26.3	26.3
	SUBURBAN	621	41.4	41.4	67.7
	RURAL	485	32.3	32.3	100.0
	Total	1500	100.0	100.0	

Prior to recoding SEX into a dichotomous variable, we need to determine what the numeric values are for Male and Female. To do this, click on SPSS **variable view** at the bottom left corner of your SPSS screen**.** Click the **Values** column for SEX variable. Click on the grey bar to reveal how Male and Female are coded. We see that Male is coded as "1" and Female is coded as "2". Click **Cancel.**

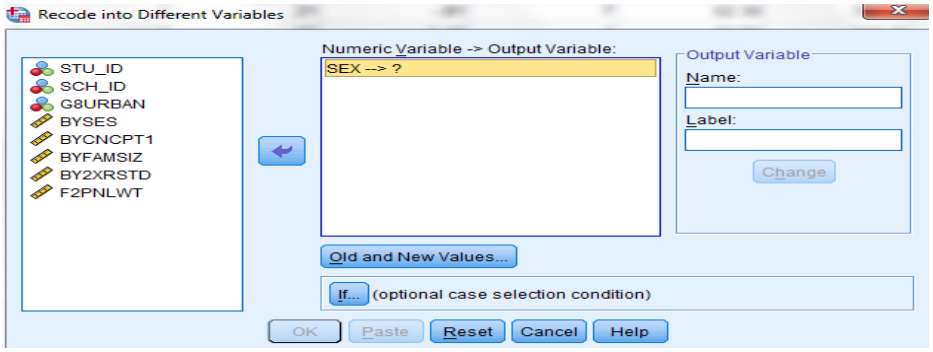




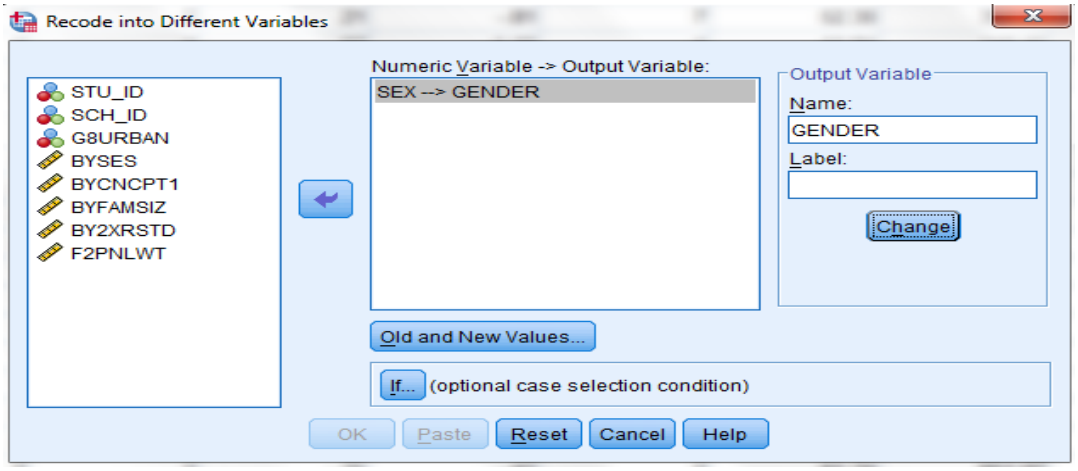
To run multiple regression analysis in SPSS, the values for the SEX variable need to be recoded from '1' and '2' to '0' and '1'. Because the value for Male is already coded 1, we only need to re-code the value for Female, from '2' to '0'. Open the **Transform** menu at the top of the SPSS menu bar. Select and click **Recode into Different Variables**.



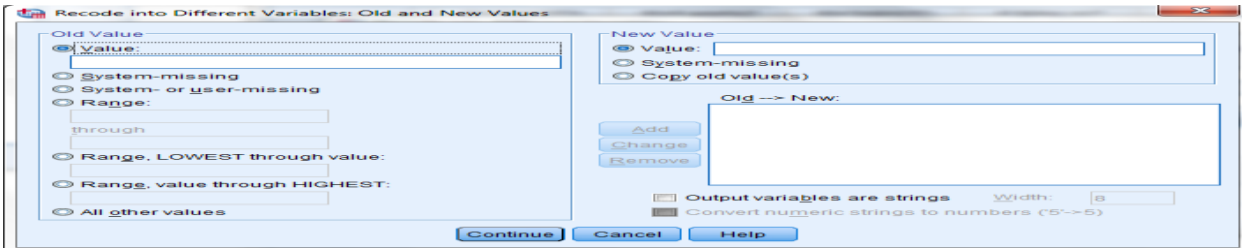
On the variables list in the box on the left, click SEX and move it to center box.



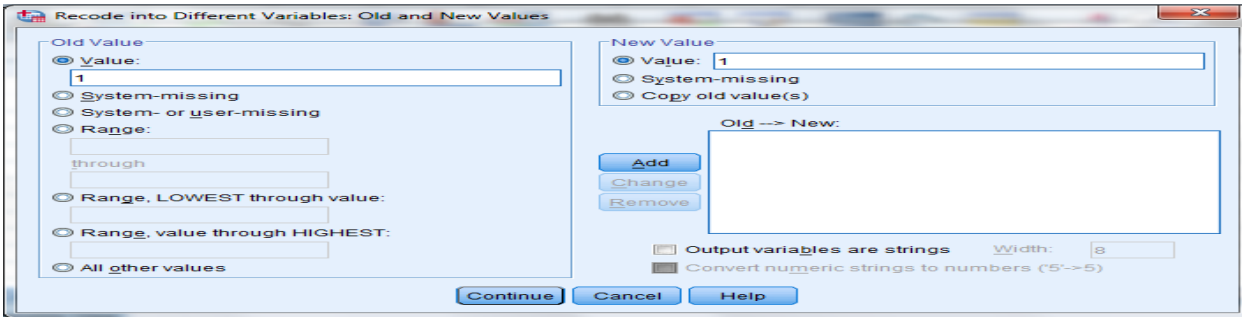
Within the Output variable **Name** type 'GENDER' to rename the SEX variable. Click **Change.**

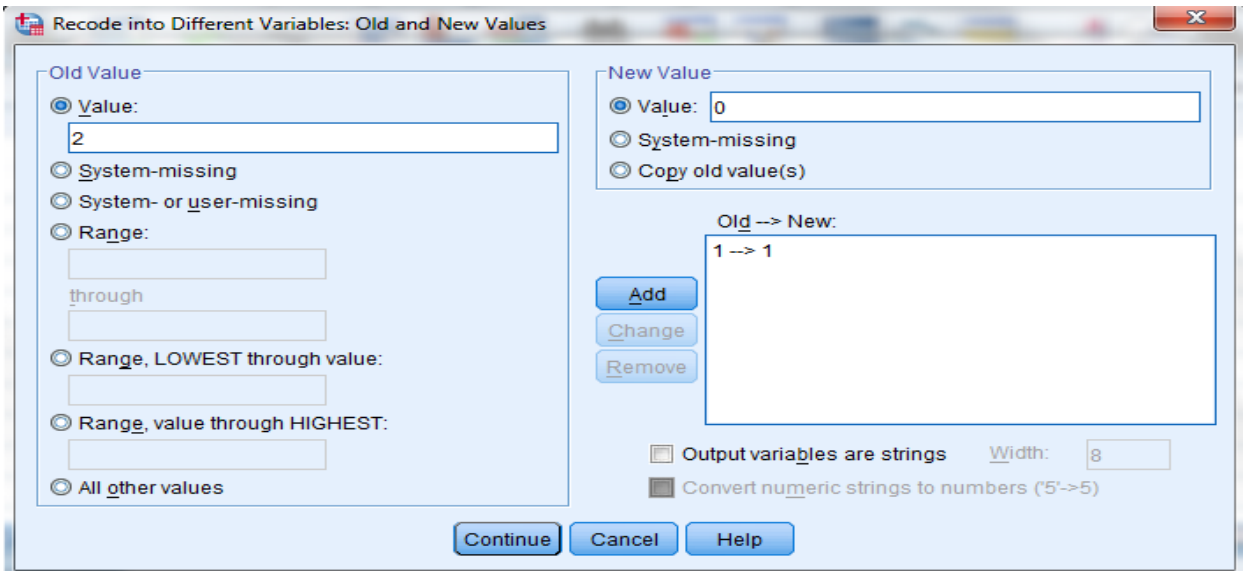


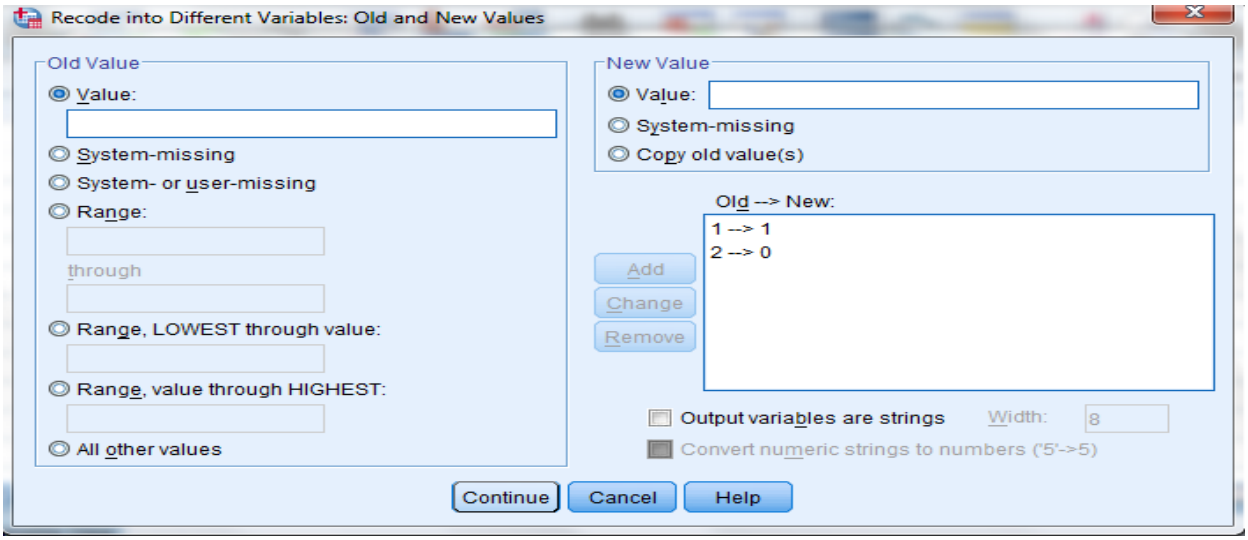
Click **Old and New Values.** The "Old Value" is the value for the level of the categorical variable (SEX) to be changed. The "New Value" is the value for the level on the recoded variable (GENDER).



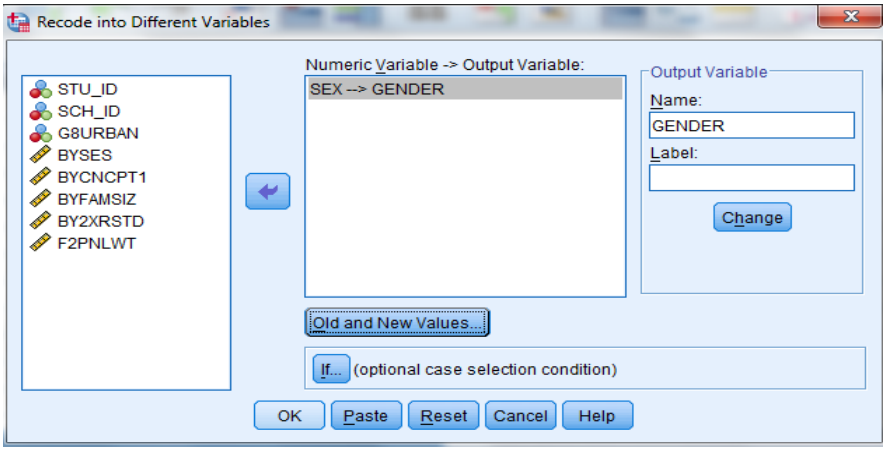
The value for Male is "1" on both SEX and GENDER. Type "**1"** in the **Old Value** box; type "1" in the **New Value** box**.** Click **Add.** The value for Female is recoded from "2" on SEX to "0" on GENDER. Type "2" in the **Old Value** box; type "0" in the **New Value** box**.** Click **Add.**



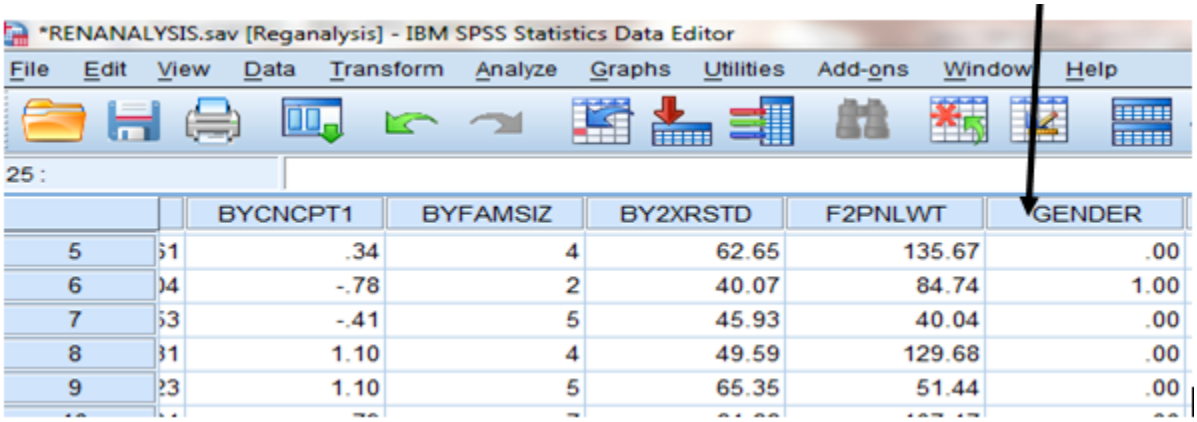




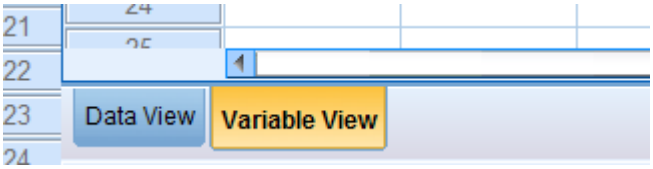
Click **Continue.** Click **OK**



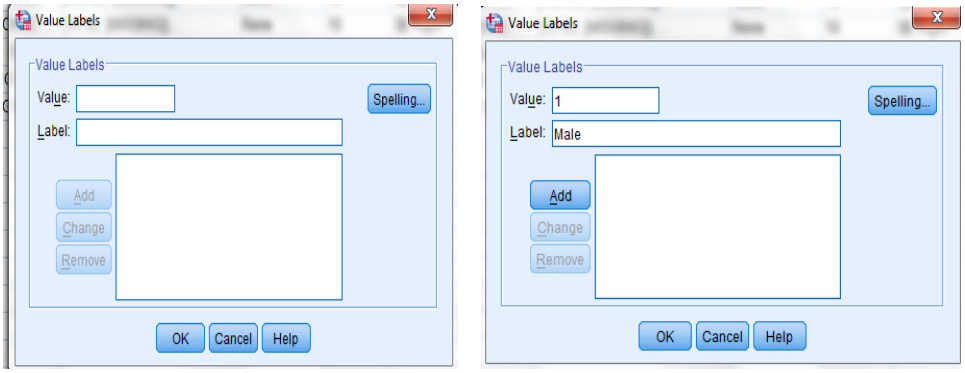
The variable GENDER will be added to the dataset.



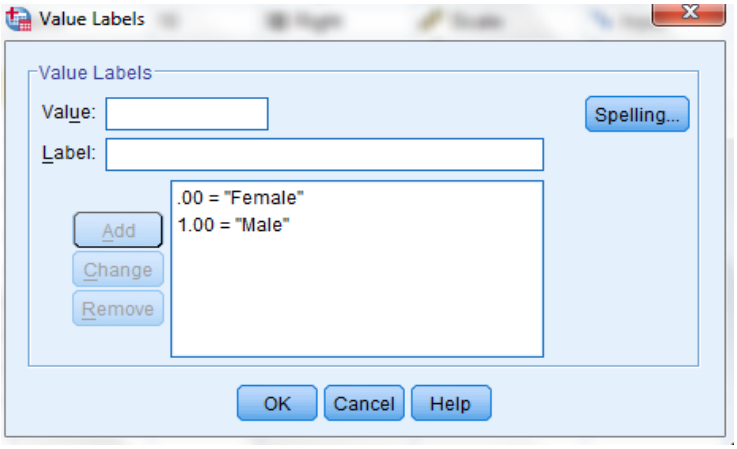
Next we label the values for the GENDER categories (Male and Female). Click **Variable View** on bottom left corner of the data spread sheet.



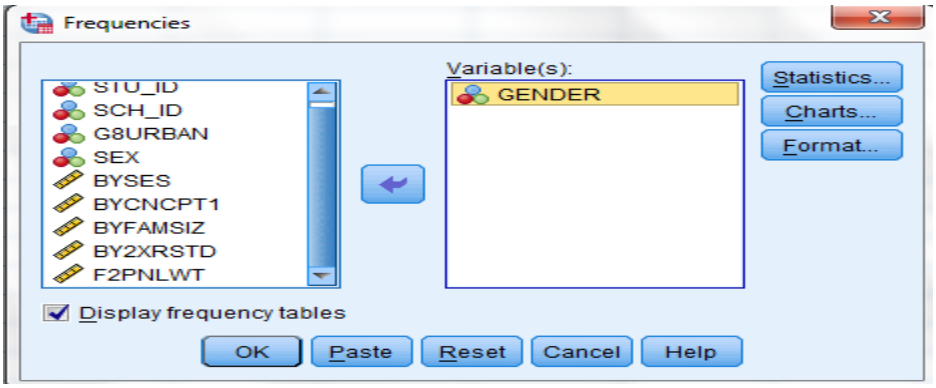
Click the grey bar on the **Values** column for GENDER



Type "1" into the Value box; type "Male" into the Label box. Click **Add.** Type **"0"** into the Value box; type "Female" into the Label Box. Click **Add.** Click **OK.**



To determine whether GENDER was created correctly, run a frequency on the newly created GENDER variable.



The values for Males and Females should match those in the SEX variable.

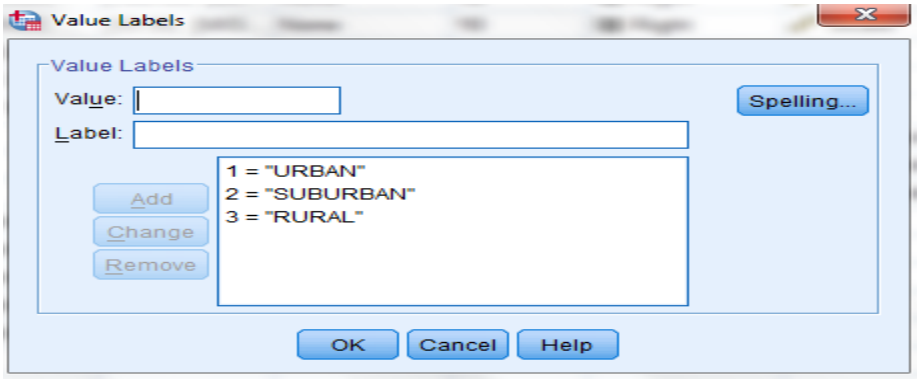
GENDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	780	52.0	52.0	52.0
	Male	720	48.0	48.0	100.0
	Total	1500	100.0	100.0	

COMPOSITE SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	720	48.0	48.0	48.0
	FEMALE	780	52.0	52.0	100.0
	Total	1500	100.0	100.0	

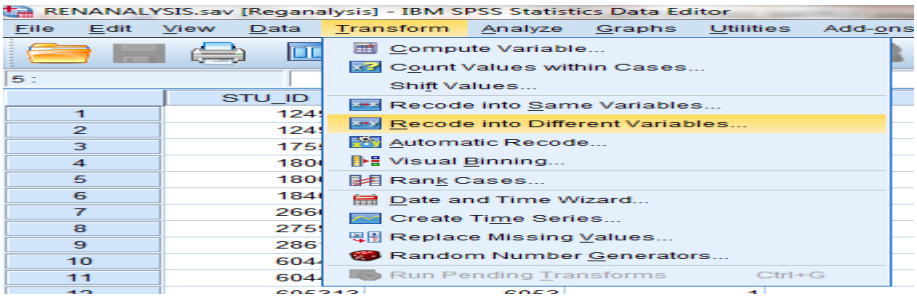
We now need to examine G8URBAN to determine its categories and how it should be recoded. Click on SPSS **variable view.** Click the **Values** column for G8URBAN variable. Click on the grey bar to reveal how G8URBAN categories are coded. We see that G8URBAN has three categories.



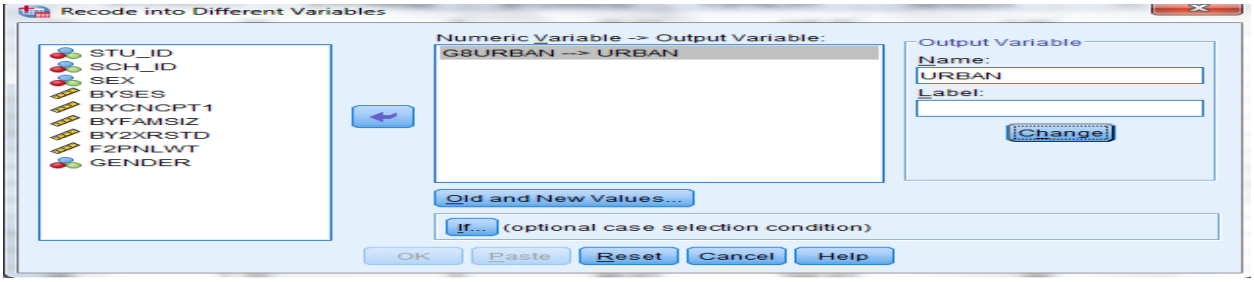
Recall that in multiple regression analysis, independent variables must be measured as either interval/ratio or nominal/ordinal with only two values (i.e., dichotomous). Thus, we need to recode G8URBAN. **A rule of thumb in creating dichotomous variables is that for categorical independent variables with more than two categories (i.e., k), we create k minus 1 dichotomous variables.** In this case, G8URBAN has 3 categories, thus, we will create 2 dichotomous variables (3 – 1 = 2). The category that is not included in the dichotomous variables is referred to as the reference category. In this example, SUBURBAN is the reference category. The two new dichotomous variables are named URBAN and RURAL. The recoding is done in the same way that we recoded SEX. The following table shows the values of the old variable (i.e., G8URBAN) and values for each of the new dichotomous variables (URBAN and RURAL):

Variables	Values		
	Urban	Suburban	Rural
Old: G8URBAN	1	2	3
New: URBAN	1	Reference Category	0
New: RURAL	0	Reference Category	1

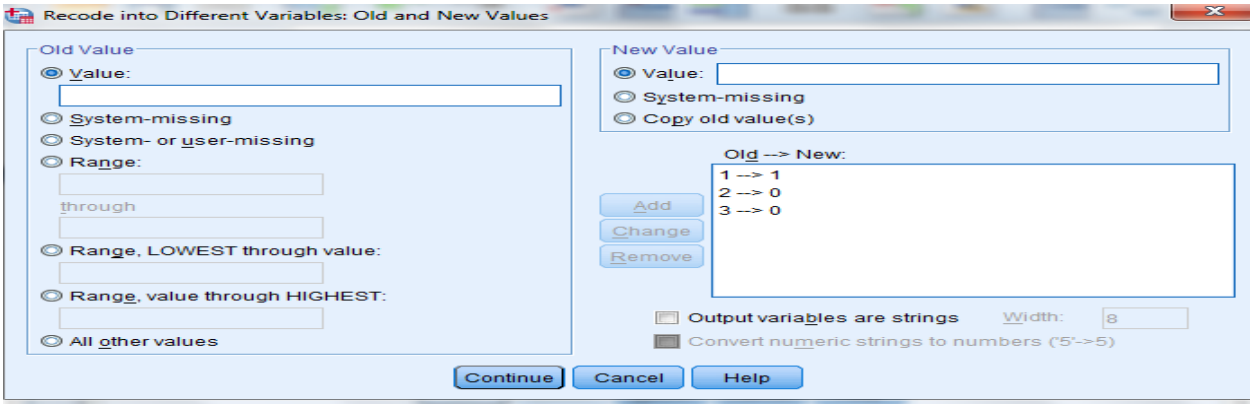
To create URBAN from G8URBAN, open the **Transform** menu at the top of the SPSS menu bar. Select and click **Recode into Different Variables**.



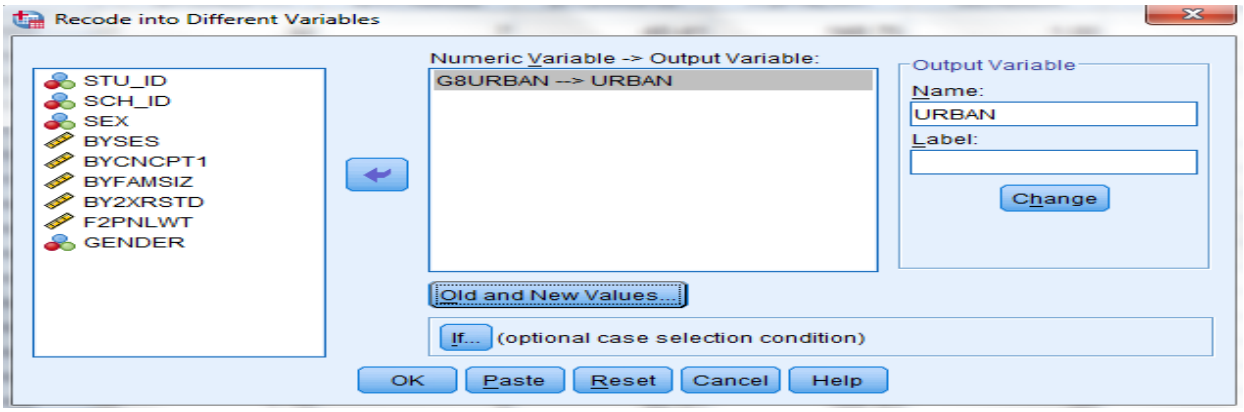
If necessary, click **Reset** to clear previous selections. On the variables list in the box on the left, click G8URBAN and move it to center box. Within the Output variable **Name** type URBAN. Click **Change.**



Click Old and New Values. Type "**1"** in the **Old Value** box; type "1" in the **New Value** box**.** Click **Add.** Type "2" in the **Old Value** box; type "0" in the **New Value** box**.** Click **Add.** Type "3" in the **Old Value** box; type "0" in the **New Value** box**.** Click **Add**. Click **Continue.**



Click **OK.**



A new variable URBAN will be added to the dataset. Run frequencies on URBAN to verify that the number of 1's in the URBAN category matches the number of URBAN in the original G8URBAN variable.

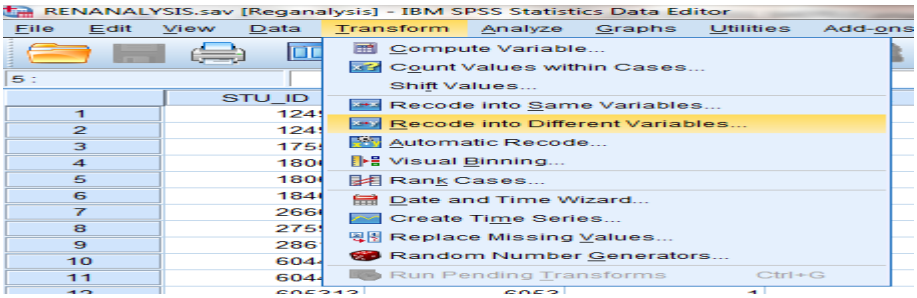
URBAN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1106	73.7	73.7	73.7
	1.00	394	26.3	26.3	100.0
	Total	1500	100.0	100.0	

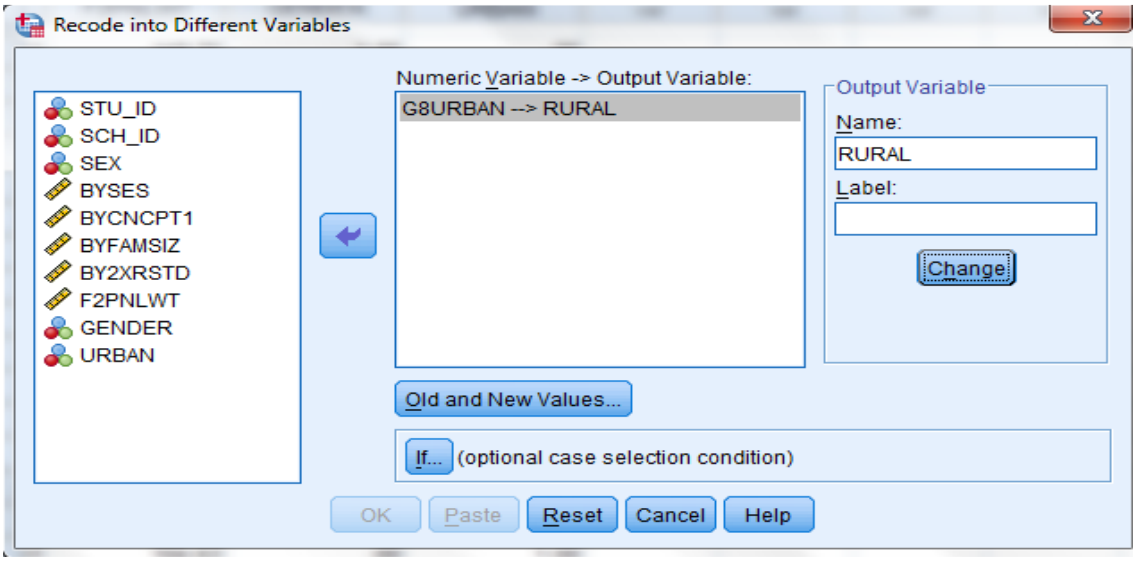
URBANICITY COMPOSITE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	URBAN	394	26.3	26.3	26.3
	SUBURBAN	621	41.4	41.4	67.7
	RURAL	485	32.3	32.3	100.0
	Total	1500	100.0	100.0	

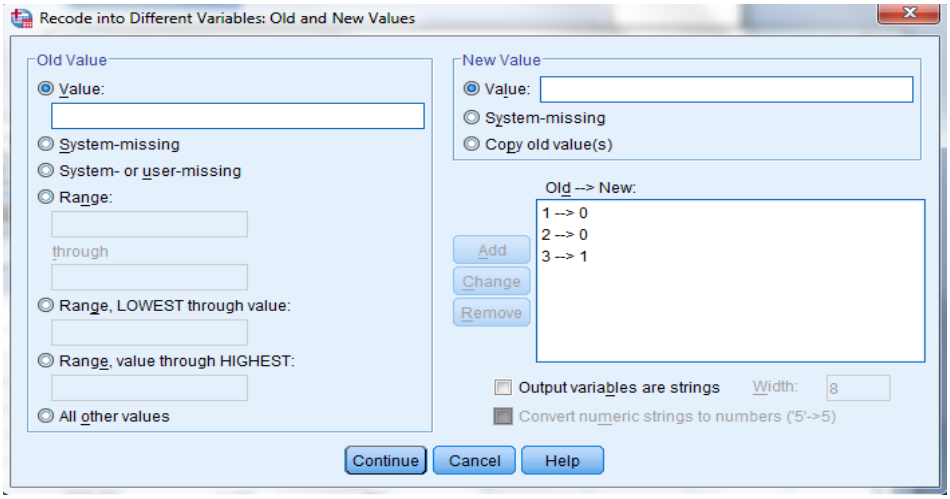
To create RURAL from G8URBAN, open the **Transform** menu at the top of the SPSS menu bar. Select and click **Recode into Different Variables**.



Click **Reset** to clear the previous selections. On the variables list in the box on the left, click G8URBAN and move it to center box. Within the Output variable **Name** type **RURAL**. Click **Change.**



Click Old and New Values. Type "**1"** in the **Old Value** box; type "0" in the **New Value** box**.** Click **Add.** Type "2" in the **Old Value** box; type "0" in the **New Value** box**.** Click **Add.** Type "3" in the **Old Value** box; type "1" in the **New Value** box**.** Click **Add**. Click **Continue.**



Click **OK.**

A new variable RURAL will be added to the dataset. Run frequencies on RURAL to verify that the number of 1's in the RURAL category matches the number of RURAL in the original G8URBAN variable.

RURAL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1015	67.7	67.7	67.7
	1.00	485	32.3	32.3	100.0
	Total	1500	100.0	100.0	

URBANICITY COMPOSITE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	URBAN	394	26.3	26.3	26.3
	SUBURBAN	621	41.4	41.4	67.7
	RURAL	485	32.3	32.3	100.0
	Total	1500	100.0	100.0	

STEP II: PRELIMINARY ANALYSES

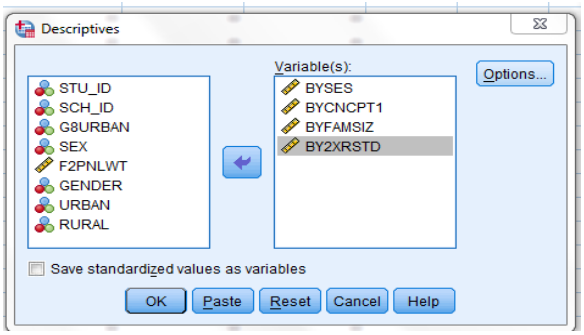
Dependent Variable: BY2XRSTD

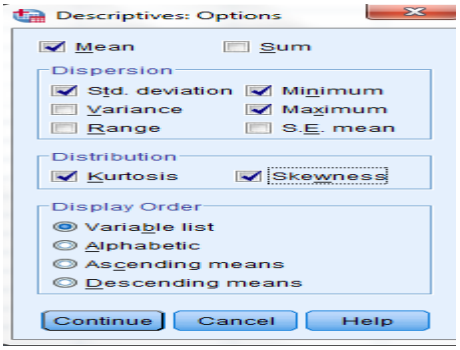
Independent Variables: BYSES, BYFAMSIZ, BYCNCPT1, GENDER, URBAN, RURAL

****NOTE: The list of independent variables does not include SEX and G8URBAN variables. This is because we replaced them with the recoded and dichotomous variables****

Examine Descriptive Statistics

In conducting preliminary analyses, the first step is to examine various descriptive statistics of the continuous variables (i.e., BY2XRSTD, BYSES, BYFAMSIZ, and BYCNCPT1). On **Analyze** menu, select **Descriptive Statistics**. Click **Descriptives.** Click on BY2XRSTD, BYSES, BYFAMSIZ, and BYCNCPT1, one at a time, and click  to add them to the **Variables** box. Click **Options.** Check Mean, Std. deviation, Minimum, Maximum, Kurtosis, and Skewness. Click **Continue.** Click **OK.**





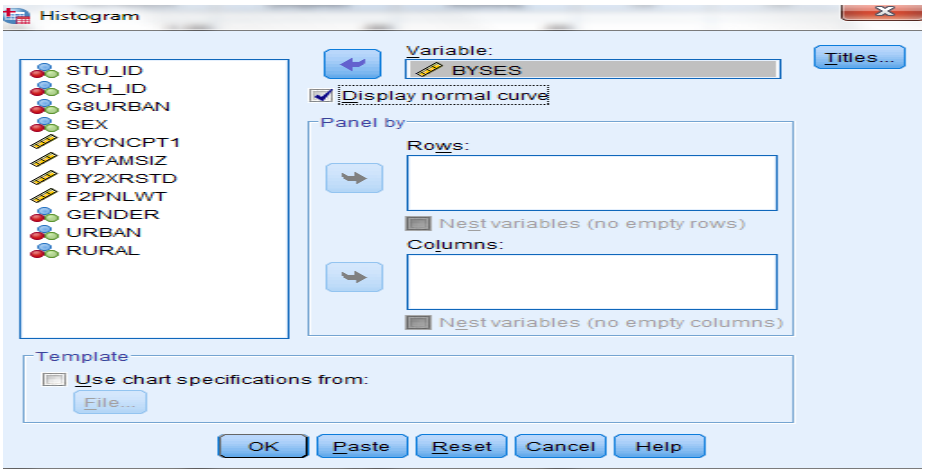
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
SOCIO-ECONOMIC STATUS COMPOSITE	1500	-2.41	1.85	-.0317	.78965	-.049	.063	-.510	.126
SELF CONCEPT 1	1500	-3.09	1.12	.0094	.72266	-.565	.063	.819	.126
FAMILY SIZE	1500	2	10	4.65	1.437	1.035	.063	1.603	.126
READING STANDARDIZED SCORE	1500	32.01	70.55	51.2095	10.26650	.215	.063	-1.051	.126
Valid N (listwise)	1500								

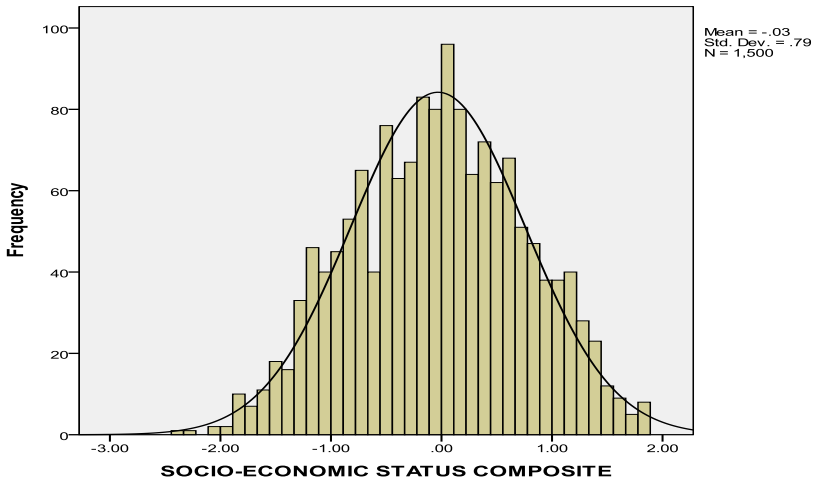
The values for Skewness and the Kurtosis indices are very small which indicates that the variables most likely do not include influential cases or outliers.

Check Normality Assumption

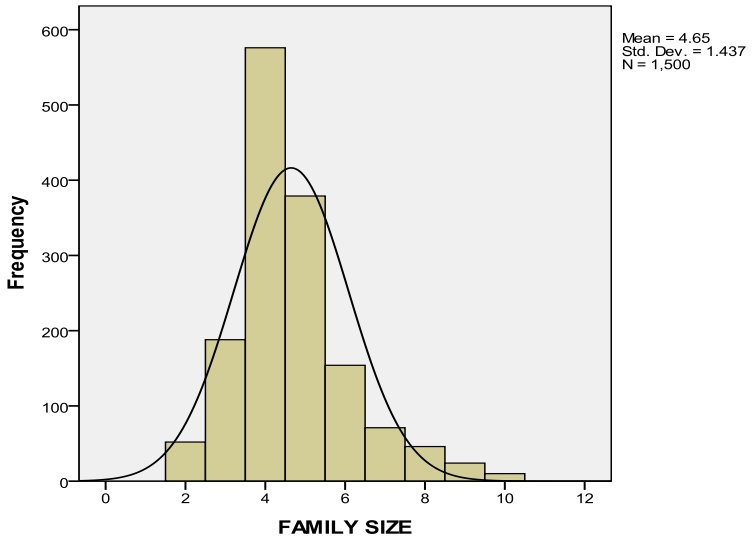
We examine the distribution of the independent variables measured at the interval/ratio-levels and dependent variables to check the normality assumption as the second step. To check BYSES, click **Graphs** menu. Select **Legacy Dialogs**. Click **Histograms.** Click and move BYSES to the **Variable** box. Check **Display normal** curve. Click **OK.**



As the graph shows, BYSES is normally distributed.

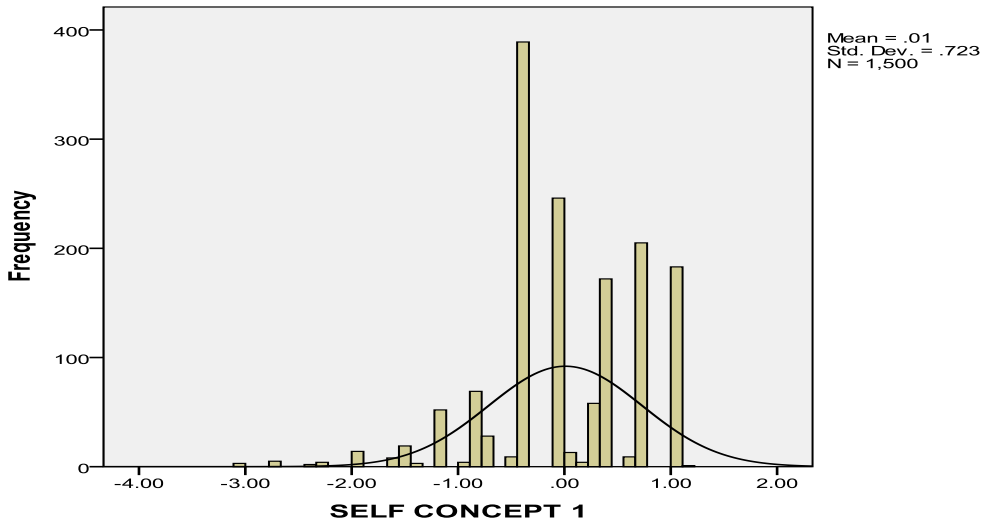


To check BYFAMSIZ, click **Graphs** menu. Select **Legacy Dialogs**. Click **Histograms.** Click **Reset** to clear the variable box. Click and move BYFAMSIZ to the **Variable** box. Check **Display normal** curve. Click **OK.**



Although BYFAMSIZ is slightly skewed in a positive direction, the normality assumption is not violated.

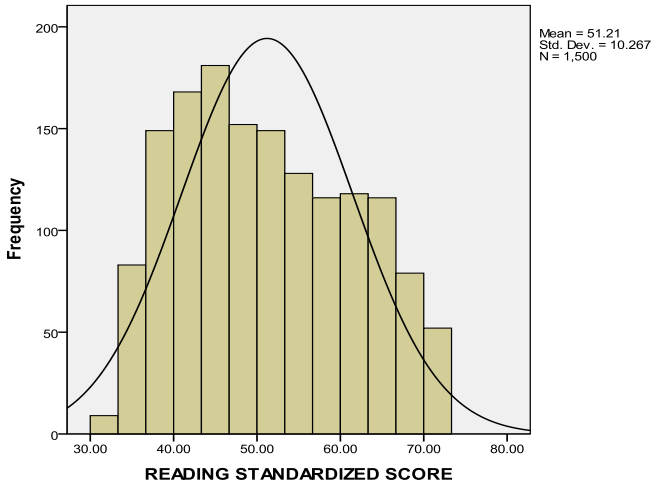
To check BYCNPT1, click **Graphs** menu. Select **Legacy Dialogs**. Click **Histograms.** Click **Reset** to clear the variable box. Click and move BYCNCPT1 to the **Variable** box. Check **Display normal** curve. Click **OK.**



Sometimes the way in which SPSS determines the number of intervals for histograms results in the creation of a histogram with a shape that is difficult to interpret (as in the preceding histogram for self concept). When this happens, we can manually adjust the number of intervals to improve interpretation. Between 15 and 20 intervals is recommended for a data set with more than 250 cases. The following revised histogram includes 15 intervals.

As the graph indicates, self concept is negatively skewed. The distribution of self-concept scores is not normal and thus violates the normality assumption.

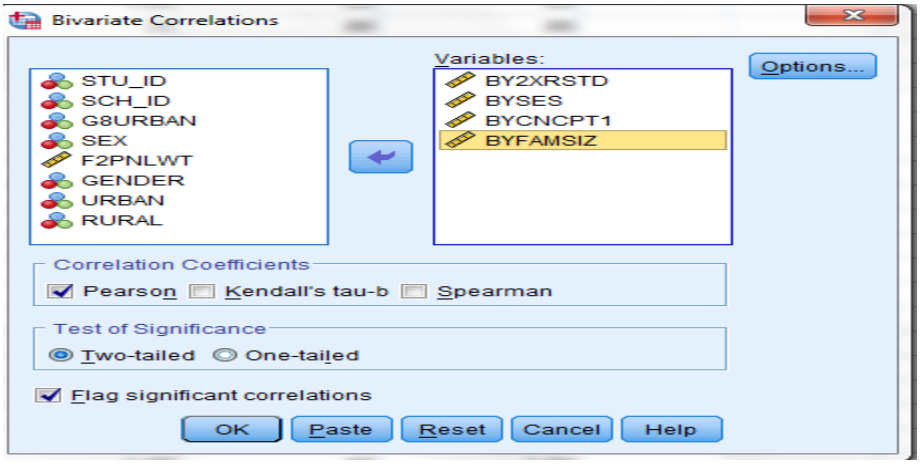
To check BY2XRSTD, click **Graphs** menu. Select **Legacy Dialogs**. Click **Histograms.** Click **Reset** to clear the variable box. Click and move BY2XRSTD to the **Variable** box. Check **Display normal** curve. Click **OK.** The number of intervals for this histogram has also been adjusted to 15 using the same procedure as self concept.



Reading Standardized score is slightly skewed in a positive direction but not enough to violate the normality assumption.

Check Linearity Assumption:

Another linear regression assumption is that the relationship between the dependent and independent variables is linear. We can check this assumption by examining scatterplots of the dependent and independent variables. First, we calculate Pearson correlation coefficients to examine relationships between the DV and the IVs measured at the interval/ratio-levels to check an indication of the magnitude of the relationship between variable pairs. Click **Analyze**. Select **Correlate**. Click **Bivariate**. Click and move the dependent and the continuous independent variables to the **Variables** box. Check **Pearson**. Select **Two-tailed significance**. Check **Flag significant correlations.** Click **OK.**



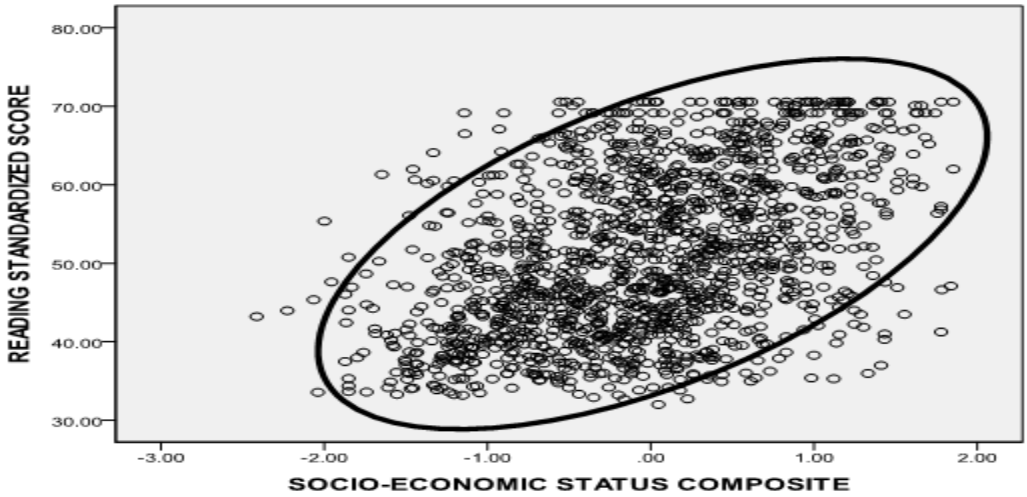
Correlations

		READING STANDARDIZED SCORE	SOCIO-ECONOMIC STATUS COMPOSITE	SELF CONCEPT 1	FAMILY SIZE
READING STANDARDIZED SCORE	Pearson Correlation	1	.453**	.099**	-.070**
	Sig. (2-tailed)		.000	.000	.007
	N	1500	1500	1500	1500
SOCIO-ECONOMIC STATUS COMPOSITE	Pearson Correlation	.453**	1	.053*	-.111**
	Sig. (2-tailed)	.000		.041	.000
	N	1500	1500	1500	1500
SELF CONCEPT 1	Pearson Correlation	.099**	.053*	1	-.037
	Sig. (2-tailed)	.000	.041		.154
	N	1500	1500	1500	1500
FAMILY SIZE	Pearson Correlation	-.070**	-.111**	-.037	1
	Sig. (2-tailed)	.007	.000	.154	
	N	1500	1500	1500	1500

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

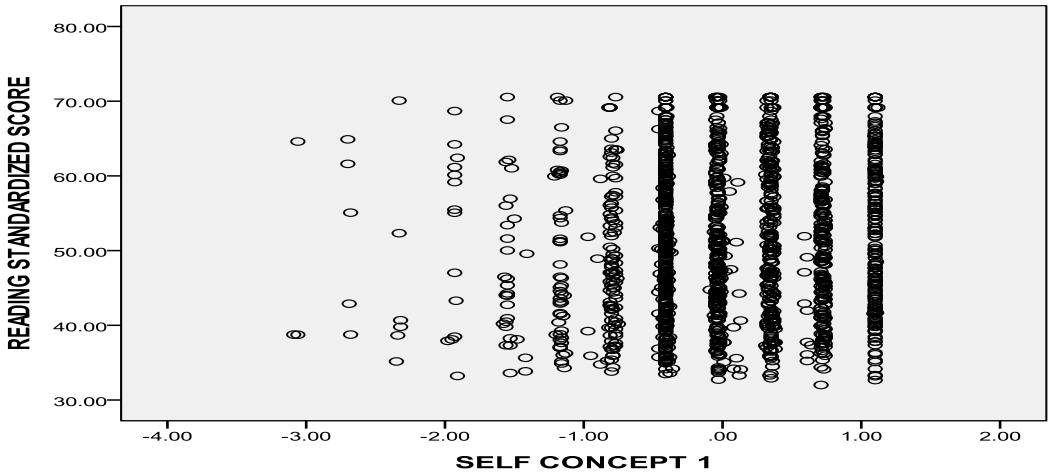
The results indicate a moderate positive correlation between socio economic status and reading scores; a weak positive correlation between self concept and reading scores, and a weak negative correlation between family size and reading scores.

To create a scatter diagram of BY2XRSTD by BYSES, click **Graphs** menu. Select **Legacy Dialogs**. Click **Scatter/Dots.** Select **Simple scatter**. Click **Define**. Click and move BY2XRSTD to the **Y Axis.** Click and move BYSES to the **X Axis**. Click **OK.**



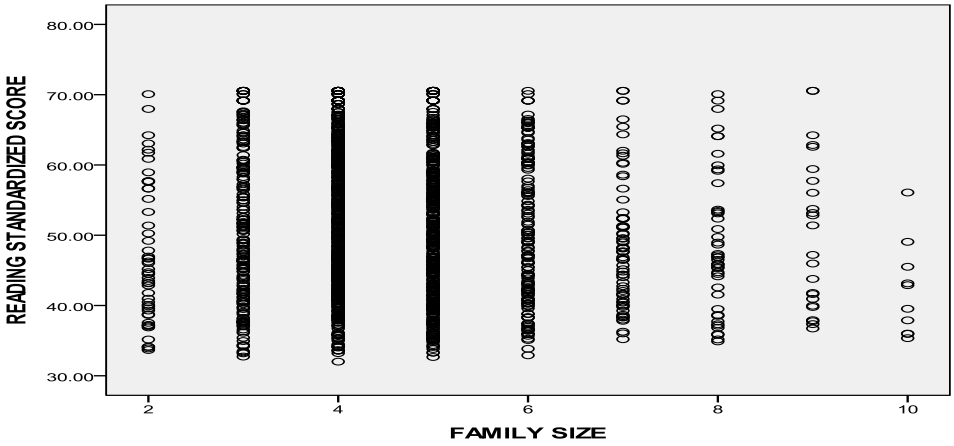
The graph shows a positive linear relationship between reading scores and socio economic status. The correlation between the two variables is significant so we can conclude that there is a linear relationship between BY2XRSTD and BYSES, thus not violating the linearity assumption.

To create a scatter diagram of BY2XRSTD by BYCNPT1, click **Graphs** menu. Select **Legacy Dialogs**. Click **Scatter/Dots.** Select **Simple scatter**. Click **Define**. Click **Reset** to clear the axes boxes. Click and move BY2XRSTD to the **Y Axis.** Click and move BYCNCPT1 to the **X Axis**. Click **OK.**



Although this graph suggests that the linearity assumption may be violated here, we will keep self concept in the model because its correlation with reading scores is significant. However, this violation must be noted as a limitation of the model.

To create a scatter diagram of BY2XRSTD by BYFAMSIZ, click **Graphs** menu. Select **Legacy Dialogs**. Click **Scatter/Dots.** Select **Simple scatter**. Click **Define**. Click **Reset** to clear the axes boxes. Click and move BY2XRSTD to the **Y Axis.** Click and move BYFAMSIZ to the **X Axis**. Click **OK.**



This graph suggests that the linearity assumption may be violated here as well. In fact, the graph shows similar reading scores among students in families with two to seven family members. The reading scores of students from families with more than seven members appear to be lower. We could recode family size to a dichotomous variable (0= less than 8 family members, 1=8 or more family members) as an alternative way to evaluate its impact. This violation must also be noted as a limitation of the model.

Based on the outcome of this assessment, we retain all continuous independent variables in the model because they are significantly correlated with the dependent variable.

STEP III: MULTIPLE LINEAR REGRESSION ANALYSIS

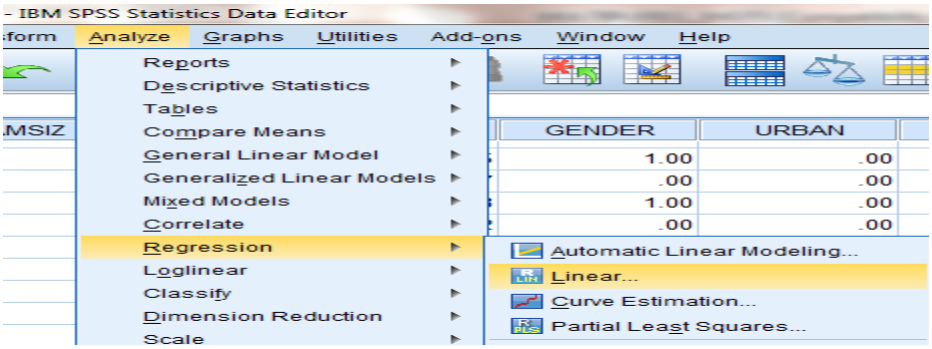
Dependent Variable: BY2XRSTD

Independent Variables: BYSES, BYFAMSIZ, BYCNCPT1, GENDER, URBAN, RURAL

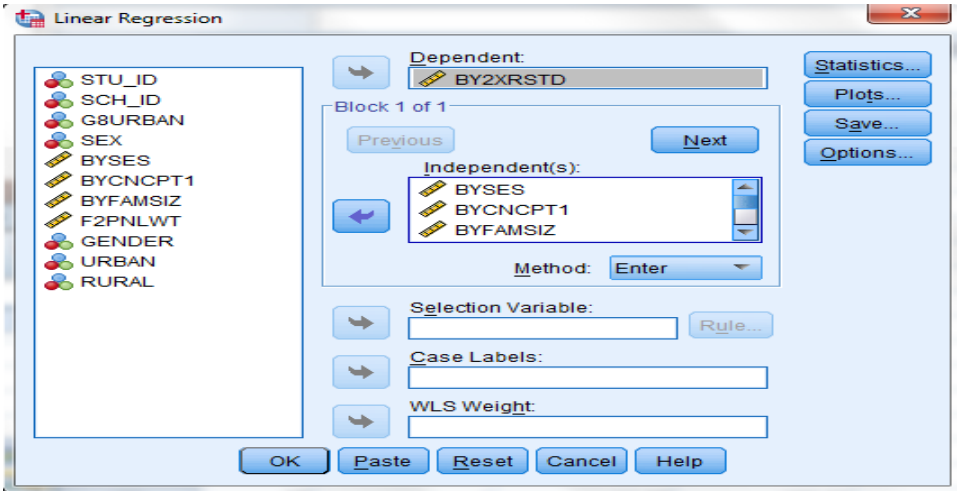
****NOTE: The list of independent variables does not include SEX and G8URBAN variables. This is because we replaced them with the recoded and dichotomous variables****

In addition to providing instructions for running the multiple regression analysis, we also include instructions for checking model requirements (i.e., outliers and influential cases and multicollinearity among the continuous independent variables) and the homogeneity of variance assumption. Recall that outliers and influential cases can negatively affect the fit of the model or the parameter estimates. We calculate residuals and dfbetas using the Influence Diagnostics procedure to check for outliers and influential cases. We calculate the Variance Inflation Factor (VIF) and Tolerance statistics to check for multicollinearity. Finally, we create homogeneity of error variance plots to check the homogeneity of variance. The dataset used to conduct these analyses is included on the companion website. It is entitled, REGRESSION.SAV.

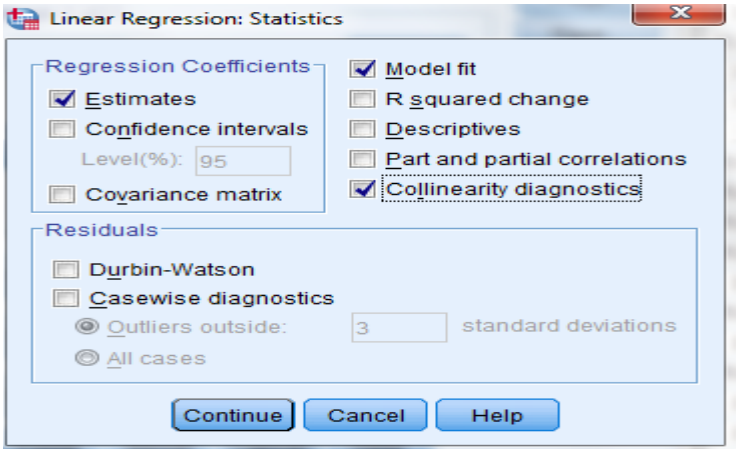
The first step is to enter the dependent and independent variables into the SPSS model. Click on **Analyze** in the menu bar at the top of the data view screen. A drop down menu will appear. Move your cursor to **Regression**. Another drop down menu will appear to the right of the first menu. Click on **Linear** to open the Linear Regression dialog box.



Click to highlight BY2XRSTD in the box on the left side. Click the top arrow to move it to the Dependent box. Click to highlight BYSES, BYFAMSIZ, BYCNCPT1, GENDER, URBAN, and RURAL and move each of them one at a time to the Independent**(**s) box.

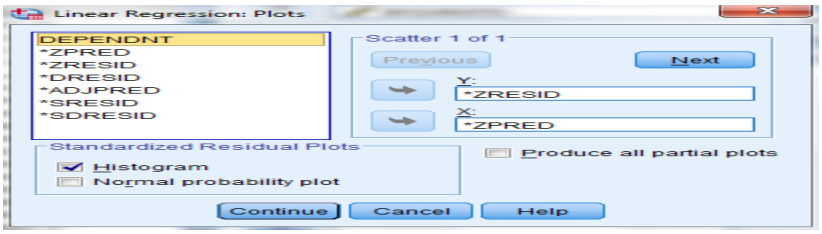


Next, we program SPSS to provide a check for multicollinearity. Click **Statistics** in the upper right corner to open the Linear Regression: Statistics dialogue box. Make sure that **Estimates**, **Model fit** and **Collinearity Diagnostics** are checked. Click **Continue** to return to the Linear Regression dialogue box**.**

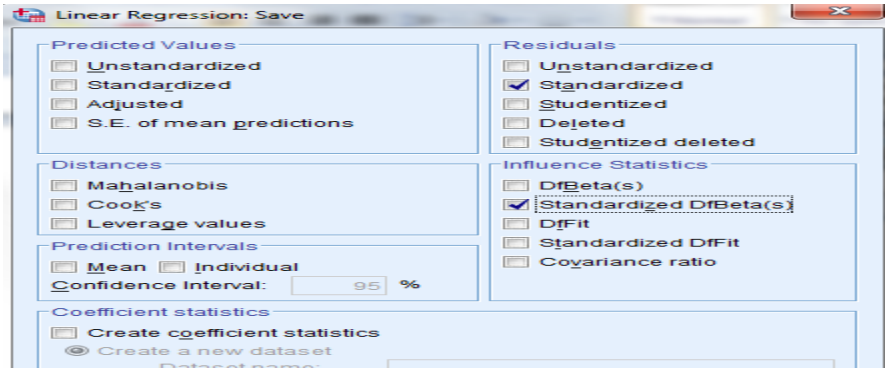


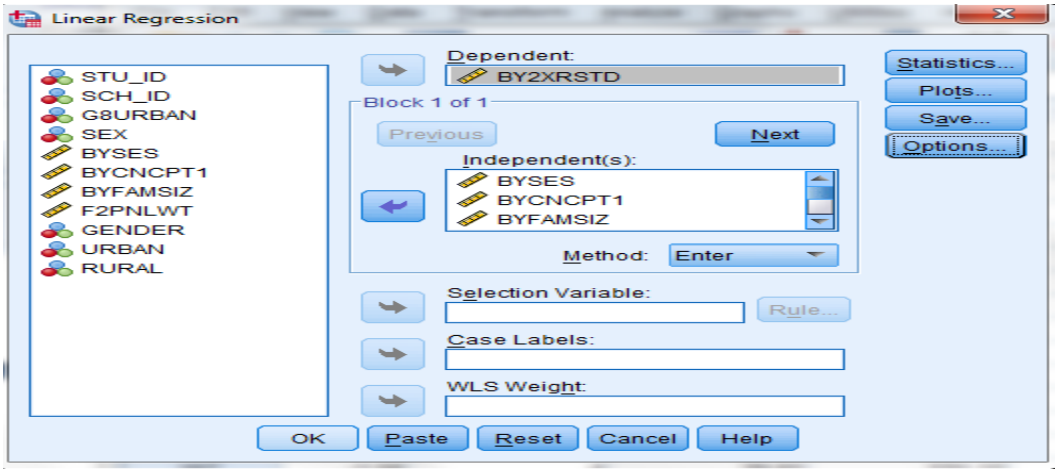
To create residual plots to check the homogeneity of variance and normality assumptions,

click **Plots** just below **Statistics** to open the Linear Regression: Plots dialogue box. Click and move "*ZRESID" to the Y box. Click and move "*ZPRED" to the X box. Check "Histogram"**.** Click **Continue** to return to the Linear Regression dialogue box**.**



To check for outliers and influential cases, click **Save** just below **Plots** to open the Linear Regression: Save dialogue box. Under Residuals, check "Standardized**"**. Under Influence Statistics, check "Standardized DfBetas". Click **Continue** to return to the Linear Regression dialogue box**.**





Click OK at the bottom of the Linear Regression dialogue box to run the multiple linear regression analysis.

We now examine the output, including findings with regard to multicollinearity, whether the model should be trimmed (i.e., removing insignificant predictors), violation of homogeneity of variance and normality assumptions, and outliers and influential cases.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.474[a]	.225	.222	9.05675

a. Predictors: (Constant), RURAL, FAMILY SIZE, GENDER, SELF CONCEPT 1, SOCIO-ECONOMIC STATUS COMPOSITE, URBAN
b. Dependent Variable: READING STANDARDIZED SCORE

The R^2 is .225. This means that the independent variables explain 22.5% of the variation in the dependent variable.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35533.450	6	5922.242	72.201	.000[a]
	Residual	122462.788	1493	82.025		
	Total	157996.238	1499			

a. Predictors: (Constant), RURAL, FAMILY SIZE, GENDER, SELF CONCEPT 1, SOCIO-ECONOMIC STATUS COMPOSITE, URBAN
b. Dependent Variable: READING STANDARDIZED SCORE

The p value for the F statistic is < .05. This means that at least one of the independent variables is a significant predictor of the DV (standardized reading scores). The "Sig." column in the Coefficients table shows which variables are significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.319	.869		61.375	.000		
	SOCIO-ECONOMIC STATUS COMPOSITE	5.820	.304	.448	19.121	.000	.947	1.056
	SELF CONCEPT 1	1.333	.329	.094	4.057	.000	.970	1.031
	FAMILY SIZE	-.111	.164	-.016	-.678	.498	.985	1.015
	GENDER	-2.341	.475	-.114	-4.931	.000	.972	1.029
	URBAN	-.393	.584	-.017	-.674	.500	.829	1.207
	RURAL	-.598	.568	-.027	-1.070	.285	.802	1.247

a. Dependent Variable: READING STANDARDIZED SCORE

Unstandardized coefficients are used in the prediction and interpretation

Standardized coefficients are used for comparing the effects of the independent variables

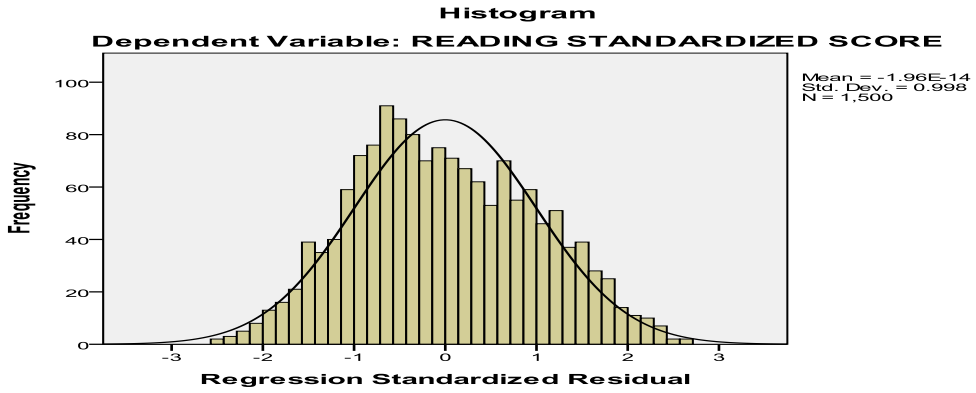
If Sig<.05 then that variable is significant and should be kept. If Sig>.05 then the variable is not a good predictor and can be removed from the model

If VIF > 3 there could be multicolinearity problems.

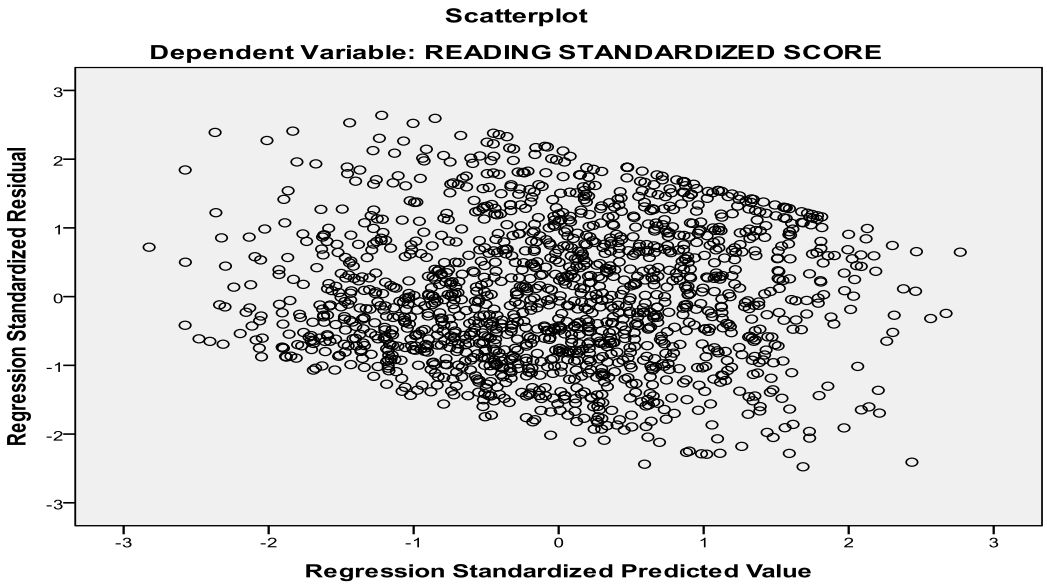
It appears **multicollinearity** is not a concern because the VIF scores are less than 3.

In terms of **model trimming**, the results also show that BYFAMSIZ, URBAN, RURAL are not significant predictors of the standardized reading scores. We will remove these variables from the model and rerun the analysis.

The histogram of residuals allows us to check the extent to which the **residuals are normally distributed**. The residuals histogram shows a fairly normal distribution. Thus, based on these results, the normality of residuals assumption is satisfied.

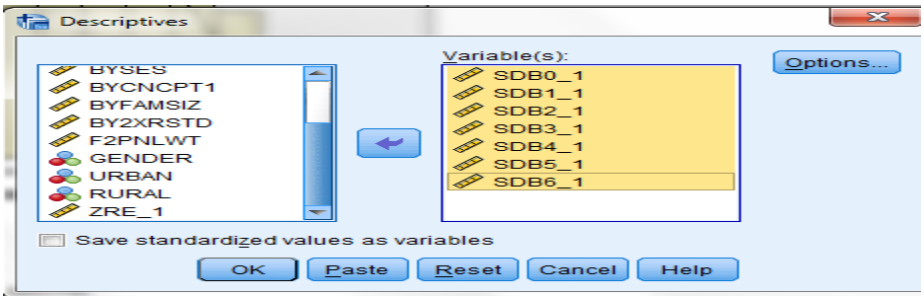


We examine a scatter plot of the residuals against the predicted values to evaluate whether the **homogeneity of variance** assumption is met. If it is met, there should be no pattern to the residuals plotted against the predicted values. In the following scatter plot, we see a slanting pattern, which suggests heteroscedasticity, (i.e., violation of the homogeneity of variance assumption).



Finally, we examine the values of the standardized DfBetas and standardized residual values to identify **outliers and influential cases.** Large values suggest outliers or influential cases. Note that the results thus far (histograms and scatter plots of the continuous variables and residuals) showed no data point(s) that stood out as outliers. Thus, it is unlikely that we will find large standardized DfBetas or standardized residual values. Nonetheless, the standardized DfBeta values can verify this. The values of the standardized DfBetas have been added as seven additional variables in the data set (SDB0_1 – SDB6_1). Outliers or influential cases have large (< -2 or >2) standardized DfBetas. Instead of manually scrolling through the values of each variable to check this, we can calculate maximum and minimum values.

Click **Analyze** in the menu bar at the top of the data view screen. A drop down menu will appear. Move your cursor to **Descriptive Statistics**. Another drop down menu will appear to the right of the first menu. Click on **Descriptives** to open the Descriptives dialog box. Highlight the seven SDB variables and move them to the **Variables** box. Click **OK.**



The results show no standardized Dfbeta values < -2 or > 2**.** We can conclude that the dataset does not include outliers or influential cases.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Standardized DFBETA Intercept	1500	-.13699	.17513	.0000011	.02631169
Standardized DFBETA BYSES	1500	-.16348	.07291	-.0000033	.02493608
Standardized DFBETA BYCNCPT1	1500	-.22988	.14657	-.0000047	.02604153
Standardized DFBETA BYFAMSIZ	1500	-.18559	.18406	-.0000011	.02655174
Standardized DFBETA GENDER	1500	-.07051	.07862	.0000013	.02593275
Standardized DFBETA URBAN	1500	-.09200	.09929	-.0000005	.02566111
Standardized DFBETA RURAL	1500	-.11076	.08567	-.0000012	.02631213
Valid N (listwise)	1500				

A copy of the output for the analyses we just conducted is provided on the companion website. It is entitled, Multiple Linear Regression Results_prelim.spv.

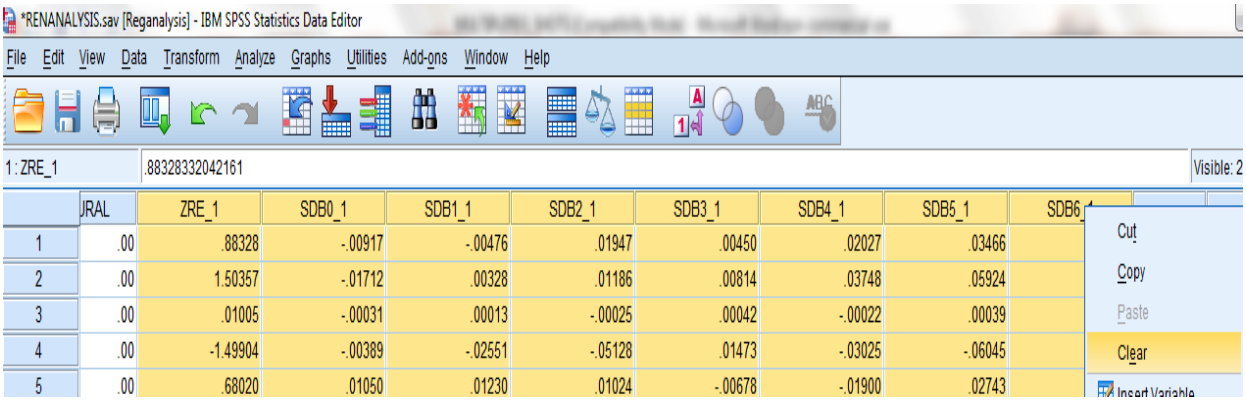
Fitting a Final Model

Given that BYFAMSIZ, URBAN, RURAL are not significant we remove them from the analysis and refit the model. The revised list variables are:

Dependent Variable: BY2XRSTD

Independent Variables: BYSES, BYCNCPT1, GENDER

First, we must delete the Residual and DfBeta variables from the dataset. This can easily be done by highlighting them, right clicking your mouse, and then clicking **Clear.**



Next, we rerun the model following the same steps used in conducting the preliminary analyses. Click **Analyze** in the menu bar at the top of the data view screen. A drop down menu will appear. Move your cursor to **Regression**. Another drop down menu will appear to the right of the first menu. Click **Linear** to open the Linear Regression dialog box. Click to highlight BY2XRSTD in the box on the left side. Click the top arrow to move it to the Dependent box. Click to highlight BYSES, BYCNCPT1, and GENDER and move each of them one at a time to the Independent**(**s) box.

Click **Statistics** in the upper right corner to open the Linear Regression: Statistics dialogue box. Make sure that **Estimates** and **Model fit** are checked. Click **Continue** to return to the Linear Regression dialogue box**.**

To check the homogeneity of variance and normality assumptions, click **Plots** just below **Statistics** to open the Linear Regression: Plots dialogue box. Click and move "*ZRESID" to the Y box. Click and move "*ZPRED" to the X box. Check "Histogram"**.** Click **Continue** to return

to the Linear Regression dialogue box**.**

Click **OK** at the bottom of the Linear Regression dialogue box to run the revised analyses. The output is provided as follows:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.473[a]	.224	.222	9.05271

a. Predictors: (Constant), GENDER, SOCIO-ECONOMIC STATUS COMPOSITE, SELF CONCEPT 1
b. Dependent Variable: READING STANDARDIZED SCORE

The R^2 = .224. This means that the independent variables explain 22.4% of the variation in the dependent variable. This value almost the same as the R^2 value from the preliminary model. This confirms that the variables removed from the preliminary model were useless in predicting reading scores.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35396.593	3	11798.864	143.974	.000[a]
	Residual	122599.645	1496	81.952		
	Total	157996.238	1499			

a. Predictors: (Constant), GENDER, SOCIO-ECONOMIC STATUS COMPOSITE, SELF CONCEPT 1
b. Dependent Variable: READING STANDARDIZED SCORE

The p value for F statistic is < .05. This means that at least one independent variable is a significant predictor of reading scores. The Sig. column in the Coefficients table shows which variables are significant.

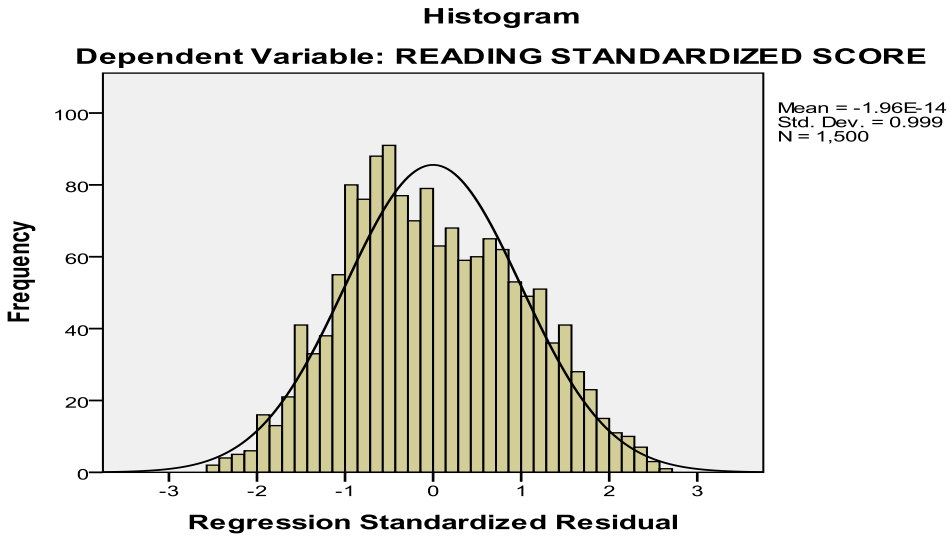
Coefficients[a]

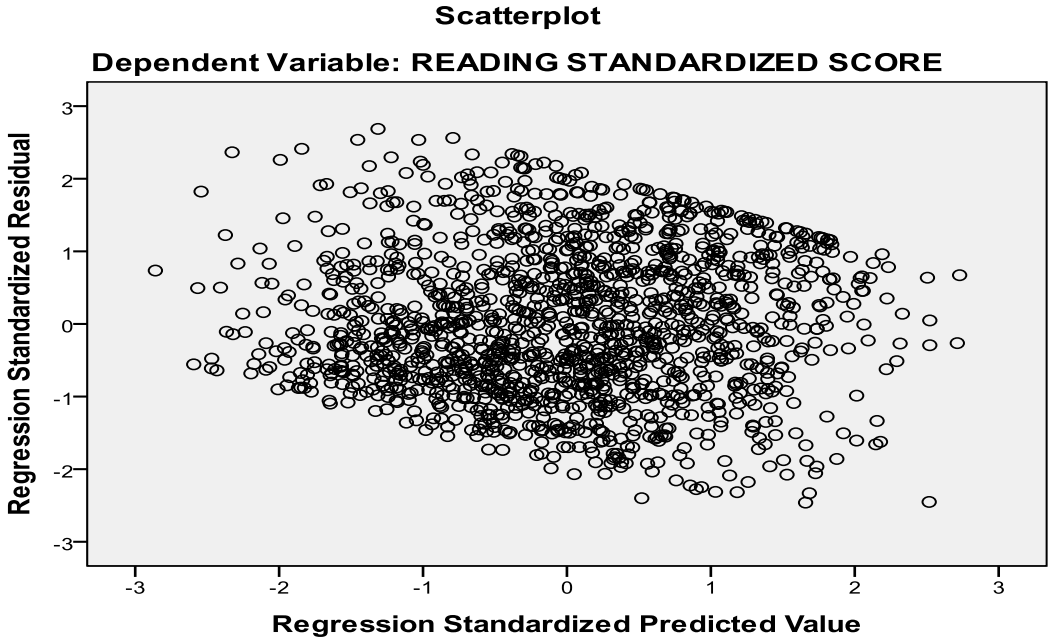
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	52.510	.326		160.878	.000		
	SOCIO-ECONOMIC STATUS COMPOSITE	5.897	.297	.454	19.875	.000	.996	1.004
	SELF CONCEPT 1	1.336	.328	.094	4.069	.000	.971	1.029
	GENDER	-2.346	.474	-.114	-4.945	.000	.972	1.029

a. Dependent Variable: READING STANDARDIZED SCORE

All Sig<.05 means all the predictors in the model are significant and should be retained.

Plots of residuals and homogeneity of error variance look identical to the plots from the preliminary model, indicating that the normality of residuals assumption is met but the homogeneity of variance assumption is not met.





Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52.510	.326		160.878	.000
	SOCIO-ECONOMIC STATUS COMPOSITE	5.897	.297	.454	19.875	.000
	SELF CONCEPT 1	1.336	.328	.094	4.069	.000
	GENDER	-2.346	.474	-.114	-4.945	.000

a. Dependent Variable: READING STANDARDIZED SCORE

We examine the coefficients table to examine and interpret the results. The prediction equation is based on the unstandardized coefficients, as follows:

$BY2XRSTD_i = 52.51 + 5.897\ BYSES_i + 1.336\ BYCNCPT1_i - 2.346\ GENDER_i$, where i=1….1500 and GENDER= 1 for Males and 0 for Females.

We can use the unstandardized coefficients to interpret the results.

- The **Constant** is the predicted value of the dependent variable when all of the independent variables have a value of zero. In the context of this analysis, the predicted

reading score for females with zero self-concept and zero socio-economic status score is 52.51.

- The slope of **socio-economic status** (BYSES) is 5.897. This means that for every one unit increase in socio economic status, predicted reading scores increase by 5.897 units, after controlling for self concept and gender.
- The slope of **self concept** (BYCNCPT1) is 1.336. This means that for every one unit increase in self concept, predicted reading scores increase by 1.336, after controlling for socio-economic status and gender.
- The slope of **gender** (GENDER) is -2.346. This means that, on average, predicted reading scores for males are 2.346 points lower than for females, after controlling for socio-economic status and self concept.