Practical : Linear Regression

Exercice 1 :

Explanations

This first example is inspired by a project we had; the data is simulated.

Context:

The cohort we have consists of 600 patients with Parkinson's disease, all carrying a specific mutation out of a total of six different mutations. In the context of our study, we follow these patients and measure their UPDRS score, a widely used cognitive indicator to assess the severity of Parkinson's disease. We have also collected other information, including the duration of the disease, age at the time of diagnosis, gender, and age at the time of participation in the study.

Primary Outcome:
We aim to understand the factors influencing the Unified Parkinson's Disease Rating Scale (UPDRS), which is an indicator of the severity of Parkinson's disease, and potentially predict it using variables such as disease duration. We have reasons to believe that the severity of the disease varies based on specific mutations (senescence vs. non-senescence mutations), and we plan to account for this variation. Additionally, we assume that the age at the time of participation in the study may influence UPDRS, and we intend to correct for this influence. Furthermore, it is possible that the gender of the patients may also have an impact, which we will examine.

Secondary Outcome:
We also aim to assess to what extent the mutation can influence the age at which the diagnosis of Parkinson's disease is made and the timing of disease onset.

Database Description:

The dataset comprises 600 patients, and the included variables are as follows:

- "Mutation": indicating the patient's specific mutation, with a total of 6 categories: "m/s", "f/s", "s/s", "f/f", "m/m", "f/m".
- "Age_at_onset": representing the age at the time of diagnosis.
- "Sex": specifying the patient's sex.
- "Disease_duration": describing the duration of the disease at the time of the study.

- "Age": indicating the age at the time of participation in the study.
- "UPDRS": reflecting the value of the UPDRS score, an indicator of the severity of Parkinson's disease.

Questions:

Data management:

1. Import the dataset 'Exercice1.txt' into R and display the first few rows of the dataset to get an overview of its content.
   Note: Use the functions read.table, head.

2. After importing the dataset, ensure that the variables 'Sex' and 'Mutations' are correctly recognized as factors.

3. Create a new variable called 'Mutation_bis' that has two categories: 'senescence' and 'non-senescence'. The 'senescence' category includes the modalities 'm/s', 'f/s', and 's/s', while the 'non-senescence' category includes 'm/m', 'f/f', and 'm/f' from the 'Mutation' variable.

Univariate Statistics:

4. Create histograms for the variables UPDRS, Age, Age at Diagnosis (Age_at_onset), and Disease Duration (Disease_duration). Comment on the shape of the distributions.
   Note: use geom_histogram

5. Create a barplot for the 'Sex' variable.
   Note: use geom_barplot

6. Calculate the mean and standard deviation for numerical variables, as well as percentages for categorical variables, and present these measures in a data table.
   Note: Use the 'table1' function from the 'furniture' package

Bivariate Statistics:

7. Create the same table as before, but this time segment the results based on the 'Sex' variable.
   Note: Use the 'splitby' and 'test' options of the 'table1' function from the 'furniture' package.

8. We want to determine if the disease duration (variable 'disease_duration') varies based on the type of mutation (variable 'Mutation').

- Write the equation of the corresponding model.
- Run this model in R. *Note: use lm, summary, Anova functions.*
- Create a figure to represent the results. *Note: use geom_boxplot a,d stat_compare_means.*

9. We want to determine if the age at the time of diagnosis (variable 'Age_at_onset') varies based on the type of mutation (variable 'Mutation').
    - Write the equation of the corresponding model.
    - Run this model in R. *Note: use lm, summary, Anova functions.*
    - Create a figure to represent the results. *Note: use geom_boxplot and stat_compare_means.*
    - Perform post hoc comparisons to identify differences in age at diagnosis among different mutations. *Note: Use the 'emmeans' package for this.*

10. We want to determine if the UPDRS (variable 'UPDRS') varies based on the type of mutation (variable 'Mutation').
    - Write the equation of the corresponding model.
    - Run this model in R. *Note: use lm, summary, Anova functions.*
    - Create a figure to represent the results. *Note: use geom_boxplot and stat_compare_means.*
    - Perform post hoc comparisons to identify differences in age at diagnosis among different mutations. *Note: Use the 'emmeans' package for this.*

<u>Régression linéaire simple :</u>

11. Using the primary criterion, identify the explanatory variables and determine the variable to be explained.

12. Perform a simple linear regression between UPDRS and age.
    - Write the equation of the corresponding model.
    - Analyze the output of the model, including the coefficients of the regression line, the intercept, and the R^2. Based on the results, assess whether there is a significant association between these two variables.
      Note: Use the 'summary' function and 'Anova' to obtain this information
    - Create a figure to represent the results. *Note: use geom_point and stat_compare_means.*

13. Perform a simple linear regression between UPDRS and the duration of the disease (Disease_duration)
    - Write the equation of the corresponding model.

- Analyze the output of the model, including the coefficients of the regression line, the intercept, and the R^2. Based on the results, assess whether there is a significant association between these two variables.
- Note: Use the 'summary' function and 'Anova' to obtain this information
- Create a figure to represent the results. Note: use geom_point and stat_compare_means.

Multiple Linear Regression:
Response to the primary outcome measure

14. Write the equation corresponding to the primary outcome measure.

15. Create a graphical representation of the model (UPDRS as a function of Disease_duration) using ggplot. Divide the 'Age' variable into categories based on its quantiles into 'Age_discrete', use 'geom_point' and 'geom_smooth', segment the Age graph with 'facet_wrap(~Age_discrete)', and assign different colors based on 'Mutation_bis' using the 'color' option.

16. Run this model in R.
    Note: use lm

17. Analyze the model output, specifically the coefficients of the regression line, the intercept, and the R^2. Interpret the model.
    Note: use summary and Anova functions

18. Conduct a contrast analysis to assess the impact of different mutations on our continuous variable. We want to perform post hoc comparisons while taking into account a continuous variable.
    Note: use the emmeans package

19. Create a comprehensive figure summarizing our model.
    Note: Use the 'interactions' package and the 'interact_plot' function

20. Perform a check of the assumptions for model validity.
    Note: Use the package ggResidpanel and the function resid_panel

    Collinearity in multiple linear regression occurs when two or more independent variables in the model are strongly correlated, making it challenging to distinguish their individual impact on the dependent variable. This can lead to instability in the estimates of regression coefficients. Check for collinearity.
    Note: use the vif function of the car package

Prediction:

21. Use the model you have constructed to predict the UPDRS value on the same dataset.
    Note: use predict.