MAST30025: Linear Statistical Models

Week 7 Lab

1. Recall the joint confidence region for the parameters of a full rank linear model:

$$(\mathbf{b} - \boldsymbol{\beta})^T X^T X (\mathbf{b} - \boldsymbol{\beta}) \le ps^2 f_\alpha.$$

   Use this to derive a test for the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. Show that this test is equivalent to the test for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$ obtained using the general linear hypothesis.

2. Load the `beef` dataset from the website:

   ```
   > beef <- read.csv('../data/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 2c 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 2d. 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 2c.

   (g) Use the diagnostic plots provided by R to assess the model from 2d.

   Refer back to 2a; 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.

3. Load and examine the dataset `trees` using

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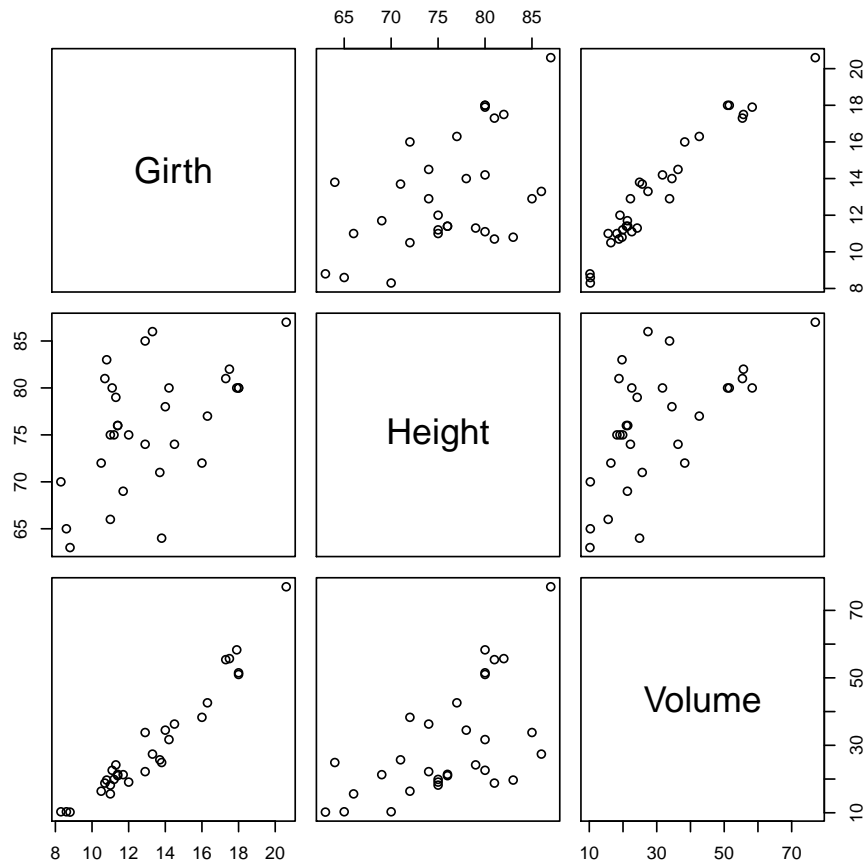

We will model the volume of a black cherry tree as a function of its girth and height.

(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.