Exploratory Data Analysis Using SAS® Software

Greg Weier, SAS Institute Inc., Cary, NC

Randall D. Tobias, SAS Institute Inc., Cary, NC

Presented at SEUGI '93 by
Gerhard Held

ABSTRACT

This tutorial describes two environments within the SAS® System for carrying out exploratory data analysis. SAS/LAB® software provides a guided, interactive environment for users who lack an extensive statistical background. Analysis results are presented graphically with interpretations, warnings about assumptions that may be violated, and suggestions for further analyses. SAS/LAB software focuses on basic methods commonly requested by researchers and engineers. SAS/INSIGHT® software provides a highly interactive graphical environment in which windows are a means of viewing and manipulating data objects. Scatter plot smoothers and numerous parametric curves with associated diagnostic plots can be displayed. Tools, palettes, and pull-down menus enable the user to further explore objects. Windows are dynamically linked so that actions on observations in one window are reflected in all windows.

INTRODUCTION

For many years the SAS System has provided a programming language interface to very powerful facilities for statistical analysis. However, these tools require you to know the basics of the programming language, as well as possess a command of statistical techniques. Moreover, interactive statistical and graphical analysis is cumbersome with these tools.

SAS/LAB and SAS/INSIGHT software were developed to simplify exploratory data analysis. Both are interactive systems for statistical analysis, both provide facilities for data summarization and modeling, and both rely heavily on high-resolution graphics. However, the two systems are intended for different audiences and thus have different interfaces.

SAS/LAB software provides a complete data analysis system for the basic statistical applications most often encountered by engineers and scientists. Incorporating tools for everything from data access to final presentation, the software guides you through an analysis, checking assumptions and making suggestions at each step. SAS/LAB software is primarily intended for researchers whose expertise is not in statistics; however, its ease of use and assumption checking will also appeal to statisticians.

SAS/INSIGHT software is a powerful tool for exploring the structure of data. The versatility of the software allows you to begin data exploration with basic tools and proceed step by step through more complex analyses. The guidance for exploration comes only from the data and your interests. With both versatile analysis options and sophisticated, interconnected visualization tools, SAS/INSIGHT software gives experienced data analysts the ability to discover relationships in the data.

The purpose of this tutorial is to demonstrate how to use these two systems for the kinds of data analysis tasks for which they are each most appropriate. The first example for each environment reviews the capabilities of the software. The remaining examples take a closer look at specific statistical techniques.

SAS/LAB SOFTWARE

Simple Linear Regression for Studying Wood Block Fabrication

Suppose you are a researcher for a company that manufactures pressed wood products, and you are interested in how the density of wood blocks affects their stiffness.

You can invoke SAS/LAB software by entering **lab** on a command line or selecting it from the **Globals** pmenu (under **Invoke Application → Data Analysis**). Then select **New ...** under the **File** pmenu to choose a data set; alternatively, you can invoke SAS/LAB software from the command line with the **data=** option to load the dataset automatically.

The main screen for SAS/LAB, shown in Figure 1, displays summary information about the data set that includes the variables DENSITY and STIFF, as well as options for performing the analysis.

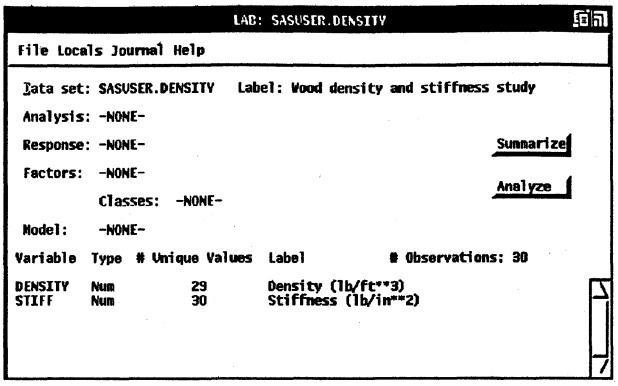


Figure 1 Primary SAS/Lab Menu

You can see a quick picture of the data by selecting the Summarize button; choosing both variables brings up a scatter-plot.

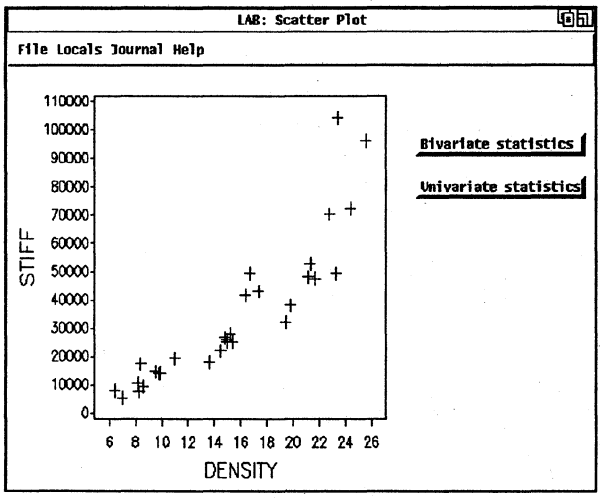


Figure 2 Scatter Plot for DENSITY

There is a strong positive correlation between the variables. To better understand the nature of the relationship, return to the main screen (select **End** under **File**) and select the Analyze button. Choose "Simple linear regression" as the type of analysis and then STIFF as the dependent variable; as the on-line help explains, the dependent variable is the one you are modeling.

A SAS/LAB analysis consists not only of displaying the results of statistical calculations but also checking the assumptions on which those calculations are based.

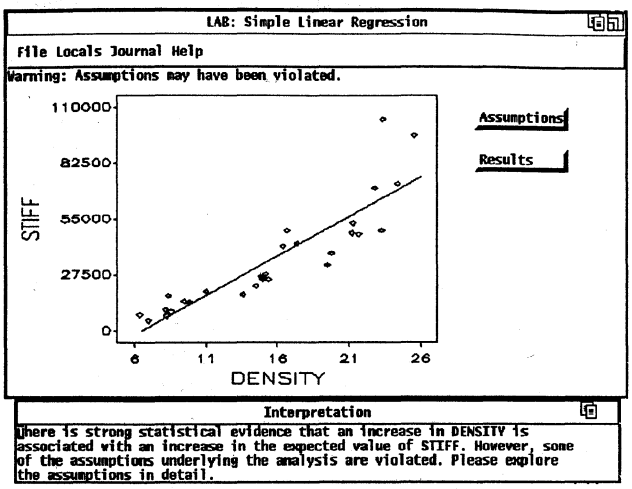


Figure 3 Analysis for DENSITY

The analysis screens for SAS/LAB typically have three parts: a graphical display of the analysis, a natural-language interpretation of the results, and tools for additional analysis. In this case, the graphical display is a scatterplot with the regression line superimposed. The interpretation notes that there is a statistically significant relationship between the variables and also that some of the assumptions underlying the analysis may be violated. When you press the Assumptions button, you see that there are several potential problems that are marked with an asterisk:

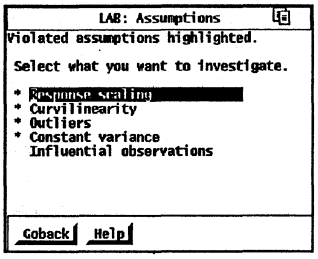


Figure 4 Assumptions for DENSITY

The highlighted "Response scaling" line indicates that it may be more appropriate to model a transformation of stiffness; "Curvilinearity" indicates that the simple straight line fit may not be sufficient; "Outliers" suggests that there are observations that do not follow the general trend of the rest of the data; and "Constant variance" indicates that the assumption that all the observations have the same level of random noise may not be tenable.

However, it turns out that you can address all of these problems in one step. Selecting the first assumption, "Response scaling," you see the suggestion to analyze the logarithm of stiffness rather than the original stiffness values.

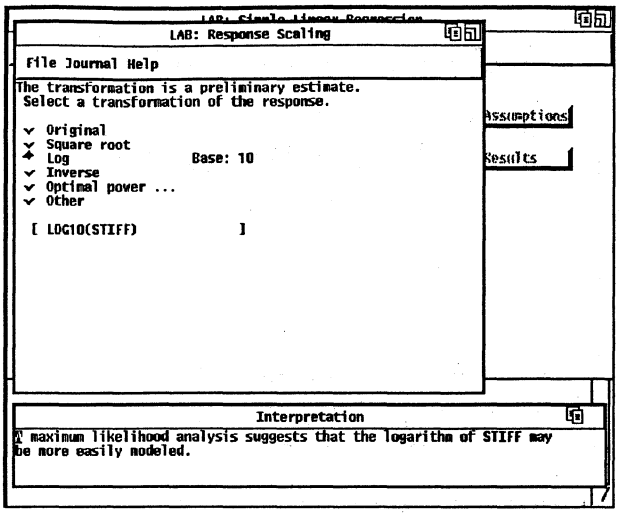


Figure 5 Response Scaling for DENSITY

The need for a transformation is common in engineering applications; refer to Box and Bisgaard (1987). When you apply it here (select **Apply** under **File**), you find that all of the other assumptions are also now satisfied.

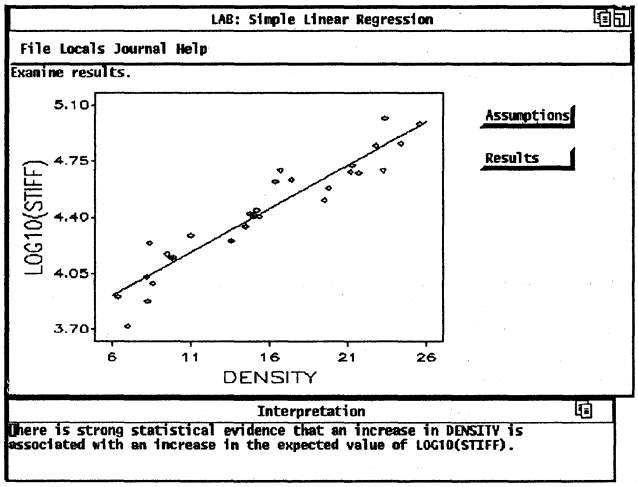


Figure 6 Analysis for DENSITY, with transformation

So what is the precise form of the estimated relationship between wood block density and stiffness? Select "Parameter estimates" under the Results button to display the following screen:

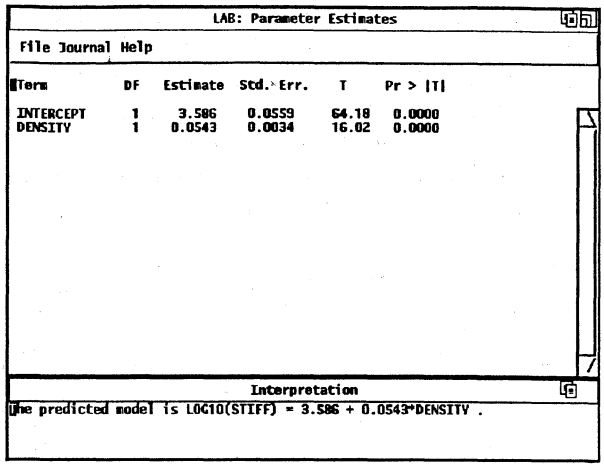


Figure 7 Parameter Estimates for DENSITY

For predicting particular values, select **End** under **File** and then select "Predicted values." Various default predictions are presented, but you can easily add more. For example, suppose you want to know what density will result in a predicted stiffness of 10,000 lbs/ft^2. Select **Change response units → Original** under **Locals** to work in the original units, then enter 10,000 in the STIFF column; the solution of the predictive equation appears in the DENSITY column.

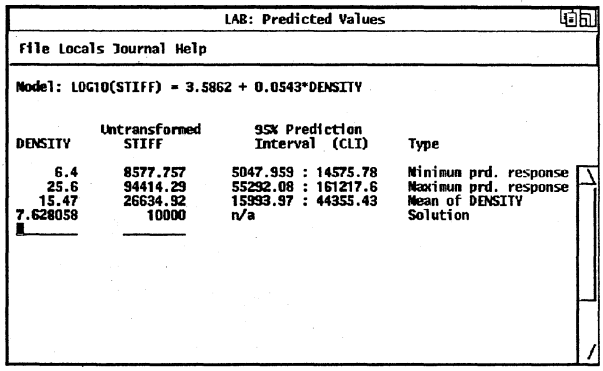


Figure 8 Predictions for DENSITY

Guided Analysis of Covariance

SAS/LAB software facilitates not only the primary task of finding and fitting a good model for your data but also secondary tasks like analysis of means. In this example, a significant difference is detected between the responses at different levels of a classification effect, and you want to explore these differences in more detail.

Hochberg and Tamhane (1987) present the results of an experiment that studied the ascorbic acid content of different varieties of lima beans. Past experience showed that older plants have less acid, so the proportion of dry matter in the harvested plants was also recorded as a measure of plant maturity. Five

observations were made for each of six varieties. Hochberg and Tamhane (1987) analyze the data using analysis of covariance on the raw acid measurements and find that variety 5, with the highest observed acid content and the lowest percentage of dry matter, also has the highest expected acid content when dry matter is taken into account.

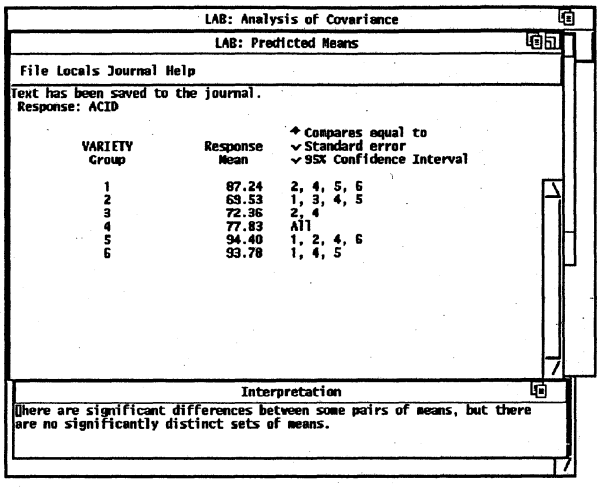


Figure 9 LSMEANS for ACID

However, while an analysis of covariance using SAS/LAB software reveals that the variety and percentage of dry matter significantly affect the expected acid content (Figure 9), as expected, there are once again problems with the underlying assumptions (Figure 10).

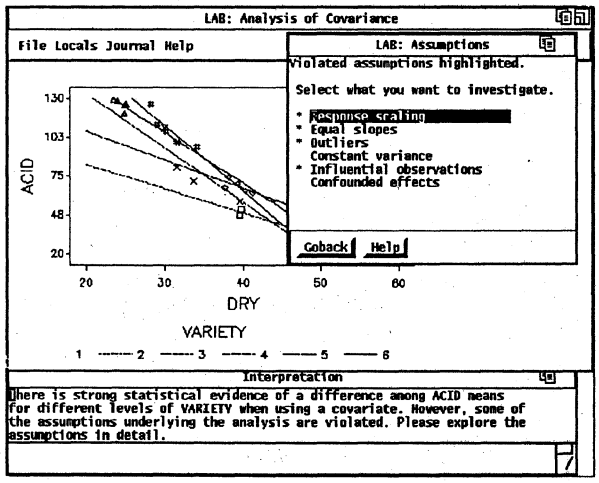


Figure 10 Assumptions for ACID

In this case, it is not clear how to solve problems with violated assumptions. If you take care of the response scaling first and apply the suggested log transformation to the measurements of acid content, there is still a violation of the assumption of constant variance.

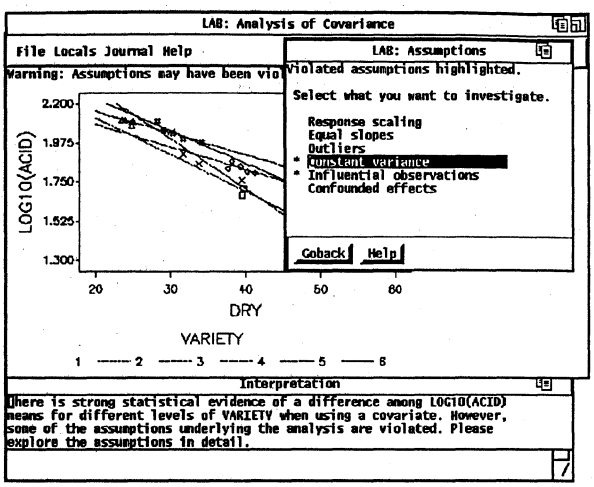


Figure 11 Assumptions for LOG10(ACID)

On the other hand, you might look at the outlier problem first. Observation 11, with an ACID value of 109 for variety 3, stands out.

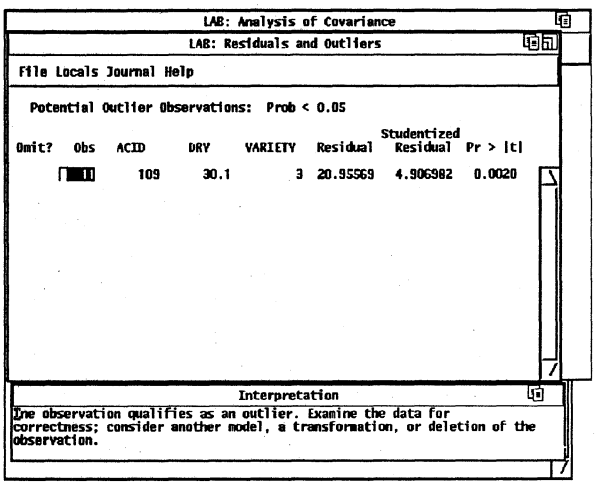


Figure 12 Outlier for ACID

If this observation is omitted, a transformation still seems required, but this time the necessary transformation is the square root of the acid content. You need to decide whether to accept observation 9 as valid and to analyze LOG10(ACID), or to reject the observation and to analyze SQRT(ACID). In both cases, the predicted means are dramatically different from the original analysis; for example, the means for the analysis of log(ACID) are shown in Figure 13.

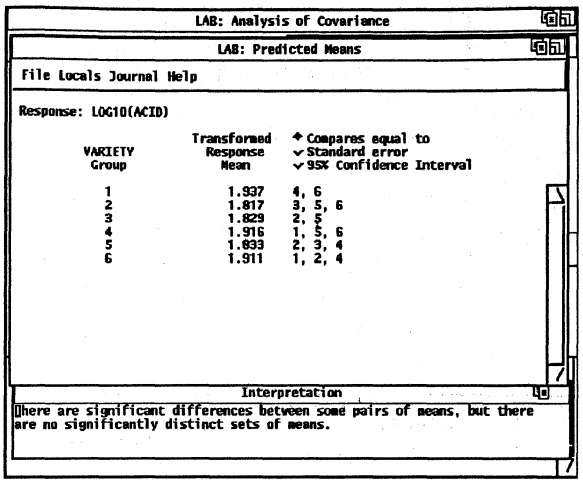


Figure 13 LSMEANS for LOG10(ACID)

Note in particular that variety 5 no longer has the highest predicted acid content. Evidently, analyzing ACID on the wrong scale masked the true effect of plant maturity as measured by the percentage of dry matter.

For analyses like this one, involving complicated modeling, the journaling facility in SAS/LAB software is a convenient way to keep track of the course of analysis. You can save the contents of any screen to the journal; both textual and graphical information can be edited or printed. For example, Figure 14 shows part of the journal with results from the original model fit.

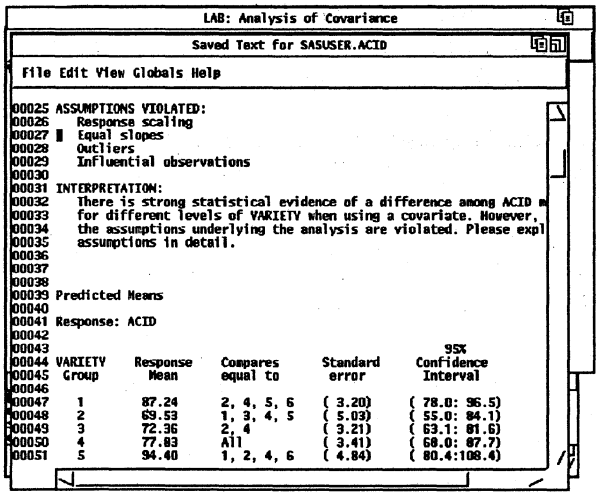


Figure 14 Journal Contents

SAS/INSIGHT SOFTWARE

SAS/INSIGHT can be invoked from a command line, from a pmenu, or by submitting a **PROC INSIGHT** statement with an optional **DATA=** specification. If no data set is specified, a data dialog appears from which you may select a SAS data set. The data set is displayed in a spreadsheet format when selected. The data window and each analysis window contain appropriate menus. You can select an object such as an observation or variable and then select a menu item to act on the object. You can also select a menu item first, which causes a dialog to appear from which you can select an observation or variable depending on the menu item selected.

Model Building with SAS/INSIGHT Software

The data for the following example is from Weisberg(1985) and was collected from the *American Almanac for 1974* and the *1974 World Almanac*. Of interest in recent political debate has been the proposal to increase the taxes on gasoline for automobiles. As a researcher you are interested in the effect of various factors on fuel consumption. Figure 15 shows a portion of the SAS/INSIGHT data window. The first column of the data window shows the observation number and the observation marker. At the top of this column are the number of variables(9) and the number of observations(50). Each column contains a variable. The header of each column contains the variable name and type. The STATE variable also is designated as an observation label variable. You can set the state name variable as a label variable by selecting **Data → Set Properties** in the data window menu and then clicking on the **Label** box under DEFAULT ROLE in the dialog that appears. Variable labels also appear in each header. You can display the variable labels when you select **Data → Labels**.

File Edit Analyze Data Help

9 / 50	Label Nom STATE state name	Int FUEL fuel consumption in gallons per person	Int TAX 1972 fuel tax rate in cent/gallon
1	MAINE	541	9.00
2	NEW HAMPSHIRE	524	9.00
3	VERMONT	561	9.00
4	MASSACHUSETTS	414	7.50
5	RHODE ISLAND	410	8.00
6	CONNECTICUT	457	10.00
7	NEW YORK	344	8.00
8	NEW JERSEY	467	8.00
9	PENNSYLVANIA	464	8.00
10	OHIO	498	7.00

Figure 15 Data Window

The first concern is whether the fuel consumption rates for all 50 states are similar. You can inspect the distribution of the FUEL variable by selecting FUEL and then selecting **Analyze → DISTRIBUTION (Y)**. The histogram from the distribution window appears in Figure 16. It appears that more states had fuel

consumption rates between 540 and 660 gallons per person than any other range. The plot also shows a normal kernel density estimate. You add this density estimate to the graph by selecting the distribution menu item **Density** and then selecting **Kernel : Normal** and a smoothing parameter of **1.0**.

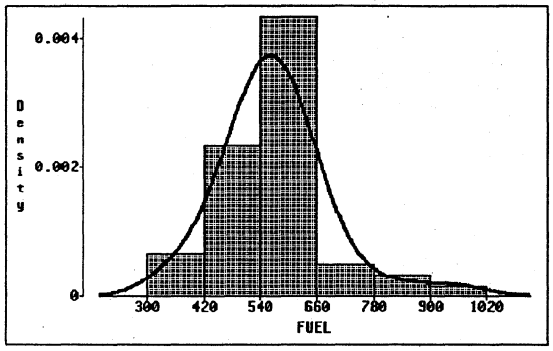


Figure 16 Density Estimation

The smoothed density estimate shows a heavy right tail, which may be due to a few states with abnormally high fuel consumption. You can examine the right tail more closely by looking at the box plot that also appears in the distribution window. Figure 17 shows the box plot for fuel with a mean diamond added. The crossbar on the diamond represents the mean (572), and the diamond extends one standard deviation on each side of the mean. The mean diamond is added by selecting **Edit → Graphs → Means**. The values of the median (566), the lower quartile (508), and the upper quartile (632) also appear. These values are added by selecting **Edit → Graphs → Values**. The thin bars or whiskers at each end of the box plot extend to a maximum of 1.5 times the Interquartile range. The two observations that fall beyond the whiskers are considered outliers. In this case, Wyoming and South Dakota have a much higher fuel consumption rate than the other states. These are large states with small populations and more fuel is required for transportation. You can remove the outlying observations from this and all calculations by selecting the observations and then selecting **Edit → Observations → Exclude in Calculations**.

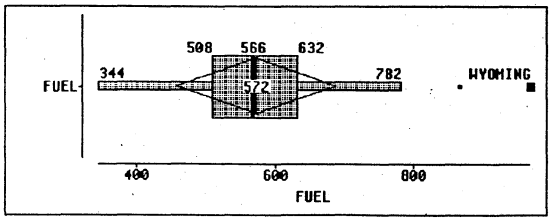
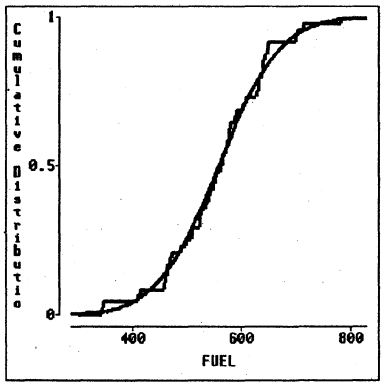


Figure 17 Box Plot

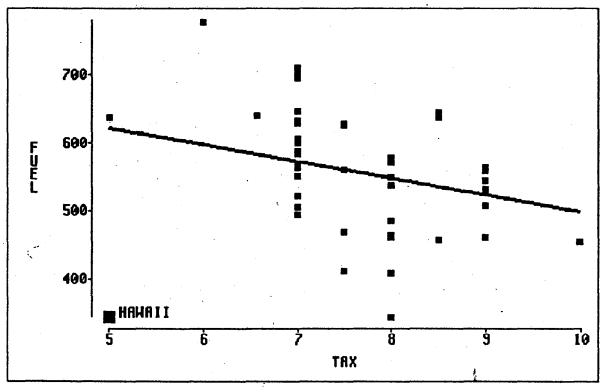
Some statistical procedures make the assumption that the data are from a normal distribution. You can test this assumption with SAS/INSIGHT by selecting **CDF → Test for Distribution → Normal** in the distribution window. This produces Figure 18, a plot of the empirical Cumulative Distribution Function and an estimate of the normal Cumulative Distribution Function. The estimated curve is much smoother than the empirical curve. The test for normality has a *p*-value >.15; thus, there is no evidence that FUEL is not normally distributed(when Wyoming and South Dakota are excluded).



Test for Distribution				
Distribution	Mean / Theta	Sigma	Kolmogorov D	Prob > D
Normal	557.2500	90.8964	0.0731	>.15

Figure 18 Test for Normality after Excluding Observations

Wyoming and South Dakota are not used in the remaining examples. You can now begin examining the relationship between rate of fuel consumption and the gas tax rate. Figure 19 shows a scatterplot of FUEL on TAX with a simple linear regression line fit. The Fit window was created by selecting the FUEL and TAX variables and then selecting **Analyze → Fit (Y X)**.



Summary of Fit			
Mean of Response	557.2500	R-Square	0.0775
Root MSE	88.2475	Adj R-Sq	0.0574

Figure 19 Simple Linear Regression

It appears that states that have higher gas tax rates tend to consume less fuel. The R-square value of .0775 in the Summary of Fit table indicates that 7.75% of the variation in fuel consumption is explained by the gas tax rate. Notice the outlying observation for Hawaii. Again, as with Wyoming and South Dakota, it is reasonable to suppose that the relationship between tax rates and fuel consumption is different for Hawaii than for the other states. When you remove Hawaii, the R-square value increases to .2237.

The data set includes three other variables that may help to explain fuel consumption. INC is per capita income in thousands of dollars. DLIC is the percentage of the population with a driver's license. ROAD contains thousands of miles of primary highways. Figure 20 shows a plot of fuel on each of these variables. You create these plots by selecting **Analyze → Scatter Plot (Y X)**, which brings up a Scatter Plot dialog. In the dialog select FUEL as the Y variable and TAX, INC, ROAD, and DLIC as the X variables. These plots show that DLIC has a strong positive relationship with fuel consumption and that as INC and TAX increase, fuel consumption appears to go down.

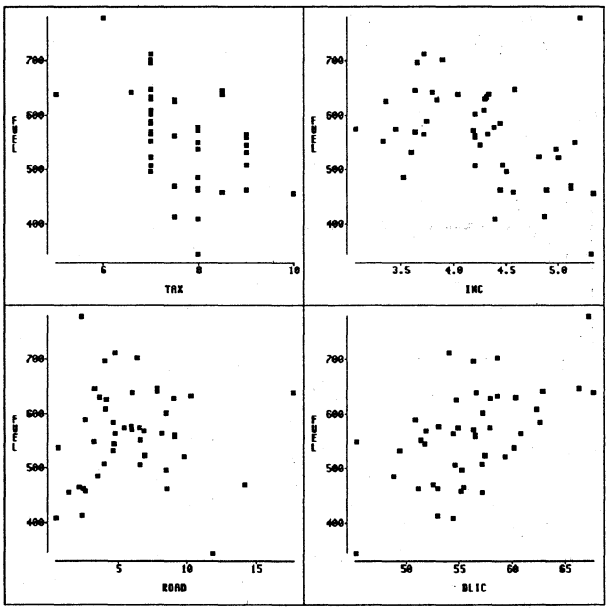


Figure 20 Scatter Plot Matrix

You can check the correlations of these variables with fuel consumption by creating a correlation matrix. Select each of the five variables in the data window and then select **Analyze → Multivariate (Y's)**. The multivariate window includes the correlation matrix that appears in Figure 21. Looking down the first column, note that DLIC has the strongest correlation with FUEL while TAX and INC have slightly weaker negative correlations. Also, notice that TAX and ROAD have a fairly strong negative correlation.

Correlation Matrix

	FUEL	TAX	INC	ROAD	DLIC
FUEL	1.0000	-.4730	-.3983	0.0672	0.5745
TAX	-.4730	1.0000	0.0388	-.5366	-.2529
INC	-.3983	0.0388	1.0000	0.0270	0.0324
ROAD	0.0672	-.5366	0.0270	1.0000	-.0242
DLIC	0.5745	-.2529	0.0324	-.0242	1.0000

Figure 21 Correlation Matrix

A multiple regression model can be fit by selecting **Analyze → Fit (Y X)**, which brings up a fit dialog. Select FUEL as the Y variable and ROAD, TAX, DLIC, and INC as the X variables in the Fit dialog. Then click on the **Run** button in the Fit dialog. The run button allows the fit dialog to remain on the screen after the model is fit. Figure 22 shows the TYPE I and TYPE III tests that result from this model. TYPE I tests are sequential tests so each sum of squares is adjusted only for the variables that appear before the current variable. TYPE III tests adjust the sum of squares for all other variables in the model. From these tables you can see that ROAD has a *p*-value of .4857 when in the model by itself but the *p*-value for ROAD decreases when adjusted for the other variables in the model. If you recall the high correlation between ROAD and TAX, this is not surprising.

Type I Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
ROAD	1	1545.6819	1545.6819	0.4948	0.4857
TAX	1	91768.3391	91768.3391	29.3787	0.0001
DLIC	1	64869.1160	64869.1160	20.7672	0.0001
INC	1	52936.6309	52936.6309	16.9471	0.0002

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
ROAD	1	3782.1223	3782.1223	1.2108	0.2774
TAX	1	35623.1413	35623.1413	11.4044	0.0016
DLIC	1	71384.4332	71384.4332	22.8530	0.0001
INC	1	52936.6309	52936.6309	16.9471	0.0002

Figure 22 Multiple Regression

You can drop ROAD from the model by going back to the fit dialog, selecting ROAD, clicking on **Remove**, and then clicking on the **Run** button. The new TYPE I and TYPE III tables appear in Figure 23. All of the remaining variables in the model are significant.

Type I Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
TAX	1	76571.3920	76571.3920	24.3940	0.0001
DLIC	1	75681.2532	75681.2532	24.1104	0.0001
INC	1	55085.0004	55085.0004	17.5489	0.0001

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
TAX	1	34786.1274	34786.1274	11.0821	0.0018
DLIC	1	81281.0740	81281.0740	25.8944	0.0001
INC	1	55085.0004	55085.0004	17.5489	0.0001

Figure 23 ROAD Dropped

To summarize, this analysis shows that there is a strong relationship between fuel consumption and the explanatory variables when certain states are

removed from the analyses. Tax rates alone are not as effective a predictor of fuel consumption as the combination of tax rates with other explanatory variables. Future research might examine the affect of changes in gas tax rates over time within states.

Nonparametric Smoothing Spline Fit

The ratio of net petroleum imports from OPEC to total U.S. petroleum product supplied(CONSUMP) was collected over a 28-year period. The data were collected from *The World Almanac 1989*. Figure 24 shows the scatterplot of CONSUMP on YEAR with a linear fit added. Obviously, a linear model is not appropriate for these data. The plot also contains two other curves that are very similar. The spline curve is added by selecting **Curves → Spline → GCV**. This is the "best" spline fit as determined by generalized cross validation. An 8th degree polynomial fit is also added by selecting **Curves → Polynomial → other...** and typing **8** in the resulting dialog. The plot shows that these two curves are very similar.

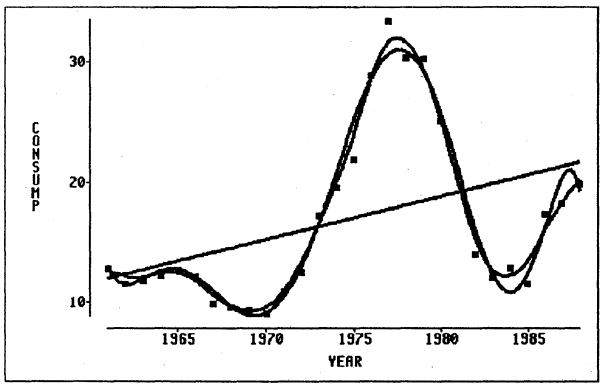


Figure 24 Spline Fit

You can examine the 8th degree polynomial fit more closely using a separate fit analysis. First, you want to scale the YEAR variable to avoid numerical precision problems. Select the YEAR variable and **Edit → Variables → other...**. In the dialog, select the standardized transformation and change the name of the variable to YEAR_STD. Then select **Analyze → Fit (Y X)**. Select CONSUMP as the Y variable and YEAR_STD as the X variable. You can type **8** in the expand field, then select YEAR_STD and click on the **Expand** button to create the 8th degree polynomial. A residual plot for this fit appears in Figure 25.

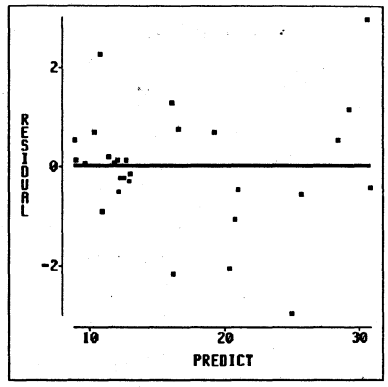


Figure 25 Residual Plot

Notice that the residuals tend to get larger in absolute value as the predicted values get larger. This nonconstant variance is often corrected by using a transformation of the response variable. By selecting CONSUMP in the fit window and then **Edit → Variables → log(X)**, you can refit the model with log consumption as the response. Figure 26 displays a residual plot of the transformed model that shows the transformation has corrected the nonconstant variance problem.

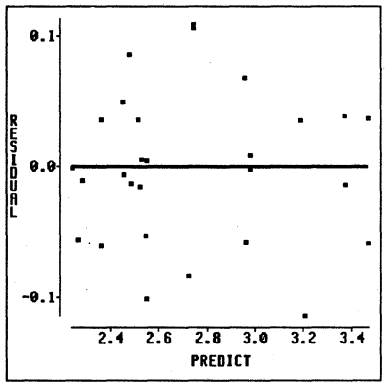


Figure 26 Residual Plot after Transformation

SAS/LAB software provides an easy way to check for a better transformation. You can fit the same 8th degree polynomial model in SAS/LAB and see what transformation is suggested. Figure 27 shows the Response Scaling window that suggests that the logarithm of CONSUMP does provide a better fit.

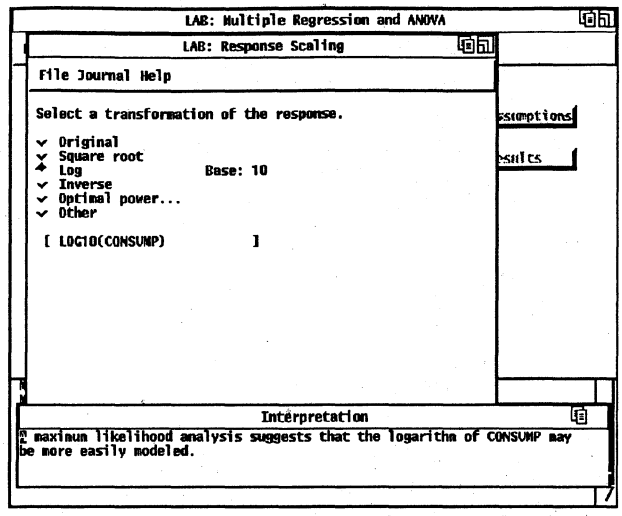


Figure 27 Residual Plot after Transformation

Generalized Linear Model

Samples of size 50 were collected for five age groups recording the number of people suffering from congenital eye disease on the Aegean island of Kalythos(Silvey, 1975). The variables are named AGE (midpoint of age group), BLIND (number in sample with disease), and TOTAL (50 for each group). A generalized linear model can be used to predict the probability of blindness for each age group. The generalized linear model is commonly used when the response variable is binomially distributed. A generalized linear model requires you to specify a link function that relates the mean of the response to a linear predictor and to specify a probability distribution for the response. In SAS/INSIGHT you can select **Analyze → Fit (Y X)** and click on **Method** in the fit dialog. In the method dialog, select **Binomial** for the Response Dist. and **Logit** for the Link Function. The logit function is defined as log(p/(1-p)), where p is the probability of disease. You must also select the denominator variable for the response ratio in the method dialog. Select TOTAL and then click on the **Binomial** variable selection box. After clicking on **OK** in the method dialog you can select BLIND as the **Y** variable and AGE as the **X** variable. Figure 28 shows the parameter estimates that result from this model. As expected, the chance of blindness increases with age.

Parameter Estimates

Variable	DF	Estimate	Std Dev	Chi-Sq	Pr > Chi-Sq
INTERCEPT	1	-3.5378	0.5023	49.6013	0.0001
AGE	1	0.0811	0.0108	56.2228	0.0001

Figure 28 Binomial Parameter Estimates

SAS/INSIGHT automatically stores the predicted probabilities for each fit in the data window. You save the probabilities in the data set permanently by selecting **File → Save → Data** in the fit window. Figure 29 shows a plot comparing the observed proportions with the predicted probabilities from the model. The observed probabilities are denoted by plus signs and the predicted probabilities by an X. You can change observation markers by selecting **Edit → Windows → Markers**. This creates a markers dialog that allows you to select observations and then click on the type of marker you want to assign to the selected observations. A SAS data step was used to append the predicted probabilities to the observed probabilities so that both variables could be shown in one plot.

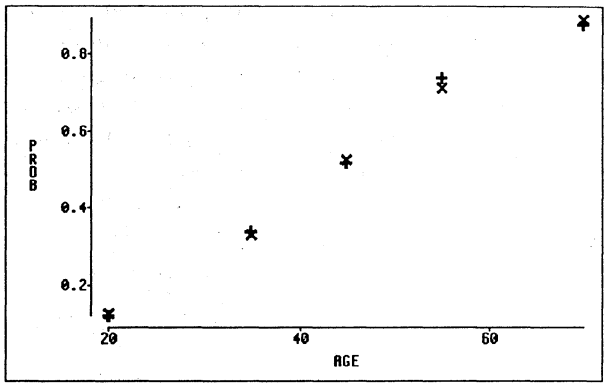


Figure 29 Probability Plot

SAS/INSIGHT also allows Poisson, gamma and inverse gaussian response distributions for fitting generalized linear models. You can also select from six different link functions as well as the canonical link for each distribution. Three different methods are available in the **Methods** dialog to correct for overdispersion in generalized linear models by controlling the estimation process for the scale parameter. You can specify a constant as the scale parameter or use the deviance or Pearson chi-square to estimate the scale parameter.

Multivariate Analysis

The World Almanac and Book of Facts 1990 contains crime rates for seven types of crime for all 50 states in 1988. You can use this data to explore which states in the country have the safest living conditions as well as the relationship between incidences of different types of crime. Figure 30 shows a plot that allows you to examine five of these variables in a single plot. You can create this plot by selecting the variables LARCENY, CARTHEFT, and BURGLARY and then selecting **Analyze → Rotating Plot (Z Y X)**. This creates a three-dimensional rotating plot. Using different symbols, you can add a fourth dimension to this plot. By selecting ROBBERY and **Analyze → Bar**

Chart (Y) you can create a histogram. Then select **Edit → Windows → Markers** to create an observation marker window. Now select the first two bars in the histogram of robbery and select the upside down triangle. Similarly select the next two bars, click on the circle, and then assign the last two bars a plus sign. To add a fifth dimension, create a box next to the rotating plot, add a histogram for AGGASULT to the box by selecting **Analyze → Bar Chart (Y)**, and then select AGGASULT for the Y variable. You can now brush the histogram of AGGASULT. You do this by drawing a small box in the histogram and dragging it through the bars to highlight each set of observations. Figure 30 shows the higlighted observations in the rotating plot associated with the first two bars in the histogram.

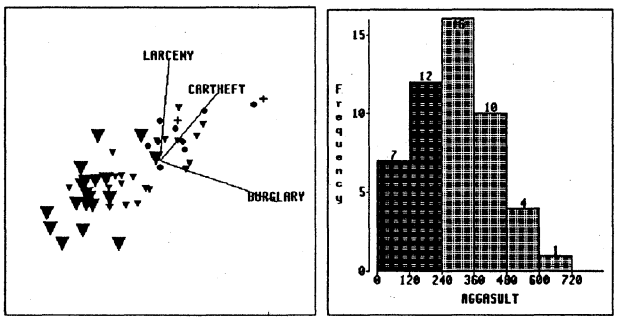


Figure 30 Exploring 5 Dimensions

A principal component analysis (PCA) is often used to reduce the dimensionality of data by finding linear combinations that maximize the variance between the original variables. The eigenvectors are used as transformations to create the orthogonal components. You can create a PCA in SAS/INSIGHT by selecting **Analyze → Multivariate (Y's)**. This creates a Multivariate dialog in which you can select all seven variables for Y. Under the **Method** button, select **Covariance Matrix** so that the PCA will use covariances in computations. The covariance matrix is preferred over the correlation matrix for this data set since all variables are measured on the same scale, occurrences per 100,000 people. Under the **Output** button select **Principal Component Analysis** and **Principal Component Options**. Select **3** in the **Eigenvectors** column of the options dialog to display only the first three eigenvectors. Figure 31 shows the eigenvalue and eigenvector tables that appear in the multivariate window.

Eigenvalues (COV)

Component	Eigenvalue	Difference	Proportion	Cumulative
PRINCOM1	784417.982	682076.420	0.8326	0.8326
PRINCOM2	102341.562	60006.8123	0.1086	0.9412
PRINCOM3	42334.7501	32714.2059	0.0449	0.9861
PRINCOM4	9620.5442	6315.2919	0.0102	0.9963
PRINCOM5	3305.2523	3172.1657	0.0035	0.9998
PRINCOM6	133.0866	119.2061	0.0001	1.0000
PRINCOM7	13.8805	.	1.E-05	1.0000

Eigenvectors (COV)

Variable	PRINCOM1	PRINCOM2	PRINCOM3
RAPE	0.0072	0.0107	-0.0102
ROBBERY	0.0968	0.2940	0.3087
AGGASULT	0.1199	0.3118	0.1169
BURGLARY	0.3952	0.5571	-0.7152
LARCENY	0.8888	-0.4321	0.1523
CARTHEFT	0.1734	0.5648	0.5967
MURDER	0.0043	0.0131	0.0098

Figure 31 Eigenvalues

The Proportion column of the eigenvalues table shows the proportion of variation explained by each component. The eigenvector table gives the loading or transformation for each variable. Notice the first component explains 83 percent ot the variation and is heavily weighted by LARCENY. The second component contributes another 11 percent of the variation and appears to be a contrast between BURGLARY and CARTHEFT versus LARCENY. With just the first three components you can explain over 98 percent of the crime rate variation among states. Figure 32 shows a plot of the first three principal components with the three highest (Arizona, Florida, and Texas) and three lowest (Pennsylvania, South Dakota, and West Virginia) values for the first component selected.

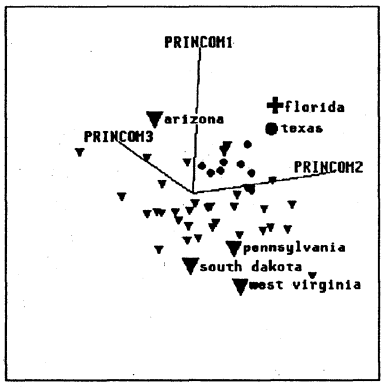


Figure 32 Principal Component Plot

SUMMARY

SAS/LAB and SAS/INSIGHT software address the needs of different users for interactive and graphical statistical software. While SAS/INSIGHT software provides a general system for exploring and visualizing data, SAS/LAB software focuses on the data analysis needs of nonstatisticians, particularly engineers and research scientists. In many cases you will want to use both SAS/LAB and SAS/INSIGHT software. For example, experienced statisticians may

want to use SAS/LAB software to get a quick analysis, with all assumptions checked. If there seem to be anomalies in the data, SAS/INSIGHT software can be used to explore them. Likewise, the interactive graphics in SAS/INSIGHT software can be useful to statisticians and nonstatisticians alike.

REFERENCES

Box, G.E.P., and Bisgaard, S. (1987). "The Scientific Context of Quality Improvement." *Quality Progress,* June 1987, **20**: 6, pp. 54-61.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparisons Procedures*, New York: John Wiley and Sons Inc.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979) *Multivariate Analysis*, Orlando, FL: Academic Press Inc.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd Edition, New York: Chapman and Hall.

Robinson, Heman (1992). "Analytic Modeling in SAS/INSIGHT® Software." *Proceedings of the Seventeenth Annual SAS® Users Group International Conference*, Cary, N.C.: SAS Institute Inc.

SAS Institute Inc. (1993), SAS/INSIGHT® User's Guide, Version 6, Second Edition, Cary, N.C.: SAS Institute Inc.

Silvey, S.D. (1975). *Statistical Inference*, New York: Chapman and Hall.

Watson, Wayne E., Roggenkamp, Kathy, and Tobias, Randall D. (1993). "SAS/LAB® Software for Guided Data Analysis." *Proceedings of the Eighteenth Annual SAS® Users Group International Conference*, Cary, N.C.: SAS Institute Inc.

Weisberg, Sanford (1985). *Applied Linear Regression* 2d Edition, New York: John Wiley and Sons, Inc.