

EXPLORATORY DATA ANALYSIS: GETTING TO KNOW YOUR DATA

Michael A. Walega
Covance, Inc.

INTRODUCTION

In broad terms, Exploratory Data Analysis (EDA) can be defined as the numerical and graphical examination of data characteristics and relationships before formal, rigorous statistical analyses are applied. Although the temptation to omit EDA in favor of delving into ANOVAs, MANOVAs and the like is great, the role of EDA cannot be underemphasized. EDA rewards the user with a better understanding of the intricacies of the relationships among data. In addition, EDA can be used to probe the validity of assumptions that are made by formal statistical tests (eg., normally distributed data/residuals, homogeneity of variances, additivity, etc.) EDA can be envisioned as the setup for more formal analyses. It provides all the prerequisites for a final, no-surprises formal analysis of the data.

As stated above, both numerical and graphical methods are employed in EDA. For numeric methods, one might begin with descriptive statistics (mean, median) and progress to more formal analyses of the relationships between data (regression, correlation). For graphical methods, one can advance from frequency distributions to simple 2-way scatter plots to n-way scatter plots to response-surface plots. A combination of numeric and graphical methods leads to normal probability (quantile) plots, histogram/distribution plots and beyond.

At this time, the usual warning must be imparted: It is up to the reader to determine the appropriate EDA methods for the data. Ignorance is not bliss!

EDA: UNIVARIATE METHODS

This section of the paper will emphasize some of the numeric and graphical methods that could be employed in EDA of a single variable. The methods described here are not meant to be all-inclusive; rather, they are intended to provide a starting point for further in-depth analyses.

Numerical Methods

Regardless of the number of variables, the most informative SAS® PROC for any numeric analysis (and the best starting point) is UNIVARIATE. This procedure contains a wealth of information that can be utilized in EDA. The procedure provides statistics that detail the central tendency, spread, extremes and distributional characteristics of the data. **Figure 1** presents an analysis of hits for American League baseball players from 1988. The code used to generate this output is shown below:

```
PROC UNIVARIATE DATA=BASEBALL PLOT NORMAL;  
  
    ID NAME;  
    VAR HITS;  
RUN;
```

Examination of the central tendency statistics (mean and median) and the spread statistics (standard deviation and interquartile range, Q3-Q1) can shed some light on the distribution of the data. The extremes can be used to identify potential outliers. Incorporation of the ID *varname* statement in PROC UNIVARIATE results in the five high and low extremes being identified by the values of *varname*. The statistics skewness and kurtosis compare the shape of the distribution of the data to that of normally distributed data. It should be noted that these statistics have been shown to be heavily influenced by sample size and outliers. Other shape statistics have been suggested (Walega, 1993).

The addition of the NORMAL option to the PROC UNIVARIATE statement generates a statistical test for normality. The statistic employed is the Shapiro-Wilk W for sample sizes ≤ 2000 , otherwise the Kolmogorov statistic is used. It is suggested that tests for normality be conducted conservatively, with significance levels of 0.10 or 0.15 used to reject the hypothesis of normally distributed data. In addition, care must be used in the interpretation of these tests, as they are sensitive to both sample size and to the presence of outliers.

Graphical Methods

As with numeric methods, graphs developed for univariate data usually identify the central tendency, spread and distribution of the data. Methods that can be used to elucidate this information include histograms and stem-and-leaf, box and empirical distribution plots. Each of these will be discussed in further detail below.

Sample histograms can be produced by PROC CHART, while more complex, high-resolution histograms can be generated by the SAS/GRAPH® GCHART procedure. Histograms are useful for displaying the distribution of data within selected intervals. They are often used to provide a preliminary estimate of the fit of the data to a normal distribution. The key to using histograms is the selection of the number of midpoints. Too few, and it is difficult to extract information about the distribution of the data. Too many, and the overall picture of the distribution becomes clouded. PROC GCHART can be used to enhance histograms through the addition of labels, specialized counting statistics, and other display

enhancing options, although a simple chart is usually sufficient.

Stem-and-leaf displays (see **Figure 2**) are similar to histograms in that both divide the data into intervals, thus providing a frequency distribution of the data. The stem-and-leaf plot goes on to provide the actual data values in the display while showing the distribution of the data. The PROC UNIVARIATE statement, in conjunction with the PLOT option, will generate stem-and-leaf plots. Although not as widely employed as a histogram, the stem-and-leaf plot is useful because: Actual data are shown, the stem-and-leaf plot is generated along with the remainder of the UNIVARIATE output and YOU DON'T HAVE TO DO ANY PROGRAMMING TO GENERATE IT!

An extremely effective method for presenting a summary of univariate data is the box plot (see **Figure 2**). As described by Tukey (1977), the box plot is a practical means for the display of the following summary statistics: central tendency (mean, median), spread (interquartile range), shape and outliers. The addition of the PLOT option to the PROC UNIVARIATE statement will generate box plots for each analysis. If a BY *by-variables* statement is included, the box plots for each level of the *by-variables* will be grouped on a separate page of output. High-resolution box plots can be generated by specifying I = BOX in the SYMBOL statement of PROC GPLOT. The stem-and-leaf plot and the box plot, when used in conjunction with the numeric analyses provided by UNIVARIATE, create a powerful analytical tool that can be used to further understand the distribution of data from a single variable.

Data Transformations

Close examination of the data may indicate the need for transformation to achieve the gold standard of normally distributed data. How to effect this transformation has been greatly simplified by the development of specific graphical methods that suggest to the analyst the appropriate power transformation.

As stated earlier, many formal statistical analyses rely on data that approximately follow a normal distribution, specifically the residuals (errors). While there are many methods that can be used to effect a transformation of the data to an approximately normally distributed form, the family of power transformations have been the most studied, and thus the best understood.

Unlike other data transformations, a power transformation of the data does not impact on the order of the data. Rather, it changes the spacing between the data. For example, $\text{SQRT}(x)$ and $\text{LOG}(x)$ pull in the upper tail of a distribution, with $\text{LOG}(x)$ being more powerful than $\text{SQRT}(x)$. The transformation X^n , where n is usually a multiple of 0.5, spaces out the upper tail. But how does one determine the 'best' power transformation?

A family of plots called symmetry plots have been shown to be very effective in providing the analyst with an assessment of the distribution of data, information as to the 'best' transformation to employ, and the ability to

assess the effect of the transformation. The plots are based on equivalent distances between corresponding points to the median value. When the data are appropriately plotted, a linear regression line with a slope of β indicates that the power for transformation to symmetry is $1 - \beta$. For example, if the slope is 0.5, then the power transformation is also 0.5, which suggests the square root of the data will effectively transform the data such that they follow approximately normal distribution. **Figure 3** shows a symmetry plot for home runs. It suggests that a square root transformation of the data will be the most appropriate.

BIVARIATE/MULTIVARIATE METHODS

Most analyses involve measuring the effect that two or more variables (independent) have on one or more dependent variable. In some cases, extensions to the methodology presented in the previous section can be employed. However, in most cases more sophisticated tools are required to effectively perform EDA.

Graphical Methods

Graphical EDA of bivariate data is just an extension of the univariate methodology presented above. Actually, the graphical display of bivariate data is probably the most frequently used aspect of EDA, and can be significantly more useful than the numeric methods to be described below.

Of the methods available, simple and enhanced scatterplots are the cornerstone of displaying data from two to many variables. Indeed, many multivariate methods for graphing data expand on the simple scatterplot. Not only can bivariate displays be used for preliminary EDA, but they are also employed as tools for regression diagnostics and are used to validate model assumptions.

The simplest way to display data from 2 variables is to use PROC PLOT to generate a scatterplot. In most instances the resolution of PROC PLOT is acceptable; PROC GPLOT does achieve higher resolution, at a cost of more programming. PROCs PLOT and GPLOT can also display the relationship between pairs of variables. Labeling observations is useful if one intends to identify outliers. PROC PLOT does have the ability to label data points; unfortunately, all data points are labeled. For this special case, the use of the Annotate facility of SAS/GRAPH to label data points on a PROC GPLOT scatterplot is recommended. This method is the best way for the user to set limits that define outlier data points.

Two extremely useful displays of bivariate data are: the marriage of the scatterplot and boxplot, and the marriage of scatterplots and confidence ellipses (Friendly, 1991). The scatterplot/boxplot display aids the user in the visualization of the shape of the distributions of the 2 variables, and provides means to estimate the summary statistics and detect the presence of outliers. The scatterplot/confidence ellipse provides information on the relationship between several groups of data. One must take care when using this display; outliers, non-normal data and non-linear relationships between the data may

distort the ellipses. In that case, a non-parametric version should be employed.

An extension of the scatterplot to multivariate analyses is the scatterplot matrix. This display enhances the ability to see the relationships between more than 2 variables. All pairs of n variables can be displayed in an $n \times n$ grid, with the $n(n-1)/2$ scatterplots shown on a single page of output. Obviously, resolution decreases as an n increases. Experience shows that $n > 6$ limits the effectiveness of this display. PROCs GPLOT and GREPLAY, in conjunction with the Annotate facility of SAS/GRAPH, are used to generate scatterplot matrices. **Figure 4** presents a scatterplot matrix for batting average, SQRT(hits), SQRT(hr), years in the major leagues (calculated as min(years,7)), SQRT(rbi) and LOG(salary).

Other multivariate plots (ie, star, profile), can also be used to display multivariate data. However, they require more programming, and the results can be more difficult to interpret. Two very useful 'multivariate' plots are the multivariate normal probability plot and the multivariate outlier plot. Although both plots collapse multi-dimensional data down to 2 dimensions, both can be extremely useful. **Figures 5** and **6** display examples of a multivariate normal probability plot and outlier plot, respectively.

As we approach the stage of building a model(s) to explain the relationships that we have observed, once again it is graphics to the rescue. A straightforward plot called the Cp plot can be used to assist the analyst in identifying models of interest that can then be formally evaluated using PROC REG.

The Cp plot relies on output from PROC RSQUARE to generate Cp values for all possible combinations of variables that entered into a model. For example, if the variables batting average, SQRT(hits), SQRT(hr) and SQRT(rbi) were of interest, RSQUARE would generate 14 different models, using any and all combinations from one to four variables. The values of Cp are plotted against 1 plus the number of predictors in the model. Models judged to have high potential have low values of Cp. **Figure 7** presents an example where batting average, years experience, SQRT(hits), SQRT(hr) and SQRT(rbi) are used to determine LOG(salary). In this example, the models with the highest potential appear below the line.

Numerical Methods

Examination of the linear relationship between two (or more) variable requires the use of PROCs CORR and REG. PROC CORR starts a preliminary investigation of the strength of the linear relationship between two variables. PROC REG allows us to delve deeper into the possible linear relationships between two or more variables, and thus build models that can help explain these linear relationships.

PROC CORR provides parametric (Pearson) and nonparametric (Spearman) measures that calculate the

correlation or partial correlation between variables. The code

```
PROC CORR DATA=BASEBALL PEARSON SPEARMAN  
NOSIMPLE;
```

```
VAR AVG SQHITS SQHR SQRBI LOGSAL;  
RUN;
```

generates a correlation analysis that examines the linear relationship between the batting average, hits, home runs and runs batted in. One must carefully interpret the results of any correlation analyses. It is suggested that graphical methods also be employed, as lack of correlation indicates only a lack of a linear relationship between two variables, not a lack of any relationship.

Once preliminary EDA analyses have been completed, PROC REG can then be used to further examine the relationships among the data, and to provide an estimate of the value of a dependent variable given the predicted value(s) of the independent variable(s). PROC REG is similar to UNIVARIATE in that the output can contain a wealth of information about the linear relationship among data (the only problem is to determine what you really need!) Unlike UNIVARIATE, use of the more advanced options of the procedure necessitates some understanding of statistics.

With PROC REG, various models can be developed that may provide further insights into the relationships among data. The MODEL statement options FORWARD, BACKWARD, and STEPWISE enter/delete variables into/from the model, while the options MAXR, MINR, RSQUARE and ADJR SQ employ the correlation coefficients to determine the "best" model. Then, after a model has been fit, PROC REG can provide diagnostic information on the fit of the data to the model and information on the reliability of the assumptions made by linear regression analyses.

DIAGNOSTICS

After a model for the data has been proposed, it is always appropriate to test that the assumptions made by the analytical method have been satisfied. For example, one basic tenant of regression and general linear models analyses is that the error terms are normally distributed with mean 0 and variance σ^2 . PROC REG, in addition to providing the results of the proposed model, also provides diagnostics in the form of statistical tests and plots that can be used to validate the assumptions made by linear regression. These assumptions can also be examined through the use of PROC UNIVARIATE and SAS/GRAPH.

Graphical Methods

Within REG, many types of diagnostic plots are available. A plot of the studentized residuals versus the predicted values help the analyst determine if the variances are homoscedastic (approximately equal). A more meaningful plot is the absolute value of the rstudent

residuals versus the predicted values, as the student residuals are, by their nature, not influenced by outliers.

An effective display that is used to examine the distribution of data (or residuals) is the quantile plot. The NORMAL option on the PROC UNIVARIATE statement (see **Figure 2**) produces such a plot. This plot provided information about the distribution of the data (or residuals) as compared to a normal distribution. Deviations from a straight line are suggestive of non-normality. For example, evidence of skewness (both ends of plot deflect in same direction) or kurtosis (ends deflect in opposite directions) can be seen in these plots. Additionally, this plot permits the user to detect potential outliers as well as systematic departures from normality. The normal quantile plot produced by UNIVARIATE is, however limited in the detail that it can provide. A better alternative would be to program the display via the GPLOT procedure of SAS/GRAPH.

Numeric Methods

The COLLIN option in PROC REG can be specified to test for collinearity (one regression variable being a linear combination of other variables in the model). Note that evaluation of collinearity is purely subjective, in that one looks for a "large" increase in the Condition Index.

The SPEC option of PROC REG can be used to test for heteroscedasticity of the variance (errors not independent, variances not constant). The results of the analysis should be examined at a conservative p-value, say from 0.10 to 0.20.

SUMMARY

This paper has attempted to give the reader a flavor for the many Exploratory Data Analysis tools that are available in Base SAS, SAS/GRAPH and SAS/STAT®. Realizing that this paper has skimmed the surface of EDA methodology available today, hopefully the ideas discussed in this paper will stimulate the reader to further explore their data.

CONTACT INFORMATION

The author can be reached at:

Covance, Inc.
210 Carnegie Center
Princeton, NJ 08540
Phone: (609) 452-4150
Email: Michael.Walega@Covance.com

REFERENCES

- Friendly, M. *SAS System for Statistical Graphics, First Edition*. SAS Institute, Inc., Cary, NC (1991).
- Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley Publishing Co., Inc., Reading, MA (1977).

Walega, M. A. *Use of PROC IML to Calculate L-Moments for the Univariate Distributional Parameters Skewness and Kurtosis*. Proceedings of the Sixth NorthEast SAS Users Group Meeting, 1993, p. 573.

SAS, SAS/GRAPH and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries.

FIGURE 1

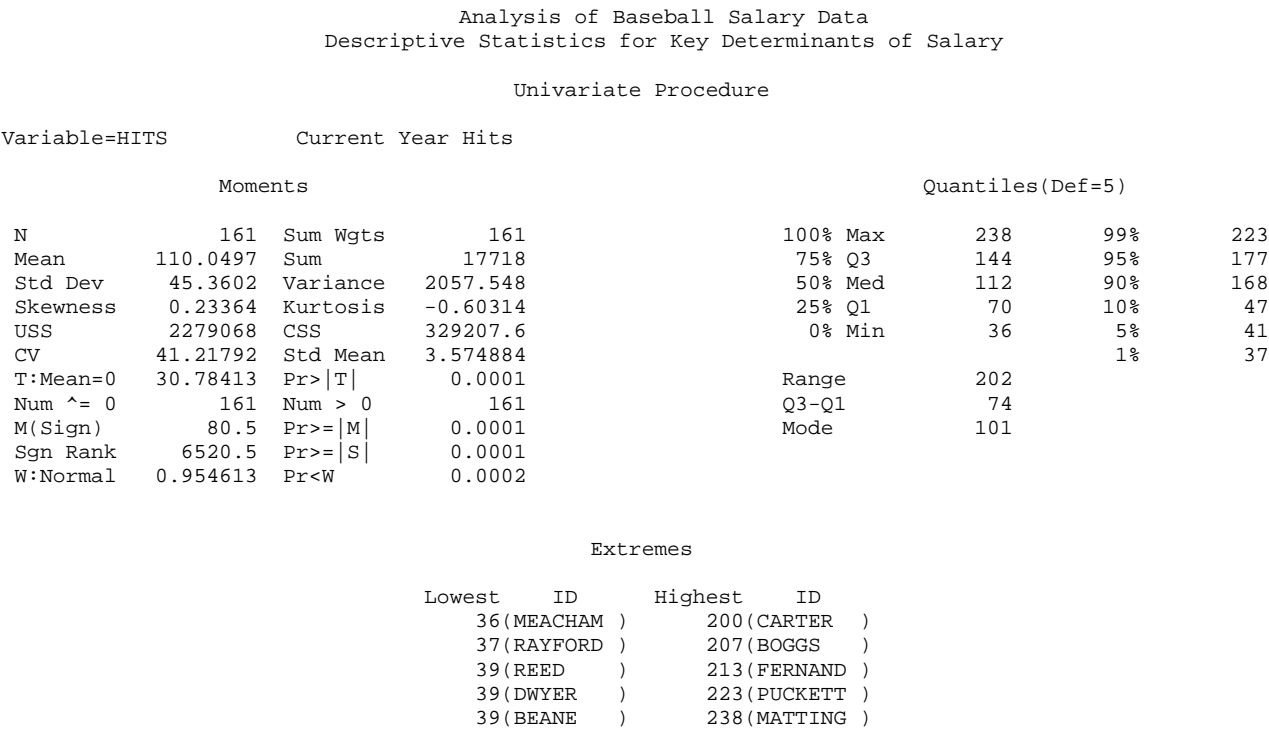


FIGURE 2

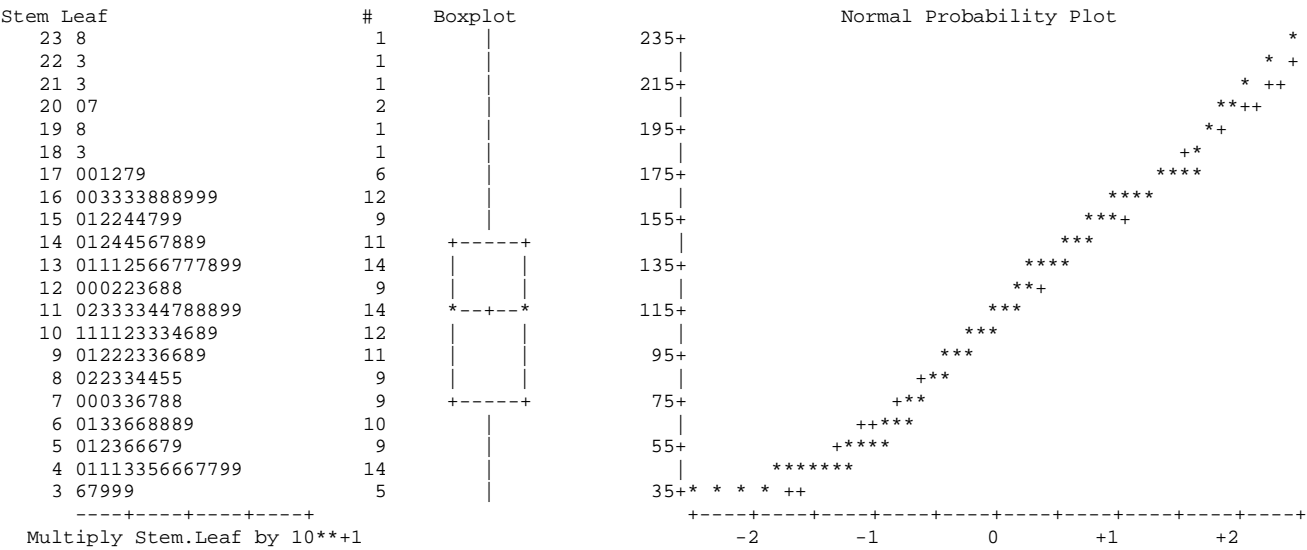


FIGURE 3
Baseball Salary Data Analysis
Symmetry Transformation Plot: hr

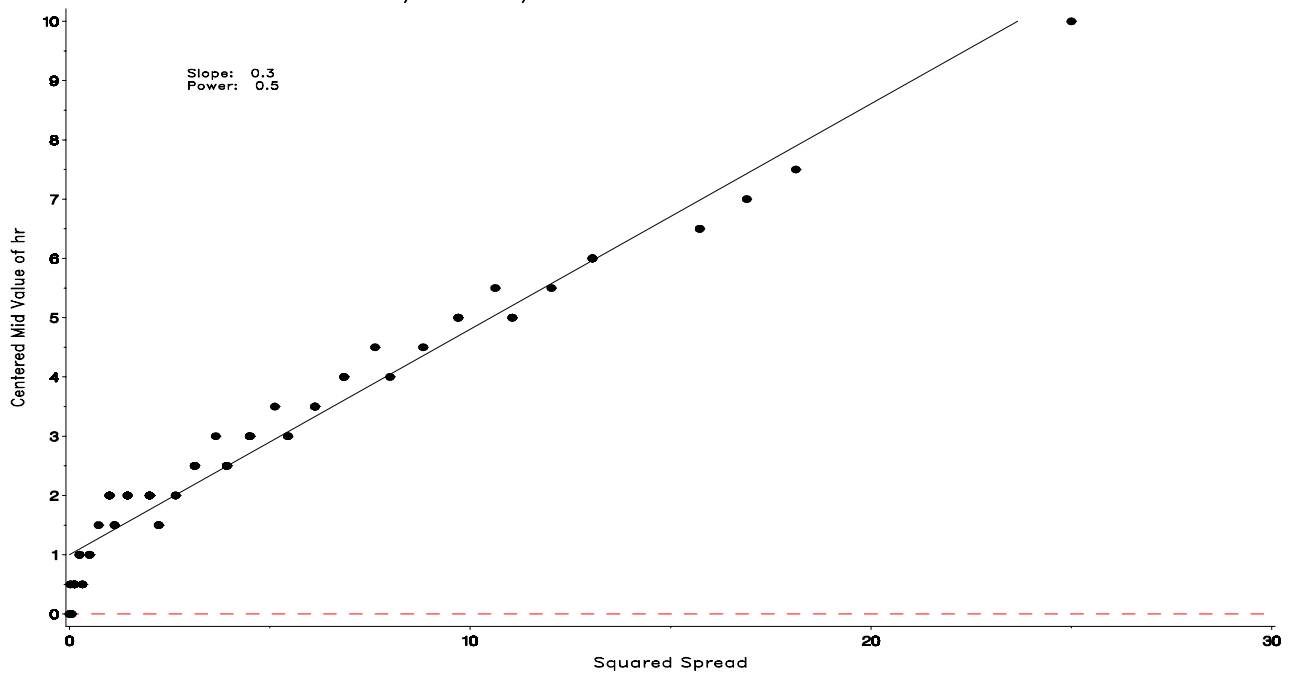


FIGURE 4

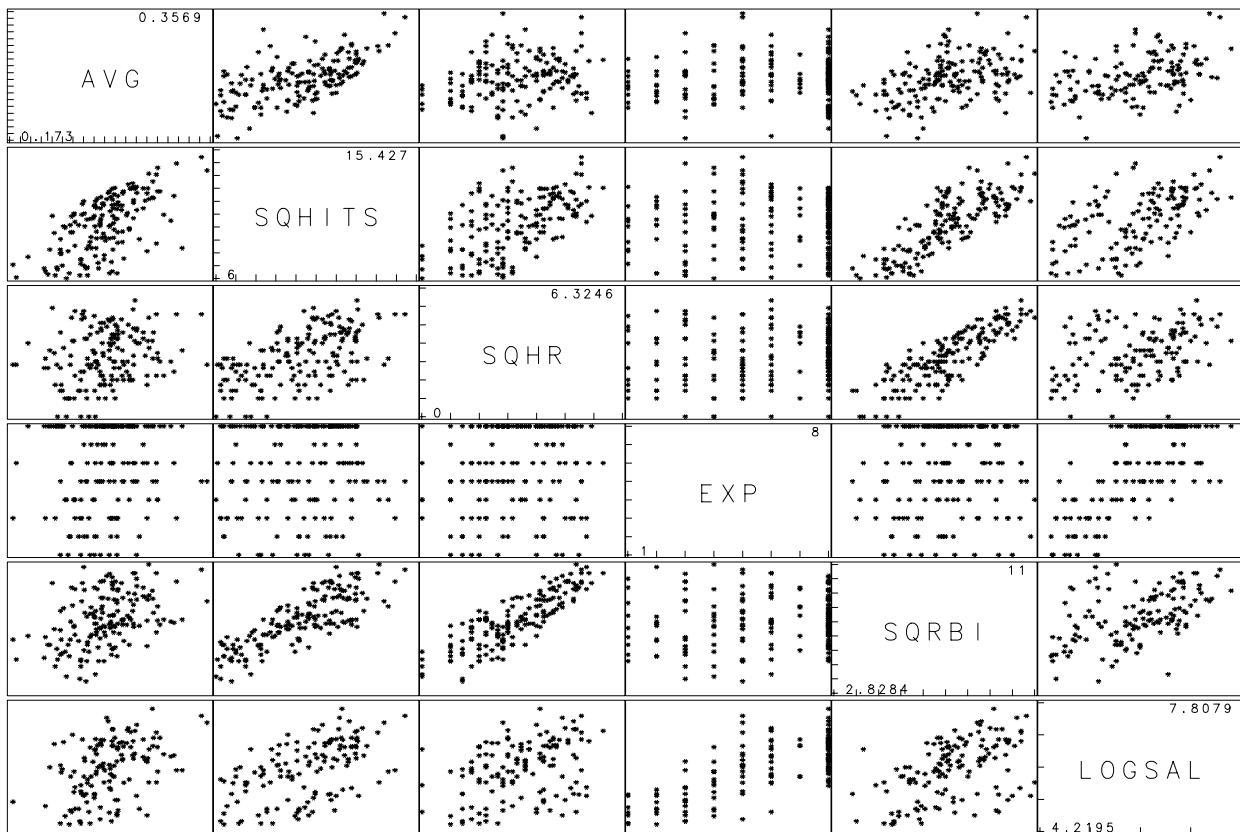


FIGURE 5

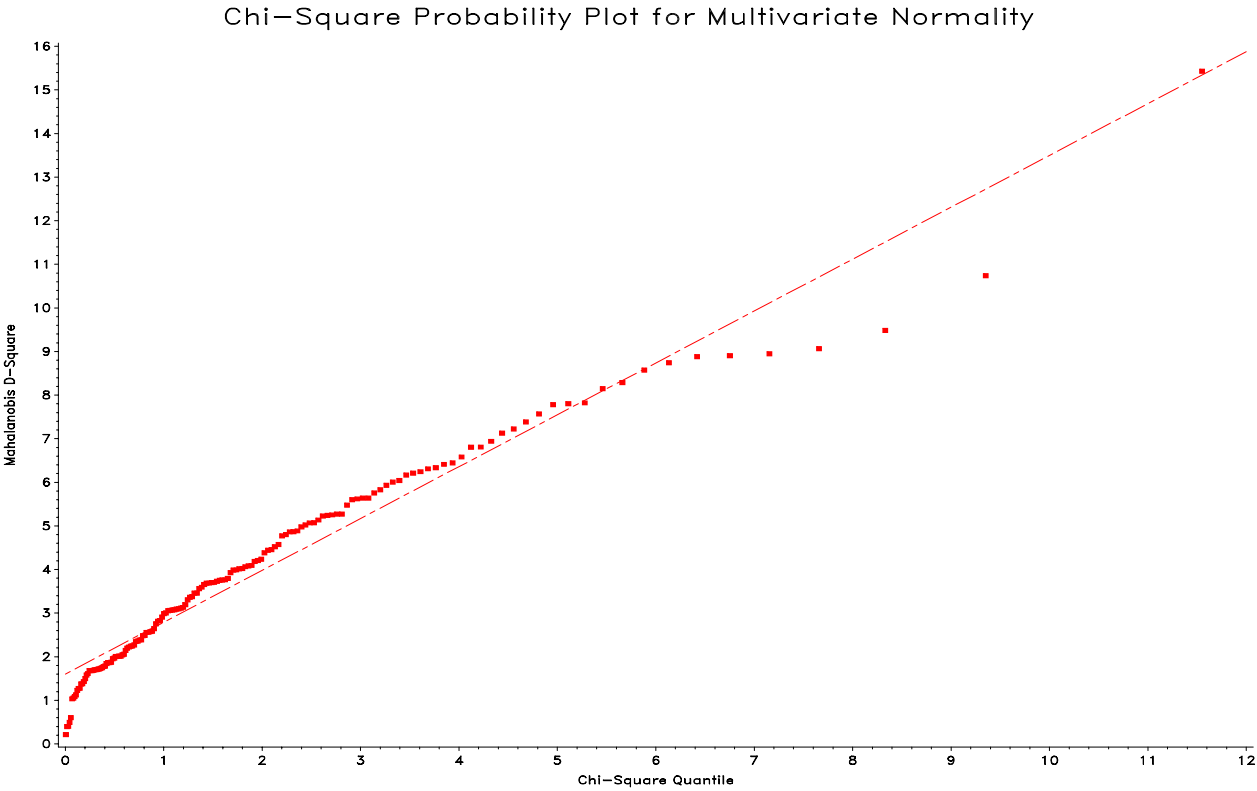


FIGURE 6

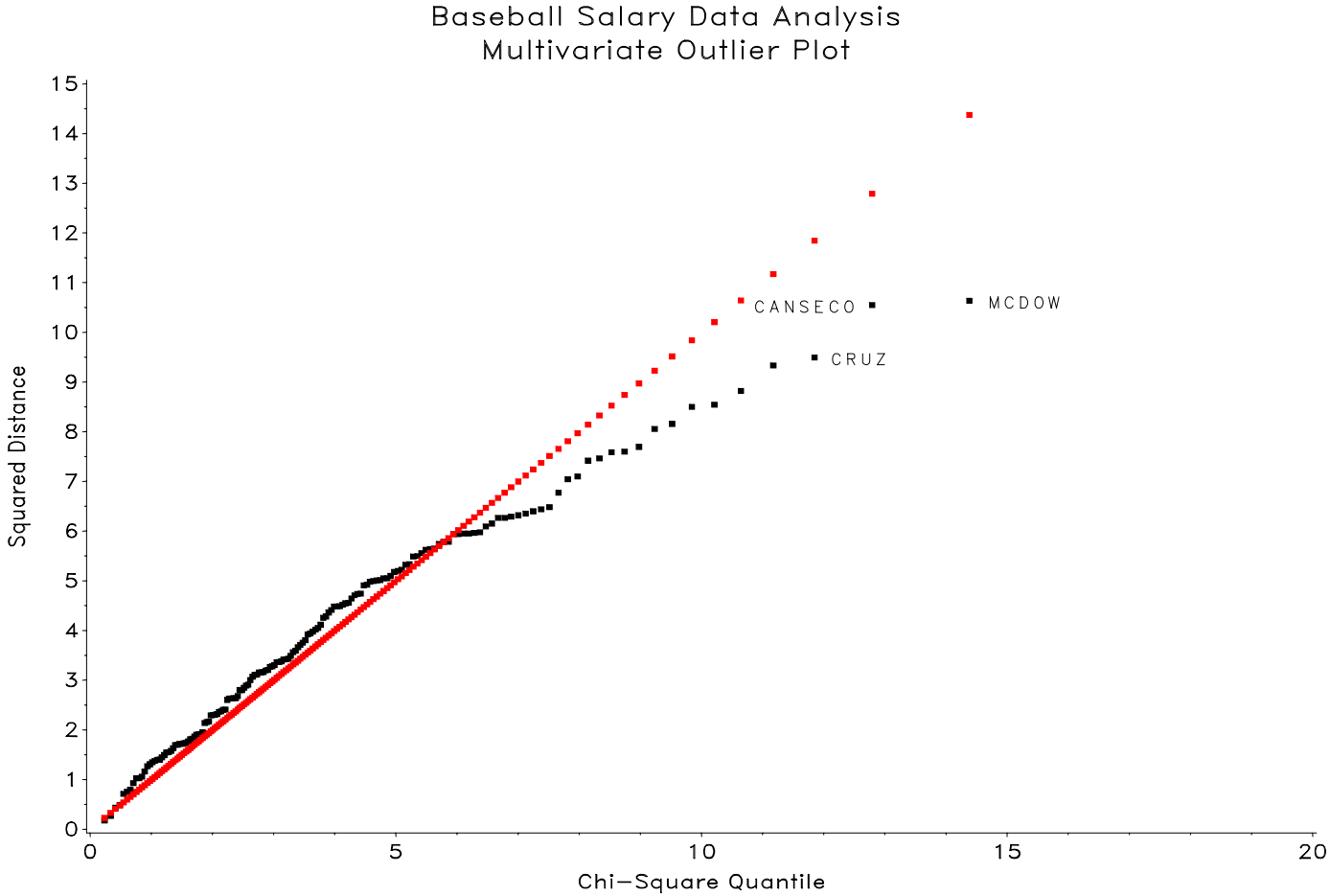


FIGURE 7

Baseball Salary Data Analysis
Cp Plot for the Predictor Variables (avg exp hitssq sqhr rbisq)

