MAST9014: Introduction to Statistical Learning

Week 6 Lab and Workshop

1 Lab

1. Load the **beef** dataset from the LMS or the server:

```
beef <- read.csv('beef.csv')
```

In the USA, the Cattlemen's Beef Board and the National Cattlemen's Beef Association promote the consumption of beef with an advertising campaign using the theme "Beef: it's what's for dinner". The campaign is paid for by the "Beef Checkoff", a law that requires all cattle producers to pay $1 per head of cattle sold to support beef/veal promotion and research. In 1988 the Missoulian newspaper surveyed the cattle growers of Montana, and for each of Montana's 56 counties reported the percent of growers voting "yes" for the checkoff.

In this question we explain the size of the yes vote in terms of the characteristics of the farms in each county. Data on farms is taken from the U.S. Bureau of the Census, City and County Data Book, 1986. The variables given in the dataset are:

yes Percentage of farmers voting "yes" for the checkoff

big Percentage of farms with 500 acres or more

prin Percentage of operators whose principle income is farming

size Average size of farm (hundreds of acres)

val Average value of products sold ($1000's)

live Percentage of products sold from livestock and poultry

sale Percentage of farms with sales of $100,000 or more

(a) Use `pairs` to plot the data. Is there any evidence of non-linearity or heteroskedasticity?

(b) Using the `add1` and `drop1` commands, use forward and backward selection to find parsimonious models for `yes`.

(c) Using the `step` command, starting from a model with just an intercept, use the AIC and stepwise selection to choose a model.

(d) Show that the model found in 1c can be improved by adding the interaction term `size*sale`. (Important here is how you judge "improved".)

Use stepwise selection again to see if adding `size*sale` can let you remove any other variables from the model.

(e) Suppose that β_1, β_2 and β_{12} are the coefficients of $x_1 =$ `size`, $x_2 =$ `sale` and `size*sale`, in the model from 1d. Plot $\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 * x_2$ as a function of (x_1, x_2), to see the combined effect of these variables on the yes vote. You may need the `wireframe` function from the `lattice` library, and also `expand.grid`.

(f) Repeat the above question using the model with no `size*sale` interaction term from 1c.

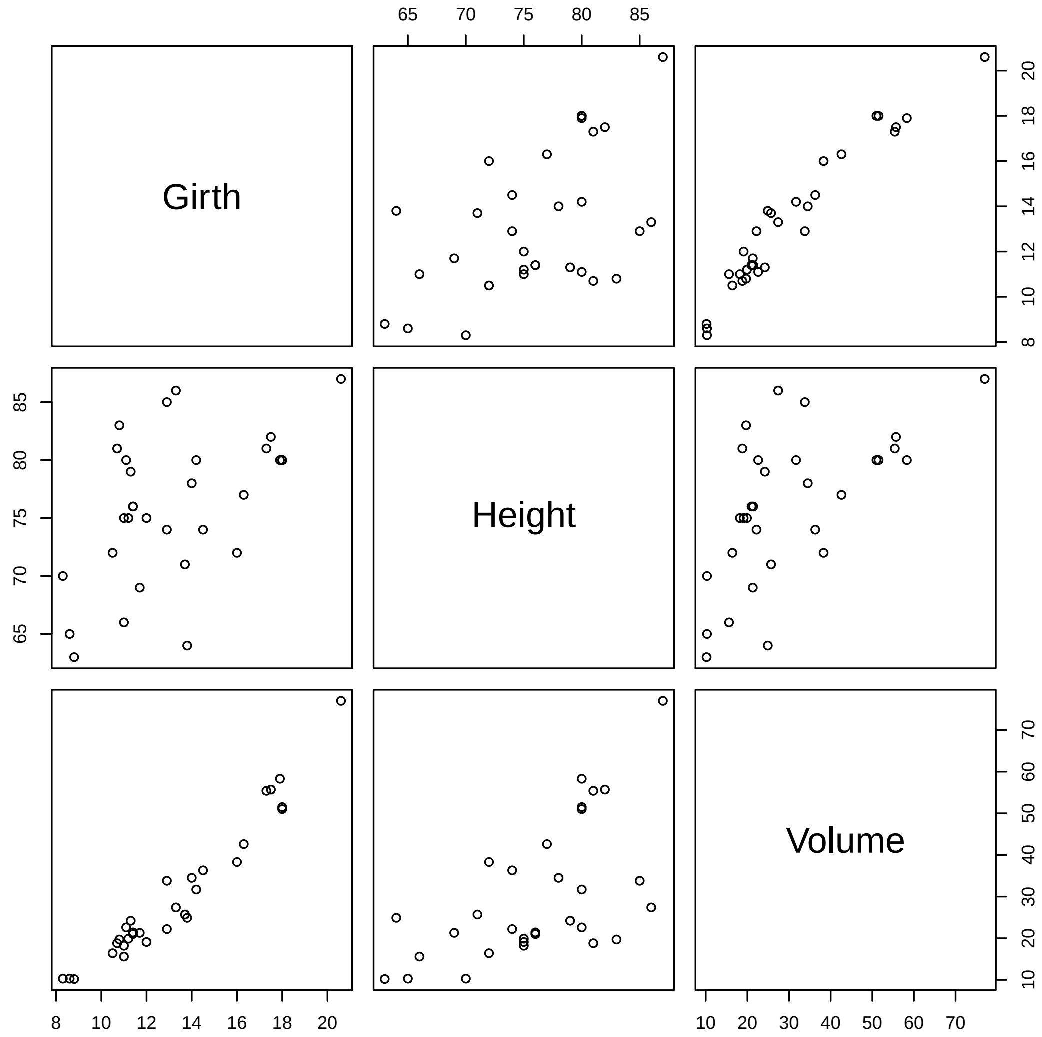
(g) Use the diagnostic plots provided by R to assess the model from 1d.

Refer back to 1a; do you need to transform the data and start again?

(h) Which are the most important variables when it comes to predicting the yes vote? In deciding this, take into account the average size of the variables as well as the size of the fitted coefficients.

2. Load and examine the dataset `trees` using

```r
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.

2 Workshop

3. If the full rank linear model has an intercept term, what is the mean of the fitted values? [*Hint: use the normal equation for the intercept term.*]

4. 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.*]

5. Show that the adjusted R^2 satisfies:

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

6. 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.

7. 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) What is the treatment contrast matrix for this model?

   (c) What is the sum contrast matrix for this model?

8. 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.