ABSTRACT

Analytic modeling software has evolved from batch-oriented systems to graphical, exploratory systems with friendly user interfaces. SAS/INSIGHT software provides a dynamic environment for exploratory data analysis and model building. This paper describes capabilities of SAS/INSIGHT software, including exploratory graphics, distribution analysis, and the generalized linear model.

EXPLORATION

Tukey states that “exploratory data analysis can never be the whole story, but nothing else can serve as the foundation” (Tukey 1977). Data exploration often suggests appropriate models.

One common exploratory technique is to compare distributions. For this purpose SAS/INSIGHT software provides dynamic implementations of Tukey’s box plots, in which you can point and click to identify extreme values. Figure 1 shows box plots of data from a physical fitness class. OXYGEN (oxygen intake) is compared by AGEGRP (age group).

Another common exploratory technique is to search for two-dimensional relationships in a scatter plot matrix (Chambers et. al. 1983). Figure 2 shows OXYGEN plotted against RUNTIME (time to run 1.5 miles), and RSTPULSE (resting pulse rate). This scatter plot matrix reveals a linear relationship between OXYGEN and RUNTIME.

Occasionally a relationship that is not evident in two dimensions becomes obvious in three. Figure 3 shows a three-dimensional plot of the same variables shown in the scatter plot matrix. You can rotate the plot about any axis to explore patterns in the data.
In all graphs, you can highlight groups of observations by selecting a rectangular area. Because all graphs are linked to the same data, observations highlighted in one window are highlighted in all windows. Highlighting observations in multiple graphs, or **brushing**, is an effective way to explore structure in higher dimensions (Becker and Cleveland 1987).

Not only brushing but coloring, marker shape, and all other observation states are linked across all graphs and analyses. For example, when you exclude outliers from calculations, all analyses are automatically recalculated. Any change you make to the data shows immediately in all windows.

**REGRESSION**

For the physical fitness data, exploration indicates a linear relationship between OXYGEN and RUNTIME. To fit this model, simply select OXYGEN, then RUNTIME, and then choose the menu “Analyze:Fit”. The result is a scatter plot with a fitted regression line.

Polynomial, kernel, and spline curves can be added to the scatter plot. In Release 6.10, loess smoothers (Cleveland 1993) are also available, and all curves are dynamically adjustable with slider controls. The plot in **Figure 6** shows spline and loess curves together with the fitted line. Tables below the scatter plot contain the slider controls and show detailed information on the curve and linear fits.

![Figure 6 OXYGEN = RUNTIME](image)

It is a simple matter to create more complex models by adding explanatory variables. By selecting OXYGEN, RUNTIME, and RSTPULSE, and then choosing “Analyze:Fit”, you can add the resting pulse rate to the model. Similarly, by selecting OXYGEN, RUNTIME, RSTPULSE, and AGE, you can add participants' ages to the model.

Creating these models requires, on average, four keystrokes per model. On a moderately fast workstation or personal computer, the models shown in Figures 6, 7, and 8 would appear in less time than it took to read this description.
### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Std Error</th>
<th>T Stat</th>
<th>Prob &gt;</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>92.4817</td>
<td>6.6222</td>
<td>13.8655</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>RUNTIME</td>
<td>1</td>
<td>3.4169</td>
<td>0.4405</td>
<td>-7.7103</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>RSTPULSE</td>
<td>1</td>
<td>-0.0121</td>
<td>0.0582</td>
<td>-0.2105</td>
<td>0.8358</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>1</td>
<td>-0.0900</td>
<td>0.0582</td>
<td>-1.5751</td>
<td>0.1234</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7** $\text{OXYGEN} = \text{RUNTIME RSTPULSE}$

**Figure 8** $\text{OXYGEN} = \text{RUNTIME RSTPULSE AGE}$

### DIAGNOSTIC OUTPUT

SAS/INSIGHT software provides a variety of diagnostic output variables including residuals, predicted values, hat diagonal, standardized and Studentized residuals, Cook's D, Dffits, Covratio, and Dfbetas (Hamilton 1992; Rawlings 1988). All output variables are immediately available for diagnostic use.

For example, **Figure 9** 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 estimates, normal Q-Q plot, and a plot of the cumulative distribution function.

**Figure 9** Distribution of Residuals

The histogram is adjustable, one of many innovations adapted from JMP® software (SAS Institute Inc. 1989). By "grabbing" and moving the bars, you can adjust bar width and position to efficiently examine the distribution of residuals. **Figure 10** shows the histogram of **Figure 9** adjusted to a smaller bar width and new positions.

**Figure 10** Adjusted Histogram
One way to test models is to plot the residuals against other variables. By marking a location for the plot, then selecting variables, you can add a residual plot to the analysis as shown in Figure 11. In the same manner, you can easily add plots of any diagnostic variables to test your models.

If you decide some of the effects are not significant, it is easy to select them and choose “Edit:Delete” to remove the effects and recalculate the model. Figure 13 and Figure 14 display the results when the effects TIME*CURRENT*NUMBER and TIME*CURRENT are removed to create two new models.

If you decide some of the effects are not significant, it is easy to select them and choose “Edit:Delete” to remove the effects and recalculate the model. Figure 13 and Figure 14 display the results when the effects TIME*CURRENT*NUMBER and TIME*CURRENT are removed to create two new models.

### ANALYSIS OF VARIANCE

Release 6.08 of SAS/INSIGHT software supports both the general linear model and the generalized linear model. The general linear model is the same class of models provided in the GLM procedure (SAS Institute Inc. 1990). One special case of the general linear model is a factorial model.

Cochran and Cox (1957) discuss analysis of variance in a study of the effects of electric current in muscle tissue. The variables are REP (replicate number), TIME (time current was applied), CURRENT (level of electric current), NUMBER (treatments per day), and Y (weight of muscle tissue). Figure 12 shows the dialog used to create a full three-way factorial model.

### ANALYSIS OF COVARIANCE

Another use 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 15 shows an analysis of covariance with a plot of residual against predicted values.
Researchers have worked to make linear modeling techniques apply to a wider class of problems. One popular extension is known as the generalized linear model (Nelder and Wedderburn 1972). The generalized linear model can accept non-normal response distributions and can use a link function to relate expected response values to explanatory variables. Release 6.08 of SAS/INSIGHT software supports normal, inverse Gaussian, gamma, Poisson, and binomial response distributions; and log, logit, probit, complementary log-log, and power link functions.

Choosing from these response distributions and link functions enables you to fit a variety of different models. One special case of the generalized linear model is logistic regression. Finney (1947) and Pregibon (1981) describe a study of a vasoconstriction reflex in the skin of the digits. Variables are RESPONSE (whether the reflex occurred) RATE and VOLUME (rate and volume of air inspired) and LOGRATE and LOGVOL (logs of the rates).

By choosing binomial response distribution and logit link function as in Figure 16, you can specify a logistic regression model. Figure 17 shows some of the model results including summary of fit, Type I and Type III likelihood ratio tests, and parameter estimates.

The generalized linear model also provides a variety of diagnostic output, including standardized and Studentized deviance and Pearson and Anscombe residuals. You can generate output variables dynamically or cause them to be generated automatically for every model.

Once models have been created and examined, it is necessary to save the results. In SAS/INSIGHT software, results can be saved in five formats:

- as SAS® data sets
- as text files
- as SAS output objects
- as SAS/GRAPH® catalogs
- as graphics files

SAVING RESULTS
All output variables from analyses can be saved to SAS data sets. Assignments of observation colors and markers are saved also.

All tables can be written to the SAS output window and printed as text.

Tables can also be saved as SAS output objects. Output objects can be translated into SAS data sets using the Output Delivery System, a feature of base SAS software (SAS Institute Inc. 1992).

All objects that appear on the display can be saved as SAS/GRAPH catalogs.

Release 6.10 of SAS/GRAPH software enables saving to graphics files in popular bitmap formats. Many hosts also provide tools to save and print graphics directly from the display.

**RECORDING ANALYSES**

In addition to saving results, it is often useful to record the steps in an analysis. Release 6.10 of SAS/INSIGHT software provides statements to record the creation of every graph and analysis window. Shown below is a record of the three fitness models and the distribution of residuals described earlier in this paper.

```sas
fit oxygen = runtime;
fraction oxygen = runtime rstpulse;
fraction oxygen = runtime rstpulse age;
dist r_oxygen;
```

SAS/INSIGHT statements support complete specification of the generalized linear model.

**CONCLUSION**

SAS/INSIGHT software offers a highly interactive environment for data exploration and modeling. You can efficiently explore data, build models, and save results, with a minimum of keystrokes and mouse movement, without learning complex commands. Minimizing the learning curve throughout SAS/INSIGHT software gives data analysts more time for exploring and understanding their data.

**REFERENCES**


SAS, SAS/GRAPH, JMP, and SAS/INSIGHT are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.