FIT2086 Assignment 2

Due Date: 11:55PM, Sunday, 29/9/2019

1 Introduction

There are total of four questions worth $10 + 10 + 9 + 10 = 39$ marks in this assignment. There is one bonus question worth an additional 2 marks. The total marks awarded will be capped at 39, but the bonus marks can compensate for marks lost in the four compulsory questions.

This assignment is worth a total of 20% of your final mark, subject to hurdles and any other matters (e.g., late penalties, special consideration, etc.) as specified in the FIT2086 Unit Guide or elsewhere in the FIT2086 Moodle site (including Faculty of I.T. and Monash University policies).

Students are reminded of the Academic Integrity Awareness Training Tutorial Activity and, in particular, of Monash University's policies on academic integrity. In submitting this assignment, you acknowledge your awareness of Monash University's policies on academic integrity and that work is done and submitted in accordance with these policies.

Submission Instructions: Please follow these submission instructions:

1. No files are to be submitted via e-mail. Correct files are to be submitted to Moodle, as given above.

2. Please provide a single file containing your report, i.e., your answers to these questions. Provide code/code fragments as required in your report, and make sure the code is written in a fixed width font such as Courier New, or similar, and is grouped with the question the code is answering. You can submit hand-written answers, but if you do, please make sure they are clear and legible. Do not submit multiple files for the written component of the assignment – all your files should be combined into a single PDF file as required. Please ensure that the written component of your assignment answers the questions in the order specified in the assignment. Multiple files and questions out of order make the life of the tutors marking your assignment much more difficult than it needs to be, so please **ensure you assignment follows these requirements**.

3. If you are completing the bonus question then please ZIP the PDF of your written answers along with your CSV of predictions and submit this single ZIP file. Please read these submission instructions carefully and take care to submit the correct files in the correct places.

Question 1 (10 marks)

It was believed for a long time by medical practitioners that the full moon influenced the expression of medical conditions including fevers, rheumatism, epilepsy and bipolar disorder – in fact, the antiquated term "lunatic" derives from the word lunar, i.e., of the moon. In the late 1990's a (tongue in cheek) study was undertaken to test if the full moon induced dogs to become more aggressive, with a resulting increased likelihood of biting people. In addition to being a little bit of fun, examining a problem like this through the lense of data science is an instructive example on how quantitative methods can be used to answer "folk-lore" questions/hypotheses.

The file `dogbites.fullmoon.csv` contains the daily number of admissions to hospital of people being bitten by dogs from 13th of June, 1997 through to 30th of June, 1998[1]. It also contains a second column indicating whether the day in question was a full moon or not. Use this data to answer the following questions. We know from Assignment 1 that the Poisson distribution is not a good fit to the daily dog-bite data: instead, for this question we will use a normal distribution as it provides an improved fit to the data due to its increased flexibility, while accepting this assumption is also not necessarily correct; to quote the famous statistician G.E.P.Box: "*all models are wrong – but some are more useful than others*".

Important: you may use R to determine the means and variances of the data, as required, and the R functions `pt()` and `pnorm()` but you must perform all the remaining steps by hand. Please provide appropriate R code fragments and all working out.

1. Calculate an estimate of the average number of dog-bites for days on which there was a full moon. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. **[4 marks]**

2. Researchers asked the question: do dogs bite more on the full moon? Using the provided data and the approximate method for difference in means with unknown variances presented in Lecture 4, calculate the estimated mean difference in mean dog bite occurences between full moon days and non-full moon days, and a 95% confidence interval for this difference. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

3. Test the hypothesis that dogs bite more frequently on full moon days than on non-full moon days. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate hypothesis test for differences in means with unknown variances presented in Lecture 5. What does this p-value suggest about the behaviour of dogs on full moon days *vs* non-full moon days? Show working as required. **[3 marks]**

[1] Data source is taken from the Australian Institute of Health and Welfare Database of Australian Hospital Statistics.

Question 2 (10 marks)

The exponential distribution is a probability distribution for non-negative real numbers. It is often used to model waiting or survival times. The version that we will look at has a probability density function of the form

$$p(y \mid v) = \exp\left(-e^{-v} y - v\right) \tag{1}$$

where $y \in \mathbb{R}_+$, i.e., y can take on the values of non-negative real numbers. In this form it has one parameters: a log-scale parameter v. If a random variable follows a gamma distribution with log-scale v we say that $Y \sim \mathrm{Exp}(v)$. If $Y \sim \mathrm{Exp}(v)$, then $\mathbb{E}[Y] = e^v$ and $\mathbb{V}[Y] = e^{2v}$.

1. Produce a plot of the exponential probability density function (1) for the values $y \in (0, 10)$, for $v = 1$, $v = 0.5$ and $v = 2$. Ensure the graph is readable, the axis are labeled appropriately and a legend is included. **[2 marks]**

2. Imagine we are given a sample of n observations $\mathbf{y} = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from an exponential distribution with log-scale parameter v (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. *(hint: remember that these samples are independent and identically distributed.)* **[2 marks]**

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data $\mathbf{y}$ under the exponential model with log-scale v. Simplify this expression. **[1 mark]**

4. Derive the maximum likelihood estimator $\hat{v}$ for v. That is, find the value of v that minimises the negative log-likelihood. You must provide working. **[2 marks]**

5. Determine the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution. *(hints: utilise techniques from Lecture 2, Slide 21 and the mean/variance of the sample mean)* **[3 marks]**

Question 3 (9 marks)

It is frequent in nature that animals express certain asymmetries in their behaviour patterns. It has been suggested that this might be nature's way of "breaking gridlocks" that might occur if we were to act purely rationally (think: why does a beetle decide to move one way over another when put in a featureless bowl?). An interesting observational study, undertaken by a European researcher in 2003 examined the head tilting preferences of humans when kissing.

The data was collected by observing kissing couples of age ranging from 13 to 70 in public places (mostly airports and train stations) in the United States, Germany and Turkey. The observational data found that of 124 kissing pairs, 80 turned their heads to the right and 44 turned their heads to the left.

You must analyse this data to see if there is an inbuilt preference in humans for the direction of head tilt when kissing. Provide working, reasoning or explanations and R commands that you have used, as appropriate.

1. Calculate an estimate of the preference for humans turning their heads to the right when kissing using the above data, and provide an approximate 95% confidence interval for this estimate. Summarise/describe your results appropriately. **[3 marks]**

2. Test the hypothesis that there is a preference in humans for tilting their head to one particular side when kissing. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? **[2 marks]**

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. **[1 mark]**

4. It is entirely possible that any preference for head turning to the right/left could be simply a product of right/left-handedness. To test this we obtain handedness of a sample of different people. It was found that 83 people were right-handed and 17 were left handed. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the rate of right-handedness in the population is the same as the preference for turning heads to the right when kissing this data. Summarise your findings. What does the p-value suggest? **[2 marks]**

5. Can you identify any possible problems with your conclusions based on the way in which the data was collected? Could there be alternative reasons for preference/lack of preference? **[1 mark]**

Question 4 (10 marks)

This question will require you to analyse a regression dataset. In particular, you will be looking at predicting the fuel efficiency of a car (in kilometers per litre) based on characteristics of the car and its engine. This is clearly an important and useful problem. The dataset `fuel2017-20.csv` contains $n = 2,000$ observations on $p = 9$ predictors obtained from actual fuel efficiency tables for car models available for sale during the years 2017 through to 2020. The target is the fuel efficiency of the car measured in kilometers per litre. The higher this score, the better the fuel efficiency of the car. The data dictionary for this dataset is given in Table 1. Provide working/R code/justifications for each of these questions as required.

1. Fit a multiple linear model to the fuel efficiency data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with fuel efficiency, and why? Which three variables appear to be the strongest predictors of fuel efficiency, and why? **[2 marks]**

2. Would your assessment of which predictors are associated change if you used the Bonferroni procedure with $\alpha = 0.05$? **[1 marks]**

3. Describe what effect the year of manufacture (`Model.Year`) appears to have on the mean fuel efficiency. Describe the effect that the number of gears (`No.Gears`) variable has on the mean fuel efficiency of the car. **[2 marks]**

4. Use the stepwise selection procedure with the BIC penalty to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning. **[1 mark]**

5. If we wanted to improve the fuel efficiency of our car, what does this BIC model suggest we could do? **[2 marks]**

6. Imagine that you are looking for a new car to buy to replace your existing car. Load the dataset `fuel2017-20.test.csv`. The characteristics of the new car that you are looking at are given by the first row of this dataset.

   (a) Use your BIC model to predict the mean fuel efficiency for this new car. Provide a 95% confidence interval for this prediction. **[1 mark]**

   (b) The current car that you own has a mean fuel efficiency of $8.5 km/l$ (measured over the life time of your ownership). Does your model suggest that the new car will have better fuel efficiency than your current car? **[1 mark]**

Bonus Question – challenge (2 marks)

Explore the fuel efficiency data further and try to build a better linear model for the fuel efficiency of a car. You could try using techniques such as interactions or other nonlinear transformations of the variables or even the target to see if you can improve your model of fuel efficiency. For this assignment, please restrict yourself to linear regression models as these provide an interpretability not available to other methods such as random forests. To obtain these extra marks you should write a short report (one page maximum) detailing the methods and models that you tried, the R commands that you used and your reasoning for including/removing various predictors or transformations of predictors, and what the resulting model suggests about fuel efficiency.

Additionally, once you have found a model that you think is the best, load the `fuel2017-20.test.csv` dataset which contains the explanatory variables for $2,352$ new cars, but is missing associated values of `Comb.FE`; use your best model to predict the fuel efficiency for each of the $2,352$ suburbs in this dataset and write your predicted fuel efficiency to a CSV file called `fuel.predictions.yourID.csv`, where `yourID` is your student ID number. To do this, use the `write.csv()` function in R. Submit this file along with your assignment. After all the assignments are submitted I will calculate prediction errors for all the people that have submitted predictions, and we will discuss briefly in class which models predicted well and why. See if you can win the FIT2086 data prediction challenge! :) *(note that the awarding of marks is not connected to how well the final model predicts – rather it is based on the things you tried and the discussion of your analysis)* **[2 marks]**

Variable name	**Description**	**Values**
`Model.Year`	Year of sale	$2017 - 2020$
`Eng.Displacement`	Engine Displacement (litres, l)	$0.9 - 8.4$
`No.Cylinders`	Number of Cylinders	$3 - 16$
`Aspiration`	Engine Aspiration (Oxygen intake)	`N`: Naturally* `OT`: Other `SC`: Supercharged `TC`: Turbocharged `TS`: Turbo+supercharged
`No.Gears`	Number of Gears	$1 - 10$
`Lockup.Torque.Converter`	Lockup torque converter present?	`N`* and `Y`
`Drive.Sys`	Drive System	`4`*: 4-wheel drive `A`:All-wheel `F`:Front-wheel `P`:Part-time 4-wheel `R`:Rear-wheel
`Max.Ethanol`	Maximum % of Ethanol allowed	$10 - 85$
`Fuel.Type`	Type of Fuel	`G`*: Regular Unleaded `GM`: Mid-grade Unleaded Recommended `GP`: Premium Unleaded Recommended `GPR`: Premium Unleaded Required
`Comb.FE`	Fuel Efficiency (km/l)	$4.974 - 26.224$

Table 1: Fuel efficiency data dictionary. The * denotes the reference category for each categorical variable.