Project 6

The database is based on the study of Steiner et al (2020). The database contains the following variables:

- Group – indicates the group each individual belongs to: “CFS” (patient with Chronic Fatigue Syndrome with an infection trigger at his/her disease onset) and “HC” (healthy control)
- Gender – Sex of each individual: “m” – male and “f” – female.
- rs2456601, rs3087243, rs3807306, rs1800629 and rs1799724 – indicate the genotype of each individual for these genetic markers.

The main objective is to identify the genetic markers potentially associated with the disease. To do that:

1. Perform a descriptive analysis (e.g., summary statistics/plots) for each variable.
2. For each genetic marker, retrieve the information about the respective genetic information (chromosome, location, alleles) and the allele distribution of different populations from the 1000 Human Genome Project (1000HGP).
3. Calculate the allele distribution of HC for each genetic marker and. Compare these allele distributions against the allele distributions of each population from the 1000HGP. What is the population from 1000HGP closest to the HC with the closest allele?
4. Calculate the genotype distribution of each genetic marker in HC and CFS patients. Do these distributions agree with the expectations from the Hardy-Weinberg Equilibrium? Answer with an appropriate test.
5. Compare the genotype distribution of CFS and HC for each genetic marker using an appropriate statistical test. Draw your conclusions.
6. Using “Group” as the outcome variable, test different additive models for binary traits that enable to test the association between “Group” and each genetic marker while adjusting for gender. Extend, if possible, the previous model in order to include the additive effects of multiple significant genetic markers. What is the accuracy of the final model? Draw your conclusions.

7. Can these genetic data be used as a diagnostic tool for CFS? Answer with this question with an appropriate generalized linear model and the estimates of the respective sensitivity/specificity based on a ROC curve analysis.

Important:

Prepare a 15-minute presentation with your main findings. There will be a penalty of 0.5 points in your project grade if you exceed the time of your presentation. Upload your R script/R Markdown for code verification. Also upload your presentation as a pdf file. Failure to upload these files before classroom evaluation leads to a penalty of 0.5 in your project grade.

Note:

The original data set was modified for the purpose of this project and, therefore, the published results should not be used as guidance.

Reference:

Steiner S, Becker SC, Hartwig J, Sotzny F, Lorenz S, Bauer S, Löbel M, Stittrich AB, Grabowski P, Scheibenbogen C. Autoimmunity-Related Risk Variants in PTPN22 and CTLA4 Are Associated With ME/CFS With Infectious Onset. Front Immunol. 2020 Apr 9;11:578. doi: 10.3389/fimmu.2020.00578. PMID: 32328064; PMCID: PMC7161310.