Advanced and Multivariate Statistical Methods

Practical Application and Interpretation

Second Edition

Craig A. Mertler
Bowling Green State University

Rachel A. Vannatta
Bowling Green State University

Pyrczak Publishing
P.O. Box 39731 • Los Angeles, CA 90039

CHAPTER 3

PRE-ANALYSIS DATA SCREENING

In this chapter, we discuss several issues related to the quality of data that a researcher wishes to subject to a multivariate analysis. These issues must be carefully considered and addressed *prior* to the actual statistical analysis—they are essentially an analysis *within* the analysis! Only after these quality assurance issues have been examined can the researcher be confident that the main analysis will be an honest one, which will ultimately result in valid conclusions being drawn from the data.

SECTION 3.1 WHY SCREEN DATA?

There are four main purposes for screening data prior to conducting a multivariate analysis. The first of these deals with the accuracy of the data that have been collected. Obviously, the results of any statistical analysis are only as good as the data that were analyzed. If inaccurate data are used, the computer program will run the analysis (in all likelihood), and the researcher will obtain her output. However, the researcher will not be able to discern the extent to which the results are valid simply by examining the output—the results will appear to be legitimate (e.g., values for test statistics will appear, accompanied by significance values, etc.). The researcher will then proceed to interpret the results and draw conclusions; however, unknown to her, they are erroneous conclusions because they have been based on the analysis of inaccurate data.

With a small data file, simply printing the entire data set and proofreading it against the actual data is probably an easy and efficient method of determining the accuracy of data. This can be accomplished by using the **SPSS List** procedure. However, if the data set is rather large, this process would be overwhelming. In this case, examination of the data using frequency distributions and descriptive statistics would be a more realistic method. Both frequency distributions and descriptive statistics can be obtained by using the **SPSS Frequencies** procedure. For quantitative variables, a researcher might examine the range of values to be sure that no cases have values outside the range of possible values. Assessment of the means and standard deviations (i.e., are they plausible?) would also be beneficial. For categorical variables, the researcher would also want to make sure that all cases have values that correspond to the coded values for the possible categories.

The second purpose deals with missing data and attempts to assess the effect of and ways to deal with incomplete data. Missing data occur when measurement equipment fails, subjects do not complete all trials or respond to all items, or errors occur during data entry. The amount of missing data is less crucial than the pattern of missing data (Tabachnick & Fidell, 1996). Missing values that are randomly scattered throughout a data set sometimes are not serious because their pattern is random. Nonrandom missing data, on the other hand, create problems with respect to the generalizability of the results. Since these missing values are nonrandom, there is likely some underlying reason as to their occurrence. Unfortunately, there are no firm guidelines for determining how much missing data is too much for a given sample size. Those decisions still rest largely on the shoulders of the researcher. Methods for dealing with missing data are discussed in Section 3.2.

The third purpose deals with assessing the effects of extreme values (i.e., *outliers*) on the analysis. Outliers are cases with such extreme values on one variable or on a combination of variables that they distort the resultant statistics. Outliers often create critical problems in multivariate data analyses.

There are several causes for a case to be defined as an extreme value, some of which are far more serious than others. These various causes and methods for addressing each will be discussed in Section 3.3.

Finally, all multivariate statistical procedures are based on assumptions, to some degree. The fourth purpose of screening data is to assess the adequacy of fit between the data and the assumptions of a specific procedure. Some multivariate procedures have "unique" assumptions (which will be discussed in those specific chapters) upon which they are based, but nearly all techniques include three basic assumptions: normality, linearity, and homoscedasticity. These assumptions will be defined and methods for assessing the adequacy of the data with respect to each will be discussed in Sections 3.4, 3.5, and 3.6, respectively. Techniques for implementing these methods using SPSS will be described in Sections 3.7 and 3.8.

SECTION 3.2 MISSING DATA

Many researchers tend to assume that any missing data that occur within their data sets is random in nature. This may or may not be the case; if it is not the case, serious problems can arise when trying to generalize to the larger population from which the sample was obtained. The best thing to do when a data set includes missing data is to examine it. Using data that are available, a researcher should conduct tests to see if patterns exist in the missing data. To do so, one could create a dichotomous dummy variable, coded so that one group includes cases with values on a given variable and the other group contains cases with missing values on that variable. For example, if respondents on an attitudinal survey are asked to provide their income and many do not supply that information (for reasons unknown to us at this time), those who provided an income level would be coded "0" and those who did not would be coded "1." Then the researcher could run a simple independent samples *t*-test to determine if there are significant mean differences in attitude between the two groups. If significant differences do exist, there is an indication that those who did not provide income information possess different attitudes than those who did report their income. In other words, there exists a pattern in the missing responses.

If a researcher decides that the missing data are important and need to be addressed, there are several alternative methods to handle these data. (For a discussion on additional techniques to use when there are missing data, the reader is advised to refer to Tabachnick and Fidell, 1996.) The first of these alternatives involves deleting the cases or variables that have created the problems. Any case that has a missing value is simply dropped from the data file. If only a few cases have missing values, this is a good alternative. Another option involves a situation where the missing values may be concentrated to only a few variables. In this case, an entire variable may be dropped from the data set, provided it is not central to the main research questions and subsequent analysis. However, if missing values are scattered throughout the data and are abundant, deletion of cases and/or variables may result in a substantial loss of data, either in the form of subjects or measures. Sample size may begin to decrease rather rapidly and, if the main analysis involves group comparisons, some groups may approach dangerously low sample sizes inappropriate for some multivariate analyses.

A second alternative to handling missing data is to estimate the missing values and then use these values during the main analysis. There are three main methods of estimating missing values. The first of these is for the researcher to use *prior knowledge*, or a well-educated guess, for a replacement value. This method should be used only when a researcher has been working in the specific research area for quite some time and is very familiar with the variables and the population being studied.

Another method of estimating missing values involves the calculation of the means, using available data, for variables with missing values. Those mean values are then used to replace the missing values prior to the main analysis. When no other information is available to the researcher, the mean is the best estimate for the value on a given variable. This is somewhat of a conservative procedure since the overall mean does not change by inserting the mean value for a case, and no guessing on the part of the researcher is required. However, the variance is reduced somewhat since the "real" value probably would not have been precisely equal to the mean. This is usually not a serious problem unless there are

numerous missing values. In this situation, a possible concession is to insert a group mean, as opposed to the overall mean, for a missing value. This procedure is more appropriate for situations involving group comparison analyses.

Finally, a third alternative to handling missing data also estimates the missing value, but does so using a *regression* approach. Regression is discussed extensively in Chapter 7. In regression, several IVs are used to develop an equation that can be used to predict the value on a DV. For missing data, the variable with missing values becomes the DV. Cases with complete data are used to develop this prediction equation. The equation is then used to predict missing values on the DV for incomplete cases. An advantage to this procedure is that it is more objective than a researcher's guess and factors in more information than simply inserting the overall mean. One disadvantage of regression is that the predicted scores are better than they actually would be; since the predicted values are based on other variables in the data set, they are more consistent with those scores than a real score would be. Another disadvantage to regression is that the IVs must be good predictors of the DV in order for the estimated values to be accurate; otherwise, this amounts to simply inserting the overall mean in place of the missing value (Tabachnick & Fidell, 1996).

If any of the above methods is used to estimate missing values, a researcher should consider repeating the analysis using only complete cases (i.e., conduct the main analysis with the missing values and repeat the analysis with no missing values). If the results are similar, one can be confident in the results. However, if they are different, an examination of the reasons for the differences should be conducted. The researcher should then determine which of the two represents the "real world" more accurately, or consider reporting both sets of results.

SECTION 3.3 OUTLIERS

Cases with unusual or extreme values at one or both ends of a sample distribution are known as *outliers*. There are three fundamental causes for outliers: (1) data entry errors were made by the researcher, (2) the subject is not a member of the population for which the sample is intended, or (3) the subject is simply different from the remainder of the sample (Tabachnick & Fidell, 1996).

The problem with outliers is that they can distort the results of a statistical test. This is due largely to the fact that many statistical procedures rely on squared deviations from the mean (Aron & Aron, 1997). If an observation is located far from the rest of the distribution (and, therefore, far from the mean), the value of its deviation would be large. Imagine by how much a deviation increases when squared! Generally speaking, statistical tests are quite sensitive to outliers. An outlier can exert a great deal of influence on the results of a statistical test. A single outlier, if extreme enough, can cause the results of a statistical test to be significant when, in fact, it would not have been if it had been based on all values other than the outlier. The complementary situation can also occur: An outlier can cause a result to be insignificant when, without the outlier, it would have been significant. Similarly, outliers can seriously affect the values of correlation coefficients. As researchers, it is vital that the results of our statistical analyses represent the majority of the data and not be largely influenced by one, or a few, extreme observations. It is for this reason that it is crucial for researchers to be able to identify outliers and decide how to handle them (Stevens, 1992).

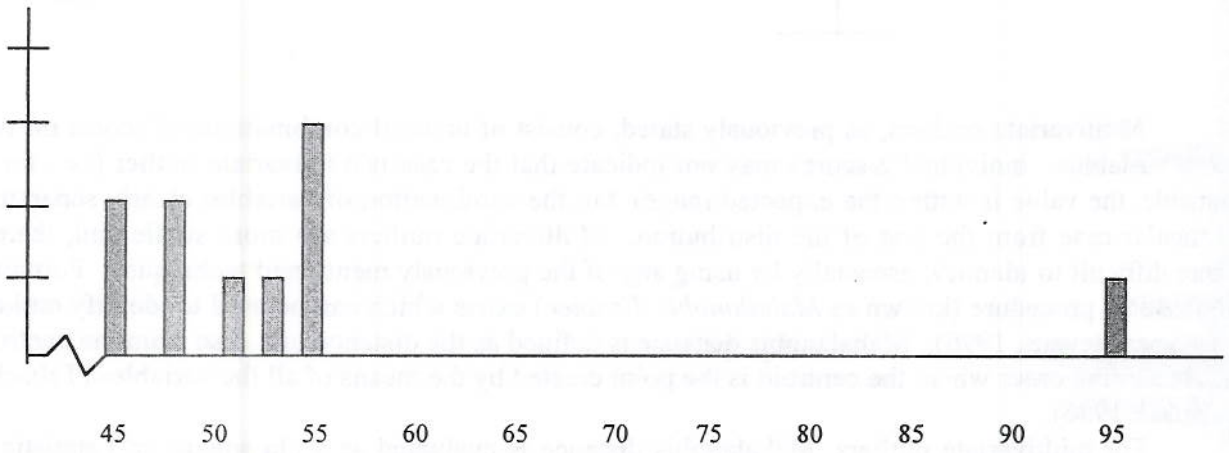
Outliers can exist in both univariate and multivariate situations, among dichotomous and continuous variables, and among IVs as well as DVs (Tabachnick & Fidell, 1996). Univariate outliers are cases with extreme values on one variable; multivariate outliers are cases with unusual combinations of scores on two or more variables. With data sets consisting of a small number of variables, detection of univariate outliers can be relatively simple. This can be accomplished by visually inspecting the data, either by examining a frequency distribution or by obtaining a histogram and looking for unusual values. One would simply look for values that appear far from the others in the data set. In Figure 3.1, Case #3 would clearly be identified as an outlier since it is located far from the rest of the observations.

Figure 3.1 Sample Data Set (a) and Corresponding Histogram (b), Indicating One Outlier.

(a)

Case No.	X_1
1	55
2	48
3	95
4	48
5	51
6	55
7	45
8	53
9	55
10	45

(b)

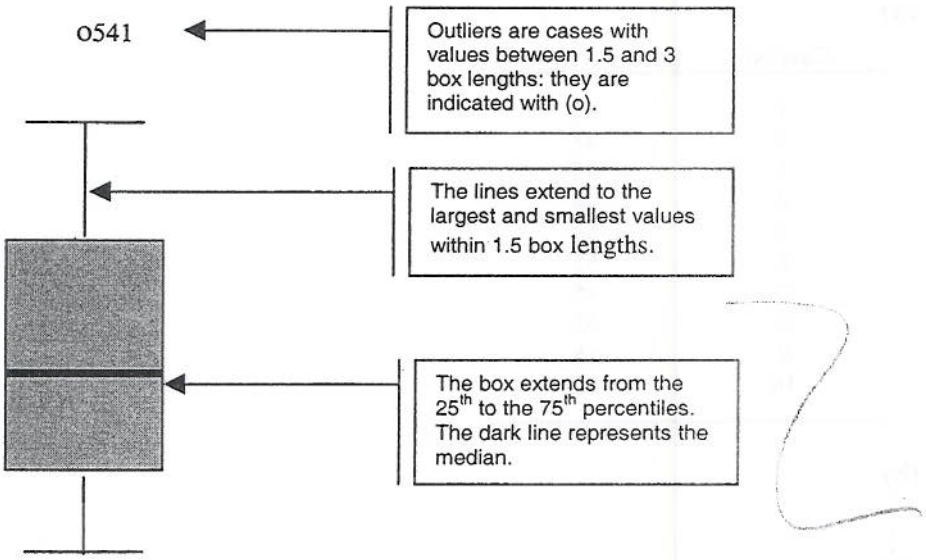


Univariate outliers can also be detected through statistical methods by standardizing all raw scores in the distribution. This is most easily accomplished by transforming the data to *z*-scores. If a normal distribution is assumed, approximately 99% of the scores will lie within three standard deviations of the mean. Therefore, any *z* value greater than +3.00 or less than -3.00 indicates an unlikely value and the case should be considered an outlier. However, with large sample sizes (e.g., $n > 100$), it is likely that a few subjects could have *z*-scores in excess of ±3.00. In this situation, the researcher might want to consider extending the rule to $z > +4.00$ and $z < -4.00$ (Stevens, 1992). For small sample sizes (e.g., $n \leq 10$), any data point with a *z* value greater than 2.50 should be considered as a possible outlier.

Univariate outliers can also be detected by means of graphical methods (Tabachnick & Fidell, 1996). Box plots literally "box in" cases that are located near the median value; extreme values are located far away from the box. Figure 3.2 presents a sample box plot.

As shown in Figure 3.2, the box portion of the plot extends from the 25th to the 75th percentiles, with the dark line representing the median value for the distribution. The lines above and below the box include all values within 1.5 box lengths. Cases with values between 1.5 and 3 box lengths from the upper or lower edges of the box are outliers and are designated by a small circle (o). The specific case number is also listed next to the symbol. Although not depicted in the figure, cases with values greater than 3 box lengths from the edges are also identified and are designated with asterisks (*) and the specific case number.

Figure 3.2 Sample Box Plot Indicating One Outlier.



Multivariate outliers, as previously stated, consist of unusual combinations of scores on two or more variables. Individual z-scores may not indicate that the case is a univariate outlier (i.e., for each variable, the value is within the expected range), but the combination of variables clearly separates the particular case from the rest of the distribution. Multivariate outliers are more subtle and, therefore, more difficult to identify, especially by using any of the previously mentioned techniques. Fortunately, a statistical procedure (known as *Mahalanobis distance*) exists which can be used to identify outliers of any type (Stevens, 1996). Mahalanobis distance is defined as the distance of a case from the centroid of the remaining cases where the centroid is the point created by the means of all the variables (Tabachnick & Fidell, 1996).

For multivariate outliers, Mahalanobis distance is evaluated as a chi-square (χ^2) statistic with degrees of freedom equal to the number of variables in the analysis (Tabachnick & Fidell, 1996). The accepted criterion for outliers is a value for Mahalanobis distance which is significant at $p < .001$, determined by comparing the obtained value for Mahalanobis distance to the chi-square critical value.

Once outliers have been identified, it is necessary to investigate them further. First, the researcher must determine whether the outlier was due to an error in data entry. In this situation, of course, the value would be corrected and the data reanalyzed. However, if the researcher determines that the extreme value was correctly entered and that it may be due to an instrumentation error or that the subject is simply different from the rest of the sample, then it is appropriate to drop the case from the analysis. If it cannot be determined that either of these situations resulted in the extreme value, one should not drop the case from the analysis, but rather should consider reporting two analyses (one with the outlying case included and the other after the case has been deleted) (Stevens, 1996). Remember that outliers should not be viewed as being "bad" because they often represent interesting cases. Care must be taken so that the outlying case is not automatically dropped from the analysis. The case and its value(s) on the variable(s) may be perfectly legitimate.

If the researcher decides that a case with unusual values is legitimate and should remain in the sample, steps may be taken to reduce the relative influence of those cases. Variables may be transformed (i.e., the scales may be changed so that the distribution appears more normal), thus reducing the impact of extreme values. Data transformations are discussed in greater detail in the next section. For a more thorough discussion of variable transformations, the reader is advised to refer to Johnson and Wichern (1998), Stevens (1992), and Tabachnick & Fidell (1996).

SECTION 3.4 NORMALITY

As previously mentioned, there are three general assumptions involved in multivariate statistical testing: normality, linearity, and homoscedasticity. There are consequences of applying statistical analyses—particularly inferential testing—to data that do not conform to these assumptions. If one or more assumptions are violated, the results of the analysis may be biased (Kennedy & Bush, 1985). It is critical then to assess the extent to which the sample data meet the assumptions. The issue at hand is one of test robustness. ***Robustness*** refers to the relative insensitivity of a statistical test to violations of the underlying inferential assumptions. In other words, it is the degree to which a statistical test is still appropriate to apply when some of its assumptions are not met:

> If in the presence of marked departures from model assumptions, little or no discrepancy between nominal and actual levels of significance occurs, then the statistical test is said to be robust with respect to that particular violation (Kennedy & Bush, 1985, p. 144).

The first of these assumptions is that of a normal sample distribution. Prior to examining multivariate normality, one should first assess univariate normality. Univariate normality refers to the extent to which all observations in the sample for a given variable are distributed normally. There are several ways, both graphical and statistical, to assess univariate normality. A simple graphical method involves the examination of the histogram for each variable. Although somewhat oversimplified, this does give an indication as to whether or not normality might be violated. One of the most popular graphical methods is the *normal probability plot*. In a normal probability plot, also known as a *normal Q-Q plot*, the observations are arranged in increasing order of magnitude and plotted against the expected normal distribution values (Stevens, 1996). The plot shows the variable's observed values along the x-axis and the corresponding predicted values from a standard normal distribution along the y-axis (Norusis, 1998). If normality is defensible, the plot should resemble a straight line.

Among the statistical options for assessing univariate normality are the use of skewness and kurtosis coefficients. As a reminder to the reader, *skewness* is a quantitative measure of the degree of symmetry of a distribution about the mean; *kurtosis* is a quantitative measure of degree of peakedness of a distribution. A variable can have significant skewness, kurtosis, or both. When a distribution is normal, the values for skewness and kurtosis are both equal to zero.[1] If a distribution has a positive skew (i.e., a skewness value > zero), there is a clustering of cases to the left and the right tail is extended with only a small number of cases. In contrast, if a distribution has a negative skew (i.e., a skewness value < zero), there is a clustering of cases to the right and the left tail is extended with only a small number of cases. Values for kurtosis that are positive indicate that the distribution is too peaked with long, thin tails (a condition known as *leptokurtosis*); kurtosis values that are negative indicate that the distribution is too flat, with many cases in the tails (a condition known as *platykurtosis*). Significance tests for both skewness and kurtosis values should be evaluated at an alpha level of .01 or .001 for small to moderate sample sizes, using a table of critical values for skewness and kurtosis, respectively. Larger samples may show significant skewness and/or kurtosis values, but often may not deviate enough from normal to make a meaningful difference in the analysis (Tabachnick & Fidell, 1996).

Another specific statistical test that is used to assess univariate normality is the *Kolmogorov-Smirnov statistic*, with Lilliefors significance level. The Kolmogorov-Smirnov statistic tests the null hypothesis that the population is normally distributed. A rejection of this null hypothesis based on the value of the Kolmogorov-Smirnov statistic and associated observed significance level serves as an indication that the variable is not normally distributed.

Multivariate normality refers to the extent to which all observations in the sample for all combinations of variables are distributed normally. Similar to the univariate examination, there are several

---

[1] The mathematical equation for kurtosis gives a value of 3 when the distribution is normal, but statistical packages subtract 3 before printing so that the expected value is equal to zero.

ways, both graphical and statistical, to assess multivariate normality. It is difficult to completely describe multivariate normality but, suffice to say, "normality on each of the variables separately is a necessary but not sufficient condition for multivariate normality to hold" (Stevens, 1996, p. 245). Since univariate normality is a necessary condition for multivariate normality, it is recommended that all variables be assessed based on values for skewness and kurtosis, as previously described.

Other characteristics of multivariate normality include:

1. Each of the individual variables must be normally distributed;
2. Any linear combination of the variables must be normally distributed; and
3. All subsets of the set of variables (i.e., every pairwise combination) must have a multivariate normal distribution (this is known as *bivariate normality*).

Bivariate normality implies that the scatterplots for each pair of variables will be elliptical. An initial check for multivariate normality would consist of an examination of all bivariate scatterplots to check that they are approximately elliptical (Stevens, 1996). A specific graphical test for multivariate normality exists, but requires a special computer program be written, as it is not available in standard statistical software packages (Stevens, 1996).

If the researcher determines that the data have substantially deviated from normal, he or she can consider transforming the data. ***Data transformations*** involve the application of mathematical procedures to the data in order to make them appear "more normal." Once data have been transformed, provided all other assumptions have been met, the results of the statistical analyses will be more accurate. It should be noted that there is nothing unethical about transforming data; transformations are nothing more than a reexpression of the data in different units (Johnson & Wichern, 1998). The transformations are performed on every subject in the data set, so the order and relative position of observations is not affected (Aron & Aron, 1997).

A variety of data transformations exists, depending on the shape (e.g., extent of deviation from normal) of the original raw data. For example, if a distribution differs only moderately from normal, a square root transformation should be tried initially. If the deviation is more substantial, a log transformation is obtained. Finally, if a distribution differs severely, an inverse transformation is tried. The direction of the deviation must also be considered. The above transformations are appropriate for distributions with positive skewness. If the distribution has a negative skew, the appropriate strategy is to "reflect" the variable and then apply the transformation procedure listed above. Reflection involves finding the largest score in the distribution and adding one to it to form a constant that is larger than any score in the distribution. A new variable is then created by subtracting each score from the constant. In effect, this process converts a distribution with negative skewness to one with positive skewness. It should be noted that interpretation of the results of analyses of this variable must also be reversed (Tabachnick & Fidell, 1996). Transformations can be easily obtained in various statistical packages, including SPSS. The transformations discussed here, along with the SPSS language for the computation of new variables, are summarized in Figure 3.3.

Once variables have been transformed, it is important to reevaluate the normality assumption. Following the confirmation of a normal or near-normal distribution, the analysis may proceed typically, resulting in vastly improved results (Tabachnick & Fidell, 1996). Additionally, the researcher should be cognizant of the fact that any transformations performed on the data must be discussed in the methods section of any research report.

It should be understood that the topic of data transformation is much too broad to be adequately addressed here. Should one require further details and examples of these various transformations, it is recommended that the reader refer to Tabachnick and Fidell (1996).

Figure 3.3 Summary of Common Data Transformations to Produce Normal Distributions.

Original Shape	Transformation	SPSS Compute Language
Moderate positive skew	Square root	NEWX=SQRT(X)
Substantial positive skew	Logarithm	NEWX=LG10(X)
With value < 0	Logarithm	NEWX=LG10($X + C$) [a]
Severe positive skew	Inverse	NEWX=1/X
With value < 0	Inverse	NEWX=1/($X + C$) [a]
Moderate negative skew	Reflect & square root	NEWX=SQRT($K - X$) [b]
Substantial negative skew	Reflect & logarithm	NEWX=LG10($K - X$) [b]
Severe negative skew	Reflect & inverse	NEWX=1/($K - X$) [b]

[a] C = a constant added to each score in order to bring the smallest value to at least 1.

[b] K = a constant from which each score is subtracted so that the smallest score equals 1.

SECTION 3.5 LINEARITY

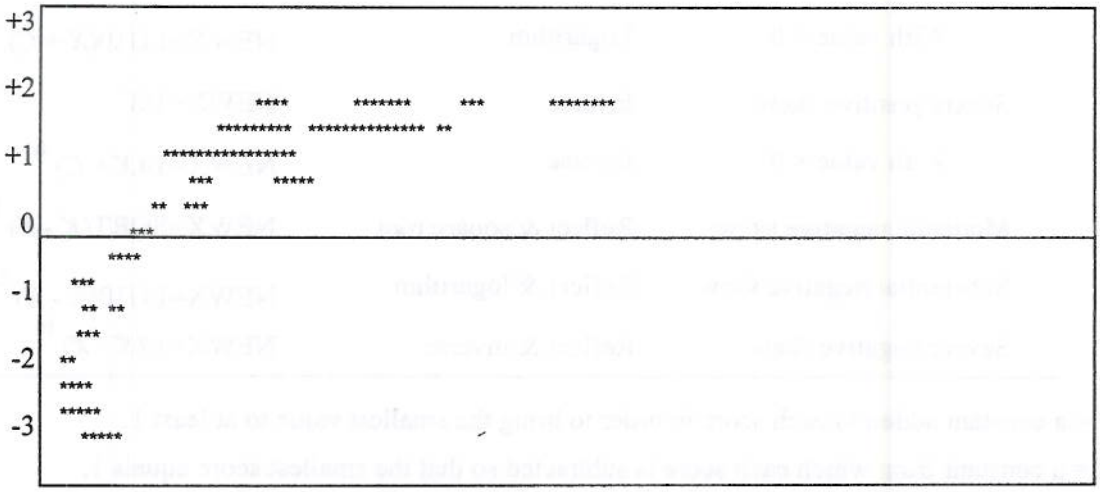
The second assumption, that of linearity, presupposes that there is a straight line relationship between two variables. These two variables can be individual raw data variables (e.g., "drug dosage" and "length of illness") or can be combinations of several raw data variables (i.e., *composite* or *subscale scores*, such as eight items additively combined to arrive at a score for "self-esteem"). The assumption of linearity is important in multivariate analyses due to the fact that many of the analysis techniques are based on linear combinations of variables. Furthermore, statistical measures of relationship such as Pearson's r capture only linear relationships between variables and ignore any substantial nonlinear relationships that may exist (Tabachnick & Fidell, 1996).

There are essentially two methods of assessing the extent to which the assumption of linearity is supported by data. In analyses that involve predicted variables (e.g., multiple regression as presented in Chapter 7), nonlinearity is determined through the examination of residuals plots. *Residuals* are defined as the portions of scores not accounted for by the multivariate analysis; they are also referred to as "prediction errors" since they serve as measures of the differences between obtained and predicted values on a given variable. If standardized residual values are plotted against the predicted values, nonlinearity will be indicated by a curved pattern to the points (Norusis, 1998). In other words, residuals will fall above the zero line for some predicted values and below the line for other predicted values (Tabachnick & Fidell, 1996). Therefore, a relationship that does not violate the linearity assumption would be indicated by the points clustering around the zero line. A nonlinear relationship is depicted in Figure 3.4 (a).

A second, and more crude, method of assessing linearity is accomplished by inspection of bivariate scatterplots. If both variables are normally distributed and linearly related, the shape of the scatterplot will be elliptical. If one of the variables is not normally distributed, the relationship will not be linear, and the scatterplot between the two variables will not be oval-shaped. Assessing linearity by means of bivariate scatterplots is an extremely subjective procedure, at best. The process can become even more cumbersome when data sets with numerous variables are being examined. In situations

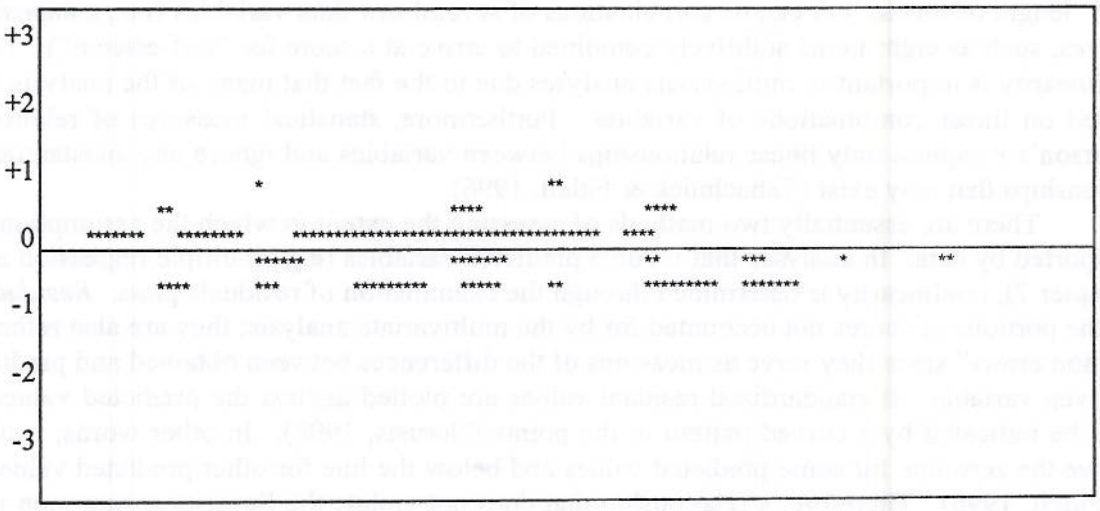
where nonlinearity between variables is apparent, the data can once again be transformed in order to enhance the linear relationship.

Figure 3.4 Sample Standardized Residuals Plots Showing a Strong Nonlinear Relationship (a) and a Linear Relationship (b).



Predicted values

(a) nonlinear relationship



Predicted values

(b) linear relationship

SECTION 3.6 HOMOSCEDASTICITY

The third and final assumption is the assumption of homoscedasticity. *Homoscedasticity* is the assumption that the variability in scores for one continuous variable is roughly the same at all values of another continuous variable. This concept is analogous to the univariate assumption of homogeneity of variance (i.e., the variability in a continuous dependent variable is expected to be roughly consistent at all levels of the independent, or discrete grouping, variable). In the univariate case, homogeneity of variances is assessed statistically with *Levene's test*. This statistic provides a test of the hypothesis that the samples come from populations with the same variances. If the observed significance level for Levene's test is small (i.e., $p < .05$), one should reject the null hypothesis that the variances are equal. It should be noted that a violation of this assumption, based on a "reject" decision of Levene's test, is not fatal to the analysis. Furthermore, Levene's test provides a sound means for assessing univariate homogeneity since it is not affected by violations of the normality assumption (Kennedy & Bush, 1985).

Homoscedasticity is related to the assumption of normality because if the assumption of multivariate normality is met, the two variables must be homoscedastic (Tabachnick & Fidell, 1996). The failure of the relationship between two variables to be homoscedastic is caused either by the nonnormality of one of the variables or by the fact that one of the variables may have some sort of relationship to the transformation of the other variable (Tabachnick & Fidell, 1996). Errors in measurement, which are greater at some levels of the independent variable than at others, may also cause a lack of homoscedasticity.

Heteroscedasticity, or the violation of the assumption of homoscedasticity, can be assessed through the examination of bivariate scatterplots. Within the scatterplot, the collection of points between variables should be approximately the same width across all values with some bulging toward the middle. Although subjective in nature, homoscedasticity is best assessed through the examination of bivariate scatterplots. In multivariate situations, homoscedasticity can be assessed statistically by using *Box's M test for equality of variance-covariance matrices.* This test allows the researcher to evaluate the hypothesis that the covariance matrices are equal. If the observed significance level for Box's M test is small (i.e., $p < .05$), one should reject the null hypothesis that the covariance matrices are equal. It should be noted, however, that Box's M test is very sensitive to nonnormality; thus one may reject the assumption that covariance matrices are equal due to a lack of multivariate normality, not because the covariance matrices are different (Stevens, 1996). Therefore, it is recommended that the tenability of the multivariate normality assumption be assessed prior to examining the results of the Box's M test as a means of assessing possible violations of the assumption of homoscedasticity. Violations of this assumption can be corrected by transformation of variables; however, it should be noted that a violation of the assumption of homoscedasticity, similar to a violation of homogeneity, will not prove fatal to an analysis (Tabachnick & Fidell, 1996; Kennedy & Bush, 1985). The linear relationship will still be accounted for, although the results will be greatly improved if the heteroscedasticity is identified and corrected (Tabachnick & Fidell, 1996).

Because screening data prior to multivariate analysis requires univariate screening, we have provided two univariate examples and one multivariate example.

SECTION 3.7 USING SPSS TO EXAMINE DATA FOR UNIVARIATE ANALYSIS

The following univariate examples explain the steps for using SPSS to examine missing values, outliers, normality, linearity, and homoscedasticity for both grouped data and ungrouped data. Both examples utilize the data set *gssft.sav* from the SPSS Web site.

Univariate Example with Grouped Data

Suppose one is interested in investigating income (*rincom91*) differences between individuals who are either satisfied or not satisfied with their job (*satjob2*). Since this research question compares groups, screening procedures must also examine data for each group.

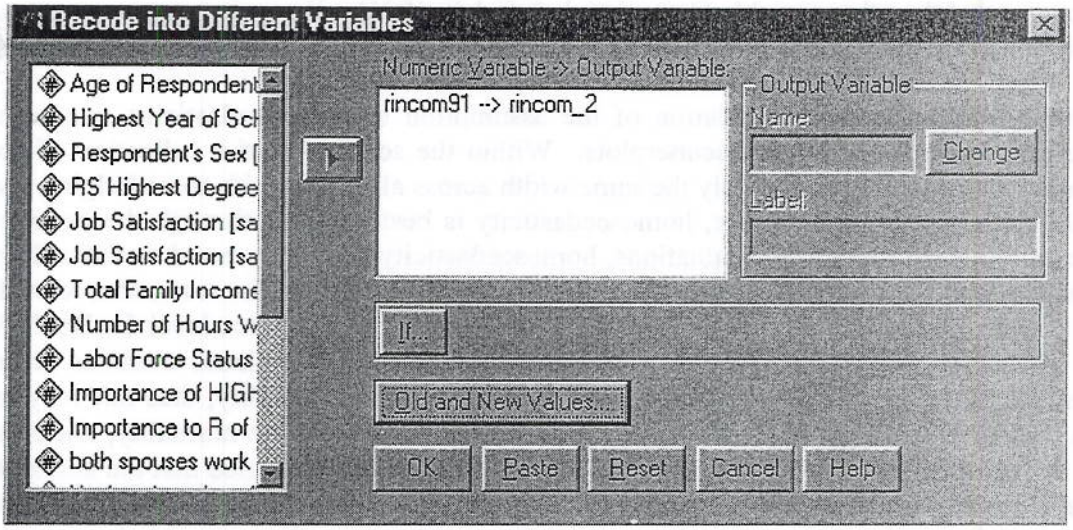
Before screening of data begins, we must first address a coding problem within the variable *rincom91*. This variable represents income levels ranging from 1-21; however, 22 represents "refusal to report" and 98 and 99 represent "not applicable." Since these values could be misinterpreted as income levels, they should be recoded as missing values. To do so, open the following menus:

Transform
Recode
Into Different Variable

Recode into Different Variables Dialogue Box (see Figure 3.5)

We recommend recoding *rincom91* into a different variable, since this provides a record of both the original and altered variables. (Since variables may be transformed numerous times, we will name our new variable *rincom_2*). Once in this dialogue box, indicate the new name for the variable, then click **Change**. Then click **Old and New Values** to specify the transformations.

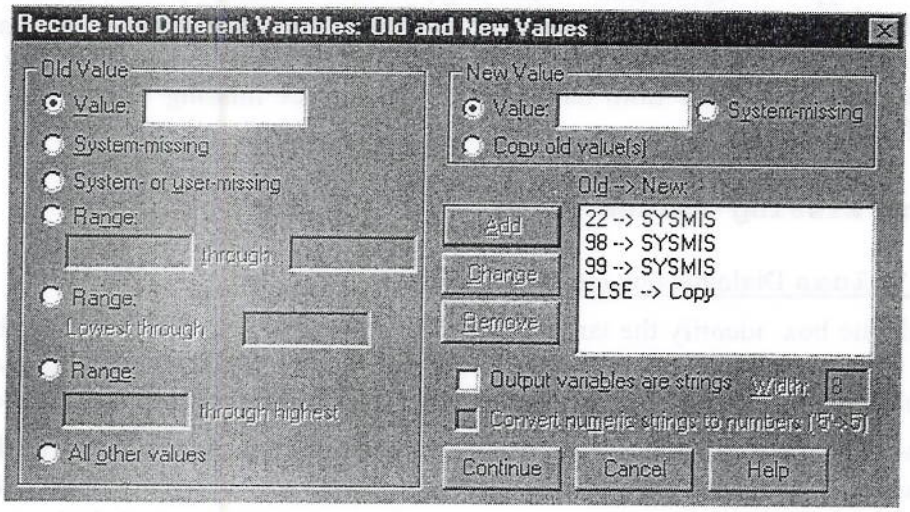
Figure 3.5 Recode Dialogue Box.



Recode into Different Variables: Old and New Values Dialogue Box (see Figure 3.6)

The only cases to be changed are those with values of 22, 98, and 99; all other values will remain the same. To indicate these transformations, click **Value** under Old Value. In the blank, type the value of 22. Under New Value, click **System Missing** then click **Add**. Continue this procedure for both the values of 98 and 99. Once this transformation has been made, be sure to indicate that all other values should be copied. (Specifically, click "All of the values," then click on "Copy old value, then click on "Add," and then click on "Continue" and "OK.") Now examination for missing data may begin.

Figure 3.6 Recode Old and New Values Dialogue Box.



Missing Data

SPSS has several procedures within the analysis process for deleting cases or subjects that have missing values. For most analyses, the **Option** Dialogue Box typically displays the default of **Listwise** in which subjects with missing values are removed only if the missing values are critical to the variables being analyzed. **Pairwise** is another method of deleting subjects with missing values. This method removes subjects with missing values from any and all analyses, even if the missing values are not critical to the variables being studied. Consequently, most researchers utilize the **Listwise** method since it allows for the maximum number of subjects within each analysis.

Examination of missing data in categorical variables can be done by creating a frequency table using **Frequencies**. To determine the extent of missing values within the variable of *satjob2*, the following menus would be selected:

Analyze
Descriptive Statistics
Frequencies

Quantitative variables with missing data can be examined by creating a table of **Descriptive Statistics**. To evaluate missing values in *rincom_2*, open the following menus:

Analyze
Descriptive Statistics
Descriptives

For our example, the frequency and descriptive tables reveal zero missing values for *satjob2* and 37 missing values for *rincom_2*. Typically, if a categorical variable has less than 5% of cases missing, the **Listwise** default would be utilized to delete the cases during the analyses. If a categorical variable has 5-15% of cases with missing data, an additional level or category would be created within the variable so that missing data would be recoded with this new level. Since SPSS no longer detects the missing values and does not recognize the new category as providing meaningful information for the variable being analyzed, these cases would not be included in the analysis.

SPSS also provides a variety of options for handling missing values in quantitative data. In our example, data is missing for 37 cases in the variable *rincom_2*. Since less than 5% of the cases have missing values, the **listwise** default will be used to delete the missing cases. If 5% or more of the values were missing, the method of replacement would be utilized. The most common method is to re-

place the missing values with the mean score of available cases for that variable. The replacement procedure also allows for other types of replacement values (e.g., median of nearby points, mean of nearby points). Typically, replacing 15% or less of the subjects will have little effect on the outcome of the analysis. However, if a certain subject or variable has more that 15% missing data, you may want to consider dropping the subject or variable from the analysis. To replace missing values with an estimated value, select the following menus:

Transform
Replace Missing Values

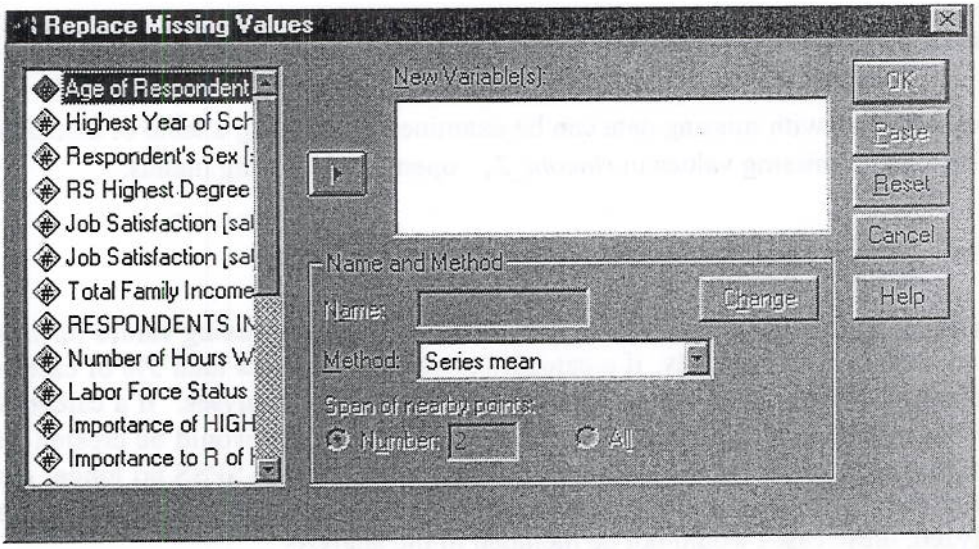
Replace Missing Values Dialogue box (see Figure 3.7)

Once in this dialogue box, identify the targeted variable and move it to the New Variable Box. Notice that a name for the new variable has been generated; this may be changed accordingly. Next, select the Method of replacement. Five options are available in which missing values are replaced by:

Series Mean—The mean of all available cases for the specific variable. This is the default.
Mean of Nearby Points—The mean of surrounding values. You can designate the number of surrounding values to use under **Span of Nearby Points.** The default span is two values.
Median of Nearby Points—The median of surrounding values, the number of which can be designated.
Linear Interpolation—The value midway between the surrounding two values.
Linear Trend at Point—A value consistent with a trend that has been established (e.g., values increasing from the first to the last case).

Once the method of replacement has been determined for the variable, you may also identify additional variables for replacement by using the **Change** button. This will allow you to identify another variable as well as another replacement method.

Figure 3.7 Replace Missing Values Dialogue Box.



Missing values in quantitative variables can also be estimated by creating a regression equation in which the variable with missing data serves as the dependent variable. Since this method is fairly sophisticated, we will discuss how to use predicted values in Chapter 7 on Multiple Regression.

Finally, if publishing the results of analyses that have utilized replacement of missing values,

one should present the procedure(s) for handling such data.

Outliers

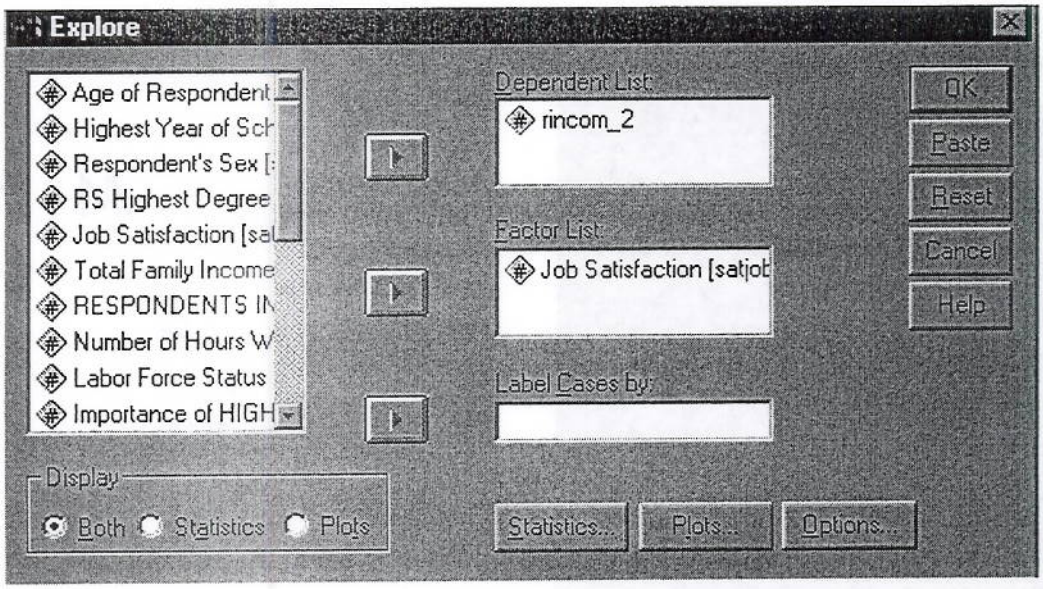
Since univariate outliers are subjects or cases with extreme values for one variable, identification of such cases is fairly easy. The **Explore** menu under **Descriptive Statistics** offers several options for such examination. To identify outliers in the categorical variable of *satjob2*, **Frequencies** could be used to detect very uneven splits in categories, splits that typically produce outliers. Categorical variables with 90-10 splits between categories are usually deleted, since scores in the category with 10% of the cases influence the analysis more than those in the category with 90% of the cases. Because our example research question investigates group differences in income, both the IV (*satjob2*) and DV (*rincom_2*) can be examined for outliers using **Explore**. This procedure will allow us to identify outliers for income within each group. To do so, select the following menus:

Analyze
Descriptive Statistics
Explore

Explore Dialogue Box (see Figure 3.8)

Within this dialogue box, move the DVs into the Dependent List. Move IVs into the Factor List. After you have defined the variables, click the **Statistics** button.

Figure 3.8 Explore Dialogue Box.



Explore Statistics Dialogue Box (see Figure 3.9)

This box provides the following options for examining outliers:

Descriptives—Calculates descriptive statistics for all subjects and identified categories in the data. This is selected by default.

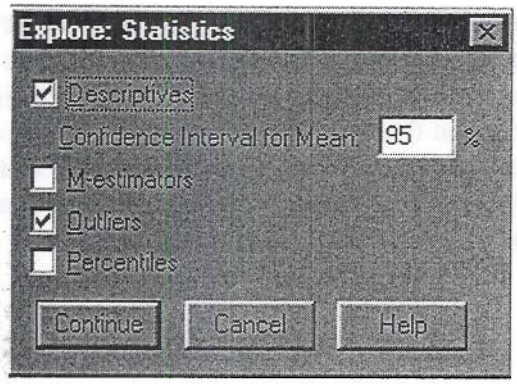
M-Estimators—Assigns weights to cases depending upon their distance from the center.

Outliers—Identifies the five highest and five lowest cases for the DV by group.

Percentiles—Displays the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for the DV by group.

For our example, we selected **Descriptives** and **Outliers.** Click **Continue,** then click **Plots.**

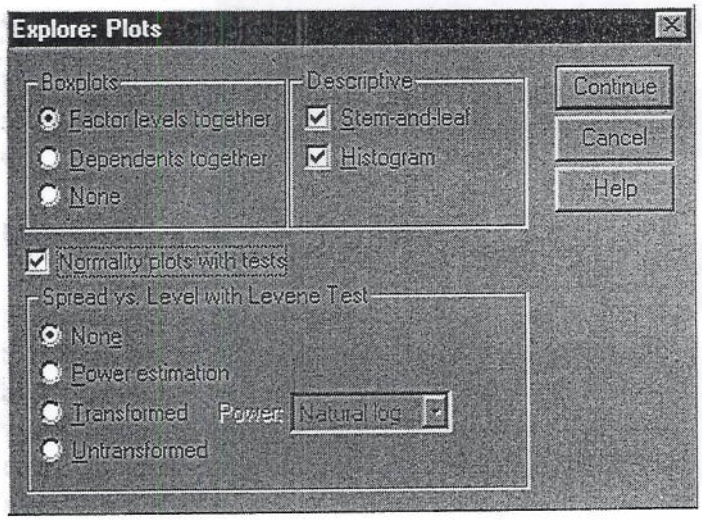
Figure 3.9 Explore Statistics Dialogue Box.



Explore Plots Dialogue Box (see Figure 3.10)

This box provides several options for creating graphic representations of the data. For our example, we will select **Boxplot** and **Stem-and-Leaf Plot**. Since it is best to examine normality after outliers have been addressed, other selections such as **Normality Plots with Tests** and **Histograms** will be conducted later.

Figure 3.10 Explore Plots Dialogue Box.



The output reveals some outlier problems within the example. The case summary shows category splits in that 44% of the sample is very satisfied while 56% is not satisfied. This split is not severe enough to delete this variable. The table generated on extreme values (see Figure 3.11) identifies the five highest and lowest scores for each group; keep in mind that these values are not necessarily outliers. The boxplot (see Figure 3.12) generated reveals that both groups have some outliers. The stem-and-leaf plots (see Figure 3.13) support this finding but provide more information regarding the number of outliers. The first plot indicates that 16 subjects who are very satisfied reported extreme income values of 3 or less. In contrast, the second plot displays 22 subjects who are not very satisfied reported extreme in-

come values of 3 or less. Since the number of outlying cases for both groups is fairly small, these outliers could either be deleted using the case numbers identified in the boxplot or be altered to a value that is within the extreme tail in the accepted distribution. In this example, outliers will be altered by replacing them with a maximum/minimum value (depending on the direction of outliers) that falls within the accepted distribution. To alter the outliers in *rincom_2*, the stem-and-leaf plot (see Figure 3.13) helps one identify the specific outlying values to be altered and the accepted minimum value to be used as the replacement value. Cases that have an income level of 3 or less will be replaced with the accepted value of 4. To alter outliers, complete the following steps:

```
Transform
    Recode
        Into Different Variable
```

Figure 3.11 Extreme Values Table for Income (*rincom_2*) by Job Satisfaction (*satjob2*).

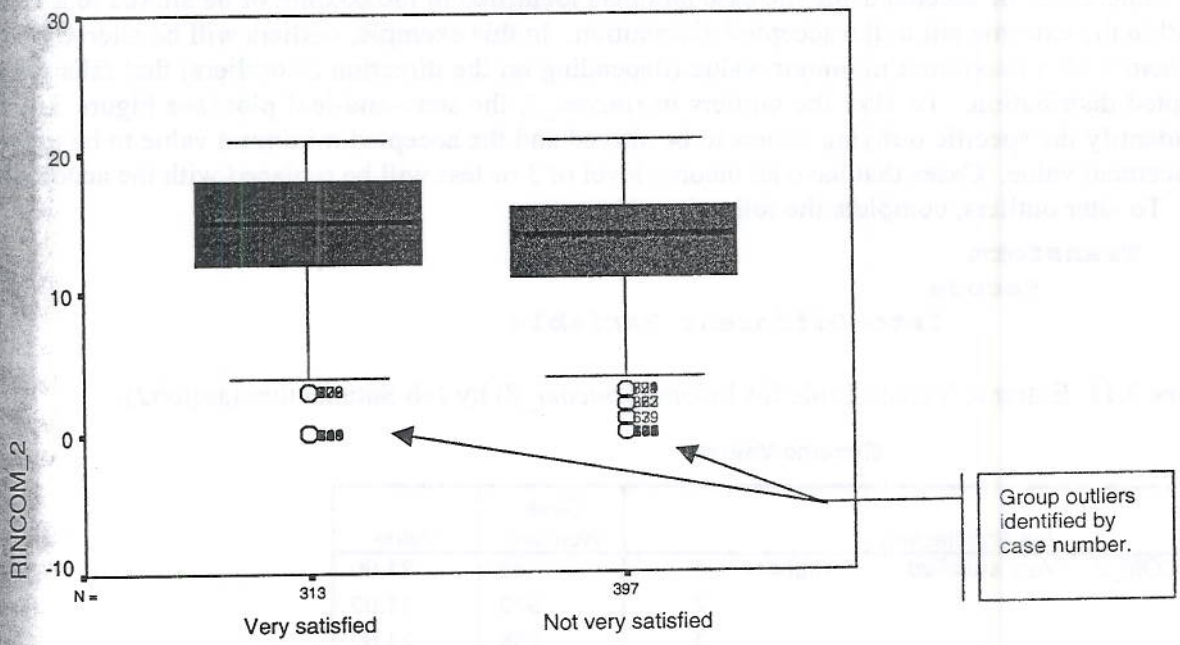
Extreme Values

	Job Satisfaction			Case Number	Value
RINCOM_2	Very satisfied	Highest	1	69	21.00
			2	579	21.00
			3	388	21.00
			4	479	21.00
			5	62	.[a]
		Lowest	1	716	.00
			2	663	.00
			3	363	.00
			4	649	.00
			5	691	.[b]
	Not very satisfied	Highest	1	463	21.00
			2	89	21.00
			3	64	21.00
			4	502	21.00
			5	41	.[a]
		Lowest	1	670	.00
			2	419	.00
			3	206	.00
			4	507	.00
			5	184	.[b]

a. Only a partial list of cases with the value 21 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

Figure 3.12 Boxplot for Income (*rincom_2*) by Job Satisfaction (*satjob2*).



Recode **into Different Variables** Dialogue Box (see Figure 3.14)

Recoding *rincom_2* into a different variable will allow us to conduct our analysis with the original variable (*rincom91),* the first altered variable (*rincom_2*) and the second altered variable (*rincom_3*). This joint analysis helps determine if altering the outliers had an impact on the results. Once in this dialogue box, identify the variable(*rincom_2*) to be altered and move to the Input (*or* Numeric Variable)→Output Variable box. Indicate a new name for the variable (*rincom_3*), then click **Change**. Click **Old and New Values** to specify the transformations.

Recode **Old and New Values** Dialogue Box (see Figure 3.15)

The only cases to be changed are those with values 3 or less; all other values will remain the same. To indicate these transformations, under Old Value, click **Range: Lowest through X.** In the blank, indicate the cutoff value of 3. Under New Value, type in 4 and click **Add**. These commands have transformed the outliers (those 3 or less) to a value of 4. The next step is to indicate that all other values will stay the same. To do so, under Old Value click **Else**.[2] Under New Value, click **Copy old value(s)**, then click **Add**. Once cases have been altered, you can proceed with further data examination and analysis. But remember, when conducting the analyses, do so with both original and altered variables.

[2]In some versions of SPSS, "**Else**" is "**All other values**"

Figure 3.13 Stem-and-Leaf Plots for Income (*rincom_2*) by Job Satisfaction (*satjob2*).

```
RINCOM_2 Stem-and-Leaf Plot for
SATJOB2= Very satisfied

 Frequency     Stem &  Leaf

    16.00 Extremes    (=<3.0)
     1.00         4 .  0
     2.00         5 .  00
     2.00         6 .  00
     3.00         7 .  000
     7.00         8 .  0000000
     9.00         9 .  000000000
    11.00        10 .  00000000000
    22.00        11 .  0000000000000000000000
    13.00        12 .  0000000000000
    17.00        13 .  00000000000000000
    21.00        14 .  000000000000000000000
    36.00        15 .  000000000000000000000000000000000000
    41.00        16 .  00000000000000000000000000000000000000000
    25.00        17 .  0000000000000000000000000
    30.00        18 .  000000000000000000000000000000
    23.00        19 .  00000000000000000000000
     8.00        20 .  00000000
    26.00        21 .  00000000000000000000000000

 Stem width:      1.00
 Each leaf:       1 case(s)
```

16 subjects are outliers with values of 3 or less.

```
RINCOM_2 Stem-and-Leaf Plot for
SATJOB2= Not very satisfied

 Frequency     Stem &  Leaf

    22.00 Extremes    (=<3.0)
     8.00         4 .  00000000
     4.00         5 .  0000
     4.00         6 .  0000
     4.00         7 .  0000
     9.00         8 .  000000000
    24.00         9 .  000000000000000000000000
    24.00        10 .  000000000000000000000000
    33.00        11 .  000000000000000000000000000000000
    29.00        12 .  00000000000000000000000000000
    27.00        13 .  000000000000000000000000000
    35.00        14 .  00000000000000000000000000000000000
    41.00        15 .  000000000000000000000000000000000000000000
    36.00        16 .  000000000000000000000000000000000000
    24.00        17 .  000000000000000000000000
    27.00        18 .  000000000000000000000000000
    15.00        19 .  000000000000000
    14.00        20 .  00000000000000
    17.00        21 .  00000000000000000

 Stem width:      1.00
 Each leaf:       1 case(s)
```

22 subjects are outliers with values of 3 or less.

Figure 3.14 Recode into Different Variables Dialogue Box.

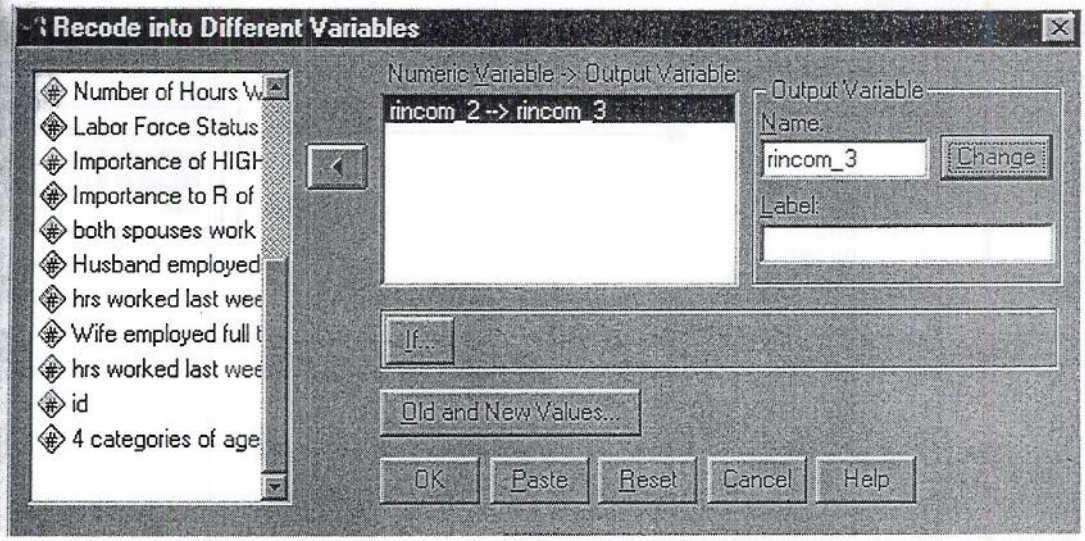
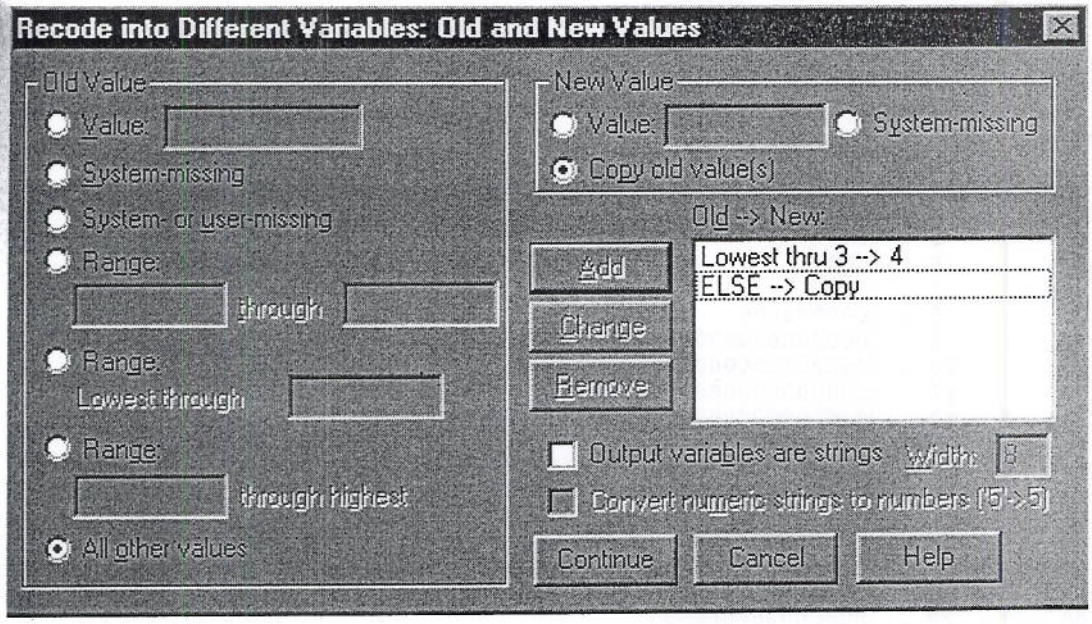


Figure 3.15 Recode Old and New Values Dialogue Box.



Normality, Linearity, and Homoscedasticity

The **Explore** procedure also provides several options for examining normality and is usually conducted after addressing outliers. To conduct this procedure using the DV (*rincom_3*) and IV (*satjob2*), return to the previous directions for Explore. Within the **Explore Statistics** Dialogue Box, be sure to check **Descriptives**. In the **Explore Plots** Dialogue Box, be sure to check **Histograms** and **Normality plots with tests.** These are most helpful in examining normality. Descriptive statistics (see Figure 3.16) present skewness and kurtosis values, which also imply negative distributions. Typically, skewness and kurtosis values should lie between +1 and –1. Histograms (see Figure 3.17) display moderate, negatively skewed distributions for both groups. The normal

Q-Q plots support this finding as the observed values deviate somewhat from the straight line (see Figure 3.18). Tests of normality were also calculated. Specifically, the Kolmogorov-Smirnov test (see Figure 3.19) significantly rejects the hypothesis of normality of income for the populations of both groups. Thus, the variable of *rincom_3* must be transformed again. To decrease the moderate negative skewness, the transformation procedure will reflect and take the square root of the variable. Steps for such transformation follow:

```
Transform
    Compute
```

Figure 3.16 Descriptive Statistics for Income (*rincom_2*) by Job Satisfaction (*satjob2*).

Descriptives

	Job Satisfaction			Statistic	Std. Error
RINCOM_3	Very satisfied	Mean		14.6166	.2475
		95% Confidence Interval for Mean	Lower Bound	14.1297	
			Upper Bound	15.1035	
		5% Trimmed Mean		14.8518	
		Median		15.0000	
		Variance		19.167	
		Std. Deviation		4.3780	
		Minimum		4.00	
		Maximum		21.00	
		Range		17.00	
		Interquartile Range		6.0000	
		Skewness		-.739	.138
		Kurtosis		.078	.275
	Not very satisfied	Mean		13.2972	.2238
		95% Confidence Interval for Mean	Lower Bound	12.8573	
			Upper Bound	13.7372	
		5% Trimmed Mean		13.3938	
		Median		14.0000	
		Variance		19.881	
		Std. Deviation		4.4588	
		Minimum		4.00	
		Maximum		21.00	
		Range		17.00	
		Interquartile Range		5.5000	
		Skewness		-.395	.122
		Kurtosis		-.404	.244

For a normal distribution, kurtosis and skewness values will be close to zero but can range between –1 and +1.

Figure 3.17 Histograms for Income (*rincom_3*) by Job Satisfaction (*satjob2*).

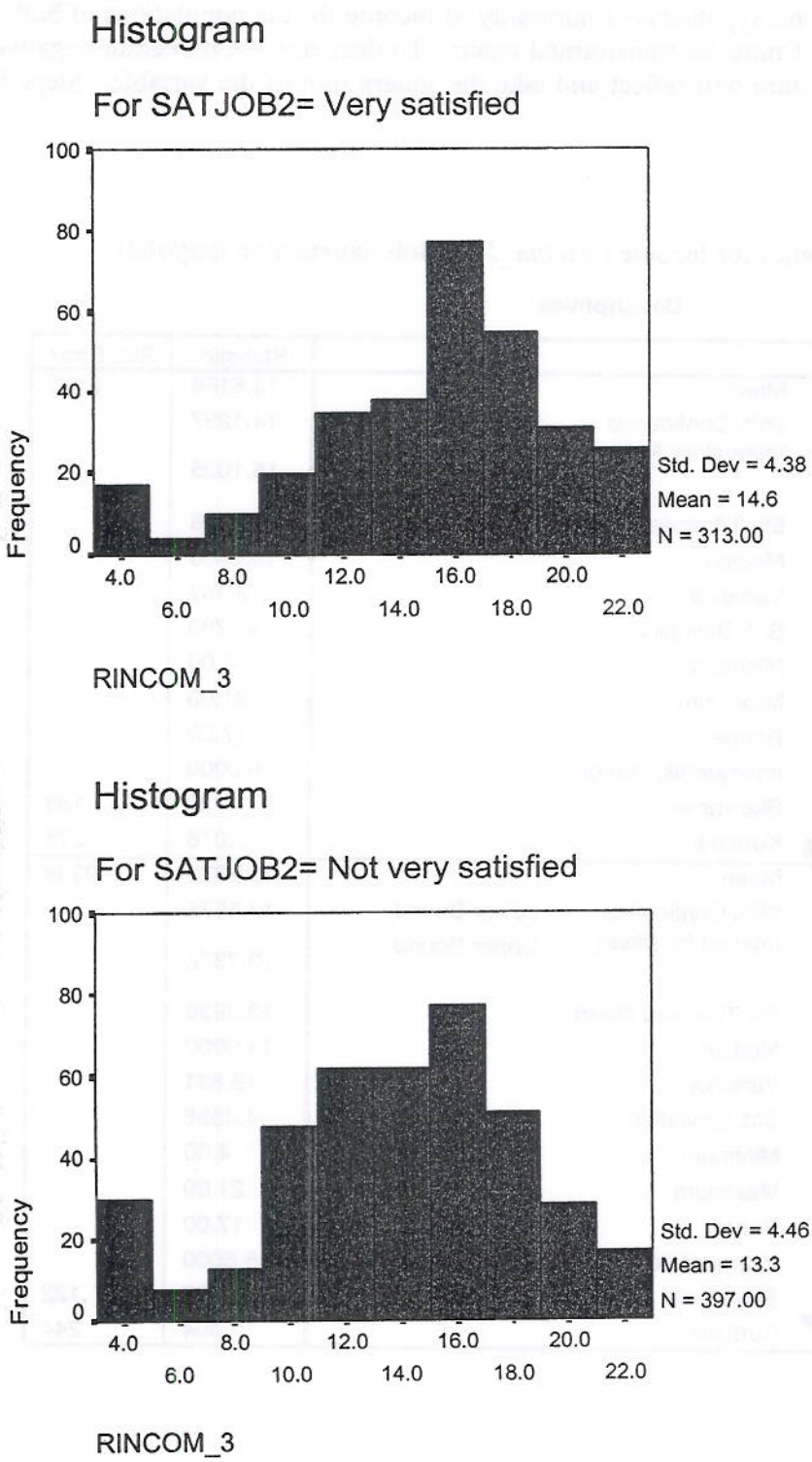


Figure 3.18 Normal Q-Q Plots for Income (*rincom_3*) by Job Satisfaction (*satjob2*).

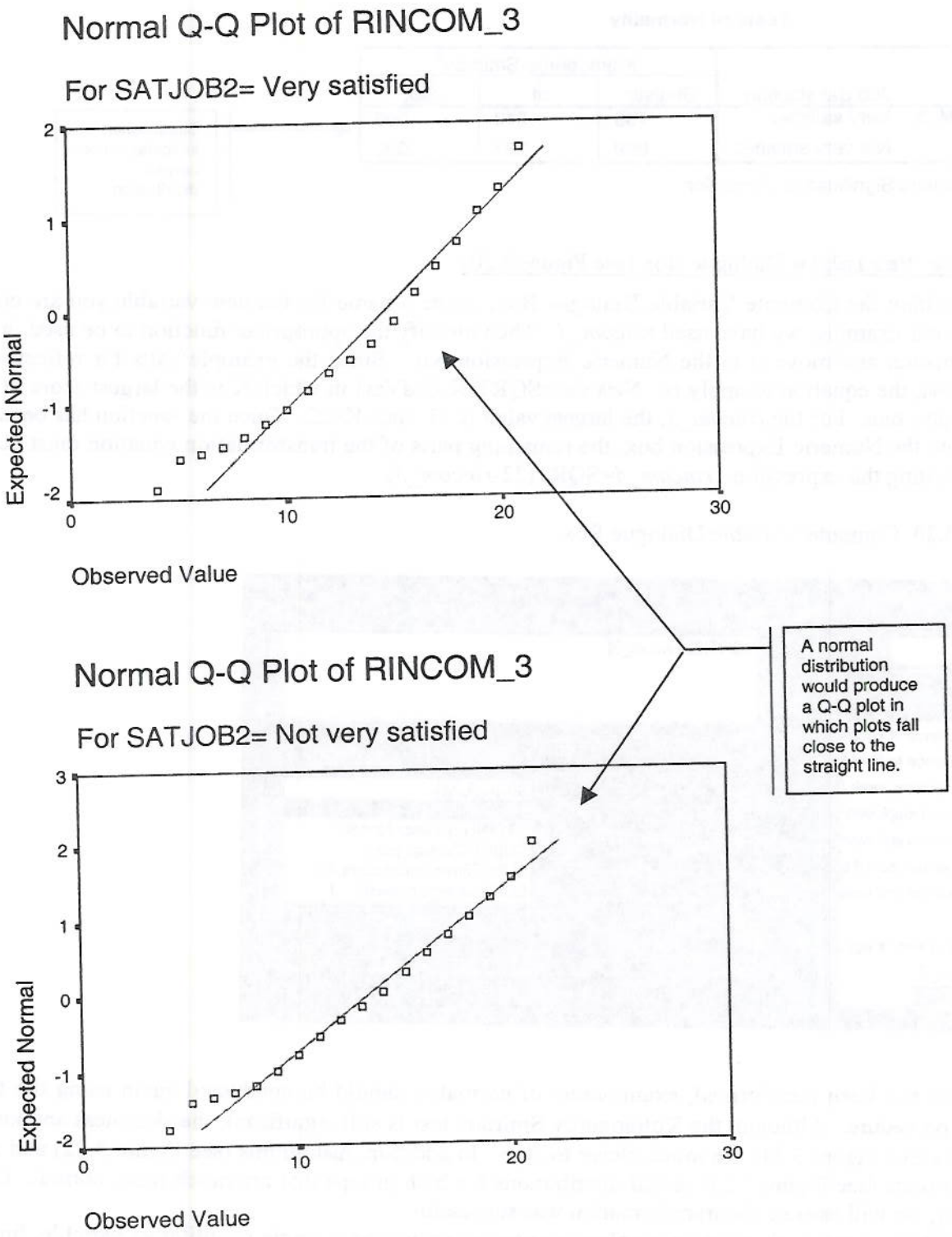


Figure 3.19 Tests for Normality for Income (*rincom-3*) by Job Satisfaction (*satjob2*).

Tests of Normality

		Kolmogorov-Smirnov[a]		
	Job Satisfaction	Statistic	df	Sig.
RINCOM_3	Very satisfied	.139	313	.000
	Not very satisfied	.089	397	.000

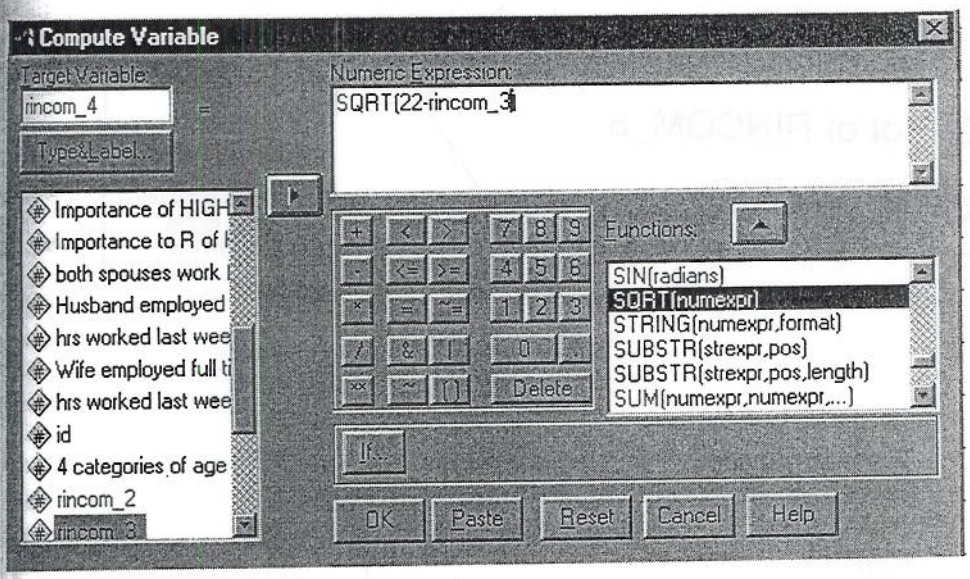
a. Lilliefors Significance Correction

Significance indicates a non-normal distribution.

`Compute Variable` Dialogue Box (see Figure 3.20)

Within the Compute Variable Dialogue Box, create a name for the new variable you are creating. For our example, we have used *rincom_4*. Then identify the appropriate function to be used in the transformation and move it to the Numeric Expression box. Since the example calls for reflect with square root, the equation to apply is: NewVar=SQRT(K-OldVar) in which K is the largest score of the OldVar plus one. For the *rincom_3*, the largest value is 21; thus K=22. Once the function has been inserted into the Numeric Expression box, the remaining parts of the transformation equation must be inserted creating the expression: *rincom_4*=SQRT(22-*rincom_3*).

Figure 3.20 Compute Variable Dialogue Box.



Once data has been transformed, examination of normality should be conducted again using the **`Explore`** procedure. Although the Kolmogorov-Smirnov test is still significant, the skewness and kurtosis values (see Figure 3.21) are much closer to zero. In addition, histograms (see Figure 3.22) and normal Q-Q plots (see Figure 3.23) reveal distributions for both groups that are much more normal. Consequently, we will assume the transformation was successful.

Since our research question involves comparing groups on a single quantitative variable, linearity cannot be examined. However, homoscedasticity, also known as *homogeneity of variance* when comparing groups, can be assessed by determining if the variability for the DV (*rincom_4*) is about the same within each category of the IV (*satjob2*). This can be completed when conducting the group comparison analyses (e.g., *t*-test, ANOVA). Within these statistical procedures, Levene's test for equal variances is automatically calculated. Figure 3.24 presents output from an independent *t*-test conducted

with our example variables. The Levene's statistic is 0.139 with a p value of 0.709. Thus, the hypothesis for equal variances is not rejected, which indicates that variances are fairly equivalent between the groups.

Figure 3.21 Descriptive Statistics for Income (*rincom_4*) by Job Satisfaction (*satjob2*).

Descriptives

	Job Satisfaction			Statistic	Std. Error
RINCOM_4	Very satisfied	Mean		2.5877	4.694E-02
		95% Confidence Interval for Mean	Lower Bound	2.4953	
			Upper Bound	2.6800	
		5% Trimmed Mean		2.5839	
		Median		2.6458	
		Variance		.690	
		Std. Deviation		.8304	
		Minimum		1.00	
		Maximum		4.24	
		Range		3.24	
		Interquartile Range		1.1623	
		Skewness		-.013	.138
		Kurtosis		-.362	.275
	Not very satisfied	Mean		2.8384	4.039E-02
		95% Confidence Interval for Mean	Lower Bound	2.7590	
			Upper Bound	2.9178	
		5% Trimmed Mean		2.8593	
		Median		2.8284	
		Variance		.648	
		Std. Deviation		.8048	
		Minimum		1.00	
		Maximum		4.24	
		Range		3.24	
		Interquartile Range		.9409	
		Skewness		-.291	.122
		Kurtosis		-.265	.244

Skewness and kurtosis values are closer to zero, indicating a more normal distribution.

Figure 3.22 Histograms for Income (*rincom_4*) by Job Satisfaction (*satjob2*).

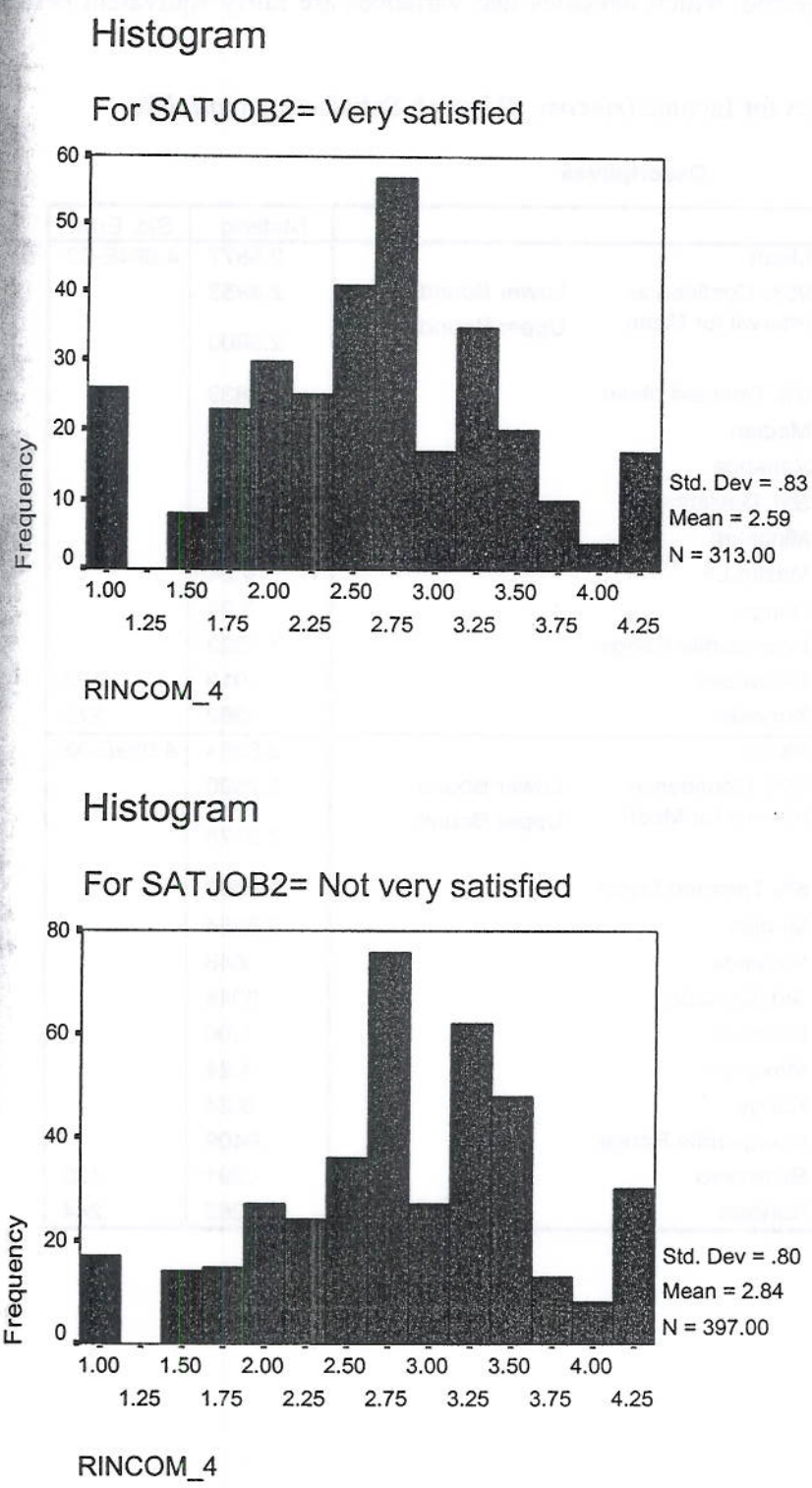
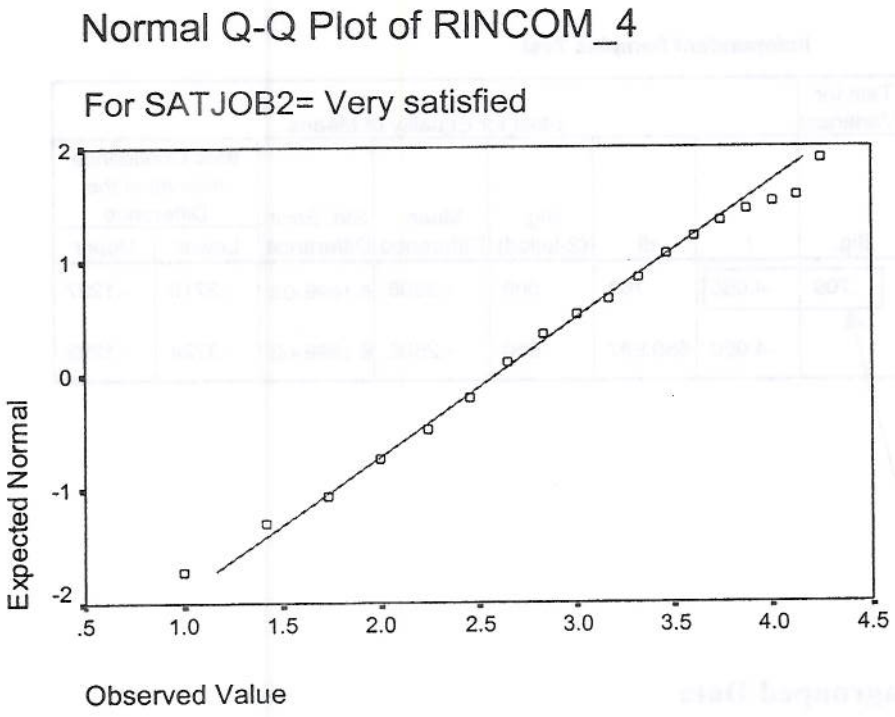


Figure 3.23 Normal Q-Q Plots for Income (*rincom_4*) by Job Satisfaction (*satjob2*).



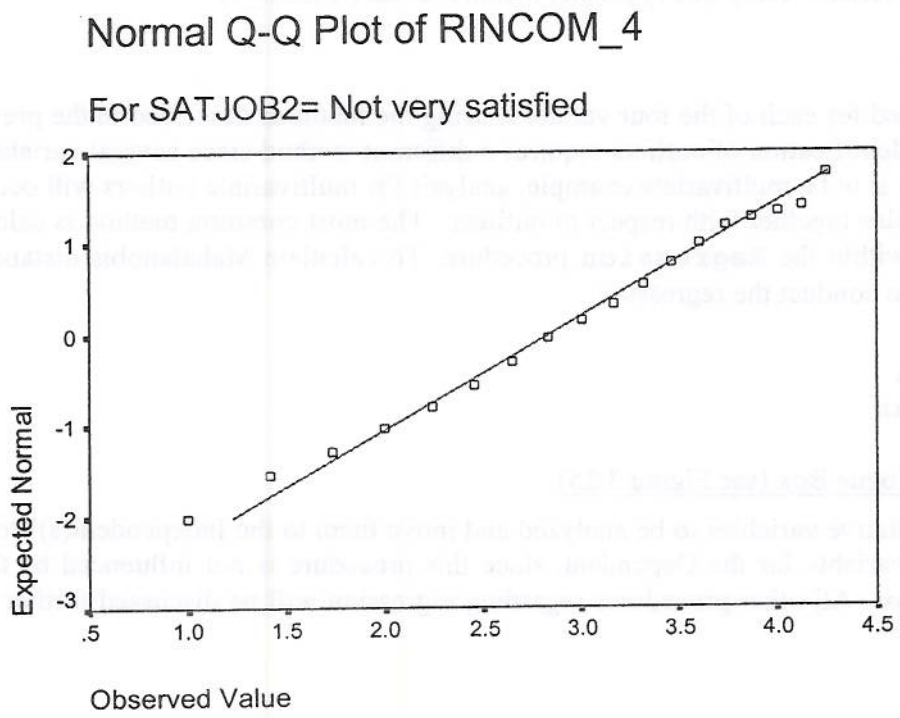


Figure 3.24 Levene's Test for Equality of Variances.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RINCOM_	Equal variance assumed	.139	.709	-4.065	708	.000	-.2508	6.169E-02	-.3719	-.1297
	Equal variance not assumed			-4.050	659.997	.000	-.2508	6.169E-02	-.3724	-.1292

Nonsignificant value indicates homogeneity of variance.

Univariate Example of Ungrouped Data

In this example, we seek to investigate the degree to which the variables of years of education (*educ*), age (*age)*, and hours worked weekly (*hrs1*) predict income levels (*rincom_4*).

Missing Data and Outliers

Missing data is analyzed for each of the four variables using the methods described in the previous example. However, the identification of outliers requires a different method since several variables are in question. Although this is not a multivariate example, analysis for multivariate outliers will occur in order to examine the variables together with respect to outliers. The most common method is calculating Mahalanobis distance within the **Regression** procedure. To calculate Mahalanobis distance, complete the following steps to conduct the regression:

Analyze
Regression
Linear

Linear Regression Dialogue Box (see Figure 3.25)

Identify all four quantitative variables to be analyzed and move them to the Independent(s) Box. Utilize a case number or ID variable for the Dependent, since this procedure is not influenced by the DV. Next, click the **Save** box. All other procedures regarding regression will be discussed further in Chapter 7.

Figure 3.25 Linear Regression Dialogue Box.

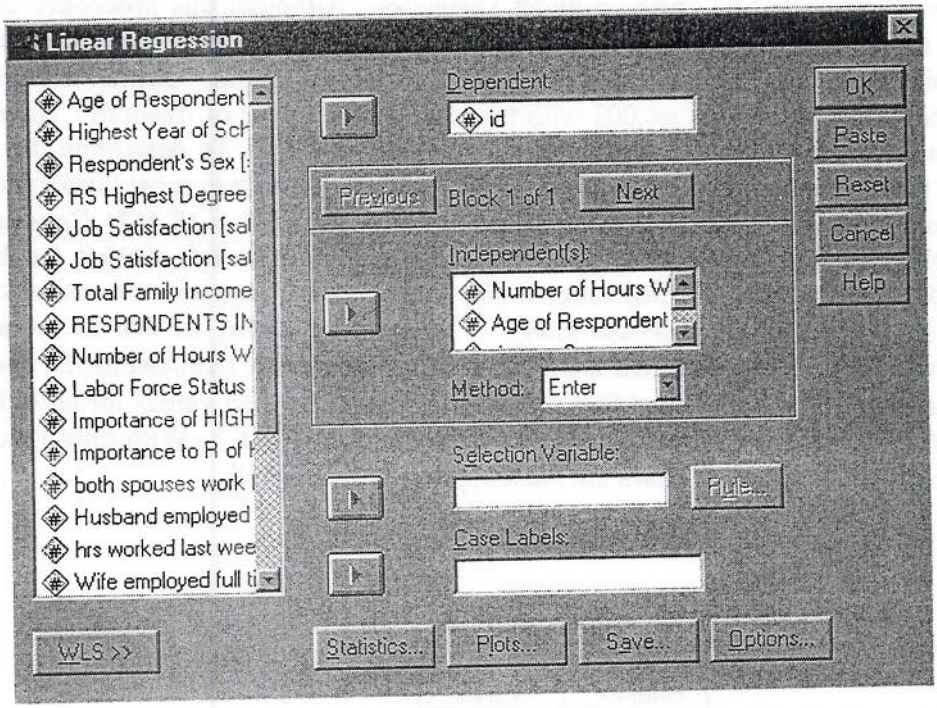
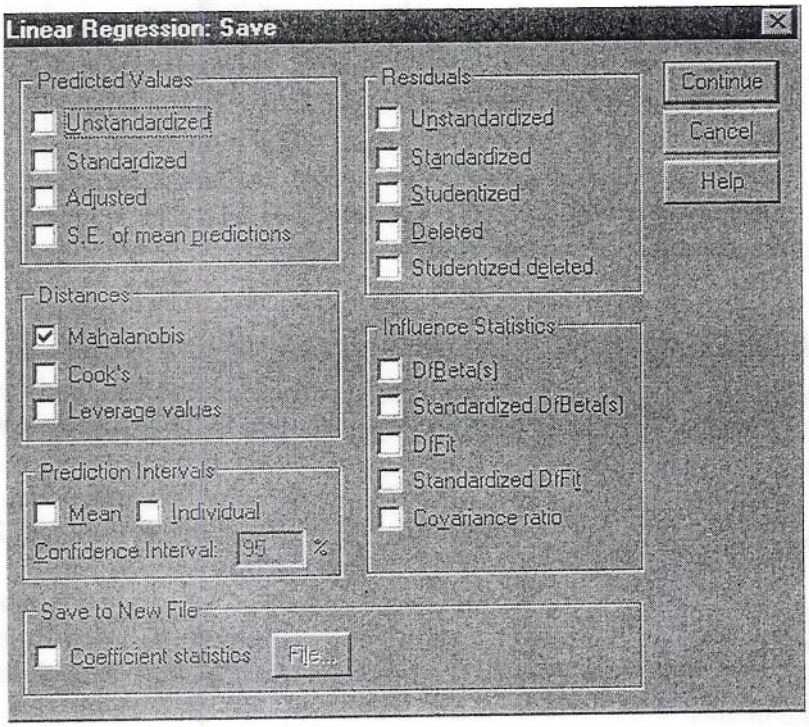


Figure 3.26 Linear Regression Save Dialogue Box.



`Linear Regression Save` Dialogue Box (see Figure 3.26)

Once in this box, check **`Mahalanobis`** under Distances. Although this procedure does not produce output that is especially helpful in identifying multivariate outliers, a new variable (*mah_1*) is created for Mahalanobis distances, which is tested using chi-square (χ^2) criteria. Outliers are indicated by chi square values that are significant at $p<.001$ with the respective degrees of freedom. The number of variables being examined for outliers is used as the degrees of freedom. To determine the critical value for χ^2, one must utilize a table of critical values for chi square, available in most introductory statistics textbooks. For our example, the critical value of χ^2 at $p<.001$ and $df=4$ is 18.467. Consequently, cases with a Mahalanobis distance greater than 18.467 are considered multivariate outliers for the variables of *rincom_4*, *age*, *educ* and *hrs1*. Identification of the outlying cases can now be easily achieved using the **`Explore`** procedure for the variable *mah_1*. Within **`Explore`**, all that is necessary is checking **`Outliers`** within the **`Statistics`** Dialogue Box, as previously demonstrated. The Table of Extreme Values (see Figure 3.27) generated lists the five highest and lowest values for *mah_1*. Four cases (#222, #24, #616, #208) are identified as outliers as they exceed 18.467. With only four multivariate outliers in the entire data set, these cases are most appropriately deleted.

Figure 3.27 Extreme Values for Mahalanobis Distance.

Extreme Values

			Case Number	Value
Mahalanobis Distance	Highest	1	222	29.93848
		2	24	22.03483
		3	616	18.71248
		4	208	18.52947
		5	729	18.15252
	Lowest	1	292	.23228
		2	146	.24275
		3	550	.30986
		4	126	.32204
		5	443	.33112

Only cases (222, 24, 616, 208) with values that exceed the critical value of chi square are considered outliers.

Normality, Linearity, and Homoscedasticity

Since our example includes several quantitative variables, univariate normality should be examined for each individual variable; however, multivariate normality will need to be assessed as well. To assess univariate normality, the **`Explore`** procedure is conducted for each of these variables. Histograms, normal Q-Q plots, and descriptive statistics reveal the following: *age* has moderate, positive skewness; *hrs1* and *educ* are fairly normal but very peaked. *Age* will be transformed into *age_2* by taking the square root of *age*. *Hrs1* and *educ* will not be transformed at this point.

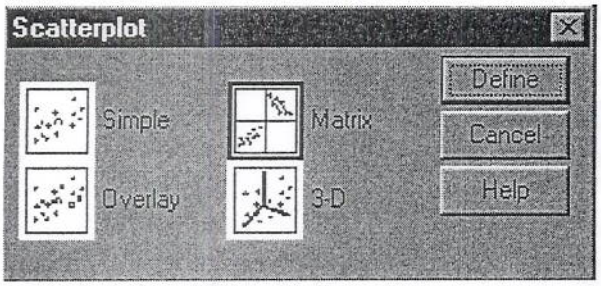
The next step is to analyze for multivariate normality and linearity. The most common method evaluating multivariate normality is creating scatterplots of all variables in relation to one another. If variable combinations are normal, scatterplots will display elliptical shapes. To create scatterplots of the four variables, open the following menus:

```
Graphs
     Scatter...
```

`Scatterplot` Dialogue Box (see Figure 3.28)

Since scatterplots will be created for several combinations of variables, click **`Matrix`**, then **`Define`**.

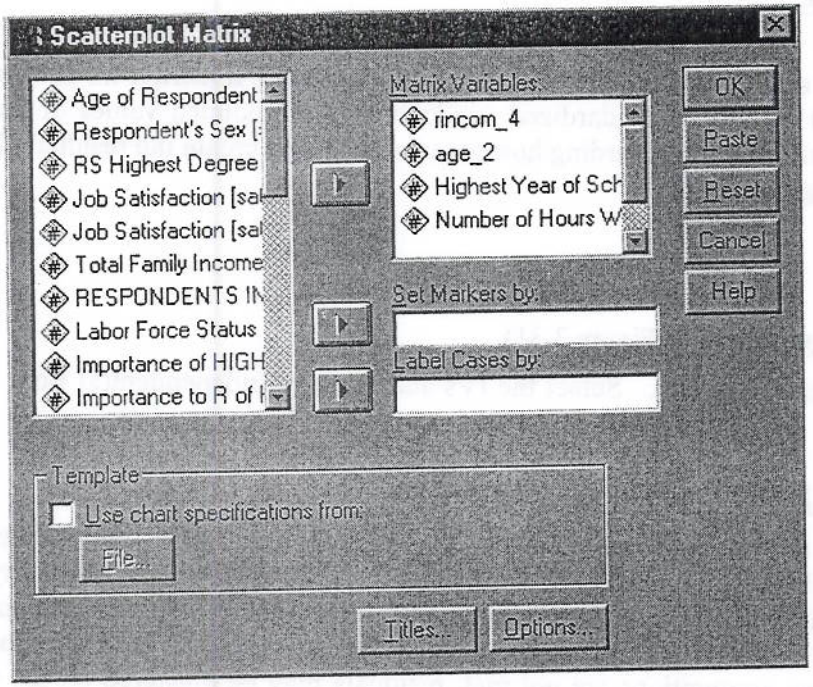
Figure 3.28 Scatterplot Dialogue Box.



`Scatterplot Matrix` Dialogue Box (see Figure 3.29)

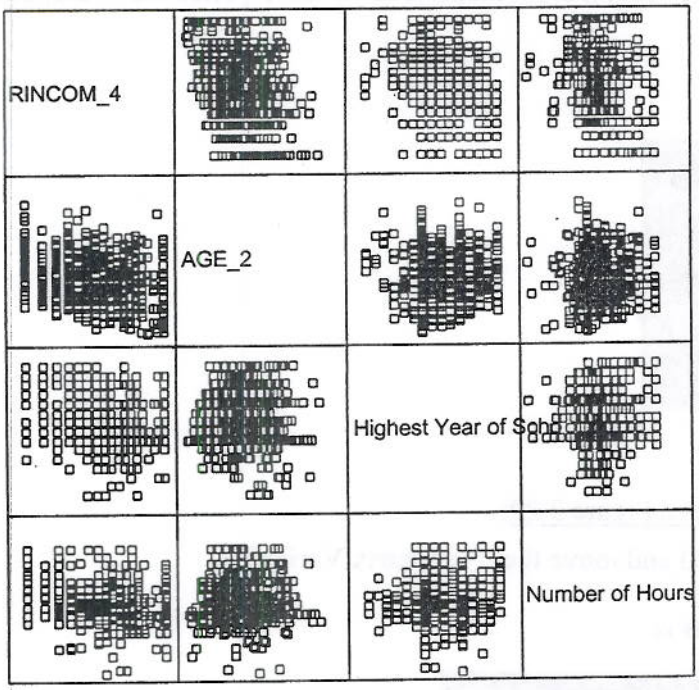
Identify the variables to be analyzed and move them to Matrix Variables.

Figure 3.29 Scatterplot Matrix Dialogue Box.



The output (see Figure 3.30) for our example displays nonelliptical shapes for all combinations, which implies failure of normality and linearity. For such a situation, two options are available: (1) re-check univariate normality for each variable; and (2) only utilize variables for univariate analyses.

Figure 3.30 Scatterplot matrix of *rincom_4, age_2, educ* and *hrs1*.



Since the use of bivariate scatterplots is fairly subjective in examining linearity, we recommend a more sophisticated method that compares standardized residuals to the predicted values of the DV. This method also provides some information regarding homoscedasticity. To create the residual plot for these variables, open the following menus:

Regression
Linear

Linear Regression Dialogue Box (see Figure 3.31)

Move *rincom_4* to the Dependent Box. Select the IVs and move to Independent(s) Box. Then click **Plot**.

Linear Regression Plot Dialogue Box (see Figure 3.32)

Within this menu, select the standardized residuals (ZRESID) for the *Y*-axis. Select the standardized predicted values (ZPRED) for the *X*-axis. When the assumptions of linearity, normality, and homoscedasticity are met, residuals will create an approximate rectangular distribution with a concentration of scores along the center. Figure 3.33 displays fairly consistent scores throughout the plot with concentration in the center. When assumptions are not met, residuals may be clustered on the top or bottom of the plot (non-normality), may be curved (nonlinearity), or may be clustered on the right or left side (heteroscedasticity). Since such extreme clustering is not displayed, we will conclude that the assumptions of normality, linearity, and homoscedasticity are met for these variables.

Figure 3.31 Linear Regression Dialogue Box.

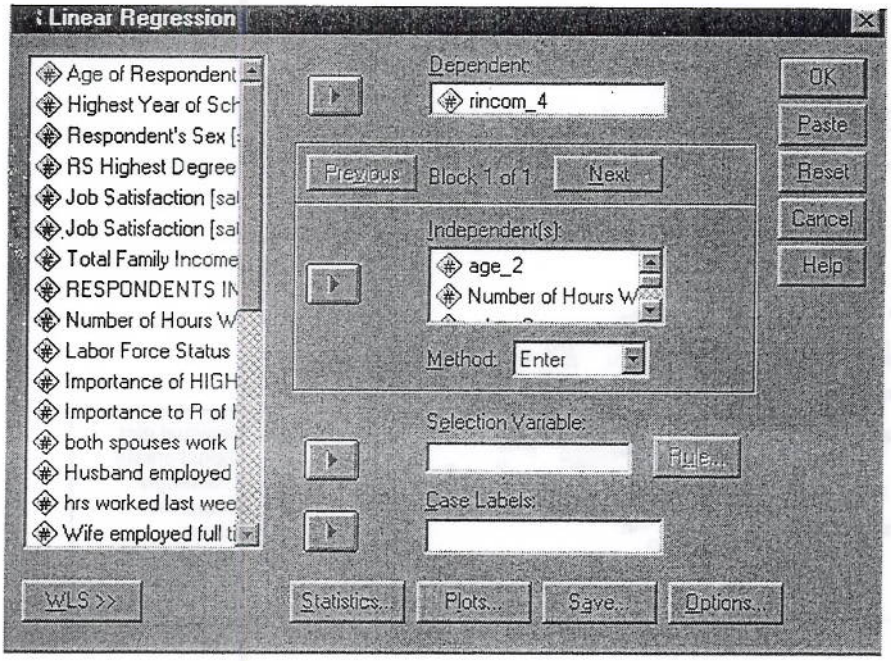


Figure 3.32 Linear Regression Plots Dialogue Box.

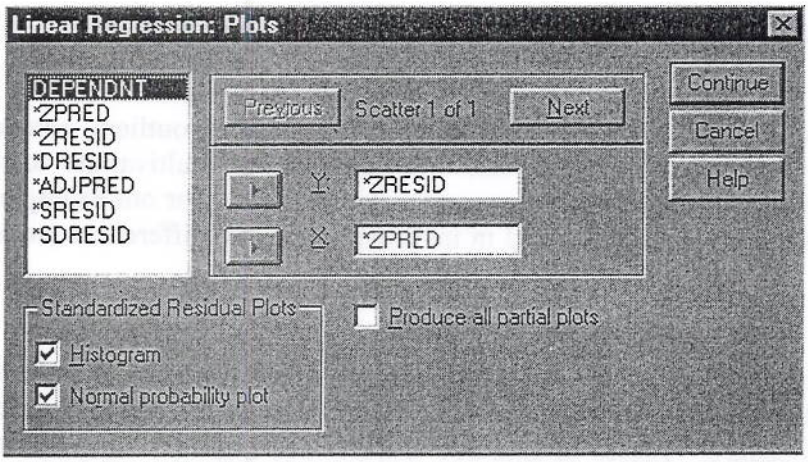
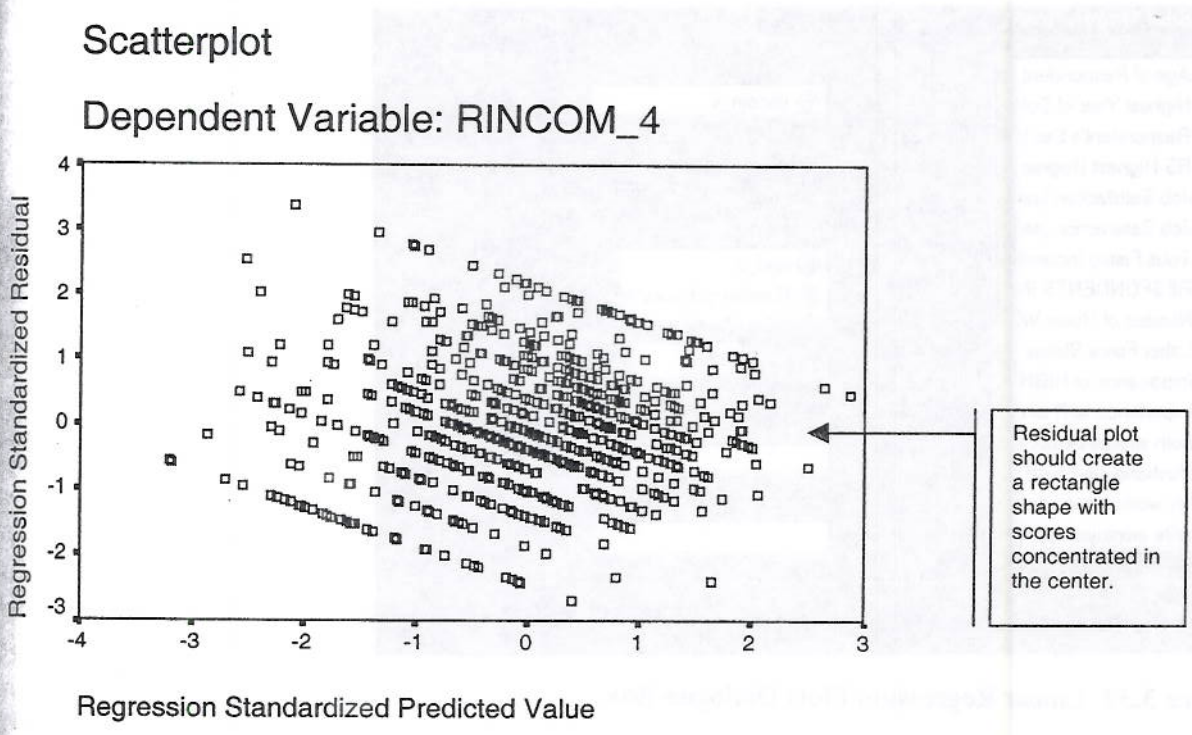


Figure 3.33 Scatterplot of Standardized Predicted Values by Standardized Residuals.



SECTION 3.8 USING SPSS TO EXAMINE GROUPED DATA FOR MULTIVARIATE ANALYSIS

The following example describes the process of examining missing values, outliers, normality, linearity, and homoscedasticity for grouped multivariate data. A non-grouped multivariate scenario would typically follow the univariate non-grouped example previously presented. For our example, we will again utilize the data set *gssft.sav*, as we are interested in investigating group differences (*satjob2*) in *rincom_4, age_2,* and *educ*.

Missing Data and Outliers

Missing data would be assessed for each variable. Multivariate outliers would be examined using Mahalanobis distances within **Regression** for *each group*. Please refer to the previous example on methods for conducting this procedure. Tables of Extreme Values (see Figure 3.34) present chi square values for each possible outlier within each group. The critical value at $p<.001$ for chi square is again 18.467 with $df=4$. Thus, the satisfied group has two outliers (#222, #24), while the unsatisfied group has three outliers (#545, #551, #575). Notice the particular cases identified as outliers are slightly different from the previous example where data was ungrouped. Identified outliers will be deleted from further analysis.

Figure 3.34 Tables of Extreme Values.

Extreme Values for *Satjob2=1*

Extreme Values

			Case Number	Value
Mahalanobis Distance	Highest	1	222	28.50217
		2	24	19.54314
		3	616	16.06943
		4	208	15.98786
		5	661	14.89095
	Lowest	1	126	.21779
		2	550	.22604
		3	741	.25174
		4	146	.25229
		5	331	.27349

Cases 222 & 24 are outliers since their values exceed chi square critical.

Extreme Values for *Satjob2=2*

Extreme Values

			Case Number	Value
Mahalanobis Distance	Highest	1	545	19.17850
		2	551	18.76909
		3	575	18.49591
		4	427	18.23602
		5	729	15.20968
	Lowest	1	292	.15434
		2	619	.37817
		3	637	.38071
		4	677	.40415
		5	527	.46097

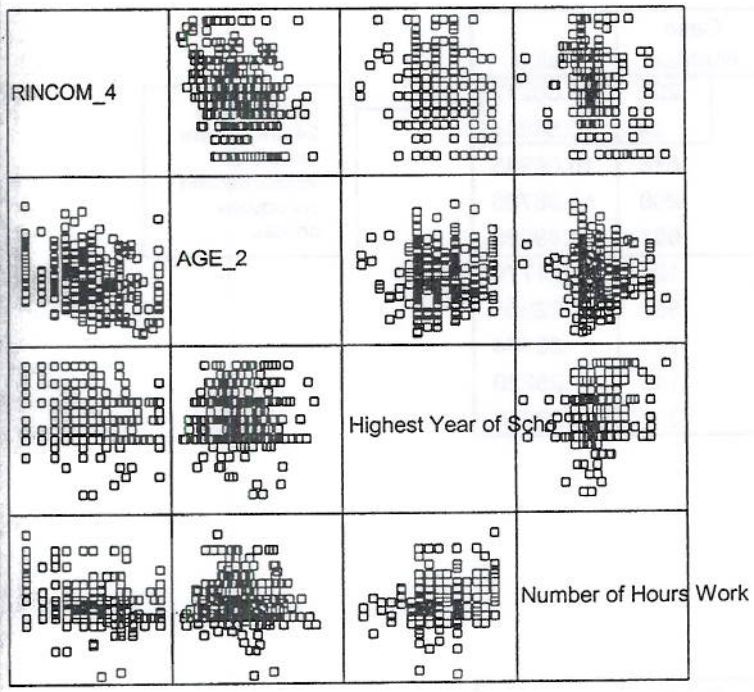
Cases 545, 551 & 575 are outliers since their values exceed chi square critical.

Normality, Linearity, and Homoscedasticity

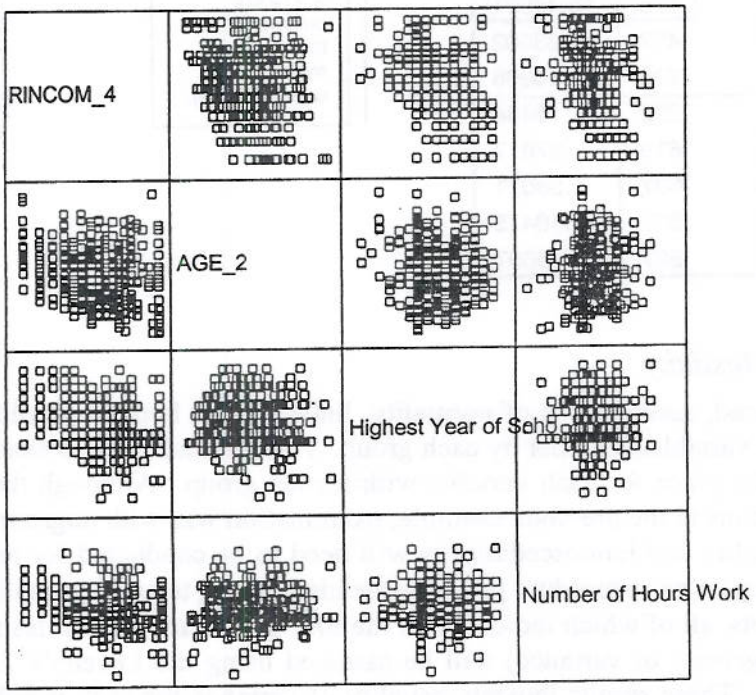
Because groups are being compared, assumptions of normality, linearity and homoscedasticity must be examined for all the quantitative variables together by each group. Prior to multivariate examination, univariate examination should take place for each variable within each group. Although these variables have been assessed for assumptions in the previous example, examination was with ungrouped data. Consequently, assessment of normality and homoscedasticity will need to be conducted for each variable within each group. Using the **`Explore`** procedure provides the histograms, tests of normality, descriptive statistics, and normal Q-Q plots, all of which indicate that the four quantitative variables are fairly normal. Homoscedasticity (homogeneity of variance) will be assessed using the Levene's Test within the *t*-test of independent samples. These results indicate equality of variance for each variable between groups.

Figure 3.35 Scatterplot Matrices of *rincom_4, age_2, educ,* and *hrs1* by *satjob2*.

satjob2=1



satjob2=2



Multivariate normality, linearity, and homoscedasticity can now be assessed. Multivariate normality and linearity are examined with a matrix of scatterplots for each group (see Figure 3.35). The results are quite similar to the previously produced scatterplot matrix of the same variables but with ungrouped data (see Figure 3.30). Although some plots display enlarged oval shapes, multivariate normality and linearity are questionable.

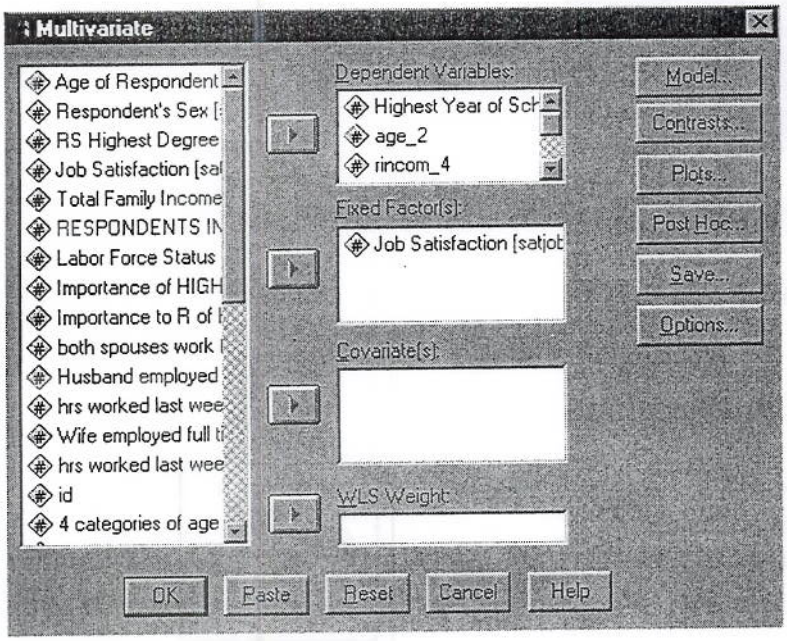
Homogeneity of variance-covariance matrices is evaluated within MANOVA by calculating Box's Test of Equality of Covariance. To do so, open the following menus:

Analyze
General Linear Model
Multivariate...

Multivariate Dialogue Box (see Figure 3.36)

Move the DVs into the Dependent Variables box. Identify the IV and move to the Fixed Factor(s) box. Once variables have been identified, click **Options**.

Figure 3.36 Multivariate Dialogue Box.



Multivariate Options Dialogue Box (see Figure 3.37)

Check Homogeneity Tests.

Since tests of homogeneity of variance-covariance matrices are quite strict, a more stringent critical value of .025 or .01 is often used rather than .05. Thus, when interpreting the results from the Box's Test (see Figure 3.38), the probability value was calculated at .044, which at the .025 level of significance would lead us to conclude that the covariance matrices for the dependent variable are fairly equivalent.

Figure 3.37 Multivariate Options Dialogue Box.

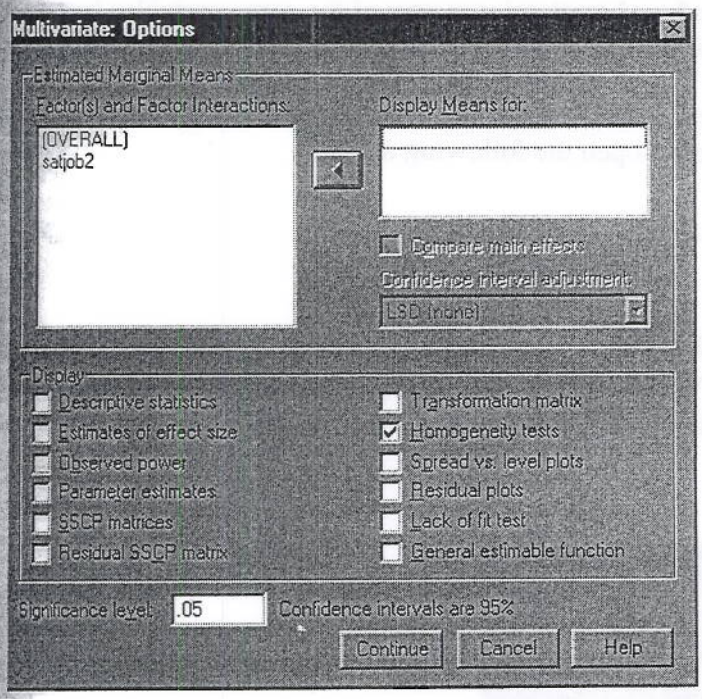


Figure 3.38 Box's Test of Equality of Covariance.

Box's Test of Equality of Covariance Matrices[a]

Box's M	18.852
F	1.873
df1	10
df2	2053031
Sig.	.044

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+SATJOB2

Significance is NOT found at .025 or .01. Equality of covariance is concluded.

Summary

Screening data for missing data, outliers, and the assumptions of normality, linearity, and homoscedasticity is an important task prior to conducting statistical analyses. If data are not screened, conclusions drawn from statistical results may be erroneous. Figure 3.39 presents the steps for univariate and multivariate examination of grouped data, while Figure 3.40 presents the process of univariate and multivariate examination of ungrouped data.

Figure 3.39 Steps for Screening Grouped Data.

Examination & Process	SPSS Procedure	Technique for "Fixing"
Missing Data • Examine missing data for each variable.	• Run **Frequency** for categorical variables 1. **Analyze...Descriptive Statistics...Frequencies** 2. Move IVs to Variables box. 3. **OK.** • Run **Descriptive** for quantitative variables 1. **Analyze...Descriptive Statistics...Frequencies** 2. Move quantitative variables to Variables box. 3. **Options.** 4. Check **Mean, Standard Deviation, Kurtosis,** and **Skewness.** 5. **Continue.** 6. **OK.**	• Less than 5% missing cases→use Listwise default. • 5-15% missing cases→replace missing values with estimated value by conducting **Transform.** 1. **Transform...Replace Missing Values.** 2. Identify variable to be transformed and move to New Variable box. 3. Identify new variable name (this occurs automatically). 4. Select method of replacement (e.g., mean, median). 5. **OK.** • More than 15% missing cases→delete variable from analysis.
Univariate Outliers • Examine outliers for quantitative variable within each group.	• Run **Explore.** 1. **Analyze...Descriptive Statistics...Explore** 2. Move DVs to Dependent Variable box. 3. Move IVs to Factor List box 4. **Statistics.** 5. Check **Descriptives** and **Outliers.** 6. **Continue.** 7. **Plots.** 8. Check **Boxplots,** and **Stem-and-leaf.** 9. **Continue.** 10. **OK.**	• More than 90-10 split between categories→ delete variable from analysis. • Small # of outliers→delete severe outliers. • Small to moderate # of outliers→replace with accepted minimum or maximum value by conducting **Recode.** 1. **Transform...Recode...Into a different variable.** 2. Select variable to be transformed and move to Input→Output Variable box. 3. Type in new variable name under Output Variable Name box. 4. **Change.** 5. **Old and New Values.** 6. Identify value to be changed under Old Value. 7. Under New Value, identify appropriate new value. 8. **Add.** 9. After all necessary values have been recoded, check **All Other Values** under Old Value. 10. Check **Copy Old Value(s)** under New Value. 11. **Add.** 12. **Continue.** 13. **OK.**
Univariate Normality • Examine normality for quantitative variable within each group.	• Run **Explore.** 1. **Analyze...Descriptive Statistics...Explore** 2. Move DVs to Dependent Variable box. 3. Move IVs to Factor List box. 4. **Statistics.** 5. Check **Descriptives.** 6. **Continue.** 7. **Plots.** 8. Check **Histograms** and **Normality plots with tests.** 9. **Continue.** 10. **OK.**	• Transform variable accordingly (see Figure 3.3) using **Compute.** 1. **Transform...Compute.** 2. Under Target, identify new variable name. 3. Identify appropriate function and move to Numeric Expression(s) box. 4. Identify variable to be transformed and move within the function equation (in place of ?). 5. **OK.**

Figure 3.39 Steps for Screening Grouped Data. *(Continued)*

Univariate Homoscedasticity • Examine homogeneity of variances between/among groups	• Conduct t-test or ANOVA using **Compare Means** to run Levene's Test.	• p value is significant at .05→ reevaluate univariate normality and consider transformations.
Multivariate Outliers • Examine quantitative variables together by group for outliers.	• Conduct **Regression** to test Mahalanobis' Distance. 1. **Analyze...Regression...Linear.** 2. Identify a variable that serves as a case number and move to Dependent Variable box. 3. Identify all appropriate quantitative variables and move to Independent(s) box. 4. **Save.** 5. Check **Mahalanobis'**. 6. **Continue.** 7. **OK.** 8. Determine chi square χ^2 critical value at p<.001. • Conduct **Explore** to test outliers for Mahalanobis chi square χ^2. 1. **Analyze...Descriptive, Statistics...Explore** 2. Move *mah_1* to Dependent Variable box. 3. Leave Factor box empty. 4. **Statistics.** 5. Check **Outliers.** 6. **Continue.** 7. **OK.**	• Delete outliers for subjects when χ^2 exceeds critical χ^2 at p<.001.
Multivariate Normality, Linearity • Examine normality and linearity of variable combinations by group.	• Create **Scatterplot Matrix** 1. **Graphs...Scatter** 2. **Matrix.** 3. **Define.** 4. Identify appropriate quantitative variables and move to Matrix Variables. 5. **OK.**	• Scatterplot shapes are not close to elliptical shapes→reevaluate univariate normality and consider transformations.
Multivariate Homogeneity of Variance-Covariance • Examine homogeneity of variance-covariance between/among groups.	• Conduct MANOVA using **Multivariate** to run homogeneity tests (see Chapter 6 for SPSS steps).	• p value is significant at .025 or .01→ reevaluate univariate normality and consider transformations.

See the next page for screening ungrouped data.

Boo !

Figure 3.40 Steps for Screening Ungrouped Data.

Examination & Process	SPSS Procedure	Technique for "Fixing"
Missing Data • Examine missing data for each variable.	• Run **Descriptive** for quantitative variables. 1. **Analyze...Descriptive Statistics...Frequencies** 2. Move quantitative variables to Variables box. 3. **Options.** 4. Check **Mean, Standard Deviation, Kurtosis,** and **Skewness.** 5. **Continue.** 6. **OK.**	• Less than 5% missing cases→use Listwise default. • 5-15% massing cases→replace missing values with estimated value by conducting **Transform.** 1. **Transform...Replace Missing Values.** 2. Identify variable to be transformed and move to New Variable box. 3. Identify new variable name (this occurs automatically). 4. Select method of replacement (e.g., mean, median). 5. **OK.** • More than 15% missing cases→delete variable from analysis.
Univariate Outliers • Examine outliers for each quantitative variable.	• Run **Explore.** 1. **Analyze...Descriptive Statistics...Explore** 2. Move DVs to Dependent Variable box. 3. Move IVs to Factor List box 4. **Statistics.** 5. Check **Descriptives** and **Outliers.** 6. **Continue.** 7. **Plots.** 8. Check **Boxplots,** and **Stem-and-leaf.** 9. **Continue.** 10. **OK.**	• Small # of outliers→delete severe outliers. • Small to moderate # of outliers→replace with accepted minimum or maximum value by conducting **Recode.** 1. **Transform...Recode...Into a different variable.** 2. Select variable to be transformed and move to Input→Output Variable box. 3. Type in new variable name under Output Variable Name box. 4. **Change.** 5. **Old and New Values.** 6. Identify value to be changed under Old Value. 7. Under New Value, identify appropriate new value. 8. **Add.** 9. After all necessary values have been recoded, check **All Other Values** under Old Value. 10. Check **Copy Old Value(s)** under New Value. 11. **Add.** 12. **Continue.** 13. **OK.**
Univariate Normality • Examine normality for each quantitative variable.	• Run **Explore.** 1. **Analyze...Descriptive Statistics...Explore** 2. Move DVs to Dependent Variable box. 3. Move IVs to Factor List box. 4. **Statistics.** 5. Check **Descriptives.** 6. **Continue.** 7. **Plots.** 8. Check **Histograms** and **Normality plots with tests.** 9. **Continue.** 10. **OK.**	• Transform variable accordingly (see Figure 3.3) using **Compute.** 1. **Transform...Compute.** 2. Under Target, identify new variable name. 3. Identify appropriate function and move to Numeric Expression(s) box. 4. Identify variable to be transformed and move within the function equation (in place of ?). 5. **OK.**

Figure 3.40 Steps for Screening Ungrouped Data. (*Continued*)

Multivariate Outliers • Examine quantitative variables together for outliers.	• Conduct **Regression** to test Mahalanobis' Distance. 1. **Analyze...Regression...Linear.** 2. Identify a variable that serves as a case number and move to Dependent Variable box. 3. Identify all appropriate quantitative variables and move to Independent(s) box. 4. **Save.** 5. Check **Mahalanobis'**. 6. **Continue.** 7. **OK.** 8. Determine chi square χ^2 critical value at $p<.001$. • Conduct **Explore** to test outliers for Mahalanobis chi square χ^2. 1. **Analyze...Descriptive Statistics...Explore** 2. Move *mah_1* to Dependent Variable box. 3. Leave Factor box empty. 4. **Statistics.** 5. Check **Outliers.** 6. **Continue.** 7. **OK.**	• Delete outliers for subjects when χ^2 exceeds critical χ^2 at $p<.001$.
Multivariate Normality, Linearity • Examine normality and linearity of variable combinations.	• Create **Scatterplot Matrix** 1. **Graphs...Scatter** 2. **Matrix.** 3. **Define.** 4. Identify appropriate quantitative variables and move to Matrix Variables. 5. **OK.**	• Scatterplot shapes are not close to elliptical shapes→reevaluate univariate normality and consider transformations.
Multivariate Homogeneity of Variance-Covariance • Examine standardized residuals to predicted values.	• Create residual plot using **Regression**. 1. **Analyze...Regression...Linear** 2. Move DV to Dependent Variable box. 3. Move IVs to Independent(s) Variable box. 4. **Plot.** 5. Select ZRESID for y-axis. 6. Select ZPRED for x-axis. 7. **Continue.** 8. **OK.**	• Residuals are clustered at the top, bottom, left, or right area in plot→ reevaluate univariate normality and consider transformations.

Exercises for Chapter 3

This exercise utilizes the data set *schools.sav*, which can be downloaded at the SPSS Web site. Open the URL: **www.spss.com/tech/DataSets.html** in your Web browser. Scroll down until you see "Data Used in SPSS Guide to Data Analysis—8.0 and 9.0" and click on the link "dataset.exe." When the "Save As" dialogue appears, select the appropriate folder and save the file. Preferably, this should be a folder created in the SPSS folder of your hard drive for this purpose. Once the file is saved, double-click the "dataset.exe" file to extract the data sets to the folder.[3]

[3]These directions have been tested on a number of computer platforms and have worked. However, it is possible that some platforms are configured in such a way that adjustments will need to be made to download the data.

1. You are interested in investigating if being above or below the median income (*medloinc*) impacts ACT means (*act94*) for schools. Complete the necessary steps to examine univariate grouped data in order to respond to the questions below. Although deletions and/or transformations may be implied from your examination, all steps will examine original variables.

   a. How many subjects have missing values for *medloinc* and *act94*?

   b. Is there a severe split in frequencies between groups?

   c. What are the cutoff values for outliers in each group?

   d. Which outlying cases should be deleted for each group?

   e. Analyzing histograms, normal Q-Q plots, and tests of normality, what is your conclusion regarding normality? If a transformation is necessary, which one would you use?

   f. Do the results from Levene's Test of Equal Variances indicate homogeneity of variance? Explain.

2. Examination of the variable of *scienc93* indicates a substantial to severe positively skewed distribution. Transform this variable using the two most appropriate methods. After examining the distributions for these transformed variables, which produced the best alteration?

3. You are interested in studying predictors (*math94me, loinc93,* and *read94me*) of the % graduating in 1994 (*grad94*).

   a. Examine univariate normality for each variable. What are your conclusions about the distributions? What transformations should be conducted?

   b. After making the necessary transformations, examine multivariate outliers using Mahalanobis' distance. What cases should be deleted?

   c. After deleting the multivariate outliers, examine multivariate normality and linearity by creating a Scatterplot Matrix.

   d. Examine the variables for homoscedasticity by creating a residuals plot (standardized vs. predicted values). What are your conclusions about homoscedasticity?