An Introduction to Applied Multivariate Analysis

Tenko Raykov ♦ George A. Marcoulides

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
2 Park Square
Milton Park, Abingdon
Oxon OX14 4RN

Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8058-6375-8 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Introduction to applied multivariate analysis / by Tenko Raykov & George A. Marcoulides.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-8058-6375-8 (hardcover)
ISBN-10: 0-8058-6375-3 (hardcover)
1. Multivariate analysis. I. Raykov, Tenko. II. Marcoulides, George A.

QA278.I597 2008
519.5'35--dc22 2007039834

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Psychology Press Web site at
http://www.psypress.com

3

Data Screening and Preliminary Analyses

Results obtained through application of univariate or multivariate statistical methods will in general depend critically on the quality of the data and on the numerical magnitude of the elements of the data matrix as well as variable relationships. For this reason, after data are collected in an empirical study and before they are analyzed using a particular method(s) to respond to a research question(s) of concern, one needs to conduct what is typically referred to as data screening. These preliminary activities aim (a) to ensure that the data to be analyzed represent correctly the data originally obtained, (b) to search for any potentially very influential observations, and (c) to assess whether assumptions underlying the method(s) to be applied subsequently are plausible. This chapter addresses these issues.

3.1 Initial Data Exploration

To obtain veridical results from an empirical investigation, the data collected in it must have been accurately entered into the data file submitted to the computer for analysis. Mistakes committed during the process of data entry can be very costly and can result in incorrect parameter estimates, standard errors, and test statistics, potentially yielding misleading substantive conclusions. Hence, one needs to spend as much time as necessary to screen the data for entry errors, before proceeding with the application of any uni- or multivariate method aimed at responding to the posited research question(s). Although this process of data screening may be quite time consuming, it is an indispensable prerequisite of a trustworthy data analytic session, and the time invested in data screening will always prove to be worthwhile.

Once a data set is obtained in a study, it is essential to begin with proofreading the available data file. With a small data set, it may be best to check each original record (i.e., each subject's data) for correct entry. With larger data sets, however, this may not be a viable option, and so one may instead arrange to have at least two independent data entry sessions followed by a comparison of the resulting files. Where discrepancies are

found, examination of the raw (original) data records must then be carried out in order to correctly represent the data into a computer file to be analyzed subsequently using particular statistical methods. Obviously, the use of independent data entry sessions can prove to be expensive and time consuming. In addition, although such checks may resolve noted discrepancies when entering the data into a file, they will not detect possible common errors across all entry sessions or incorrect records in the original data. Therefore, for any data set once entered into a computer file and proofread, it is recommended that a researcher carefully examine frequencies and descriptive statistics for each variable across all studied persons. (In situations involving multiple-population studies, this should also be carried out within each group or sample.) Thereby, one should check, in particular, the range of each variable, and specifically whether the recorded maximum and minimum values on it make sense. Further, when examining each variable's frequencies, one should also check if all values listed in the frequency table are legitimate. In this way, errors at the data-recording stage can be spotted and immediately corrected.

To illustrate these very important preliminary activities, let us consider a study in which data were collected from a sample of 40 university freshmen on a measure of their success in an educational program (referred to below as ''exam score'' and recorded in a percentage correct metric), and its relationship to an aptitude measure, age in years, an intelligence test score, as well as a measure of attention span. (The data for this study can be found in the file named ch3ex1.dat available from www.psypress.com/applied-multivariate-analysis.) To initially screen the data set, we begin by examining the frequencies and descriptive statistics of all variables.

To accomplish this initial data screening in SPSS, we use the following menu options (in the order given next) to obtain the variable frequencies:

Analyze → Descriptive statistics → Frequencies,

and, correspondingly, to furnish their descriptive statistics:

Analyze → Descriptive statistics → Descriptives.

In order to generate the variable frequencies and descriptive statistics in SAS, the following command file can be used. In SAS, there are often a number of different ways to accomplish the same aim. The commands provided below were selected to maintain similarity with the structure of the output rendered by the above SPSS analysis session. In particular, the order of the options in the SAS PROC MEANS statement is structured to create similar output (with the exception of fw=6, which requests the field width of the displayed statistics be set at 6—alternatively, the command ''maxdec=6'' could be used to specify the maximum number of decimal places to output).

```
DATA CHAPTER3;
INFILE 'ch3ex1.dat';
INPUT id Exam_Score Aptitude_Measure Age_in_Years
  Intelligence_Score Attention_Span;
PROC MEANS n range min max mean std fw=6;
  var Exam_Score Aptitude_Measure Age_in_Years
  Intelligence_Score Attention_Span;
RUN;
PROC FREQ;
TABLES Exam_Score Aptitude_Measure Age_in_Years
  Intelligence_Score Attention_Span;
RUN;
```

The resulting outputs produced by SPSS and SAS are as follows:

SPSS descriptive statistics output

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Exam Score	40	102	50	152	57.60	16.123
Aptitude Measure	40	24	20	44	23.12	3.589
Age in Years	40	9	15	24	18.22	1.441
Intelligence Score	40	8	96	104	99.00	2.418
Attention Span	40	7	16	23	20.02	1.349
Valid N (listwise)	40					

SAS descriptive statistics output

The SAS System

The MEANS Procedure

Variable	N	Range	Min	Max	Mean	Std Dev
Exam_Score	40	102.0	50.00	152.0	57.60	16.12
Aptitude _Measure	40	24.00	20.00	44.00	23.13	3.589
Age_in_Years	40	9.000	15.00	24.00	18.23	1.441
Intelligence _Score	40	8.000	96.00	104.0	99.00	2.418
Attention _Span	40	7.000	16.00	23.00	20.03	1.349

By examining the descriptive statistics in either of the above tables, we readily observe the high range on the dependent variable Exam Score. This apparent anomaly is also detected by looking at the frequency distribution of each measure, in particular of the same variable. The pertinent output sections are as follows:

SPSS frequencies output

Frequencies

Exam Score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	5	12.5	12.5	12.5
	51	3	7.5	7.5	20.0
	52	8	20.0	20.0	40.0
	53	5	12.5	12.5	52.5
	54	3	7.5	7.5	60.0
	55	3	7.5	7.5	67.5
	56	1	2.5	2.5	70.0
	57	3	7.5	7.5	77.5
	62	1	2.5	2.5	80.0
	63	3	7.5	7.5	87.5
	64	1	2.5	2.5	90.0
	65	2	5.0	5.0	95.0
	69	1	2.5	2.5	97.5
	152	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Note how the score 152 ''sticks out'' from the rest of the values observed on the Exam Score variable—there is no one else having a score even close to 152; the latter finding is also not unexpected because as mentioned this variable was recorded in the metric of percentage correct responses. We continue our examination of the remaining measures in the study and return later to the issue of discussing and dealing with found anomalous, or at least apparently so, values.

Aptitude Measure

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	2	5.0	5.0	5.0
	21	6	15.0	15.0	20.0
	22	8	20.0	20.0	40.0
	23	14	35.0	35.0	75.0
	24	8	20.0	20.0	95.0
	25	1	2.5	2.5	97.5
	44	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Here we also note a subject whose aptitude score tends to stand out from the rest: the one with a score of 44.

Age in Years

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	15	1	2.5	2.5	2.5
	16	1	2.5	2.5	5.0
	17	9	22.5	22.5	27.5
	18	15	37.5	37.5	65.0
	19	9	22.5	22.5	87.5
	20	4	10.0	10.0	97.5
	24	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

On the age variable, we observe that a subject seems to be very different from the remaining persons with regard to age, having a low value of 15. Given that this is a study of university freshmen, although not a common phenomenon to encounter someone that young, such an age per se does not seem really unusual for attending college.

Intelligence Score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	96	9	22.5	22.5	22.5
	97	4	10.0	10.0	32.5
	98	5	12.5	12.5	45.0
	99	5	12.5	12.5	57.5
	100	6	15.0	15.0	72.5
	101	5	12.5	12.5	85.0
	102	2	5.0	5.0	90.0
	103	2	5.0	5.0	95.0
	104	2	5.0	5.0	100.0
	Total	40	100.0	100.0	

The range of scores on this measure also seems to be well within what could be considered consistent with expectations in a study involving university freshmen.

Attention Span

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16	1	2.5	2.5	2.5
	18	6	15.0	15.0	17.5
	19	2	5.0	5.0	22.5
	20	16	40.0	40.0	62.5
	21	12	30.0	30.0	92.5
	22	2	5.0	5.0	97.5
	23	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Finally, with regard to the variable attention span, there is no subject that appears to have an excessively high or low score compared to the rest of the available sample.

SAS frequencies output

Because the similarly structured output created by SAS would obviously lead to interpretations akin to those offered above, we dispense with inserting comments in the next presented sections.

The SAS System

The FREQ Procedure

Exam_Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
50	5	12.50	5	12.50
51	3	7.50	8	20.00
52	8	20.00	16	40.00
53	5	12.50	21	52.50
54	3	7.50	24	60.00
55	3	7.50	27	67.50
56	1	2.50	28	70.00
57	3	7.50	31	77.50
62	1	2.50	32	80.00
63	3	7.50	35	87.50
64	1	2.50	36	90.00
65	2	5.00	38	95.00
69	1	2.50	39	97.50
152	1	2.50	40	100.00

Aptitude_Measure	Frequency	Percent	Cumulative Frequency	Cumulative Percent
20	2	5.00	2	5.00
21	6	15.00	8	20.00
22	8	20.00	16	40.00
23	14	35.00	30	75.00
24	8	20.00	38	95.00
25	1	2.50	39	97.50
44	1	2.50	40	100.00

Age_in_Years	Frequency	Percent	Cumulative Frequency	Cumulative Percent
15	1	2.50	1	2.50
16	1	2.50	2	5.00
17	9	22.50	11	27.50
18	15	37.50	26	65.00
19	9	22.50	35	87.50
20	4	10.00	39	97.50
24	1	2.50	40	100.00

Intelligence_Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
96	9	22.50	9	22.50
97	4	10.00	13	32.50
98	5	12.50	18	45.00
99	5	12.50	23	57.50
100	6	15.00	29	72.50
101	5	12.50	34	85.00
102	2	5.00	36	90.00
103	2	5.00	38	95.00
104	2	5.00	40	100.00

Attention_Span	Frequency	Percent	Cumulative Frequency	Cumulative Percent
16	1	2.50	1	2.50
18	6	15.00	7	17.50
19	2	5.00	9	22.50
20	16	40.00	25	62.50
21	12	30.00	37	92.50
22	2	5.00	39	97.50
23	1	2.50	40	100.00

Although examining the descriptive statistics and frequency distributions across all variables is highly informative, in the sense that one learns what the data actually are (especially when looking at their frequency tables), it is worthwhile noting that these statistics and distributions are only available for each variable when considered separately from the others. That is, like the descriptive statistics, frequency distributions provide only univariate information with regard to the relationships among the values that subjects give rise to on a given measure. Hence, when an (apparently) anomalous value is found for a particular variable, neither descriptive statistics nor frequency tables can provide further information about the person(s) with that anomalous score, in particular regarding their scores on some or all of

the remaining measures. As a first step toward obtaining such information, it is helpful to extract the data on all variables for any subject exhibiting a seemingly extreme value on one or more of them. For example, to find out who the person was with the exam score of 152, its extraction from the file is accomplished in SPSS by using the following menu options/sequence (the variable Exam Score is named ''exam_score'' in the data file):

Data → Select cases → If condition ''exam_score=152'' (check ''delete unselected cases'').

To accomplish the printing of apparently aberrant data records, the following command line would be added to the above SAS program:

```
IF Exam_Score=152 THEN LIST;
```

Consequently, each time a score of 152 is detected (in the present example, just once) SAS prints the current input data line in the SAS log file.

When this activity is carried out and one takes a look at that person's scores on all variables, it is readily seen that apart from the screening results mentioned, his/her values on the remaining measures are unremarkable (i.e., lie within the variable-specific range for meaningful scores; in actual fact, reference to the original data record would reveal that this subject had an exam score of 52 and his value of 152 in the data file simply resulted from a typographical error).

After the data on all variables are examined for each subject with anomalous value on at least one of them, the next question that needs to be addressed refers to the reason(s) for this data abnormality. As we have just seen, the latter may result from an incorrect data entry, in which case the value is simply corrected according to the original data record. Alternatively, the extreme score may have been due to a failure to declare to the software a missing value code, so that a data point is read by the computer program as a legitimate value while it is not. (Oftentimes, this may be the result of a too hasty move on to the data analysis phase, even a preliminary one, by a researcher skipping this declaration step.) Another possibility could be that the person(s) with an out-of-range value may actually not be a member of the population intended to be studied, but happened to be included in the investigation for some unrelated reasons. In this case, his/her entire data record would have to be deleted from the data set and following analyses. Furthermore, and no less importantly, an apparently anomalous value may in fact be a legitimate value for a sample from a population where the distribution of the variable in question is highly skewed. Because of the potential impact such situations can have on data analysis results, these circumstances are addressed in greater detail in a later section of the chapter. We move next to a more formal discussion of extreme scores, which helps additionally in the process of handling abnormal data values.

3.2 Outliers and the Search for Them

As indicated in Section 3.1, the relevance of an examination for extreme observations, or so-called outliers, follows from the fact that these may exert very strong influence upon the results of ensuing analyses. An outlier is a case with (a) such an extreme value on a given variable, or (b) such an abnormal combination of values on several variables, which may render it having a substantial impact on the outcomes of a data analysis and modeling session. In case (a), the observation is called univariate outlier, while in case (b) it is referred to as multivariate outlier. Whenever even a single outlier (whether univariate or multivariate) is present in a data set, results generated with and without that observation(s) may be very different, leading to possibly incompatible substantive conclusions. For this reason, it is critically important to also consider some formal means that can be used to routinely search for outliers in a given data set.

3.2.1 Univariate Outliers

Univariate outliers are usually easier to spot than multivariate outliers. Typically, univariate outliers are to be sought among those observations with the following properties: (a) the magnitude of their z-scores is greater than 3 or smaller than -3; and (b) their z-scores are to some extent ''disconnected'' from the z-scores of the remaining observations. One of the easiest ways to search for univariate outliers is to use descriptive methods and/or graphical methods. The essence of using the descriptive methods is to check for individual observations with the properties (a) and (b) just mentioned. In contrast, graphical methods involve the use of various plots, including boxplots, steam-and-leaf plots, and normal probability (detrended) plots for studied variables. Before we discuss this topic further, let us mention in passing that often with large samples (at least in the hundreds), there may occasionally be a few apparent extreme observations that need not necessarily be outliers. The reason is that large samples have a relatively high chance of including extreme cases in a studied population that are legitimate members of it and thus need not be removed from the ensuing analyses.

To illustrate, consider the earlier study of university freshmen on the relationship between success in an educational program, aptitude, age, intelligence, and attention span (see data file ch3ex1.dat available from www.psypress.com/applied-multivariate-analysis). To search for univariate outliers, we first obtain the z-scores for all variables. This is readily achieved with SPSS using the following menu options/sequence:

Analyze → Descriptive statistics → Descriptives (check ''save standardized values'').

With SAS, the following PROC STANDARD command lines could be used:

```
DATA CHAPTER3;
INFILE 'ch3ex1.dat';
INPUT id Exam_Score Aptitude_Measure Age_in_Years
   Intelligence_Score Attention_Span;
zscore=Exam_Score;
PROC STANDARD mean=0 std=1 out=newscore;
  var zscore;
RUN;
PROC print data=newscore;
  var Exam_Score zscore;
  title 'Standardized Exam Scores';
RUN;
```

In these SAS statements, PROC STANDARD standardizes the specified variable from the data set (for our illustrative purposes, in this example only the variable exam_score was selected), using a mean of 0 and standard deviation of 1, and then creates a new SAS data set (defined here as the outfile ''newscore'') that contains the resulting standardized values. The PROC PRINT statement subsequently prints the original values alongside the standardized values for each individual on the named variables.

As a result of these software activities, SPSS and SAS generate an extended data file containing both the original variables plus a ''copy'' of each one of them, which consists of all subjects' *z*-scores; to save space, we only provide next the output generated by the above SAS statements (in which the variable ''Exam Score'' was selected for standardization).

Standardized Test Scores

Obs	Exam_Score	zscore
1	51	−0.40936
2	53	−0.28531
3	50	−0.47139
4	63	0.33493
5	65	0.45898
6	53	−0.28531
7	52	−0.34734
8	50	−0.47139
9	57	−0.03721
10	54	−0.22329
11	65	0.45898
12	50	−0.47139
13	52	−0.34734
14	63	0.33493
15	52	−0.34734
16	52	−0.34734
17	51	−0.40936
18	52	−0.34734
19	55	−0.16126

20	55	−0.16126
21	53	−0.28531
22	54	−0.22329
23	152	5.85513
24	50	−0.47139
25	63	0.33493
26	57	−0.03721
27	52	−0.34734
28	62	0.27291
29	52	−0.34734
30	55	−0.16126
31	54	−0.22329
32	56	−0.09924
33	52	−0.34734
34	53	−0.28531
35	64	0.39696
36	57	−0.03721
37	50	−0.47139
38	51	−0.40936
39	53	−0.28531
40	69	0.70708

Looking through the column labeled ''zscore'' in the last output table (and in general each of the columns generated for the remaining variables under consideration), we try to spot the z-scores that are larger than 3 or smaller than -3 and at the same time ''stick out'' of the remaining values in that column. (With a larger data set, it is also helpful to request the descriptive statistics for each variable along with their corresponding z-scores, and then look for any extreme values.) In this illustrative example, subject #23 clearly has a very large z-score relative to the rest of the observations on exam score (viz. larger than 5, although as discussed above this was clearly a data entry error). If we similarly examined the z-scores on the other variables (not tabled above), we would observe no apparent univariate outliers with respect to the variables Intelligence and Attention Span; however, we would find out that subject #40 had a large z-score on the Aptitude measure (z-score $= 5.82$), like subject #8 on age (z-score $= 4.01$).

Once possible univariate outliers are located in a data set, the next step is to search for the presence of multivariate outliers. We stress that it may be premature to make a decision for deleting a univariate outlier before examination for multivariate outliers is conducted.

3.2.2 Multivariate Outliers

Searching for multivariate outliers is considerably more difficult to carry out than examination for univariate outliers. As mentioned in the

preceding section, a multivariate outlier is an observation with values on several variables that are not necessarily abnormal when each variable is considered separately, but are unusual in their combination. For example, in a study concerning income of college students, someone who reports an income of $100,000 per year is not an unusual observation per se. Similarly, someone who reports that they are 16 years of age would not be considered an unusual observation. However, a case with these two measures in combination is likely to be highly unusual, that is, a possible multivariate outlier (Tabachnick & Fidell, 2007).

This example shows the necessity of utilizing such formal means when searching for multivariate outliers, which capitalize in an appropriate way on the individual variable values for each subject and at the same time also take into consideration their interrelationships. A very useful statistic in this regard is the Mahalanobis distance (MD) that we have discussed in Chapter 2. As indicated there, in an empirical setting, the MD represents the distance of a subject's data to the centroid (mean) of all cases in an available sample, that is, to the point in the multivariate space, which has as coordinates the means of all observed variables. That the MD is so instrumental in searching for multivariate outliers should actually not be unexpected, considering the earlier mentioned fact that it is the multivariate analog of univariate distance, as reflected in the z-score (see pertinent discussion in Chapter 2). As mentioned earlier, the MD is also frequently referred to as statistical distance since it takes into account the variances and covariances for all pairs of studied variables. In particular, from two variables with different variances, the one with larger variability will contribute less to the MD; further, two highly correlated variables will contribute less to the MD than two nearly uncorrelated ones. The reason is that the inverse of the empirical covariance matrix participates in the MD, and in effect assigns in this way weights of ''importance'' to the contribution of each variable to the MD.

In addition to being closely related to the concept of univariate distance, it can be shown that with multinormal data on a given set of variables and a large sample, the Mahalanobis distance follows approximately a chi-square distribution with degrees of freedom being the number of these variables (with this approximation becoming much better with larger samples) (Johnson & Wichern, 2002). This characteristic of the MD helps considerably in the search for multivariate outliers. Indeed, given this distributional property, one may consider an observation as a possible multivariate outlier if its MD is larger than the critical point (generally specified at a conservative recommended significance level of $\alpha = .001$) of the chi-square distribution with degrees of freedom being the number of variables participating in the MD. We note that the MDs for different observations are not unrelated to one another, as can be seen from their formal definition in Chapter 2. This suggests the need for some caution

when using the MD in searching for multivariate outliers, especially with samples that cannot be considered large.

We already discussed in Chapter 2 a straightforward way of computing the MD for any particular observation from a data set. Using it for examination of multivariate outliers, however, can be a very tedious and time-consuming activity especially with large data sets. Instead, one can use alternative approaches that are readily applied with statistical software. Specifically, in the case of SPSS, one can simply regress a variable of no interest (e.g., subject ID, or case number) upon all variables participating in the MD; requesting thereby the MD for each subject yields as a byproduct this distance for all observations (Tabachnick & Fidell, 2007). We stress that the results of this multiple regression analysis are of no interest and value per se, apart from providing, of course, each individual's MD.

As an example, consider the earlier study of university freshmen on their success in an educational program in relation to their aptitude, age, intelligence, and attention span. (See data file ch3ex1.dat available from www.psypress.com/applied-multivariate-analysis.) To obtain the MD for each subject, we use in SPSS the following menu options/sequence:

Analyze → Regression → Linear → (ID as DV; all others as IVs)
→ Save ''Mahalanobis Distance''

At the end of this analysis, a new variable is added by the software to the original data file, named MAH_1, which contains the MD values for each subject. (We note in passing that a number of SPSS macros have also been proposed in the literature for the same purposes, which are readily available.) (De Carlo, 1997).

In order to accomplish the same goal with SAS, several options exist. One of them is provided by the following PROC IML program:

```
title 'Mahalanobis Distance Values';
DATA CHAPTER3;
INFILE 'ch3ex1.dat';
INPUT id $ y1 y2 y3 y4 y5;
%let id=id; /* THE %let IS A MACRO STATEMENT*/
%let var=y1 y2 y3 y4 y5; /* DEFINES A VARIABLE */
PROC iml;
  start dsquare;
    use _last_;
    read all var {&var} into y [colname=vars rowname=&id];
    n=nrow(y);
    p=ncol(y);
    r1=&id;
    mean=y[ :,];
```

```
    d=y-j(n,1)*mean;
    s=d'* d/(n-1);
    dsq=vecdiag(d* inv(s) * d');
    r=rank(dsq); /* ranks the values of dsq */
    val=dsq; dsq[r, ]=val;
    val=r1; &id [r]=val;
    result=dsq;
    cl={'dsq'};
  create dsquare from result [colname=cl rowname=&id];
  append from result [rowname=&id];
finish;
print dsquare;
run dsquare;
quit;
PROC print data=dsquare;
    var id dsq;
run;
```

The following output results would be obtained by submitting this command file to SAS (since the resulting output from SPSS would lead to the same individual MDs, we only provide next those generated by SAS); the column headings ''ID'' and ''dsq'' below correspond to subject ID number and MD, respectively. (Note that the observations are rank ordered according to their MD rather than their identification number.)

Mahalanobis Distance Values

Obs	ID	dsq
1	6	0.0992
2	34	0.1810
3	33	0.4039
4	3	0.4764
5	36	0.6769
6	25	0.7401
7	16	0.7651
8	38	0.8257
9	22	0.8821
10	32	1.0610
11	27	1.0714
12	21	1.1987
13	7	1.5199
14	14	1.5487
15	1	1.6823
16	2	2.0967
17	30	2.2345
18	28	2.5811
19	18	2.7049

20	13	2.8883
21	10	2.9170
22	31	2.9884
23	5	3.0018
24	29	3.0367
25	9	3.1060
26	19	3.1308
27	35	3.1815
28	12	3.6398
29	26	3.6548
30	15	3.8936
31	4	4.1176
32	17	4.4722
33	39	4.5406
34	24	4.7062
35	20	5.1592
36	37	13.0175
37	11	13.8536
38	8	17.1867
39	40	34.0070
40	23	35.7510

Mahalanobis distance measures can also be obtained in SAS by using the procedure PROC PRINCOMP along with the STD option. (These are based on computing the uncorrected sum of squared principal component scores within each output observation; see pertinent discussion in Chapters 1 and 7.) Accordingly, the following SAS program would generate the same MD values as displayed above (but ordered by subject ID instead):

```
PROC PRINCOMP std out=scores noprint;
  var Exam_Score Aptitude_Measure Age_in_Years
    Intelligence_Score Attention_Span;
RUN;
DATA mahdist;
    set scores;
    md=(uss(of prin1-prin5));
RUN;
PROC PRINT;
  var md;
RUN;
```

Yet another option available in SAS is to use the multiple regression procedure PROC REG and, similarly to the approach utilized with SPSS

above, regress a variable of no interest (e.g., subject ID) upon all variables participating in the MD. The information of relevance to this discussion is obtained using the INFLUENCE statistics option as illustrated in the next program code.

```
PROC REG;
  model id=Exam_Score Aptitude_Measure Age_in_Years
          Intelligence_Score Attention_Span/INFLUENCE;
RUN;
```

This INFLUENCE option approach within PROC REG does not directly provide the values of the MD but a closely related individual statistic called leverage—commonly denoted by h_i and labeled in the SAS output as HAT DIAG H (for further details, see Belsley, Kuh, & Welsch, 1980). However, the leverage statistic can easily be used to determine MD values for each observation in a considered data set. In particular, it has been shown that MD and leverage are related (in the case under consideration) as follows:

$$\text{MD} = (n-1)(h_i - 1/n), \tag{3.1}$$

where n denotes sample size and h_i is the leverage associated with the ith subject (i=1, . . . , n) (Belsley et al., 1980).

Note from Equation 3.1 that MD and leverage are directly proportional to one another—as MD grows (decreases) so does leverage.

The output resulting from submitting these PROC REG command lines to SAS is given below:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: id
Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	−14.9860	−1.5872	0.0681	0.8256	−0.4292
2	−14.9323	−1.5908	0.0788	0.8334	−0.4652
3	−18.3368	−1.9442	0.0372	0.6481	−0.3822
4	−9.9411	−1.0687	0.1306	1.1218	−0.4142
5	−13.7132	−1.4721	0.1020	0.9094	−0.4961
6	−14.5586	−1.5039	0.0275	0.8264	−0.2531
7	−6.2042	−0.6358	0.0640	1.1879	−0.1662

8	−2.2869	−0.3088	0.4657	2.2003	−0.2882
9	−7.0221	−0.7373	0.1046	1.2112	−0.2521
10	−7.2634	−0.7610	0.0998	1.1971	−0.2534
11	3.0439	0.3819	0.3802	1.8796	0.2991
12	−4.5687	−0.4812	0.1183	1.3010	−0.1763
13	4.0729	0.4240	0.0991	1.2851	0.1406
14	−9.3569	−0.9669	0.0647	1.0816	−0.2543
15	−1.8641	−0.1965	0.1248	1.3572	−0.0742
16	−4.7932	−0.4850	0.0446	1.1998	−0.1048
17	−0.5673	−0.0603	0.1397	1.3894	−0.0243
18	0.9985	0.1034	0.0944	1.3182	0.0334
19	−11.8243	−1.2612	0.1053	1.0079	−0.4326
20	2.4913	0.2677	0.1573	1.4011	0.1157
21	6.9400	0.7092	0.0557	1.1569	0.1723
22	4.7030	0.4765	0.0476	1.2053	0.1066
23	0.2974	0.1214	0.9417	20.4599	0.4880
24	−8.1462	−0.8786	0.1457	1.2187	−0.3628
25	1.8029	0.1818	0.0440	1.2437	0.0390
26	10.6511	1.1399	0.1187	1.0766	0.4184
27	12.1511	1.2594	0.0525	0.9525	0.2964
28	7.7030	0.8040	0.0912	1.1715	0.2547
29	0.6869	0.0715	0.1029	1.3321	0.0242
30	6.6844	0.6926	0.0823	1.1953	0.2074
31	2.0881	0.2173	0.1016	1.3201	0.0731
32	9.5648	0.9822	0.0522	1.0617	0.2305
33	12.4692	1.2819	0.0354	0.9264	0.2454
34	14.2581	1.4725	0.0296	0.8415	0.2574
35	4.2887	0.4485	0.1066	1.2909	0.1549
36	15.9407	1.6719	0.0424	0.7669	0.3516
37	4.0544	0.5009	0.3588	1.7826	0.3746
38	19.1304	2.0495	0.0462	0.6111	0.4509
39	6.8041	0.7294	0.1414	1.2657	0.2961
40	−0.4596	−0.1411	0.8970	11.5683	−0.4165

As can be readily seen, using Equation 3.1 with, say, the obtained leverage value of 0.0681 for subject #1 in the original data file, his/her MD is computed as

$$\text{MD} = (40 - 1)(0.0681 - 1/40) = 1.681, \tag{3.2}$$

which corresponds to his or her MD value in the previously presented output.

By inspection of the last displayed output section, it is readily found that subjects #23 and #40 have notably large MD values—above 30—that may fulfill the above-indicated criterion of being possible multivariate outliers. Indeed, since we have analyzed simultaneously $p = 5$ variables, we are dealing with 5 degrees of freedom for this evaluation, and at

a significance level of $\alpha = .001$, the corresponding chi-square cutoff is 20.515 that is exceeded by the MD of these two cases. Alternatively, requesting extraction from the data file of all subjects' records for whom their MD value is larger than 20.515 (see preceding section) would yield only these two subjects with values beyond this cutoff that can be, thus, potentially considered as multivariate outliers.

With respect to examining leverage values, we note in passing that they range from 0 to 1 with $(p+1)/n$ being their average (in this empirical example, 0.15). Rules of thumb concerning high values of leverage have also been suggested in the literature, whereby in general observations with leverage greater than a certain cutoff may be considered multivariate outliers (Fung, 1993; Huber, 1981). These cutoffs are based on the above-indicated MD cutoff at a specified significance level α (denoted MD_α). Specifically, the leverage cutoffs are

$$h_{\text{cutoff}} = (MD_\alpha)/(n-1) + 1/n, \tag{3.3}$$

which yields $20.515/39 + 1/40 = .551$ for the currently considered example. With the use of Equation 3.3, if one were to utilize the output generated by PROC REG, there is no need to convert to MD the then reported leverage values to determine the observations that may be considered multivariate outliers. In this way, it can be readily seen that only subjects #23 and #40 could be suggested as multivariate outliers.

Using diagnostic measures to identify an observation as a possible multivariate outlier depends on a potentially rather complicated correlational structure among a set of studied variables. It is therefore quite possible that some observations may have a masking effect upon others. That is, one or more subjects may appear to be possible multivariate outliers, yet if one were to delete them, other observations might emerge then as such. In other words, the former group of observations, while being in the data file, could mask the latter ones that, thus, could not be sensed at an initial inspection as possible outliers. For this reason, if one eventually decides to delete outliers masked by previously removed ones, ensuing analysis findings must be treated with great caution since there is a potential that the latter may have resulted from capitalization on chance fluctuations in the available sample.

3.2.3 Handling Outliers: A Revisit

Multivariate outliers may be often found among those that are univariate outliers, but there may also be cases that do not have extreme values on separately considered variables (one at a time). Either way, once an

observation is deemed to be a possible outlier, a decision needs to be made with respect to handling it. To this end, first one should try to use all available information, or information that it is possible to obtain, to determine what reason(s) may have led to the observation appearing as an outlier. Coding or typographical errors, instrument malfunction or incorrect instructions during its administration, or being a member of another population that is not of interest are often sufficient grounds to correspondingly correct or consider removing the particular observation(s) from further analyses. Second, when there is no such relatively easily found reason, it is important to assess to what degree the observation(s) in question may be reflecting legitimate variability in the studied population. If the latter is the case, instead of subject removal variable transformations may be worth considering, a topic that is discussed later in this chapter.

There is a growing literature on robust statistics that deals with methods aimed at down-weighting the contribution of potential outliers to the results of statistical analyses (Wilcox, 2003). Unfortunately, at present there are still no widely available and easily applicable multivariate robust statistical methods. For this reason, we only mention here this direction of current methodological developments that is likely to contribute in the future readily used procedures for differential weighting of observations in multivariate analyses. These procedures will also be worth considering in empirical settings with potential outliers.

When one or more possible outliers are identified, it should be borne in mind that any one of these may unduly influence the ensuing statistical analysis results, but need not do so. In particular, an outlier may or may not be an influential observation in this sense. The degree to which it is influential is reflected in what are referred to as influence statistics and related quantities (such as the leverage value discussed earlier) (Pedhazur, 1997). These statistics have been developed within a regression analysis framework and made easily available in most statistical software. In fact, it is possible that keeping one or more outliers in the subsequent analyses will not change their results appreciably, and especially their substantive interpretations. In such a case, the decision regarding whether to keep them in the analysis or not does not have a real impact upon the final conclusions. Alternatively, if the results and their interpretation depend on whether the outliers are retained in the analyses, while a clear-cut decision for removal versus no removal cannot be reached, it is important to provide the results and interpretations in both cases. For the case where the outlier is removed, it is also necessary that one explicitly mentions, that is, specifically reports, the characteristics of the deleted outlier(s), and then restricts the final substantive conclusions to a population that does not contain members with the outliers' values on the studied variables. For example, if one has good reasons to exclude the subject with ID = 8 from

the above study of university freshmen, who was 15 years old, one should also explicitly state in the substantive result interpretations of the following statistical analyses that they do not necessarily generalize to subjects in their mid-teens.

3.3 Checking of Variable Distribution Assumptions

The multivariate statistical methods we consider in this text are based on the assumption of multivariate normality for the dependent variables. Although this assumption is not used for parameter estimation purposes, it is needed when statistical tests and inference are performed. Multivariate normality (MVN) holds when and only when any linear combination of the individual variables involved is univariate normal (Roussas, 1997). Hence, testing for multivariate normality per se is not practically possible, since it involves infinitely many tests. However, there are several implications of MVN that can be empirically tested. These represent necessary conditions, rather than sufficient conditions, for multivariate normality. That is, these are implied by MVN, but none of these conditions by itself or in combination with any other(s) condition(s) entails multivariate normality.

In particular, if a set of p variables is multivariate normally distributed, then each of them is univariate normal ($p > 1$). In addition, any pair or subset of k variables from that set is bivariate or k-dimensional normal, respectively ($2 < k < p$). Further, at any given value for a single variable (or values for a subset of k variables), the remaining variables are jointly multivariate normal, and their variability does not depend on that value (or values, $2 < k < p$); moreover, the relationship of any of these variables, and a subset of the remaining ones that are not fixed, is linear.

To examine univariate normality, two distributional indices can be judged: skewness and kurtosis. These are closely related to the third and fourth moments of the underlying variable distribution, respectively. The skewness characterizes the symmetry of the distribution. A univariate normally distributed variable has a skewness index that is equal to zero. Deviations from this value on the positive or negative side indicate asymmetry. The kurtosis characterizes the shape of the distribution in terms of whether it is peaked or flat relative to a corresponding normal distribution (with the same mean and variance). A univariate normally distributed variable has a kurtosis that is (effectively) equal to zero, whereby positive values are indicative of a leptokurtic distribution and negative values of a platykurtic distribution. Two statistical tests for evaluating univariate normality are also usually considered, the Kolmogorov–Smirnov Test

and the Shapiro–Wilk Test. If the sample size cannot be considered large, the Shapiro–Wilk Test may be preferred, whereas if the sample size is large the Kolmogorov–Smirnov Test is highly trustworthy. In general terms, both tests consider the following null hypothesis H_0: ''The sampled data have been drawn from a normally distributed population.'' Rejection of this hypothesis at some prespecified significance level is suggestive of the data not coming from a population where the variable in question is normally distributed.

To examine multivariate normality, two analogous measures of skewness and kurtosis—called Mardia's skewness and kurtosis—have been developed (Mardia, 1970). In cases where the data are multivariate normal, the skewness coefficient is zero and the kurtosis is equal to $p(p+2)$; for example, in case of bivariate normality, Mardia's skewness is 0 and kurtosis is 8. Consequently, similar to evaluating their univariate counterparts, if the distribution is, say, leptokurtic, Mardia's measure of kurtosis will be comparatively large, whereas if it is platykurtic, the coefficient will be small. Mardia (1970) also showed that these two measures of multivariate normality can be statistically evaluated. Although most statistical analysis programs readily provide output of univariate skewness and kurtosis (see examples and discussion in Section 3.4), multivariate measures are not as yet commonly evaluated by software. For example, in order to obtain Mardia's coefficients with SAS, one could use the macro called %MULTNORM. Similarly, with SPSS, the macro developed by De Carlo (1997) could be utilized. Alternatively, structural equation modeling software may be employed for this purpose (Bentler, 2004; Jöreskog & Sörbom, 1996).

In addition to examining normality by means of the above-mentioned statistical tests, it can also be assessed by using some informal methods. In case of univariate normality, the so-called normal probability plot (often also referred to as Q–Q plot) or the detrended normal probability plot can be considered. The normal probability plot is a graphical representation in which each observation is plotted against a corresponding theoretical normal distribution value such that the points fall along a diagonal straight line in case of normality. Departures from the straight line indicate violations of the normality assumption. The detrended probability plot is similar, with deviations from that diagonal line effectively plotted horizontally. If the data are normally distributed, the observations will be basically evenly distributed above and below the horizontal line in the latter plot (see illustrations considered in Section 3.4).

Another method that can be used to examine multivariate normality is to create a graph that plots the MD for each observation against its ordered chi-square percentile value (see earlier in the chapter). If the data are multivariate normal, the plotted values should be close to a straight line, whereas points that fall far from the line may be multivariate

outliers (Marcoulides & Hershberger, 1997). For example, the following PROC IML program could be used to generate such a plot:

```
TITLE 'Chi-Square Plot';
DATA CHAPTER3;
INFILE 'ch3ex1.dat';
INPUT id $ y1 y2 y3 y4 y5;
%let id=id;
%let var=y1 y2 y3 y4 y5;
PROC iml;
  start dsquare;
    use_last_;
    read all var {&var} into y [colname=vars rowname=&id];
    n=nrow(y);
    p=ncol(y);
    r1=&id;
    mean=y[ :,];
    d=y - j(n,1)*mean;
    s=d'* d / (n - 1);
    dsq=vecdiag(d* inv(s) * d');
    r=rank(dsq);
    val=dsq; dsq[r, ]=val;
    val=r1; &id [r]=val;
    z=((1:n)' - .5)/n;
    chisq=2 * gaminv(z, p/2);
    result=dsq||chisq;
    cl={'dsq' 'chisq'};
  create dsquare from result [colname=cl rowname=&id];
  append from result [rowname=&id];
finish;
print dsquare; /* THIS COMMAND IS ONLY NEEDED IF YOU WISH TO PRINT THE MD */
RUN dsquare;
quit;
PROC print data=dsquare;
   var id dsq chisq;
RUN;
PROC gplot data=dsquare;
plot chisq*dsq;
RUN;
```

This command file is quite similar to that presented earlier in Section 3.2.2, with the only difference being that now, in addition to the MD values, ordered chi-square percentile values are computed. Submitting this PROC IML program to SAS for the last considered data set generates the

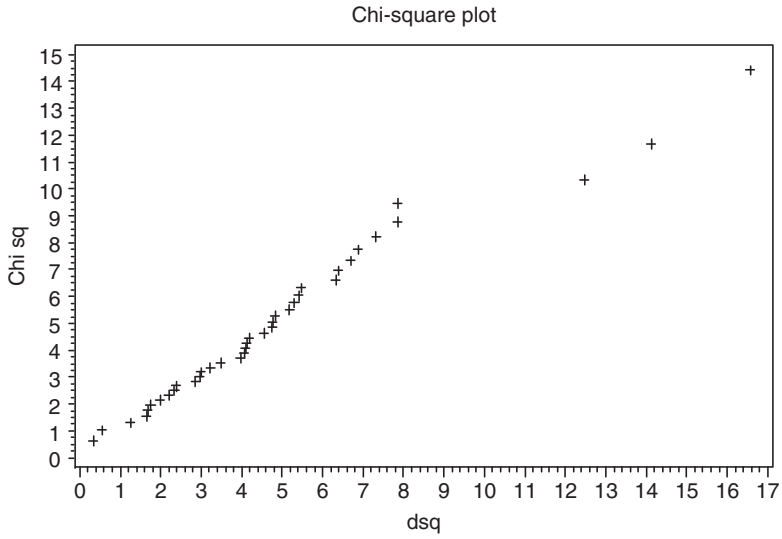


FIGURE 3.1
Chi-square plot for assessing multivariate normality.

above multivariate probability plot (if first removing the data lines for subjects #23 and #40 suggested previously as multivariate outliers).

An examination of Figure 3.1 reveals that the plotted values are reasonably close to a diagonal straight line, indicating that the data do not deviate considerably from normality (keeping in mind, of course, the relatively small sample size used for this illustration).

The discussion in this section suggests that examination of MVN is a difficult yet important topic that has been widely discussed in the literature, and there are a number of excellent and accessible treatments of it (Mardia, 1970; Johnson & Wichern, 2002). In conclusion, we mention that most MVS methods that we deal with in this text can tolerate minor nonnormality (i.e., their results can be viewed also then as trustworthy). However, in empirical applications it is important to consider all the issues discussed in this section, so that a researcher becomes aware of the degree to which the normality assumption may be violated in an analyzed data set.

3.4 Variable Transformations

When data are found to be decidedly nonnormal, in particular on a given variable, it may be possible to transform that variable to be closer to

normally distributed whereupon the set of variables under consideration would likely better comply with the multivariate normality assumption. (There is no guarantee for multinormality as a result of the transformation, however, as indicated in Section 3.3.) In this section, we discuss a class of transformations that can be used to deal with the lack of symmetry of individual variables, an important aspect of deviation from the normal distribution that as well known is symmetric. As it often happens, dealing with this aspect of normality deviation may also improve variable kurtosis and make it closer to that of the normal distribution. Before we begin, however, let us emphasize that asymmetry or skewness as well as excessive kurtosis—and consequently nonnormality in general—may be primarily the result of outliers being present in a given data set. Hence, before considering any particular transformation, it is recommended that one first examines the data for potential outliers. In the remainder of this section, we assume that the latter issue has been already handled.

We start with relatively weak transformations that are usually applicable with mild asymmetry (skewness) and gradually move on to stronger transformations that may be used on distributions with considerably longer and heavier tails. If the observed skewness is not very pronounced and positive, chances are that the square root transformation, $Y' = \sqrt{Y}$, where Y is the original variable, will lead to a transformed measure Y' with a distribution that is considerably closer to the normal (assuming that all Y scores are positive). With SPSS, to obtain the square-rooted variable Y', we use

Transform → Compute,

and then enter in the small left- and right-opened windows correspondingly

SQRT_Y=SQRT(Y),

where Y is the original variable. In the syntax mode of SPSS, this is equivalent to the command

COMPUTE SQRT_Y=SQRT(Y).

(which as mentioned earlier may be abbreviated to COMP SQRT_Y=SQRT(Y).)

With SAS, this can be accomplished by inserting the following general format data-modifying statement immediately after the INPUT statement (but before any PROC statement is invoked):

New-Variable-Name=Formula-Specifying-Manipulation-of-an-Existing-Variable

For example, the following SAS statement could be used in this way for the square root transformation:

SQRT_Y=SQRT(Y),

which is obviously quite similar to the above syntax with SPSS.

If for some subjects $Y < 0$, since a square root cannot be taken then, we first add the absolute value of the smallest of them to all scores, and then proceed with the following SPSS syntax mode command that is to be executed in the same manner as above:

COMP SQRT_Y=SQRT(Y + |MIN(Y)|).

where |MIN(Y)| denotes the absolute value of the smallest negative Y score, which may have been obtained beforehand, for example, with the descriptives procedure (see discussion earlier in the chapter). With SAS, the same operation could be accomplished using the command:

SQRT_Y=SQRT(Y + ABS(min(Y)),

where ABS(min(Y)) is the absolute value of the smallest negative Y score (which can either be obtained directly or furnished beforehand, as mentioned above).

For variables with more pronounced positive skewness, the stronger logarithmic transformation may be more appropriate. The notion of ''stronger'' transformation is used in this section to refer to a transformation with a more pronounced effect upon a variable under consideration. In the presently considered setting, such a transformation would reduce more notably variable skewness; see below. The logarithmic transformation can be carried out with SPSS using the command:

COMP LN_Y=LN(Y).

or with SAS employing the command:

LN_Y=log(Y);

assuming all Y scores are positive since otherwise the logarithm is not defined. If for some cases $Y = 0$ (and for none $Y < 0$ holds), we add 1 first to Y and then take the logarithm, which can be accomplished in SPSS and SAS using respectively the following commands:

COMP LN_Y=LN(Y + 1).

LN_Y=log(Y + 1);

If for some subjects $Y < 0$, we first add to all scores $1 + |\text{MIN}(Y)|$, and then take the logarithm (as indicated above).

A stronger yet transformation is the inverse, which is more effective on distributions with larger skewness, for which the logarithm does not render them close to normality. This transformation is obtained as follows using either of the following SPSS or SAS commands, respectively:

COMP INV_Y=1/Y.

INV_Y=1/Y;

in cases where there are no zero scores. Alternatively, if for some cases $Y = 0$, we add first 1 to Y before taking inverse:

COMPUTE INV_Y=1/(*Y* + 1).

or

INV_Y=1/(Y + 1);

(If there are zero and negative scores in the data, we add first to all scores 1 plus the absolute value of their minimum, and then proceed as in the last two equations.) An even stronger transformation is the inverse squared, which under the assumption of no zero scores in the data can be obtained using the commands:

COMPUTE INVSQ_*Y*=1/Y^2.

or

INV_Y=1/(Y**2);

If there are some cases with negative scores, or zero scores, first add the constant 1 plus the absolute value of their minimum to all subjects' data, and then proceed with this transformation.

When a variable is negatively skewed (i.e., its left tail is longer than its right one), then one needs to first ''reflect'' the distribution before conducting any further transformations. Such a reflection of the distribution can be accomplished by subtracting each original score from 1 plus their maximum, as illustrated in the following SPSS statement:

COMPUTE Y_NEW=MAX(Y) + 1 − Y.

where MAX(Y) is the highest score in the sample, which may have been obtained beforehand (e.g., with the descriptives procedure). With SAS, this operation is accomplished using the command:

SQRT_Y=max(Y) + 1 − Y;

where max(Y) returns the largest value of *Y* (obtained directly, or using instead that value furnished beforehand via examination of variable descriptive statistics). Once reflected in this way, the variable in question is positively skewed and all above discussion concerning transformations is then applicable.

In an empirical study, it is possible that a weaker transformation does not render a distribution close to normality, for example, when the transformed distribution still has a significant and substantial skewness (see below for a pertinent testing procedure). Therefore, one needs to examine the transformed variable for normality before proceeding with it in any analyses that assume normality. In this sense, if one transformation is not strong enough, it is recommendable that a stronger transformation be chosen. However, if one applies a stronger than necessary transformation, the sign of the skewness may end up being changed (e.g., from positive to negative). Hence, one might better start with the weakest transformation

that appears to be worthwhile trying (e.g., square root). Further, and no less important, as indicated above, it is always worthwhile examining whether excessive asymmetry (and kurtosis) may be due to outliers. If the transformed variable exhibits substantial skewness, it is recommendable that one examines it, in addition to the pretransformed variable, also for outliers (see Section 3.3).

Before moving on to an example, let us stress that caution is advised when interpreting the results of statistical analyses that use transformed variables. This is because the units and possibly origin of measurement have been changed by the transformation, and thus those of the transformed variable(s) are no longer identical to the variables underlying the original measure(s). However, all above transformations (and the ones mentioned at the conclusion of this section) are monotone, that is, they preserve the rank ordering of the studied subjects. Hence, when units of measurement are arbitrary or irrelevant, a transformation may not lead to a considerable loss of substantive interpretability of the final analytic results. It is also worth mentioning at this point that the discussed transformed variables result from other than linear transformations, and hence their correlational structure is in general different from that of the original variables. This consequence may be particularly relevant in settings where one considers subsequent analysis of the structure underlying the studied variables (such as factor analysis; see Chapter 8). In those cases, the alteration of the relationships among these variables may contribute to a decision perhaps not to transform the variables but instead to use subsequently specific correction methods that are available within the general framework of latent variable modeling, for which we refer to alternative sources (Muthén, 2002; Muthén & Muthén, 2006; for a nontechnical introduction, see Raykov & Marcoulides, 2006).

To exemplify the preceding discussion in this section, consider data obtained from a study in which $n = 150$ students were administered a test of inductive reasoning ability (denoted IR1 in the data file named ch3ex2.dat available from www.psypress.com/applied-multivariate-analysis). To examine the distribution of their scores on this intelligence measure, with SPSS we use the following menu options/sequence:

Analyze → Descriptive statistics → Explore,

whereas with SAS the following command file could be used:

```
DATA Chapter3EX2;
INFILE 'ch3ex2.dat';
INPUT ir1 group gender sqrt_ir1 ln_ir1;
PROC UNIVARIATE plot normal;
```

```
/* Note that instead of the "plot" statement, additional
 commands like "QQPLOT", "PROBPLOT" or "HISTOGRAM" can be
 provided in a line below to create separate plots */
var ir1;
RUN;
```

The resulting outputs produced by SPSS and SAS are as follows (provided in segments to simplify the discussion).

SPSS descriptive statistics output

Descriptives

			Statistic	Std. Error
IR1	Mean		30.5145	1.20818
	95% Confidence	Lower Bound	28.1272	
	Interval for Mean	Upper Bound	32.9019	
	5% Trimmed Mean		29.9512	
	Median		28.5800	
	Variance		218.954	
	Std. Deviation		14.79710	
	Minimum		1.43	
	Maximum		78.60	
	Range		77.17	
	Interquartile Range		18.5700	
	Skewness		.643	.198
	Kurtosis		.158	.394

Extreme Values

			Case Number	Value
IR1	Highest	1	100	78.60
		2	60	71.45
		3	16	64.31
		4	107	61.45
		5	20	60.02[a]
	Lowest	1	22	1.43
		2	129	7.15
		3	126	7.15
		4	76	7.15
		5	66	7.15[b]

a. Only a partial list of cases with the value 60.02 are shown in the table of upper extremes.
b. Only a partial list of cases with the value 7.15 are shown in the table of lower extremes.

SAS descriptive statistics output

The SAS System
The UNIVARIATE Procedure
Variable: ir1

Moments

N	150	Sum Weights	150
Mean	30.5145333	Sum Observations	4577.18
Std Deviation	14.7971049	Variance	218.954312
Skewness	0.64299511	Kurtosis	0.15756849
Uncorrected SS	172294.704	Corrected SS	32624.1925
Coeff Variation	48.4919913	Std Error Mean	1.20817855

Basic Statistical Measures

Location		Variability	
Mean	30.51453	Std Deviation	14.79710
Median	28.58000	Variance	218.95431
Mode	25.72000	Range	77.17000
		Interquartile Range	18.57000

Extreme Observations

Lowest		Highest	
Value	Obs	Value	Obs
1.43	22	60.02	78
7.15	129	61.45	107
7.15	126	64.31	16
7.15	76	71.45	60
7.15	66	78.60	100

As can be readily seen by examining the skewness and kurtosis in either of the above sections with descriptive statistics, skewness of the variable under consideration is positive and quite large (as well as significant, since the ratio of its estimate to standard error is larger than 2; recall that at α=.05, the cutoff is $\pm$1.96 for this ratio that follows a normal distribution). Such a finding is not the case for its kurtosis, however. With respect to the listed extreme values, at this point, we withhold judgment about any of these 10 cases since their being apparently extreme may actually be due to lack of normality. We turn next to this issue.

SPSS tests of normality

Tests of Normality

	Kolmogoroy-Smirnov[a]			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
IR1	.094	150	.003	.968	150	.002

a. Lilliefors Significance Correction

SAS tests of normality

Tests for Normality

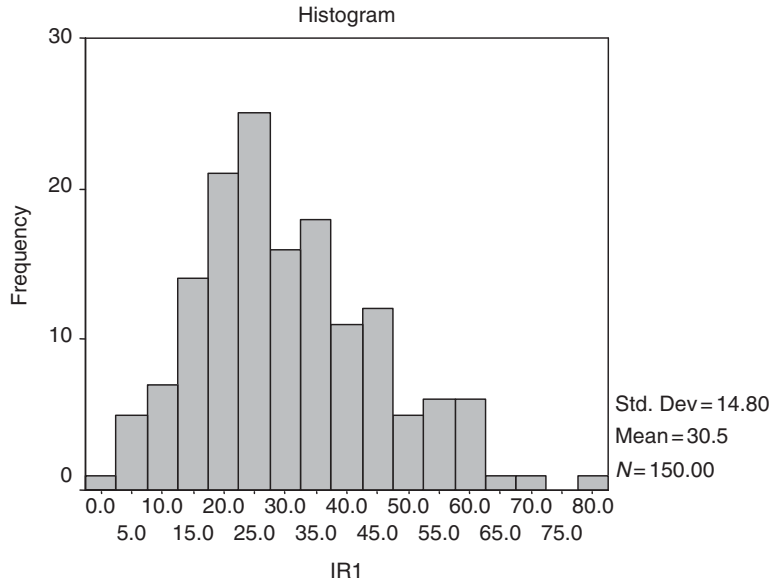
Test	Statistic		p Value	
Shapiro-Wilk	W	0.96824	Pr < W	0.0015
Kolmogorov-Smirnov	D	0.093705	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.224096	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.348968	Pr > A-Sq	<0.0050

As mentioned in Section 3.3, two statistical means can be employed to examine normality, the Kolmogorov–Smirnov (K–S) and Shapiro–Wilk (S–W) tests. (SAS also provides the Cramer–von Mises and the Anderson–Darling tests, which may be viewed as modifications of the K–S Test.) Note that both the K–S and S–W tests indicate that the normality assumption is violated.

The graphical output created by SPSS and SAS would lead to essentially identical plots. To save space, below we only provide the output generated by invoking the SPSS commands given earlier in this section.

Consistent with our earlier findings regarding skewness, the positive tail of the distribution is considerably longer, as seen by examining the following histogram, stem-and-leaf plot, and box plot. This can also be noticed when inspecting the normal probability plots provided next. The degree of skewness is especially evident when examining the detrended plot next, in which the observations are not close to evenly distributed following and below the horizontal line.

So far, we have seen substantial evidence for pronounced skewness of the variable in question to the right. In an attempt to deal with this skewness, which does not appear to be excessive, we try first the square root transformation on this measure, which is the weakest from the ones

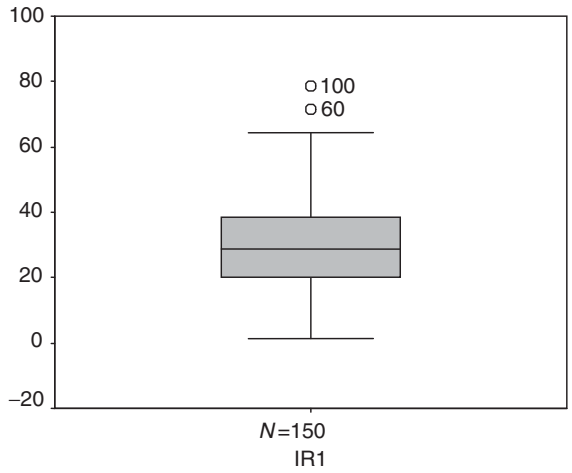


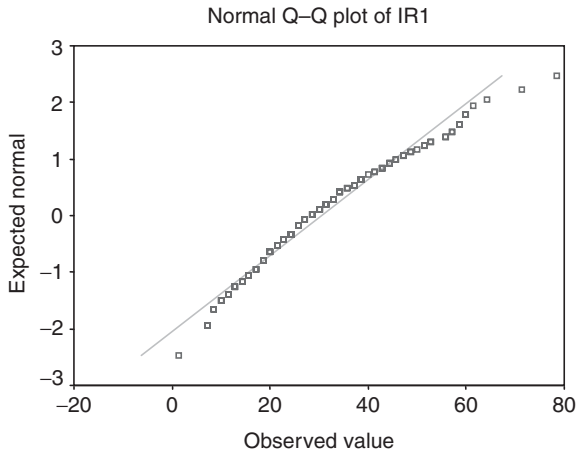
discussed above. To this end, we use with SPSS the following menu options/command (which as illustrated earlier, could also be readily implemented with SAS):

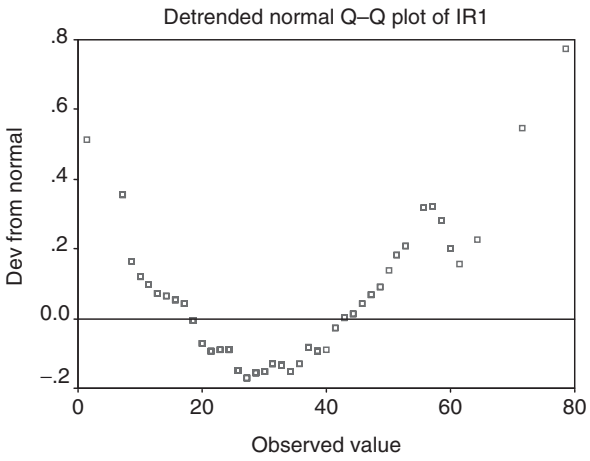
IR1 Stem-and-Leaf Plot

Frequency	Stem & Leaf
1.00	0 . 1
7.00	0 . 7777788
11.00	1 . 00011222244
17.00	1 . 55555777888888888
23.00	2 . 00000011111122224444444
21.00	2 . 555555555557778888888
22.00	3 . 0000111112222222444444
13.00	3 . 5577788888888
10.00	4 . 0112222444
7.00	4 . 5577788
5.00	5 . 01122
7.00	5 . 5577888
4.00	6 . 0014
2.00 Extremes	(>=71)

Each leaf: 1 case(s)
Stem width: 10.00


100
80
60
40
20
0
–20
100
60
N=150
IR1


Normal Q–Q plot of IR1
Expected normal
3
2
1
0
–1
–2
–3
–20
0
20
40
60
80
Observed value


Detrended normal Q–Q plot of IR1
Dev from normal
.8
.6
.4
.2
0.0
–.2
0
20
40
60
80
Observed value

Transform → Compute

(SQRT_IR1=SQRT(IR1))

or COMP SQRT_IR1=SQRT(IR1)

in the syntax mode. Now, to see whether this transformation is sufficient to deal with the problem of positive and marked skewness, we explore the distribution of the so-transformed variable and obtain the following output (presented only using SPSS, since that created by SAS would lead to the same results).

Descriptives

			Statistic	Std. Error
SQRT_IR1	Mean		5.3528	.11178
	95% Confidence Interval for Mean	Lower Bound	5.1319	
		Upper Bound	5.5737	
	5% Trimmed Mean		5.3616	
	Median		5.3460	
	Variance		1.874	
	Std. Deviation		1.36905	
	Minimum		1.20	
	Maximum		8.87	
	Range		7.67	
	Interquartile Range		1.7380	
	Skewness		−.046	.198
	Kurtosis		−.058	.394

As seen by examining this table, the skewness of the transformed variable is no longer significant (like its kurtosis), and the null hypothesis of its distribution being normal is not rejected (see tests of normality in the next table).

Tests of Normality

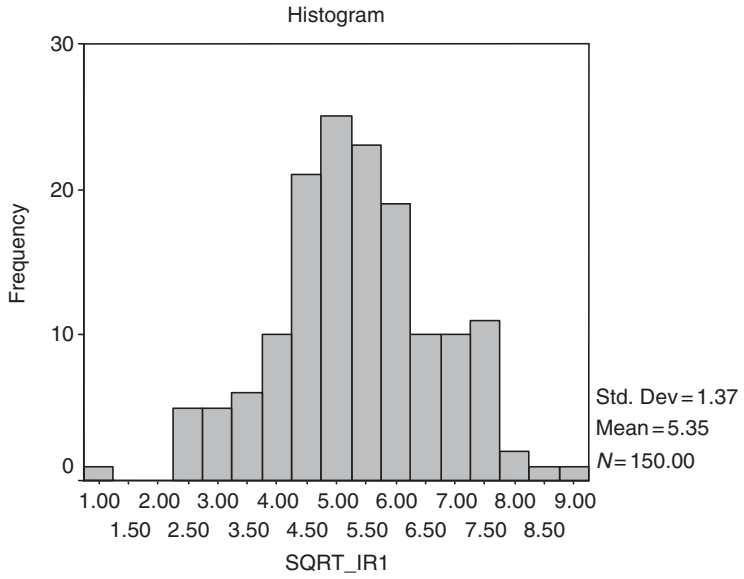
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SQRT_IR1	.048	150	.200*	.994	150	.840

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction.

With this in mind, examining the histogram, stem-and-leaf plot, and box plot presented next, given the relatively limited sample size, it is plausible to consider the distribution of the square-rooted inductive reasoning score as much closer to normal than the initial variable. (We should not overinterpret the seemingly heavier left tail in the last histogram, since its appearance is in part due to the default intervals that the software selects

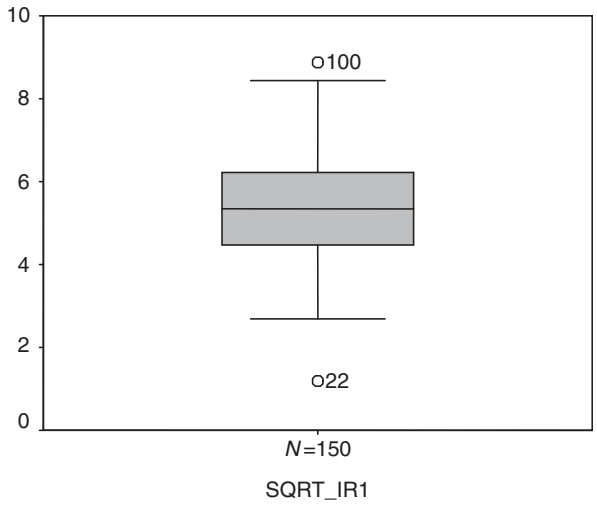
internally.) We stress that with samples that are small, some (apparent) deviations from normality may not result from inherent lack of normality of a studied variable in the population of concern, but may be consequences of the sizable sampling error involved. We therefore do not look for nearly ''perfect'' signs of normality in the graphs to follow, but only for strong and unambiguous deviation patterns (across several of the plots).



SQRT_IR1 Stem-and-Leaf Plot

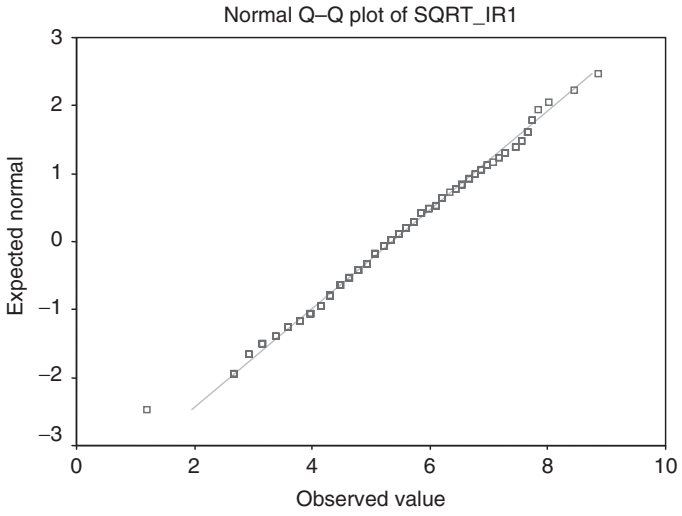
Frequency	Stem & Leaf
1.00	Extremes (=<1.2)
7.00	2. 6666699
5.00	3. 11133
11.00	3. 55557799999
18.00	4. 111333333333444444
17.00	4. 66666777779999999
25.00	5. 000000000002223333334444
20.00	5. 66666777777788888899
14.00	6. 00022222222344
14.00	6. 55556667788899
7.00	7. 0112244
8.00	7. 55666778
2.00	8. 04
1.00	Extremes (>=8.9)

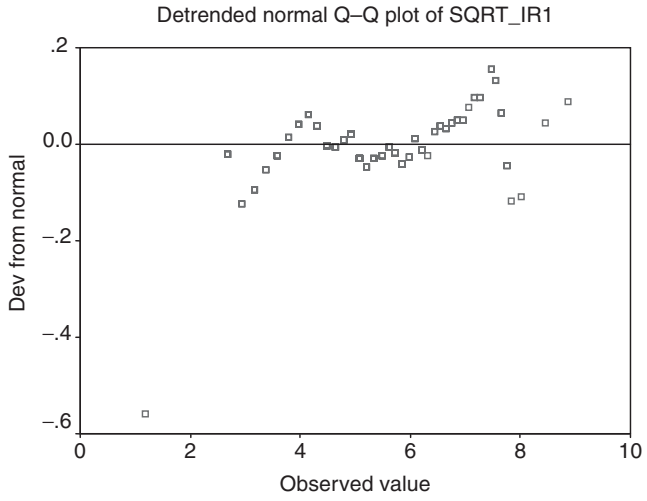
Stem width: 1.00
Each leaf: 1 case(s)



In addition to the last three plots, plausibility of the normality assumption is also suggested from an inspection of the next presented normal probability plots.

As a side note, if we had inadvertently applied the stronger logarithmic transformation instead of the square root, we would have in fact induced negative skewness on the distribution. (As mentioned before, this can happen if too strong a transformation is used.) For illustrative purposes, we present next the relevant part of the data exploration descriptive output that would be obtained then.





Descriptives

			Statistic	Std. Error
LN_IR1	Mean		3.2809	.04700
	95% Confidence Interval for Mean	Lower Bound	3.1880	
		Upper Bound	3.3738	
	5% Trimmed Mean		3.3149	
	Median		3.3527	
	Variance		.331	
	Std. Deviation		.57565	
	Minimum		.36	
	Maximum		4.36	
	Range		4.01	
	Interquartile Range		.6565	
	Skewness		−1.229	.198
	Kurtosis		3.706	.394

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LN_IR1	.091	150	.004	.933	150	.000

a. Lilliefors Significance Correction.

This example demonstrates that considerable caution is advised whenever transformations are used, as one also runs the potential danger of ''overdoing'' it if an unnecessarily strong transformation is chosen. Although in many cases in empirical research, some of the above-mentioned transformations will often render the resulting variable distribution close to normal, this need not always happen. In the latter cases, it may be recommended that one use the so-called likelihood-based method to determine an appropriate power to which the original measure could be raised in order to achieve closer approximation by the normal distribution. This method yields the most favorable transformation with regard to univariate normality, and does not proceed through examination in a step-by-step manner of possible choices as above. Rather, that transformation is selected based on a procedure considering the likelihood function of the observed data. This procedure is developed within the framework of what is referred to as Box–Cox family of variable transformations, and an instructive discussion of it is provided in the original publication by Box and Cox (1964).

In conclusion, we stress that oftentimes in empirical research, a transformation that renders a variable closer to normality may also lead to comparable variances of the resulting variable across groups in a given study. This variance homogeneity result is then an added bonus of the utilized transformation, and is relevant because many univariate as well as multivariate methods are based on the assumption of such homogeneity (and specifically, as we will see in the next chapter, on the more general assumption of homogeneity of the covariance matrix of the dependent variables).