STATS 767 Project Proposal

Stephen Su, ID 253222842

0.1 The Data

The dataset is sourced from ggeop/Myopia-Study (Papachristou, 2018) for a study of the risk factors of developing myopia in children, which is a subset from the Orinda Longitudinal Study of Myopia (OLSM) collected in 1989-1990 and 2000-2001. The dataset consists of 618 observations and 17 variables. In particular, the project will focus on the following six numeric variables and one categorical variable:

Variable_Name	Unit	Description
myopic	boolean	Myopia within the first five years of follow up
age	years	Age at first visit
sporthr	hours per week	Time spent engaging in sports/outdoor activities
readhr	hours per week	Time spent reading for pleasure
comphr	hours per week	Time spent playing video/computer games or working on the computer
studyhr	hours per week	Time spent reading or studying for school assignments
tvhr	hours per week	Time spent watching television

0.2 Question of Interest

The project aims to explore the relationship between the onset of myopia in children and lifestyle risk factors as per the mentioned variables and produce a statistical predictive model.

0.3 Methodologies

The project will utilise the methods from STATS 767. A linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are considered suitable for modelling the multivariate relationship for the purpose of establishing grouping for the characteristics of myopic and non-myopic children. The LDA and QDA models will also be compared to other methods, including PLS-DA and the logistic regression. The statistical methods used in this project are from **base R** (R Core Team, 2021), the **MASS** (Venables & Ripley, 2002) and **mixOmics** (F et al., 2017) packages.

0.4 Deliverables of the Project

The final deliverables of the project consist of an RMarkdown report and an HTML slideshow for the presentation using **rmarkdown** (Xie et al., 2018) and **xaringan** (Xie, 2021). All materials required to reproduce this project can be found at the Github repository szmsu2011/stats767proj.

Bibliography

F, R., B, G., A, S., & K-A, L. C. (2017). MixOmics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, *13*(11), e1005752. http://www.mixOmics.org

Papachristou, G. (2018). *Myopia study*. https://github.com/ggeop/Myopia-Study

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. http://www.stats.ox.ac.uk/pub/MASS4

Xie, Y. (2021). *Xaringan: Presentation ninja*. https://CRAN.R-project.org/package=xaringan

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown