MTH2006: Statistical Modelling and Inference
Assessed Coursework Assignment 1

October 25, 2024

Your solutions for this assignment should be written up and uploaded as a PDF file on the MTH2006 ELE coursework page before

12 noon on Friday 13 December 2024.

This assignment has a total of 100 marks, which contribute 15% to the final module mark. Usual penalties (marks capped at 40 for work submitted up to 2 weeks after the deadline, zero thereafter) will apply for late submission unless mitigation is approved.

This assessed coursework should be completed by you and you alone - strict disciplinary action will be taken for any collusion or plagiarism. Furthermore, this assessment is AI-prohibited. This is because you will demonstrate that you have achieved the intended learning outcomes only if you complete the assessment without using GenAI tools such as ChatGPT. If markers think that you might have committed an academic offence then you will be required to attend a viva (oral exam) in order to establish the legitimacy of your work.

The report should clearly and concisely answer the questions. It should be written up using word processing sofware e.g. LaTeX, R Markdown, or Word with a font size of at least 11 point. Approximately a third of the marks will be awarded for correct calculation, a third for good presentation of results, and a third for reliable interpretation. All plots should have meaningful titles and appropriately labelled axes and decimal numbers should be reported to a sensible number of significant figures. Unless requested in the question, do not include raw R output or blocks of R code in your report, e.g. the output of '`summary(model)`'.

The data frames `alfheim` and `whra` required for these questions can be obtained by loading the workspace `Aut2024.RData` available to download from the ELE module page, i.e. `load("Aut2024.RData")` in R.

1. The Elves in the village of Alfheim have recorded 10 years ($n = 3650$) of daily rainfall totals $\{y_1, y_2, \ldots, y_n\}$ (in mm), which they have kindly provided in data frame `alfheim`.

   (a) Make plots to show the rainfall data and its distribution and briefly discuss the main features.

   (10 marks)

   (b) Suppose that daily rainfall totals can be considered to be uniformly distributed for amounts less than 1mm (quite dry days) and the probability of having such days is $1 - \phi$ with $0 \leq \phi \leq 1$. On wetter days having daily rainfall amounts greater than 1mm, assume that the amounts are exponentially distributed, i.e. they have the p.d.f. $f(y) = c\exp(-\theta y)$.

   Sketch the p.d.f. $f(y)$ for $y > 0$ and by suitable integration find an expression for c in terms of parameters ϕ and θ.

   (8 marks)

   (c) Write down the likelihood function for the following small sample of rainfall totals: $\{30, 0.2, 10, 0\}$ stating any additional assumptions you have had to make.

   (8 marks)

   (d) By partitioning days into wet days with totals $Y > 1$ and dry days with $Y \leq 1$ derive an expression for the likelihood from 1c) that depends on the number of wet days $m = \sum_{y_i > 1} 1$. By partial differentiation of the logarithm of this likelihood function find maximum likelihood estimators for ϕ and θ. Use these estimators and the `alfheim` data to obtain estimates for ϕ and θ. By calculating the expected information, comment on the correlation expected between the $\hat{\phi}$ and $\hat{\theta}$ estimates.

   (14 marks)

   (e) Use numerical optimization in R to find maximum likelihood estimates of ϕ and θ based on the data in `alfheim`. Show the code you used to do this. Briefly compare your estimates to those found from differentiation of the log likelihood.

   (10 marks)

   (Total 50/100 marks)

2. The World Happiness Report 2023 ranked 136 countries by the average happiness levels of a representative sample of citizens. A subset of the data have been extracted into the R data frame `whra` which contains observations of the following variables:

   - `Country` - name of the country.
   - `Happiness` - measure of national happiness based on average response to question about which step of the happiness ladder you feel you are on: 0 (worst) to 10 (best).
   - `GDP` - natural logarithm of the Gross Domestic Product (in USD) per capita.
   - `SocialSupport` - the mean response (0 - "No", 1 - "Yes") to"If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them?"
   - `LifeExpectancy`- estimate of mean life span in years.
   - `Freedom` - the mean response (0 - "No", 1 - "Yes") to "Are you satisfied with your freedom to choose what you do with your life?"
   - `Generosity` - the residual of regressing the national average of responses (0 - "No", 1 - "Yes") to the question "Have you donated money to a charity in the past month?" on GDP per capita.
   - `Corruption` - the mean response (0 - "No", 1 - "Yes") to"Is corruption widespread throughout the government or businesses?"

   Analyse the data to address the questions below:

   (a) Fit a simple linear regression model to assess how Happiness depends upon (the natural logarithm of) Gross Domestic Product (GDP) with the inclusion of no other explanatory variables. Make a scattter plot showing the data and the line of best fit together with a 95% confidence band.

   (7 marks)

   (b) Quantify the goodness of fit and assess how statistically significant the fit is stating clearly your hypotheses and chosen level of significance.

   (10 marks)

   (c) State the assumptions about the standardised residuals of the normal linear model and present appropriate diagnostic plots to assess these. Also comment on which (if any) countries have unusual outliers or overly influential observations.

   (20 marks)

   (d) Calculate a 95% prediction interval using this model fit for the United Kingdom assuming that UK GDP will increase in the future by 10%.

   (6 marks)

   (e) Use stepwise selection of variables find the best subset of (the natural logarithm of) GDP, LifeExpectancy, SocialSupport, Freedom, Generosity, and Corruption for explaining Happiness. Perform an ANOVA F-test to assess whether the model with the selected explanatory variables is significantly better than the simple linear model from 2a) with only (the natural logarithm of) GDP.

   (7 marks)

   (Total 50/100 marks)