11
Description of Data Sets

We describe the data sets used in the book

Alcohol

The solubility of alcohols in water is important in understanding alcohol transport in living organisms. This dataset from (Romanelli et al., 2001) contains physico-chemical characteristics of 44 aliphatic alcohols. The aim of the experiment was the prediction of the solubility on the basis of molecular descriptors. The columns are:

1. SAG=solvent accessible surface-bounded molecular volume
2. V= volume
3. Log PC (PC=octanol-water partitions coefficient)
4. P=polarizability
5. RM=molar refractivity
6. Mass
7. ln(Solubility) (response)

Algae

This dataset is part of a larger one (http://kdd.ics.uci.edu/databases/coil/coil.html), which comes from a water quality study where samples were taken from sites on different European rivers of a period of approximately one year. These samples were analyzed for various chemical substances. In parallel, algae samples were collected to determine the algae population distributions. The columns are:

1. season (1,2,3,4 for winter, spring, summer and autumn)
2. river size (1,2,3 for small, medium and large)
3. fluid velocity (1,2,3 for low, medium and high)

4-11 content of nitrogen in the form of nitrates, nitrites and ammonia, and other chemical compounds.

The response is the abundance of a type of algae (type 6 in the complete file). For simplicity we deleted the rows with missing values, or with null response values, and took the logarithm of the response.

Aptitude

There are three variables observed on 27 subjects:

Score: numeric, represents scores on an aptitude test for a course

Exp: numeric represents months of relevant previous experience

Pass: binary response, 1 if the subject passed the exam at the end of the course and

0 otherwise.

The data may be downloaded as dataset 6.2 from the site
http://www.jeremymiles.co.uk/regressionbook/data/

Bus

This dataset from the Turing Institute, Glasgow, Scotland, contains measures of shape features extracted from vehicle silhouettes. The images were acquired by a camera looking downwards at the model vehicle from a fixed angle of elevation

The following features were extracted from the silhouettes.

1. compactness
2. circularity
3. distance circularity
4. radius ratio
5. principal axis aspect ratio
6. maximum length aspect ratio
7. scatter ratio
8. elongatedness
9. principal axis rectangularity
10. maximum length rectangularity
11. scaled variance along major axis
12. scaled variance along minor axis
13. scaled radius of gyration
14. skewness about major axis
15. skewness about minor axis
16. kurtosis about minor axis
17. kurtosis about major axis
18. hollows ratio

Glass

This is part of a file donated by Vina Speihler, describing the composition of glass pieces from cars.

The columns are:

1. RI refractive index
2. Na2O Sodium oxide (unit measurement: weight percent in corresponding oxide, as are the rest of attributes)
3. MgO magnesium oxide
4. Al2O3 aluminum oxide
5. SiO2 silcon oxide
6. K2O potassium oxide
7. CaO calcium oxide

Hearing

Prevalence rates in percent for men aged 55-64 with hearing levels 16 decibels or more above the audiometric zero,

The rows correspond to different frequencies and to normal speech.

1. 500 herz
2. 1000 herz
3. 2000 herz

4. 3000 herz
5. 4000 herz
6. 6000 herz
7. Normal speech

The columns classify the data in seven occupational groups:

1. professional-managerial
2. farm
3. clerical sales
4. craftsmen
5. operatives
6. service
7. laborers.

Image

The data were supplied by A. Frery. They are a part of a synthetic aperture satellite radar image corresponding to a suburb of Munich.

Krafft

The Krafft point is an important physical characteristic of the compounds called *surfactants*, establishing the minimum temperature at which a surfactant can be used. The purpose of the experiment was to estimate the Krafft point of compounds as a function of their molecular structure.

The columns are:

1. Randiç index
2. Volume of tail of molecule
3. Dipole moment of molecule
4. Heat of formation
5. Krafft point (response)

Neuralgia

The data come from a study on the effect of iontophoretic treatment on elderly patients complaining of post-herpetic neuralgia. There were eighteen patients in the study, who were interviewed six weeks after the initial treatment and were asked if the pain had been reduced.

There are 18 observations on five variables:

Pain: binary response: 1 if the pain eased, 0 otherwise.

Treatment: binary variable: 1 if the patient underwent treatment, 0 otherwise.

Age: the age of the patient in completed years.

Gender: M (male) or F (female).

Duration: pretreatment duration of symptoms (in months)

Oats

Yield of grain in grams per 16-foot row for each of eight varieties of oats in five replications in a randomized-block experiment.

Solid waste

The original data are the result of a study on production waste and land use by Golueke and McGauhey (1970), and contains nine variables. Here we consider the following six.

1. industrial land (acres)

2. fabricated metals (acres)
3. trucking and wholesale trade (acres)
4. retail trade (acres)
5. restaurants and hotels (acres)
6. solid waste (millions of tons), response.

Stack loss

The columns are:

1. air flow
2. cooling water inlet temperature (^{0}C)
3. acid concentration (%)
4. Stack loss, defined as the percentage of ingoing ammonia that escapes unabsorbed (response).

Toxicity

The aim of the experiment was to predict the toxicity of carboxylic acids on the basis of several molecular descriptors. The attributes for each acid are:

1. log(IGC_{50}^{-1}): aquatic toxicity (response)
2. log Kow: Partition coefficient
3. pKa: Dissociation constant
4. ELUMO: Energy of the lowest unoccupied molecular orbital
5. Ecarb: Electrotopological state of the carboxylic group
6. Emet: Electrotopological state of the methyl group
7. RM: Molar refractivity
8. IR: Refraction index
9. Ts: Surface tension
10. P: Polarizability

Wine

This dataset, which is part of a larger one donated by Riccardo Leardi, gives the composition of several wines. The attributes are:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline