

MATH48011/68011 – Linear Models with Nonparametric Regression

October 30, 2025

Coursework 2025-26

Deadline: 11am (UK time) Monday 14th November 2025

Information: This piece of coursework accounts for 20% of your overall mark for this course and it may take around 10 hours to complete.

Instructions: Please present your solution in the form of a report, which you should upload on Blackboard as a single file before the deadline. RMarkdown, LaTeX, or Microsoft Word would be good tools to produce your report.

Although you should use R to perform analyses and computations, you must present the formulas used in the text in mathematical notation (not as R code).

Marks will be awarded for correct and accurate calculations and their interpretation. Interpretations should be explained in words, referring to the context, rather than naming generic symbols only. High marks will be less likely if the presentation of the results is unclear, too short or unnecessarily long and confusing, or if any formulas used in the calculations are missing from the text.

Report length Around 6 A4 sides should be enough to produce a complete solution, using 11pt font. The maximum page limit for submissions is 12 A4 sides in 11pt font.

Academic integrity: The submitted work should be your own. Plagiarism may result in zero marks for all parties involved and/or disciplinary action.

Questions

Semaglutide is a glucagon-like peptide-1 receptor agonist that was recently approved for use in the treatment of type 2 diabetes, and for chronic weight management.

In adults, a dose of 2.4 mg/week is prescribed f is used to indirectly measure the amount of iron in the blood (measured in ng/mL).

A medical research team has been asked to examine the effect of taking 2.4 mg/week of Semaglutide on walking capacity. The team conducted a clinical trial with $n = 280^1$ participants, and examined the change between maximum walking distance, before (MWDB), and after 59 weeks of treatment with Semaglutide (MWDA), by examining their ratio: $Y = \text{WMDA}/\text{MWDB}$.

It is known that maximum walking distance, even under fixed conditions, varies from person to person, that they often change with age and differ between men and women, and suspect

¹Typo corrected, this was originally printed as $n = 200$.

the effect of Semaglutide on walking will vary similarly. The team is interested in establishing the relationship between the maximum walking distance ratio (Y) and the age (x), and sex (z), of a participant. Due to historical data limitations the sex of the patient was recorded as either female (F) or male (M). The data are given in the file `Semaglutide.csv` available from Canvas.

1. Load the data into R. Draw exploratory plots of the relationship between x, z, and Y. Comment on any interesting features in your findings. [4 marks]
2. Create a dummy variable w using male as the reference level. Fit a model of the form

$$F : E[Y] = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 w + \theta_4 xw.$$

Give the equation of the fitted model. Plot the fitted ‘lines’ for each of the two genders separately. [4 marks]

3. Give the interpretation of the parameters θ_0 , $\theta_0 + \theta_3$, θ_4 . [3 marks]
4. What assumptions are required to carry out statistical tests using your model? Use graphical diagnostics to assess the validity of these assumptions. [3 marks]
5. At 5% significant level, test the null hypothesis that, the expected maximum walking distance ratio for 40 year old females is the same as 40 year old males . [3 marks]
6. Using the model fitted, what is the prediction of expected maximum walking distance ratio for males aged 22? Give the 95% prediction interval. for your prediction. [3 marks] [Total 20 marks]