LAB 4 INSTRUCTIONS

MULTIPLE LINEAR REGRESSION

Multiple linear regression is a straightforward extension of the simple linear regression model. It models the mean of a response variable as a function of several explanatory variables. In this lab you will learn how to use linear regression tools in SPSS to obtain the estimated regression equation and make inferences associated with regression analysis. You will also study variable selection techniques, regression diagnostic tools and case-influence statistics. We will omit or spend relatively little time on those SPSS linear regression tools that are natural extensions of the tools employed for simple linear regression model and were already discussed in *Lab 3 Instructions*.

We will demonstrate multiple linear regression tools in SPSS using a simple example with three explanatory variables.

Example: A survey was carried out to study television viewing habits among senior citizens. Twenty-five subjects over the age of 65 were sampled and the following variables were recorded:

Column	Variable Name	Description of Variable
1	TV	average number of hours per day that the subject watches television,
2	MARRIED	marital status, (1 if living with a spouse, 0 otherwise);
3	AGE	age of subject in years,
4	EDUC	number of years of formal education.

The data are saved in SPSS file *tv.sav* and can be downloaded by clicking the link below:

DOWNLOAD DATA

We will use the data to examine the relationship between the average number of hours per day spent watching television (response variable) and the marriage status (living with a spouse or not), age, and education.

1.   Multiple Linear Regression Model

In multiple regression model there is a single response variable and several explanatory variables and we are interested in the distribution of the response variable as a function of explanatory variables.

In our example *TV* is the response Assume that the relationship between the response variable *TV* and each of the three explanatory variables *MARRIED, AGE,* and *EDUC* is linear. We would like to determine how the number of hours spent watching television is affected by the subjects' age, marital status, and education.

Define a multiple regression as follows:

$$TV = \beta_0 + \beta_1 MARRIED + \beta_2 AGE + \beta_3 EDUC + ERROR.$$

We assume here that the variable *ERROR* follows a normal distribution for each combination of values of the explanatory variables and the mean of *ERROR* is zero. Moreover, we assume that the variance of the *ERROR* variable is constant for each combination of values of the explanatory variables.

The above model can be equivalently rewritten as

$$\mu(TV \mid MARRIED, AGE, EDUC) = \beta_0 + \beta_1 MARRIED + \beta_2 AGE + \beta_3 EDUC.$$
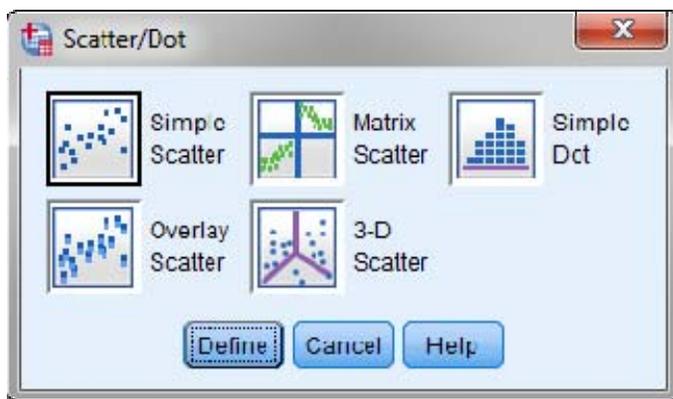
2.   Matrix of Scatterplots

Before you apply the regression tool in SPSS for your data, you must make sure that the explanatory variables are linearly related to the response variable. If they are not, you may have to transform the original data. For example, you may have to apply log or square root transformation to make the relationship approximately linear.

Scatterplots are very useful in visualizing the relationship between the response variable and a single explanatory variable. However, it is much harder to display the relationship among the response variable and several explanatory variables in multiple regression problems.

A matrix of scatterplots is an array of scatterplots displaying all possible pairwise combinations of the response and explanatory variables. The number of rows and columns in the matrix is equal to the number of all variables in the model (with the response variable included) and the scatterplot of one variable versus another is at the intersection of the appropriate column and the row in the matrix. Briefly, scatterplot matrix is a matrix whose elements are scatterplots of each pair of variables. The scatterplot may be useful to evaluate the strength of the relationship between the response variable and each of several explanatory variables, its direction and to bring your attention to unusual observations.
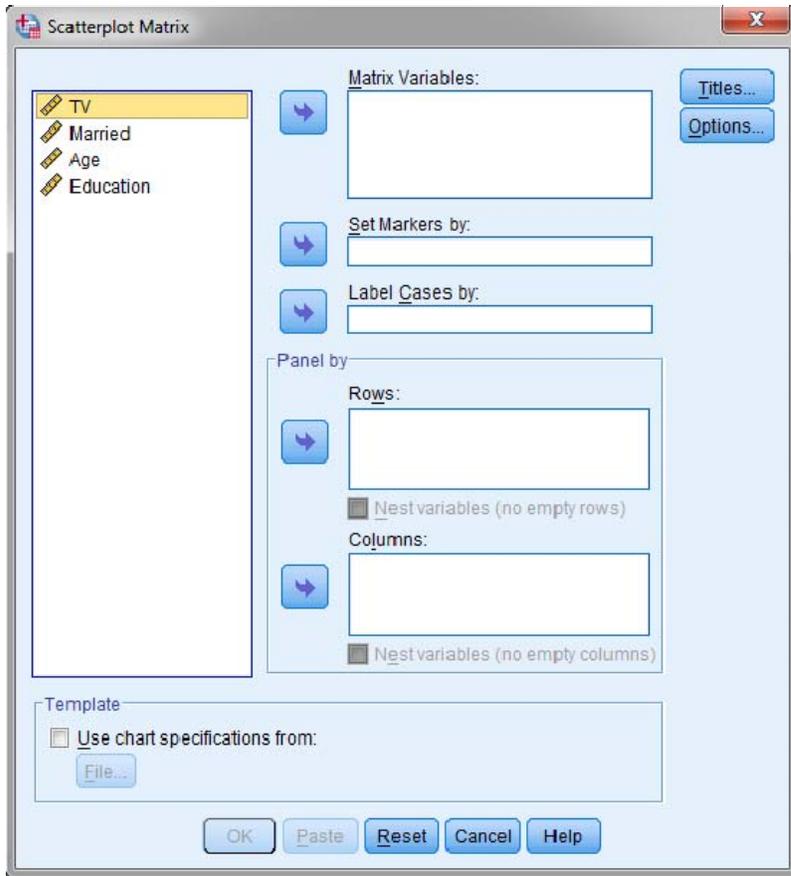
Obtaining a matrix of scatterplots is usually the first step in examining the relationships among several variables in multiple regression problems. However, it is important to realize that though simple scatterplots are very useful in exploring the relationship between a response and a single explanatory variable in simple regression problems, matrix of scatterplots is not always effective in revealing the complex relationships among the variables or detecting unusual observations in multiple regression problems.

To illustrate the above concepts we will obtain a matrix of scatterplots for our data. To access the *Matrix of Scatterplots* feature in SPSS, select *Scatter/Dot* option in the *Graphs* menu. It opens the *Scatter /Dot* dialog box shown below.

Click the *Matrix Scatter* icon and then the *Define* button. You will obtain *Scatterplot Matrix* dialog box displayed on the next page.

Select and move all variables, including the response (*TV*) and three predictors (*MARRIED, AGE* and *EDUC*) into the *Matrix Variables* box. Put the response variable on the top of the list to make sure that the variable will appear on the vertical axis in the first row of the matrix of scatterplots. Then click *OK*.

The matrix of scatterplots will be displayed in O*utput View* window.

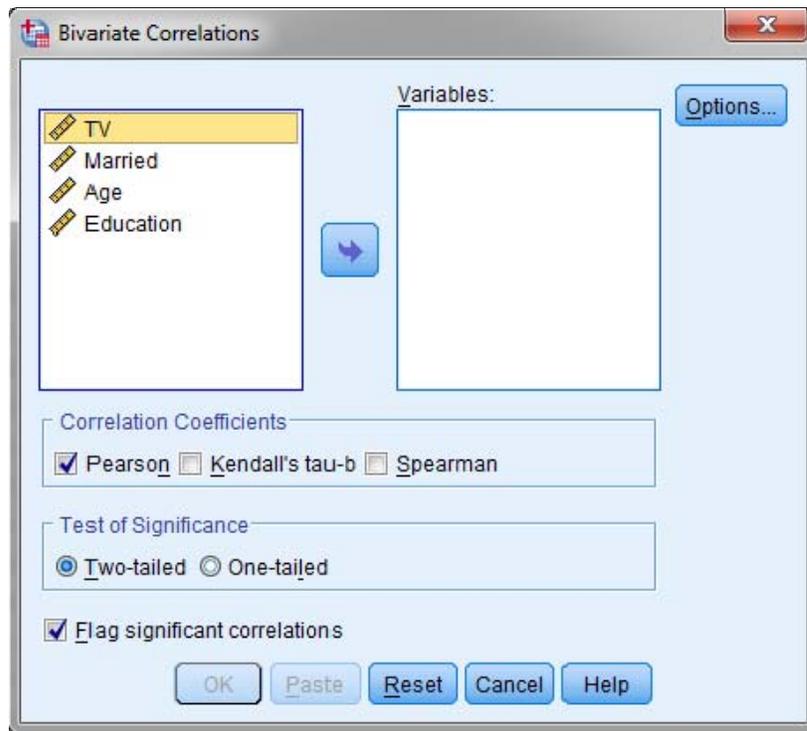Notice that the matrix of scatterplots consists of 12 smaller plots describing 6 different relationships. The first row shows the relationship between *TV* with each of the three predictors (*MARRIED*, *AGE* and *EDUC*) respectively. Notice that the above matrix of scatterplots is symmetric; the upper right triangle contains the same plots that the lower left one but with axes reversed. This provides a different perspective and may be useful in evaluating the pattern in the plots.

First we evaluate the relationship between the response variable TV and each of the three explanatory variables, *MARRIED*, *AGE*, and *EDUC*. There is a negative linear relationship between *TV* and *EDUC*. The linear relationship between *TV* and *EDUC* is strong; the one outlier at the left bottom of the plot makes the relationship weaker. The scatterplot matrix also shows a moderate tendency for the number of TV hours to increase as age of the subject increases (a positive relationship); the pattern is disturbed by a few points at the right bottom of the plot. *MARRIED* is in fact a categorical variable with only two possible values (0 or 1), and therefore the relationship between marital status and number of hours spent watching TV can be displayed better with other displays (for example, boxplots). Nevertheless, the scatterplot of *TV* versus *MARRIED* shows that singles tend to spend more time watching television than subjects living with a spouse. Notice also a considerably smaller spread for the TV values for singles than that one for married people.

Now we comment about the relationship between pairs of explanatory variables. Scatterplots of one explanatory variable versus another explanatory variable may be useful in detecting possible multicollinearity (when the variables are highly correlated). There is a moderate negative linear relationship between age and education.

3. Matrix of Correlations

In this section you will learn how to measure the association between interval variables (their values represent ordered categories, so that distance comparisons or ratios are appropriate; for example, income or age) or ordinal variables (their values represent only order or ranking; for example, levels of satisfaction).

3.1 Bivariate Correlations

First we will discuss here bivariate correlation (i.e. dealing with the relationship between two variables regardless of the influence of other variables). When you want to control for the effect of a third variable, you need to apply partial correlation. The *Bivariate Correlations* procedure in SPSS computes the pairwise associations for a set of variables and displays the results in a matrix.

Click *Analyze* in the main menu and then *Correlate* from the pull-down menu, and finally *Bivariate*...

The *Bivariate Correlations* dialog box is displayed on the next page. In the *Bivariate Correlations* dialog box, select two or more variables and move them into *Variables* list. This produces a matrix of correlation coefficients for all pairs of the selected variables.


The *Correlation Coefficients* group consists of three check boxes: *Pearson*, *Kendall's tau-b*, and *Spearman*. The Pearson correlation coefficient is appropriate for interval variables, while the Kendall and Spearman coefficients can be obtained for variables measured on ordinal scale. The Pearson correlation describes the strength of the linear association between two interval variables. Spearman correlation is simple the Pearson correlation when the data values are replaced by their ranks. *Pearson* is the default option in the *Bivariate Correlations* dialog box.

Once you've computed a correlation coefficient, you would like to know how likely the observed correlation has occurred by chance i.e. the observed value of the sample correlation coefficient is the result of sampling error. Indeed, if an outcome could have occurred by chance with considerable probability, this outcome would not be considered trustworthy. The probability that the observed outcome occurred by chance (p-value) is also produced by the feature. In the *Test of Significance* group you may select either two-tailed or one-tailed tests of statistical significance of the observed correlation. A one-tailed test is needed, if we assume that the relationship between two variables has a certain direction. If you have no prior expectation regarding a positive or negative association between the two variables, you should use a two-tailed test. By default, SPSS performs a two-tailed test.

SPSS marks all significant correlations. If you leave the *Flag significant correlations* check box selected (default option), correlations significant at the 0.05 level are marked with an asterisk, and those significant at the 0.01 level with two asterisks.

The *Bivariate Correlations* procedure has some additional features. If you click the *Options* button, you can request some additional statistics or specify how the missing data are to be treated. The optional statistics are means and standard deviations displayed for the variables in the correlation matrix.
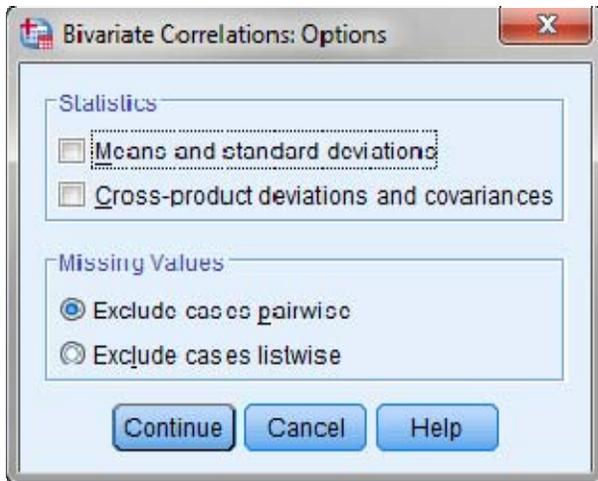
Now we obtain the matrix of correlations for our data. Select and move all three interval variables TV, AGE, and EDUC into *Variables* box. Make sure that the *Pearson* and *Flag significant correlations* boxes are checked. Click *OK* to run the procedure. The matrix of correlations is shown below.

The Pearson correlation coefficient indicates the direction and strength of the relationship between two variables. Correlation coefficients range from –1 to 1, where 1 or -1 is a perfect correlation (straight-line relationship) and 0 is no correlation. A negative coefficient means that one variable tends to increase as the other decreases. A positive coefficient means that both variables tend to increase or decrease together.

The signs and magnitudes of the correlation coefficients in our case, confirm the conclusions based on the examination of the matrix of scatterplots in Section 1. For example, there is a significant correlation (-0.634) exists between the response variable (*TV*) and the variable *EDUC*. The correlation between *TV* and *AGE* is weak and not significant.

3.2 Partial Correlations

Bivariate correlation analysis provides information about the strength and direction of the linear relationship between two variables. Sometimes, however, the relationship between two variables is obscured by the influence of a third variable. The partial correlation coefficient is the correlation between two variables when the linear effects of other variables are removed. With the option *Partial Correlations* you can investigate the true relationship between two variables controlling for the effects of other variables.

Click *Analyze* in the main menu and then *Correlate* from the pull-down menu, and finally *Partial....*

The following dialog box opens.
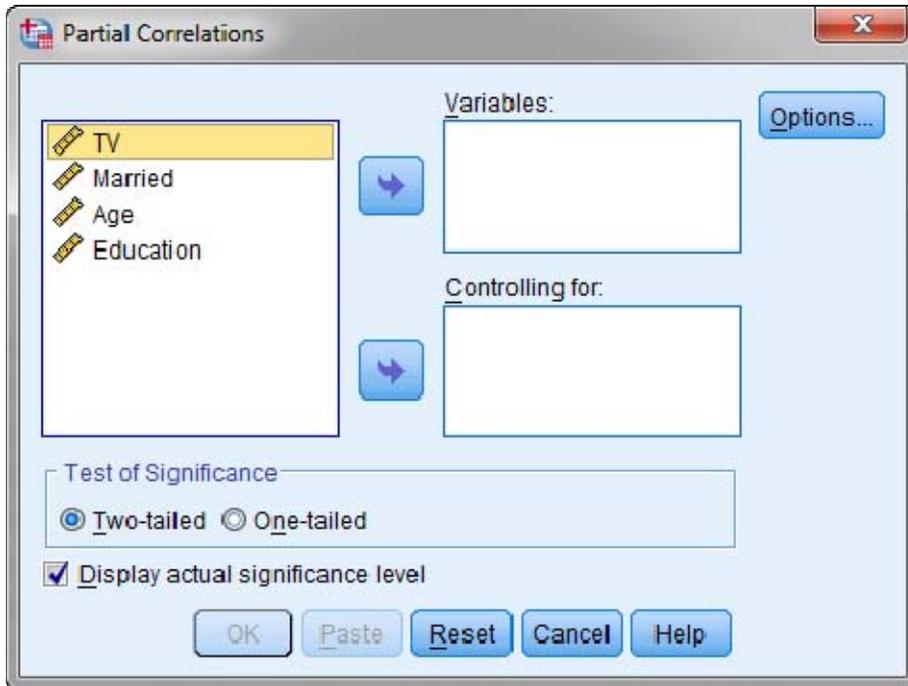
You need to specify the variables whose relationships you want to evaluate and the variable(s) you want to control. For instance, the matrix of correlations in Section 3.1 shows that the correlation coefficient between the number of hours spent watching television (*TV*) and age (*Age*) is not significant (p-value = 0.200). Nevertheless, it is possible that older people may spend more time watching television due to their restricted mobility. Therefore the relationship between *TV* and *Age* may be masked by the influence of a third variable, for example mobility score. Here mobility is a fictitious variable that can be added to the data file.

To determine the true strength of the relationship between *TV* and *Age* (not affected by the third variable, mobility score), we obtain the partial correlation coefficient between *TV* and *Age*.

The dialog box of the *Partial Correlations* procedure requesting for the correlation between *TV* and *Age* when controlling for the effect of the third fictional variable (mobility, not included in the actual data file) is shown below:
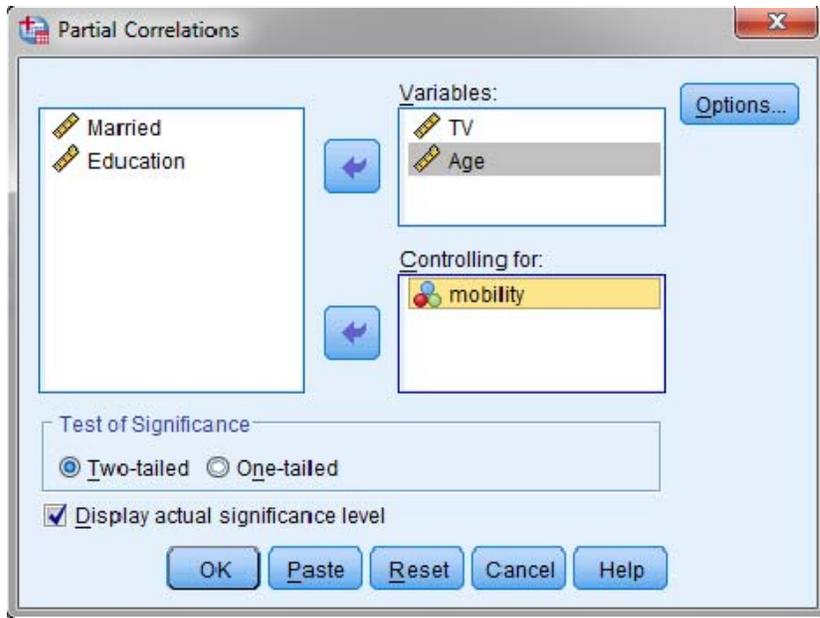
Suppose the output of the *Partial Correlations* procedure is as follows:

The bivariate correlation coefficient between TV and Age was 0.200 (see the correlation table in Section 3.1) and has now increased to 0.489. The correlation between *TV* and *Age* is now significant, with a significance level of 0.015. Se we indeed find evidence that older people spend significantly more time watching television.

4.   Multiple Linear Regression in SPSS

Multiple regression is a natural extension of simple linear regression model discussed in *Lab 3 Instructions*. In this section we will discuss the Linear Regression tool in SPSS in more detail. We will demonstrate how to build a linear regression model that has more than one explanatory variable and how to interpret the corresponding SPSS output.
Click *Regression* in the *Analyze* menu, then on *Linear*.
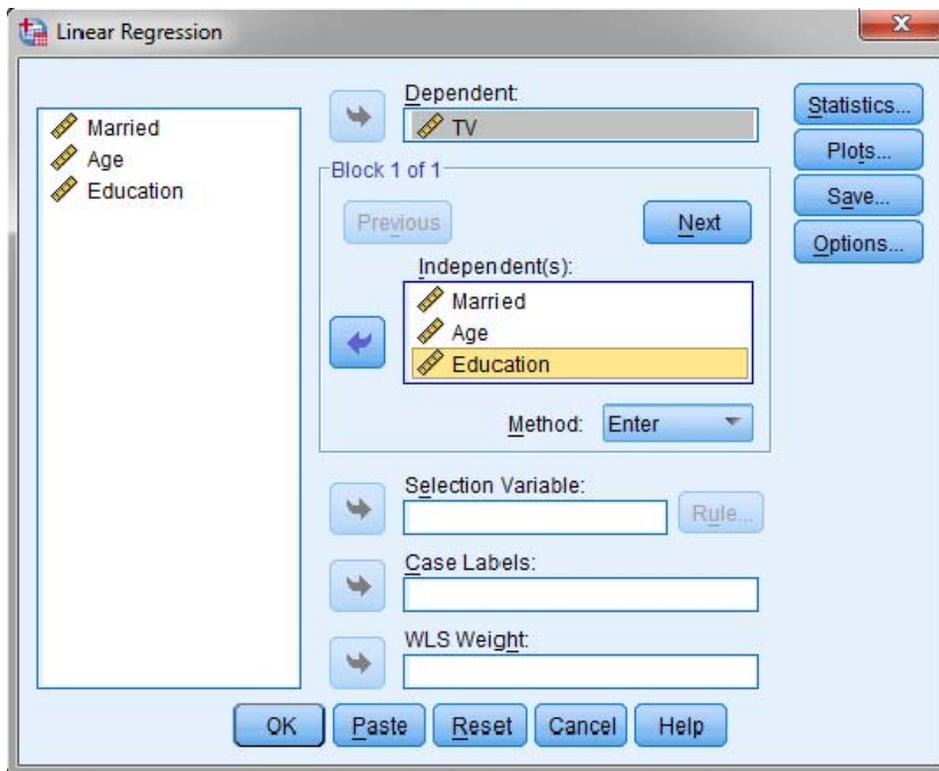
Move the response variable TV into the *Dependent* list and the explanatory variables MARRIED, AGE and EDUC into the *Independent(s)* list. Click *OK*. This produces the basic regression output shown below.

Variables Entered/Removed[b]

Mode l	Variables Entered	Variables Removed	Method
1	Education, Age, Married[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: TV

Model Summary

Mode l	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.737[a]	.543	.478	.8719

a. Predictors: (Constant), Education, Age, Married

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.975	3	6.325	8.321	.001[a]
	Residual	15.963	21	.760		
	Total	34.938	24			

a. Predictors: (Constant), Education, Age, Married

b. Dependent Variable: TV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.443	2.904		.841	.410
	Married	-1.078	.433	-.452	-2.487	.021
	Age	.017	.036	.084	.473	.641
	Education	-.103	.054	-.383	-1.915	.069

a. Dependent Variable: TV

The basic output of Linear Regression in SPSS consists of the following four components: an overview of the variables included in the regression model (*Variables Entered/Removed*), the overall results of the regression analysis (*Model Summary*), the F test for the estimated model (*ANOVA*) and an overview of the estimated regression coefficients and the corresponding statistics (*Coefficients*).

4.1 Multiple Regression Basic Output Interpretation

Now we will discuss in detail all components of the SPSS output obtained above.

(a) Variables Entered/Removed

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Education, Age, Married[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: TV

Model: SPSS allows you to specify multiple models in a single regression command. This column specifies the number of the model being reported in the output.

Variables Entered: SPSS allows you to enter the explanatory variables in blocks, and you can specify different entry methods for different blocks of variables (Enter, Stepwise, Backward,…). For example, you can enter one block of variables into the regression model using stepwise selection and a second block using forward selection.

If you did not block your explanatory variables or use stepwise regression, this column should list all of the explanatory variables that you specified. In case of our data, the table lists all explanatory variables in the model: *MARRIED*, *AGE* and *EDUC*.

Variables Removed: This column listed the variables that were removed from the current regression. This column is empty in our case (Enter method).

Method: Method selection allows you to specify how explanatory variables are entered into the model. Using different methods, you can construct various regression models from the same set of variables. The methods will be discussed in detail later.

 (b)   **Model Summary**

In the standard form (with the default options: *Estimates* and *Model fit* checked in *Linear Regression: Statistics* dialog box) the *Model Summary* produces four values: the value of the multiple correlation coefficient R, the coefficient of determination R^2, Adjusted R^2, and Standard Error of the Estimate.

Model Summary^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.737^a	.543	.478	.8719
a. Predictors: (Constant), Education, Age, Married				
b. Dependent Variable: TV				

The *Model Summary* produces four values: the value of the multiple correlation coefficient R, the coefficient of determination R^2, Adjusted R^2, and Standard Error of the Estimate.

R (multiple correlation coefficient) is the correlation between the observed values and the predicted values. It is used in multiple regression analysis to assess the quality of the prediction of the response variable in terms of the explanatory variables. The value of R ranges from 0 to 1. The higher the value of R, the better the predictive power of the regression model. The value of R for our data is 0.737.

R Square is the squared multiple correlation coefficient. It is also called the coefficient of determination. R^2 is the ratio of the sum of squares of residuals due to regression to the total sum of squares for the regression model. R^2 indicates what proportion of the variation in the response variable is explained by the explanatory variables. If all cases are exactly on the regression line (the residuals are all zero), R^2 equals 1. If R^2 is zero, the model has no predictive capability (this does not automatically mean that there is no relationship between the response and explanatory variables. It only means that no linear relationship exists). The higher the value of the coefficient of determination, the better predictions can be obtained with the regression model. The R^2 value in our example is 0.543 which means that 54.3% of the variation in TV is explained by the three explanatory variables. That high percentage makes the model really useful.

Adjusted R Square is a modified measure of the coefficient of determination that takes into account the sample size and the number of explanatory variables in the model. The rationale for this statistic is that, if the number of explanatory variables is large relative to the sample size, the unadjusted R^2 value may be unrealistically high (the addition of explanatory variables will always cause the R^2 value to rise, even if the variables have no real predictive capability). However, when variables are added to the model, adjusted R^2 doesn't increase unless the new variables have additional predictive capability; adjusted R^2 may even fall if the added explanatory variables have no explanatory power and are statistically insignificant. The statistic is quite useful to compare regression models with different numbers of explanatory variables. Data sets

with a small sample size and a large number of predictors will have a greater difference between the obtained and adjusted R square. The adjusted R^2 for our data is 0.478 and it is smaller than unadjusted R^2 value.

The Standard Error of the Estimate is an estimate of the standard deviation σ of the ERROR term in the regression model. The standard error of the estimate is a measure of the accuracy of predictions made with the regression model; the smaller the standard error of estimate, the better the prediction. It is obtained as the square root of the Residual Mean Square (sum of squares of residuals divided by their respective degrees of freedom).

Now we will discuss the *Model Summary* output when additionally to the default options (*Estimates* and *Model fit*) the *R squared change* and *Descriptives* boxes are also checked in *Linear Regression: Statistics* dialog box. If *R squared change* option is selected the following columns are added to the standard output:

Model Summary[b]

		Change Statistics		
R Square Change	F Change	df1	df2	Sig. F Change
.543	8.321	3	21	.001

The extra columns display R^2 and F change and the corresponding significance level. R^2 change refers to the amount R^2 increases or decreases when a variable is added to or deleted from the equation as is done in stepwise regression or if the explanatory variables are entered in blocks. If R^2 change associated with a specific explanatory variable is large, that means that the variable is a good predictor of the response variable. If the "Enter" method is used to enter all explanatory variables at once in a single model, R^2 change for that model will reflect change from the intercept-only model.

In the above output, R^2 change is the same as R^2 because the variables were entered at the same time (not stepwise or in blocks); there is only one regression model to report, and R^2 change is change from the intercept-only model, which is also what R^2 is. R^2 change is tested by F test.

If the *Descriptives* option is selected, then the mean and the standard deviation for each variable in the analysis and the correlation matrix are also displayed. If the *Durbin-Watson* box is checked, the results of the Durbin-Watson test for serial correlation of the residuals are provided. The Durbin Watson test is of particular importance for data collected over time.

(c)   The Analysis of Variance Table

The third block of the regression output contains the results of the analysis of variance.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.975	3	6.325	8.321	.001[a]
	Residual	15.963	21	.760		
	Total	34.938	24			

a. Predictors: (Constant), Education, Age, Married

b. Dependent Variable: TV

The analysis of variance table is used to test the null hypothesis that all of the population regression coefficients, with the exception of the intercept, are equal to zero (equivalently, there is no linear relationship in the between the response variable and the explanatory variables). In other words, this hypothesis states that none of the considered explanatory variables is useful in explaining the response. The F test is used to test the hypothesis; this test is often called the F-test for overall significance of the regression.

Now we will discuss all columns in the above ANOVA output in detail. In testing the null hypothesis, the total variation of the response variable is divided into two components. One part of the variation is explained by the explanatory variables in the regression model, while the other part is not explained, which is the residual. Analysis of variance compares the explained variation, which SPSS labels as "Regression", with the unexplained variation, called "Residual".

The second column in the ANOVA output displays the sums of squares (Sum of Squares) for the two components (the regression equation and the residual) and also their total. The following column contains the degrees of freedom (df). For the regression model these equal the number of explanatory variables, k=three in our example. The number of degrees of freedom for the residual is equal to the number of cases (n=25) minus the number of explanatory variables minus 1 (n-k-1= 25-3-1=21).

The following column contains mean sum of squares for the regression model and the residual (Mean Square).The mean sum of squares is obtained by dividing the sum of squares by the corresponding degrees of freedom. The square root of the Mean Square for the residual component is equal to the standard error of the estimate discussed above. In our example, based on the *Model Summary* part of the output the standard error of the estimate is 0.8719 and indeed it is equal to the square root of the mean square for the residual of 0.760 (any discrepancy is due to rounding).

The final two columns show the results of the F-test. The F-value is obtained as the ratio of the mean sums of squares of the regression model and the residual. In our example, the F value equals 6.325/0.760=8.321. SPSS also displays the level of significance of the F test and the degrees of freedom. Under the null hypothesis, the F statistic follows an F distribution with the degrees of freedom for the numerator equal to the number of explanatory variables (k) and the number of degrees of freedom for the denominator equal to the number of cases minus the number of explanatory variables minus 1. In our example, F statistic follows an F distribution with 3 degrees of freedom for the numerator and 21 degrees of freedom for the denominator. The p-value reported by SPSS is 0.001. Using the threshold value of α=0.05, we reject the null hypothesis that all regression coefficients are zero. In other words, al least one regression coefficient significantly differs from zero; the regression model is useful.

(d)   Coefficients

The *Coefficients* table follows the ANOVA output.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.443	2.904		.841	.410
	Married	-1.078	.433	-.452	-2.487	.021
	Age	.017	.036	.084	.473	.641
	Education	-.103	.054	-.383	-1.915	.069
a. Dependent Variable: TV						

In the standard form (with the default options: *Estimates* and *Model fit* checked in *Linear Regression: Statistics* dialog box) the table contains the estimates of the population regression coefficients, their standard errors, the standardized coefficients and the values of the t statistics to test the regression coefficients with the corresponding two-sided p-values.

This first column *Model* shows the number of the model being reported (1) and the predictor variables (Constant, Married, Age, Education). The first variable (Constant) represents the constant, also referred to in the text as the y intercept, the height of the regression line when it crosses the Y axis.

The column B contains the coefficients for the regression equation for predicting the response variable from the explanatory variables and the values of the other explanatory variables do not change. Thus, the estimated regression equation is

$$\mu(TV) = 2.443 - 1.078 \cdot MARRIED + 0.017 \cdot AGE - 0.103 \cdot EDUC.$$

The interpretation of the coefficient of an explanatory variable in a regression model depends on what other explanatory variables are included in the model. A regression coefficient indicates the number of units of change (increase or decrease) in the response variable caused by an increase of one unit in the explanatory variable (with the constant values of the other variables). A positive coefficient means that the predicted value of the response variable increases when the value of the explanatory variable increases. A negative coefficient means that the predicted value of the response variable decreases when the value of the explanatory variable increases. Thus according to the above equation, married seniors watch on average 1.078 hours less television per day than not married people with the constant values of the other variables.

The regression coefficient does not reflect the relative importance of a variable because the magnitude of the coefficients depends on the units used for measuring the variables. The regression coefficients reflect the relative importance of the variables only if all explanatory variables are measured in the same units. In order to make meaningful comparisons among the regression coefficients, SPSS also displays the coefficients that would be obtained if the variables were standardized before the regression analysis. These regression coefficients, shown as *Standardized Coefficients-Beta*, reflect the relative importance of the explanatory variables. To compare the relative importance of two variables, you have to use the absolute values of the beta coefficients.

The last two columns *t* and *Sig.* display the results of the t-test of the regression coefficients. In this test, the null hypothesis states that a regression coefficient is zero. The alternative hypothesis states that a regression coefficient differs from zero. The value of the t statistic for a regression coefficient is displayed in the column *t.* This value is obtained by dividing the estimated regression coefficient in the B column by the corresponding standard error. The standard error is found in the output under *Std. Error.* The last column of the table contains two-tailed p-value for the computed t value. If the p-value is smaller than α=0.05, the coefficient is significantly different from 0.

The p-value of the t test for an explanatory variable must also be interpreted in terms of other variables included in the model. Therefore the variable MARRIED is significant (p-value of 0.021) given the other variables in the model.

The coefficient for AGE (0.017) is not significantly different from 0 using alpha of 0.05 because its p-value is 0.641, which is larger than 0.05.

95% confidence intervals for each regression coefficient will also be displayed if the *Confidence intervals* box is checked in *Linear Regression: Statistics* dialog box. They are very useful as help you understand how high and how low the actual values of the population parameters might be. The confidence intervals are related to the p-values such that the coefficient will not be statistically significant if the confidence interval includes 0.

95% Confidence Interval for B	
Lower Bound	Upper Bound
-3.596	8.482
-1.979	-.177
-.057	.091
-.214	.009

(e) Collinearity Diagnostics

Multicollinearity is a situation in which two or more explanatory variables in a multiple regression model are highly correlated. When correlation is excessive, standard errors of the estimated regression coefficients become large, making it difficult or impossible to assess the relative importance of the predictor variables. Multicollinearity is a matter of degree: there is no irrefutable test that it is or it is not a problem. Nevertheless, there exist diagnostic tools to detect excessive multicollinearity in the regression model.

Multicollinearity does not diminish the predictive power of the regression model; it only affects calculations regarding individual explanatory variables. A natural solution to this problem is to remove some variables from the model. When two explanatory variables are involved, multicollinearity is called collinearity (means that strong correlation exists between them).

In order to obtain an extensive collinearity statistics (tolerance, VIF, regression coefficient variance-decomposition matrix), make sure that the *Collinearity diagnostics* box is checked in *Linear Regression: Statistics* dialog box.
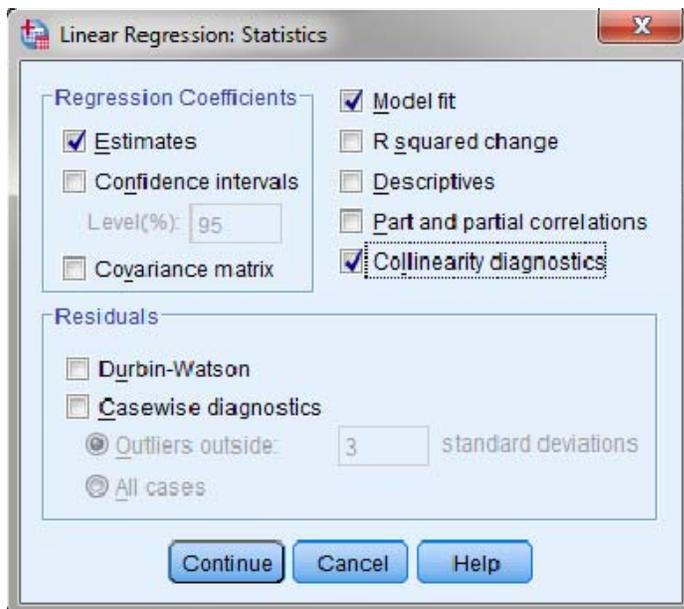
The first table in the output displays two collinearity measures, the tolerance and the VIF. To determine the tolerance, SPSS computes the R^2 of the regression model in which one of the explanatory variables is treated as the response variable and is explained by the other explanatory variables. The bigger R^2 is (i.e. the more highly the explanatory variable is with the other explanatory variables in the model), the bigger the standard error will be. In consequence, confidence intervals for the coefficients tend to be very wide and t-statistics tend to be very small.

The tolerance equals one minus this R^2 expressing that fraction of the variation of an explanatory variable that is not explained by the other explanatory variables. Since tolerance is a proportion, its values range from 0 to 1. A value close to 1 indicates that an explanatory variable has little of its variability explained by the other explanatory variables. A value of tolerance close to 0 indicates that a variable is almost a linear combination of the other explanatory variables. If any of the tolerances are small (less than 0.10), multicollinearity may be a problem.

The VIF (Variance Inflation Factor) is the reciprocal of the tolerance (i.e. 1 divided by the tolerance). A rule of thumb often used is that VIF values greater than 10 signal multicollinearity.

If multicollinearity is a problem in your data, you may find that although you can reject the null hypothesis that all population coefficients are 0 based on the F statistic, none of the individual coefficients in the model is significantly different from 0 based on the t statistic. The collinearity diagnostics output for our data is shown below.

Coefficients[a]

		Collinearity Statistics	
Model		Tolerance	VIF
1	Married	.657	1.521
	Age	.696	1.437
	Education	.544	1.839

a. Dependent Variable: TV

Collinearity Diagnostics[a]

				Variance Proportions			
Model	Dimension	Eigenvalue	Condition Index	(Constant)	Married	Age	Education
1	1	3.444	1.000	.00	.02	.00	.01
	2	.434	2.818	.00	.63	.00	.00
	3	.120	5.352	.00	.19	.01	.63
	4	.002	42.213	1.00	.16	.99	.36

a. Dependent Variable: TV

As none of the tolerance values the table above is smaller than 0.10, there is no evidence of multicollinearity in our data based on the tolerance values alone.

The *Collinearity Diagnostics* table in SPSS is an alternative method of assessing if there is too much multicollinearity in the model. In particular, condition indices are used to flag excessive collinearity in the data. A condition index over 30 may suggest serious collinearity problems and an index over 15 may indicate possible collinearity problems.

If a component (dimension) has a high condition index, one looks in the variance proportions column. If two or more variables have a variance proportion of .50 or higher on a factor with a high condition index, these variables have high linear dependence and multicollinearity is a problem, with the effect that small data changes or arithmetic errors may translate into very large changes or errors in the regression analysis.

Note that it is possible for the rule of thumb for condition indices (index over 30) to indicate multicollinearity (variance proportions ignored), even when the rules of thumb for tolerance > .10 suggest no multicollinearity. In our data, one of the condition indices is 42.213 (over 30), but the variance proportions as well the tolerance values do not indicate multicollinearity.

4.2   Building Regression Models

A common problem in regression analysis is that there are many variables that can be potentially good predictors of the response variable and that you have to determine how many, and which of those variables have to be included in the final regression model. The final model should provide the best possible explanation of the response variable and also be easy to interpret.

From an interpretation point of view, a regression model with as few as possible explanatory variables is preferred. On the other hand, a regression model with the highest value of R^2 provides the best possible explanation of the response variable. When more explanatory variables are added to the model, the R^2 value tends to increase. In an extreme situation you obtain a model that gives an excellent explanation of the response variable, but which contains so many explanatory variables that the interpretation becomes very difficult. Therefore, we have to find a balance between the ease of interpretation and the predictive power of the model.

One possible way of finding this balance is to include in the final model only those variables whose estimated regression coefficients are significant. A variant of this is not to include all variables in the model straight away, but to include them one by one. A new variable will be added to the model when the change in the R^2 value is sufficiently large to make the addition justified. Another opposite approach is to start with a model containing all possible variables and then decide which variables can be left out without substantially affecting the value of R^2. Selecting the explanatory variable requires a criterion to determine whether the change in the R^2 resulting from adding or removing a variable is significant. The F test is used for this purpose. The value of F is computed as the change in the R^2 value relative to R^2.

SPSS has two statistics that can be used as threshold values, both for adding and removing variables from the model. For adding (removing) variables:

(a)   The probability of F to enter (remove) is the maximum acceptable level of significance. If the computed p-value is lower (higher) than the entered value, the variable is added (removed); otherwise it is not.

(b)   The F to enter (remove) is the minimum F value. If the computed F is higher (lower) than this value, the variable is added (removed); otherwise it is not.

The two criteria can lead to different results because the degrees of freedom depend upon the number of variables in the model.
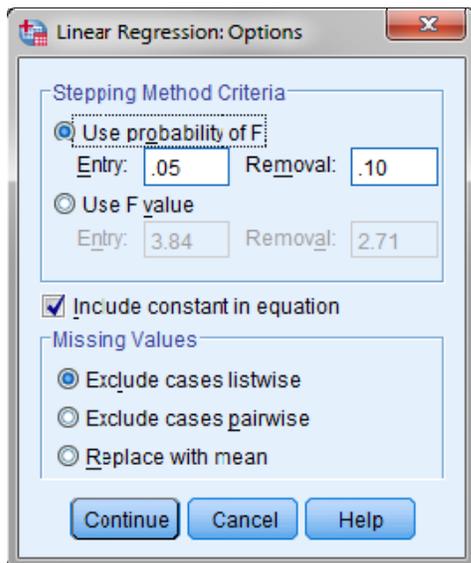
The button *Options* (see the picture above) in *Linear Regression* dialog box allows you to specify whether SPSS is to use the significance level or the F value as the criterion and change the default threshold values.

SPSS contains five methods that determine which independent variables are included in the regression model: Enter, Forward, Backward, Stepwise and Remove.

1. The **Enter** method: The model is obtained with all specified variables. This is the default method.

2. The **Forward** method: The variables are added to the model one by one if they meet the criterion for entry (a maximum significance level or a minimum F value). SPSS starts with the variable that has the largest correlation with the response variable. If this variable meets the criterion for entry, a regression analysis is performed with only this variable. Then SPSS determines which of the variables not yet included has the strongest partial correlation with the response variable. If this variable also meets the criterion for entry, it is included into the regression equation as the second variable. This process is continued until a variable no longer satisfies the criterion for entry or all variables have been included.

3. The **Backward** method: The variables are removed from the model one by one if the meet the criterion for removal (a maximum significance level or a minimum F value). SPSS starts with a model containing all explanatory variables. Next it finds the variable with the smallest partial correlation with the response variable and determines whether the variable meets the criterion for removal. If that is the case, this variable is removed and a new model is estimated. This process is continued until a variable no longer satisfies the criterion for removal or all variables have been removed.

4. The **Stepwise** method: This method is a combination of Forward and Backward. The variables are added in the same way as in Forward method. The difference with the Forward method is that when variables are added, the variables already in the model are also assessed based on the criterion for removal, as for Backward method. In order to prevent the same variable being alternately added or removed, make sure that the Probability of F to remove is always higher than the Probability of F to enter (or the F to remove is lower than F to enter).

5. The **Remove** method: Remove can be used when you have specified the explanatory variables in blocks (buttons Previous and Next in Linear Regression dialog box). All variables belonging to the same block are removed from the model in one step.

To illustrate the above methods, we will apply some of them to our data in Section 4.4

4.3   Multiple Regression Diagnostics

The linear regression model is valid under the assumption of a linear relationship between the response variable and each explanatory variable and the ERROR variable following a normal distribution for each combination of values of the independent variables with the mean zero. Moreover, we assume that the variance of the *ERROR* variable is constant for each combination of values of the explanatory variables.

The above assumptions can be tested by an examination of the residuals as they should reflect the properties assumed for the unknown ERROR term. The residuals are expected to follow a normal distribution with the mean 0.

We used residual plots to check the regression model assumptions for a simple linear regression. However, there are more plots to be examined in a multiple linear regression. Residuals can be plotted against each explanatory variable, against the predicted values for the response variable, and against the order in which the observations were obtained (if suitable). Often, the residuals in the plot are standardized by dividing each one by the standard deviation of the residuals.

The following diagnostic tools can be used to test the validity of the assumptions:

1. **Linearity:** The plot of residuals versus predicted values and versus explanatory variables. If the model is correct, there should be no pattern in these plots. Transformations of one or more of the variables can be tried (logarithm or square root) or polynomial terms (e.g., squares) of one or more of the variables can be added to try to remedy nonlinearity.

2. **Constant Variance:** The plot of residuals versus predicted values and versus explanatory variables. A pattern in the spread of the residuals (e.g., fan or funnel pattern) indicates nonconstant variance. If the assumption is satisfied, most of the residuals should fall in a horizontal band around 0. The spread of points should be approximately the same across all the predicted or explanatory variable values. Transformations of the response variable can be tried to correct nonconstant variance.

3. **Normality:** Normal probability plot (Q-Q plot of residuals; if the residuals come from a normal population, the points in the plot should fall close to a straight line) or histogram of the residuals (it should be approximately bell shaped).

4. **Independence:** Plot the residuals versus the time order of the observations. If the observations are independent, there should be no pattern in the plot over time.

The plot of standardized residuals versus standardized predicted values and normal probability plot of residuals can be obtained in SPSS by clicking the button *Plots* in *Linear Regression* dialog box and filling out *Linear Regression: Plots* dialog box as follows:
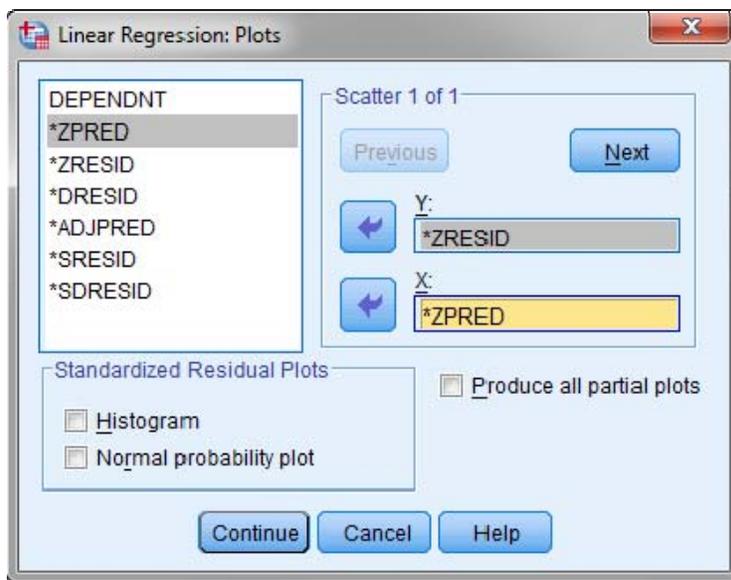
Outliers in regression are observations with residuals of large magnitude (in absolute value), i.e., observation's y value is unusual given its explanatory variable values. Least squares regression is not resistant to outliers. One or two observations with large residuals can strongly influence the analysis and may significantly change the answers to the questions of interest.

An observation is influential if removing it markedly changes the estimated coefficients of the regression model. An outlier may be an influential observation.

To identify outliers and/or influential observations, the following three case statistics can be calculated: studentized residual, leverage value, and Cook's distance.

1. **Studentized Residuals:** A studentized residual is a residual divided by its estimated standard deviation. Studentized residuals are used for flagging outliers. A case may be considered an outlier if the absolute value of its studentized residual exceeds 2.

2. **Leverage Values**: They are used for finding cases with unusual explanatory variable values: If the leverage for an observation is larger than $2p/n$, then the observation has a high potential for influence, where p is the number of regression coefficients and n is the number of data in the study.

   In general, an observation that has high leverage and of a large studentized residual will often be influential. Cook's Distance can be used to find observations that are influential.

3. **Cook's Distances:** They are used for flagging influential cases: If Cook's distance is close to or larger than 1, the case may be considered influential.

These three case-influence statistics: leverage values, studentized residuals and Cook's distance, can be requested in your regression analysis in SPSS by clicking on the *Save* button in *Linear Regression* dialog box. By identifying potentially influential cases, we can refit the model with and without the flagged cases to see whether the answers to the questions of interest change.

4.4   Applications: Example Data

Now we will use the backward elimination procedure to obtain the estimated linear regression model for our data and identify the possible influential cases.

Select the *Analyze* option in the main menu bar, then click on *Regression* from the pull-down menu, and finally on *Linear*. This opens *Linear Regression* dialog box. Select and move the variable *TV* into the *Dependent* box and the three explanatory variables (*MARRIED, AGE* and *EDUC*) into the *Independent(s)* box. Then click the *Method* button and select *Backward*. To identify the influential cases, click the *Save* button in *Linear Regression* dialog box and check the *Cook's*, *Leverage values*, and *Studentized* boxes in *Linear Regression: Save* dialog box.

To check the model assumptions, we will obtain the normal probability plot of residuals and a plot of standardized residuals versus standardized predicted values. Click the *Plots* button in *Linear Regression* dialog box. It opens *Linear Regression: Plots* dialog box. Check the *Normal probability plot* box and specify the type of residual plot. Click *Continue* to close the dialog box and then click *OK* to run the procedure.

The outputs are shown below.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	EDUC, AGE, MARRIED[a]	.	Enter
2	.	AGE	Backward (criterion: Probability of F-to-remove >= .100).

a. All requested variables entered.

b. Dependent Variable: TV

The table above shows that variable *AGE* has been eliminated by the backward elimination procedure. Only the two variables are included in the final model: *EDUC* and *MARRIED*.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.737[a]	.543	.478	.87187
2	.734[b]	.538	.496	.85635

a. Predictors: (Constant), EDUC, AGE, MARRIED

b. Predictors: (Constant), EDUC, MARRIED

c. Dependent Variable: TV

The R square value for the final model is 53.8%. It means that 53.8% of the variation in *TV* is explained by the two explanatory variables *MARRIED* and *EDUC* in the fitted regression model.

We define the regression model with the two explanatory variables as follows:
$$\mu(TV) = \beta_0 + \beta_1 MARRIED + \beta_2 EDUC$$

Then the null and alternative hypotheses to test the overall significance of the model are

$H_0 : \beta_1 = \beta_2 = 0$ vs. $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.975	3	6.325	8.321	.001[a]
	Residual	15.963	21	.760		
	Total	34.938	24			
2	Regression	18.805	2	9.403	12.822	.000[b]
	Residual	16.133	22	.733		
	Total	34.938	24			

a. Predictors: (Constant), EDUC, AGE, MARRIED

b. Predictors: (Constant), EDUC, MARRIED

c. Dependent Variable: TV

According to the ANOVA table above, the test statistic F follows the F-distribution with 2 degrees for numerator and 22 degrees for the denominator. The value of F = 12.822 and the corresponding p-value is reported as zero. This provides very strong evidence against the null hypothesis. Thus the regression model is highly significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.443	2.904		.841	.410
	MARRIED	-1.078	.433	-.452	-2.487	.021
	AGE	.017	.036	.084	.473	.641
	EDUC	-.103	.054	-.383	-1.915	.069
2	(Constant)	3.803	.392		9.705	.000
	MARRIED	-.998	.392	-.419	-2.544	.018
	EDUC	-.116	.044	-.435	-2.637	.015

a. Dependent Variable: TV

According to the *Coefficients* table above, the estimated regression equation is

$$\mu(TV) = 3.803 - 0.998 \times MARRIED - 0.116 \times EDUC$$

In order to see how significantly each explanatory variable contributes individually given the other variables in the model, we define the null and alternative hypotheses as follows:

$$H_0 : \beta_i = 0 \text{ vs. } H_a : \beta_i \neq 0 \text{ , } i = 1, 2$$

The p-values for the *MARRIED* are reported as 0.018 and the p-value for the *EDUC* is reported as 0.015. Thus each of the two variables is very significant in predicting *TV* given the other variable in the model.

The normal probability plot of standardized residuals is displayed below:

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: TV
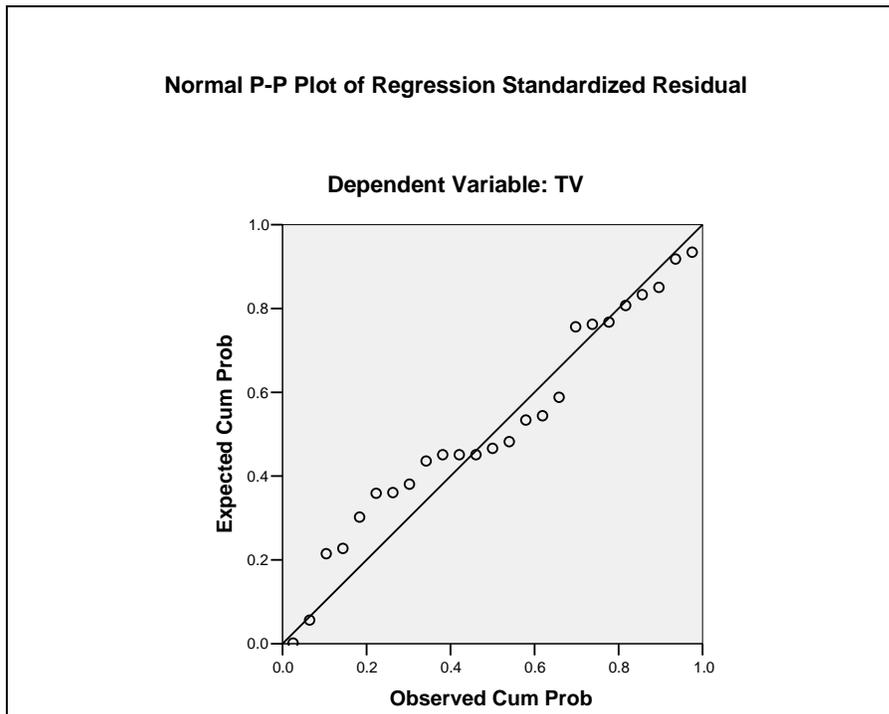
The above plot indicates some slight but not serious departures from the normality assumption. In general, regression models are quite robust for slight departures from the assumption of normality for residuals.

The plot of standardized residuals versus standardized predicted values is shown below.

Scatterplot

Dependent Variable: TV
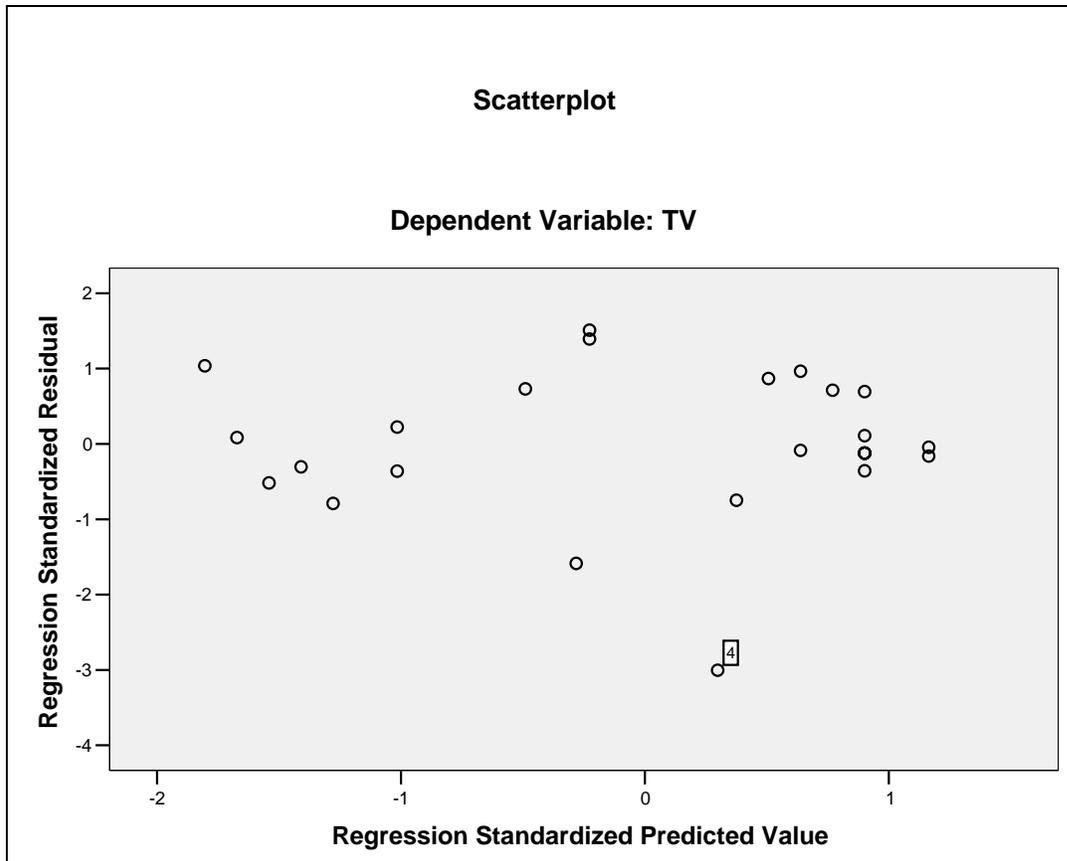
The pattern in the plot causes some concern, because the spread of residuals indicates that the assumption of equal variance may be violated. As a consequence, the inferences based on the model including tests and confidence intervals can be invalid.

There is an outlier, which is case #4. This case refers to a senior whose values of TV (0), AGE (90) and EDUCATION (2) are unusual relative to the other subjects in the sample.

	TV	Married	Age	Education	SRE_1	COO_1	LEV_1
1	.5	1	65	14	-.83671	.02828	.06809
2	.5	1	66	16	-.56105	.01789	.10570
3	.7	0	65	15	-1.81079	.33051	.19218
4	.0	1	90	2	-3.66420	2.18406	.28796
5	.8	1	68	15	-.32491	.00499	.08424
6	.9	1	79	17	.09272	.00060	.13246
7	1.1	1	70	12	-.37896	.00483	.05170
8	1.6	1	83	18	1.16336	.11599	.16453
9	1.6	1	81	12	.23367	.00184	.05170
10	2.0	0	72	10	-.78467	.02080	.05201
11	2.5	1	74	8	.77977	.02831	.08256
12	2.8	0	71	8	-.08843	.00021	.03307
13	2.8	0	71	6	-.37101	.00374	.03534
14	3.0	0	80	6	-.12813	.00045	.03534
15	3.0	0	73	6	-.12813	.00045	.03534
16	3.0	0	75	6	-.12813	.00045	.03534
17	3.2	0	76	4	-.16990	.00106	.05883
18	3.2	0	78	6	.11475	.00036	.03534
19	3.3	1	79	6	1.52883	.15936	.12981
20	3.3	0	79	4	-.04689	.00008	.05883
21	3.4	1	78	6	1.65700	.18720	.12981
22	3.5	0	76	9	.90501	.02370	.03988
23	3.6	0	75	7	.74023	.01408	.03155
24	3.7	0	72	8	1.00319	.02644	.03307
25	3.7	0	80	6	.72194	.01416	.03534

The case #4 has a relatively high studentized residual of -3.66 (<-2), large Cook's distance of 2.18 (>1), and high leverage value of 0.29 (>2×3/25=0.24). Therefore, the case #4 may be considered an outlier (high studentized residual) and has high potential to be an influential case.

In order to see whether the case is indeed influential, let us remove the case, and rerun the regression. This time, we will try the forward selection procedure. You will see that indeed excluding the case 4 will change substantially the estimates and inferences. Regression outputs are shown below.

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	EDUC	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	AGE	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: TV

The following predictors are added to the model: *EDUC* and then *AGE*.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.882[a]	.779	.768	.54392
2	.919[b]	.844	.829	.46755

a. Predictors: (Constant), EDUC

b. Predictors: (Constant), EDUC, AGE

R square has increased to 0.844 after removing the influential case.

ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22.881	1	22.881	77.341	.000[a]
	Residual	6.509	22	.296		
	Total	29.390	23			
2	Regression	24.799	2	12.399	56.722	.000[b]
	Residual	4.591	21	.219		
	Total	29.390	23			

a. Predictors: (Constant), EDUC

b. Predictors: (Constant), EDUC, AGE

c. Dependent Variable: TV

Overall, the final model is significant with $F_{2,21} = 56.722$ and P value = 0.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.553	.268		16.965	.000
	EDUC	-.229	.026	-.882	-8.794	.000
2	(Constant)	-.020	1.561		-.013	.990
	EDUC	-.207	.024	-.797	-8.764	.000
	AGE	.059	.020	.269	2.962	.007

a. Dependent Variable: TV

According to the Coefficients table above, the estimated regression equation is

$$\mu(TV) = -0.02 - 0.207 \times EDUC + 0.059 \times AGE$$

The p-value for the *EDUC* is reported as 0.00 and the p-value for the *AGE* is reported as 0.007. Thus each of the two variables is very significant in predicting the response variable *TV* given the other variable in the model.

The normality plot of residuals and the plot of standardized residuals versus standardized predicted values should be obtained to verify the model assumptions.

Henryk Kolacz
University of Alberta
September 2012