MTH2006: Statistical Modelling and Inference
Assessed Coursework Assignment 2

February 26, 2025

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 28 March 2025.

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 be clear and concise and formatted with a font size of at least 11 point - it should be written up using word processing sofware e.g. LaTeX, R Markdown, or Word. Approximately a third of the marks will be awarded for correct calculation, a third for good presentation of results, and a third for suitable interpretation. Do not include raw R output as part of your solutions (e.g. the output of '`summary(model)`'). All plots should have meaningful titles and appropriately labelled axes and decimal numbers should be reported to a sensible number of significant figures.

©2023 University of Exeter. All rights reserved.

1. Heart rate recovery (HRR) measures how quickly heart rate decreases after exercise and is closely related to the cardiorespiratory fitness (CRF) of a person. It is widely used by sports professionals to adjust the training of professional athletes. There are various ways to calculate HRR. Mongin *et al.* (2023) [1] performed a study to explore the relationships between various HRR measures and cardiorespiratory fitness (CRF). In this study, the athletes performed a Graded Exercise Testing (GET). During each test, the cardiorespiratory fitness measure, i.e., maximum aerobic speed (km per hour), `MAS` was recorded. In addition, the following HRR indices related to the cardiorespiratory measurements were recorded during each test, where t_0 denotes a time after exercise cessation and HRpeak represents the maximum heart rate (beats per minute):

   - `HR10`, `HR180`: HR measured at $t_0 = 10$ and 180 seconds respectively (beats per minute).
   - `HRR_delta10`, `HRR_delta180`: the difference between HRpeak and HR $t_0 = 10$ and 180 seconds respectively (beats per minute).
   - `HRR10per`, `HRR180per`: the ratio between HR at $t_0 = 10$ and 180 seconds and HRpeak, expressed as percentage.

   The dataset also includes the following variables:

   - `ID`: Identification of the athlete.
   - `Age`: age of the athlete (years).
   - `BMI`: body mass index of the atlete (km/m^2).
   - `Sex`: categorical variable with 1 indicating Female and 0 indicating Male.

   You can download the data from the assessment tile on the ELE page (`experiment.csv`).

   (a) Produce and include a scatter plot of the `MAS` values against `HRR_delta180` with appropriate labelling. Briefly discuss the observed trends or patterns.

   (7 marks)

   (b) Fit a linear regression model to examine the relationship between `HRR_delta180` and `MAS`. In your analysis, to account for potential confounding factors, consider extending the model by including relevant covariates, `BMI`, `Age` and `Sex`, which may influence both HRR measure and `MAS`. You should conduct and interpret hypothesis tests at the 5% significance level. Make sure to report the results of your linear model in a table, along with an interpretation of the table.

   (13 marks)

©2023 University of Exeter. All rights reserved.

(c) Mongin *et al.* (2023) [1] stratified the analysis by sex to assess the potential sex differences. Instead of stratifying by sex, account for sex differences in your linear model for `MAS` by introducing interaction terms between sex and other predictors. Present the results of your modified model in a table and provide a statement on whether there is evidence of interactions between sex and the predictor variables. You should conduct and interpret hypothesis tests at the 5% significance level.

(5 marks)

(d) Present residual diagnostic plots for your linear model from part (c) and discuss whether you think this model is well-specified, and if not, which assumptions are not valid.

(10 marks)

(e) In this part of the question, consider only females. Explore the effects of all HRR indices provided in this data set on `MAS` by considering linear regression models of the following form:

$$\texttt{MAS} \sim \beta_0 + \beta_1\texttt{Age} + \beta_2\texttt{BMI} + \beta_3\texttt{HRR}.$$

i Apply a Bonferroni correction to preserve a 5% significance threshold and interpret the results.

ii Apply the Holm method (for the same threshold) and interpret the results.

(15 marks)

(Total 50/100 marks)

©2023 University of Exeter. All rights reserved.

2. In this question, you will investigate and compare the audio features of Kendrick Lamar's and Drake's songs using methods from nonparametric statistics. You can download the data from the assessment tile on the ELE page. Two data files, `kendrick_music.csv` and `drake_music.csv`, contain audio features of Kendrick Lamar's and Drake's songs featured on studio albums and available on Spotify:

   - `speechiness`: a measure of how much spoken word content is present in a track. This value is between 0 and 1, with lower values indicating tracks that are mostly musical and higher values representing a greater presence of spoken words.
   - `danceability`: a measure of how suitable a track is for dancing based on a combination of musical elements. This value is between 0 and 1, with lower values indicating tracks that are less suitable for dancing. Higher values represent tracks that are more dance-friendly.

   (a) Consider a set of variables $\{Y_1, Y_2, \ldots, Y_n\}$ that are independently distributed with a probability density function (pdf) following the Beta distribution, defined as

   $$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \quad 0 \leq y \leq 1.$$

   Here, $\mathrm{B}(\alpha, \beta)$ is the Beta function, defined as $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where $\Gamma(\cdot)$ denotes the Gamma function, and $\alpha, \beta > 0$. Write down the log-likelihood function and then use numerical optimisation in `R` to find the Maximum Likelihood Estimates of α and β using Kendrick Lamar's `danceability` data. Show the code you used to do this.

   (10 marks)

   (b) Use a Q-Q plot to judge whether the beta probability distribution produces a good model for the `danceability` data from part (a). The Q-Q plot should be included. Show the code you used to do this.

   (5 marks)

   (c) Test at the 5% significance level whether cdf $F_Y(y; \hat{\alpha}, \hat{\beta})$ is a good model for the danceability measure with Pearson's chi-squared test. Use bins

   $$(0, 0.15], (0.15, 0.3], (0.3, 0.45], (0.45, 0.6], (0.6, 0.75], (0.75, 1].$$

   State the hypotheses that you use and present the observed and expected values tidily in a table. Compare your findings to the Q-Q plot of 2(b). Show the code you used to do this.

   (10 marks)

©2023 University of Exeter. All rights reserved.

(d) Let $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$ be two-dimensional random vectors, where X_1 and Y_1 represent the speechiness scores, and X_2 and Y_2 represent the danceability measures for Kendrick Lamar's and Drake's songs respectively. In this question, you will examine the **dispersion ordering $\mathbf{X} \preceq_d \mathbf{Y}$**, which says that $\mathbf{X}$ is less dispersive than $\mathbf{Y}$. Assess this dispersion ordering by completing the following tasks:

i Write a function in `R` to compute the area of a 2-dimensional simplex (triangle) defined by three vertices in the 2D plane, which is a widely used measure of dispersion. The vertices are given by the points $\mathbf{x}_1 = (x_{11}, x_{12})$, $\mathbf{x}_2 = (x_{21}, x_{22})$ and $\mathbf{x}_3 = (x_{31}, x_{32})$. The area of the triangle can be computed using the following formula:

$$\text{Area} = \frac{1}{2}|x_{11}(x_{22} - x_{32})| + x_{21}(x_{32} - x_{12}) + x_{31}(x_{12} - x_{22})|.$$

ii For each artist, randomly sample three songs and compute the area of a 2D triangle using their `danceability` and `speechiness` scores as the coordinates and using the function from part (i). Repeat this process for 1000 iterations and store these values in a vector. You can use `combn()` function in `R` to generate random samples of songs. Show the code you used to do this.

iii Produce a plot of the empirical cumulative distribution function (ecdf) for the triangle area measure for each artist from part (ii), with both ecdfs displayed on the same plot. Discuss the relative positions of these ecdfs.

iv Consider the Kolmogorov-Smirnov (K-S) test at the 1% significance level for assessing whether the triangle area measures for Drake are greater than those for Kendrick Lamar. Show the code you used to do this. Comment on the **dispersion ordering**, where a greater dispersion for Drake would indicate that his triangle area measure stochastically dominates Kendrick Lamar's.

(25 marks)

(Total 50/100 marks)

References

[1] Denis Mongin, Clovis Chabert, Delphine Sophie Courvoisier, Jeronimo García-Romero, and Jose Ramon Alvero-Cruz. Heart rate recovery to assess fitness: comparison of different calculation methods in a large cross-sectional study. *Research in Sports Medicine*, 31(2):157–170, 2023.

©2023 University of Exeter. All rights reserved.