Department of Statistics
STATS 330: Statistical Modelling
Assignment 1
Semester 1, 2020

Total: 100 marks Due: 12:00 (noon) NZDT, Friday 27 March 2020

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each question. Include all relevant code and output in the final document.

(iii) Marks may be deducted for poor style. Please keep your code and plots neat.

(iv) Please remember to hand in your hard copy, with signed cover sheet, by the due date.

(v) Please remember to upload your R Markdown file to Canvas before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong.

Introduction

The *Fletcher Challenge – University of Auckland Heart & Health Study* was established in the mid 1990s with the aim of determining the relationship of sociodemographic factors, psychological factors and several other factors measured in blood with the risk of coronary heart disease (CHD) in a New Zealand population. Participants were recruited from two sources: employees of the *Fletcher Challenge Group* and individuals listed on the general electoral roll for the Auckland region. Baseline and follow-up risk factor data were obtained from a questionnaire, blood samples and a simple physical examination[1].

[1] MacMahon, S. et al, "Fletcher Challenge - University of Auckland Heart & Health Study: design and baseline findings." *The New Zealand Medical Journal*, 1995; 108(1013): 499–502.

You have access to a subset of the study data for the purposes of this assignment. You can find it in a file called `hearthealth.csv` on CANVAS (Assignments > Assignment 1).

The data set `hearthealth.csv` contains the following variables:

- `age`: the age of the participant (in years)
- `sex`: the gender of the participant (either `M` or `F`)
- `ethnicity`: the ethnicity of the participant (either `European`, `Māori`, `Polynesian`, or `Other`)
- `height`: the height of the participant (in metres)
- `weight`: the weight of the participant (in kg)
- `sbp`: the systolic pressure of the participant (in mm Hg)
- `dbp`: the diastolic pressure of the participant (in mm Hg)
- `pulse`: the pulse of the participant (in beats per minute)
- `chol`: cholesterol level status (either `High` or `Low`)
- `heartattack`: heart attack (either `Yes` or `No`)
- `drinkmaxday`: the maximum number of alcoholic drinks consumed per day (within the last 3 months)
- `smoke`: ever smoked once a week or more (either `Yes` or `No`)
- `exerday`: typical number of days a week engaged in vigorous activity
- `exerhour`: typical number of hours a day engaged in vigorous activity
- `exermin`: typical number of minutes a day engaged in vigorous activity
- `eggs`: typical number of eggs consumed per week

(1) **Data Cleaning.** Clean the data set. You will need to explore the following variables: `height`, `weight`, `sbp`, and `dbp`, perhaps plotting some variables against other variables. Identify any problems you encounter and make sensible suggestions about what might have happened when these variables were recorded. Make decisions about whether or not to remove a rogue observation, or modify it in a sensible way. Whatever you decide to do, make sure you justify your actions.

[10 marks]

(2) **Eggs exploration**. We are interested in the variable `eggs`. In particular, does `eggs` vary according to `ethnicity`. Produce some informative exploratory plots, identify if there are any observations that need more investigation and then fit an appropriate model. Justify your choice of model and interpret the results of your model.

[12 marks]

(3) **Drinking**.

We are interested in exploring the relationship between a person's age and gender and the maximum number of alcoholic drinks they consume in a day (within the last 3 months).

(a) Identify the response and explanatory variables for this exploration.

[3 marks]

(b) What model would be appropriate to fit? Choose from linear regression, Poisson regression and logistic regression. Justify your choice.

[3 marks]

(c) Using your answers to parts (i) and (ii), fit a model to explore the relationship between age, gender and maximum number of alcoholic drinks consumed per day (within the last 3 months). Interpret the model. Is there evidence to suggest that age and/or gender are somehow related to the maximum number of alcoholic drinks consumed per day (within the last 3 months)? Justify your answer[2].

[8 marks]

(d) (i) Construct a plot displaying the estimated effects of your model on the response variable[3].

[7 marks]

(ii) Clearly communicate what your plot displays.

[3 marks]

(e) Compare the estimated `drinkmaxday` values for the following values of `age` and `sex`. Interpret these estimates in context.

- (`age` = 30, `sex` = M) and (`age` = 50, `sex` = M)
- (`age` = 40, `sex` = M) and (`age` = 40, `sex` = F)

[4 marks]

(4) **Exercise**.

We are interested in exploring the relationship between a person's age and their smoking history (as measured by `smokeever`) and the typical number of hours that they engage in vigorous exercise in a week.

(a) Identify the response and explanatory variables for this exploration.

[5 marks]

(b) What model would be appropriate to fit? Choose from linear regression, Poisson regression and logistic regression. Justify your choice.

[3 marks]

(c) Using your answers to parts (i) and (ii), fit a model to explore the relationship between age, smoking history and the typical number of hours engaged in vigorous exercise in a week. Interpret the model. Is there evidence to suggest that age and/or smoking history are somehow related to the typical number of hours of vigorous exercise undertaken in a week? Justify your answer[2].

[8 marks]

(d) (i) Construct a plot displaying the estimated effects of your model on the response variable[3].

[7 marks]

(ii) Clearly communicate what your plot displays.

[3 marks]

(5) **Heart Attacks**.

We are interested in exploring the relationship between a person's age and cholesterol level status and whether or not they have had a heart attack.

(a) Identify the response and explanatory variables for this exploration.

[3 marks]

(b) What model would be appropriate to fit? Choose from linear regression, Poisson regression and logistic regression. Justify your choice.

[3 marks]

(c) Using your answers to parts (i) and (ii), fit a model to explore the relationship between age, cholesterol level status and heart attack risk. Interpret the model. Is there evidence to suggest that age and/or cholesterol level status have an effect? Justify your answer[2].

[8 marks]

(d) (i) Construct a plot displaying the estimated effects of your model on the response variable[3].

[7 marks]

(ii) Clearly communicate what your plot displays.

[3 marks]

[2] Don't worry about testing for goodness-of-fit or doing model selection - these topics were not covered when this assignment was released.

[3] See the plots in Handout 2 for a good example of something similar from the chicken bacteria data analysis. Also see Tutorial 2.