

MTHM502 Introduction to Data Science and Statistical Modelling

Assignment

1. The offspring of an animal can be split into three distinct types: type BB, type Bb and type bb. If the offspring is of type bb it has brown fur, otherwise it has black fur. Suppose that the initial proportion of type BB offspring is $1/8$, the initial proportion of type bb offspring is $1/4$, and the rest of the offspring are of type Bb. Suppose that any offspring of type BB, type Bb and type bb will independently survive to maturity with probability $1/4$, $1/8$ and $3/4$ respectively.
 - (a) [5 marks] Find the probability that a randomly chosen offspring survives to maturity.
 - (b) [7 marks] Given that an offspring has survived to maturity, find the probability that it has black fur.
2. We toss a coin which shows a head with probability p . If the coin shows a head, we toss it again, but if it shows a tail we stop (so the coin is tossed either once or twice).
 - (a) [4 marks] Find the probability that we get a head at the end of this experiment.
 - (b) [5 marks] Now we toss 10 coins, where each coin shows a head with probability p , independently of each other. Each coin which shows a head is tossed again. What is the probability mass function of the number of heads we have after the second round of tosses?
 - (c) [4 marks] Find the expected number of heads after the first round of tosses in Q2(b). How does this number change after the second round of tosses?
 - (d) [6 marks] **R:** Use simulation to find the probability that we have no heads at the end of the experiment in Q2(b) when $p = 0.3$.
3. The following data are the observed frequencies of the number of strikes y in a ship building company in the UK during 1948-1961:

No. of Strikes	0	1	2	3	4
Frequency	137	33	10	1	1

Thus $n = 182$ and $\sum_{i=1}^{182} y_i = 60$. Assume a Poisson model for these data with parameter λ .

- (a) [7 marks] Suggest a point estimator for the Poisson parameter λ , and obtain a standard error for this estimator. Next, use the provided data to calculate an approximate 95% confidence interval for λ . Does the confidence interval support the hypothesis at the 5%-level that $\lambda = 1$?
- (b) [8 marks] Suppose only those periods during which there was at least one strike are of interest (note that in this case the zero entries in the dataset are not considered!). An appropriate model here has the following probability mass function

$$f(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{(1 - e^{-\lambda}) y!}, \quad y = 1, 2, 3, \dots \text{ and } \lambda > 0.$$

Obtain the log-likelihood function for this case and show that the equation for the maximum likelihood estimator $\hat{\lambda}$ reduces to solving:

$$\frac{\hat{\lambda}}{1 - e^{-\hat{\lambda}}} = \frac{4}{3}.$$

That is, the maximum likelihood estimate is the root of the function

$$f(\lambda) = \frac{\hat{\lambda}}{1 - e^{-\hat{\lambda}}} - \frac{4}{3}. \quad (1)$$

Note that you're not expected to solve the above equation manually.

- (c) [5 marks] **R:** Plot the derivative of the loglikelihood function from Q3(b) and identify an interval of length one that contains the root of this function.
- (d) [6 marks] **R:** Use the `optimise` function to determine the maximum likelihood estimate $\hat{\lambda}$ to three decimal places. (Looking at the help page of `optimise` reveals that `optimise` needs a function to minimise, furthermore it has a `lower` and `upper` argument that specifies the two endpoints of the interval to be searched). The function that needs to be minimised in order to obtain the maximum likelihood estimate is the **absolute value of the function (1) from Q3(b)** (you need to define this function), while a suitable interval can be the one identified in Q3(c).
4. The random variable Y has a $\text{Binom}(n, p)$ distribution where n is known and $p \in [0, 1]$ is to be estimated.
- (a) [3 marks] Show that the method of moments estimate of p on observing $Y = y$ is $\hat{p} = \frac{y}{n}$.

From now on we shall write the method of moments estimator of p as $T = T(Y) = \frac{Y}{n}$.

- (b) [6 marks] Find the mean and variance of T . What are the bias and mean square error of T as an estimator of p ?
- (c) [4 marks] An investigator is interested in the parameter $\theta = p^2$ and proposes to use $T^2 = \frac{Y^2}{n^2}$ to estimate θ . Show that T^2 is a biased estimator of θ .
- (d) [5 marks] Find an expression of the form $aT^2 + bT$ which gives an unbiased estimator of θ .
5. Plankton are key organisms in many aquatic systems. As such, it is essential to understand how different abiotic quantities affect plankton levels.

Part of an investigation into the factors affecting the productivity of plankton in the River Thames at Reading consisted of taking measurements at 30 monthly intervals of the production of oxygen (in milligrams per cubic metre per day) and potentially useful explanatory variables in predicting its level. The aim of this assignment is to investigate the relationship between the explanatory variables and the production of oxygen. The data file (`oxygenfull.csv`) contains one line of data per month as follows:

- OXY – Production of oxygen (in milligrams per cubic metre per day)
- CHLOR – The amount of chlorophyll (in micrograms per cubic metre per day)
- LIGHT -The amount of light (in calories per square centimetre per day)
- TEMP -The temperature of the water (in degrees C)
- SPEED -The speed of the river flow (in metres per second)
- SUN - The amount of sun light per day (hours)
- ALT - The altitude above sea level (in metres)

- (a) [6 marks] Read in and check the data. Carry out exploratory data analysis, and comment on your findings.
- (b) [10 marks] Build a regression model during which you will have to decide which of the possible explanatory variables should be included in the model. Explain the key steps of your analysis (in particular why you decided to include certain explanatory variables, but not others).
- (c) [9 marks] Comment on the fit and interpretation of the final model, explaining what conclusions you can draw from your chosen model. Include suitable residual plots, commenting as appropriate.

Total for paper = 100 marks.

The submitted work should be your own work! The questions apart from Q2(d), Q3(c), Q3(d) and Q5 are theoretical exercises, and should be solved using results we covered in lectures. Make sure you justify each step of the theoretical reasoning by clearly stating the theorem/property you are using (marks will be awarded for these). Also make sure that you add comments to each section of your R code, explaining what you're doing. A pdf document with your R code and the solutions to the theoretical exercises should be submitted through EBART by Noon (12pm), 22nd November. Note that late submissions will be penalised.