4 Cleaning Up Your Act

Screening Data Prior to Analysis

This chapter deals with a set of issues that are resolved after data are collected but before the main data analysis is run. Careful consideration of these issues is time-consuming and sometimes tedious; it is common, for instance, to spend many days in careful examination of data prior to running the main analysis that, itself, takes about 5 min. But consideration and resolution of these issues before the main analysis are fundamental to an honest analysis of the data.

The first issues concern the accuracy with which data have been entered into the data file and consideration of factors that could produce distorted correlations. Second, missing data—the bane of (almost) every researcher—are assessed and dealt with. Next, many multivariate procedures are based on assumptions; the fit between your data set and the assumptions is assessed before the procedure is applied. Transformations of variables to bring them into compliance with the requirements of analysis are considered. Outliers—cases that are extreme (outlandish)—create other headaches because solutions are unduly influenced and sometimes distorted by them. Last, perfect or near-perfect correlations among variables can threaten a multivariate analysis.

This chapter deals with issues that are relevant to most analyses. However, the issues are not all applicable to all analyses all the time; for instance, multiway frequency analysis (Chapter 16) and logistic regression (Chapter 10)—procedures that use log-linear techniques—have far fewer assumptions than the other techniques. Other analyses have additional assumptions that are not covered in this chapter. For these reasons, assumptions and limitations specific to each analysis are reviewed in the third section of the chapter describing that analysis.

There are differences in data screening for grouped and ungrouped data. If you are performing multiple regression, canonical correlation, factor analysis, or structural equation modeling, where subjects are not subdivided into groups, there is one procedure for screening data. If you are performing analysis of covariance, multivariate analysis of variance or covariance, profile analysis, discriminant analysis, or multilevel modeling where subjects are in groups, there is another procedure for screening data. Differences in these procedures are illustrated by example in Section 4.2. Other analyses (survival analysis and time-series analysis) sometimes have grouped data and often do not, so screening is adjusted accordingly.

You may find the material in this chapter difficult from time to time. Sometimes it is necessary to refer to material covered in subsequent chapters to explain some issue, material that is more understandable after those chapters are studied. Therefore, you may want to read this chapter now to get an overview of the tasks to be accomplished prior to the main data analysis and then read it again after mastering the remaining chapters.

4.1 Important Issues in Data Screening

4.1.1 Accuracy of Data File

The best way to ensure the accuracy of a data file is to proofread the original data against the computerized data file in the data window. In SAS, data are most easily viewed in the Interactive Data Analysis window. With a small data file, proofreading is highly recommended, but with a large data file, it may not be possible. In this case, screening for accuracy involves examination of descriptive statistics and graphic representations of the variables.

The first step with a large data set is to examine univariate descriptive statistics through one of the descriptive programs such as IBM SPSS FREQUENCIES, or SAS MEANS or UNIVARIATE or Interactive Data Analysis. For continuous variables, are all the values within range? Are means and standard deviations plausible? If you have discrete variables (such as categories of religious affiliation), are there any out-of-range numbers? Have you accurately programmed your codes for missing values?

4.1.2 Honest Correlations

Most multivariate procedures analyze patterns of correlation (or covariance) among variables. It is important that the correlations, whether between two continuous variables or between a dichotomous and continuous variable, be as accurate as possible. Under some rather common research conditions, correlations are larger or smaller than they should be.

4.1.2.1 Inflated Correlation

When composite variables are constructed from several individual items by pooling responses to individual items, correlations are inflated if some items are reused. Scales on personality inventories, measures of socioeconomic status, health indices, and many other variables in social and behavioral sciences are often composites of several items. If composite variables are used and they contain, in part, the same items, correlations are inflated; do not overinterpret a high correlation between two measures composed, in part, of the same items. If there is enough overlap, consider using only one of the composite variables in the analysis.

4.1.2.2 Deflated Correlation

Sample correlations may be lower than population correlations when there is restricted range in sampling of cases or very uneven splits in the categories of dichotomous variables.[1] Problems with distributions that lead to lower correlations are discussed in Section 4.1.5.

A falsely small correlation between two continuous variables is obtained if the range of responses to one or both of the variables is restricted in the sample. Correlation is a measure of the extent to which scores on two variables go up together (positive correlation) or one goes up while the other goes down (negative correlation). If the range of scores on one of the variables is narrow due to

[1] A very small coefficient of determination (standard deviation/mean) was also associated with lower correlations when computers had less computational accuracy. However, computational accuracy is so high in modern statistical packages that the problem is unlikely to occur, unless, perhaps, to astronomers.

restricted sampling, then it is effectively a constant and cannot correlate highly with another variable. In a study of success in graduate school, for instance, quantitative ability could not emerge as highly correlated with other variables if all students had about the same high scores in quantitative skills.

When a correlation is too small because of restricted range in sampling, you can estimate its magnitude in a nonrestricted sample by using Equation 4.1 if you can estimate the standard deviation in the nonrestricted sample. The standard deviation in the nonrestricted sample is estimated from prior data or from knowledge of the population distribution.

$$\tilde{r}_{xy} = \frac{r_{t(xy)}\left[S_x / S_{t(x)}\right]}{\sqrt{1 + r^2_{t(xy)}\left[S_x / S_{t(x)}\right] - r^2_{t(xy)}}} \tag{4.1}$$

where $r_{t(xy)}$ = adjusted correlation
$r^2_{t(xy)}$ = correlation between X and Y with the range of X truncated
S_x = unrestricted standard deviation of X
$S_{t(x)}$ = truncated standard deviation of X

Many programs allow analysis of a correlation matrix instead of raw data. The estimated correlation is inserted in place of the truncated correlation prior to analysis of the correlation matrix. (However, insertion of estimated correlations may create internal inconsistencies in the correlation matrix, as discussed in Section 4.1.3.3.)

The correlation between a continuous variable and a dichotomous variable, or between two dichotomous variables (unless they have the same peculiar splits), is also too low if most (say, over 90%) responses to the dichotomous variable fall into one category. Even if the continuous and dichotomous variables are strongly related in the population, the highest correlation that could be obtained is well below 1. Some recommend dividing the obtained (but deflated) correlation by the maximum it could achieve given the split between the categories and then using the resulting value in subsequent analyses. This procedure is attractive, but not without hazard, as discussed by Comrey and Lee (1992).

4.1.3 Missing Data

Missing data is one of the most pervasive problems in data analysis. The problem occurs when rats die, equipment malfunctions, respondents become recalcitrant, or somebody goofs. Its seriousness depends on the pattern of missing data, how much is missing, and why it is missing. Summaries of issues surrounding missing data are provided by Schafer and Graham (2002) and by Graham, Cumsille, and Elek-Fisk (2003).

The pattern of missing data is more important than the amount missing. Missing values scattered randomly through a data matrix pose less serious problems. Nonrandomly missing values, on the other hand, are serious no matter how few of them there are because they affect the generalizability of results. Suppose that in a questionnaire with both attitudinal and demographic questions, several respondents refuse to answer questions about income. It is likely that refusal to answer questions about income is related to attitude. If respondents with missing data on income are deleted, the sample values on the attitude variables are distorted. Some method of estimating income is needed to retain the cases for analysis of attitude.

Missing data are characterized as MCAR (missing completely at random), MAR (missing at random, called ignorable nonresponse), and MNAR (missing not at random or nonignorable).

The distribution of missing data is unpredictable in MCAR, the best of all possible worlds if data must be missing. The pattern of missing data is predictable from other variables in the data set when data are MAR. In NMAR, the missingness is related to the variable itself and, therefore, cannot be ignored.

If only a few data points—say, 5% or less—are missing in a random pattern from a large data set, the problems are less serious and almost any procedure for handling missing values yields similar results. If, however, a lot of data are missing from a small to moderately sized data set, the problems can be very serious. Unfortunately, there are as yet no firm guidelines for how much missing data can be tolerated for a sample of a given size.

Although the temptation to assume that data are missing randomly is nearly overwhelming, the safest thing to do is to test it. Use the information you have to test for patterns in missing data. For instance, construct a dummy variable with two groups, cases with missing and nonmissing values on income, and perform a test of mean differences in attitude between the groups. If there are no differences, decisions about how to handle missing data are not so critical (except, of course, for inferences about income). If there are significant differences and η^2 is substantial (cf. Equation 3.25), care is needed to preserve the cases with missing values for other analyses, as discussed in Section 4.1.3.2.

IBM SPSS MVA (Missing Values Analysis) is specifically designed to highlight patterns of missing values as well as to replace them in the data set. Table 4.1 shows syntax and output for a data set with missing values on ATTHOUSE and INCOME. A TTEST is requested to see if missingness is related to any of other variables, with $\alpha = .05$ and tests done only for variables with at least 5 PERCENT of data missing. The EM syntax requests a table of correlations and a test of whether data are missing completely at random (MCAR).

The Univariate Statistics table shows that there is one missing value on ATTHOUSE and 26 missing values on INCOME. Separate Variance *t* Tests show no systematic relationship between missingness on INCOME and any of the other variables. ATTHOUSE is not tested because fewer than 5% of the cases have missing values. The Missing Patterns table shows that Case number 52, among others, is missing INCOME, indicated by an S in the table. Case number 253 is missing ATTHOUSE. The last table shows EM Correlations with missing values filled in using the EM method, to be discussed. Below the table is Little's MCAR test of whether the data are missing completely at random. A statistically nonsignificant result is desired: $p = .76$ indicates that the probability that the pattern of missing diverges from randomness is greater than .05, so that MCAR may be inferred.

MAR can be inferred if the MCAR test is statistically significant but missingness is predictable from variables (other than the DV) as indicated by the Separate Variance t Tests. MNAR is inferred if the *t* test shows that missingness is related to the DV.

The decision about how to handle missing data is important. At best, the decision is among several bad alternatives, several of which are discussed in the subsections that follow.

4.1.3.1 Deleting Cases or Variables

One procedure for handling missing values is simply to drop any cases with them. If only a few cases have missing data and they seem to be a random subsample of the whole sample, deletion is a good alternative. Deletion of cases with missing values is the default option for most programs in the IBM SPSS and SAS packages.[2]

[2]Because this is the default option, numerous cases can be deleted without the researcher's knowledge. For this reason, it is important to check the number of cases in your analyses to make sure that all of the desired cases are used.

TABLE 4.1 IBM SPSS MVA Syntax and Output for Missing Data

```
MVA VARIABLES=timedrs attdrug atthouse income emplmnt mstatus race
/TTEST NO PROB PERCENT=5
/MPATTERN
/EM(TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25).
```

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a,b] Low	No. of Extremes[a,b] High
TIMEDRS	465	7.90	10.948	0	.0	0	34
ATTDRUG	465	7.69	1.156	0	.0	0	0
ATTHOUSE	464	23.54	4.484	1	.2	4	0
INCOME	439	4.21	2.419	26	5.6	0	0
EMPLMNT	465	.47	.500	0	.0	0	0
MSTATUS	465	1.78	.416	0	.0	.	.
RACE	465	1.09	.284	0	.0	.	.

a. Indicates that the interquartile range (IQR) is zero.

b. Number of cases outside the range (Q1 − 1.5*IQR, Q3 + 1.5*IQR)

Separate Variance t Tests[a]

		TIMEDRS	ATTDRUG	ATTHOUSE	INCOME	MSTATUS	RACE	EMPLMNT
INCOME	t	.2	-1.1	-.2	.	-1.0	-.4	-1.1
	df	32.2	29.6	28.6	.	29.0	27.3	28.0
	# Present	439	439	438	439	439	439	439
	# Missing	26	26	26	0	26	26	26
	Mean (Present)	7.92	7.67	23.53	4.21	1.77	1.09	.46
	Mean (Missing)	7.62	7.88	23.69	.	1.85	1.12	.58

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

TABLE 4.1 Continued

Missing Patterns (cases with missing values)

Case	# Missing	% Missing	Missing and Extreme Value Patterns[a] TIMEDRS	ATTDRUG	EMPLMNT	MSTATUS	RACE	ATTHOUSE	INCOME
52	1	14.3				−	−		S
64	1	14.3	+			−	−		S
69	1	14.3				−	−		S
77	1	14.3				−	−		S
118	1	14.3				−	−		S
135	1	14.3				−	−		S
161	1	14.3				−	−		S
172	1	14.3				−	−		S
173	1	14.3				−	−		S
174	1	14.3				−	−		S
181	1	14.3				−	−		S
196	1	14.3				−	+		S
203	1	14.3	+			−	−		S
236	1	14.3				−	−		S
240	1	14.3				−	−		S
258	1	14.3				−	+		S
304	1	14.3				−	−		S
321	1	14.3				−	−		S
325	1	14.3				−	−		S
352	1	14.3				−	−		S
378	1	14.3				−	−		S
379	1	14.3				−	−		S
409	1	14.3	+			−	−		S
419	1	14.3				−	−		S
421	1	14.3				−	−		S
435	1	14.3				−	+	−	S
253	1	14.3				−	+	S	

− Indicates an extreme low value, and + indicates an extreme high value. The range used is (Q1 − 1.5*IQR, Q3 + 1.5*IQR).

a.Cases and variables are sorted on missing patterns.

(*continued*)

TABLE 4.1 Continued

EM Correlations[a]

	timedrs	attdrug	atthouse	income	emplmnt	mstatus	race
timedrs	1						
attdrug	.104	1					
atthouse	.128	.023	1				
income	.050	−.005	.002	1			
emplmnt	.059	.085	−.023	−.006	1		
mstatus	−.065	−.006	−.030	−.466	.234	1	
race	−.035	.019	−.038	.105	−.081	−.035	1

a. Little's MCAR test: Chi-Square = 19.550, DF = 12, Sig. = 0.76.

If missing values are concentrated in a few variables and those variables are not critical to the analysis, or are highly correlated with other, complete variables, the variable(s) with missing values are profitably dropped.

But if missing values are scattered throughout cases and variables, deletion of cases can mean substantial loss of subjects. This is particularly serious when data are grouped in an experimental design because loss of even one case requires adjustment for unequal *n* (see Chapter 6). Further, the researcher who has expended considerable time and energy collecting data is not likely to be eager to toss some out. And as previously noted, if cases with missing values are not randomly distributed through the data, distortions of the sample occur if they are deleted.

4.1.3.2 Estimating Missing Data

A second option is to estimate (impute) missing values and then use the estimates during data analysis. There are several popular schemes for doing so: using prior knowledge; inserting mean values; using regression; expectation-maximization; and multiple imputation.

Prior knowledge is used when a researcher replaces a missing value with a value from an educated guess. If the researcher has been working in an area for a while, and if the sample is large and the number of missing values small, this is often a reasonable procedure. The researcher is often confident that the value would have been about at the median, or whatever, for a particular case. Alternatively, the researcher can downgrade a continuous variable to a dichotomous variable (e.g., "high" vs. "low") to predict with confidence into which category a case with a missing value falls. The discrete variable replaces the continuous variable in the analysis, but it has less information than the continuous variable. An option with longitudinal data is to apply the last observed value to fill in data missing at a later point in time. However, this requires the expectation that there are no changes over time.

Remaining options for imputing missing data are available through software. Table 4.2 shows programs that implement missing data procedures.

TABLE 4.2 Missing Data Options Available in Some Computer Programs

Strategy		Program						
		SPSS MVA	SOLAS MDA	SPSS REGRESSION	NORM	SAS STANDARD	SAS MI and MIANALYZE	AMOS, EQS, and LISREL
Mean substitution	Grand mean	No	Group Means[a]	MEAN SUBSTITUTION	No	REPLACE	No	No
	Group mean	No	Group Means	No	No	No	No	No
Regression		Regression	No	No	No	No	No	No
Expectation Maximization (EM/FIML)		EM	No	No	Yes[c]	No	PROC MI with NIMPUTE =	Yes
Multiple imputation		No[b]	Multiple Imputation	No	Yes	No	Yes	No

[a]With omission of group identification.

[b]May be done by generating multiple files through the EM method and computing additional statistics.

[c]For preparation of data prior to multiple imputation; provides missing values for one random imputation.

Mean substitution has been a popular way to estimate missing values, although it is less commonly used now that more desirable methods are feasible through computer programs. Means are calculated from available data and are used to replace missing values prior to analysis. In the absence of all other information, the mean is the best guess about the value of a variable. Part of the attraction of this procedure is that it is conservative; the mean for the distribution as a whole does not change and the researcher is not required to guess at missing values. On the other hand, the variance of a variable is reduced because the mean is closer to itself than to the missing value it replaces, and the correlation the variable has with other variables is reduced because of the reduction in variance. The extent of loss in variance depends on the amount of missing data and on the actual values that are missing.

A compromise is to insert a group mean for the missing value. If, for instance, the case with a missing value is a Republican, the mean value for Republicans is computed and inserted in place of the missing value. This procedure is not as conservative as inserting overall mean values and not as liberal as using prior knowledge. However, the reduction in within-group variance can make differences among groups spuriously large.

Many programs have provisions for inserting mean values. SAS STANDARD allows a data set to be created with missing values replaced by the mean on the variable for complete cases. SOLAS MDA, a program devoted to missing data analysis, produces data sets in which group means are used to replace missing values IBM SPSS REGRESSION permits MEANSUBSTITUTION. And, of course, transformation instructions can be used with any program to replace any defined value of a variable (including a missing code) with the mean.

Regression (see Chapter 5) is a more sophisticated method for estimating missing values. Other variables are used as IVs to write a regression equation for the variable with missing data

serving as DV. Cases with complete data generate the regression equation; the equation is then used to predict missing values for incomplete cases. Sometimes, the predicted values from a first round of regression are inserted for missing values and then all the cases are used in a second regression. The predicted values for the variable with missing data from round two are used to develop a third equation, and so forth, until the predicted values from one step to the next are similar (they converge). The predictions from the last round are the ones used to replace missing values.

An advantage of regression is that it is more objective than the researcher's guess but not as blind as simply inserting the grand mean. One disadvantage to the use of regression is that the scores fit together better than they should; because the missing value is predicted from other variables, it is likely to be more consistent with them than a real score is. A second disadvantage is reduced variance because the estimate is probably too close to the mean. A third disadvantage is the requirement that good IVs be available in the data set; if none of the other variables is a good predictor of the one with missing data, the estimate from regression is about the same as simply inserting the mean. Finally, estimates from regression are used only if the estimated value falls within the range of values for complete cases; out-of-range estimates are not acceptable. Using regression to estimate missing values is convenient in IBM SPSS MVA. The program also permits adjustment of the imputed values so that overconsistency is lessened.

Expectation maximization (EM) methods are available for randomly missing data. EM forms a missing data correlation (or covariance) matrix by assuming the shape of a distribution (such as normal) for the partially missing data and basing inferences about missing values on the likelihood under that distribution. It is an iterative procedure with two steps—expectation and maximization—for each iteration. First, the E step finds the conditional expectation of the "missing" data, given the observed values and current estimate of the parameters, such as correlations. These expectations are then substituted for the missing data. Second, the M step performs maximum likelihood estimation (often referred to as FIML—full information maximum likelihood) as though the missing data had been filled in. Finally, after convergence is achieved, the EM variance–covariance matrix may be provided and/or the filled-in data saved in the data set.

However, as pointed out by Graham et al. (2003), analysis of an EM-imputed data set is biased because error is not added to the imputed data set. Thus, analyses based on this data set have inappropriate standard errors for testing hypotheses. The bias is greatest when the data set with imputed values filled in is analyzed, but bias exists even when a variance–covariance or correlation matrix is used as input. Nevertheless, these imputed data sets can be useful for evaluating assumptions and for exploratory analyses that do not employ inferential statistics. Analysis of EM-imputed data sets can also provide insights when amounts of missing data are small if inferential statistics are interpreted with caution.

IBM SPSS MVA performs EM to generate a data set with imputed values as well as variance–covariance and correlation matrices, and permits specification of some distributions other than normal. IBM SPSS MVA also is extremely helpful for assessing patterns of missing data, providing *t* tests to predict missingness from other variables in the data set, and testing for MCAR, as seen in Table 4.1.

Structural equations modeling (SEM) programs (cf. AMOS, EQS, and LISREL in Chapter 14) typically have their own built-in imputation procedures which are based on EM. The programs do not produce data sets with imputed values but utilize appropriate standard errors in their analyses.

SAS, NORM, and SOLAS MDA can be used to create an EM-imputed data set by running the MI (multiple imputation, see below) procedure with *m* (number of imputations) = 1, but analysis

of the imputed data set is subject to the same cautions as noted for IBM SPSS MVA. The EM variance–covariance matrix produced by NORM builds in appropriate standard errors, so that analyses based on those matrices are unbiased (Graham et al., 2003). Little and Rubin (1987) discuss EM and other methods in detail. EM through IBM SPSS MVA is demonstrated in Section 10.7.1.1.

Multiple imputation also takes several steps to estimate missing data. First, logistic regression (Chapter 10) is used when cases with and without a missing value on a particular variable form the dichotomous DV. You determine which other variables are to be used as predictors in the logistic regression, which in turn provides an equation for estimating the missing values. Next, a random sample is taken (with replacement) from the cases with complete responses to identify the distribution of the variable with missing data.

Then several (*m*) random samples are taken (with replacement) from the distribution of the variable with missing data to provide estimates of that variable for each of *m* newly created (now complete) data sets. Rubin (1996) shows that five (or even three in some cases) such samples are adequate in many situations. You then perform your statistical analysis separately on the *m* new data sets and report the average parameter estimates (e.g., regression coefficients) from the multiple runs in your results.

Advantages of multiple imputation are that it can be applied to longitudinal data (e.g., for within-subjects IVs or time-series analysis) as well as data with single observations on each variable, and that it retains sampling variability (Statistical Solutions, Ltd., 1997). Another advantage is that it makes no assumptions about whether data are randomly missing. This is the method of choice for databases that are made available for analyses outside the agency that collected the data. That is, multiple data sets are generated, and other users may either make a choice of a single data set (with its inherent bias) or use the multiple data sets and report combined results. Reported results are the mean for each parameter estimate over the analyses of multiple data sets as well as the total variance estimate, which includes variance within imputations and between imputations—a measure of the true uncertainty in the data set caused by missing data (A. McDonnell, personal communication, August 24, 1999).

SOLAS MDA performs multiple imputation directly and provides a ROLLUP editor that combines the results from the newly created complete data sets (Statistical Solutions, Ltd., 1997). The editor shows the mean for each parameter and its total variance estimate, as well as within- and between-imputations variance estimates. The SOLAS MDA manual demonstrates multiple imputations with longitudinal data. Rubin (1996) provides further details about the procedure. With IBM SPSS MVA, you apply your method *m* times via the EM procedure, using a random-number seed that changes for each new data set. Then you do your own averaging to establish the final parameter estimate.

NORM is a freely distributed program for multiple imputation available on the Internet (Schafer, 1999). The program currently is limited to normally distributed predictors and encompasses an EM procedure to estimate parameters, provide start values for the data augmentation step (multiple imputation), and help determine the proper number of imputations. A summary of the results of the multiple data sets produced by data augmentation is available as are the multiple data sets.

Newer versions of SAS use a three-step process to deal with multiple imputation of missing data. First, PROC MI provides an analysis of missing data patterns much like that of IBM SPSS MVA (Table 4.1) but without the MCAR diagnosis or *t* tests to predict missingness from other variables. At the same time, a data set is generated with *m* subsets (default $m = 5$) with different imputations of missing data in each subset. A column indicating the imputation number (e.g., 1 to 5)

is added to the data set. Next, the desired analysis module is run (e.g., REG, GLM, or MIXED) with syntax that requests a separate analysis for each imputation number and with some (but not all) results added to a data file. Finally, PROC MIANALYZE is run on the data file of results which combines the m sets of results into a single summary report.[3] Graham and colleagues (2003) report that m typically ranges from 5 to 20. Rubin (1996) suggests that often 3 to 5 imputations are adequate, as long as the amount of missing data is fairly small. He also asserts that any $m > 1$ is better than just one imputed data set. SAS MI and MIANALYZE are demonstrated in Section 5.7.4, where guidelines are given for choice of m.

IBM SPSS MVA may be used to create multiple data sets with different imputations by running EM imputation m times with different random number seeds. However, you are then on your own for combining the results of the multiple analyses, and there is no provision for developing appropriate standard errors. Rubin (1987) discusses multiple imputation at length.

Other methods, such as *hot decking,* are available but they require specialized software and have few advantages in most situations over other imputation methods offered by SAS, SOLAS, and NORM.

4.1.3.3 *Using a Missing Data Correlation Matrix*

Another option with randomly missing data involves analysis of a missing data correlation matrix. In this option, all available pairs of values are used to calculate each of the correlations in **R**. A variable with 10 missing values has all its correlations with other variables based on 10 fewer pairs of numbers. If some of the other variables also have missing values, but in different cases, the number of complete pairs of variables is further reduced. Thus, each correlations in **R** can be based on a different number and a different subset of cases, depending on the pattern of missing values. Because the standard error of the sampling distribution for r is based on N, some correlations are less stable than others in the same correlation matrix.

But that is not the only problem. In a correlation matrix based on complete data, the sizes of some correlations place constraints on the sizes of others. In particular,

$$r_{13}r_{23} - \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \leq r_{12} \leq r_{13}r_{23} + \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \qquad (4.2)$$

The correlation between variables 1 and 2, r_{12}, cannot be smaller than the value on the left or larger than the value on the right in a three-variable correlation matrix. If $r_{13} = .60$ and $r_{23} = .40$ then r_{12} cannot be smaller than $-.49$ or larger than .97. If, however, r_{12}, r_{23}, and r_{13} are all based on different subsets of cases due to missing data, the value for r_{12} can go out of range.

Most multivariate statistics involve calculation of eigenvalues and eigenvectors from a correlation matrix (see Appendix A). With loosened constraints on size of correlations in a missing data correlation matrix, eigenvalues sometimes become negative. Because eigenvalues represent variance, negative eigenvalues represent something akin to negative variance. Moreover, because the total variance that is partitioned in the analysis is a constant (usually equal to the number of variables), positive eigenvalues are inflated by the size of negative eigenvalues, resulting in inflation of variance. The statistics derived under these conditions can be quite distorted.

However, with a large sample and only a few missing values, eigenvalues are often all positive even if some correlations are based on slightly different pairs of cases. Under these conditions, a missing data correlation matrix provides a reasonable multivariate solution and has the advantage of using all available data. Use of this option for the missing data problem should not be rejected out of hand but should be used cautiously with a wary eye to negative eigenvalues.

A missing value correlation matrix is prepared through the PAIRWISE deletion option in some of the IBM SPSS programs. It is the default option for SAS CORR. If this is not an option of the program you want to run, then generate a missing data correlation matrix through another program for input to the one you are using.

[3]Output is especially sparse for procedures other than SAS REG.

4.1.3.4 *Treating Missing Data as Data*

It is possible that the fact that a value is missing is itself a very good predictor of the variable of interest in your research. If a dummy variable is created when cases with complete data are assigned 0 and cases with missing data 1, the liability of missing data could become an asset. The mean is inserted for missing values so that all cases are analyzed, and the dummy variable is used as simply another variable in analysis, as discussed by Cohen, Cohen, West, and Aiken (2003, pp. 431–451).

4.1.3.5 *Repeating Analyses With and Without Missing Data*

If you use some method of estimating missing values or a missing data correlation matrix, consider repeating your analyses using only complete cases. This is particularly important if the data set is small, the proportion of missing values high, or data are missing in a nonrandom pattern. If the results are similar, you can have confidence in them. If they are different, however, you need to investigate the reasons for the difference, and either evaluate which result more nearly approximates "reality" or report both sets of results.

4.1.3.6 *Choosing Among Methods for Dealing With Missing Data*

The first step in dealing with missing data is to observe their pattern to try to determine whether data are randomly missing. Deletion of cases is a reasonable choice if the pattern appears random and if only a very few cases have missing data and those cases are missing data on different variables. However, if there is evidence of nonrandomness in the pattern of missing data, methods that preserve all cases for further analysis are preferred.

Deletion of a variable with a lot of missing data is also acceptable as long as that variable is not critical to the analysis. Or, if the variable is important, use a dummy variable that codes the fact that the scores are missing coupled with mean substitution to preserve the variable and make it possible to analyze all cases and variables.

It is best to avoid mean substitution unless the proportion of missing values is *very* small and there are no other options available to you. Using prior knowledge requires a great deal of confidence on the part of the researcher about the research area and expected results. Regression methods may be implemented (with some difficulty) without specialized software but are less desirable than EM methods.

EM methods sometimes offer the simplest and most reasonable approach to imputation of missing data, as long as your preliminary analysis provides evidence that scores are missing randomly (MCAR or MAR). Use of an EM covariance matrix, if the technique permits it as input, provides a less biased analysis a data set with imputed values. However, unless the EM program provides appropriate standard errors (as per the SEM programs of Chapter 14 or NORM), the strategy should be limited to data sets in which there is not a great deal of missing data, and inferential

results (e.g., p values) are interpreted with caution. EM is especially appropriate for techniques that do not rely on inferential statistics, such as exploratory factor analysis (Chapter 13). Better yet is to incorporate EM methods into multiple imputation.

Multiple imputation is currently considered the most respectable method of dealing with missing data. It has the advantage of not requiring MCAR (and perhaps not even MAR) and can be used for any form of GLM analysis, such as regression, ANOVA, and logistic regression. The problem is that it is more difficult to implement and does not provide the full richness of output that is typical with other methods.

Using a missing data correlation matrix is tempting if your software offers it as an option for your analysis because it requires no extra steps. It makes most sense to use when missing data are scattered over variables, and there are no variables with a lot of missing values. The vagaries of missing data correlation matrices should be minimized as long as the data set is large and missing values are few.

Repeating analyses with and without missing data is highly recommended whenever any imputation method or a missing data correlation matrix is used and the proportion of missing values is high—especially if the data set is small.

4.1.4 Outliers

An outlier is a case with such an extreme value on one variable (a univariate outlier) or such a strange combination of scores on two or more variables (multivariate outlier) that it distorts statistics. Consider, for example, the bivariate scatterplot of Figure 4.1 in which several regression lines, all with slightly different slopes, provide a good fit to the data points inside the swarm. But when the data point labeled A in the upper right-hand portion of the scatterplot is also considered, the regression coefficient that is computed is the one from among the several good alternatives that provides the best fit to the extreme case. The case is an outlier because it has much more impact on the value of the regression coefficient than any of those inside the swarm.

Outliers are found in both univariate and multivariate situations, among both dichotomous and continuous variables, among both IVs and DVs, and in both data and results of analyses. They lead to both Type I and Type II errors, frequently with no clue as to which effect they have in a

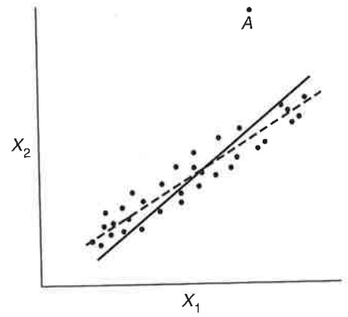


FIGURE 4.1 Bivariate scatterplot for showing impact of an outlier.

particular analysis. And they lead to results that do not generalize except to another sample with the same kind of outlier.

There are four reasons for the presence of an outlier. First is incorrect data entry. Cases that are extreme should be checked carefully to see that data are correctly entered. Second is failure to specify missing-value codes in computer syntax so that missing-value indicators are read as real data. Third is that the outlier is not a member of the population from which you intended to sample. If the case should not have been sampled, it is deleted once it is detected. Fourth is that the case is from the intended population but the distribution for the variable in the population has more extreme values than a normal distribution. In this event, the researcher retains the case but considers changing the value on the variable(s) so that the case no longer has as much impact. Although errors in data entry and missing values specification are easily found and remedied, deciding between alternatives three and four, between deletion and retention with alteration, is difficult.

4.1.4.1 Detecting Univariate and Multivariate Outliers

Univariate outliers are cases with an outlandish value on one variable; multivariate outliers are cases with an unusual combination of scores on two or more variables. For example, a 15-year-old is perfectly within bounds regarding age, and someone who earns $45,000 a year is in bounds regarding income, but a 15-year-old who earns $45,000 a year is very unusual and is likely to be a multivariate outlier. Multivariate outliers can occur when several different populations are mixed in the same sample or when some important variables are omitted that, if included, would attach the outlier to the rest of the cases.

Univariate outliers are easier to spot. Among dichotomous variables, the cases on the "wrong" side of a very uneven split are likely univariate outliers. Rummel (1970) suggests deleting dichotomous variables with 90–10 splits between categories, or more, both because the correlation coefficients between these variables and others are truncated and because the scores for the cases in the small category are more influential than those in the category with numerous cases. Dichotomous variables with extreme splits are easily found in the programs for frequency distributions (IBM SPSS FREQUENCIES, or SAS UNIVARIATE or Interactive Data Analysis) used during routine preliminary data screening.

Among continuous variables, the procedure for searching for outliers depends on whether data are grouped. If you are going to perform one of the analyses with ungrouped data (regression, canonical correlation, factor analysis, structural equation modeling, or some forms of time-series analysis), univariate and multivariate outliers are sought among all cases at once, as illustrated in Sections 4.2.1.1 (univariate) and 4.2.1.4 (multivariate). If you are going to perform one of the analyses with grouped data (ANCOVA, MANOVA or MANCOVA, profile analysis, discriminant analysis, logistic regression, survival analysis, or multilevel modeling), outliers are sought separately within each group, as illustrated in Sections 4.2.2.1 and 4.2.2.3.

Among continuous variables, univariate outliers are cases with very large standardized scores, z scores, on one or more variables, that are disconnected from the other z scores. Cases with standardized scores in excess of 3.29 ($p < .001$, two-tailed test) are potential outliers. However, the extremeness of a standardized score depends on the size of the sample; with a very large N, a few standardized scores in excess of 3.29 are expected. Z scores are available through IBM SPSS EXPLORE or DESCRIPTIVES (where z scores are saved in the data file), and SAS STANDARD (with MEAN = 0 and STD = 1). Or you can hand-calculate z scores from any output that provides means, standard deviations, and maximum and minimum scores.

As an alternative or in addition to inspection of z scores, there are graphical methods for finding univariate outliers. Helpful plots are histograms, box plots, normal probability plots, or detrended normal probability plots. Histograms of variables are readily understood and available and may reveal one or more univariate outliers. There is usually a pileup of cases near the mean with cases trailing away in either direction. An outlier is a case (or a very few cases) that seems to be unattached to the rest of the distribution. Histograms for continuous variables are produced by IBM SPSS FREQUENCIES (plus SORT and SPLIT for grouped data), and SAS UNIVARIATE or CHART (with BY for grouped data).

Box plots are simpler and literally box in observations that are around the median; cases that fall far away from the box are extreme. Normal probability plots and detrended normal probability plots are very useful for assessing normality of distributions of variables and are discussed in that context in Section 4.1.5.1. However, univariate outliers are visible in these plots as points that lie a considerable distance from others.

Once potential univariate outliers are located, the researcher decides whether transformations are acceptable. Transformations (Section 4.1.6) are undertaken both to improve the normality of distributions (Section 4.1.5.1) and to pull univariate outliers closer to the center of a distribution, thereby reducing their impact. Transformations, if acceptable, are undertaken prior to the search for multivariate outliers because the statistics used to reveal them (Mahalanobis distance and its variants) are also sensitive to failures of normality.

Mahalanobis distance is the distance of a case from the centroid of the remaining cases where the centroid is the point created at the intersection of the means of all the variables. In most data sets, the cases form a swarm around the centroid in multivariate space. Each case is represented in the swarm by a single point at its own peculiar combination of scores on all of the variables, just as each case is represented by a point at its own X, Y combination in a bivariate scatterplot. A case that is a multivariate outlier, however, lies outside the swarm, some distance from the other cases. Mahalanobis distance is one measure of that multivariate distance and it can be evaluated for each case using the χ^2 distribution.

Mahalanobis distance is tempered by the patterns of variances and covariances among the variables. It gives lower weight to variables with large variances and to groups of highly correlated variables. Under some conditions, Mahalanobis distance can either "mask" a real outlier (produce a false negative) or "swamp" a normal case (produce a false positive). Thus, it is not a perfectly reliable indicator of multivariate outliers and should be used with caution.

Mahalanobis distances are requested and interpreted in Sections 4.2.1.4 and 4.2.2.3 and in numerous other places throughout the book. A very conservative probability estimate for a case being an outlier, say, $p < .001$ for the χ^2 value, is appropriate with Mahalanobis distance.

Other statistical measures used to identify multivariate outliers are leverage, discrepancy, and influence. Although developed in the context of multiple regression (Chapter 5), the three measures are now available for some of the other analyses. Leverage is related to Mahalanobis distance (or variations of it in the "hat" matrix) and is variously called HATDIAG, RHAT, or h_{ii}. Although leverage is related to Mahalanobis distance, it is measured on a different scale so that significance tests based on a χ^2 distribution do not apply.[4] Equation 4.3 shows the relationship between leverage, h_{ii}, and Mahalanobis distance.

[4]Lunneborg (1994) suggests that outliers be defined as cases with $h_{ii} \geq 2(k/N)$.

$$\text{Mahalanobis distance} = (N - 1)(h_{ii} - 1/N) \tag{4.3}$$

Or, as is sometimes more useful,

$$h_{ii} = \frac{\text{Mahalanobis distance}}{N - 1} + \frac{1}{N}$$

The latter form is handy if you want to find a critical value for leverage at $\alpha = .001$ by translating the critical χ^2 value for Mahalanobis distance.

Cases with high leverage are far from the others, but they can be far out on basically the same line as the other cases, or far away and off the line. *Discrepancy* measures the extent to which a case is in line with the others. Figure 4.2(a) shows a case with high leverage and low discrepancy; Figure 4.2(b) shows a case with high leverage and high discrepancy. In Figure 4.2(c) is a case with low leverage and high discrepancy. In all of these figures, the outlier appears disconnected from the remaining scores.

Influence is a product of leverage and discrepancy (Fox, 1991). It assesses change in regression coefficients when a case is deleted; cases with influence scores larger than 1.00 are suspected of being outliers. Measures of influence are variations of Cook's distance and are identified in output as Cook's distance, modified Cook's distance, DFFITS, and DBETAS. For the interested reader, Fox (1991, pp. 29–30) describes these terms in more detail.

Leverage and/or Mahalanobis distance values are available as statistical methods of outlier detection in both statistical packages. However, research (e.g., Egan & Morgan, 1998; Hadi & Simonoff, 1993; Rousseeuw & van Zomeren, 1990) indicates that these methods are not perfectly reliable. Unfortunately, alternative methods are computationally challenging and not readily available in statistical packages. Therefore, multivariate outliers are currently most easily detected through Mahalanobis distance, or one of its cousins, but cautiously.

Statistics assessing the distance for each case, in turn, from all other cases, are available through IBM SPSS REGRESSION by evoking Mahalanobis, Cook's, or Leverage values through the Save command in the Regression menu; these values are saved as separate columns in the data file and examined using standard descriptive procedures. To use the regression program just to find outliers, however, you must specify some variable (such as the case number) as DV, to find

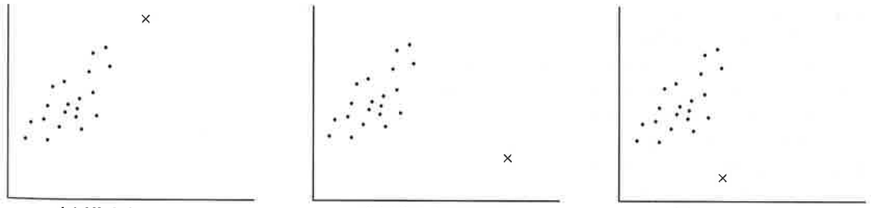


FIGURE 4.2 The relationships among leverage, discrepancy, and influence.

outliers among the set of variables of interest, considered IVs. Alternatively, the 10 cases with largest Mahalanobis distance are printed out by IBM SPSS REGRESSION using the RESIDUALS subcommand, as demonstrated in Section 4.2.1.4.

SAS regression programs provide a leverage, h_{ii}, value for each case that converts easily to Mahalanobis distance (Equation 4.3). These values are also-saved to the data file and examined using standard statistical and graphical techniques.

When multivariate outliers are sought in grouped data, they are sought within each group separately. IBM SPSS and SAS REGRESSION require separate runs for each group, each with its own error term. Programs in other packages, such as SYSTAT DISCRIM and BMDP7M, provide Mahalanobis distance for each case using a within-groups error term, so that outliers identified through those programs may be different from those identified by IBM SPSS and SAS REGRESSION.

IBM SPSS DISCRIMINANT provides outliers in the solution. These are not particularly helpful for screening (you would not want to delete cases just because the solution doesn't fit them very well), but are useful to evaluate generalizability of the results.

Frequently, some multivariate outliers hide behind other multivariate outliers—outliers are known to mask other outliers (Rousseeuw & van Zomeren, 1990). When the first few cases identified as outliers are deleted, the data set becomes more consistent and then other cases become extreme. Robust approaches to this problem have been proposed (e.g., Egan & Morgan, 1998; Hadi & Simonoff, 1993; Rousseeuw & van Zomeren, 1990), but these are not yet implemented in popular software packages. These methods can be approximated by screening for multivariate outliers several times—each time dealing with cases identified as outliers on the last run, until finally no new outliers are identified. But if the process of identifying ever more outliers seems to stretch into infinity, do a trial run with and without outliers to see if ones identified later are truly influencing results. If not, do not delete the later-identified outliers.

4.1.4.2 *Describing Outliers*

Once multivariate outliers are identified, you need to discover why the cases are extreme. (You already know why univariate outliers are extreme.) It is important to identify the variables on which the cases are deviant for three reasons. First, this procedure helps you decide whether the case is properly part of your sample. Second, if you are going to modify scores instead of delete cases, you have to know which scores to modify. Third, it provides an indication of the kinds of cases to which your results do not generalize.

If there are only a few multivariate outliers, it is reasonable to examine them individually. If there are several, you can examine them as a group to see if there are any variables that separate the group of outliers from the rest of the cases.

Whether you are trying to describe one or a group of outliers, the trick is to create a dummy grouping variable where the outlier(s) has one value and the rest of the cases another value. The dummy variable is then used as the grouping DV in discriminant analysis (Chapter 9) or logistic regression (Chapter 10), or as the DV in regression (Chapter 5). The goal is to identify the variables that distinguish outliers from the other cases. Variables on which the outlier(s) differs from the rest of the cases enter the equation; the remaining variables do not. Once those variables are identified, means on those variables for outlying and nonoutlying cases are found through any of the routine descriptive programs. Description of outliers is illustrated in Sections 4.2.1.4 and 4.2.2.3.

4.1.4.3 *Reducing the Influence of Outliers*

Once univariate outliers have been identified, there are several strategies for reducing their impact. But before you use one of them, check the data for the case to make sure that they are accurately entered into the data file. If the data are accurate, consider the possibility that one variable is responsible for most of the outliers. If so, elimination of the variable would reduce the number of outliers. If the variable is highly correlated with others or is not critical to the analysis, deletion of it is a good alternative.

If neither of these simple alternatives is reasonable, you must decide whether the cases that are outliers are properly part of the population from which you intended to sample. Cases with extreme scores, which are, nonetheless, apparently connected to the rest of the cases, are more likely to be a legitimate part of the sample. If the cases are not part of the population, they are deleted with no loss of generalizability of results to your intended population.

If you decide that the outliers are sampled from your target population, they remain in the analysis, but steps are taken to reduce their impact—variables are transformed or scores changed.

A first option for reducing impact of univariate outliers is variable transformation—undertaken to change the shape of the distribution to more nearly normal. In this case, outliers are considered part of a nonnormal distribution with tails that are too heavy so that too many cases fall at extreme values of the distribution. Cases that were outliers in the untransformed distribution are still on the tails of the transformed distribution, but their impact is reduced. Transformation of variables has other salutary effects, as described in Section 4.1.6.

A second option for univariate outliers is to change the score(s) on the variable(s) for the outlying case(s) so that they are deviant, but not as deviant as they were. For instance, assign the outlying case(s) a raw score on the offending variable that is one unit larger (or smaller) than the next most extreme score in the distribution. Because measurement of variables is sometimes rather arbitrary anyway, this is often an attractive alternative to reduce the impact of a univariate outlier.

Transformation or score alteration may not work for a truly multivariate outlier because the problem is with the combination of scores on two or more variables, not with the score on any one variable. The case is discrepant from the rest in its combinations of scores. Although the number of possible multivariate outliers is often substantially reduced after transformation or alteration of scores on variables, there are sometimes a few cases that are still far away from the others. These cases are usually deleted. If they are allowed to remain, it is with the knowledge that they may distort the results in almost any direction. Any transformations, changes of scores, and deletions are reported in the Results section together with the rationale.

4.1.4.4 *Outliers in a Solution*

Some cases may not fit well within a solution; the scores predicted for those cases by the selected model are very different from the actual scores for the cases. Such cases are identified *after* an analysis is completed, not as part of the screening process. To identify and eliminate or change scores for such cases, before conducting the major analysis, make the analysis look better than it should. Therefore, conducting the major analysis and then "retrofitting" is a procedure best limited to exploratory analysis. Chapters that describe techniques for ungrouped data deal with outliers in the solution when discussing the limitations of the technique.

4.1.5 Normality, Linearity, and Homoscedasticity

Underlying some multivariate procedures and most statistical tests of their outcomes is the assumption of multivariate normality. Multivariate normality is the assumption that each variable and all linear combinations of the variables are normally distributed. When the assumption is met, the residuals[5] of analysis are also normally distributed and independent. The assumption of multivariate normality is not readily tested because it is impractical to test an infinite number of linear combinations of variables for normality. Those tests that are available are overly sensitive.

The assumption of multivariate normality is made as part of derivation of many significance tests. Although it is tempting to conclude that most inferential statistics are robust[6] to violations of the assumption, that conclusion may not be warranted.[7] Bradley (1982) reports that statistical inference becomes less and less robust as distributions depart from normality, rapidly so under many conditions. And even when the statistics are used purely descriptively, normality, linearity, and homoscedasticity of variables enhance the analysis. The safest strategy, then, is to use transformations of variables to improve their normality unless there is some compelling reason not to.

The assumption of multivariate normality applies differently to different multivariate statistics. For analyses when subjects are not grouped, the assumption applies to the distributions of the variables themselves or to the residuals of the analyses; for analyses when subjects are grouped, the assumption applies to the sampling distributions[8] of means of variables.

If there is multivariate normality in ungrouped data, each variable is itself normally distributed and the relationships between pairs of variables, if present, are linear and *homoscedastic* (i.e., the variance of one variable is the same at all values of the other variable). The assumption of multivariate normality can be partially checked by examining the normality, linearity, and homoscedasticity of individual variables or through examination of residuals in analyses involving prediction.[9] The assumption is certainly violated, at least to some extent, if the individual variables (or the residuals) are not normally distributed or do not have pairwise linearity and homoscedasticity.

For grouped data, it is the sampling distributions of the means of variables that are to be normally distributed. The Central Limit Theorem reassures us that, with sufficiently large sample sizes, sampling distributions of means are normally distributed regardless of the distributions of variables.

[5]Residuals are leftovers. They are the segments of scores not accounted for by the multivariate analysis. They are also called "errors" between predicted and obtained scores where the analysis provides the predicted scores. Note that the practice of using a dummy DV such as case number to investigate multivariate outliers will *not* produce meaningful residuals plots.

[6]Robust means that the researcher is led to correctly reject the null hypothesis at a given alpha level the right number of times even if the distributions do not meet the assumptions of analysis. Often, Monte Carlo procedures are used where a distribution with some known properties is put into a computer, sampled from repeatedly, and repeatedly analyzed; the researcher studies the rates of retention and rejection of the null hypothesis against the known properties of the distribution in the computer.

[7]The univariate *F* test of mean differences, for example, is frequently said to be robust to violation of assumptions of normality and homogeneity of variance with large and equal samples, but Bradley (1984) questions this generalization.

[8]A sampling distribution is a distribution of statistics (not of raw scores) computed from random samples of a given size taken repeatedly from a population. For example, in univariate ANOVA, hypotheses are tested with respect to the sampling distribution of means (Chapter 3).

[9]Analysis of residuals to screen for normality, linearity, and homoscedasticity in multiple regression is discussed in Section 5.3.2.4.

For example, if there are at least 20 degrees of freedom for error in a univariate ANOVA, the *F* test is said to be robust to violations of normality of variables (provided that there are no outliers).

These issues are discussed again in the third sections of Chapters 5 through 16 and 18 (online) as they apply directly to one or another of the multivariate procedures. For nonparametric procedures such as multiway frequency analysis (Chapter 16) and logistic regression (Chapter 10), there are no distributional assumptions. Instead, distributions of scores typically are hypothesized and observed distributions are tested against hypothesized distributions.

4.1.5.1 Normality

Screening continuous variables for normality is an important early step in almost every multivariate analysis, particularly when inference is a goal. Although normality of the variables is not always required for analysis, the solution is usually quite a bit better if the variables are all normally distributed. The solution is degraded, if the variables are not normally distributed, and particularly if they are nonnormal in very different ways (e.g., some positively and some negatively skewed).

Normality of variables is assessed by either statistical or graphical methods. Two components of normality are skewness and kurtosis. Skewness has to do with the symmetry of the distribution; a skewed variable is a variable whose mean is not in the center of the distribution. Kurtosis has to do with the peakedness of a distribution; a distribution is either too peaked (with short, thick tails) or too flat (with long, thin tails).[10] Figure 4.3 shows a normal distribution, distributions with skewness, and distributions with nonnormal kurtosis. A variable can have significant skewness, kurtosis, or both.

When a distribution is normal, the values of skewness and kurtosis are zero. If there is positive skewness, there is a pileup of cases to the left and the right tail is too long; with negative skewness, there is a pileup of cases to the right and the left tail is too long. Kurtosis values above zero indicate a distribution that is too peaked with short, thick tails, and kurtosis values below zero indicate a distribution that is too flat (also with too many cases in the tails).[11] Nonnormal kurtosis produces an underestimate of the variance of a variable.

There are significance tests for both skewness and kurtosis that test the obtained value against null hypotheses of zero. For instance, the standard error for skewness is approximately

$$s_s = \sqrt{\frac{6}{N}} \tag{4.4}$$

where N is the number of cases. The obtained skewness value is then compared with zero using the z distribution, where

$$z = \frac{S - 0}{s_s} \tag{4.5}$$

and S is the value reported for skewness. The standard error for kurtosis is approximately

$$s_k = \sqrt{\frac{24}{N}} \tag{4.6}$$

[10]If you decide that outliers are sampled from the intended population but that there are too many cases in the tails, you are saying that the distribution from which the outliers are sampled has kurtosis that departs from normal.

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

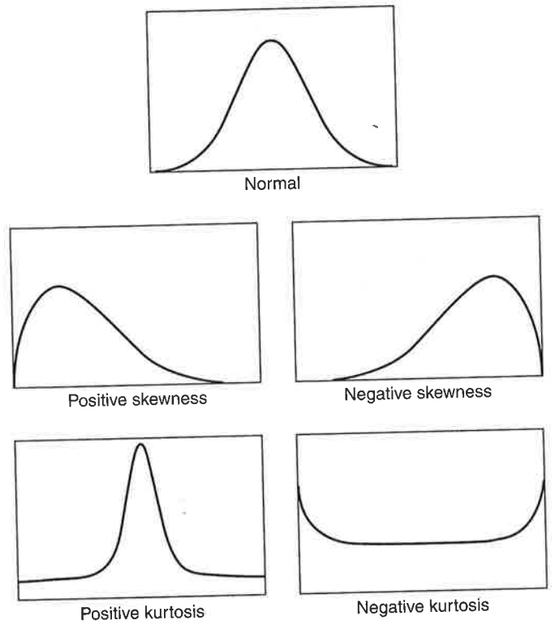


FIGURE 4.3 Normal distribution, distributions with skewness, and distributions with kurtoses.

and the obtained kurtosis value is compared with zero using the z distribution, where

$$z = \frac{K - 0}{S_k} \tag{4.7}$$

and K is the value reported for kurtosis.

Conventional but conservative (.01 or .001) alpha levels are used to evaluate the significance of skewness and kurtosis with small to moderate samples, but if the sample is large, it is a good idea to look at the shape of the distribution instead of using formal inference tests. Because the standard errors for both skewness and kurtosis decrease with larger N, the null hypothesis is likely to be rejected with large samples when there are only minor deviations from normality.

In a large sample, a variable with statistically significant skewness often does not deviate enough from normality to make a substantive difference in the analysis. In other words, with large samples, the significance level of skewness is not as important as its actual size (worse the farther from zero) and the visual appearance of the distribution. In a large sample, the impact of departure from zero kurtosis also diminishes. For example, underestimates of variance associated with positive kurtosis (distributions with short, thick tails) disappear with samples of 100 or more cases; with negative kurtosis, underestimation of variance disappears with samples of 200 or more (Waternaux, 1976).

Values for skewness and kurtosis are available in several programs. IBM SPSS FREQUENCIES, for instance, prints as options skewness, kurtosis, and their standard errors, and, in addition, superimposes a normal distribution over a frequency histogram for a variable if HISTOGRAM = NORMAL is specified. DESCRIPTIVES and EXPLORE also print skewness and kurtosis statistics. A histogram or stem-and-leaf plot is also available in SAS UNIVARIATE.[12]

Frequency histograms are an important graphical device for assessing normality, especially with the normal distribution as an overlay, but even more helpful than frequency histograms are expected normal probability plots and detrended expected normal probability plots. In these plots, the scores are ranked and sorted; then an expected normal value is computed and compared with the actual normal value for each case. The expected normal value is the z score that a case with that rank holds in a normal distribution; the normal value is the z score it has in the actual distribution. If the actual distribution is normal, then the points for the cases fall along the diagonal running from lower left to upper right, with some minor deviations due to random processes. Deviations from normality shift the points away from the diagonal.

Consider the expected normal probability plots for ATTDRUG and TIMEDRS through IBM SPSS PPLOT in Figure 4.4. Syntax indicates the VARIABLES we are interested in are attdrug and timedrs. The remaining syntax is produced by default by the IBM SPSS Windows menu system. As reported in Section 4.2.1.1, ATTDRUG is reasonably normally distributed (kurtosis = −0.447, skewness = −0.123) and TIMEDRS is too peaked and positively skewed (kurtosis = 13.101, skewness = 3.248, both significantly different from 0). The cases for ATTDRUG line up along the diagonal, whereas those for TIMEDRS do not. At low values of TIMEDRS, there are too many cases above the diagonal, and at high values, there are too many cases below the diagonal, reflecting the patterns of skewness and kurtosis.

Detrended normal probability plots for TIMEDRS and ATTDRUG are also in Figure 4.4. These plots are similar to expected normal probability plots except that deviations from the diagonal are plotted instead of values along the diagonal. In other words, the linear trend from lower left to upper right is removed. If the distribution of a variable is normal, as is ATTDRUG, the cases distribute themselves evenly above and below the horizontal line that intersects the Y-axis at 0.0, the line of zero deviation from expected normal values. The skewness and kurtosis of TIMEDRS are again apparent from the cluster of points above the line at low values of TIMEDRS and below the line at high values of TIMEDRS. Normal probability plots for variables are also available in SAS UNIVARIATE and IBM SPSS MANOVA. Many of these programs also produce detrended normal plots.

If you are going to perform an analysis with ungrouped data, an alternative to screening the variables prior to analysis is conducting the analysis and then screening the residuals (the differences between the predicted and obtained DV values). If normality is present, the residuals are normally and independently distributed. That is, the differences between predicted and obtained scores—the errors—are symmetrically distributed around a mean value of zero and there are no contingencies among the errors. In multiple regression, residuals are also screened for normality through the expected normal probability plot and the detrended normal probability plot.[13] IBM SPSS REGRESSION

[12]In structural equation modeling (Chapter 14), skewness and kurtosis for each variable are available in EQS and Mardia's coefficient (the multivariate kurtosis measure) is available in EQS, PRELIS, and CALIS. In addition, PRELIS can be used to deal with nonnormality through alternative correlation coefficients, such as polyserial or polychoric (cf. Section 14.5.6).

[13]For grouped data, residuals have the same shape as within-group distributions because the predicted value is the mean, and subtracting a constant does not change the shape of the distribution. Many of the programs for grouped data plot the within-group distribution as an option, as discussed in the next few chapters when relevant.

```
PPLOT
  /VARIABLES = attdrug timedrs
  /NOLOG
  /NOSTANDARDIZE
  /TYPE = P-P
  /FRACTION = BLOM
  /TIES = MEAN
  /DIST = NORMAL.
```

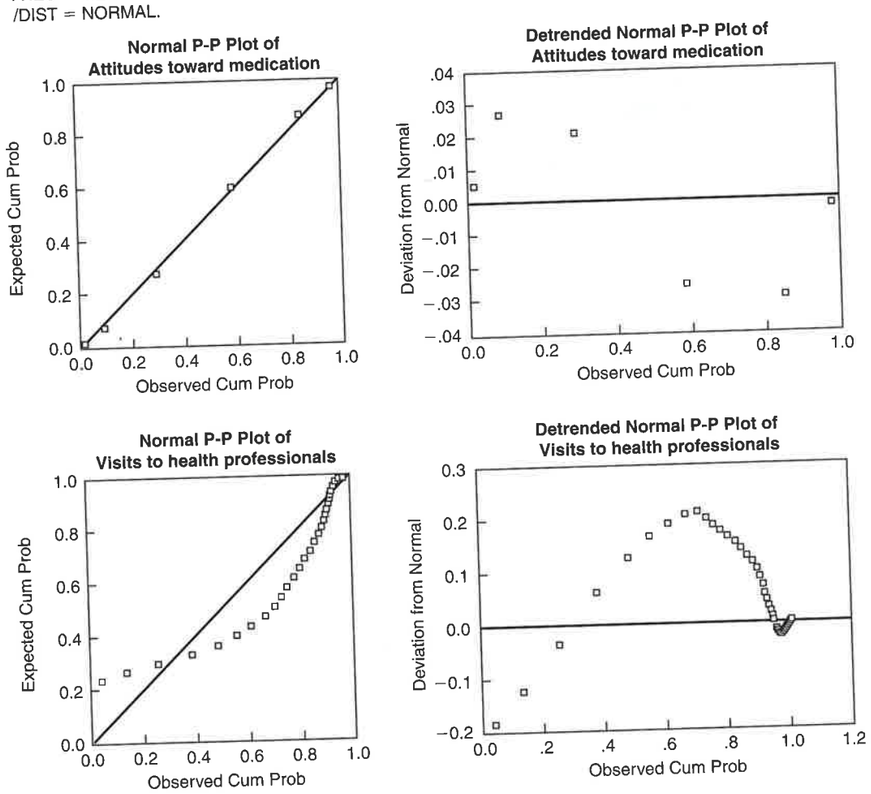


FIGURE 4.4 Expected normal probability plot and detrended normal probability plot for ATTDRUG and TIMEDRS. IBM SPSS PPLOT syntax and output.

provides this diagnostic technique (and others, as discussed in Chapter 5). If the residuals are normally distributed, the expected normal probability plot and the detrended normal probability plot look just the same as they do if a variable is normally distributed. In regression, if the residuals plot looks normal, there is no reason to screen the individual variables for normality.

Although residuals will reveal departures from normality, the analyst has to resist temptation to look at the rest of the output to avoid "tinkering" with variables and cases to produce an anticipated result. Because screening the variables should lead to the same conclusions as screening residuals, it may be more objective to make one's decisions about transformations, deletion of outliers, and the like, on the basis of screening runs alone rather than screening through the outcome of analysis.[14]

With ungrouped data, if nonnormality is found, transformation of variables is considered. Common transformations are described in Section 4.1.6. Unless there are compelling reasons not to transform, it is probably better to do so. However, realize that even if each of the variables is normally distributed, or transformed to normal, there is no guarantee that all linear combinations of the variables are normally distributed. That is, if variables are each univariate normal, they do not necessarily have a multivariate normal distribution. However, it is more likely that the assumption of multivariate normality is met if all the variables are normally distributed.

4.1.5.2 Linearity

The assumption of linearity is that there is a straight-line relationship between two variables (where one or both of the variables can be combinations of several variables). Linearity is important in a practical sense because Pearson's *r* only captures the linear relationships among variables; if there are substantial nonlinear relationships among variables, they are ignored.

Nonlinearity is diagnosed either from residuals plots in analyses involving a predicted variable or from bivariate scatterplots between pairs of variables. In plots where standardized residuals are plotted against predicted values, nonlinearity is indicated when most of the residuals are above the zero line on the plot at some predicted values and below the zero line at other predicted values (see Chapter 5).

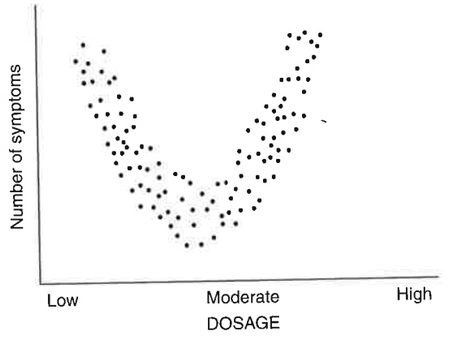
Linearity between two variables is assessed roughly by inspection of bivariate scatterplots. If both variables are normally distributed and linearly related, the scatterplot is oval-shaped. If one of the variables is nonnormal, then the scatterplot between this variable and the other is not oval. Examination of bivariate scatterplots is demonstrated in Section 4.2.1.2, along with transformation of a variable to enhance linearity.

However, sometimes the relationship between variables is simply not linear. Consider, for instance, the number of symptoms and the dosage of drug, as shown in Figure 4.5(a). It seems likely that there are lots of symptoms when the dosage is low, only a few symptoms when the dosage is moderate, and lots of symptoms again when the dosage is high. Number of symptoms and drug dosage are curvilinearly related. One alternative in this case is to use the square of number of symptoms to represent the curvilinear relationship instead of number of symptoms in the analysis. Another alternative is to recode dosage into two dummy variables (high vs. low on one dummy variable and a combination of high and low vs. medium on another dummy variable) and then use the dummy variables in place of dosage in analysis.[15] The dichotomous dummy variables can only have a linear relationship with other variables, if, indeed, there is any relationship at all after recoding.

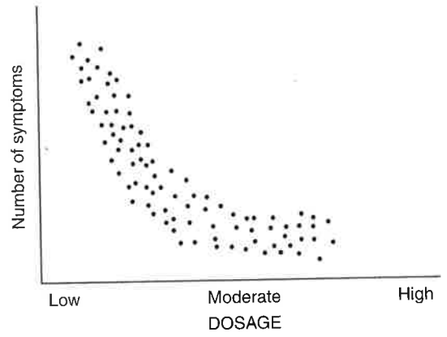
Often, two variables have a mix of linear and curvilinear relationships, as shown in Figure 4.5(b). One variable generally gets smaller (or larger) as the other gets larger (or smaller) but there

[14]We realize that others (e.g., Berry, 1993; Fox, 1991) have very different views about the wisdom of screening from residuals.

[15]A nonlinear analytic strategy is most appropriate here, such as nonlinear regression through SAS NLIN, but such strategies are beyond the scope of this book.



(a) Curvilinear



(b) Curvilinear + linear

FIGURE 4.5 Curvilinear relationship and curvilinear plus linear relationship.

is also a curve to the relationship. For instance, symptoms might drop off with increasing dosage, but only to a point; increasing dosage beyond the point does not result in further reduction or increase in symptoms. In this case, the linear component may be strong enough that not much is lost by ignoring the curvilinear component unless it has important theoretical implications.

Assessing linearity through bivariate scatterplots is reminiscent of reading tea leaves, especially with small samples. And there are many cups of tea if there are several variables and all possible pairs are examined, especially when subjects are grouped and the analysis is done separately within each group. If there are only a few variables, screening all possible pairs is not burdensome; if there are numerous variables, you may want to use statistics on skewness to screen only pairs that are likely to depart from linearity. Think, also, about pairs of variables that might have true nonlinearity and examine them through bivariate scatterplots. Bivariate scatterplots are produced by IBM SPSS GRAPH, and SAS PLOT, among other programs.

4.1.5.3 Homoscedasticity, Homogeneity of Variance, and Homogeneity of Variance–Covariance Matrices

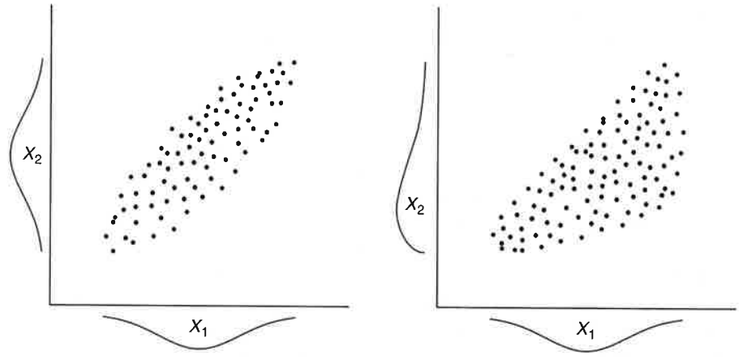
For ungrouped data, the assumption of homoscedasticity is that the variability in scores for one continuous variable is roughly the same at all values of another continuous variable. For grouped data, this is the same as the assumption of homogeneity of variance when one of the variables is discrete (the grouping variable), the other is continuous (the DV); the variability in the DV is expected to be about the same at all levels of the grouping variable.

Homoscedasticity is related to the assumption of normality because when the assumption of multivariate normality is met, the relationships between variables are homoscedastic. The bivariate scatterplots between two variables are of roughly the same width all over with some bulging toward the middle. Homoscedasticity for a bivariate plot is illustrated in Figure 4.6(a).

Heteroscedasticity, the failure of homoscedasticity, is caused either by nonnormality of one of the variables or by the fact that one variable is related to some transformation of the other. Consider, for example, the relationship between age (X_1) and income (X_2) as depicted in Figure 4.6(b). People start out making about the same salaries, but with increasing age, people spread farther apart on income. The relationship is perfectly lawful, but it is not homoscedastic. In this example, income is likely to be positively skewed and transformation of income is likely to improve the homoscedasticity of its relationship with age.

Another source of heteroscedasticity is a greater error of measurement at some levels of an IV. For example, people in the age range 25 to 45 might be more concerned about their weight than people who are younger or older. Older and younger people would, as a result, give less reliable estimates of their weight, increasing the variance of weight scores at those ages.

It should be noted that heteroscedasticity is not fatal to an analysis of ungrouped data. The linear relationship between variables is captured by the analysis, but there is even more predictability if the heteroscedasticity is accounted for. If it is not, the analysis is weakened, but not invalidated.



(a) Homoscedasticity with both variables normally distributed

(b) Heteroscedasticity with skewness on X_2

FIGURE 4.6 Bivariate scatterplots under conditions of homoscedasticity and heteroscedasticity.

When data are grouped, homoscedasticity is known as homogeneity of variance. A great deal of research has assessed the robustness (or lack thereof) of ANOVA and ANOVA-like analyses to violation of homogeneity of variance. Recent guidelines have become more stringent than earlier, more cavalier ones. There are formal tests of homogeneity of variance but most are too strict because they also assess normality. (An exception is Levene's test of homogeneity of variance, which is not typically sensitive to departures from normality.) Instead, once outliers are eliminated, homogeneity of variance is assessed with F_{max} in conjunction with sample-size ratios.

F_{max} is the ratio of the largest cell variance to the smallest. If sample sizes are relatively equal (within a ratio of 4 to 1 or less for largest to smallest cell size), an F_{max} as great as 10 is acceptable. As the cell size discrepancy increases (say, goes to 9 to 1 instead of 4 to 1), an F_{max} as small as 3 is associated with inflated Type I error if the larger variance is associated with the smaller cell size (Milligan, Wong, & Thompson, 1987).

Violations of homogeneity usually can be corrected by transformation of the DV scores. Interpretation, however, is then limited to the transformed scores. Another option is to use untransformed variables with a more stringent α level (for nominal α, use .025 with moderate violation and .01 with severe violation).

The multivariate analog of homogeneity of variance is homogeneity of variance–covariance matrices. As for univariate homogeneity of variance, inflated Type I error rate occurs when the greatest dispersion is associated with the smallest sample size. The formal test used by IBM SPSS, Box's *M*, is too strict with the large sample sizes usually necessary for multivariate applications of ANOVA. Section 9.7.1.5 demonstrates an assessment of homogeneity of variance–covariance matrices through SAS DISCRIM using Bartlett's test. SAS DISCRIM permits a stringent α level for determining heterogeneity, and bases the discriminant analysis on separate variance–covariance matrices when the assumption of homogeneity is violated.

4.1.6 Common Data Transformations

Although data transformations are recommended as a remedy for outliers and for failures of normality, linearity, and homoscedasticity, they are not universally recommended. The reason is that an analysis is interpreted from the variables that are in it, and transformed variables are sometimes harder to interpret. For instance, although IQ scores are widely understood and meaningfully interpreted, the logarithm of IQ scores may be harder to explain.

Whether transformation increases difficulty of interpretation often depends on the scale in which the variable is measured. If the scale is meaningful or widely used, transformation often hinders interpretation, but if the scale is somewhat arbitrary anyway (as is often the case), transformation does not notably increase the difficulty of interpretation.

With ungrouped data, it is probably best to transform variables to normality unless interpretation is not feasible with the transformed scores. With grouped data, the assumption of normality is evaluated with respect to the sampling distribution of means (not the distribution of scores) and the Central Limit Theorem predicts normality with decently sized samples. However, transformations may improve the analysis and may have the further advantage of reducing the impact of outliers. Our recommendation, then, is to consider transformation of variables in all situations unless there is some reason not to.

If you decide to transform, it is important to check that the variable is normally or near-normally distributed after transformation. Often you need to try first one transformation and then another until you find the transformation that produces the skewness and kurtosis values nearest zero, the prettiest picture, and/or the fewest outliers.

With almost every data set in which we have used transformations, the results of analysis have been substantially improved. This is particularly true when some variables are skewed and others are not, or variables are skewed very differently prior to transformation. However, if all the variables are skewed to about the same moderate extent, improvements of analysis with transformation are often marginal.

With grouped data, the test of mean differences after transformation is a test of differences between medians in the original data. After a distribution is normalized by transformation, the mean is equal to the median. The transformation affects the mean but not the median because the median depends only on rank order of cases. Therefore, conclusions about means of transformed distributions apply to medians of untransformed distributions. Transformation is undertaken because the distribution is skewed and the mean is not a good indicator of the central tendency of the scores in the distribution. For skewed distributions, the median is often a more appropriate measure of central tendency than the mean, anyway, so interpretation of differences in medians is appropriate.

Variables differ in the extent to which they diverge from normal. Figure 4.7 presents several distributions together with the transformations that are likely to render them normal. If the distribution differs moderately from normal, a square root transformation is tried first. If the distribution

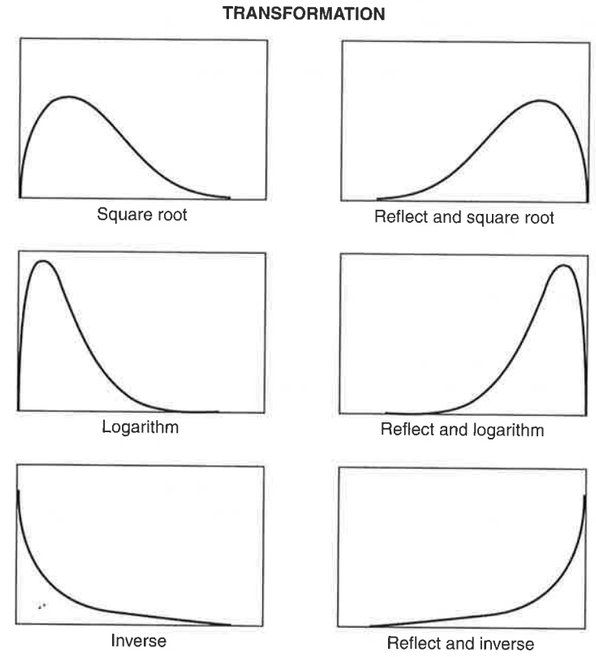


FIGURE 4.7 Original distributions and common transformations to produce normality.

differs substantially, a log transformation is tried. If the distribution differs severely, the inverse is tried. According to Bradley (1982), the inverse is the best of several alternatives for J-shaped distributions, but even it may not render the distribution normal. Finally, if the departure from normality is severe and no transformation seems to help, you may want to try dichotomizing the variable.

The direction of the deviation is also considered. When distributions have positive skewness, as discussed earlier, the long tail is to the right. When they have negative skewness, the long tail is to the left. If there is negative skewness, the best strategy is to *reflect* the variable and then apply the appropriate transformation for positive skewness.[16] To reflect a variable, find the largest score in the distribution. Then create a new variable by subtracting each score from the largest score plus 1. In this way, a variable with negative skewness is converted to one with positive skewness prior to transformation. When you interpret a reflected variable, be sure to reverse the direction of the interpretation as well (or consider re-reflecting it after transformation).

Remember to check your transformations after applying them. If a variable is only moderately positively skewed, for instance, a square root transformation may make the variable moderately negatively skewed, and there is no advantage to transformation. Often you have to try several transformations before you find the most helpful one.

Syntax for transforming variables in IBM SPSS and SAS is given in Table 4.3.[17] Notice that a constant is also added if the distribution contains a value less than one. A constant (to bring the smallest value to at least one) is added to each score to avoid taking the log, square root, or inverse of zero.

Different software packages handle missing data differently in various transformations. Be sure to check the manual to ensure that the program is treating missing data the way you want it to in the transformation.

It should be clearly understood that this section merely scratches the surface of the topic of transformations, about which a great deal more is known. The interested reader is referred to Box and Cox (1964) or Mosteller and Tukey (1977) for a more flexible and challenging approach to the problem of transformation.

4.1.7 Multicollinearity and Singularity

Multicollinearity and singularity are problems with a correlation matrix that occur when variables are too highly correlated. With multicollinearity, the variables are very highly correlated (say, .90 and above); with singularity, the variables are redundant; one of the variables is a combination of two or more of the other variables.

For example, scores on the Wechsler Adult Intelligence Scale (the WAIS) and scores on the Stanford–Binet Intelligence Scale are likely to be *multicollinear* because they are two similar measures of the same thing. But the total WAIS IQ score is *singular* with its subscales because the total score is found by combining subscale scores. When variables are multicollinear or singular, they

[16]Remember, however, that the interpretation of a reflected variable is just the opposite of what it was; if big numbers meant good things prior to reflecting the variable, big numbers mean bad things afterward.

[17]Logarithmic (LO) and power (PO) transformations are also available in PRELIS for variables used in structural equation modeling (Chapter 14). A λ (GA) value is specified for power transformations; for example, λ =1/2 provides a square root transform (PO GA =.5).

TABLE 4.3 Syntax for Common Data Transformations

	IBM SPSS COMPUTE	SAS[a] DATA Procedure
Moderate positive skewness	NEWX=SQRT(X)	NEWX=SQRT(X)
Substantial positive skewness	NEWX=LG10(X)	NEWX=LOG10(X)
With zero	NEWX=LG10(X + C)	NEWX=LOG10(X+C)
Severe positive skewness	NEWX=1/X	NEWX=1/X
L-shaped With zero	NEWX=1/(X + C)	NEWX=1/(X+C)
Moderate negative skewness	NEWX=SQRT(K – X)	NEWX=SQRT(K-X)
Substantial negative skewness	NEWX=LG10(K – X)	NEWX=LOG10(K-X)
Severe negative skewness J-shaped	NEWX=1/(K – X)	NEWX=1/(K-X)

C = a constant added to each score so that the smallest score is 1.

K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

[a]Also may be done through SAS Interactive Data Analysis.

contain redundant information and they are not all needed in the same analysis. In other words, there are fewer variables than it appears and the correlation matrix is not full rank because there are not really as many variables as columns.

Either bivariate or multivariate correlations can create multicollinearity or singularity. If a bivariate correlation is too high, it shows up in a correlation matrix as a correlation above .90, and, after deletion of one of the two redundant variables, the problem is solved. If it is a multivariate correlation that is too high, diagnosis is slightly more difficult because multivariate statistics are needed to find the offending variable. For example, although the WAIS IQ is a combination of its subscales, the bivariate correlations between total IQ and each of the subscale scores are not all that high. You would not know there was singularity by examination of the correlation matrix.

Multicollinearity and singularity cause both logical and statistical problems. The logical problem is that unless you are doing analysis of structure (factor analysis, principal components analysis, and structural-equation modeling), it is not a good idea to include redundant variables in the same analysis. They are not needed, and because they inflate the size of error terms, they actually weaken an analysis. Unless you are doing analysis of structure or are dealing with repeated measures of the same variable (as in various forms of ANOVA including profile analysis), think carefully before

including two variables with a bivariate correlation of, say, .70 or more in the same analysis. You might omit one of the variables or you might create a composite score from the redundant variables.

The statistical problems created by singularity and multicollinearity occur at much higher correlations (.90 and higher). The problem is that singularity prohibits, and multicollinearity renders unstable, matrix inversion. Matrix inversion is the logical equivalent of division; calculations requiring division (and there are many of them—see the fourth sections of Chapters 5 through 16 and 18) cannot be performed on singular matrices because they produce determinants equal to zero that cannot be used as divisors (see Appendix A). Multicollinearity often occurs when you form cross products or powers of variables and include them in the analysis along with the original variables, unless steps are taken to reduce the multicollinearity (Section 5.6.6).

With multicollinearity, the determinant is not exactly zero, but it is zero to several decimal places. Division by a near-zero determinant produces very large and unstable numbers in the inverted matrix. The sizes of numbers in the inverted matrix fluctuate wildly with only minor changes (say, in the second or third decimal place) in the sizes of the correlations in **R**. The portions of the multivariate solution that flow from an inverted matrix that is unstable are also unstable. In regression, for instance, error terms get so large that none of the coefficients is significant (Berry, 1993). For example, when r is .9, the precision of estimation of regression coefficients is halved (Fox, 1991).

Most programs protect against multicollinearity and singularity by computing SMCs for the variables. SMC is the squared multiple correlation of a variable where it serves as DV with the rest as IVs in multiple correlation (see Chapter 5). If the SMC is high, the variable is highly related to the others in the set and you have multicollinearity. If the SMC is 1, the variable is perfectly related to others in the set and you have singularity. Many programs convert the SMC values for each variable to tolerance (1 − SMC) and deal with tolerance instead of SMC.

Screening for singularity often takes the form of running your main analysis to see if the computer balks. Singularity aborts most runs except those for principal components analysis (see Chapter 13), where matrix inversion is not required. If the run aborts, you need to identify and delete the offending variable. A first step is to think about the variables. "Did you create any of the variables from other variables; for instance, did you create one of them by adding two others?" If so, deletion of one removes singularity.

Screening for multicollinearity that causes statistical instability is also a routine with most programs because they have tolerance criteria for inclusion of variables. If the tolerance (1 – SMC) is too low, the variable does not enter the analysis. Default tolerance levels range between .01 and .0001, so SMCs are .99 to .9999 before variables are excluded. You may wish to take control of this process, however, by adjusting the tolerance level (an option with many programs) or deciding yourself which variable(s) to delete instead of letting the program make the decision on purely statistical grounds. For this you need SMCs for each variable. Note that SMCs are *not* evaluated separately for each group if you are analyzing grouped data.

SMCs are available through factor analysis and regression programs in all packages. PRELIS provides SMCs for structural equation modeling. SAS and IBM SPSS have incorporated *collinearity diagnostics* proposed by Belsley, Kuh, and Welsch (1980) in which a conditioning index is produced, as well as variance proportions associated with each variable, after standardization, for each root (see Chapters 12 and 13 and Appendix A for a discussion of roots and dimensions). Two or more variables with large variance proportions on the same dimension are those with problems.

Condition index is a measure of tightness or dependency of one variable on the others. The condition index is monotonic with SMC, but not linear with it. A high condition index is associated with variance inflation in the standard error of the parameter estimate for a variable. When its standard error becomes very large, the parameter estimate is highly uncertain. Each root (dimension) accounts for some proportion of the variance of each parameter estimated. A collinearity problem occurs when a root with a high condition index contributes strongly (has a high variance proportion) to the variance of two or more variables. Criteria for multicollinearity suggested by Belsely et al. (1980) are a conditioning index greater than 30 for a given dimension coupled with variance proportions greater than .50 for at least two different variables. Collinearity diagnostics are demonstrated in Section 4.2.1.6.

There are several options for dealing with collinearity if it is detected. First, if the only goal of analysis is prediction, you can ignore it. A second option is to delete the variable with the highest variance proportion. A third option is to sum or average the collinear variables. A fourth option is to compute principal components and use the components as the predictors instead of the original variables (see Chapter 13). A final alternative is to center one or more of the variables, as discussed in Chapters 5 and 15, if multicollinearity is caused by forming interactions or powers of continuous variables.

4.1.8 A Checklist and Some Practical Recommendations

Table 4.4 is a checklist for screening data. It is important to consider all the issues prior to the fundamental analysis lest you be tempted to make some of your decisions based on their influence on the analysis. If you choose to screen through residuals, you cannot avoid doing an analysis at the same time; however, in these cases, you concentrate on the residuals and not on the other features of the analysis while making your screening decisions.

TABLE 4.4 Checklist for Screening Data

1. Inspect univariate descriptive statistics for accuracy of input
   a. Out-of-range values
   b. Plausible means and standard deviations
   c. Univariate outliers
2. Evaluate amount and distribution of missing data; deal with problem
3. Check pairwise plots for nonlinearity and heteroscedasticity
4. Identify and deal with nonnormal variables and univariate outliers
   a. Check skewness and kurtosis, probability plots
   b. Transform variables (if desirable)
   c. Check results of transformation
5. Identify and deal with multivariate outliers
   a. Variables causing multivariate outliers
   b. Description of multivariate outliers
6. Evaluate variables for multicollinearity and singularity

The order in which screening takes place is important because the decisions that you make at one step influence the outcomes of later steps. In a situation where you have both nonnormal variables and potential univariate outliers, a fundamental decision is whether you would prefer to transform variables, delete cases, or change scores on cases. If you transform variables first, you are likely to find fewer outliers. If you delete or modify the outliers first, you are likely to find fewer variables with nonnormality.

Of the two choices, transformation of variables is usually preferable. It typically reduces the number of outliers. It is likely to produce normality, linearity, and homoscedasticity among the variables. It increases the likelihood of multivariate normality to bring the data into conformity with one of the fundamental assumptions of most inferential tests. And on a very practical level, it usually enhances the analysis even if inference is not a goal. On the other hand, transformation may threaten interpretation, in which case all the statistical niceties are of little avail.

Or, if the impact of outliers is reduced first, you are less likely to find variables that are skewed because significant skewness is sometimes caused by extreme cases on the tails of the distributions. If you have cases that are univariate outliers because they are not part of the population from which you intended to sample, by all means delete them before checking distributions.

Last, as will become obvious in the next two sections, although the issues are different, the runs on which they are screened are not necessarily different. That is, the same run often provides you with information regarding two or more issues.

4.2 Complete Examples of Data Screening

Evaluation of assumptions is somewhat different for ungrouped and grouped data. That is, if you are going to perform multiple regression, canonical correlation, factor analysis, or structural equation modeling on ungrouped data, there is one approach to screening. If you are going to perform univariate or multivariate analysis of variance (including profile analysis), discriminant analysis, or multilevel modeling on grouped data, there is another approach to screening.[18]

Therefore, two complete examples are presented that use the same set of variables taken from the research described in Appendix B: number of visits to health professionals (TIMEDRS), attitudes toward drug use (ATTDRUG), attitudes toward housework (ATTHOUSE), INCOME, marital status (MSTATUS), and RACE. The grouping variable used in the analysis of grouped data is current employment status (EMPLMNT).[19] Data are in files labeled SCREEN.* (e.g. SCREEN.sav for IBM SPSS or SCREEN.sas7bdat for SAS Version 8 or newer).

Where possible in these examples, and for illustrative purposes, screening for ungrouped data is performed using IBM SPSS, and screening for grouped data is performed using SAS programs.

4.2.1 Screening Ungrouped Data

A flow diagram for screening ungrouped data appears as Figure 4.8. The direction of flow assumes that data transformation is undertaken, as necessary. If transformation is not acceptable, then other procedures for handling outliers are used.

[18]If you are using multiway frequency analysis or logistic regression, there are far fewer assumptions than with these other analyses.

[19]This is a motley collection of variables chosen primarily for their statistical properties.

Searching for:	*Programs*:
Plausible Range Missing Values Normality Univariate Outliers	IBM SPSS FREQUENCIES DESCRIPTIVES EXPLORE SAS MEANS UNIVARIATE
Pairwise Linearity	IBM SPSS PLOT SAS PLOT
Or Check Residuals	IBM SPSS REGRESSION SAS REG
Deal with Missing Data	IBM SPSS MVA SAS MI
Multivariable Outliers	IBM SPSS REGRESSION SAS REG
(Describe Outliers)	IBM SPSS DISCRIMINANT LIST VARIABLES SAS MEANS DATA

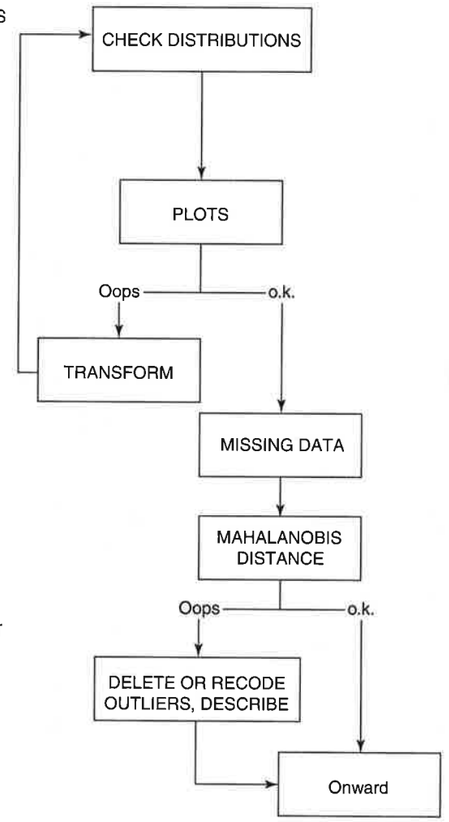


FIGURE 4.8 Flow diagram for screening ungrouped data.

4.2.1.1 Accuracy of Input, Missing Data, Distributions, and Univariate Outliers

A check on accuracy of data entry, missing data, skewness, and kurtosis for the data set is done through IBM SPSS FREQUENCIES, as shown in Table 4.5.

The minimum and maximum values, means, and standard deviations of each of the variables are inspected for plausibility. For instance, the Minimum number of visits to health professionals (TIMEDRS) is 0 and the Maximum is 81, higher than expected but found to be accurate on checking the data sheets.[20] The Mean for the variable is 7.901, higher than the national average but not

[20]The woman with this number of visits was terminally ill when she was interviewed.

TABLE 4.5 Syntax and IBM SPSS FREQUENCIES Output Showing Descriptive Statistics and Histograms for Ungrouped Data

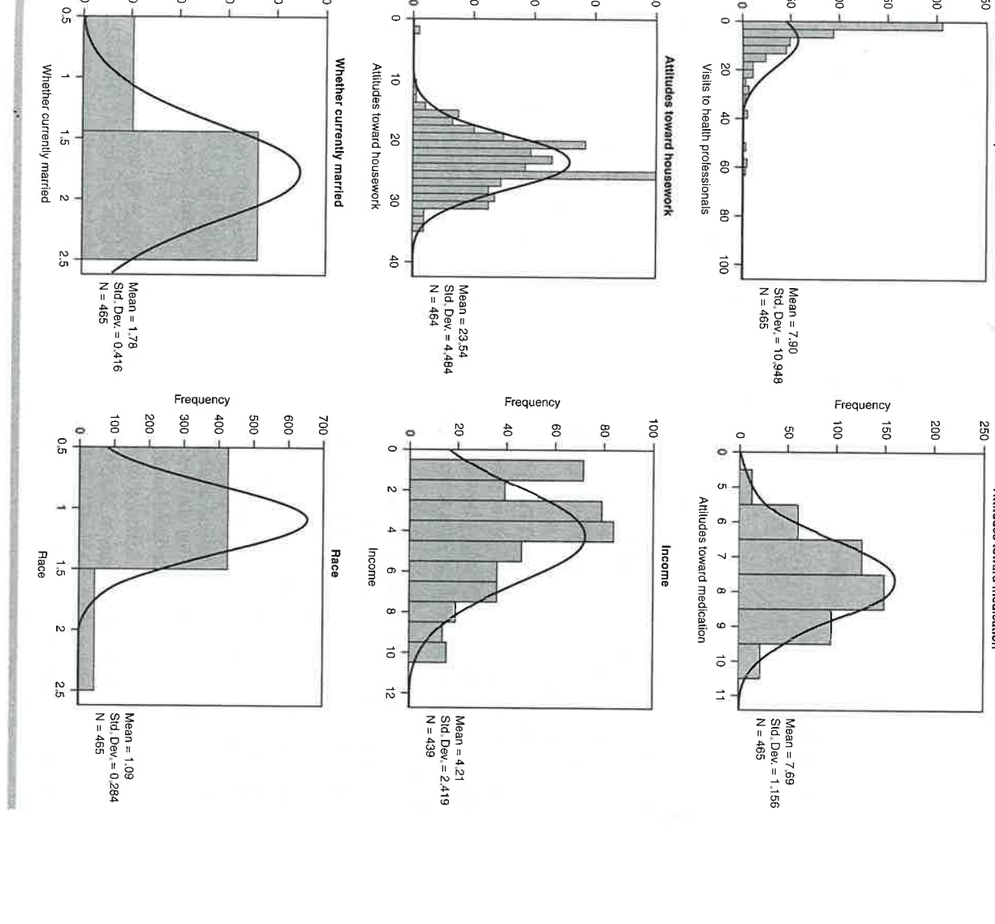
```
FREQUENCIES
  VARIABLES=timedrs attdrug atthouse income mstatus race
  /FORMAT=NOTABLE
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM MEAN SKEWNESS SESKEW KURTOSIS
  SEKURT
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS
```

Statistics

		Visits to health professionals	Attitudes toward medication	Attitudes toward housework	Income	Whether currently married	Race
N	Valid	465	465	464	439	465	465
	Missing	0	0	1	26	0	0
Mean		7.90	7.69	23.54	4.21	1.78	1.09
Std. Deviation		10.948	1.156	4.484	2.419	.416	.284
Variance		119.870	1.337	20.102	5.851	.173	.081
Skewness		3.248	−.123	−.457	.582	−1.346	2.914
Std. Error of Skewness		.113	.113	.113	.117	.113	.113
Kurtosis		13.101	−.447	1.556	−.359	−.190	6.521
Std. Error of Kurtosis		.226	.226	.226	.233	.226	.226
Minimum		0	5	2	1	1	1
Maximum		81	10	35	10	2	2

TABLE 4.5 Continued

Histogram



extremely so, and the standard deviation (Std. Deviation) is 10.948. These values are all reasonable, as are the values on the other variables. For instance, the ATTDRUG variable is constructed with a range of 5 to 10, so it is reassuring to find these values as Minimum and Maximum.

TIMEDRS shows no Missing cases but has strong positive Skewness (3.248). The significance of Skewness is evaluated by dividing it by Std. Error of Skewness, as in Equation 4.5,

$$z = \frac{3.248}{.113} = 28.74$$

to reveal a clear departure from symmetry. The distribution also has significant Kurtosis as evaluated by Equation 4.7,

$$z = \frac{13.101}{.22} = 57.97$$

The departures from normality are also obvious from inspection of the difference between frequencies expected under the normal distribution (the superimposed curve) and obtained frequencies. Because this variable is a candidate for transformation, evaluation of univariate outliers is deferred.

ATTDRUG, on the other hand, is well behaved. There are no Missing cases, and Skewness and Kurtosis are well within expected values. ATTHOUSE has a single missing value but is otherwise well distributed except for the two extremely low scores. The score of 2 is 4.8 standard deviations below the mean of ATTHOUSE (well beyond the $p = .001$ criterion of 3.29, two-tailed) and is disconnected from the other cases. It is not clear whether these are recording errors or if these two women actually enjoy housework that much. In any event, the decision is made to delete from further analysis the data from the two women with extremely favorable attitudes toward housework.

Information about these deletions is included in the report of results. The single missing value is replaced with the mean. (Section 10.7.1.1 illustrates a more sophisticated way of dealing with missing data in IBM SPSS when the amount missing is greater than 5%.)

On INCOME, however, there are 26 cases with Missing values—more than 5% of the sample. If INCOME is not critical to the hypotheses, we delete it in subsequent analyses. If INCOME is important to the hypotheses, we could replace the missing values.

The two remaining variables are dichotomous and not evenly split. MSTATUS has a 362 to 103 split, roughly a 3.5 to 1 ratio, that is not particularly disturbing. But RACE, with a split greater than 10 to 1 is marginal. For this analysis, we choose to retain the variable, realizing that its association with other variables is deflated because of the uneven split.

Table 4.6 shows the distribution of ATTHOUSE with elimination of the univariate outliers. The mean for ATTHOUSE changes to 23.634, the value used to replace the missing ATTHOUSE score in subsequent analyses. The case with a missing value on ATTHOUSE becomes complete and available for use in all computations. The COMPUTE instructions filter out cases with values equal to or less than 2 on ATTHOUSE (univariate outliers) and the RECODE instruction sets the missing value to 23.63.

At this point, we have investigated the accuracy of data entry and the distributions of all variables, determined the number of missing values, found the mean for replacement of missing data, and found two univariate outliers that, when deleted, result in $N = 463$.

TABLE 4.6 Syntax and IBM SPSS FREQUENCIES Output Showing Descriptive Statistics and Histograms for ATTHOUSE With Univariate Outliers Deleted

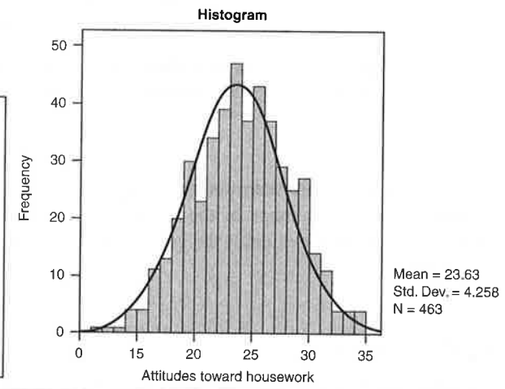
```
USE ALL.
COMPUTE filter_$=(atthouse > 2).
VARIABLE LABEL filter_$ 'atthouse > 2 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
RECODE
 atthouse (SYSMIS=23.63).
EXECUTE.
FREQUENCIES
 VARIABLES=atthouse /FORMAT=NOTABLE
 /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM MEAN SKEWNESS SESKEW KURTOSIS
 SEKURT
 /HISTOGRAM NORMAL
 /ORDER= ANALYSIS.
```

Frequencies

Statistics

Attitudes toward housework

N	Valid	463
	Missing	0
Mean		23.63
Std. Deviation		4.258
Variance		18.128
Skewness		−.038
Std. Error of Skewness		.113
Kurtosis		−.254
Std. Error of Kurtosis		.226
Minimum		11
Maximum		35



4.2.1.2 *Linearity and Homoscedasticity*

Owing to nonnormality on at least one variable, IBM SPSS GRAPH is run to check the bivariate plots for departures from linearity and homoscedasticity, as reproduced in Figure 4.9. The variables picked as *worst case* are those with the most discrepant distributions: TIMEDRS, which has the greatest departure from normality, and ATTDRUG, which is nicely distributed. (The SELECT IF instruction eliminates the univariate outliers on ATTHOUSE.)

In Figure 4.9, ATTDRUG is along the *Y* axis; turn the page so that the *Y* axis becomes the *X* axis and you can see the symmetry of the ATTDRUG distribution. TIMEDRS is along the *X* axis.

```
GRAPH
 /SCATTERPLOT(BIVAR)=timedrs WITH attdrug
 /MISSING=LISTWISE.
```

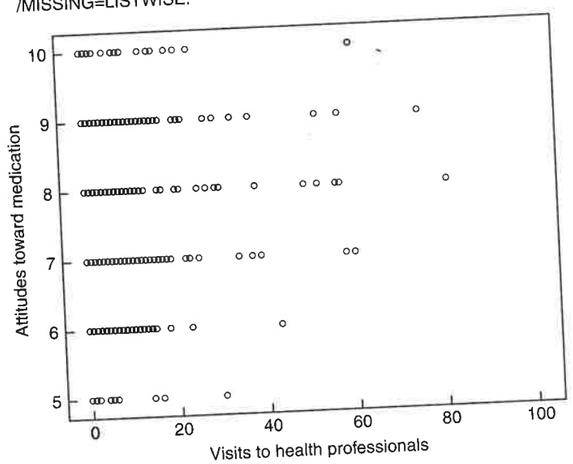


FIGURE 4.9 Assessment of linearity through bivariate scatterplots, as produced by IBM SPSS GRAPH. This indicates ATTDRUG is normal; TIMEDRS is nonnormal.

The asymmetry of the distribution is apparent from the pileup of scores at low values of the variable. The overall shape of the scatterplot is not oval; the variables are not linearly related. Heteroscedasticity is evident in the greater variability in ATTDRUG scores for low than high values of TIMEDRS.

4.2.1.3 *Transformation*

Variables are transformed prior to searching for multivariate outliers. A logarithmic transformation is applied to TIMEDRS to overcome strong skewness. Because the smallest value on the variable is zero, one is added to each score as the transformation is performed, as indicated in the COMPUTE statement. Table 4.7 shows the distribution of TIMEDRS as transformed to LTIMEDRS.

Skewness is reduced from 3.248 to 0.221 and Kurtosis is reduced from 13.101 to −0.183 by the transformation. The frequency plot is not exactly pleasing (the frequencies are still too high for small scores), but the statistical evaluation of the distribution is much improved.

Figure 4.10 is a bivariate scatterplot between ATTDRUG and LTIMEDRS. Although still not perfect, the overall shape of the scatterplot is more nearly oval. The nonlinearity associated with nonnormality of one of the variables is "fixed" by transformation of the variable.

TABLE 4.7 Syntax and IBM SPSS FREQUENCIES Output Showing Descriptive Statistics and Histograms for TIMEDRS After Logarithmic Transform

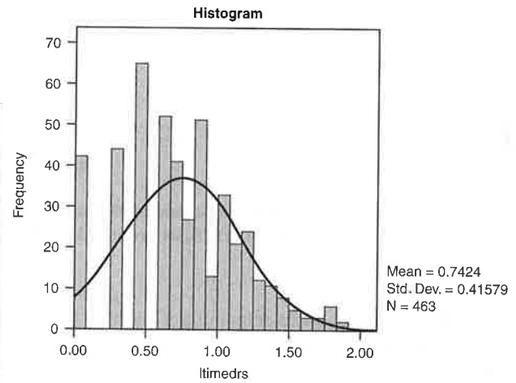
```
COMPUTE ltimedrs=lg10(timedrs+1).
EXECUTE.
FREQUENCIES
 VARIABLES=ltimedrs/FORMAT=NOTABLE
 /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM MEAN SKEWNESS SESKEW KURTOSIS
 SEKURT
 /HISTOGRAM NORMAL
 /ORDER=ANALYSIS.
```

Frequencies

Statistics

ltimedrs

N	Valid	463
	Missing	0
Mean		.7424
Std. Deviation		.41579
Variance		.173
Skewness		.221
Std. Error of Skewness		.113
Kurtosis		−.183
Std. Error of Kurtosis		.226
Minimum		.00
Maximum		1.91



4.2.1.4 *Detecting Multivariate Outliers*

The 463 cases, with transformation applied to LTIMEDRS, are screened for multivariate outliers through IBM SPSS REGRESSION (Table 4.8) using the RESIDUALS=OUTLIERS(MAHAL) syntax added to menu choices. Case labels (SUBNO) are used as the dummy DV—convenient because multivariate outliers among IVs are unaffected by the DV.[21] The remaining VARIABLES are considered independent variables.

The criterion for multivariate outliers is Mahalanobis distance at $p < .001$. Mahalanobis distance is evaluated as χ^2 with degrees of freedom equal to the number of variables, in this case five: LTIMEDRS, ATTDRUG, ATTHOUSE, MSTATUS, and RACE. Any case with a Mahal. Distance in Table 4.8 greater than $\chi^2_5 = 20.515$ (cf. Appendix C, Table C.4), then, is a multivariate outlier. As shown in Table 4.8, cases 117 and 193 are outliers among these variables in this data set.

[21]For a multiple regression analysis, the actual DV would be used here rather than SUBNO as a dummy DV.

```
GRAPH
 /SCATTERPLOT(BIVAR)=ltimedrs WITH attdrug
 /MISSING=LISTWISE.
```

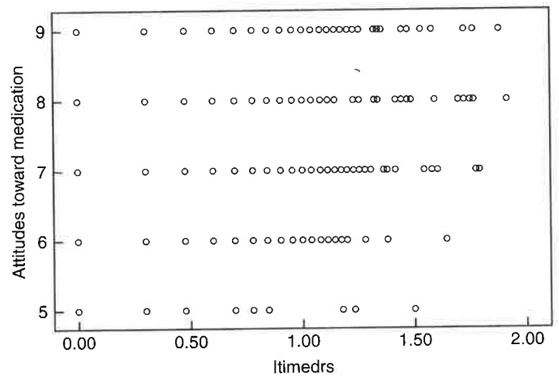


FIGURE 4.10 Assessment of linearity after log transformation of TIMEDRS, as produced by IBM SPSS GRAPH.

There are 461 cases remaining if the two multivariate outliers are deleted. Little is lost by deleting the additional two outliers from the sample. It is still necessary to determine why the two cases are multivariate outliers, to better understand how their deletion limits generalizability, and to include that information in the Results section.

4.2.1.5 Variables Causing Cases to Be Outliers

IBM SPSS REGRESSION is used to identify the combination of variables on which case 117 (subject number 137 as found in the data editor) and case 193 (subject number 262) deviate from the remaining 462 cases. Each outlying case is evaluated in a separate IBM SPSS REGRESSION run

TABLE 4.8 Syntax and Selected IBM SPSS REGRESSION Output for Multivariate Outliers and Multicollinearity

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT subno
 /METHOD=ENTER attdrug atthouse mstatus race ltimedrs
 /RESIDUALS=OUTLIERS(MAHAL).
```

TABLE 4.8 Continued

Regression

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions: (Constant)	Attitudes toward medication	Attitudes toward housework	Whether currently married	Race	ltimedrs
1	1	5.656	1.000	.00	.00	.00	.00	.00	.01
	2	.210	5.193	.00	.00	.00	.01	.02	.92
	3	.060	9.688	.00	.00	.01	.29	.66	.01
	4	.043	11.508	.00	.03	.29	.46	.16	.06
	5	.025	15.113	.00	.53	.41	.06	.04	.00
	6	.007	28.872	.99	.43	.29	.18	.12	.00

a. Dependent Variable: Subject number

Outlier Statistics[a]

	Case Number	Statistic
Mahal. Distance 1	117	21.837
2	193	20.650
3	435	19.968
4	99	18.499
5	335	18.469
6	292	17.518
7	58	17.373
8	71	17.172
9	102	16.942
10	196	16.723

a. Dependent Variable: Subject number

TABLE 4.9 IBM SPSS REGRESSION Syntax and Partial Output Showing Variables Causing the 117th Case to Be an Outlier

```
COMPUTE dummy = 0.
EXECUTE.
IF (subno=137) dummy = 1.
EXECUTE.
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT dummy
 /METHOD=STEPWISE attdrug atthouse emplmnt mstatus race ltimedrs.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−.024	.008		−2.881	.004
	race	.024	.008	.149	3.241	.001
2	(Constant)	.009	.016		.577	.564
	race	.025	.007	.151	3.304	.001
	Attitudes toward medication	−.004	.002	−.111	−2.419	.016
3	(Constant)	.003	.016		.169	.866
	race	.026	.007	.159	3.481	.001
	Attitude toward medication	−.005	.002	−.123	−2.681	.008
	ltimedrs	.012	.005	.109	2.360	.019

a. Dependent Variable: DUMMY

after a dummy variable is created to separate the outlying case from the remaining cases. In Table 4.9, the dummy variable for subject 137 is created in the COMPUTE instruction with dummy = 0 and if (subno=137) dummy = 1. With the dummy variable as the DV and the remaining variables as IVs, you can find the variables that distinguish the outlier from the other cases.

For the 117th case (subno=137), RACE, ATTDRUG, and LTIMEDRS show up as significant predictors of the case (Table 4.9).

Variables separating case 193 (subno=262) from the other cases are RACE and LTIMEDRS, as seen in Table 4.10. The final step in evaluating outlying cases is to determine how their scores on the variables that cause them to be outliers differ from the scores of the remaining sample. The IBM SPSS LIST and DESCRIPTIVES procedures are used, as shown in Table 4.11. The LIST

TABLE 4.10 IBM SPSS REGRESSION Syntax and Partial Output Showing Variables Causing the 193rd Case to Be an Outlier

```
COMPUTE dummy = 0.
EXECUTE.
IF (subno=262) dummy = 1.
EXECUTE.
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT dummy
 /METHOD=STEPWISE attdrug atthouse emplmnt mstatus race ltimedrs.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−.024	.008		−2.881	.004
	race	.024	.008	.149	3.241	.001
2	(Constant)	−.036	.009		−3.787	.000
	race	.026	.007	.158	3.436	.001
	ltimedrs	.014	.005	.121	2.634	.009

a. Dependent Variable: DUMMY

TABLE 4.11 Syntax and IBM SPSS Output Showing Variable Scores for Multivariate Outliers and Descriptive Statistics for All Cases

```
LIST VARIABLES=subno attdrug atthouse mstatus race ltimedrs
 /CASES FROM 117 TO 117.
```

List

subno	attdrug	atthouse	mstatus	race	ltimedrs
137	5	24	2	2	1.49

```
LIST VARIABLES=subno attdrug atthouse mstatus race ltimedrs
 /CASES FROM 193 TO 193.
```

List

subno	attdrug	atthouse	mstatus	race	ltimedrs
262	9	31	2	2	1.72

TABLE 4.11 Continued

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Attitudes toward medication	463	5	10	7.68	1.158
Attitudes toward housework	463	11	35	23.63	4.258
Whether currently married	463	1	2	1.78	.413
race	463	1	2	1.09	.284
ltimedrs	463	.00	1.91	.7424	.41579
Valid N (listwise)	463				

procedure is run for each outlying case to show its values on all the variables of interest. Then DESCRIPTIVES is used to show the average values for the 463 cases remaining after deletion of the two univariate outliers against which the values for the outlying cases are compared.[22]

The 117th case is nonwhite on RACE, has very unfavorable attitudes regarding use of drugs (the lowest possible score on ATTDRUG), and has a high score on LTIMEDRS. The 193rd case is also nonwhite on RACE and has a very high score on LTIMEDRS. There is some question, then, about the generalizability of subsequent findings to nonwhite women who make numerous visits to physicians, especially in combination with unfavorable attitude toward use of drugs.

4.2.1.6 Multicollinearity

Evaluation of multicollinearity is produced in IBM SPSS through the STATISTICS COLLIN instruction. As seen by the Collinearity Diagnostics output of Table 4.8, no multicollinearity is evident. Although the last root has a Condition Index that approaches 30, no dimension (row) has more than one Variance Proportion greater than .50.

Screening information as it might be described in a Results section of a journal article appears next.

> Results
>
> Prior to analysis, number of visits to health professionals, attitude toward drug use, attitude toward housework, income, marital status, and race were examined through various IBM SPSS programs for accuracy of data entry, missing values, and fit

[22]These values are equal to those shown in the earlier FREQUENCIES runs but for deletion of the two univariate outliers.

> between their distributions and the assumptions of multivariate analysis. The single missing value on attitude toward housework was replaced by the mean, while income, with missing values on more than 5% of the cases, was deleted. The poor split on race (424 to 41) truncates its correlations with other variables, but it was retained for analysis. To improve pairwise linearity and to reduce the extreme skewness and kurtosis, visits to health professionals was logarithmically transformed.
>
> Two cases with extremely low z scores on attitude toward housework were found to be univariate outliers; two other cases were identified through Mahalanobis distance as multivariate outliers with $p < .001$.[23]
>
> All four outliers were deleted, leaving 461 cases for analysis.

4.2.2 Screening Grouped Data

For this example, the cases are divided into two groups according to the EMPLMNT (employment status) variable; there are 246 cases who have paid work (EMPLMNT = 0) and 219 cases who are housewives (EMPLMNT = 1). For illustrative purposes, variable transformation is considered inappropriate for this example, to be undertaken only if proved necessary. A flow diagram for screening grouped data appears in Figure 4.11.

4.2.2.1 Accuracy of Input, Missing Data, Distributions, Homogeneity of Variance, and Univariate Outliers

SAS MEANS and Interactive Data Analysis provide descriptive statistics and histograms, respectively, for each group separately, as shown in Table 4.12. Menu choices are shown for SAS Interactive Data Analysis because no SAS log file is provided. Note that data must be sorted by EMPLMNT group before analyzing separately by groups. As with ungrouped data, accuracy of input is judged by plausible **Means** and **Std Devs** and reasonable **Maximum** and **Minimum** values. The distributions are judged by their overall shapes within each group. TIMEDRS is just as badly skewed when grouped as when ungrouped, but this is of less concern when dealing with sampling distributions based on over 200 cases unless the skewness causes nonlinearity among variables or there are outliers.

[23]Case 117 was nonwhite with very unfavorable attitudes regarding use of drugs but numerous visits to physicians. Case 193 was also nonwhite with numerous visits to physicians. Results of analysis may not generalize to nonwhite women with numerous visits to physicians, particularly if they have very unfavorable attitude toward use of drugs.

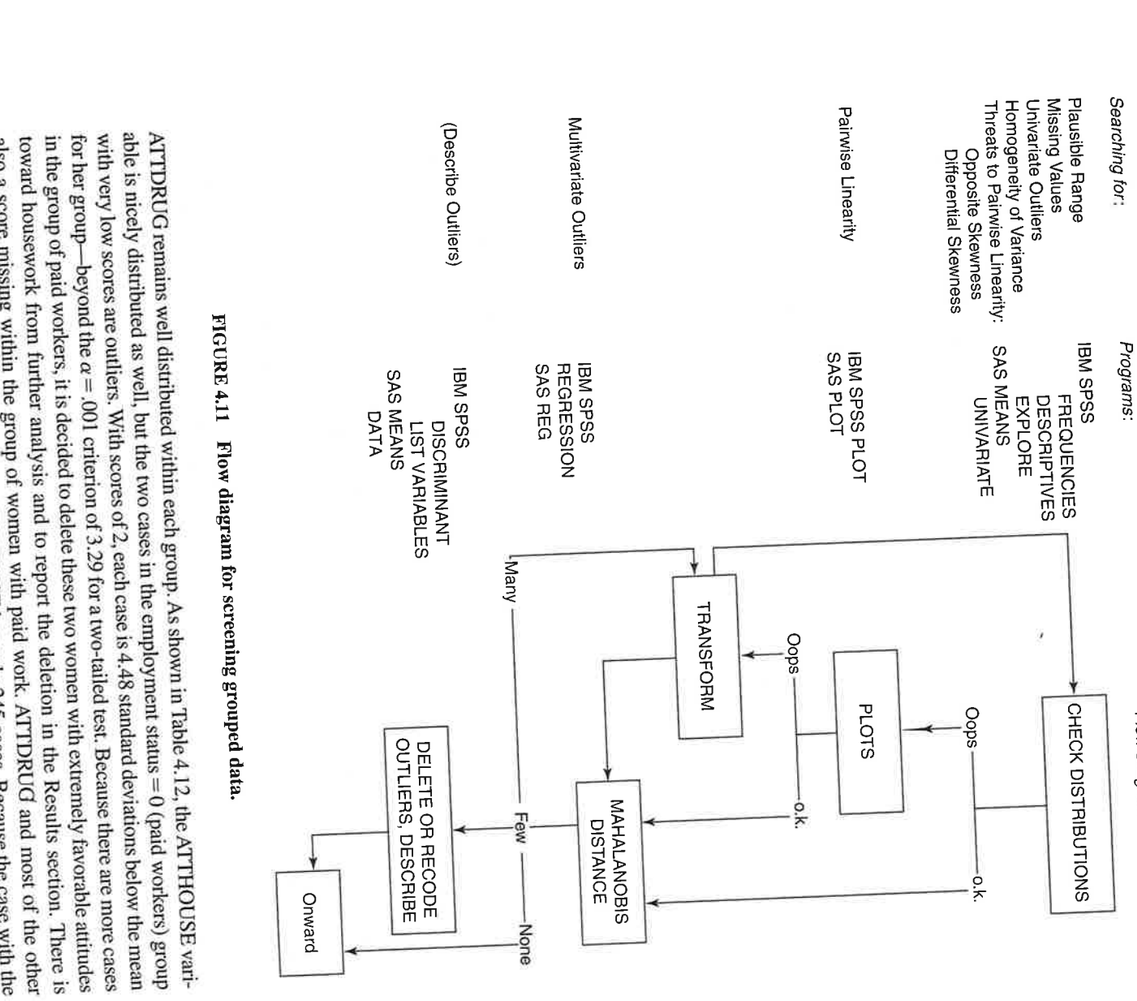


FIGURE 4.11 Flow diagram for screening grouped data.

ATTDRUG remains well distributed within each group. As shown in Table 4.12, the ATTHOUSE variable is nicely distributed as well, but the two cases in the employment status = 0 (paid workers) group with very low scores are outliers. With scores of 2, each case is 4.48 standard deviations below the mean for her group—beyond the $\alpha = .001$ criterion of 3.29 for a two-tailed test. Because there are more cases in the group of paid workers, it is decided to delete these two women with extremely favorable attitudes toward housework from further analysis and to report the deletion in the Results section. There is also a score missing within the group of women with paid work. ATTDRUG and most of the other variables have 246 cases in this group, but ATTHOUSE has only 245 cases. Because the case with the missing value is from the larger group, it is decided to delete the case from subsequent analyses.

TABLE 4.12 Syntax and Selected SAS MEANS and SAS Interactive Data Analysis Output Showing (a) Descriptive Statistics and (b) Histograms for Grouped Data

```
proc sort data = SASUSER.SCREEN;
  by EMPLMNT;
run;
proc means data=SASUSER.SCREEN    vardef=DF
  N NMISS MIN MAX MEAN VAR STD SKEWNESS KURTOSIS;
  Var TIMEDRS ATTDRUG ATTHOUSE INCOME MSTATUS RACE;
by EMPLMNT;
run;
```

---------- Current employment status=0 ----------

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum	Mean
TIMEDRS	Visits to health professionals	246	0	0	81.0000000	7.2926829
ATTDRUG	Attitudes toward use of medication	246	0	5.0000000	10.0000000	7.5934959
ATTHOUSE	Attitudes toward housework	245	1	2.0000000	34.0000000	23.6408163
INCOME	Income code	235	11	1.0000000	10.0000000	4.2382979
MSTATUS	Current marital status	246	0	1.0000000	2.0000000	1.6869919
RACE	Ethnic affiliation	246	0	1.0000000	2.0000000	1.1097561

Variable	Label	Variance	Std Dev	Skewness	Kurtosis
TIMEDRS	Visits to health professionals	122.4527626	11.0658376	3.8716749	18.0765985
ATTDRUG	Attitudes toward use of medication	1.2381616	1.1127271	-0.1479593	-0.4724261
ATTHOUSE	Attitudes toward housework	23.3376715	4.8309079	-0.6828286	2.1614074
INCOME	Income code	5.9515185	2.4395734	0.5733054	-0.4287488
MSTATUS	Current marital status	0.2159117	0.4646630	-0.8114465	-1.3526182
RACE	Ethnic affiliation	0.0981085	0.3132228	2.5122223	4.3465331

---------- Current employment status = 1 ----------

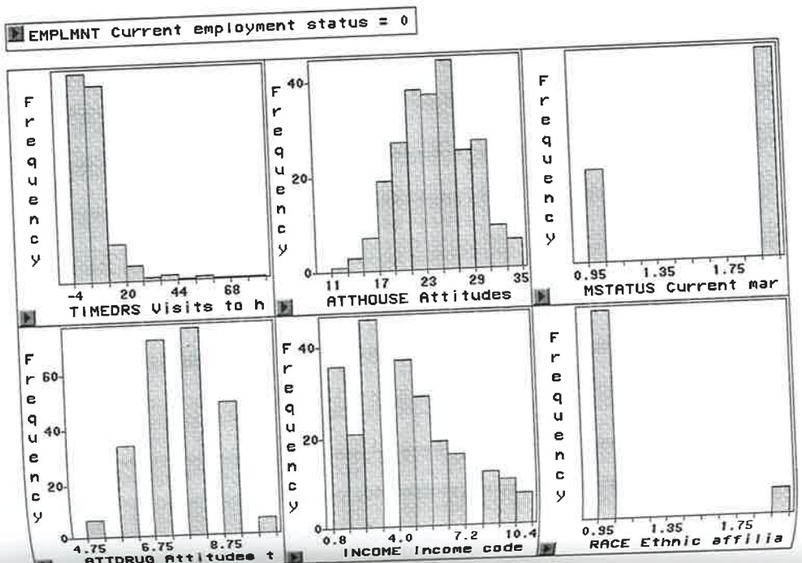
Variable	Label	N	N Miss	Minimum	Maximum	Mean
TIMEDRS	Visits to health professionals	219	0	0	60.0000000	8.5844749
ATTDRUG	Attitudes toward use of medication	219	0	5.0000000	10.0000000	7.7899543
ATTHOUSE	Attitudes toward housework	219	0	11.0000000	35.0000000	23.4292237
INCOME	Income code	204	15	1.0000000	10.0000000	4.1764706
MSTATUS	Current marital status	219	0	1.0000000	2.0000000	1.8812785
RACE	Ethnic affiliation	219	0	1.0000000	2.0000000	1.0639269

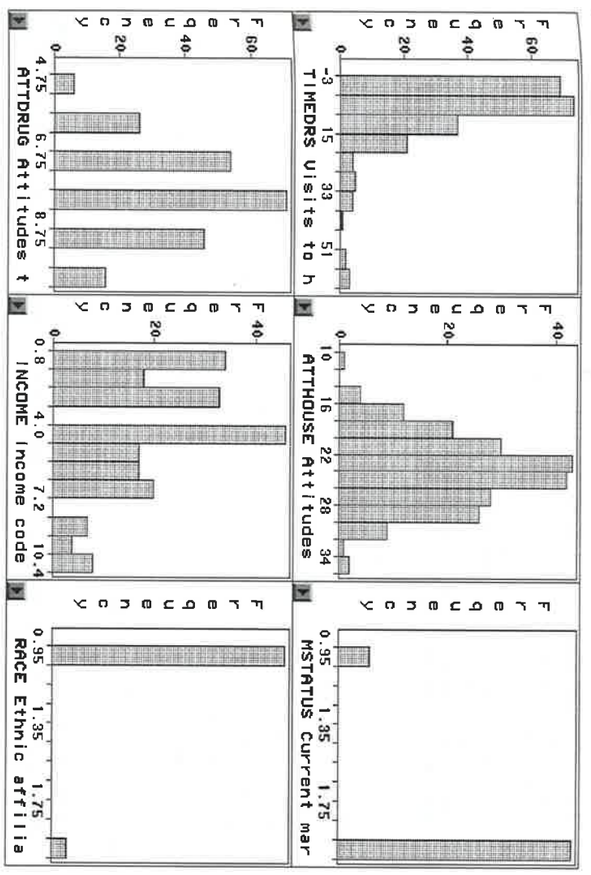
(*continued*)

TABLE 4.12 Continued

Variable	Label	Variance	Std Dev	Skewness	Kurtosis
TIMEDRS	Visits to health professionals	116.6292991	10.7995046	2.5624008	7.8645362
ATTDRUG	Attitudes toward use of medication	1.4327427	1.1969723	-0.1380901	-0.4417282
ATTHOUSE	Attitudes toward housework	16.5488668	4.0680298	-0.0591932	0.0119403
INCOME	Income code	5.7618082	2.4003767	0.5948250	-0.2543926
MSTATUS	Current marital status	0.1051066	0.3242015	-2.3737868	3.6682817
RACE	Ethnic affiliation	0.0601148	0.2451832	3.5899053	10.9876845

(b) 1. Open SAS Interactive Data Analysis with appropriate data set (here Sasuser.Screen).
2. Choose Analyze and then Histogram/Bar Chart(Y).
3. Select Y variables: TIMEDRS ATTDRUG ATTHOUSE INCOME MSTATUS RACE.
4. In Group box, select EMPLMNT.
5. In Output dialog box, select Both, Vertical Axis at Left, and Horizontal Axis at Bottom.





On INCOME, however, it is the smaller group, housewives, with the greater number of missing values; within that group almost 7% of the cases do not have INCOME scores. INCOME, then, is a good candidate for variable deletion, although other remedies are available should deletion seriously interfere with hypothesis testing.

The splits in the two dichotomous variables, MSTATUS and RACE, are about the same for grouped as for ungrouped data. The splits for both MSTATUS (for the housewife group) and for RACE (both groups) are disturbing, but we choose to retain them here.

For the remaining analyses, INCOME is deleted as a variable, and the case with the missing value as well as the two univariate outliers on ATTHOUSE are deleted, leaving a total sample size of 462, 243 cases in the paid work group and 219 cases in the housewife group.

Because cell sample sizes are not very discrepant, variance ratios as great as 10 can be tolerated. All F_{max} are well below this criterion. As an example, for the two groups for ATTDRUG, $F_{max} = 1.438/1.238 = 1.16$. Thus, there is no concern about violation of homogeneity of variance–covariance matrices.

```
data SASUSER.SCREENX;
  set SASUSER.SCREEN;
  if ATTHOUSE < 3 then delete;
run;

1. Open SAS Interactive Data Analysis with appropriate data set (here-
   SASUSER.SCREENX).
2. Choose Analyze and the Scatter Plot (YX).
3. Select Y variable: ATTDRUG.
4. Select X variable: TIMEDRS.
5. In Group box, select EMPLMNT.
6. In Output dialog box, select Names, Y Axis Vertical, Vertical Axis
   at Left, and Horizontal Axis at Bottom.
```

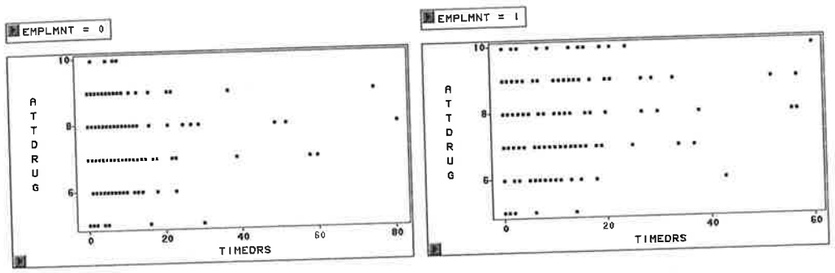


FIGURE 4.12 Syntax for creating reduced data set through SAS DATA and SAS Interactive Data Analysis setup and output showing within-group scatterplot of ATTDRUG versus TIMEDRS.

4.2.2.2 *Linearity*

Owing to the poor distribution on TIMEDRS, a check of scatterplots is warranted to see if TIMEDRS has a linear relationship with other variables. There is no need to check for linearity with MSTATUS and RACE because variables with two levels have only linear relationships with other variables. Of the two remaining variables, ATTHOUSE and ATTDRUG, the distribution of ATTDRUG differs most from that of TIMEDRS after univariate outliers are deleted.

Appropriately checked first, then, are within-group scatterplots of ATTDRUG versus TIMEDRS. Figure 4.12 shows syntax for creating the data set with extreme cases and missing data on ATTHOUSE deleted (SASUSER.SCREENX) and the setup and output for creating the scatterplots.

In the within-group scatterplots of Figure 4.12, there is ample evidence of skewness in the bunching up of scores at low values of TIMEDRS, but no suggestion of nonlinearity for these variables in the group of paid workers or housewives. Because the plots are acceptable, there is no evidence that the extreme skewness of TIMEDRS produces a harmful departure from linearity. Nor is there any reason to expect nonlinearity with the symmetrically distributed ATTHOUSE.

TABLE 4.13 Syntax for SAS REG and Selected Portion of Output File for Identification of Multivariate Outliers

```
proc reg data=SASUSER.SCREENX;
  by EMPLMNT
  model SUBNO = TIMEDRS ATTDRUG ATTHOUSE MSTATUS RACE;
  output out=SASUSER.SCRN_LEV H=H;
run;
```

SASUSER.SCRN_LEV

462	SUBNO	TIMEDRS	ATTDRUG	ATTHOUSE	INCOME	EMPLMNT	MSTATUS	RACE	H
1	2	3	7	20	6	0	2	1	0.0106
2	3	0	8	23	3	0	2	1	0.0089
3	6	3	8	25	4	0	2	1	0.0080
4	10	4	8	30	8	0	1	1	0.0231
5	16	2	8	19	3	0	2	2	0.0440
6	22	2	6	21	7	0	1	1	0.0256
7	24	5	10	25	9	0	2	1	0.0269
8	25	3	6	19	4	0	2	1	0.0188
9	26	4	5	31	5	0	2	1	0.0418
10	27	2	8	25	2	0	2	1	0.0084
11	29	13	9	26	2	0	2	1	0.0150
12	30	7	9	33	1	0	2	1	0.0302
13	33	2	6	27	3	0	1	1	0.0273
14	34	5	9	30	1	0	2	1	0.0212
15	35	4	7	31	4	0	2	1	0.0197
16	36	6	6	25	5	0	2	1	0.0149
17	37	2	9	27	5	0	2	1	0.0165
18	40	7	7	32	3	0	2	1	0.0222
19	46	1	8	28	6	0	2	2	0.0478
20	47	3	8	16	1	0	2	1	0.0207
21	48	60	7	24	1	0	2	1	0.1036
22	49	5	8	19	1	0	2	1	0.0124
23	50	3	9	23	6	0	2	1	0.0144

4.2.2.3 *Multivariate Outliers*

Multivariate outliers within the groups are sought using SAS REG with EMPLMNT as the DV. Because outliers are sought only among IVs, the choice of the DV is simply one of convenience. Mahalanobis distance is unavailable directly but may be calculated from leverage values, which are added to the data file. Table 4.13 shows SAS REG syntax and a selected portion of the output data file (SASUSER.SCRN_LEV) that provides leverage (H) for each case from the centroid of each group. Missing data and univariate outliers have already been omitted (see syntax for Figure 4.12).

Critical values of Mahalanobis distance with five variables at $\alpha = .001$ is 20.515 (Appendix C, Table C.4). Equation 4.3 is used to convert this to a critical leverage value for each group.

For housewives,

$$h_{ii} = \frac{20.515}{219 - 1} + \frac{1}{219} = 0.099$$

TABLE 4.14 SAS DATA Syntax for Transformation of TIMEDRS; Syntax For SAS REG and Selected Portion of Output File

```
data SASUSER.SCREENT;
 set SASUSER.SCREENX;
 LTIMEDRS = log10(TIMEDRS + 1) ;
run;
proc reg data=SASUSER.SCREENT;
 by EMPLMNT;
 model SUBNO = LTIMEDRS ATTDRUG ATTHOUSE MSTATUS RACE;
 output out=sasuser.Scrnlev2 H=H;
run;
```

SASUSER.SCRNLEV2

462	SUBNO	TIMEDRS	ATTDRUG	ATTHOUSE	INCOME	EMPLMNT	MSTATUS	RACE	LTIMEDR	H
1	2	3	7	20	6	0	2	1	0.6021	0.0105
2	3	0	8	23	3	0	2	1	0.0000	0.0213
3	6	3	8	25	4	0	2	1	0.6021	0.0078
4	10	4	8	30	8	0	1	1	0.6990	0.0219
5	16	2	8	19	3	0	2	2	0.4771	0.0446
6	22	2	6	21	7	0	1	1	0.4771	0.0261
7	24	5	10	25	9	0	2	1	0.7782	0.0266
8	25	3	6	19	4	0	2	1	0.6021	0.0188
9	26	4	5	31	5	0	2	1	0.6990	0.0412
10	27	2	8	25	2	0	2	1	0.4771	0.0091
11	29	13	9	26	2	0	2	1	1.1461	0.0184
12	30	7	9	33	1	0	1	1	0.9031	0.0300
13	33	2	6	27	3	0	2	1	0.4771	0.0277
14	34	5	9	30	1	0	2	1	0.7782	0.0205
15	35	4	7	31	4	0	2	1	0.6990	0.0189
16	36	6	6	25	5	0	2	1	0.8451	0.0153
17	37	2	9	27	5	0	2	1	0.4771	0.0173
18	40	7	7	32	3	0	2	1	0.9031	0.0223
19	46	1	8	28	6	0	2	2	0.3010	0.0514
20	47	3	8	16	1	0	2	1	0.6021	0.0207
21	48	60	7	24	1	0	2	1	1.7853	0.0383
22	49	5	8	19	1	0	2	1	0.7782	0.0127
23	50	3	9	23	6	0	2	1	0.6021	0.0142

and for the paid work group,

$$h_{ii} = \frac{20.515}{243 - 1} + \frac{1}{243} = 0.089$$

The data set is shown for the first cases for the paid workers (group = 0). The last column, labeled H, shows leverage values. The 21st case, SUBNO #48 is a multivariate outlier among paid workers.

Altogether, 12 cases (about 2.5%) are identified as multivariate outliers: 4 paid workers and 8 housewives. Although this is not an exceptionally large number of cases to delete, it is worth investigating alternative strategies for dealing with outliers. The univariate summary statistics for TIMEDRS in Table 4.12 show a Maximum score of 81, converting to a standard score of $z = (81 - 7.293)/11.066 = 6.66$ among those with PAIDWORK and $z = 4.76$ among HOUSEWIVEs; the poorly distributed variable produces univariate outliers in both groups. The skewed histograms of Table 4.13 suggest a logarithmic transformation of TIMEDRS.

TABLE 4.15 SAS DATA Syntax for Forming the Dummy DV and Limiting Data to Housewives; Syntax and Partial SAS REG Output Showing Variables Causing SUBNO #262 to Be an Outlier Among Housewives

```
data SASUSER.SCREEND;
 set SASUSER.SCREENT;
  DUMMY = 0;
 if SUBNO=262 then DUMMY=1;
 if EMPLMNT=0 then delete;
 run;
proc reg DATA=SASUSER.SCREEND;
model DUMMY = LTIMEDRS ATTDRUG ATTHOUSE MSTATUS RACE/
 selection=FORWARD SLENTRY=0.05;
run;
```

```
                    The REG Procedure
                     Model: MODEL1
                Dependent Variable: DUMMY
                Forward Selection: Step 2

                  Parameter      Standard
Variable           Estimate         Error    Type II SS    F Value    Pr > F

Intercept           -0.09729       0.02167       0.08382      20.16    <.0001
LTIMEDR              0.02690       0.00995       0.03041       7.31    0.0074
RACE                 0.07638       0.01791       0.07565      18.19    <.0001

          Bounds on condition number: 1.0105, 4.0422

No other variable met the 0.0500 significance level for entry into
the model.
```

Table 4.14 shows output from a second run of SAS REG identical to the run in Table 4.13 except that TIMEDRS is replaced by LTIMEDR, its logarithmic transform (transformation syntax is shown). The 21st case no longer is an outlier. With the transformed variable, the entire data set contains only five multivariate outliers, all of them in the housewife group. One of these, SUBNO #262 was also identified as a multivariate outlier in the ungrouped data set.[24]

4.2.2.4 Variables Causing Cases to Be Outliers

Identification of the variables causing outliers to be extreme proceeds in the same manner as for ungrouped data except that the values for the case are compared with the means for the group the case comes from. For subject number 262, a paid worker, the regression run is limited to paid workers, as shown in Table 4.15. First, Table 4.15 shows the SAS DATA syntax to create a dummy variable in which SUBNO #262 forms one code of the dummy variable and the remaining housewives form the

[24]Note that these are different multivariate outliers than found by software used in earlier editions of this book.

TABLE 4.16 Syntax and SAS MEANS Output Showing Descriptive Statistics for Housewife Group

```
data SASUSER.SCREEND;
  set SASUSER.SCREENT;
 if EMPLMNT=0 then delete;
 run;
proc means data=SASUSER.SCREEND
  N MEAN STD;
 var LTIMEDRS ATTDRUG MSTATUS RACE;
 by EMPLMNT;
run;
```

---------- Current employment status=1 ----------

The MEANS Procedure

Variable	Label	N	Mean	Std Dev
LTIMEDR		219	0.7657821	0.4414145
ATTDRUG	Attitudes toward use of medication	219	7.7899543	1.1969723
MSTATUS	Current marital status	219	1.8812785	0.3242015
RACE	Ethnic affiliation	219	1.0639269	0.2451832

other code. The dummy variable then serves as the DV in the SAS REG run. The table shows that the same variables cause this woman to be an outlier from her group as from the entire sample: She differs on the combination of RACE and LTIMEDRS.

Regression runs for the other four cases are not shown; however, the remaining four cases all differed on RACE. In addition, one case differed on ATTDRUG and another on MSTATUS.

As with ungrouped data, identification of variables on which cases are outliers is followed by an analysis of the scores on the variables for those cases. First, Table 4.16 shows the means on the three variables involved in outlying cases, separately by employment group. The data set is consulted for these values for the five outliers.

The data set shows that scores for subject number 262 are 2 for RACE and 1.763 for LTIMEDRS. For subject number 45, they are 2 for RACE and 1 (single) for MSTATUS. For subject number 119, they are 2 for RACE and 10 for ATTDRUG. For subject numbers 265 and 584, they are 2 for RACE. Thus, all of the outliers are non-Caucasian housewives. A separate run of descriptive statistics for housewives (not shown) reveals that only 14 of the housewives in the sample of 219 are non-Caucasian. One of them also has an unusually large number of visits to health professionals, one is unmarried, and one has exceptionally unfavorable attitudes regarding use of drugs.

4.2.2.5 Multicollinearity

The collinearity diagnostics of a SAS REG run are used to assess multicollinearity for the two groups, combined (Table 4.17) after deleting the five multivariate outliers.

Screening information as it might be described in a Results section of a journal article appears next.

TABLE 4.17 Sytnax and Selected Multicollinearity Output From SAS REG

```
data SASUSER.SCREENF;
   set SASUSER.SCREENT;
   if subno=45 or subno=265 or subno=119 or subno=262 or
      subno=584 then delete;
run;
proc reg data=SASUSER.SCREENF;
   model SUBNO= ATTDRUG ATTHOUSE MSTATUS RACE LTIMEDRS/COLLIN;
run;
```

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	5.66743	1.00000
2	0.20446	5.26483
3	0.05466	10.18261
4	0.04223	11.58407
5	0.02453	15.19939
6	0.00668	29.13622

The REG Procedure
Model: MODEL1
Dependent Variable: SUBNO

Collinearity Diagnostics

	Proportion of Variation					
Number	Intercept	ATTDRUG	ATTHOUSE	MSTATUS	RACE	LTIMEDRS
1	0.00026439	0.00066304	0.00090957	0.00148	0.00169	0.00581
2	0.00093585	0.00162	0.00193	0.01207	0.01614	0.92916
3	0.00033355	0.00118	0.00259	0.35672	0.62656	0.00305
4	0.00401	0.04221	0.28153	0.40511	0.20211	0.05695
5	0.00391	0.53054	0.43329	0.04863	0.03159	0.00408
6	0.99055	0.42378	0.27975	0.17598	0.12191	.00094416

Results

Prior to analysis, number of visits to health professionals, attitude toward drug use, attitude toward housework, income, marital status, and race were examined through various SAS programs for accuracy of data entry, missing values, and fit between their distributions and the assumptions of multivariate analysis. The variables were examined separately for the 246 employed women and the 219 housewives.

A case with a single missing value on attitude toward housework was deleted from the group of employed women, leaving 245 cases in that group. Income, with missing values on more than 5% of the cases, was deleted. Pairwise linearity was checked using within-group scatterplots and found to be satisfactory.

Two cases in the employed group were univariate outliers due to their extremely low *z* scores on attitude toward housework; these cases were deleted. By using Mahalanobis distance with $p < .001$, derived from leverage scores, 15 cases (about 3%) were identified as multivariate outliers in their own groups. Because several of these cases had extreme *z* scores on visits to health professionals and because that variable was severely skewed, a logarithmic transformation was applied. With the transformed variable in the variable set, only five cases were identified as multivariate outliers, all from the employed group.[25] With all seven outliers and the case with missing values deleted, 238 cases remained in the employed group and 214 in the group of housewives.

[25]All the outliers were non-Caucasian housewives. Thus, 36% (5/14) of the non-Caucasian housewives were outliers. One of them also had an unusually large number of visits to health professionals, one was unmarried, and one had exceptionally unfavorable attitudes regarding use of drugs. Thus, results may not generalize to non-Caucasian housewives, particularly those who are unmarried, make frequent visits to physicians, and have very unfavorable attitudes toward the use of drugs.

5 Multiple Regression

5.1 General Purpose and Description

Regression analyses are a set of statistical techniques that allow one to assess the relationship between one dependent variable (DV) and several independent variables (IVs). For example, is reading ability in primary grades (the DV) related to several IVs such as perceptual development, motor development, and age? The terms *regression* and *correlation* are used more or less interchangeably to label these procedures, although the term *regression* is often used when the intent of the analysis is prediction, and the term *correlation* is used when the intent is simply to assess the relationship between the DV and the IVs.

Multiple regression is a popular technique in many disciplines. For example, Stefl-Mabry (2003) used standard multiple regression to study satisfaction derived from various sources of information (word-of-mouth, expert oral advice, Internet, print news, nonfiction books, and radio and television news). Forty vignettes were developed with high and low levels of information and high and low consistency; individual regression analyses were performed for each of the 90 professional participants, and then their standardized regression coefficients were averaged. The normative participant was most satisfied by expert oral advice, with nonfiction books and word-of-mouth next in order to produce satisfaction. Participants were consistent in the satisfaction they derived from various sources in various vignettes.

Peguero and Bondy (2011) used data from the public-use Educational Longitudinal Study of 2002 (ELS:2002) to study the relationship between a student's immigration status and his/her relationship with teachers. The study used a weighted sample of 9,874 10th grade students (1,628 Latino/a, 1,129 Asian American, 1,491 Black/African American, and 5,626 White American students in public schools). Immigration status was dummy variable coded as first generation (the student is born outside the United States), second generation (the student is born in the United States but one or both parents are not), and third generation (both the student and parents are born in the United States). A sequential multiple-regression strategy was used where race/ethnicity, student achievement (math and verbal ETS scores), family SES, gender, region of country (Midwest, South, Northeast, West), and locale (urban and rural) were entered as Model 1; Model 2 added immigration status. In Model 1, racial/ethnic status, student achievement, and gender were significant predictors of the relationship with a teacher. In Model 2, those predictors remained and immigration status also entered as a powerful predictor. For Latino, Asian American, and Black/African American immigrants (but particularly Latino students), there was a strong positive relationship with teachers for first-generation immigrants, but the relationship "fell off" sharply for second- and third-generation immigrants. The deterioration in the relationship was not seen for White American immigrants.