CHAPTER TWO

REGRESSION

2.0 Introduction

The second chapter, Regression analysis is an extension of correlation. The aim of the discussion of exercises is to enhance students' capability to assess the effect of one variable *x* on another variable *y* which is known as the bivariate regression. Meanwhile students will also be exposed to the discussion of multiple regressions; that is the effect of several variables.

Regression statistical techniques attempt to investigate the best way of describing the relationship between a dependent variable and one or more independent variables. The regression line represents the best fit straight line through a set of coordinates *X* and *Y*.

Generally independent variable is also known as predictor variable will be assigned as *X* variable, whereas dependent variable will be assigned as *Y*. When explaining the relationship between *X* and *Y*, it is often said as *X* predicting *Y*. The mathematical formula is as follows:

$Y = a + bX$

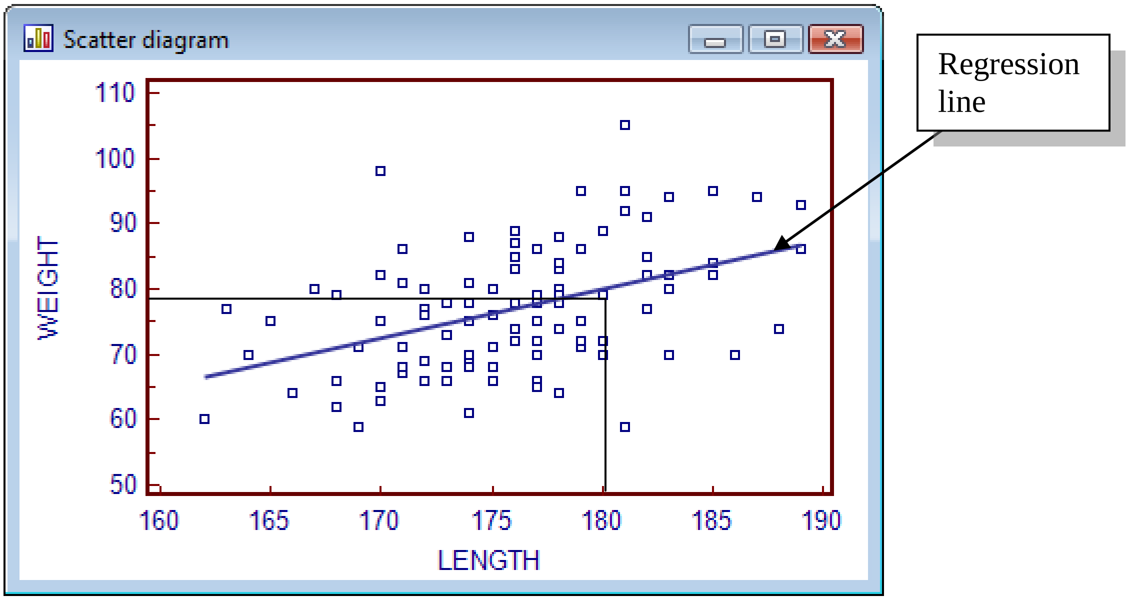
Where

Y = predicted value for dependent variable Y
a = value of Y intercept (point cut at Y axis)
b = regression coefficient (gradient of the line)
X = value for independent or predictor variable X

Regression Line

Regression is used to make predictions based on linear relationship. It is a statistical technique for finding the best-fitting straight line for a set of data. The resulting straight line is called regression line. An example of a regression line is shown in Figure 2.1 (Medcalc, 2009).

.



Regression line

Figure 2.1: Regression Line (Medcalc, 2009)

2.1 Characteristics of Regression Line

Based on Figure 2.1, the following characteristics of the regression line can be observed:

- The line makes the relationship between length and weight easier to see.
- The line identifies the center, or central tendency, of the relationship, just as the mean describes central tendency for a set of scores. Therefore, regression line provides a simplified description of the relationship. Even when all the data points on the scatterplot were removed, the regression line will still give a general picture of the relationship between length and weight.
- The regression line can be used to make prediction. According to Gravetter (2005), the line establishes a precise, one-to-one relationship between each X value (length) and a corresponding Y value (weight). For example, in the scatterplot above, when $X = 180$, Y can be predicted as 79 with the presence of the regression line.

However, regression line does not have to be drawn on a graph; it can be presented in a simple equation $Y = bX + a$, where b and a are fixed constants (Gravetter & Wallnau, 2005).

2.2 Types of Regression

There are 2 types of regression analysis, Bivariate Regression and Multiple Regression. Bivariate Regression involves analyzing the relationship between the dependent variable and one independent variable. Multiple Regression involves the relationship between dependent variable and more than one independent variable (such as $X_{1,}$ $X_{2,}$ $X_{3,}$ etc).

2.2.1 Bivariate Regression

Bivariate Regression is used to examine the relationship between two variables (*X*) and (*Y*). The results indicated in the study of regression are then used to make predictions. In other words, we can use regression for a study when we have knowledge of one variable, while trying to predict the other.

For discussion purposes, let's take example of a study to find out whether the increase in sales of a product is due to the recent advertising on radio. In order to predict the result of this study, regression is an appropriate approach to use. In this situation, we do have knowledge of one variable – sales of the product have increased. However, what we do not know is the reasons behind the increase in sales. As such, we can test and predict if the increase of sales was due to the recent radio advertising. Nevertheless, the results may also show otherwise.

The two variables mentioned earlier ('*X*' and '*Y*'), one represents an independent variable while the other is an independent variable. An independent variable is a factor which is selected by the researcher of which he/she has control of. A dependent variable is the result of which the researcher wants to find.

As mentioned earlier, the formula for linear regression is $Y = a + bX$, whereby *X* is the independent variable and *Y* is the dependent variable. The slope of the line is *b*, and *a* is the point where *X* and *Y* intercept (the value of *Y* when $X = 0$).

Example:
The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (*Y*) can be predicted from the entrance test score (*X*). The results of the study are as follow:

Table 2.1: Entrance Test Score and GPA

Entrance Test Score (X)	GPA (Y)	XY	X^2
5.50	3.10	17.05	30.25
4.80	2.30	11.04	23.04
4.70	3.00	14.10	22.09
3.90	1.90	7.41	15.21
4.50	2.50	11.25	20.25
6.20	3.70	22.94	38.44
6.00	3.40	20.40	36.00
5.20	2.60	13.52	27.04
4.70	2.80	13.16	22.09
4.30	1.60	6.88	18.49
4.90	2.00	9.80	24.01
5.40	2.90	15.66	29.16
5.00	2.30	11.50	25.00
6.30	3.20	20.16	39.69
4.60	1.80	8.28	21.16
4.30	1.40	6.02	18.49
5.00	2.00	10.00	25.00
5.90	3.80	22.42	34.81
4.10	2.20	9.02	16.81
4.70	1.50	7.05	22.09
ΣX=100.00	ΣY=50.00	ΣXY=257.66	ΣX^2 =509.12

Slope (b) $= [N\Sigma XY - (\Sigma X)(\Sigma Y)] / [N\Sigma X^2 - (\Sigma X)^2]$

$= (5153.32-5000) / (10182.4-10000)$
$= 153.32 / 182.4$
$= 0.84$

Intercept (a) $= (\Sigma Y - b(\Sigma X)) / N$
$= (50-0.84(100)) / 20$
$= -34 / 20$
$= -1.7$

Regression Equation $Y = a + bX$
$Y = -1.7 + 0.84X$

Find out the Predicted Grade Point Average (GPA) of a student if the entrance test score is 5.8.

$Y = -1.7 + 0.84X$

$= -1.7 + 0.84(5.8)$
$= -1.7 +4.872$
$= 3.172$

2.2.2 Multiple Regression

Multiple Regression is used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors. It is based on correlation, but allows a more sophisticated exploration of the interrelationship among a set of variables. One shouldn't use multiple regressions as a fishing expedition. You must support your analysis with theoretical reason.

Mathematical formula for multiple regression is as follow:

$Y = a + b_1X_1 + b_2X_2$

Y = predicted value for dependent variable Y
a = value of Y intercept (point cut at Y axis)
b_1= regression coefficient (gradient of the line) for the first independent variable
X_1 = value for first independent or predictor variable X_1
b_2= regression coefficient (gradient of the line) for the second independent variable
X_2 = value for second independent or predictor variable X_2

Multiple regression can tell you how well a set of variables is able to predict a particular outcome. For example, you may be interested how well a set of subscales on predicting the academic performance among students. In addition, it will provide you information about the contribution of total subscales, and the relative contribution of each subscale. Main types of research questions that can be used:

- How well a set of variable is able to predict a particular outcome
- Which variable in a set of variables is the best predictor of an outcome

Assumptions of Multiple Regression

Sample Size

Sample size need to be big enough to be generalized to a bigger population. What should be the ideal sample size? Different authors will have different criteria. For example, Steven (1996), recommends 15 subjects per variable for social science research. Whereas Tabachnick and Fidell (2001) give a formula to calculate the size of sample, $N > 50 + 8m$

(where *m* = number of independent variables). If you have 3 independent variables, you will need 74 subjects in a particular study.

Outliers

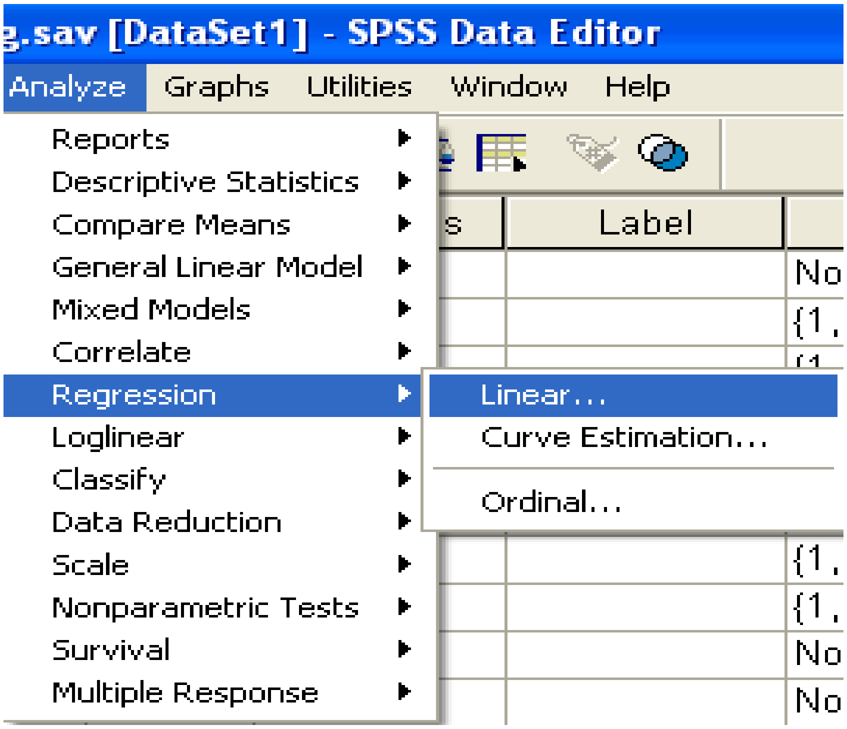
Outliers should be excluded in multiple regression. Extreme scores which are too high or too low should be deleted from both dependent and independent variables.

Multiple Regression analyses

- Standard Multiple Regression: All the independent (or predictor) variables are entered into the equation simultaneously. Each independent variable is evaluated in terms of its predictive power, over and above that offered by all the other independent variables.
- Hierarchical multiple Regression: In hierarchical regression (also called sequential) the independent variables are entered into the equation in the order specified by the researcher based on theoretical grounds.
- Stepwise multiple regression: In stepwise regression the researcher provides SPSS with a list of independent variables and then allows the program to select which variables it will enter, and in which order they go into the equation, based on a set of statistical criteria.
- Multiple regression would be used if you had a set of variables (eg., various personality scales) and wanted to know how much variance in a dependent variable (eg., anxiety) they were able to explain as a group or block.
- Multiple regression approach would also tell you how much unique variance in the dependent variable that each of the independent variables explained.
- Hierarchical multiple regression would be used if you wanted to know how much a variable predicts another variable.
- Once all sets of variables are entered, the overall model is assessed in terms of its ability to predict the dependent measure. The relative contribution of each block of variables is also assessed.
- Stepwise multiple regression would be used when you have a large number of predictor variables.
- Stepwise multiple regression would be used when no underlying theory concerning their possible predictive power.

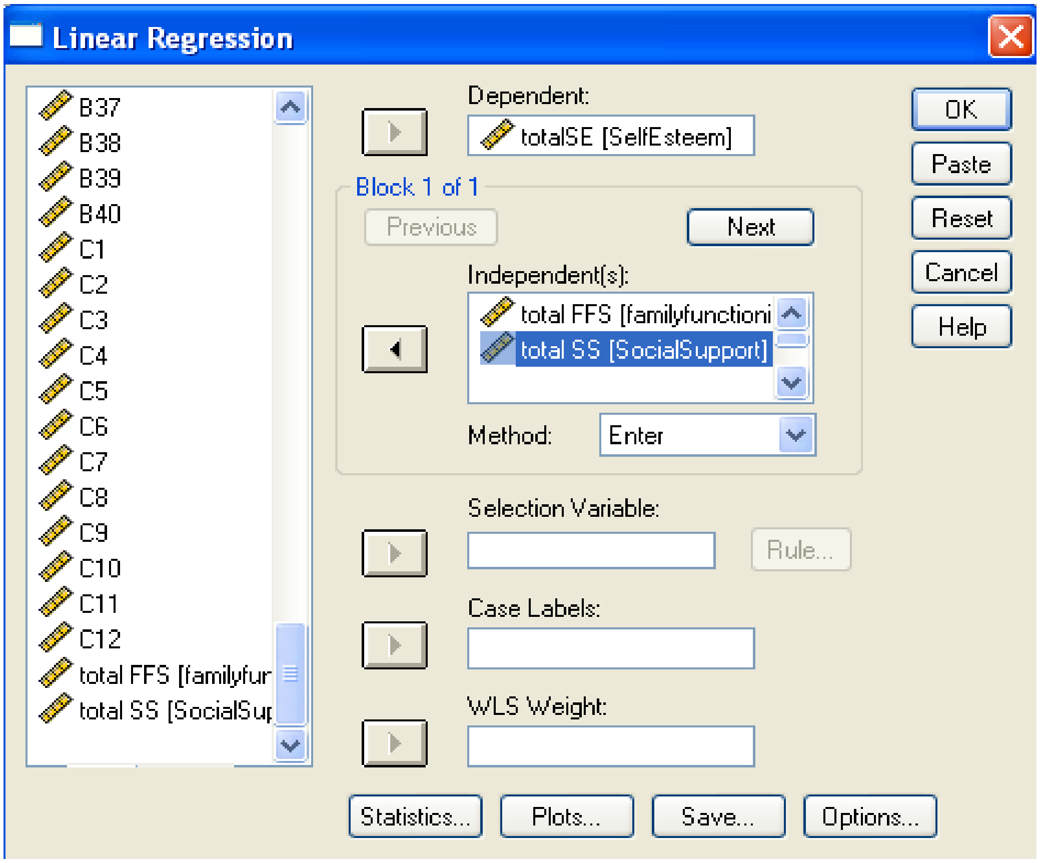
2.3 Procedure for Generating Multiple Regression

- Select **Analyse,** click on **Regression**, then on **Linear.**



- Select dependent variable (e.g Self-esteem) and move it to **Dependent** box..

- Select independent variable (e.g Family Functioning and Social Support) and move them to **Independent** box.



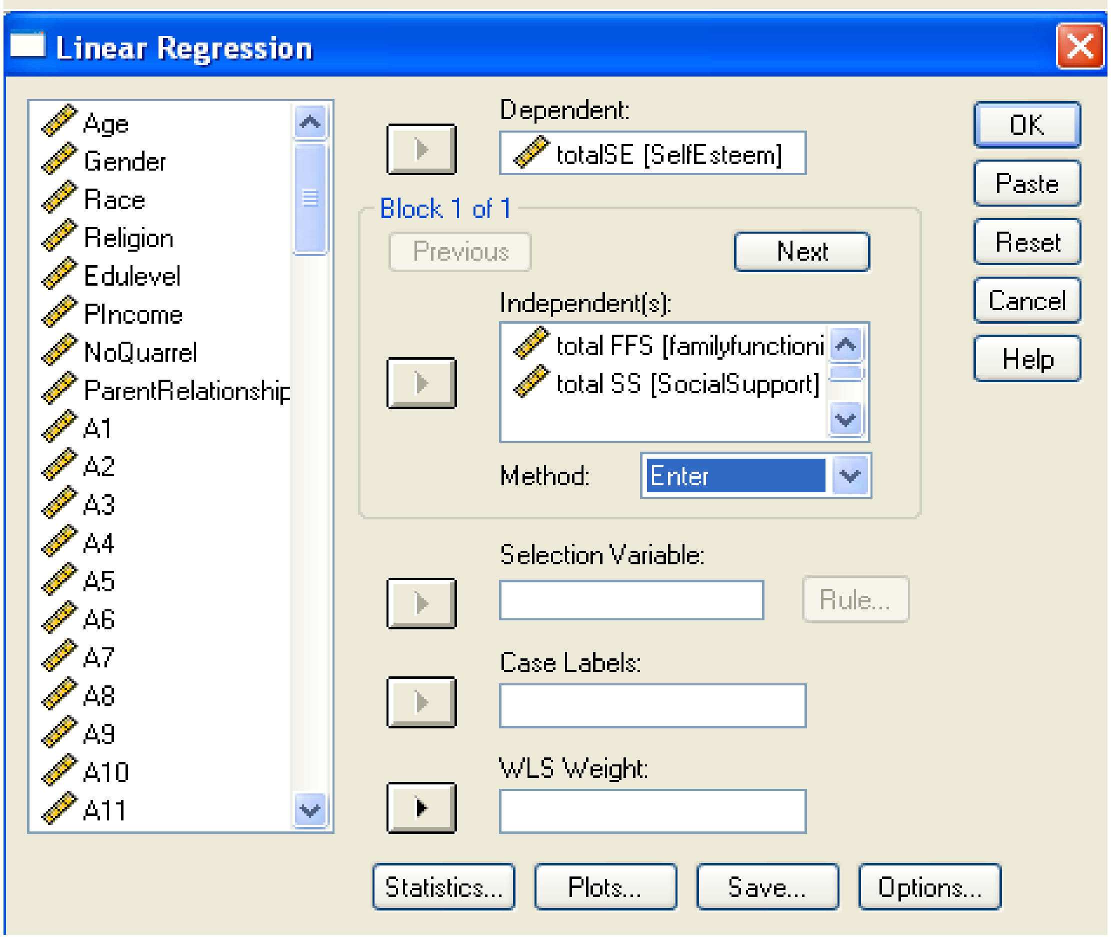
Dependent: totalSE [SelfEsteem]

Independent(s): total FFS [familyfunctioni
total SS [SocialSupport]

Method: Enter

- For **Method**, select **Enter.**

- Click on the **Option** button, in the **Missing Values** section click on **Exclude cases pairwise.**

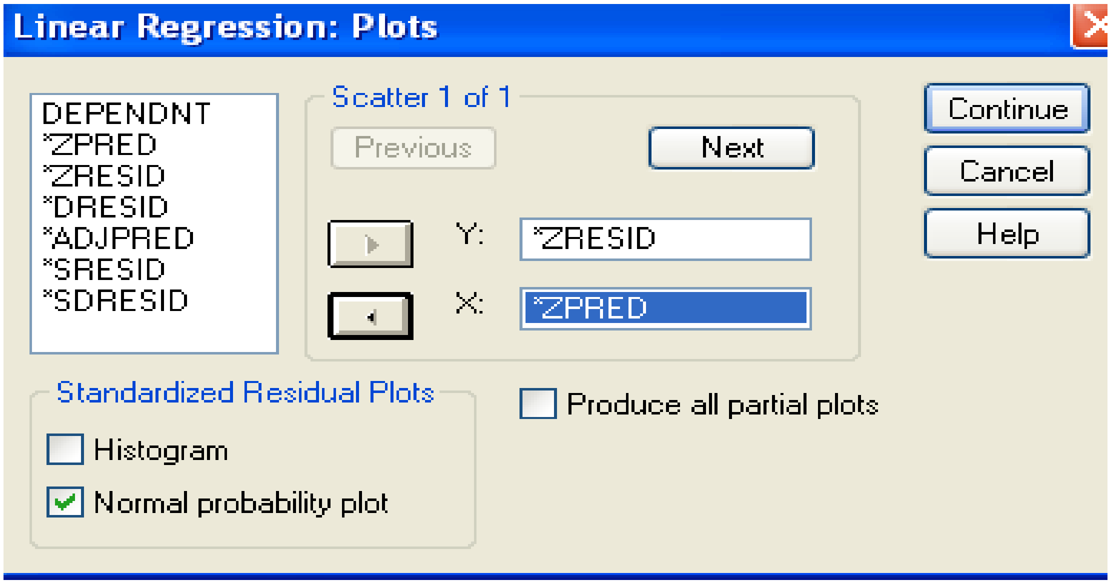


Dependent: totalSE [SelfEsteem]

Independent(s): total FFS [familyfunctioni
total SS [SocialSupport]

Method: Enter

- Click on the **Plots** button.
- Click on **“ZRESID**” and move this into the **Y** box.
- Click on the “**ZPRED**” and move this into the **X** box.
- Select **Standardized Residual Plots**, tick the **Normal probability plot.**
- Click **Continue.**



Linear Regression: Plots

DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

Standardized Residual Plots

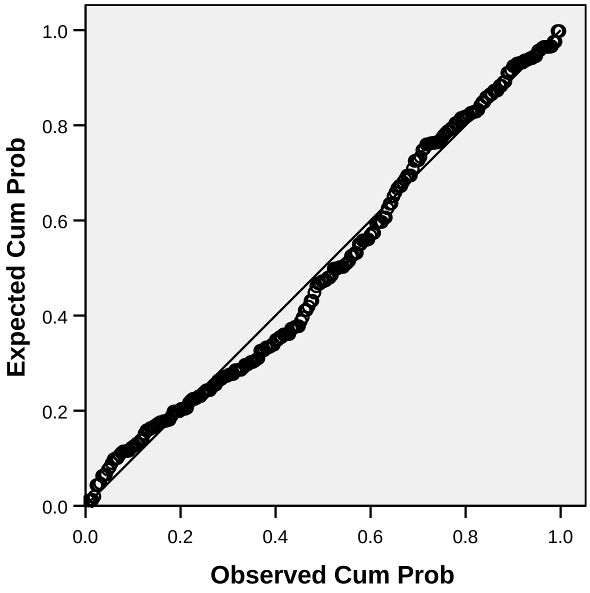
□ Histogram

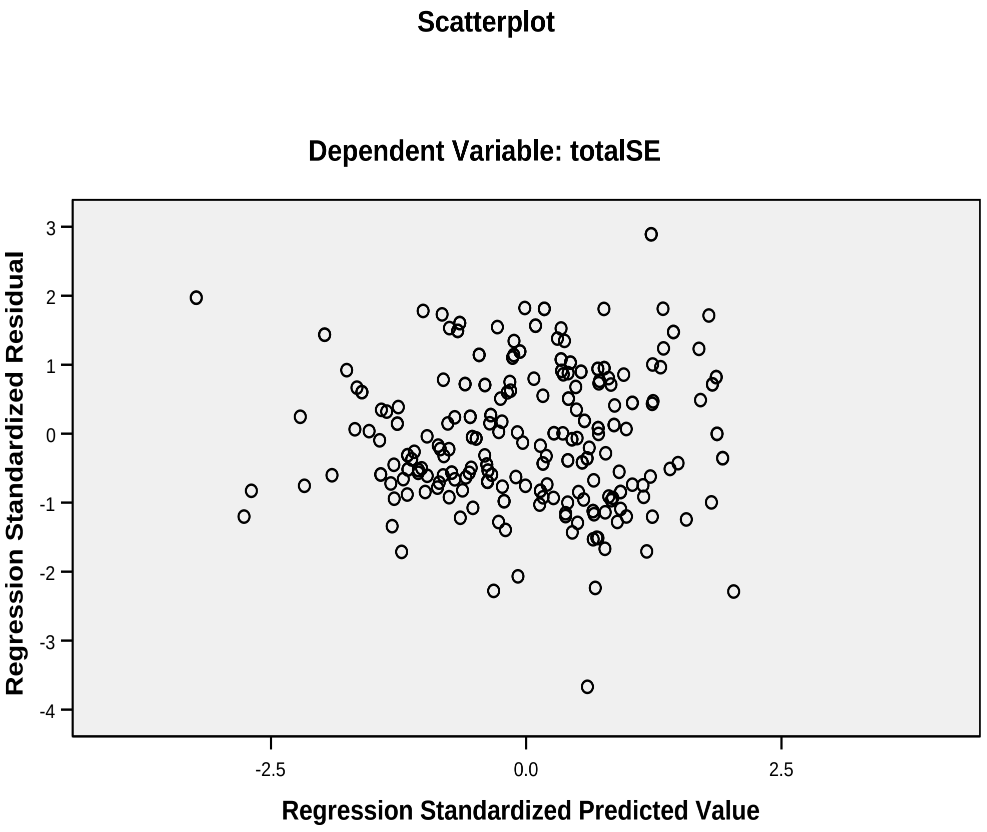
☑ Normal probability plot

□ Produce all partial plots

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: totalSE





2.3.1 Outliers, Normality, Linearity and Independence of Residuals

- In the Normal Probability Plot you are hoping that your points will lie in a reasonably straight diagonal line from bottom left to top right.
- This would be no major deviations from normality.
- In the Scatterplot of the Standardised residuals, the residuals will be roughly rectangular distributed, with most of the scores in the centre.
- Tabachinick and Fidell (2001) define outliers as cases that have a standardized residual more than 3.3 or less than -3.3.
- If there are only a few outlying residuals in a large sample, it may not be necessary to take any action.

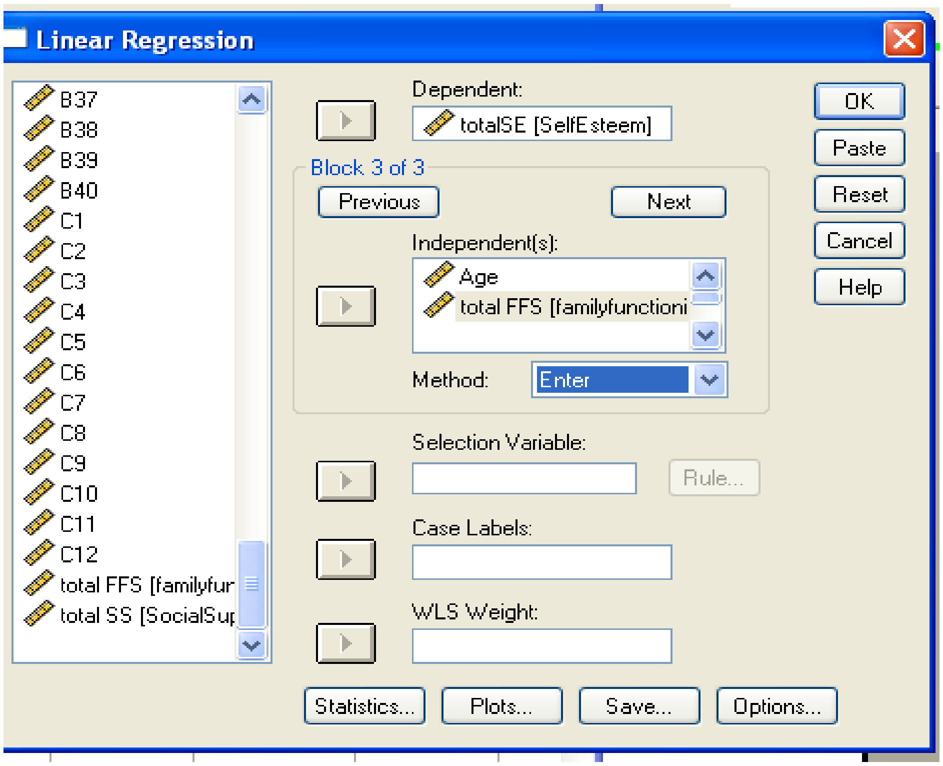
2.3.2 Hierarchical Multiple Regression

To illustrate the use of hierarchical multiple regression, we have to evaluate the effectiveness of the model. For example, after controlling for age and Total Family Functioning, will Total Perceived Social Support still able to predict a significant amount the variance in self-esteem? To answer this question, we need to use hierarchical multiple regression (also known as sequential regression). In the first

block, we have to “force” age and Total Family Functioning responding into the analysis. This has the effect of statistically controlling for the variables.
The next step we enter Total Perceived Social Support (independent) into the equation as a block where the possible effect of age and Total Family Functioning has been removed.

2.3.3 Procedure for Generating Hierarchical Multiple Regression

- Select **Analyze**, click on **Regression** and **Linear.**
- Select dependent variable (self-esteem) and move it into the **Dependent** box**.**
- Move Total Family Functioning and age (independent variable) that you wish to control into **Independent** box.
- Click on **Next**.



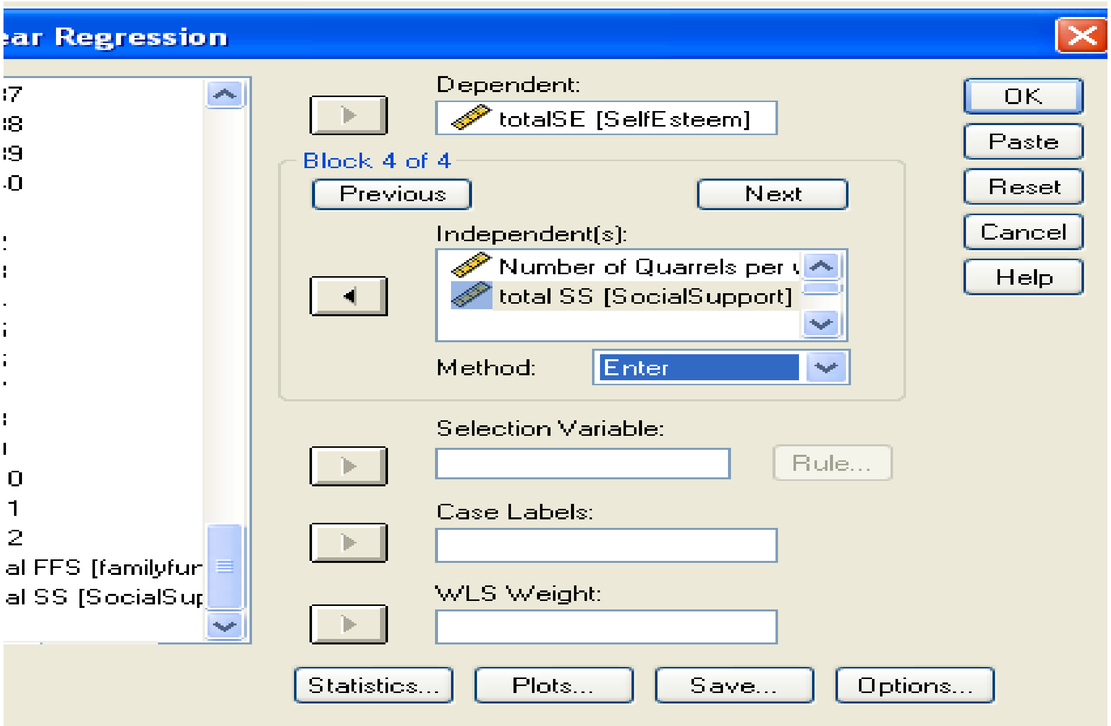
Dependent: totalSE [SelfEsteem]

Block 3 of 3

Independent(s): Age, total FFS [familyfunctioni

Method: Enter

- Select the next block of independent variable variables (Number of Quarrels and Total Perceived Social Support). Move the independent variables into the **Independent** box.
- Set default (**Enter**) in the **Method** box.

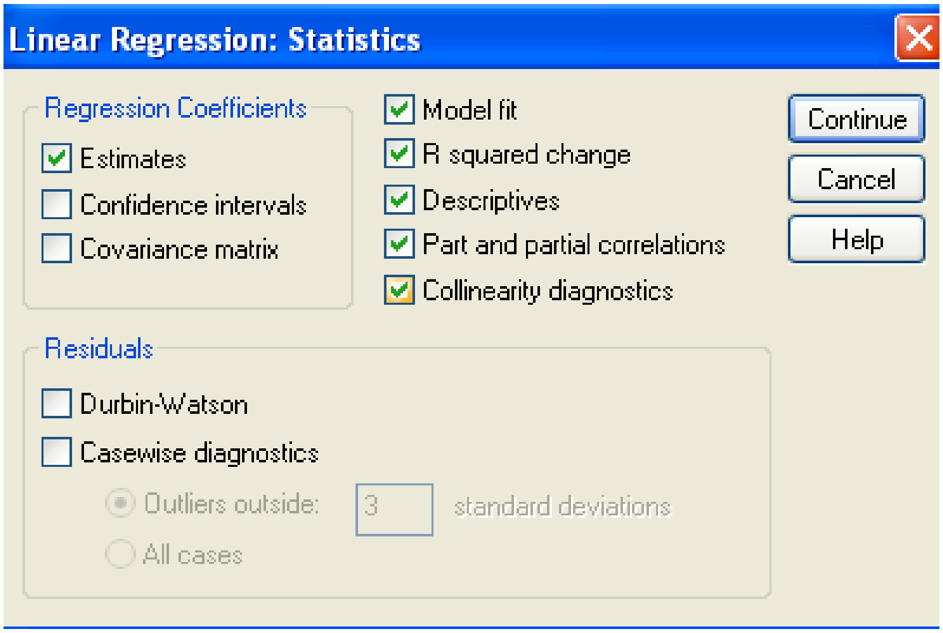


Dependent: totalSE [SelfEsteem]

Block 4 of 4

Independent(s): Number of Quarrels per ι
total SS [SocialSupport]

Method: Enter



Linear Regression: Statistics

Regression Coefficients

☑ Estimates

☐ Confidence intervals

☐ Covariance matrix

☑ Model fit

☑ R squared change

☑ Descriptives

☑ Part and partial correlations

☑ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

- Click on **Statistic. Tick Estimates, Model fit, R squared change, Descriptive, Part and partial correlations and Collinearity diagnostics.**
- Click on **Continue.**
- Click on the **Options** button. Select **Missing Values** and Exclude **cases pair wise.**
- Click on **Save**.
- Select **Mahalonobis** and **Cook's.**
- Click on **Continue** and **OK**.

2.3.4 Interpretation of Hierarchical Multiple Regression

Output of SPSS

Model Summary

				Change Statistics				
Model	R	R Square	Std. Error of the Estimate	R Square Change	F Change	Df1	Df2	Sig F Change
1	.120(a)	.014	27.079	.014	1.786	2	243	.170
2	.456(b)	.208	24.375	.194	29.461	2	241	.000

a Predictors: (Constant), total FFS, Age
b Predictors: (Constant), total FFS, Age, Contineous data of Number of Quarrels per week, total SS

ANOVA(c)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	2619.854	2	1309.927	1.786	.170(a)
	Residual	178191.105	243	733.297		
	Total	180810.959	245			
2	Regression	37627.040	4	9406.760	15.833	.000(b)
	Residual	143183.919	241	594.124		
	Total	180810.959	245			

a Predictors: (Constant), total FFS, Age
b Predictors: (Constant), total FFS, Age, Continouse data of Number of Quarrels per week, total SS
c Dependent Variable: totalSE

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	Zero-order	Partial
1	(Constant)	6.785	26.902		.252	.801
	Age	-.444	1.047	-.027	-.424	.672
	total FFS	.169	.091	.117	1.844	.066
2	(Constant)	-40.407	25.343		-1.594	.112
	Age	.809	.959	.049	.844	.400
	total FFS	-.080	.089	-.056	-.900	.369
	total SS	1.127	.151	.468	7.442	.000
	Continouse data of Number of Quarrels per week	-2.159	1.238	-.100	-1.744	.082

a Dependent Variable: totalSE

In the model summary, there are two models listed. Model **Regression**
Regression is used to make predictions based on linear relationship. It is a statistical technique for finding the best-fitting straight line for a set of data. The resulting straight line is called regression line.
Model 1 refers to the first block of independent variables (Age and Total Family Functioning), whereas Model 2 includes the second block of independent variables (Number of quarrels per week and Total Perceived Social Support).
To find out how much of this overall variance is explained by independent of interest (Number of quarrels per week and Total Perceived Social Support) after the effects of age and Total Family Functioning responding are removed, we need to look in the column labeled R Square change. R Square change value is .194. This means that Number of Quarrels per week and Total Perceived Social Support explain an additional 19.4 % of the variance in self-esteem, even when the effects of age and Total Family Functioning responding are controlled. The ANOVA table indicates that the model as a whole is significant [F (4, 241) = 15.833, p < .01].