FIT2086 Assignment 2

Due Date: 11:55PM, Friday, 21/9/2018

1 Introduction

There are total of four questions worth $10 + 9 + 7 + 10 = 36$ marks in this assignment. There is one bonus question worth an additional 4 marks. The total marks awarded will be capped at 36, 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

The Charles River is a major feature of the city of Boston, but it only runs through a small number of the suburbs in the city. A real estate agent would like to evaluate the effect of the presence of the Charles River within a suburb on its median house price. To accomplish this, the agent collected the median house prices from eight suburbs which have houses along the river, and eight suburbs which do not. The agent has asked us (as data scientists) for help with this, and we have decided that we will compare the average of the median house prices in each set (the "average median house price"), using statistical tests to determine whether each set might have the same average or not. The median house prices collected on those suburbs which do not have houses along the Charles river were:

$$\mathbf{y}_n = (270.5, 230.9, 180.2, 230.3, 210.6, 200.6, 160.8, 220.9)$$

1. Calculate an estimate of the average median house price for suburbs of Boston which do not have houses on the Charles river. 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. The same real estate agency has also collected median houses prices for eight suburbs that do have houses on the Charles river. These are:

   $$\mathbf{y}_c = (290.0, 500.0, 500.0, 210.7, 130.4, 330.1, 230.3, 150.3)$$

   The real estate agents want to know if there is a difference, at the population level, between average median houses prices in suburbs that do, and do not, have houses on the Charles river. Use this sample to answer this question. Using the approximate method for difference in means with unknown variances presented in Lecture 4, calculate the estimated mean difference in median house price between the suburbs with and without houses on the Charles River, and a 95% confidence interval for this difference. Summarise/describe your results appropriately. Show working as required. **[4 marks]**

3. Test the hypothesis that the two groups are the same. Using the approximate hypothesis test for difference in means with unknown variances presented in Lecture 5, calculate an appropriate p-value under the null hypothesis that the average median house price for suburbs with and without houses on the Charles River are the same, showing working as required. Interpret this p-value; do you think the two groups of suburbs (with and without houses on the Charles River) have the different average median house price at the population level? **[2 marks]**

Question 2

The gamma 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

$$p(y \mid \mu, k) = \left(\frac{k}{\mu}\right)^k \left(\frac{1}{(k-1)!}\right) y^{k-1} \exp\left(-\frac{k\,y}{\mu}\right) \tag{1}$$

where $y \in \mathbb{R}_+$, i.e., y can take on the values of non-negative real numbers. In this form it has two parameters: a mean parameter μ, and a shape parameter k. If a random variable follows a gamma distribution with mean μ and shape k we say that $Y \sim \text{Ga}(\mu, k)$. If $Y \sim \text{Ga}(\mu, k)$, then $\mathbb{E}[Y] = \mu$ and $\mathbb{V}[Y] = \mu^2/k$.

1. Produce a plot of the gamma probability density function (1) for the values $y \in (0, 10)$, for ($k = 1$, $\mu = 1$), ($k = 2$, $\mu = 1$) and ($k = 2$, $\mu = 2$). 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 a gamma distribution with mean parameter μ and shape parameter k (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.)* **[1 mark]**

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data $\mathbf{y}$ under the gamma model with mean parameter μ and shape parameter k. Simplify this expression. **[1 mark]**

4. Derive the maximum likelihood estimator $\hat{\mu}$ for μ, under the assumption that k is known. That is, find the value of μ that minimises the negative log-likelihood with k assumed to be an arbitrary, known constant. You must provide working. **[2 marks]**

5. What is the bias and variance of the maximum likelihood estimator $\hat{\mu}$ of μ for the gamma distribution, assuming that the population value of k is known? Explain how you obtained your answer. **[1 mark]**

6. So far we have treated k as known. We can also estimate this by using the method of maximum likelihood, although we cannot do it by the usual procedure as no closed form solution exists. However, as k is an integer in our setting, we can instead use R to search for it by trying all values of $k = 1, \ldots, 100$ and searching for the one that minimises the negative log-likelihood of the data. Implement this idea in a function `gamma.ml(y)` that takes a data vector `y` and returns the maximum likelihood estimates $\hat{\mu}$ and $\hat{k}$. Once you have coded this, run your code on the data

   $$\mathbf{y} = (4.81, 4.28, 7.04, 2.37, 7.30, 3.66, 2.33, 6.38)$$

   and report the values of the maximum likelihood estimates for μ and k. **[2 marks]**

Tournament	Opponent	Scored/Taken	Conceded/Faced
WC 1982	France	5/6	4/6
WC 1986	Mexico	4/4	1/3
WC 1990	England	4/4	3/5
WC 2006	Argentina	4/4	2/4

Table 1: Penalty shootouts involving the German national team at the World Cup.

Question 3

In the game of football (soccer) a penalty shoot-out is frequently used to determine the outcome of a drawn game, particularly in major tournaments. In a penalty shoot-out, the individual players from the two opposing teams repeatedly take turns at trying to score a goal past a goalkeeper from the penalty-spot ("converting a penalty"), which is 11 meters away from the goal-line. The penalty shoot-out begins with both teams having five attempts to convert penalties. If, by the conclusion of this first phase, one team converts more penalties than the other – or one team gets so far ahead in penalties scored that the other team cannot catch up – they are deemed the winner of the match. If both teams are drawn on the same number of penalties scored at the end of the first phase, the penalty shoot-out then switches to a "sudden death" format, in which the first team to fall behind in penalties scored is immediately deemed to lose the match.

During the history of the World Cup, the German national football team has been involved in four penalty shoot-outs, and has won all four. The data, recording the number of penalties taken and successfully scored in each shoot-out, as well as the number of penalties faced and conceded (i.e., scored by their opponents), is summarised in Table 1. The ability of the German national team to convert penalties is an "established" part of the world-cup folklore. By analysing the data we can try and assess this in a more formal fashion. Provide working, reasoning or explanations and R commands that you have used, as appropriate.

1. Calculate an estimate of the German national team's success rate at converting penalties in World Cup penalty shoot-outs. **[1 mark]**

2. The average rate of penalty conversion across all games at the world cup is 71%. Using hypothesis testing, test the hypothesis that the German national team has a penalty conversion rate that is better than the world cup average. Write down explicitly the hypothesis that 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. Part of winning a penalty shoot-out is denying your opponent from scoring penalties. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the German penalty conversion rate is different to the penalty conversion rate of their opponents – at least in shoot-outs against Germany – using the data provided in Table 1. Summarise your findings. What does the p-value suggest? **[2 marks]**

5. Can you identify any possible problems with the way in which the penalty data is sampled that might introduce some biases into your analysis? **[1 marks]**

Question 4

This question will require you to analyse a regression dataset. In particular, you will be looking at trying to predict the compressive strength of concrete from various measurements of the various components used in the concrete mixture. Obviously this is an extremely important problem as concrete is the single most important material in civil engineering and construction. The file `concrete.csv` contains the data you will be analysing. There are $n = 250$ observations on $p = 8$ predictors, seven of which measure the amount of various component substances within the concrete mixture. The target is the compressive strength of the resulting concrete mixture in megapascals. The higher the compressive strength, the better the concrete mixture is. The data dictionary for this dataset is given in Table 2. Provide working/R code/justifications for each of these questions as required.

1. Fit a multiple linear model to the concrete data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with compressive strength, and why? Which three variables appear to be the strongest predictors of compressive strength, 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 water (`Water`) in the concrete mix appears to have on the mean compressive strength. Describe the effect that the `Age` variable has on the mean compressive strength of the concrete. **[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 strength of our concrete, what does this model suggest we could do? Can we use this model, as it stands, to find the "optimal" mixture? **[2 marks]**

6. Imagine that a civil engineer proposes to use a new mix of concrete for a project with the mixture given in Table 3. The engineer asks you to predict the mean compressive strength of this new concrete mix after it has set for 28 days.

   (a) Use your model to predict the mean compressive strength for this mix. Provide a 95% confidence interval for this prediction. **[1 mark]**

   (b) The current mix of concrete the engineer is using has a mean compressive strength of $52.35mPa$ after setting for 28 days. Does your model suggest that the newly proposed mix is better than the current mix? **[1 mark]**

Variable name	Description	Values
`Cement`	Cement (kg in a m^3 mixture)	$139.6 - 540$
`Blast.Furnace.Slag`	Blast furnace slag (kg in a m^3 mixture)	$0 - 282.8$
`Fly.Ash`	Fly ash (kg in a m^3 mixture)	$0 - 163.8$
`Water`	Water (kg in a m^3 mixture)	$121.8 - 228$
`Superplasticizer`	Superplasticizer (kg in a m^3 mixture)	$0 - 32$
`Coarse.Aggregate`	Coarse aggregate (kg in a m^3 mixture)	$852.1 - 1134.3$
`Fine.Aggregate`	Fine aggregate (kg in a m^3 mixture)	$594 - 992.6$
`Age`	Age of concrete (days since pour)	$3 - 365$
`Strength`	Compressive strength of concrete (in MPa)	$7.75 - 82.6$

Table 2: Concrete Compressive Strength Data Dictionary.

Variable	Cement	Blast.Furnace.Slag	Fly.Ash	Water	Superplasticizer	Coarse.Aggregate	Fine.Aggregate	Age
Value	491	26	123	210	3.9	882	699	28

Table 3: Example Concrete Mix.

Bonus Question – challenge

Explore the concrete data further and try to build a better model for the compressive strength. You could try using techniques such as interactions or nonlinear transformations of the variables to see if you can improve your model of compressive strength of concrete – or you could use another data science technique altogether! To obtain these extra marks you should write a short report (around one page) detailing the methods that you tried, the R commands that you used and your reasoning for including/removing various predictors or transformations of predictors.

Additionally, once you have found a model that you think is the best, load the `concrete.test.csv` dataset which contains the explanatory variables for 780 new concrete mixes, but is missing associated values of `Strength`; use your best model to predict compressive strengths for each of the 780 suburbs in this dataset and write your predicted compressive strength to a CSV file called `concrete.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)* **[4 marks]**