

# **SAS/STAT 15.2<sup>®</sup> User's Guide Sashelp Data Sets**

This document is an individual chapter from *SAS/STAT® 15.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2020. *SAS/STAT® 15.2 User's Guide*. Cary, NC: SAS Institute Inc.

### **SAS/STAT® 15.2 User's Guide**

Copyright © 2020, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government License Rights; Restricted Rights:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2020

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

# Appendix B

## Sashelp Data Sets

### Contents

---

|                                                         |       |
|---------------------------------------------------------|-------|
| Overview of Sashelp Data Sets . . . . .                 | 10870 |
| Baseball Data . . . . .                                 | 10871 |
| Tropical Rain Forest Tree Data . . . . .                | 10873 |
| 2003 Birth Weight Data Set . . . . .                    | 10875 |
| Body Mass Index for Men . . . . .                       | 10876 |
| Isopod Burrows Data . . . . .                           | 10877 |
| Bone Marrow Transplant Data . . . . .                   | 10879 |
| 1997 Birth Weight Data . . . . .                        | 10880 |
| Class Data . . . . .                                    | 10881 |
| Comet Data . . . . .                                    | 10882 |
| El Niño–Southern Oscillation Data . . . . .             | 10883 |
| Finland’s Lake Laengelmaevesi Fish Catch Data . . . . . | 10884 |
| Exhaust Emissions Data . . . . .                        | 10885 |
| Fisher (1936) Iris Data . . . . .                       | 10886 |
| Junk Email Data . . . . .                               | 10888 |
| Leukemia Data Sets . . . . .                            | 10890 |
| Margarine Data . . . . .                                | 10892 |
| Flying Mileages between 10 US Cities Data . . . . .     | 10893 |
| Earthquake Locations in the United States . . . . .     | 10895 |
| Hot-Spring Locations in the United States . . . . .     | 10896 |
| Coal Seam Thickness Data . . . . .                      | 10897 |
| 1980 US Presidential Election Data . . . . .            | 10898 |
| References . . . . .                                    | 10898 |

---

## Overview of Sashelp Data Sets

SAS provides more than 200 data sets in the Sashelp library. These data sets are available for you to use for examples and for testing code. For example, the following step uses the Sashelp.Class data set:

```
proc reg data=sashelp.Class;
    model weight = height;
quit;
```

You do not need to provide a DATA step to use Sashelp data sets.

The following steps list all the data sets that are available in Sashelp:

```
ods select none;
proc contents data=sashelp._all_;
    ods output members=m;
run;
ods select all;

proc print;
    where memtype = 'DATA';
run;
```

The results of these steps (more than 200 data set names) are not displayed.

The following steps provide detailed information about the Sashelp data sets:

```
proc contents data=sashelp._all_;
run;
```

The results of this step (hundreds of pages of PROC CONTENTS output) are not displayed.

Twenty-two Sashelp data sets are used in SAS/STAT documentation, and the following sections describe these data sets:

|                  |                                                                              |
|------------------|------------------------------------------------------------------------------|
| Sashelp.Baseball | <a href="#">“Baseball Data”</a> on page 10871                                |
| Sashelp.BE1      | <a href="#">“Tropical Rain Forest Tree Data”</a> on page 10873               |
| Sashelp.BirthWgt | <a href="#">“2003 Birth Weight Data Set”</a> on page 10875                   |
| Sashelp.BMIMen   | <a href="#">“Body Mass Index for Men”</a> on page 10876                      |
| Sashelp.BMT      | <a href="#">“Bone Marrow Transplant Data”</a> on page 10879                  |
| Sashelp.Burrows  | <a href="#">“Isopod Burrows Data”</a> on page 10877                          |
| Sashelp.BWeight  | <a href="#">“1997 Birth Weight Data”</a> on page 10880                       |
| Sashelp.Class    | <a href="#">“Class Data”</a> on page 10881                                   |
| Sashelp.Comet    | <a href="#">“Comet Data”</a> on page 10882                                   |
| Sashelp.ENS0     | <a href="#">“El Niño–Southern Oscillation Data”</a> on page 10883            |
| Sashelp.Fish     | <a href="#">“Finland’s Lake Laengelmävesi Fish Catch Data”</a> on page 10884 |
| Sashelp.Gas      | <a href="#">“Exhaust Emissions Data”</a> on page 10885                       |
| Sashelp.Iris     | <a href="#">“Fisher (1936) Iris Data”</a> on page 10886                      |
| Sashelp.JunkMail | <a href="#">“Junk Email Data”</a> on page 10888                              |
| Sashelp.LeuTest  | <a href="#">“Leukemia Data Sets”</a> on page 10890                           |
| Sashelp.LeuTrain | <a href="#">“Leukemia Data Sets”</a> on page 10890                           |
| Sashelp.Margarin | <a href="#">“Margarine Data”</a> on page 10892                               |

|                  |                                                           |
|------------------|-----------------------------------------------------------|
| Sashelp.Mileages | “Flying Mileages between 10 US Cities Data” on page 10893 |
| Sashelp.Quakes   | “Earthquake Locations in the United States” on page 10895 |
| Sashelp.Springs  | “Hot-Spring Locations in the United States” on page 10896 |
| Sashelp.Thick    | “Coal Seam Thickness Data” on page 10897                  |
| Sashelp.Vote1980 | “1980 US Presidential Election Data” on page 10898        |

---

## Baseball Data

The Sashelp.Baseball data set contains salary and performance information for Major League Baseball players (excluding pitchers) who played at least one game in both the 1986 and 1987 seasons (Time Inc. 1987). The salaries are for the 1987 season, and the performance measures are from the 1986 season. The following steps display information about the Sashelp.Baseball data set and create [Figure B.1](#):

```
title 'Baseball Data';
proc contents data=sashelp.Baseball varnum;
    ods select position;
run;

title 'The First Five Observations Out of 322';
proc print data=sashelp.Baseball(obs=5);
run;
```

**Figure B.1** Baseball Data**Baseball Data**

| Variables in Creation Order |           |      |                               |
|-----------------------------|-----------|------|-------------------------------|
| #                           | Variable  | Type | Len Label                     |
| 1                           | Name      | Char | 18 Player's Name              |
| 2                           | Team      | Char | 14 Team at the End of 1986    |
| 3                           | nAtBat    | Num  | 8 Times at Bat in 1986        |
| 4                           | nHits     | Num  | 8 Hits in 1986                |
| 5                           | nHome     | Num  | 8 Home Runs in 1986           |
| 6                           | nRuns     | Num  | 8 Runs in 1986                |
| 7                           | nRBI      | Num  | 8 RBIs in 1986                |
| 8                           | nBB       | Num  | 8 Walks in 1986               |
| 9                           | YrMajor   | Num  | 8 Years in the Major Leagues  |
| 10                          | CrAtBat   | Num  | 8 Career Times at Bat         |
| 11                          | CrHits    | Num  | 8 Career Hits                 |
| 12                          | CrHome    | Num  | 8 Career Home Runs            |
| 13                          | CrRuns    | Num  | 8 Career Runs                 |
| 14                          | CrRbi     | Num  | 8 Career RBIs                 |
| 15                          | CrBB      | Num  | 8 Career Walks                |
| 16                          | League    | Char | 8 League at the End of 1986   |
| 17                          | Division  | Char | 8 Division at the End of 1986 |
| 18                          | Position  | Char | 8 Position(s) in 1986         |
| 19                          | nOuts     | Num  | 8 Put Outs in 1986            |
| 20                          | nAssts    | Num  | 8 Assists in 1986             |
| 21                          | nError    | Num  | 8 Errors in 1986              |
| 22                          | Salary    | Num  | 8 1987 Salary in \$ Thousands |
| 23                          | Div       | Char | 16 League and Division        |
| 24                          | logSalary | Num  | 8 Log Salary                  |

**The First Five Observations Out of 322**

| Obs | Name              | Team      | nAtBat | nHits | nHome | nRuns | nRBI | nBB | YrMajor | CrAtBat | CrHits | CrHome |
|-----|-------------------|-----------|--------|-------|-------|-------|------|-----|---------|---------|--------|--------|
| 1   | Allanson, Andy    | Cleveland | 293    | 66    | 1     | 30    | 29   | 14  | 1       | 293     | 66     | 1      |
| 2   | Ashby, Alan       | Houston   | 315    | 81    | 7     | 24    | 38   | 39  | 14      | 3449    | 835    | 69     |
| 3   | Davis, Alan       | Seattle   | 479    | 130   | 18    | 66    | 72   | 76  | 3       | 1624    | 457    | 63     |
| 4   | Dawson, Andre     | Montreal  | 496    | 141   | 20    | 65    | 78   | 37  | 11      | 5628    | 1575   | 225    |
| 5   | Galarraga, Andres | Montreal  | 321    | 87    | 10    | 39    | 42   | 30  | 2       | 396     | 101    | 12     |

  

| Obs | CrRuns | CrRbi | CrBB | League   | Division | Position | nOuts | nAssts | nError | Salary | Div | logSalary |
|-----|--------|-------|------|----------|----------|----------|-------|--------|--------|--------|-----|-----------|
| 1   | 30     | 29    | 14   | American | East     | C        | 446   | 33     | 20     | .      | AE  | .         |
| 2   | 321    | 414   | 375  | National | West     | C        | 632   | 43     | 10     | 475.0  | NW  | 6.16331   |
| 3   | 224    | 266   | 263  | American | West     | 1B       | 880   | 82     | 14     | 480.0  | AW  | 6.17379   |
| 4   | 828    | 838   | 354  | National | East     | RF       | 200   | 11     | 3      | 500.0  | NE  | 6.21461   |
| 5   | 48     | 46    | 33   | National | East     | 1B       | 805   | 40     | 4      | 91.5   | NE  | 4.51634   |

**NOTE:** Multiple versions of the baseball data set, including several revisions, are available from other sources.

---

## Tropical Rain Forest Tree Data

The Sashelp.BEI data set contains the locations of 3,604 trees in a number of tropical rain forests (Condit 1998; Hubbell and Foster 1983; Condit, Hubbell, and Foster 1996). A study window of  $1,000 \times 500$  square kilometers is used. The data set also contains covariates, represented by the variables Gradient and Elevation, which were collected at over 20,301 locations on a regular grid across the study region. The variable Trees distinguishes the event observations in the data set. The following steps display information about the data set Sashelp.BEI and create [Figure B.2](#):

```
title 'BEI Data';
proc contents data=sashelp.bei varnum;
    ods select position;
run;

title 'The First Five Observations Out of 23,905';
proc print data=sashelp.bei (obs=5);
run;

title 'The Trees Variable';
proc freq data=sashelp.bei;
    tables Trees;
run;
```

### Figure B.2 Tropical Rain Forest Tree Data

## BEI Data

| Variables in Creation Order |             |      |     |
|-----------------------------|-------------|------|-----|
| #                           | Variable    | Type | Len |
| 1                           | X           | Num  | 8   |
| 2                           | Y           | Num  | 8   |
| 3                           | Elevation   | Num  | 8   |
| 4                           | Gradient    | Num  | 8   |
| 5                           | Aluminum    | Num  | 8   |
| 6                           | Boron       | Num  | 8   |
| 7                           | Calcium     | Num  | 8   |
| 8                           | Copper      | Num  | 8   |
| 9                           | Iron        | Num  | 8   |
| 10                          | Potassium   | Num  | 8   |
| 11                          | Magnesium   | Num  | 8   |
| 12                          | Manganese   | Num  | 8   |
| 13                          | Sodium      | Num  | 8   |
| 14                          | Phosphorus  | Num  | 8   |
| 15                          | Sulphur     | Num  | 8   |
| 16                          | Zinc        | Num  | 8   |
| 17                          | Ammonia     | Num  | 8   |
| 18                          | Nitrate     | Num  | 8   |
| 19                          | TotalNitro  | Num  | 8   |
| 20                          | MinAmmonia  | Num  | 8   |
| 21                          | MinNitrate  | Num  | 8   |
| 22                          | MinNitrogen | Num  | 8   |
| 23                          | pH          | Num  | 8   |
| 24                          | Trees       | Num  | 8   |

## The First Five Observations Out of 23,905

[illegible][illegible]

**Figure B.2** *continued***The Trees Variable**

| Trees | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-------|-----------|---------|----------------------|--------------------|
| 0     | 20601     | 85.11   | 20601                | 85.11              |
| 1     | 3604      | 14.89   | 24205                | 100.00             |

---

## 2003 Birth Weight Data Set

The Sashelp.BirthWgt data set contains 100,000 random observations about infant mortality in 2003 from the US National Center for Health Statistics. Each observation records infant death within one year of birth, birth weight, maternal smoking and drinking behavior, and other background characteristics of the mother. The following steps display information about the Sashelp.BirthWgt data set and create [Figure B.3](#):

```

title '2003 Birth Weight Data';
proc contents data=sashelp.BirthWgt varnum;
    ods select position;
run;

title 'The First Five Observations Out of 100,000';
proc print data=sashelp.BirthWgt (obs=5);
run;

```

**Figure B.3** 2003 Birth Weight Data**2003 Birth Weight Data**

| Variables in Creation Order |             |      |     |
|-----------------------------|-------------|------|-----|
| #                           | Variable    | Type | Len |
| 1                           | LowBirthWgt | Char | 3   |
| 2                           | Married     | Char | 3   |
| 3                           | AgeGroup    | Num  | 8   |
| 4                           | Race        | Char | 9   |
| 5                           | Drinking    | Char | 3   |
| 6                           | Death       | Char | 3   |
| 7                           | Smoking     | Char | 3   |
| 8                           | SomeCollege | Char | 3   |

**The First Five Observations Out of 100,000**

| Obs | LowBirthWgt | Married | AgeGroup | Race   | Drinking | Death | Smoking | SomeCollege |
|-----|-------------|---------|----------|--------|----------|-------|---------|-------------|
| 1   | No          | No      | 3        | Asian  | No       | No    | No      | Yes         |
| 2   | No          | No      | 2        | White  | No       | No    | No      | No          |
| 3   | Yes         | Yes     | 2        | Native | No       | Yes   | No      | No          |
| 4   | No          | No      | 2        | White  | No       | No    | No      | No          |
| 5   | No          | No      | 2        | White  | No       | No    | No      | Yes         |

## Body Mass Index for Men

The Sashelp.BMIMen data set was collected from the 1999–2000 and 2001–2002 surveys of men that are published by the National Center for Health Statistics. This data set contains the variables BMI and Age. Body mass index (BMI) is defined as the ratio of weight in kilograms to the square of height in meters. BMI is widely used to categorize individuals as overweight or underweight. The following steps display information about the data set Sashelp.BMIMen and create [Figure B.4](#):

```
title 'BMIMen Data';
proc contents data=sashelp.bmimen varnum;
    ods select position;
run;

title 'The First Five Observations Out of 3,264';
proc print data=sashelp.bmimen(obs=5);
run;
```

**Figure B.4** Body Mass Index for Men

### BMIMen Data

| Variables in Creation<br>Order |          |      |     |
|--------------------------------|----------|------|-----|
| #                              | Variable | Type | Len |
| 1                              | Age      | Num  | 8   |
| 2                              | BMI      | Num  | 8   |

### The First Five Observations Out of 3,264

| Obs | Age | BMI  |
|-----|-----|------|
| 1   | 2.0 | 18.6 |
| 2   | 2.0 | 17.1 |
| 3   | 2.0 | 19.0 |
| 4   | 2.0 | 16.8 |
| 5   | 2.1 | 19.0 |

---

## Isopod Burrows Data

The Sashelp.Burrows data set contains the locations of 2,089 burrows, which are the habitat of isopods (“pill bugs”) in the Negev desert in Israel (Banerjee and Gelfand 2002). Each burrow has a Status variable that records whether a burrow still contains live isopods, has been emptied by an isopod predator, or has failed for some other environmental reason. The data set also contains covariates, which are measured on a square-meter grid over the study region. These covariates include topographical characteristics such as Elevation, Slope, Aspect, and PercentRock (the percentage of rocks). The data set also contains environmental characteristics such as PercentShrub (the percentage of shrubs) and Dew (the time in hours since midnight required to evaporate the dew). The following steps display information about the data set Sashelp.Burrows and create Figure B.5:

```
title 'Burrows Data';
proc contents data=sashelp.burrows varnum;
    ods select position;
run;

title 'The First Five Observations Out of 24,591';
proc print data=sashelp.burrows(obs=5);
run;

proc format;
    value bf -1 = "Killed by Predator" 0 = "Failed" 1 = "Successful";
run;

title 'The Status Variable';
proc freq data=sashelp.burrows;
    tables status;
    format status bf.;
run;
```

**Figure B.5** Isopod Burrows Data**Burrows Data**

| Variables in Creation Order |              |      |     |
|-----------------------------|--------------|------|-----|
| #                           | Variable     | Type | Len |
| 1                           | ID           | Num  | 8   |
| 2                           | X            | Num  | 8   |
| 3                           | Y            | Num  | 8   |
| 4                           | Z            | Num  | 8   |
| 5                           | Status       | Num  | 8   |
| 6                           | Watershed    | Num  | 8   |
| 7                           | BurrowCount  | Num  | 8   |
| 8                           | Elevation    | Num  | 8   |
| 9                           | Aspect       | Num  | 8   |
| 10                          | Slope        | Num  | 8   |
| 11                          | Dew          | Num  | 8   |
| 12                          | PercentRock  | Num  | 8   |
| 13                          | PercentShrub | Num  | 8   |
| 14                          | Burrow       | Num  | 8   |

**The First Five Observations Out of 24,591**

| Obs | ID  | X         | Y          | Z      | Status | Watershed | BurrowCount | Elevation | Aspect |
|-----|-----|-----------|------------|--------|--------|-----------|-------------|-----------|--------|
| 1   | 448 | 128324.66 | 1031421.00 | 531.75 | 1      | 5         | .           | .         | .      |
| 2   | 679 | 128344.23 | 1031391.63 | 530.79 | 0      | 5         | .           | .         | .      |
| 3   | 680 | 128344.77 | 1031391.31 | 530.86 | 1      | 5         | .           | .         | .      |
| 4   | 683 | 128343.09 | 1031402.81 | 531.24 | 0      | 5         | .           | .         | .      |
| 5   | 686 | 128343.84 | 1031415.25 | 532.11 | -1     | 5         | .           | .         | .      |

| Obs | Slope | Dew | PercentRock | PercentShrub | Burrow |
|-----|-------|-----|-------------|--------------|--------|
| 1   | .     | .   | .           | .            | 1      |
| 2   | .     | .   | .           | .            | 1      |
| 3   | .     | .   | .           | .            | 1      |
| 4   | .     | .   | .           | .            | 1      |
| 5   | .     | .   | .           | .            | 1      |

**The Status Variable**

| Status                    | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---------------------------|-----------|---------|----------------------|--------------------|
| Killed by Predator        | 88        | 4.21    | 88                   | 4.21               |
| Failed                    | 1443      | 69.04   | 1531                 | 73.25              |
| Successful                | 559       | 26.75   | 2090                 | 100.00             |
| Frequency Missing = 22501 |           |         |                      |                    |

## Bone Marrow Transplant Data

The Sashelp.BMT (bone marrow transplant) data set is used to illustrate survival analysis methods (Klein and Moeschberger 1997). At the time of transplant, each patient is classified into one of three risk categories: ALL (acute lymphoblastic leukemia), AML-Low Risk (acute myelocytic leukemia, low risk), and AML-High Risk. The endpoint of interest is the disease-free survival time, which is the time in days to death, relapse, or the end of the study. In this data set, the variable Group represents the patient's risk category, the variable T represents the disease-free survival time, and the variable Status is the censoring indicator such that the value 1 indicates an event time and the value 0 indicates a censored time. The following steps display information about the Sashelp.BMT data set and create [Figure B.6](#):

```
title 'Bone Marrow Transplant Data';
proc contents data=sashelp.BMT varnum;
    ods select position;
run;

title 'The First Five Observations Out of 137';
proc print data=sashelp.BMT(obs=5);
run;

title 'The Risk Group Variable';
proc freq data=sashelp.BMT;
    tables group;
run;
```

**Figure B.6** Bone Marrow Transplant Data

### Bone Marrow Transplant Data

| Variables in Creation Order |          |      |     |                                     |
|-----------------------------|----------|------|-----|-------------------------------------|
| #                           | Variable | Type | Len | Label                               |
| 1                           | Group    | Char | 13  | Disease Group                       |
| 2                           | T        | Num  | 8   | Disease-Free Survival Time          |
| 3                           | Status   | Num  | 8   | Event Indicator: 1=Event 0=Censored |

### The First Five Observations Out of 137

| Obs | Group | T    | Status |
|-----|-------|------|--------|
| 1   | ALL   | 2081 | 0      |
| 2   | ALL   | 1602 | 0      |
| 3   | ALL   | 1496 | 0      |
| 4   | ALL   | 1462 | 0      |
| 5   | ALL   | 1433 | 0      |

**Figure B.6** *continued***The Risk Group Variable**

| Disease Group |           |         |                      |                    |
|---------------|-----------|---------|----------------------|--------------------|
| Group         | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| ALL           | 38        | 27.74   | 38                   | 27.74              |
| AML-High Risk | 45        | 32.85   | 83                   | 60.58              |
| AML-Low Risk  | 54        | 39.42   | 137                  | 100.00             |

---

## 1997 Birth Weight Data

The Sashelp.BWeight data set provides 1997 birth weight data from National Center for Health Statistics (Koenker and Hallock 2001; Abreveya 2001). The data record live, singleton births to mothers between the ages of 18 and 45 in the United States who were classified as black or white. The following steps display information about the Sashelp.BWeight data set and create [Figure B.7](#):

```

title '1997 Birth Weight Data';
proc contents data=sashelp.BWeight varnum;
    ods select position;
run;

title 'The First Five Observations Out of 50,000';
proc print data=sashelp.BWeight (obs=5);
run;

```

**Figure B.7** Birth Weight Data**1997 Birth Weight Data**

| Variables in Creation Order |            |      |                                  |
|-----------------------------|------------|------|----------------------------------|
| #                           | Variable   | Type | Len Label                        |
| 1                           | Weight     | Num  | 8 Infant Birth Weight            |
| 2                           | Black      | Num  | 8 Black Mother                   |
| 3                           | Married    | Num  | 8 Married Mother                 |
| 4                           | Boy        | Num  | 8 Baby Boy                       |
| 5                           | MomAge     | Num  | 8 Mother's Age                   |
| 6                           | MomSmoke   | Num  | 8 Smoking Mother                 |
| 7                           | CigsPerDay | Num  | 8 Cigarettes Per Day             |
| 8                           | MomWtGain  | Num  | 8 Mother's Pregnancy Weight Gain |
| 9                           | Visit      | Num  | 8 Prenatal Visit                 |
| 10                          | MomEdLevel | Num  | 8 Mother's Education Level       |

**Figure B.7** *continued***The First Five Observations Out of 50,000**

| Obs | Weight | Black | Married | Boy | MomAge | MomSmoke | CigsPerDay | MomWtGain | Visit | MomEdLevel |
|-----|--------|-------|---------|-----|--------|----------|------------|-----------|-------|------------|
| 1   | 4111   | 0     | 1       | 1   | -3     | 0        | 0          | -16       | 1     | 0          |
| 2   | 3997   | 0     | 1       | 0   | 1      | 0        | 0          | 2         | 3     | 2          |
| 3   | 3572   | 0     | 1       | 1   | 0      | 0        | 0          | -3        | 3     | 0          |
| 4   | 1956   | 0     | 1       | 1   | -1     | 0        | 0          | -5        | 3     | 2          |
| 5   | 3515   | 0     | 1       | 1   | -6     | 0        | 0          | -20       | 3     | 0          |

---

## Class Data

The Sashelp.Class data set provides information about a small fictitious class of students. Variables include Sex, Age, Height, and Weight. This data set is frequently used in SAS documentation to illustrate basic SAS coding. The following steps display information about the Sashelp.Class data set and create [Figure B.8](#):

```

title 'Class Data';
proc contents data=sashelp.Class varnum;
    ods select position;
run;

title 'The Full Data Set';
proc print data=sashelp.Class;
run;

```

**Figure B.8** Class Data**Class Data**

| Variables in Creation Order |          |      |     |
|-----------------------------|----------|------|-----|
| #                           | Variable | Type | Len |
| 1                           | Name     | Char | 8   |
| 2                           | Sex      | Char | 1   |
| 3                           | Age      | Num  | 8   |
| 4                           | Height   | Num  | 8   |
| 5                           | Weight   | Num  | 8   |

**Figure B.8** *continued***The Full Data Set**

| Obs | Name    | Sex | Age | Height | Weight |
|-----|---------|-----|-----|--------|--------|
| 1   | Alfred  | M   | 14  | 69.0   | 112.5  |
| 2   | Alice   | F   | 13  | 56.5   | 84.0   |
| 3   | Barbara | F   | 13  | 65.3   | 98.0   |
| 4   | Carol   | F   | 14  | 62.8   | 102.5  |
| 5   | Henry   | M   | 14  | 63.5   | 102.5  |
| 6   | James   | M   | 12  | 57.3   | 83.0   |
| 7   | Jane    | F   | 12  | 59.8   | 84.5   |
| 8   | Janet   | F   | 15  | 62.5   | 112.5  |
| 9   | Jeffrey | M   | 13  | 62.5   | 84.0   |
| 10  | John    | M   | 12  | 59.0   | 99.5   |
| 11  | Joyce   | F   | 11  | 51.3   | 50.5   |
| 12  | Judy    | F   | 14  | 64.3   | 90.0   |
| 13  | Louise  | F   | 12  | 56.3   | 77.0   |
| 14  | Mary    | F   | 15  | 66.5   | 112.0  |
| 15  | Philip  | M   | 16  | 72.0   | 150.0  |
| 16  | Robert  | M   | 12  | 64.8   | 128.0  |
| 17  | Ronald  | M   | 15  | 67.0   | 133.0  |
| 18  | Thomas  | M   | 11  | 57.5   | 85.0   |
| 19  | William | M   | 15  | 66.5   | 112.0  |

---

## Comet Data

The Sashelp.Comet data set provides information from the following experiment. Twenty-four male rats were divided into four groups. Three groups received a daily oral dose of a 1,2-dimethylhydrazine dihydrochloride in three dose levels (low, medium, and high, respectively); the fourth group was a control group. Three additional animals received a positive control. Cell suspensions for each animal were scored for DNA damage by using a comet assay (Ghebretinsae et al. 2013). The following steps display information about the Sashelp.Comet data set and create [Figure B.9](#):

```

title 'Comet Data';
proc contents data=sashelp.Comet varnum;
  ods select position;
run;

title 'The First Five Observations Out of 4,050';
proc print data=sashelp.Comet (obs=5);
run;
```

**Figure B.9** Comet Data**Comet Data**

| Variables in Creation Order |          |      |     |                                                  |
|-----------------------------|----------|------|-----|--------------------------------------------------|
| #                           | Variable | Type | Len | Label                                            |
| 1                           | Dose     | Num  | 8   | 1,2 Dimethylhydrazine dihydrochloride Dose Level |
| 2                           | Rat      | Num  | 8   | Rat Index                                        |
| 3                           | Sample   | Num  | 8   | Slide Index of Grouped Cells from a Rat          |
| 4                           | Length   | Num  | 8   | Tail Length of the Comet                         |

**The First Five Observations Out of 4,050**

| Obs | Dose | Rat | Sample | Length  |
|-----|------|-----|--------|---------|
| 1   | 0    | 1   | 1      | 15.3527 |
| 2   | 0    | 1   | 1      | 16.1826 |
| 3   | 0    | 1   | 1      | 14.9378 |
| 4   | 0    | 1   | 1      | 12.4481 |
| 5   | 0    | 1   | 1      | 12.8631 |

## El Niño–Southern Oscillation Data

The Sashelp.ENS0 (El Niño–Southern Oscillation) data set contains measurements of monthly averaged atmospheric pressure differences between Easter Island and Darwin, Australia, for a period of 168 months (National Institute of Standards and Technology 1998). These pressure differences drive the southern trade winds. This data set is used to illustrate fitting nonlinear functions to a scatter plot by using methods such as loess and penalized B-splines. These data show both seasonal variations and variations due to El Niño. The following steps display information about the Sashelp.ENS0 data set and create [Figure B.10](#):

```

title 'El Nino Southern Oscillation Data';
proc contents data=sashelp.ENS0 varnum;
    ods select position;
run;

title 'The First Five Observations Out of 168';
proc print data=sashelp.ENS0 (obs=5);
run;

```

**Figure B.10** El Niño–Southern Oscillation Data**El Nino Southern Oscillation Data**

| Variables in Creation Order |          |      |     |
|-----------------------------|----------|------|-----|
| #                           | Variable | Type | Len |
| 1                           | Month    | Num  | 8   |
| 2                           | Year     | Num  | 8   |
| 3                           | Pressure | Num  | 8   |

**Figure B.10** *continued***The First Five Observations Out of 168**

| Obs | Month | Year    | Pressure |
|-----|-------|---------|----------|
| 1   | 1     | 0.08333 | 12.9     |
| 2   | 2     | 0.16667 | 11.3     |
| 3   | 3     | 0.25000 | 10.6     |
| 4   | 4     | 0.33333 | 11.2     |
| 5   | 5     | 0.41667 | 10.9     |

---

## Finland's Lake Laengelmävesi Fish Catch Data

The Sashelp.Fish catch data set contains measurements of 159 fish that were caught in Finland's Lake Laengelmävesi (Puranen 1917); it is used to illustrate discriminant analysis. For each of the seven species (bream, roach, whitefish, parkki, perch, pike, and smelt), the weight, length, height, and width of each fish are tallied. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail, and from the nose to the end of its tail. The height and width are recorded as percentages of the third length variable. The following steps display information about the Sashelp.Fish data set and create [Figure B.11](#):

```

title 'Finland's Lake Laengelmävesi Fish Catch Data';
proc contents data=sashelp.Fish varnum;
    ods select position;
run;

title 'The First Five Observations Out of 159';
proc print data=sashelp.Fish(obs=5);
run;

title 'The Fish Species Variable';
proc freq data=sashelp.Fish;
    tables species;
run;

```

**Figure B.11** Finland's Lake Laengelmävesi Fish Catch Data

**Finland's Lake Laengelmävesi Fish Catch Data**

| Variables in Creation Order |          |      |     |
|-----------------------------|----------|------|-----|
| #                           | Variable | Type | Len |
| 1                           | Species  | Char | 9   |
| 2                           | Weight   | Num  | 8   |
| 3                           | Length1  | Num  | 8   |
| 4                           | Length2  | Num  | 8   |
| 5                           | Length3  | Num  | 8   |
| 6                           | Height   | Num  | 8   |
| 7                           | Width    | Num  | 8   |

**Figure B.11** *continued***The First Five Observations Out of 159**

| Obs | Species | Weight | Length1 | Length2 | Length3 | Height  | Width  |
|-----|---------|--------|---------|---------|---------|---------|--------|
| 1   | Bream   | 242    | 23.2    | 25.4    | 30.0    | 11.5200 | 4.0200 |
| 2   | Bream   | 290    | 24.0    | 26.3    | 31.2    | 12.4800 | 4.3056 |
| 3   | Bream   | 340    | 23.9    | 26.5    | 31.1    | 12.3778 | 4.6961 |
| 4   | Bream   | 363    | 26.3    | 29.0    | 33.5    | 12.7300 | 4.4555 |
| 5   | Bream   | 430    | 26.5    | 29.0    | 34.0    | 12.4440 | 5.1340 |

**The Fish Species Variable**

| Species   | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----------|-----------|---------|----------------------|--------------------|
| Bream     | 35        | 22.01   | 35                   | 22.01              |
| Parkki    | 11        | 6.92    | 46                   | 28.93              |
| Perch     | 56        | 35.22   | 102                  | 64.15              |
| Pike      | 17        | 10.69   | 119                  | 74.84              |
| Roach     | 20        | 12.58   | 139                  | 87.42              |
| Smelt     | 14        | 8.81    | 153                  | 96.23              |
| Whitefish | 6         | 3.77    | 159                  | 100.00             |

---

## Exhaust Emissions Data

The Sashelp.Gas data set contains data from an experiment about gasoline engine exhaust emissions (Brinkman 1981). Nitrogen oxide emissions from a single-cylinder engine are measured for various combinations of fuel, compression ratio, and equivalence ratio. This data set is used to illustrate how to fit models by using nonlinearly transformed data. The following steps display information about the Sashelp.Gas data set and create Figure B.12:

```

title 'Exhaust Emissions Data';
proc contents data=sashelp.Gas varnum;
    ods select position;
run;

title 'The First Five Observations Out of 171';
proc print data=sashelp.Gas (obs=5);
run;

title 'The Fuel Type Variable';
proc freq data=sashelp.Gas;
    tables fuel;
run;
```

**Figure B.12** Exhaust Emissions Data**Exhaust Emissions Data**

| Variables in Creation Order |          |      |     |                   |
|-----------------------------|----------|------|-----|-------------------|
| #                           | Variable | Type | Len | Label             |
| 1                           | Fuel     | Char | 8   |                   |
| 2                           | CpRatio  | Num  | 8   | Compression Ratio |
| 3                           | EqRatio  | Num  | 8   | Equivalence Ratio |
| 4                           | NOx      | Num  | 8   | Nitrogen Oxide    |

**The First Five Observations Out of 171**

| Obs | Fuel    | CpRatio | EqRatio | NOx   |
|-----|---------|---------|---------|-------|
| 1   | Ethanol | 12      | 0.907   | 3.741 |
| 2   | Ethanol | 12      | 0.761   | 2.295 |
| 3   | Ethanol | 12      | 1.108   | 1.498 |
| 4   | Ethanol | 12      | 1.016   | 2.881 |
| 5   | Ethanol | 12      | 1.189   | 0.760 |

**The Fuel Type Variable**

| Fuel     | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|----------|-----------|---------|----------------------|--------------------|
| 82rongs  | 9         | 5.26    | 9                    | 5.26               |
| 94%Eth   | 25        | 14.62   | 34                   | 19.88              |
| Ethanol  | 90        | 52.63   | 124                  | 72.51              |
| Gasohol  | 13        | 7.60    | 137                  | 80.12              |
| Indolene | 22        | 12.87   | 159                  | 92.98              |
| Methanol | 12        | 7.02    | 171                  | 100.00             |

## Fisher (1936) Iris Data

The Sashelp.Iris data set (Fisher 1936) is widely used for examples of discriminant analysis and cluster analysis. The data are measurements in millimeters of the sepal length, sepal width, petal length, and petal width of 50 iris specimens from each of three species: *Iris setosa*, *I. versicolor*, and *I. virginica*. The following steps display information about the Sashelp.Iris data set and create [Figure B.13](#):

```

title 'Fisher (1936) Iris Data';
proc contents data=sashelp.Iris varnum;
    ods select position;
run;

title 'The First Five Observations Out of 150';
proc print data=sashelp.Iris(obs=5);
run;

title 'The Iris Species Variable';
proc freq data=sashelp.Iris;
```

```
tables species;
run;
```

**Figure B.13** Fisher (1936) Iris Data**Fisher (1936) Iris Data**

| Variables in Creation Order |             |      |     |                   |
|-----------------------------|-------------|------|-----|-------------------|
| #                           | Variable    | Type | Len | Label             |
| 1                           | Species     | Char | 10  | Iris Species      |
| 2                           | SepalLength | Num  | 8   | Sepal Length (mm) |
| 3                           | SepalWidth  | Num  | 8   | Sepal Width (mm)  |
| 4                           | PetalLength | Num  | 8   | Petal Length (mm) |
| 5                           | PetalWidth  | Num  | 8   | Petal Width (mm)  |

**The First Five Observations Out of 150**

| Obs | Species | SepalLength | SepalWidth | PetalLength | PetalWidth |
|-----|---------|-------------|------------|-------------|------------|
| 1   | Setosa  | 50          | 33         | 14          | 2          |
| 2   | Setosa  | 46          | 34         | 14          | 3          |
| 3   | Setosa  | 46          | 36         | 10          | 2          |
| 4   | Setosa  | 51          | 33         | 17          | 5          |
| 5   | Setosa  | 55          | 35         | 13          | 2          |

**The Iris Species Variable**

| Iris Species |           |         |                      |                    |
|--------------|-----------|---------|----------------------|--------------------|
| Species      | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Setosa       | 50        | 33.33   | 50                   | 33.33              |
| Versicolor   | 50        | 33.33   | 100                  | 66.67              |
| Virginica    | 50        | 33.33   | 150                  | 100.00             |

## Junk Email Data

The Sashelp.JunkMail data set comes from a study that classifies whether an email is junk email (coded as 1) or not (coded as 0). The data were collected in Hewlett-Packard labs and donated by George Forman. The data set contains 4,601 observations with 59 variables. The response variable is a binary indicator of whether an email is considered spam or not. There are 57 predictor variables that record frequencies of some common words and characters and lengths of uninterrupted sequences of capital letters in emails. The following steps display information about the Sashelp.JunkMail data set and create [Figure B.14](#):

```
title 'Junk Email Data';
proc contents data=sashelp.JunkMail varnum;
    ods select position;
run;

title 'The First Five Observations Out of 4,601';
proc print data=sashelp.JunkMail(obs=5);
run;
```

**Figure B.14** Junk Email Data

| Variables in Creation Order |           |      |     |                        |
|-----------------------------|-----------|------|-----|------------------------|
| #                           | Variable  | Type | Len | Label                  |
| 1                           | Test      | Num  | 8   | 0 - Training, 1 - Test |
| 2                           | Make      | Num  | 8   |                        |
| 3                           | Address   | Num  | 8   |                        |
| 4                           | All       | Num  | 8   |                        |
| 5                           | _3D       | Num  | 8   | 3D                     |
| 6                           | Our       | Num  | 8   |                        |
| 7                           | Over      | Num  | 8   |                        |
| 8                           | Remove    | Num  | 8   |                        |
| 9                           | Internet  | Num  | 8   |                        |
| 10                          | Order     | Num  | 8   |                        |
| 11                          | Mail      | Num  | 8   |                        |
| 12                          | Receive   | Num  | 8   |                        |
| 13                          | Will      | Num  | 8   |                        |
| 14                          | People    | Num  | 8   |                        |
| 15                          | Report    | Num  | 8   |                        |
| 16                          | Addresses | Num  | 8   |                        |
| 17                          | Free      | Num  | 8   |                        |
| 18                          | Business  | Num  | 8   |                        |
| 19                          | Email     | Num  | 8   |                        |
| 20                          | You       | Num  | 8   |                        |
| 21                          | Credit    | Num  | 8   |                        |
| 22                          | Your      | Num  | 8   |                        |
| 23                          | Font      | Num  | 8   |                        |
| 24                          | _000      | Num  | 8   | 000                    |
| 25                          | Money     | Num  | 8   |                        |

**Figure B.14** *continued***Junk Email Data**

| Variables in Creation Order |             |      |                              |
|-----------------------------|-------------|------|------------------------------|
| #                           | Variable    | Type | Len Label                    |
| 26                          | HP          | Num  | 8                            |
| 27                          | HPL         | Num  | 8                            |
| 28                          | George      | Num  | 8                            |
| 29                          | _650        | Num  | 8 650                        |
| 30                          | Lab         | Num  | 8                            |
| 31                          | Labs        | Num  | 8                            |
| 32                          | Telnet      | Num  | 8                            |
| 33                          | _857        | Num  | 8 857                        |
| 34                          | Data        | Num  | 8                            |
| 35                          | _415        | Num  | 8 415                        |
| 36                          | _85         | Num  | 8 85                         |
| 37                          | Technology  | Num  | 8                            |
| 38                          | _1999       | Num  | 8 1999                       |
| 39                          | Parts       | Num  | 8                            |
| 40                          | PM          | Num  | 8                            |
| 41                          | Direct      | Num  | 8                            |
| 42                          | CS          | Num  | 8                            |
| 43                          | Meeting     | Num  | 8                            |
| 44                          | Original    | Num  | 8                            |
| 45                          | Project     | Num  | 8                            |
| 46                          | RE          | Num  | 8                            |
| 47                          | Edu         | Num  | 8                            |
| 48                          | Table       | Num  | 8                            |
| 49                          | Conference  | Num  | 8                            |
| 50                          | Semicolon   | Num  | 8                            |
| 51                          | Paren       | Num  | 8                            |
| 52                          | Bracket     | Num  | 8                            |
| 53                          | Exclamation | Num  | 8                            |
| 54                          | Dollar      | Num  | 8                            |
| 55                          | Pound       | Num  | 8                            |
| 56                          | CapAvg      | Num  | 8 Capital Run Length Average |
| 57                          | CapLong     | Num  | 8 Capital Run Length Longest |
| 58                          | CapTotal    | Num  | 8 Capital Run Length Total   |
| 59                          | Class       | Num  | 8 0 - Not Junk, 1 - Junk     |

Figure B.14 continued

## The First Five Observations Out of 4,601

| Obs | Test | Make | Address | All  | _3D | Our  | Over | Remove | Internet | Order | Mail | Receive | Will | People | Report |
|-----|------|------|---------|------|-----|------|------|--------|----------|-------|------|---------|------|--------|--------|
| 1   | 1    | 0.00 | 0.64    | 0.64 | 0   | 0.32 | 0.00 | 0.00   | 0.00     | 0.00  | 0.00 | 0.00    | 0.64 | 0.00   | 0.00   |
| 2   | 0    | 0.21 | 0.28    | 0.50 | 0   | 0.14 | 0.28 | 0.21   | 0.07     | 0.00  | 0.94 | 0.21    | 0.79 | 0.65   | 0.21   |
| 3   | 1    | 0.06 | 0.00    | 0.71 | 0   | 1.23 | 0.19 | 0.19   | 0.12     | 0.64  | 0.25 | 0.38    | 0.45 | 0.12   | 0.00   |
| 4   | 0    | 0.00 | 0.00    | 0.00 | 0   | 0.63 | 0.00 | 0.31   | 0.63     | 0.31  | 0.63 | 0.31    | 0.31 | 0.31   | 0.00   |
| 5   | 0    | 0.00 | 0.00    | 0.00 | 0   | 0.63 | 0.00 | 0.31   | 0.63     | 0.31  | 0.63 | 0.31    | 0.31 | 0.31   | 0.00   |

  

| Obs | Addresses | Free | Business | Email | You  | Credit | Your | Font | _000 | Money | HP | HPL | George | _650 | Lab | Labs |
|-----|-----------|------|----------|-------|------|--------|------|------|------|-------|----|-----|--------|------|-----|------|
| 1   | 0.00      | 0.32 | 0.00     | 1.29  | 1.93 | 0.00   | 0.96 | 0    | 0.00 | 0.00  | 0  | 0   | 0      | 0    | 0   | 0    |
| 2   | 0.14      | 0.14 | 0.07     | 0.28  | 3.47 | 0.00   | 1.59 | 0    | 0.43 | 0.43  | 0  | 0   | 0      | 0    | 0   | 0    |
| 3   | 1.75      | 0.06 | 0.06     | 1.03  | 1.36 | 0.32   | 0.51 | 0    | 1.16 | 0.06  | 0  | 0   | 0      | 0    | 0   | 0    |
| 4   | 0.00      | 0.31 | 0.00     | 0.00  | 3.18 | 0.00   | 0.31 | 0    | 0.00 | 0.00  | 0  | 0   | 0      | 0    | 0   | 0    |
| 5   | 0.00      | 0.31 | 0.00     | 0.00  | 3.18 | 0.00   | 0.31 | 0    | 0.00 | 0.00  | 0  | 0   | 0      | 0    | 0   | 0    |

  

| Obs | Telnet | _857 | Data | _415 | _85 | Technology | _1999 | Parts | PM | Direct | CS | Meeting | Original | Project | RE   | Edu  |
|-----|--------|------|------|------|-----|------------|-------|-------|----|--------|----|---------|----------|---------|------|------|
| 1   | 0      | 0    | 0    | 0    | 0   | 0          | 0.00  | 0     | 0  | 0.00   | 0  | 0       | 0.00     | 0       | 0.00 | 0.00 |
| 2   | 0      | 0    | 0    | 0    | 0   | 0          | 0.07  | 0     | 0  | 0.00   | 0  | 0       | 0.00     | 0       | 0.00 | 0.00 |
| 3   | 0      | 0    | 0    | 0    | 0   | 0          | 0.00  | 0     | 0  | 0.06   | 0  | 0       | 0.12     | 0       | 0.06 | 0.06 |
| 4   | 0      | 0    | 0    | 0    | 0   | 0          | 0.00  | 0     | 0  | 0.00   | 0  | 0       | 0.00     | 0       | 0.00 | 0.00 |
| 5   | 0      | 0    | 0    | 0    | 0   | 0          | 0.00  | 0     | 0  | 0.00   | 0  | 0       | 0.00     | 0       | 0.00 | 0.00 |

  

| Obs | Table | Conference | Semicolon | Paren | Bracket | Exclamation | Dollar | Pound | CapAvg | CapLong | CapTotal | Class |
|-----|-------|------------|-----------|-------|---------|-------------|--------|-------|--------|---------|----------|-------|
| 1   | 0     | 0          | 0.00      | 0.000 | 0       | 0.778       | 0.000  | 0.000 | 3.756  | 61      | 278      | 1     |
| 2   | 0     | 0          | 0.00      | 0.132 | 0       | 0.372       | 0.180  | 0.048 | 5.114  | 101     | 1028     | 1     |
| 3   | 0     | 0          | 0.01      | 0.143 | 0       | 0.276       | 0.184  | 0.010 | 9.821  | 485     | 2259     | 1     |
| 4   | 0     | 0          | 0.00      | 0.137 | 0       | 0.137       | 0.000  | 0.000 | 3.537  | 40      | 191      | 1     |
| 5   | 0     | 0          | 0.00      | 0.135 | 0       | 0.135       | 0.000  | 0.000 | 3.537  | 40      | 191      | 1     |

## Leukemia Data Sets

The Sashelp.LeuTrain and Sashelp.LeuTest data sets provide microarray data from (Golub et al. 1999; Zou and Hastie 2005). The Sashelp.LeuTrain data set consists of 7,129 genes and 38 training samples, and the Sashelp.LeuTest data set consists of the same 7,129 genes and 34 testing samples. Among the 38 training samples, 27 are type 1 leukemia (acute lymphoblastic leukemia, coded in the data as 1) and 11 are type 2 leukemia (acute myeloid leukemia, coded in the data as -1).

The following steps display information about Sashelp.LeuTrain data set and create Figure B.15:

```

title 'Leukemia Training Data';
proc contents data=sashelp.LeuTrain varnum;
    ods select position;
run;

title 'The First Five Observations and 11 Variables';
proc print data=sashelp.LeuTrain(obs=5);

```

```

var y x1-x10;
run;

title 'Leukemia Type Variable';
proc freq data=sashelp.LeuTrain;
  tables y;
run;

```

**Figure B.15** Leukemia Training Data**The First Five Observations and 11 Variables**

| Obs | y | x1       | x2       | x3       | x4       | x5       | x6       | x7       | x8       | x9       | x10      |
|-----|---|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1   | 1 | -1.46240 | -0.64514 | -0.83593 | -1.47040 | -0.91997 | -1.58430 | 0.71239  | -0.54229 | 1.05090  | 0.23649  |
| 2   | 1 | -0.66480 | 0.20615  | -0.36857 | 0.25822  | -0.47567 | -0.35497 | -1.11940 | -0.29251 | -0.37542 | -0.38760 |
| 3   | 1 | -0.20049 | 0.37994  | -2.38280 | 0.43960  | -1.22700 | -1.76220 | 0.10464  | -1.80750 | 0.49292  | -1.67000 |
| 4   | 1 | -0.25776 | 0.27994  | 1.83920  | -1.62950 | -1.28750 | -1.26510 | 0.76334  | -0.61645 | -0.31578 | -0.32193 |
| 5   | 1 | -0.56457 | -0.39588 | -0.98372 | -0.83741 | -0.41477 | 0.14834  | -0.03550 | -0.10022 | -0.75753 | 0.37068  |

**Leukemia Type Variable**

| y  | Frequency | Percent | Cumulative |         |
|----|-----------|---------|------------|---------|
|    |           |         | Frequency  | Percent |
| -1 | 11        | 28.95   | 11         | 28.95   |
| 1  | 27        | 71.05   | 38         | 100.00  |

The results of the PROC CONTENTS step are not displayed. The results show that there are 7,130 variables, y and x1-x7129.

The following steps display information about Sashelp.LeuTest data set and create [Figure B.16](#):

```

title 'Leukemia Test Data';
proc contents data=sashelp.LeuTest varnum;
  ods select position;
run;

title 'The First Five Observations and 11 Variables';
proc print data=sashelp.LeuTest(obs=5);
  var y x1-x10;
run;

title 'Leukemia Type Variable';
proc freq data=sashelp.LeuTest;
  tables y;
run;

```

**Figure B.16** Leukemia Test Data**The First Five Observations and 11 Variables**

| Obs | y | x1       | x2       | x3       | x4       | x5      | x6       | x7       | x8       | x9       | x10      |
|-----|---|----------|----------|----------|----------|---------|----------|----------|----------|----------|----------|
| 1   | 1 | -1.38240 | 0.06288  | 0.62252  | 1.61210  | 0.52179 | 0.11516  | -1.85270 | -0.39956 | 0.88007  | -0.86565 |
| 2   | 1 | 0.65192  | -0.35476 | 2.29630  | 1.64980  | 0.50211 | -0.37315 | 1.76820  | -1.74270 | 1.63080  | 0.60171  |
| 3   | 1 | 0.65409  | 1.41340  | 0.22593  | -0.06719 | 0.30015 | 0.76964  | -0.26212 | 0.94481  | -0.51884 | -0.60999 |
| 4   | 1 | 1.07220  | 0.01959  | 0.16875  | 0.84779  | 0.24533 | 0.79682  | 0.41442  | 0.35122  | -0.70177 | 1.85410  |
| 5   | 1 | 2.12480  | 1.66370  | -0.35986 | 1.15850  | 0.89379 | 0.56310  | -0.92476 | 0.56790  | -0.56039 | -2.12400 |

**Leukemia Type Variable**

| y  | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|----|-----------|---------|----------------------|--------------------|
| -1 | 14        | 41.18   | 14                   | 41.18              |
| 1  | 20        | 58.82   | 34                   | 100.00             |

The results of the PROC CONTENTS step are not displayed. The results show that there are 7,130 variables, y and x1–x7129.

---

## Margarine Data

The Sashelp.Margarin data set is a scanner panel data set that lists purchases of margarine (Rossi, Allenby, and McCulloch 2005). There are 313 households and a total of 3,405 purchases. The variable Household represents the household ID; each household made at least five purchases, which are defined by the choice set variable Set. The variable Choice represents the choice that households made among the six margarine brands for each purchase or choice set. The variable Brand has the value PPK for Parkay stick, PBB for Blue Bonnet stick, PFL for Fleischmann's stick, PHse for the house brand stick, PGen for the generic stick, and PSS for Shedd's Spread tub. The variable LogPrice is the logarithm of the product price. The variables LogInc and FamSize provide information about household income and family size, respectively. The following steps display information about the Sashelp.Margarin data set and create Figure B.17:

```

title 'Margarine Data';
proc contents data=sashelp.Margarin varnum;
    ods select position;
run;

title 'The First Six Observations Out of 20,430';
proc print data=sashelp.Margarin(obs=6);
run;

```

**Figure B.17** Margarine Data**Margarine Data**

| Variables in Creation Order |          |      |     |
|-----------------------------|----------|------|-----|
| #                           | Variable | Type | Len |
| 1                           | HouseID  | Num  | 8   |
| 2                           | Set      | Num  | 8   |
| 3                           | Choice   | Num  | 8   |
| 4                           | Brand    | Char | 8   |
| 5                           | LogPrice | Num  | 8   |
| 6                           | LogInc   | Num  | 8   |
| 7                           | FamSize  | Num  | 8   |

**The First Six Observations Out of 20,430**

| Obs | HouseID | Set | Choice | Brand | LogPrice | LogInc  | FamSize |
|-----|---------|-----|--------|-------|----------|---------|---------|
| 1   | 2100016 | 1   | 1      | PPk   | -0.41552 | 3.48124 | 2       |
| 2   | 2100016 | 1   | 0      | PBB   | -0.40048 | 3.48124 | 2       |
| 3   | 2100016 | 1   | 0      | PFI   | 0.08618  | 3.48124 | 2       |
| 4   | 2100016 | 1   | 0      | PHse  | -0.56212 | 3.48124 | 2       |
| 5   | 2100016 | 1   | 0      | PGen  | -1.02165 | 3.48124 | 2       |
| 6   | 2100016 | 1   | 0      | PSS   | -0.16252 | 3.48124 | 2       |

---

## Flying Mileages between 10 US Cities Data

The Sashelp.Mileages data set contains a table of flying mileages between 10 US cities. This data set is frequently used to illustrate cluster analysis and multidimensional scaling. The following steps display information about the Sashelp.Mileages data set and create [Figure B.18](#):

```

title 'Flying Mileages between 10 US Cities Data';
proc contents data=sashelp.Mileages varnum;
    ods select position;
run;

title 'The Full Data Set';
proc print data=sashelp.Mileages noobs;
    id city;
run;
```

**Figure B.18** Flying Mileages between 10 US Cities Data**Flying Mileages between 10 US Cities Data**

| Variables in Creation Order |              |      |     |
|-----------------------------|--------------|------|-----|
| #                           | Variable     | Type | Len |
| 1                           | Atlanta      | Num  | 8   |
| 2                           | Chicago      | Num  | 8   |
| 3                           | Denver       | Num  | 8   |
| 4                           | Houston      | Num  | 8   |
| 5                           | LosAngeles   | Num  | 8   |
| 6                           | Miami        | Num  | 8   |
| 7                           | NewYork      | Num  | 8   |
| 8                           | SanFrancisco | Num  | 8   |
| 9                           | Seattle      | Num  | 8   |
| 10                          | WashingtonDC | Num  | 8   |
| 11                          | City         | Char | 15  |

**The Full Data Set**

| City            | Atlanta | Chicago | Denver | Houston | LosAngeles | Miami | NewYork | SanFrancisco | Seattle | WashingtonDC |
|-----------------|---------|---------|--------|---------|------------|-------|---------|--------------|---------|--------------|
| Atlanta         | 0       | .       | .      | .       | .          | .     | .       | .            | .       | .            |
| Chicago         | 587     | 0       | .      | .       | .          | .     | .       | .            | .       | .            |
| Denver          | 1212    | 920     | 0      | .       | .          | .     | .       | .            | .       | .            |
| Houston         | 701     | 940     | 879    | 0       | .          | .     | .       | .            | .       | .            |
| Los Angeles     | 1936    | 1745    | 831    | 1374    | 0          | .     | .       | .            | .       | .            |
| Miami           | 604     | 1188    | 1726   | 968     | 2339       | 0     | .       | .            | .       | .            |
| New York        | 748     | 713     | 1631   | 1420    | 2451       | 1092  | 0       | .            | .       | .            |
| San Francisco   | 2139    | 1858    | 949    | 1645    | 347        | 2594  | 2571    | 0            | .       | .            |
| Seattle         | 2182    | 1737    | 1021   | 1891    | 959        | 2734  | 2408    | 678          | 0       | .            |
| Washington D.C. | 543     | 597     | 1494   | 1220    | 2300       | 923   | 205     | 2442         | 2329    | 0            |

## Earthquake Locations in the United States

The Sashelp.Quakes data set provides locations and measurements of earthquakes in the United States. These data are provided courtesy of the United States Geological Service. The following steps display information about the Sashelp.Quakes data set and create [Figure B.19](#):

```
title 'Earthquake Locations in the United States';
proc contents data=sashelp.Quakes varnum;
    ods select position;
run;

title 'The First Five Observations Out of 15,578';
proc print data=sashelp.Quakes(obs=5);
run;
```

**Figure B.19** Earthquake Data

### Earthquake Locations in the United States

| Variables in Creation Order |                    |      |     |
|-----------------------------|--------------------|------|-----|
| #                           | Variable           | Type | Len |
| 1                           | Latitude           | Num  | 8   |
| 2                           | Longitude          | Num  | 8   |
| 3                           | Depth              | Num  | 8   |
| 4                           | Magnitude          | Num  | 8   |
| 5                           | dNearestStation    | Num  | 8   |
| 6                           | RootMeanSquareTime | Num  | 8   |
| 7                           | Type               | Char | 10  |

### The First Five Observations Out of 15,578

| Obs | Latitude | Longitude | Depth   | Magnitude | dNearestStation | RootMeanSquareTime | Type       |
|-----|----------|-----------|---------|-----------|-----------------|--------------------|------------|
| 1   | 35.6880  | -121.129  | 6.2500  | 2.75      | 0.02121         | 0.0700             | earthquake |
| 2   | 41.9009  | -119.622  | 0.0000  | 4.59      | 0.50700         | 0.0742             | earthquake |
| 3   | 41.8836  | -119.641  | 0.5329  | 3.19      | 0.49300         | 0.0000             | earthquake |
| 4   | 41.8897  | -119.640  | 0.5077  | 3.12      | 0.49400         | 0.0119             | earthquake |
| 5   | 40.8783  | -123.270  | 33.5700 | 2.60      | 0.22160         | 0.0600             | earthquake |

## Hot-Spring Locations in the United States

The Sashelp.Springs data set provides locations and measurements of hot springs in the United States. These data are provided courtesy of the National Oceanic and Atmospheric Administration. The following steps display information about the Sashelp.Springs data set and create [Figure B.20](#):

```
title 'Hot Spring Locations in the United States';
proc contents data=sashelp.Springs varnum;
    ods select position;
run;

title 'The First Five Observations Out of 1,587';
proc print data=sashelp.Springs (obs=5);
run;
```

**Figure B.20** Hot Spring Data

### Hot Spring Locations in the United States

| Variables in Creation<br>Order |            |      |     |
|--------------------------------|------------|------|-----|
| #                              | Variable   | Type | Len |
| 1                              | Latitude   | Num  | 8   |
| 2                              | Longitude  | Num  | 8   |
| 3                              | Name       | Char | 34  |
| 4                              | Area       | Char | 14  |
| 5                              | Type       | Char | 9   |
| 6                              | Fahrenheit | Num  | 8   |
| 7                              | Celsius    | Num  | 8   |

### The First Five Observations Out of 1,587

| Obs | Latitude | Longitude | Name                              | Area       | Type      | Fahrenheit | Celsius |
|-----|----------|-----------|-----------------------------------|------------|-----------|------------|---------|
| 1   | 51.925   | -177.160  | Fumaroles on Kanaga Island        | Adak       | Hotspring | 219        | 104     |
| 2   | 65.858   | -164.710  | Serpentine Hot Springs            | Bendeleben | Hotspring | 171        | 77      |
| 3   | 53.851   | -166.918  | Hot Springs Near Makushin Volcano | Unalaska   | Hotspring | H          | H       |
| 4   | 53.877   | -166.448  | Hot Springs Near Summer Bay       | Unalaska   | Hotspring | H          | H       |
| 5   | 52.270   | -174.042  | Hot Springs on Atka Island        | Atka       | Hotspring | H          | H       |

## Coal Seam Thickness Data

The Sashelp.Thick data set simulates measurements of coal seam thickness (in feet) taken over an approximately square area. The variable Thick contains the thickness values. The coordinates are offsets from a point in the southwest corner of the measurement area, where the unit for the north and east distances is 1,000 feet. The following steps display information about the Sashelp.Thick data set and create [Figure B.21](#):

```
title 'Coal Seam Thickness Data';
proc contents data=sashelp.Thick varnum;
    ods select position;
run;

title 'The First Five Observations Out of 75';
proc print data=sashelp.Thick (obs=5);
run;
```

**Figure B.21** Coal Seam Thickness Data

### Coal Seam Thickness Data

| Variables in Creation Order |          |      |                       |
|-----------------------------|----------|------|-----------------------|
| #                           | Variable | Type | Len Label             |
| 1                           | East     | Num  | 8                     |
| 2                           | North    | Num  | 8                     |
| 3                           | Thick    | Num  | 8 Coal Seam Thickness |

### The First Five Observations Out of 75

| Obs | East | North | Thick |
|-----|------|-------|-------|
| 1   | 0.7  | 59.6  | 34.1  |
| 2   | 2.1  | 82.7  | 42.2  |
| 3   | 4.7  | 75.1  | 39.5  |
| 4   | 4.8  | 52.8  | 34.3  |
| 5   | 5.9  | 67.1  | 37.0  |

## 1980 US Presidential Election Data

The Sashelp.Vote1980 data set contains US county votes-cast proportions and demographic and geographic characteristics for 3,107 US counties in the 1980 presidential election (Pace and Barry 1997). The six explanatory variables are as follows: the population 18 years of age or older (Pop), the population with 12th-grade or higher education (Edu), the number of owned housing units (Houses), the aggregate income (Income), and scaled longitude and latitude of geographic centroids (Longitude, Latitude). The dependent variable LogVoteRate is the logarithm of the proportion of votes cast divided by the variable Pop. The following steps display information about the data set Sashelp.Vote1980 and create Figure B.22:

```
title 'US 1980 Presidential Election Data';
proc contents data=sashelp.vote1980 varnum;
    ods select position;
run;

title 'The First Five Observations Out of 3,107';
proc print data=sashelp.vote1980 (obs=5);
run;
```

**Figure B.22** US 1980 Presidential Election Data

### US 1980 Presidential Election Data

| Variables in Creation Order |             |      |                                         |
|-----------------------------|-------------|------|-----------------------------------------|
| #                           | Variable    | Type | Len Label                               |
| 1                           | LogVoteRate | Num  | 8 Log Votes Cast per County             |
| 2                           | Pop         | Num  | 8 Population of 18 Years and Older      |
| 3                           | Edu         | Num  | 8 Population with 12th Grade and Higher |
| 4                           | Houses      | Num  | 8 Number of Owned Housing Units         |
| 5                           | Income      | Num  | 8 Aggregate Income                      |
| 6                           | Longitude   | Num  | 8 Scaled Longitude                      |
| 7                           | Latitude    | Num  | 8 Scaled Latitude                       |

### The First Five Observations Out of 3,107

| Obs | LogVoteRate | Pop     | Edu     | Houses  | Income  | Longitude | Latitude |
|-----|-------------|---------|---------|---------|---------|-----------|----------|
| 1   | -0.66156    | 9.9729  | 9.2463  | 9.00405 | 12.1349 | -0.86641  | 0.32542  |
| 2   | -0.65086    | 10.9033 | 10.2212 | 9.96576 | 13.0566 | -0.87755  | 0.30655  |
| 3   | -0.61711    | 9.7222  | 8.7535  | 8.70765 | 11.6306 | -0.85389  | 0.31863  |
| 4   | -0.63907    | 9.2737  | 8.1831  | 8.27741 | 11.2437 | -0.87127  | 0.32997  |
| 5   | -0.70027    | 10.1515 | 9.2077  | 9.24068 | 12.1551 | -0.86566  | 0.33980  |

## References

Abreveya, J. (2001). “The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes.” *Journal of Economics* 26:247–257.

- Banerjee, S., and Gelfand, A. E. (2002). “Prediction, Interpolation, and Regression for Spatially Misaligned Data.” *Sankhyā* 64:227–245.
- Brinkman, N. D. (1981). “Ethanol Fuel: A Single-Cylinder Engine Study of Efficiency and Exhaust Emissions.” *Society of Automotive Engineers Transactions* 90:1410–1424.
- Condit, R. (1998). *Tropical Forest Census Plots: Methods and Results from Barro Colorado Island, Panama, and a Comparison with Other Plots*. Berlin: Springer-Verlag.
- Condit, R., Hubbell, S. P., and Foster, R. B. (1996). “Changes in Tree Species Abundance in a Neotropical Forest: Impact of Climate Change.” *Journal of Tropical Ecology* 12:231–256.
- Fisher, R. A. (1936). “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7:179–188.
- Ghebretinsae, A. H., Faes, C., Molenberghs, G., De Boeck, M., and Geys, H. (2013). “A Bayesian, Generalized Frailty Model for Comet Assays.” *Journal of Biopharmaceutical Statistics* 23:618–636.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., et al. (1999). “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression.” *Science* 286:531–537.
- Hubbell, S. P., and Foster, R. B. (1983). “Diversity of Canopy Trees in a Neotropical Forest and Implications for the Conservation of Tropical Trees.” In *Tropical Rain Forest: Ecology and Management*, edited by S. J. Sutton, T. C. Whitmore, and A. C. Chadwick, 25–41. Oxford: Blackwell.
- Klein, J. P., and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.
- Koenker, R., and Hallock, K. (2001). “Quantile Regression: An Introduction.” *Journal of Economic Perspectives* 15:143–156.
- National Institute of Standards and Technology (1998). “Statistical Reference Data Sets.” Accessed June 6, 2011. <http://www.itl.nist.gov/div898/strd/general/dataarchive.html>.
- Pace, R. K., and Barry, R. (1997). “Quick Computation of Spatial Autoregressive Estimators.” *Geographical Analysis* 29:232–247.
- Puranen, J. (1917). “Fish Catch data set (1917).” Journal of Statistics Education Data Archive. Accessed October 29, 2019. <https://www.amstat.org/publications/jse/datasets/fishcatch.txt>.
- Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Chichester, UK: John Wiley & Sons.
- Time Inc. (1987). “What They Make.” *Sports Illustrated* (April 20): 54–81.
- Zou, H., and Hastie, T. (2005). “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society, Series B* 67:301–320.