BINF-F401 Project

Title

Draguet Simon Godin Maximilien Guyot Léopold

May 8, 2024

1 Introduction

2 exploration of clinical variables

Description des variables .....

2.1 Distribution of the variables

Associated script : .R

By examining the clinical variables through histograms, we can visualize their distribution. Most of the continuous variables do not seem to follow a normal distribution. For instance, the variable AGE appears to be asymmetric, exhibiting a right skew (fig X). To now if the continuous variable were normally distributed, we used a Shapiro-Wilk test. The results showed that AGE, HGHT, BMI, TRISCHD all had p-value under 5, leading us to reject the null hypothesis, which say that the sample is drawn form a normally distributed population. However, for WGHT, it was higher than 5 , so we didn't reject the null hypothesis (table X) .

To do further analysis, we need them to be normally distributed. We tried to apply various transformations like log, square root, square. Only the square transformation successfully normalized the AGE sample. Because we needed it to work on all the continuous variable, we finally chose to use the rank-based inverse normal transformation (INT). that first convert the variable into ranks, then map it to a normal distribution.

$Y_i^t = \Phi^{-1}(r - C/N - 2C + 1)$

where r_i is the ordinary rank of the $_ith$ case among the N observations and Φ^{-1} denotes the standard normal quantile (or probit) function.For the value of C we use C=3/8(ref).

If we look at the discrete variable, we can see that we don't have a balanced distribution. For example, for the variable AGE, there are more than twice as many males as females (fig x). For DTHHRDY, we can observe that most of the deaths occurred in ventilator cases (fig x). When conducting analysis, it's essential to keep these observations in mind as they can significantly influence the interpretation.

2.2 Impact of technical variables on clinical variables

In this section, we investigate the influence of technical variables on clinical data and describe the methodology employed to isolate their effects. To decouple the impact of technical variables from the data of interest, linear regression analyses were conducted. Prior to regression, the data underwent transformation using inverse data normalization for numerical variables (cf. section 1.1).

From the preceding section where correlations between variables were examined, certain expectations were established. Age was anticipated to be associated with TRISCHD, DTHHRDY, and COHORT, while height and weight were expected to be linked to the same variables. Additionally, sex appeared correlated with DTHHRDY and COHORT. Surprisingly, BMI seemed uncorrelated with any technical variables.

For the technical variable DTHHRDY, a decision was made to group certain categories into three new categories: slow, intermediate, and fast. This choice was made since more observations by categories would enhance statistical power.

Upon conducting linear regression (generalized linear regression in the case of the categorical variable sex), it was observed that age was dependent on cohort, sex was dependent on DTHHRDY, height was dependent on DTHHRDY, weight was dependent on DTHHRDY, and, surprisingly, BMI was dependent on both DTHHRDY and COHORT.

Subsequently, linear regression (or generalize d linear regression) was rerun for each variable of interest, utilizing only the significant technical variables from the overall models. The residuals of these models were stored for further analysis. These residuals represent the data of interest with the unwanted effects of the technical variables removed, providing a clearer understanding of the clinical data.

References