MAST90104: A First Course in Statistical Learning

Week 7 Practical and Workshop

1 Practical questions

1. The data *teengamb* from the `faraway` package contains data from a survey in Britain to study teenage gambling. The variables are

   - **sex**: 0=male, 1=female
   - **status**: Socioeconomic status score based on parents' occupation income in pounds per week
   - **verbal**: verbal score in words out of 12 correctly defined
   - **gamble**: expenditure on gambling in pounds per year

   We can import the data by

   ```
   data(teengamb)
   ```

   We are interested in predicting **gamble** using the other variables. Implement the following variable selection methods to determine the "best" model

   (a) Backward elimination

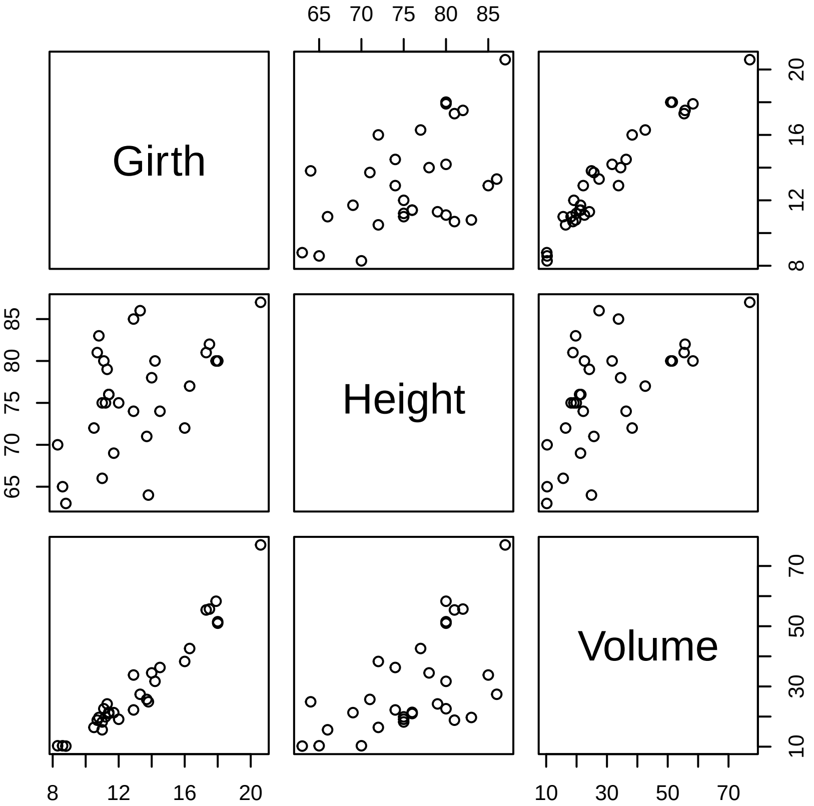
   (b) Forward selection

   (c) Stepwise selection

   Comment on the AIC of the final model chosen by the 3 methods.

2. Load and examine the dataset `trees` using

   ```
   data(trees)
   ?trees
   pairs(trees)
   ```



We will model the volume of a black cherry tree as a function of its girth and height.

(a) By calculating $R(\gamma_1|\gamma_2)$ and SS_{Res} from the data $\mathbf{y}$ and design matrix X, use an F test to determine if including the variable `Height` significantly improves the model fitted using only `Girth` (and an intercept).

Repeat the test using the `lm` and `anova` commands, to see if you get the same numbers.

(b) Add variables `Girth` squared and `Girth` squared times `Height` to the model, then use stepwise selection to simplify the model. (You can use `step` for this step.)

Comment on the form of your final model.

(c) Use diagnostic plots to check the fit of your final model.

(d) What transformation might be indicated from the plot of residuals versus fitted values? Transform all variables with this transformation. What might the appropriate model be? Fit it and comment on the resulting residuals.

3. In a manufacturing plant, filters are used to remove pollutants. We are interested in comparing the lifespan of 5 different types of filters. Six filters of each type are tested, and the time to failure in hours is given in the dataset (on the website) `filters` (in `csv` format).

(a) Use the `read.csv` function to read the data. Then convert the `type` component into a factor.

(b) Using only the `matrix` command, construct a $\mathbf{y}$ vector and a full rank X for this linear model, corresponding to `cont.treatment`

(c) Fit the models and compare with the `lm` output.

(d) Calculate s^2 using the residuals

2 Workshop questions

1. Show that R^2 is the square of the correlation coefficient between the data and the fitted values. [*Hint: write the response as fit + residual. Express the correlation coefficient as that of a random vector which chooses randomly from the data (both y and corresponding row of X) and has values of response and fit. Use rules for covariance and the correlation between fitted values and residuals.*]

2. Show that the adjusted R^2 satisfies:

$$\text{adjusted } R^2 = 1 - \frac{\text{estimate of } \sigma^2 \text{using the model}}{\text{estimate of } \sigma^2 \text{assuming equal means}}$$

3. Suppose each row of a dataset has a response variable and two factors, which have 2 and 3 possible levels respectively. The dataset has 2 rows for each possible combination of factor levels. We model this with a less than full rank model with one parameter for the overall mean, and one parameter for each level of each factor, assuming that the overall mean is adjusted additively by each factor. Write down the linear model in both equation and matrix form.

4. Let

$$A = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ 0 & 1 & -3 & -2 \end{bmatrix}.$$

   (a) Show that $r(A) = 2$.

   (b) Construct two different full rank matrices which generate the same column space as A.

5. It is known that toxic material was dumped into a river that flows into a large salt-water commercial fishing area. We are interested in the amount of toxic material (in parts per million) found in oysters harvested at three different locations in this area. A study is conducted and the following data obtained:

Site 1	Site 2	Site 3
15	19	22
26	15	26

   (a) Write down the linear model in matrix form.

   (b) Write down the normal equations.

   (c) Reparameterize the model to a full rank model.

   (d) Find a solution for the normal equations.

6. In the one-way classification model, show that any linear combination of $\bar{y}_1 - \bar{y}., \ldots, \bar{y}_k - \bar{y}.$ can be written as a linear combination of $\bar{y}_1, \ldots, \bar{y}_k$. Does the converse hold?