Appendix C

Data Sets

Data Set C.1 SENIC

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (**SENIC** Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed.

Each line of the data set has an identification number and provides information on 11 other variables for a single hospital. The data presented here are for the 1975–76 study period. The 12 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–113
2	Length of stay	Average length of stay of all patients in hospital (in days)
3	Age	Average age of patients (in years)
4	Infection risk	Average estimated probability of acquiring infection in hospital (in percent)
5	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
6	Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
7	Number of beds	Average number of beds in hospital during study period
8	Medical school affiliation	1 = Yes, 2 = No
9	Region	Geographic region, where: 1 = NE, 2 = NC, 3 = S, 4 = W
10	Average daily census	Average number of patients in hospital per day during study period
11	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full time plus one half the number part time)
12	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

Reference: Special Issue, "The SENIC Project," *American Journal of Epidemiology* 111 (1980), pp. 465–653. Data obtained from Robert W. Haley, M.D., Hospital Infections Program, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333.

1	2	3	4	5	6	7	8	9	10	11	12
1	7.13	55.7	4.1	9.0	39.6	279	2	4	207	241	60.0
2	8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40.0
3	8.34	56.9	2.7	8.1	74.0	107	2	3	82	54	20.0
...	...	...	...	...	...	...	...	...	...	...	...
111	7.70	56.9	4.4	12.2	67.9	129	2	4	85	136	62.9
112	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9
113	9.41	59.5	3.1	20.6	91.7	29	2	3	20	22	22.9

Data Set C.2 CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

Source: Geospatial and Statistical Data Center, University of Virginia.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Los_Angeles	CA	4060	8863164	32.1	9.7	23677	27700	688936	70.0	22.3	11.6	8.0	20786	184230	4
2	Cook	IL	946	5105067	29.2	12.4	15153	21550	436936	73.4	22.8	11.1	7.2	21729	110928	2
3	Harris	TX	1729	2818199	31.3	7.1	7553	12449	253526	74.9	25.4	12.5	5.7	19517	55003	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
438	Montgomery	TN	539	100498	35.7	7.9	87	188	6537	77.9	16.5	10.8	8.0	13169	1323	3
439	Maui	HI	1159	100374	26.2	11.3	192	182	7130	77.0	17.8	5.7	3.2	18504	1857	4
440	Morgan	AL	582	100043	26.3	11.7	122	464	4693	69.4	15.5	9.4	7.1	16458	1647	3

Data Set C.3 Market Share

Company executives from a large packaged foods manufacturer wished to determine which factors influence the market share of one of its products. Data were collected from a national database (Nielsen) for 36 consecutive months. Each line of the data set has an identification number and provides information on 6 other variables for each month. The data presented here are for September, 1999, through August, 2002. The variables are:

Variable Number	Variable Name	Description
1	Identification number	1–36
2	Market share	Average monthly market share for product (percent)
3	Price	Average monthly price of product (dollars)
4	Gross Nielsen rating points	An index of the amount of advertising exposure that the product received
5	Discount price	Presence or absence of discount price during period: 1 if discount, 0 otherwise
6	Package promotion	Presence or absence of package promotion during period: 1 if promotion present, 0 otherwise
7	Month	Month (Jan–Dec)
8	Year	Year (1999–2002)

1	2	3	4	5	6	7	8
1	3.15	2.198	498	1	1	Sep	1999
2	2.52	2.186	510	0	0	Oct	1999
3	2.64	2.293	422	1	1	Nov	1999
...	...	...	...	...	...	...	...
34	2.80	2.518	270	1	0	Jun	2002
35	2.48	2.497	322	0	1	Jul	2002
36	2.85	2.781	317	1	1	Aug	2002

Data Set C.4 University Admissions

The director of admissions at a state university wanted to determine how accurately students' grade-point averages at the end of their freshman year could be predicted by entrance test scores and high school class rank. The academic years cover 1996 through 2000. Each line of the data set has an identification number and information on 4 other variables for each student. The 5 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–705
2	GPA	Grade-point average following freshman year
3	High school class rank	High school class rank as percentile: lower percentiles imply higher class ranks
4	ACT score	ACT entrance examination score
5	Academic year	Calendar year that freshman entered university

1	2	3	4	5
1	0.980	61	20	1996
2	1.130	84	20	1996
3	1.250	74	19	1996
...	...	...	...	...
703	4.000	97	29	2000
704	4.000	97	29	2000
705	4.000	99	32	2000

Data Set C.5 Prostate Cancer

A university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. Each line of the data set has an identification number and provides information on 8 other variables for each person. The 9 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–97
2	PSA level	Serum prostate-specific antigen level (mg/ml)
3	Cancer volume	Estimate of prostate cancer volume (cc)
4	Weight	Prostate weight (gm)
5	Age	Age of patient (years)
6	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
7	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 otherwise
8	Capsular penetration	Degree of capsular penetration (cm)
9	Gleason score	Pathologically determined grade of disease using total score of two patterns (summed scores were either 6, 7, or 8 with higher scores indicating worse prognosis)

1	2	3	4	5	6	7	8	9
1	0.651	0.5599	15.959	50	0	0	0	6
2	0.852	0.3716	27.660	58	0	0	0	7
3	0.852	0.6005	14.732	74	0	0	0	7
...	...	...	...	...	...	...	...	...
95	170.716	18.3568	29.964	52	0	1	11.7048	8
96	239.847	17.8143	43.380	68	4.7588	1	4.7588	8
97	265.072	32.1367	52.985	68	1.5527	1	18.1741	8

Adapted in part from: Hastie, T. J.; R. J. Tibshirani; and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

Data Set C.6 Website Developer

Management of a company that develops websites was interested in determining which variables have the greatest impact on the number of websites developed and delivered to customers per quarter. Data were collected on website production output for 13 three-person website development teams, from January 2001 through August 2002. Each line of the data set has an identification number and provides information on 6 other variables for thirteen teams over time. The 8 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–73
2	Websites delivered	Number of websites completed and delivered to customers during the quarter
3	Backlog of orders	Number of website orders in backlog at the close of the quarter
4	Team number	1–13
5	Team experience	Number of months team has been together
6	Process change	A change in the website development process occurred during the second quarter of 2002: 1 if quarter 2 or 3, 2002; 0 otherwise
7	Year	2001 or 2002
8	Quarter	1, 2, 3, or 4

1	2	3	4	5	6	7	8
1	1	12	1	3	0	2001	1
2	2	18	1	6	0	2001	2
3	7	26	1	9	0	2001	3
...	...	...	...	...	...	...	...
71	7	36	13	14	0	2002	1
72	19	37	13	17	1	2002	2
73	12	26	13	20	1	2002	3

Data Set C.7 Real Estate Sales

The city tax assessor was interested in predicting residential home sales prices in a midwestern city as a function of various characteristics of the home and surrounding property. Data on 522 arms-length transactions were obtained for home sales during the year 2002. Each line of the data set has an identification number and provides information on 12 other variables. The 13 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–522
2	Sales price	Sales price of residence (dollars)
3	Finished square feet	Finished area of residence (square feet)
4	Number of bedrooms	Total number of bedrooms in residence
5	Number of bathrooms	Total number of bathrooms in residence
6	Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
7	Garage size	Number of cars that garage will hold
8	Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
9	Year built	Year property was originally constructed

10	Quality	Index for quality of construction:
		1 indicates high quality;
		2 indicates medium quality;
		3 indicates low quality

11	Style	Qualitative indicator of architectural style
12	Lot size	Lot size (square feet)
13	Adjacent to highway	Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

1	2	3	4	5	6	7	8	9	10	11	12	13
1	360000	3032	4	4	1	2	0	1972	2	1	22221	0
2	340000	2058	4	2	1	2	0	1976	2	1	22912	0
3	250000	1780	4	3	1	2	0	1980	2	1	21345	0
...	...	...	...	...	...	...	...	...	...	...	...	...
520	133500	1922	3	1	0	2	0	1950	3	1	14805	0
521	124000	1480	3	2	1	2	0	1953	3	1	28351	0
522	95500	1184	2	1	0	1	0	1951	3	1	14786	0

Data Set C.8 Heating Equipment

A manufacturer of heating equipment was interested in forecasting the volume of monthly orders as a function of various economic indicators, supply-chain factors, and weather in a particular sales region. Data by month over a four-year period (1999–2002) for this region were available for analysis. Each line of the data set has an identification number and provides information on 9 other variables. The 10 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–43
2	Number of orders	Number of heating equipment orders during month
3	Interest rate	Prime rate in effect during month
4	New homes	Number of new homes completed and for sale in sales region during month
5	Discount	Percent discount (0–5) offered to distributors during month; value is usually 0, indicating no discount
6	Inventories	Distributor inventories in warehouses during month
7	Sell through	Number of units sold by distributor to contractors in previous month
8	Temperature deviation	Difference between average temperature for month and 30-year average for that month
9	Year	1999, 2000, 2001, or 2002
10	Month	Coded 1–12

1	2	3	4	5	6	7	8	9	10
1	121	0.0750	64	0	3536	615	2.22	1999	1
2	227	0.0750	64	0	3042	813	0.28	1999	2
3	446	0.0750	65	0	2456	704	0.79	1999	3
...	...	...	...	...	...	...	...	...	...
41	754	0.0475	64	0	1417	927	0.81	2002	6
42	1098	0.0475	65	0	1244	877	0.28	2002	7
43	1158	0.0475	65	0	1465	809	0.50	2002	8

Data Set C.9 Ischemic Heart Disease

A health insurance company collected information on 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. Data were obtained on total costs of services provided for these 788 subscribers and the nature of the various services for the period of January 1, 1998 through December 31, 1999. Each line in the data set has an identification number and provides information on 9 other variables for each subscriber. The 10 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–788
2	Total cost	Total cost of claims by subscriber (dollars)
3	Age	Age of subscriber (years)
4	Gender	Gender of subscriber: 1 if male; 0 otherwise
5	Interventions	Total number of interventions or procedures carried out
6	Drugs	Number of tracked drugs prescribed
7	Emergency room visits	Number of emergency room visits
8	Complications	Number of other complications that arose during heart disease treatment
9	Comorbidities	Number of other diseases that the subscriber had during period
10	Duration	Number of days of duration of treatment condition

1	2	3	4	5	6	7	8	9	10
1	179.1	63	0	2	1	4	0	3	300
2	319.0	59	0	2	0	6	0	0	120
3	9310.7	62	0	17	0	2	0	5	353
...	...	...	...	...	...	...	...	...	...
786	2677.7	68	0	3	2	6	0	10	303
787	1282.2	58	0	7	2	2	0	7	244
788	586.0	56	0	4	4	6	0	3	336

Data Set C.10 Disease Outbreak

This data set provides information from a study based on 196 persons selected in a probability sample within two sectors in a city. Each line of the data set has an identification number and provides information on 5 other variables for a single person. The 6 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–196
2	Age	Age of person (in years)
3	Socioeconomic status	1 = upper, 2 = middle, 3 = lower
4	Sector	Sector within city, where: 1 = sector 1, 2 = sector 2
5	Disease status	1 = with disease, 0 = without disease
6	Savings account status	1 = has savings account, 0 = does not have savings account

Adapted in part from H. G. Dantes, J. S. Koopman, C. L. Addy, et al., "Dengue Epidemics on the Pacific Coast of Mexico," *International Journal of Epidemiology* 17 (1988), pp. 178–86.

1	2	3	4	5	6
1	33	1	1	0	1
2	35	1	1	0	1
3	6	1	1	0	0
...	...	...	...	...	...
194	31	3	1	0	0
195	85	3	1	0	1
196	24	2	1	0	0

Data Set C.11 IPO

Private companies often go public by issuing shares of stock referred to as initial public offerings (IPOs). A study of 482 IPOs was conducted to determine what are the characteristics of companies that attract venture capital funding. The response of interest is whether or not a company was financed with venture capital funds. Potential predictors include the face value of the company, the number of shares offered, and whether or not the company underwent a leveraged buyout. Each line of the data set has an identification number and provides information on 4 other variables for a single person. The 5 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–482
2	Venture capital funding	Presence or absence of venture capital funding: 1 if yes; 0 otherwise
3	Face value of company	Estimated face value of company from prospectus (in dollars)
4	Number of shares offered	Total number of shares offered
5	Leveraged buyout	Presence or absence of leveraged buyout: 1 if yes; 0 otherwise

1	2	3	4	5
1	0	1,200,000	3,000,000	0
2	0	1,454,000	1,454,000	1
3	0	1,500,000	300,000	0
...	...	...	...	...
480	0	159,500,000	7,250,000	0
481	0	165,000,000	11,000,000	0
482	0	234,600,000	9,200,000	0

Data Set C.12 Drug Effect Experiment

This data set provides results adapted from an experiment in which the effects of a drug on the behavior of rats were studied. The behavior under consideration was the rate at which a rat deprived of water presses a lever to obtain water. The experiment was carried out in two parts. Variable 2 identifies the two parts of the study (1, 2).

In Part I of the study, 12 male albino rats of the same strain and approximately the same weight were utilized. Variable 3 identifies each rat (1, . . . , 12). Prior to the experiment, each rat was trained to press a lever for water until a stable rate of pressing was reached. Two factors were studied in this experiment—initial lever press rate (factor A) and dosage of the drug (factor B). The 12 rats were classified into one of three groups according to their initial lever press rate. Variable 4 identifies the level of the initial lever press rate (1, 2, 3). Level 1 is a slow rate, level 2 a moderate rate, and level 3 a fast rate. The levels were defined such that one third of the rats were classified into each of the three levels.

Four dosage levels of the drug were studied, including a zero level consisting of a saline solution. Variable 5 identifies the drug dosage (1, . . . , 4). All dosage levels were specified in terms of milligrams of drug per kilogram of weight of the rat.

One hour after a drug dosage injection was administered, an experimental session began during which the rat received water each time after the second lever press. This reinforcement schedule will be denoted by FR-2. Each rat received all four drug dosage levels in a random order. Each of the four drug dosages was administered twice, thus providing two observation units for each treatment. Variable 6 identifies the observation unit (1, 2).

The response variable was defined as the total number of lever presses divided by the elapsed time (in seconds) during a session for the given treatment. Variable 7 is the response variable.

In Part II of the study, another 12 albino male rats of the same strain and approximately the same weight as the rats used in Part I were used. Variable 2 identifies this part of the study, and variable 3 identifies the 12 additional rats (13, . . . , 24). The experimental design for Part II of the study was exactly the same as for Part I, except that each rat received water each time after the fifth lever press. This reinforcement schedule will be denoted by FR-5. Variable 2 identifies the reinforcement schedule since Part I of the study used schedule FR-2 while Part II of the study used schedule FR-5. The reinforcement schedule thus is another factor (factor C) that was studied in the combined experiment.

To summarize, the variables for this experimental design are:

Variable Number	Variable Name	Description
1	Identification number	1–192
2	Part of study (factor *C*: reinforcement schedule)	1:Part I (FR-2) 2:Part II (FR-5)
3	Rat identification	1–24
4	Initial lever press rate (factor *A*)	1:Slow 2:Moderate 3:Fast
5	Dosage level (mg/kg) (factor *B*)	1:0 (saline solution) 2:.5 3:1.0 4:1.8
6	Observation unit	1, 2
7	Response variable—lever press rate	Total number of lever presses divided by elapsed time in seconds

Reference: T. G. Heffner; R. B. Drawbaugh; and M. J. Zigmond. "Amphetamine and Operant Behavior in Rats: Relationship between Drug Effect and Control Response Rate," *Journal of Comparative and Physiological Psychology* 86 (1974), pp. 1031–43.

1	2	3	4	5	6	7
1	1	1	1	1	1	.81
2	1	1	1	2	1	.80
3	1	1	1	3	1	.82
...	...	...	...	...	...	...
190	2	24	3	2	2	2.98
191	2	24	3	3	2	2.47
192	2	24	3	4	2	1.51