Myopia Study

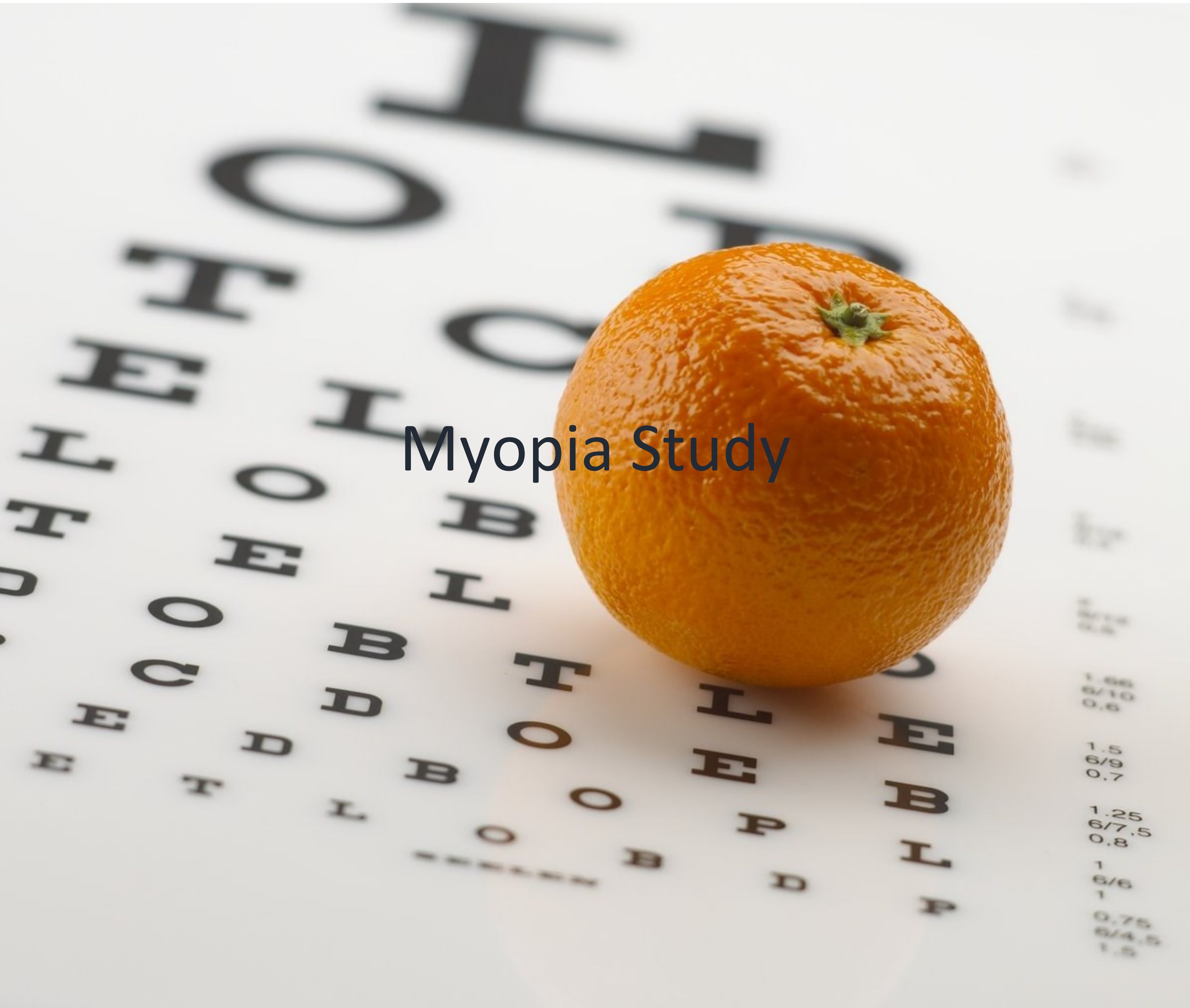

Table of Contents

Introduction – Description of the Problem

Myopia (near-sightedness or short-sightedness) is the most common eye problem and is estimated to affect 1.5 billion people (22% of the population). It is a condition of the eye where light focuses in front, instead of on the retina (see Figure 1 - Myopia Focus Problem). This causes distant objects to be blurry while close objects appear normal.

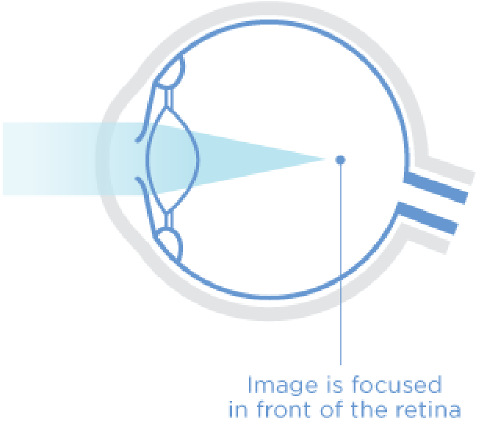

Figure 1 - Myopia Focus Problem[1]

The underlying cause is believed to be a combination of genetic and environmental variables. A family history of the condition is likely to play a role. Environmental risk factors include doing work that involves focusing on close objects and greater time spent indoors. There is tentative evidence that near-sightedness can be prevented by having young children spend more time outside. This may be related to natural light exposure.

Here we study myopia's various likely influencing variables using General Linear Models (GLM) and especially Logistic Regression. We examine Physiological variables (age, gender, eyeball parameters), Environmental variables (time spent on near-work and outdoor activities) as well as Hereditary variables (myopic mother or/and father). By doing the analysis, we examine the validity of the aforementioned hypotheses.

[1] Image from: http://www.cliniqueinvisia.com/en/vision/

Dataset

The data refers 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. 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 (**Spherical Equivalent Refraction > -0.75 D**) when they entered the study. All data are from their initial exam and includes 17 variables (see Table 1 - Variables Table Description). Subjects are characterized as myopic if they become myopic at any time during the first five years of follow-up. Also I would like to mention that a person with Spherical Equivalent Refraction between -0.25 D and +1.00 D doesn't have effective focusing power, maybe need glasses or contact lenses but he's **not** a myopic person.

Col.	Description	Value/Unit	Name
1	**Year subject entered the study** (numerical variable)	year	STUDYYEAR
2	**Myopia within the first five years of follow up** (categorical variable)	0 = No; 1 = Yes	MYOPIC
3	**Age at first visit** (numerical variable)	years	AGE
4	**Gender** (categorical variable)	0 = Male; 1 = Female	GENDER
5	**Spherical Equivalent Refraction** (numerical variable)	diopter	SPHEQ
6	**Axial Length** (numerical variable)	mm	AL
7	**Anterior Chamber Depth** (numerical variable)	mm	ACD
8	**Lens Thickness** (numerical variable)	mm	LT
9	**Vitreous Chamber Depth** (numerical variable)	mm	VCD
10	**Time spent engaging in sports/outdoor activities** (numerical variable)	hours per week	SPORTHR
11	**Time spent reading for pleasure** (numerical variable)	hours per week	READHR
12	**Time spent playing video games/working on the pc** (numerical variable)	hours per week	COMPHR
13	**Time spent reading/studying for school assignments** (numerical variable)	hours per week	STUDYHR
14	**Time spent watching television** (numerical variable)	hours per week	TVHR
15	**Composite of near-work activities** (numerical variable)	hours per week	DIOPTERHR
16	**Myopic Mother** (categorical variable)	0 = No; 1 = Yes	MOMMY
17	**Myopic Father** (categorical variable)	0 = No; 1 = Yes	DADMY

Table 1 - Variables Table Description

Descriptive Analysis

At this stage we import data in **R studio**. We observed that the ID column is the id of the records, so we excluded it, because it's useless information for our analysis. Variable STUDYYEAR is a chronological trend and sampling process and AGE is the age of subjects; children **between 5 to 9 years old**.

About the physiological variables axial length (AL), anterior chamber depth (ACD), lens thickness (LT) and vitreous chamber depth (VCD), there is a full linear relationship between them: (see Figure 2 - Human Eye Anatomy)

$$AL = ACD + VCD + LT$$

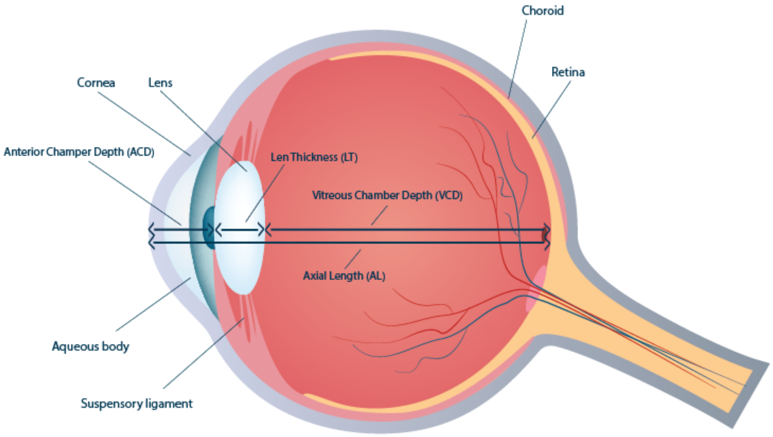

Figure 2 - Human Eye Anatomy[2]

In model we have to select if we want to keep the axial length or the rest variables. If we keep all the variables our model will suffer from multicollinearity (VIF>>1).

About the environmental variables, we have the Time spend reading for pleasure (READHR), for playing video/computer games or working on the computer (COMPHR), for reading or studying (STUDYHR) and the Time spend watching television (TVHR). But we also have the variable DIOPTERHR which is a linear combination of all these 4 variables:

$$DIOPTERHR = 3 * (READHR + STUDYHR) + 2 * COMPHR + TVHR$$

[2] Background image from: http://www.allaboutvision.com/resources/anatomy.htm

So, we can’t have to our model all these environmental variables and DIOPTERHR simultaneously. About hereditary variables, we observe that if mother or father are myopic then the approximately 70% of the children are myopic. Hence, it could be more meaningful to our analysis to combine them to one new variable PARENTSMY (at least on parent is myopic=1, neither parent is myopic= 0)

At this stage we can make a pairwise comparison of all the numerical variables, to observe if we have any correlations between them (see Figure 3 - Numeric Variables Correlations). In the following figure there is a high correlation between AL and VCD (|values| ≈1), all the other correlations are not strong (|values| <0.5).

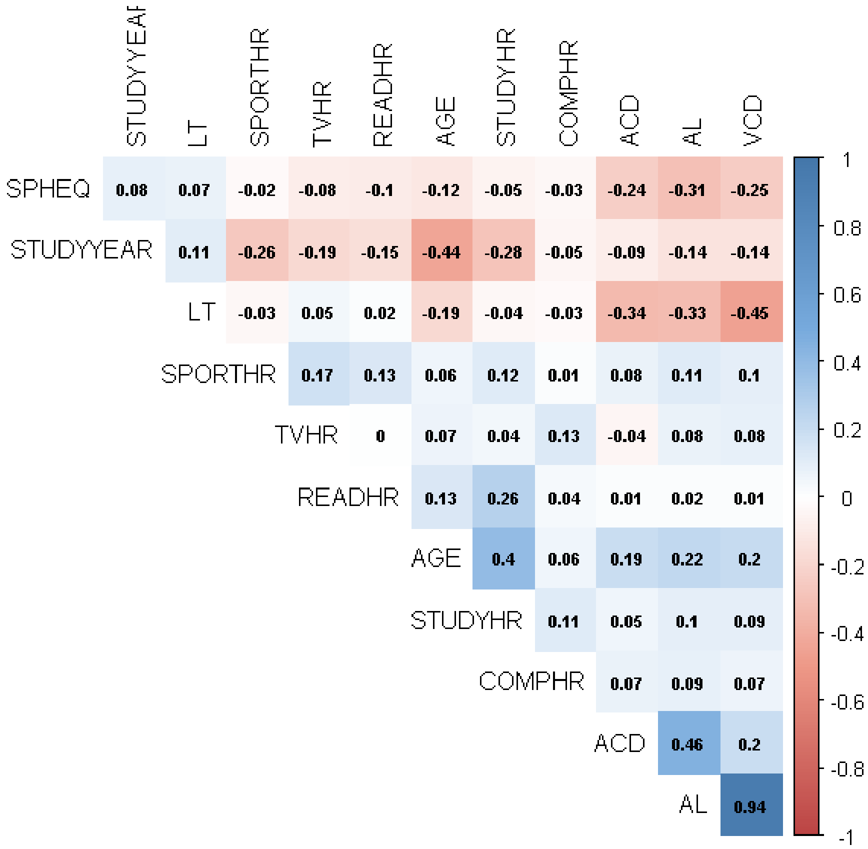

Figure 3 - Numeric Variables Correlations

At this stage we have to investigate the relation between the numerical variables and the response (MYOPIC). Hence, we create **boxplots** between numerical variables and MYOPIC (see Figure 4 - Numerical Variables Distributions). So, we can understand the distribution of these variables in myopic and healthy subjects. In the following plots, all variables have homogeneity between healthy and myopic subjects, except the Spherical Equivalent Refraction (SPHEQ), people with low values of SPHEQ tend to be myopic.

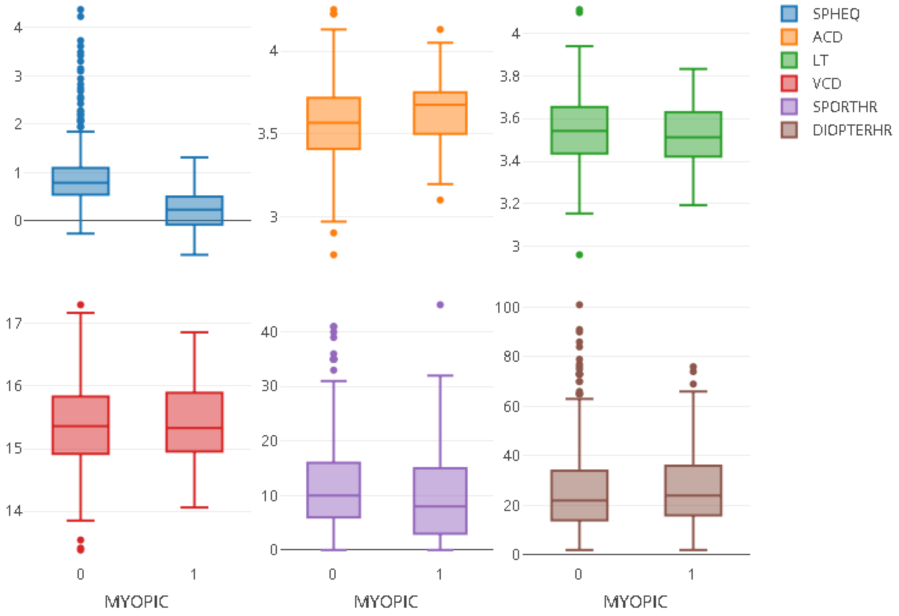

Figure 4 - Numerical Variables Distributions

Finally, in order to have a better picture of our predictors in our model, we create bar plots for our both categorical variables GENDER and PARENTS (see in Figure 5 - Categorical Variables Distribution). Myopia doesn't differ significantly between males and females (X-Square test- Pr>0.05), on the other hand, approximately the 70% of the myopic people tend to have one of their parents myopic.

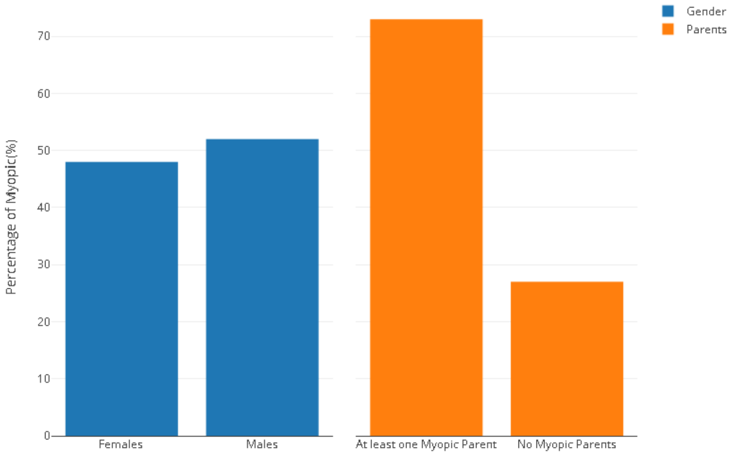

Figure 5 - Categorical Variables Distribution

Model Selection

Our dependent variable (MYOPIC) is binary; **Logistic regression** is the appropriate statistical method to conduct. Our model will be able to explain the relationship between one dependent binary variable and other independent numeric and categorical variables. Hence, we are searching for a model with outcome (response) a binary variable (0=No Myopic, 1=Myopic) that will have the form:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 \dots + b_kX_k\ , with\ k \leq 15\ and\ \ logit(p) = \ln(\frac{p}{1-p})$$

In R we used GLM package to create our model. We created our initial model (see Table 4 - Coefficients in Logistic Regression Model). In this model we observe underdispersion, that means that in our data the variability would be significant smaller than we expected in our model (mathematically, Variance of residuals/ Degrees of Freedom<1). The reason for underdispersion phenomenon could be adjusted to subgroups that are correlated with each other, also known as autocorrelation. Also, we observe that in our initial model we have a lot of variables that are not significant, that means that the Pr is greater than 0.05 (Null Hypothesis: zero correlation between covariates and response). So, we run the **step** function in R (Stepwise Regression method with AIC and BIC criterion). The AIC or BIC for a model is usually written in the form (-2logL + kp), where *L* is the likelihood function, *p* is the number of parameters in the model, and *k* is 2 for AIC and log(*n*) for BIC. So both methods suggest models with balance of fit and parsimony. Now the remaining variables after stepwise are only the following variables:

(AIC Criterion)	GENDER+SPHEQ+ACD+SPORTHR+PARENTMY
(BIC Criterion)	SPHEQ+ SPORTHR+PARENTMY

Lasso Model

At that stage we create another model by using **Lasso Regression** (least absolute shrinkage and selection operator). Lasso performs model selection. In Lasso we have to tune the parameter lambda (λ) that controls the amount of regularization. As the λ increases, Lasso sets more coefficients to zero (see Figure 7 - Lasso Coefficient Shrinkage). We choose the largest value of λ, in order to limit the error within 1 standard error of the minimum. In R we create a matrix (without the intercept) and by using the **glmnet** library we create the relation between λ and Lasso coefficients. After Lasso Method only 3 variables selected as significant, all the other have shrink to zero(see Table 7 - Lasso Variables Selection,). Now we have as significant only the following variables:

SPHEQ+ PARENTMY

If we create our model with only the Lasso coefficients we create a model with AIC = 328.53 but it also suffers from underdispersion. It's a model more simple than the model from stepwise method but according to AIC criterion we have a worse model. Lasso method creates a model according to regularization technique and it's better for prediction, so in this case study, this kind of information is useless.

Final Model & Interpretation

We have already observed the correlations between the variables and how they affect the logistic regression model. Our goal is to find which of all the variables influence the creation of myopia. In this final stage we have to compare all the created models. Our selection criteria for variable selection for this analysis is based on our initial descriptive analysis, AIC criterion (lower AIC value is better), interpretation difficulty level , low VIF values <10 and also we take into consideration that the Residuals Deviance / Degrees of Freedom ratio (Res. Dev./ Df) should be as much closer to one. In our analysis we investigated if there is 2nd term interactions, but that model is very complicated, higher AIC (≈ 332) and difficult interpretation. In the following table we gathered the most important models from our analysis:

Model Variables	Comments	AIC	Res. Dev./Df
GENDER+SPHEQ+ACD+LT+VCD+SPORTHR+DIOPTERHR+PARENTMY STUDYYEAR+READHR+COMPHR+TVHR+AGE+AL+STUYHR	FULL MODEL	332.1	0.49
GENDER+SPHEQ+ACD+SPORTHR+PARENTMY	STEPWISE (AIC)	324.7	0.51
SPHEQ +SPORTHR+PARENTMY	STEPWISE (BIC)	328.5	0.52
SPHEQ+ PARENTMY	FROM LASSO	333.6	0.53

We started analysis with the Full model. This model is too complicated, has 8 variables that are not all significant. For descriptive models we prefer model with significant variables. After we ameliorate the AIC criterion by running the stepwise method with AIC & BIC; with BIC we have higher penalty and as a result a more parsimony model. Finally, we implement Lasso variable selection; Lasso method gives a higher penalty than the other methods, because it tries to regularize the model (Lasso model has better prediction skill than a model from stepwise). In all models Spherical Equivalent Refraction (SPHEQ) and at least one myopic parent (PARENTMY) are significant variables in all selected models, so it make sense that these variables can not be excluded from the model. We prefer the model from stepwise which is the model with the lowest AIC value. Also, I would like to mention, that the Residuals Deviance / Degrees of Freedom ratio is approximately the same in all models. In realistic studies, it's very difficult to have this ratio equals to 1. So, the model is:

$$\boxed{\text{Logit}(p)= -5.90 + 0.73\times\text{FEMALE} + -3.83\times\text{SPHEQ} + 1.37\times\text{ACD}-0.05\times\text{SPORTHR}+1.39\times(\text{ONE PARENT IS MYOPIC})}$$

All p-values are significant (Pr<0.05). Here we examine all the above variables[3]:

- -5.90 is the intercept, this value doesn't have any physical meaning, but we keep this constant in our model because it helps for better fitting
- If the **gender is female (FEMALE)** then the **probability increases** 107% (In this model the reference level is male)
- If the **spherical equivalent refraction (SPHEQ)** decreases 0.05 D then **probability increases 21%** (the more negative the spherical equivalent, the more myopic the subject)
- If the **anterior chamber depth (ACD)** increases 0.1 mm then the **probability increases 10%**
- If the subject increases **the engaging time spend (SPORTHR)** by 1hour per week, then the **probability decreases 5%**
- If one of both parents is myopic(**ONE PARENT MYOPIC**) then the children will have **300% higher probability** to be myopic

[3] For variables interpretation we change only one variable at the time and we keep all the other constant. All the coefficients are the difference in log odds. So, we recall that logarithm converts multiplication and division to addition and subtraction. Its inverse, the exponentiation converts addition and subtraction back to multiplication and division.

Conclusion

We examined various physiological, environmental and hereditary variables using logistic regression. The most important variable is the **spherical equivalent refraction.** People with low values in spherical equivalent refraction in young ages have extremely higher changes to suffer from myopia in the following years. Additionally, another physiological variable that looks to affect myopia, is the **anterior chamber depth**. About hereditary variables, a high important parameter is the myopia of at **least one of the parents**. Also there is a **difference between genders**; females have much higher changes to have myopia than males. On the other hand, spending time in **sport/outdoor activities,** it looks to prevent from myopia in a small extent but maybe, if we had a longer period study, we could have a more clear view in this effect.

Appendix I

Table 2 - Gender Myopia Table

		Percentage (%)
Myopic	Man	52
	Woman	48
Healthy	Man	43
	Woman	57

Table 3 - Parents Myopia Table

		Percentage (%)
Myopic	Myopic children	73
Parents	Healthy children	27
Healthy	Myopic children	94
Parents	Healthy children	6

Table 4 - Coefficients in Logistic Regression Model

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -219.59299  218.50651  -1.005   0.3149
STUDYYEAR      0.11146    0.10938   1.019   0.3082
AGE            0.03729    0.25311   0.147   0.8829
GENDER1        0.65312    0.34353   1.901   0.057
SPHEQ         -4.13571    0.47125  -8.776   <2e-16
AL           -31.66998   38.92884  -0.814   0.4159
ACD           32.95501   39.00219   0.845   0.3981
LT            30.85557   39.03947   0.790   0.4293
VCD           31.33732   38.95691   0.804   0.4212
SPORTHR       -0.04623    0.02121  -2.179   0.0293
READHR         0.07068    0.04951   1.428   0.1534
COMPHR         0.04604    0.04561   1.010   0.3127
STUDYHR       -0.18133    0.09945  -1.823   0.0683
TVHR          -0.01324    0.02869  -0.461   0.6445
DIOPTERHR           NA         NA      NA       NA
PARENTMY1      1.20824    0.51800   2.333   0.0197

Residual deviance: 302.10  on 603  degrees of freedom

AIC: 332.1
```

Table 5 - Coefficients after Stepwise Method (AIC Criterion)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.90146    2.60855  -2.262  0.02368 *
GENDER1      0.73277    0.30607   2.394  0.01666 *
SPHEQ       -3.82779    0.43598  -8.780  < 2e-16 ***
ACD          1.36918    0.69187   1.979  0.04782 *
SPORTHR     -0.05362    0.02027  -2.645  0.00817 **
PARENTMY1    1.39899    0.51808   2.700  0.00693 **

Residual deviance: 312.70  on 612  degrees of freedom

AIC: 324.7
```

Table 6 - Coefficients after Stepwise Method (BIC Criterion)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.54529    0.56262  -0.969  0.33244
SPORTHR     -0.05045    0.01973  -2.557  0.01056 *
PARENTMY1    1.34430    0.51039   2.634  0.00844 **
SPHEQ       -3.83186    0.43398  -8.830  < 2e-16 ***

Residual deviance: 320.53  on 614  degrees of freedom

AIC: 328.53
```

Table 7 - Lasso Variables Selection

```
(Intercept)                          -0.664
GENDER1                               .
SPHEQ                                -2.619
ACD                                   .
LT                                    .
VCD                                   .
SPORTHR                               .
DIOPTERHR                             .
PARENTMY1                             0.063
```

Table 8 - Coefficients after Lasso Method

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.54529    0.56262  -0.969  0.02295
SPHEQ       -3.83186    0.43398  -8.830  < 2e-16
PARENTMY1    1.34430    0.51039   2.634  0.00751

Residual deviance: 320.53  on 614  degrees of freedom
```

```
AIC: 333.95
```

Appendix II

Figures

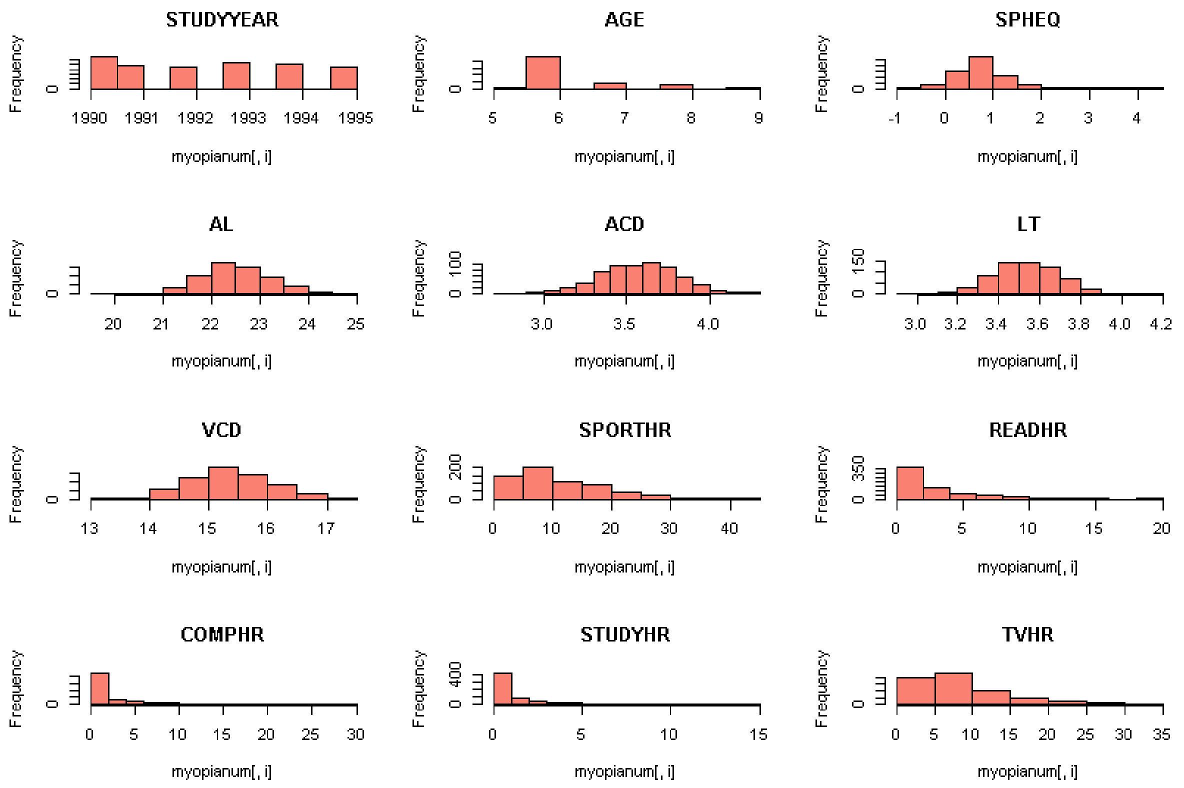

Figure 6- Numerical Variables Distributions

In the above graphs we have created a bar graph for each numerical variable. STUDYYEAR and AGE take discrete values and all the others take continues values. The distribution of STUDYYEAR is approximately the same in the period of the 5 years study. In AGE graph we observe that the majority of subjects are 6 year children, this information will be useful in our final conclusions. AL, ACD, LT, VCD are linear related AL=ACD+LT+VCD and they have approximately normal-symmetric distributions. On the other hand SPORTHR, READHR, COMPHR, STUDYHR and TVHR are distributions with very high positive skewness.

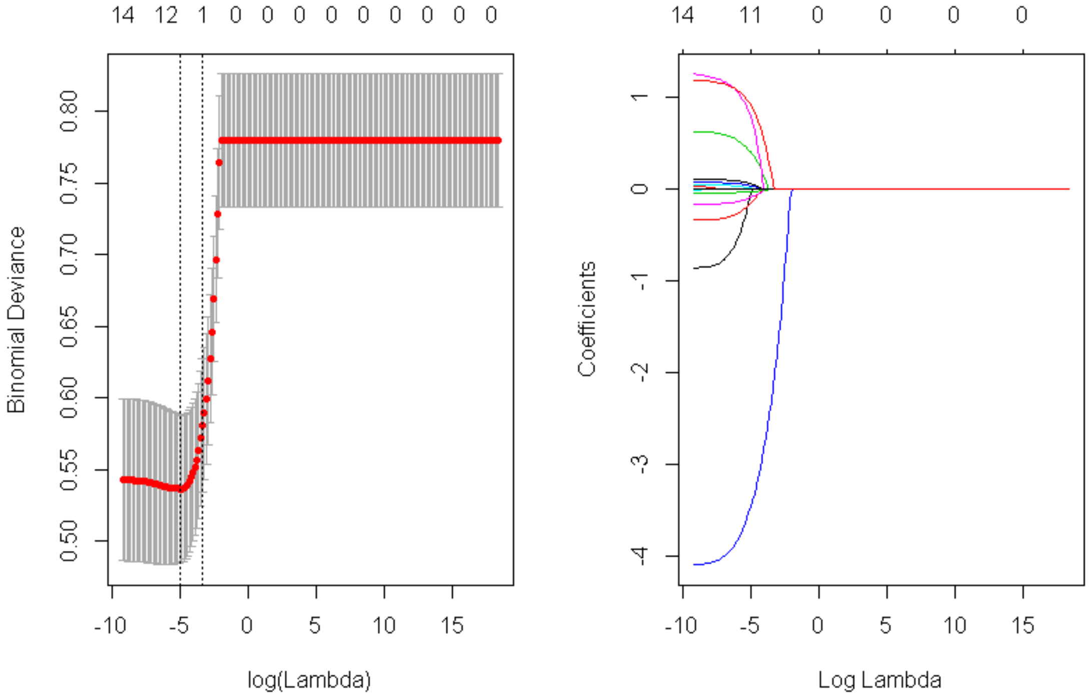

Figure 7 - Lasso Coefficient Shrinkage