STAT3500
Problems and Applications in Modern Statistics
Assignment 1

Due: 2pm on Friday 20 August via Blackboard
Weighting: 25%

Instructions: There are **five** 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. Questions 1 - 3 are probably best answered with pen and paper and scanning in your answers. For Questions 4 and 5, as you should produce a document that displays your relevant and non-excessive R code along with comments and interpretation, 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 two parts of questions involve interpretation and communication of results in the form of an audio recording which you should upload onto Blackboard as an audio file.

1. [**2 marks**]
   Let X_1, X_2 and X_3 be a random sample of size three from a $Uniform[\theta, 2\theta]$ distribution, where $\theta > 0$. Find the MLE, $\hat{\theta}$, and find a constant k such that $\mathbb{E}_\theta(k\hat{\theta}) = \theta$.

2. [**2 marks**]
   For the general exponential family given by

$$f(y;\theta) = \exp\left\{\frac{a(y)b(\theta) + c(\theta)}{e(\phi)} + d(y;\phi)\right\},$$

   find expressions for

   (a) $E[a(Y)]$; and

   (b) $\text{Var}[a(Y)]$

   in terms of the known functions $a(\cdot), b(\cdot), c(\cdot), d(\cdot)$ and $e(\cdot)$

3. [**3 marks**]
   For the Pareto distribution with probability density function defined as

$$f_Y(y;\theta) = \theta y^{-\theta-1}, \quad y > 1, \theta > 2,$$

   (a) show that it is a member of the exponential family and state the natural parameter;

   (b) using Question 2, or otherwise, find $E[\log(Y)]$ and $\text{Var}[\log(Y)]$; and

   (c) find $E(Y)$ and $\text{Var}(Y)$.

4. [**9 marks**]
The data in **parasite.csv** concern the prevalence of a parasite infection in villagers across 181 villages. Data about the village locations are recorded, such as their ELEVATION (in metres above sea level) and LATITUDE and LONGITUDE (both recorded in degrees) as well as an average status of plant health around the village MEAN.NDVI (where NDVI is an index that ranges from -1 to 1, with higher values indicating more greenness in plants). The number of villagers that were examined is recorded as NUM.EXAM and the number of positive test results as NUM.INF .

(a) Create a new variable, which is the proportion of villagers infected by the parasite, and describe its distribution by using appropriate numerical and graphical summaries. Apply a transformation if appropriate.

(b) Treating the (transformed) variable created in (a) as the response (dependent) variable, fit a kitchen-sink linear model that includes all available explanatory variables (but excludes those variables used in (a)). Make 2-3 relevant statements about the model parameter estimates and goodness-of-fit.

(c) If any covariates are not significant in your kitchen-sink model, remove them and refit the model. Comment on any changes that you notice in the output. Check the model assumptions using appropriate plots of the residuals and make conclusions based on these.

(d) **Audio question**: You are meeting with the chief investigator of this project who has a background in parasitology but not in statistics. Briefly explain to them how you would interpret the final model and how confident you feel about this model adequately describing the data.

5. [**9 marks**]
The dataset **cholest.csv** documents the incidence of disease for, and cholesterol level and gender of 100 subjects.

(a) Plot the data with gender identified using variable *genderS*.

(b) Fit a logistic model with main effects for gender and cholesterol levels *and their interaction*. Comment on the fitted interaction term. Remove any effects that are insignificant and add the fitted line from the final model to the plot produced in (a)

(c) Fit an appropriate probit model to the data. Write down an expression for this fitted model.

(d) Add the fitted probit mean curve to the graph and make a statement about their difference.

(e) **Audio question**: Briefly comment on the goodness-of-fit of the probit model in comparison to the logistic model and on which model you prefer in explaining this data and why.