STAT270: Applied Statistics - Assignment 2

Semester 2 – 2017. Due : Friday 3rd November 5:00pm

You are expected to write the assignment using MS Word or similar word processor (or in PDF format). You are required to write your name, student ID and Unit Code written on the first page. Please submit your assignment via the provided submission link on iLearn.

You may discuss the assignment in the early stages with your fellow students. However, the assignment submitted should be your own individual work.

You are encouraged to use R Markdown if you are familiar but this is not required. A 'Cheatsheet' from the RStudio team is given here.

In your answers to the questions below, produce the appropriate `R` output and/or explanation of the steps and results. Don't include any more `R` output than necessary and include only concise explanations.

Question 1

Chief executive office (CEO) compensation varies significantly from firm to firm. For this question, you will report on a sample of firms from a survey by *Forbes* magazine to establish important pattern in the compensation of CEOs. The data is available in the file `CeoCompensation.csv` on `iLearn`. This data will be used to study CEO and firm characteristics to determine the important factors influencing CEO compensation.

COMP	Sum of salary, bonus and other compensation, in thousands of dollars. Other compensation does not include stock gains.
AGE	The CEOs age, in years.
EXPER	Number of years as the firm CEO
logSALES	$\log_e$ transform of Sales revenues, in log(millions of dollars)
PROF	Profit of the firm, before taxes, in millions of dollars.

a. Produce a matrix scatterplot of the data and describe the possible relationships between the response and predictors and relationships between the predictors themselves. Explain if the data is suitable for a multiple regression model.

b. Compute the correlation matrix of the dataset and reconcile the correlation matrix with the matrix scatterplot in part a) above.

c. Fit a multiple model using all the predictors to explain the `COMP` response. Conduct an F-test for the overall regression i.e. is there **any** relationship between the response and the predictors. In your answer:

- Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.
- Write down the Hypothesis for the Overall ANOVA test of multiple regression.
- Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).
- Compute the F statistics for this test.
- State the Null distribution.
- Compute the P-value.

- State your conclusion (both statistical and contextual conclusions).

d. Using the backward model selection procedure discussed in the course, find the best multiple regression model that explains the data by using `COMP` as the response and start with `all` the predictors provided.

e. Validate your final model and explain why it is **not** appropriate to use the multiple regression model to explain the `COMP` response.

f. Re-fit the model using `log(COMP)` as the new response variable. In your answer, use the backward selection procedure discussed in the course to find the best multiple regression model to explain `log(COMP)`. You should again include `all` the predictors provided in your initial model.

g. Validate your final model with the `log(COMP)` response. In particular, in your answer, explain why the regression model with `log(COMP)` response variable is superior to the model with the `COMP` response variable.

Question 2

Berger and Walker (1972) monitored the heart rates for six randomly selected tree shrews (*Tupaia glis*) during three different stages of sleep: LSWS (light slow-wave sleep); DSWS (deep-slow-wave sleep) and REM (rapid eye movement sleep). The authors wish to account for any differences between the tree shrews and stages of sleep.

HeartRates	The heart rates of the tree shrew;
Shrews	The specific tree shrews (labelled as 1, 2, 3, 4, 5 and 6);
Sleep	The stages of sleep (labelled as "LSWS", "DSWS", "REM").

The data is available in the file `TreeShrews.dat` on `iLearn`.

a. For this study, is the design balanced or unbalanced? Explain why.

b. Construct two different preliminary graphs that investigate different features of the data and comment.

c. Explain why we **cannot** fit a Two-Way ANOVA with interaction model to this dataset.

d. Analyse the data, stating the null and alternative hypothesis for each test, and check assumptions. In your answer:

- Write down the mathematical model for this situation, defining all appropriate parameters.
- State the appropriate hypotheses.
- Compute an appropriate ANOVA table.
- Check assumptions.

e. State your conclusions about the effect of `Shrews` and `Sleep` on `HeartRates`. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests in c. and the preliminary plots in b.. You do not need to statistically examine the multiple comparisons.