Table of Contents

Table of Figures

1. Background information

The data refer to the Myopia study. These data are a subset of data from the Orinda Longitudinal Study of Myopia (OLSM), a cohort study of ocular component development and risk factors for the onset of myopia in children. Data collection began in the 1989–1990 school year and continued annually through the 2000–2001 school year. All data about the parts that make up the eye (the ocular components) were collected during an examination during the school day. Data on family history and visual activities were collected yearly in a survey completed by a parent or guardian.

The dataset used in this text is from 618 of the subjects who had at least five years of follow-up and were not myopic when they entered the study. All data are from their initial exam and includes 17 variables. In addition to the ocular data there is information on age at entry, year of entry, family history of myopia and hours of various visual activities. The ocular data come from a subject's right eye.

A subject was coded as myopic if they became myopic at any time during the first five years of follow-up. A detailed description of all the variables can be found in the attached pdf. Also the data are attached

The purpose of the project is to examine which variables contribute to the development of "Myopia within the first five years of follow up", measure by variable MYOPIC. The rest variables are potential candidates for examining the variable under study.

2. Descriptive analysis and exploratory data analysis

We read the dataset and understand its structure.

```
## 'data.frame':    618 obs. of  18 variables:
##  $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ STUDYYEAR: int  1992 1995 1991 1990 1995 1995 1993 1991 1991 1991 ...
##  $ MYOPIC    : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ AGE       : int  6 6 6 6 5 6 6 6 7 6 ...
##  $ GENDER    : int  1 1 1 1 0 0 1 1 0 1 ...
##  $ SPHEQ     : num  -0.052 0.608 1.179 0.525 0.697 ...
##  $ AL        : num  21.9 22.4 22.5 22.2 23.3 ...
##  $ ACD       : num  3.69 3.7 3.46 3.86 3.68 ...
##  $ LT        : num  3.5 3.39 3.51 3.61 3.45 ...
##  $ VCD       : num  14.7 15.3 15.5 14.7 16.2 ...
##  $ SPORTHR   : int  45 4 14 18 14 10 12 12 4 30 ...
##  $ READHR    : int  8 0 0 11 0 6 7 0 0 5 ...
##  $ COMPHR    : int  0 1 2 0 0 2 2 0 3 1 ...
##  $ STUDYHR   : int  0 1 0 0 0 1 1 0 1 0 ...
##  $ TVHR      : int  10 7 10 4 4 19 8 8 3 10 ...
##  $ DIOPTERHR: int  34 12 14 37 4 44 36 8 12 27 ...
##  $ MOMMY     : int  1 1 0 0 1 0 0 0 0 0 ...
##  $ DADMY     : int  1 1 0 1 0 1 1 0 0 0 ...
```

There are 618 observations in the dataset analyzed which concern the onset of myopia in children. The MYOPIC variable is considered the response or dependent variable. It falls into one of two categories, Yes (1) or No (0). Rather than modeling this response MYOPIC directly, the logistic regression model that will be used will measure the probability that Y belongs to a particular category. The rest of the variables are controlled and we will examine their impact and contribution to the myopic status of the subjects.

The variables regarding the ocular components (Spherical Equivalent refraction-SPHEQ-, Axial length-AL-, Anterior Chamber Depth-ACD-, Lens Thickness-LT- and Vitreous Chamber Depth-VCD-) are considered continuous variables.

The variables regarding the time spent on reading or other leisure activities (SPORTHR, READHR, COMPHR, STUDYHR, TVHR and DIOPTERHR) as well as the STUDYYEAR and AGE are discrete variables. GENDER and MOMMY, DADMY are categorical variables.

The initial transformations that will be performed to the dataset are the following:

- Excluding the ID variable as it has no valuable meaning for our analysis.
- Creation of a new variable called HEREDITY, out of the MOMMY and DADMY variables with the following 3 levels:
    - NONE: for those children where none of their grandparents are myopic,
    - ONE: for those children where one out of the two parents are myopic and
    - BOTH: for those children where both of their parents are myopic.
- The composite variable DIOPTERHR will not be used as the information that it provides is already included in the variables that is composed of (READHR,

STUDYHR, COMPHR, TVHR). This is because we want to examine the effect that each of the variables have on the MYOPIC variable.

- Conversion of variables MYOPIC, GENDER to factors and numeric the rest.

2.1 Univariate analysis of the variables

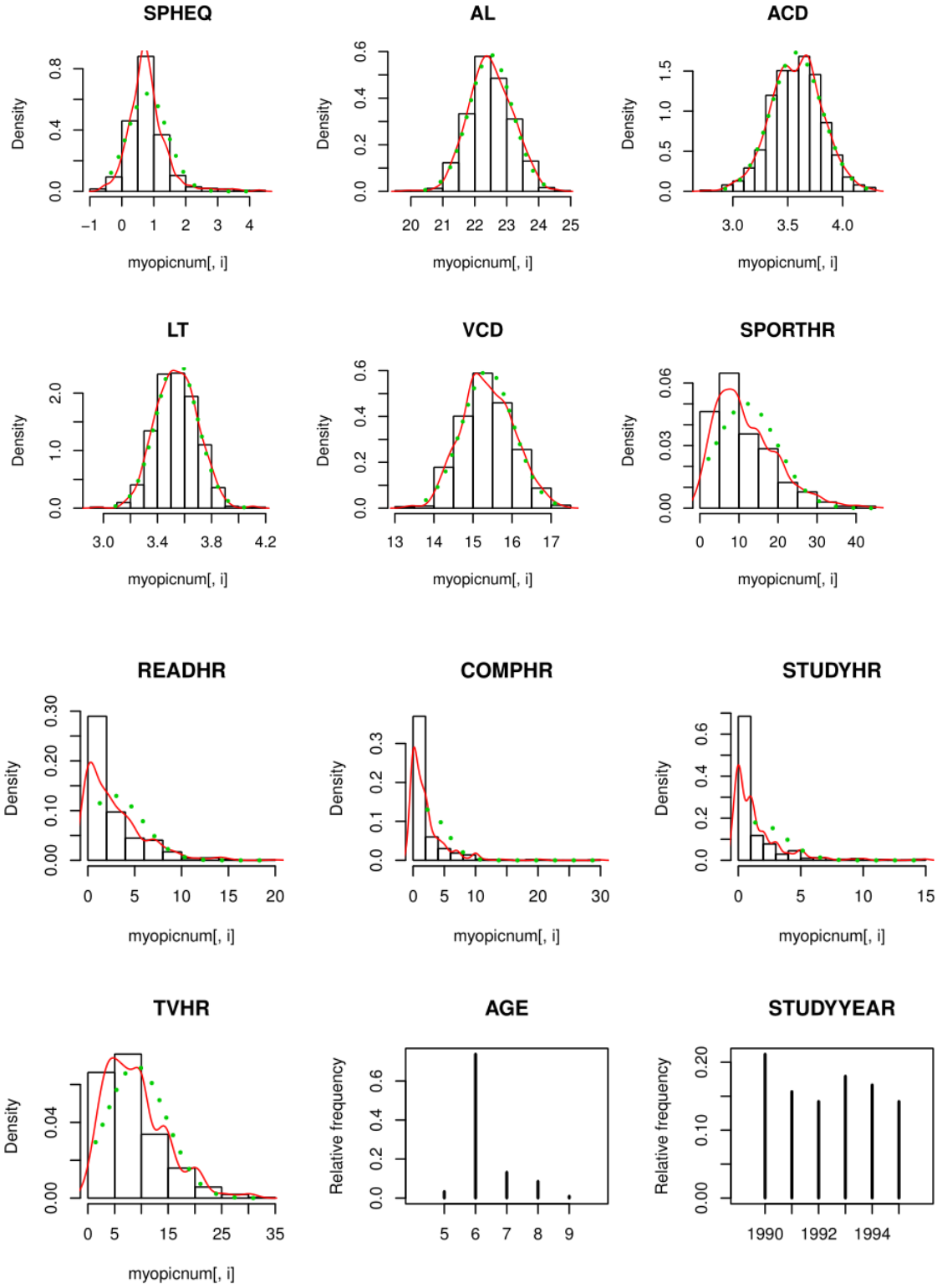


Figure 1: Univariate Analysis Histograms for the continuous and Frequency Plots for the discrete

The distribution of the spherical equivalent refraction (SPHEQ), the focusing power of the eye, appears to be right skewed a little with the median < mean (0.73 < 0.80). 50% of the subjects had for this measure values ranging from -0.69 to 0.73 and 50% from 0.73-4.37. In the beginning none of the subjects were myopic (SPHEQ <= -0.75).

The rest of the continuous variables concerning other ocular measures, i.e. the length of the eye from front to back (AL), the length of the containing space of the eye between the cornea and the iris (ACD), the length of the crystalline lens (LT) and the length of the containing space of the eye in front of the retina (VCD) seem to follow a normal distribution with the mean/median of the measurements to be approx. 22.5, 3.58, 3.54 and 15.4 respectively.

Regarding the variables that indicate how the children spent their time the distributions of the variables are right skewed. More specifically as far as the leisure time is concerned, 75% of the children spent up to 16 hours per week on sports followed by 4 hours reading for pleasure. The rest of the time, 75% of the children spent up to 2 hours per week on studying, up to 3 hours using computers and up to 12 hours watching TV.

Most of the children examined (ages from 5 to 9), were at the age of 6. Also, around equal number of observations took place through ’90-’95 with slightly large number to be the observations taken in ’90.

Let’s know check the bar plots of the categorical variables.

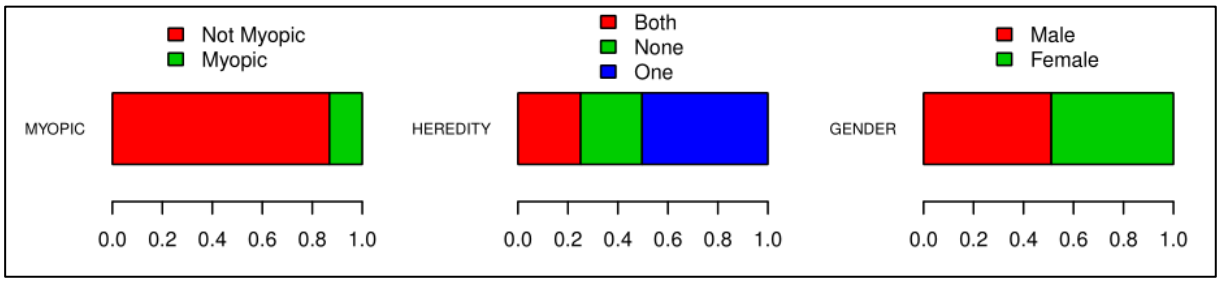


Figure 2: Bar plots of Categorical Variables

Considering the bar plots of the categorical variables, around 85% of the kids examined, didn’t develop myopia within the first five years of study, whereas the rest developed. As far as the heredity factor in myopia is concerned, approx.50% of the subjects had one out of the two parents with myopia, 25% had both parents with myopia and for the remaining, none of the parents had myopia. In the sample around 50% of the kids where males and 50% females.

2.2 Pairwise comparisons

The scatterplots below are used for pairwise comparisons among the numeric variables of our dataset.

Scatterplots

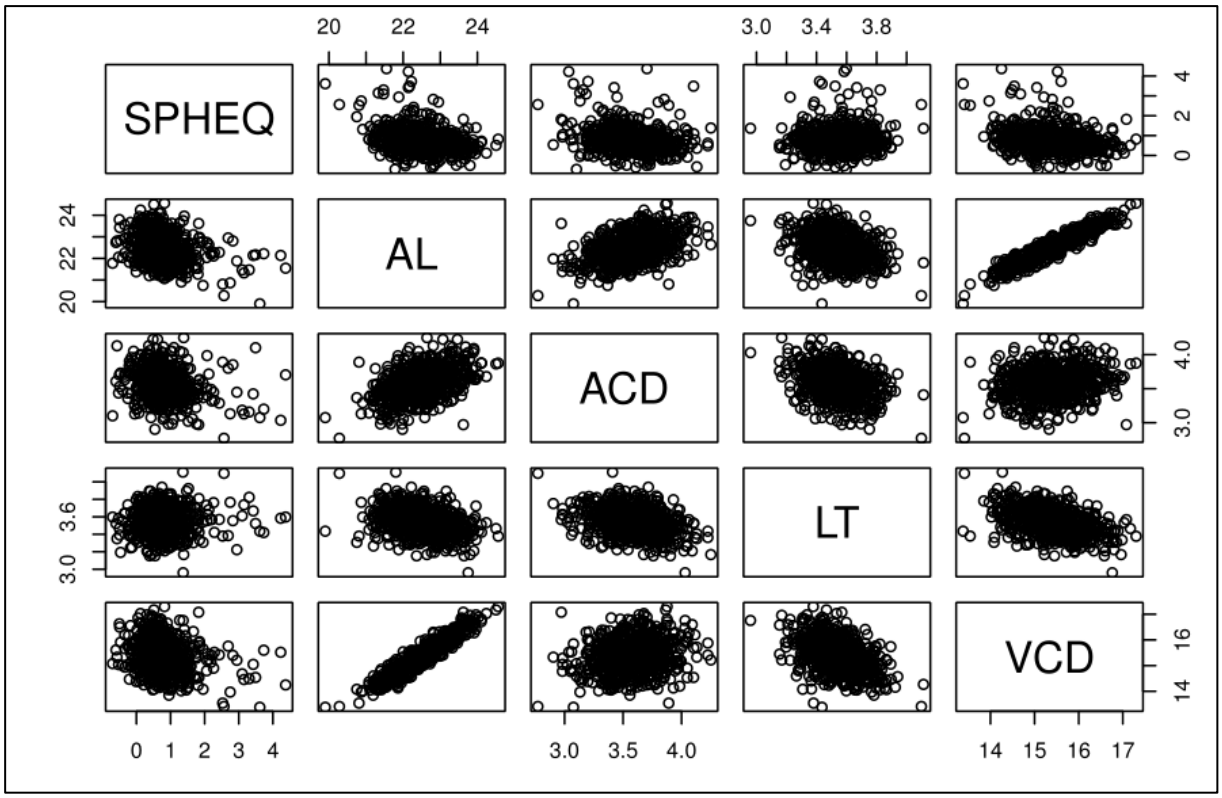


Figure 3: Scatterplots of the continuous numeric variables

The scatterplots above are used for pairwise comparisons among the numeric variables of our dataset. We can see that the measure for the eyes effective focusing power (SPHEQ) has a weak negative relationship with the AL, ACD and it doesn't seem to be related linearly with LT and VCD.

Regarding the relationships between the other explanatory variables it is worth mentioning that there is an almost perfect linear relationship between the variables AL- the length of the eye from front to back- and VCD- the containing space of the eye in front of the retina-. This could be an indication of multi-collinearity. This means that the two variables provide the same information and following one of the two might be excluded from the model.

2.3 Pearson's correlation coefficient for numerical variables

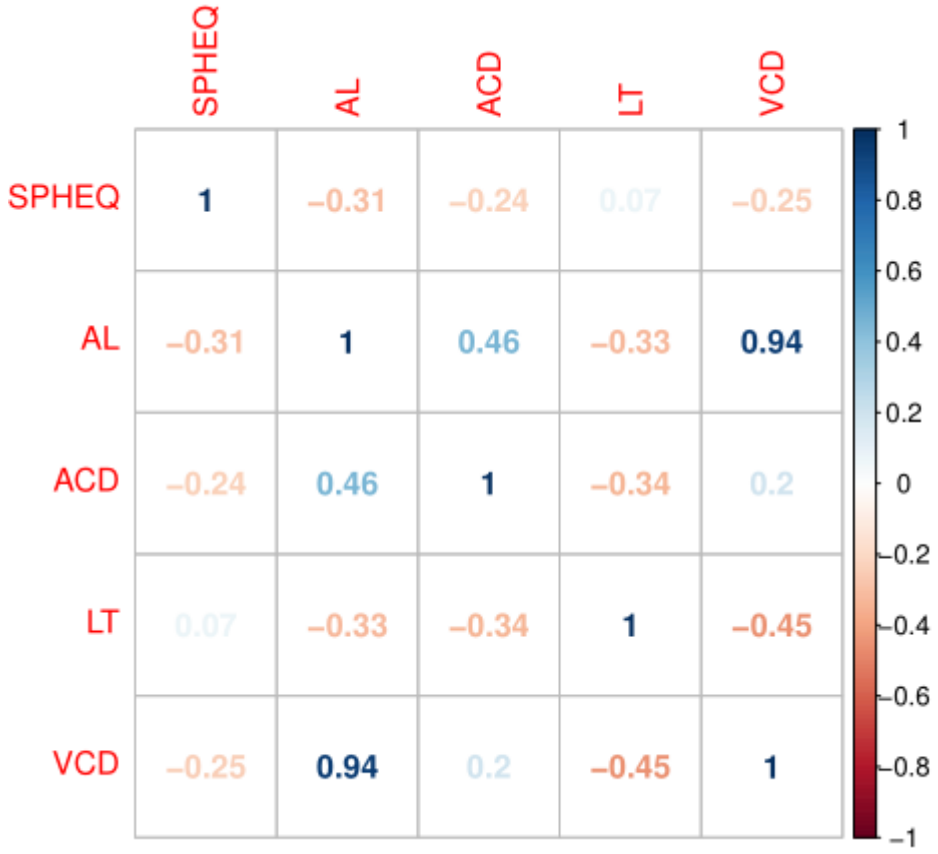


Figure 4: Pearson's Correlation for numerical Variables

Running the Pearson's correlation coefficient, we measure the degree of linear dependence between the variables. It seems that indeed there is a strong linear relationship between the variables AL -the length of the eye from front to back- and VCD - the containing space of the eye in front of the retina- (0.94), which is an indication of multi-collinearity. Additionally, AL has a weak positive linear relationship with the ACD variable- the length between the cornea and the iris- (0.46). There is a week negative linear relationship between the effective focusing power of the eye (SPHEQ) and the AL (-0.31). Similarly, the LT- the length of the crystalline lens - has weak negative linear relationships with the AL (-0.33), the ACD (-0.34) and the VCD (-0.45).

We have to mention here that if two explanatory variables are highly correlated between them, they would carry similar information and cause misleading results in predicting the response variable. Therefore, before proceeding with finding the best model for predicting MYOPIA, we will perform vif test to check the variable that causes the multi-collinearity between the explanatory variables and exclude it of the full model.

2.4 Boxplots for categorical variables

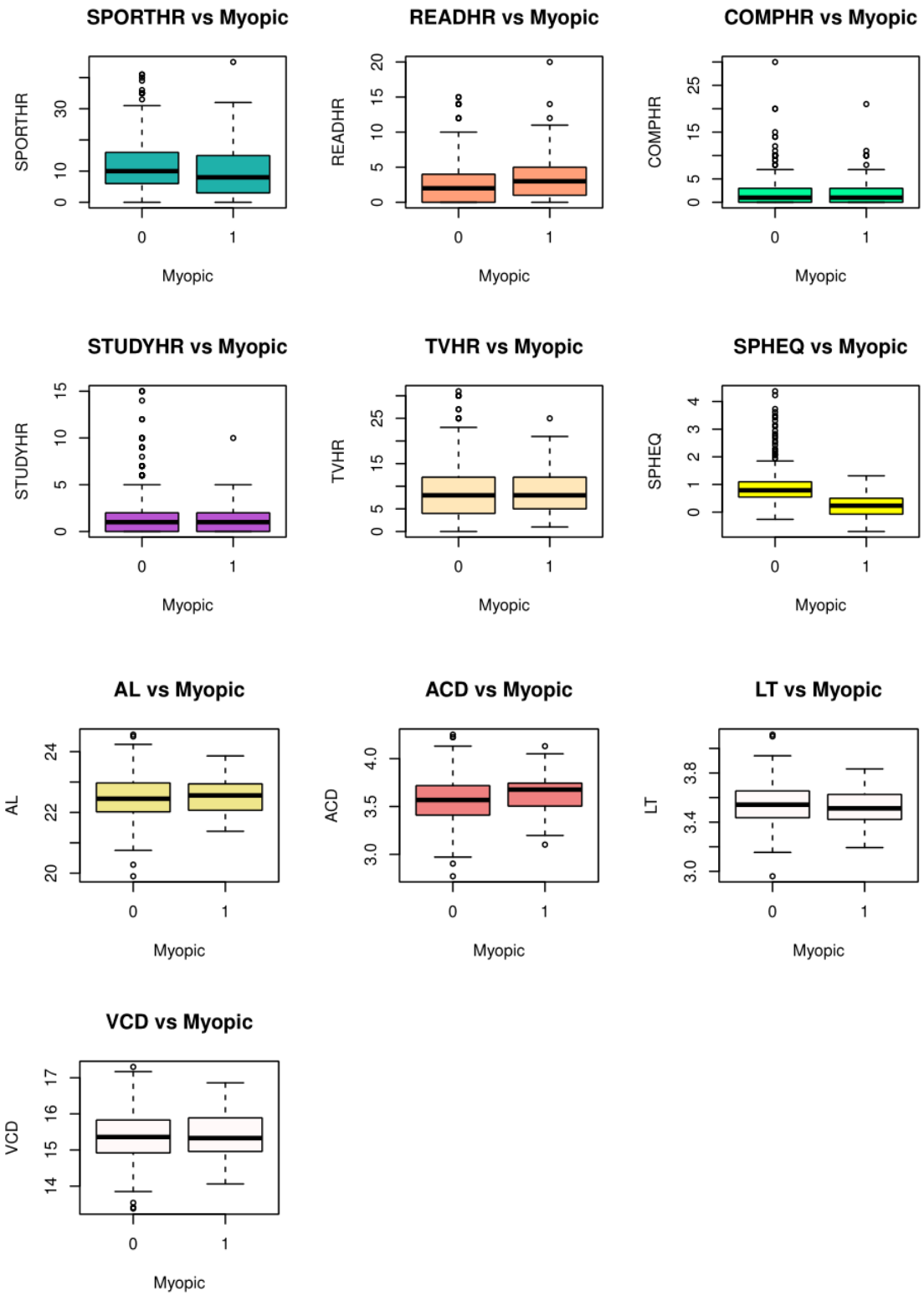
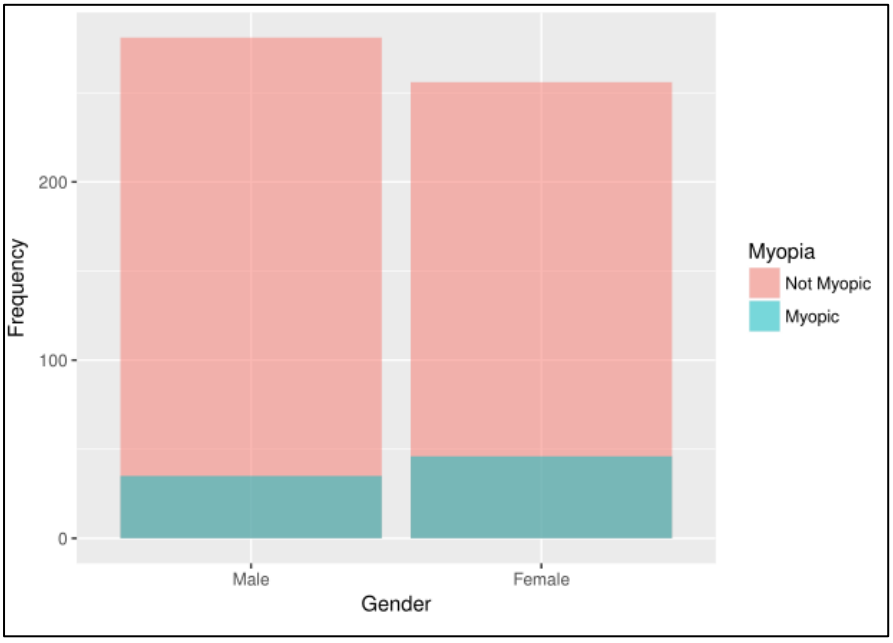


Figure 5: Boxplots for the response (categorical variable)

We plotted the boxplots above because we are particularly interested to check how the explanatory variables affect the response variable MYOPIC. We can say the following based on the boxplots above:

- Children that developed myopia, spent slightly less time (based on the median) doing sport activities than children who devoted more.
- Myopic children spent slightly more hours reading (based on the median).
- Time spent on studying or on playing/working on the computer doesn't seem to influence myopia as the boxplots are similar for myopic and non-myopic children.
- The variation of hours of the non-myopic children who watch TV is slightly larger than that of the myopic children.
- Children who developed Myopia seem to have initially less effective focusing power (SPHEQ) than those children who didn't develop.
- Myopic children appear to have lower variation of AL, ACD, LT, VCD values. Myopic children have higher ACD values (larger median) than non-myopic.



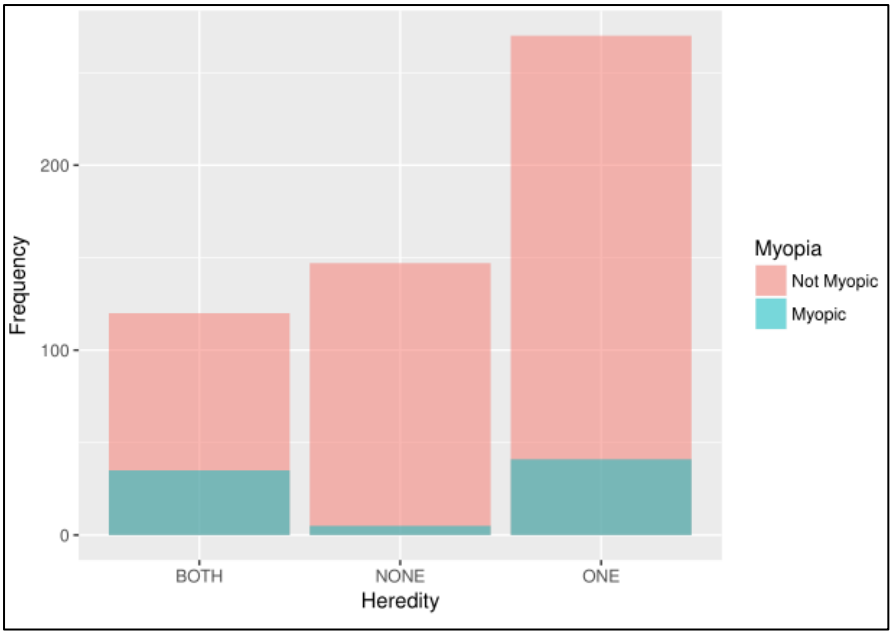


Figure 6: Bar plots for Categorical Variables (Gender and Heredity) and the response

Regarding the gender, slightly more girls developed myopia comparing to boys (top bar plot). Also, considering the heredity (bottom bar plot), similar number of children developed myopia who had either one or both of their parents with myopia. However, much less children (approx.. 1/9 of the other respective numbers of BOTH and ONE) developed myopia despite of the fact that none of their parents had.

3. Descriptive and Predictive models

Due to the fact that the response variable is binary (0 and 1) we will perform logistic regression. Rather than modeling the response MYOPIC directly, the logistic regression model that will be used will measure the probability that Y belongs to one out of the two particular categories. We don't use a linear regression model, because for response variables close to zero we predict a negative probability of MYOPIC; if we were to predict for very large response variables, we would get values bigger than 1. These predictions are not sensible, since of course the true probability of MYOPIC, must fall between 0 and 1. That is why we use the logistic function.

Before proceeding to run the full model, we will check the multicollinearity among the explanatory variables:

	GVIF	Df	GVIF^(1/(2*Df))
GENDER	1.319020	1	1.148486
HEREDITY	1.115415	2	1.027683
AGE	1.762798	1	1.327704
STUDYYEAR	1.626800	1	1.275461
SPHEQ	1.160406	1	1.077221
AL	29332.345186	1	171.266883
ACD	3371.604808	1	58.065522
LT	1524.011873	1	39.038595
VCD	27728.859930	1	166.519848
SPORTHR	1.134895	1	1.065315
READHR	1.222779	1	1.105793
COMPHR	1.170368	1	1.081835
STUDYHR	1.319566	1	1.148723
TVHR	1.163895	1	1.078839

	GVIF	Df	GVIF^(1/(2*Df))
GENDER	1.314582	1	1.146552
HEREDITY	1.114507	2	1.027474
AGE	1.766035	1	1.328922
STUDYYEAR	1.622657	1	1.273835
SPHEQ	1.151812	1	1.073225
ACD	1.292905	1	1.137060
LT	1.436119	1	1.198382
VCD	1.390096	1	1.179023
SPORTHR	1.136859	1	1.066236
READHR	1.216499	1	1.102950
COMPHR	1.166356	1	1.079980
STUDYHR	1.325720	1	1.151399
TVHR	1.157535	1	1.075888

Figure 7: VIF with and without the variable AL that causes multi-collinearity

We run the VIF test and find that the AL variable causes collinearity. Its value has the largest (GVIF^(1/2*Df value>3.16). Also, this finding is aligned with the results from the exploratory data analysis. Thus, we remove this variable from the full model and run VIF again to check if we still have a problem regarding multi-collinearity. After running again, the VIF test (2nd graph of Figure 7) we see that the problem corrected. We now run the generalized linear model, as the full model without the AL variable and check the results.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6869  -0.3913  -0.2043  -0.0654   3.1899

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.443759   8.316394   0.655  0.51274
GENDER1       0.605961   0.343588   1.764  0.07780 .
HEREDITYNONE -1.596244   0.554932  -2.876  0.00402 **
HEREDITYONE  -0.656504   0.328410  -1.999  0.04560 *
AGE           0.085374   0.254087   0.336  0.73687
STUDYYEAR     0.125286   0.110128   1.138  0.25527
SPHEQ        -4.105049   0.468394  -8.764  < 2e-16 ***
ACD           1.108876   0.765488   1.449  0.14745
LT           -0.967589   1.201649  -0.805  0.42069
VCD          -0.396468   0.275945  -1.437  0.15079
SPORTHR      -0.046915   0.021546  -2.177  0.02945 *
READHR        0.085515   0.050634   1.689  0.09124 .
COMPHR        0.043609   0.046915   0.930  0.35261
STUDYHR      -0.171820   0.099399  -1.729  0.08388 .
TVHR         -0.008584   0.028856  -0.297  0.76609
---
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: 298.77  on 603  degrees of freedom
AIC: 328.77
```

Figure 8: Full Model

The figure above shows the coefficient estimates and related information that result from fitting a regression model on the Myopic data in order to predict the probability that the children are myopic using all the other explanatory variables. We see that the coefficient of the effective eye focusing power SPHEQ is -4.10. This indicates that a decrease in SPHEQ is associated with an increase in the probability of myopia. To be more precise, a one-unit decrease in SPHEQ is associated with an increase in the log odds of myopia by 4.10 units, considering that all other predictor variables are constant. These odds are low (exp(-4.10)=0.01657268). Regarding the z-statistic a large (absolute) value of it indicates evidence against the null hypothesis that the covariates can be set equal to 0 (i.e. that the Myopia doesn’t depend on the specific explanatory variables). Also, by checking the p-values the variables SPHEQ, HEREDITY and SPORTHR are considered to be statistically significant. Considering the HEREDITY, the coefficient is negative. This means that those kids that none of their parents have myopia are less likely to have myopia to those kids that both parents have myopia, taking into consideration that all other variables are fixed. For the rest of the coefficients whose p-value is more than 0.05, we cannot reject the Ho that they can be set equal to 0.

3.1 Stepwise procedure

We perform stepwise method on the full model to check whether we can use a simpler function with less coefficients.

```
Call:
glm(formula = MYOPIC ~ GENDER + HEREDITY + SPHEQ + ACD + SPORTHR +
    READHR + STUDYHR, family = binomial, data = myopic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7280  -0.4068  -0.2099  -0.0644   3.1871

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.37858    2.60876  -1.295  0.19529
GENDER1       0.64426    0.31265   2.061  0.03934 *
HEREDITYNONE -1.70031    0.54960  -3.094  0.00198 **
HEREDITYONE  -0.61409    0.32167  -1.909  0.05625 .
SPHEQ        -3.93212    0.44939  -8.750  < 2e-16 ***
ACD           1.18550    0.70117   1.691  0.09088 .
SPORTHR      -0.05353    0.02065  -2.593  0.00953 **
READHR        0.07596    0.04826   1.574  0.11549
STUDYHR      -0.17417    0.09042  -1.926  0.05406 .
---
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: 303.28  on 609  degrees of freedom
AIC: 321.28

Number of Fisher Scoring iterations: 7
```

Figure 9: After stepwise

The stepwise procedure resulted in seven variables with better AIC, comparing to the previous one (321.28 instead of 328.77). This model is one that derived by rewarding the goodness of fit but minimizing at the same time the information loss.

Goodness of fit of the model after conducting stepwise procedure:

	actual	
predicted	0	1
0	521	49
1	16	32

We have 553/618 correct predictions (89.4%). We are thinking now that again out of 8 only 4 predictors are significant. The other may contribute to overfitting. They should, therefore, be eliminated. In order to do that we will use LASSO as an alternative for model selection.

3.2 LASSO

We create the model matrix of the full model (before excluding the multicollinear variable) and remove the intercept. In the Lasso lambda quantity is used as a penalty. We can visualize the coefficients. As the amount of penalty increases, all coefficients go to 0.

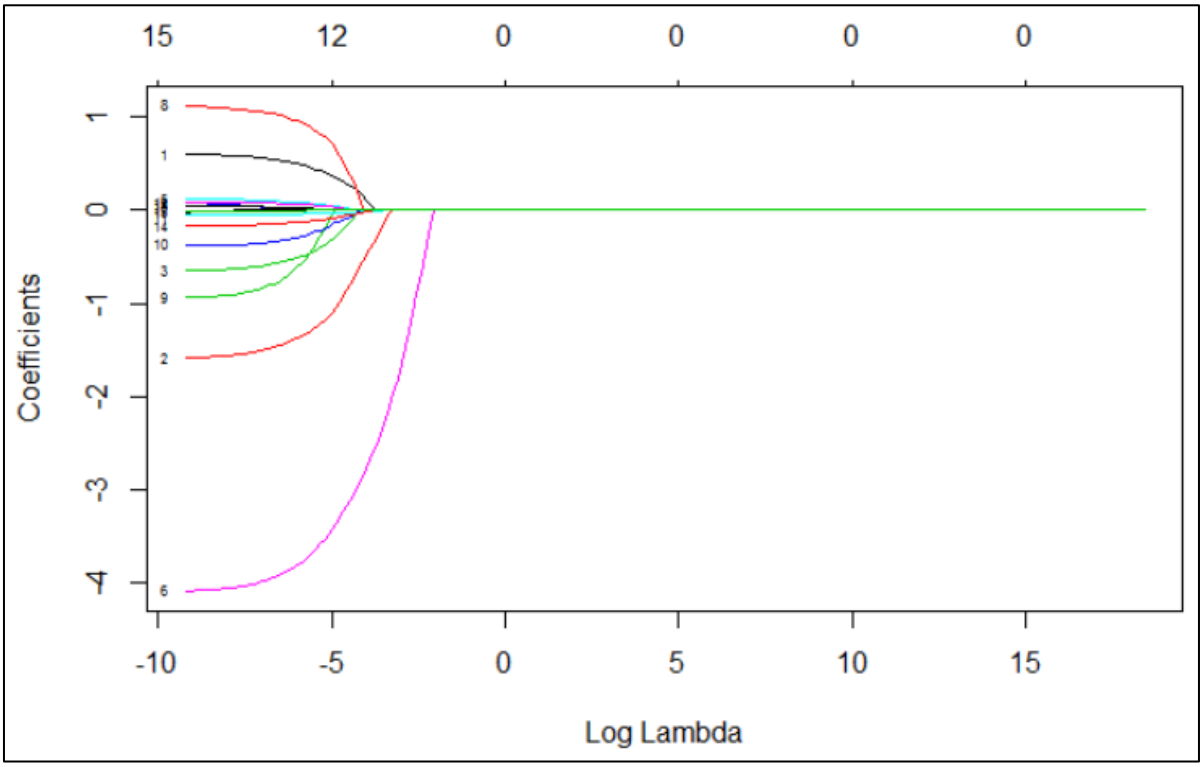


Figure 10: λ and coefficients

The question that arises is which is the appropriate value to choose for λ.

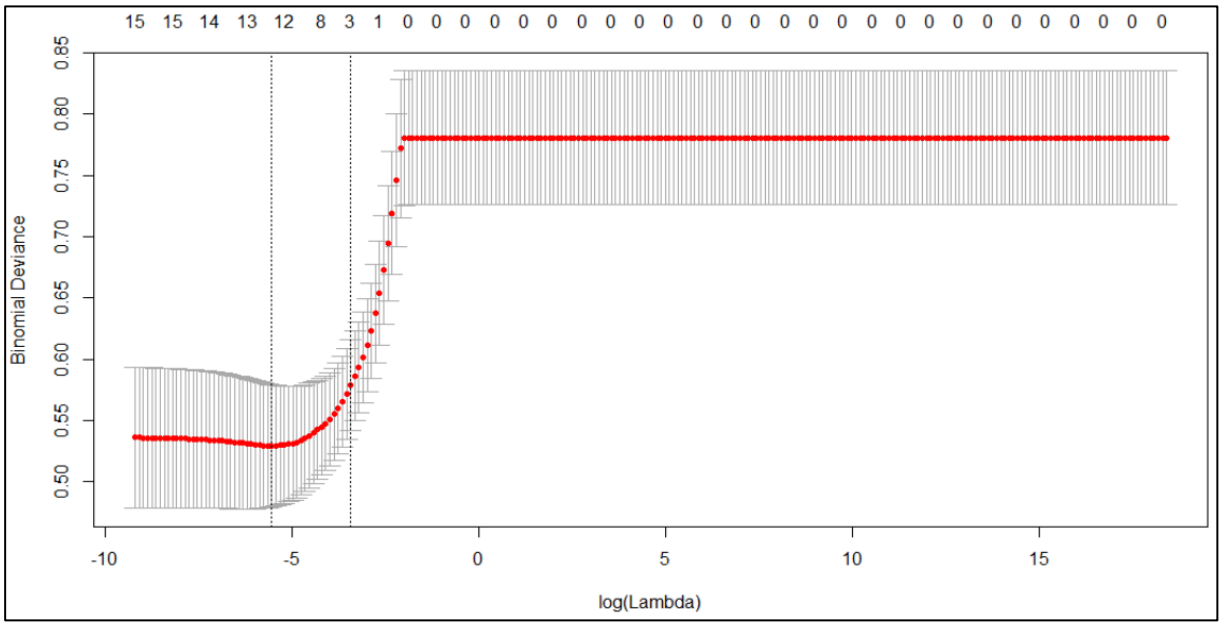


Figure 11: Cross Validation curve with error bars

The plot in Figure 11 illustrates the cross-validation curve (red dotted line), and upper and lower standard deviation curves along the lambda sequence (error bars). Two selected lambdas are indicated by the vertical dotted lines:

- **lambda.min** which is the value of lambda that gives minimum mean cross-validated error (log(lasso1$lambda.min)= -5.54).
- **lambda.1se** which gives the most regularized model such that, error is within one standard error of the minimum (log(lasso1$lambda.1se)= -3.44. This is done in order to be inside the confidence interval of the minimum and simultaneously don't lose accuracy.
    - The resulted model is the following.

```
> coef(lasso.cv, s = "lambda.1se")
16 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)   -0.548561491
GENDER1        .
HEREDITYNONE  -0.109801997
HEREDITYONE    .
AGE            .
STUDYYEAR      .
SPHEQ         -2.236184354
AL             .
ACD            .
LT             .
VCD            .
SPORTHR       -0.001121784
READHR         .
COMPHR         .
STUDYHR        .
TVHR           .
```

    - Goodness of fit of the model derived by LASSO with lambda.1se. We have 546/618 correct predictions (88.3%).

	actual	
predicted	0	1
0	535	70
1	2	11

3.3 Comparison of results between Stepwise and LASSO

The variables that contribute to overfitting can be eliminated using lasso. Thus, we get a more parsimonious model without compromising accuracy (in-sample). This accuracy is a bit less comparing to the value that we got by the more complex model (88.3% instead of 89.4%), resulted from the stepwise method. In other words, we get an almost similar accuracy by using a much simpler function (3 non-zero coefficients) than the original one (8 non-zero coefficients).

In Conclusion, we select as final model the model derived from LASSO. We will re-run the model in order to get the correct value of coefficients and interpret them accordingly.

3.4 Interpretation of the final model

In order to interpret the coefficients more easily, we will change the reference level of the HEREDITY variable from HEREDITY BOTH to HEREDITY NONE (the coefficients of the other explanatory variables will remain the same. The only change is on the coefficients of the levels of heredity(will become positive) and the intercept(will become negative). Before the change the interpretation of the intercept was exp(1.15747) = 3.18 i.e. the log odds of a kid with zero time spent on sports, 0 eye focusing power and that has both of his parents myopic had rather increased odds to develop myopia, which sounds sensible based on the exploratory analysis we have conducted.)

```
Call:
glm(formula = MYOPIC ~ HEREDITY + SPHEQ + SPORTHR, family = binomial,
    data = myopic2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.77270  -0.43454  -0.23257  -0.08022   3.05314

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.57173    0.56325  -1.015   0.3101
HEREDITYBOTH  1.72920    0.54112   3.196   0.0014 **
HEREDITYONE   1.09484    0.52709   2.077   0.0378 *
SPHEQ        -3.82479    0.43389  -8.815   <2e-16 ***
SPORTHR      -0.04828    0.01977  -2.441   0.0146 *
---
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: 316.32  on 613  degrees of freedom
AIC: 326.32
```

Figure 12: Final Model

The mathematical formulation of the model is the following:

$$p(MYOPIC = 1) = \frac{e^{-0.57+1.73 \cdot HEREDITYBOTH+1.09 \cdot HEREDITYONE-3.82 \cdot SPHEQ-0.04 \cdot SPORTHR}}{1 + e^{1.16+1.73 \cdot HEREDITYBOTH+1.09 \cdot HEREDITYONE-3.82 \cdot SPHEQ-0.04 \cdot SPORTHR}}$$

In this case, the estimated coefficient for the intercept is the log odds of a kid with zero time spent on sports, 0 eye focusing power and that has none of his parents myopic. In other words, the odds of being myopic when the all the other variables are set to zero, is exp(-0.57173) = 0.5645479 which are below 50%.

Regarding how heredity affects myopia, odds of a child to develop myopia when both of his parents have myopia are (exp(1.72920)= 5.636143) compared to a child that none of his parents has myopia. This is an increase of 463% in the overall odds towards the development of myopia on top of the odds of a child that none of his parents has myopia.

Similarly, odds of a child to develop myopia when one out of his parents have myopia are (exp(1.09484)= 2.988704) compared to a child that none of his parents have myopia. This is an increase of 198% in the overall odds towards the development of myopia on top of the odds of a child that none of his parents has myopia.

As far as the interpretation of SPHEQ variable is concerned, if we compare two children that one has 1 diopter more focusing power than the other and the rest of the variables same the odds that this child will develop myopia within five years is a lot less likely(exp(-3.82479)=0.0218230).

Regarding the interpretation of the SPORTHR variable, engagement in an additional hour of s ports activities per week will cause the event of myopia within five years to become less likely . However, the odds are close to 1 exp(-0.04828)=0.9528669 which means that developing my opia is close to 50-50 percent chance when all other variables don't change.