APPENDIX **2**
Data Sets

INTRODUCTION

This appendix provides information about the data sets used as examples in this book. They are listed in the order in which they are discussed in the main text. All data are written in free format, with the variables appearing in the order given in each of the following sections. These data sets are available on the Web at support.sas.com/publishing/authors/allison.html.

THE MYEL DATA SET: MYELOMATOSIS PATIENTS

The MYEL data set contains survival times for 25 patients diagnosed with myelomatosis (Peto et al., 1977). The patients were randomly assigned to two drug treatments. The entire data set is listed in Output 2.1. These data are used extensively in Chapter 3 to illustrate the LIFETEST procedure and briefly in Chapter 5. The variables are as follows:

DUR | is the time in days from the point of randomization to either death or censoring (which could occur either by loss to follow up or termination of the observation).

STATUS | has a value of 1 if dead and a value of 0 if censored.

TREAT | specifies a value of 1 or 2 to correspond to the two treatments.

RENAL   has a value of 1 if renal functioning was normal at the time of randomization; it has a value of 0 for impaired functioning.

THE RECID DATA SET: ARREST TIMES FOR RELEASED PRISONERS

The RECID data set contains information about 432 inmates who were released from Maryland state prisons in the early 1970s. This data set is used in Chapters 3, 4, and 5. The first 20 cases are listed in Output 3.10. The data were kindly provided to me by Dr. Kenneth Lenihan who was one of the principal investigators. The aim of this research was to determine the efficacy of financial aid to released inmates as a means of reducing recidivism. Results from the study are described in Rossi, Berk, and Lenihan (1980). (This book also reports results from a much larger follow-up study done in Texas and Georgia.) Half the inmates were randomly assigned to receive financial aid (approximately the same amount as unemployment compensation). They were followed for 1 year after their release and were interviewed monthly during that period. Data on arrests were taken from police and court records. The data set used here contains the following variables:

WEEK   is the week of first arrest; WEEK has a value of 52 if not arrested.

ARREST   has a value of 1 if arrested; otherwise, ARREST has a value of 0.

FIN   has a value of 1 if the inmate received financial aid after release; otherwise, FIN has a value of 0. FIN is randomly assigned, with equal numbers in each category.

AGE   is the age in years at the time of release.

RACE   has a value of 1 if the inmate is black; otherwise, RACE has a value of 0.

WEXP   has a value of 1 if the inmate had full-time work experience before incarceration; otherwise, WEXP has a value of 0.

MAR   has a value of 1 if the inmate was married at the time of release; otherwise, MAR has a value of 0.

PARO   has a value of 1 if released on parole; otherwise, PARO has a value of 0.

PRIO   is the number of convictions before the current incarceration.

EDUC   is the highest level of completed schooling, coded as
2 = 6th grade or less
3 = 7th to 9th grade
4 = 10th to 11th grade
5 = 12th grade
6 = some college

EMP1–EMP52 represents the employment status in each of the first 52 weeks after release. These variables have values of 1 if the inmate was employed full time; otherwise, EMP1–EMP52 have values of 0. Data are missing for weeks after the first arrest.

THE STAN DATA SET: STANFORD HEART TRANSPLANT PATIENTS

Various versions of data from the Stanford Heart Transplant study have been reported in a number of publications. The data set used in Chapter 5, "Estimating Cox Regression Models with PROC PHREG," appeared in Crowley and Hu (1977). The sample consisted of 103 cardiac patients who were enrolled in the transplantation program between 1967 and 1974. After enrollment, patients waited varying lengths of time until a suitable donor heart was found. Patients were followed until death or until the termination date of April 1, 1974. Of the 69 transplant recipients, only 24 were still alive at termination. At the time of transplantation, all but four of the patients were tissue typed to determine the degree of similarity with the donor. The data set contains the following variables:

DOB   is the date of birth.
DOA   is the date of acceptance into the program.
DOT   is the date of transplant.
DLS   is the date last seen (death date or censoring date).
DEAD   has a value of 1 if the patient is dead at DLS; otherwise, DEAD has a value of 0.
SURG   has a value of 1 if the patient had open-heart surgery before DOA; otherwise, SURG has a value of 0.
M1   is the number of donor alleles with no match in the recipient (1 through 4).
M2   has a value of 1 if the donor and recipient mismatch on the HLA-A2 antigen; otherwise, M2 has a value of 0.
M3   is the mismatch score.

All date variables are in the *mm*/*dd*/*yy* format.