ABSTRACT
Analytic modeling software has evolved from line-mode, batch-oriented systems to graphical, direct-manipulation interfaces. Mouse- and menu-based input eliminates complex command syntax and makes users of these new interfaces highly efficient. SAS/INSIGHT software provides an efficient environment for exploratory data analysis and model building. This paper describes exploration and modeling capabilities in current and future releases of SAS/INSIGHT software.

EXPLORATION
A typical data exploration technique is to plot all variables pairwise. This display, often called a scatter plot matrix, will show two-dimensional relationships in the data if any exist (Chambers, Cleveland, Kleiner, and Tukey, 1983). In SAS/INSIGHT software, no complex command syntax is needed to create a scatter plot matrix. Simply select the variables of interest and choose the menu "Graph:Scatter Plot".

The scatter plot matrix in Figure 1 shows data from a physical fitness class. Variables are OXYGEN (oxygen intake), RUNTIME (time in minutes to run 1.5 miles), and RSTPULSE (resting pulse rate). The scatter plot matrix shows clearly the relationship between OXYGEN and RUNTIME.

Occasionally a relationship that will not show in two dimensions will become obvious in three. Figure 2 shows a three-dimensional plot of the same variables shown in the scatter plot matrix. To create the three-dimensional plot, just select the variables of interest and choose the menu "Graph:Rotating Plot".

You can rotate the plot about any axis to explore patterns in the data. To rotate the plot you "grab" it and turn it, just as you would rotate any spherical object. A direct-manipulation interface such as this takes advantage of knowledge you already have, rather than requiring you to learn more.

In both two- and three-dimensional plots, you can identify observations by pointing at them and clicking. This displays either the observation number or the value of a specified variable. Identifying observations by pointing and clicking allows quick examination of potential outliers.
When you identify observations by pointing and clicking, you highlight observations in all plots, not just at the location of the click. Similarly, you can highlight groups of observations in all plots by selecting a group in any plot. Highlighting observations in multiple plots, or brushing, is an effective way to explore the structure of the data in higher dimensions (Becker and Cleveland, 1987). Figure 4 shows a rectangular brush drawn in one plot highlighting the same observations in all plots.

Every graph in SAS/INSIGHT software is a view of the same data. Not only highlighting but color, marker shape, and observation values are all held in common, and the displays are updated automatically. Thus, any change to the data shows immediately in all graphs and analyses.

REGRESSION

Data exploration can often suggest which models to fit. For this data, OXYGEN is the variable of interest and RUNTIME an obvious explanatory variable. To fit a model, simply select OXYGEN, then RUNTIME, and then choose the menu "Analyze:Fit". The result is a regression fit with a scatter plot.

You can add various curves to the scatter plot. The plot in Figure 5 shows fitted spline and polynomial curves together with the fitted line. Tabular output below the scatter plot shows detailed information on the curve and linear fits.

To create other models, you can easily add additional explanatory variables. By selecting OXYGEN, RUNTIME, and RSTPULSE, and then choosing Analyze:Fit, you can add the resting pulse rate to the model.
By selecting OXYGEN, RUNTIME, RSTPULSE, and AGE, you can add the age of the participant to the model.

Creating these three models requires roughly four keystrokes per model. On a moderately fast workstation, these models would appear in less time than it took to read this description.

**DIAGNOSTIC OUTPUT**

Creating models produces output variables such as residuals, predicted values, and normal scores. In SAS/INSIGHT software, all such output variables are immediately available for diagnostic use.

For example, Figure 8 shows a distribution analysis used to test for normal distribution of residuals from one of the fitness models described earlier. The distribution analysis includes a histogram with Kernel Density Estimate, Moments and Quantiles tables, Normal Q-Q Plot, and a plot of the Cumulative Distribution Function.

The histogram is adjustable, an innovation adapted from JMP® software (SAS Institute Inc., 1989). By 'grabbing' the bars, you can directly manipulate the bar width and position to efficiently examine the distribution of residuals.

Figure 9 shows the histogram of Figure 8 adjusted with smaller bar width and new positions.
It is often desirable to test models by constructing plots of residuals against some other variable. By marking a location for the plot, selecting the residual and one of the model variables, and choosing "Graph: Scatter Plot", you can add a scatter plot to the analysis as shown in Figure 10.

One special case of the general linear model is the analysis of variance. Cochran and Cox (1957) discuss a study on the effects of electric current in denervated muscle tissue. The variables are REP, the replicate number; TIME, the length of time current was applied; CURRENT, the level of electric current; NUMBER, the number of treatments per day; and Y, the weight of the muscle tissue.

If you choose "Analyze: Fit" without selecting variables, you will be prompted to enter the model in a dialog. Figure 11 shows the dialog used to create the full three-way factorial model.

If you decide some of the higher-order effects are not significant, it is quick and easy to select them, click the "Remove" button, and create new models. Figure 12 and Figure 13 display the results when TIME*CURRENT*NUMBER and then TIME*CURRENT are removed to create two new models.

In the same manner, you can quickly add any number of plots to the analysis to facilitate examination of your models.

**ANALYSIS OF VARIANCE**

All of the features described above are available in the first release of SAS/INSIGHT software. In the second release more statistical models are available, including the General Linear Model. All models supported in the GLM procedure (SAS Institute, Inc., 1990) are supported in the second release of SAS/INSIGHT software.
Creating these three models takes on average six keystrokes per model. Again the models can be created in less time than it took to read these pages.

Counting keystrokes is a common technique in software design to optimize important features. Early in the design of SAS/INSIGHT software, statisticians identified modeling as a frequent, iterative activity, one that was critical to efficient use of the product. Now the ability to create models quickly and get diagnostic output easily is a feature of SAS/INSIGHT software.

ANALYSIS OF COVARIANCE

Another special case of the general linear model is analysis of covariance. Snedecor and Cochran (1967) use the example of a drug study where DRUG represents two antibiotics, A and D, and a control F, X represents a score of bacilli present before treatment, and Y a score of bacilli present after treatment. Figure 14 shows an analysis of covariance with a plot of residual against predicted values.

GENERALIZED LINEAR MODEL

In recent years researchers have worked to make linear modeling techniques apply to a wider class of problems. This more general class of models is usually known as the Generalized Linear Model (Nelder and Wedderburn, 1972). The generalized linear model can accept non-normal error distributions from the exponential family and can use a link function to relate the expected response value to a linear form of the explanatory variables.

The second release of SAS/INSIGHT software supports error distributions

- Normal
- Inverse Gaussian
- Gamma
- Poisson
- Binomial

and link functions

- Canonical
- Identity
- Log
- Logit
- Probit
- Complementary log-log
- Power

By choosing appropriate error distributions and link functions, you can specify a variety of different models. For example, by choosing a Binomial error distribution and Logit link function as in Figure 15, you can specify a logistic regression model.

Using the generalized linear model you can easily add or remove effects to create and examine new models as needed, in the same way you do for the general linear model.
SAVING RESULTS

Once models have been created and examined it is necessary to save the results. In SAS/INSIGHT software, you can save results in three ways:

- as SAS® data sets
- as SAS/GRAPH® output
- as ASCII text files

All output variables from analyses, including residuals, predicted values, hat diagonal, normal quantile scores, and others, can be saved to SAS data sets.

All objects that appear on the screen can be saved as SAS/GRAPH catalogs or graphics stream files. For presentation purposes you can enhance graphs interactively with the Graphics Editor (SAS Institute Inc., 1991).

In the second release of SAS/INSIGHT software, all tabulated output, including Parameter Estimates, Analysis of Variance, Summary of Fit, and other output, can be saved as SAS Output Objects. You can translate Output Objects into SAS data sets or ASCII text with PROC OUTPUT, a recent enhancement to base SAS software (SAS Institute Inc., 1992).

SCRIPTING

Data analysts must be able to record and reproduce their results. In the second release of SAS/INSIGHT software, the creation of every graph or analysis can be logged to a script. You can use scripts to record the models you create. You can also use scripts as input to SAS/INSIGHT software to re-create any phase of your analysis.

Shown below is a script to create the three fitness models and the distribution analysis described earlier in this paper.

```plaintext
menu Analyze:Fit Y = OXYGEN,
   Effects = 'RUNTIME';
menu Analyze:Fit Y = OXYGEN,
   Effects = 'RUNTIMERSTPULSE';
menu Analyze:Fit Y = OXYGEN,
   Effects = 'RUNTIMERSTPULSEAGE';
menu Analyze:Distribution
   Y = R_OXYGEN;
```

The script language also supports complete specification of the generalized linear model.

CONCLUSION

SAS/INSIGHT software offers a highly interactive environment for data exploration and modeling. Without learning complex command syntax, with a minimum of keystrokes and mouse movement, you can efficiently explore data, build models, and save results. Minimizing user effort throughout SAS/INSIGHT software frees data analysts to spend their time exploring and understanding their data.

REFERENCES


SAS, SAS/GRAPH, JMP, and SAS/INSIGHT are registered trademarks or trademarks of SAS Institute Inc. ® indicates USA registration.