MAST90104: A First Course in Statistical Learning

Assignment 3, 2022

Due: 11:59 pm Monday 3 October. Please submit a scanned or other electronic copy of your work via the Learning Management System

This assignment is worth 5% of your total mark. The total point is 40. Include your R code with your submission (either in your answer or add them as a separated file). Late submissions will have their mark deducted.

1. (12 pt) A study was conducted to determine the effect of the size of the root system on the growth of Douglas-fir seedlings when they are planted out. Seedlings were obtained from three seed lots, and when they were planted out their root volume was classified as small (RV1), medium (RV2), or large (RV3). The heights of the seedlings were then measured at the end of the first growing season. The data from the experiment is given in the file `douglas.csv`.

   (a) Fit an additive two-factor model to the data.

   (b) From this model, estimate the difference between the heights of the J052 lot and B349 lot.

   (c) Fit a linear model with interaction to the data. Calculate a confidence interval for the difference between the heights of large (RV3) and medium (RV2) seedlings in the B349 seed lot.

   (d) Test the hypothesis that the height of seedlings from the J052 plot has no dependence on root volume.

   (e) Generate an interaction plot for the data. Is there any evidence of an interaction?

   (f) Test for the presence of interaction between root volume and seed lot.

2. (8 pt) Depletion of the ozone layer allows the most damaging ultraviolet radiation to reach the Earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable). To measure the relationship, researchers sampled the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. The data from this study is given in the file `ozone.csv`. There are 3 variables:

   - **Inhibit**: percent inhibition of primary phytoplankton production in water
   - **UVB**: UVB exposure
   - **Surface**: a factor with levels ”Deep” and ”Surface”

   Hint: When importing the data to R, you will need to specify `Surface` as factor

   (a) Plot the percentage inhibition against UVB exposure, use different colours for observations at the surface and in the deep.

   (b) Test whether the effect of UVB exposure on percentage inhibition differ at the surface and in the deep.

3. (8 pt) You wish to perform a study with 4 levels of a new treatment and one placebo using a completely randomised design. The sample units are to be divided into 5 groups, with the last group being given the placebo. You want to study the contrasts $\tau_1 - \tau_5, \tau_2 - \tau_5, \tau_3 - \tau_5, \tau_4 - \tau_5$. You are given resources to study 30 sample units.

   (a) Determine the optimal allocation of the number of units to assign to each treatment.

(b) Perform the random allocation. You must use R for randomisation and include your R commands and output.

4. (12 pt) **You should not use R's `glm` command for this question.** Aflatoxin B1 was fed to lab animals at various doses. The dose in ppb (`dose`), number of test animals (`total`) and number with liver cancer (`tumor`) is recorded. The data is presented in the following table

dose	total	tumor
0	18	0
1	22	2
5	22	1
15	21	4
50	25	20
100	28	28

We are interested in building a model to predict the occurrence of liver cancer.

(a) Fit a binomial regression model to the data using a logit link.

(b) Construct the 95% CIs for the parameter estimates.

(c) Perform a likelihood ratio test for the significance of the dose coefficient.

(d) Estimate the probability of developing liver cancer when the dose is 70 ppb together with a 95% CI.

(e) Refit the binomial regression model using a probit link.

(f) Create a plot comparing the fitted probit model to the fitted logit model.