Assignment 1 Questions

Student name and ID

Due: 6 April 2023

Your assignment should be submitted as a Rmarkdown document with all analysis and graphics included as R-chunks, together with the compiled pdf file. The Rmarkdown document should compile without error. You can assume that the data files are in the same folder as your `Rmd` file. This assignment has 40 marks in total.

Consider the question whether there is a link between maternal smoking and the baby's health. The goal of this assignment is to model birth weight, which is an important indicator of a baby's health. The data set `dataA1_etc3580.csv` includes the following variables:

- bwt: birth weight in grams
- age: mother's age in years
- race: indicator for whether the mother is white, black, or other
- smoke: indicator for whether the mother is a smoker or nonsmoker

Please follow the following steps and answer the questions:

1. Do some data preparation: Read in the data set, convert categorical variables to factors, and define the labels of the categories. (2 marks)

2. Use `ggplot()` to produce three appropriate plots of respectively `age`, `race` and `smoke` against `bwt`. What do you learn about the relation between `bwt` and each of these predictors? (4 marks)

3. Fit a linear regression model for `bwt` with the main effects of the predictors `age`, `race` and `smoke`. Provide an interpretation of the coefficients for `age` and `smoke`. (4 marks)

4. Fit a linear regression model for `bwt` that includes an interaction between 'race' and 'smoke', together with their main effects and the main effect for `age`. Use visreg to visualize the interaction term in the model, and describe what you learn from this. (4 marks)

5. Test if the interaction in Question 4 is significant. (2 marks)

6. Fit a regression model for `bwt` with the main predictors `age`, `race` and `smoke`. Use the AIC to select a linear or quadratic relation between `age` and `bwt`. (4 marks)

7. Instead of the exact birth weight, predictions for the probability that the birth weight is too low may be of more interest to a nurse. Estimate a model that can predict the probability that birth weight is smaller than 2500 gram, using the predictors `age`, `race`, and `smoke`. What is the predicted probability of low birth weight for a baby with a 30 year old black mother who does not smoke? (4 marks)

8. Individual probabilities for each baby may be too much information for a nurse. In this case, group-level probabilities are more practical. Construct a categorical variable `agecat` that indicates whether a baby's mother is 20 years or younger, 30 years or younger but older than 20, or older than 30. Now construct a data set with at most 18 groups (3 age categories times 3 races times 2 smoke types), and save for each group the number of babies with birth weight smaller than 2500 gram and higher. (4 marks)

9. Fit a regression model for the number of low birth weight babies in each group, with the main effects of the predictors `agecat`, `race`, and `smoke`. Provide an interpretation of the coefficients for `race` and `smoke` in terms of the effect of each variable on the odds of a low birth weight. (4 marks)

10. Provide three different reasons that may explain why the estimated coefficients have different signs across the estimated regressions in questions 3, 7, and 9, while they all explain birth weight using age, race, and smoke. (4 marks)

11. Fit a regression model for the number of low birth weight babies in each group, with the main effects of the predictors `agecat`, `race`, and `smoke`, that takes overdispersion into account. Use this model and the estimated model from question 9 to provide two reasons why overdispersion may not be a problem here. (4 marks)