Chapter 1

Using General Linear Regression Models (GLM)

Background Information

MANOVA, ANOVA, ANCOVA, and Multiple Regression models belong to a family of models known as Generalized Linear Models (GLM), which were first introduced by Nelder and Wedderburn (1972). General linear models are a common set of statistical assumptions upon which regression, correlation, and analysis of variance are based; that is, the full range of methods analysts use to study one or more continuous dependent variables and one or more independent variables, whether they are continuous or categorical.

The basic concept of GLM is that the relationship between the dependent variables and the independent variables is expressed as an equation that contains a term for the weighted sum of the values of the independent variables, plus an error term for unexplained effects.

General Linear Models in STATGRAPHICS *Plus*

The General Linear Models Analysis in STATGRAPHICS *Plus* allows you to estimate linear statistical models that relate one or more dependent variables to one or more independent variables. You can use the analysis for regression, analysis of variance, multifactor ANOVA, or for an analysis of covariance.

The independent variables can be continuous or categorical; categorical variables can be crossed or nested. While STATGRAPHICS *Plus* contains other analyses such as Multiple Regression and Multifactor ANOVA, which can fit certain types of linear models, the General Linear Models Analysis is more flexible, particularly for models that involve both continuous and categorical factors.

The General Linear Models Analysis in STATGRAPHICS *Plus* contains many features, including the following.

- The program allows you to specify more than one dependent variable. It fits the model you choose to each of the dependent variables. You can also perform a MANOVA. The program allows for crossed and nested effects.

- The program can create tables that contain the predicted and the residual values from an analysis.

- The program allows you to choose the error term that is associated with each effect in the model.

- The program allows you to specify which effects are random rather than fixed.

You use the General Linear Models Analysis:

- when you have both crossed and nested factors, such as in split plot and repeated measures designs

- when you want to enter your own contrasts in an analysis of variance

- when you want to obtain MANOVA statistics for multiple dependent variables.

Table 1-1 lists some of the analyses and types of variables you use with them, provides samples of effects, and also indicates when there are other analyses you might want to use instead of the General Linear Models Analysis.

Table 1-1. General Linear Models Analyses in STATGRAPHICS Plus

Type of Model	*Type of Dependent Variable*	*Example of Sample Effects*
Simple Regression*	D1	Q1
Multiple Regression*	D1	Q1, Q2, Q3...
Polynomial Regression*	D1	Q1, Q1*Q1, Q1*Q1*Q1 ...
Linear Regression*	D1	Q1, Q1*Q2, Q2*Q2 ...
One-Way ANOVA	D1	C1
Multifactor ANOVA:		
First Order*	D1	C1, C2, ...
Second Order*	D1	C1, C2, C1*C2, ...
Sample Effects:		
Nested*	D1	C1, C2(C1), C3(C1 C2), ...
Three-Way Factorial*	D1	C1, C2, C3, C1*C2, C2*C3, C1*C3, C1*C2*C3
Designed Experiments**	D1	Screening, Response Surface, and Mixture Designs
More General Models:		
Analysis of Covariance*	D1	C1, C2, C1*C2, ..., Q2, Q3, ...

Table 1-1. Continued

Type of Model	*Type of Dependent Variable*	*Example of Sample Effects*
Separate Slopes	D1	C1, Q1(C1) - Quantitative Nested in Categorical
Quantitative by Categorical***	D1	Q1*C1... - Quantitative Crossed with Categorical
MANOVA	D1, D2, ...	Any of the Above

*You might want to use analyses from the Compare or Relate menus in the Standard Edition instead.

**You might want to use the Experimental Design portion of the Quality and Design product instead.

***You might want to use the Comparison of Regression Lines Analysis instead.

D*n* represents a dependent variable, Q*n* represents a quantitative variable, C*n* represents a categorical variable.

All models include a constant unless you remove it using Analysis Options.

The next two sections contain information about how you specify effects and how the program handles missing values.

Specifying Effects

General linear models are made up of terms you define using special annotations that represent the names of variables and operators. In the General Linear Models Analysis, STATGRAPHICS *Plus* uses two types of independent variables: categorical and quantitative, and two operators, * (Cross) and ( ) (Nest). How you use the operators depends on whether the factors are categorical or quantitative. For example, if the model contains three categorical factors — C1, C2, and C3 — and three quantitative factors — Q1, Q2, and Q3 — the factors are named A, B, C, D, E, and F according to the order you enter them into the General Linear Models Analysis dialog box. The Factors list box on the GLM Model Specification dialog box displays the names of the factors and their assigned letters; the Effects list box, by default, contains the letters for all the main effects (see Figure 1-2). The GLM Model Specification dialog box also allows the user to specify which effects are random.

You can delete a main effect and you can add additional terms to a model by typing them into the Effects list box. You also can create many different types of models by

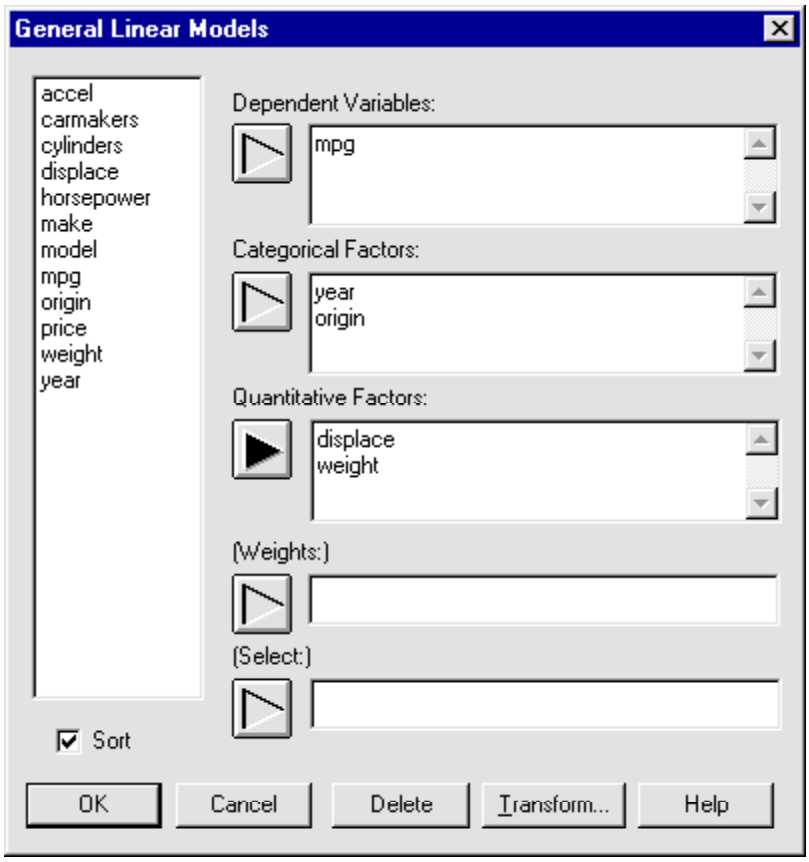


Figure 1-1. Main Effects Listed by Default

using a variety of effects. For example, for these four rows of effects: A, B, A*B, and C(A), the model looks like this:

A + B + A*B + C(A)

By default, the program adds a constant term to the model. To remove the constant, use the GLM Model Specification dialog box. You can specify up to 26 factors and 256 effects. Notice that Figure 1-2 shows an example of the GLM Model Specification dialog box with various types of effects entered into it.

The seven types of effects you can use with the factors are:

- regressor effects that involve a single quantitative factor, such as D or E
- polynomial effects that are products of two or more quantitative factors, such as D*E or E*E

- main effects that involve a single categorical factor, such as A

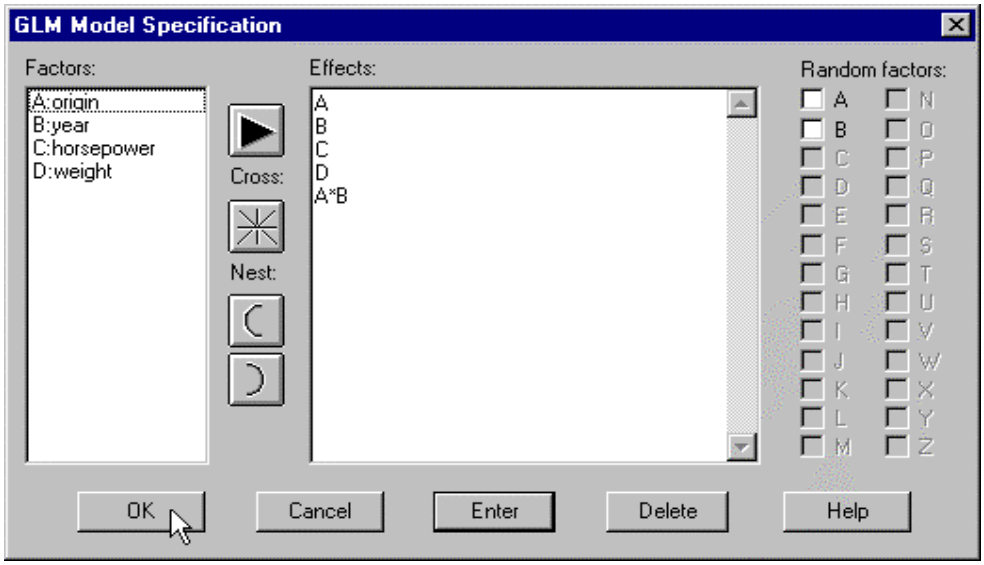


Figure 1-2. GLM Model Specification Dialog Box

- crossed effects that involve two or more categorical factors, such as A*C or A*B*C
- a quantitative variable that is crossed with a categorical variable, such as A*D
- nested categorical effects, such as C(B) or C(B A)
- quantitative variables, nested within categorical variables, such as F(B).

Treating Missing Values

When you choose more than one dependent variable, the program eliminates the missing values from each variable separately, unless you are performing a MANOVA. If you are performing a MANOVA, the program bases all the models on cases that have no missing values for any of the dependent variables.

To access the analysis, choose: SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display the Analysis dialog box (see Figure 1-1, above). After you complete this dialog box and click OK, the GLM Model Specification Dialog Box displays (see Figure 1-2, above).

Tabular Options

Analysis Summary

The Analysis Summary option displays the results of fitting the general linear statistical model (see Figure 1-3). If any *p*-value in the first ANOVA table is less than 0.10, there is a statistically significant relationship between the variables. The second ANOVA table tests the statistical significance for each of the factors. If any *p*-value in this ANOVA table is greater than or equal to 0.10, that term is not statistically significant at the 90 percent or higher confidence level and you should consider removing it from the model.

The top portion of the resulting Analysis Summary shown in Figure 1-3, contains the following information:

Analysis of Variance - shows a decomposition of the sum of squares into components for the model and the residuals. The F-test tests the statistical significance of the model as a whole. A small P-value (less than 0.05) indicates that at least one factor in the model is significantly related to the dependent variable at the 5% significance level. In the current example, the model is highly significant.

Type III Sums of Squares - shows a decomposition of the model sum of squares into components for each factor. Based on the settings specified on the Analysis Options dialog box, either Type III or Type I sums of squares are displayed. Small P-values indicate significant effects. In this example, all four effects are highly significant

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 3
Number of quantitative factors: 0


Analysis of Variance for Heart Rate
-------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------
Model                     4487.94     32      140.248      18.83       0.0000
Residual                  469.219     63      7.44792
-------------------------------------------------------------------------------
Total (Corr.)             4957.16     95

Type III Sums of Squares
-----------------------------------------------------------------------------------
Source                      Sum of Squares     Df   Mean Square    F-Ratio    P-Value
-----------------------------------------------------------------------------------
Drug                                1333.0      2         666.5       5.99     0.0088
Person(Drug)                       2337.91     21       111.329      14.95     0.0000
Time                               289.615      3       96.5382      12.96     0.0000
Drug*Time                          527.417      6       87.9028      11.80     0.0000
Residual                           469.219     63       7.44792
-----------------------------------------------------------------------------------
Total (corrected)                  4957.16     95
```

Figure 1-3. Top Portion of the Analysis Summary

The middle portion of the *Analysis Summary* includes information about how the F-tests for each factor were constructed (see Figure 1-4).

It includes:

Expected Mean Squares - the expected mean square for each factor is determined using Hartley's (1967) *synthesis* method. The mean squares are labeled from top to bottom as (1), (2), etc. through (5), which corresponds to the mean square for the residuals. A term such as Q1 indicates a quantity unique to the factor in which it appears. The expected mean squares are important in constructing proper F-tests for models involving random factors.

F-Test Denominators - the mean square used as the denominator of the F-test for each factor, together with its degrees of freedom and how it was determined.

Variance Components - for models with random factors, estimates the variance component σ_j of each random effect. These are derived by equating the mean squares with their expected values, which is the method of moments described by Milliken and Johnson (1996). Variance components measure the variability in the response induced by variation in the random factors.

Expected Mean Squares

Source	EMS
Drug	(5)+4.0(2)+Q1
Person(Drug)	(5)+4.0(2)
Time	(5)+Q2
Drug*Time	(5)+Q3
Residual	(5)

F-Test Denominators

Source	Df	Mean Square	Denominator
Drug	21.00	111.329	(2)
Person(Drug)	63.00	7.44792	(5)
Time	63.00	7.44792	(5)
Drug*Time	63.00	7.44792	(5)

Variance Components

Source	Estimate
Person(Drug)	25.9702
Residual	7.44792

Figure 1-4. Middle Portion of the Analysis Summary

Other statistics include:

- The R-Squared statistic, which indicates the percentage of variability in the dependent variable for which the model accounts.

- The Adjusted R-Squared statistic, which is more suitable for comparing models that have different numbers of independent variables and indicates the percentage of variability represented by the model.

- The Mean Absolute Error (MAE) statistic, which is the average absolute value of the residuals.

- The Durbin-Watson statistic, which tests the residuals to determine if there is significant serial correlation based on the order in which the values fall in the file.

If you used the optional Select text box on the General Linear Models Analysis dialog box, the program uses the prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use one of the methods discussed in the topic, “Overview of the Model-Building Process," in Online Help.

The Residuals Table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.

- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the set of data, then finding the average or mean value for the sum of those squares. If the result is a small value, you can predict performance more precisely; if the result is a large value, you may want to use a different model.

- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more precisely; if the result is a large value, you may want to use a different model.

- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.

- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Each of these statistics is based on the residuals. The first three statistics measure the magnitude of the errors; better models have smaller values. The last two statistics measure the bias; better models have a value close to 0.0.

Use the *General Linear Models Options* dialog box to indicate the type of sums of squares that will be used, to choose the dependent variable that will display in certain panes, to indicate if the constant and the MANOVA values in the analysis will be included in the model, to allow for automatic transformation of the dependent variable using the Box-Cox transformation, to choose the error terms, and to display your choice of factors and error terms (see Figure 1-5).

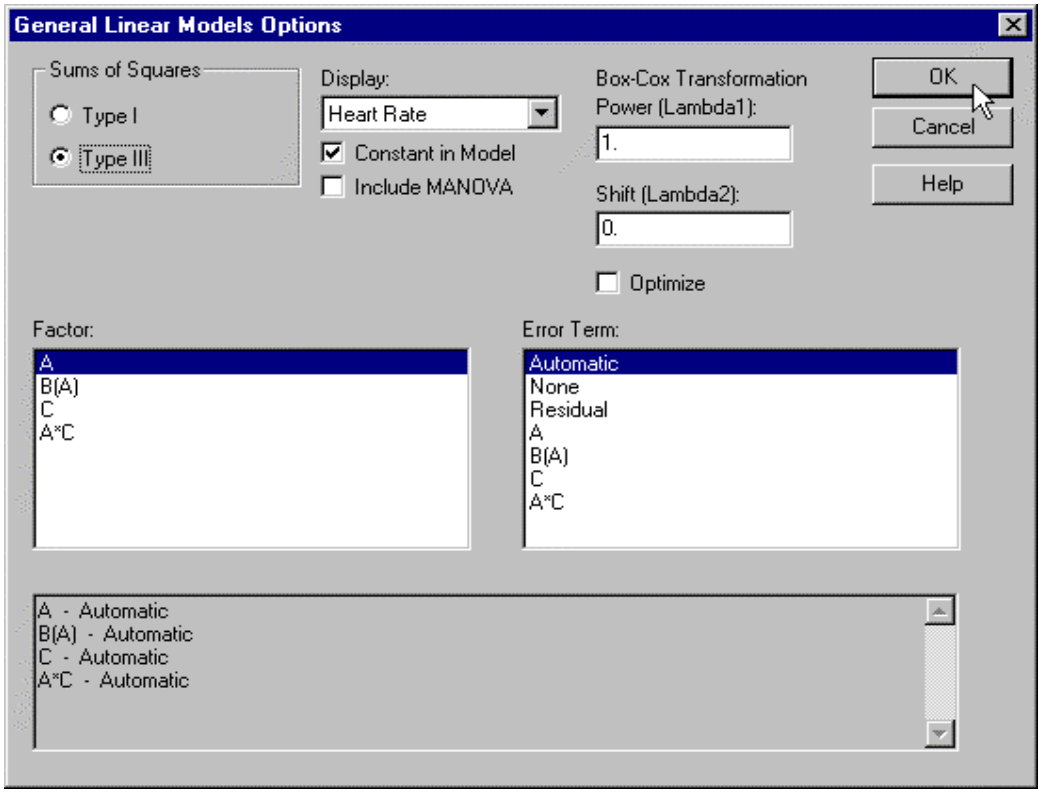
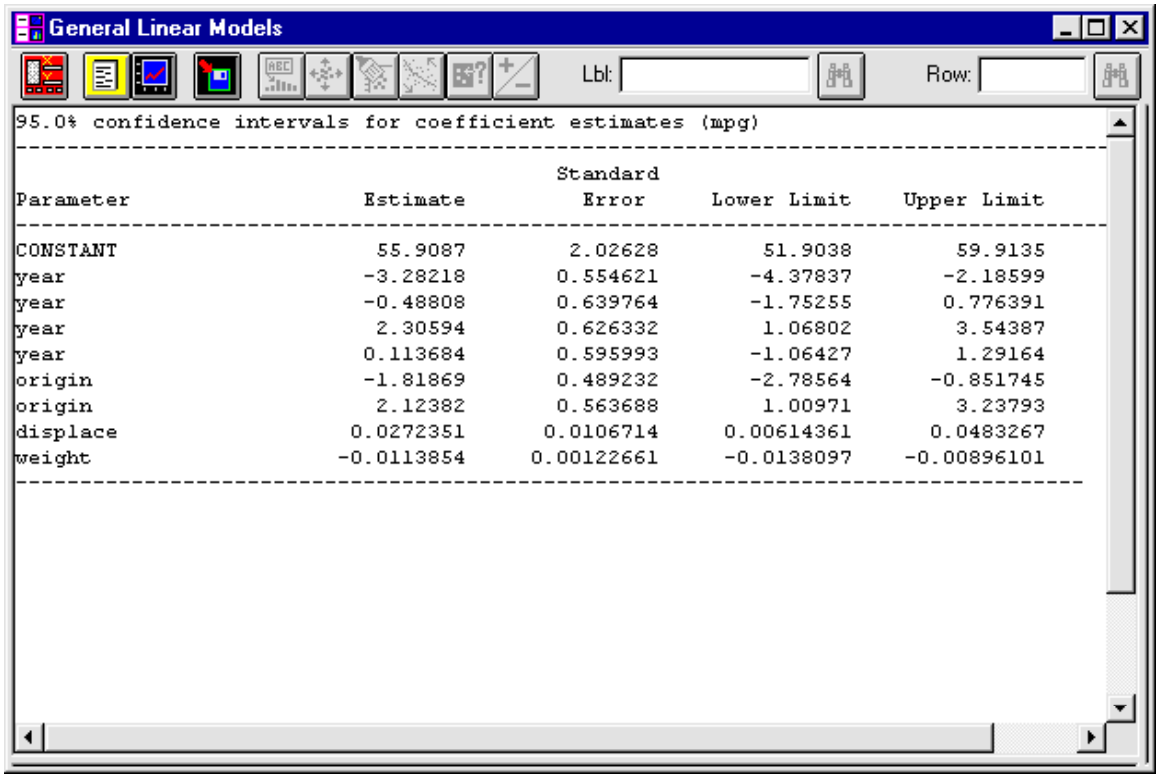


Figure 1-5. General Linear Models Options Dialog

Model Coefficients

The Model Coefficients option displays the coefficients in the model, including the confidence intervals (see Figure 1-6). The confidence intervals show the preciseness of the estimated coefficients given the amount of available data and noise in the model. The table also includes the Variance Inflation Factors (VIFs), which you can use to measure the extent to which the explanatory variables are correlated among themselves. VIF values above 10 usually indicate serious multicollinearity, which greatly increases the estimation error of the model coefficients when you compare them with an orthogonal sample.



Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	55.9087	2.02628	51.9038	59.9135
year	-3.28218	0.554621	-4.37837	-2.18599
year	-0.48808	0.639764	-1.75255	0.776391
year	2.30594	0.626332	1.06802	3.54387
year	0.113684	0.595993	-1.06427	1.29164
origin	-1.81869	0.489232	-2.78564	-0.851745
origin	2.12382	0.563688	1.00971	3.23793
displace	0.0272351	0.0106714	0.00614361	0.0483267
weight	-0.0113854	0.00122661	-0.0138097	-0.00896101

Figure 1-6. Model Coefficients Table

Use the *Model Coefficients Options* dialog box to choose the type of interval that will be used, to enter a value for the confidence level, and to indicate if the report should show the correlations.

Table of Means

The Table of Means option displays the least squares means for each level of the categorical factors as well as the standard error for each mean, which is a measure of its sampling variability (see Figure 1-7). The Lower Limit and Upper Limit columns display the confidence intervals.

Use the *Confidence Intervals Options* dialog box to change the confidence level.

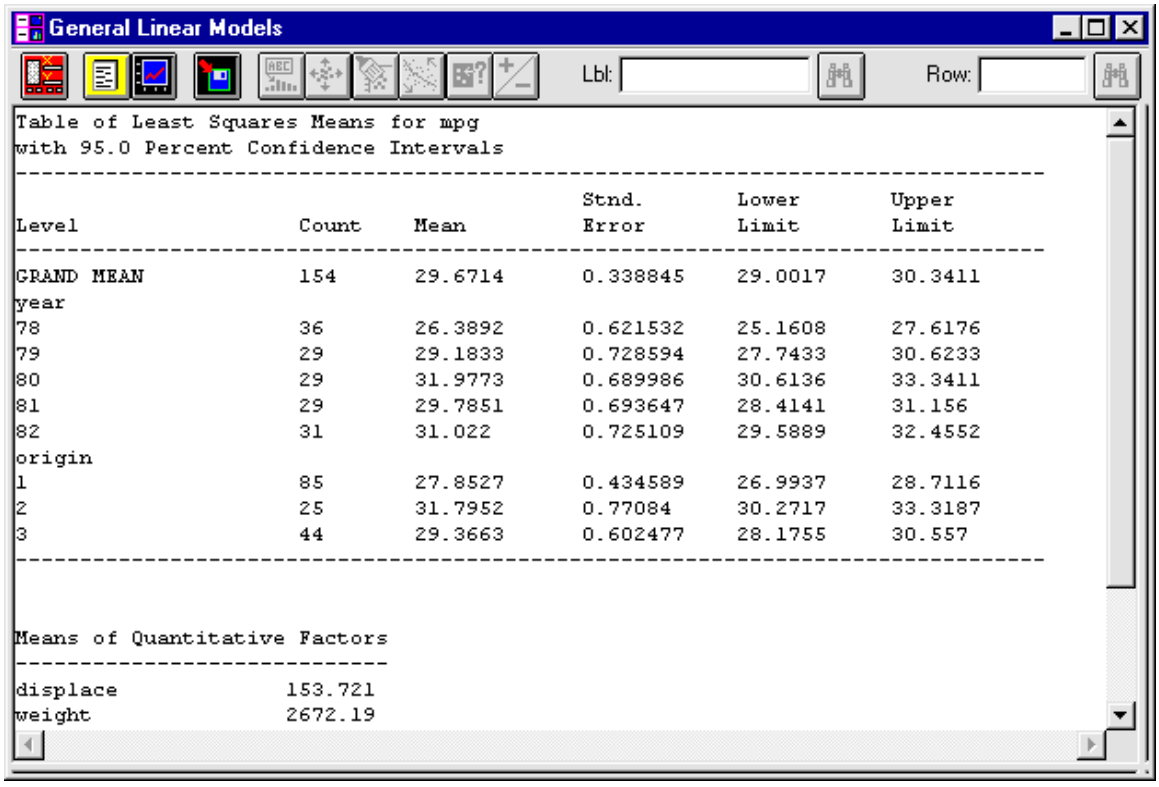


Figure 1-7. Table of Means Report

Multiple Range Tests

The Multiple Range Tests option displays the means that are significantly different from others in the analysis (see Figure 1-8). The program uses a multiple comparison procedure that you choose to calculate the limits. The upper portion of the table identifies the homogenous groups as columns of Xs. The levels within each column that contain an X form a group of means within which there are no statistically significant differences. The lower portion of the table shows the estimated difference between each pair of means. Pairs for which the differences are statistically significant at the 95 percent confidence level are marked with an asterisk. The program uses the currently chosen method to discriminate among the means. Determining the percentage of risk for each pair of means varies depending upon the method you use.

Use the *Multiple Comparison Options* dialog box to choose the options that will be used to perform the multiple range tests when the model includes at least one categorical factor.

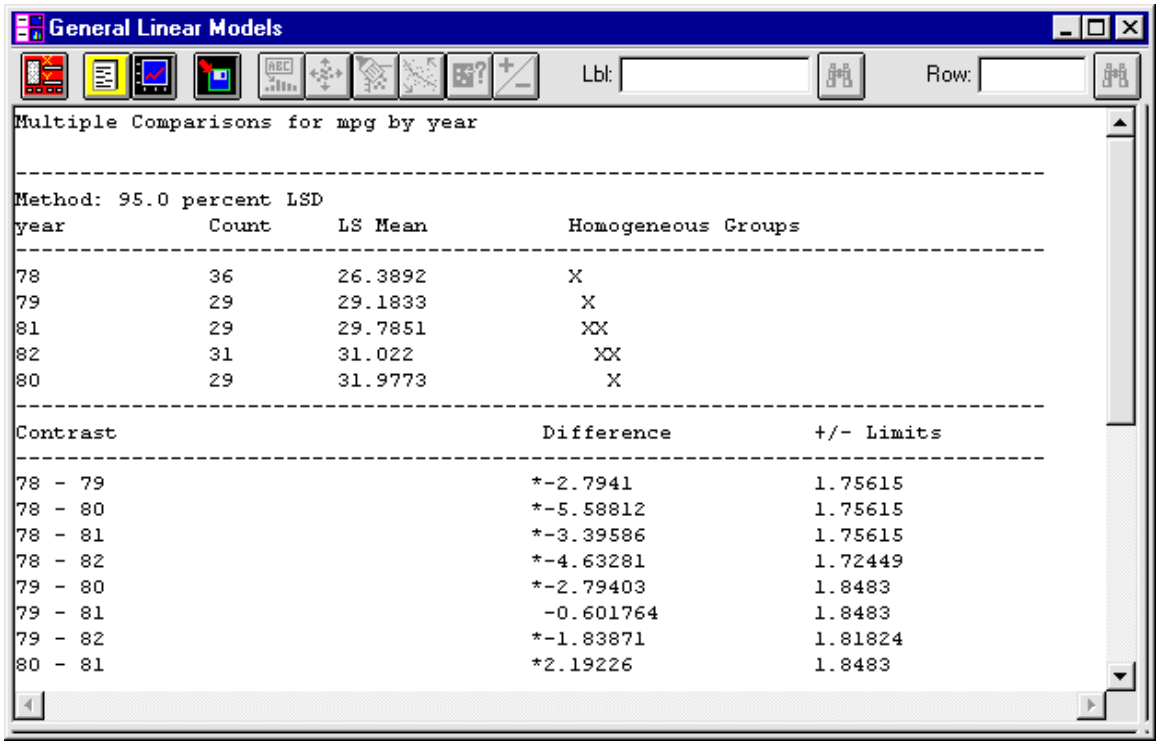


Figure 1-8. Multiple Range Tests Report

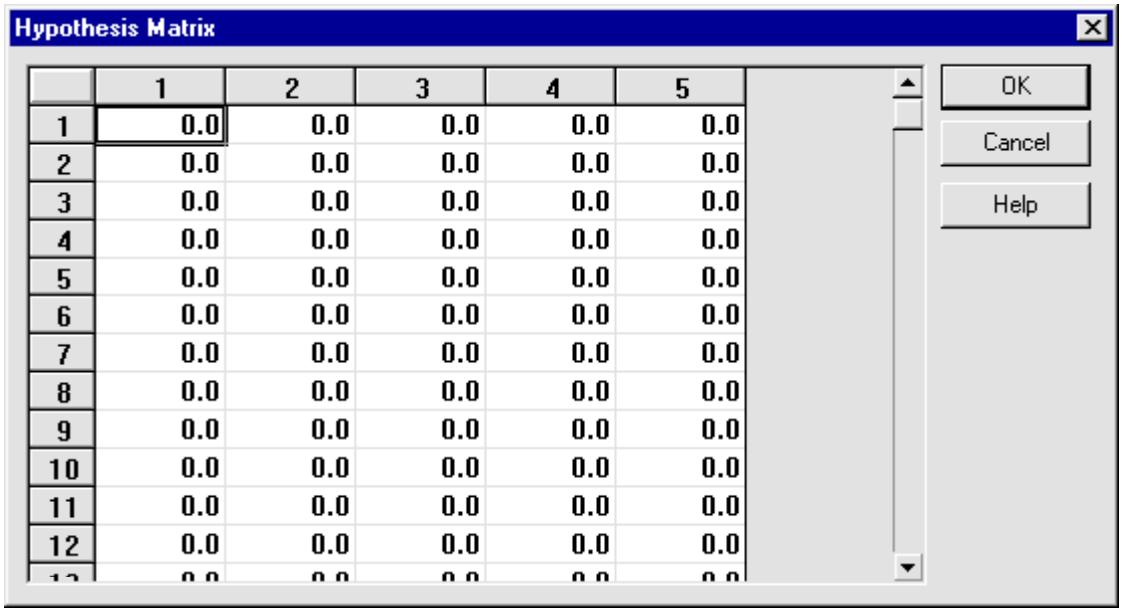
If you choose the User-Specified Option, you can choose contrasts for the groups using a Hypothesis Matrix shown in Figure 1-9. Use the matrix to enter the contrasts that will be used to compare the means among the groups. See the Hypothesis Matrix topic in Online Help for further information.

Reports

The Reports option displays information about the results generated for the regression using the fitted model (see Figure 1-10).

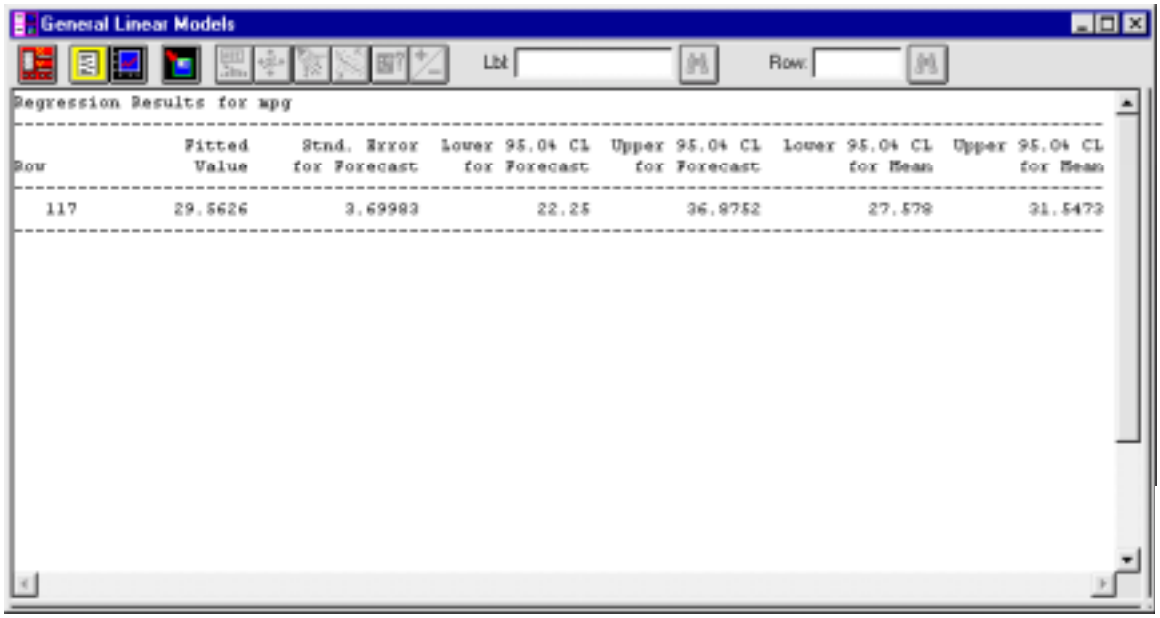
Depending on the options that appear on the Reports Options dialog box, the table includes the predicted value for the dependent variable, the predicted value for the standard error, the 95 percent prediction limits for new observations, and the 95 percent confidence limits for the mean response. Each item corresponds to the values of the independent variables in a specific row of the file. To create forecasts (predictions) for additional combinations of the variables, add additional rows to the bottom of the file. In each new row, enter values for the independent variables but leave the cell for the dependent variable empty. The program adds the predicted values for the new rows to the table, but leaves the model unchanged. Use the Analysis Options to choose the dependent variable that will be used to create the report.

Use the *Reports Options* Dialog Box to choose the values that will be included in the report. The options are: Observed Y, Fitted Y, Residuals, Studentized Residuals, Standard Errors for Forecasts, Confidence Limits for Individual Forecasts, and Confidence Limits for Forecast Means.



	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0

Figure 1-9. Hypothesis Matrix

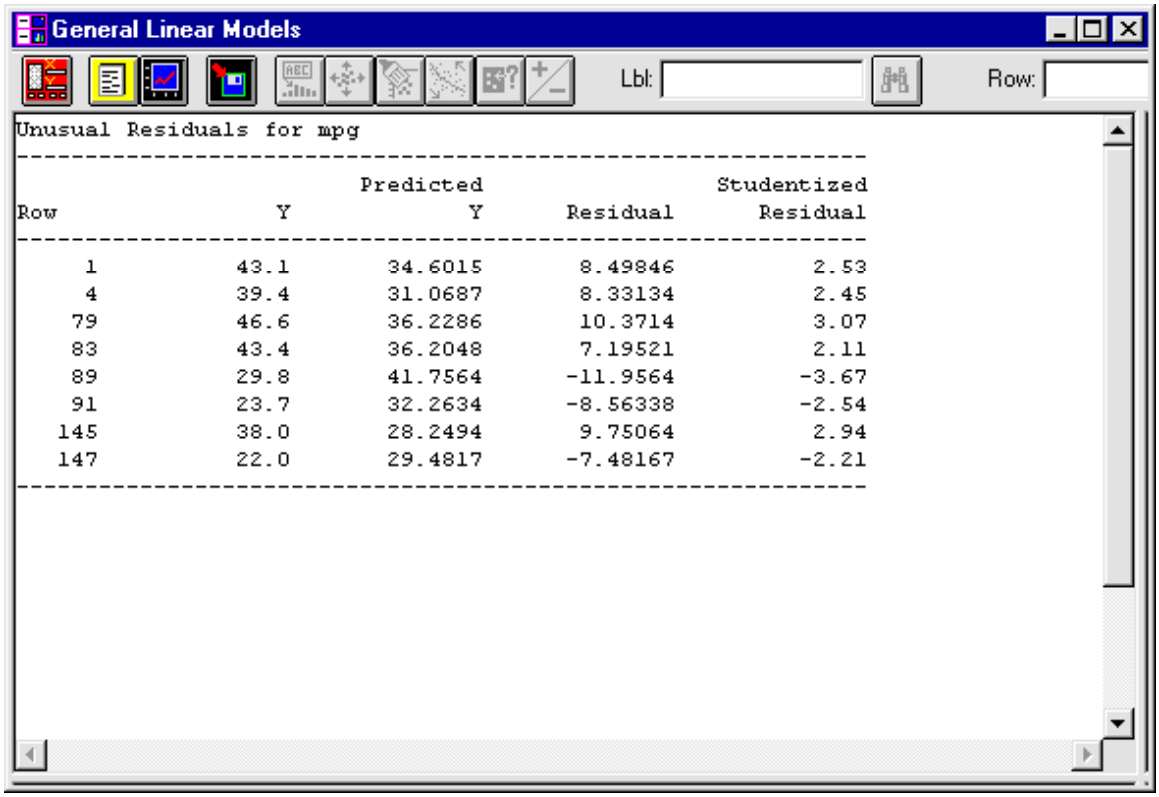


Row	Fitted Value	Stnd. Error for Forecast	Lower 95.04 CL for Forecast	Upper 95.04 CL for Forecast	Lower 95.04 CL for Mean	Upper 95.04 CL for Mean
117	29.5626	3.69983	22.25	36.8752	27.579	31.5473

Figure 1-10. Reports Table

Unusual Residuals

The Unusual Residuals option displays a table that lists all the observations that have studentized residuals with values greater than 2.0 in absolute value (see Figure 1-11). Studentized residuals measure the number of standard deviations that each observed value of the dependent variable deviates from the model that was fitted using all the data except that observation.



Row	Y	Predicted Y	Residual	Studentized Residual
1	43.1	34.6015	8.49846	2.53
4	39.4	31.0687	8.33134	2.45
79	46.6	36.2286	10.3714	3.07
83	43.4	36.2048	7.19521	2.11
89	29.8	41.7564	-11.9564	-3.67
91	23.7	32.2634	-8.56338	-2.54
145	38.0	28.2494	9.75064	2.94
147	22.0	29.4817	-7.48167	-2.21

Figure 1-11. Unusual Residuals Table

Influential Points

The Influential Points option displays a table that lists the observations that have leverage values greater than three times that of an average point, or that have unusually large DFITS or Cook's distance values (see Figure 1-12). Leverage is a statistic that measures the amount each estimated coefficient would change if each observation was removed from the data. The Cook's distance statistic measures the distance between the estimated coefficients with and without each observation.

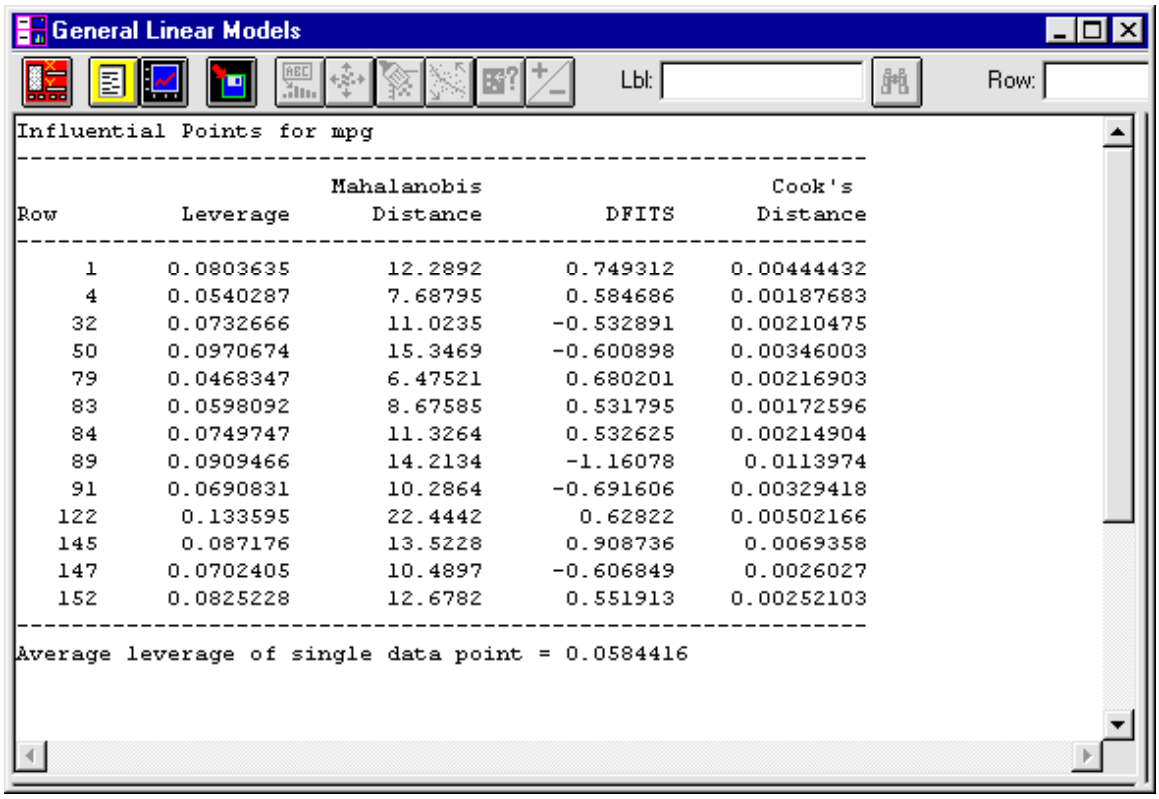


Figure 1-12. Influential Points Table

Graphical Options

Scatterplot

The Scatterplot option displays a scatterplot of one dependent variable versus one independent variable (see Figure 1-13).

Use the *General Linear Models Options* dialog box to choose the dependent variable; use the *Scatterplot Options* dialog box to choose the independent variable.

Means Plot

The Means Plot option displays a plot of the mean for a dependent variable for each level of an independent categorical variable, if there is one (see Figure 1-14). The plot displays the confidence intervals for each of the means separately. Or, if you choose a quantitative variable, the program displays a Plot of Fitted Model for the regression versus the quantitative variable you choose. The Hold... command retains any other quantitative variables at the value you enter.

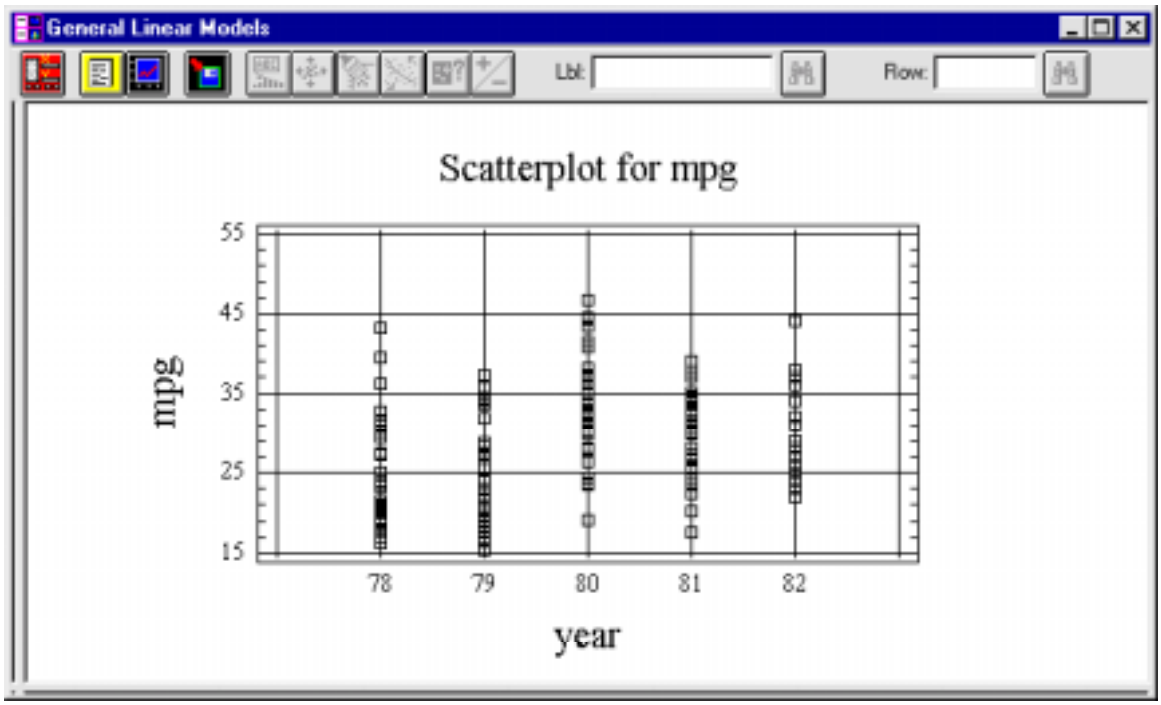


Figure 1-13. Scatterplot

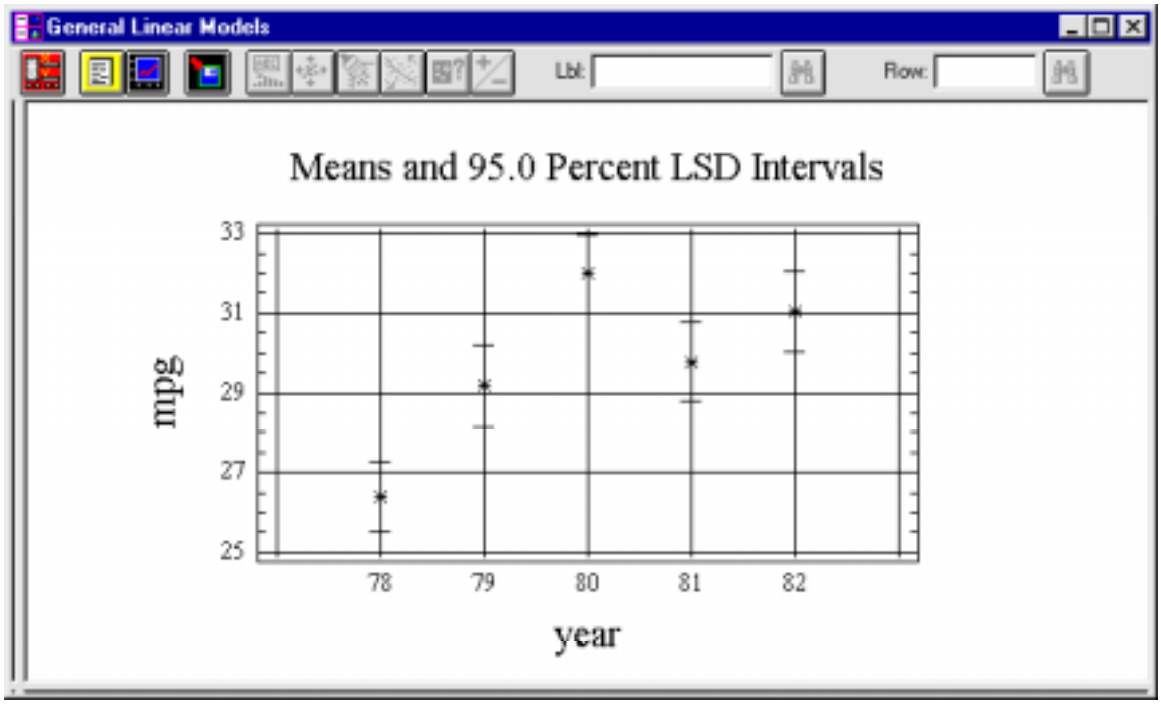


Figure 1-14. Means Plot

Use the *Means Plot Options* dialog box to choose the type of interval that will be used and the factor that will be displayed on the plot. You can also enter a different value for the confidence level.

Interaction Plot

The Interaction Plot option shows the two-factor interactions, if there are any, that were estimated in the analysis using the current model (see Figure 1-15)

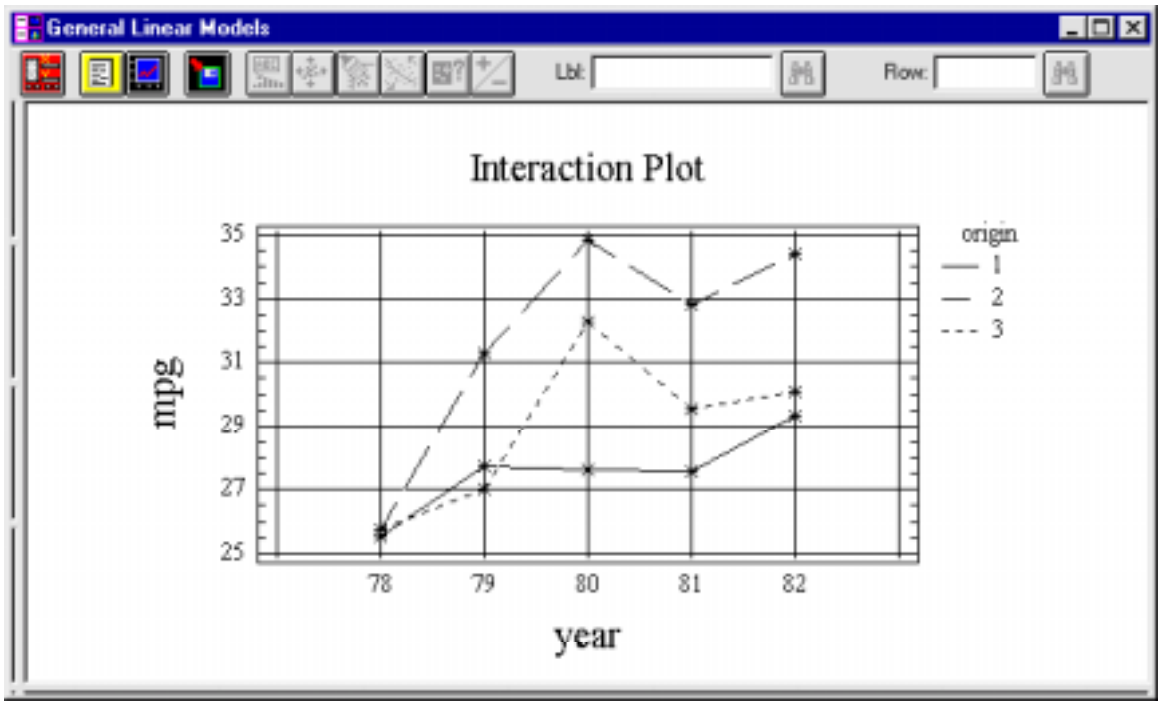


Figure 1-15. Interaction Plot

Use the *Interaction Plot Options* dialog box to choose the type of interval that will display, to choose an interaction for the plot, to enter a value for the confidence level, and to choose the factor that will be plotted on the axis.

Surface Plot

The Surface Plot option creates a Surface Plot, which is a three-dimensional plot of the relationship between the estimated dependent variable and the other two chosen variables (see Figure 1-16).

Use the *Response Plot Options* Dialog Box to choose the type of plot you want to create (Surface, Contour, or Square); choose settings for a Surface Plot; enter a value for the resolution; and access the Plot of Fitted Model Options dialog box using the Factors... command.

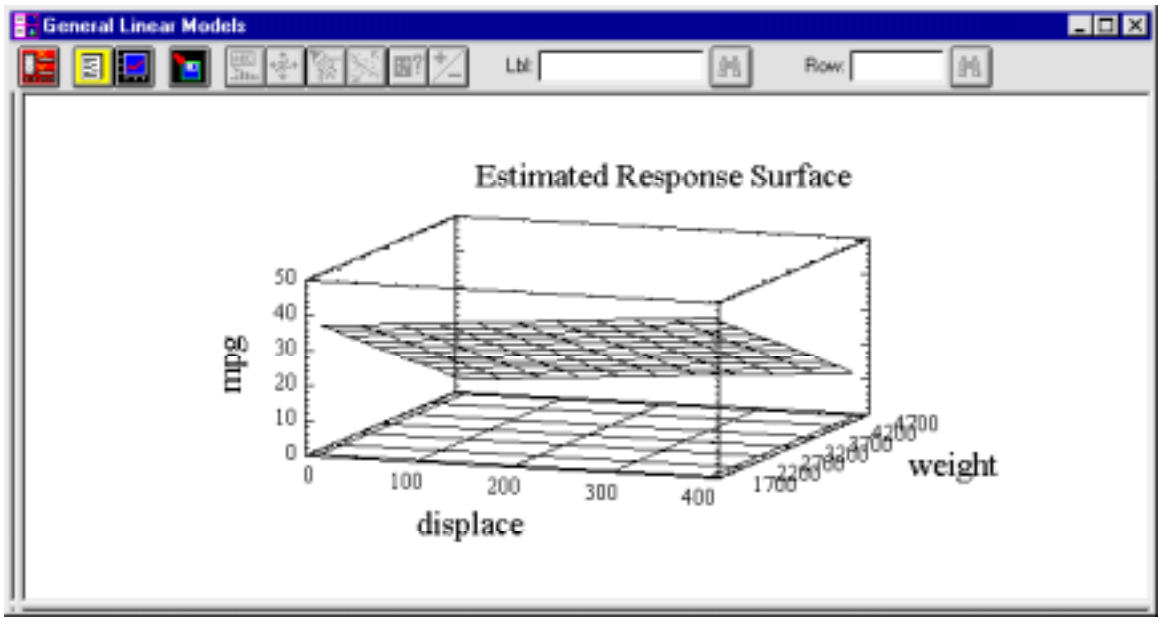


Figure 1-16. Surface Plot

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model; to enter values for upper and lower limits of the axis for the chosen variable; and enter a value that will determine the level at which all of the other variables will be held.

Contour Plot

The Contour Plot option creates a Contour Plot, which is a two-dimensional plot that traces the contours of the estimated dependent variable as a function of the other variables (see Figure 1-17).

Each contour line represents combinations of the independent variables, which have a value you chose for the estimated dependent variable. You can predict the next value for the dependent variable by following the ridge of the contour.

Use the *Response Plot Options* dialog box to choose the type of plot you want to create (Surface, Contour, or Square); choose settings for a Contour Plot; enter a value for the resolution; and access the Plot of Fitted Model Options dialog box using the Factors... command.

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model; to enter values for upper and lower limits of the axis for the chosen variable; and enter a value that will determine the level at which all of the other variables will be held.

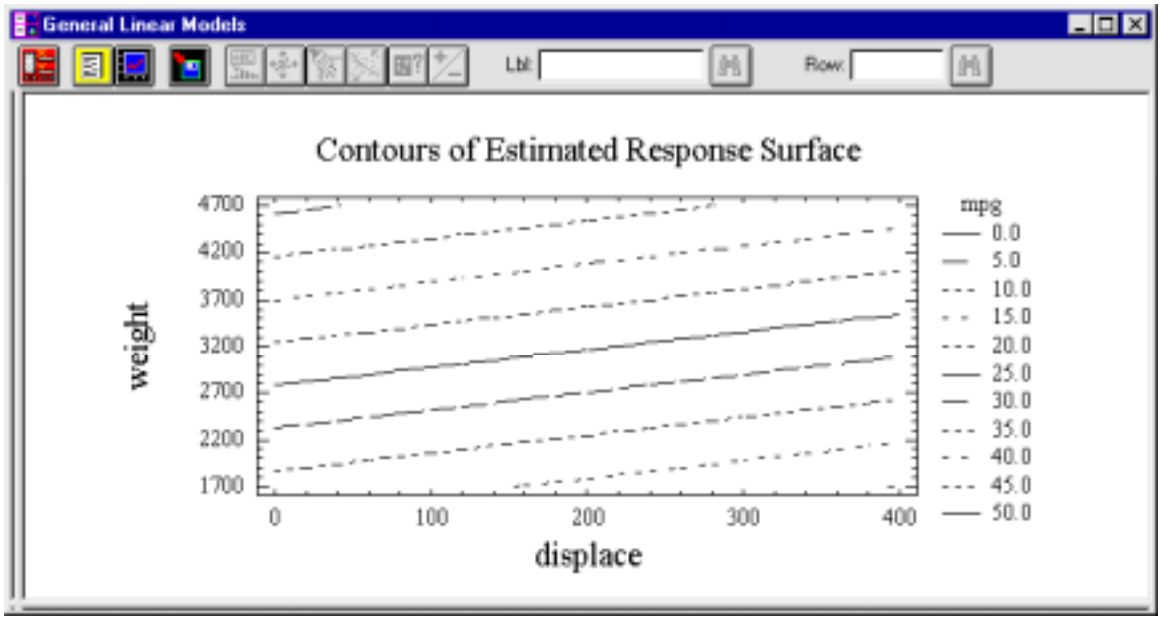


Figure 1-17. Contour Plot

Observed versus Predicted

The Observed versus Predicted option displays a plot of the observed values versus the values predicted by the fitted model (see Figure 1-18). The plot includes a line with slope equal to one. Points close to the diagonal line are those best predicted by the model.

Use the plot to detect situations in which the error variance is not constant, which indicates that you should probably transform the values for the dependent variable.

Residual Plots

The Residual Plots option displays one of three different plots: a Scatterplot of the Residual versus the values predicted by the fitted model, the row number, or by X; a Normal Probability Plot; or an Autocorrelation Function.

Use the *Residual Plots Options* dialog box to choose one of the plots and, if applicable, its options.

Residual versus Predicted

The Residual versus Predicted scatterplot displays a plot of the residual or the studentized residual versus the predicted for the observed variable (see Figure 1-19). A nonrandom pattern indicates that the model does not adequately describe the observed data. The plot is helpful in showing heteroscedasticity, an indication that the variability changes as the values of the dependent variable change.

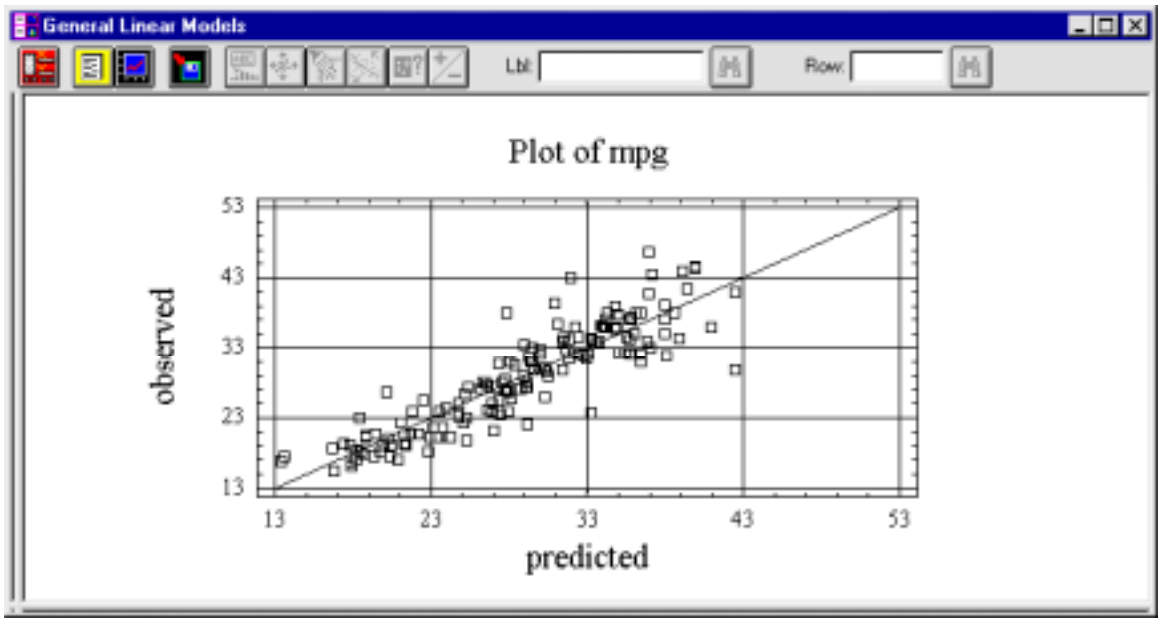


Figure 1-18. Observed versus Predicted Plot

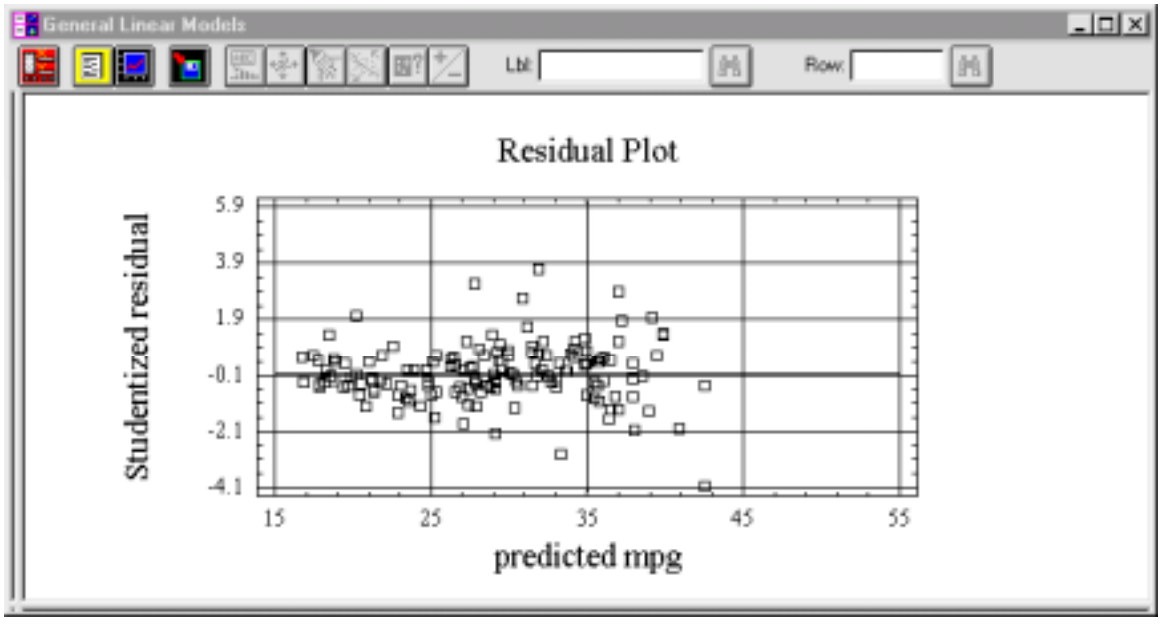


Figure 1-19. Residual versus Predicted Scatterplot

Residual versus Row Number

The Residual versus Row Number scatterplot displays a plot of the residual or the studentized residual versus the row number (see Figure 1-20). Any nonrandom pattern indicates serial correlation in the data, particularly if the row order corresponds to the order in which the data were collected.

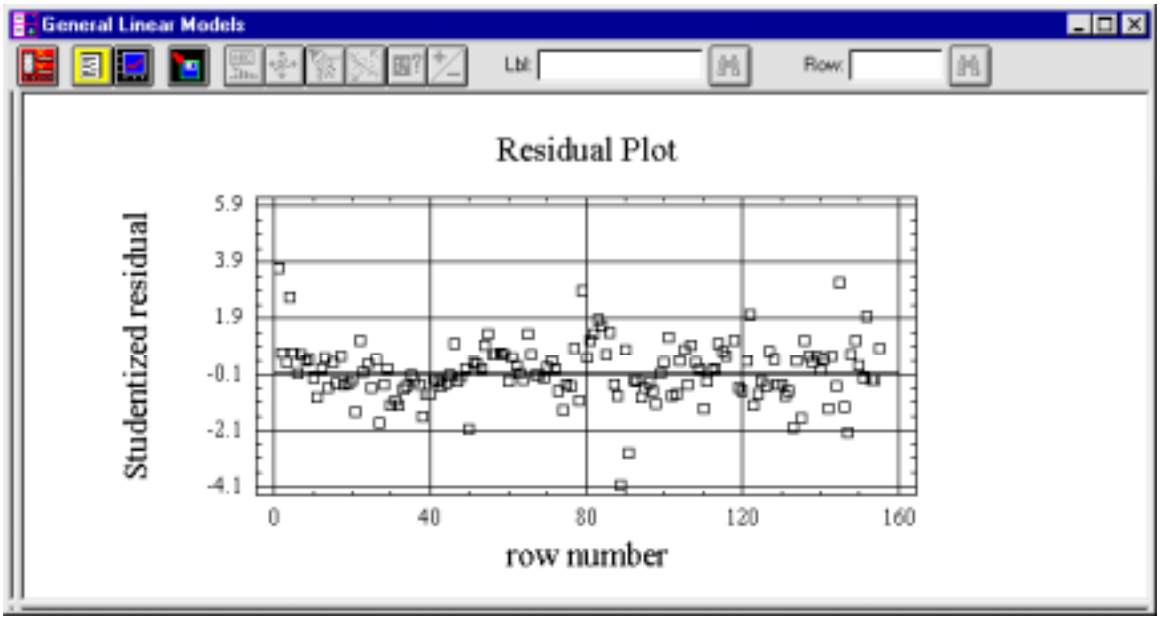


Figure 1-20. Residual versus Row Number Scatterplot

Residual versus X

The Residual versus X scatterplot displays the residual or studentized residual versus the independent variable (X). Use this plot to detect the nonlinear relationship between the dependent and independent variables (see Figure 1-21). You can also use the plot to determine if the variance of the residual is constant. Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may an outlier.

You can also use the plot to determine if the variance of the residual is constant. Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may be an outlier.

Normal Probability Plot

The Normal Probability Plot option displays the residual used to determine if the errors follow a normal distribution (see Figure 1-22). The plot consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as

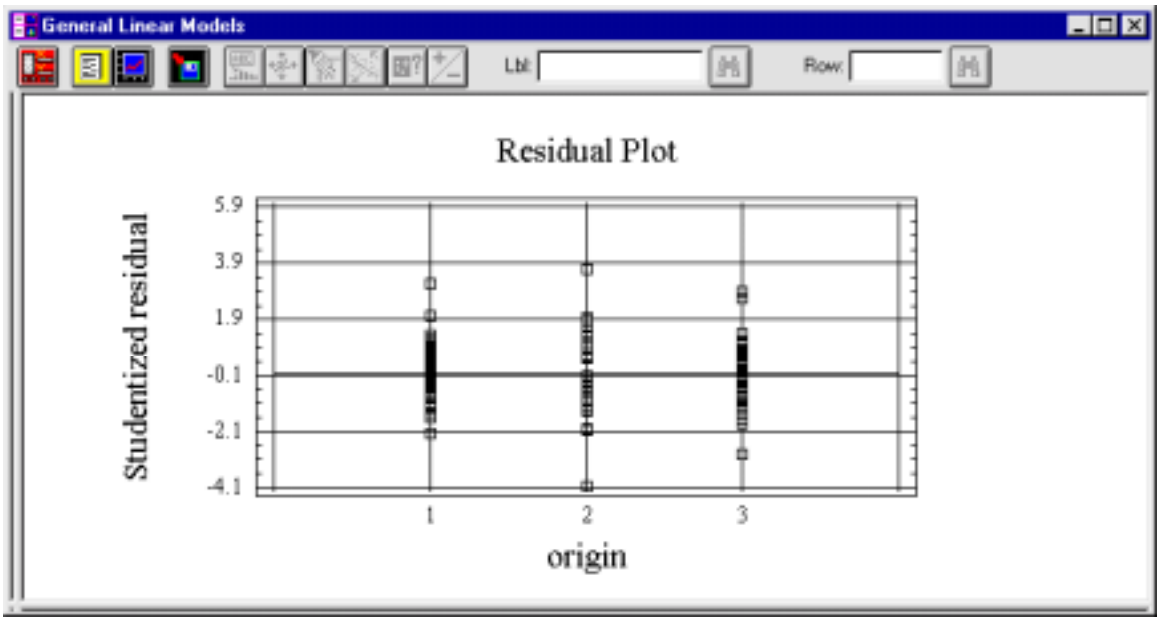


Figure 1-21. Residual versus X Scatterplot

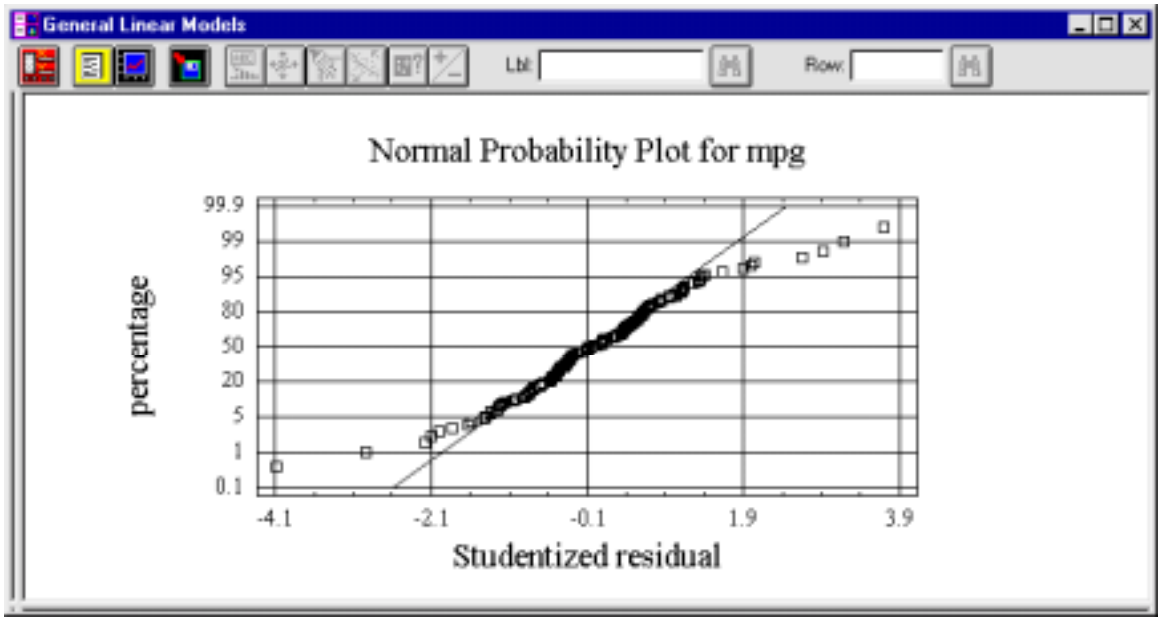


Figure 1-22. Normal Probability Plot

a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution.

Autocorrelation Function Plot

The Autocorrelation Function plot creates a graph of the estimated autocorrelations between the residuals at various lags (see Figure 1-23). The lag *k* autocorrelation of coefficient measures the correlation between the residuals at time *t* and time *t-k*. If the probabiity limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlation at that lag. The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag.

Significant autocorrelations extend above or below the probability limits. When you choose this option, the Number of Lags and Confidence Level text boxes become active.

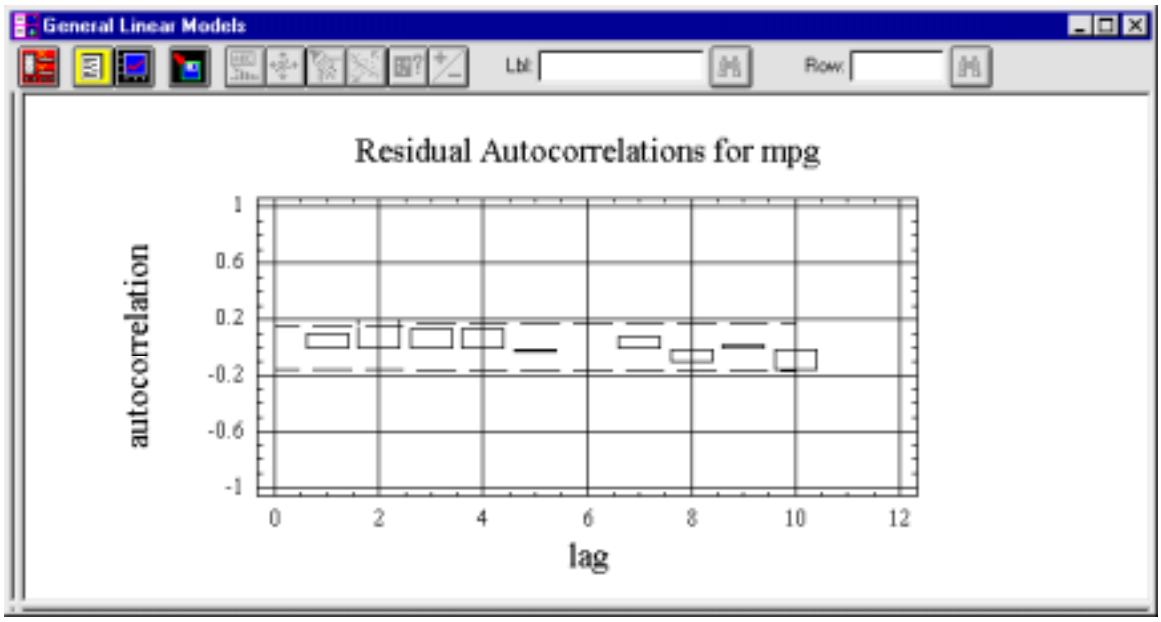


Figure 1-23. Autocorrelation Function Plot

Saving the Results

Use the Save Results Options dialog box to choose the results you want to save. There are 13 options: Predicted Values, Standard Errors of Predictions, Lower Limits for Predictions, Upper Limits for Predictions, Standard Errors of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, DFITS Statistics, Mahalanobis Distances, and Coefficients.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results dialog box, click the Save Results Option button on the Analysis toolbar (the fourth button from the left).

References

Graybill, F. A. 1976. *Theory and Application of the Linear Model*. Belmont, California: Wadsworth.

Hartley, H.O. 1967. Expectations, variances and covariances of ANOVA mean squares by 'synthesis'. *Biometrics* 23: 105-14; Corrigenda, 853.

McCullash, P. and Nelder, J. A. 1989. *Generalized Linear Models*, second edition. London: Chapman & Hall.

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data*. Volume I, *Designed Experiments*. New York: Van Nostrand Reinhold.

Morrison, D. F. 1983. *Applied Linear Statistical Methods*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Nelder, J. A. and Wedderburn, R. W. M. 1972. "Generalized Linear Models," *Journal of the Royal Statistical Society*, **A135:**370-84.

Neter, J., Kutner, M. H., Nachsheim, C. J. and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago: Richard D. Irwin, Inc.

Scheffe, H. 1959. *The Analysis of Variance*. New York: John Wiley & Sons.

Chapter 2

Using Calibration Models

Background Information

With quality improvement increasingly becoming a major business strategy, ensuring that parts and processes conform to product specifications, one important concept of quality improvement, means that companies must continuously analyze data and feed it back into the manufacturing process to prevent production problems. Quantifying this type of data involves defining standard units, calibrating instruments to meet the standard units, and using the instruments to quantify the parts and processes (DataMyte Handbook, 1987).

Today companies involved in international commerce use several systems of international units of measure: English, metric, and the Systeme International d'Unities (SI); the preferred system is the SI. All industrialized countries maintain primary reference standards through an institute or bureau whose purpose is to construct and maintain the standards, which are then used as a basis for calibrating equipment. Because it is impossible and impractical for these institutes or bureaus to calibrate every piece of equipment, second- and third-order standards were developed for calibrating instruments that are used in general laboratories and manufacturing areas.

Technicians and inspectors calibrate their equipment against a set of working standards, which are directly tied to the primary standards through the use of transfer standards. Transferring from one standard to a higher, more accurate standard, is known as calibration. When a piece of equipment is calibrated and transferred back to a primary standard through transfer standards, the process is known as traceability. Quality standards such as the ISO 9000 and ASQC, often require that equipment be calibrated against a set of working standards that are referred back to primary standards; perhaps even all the way back to the standards maintained by an agency such as the National Institute of Standards and Technology in the United States.

Regression analysis is a mathematical tool that quantifies the relationship between two methods of measurement and provides information that helps researchers determine if a less accurate tool might still be adequate for its intended purpose. At times, a regression model of Y is used to make a prediction for X, given a new observation Y. This is known as an *inverse prediction problem* or *a calibration problem*.

Calibration problems often occur when the following takes place: A researcher finds that there are two ways to measure the same quantity. One is extremely accurate,

very time-consuming, and possibly expensive — the "gold" standard. The other is more alterable, easier to obtain, and usually less expensive. One may be a direct measurement; the other an indirect measurement.

You can use the regression model Y = aX + b to predict values for X rather than for Y; that is, you cannot take the prediction intervals directly from the intervals that you would use for predicting Y values. For example, analytical chemists and engineers use the calibration line to translate measurements from the scale of a measuring device or process to an interval estimate for a true quantity. The calibration line is then determined by running an experiment, taking several samples where the true quantity is known (X), measuring those samples with the instrument or process being calibrated at (Y), and plotting the pairs on the XY plane. The equation of the line that best fits through these points is the calibration line; the mathematical tool that provides the equation of that line is least squares regression.

Calibration Models in STATGRAPHICS *Plus*

The Calibration Models Analysis in STATGRAPHICS *Plus* allows you to find a calibration line that fits the data, letting you choose from several transformations when the points do not follow a straight line. You can choose to eliminate the constant from the model, which forces the line through the point (0,0). You also can associate a weight with each observation to perform a weighted regression. If the calibration experiment includes replicate observations for any of the X values, you can use a lack-of-fit test in the ANOVA table. The tables and graphs in the analysis help you to determine the validity of the calibration line by completely testing the hypotheses, model diagnostics, and residuals.

Once you are completely satisfied with the calibration line, the program can add values to the equation for either Y or X to calculate estimates with intervals for the opposite variable. Corresponding values for the measurement and the converted quantity with intervals are shown in tables or directly on an XY Plot. In addition, you can save a calibration line for later use when you convert additional measurements into interval estimates. Finally, if you have two sets of experimental data, you can compare the two calibration lines with each other using a hypothesis test to see if the results are consistent.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... CALIBRATION MODELS... from the Menu bar to display the Calibration Models Analysis dialog box (see Figure 2-1).

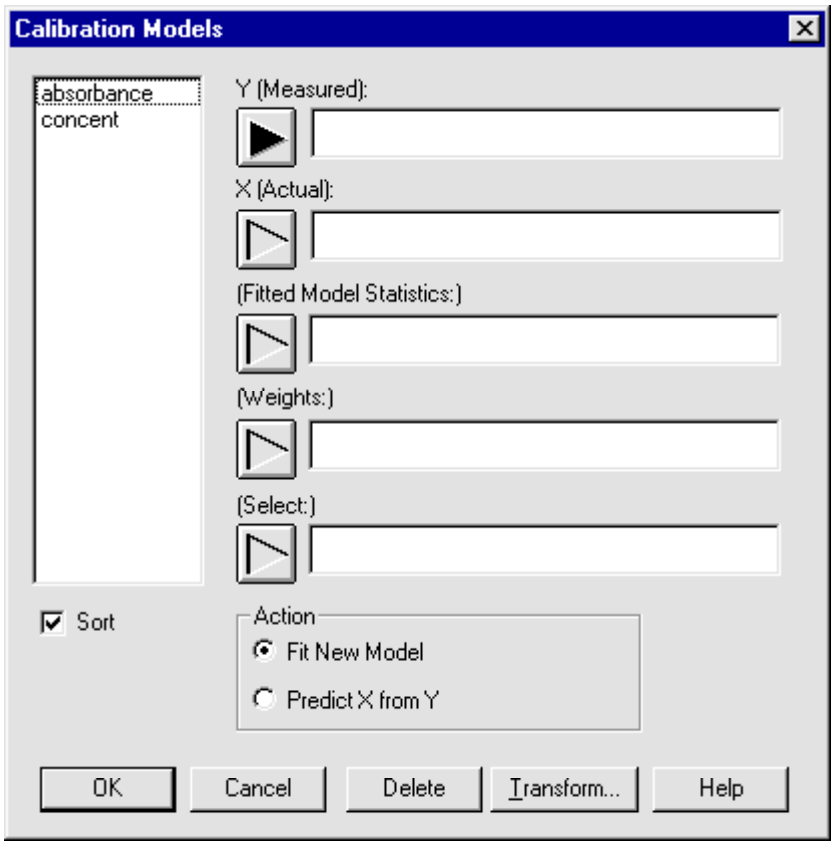


Figure 2-1. Calibration Models Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option shows the results of fitting a linear model to describe the relationship between the dependent (measured) and independent (actual) variables (see Figure 2-2). The StatAdvisor includes the equation for the fitted model. If the data contain replicate values for X, the program also performs a lack-of-fit test.

- If the *p*-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the two variables at the 99 percent confidence level.

- The R-Squared statistic explains the percentage of variability represented by the dependent variable as it was fitted in the full model.

- The correlation coefficients value explains the strength of the linear relationship between the variables.

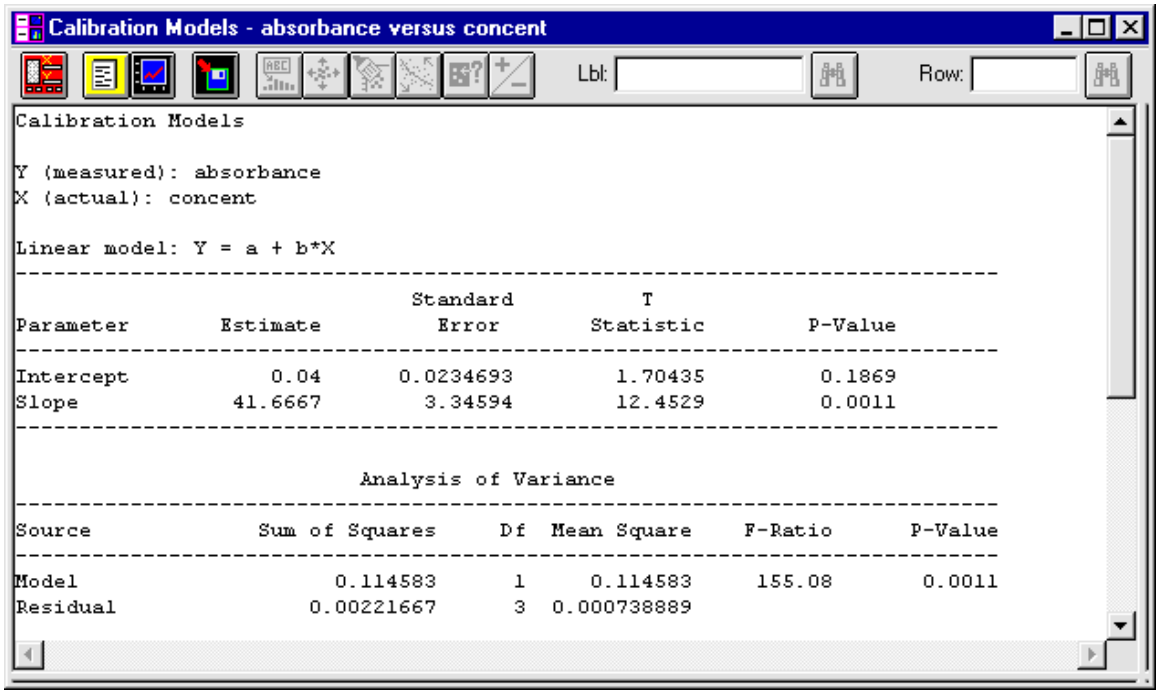


Figure 2-2. Analysis Summary

- The Standard Error of the Estimate statistic explains the value for the standard deviation of the residuals. You can use this value to construct prediction limits for new observations.

- If you used the optional Select text box on the Calibration Models Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the models and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use the methods discussed in the topic, "Understanding Advanced Regression."

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.

- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the set of data, then finding the average or mean value for the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small

value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.

- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Confidence Intervals

The Confidence Intervals option calculates and displays the confidence intervals or bounds for the parameters of the fitted model (see Figure 2-3). In repeated sampling, 95 percent of these intervals will contain the true values for the parameters, provided the model has the proper form and the errors are distributed normally and independently.

Hypothesis Tests

The Hypothesis Tests option allows you to test hypotheses about the intercept, slope, or difference between the current slope and one from a previous model (if you chose the Fit New Model option and entered the saved model statistics on the Calibration Models Analysis dialog box) (see Figure 2-4).

The program performs the calculations and displays the results of the two hypothesis tests for the parameters of the fitted calibration model. The results of the first test contain information about the intercept. If the *p*-value is greater than 0.10, you cannot reject the hypothesis that the intercept equals 0.0 at a 90 percent or higher confidence level. The results of the second test contain information about the slope. If the *p*-value is less than 0.01, reject the hypothesis that the slope equals 1.0 at the 99 percent confidence level.

Use the *Hypothesis Tests Options* dialog box to enter values used to test the intercept and slope, and to enter a percentage for the alpha level. You can also choose an alternative hypothesis.

Predictions

The Predictions option allows you to predict the values for Y, given X; or the values 0 for X, given Y (see Figure 2-5) . The results display the values for the predictions using the fitted model. The table shows the best predictions and (1) the 95 percent

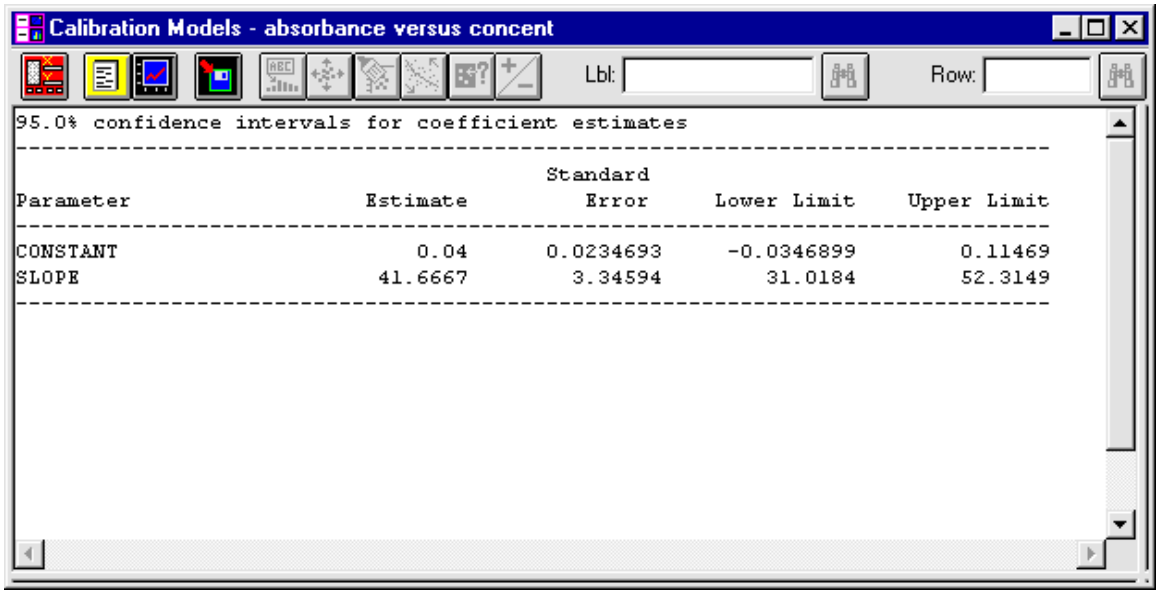


Figure 2-3. Confidence Intervals

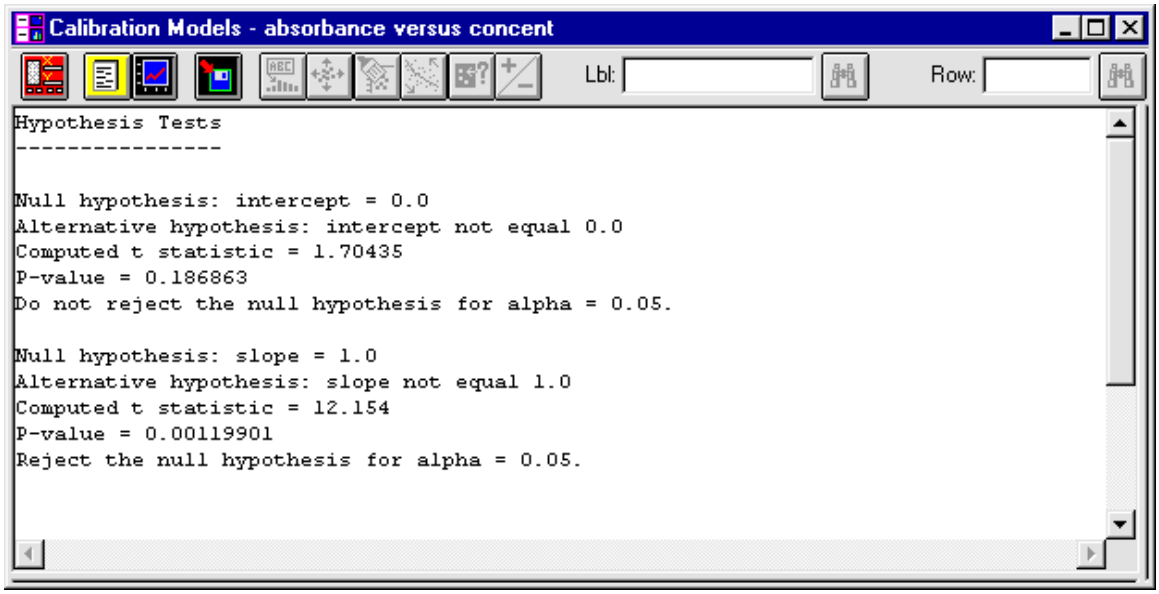


Figure 2-4. Hypothesis Tests

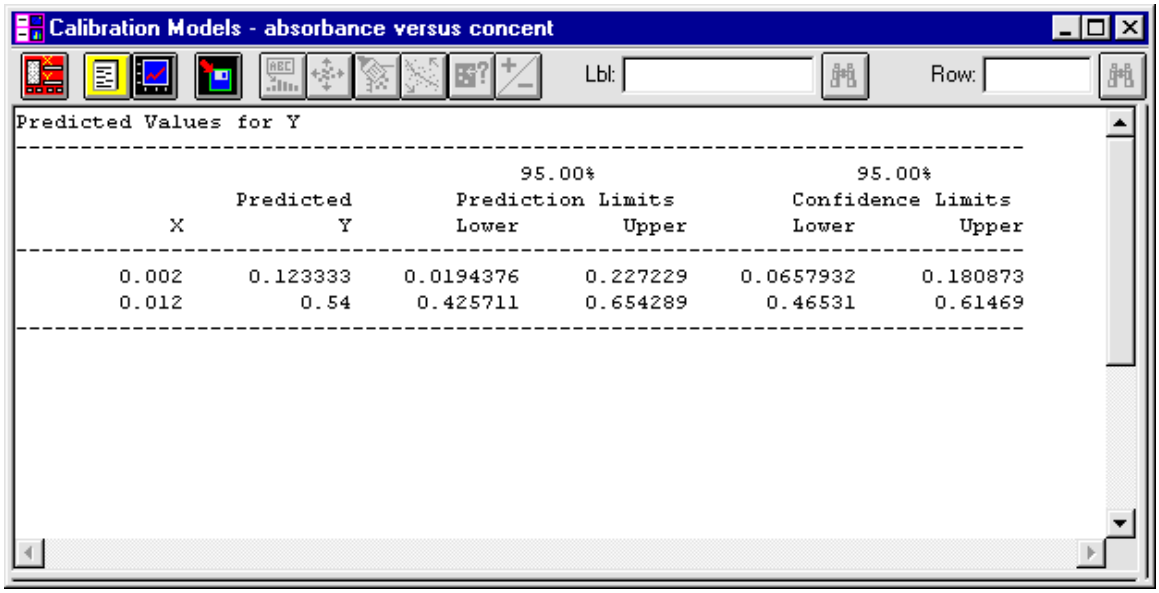


Figure 2-5. Predictions Table

prediction intervals for new observations, and (2) the 95 percent confidence intervals for the mean of many observations. The prediction and confidence intervals correspond to the inner and outer bounds on the Plot of Fitted Model.

Use the *Predictions Options* dialog box to choose the value that will be predicted: Y, given X, or X, given Y; to choose values for the confidence level and mean size or weight; and to choose values for the opposite variable for which the program will display the predictions.

Comparison of Alternative Models (Available for only Fit New Model action)

The Comparison of Alternative Models option fits the curvilinear models to the data and displays the values for the correlations and the R-Squared statistics, arranging them by the highest value (see Figure 2-6). To choose a different model, use the Calibration Model Options dialog box.

Unusual Residuals (Available for only Fit New Model action)

The Unusual Residuals option lists all of the observations that have unusually large studentized residuals — those greater than 2.0 in absolute value (see Figure 2-7). Studentized residuals measure the number of standard deviations each observed value deviates from the model fitting using all the data except that observation.

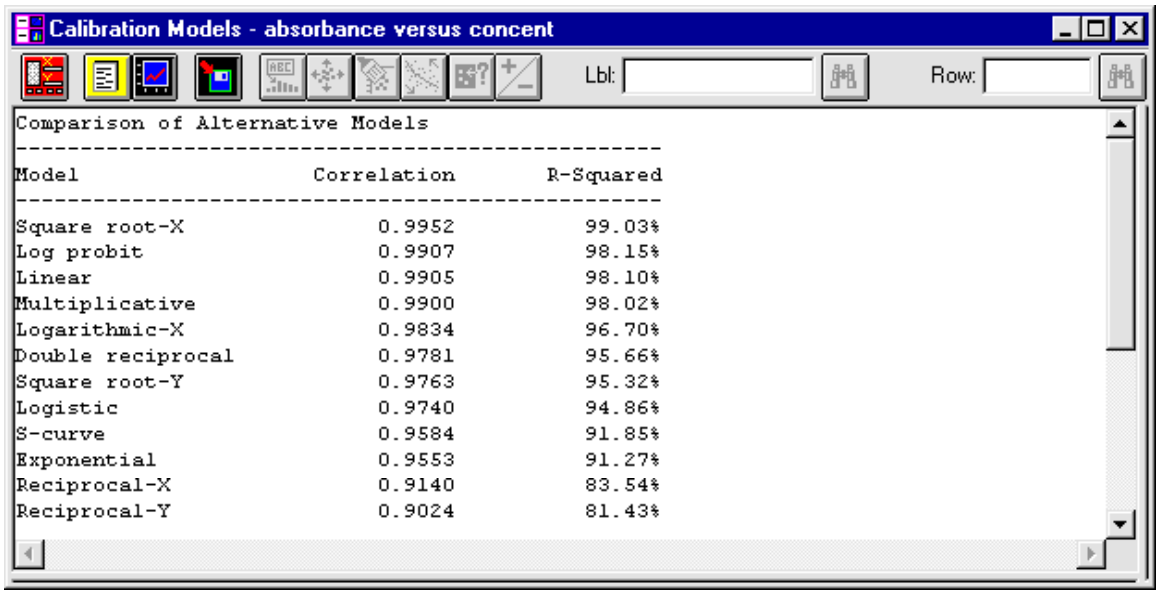


Figure 2-6. Comparison of Alternative Models Report

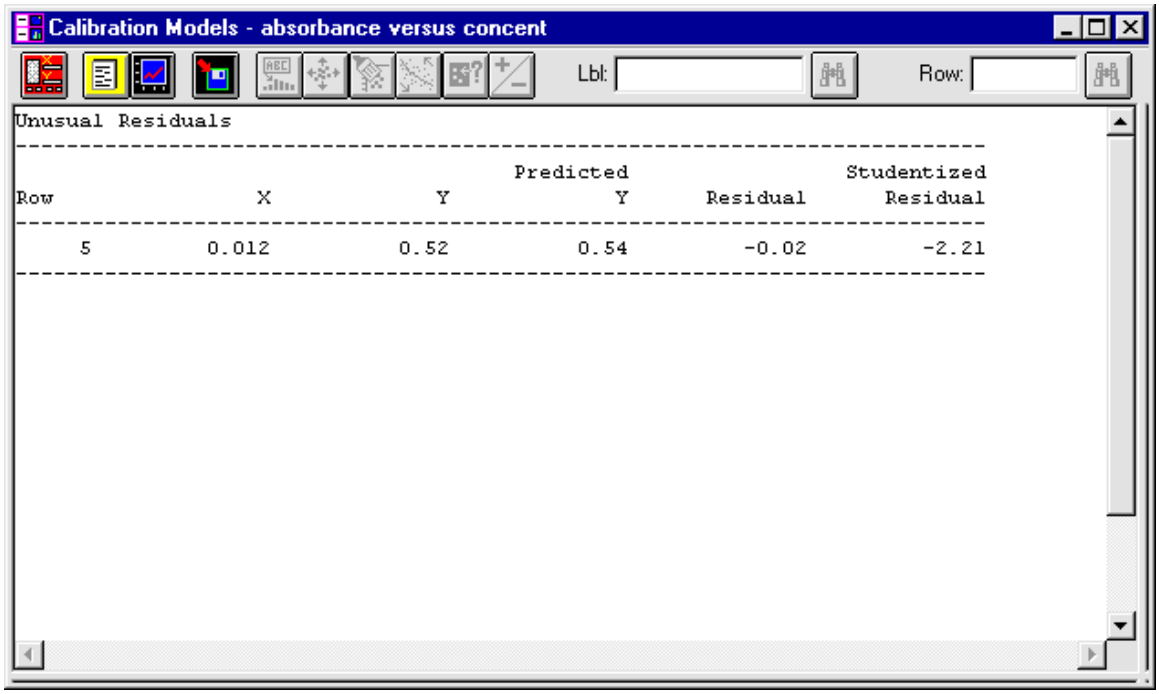


Figure 2-7. Unusual Residuals Table

Influential Points
(Available for only Fit New Model action)

The Influential Points option identifies all the observations that have leverage values greater than three times that of an average point, or that have unusually large DFITS or Cook's distance values (see Figure 2-8). The leverage statistic helps determine the coefficients of the estimated model by measuring the amount of influence that can be attributed to each observation. The DFITS statistic measures the amount of change for each estimated coefficient if the observation is removed from the data. The Cook's distance statistic measures the distance between the estimated coefficients with and without each observation.

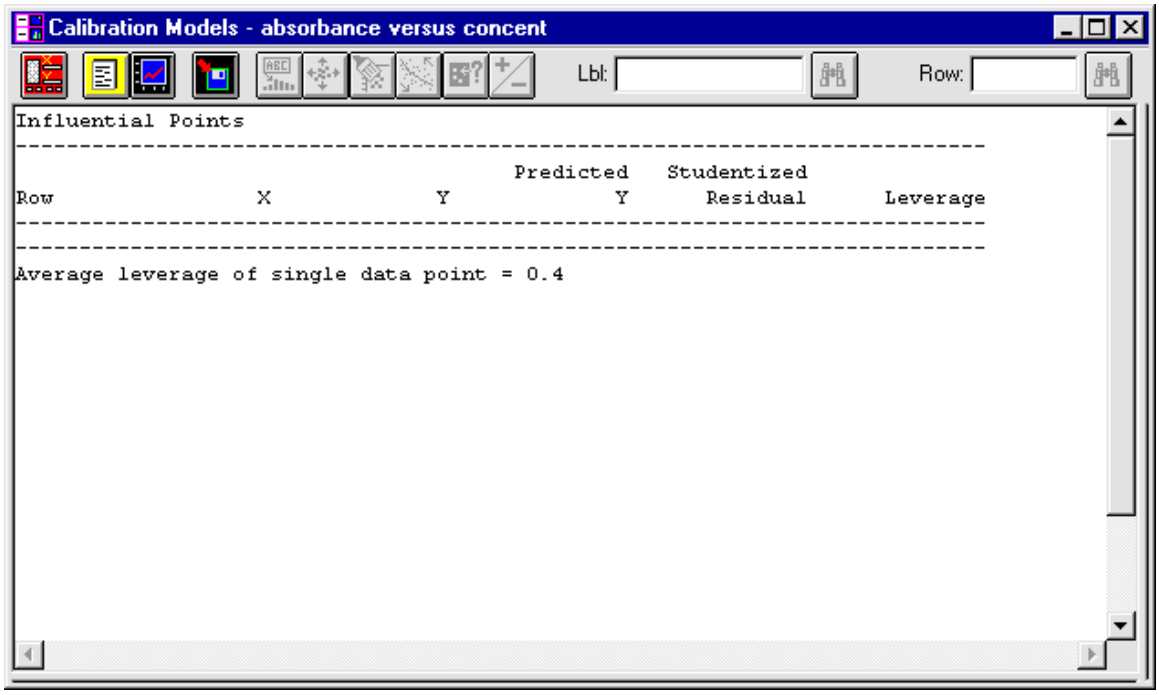


Figure 2-8. Influential Points Table

Graphical Options

Plot of Fitted Model

If you choose the Fit New Model action on the Calibration Models Analysis dialog box, the Plot of Fitted Model graphical option displays a plot of the fitted model with confidence and/or prediction limits (see Figure 2-9). You can also plot a single prediction.

If you choose the Predict X from Y action on the Calibration Models Analysis dialog box, the Plot of Fitted Model graphical option displays a plot of the fitted model with prediction limits for the mean of m observations, where m is the number of Y values you entered on the Calibration Models Analysis dialog box.

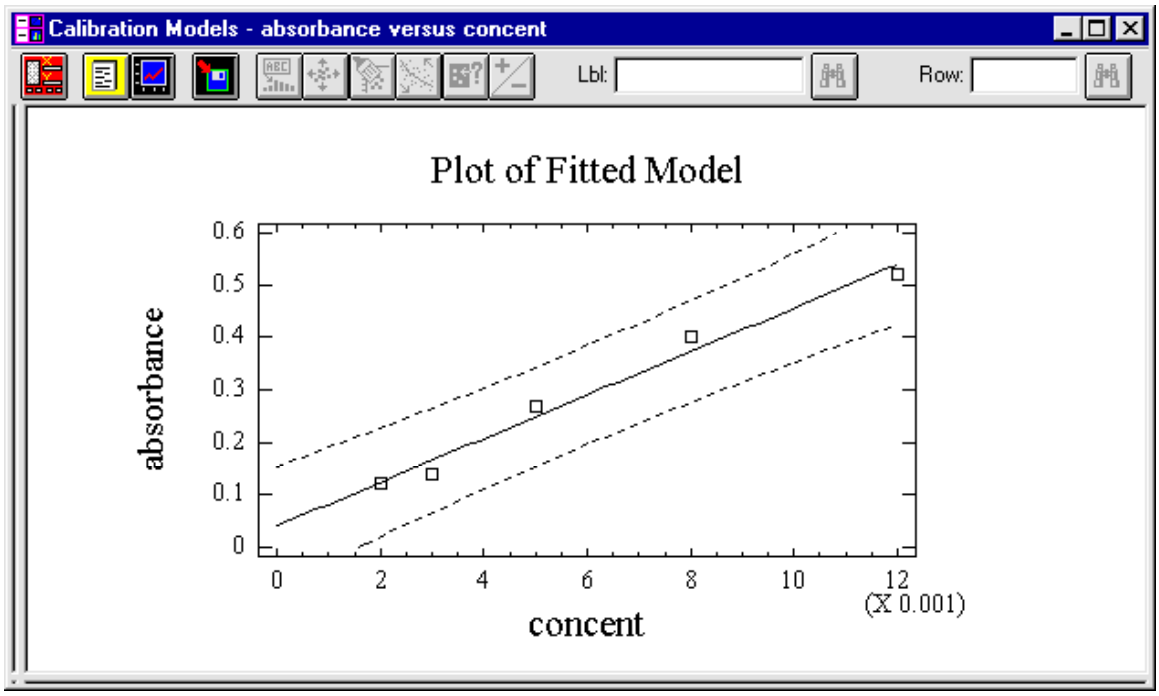


Figure 2-9. Plot of Fitted Model

Observed versus Predicted (Available for Only Fit New Model Action)

The Observed versus Predicted Plot option displays the observed values of Y versus the values predicted by the predicted model (see Figure 2-10). The closer the points lie to the diagonal line, the better the model predicts the observed data. Look for various anomalies, such as increases in variability around the line as the value of Y increases (heteroscedasticity), or points that lie far away from the line (outliers).

Scatterplot

The Scatterplot option displays your choice of up to three scatterplots: Residual versus Predicted, Residual versus Row Number, and Residual versus X; as well as a Normal Probability Plot; and an Autocorrelation Function Plot.

Use the *Residual Plots Options* dialog box to choose one of the plots and, if applicable, its options.

Residuals versus Predicted

The Residuals versus Predicted plot displays the residuals or the studentized residuals versus the predicted values for the observed variable (Y) (see Figure 2-11). A nonrandom pattern indicates that the model you chose does not adequately describe the observed data. The plot is helpful in showing

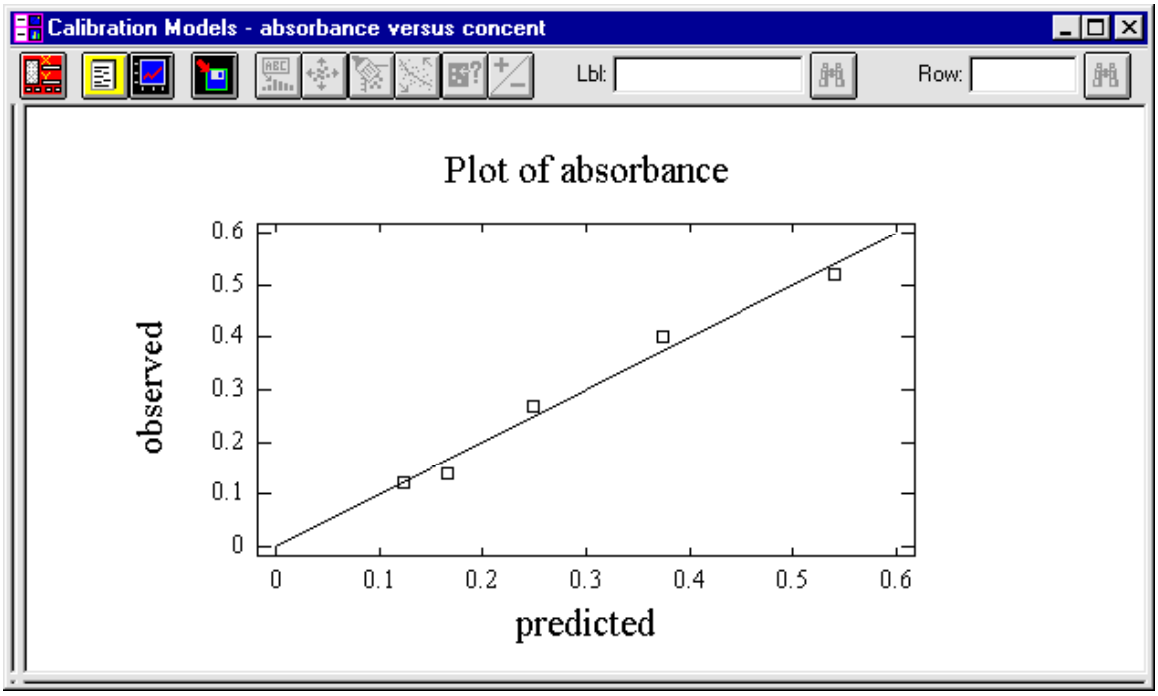


Figure 2-10. Observed versus Predicted Plot

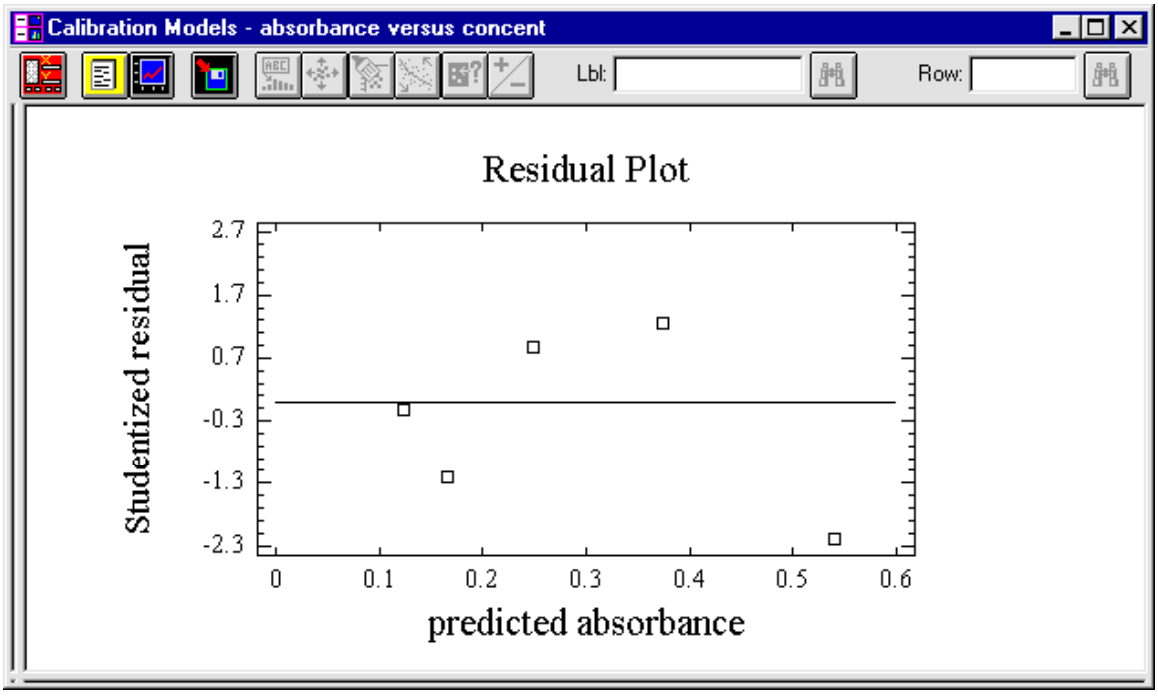


Figure 2-11. Residual versus Predicted

heteroscedasticity; an indication of the variability in the residuals as the values of the dependent variable change.

Residual versus Row Number

The Residual versus Row Number plot displays the residual or the studentized residual versus the row number (see Figure 2-12). The program plots the residuals in the order that the observations appear in the dependent variable. The plot is helpful in determining sequential correlations among the residuals. Any nonrandom pattern indicates serial correlation in the data, particularly if the row order corresponds to the order in which the data were collected.

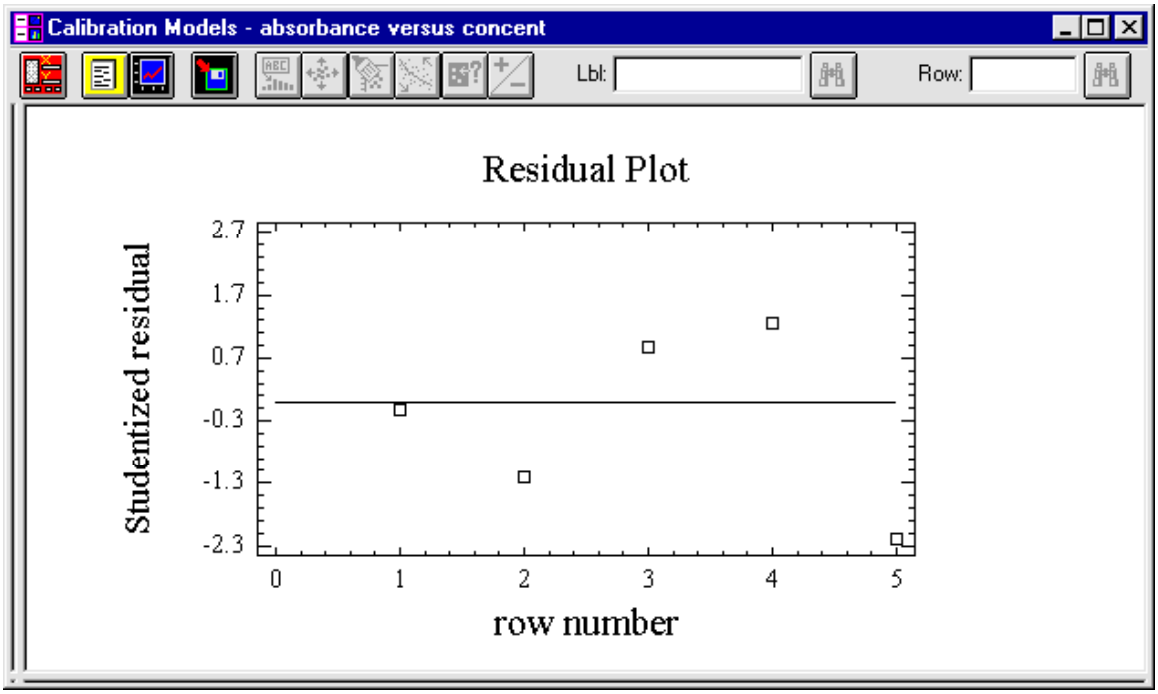


Figure 2-12. Residual versus Row Number

Residuals versus X
(Where X indicates the name of each independent X variable)

The Residual versus X Plot displays the residual or studentized residual versus the independent variable (X), where X equals the name of the independent variable you chose (see Figure 2-13). Use this plot to detect the nonlinear relationship between the dependent and independent variables. You can also use the plot to determine if the variance of the residuals is constant. Nonrandom patterns indicate that the model you chose does not adequately describe the data. Any studentized residual values outside the range of -3 to +3 might be outliers.

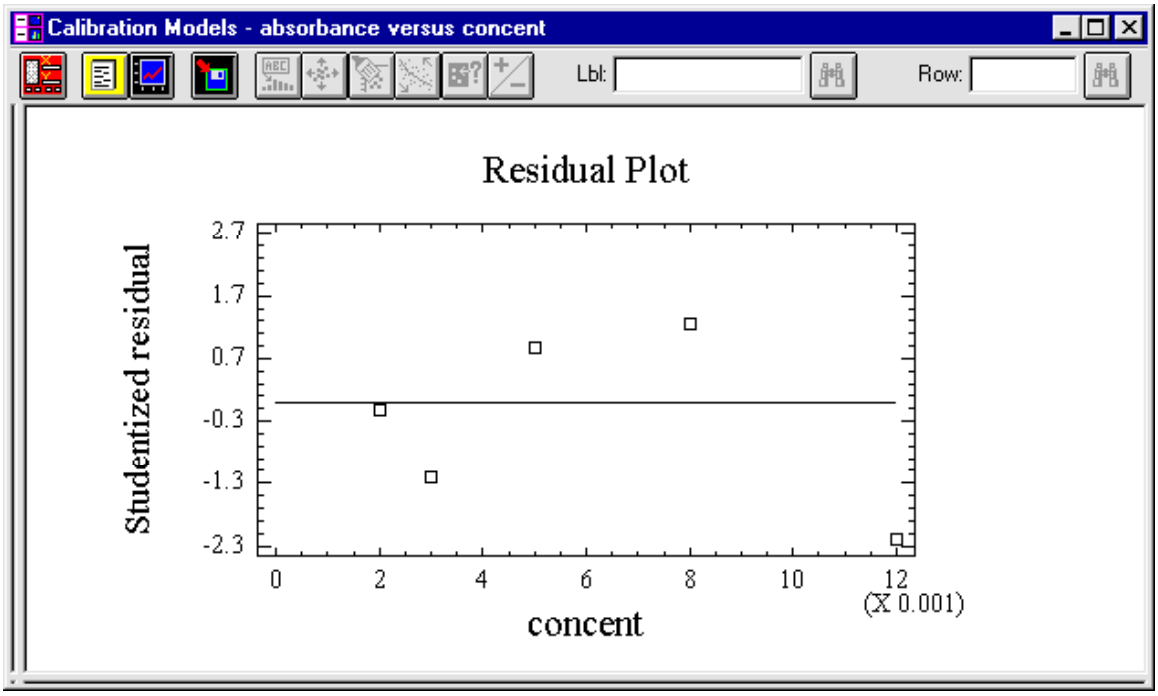


Figure 2-13. Residual versus X Plot

Normal Probability Plot

The Normal Probability Plot option displays the residual used to determine if the data follow a normal distribution (see Figure 2-14). A Normal Probability Plot consists of an arithmetic (interval) horizontal axis scaled for the data, and a vertical axis scaled so the cumulative distribution function plots as a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution. When you choose this option, the Direction and Fitted Line options become active, allowing you to choose the direction of the plot and the type of values that will be used for the fitted line.

Autocorrelation Function Plot

The Autocorrelation Function option displays a plot of the autocorrelation estimates for the residuals (see Figure 2-15). The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits. When you choose this option, the Number of Lags and Confidence Level text boxes are activated.

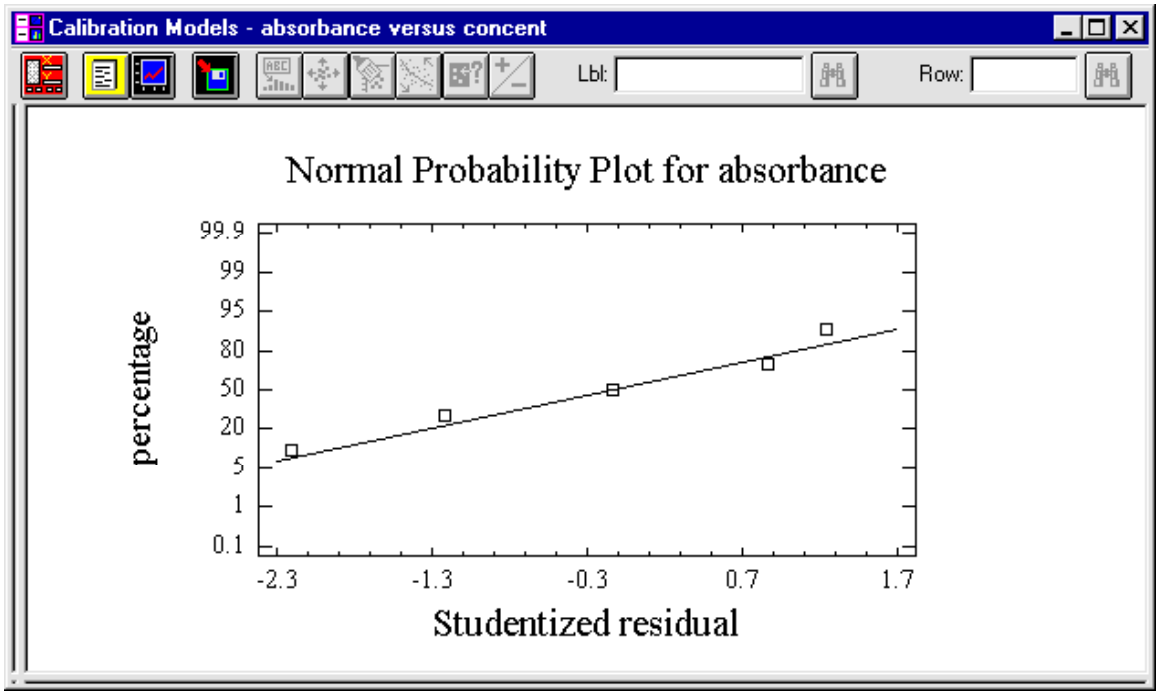


Figure 2-14. Normal Probability Plot

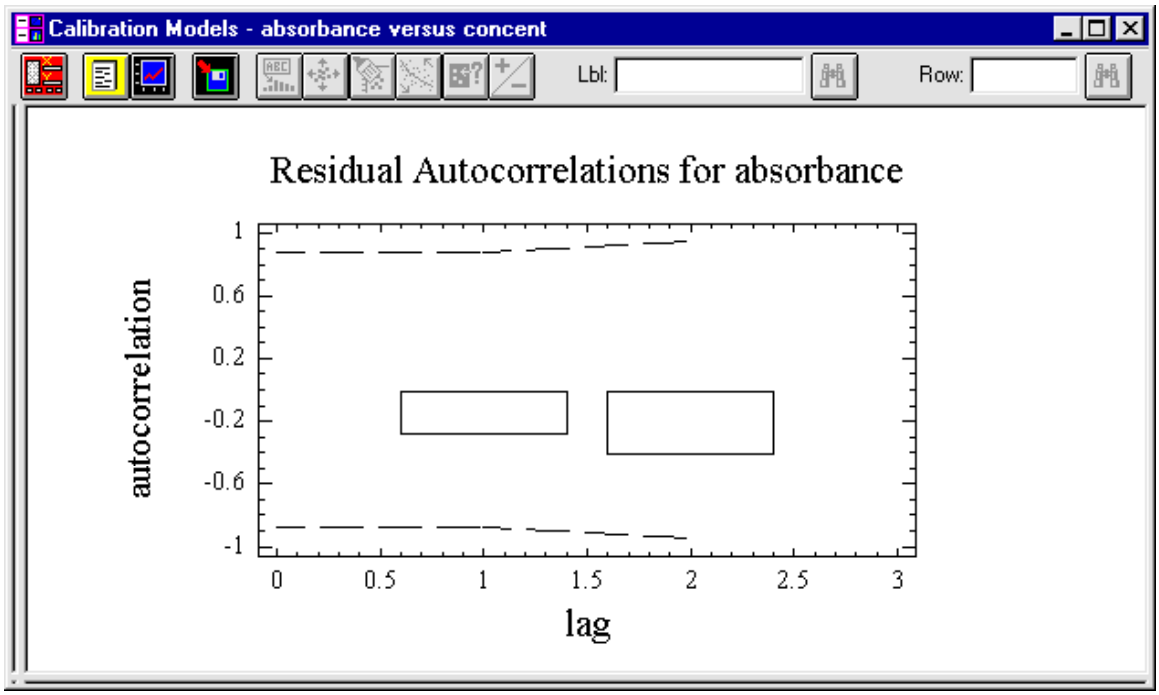


Figure 2-15. Autcorrelation Function Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 10 selections: Model Statistics, Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, and Coefficients.

You can also use the Target Variables text boxes on the dialog box to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Caulcutt, R. and Boddy, R. 1995. *Statistics for Analytical Chemists*. London: Chapman & Hall.

DataMyte Corporation. 1987. *DataMyte Handbook*, third edition. Minnetonka, Minnesota: DataMyte Corporation.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: John Wiley & Sons.

Chapter 3

COMPARING REGRESSION LINES

Background Information

An important technique in regression analysis involves the use of indicator variables as well as the usual quantitative variables. Consider an example given in Myers (1990): a chemical engineer is modeling yield of a reaction y as a function of temperature (x_1) and pressure (x_2); the goal is a model that accounts for the fact that two different catalysts are used. Catalyst would qualify as a categorical variable; an indicator variable could be used to help understand its role in the relationship between the y and x variables.

A convenient way to study the effect of a qualitative variable is to use indicator or "dummy" variables that take on the values of 0 and 1; that is, a variable usually coded 1 to indicate the presence of an attribute and 0 to indicate its absence. The indicator or dummy variables may sometimes appear as unplanned nuisance variables; at other times they may be planned as part of the research.

Neter et al. (1996) suggest examples of qualitative explanatory variables as gender (male, female), purchase status (purchase, no purchase), and disability status (not disabled, partly disabled, fully disabled). Researchers in business, economics, and the social and biological sciences use many indicator variables to compare regression models across groups because they want to simplify a model by determining that grouping is not significant, or they want to improve the prediction capability of a model by using groups.

Comparison of Regression Lines Analysis in STATGRAPHICS *Plus*

The Comparison of Regression Lines Analysis in STATGRAPHICS *Plus* automatically constructs the necessary indicator variables so you can compare two or more simple regression models, usually to see whether you can use a single model across groups. You first enter a model of the form Y = a + bX and a grouping variable, and STATGRAPHICS *Plus* adds the indicator variables to the model. In effect, this fits separate lines for each level of the categorical variable and allows you to determine if the intercepts and/or the slopes differ among the levels of that variable.

The tables and graphs in the program are designed specifically for comparing slopes and intercepts. The tabular and graphical options are helpful when either the slopes or intercepts do not differ significantly and you want to force them to be

equal to simplify the model. The final step in the model-building process is to validate the model. For information about model validation, see the topic "Overview of the Model-Building Process," in Online Help.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... COMPARISON OF REGRESSION LINES... from the Menu bar to display the Comparison of Regression Lines Analysis dialog box (see Figure 3-1).

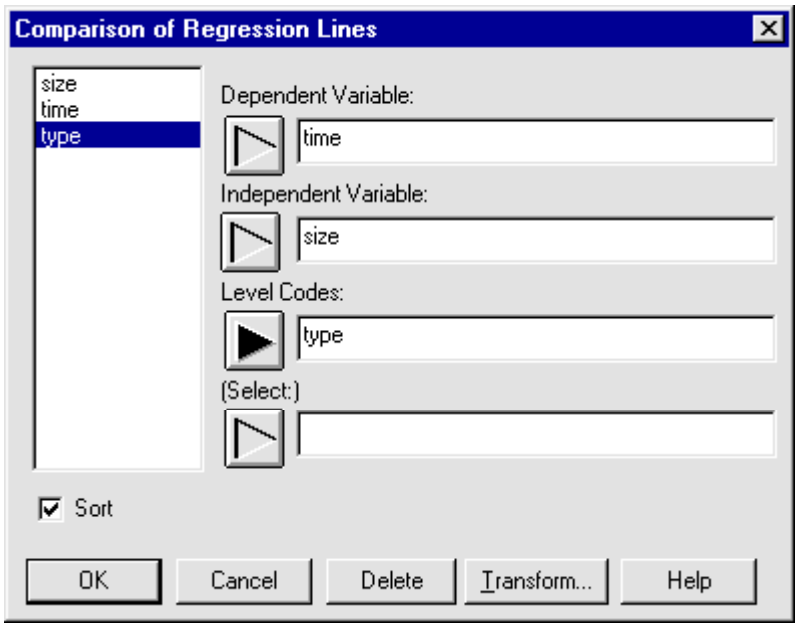


Figure 3-1. Comparison of Regression Lines Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option shows the results of the multiple regression analysis (see Figure 3-2). If the display includes the explanation provided by the StatAdvisor, the equation of the fitted model and an explanation of the terms in the equation are shown. By default, the model contains an intercept, a coefficient for the X variable, a coefficient for the indicator variable for each group except the first, and for products of the X variable with each group indicator.

The table then shows the results of the fit:

- If the p-value in the ANOVA table is less than 0.01, the results indicate that there is a statistically significant relationship among the variables at the 99 percent confidence level.

- The R-Squared statistic indicates the percentage of variability the dependent variable accounts for as it was fitted in the full model.

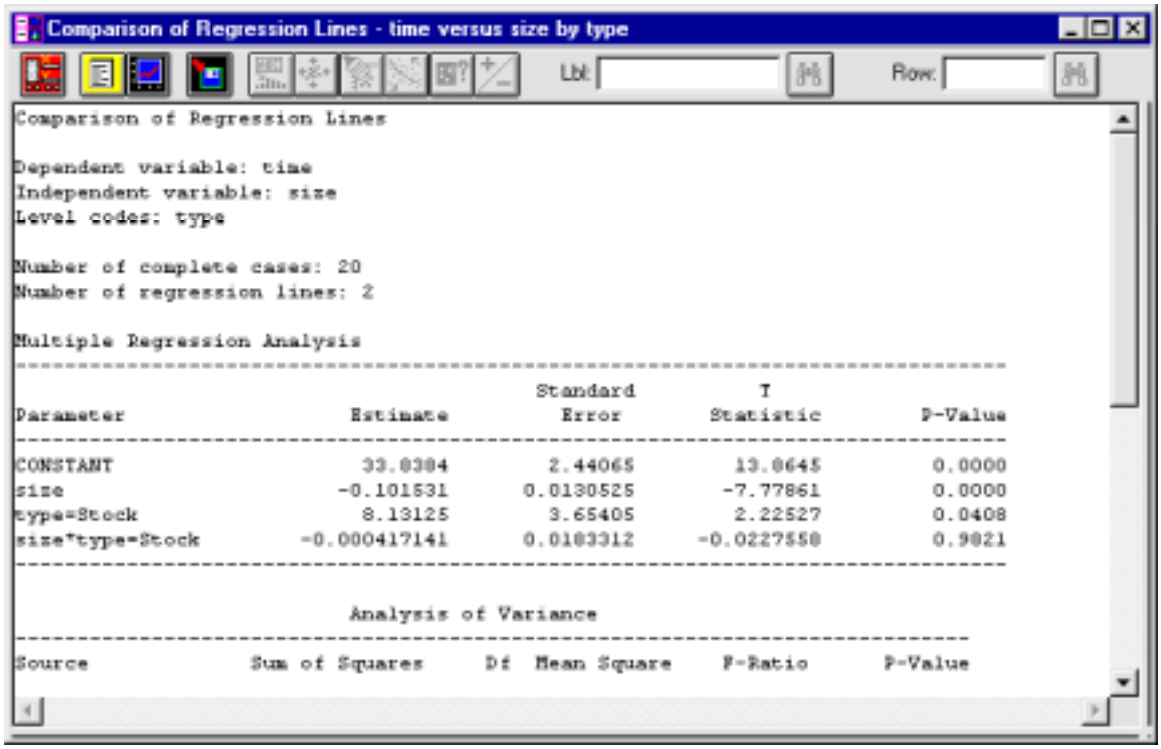


Figure 3-2. Analysis Summary

- The Adjusted R-Squared statistic, which is more suitable for comparing models that have different numbers of independent variables and indicates the percentage of variability represented by the model after accounting for the number of parameters that were estimated.

- The Standard Error of the Estimate shows the value for the standard deviation of the residuals. You can use this value to construct prediction limits for new observations by selecting the Forecasts tabular option.

- The value for the Mean Absolute Error (MAE) is the average absolute value of the residuals.

- You can use the value for the Durbin-Watson statistic to test the residuals to determine if there is significant serial correlation based on the order in which the values fall in the file. The StatAdvisor contains recommendations for further tests and plots.

If you used the optional Select text box on the Comparison of Regression Lines Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis

Summary. If you decide to validate the model, use the methods discussed in the topic, Overview of the Model-Building Process," in Online Help.

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.
- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the set of data, then finding the average or mean value for the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.
- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.
- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.
- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.
- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Use the *Comparison of Regression Lines Options* dialog box to force equal slopes or intecepts for all the groups.

Conditional Sums of Squares

The Conditional Sums of Squares option allows you to test the statistical significance of the terms in the model (see Figure 3-3). The option produces an analysis of variance table that shows the values for additional sums of squares as the program adds various terms to the model. The program adds the sums of squares for the linear indicator variable terms, labels them as intercepts, and groups the sums of squares for the interaction terms of the X variable with the indicators (slopes).

Confidence Intervals

The Confidence Intervals option displays the confidence intervals for the coefficients in the model (see Figure 3-4). You use confidence intervals to see the preciseness of

the estimates for the coefficients, given the amount of available data and the amount of noise.

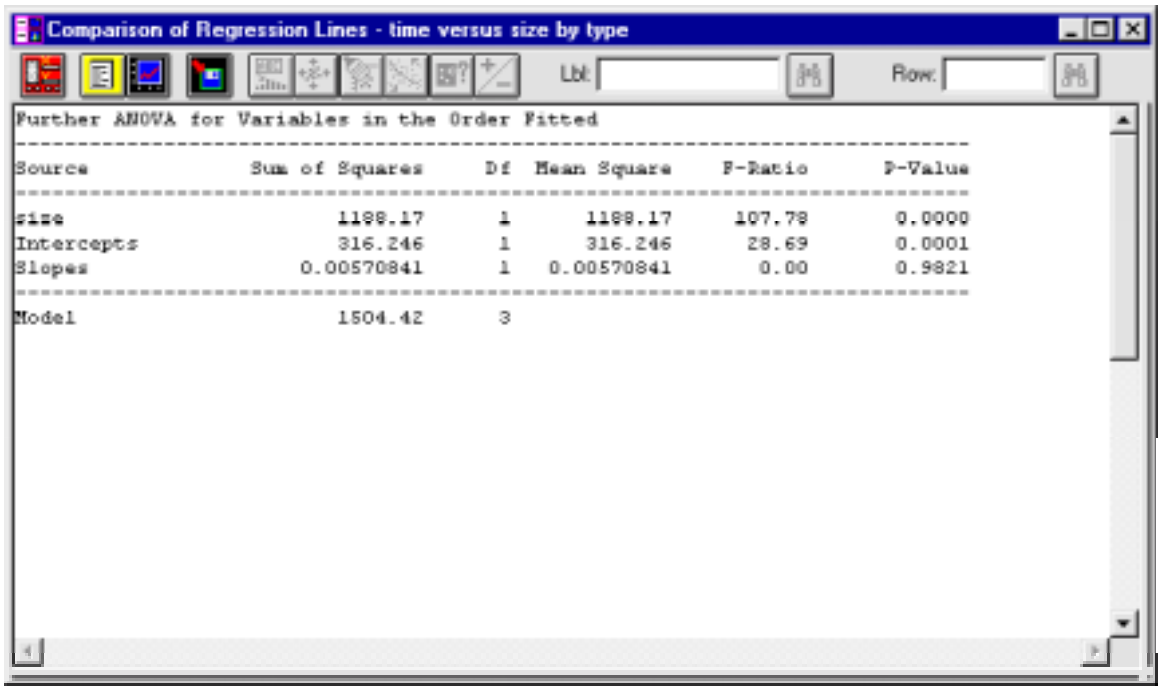


Figure 3-3. Conditional Sums of Squares

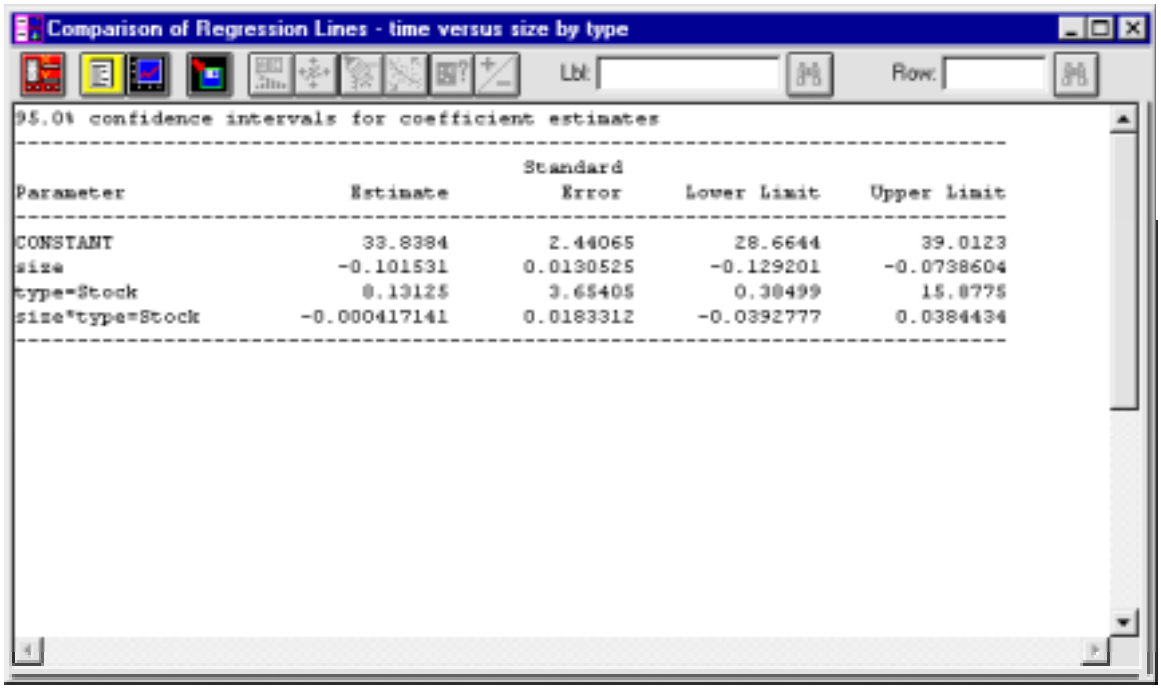


Figure 3-4. Confidence Intervals

Use the *Confidence Intervals Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals for the difference between the estimates.

Forecasts

The Forecasts option generates and displays predictions for the variable using the fitted model (see Figure 3-5). The display includes predictions for the values of the Y variable at each level code and includes values for the prediction and confidence limits. Default predictions are shown for the minimum and maximum values of the X variable.

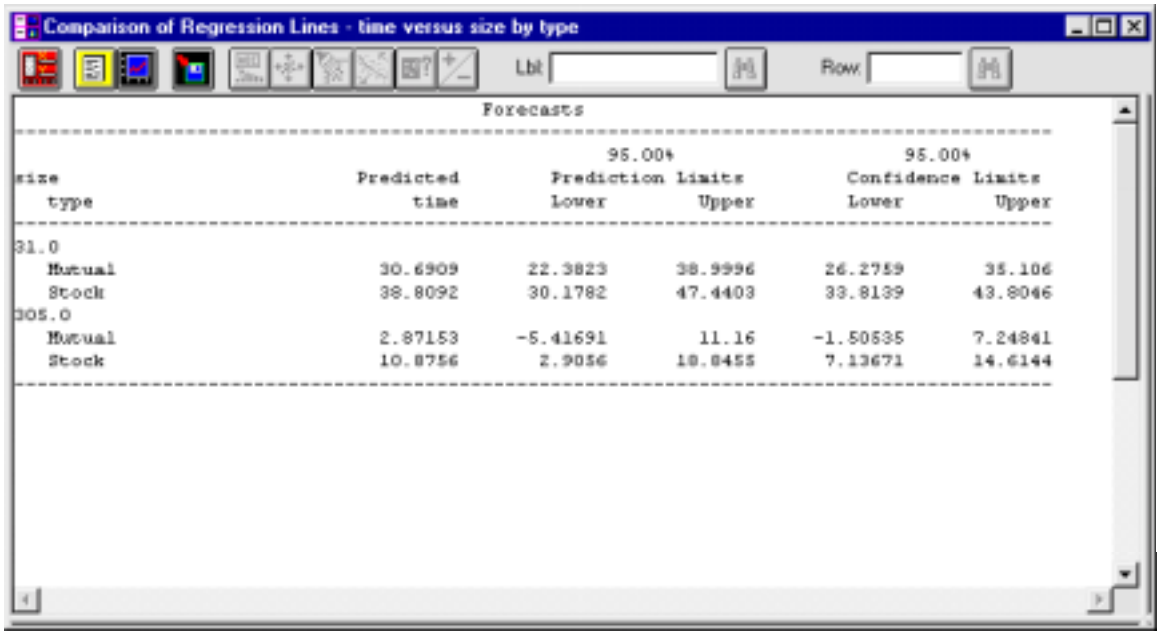


Figure 3-5. Forecasts

Use the *Forecasts Options* dialog box to choose a value for the confidence level and for the X variable.

Unusual Residuals

The Unusual Residuals option lists all the observations that have unusually large studentized residuals — greater than 2.0 in absolute value (see Figure 3-6). Studentized residuals are used to measure the number of standard deviations each observed value deviates from the model fitting using all the data except that observation.

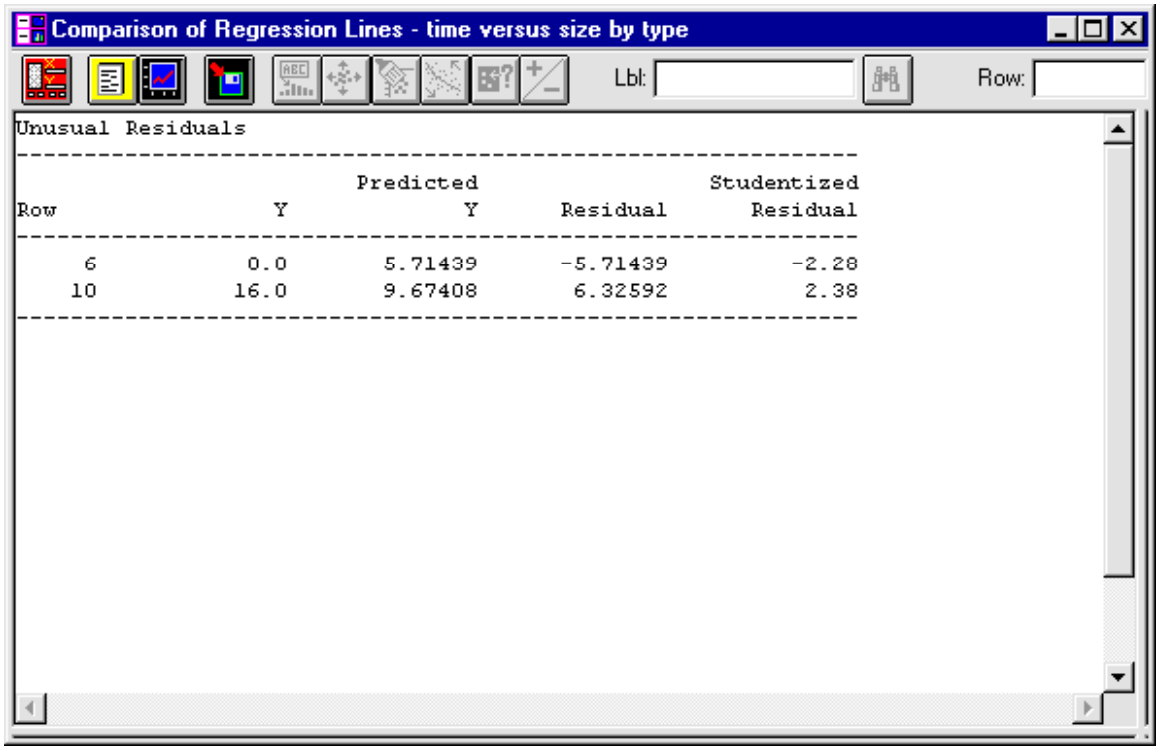


Figure 3-6. Unusual Residuals

Influential Points

The Influential Points option identifies all the observations that have leverage values greater than three times that of an average point, or that have unusually large DFITS or Cook's distance values (see Figure 3-7).

The leverage statistic helps determine the coefficients of the estimated model by measuring the amount of influence that can be attributed to each observation. The DFITS statistic measures the amount of change for each estimated coefficient if the observation is removed from the data. Cook's distance measures the distance between the estimated coefficients with and without each observation.

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model option displays a plot of the fitted model (see Figure 3-8). A separate line is shown for each level code.

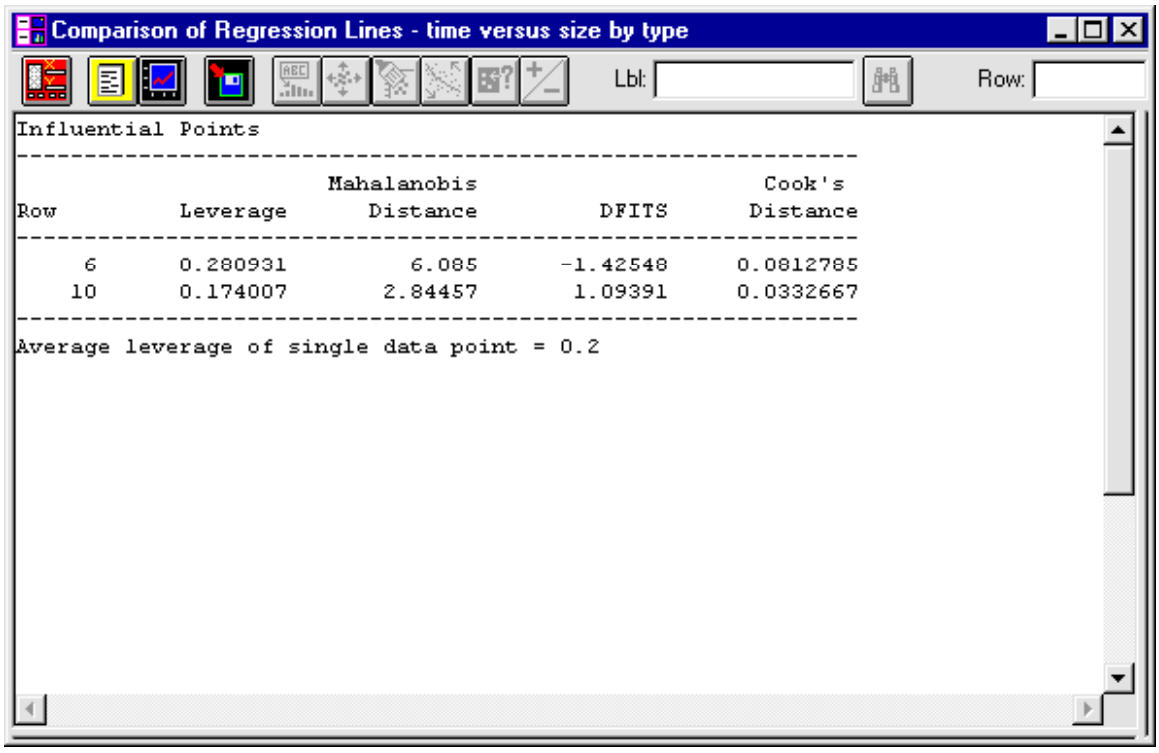


Figure 3-7. Influential Points

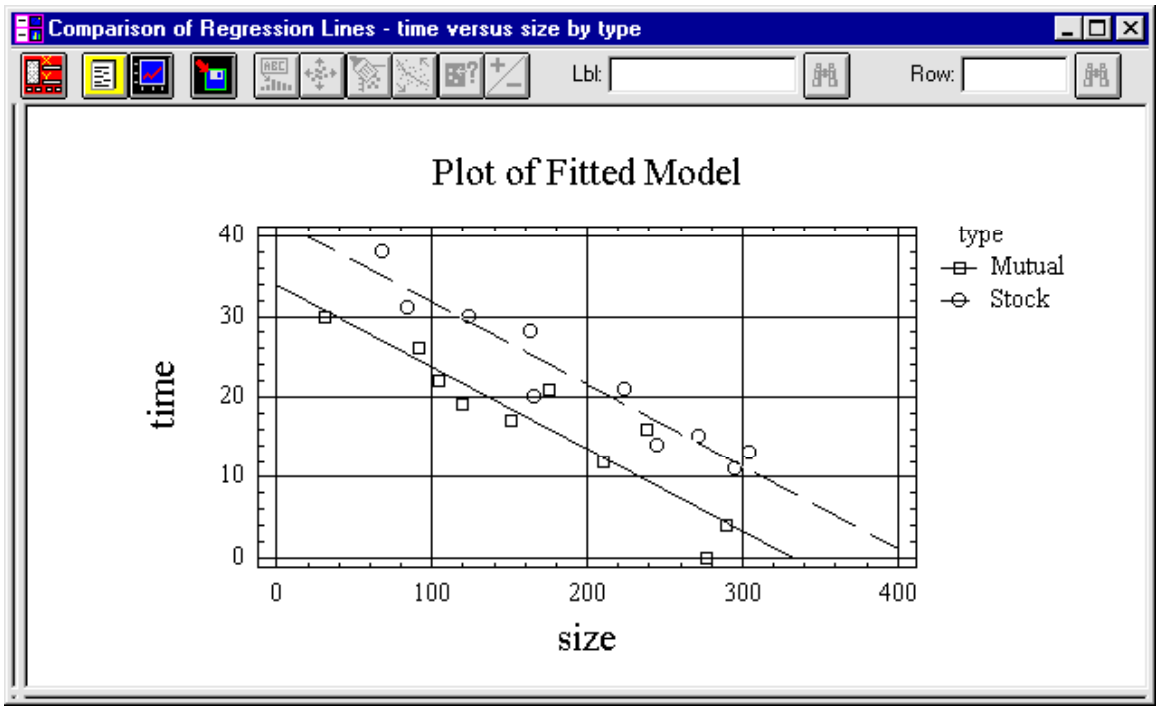


Figure 3-8. Plot of Fitted Model

Observed versus Predicted

The Observed versus Predicted option displays a plot of the observed values (Y) versus the values predicted by the fitted model (see Figure 3-9). The plot includes a line with slope equal to one. The closer the points lie to the diagonal line, the better the model predicts the observed data. Use the plot to determine cases in which the variance is not constant, which indicates that you should probably transform the values for the dependent variable.

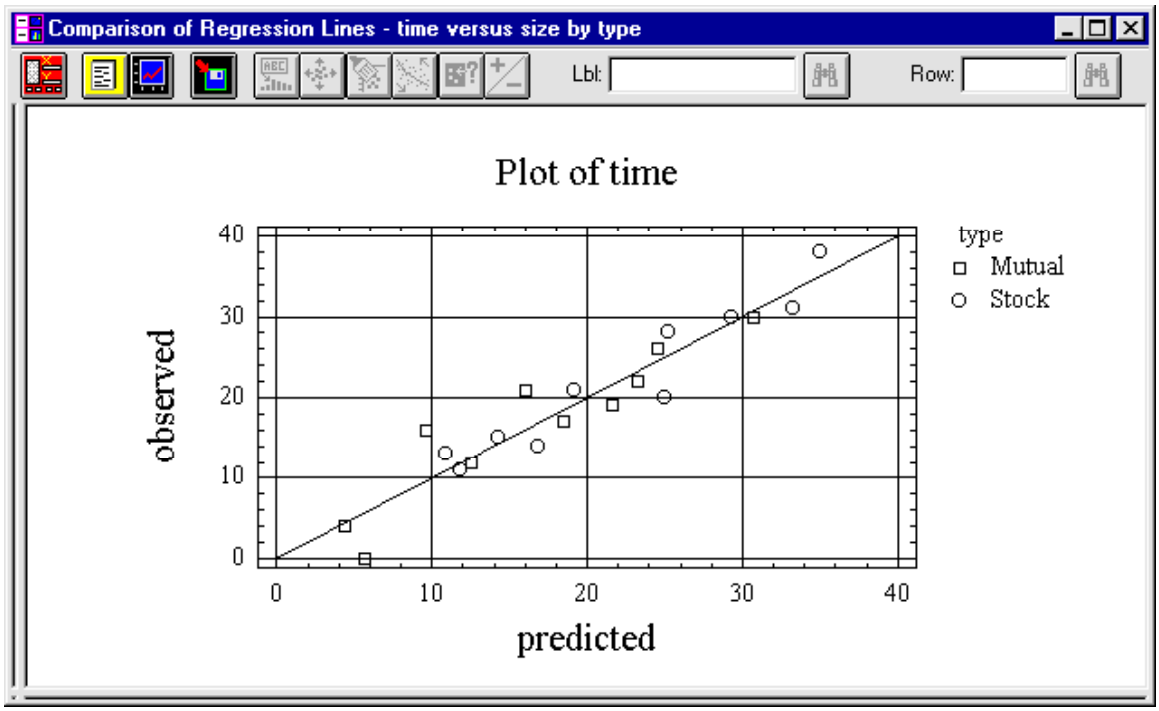


Figure 3-9. Observed versus Predicted

Residual Plots

The Residual Plots option displays one of three different plots: a scatterplot of the residual versus the values predicted by the fitted model, the row number, or by X; a Normal Probability Plot; and an Autocorrelation Function Plot.

Use the *Residual Plots Options* dialog box to choose a plot and, if applicable, its options.

Residual versus Predicted

The Residual versus Predicted options displays the residual or the studentized residual versus the predicted values for the observed variable (Y) (see Figure 3-10). A nonrandom pattern indicates that the chosen model does not adequately describe the observed data. The plot is helpful in showing heteroscedasticity; an indication of the variability of the changes in the residuals as the values of the dependent variable change.

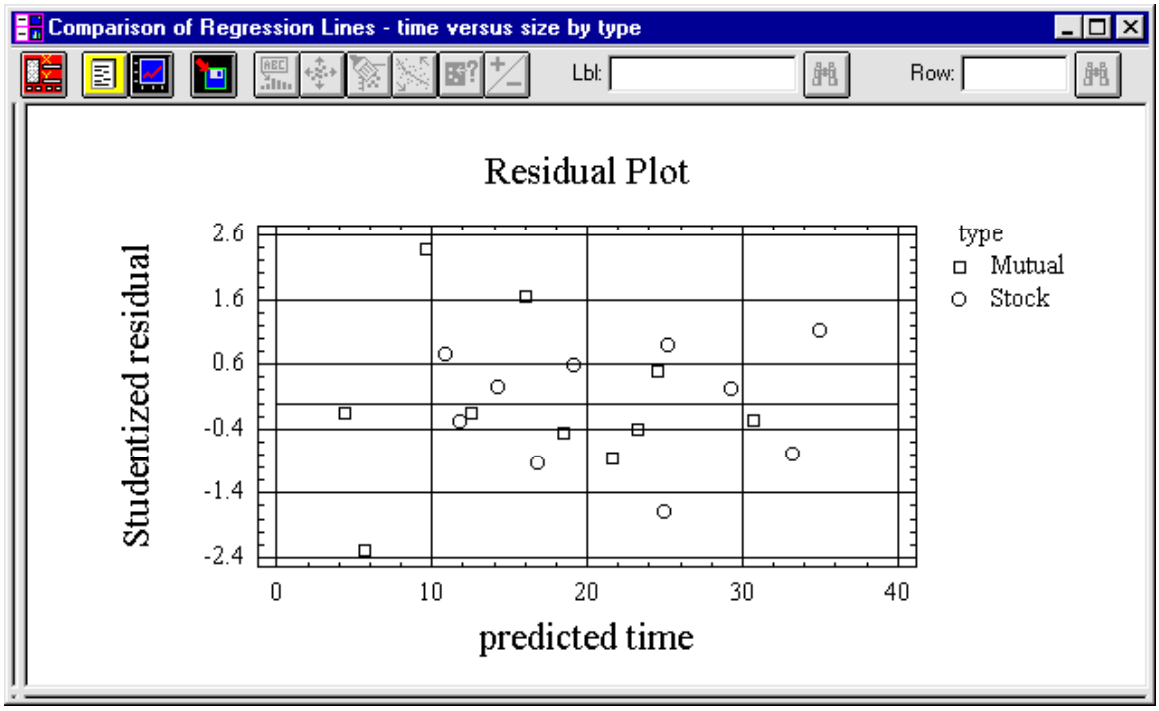


Figure 3-10. Residual versus Predicted Scatterplot

Residual versus Row Number

The Residual versus Row Number plot displays the residual or the studentized residual versus the row number (see Figure 3-11). The plot is helpful in determining sequential correlations. Any nonrandom pattern indicates serial correlation in the data, particularly if the row order corresponds to the order in which the data were collected.

Residual versus X

The Residual versus X plot displays the residual or studentized residual versus the independent variable (X), where X equals the name of the independent variable (see Figure 3-12). Use this plot to detect the nonlinear relationship between the dependent and independent variables. You can also use the plot to determine if the variance of the residual is constant. Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual values outside the range of -3 to +3 might be outliers.

Normal Probability Plot

The Normal Probability Plot displays the residuals the program uses to determine if the data follow a normal distribution (see Figure 3-13). A Normal Probability Plot consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as

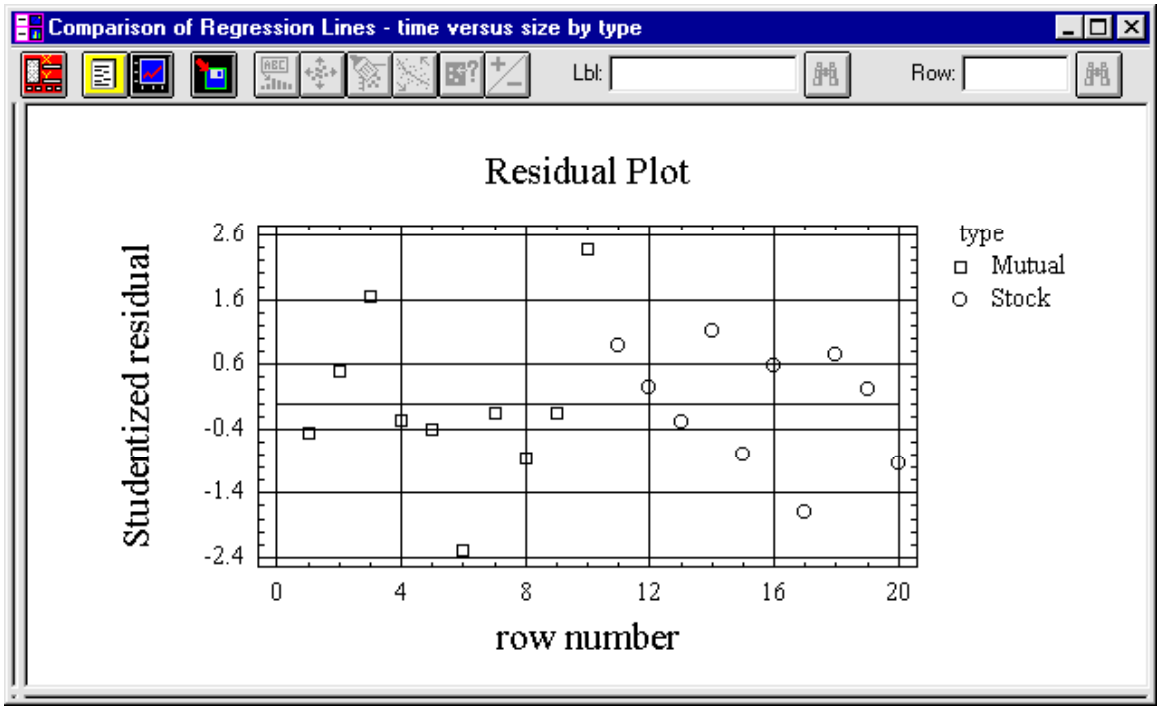


Figure 3-11. Residual versus Row Number Scatterplot

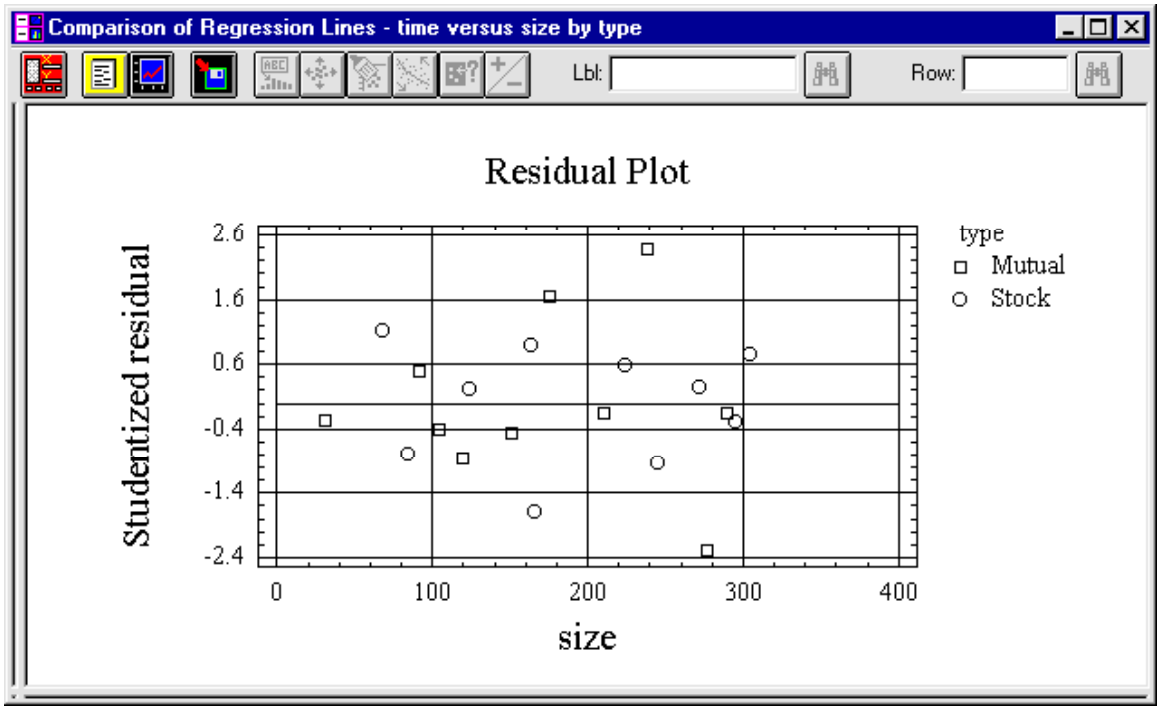


Figure 3-12. Residual versus X Scatterplot

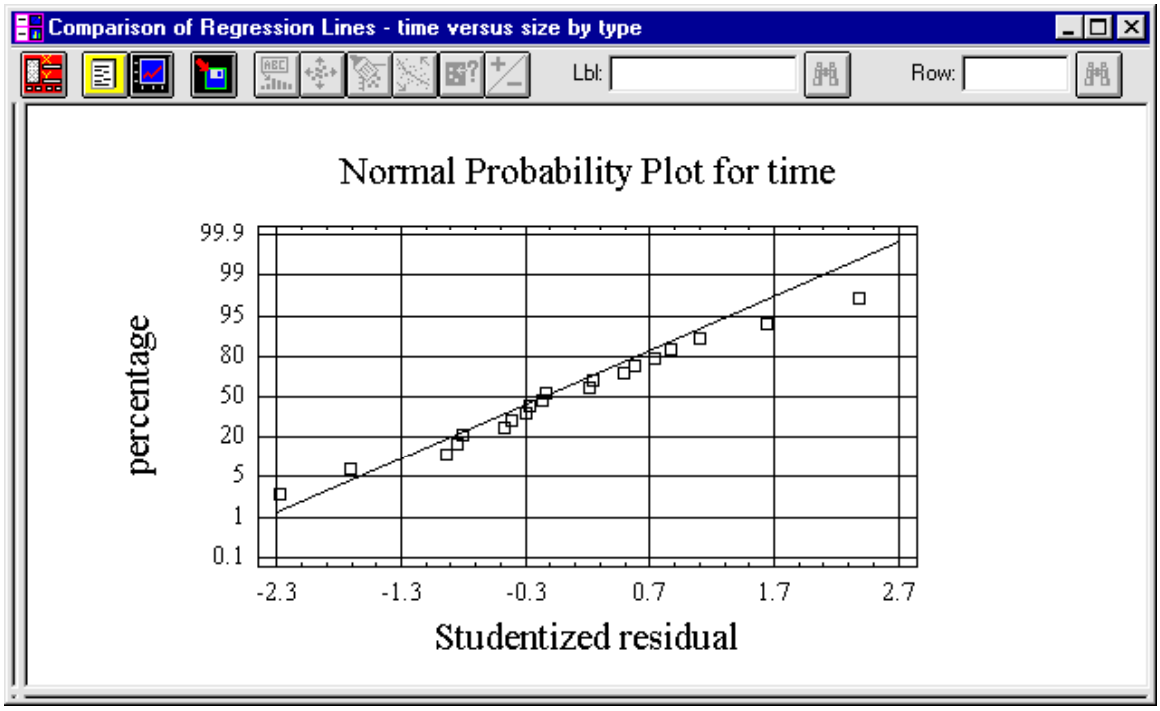


Figure 3-13. Normal Probability Plot

a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution.

Autocorrelation Function

The Autocorrelation Function plot displays a graph of the autocorrelation estimates for the residuals (see Figure 3-14). The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits. When you choose this option, the Number of Lags and Confidence Level text boxes are activated.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 13 selections: Predicted Values, Standard Errors of Predictions, Lower Limits for Predictions, Upper Limits for Predictions, Standard Errors of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, DFITS Statistics, Mahalanobis Distances, and Coefficients.

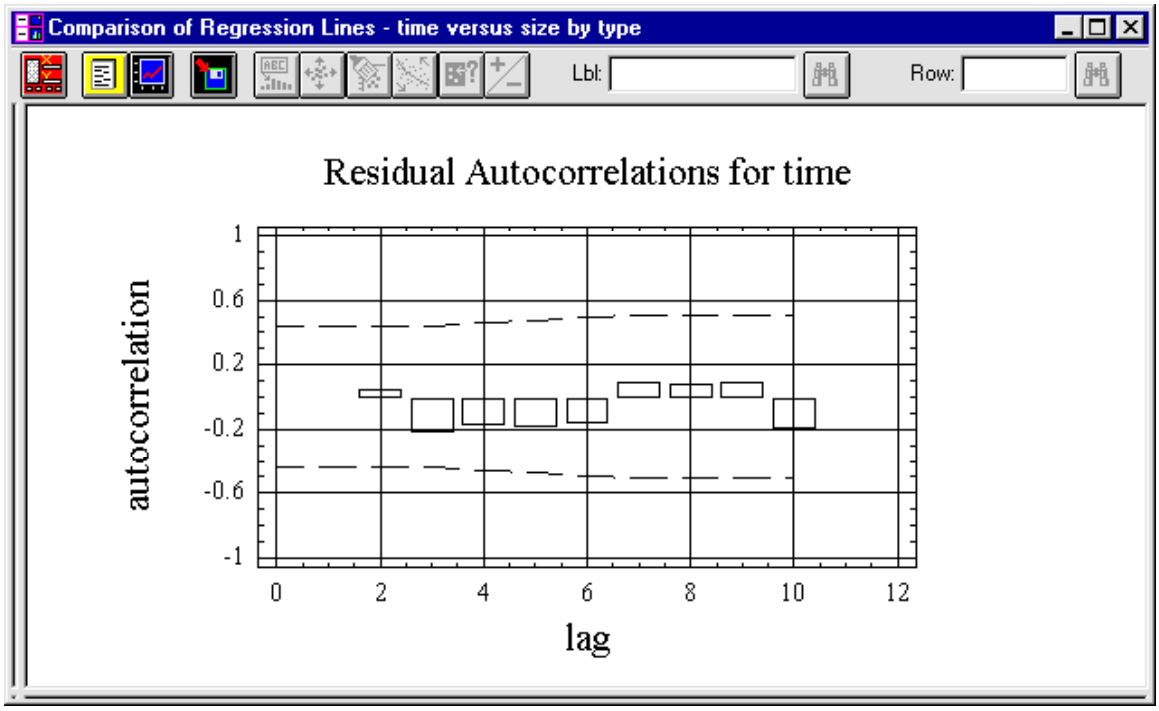


Figure 3-14. Autocorrelation Function Plot

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: John Wiley & Sons.

Chatterjee, S. and Price, B. 1991. *Regression Analysis by Example*, second edition. New York: John Wiley & Sons.

Draper, N. R. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: John Wiley & sons.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, **38.**

Montgomery, D. C. 1991. *Design and Analysis of Experiments*, third edition. New York: John Wiley & Sons.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago: Richard D. Irwin, Inc.

Vogt, W. P. 1993. *Dictionary of Statistics and Methodology*. Newbury Park, California: Sage Publications, Inc.

Chapter 4

Choosing Regression Models

Background Information

Multiple regression analysis involves fitting a random variable to a set of explanatory variables to provide regression coefficients for a linear model. The resulting model yields many different statistics you can use to determine how well the model fits the data; that is, to determine the "best" model. How well the model describes the data determines the degree to which the conclusions are believed reliable. However, it is important to note that a model accepted for one scenario will not necessarily provide an acceptable solution for another.

Frequently a problem arises when you are faced with determining which of several candidate variables to include from the model. Further complications occur when multicollinearity exists, when the quality of the data is suspect, or when outside factors such as budget or security issues arise during data collection.

Myers (1990), notes that prior to beginning the analysis it is a good idea to learn something about the nature of the data; for example, can you expect a positive or negative sign for a particular coefficient, or can you determine which regressor variables are important and which are not. Most importantly, your goal should be to create an uncomplicated model.

Neter et al. (1996) also notes that it is important to remember that linear models are experimental approximations and that with many sets of data you can fit different models that are nearly equal in effectiveness. The problem is to select the one "best" model from a pool of candidate models.

Regression Model Selection in STATGRAPHICS *Plus*

The Regression Model Selection Analysis in STATGRAPHICS *Plus* ranks the best subsets of the explanatory variables, based on the criterion you enter in the Regression Model Selection Analysis dialog box, calculates the statistics for all possible linear regression models, and sorts the values so you can choose the "best" model. After you decide upon and thoroughly check the final regression model, you should validate it using the the Multiple Regression Analysis to examine the relationship between the dependent variable and the final set of independent variables.

Note: While the Regression Model Selection Analysis is a useful tool for exploratory model-building, you cannot rely on any one statistical test to identify a "true" model. Remember that using the values of the C_p and R-Squared statistics to select variables has long been a subject of controversy in statistical literature.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... REGRESSION MODEL SELECTION... from the Menu bar to display the Regression Model Selection Analysis dialog box (see Figure 4-1).

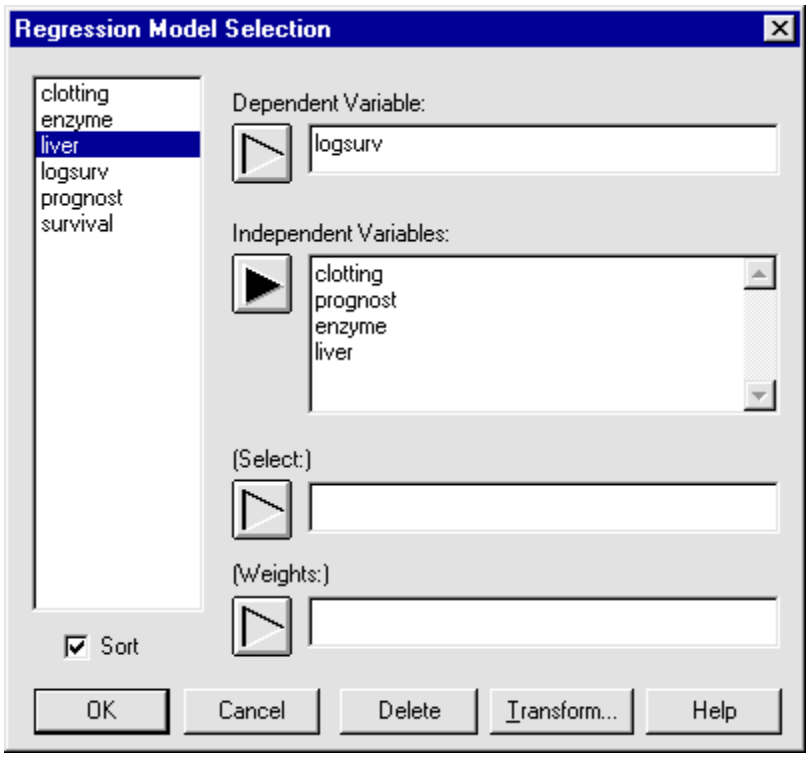


Figure 4-1. Regression Model Selection Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option displays statistics you can use to compare various multiple regression models with the dependent variable and the subsets of the independent (predictor) variables (see Figure 4-2). It also provides a key for the labels on all the tables and graphs; for example, A = Variable1, B = Variable2, and so on. Statistics in the table include values for the Mean Squared Error (MSE), R-Squared, Adjusted R-Squared, Mallows' C_p, and a model identifier for each of the subsets of variables.

Better models have low MSE values, low Mallows' C_p values (close to p — the number of coefficients in the model including the constant), and high Adjusted R-Squared values. The program sorts the models by the number of variables. The Adjusted R-Squared statistic measures the proportion of variability in the independent variable for which the model is accountable. This statistic is useful for comparing regression models that have different numbers of independent variables.

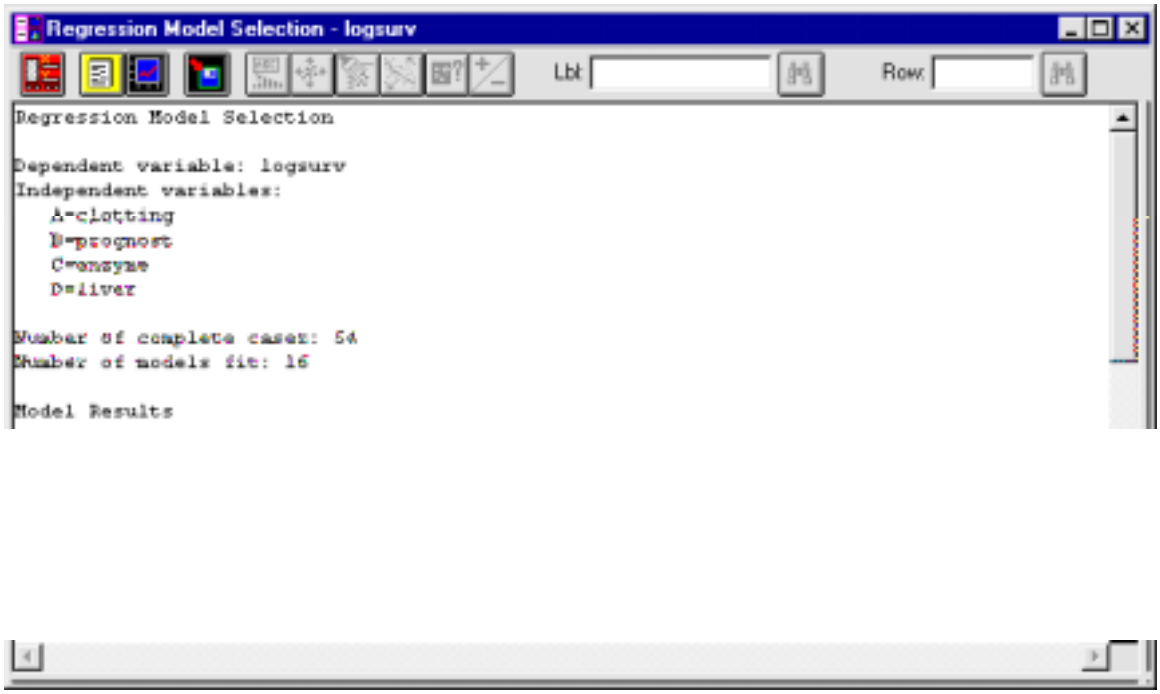


Figure 4-2. Analysis Summary

Use the *Regression Model Selection Options* dialog box to enter a number that will be used to limit the number of subsets that will be ranked. This is done by placing restrictions on the number of independent variables in the smallest and largest models.

Best Adjusted R-Squared

The Best Adjusted R-Squared option lists the models that give the largest Adjusted R-Squared values (see Figure 4-3). The table lists the models in order of decreasing value, which places the "best" models near the top of the list. The Adjusted R-Squared statistic measures the proportion of variability in the dependent variable for which the model is accountable. Larger values of the Adjusted R-Squared statistic correspond to smaller values of the Mean Squared Error (MSE).

Use the *Best Adjusted R-Squared Options* dialog box to enter a number that will determine the maximum number of models that will be listed for each subset size. For example, if you enter the names of 10 independent variables in the Regression

Model Selection Analysis dialog box, and enter 3 in this text box, up to the three "best" models for each subset size — three with nine variables, three with eight variables, and so on — will be listed.

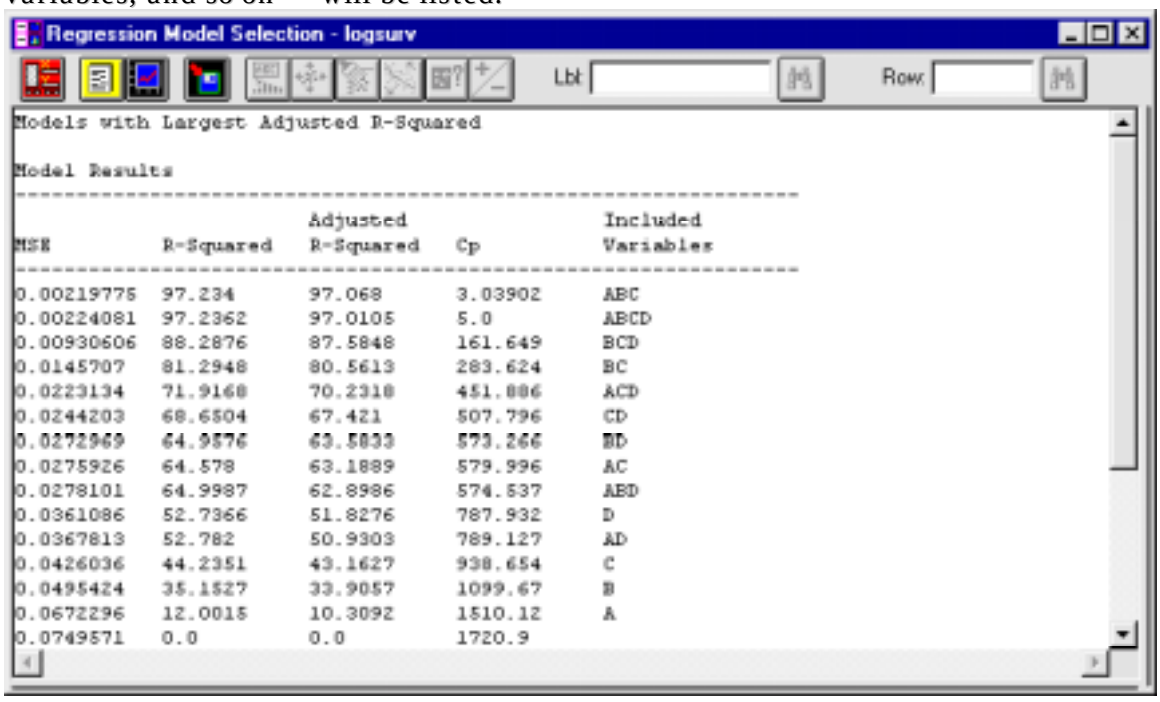


Figure 4-3. Best Adjusted R-Squared

Best C_p

The Best C_p option displays the models that have the smallest values of the Mallows' C_p statistic (see Figure 4-4). C_p is a measure of the bias in the model based on a comparison of total Mean Squared Error to the true error variance. Unbiased models have an expected C_p value of approximately p, where p is the number of coefficients in the fitted model. C_p is based on the assumption that the model that contains all the candidate variables is unbiased; therefore, the full model will always have C_p = p. Look for models that have C_p values close to p. The models are listed in order of increasing value, which places the "best" models near the top of the list. For more information, see Mallows (1973, 1995).

Use the *Best C_p Options* dialog box to enter a number that will determine the maximum number of models that will be listed for each subset size.

Graphical Options

Adjusted R-Squared Plot

The Adjusted R-Squared Plot option displays a plot of the Adjusted R-Squared values for each model plotted against the value of p (the number of independent

variables, plus one) (see Figure 4-5). A line connects each number of coefficients with the models that have the highest values of the Adjusted R-Squared statistic.

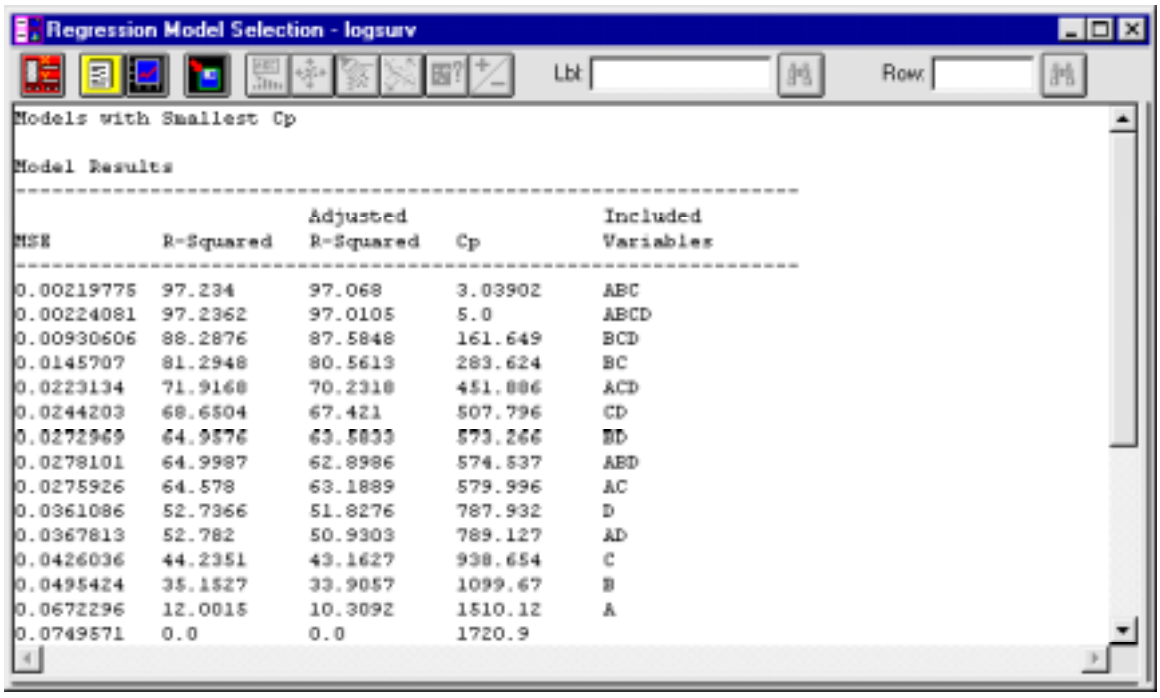


Figure 4-4. Best C_p

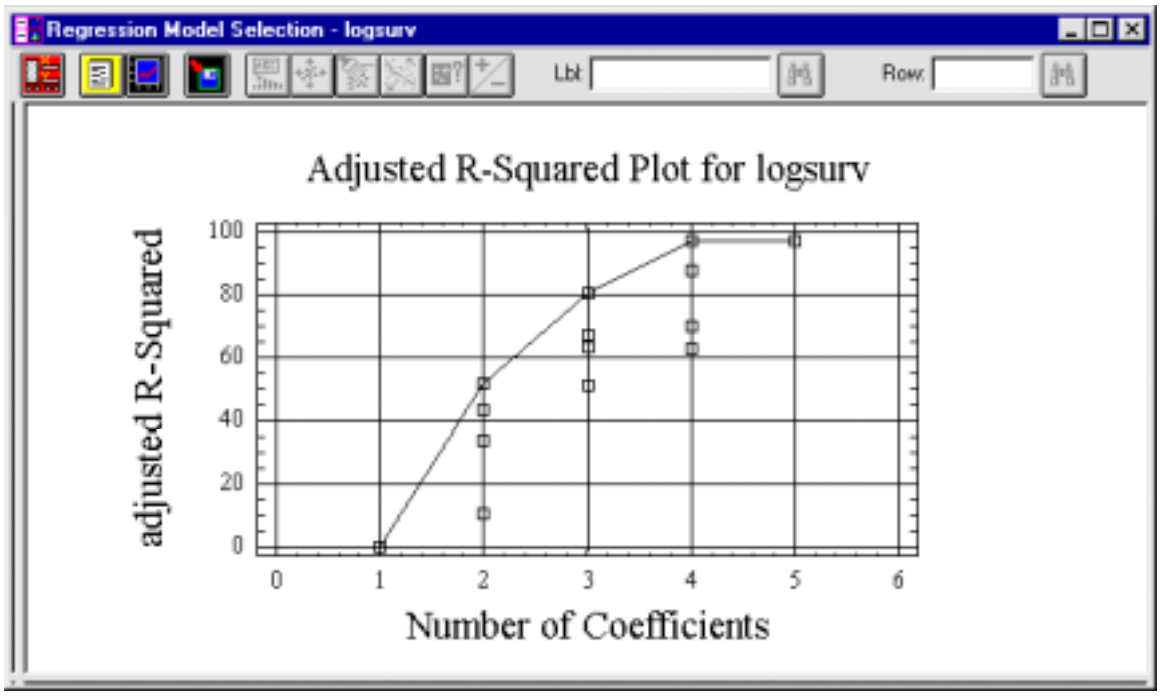


Figure 4-5. Adjusted R-Squared Plot

Use the *Adjusted R-Squared Plot Options* dialog box to indicate if the points on the plot should be labeled with the names of the models according to the key in the Analysis Summary — A, AB, and so on, and to enter a number that determines the maximum number of models that will be shown per subset size.

Mallows' C_p Plot

The Mallows' C_p Plot option displays the C_p statistic for each model plotted with the number of coefficients in the model (see Figure 4-6). The program draws a reference line for $C_p = p$. C_p is a measure of the bias in the model, based on a comparison of the total Mean Squared Error with the true error variance. The program assumes that the full model is unbiased. Unbiased models have an expected value of approximately p, where p is the number of coefficients in the fitted model. Look for models that have C_p values close to p. Unbiased models should be close to the line on the plot; the full model will always have $C_p = p$. Mallows (1995) warns that for some cases, "picking the minimum C_p subset and fitting by least squares will not give a good prediction formula." If necessary, you can rescale the C_p axis to get a better view.

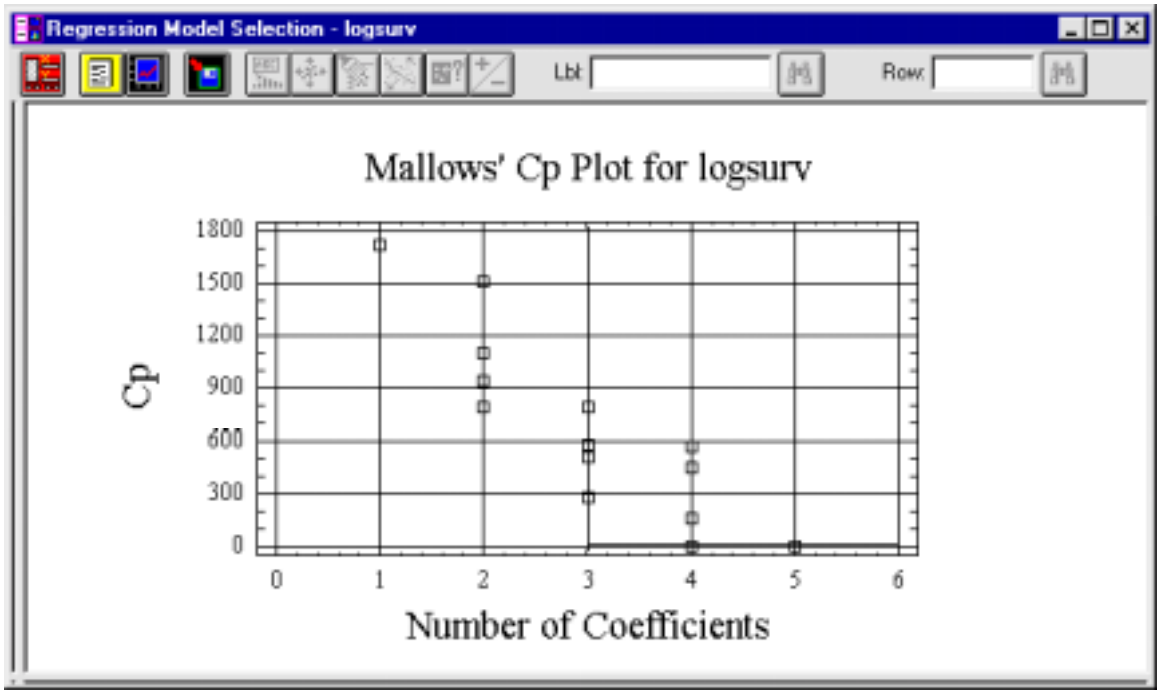


Figure 4-6. Mallows' C_p Plot

Use the *Mallows' C_p Plot Options* dialog box to indicate if the points on the plot should be labeled with the names of the models and to enter a number that will determine the maximum number of models that will be listed for each subset size.

MSE Plot

The MSE Plot option displays the Mean Squared Error (MSE) for each model plotted with the number of coefficients in the model (see Figure 4-7). A line connects each number of coefficients with the models that have the smallest values of MSE. When the line levels off, little is gained by adding variables to the model.

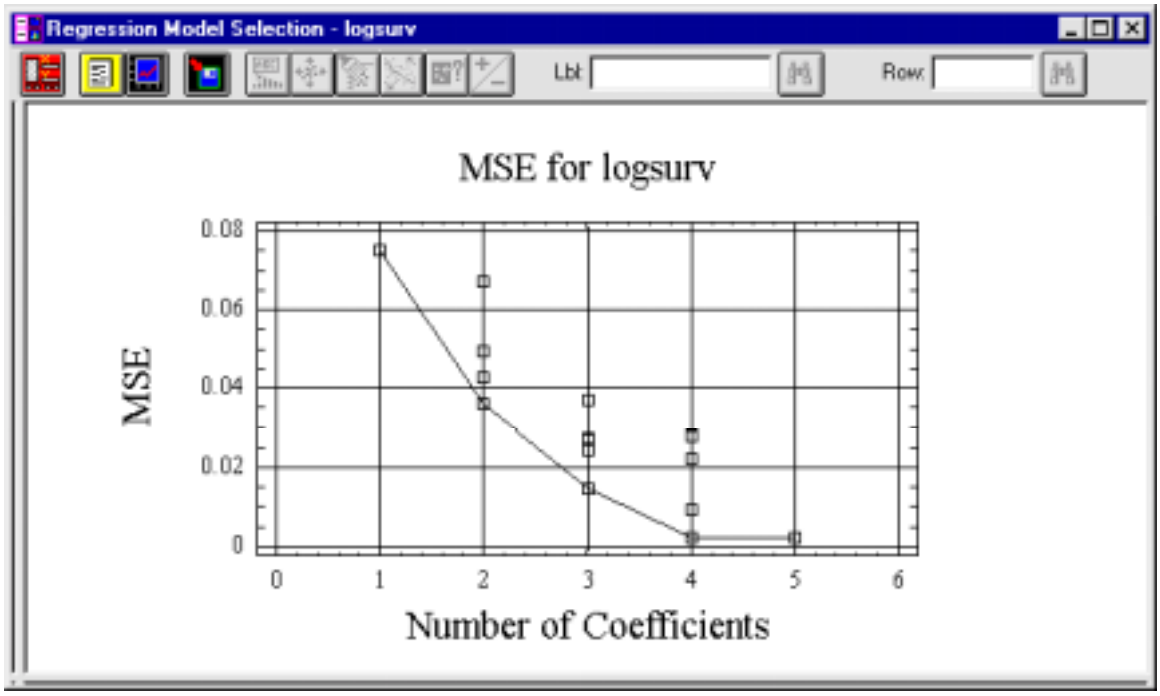


Figure 4-7. MSE Plot

Use the *MSE Plot Options* dialog box to indicate if the points on the plot will be labeled with the names of the models and to enter a number that will determine the maximum number of models that will be listed for each subset size.

R-Squared Plot

The R-Squared Plot option displays the value of the R-Squared statistic for each model plotted with the number of coefficients (see Figure 4-8). A line connects each number of coefficients with the models that have the largest R-Squared values. When the line levels off, little is gained by adding variables to the model.

Use the *R-Squared Plot Options* dialog box to indicate if the points on the plot will be labeled with the names of the models and to enter a number that will determine the maximum number of models that will be shown in each subset size.

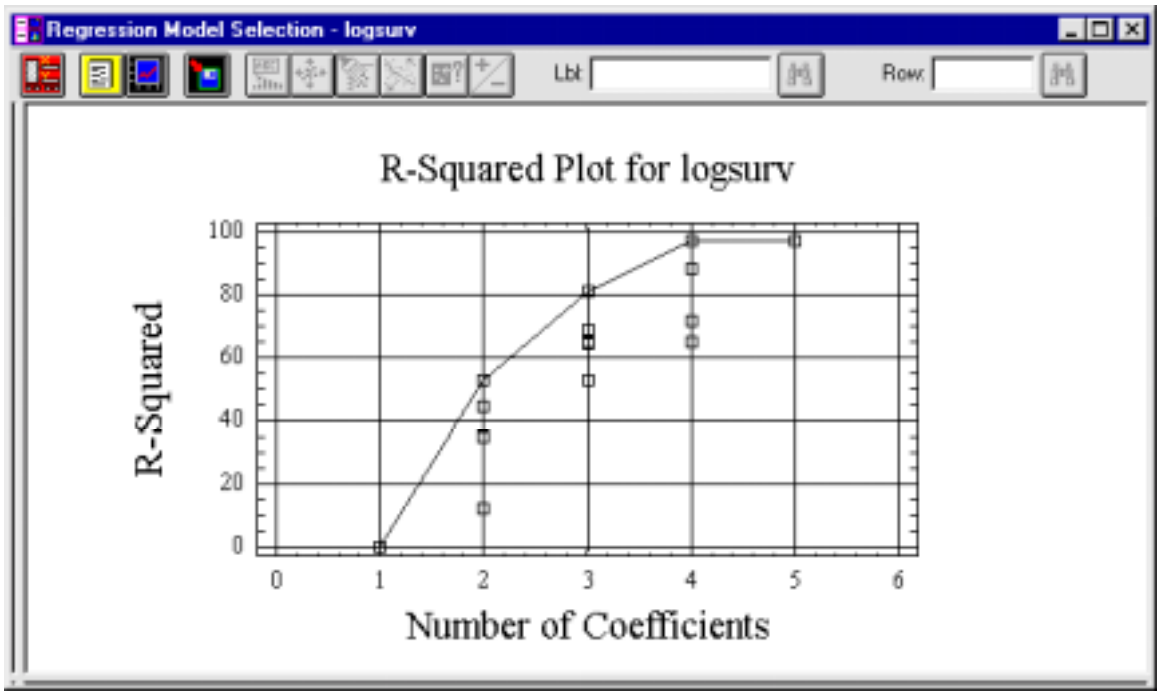


Figure 4-8. R-Squared Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Model Identifiers, Adjusted R-Squared, C_p, MSE, and R-Squared.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: John Wiley & Sons.

Mallows, C. L. 1973. "Some Comments on C_p." *Technometrics*, **15**:661-75.

Mallows, C. L. 1995. "More Comments on C_p." *Technometrics*, **37**:362-72.

Montgomery, D. C. and Peck, E. A. 1992. *Introduction to Linear Regression Analysis*, second edition. New York: John Wiley & Sons.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago: Richard D. Irwin, Inc.

Chapter 5

Using Nonlinear Regression

Background Information

The goal of nonlinear regression is to find a least squares solution for a nonlinear model, which cannot be done using matrix algebra as it is in linear regression. In linear regression, the parameters enter the model in a linear fashion and least squares estimates are calculated using matrix algebra. A least squares solution has an additional property of being the unbiased and minimum variance estimator. A regression function that is not a straight line can be linear in the parameters, such as polynomial functions and models that can be linearized by transforming the variables.

Often the form of a function that is known to be nonlinear in the parameters cannot be algebraically linearized or, if it can be, the error term is distorted while it should be normally distributed with constant variance. Neter et al. (1996) and Myers (1990) list several types of nonlinear models that regularly occur in various scientific and other applications.

Unlike linear regression, it is usually possible to find analytical expressions for the least squares by using numerical search procedures (iterations). In an iterative search, the program starts with an initial estimated value for each parameter in the equation. It calculates the sum of squares (the sum of the squares of the vertical distances of the points from the curve), then adjusts the parameters to make the curve come closer to the points. It then adjusts the parameters again so the curve comes even closer to the points. The program keeps adjusting the parameters until the adjustments virtually make no difference, then it reports the best-fit results.

The preciseness of the values depend in part on the initial values you choose and the stopping criteria. This means that if you repeat the analysis on the same data you will not always get the same exact information. It is important to begin the search using reasonable values for each parameter.

Nonlinear Regression in STATGRAPHICS *Plus*

In STATGRAPHICS *Plus*, you first define the function that will be used to fit the data. Then you choose an iterative search algorithm to determine the estimates that minimize the residual sum of squares. The methods are: Marquardt, Gauss-Newton, and Steepest Descent. All three require you to supply initial values for the estimates of the parameters; see Draper and Smith (1981), Neter et al. (1996), and

Myers (1990). Because the method you use depends on these initial parameters, you should take time to develop reasonable initial estimates. One caution regarding the initial estimates is that the model may converge to a local minimum instead of a global minimum. To partially guard against this situation, you can perform the analysis several times using widely differing initial estimates. Sometimes you can obtain initial estimates from estimates for a linearized form of the regression equation, or by using your own observations. Either way, remember that different starting values can result in very different estimates for the parameters.

Estimation also sometimes fails to converge due to misplaced parentheses or other errors in the functions; at other times, adjusting the convergence control details will help, however, there will be cases when function estimation fails because the search algorithm simply does not converge on a solution because it is not completely foolproof.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... NONLINEAR REGRESSION... from the Menu bar to display the Nonlinear Regression Analysis dialog box (see Figure 5-1). After you complete the dialog box and click OK, the Initial Parameter Estimates dialog box displays (see Figure 5-2). After you complete and process this dialog box the Analysis Summary and Plot of Fitted Model appear in the Analysis window.

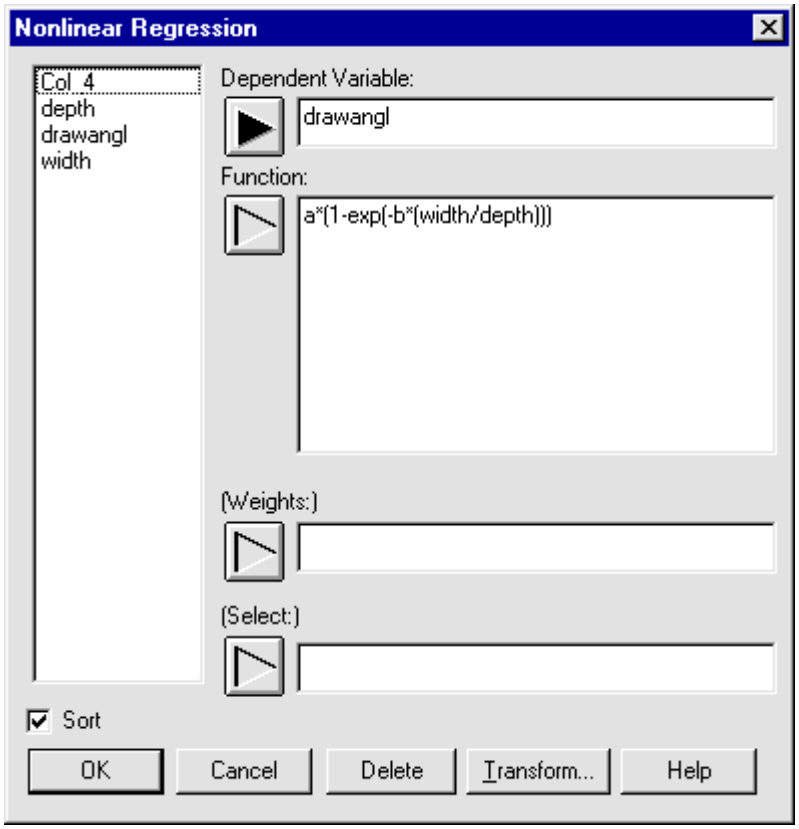


Figure 5-1. Nonlinear Regression Analysis Dialog Box

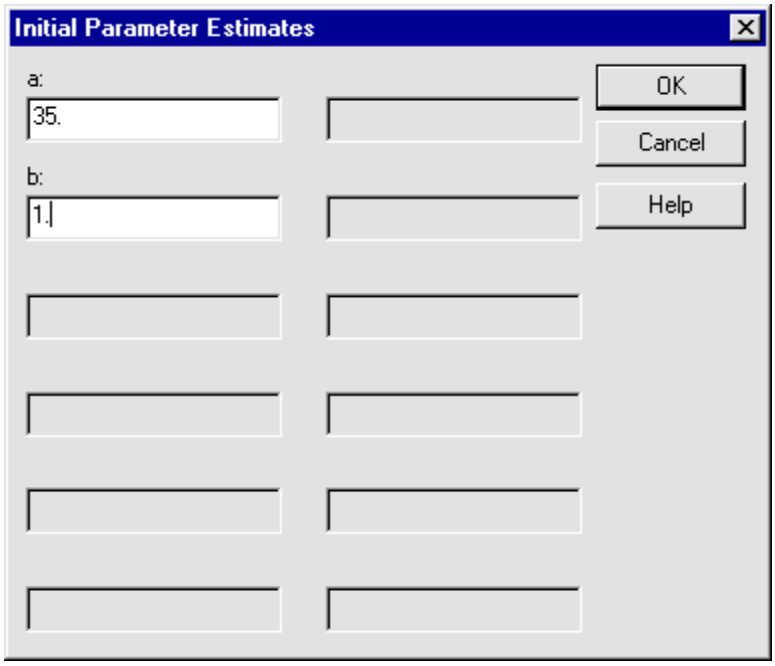


Figure 5-2. Initial Parameter Estimates Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option shows the results of fitting a nonlinear regression model that describes the relationship between the dependent variable and the estimated function (see Figure 5-3). Also shown is the number of iterations the program performed before it reached the minimum residual sum of squares. The StatAdvisor includes the equation for the fitted model.

The statistics include the following:

- The R-Squared statistic is shown, but it may not be meaningful for a linear model.
- The Adjusted R-Squared statistic is used to compare models that have different numbers of independent variables.
- The Standard Error of the Estimate shows the standard deviation of the residuals.
- The Mean Absolute Error (MAE) is the average value of the residuals.

- The Durbin-Watson statistic tests the order of the residuals as they occur in the file. If the Durbin-Watson statistic is greater than 1.4, there is probably no serious autocorrelation in the residuals.

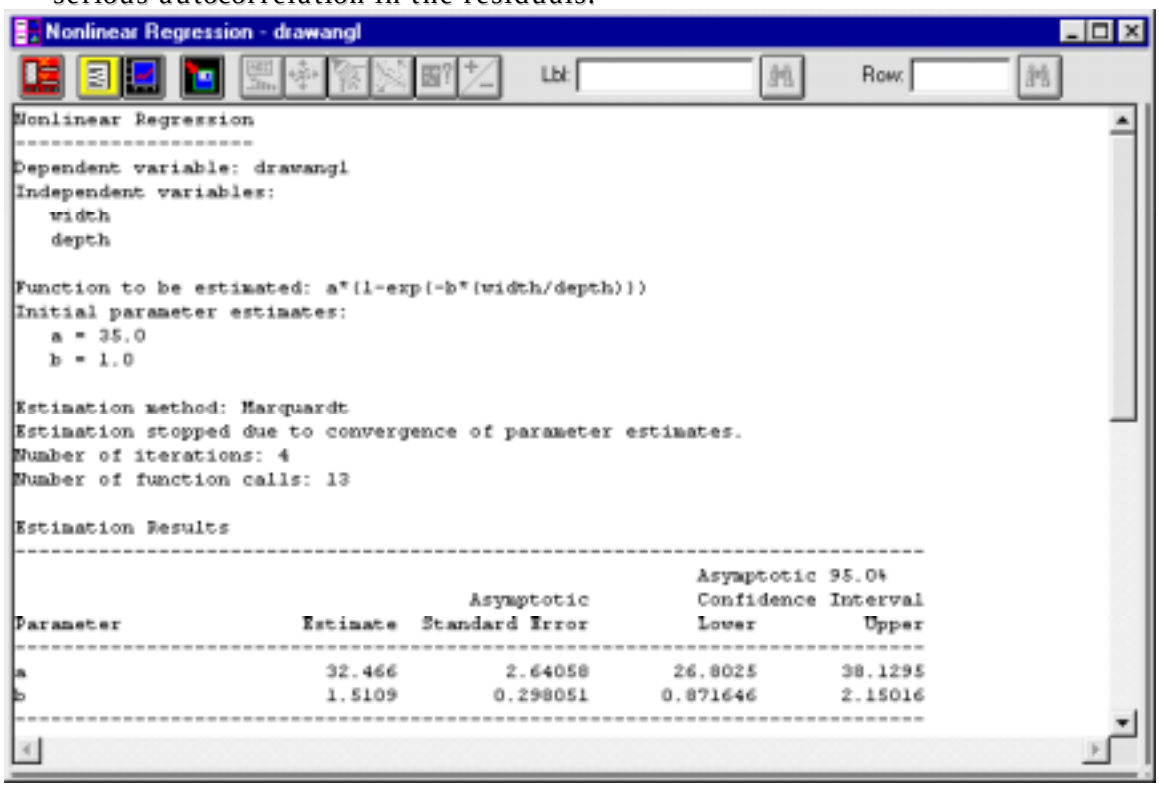


Figure 5-3. Analysis Summary

- The values for the asymptotic confidence intervals are shown for each of the unknown parameters. The intervals are approximate and most accurate for large sample sizes. A value of 0.0 in each interval indicates whether or not an estimate is statistically significant. Intervals that cover 0.0 correspond to coefficients that could be removed from the model without substantially affecting the fit.

If you used the optional Select text box on the Nonlinear Regression Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use one of the methods discussed in the topic, "Ov0erview of the Model-Building Process," in Online Help.

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.

- MSE (Mean Square Error) a measure of accuracy computed by squaring the individual error for each item in the data, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.
- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.

- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Use the *Nonlinear Regression Options* dialog box to enter values for the stopping criterions, maximum iterations, maximum function calls, and the confidence level. You can also choose the method that will be used to estimate the parameters. If you choose the Marquardt method, you can enter the initial value, the scaling factor, and the maximum value.

Correlation Matrix

The Correlation Matrix option displays the estimated correlations between the coefficients in the fitted model (see Figure 5-4). You can use the correlations to detect the presence of serious multicollinearity; that is, to see if there is correlation among the explanatory variables.

Reports

The Reports option displays information about the results using the current parameter and depending on your choices on the Reports Options dialog box (see Figure 5-5). Each item in the table corresponds to the values of the independent variables in a specific row of the file.

To create forecasts (predictions) for additional combinations of values for the variables, you can add additional rows to the bottom of the file. In each new row, enter values for the independent variables, but leave the cell for the dependent variable empty. The program adds the predicted values for the new rows to the table, and leaves the model unchanged.

Use the *Reports Options* dialog box to choose the results that will be included in the report.

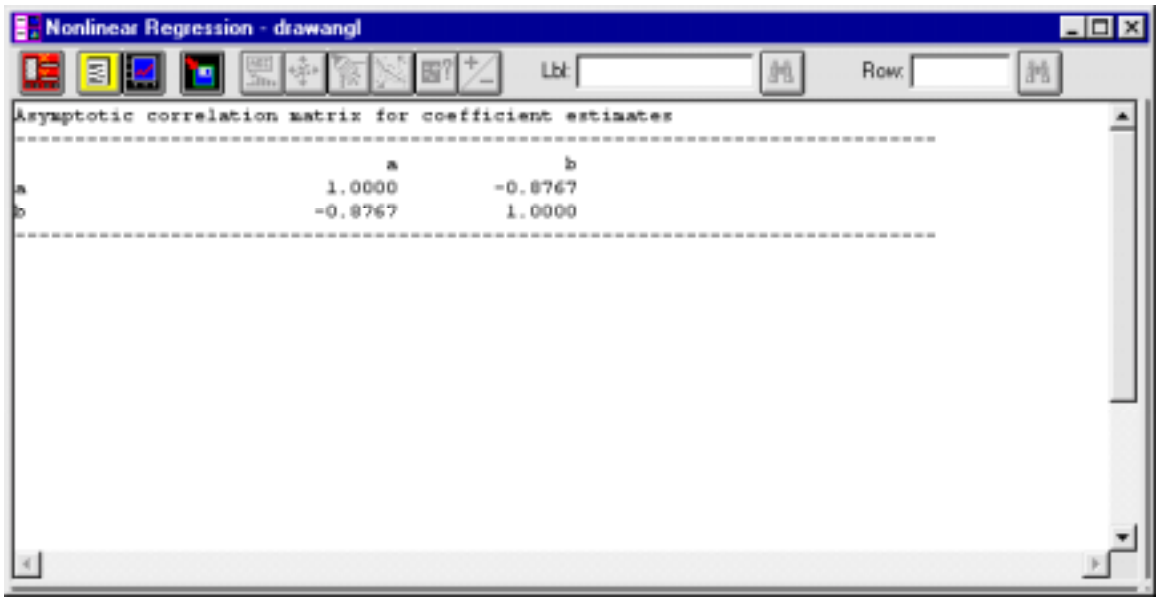


Figure 5-4. Correlation Matrix

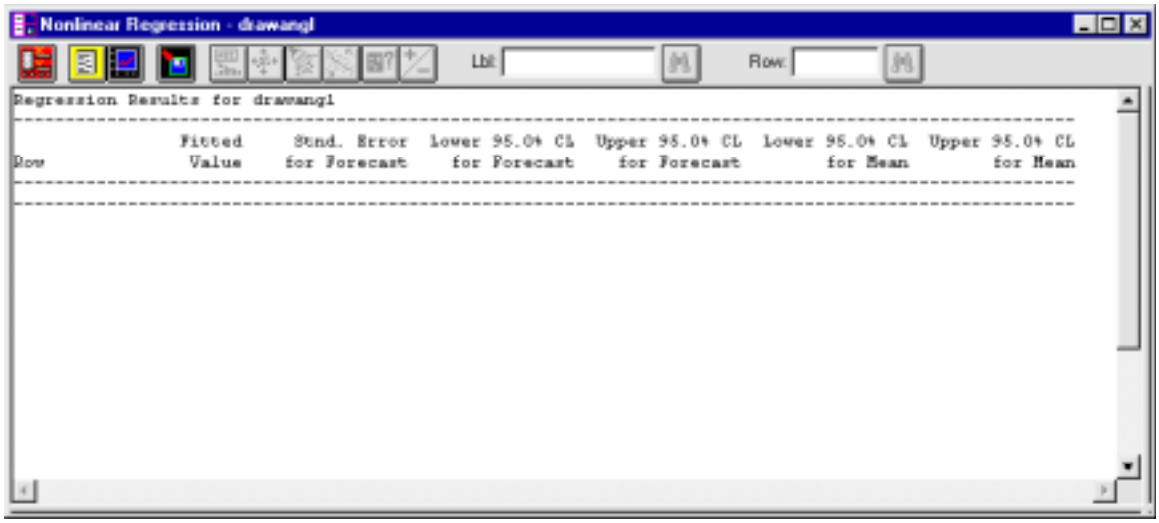


Figure 5-5. Reports Table

Unusual Residuals

The Unusual Residuals option provides a list of all the observations that have unusually large studentized residuals (greater than 2.0 in absolute value) (see Figure 5-6). Studentized residuals measure the number of standard deviations each observed value deviates from the model fitting using all the data except that observation.

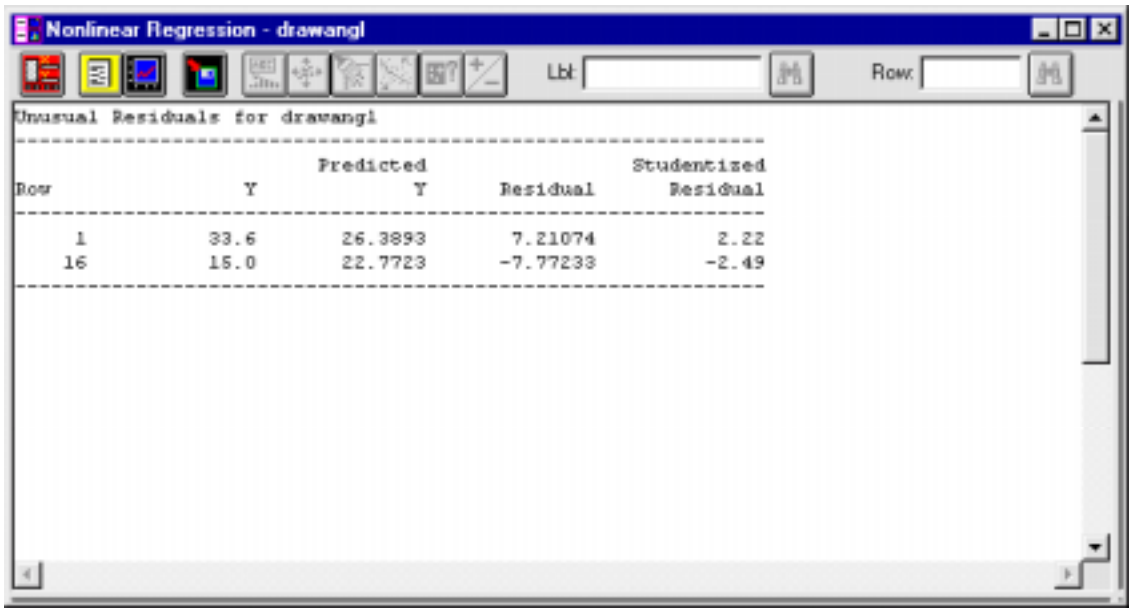


Figure 5-6. Unusual Residuals

Influential Points

The Influential Points option identifies all the observations that have leverage values greater than three times that of an average point, or that have unusually large DFITS or Cook's distance values (see Figure 5-7).

The leverage statistic helps determine the coefficients of the estimated model by measuring the amount of influence that can be attributed to each observation. The DFITS statistic measures the amount of change for each estimated coefficient if the observation is removed from the data. Cook's distance measures the distance between the estimated coefficients with and without each observation.

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model option displays a plot of the fitted function versus the chosen independent variable (see Figure 5-8). A line is drawn over the range of the chosen independent variable (X). If there is only one independent variable, the

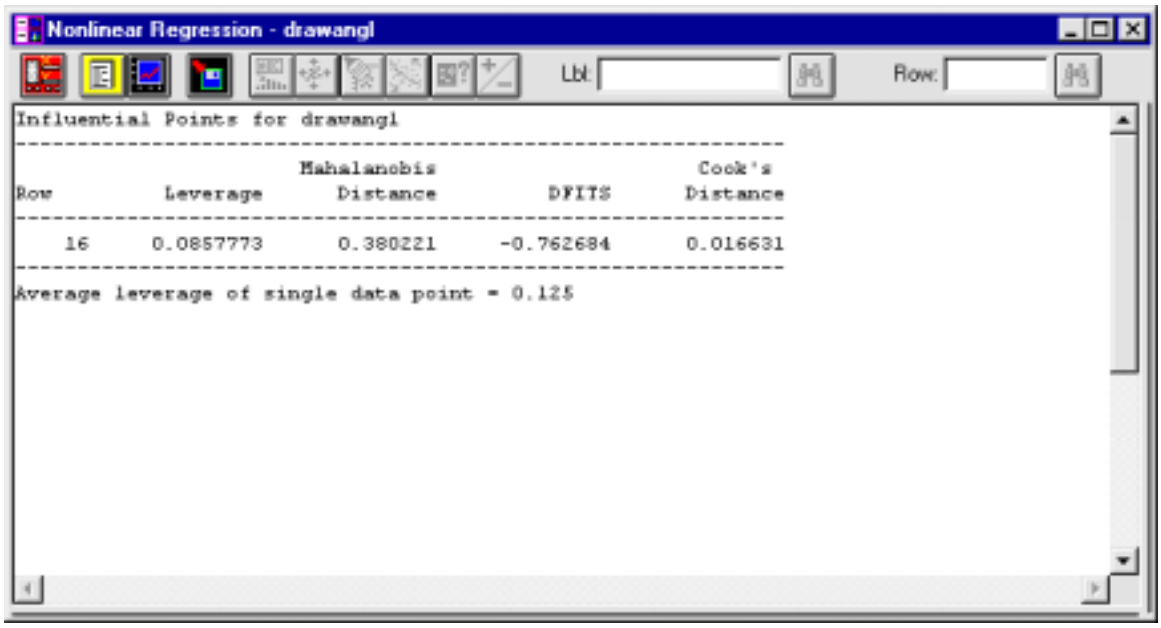


Figure 5-7. Influential Points

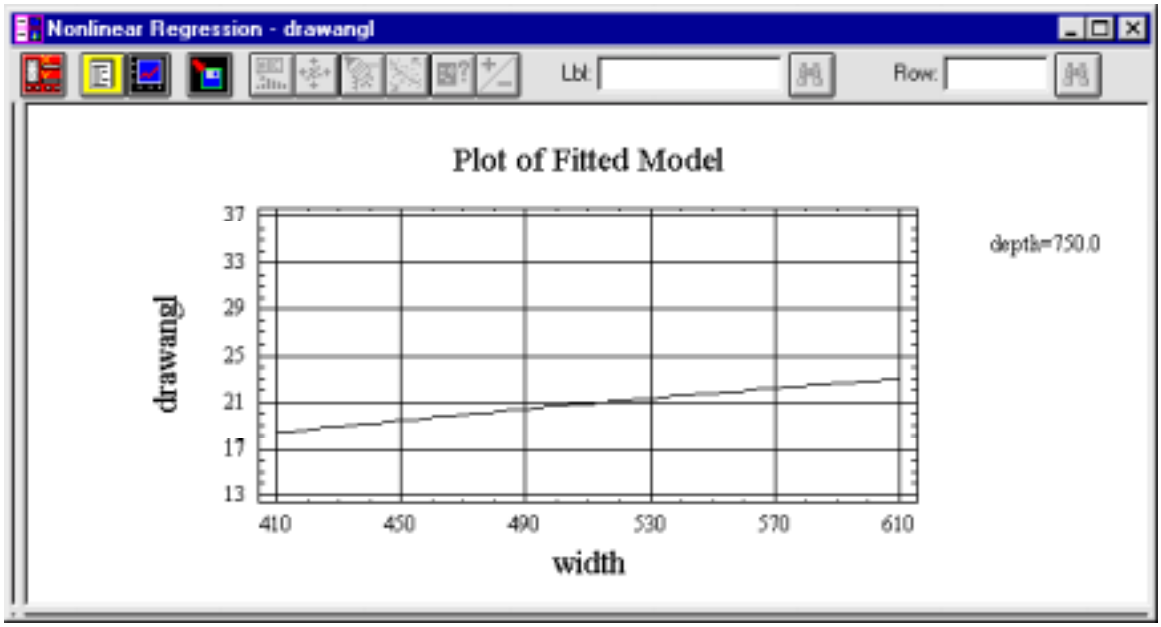


Figure 5-8. Plot of Fitted Model

program also plots the values of the points. Other independent variables are held constant at the levels shown in the legend on the plot.

Use the *Plot of Fitted Model Options* dialog box to change the levels for the other variables or to change the variable shown on the X-axis. The dialog box allows you to choose a variable to plot against the fitted model, to change the limits for that variable, and to choose a level at which to hold all of the other variables.

Response Surface Plots

Note: The two separate options for Response Surface Plots allow you to create two different plots and display the two views at once; for example, you can create both a Surface Plot and a Contour Plot of the same data.

The Response Surface Plots option allows you to create one of four different types of plot: Surface, Contour, Square, or Cube. The Square Plot is available if you have two or more variables in the function; the Cube Plot is available when you have at least three variables in the function.

Surface Plot

The Surface Plot option creates a three-dimensional model of the relationship between the estimated dependent variable as a function of the other variables (see Figure 5-9). The height of the surface represents the value of the estimated dependent variable.

When you choose this option, the Surface options portion of the dialog box becomes active, which allows you to customize the plot. You can also access the Plot of Fitted Model Options dialog box by clicking the Factors... command button. The Factors... command lets you choose another variable to plot against, to change the limits for that variable, and to enter a level at which the values of the other variables will be retained.

To view any of the plots from a different angle, click the Rotate button on the Analysis toolbar.

Contour Plot

The Contour Plot option creates a two-dimensional plot that traces the contours of the estimated dependent variable as a function of the other variables (see Figure 5-10).

Each contour (line) represents combinations of the independent variables, which have a value for the estimated dependent variable. You can predict the next value for the dependent variable by following the ridge of the contour.

When you choose this option, the Contours options portion of the dialog box becomes active, which allows you to customize the plot.

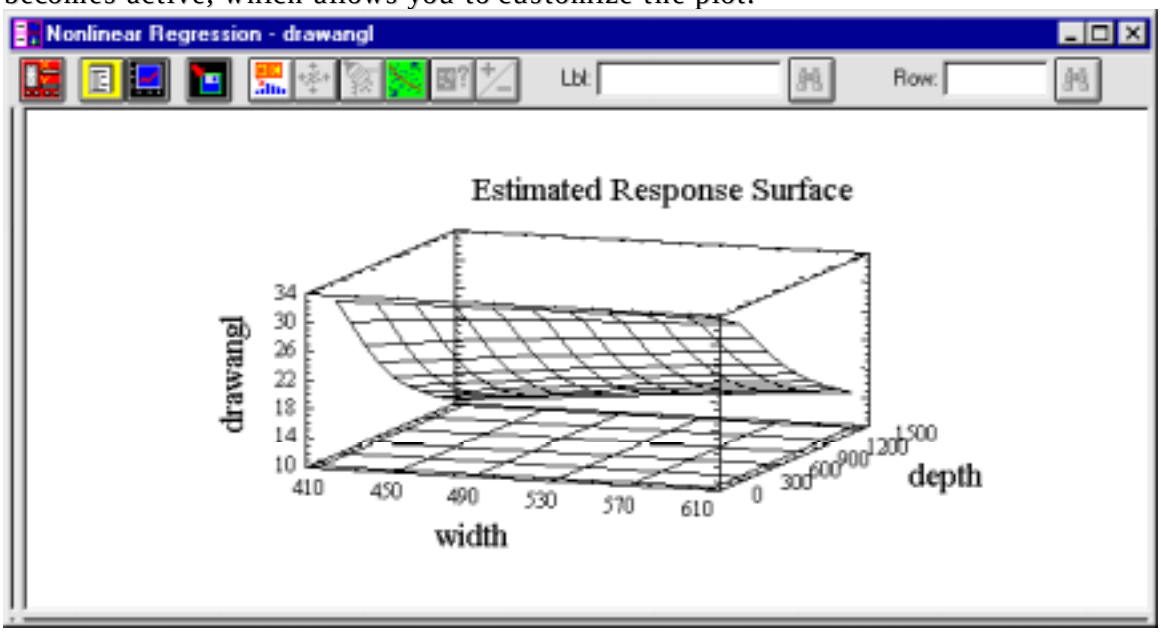


Figure 5-9. Surface Plot

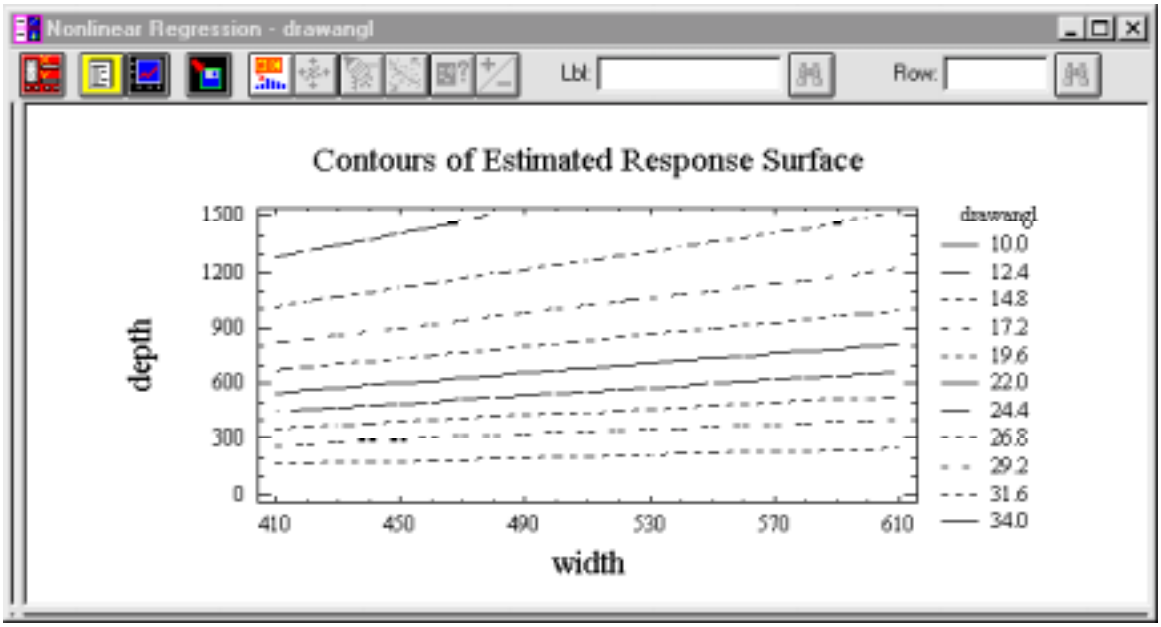


Figure 5-10. Contour Plot

Square Plot

The Square Plot option creates a plot of the observed values of the dependent variable versus the values predicted by the fitted model (see Figure 5-11). The program plots the values for the dependent variable for all combinations of the high and low values for each independent variable you select. Points close to the diagonal line are those that best predict the observed data.

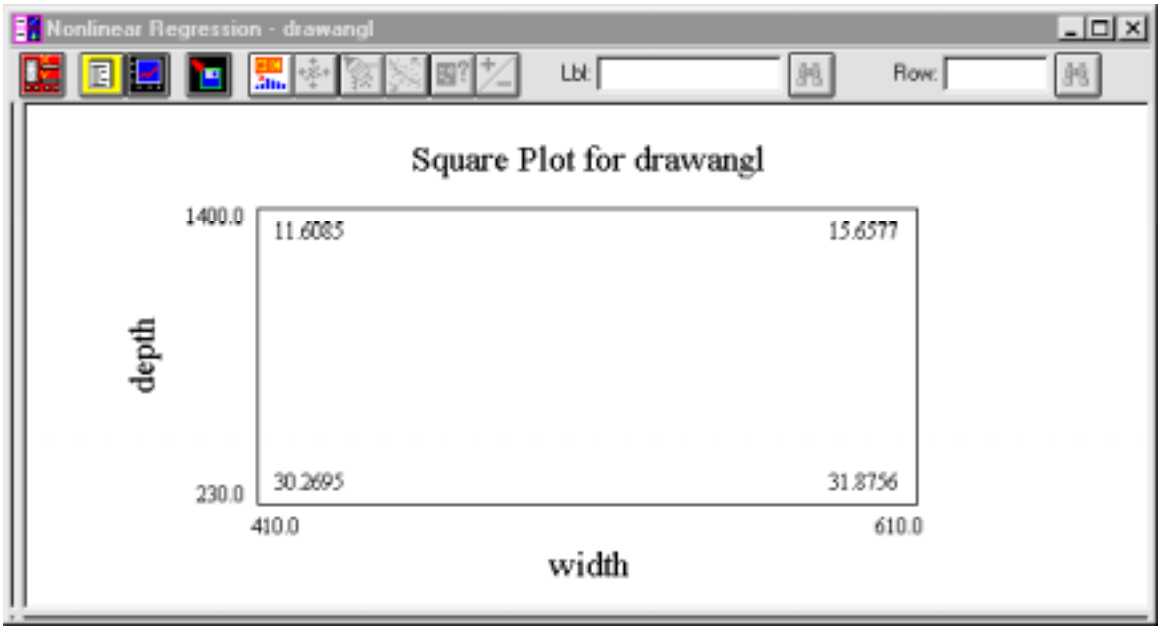


Figure 5-11. Square Plot

Cube Plot

The Cube Plot option creates a plot of the estimated effects for three factors. The program plots the values for the response variable for all combinations of the high and low values for each factor you select.

Use the *Response Plot Options* dialog box to choose the type of plot that will be created, to enter settings for Contour and Surface plots, and to access the Plot of Fitted Model Options dialog box, which lets you to choose another variable.

Observed versus Predicted

The Observed versus Predicted option displays a plot of the observed values (Y) versus the values predicted by the fitted model (see Figure 5-12). The plot includes a line with slope equal to one. Points close to the diagonal line are those that best predict the observed data. Use the plot to determine cases in which the variance is

not constant, which indicates that you should probably transform the values for the dependent variable.

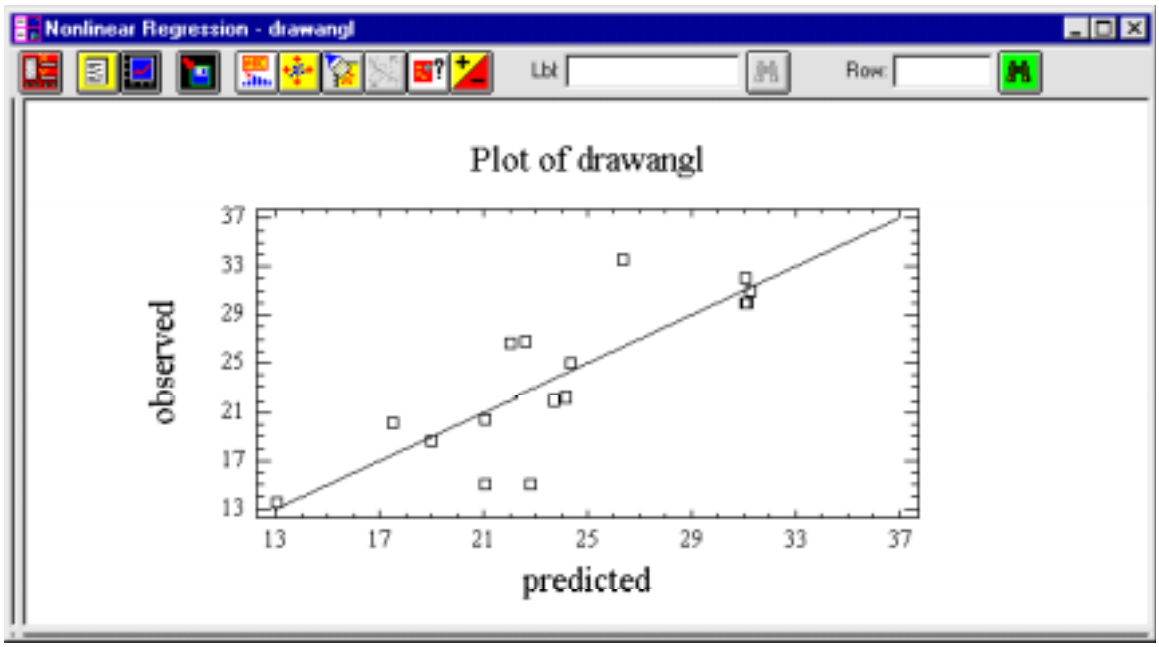


Figure 5-12. Observed versus Predicted Plot

Residual Plots

The Residual Plots option displays three different types of plots: Scatterplots, including Residual versus Predicted, Residual versus Row Number, and Residual versus X; as well as a Normal Probability Plot, and an Autocorrelation Function.Plot.

Use the *Residual Plots Options* dialog box to choose one of the plots, and, if applicable, its options.

Residual versus Predicted

The Residual versus Predicted scatterplot displays the residual or the studentized residual versus the predicted for the observed variable (Y) (see Figure 5-13). A nonrandom pattern indicates that the model does not adequately describe the observed data.

The plot is helpful in showing heteroscedasticity, an indication that the variability changes as the values of the dependent variable change.

Residual versus Row Number

The Residual versus Row Number scatterplot displays the residual or the studentized residual versus the row number (see Figure 5-14). The program

plots the residuals in the order that the observations appear in the dependent variable.

The plot is helpful in determining sequential correlations among the residuals. Any nonrandom pattern indicates serial correlation in the data, particularly if the row order corresponds to the order in which the data were collected.

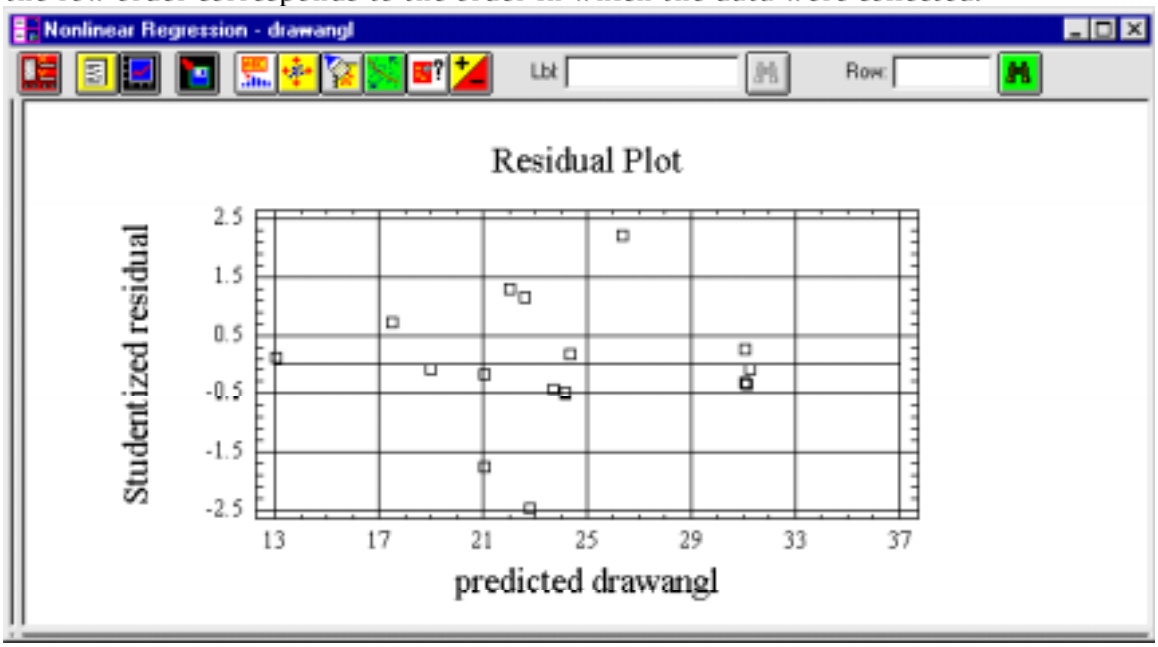


Figure 5-13. Residual versus Predicted Scatterplot

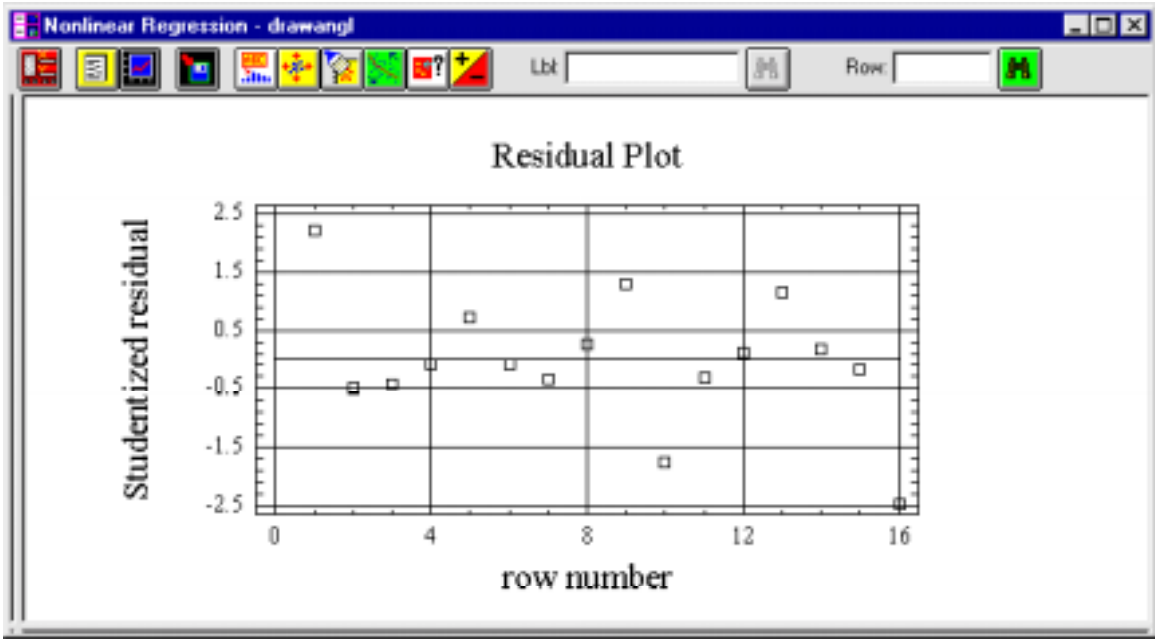


Figure 5-14. Residual versus Row Number Scatterplot

Residual versus X

The Residual versus X scatterplot displays the residual or studentized residual versus the independent variable (X). Use this plot to detect the nonlinear relationship between the dependent and independent variables (see Figure 5-15).

Nonrandom patterns indicate that the chosen model does not adequately the data. Any residual value outside the range of -3 to +3 may be an outlier.

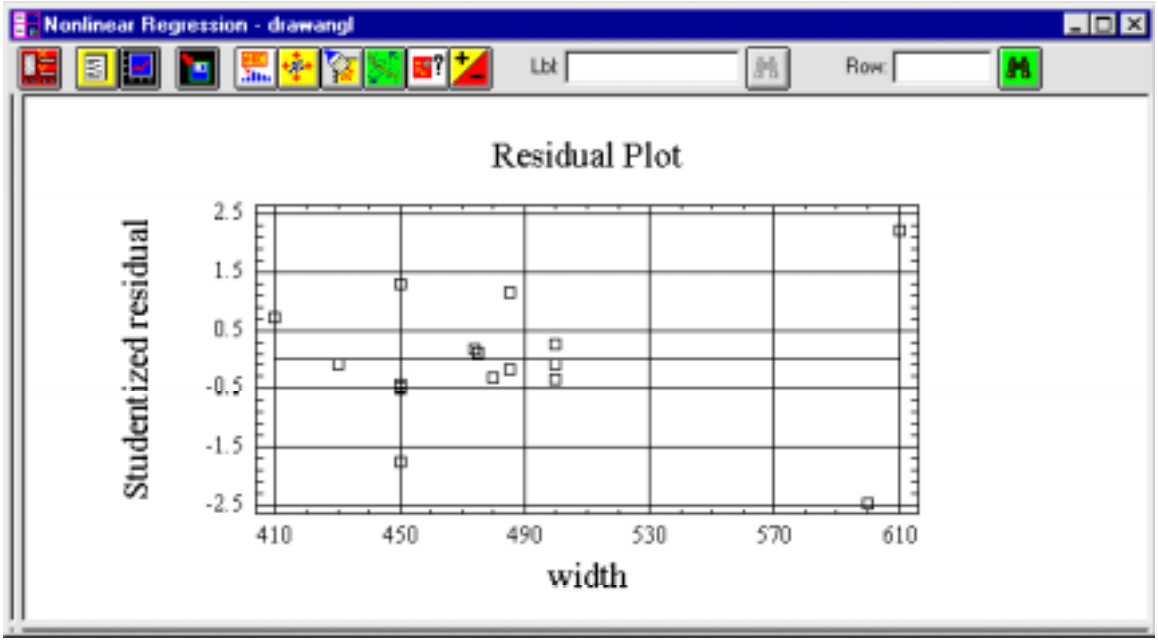


Figure 5-15. Residual versus X Scatterplot

Normal Probability Plot

The Normal Probability Plot option displays a plot to determine if the errors follow a normal distribution. (see Figure 5-16).

The plot consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution.

Autocorrelation Function Plot

The Autocorrelation Function plot option displays a graph of the estimated autocorrelations for the residuals (see Figure 5-17).

The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

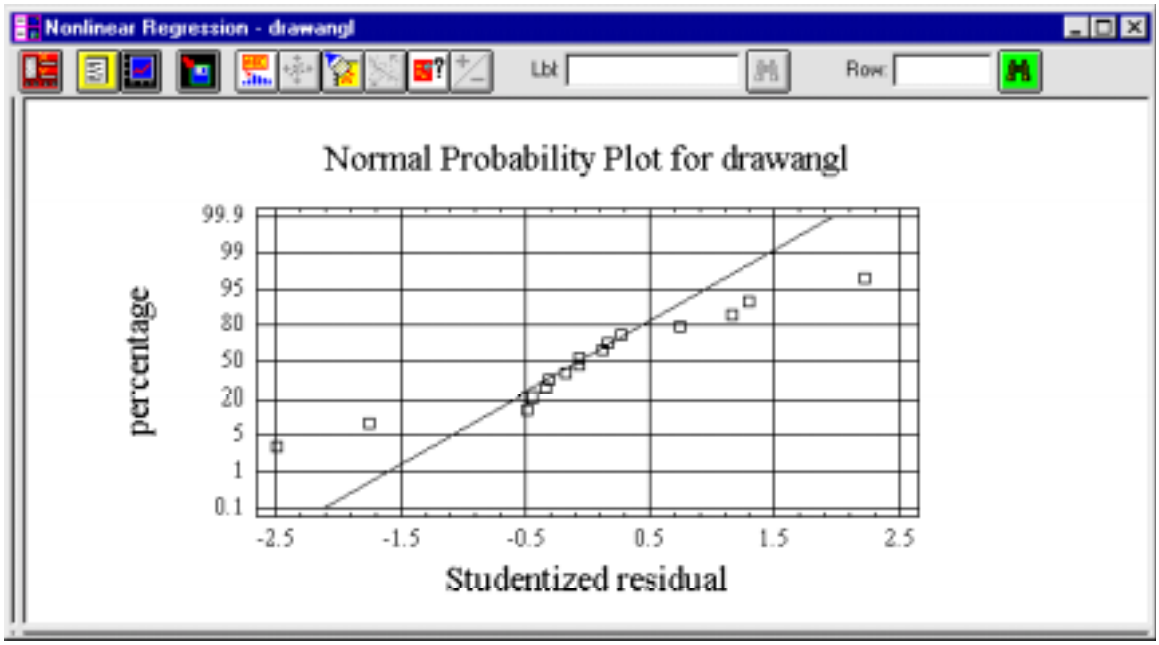


Figure 5-16. Normal Probability Plot

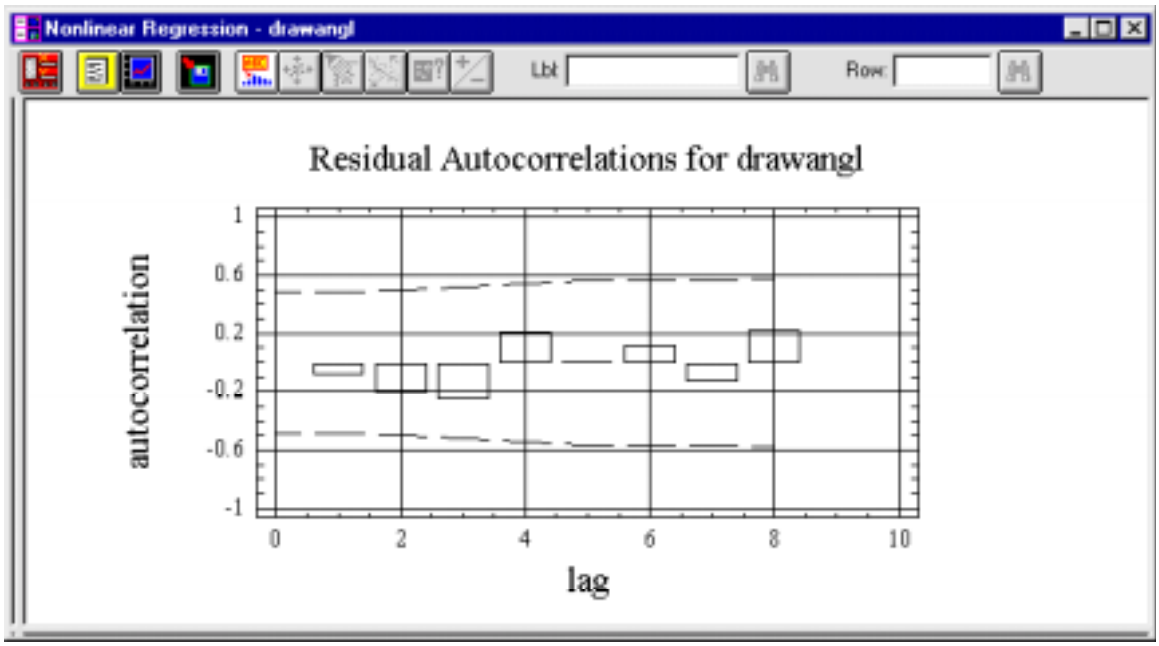


Figure 5-17. Autocorrelation Function Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 14 selections: Predicted Values, Standard Errors of Predictions, Lower Limits for Predictions, Upper Limits for Predictions, Standard Error of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, DFITS Statistics, Mahalanobis Distances, Coefficients, and Function.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Options Dialog Box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Cox, D. R. 1970. *Analysis of Binary Data*. London: Chapman and Hall.

Draper, N. R. and Smith, H. *Applied Regression Analysis*, second edition. New York: John Wiley & Sons.

Hartley, H. O. 1961. "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," *Technometrics*, **3**:269-80.

Marquardt, D. W. 1963. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal for the Society of Industrial and Applied Mathematics*, **11**:431-41.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago, Illinois: Richard D. Irwin, Inc.

Chapter 6

Using Ridge Regression

Background Information

Neter et al. (1996) define ridge regression as one of several methods that have been proposed to remedy multicollinearity problems by modifying the method of least squares to allow biased estimators of the regression coefficients. In multiple regression analysis, multicollinearity exists when two or more independent variables are highly correlated, which makes it difficult, if not impossible, to determine their separate effects on the dependent variable.

This method falls into the category of biased estimation techniques; that is, though ordinary least squares give unbiased and minimum variance regression estimates, there is no upper bound on the variance of the estimators, and when multicollinearity occurs it may produce large variances. When the variance is reduced using biased estimation, a noticeable increase in stability occurs in the regression coefficients. Reducing variance may provide benefits that offset any loss suffered due to using biased estimates. An estimator that has only a small bias, and that is substantially more precise, is the preferred estimator because it more likely will be closest to the true value of the parameter (Myers, 1990).

Small changes in the data used for fitting the regression usually do not affect the ridge regression estimates, which tend to be quite stable. However, under the same conditions, and when the predictor variables are highly multicollinear, the least squares estimates may be very unstable. Neter et al. (1996) notes that predictions of new observations made from ridge-estimated regression functions tend to be more precise than predictions made from ordinary least squares regression functions when the explanatory variables are correlated and the new observations follow the same multicollinearity pattern. The advantage in making predictions using ridge regression becomes even greater when intercorrelations among the explanatory variables are high.

Analyzing the ridge trace is helpful when you need to reduce the number of potential explanatory variables in exploratory observational studies. When the ridge trace is unstable, variables whose coefficients tend toward the value of zero are dropped from the study as are variables whose ridge trace is stable but at a very small value. Variables that have unstable ridge traces that do not tend toward zero are also likely candidates for dropping.

Ridge Regression Analysis in STATGRAPHICS *Plus*

The Ridge Regression Analysis in STATGRAPHICS *Plus* allows you to modify the least squares method to estimate the multiple regression model using slightly modified normal equations. The estimates that result are often more precise than those that result when you use ordinary least squares and, in the case of correlated independent variables, may be closer to the "true value" of the parameter.

The analysis provides tools for selecting a value for theta (the ridge parameter), which controls the extent of the bias that is introduced. When theta equals zero, the estimates that result are the same as those for least squares. As theta increases, but usually remains less than one, bias increases; however, the coefficients usually become more precise. Ordinarily, an appropriate value for theta is the smallest value that occurs before the estimates begin to slowly change.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... RIDGE REGRESSION... from the Menu bar to display the Ridge Regression Analysis dialog box (see Figure 6-1).

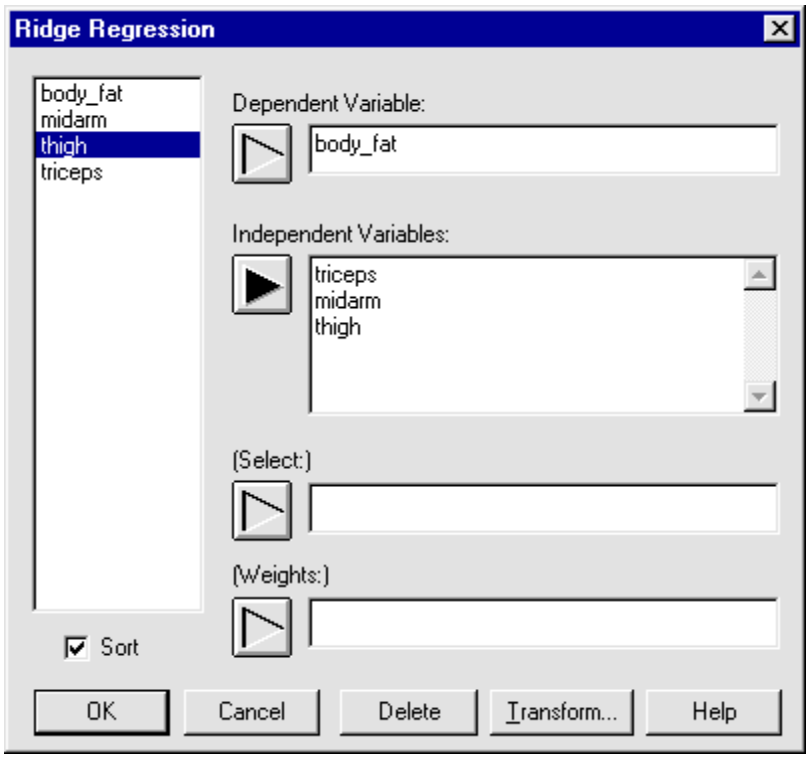


Figure 6-1. Ridge Regression Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option shows the current value of the ridge parameter (see Figure 6-2). By allowing for a small amount of bias, the precision of the estimates is greatly increased. The summary shows the current value of the ridge parameter to be 0.0, which is equivalent to ordinary least squares.

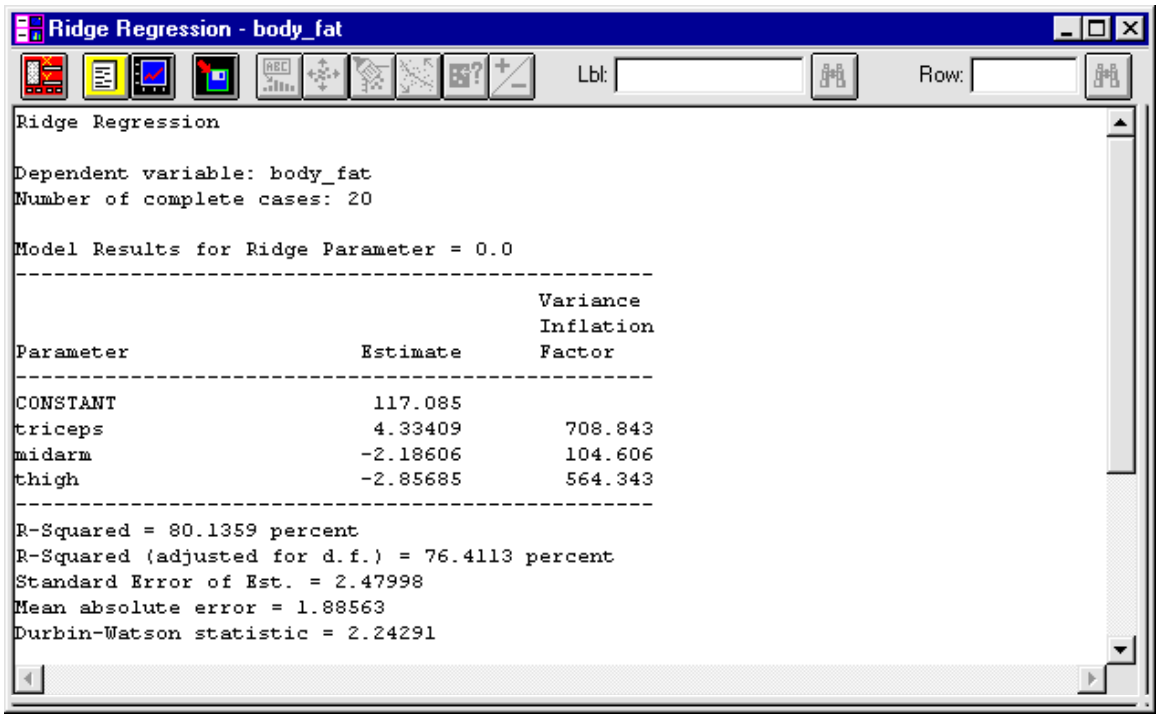


Figure 6-2. Analysis Summary

Other results include the following:

- The R-Squared statistic, which indicates the percentage of variability the dependent variable accounts for as it was fitted in the model.
- The Adjusted R-Squared statistic, which is more suitable for comparing models that have different numbers of independent variables, indicates the percentage of variability represented by the model.
- The Standard Error of the Estimate, which shows the value for the standard deviation of the residuals.
- The Mean Absolute Error (MAE) statistic, which is the average absolute value of the residuals.

- The Durbin-Watson statistic, which tests the order of the residuals as they occur in the file. If the Durbin-Watson statistic is greater than 1.4, there is probably no serious autocorrelation in the residuals.

If you used the optional Select text box on the Ridge Regression Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use the methods discussed in the topic, "Overview of the Model-Building Process," in Online Help.

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.

- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the data, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.

- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Use the *Ridge Regression Options* dialog box to change the value for the ridge parameter. The value is usually between 0 and 1.0. To determine a good value, examine the standardized regression coefficients or the variance inflation factors.

Regression Coefficients

The Regression Coefficients option displays the estimated coefficients for values of the ridge parameter that are between 0.0 and 0.1 for variables that are unscaled and uncentered; that is, the variables are expressed in their original natural units (see Figure 6-3).

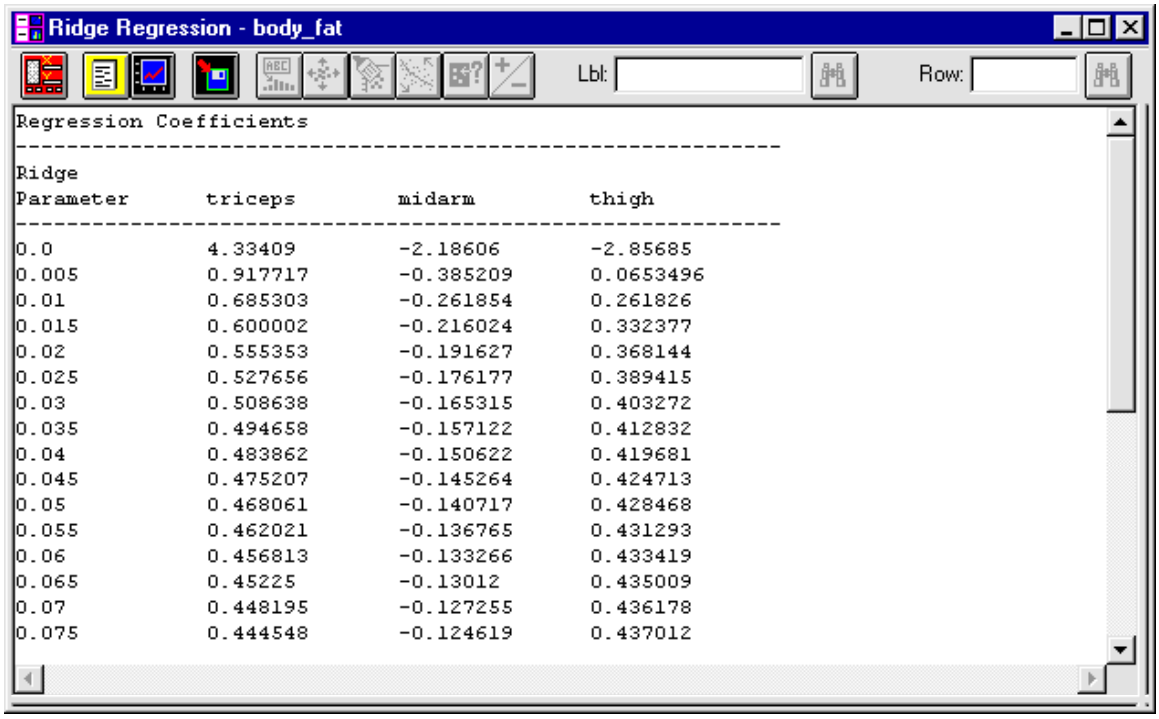


Figure 6-3. Regression Coefficients

As the ridge parameter increases from 0.0, the coefficients at first change dramatically, then become relatively stable. Sometimes it is reasonable to use the natural units for the variables to find a good ridge parameter. However, to eliminate problems due to extreme differences in the magnitudes of the values in the different variables, use the standardized coefficients.

A good value for the ridge parameter is the smallest value that occurs before the estimates slowly change. This technique is subjective, but the Ridge Trace graphical option can help you make a good choice.

Use the *Ridge Regression Options* dialog box to change the range of the ridge parameters that will display.

Standardized Regression Coefficients

The Standardized Regression Coefficients option displays the standardized coefficients for values of the ridge parameter that are between 0.0 and 0.1 (see Figure 6-4). These are coefficients of the regression model when the variables are expressed in standardized form.

As the ridge parameter increases from 0.0, the coefficients at first change dramatically, then become relatively stable. A good value for the ridge parameter

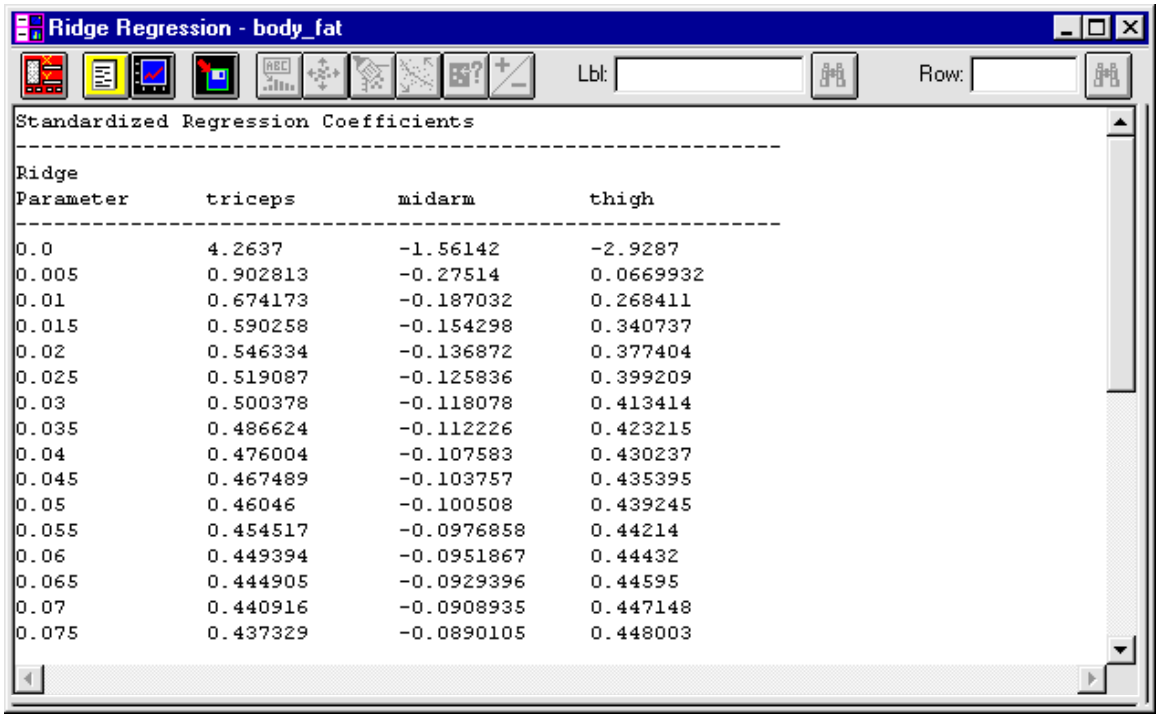


Figure 6-4. Standardized Regression Coefficients

is the smallest value that occurs before the estimates slowly change. This technique is subjective, but the Ridge Trace graphical option can help you make a good choice.

Use the Ridge Regression Options dialog box to change the range of the ridge parameters that will display.

Variance Inflation Factors

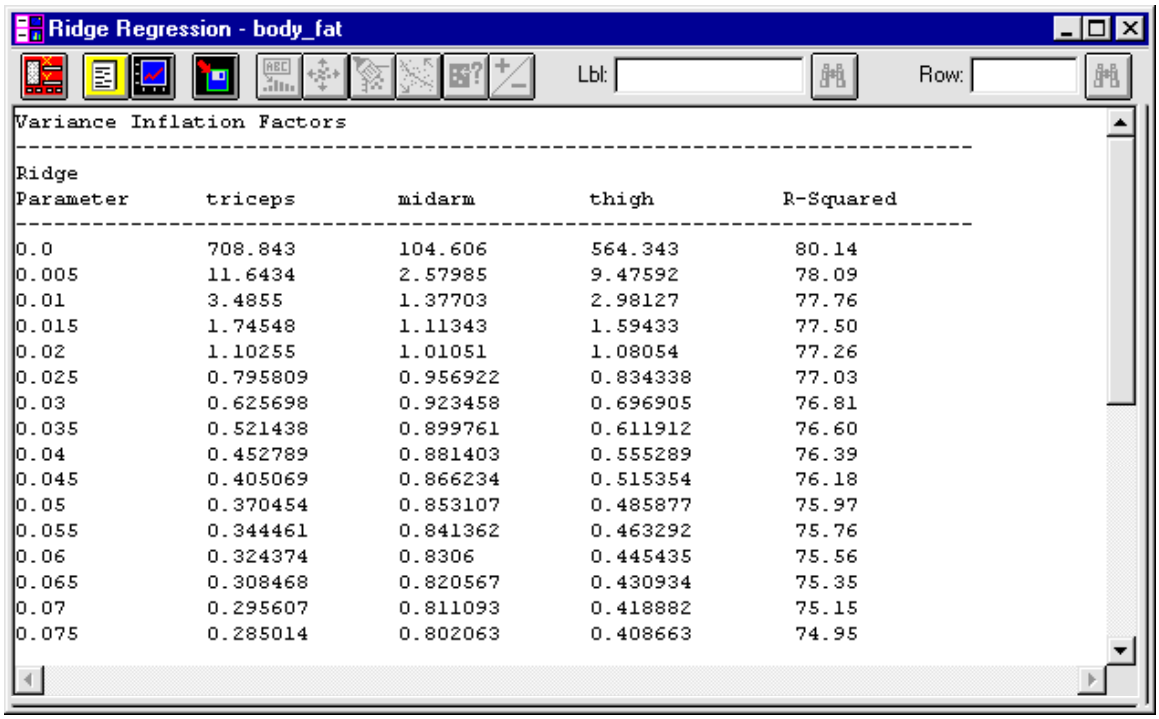
The Variance Inflation Factors (VIF) option displays the VIFs and the R-Squared statistics for each value of the ridge parameter (see Figure 6-5).

The VIFs measure the amount by which the variance of the estimated coefficients would be inflated when compared with the ideal case — having all the independent variables uncorrelated. As the ridge parameter increases from 0.0, the VIFs at first change dramatically, then become relatively stable. This technique is subjective, but the Variance Inflation Factors graphical option can help you make a good choice.

Reports

The Reports option displays information about the results generated for the regression using the current ridge parameter (see Figure 6-6). Depending on your selections on the Reports Options dialog box, the table includes the observed value

for the dependent variable (if any), the predicted value for the dependent variable using the fitted model, and the residual (observed value minus the predicted value).



Variance Inflation Factors

Ridge Parameter	triceps	midarm	thigh	R-Squared
0.0	708.843	104.606	564.343	80.14
0.005	11.6434	2.57985	9.47592	78.09
0.01	3.4855	1.37703	2.98127	77.76
0.015	1.74548	1.11343	1.59433	77.50
0.02	1.10255	1.01051	1.08054	77.26
0.025	0.795809	0.956922	0.834338	77.03
0.03	0.625698	0.923458	0.696905	76.81
0.035	0.521438	0.899761	0.611912	76.60
0.04	0.452789	0.881403	0.555289	76.39
0.045	0.405069	0.866234	0.515354	76.18
0.05	0.370454	0.853107	0.485877	75.97
0.055	0.344461	0.841362	0.463292	75.76
0.06	0.324374	0.8306	0.445435	75.56
0.065	0.308468	0.820567	0.430934	75.35
0.07	0.295607	0.811093	0.418882	75.15
0.075	0.285014	0.802063	0.408663	74.95

Figure 6-5. Variance Inflation Factors

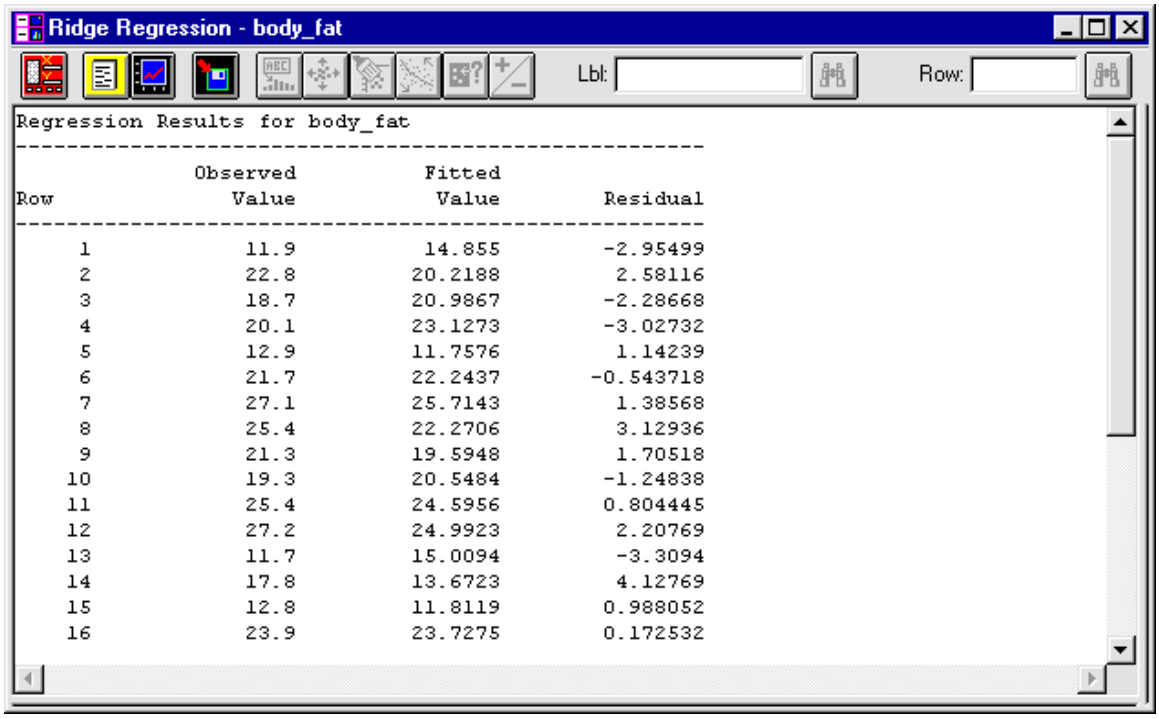


Figure 6-6. Reports

Each item corresponds to the values of the independent variables in a specific row of the file.

To create forecasts (predictions) for additional combinations of the variables, add additional rows to the bottom of the file. In each new row, enter values for the independent variables but leave the cell for the dependent variable empty. The program adds the predicted values for the new rows to the table, but leaves the model unchanged.

Use the *Reports Options* dialog box to choose the values that will be included in the report.

Graphical Options

Ridge Trace

The Ridge Trace option creates a plot of the values for the standardized or unstandardized coefficients versus the values for the ridge parameter (see Figure 6-7). As the ridge parameter increases from 0.0, the coefficients at first change dramatically, then become relatively stable.

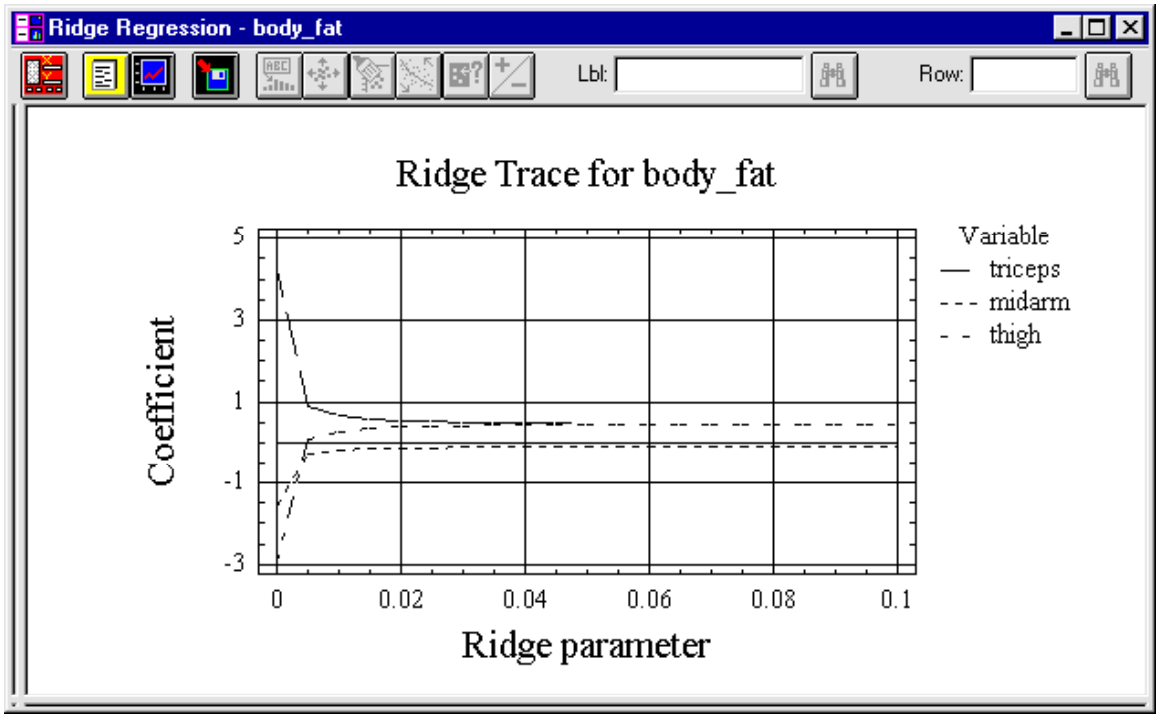


Figure 6-7. Ridge Trace

Use the *Ridge Trace Options* dialog box to choose the type of coefficients that will appear on the plot.

Variance Inflation Factors

The Variance Inflation Factors option creates a plot of the values for the standardized or unstandardized regression coefficients versus the values for the ridge parameter (see Figure 6-8). As the value for the ridge parameter increases from 0.0, the coefficients at first change dramatically, then become relatively stable.

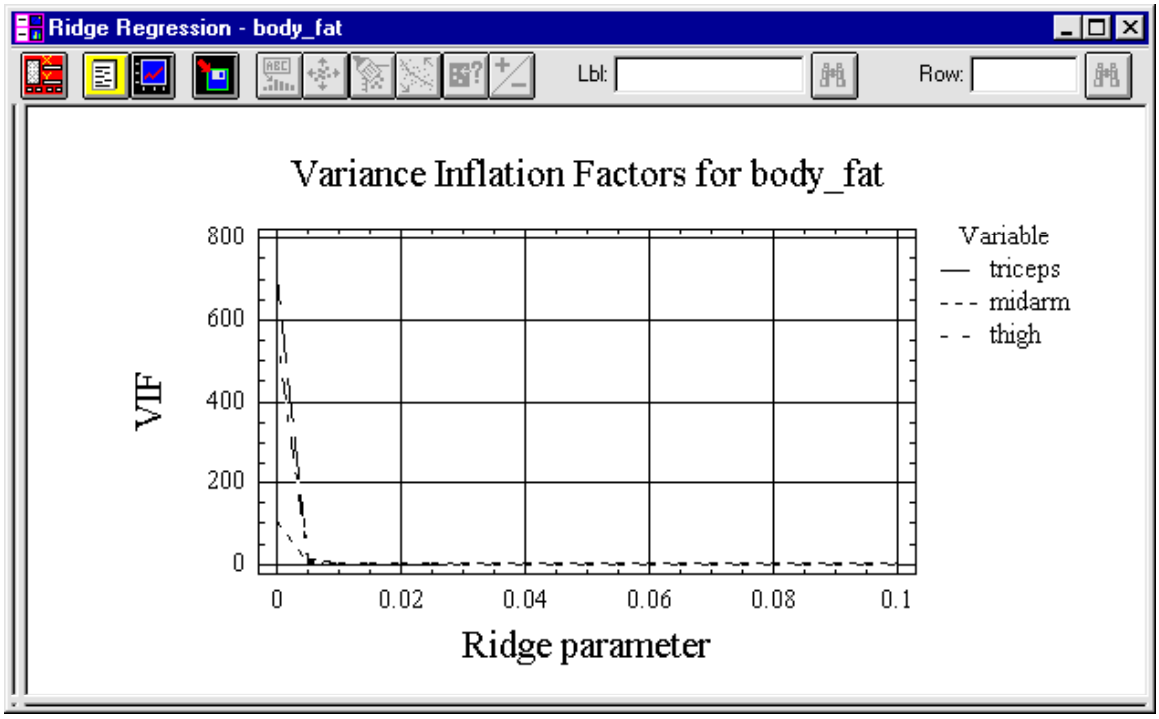


Figure 6-8. Variance Inflation Factors Plot

Observed versus Predicted

The Observed versus Predicted option creates a plot of the observed values for the dependent variable versus the values predicted by the fitted model for the current ridge parameter (see Figure 6-9). The closer the points lie to the diagonal line, the better the model is at predicting the observed data.

Residual Plots

The Residual Plots option displays three different types of plots: Scatterplots, including Residual versus Predicted, Residual versus Row Number, and Residual versus X; as well as a Normal Probability Plot, and an Autocorrelation Function Plot.

Use the *Residual Plots Options* dialog box to choose one of the plots, and, if applicable, its options.

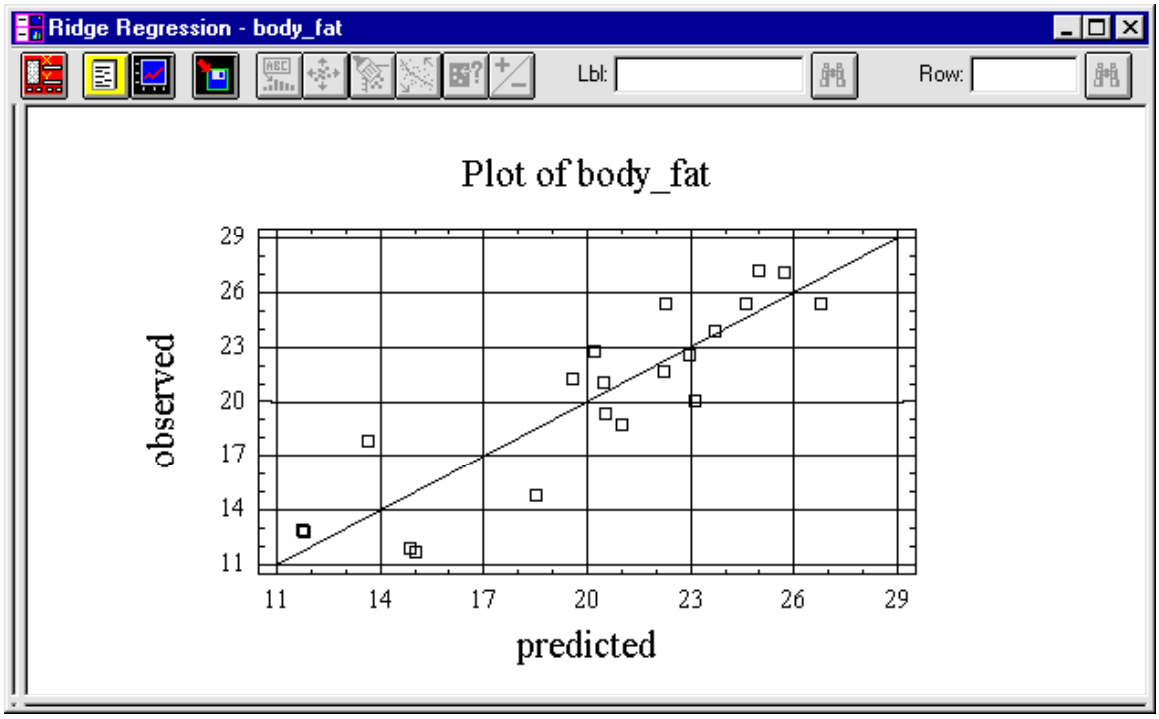


Figure 6-9. Observed versus Predicted Plot

Residual versus Predicted

The Residual versus Predicted scatterplot displays the residual or the studentized residual versus the predicted values for the observed variable (Y) (see Figure 6-10). A nonrandom pattern indicates that the model does not adequately describe the observed data.

The plot is helpful in showing heteroscedasticity; an indication that the variability changes in the values of the dependent variable change.

Residual versus Row Number

The Residual versus Row Number scatterplot displays the residual or the studentized residual versus the row number (see Figure 6-11). The program plots the residuals in the order that the observations appear in the dependent variable.

The plot is helpful in determining sequential correlations among the residuals. Any nonrandom pattern indicates serial correlation in the data, particularly if the row number corresponds to the order in which the data were collected.

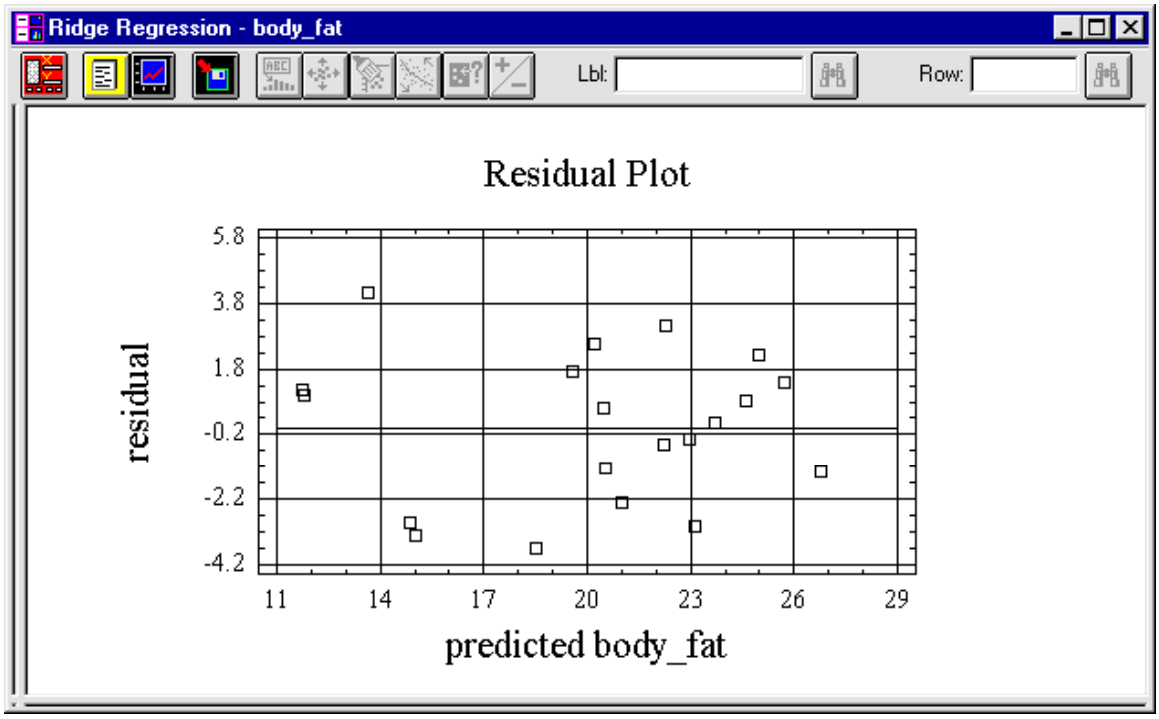


Figure 6-10. Residual versus Predicted Scatterplot

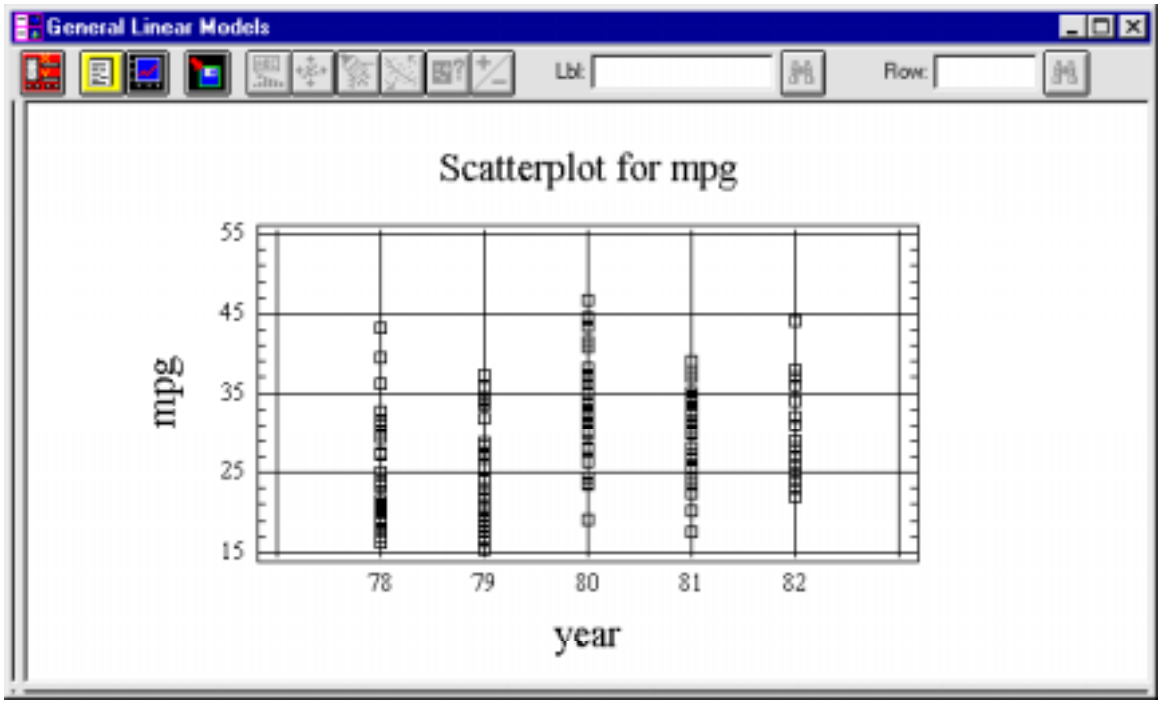


Figure 6-11. Residual versus Row Number Scatterplot

Residual versus X

The Residual versus X scatterplot displays the residual versus the independent variable (X) (see Figure 6-12). Use this plot to detect the nonlinear relationship between the dependent and independent variables.

Nonrandom patterns indicate that the chosen model does not adequately describe the data.

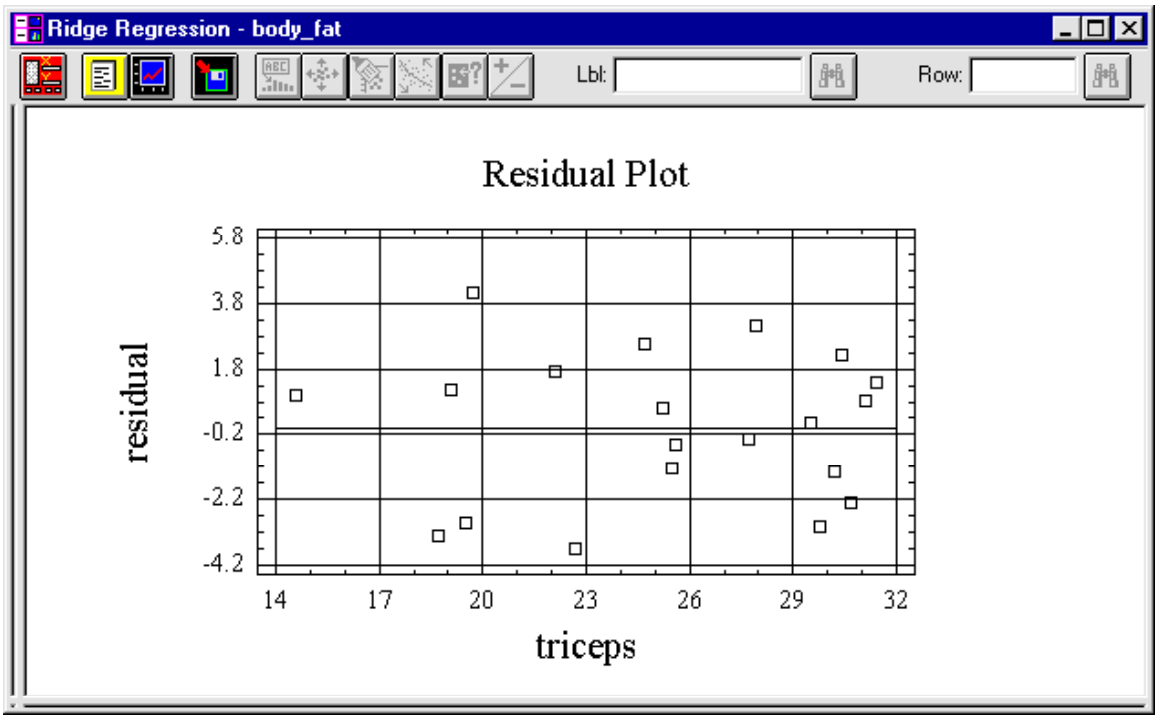


Figure 6-12. Residual versus X Scatterplot

Normal Probability Plot

The Normal Probability Plot option displays a plot that determines if the errors follow a normal distribution (see Figure 6-13).

The plot consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution.

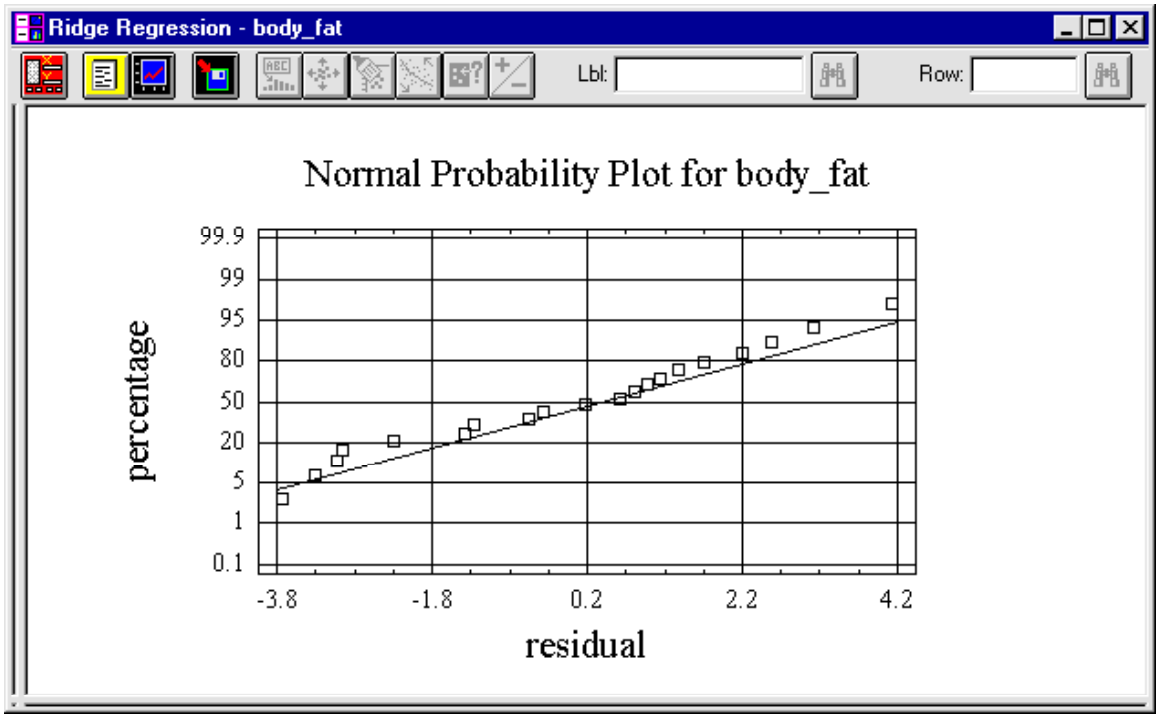


Figure 6-13. Normal Probability Plot

Autocorrelation Function

The Autocorrelation Function option displays a graph of the estimated autocorrelations for the residuals (see Figure 6-14).

The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are three selections: Predicted Values, Residuals, and Coefficients.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

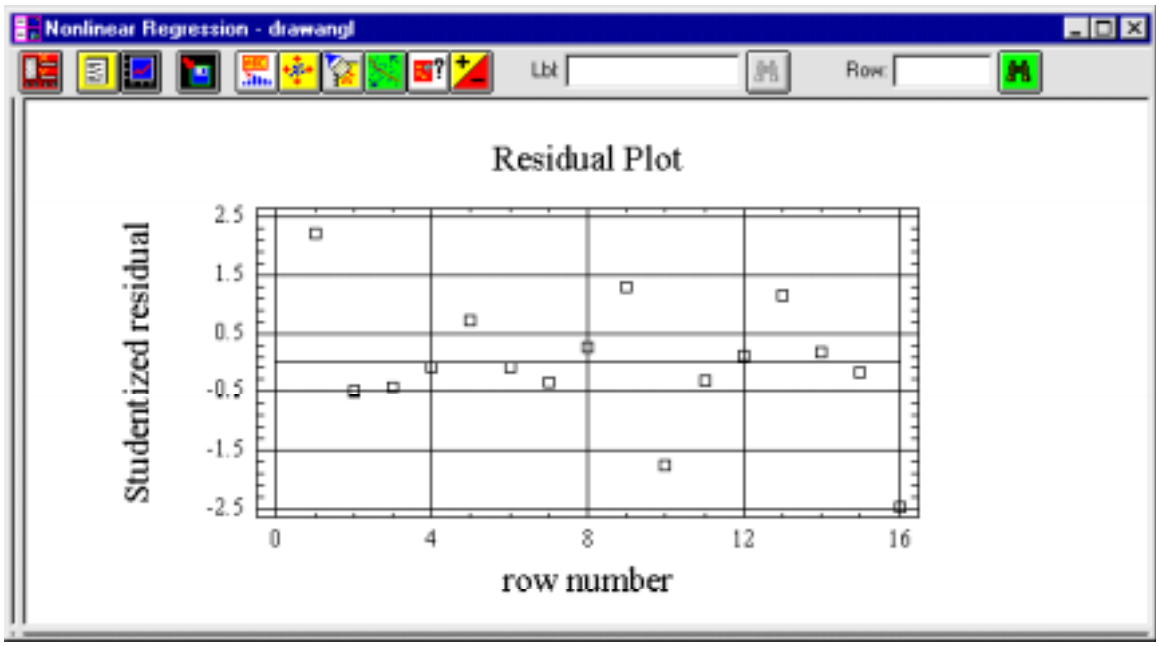


Figure 6-14. Autocorrelation Function Plot

References

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: John Wiley & Sons.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*. Belmont, California: Duxbury Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago: Richard D. Irwin, Inc.

Vogt, W. P. 1993. *Dictionary of Statistics and Methodology*. New York: Sage Publications.

Chapter 7

Using Logistic Regression

Background Information

Logistic regression analysis allows analysts to estimate multiple regression models when the response being modeled is dichotomous and can be scored 0,1; that is, the outcome must be one of two choices. Analysts use logistic regression when they want to model something that can be expressed as Event/Nonevent or something that has two possible outcomes, such as the examples given by Neter et al. (1996) as financial status of a firm (sound status, headed toward insolvency) or blood pressure status (high blood pressure, not high blood pressure).

The logistic transformation in this manual provides the basis for the linear logistic model. It is ideal in that it restricts predictions to the interval (0,1), which is suitable for proportions and can be easily transformed into a linear form. The function takes on a monotonically increasing or decreasing S shape that frequently occurs in Dose/Response and other survival rate studies. It is an increasingly popular alternative to discriminant analysis because it relaxes some of the assumptions.

Logistic Regression in STATGRAPHICS *Plus*

STATGRAPHICS *Plus* allows you to enter dichotomous response data in one of two forms:

- the dependent variable can consist of individual observations of either a 0 or a 1, representing the actual outcome for each individual case

- the dependent variable can contain probabilities between 0 and 1 that are the proportions of successes for groups of observations whose sizes can vary.

The independent or explanatory variables can be either categorical or continuous. The program creates indicator variables for categorical factors that can simplify the use of the model when making predictions for future cases. There are several statistics and graphs that help to assess the accuracy and usefulness of the model.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... LOGISTIC REGRESSION... from the Menu bar to display the Logistic Regression Analysis dialog box (see Figure 7-1).

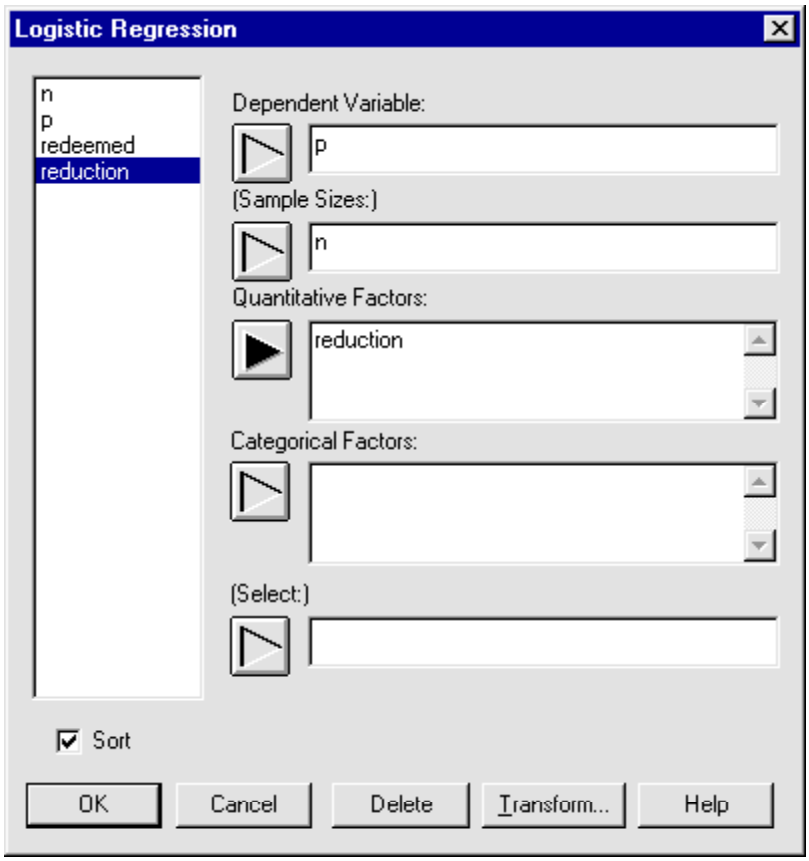


Figure 7-1. Logistic Regression Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option displays the results of fitting a linear logistic regression model that describes the relationship between the dependent and independent variables (see Figure 7-2).

The results include estimates for each of the coefficients, approximate standard errors, and the estimated odds ratios. If the *p*-value is less than 0.01, there is a statistically significant relationship among the variables. If the *p*-value for the residuals is greater than or equal to 0.10, it indicates that the model is not significantly worse than the best possible model for the data you are currently using.

If the model was fit using weighted least squares, the results include: *t*-tests with associated *p*-values; an ANOVA table; summary statistics, such as R-Squared; and the Type III sums of squares for each of the factors.

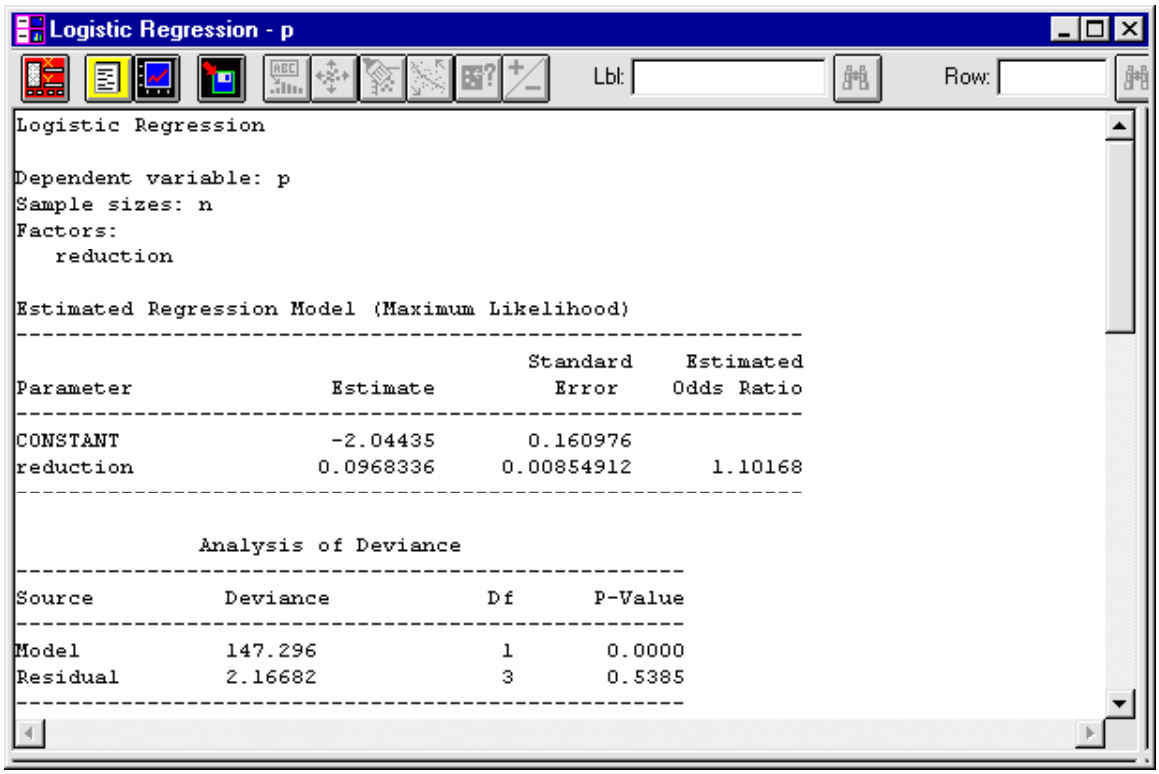


Figure 7-2. Analysis Summary

If the model was fit using maximum likelihood, the results include an analysis of model deviance and the percentage of deviance for which the model accounts. This value is similar to the R-Squared statistic. The results also include values for tests of likelihood ratio for each of the factors.

If you used the optional Select text box on the Logistic Regression Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use the methods discussed in the topic, “Overview of the Model-Building Process," in Online Help.

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.
- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the data, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.

- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.

- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Use the *Logistic Regression Options* dialog box to choose the estimation method, to enter a value for the smallest proportion (for weighted least squares), to choose the type of model to be fit, to choose a selection procedure, to enter values for *p* to enter or remove, to enter a value for the maximum number of steps, and to indicate how the results will be displayed (see Figure 7-3).

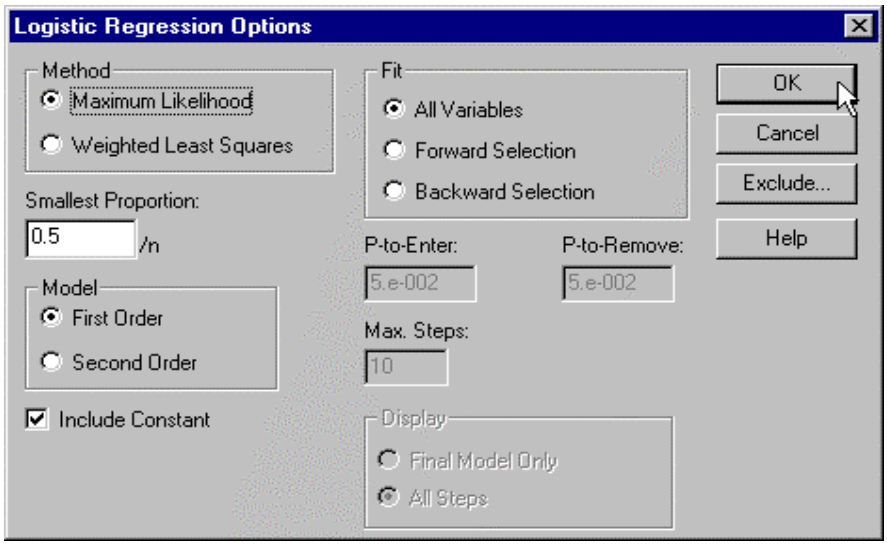


Figure 7-3. Logistic Regression Options Dialog Box.

Goodness-of-Fit

The Goodness-of-Fit option displays a table that divides the logistic scale into intervals, each of which contains approximately the same number of observations (see Figure 7-4).

The program compares the observed versus the predicted number of True and False observations in each interval of the observed data with those predicted by the model to determine if the function adequately fits the observed data. Small *p*-values indicate a significant lack of fit.

Use the *Goodness-of-Fit Options* dialog box to enter a value for the number of classes into which the data will be grouped.

Confidence Intervals

The Confidence Intervals option displays confidence intervals for the coefficients in the model and the odds ratios using a confidence level of 95 percent (see Figure 7-5). The confidence intervals illustrate the preciseness with which the coefficients were estimated, given the amount of available data and the noise present. The odds ratios equal the inverse natural logarithm of the coefficient and show the proportional change in the response variable as the independent variable is increased by one unit.

Use the *Confidence Intervals Options* dialog box to enter a number that will be used to calculate the confidence intervals for the mean and standard deviation.

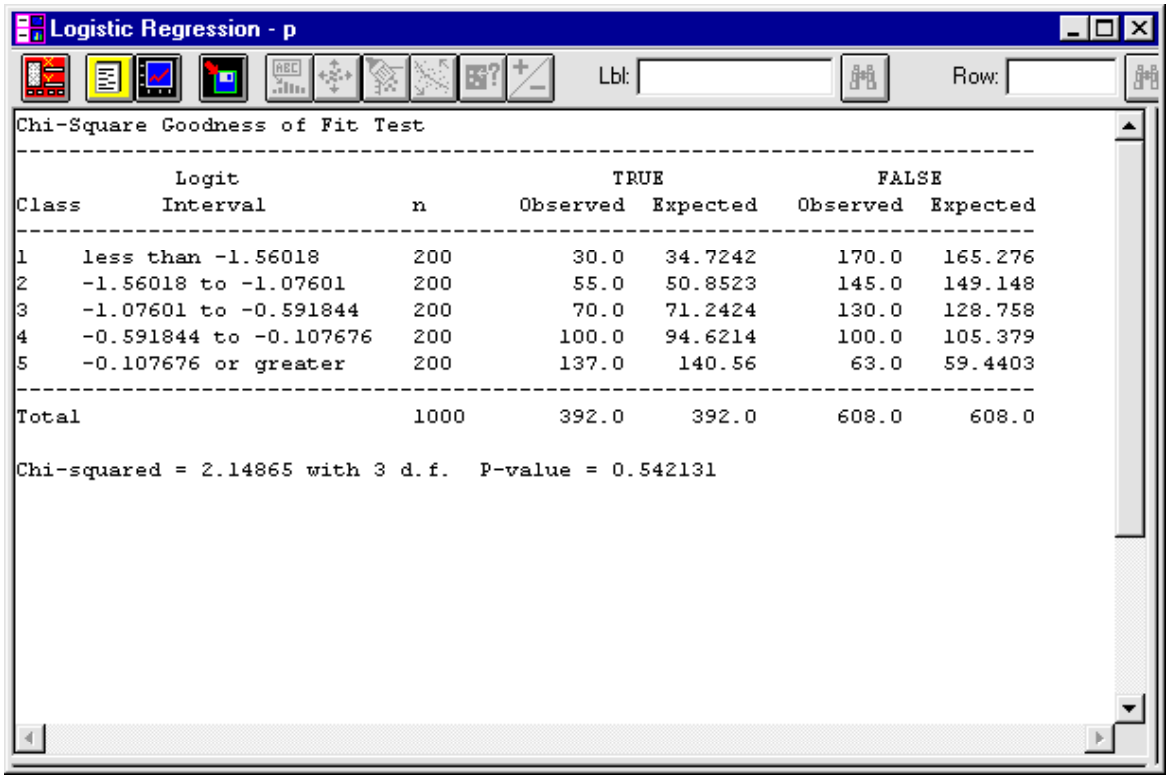


Figure 7-4. Goodness-of-Fit Results

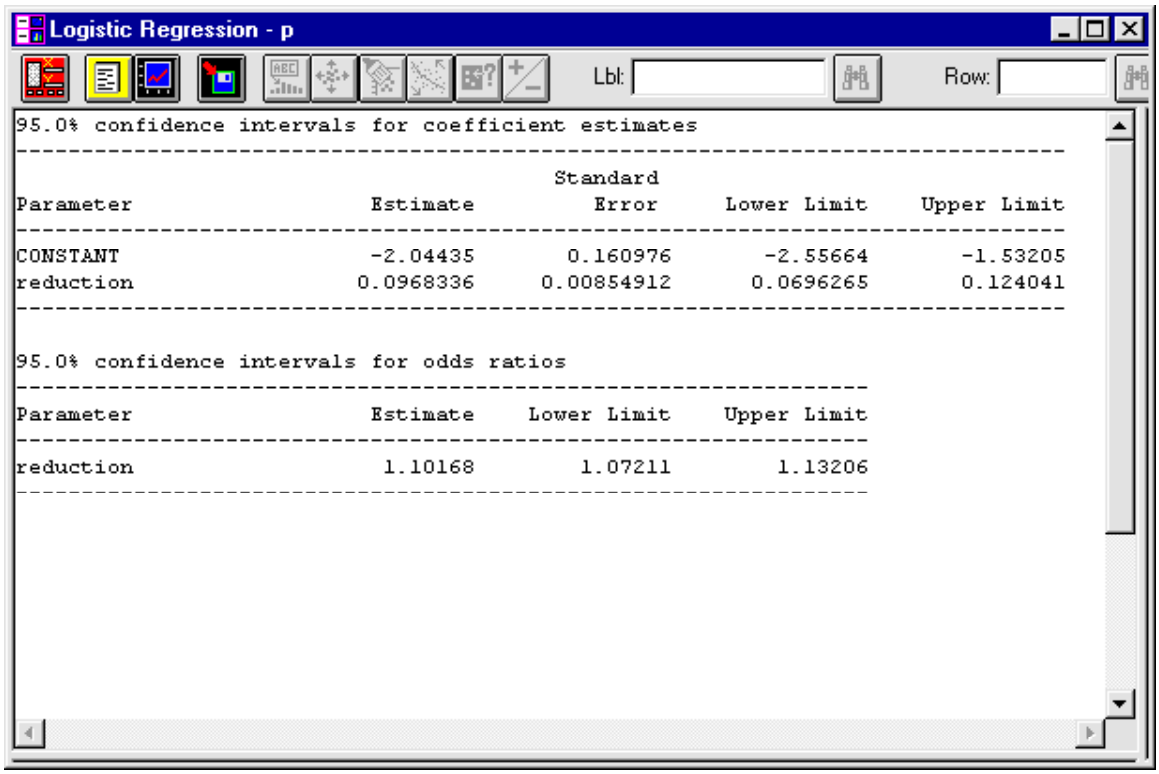


Figure 7-5. Confidence Intervals

Correlation Matrix

The Correlation Matrix option displays a table of the estimated correlations between the coefficients in the fitted model (see Figure 7-6). The correlations are helpful in detecting the presence of serious multicollinearity.

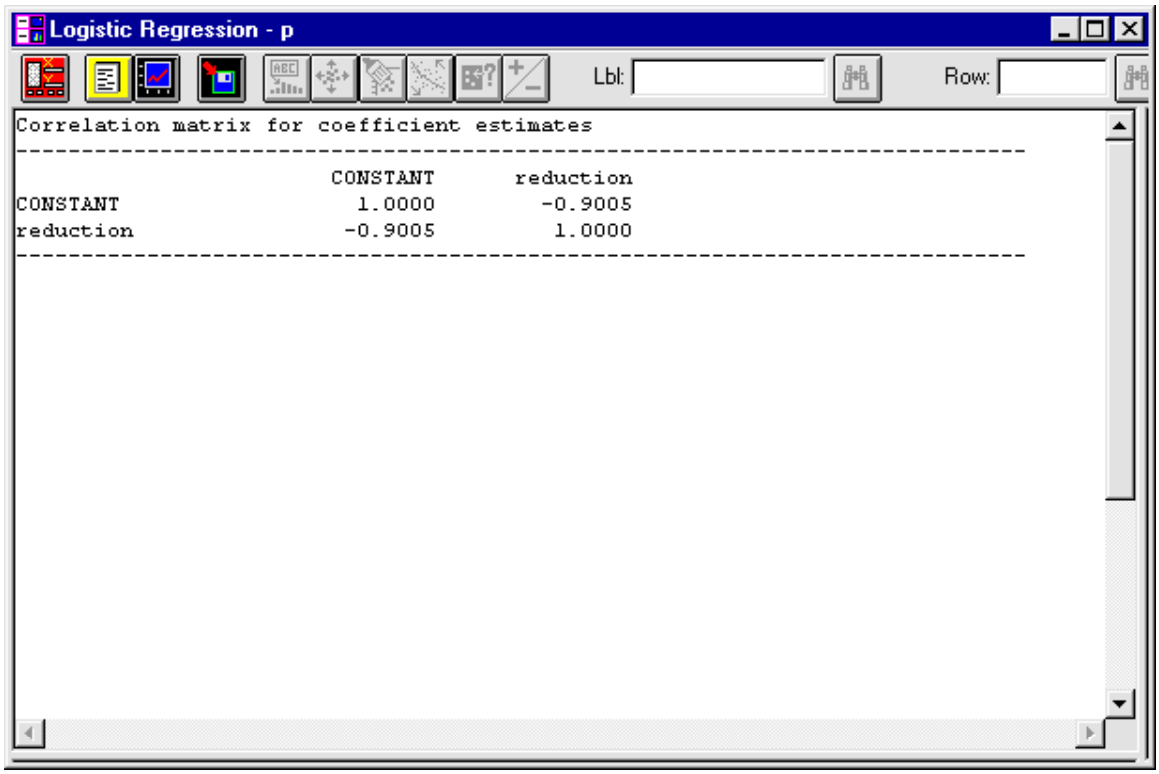


Figure 7-6. Correlation Matrix

Predictions

The Predictions option displays a summary of the prediction capability of the fitted model (see Figure 7-7). The program first uses the model to predict the response, working with the information in each row of the file. If the predicted value is larger than the cutoff value, the response is predicted to be True. If the predicted value is less than or equal to the cutoff value, the response is predicted to be False.

Also shown is the percentage of observed data that were correctly predicted at various cutoff values. Using the cutoff value that maximizes the total percentage of correct predictions provides a good value when you need to predict additional individuals. Another approach is to predict "Success" for any observations with a predicted probability greater than .05.

Use the *Predictions Options* dialog box to set the range of values that will be displayed in the table, to indicate the values that will be displayed in the table, and

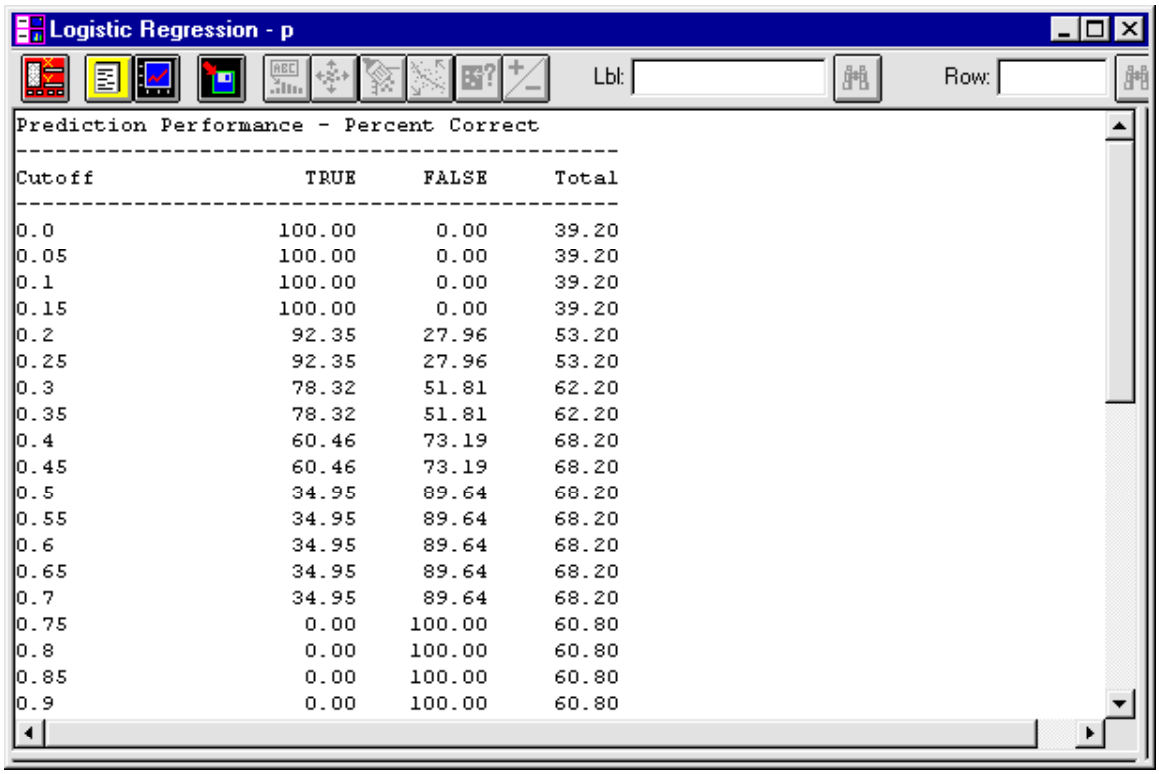


Figure 7-7. Predictions

to enter a number for the confidence level that will be used to calculate the confidence intervals.

Unusual Residuals

The Unusual Residuals option displays a table that lists all the observations with studentized residuals with values greater than 2.0 in absolute value (see Figure 7-8). Studentized residuals measure the number of standard deviations each observed value deviates from the model that was fitted using all the data except that observation.

Percetiles

The Percentiles tabular option displays a table of estimated percentiles based on the fitted model. Confidence limits are also displayed, corresponding to locations on the logit plot which the upper and lower confidence bands cross the indicated percentile (see Figure 7-9).

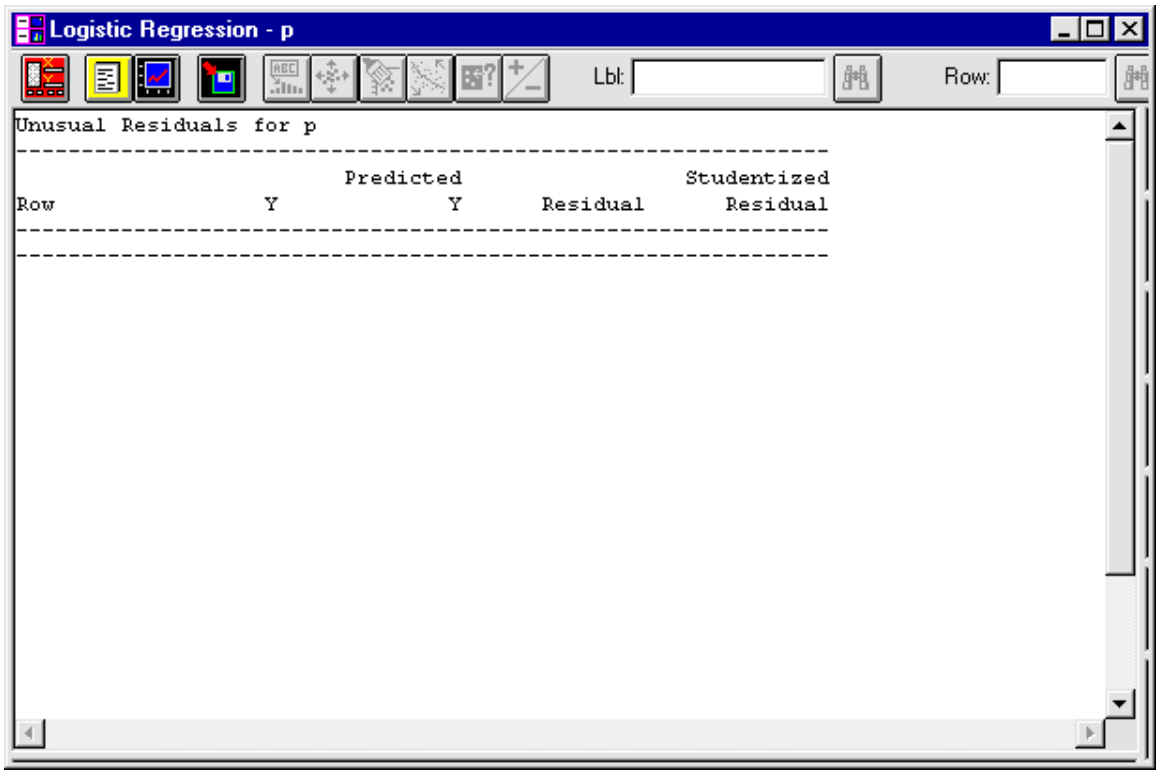


Figure 7-8. Unusual Residuals

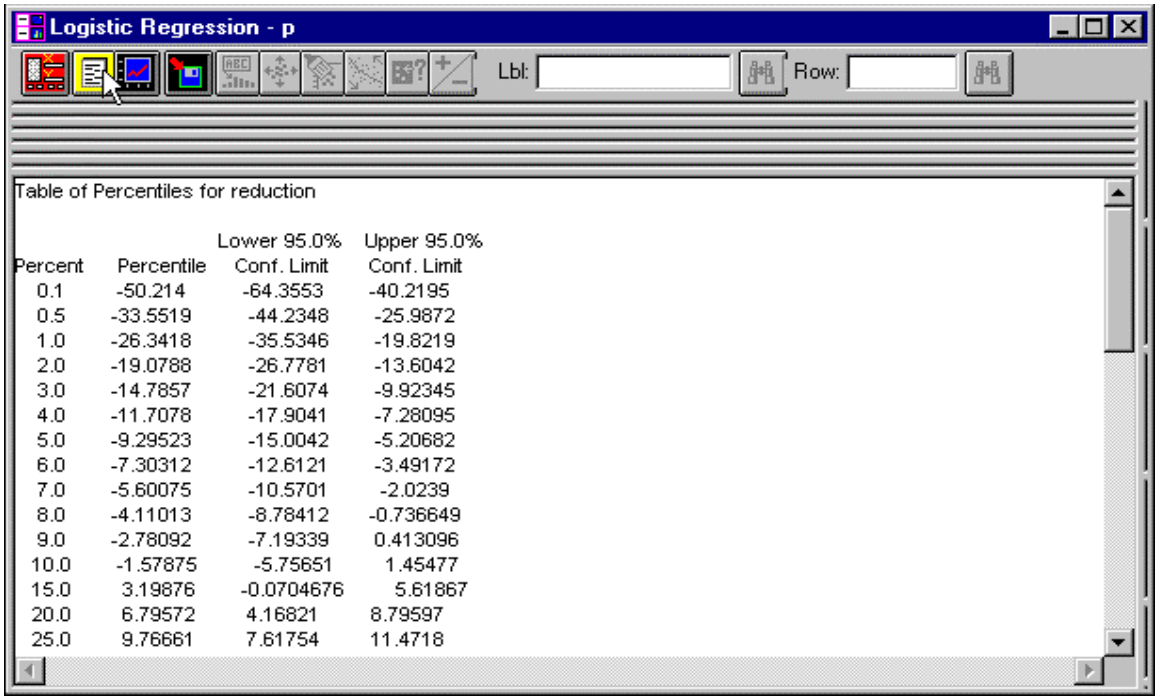


Figure 7-9. Percentiles

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model option displays a plot of the fitted model versus the chosen independent variable (see Figure 7-10).

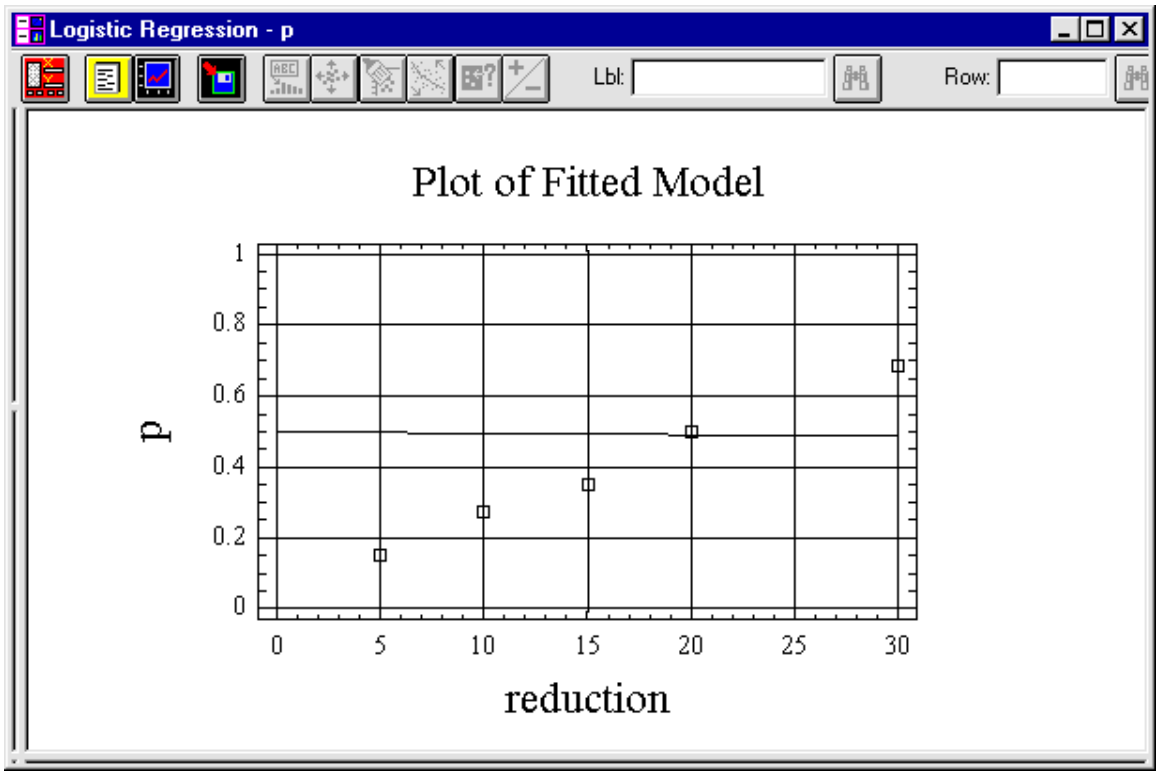


Figure 7-10. Plot of Fitted Model

If there is only one independent variable, a curve is drawn over the range of X, and points are plotted for the values of the dependent variables. The StatAdvisor shows the equation for the fitted model.

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model, and to enter values for the levels for holding the other variables, or new limits for the axis for the chosen variable. The Plot of Fitted Model Options now includes confidence limits for the model or choose to plot the survivor function (see Figure 7-11).

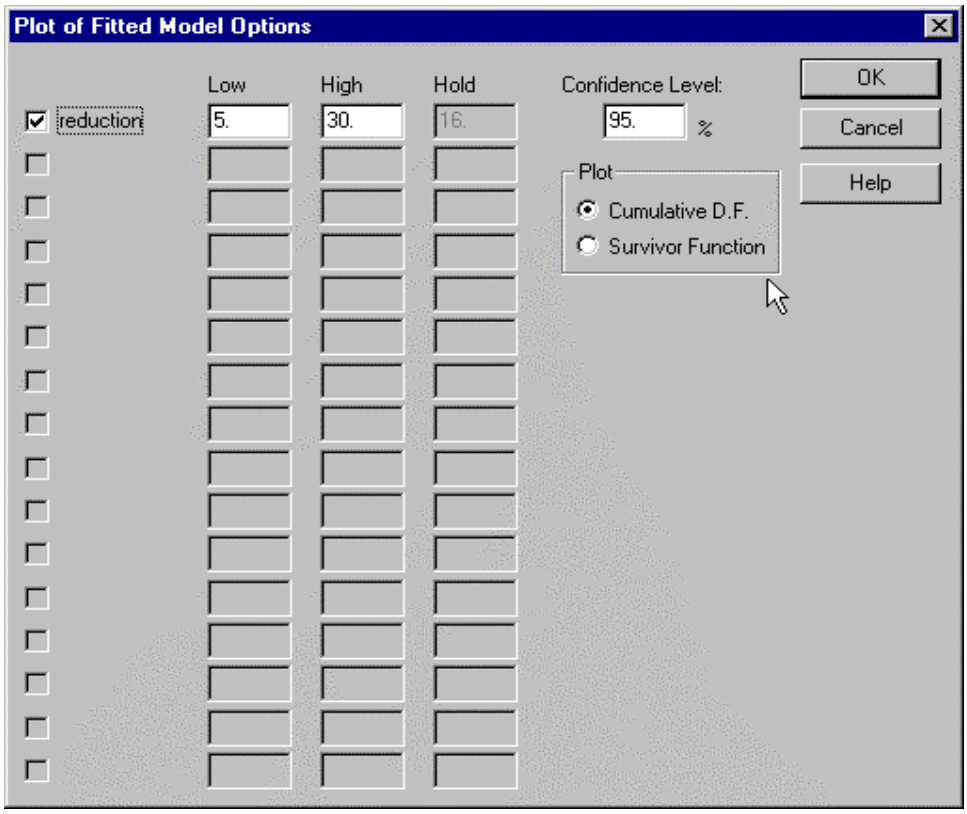


Figure 7-11. Plot of Fitted Model Options Dialog Box

Logit Plot

The Logit Plot option displays a plot with a probability scale on the vertical axis and optional confidence limits (see Figure 7-12).

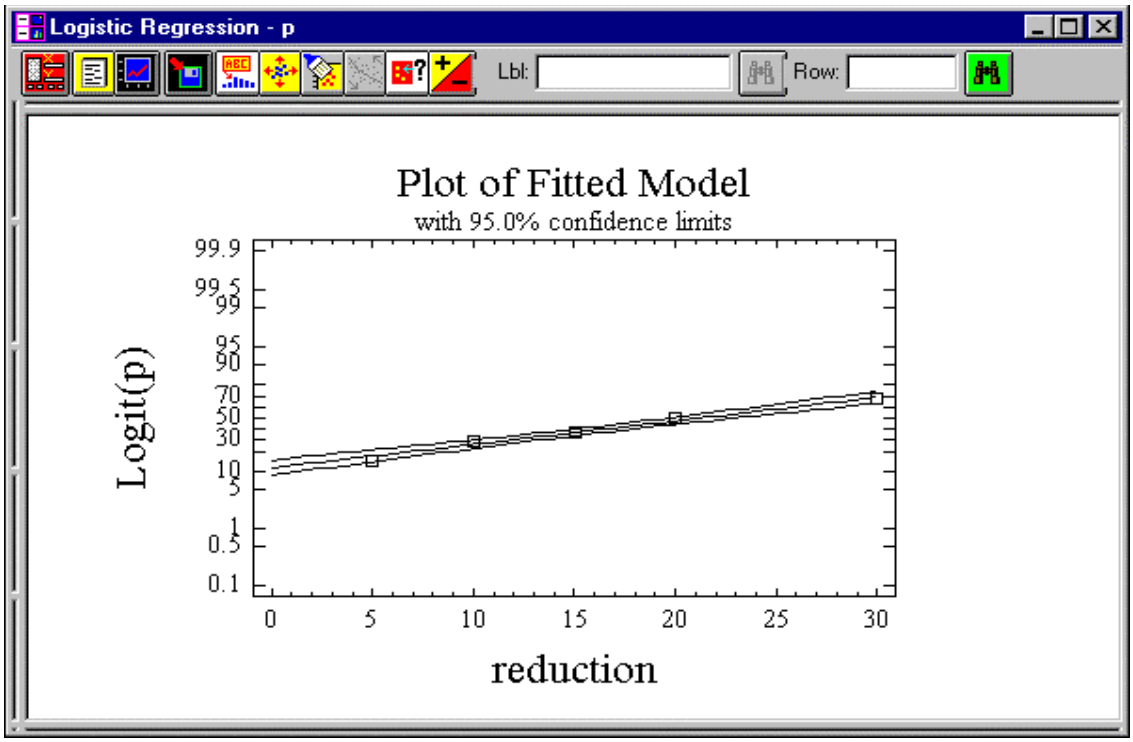


Figure 7-12. Logit Plot

Points appear on the plot only if there is a single factor and you enter data as proportions and sample sizes. The StatAdvisor displays the equation for the logit transformation.

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model, and to enter values for the levels that will hold the other variables, or for new limits for the axis for the chosen variable.

Observed versus Predicted

The Observed versus Predicted option displays a plot of the residuals from the fitted model plotted against the predicted values of the dependent variable (see Figure 7-13). Use the plot to determine any unusual patterns in the residuals.

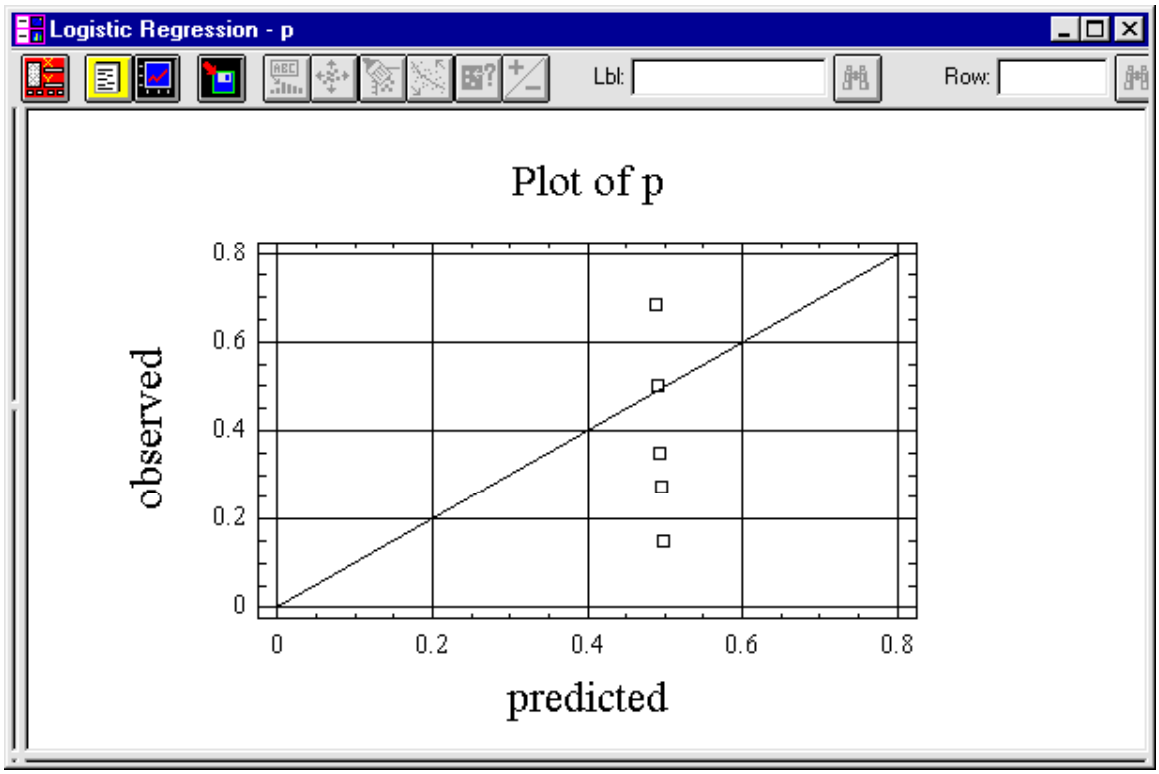


Figure 7-13. Observed versus Predicted Plot

Observed versus Log Odds

The Observed versus Log Odds option displays a plot of the values of Y versus the values of the log odds as predicted by the fitted model (see Figure 7-14).

The StatAdvisor displays the equation for the logit transformation. Use the plot to detect cases in which the variance is not constant or to determine if the dependent variables should be transformed.

Prediction Capability

The Prediction Capability option displays a plot that summarizes the prediction capability of the fitted logistic model: the percentages of correct values versus the cutoff values for the dependent variable that were Total, True, or False (see Figure 7-15).

The program first uses the model to predict the response using the information in each row of the file. If the predicted value is larger than the cutoff, the response is predicted to be True. If the predicted value is less than or equal to the cutoff v alue, the response is predicted to be False. The plot shows the percentage of observed data that were correctly predicted at various cutoff values.

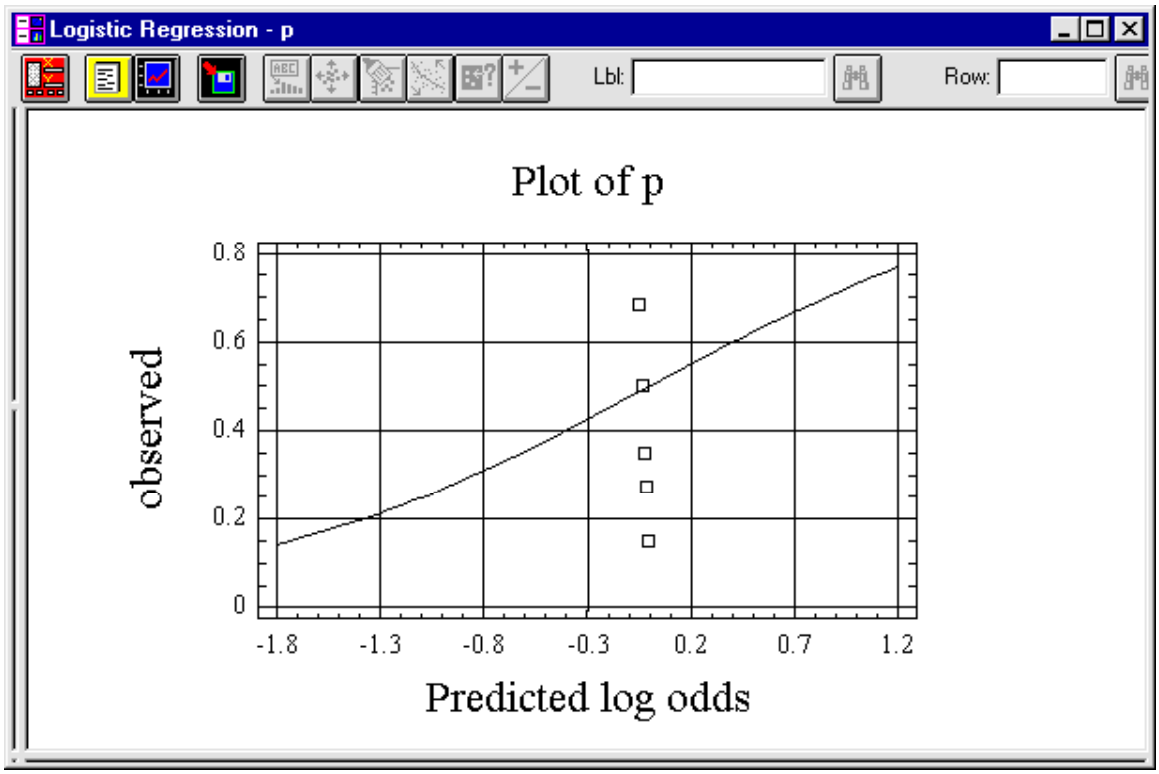


Figure 7-14. Observed versus Log Odds Plot

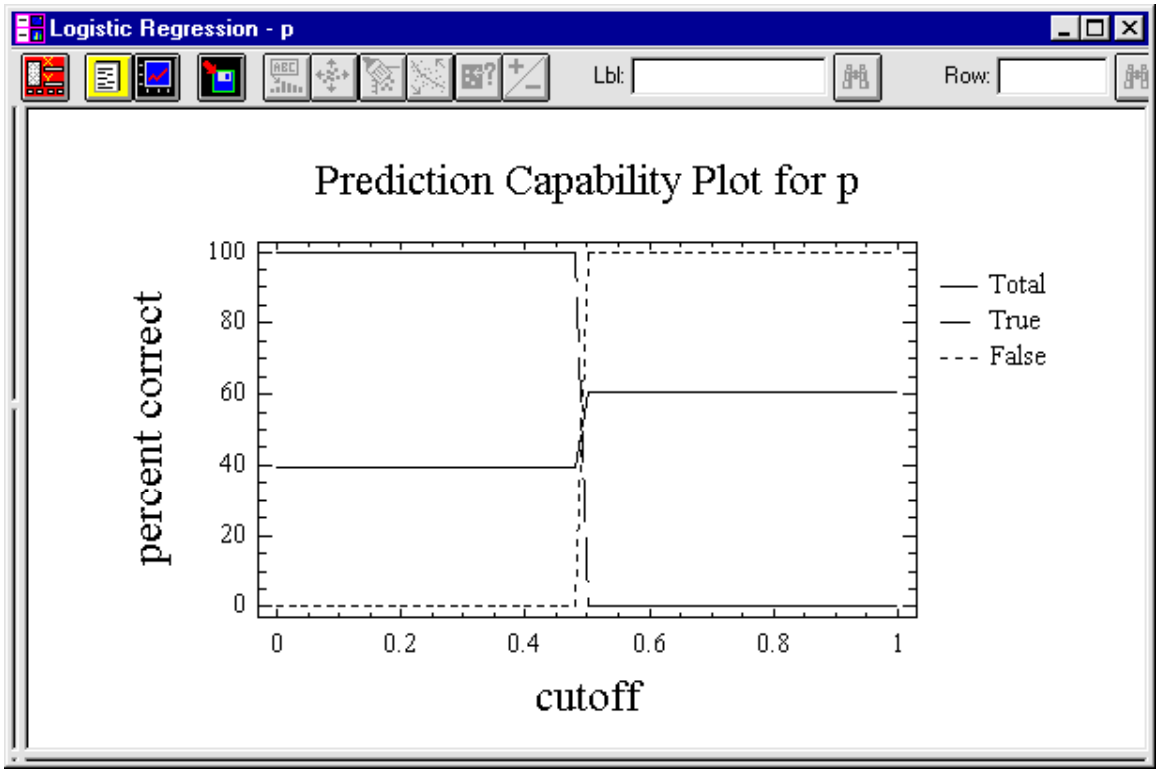


Figure 7-15. Prediction Capability Plot

Prediction Histograms

The Prediction Histograms option displays a plot that demonstrates the ability of the fitted logistic model to distinguish between cases when the dependent variable is True or False (see Figure 7-16).

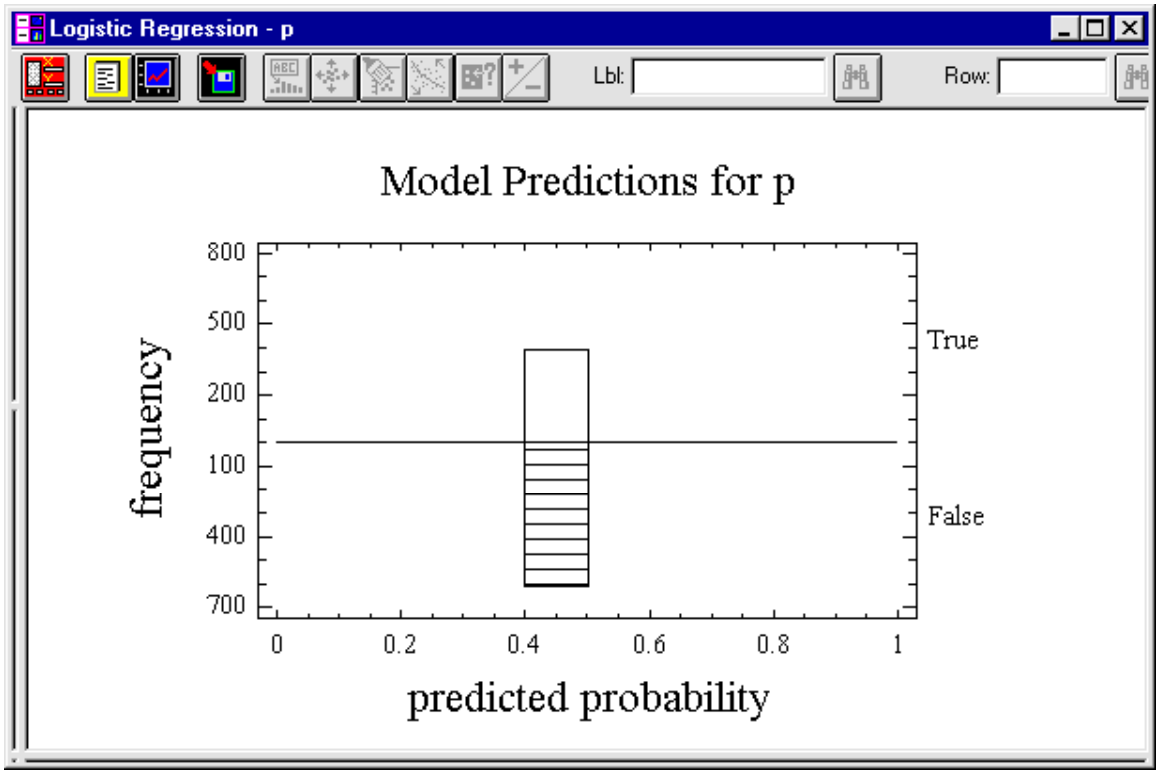


Figure 7-16. Prediction Histograms

The plot shows the frequency distribution of True and False cases versus the probability predicted by the fitted model. In an ideal situation, the model will predict a small probability for the False cases and a large probability for the True cases. If the model works well, the larger frequencies above the line will appear to the right; larger frequencies below the line will appear to the left.

Use the *Histogram Options* dialog box to enter the number of classes into which the data will be divided, to enter values for the lower and upper limits, and to indicate the type of counts that will be included in the plot. You can also indicate if the current scaling should be retained if you change the values on the Analysis dialog box.

Residual Plots

The Residual Plots option displays three different types of plots: Scatterplots, including Residual versus Predicted, Residual versus Row Number, and Residual

versus X; as well as a Normal Probabiity Plot, and an Autocorrelation Function Plot. The Residual Plots allow the user to plot any of three different types of residuals, controlled by Pane Options (see Figure 7-17).

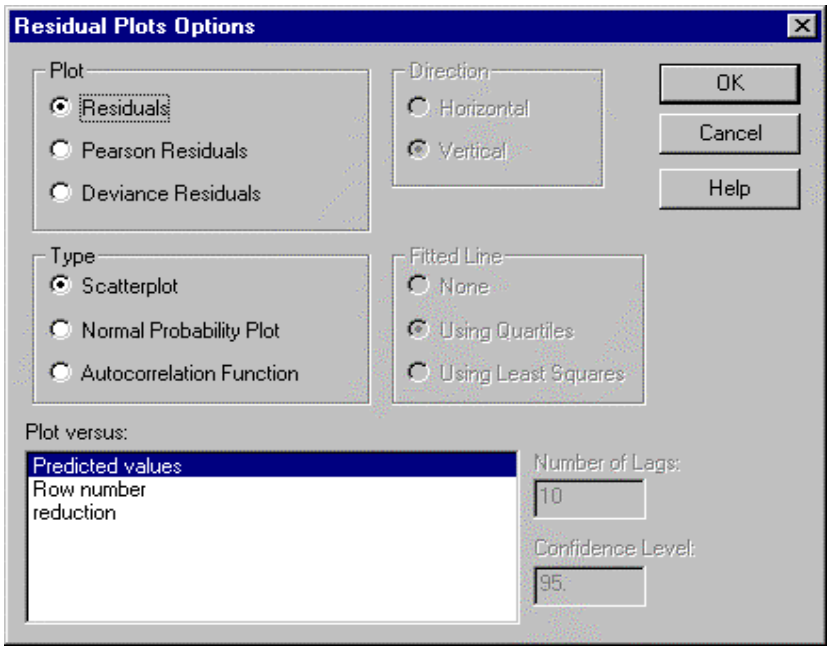


Figure 7-17. Residual Plots Options Dialog Box

Use the *Residual Plots Options* dialog box to choose one of the plots, and, if applicable, its options.

Residuals versus Predicted

The Residual versus Predicted scatterplot displays the residual or the studentized residual versus the observed variable (see Figure 7-18). Examine the residuals to look for any unusual patterns. The response is limited to be between 0 and 1, so the residuals will not follow a normal distribution.

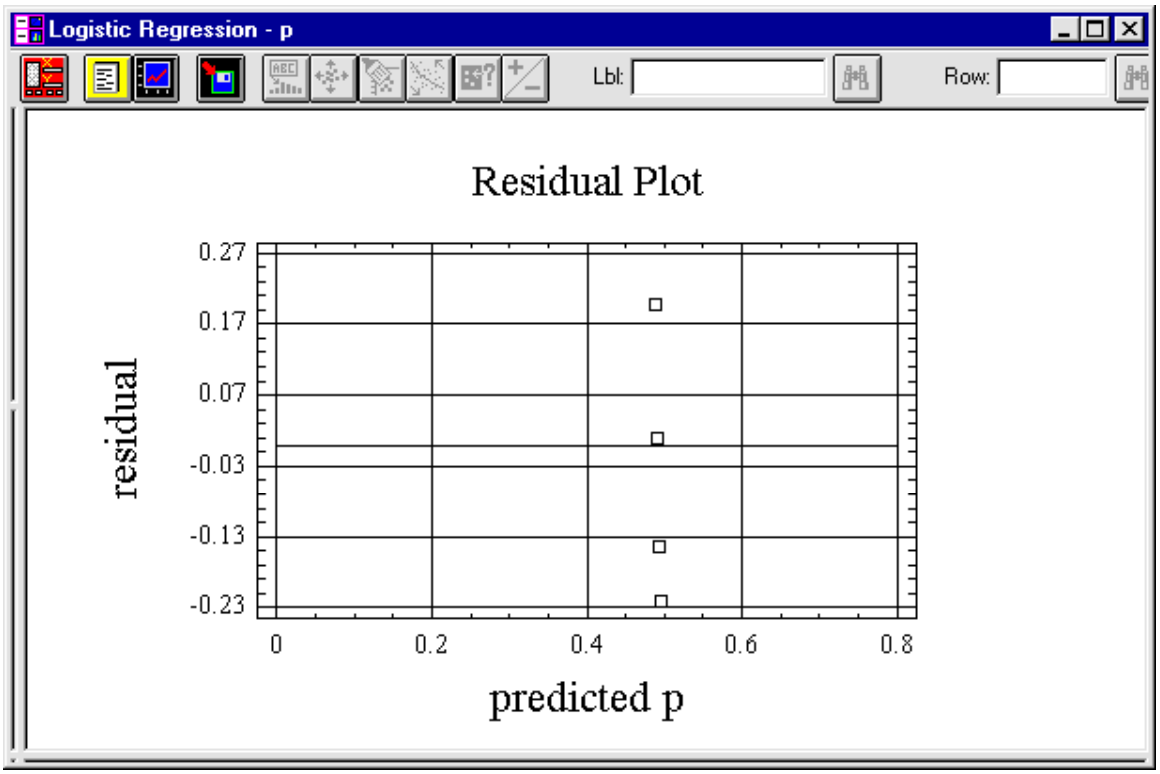


Figure 7-18. Residual versus Predicted Scatterplot

Residual versus Row Number

The Residual versus Row Number scatterplot displays the residual or the studentized residual versus the row number (see Figure 7-19). The program plots the residuals in the order that the observations appear in the dependent variable.

The plot is helpful in determining sequential correlations among the residuals. Any nonrandom pattern indicates serial correlation in the data, particularly if the row number corresponds with the order in which the data were collected.

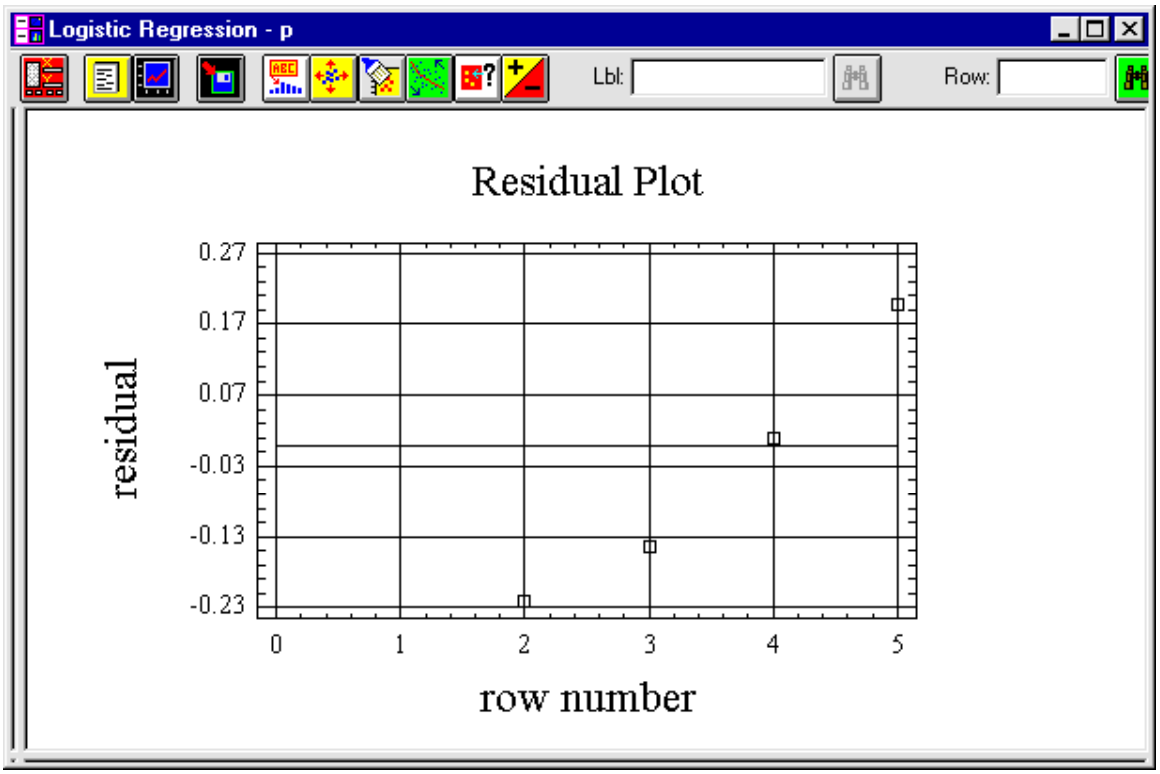


Figure 7-19. Residual versus Row Number Scatterplot

Residual versus X

The Residual versus X scatterplot displays the residual or studentized residual versus the independent variable (X). Use this plot to detect the nonlinear relationship between the dependent and independent variables (see Figure 7-20).

Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may be an outlier.

Normal Probability Plot

The Normal Probability Plot option displays a plot that determines if the errors follow a normal distribution (see Figure 7-21). The program sorts the residuals from smallest to largest, then plots them versus the values. If the data come from a normal distribution, the points should fall approximately along a straight line. To help determine how close the points are to the line, the program draws a reference line on the plot, which passes through the median with the slope determined from the interquartile range. If the points show a significant curvature, it may indicate that the residuals are not from a normal distribution.

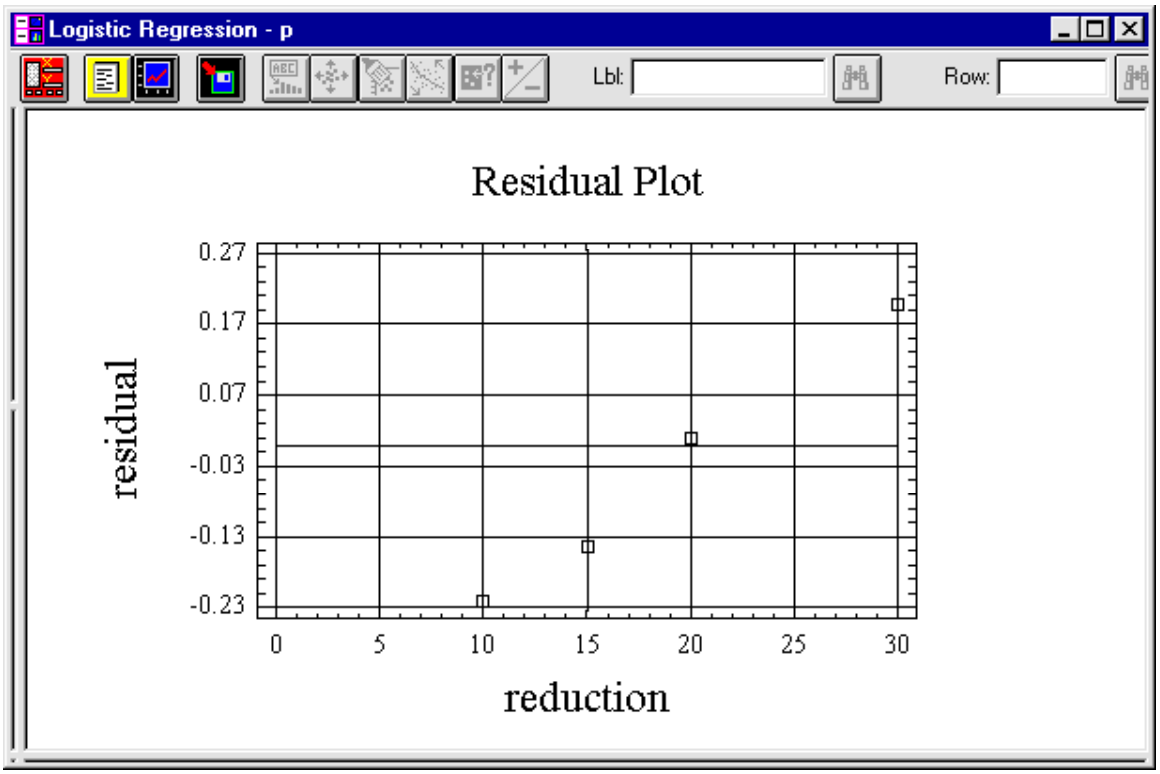


Figure 7-20. Residual versus X Scatterplot

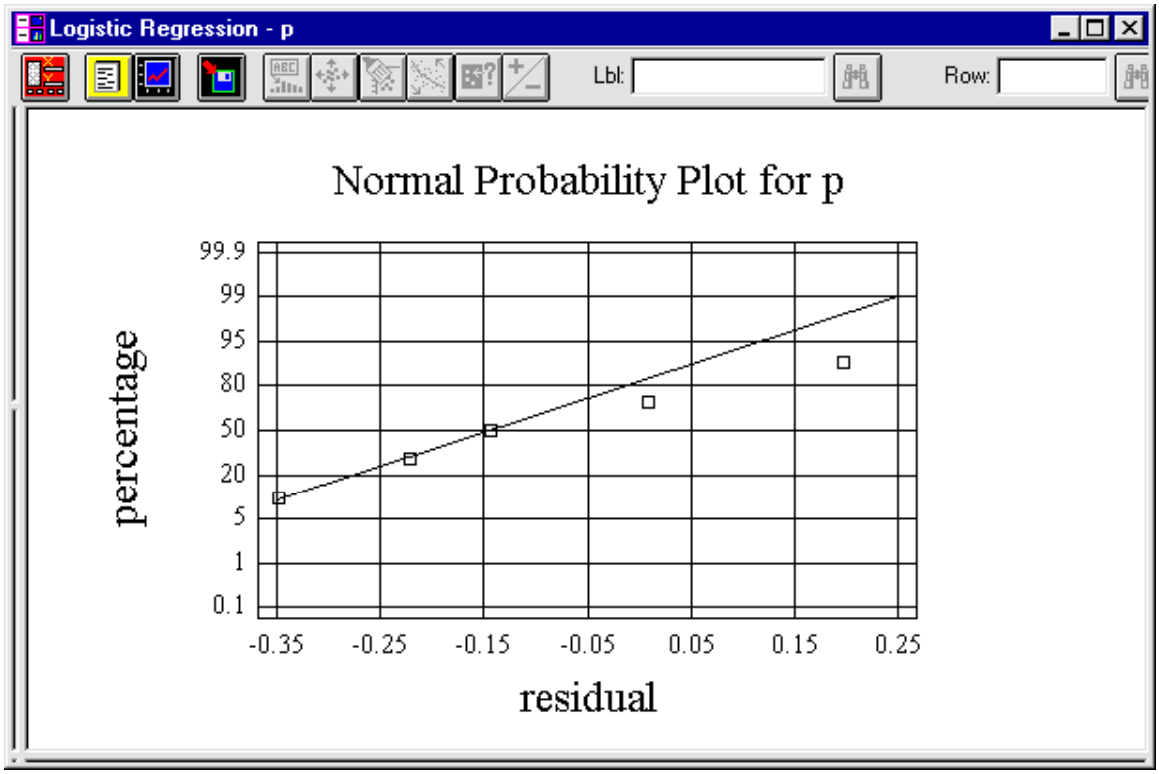


Figure 7-21. Normal Probability Plot

Autocorrelation Function

The Autocorrelation Function option displays a plot of the autocorrelation estimates for the residuals (see Figure 7-22).

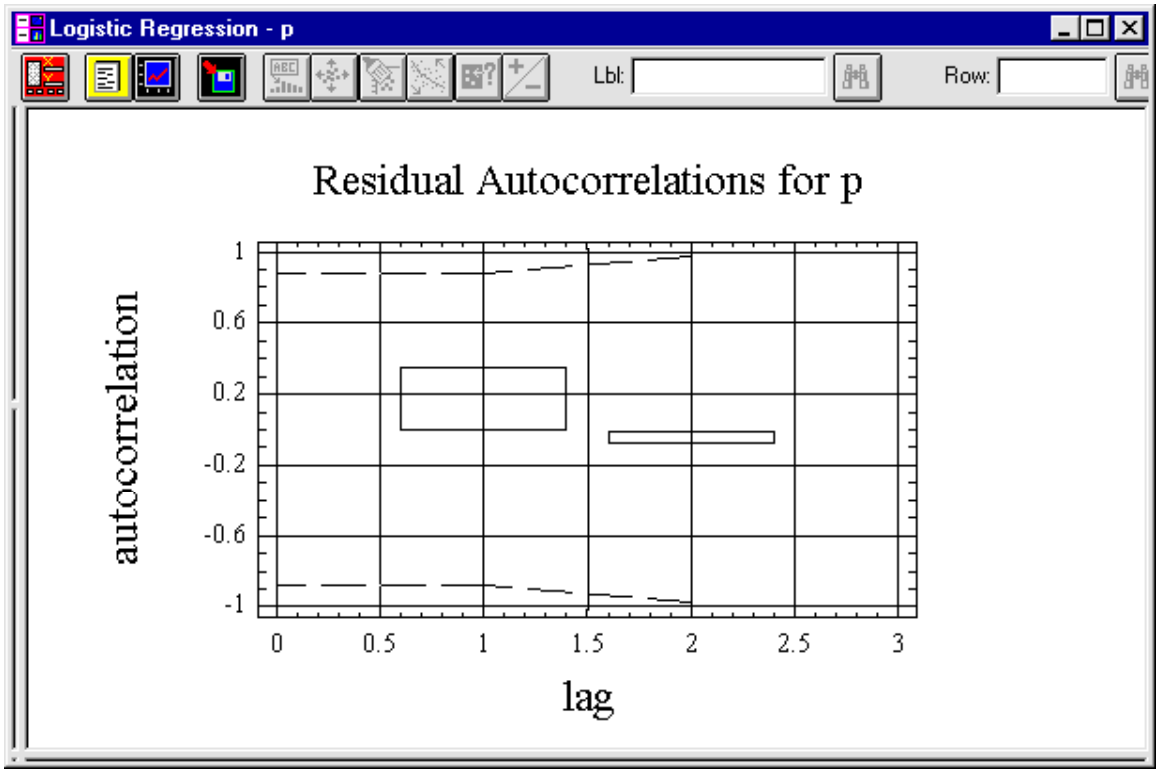


Figure 7-22. Autocorrelation Function Plot

The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

Use the *Autocorrelation Function Plot Options* dialog box to choose the type of residual that will be plotted, to choose the direction of the plot, to choose the type of values that will be used for the fitted line, and to enter values for the maximum lag and confidence level.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 11 selections: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Residuals, Pearson residuals, Deviance Residuals,

Leverages, Percentages, Percentiles, Lower Fiducial Limits, and Upper Fiducial Limits.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Bowerman, B.L., and O'Connell,R. T. 1990. *Linear Statistical Models: An Applied Approach*, second edition. PWS-Kent.

Cox, D. R. 1970. *The Analysis of Binary Data*. London: Methuen and Co. Ltd.

Chatterjee, S. and Price, B. 1991. *Regression Analysis by Example*, second edition. New York: John Wiley & Sons, Inc.

Collett, D. 1991. *Modelling Binary Data*. London: Chapman and Hall.

Hosmer, D. W. and Lemeshow, S. 1989. *Applied Logistic Regression*. Wiley.

Kleinbaum D. G., Kupper L. L., and Muller, K.E. 1997. *Applied Regression Analysis and Other Multivariable Methods*, third edition. PWS-Kent.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Nelder, J.A., and McCullagh, P. 1989. *Generalized Linear Models*, second edition. Chapman and Hall.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. Chicago, Illinois: Richard D. Irwin, Inc.

Chapter 8

Using Probit Analysis

Background Information

Chapter 7 describes a class of logistic regression models designed for situations where the outcome of an experiment results in a binary response such as a success or failure. In that chapter, a logistic transformation was applied to the response, resulting in a linear model of the form:

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X$$

This chapter describes an alternative form of model which involves the *probit transformation*

$$\Phi^{-1}(Y) = \beta_0 + \beta_1 X$$

in which an observed proportion Y is transformed using the inverse cumulative standard normal distribution function $\Phi^{-1}(Y)$. Such models are commonly used to describe the response of experimental subjects to varying doses of a drug.
This chapter discusses the estimation of probit models of the form

$$Y = \Phi(\beta_0 + \beta_1 X)$$

and multifactor models of similar form. It also describes the estimation of important quantities such as percentiles (levels of dose at which the response reaches a specified level). For a full discussion of other *Probit Analysis* options, refer to the *Logistic Regression* chapter.

Probit Analysis in STATGRAPHICS *Plus*

As is the case with the *Logistic Regression* procedure, *Probit Analysis* is designed to handle data in either of two formats:

1. *Aggregated data* consisting of proportions and sample sizes.
2. *Binary data* (0's and 1's), where each experimental subject is classified as an occurrence or non-occurrence.

The current example contains data in the first format and would be entered into the data input dialog box as shown in Figure 8-1.
The independent or explanatory variables can be either categorical or continuous. The program creates indicator variables for categorical factors that can simplify the use of the model when making predictions for future cases. There are several statistics and graphs that help to assess the accuracy and usefulness of the model.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... PROBIT ALANYSIS... from the Menu bar to display the Probit Analysis dialog box (see Figure 8-1).

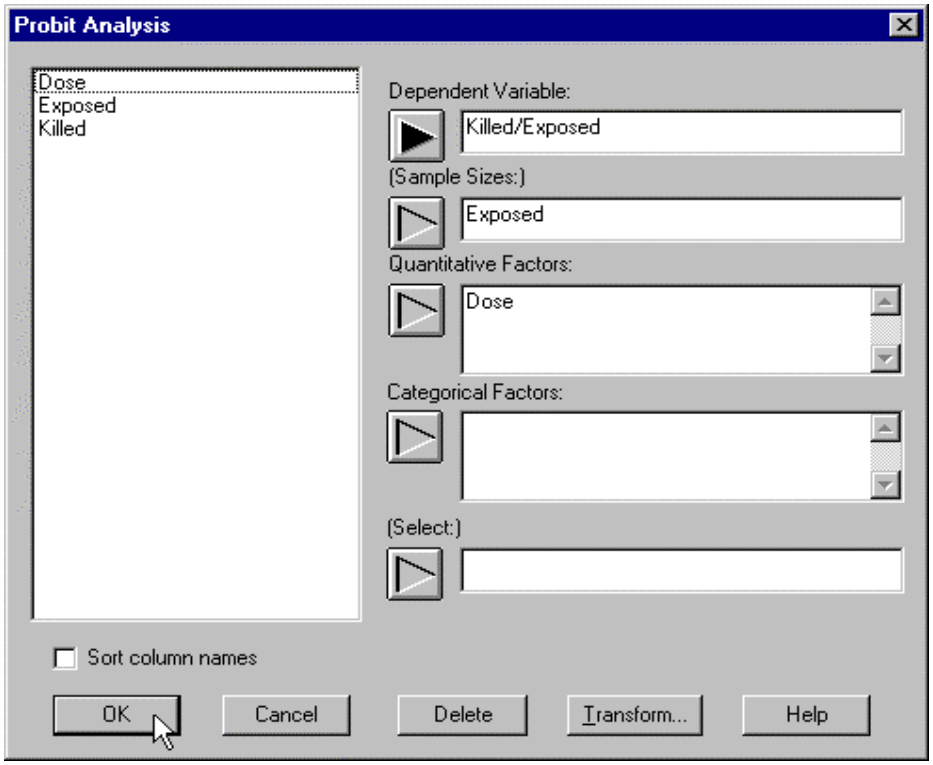


Figure 8-1. Probit Analysis Dialog

Tabular Options

Analysis Summary

The Analysis Summary option displays the results of fitting a probit regression model that describes the relationship between the dependent and independent variables (see Figure 8-2).

The results include estimates for each of the coefficients, approximate standard errors, and the estimated odds ratios. If the *p*-value is less than 0.01, there is a statistically significant relationship among the variables. If the *p*-value for the residuals is greater than or equal to 0.10, it indicates that the model is not significantly worse than the best possible model for the data you are currently using.

If the model was fit using weighted least squares, the results include: *t*-tests with associated *p*-values; an ANOVA table; summary statistics, such as R-Squared; and the Type III sums of squares for each of the factors.

If the model was fit using maximum likelihood, the results include an analysis of model deviance and the percentage of deviance for which the model accounts. This value is similar to the R-Squared statistic. The results also include values for tests of likelihood ratio for each of the factors.

If you used the optional Select text box on the Probit Analysis dialog box, the program uses prediction error statistics to validate the accuracy of the model and displays the Residuals Table at the end of the Analysis Summary. If you decide to validate the model, use the methods discussed in the topic, "Overview of the Model-Building Process," in Online Help.

The table includes values for the following statistics for the validation and estimation data:

- n — the number of observations.
- MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in the data, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.
- MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different model.
- MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given set of data without regard to sign (that is, their absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.
- ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the prediction.
- MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. Like MAPE, it is independent of scale.

Use the *Probit Analysis Options* dialog box to choose the estimation method, to enter a value for the smallest proportion (for weighted least squares), to choose the type of model to be fit, to choose a selection procedure, to enter values for *p* to enter

or remove, to enter a value for the maximum number of steps, and to indicate how the results will be displayed.

```
Probit Analysis

Dependent variable: Killed/Exposed
Sample sizes: Exposed
Factors:
   Dose

Estimated Regression Model (Maximum Likelihood)
-----------------------------------------------
                                       Standard
Parameter               Estimate         Error
-----------------------------------------------
CONSTANT                -34.9349        2.65395
Dose                     19.7277        1.49062
-----------------------------------------------

              Analysis of Deviance
--------------------------------------------------
Source          Deviance            Df      P-Value
--------------------------------------------------
Model           274.083              1       0.0000
Residual        10.1198              6       0.1197
--------------------------------------------------
Total (corr.)   284.202              7

Percentage of deviance explained by model = 96.4392
Adjusted percentage = 95.0318

Likelihood Ratio Tests
----------------------------------------------------------
Factor                   Chi-Square          Df   P-Value
----------------------------------------------------------
Dose                     274.083              1    0.0000
----------------------------------------------------------

Residual Analysis
---------------------------------
      Estimation       Validation
n     8
MSE   0.131797
MAE   0.0562163
MAPE  17.4188
ME    -0.0211148
MPE   -3.25668
```

Figure 8-2. Analysis Summary

Goodness-of-Fit

The Goodness-of-Fit option displays a table that divides the logistic scale into intervals, each of which contains approximately the same number of observations (see Figure 8-3).

The program compares the observed versus the predicted number of True and False observations in each interval of the observed data with those predicted by the model to determine if the function adequately fits the observed data. Small *p*-values indicate a significant lack of fit.

Use the *Goodness-of-Fit Options* dialog box to enter a value for the number of classes into which the data will be grouped.

Confidence Intervals

The Confidence Intervals option displays confidence intervals for the coefficients in the model and the odds ratios using a confidence level of 95 percent (see Figure 8-4). The confidence intervals illustrate the preciseness with which the coefficients were estimated, given the amount of available data and the noise present. The odds ratios equal the inverse natural logarithm of the coefficient and show the proportional change in the response variable as the independent variable is increased by one unit.

Use the *Confidence Intervals Options* dialog box to enter a number that will be used to calculate the confidence intervals for the mean and standard deviation.

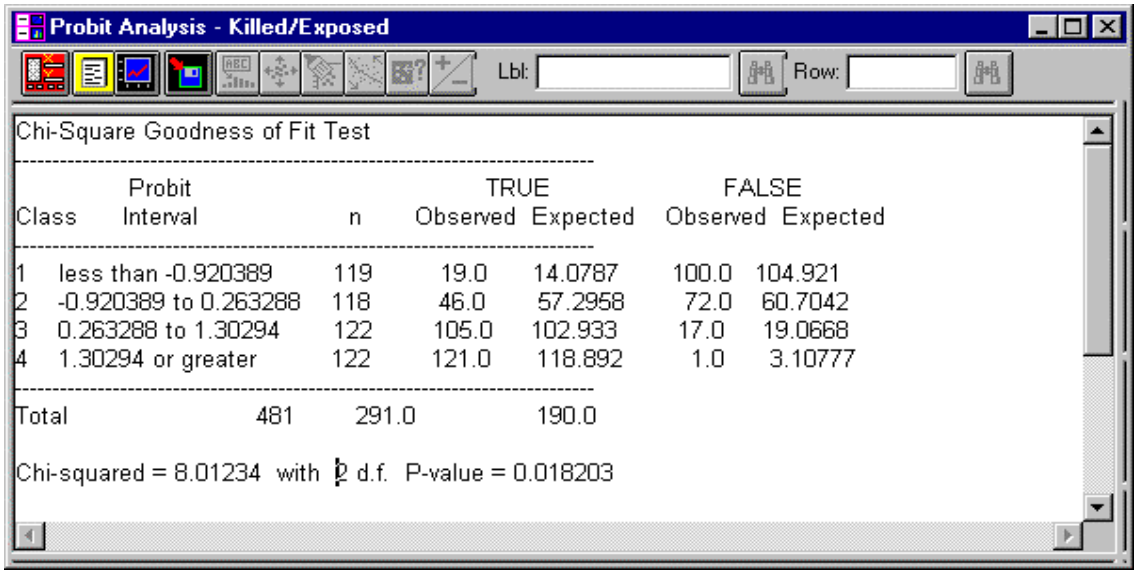


Figure 8-3. Goodness-of-Fit Results

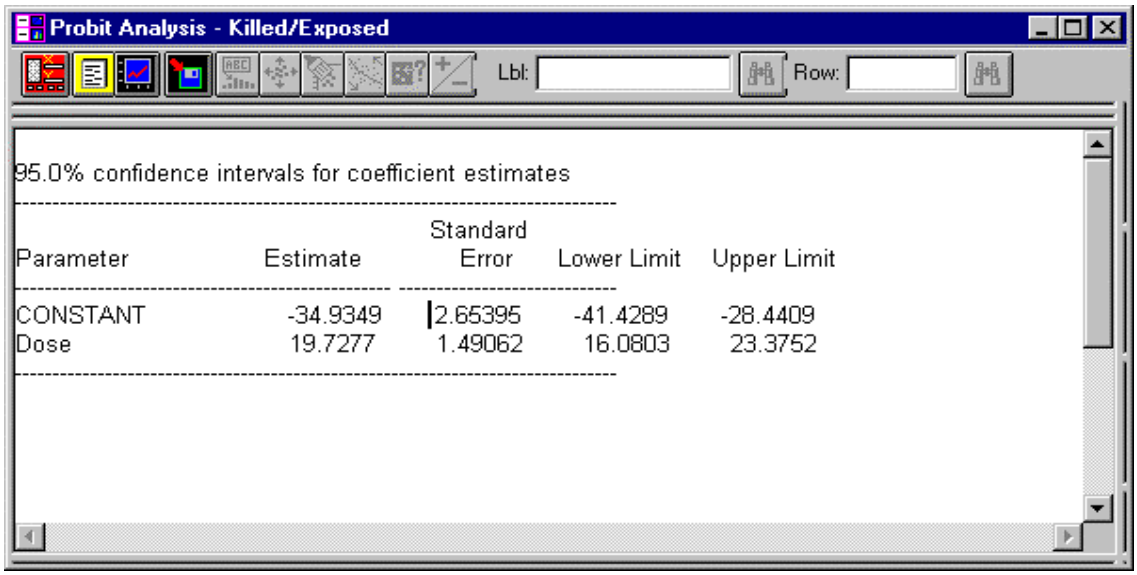


Figure 8-4. Confidence Intervals

Correlation Matrix

The Correlation Matrix option displays a table of the estimated correlations between the coefficients in the fitted model (see Figure 8-5). The correlations are helpful in detecting the presence of serious multicollinearity.

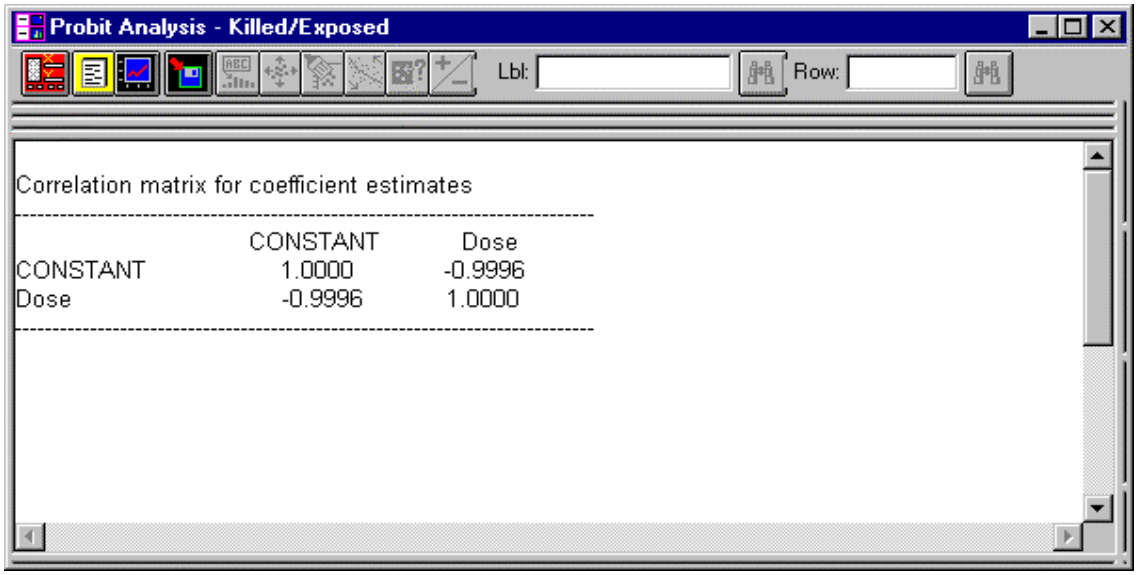


Figure 7-5. Correlation Matrix

Predictions

The Predictions option displays a summary of the prediction capability of the fitted model (see Figure 8-6). The program first uses the model to predict the response, working with the information in each row of the file. If the predicted value is larger

than the cutoff value, the response is predicted to be True. If the predicted value is less than or equal to the cutoff value, the response is predicted to be False.

Also shown is the percentage of observed data that were correctly predicted at various cutoff values. Using the cutoff value that maximizes the total percentage of correct predictions provides a good value when you need to predict additional individuals. Another approach is to predict "Success" for any observations with a predicted probability greater than .05.

Use the *Predictions Options* dialog box to set the range of values that will be displayed in the table, to indicate the values that will be displayed in the table, and to enter a number for the confidence level that will be used to calculate the confidence intervals.

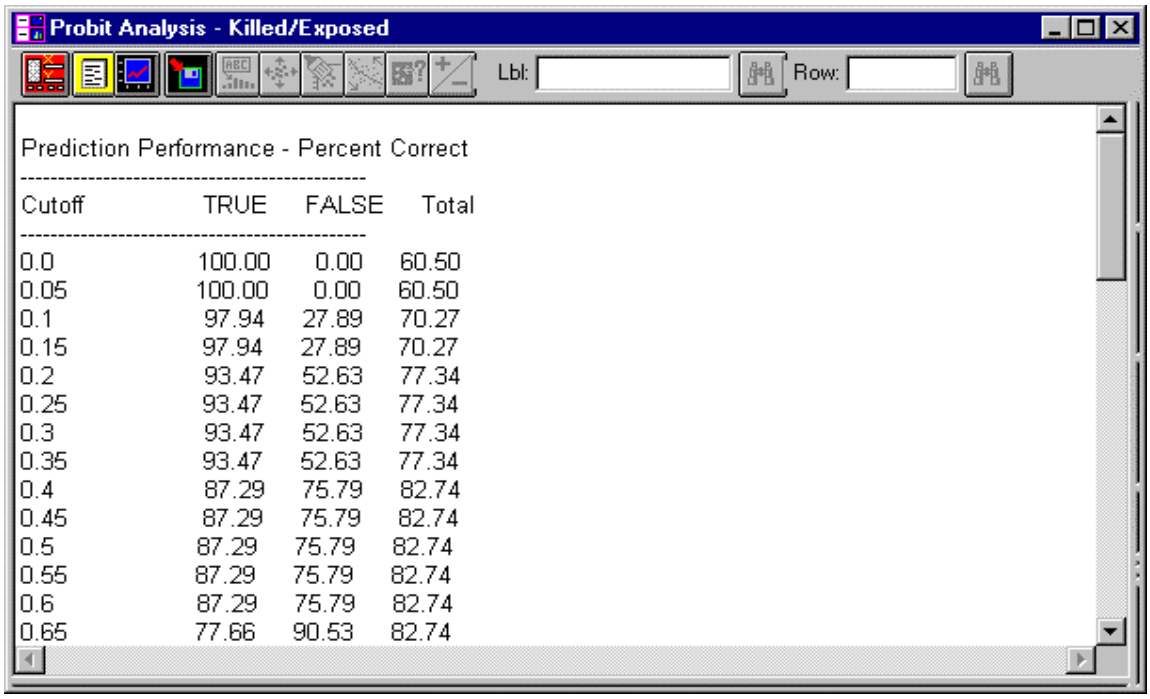


Figure 8-6. Predictions

Unusual Residuals

The Unusual Residuals option displays a table that lists all the observations with studentized residuals with values greater than 2.0 in absolute value (see Figure 8-7). Studentized residuals measure the number of standard deviations each observed value deviates from the model that was fitted using all the data except that observation.

Influential Points

The Influential Points option displays a table that lists observations with leverage values greater than three times that of an average point, or with unusually large values for the DFITS or Cook's distance statistics (see Figure 8-8).

Leverage is a statistic that measures the amount of influence each observation has when determining the coefficients for the estimated model. DFITS is a statistic that measures the amount each estimated coefficient would change if each observation was removed from the data. The Cook's distance statistic measures the distance between the estimated coefficients with and without each observation.

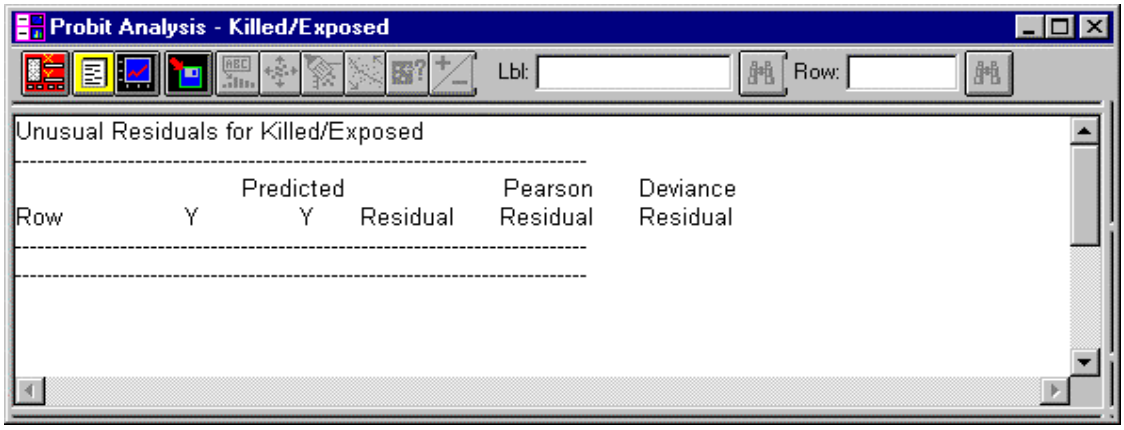


Figure 8-7. Unusual Residuals

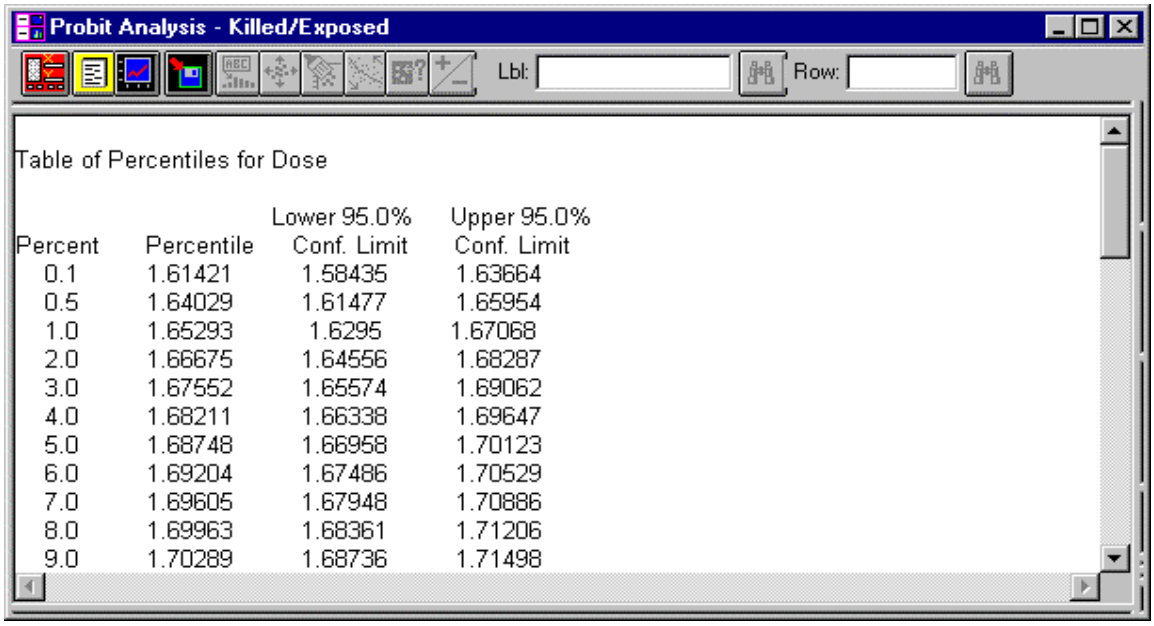


Figure 8-8. Influential Points

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model option displays a plot of the fitted model versus the chosen independent variable (see Figure 8-9).

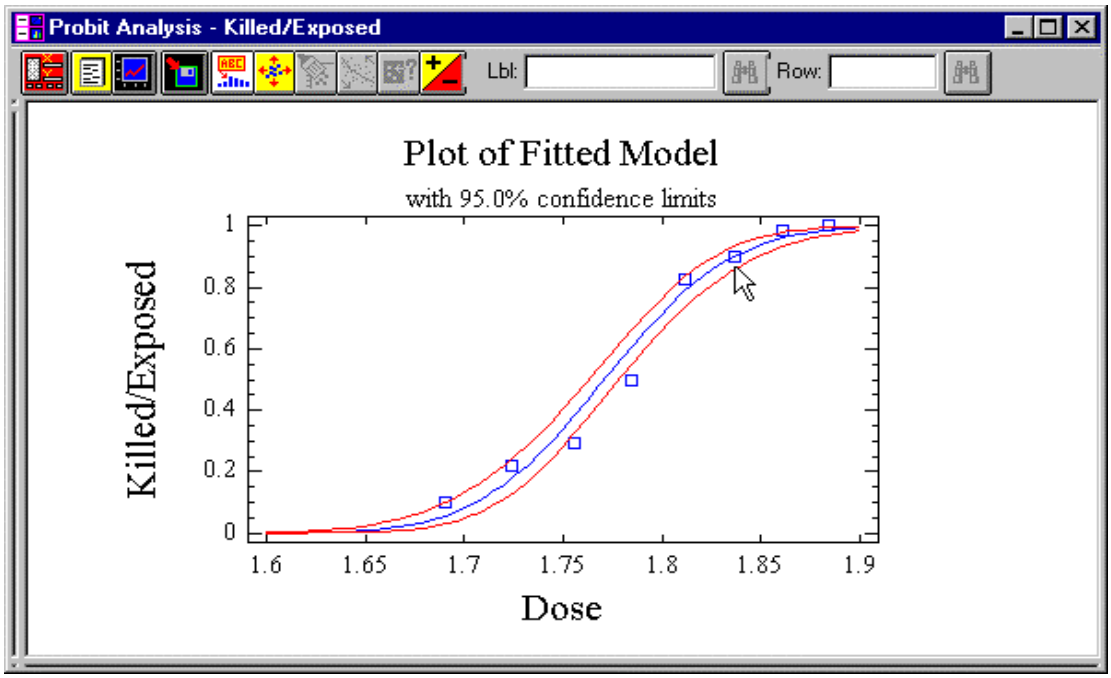


Figure 8-9. Plot of Fitted Model

If there is only one independent variable, a curve is drawn over the range of X, and points are plotted for the values of the dependent variables. The StatAdvisor shows the equation for the fitted model.

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model, and to enter values for the levels for holding the other variables, or new limits for the axis for the chosen variable.

Probit Plot

The Probit Plot option displays a plot similar to a plot of the fitted model except that the scale is transformed using the probit transformation, which results in a straight regression line for a first-order model (see Figure 8-10).

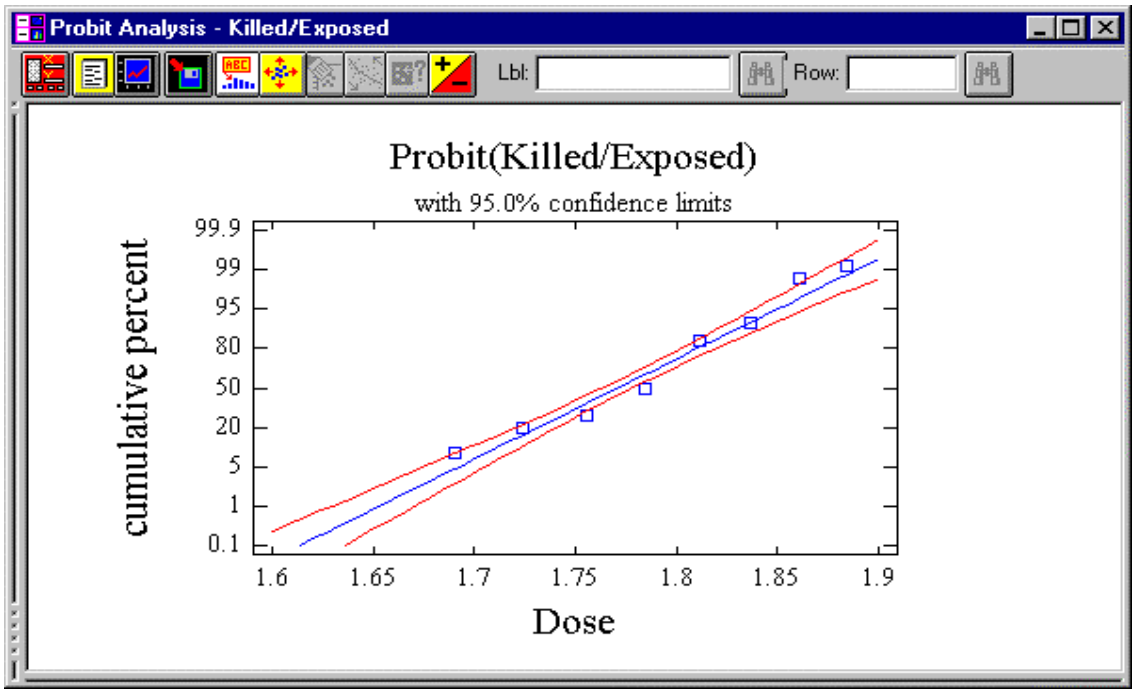


Figure 8-10. Probit Plot

Points appear on the plot only if there is a single factor and you enter data as proportions and sample sizes. The StatAdvisor displays the equation for the probit transformation.

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model, and to enter values for the levels that will hold the other variables, or for new limits for the axis for the chosen variable.

Observed versus Predicted

The Observed versus Predicted option displays a plot of the residuals from the fitted model plotted against the predicted values of the dependent variable (see Figure 8-11). Use the plot to determine any unusual patterns in the residuals.

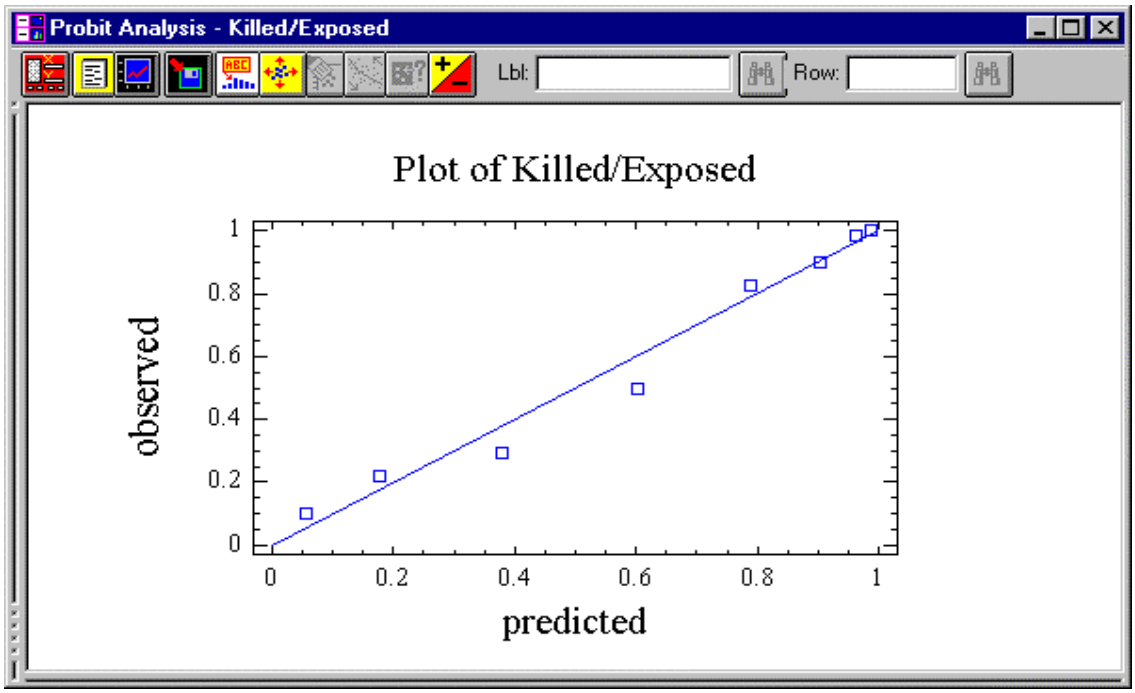


Figure 8-11. Observed versus Predicted Plot

Observed versus Log Odds

The Observed versus Log Odds option displays a plot of the values of Y versus the values of the log odds as predicted by the fitted model (see Figure 8-12).

The StatAdvisor displays the equation for the logit transformation. Use the plot to detect cases in which the variance is not constant or to determine if the dependent variables should be transformed.

Prediction Capability

The Prediction Capability option displays a plot that summarizes the prediction capability of the fitted logistic model: the percentages of correct values versus the cutoff values for the dependent variable that were Total, True, or False (see Figure 8-13).

The program first uses the model to predict the response using the information in each row of the file. If the predicted value is larger than the cutoff, the response is predicted to be True. If the predicted value is less than or equal to the cutoff value, the response is predicted to be False. The plot shows the percentage of observed data that were correctly predicted at various cutoff values.

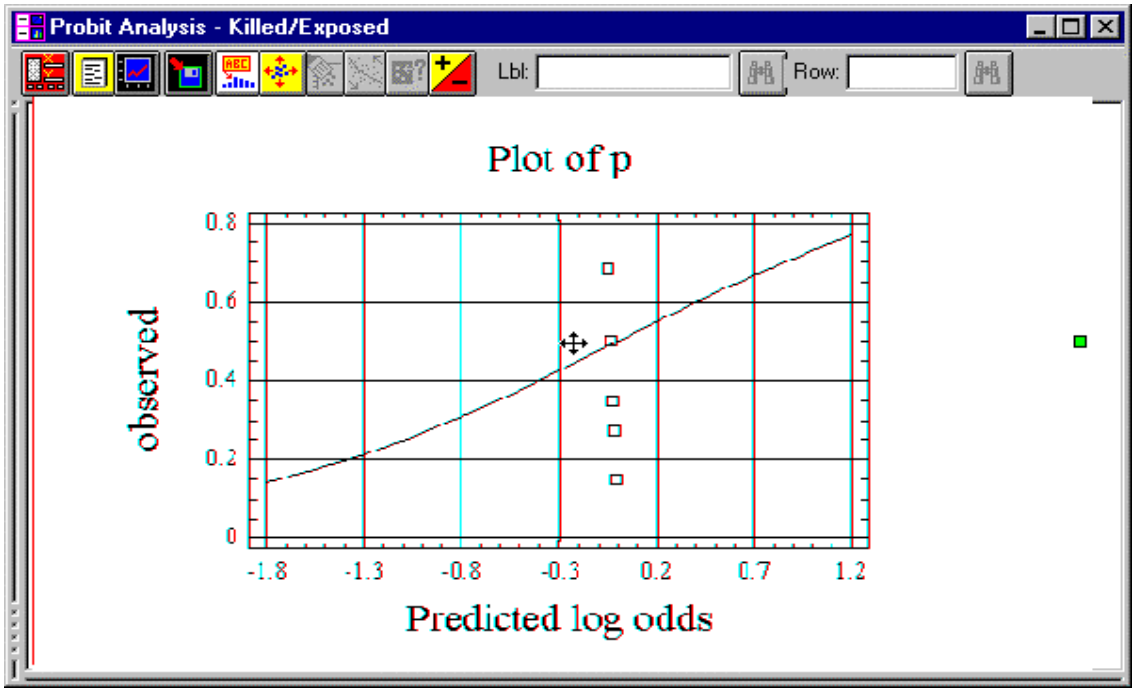


Figure 8-12. Observed versus Log Odds Plot

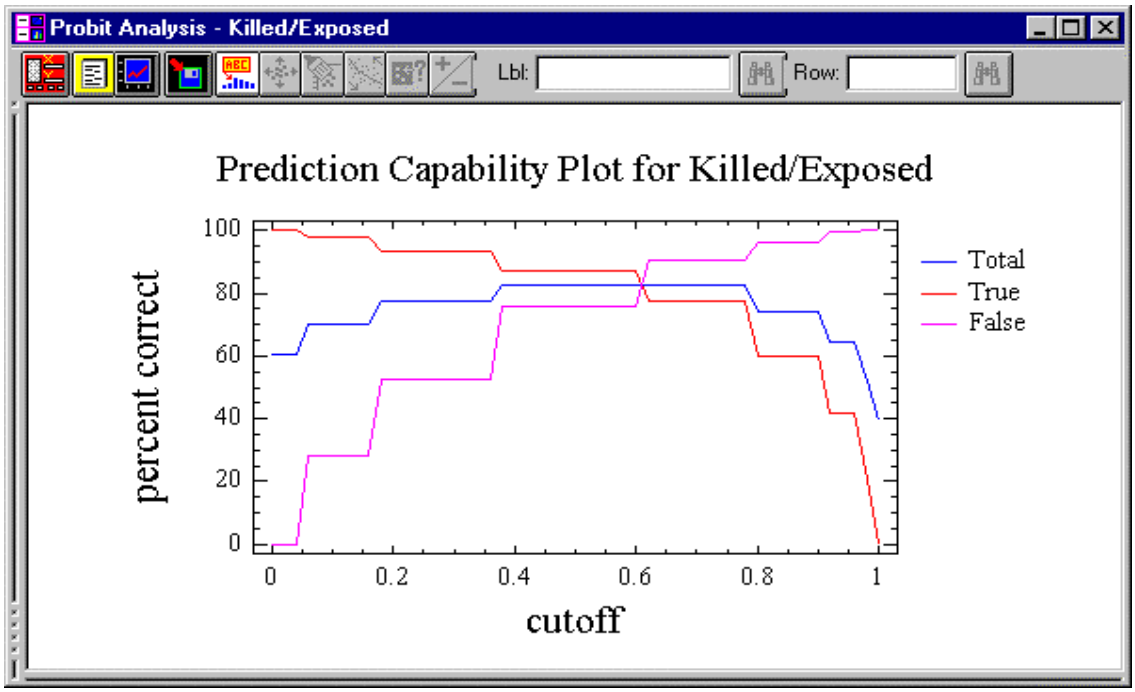


Figure 8-13. Prediction Capability Plot

Prediction Histograms

The Prediction Histograms option displays a plot that demonstrates the ability of the fitted logistic model to distinguish between cases when the dependent variable is True or False (see Figure 8-14).

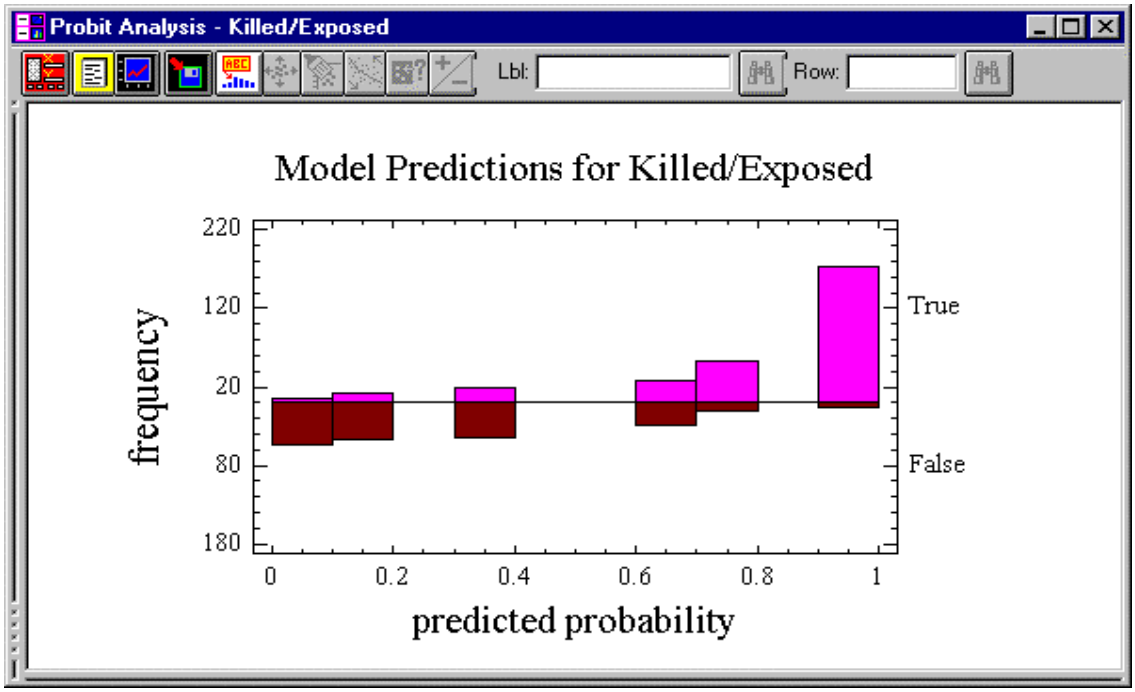


Figure 8-14. Prediction Histograms

The plot shows the frequency distribution of True and False cases versus the probability predicted by the fitted model. In an ideal situation, the model will predict a small probability for the False cases and a large probability for the True cases. If the model works well, the larger frequencies above the line will appear to the right; larger frequencies below the line will appear to the left.

Use the *Histogram Options* dialog box to enter the number of classes into which the data will be divided, to enter values for the lower and upper limits, and to indicate the type of counts that will be included in the plot. You can also indicate if the current scaling should be retained if you change the values on the Analysis dialog box.

Residual Plots

The Residual Plots option displays three different types of plots: Scatterplots, including Residual versus Predicted, Residual versus Row Number, and Residual versus X; as well as a Normal Probabiity Plot, and an Autocorrelation Function Plot. Use the *Residual Plots Options* dialog box to choose one of the plots, and, if applicable, its options.

Residuals versus Predicted

The Residual versus Predicted scatterplot displays the residual or the studentized residual versus the observed variable (see Figure 8-15). Examine the residuals to look for any unusual patterns. The response is limited to be between 0 and 1, so the residuals will not follow a normal distribution.

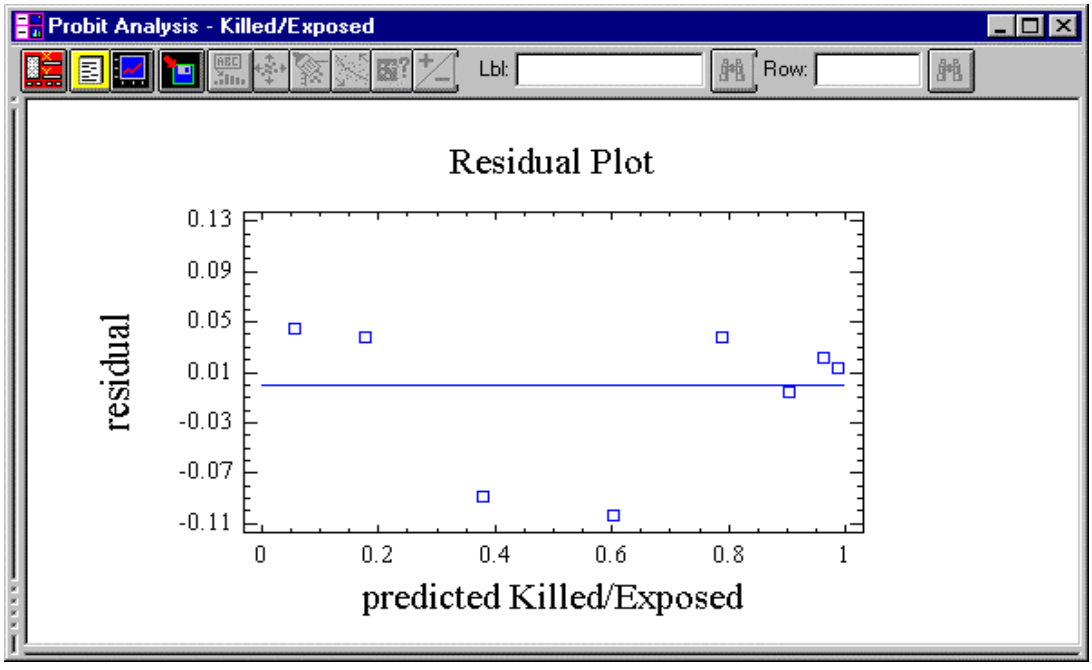


Figure 8-15. Residual versus Predicted Scatterplot

Residual versus Row Number

The Residual versus Row Number scatterplot displays the residual or the studentized residual versus the row number (see Figure 8-16). The program plots the residuals in the order that the observations appear in the dependent variable.

The plot is helpful in determining sequential correlations among the residuals. Any nonrandom pattern indicates serial correlation in the data, particularly if the row number corresponds with the order in which the data were collected.

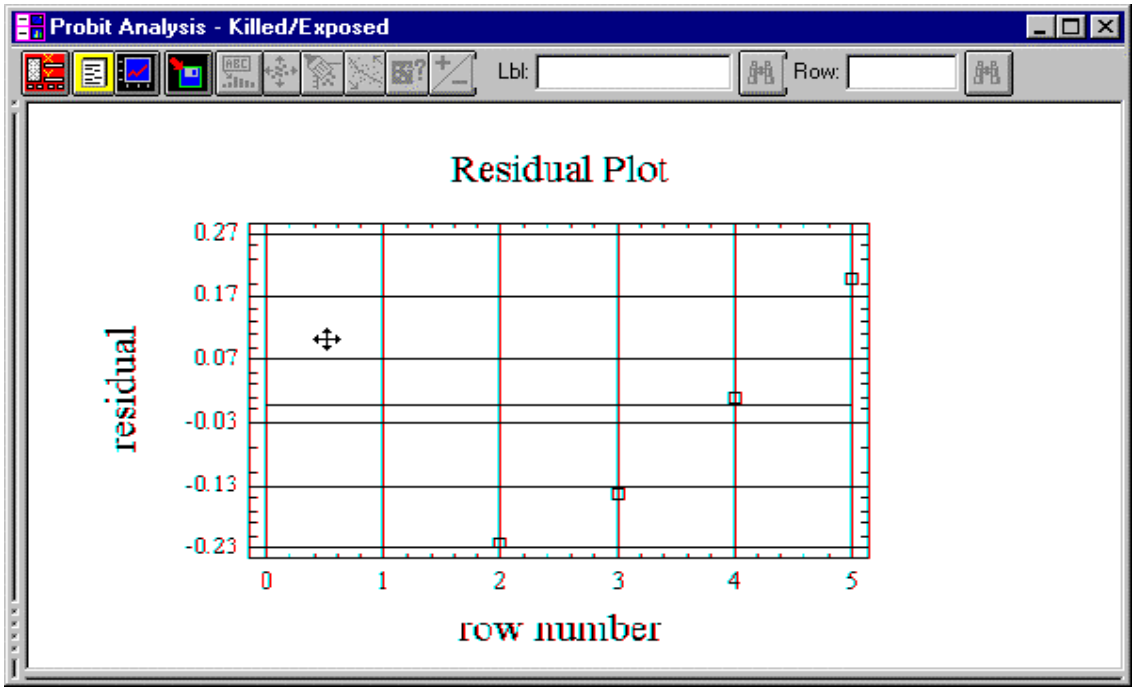


Figure 8-16. Residual versus Row Number Scatterplot

Residual versus X

The Residual versus X scatterplot displays the residual or studentized residual versus the independent variable (X). Use this plot to detect the nonlinear relationship between the dependent and independent variables (see Figure 8-17).

Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may be an outlier.

Normal Probability Plot

The Normal Probability Plot option displays a plot that determines if the errors follow a normal distribution (see Figure 8-18). The program sorts the residuals from smallest to largest, then plots them versus the values. If the data come from a

normal distribution, the points should fall approximately along a straight line. To help determine how close the points are to the line, the program draws a reference line on the plot, which passes through the median with the slope determined from the interquartile range. If the points show a significant curvature, it may indicate that the residuals are not from a normal distribution.

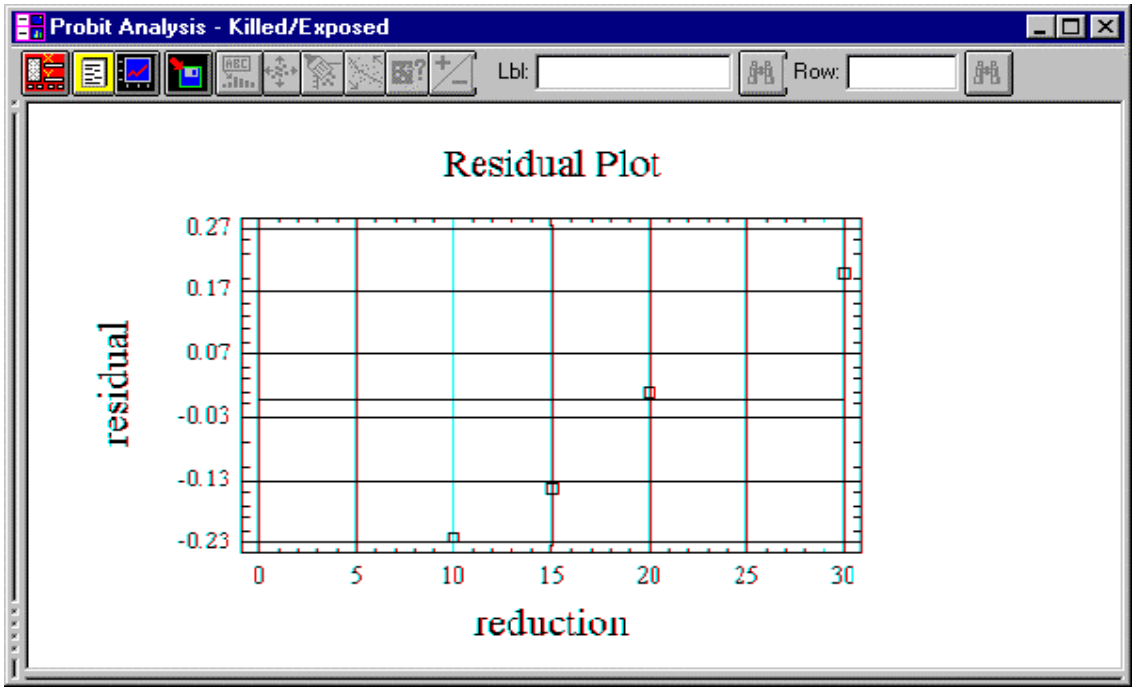


Figure 8-17. Residual versus X Scatterplot

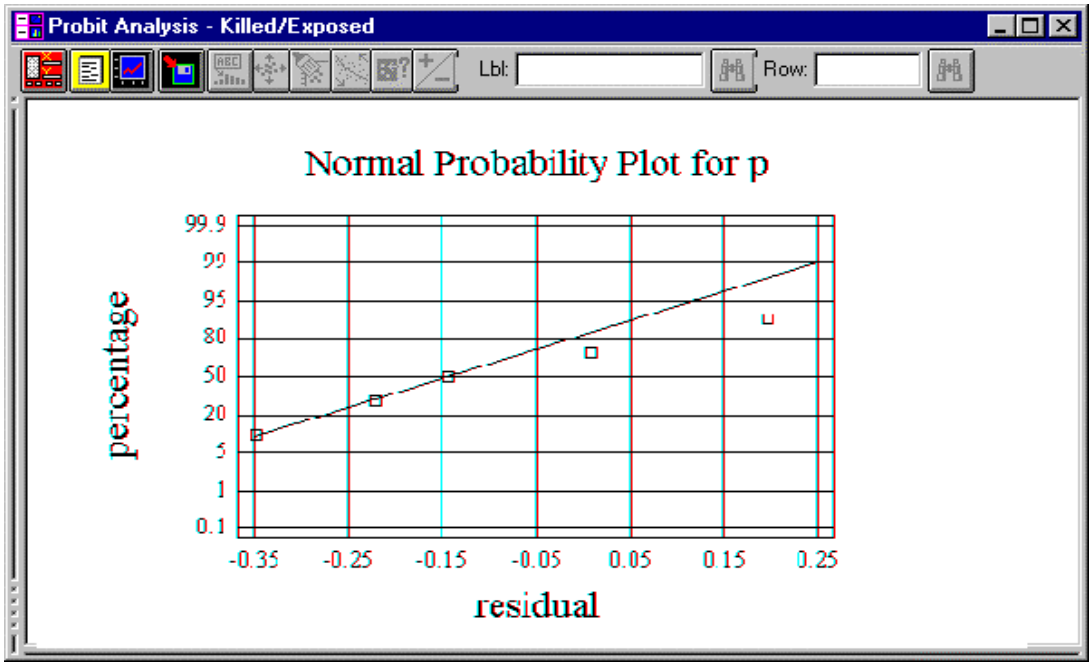


Figure 8-18. Normal Probability Plot

Autocorrelation Function

The Autocorrelation Function option displays a plot of the autocorrelation estimates for the residuals (see Figure 8-19).

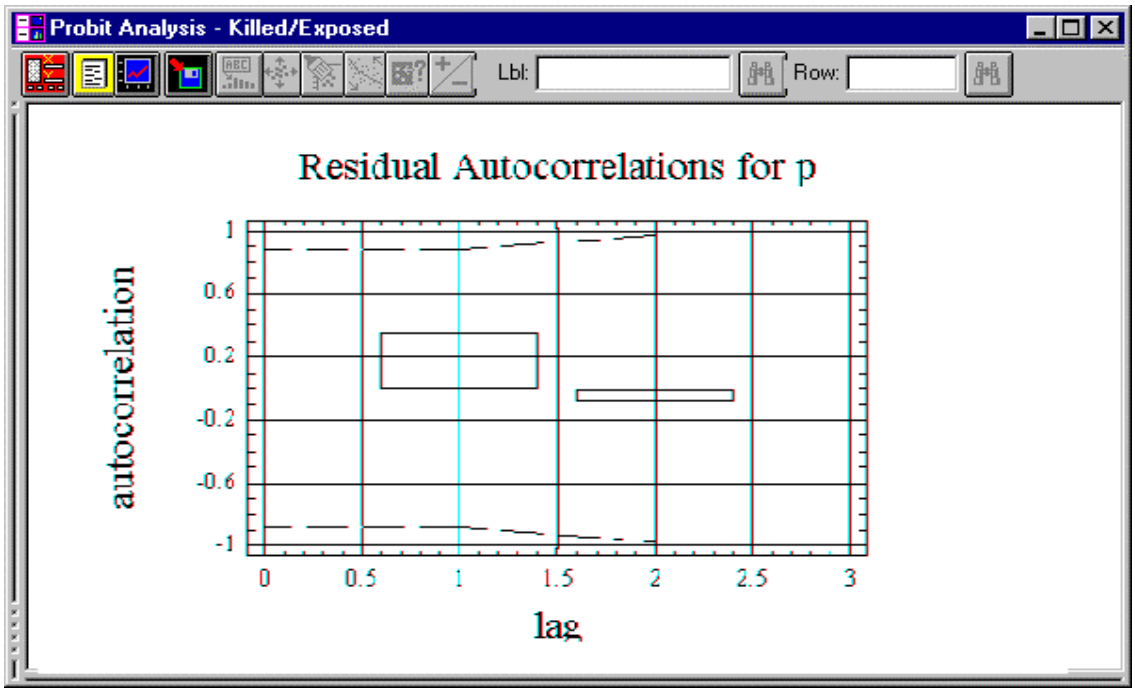


Figure 8-19. Autocorrelation Function Plot

The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

Use the *Autocorrelation Function Plot Options* dialog box to choose the type of residual that will be plotted, to choose the direction of the plot, to choose the type of values that will be used for the fitted line, and to enter values for the maximum lag and confidence level.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 11 selections: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Residuals, Pearson Residuals, Deviance Residuals, Leverages, Percentages, Percentiles, Lower Fiducial Limits, and Upper Fiducial Limits.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Generalized Linear Models, second edition by P. McCullagh and J.A. Nelder. Chapman and Hall, 1989.
Probit Analysis by D.J. Finney. 1971, Cambridge University Press.

Chapter 9

Using Poisson Regression

Background Information

Standard regression models assume that the dependent variable Y is measured on a continuous scale. In fact, many applications exist where Y is a count and thus discrete. If Y represents the number of events observed over a period of length t where events occur at the rate λ, it is common to assume that Y follows a Poisson distribution

$$p(y) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}$$

with mean

$$\mu = \lambda t$$

The rate is then linked to the regressor variables through a loglinear function of the form

$$\lambda = \exp(B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots)$$

The loglinear model assures (among other properties) that the rate is always positive.

Poisson Regression in STATGRAPHICS *Plus*

Data Input

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... POISSON REGRESSION... from the Menu bar to display the Poisson Regression Analysis dialog box (see Figure 9-1).

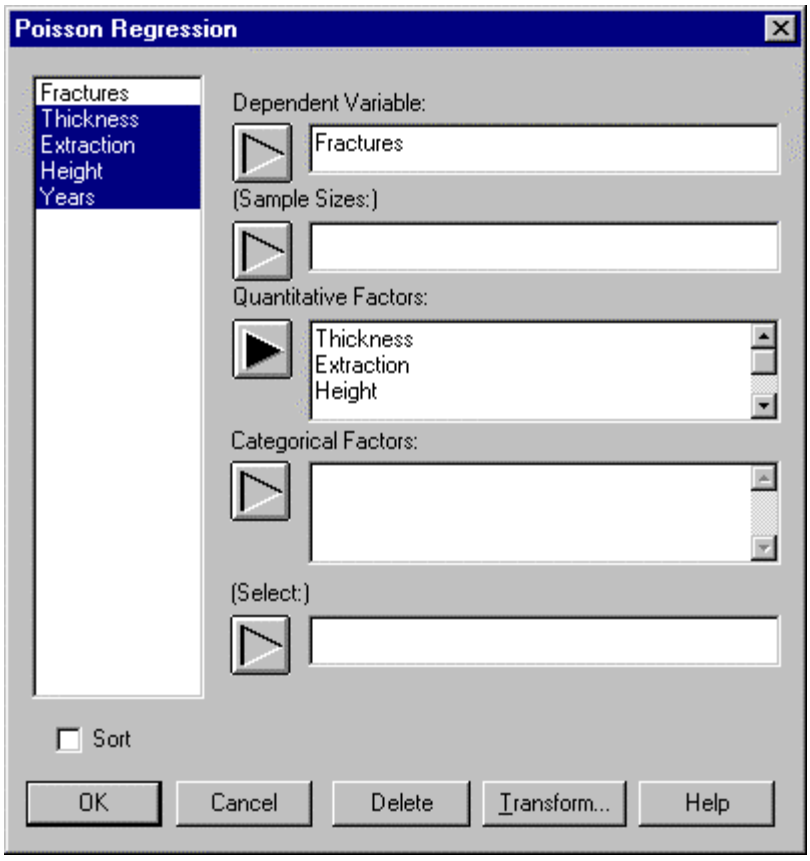


Figure 9-1. Poisson Regression Analysis Dialog Box

The fields include:

- ***Dependent Variable*** - the column name or STATGRAPHICS expression containing Y, which consists of counts.
- ***Sample Sizes*** - an optional name for the column containing the length of the sample periods t_i . If no entry is made in this field, all sample periods are set to the value 1.
- ***Quantitative Factors*** - the names of any quantitative factors to be included in the model.
- ***Categorical Factors*** - the names of any non-quantitative factors to be included in the model.
- ***Select*** - an optional row selection.

Tabular Options

Analysis Summary

The Analysis Summary option displays the results of the fitted model that describes the relationship between the dependent and independent variables (see Figure 9-2).

The results include estimates for each of the coefficients, approximate standard errors, and the estimated rate ratios. If the *p*-value is less than 0.01, there is a statistically significant relationship among the variables. If the *p*-value for the residuals is greater than or equal to 0.10, it indicates that the model is not significantly worse than the best possible model for the data you are currently using.

At the top of table are the estimated model coefficients together with their standard errors. In this case, the fitted model is

$$\log(Y) = -3.79 - 0.00141X_1 + 0.0623X_2 - 0.00208X_3 - 0.0308X_4$$

The output also displays the *estimated rate ratio*, calculated from each estimated coefficient according to

$$\text{rate ratio}_j = \exp(\hat{\beta}_j)$$

Since the log rate is a linear function of X for the Poisson regression model, the rate ratio shows the percentage increase in the rate of events for each unit increase in X. In the current example, a unit increase in *extraction* increases the rate of fractures by slightly more than 6.4%.

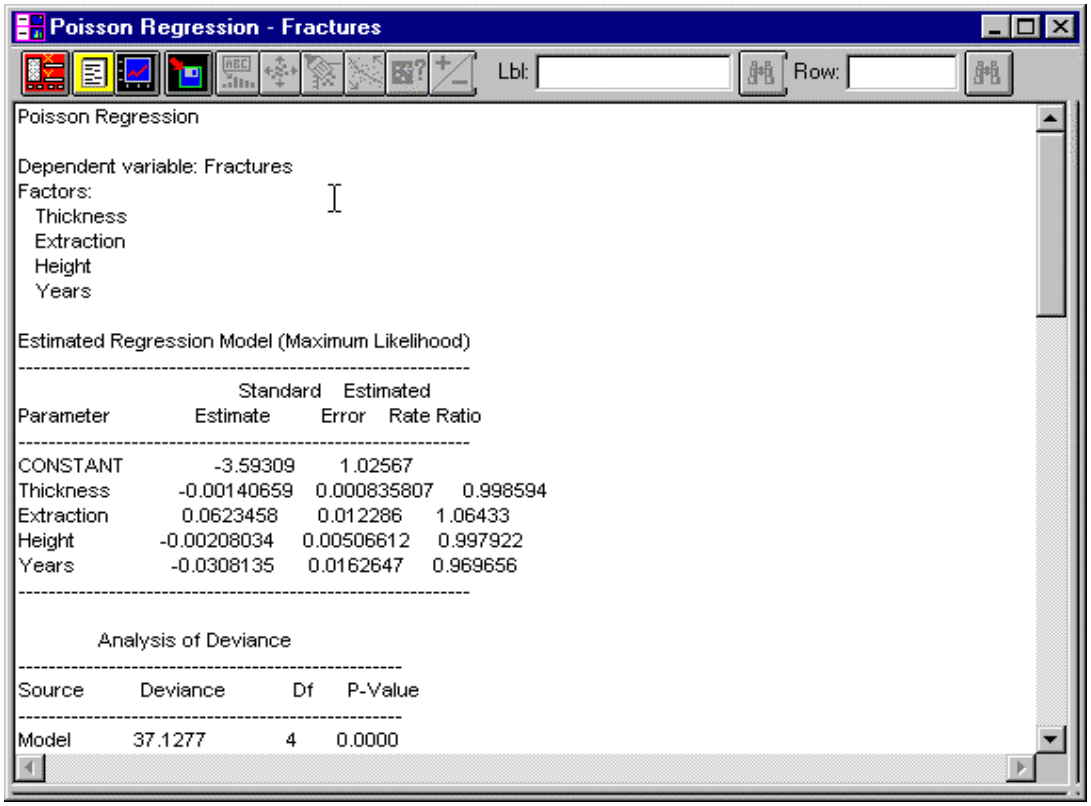


Figure 9-2. Analysis Summary

The *Analysis Options* dialog box controls how the coefficients are estimated (see Figure 9-3).

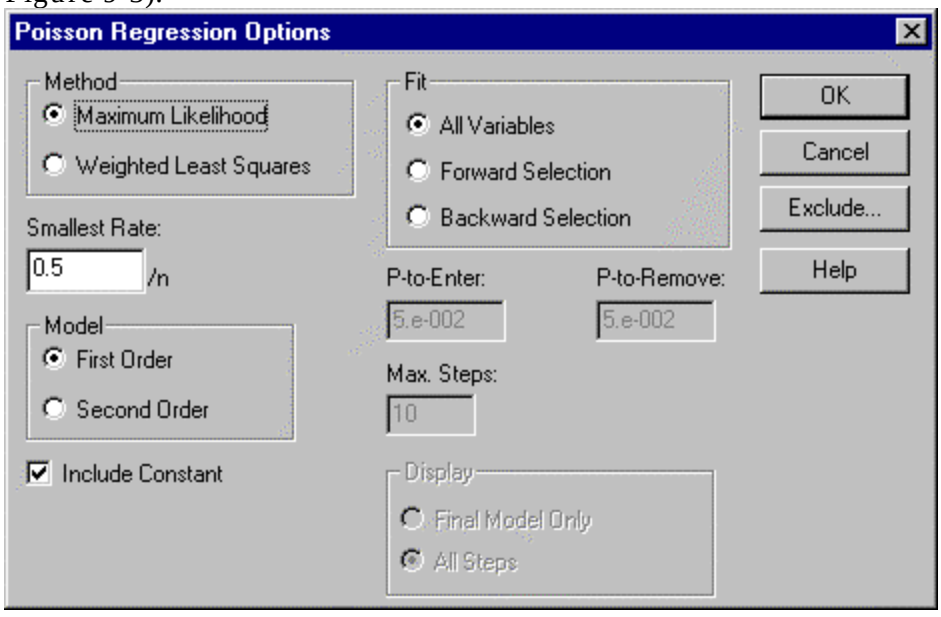


Figure 9-3 Poisson Regression Analysis Options Dialog Box

The fields include:

Method - by default, *maximum likelihood* estimates of the coefficients are obtained using iteratively reweighted least squares. In that case, the standard errors of the model coefficients are obtained from the matrix of second partial derivatives of the log likelihood function. For aggregated data, *weighted least squares* may be selected instead, in which case the model is estimated using ordinary weighted least squares, where the weights are equal to

$$w_i = \frac{1}{y_i}$$

In that case, the standard errors are obtained from the usual weighted regression formulas.

Smallest Rate - when the *weighted least squares* method is selected, responses equal to 0 would cause the weights to be undefined. This field indicates the smallest acceptable rate. By default, all observed counts less than 0.5 are replaced by 0.5. This field is ignored when estimating the model using maximum likelihood.

Model - if *first order* is selected, only linear terms are included in the model. If *second order* is selected, quadratic terms for the quantitative factors and two-factor interactions for both quantitative and categorical factors are also included.

Include Constant - turn off this checkbox to exclude $_0$ from the model.

Fit - this field and the fields below it allow for stepwise selection of variables in a manner similar to the *Multiple Regression* procedure.

In the center of the Analysis Summary is a table showing an *Analysis of Deviance.* The **deviance** of the fitted logistic model is defined by the likelihood ratio statistic

$$\lambda(\beta) = -2\ln\left[\frac{L(\hat{\beta})}{L(Y)}\right]$$

where $L(\hat{\beta})$ is the value of the likelihood function corresponding to the fitted model and $L(Y)$ is the greatest possible value of the likelihood function. The table shows a partition of the deviance in the following manner:

- **Total (corr.)** - the total deviance when using a model involving only a constant term.
- **Residual** - the deviance when using the full model. This value is compared to a chi-square distribution to determine whether the fitted model is significantly worse than the best possible model. A large P-value as in the table above indicates no significant *lack-of-fit* remaining in the residuals.
- **Model** - the reduction in the deviance achieved by adding the selected factors to the model. This value is compared to a chi-square distribution with the indicated number of degrees of freedom to determine whether or not adding the factors to the model significantly reduces the deviance. A small value as in the table above indicates that the factors have a statistically significant effect.

The table also computes the percentage of deviance explained by the model, a statistic similar to R-squared in ordinary regression. It is computed by

$$R^2 = 100\left(1 - \frac{\lambda(residual)}{\lambda(total)}\right)\%$$

An adjusted R-squared is also computed by

$$R^2_{adj} = 100\left[1 - \left(\frac{n-1}{n-p}\right)\frac{\lambda(residual)}{\lambda(total)}\right]\%$$

where p is the number of coefficients in the fitted model. In the current example, the fitted model accounts for approximately 50% of the observed deviance.

To test the significance of each term in the model, likelihood ratio tests are performed. These tests are computed for each factor by removing that factor from the fitted model while leaving all other factors in the model. The test statistic for factor j is the increase in deviance when factor j is removed, i.e.,

$$\lambda(\beta_0, \beta_1, \beta_2, ..., \beta_{j-1}, \beta_{j+1}, ...) - \lambda(\beta)$$

which is compared to a chi-square distribution with 1 degree of freedom. Small P-values for a factor indicate that the factor adds significantly to the fit, as in the current example.

When a factor is categorical, with m levels, m-1 coefficients are added to the model to represent that factor. Indicator variables are used in the same manner as for the *General Linear Models* procedure. The likelihood ratio test for the factor is then based on removing all m-1 coefficients simultaneously.

Confidence Intervals

The *Confidence Intervals* tabular option calculates confidence intervals for the coefficients in the fitted model for the computed rate ratios (see Figure 9-4).

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	-3.59309	1.02567	-5.6677	-1.51848
Thickness	-0.00140659	0.000835807	-0.00309717	0.000283995
Extraction	0.0623458	0.012286	0.0374949	0.0871967
Height	-0.00208034	0.00506612	-0.0123276	0.00816687
Years	-0.0308135	0.0162647	-0.063712	0.00208506

95.0% confidence intervals for rate ratios

Parameter	Estimate	Lower Limit	Upper Limit
Thickness	0.998594	0.996908	1.00028
Extraction	1.06433	1.03821	1.09111
Height	0.997922	0.987748	1.0082
Years	0.969656	0.938275	1.00209

Figure 9-4. 95% Confidence Limits for Estimated Coefficients and Rate Ratios

A normal approximation is used to form the intervals. In the above example, one may conclude with 95% confidence that increasing *extraction* by 1 unit increases the rate by an amount somewhere between 3.8% and 9.1%.

Use the *Confidence Intervals Options* dialog box to enter a number for the confidence level.

Correlation Matrix

The Correlation Matrix option displays a table of the estimated correlations between the coefficients in the fitted model (see Figure 9-5). The correlations are helpful in detecting the presence of serious multicollinearity.

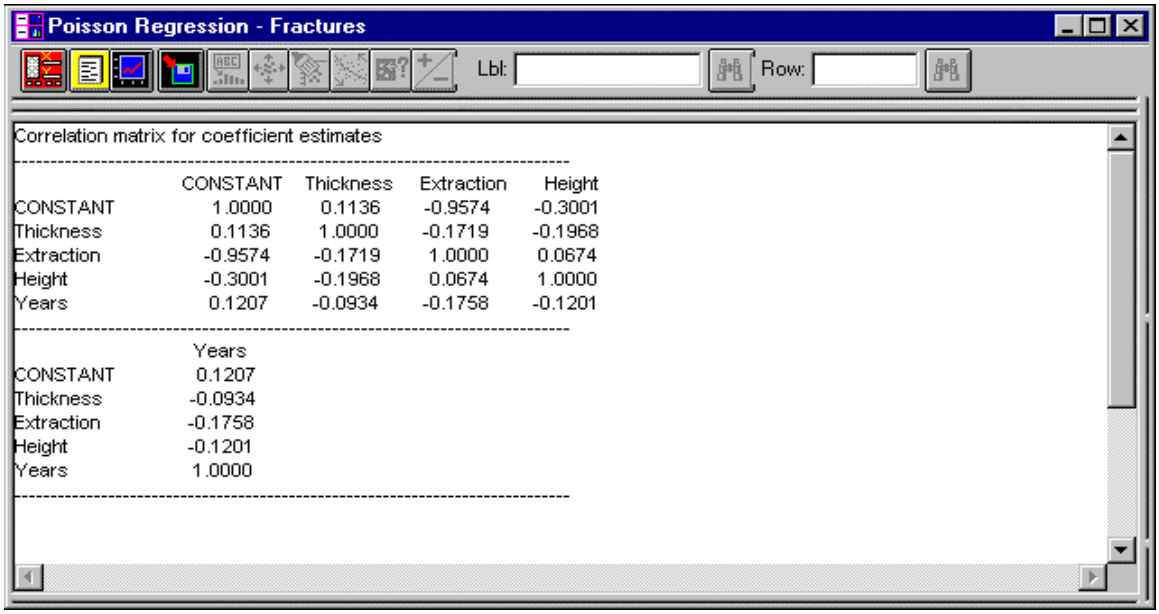


Figure 9-5. Correlation Matrix

Predictions

The Predictions option displays a summary of the prediction capability of the fitted model (see Figure 9-6). The fitted model may be used to predict the dependent variable at each given combination of the experimental factors. Selecting *Predictions* from the list of tabular options creates a table similar to that shown in figure 9-6.

Predictions for Fractures

Row	Observed Value	Fitted Value	Lower 95.0% CL for Prediction	Upper 95.0% CL for Prediction
1	2.0	1.75381	1.2265	2.50783
2	1.0	0.872506	0.521163	1.46071
3	0.0	1.60137	1.11726	2.29525
4	4.0	1.21777	0.763246	1.94297
5	1.0	1.26524	0.812808	1.9695
6	2.0	1.63485	1.15503	2.31398

Figure 9-6. Predictions and Confidence Limits for Mean Rate

Included in the table are:

- the observed values y_i.
- the fitted values $\hat{\mu}_i = t_i \exp(\hat{\eta}_i)$ where $\hat{\eta} = X\hat{B}$.

- upper and lower confidence limits for the mean rate computed from the final weighted least squares fit.

For example, the first mine in the data file had 2 fractures. At the combination of the X variables in that mine, the model predicts a mean of 1.75 fractures, although the mean may well lie anywhere between 1.23 and 2.51. The number of fractures in similar mines would be expected to follow a Poisson distribution with that mean.
Use the *Predictions Options* dialog box to set a range of values for the cutoff, to indicate the values that will be displayed, and to set the percentage for the confidence level that will be used to calculate the confidence intervals.

Unusual Residuals

The Unusual Residuals option displays a table that lists three different types of residuals:
Ordinary residuals:

$$r_i = y_i - \hat{\mu}_i$$

Pearson residuals:

$$\frac{r_i}{\sqrt{\hat{\mu}_i}}$$

Deviance residuals:

$$sign(r_i)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right\}}$$

The *ordinary residuals* quantify the difference between the observed data values and the predictions from the fitted model. The *Pearson residuals* are a form of standardized residual in which each residual is divided by its estimated standard error. The *deviance residuals* measure each observation's contribution to the residual deviance, with their sum of squares equaling the *Residual Deviance* displayed on the *Analysis Summary* pane.
In addition, the *Unusual Residuals* tabular option displays a list of all observations for which the Pearson residual or deviance residual exceeds 2.0 in absolute value (see Figure 9-7)

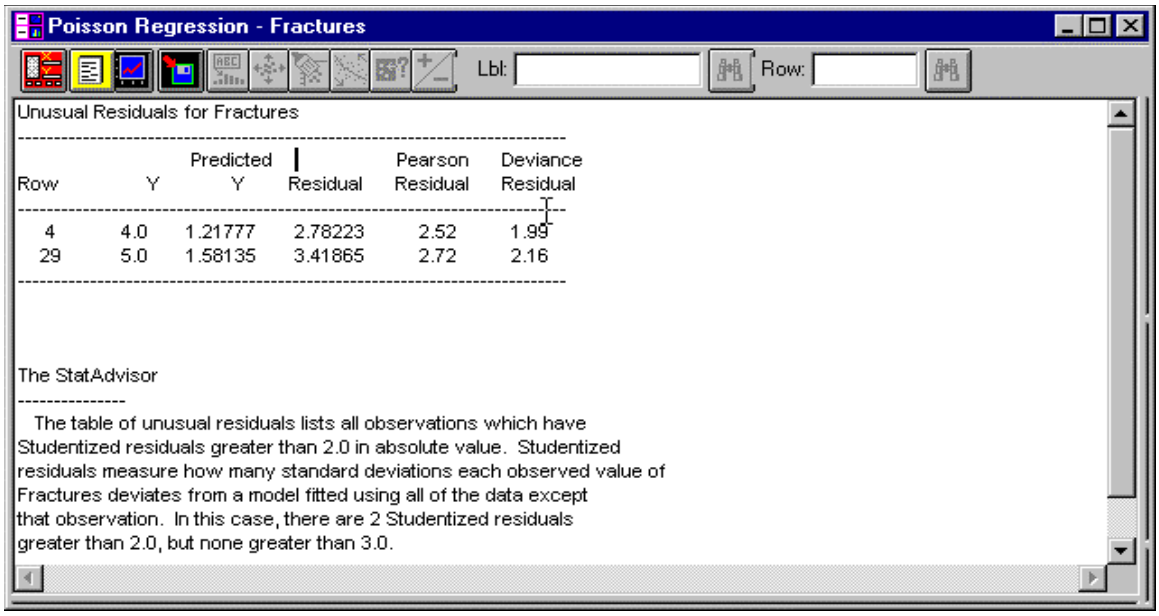


Figure 9-7. Unusual Residuals

Influential Points

The Influential Points option displays a table that lists observations with leverage values greater than three times that of an average point, or with unusually large values for the DFITS or Cook's distance statistics (see Figure 9-8).

Leverage is a statistic that measures the amount of influence each observation has when determining the coefficients for the estimated model. DFITS is a statistic that measures the amount each estimated coefficient would change if each observation was removed from the data. The Cook's distance statistic measures the distance between the estimated coefficients with and without each observation.

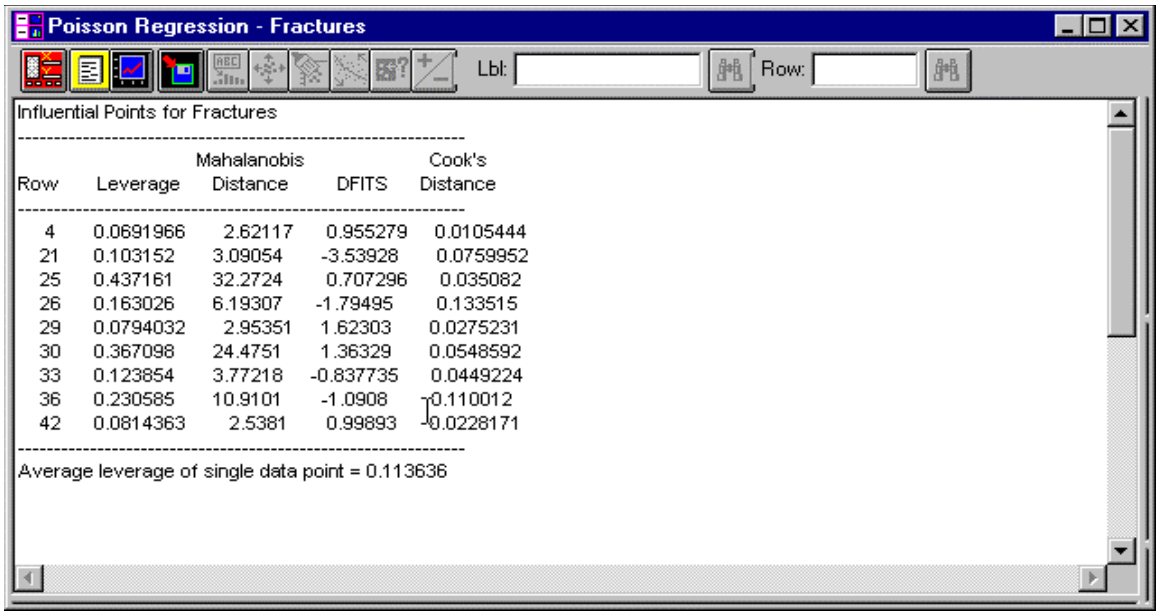


Figure 9-8. Influential Points

Graphical Options

Plot of Fitted Model

The *Logistic Regression* Plot of Fitted Model option displays a plot of the fitted model versus the chosen independent variable (see Figure 9-9). The procedure generates by default a plot of the fitted model versus any selected factor, with all other factors held constant. By default, confidence bands for the fitted model are also displayed, using a normal approximation based on the final weighted least squares results.

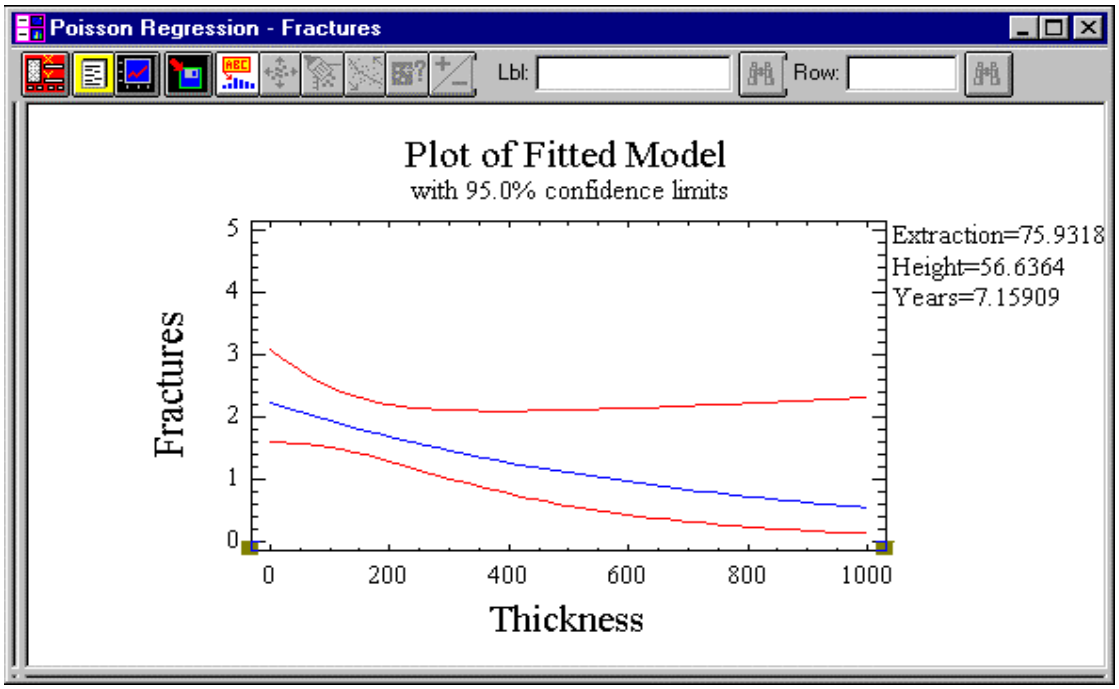


Figure 9-9. Plot of Fitted Model

Use the *Plot of Fitted Model Options* dialog box to choose the variable that will be plotted against the fitted model, and to enter values for the levels for holding the other variables, or new limits for the axis for the chosen variable.

Residuals versus Predicted Plot

The Residual versus Predicted scatterplot displays the residual or the studentized residual versus the observed variable (see Figure 9-10). Examine the residuals to look for any unusual patterns. The response is limited to be between 0 and 1, so the residuals will not follow a normal distribution.

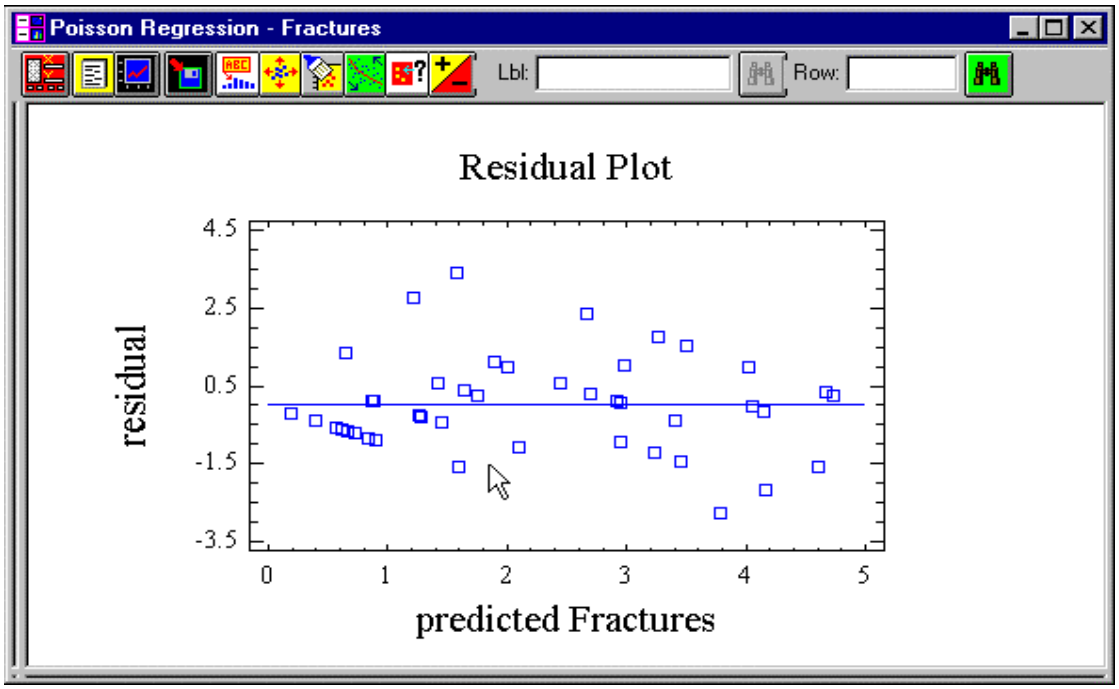


Figure 9-10. ***Residuals versus Predicted Plot***

Observed versus Predicted

The Observed versus Predicted option displays a plot of the residuals from the fitted model plotted against the predicted values of the dependent variable (see Figure 9-11). Use the plot to determine any unusual patterns in the residuals.

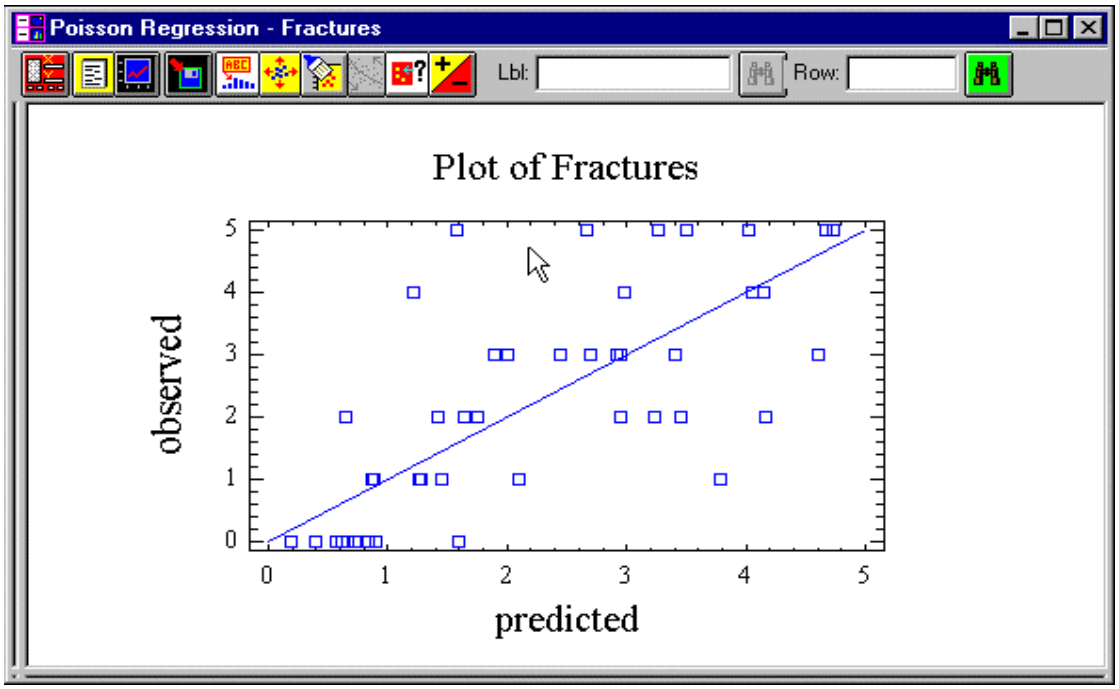


Figure 9-11. Observed versus Predicted *Plot*

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are 8 selections: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Residuals, Pearson Residuals, Deviance Residuals, and Leverages.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

***Classical and Modern Regression with Applications* by Raymond H. Myers, second edition, 1990. PWS-Kent.**
***Generalized Linear Models*, second edition by P. McCullagh and J.A. Nelder. Chapman and Hall, 1989.**
***Applied Regression Analysis and Other Multivariable Methods*, third edition by David G. Kleinbaum, Lawrence L. Kupper, and Keith E. Muller. PWS-Kent, 1997.**
***Regression Analysis of Count Data* by A. Colin Cameron and Pravin K. Trivedi. Cambridge University Press, 1998.**

Chapter 10

Using Life Data Regression

Background Information

In recent years, there has been great interest in the analysis of data where the response variable is time to failure. In medical research, the data may consist of survival times for patients given a clinical treatment. In manufacturing, the data may be the failure times of a manufactured component. In either setting, the data usually contain incomplete or censored measurements, due to removal of patients from the study or study termination before all units have failed. In addition, failure time distributions are rarely normal, usually showing marked skewness.
In this chapter, we consider the problem of modeling failure times when the mean failure time (or log failure time) μ is a function of one or more predictor variables. We will consider both the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \Phi^{-1}(p)\sigma$$

and the loglinear model

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \Phi^{-1}(p)\sigma$$

where Φ(p) is either the standard normal distribution, logistic distribution, or smallest extreme value distribution. As with other models described in earlier chapters, the predictors may be either quantitative or categorical.

Life Data Regression in STATGRAPHICS *Plus*

This procedure is designed for situations when the dependent variable is a failure time. It assumes that the failure times or their log can be expressed as a linear function of one or more independent variables. The data may or may not be censored, defined in the same manner as for the *Distribution Fitting - Censored Data* analysis. As in GLM and Logistic, both quantitative and categorical factors may be included in the model.
The output is similar to that of the logistic procedure, except that no odds ratio is displayed.

To access the analysis, choose SPECIAL... ADVANCED REGRESSION... LIFE DATA REGRESSION... from the Menu bar to display the Life Data Regression Analysis dialog box (see Figure 10-1).

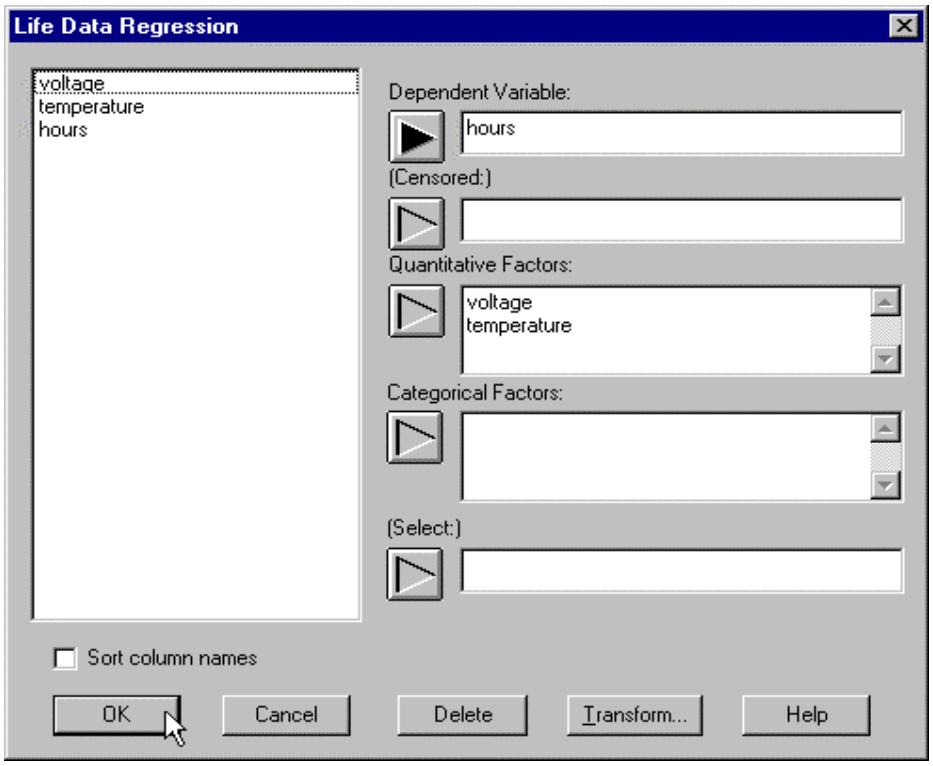


Figure 10-1. Life Data Regression Dialog

The data input fields include:

- **Dependent Variable** - the column name or STATGRAPHICS expression containing Y, the failure times (for uncensored data) or censoring times (for censored data).
- **(Censoring)** - an optional column indicating whether or not each data value has been censored. Enter 0 if the value of the dependent variable represents an uncensored failure time. Enter a number greater than 0 (usually 1) if the value has been right-censored (the true failure time is greater than the value entered). Enter a number less than 0 (usually -1) if the value has been left-censored (the true failure time is less than the value entered).
- **Quantitative Factors** - the names of any quantitative factors to be included in the model.
- **Categorical Factors** - the names of any non-quantitative factors to be included in the model.
- **Select** - an optional row selection.

Tabular Options

Analysis Summary

The *Analysis Summary* pane generated by the *Life Data Regression* procedure shows the fitted model and other summary information (see Figure 10-2). By default, a Weibull distribution is selected. At the top of table are the estimated

model coefficients together with their standard errors and approximate confidence limits. The model coefficients are estimated by maximum likelihood, with standard errors for the model coefficients (including log(σ)) obtained using a normal approximation based on the final information matrix. To test the significance of each term in the model, likelihood ratio tests are performed. These tests are computed for each factor by removing that factor from the fitted model while leaving all other factors in the model. The test statistic for factor j is the decrease in -2 times the log likelihood function when the factor is removed, which is compared to a chi-square distribution with 1 degree of freedom for a quantitative factor or k-1 degrees of freedom for a categorical factor with k levels. Small P-values for a factor indicate that the factor adds significantly to the fit.

```
Life Data Regression

Dependent variable: hours
Factors:
   voltage
   temperature

Number of uncensored values: 32
Number of right-censored values: 0
Number of left-censored values: 0

Estimated Regression Model - Weibull
-------------------------------------------------------------------------------
                                          Standard     Lower 95.0%     Upper 95.0%
Parameter                  Estimate         Error      Conf. Limit     Conf. Limit
-------------------------------------------------------------------------------
CONSTANT                     11.6981         1.9545         7.86738         15.5289
voltage                  -0.00660564     0.000989638     -0.00854529     -0.00466598
temperature               -0.0200546       0.0110578      -0.0417275      0.00161834
SIGMA                       0.312591       0.0481357        0.231154         0.42272
-------------------------------------------------------------------------------
Log likelihood = -211.019

Likelihood Ratio Tests
-----------------------------------------------------------
Factor                  Chi-Square          Df   P-Value
-----------------------------------------------------------
voltage                 29.3505              1    0.0000
temperature             3.06457              1    0.0800
-----------------------------------------------------------
```

Figure 10-2. Analysis Summary

Use the *Life Data Regression Options* dialog box to indicate the type of model that will be used, to choose the distribution, and choose the confidence level used to form the confidence interval for the model parameters (see Figure 10-3).

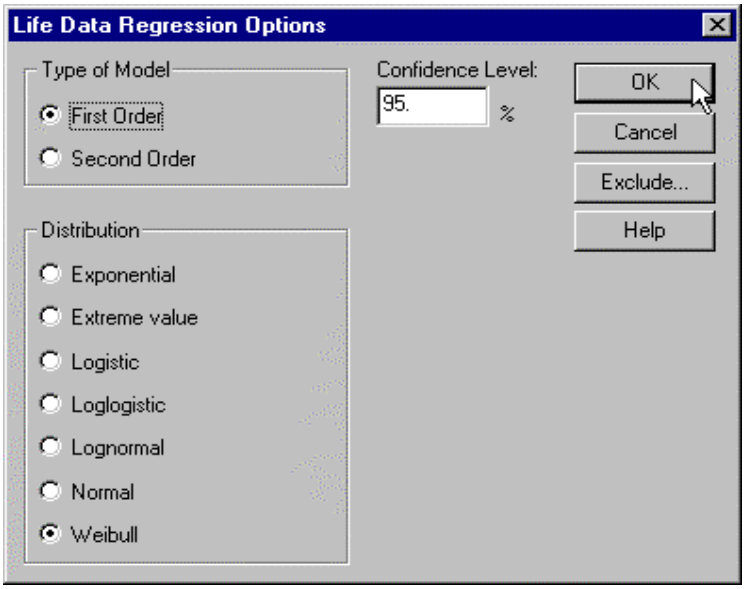


Figure 10-3. Life Data Regression Analysis Options Dialog

Correlation Matrix

The Correlation Matrix option displays a table of the estimated correlations between the coefficients in the fitted model (see Figure 10-4). The correlations are helpful in detecting the presence of serious multicollinearity.

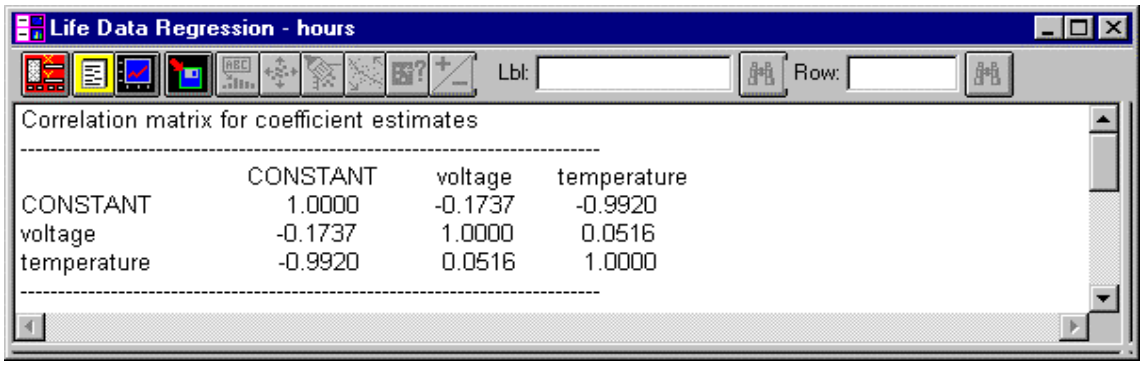


Figure 10-4. Life Data Regression Correlation Matrix

Predictions

The fitted model may also be used to predict the mean value of the failure time distribution corresponding to each observation in the file. Selecting *Predictions* from the list of tabular options displays the observed failure times y_i, the predicted mean failure time $\hat{\mu}_i$ for a linear model, or $\exp(\hat{\mu}_i)$ for a loglinear model and upper

and lower confidence limits for $\hat{\mu}_i$ or $\exp(\hat{\mu}_i)$ based on the estimated variance-covariance matrix of the coefficients (see Figure 10-5).

Predictions for hours

Row	Observed Value	Fitted Value	Standard Error	Lower 95.0% CL for Mean	Upper 95.0% CL for Mean
1	439.0	1061.8	0.10981	856.194	1316.78
2	904.0	1061.8	0.10981	856.194	1316.78
3	1092.0	1061.8	0.10981	856.194	1316.78
4	1105.0	1061.8	0.10981	856.194	1316.78
5	572.0	763.139	0.0853193	645.624	902.044
6	690.0	763.139	0.0853193	645.624	902.044
7	904.0	763.139	0.0853193	645.624	902.044
8	1090.0	763.139	0.0853193	645.624	902.044
9	315.0	548.484	0.0860082	463.397	649.194
10	315.0	548.484	0.0860082	463.397	649.194
11	439.0	548.484	0.0860082	463.397	649.194
12	628.0	548.484	0.0860082	463.397	649.194
13	258.0	394.207	0.111411	316.877	490.408
14	258.0	394.207	0.111411	316.877	490.408
15	347.0	394.207	0.111411	316.877	490.408
16	588.0	394.207	0.111411	316.877	490.408
17	959.0	868.855	0.110577	699.558	1079.12
18	1065.0	868.855	0.110577	699.558	1079.12
19	1065.0	868.855	0.110577	699.558	1079.12
20	1087.0	868.855	0.110577	699.558	1079.12
21	216.0	624.464	0.0864042	527.181	739.699
22	315.0	624.464	0.0864042	527.181	739.699
23	455.0	624.464	0.0864042	527.181	739.699
24	473.0	624.464	0.0864042	527.181	739.699
25	241.0	448.816	0.0871839	378.318	532.45
26	315.0	448.816	0.0871839	378.318	532.45
27	332.0	448.816	0.0871839	378.318	532.45
28	380.0	448.816	0.0871839	378.318	532.45
29	241.0	322.573	0.112398	258.794	402.07
30	241.0	322.573	0.112398	258.794	402.07
31	435.0	322.573	0.112398	258.794	402.07
32	455.0	322.573	0.112398	258.794	402.07

Figure 10-5. Predictions and Confidence Limits for Mean Failure Time

Unusual Residuals

The Unusual Residuals option displays a table that lists all the observations for which the Cox-Snell residuals are less than 0.025 or greater than 0.975. Cox-Snell residuals are scaled to lie between 0 and 1, representing the area in the error distribution below that residual (see Figure 10-6).

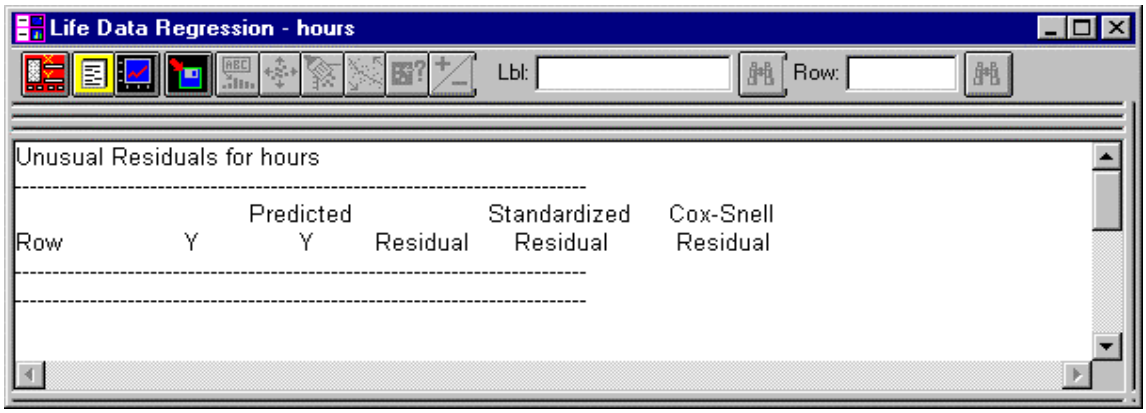
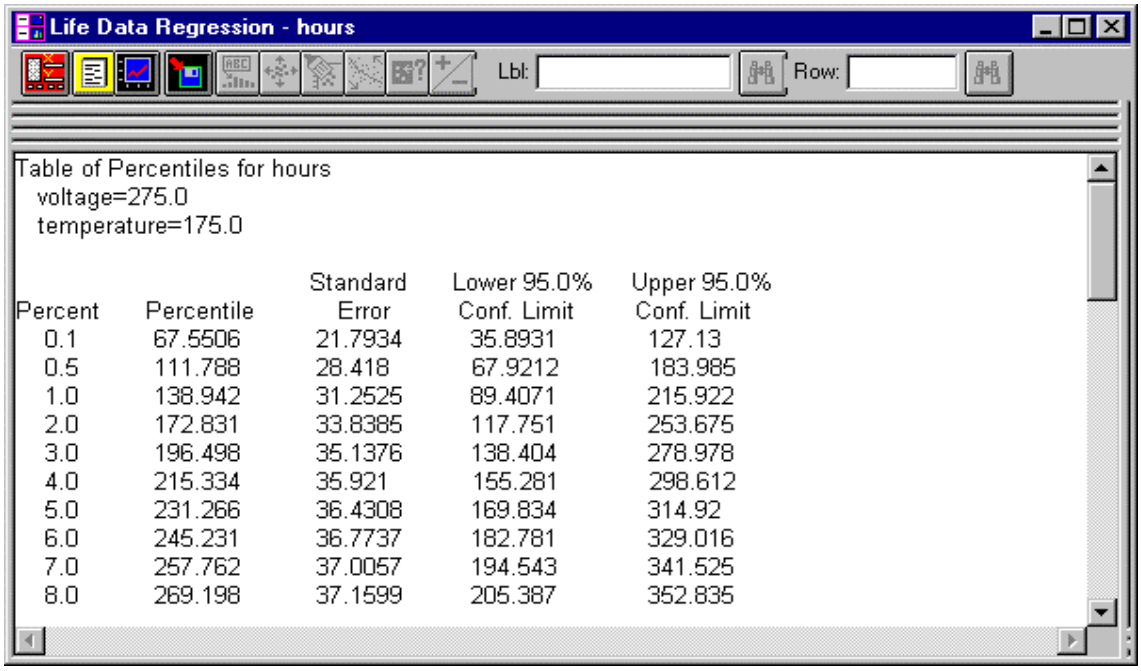


Figure 10-6. Unusual Residuals Table

Percentiles

A table of percentiles may also be displayed by selecting *Percentiles* from the list of tabular options (see Figure 10-7). The percentiles correspond to a specific level of the X variables, which may be changed using *Pane Options* .



Percent	Percentile	Standard Error	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
0.1	67.5506	21.7934	35.8931	127.13
0.5	111.788	28.418	67.9212	183.985
1.0	138.942	31.2525	89.4071	215.922
2.0	172.831	33.8385	117.751	253.675
3.0	196.498	35.1376	138.404	278.978
4.0	215.334	35.921	155.281	298.612
5.0	231.266	36.4308	169.834	314.92
6.0	245.231	36.7737	182.781	329.016
7.0	257.762	37.0057	194.543	341.525
8.0	269.198	37.1599	205.387	352.835

Figure 10-7. Table of Percentiles

Graphical Options

Plot of Fitted Model

The *Life Data Regression* procedure also generates by default a plot of the fitted model versus any selected factor, with all other factors held constant (see Figure 10-8). By default, bounds are placed around the fitted model at the 5th and 95th percentiles of the estimated failure time distribution. Up to four percentiles may be selected using *Pane Options*. Use the Pane Options dialog box to choose the variable that will be plotted against the fitted model; to enter values for upper and lower limits of the axis for the chosen variable; and enter a value that will determine the level at which all of the other variables will be held.

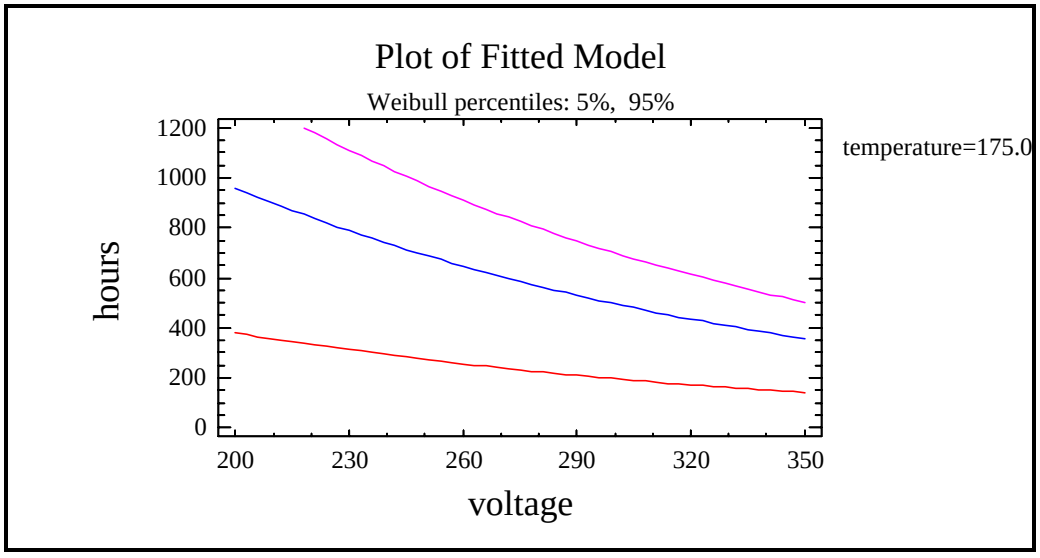


Figure 10-8. Plot of Fitted Model

Percentile Plot

The *Percentile Plot* shows failure probabilities in hours at selected values of the predictive factors. Use Pane Options to change the factor levels (see Figure 10-9).

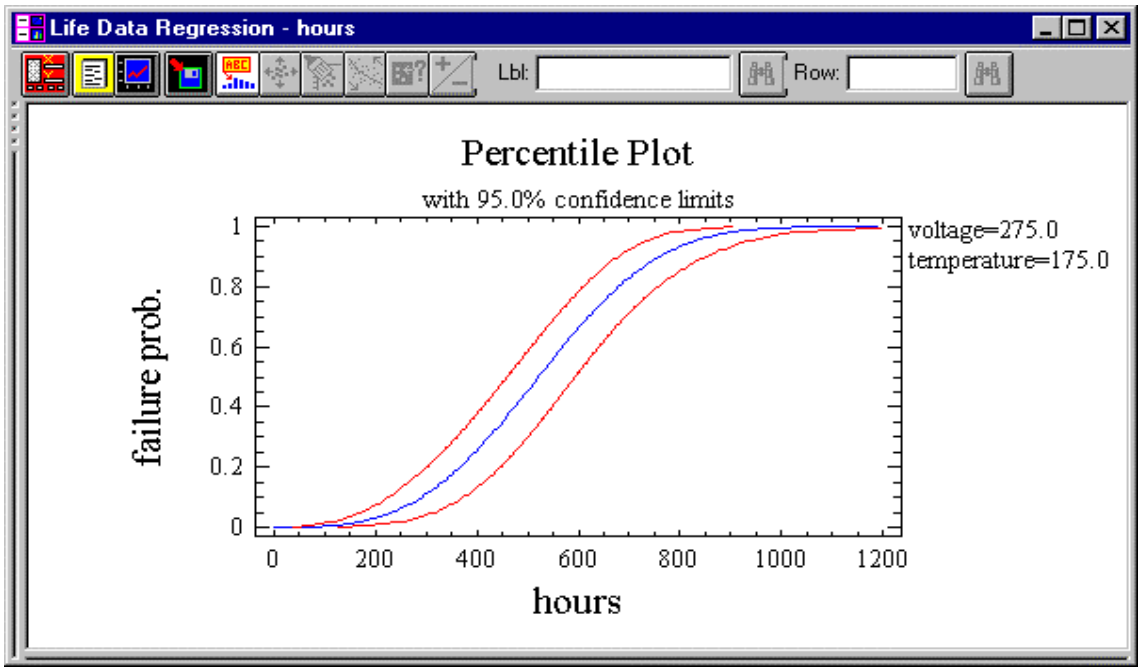


Figure 10-9. Percentile Plot

Percentile Probability Plot

The *Percentile Probability Plot* shows failure probabilities in hours at selected values of the predictive factors (see Figure 10-10). Use Pane Options to change the factor levels.

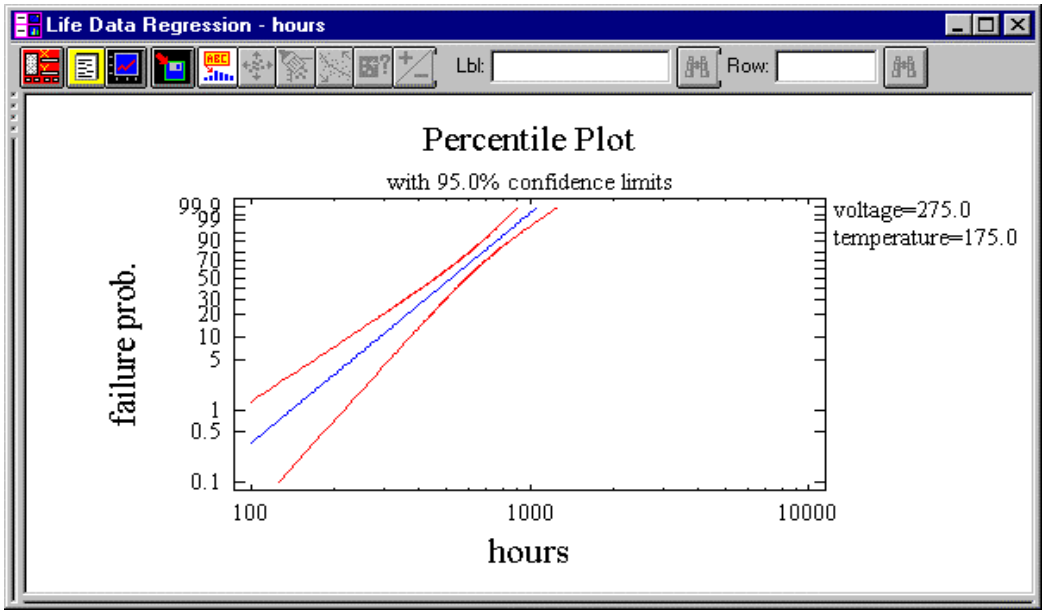


Figure 10-10. Percentile Probability Plot

Observed versus Predicted

The Observed versus Predicted option displays a plot of the observed values versus the values predicted by the fitted model (see Figure 10-11). The plot includes a line with slope equal to one. Points close to the diagonal line are those best predicted by the model.

Use the plot to detect situations in which the error variance is not constant, which indicates that you should probably transform the values for the dependent variable.

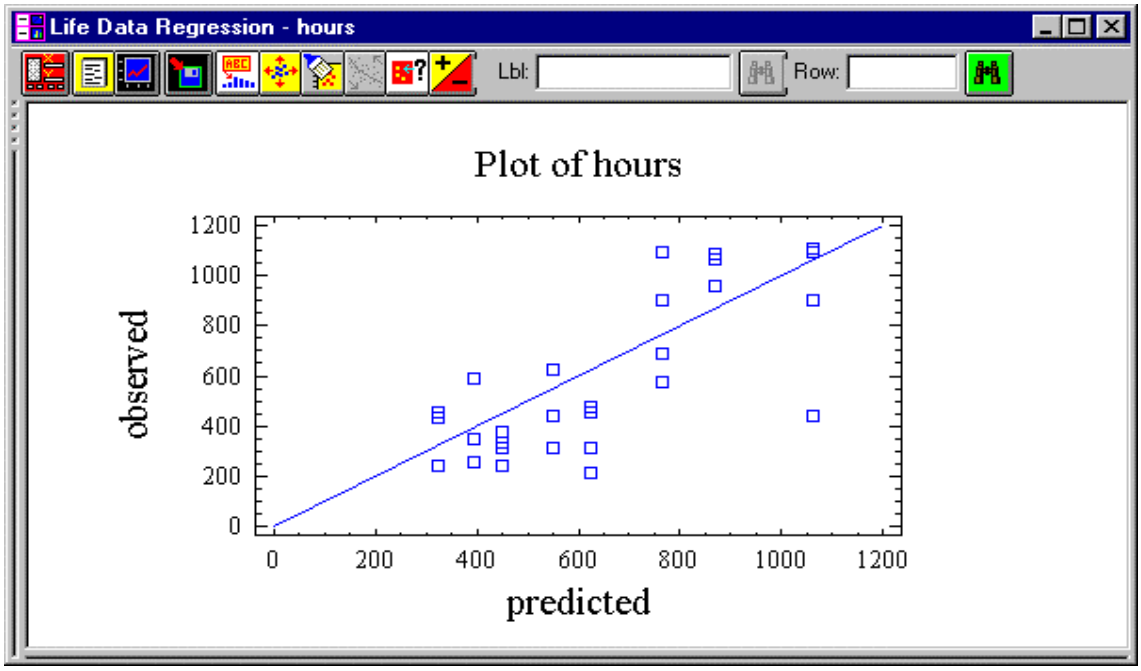


Figure 10-11. Observed versus Predicted

Residual Probability Plot

The Residual Probability Plot option the standardized residuals on a probability scale. If the selected distribution is appropriate, the residuals should fall approximately along a straight line (see Figure 10-12).

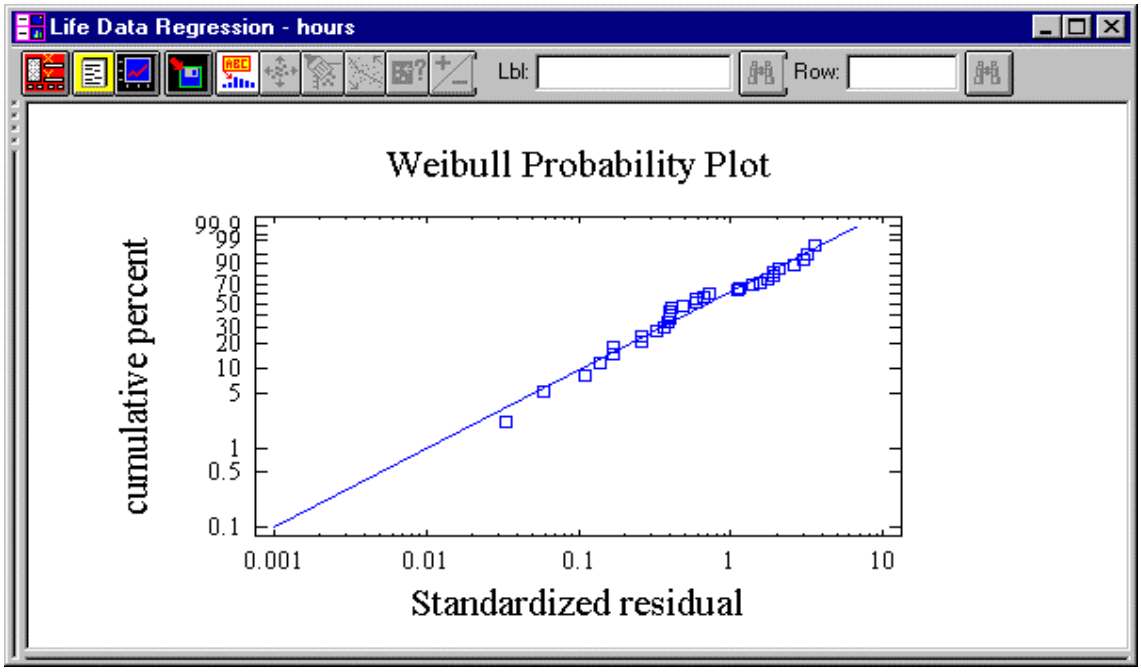


Figure 10-12. Residual Probability Plot

Residual Plots

The Residual Plots option displays one of three different plots: a Scatterplot of the Residual versus the values predicted by the fitted model, the row number, or by X; a Normal Probability Plot; or an Autocorrelation Function.

Use the *Residual Plots Options* dialog box to choose one of the plots and, if applicable, its options.

Residual versus Predicted

The Residual versus Predicted scatterplot displays a plot of the residual or the studentized residual versus the predicted for the observed variable. A nonrandom pattern indicates that the model does not adequately describe the observed data. The plot is helpful in showing heteroscedasticity, an indication that the variability changes as the values of the dependent variable change.

Residual versus Row Number

The Residual versus Row Number scatterplot displays a plot of the residual or the studentized residual versus the row number. Any nonrandom pattern indicates serial correlation in the data, particularly if the row order corresponds to the order in which the data were collected.

Residual versus X

The Residual versus X scatterplot displays the residual or studentized residual versus the independent variable (X). Use this plot to detect the nonlinear relationship between the dependent and independent variables . You can also use the plot to determine if the variance of the residual is constant. Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may an outlier.

You can also use the plot to determine if the variance of the residual is constant. Nonrandom patterns indicate that the chosen model does not adequately describe the data. Any residual value outside the range of -3 to +3 may be an outlier.

Normal Probability Plot

The Normal Probability Plot option displays the residual used to determine if the errors follow a normal distribution. The plot consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as a straight line. The closer the data are to the reference line, the more likely they follow a normal distribution.

Autocorrelation Function Plot

The Autocorrelation Function plot creates a graph of the estimated autocorrelations between the residuals at various lags. The lag *k* autocorrelation of coefficient measures the correlation between the residuals at time *t* and time *t-k*. If the probabiity limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlation at that lag. The plot contains a pair of dotted lines and vertical bars that represent the coefficient for each lag. The distance from the baseline is a multiple of the standard error at each lag.

Significant autocorrelations extend above or below the probability limits. When you choose this option, the Number of Lags and Confidence Level text boxes become active.

Saving the Results

Use the Save Results Options dialog box to choose the results you want to save. There are 13 options: Predicted Values, Standard Errors of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Standardized Residuals, Cox-Snell Residuals, Coefficients, Percentages, Percentiles, Standard Errors of Percentiles, Lower Percentile Confidence Limits, and Upper Percentile Confidence Limits.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results dialog box, click the Save Results Option button on the Analysis toolbar (the fourth button from the left).

References

Statistical Methods for Reliability Data by William Q. Meeker and Luis A. Escobar, Wiley, 1998.

Statistical Models and Methods for Lifetime Data by J.F. Lawless. Wiley, 1982.

Analysis of Survival Data by D.R. Cox and D. Oakes. Chpamna and Hall, 1984.

Practical Methods for Reliability Data Analysis by J.I. Ansell and M.J. Phillips. Clarendon Press, 1994.

Applied Life Data Analysis by Wayne Nelson. Wiley, 1982.

The Statistical Analysis of Failure Time Data by John D. Kalbfleisch and Ross L. Prentice. Wiley, 1980.