SEX DISCRIMINATION PROBLEM

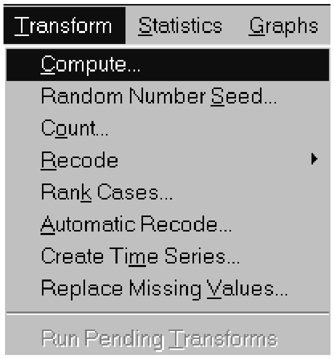
12. Multiple Linear Regression in SPSS

In this section we will demonstrate how to apply the multiple linear regression procedure in SPSS to the sex discrimination data. The numerical outputs and graphical displays for the data are displayed in **Section 8**.

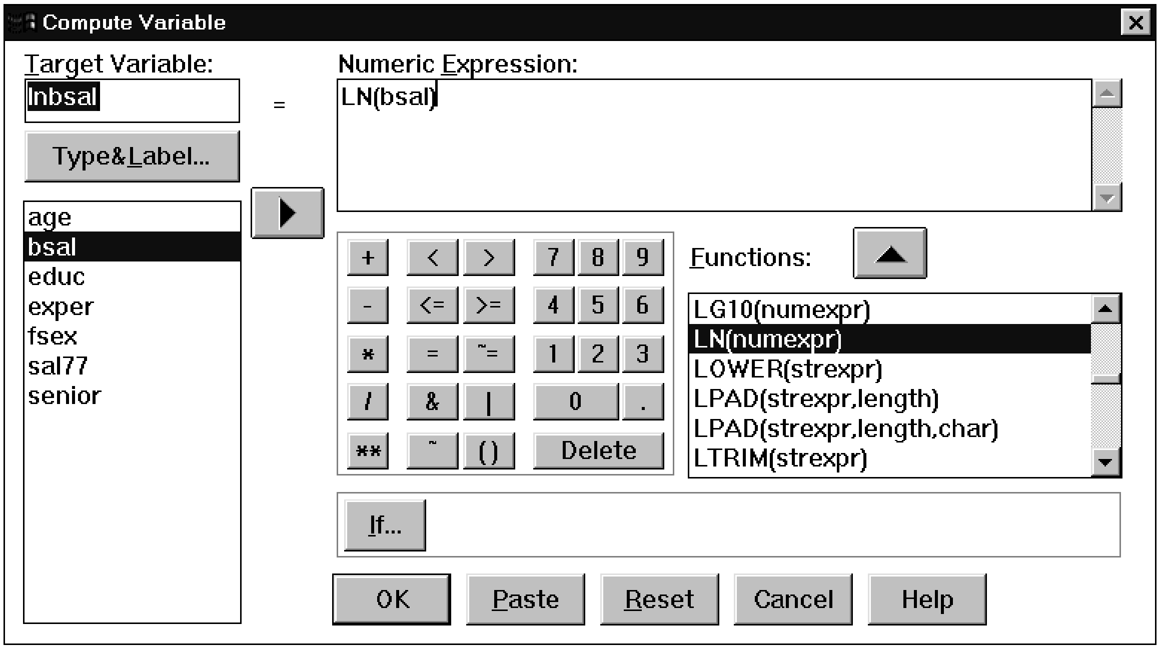
The data for the problem are available in the SPSS file *sex.sav* located on the FTP server in the STAT 252 directory. The instructions how to download the data files using FTP are available in the ***Introduction to SPSS*** module in STAT 252 Web site (*Appendices*).

The preliminary analysis indicated that in order to apply a linear regression procedure, we have to express the dependent variable *BSAL* on a log scale (*LNBSAL*). Moreover, it is necessary to transform the experience variable EXPER into a new variable TREXP that reveals a linear relationship with the log-transformed dependent variable *LNBSAL*.

As the linear regression model for the data will be developed for the log-transformed values of the response variable (*BSAL)*, we will demonstrate first how to carry out the transformation with SPSS. Click on *Transform* in the menu and then on *Compute…*



The *Compute Variable* dialog box is displayed. In the *Target Variable* box enter the name of the new variable as *LNBSAL*. Transfer the natural logarithm function LN from the *Functions* box to the *Numeric Expression* box by clicking on the function name and then on >. Then move in the same way the variable *BSAL* to the *Numeric Expression* box. The dialog box should have the following form:

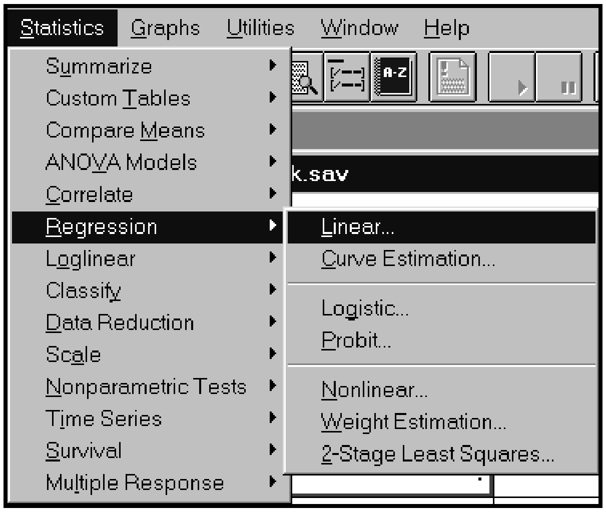


Compute Variable

Target Variable: lnbsal = Numeric Expression: LN(bsal)

Click on OK. As a result, the new variable *LNBSAL* is obtained in one of the columns of the data file. In the same way, you can obtain the new variable TREXP, which is defined as *1/(EXPER+12).*

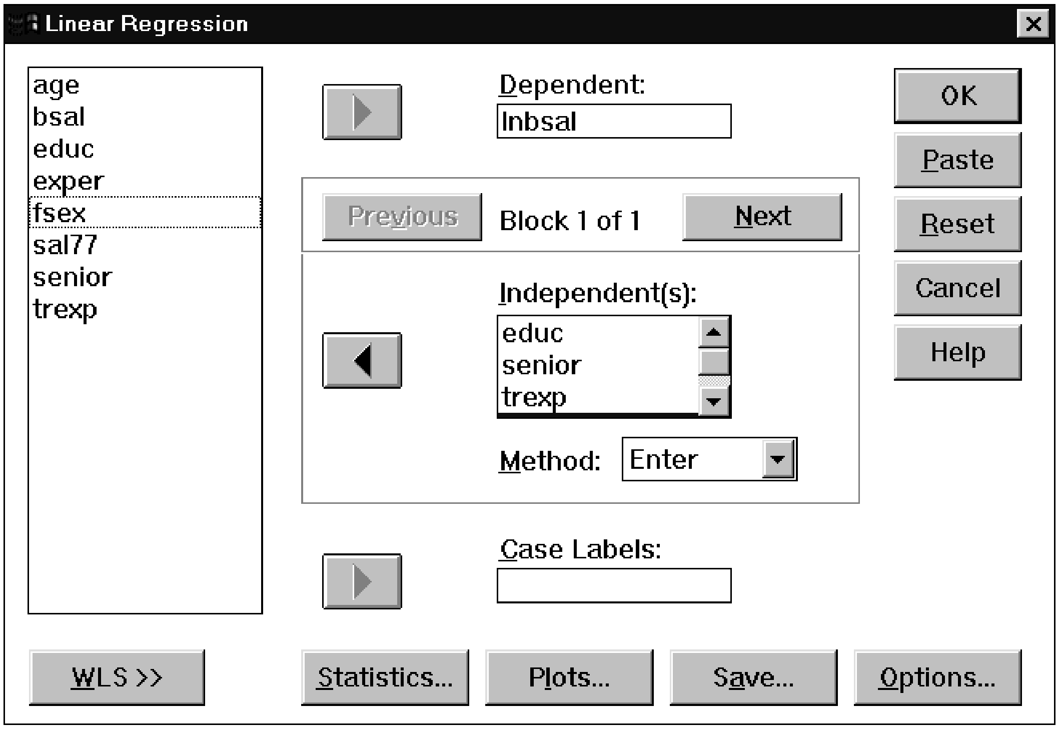
In order to access the simple linear regression procedure in SPSS, choose *Statistics* and then *Regression*. The following menu will be displayed:



Click on *Linear* to open the *Linear Regression* dialog box. The nine variable names in the data file will appear in the left-hand box: *BSAL, SAL77, FSEX, SENIOR, AGE, EDUC, EXPER*, *LNBSAL*, and *TREXP*.

We will carry out the regression of *LNBSAL* on the following independent variables: *EDUC, SENIOR*, *TREXP*, and *FSEX*. The variable *AGE* is not included in the list of the independent variables due to its nonlinear relationship with the response variable and relatively high correlation with the transformed experience.

Move the variable *LNBSAL* to the *Dependent* variable dialog box by clicking on it and then on the arrow at the box. In the same way move each of the four independent variables listed above to the *Independent(s)* variables dialog box. The Linear Regression dialog box should be filled out as follows:



Linear Regression

age
bsal
educ
exper
fsex
sal77
senior
trexp

Dependent: lnbsal

Block 1 of 1

Independent(s):
educ
senior
trexp

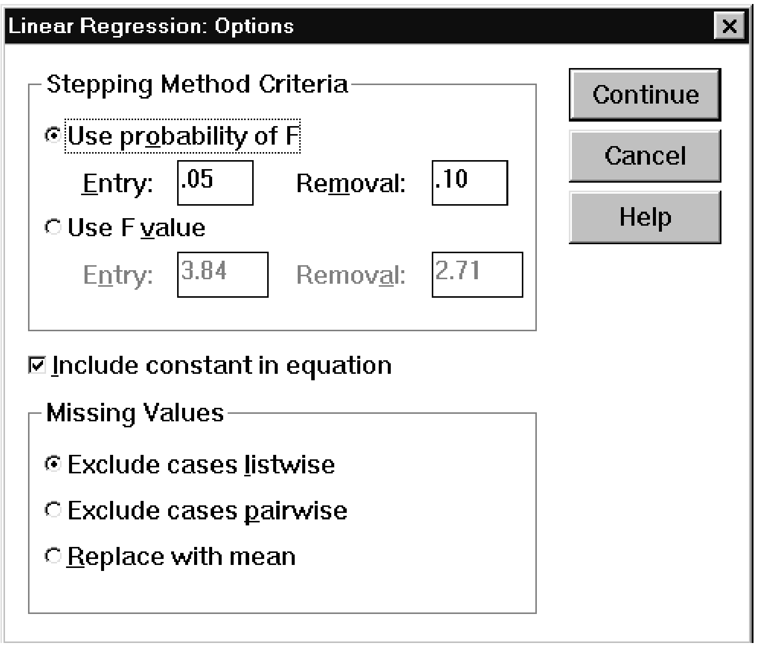
Method: Enter

Case Labels:

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

To enter the variables in the block in a single step, select *Enter*. To remove the variables in the block in a single step, select *Remove*. *Forward* variable selection enters the variables in the block one at a time based on entry criteria. *Backward* variable elimination enters all of the variables in the block in a single step and then removes them one at a time based on removal criteria. *Stepwise* variable entry and removal examines the variables in the block at each step for entry or removal. This is a forward stepwise procedure.

Variables can be entered or removed from the model depending on either the significance level (probability) of the F value, or the F value itself. In order to specify the significance levels for entry and removal of a variable click on *Options:*



Linear Regression: Options

Stepping Method Criteria

● Use probability of F

Entry: .05 Removal: .10

○ Use F value

Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values

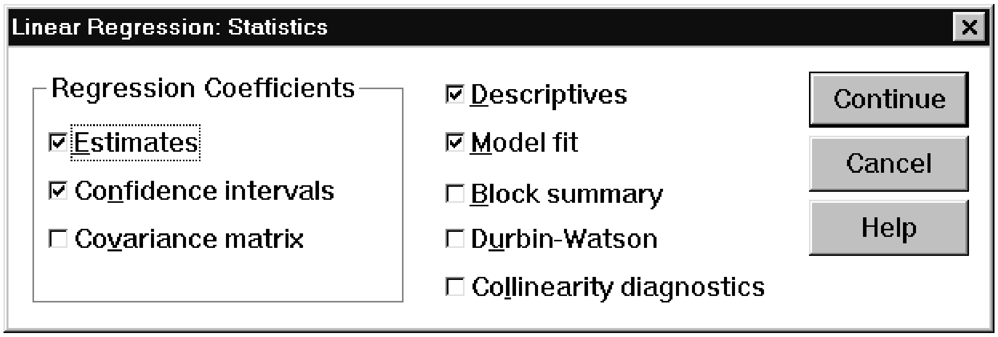
● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

The default significance values are 0.05 to enter a variable into the model, and 0.10 to remove a variable from the model. You may wish to specify different significance levels. These options apply when either the forward, backward, or stepwise variable selection has been specified.

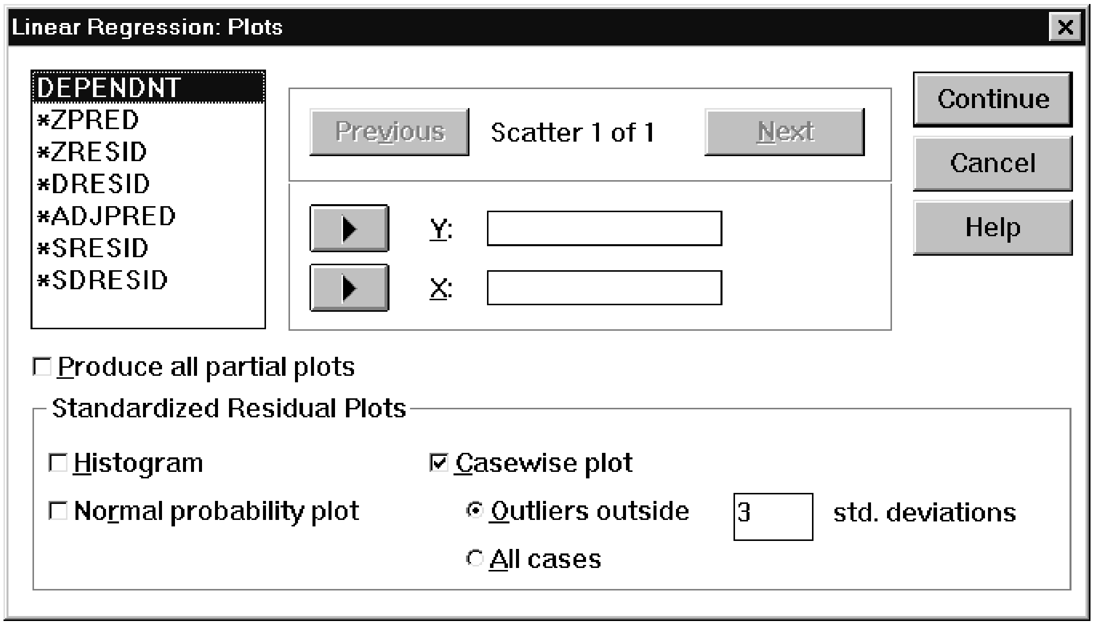
Click on *Statistics…* tab to obtain statistics for the linear regression. *Estimates* of the regression coefficients and *Model fit* are default options. In order to obtain the mean and standard deviation for each variable and a correlation matrix, check the *Descriptives* box. It is also very useful to obtain the 95% confidence intervals for the regression coefficients.



Regression Coefficients
- ☑ Estimates
- ☑ Confidence intervals
- □ Covariance matrix

- ☑ Descriptives
- ☑ Model fit
- □ Block summary
- □ Durbin-Watson
- □ Collinearity diagnostics

Click on *Collinearity diagnostics* to obtain an extensive collinearity statistics (tolerance, VIF, regression coefficient variance-decomposition matrix). Then click on *Continue* to return to the *Linear Regression* dialog box. Information about residuals is obtained by clicking on the *Plots* tab. The following *Linear Regression: Plots* dialog box opens.



□ Produce all partial plots

Standardized Residual Plots
- □ Histogram
- □ Normal probability plot
- ☑ Casewise plot
  - ● Outliers outside 3 std. deviations
  - ○ All cases

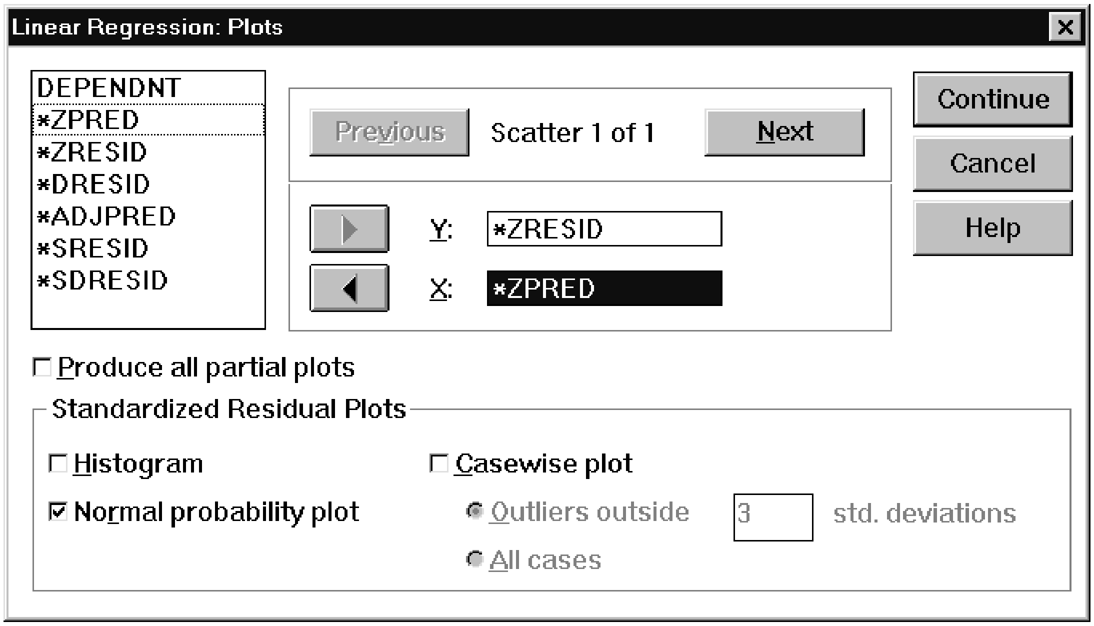
Plots can aid in the validation of the assumptions of normality, linearity, and equality of variances. Plots also allow you to check whether there are any cases, which might be considered as outliers and so dropped from the analysis. Click on the *Casewise plot* check box to obtain a listing of any exceptionally large residuals. We recommend that this is done for an initial run of the procedure. Click on *Continue* and the on OK to run the regression for the first time.

No outliers have been found in the sex discrimination data. The listing states: *No outliers found. No casewise plot produced.*

Now we can request a plot of residuals against the predicted values. The plot allows you to check whether the assumption of constant variance is not violated. We then look for a change in the spread or dispersion of the plotted points.

SPSS creates several temporary variables (prefaced with *) during execution of a regression analysis. *PRED comprises the unstandardized predicted values, *RESID is the set of unstandardized residuals, *ZPRED contains the standardized predicted values (i.e. *PRED has been transformed to a scale with mean 0 and standard deviation of 1), and *ZRESID comprises the standardized residuals (i.e. *RESID standardized to a scale with mean 0 and standard deviation of 1).

In order to obtain a plot of residuals against the predicted values, click on *ZRESID and then on > to transfer the variable to the Y: box, and on *ZPRED and then on > to transfer the variable to the X: box. The completed *Linear Regression: Plots* dialog box is shown below.



Linear Regression: Plots

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

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

□ Produce all partial plots

Standardized Residual Plots

□ Histogram

☑ Normal probability plot

□ Casewise plot

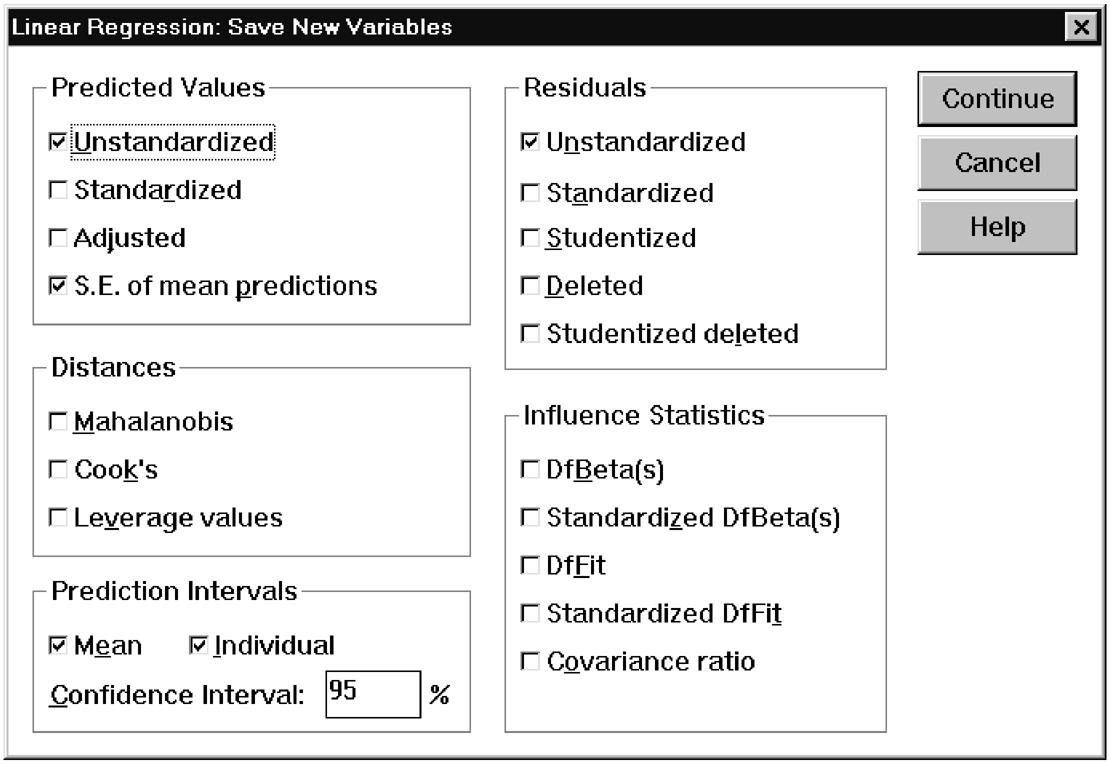
● Outliers outside 3 std. deviations

○ All cases

The normality assumption can be verified by looking at the plot of residuals. In order to assess whether the normality assumption is not violated with SPSS, the normal P-P plot of regression standardized residuals is obtained. The plot plots the cumulative proportions of standardized residuals against the cumulative proportions of the normal distribution. If the normality assumption is not violated, points will cluster around a straight line. In order to obtain the plot, check the *Normal probability plot* box.

Click on *Continue* to return to the *Linear Regression* dialog box. Click on *Save…* to save predicted values, residuals, and other statistics useful for diagnostics. Each selection adds one or more new variables to your active data file. In particular, the predicted values (mean estimates) are obtained as the variable pre_1. Studentized residuals are used for flagging outliers, and leverages and Cook's distances for flagging influential cases. Clicking on *Studentized* creates a new variable sre_1 in the original data file containing the studentized residuals. Clicking on *Cook's* and *Leverage values* produces two other variables: cook_1 and lev_1 containing the Cook's distances and leverage values for the data, respectively.

The *Linear Regression: Save New Variables* dialog box is displayed below.



Linear Regression: Save New Variables

Predicted Values

☑ Unstandardized
□ Standardized
□ Adjusted
☑ S.E. of mean predictions

Distances

□ Mahalanobis
□ Cook's
□ Leverage values

Prediction Intervals

☑ Mean ☑ Individual
Confidence Interval: 95 %

Residuals

☑ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Influence Statistics

□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

Click on *Continue* to return to the *Linear Regression* dialog box. Click on *OK* to obtain the linear regression output. The numerical output is discussed in **Section 8.** The plots obtained to validate the assumptions of the multiple linear regression model are discussed in **Section 9**.