# Chapter 22
# Data

All data sets are available on the Springer webpage or at the authors' home pages.

## 22.1 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 analysed 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 \$1,000.

© Springer-Verlag Berlin Heidelberg 2015

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_22

## 22.2 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 to the following six 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.

## 22.3 Car Data

The car data set (Chambers, Cleveland, Kleiner & Tukey, 1983) consists of 13 variables measured for 74 car types. The abbreviations in this section are as follows:

| | | |
|---|---|---|
| $X_1$: | P | Price, |
| $X_2$: | M | Mileage (in miles per gallon), |
| $X_3$: | 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 USA, 2 for Japan, 3 for Europe). |

## 22.4 Classic Blue Pullovers Data

This is a data set consisting of ten measurements of four variables. The story: A textile shop manager is studying the sales of "classic blue" pullovers over ten 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).

## 22.5 US Companies Data

The data set consists of measurements for 79 US companies. The abbreviations in this section 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. |

## 22.6 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).

## 22.7 Car Marks

The data are averaged marks for 23 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.

## 22.8 French Baccalauréat Frequencies

The data consist of observations of 202,100 baccalauréats from France in 1976 and give the frequencies for different sets of modalities classified into regions. For a reference see Bouroche 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.

## 22.9 Journaux Data

This is a data set that was created from a survey completed in the 1980s 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 |

## 22.10 US Crime Data

This is a data set consisting of 50 measurements of 7 variables. It states for 1 year (1985) the reported number of crimes in the 50 states of the US 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 | | |

## 22.11 Plasma Data

In Olkin and Veath (1980), the evolution of citrate concentration in the plasma is observed at three 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 |

## 22.12 WAIS Data

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

WAIS subtests:

| | |
|---|---|
| $X_1$: | Information |
| $X_2$: | Similarities |
| $X_3$: | Arithmetic |
| $X_4$: | Picture completion |

## 22.13 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 fertilisers A, B and C.

| | |
|---|---|
| $X_1$: | Fertiliser A |
| $X_2$: | Fertiliser B |
| $X_3$: | Fertiliser C |

## 22.14 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 ($100 \cdot 24\,\text{h} = 2{,}400\,\text{h}$ 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 USA |
| waus: | Active women in the USA |
| wnus: | Nonactive women in the USA |
| mmus: | Married men in USA |
| wmus: | Married women in USA |
| msus: | Single men in USA |
| wsus: | Single women in USA |
| 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 |

## 22.15 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 realised |
| $X_9$: | nunh | Number of newspapers and magazines per 1,000 habitants |
| $X_{10}$: | nuth | Number of television per 1,000 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 | | |

## 22.16 US Health Data

This is a data set consisting of 50 measurements of 13 variables. It states for 1 year (1985) the reported number of deaths in the 50 states of the US 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)
$X_{12}$: US states region number (r)
$X_{13}$: US 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 | | |

## 22.17 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 Cooperative Reading Test. It provides the following scaled scores shown for the sample of 64 subjects (the origin and units are fixed arbitrarily).

## 22.18 Athletic Records Data

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

## 22.19 Unemployment Data

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

## 22.20 Annual Population Data

The data shows yearly average population rates for Former territory of the Federal Republic of Germany incl. Berlin-West (given in 1,000 inhabitants).

## 22.21 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 Chap. 11 of the US Bankruptcy Code in 2001–2002 after the stock market crash of 2000. The bankrupt companies were matched with 42 surviving companies with the closest capitalisations and the same US industry classification codes available through the Division of Corporate Finance of the Securities and Exchange Commission (SEC, 2004).

The information for each company was collected from the annual reports for 1998–1999 (SEC, 2004), i.e. 3 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).

## 22.22 Bankruptcy Data II

Altman (1968), quoted by Morrison (1990), 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*