

INTRODUCTION TO RELEASE 6.08 SAS/INSIGHT™ AND SAS/LAB™ SOFTWARE

Mark Brown, SAS Institute Inc.
Marshall P. Scott, SAS Institute Inc.

ABSTRACT

Recent trends in data analysis techniques are now available to SAS software users with SAS/LAB and SAS/INSIGHT software, two key components of a data analysis strategy that are part of the SAS System for Information Delivery. SAS/LAB software is a data analysis advisor for users without an extensive statistical background. As an analysis proceeds, SAS/LAB software offers explanations of the results and suggestions for further analysis. SAS/INSIGHT software is a graphical data analysis tool that offers interactive statistical graphics combined with traditional statistical techniques. This paper provides an introduction to SAS/LAB and SAS/INSIGHT software.

INTRODUCTION

SAS/INSIGHT software provides the statistician with an extensive repertoire of methods for the detailed study of a set of data. The emphasis is on flexible probing of the data, often before comparing the data to a probabilistic model. A cycle of alternating uses of the exploratory and confirmatory techniques found within SAS/INSIGHT software, either on successive smaller bodies of data or on a single substantial one, is not uncommon and is often desirable. SAS/INSIGHT software enables users in technical, scientific, and commercial fields to understand the implications of their data quickly and to verify their impressions with the facts. In effect, SAS/INSIGHT software speeds up the discovery process.

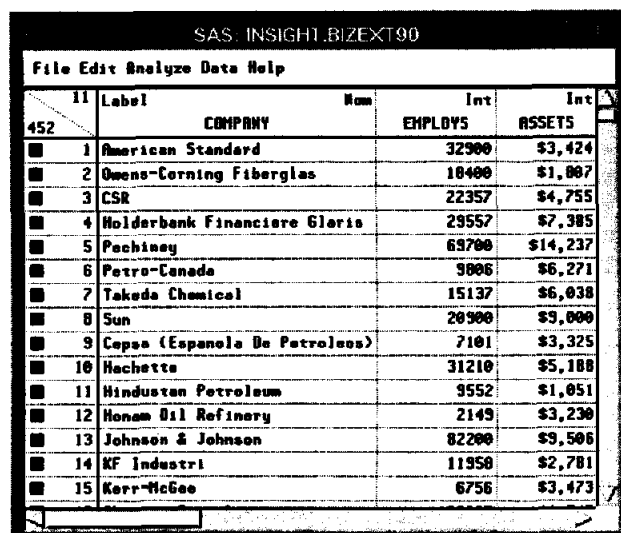
SAS/LAB software provides a tool for a broad range of users with different degrees of statistical expertise. For the statistical novice, SAS/LAB software serves as a data analysis advisor by guiding the user through the data analysis process with prompting, warnings about assumptions, and interpretations. For the professional statistical consultant, SAS/LAB software represents a solution for routine analyses that can be used by the general user community, freeing the consultant for more complex tasks. The SAS/LAB software interface is designed to be easy to use. To that extent, the software contains an extensive online help facility as well as a journaling facility that enables you to save, edit, and manage text and graphics.

SAS/INSIGHT SOFTWARE: GRAPHICAL DATA ANALYSIS

SAS/INSIGHT software is designed to offer users an environment for Exploratory Data Analysis (EDA). EDA grew from the work of J.W. Tukey who argued that the data analyst should proceed much in the same manner as a detective trying to solve a mystery. Instead of investigating hypotheses formulated independently of the data, the data analyst uses an open-ended exploratory approach with few preconceptions. Analyses are interactive and information gathered at each stage is used to refine subsequent analyses. This approach requires a project like SAS/INSIGHT software, one that supports interactive and dynamic graphical methods.

Graphical Exploration Techniques

SAS/INSIGHT software has three types of windows: a data window that displays a SAS data set in tabular format; a graph window that displays graphs; and an analysis window that displays graphs and statistical tables. An example of a data window is shown in Figure 1. This SAS data set contains information about large international companies. In the data window you can move, sort, subset, extract, and set properties for the data. In this example, the variable COMPANY is assigned the property of a label variable so the name of the company associated with each observation will be displayed when that observation is selected in any of the other windows.



The screenshot shows the SAS/INSIGHT interface with a menu bar (File, Edit, Analyze, Data, Help) and a data table. The table has columns for observation number, label, company name, number of employees, and assets. The data is sorted by company name.

	Label	Num	Int	Int
452		COMPANY	EMPLOYEES	ASSETS
1	American Standard		32900	\$3,424
2	Owens-Corning Fiberglas		18400	\$1,007
3	CSR		22357	\$4,755
4	Holderbank Financiere Glaris		29557	\$7,385
5	Pechiney		63700	\$14,237
6	Petro-Canada		9006	\$6,271
7	Takeda Chemical		15137	\$6,038
8	Sun		20900	\$3,000
9	Cepsa (Esparola De Petroleos)		7101	\$3,325
10	Hachette		31210	\$5,188
11	Hindustan Petroleum		9552	\$1,051
12	Honam Oil Refinery		2143	\$3,230
13	Johnson & Johnson		82200	\$9,506
14	KF Industri		11950	\$2,781
15	Kerr-McGee		6756	\$3,473

Figure 1 Data Window

SAS/INSIGHT software offers techniques for graphically exploring data with bar charts, box plots, scatter plots, and rotating plots. No complex command syntax is needed to produce graphical displays. Simply select the variables of interest and choose the type of graph from the graph menu. A useful graphical tool for comparing distributions is a box plot. Release 6.08 SAS/INSIGHT software supports interactive box plots with selection of all observations and identification of extreme values. Figure 2 shows how a box plot is used to explore the profits of various types of corporations. Extreme observations can be selected and hidden to reduce possible skewness. The box plot rescales automatically.

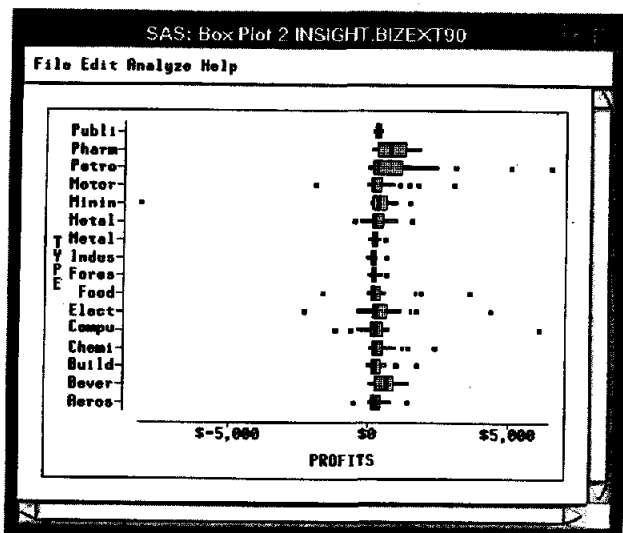


Figure 2 Comparing Distributions with a Box Plot

Another typical data exploration technique is to plot all variables pairwise. This graphical technique, often called a *scatter plot matrix*, is a representation of the relationship between two quantitative variables. You can label observation markers in scatter plots, brush the point cloud, and customize the markers and axes. A scatter plot matrix is shown in Figure 3. In this scatter plot, the log values of the three variables are used to spread the data values and compensate for valid extreme observations. A rectangle is drawn around some observations in the plot of `L_EMPLOY` and `L_SALES`. This creates a *brush* for exploring observations dynamically. A *brush* in one plot highlights the same observations in all plots.

You can add information to the the scatter plot matrix to reveal patterns in the data. SAS/INSIGHT software enables you to assign a unique color-marker combination to your data. Highlighting, colors, markers, and observation values are updated automatically in SAS/INSIGHT displays. Thus, any change to the data shows immediately in all graphs and analyses.

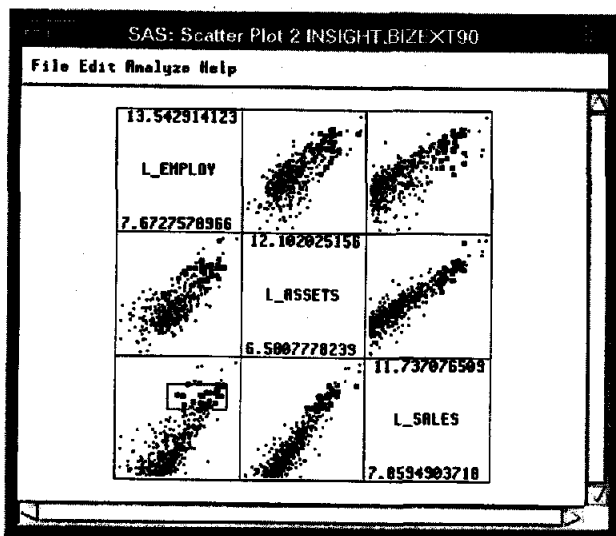


Figure 3 Exploring Observations with a Scatter Plot Matrix

Sometimes a relationship that does not show in two dimensions will become more evident in three with a *rotating plot*. Figure 4 shows a three-dimensional plot of the same variables shown in the scatter plot matrix. The box drawn around the data points is added to make the effects of the rotation clearer. You can rotate the plot by clicking, dragging, pressing, or throwing. In this example, the data is roughly ellipsoid, and you can identify a few extreme observations. In addition to adding a box to a rotating plot, you can use rays to draw a line from each observation marker to the origin of the plot. These lines help show the shape of the point cloud.

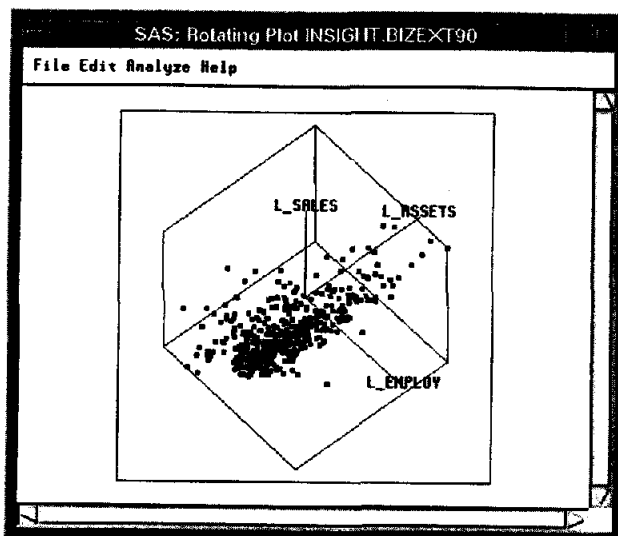


Figure 4 Identifying Observations with a Rotating Plot

Release 6.08 SAS/INSIGHT software provides the equivalent of BY-group processing. Data for each graph and analysis can be processed by groups, using any number of group variables. The scatter plot shown in Figure 5 shows `L_SALES` assigned the Y role, `L_EMPLOY` the X role, and `TYPE` the group role. There is one scatter plot for each type. To explore relationships dependent on `TYPE`, scroll through the grouped scatter plots and examine the different intercepts for different types. The scatter plot in Figure 5 shows that for `TYPE=Petroleum Refining`, there is a near linear relationship.

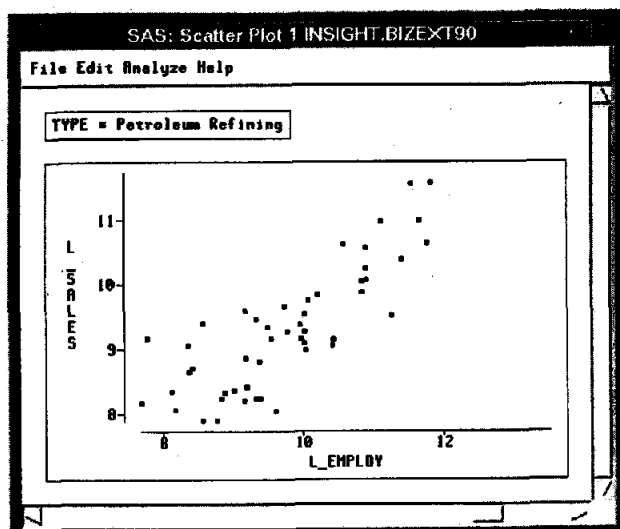


Figure 5 Using BY-Groups In a Scatter Plot

Data Analysis Techniques

The graphical exploration techniques in SAS/INSIGHT software often lead to the use of the data analysis techniques. You can examine distributions and correlations and calculate principal components to find the structure of your data. In addition, you can construct predictive models based on relationships you find in the data. The fit analyses in SAS/INSIGHT software provide methods for examining the relationship between a response (dependent) variable and a set of explanatory (independent) variables. The methods available include simple and multiple linear regression with various diagnostic capabilities. Release 6.08 SAS/INSIGHT software adds support for models with nominal effects, including the general linear model and the generalized linear model.

To quantify the relationship between sales and the number of employees, you can use SAS/INSIGHT software to fit a simple regression model, as shown in Figure 6.

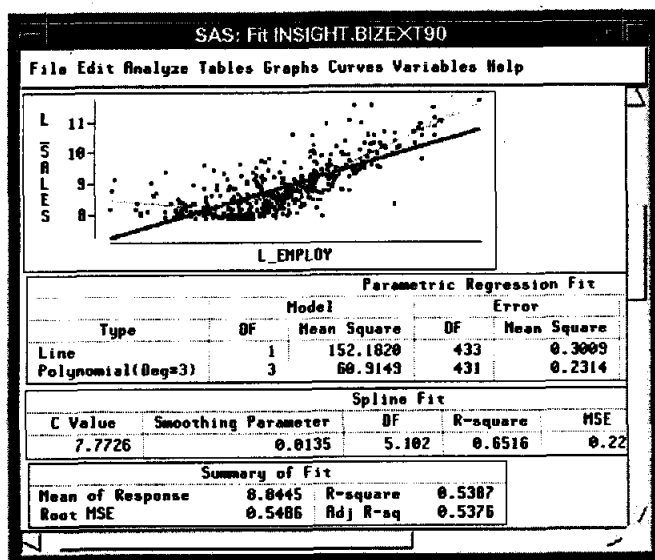


Figure 6 Fitting Simple Regression Models with SAS/INSIGHT Software

To fit a model, simply select `L_SALES` and `L_EMPLOY`, choose the menu item **Analyze**, then choose the item **Fit**. The result is a regression fit with a scatter plot. By using the pop-up menus, you can add lines and curves to the scatter plot as well as tabular output. In this example, a polynomial curve and a spline curve are easily added with the **Curves** pop-up menu. The fitted regression line is also highlighted. Additional features include mean and predicted confidence ellipses, mean and predictive confidence curves, and kernel capabilities.

The curves fit the model fairly well, but taking more variables into consideration might improve the model. For example, to use a nominal variable such as `TYPE` in a model, choose **Analyze**, then **Fit**. Assign `L_SALES` the Y role, `L_EMPLOY` the X role, and `TYPE` the X role also. Add the crossed effect `TYPE*L_EMPLOY` to explore interaction between these variables, as shown in Figure 7.

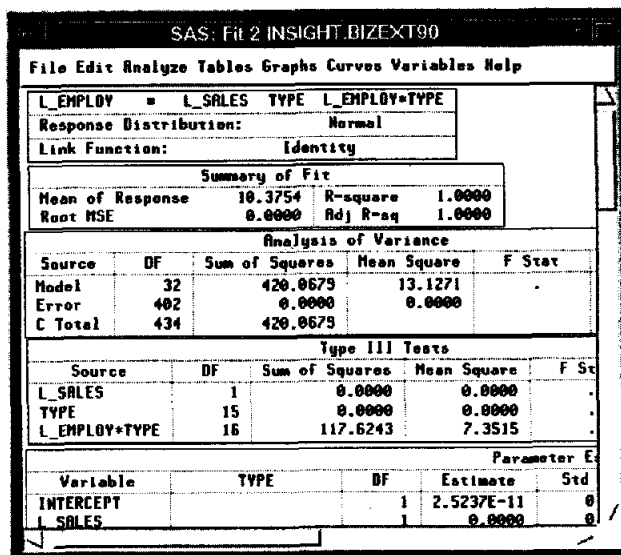


Figure 7 Adding Cross Effects for General Linear Models

All of the models supported in the GLM procedure are available in Release 6.08 SAS/INSIGHT software. Statistical researchers have worked to make linear modeling techniques apply to a wider class of problems than those supported by the general linear model. This more general class of models is usually known as the *generalized linear model*. The generalized linear model accepts nonnormal error distributions from the exponential family and uses link functions to relate the expected response value to a linear form of the explanatory variables. In Release 6.08 SAS/INSIGHT software you can create models using Normal, Inverse Gaussian, Gamma, Poisson, and Binomial error distributions as well as with Canonical, Identity, Log, Logit, Probit, Complementary log-log, and Power link functions.

When models are created, output variables such as residuals, predicted values, and normal scores are created. These output variables are available for immediate use. Figure 8 shows a distribution analysis that contains a histogram and tables of moments and quantiles. For the model produced earlier, you should check to see if the residuals are normally distributed.

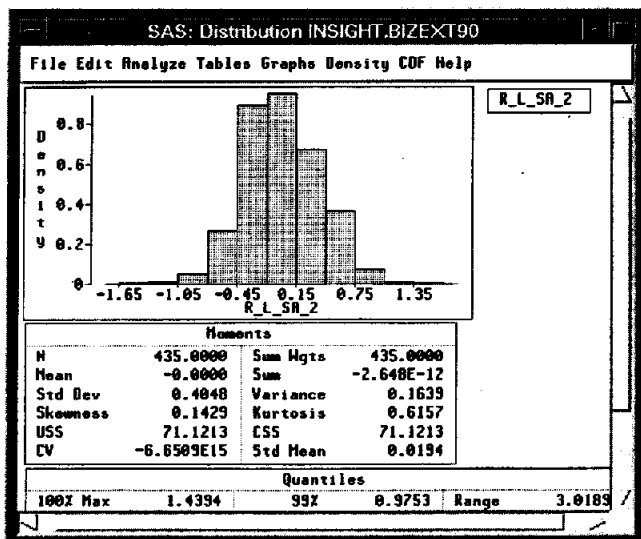


Figure 8 Distribution Analysis of Residuals

Via the pop-up menus, you can add additional information, such as normal curves or Kernel Density Estimates, to the histogram. In addition, you can produce QQ plots and use the Kolmogorov's D statistic to test the normality of the distribution. The initial display of the histogram is static, showing only one possible bin width and offset. To explore the histogram interactively, use the hand tool to manipulate the bar width and position to examine the distribution more efficiently.

Saving Results

Results can be saved in three ways with SAS/INSIGHT software: as SAS data sets, as SAS/GRAPH output, and as ASCII text files. All the output variables that are created during your session are added to the data window automatically and can be saved to SAS data sets.

SAS/LAB SOFTWARE: GUIDED DATA ANALYSIS

SAS/LAB software is designed to meet the needs of two primary audiences: users and statistical consulting professionals. SAS/LAB software is a menu-driven, task-oriented product designed to guide the analysis of data and the interpretation of results. You can conduct analyses with minimal statistical training.

SAS/LAB takes a three-step approach to data analysis:

1. The first step is entering data.
2. The second step is summarizing data.
3. The third step is analyzing data.

SAS/LAB software users are greeted by the Welcome window the first time the software is invoked. The Welcome window is designed to be user-friendly with an extensive help facility. Sample data sets are created that are used in SAS/LAB software documentation.

Beginning an Analysis

When a data set is selected for analysis, SAS/LAB software displays information about the data. Figure 9 shows the data set SASUSER.DENSITY selected for use. This data set contains mea-

surements from a study of the stiffness and density of wooden particle boards. To simulate a real study, assume that you want to predict the value of the variable STIFFNESS at a density of 25 lbs/cubic foot. The number of observations and the number of unique values for each variable is displayed. You can examine the distribution of the variables in the data set by clicking on the variable name to produce a histogram showing the distribution graphically with a superimposed curve. A *box and whisker plot* is also displayed to show the quartiles, the mean, and any outliers. In addition, you can change the graphical display to a normal *probability plot*. Besides displaying univariate plots, you can also compute descriptive statistics that provide various types of summary information.

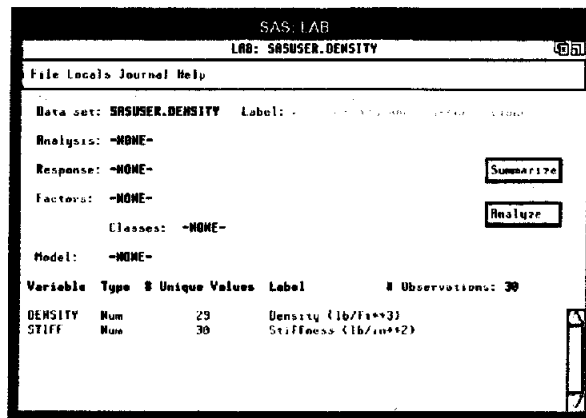


Figure 9 Computing Descriptive Statistics to Obtain Data Summaries

SAS/LAB software also provides a complete online journaling facility that is available on a window-by-window basis. You can save either text or graphics. Later, you can edit the contents of the journal to add comments or enhance graphs. Journal files, individual windows, or graphs can be printed immediately.

Data Analysis Techniques

SAS/LAB software offers four ways to analyze data:

- one-way analysis of variance
- simple linear regression
- multiple regression and analysis of variance
- analysis of covariance.

When the **Analyze** button is selected in the main window, a list of the analyses is displayed, as shown in Figure 10. Perhaps the easiest way to analyze the data is to select **Not Sure**. When this choice is selected, SAS/LAB software prompts you for more information until the appropriate analysis technique is determined. In this example, the software prompts you to choose the response variable (STIFFNESS), and then it assigns the factors variable automatically (DENSITY). The software is designed to guide you through the analysis process. You are presented only with those choices that are appropriate for your data.

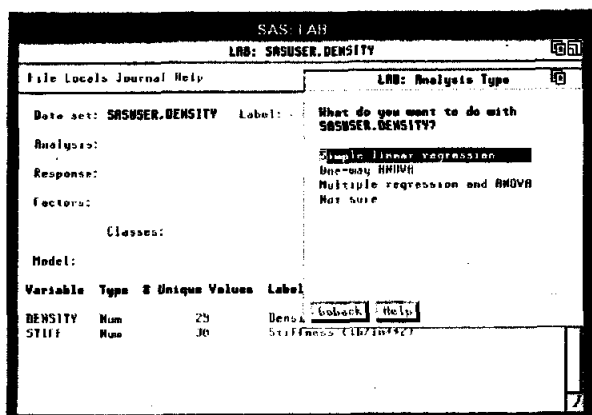


Figure 10 Choosing an Analysis Type

The analysis is performed and SAS/LAB software checks several assumptions underlying the validity of the analysis. SAS/LAB software warns you if these assumptions are violated. Some assumptions are specific for certain analysis types, and some checks are made for all analysis types. The analysis results are displayed with a list of the assumptions that are violated. An interpretation window summarizes the analysis. If you decide to explore the assumptions, a list is displayed. You can investigate each in turn, and SAS/LAB software offers you an explanation of the violation and a possible remedy in the interpretation window. In this example, several assumptions are violated during the analysis and a log transformation of the variables is recommended to produce a better model. When the log transformation is applied, all of the other violations are corrected. Choosing a remedy as suggested by the software causes the analysis to be recalculated and the assumptions to be rechecked.

When you are satisfied with the state of the assumptions, you can view the results of the analysis. As with the assumptions, the results are dependent on the type of analysis performed. Based on the data in this example, a simple linear regression is the analysis type chosen by SAS/LAB software. The overall model fit is shown in Figure 11.

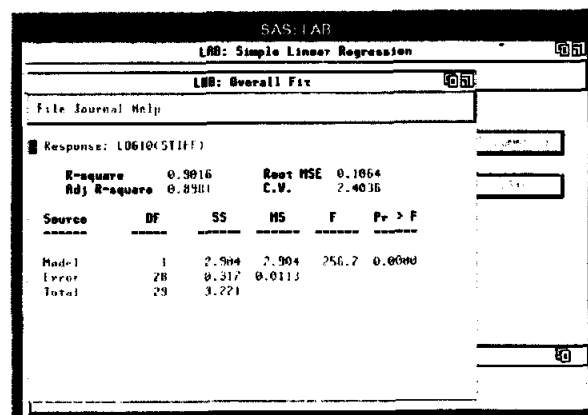


Figure 11 Viewing the Overall Model Fit

The Overall Fit window displays the significance of each effect on the model. The Parameter Estimates window offers additional information such as the parameter estimates, standard errors, and *t* statistics for each term in the model.

A facility for analyzing group means or calculating predicted response values is available. Figure 12 shows how, if a value is entered for the response variable (25 in this example), SAS/LAB software calculates the predicted value. You can display a predicted value of the response variable for any value of the factor. Or you can specify a value for the response variable for any value of the factor.

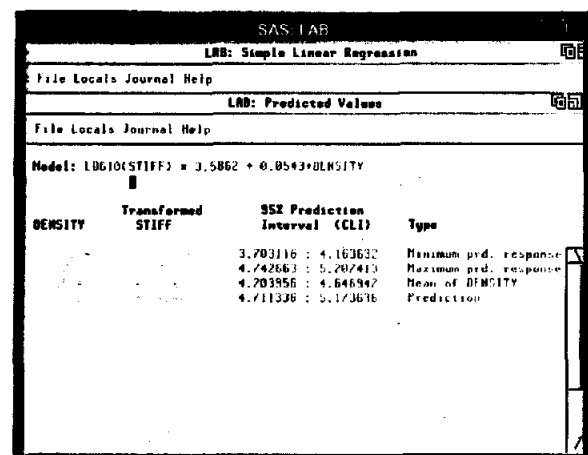


Figure 12 Calculating Predicted Values

Help

Online help is available for all windows in SAS/LAB software and can be saved to the journal. In addition, technical information is available through the help facility that explains statistical concepts and terms. For example, if you choose Help from the pop-up window that lists the type of analyses available for your data, you can get

a detailed explanation of the conditions for which each type of analysis is appropriate.

CONCLUSION

SAS/INSIGHT software and SAS/LAB software represent new capabilities that are part of an overall information delivery system provided by the SAS System. SAS/INSIGHT software offers users a unique interactive graphical and exploratory view of their data. By speeding up the discovery process, this software provides an envi-

ronment for examining implications of data and interactively constructing predictive models. Similarly, SAS/LAB software also provides an environment for examining data. However, unlike SAS/INSIGHT software, SAS/LAB software offers users a guided approach to data analysis using reliable statistical methods. SAS/LAB software helps to automate the data analysis process by prompting you for necessary information, offering suggestions in an interpretation window for further analyses, and warning you of potential problems.