```
data test;
  input X Y @@;
  datalines;
1 2 3 2 5 7
;
run;
```

This method is useful for small data sets, but not for large data sets. SAS data can be entered from an external file using the INFILE statement shown in Display 1.1.

A SAS library is a collection of SAS data sets (and possibly other SAS objects) in a common location such as a directory or folder. SAS data can be read from an existing SAS library using the LIBNAME or SET statement. In the following code, a permanent library is created in a directory on a D drive with the SUMMAR98 data set. This data set will be read and later analyzed.

```
Libname insas "D:\baileraj\grants\odhs-oct98\Datasets";

data mdsmar98; set insas.summar98;
```

In this case, the permanent SAS data set is a file named summar98.sas7bdat, located in the folder specified by the path **`D:\baileraj\grants\odhs-oct98\Datasets`**. (The file extension .sas7bdat is used in Windows and specifies a SAS data set.) This permanent SAS data set is copied to a temporary data set named MDSMAR98 and is located in the WORK library. The temporary data set will exist for the duration of this SAS session.

As shown in Display 1.2, you can generate data in a DATA step for later use. This is a common task in simulation studies. In the following code, a linear regression data set is generated with $Y \sim N(\mu=3 + 2\ X, \sigma=2)$:

```
data lin_reg_data;
   call streaminit(32123);
   do X = 1 to 10 by 1;
      Y = 3 + 2*X + 2*RAND('NORMAL');
      output;
   end;
run;
```

Data can also be imported into SAS from other applications such as Microsoft Excel (using the Import Wizard or PROC IMPORT) or Oracle (using PROC SQL). Selecting **Import Data** from the **File** menu in Windows launches the Import Wizard. You can use the Import Wizard to import Microsoft Excel spreadsheets or other PC file formats if you have licensed SAS/ACCESS Interface to PC File Formats. If you do not have a license to

use this software, then the Import Wizard can still be used to import text files or comma-separated, tab-delimited, or any type of delimited files. For more information about importing data from external sources and manipulating it to build an analysis data set, see Chapter 2.

1.5 Internally Documenting SAS Programs

Comments can be incorporated into SAS programs using several strategies. Commenting throughout a program makes it easier to read.

1. Comments can be statements beginning with an asterisk and ending with a semicolon. For example:

```
proc glm data=health_data;
* fit ancova model with covariate x and
                  classification variable SMOKE;
    class SMOKE;
    model SBP = x | SMOKE;
run;
```

2. A block of comments is often enclosed within /* and */. For example:

```
/* fit ancova model
   covariate: x
   classification variable: SMOKE
*/
proc glm data=health_data;
    class SMOKE;
         model SBP = x | SMOKE;
run;
```

   I use both strategies, although the **`/* <comment> . . . */`** strategy is nice for commenting out blocks of code and defining the header information for a program.

3. You can use TITLE statements to annotate output and implicitly comment a program. For example:

```
proc glm;
    title  "fit ancova model with covariate X";
    title2 "   and classification variable SMOKE";
    class SMOKE;
    model SBP = x | SMOKE;
run;
```

1.6 Summary

This chapter provides an introduction to SAS. But, more important, it suggests coding habits for structuring and documenting a program to make it easier to read, debug, maintain, and modify. In the chapters that follow, the use of SAS to address data analytic problems is discussed in detail. Chapter 2 discusses creating an analysis data set, which typically involves reading an external data source into SAS, and then processing the data set. Chapter 3 presents basic descriptive and graphical statistics, which is followed by creating fancier, more customized tables and output in Chapter 4. The first section of the book concludes with chapters addressing the specifications of statistical models in SAS, and the production of statistical graphics.

1.7 References

Allison, T., and D. Cicchetti "Sleep in Mammals: Ecological and Constitutional Correlates." *Science* 194 (1976): No. 4266, pp. 732–734.

Cody, Ronald, and Ray Pass. 1995. *SAS Programming by Example*. Cary, NC: SAS Institute Inc.

Delwiche, Lora D., and Susan J. Slaughter. 2008. *The Little SAS Book: A Primer*, *Fourth Edition*. Cary, NC: SAS Institute Inc.

Givens, Geof H., and Jennifer A. Hoeting. 2005. *Computational Statistics*. Hoboken, NJ: Wiley Interscience.

Moore, David S., and George P. McCabe. 2003. *Introduction to the Practice of Statistics*. 4th Ed. New York: NY: W. H. Freeman and Co.

1.8 Exercises

1. In the following code, Y=year, X=number of boats registered (in multiples of 1000), and Z=number of manatees killed.[1] Modify this program to make it more readable. Add an introductory comment block to describe this program.

[1] Manatee data set usage based on "Florida powerboats and manatee deaths, 1977–1990" data set from Moore, David S., and George P. McCabe. 2003. *Introduction to the Practice of Statistics*. 4th Ed. New York, NY: W. H. Freeman and Co.

```
OPTIONS LS=75;
DATA EXAMPLE1; INPUT Y X Z; DATALINES;
77    447   13
78    460   21
79    481   24
80    498   16
81    513   24
82    512   20
83    526   15
84    559   34
85    585   33
86    614   33
87    645   39
88    675   43
89    711   50
90    719   47
;
PROC REG; MODEL Z = X / P R CLI CLM;
PLOT Z*X P.*X / OVERLAY;
PLOT R.*X R.*P.; RUN;
```

2. Download the data http://lib.stat.cmu.edu/datasets/sleep.[2] Read the data into a SAS data set. Make sure you reassign the missing code from -999.0 to . (**Hint:** Use an IF statement in a DATA step.) Make sure that you identify animal name as a character variable (using $ after the variable declaration).
3. Read the fitness data from SAS online Help (**Help → Getting Started with SAS Software → Learning to Use SAS → Sample SAS Programs → SAS/STAT → Sample** and select **Example 2 for PROC REG**). The description of this data set (reproduced from this sample page) follows:

   These measurements were made on men involved in a physical fitness course at N.C. State University. The variables are Age (years), Weight (kg), Oxygen intake rate (ml per kg body weight per minute), time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (same time Oxygen rate measured), and maximum heart rate recorded while running. Note: the data for two individuals are found on each line below.

```
44 89.47 44.609 11.37 62 178 182  40 75.07 45.313 10.07 62 185 185
44 85.84 54.297  8.65 45 156 168  42 68.15 59.571  8.17 40 166 172
38 89.02 49.874  9.22 55 178 180  47 77.45 44.811 11.63 58 176 176
40 75.98 45.681 11.95 70 176 180  43 81.19 49.091 10.85 64 162 170
44 81.42 39.442 13.08 63 174 176  38 81.87 60.055  8.63 48 170 186
44 73.03 50.541 10.13 45 168 168  45 87.66 37.388 14.03 56 186 192
```

[2] Allison, T., and D. Cicchetti. "Sleep in Mammals: Ecological and Constitutional Correlates." *Science* 194 (1976): No. 4266, pp. 732–734. Reprinted with permission from AAAS.

```
45 66.45 44.754 11.12 51 176 176  47 79.15 47.273 10.60 47 162 164
54 83.12 51.855 10.33 50 166 170  49 81.42 49.156  8.95 44 180 185
51 69.63 40.836 10.95 57 168 172  51 77.91 46.672 10.00 48 162 168
48 91.63 46.774 10.25 48 162 164  49 73.37 50.388 10.08 67 168 168
57 73.37 39.407 12.63 58 174 176  54 79.38 46.080 11.17 62 156 165
52 76.32 45.441  9.63 48 164 166  50 70.87 54.625  8.92 48 146 155
51 67.25 45.118 11.08 48 172 172  54 91.63 39.203 12.88 44 168 172
51 73.71 45.790 10.47 59 186 188  57 59.08 50.545  9.93 49 148 155
49 76.32 48.673  9.40 56 186 188  48 61.24 47.920 11.50 52 170 176
52 82.78 47.467 10.50 53 170 172
```

a. Create a text file containing these data.

b. Use the INFILE statement to input the data into SAS. For more information about this task, see Chapter 2.

c. Produce summary statistics for each variable (using PROC UNIVARIATE). For more information about this task, see Chapter 3.

d. Generate a plot of Time to run 1.5 miles as a function of Resting Pulse (using PROC SGPLOT). For more information about this task, see Chapter 6.

4. Generate data from an equally replicated one-way ANOVA model with four groups and means (15, 15, 18, 15) with a standard deviation of 1.5. Assume that each group was replicated five times. **Hint:** SAS Help can suggest syntax. Remember that if $Z \sim N(0,1)$, then $\mu+\sigma*Z \sim N(\mu, \sigma^2)$. Test the mean equality in the four groups by fitting an ANOVA model using PROC GLM. Report the results. **Hint:** You need to use a variable that identifies which group is being generated, and you need to generate a random response from the specified distributions. This group variable is part of the CLASS statement that is used with PROC GLM. For more information about linear models, see Chapter 5.

5. Generate pairs of observations (X,Y), where $Y \sim N(3 + X - 0.5*X^2, \sigma=2)$, and X ranges from 0 to 10 by increments of 0.05. Produce a scatter plot of the data with a superimposed regression line and LOESS fit. Do you have evidence that a linear regression model would be inadequate here? **Hint:** The X^2 term can be coded using either X*X or X**2. **Hint:** DO loops can be specified with noninteger increments. (See SAS Help about DO loops.) For more information about producing statistical graphics, see Chapters 6 and 7.