STAT2402: Week 2 Computer Laboratory

In this lab we will first examine the data in the first week's lectures, specifically:

1. some exercises from the R lecture notes; and
2. fitting a linear regression model.

Getting Started:

Log in at one of the PCs and start up the software package R; either directly or via RStudio. If you have problems either logging in or starting R ask for help.

Recall: when typing in R commands you can use the arrow keys to speed things up. The 'up' arrow gives you the previous command that you typed. The usual prompt sign for R is >. If you get a + prompt sign instead, it means that R is awaiting the completion of the previous command that you typed in. This can happen because you have forgotten to close parentheses, for instance. Just type in the remainder of the command. **Note also that R is case sensitive.**

Use scripting an save your code.

Exercise 1: Battery lifetimes—from R lecture notes An engineer is designing a battery for use in a device that will be subjected to some extreme variations in temperature. He has three possible choices for the plate material. For testing purposes he selects three temperatures. Four batteries are tested at each combination of plate material and temperature and the tests are run in random order. The battery life (hours) under each set of conditions is given in Table 1.

	Temperature (°C)					
Material	-10		20		55	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

Table 1: Life (in hours) data for the battery design example.

1. Examine the data and report your observations.

2. Write the R code to read this data into R.

3. Find the summary statistics for this data. First think about what sort of statistics you should be interested in.

4. Now fit a linear model to the battery lifetimes. Investigate any interaction terms.

5. Select the best model based on your analysis.

6. Perform appropriate model diagnostics.

7. Produce an interaction plot for the mean battery lifetimes.

8. Interpret your model.

9. Which material would you recommend for the batteries? Justify your selection.

Exercise 2: Data manipulation Consider the following grouped data on seatbelt use and the severity of injury in an accident.

	worn	not worn	unknown
fatal	35	6	15
severe	1142	48	328
minor	7969	76	764
unknown	11404	24	38570

1. Enter the data into R using the variables `Injury, SeatBelt` and `Frequency`.

2. Now we want to create date that contains one record for each case. That is, we need to create 35 entries corresponding to a fatal injury where the seat belt was worn, 6 for when the seat belt was worn, and 15 for unknown. Similarly for the other levels of injuries. Write a short (2 lines!) of R code to achieve this, and test your code (for example, by producing a table from your new data).

Exercise 3: Fish data The folder Data in the Computer Labs folder contains the data set `Fish.txt`. Download the file and read the data into R. The variables are:

- Code: fish species code
- Weight: weight of the fish in grammes
- Length1: length from the nose to the beginning of the tail (cm)
- Length2: length from nose to notch of tail (cm)
- Length3: length from nose to the end of the tail (cm)
- Height: maximum height as a percentage of Length3

- Width: maximum width as a percentage of Length3

1. Summarise the data and check for any data errors.
2. You will note a weight of 0. Determine which data record this corresponds to and omit it. Use the commands `which` and `fish1 <- fish[-x,]`, where `x` corresponds to the number of the record in error. Check that the record with the error has been removed.
3. Note that `Code` for the species of fish. This is currently numerical and needs to be converted to a factor. Use the code `fish1$Code <- factor(fish1$Code)`. (Note that if `Code` is left as numerical, the model will estimate a single coefficient for it. The contribution of this variable will then be linear in this coefficient. So for example, the effect of a value 2 for `Code` is twice that for a value 1. This is not correct.)
4. Fit a linear regression model with `Weight` as response against the other covariates.
5. Investigate interaction terms in the model.
6. Perform model diagnostics. In particular, examine the plot of residuals against fitted values for any patterns (indicating issues with a linear model fit) or change in spread (indicating a violation of homogeneous variance assumption).
7. By examining plots of the explanatory variables against the response variable, determine an appropriate transformation of data to improve the model for weight against the other morphological measurements.
8. Fit your selected model.
9. Reduce the model removing non-significant variables one by one, until a model with only significant terms is left. Use `update` command. For example, `fish.lm1 <- update(fish.lm .∼.-Length2)`.
10. Perform model diagnostics. For this, plot a histogram of the residuals and a scatter plot of the residuals against the fitted values. Comment on whether the model assumptions are satisfied.
11. Explore the data further and decide how the model can be improved.
12. Report your findings on the dependence of the weight of the fish on the explanatory variables.

Exercise 4: Bank data The folder Data in the Computer Labs folder contains the data set `Bank.txt`. The female employees are suing the bank for gender discrimation in salary. Download the file and read the data into `R`. For each employee the Bank Data as the following variables.

- EducLev: education level, a categorical variable with categories 1 (finished high school), 2 (finished some tertiary education), 3 (obtained a bachelor's degree), 4 (took some postgraduate courses), 5 (obtained a postgraduate degree).

- Job Grade: a categorical variable indicating the current job level, the possible levels being 1 (lowest) to 6 (highest).
- YrHired: year employee was hired.
- YrBorn: year employee was born.
- Gender: a categorical variable with values "Female" and "Male".
- YrsPrior: number of years of work experience at another bank prior to working at First National.
- PCJob: a categorical yes/no variable indicating whether the employees current job is PC related.
- Salary: current salary in thousands of dollar.

1. Summarise the data and check for any data errors.
2. Fit a linear model to the Salary. Do not include any interactions.
3. Reduce the model to only significant terms.
4. Perform appropriate model diagnostics.
5. Is there a gender bias in salaries? Justify your decision.
6. Now include appropriate interaction terms. You may have to consider which interactions are meaningful.
7. Again reduce the model to only significant terms.
8. Perform model diagnostics.
9. Under this new model, is there gender bias in salaries? Justify your decision.
10. Produce a scatterplot of fitted salaries against observed salaries. The plotting character should be Sex (M or F), and the colour code should be by education level.
11. Comment your findings from the plot.
12. What form of discrimination can you detect from your analysis?

Finishing Off:

When you've finished, close down R by typing `q()`. Choose 'Save' when prompted as to whether you want to retain your workspace. Remember to log off from your computer before leaving.