UNIVERSITY OF NEW SOUTH WALES
SCHOOL OF MATHEMATICS AND STATISTICS

MATH3821 Statistical Modelling and Computing
Term Two 2020

Assignment One

Given: 19th June 2020 **Due date: 12 pm (midday) Friday 3rd July 2020**

Number of exercises: 5 (one per page)

INSTRUCTIONS: This assignment is to be done by a group of **at most 5** students. The same mark will be given to each student within the group, unless I have good reasons to believe that somebody did not contribute appropriately. It is **strongly** advised that you use the RStudio software and its File/New File/R Markdown. . . /PDF capability to produce a PDF file that you will submit on Moodle (see instructions on Moodle close to due date). The computing language you will be using is called RMarkdown (see the first few lessons starting here https://rmarkdown.rstudio.com/lesson-1.html for a quick introduction). For typesetting mathematical formulae, you will need to have a distribution of the LaTeX software installed on your computer (e.g., TeX Live or MikTeX). For Microsoft Windows and Unix users, you might consider using the `install_tinytex()` function from the R package `tinytex`. Another function (Microsoft Windows only) is `install.MikTeX()` from the `installr` R package. MacOS users should consider installing the `MacTeX` software directly; this is **not** an R package (see https://www.tug.org/mactex/).

Only one of the five students should submit the PDF file, with the names of the other students in the group clearly indicated in the document.

We declare that this assessment item is our own work, except where acknowledged, and has not been submitted for academic credit elsewhere. We acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking). We certify that we have read and understood the University Rules in respect of Student Academic Misconduct.

Name Student No Signature Date

Question One

Recall the Simple Linear Regression (SLR) model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

(a) Show that the SLR model can be expressed in the following form

$$Y_i = \alpha + \beta_1(x_i - \bar{x}) + \epsilon_i.$$

(b) Provide an interpretation for the parameter α.

(c) Find a closed form formula for the least square parameter estimates $\hat{\alpha}$ and $\hat{\beta}_1$.

(d) Find the variance of the estimates $\hat{\alpha}$ and $\hat{\beta}_1$ and the covariance between them $\text{Cov}(\hat{\alpha}, \hat{\beta}_1)$.

(e) Using the method of gradient descent with $\gamma = 0.00001$, find the estimates for the following simulated data:

```
set.seed(1234567)
x = runif(1000)
eps = rnorm(1000)
y = 5 + 10*x + eps
```

Start the algorithm at the initial value $(\alpha^{[0]}, \beta_1^{[0]}) = (0, 0)$. Use the convergence criteria that the L2 norm of the score is less than 0.00001. Show that the results are comparable to the closed formed solutions found in (c). Report the number of iterations required. Make sure that you provide all the workings/derivations that are needed to implement the above algorithm.

(f) Plot the data and include the fitted regression line.

(g) Using the Newton-Raphson method find the estimates for the same simulated data. Use the same initial value $(\alpha^{[0]}, \beta_1^{[0]}) = (0, 0)$ and convergence criteria as (e). Did it take more or less iterations than part (e). Why? Make sure that you provide all the workings/derivations that are needed to implement the above algorithm.

Quesion Two

Consider n independent binary random variables $Y_1, ..., Y_n$ with

$$P(Y_i = 1) = \pi_i \quad \text{and} \quad P(Y_i = 0) = 1 - \pi_i.$$

The probability function of Y_i is:

$$\pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

where $Y_i = 0$ or 1.

(a) Show that this probability function belongs to the exponential family of distributions.

(b) Show that the natural parameter is

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right).$$

(c) Show that $\mathrm{E}(Y_i) = \pi_i$ using the cumulant generator $c(\theta)$ in the definition of the exponential family.

(d) Suppose the link function is

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = x^T\beta.$$

Show that this is equivalent to modelling the probability π as

$$\pi = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}.$$

(e) Sketch the graph of π aganist x for the particular case $x^T\beta = \beta_1 + \beta_2 x$ where β_1 and β_2 are constants. How would you interpret this graph if x is the dose of an insecticide and π is the probability of an insect dying?

(f) Does the following probability density function

$$f(y;\theta) = \frac{1}{\phi}\exp\left\{\frac{(y-\theta)}{\phi} - \exp\left[\frac{(y-\theta)}{\phi}\right]\right\}$$

where $\phi > 0$ is regarded as a nuisance parameters, belong to the exponential family?

Question Three

The Titanic was a British luxury passenger liner that sank when it struck an iceberg about 640 km south of Newfoundland on April 14–15, 1912, on its maiden voyage to New York City from Southampton, England. The data in the file `titanic.txt` **(from the assignment section on Moodle!)** classify the people on board the ship according to their `Sex`, `Age`, and `Class`, either first, second, third.

(a) Read the file `titanic.txt` (see Moodle) into a variable called `titanic`. Display the first six lines of `titanic` and then provide a summary of the variables in the dataset using `summary`.

(b) Compute the number of men and women on the Titanic. Calculate the survivial rates for each sex. Conduct a test which tests whether the survivial rates for men and women are the same aganist the alternative that they are different. What is the hypothesis, test statistic, p-value and conclusion from the test?

(c) Fit a logistic regression model with response `Survived` and predictor `Age`, and provide an interpretation for the fitted coefficient for `Age` using the odds ratio with a factor change and a standardized factor change in the variable `Age`.

(d) Plot the graph of `Survived` versus `Age`. Then add both a fitted logistic curve and a loess smoother to the graph. Explain what the differences are betwen these two fits. Fit again, but this time, add a quadratic term in `Age`. Does the fitted curve now match the smoother more accurately? Provide all plots in a single graph, with correctly defined labels, titles and a legend.

(e) Use the method of scoring algorithm to compute an estimate of the parameters of the logistic regression model with survived as the response and age and a quadratic term in `age` as explantory variables and provide your `R` code for it. You must also present the calculations that you used to come up with your algorithm.

(f) Check that, using the code in (e), you obtain estimates of the coefficients numerically close to the ones given by the `glm()` function.

(g) Create an R code, and provide it, to compute the estimation of the variances-covariances matrix of the corresponding estimators (using the first approach presented in the slide entitled "Estimation of the variance" in Chapter 2).

(h) Check the numerical closeness of the result obtained using your code from (g) to the one you get when using the `vcov()` function.

(i) Fit the logistic regression model with terms for an intercept, `Age`, Age^2, `Sex`, and `PClass`. Obtain tests on the basis of the deviance for adding each of the terms to a mean function that already includes the other terms (in the order given above), and summarize the results of each of the tests via a p-value and a one-sentence summary of the results.

(j) Provide a plot that interprets the relationship between `Age`, `Sex` and their `Survival` rates. Make sure that you include titles with a legend.

Question Four

In this question we will examine binomial response data. Consider the single response Y with $Y \sim binomial(n, \pi)$.

(a) Find the Wald statistic $(\hat{\pi} - \pi)I(\pi)(\hat{\pi} - \pi)$ where $\hat{\pi}$ is the maximum likelihood estimator of π and $I(\pi)$ is the information.

(b) Verify that the Wald statistic is the same as the score statistic $U^\top I(\pi)^{-1}U$ in this case.

(c) Find the deviance

$$2\big[\log L(\hat{\pi}; y) - \log L(\pi; y)\big].$$

(d) For large samples, both the Wald/score statistic and the deviance approximately have the $\chi^2(1)$ distribution. For $n = 10$ and $y = 3$ use both statistics to assess the adequacy of the models:

(i) $\pi = 0.1$;
(ii) $\pi = 0.3$;
(iii) $\pi = 0.5$. Do the two statistics lead to the same conclusions.

(e) Give the three parts of the GLM for the binomial regression model with a fixed number of trials:

- state the law of Y;
- prove it is a member of the exponential family;
- give the parameters (notably the mean μ_i) and the canonical link function.

Question Five

In this question, we will analyse the following data set, which contain the quaterly number of cases of AIDS in Australia for successive quarter from 1984 to 1988.

	Quarter			
Year	1	2	3	4
1984	1	6	16	23
1985	27	39	31	30
1986	43	51	63	70
1987	88	97	91	104
1988	110	113	149	159

(a) Provide a boxplot of the number of cases by quater and then another boxplot of the number of cases by year. Comment on the trend that is evident in those plots.

(b) Conduct a test based on the reduction in deviance to assess whether an interaction term is required when fitting a Poisson regression model with `year` and `quarter` as predictors. What is the conclusion.

(c) For the appropriate model in part (b) use the `summary` and `anova` functions to comment on the suitability of each predictor in the model.

(d) Compute the expected the number of cases in 1987 for the third quarter and also, the expected number of cases in 1987 for the fourth quarter. By what percent has the expected number of cases changed?