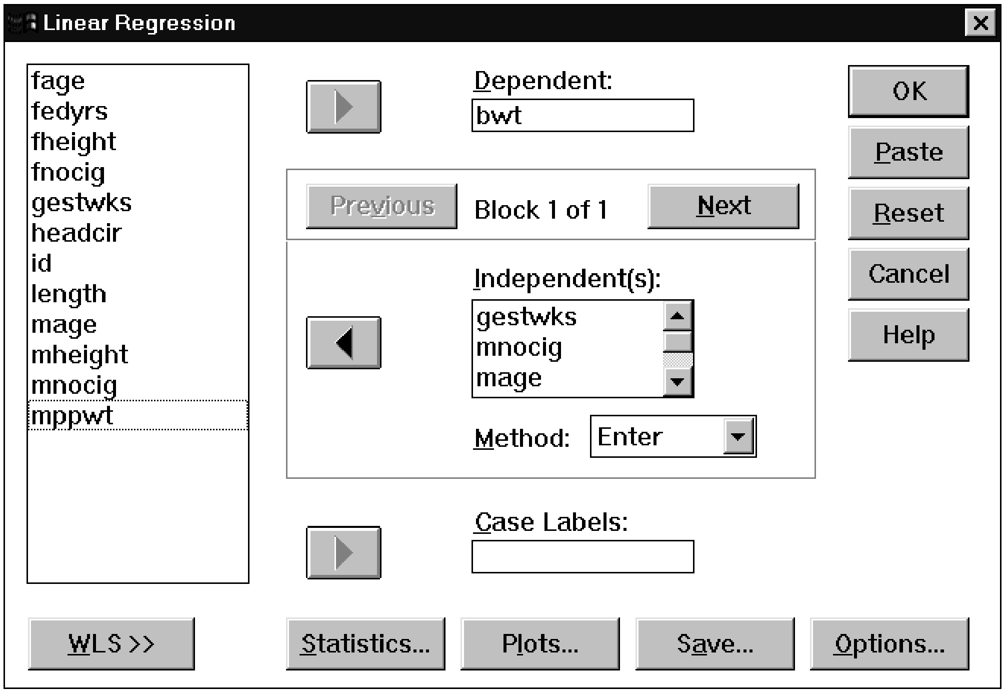
13. Multiple Regression Diagnostics in SPSS

In this section we will summarize various diagnostic tools available in SPSS to evaluate the adequacy of a multiple linear regression model. These tools include residual plots to investigate whether the assumptions of linearity of the dependent variable, and normality and constant variance of the error appear to be met. Some of these tools were discussed in the previous section for the child health and development data.

The regression diagnostics in SPSS can be requested from the *Linear Regression* dialog box. The box for the child health and development data is displayed below:


Linear Regression

Dependent: bwt

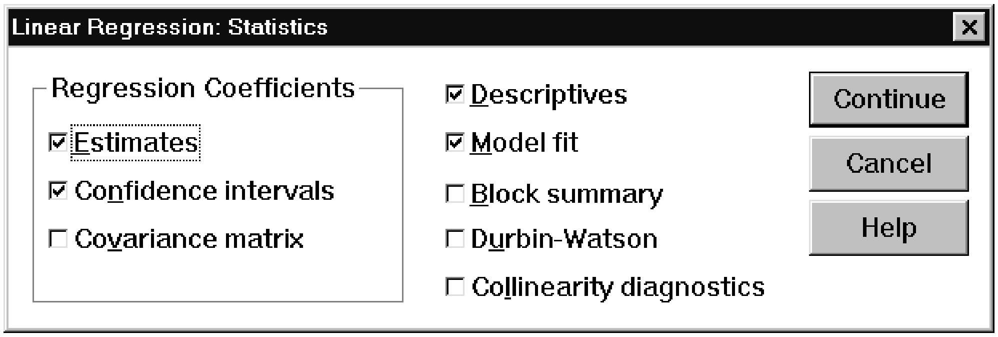
Block 1 of 1

Independent(s): gestwks, mnocig, mage

Method: Enter

Case Labels:

Click on *Statistics…* tab to obtain *Linear Regression: Statistics* dialog box displayed below:


Linear Regression: Statistics

Regression Coefficients

☑ Estimates

☑ Confidence intervals

□ Covariance matrix

☑ Descriptives

☑ Model fit

□ Block summary

□ Durbin-Watson

□ Collinearity diagnostics

In order to obtain some statistics useful for diagnostics, check the *Collinearity diagnostics* box. As a result, eigenvalues condition indices, tolerances, VIF values, and regression coefficient variance-decomposition matrix will be provided.

Eigenvalues provide an indication of how many distinct dimensions there are among the independent variables. When several eigenvalues are close to 0, the variables are highly intercorrelated and the matrix is said to be *ill-conditioned*; small changes in the data values may lead to large changes in the estimates of the coefficients.

Condition indices are the square roots of the ratios of the largest eigenvalue to each successive eigenvalue. A condition index greater than 15 indicates a possible problem and an index greater than 30 suggests a serious problem with collinearity.

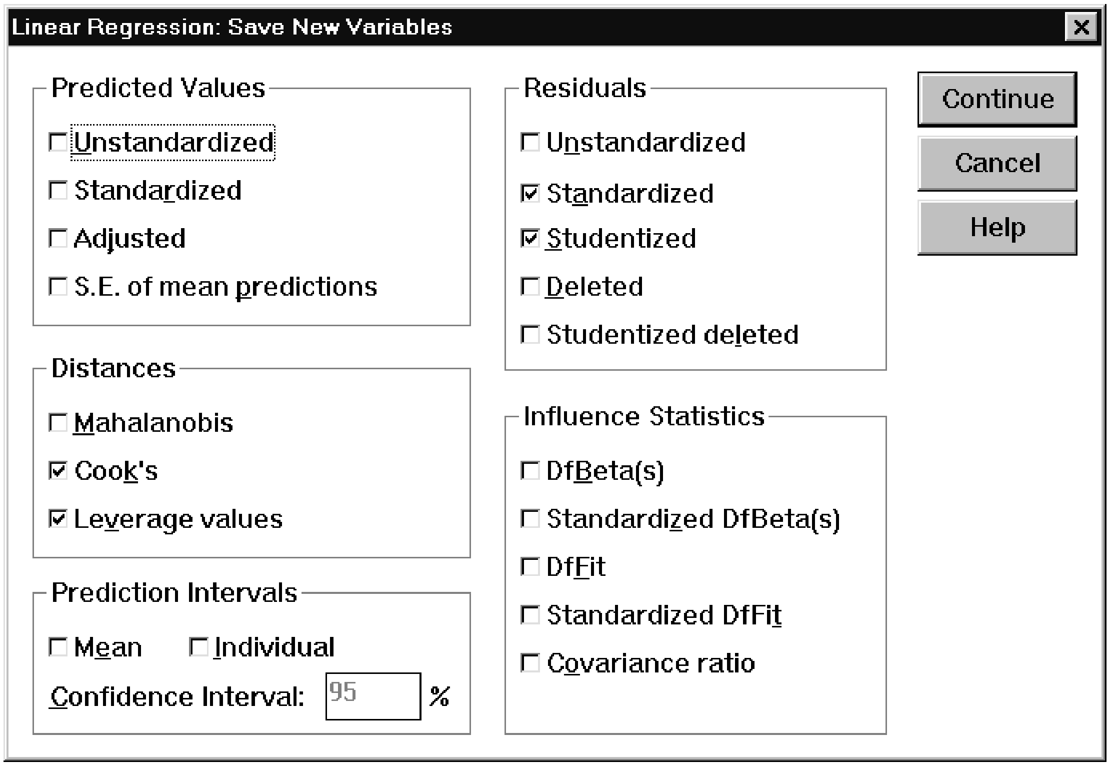
Tolerance is the amount of variability of the selected independent variable not explained by the other independent variables. It is obtained by making each independent variable a dependent variable and regressing it against the remaining independent variables. Tolerance values approaching zero indicate that the variable is highly collinear with the other predictor variables.

Variance inflation factor (VIF) is inversely related to the tolerance value: $VIF = 1/TOLERANCE$. Large VIF values (a usual threshold is 10.0, which corresponds to a tolerance of .10) indicate a high degree of collinearity or multicollinearity among the independent variables.

Regression coefficient variance-decomposition matrix shows the proportion of variance for each regression coefficient (and its associated variable) attributable to each condition index.

In order to examine collinearity, we first identify all condition indices above the threshold value of 30. Then for all condition indices exceeding the threshold, we identify variables with variance proportions above 0.90. A collinearity problem is indicated when a condition index identified as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more coefficients*. Thus each row in the matrix with the proportions exceeding 0.90 for at least two coefficients indicates significant correlations among the corresponding variables.

Click on *Continue* to return to the *Linear Regression* dialog box. Then click on *Save…* tab to obtain *Linear Regression: Save* dialog box displayed below:



Linear Regression: Save New Variables

Predicted Values
□ Unstandardized
□ Standardized
□ Adjusted
□ S.E. of mean predictions

Distances
□ Mahalanobis
☑ Cook's
☑ Leverage values

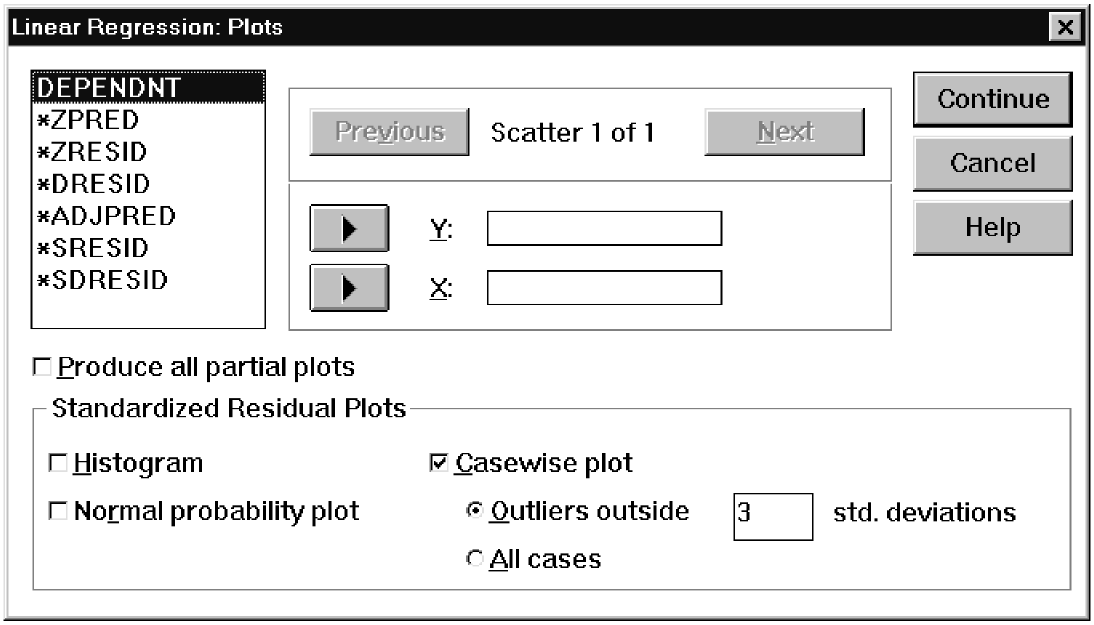
Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Residuals
□ Unstandardized
☑ Standardized
☑ Studentized
□ Deleted
□ Studentized deleted

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

This dialog box allows you to save predicted values, residuals, distances, prediction intervals, and influence statistics as new variables in the *Data Editor*. 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. You can obtain plots of some of these variables by using *Scatter* option on the *Graphs* menu.

Now 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

□ Histogram

□ Normal probability plot

☑ Casewise plot

● Outliers outside 3 std. deviations

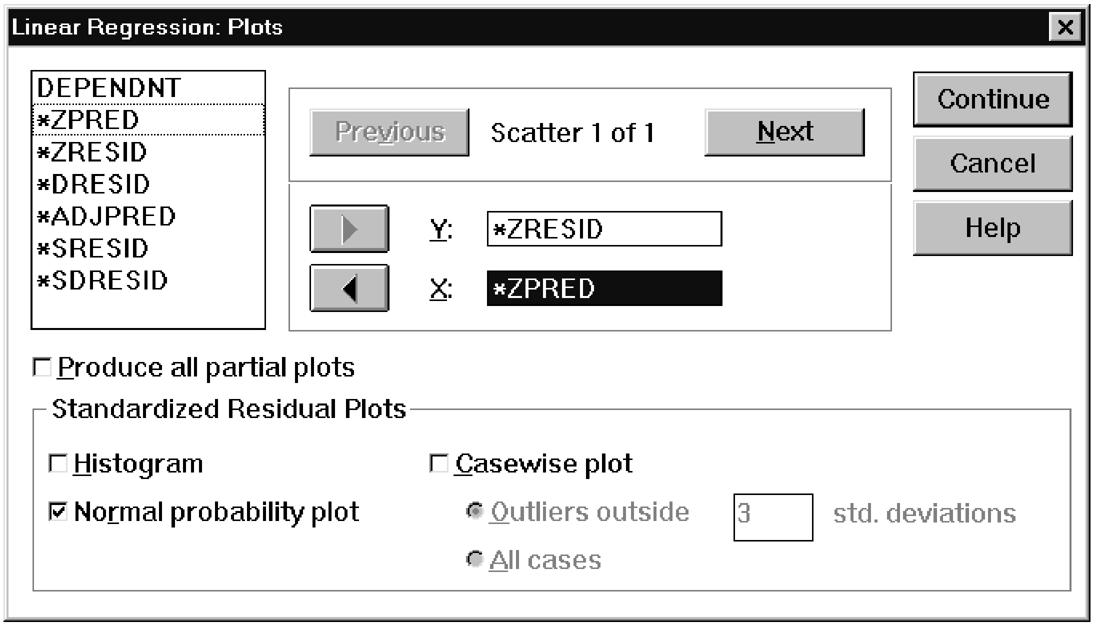
○ All cases

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.

Now we can request a plot of residuals against the predicted values. The plot allows you to check whether the assumptions of linearity and constant variance are not violated. We then look for any departures from a linear pattern and 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.



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 *OK* to obtain the linear regression output. The numerical output is discussed in **Section 7.** The plots obtained to validate the assumptions of the multiple linear regression model are discussed in **Section 9**.