MATH363: Group Project part 2

Group project contributes 60% to your mark for Math363.
Part 1 = 15 marks; Part 2 = 75 marks; Minutes = 10 marks

You will work on all parts of this project as a group and submit your answers as a group. In addition to the project you need to submit minutes from your meetings and provide the peer assessment of everyone's contribution through Buddycheck. The final mark for each member of the group will be adjusted according to these peer scores.

The project needs to be submitted on Canvas and will be checked for potential plagiarism using turnitin. A high similarity score might result in the project being investigated for suspected plagiarism; see Code of Practice on Assessment Appendix L.

DATA SETS FOR EACH GROUP ARE DIFFERENT and THEREFORE YOUR RESULTS AND ANSWERS SHOULD BE DIFFERENT

Please only discuss your project with your group members. Any similarity between different project submission might be investigated for suspected plagiarism.

You must submit your minutes and buddycheck scores at the end of every three week period via Canvas page.

All relevant code must be included in the appendix **[10 marks]**.

Part II

1. **[40 marks]** A six minute walking test (6MWT) is used to assess functional exercise performance. In this test the subject is asked to walk for 6 minutes on a level course and the distance covered is recorded in meters (m). The pace is set by the subject and breaks are allowed if needed. The data were collected on healthy subjects and includes information about

    - sex (Female = 1; Male = 0);
    - age (in years);
    - height (cm);
    - weight (kg);
    - BMI (= body mass index);

- resting heart rate (beats per minute, bmp);
- smoking status (1 if smoker);
- usual activity level: 0 =sedentary (less then 30 min physical activity a day), 1 = moderately active (30-60 min physical activity a day), 2 = active (more than 60 min physical activity a day);
- heart rate at the end of the 6-minute walk;

You are asked to find a model for dependance of the distance travelled in 6MWT on the explanatory variables provided. The model will be used to predict the average distance travelled for a person and should only include variables available before the test is taken (that is, not their heart rate at the end of the test).

(a) The researchers think that different models might be needed for smokers and non-smokers subjects. For each of the covariates separately, fit a linear model which has different intercept and slope for the two groups. Using these models and appropriate plots, write a short report discussing if separate models are indeed required.

(b) Propose a model for these data that will allow the researchers to predict patient's performance in a 6MWT. Some possible approaches here are:
   - You could start by fitting separate models for each variable, and choosing the model you think is most useful from these single variable models. Then you can try adding another variable one at a time and so on until no new variables improve the model.
   - Include all variables and any additional terms such as interactions and higher order terms in the initial model. Then remove variables that don't appear useful (for example, because their coefficient is not significant) one at a time, starting from the least useful. Continue until all remaining variables are useful (or significant).
   - Combine the two approaches above and add and remove terms till you get your 'best' model.

(c) For the model chosen in (b), perform residual analysis and decide if the model fits well. If it does not, suggest changes that can be made to the model to address these issues.

(d) An additional data set was obtained and in included in `Part2Validation.rds`. Use your proposed model to make predictions for the new data set. Comment on the results. You should also discuss the uncertainty of your predictions and any limitations of your model here.

2. **[10 marks]** In Math363 lectures we discussed ANOVA models but focused only on balanced design. In reality, data often come from various incomplete and unbalanced designs, for which the ANOVA models is not orthogonal. One such example would be your data, when we only analyse the influence of the three factors on the response. For such data, different types of sums of squares have been defined. Provide a short summary about the different types of sums of squares, their advantages and limitations. Decide which type is most appropriate for your data and justify your choice. Create an appropriate ANOVA table for your data and fully analyse it. You should use Generative AI (for example Microsoft Copilot) to obtain your initial answer and then refine and verify this answer using published resources. YOU NEED TO INCLUDE THE GAI GENERATED REPORT IN THE APPENDIX; no marks will be give

for this part if this is missing. *Word limit 1000; tables are not included in word limit*

3. **[15 marks]** It is also of interest to model the probability that a subject reached the threshold heart rate (HR) for vigorous-intensity physical activity (VIPA), that is a heart rate of at least 77% of maximum HR. Maximum HR is calculated as 220 minus age. For example, for a person who is 35, maximum HR is 185 bpm and VIPA is when their HR is at least $0.77(220 - 35) \approx 143$bpm. The researchers want to use the data set used in Q1 to estimate the probability of VIPA during 6MWT. The distance travelled should be one of the explanatory variables in this case.

   (a) Create a new variable which is 1 if a person's HR reached VIPA, and 0 otherwise (you might want to first create a variable with maximum HR for each person).

   (b) Propose a model for the probability of VIPA. You do not need to consider interactions or higher order terms in this case but you need to consider which link function is most appropriate.

   (c) Interpret your model parameters, discuss its fit and any limitations.