MAST90104: A First Course in Statistical Learning

Week 6 Practical and Workshop

1 Practical questions

1. The data set `ufc.csv` contains forest inventory observations from the University of Idaho Experimental Forest. In the experiment, scientists randomly selected a number of plots and then from each plot selected a number of trees. For each tree they measured its height and diameter (which are numeric), and also the species of tree (which is a character string). Answer the following questions:

   (a) What are the species of the three tallest trees? Of the five fattest trees? (Use the `order` command.)

   (b) What are the mean diameters by species?

   (c) What are the two species that have the largest third quartile diameters?

   (d) What is the identity of the tallest tree of the species that was the fattest on average?

2. The following questions use the 'sleep' dataset, which you can download from the course website. This dataset contains (among other things) data on the body weight (kg) and brain weight (g) of 62 mammals. Use the following commands to read the data (make sure the data file is in your working directory, or change to the correct path):

```
mammals <- read.csv("sleep.csv")
```

   This creates a data frame, `mammals`, with components (among others) named `BodyWt` and `BrainWt`We are interested in predicting brain weight from body weight.

   (a) Plot the data. Fit the model of brain weight vs. body weight using the `lm` function. Plot the diagnostics plots and comment on the plots. Is the model appropriate ?

   (b) Apply a logarithmic transformation to both `BodyWt` and `BrainWt`.

```
mammals$BodyWt <- log(mammals$BodyWt)
mammals$BrainWt <- log(mammals$BrainWt)
```

   Fit a linear model explaining (transformed) brain weight from body weight, using the `lm` command.

   Display the summary of the fitted model, and then create a scatter plot of the data and superimpose the fitted regression line on it. Does it look like a reasonable fit?

   Use diagnostic plots to assess if the model assumptions are satisfied.

   (c) Find a 95% confidence interval for a mammal weighing 50 kg.

   (d) Find a 95% prediction interval for a mammal weighing 50 kg.

   (e) Test the following hypotheses, using the `anova` function.

      i. $H_0 : \boldsymbol{\beta} = 0$

      ii. $H_0 : \beta_1 = 0$

      iii. $H_0 : \beta_0 = 0$

      iv. $H_0 : \boldsymbol{\beta} = (2, 1)$

   (f) Write down the final model for the untransformed data.

2 Workshop questions

1. Suppose X is $n \times p$ of full rank and C is $r \times p, r \leq p$ also of full rank.

   (a) Show that $X^T X$ is positive definite (hint: use the definition).

   (b) Show that $C(X^T X)^{-1} C^T$ is positive definite (hint: why does $(X^T X)^{-1}$ have a matrix square root?).

   (c) Show that $C(X^T X)^{-1} C^T$ is invertible.

   (d) Show that $[C(X^T X)^{-1} C^T]^{-1}$ is positive definite.

2. In this question we consider the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. The test statistic for this hypothesis is

$$\frac{(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*)/p}{SS_{Res}/(n-p)}.$$

   (a) Show that

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*) = (\mathbf{y} - X\boldsymbol{\beta}^*)^T (\mathbf{y} - X\boldsymbol{\beta}^*) - (\mathbf{y} - X\mathbf{b})^T (\mathbf{y} - X\mathbf{b}).$$

   That is, it is the SS_{Res} for the null model minus the SS_{Res} for the full model.
   Also show that, in general,

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*) \neq \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y} - \boldsymbol{\beta}^{*T} X^T X \boldsymbol{\beta}^*.$$

   That is, in this case we can not write it as the SS_{Reg} for the full model minus the SS_{Reg} for the model under H_0.

   (b) Show directly that $(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*)$ and SS_{Res} are independent, that is without using our existing results that $\mathbf{b}$ and SS_{Res} are independent.
   Hint: set $\mathbf{q} = \mathbf{y} - X\boldsymbol{\beta}^*$ then

      i. Show that $(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*) = \mathbf{q}^T X (X^T X)^{-1} X^T \mathbf{q}$.

      ii. Show that $SS_{Res} = \mathbf{q}^T [I - X(X^T X)^{-1} X^T] \mathbf{q}$ and hence that these two quadratic forms are independent.

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

$$(\mathbf{b} - \boldsymbol{\beta})^T X^T X (\mathbf{b} - \boldsymbol{\beta}) \leq p s^2 f_\alpha.$$

   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.