DOCUMENT RESUME

ED 466 781 TM 034 292

AUTHOR Odom, Leslie R.; Henson, Robin K.

TITLE Data Screening: Essential Techniques for Data Review and Preparation.

PUB DATE 2002-02-15

NOTE 37p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Data Analysis; Data Collection; Heuristics; *Regression (Statistics)

IDENTIFIERS Outliers; *Screening Procedures

ABSTRACT

Prior to conducting a statistical analysis, sufficient data screening methods should be used for all research variables to identify miscoded, missing, or otherwise messy data. The primary purpose of these exercises was to demonstrate the role of data screening techniques and their potential to improve the performance of statistical methods. A heuristic data set was used to make the discussion more concrete, and the Statistical Package for the Social Sciences (SPSS) was used to screen the data. Overall, cleaning raw data by determining normality and linearity problems, outlier influences, and missing value presence proved to increase the R squared values if only by very small increments. One of the most interesting findings in this exercise was the performance of the regression models when outliers were taken into consideration, without respect to any additional data cleaning procedures. These screening procedures, if used properly, assist the researcher in optimizing data so that the analysis procedure will produce the most accurate and efficient estimates. Two appendixes contain SPSS syntax for the analyses. (Contains 1 table, 9 figures, and 15 references.) (Author/SLD)

Running head: DATA SCREENING: ESSENTIAL TECHNIQUES

Data Screening: Essential Techniques for Data Review and Preparation

Leslie R. Odom

Robin K. Henson

University of North Texas

Paper presented at the annual meeting of the Southwest Educational Research Association, February 15, 2002, Austin, Texas. Correspondence concerning this manuscript may be sent to the senior author at lro0001@unt.edu.

Abstract

Prior to conducting a statistical analysis, sufficient data screening methods should be employed for all research variables to identify miscoded, missing, or otherwise messy data. The primary purpose of these exercises was to demonstrate the role of data screening techniques and their potential to improve the performance of statistical methods. Overall, cleaning raw data by determining normality and linearity problems, outlier influences, and missing value presence proved to increase the R^2 values, if only by very small increments. One of the most interesting findings in this exercise was the performance of the regression models when outliers were taken into consideration, irrespective of any additional data cleaning procedures. These screening procedures, if used properly, assist the researcher in optimizing the data so that the analysis procedure will produce the most accurate and efficient estimates.

Data Screening: Essential Techniques for Data Review and Preparation

Beginning a statistical analysis without a careful inspection of the research data may result in erroneous findings and/or conclusions. Data screening methods provide the researcher with a means to detect potential data problems by identifying data entry errors, missing values, possible outliers, non-normal distributions, and other data features. The recent report of the American Psychological Association's (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999) expressed the importance of data screening to the research community. In this report, the Task Force included data screening as a vital step in statistical analysis and emphasized the reporting of unexpected data complications when presenting research findings. Particularly, the Task Force stated: "The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, nonrandom missing data, selection bias, attrition problems) should be a standard component of all analyses" (Wilkinson & Task Force on Statistical Inference, 1999, p. 599, emphasis added).

With the expectation that data screening methods become routine, how does a researcher determine the proper screening techniques? Tabachnick and Fidell (2001) suggested two factors for initial consideration. First, data screening techniques relate directly to the statistical method of choice and the assumptions of that method. For example, analyzing data using logistic regression (a log-linear method) will not require the same screening process as a multiple regression analysis (a linear method) due to the different assumptions that are required for each procedure. A second consideration for data screening procedures is data grouping. Group data, such as attitudinal differences based on an ethnic classification, may necessitate different screening procedures as compared to data that does not compare groups, such as the relationship between two attitudinal measures in a validity study.

Purpose

It is important to understand that certain data features can have tremendous impact on analysis results. Further, certain data features may impact some analyses but not others. The purpose of the present paper is to discuss and demonstrate several important screening issues. Because screening procedures are dependent on the analysis used (Tabachnick and Fidell, 2001), the paper will be confined to screening techniques applicable to multiple regression analysis for ungrouped data. Because multiple regression is the univarite general linear mode (Cohen, 1968), the techniques discussed herein are generalizable to many other correlational methods in the general linear model. Specifically, the paper will (a) review several of the most common issues that should be addressed in a multiple regression analysis, (b) detail the data irregularities that can be identified by using specific screening methods, and (c) demonstrate several methods and illustrate the effect of data irregularities on results. A heuristic data set was used to make the demonstration concrete; SPSS (v 11.0) was used to screen the data. The SPSS syntax for the analysis is provided in Appendix A and Appendix B for the reader's consideration. The present paper is intended to provide initial guidance on data screening procedures in accessible format for applied researchers.

Variable Coding and Outlier Detection

At a minimum, the research data should be checked to verify variable coding or data entry. Determining the range of values, central tendency measures, and simple frequency counts will identify cases that do not fit (or are not possible) with the other observations. For example, if math achievement scores on a measure ordinarily range from 0 to 100, then reported scores of 101, 102, and 104 appear anomalous. If the measure has maximum scores of 100, then these scores are obvious entry errors and should be corrected or deleted. If a wider range of scores is possible, then the extreme scores may be real scores, but they may not be probable and

considered outliers. If fact, with asymptotic distributions, we have the expectation that 5% of the values will exceed roughly two standard deviations above and below the mean – provided the variable simulates a normal distribution (Lewis-Beck, 1995).

What should be done, if anything, to these scores? Stevens (1996) provides several alternatives. If the data recording process is at fault, then the scores are simply recoded to their proper value. Should there be an instrumentation malfunction, then it is acceptable to remove the resulting outliers from the analysis. However, if these scores appear to be valid data points, some may argue that they need to be a part of the analysis. Reporting two analysis results, one conducted with the outliers and another without the outlier scores, provides a means to compare the effects of the outliers. As Stevens (1996) noted, "Outliers should not necessarily be regarded as 'bad'. As a matter of fact, it has been argued that outliers can provide some of the most interesting cases for further study" (p. 18).

Normality Issues: Statistical Tests and Verification

Distribution of error scores.

One important characteristic of a variable involves the distribution of the variable's scores and the approximation of those scores to a normal distribution. However, the effect of a non-normal variable distribution on a multiple regression analysis is frequently misunderstood by many researchers. According to Allison (1999),

> Many people think that all the variables in a regression equation must be normally distributed. Nothing could be further from the truth. The only variable that is assumed to have a normal distribution is the disturbance [error] term U, which is something we can't observe directly. The x variables can have any kind of distribution. Because y is a linear function of both the x's and U, there's no requirement that y be normally distributed either. (p. 130)

The general approach to teaching this topic often does not involve a proper discussion of the distribution of the error term. In regression, as in all general linear model analyses, the focus of the analysis is actually on the unobserved synthetic variables that are created from the observed variables. The synthetic variables for regression include $\hat{Y}$ (the predicted variable) and e (the error scores). The regression normality assumption states that the error scores should be normally distributed and does not assume such for the observed variables directly.

Often, however, an emphasis is placed on the dependent variable (and the continuous predictors) meeting the normality assumption, with a hoped-for by-product being that the resulting error term will be normally distributed - a somewhat backwards approach. Nevertheless, if the continuous predictors are normal, the likelihood that the error scores are normal increases, although this is not a certainty. Therefore, normality of the observed variables is sometimes used as a proxy for normality of the error term.

It is important to note that if the sample size is fairly large (200 or more cases), then the question of normality is less relevant because the central limit theorem states that the sample statistics will be good estimates even if the error term is not normally distributed. However, should the sample size drop below 100 cases, the residuals will need to be examined for deviations from a normal distribution (Allison, 1999). Such smaller samples are closer to the norm rather than the exception in applied research.

Skewness and kurtosis.

Two statistics that describe distributional shape are skewness and kurtosis. The skewed nature of a variable relates to the clustering of data points at one end of the distribution. Positively skewed variables will have observations piled towards the lower values with a moderate number of cases trailing-off in the direction of the higher values. Negatively skewed variables exhibit grouping toward the higher values and cases trail-off toward the lower end.

The kurtosis statistic is sometimes described as a measure of the peakedness of a distribution curve (Everitt, 1998), but Hopkins and Weeks (1990) further clarified this indicator: "The kurtosis index . . . reflects the extent to which the density of observations differs from the probability densities of the normal curve (Kotz & Johnson, 1982) (p. 102)". A concentrated group of observations around the mean of a variable, which is higher than the density of observations within a normal distribution, results in a highly peaked (leptokurtic) distribution. A lower concentration of values about the mean results in a lower peaked (platykurtic) distribution. Because kurtosis concerns itself with the density of observations under the curve as they relate to a normal distribution, it stands to reason that many shapes may have similar densities relative to the whole distribution. As Henson (1999) explained:

> Consider your evaluation of the shape of a man weighing 170 pounds. Your evaluation of the "normality" of this man's "shape" will largely depend on how tall he is. It could be said that the man should be about six feet tall to be normal, speaking purely in heuristic terms, of course! As the 170-pound man's height decreases, his relative width increases and his shape begins to depart from normality. Similarly, as his height increases, his relative width decreases resulting in a very tall but extraordinarily thin person weighing 170 pounds. (p. 196)

Kurtosis measures a similar dynamic concerning relative height to width in distributions, and because there are many appropriate relative proportions (i.e., density of scores under the curve), "there are infinitely many different distribution shapes that may be normal" (Henson, 1999, p. 196, emphasis in original).

Assessing the normality assumption for a variable is best accomplished through a combination of statistical and graphical procedures. Again, the goal is the normal distribution of the error term, and not necessarily the independent or dependent variables. But review of the

observed variable distributions is presented later with the tentative assumption that normally distributed observed variables result in a normally distributed error term.

For a completely normal distribution, the skewness and kurtosis values are zero. The more a distribution deviates from normal, the farther from zero the statistic travels. Speaking in terms of statistical significance testing, the ratio of the skewness or kurtosis statistic to its standard error (parameter divided by standard error) should not be less than negative two or greater than positive two (SPSS, 1999). However, like all statistical significance tests, the sample size can have an impact on the standard errors of these statistics, so it is possible that the normality assumption would be rejected when the case distribution is actually closer to normal than the statistics indicate (due to excessive power). Comparing the central tendency measures in addition to the review of the skewness and kurtosis statistics provide a second test for the verification of normality. As Lewis-Beck (1995) noted, "If a variable is normal, then the mean, median, and the mode are equal" (p. 15). Reviewing this information in conjunction with a histogram of the variable will provide information regarding the case distribution of the variable as compared to the expected case distribution of a normal curve.

Data Transformations

When a variable does not meet the normality assumption, transformation procedures may be applied to the data so that the observations more closely approximate a normal distribution. Several of the most common distributions that may benefit from transformation procedures are presented in Figure 1. The degree and direction of a distribution's departure from normality helps to determine the initial transformation that may create a more normal distribution. One group of transformations, the family of powers and roots, is especially useful. Square root transformations can correct for mild positively skewed distributions. Log transformations are utilized for more severe positively skewed distributions. These

transformations, however, have certain requirements. First, they are most effective when the ratio of the largest to the smallest value is very large. Second, the order of the original values (not the distance between them) is retained only if the values are positive. If these conditions are not met, then a positive or negative constant (a start value) added to each value will "correct" the data so that all values have the same sign prior to transformation (Fox, 1997).

Insert Figure 1 about here

Mild or severe negatively skewed distributions are treated in the same manner, except the variables are first reflected. A variable is reflected by identifying the largest value of the distribution and then adding one to it, thus forming a constant. A new variable is created as the original values are subtracted from the constant. Care should be taken regarding the interpretation of these new variables. Since the reflection reversed the direction of the distribution's skew, ". . .if big numbers meant good things prior to reflecting the variable, big numbers mean bad things afterward" (Tabachnick and Fidell, 2001, p. 81).

The decision to transform data points is highly dependent on variable interpretation. Taking the logarithm or square root of an achievement variable is more difficult to interpret than the original score. Other variables with subjective scales may have less interpretation difficulties after transformations are applied. Transforming data must also be performed within the context of the original data. Data that exhibits mild deviations from normality or a series of variables, all of which are non-normal to the same extent, will not be dramatically improved with transformations. Data transformation is not an exact science. Different transformations may need to be applied to a variable and then the results compared to one another in order to determine which method yielded the best distribution correction.

determine which method yielded the best distribution correction.

Linearity

Assessing the linear relationship among the variables in a multiple regression analysis is important because a nonlinear relationship ". . .implies that the [regression] model fails to capture the systematic pattern relationship between the dependent and independent variables" (Fox, 1991, p. 54). A general determination of variable relationship and linearity is accomplished by reviewing scatterplots. If two variables are approximately normal and have a linear relationship, the bivariate scatterplot is cigar or oval-shaped. A nonlinear relationship results in a scatterplot characterized by one or more dips and/or bulges in the representation or a circular pattern. Although multiple regression can be used with non-linear data, it will not provide an accurate illustration of the relationship between the independent and dependent variables. If preliminary analysis indicates the presence of a highly nonlinear relationship, another statistical method should be used.

Nonlinear relationships can be made more linear by power transformations as previously discussed. This is because when one changes the shape of a variable, you also will affect its relationship with other variables. However, the application of Tukey and Mosteller's (1977) "Bulging Rule" (see Figure 2) provides assistance in selecting which transformation(s) will be better suited for the direction and degree of nonlinearity. This bulging rule is easily applied because a comparison is made between the nonlinear relationship depicted in the scatterplot and the graphical representation of the bulging rule principal. The direction of the scatterplot bulge determines if the X-variable, the Y-variable, or both variables should be transformed (Fox, 1997). Post-analysis screening using plots between standardized residuals and standardized

predicted values will identify unresolved nonlinear relationships as some residuals will fall well above and well below the zero line for certain predicted values (Tabatchnick & Fidell, 2001).

Insert Figure 2 about here

Homoscedasticity

A term derived from Latin meaning "same variance", the homoscedasticity assumption of regression refers to the fact that the degree of random noise in the linear equation does not vary with the values of the independent variables (Allison, 1999). The concept of homoscedasticity is best illustrated through a scatterplot showing the relationship between two variables. In Figure 3, the plot of math achievement (Y) and math aptitude score (X) is displayed. The predicted scores are on the regression line and the actual scores are either above or below the line. For any given value of X, there is the assumption that the Y scores are normally distributed about the given X value and that the standard deviations of these distributions are equal, or homoscedastic. When this assumption is met, probability statements regarding the predicted score can be applied (Hinkle, Wiersma & Jurs, 1997).

Insert Figure 3 about here

Detecting the presence of homoscedasticity in an actual data set can be accomplished by reviewing scatterplots of the independent/dependent variables and the standardized residuals/standardized predicted values. Figure 4 depicts the math achievement scores regressed on math aptitude. Notice the relative even spread of the black data points about the regression line. This is an example of homoscedasticity. For contrast, red data points are also

distributed about the regression line but the degree of variance increases as both the math achievement scores and math aptitude scores increase. These scores represent heteroscedasticity and a violation of the homoscedasticity assumption. In conjunction with this scatterplot, a review of the plot between the standardized residuals and the standardized predicted values is also beneficial. In Figure 5, the standardized residuals and standardized predicted scores for the math achievement variable are graphed and indicate two distributions. The black data points indicate more error associated with the predicted values as the math achievement scores increased, an indicator of the violation of the homoscedasticity assumption. Again for comparison purposes, the red data points indicate a plot between the residuals and predicted values that has a more uniform variance distribution.

Insert Figures 4 and 5 about here

Allison (1999) cites two noteworthy difficulties regarding the violation of the homoscedasticity assumption. The most serious consequence is biased standard errors. Since heteroscedasticity may cause biased standard errors, this will naturally lead to a bias in the test statistics and confidence intervals which in-turn can lead to inaccurate conclusions. Inefficiency is the second outcome of the violation. Heteroscedasticity no longer allows for the least squares estimates to have the smallest standard errors. The phenonmenon gives equal weight to all of the observations, when actually only the observations with the smallest of variances contain the most information. In order to curb the effects of the homoscedasticity violation, variable transformations may be applied to the dependent variable. The available techniques were presented in a previous discussion.

Missing Data Identification Methods and Treatments

Coming to terms with the inevitability of missing data in a research study is an early lesson learned by novice researchers. Deciding how to best contend with the missing data can be a challenging task. Allison (2001) reported a variety of possible explanations for the absence of data in a research project. Three common scenarios are: (a) Subjects may refuse to reveal sensitive or personal information (e.g., income), (b) Some respondents may inadvertently skip one or more questions, and (c) Certain questions may not be applicable to all participants. Of course, other explanations are possible.

Missing data is a common problem regardless of the source. However, identifying any patterns related to the missing cases can be more informative than the amount of missing data. In the simplest of case, missing data is randomly missing or nonrandomly missing. Randomly missing data can be further described as missing at random (MAR) or missing completely at random (MCAR). The distinction between the latter two terms is one not often detailed, but reviewed in this discussion. MCAR data refers to dependent or independent variable data that is no more or no less likely to be missing due to the values of the variable (Allison, 2001). For example, the likelihood that scores for a self-reported math achievement variable would be missing is equal for students of both high and low achievement scores. There is no violation of the MCAR assumption in this instance. The observed, non-missing data cases act as a simple random sample for the study subjects (Everitt, 1998).

MAR data is more of a theoretical conundrum. Data for one variable can be termed missing at random if the probability of data missing for that variable is distinct to the values of that variable with all other analysis variables being controlled (Allison, 2001). Referring back to the math achievement data example, suppose there is the analysis of the self-reported math achievement score and a math attitude score. If the probability of missing data on math

achievement depended on a subject's math attitude score (a variable ranging from 4="I love math" to 1="I hate math"), but within each attitude category, the probability of missing math achievement data was not related to attitude – then the MAR assumption would be met. However, note that this assumption is not quantifiable. The reason is that the missing values are unidentified and a comparison between the values with and without missing data is unobtainable (Allison, 2001).

A violation of the MCAR or MAR assumptions indicates nonrandomly missing data. One way to evaluate the relationship patterns of missing data within the research study, and thereby assessing whether there is a violation of the randomness assumption, is to compare the mean differences between the categories of a grouping variable based on a new variable (a dummy variable). To illustrate this procedure, assume a researcher is testing for a relationship between a math achievement score and a math attitude score. The math achievement scores have missing data throughout the data file, while the math attitude scores are complete. The investigator can generate a dummy variable (DMTACH) based on the achievement score where: missing data cases in the math achievement variable receive a value of "0" for DMTACH, and any non-missing value in the research variable receives a "1" for DMTACH. A test for mean differences (and examination of an appropriate effect size, such as Cohen's d) for the DMTACH variable based on the values of the math attitude groups may reveal patterns between the two variables. If there is no substantive difference between the two groups, then case deletion based on the missing values may be acceptable (assuming sample size is not an issue). Any disparity between the two group means may result in the need to apply additional statistical methods to retain the cases with missing values (Tibatchnick & Fidell, 2001).

One simple method of contending with missing data is listwise deletion. Only records with valid data for all model variables are included in the analysis. Allison (2001) provides two

advantages to this method: (a) it can be used for any statistical method and (b) it does not require a complicated algorithm. From an analysis perspective, if the missing data is MCAR, then the parameter estimates and standard errors will be unbiased and appropriate for the listwise deleted data set as they would be for a complete data set (Allison, 2001). However, this option may not always be the most appropriate technique. If the number of subjects is already fewer than desired, the removal of data will further decrease the sample size (Tibatchnick & Fidell, 2001).

Other procedures contend with missing data when the retention of cases with missing data is essential to the study. These procedures are typically categorized as imputation methods. The basic concept for these procedures is to "substitute some reasonable guess (imputation) for each missing value and then proceed to do the analysis as if there were no missing data" (Allison, 2001, p. 10). One of these imputation methods is termed marginal mean substitution, and it is one of the most common methods to use. In this method, all case data with no missing data are used to generate a variable mean. This variable mean value is then substituted for those cases with missing values within the same variable. This method will have an impact on the variance of a variable with the substitutions since the substituted mean is closer to itself than the would-be (missing) value (Tibatchnick & Fidell, 2001). The consequences of the reduction in variance depends on (a) the degree of reduction and (b) the statistic used.

Another imputation method is termed multiple regression imputation, or conditional mean imputation. For this procedure, suppose one of the independent variables has missing data. The complete data for this single variable is regressed on all other independent/dependent variables. The regression equation is then applied to create predicted values for the missing cases. Regression imputation will provide more accurate "guesses" for the missing values because it uses information provided by the other variables. The underlying problem for these

imputation methods, is the use of imputed data as if it were complete research data. Allison (2001) suggested that these methods generated underestimated standard errors and overestimated test statistics and states that these methods cannot contend with the uncertainty aspect of missing data.

Method

Hueristic Data

The data to be analyzed for this paper is an SPSS data set called hsb500.sav and accompanies the text, *Applied Statistics for the Behavioral Sciences (4th ed.)* by Hinkle, Wiersma and Jurs (1998). This data set is a random sample of 500 senior respondents from a total of 28,240 surveyed in the High School and Beyond National Survey. Specific information regarding the selection process of these students is unknown and irrelevant for the present illustration.

Variables

The file contains data regarding student gender (SEX), father's and mother's educational level (FAED, MAED), math courses taken (ALG, ALG2, GEO, TRIG, CALC), self-reported grades for all high school courses completed (GRADES) and for math courses completed (MATHGR), achievement scores for a 25-item math test (MATHACH), scores for a 16-item visualization/spatial perception inventory (VISUAL), and scores from a 56-item test to determine pattern and relationship detection ability (MOSAIC). Because neither the survey nor the achievement tests themselves are included, data coding accuracy was determined from a report of the "acceptable" score ranges provided with the data set. We added the variables, HSGRDAVG ("actual" overall high school grade average) and MGRDAVG ("actual" overall math grade average) for illustration purposes. These variables are based on the self-reported grades, but are scale variables. A variable describing a student's self-report for the number of

study hours dedicated to their math class (MAVGSTDY), a math attitude score (MATHATTD), and the student's most recent math score earned (LSTSTGRD) were also added to the data file, but are completely fictitious. These scores were added to depict problematic data situations.

Results

The data file with 500 complete records was first analyzed using multiple regression. This initial procedure was conducted prior to any data modifications to produce preliminary results for comparison purposes. The regression model included three predictor variables (MOSAIC, MATHACH, and MAVGSTDY) and one dependent variable (MGRDAVG). With an alpha level of .05, the results for this analysis were statistically significant, $F(3,496)=2.60, p<.001$ and $R^2=.236$. However, through the procedural diagnostics, two outliers were identified as having standardized residuals more than three standard deviations below the mean. To determine the effects of the outliers, both observations (449 and 450) were removed and the regression analysis was again conducted. The R^2 value rose to .257, and, of course, the analysis still produced statistically significant results for an alpha level of .05 and $F(3,494)=2.60, p<.001$). These results point to the impact only a few outliers can have in a regression (here $\Delta R^2 = .021$due to only 2 out of 500 cases).

Data screening began with the calculation of univariate statistics for seven research variables: MGRDAVG, VISUAL, MOSAIC, MATHACH, MATHATTD, LSTSTGRD, and MAVGSTDY. Measures of central tendency and the skewness and kurtosis statistics were considered *in toto* to identify variable distributions that departed from normality. Confirmation of nonnormality was determined by a visual inspection of the variable histograms. One variable in particular, MAVGSTDY, indicated a nonnormal distribution with skewness and kurtosis statistics of 1.550 (standard error of .109) and 1.718 (standard error of .218) respectively. The histogram graphically displays the distributional problems of this variable (see Figure 6).

Insert Figure 6 about here

In addition to displaying problems with normality, this variable also has a nonlinear relationship with another variable. Scatterplots were produced to generate graphical representations between the seven selected research variables. Figure 7 shows the entire scatterplot with two variable relationships highlighted: LSTSTGRD-MAVGSTDY and MGRDAVG-MAVGSTDY. The bivariate scatterplot between LSTSTGRD-MAVGSTDY is characteristically indicative of a nonlinear relationship. A less obvious nonlinear relationship is portrayed by the scatterplot of MGRDAVG-MAVGSTDY. The variables involved in the more extreme case of nonlinearity were transformed in an attempt to make the relationship more linear. According to Tukey and Mostetler's "Bulging Rule", a transformation of LSTSTGRD and/or a transformation of MAVGSTDY will improve nonlinearity. Figure 8 illustrates the new relationships after transformation. An additional effect was noted for the transformed MAVGSTDY variable. Taking the logarithm of MAVGSTDY (LMAS) resulted in both statistical and graphical improvements. The skewness statistic improved from 1.550 to .291 (standard error of .109), but the kurtosis statistic diminished from 1.718 to -.778 (standard error of .218). The histogram for LMAS (see Figure 9), displays the distributional change after transformation.

Insert Figures 7, 8 and 9 about here

Once the screening techniques were applied, a second regression analysis with the transformed variable (LMAS), the two untransformed variables (MOSAIC and MATHACH), and

the single dependent variable (MGRDAVG) was conducted. The outliers previously identified were removed from this second regression as well. Again, the results for this analysis were statistically significant, $F(3,494)=2.60, p<.001$ with $R^2=.268$. However, compared to the results of the first analysis, the R^2 value showed slight improvement. For the present data, then, the impact of the positive skewness of the variable was minimal.

To compare calculated statistics due to missing data, a second data file was created with randomly missing cases for the MOSAIC variable. Since MOSAIC was the only variable known to have a number of observations deleted from the file (105 observations), univariate statistics were generated for only this variable. Central tendency measures were reasonable with a mean of 28.5 and a median of 27, but the skewness statistic of 1.359 (standard error of .123) and a kurtosis statistic of 2.238 (standard error of .245) in conjunction with a histogram indicated a nonnormal distribution. No transformations were performed at this time. Missing values for the MOSAIC variable were analyzed by SPSS Missing Value Analysis (MVA). This module performs missing value diagnostics and uses the maximum likelihood method to generate data for missing cases. The regression imputation option of the MVA module was used, as it's approach is similar to the regression imputation method for missing data discussed earlier. No statistically significant differences were detected in the patterning of the missing variables, so the missing MOSAIC values were determined to be randomly missing. SPSS MVA also produced estimates for the missing MOSAIC values. These were copied to the data file to be used for further analysis.

Determining the most effective procedure for replicating the regression estimates from the full data set analysis was accomplished by reviewing the results of three popular methods. The two outliers identified through the first regression analysis were also removed from these analyses. This was done in an attempt to maintain similar sample sizes for comparison

purposes. The listwise deletion procedure indicated statistical significance, $F(3,389)=2.60, p<.001$ and $R^2=.193$. Mean substitution improved R^2 to .218, with $F(3,494)=2.60, p<.001$. Regression imputation produced the following statistics: $F(3,494)=2.60, p<.001$ with an R^2 of .222. These findings point to the fact that different methods for handling missing data can lead to different results. Therefore, researchers should make their decisions thoughtfully and report decisions in their articles. Additional outliers emerged through all three of the regression diagnostics, but were not removed due to the desired consistency between the data files.

As previously discussed, the MOSAIC variable exhibited nonnormal tendencies, so a square root transformation was applied. Conducting the regression analysis with the transformed variable resulted in almost no improvement. The listwise deletion procedure indicated statistical significance, $F(3,389)=2.60, p<.001$ and $R^2=.194$. Mean substitution showed no change with an R^2 of .218, and $F(3,494)=2.60, p<.001$. Regression imputation was not repeated, since the imputed values were relatively normally distributed and the analysis was already complete.

Discussion

The primary purpose of these exercises was to demonstrate the role of data screening techniques and their potential to improve the performance of statistical methods or lead to different results. Overall, cleaning raw data by affecting normality and linearity problems, removal of outliers, and methods for handling missing value presence proved to increase the R^2 values, if only by very small increments (see Table 4). However, one of the most interesting findings in this exercise was the performance of the regression models when outliers were taken into consideration, irrespective of any additional data cleaning procedures. The second regression analysis performed with the full data set and with two extreme cases removed produced the highest R and R^2 values of .507 and .257 respectively. This appears to be a

Insert Table 1 about here

Data screening procedures make a difference in data analysis and statistical outcomes. However, the researcher must resist the temptation to perform data cleaning simply for the sake of improved effects. These screening procedures, if used properly, assist the researcher in optimizing the data so that the analysis procedure will produce the most accurate and efficient estimates. Analyzing data without conducting some form of preliminary screening is certainly not wise, as this scenario would be a prime example of the adage: "garbage in, garbage out". The regular practice of data screening is worth the time and effort to achieve the most reliable and accurate results, and is consistent with best statistical practice (Wilkinson & APA Task Force on Statistical Inference, 1999).

References

Allison, P., (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P., (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.

Berry, W., (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.

Cohen, J., (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426-443.

Everitt, B.S., (1998). *The Cambridge dictionary of statistics*. Cambridge, UK: Cambridge University Press.

Fox, J., (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Fox, J., (1991). *Regression diagnostics*. Thousand Oaks, CA: Sage.

Henson, R. K., (1999). Multivariate normality: What is it and how is it assessed? In B, Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 193-211). Stamford, CT: JAI Press.

Hinkle, Dennis, Wiersma, William and Jurs, Stephen, (1998). *Applied Statistics for the Behavioral Sciences 4th ed.* Boston: Houghton Mifflin.

Hopkins, K. D. & Weeks, D., (1990, Winter). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement*, 50,717-726.

Lewis-Beck, M. S., (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

SPSS Inc, (1999). *SPSS Base 9.0 Applications Guide*. Chicago: SPSS Inc.

Stevens, J., (1996). *Applied Multivariate Statistics for the Social Sciences (3th ed.)* Mahwah, NJ: Erlbaum.

Tabachnick, B. G. & Fidell, L., (2001). *Using multivariate statistics (4th ed.)* Boston: Allyn & Bacon.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, 54, 594-604.

Appendix A

```
*Work with the FULL DATA SET - Hsb500Rev.sav.

*Open the data file at the specified location.
GET
  FILE='C:\My Documents\SERA Research\Hsb500Rev.sav'.
EXECUTE .

*Run a regression for the complete data file to get preliminary results - use only to compare statistics.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER  mosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Remove two scores identified as outliers.
USE ALL.
COMPUTE filter_$=((id  ~=  449) And (id ~=  450)).
VARIABLE LABEL filter_$ '(id  ~=  449) And (id ~=  450) (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
*Re-run a regression for the complete data file.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER mosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Remove filter.
USE ALL.
*Perform basic variable descriptives for the variables of interest.
FREQUENCIES
  VARIABLES=mgrdavg visual mosaic mathach mathattd lststgrd mavgstdy
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SKEWNESS
  SESKEW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL
  /ORDER  VARIABLES .
*Run a Scatterplot to get a visual of the variable relationships.
GRAPH
  /SCATTERPLOT(MATRIX)= mathattd mgrdavg lststgrd mavgstdy visual mosaic mathach
  /MISSING=LISTWISE .
*Try some transformations to see effect on variable.
COMPUTE smas = SQRT(mavgstdy) .
COMPUTE lmas = LG10(mavgstdy) .
EXECUTE .
FREQUENCIES
```

```
 VARIABLES=smas lmas
 /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SKEWNESS
 SESKEW KURTOSIS SEKURT
 /HISTOGRAM  NORMAL
 /ORDER  VARIABLES .
*Run a Scatterplot to get a visual of the variable relationships - includes the selected transformed
variables.
GRAPH
 /SCATTERPLOT(MATRIX)= mathattd mgrdavg lststgrd lmas visual mosaic mathach
 /MISSING=LISTWISE .
*Run a regression for the complete data file - with a few transformed variables.
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT mgrdavg
 /METHOD=ENTER mosaic mathach lmas
 /SCATTERPLOT=(*ZRESID ,*ZPRED )
 /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

Appendix B

```
*Work with the ALTERNATE DATA SET - Hsb500Alt.sav.
GET
  FILE='C:\My Documents\SERA Research\Hsb500Alt.sav'.
EXECUTE .
*Perform basic variable descriptives for the variable of interest - mosaic had cases deleted.
FREQUENCIES
  VARIABLES=mosaic
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SKEWNESS
  SESKEW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL
  /ORDER  VARIABLES .
*SPSS Missing Value Analysis - must have MVA module istalled for the following lines of syntax to work.
*Copied RI figures to Hsb500 data file for analysis.
MVA
  mosaic mavgstdy mathach
  /TTEST PROB PERCENT=5
  /CROSSTAB PERCENT=5
  /MISMATCH PERCENT=5
  /DPATTERN
  /MPATTERN
  /TPATTERN PERCENT=1
  /REGRESSION ( TOLERANCE=0.001 FLIMIT=4.0 ADDTYPE= T(5) OUTFILE='C:\My
Documents\SERA Research\MVA.sav'
  ) .

*Perform basic variable descriptives for the variable of interest - imputed mosaic cases.
FREQUENCIES
  VARIABLES=mosaice
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SKEWNESS
  SESKEW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL
  /ORDER  VARIABLES .
*Remove two scores identified as outliers - with a few transformed variables & known outliers removed.
USE ALL.
COMPUTE filter_$=((id  ~=  449) And (id ~=  450)).
VARIABLE LABEL filter_$ '(id  ~=  449) And (id ~=  450) (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
*Run a regression analysis using listwise deletion.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER mosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Run a regression anlaysis using mean substitution from the missing case file.
```

```
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER mosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Run a regression anlaysis using values estimated from MVA from the missing case file.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER mosaice mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Try a variety of transformations to see effect on variables.
COMPUTE smosaic = SQRT(mosaic) .
COMPUTE lmosaic = LG10(mosaic) .
EXECUTE .
*Perform basic variable descriptives for the variable of interest - mosaic after trasformation.
FREQUENCIES
  VARIABLES=smosaic lmosaic
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SKEWNESS
  SESKEW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL
  /ORDER  VARIABLES .
*Run a regression analysis using listwise deletion.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER smosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
*Run a regression anlaysis using mean substitution from the missing case file.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mgrdavg
  /METHOD=ENTER smosaic mathach mavgstdy
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

Table 1

Effect size Comparisons for Full Data Sets vs. Missing Values Data Set

	Full Data Set						**Incomplete Data Set**					
	All Data		Outliers Removed		Transformation(s) Applied		All Data		Outliers Removed		Transformation(s) Applied	
Method	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
Original Model	.486	.236	.507	.257	.518	.268						
Listwise Deletion									.439	.193	.440	.194
Mean Substitution									.447	.218	.467	.218
Regression Imputation									.471	.222		

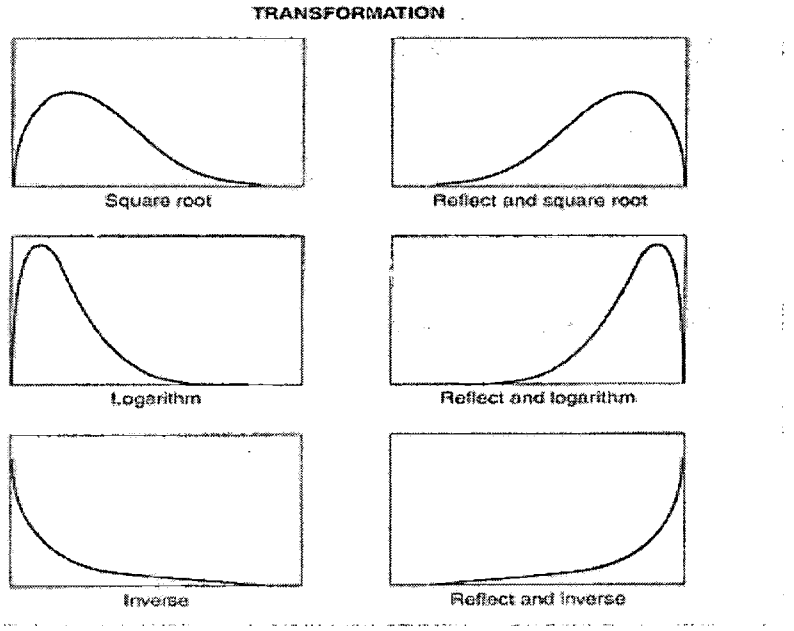

Figure 1. Examples of nonnormal distributions and the transformation methods that may be applied that will render the resulting distributions more normal.

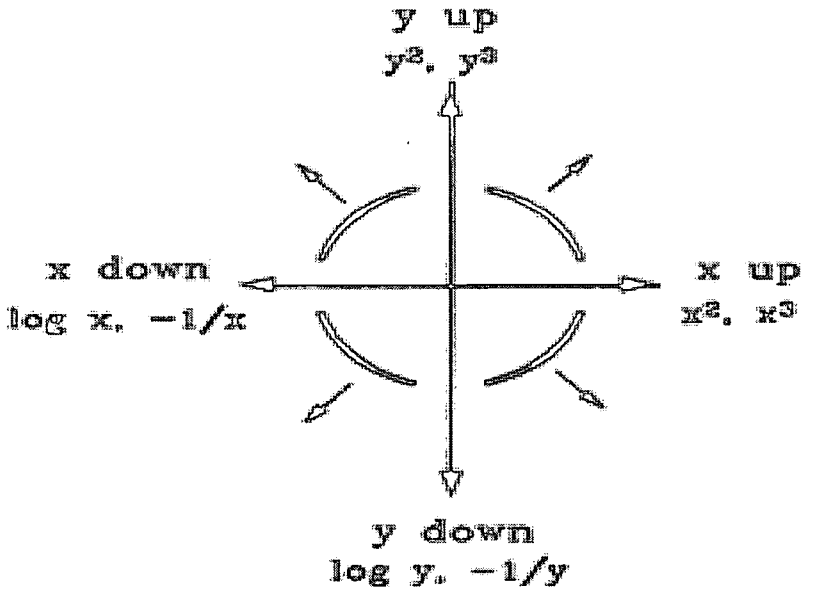

Figure 2. Tukey And Mosteller's (1977) "Bulging Rule". Nonlinear relationships may be improved with the application of the power transformations.

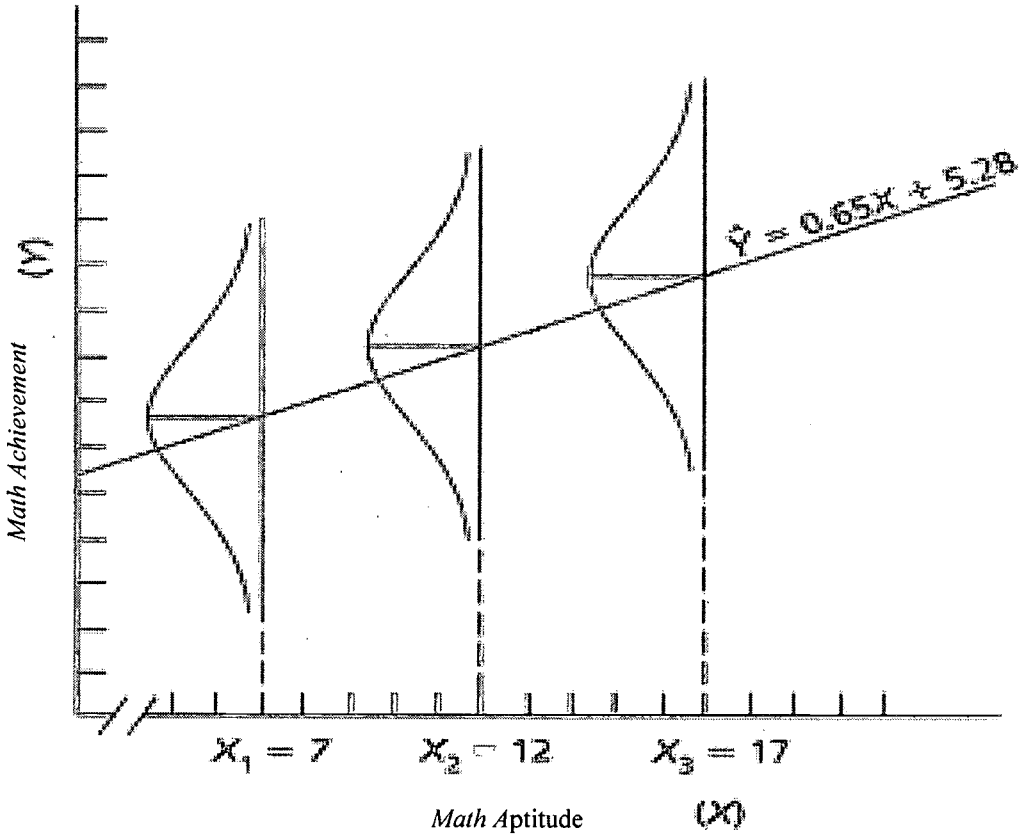

Figure 3. An example of homoscedasticity.

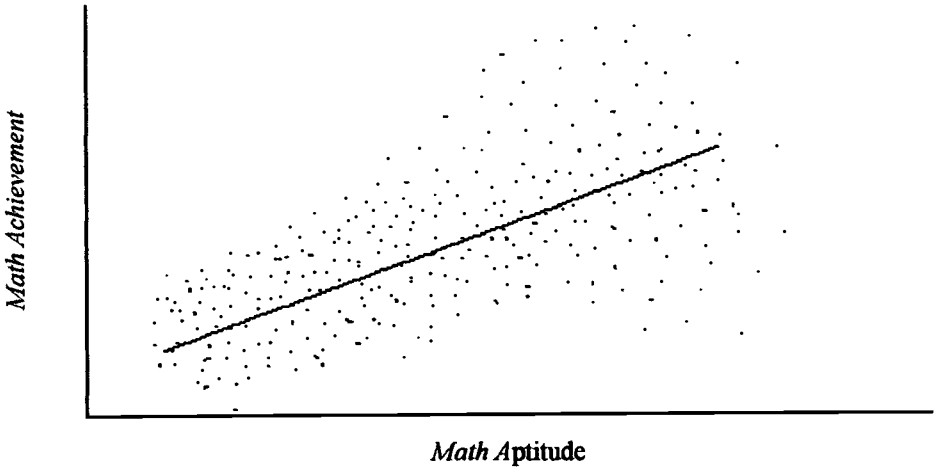

Figure 4. A regression of Math achievement on Math aptitude comparing homoscedasticity (the black data points) vs. heteroscedasticity (red data points).

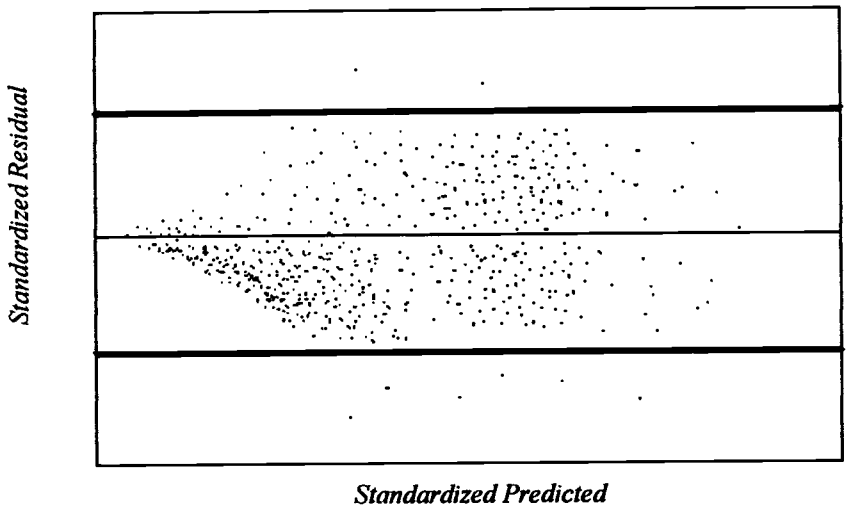

Figure 5. A scatterplot of the standardized residual scores and the standardized predicted scores for two instances of two sets of variables. Homoscedasticity is portrayed by the black data points and heteroscedasticity is detailed by the red data points.

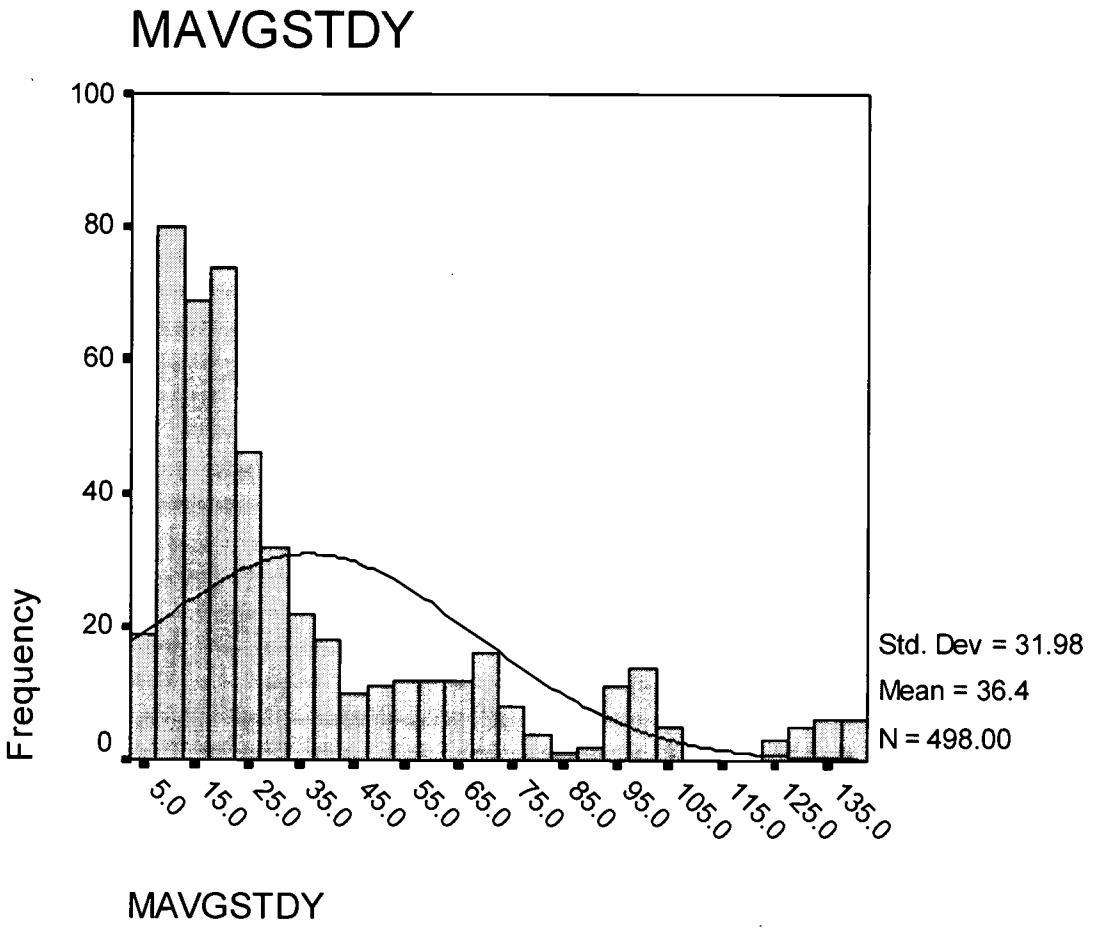

Figure 6. An example of a positively skewed distribution that may be improved with the application of a power transformation.

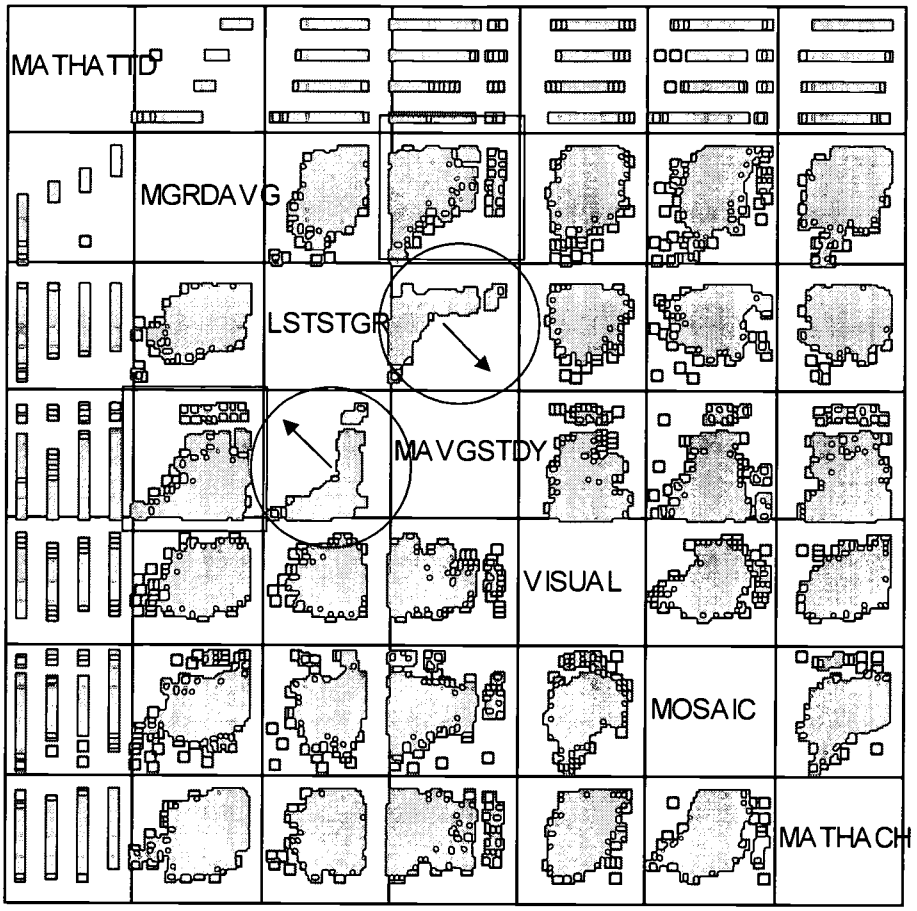

Figure 7. A 3D scatterplot matrix detailing variable relationship. The two nonlinear relationships are highlighted. The arrows indicate the direction a transformation would need to "pull" the curve.

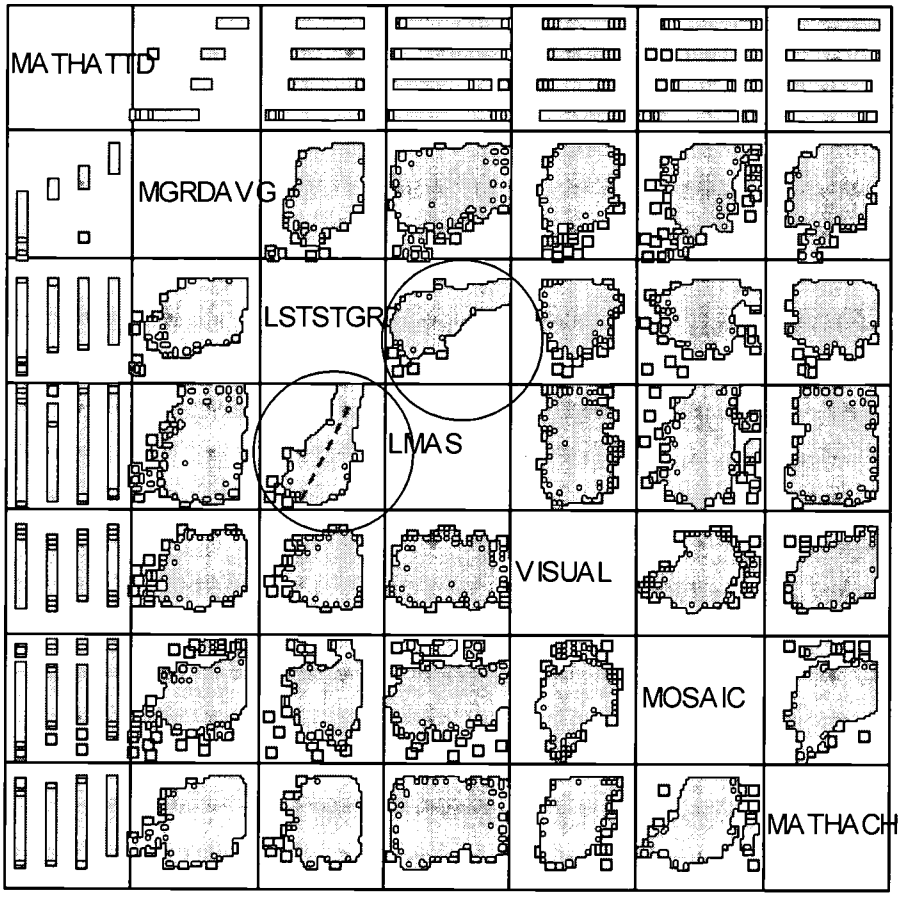

Figure 8. A 3D scatterplot matrix detailing variable relationship after transformation. Note the more linear relationships.

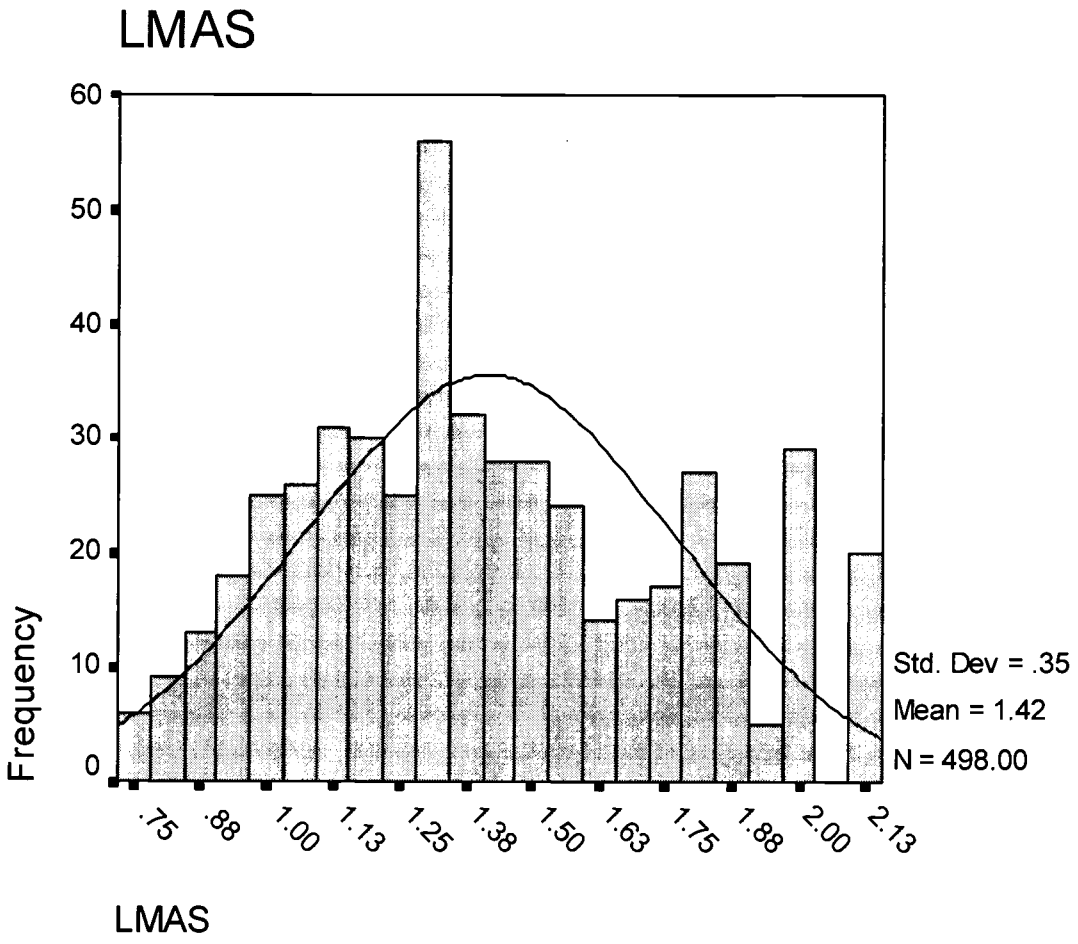

Figure 9. The MAVGSTDY variable after a logarithmic transformation.

TM034292

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Data screening: Essential techniques for data review and prepartation	
Author(s): Leslie R. Odom and Robin K. Henson	
Corporate Source: University of North Texas	Publication Date: Feb. 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only
Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.		

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Robin K. Henson/Assistant Professor	
Organization/Address: Dept. of Technology and Cognition, P.O. Box 311337, Denton, TX 76203-1337	Telephone: 940-369-8385	Fax: 940-565-2185
	E-mail Address: rhenson@unt.edu	Date: 6/20/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:	
ERIC Clearinghouse on Assessment and Evaluation **1129 Shriver Laboratory (Bldg 075)** **College Park, Maryland 20742**	**Telephone: 301-405-7449** **Toll Free: 800-464-3742** **Fax: 301-405-8134** **ericae@ericae.net** **http://ericae.net**

EFF-088 (Rev. 9/97)