TUTORIALS FOR
Advanced Regression Analyses

STATGRAPHICS® *PLUS* FOR WINDOWS

SEPTEMBER 1999

Manugistics
2115 East Jefferson Street
Rockville, Maryland 20852

Introduction

The manual of tutorials for the Advanced Regression analyses in STATGRAPHICS *Plus* is broken into two parts. Part I consists of five individual tutorials that all pertain to the General Linear Models Analysis. Part II consists of six individual tutorials; one each for the remaining analyses.

For information about advanced regression in general, see the section, *Overview of the Model-Building Process*, in Chapter 1 of the online Advanced Regression User Manual as well as the online help system.

Tutorials in Part I

The tutorials for the General Linear Models Analysis are:

- *Using Two Covariates in a Two-Way Analysis of Variance*
- *Using MANOVA and Entering User-Specified Contrasts*
- *Using Nested and Crossed Factors in a Model*
- *Creating and Using a Split Plot Design*
- *Creating and Using Repeated Measures Designs*

TUTORIAL 1

Using Two Covariates in a Two-Way Analysis of Variance

This tutorial uses sample data to illustrate using two covariates in a two-way analysis of variance. The purpose of the tutorial is to:

- illustrate how to use the Select text box on the General Linear Models Analysis dialog box to select or remove a random sample
- create an Interaction Plot
- create a Table of Least Squares Means and a Means Plot
- test for differences among group means.

To begin the tutorial, open STATGRAPHICS *Plus* and the **Cardata** data file.

Completing the Analysis Dialog Box

1. Choose SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display the analysis dialog box.

As you complete the dialog box, you will randomly choose a subset of observations that the program will use to estimate the model; it will use the remaining complete observations to validate the model. You will use the optional Select text box to hold out a random sample of about one-third of the observations.

The two categorical factors you will use are **Year** and **Origin**; **Year** contains five different values while **Origin** contains three. The variables that represent the discrete groups of data are categorical. If a categorical variable has n levels, the program will create $n - 1$ indicator variables.

2. Enter **Mpg** into the Dependent Variables text box.

3. Enter **Year** and **Origin** into the Categorical Factors text box.

4. Enter **Displace** and **Weight** into the Quantitative Factors text box.

These two quantitative factors are the covariates.

5. Click the mouse pointer in the Select text box, then type, ***Random(100)*** (see Figure 1-1).

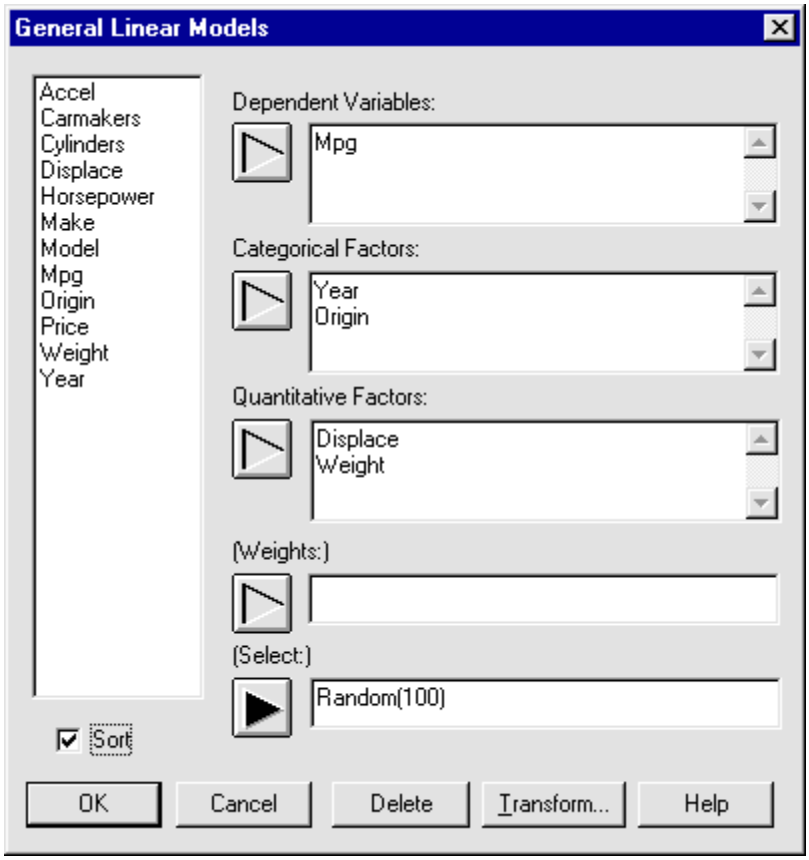


Figure 1-1. Completed General Linear Model Analysis Dialog Box

6. Click OK to display the GLM Model Specification dialog box shown in Figure 1-2.

   Notice that the names of the factors have been given letter designations; for example, A:year. Additionally, the Effects are shown in the Effects list box.

   You are interested in seeing the main effects and the interaction between the **Year** and **Origin** factors, so you will enter A*B into the Effects list box.

7. Click in the Effects text box on the line immediately under the D effect, then type, ***A*B*** (see Figure 1-3).

8. Click OK to display the Analysis Summary and Scatterplot in the Analysis window, then maximize the Analysis Summary (see Figure 1-4).

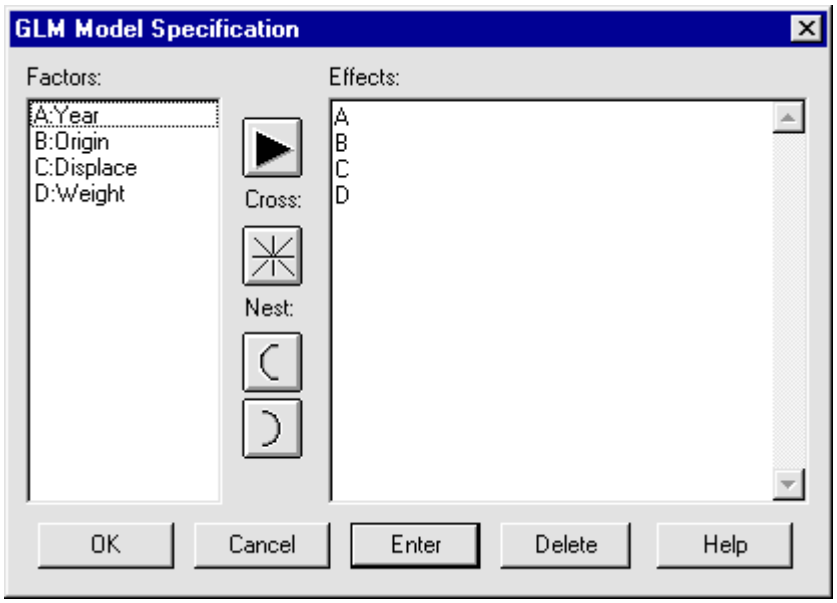


Figure 1-2. The GLM Model Specification Dialog Box

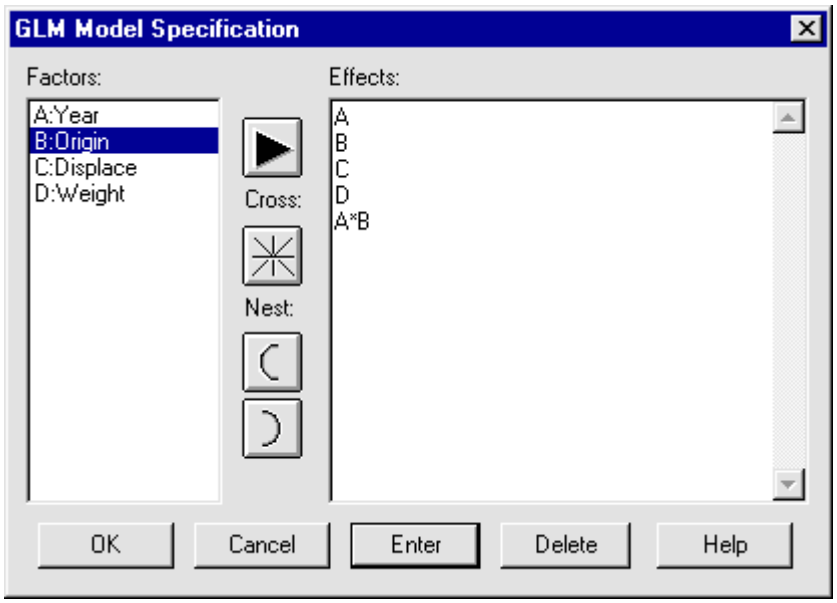


*Figure 1-3. The GLM Model Specification Dialog Box with An Interaction Added (A*B)*

Note: Because the program is randomly selecting observations, the data and their interpretations will differ. The interpretation is provided here to coincide with the results shown in this tutorial.

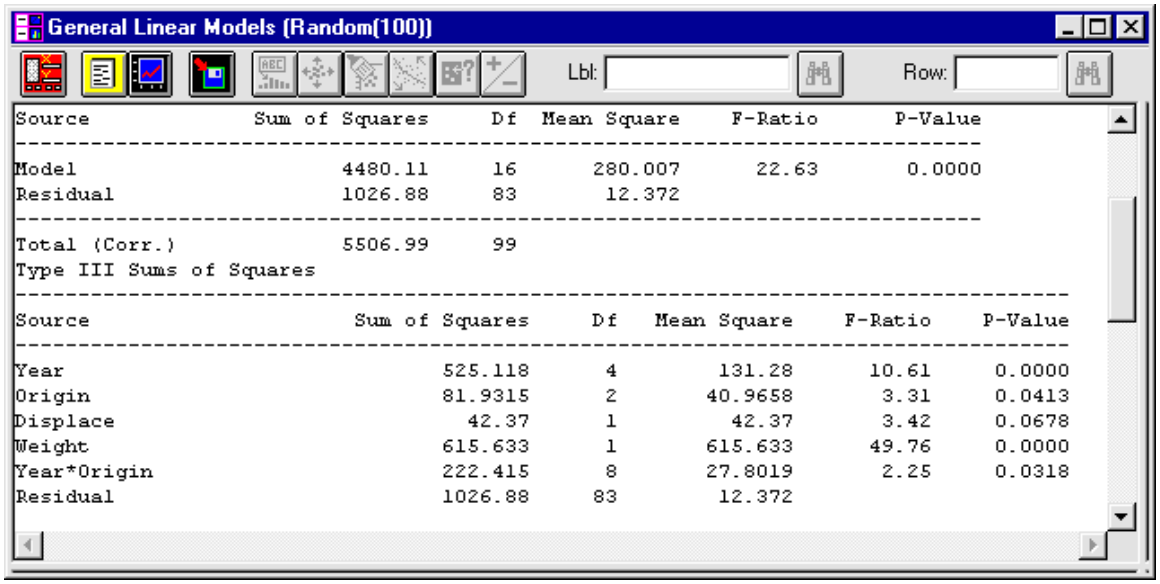


Figure 1-4. The Analysis Summary for a Random Sample

In Figure 1-4, the first ANOVA Table shows that the *p*-value for the **Mpg** variable is less than 0.01, so there is a statistically significant relationship between that variable and the explanatory variables at the 99 percent confidence level.

The second ANOVA in the figure shows the results from testing the statistical significance of each factor in the order the factor was entered into the model.

Notice also that the highest *p*-value in this example is 0.0678 for the **Displace** variable. Because that value is greater than or equal to 0.10, that term is not statistically significant at the 90 percent or higher confidence level, which indicates you should remove it from the model. The validation column on the Residual Analysis Table shows that the results are reasonably consistent with the data you withheld.

Now you will return to the General Linear Models Analysis dialog box and remove the expression from the Select text box.

Removing a Random Sample

1. Click the Return to Analysis Dialog Box button (the left-most button on the Analysis toolbar) to redisplay the General Linear Models Analysis dialog box.

2. Click the mouse pointer in the left-most corner of the Select text box (over the letter r), then hold down the left mouse button and drag the pointer over ***Random(100)*** to highlight it.

3. Click the Delete button, then OK to display the GLM Model Specification dialog box.

4. Click OK to display the Analysis Summary and Scatterplot in the Analysis window, then maximize the Analysis Summary to see it without the random sample (see Figure 1-5).

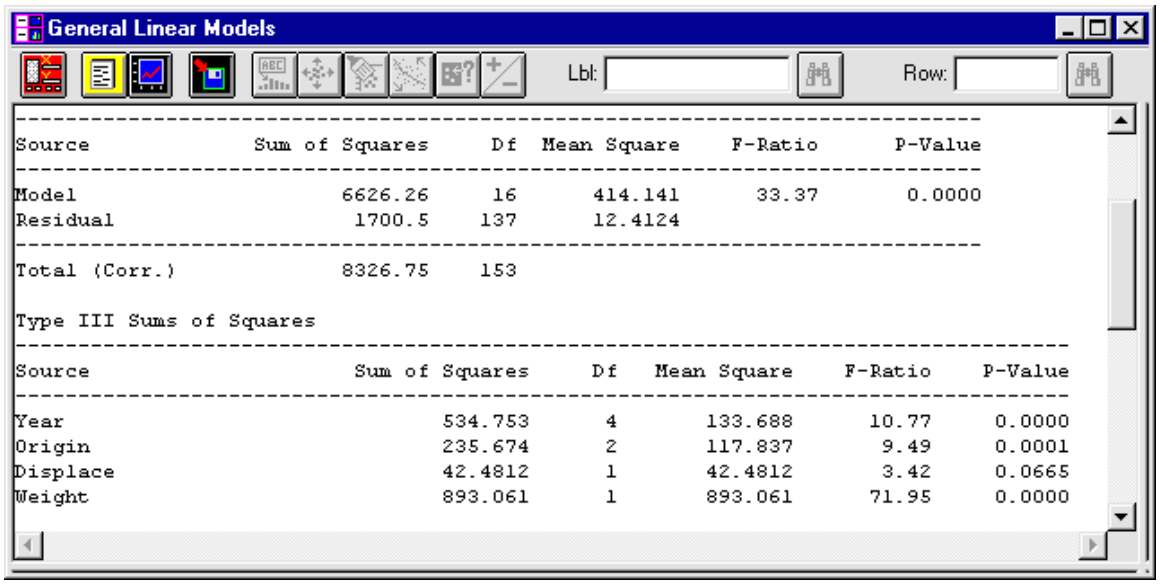


Figure 1-5. The Analysis Summary with the Random Sample Removed

Using the full set of data, the **Displace** variable is significant at the 90 percent confidence level, but not at 95 percent.

Creating an Interaction Plot

1. Click the Graphical Options button to display the dialog box, then click the Interaction Plot check box, and OK to display the plot in a graphics pane. Maximize the plot (see Figure 1-6).

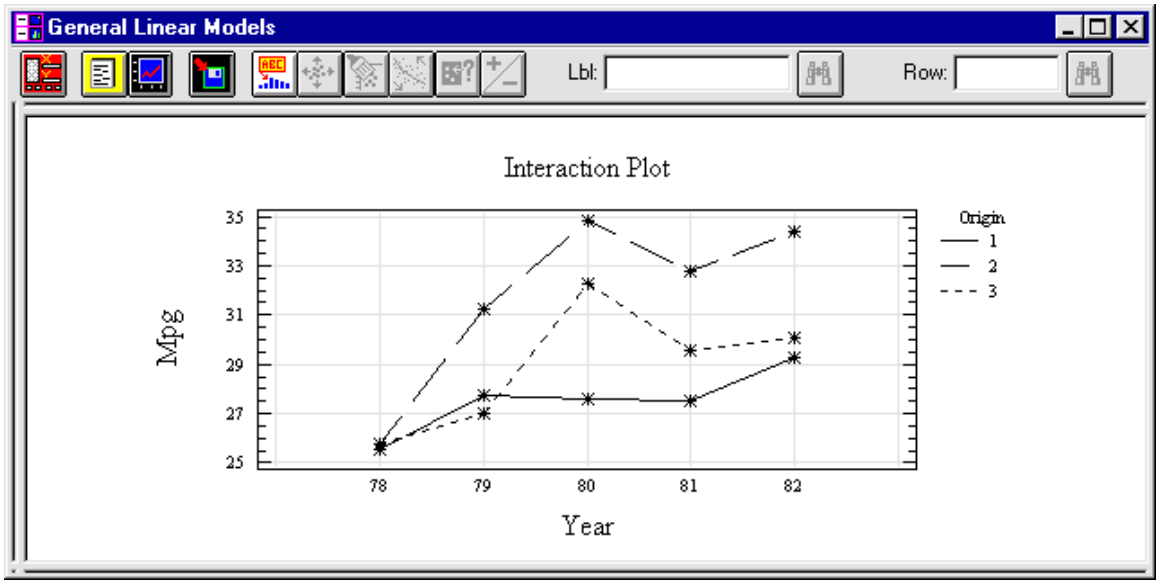


Figure 1-6. The Interaction Plot

The plot contains one line for each level of **Year**, which illustrates the change in estimated miles per gallon ratings over five years. The three lines rise and fall together, which confirms that the interaction effect is not strong.

Creating a Table of Least Squares Means and a Means Plot

The data are made up of different observations in various combinations of unbalanced data. The best estimates of marginal means in unbalanced data are known as least squares means.

You will create a Table of Means and a Means Plot to see the values for the least squares means.

1. Click the Tabular Options button, then the Table of Means check box, and OK to display the table in the second text pane.

2. Click the Graphical Options button, then the Means Plot check box, and OK to display the plot in a graphics pane.

3. Maximize the Table of Means (see Figure 1-7).

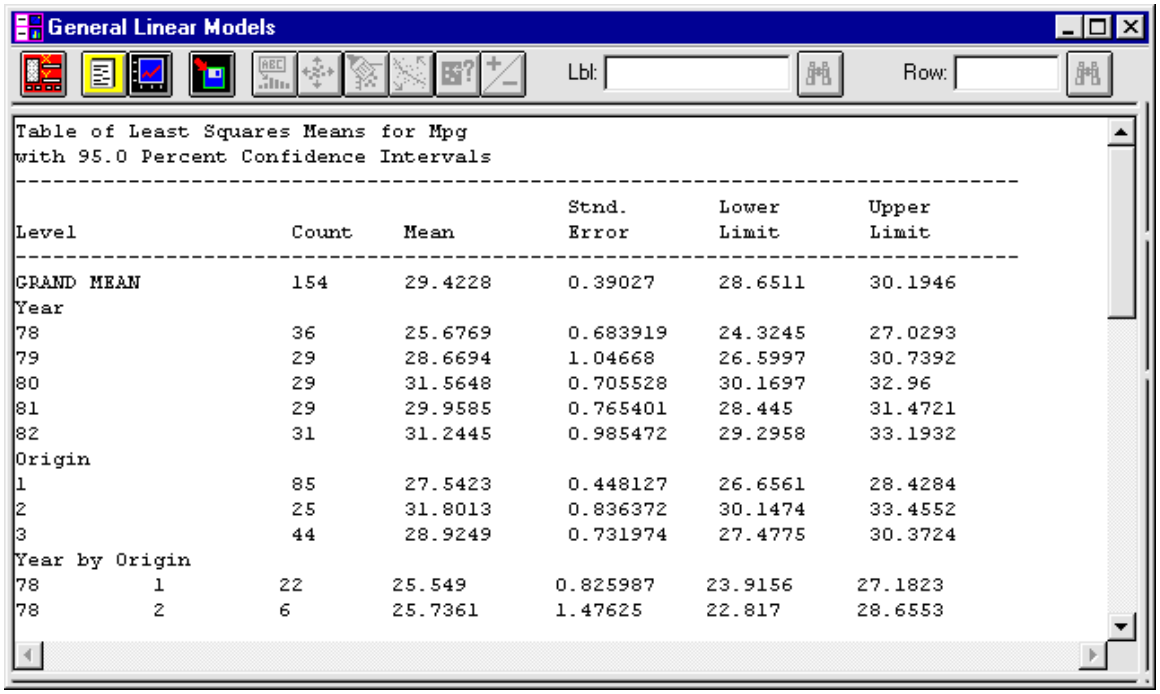


Figure 1-7. The Table of Least Squares Means for ***Mpg***

The least squares means in the table will differ from the simple means for each group. Milliken and Johnson (1984) in their book, *Analysis of Messy Data*, provide detailed explanations about why least squares means are preferred for analyzing unbalanced designs.

4. Minimize the Table of Means and maximize the Means Plot (see Figure 1-8).

The plot substantiates the results shown in the Table of Least Squares Means for the **Year** factor. To see the Means Plot for the **Origin** factor, use the Means Plot Options dialog box.

5. Click the right mouse button on the graphics pane, then the left on Pane Options to display the Means Plot Options dialog box.

6. Click the **Origin** factor to highlight it, accept the defaults for the remainder of the options on the dialog box, and click OK to redisplay the plot.

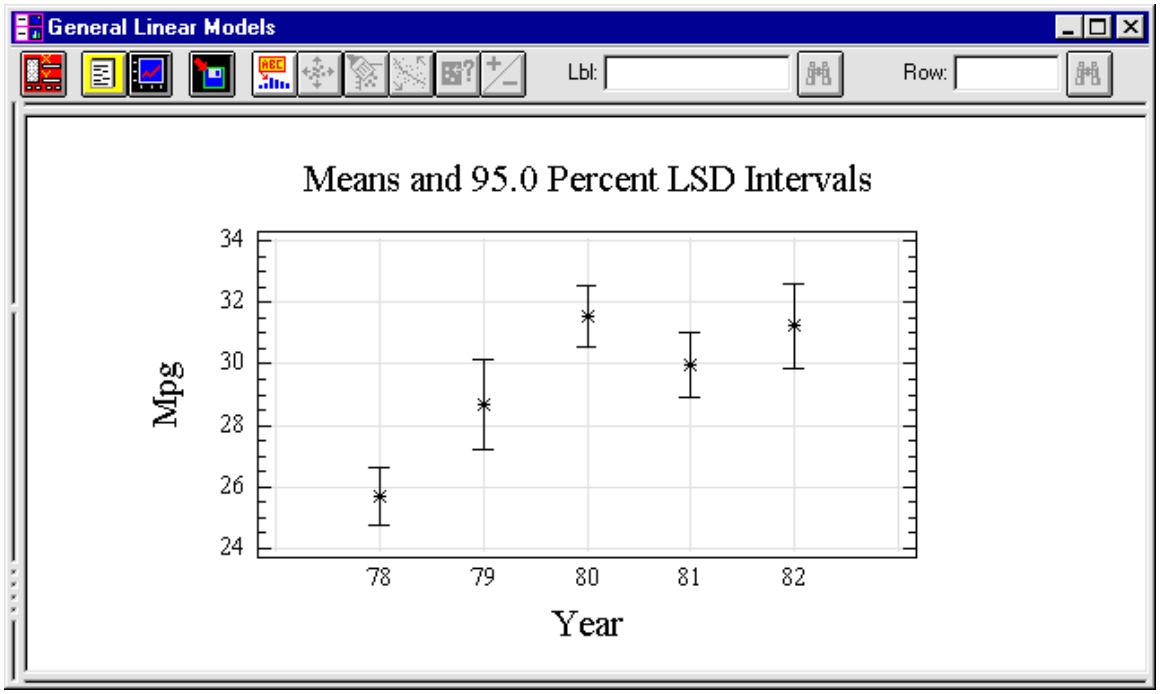


*Figure 1-8. The Means Plot for the **Year** Variable*

Testing for Differences Among Group Means

You can also test for significant differences among group means. This is done for the various levels of each categorical factor.

1. Click the Tabular Options button to display the dialog box, then click the Multiple Range Tests check box, and OK to display the table in the third text pane.

2. Maximize the pane (see Figure 1-9).

To calculate the results, a multiple comparison analysis is applied to the data to determine which means are significantly different. The top portion of the table identifies the homogenous groups by using columns of Xs. Within each column, the levels containing Xs form a group of means for the statistically significant differences, if any.

The bottom portion of the table shows the estimated differences between each pair of means. An asterisk identifies each statistically significant pair.

Now you will use Pane Options to compare the means among the three origins using Scheffe intervals.

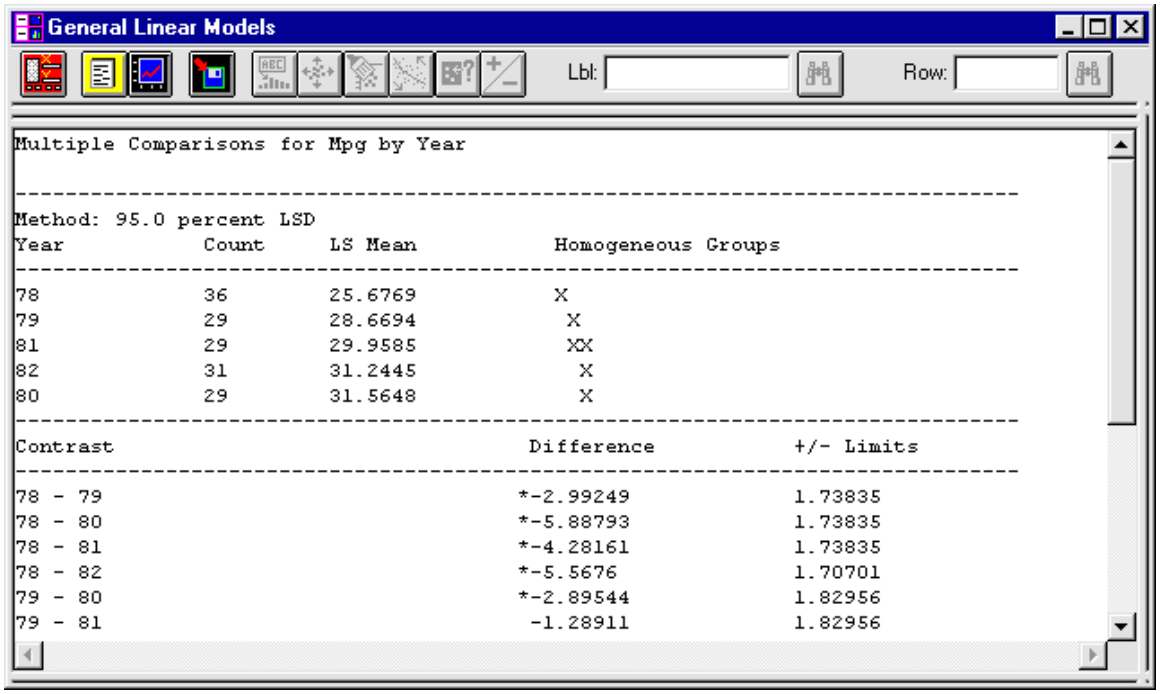


*Figure 1-9. Multiple Comparisons for **Mpg** by **Year***

3. Click the right mouse button on the text pane, then the left on Pane Options to display the Multiple Comparisons Options dialog box.

4. Click the Origin factor to select it.

5. Click the Scheffe option and accept the defaults for the remainder of the options (see Figure 1-10).

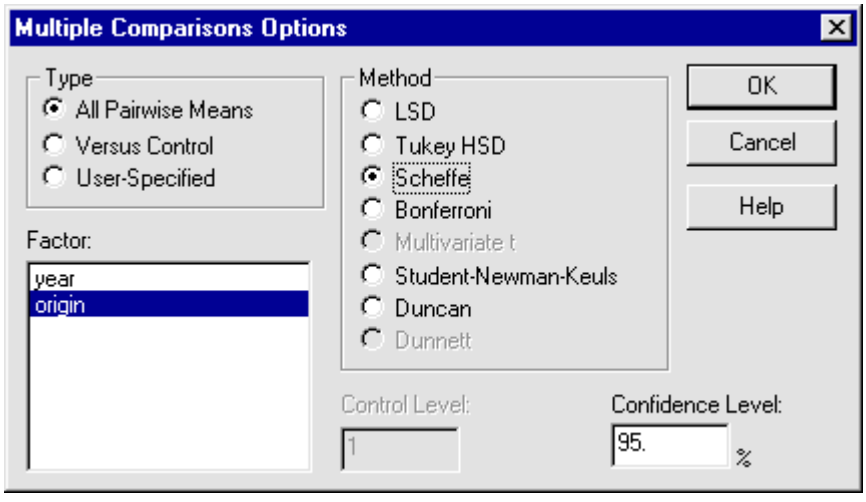


Figure 1-10 Completed Multiple Comparisons Options Dialog Box

6. Click OK to redisplay the table (see Figure 1-11).

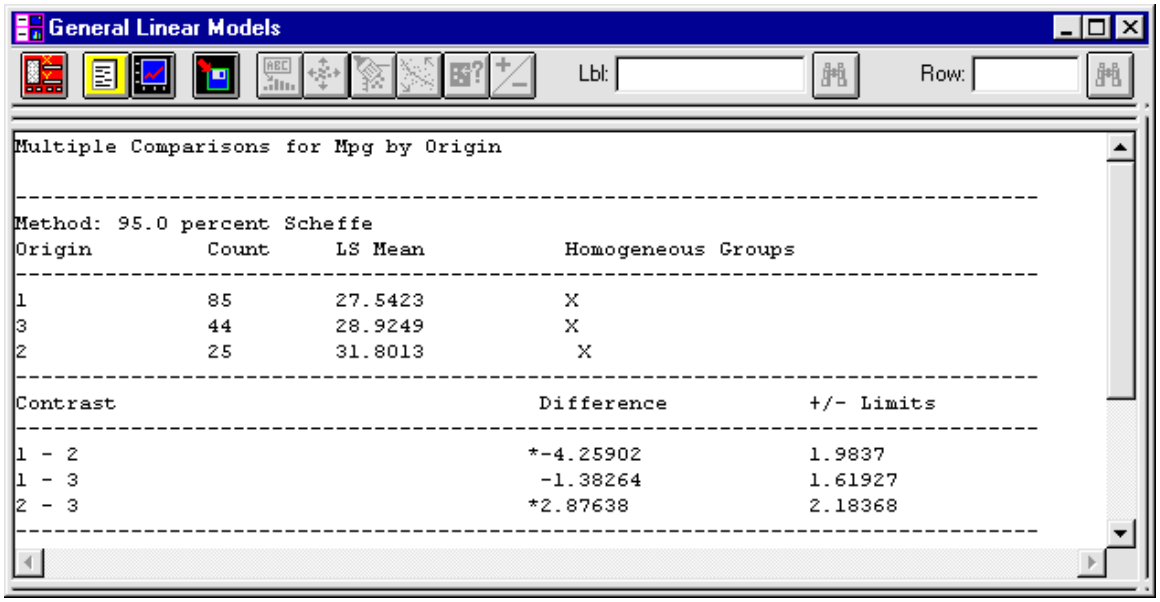


*Figure 1-11. Multiple Comparisons Tests for **Mpg** by **Origin***

The table shows that when you use the Scheffe Method for the comparison and group the **Mpg** factor with the **Origin** factor, they become two homogenous groups.

Optional Exercise

For an optional exercise, continue the tutorial above, creating an Unusual Residuals Table, which will confirm that the model, although good for illustrative purposes, is not necessarily, the best one to use.

References

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data.* Volume 1. *Designed Experiments*. New York: Van Nostrand Reinhold.

TUTORIAL 2

Using MANOVA and Entering User-Specified Contrasts

The purpose of this tutorial is to introduce you to MANOVA in the General Linear Models Analysis and to show you how to enter user-specified contrasts.

You will use data collected and adapted from soil samples taken from four different locations in California. Each location was sampled at 12 different depths and the percentage of sand, silt, and clay was determined for each sample.

There are three response variables: **Sand**, **Silt**, and **Clay**. For the first portion of the tutorial, you will focus only on **Sand**. Later you will apply the MANOVA capabilities in STATGRAPHICS *Plus* to all three variables.

Before you begin, open STATGRAPHICS *Plus* and the **Soil** data file.

Completing the Analysis Dialog Box

1. Choose SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display the dialog box.

2. Enter **Sand** into the Dependent Variables text box.

3. Enter **Location** and **Depth** into the Categorical Factors text box (see Figure 2-1).

4. Click OK to display the GLM Model Specification dialog box.

5. Accept the defaults and click OK to display the Analysis Summary and Scatterplot in the Analysis window, then maximize the Analysis Summary (see Figure 2-2).

   The results show a summary of fitting a general linear statistical model that relates the **Sand** variable to two predictive factors. The first ANOVA Table shows that the *p*-value is less than 0.01, which indicates that there is a statistically significant relationship between **Sand** and the predictor variables at the 99 percent confidence level.

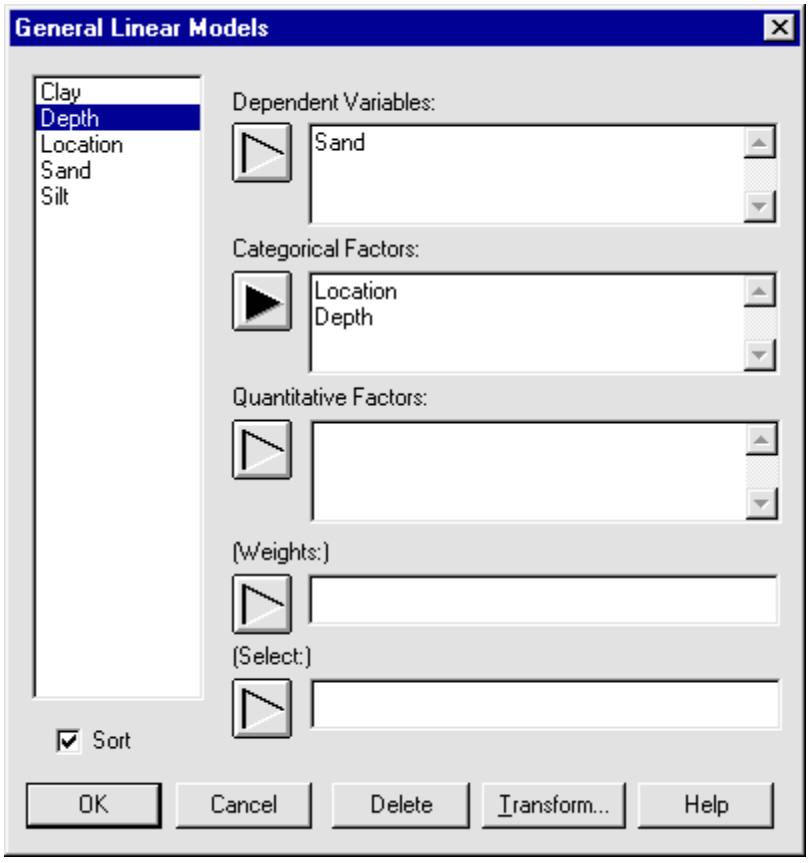


Figure 2-1. Completed General Linear Models Analysis Dialog Box

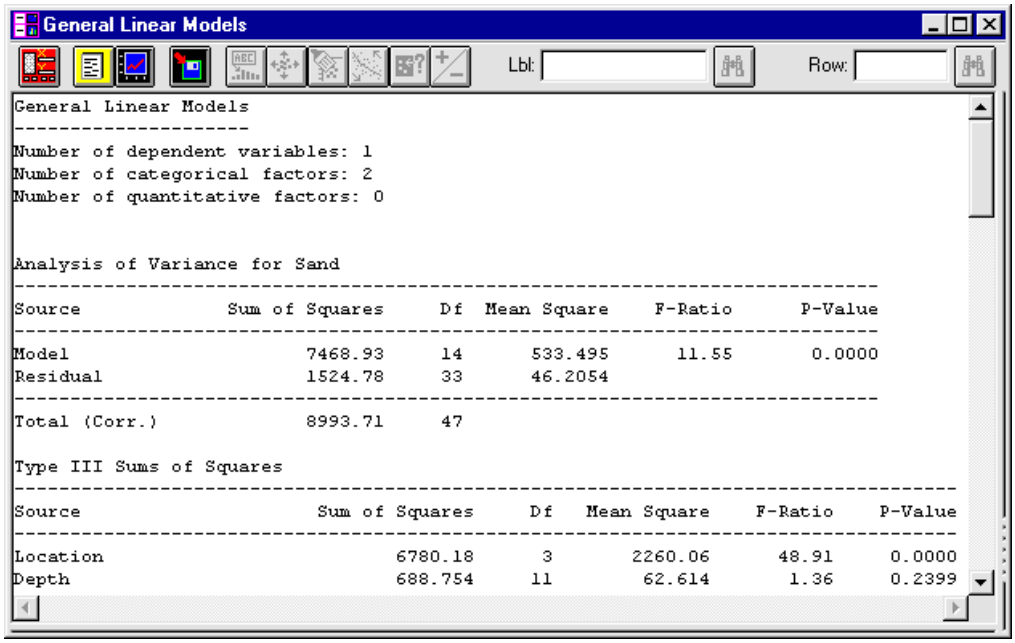


Figure 2-2. The Analysis Summary

The second ANOVA Table shows the results of testing the statistical significance of each factor as it was entered into the model. The highest *p*-value is 0.2399, which corresponds to the **Depth** variable. Because the value is greater than or equal to 0.10, the term is not statistically significant at the 90 percent or higher confidence level.

You can read the interpretation for the remaining statistics in the StatAdvisor.

Testing Hypotheses Using the GLM Analysis

What you really want to do is to compare the means for the **Sand** variable with different levels of the **Location** variable, so you will create Multiple Range Tests.

1. Click the Tabular Options button to display the dialog box, then click the Multiple Range Tests check box, and OK to display the table in the second text pane.

2. Maximize the text pane (see Figure 2-3).

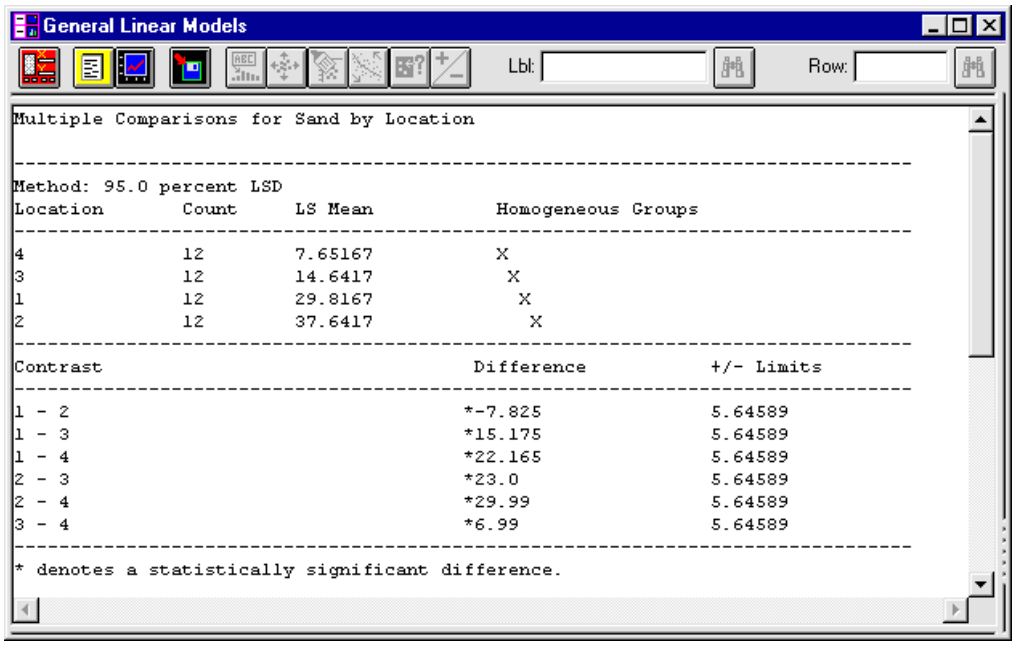


Figure 2-3. The Table of Results for the Multiple Range Tests

First, look at the results for **Sand** by **Location**. The table illustrates that the mean for each of the four locations is significantly different from each of

the other means. Therefore, each location forms a homogenous group by itself. Also, notice that in the second half of the table, the program lists the pairwise differences, which are known as *contrasts*. If they are significantly different from zero, they are marked with asterisks.

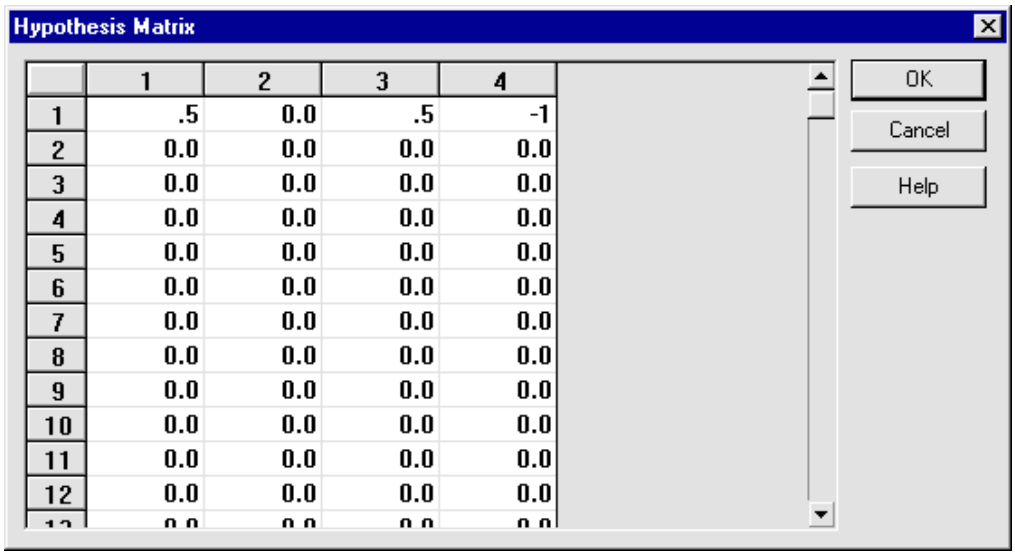
To this point, all the steps you have taken and the results you have generated could have been accomplished using the Multifactor ANOVA Analysis in the base program.

The General Linear Models Analysis lets you formulate your own contrasts to test a more complicated hypothesis. For example, Is the estimated mean for **Sand**, when averaged over the first and third locations, significantly different from the mean for the fourth location? To test this hypothesis, use Pane Options.

3. Click the right mouse button on the text pane, then the left on Pane Options to display the Multiple Comparisons Options dialog box.

4. Click the User-Specified button, accept the defaults for the remaining options on the dialog box, and click OK to display the Hypothesis Matrix dialog box.

5. Using the Tab key to move from left to right in the matrix, type the following in the first through fourth cells: ***.5, 0.0***, ***.5***, and ***-1*** (see Figure 2-4).



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

Figure 2-4. The Completed Hypothesis Matrix Dialog Box

6. Click OK to redisplay the Multiple Range Tests Table showing the results obtained using the contrasts you entered (see Figure 2-5).

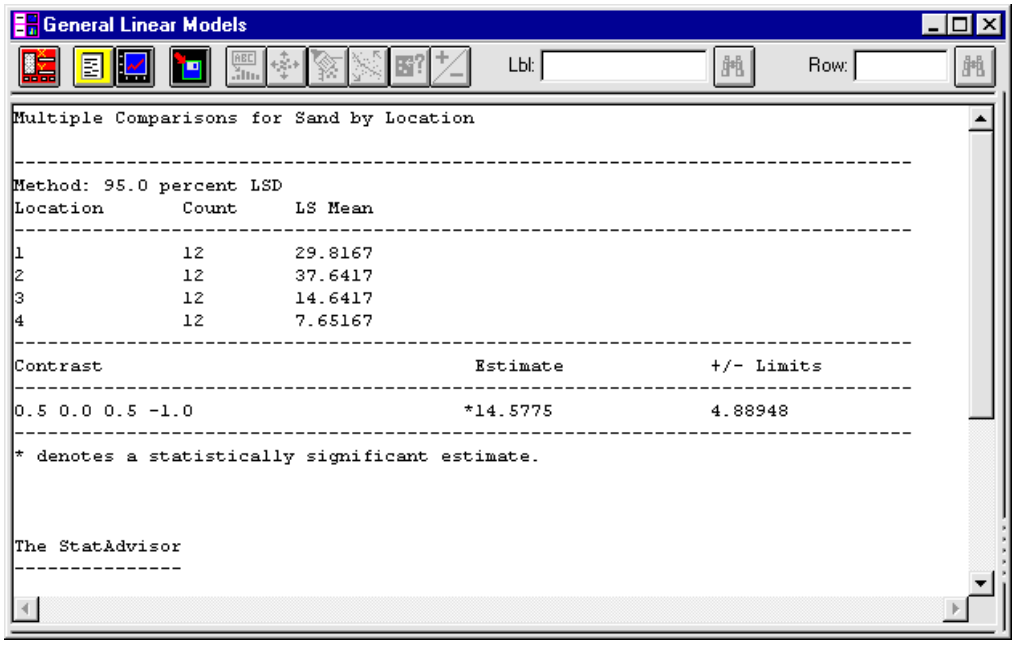


Figure 2-5. The Redisplayed Table of Results for the User-Defined Contrasts

The table shows the results of testing the contrasts. The asterisk indicates that the contrast is statistically different from 0.0 at the 95 percent confidence level.

The program used Fisher's Least Significant Difference (LSD) method to discriminate among the means. Using this method, there is a 5 percent risk that each pair of means will be significantly different when the actual difference is zero.

You can use a Means Plot to verify the results.

7. Click the Graphical Options button to display the dialog box, then the Means Plot check box, and OK to display the Means Plot in the graphics pane. Maximize the plot (see Figure 2-6).

Notice that the averages of Locations 1 and 3 are higher than that of Location 4.

Creating MANOVA Statistics

As the conclusion for this tutorial, you will add the two remaining response variables then create MANOVA statistics. The effect of each factor on the dependent variables is simultaneously quantified by the MANOVA.

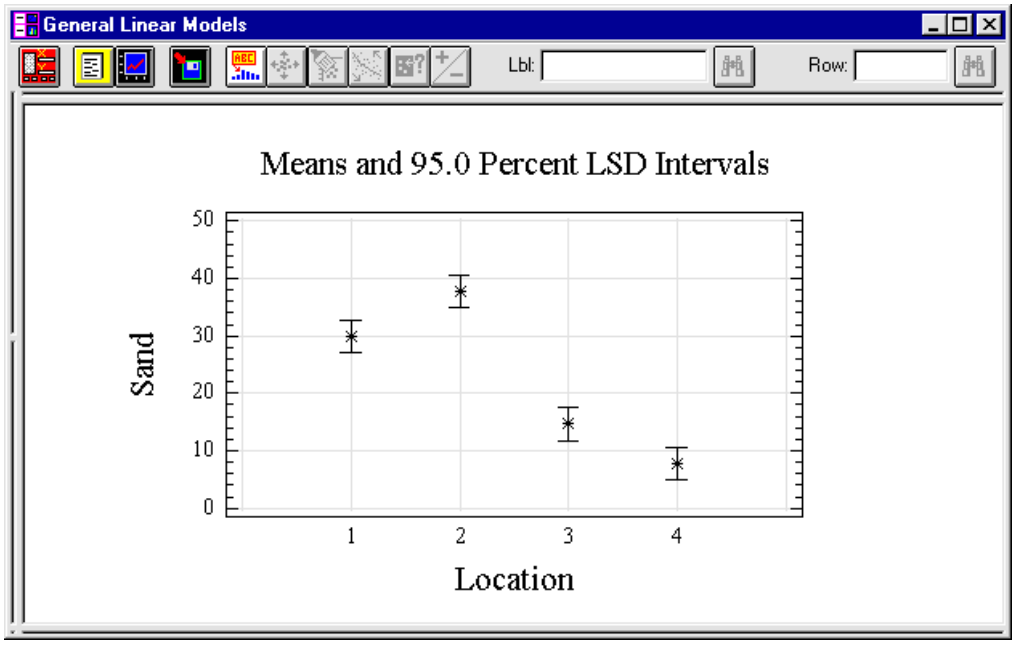


Figure 2-6. The Means Plot

1. Click the Return to Analysis Dialog Box button on the Analysis toolbar to redisplay the Analysis dialog box.

2. Enter **Silt**, **Clay**, and **Sand** into the Dependent Variables text box.

3. Enter **Location** and **Depth** into the Categorical Variables text box (see Figure 2-7).

4. Click OK to display the GLM Specification dialog box.

5. Accept the defaults and click OK to redisplay the Analysis Summary and the Means Plot in the Analysis window.

6. Maximize the Analysis Summary.

7. Click the right mouse button on the Analysis Summary then the left on Analysis Options to display the General Linear Models Options dialog box.

8. Click the Include MANOVA check box and accept the defaults for the remaining options on the dialog box.

9. Click OK to add the MANOVA statistics and to redisplay the Analysis Summary.

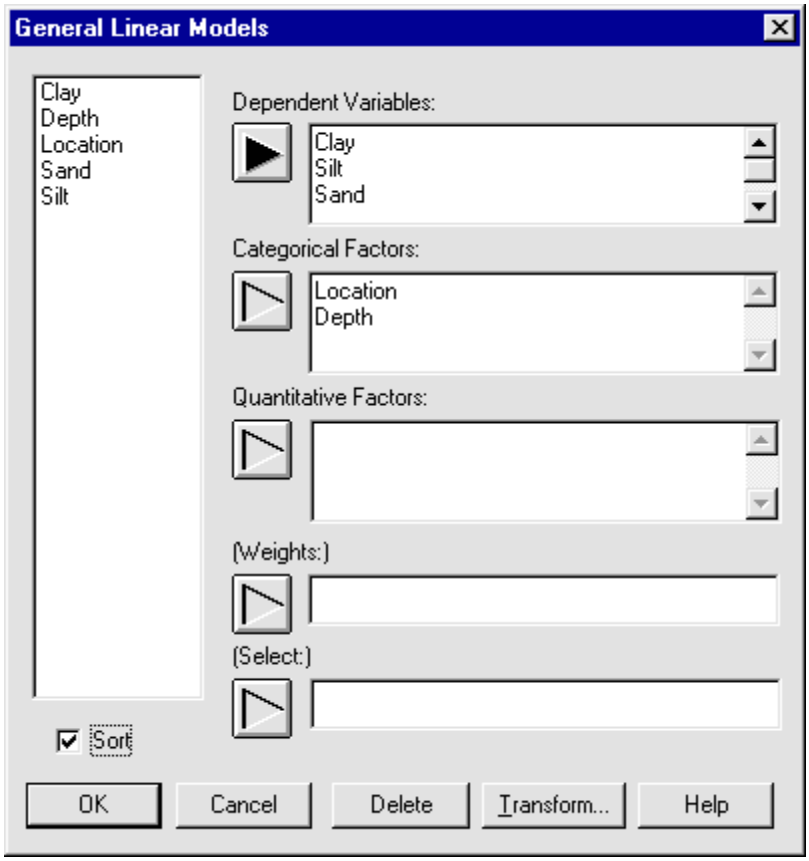


Figure 2-7. Completed General Linear Models Dialog Box

As you scroll through the Analysis Summary, notice that there are separate analysis of variance results for each of the three dependent variables. The MANOVA statistics appear at the end of the summary, including one set of statistics for each factor (see Figure 2-8).

Because *p*-values below 0.10 indicate that an effect is statistically significant at the 90 percent or higher confidence level, the most significant factor is B, **Depth**. The test was conducted using Wilks lambda.

To read the interpretations for other values in the table, see the StatAdvisor.

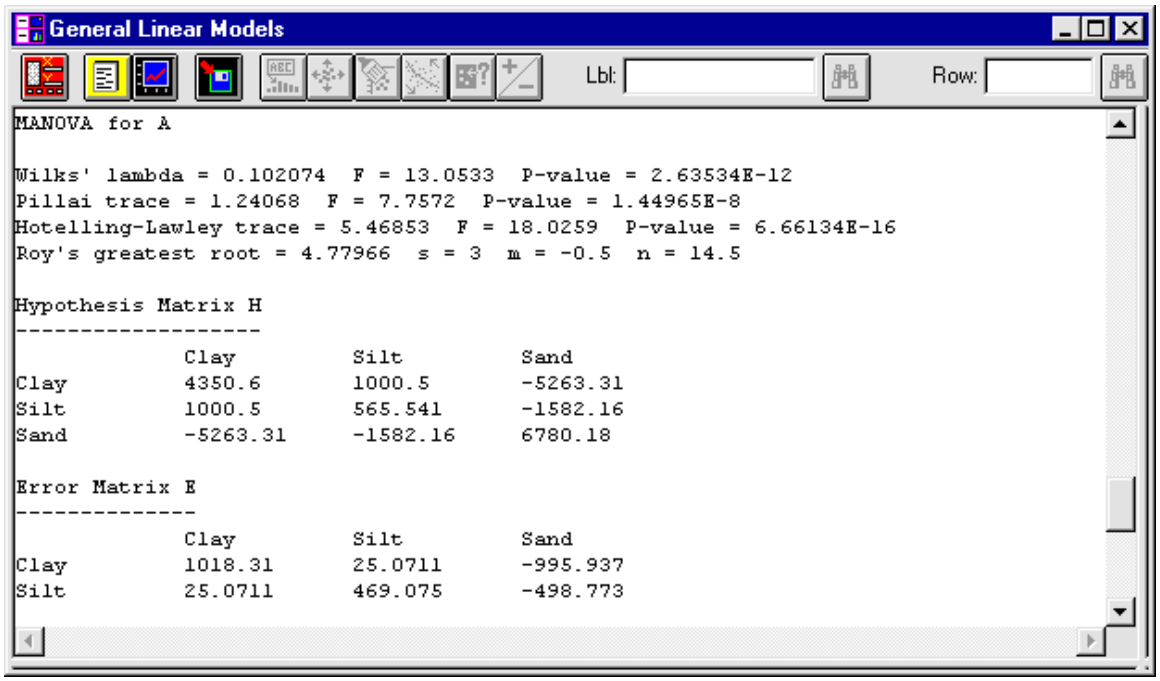


Figure 2-8. MANOVA Statistics in the Analysis Summary

TUTORIAL 3

Using Nested and Crossed Factors in a Model

This tutorial illustrates how you use nested and crossed factors in a model. The lesson is adapted from an example in Milliken and Johnson (1984), titled, Simple Comfort Experiment.

The comfort experiment studied the effects of three temperature levels and the gender of a person (male/female) in a two-way treatment structure based on each person's comfort level. The participants were each randomly assigned to three of nine available environmental chambers, numbered 1 to 3 for each of the three temperatures.

In the tutorial, you will analyze the environmental chambers nested within temperatures, as well as the effects of the **Temperatur*Gender** (temperature) interaction, which is a between-person comparison.

Before you begin, open STATGRAPHICS *Plus* and the **Comfort** data file.

Completing the General Linear Models Analysis Dialog Box

1. Choose SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display the General Linear Models Analysis dialog box.

2. Enter **Comfort** into the Dependent Variables text box.

3. Enter **Chamber**, **Temperatur**, and **Gender** into the Categorical Factors text box (see Figure 3-1).

4. Click OK to display the GLM Specification dialog box (see Figure 3-2).

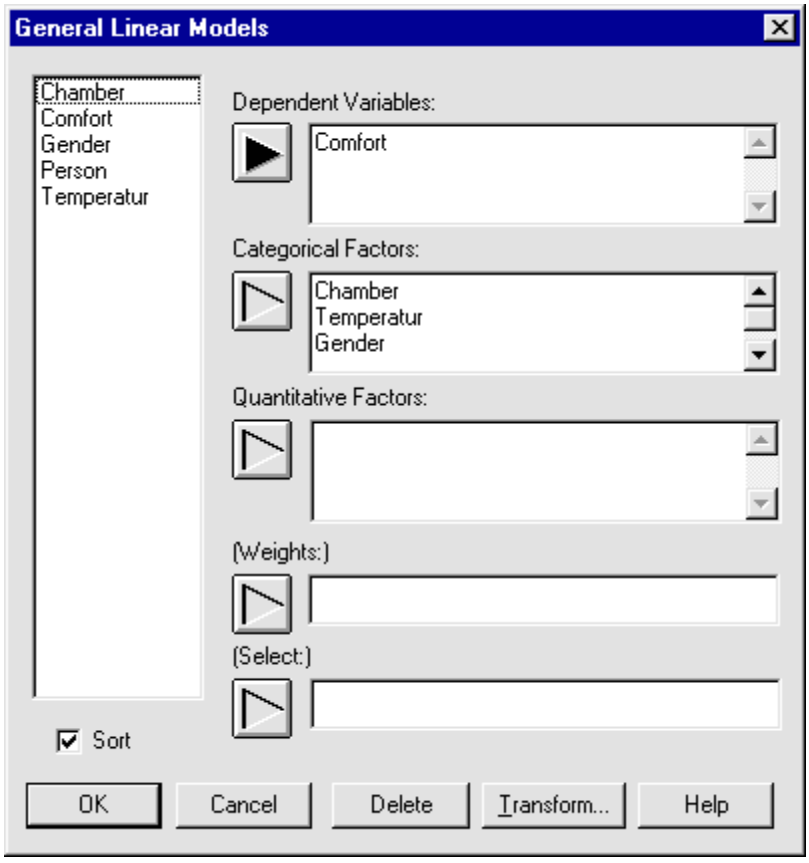


Figure 3-1. The Completed General Linear Models Analysis Dialog Box

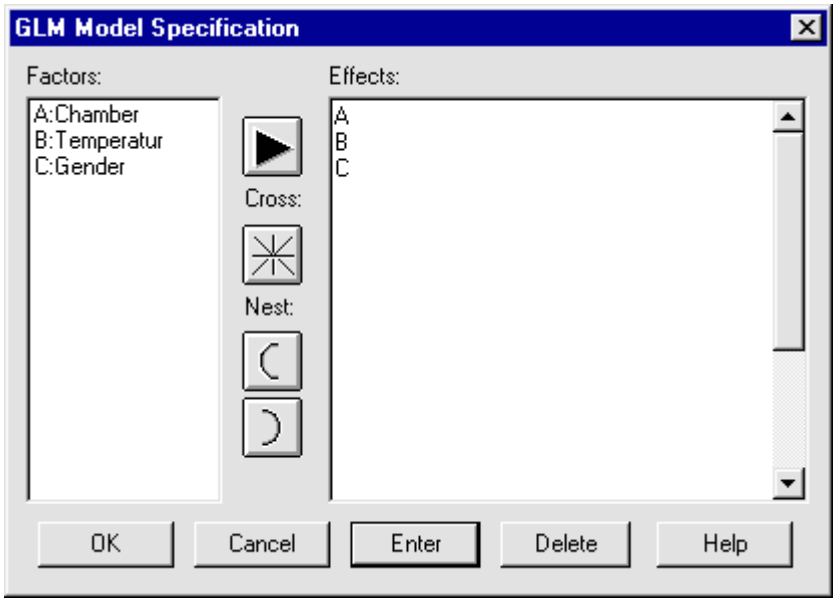


Figure 3-2. GLM Model Specification Dialog Box

Creating and Analyzing Nested Effects

You can use either the mouse or the keyboard to add terms to the Effects text box. The steps below, and throughout the remainder of these tutorials, use the mouse and the keyboard. If you vary from any of the steps, your results may not match the example.

Because you will use Factor A only in the nested factor, you will first delete it from the Effects list box. Then you will create the model to include the **Temperatur*Gender** effect (B*C) and the nested factor (A(B)). The nesting occurs in the design structure with **Chamber** nested within **Temperatur** (Milliken and Johnson, 1984).

The A(B) nested effect nests **Chamber** within **Temperatur**; that is, Chambers 1, 2, and 3 for the lowest temperature (65°) are not the same as Chambers 1, 2, and 3 for the highest temperature (75°). For clarification, you may want to look at how the data are entered into the DataSheet.

1. Click the mouse pointer in the left-most corner of the letter A in the Effects list box, hold down the left button, and drag the pointer over A to highlight it.

2. Click Delete to delete the factor.

3. Click the mouse pointer in the Effects text box on the line immediately under the C effect.

4. Click B in the Factors list box, then click the arrow button to move the factor to the Effects list box.

5. Click * (the asterisk- Cross) to move the asterisk to the right of the B factor in the Effects list box.

6. Click C in the Factors list box, then click the arrow button to move the factor to the Effects list box.

7. Click Enter to enter the nested factor on the next line.

8. Click A in the Factors list box, then click the arrow button to move the factor to the Effects list box.

9. Click ((the left parenthesis - Nest) to move the left parenthesis to the right of the A factor in the Effects list box.

10. Click B in the Factors list box, then click the arrow button to move the factor to the Effects list box.

11. Click) (the right parenthesis- Nest) to move the right parenthesis to the right of the B factor in the Effects list box (see Figure 3-3).

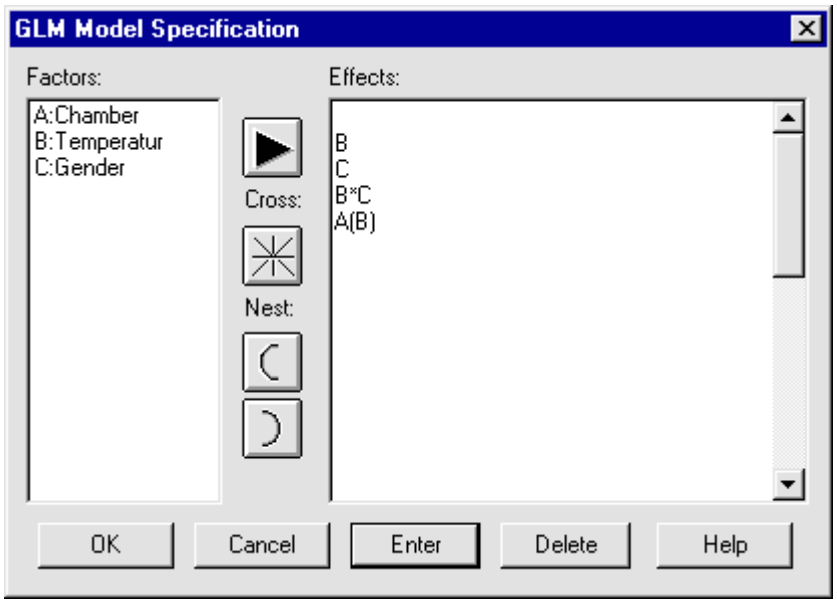


Figure 3-3. The Completed GLM Model Specification Dialog Box

If you make an error, highlight the text you want to remove, then click the Delete button.

12. Click OK to display the Analysis Summary and the Scatterplot in the Analysis window. Maximize the Analysis Summary (see Figure 3-4).

The values for the mean squares agree with the results shown in Milliken and Johnson (1984).

To see the results of the **Temperatur*Gender** interaction, you will create an Interaction Plot.

13. Click the Graphical Options button to display the dialog box, then the Interaction Plot check box, and OK to display the Interaction Plot in a graphics pane.

14. Maximize the plot (see Figure 3-5).

The interaction effect is shown by the crossed lines on the plot. The **Comfort** variable indicates that comfort level is not an additive function of the interaction effect **Gender** and **Temperatur**.

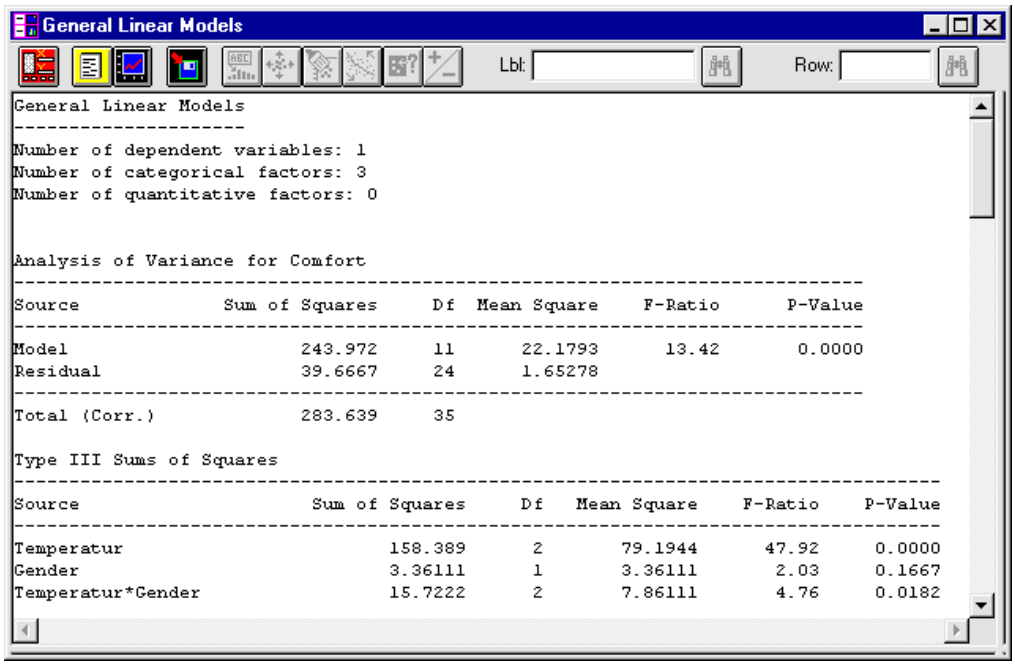


Figure 3-4. The Analysis Summary

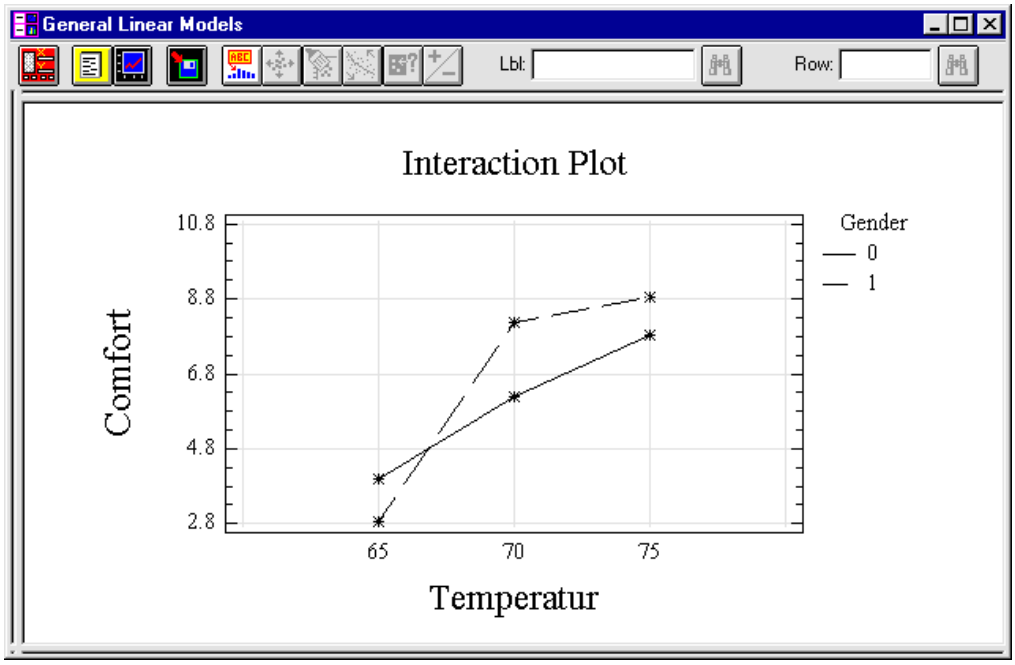


Figure 3-5. The Interaction Plot

Because the purpose of this tutorial was to introduce the concept of using nested and crossed effects, no further interpretation of the results is included

here. If you are interested in additional results, read the comments provided by the StatAdvisor.

References

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data*. Volume 1, *Designed Experiments*. New York: Van Nostrand Reinhold.

TUTORIAL 4

Creating and Using a Split Plot Design

The purpose of this tutorial is to demonstrate how to specify a model and create proper tests and plots. You will create and use a Split Plot design.

The tutorial is adapted from Milliken and Johnson (1984). When you are constructing a model, they provide two important reminders about design and concept:

- recognize that there are different sizes of experimental units
- identify the corresponding design and treatment structures.

The model is built by incorporating models for each size of experimental unit.

Milliken and Johnson′s example concerns yield, in pounds, for two varieties of wheat grown according to four fertility regimes. The researchers divided the fields into two blocks, each of which contained four whole plots. To each of the four plots within each block, they randomly assigned four fertilizing regimes to one whole plot. Then they divided each whole plot into two parts (subplots), and randomly assigned each variety of wheat to one subplot within each whole plot.

Before you begin, open STATGRAPHICS *Plus* and the **Wheat** data file.

Completing the General Linear Models Analysis Dialog Box

1. Choose SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display Analysis dialog box.

2. Enter **Yield** into the Dependent Variables text box.

3. Enter **Block**, **Regime**, and **Variety** into the Categorical Factors text box (see Figure 4-1).

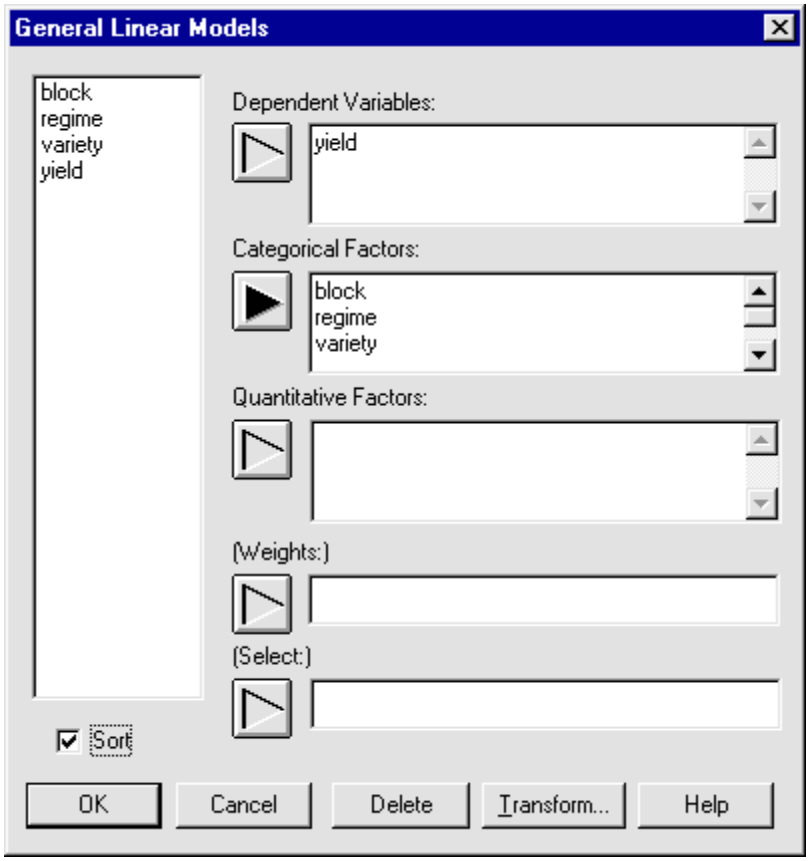


Figure 4-1. The Completed General Linear Models Analysis Dialog Box

Creating a Split Plot Model

1. Click OK to display the GLM Model Specification dialog box.

   You will enter two interaction effects: A*B and B*C.

2. Position the mouse pointer in the Effects text box on the line immediately under the C effect.

3. Click A in the Factors list box, then the arrow button to move the factor to the Effects list box.

4. Click * (the Cross button) to place the asterisk to the right of the A factor in the Effects text box.

5. Click B in the Factors list box, then the arrow button to move the factor to the Effects list box.

6. Position and click the mouse pointer in the Effects text box on the line immediately under the C effect.

7. Follow Steps 3, 4, and 5 above, using the B and C factors.

Figure 4-2 illustrates how the GLM Model Specification dialog box will look when you have completed these steps.

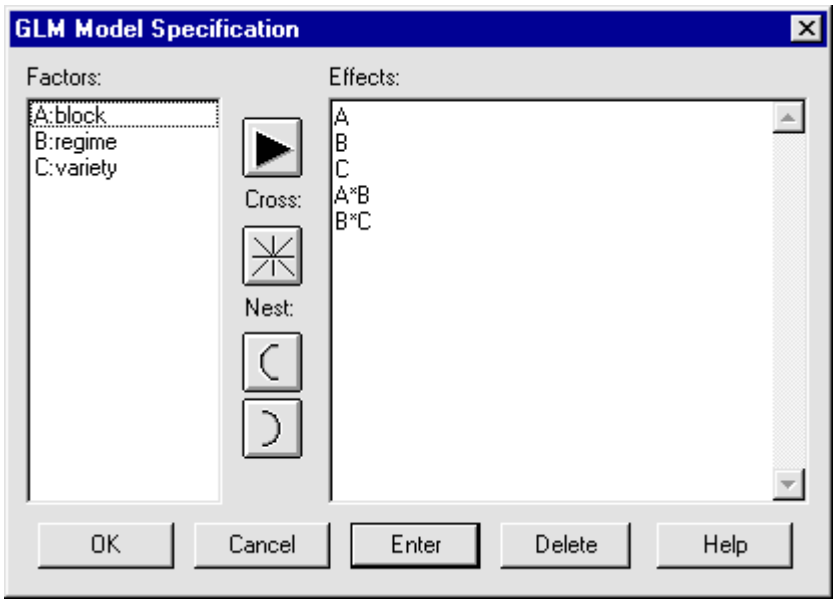


Figure 4-2. Completed GLM Model Specification Dialog Box

1. Click OK to display the Analysis Summary and Scatterplot in the Analysis window. Maximize the Analysis Summary (see Figure 4-3).

The table summarizes the results of fitting a general linear model that relates the **Yield** variable to three predictive factors: **Block**, **Regime**, and **Variety,** where **Regime** is the whole plot factor and **Variety** is the subplot factor. The table also summarizes how well the model performed when it fit the data and predicted the values that were withheld from the fitting process. Each of these statistics is based on the residuals.

To continue with the analysis, you will change the error terms to account for the Split Plot design.

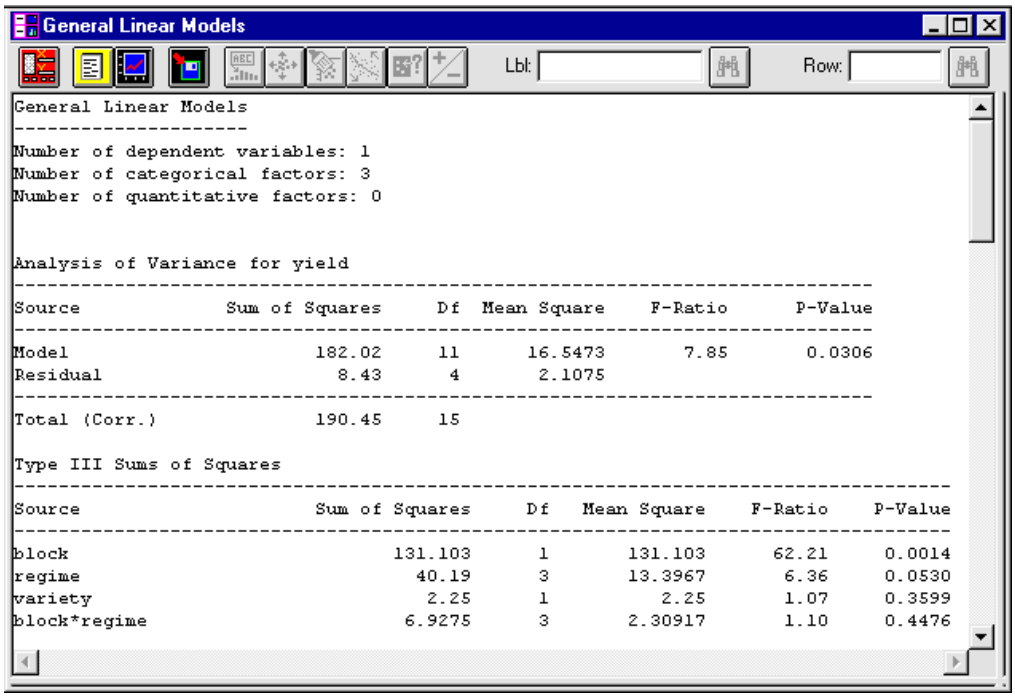


Figure 4-3. The Analysis Summary

Modifying the Error Terms

1. Click the right mouse button on the text pane then the left on Analysis Options to display the General Linear Models Options dialog box.

2. Click A in the Factors list box then None in the Error Term list box to display A - None in the Selections list box (you are not conducting a test on **Block).**

3. Click B in the Factor list box then A*B in the Error Term list box to display B - A*B in the Selections list box.

   The F test compares the mean squares for **Regime** (B) with the whole plot mean squares, A*B.

4. Click A*B in the Factor list box then None in the Error Term list box to display A*B - None in the Selections list box.

   Notice that the first factor, A, is the **Block** factor; B versus A*B is the whole plot design; and B*C is the subplot design (see Figure 4-4).

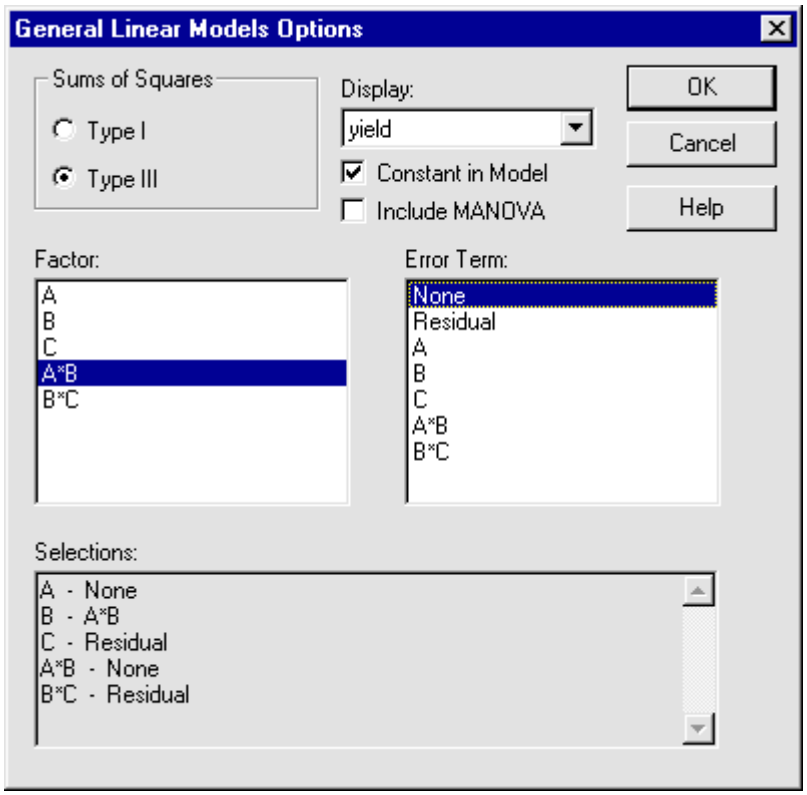


Figure 4-4. The Completed Options Dialog Box

5. Click OK to redisplay the Analysis Summary shown in Figure 4-5.

The values shown in the second ANOVA table match the results in the Milliken and Johnson (1984) study.

References

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data*. Volume 1, *Designed Experiments*. New York: Van Nostrand Reinhold.

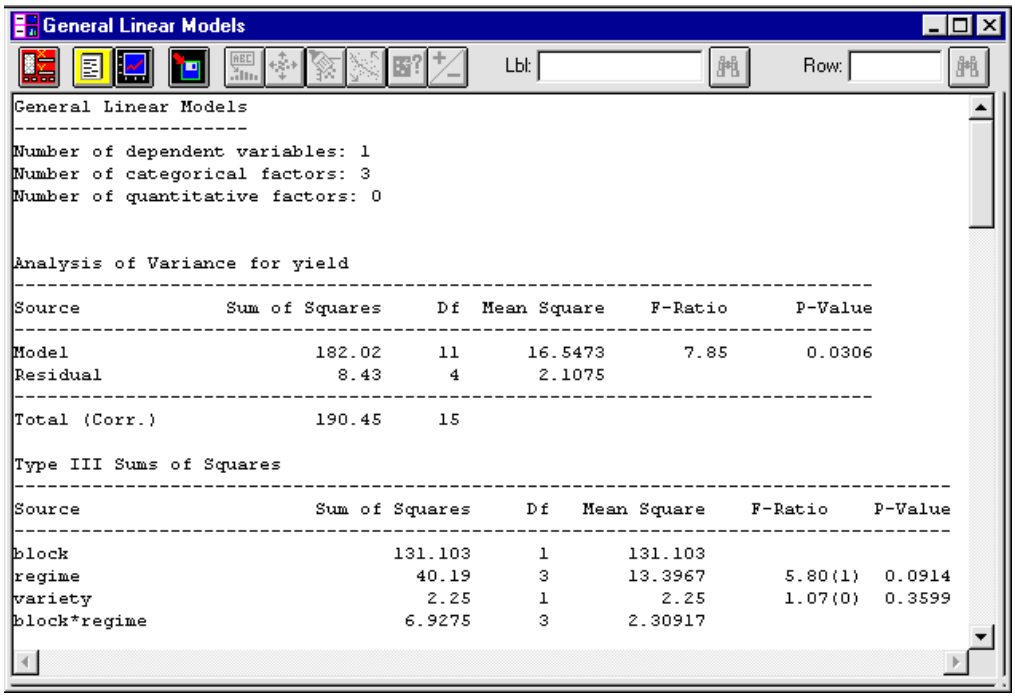


Figure 4-5. Redisplay of the Analysis Summary with Results that Match Milliken and Johnson (1984)

TUTORIAL 5

Creating and Using Repeated Measures Designs

Repeated measures designs, like split plot designs, contain structures that involve more than one size of experimental unit. For example, you might measure one subject over a timeframe where time is one of the factors in the treatment structure.

In repeated measures designs you cannot randomly assign levels of factors, which means that errors that correspond to the experimental units may have a covariance matrix.

This tutorial, adapted from Milliken and Johnson (1984), illustrates how you create and enter data for a repeated measures design, how you enter model effects, change error terms, and create the subsequent report and plot.

The Milliken and Johnson investigation examines the effects of three drugs, which were administered to eight subjects. The researchers measured the heart rate for each subject every five minutes for four time intervals. The larger experimental unit is the subject; the smaller is the time interval.

Before you begin, open STATGRAPHICS *Plus* and the **Heart** data file. When the DataSheet appears, notice how the entries are coded by scrolling through the file. Notice that the **Drug** column contains three different drugs: ax23, bww9, and the control (see Figure 5-1).

In the analysis, the **Person** variable is nested within the **Drug** factor. This is because Subject 1 for the ax23 drug is not the same as Subject 1 for the bww9 drug, or the same as Subject 1 for the control drug. Although you could code each subject with a different number (1 to 24), coding them 1 through 8 within each drug, then specifying them as a nested factor in the model is faster and easier.

After you view the DataSheet, minimize it.

Heart.sf

	drug	person	time	response	Col_5
32	ax23	8	4	70	
33	bww9	1	1	85	
34	bww9	1	2	86	
35	bww9	1	3	83	
36	bww9	1	4	80	
37	bww9	2	1	82	
38	bww9	2	2	86	
39	bww9	2	3	80	
40	bww9	2	4	84	
41	bww9	3	1	71	
42	bww9	3	2	78	
43	bww9	3	3	70	
44	bww9	3	4	75	
45	bww9	4	1	83	
46	bww9	4	2	88	
47	bww9	4	3	79	
48	bww9	4	4	81	
49	bww9	5	1	86	

*Figure 5-1. The DataSheet for the **Heart** Data File*

Completing the General Linear Models Analysis Dialog Box

1. Choose SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS... from the Menu bar to display the Analysis dialog box.

2. Enter **Response** into the Dependent Variables text box.

3. Enter **Drug**, **Person**, and **Time** into the Categorical Factors text box (see Figure 5-2).

4. Click OK to display the GLM Model Specification dialog box.

Entering the Model Effects

Now you will specify the model effects B(A), which is **Person** within **Drug**. However, first you will add the **Time*Drug** interaction, C*A, because you suspect that it might also be important.

1. Position and click the mouse pointer under C in the Effects list box.

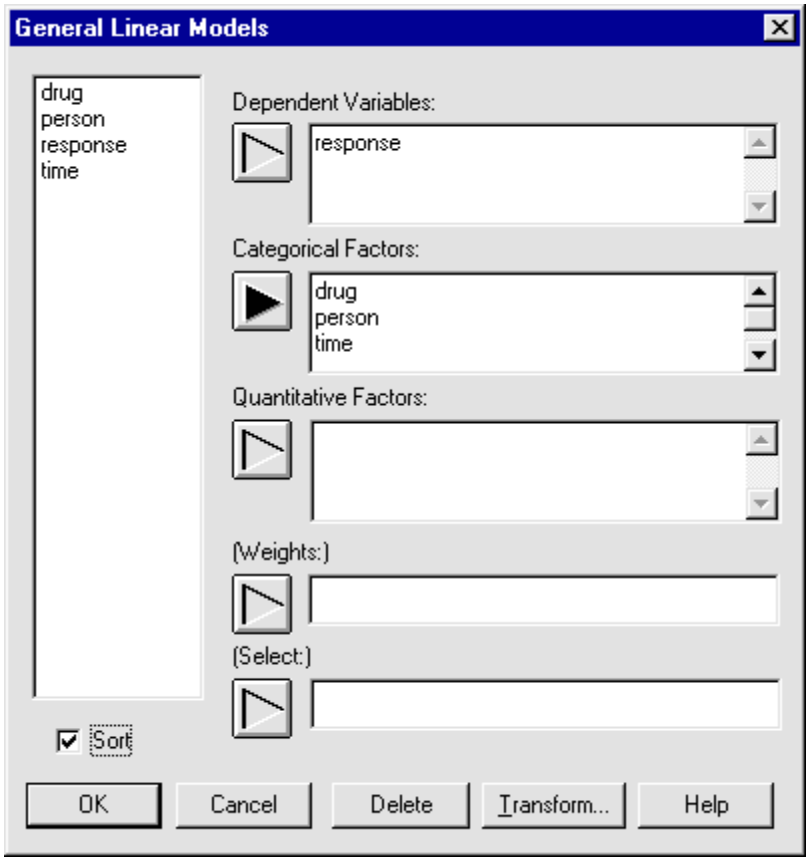


Figure 5-2. The Completed General Linear Models Analysis Dialog Box

2. Click C in the Factors list box, then the arrow button to move the factor to the Effects list box.

3. Click * (the asterisk - Cross button) to place the asterisk to the right of the C in the Effects text box.

4. Click A in the Factors list box, then the arrow button to move the factor to the Effects list box.

5. Position and click the mouse pointer next to the B factor in the Effects list box A.

6. Click ((the left parenthesis - Nest button) to place the parenthesis to the right of the B factor.

7. Click A in the Factors list box, then the arrow button to move the factor to the Effects list box

8. Click) (the right parenthesis - Nest button) to move the parenthesis to the right of the A factor in the Effects list box.

   Your computer screen should now look like the one shown in Figure 5-3.

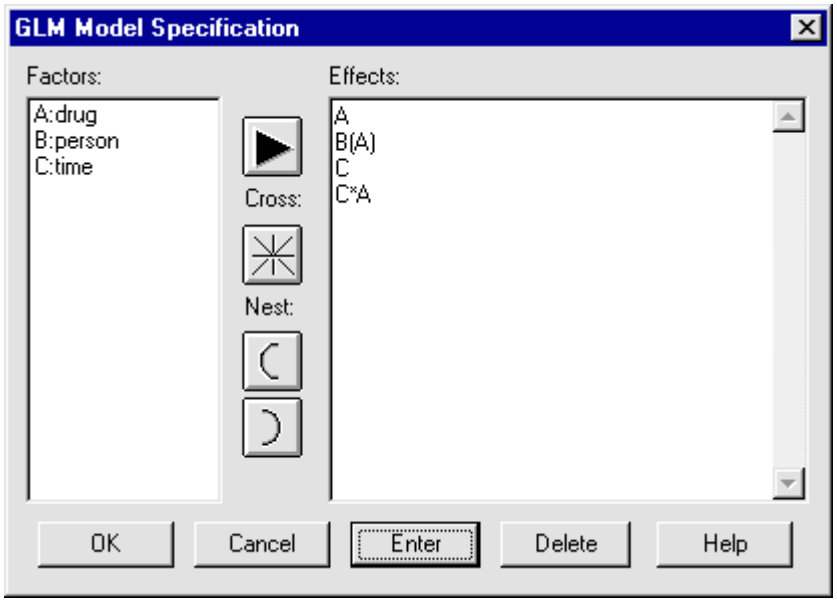


Figure 5-3. The Completed GLM Model Specification Dialog Box

Changing Error Terms

1. Click OK to display the Analysis Summary and Scatterplot in the Analysis window. Maximize the Analysis Summary.

2. Click the right mouse button on the text pane, then the left on Analysis Options to display the General Linear Models Options dialog box.

3. Click A in the Factor list box, then the B(A) interaction in the Error term list box to display A - B(A) in the Selections list box.

4. Click the B(A) interaction in the Factor list box, then click None in the Error Term list box to display B(A) - None in the Selections list box.

   The two changes appear as A - B(A) and B(A) - None (see Figure 5-4).

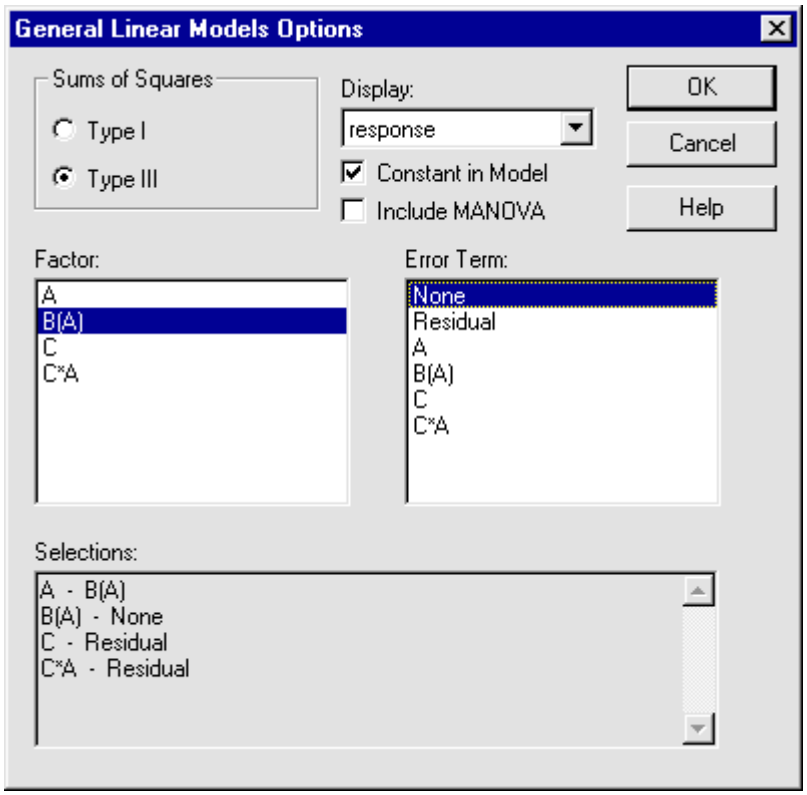


Figure 5-4. The Completed General Linear Models Options Dialog Box

Creating a Report and a Plot

1. Click OK to recalculate and redisplay the Analysis Summary (see Figure 5-5).

The second ANOVA Table shows the Type III Sums of Squares. Notice that the highest *p*-value is 0.0088 for Factor A, which is less than 0.01, the highest order term that is statistically significant at the 99 percent confidence level. This means that you probably will not want to remove any variables from the model. The new error term definitions are shown in the footnote on the table.

It would be interesting to see the Interaction Plot.

2. Click the Graphical Options button to display the dialog box, then the Interaction Plot check box, and OK to display the plot in the graphics pane. Maximize the plot (see Figure 5-6).

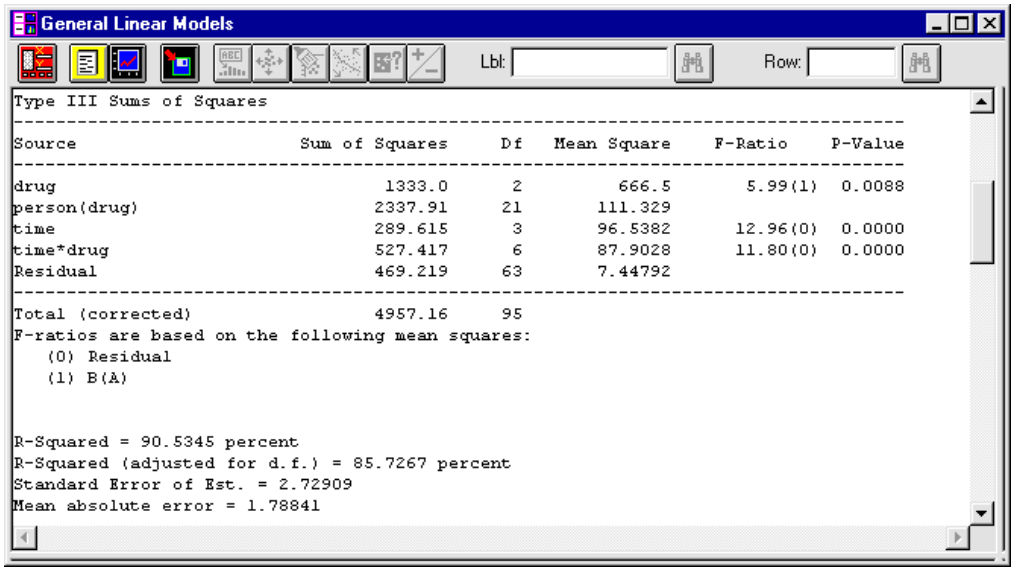


Figure 5-5. Redisplay of the Analysis Summary

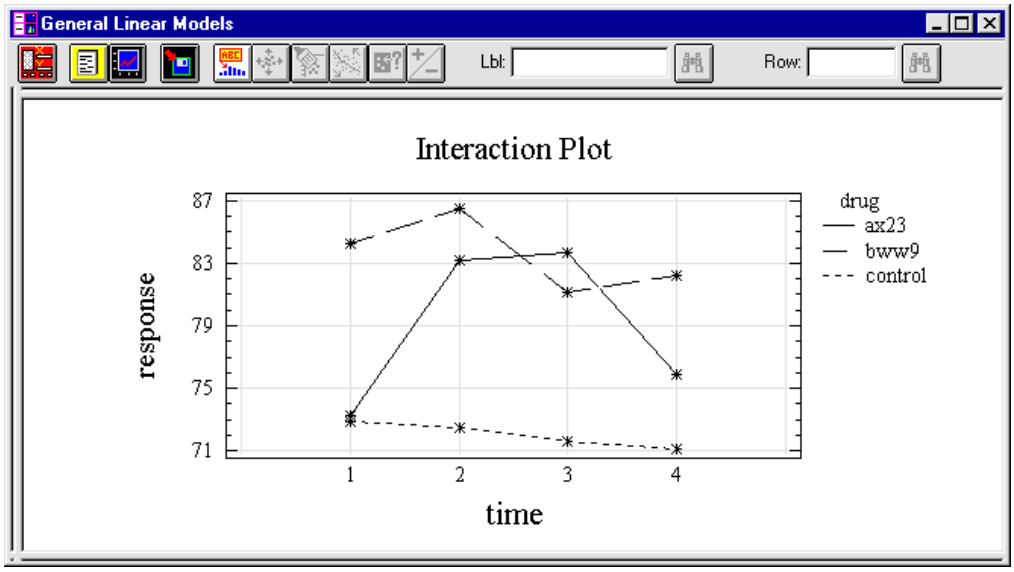


Figure 5-6. The Interaction Plot

The plot shows the interaction between **Time** and **Drug**. The three lines on the plot represent each of the three drugs. The lines connect the least squares means for the four levels of **Time**. The lines will be parallel if an interaction does not occur.

You can see from the plot that the **Time** trend is very different among the three drugs.

References

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

McCullagh, 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 1, *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**:370-384.

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.

Introduction

This portion of the online manual of tutorials for the Advanced Regression analyses in STATGRAPHICS *Plus* contains Part II -- tutorials for these analyses:

- Calibration Models
- Comparison of Regression Lines
- Regression Model Selection
- Nonlinear Regression
- Ridge Regression
- Logistic Regression.

To use the tutorials for the General Linear Models Analysis, see Part I of this manual.

Tutorials in this Manual

The tutorials for the remaining analyses are:

- *Fitting a Calibration Line* (Calibration Models Analysis)
- *Analyzing an Insurance Innovation Study (*Comparison of Regression Lines Analysis)
- *Illustrating Model-Building Techniques (*Regression Model Selection Analysis)
- *West Virginia Mining Excavation Study (*Nonlinear Regression Analysis)
- *Studying the Relationship of Body Fat to Explanatory Variables* (Ridge Regression Analysis)
- *Analyzing Coupon Redemption Rate with Logistic Regression* (Logistic Regression Analysis)

TUTORIAL 6

Fitting a Calibration Line

This tutorial is adapted from a study reported in Caulcutt and Boddy (1995) in which four analysts participated. Their employer, Indichem Ltd., uses large quantities of ammonia solution that they purchase from several vendors.

Although a new and less expensive supply source looks promising, the chief chemist suspects that this ammonia solution might be contaminated with cuprammonium. He devised a spectrophotometric method for determining the concentration of cuprammonium ion in the solution, which involves measuring the absorbency of a sample at a wavelength of 600 nm.

The chemist was quite sure there was a relationship between the absorbency reading of the spectrophotometer and the concentration of cuprammonium ion in the sample. The relationship might have been influenced by other compounds present in the sample, however, he chose not to investigate those effects now.

Instead, he will calibrate the test method by quantifying the relationship between absorbency (Y) and concentration (X), using a range of concentration that is likely to be found when deliveries of the solution are monitored.

Five samples of known concentration were prepared, the absorbency of each was recorded, and stored in a data file.

Before you begin, open STATGRAPHICS *Plus* and the **Calib** (calibration) data file.

Determining the Relationship between Two Variables

1. Choose SPECIAL... ADVANCED REGRESSION... CALIBRATION MODELS... from the Menu bar to display the Analysis dialog box.

The concentrations are known, therefore, you will use them for the X variable. The chief chemist analyzed each sample with the spectrophotometer; the **Absorbence** variable contains these measurements. You will use them for the Y variable. It is very important that you enter the variables in the correct text boxes; do not reverse them.

Removing Intercepts from a Model

1. Click the right mouse button on the text pane, then the left on Analysis Options to display the Calibration Model Options dialog box.

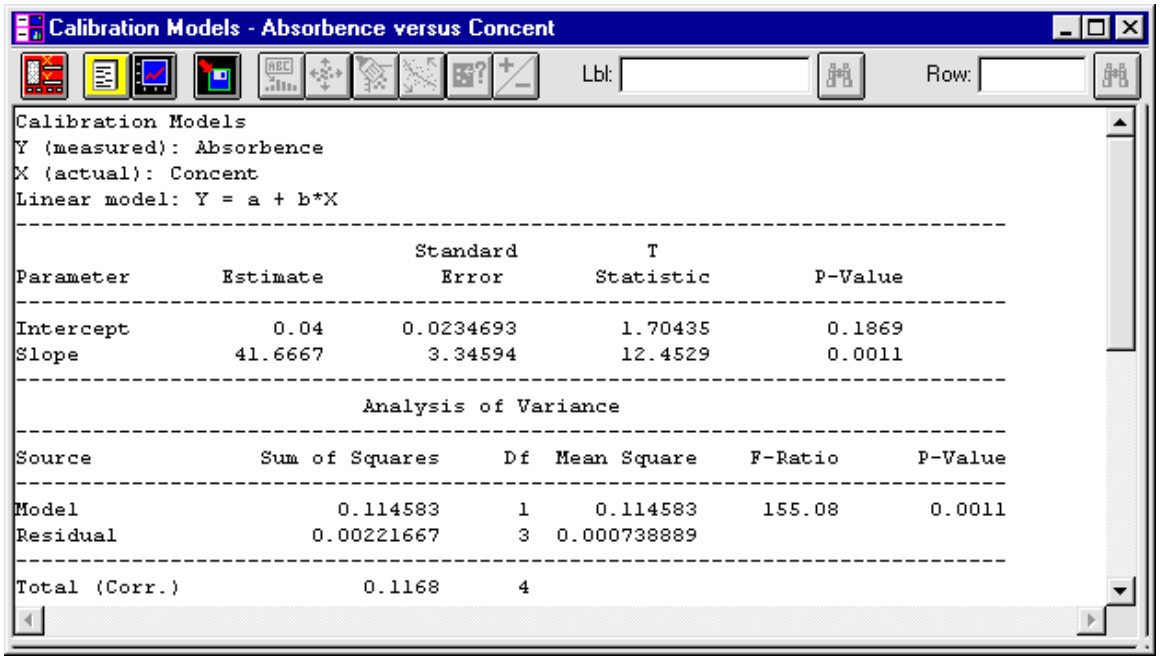


Figure 6-2. The Analysis Summary Results

2. Click the Include Constant check box to turn the option off as shown in Figure 6-3.

3. Click OK to recalculate and redisplay the data with the constant removed from the model (see Figure 6-4).

   The results are significant, so you will create a plot of the model and look at the calibration line.

Plotting the Model

1. Minimize the Analysis Summary and maximize the Plot of Fitted Model (see Figure 6-5).

   Notice that when you remove the constant from the plot, the prediction limits are not parallel; instead, the lower values for the concentration are slightly

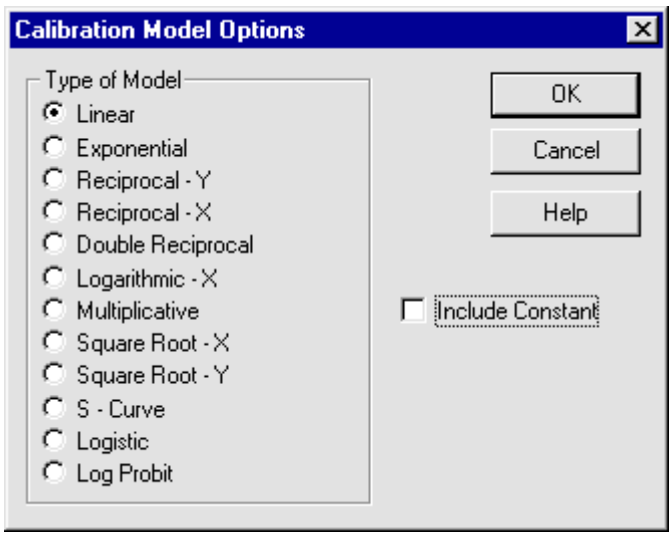


Figure 6-3. The Calibration Model Options Dialog Box

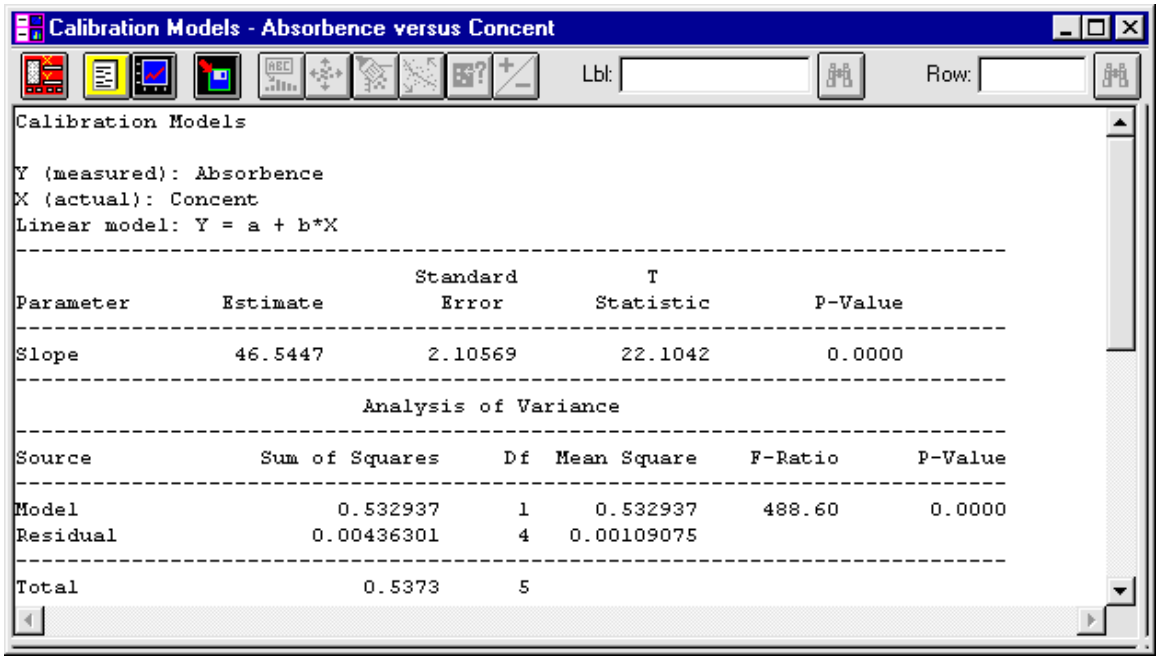


Figure 6-4. The Analysis Summary Recalculated After Removing the Constant

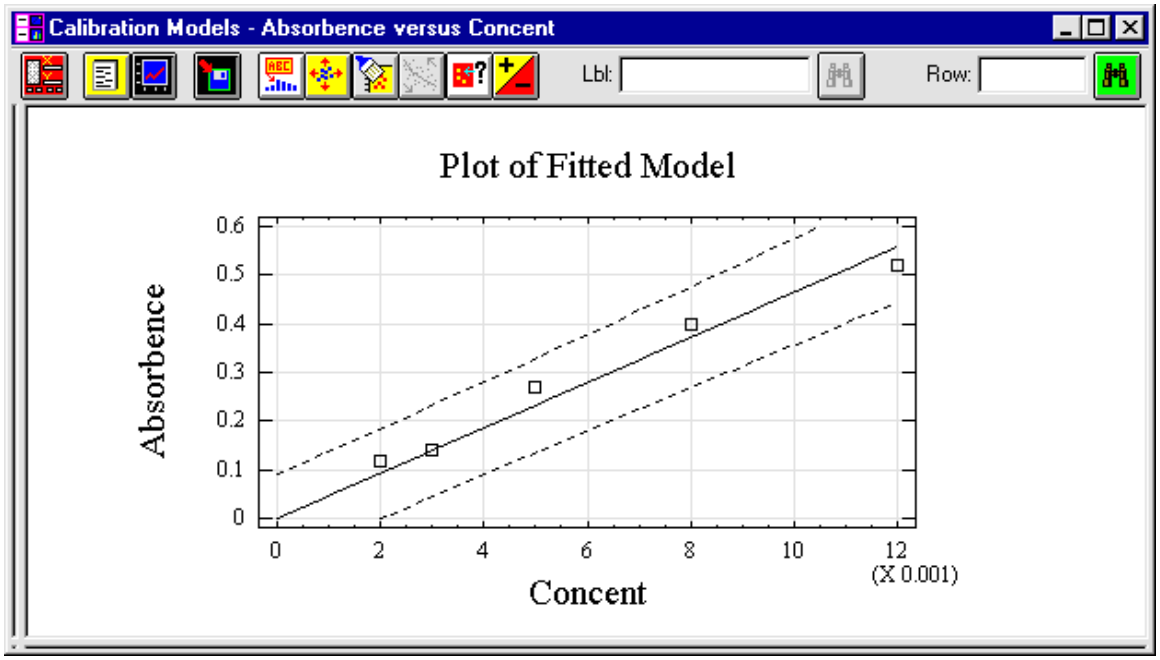


Figure 6-5. The Plot of Fitted Model

closer together than they were for the higher values. This is particularly evident for the confidence limits, which you will now add to the plot.

2. Click the right mouse button on the graphics pane, then the left on Pane Options to display the Plot of Fitted Model Options dialog box.

3. Click the Confidence Limits check box and accept the defaults for the other options on the dialog box (see Figure 6-6).

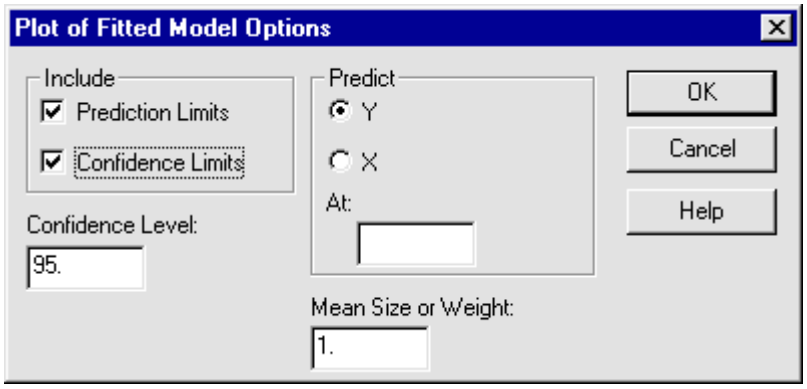


Figure 6-6. Completed Plot of Fitted Model Options Dialog Box

4. Click OK to display the plot with the confidence limits added (see Figure 6-7).

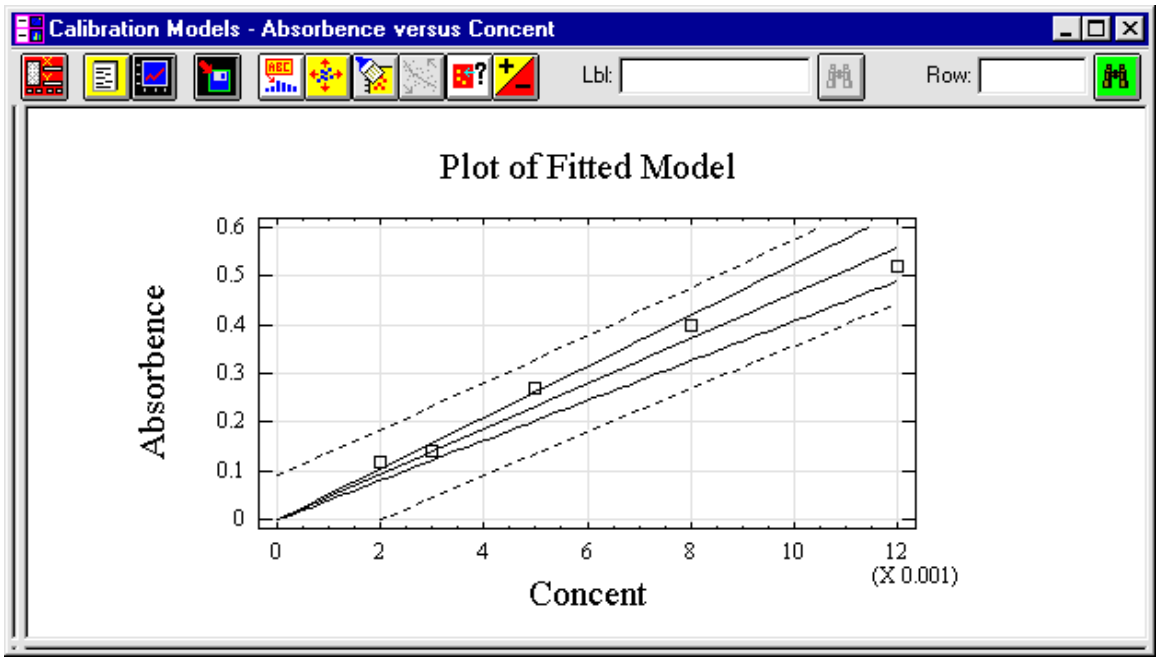


Figure 6-7. Plot of Fitted Model with the Confidence Limits Added

Instead of using a laborious hand calculation to calculate the prediction for Y, you will again use the Plot of Fitted Model Options dialog box.

5. Click the right mouse button on the graphics pane, then the left on Pane Options to display the Plot of Fitted Model Options dialog box.

6. Click the Confidence Limits check box to turn it off, then click the X predictions option, enter ***0.352*** into the At text box, and accept the defaults for the remaining options (see Figure 6-8).

7. Click OK to plot the prediction limits and the concentration values, and redisplay the plot (see Figure 6-9).

The results shown in the plot are fine, but the chemist decides it would be helpful to have a table that lists several predictions at the same time. To do this, create a Predictions Table.

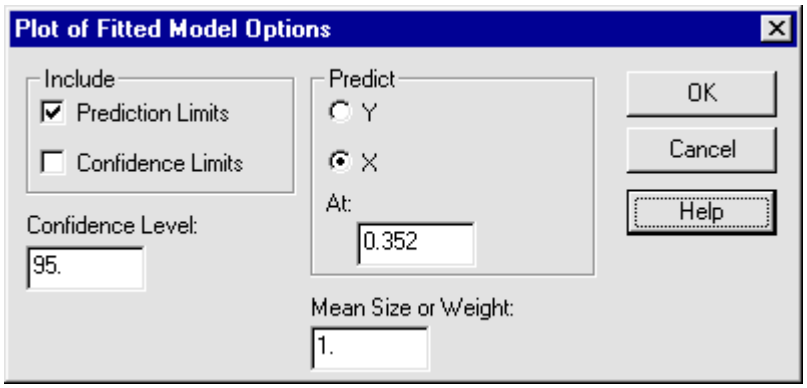


Figure 6-8. Completed Dialog Box for Predicting X

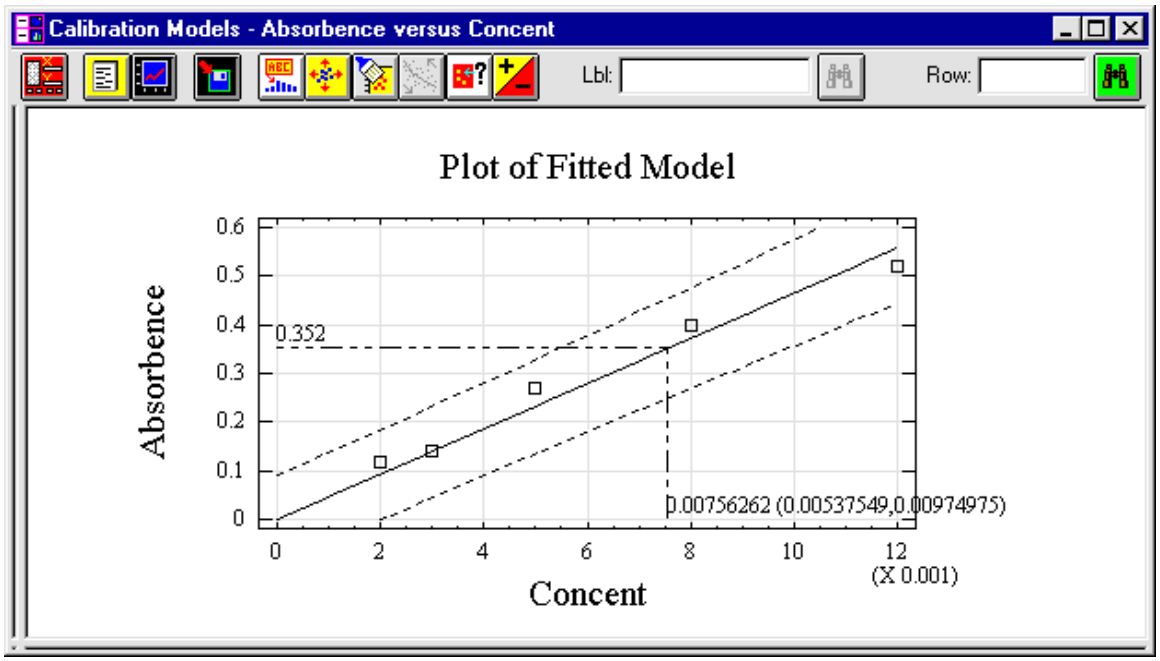


Figure 6-9. The Plot of Fitted Model with Confidence Limits and Prediction Lines Added

Creating a Predictions Table

1. Click the Tabular Options button to display the dialog box, then click the Predictions check box, and OK to display the Predictions Table in the second text pane.

2. Maximize the table (see Figure 6-10).

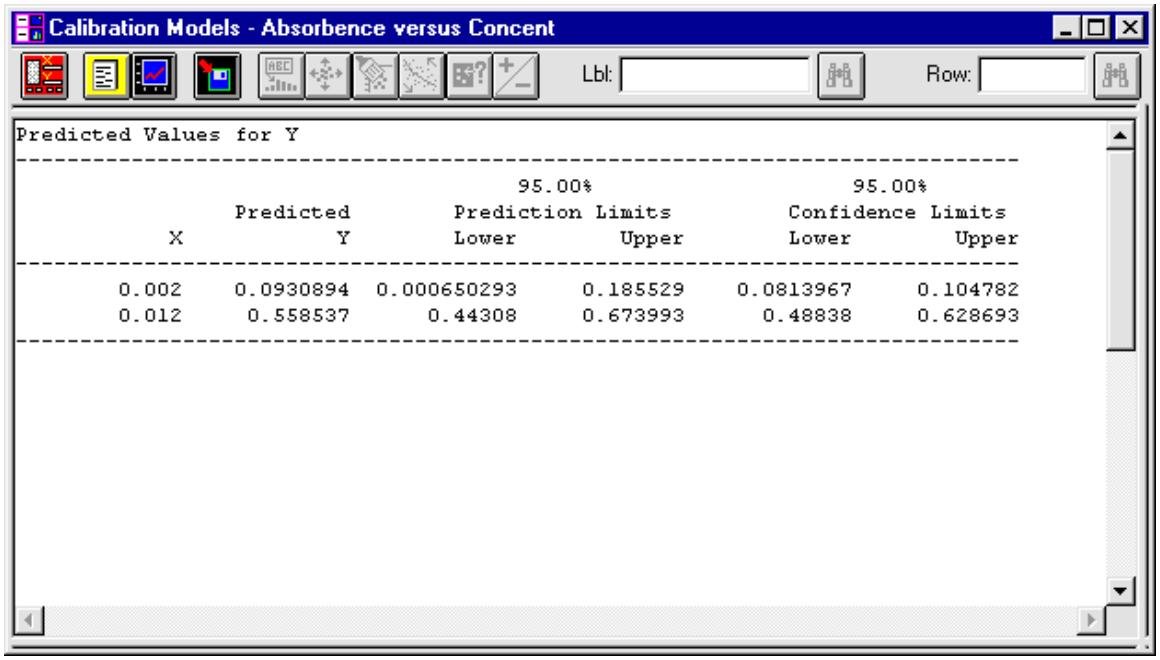


Figure 6-10. Predictions Table for Y

Notice that the table lists predictions for Y for the Upper and Lower values of X. A regression equation is used to predict either X or Y, given a value for the other variable. You will predict X values for six different values of Y.

3. Click the right mouse button on the text pane, then the left on Pane Options to display the Predictions Options dialog box.

4. Click X in the Prediction portion of the dialog box to change the prediction limits.

5. Enter ***0***, ***.1***, ***.2***, ***.3***, ***.4***, and ***.5*** in the first six Predict At text boxes.

6. Accept the defaults in the Confidence Level and Mean Size, or Weight text boxes (see Figure 6-11).

   The program calculates and displays the values for the Predictions Table (see Figure 6-12).

   The table shows predictions for the concentrations of six absorbency readings as well as for the prediction limits.

Saving Statistics for a Calibration Line

The chemist is now satisfied with the results and wants to save the calibration line for later use.

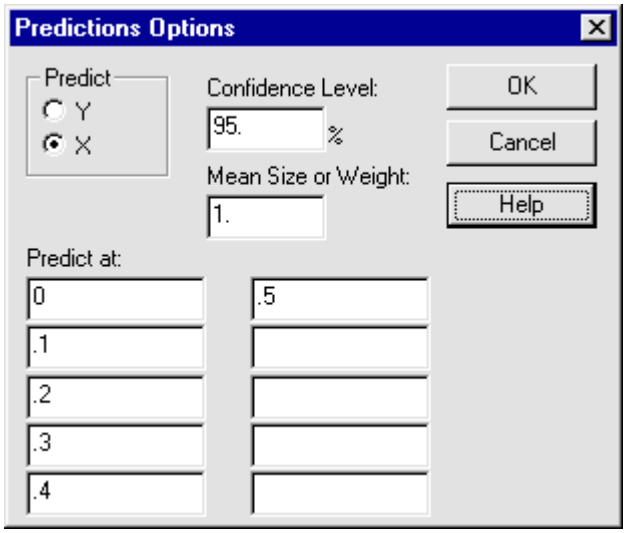


Figure 6-11. Completed Dialog Box for Predictions Options

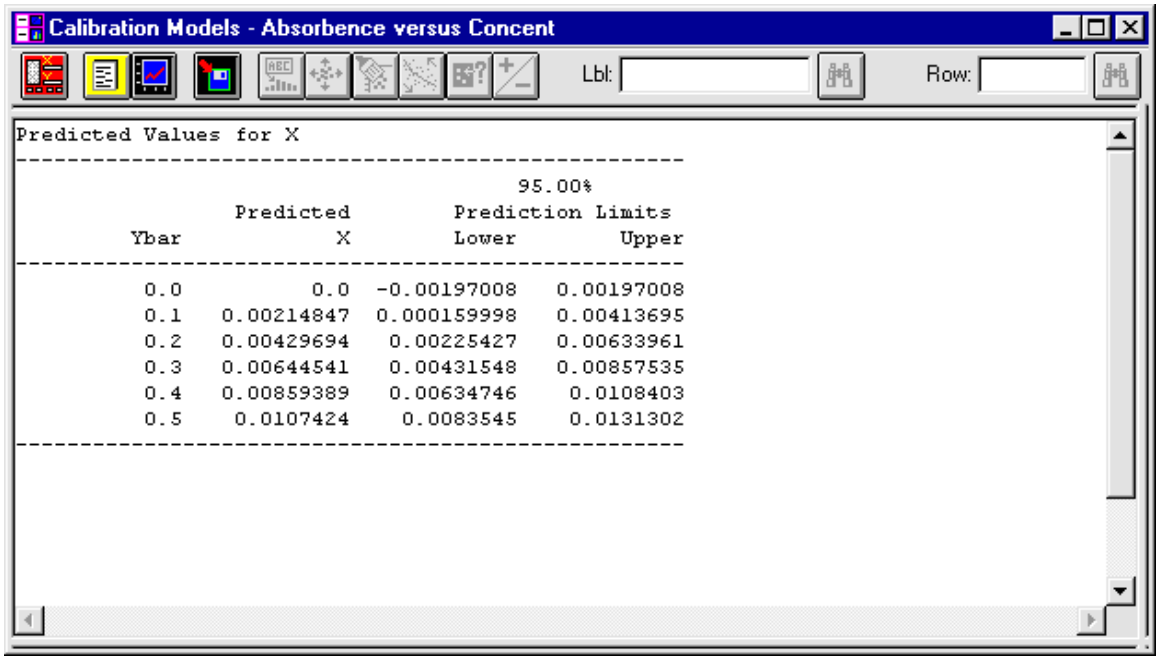


Figure 6-12. Predictions Table for X

1. Click the Save Results button on the Analysis toolbar (the fourth button from the left) to display the Save Results Options dialog box.

2. Click the Model Statistics check box under the Save options and type ***MYMODEL*** in the first Target Variables text box (see Figure 6-13).

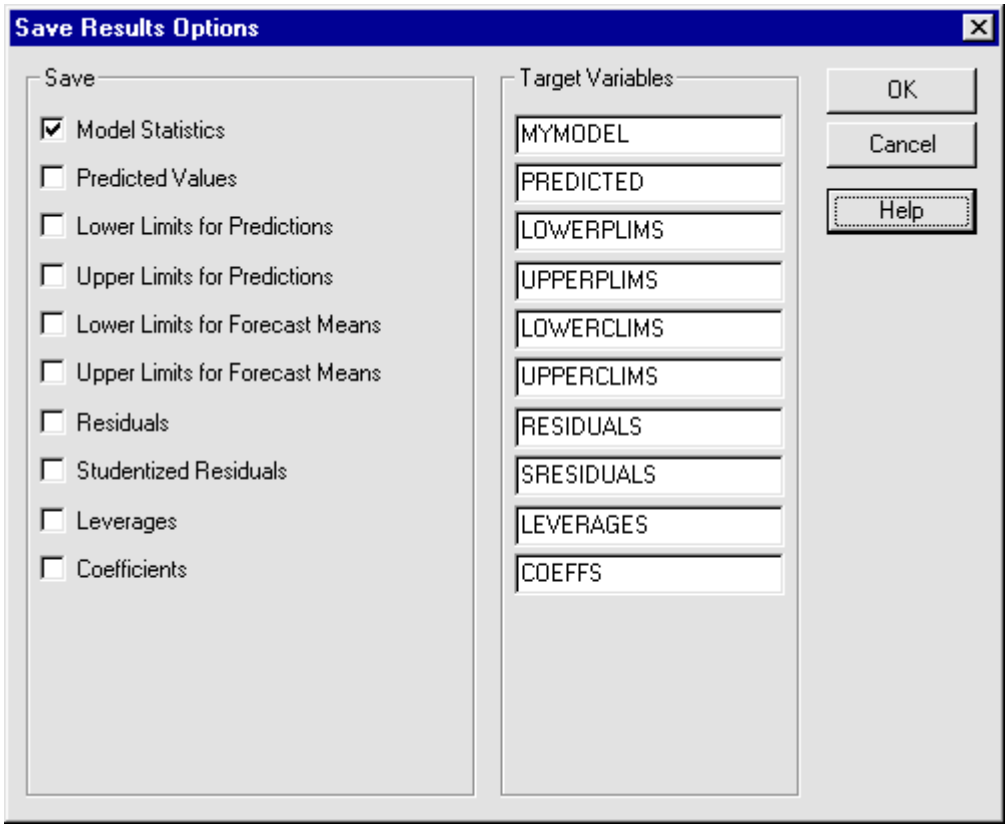


Figure 6-13. Completed Save Results Options Dialog Box

3. Click OK to save the column of numbers that define the model.

4. Click FILE... SAVE... SAVE DATA FILE... from the Application toolbar to save the file.

Using Previously Saved Data

Two weeks after completing the first analysis, the chemist repeated the new analysis on a new sample, which resulted in an absorbency measurement of .44. He now wants to use the calibration line that was saved in the above analysis to determine the concentration of cuprammonium ion in the new sample.

1. Restore the **Calib** data file.

Notice that the results you saved now appear in a new column titled, "MYMODEL."

2. Choose SPECIAL... ADVANCED REGRESSION... CALIBRATION MODELS... from the Menu bar to display the Calibration Models Analysis dialog box.

3. Enter ***.44*** into the Y (Measured) text box.

4. Enter ***MY MODEL*** into the Fitted Model Statistics text box.

5. Click the Predict X from Y button under the Action portion of the dialog box to turn it "on" (see Figure 6-14)

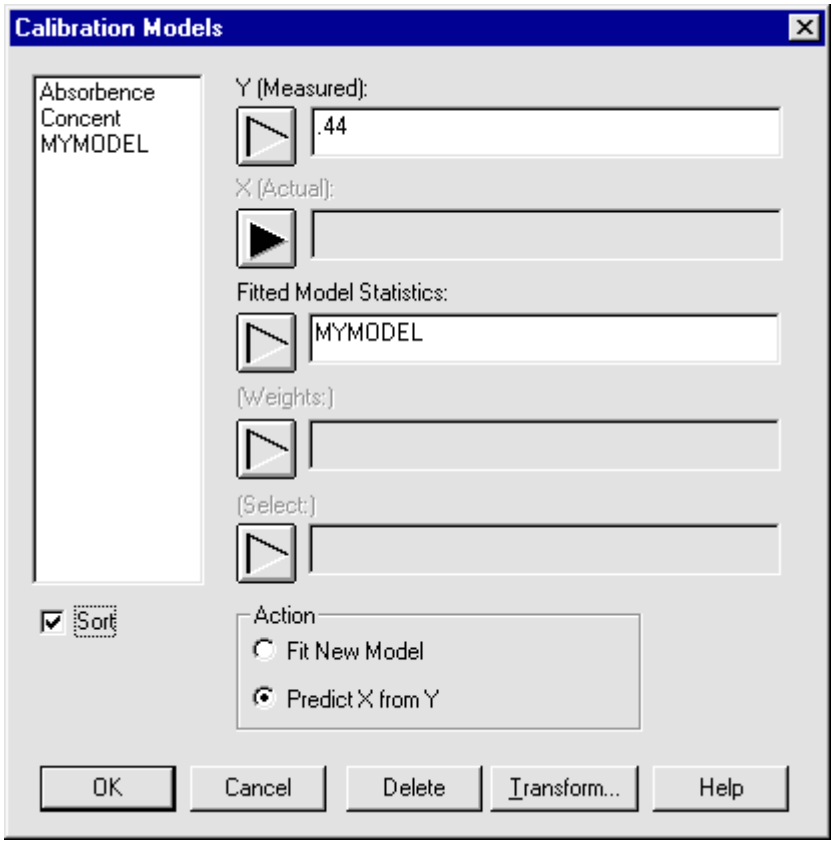


Figure 6-14. Completed Dialog Box for Predicting X from Y

6. Click OK to display the Analysis Summary and the Plot of Fitted Model in the Analysis window.

7. Click the Tabular Options button to display the dialog box, then the Predictions check box, and OK to display the Predictions Table.

8. Maximize the Predictions Table (see Figure 6-15).

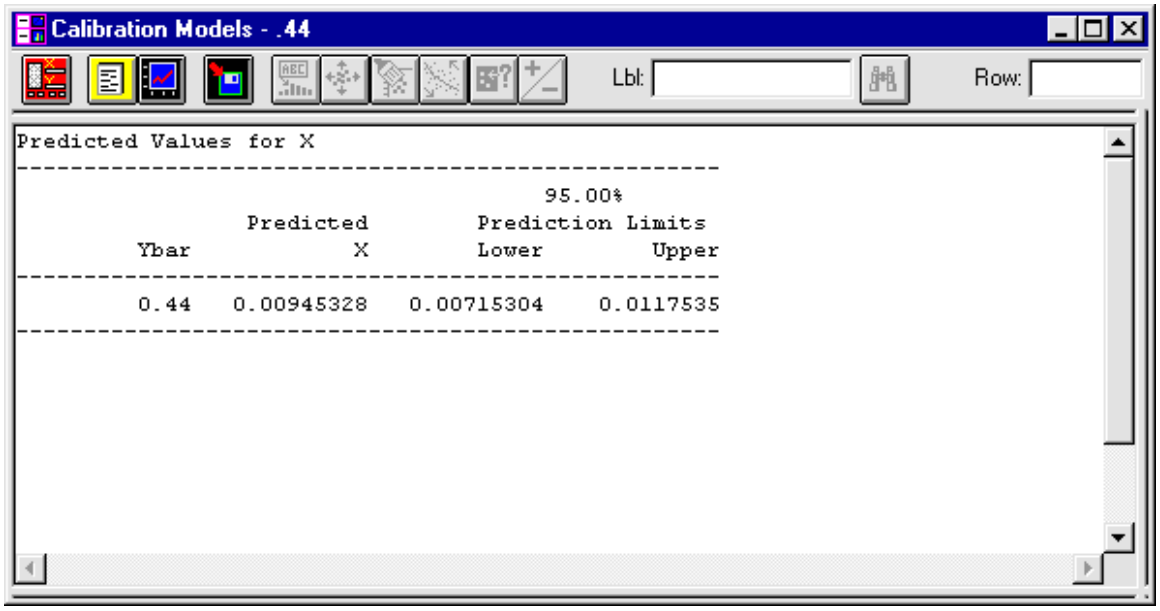


Figure 6-15. The Predictions Table for X

Optional Exercises

In real-life situations, a single measurement for each standard and each new sample would probably not provide adequately tight intervals.

1. Continue the above tutorial using the Plot of Fitted Models Options dialog box and the Predictions Table option to determine the effect of increasing the number of trials. Increase the mean size or weight to ***5***, and compare the interval widths with those for a single measurement.

2. Refit the data using the next best model from the Calibration Model Options dialog box.

3. Create other tabular and graphical options, especially Hypothesis Tests and Unusual Residuals.

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.

TUTORIAL 7

Analyzing An Insurance Innovation Study

This tutorial was adapted from Neter et al. (1996), where an economist decided to compare the speed at which a particular insurance innovation was accepted (Y) with the size of the insurance firm (X_1) and the type of firm. The economist measured the response variable by the number of months that elapsed before the firm accepted the innovation.

The study included three variables:

- Size of the firm, which is quantitative, and measured by the amount of the firm's total assets (in millions of dollars).
- Type of firm, which is qualitative, and identifies two classes stock companies and mutual fund companies.
- Time, which represents the speed with which a firm initiated a particular innovation.

The economist wanted to compare regression model estimates across groups 10 mutual fund firms and 10 stock firms. The data for these firms are in the **Insurance** file. **Time** is the dependent variable (Y), **Size** is the independent variable (X), and **Type** is the level code, a character variable that represents the type of firm, either a stock or a mutual fund brokerage.

The purpose of the tutorial is to determine if a regression analysis performed on the **Time** versus **Size** variables can be improved by taking into account the effect of the type of firm. You will complete a regression analysis on each group to see if the slopes and/or intercepts differ significantly between the groups.

To begin the analysis, open the **Insurance** data file.

Estimating the Model without Groups

It is usually a good idea to look for bias in a model when the data come from, or are suspected to come from, distinct groups. The first step is to run a simple regression analysis on all the data.

1. Choose SPECIAL... ADVANCED REGRESSION... COMPARISON OF REGRESSION LINES... from the Menu bar to display the Comparison of Regression Lines Analysis dialog box.

2. Enter **Time** into the Dependent Variable text box.

3. Enter **Size** into the Independent Variable text box.

4. Enter **Type** into the Level Codes text box (see Figure 7-1).

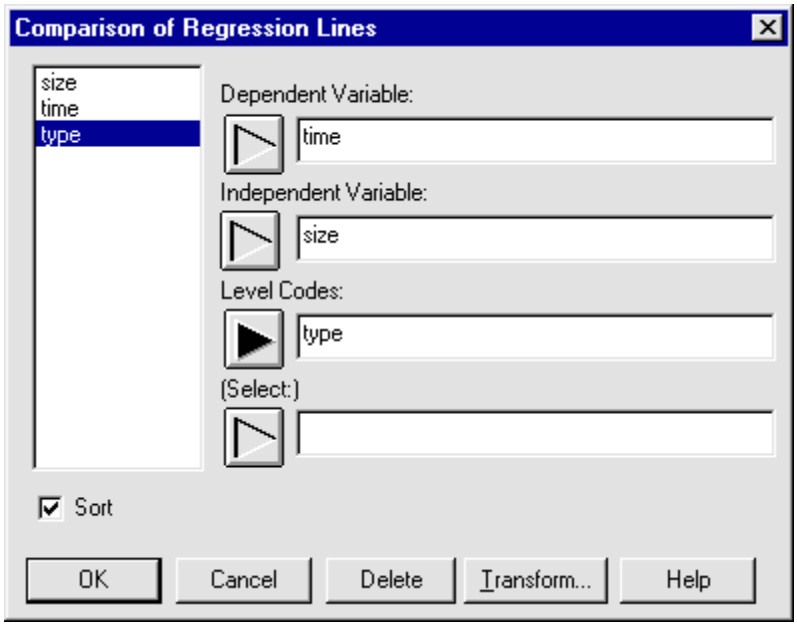


Figure 7-1. Completed Dialog Box for the Comparison of Regression Lines Analysis

5. Click OK to display the Analysis Summary and the Plot of Fitted Model in the Analysis window. Maximize the Analysis Summary.

6. Click the right mouse button on the Analysis Summary pane, then the left on Analysis Options to display the Comparison of Regression Lines Options dialog box.

7. Click the Assume Equal Intercepts and Assume Equal Slopes check boxes to turn on both of the options, which results in a single regression line (see Figure 7-2).

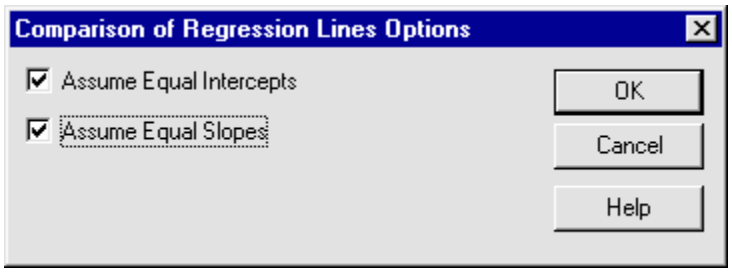


Figure 7-2. Comparison of Regression Lines Options Dialog Box

8. Click OK to redisplay the Analysis Summary (see Figure 7-3).

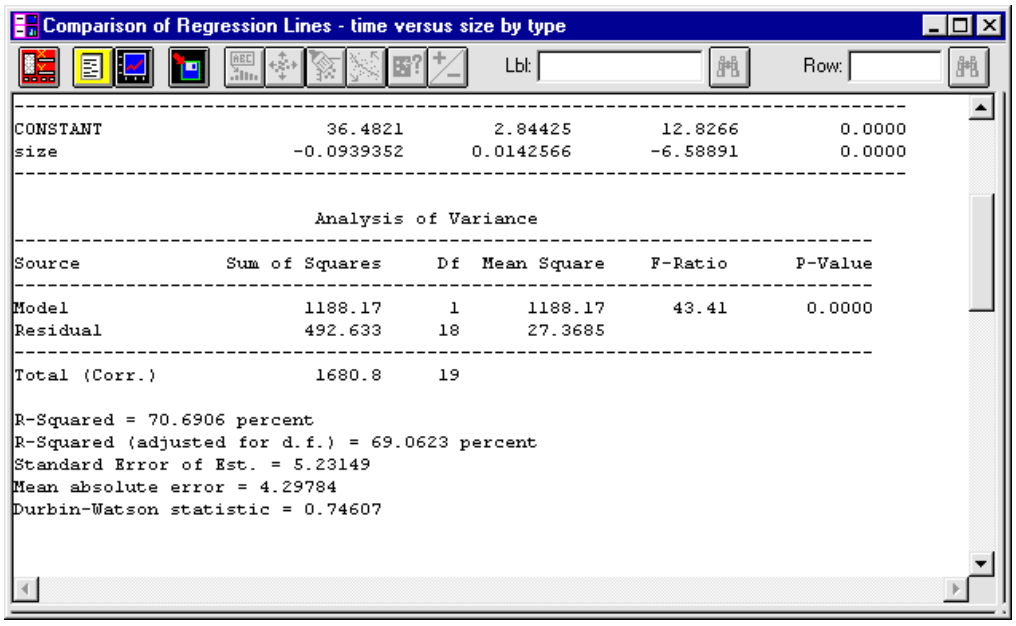


Figure 7-3. The Analysis Summary Showing Various Statistics

The values from the regression analysis and the analysis of variance seem to be within normal ranges; however, note that the value for the Durbin-Watson statistic is 0.74607, which indicates possible serial correlation (the value is less than 1.4). To confirm this, you will create a Residual versus Row Number Plot and an Autocorrelation Function Plot.

9. Click the Graphical Options button to display the dialog box, then the Residual Plots check box, and OK to display the Residual Plot in the second graphics pane. Maximize the plot.

10. Click the right mouse button, then the left on Pane Options to display the Residual Plots Options dialog box.

11. Accept the default options, Studentized Residuals and Scatterplot; move to the Plot versus: list box and choose Row Number. Your screen should look like that shown in Figure 7-4.

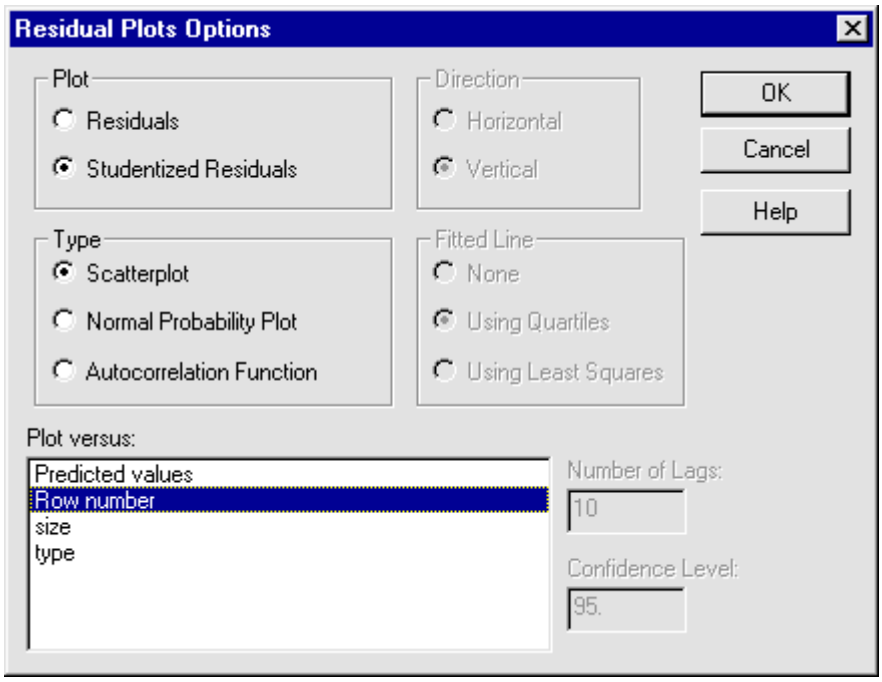


Figure 7-4. The Completed Residual Plots Options Dialog Box

12. Click OK to display the Residual versus Row Number Plot (see Figure 7-5).

The Residual versus Row Number Plot reveals a pattern; most of the first half of the data is below the zero line while most of the second half is above the zero line. This indicates a biased model and confirms that using a single regression from the two groups is probably not adequate.

Now create an Autocorrelation Function Plot.

13. Click the right mouse button on the graphics pane, then the left on Pane Options to display the Residual Plots Options dialog box.

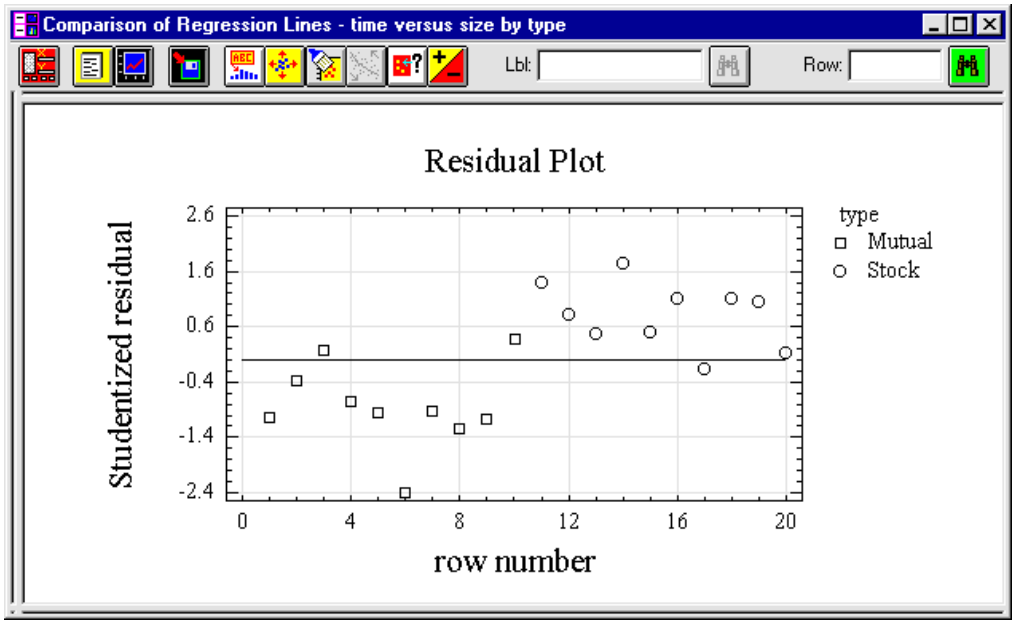


Figure 7-5. The Residual versus Row Number Plot

14. Accept Studentized Residuals as the default for the type of data that will appear in the plot.

15. Click the Autocorrelation Function Plot check box; accept the defaults in the Number of Lags and Confidence Level text boxes, then click OK to display the Autocorrelation Function Plot shown in Figure 7-6.

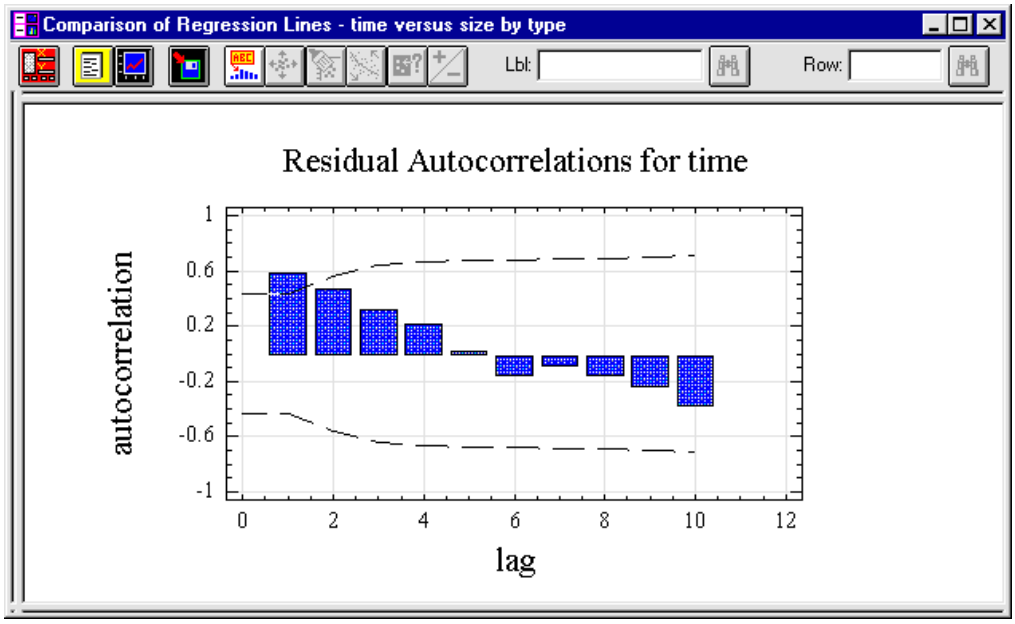


Figure 7-6. The Autocorrelation Function Plot

Remember that the Durbin-Watson statistic was less than 1.4, which raised the suspicion that there was serial correlation. The Autocorrelation Function Plot confirms that suspicion; the first bar extends above the upper probability limit.

The results from the analysis support the notion that analysts often group data in an effort to improve the prediction capability of a model. You can safely conclude that the model is inadequate based on these results:

- the data were presorted by type of firm
- the Durbin-Watson statistic was less than 1.4, indicating the possibility of serial correlation (confirmed by the Autocorrelation Function Plot).

Detecting bias would be more difficult if unsorted data had been used.

Determining How Groups Affect Regression

In this portion of the tutorial, you will determine exactly how regression lines differ between the groups.

1. Minimize the Autocorrelation Function Plot.
2. Maximize the Analysis Summary.
3. Click the right mouse button, then the left on Analysis Options to display the Comparison of Regression Lines Options dialog box.
4. Click the Assume Equal Intercepts and Assume Equal Slopes check boxes to turn off both the options, allowing a separate regression line to be fit for each group.
5. Click OK to redisplay the Analysis Summary (see Figure 7-7).

The results describe the relationship among the variables **Time**, **Size**, and **Type**. The equation for the fitted model is:

Time = 33.8384 - 0.101531***Size***
+ 8.13125***(Type = Stock)***
- 0.000417141***Size*(Type = Stock)**

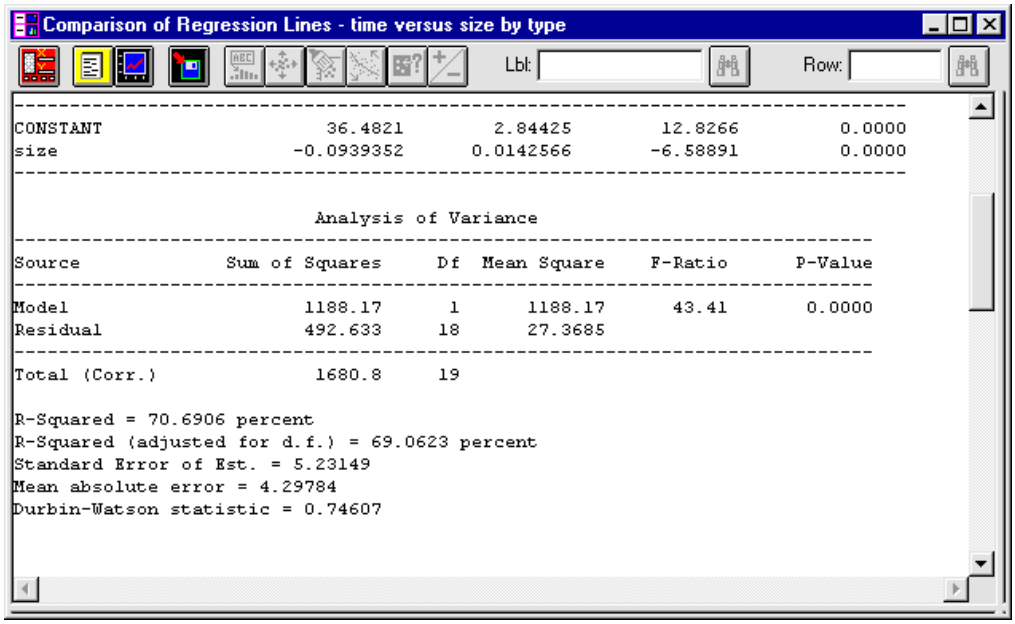


Figure 7-7. The Analysis Summary Results for Fitting a Linear Regression Model

where the term **Type = Stock** is the indicator variable that takes the value 1 if true and 0 if false. This corresponds to two separate lines, one for each value of **Type**. For example, when **Type = Mutual**, the model reduces to

Time = 33.8384 = 0.101531***Size**

because the other two terms become zero. Similarly, when **Type = Stock**, the model reduces to

Time = 41.9696 - 0.101948***Size**

The R-Squared statistic indicates that the model as it was fitted explains 90.5061 percent of the variability in the **Size** variable. The Adjusted R-Squared statistic, which is more suitable for comparing models that have different numbers of independent variables, is 87.5385 percent. For an explanation of the other statistics, read the explanation offered by the StatAdvisor.

Now compare the data from the two analyses. Notice that the values in the second Analysis Summary for the R-Squared, Adjusted R-Squared, Standard Error of Estimates, Mean Absolute Error, and the Durbin-Watson statistics have all improved over the values in the first analysis (refer to Figure 7-2). Also notice that the value for the parameter, **Size*Type** = Stock, -0.000417141, is very close to zero and has a *p*-value of .9821, which means that it is an unnecessary term (the slopes are virtually equal).

To test for statistically significant differences between the two intercepts and the two slopes, you will create a Conditional Sums of Squares Table.

6. Click the Tabular Options button, then the Conditional Sums of Squares check box, and OK to display the table in the text pane.

7. Maximize the table (see Figure 7-8).

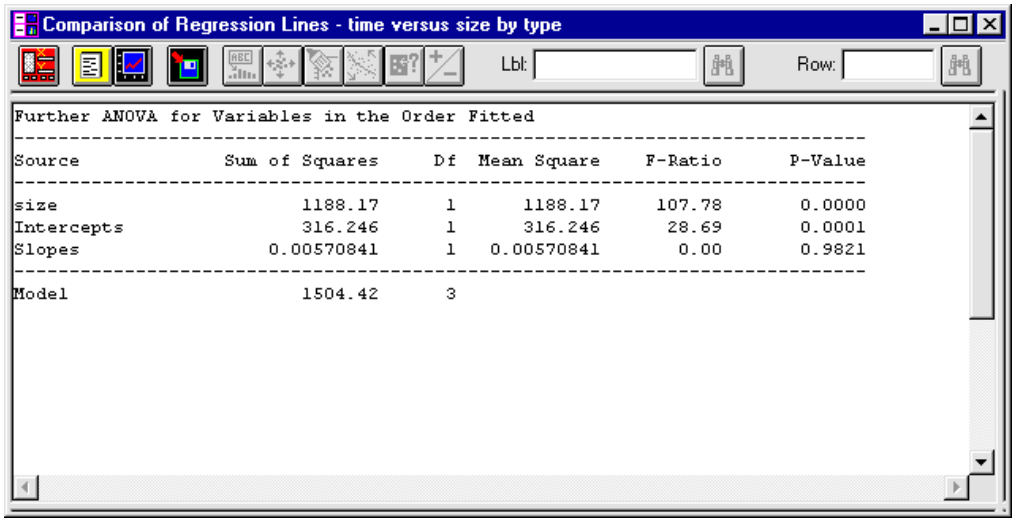


Figure 7-8. The Conditional Sums of Squares Table Showing the Statistically Significant Differences among the Intercepts

Refining the Model

The *p*-value for the intercepts is less than 0.01, which indicates that there is a statistically significant difference among the intercepts at the 99 percent confidence level. However, the *p*-value for the slopes is greater than 0.10, which indicates that there is not a statistically significant difference between the slopes for the two values for the **Type** variable at 90 percent or higher confidence level. Because the slopes are not significantly different, it makes sense to simplify the model by forcing equal slopes.

1. Click the right mouse button on the Conditional Sums of Squares pane, then the left on Analysis Options to display the Comparison of Regression Lines Options dialog box.

2. Click the Assume Equal Slopes check box, then OK to recalculate and redisplay the table with slopes equal (see Figure 7-9).

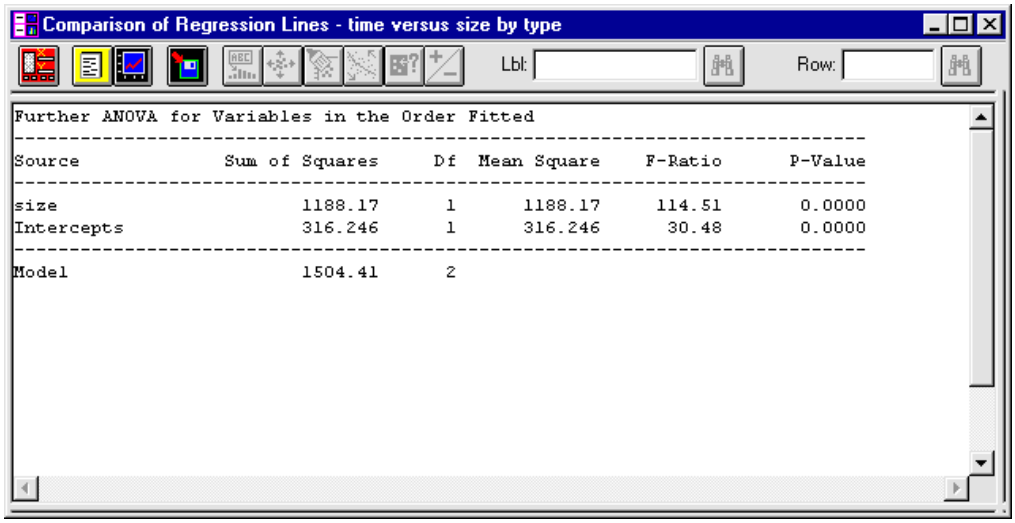


Figure 7-9. The Conditional Sums of Squares for the Assume Equal Slopes Option

The conclusion for this section of the tutorial is that only the intercepts differ significantly; therefore, you can conclude that there is an additive relationship due to the type of firm -- the two regression lines are parallel.

Using the Model

You forced equal slopes by eliminating the unnecessary interaction term from the model (**Time**). Look at the Analysis Summary again.

1. Minimize the Conditional Sums of Squares Table, then maximize the Analysis Summary pane (see Figure 7-10).

   The R-Squared and Adjusted R-Squared statistics reveal that the reduction in the R-Squared statistic is minor, which is worth the tradeoff for a simpler model. The value for the Standard Error of the Estimate has improved, indicating less bias in the model. The value for the Durbin-Watson statistic is still reliable. The Residual plots no longer show any problems (to verify this, you can optionally generate the Residual plots).

   The equation for the final model is:

   Time = 33.8741 - 0.101742***Size**
   + 8.05547*(**Type** = **Stock**)

   The conclusion is that a stock company will take about eight months longer to accept an innovation, than will a mutual fund company of the same size.

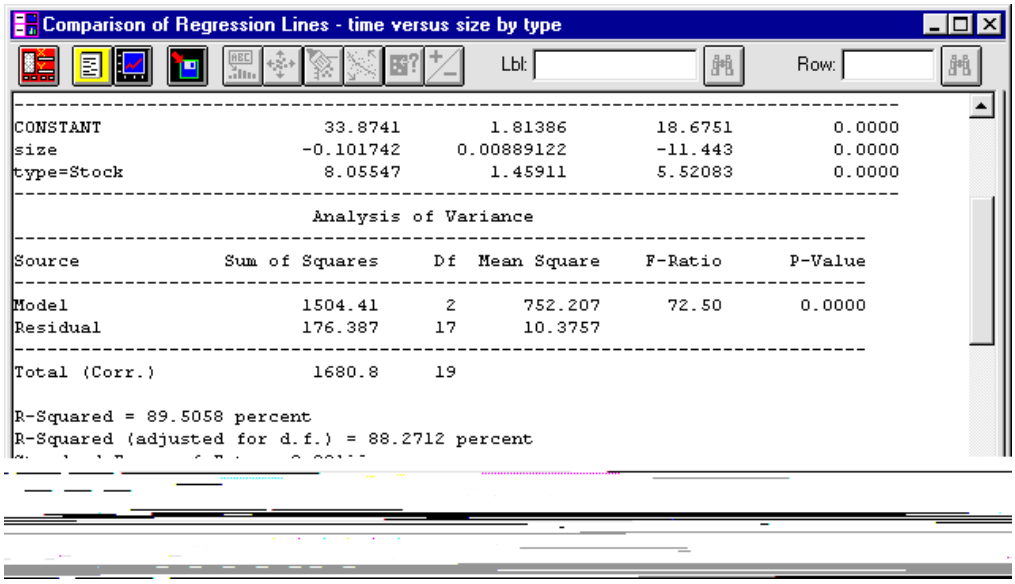


Figure 7-10. The Results of Forced Equal Slopes

To make predictions using this model, notice that when **Type** = **Mutual**, the model reduces to:

Time = 33.8741 - 0.101742***Size**

When **Type** = **Stock**, the model reduces to:

Time = 41.9295 - 0101742***Size**

You now decide to estimate the time it takes for firms with revenues between $100 and $200 million to accept the innovation.

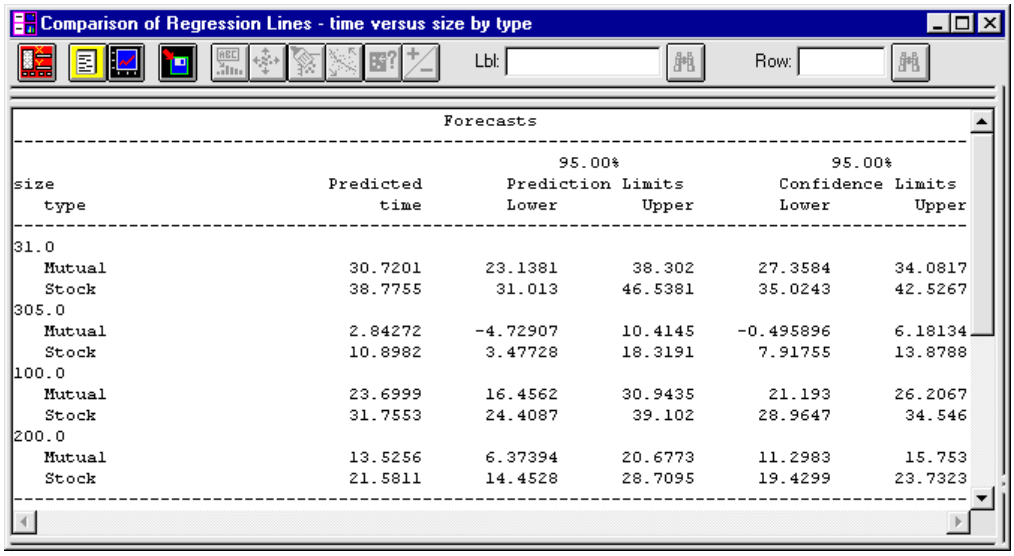
To calculate the predictions, you will generate a Forecasts Table, then use the Forecasts Options dialog box to add the two new observations to the Forecasts Table.

2. Click the Tabular Options button to display the dialog box, then click the Forecasts check box, and OK to display the Forecasts Table. Maximize the table (see Figure 7-11).

Forecasts are shown for the minimum and maximum values of the **Size** variable.

3. Click the right mouse button, then the left on Pane Options to display the Forecasts Options dialog box.

4. Type ***100*** in the third text box, ***200*** in the fourth text box. The dialog box should look like the one shown in Figure 7-12.



Forecasts

size type	Predicted time	95.00% Prediction Limits Lower	 Upper	95.00% Confidence Limits Lower	 Upper
31.0					
Mutual	30.7201	23.1381	38.302	27.3584	34.0817
Stock	38.7755	31.013	46.5381	35.0243	42.5267
305.0					
Mutual	2.84272	-4.72907	10.4145	-0.495896	6.18134
Stock	10.8982	3.47728	18.3191	7.91755	13.8788
100.0					
Mutual	23.6999	16.4562	30.9435	21.193	26.2067
Stock	31.7553	24.4087	39.102	28.9647	34.546
200.0					
Mutual	13.5256	6.37394	20.6773	11.2983	15.753
Stock	21.5811	14.4528	28.7095	19.4299	23.7323

Figure 7-13. The Results of the Recalculated Forecasts

Now that you have examined all of the tabular data, you will create a Plot of Fitted Model to view the results graphically.

6. Minimize the tabular options, then maximize the Plot of Fitted Model (see Figure 7-14).

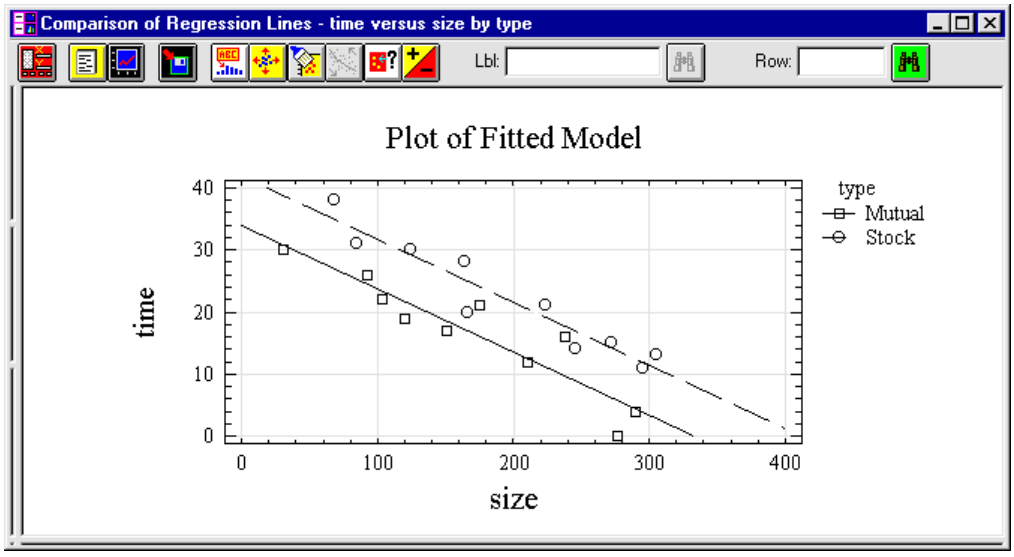


Figure 7-14. The Plot of Fitted Model

The plot shows the two parallel regression lines, one for each value of the **Type** variable. If you look at the Residual plots again, you will find that the earlier symptoms of bias have been resolved.

References

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

TUTORIAL 8

Illustrating Model-Building Techniques

This tutorial illustrates model-building techniques by working through the surgical unit example in Neter et al. (1996). The simple example is based on an exploratory observational study that contained four potential explanatory variables. Limiting the number of potential explanatory variables helps to illustrate the process.

The focus of the example was predicting the survival rate for patients who were undergoing a particular type of liver surgery. The hospital surgical unit randomly selected 54 patients; from each patient record, they extracted the following preoperative information:

- A = Blood clotting test score
- B = Prognostic index, which included the patient s age
- C = Enzyme function test score
- D = Liver function test score

This information makes up a pool of potential explanatory variables for a predictive regression model. The response variable is **Logsurv (**log survival time), which was determined in a follow-up study.

Because the pool of explanatory variables is small, at this stage you can fully explore the relationships and possible strong interaction effects. The researchers first prepared a Stem-and-Leaf display (optional exercise) for each of the explanatory variables. This highlighted several cases as outliers, with respect to the dependent variable, and reminded the researchers that they would later need to examine these cases. They examined the full model and decided to use a log transformation as the survival variable to allow for a first-order model. Next, they produced a Scatterplot Matrix and a Correlation Matrix to check for multi-collinearity and bias (optional exercises).

To begin the analysis, open STATGRAPHICS *Plus* and the **Surgery** data file, then continue with the analysis.

Building the Model

1. Choose SPECIAL... ADVANCED REGRESSION... REGRESSION MODEL SELECTION... from the Menu bar to display the Regression Model Selection Analysis dialog box.

2. Enter **Logsurv** into the Dependent Variable text box.

The name of the variable is the logarithmic transformation Y = $\log_{10}$Y, which the researchers used to make the distribution of the error terms more nearly normal and to reduce the BC interaction effect.

3. Enter **Clotting**, **Prognost**, **Enzyme**, and **Liver** into the Independent Variables text box.

Note: The variables are labeled alphabetically in the text and graphs. For example, **Clotting** becomes variable A; **Prognost** becomes B, and so on (see Figure 8-1).

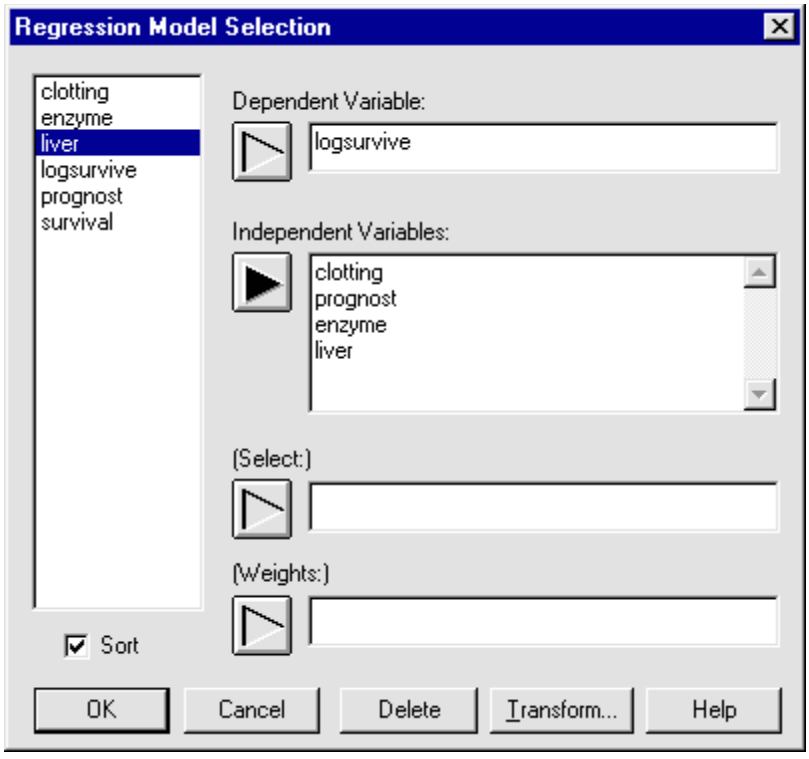


Figure 8-1. Completed Dialog Box for the Regression Model Selection Analysis

4. Click OK to display the Analysis Summary and the Adjusted R-Squared Plot in the Analysis window. Maximize the Analysis Summary.

The Analysis Summary includes values for the single-variable models (labeled as A, B, C, and D). Now you will eliminate these models from the analysis.

5. Click the right mouse button on the Analysis Summary pane, then the left on Analysis Options to display the Regression Model Selection Options dialog box.

6. Enter ***2*** into the Minimum text box to change the minimum number of variables that will be included in the study, accept the default in the Maximum text box. The dialog box on your screen should look like the one shown in Figure 8-2.

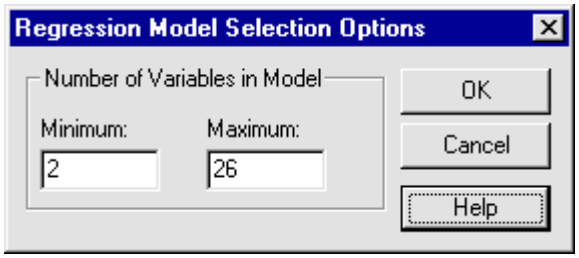


Figure 8-2. Completed Dialog Box

7. Click OK to remove the single-variable models from the analysis and to redisplay the summary (see Figure 8-3).

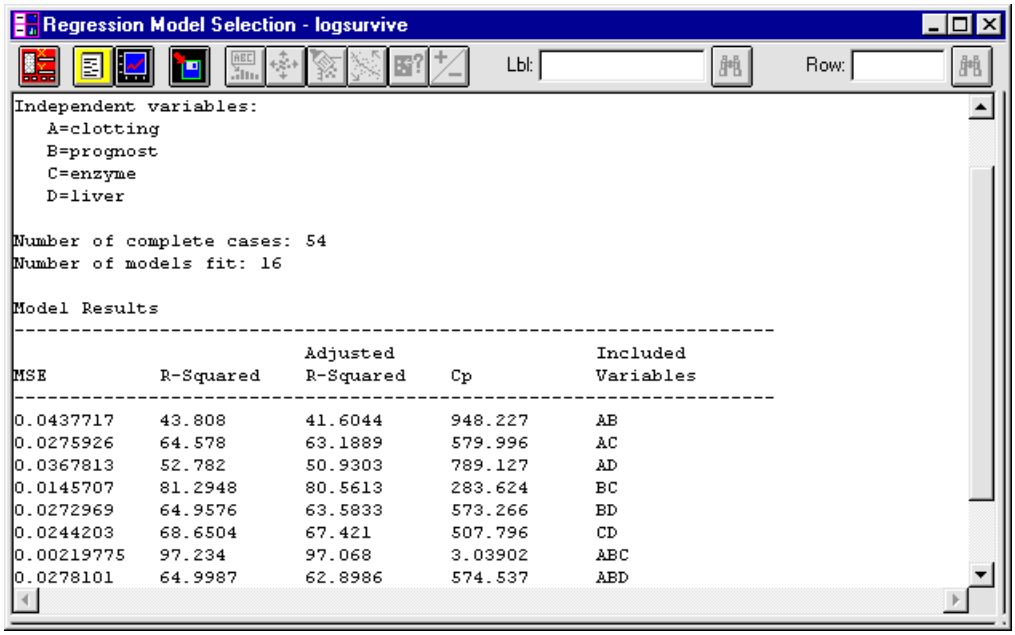


Figure 8-3. The Analysis Summary Redisplayed

The Analysis Summary lists the results of fitting various multiple regression models to describe the relationship among the dependent variable and the different possible subsets of explanatory variables.

The models that were fit contain all possible combinations of two to four variables; the four single-model variables were removed. To determine which of these models is best, you will use the other tabular options.

8. Click the Tabular Options button to display the dialog box, click the All button to choose the remaining options, then OK to display the three tabular options in the text panes. The remaining options are: Best Adjusted R-Squared and Best C_p.

9. Maximize the Best Adjusted R-Squared Table (the second text pane) (see Figure 8-4).

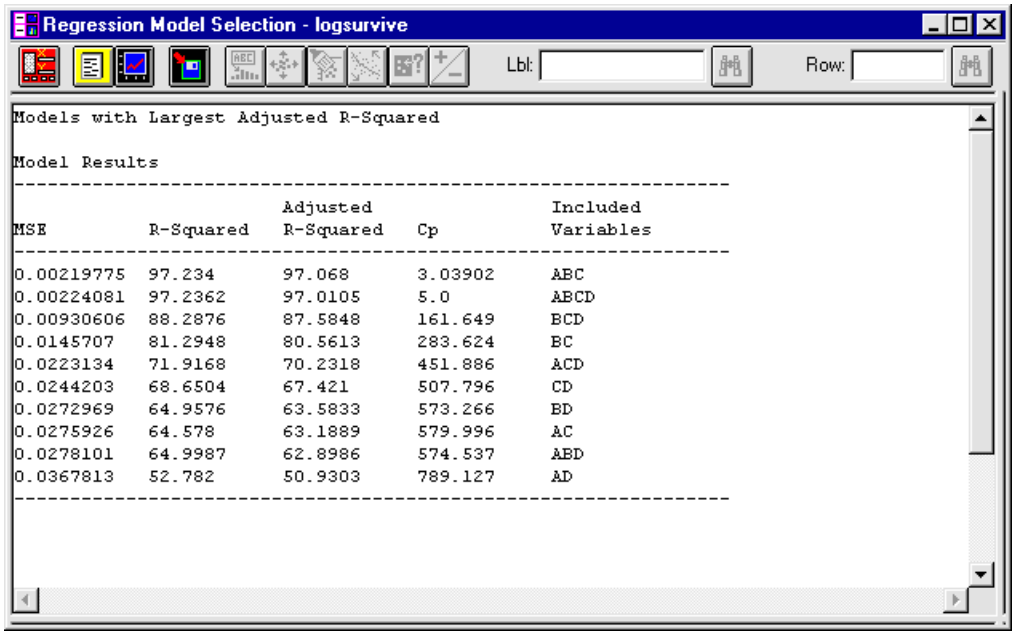


Figure 8-4. The Best Adjusted R-Squared Table

The results show the models sorted by Adjusted R-Squared values with the four single-variable models removed from the analysis. This table indicates that the best model contains three variables: (A) **Clotting**, (B) **Prognost**, and (C) **Enzyme**.

10. Minimize the Best Adjusted R-Squared Table. Maximize the Best C_p Table (the third text pane) (see Figure 8-5).

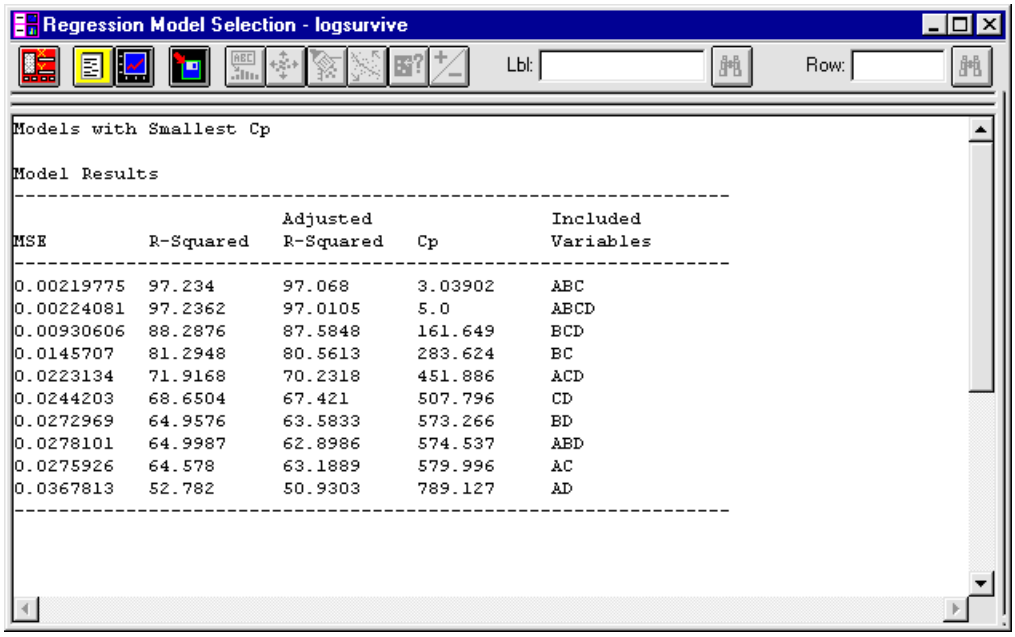


Figure 8-5. The Best C_p Table.

The results show the models sorted by the smallest or best values for the Mallows C_p statistic. Look for models other than the full four-variable model with C_p values that are close to *p*.

Now you will look at one model for each subset size.

11. Minimize the Best C_p Table then maximize the Best Adjusted R-Squared pane.

12. Click the right mouse button on the table pane, then the left on Pane Options to display the Best Adjusted R-Squared Options dialog box.

13. Enter ***1*** into the Maximum Number per Subset text box to include one model for each subset size, then click OK to redisplay the table (see Figure 8-6).

The results show the models that give the largest Adjusted R-Squared values. Values are included for the best model of each size. The best model contains three variables: (A) **Clotting**, (B) **Prognost**, and **Enzyme**.

14. Minimize the table then maximize the Best C_p Table (see Figure 8-7).

The results show the best for each size; again, the ABC model rates best. Note that changing the minimum subset size for the Best Adjusted R-Squared Table automatically changed it in the Best C_p Table.

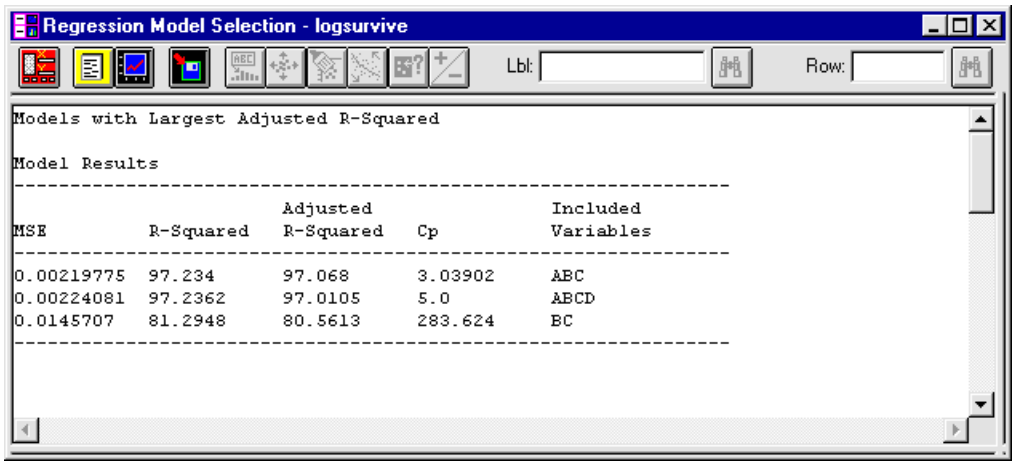


Figure 8-6. Recalculated Best Adjusted R-Squared Results

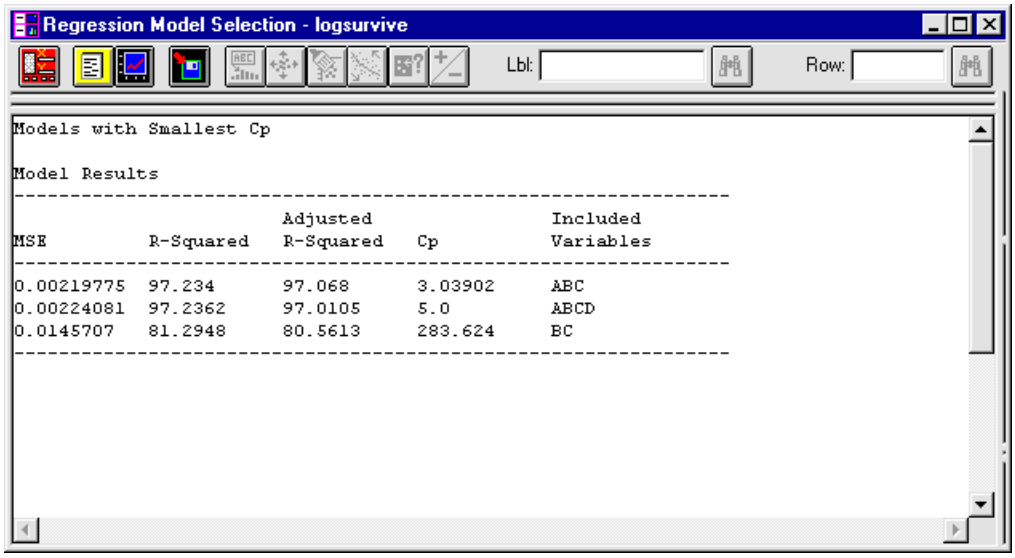


Figure 8-7. The Best C_p Table Shown with the Recalculated Results

Sometimes using the plots will help to confirm which model is best. Create a Mallows C_p Plot to test this.

15. Click the Graphical Options button to display the dialog box, then click the Mallows C_p Plot check box, and OK to display the plot in the second graphics pane.

16. Maximize the plot (see Figure 8-8).

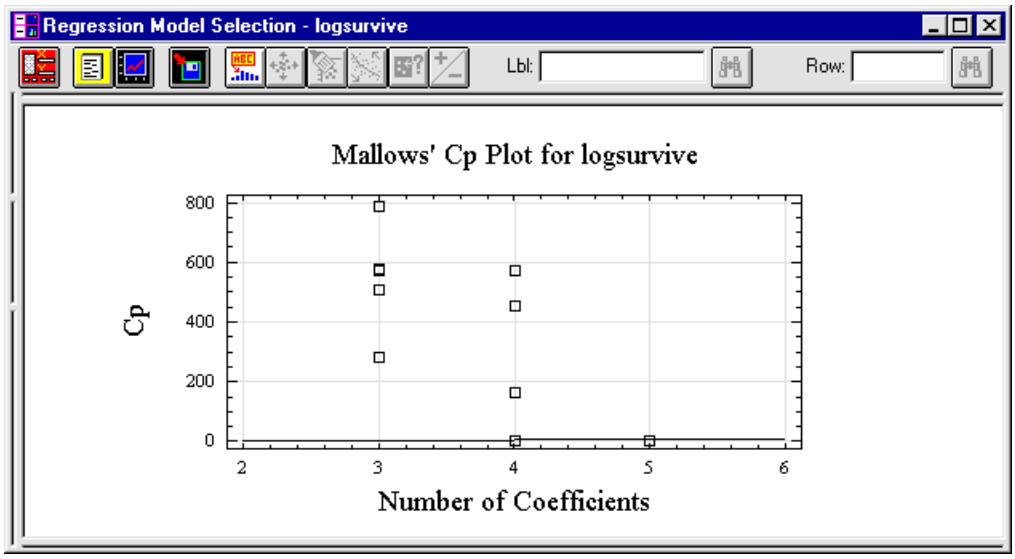


Figure 8-8. Mallows' C_p Plot

To see how close the C_p values (the fourth and fifth coefficients) are to the line, scale the C_p axis from 0 to 10 by 1.

17. Place the mouse pointer on one number of the X-axis scale. Click the left button to place markers around the scale, then click the right to display the pop-up menu. Click Graphics Options to display the Graphics Options dialog box opened to the X-Axis tab page.

18. Enter ***0*** in the From text box, ***10*** in the To text box, ***1*** in the By text box, and accept the defaults for the other text boxes. The dialog box should look like the one shown in Figure 8-9.

19. Click OK to rescale the axis and redisplay the plot (see Figure 8-10).

The plot shows the C_p values for all possible regression models. The three-variable subset (ABC) has the smallest C_p value without an indication of severe bias compared with the full four-variable model. The fact that the C_p measure for this model is below the line $p = C_p$, is the result of random

Neter et al. (1996) noted that, although C_p is on the line for $p = 5$, it is due to the definition of C_p, not because the model that contains all four variables is considered best.

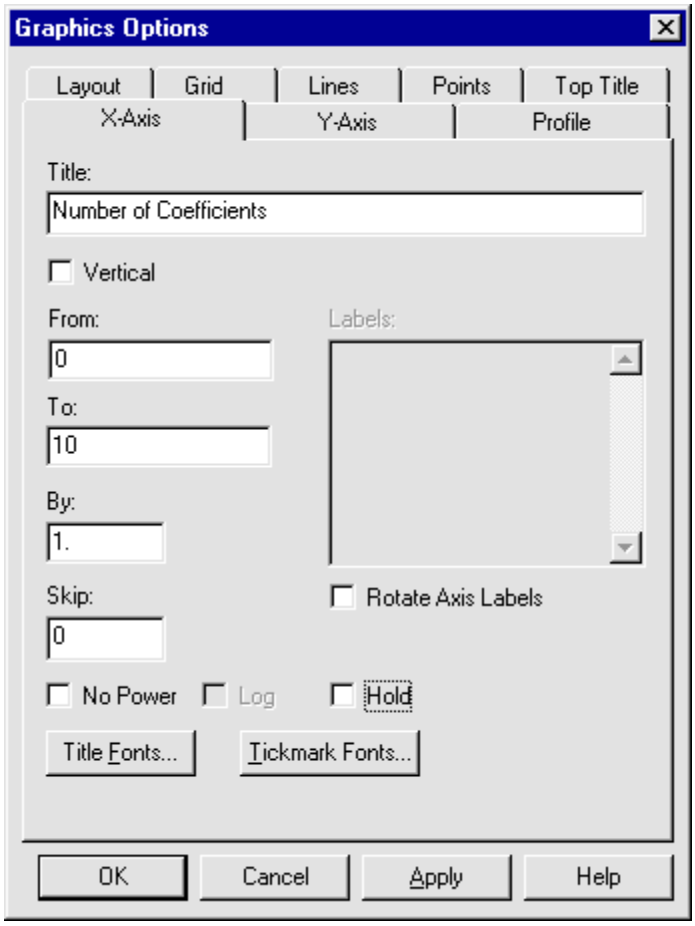


Figure 8-9. The Completed Tab Page

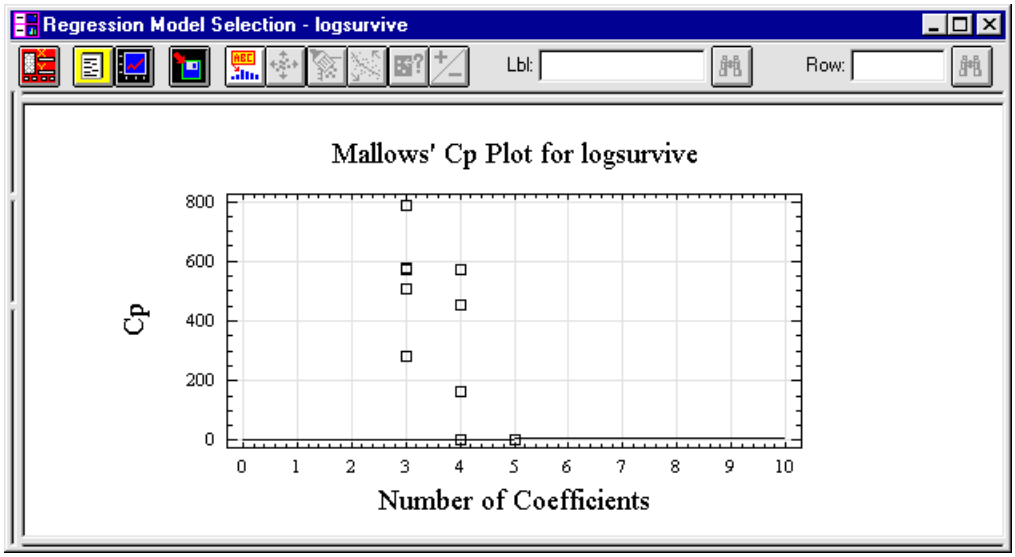


Figure 8-10. Mallows' C_p Plot with the X-Axis Rescaled

Optional Exercise

As an optional exercise, generate the Adjusted R-Squared, R-Squared, and MSE plots. All three plots illustrate that adding the fourth variable provides little or no improvement.

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**:661-675.

Mallows, C. L. 1995. More Comments on C_p. *Technometrics*, **37**:362-372.

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

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

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

TUTORIAL 9

West Virginia Mining Excavation Study

This tutorial is adapted from Myers (1990), who used data collected by the Mining Engineering Department and analyzed by the Statistical Consulting Center at Virginia Polytechnic Institute and State University, Blacksburg, Virginia (1982).

In the study, Myers noted that a major problem connected with mining projects was ground sinking above the excavation or subsidence. To make sure that existing structures did not collapse during an excavation, a mining engineer was responsible for controlling the amount and distribution of the subsidence.

The amount and nature of the subsidence is affected by several factors: the depth of the mine and the width of the excavation. An important variable, known as the angle of draw (y), was identified as an aid in characterizing the condition. Myers defines it as the angle between the perpendicular at the edge of the excavation and the line that connects the same edge of excavation with the point on the surface for which there is zero subsidence.

In the study, the engineers felt that the angle of draw should relate to the ratio of the width (w) of the excavation and the depth (d) of the mine. They also knew that any relationship would be nonlinear. You will repeat the analysis using the data collected at Blacksburg, Virginia.

Begin the analysis by opening STATGRAPHICS *Plus* and the **Mining** data file.

Preparing for the Analysis

1. Choose SPECIAL... ADVANCED REGRESSION... NONLINEAR REGRESSION... from the Menu bar to display the Nonlinear Regression dialog box.

2. Enter **Drawangl** into the Dependent Variable text box.

3. Move the mouse pointer inside the Function text box, then type: ***a*(1-exp(-b*(width/depth)))*** (see Figure 9-1).

4. Click OK to display the Initial Parameter Estimates dialog box.

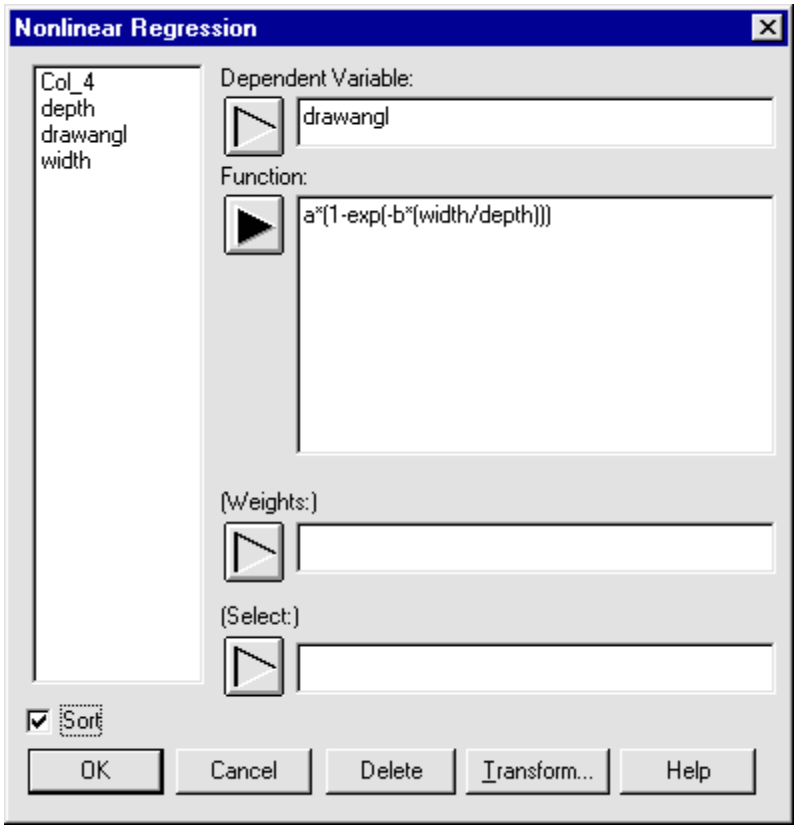


Figure 9-1. Completed Analysis Dialog Box

You need to enter a starting value for each of the parameters. Notice that the two active text boxes, a: and b: are the unknown variables in the function expression.

If you spell the name of a variable incorrectly, the program will interpret the misspelled word as a parameter and enter it into one of the text boxes. For example, you might accidentally spell depth as detph. Remember that if an unexpected parameter name appears in this dialog box, the name is in the function expression but not in the file. See Myers (1990) for the rationale for starting values.

5. Type ***35*** in the a: text box and ***1*** in the b: text box (see Figure 9-2).

Interpreting the Results

1. Click OK to display the Analysis Summary and the Plot of Fitted Model in the Analysis window.

2. Maximize the Analysis Summary (see Figure 9-3).

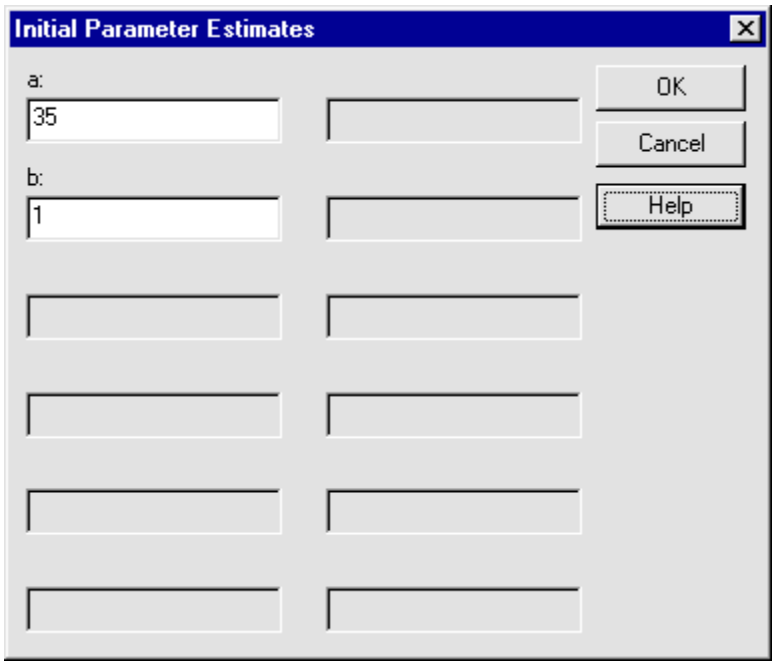


Figure 9-2. Completed Initial Parameter Estimates Dialog Box

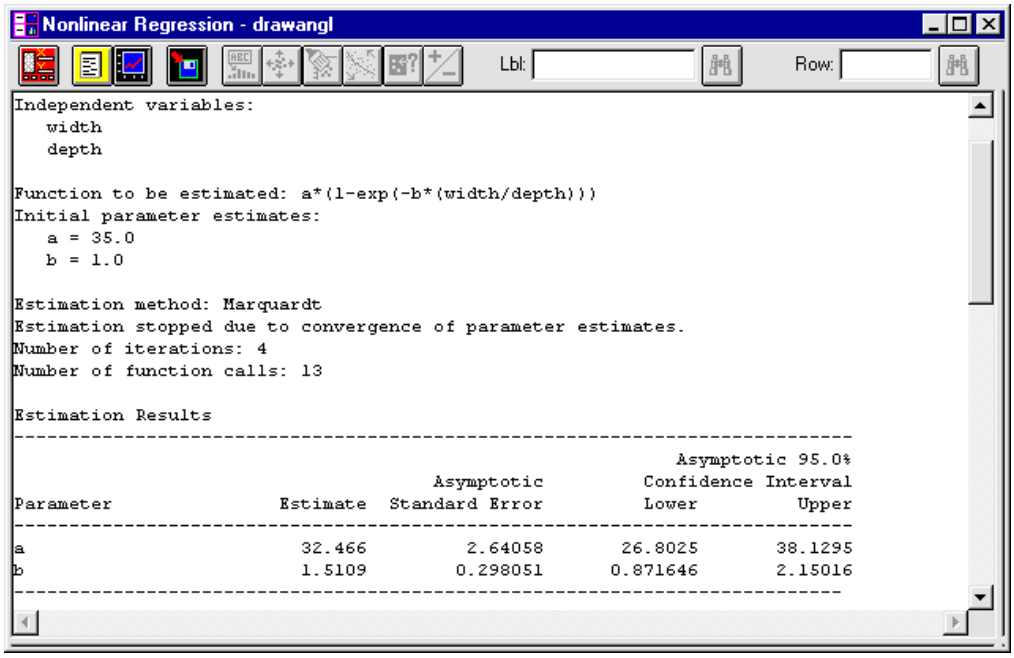


Figure 9-3. The Analysis Summary

The Analysis Summary shows that the estimation process was terminated when it successfully completed four iterations. At this point, the residual sum of squares appeared to approach a minimum.

The R-Squared statistic shows that as the model was fit, it explained 67.2795 percent of the variability in the dependent variable. The value of the Adjusted R-Squared statistic is 64.9423 percent; the Durbin-Watson statistic is greater than 1.4, which indicates that there are no serious autocorrelations in the residuals.

Now you will use the Nonlinear Regression Options dialog box to change the estimation method from the default, Marquardt, to Steepest Ascent. You will do this to increase the number of iterations it takes to get convergence.

3. Click the right mouse button on the text pane, then the left on Analysis Options to display the Nonlinear Regression Options dialog box.

4. Click Steepest Descent in the Method portion of the dialog box, accept the defaults for the remaining options (see Figure 9-4).

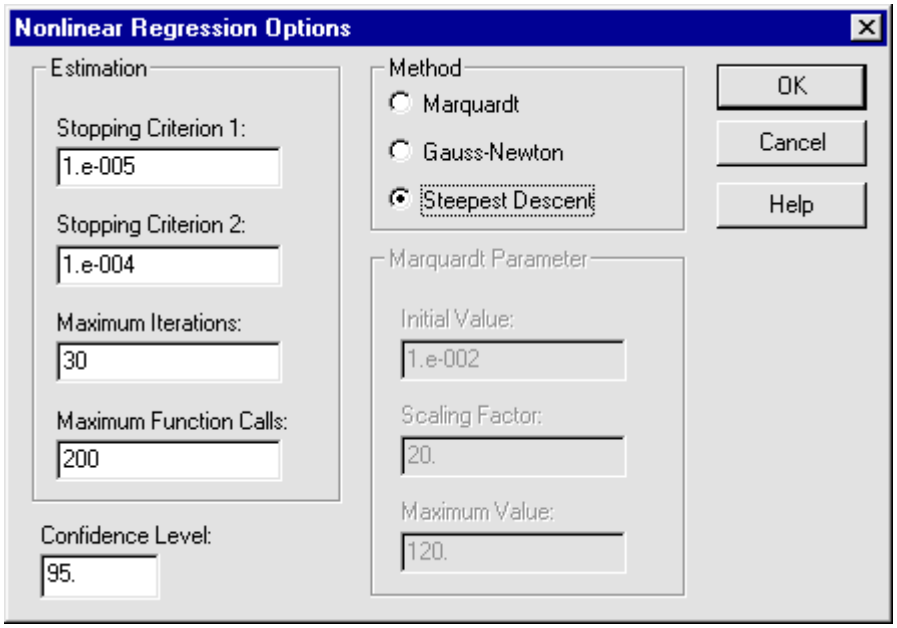


Figure 9-4. Completed Nonlinear Regression Options Dialog Box

5. Click OK to recalculate and redisplay the Analysis Summary using the new method (see Figure 9-5).

Notice that the Analysis Summary contains the name of the estimation method that was used, the reason for the estimation stopping, the number of

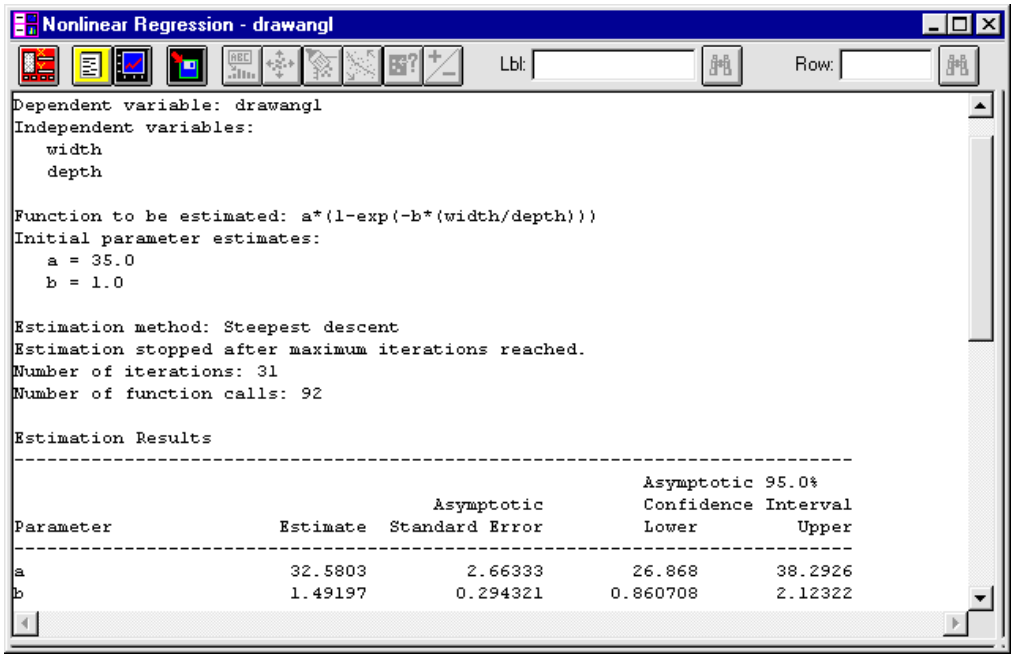


Figure 9-5. The Analysis Summary with Recalculated Results

iterations completed, and the number of function calls. An important factor in the report is the primary reason for the estimation stopping. In this case, the estimation stopped after the program reached the maximum number of iterations; therefore, the methods did not converge.

Now return to the Nonlinear Regression Options dialog box where you will increase the maximum number of iterations.

5. Click the right mouse button on the text pane, then the left on Analysis Options to display the Nonlinear Regression Options dialog box.

6. Type ***50*** in the Maximum Iterations text box; leave the remaining options as they are currently set (see Figure 9-6).

7. Click OK to recalculate and redisplay the Analysis Summary using the new values for the Maximum Iterations (see Figure 9-7).

This time the estimation stopped due to convergence of the residual sum of squares. The summary shows that 40 iterations and 122 function calls were performed.

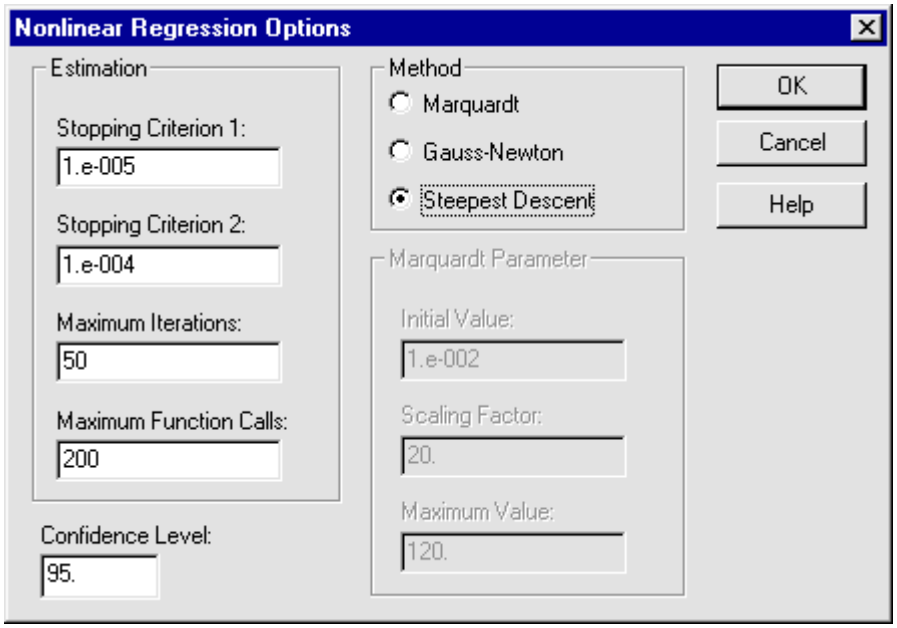


Figure 9-6. Completed Nonlinear Regression Options Dialog Box

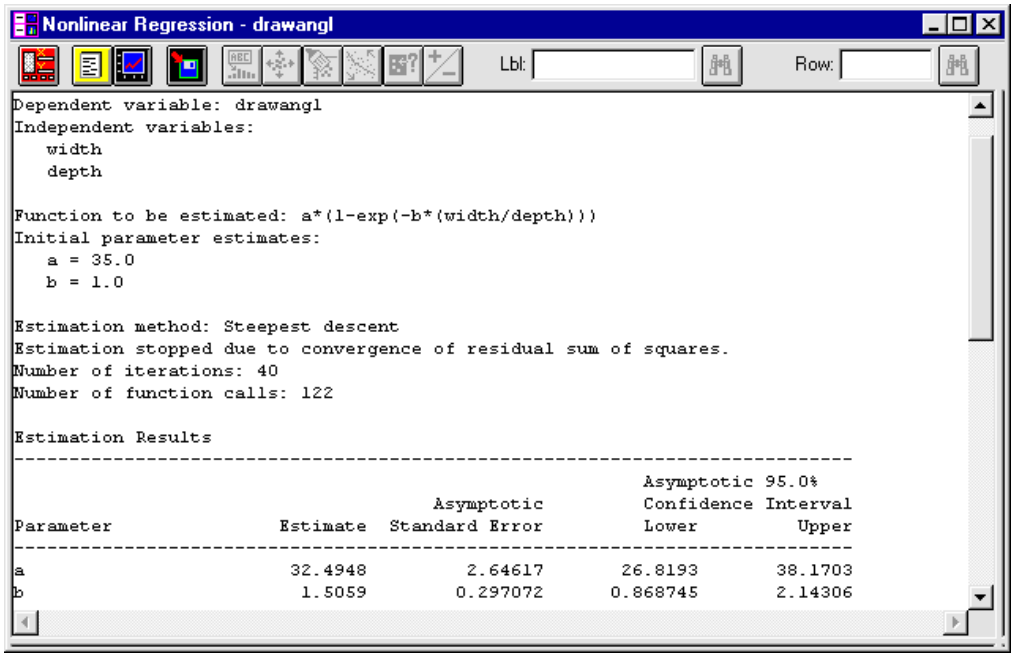


Figure 9-7. Analysis Summary Results Recalculated Using New Iterations

Interpreting Data on Plots

Now you will create a Plot of Fitted Model to see if there is a relationship between the **Drawangl** and **Width** variables.

1. Minimize the Analysis Summary then maximize the Plot of Fitted Model Plot (see Figure 9-8).

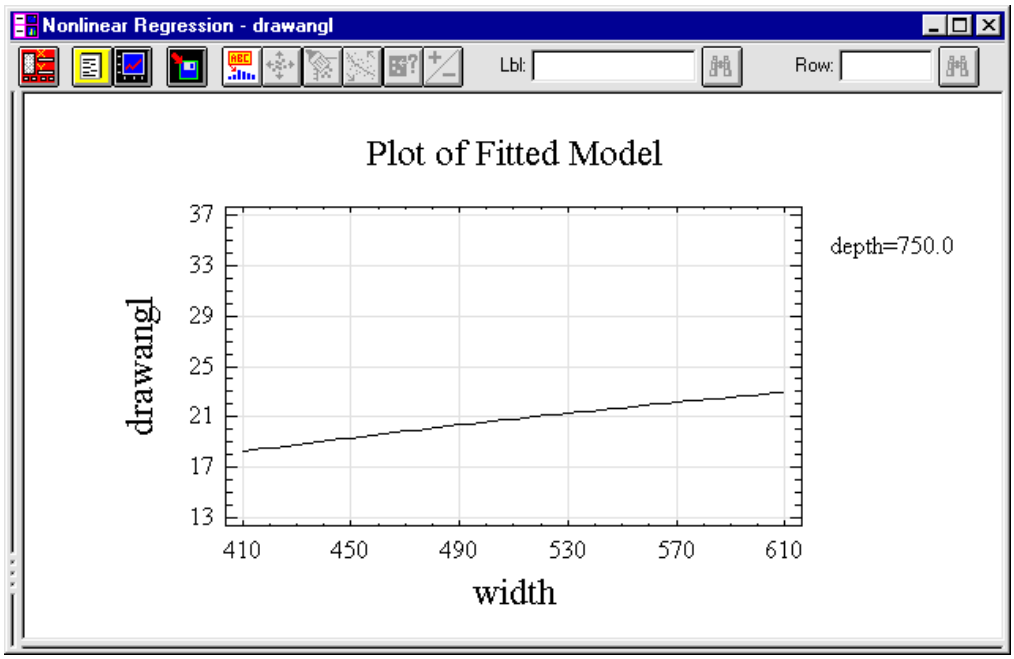


Figure 9-8. Plot of Fitted Model

The plot shows that the relationship between the **Drawangle** and **Width** variables is almost linear over the range of **Width**, when **Depth** is held at 750, which is the middle of the **Depth** range.

Now you will use the options to plot the function versus the other variable.

3. Click the right mouse button, then the left on Pane Options to display the Plot of Fitted Model Options dialog box.

4. Click the **Depth** and **Width** check boxes to select and deselect them, respectively (see Figure 9-9). Notice that when you click the check boxes, the Low, High, and Hold text boxes switch between active and inactive.

5. Click OK to display the Plot of Fitted Model (see Figure 9-10).

The plot is nonlinear but shows a monotonic decrease in the variable **Drawangl** as **Depth** increases over the range of data.

Now you will create a Response Surface Plot to see both of these effects at the same time.

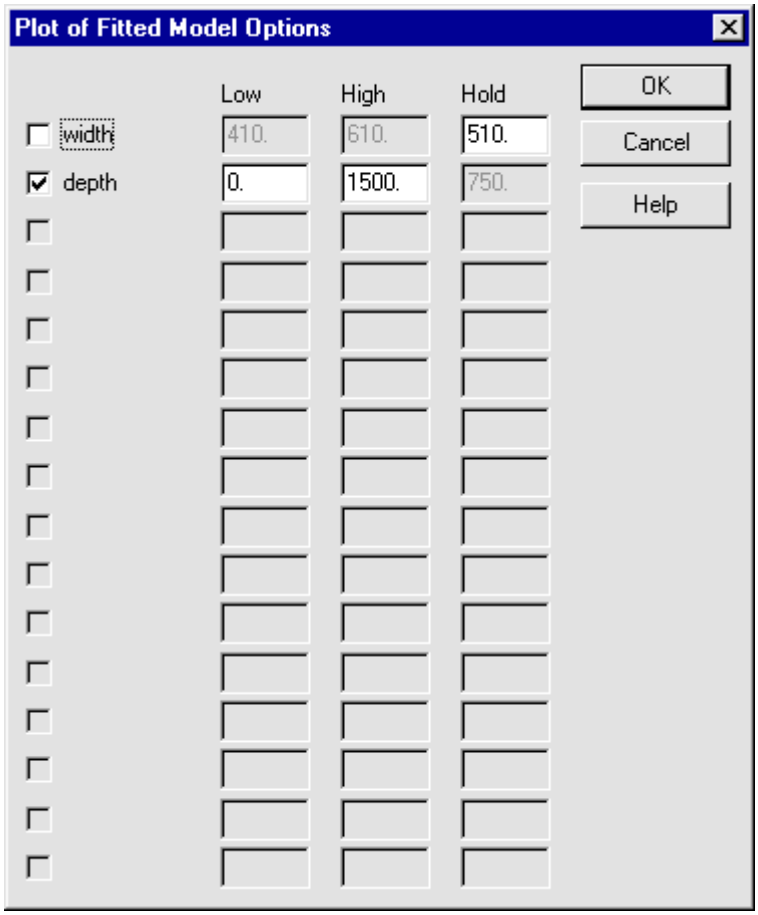


Figure 9-9. Completed Plot of Fitted Model Options Dialog Box

6. Click the Graphical Options button to display the dialog box, then the Response Surface Plot check box, and OK to display the plot in the second graphics pane.

7. Maximize the plot (see Figure 9-11).

The plot shows the fitted surface. Now create a Square Plot to see the results in yet another way.

8. Click the right mouse button on the graphics pane, then the left on Pane Options to display the Response Plot Options dialog box.

9. Click the Square Plot check box; leave the remaining options on the dialog box as they are (see Figure 9-12).

10. Click OK to replace the Square Plot with the Surface Plot (see Figure 9-13).

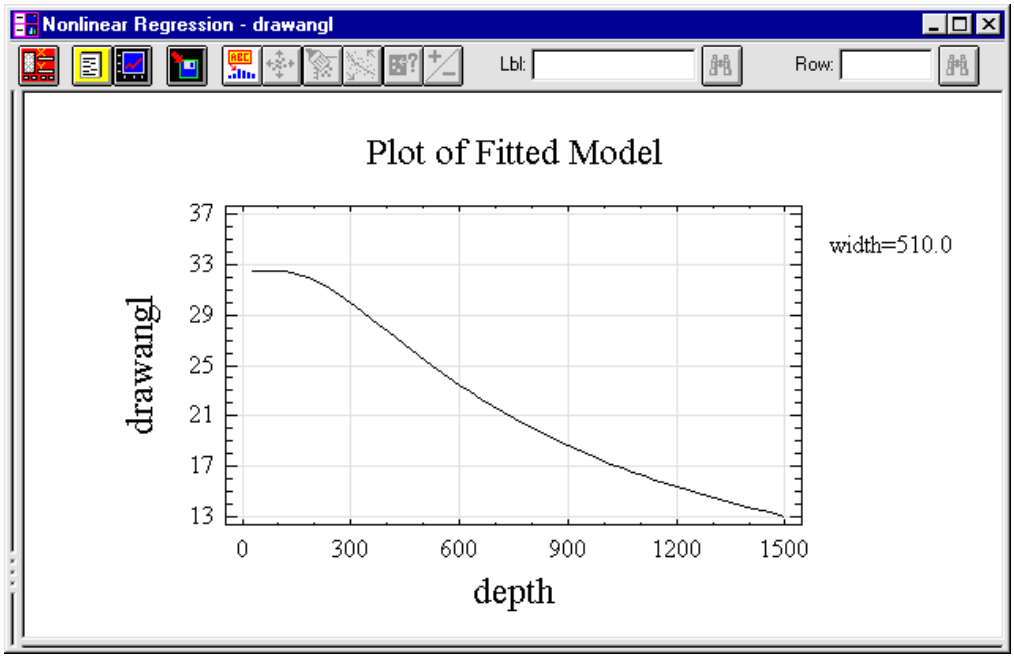


Figure 9-10. Plot of Fitted Model

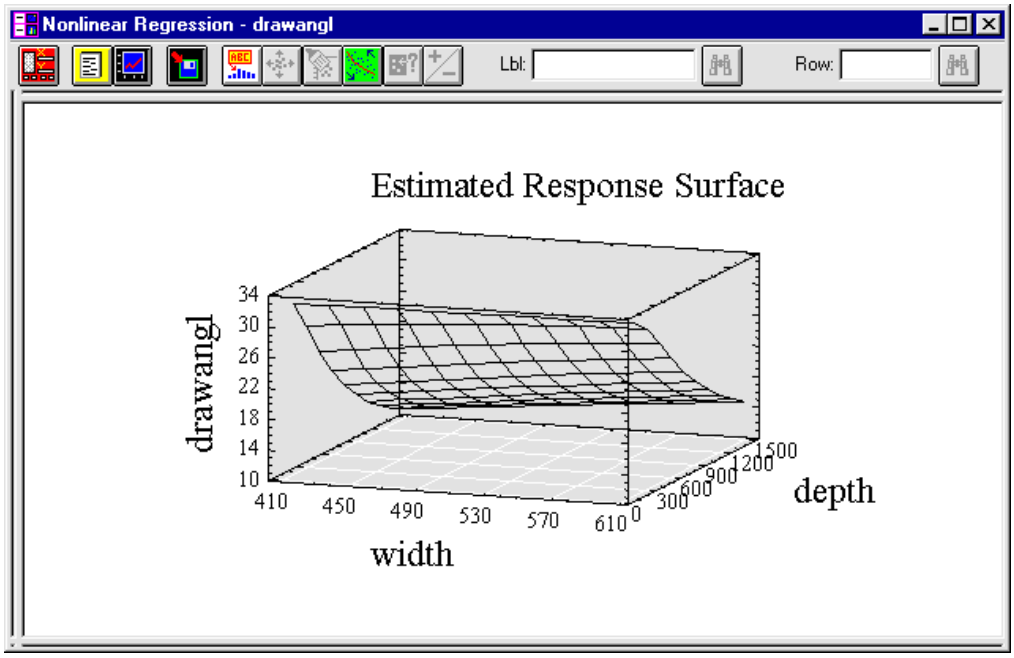


Figure 9-11. The Response Surface Plot

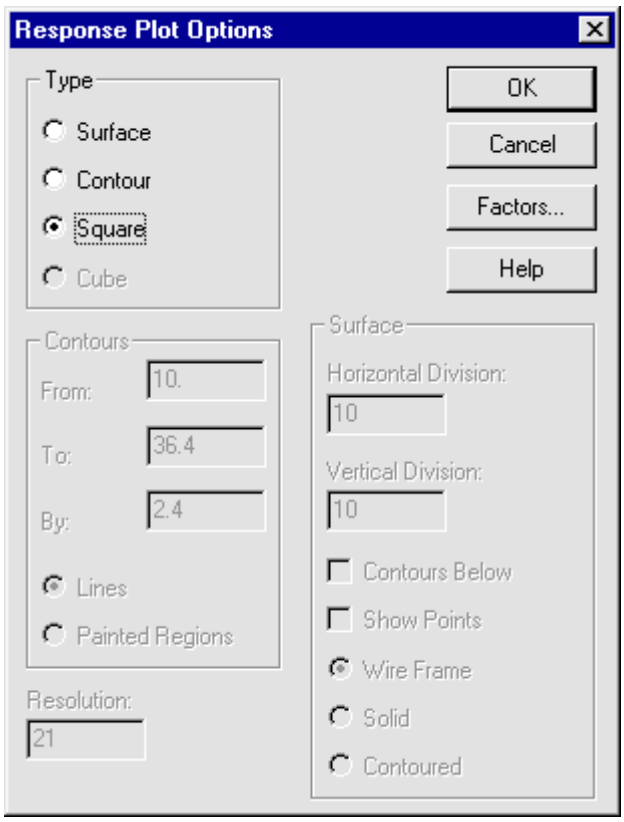


Figure 9-12. Completed Response Plot Options Dialog Box

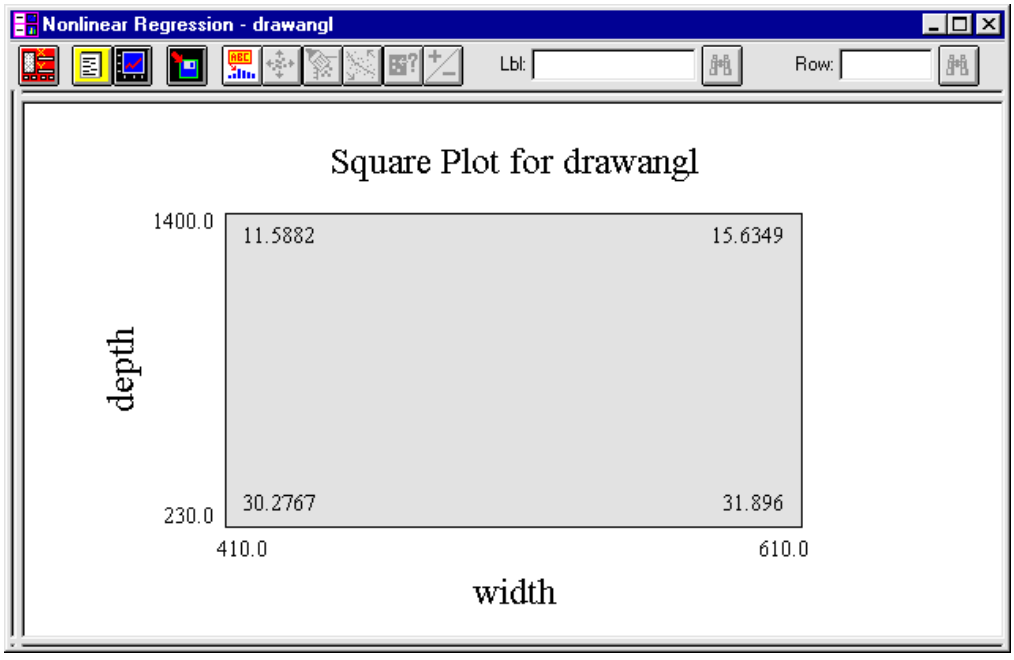


Figure 9-13. Square Plot

Estimating Predictions

1. Click the Tabular Options button to display the dialog box, then the Reports check box, and OK to display the report in the second text pane.

2. Maximize the table (see Figure 9-14).

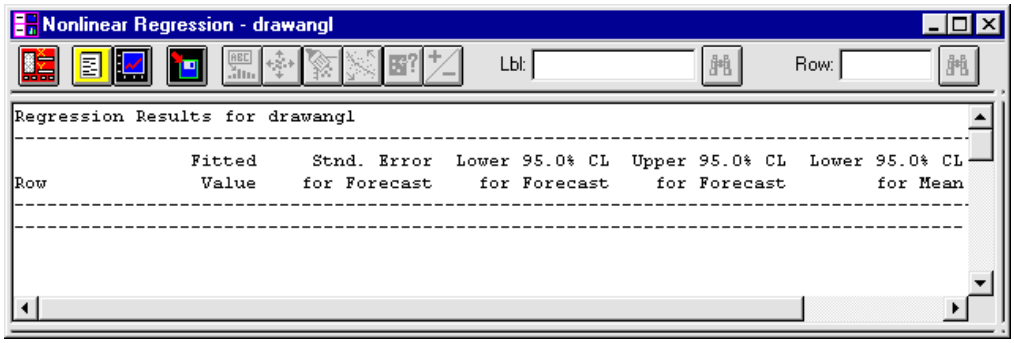


Figure 9-14. The Reports Table

Notice that the table in Figure 9-14 is blank. To correct this, you need to make a change on the Reports Options dialog box.

3. Click the right mouse button on the text pane, then the left on Pane Options to display the Reports Options dialog box.

   Four of the options are currently chosen. You will add one more.

4. Click the Observed Y check box (see Figure 9-15).

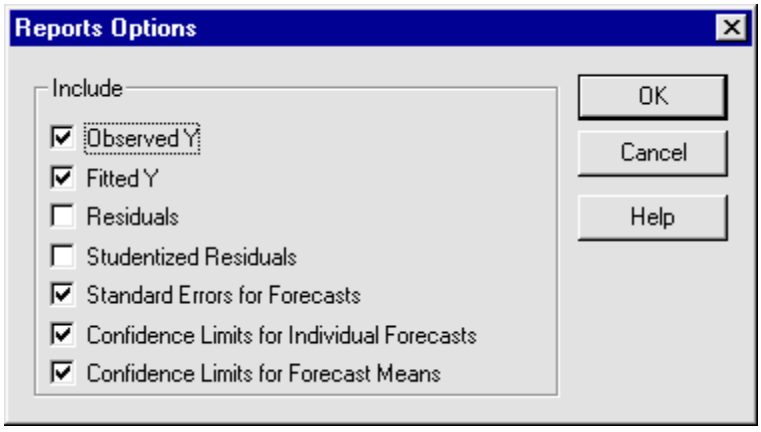
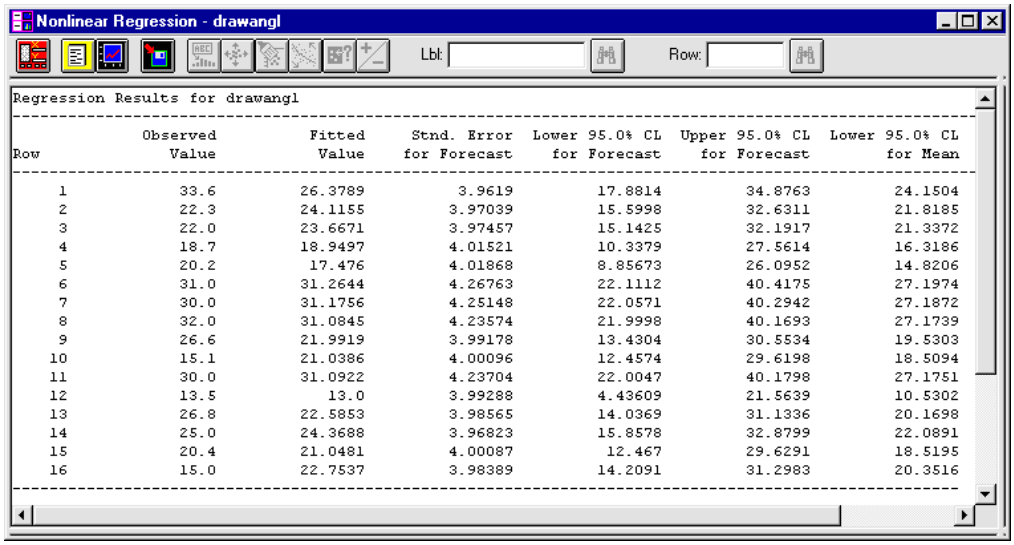


Figure 9-15. The Completed Reports Options Dialog Box

5. Click OK to redisplay the Reports Table (see Figure 9-16).



Regression Results for drawangl

Row	Observed Value	Fitted Value	Stnd. Error for Forecast	Lower 95.0% CL for Forecast	Upper 95.0% CL for Forecast	Lower 95.0% CL for Mean
1	33.6	26.3789	3.9619	17.8814	34.8763	24.1504
2	22.3	24.1155	3.97039	15.5998	32.6311	21.8185
3	22.0	23.6671	3.97457	15.1425	32.1917	21.3372
4	18.7	18.9497	4.01521	10.3379	27.5614	16.3186
5	20.2	17.476	4.01868	8.85673	26.0952	14.8206
6	31.0	31.2644	4.26763	22.1112	40.4175	27.1974
7	30.0	31.1756	4.25148	22.0571	40.2942	27.1872
8	32.0	31.0845	4.23574	21.9998	40.1693	27.1739
9	26.6	21.9919	3.99178	13.4304	30.5534	19.5303
10	15.1	21.0386	4.00096	12.4574	29.6198	18.5094
11	30.0	31.0922	4.23704	22.0047	40.1798	27.1751
12	13.5	13.0	3.99288	4.43609	21.5639	10.5302
13	26.8	22.5853	3.98565	14.0369	31.1336	20.1698
14	25.0	24.3688	3.96823	15.8578	32.8799	22.0891
15	20.4	21.0481	4.00087	12.467	29.6291	18.5195
16	15.0	22.7537	3.98389	14.2091	31.2983	20.3516

Figure 9-16. The Recalculated Reports Table

The table presents the results from the nonlinear regression equation. The conclusions show the confidence limits to be somewhat wide so you will save these results for use after you complete further research.

Saving the Results

What is noteworthy here is that you can save the Coefficients -- the parameter estimates. In addition, you can save the function as a character variable so you will be able to edit and use it in future analyses.

1. Click the Save Results button (the fourth button from the left) to display the Save Results Options dialog box.

2. Click the check boxes for the following options: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Coefficients, and Function.

   You want to save the function so you will change the name in the Target Variables text box.

3. Click the mouse pointer in the last text box (FUNCTION), and type ***MY FUNCTION*** (see Figure 9-17).

4. Click OK to save the values and the function.

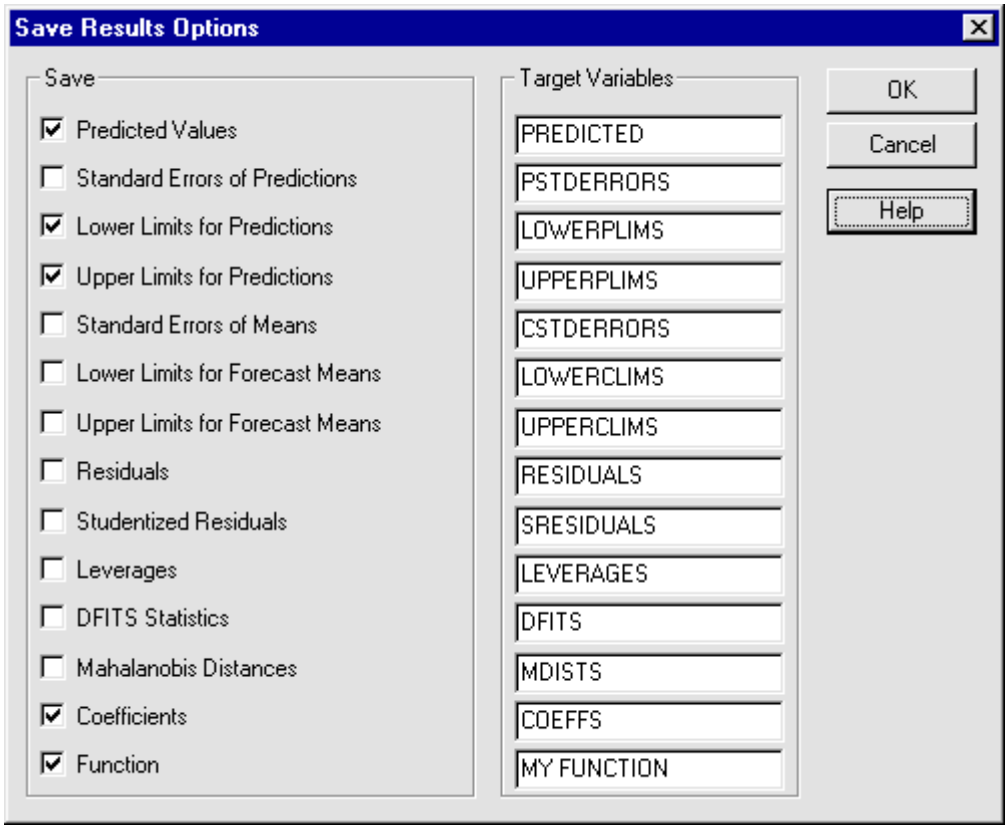


Figure 9-17. The Save Results Options Dialog Box Showing All the Results that Will Be Saved

References

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

Draper, N. R. and Smith, H. J. *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**:269-280.

Marquardt, D. W. 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *Journal for the Society of Industrial and Applied Mathematics*, **11**:431-441.

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

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

TUTORIAL 10

Studying the Relationship of Body Fat to Explanatory Variables

This tutorial is adapted from Neter et al. (1996). You will use a portion of the data that were collected to study the relationship of the amount of body fat (Y) to several possible explanatory variables, based on a sample of 20 healthy females, 25-34 years old. The variables are: **Triceps** skin-fold thickness (X_1), **Thigh** circumference (X_2), and **Midarm** circumference (X_3).

The measurements for each of the 20 persons were obtained by immersing each person in water, a cumbersome and expensive procedure. In the Neter et al. example, the researchers thought it would be more helpful to use skin-fold and tape measurements, which are easy to obtain, to provide reliable estimates.

It was also noted that there were informal indications of severe multi-collinearity in the data. When the model was fit using all three explanatory variables, the estimated regression coefficient for the **Thigh** variable was negative although it was expected that the amount of body fat was positively related to **Thigh** circumference.

You will use the Ridge Regression Analysis to try to overcome the multi-collinearity, and to evaluate the data in the body fat example.

To begin, open STATGRAPHICS *Plus* and the **Bodyfat** data file.

Finding Good Values for Parameters

1. Choose SPECIAL... ADVANCED REGRESSION... RIDGE REGRESSION... from the Menu bar to display the Analysis dialog box.

2. Enter **Bodyfat** into the Dependent Variable text box.

3. Enter **Triceps**, **Thigh**, and **Midarm** into the Independent Variables text box (see Figure 10-1).

4. Click OK to display the Analysis Summary and the Ridge Trace in the Analysis window. Maximize the Analysis Summary (see Figure 10-2).

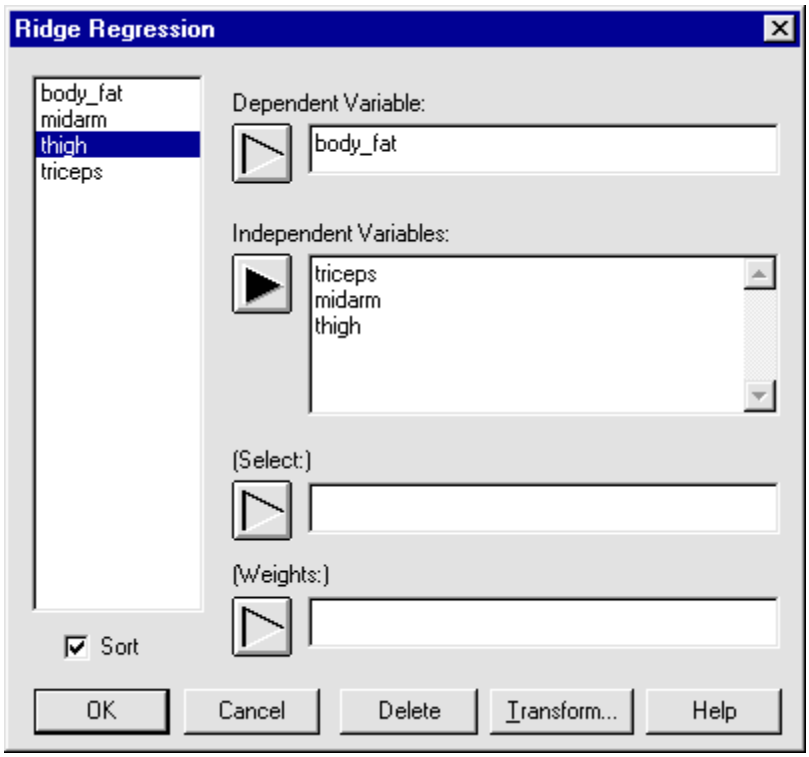


Figure 10-1. Completed Analysis Dialog Box

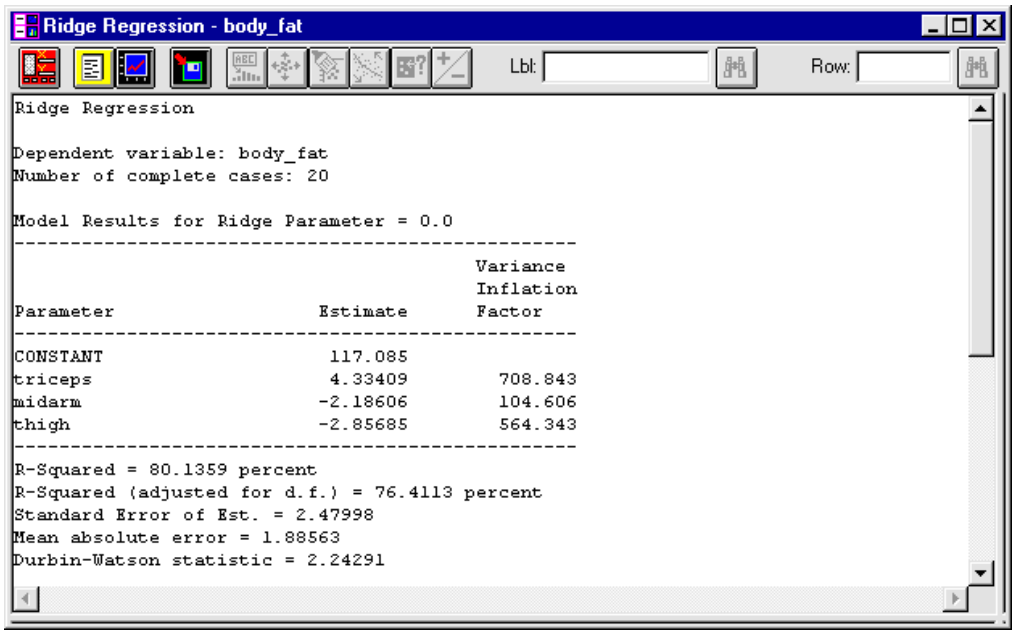


Figure 10-2. The Analysis Summary

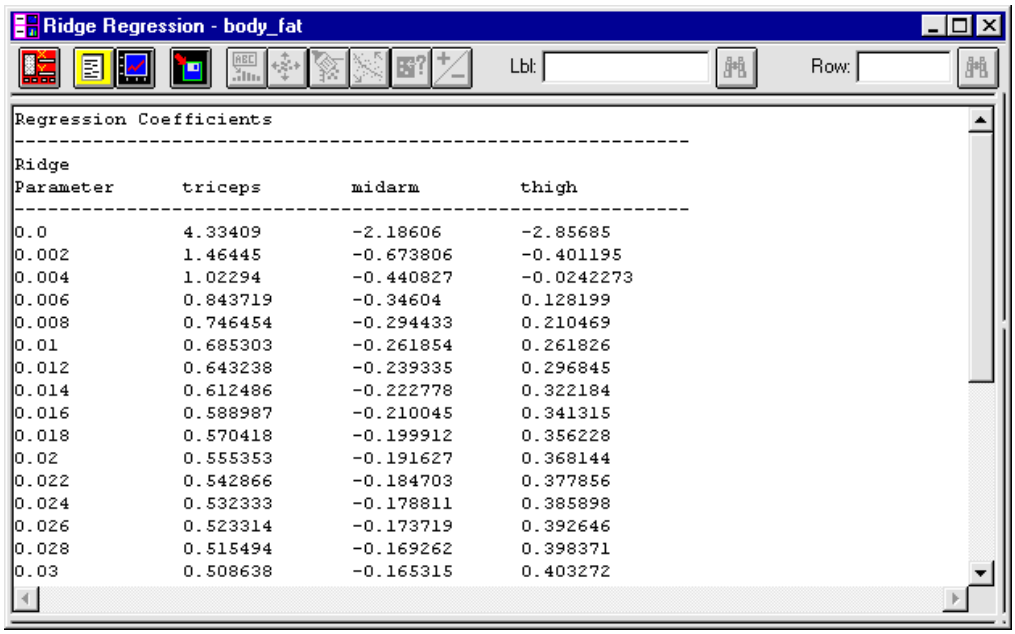
The Analysis Summary displays the natural (Unstandardized) regression coefficient estimates, which correspond to the ridge parameter, theta = 0. The large variance inflation factors are due to high correlation among the three independent variables: **Triceps**, **Thigh**, and **Midarm**. The coefficient for **Thigh** (-2.85685) is negative. The researchers felt that this was incorrect, and probably due to ill-conditioned data.

Now you will find a value for the ridge parameter that stabilizes the coefficient estimates by creating a table of regression coefficients for several values of the ridge parameter.

Creating Regression Coefficients

1. Click the Tabular Options button to display the dialog box, then click the Regression Coefficients check box, and OK to display the table in the second text pane of the Analysis window.

2. Maximize the table (see Figure 10-3).



Regression Coefficients

Ridge Parameter	triceps	midarm	thigh
0.0	4.33409	-2.18606	-2.85685
0.002	1.46445	-0.673806	-0.401195
0.004	1.02294	-0.440827	-0.0242273
0.006	0.843719	-0.34604	0.128199
0.008	0.746454	-0.294433	0.210469
0.01	0.685303	-0.261854	0.261826
0.012	0.643238	-0.239335	0.296845
0.014	0.612486	-0.222778	0.322184
0.016	0.588987	-0.210045	0.341315
0.018	0.570418	-0.199912	0.356228
0.02	0.555353	-0.191627	0.368144
0.022	0.542866	-0.184703	0.377856
0.024	0.532333	-0.178811	0.385898
0.026	0.523314	-0.173719	0.392646
0.028	0.515494	-0.169262	0.398371
0.03	0.508638	-0.165315	0.403272

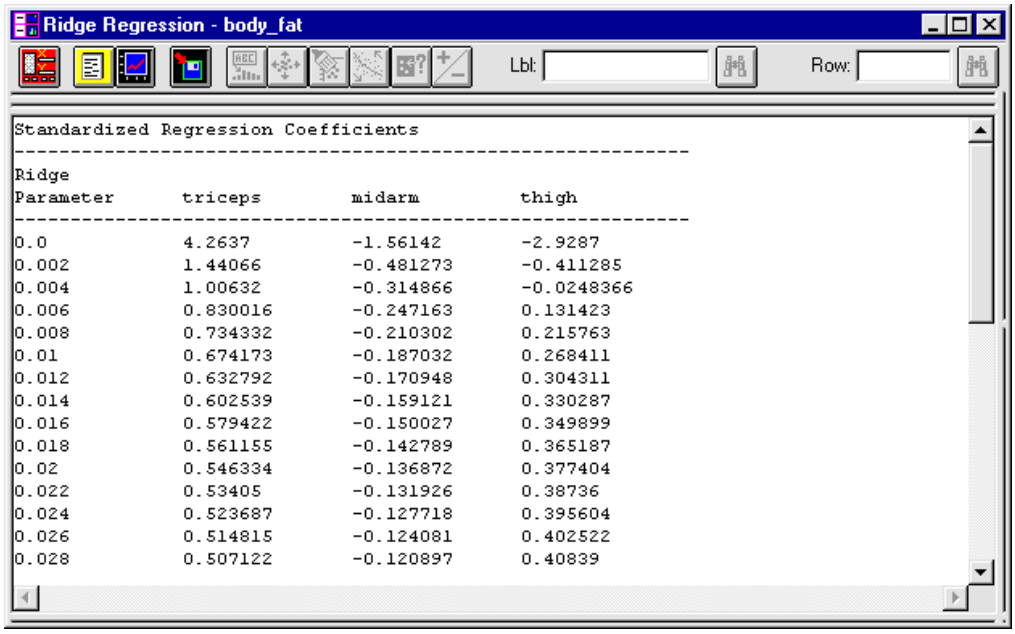
Figure 10-3. Regression Coefficients Table

The table shows the natural coefficient (Unstandardized) estimates for increments of theta from 0 to 0.1. As suspected, the coefficient for **Thigh** becomes positive even for very small values of theta (0.005).

To look for stabilization, you will create and examine the Standardized Regression Coefficients.

3. Click the Tabular Options button to display the dialog box, then click the Standardized Regression Coefficients check box, and OK to display the table in the third text pane.

4. Maximize the table (see Figure 10-4).



Standardized Regression Coefficients

Ridge Parameter	triceps	midarm	thigh
0.0	4.2637	-1.56142	-2.9287
0.002	1.44066	-0.481273	-0.411285
0.004	1.00632	-0.314866	-0.0248366
0.006	0.830016	-0.247163	0.131423
0.008	0.734332	-0.210302	0.215763
0.01	0.674173	-0.187032	0.268411
0.012	0.632792	-0.170948	0.304311
0.014	0.602539	-0.159121	0.330287
0.016	0.579422	-0.150027	0.349899
0.018	0.561155	-0.142789	0.365187
0.02	0.546334	-0.136872	0.377404
0.022	0.53405	-0.131926	0.38736
0.024	0.523687	-0.127718	0.395604
0.026	0.514815	-0.124081	0.402522
0.028	0.507122	-0.120897	0.40839

Figure 10-4. Standardized Regression Coefficients Table

As you review the ridge parameters look for the smallest value that occurs before the estimates begin to slowly change (after the standardized coefficients have begun to level off). Neter et al. recommend using theta = 0.02.

Before deciding that 0.02 is the best choice, check the values of that parameter to see if the variance inflation factors are small. Do this by creating a Variance Inflation Factors Table.

Creating a Variance Inflation Factors Table

1. Click the Tabular Options button, then the Variance Inflation Factors check box, and OK to display the Variance Inflation Factors Table in the fourth text pane.

To get a closer look at the values near 0.02, change the maximum value for the ridge parameter.

2. Maximize the table.

3. Click the right mouse button on the text pane, then the left on Analysis Options to display the Ridge Regression Options dialog box.

4. Enter ***0.04*** into the Maximum text box (see Figure 10-5).

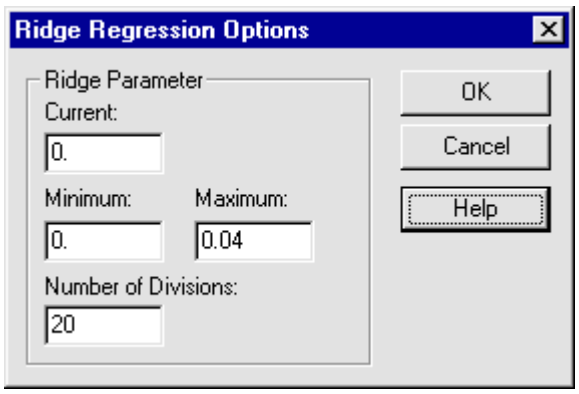


Figure 10-5. Completed Ridge Regression Options Dialog Box

5. Click OK to recalculate the parameter and redisplay the table (see Figure 10-6).

   Two items in Figure 10-6 are worth noting. At a value of 0.02 for the ridge parameter, the R-Squared value of 77.26 indicates that there is little decrease in the R-Squared statistic compared with the gain in stabilization of the coefficient estimates. Most analysts would be willing to accept this tradeoff. Also, note that now all the values for the variance inflation factors are near 1, which is desirable.

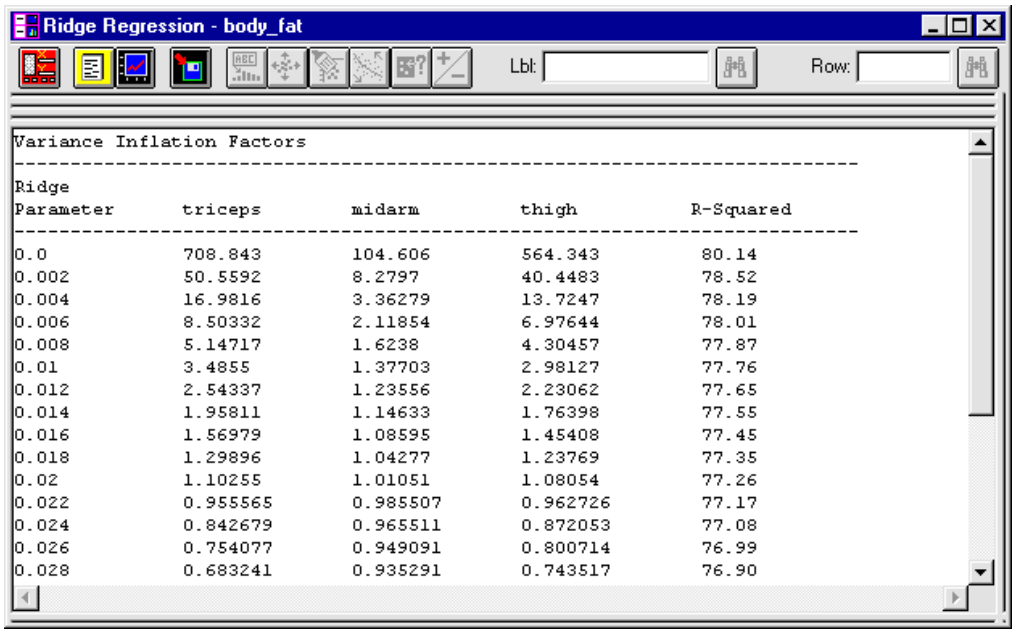
   To see these results graphically, look first at the Ridge Trace, then the Variance Inflation Factors Plot.

6. Minimize the Variance Inflation Factors Table, then maximize the Ridge Trace (see Figure 10-7).

   Figure 10-7 makes it easier to see the meaning of the stabilized coefficients -- (the lines for the three variables (**Triceps**, **Thigh**, and **Midarm**) -- that become horizontal and parallel.

   As an optional exercise, you can use the Ridge Trace Options dialog box to change the coefficients to Unstandardized, and to compare the graphical results with the results in the Regression Coefficients Table.

   Now create the Variance Inflation Factors Plot.



Variance Inflation Factors

Ridge Parameter	triceps	midarm	thigh	R-Squared
0.0	708.843	104.606	564.343	80.14
0.002	50.5592	8.2797	40.4483	78.52
0.004	16.9816	3.36279	13.7247	78.19
0.006	8.50332	2.11854	6.97644	78.01
0.008	5.14717	1.6238	4.30457	77.87
0.01	3.4855	1.37703	2.98127	77.76
0.012	2.54337	1.23556	2.23062	77.65
0.014	1.95811	1.14633	1.76398	77.55
0.016	1.56979	1.08595	1.45408	77.45
0.018	1.29896	1.04277	1.23769	77.35
0.02	1.10255	1.01051	1.08054	77.26
0.022	0.955565	0.985507	0.962726	77.17
0.024	0.842679	0.965511	0.872053	77.08
0.026	0.754077	0.949091	0.800714	76.99
0.028	0.683241	0.935291	0.743517	76.90

Figure 10-6. Variance Inflation Factors Table Redisplayed

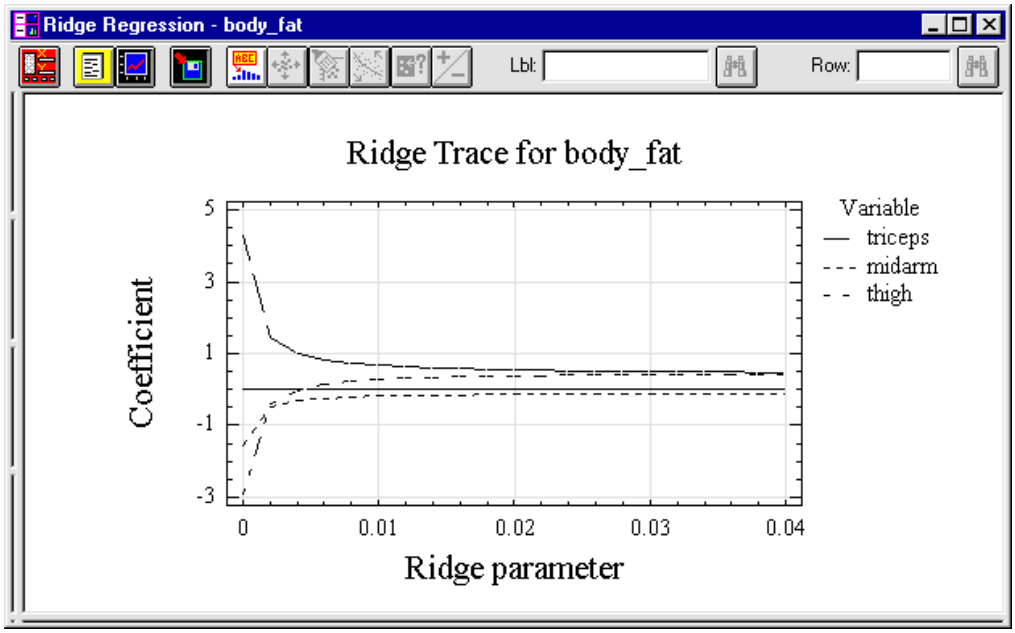


Figure 10-7. Ridge Trace

As an optional exercise, you can use the Ridge Trace Options dialog box to change the coefficients to Unstandardized, and to compare the graphical results with the results in the Regression Coefficients Table.

Now create the Variance Inflation Factors Plot.

7. Click the Graphical Options button to display the dialog box, then click the Variance Inflation Factors option, and OK to display the plot.

8. Maximize the plot (see Figure 10-8).

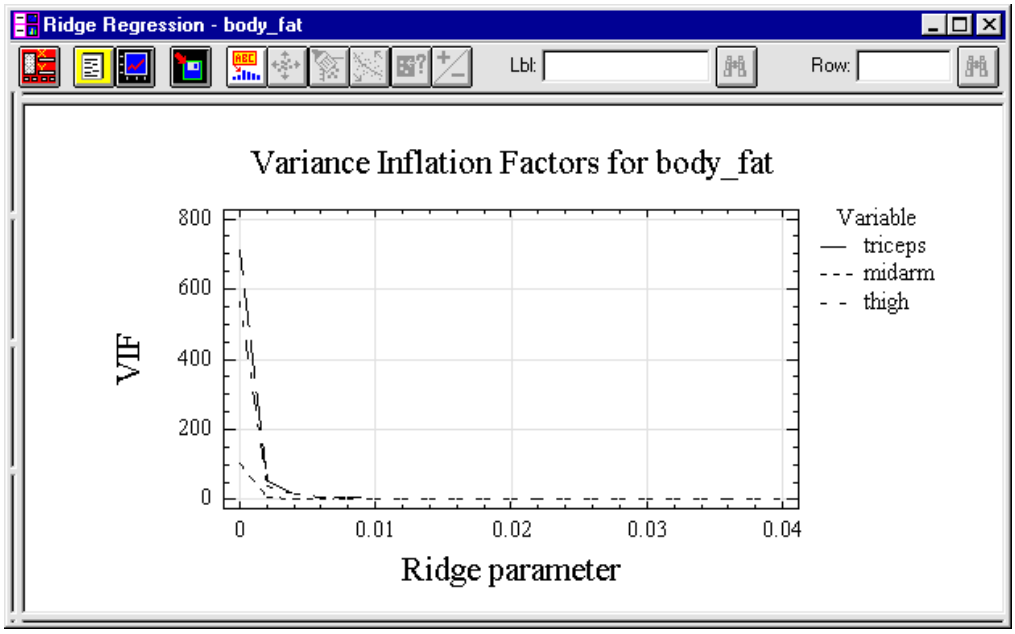


Figure 10-8. Variance Inflation Factors Plot

Notice that the Y-axis scaling ranges from 0 to 800. Because you are interested only in small VIF values, you need to change the scaling.

9. Place the mouse pointer on one of the numbers in the Y-axis and, click the left mouse button to place markers at the corners of the scale. Click the right button on Graphics Options to display the Graphics Options dialog box opened to the Y-Axis tab page.

10. Accept the default, 0.0, in the From text box, enter ***20.0*** in the To text box, and ***2*** in the By text box. Accept the defaults for the remaining check boxes (see Figure 10-9).

11. Click OK to recalculate the Y-Axis scale and redisplay the plot (see Figure 10-10).

Looking at Figure 10-10, it is evident that the variance inflation factors appear to be stabilized at theta = .02, confirming that is the value that should be used to estimate the regression coefficients.

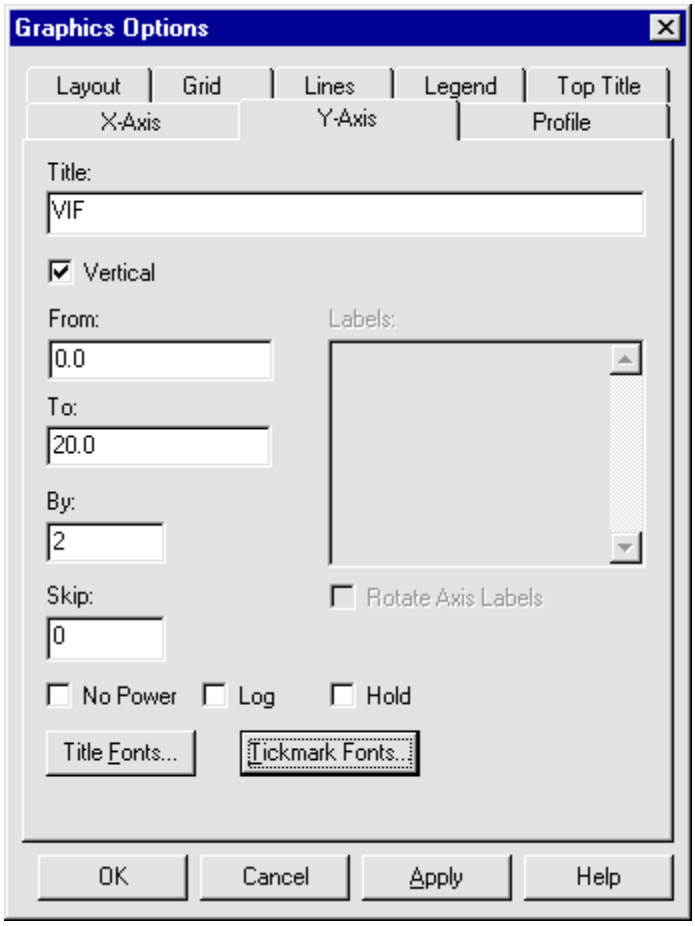


Figure 10-9. The Completed Y-Axis Tab Page

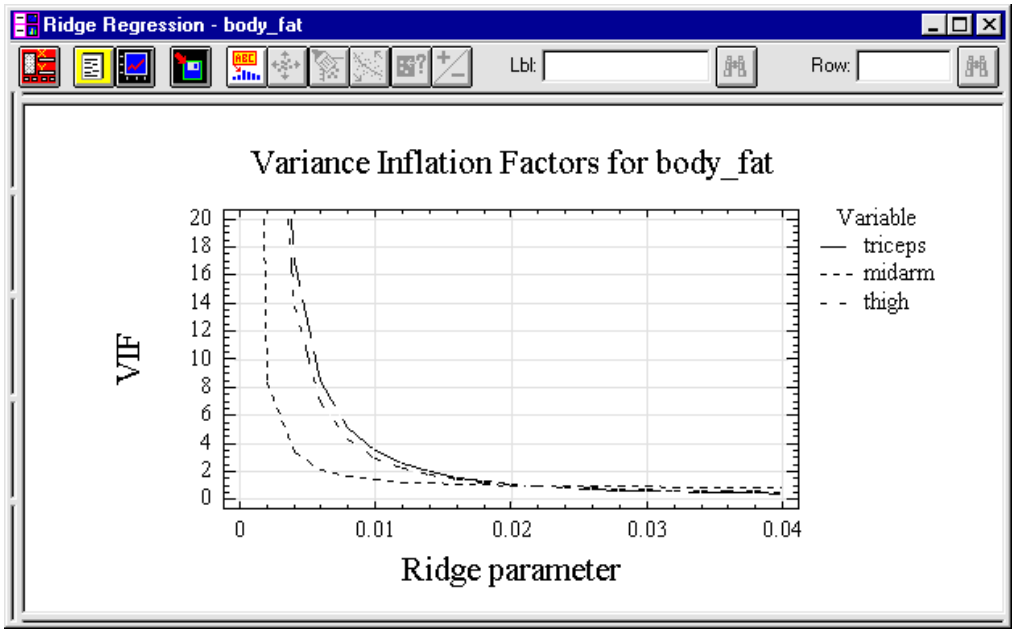


Figure 10-10. Redisplay of the Variance Inflation Factors Plot

Estimating Regression Coefficients Using New Values

1. Minimize the Variance Inflation Factors Plot, then maximize the Analysis Summary.

2. Click the right mouse button, then the left on Analysis Options to display the Ridge Regression Options dialog box.

3. Enter ***0.02*** in the Current text box, but make no other changes (see Figure 10-11).

4. Click OK to recalculate and redisplay the Analysis Summary (see Figure 10-12).

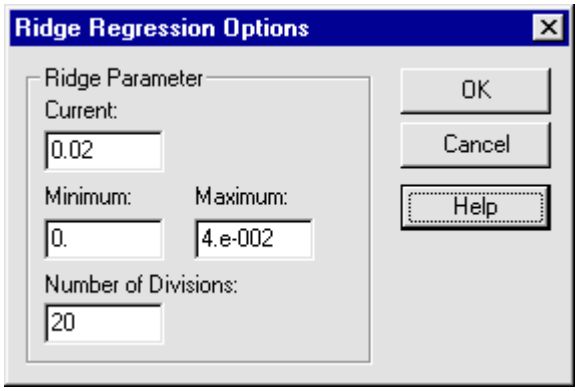


Figure 10-11. Completed Ridge Regression Options Dialog Box

Compare the values for the R-Squared, Adjusted R-Squared, Standard Error of Estimate, Mean Absolute Error, and the Durbin-Watson statistics in Figure 10-12 with those in Figure 10-2. Using the value of 0.02 for the ridge parameter gives error statistics that are slightly inferior for biased estimates; however, because the estimates for the coefficients have been stabilized, the slight changes are acceptable.

Conclusions

The tables and plots you have created all help you find a reliable value for the ridge parameter. The results indicate that, in spite of the ill-conditioned data, you could use the model to estimate skin-fold measurements to estimate body fat.

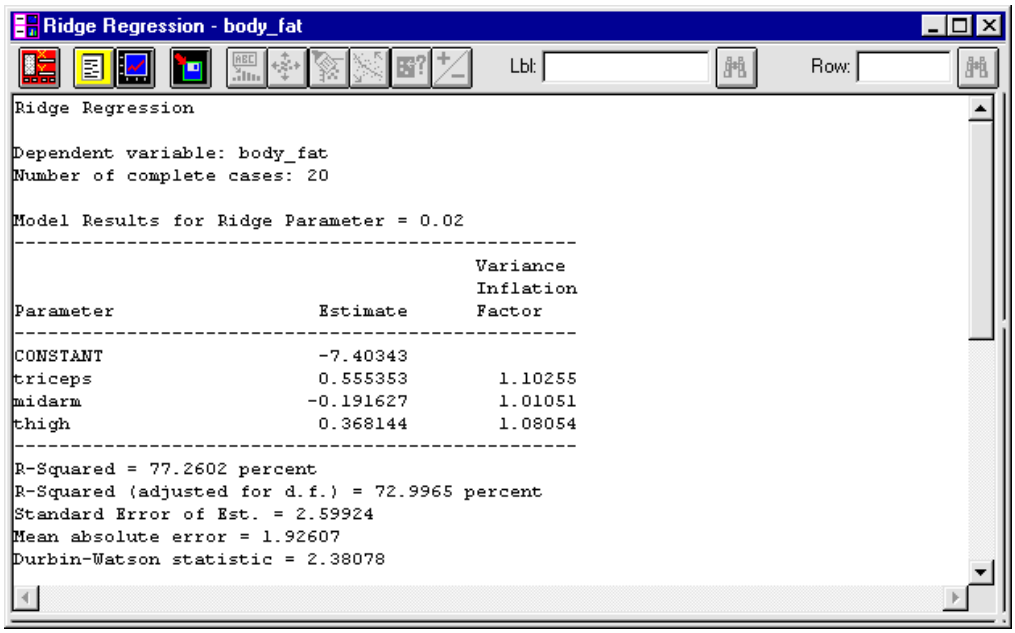


Figure 10-12. Redisplayed Analysis Summary After Using New Ridge Parameter

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.

TUTORIAL 11

Analyzing Coupon Redemption Rate with Logistic Regression

This tutorial is a study of the effectiveness of price-reduction coupons on a given product using logistic regression; it is adapted from Neter et al. (1996).

The analysts selected 1,000 homes and mailed product advertising material and coupons to each home. Two hundred selected homes were randomly assigned to each of the price-reduction categories. The coupons offered five price reductions: $5, $10, $15, $20, and $30.

The explanatory variable for this study was the X variable, the amount of the price reduction. The response variable (Y), was the proportion of coupons redeemed within six months for each group. Your goal is to quantify the relationship between the level of price reduction and the probability that a coupon would be redeemed.

The data are stored in a file, where *p* is the dependent variable, *n* is the sample size, and X is the quantitative factor (reduction). You will create the analysis by first finding the fitted response function.

To begin, open STATGRAPHICS *Plus* and the **Coupons** data file.

Fitting the Regression Model

1. Choose SPECIAL... ADVANCED REGRESSION... LOGISTIC REGRESSION... from the Menu bar to display the Analysis dialog box.

2. Enter ***p*** into the Dependent Variable text box.

3. Enter ***n*** into the Sample Sizes text box.

4. Enter **Reduction** into the Quantitative Factors text box (see Figure 11-1).

5. Click OK to display the Analysis Summary and Plot of Fitted Model in the Analysis window. Maximize the Analysis Summary (see Figure 11-2).

   In Figure 11-2, the *p*-value for the model is less than 0.01, which indicates a statistically significant relationship between the variables at the 99 percent

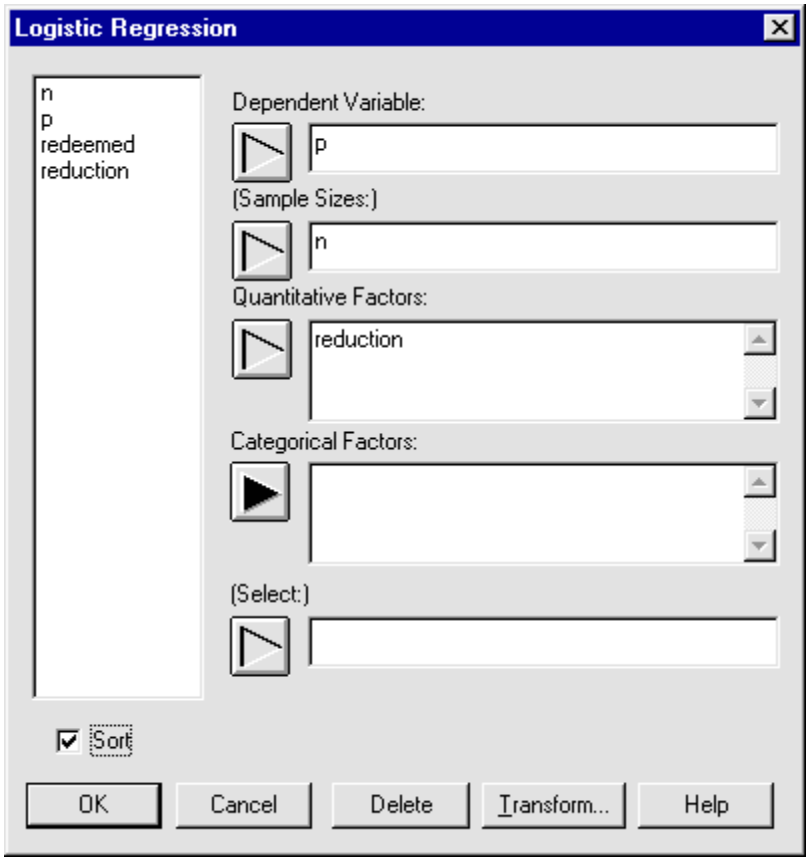


Figure 11-1. Completed Analysis Dialog Box

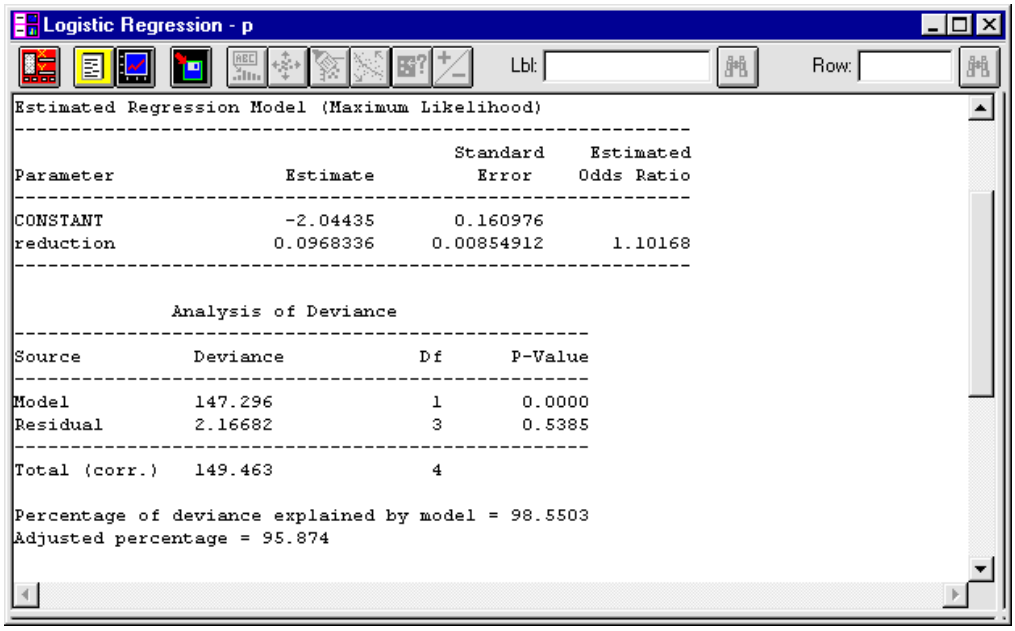


Figure 11-2. Analysis Summary Results

confidence level -- a very good fit. Additionally, the *p*-value for the residuals is greater than or equal to 0.10, which indicates that the model is not significantly worse than the best possible model at the 90 percent or higher confidence level.

The estimated odds ratio indicates that the odds of a household redeeming a coupon increase by about 10 percent with each $1.00 decrease in price allowed by the coupon (see Neter et al. 1996).

The logistic response function is:

$$\frac{e^{-2.04435 + 0968336*\text{Reduction}}}{1+e^{-2.04435 + .0968836*\text{Reduction}}}$$

where -2.04435 + .0968336*Reduction is called the *logit*. You can see the logistic curve and the logit on the Plot of Fitted Model and the Logit Plot.

6. Click the Graphical Options button to display the dialog box, then click the Logit Plot check box (the Plot of Fitted Model automatically displays), and OK to display the plots in the first and second graphics panes.

7. Maximize the Plot of Fitted Model (see Figure 11-3).

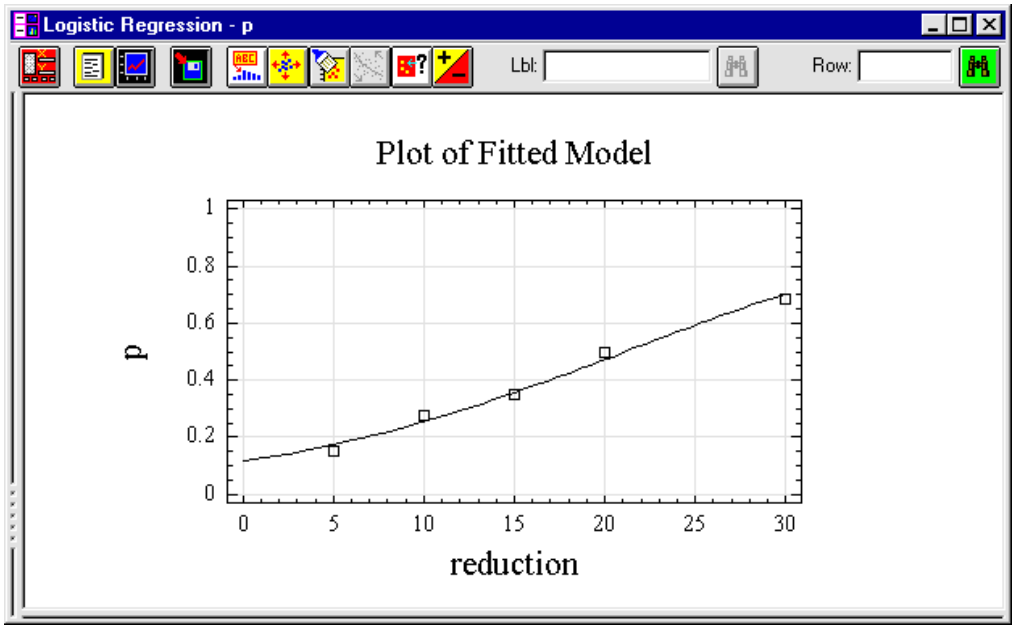


Figure 11-3. The Plot of Fitted Model

The plot shows that at a price reduction of $25, the predicted redemption rate is 60 percent. It also shows the fitted logistic regression model and the

proportions of coupons that are predicted for redemption at each of the X levels.

8. Minimize the plot, then maximize the Logit Plot (see Figure 11-4).

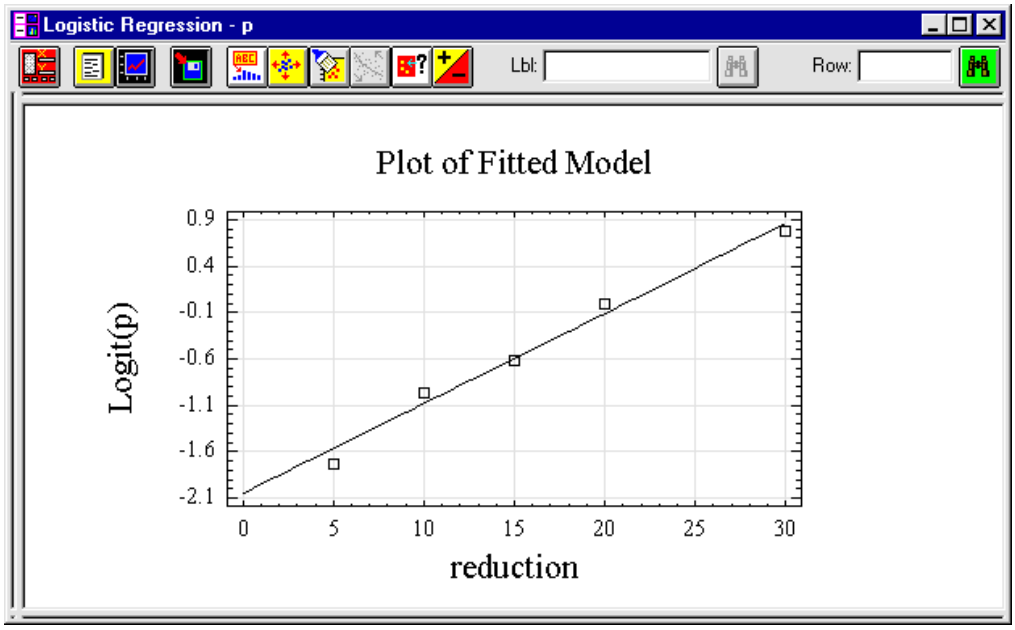


Figure 11-4. The Logit Plot

The plot shows a straight line -- the logit portion of the response function. The line is commonly used to find the median effective dose, which for this example, would be the coupon value that has a fifty/fifty chance of being redeemed.

Notice that the scaling on the plot is arranged according to the default settings. You will change the scaling so the plot will have a grid line at zero on the Y-Axis.

9. Place the mouse pointer on one of the points on the Y-Axis. Click the left mouse button to place markers at each end of the axis, then click the right button on Graphics Options to display the Y-Axis Tab page.

10. Enter ***-2*** into the From text box, ***1*** into the To text box, and accept the default in the By text box as well as the remainder of the options (see Figure 11-5).

11. Click OK to rescale the axis and redisplay the plot (see Figure 11-6).

The line is -2.04435 + .968536***Reduction.*** You can easily see the median effective value by finding the price reduction that corresponds with logit(p) =

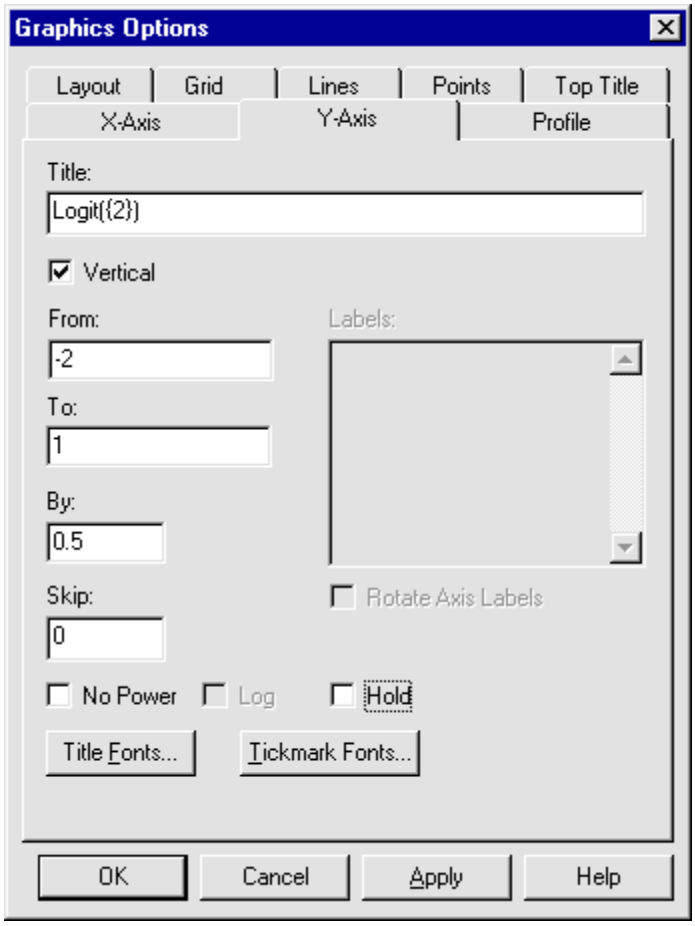


Figure 11-5. Completed Y-Axis Tab Page

0, at about $21. The price reduction that gives a probability greater than .05 (**Redemption**) is at least $21.

Now you will create the confidence intervals for the estimated coefficients.

Creating Confidence Intervals

1. Click the Tabular Options button to display the dialog box, then click the Confidence Intervals check box, and OK to display the table. Maximize the table (see Figure 11-7).

The confidence intervals for the coefficient estimates show that neither interval includes 0 so each parameter is significant at 95 percent. The confidence intervals for the odds ratios indicate that for the long term, a range of about 7 to 13 percent will cover 95 percent of the estimated odds ratios.

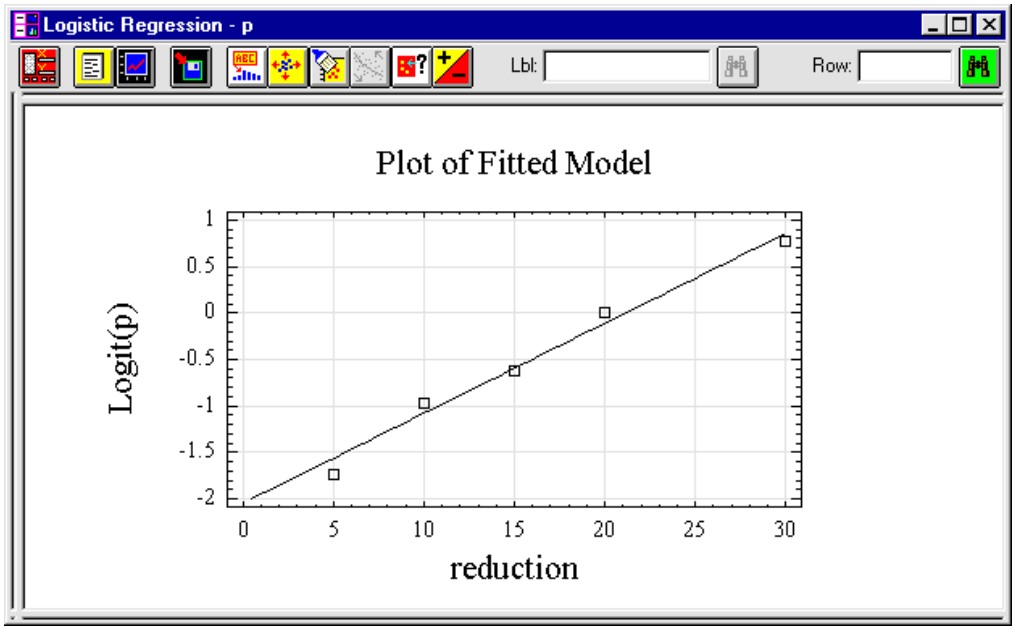


Figure 11-6. The Logit Plot with New Y-Axis Scaling

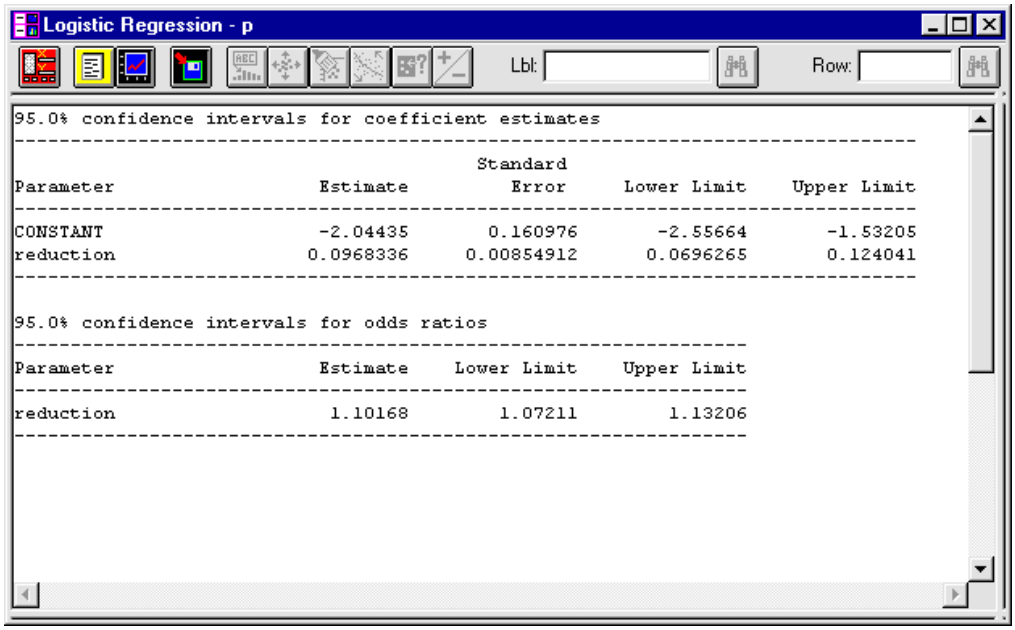


Figure 11-7. The Confidence Intervals Table

At this point, it would be helpful to look at the predictions.

Creating and Viewing Predictions

1. Click the Tabular Options button to display the dialog box, then click the Predictions check box, and OK to display the table. Maximize the table.

The program calculated the predictions using the default values, but you decide to change them to shorten the table and to eliminate the display of prediction performance results for the tails of the logistic curve.

2. Click the right mouse button, then the left on Pane Options to display the Predictions Options dialog box.

3. Enter ***0.2*** into the From text box, ***0.8*** into the To text box, accept the default in the By text box, then click the All Values option to change the values that will display in the Predictions Table. Accept the default in the Confidence Level text box (see Figure 11-8).

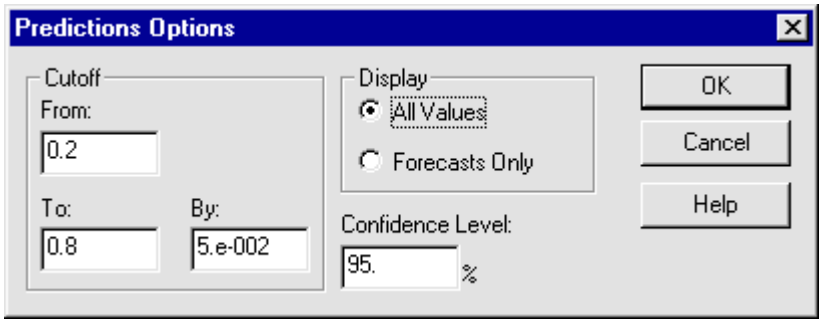


Figure 11-8. Completed Predictions Options Dialog Box

4. Click OK to recalculate and redisplay the Predictions Table (see Figure 11-9).

The Predictions Performance Table shows how well the function performs at various cut-off values when you use it to predict True or False (Success/Failure) for the estimation data. If a prediction of success is made whenever the fitted value is greater than 0.5, 68.2 percent of the sample data are correctly classified.

To graphically view the model s capability of correctly predicting success or failure, you will examine two more plots: the Prediction Capability Plot and Predictions Histograms.

5. Click the Graphical Options button to display the dialog box, then the Prediction Capability Plot and Prediction Histograms check boxes, and OK to display the two plots in the third and fourth graphics panes.

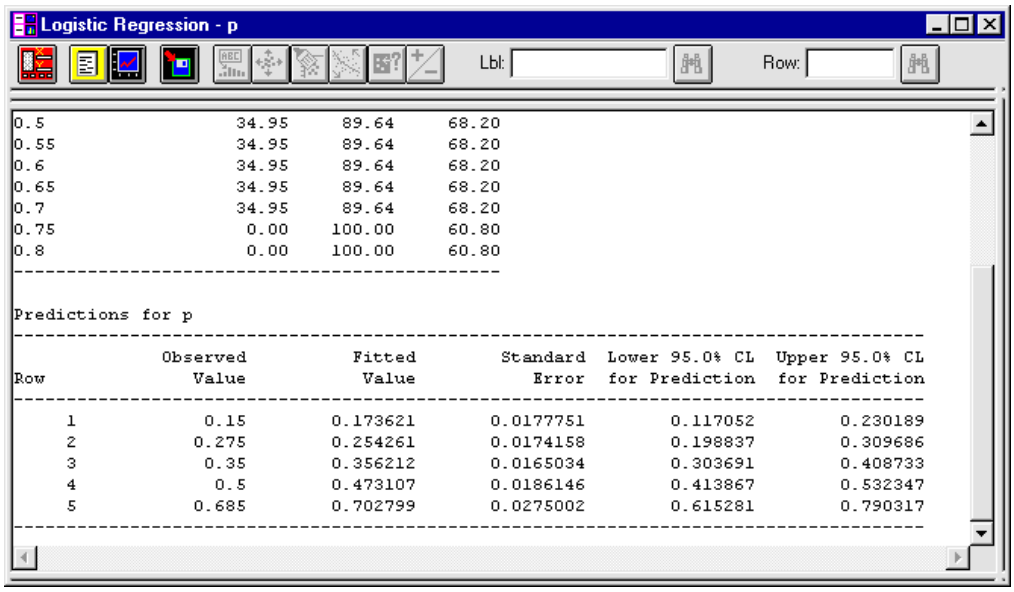


Figure 11-9. The Recalculated Predictions Performance Table

6. Maximize the Prediction Capability Plot (the third graphics pane) (see Figure 11-10).

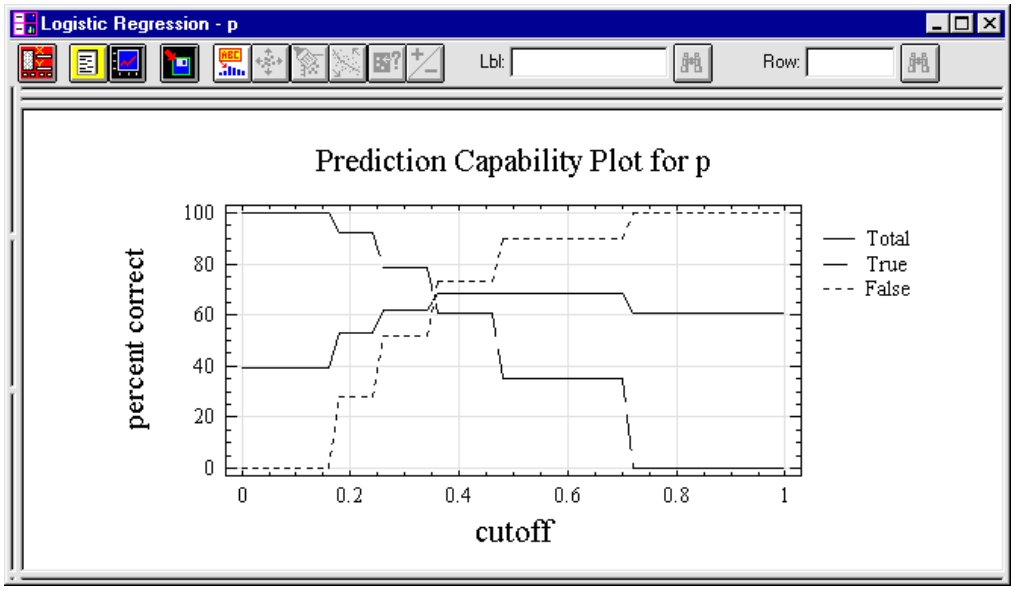


Figure 11-10. The Prediction Capability Plot

The plot shows a summary of the prediction capability of the fitted logistic model. The model first predicts 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, the response is predicted to be False.

The plot shows the percent of observed data that were correctly predicted at each cutoff value. For example, using a cutoff equal to 0.36, 60.4592 percent of all the True responses were correctly predicted, while 73.1908 percent of all the False responses were correctly predicted for a total of 68.2 percent. This cutoff value may be a good value to use to make additional predictions.

Now look at the Prediction Histograms.

7. Maximize the Prediction Histograms (see Figure 11-11).

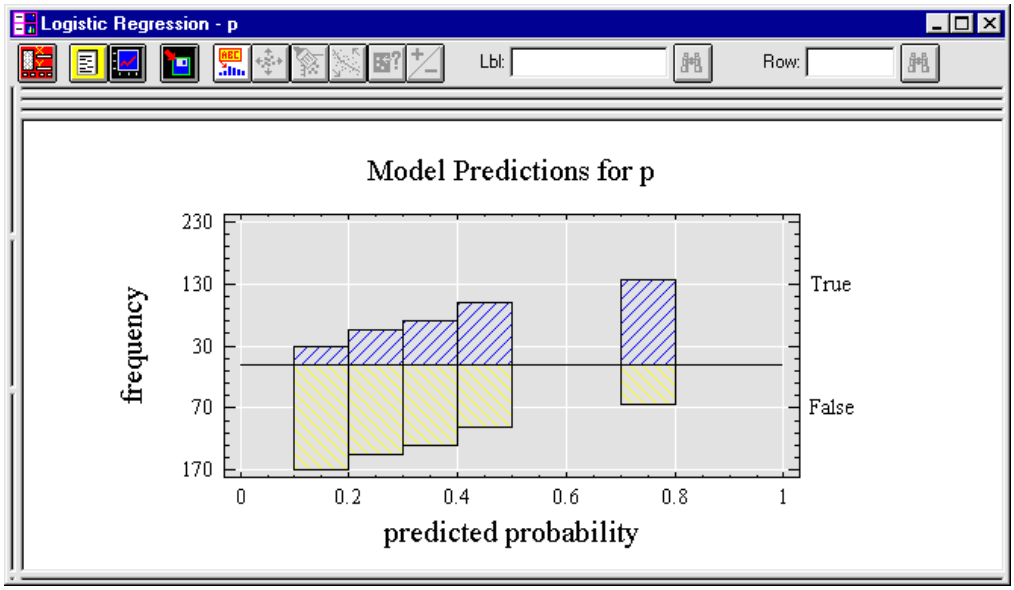


Figure 11-11. The Prediction Histograms Plot

The plot shows the ability of the fitted logistic model to distinguish between cases when the outcome is True or False, and shows the frequency distribution of the True and False cases versus the probability predicted by the fitted model. Ideally, the model predicts a small probability for the False cases and a large probability for the True cases. Notice that the large frequencies above the line are plotted on the far right and large frequencies below the line are plotted on the left, which indicates that the model works reasonably well.

Additional Exercise

1. As an additional exercise, access the Logistic Regression Options dialog box, change the method to Weighted Least Squares, and compare the Analysis Summary to the one you created using Maximum Likelihood.

References

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 & Hall.

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.