A User's Guide to *SIDE*

Software for Intake Distribution Estimation

Version 1.0

Technical Report 96-TR 30

August 1996

Department of Statistics and
Center for Agricultural and Rural Development
Iowa State University

A User's Guide to *SIDE*

Software for Intake Distribution Estimation

Version 1.0

Technical Report 96-TR 30

August 1996

Department of Statistics and
Center for Agricultural and Rural Development (CARD)
Iowa State University

Funding support for the development of this software has come from Cooperative Agreement No. 58-3198-2-006 between Iowa State University and the Agricultural Research Service, U.S. Department of Agriculture.

CONTENTS

Foreword

This book is the User's Guide to SIDE (Software for Intake Distribution Estimation) Version 1.0.

The present version of the manual has been substantially revised based on feedback we received from several beta testers, and several of the more technical sections from earlier versions of the manual have been eliminated. In their place, we have included more how-to sections to help users along as they venture into the program. We have also added figures showing what screens should look like at each step.

Three example analyses were conducted with the software. They appear in step-by-step form at the end of the manual. These examples illustrate the use of the software and its many options, and they provide a guide for analyzing data.

We gratefully acknowledge input from early users, including Patricia Guenther, Adeline Wilcox and Phil Kott from the Agricultural Research Service (ARS) of the U.S. Department of Agriculture; Professor George Beaton, from the University of Toronto; and Gerry Parise and Mary Jane Novenario from the Center for Agricultural and Rural Development (CARD) at Iowa State University. Paul Sisler from CARD provided invaluable support with technical editing and formatting of this edition of the manual. Judy Shafer and Karleen Gillen contributed their usual superb typing expertise.

Kevin Dodd was primarily responsible for the computer program. Sarah Nusser, Alicia Carriquiry, and Wayne Fuller developed the methodology. Helen Jensen and Todd Krueger contributed to the manual.

Iowa State University
August 1996

Preface

Software for Intake Distribution Estimation (*SIDE*) can help you obtain estimates of usual nutrient intake distributions, their moments, and their percentiles. The software uses a method developed at Iowa State University described in Nusser, et al. (1996).

With *SIDE* you can

❐ input SAS© data sets with multiple intake observations on at least some individuals in the sample,

❐ adjust dietary intake data for discrete or continuous factors of your choice, such as day of the week or age of the individual, and make adjustments to any day,

❐ accommodate different sampling weights associated with each individual,

❐ analyze data collected on consecutive days using an estimate of the correlation among intake days for each different dietary component,

❐ estimate the distribution of usual nutrient intake for any number of nutrients and for subpopulations, such as gender,

❐ compute the proportion of all individuals with intake above or below any number of given thresholds for each nutrient, including those not equally spaced,

❐ estimate the standard error of each estimated percentile of the distribution for simple random samples

❐ output necessary data for producing plots of estimated densities. (Plots must be produced with another software package).

This manual is divided into three main parts:

1. An introduction which provides a technical overview of the methodology and description of the kinds of output you can get
2. A guide for installing *SIDE*, including a description of the equipment you will read. A step-by-step description of the use of the software, including default settings and how to change them

    Using *SIDE*

    Description Data Sets and Keywords

    Output Data Sets
3. An examples section, with analyses of data sets that require the use of default and non default settings

    Additional Examples

    SIDE Warning Codes.

We designed the manual for a user who is familiar with standard statistical techniques and terminology. However, we made every attempt to minimize our use of "statistics jargon" throughout this manual.

Our goal is to provide you with a package that will allow you to apply the method developed at Iowa State University for estimation of usual intake distributions of dietary components. While we have extensively tested this package, you may still find errors. Always check your results against your actual data before drawing any conclusions. Should you find errors, or should you have any comments or suggestions, please feel free to contact us.

Department of Statistics
Iowa State University
Ames, IA 50011

Phone: (515) 294-5242
Fax: (515) 294-2456

.

I: Introduction

Technical overview of methodology

The *SIDE* program implements methodology described in Nusser, et al. (1996) to estimate usual intake distributions using dietary data. Several steps are involved in the analysis of any data set.

Step 1

Step 1. Preliminary adjustments are applied to the observed intake variables. The output of those operations are called smoothed daily intakes. These procedures may perform some or all of the following tasks:

1. Adjust for observed zero intakes by adding a fraction of the mean to each observation.

2. Apply an initial power or log transformation to the shifted observed intakes to produce observations that are as nearly normally distributed as possible. The procedures choose the transformation so that the normal probability plot of a random sample from the empirical distribution of the observations is as close to a straight line as possible. In this context, a *normal probability plot* of a sample of size *n* is defined to be a two-dimensional plot of *X* versus *Y*, where *Y* denotes the ordered observations in the sample (the *order statistics* of the sample) and *X* denotes the expected values of the order statistics from a random

sample of size *n* from the standard normal distribution (the *normal scores* for a sample of size *n*).

3. Ratio-adjust the shifted, power-transformed observed intakes to account for the effects of survey-related and other variables (discrete or continuous variables) present in the data set.

4. Ensure similar distributions across days by matching the mean and variance of each day's shifted, power-transformed data to the mean and variance of the first day.

5. Incorporate sampling weights into the analysis of a particular dietary component by creating an equally-weighted sample of observations that yields essentially the same empirical distribution as the weighted sample of observations.

6. Construct smoothed daily intakes by undoing the initial power transformation and shifting for the adjusted observations.

Step 2

Step 2. A transformation that carries the smoothed daily intakes into normality is developed in two steps:

1. An initial power or log transformation is applied to the smoothed daily intakes to produce observations that are as nearly normally distributed as possible. The transformation is chosen so that the normal probability plot of a random sample from the empirical distribution of the observations is as close to a straight line as possible.

2. A *grafted polynomial* function is fit to the normal probability plot of the smoothed intakes using least-squares. In this fitting, the normal scores are treated as the independent variable, and the order statistics of the smoothed data are treated as the dependent variable. The number of parameters estimated via least squares is determined by a stepwise fitting procedure. The grafted polynomial function is constrained to be monotone increasing, with continuous first and second derivatives, and is defined in a piecewise manner. The range of the smoothed daily intakes is divided into $p + 1$ segments where p is the number of parameters estimated. The grafted polynomial is linear over the first and last segments, and cubic over the other $p - 1$ segments. The inverse of the fitted function defines the transformation to normality, and the values of the inverse transformation applied to the smoothed daily intake are the transformed daily intakes.

Step 3

Step 3. The transformed daily intakes are assumed to follow an additive measurement error model in normal space.

Moment estimates of variance components are computed in the normal space and an estimate of the normal-scale usual intake distribution is obtained. If necessary, corrections for heterogeneity of within-individual variances and/or nonindependent replicate observations are made.

A set of observations representative of the normal-scale usual intake distribution is constructed, and estimates of the corresponding original-scale observations are obtained. The set of original-scale observations constitutes an estimator of the original scale usual intake distribution. Characteristics of the usual intake distribution are estimated by the corresponding characteristics of the set of original scale observations.

What kinds of output can you get?

With *SIDE*, you can control how much output is produced from your analyses. Detailed instructions for choosing what to output appear later in this manual. *SIDE* allows you to print

1. moments of the observed intake distribution estimated either from the input data or from any of the various adjusted data sets,
2. selected percentiles of the observed intake distribution,
3. statistics and diagnostics for the different steps in the *SIDE* method for estimating usual intake distributions,
4. moments of the estimated usual intake distribution,
5. proportion of individuals in the sample with observed intake above or below one or several given thresholds and the corresponding standard errors.

SIDE can also produce SAS data sets containing

1. data created during the initial smoothing process applied to the observed intakes,
2. data used to construct the semiparametric normality transformations,
3. percentiles and estimated function density values for the distribution of observed and usual intakes.

II: Installing *SIDE*

SIDE (Version 1.0) is written in the SAS/IML© language, and it runs under the SAS© system. The estimation of usual intake distributions is performed by a set of 50 SAS/IML© modules residing in executable form in a permanent SAS© storage catalog. The source code for these modules is stored in a file called **side.sas** and consists of about 5800 lines of code.

These modules are called from a SAS/IML© program that you create. Because the modules are stored in executable form, you must process the **side.sas** program only once, no matter how many data sets you plan to analyze.

***Note*:** The source code that accompanies this manual has been tested on SAS© versions 6.07, 6.08, 6.09, 6.10, and 6.11. No downward compatibility to earlier versions of SAS© is implied, nor should it be inferred.

What equipment will you need?

The current version of *SIDE* (Version 1.0) is written in the SAS/IML© language, and it runs under the SAS© system. To run *SIDE*, you will need

- ❐ A workstation running some version of UNIX
  or an IBM PC or compatible machine running Windows

- ❐ SAS© version 6.07 or later for UNIX
  or Windows SAS© version 6.08 or later for PC

SIDE will not run under SAS© version 6.04 or any earlier version. We have not conducted tests on platforms other than those mentioned.

The user must understand basic SAS© system concepts and operation in order to use *SIDE*.

Installing the source files

As a *SIDE* user, you should have received a 3.5" high-density diskette containing the **side.sas** file and a collection of example data files.

The distribution diskette included with the release of version 1.0 of the *SIDE* software contains the files listed in Table 2.1. Copy all the files on the distribution diskette to the directory on your computer that will contain the permanent catalog.

In the examples that follow, we have assumed that the files on the distribution diskette are stored in the directory '/home/SIDE'. The directory name on your computer may be different.

The directory names used in the examples follow standard UNIX conventions. If you are installing *SIDE* on an MS-DOS computer, follow the MS-DOS file-naming conventions.

File Name	Description
packing.lst	Listing of the files contained in the distribution diskette
example1.sas	SAS© program to perform the analysis discussed in **Example 1**
example2.sas	SAS© program to perform the analysis discussed in **Example 2**
example3.sas	SAS© program to perform the analysis discussed in **Example 3**
side.sas	Source code for the SAS/IML© modules
makeds.sas	SAS© program to convert the plain text data files into the SAS© data sets used in the examples
om89_91.txt	Plain text file containing dietary intake data used in **Examples 1 and 2**
yw85.txt	Plain text file containing dietary intake data used in **Example 3**

Table 2.1. Files contained on the distribution diskette

Creating the *SIDE* storage catalog

The *SIDE* storage catalog is created when the SAS© system processes the IML source code contained in the **side.sas** file. The size of the resulting *SIDE* storage catalog varies depending on the SAS© version used, but should rarely exceed 1.5 megabytes.

To create the *SIDE* storage catalog, please refer to Figure 2.1.

Figure 2.1 contains a partial listing of **side.sas.** The first line of the source code tells the SAS© system where to store the permanent catalog of modules.

You will need to modify the first line of the side.sas file according to your particular system configuration with an appropriate text editor. Change the LIBNAME statement on the first line so that it assigns the libref SIDE to the directory containing the files that you copied from the distribution diskette. **You should change only the first line**.

Once you have changed the first line, submit the modified **side.sas** file to the SAS© system for processing. The SAS© system will then create the SAS© catalog SIDE.OBJ and print a listing of the 50 modules included in that library. The SAS© output generated by **side.sas** appears in Figure 2.2. If the SIDE.OBJ catalog is ever damaged or deleted, you will need to reinstall the *SIDE* software.

```
LIBNAME SIDE '/home/SIDE';
OPTIONS NOTES;
PROC IML;
/*------------------------------------------------------------*/
/*                   SIDE/IML Version 1.0                     */
/*                      Copyright 1996                        */
/*       Iowa State University Statistical Laboratory         */
/*                   All rights reserved.                     */
```

Libref Assignment

Figure 2.1. Partial listing of ***side.sas***

```
                          The SAS System                            1

Contents of storage library = SIDE.OBJ

Matrices:

Modules:
ADJUST   ADJWTS   ALT_EST  BEST_FIT CHECK_G_ DENSITY  DMP_PID
ECDF     EQWT     EVALH    EVALHP   FILL_DS  FILL_GRP FILL_ORI
FITPOLY  GETMEANS GET_GRP  GET_NORM GRFTPNML INDIVUI  INIT_DS
INIT_GP  INVERTH  JOINPT   JUSTMEAN MPE      NEW_HOMO NORMTEST
NPARPCT  ONE_NUT  PCTILE   POWRTRAN PROC_BY  PROC_DS  QUIKHP
RADJ     READDATA READ_DS  READ_KW  SETUP    SIDE     SIMPLE
SPEC_REG SQZ      STORDATA TRANDIAG VARCOMP  VARCOMPR WARN
_VTYPE_
```

Figure 2.2. Output after submitting ***side.sas***

III: Using *SIDE*

What about your data?

Application of the method developed at ISU for estimating usual intake distributions requires that *multiple* intake days of data be available on at least some individuals in the sample because the method estimates the within-individual variation of intakes. This estimation is not possible when only one observation is available on each individual in the sample.

Required data characteristics

You must be sure that your data set has the following characteristics:

❐ All recorded intakes must be nonnegative. If you have transformed your intake data in any way, please verify that transformed intakes are also nonnegative.

❐ The observed distribution of intakes must be smooth. No one intake value should appear with frequency significantly higher than the rest of the values. An example of a nonsmooth distribution is the observed distribution of alcohol, or of foods, where it is common to observe many zeros. This version of the software is not designed to handle such distributions.

❐ At least some individuals in the sample must have multiple intake observations. Two observations on a portion of sampled individuals is the minimum requirement for successful application of the *SIDE* method.

❒ The input dataset must contain a total of *mn* observations, where *m* is the number of individuals in the dataset and *n* is the maximum number of observations present for each individual.

❒ If intake data were collected on consecutive days and if you want to account for the day-to-day correlation, then you must supply an estimate of the correlation. Please refer to Carriquiry, et. al. (1995) for a discussion of this problem and correlation estimates for several dietary components. A table of estimated day-to-day correlations estimated from the 1989-91 CSFII for different sex-age groups is presented in the Appendix.

Optional data characteristics

You may also consider data with the following *optional* characteristics:

❒ Unequal sampling weights for each individual in the sample may be used. *SIDE* will produce a weighted analysis of your data.

❒ If your intake data were collected on different days of the week, during different seasons, or from individuals of different ages, ethnic groups, or income level, you can partially remove the effect of those variables from your intake data using *SIDE*'s adjustment procedure. You may use continuous or qualitative variables for data adjustment.

❒ Data may be organized in subpopulations, such as ethnic groups or groups divided by location or income level. *SIDE* will provide an analysis within the subpopulations that you choose.

Organizing a data set for analysis with *SIDE*

SIDE requires data stored in a SAS© data set. In the context of *SIDE,* a SAS© data set contains several distinct kinds of variables: analysis variables, by variables, class variables, continuous variables, ID variables, and weight variables.

Analysis variables

Analysis variables have values that correspond to dietary components. At least one analysis variable must be present in a data set that is to be analyzed with *SIDE.* Further, the data set must be sorted so that the observations on each individual appear consecutively. *SIDE* assumes that the first observation found for an individual corresponds to intake on the first of the survey days, the second observation corresponds to intake on the second of the survey days, etc. *SIDE* matches the first two moments of each day's distribution to the distribution of the data for the first day for each individual. If adjustment to days other than the first is desired, simply sort the data set accordingly. For example, if the second intake day is to be used as the standard, sort the data set so that the first observation for an individual corresponds to intake on the second survey day.

By variables

By variables are categorical variables that define subpopulations of interest in the data set. *SIDE* can perform estimations of usual intake distributions for a dietary component separately for groups defined by a *single* by variable. The by variable should be numeric, with levels indexed by the numbers 1, 2, . . . , *g*, where *g* is the number of levels of the by variable. An example of a by variable is one which takes on the value 1 if the corresponding individual is female and 2 if the corresponding individual is male.

Class variables

Class variables are categorical variables for which the analysis variables are to be adjusted in the preliminary smoothing procedures. All class variables should be numeric, but the levels of the class variables need not be consecutive integers. Multiple class variables can be included in an analysis. However, including many class variables in an analysis can lead to execution errors caused by attempts to invert large matrices. The combined number of levels for all class variables used in an analysis should not exceed 50 minus the number of continuous variables (see below) included in the analysis. Examples of common class variables are indicators for weekday versus weekend, month of the year, and interview sequence.

Continuous variables

Continuous variables are variables with a large number of distinct values for which the analysis variables are to be adjusted in the preliminary smoothing procedures. Continuous variables must be numeric but can take on negative values and have high frequencies of extreme observations. Examples of continuous variables are age and income.

ID variables

ID variables are variables that are not used in any analysis but are included in some of the output data sets produced by *SIDE.* Multiple ID variables can be included in an analysis. Examples of ID variables include identifiers for stratum and cluster.

Weight variables

Weight variables are nonnegative, numeric variables whose values are individual survey sampling weights. A common set of weights (i.e., from one weight variable) can be used for all dietary components in an analysis, or a different weight variable may be specified for each dietary component.

Missing values

The input data set **must** contain a total of *mn* observations, where *m* is the number of individuals in the data set and *n* is the maximum number of observations present for each individual. It is possible that not every individual will have the same number of observations for a particular dietary component. If this is the case, any missing observation must be coded as a missing value. Class variables, continuous variables, by variables, and weight variables may **not** contain missing values. Missing values for ID variables are retained in output data sets created by *SIDE.* Observations with missing values are retained in output data sets, but are not used in usual intake estimation for the corresponding dietary component.

Creating SAS© data sets for example programs

The data analyzed in Examples 1-3 are contained on the distribution diskette in the form of two plain text files named

- **om89_9l.txt**
- **yw85.txt**

The SAS© program **makeds.sas** (shown in Figure 3.1) will convert these plain text files into the SAS© data sets used in the examples.

Using an appropriate text editor, modify the first three lines of the **makeds.sas** file according to your particular system configuration.

1. Change the LIBNAME statement on the first line so that it assigns the libref IN to the directory containing the files copied from the distribution diskette.
2. Change the FILENAME statements on Lines 2, and 3 so that they refer to the files that you copied from the distribution diskette.
3. After making the changes, submit the modified **makeds.sas** file to the SAS© system for processing.

The SAS© system will then create the SAS© data sets IN.OM89_91 and IN.YW85, but will produce no printed output.

***SIDE* will only accept data that has been stored in a SAS$^{©}$-formatted data set.** You may create the necessary SAS$^{©}$ data sets from plain text data sets with a program similar to **makeds.sas**.

```
LIBNAME IN '/home/SIDE';
FILENAME FILE1 '/home/SIDE/om89_91.txt';
FILENAME FILE2 '/home/SIDE/yw85.txt';
DATA IN.OM89_91;
     INFILE FILE1;
     INPUT YEAR ID SEQ MONTH WEEKDAY WEIGHT CALCIUM CALORIES
           IRON PROTEIN;
RUN;

DATA IN.YW85;
     INFILE FILE2;
     INPUT STRATUM CLUSTER INDIV DAY DOW WEIGHT CALC KCAL
           VITA VITC INREP1-INREP16;
     ARRAY INREP (16) INREP1-INREP16;
     ARRAY ACALC (16) CALC1-CALC16;
     ARRAY AKCAL (16) KCAL1-KCAL16;
     ARRAY AIRON (16) IRON1-IRON16;
     ARRAY APROT (16) PROT1-PROT16;
     ARRAY AVITA (16) VITA1-VITA16;
     ARRAY AVITC (16) VITC1-VITC16;
     DO I = 1 TO 16;
        IF INREP(I) = 0
        THEN DO;
             ACALC(I) = .;
             AKCAL(I) = .;
             AIRON(I) = .;
             APROT(I) = .;
             AVITA(I) = .;
             AVITC(I) = .;
             END;
        ELSE DO;
             ACALC(I) = CALC;
             AKCAL(I) = KCAL;
             AIRON(I) = IRON;
             APROT(I) = PROT;
             AVITA(I) = VITA;
             AVITC(I) = VITC;
             END;
     END;
DROP I INREP1-INREP16;
RUN;
```

Libref Assignment

Filename Assignment

Figure 3.1. Listing of ***makeds.sas***

Example 1

This example uses the SAS$^{©}$ file **example1.sas** to illustrate the operation of *SIDE* . Before attempting to work through this example, you must create the example data sets as described in "Creating SAS data sets for example programs".

The example data set IN.OM89_91 is a subset of the 1989-91 CSFII containing observations for male respondents age 60 years old and older. The variables included in the data set are described in Table 3.1.

Variable	Description
YEAR	Year of survey: 1=1989, 2=1990, 3=1991
ID	Individual ID
SEQ	Interview sequence
MONTH	Month of first interview
WEEKDAY	Day of week
WEIGHT	Sampling weight
CALCIUM	Calcium intake (milligrams)
CALORIES	Energy intake (kilocalories)
IRON	Iron intake (milligrams)
PROTEIN	Protein intake (grams)

Table 3.1. Variable listing for ***IN.OM89_91***

These data were collected on three consecutive days. The individuals sampled were selected according to a complex survey design. The analysis performed in this example incorporates sampling weights but ignores the correlation between multiple observations on an individual. Example 2, in the chapter "Additional Examples," shows how to analyze this same data set while accounting for the correlation among intake days. Only one dietary component, caloric intake, is analyzed in this example. Ratio-adjustment for the effects of the survey-related class variables SEQ, MONTH, WEEKDAY, and YEAR is performed.

The program **example1.sas** is shown in Figure 2.2. This program uses *SIDE* to analyze the dietary component CALORIES from the data set IN.OM89_91. Users familiar with basic system concepts of SAS© can see that the program consists of several DATA steps, followed by a PROC IML step. *SIDE* is designed to be run in batch mode, so you have to give a complete description of your particular problem before the PROC IML starts. You do this by creating small SAS© data sets that contain the required information. *SIDE* reads these specially-named SAS© data sets and extracts the information it needs. Refer to "Supplementary Description Data Sets" in the chapter "Description Data Sets and Keywords" for more information on the data sets that *SIDE* recognizes as special.

Libref Assignment

Line 1 in Figure 3.2 associates the directory containing the *SIDE* subroutine library with the libref SIDE. *SIDE* always expects to find its executable modules in the SAS© data catalog SIDE. The SAS© system must have the SIDE libref properly associated before *SIDE* can be loaded and used. If you run **example1.sas** on your own system, then you will need to modify the directory names on lines 1-3. The example code assigns the same directory to the librefs SIDE and IN, because the example data sets are stored in the same directory as the SAS© data catalog SIDE. In actual use, putting the program, input files, and output files in different directories will allow for better data management.

Lines 2 and 3 associate the librefs referred to in lines 6 and 7 of **example1.sas.** Note that line 6 assigns the value 'IN' to the variable **INLIB.** Hence, *SIDE* will look in the SAS© library IN for the data set to be analyzed. Similarly, since line 7 assigns the value 'OUT' to the variable **OUTLIB,** *SIDE* will create output data sets in the SAS© library OUT.

Data Sets

The following is an explanation of each DATA set created in **example1.sas**.

Note: If a particular supplementary description data set is not required in the analysis, it *must* not exist in the same library as the required data sets DESC and ANALYVAR.

DESC (Lines 5-13)

The DESC data set is required. When *SIDE* is called, it looks in the DESC data set to find out where the input data set is stored, what analyses are to be performed, and other necessary information. Lines 6-12 each assign a value to specially-named variables (hereafter referred to as *keyword*). See "Keywords Used in the DESC Data Set" in the chapter "Description Data Sets and Keywords" for more information on the keywords that are recognized by the software. When string or character values are assigned to a keyword (as is the case for the **INLIB, OUTLIB,** and **SAVE_NPP** keywords in the **example1.sas** program)**,** they must be given in upper case. The effect of each keyword in the context of this example is explained below.

The **INLIB, OUTLIB,** and **DATASET** keywords relate to the libname assignments from lines 2-3, and tell *SIDE* to look for the 'OM89_91' data set in the IN library, and to store any output data sets created in the OUT library.

There are three observations per individual in the data set IN.OM89_91, so **NDAY** is assigned the value 3.

The values of the three keywords shown on Lines 10-12 determine which SAS© data sets containing intermediate and final results of the analysis will be created in the OUT library. Setting **SAVE_NPP** to 'Y' requests the creation of output data sets used to generate normal probability plots for the normality transformations estimated in the analysis. Setting **SAVE_SMO** to 'Y' requests the creation of the output data set SMOOTH to contain dietary component data at each stage of the preliminary smoothing procedure. Setting **SAVE_PAR** to 'Y' requests the creation of data sets containing information on the grafted polynomial model fitting performed in the analysis.

ANALYVAR
(Lines 14-18)

The ANALYVAR data set is required. The INPUT statement on Line 15 reads in a single variable, NAME, that is formatted as an 8-character string. The data set has only one observation, with value CALORIES, signifying that the variable CALORIES in the SAS© data set IN.OM89_91 (the data set specified by the **INLIB** and **DATASET** keywords described above) contains the values of the dietary component to be analyzed.

Note: The variable NAME must have only upper case values.

RTNAMES
(Lines 19-23)

In the context of *SIDE,* RTNAMES is an example of a supplementary description data set. Data sets of this type control various aspects of the analysis that *SIDE* performs. In this example, printed output for the variable CALORIES will be identified by the name KCAL and the label Calories. In the output data sets created by *SIDE,* variables corresponding to this dietary component will have KCAL as part of their names.

Note: The variable NAME must have only upper case values, but values for the variable LABEL can have both upper and lower case characters.

```
001 LIBNAME SIDE '/home/SIDE';
002 LIBNAME IN '/home/SIDE';
003 LIBNAME OUT '/home/SIDE/EX1';
004 OPTIONS LS = 69;
005 DATA DESC;
006      INLIB = 'IN';
007      OUTLIB = 'OUT';
008      DATASET = 'OM89_91';
009      NDAY = 3;
010      SAVE_NPP = 'Y';
011      SAVE_SMO = 'Y';
012      SAVE_PAR = 'Y';
013 ;
014 DATA ANALYVAR;
015      INPUT NAME $8.;
016      CARDS;
017 CALORIES
018 ;
019 DATA RTNAMES;
020      INPUT NAME $4. LABEL $32.;
021      CARDS;
022 KCAL Calories
023 ;
024 DATA IDVAR;
025      INPUT NAME $8.;
026      CARDS;
027 ID
028 ;
029 DATA CLASSVAR;
030      INPUT NAME $8.;
031      CARDS;
032 SEQ
033 WEEKDAY
034 MONTH
035 YEAR
```

Figure 3.2. Listing of ***example1.sas***

```
036 ;
037 DATA WTVAR;
038      INPUT NAME $8.;
039      CARDS;
040 WEIGHT
041 ;
042 DATA CUTOFFS;
043      INPUT CO1;
044      CARDS;
045 2000
046 ;
047 PROC IML;
048 RESET STORAGE = SIDE.OBJ FW = 7;
049 LOAD MODULE = (SIDE);
050 RUN SIDE('WORK',0);
051 QUIT;
052 RUN;
```

Figure 3.2. Listing of ***example1.sas*** *(cont.)*

IDVAR
(Lines 24-28)

In this example, the variable ID in the data set IN.OM89_91 will be included in the data set OUT.SMOOTH, even though the variable will play no part in the analysis.

CLASSVAR
(Lines 29-36)

In this example, the data for CALORIES will be ratio-adjusted for the effects of the variables SEQ, WEEKDAY, MONTH, and YEAR. The variables SEQ, WEEKDAY, MONTH, and YEAR will also be included in the data set OUT.SMOOTH .

WTVAR (Lines 37-41)

In this example, the values of the variable WEIGHT will be used as sampling weights in the analysis. The variable WEIGHT will also be included in the data set OUT.SMOOTH.

CUTOFFS (Lines 42-46)

In this example, the proportion of individuals with intake below 2000 calories will be printed for the distribution of usual intakes.

Calling *SIDE*

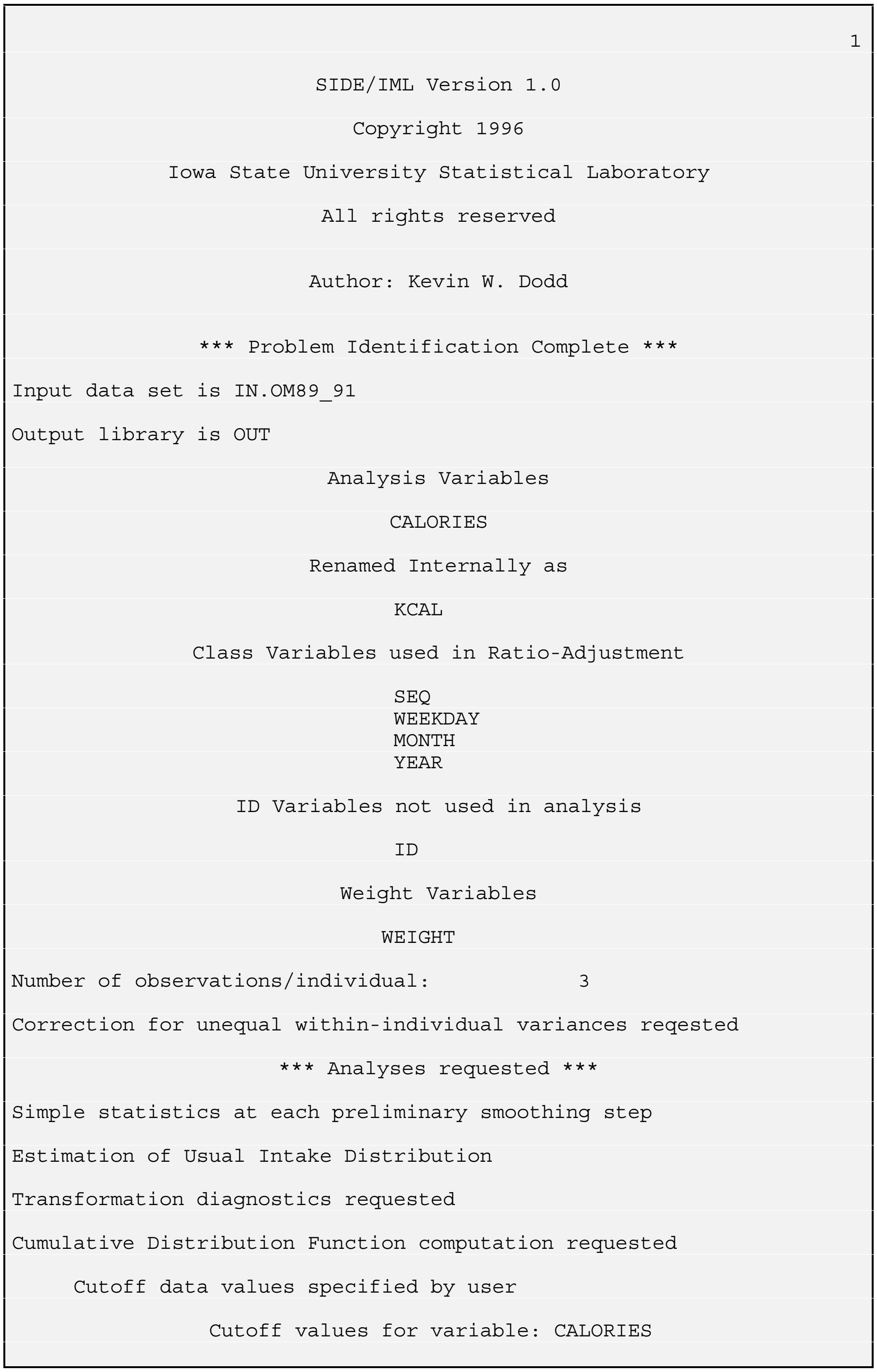
Lines 47-52 start the IML subsystem, and then load and execute the main module of the *SIDE* software. On line 50, the argument 'WORK' that is passed to the module SIDE is the name of the libref for the SAS© library that contains the data sets created in lines 5-46. If the default libref is, for example, USER, then Line 50 should be altered to read "RUN SIDE('USER',0)."

In some cases, the user may wish to increase the amount of memory available to SAS©/IML, by using the WORKSIZE and SYMSIZE options in the PROC IML statement. See SAS©/IML *Software: Usage and Reference* (1990) for a description of these options.

Output

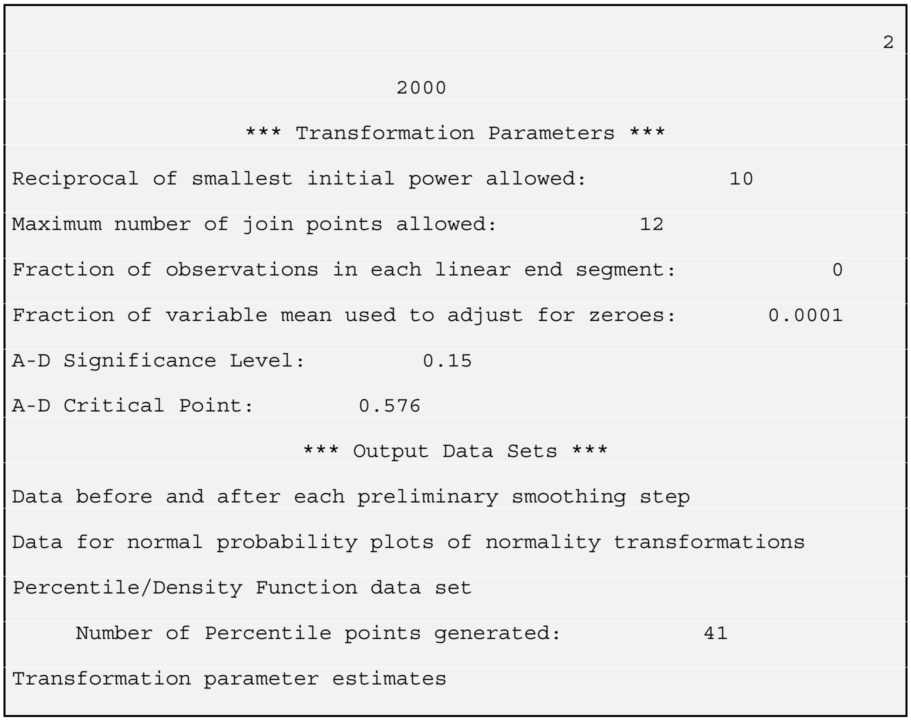
SIDE prints program trace information on the SAS© log. If *SIDE* encounters any errors, a message will appear on the SAS© log prefaced by the phrase "*** SIDE WARNING #," followed by an error code. The error codes reported by *SIDE* are described in the chapter "*SIDE* Warning Codes."

The **example1.sas** program produces the seven pages of output shown in Figures 3.3-3.9. See the chapter "Printed Output" for a description of the printed output.



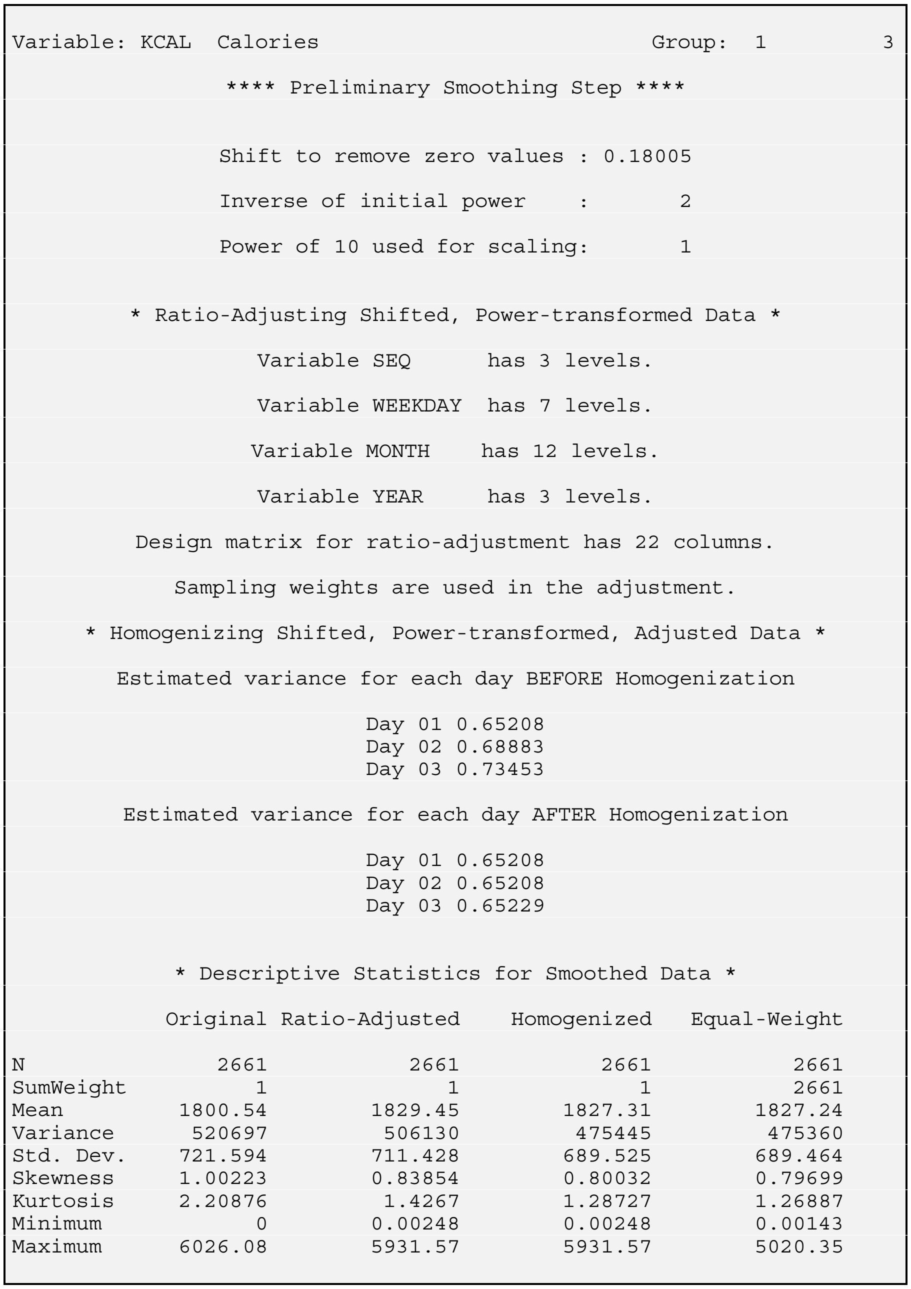
```
                                                                    1
                         SIDE/IML Version 1.0
                            Copyright 1996
             Iowa State University Statistical Laboratory
                         All rights reserved

                        Author: Kevin W. Dodd

               *** Problem Identification Complete ***
Input data set is IN.OM89_91
Output library is OUT
                          Analysis Variables
                               CALORIES
                        Renamed Internally as
                               KCAL
               Class Variables used in Ratio-Adjustment
                               SEQ
                               WEEKDAY
                               MONTH
                               YEAR
                  ID Variables not used in analysis
                               ID
                           Weight Variables
                              WEIGHT
Number of observations/individual:            3
Correction for unequal within-individual variances reqested
                      *** Analyses requested ***
Simple statistics at each preliminary smoothing step
Estimation of Usual Intake Distribution
Transformation diagnostics requested
Cumulative Distribution Function computation requested
     Cutoff data values specified by user
                Cutoff values for variable: CALORIES
```

Figure 3.3. Output from ***example1.sas***



```
                                                                    2
                              2000
                  *** Transformation Parameters ***
Reciprocal of smallest initial power allowed:           10
Maximum number of join points allowed:           12
Fraction of observations in each linear end segment:            0
Fraction of variable mean used to adjust for zeroes:       0.0001
A-D Significance Level:         0.15
A-D Critical Point:        0.576
                       *** Output Data Sets ***
Data before and after each preliminary smoothing step
Data for normal probability plots of normality transformations
Percentile/Density Function data set
     Number of Percentile points generated:           41
Transformation parameter estimates
```

Figure 3.4. Output from ***example1.sas***



```
Variable: KCAL  Calories                          Group:  1         3
                **** Preliminary Smoothing Step ****


                Shift to remove zero values : 0.18005

                Inverse of initial power    :       2

                Power of 10 used for scaling:       1


        * Ratio-Adjusting Shifted, Power-transformed Data *

                   Variable SEQ      has 3 levels.

                   Variable WEEKDAY  has 7 levels.

                   Variable MONTH    has 12 levels.

                   Variable YEAR     has 3 levels.

         Design matrix for ratio-adjustment has 22 columns.

            Sampling weights are used in the adjustment.

     * Homogenizing Shifted, Power-transformed, Adjusted Data *

        Estimated variance for each day BEFORE Homogenization

                           Day 01 0.65208
                           Day 02 0.68883
                           Day 03 0.73453

        Estimated variance for each day AFTER Homogenization

                           Day 01 0.65208
                           Day 02 0.65208
                           Day 03 0.65229


            * Descriptive Statistics for Smoothed Data *

            Original Ratio-Adjusted    Homogenized   Equal-Weight

N               2661          2661           2661           2661
SumWeight          1             1              1           2661
Mean         1800.54       1829.45        1827.31        1827.24
Variance      520697        506130         475445         475360
Std. Dev.    721.594       711.428        689.525        689.464
Skewness     1.00223       0.83854        0.80032        0.79699
Kurtosis     2.20876        1.4267        1.28727        1.26887
Minimum            0       0.00248        0.00248        0.00143
Maximum      6026.08       5931.57        5931.57        5020.35
```

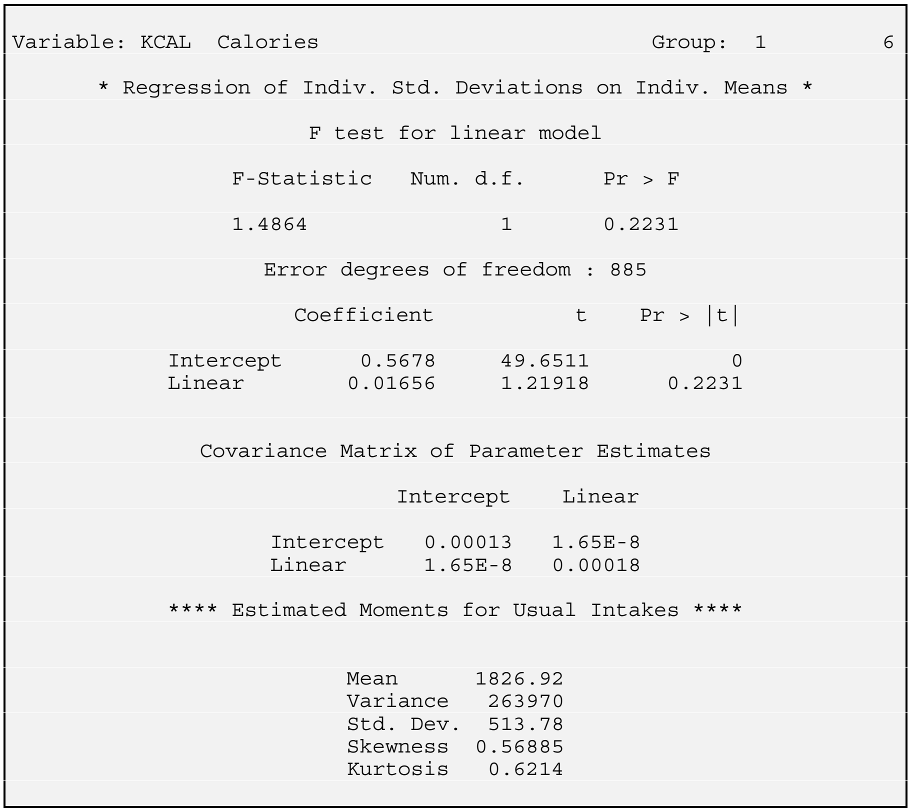
Figure 3.5. Output from ***example1.sas***

```
Variable: KCAL  Calories                          Group:  1         4

          **** Percentile Values for Observed Intakes ****


                       Probability  Percentile

                          0.010     513.317
                          0.025     735.100
                          0.050     891.994
                          0.075     981.675
                          0.100     1047.96
                          0.125     1103.70
                          0.150     1154.29
                          0.175     1202.04
                          0.200     1247.72
                          0.225     1291.81
                          0.250     1334.69
                          0.275     1376.65
                          0.300     1417.90
                          0.325     1458.65
                          0.350     1499.06
                          0.375     1539.28
                          0.400     1579.43
                          0.425     1619.66
                          0.450     1660.06
                          0.475     1700.77
                          0.500     1741.90
                          0.525     1783.58
                          0.550     1825.99
                          0.575     1869.33
                          0.600     1913.82
                          0.625     1959.72
                          0.650     2007.30
                          0.675     2056.91
                          0.700     2108.98
                          0.725     2164.00
                          0.750     2222.64
                          0.775     2285.72
                          0.800     2354.39
                          0.825     2430.21
                          0.850     2515.49
                          0.875     2613.77
                          0.900     2730.92
                          0.925     2877.87
                          0.950     3079.04
                          0.975     3411.96
                          0.990     3839.61

              **** CDF Values for Observed Intakes ****


                    Value Prob. Below Prob. Above

                     2000     0.64623     0.35377
```

Figure 3.6. Output from ***example1.sas***

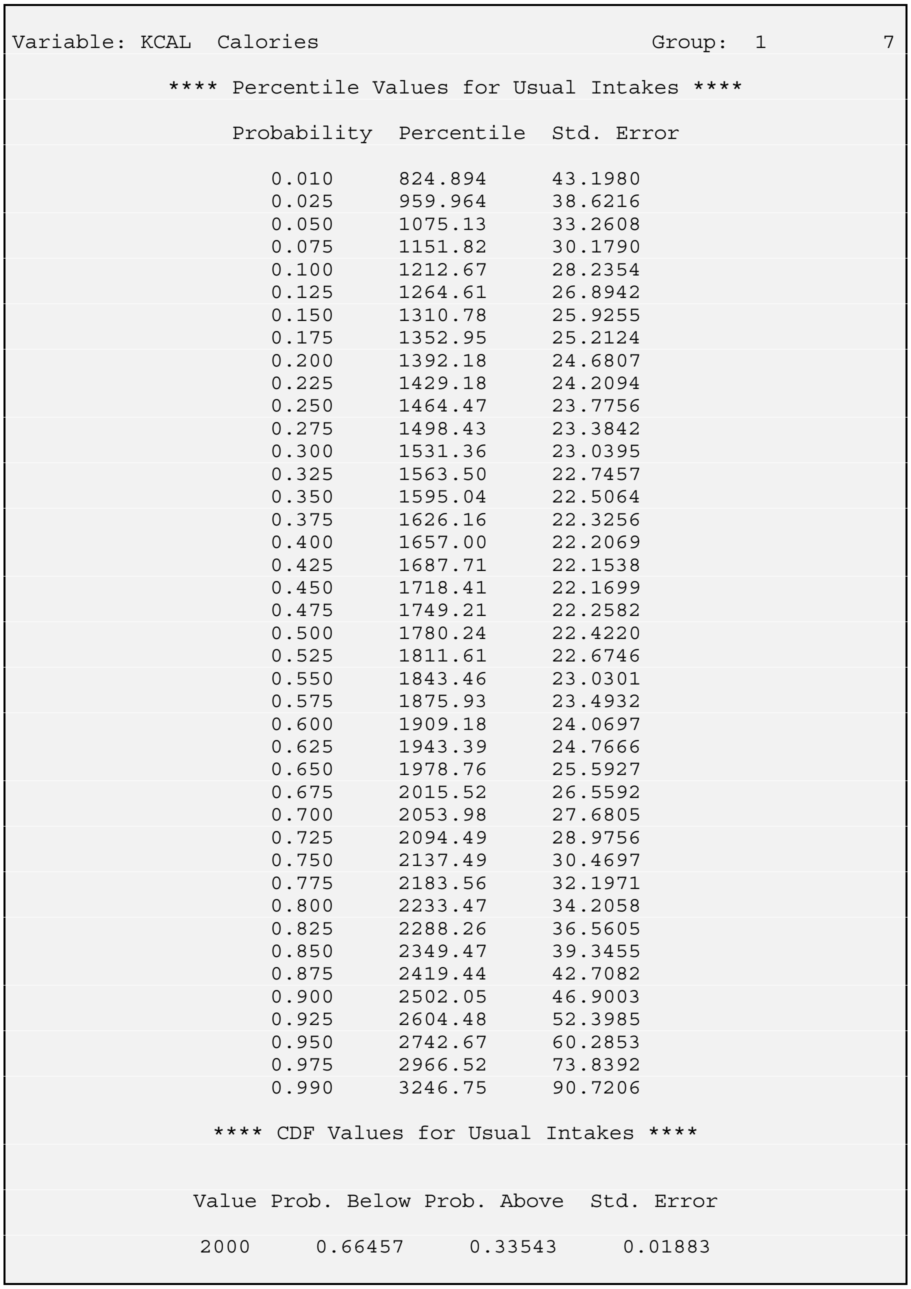
```
Variable: KCAL  Calories                         Group:  1         5

          **** Variance Components for Transformed Data ****


                Number of Individuals             887
                Mean                          -0.0001
                Variance                      1.00102
                Variance of Usual Intakes     0.56295
                Variance of Measurement Error 0.43882
                Measurement Error d.f.           1772
                Scaled 4th Moment of M.E.     3.75232
                Variance of Ind. Variances    0.04829
                t for Heterogeneous Variance  3.34009
                # of Individuals with Ni > 1      887
                Variance of Mean              0.00113
                Variance of U.I. Variance     0.00116


            **** Diagnostics for Estimation Procedure ****


             * Normality Transformation Characteristics *

                                           Initial   Usual

           Inverse of Power used                 2       2
           Power of 10 used for Scaling          1       1
           Number of Parameters/Join Points      7       7

               * Normality Transformation Diagnostics *

                ADS Critical Value (5%)         0.787
                ADS Critical Value (15%)        0.576
                ADS for Observed Intakes      0.41618
                ADS for Day 01 Intakes        0.49491
                ADS for Day 02 Intakes        0.26648
                ADS for Day 03 Intakes        0.45898
                ADS for Mean Intakes          0.16508
```

Figure 3.7. Output from ***example1.sas***



```
Variable: KCAL  Calories                          Group:  1         6

      * Regression of Indiv. Std. Deviations on Indiv. Means *

                       F test for linear model

                 F-Statistic   Num. d.f.      Pr > F

                 1.4864              1       0.2231

                    Error degrees of freedom : 885

                      Coefficient           t    Pr > |t|

            Intercept      0.5678     49.6511           0
            Linear        0.01656     1.21918      0.2231


               Covariance Matrix of Parameter Estimates

                             Intercept    Linear

                    Intercept   0.00013   1.65E-8
                    Linear      1.65E-8   0.00018

            **** Estimated Moments for Usual Intakes ****


                          Mean      1826.92
                          Variance   263970
                          Std. Dev.  513.78
                          Skewness  0.56885
                          Kurtosis   0.6214
```

Figure 3.8. Output from ***example1.sas***



```
Variable: KCAL  Calories                         Group:  1         7

            **** Percentile Values for Usual Intakes ****

                 Probability  Percentile  Std. Error

                    0.010     824.894     43.1980
                    0.025     959.964     38.6216
                    0.050     1075.13     33.2608
                    0.075     1151.82     30.1790
                    0.100     1212.67     28.2354
                    0.125     1264.61     26.8942
                    0.150     1310.78     25.9255
                    0.175     1352.95     25.2124
                    0.200     1392.18     24.6807
                    0.225     1429.18     24.2094
                    0.250     1464.47     23.7756
                    0.275     1498.43     23.3842
                    0.300     1531.36     23.0395
                    0.325     1563.50     22.7457
                    0.350     1595.04     22.5064
                    0.375     1626.16     22.3256
                    0.400     1657.00     22.2069
                    0.425     1687.71     22.1538
                    0.450     1718.41     22.1699
                    0.475     1749.21     22.2582
                    0.500     1780.24     22.4220
                    0.525     1811.61     22.6746
                    0.550     1843.46     23.0301
                    0.575     1875.93     23.4932
                    0.600     1909.18     24.0697
                    0.625     1943.39     24.7666
                    0.650     1978.76     25.5927
                    0.675     2015.52     26.5592
                    0.700     2053.98     27.6805
                    0.725     2094.49     28.9756
                    0.750     2137.49     30.4697
                    0.775     2183.56     32.1971
                    0.800     2233.47     34.2058
                    0.825     2288.26     36.5605
                    0.850     2349.47     39.3455
                    0.875     2419.44     42.7082
                    0.900     2502.05     46.9003
                    0.925     2604.48     52.3985
                    0.950     2742.67     60.2853
                    0.975     2966.52     73.8392
                    0.990     3246.75     90.7206

                **** CDF Values for Usual Intakes ****


              Value Prob. Below Prob. Above  Std. Error

               2000     0.66457     0.33543     0.01883
```

Figure 3.9. Output from ***example1.sas***

IV: Description Data Sets and Keywords

In performing an analysis, *SIDE* must be told such things as where to look for the data set to be analyzed, how many intake days are represented in the data set, what variables correspond to nutrients, and so on. The user supplies this information in the form of keywords, numbers, and variable names placed in specially-named SAS© data sets.

Table 4.1 lists the names of SAS© data sets that *SIDE* recognizes as "special.” These specially-named data sets are hereafter referred to as **description data sets**. This chapter details what can be specified in these description data sets. The DESC data set contains keywords and is presented first. The rest of the description data sets are described in the section “Supplementary Description Data Sets.”

In the **example1.sas** program, the description data sets DESC, ANALYVAR, RTNAMES, IDVAR, CLASSVAR, WTVAR, and CUTOFFS are created in the default library (usually called WORK). Near the end of the program is the command "RUN SIDE('WORK',0);" which tells *SIDE* to execute and to find the description data sets in the 'WORK' library. If the default library is not 'WORK', or if the description data sets are created in a library other than the default, then this line must be modified. Since the description library may or may not be 'WORK', depending on your setup, the library that contains the description data sets is hereafter referred to as

’DESCLIB’. For example, DESCLIB.CLASSVAR will be used to refer to the data set CLASSVAR in the description library. The library names ’INLIB’ and ’OUTLIB’ will be used to identify data sets in the input library and output library, respectively.

Note: When *SIDE* is called, it will attempt to process every specially-named data set in the description library. If a particular description data set is not required for a particular analysis, no data set with the same name may exist in the description library. For example, if you want to run an unweighted analysis of a data set, make sure that the description library does not contain a data set called WTVAR. If you are running SAS© in batch mode, you can avoid problems caused by the presence of extraneous data sets by writing programs in the style of **example1.sas**, where all the required data sets are created in the batch program. However, if you are running SAS© in Display Manager mode with interactive windows, you must be aware that data sets in the WORK library remain until the end of the session or until explicitly deleted with PROC DATASETS.

Data Sets	Comments
DESC	*Required,* variables are keywords (see Table 4.2)
ANALYVAR	*Required,* variable NAME must be formated as $8.
BYVAR	Variable NAME must be formatted as $8.
CLASSVAR	Variable NAME must be formatted as $8.
CONTSVAR	Variable NAME must be formatted as $8.
CUTOFFS	Number of variables must match number of observations in ANALYVAR, names of variables do not matter
FXDCORRS	Number of variables must match number of observations in ANALYVAR, names of variables do not matter, number of observations must match number of by variable levels or be exactly one
FXDJP	Reserved for future development
FXDROOTS	Reserved for future development
IDVAR	Variable NAME must be formatted as $8.
RTNAMES	Variable NAME must be formatted as $4. , variable LABEL must be formatted as $32.
WTVAR	Variable NAME must be formatted as $8. , number of observations must match number of observations in ANALYVAR or be exactly one

Table 4.1. Description data sets

Table 4.2 summarizes the key words that may appear in the DESC data set. The use of each key word is described in “Keywords used in the DESC data set.” Some of the keywords listed in Table 4.2 are reserved for future development, and you should not assign values to them.

Purpose	**Keyword**	**Type**	**Comments**
Input specification	DATASET INLIB OUTLIB	String String String	
Data set characteristics	CORRDAYS NDAY	Y/N Integer	FXDCORRS data set must be present
Analysis parameters	ADALPHA FXHETVAR LINFRAC MAXJP MAXROOT MEANFRAC	Special Y/N Decimal Integer Integer Decimal	
Analysis requests	EST_DAY1 EST_ISU INDIVUI SIMPLE TRANDIAG	Y/N Y/N Y/N Y/N Y/N	SAVE_SMO and EST_ISU must be ‘Y’
Output data set requests	SAVE_PAR SAVE_PCT SAVE_SMO SAVE_NPP NPTS	Y/N Y/N Y/N Y/N Integer	
Reserved for future development	EST_BCP EST_MEAN EST_NRC FITCRIT IMOMENTS INPTCORR MAXITER SAVE_CDF TOLRANCE UIMETHOD VCMETHOD WTDSAMP		WARNING Assigning values to these variables in the DESC data set may cause *SIDE* to crash

Table 4.2. Recognized keywords

Keywords used in the DESC data set

ADALPHA

Related Keywords: none
Default Value: .15

This keyword effects the fitting of the grafted polynomial function to the normal probability plot in the process of transforming the intakes into normality. The value of **ADALPHA** sets the Type I error rate (α in most statistics textbooks) used to test significance with an Anderson-Darling test for normality. Changing the value of **ADALPHA** may cause a different number of parameters to be used in the grafted polynomial fit. Acceptable values are .99, .90, .75, .5, .25, .15, .10, .05, .025, and .01. Note: **ADALPHA** values larger than .15 correspond to approximate critical values for the Anderson-Darling statistic, but the other **ADALPHA** values correspond to the tabled critical values given in Stephens (1974).

CORRDAYS

Related Keywords: none
Default Value: 'N'

Setting **CORRDAYS** to 'Y' indicates that the user has provided estimates of the day-to-day correlation ρ between successive observations in the FXDCORRS data set. These correlations are provided in order to more correctly estimate variance components in cases where daily intake observations for an individual are correlated (when intake observation are taken on consecutive days). *SIDE* can account for day-to-day correlation only if your data set contains either 2 or 3 observations per individual. *SIDE* assumes a special correlation structure for the observations on a particular individual; the correlation between observations separated by *i* days is assumed to be ρ^i for *i* = 1 or 2. Estimated correlations are provided in the Appendix.

DATASET

Related Keywords: INLIB
Default Value: none

You must set the value of **DATASET** to the name of the SAS© data set that contains the data for the dietary components to be analyzed. The data set must be present in the library referenced by the value of the keyword **INLIB**. If you want to analyze the SAS© data set NUTRIENT.RAWDATA, for example, you should set the value of this keyword to 'RAWDATA'. Be sure to use only *upper case* letters when specifying values for this keyword!

EST_DAY1

Related Keywords: EST_ISU, SAVE_PAR, SAVE_PCT, SAVE_NPP
Default value: ’N’

Used in combination with other keywords to create output data sets with information about the first semiparametric transformation into normality (which transforms the distribution of the adjusted daily intakes to a N(0,1) distribution). The information contained in these output data sets is typically used to produce plots of the probability density function, cumulative distribution function, and the semiparametric normality transformation for daily intakes.

EST_ISU

Related Keywords: EST_DAY1, SAVE_PAR, SAVE_PCT, SAVE_NPP, TRANDIAG
Default value: ’Y’

Used in combination with other keywords to create output data sets with information about the second semiparametric transformation into normality (which transforms the distribution of usual intakes in the original scale to a N(0, σ^2_{xx}) distribution, where σ^2_{xx} is the variance of the normal-scale usual intakes). The information contained in these output data sets is typically used to produce plots of the probability density function, cumulative distribution function, and the semiparametric normality transformation for usual intakes.

FXHETVAR

Related Keywords: TRANDIAG
Default value: ’Y’

A value of ’Y’ indicates that *SIDE* should adjust for the effect of heterogeneous within-individual variances when estimating the usual intake distribution. (In other words, the software should **FIX** the **HET**erogeneous **VAR**iances.) In practice, you should first perform the analysis of a data set with **FXHETVAR** set to ’Y’ (the default). If the t-test for heterogeneity of variances is not significant (i.e., the reported test statistic is less than 2), run the analysis a second time, with **FXHETVAR** set to ’N’.

INDIVUI

Related Keywords: SAVE_SMO, EST_ISU
Default value: 'N'

Setting **INDIVUI, SAVE_SMO**, and **EST_ISU** all to 'Y' causes *SIDE* to store estimated individual usual intakes in the output data set OUTLIB.SMOOTH. Each individual represented in the input data set will have one estimated usual intake for each analysis variable.
Note: By their nature, estimates of individual usual intakes are much less precise than estimates of overall characteristics of the usual intake distribution (e.g., moments and percentiles). Keep this fact in mind when interpreting the individual usual intakes *SIDE* provides.

INLIB

Related Keywords: DATASET, OUTLIB
Default value: 'WORK'

This keyword specifies the name of the SAS© library where the input data set is stored. If you want to analyze the SAS© data set NUTRIENT.RAWDATA, for example, you should set the value of this keyword to 'NUTRIENT'. Be sure to use only upper case letters when specifying values for this keyword!

LINFRAC

Related Keywords: MAXJP, MAXROOT, MEANFRAC
Default value: 0 (maximum value is .25)

This keyword controls the size of the first and last segments in grafted polynomial fitting. A value of .01 specifies that the fitted functions should be linear for the smallest 1% of the observations and for the largest 1% of the observations. The remaining 98% of the observations are allocated to cubic segments. If the specified value of **LINFRAC** would place fewer than two distinct observations in each linear segment, exactly two distinct observations are used in each linear segment. The default value of 0 places the fewest observations in the linear segments.

MAXJP

Related Keywords: LINFRAC, MAXROOT, MEANFRAC
Default value: 12

The **MAXJP** keyword specifies the **MAX**imum number of **J**oin **P**oints allowed when fitting a grafted polynomial function to a normal probability plot. A grafted polynomial defined over p + 1 segments has p join points and p coefficients. If the normal probability plot of your data has sharp curves near the middle of the data, or if you have specified a very large value for **ADALPHA**, you may have to use many join points to get an acceptable semiparametric normality transformation.

MAXROOT

Related Keywords: LINFRAC, MAXJP, MEANFRAC
Default value: 10

In the first step of the semiparametric normality transformation, the observed intakes are raised to a power in an attempt to make the data as normal as possible. *SIDE* selects the best power after transforming the data using the following exponents: 0 (the natural log transformation), $1,\frac{1}{1.5},\frac{1}{2},\frac{1}{2.5},\ldots,\frac{1}{\text{MAXROOT}-.5},\frac{1}{\text{MAXROOT}}$. Nutrients with highly skewed, heavy-tailed distributions may require more extreme power transformation than the tenth root or natural log to attain even approximate normality. For most nutrients, however, the default **MAXROOT** value of 10 is more than sufficient.

MEANFRAC

Related Keywords: LINFRAC, MAXJP, MAXROOT
Default value: .0001

This keyword specifies the fraction of the mean of the raw data that is added to each observed value to shift the entire distribution of intakes away from zero. (This is done to avoid taking the logarithm of zero in the preliminary adjustment procedures or the first step in the semiparametric normality transformations.) When the software uses the default value of **MEANFRAC,** .0001, one ten-thousandth of the mean of the observed intakes is added to each observation for that nutrient. You should never need to use any value for **MEANFRAC** larger than the default, unless your data has no observed zeroes, in which case you might want to change the value of **MEANFRAC** to 0.

NDAY

Related Keywords: none
Default value: 1

NDAY specifies the maximum number of observed daily intakes for an individual in the input data set. If some observations are missing due to nonresponse or other reasons, but at least one individual has a complete set of *N* observed daily intakes, you must still set **NDAY** to *N*. Recall that your data set must contain multiple observations on at least some individuals for *SIDE* to estimate usual intake distributions; in practice, you must always specify a value for **NDAY**.

NPTS

Related Keywords: SAVE_PCT
Default value: 41

NPTS specifies the number of additional percentiles n (beyond the default 41 preselected percentiles) to compute and store in the optional output data sets. The additional percentiles are chosen to be equally-spaced in probability from $\frac{1}{n+1}$ to $\frac{n}{n+1}$ in steps of $\frac{1}{n+1}$. Hence, if you want the 1st through 99th percentiles, set NPTS to 99.

OUTLIB

Related Keywords: INLIB, SAVE_PAR, SAVE_PCT, SAVE_SMO, SAVE_NPP
Default value: 'WORK'

This keyword specifies the name of the SAS© library where optional output data sets are created. If you set **OUTLIB** to 'RESULTS', then the first-level name of all output data sets will be RESULTS. Be sure to use only *upper case* letters when specifying values for this keyword.

SAVE_NPP

Related Keywords: OUTLIB, EST_DAY1, EST_ISU
Default value: 'N'

This keyword allows you to create output data sets with the data necessary to construct a normal probability plot of daily and/or usual intakes. If **EST_DAY1** is 'Y', the data set OUTLIB.NPP1 will be created with data for daily intakes. If **EST_ISU** is 'Y', the data set OUTLIB.NPPU will be created with data for usual intakes. With this data, you can see how well the grafted polynomial created in the semiparametric normality transformation fits the normal probability plot by plotting the fitted function on top of the normal probability plot for the transformed data. See the chapter "Output Data Sets" for an explanation of the variables written to these data sets.

SAVE_PAR

Related Keywords: OUTLIB, EST_DAY1, EST_ISU
Default value: 'N'

This keyword allows you to create output data sets with grafted polynomial parameter estimates, join points, and other information about the initial and/or usual intake normality transformations. If **EST_DAY1** is 'Y', the data set OUTLIB.GP1 will be created with information about the semiparmetric normality transformation for the distribution of daily intakes. If **EST_ISU** is 'Y', the data set OUTLIB.GPU will be created with information about the semiparametric normality transformation for the distribution of usual intakes. See the chapter "Output Data Sets" for an explanation of the variables written to these data sets.

SAVE_PCT

Related Keywords: OUTLIB, EST_DAY1, EST_ISU, NPTS
Default value: 'Y'

This keyword allows you to create output data sets with computed percentiles and density function values for the distribution of daily and/or usual intakes. If **EST_DAY1** is 'Y', the data set OUTLIB.PCT1 will be created with information about the distribution of daily intakes. If EST_ISU is 'Y', the data set OUTLIB.PCTU will be created with information about the distribution of usual intakes (including estimated standard errors valid for simple random sampling). See the chapter "Output Data Sets" for an explanation of the variables written to these data sets.

SAVE_SMO

Related Keywords: OUTLIB
Default value: 'N'
This keyword allows you to create an output data set with results of each step of the preliminary smoothing procedure. If **SAVE_SMO** is 'Y', the data set OUTLIB.SMOOTH will be created containing intermediate results taken at each stage of the smoothing operation. If CLASSVAR and/or CONTSVAR data sets were provided, values of the ratio-adjustment variables and the values of the ratio-adjusted daily intakes will be included. If you had more than one day of intake per individual, the homogenized daily intakes will also be included. If you supplied the data set WTVAR, then equal-weight daily observations will be included as well. Although the smoothing procedures are carried out on power-transformed, shifted data, the variables in the SMOOTH data set are untransformed, so that you may more easily compare the original data to the smoothed data at each stage. See the chapter "Output Data Sets" for more information about the variables written to this data set.

SIMPLE

Related Keywords: none
Default value: 'Y'

Setting **SIMPLE** to 'Y' makes the software print out simple statistics for the untransformed daily observations after each smoothing procedure has been applied.

TRANDIAG

Related Keywords: EST_ISU, FXHETVAR
Default value: 'Y'

Setting **TRANDIAG** to 'N' suppresses the printing of the output that appears under the headings * Normality Transformation Diagnostics * and * Regression of Indiv. Std. Deviations on Indiv. Means * as shown in Figures 3.6 and 3.7, respectively.

Supplementary description data sets

The easiest way to run *SIDE* is to create a file in which you build the required data sets in the default library (typically the WORK library) using SAS© DATA step commands. At the end of the file, include the call to *SIDE* in a PROC IML block, passing the name of the default library as an argument to the *SIDE* module call. The program **example1.sas** is an example of this technique. Once you have gotten used to the way the description data sets are used by SIDE, you can experiment with clever ways to set up and run an analysis. Examples of the SAS© code required to create each supplementary data set are included in the descriptions that follow.

DESC

Related Keywords: all keywords
Example code:

```
DATA DESC;
        NDAY=3;
        DATASET='RAWDAT';
        INLIB='NUTRIENT';
        CORRDAYS='Y';
```

Each line below the DATA DESC; line has the form *keyword=value*. You must supply values for **NDAY** and **DATASET**. See the previous section for an explanation of what the various keywords do. The example code tells the software that the data set NUTRIENT.RAWDAT contains 3 observations per individual, and (since the value of **CORRDAYS** is 'Y') that the observations on each individual are taken on consecutive days. Remember that the data set DESCLIB.FXDCORRS must also be created if **CORRDAYS** is set to 'Y'. Values for **INLIB, OUTLIB,** and **DATASET** must be *upper-case* character strings of length 8 or less. Any character constants (such as 'Y' or 'N') must be entered in upper case. If you do not supply this data set, *SIDE* will exit with warning code 1.

ANALYVAR

Related Keywords: none
Example code:

```
DATA ANALYVAR;
        INPUT NAME $8.;
        CARDS;
FOLATE
CALCIUM
;
```

Give the names of the analysis variables in DESCLIB.ANALYVAR. The example code tells the software that the two variables FOLATE and CALCIUM are to be analyzed. The values supplied after the CARDS; statement must be **upper-case** character strings of length 8 or less. If you do not supply this data set, *SIDE* will exit with warning code 1.

BYVAR

Related Keywords: none
Example code:

```
DATA BYVAR;
        INPUT NAME $8.;
        CARDS;
GENDER
;
```

If you want to perform an analysis separately for groups defined by some variables, give the names of the subpopulation variables in DESCLIB.BYVAR. The example code tells the software that the variable GENDER is a subpopulation variable. The values supplied after the CARDS; statement must be *upper-case* character strings of length 8 or less. If possible, use only one subpopulation variable. If you must use more than one subpopulation variable, *SIDE* will convert the multiple variables into a single variable with levels 1, 2,..., *b*, where *b* is determined by the total number of groups defined by all possible combinations of the variables. The software will print an explanation of the new coding on the SAS© log. If you supply the supplementary data set FXDCORRS, you must give the correlation(s) for each group defined by the subpopulation variable, unless you give only one correlation per dietary component, in which case the same correlation will be used for all groups for a particular dietary component.

CLASSVAR

Related Keywords: SAVE_SMO
Example code:

```
DATA CLASSVAR;
        INPUT NAME $8.;
        CARDS;
YEAR
DAYOFWK
;
```

To ratio-adjust the analysis variables for the effects of classification variables, supply the appropriate variable names in DESCLIB.CLASSVAR. The example code tells the software that the two variables YEAR and DAYOFWK are classification variables. The values supplied after the CARDS; statement must be *upper-case* character strings of length 8 or less. If you set **SAVE_SMO** to ’Y’, then the variables named in the CLASSVAR data set will be included in the data set OUTLIB.SMOOTH.

CONTSVAR

Related Keywords: SAVE_SMO
Example code:

```
DATA CONTSVAR;
        INPUT NAME $8.;
        CARDS;
AGE
;
```

To ratio-adjust the analysis variables for the effects of continuous variables, supply the appropriate variable names in DESCLIB.CLASSVAR. The example code tells the software that the variable AGE is a continuous variable. The values supplied after the CARDS; statement must be *upper-case* character strings of length 8 or less. If you set **SAVE_SMO** to ’Y’, then the variables named in the CONTSVAR data set will be included in the data set OUTLIB.SMOOTH.

CUTOFFS

Related Keywords: NPTS
Example code:

```
DATA CUTOFFS;
        INPUT FOLATECO CALCCO;
        CARDS;
160 400
600 1600
;
```

If you want the software to estimate the proportion of the population with intake below certain cutoff values, you must specify them in the data set DESCLIB.CUTOFFS. Think of the data set as a matrix of threshold values, with one column of numbers for each nutrient. The number of variables in DESCLIB.CUTOFFS must match the number of analysis variables in DESCLIB.ANALYVAR. You must make sure that the input statement for the CUTOFFS data set defines the proper number of variables for your particular analysis. The actual names of the variables (FOLATECO and CALCCO are used in the example above) do not matter. The example code supplies two cutoff values for each of two nutrients. Any number of observations can be included in the CUTOFFS data set. The example programs **example1.sas** and **example2.sas** further demonstrate the use of the CUTOFFS data set.

FXDCORRS

Related Keywords: CORRDAYS
Example code:

```
DATA FXDCORRS;
        INPUT RFOLATE RCALC;
        CARDS;
.06 .035
.12 .01
;
```

If your data set consists of two or three daily observations taken consecutively and if you want the software to account for the correlation between observations, you can supply the values of the between-day correlations in this data set.

Note: You must also set the value of **CORRDAYS** to 'Y' in DESCLIB.DESC. As with DESCLIB.CUTOFFS described above, think of the data set as a matrix of correlation values, with one column of numbers for each nutrient. The number of variables in DESCLIB.FXDCORRS must match the number of observations in DESCLIB.ANALYVAR. You must make sure that the input statement for the FXDCORRS data set defines the proper number of variables for your particular analysis, but you do not have to name the variables in any special way. If the BYVAR data set is present, you must supply a correlation coefficient for each combination of nutrient and subpopulation groups, unless you want the same correlation to be used for all subpopulation groups for a given nutrient, in which case you need only supply one correlation coefficient for each nutrient. Estimated correlations for nutrients are provided in the Appendix. The example code supplies two correlation values for each of two nutrients for a case where one subpopulation variable with two possible values was specified.

IDVAR

Related Keywords: SAVE_SMO
Example code:

```
DATA IDVAR;
        INPUT NAME $8.;
        CARDS;
STRATUM
;
```

This data set allows you to specify additional variables from the input data set to include in the OUTLIB.SMOOTH data set (see the **SAVE_SMO** keyword). These are variables that were not used in any other way in the analysis, but need to be retained for purposes of conducting further analysis on the OUTLIB.SMOOTH data set. The values supplied after the CARDS; statement must be *upper-case* character strings of length 8 or less.

RTNAMES

Related Keywords: none
Example code:

```
DATA RTNAMES;
        INPUT NAME $4.  LABEL $32.;
        CARDS;
FOLA Intake of Folic Acid, in mcg
CALC Intake of Calcium, in mg
;
```

The optional output data sets created by *SIDE* contain a variety of information about each nutrient analyzed. For example, the data set OUTLIB.PCTU contains percentiles, the associated standard errors, density function values, and inverse CDF values for each nutrient analyzed. The software names the variables in the output data sets by appending a prefix of up to four letters to the rootnames you supply in this data set. The variable NAME in this data set specifies the rootnames, and is formatted as a 4-character string. The first rootname listed in DESCLIB.RTNAMES corresponds to the first variable listed in DESCLIB.ANALYVAR, the second rootname listed in DESCLIB.RTNAMES corresponds to the second variable listed in DESCLIB.ANALYVAR, and so forth. The example code specifies that variables in the output data sets that correspond to the variable FOLATE will have the word FOLA as part of their names, while variables that correspond to the variable CALCIUM will have the word CALC as part of their names. Values of NAME must be given in upper case. The

RTNAMES data set can also contain a variable called LABEL, formatted as a 32-character string, that specifies descriptive labels to be printed on the output for each nutrient analyzed. The example code specifies that printed output for the analysis of FOLATE will have a header that includes the label 'Intake of Folic Acid, in mcg', and printed output for the analysis of CALCIUM will have a header that includes the label 'Intake of Calcium, in mg'. The values of LABEL may include both upper and lower case letters.

WTVAR

Related Keywords: none
Example code:

```
DATA WTVAR;
        INPUT NAME $8.;
        CARDS;
FOLAWT
CALCWT
;
```

If your data comes from a complex survey, you will typically have sampling weights assigned to each individual represented in your data set. If you want the software to use these weights in an analysis, you need to give the names of the weight variables in DESCLIB.WTVAR. The values supplied after the CARDS; statement must be *upper-case* character strings of length 8 or less. You have two options when specifying the names of the weight variables:

> Supply the name of one weight variable. The values of the named variable will be used as sampling weights for all nutrients analyzed.
>
> Supply a name for a weight variable for each nutrient analyzed. The first variable listed in DESCLIB.ANALYVAR will be weighted by the first variable listed in DESCLIB.WTVAR, the second variable listed in DESCLIB.ANALYVAR will be weighted by the second variable listed in DESCLIB.WTVAR, and so forth. If you choose this option, you must make sure that the number of weight variables matches the number of nutrients to be analyzed.

The example code demonstrates the use of the second option. The software will use the values of the weight variable FOLAWT when analyzing the first nutrient, and will use the values of the weight variable CALCWT when analyzing the second nutrient.

V: Printed Output

S*IDE* produces printed output on each dietary component sequentially. If a subpopulation variable is specified, output for all dietary components for the first group is given first, followed by output for all dietary components for each subsequent group. Figures 3.3 through 3.9 contain examples of each category of printed output available from *SIDE.* The following section and subsection headings correspond to the headings printed on the output.

Problem identification

The first two pages of output from a particular analysis contain the problem identification. The input data set is listed first, followed by the name of the output libref. The names of the variables from the input data set that are to be used in the analysis are printed, and the analysis configuration determined by the keywords from the DESC data set is reported. This output, in conjunction with the program tracking information printed on the SAS© log, is often helpful in diagnosing errors in the program used to invoke *SIDE.*

Preliminary smoothing step

This section of the output (shown in Figure 3.5) always appears. The output consists of

> The amount of shift used to remove zero values. This amount is added to each observation for the dietary component to ensure that the smallest observed intake is strictly positive.
>
> The inverse of the initial power applied to the observed intakes to transform them to approximate normality. This initial transformation is performed so that the smoothing procedures are carried out on "nearly-normal" data. A value of zero indicates that the natural logarithm transformation was performed.
>
> The power of 10 used to scale the data in the power-transformed scale so that the magnitude of the extreme observations is less than 10.

Ratio-adjusting shifted, power-transformed data

This subsection of the output appears only if one or more of the description data sets CLASSVAR and CONTSVAR are present. The output consists of

> The names of the variables used in the ratio adjustment, and the number of levels each variable has.
>
> The total number of parameters estimated in the ratio adjustment.
>
> A message confirming that weights were used in the analysis, if sampling weights were named in the data set WTVAR.

Homogenizing shifted, power-transformed data

This subsection of the output appears only if the number of intake days per individual is greater than one. The variance of each day's data is standardized to that of the first day. When the variances for each intake day are very different, the homogenization procedure does not completely standardize the variances. The output consists of

> Estimated variance of each day's shifted, power-transformed, scaled data before the homogenization procedure is performed.
>
> Estimated variance of each day's shifted, power-transformed, scaled data after the homogenization procedure is performed.

Descriptive statistics for smoothed data

You can suppress this subsection by setting the keyword **SIMPLE** to 'N'. The output consists of

N - The number of observations present for the dietary component/group combination under consideration.

SumWeight - The sum of the sampling weights present for the dietary component/group combination under consideration. The weights are standardized to sum to unity. If the data is equally weighted, the **SumWeight** value is equal to the number of observations.

The rest of the output in this subsection is self-explanatory.

Percentile and CDF values for observed intakes

This section of the output (Figure 3.6) appears if either keyword **EST_DAY1** or **EST_ISU** has the value 'Y'. If the data set CUTOFFS is not present, the CDF value output will not appear.

Variance components for transformed data

This section of the output (shown in Figure 3.7) appears only if the keyword **EST_ISU** has the value 'Y'. The output consists of

Mean, Variance, Variance of Usual Intakes, Variance of Measurement Error -

Estimates of the corresponding quantities for the transformed observed intakes.

Measurement Error d.f. -

Degrees of freedom used in estimating the measurement error variance.

Scaled 4th Moment of M.E. -

An estimate for the fourth moment of the measurement error distribution.

Variance of Ind. Variances -

An estimate of the variance of the within-individual variances.

t for Heterogeneous Variance -

A t-statistic for testing the null hypothesis of homogeneous within-individual variances is given. This statistic is based upon the degrees of freedom labeled as **# of Individuals with Ni > 1.** Values larger than 2 indicate that some evidence of heterogeneous within-individual variances is present.

Variance of Mean, and **Variance of U.I. Variance** - Estimates of the variance of the estimated mean, and of the estimated usual intake variance.

Diagnostics for estimation procedure

This section of the output (Figures 3.7 and 3.8) appears only if both of the keywords **EST_ISU** and **TRANDIAG** have value ’Y’. The three subsections that comprise this output are explained below.

Normality Transformation Characteristics

The column headings *Initial* and *Usual* refer to the initial and usual intake normality transformation respectively. The output consists of

The inverse of the power used in the first step of each normality transformation and the power of 10 used to scale the power-transformed data.

The number of parameters estimated for each grafted polynomial model (fit as the second step of each normality transformation).

Normality Transformation Diagnostics

This subsection contains information concerning the Anderson-Darling tests for normality that are applied to the transformed daily intakes. The output consists of

The 5% and 15% critical values for the Anderson-Darling test (Stephens 1974).

The Anderson-Darling test statistic for the complete set of transformed observations.

The Anderson-Darling test statistic for each day’s transformed observations.

The Anderson-Darling test statistic for the individual means of the transformed observations.

Regression of Indiv. Std. Deviations on Indiv. Means

SIDE computes the sample standard deviation and sample mean of the transformed observations for each individual with more than one observed intake. A regression model containing constant and linear terms is fit to the resulting paired data, with the standard deviations as the dependent variable and the means as the independent variable. The output consists of

> The F-test for the linear model and the corresponding degrees of freedom and significance level.
>
> The value of the error degrees of freedom (equal to *n* - 2, where *n* is the number of individuals with more than one observed intake).
>
> The parameter estimates and tests of significance, arranged in the customary table.
>
> The covariance matrix of the parameter estimates.

Estimated moments for usual intakes

This section of the output (shown in Figure 3.8) only appears if the keyword **EST_ISU** has the value 'Y'. The output contains the estimated mean, variance, standard deviation, skewness, and kurtosis for the usual intakes.

Percentile and CDF values for usual intakes

This section of the output (Figure 3.9) appears only if the keyword **EST_ISU** has the value 'Y'. Estimates and standard errors for each percentile and CDF value are printed. These standard errors are approximations to the standard errors of the appropriate quantities computed under the assumption of simple random sampling of individuals on nonconsecutive days. If sampling weights were used in the analysis, or if the observations on each individual were taken on consecutive days, correct standard errors may be obtained using balanced repeated replication Wolter (1985) as demonstrated in the discussion of the **example3.sas** program included in the chapter "Additional examples."

If the data set CUTOFFS is not present, the CDF value output will not appear.

VI: Output Data Sets

S*IDE* has the capability to create SAS© data sets containing various results generated by different parts of the *SIDE* method. Setting the proper keywords in the description data set will cause these output data sets to be created. The number of variables in many of the output data sets depends on the number of analysis variables used. SAS variable names are created by adding a prefix denoting the type of variable to a four-letter *rootname*. These rootnames are provided by the user by means of the data set RTNAMES. For example, in the **example1.sas** program, the rootname KCAL was assigned to the analysis variable CALORIES. If you do not supply rootnames by means of the RTNAMES data set, *SIDE* uses the default rootnames VOO1, VOO2,...,VNNN. The combination of prefix and rootname uniquely names each variable corresponding to a particular dietary component. For example, in the OUT.PCTU dataset the prefix PCTU denotes a variable whose values are percentiles for a usual intake distribution. Combining PCTU with the rootname KCAL results in the variable named PCTUKCAL. This variable has values that are percentiles of the usual intake distribution for the variable CALORIES. In this chapter, variable prefixes that are combined with rootnames are labeled with four asterisks representing the rootname, i.e. PCTU****.

Data Sets	Comments
BIASADJ	Offsets and weights for 9-point bias adjustment
GP1	Grafted polynomial parameter estimates for initial normality transformation
GPU	Grafted polynomial parameter estimates for usual intake normality transformation
PCT1	Percentiles and density function values for smoothed daily intakes
PCTU	Percentiles, standard errors, and density function values for usual intakes
NPP1	Data for normal probability plot of daily intakes
NPPU	Data for normal probability plot of usual intakes
SMOOTH	Adjusted daily intakes and estimated individual usual intakes
VARCOMP	Variance components for transformed daily intakes

Table 6.1. Data sets created by SIDE

Table 6.1 summarizes the data sets that *SIDE* can create. Each data set is described individually in the following sections.

Note: The variable named _INT_BY_ is included in each output data set, and contains the values of the subpopulation variable used internally by *SIDE* to index the groups defined by the variables you named in the description data set BYVAR. If you do not specify any subpopulation variables, the value of _INT_BY_ is always 1. All output data sets except SMOOTH and NPP1 are sorted in increasing order of the _INT_BY_ variable, so that all observations for a particular group appear together.

BIASADJ

Offsets and weights for 9-point bias adjustment.

Created if: **EST_ISU** = 'Y'

Variables:

_INT_BY_	Subpopulation index.
OFST****	Offsets used to estimate the original-scale usual intake corresponding to an observation in the transformed usual intake space. These offsets are multiplied by the measurement error standard deviation to obtain the

quantities denoted by c_j in the notation of Nusser, et al. (1996).

WGHT**** The corresponding weights w_j.

Note: The ability to create this data set is provided solely for the purposes of refining the software and is of interest only to the *SIDE* development team. However, it is mentioned here for the sake of completeness.

GP1

Parameter estimates, join points, and other information about the semiparametric normality transformations for the distributions of daily intakes.

Created if: **EST_DAY1** = ’Y’ and **SAVE_PAR** = ’Y’

Variables:

_INT_BY_ Subpopulation index.

GP1**** (See below.)

The size of this data set varies according to the value of the keyword **MAXJP**. This data set has 3*P* + 15 observations on *a* + 1 variables, where *P* is the value of **MAXJP**, and a is the number of nutrients analyzed.

❒ The first group of *P* observations contains parameter estimates for the grafted polynomial.

❒ The second group of *P* observations contains the values of the join points in the normal scale for the grafted polynomial.

❒ The third group of *P* observations contains the values of the join points in the original scale for the grafted polynomial.

If the grafted polynomial required only *p* join points to achieve an acceptable transformation, the last *P* - *p* observations in each of the three groups are zeros. The remaining 15 observations contain additional information about the grafted polynomial.

Of these, the most important are the 3*P* + 13th and 3*P* + l2th observations, which are the number of join points *p* and the inverse of the power used in the first step of the semiparametric transformation.

Note: The ability to create this data set is provided primarily for the purposes of refining the *SIDE* software. Otherwise, it is not terribly useful. However, it is described here for the sake of completeness.

GPU

Parameter estimates, join points, and other information about the semiparametric normality transformations for the distributions of usual intakes.

Created if: **EST_ISU** = 'Y' and **SAVE_PAR** = 'Y'

Variables:

_INT_BY_	Subpopulation index.
GPU****	(See below.)

The size of this data set varies according to the value of the keyword **MAXJP**. This data set has 3*P* + 15 observations on *a* + 1 variables, where *P* is the value of **MAXJP**, and *a* is the number of nutrients analyzed.

- ❐ The first group of *P* observations contains parameter estimates for the grafted polynomial.
- ❐ The second group of *P* observations contains the values of the join points in the original scale for the grafted polynomial.
- ❐ The third group of *P* observations contains the values of the join points in the normal scale for the grafted polynomial.

If the grafted polynomial required only *p* join points to achieve an acceptable transformation, the last *P* - *p* observations in each of the three groups are zeros. The remaining 15 observations contain additional information about the grafted polynomial. Of these, the most important are the 3*P* + 13th and 3*P* + 11th observations, which are the number of join points *p* and the inverse of the power used in the first step of the semiparametric transformation.

Note: The ability to create this data set is primarily for the purposes of refining the *SIDE* software. It is described here for the sake of completeness.

PCT1

Percentiles and density function values for smoothed daily intakes.

Created if: **EST_DAY1** = ’Y’ and **SAVE_PCT** = ’Y’

Variables:

_INT_BY_	Subpopulation index.
PROB	Probability values for which the percentiles are computed.
PCT1****	Percentiles for the smoothed daily intake distribution.
DEN1****	Density function for the smoothed daily intake distribution.

The program will always generate 41 preselected percentiles (increments of 0.025 with the addition of 0.01 and 0.99). To create additional percentiles, set the keyword **NPTS** to a number greater than 41.

The additional percentiles are chosen to be equally-spaced in probability from $\frac{1}{n+1}$ to $\frac{n}{n+1}$ in steps of $\frac{1}{n+1}$. Hence, if you want the lst through 99th percentiles, set **NPTS** to 99.

Suggested Use: Plotting PROB against PCT1**** produces a plot of the CDF for the distribution of daily intakes. Plotting DEN1**** against PCT1**** produces a plot of the probability density function (pdf) for the distribution of daily intakes.

PCTU

Percentiles, standard errors, and density function values for usual intakes.

Created if: **EST_ISU** = ’Y’ and **SAVE_PCT** = ’Y’

Variables:

_INT_BY_	Subpopulation index.
PROB	Probability values for which the percentiles are computed.
PCTU****	Percentiles for the usual intake distribution.
DENU****	Density function for the usual intake distribution.
SEPU****	Estimated standard errors of the percentiles for the usual intake distribution.

The standard errors SEPU**** are computed under the assumptions of simple random sampling and nonconsecutive intake days. A balanced repeated replication method [Wolter (1985)] may be used to compute appropriate standard errors when these assumptions are not met.

The program will always generate 41 pre-selected percentiles (increments of 0.025 with the addition of 0.01 and 0.99). To create additional percentiles, set the keyword **NPTS** to a number greater than 41.

The additional percentiles are chosen to be equally-spaced in probability from $\frac{1}{n+1}$ to $\frac{n}{n+1}$ in steps of $\frac{1}{n+1}$. Hence, if you want the lst through 99th percentiles, set NPTS to 99.

Suggested Use: Plotting PROB against PCTU**** produces a plot of the **CDF** for the distribution of usual intakes. Plotting DENU**** against PCTU**** produces a plot of the probability density function (pdf) for the distribution of usual intakes.

NPP1

Data for normal probability plot of daily intakes.

Created if: **EST_DAY** = 'Y' and **SAVE_NPP** = 'Y'

Variables:

_INT_BY_ Subpopulation index.

TYE**** Shifted, power-transformed, equal-weight daily intakes used as the independent variable in the grafted polynomial fit for the first semiparametric normality transformation.

ZE_**** Blom normal scores used as the dependent variable in the grafted polynomial fit for the inital normality transformation.

ZEP**** Predicted Blom scores computed from the grafted polynomial fit for the inital normality transformation.

Note: Unlike most other output data sets, NPP1 is not sorted by the values of _INT_BY_. The observations in this data set are transformed values of observations in the input data set, and are arranged in the same manner.

Suggested Use: Plotting ZE_**** against TYE**** produces a normal probability plot of the shifted, power transformed, equal-weight daily intakes. Plotting ZEP**** against TYE**** produces a plot of the fitted grafted polynomial used as the first semiparametric normality transformation. If you overlay the plots just described, you can see how well the grafted polynomial fit the data. If the fit looks poor, you can repeat the analysis using different values of **ADALPHA, MAXJP,** and **LINFRAC**.

NPPU

Data for normal probability plot of usual intakes.

Created if: **EST_ISU** = 'Y' and **SAVE_NPP** = 'Y'

Variables:

_INT_BY_ Subpopulation index.

YU_**** Estimated usual intakes used as the dependent variable in the grafted polynomial fit for the usual intake transformation.

TYU**** Shifted, power-transformed, estimated usual intakes used as the dependent variable in the grafted polynomial fit for the usual intake transformation.

ZU_**** Blom normal scores used as the independent variable in the grafted polynomial fit for the usual intake transformation.

YUP**** Predicted usual intakes computed from the grafted polynomial fit for the usual intake transformation.

Note: The observations in this data set are representative of a random sample from the usual intake distribution. They do not correspond to estimates of usual intakes for the individuals in the input data set.

Suggested Use: Plotting YU_**** against ZU_**** produces a normal probability plot of the estimated usual intakes. Plotting YUP**** against ZU_**** produces a plot of the semiparametric transformation that takes the usual intakes into normality.

SMOOTH

Adjusted daily intakes and estimated individual usual intakes.

Created if: **SAVE_SMO** = 'Y'

Variables:

_INT_BY_ Subpopulation index.
(CLASSVAR variables)
(CONTSVAR variables)
(ID variables)
(WTVAR variables)

YR**** Ratio-adjusted dietary components (appears only if at least one CLASSVAR or CONTSVAR variable is specified).

YH**** Homogenized dietary components (appears only if NDAY is larger than 1).

YE**** Equal-weight dietary components (appears only if one or more WTVAR variables are specified).

UI**** Individual usual intakes for dietary components (appears only if **INDIVUI** and **EST_ISU** are both ’Y’).

This data set contains results of the preliminary smoothing procedures that are applied to the observed intakes before the usual intake estimation is performed. If both **INDIVUI** and **EST_ISU** are set to ’Y’, this data set also contains estimated usual intakes of dietary components for the individuals represented in the input data set.

This data set contains the same number of observations as the input data set, and is arranged in the same manner. Hence, unlike most other output data sets, SMOOTH is not necessarily sorted by the values of _INT_BY_. If individual usual intakes are present, each individual’s usual intake for a dietary component is repeated for each observation concerning that inidividual. To illustrate, suppose individual 1 has three observed intakes for the variable CALC. Then three observations corresponding to individual 1 will appear in the SMOOTH data set. Each of these three observations will have the same value of UICALC, namely, the estimated usual intake of CALC for individual 1.

Note: If individual sampling weights are available, they may be used when analyzing the estimated individual usual intakes.

VARCOMP

Variance components for transformed daily intakes.

Created if: **EST_ISU** = ’Y’

Variables:

_INT_BY_ Subpopulation index.

VC**** Values of the variance components reported on the printed output.

Parameter estimates for the regression of individual standard deviations on individual means in the normal scale are also included, along with the covariance matrix of the estimates. If **TRANDIAG** is set to 'Y', this information is also printed on the output. The estimated regression equation is:

$$s_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{i.} \ ,$$

where *i* ranges from 1 to *ñ*, the number of individuals in the data set with more than 1 observed daily intake, and

n_i = the number of observed intakes for individual *i*

x_{ij} = the *j*-th observed intake for individual *i*

$\bar{x}_{i.}$ = $n_i^{-1} \sum_{i=1}^{n_i} x_{ij}$

s_i = $\sqrt{(n_i - 1)^{-1} \ \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}$

Some additional information is also provided as an aid to the *SIDE* development team. This data set always contains 21 observations. The observations for each component variable should be thought of as a column vector, with entries arranged as shown in Table 6.2.

Position	Value Reported on Output As
1	Mean
2	Variance
3	Variance of Usual Intakes
4	Variance of Measurement Error
5	t for Heterogeneous Variance
6	Scaled 4th Moment of M.E.
7	Variance of Ind. Variances
8	# of Individuals with Ni > 1
9	Variance of Mean
10	Variance of U.I. Variance
11	Number of individuals
12	Measurement Error d.f.
13	Not reported
14	$\hat{\beta}_0$
15	$\hat{\beta}_1$
16	$Var\ (\hat{\beta}_0)$
17	$Cov\ (\hat{\beta}_0, \hat{\beta}_1)$
18	$Var\ (\hat{\beta}_1)$
19	Not reported
20	F ratio for linear model (1 d.f.)
21	p-value for F test

Table 6.2. Arrangement of variance component vector

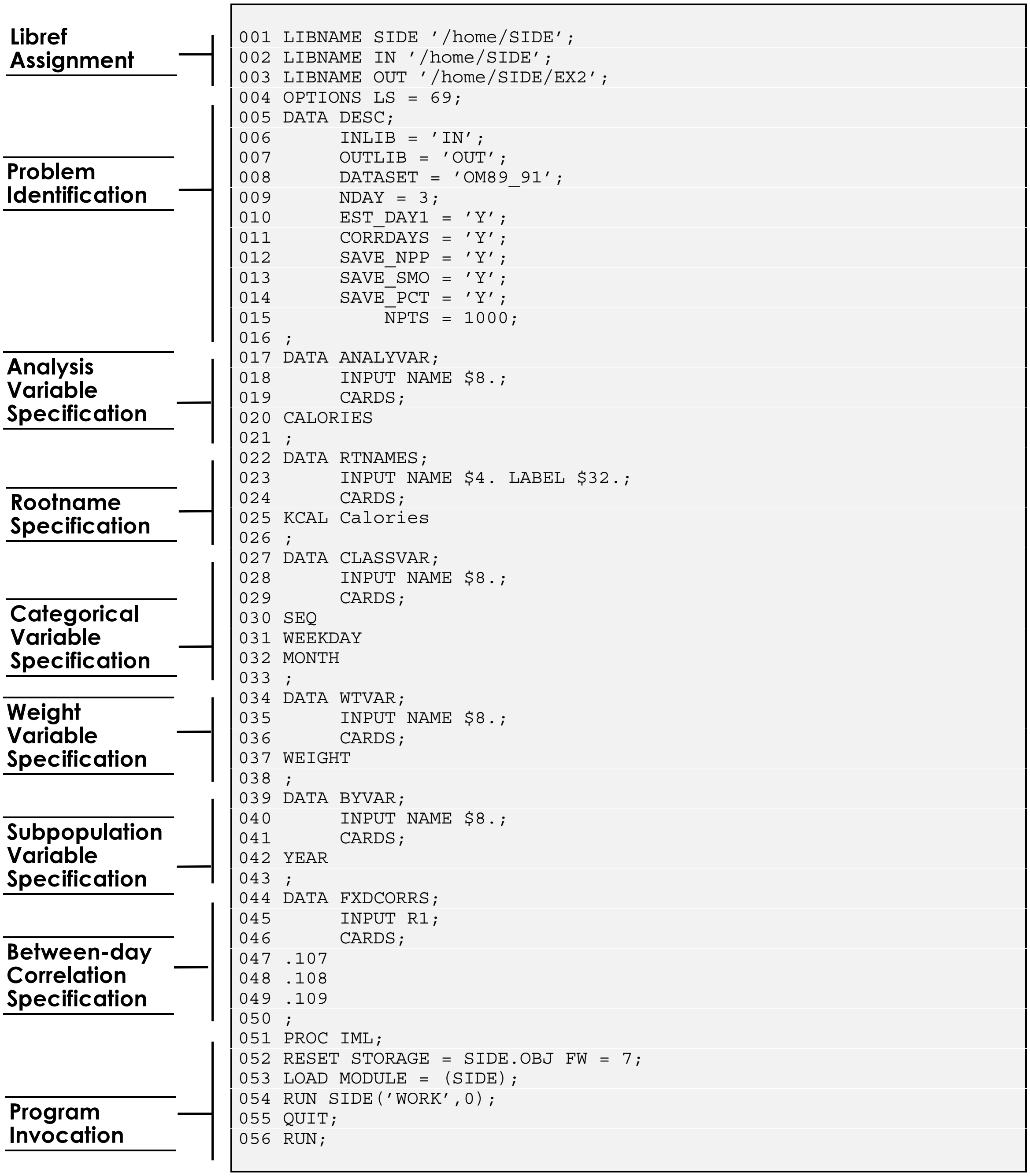
VII: Additional Examples

This section contains two SAS© programs, **example2.sas** and **example3.sas**, that demonstrate some of the advanced features of *SIDE.* Before attempting to work through this example, the user must have created the data sets as described in “Creating the SAS Data Sets for Example Programs” in the chapter “Installing *SIDE*.” Due to space considerations, only selected portions of the printed output generated by the example programs are reproduced in the manual.

Example 2

This example performs a more complex analysis on the data contained in set IN.OM89_91, the same data set used in Example 1. Refer to the chapter “Installing *SIDE*” for a description of the variables in this data set. As mentioned in Example 1, the data for each individual were collected on consecutive days; so it is reasonable to assume that the observations for each individual are correlated. The **example2.sas** program demonstrates how to use the FXDCORRS data set to account for this correlation. The program also demonstrates how to use the BYVAR data set and the **NPTS** keyword.

The two pages of the program **example2.sas** are shown in Figures 7.1 and 7.2, respectively. Note that the specification of the DESC data set in


Libref Assignment
Problem Identification
Analysis Variable Specification
Rootname Specification
Categorical Variable Specification
Weight Variable Specification
Subpopulation Variable Specification
Between-day Correlation Specification
Program Invocation

```
LIBNAME SIDE '/home/SIDE';
LIBNAME IN '/home/SIDE';
LIBNAME OUT '/home/SIDE/EX2';
OPTIONS LS = 69;
DATA DESC;
     INLIB = 'IN';
     OUTLIB = 'OUT';
     DATASET = 'OM89_91';
     NDAY = 3;
     EST_DAY1 = 'Y';
     CORRDAYS = 'Y';
     SAVE_NPP = 'Y';
     SAVE_SMO = 'Y';
     SAVE_PCT = 'Y';
         NPTS = 1000;
;
DATA ANALYVAR;
     INPUT NAME $8.;
     CARDS;
CALORIES
;
DATA RTNAMES;
     INPUT NAME $4. LABEL $32.;
     CARDS;
KCAL Calories
;
DATA CLASSVAR;
     INPUT NAME $8.;
     CARDS;
SEQ
WEEKDAY
MONTH
;
DATA WTVAR;
     INPUT NAME $8.;
     CARDS;
WEIGHT
;
DATA BYVAR;
     INPUT NAME $8.;
     CARDS;
YEAR
;
DATA FXDCORRS;
     INPUT R1;
     CARDS;
.107
.108
.109
;
PROC IML;
RESET STORAGE = SIDE.OBJ FW = 7;
LOAD MODULE = (SIDE);
RUN SIDE('WORK',0);
QUIT;
RUN;
```

Figure 7.1. Page 1 of ***example2.sas***

```
/* Generate normal probability and density function plots */;
AXIS1 LABEL=(’Normal Score’) OFFSET=(0,.1) IN MINOR=NONE;
AXIS2 LABEL=(A = 90 ’Power-Transformed Daily Intakes of KCAL’)
      MINOR=NONE OFFSET=(0,.1) IN;
AXIS3 LABEL=(’Usual Intake of KCAL’) MINOR=NONE OFFSET=(0,.1) IN;
AXIS4 LABEL=(A=90 ’Probability Density’) MINOR=NONE
      OFFSET=(0,.1) IN;
LEGEND1 SHAPE = LINE(8 PCT) CBORDER = BLACK LABEL = NONE
        ACROSS = 3 OFFSET = (,6 PCT);
SYMBOL1 V=NONE I=SPLINES L=1;
SYMBOL2 V=NONE I=SPLINES L=3;
SYMBOL3 V=NONE I=SPLINES L=8;
SYMBOL4 V=DOT I=NONE H=.7;
GOPTIONS GSFMODE = REPLACE DEVICE = EPS45
         TARGET = EPS45 HSIZE = 4.34 IN VSIZE = 4.34 IN
         HORIGIN = 0.16 IN VORIGIN = 0.16 IN BORDER
         FTEXT = DUPLEX HTEXT = 2.6 FTITLE = SWISS
         HTITLE = 2.6;
TITLE ’Normal Probability Plot of Daily Intake Transformation’;

FILENAME FIG1 ’/home/SIDE/EX2/fig7_5.eps’;
GOPTIONS GSFNAME=FIG1;
PROC GPLOT DATA=OUT.NPP1(WHERE =( _INT_BY_  =  1));
     PLOT TYEKCAL*ZE_KCAL=4 TYEKCAL*ZEPKCAL=1/OVERLAY HAXIS=AXIS1
                                              VAXIS=AXIS2;
TITLE2 H = 2.6 ’Year  =  1989’;
RUN;

FILENAME FIG2 ’/home/SIDE/EX2/fig7_6.eps’;
GOPTIONS GSFNAME = FIG2;
PROC GPLOT DATA=OUT.NPP1(WHERE =( _INT_BY_  =  2));
     PLOT TYEKCAL*ZE_KCAL=4 TYEKCAL*ZEPKCAL=1/OVERLAY HAXIS=AXIS1
                                              VAXIS=AXIS2;
TITLE2 H = 2.6 ’Year  =  1990’;
RUN;

FILENAME FIG3 ’/home/SIDE/EX2/fig7_7.eps’;
GOPTIONS GSFNAME = FIG3;
PROC GPLOT DATA=OUT.NPP1(WHERE =( _INT_BY_  =  3));
     PLOT TYEKCAL*ZE_KCAL=4 TYEKCAL*ZEPKCAL=1/OVERLAY HAXIS=AXIS1
                                              VAXIS=AXIS2;
TITLE2 H = 2.6 ’Year  =  1991’;
RUN;

PROC FORMAT;
     VALUE YR 1 = ’1989’
              2 = ’1990’
              3 = ’1991’;
FILENAME FIG4 ’/home/SIDE/EX2/fig7_8.eps’;
GOPTIONS GSFNAME = FIG4;
PROC GPLOT DATA=OUT.PCTU;
     PLOT DENUKCAL*PCTUKCAL=_INT_BY_/VAXIS=AXIS4 HAXIS=AXIS3
                                     LEGEND = LEGEND1;
     FORMAT _INT_BY_ YR.;
TITLE ’Usual Intake Distributions for KCAL’;
RUN;
```

SAS/GRAPH© Commands to Create Plots

Figure 7.2. Page 2 of ***example 2.sas***

example2.sas is identical to the specification of the DESC data set in **example1.sas**, except that the **NPTS** keyword appears in **example2.sas**, specifying that an additional 1000 percentiles per group are to be stored in the percentile data sets PCT1 and PCTU.

The ANALYVAR, RTNAMES, IDVAR, and WTVAR data sets are identical in the two programs. Note, however, that the variable name YEAR now appears in the BYVAR dataset, instead of the CLASSVAR data set, specifying that the usual intake distribution for calories is to be estimated separately for the three years represented in the survey.

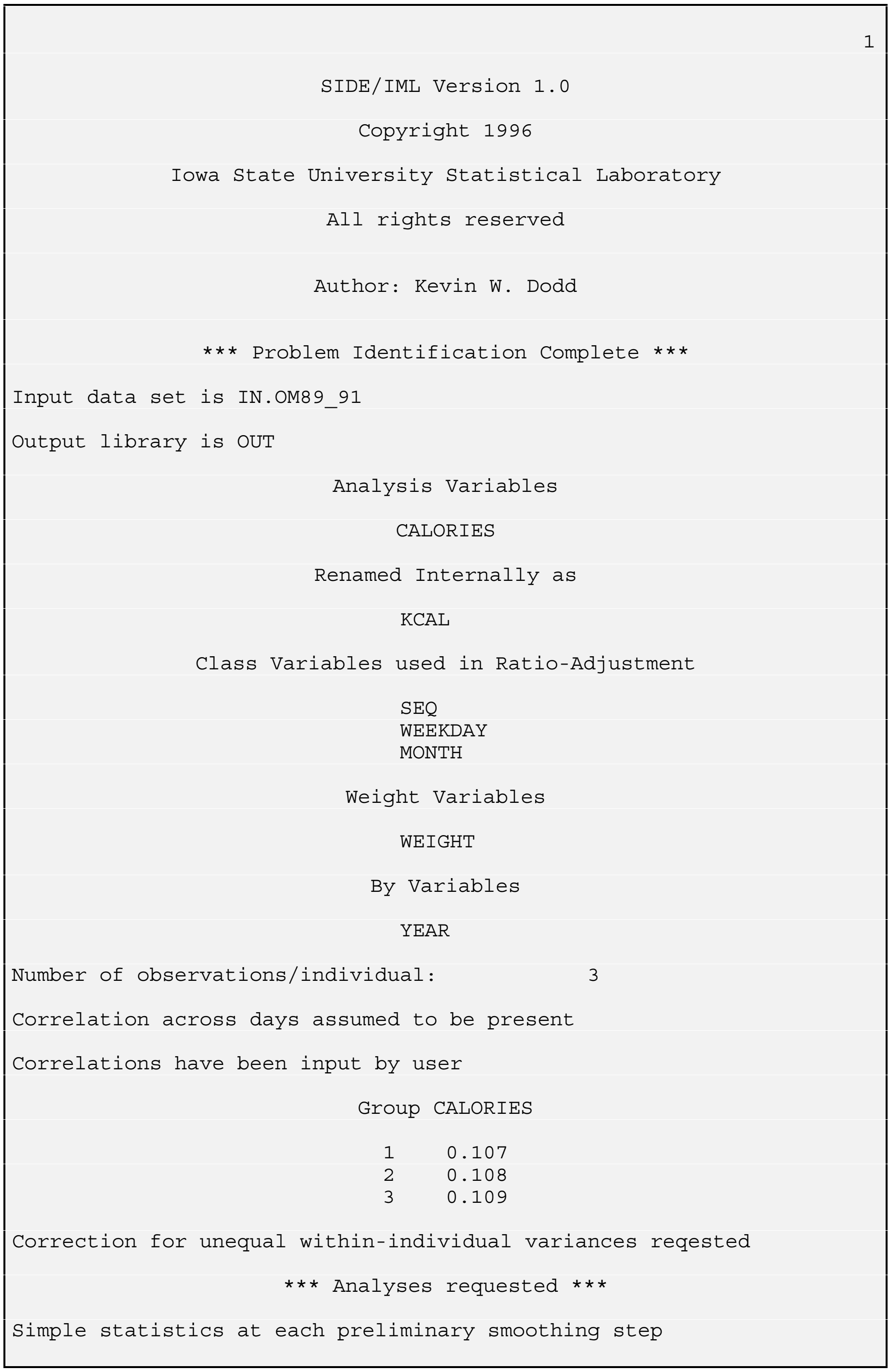
Lines 44-50 create the data set FXDCORRS. In this example, the within-individual correlations specified for years 1, 2, and 3 are 0.107, 0.108, and 0.109, respectively. Note that if only one correlation value were to be specified in the FXDCORRS data set, the specified value would be used as the common correlation for all three years.

The first two pages of the printed output from the **example2.sas** program are shown in Figures 7.3 and 7.4. The SAS©/GRAPH commands that appear on lines 59-113 generate the plots shown in Figures 7.5-7.8. See *SAS©/GRAPH Software: Reference* (1990). A description of the variables referenced in the PLOT statements on lines 81, 89, 97, and 109 can be found in the chapter "Output Data Sets." Plots of the type shown in Figures 7.5-7.7 (known as *normal probability plots*) are useful in assessing the performance of the normality transformations used in the usual intake estimation procedure. The points connected by the solid line should appear to be a very close, smooth fit to the Blom normal scores (the dots). Sometimes a relatively poor fit in the outermost segments can be improved by altering the fraction of observations allocated to the linear segments (by changing the value of the **LINFRAC** keyword). The plot shown in Figure 7.8 is a convenient way to graphically display characteristics of the usual intake distributions.

If you wish to run this example on your system, be sure to change the LIBNAME statements on program lines 1, 2,and 3, and if necessary, the FILENAME statements on lines 78, 86, 94, and 106. By default, the plots are produced as Encapsulated PostScript files. You may need to change the GOPTIONS statements on lines 71-75, 79, 87, 95, and 107 to values appropriate to your computer system.[1]

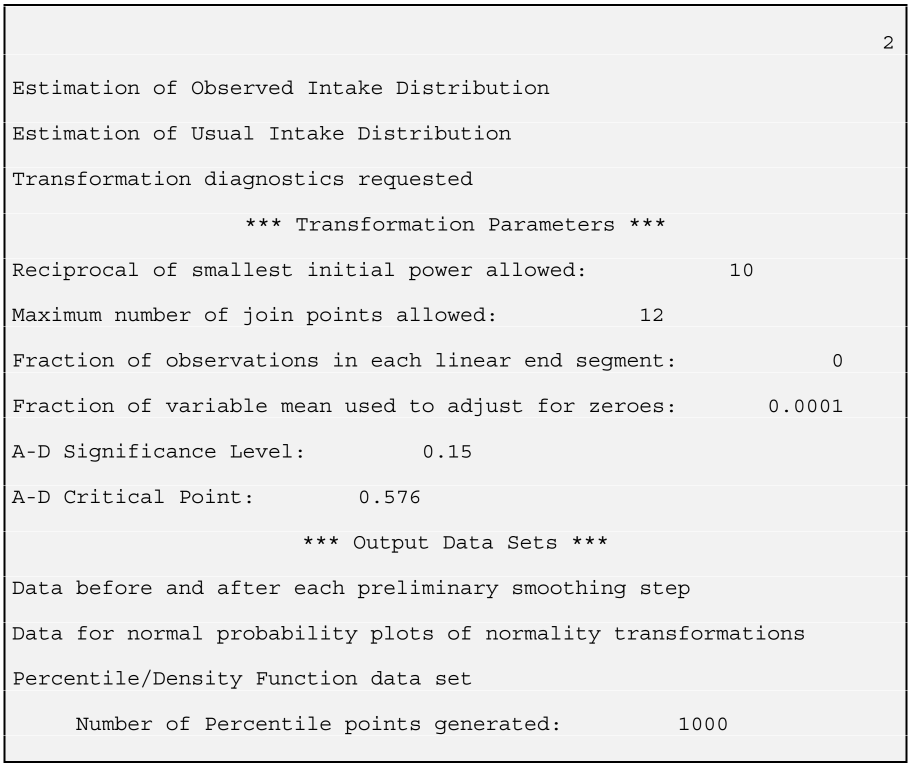
The example code assigns the same directory to the librefs SIDE and IN, because the example data sets are stored in the same directory as the SAS© data catalog SIDE. In actual use, putting the program, input files, and output files in different directories will allow for better data management.

[1] The EPS45 device driver is a user-modified device description based on the institute-supplied driver PSLEPSF. Use PROC GDEVICE to create your own version with the following parameters: XMAX=YMAX=4.5 IN, XPIXELS=YPIXELS=2700, HSIZE=VSIZE=4.34 IN, HORIGIN=VORIGIN=0.16 IN, PROWS=PCOLS=75, LROWS=LCOLS=0, ROTATE=LANDSCAPE.



```
                                                                    1

                         SIDE/IML Version 1.0

                            Copyright 1996

             Iowa State University Statistical Laboratory

                         All rights reserved


                        Author: Kevin W. Dodd


               *** Problem Identification Complete ***

Input data set is IN.OM89_91

Output library is OUT

                          Analysis Variables

                               CALORIES

                        Renamed Internally as

                               KCAL

               Class Variables used in Ratio-Adjustment

                               SEQ
                               WEEKDAY
                               MONTH

                           Weight Variables

                               WEIGHT

                             By Variables

                               YEAR

Number of observations/individual:            3

Correlation across days assumed to be present

Correlations have been input by user

                            Group CALORIES

                              1    0.107
                              2    0.108
                              3    0.109

Correction for unequal within-individual variances reqested

                      *** Analyses requested ***

Simple statistics at each preliminary smoothing step
```

Figure 7.3. First page of output from ***example2.sas***



```
                                                                    2
Estimation of Observed Intake Distribution
Estimation of Usual Intake Distribution
Transformation diagnostics requested
                  *** Transformation Parameters ***
Reciprocal of smallest initial power allowed:           10
Maximum number of join points allowed:           12
Fraction of observations in each linear end segment:           0
Fraction of variable mean used to adjust for zeroes:       0.0001
A-D Significance Level:         0.15
A-D Critical Point:        0.576
                       *** Output Data Sets ***
Data before and after each preliminary smoothing step
Data for normal probability plots of normality transformations
Percentile/Density Function data set
     Number of Percentile points generated:         1000
```

Figure 7.4. Second page of output from ***example2.sas***

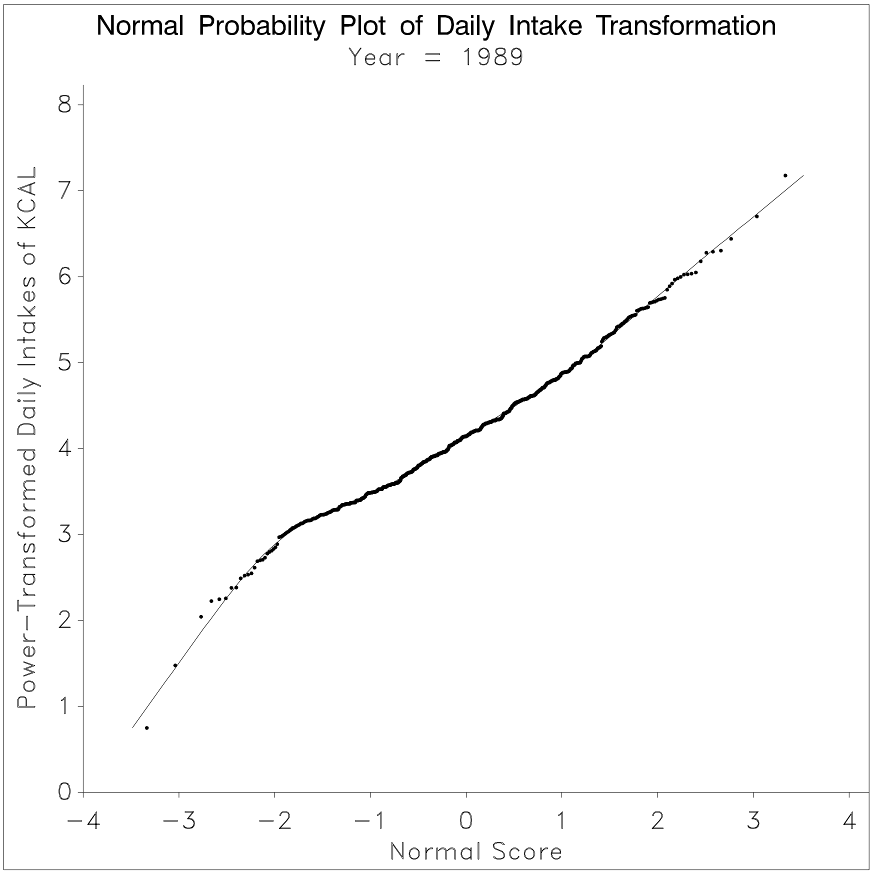


Figure 7.5. Plot of initial semiparametric normality transformation for data corresponding to the first survey year.

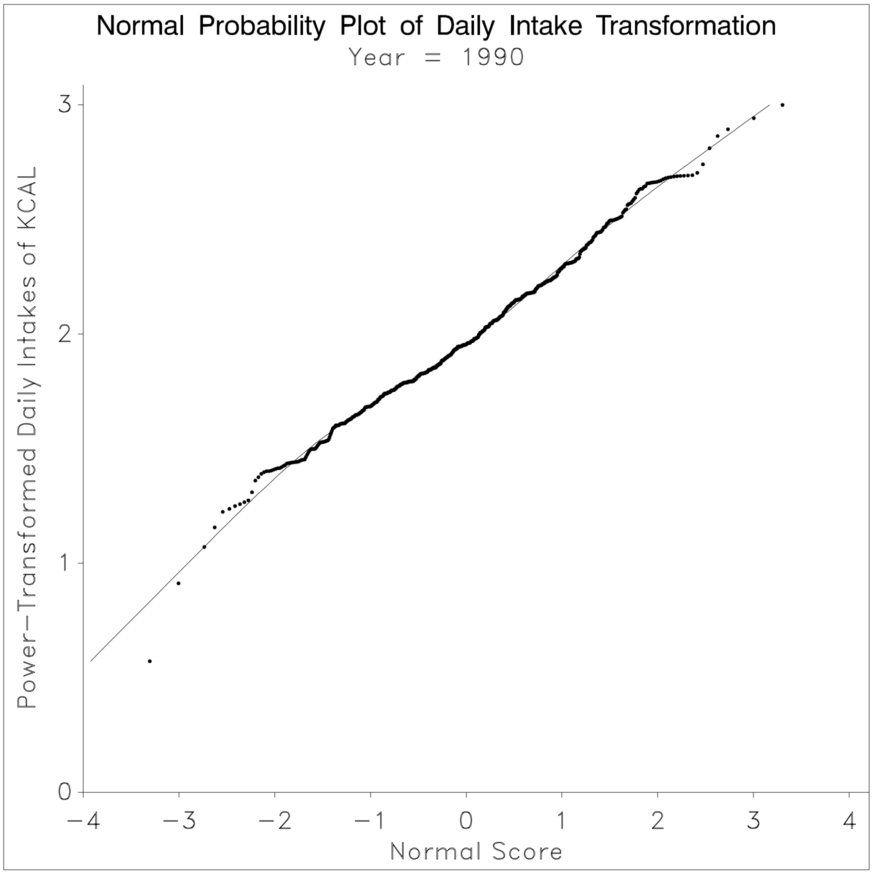


Figure 7.6. Plot of initial semiparametric normality transformation for data corresponding to the second survey year.

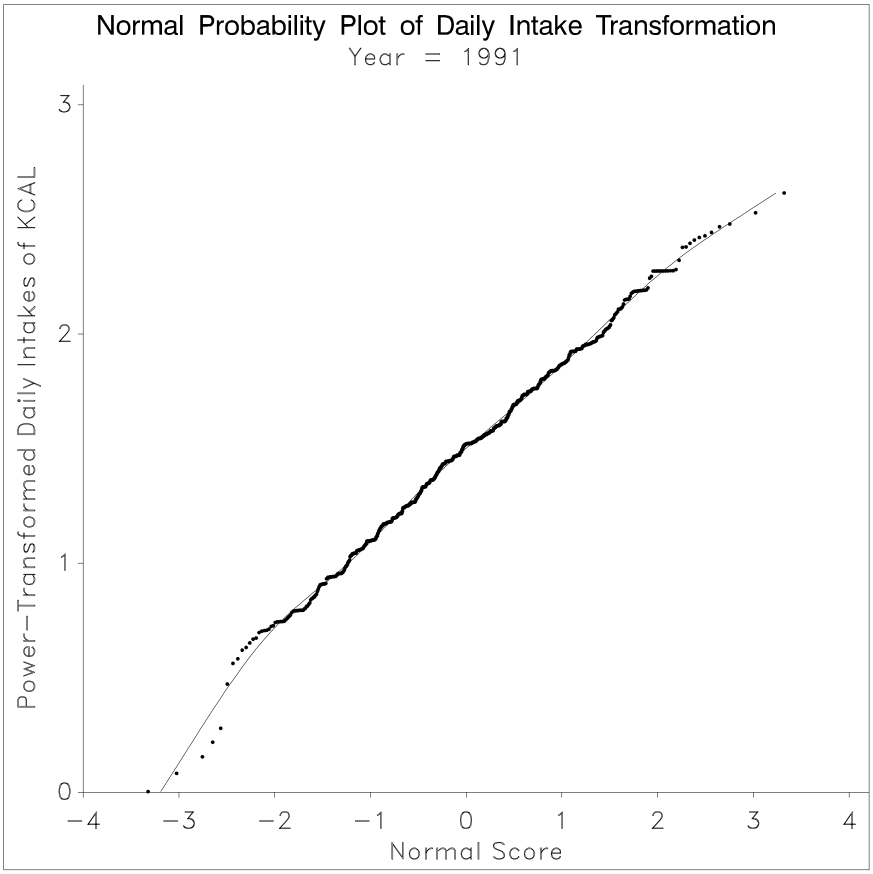


Figure 7.7. Plot of initial semiparametric normality transformation for data corresponding to the third survey year.

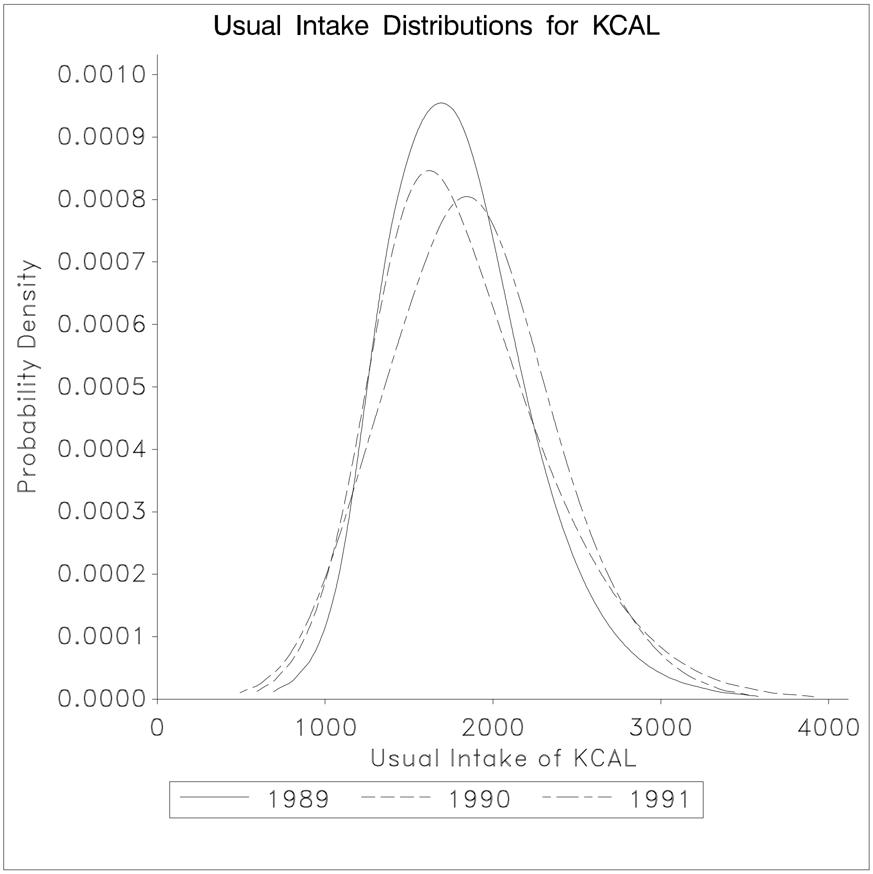


Figure 7.8. Estimated densities of usual intake distributions for each survey year.

Example 3

This example demonstrates the use of the balanced repeated replication method [Wolter (1985)] to estimate standard errors of percentiles of usual intake distributions. To use the balanced repeated replication (BRR) method to estimate the standard errors for the percentiles of the usual intake distribution for a nutrient, it is necessary to estimate the usual intake distribution using the data for the full sample and data for each of *R* half-samples that are created using information in the sampling design.

The estimated standard error of an estimated percentile, $\hat{\theta}_0$, is the square root of $\hat{V}\{\hat{\theta}_0\} = (R)^{-1}\sum_{i=1}^{R} (\hat{\theta}_i - \hat{\theta}_0)^2$, where $\hat{\theta}_i$ denotes the estimated percentile computed using the *i*-th half-sample and $\hat{\theta}_0$ denotes the estimated percentile computed using the full sample. In practice, you must define the half-samples as explained in the Technical Note at the end of this example.

The program **example3.sas** performs an analysis of the dietary component IRON from the example data set IN.YW85. This data set should be created as described in "Creating the SAS Data Sets for Example Programming" in the chapter "Installing *SIDE*" before this example is attempted.

The example data set IN.YW85 is a subset of the 1985 CSFII (Continuing Survey of Food Intakes by Individuals) containing observations for 737 non-pregnant, non-lactating women who were main meal planners. Observations were taken on four non-consecutive days for the variables listed in Table 7.1. The individuals sampled were selected according to a complex survey design. Sampling weights are provided in the input data set, contained in the variable WEIGHT.

Variable	**Description**
INDIV	Individual ID
DAY	Interview sequence: 1, 2, 3, or 4
DOW	Day of week
CALC	Calcium intake (milligrams)
KCAL	Energy intake (kilocalories)
IRON	Iron intake (milligrams)
PROT	Protein intake (grams)
VITA	Vitamin A intake (µg/RE)
VITC	Vitamin C intake (milligrams)
WEIGHT	Sampling weight

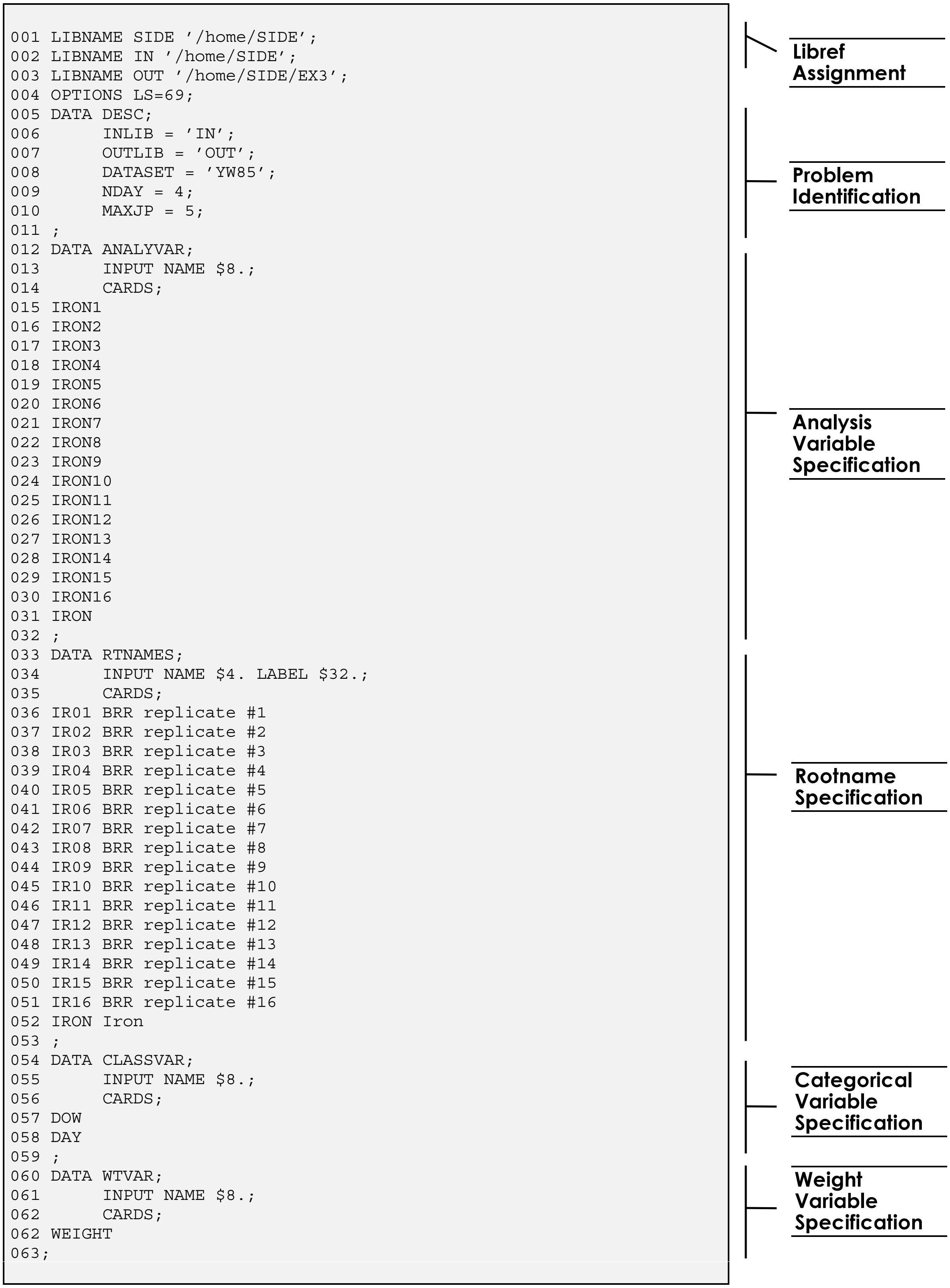
Table 7.1. Variable listing for IN.YW85

The 1985 CSFII sampling design is a stratified sample with two primary sampling units per stratum. Some strata were combined to create a sample of 48 strata, each with two primary sampling units. The text file **yw85.txt** contains indicator variables that are used in the **makeds.sas** program to create *R*=16 balanced half-samples by selectively deleting all observations from one primary sampling unit per stratum. Because *SIDE* requires the data set to have $737 \times 4 = 2948$ observations, the deletion is accomplished by setting the appropriate nutrient variables to missing values. The sixteen half-samples for IRON are contained in the variables IRON1-IRON16.

The program **example3.sas** is shown in Figures 7.9 and 7.10. The most important sections of the program are lines 12-32, where the ANALYVAR data set is created, lines 33-53, where the RTNAMES data set is created, and lines 60-63, where the WTVAR data set is created. *SIDE* is told to analyze the sixteen half-samples and then the full sample using the same weight variable for each analysis. Note that the **MAXJP** keyword on line 10 has the value 5. None of the seventeen analysis variables requires a larger value, and *SIDE* runs faster if **MAXJP** is set to small values. The problem identification output is reproduced in Figures 7.11 and 7.12.

After the seventeen separate analyses are performed, the simple SAS© data manipulations shown on lines 72-85 are used to compute the BRR standard errors shown in Figure 7.14. These are different from the estimated standard errors shown in Figure 7.13, which are based on the assumption of simple random sampling.

If you wish to run this example on your system, be sure to change the LIBNAME statements on lines 1-3. The example code assigns the same directory to the librefs SIDE and IN, because the example data sets are stored in the same directory as the SAS© data catalog SIDE. In actual use, putting the program, input files, and output files in different directories will allow for better data management.



```
LIBNAME SIDE '/home/SIDE';
LIBNAME IN '/home/SIDE';
LIBNAME OUT '/home/SIDE/EX3';
OPTIONS LS=69;
DATA DESC;
     INLIB = 'IN';
     OUTLIB = 'OUT';
     DATASET = 'YW85';
     NDAY = 4;
     MAXJP = 5;
;
DATA ANALYVAR;
     INPUT NAME $8.;
     CARDS;
IRON1
IRON2
IRON3
IRON4
IRON5
IRON6
IRON7
IRON8
IRON9
IRON10
IRON11
IRON12
IRON13
IRON14
IRON15
IRON16
IRON
;
DATA RTNAMES;
     INPUT NAME $4. LABEL $32.;
     CARDS;
IR01 BRR replicate #1
IR02 BRR replicate #2
IR03 BRR replicate #3
IR04 BRR replicate #4
IR05 BRR replicate #5
IR06 BRR replicate #6
IR07 BRR replicate #7
IR08 BRR replicate #8
IR09 BRR replicate #9
IR10 BRR replicate #10
IR11 BRR replicate #11
IR12 BRR replicate #12
IR13 BRR replicate #13
IR14 BRR replicate #14
IR15 BRR replicate #15
IR16 BRR replicate #16
IRON Iron
;
DATA CLASSVAR;
     INPUT NAME $8.;
     CARDS;
DOW
DAY
;
DATA WTVAR;
     INPUT NAME $8.;
     CARDS;
WEIGHT
;
```

Libref Assignment
Problem Identification
Analysis Variable Specification
Rootname Specification
Categorical Variable Specification
Weight Variable Specification

Figure 7.9. First page of ***example3.sas***

Program Invocation

```
PROC IML;
RESET STORAGE = SIDE.OBJ FW = 7;
LOAD MODULE = (SIDE);
RUN SIDE('WORK',0);
QUIT;
RUN;

```

Calculation of BRR Standard Errors

```
/* Compute BRR standard errors for usual intake percentiles */;
DATA ONE;
     SET OUT.PCTU;
     ARRAY REST (16) PCTUIR01-PCTUIR16;
     DO I = 1 TO 16;
        REST(I) = REST(I) - PCTUIRON;
     END;
     BRRSE = SQRT(USS(OF PCTUIR01-PCTUIR16)/16);
     KEEP PCTUIRON PROB BRRSE;
RUN;
OPTIONS CENTER;
TITLE 'Percentiles and Standard Errors Estimated via BRR';
PROC PRINT;
     VAR PROB PCTUIRON BRRSE;
RUN;
```

Figure 7.10. Second page of ***example3.sas***

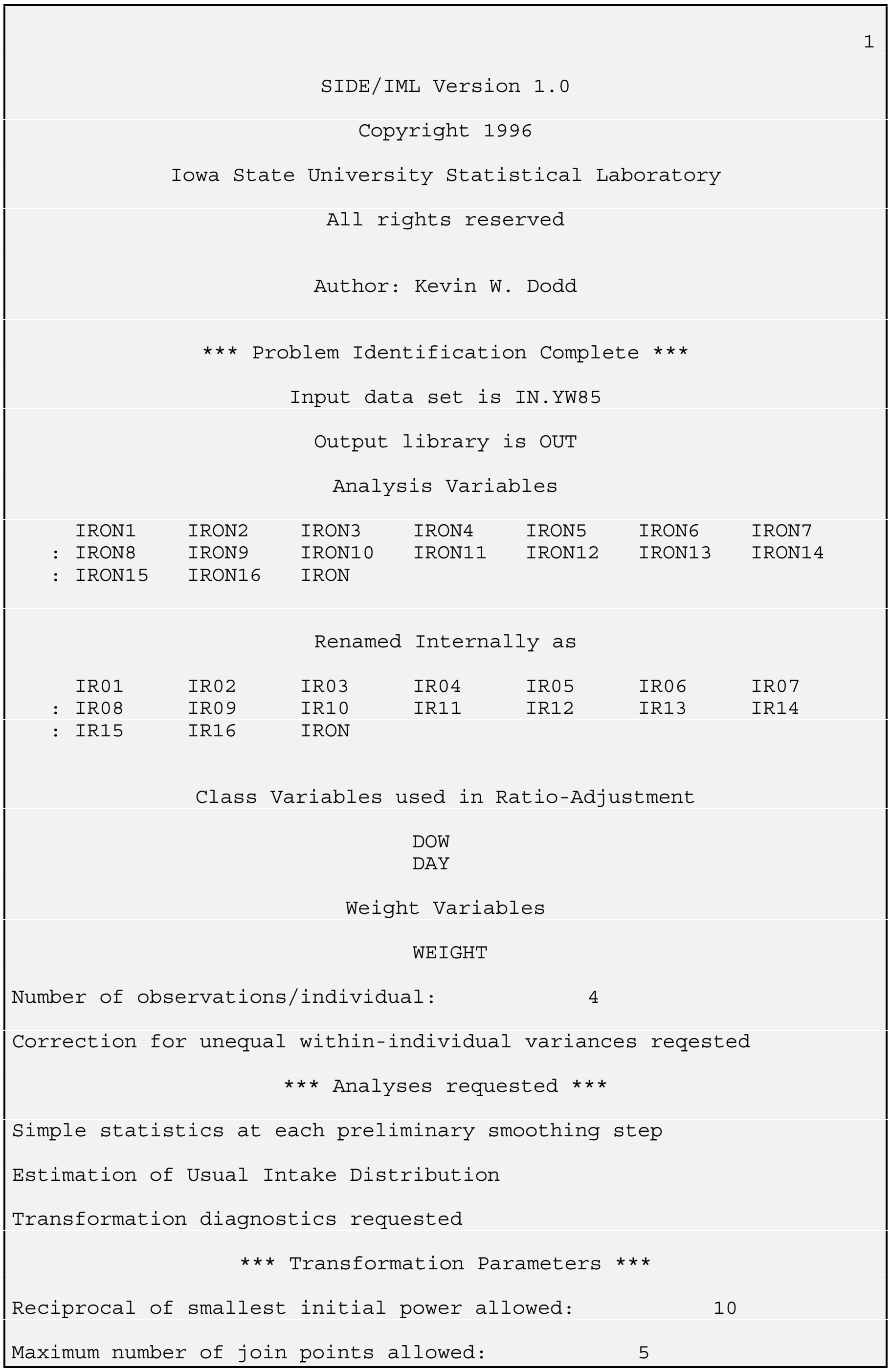
Technical Note:

The balanced repeated replication method demonstrated in this example works for the special case where each stratum represented in the survey has two primary sampling units (PSUs). If you wish to apply BRR to your own data, you must first make sure that each stratum contains two PSUs (in some cases, you may have to combine or split the existing strata into *variance estimation strata*, each with two PSUs). Suppose that you end up with L variance estimation strata. You may use any set of R orthogonal contrasts $\left\{ c_r \right\}_{r=1}^{R}$ of length L with elements ±1 to construct R balanced half-samples as follows: an observation from the k-th PSU in the l-th stratum (k = 1 or 2, l = 1,2,...,L) is included in the r-th half-sample if

- $k = 1$ and the l-th element of c_r is 1, or
- $k = 2$ and the l-th element of c_r is -1.

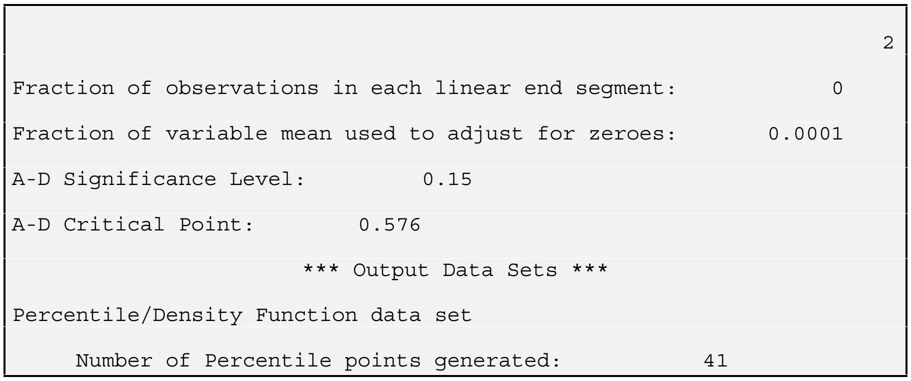
Let $\delta_j^{(r)} = \begin{cases} 1 & \text{if observation } j \text{ is included in the } r\text{-th half-sample} \\ 0 & \text{otherwise} \end{cases}$

for r=1,2,...,R, j=1,2,...,N, where N is the total number of observations in your data set. Once you have included the $\delta_j^{(r)}$ variables in your data set, lines 10-41 of the **makeds.sas** program can be modified to create the balanced half-samples for *SIDE* to analyze. In the **makeds.sas** program, the $\delta_j^{(r)}$ are named INREP1-INREP16, because in Example 3, R=16. For your data set, change all occurrences of '16' to your value of R. Lines 14-20, 24-29, and 32-37 create half-sample replicates for all of the analysis variables in the data set IN.YW85. You must modify these groups of lines to reflect the names and number of the analysis variables in your data set.



```
                                                                      1

                         SIDE/IML Version 1.0

                            Copyright 1996

              Iowa State University Statistical Laboratory

                         All rights reserved

                        Author: Kevin W. Dodd

                *** Problem Identification Complete ***

                       Input data set is IN.YW85

                         Output library is OUT

                           Analysis Variables

     IRON1    IRON2    IRON3    IRON4    IRON5    IRON6    IRON7
   : IRON8    IRON9    IRON10   IRON11   IRON12   IRON13   IRON14
   : IRON15   IRON16   IRON


                        Renamed Internally as

     IR01     IR02     IR03     IR04     IR05     IR06     IR07
   : IR08     IR09     IR10     IR11     IR12     IR13     IR14
   : IR15     IR16     IRON


               Class Variables used in Ratio-Adjustment

                                DOW
                                DAY

                           Weight Variables

                                WEIGHT

Number of observations/individual:            4

Correction for unequal within-individual variances reqested

                      *** Analyses requested ***

Simple statistics at each preliminary smoothing step

Estimation of Usual Intake Distribution

Transformation diagnostics requested

                   *** Transformation Parameters ***

Reciprocal of smallest initial power allowed:           10

Maximum number of join points allowed:           5
```

Figure 7.11. First page of output from ***example3.sas***

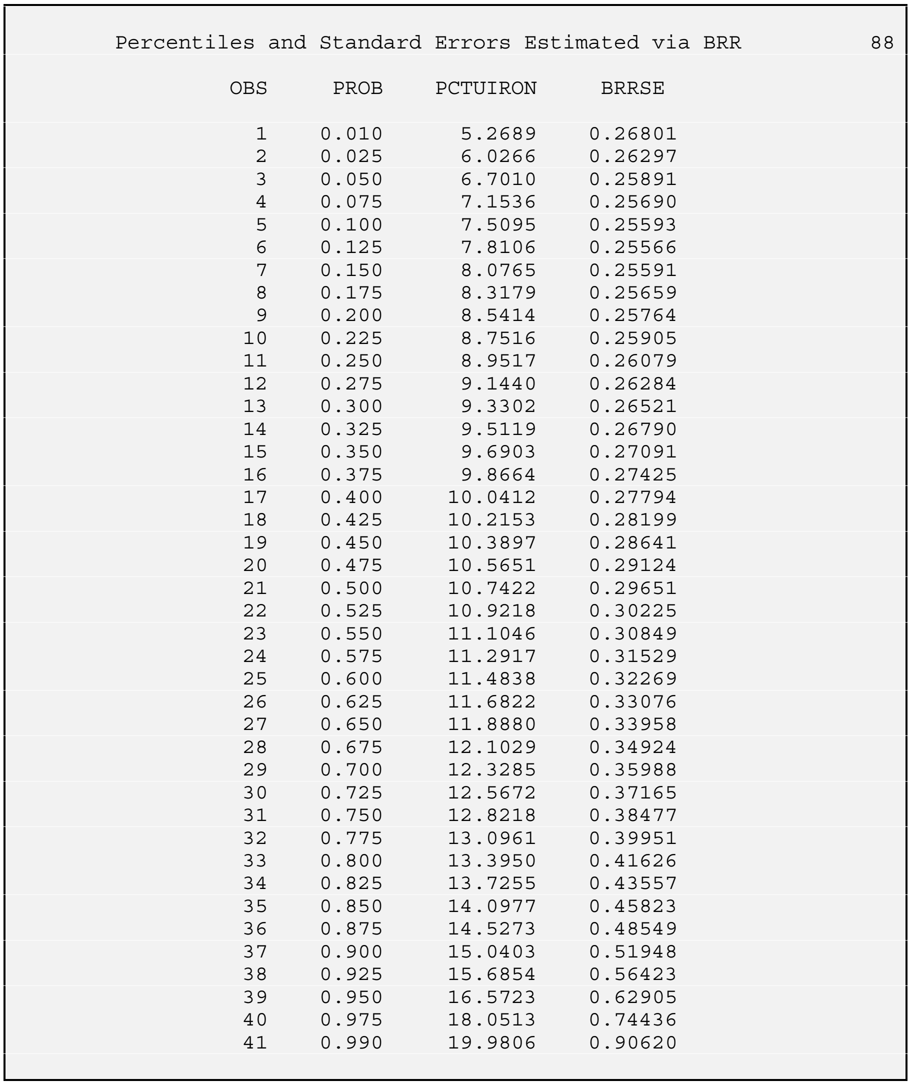


```
                                                                    2

Fraction of observations in each linear end segment:            0

Fraction of variable mean used to adjust for zeroes:       0.0001

A-D Significance Level:         0.15

A-D Critical Point:        0.576

                      *** Output Data Sets ***

Percentile/Density Function data set

     Number of Percentile points generated:           41
```

Figure 7.12. Second page of output from ***example3.sas***

```
Variable: IRON  Iron                                 Group:  1        87

            **** Percentile Values for Usual Intakes ****

                 Probability  Percentile  Std. Error

                    0.010     5.26892     0.24669
                    0.025     6.02660     0.23470
                    0.050     6.70105     0.22126
                    0.075     7.15356     0.21140
                    0.100     7.50948     0.20349
                    0.125     7.81063     0.19725
                    0.150     8.07650     0.19233
                    0.175     8.31787     0.18839
                    0.200     8.54142     0.18521
                    0.225     8.75165     0.18267
                    0.250     8.95171     0.18068
                    0.275     9.14396     0.17918
                    0.300     9.33021     0.17813
                    0.325     9.51193     0.17750
                    0.350     9.69032     0.17728
                    0.375     9.86643     0.17746
                    0.400     10.0412     0.17804
                    0.425     10.2153     0.17901
                    0.450     10.3897     0.18039
                    0.475     10.5651     0.18218
                    0.500     10.7422     0.18442
                    0.525     10.9218     0.18712
                    0.550     11.1046     0.19034
                    0.575     11.2917     0.19410
                    0.600     11.4838     0.19847
                    0.625     11.6822     0.20350
                    0.650     11.8880     0.20926
                    0.675     12.1029     0.21585
                    0.700     12.3285     0.22340
                    0.725     12.5672     0.23206
                    0.750     12.8218     0.24204
                    0.775     13.0961     0.25361
                    0.800     13.3950     0.26716
                    0.825     13.7255     0.28323
                    0.850     14.0977     0.30266
                    0.875     14.5273     0.32673
                    0.900     15.0403     0.35770
                    0.925     15.6854     0.39913
                    0.950     16.5723     0.45751
                    0.975     18.0513     0.55574
                    0.990     19.9806     0.67995
```

Figure 7.13. SRS standard errors for iron usual intake percentiles



```
         Percentiles and Standard Errors Estimated via BRR          88

                  OBS      PROB     PCTUIRON      BRRSE

                    1     0.010       5.2689     0.26801
                    2     0.025       6.0266     0.26297
                    3     0.050       6.7010     0.25891
                    4     0.075       7.1536     0.25690
                    5     0.100       7.5095     0.25593
                    6     0.125       7.8106     0.25566
                    7     0.150       8.0765     0.25591
                    8     0.175       8.3179     0.25659
                    9     0.200       8.5414     0.25764
                   10     0.225       8.7516     0.25905
                   11     0.250       8.9517     0.26079
                   12     0.275       9.1440     0.26284
                   13     0.300       9.3302     0.26521
                   14     0.325       9.5119     0.26790
                   15     0.350       9.6903     0.27091
                   16     0.375       9.8664     0.27425
                   17     0.400      10.0412     0.27794
                   18     0.425      10.2153     0.28199
                   19     0.450      10.3897     0.28641
                   20     0.475      10.5651     0.29124
                   21     0.500      10.7422     0.29651
                   22     0.525      10.9218     0.30225
                   23     0.550      11.1046     0.30849
                   24     0.575      11.2917     0.31529
                   25     0.600      11.4838     0.32269
                   26     0.625      11.6822     0.33076
                   27     0.650      11.8880     0.33958
                   28     0.675      12.1029     0.34924
                   29     0.700      12.3285     0.35988
                   30     0.725      12.5672     0.37165
                   31     0.750      12.8218     0.38477
                   32     0.775      13.0961     0.39951
                   33     0.800      13.3950     0.41626
                   34     0.825      13.7255     0.43557
                   35     0.850      14.0977     0.45823
                   36     0.875      14.5273     0.48549
                   37     0.900      15.0403     0.51948
                   38     0.925      15.6854     0.56423
                   39     0.950      16.5723     0.62905
                   40     0.975      18.0513     0.74436
                   41     0.990      19.9806     0.90620
```

Figure 7.14. BRR standard errors for iron usual intake percentiles

VIII: *SIDE* Warning Codes

Table 8.1 gives a listing of the warning codes reported by *SIDE*. Warnings are printed on the SAS© log, and begin with the characters *** SIDE WARNING # , folllowed by the error code. Table 8.1 lists each error code and its most common cause. The error codes are not numbered consecutively, but do appear in the table ordered from smallest to largest. In practice, the program trace information printed on the SAS© log will help pinpoint the source of the problem. Errors in setting up the description data sets are the most common.

Code	Reason for Error
1	Either DESCLIB.DESC or DESCLIB.ANAYVAR is missing
2	Unknown keyword(s) found in DESCLIB.DESC
3	No known keywords found in DESCLIB.DESC
4	No variables found in DESCLIB.DESC
5	Invalid data found for keyword
6	Input data set has not been specified
7	Analysis data set not in INLIB
8	No observations found in supplementary description data set
9	Missing values found in supplementary description data set
10	Invalid data found for NAME
11	Invalid data found for LABEL
12	Variable NAME not found in data set

Table 8.1. SIDE warning codes

Code	Reason for Error
13	Number of rootnames does not match number of analysis variables
14	Invalid data for rootname. Contains more than 4 characters
15	Number of weight variables does not match number of analysis variables
16	Expected number of variables not found in supplementary data set
17	No valid values found in supplementary data set
18	Invalid value found for cutoffs
20	Too many correlations per variable found in FXDCORRS
21	Number of observations in data set does not match number of analysis variables
22	Missing or invalid value found for correlation
25	Unsupported value found for **ADALPHA**
29	Too few replicate observations for usual intake estimation
30	Must have at least 1 observation per individual for smoothing
35	FXDCORRS data set found, but **CORRDAYS** = "N"
37	Unreasonable value found for **MAXJP**
38	Unreasonable value found for **MAXROOT**
39	Invalid value found for **LINFRAC**
40	Invalid value found for **MEANFRAC**
41	Unreasonable value found for **NPTS**
42	Requested data sets cannot be created; no analysis requested
44	Requested diagnostics cannot be run; no analysis requested
52	No observations found in input data set
53	All expected variables not found in input data set
54	Number of replicates does not evenly divide number of observations
55	Character-valued variable found in input data set
56	Multiple by variables specified
57	Variable name used in more than one supplementary description data set
59	User-specified value of **LINFRAC** to small
61	Negative derivatives detected in grafted polynomial fit
62	Too few correlations per variable found in FXDCORRS
63	Maximum number of join points did not produce normality
64	Not enough replicate observations to estimate 4th moment of measurement error
65	Negative estimate for usual intake variance

Table 8.1. SIDE warning codes (cont.)

APPENDIX : Correlation Estimates

Estimated day-to-day correlations based upon measurement error model. (See Carriquiry, et al. (1995).)

Component	Males ≥ 20	Females ≥ 20	Persons 0-19
Calcium	0.119	0.131	0.119
Carbohydrates	0.138	0.178	0.164
Carotene	0.097	0.078	0.045
Cholesterol	0.069	0.005	0.027
Copper	0.126	0.150	0.106
Energy	0.126	0.150	0.120
Fiber	0.133	0.167	0.092
Folate	0.111	0.110	0.100
Iron	0.108	0.105	0.020
Magnesium	0.146	0.199	0.129
Monounsaturated fat	0.092	0.064	0.066
Niacin	0.106	0.100	-0.011
Polyunsaturated fat	0.093	0.065	0.045
Phosphorus	0.127	0.153	0.068
Potassium	0.137	0.178	0.122
Protein	0.106	0.100	-0.004
Riboflavin	0.119	0.132	0.100
Saturated fat	0.107	0.100	0.103
Sodium	0.098	0.078	0.026
Total fat	0.097	0.076	0.076
Thiamin	0.117	0.130	0.001

Component	Males ≥20	Females ≥ 20	Persons 0-19
Vitamin B6	0.124	0.144	0.077
Vitamin B12	0.118	0.127	0.104
Vitamin A	0.090	0.058	0.062
Vitamin C	0.109	0.108	0.098
Vitamin E	0.106	0.099	0.070
Water	0.141	0.187	0.150
Zinc	0.103	0.093	-0.006
Percent tfat	0.081	0.036	0.035
Percent mfat	0.072	0.014	0.007
Percent sfat	0.105	0.096	0.131
Percent pfat	0.087	0.052	0.034
Percent carb	0.086	0.048	0.068
Percent prot	0.096	0.076	0.017

tfat	= total fat
mfat	= monounsaturated fat
sfat	= saturated fat
pfat	= polyunsaturated fat
carb	= calories from carbohydrates
prot	= calories from protein

References

Carriquiry, A.L. W.A. Fuller, J.J. Goyeneche, and H.H. Jensen (1995). *Estimated correlations among days for the combined 1989-91 CSFII*, Dietary Assessment Research Series 4. CARD Staff Report 95-SR 77. Center for Agricultural and Rural Development, Iowa State University, Ames.

Nusser, S. M., A.L. Carriquiry, K.W. Dodd, and W.A. Fuller, (1996). "A semiparametric transformation approach to estimating usual daily intake distributions," accepted with revisions in *Journal of the American Statistical Association.* An earlier version is available in Dietary Assessment Research Series Report 2, CARD Staff Report 95-SR74. Center for Agricultural and Rural Development, Iowa State University, Ames, Iowa.

SAS© Institute (1990). *SAS©/IML Software: Usage and Reference, Version 6, First Edition.* SAS© Institute Inc.: Cary, NC.

_____________________. (1990*). SAS© Language: Reference, Version 6, First Edition.* SAS© Institute Inc.: Cary, NC:

_____________________. (1990). *SAS©/GRAPH Software: Reference, Version 6, First Edition.* Vol. 1 and 2. SAS© Institute Inc.: Cary, NC:

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69:730-737.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.