A Multivariate Approach to Modelling Lifestyle Risk Factors of Children Myopia in the US

Stephen Su, STATS 767

1 Introduction

1.1 Background

The association between lifestyle factors and the development and subsequent progression of myopia among children has been long discussed and researched within the academic area. Among them, the Orinda Longitudinal Study of Myopia conducted research on children myopia spanning over 10 years. The research produced data that are both useful in exploring the lifestyle risk factors of myopia and a valuable case study for building and testing multivariate data models and analysis. This paper attempts to analyse the OLSM myopia dataset and produce a multivariate model.

1.2 The Data

The dataset is from ggeop/Myopia-Study (Papachristou, 2018), which is a subset from the original data collected in 1989-1990 and 2000-2001. The dataset consists of 618 observations and 17 variables. The main focus of the paper is around numeric, lifestyle-related (non-definitional) variables and the logical variable indicating the prevalence of myopia, which are represented by the following variables:

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

Note: "Non-definitional" means the variables do not serve as an optometrical reference to myopia.

1.3 Outputs and Deliverables

This paper and the project presentation only includes selective outputs serving as the final deliverable, including a **GGally** plot (Schloerke et al., 2020) and model outputs from **base R** (R Core Team, 2021) and the R **MASS** (Venables & Ripley, 2002) package. The detailed model building steps, intermediary models, and materials, such as the R source code, required to reproduce this paper can be found at the Github repository szmsu2011/stats767proj.

2 Exploratory Analysis

A preliminary visualisation of the selected data suggests heavy right-skewness except for the numeric variable `age`. As a convention, a log-transformation to all numeric variables except for `age` attempts to mitigate the skewness, yet the minima of the variables are zero. Instead, the `log1p` transformation is applied across the variables. Nevertheless, a subsequent plot of the transformed data indicates the `log1p` transformation seems to over-correct the skewness of variables `sporthr` and `tvhr`. Therefore, a final decision is made to transform variables `readhr`, `comphr` and `studyhr` by $x \to log(1+x)$ and `sporthr` and `tvhr` by $x \to \sqrt{x}$.

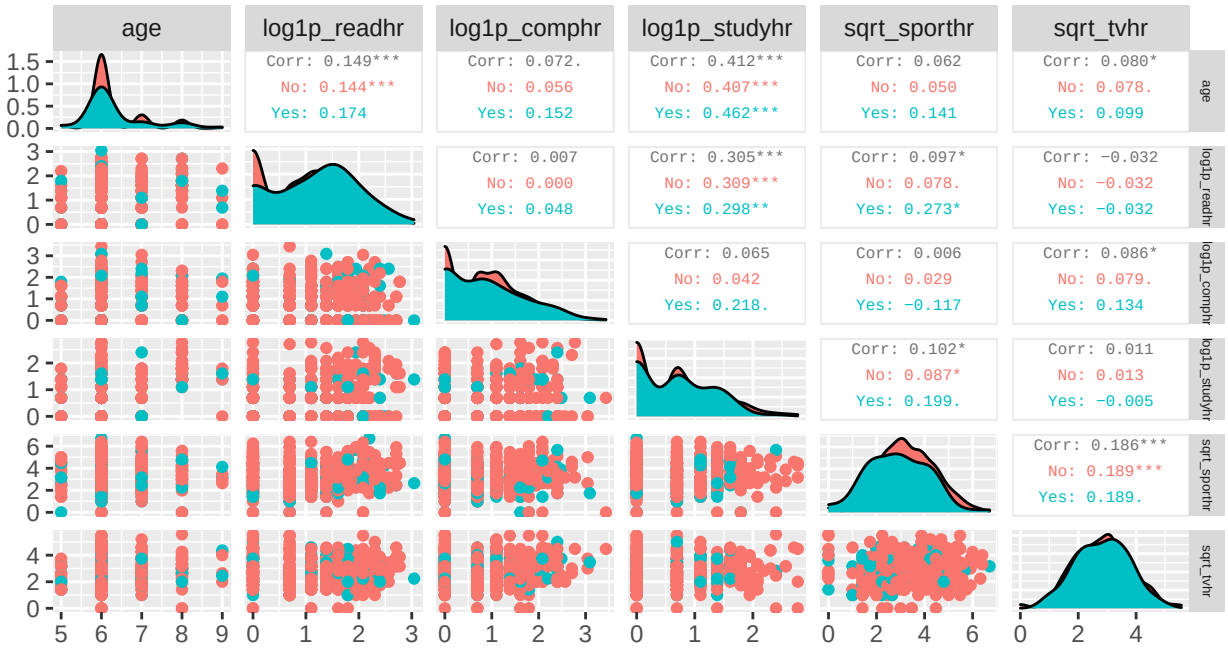

Figure 1: The transformed data is a considerable improvement from the original, notwithstanding a substantial departure from normality. All subsequent discussions are based on the transformed data.

A quantile plot, which is an extension to a boxplot, indicates a weak positive bivariate correlation with `log1p_readhr` and `log1p_studyhr` to myopia as well as a negative correlation with `sqrt_sporthr`. The factors affecting myopia development include complicated aspects, including lifestyle, gene, other conditions and their complications. Therefore, without surprise, bivariate correlations of each variable are small, which is typical in medical statistics. A pairs plot shows only a moderate correlation between `log1p_studyhr` and `age` ($\rho \approx 0.412$); that is, the expected study hours of typical children increase with age. In conjunction with all the variance inflation factors being below 1.5, it is reasonable to dismiss the concern of multicollinearity. The within-group variances are similar, even with the `age` coincidentally, except for the two square-rooted variables. The equality of covariance is not satisfied.

3 Methodologies and Model Diagnostics

3.1 Candidate Models

Upon approaching multivariate data, intuitively, Principal Component Analysis, a dimension reduction technique, is considered first. Principal Component Analysis (abbr. PCA) produces a linear space with a dimension equalling to the rank of the design matrix $\mathbf{X}$, by consecutively maximising the variance of each principal component under the constraint that the sum of squares of coefficients equal one and each principal component is orthogonal to all previous ones. As the most fundamental multivariate statistical model, PCA does

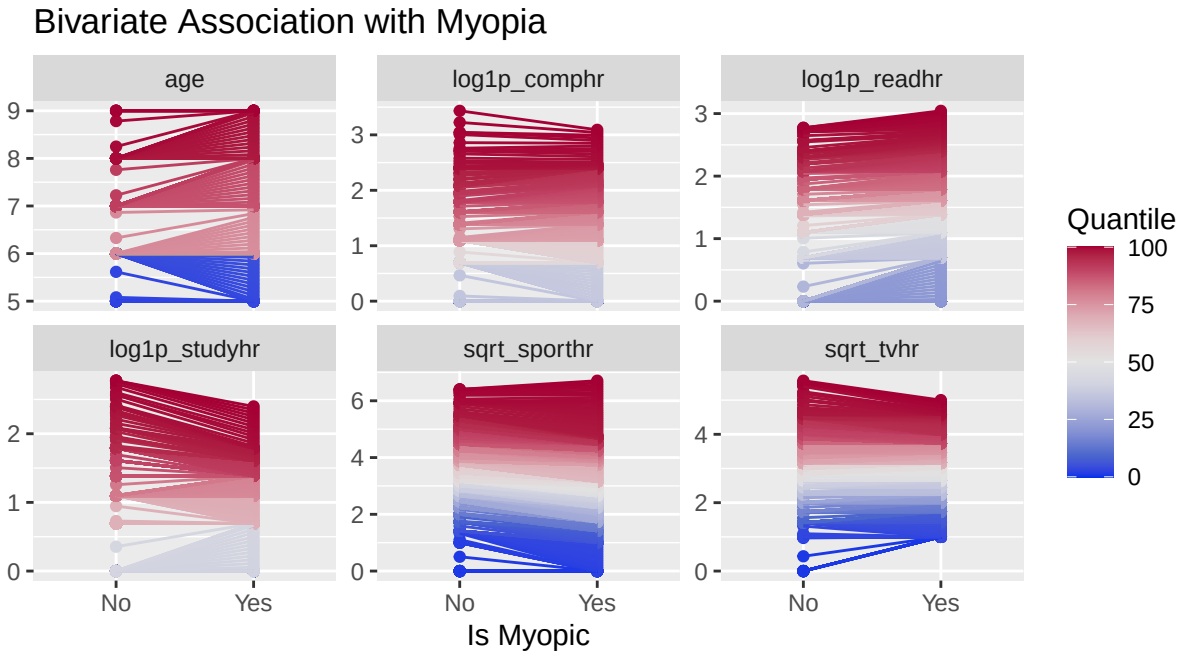

Figure 2: The bivariate correlations with myopia of the numeric variables are relatively small.

not preserve the relationship of the variables and thus ignores the features other than the maximal variance of the principal components upon dimension reduction. As a result, PCA might not separate the myopic groups well enough in achieving the main objectives.

As the objective is in an attempt to model and classify a single categorical variable with a set of numeric variables, The Linear Discriminant Analysis (abbr. LDA) is an alternative to PCA. Similar to PCA, LDA is a dimension reduction technique producing linear discriminant functions - linear combinations of the original variables. Instead of maximising the variability of each dimension, LDA attempts to maximise the *ANOVA F-statistic*, which represents the ratio of the between-group variance against the within-group variance. The maximum dimension of linear discriminant functions is one less of the number of levels. The implication for the myopic category is a single-dimension linear discriminant function. Also, LDA is particularly suitable for data with numeric variables on the same *natural* scale (i.e., the units of variables), which is not the case due to different transformations. The classification of categories is based on the Bayes Discriminant Rule, which compares the posterior probability, satisfying

$$f(\theta_i|\mathbf{x}) \propto f(\theta_i)f(\mathbf{x}|\theta_i) = \pi_i f(\mathbf{x}) \tag{1}$$

π_i represents the prior probability of developing myopia during childhood, $f(\mathbf{x})$ is the likelihood (conditional probabilities given the myopia status) from the sample. Without specifying the prior probability, the `lda()` function assumes a prior equalling sample proportion (Venables & Ripley, 2002), which is probably not a prior potentially giving the most accurate predictions. On the other hand, a US research on children myopia suggests a prevalence of 24% (Theophanous et al., 2018). As such, the first LDA fit takes the prior (0.76, 0.24). Nevertheless, it is precarious to assume this is the correct prior, and adjustments may be needed upon evaluating the predictive power of the model, including techniques like cross-validation.

Other alternative multivariate models to PCA and LDA are Quadratic Discriminant Analysis (abbr. QDA) and PLS-Discriminant Analysis (abbr. PLS-DA). QDA relaxes the assumption of equality of within-group covariance required for LDA by separating the groups with a quadratic surface. Under the expectation that the equality of covariance is violated, QDA may slightly out-perform and is potentially a better fit. Nonetheless, as QDA models the whole covariance matrix without dimension reduction as well as the use of

quadratic combinations, the selective interpretation of loadings for PCA and LDA is not possible. On the contrary, PLS-DA is a comprise of LDA towards PCA in rank deficiency and is probably unneeded.

It is tempting to compare the predictive power of the final model (from STATS 767) to other commonly used statistical models. As the question of interest is to model a Bernoulli (Yes/No) categorical variable with a set of numeric variables, the classical frequentist approach is multiple logistic regression. Hence, the performance of the final multivariate model (LDA) is compared against multiple logistic regression.

3.2 Model Fitting and Evaluation

This section compares the performance of LDA and QDA using leave-1-out cross-validation. A prior of (0.76, 0.24) is used for the preliminary models for evaluation and subsequent adjustments.

The results of leave-1-out cross-validation show a high specificity yet extremely low sensitivity for both LDA and QDA, notwithstanding a high overall correct classification rate, at (0.99, 0, 0.86) for LDA and (0.97, 0.02, 0.84) for QDA. The ridiculous results are partially due to the low sample likelihood of being myopic. Besides, the setting of the prior is also a major contributing factor, as $sensitivity \to 1$ as $\pi_1 \to 1$ yet $specificity \to 1$ as $\pi_0 \to 1$. On the other hand, $\mathbb{P}(Type\, I) \to 1$ as $\pi_1 \to 1$ and $\mathbb{P}(Type\, II) \to 1$ as $\pi_0 \to 1$. Therefore, instead of sole sensitivity or specificity, the goal is to search for a prior, which minimises the *balanced error rate*. Achieving such is equivalent to maximising the value of $sensitivity + specificity$, also known as the *sensitivity-specificity trade-off*. The area under curve of the *receiver operating characteristic* (ROC) curve computed using **DescTools** (Andri et mult. al., 2021) of LDA and QDA are 0.628 and 0.672, respectively, which are relatively similar. Given that both models possess similar predictive power, yet LDA allows selective interpretation of loadings which QDA does not, the decision is to select LDA, under a prior such that the *BER* is minimised, as the final model.

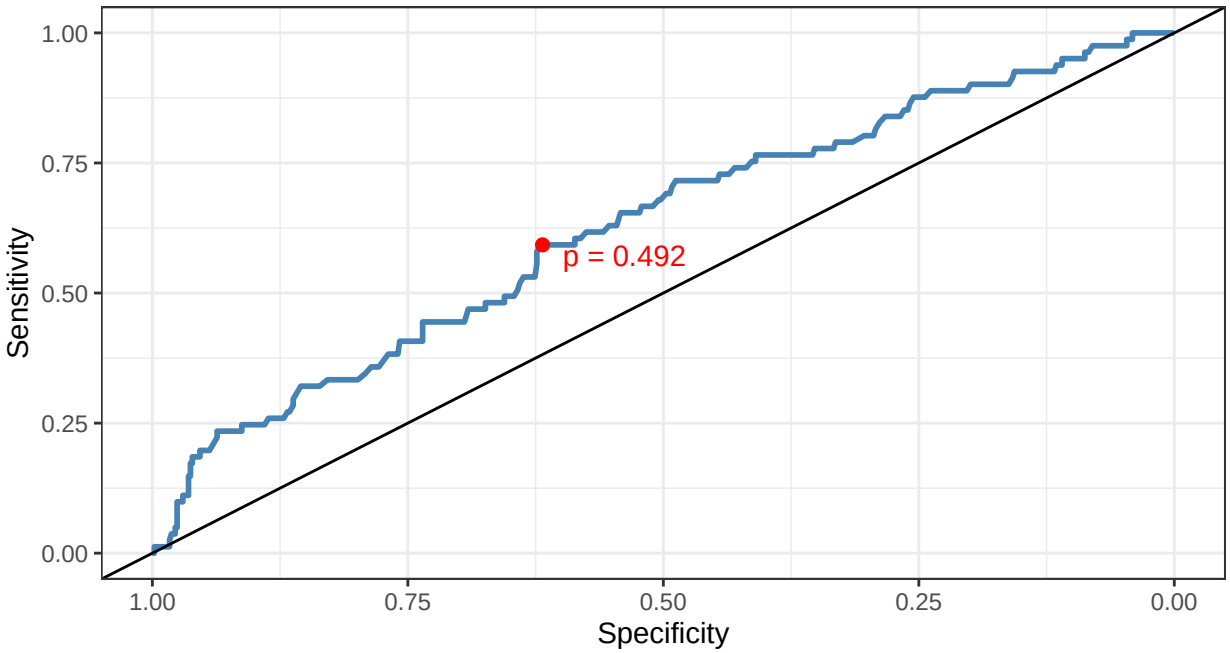

Figure 3: The receiver operating characteristic (ROC) curve of LDA

```
#> $sensitivity
#> [1] 0.5925926
#>
#> $specificity
```

```
#> [1] 0.6182495
#>
#> $auc
#> [1] 0.6276755
#>
#> $p_optim
#> [1] 0.492
```

The optimal prior (π_0, π_1) minimising BER is (0.508, 0.492), which is close to a uniform prior, giving a sensitivity of 0.593 and specificity of 0.618. The area under curve of the ROC curve is 0.628, indicating a weak-moderate predictive power ($AUC = 0.5$ is a random predictor). Again, this is typical and unsurprising in medical statistics. A weak relationship is claimable between lifestyle and myopia.

3.3 Model Output and Interpretation

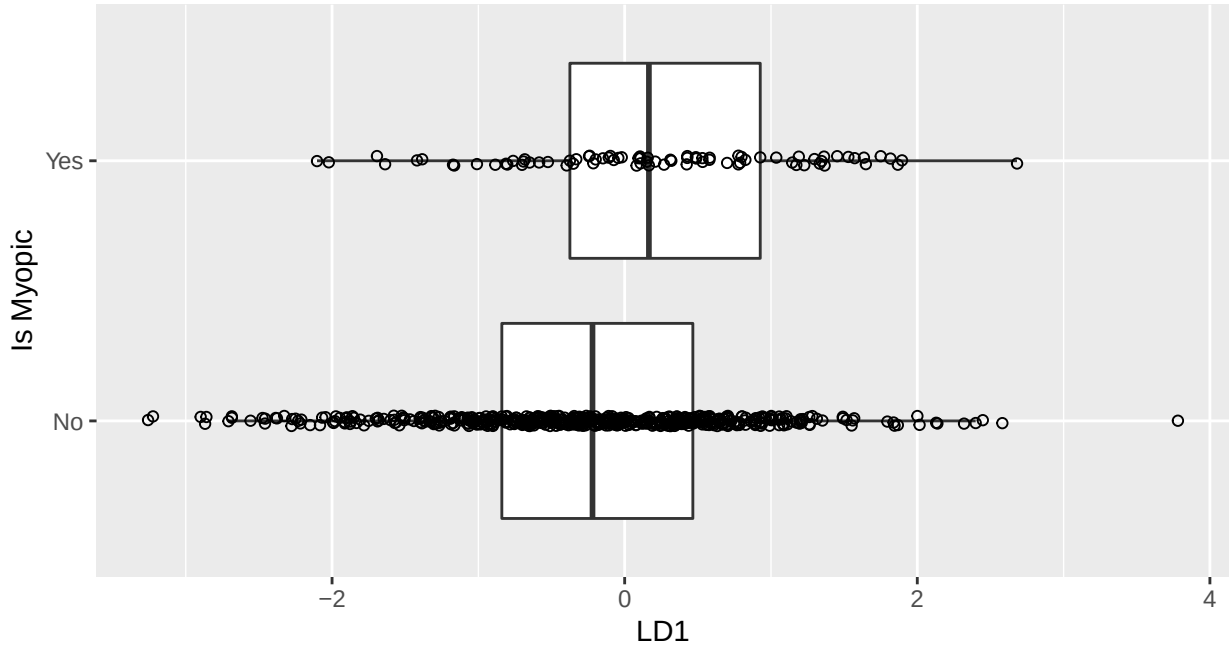

Figure 4: The plot of the linear discriminant scores

```
#>
#> Loadings:
#>               LD1
#> age            0.121
#> log1p_readhr   0.450
#> log1p_comphr
#> log1p_studyhr
#> sqrt_sporthr  -0.789
#> sqrt_tvhr
#>
#>                  LD1
#> SS loadings    0.849
#> Proportion Var 0.141
```

As the categorical variable is Bernoulli, there exists up to and only one dimension of linear discriminant function. The scatter-box plot of the one-dimensional linear discriminant scores indicates that `LD1` is higher in myopic children than not. The loadings of linear discriminant function show a weak positive correlation with `age`, moderate positive correlation with `log1p_readr` and high negative correlation with `sqrt_sporthr`. The implication is that the risk of myopia increases mildly with age, is moderately associated with a longer focused reading time, and strongly (negatively) associated with time spent in engaging with sports or outdoor activities. Although such findings cannot establish causation, they are on par with our typical belief about myopia. Surprisingly, the relationship of the time spent on studying, computer and television to myopia is unclear, given the linear discriminant scores and loadings.

4 Comparison with Classical Frequentist Approach

4.1 The Multiple Logistic Regression

Upon approaching a Bernoulli random variable of interest given a set of numeric explanatory variables, one would immediately think of a multiple logistic regression, whose model expression is given by

$$\mathbf{Y} \overset{iid}{\sim} Bernoulli(\boldsymbol{\theta}) \mid logit(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta};\ \boldsymbol{\theta} \in (0,1)^n,\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \mathbf{X} \in \mathbb{R}^n \times \mathbb{R}^p \tag{2}$$

A *full logistic regression model* `Y ~ .` is fitted, and the information-theoretic is adopted by searching every sub-model of the full model exhaustively and select the best model ranked by *AICc* (Barton, 2020). As of the LDA findings of linear subspace, quadratic transformation is unnecessary. The best model is

```
#>
#> Call:
#> glm(formula = MuMIn::get.models(logis_all, 1)[[1]][["formula"]],
#>     family = "binomial", data = myopia_2)
#>
#> Deviance Residuals:
#>     Min       1Q   Median       3Q      Max
#> -0.9656  -0.5704  -0.4959  -0.3943   2.3718
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept)   -1.2054     0.3515  -3.430 0.000604 ***
#> log1p_readhr   0.3291     0.1567   2.100 0.035712 *
#> sqrt_sporthr  -0.3382     0.1051  -3.218 0.001291 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 480.08  on 617  degrees of freedom
#> Residual deviance: 466.43  on 615  degrees of freedom
#> AIC: 472.43
#>
#> Number of Fisher Scoring iterations: 5
```

The *residual deviance* ($\mathcal{D} \approx 466.43$) is less than the model's degrees of freedom ($\nu = 615$), which dismisses the concern of an inadequate fit (overdispersion). The model provides moderate evidence for a positive association between time spent in focused reading and myopia and very strong evidence for a negative

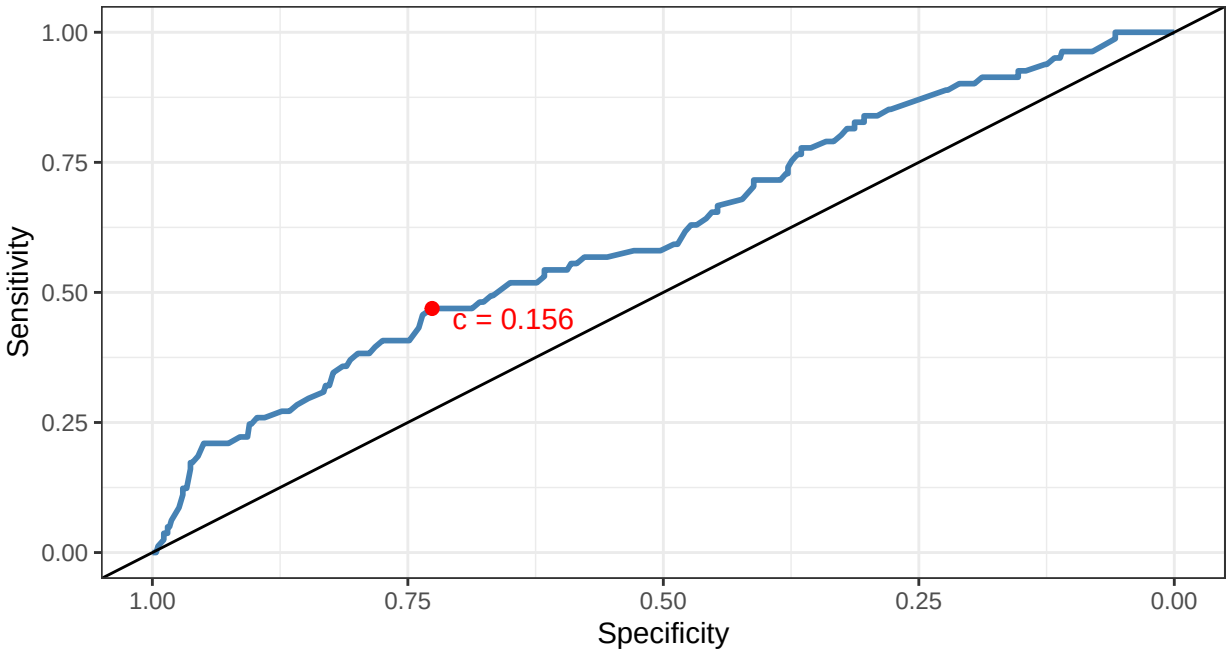

Figure 5: The receiver operating characteristic (ROC) curve of the multiple logistic regression

association between time spent in engaging with sports or outdoor activities and myopia. The findings from the logistic regression partially agree with LDA.

```
#> $sensitivity
#> [1] 0.4691358
#>
#> $specificity
#> [1] 0.726257
#>
#> $auc
#> [1] 0.6176058
#>
#> $c_optim
#> [1] 0.1555962
```

Using **pROC** package (Robin et al., 2011), the multiple logistic regression, at the optimal threshold for θ: $c_\theta \approx 0.156$, gives a sensitivity of 0.469 and specificity of 0.726, with an area under curve of *ROC* curve of 0.618, which is similar but slightly less than LDA and QDA. The classical frequentist method possesses similar yet ever so slightly weaker predictive power than the two multivariate ones.

4.2 Linear Discriminant Analysis cf. Multiple Logistic Regression

To avoid over-fitting, the information-theoretic approach to *GLMs* penalises additional explanatory variables to preserve degrees of freedom, which reflects upon *AICc* (similar to backward elimination of nonsignificant covariates). The multiple logistic regression, the *Binomial* family of *GLM*, adopts an include-or-exclude approach to covariates, completely dropping "unuseful" explanatory variables.

On the contrary, linear discriminant functions are linear combinations of all the variables, with weights such that the *ANOVA F-statistic* is maximised under certain constraints. Such an approach takes into account

the information from all (useful and useless) variables while assigning weights (LD coefficients) to them, which avoids the dropping of less significant variables yet with useful information completely.

Both LDA and the multiple logistic regression agree upon the significance of reading and outdoor time on children myopia, and the exception is age. While the multiple logistic regression drops all the other covariates, the linear discriminant loadings and coefficients of the numeric variables are non-zero. As the paper repeatedly emphasises, confounding effects of variables in medical statistics are extremely complicated, such that every single variable is potentially useful (or useless). Provided that the dataset has a considerably limited number of variables, dropping any variable should be considered cautiously.

5 Conclusion

The risk of developing myopia for children increases with age and time spent in focused reading, while the time spent on sports or outdoor activities is a negative risk factor. Such findings appear to coincide with our typical belief in children myopia. However, the relationship between computer and television use time and myopia is unclear. The *minimal-BER* LDA model possesses a weak-moderate predictive power with a sensitivity of 0.593 and a specificity of 0.618, as well as an area under curve of the *ROC* curve of 0.628. The dataset is better served for building an explanatory model than a predictive model.

6 Further Research

Recalling the linear discriminant loadings from the LDA model, the linear discriminant function, `LD1`, *appears* to be a measure of an "unhealthy lifestyle index", such that it appears to be correlated with perceivably unhealthy behaviours. Although such a claim is arguable, such as if studying or reading for too long is unhealthy remains highly questionable, as well as the need for additional lifestyle-reflecting explanatory variables, a hypothesis can be drawn from the preliminary findings for future research:

$$\textit{An overall unhealthy lifestyle is associated with the risk of developing myopia during childhood.} \quad (3)$$

In addition, Dr Beatrix Jones kindly advised the plausibility of modelling the relationship between the set of lifestyle-related variables and the prescription variables, such as Spherical-Equivalent Refraction (SPHEQ). Nonetheless, the majority of the sample has a spherical lens of below 0.75 (in absolute value), classified as non-myopic, as well as the maximum of SPHEQ is below 2, most of the variability of SPHEQ is meaningless for modelling myopia. On the contrary, LDA's factorisation and categorisation of SPHEQ remove the meaningless variation. Therefore, Canonical Correlation (abbr. CCorA) and PLS models are unsuitable for the data. Instead, they should be applied to studies for the *progression* of myopia; in the context that the data is sampled from myopic individuals of different degrees (e.g., low and high).

7 Reproducing The Paper

The paper can be reproduced with **rmarkdown** (Xie et al., 2018) from report.Rmd. Please install the required packages and their dependencies prior to knitting the document. Initial exploratory analyses, model building and diagnostics can be found in szmsu2011/stats767proj/prelim.

Bibliography

Andri et mult. al., S. (2021). *DescTools: Tools for descriptive statistics*. https://cran.r-project.org/package=DescTools

Barton, K. (2020). *MuMIn: Multi-model inference*. https://CRAN.R-project.org/package=MuMIn

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/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*, 77.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2020). *GGally: Extension to 'ggplot2'*. https://CRAN.R-project.org/package=GGally

Theophanous, C., Modjtahedi, B. S., Batech, M., Marlin, D. S., Luong, T. Q., & Fong, D. S. (2018). Myopia prevalence and risk factors in children. *Clinical Ophthalmology*, *12*. https://doi.org/10.2147/OPTH.S164641

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

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