

Using SAS/Insight as an Introductory Data Mining Platform

Robin Way
Director, Market Research
NW Natural
Portland, Oregon

Abstract

SAS software has been recognized by the broad IT community as an outstanding platform for data mining, given its combination of data warehousing and analytic techniques. A very powerful way to conduct data mining is available today using SAS/Insight software. While not as comprehensive as the soon-to-arrive SAS Enterprise Data Miner, the SEMMA approach promoted by SAS Institute applies very well to the suite of data analysis and visualization techniques available in SAS/Insight. This paper will review the usefulness of SAS/Insight for exploratory data analysis, interactive regression modeling, and advanced multidimensional data visualization. Along the way, we'll explore how to win baseball games and still save enough money to build the stadium luxury boxes next year.

Review of the SEMMA process

SAS Institute has done a fine job of popularizing this acronym, so discussion of SEMMA here will be brief. I believe the idea behind the acronym is a sound and accurate reflection of how data mining always has been conducted, whether recognized (currently) as a distinct field of inquiry or (traditionally) as one group of skills required for data management and analysis.

Sample

As disk storage costs decline and corporate data capture grows, the capability of the firm to store increased amounts of data in locations featuring rapid access will grow. As analysts' capability to handle and manipulate increasing stores of data also improve, we will eventually encounter diminishing returns on analyzing the next piece

of data. This will occur because the incremental addition to knowledge represented by the marginal piece of data will become less and less, to the point where the costs of analyzing will grow faster than the increase in utility of doing so.

This is why sampling is an important part of data mining. SAS software is full capable of helping to draw an optimally sized sample that features the appropriate mix of characteristics representative of the population. Analyzing this sample instead of the entire population will enable analysts to reach conclusions faster and exploit the leaps of logic that occur when that rare and precious experience—personal immersion in the data—arises to close the gap between data and knowledge.

Explore

SAS/Insight offers a variety of statistical visualization techniques for exploring data. In the same way that statistics summarize characteristics of a variable, statistical visualization summarizes characteristics of one or more variables in one or more dimensions at the same time. While initial exploration of data may require one or two-dimensional views of data, such as histograms, mosaic plots, boxplots and scatter plots, more sophisticated approaches are also available. Rotating plots, multiple scatter plots, color and animation vastly increase the speed with which analysts can grasp fundamental relationships in the data.

Model

Insights gained through exploration transition easily into hypothesis testing in SAS/Insight. Measures of centrality and dispersion, linear regression, and multivariate methods like principal components can all be examined and tested in SAS/Insight. Neural nets are available

in Enterprise Miner, but some of the mathematical functionality offered by neural nets is available in SAS/Insight through linear-in-parameters regression.

Modify

Of course, our initial models will give rise to new ways of considering relationships among variables. The "modify" step is a good time to consider and test alternatives. For example, what would happen if you instead analyzed the log transform of the dependent variable in a regression model? Or entered a new variable into a principal components analysis? Or created an interaction term between two existing independent variables in a logistic regression? All these activities can be conducted with ease and speed in SAS/Insight.

Another activity familiar to practitioners is the identification and elimination of outliers and model estimation issues such as heteroscedasticity and multicollinearity. SAS/Insight offers not only a good variety of diagnostics for identifying outliers in models (hat matrix, DFfits, DFbetas, Q-Q plots, partial leverage plots) but also allows the analyst to judge the effects of outlier removal interactively and in real time. Collinearity diagnostics like tolerance and variance inflation are available, as are plots of regression predictions vs. residuals for identifying non-constant variance.

Assess

If there were one aspect of SEMMA that SAS/Insight does not tackle directly, it would be in comparing cross-model performance in a way that translates to the bottom line. The Enterprise Miner package will offer ROI calculations across broad modeling approaches (i.e., regression, neural nets, CHAID) as well as across variations within models of the same approach. Nonetheless, a fair amount of this activity is subject to human judgment and experience, and all modeling approaches need to connect with the firm's strategy.

Business Need

Every business initiative should relate back to a demonstrable business need. The best way to

demonstrate data mining without having to provide lots of background on a specific business is to use an example with which everyone's familiar—major league baseball!

So, imagine you are the Director of Franchise Development for the local major league baseball team: the invisible part of the team that decides how to balance the franchise's investments in recruiting and keeping talent, advertising & public relations, concessions, and stadium development. Marketing reports indicate that fans are pretty happy with the beer and hot dogs, with the new retractable dome, and with television coverage. Attendance at games is strong, as is the pitching staff. You've got a manager that motivates and challenges the team. However, some of the recently hired free agents aren't producing as many home runs and RBI's as expected, and your win/loss record is suffering. What's more, satisfaction is down among your season-ticket holders because all your existing luxury boxes are sold out for the next ten years, and funds are tight due to the free agent salaries.

In summary, you've got to win more games to keep the fans happy, but you can't go spending an arm and a leg for new talent. The data mining objectives therefore:

- Hire more effective batting and fielding talent
- Win more games by bringing in more runs
- Keep new salary costs down

The objectives for this project are to select a subset of players to recruit who offer better than average value for their current salary. In order to accomplish this end result, we will have to first get a lay of the land, and understand what factors contribute to player salaries. Then by comparing what players should be paid with their actual salaries, we can identify our set of recruits.

The dataset for this project is sasuser.baseball. This allows all SAS users to repeat the project described in this paper and compare alternative results and techniques with those described below. I've made a few additions to the dataset, by creating new variables for current season batting average and career batting average. I've

also color-coded all players' observations depending on whether they play infield, outfield, or other (utility players and designated hitters).

Exploratory statistical analysis

Exploratory analysis helps us get our footing in a new dataset by analyzing variable distributions and inter-variable relationships.

One-dimensional

Histograms and boxplots are a quick way of describing a continuous scale variable's distribution, and enable the analyst to quickly assess the distribution's moments and quantiles. Nominal scale variables can be analyzed using mosaic charts, which are essentially visual versions of a frequency table.

A SAS/Insight technique called "brushing" allows the analyst to select portions of one variable's distribution with the mouse, and instantly view in a different window how the observations underlying the selection are related to the distribution of another variable. The brush can be re-sized and moved across any variable's chart, enabling a real-time analysis of variable relationships.

A histogram of player salaries indicates a distribution skewed to the right; a few players receive high salaries while most (72%) make less than \$1 million. Another histogram indicates career length (years in the major league) is also skewed to the right, but to a less severe degree. There is a distinct decline in the distribution after seven years.

It is not necessary to request one histogram at a time. We have constructed two more histogram charts, each containing four variables in their own section of the window. This allows the analyst to view distributions of multiple variables quickly and in a single window, which is useful in exploration mode. The first histogram window shows variable distributions for 1986 single-season hitting statistics (batting average, runs, home runs, and runs batted in). A second histogram window shows the distributions for the same variables representing career statistics for the same measurements. Distributions of

corresponding variables are similar between single-season and career statistics. Batting averages follow a roughly normal distribution, while other batting statistics (counts, rather than ratios) are skewed to the right. The career statistics for runs, homers, and RBIs are more sharply skewed, suggesting players with distinctively successful batting performance over their career number relatively fewer than those who hit well in a single-season.

By brushing the salary distribution, we can see players with high salaries tend to have high numbers of RBIs but don't stand out in terms of batting average. However, career statistics do identify a distinct group of highly paid players who have high career batting averages.

Two-dimensional

Two-dimensional charts allow the analyst to visualize inter-variable relationships in a single chart, instead of brushing one chart and viewing relationships in another window. There are also a variety of two-dimensional charting techniques in SAS/Insight, including scatter and line plots, as well as two-dimensional boxplots of a continuous variable compared against classes of a categorical variable. Mosaic charts can also be used to compare two categorical variables.

Scatter plots identify the relationship of salary with both single-season and career batting average. It appears career batting averages have a stronger correlation with salaries than single-season averages, although in both cases the correlation is clearly positive and non-zero. Scatter plots can also be used to compare a nominal-level variable with several levels against a continuous variable, such as the scatter plot comparison of salaries by team. Quickly a number of very highly paid players are identified, such as Eddie Murray, Jim Rice, Wade Boggs, Mike Schmidt and Dale Murphy. A similar result comes from a boxplot of salary statistics by team; the Boston Red Sox paid the highest mean salary (\$961,000 per year). (One small but confounding problem with the data is that all teams in `sasuser.baseball` are categorized by their host city, and some cities host more than one team; the only way to tell the Yankees from the Mets is by their league affiliation.)

Three-dimensional

Even more powerful visualization is available by using three-dimensional charts. Multivariate relationships can be explored and analyzed, usually as a precursor to regression analysis and multivariate procedures such as principal components and factor analysis.

Multiple scatter plots can be used to quickly assess relationships among many variables at once. These multiple scatter plots bear a lot in common with crosstabulations, where the shape of the scatter has a direct relationship with measures of association like the Pearson product-moment correlation coefficient. The first multiple scatter plot compares salary with single-season fielding statistics. It is difficult to conclude from these plots that salary has much of a linear relationship with any fielding statistics; one interesting trend is between put outs and assists. (It turns out this relationship is an artifact of the way baseball fielding statistics are tabulated; infielders like first and second basemen acquire many more put outs than other players, while shortstops and outfielders have many more assists.)

A second set of multiple scatter plots compares salary with single-season and career batting statistics. It is less difficult to draw conclusions from comparisons within like variables than with salary. Salary appears to be most strongly correlated with career runs and career batting average. Strong correlations appear between career homers, runs, and RBIs. A similar but less pronounced correlation is present for the same measurements at the single-season level.

We will explore rotating three-dimensional plots in the subsequent section of this paper on multi-dimensional visualization.

Fourth dimension

Formerly the hallowed ground of rocket scientists, SAS/Insight gives any analyst the ability to explore the mysterious "fourth dimension" and beyond. The application of color and symbology explores extra-dimensional patterns indirectly, while animation provides a more direct analysis, particularly when applied to time and date variables.

The business simulation used in this paper does not take direct advantage of animation, but the use of color is employed to designate the field position of each player. Red is used for infield players (catcher, first, second, and third bases, and shortstop), green for outfield players (left, center, and right), and blue for other positions (utility players, designated hitters, and combination position players).

If we had multiple years of player statistics, or monthly statistics, it might be appropriate to use animation to view the trends in hitting or fielding over the course of a season or across seasons. The choice of animation versus a multi-dimensional chart depends on the context—generally I have found it helpful to use color or symbology to indicate multiple time periods within a standard chart without animation. However, as a single chart becomes more complex and thus harder to interpret, the judicious use of animation has payoffs.

Linear regression modeling & modification

The process of building a regression model in SAS/Insight is inherently interactive and offers significant advantages over interactive model building in traditional regression procedures in SAS/STAT. Diagnostics are plentiful and easy to access, making model modification quick and painless.

Building a model

Generally regression models are developed using the dialog box-based "Fit" application that comes with SAS/Insight. The Fit menu allows the analyst to specify dependent variables and independent variables, to interactively cross and nest independent variables, and to specify model weights and by-variables. Dependent variables can be continuous or categorical and there are a variety of model fitting options, including response distribution, link function, and scaling parameters. The menu also allows the analyst to select new variable to be output from the model, such as predicted and residual values of the dependent variable, and to select from a variety of tabular and graphical fit diagnostics.

Diagnostics

The diagnostics that I consult most regularly for multiple linear regression include the traditional r-squared, F, t, and variance inflation statistics. I also find very useful the following charts:

- Residual versus predicted dependent variable plots for testing error variance for heteroscedasticity
- Residual normal quantile-quantile plots for testing normal distribution of residuals
- Partial leverage plots for examining the partial correlations of each independent variable and the likely effects of outlying observations

Modifying the model

After an initial model has been specified and run, the analyst will generally rely on interpretation of estimated parameters and fit diagnostics to assess model performance. At this point, several options are available. Using the Fit menu, independent variables can be added to the model to capture unexplained variance; independent variables that don't appear to explain any variance can be dropped from the model. If an independent variable's relationship with the dependent variable appears to be curvilinear or non-linear, independent variables can be transformed in SAS/Insight with one or two mouse clicks and added to the current model with two more.

Outlying observations that affect model performance can be examined and dropped, or a new variable that explicitly captures its effects can be created and added to the model. A very handy part of SAS/Insight's fit charting capabilities is that observations dropped from a model can be removed from calculations but still displayed (and marked accordingly). This permits the analyst to evaluate in real time whether or not the observation's effect is significant.

Model modification is faster when the analyst using the Fit menu uses the "Apply" button rather than the "OK" button to run a new model—this leaves the menu intact rather than closing it before the fit is displayed.

Using regression modeling to develop baseball recruiting profile

We used SAS/Insight to fit four multiple linear regression models. The first three models examined separate sets of independent variables, and the final model brought together the best combination of independent variables from the previous specifications.

The first model examined the relationship of single-season fielding statistics with salary. While the F statistic indicates that the model explains more variance in the dependent variable than its expected (i.e., mean) value, the adjusted r-squared statistic (0.099) indicates that fielding statistics alone do not explain enough variation to be satisfactory. The number of put outs and the interaction term between put outs and errors were significant in this model (based on t tests) and had the expected sign.

The second examined salary as a function of single-season batting statistics. The adjusted r-squared statistic (0.35) suggests an improvement over the single-season fielding model. Batting average, RBIs, and bases on balls all were statistically significant and had the correct sign, but runs and home runs were not significant. Multi-collinearity is not present in this model, but error variance appears to increase with increasing values of the predicted dependent variable, a signal of heteroscedasticity.

The third model examined salary as a function of career batting statistics. This model improved yet again over the single-season batting average model, with an adjusted r-squared of 0.49. This appears to be due to the relationship between career batting average and salary, an effect we had seen previously in the exploratory phase of the project. Other variables were not statistically significant (based on t statistics); an interaction variable between career batting average and years in the major leagues was not significant and yet was also highly collinear with career batting average. This is surprising given our expectation that players with longer histories would continue to be more highly paid than other players, whether due to continued excellent performance or to free agency. However, it appears career batting average explains the majority of the variation in salary.

The fourth model brings together the best parts of each previous individual model. Model fit improved again (adjusted r -squared = 0.55) and several variables had a significant relationship (and correct sign) with the salary variable. Other variables with non-significant t statistics were left in the model based on theoretical grounds. While residuals in this model were fairly normally distributed, heteroscedasticity continues to be a problem. Several transforms and re-weighting of independent variables were examined with inconclusive results.

Furthermore, the interpretation of the parameter estimates still left us feeling like this model did not perform well based on theoretical assumptions about the components of player salaries. It is possible that salary negotiations and unavailable data might capture more variance and develop a more satisfactory model. However, another modeling technique might reduce the natural covariance among the large set of independent variables into a smaller, more manageable set of underlying constructs. For this reason, we pursued a principal components analysis, described in the subsequent section.

Multivariate analysis & visualization

Multivariate analyses are appropriate when an analyst is faced with a sizable set of measurable variables that capture aspects underlying constructs (the basis of which is theoretical or conceptual). Approaches like principal components and factor analysis are sometimes referred to as “data reduction” techniques, which is technically accurate but perhaps an oversimplification of what actually happens in the course of applying these approaches. SAS/Insight includes a “Multivariate” analysis feature that includes multi-way scatter plots, confidence interval ellipses and principal component analysis. However, SAS/Insight can also aid in the visualization of output from other multivariate procedures, as we shall see in our continuing baseball analysis.

Principal components

Principal components analysis reduces the covariance among a set of n measurable variables into one or more linear equations, each of which

explains the largest possible amount of this covariance in n -dimensional space. Principal components by definition are uncorrelated, such that the first principal component explains the most covariance in the data matrix, then the second explains the most covariance of what is left subject to being uncorrelated with the first component, and so on. There are as many principal components as variables, but the approach (well-applied) allows the analyst to explain a substantial degree of the total covariance with a smaller number of components than the number of original variables.

An advantage of principal components over factor analysis is that the system of equations is solvable—there is no need for estimation, and therefore results can be applied back to the original observations to assist in prediction. However, the notion that all principal components are uncorrelated leaves some analysts uneasy—they reason that most sets of real-world variables, such as baseball statistics, must be correlated in some way, and the orthogonal nature of principal components unnecessarily simplifies the nature of real things.

We used SAS/Insight’s multivariate features to build principal components from single-season batting, single-season fielding, and career batting statistics. An examination of eigenvalues of the correlation matrix suggests four principal components explain a larger proportion of the correlation among these variables than the variables themselves (i.e., the first four eigenvalues are all greater than one). However, the pattern matrix for the first four principal components fails to deliver an easily interpretable profile, what statisticians call “simple structure.” Simple structure means that each component exhibits a cluster of high loadings on a set of variables and low loadings on the remaining variables. It also means that no one variable has high loadings on more than one principal component. This analysis fails the test for simple structure, and we conclude that principal components may not be an appropriate approach for this data. Instead we pursue factor analysis.

Factor analysis

Factor analysis is not as straightforward as principal components analysis, but will permit

(through a technique called oblique rotation) each underlying construct to be correlated. Given our earlier analyses in the baseball data, it seems intuitive that great athletes will have both good fielding statistics and good batting statistics. We ran a factor analysis on the same set of variables as in the principal components analysis (for the statisticians out there, we used iterated principal factor extraction with promax rotation on a correlation matrix with communalities on the diagonal estimated via the squared multiple correlation method). We also used the factor analysis results to create a set of output observations (appended to the original input dataset) that places each observation in the p-dimensional space (given p factors retained by the analysis).

The results of this factor analysis again suggest that four factors should be retained, but this time simple structure (following oblique rotation) suggests clearly that the first factor loads most strongly on career batting counts (at bats, hits, runs, home runs, RBIs and bases on balls). The second factor loads on single-season batting counts, the third on career and single-season batting averages, and the fourth on fielding statistics.

Using a SAS/Insight rotating three-dimensional plot to visualize the results of the factor analysis, we see a large contingent of players with relatively low career batting counts, mostly due to their short careers to date. The factor scores for each observation can be used in a regression to capture the variance in salaries explained by these underlying constructs of player performance. It turns out that the fielding construct has little to do with salary variations, and that all three other factors (career batting, single-season batting, and batting averages) explain a significant portion of salary variance. The regression model with only the first three factors as dependent variables has an adjusted r-squared of .54, all parameters are statistically significant with correct sign. While some heteroscedasticity is still a problem, multicollinearity is not a problem (which we know from the inter-factor correlation table, where correlations between factors were around 0.15).

Finally, we can select players with strong batting statistics who are paid less than their peers by comparing a salary histogram with scatter plots

of the first four estimated factor scores. We can brush over the salary histogram and see that players with higher salaries score in the upper portions of the factor score distributions. However, we can also select players who are in the upper portions of the first two factor score distributions to find players who currently are making salaries in the lower tiers of the salary distribution. It stands to reason these are the top players for recruitment, since they are relatively low-paid but have excellent performance on the basis of the first two factors. We can be fairly certain these players will make a strong contribution to the team yet are under-appreciated in terms of league value. This list of players can be quickly extracted to a new dataset and distributed to field scouts and the rest of the franchise management.

Conclusions

This paper has explored statistical visualization and modeling capabilities of SAS/Insight software in the context of data mining. SAS/Insight provides a simple and inexpensive option for firms and organizations to get accustomed to data mining endeavors. Further functionality and power is available in the SAS Enterprise Miner package, soon to be released. However, smaller problems that do not require this power may find SAS/Insight to be a fully satisfactory platform for statistical visualization and modeling.