STAT3500
Problems and Applications in Modern Statistics
Assignment 2

Due: 2pm on Friday 3 September via Blackboard
Weighting: 25%

Instructions: There are **three** questions in this assignment. The associated maximum mark for each question is provided for your information noting that marks will be allocated for correctness, clarity and completeness. You should produce and submit a document that displays your relevant and non-excessive R code along with comments and interpretation. (i.e. it is recommended you use R Markdown (.Rmd) to create it). Submit your work as a single PDF, Word or HTML document to Blackboard by the due date and time. Note that one part of Question 1 involves interpretation and communication of results in the form of an audio recording which you should upload onto Blackboard as an audio file.

1. [**11 marks**]
   An experiment was established to investigate the toxicity of a chemical spray applied to insects as a result of them starting to show resistance. Twenty insects (*Total*) of each sex (*Sex*) were exposed at each of six doses (*Dose*) of the chemical spray, and the number that were killed (*Dead*) were recorded. The data in **Insects.csv** summarise the results.

   (a) Read the data into R and produce a plot that demonstrates the proportion of insects killed against dose for each sex. Produce the same plot but using the logarithm of dose (logdose) and comment on the nature of the relationship between the variables.

   (b) Fit a logistic regression model to this data, with logdose as a predictor, allowing for a different response to the spray for each sex, and ensuring that the model is forced through the origin (i.e. no explicit intercept term). Make 2-3 relevant statements about the model parameter estimates and goodness-of-fit.

   (c) Now consider the simpler model without the interaction between sex and logdose. Perform an analysis of deviance to determine the significance of the interaction between logdose and sex. You can do this using anova(M2,M1,test="Chisq") where M1 and M2 are the two model outputs, with M2 being the model output for the more complex model. Record the p-value and comment on its interpretation.

   (d) Does sex have a significant effect on the probability of an insect dying? Explain your response.

   (e) **Audio question**: Explain the difference in the interpretation of the p-value associated with the interaction observed in the analysis of deviance (Part c) and that reported in the summary of model output (Part b)?

2. [**8 marks**]

Suppose we are interested in determining the volume of timber in a tree V_i (in cubic feet) based on measurements of the tree's height H_i (in feet) and diameter D_i (in inches). A suitable model might be

$$V_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 H_i^{\beta_1} D_i^{\beta_2}$$

(a) Discuss whether this is a GLM and specify which link function might be appropriate.

(b) For the dataset **treedims.csv**, graphically explore the data before fitting the model specified above. Comment on the model output. Write down an expression for the fitted model.

(c) If you weren't provided with the family specification for V_i, discuss an alternative suitable modelling approach.

3. [**6 marks**]

Data were collected from two cancer treatment centers (*center*) with the intention of investigating whether the survival rate is different between the two treatment centers or if any apparent difference can be explained by differences in the grade of cancer (*malignant*) of the respective patients. The data are provided in **cancer.csv**.

(a) Read the data into R and ensure all predictors are factors. Fit a saturated model to the counts, that is the mean model should include all possible interactions between the factors. Comment on the model output.

(b) By removing non-significant effects one at a time, starting with the highest order interactions, find an optimal model. Justify your decisions.

(c) What is the estimated dispersion parameter when fitting a quasi-Poisson model to the optimal model found in Part (b)? Provide a brief interpretation.