Appendix B
Data

All data sets are available on the Springer webpage or at the authors' home pages. More detailed information on the data sets may be found there.

B.1 Boston Housing Data

The Boston housing data set was collected by Harrison and Rubinfeld (1978). It comprise 506 observations for each census district of the Boston metropolitan area. The data set was analyzed in Belsley, Kuh and Welsch (1980).

X_1: per capita crime rate
X_2: proportion of residential land zoned for large lots
X_3: proportion of nonretail business acres
X_4: Charles River (1 if tract bounds river, 0 otherwise)
X_5: nitric oxides concentration
X_6: average number of rooms per dwelling
X_7: proportion of owner-occupied units built prior to 1940
X_8: weighted distances to five Boston employment centers
X_9: index of accessibility to radial highways
X_{10}: full-value property tax rate per \$10,000
X_{11}: pupil/teacher ratio
X_{12}: $1000(B - 0.63)^2 \boldsymbol{I}(B < 0.63)$ where B is the proportion of African American
X_{13}: % lower status of the population
X_{14}: median value of owner-occupied homes in \$1000

B.2 Swiss Bank Notes

Six variables measured on 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes. The data stem from Flury and Riedwyl (1988). The columns correspond

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-642-17229-8_21, © Springer-Verlag Berlin Heidelberg 2012

to the following 6 variables.

X_1: Length of the bank note
X_2: Height of the bank note, measured on the left
X_3: Height of the bank note, measured on the right
X_4: Distance of inner frame to the lower border
X_5: Distance of inner frame to the upper border
X_6: Length of the diagonal

Observations 1–100 are the genuine bank notes and the other 100 observations are the counterfeit bank notes.

B.3 Car Data

The car data set (Chambers, Cleveland, Kleiner and Tukey, 1983) consists of 13 variables measured for 74 car types. The abbreviations in Table B.3 are as follows:

X_1:	P	Price
X_2:	M	Mileage (in miles per gallone)
X_3:	R78	Repair record 1978 (rated on a 5-point scale; 5 best, 1 worst)
X_4:	R77	Repair record 1977 (scale as before)
X_5:	H	Headroom (in inches)
X_6:	R	Rear seat clearance (distance from front seat back to rear seat, in inches)
X_7:	Tr	Trunk space (in cubic feet)
X_8:	W	Weight (in pound)
X_9:	L	Length (in inches)
X_{10}:	T	Turning diameter (clearance required to make a U-turn, in feet)
X_{11}:	D	Displacement (in cubic inches)
X_{12}:	G	Gear ratio for high gear
X_{13}:	C	Company headquarter (1 for U.S., 2 for Japan, 3 for Europe)

B.4 Classic Blue Pullovers Data

This is a data set consisting of 10 measurements of 4 variables. The story: A textile shop manager is studying the sales of "classic blue" pullovers over 10 periods. He uses three different marketing methods and hopes to understand his sales as a fit of these variables using statistics. The variables measured are

X_1: Numbers of sold pullovers
X_2: Price (in EUR)
X_3: Advertisement costs in local newspapers (in EUR)
X_4: Presence of a sales assistant (in hours per period)

B.5 U.S. Companies Data

The data set consists of measurements for 79 U.S. companies. The abbreviations in Table B.5 are as follows:

X_1:	A	Assets (USD)
X_2:	S	Sales (USD)
X_3:	MV	Market Value (USD)
X_4:	P	Profits (USD)
X_5:	CF	Cash Flow (USD)
X_6:	E	Employees

B.6 French Food Data

The data set consists of the average expenditures on food for several different types of families in France (manual workers = MA, employees = EM, managers = CA) with different numbers of children (2, 3, 4 or 5 children). The data is taken from Lebart, Morineau and Fénelon (1982).

B.7 Car Marks

The data are averaged marks for 24 car types from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad) like German school marks. The variables are:

X_1:	A	Economy
X_2:	B	Service
X_3:	C	Non-depreciation of value
X_4:	D	Price, Mark 1 for very cheap cars
X_5:	E	Design
X_6:	F	Sporty car
X_7:	G	Safety
X_8:	H	Easy handling

B.8 French Baccalauréat Frequencies

The data consist of observations of 202100 baccalauréats from France in 1976 and give the frequencies for different sets of modalities classified into regions. For a

reference see Bourouche and Saporta (1980). The variables (modalities) are:

X_1:	A	Philosophy-Letters
X_2:	B	Economics and Social Sciences
X_3:	C	Mathematics and Physics
X_4:	D	Mathematics and Natural Sciences
X_5:	E	Mathematics and Techniques
X_6:	F	Industrial Techniques
X_7:	G	Economic Techniques
X_8:	H	Computer Techniques

B.9 Journaux Data

This is a data set that was created from a survey completed in the 1980's in Belgium questioning people's reading habits. They were asked where they live (10 regions comprised of 7 provinces and 3 regions around Brussels) and what kind of newspaper they read on a regular basis. The 15 possible answers belong to 3 classes: Flemish newspapers (first letter v), French newspapers (first letter f) and both languages (first letter b).

X_1:	WaBr	Walloon Brabant
X_2:	Brar	Brussels area
X_3:	Antw	Antwerp
X_4:	FlBr	Flemish Brabant
X_5:	OcFl	Occidental Flanders
X_6:	OrFl	Oriental Flanders
X_7:	Hain	Hainaut
X_8:	Lièg	Liège
X_9:	Limb	Limburg
X_{10}:	Luxe	Luxembourg

B.10 U.S. Crime Data

This is a data set consisting of 50 measurements of 7 variables. It states for one year (1985) the reported number of crimes in the 50 states of the U.S. classified according to 7 categories (X_3–X_9).

X_1:	land area (land)
X_2:	population 1985 (popu 1985)
X_3:	murder (murd)
X_4:	rape
X_5:	robbery (robb)

X_6: assault (assa)
X_7: burglary (burg)
X_8: larcery (larc)
X_9: autothieft (auto)
X_{10}: US states region number (reg)
X_{11}: US states division number (div)

division numbers		*region numbers*	
New England	1	Northeast	1
Mid Atlantic	2	Midwest	2
E N Central	3	South	3
W N Central	4	West	4
S Atlantic	5		
E S Central	6		
W S Central	7		
Mountain	8		
Pacific	9		

B.11 Plasma Data

In Olkin and Veath (1980), the evolution of citrate concentration in the plasma is observed at 3 different times of day, X_1 (8 am), X_2 (11 am) and X_3 (3 pm), for two groups of patients. Each group follows a different diet.

X_1: 8 am
X_2: 11 am
X_3: 3 pm

B.12 WAIS Data

Morrison (1990b) compares the results of 4 subtests of the Wechsler Adult Intelligence Scale (WAIS) for 2 categories of people: in group 1 are $n_1 = 37$ people who do not present a senile factor, group 2 are those ($n_2 = 12$) presenting a senile factor.

WAIS subtests:
X_1: information
X_2: similarities
X_3: arithmetic
X_4: picture completion

B.13 ANOVA Data

The yields of wheat have been measured in 30 parcels which have been randomly attributed to 3 lots prepared by one of 3 different fertilizers A, B, and C.

X_1: fertilizer A
X_2: fertilizer B
X_3: fertilizer C

B.14 Timebudget Data

In Volle (1985), we can find data on 28 individuals identified according to sex, country where they live, professional activity and matrimonial status, which indicates the amount of time each person spent on ten categories of activities over 100 days ($100 \cdot 24$ h $= 2400$ hours total in each row) in the year 1976.

X_1:	prof:	professional activity
X_2:	tran:	transportation linked to professional activity
X_3:	hous:	household occupation
X_4:	kids:	occupation linked to children
X_5:	shop:	shopping
X_6:	pers:	time spent for personal care
X_7:	eat:	eating
X_8:	slee:	sleeping
X_9:	tele:	watching television
X_{10}:	leis:	other leisures

maus:	active men in the U.S.
waus:	active women in the U.S.
wnus:	nonactive women in the U.S.
mmus:	married men in U.S.
wmus:	married women in U.S.
msus:	single men in U.S.
wsus:	single women in U.S.
mawe:	active men from Western countries
wawe:	active women from Western countries
wnwe:	nonactive women from Western countries
mmwe:	married men from Western countries
wmwe:	married women from Western countries
mswe:	single men from Western countries
wswe:	single women from Western countries
mayo:	active men from Yugoslavia
wayo:	active women from Yugoslavia

wnyo: nonactive women from Yugoslavia
mmyo: married men from Yugoslavia
wmyo: married women from Yugoslavia
msyo: single men from Yugoslavia
wsyo: single women from Yugoslavia
maes: active men from Eastern countries
waes: active women from Eastern countries
wnes: nonactive women from Eastern countries
mmes: married men from Eastern countries
wmes: married women from Eastern countries
mses: single men from Eastern countries
wses: single women from Eastern countries

B.15 Geopol Data

This data set contains a comparison of 41 countries according to 10 different political and economic parameters.

X_1: popu population
X_2: giph Gross Internal Product per habitant
X_3: ripo rate of increase of the population
X_4: rupo rate of urban population
X_5: rlpo rate of illiteracy in the population
X_6: rspo rate of students in the population
X_7: eltp expected lifetime of people
X_8: rnnr rate of nutritional needs realized
X_9: nunh number of newspapers and magazines per 1000 habitants
X_{10}: nuth number of television per 1000 habitants

AFS	South Africa	DAN	Denmark	MAR	Marocco
ALG	Algeria	EGY	Egypt	MEX	Mexico
BRD	Germany	ESP	Spain	NOR	Norway
GBR	Great Britain	FRA	France	PER	Peru
ARS	Saudi Arabia	GAB	Gabun	POL	Poland
ARG	Argentine	GRE	Greece	POR	Portugal
AUS	Australia	HOK	Hong Kong	SUE	Sweden
AUT	Austria	HON	Hungary	SUI	Switzerland
BEL	Belgium	IND	India	THA	Tailand
CAM	Cameroon	IDO	Indonesia	URS	USSR
CAN	Canada	ISR	Israel	USA	USA
CHL	Chile	ITA	Italia	VEN	Venezuela
CHN	China	JAP	Japan	YOU	Yugoslavia
CUB	Cuba	KEN	Kenia		

B.16 U.S. Health Data

This is a data set consisting of 50 measurements of 13 variables. It states for one year (1985) the reported number of deaths in the 50 states of the U.S. classified according to 7 categories.

X_1: land area (land)
X_2: population 1985 (popu)
X_3: accident (acc)
X_4: cardiovascular (card)
X_5: cancer (canc)
X_6: pulmonar (pul)
X_7: pneumonia flu (pnue)
X_8: diabetis (diab)
X_9: liver (liv)
X_{10}: Doctors (doc)
X_{11}: Hospitals (hosp)
X_{12}: U.S. states region number (r)
X_{13}: U.S. states division number (d)

division numbers		*region numbers*	
New England	1	Northeast	1
Mid Atlantic	2	Midwest	2
E N Central	3	South	3
W N Central	4	West	4
S Atlantic	5		
E S Central	6		
W S Central	7		
Mountain	8		
Pacific	9		

B.17 Vocabulary Data

This example of the evolution of the vocabulary of children can be found in Bock (1975). Data are drawn from test results on file in the Records Office of the Laboratory School of the University of Chicago. They consist of scores, obtained from a cohort of pupils from the eighth through eleventh grade levels, on alternative forms of the vocabulary section of the Coorperative Reading Test. It provides the following scaled scores shown for the sample of 64 subjects (the origin and units are fixed arbitrarily).

B.18 Athletic Records Data

This data set provides data on Men's athletic records for 55 countries in 1984 Olympic Games.

B.19 Unemployment Data

This data set provides unemployment rates in all federal states of Germany in November 2005.

B.20 Annual Population Data

The data shows yearly average population rates for Former territory of the Federal Republic of Germany incl. Berlin-West (given in 1000 inhabitants).

B.21 Bankruptcy Data I

The data are the profitability, leverage, and bankruptcy indicators for 84 companies.

The data set contains information on 42 of the largest companies that filed for protection against creditors under Chapter 11 of the U.S. Bankruptcy Code in 2001–2002 after the stock market crash of 2000. The bankrupt companies were matched with 42 surviving companies with the closest capitalizations and the same US industry classification codes available through the Division of Corporate Finance of the Securities and Exchange Commission (CF SEC, 2004).

The information for each company was collected from the annual reports for 1998–1999 (CF SEC, 2004), i.e., three years prior to the defaults of the bankrupt companies. The following data set contains profitability and leverage ratios calculated, respectively, as the ratio of net income (NI) and total assets (TA) and the ratio of total liabilities (TL) and total assets (TA).

B.22 Bankruptcy Data II

Altman (1968), quoted by Morrison (1990a), reports financial data on 66 banks.

X1 = (working capital)/(total assets)
X2 = (retained earnings)/(total assets)
X3 = (earnings before interest and taxes)/(total assets)
X4 = (market value equity)/(book value of total liabilities)
X5 = (sales)/(total assets)

The first 33 observations correspond to bankrupt banks and the last 33 for solvent banks as indicated by the last columns: values of y.

Original Data:

	X1	X2	X3	X4	X5	y
1	36.70	-62.80	−89.50	54.10	1.70	1
2	24.00	3.30	−3.50	20.90	1.10	1
3	−61.60	−120.80	−103.20	24.70	2.50	1
4	−1.00	−18.10	−28.80	36.20	1.10	1
5	18.90	−3.80	−50.60	26.40	0.90	1
6	−57.20	−61.20	−56.60	11.00	1.70	1
7	3.00	−20.30	−17.40	8.00	1.00	1
8	−5.10	−194.50	−25.80	6.50	0.50	1
9	17.90	20.80	−4.30	22.60	1.00	1
10	5.40	−106.10	−22.90	23.80	1.50	1
11	23.00	−39.40	−35.70	69.10	1.20	1
12	−67.60	−164.10	−17.70	8.70	1.30	1
13	−185.10	−308.90	−65.80	35.70	0.80	1
14	13.50	7.20	−22.60	96.10	2.00	1
15	−5.70	−118.30	−34.20	21.70	1.50	1
16	72.40	−185.90	−280.00	12.50	6.70	1
17	17.00	−34.60	−19.40	35.50	3.40	1
18	−31.20	−27.90	6.30	7.00	1.30	1
19	14.10	−48.20	6.80	16.60	1.60	1
20	−60.60	−49.20	−17.20	7.20	0.30	1
21	26.20	−19.20	−36.70	90.40	0.80	1
22	7.00	−18.10	−6.50	16.50	0.90	1
23	−53.10	−98.00	−20.80	26.60	1.70	1
24	−17.20	−129.00	−14.20	267.90	1.30	1
25	32.70	−4.00	−15.80	177.40	2.10	1
26	26.70	−8.70	−36.30	32.50	2.80	1
27	−7.70	−59.20	−12.80	21.30	2.10	1
28	18.00	−13.10	−17.60	14.60	0.90	1
29	2.03	−38.00	1.60	7.70	1.20	1
30	−35.30	−57.90	0.70	13.70	0.80	1
31	5.10	−8.80	−9.10	100.90	0.90	1
32	0.01	−64.70	−4.00	0.70	0.10	1
33	25.20	−11.40	4.80	7.00	0.90	1
34	35.20	43.00	16.40	99.10	1.30	0
35	38.80	47.00	16.00	126.50	1.90	0
36	14.00	−3.30	4.00	91.70	2.70	0
37	55.10	35.00	20.80	72.30	1.90	0
38	59.30	46.70	12.60	724.10	0.90	0
39	33.60	20.80	12.50	152.80	2.40	0
40	52.80	33.00	23.60	475.90	1.50	0
41	45.60	26.10	10.40	287.90	2.10	0
42	47.40	68.60	13.80	581.30	1.60	0
43	40.00	37.30	33.40	228.80	3.50	0
44	69.00	59.00	23.10	406.00	5.50	0
45	34.20	49.60	23.80	126.60	1.90	0
46	47.00	12.50	7.00	53.40	1.80	0
47	15.40	37.30	34.10	570.10	1.50	0
48	56.90	35.30	4.20	240.30	0.90	0

	X1	X2	X3	X4	X5	y
49	43.80	49.50	25.10	115.00	2.60	0
50	20.70	18.10	13.50	63.10	4.00	0
51	33.80	31.40	15.70	144.80	1.90	0
52	35.30	21.50	−14.40	90.00	1.00	0
53	24.40	8.50	5.80	149.10	1.50	0
54	48.90	40.60	5.80	82.00	1.80	0
55	49.90	34.60	26.40	310.00	1.80	0
56	54.80	19.90	26.70	239.90	2.30	0
57	39.00	17.40	12.60	60.50	1.30	0
58	53.00	54.70	14.60	771.70	1.70	0
59	20.10	53.50	20.60	307.50	1.10	0
60	53.70	35.90	26.40	289.50	2.00	0
61	46.10	39.40	30.50	700.00	1.90	0
62	48.30	53.10	7.10	164.40	1.90	0
63	46.70	39.80	13.80	229.10	1.20	0
64	60.30	59.50	7.00	226.60	2.00	0
65	17.90	16.30	20.40	105.60	1.00	0
66	24.70	21.70	−7.80	118.60	1.60	0