Data tables

Data comes in many forms. Initially we will consider data that can be represented as a data table.

• Rows correspond to observations; columns to variables.
• Each column is homogeneous (all the same type) but different columns can be different types of data.
• In SAS a data table is called a data set.
• In S a data table is called a data.frame.
• In relational databases like PostgreSQL, Oracle, MySQL, Access, . . ., a data table is called a relation.

Types of data

Typical kinds of univariate data that constitute the columns of the table are:

Continuous Numerical data that are represented by real numbers. Typically these are measurements of some physical property.

Ordinal Classifications into different groups where the groups have a natural ordering. For example, poor < average < good

Categorical Classification into groups without an ordering.

Further classification of numeric data

Occasionally numeric or continuous data are further classified as being on different scales, such as

interval scale i.e. a change of 0.1 units has the same interpretation across the range of the data.

ratio scale i.e. a doubling of the data value has the same interpretation across the range of the data

Temperatures in °F or °C are on an interval scale but not a ratio scale. Data being on a ratio scale implies that the zero value is “meaningful”.

These distinctions are not terribly important to us right now. However, it is always a good idea to consider whether zero is meaningful on the scale of the data. In particular, are the data restricted to positive or to non-negative values?
Data representations in S

Columns in an S data frame are typically

- **numeric**: Numeric data.
- **ordered**: Ordered categorical data.
- **factors**: Categorical data without an ordering.

There are other vector data types, such as logical and character but they typically are not used as columns of a data frame.

Data representations in SAS

SAS has two primary data types

- **numeric**: Double precision floating point data
- **character**: Fixed-length character strings

Model specification in a SAS procedure (or PROC) typically allows a class statement to specify which variables in the data set should be treated as classifications.

Data representations in databases

There are many, many types of representations of data in databases including various numeric and character formats, times, dates, datetimes, . . .

Although important in many situations, use of relational databases is beyond the scope of this course.

Accessing a data frame

Many sample data sets are available with R and with its add-on packages. The `data` function is used to access these data sets. If no argument is given, a list of the available data sets is returned. You must use, e.g. `data(trees)` before you can access the `trees` data.

The `help` function, as in `help(trees)` provides the description of the data. A synonym is `?trees`. The summary function provides a summary by column. The `str` function provides a very brief description of the structure of the data.

```
R> data(trees)
R> str(trees)
'data.frame': 31 obs. of 3 variables:
  $ Girth : num 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
  $ Height: num 70 65 63 72 81 83 66 75 80 75 ...
  $ Volume: num 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```
Summarizing a data frame

\[
R> \text{summary(trees)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Girth</th>
<th>Height</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>8.30</td>
<td>63</td>
<td>10.20</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>11.05</td>
<td>72</td>
<td>19.40</td>
</tr>
<tr>
<td>Median</td>
<td>12.90</td>
<td>76</td>
<td>24.20</td>
</tr>
<tr>
<td>Mean</td>
<td>13.25</td>
<td>76</td>
<td>30.17</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>15.25</td>
<td>80</td>
<td>37.30</td>
</tr>
<tr>
<td>Max.</td>
<td>20.60</td>
<td>87</td>
<td>77.00</td>
</tr>
</tbody>
</table>

Scatterplots

For a continuous response and a continuous covariate, the most common plotting technique is a scatterplot.

\[
R> \text{plot(Volume} \sim \text{Girth, data = trees)}
\]

Scatterplots with lattice

Deepayan Sarkar's lattice package provides xyplot, a more general plotting system based on Bill Cleveland's trellis graphics.

\[
R> \text{library(lattice)}
R> \text{print(xyplot(Volume ~ Girth, data = trees))}
\]

Scatterplot matrices

With more than one covariate, a scatterplot matrix or draftsman's plot provides all possible combinations of pairwise plots.

\[
R> \text{pairs(trees)}
\]
A data frame can be indexed as if it were a matrix, using `[' and ']' with two sets of indices separated by a comma. An omitted index is taken to be all the values of that index.

- **Rows 1 through 3**
  ```r
  R> trees[1:3,]
  Girth Height Volume
  1 8.3 70 10.3
  2 8.6 65 10.3
  3 8.8 63 10.2
  ```

- **Rows 4 through 6, columns 2 and 3 only**
  ```r
  R> trees[4:6, c(2, 3)]
  Height Volume
  4 72 16.4
  5 81 18.8
  6 83 19.7
  ```

### General definition of regression modelling

For the purposes of this course we assume there is one variable that constitutes the response in the data table. We call the other columns covariates.

In a broad sense, Regression is the study of how a response depends statistically upon a set of predictors. The definition we will use is

Regression is the study of the conditional distribution of \( y|x \) of a response variable \( y \) as the values of the predictors \( x \) are varied.

- We have measurements on a number of experimental units.
- Each measurement includes the response variable \( y \) and a set of predictor variables or covariates, \( x_1, x_2, \ldots, x_p \) where \( p \) is the number of predictors.
- The distribution could vary in mean, or in variance, or in skewness, or . . . or any combination of these.
Purpose of regression modeling

Regression modeling is used for several purposes, including:

- **Summary**: To summarize the relationship between the response and the covariates.
- **Prediction**: To predict the value the response will take on given particular values of the covariates. We must also quantify the precision of these predictions.
- **Hypothesis testing**: To test hypotheses on the relationship between the response and the covariates.
- **Comparing models**: Comparing different possible models for the relationship. Generally we are seeking a simple but adequate model.

Everything should be made as simple as possible, but not simpler.  
Albert Einstein

Variables in the trees data

For example, the trees data are described as:

This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

and the individual variables are:

- **Girth**: Tree diameter in inches
- **Height**: Height in ft
- **Volume**: Volume of timber in cubic ft

The diameter is called Girth because the actual measurement is the girth of the tree, measured by wrapping a tape around the tree. This measurement is converted to a diameter assuming a circular cross-section.

Response for the cherry trees

The purpose of collecting the data on the cherry trees is to be able to predict the volume of usable lumber from a tree while it is still standing.

The response is Volume. The covariates are Girth and Height.

There is an “interesting twist” here in that Girth is much easier to measure than Height. If there are two comparable models: one involving Height and one involving Girth only, the model involving Girth only could be preferred.

Notice that we must know the context of the data collection to decide how to model the data.