FIT2086 Assignment 2

Due Date: 6PM, Thursday, 21/9/2017

1 Introduction

There are total of four questions worth $10 + 7 + 5 + 8 = 30$ marks in this assignment. There is one bonus question worth an additional 4 marks. The total marks awarded will be capped at 30, 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: No files are to be submitted via e-mail. Correct files are to be submitted to Moodle, as given above. You can submit hand-written answers, but if you do, then please make sure it is clear and legible. Do not submit multiple files for the written component of the assignment – these should all be bundled into a single PDF file. 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.

2 Questions

1. A researcher has collected blood pressure measurements from a small sample of Pima people who have diabetes. The recorded blood pressure measurements (measured in millimeters of mercury, $mmHg$) were:

   $$\mathbf{y}_d = (72, 64, 40, 50, 70, 96, 74, 60)$$

   A previous study, based on a very large number of individuals, has found the population standard deviation of blood pressure of Pima people with diabetes to be $\sigma_d = 15.23mmHg$ (i.e., we can treat the population standard deviation for Pima people with diabetes as known).

   (a) Calculate an estimate of the mean blood pressure for Pima people with diabetes. Calculate a 95% confidence interval for this estimate, and summarise/describe your results appropriately. Show working as required. **[4 marks]**

(b) The same research has also collected blood pressure measurements on a sample of Pima people who do not have diabetes. The recorded blood pressures for this sample were:

$$\mathbf{y}_n = (66, 66, 74, 35, 92, 80, 30, 88, 84, 66)$$

A previous study, based on a very large number of individuals, has found the population standard deviation of blood pressure of Pima people without diabetes to be $\sigma_n = 14.69 mmHg$ (i.e., we can treat the population standard deviation for Pima people without diabetes as known).

The researchers want to know if there is a difference at the population level in blood pressure between Pima people with and without diabetes. Use this sample to answer this question. Calculate the estimated mean difference in blood pressure between the Pima people with and without diabetes, and a confidence interval for this difference. Summarise/describe your results appropriately. Show working as required. **[4 marks]**

(c) Test the hypothesis that the two groups are the same. Calculate an appropriate p-value for the null hypothesis that the population means of diabetic and non-diabetic Pima people are the same, showing working as required. Interpret this p-value; do you think the two groups (diabetics and non-diabetics) have the same blood pressure at the population level? **[2 marks]**

2. The geometric distribution is a probability distribution for counts. It models the number of tosses of a coin required to see a success, and is usually parameterised in terms of the success probability θ of the coin. However, it is also possible to reparameterise it in terms of the mean number of throws needed. In this form we have

$$p(y \,|\, \mu) = \left(\frac{\mu}{1+\mu}\right)^y \left(\frac{1}{1+\mu}\right) \tag{1}$$

where $y \in \{0, 1, 2, \ldots, \}$, i.e., y can take on the values of non-negative integers. If a random variable follows a geometric distribution with mean μ we say that $Y \sim \text{Geo}(\mu)$. If $Y \sim \text{Geo}(\mu)$, then $\mathbb{E}[Y] = \mu$ and $\mathbb{V}[Y] = \mu(1+\mu)$.

(a) Produce a plot of the geometric probability distribution (1) for the values $y = 0, 1, \ldots, 10$, for $\mu = 1$, $\mu = 2$ and $\mu = 4$. Ensure the graph is readable, the axis are labelled appropriately and a legend is included. **[2 marks]**

(b) Imagine we are given a sample of n counts $\mathbf{y} = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from a geometric distribution with mean parameter μ (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]**

(c) Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data $\mathbf{y}$ under the geometric model with mean parameter μ. **[1 mark]**

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

(e) What is the bias and variance of the maximum likelihood estimator $\hat{\mu}$ of μ for the geometric distribution? Explain how you obtain your answer. **[1 mark]**

Tournament	Stage	Team 1	Team 2	Prediction	Correct?
Euro 2008	Group stage	Germany	Poland	Germany	✓
Euro 2008	Group stage	Germany	Croatia	Germany	✗
Euro 2008	Group stage	Germany	Austria	Germany	✓
Euro 2008	Quarter-finals	Germany	Portugal	Germany	✓
Euro 2008	Semi-finals	Germany	Turkey	Germany	✓
Euro 2008	Final	Germany	Spain	Germany	✗
World Cup 2010	Group stage	Germany	Australia	Germany	✓
World Cup 2010	Group stage	Germany	Serbia	Serbia	✓
World Cup 2010	Group stage	Germany	Ghana	Germany	✓
World Cup 2010	Round of 16	Germany	England	Germany	✓
World Cup 2010	Quarter-finals	Germany	Argentina	Germany	✓
World Cup 2010	Semi-finals	Germany	Spain	Spain	✓
World Cup 2010	3rd-4th playoff	Germany	Uruguay	Germany	✓
World Cup 2010	Final	Spain	Netherlands	Spain	✓

Table 1: Predictions made by Paul the Octopus at two major international football tournaments.

3. During the European football championships in 2008, and the football World Cup in 2010, an octopus called Paul living at an aquarium in Oberhausen, Germany, was used to predict the outcome of football matches, mostly involving the German national football team. To obtain Pauls' predictions, his keepers at the aquarium would present him with two boxes of food before each match. Each box was covered in the flag of the two nations that were participating, and the box that Paul chose to feed from first determined which nation he predicted would win.

   Paul was asked to predict the outcome of 14 matches, 12 of which involved Germany. He correctly predicted the outcomes of 12 matches, only incorrectly guessing that Germany would beat Croatia in the Euro 2008 group stage, and that Germany would beat Spain in the Euro 2008 final. The data regarding his predictions is presented in Table 1. Some people claimed he was an "animal oracle".

   Answer the questions below regarding the claim that Paul the Octopus is an oracle. Provide working, reasoning or explanations and R commands that you have used, as appropriate.

   (a) Calculate an estimate of Paul's success rate at predicting football matches. **[1 mark]**

   (b) Using hypothesis testing, test the hypothesis that Paul just "got lucky" and was randomly guessing the outcomes of the football matches. 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]**

   (c) 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]**

   (d) Given your analysis, do you think that Paul is an oracle (can see the future!), or can you identify any potential weaknesses in the experimental setup or way in which the data was sampled? **[1 marks]**

4. The file `housing.csv` contains the data that we will use for this question. This dataset is a modified version of the Boston housing data which was collected to study house prices in the metropolitan region of Boston. In this data set, each observation represents a particular suburb from the Boston region. The outcome, `medv`, is the median value of owner-occupied homes in $1,000$ in the suburb. The variables are summarised in Table 2. We are interested in discovering which predictors are good determinants of housing price.

   (a) Fit a multiple linear model to the housing data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with median house value, and why? Which three variables appear to be the strongest predictors of housing price, and why? **[2 marks]**

   (b) Describe what effect the per-capita crime rate (`crim`) appears to have on the median house price. Describe the effect of a suburb having frontage on the Charles River has on the median house price for that suburb. **[2 marks]**

   (c) Use the stepwise selection procedure to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning. **[1 mark]**

   (d) If a council wanted to try and improve the median house value in their suburb, what does this model suggest they could try and do? **[2 marks]**

   (e) Table 3 gives the values of predictors for a new suburb. Use your model to predict the median house price for this suburb. **[1 mark]**

5. **Bonus Question – challenge**. Explore the housing data further and try to build a better model for the housing data. You should try using techniques such as interactions or nonlinear transformations of the variables to see if you can improve your model of housing prices. Remember to utilise p-values, R^2 values and model selection methods as part of this process. To obtain these extra marks you should write a short report (around one to one and a half pages) 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 `housing_test.csv` dataset which contains the explanatory variables for 106 new suburbs but is missing associated values of `medv`; use your best model to predict housing prices for each of these 106 suburbs in this dataset and write your predicted median house prices to a CSV file called `housing_predictions.csv` using 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 first 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]**

Variable name	Description	Values
`crim`	Per-capita crime rate	> 0
`zn`	Proportion of residential land zoned for lots over 25,000 sq. ft.	$0 - 100$
`indus`	Proportion of non-retail business acres per town	$0 - 100$
`chas`	Does the suburb front the Charles River?	0 = No, 1 = Yes
`nox`	Nitric oxides concentration (parts per 10 million)	> 0
`rm`	Average number of rooms per dwelling	≥ 1
`age`	Proportion of owner-occupied units built prior to 1940	$0 - 100$
`dis`	Weighted distances to five Boston employment centres	> 0
`rad`	Index of accessibility to radial highways	> 0
`tax`	Full-value property-tax rate per \$10,000	$187 - 711$
`ptratio`	Pupil-teacher ratio	> 0
`lstat`	Percentage of lower status of the population	$0 - 100$
`medv`	Median value of owner-occupied homes in $\$1,000$s	> 0

Table 2: Boston Housing Data Dictionary.

Variable	`crim`	`zn`	`indus`	`chas`	`nox`	`rm`	`age`	`dis`	`rad`	`tax`	`ptratio`	`lstat`
Value	0.04741	0	11.93	0	0.573	6.03	80.8	2.505	1	273	21	7.88

Table 3: Boston Housing Data Dictionary.