MYOPIA STUDY

STATISTICS II FOR BUSINESS ANALYTICS

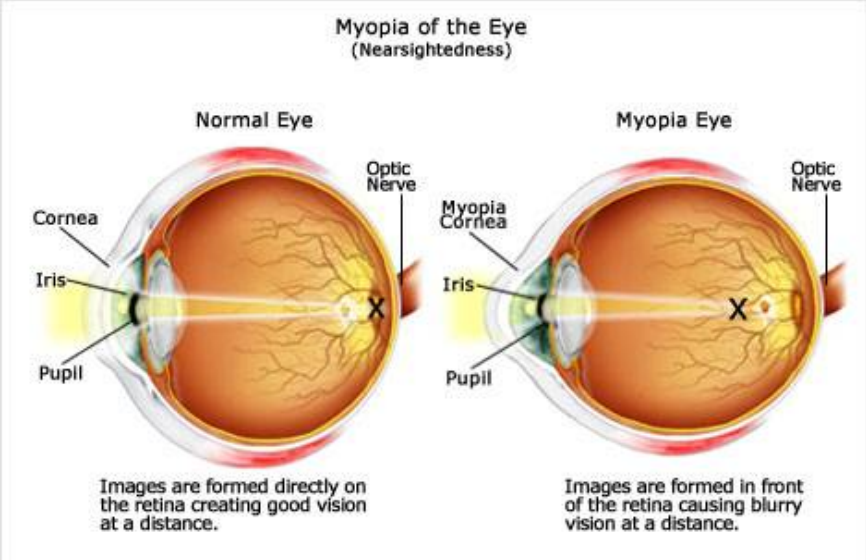
1. INTRODUCTION

1.1. DESCRIPTION OF THE PROBLEM AND STUDY AIM

Myopia (also called nearsightedness) is the most common cause of impaired vision in people and occurs when the eyeball is too long, relative to the focusing power of the cornea and lens of the eye. This causes light rays to focus at a point in front of the retina, rather than directly on its surface.

The underlying cause is believed to be a combination of genetic and environmental factors. Risk factors include doing work that involves focusing on close objects, greater time spent indoors, and a family history of the condition. It is also associated with a high socioeconomic class. The underlying mechanism involves the length of the eyeball growing too long or less commonly the lens being too strong. It is a type of refractive error.

The purpose of the current study is to examine by using general linear models (GLM) which are the influential variables that contribute to the development of myopia.



1.2. DATA CHARACTERISTICS

The dataset used is 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.

This dataset refers to 618 cases of people who had at least five years of follow-up and were not myopic. A person is considered myopic when spherical equivalent refraction is equal or less than -0.75 D. 17 variables are included in the dataset, the detailed description of which can be found in the below Table 1.

Variable description	Variable name	Value

Myopia within the first five years of follow up	MYOPIC	0 = No, 1 = Yes
Year subject entered the study	STUDYYEAR	Year
Age at first visit	AGE	Years
Gender	GENDER	0 = Male, 1 = Female
Spherical equivalent refraction	SPHEQ	Diopter
Axial length	AL	mm
Anterior chamber depth	ACD	mm
Lens thickness	LT	mm
Vitreous chamber depth	VCD	mm
Time spent engaging in sports/outdoor activities	SPORTHR	Hours per week
Time spent reading for pleasure	READHR	Hours per week
Time spent playing video games/working on the pc	COMPHR	Hours per week
Time spent reading/studying for school assignments	STUDYHR	Hours per week
Time spent watching television	TVHR	Hours per week
Composite of near-work activities	DIOPTERHR	Hours per week
Myopic mother	MOMMY	0 = No, 1 = Yes
Myopic father	DADMY	0 = No, 1 = Yes

Table 1 – Dataset Variables

2. DESCRIPTIVE ANALYSIS

All the data used were imported in Rstudio. A number of data preprocessing steps was initially followed to clear and transform the data so as to proceed to the required analysis of the current section. Note that for convenience purposes the MOMMY and DADMY variables initially included in the dataset were combined in one new variable named MYOPIC_PARENTS that indicates if any myopia exists for each of the person's parents (No myopic parents = 0, Myopic mother = 1, Myopic father=2, Both myopic parents = 3). After this process is completed, 16 variables in total remain in the data set, 13 numeric and 3 factors.

Once further analyzing the data, it is observed that as far as the axial length variable (AL) is concerned, there is a linear relationship with the variables of anterior chamber depth (ACD), lens thickness (LT) and vitreous chamber depth (VCD) since

$$AL = ACD + VCD + LT$$

A similar linear relationship also exists between the composite of near-work activities variable (DIOPTERHR) and the rest 4 environmental variables: the time spend for reading pleasure (READHR), for playing video/computer games or working on the computer (COMPHR), for reading or studying (STUDYHR) and watching television (TVHR) as

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

This kind of detected relation should be taken into account as these variables cannot be included simultaneously in the selected final model. If they do, then multicollinearity problem will exist misleading us to the appropriate interpretation of the results. When performing a pairwise comparison for the numerical variables (see Diagram 1 - Numeric variables correlations), it is indeed verified the statement mentioned above that there is a high correlation between AL - VCD (94%) and DIOPTERHR - STUDYHR (62%) while no particular relations exist among the other variables.

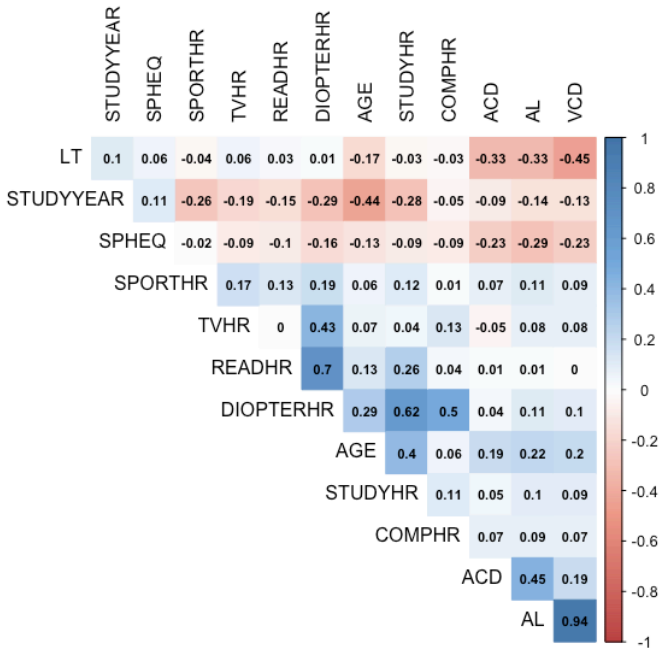


Diagram 1 - Numeric Variables Correlations

After proceeding to the visualization of the numeric variables (See *Diagram 2 - Numeric Variables Distribution*), it is observed that:

- The AL and the VCD variables tend to approximately follow normal distribution.
- None of the rest variables indicates such high symmetry in its distribution, though. Worth to be mentioned is that the heavily left-skewed variables are these of READHR, COMPHR, STUDYHR, SPORTHR, TVHR, DIOPTERHR and AGE.

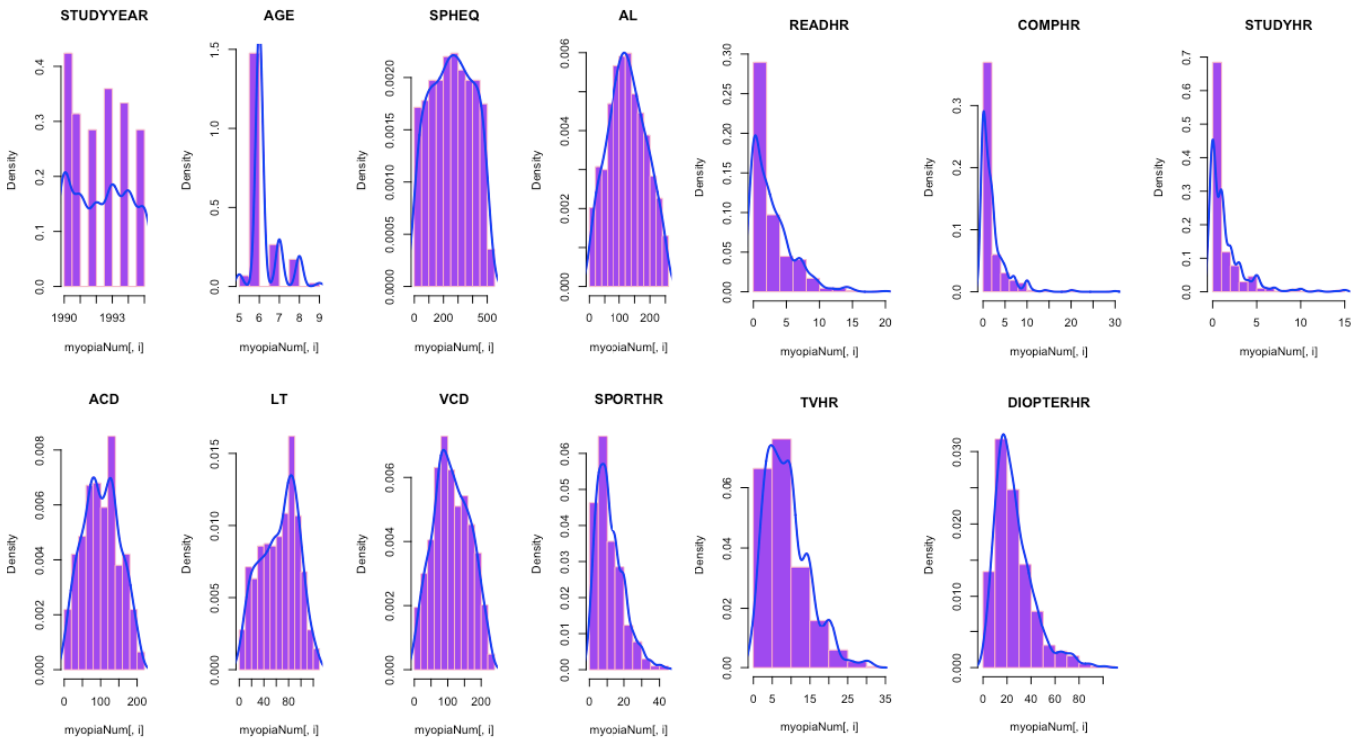


Diagram 2 - Numeric Variables Distribution

In order to further investigate the relation between the person's myopia and the categorical variables of parents' myopia and gender, the relevant bar plots were created (See Diagram 3 – Parents' Myopia Distribution and Diagram 4 – Myopia Per Gender respectively). To be highlighted that the percentage of the myopic people who have at least one myopic parent (mother or father) is almost the same (≈ 25%) while myopic men and women present almost the same percentages (≈ 80%), an evidence which proves that myopia doesn't differ significantly between genders (X-squared test: Pr > 0.05).

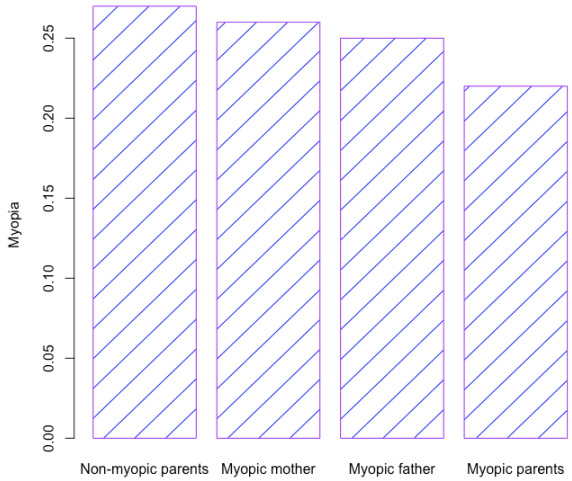


Diagram 3 – Parents' Myopia Distribution

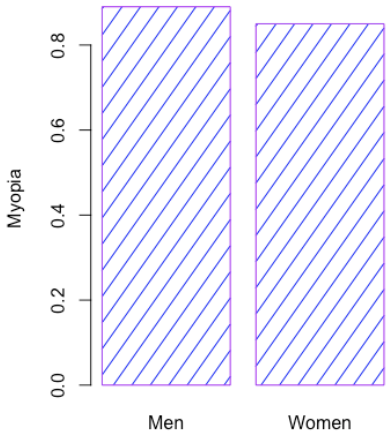


Diagram 4 – Myopia Per Gender

3. MODEL SELECTION

3.1. LOGISTIC MODEL

The dependent variable of this study is a binary variable (MYOPIC - 0: No myopic,1: Myopic) and the independent variables are all the other numeric and categorical variables already mentioned above. Hence, logistic regression is the suitable statistical method to conduct and as such the model which is searched with outcome (response) a binary variable will be of the form:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m \text{ with } m \leq 16 \text{ and } \text{logit}(p) = \ln(p/(1-p))$$

The initial model is created (See *Table 2 – Initial Logistic Model*) by using the GLM package of R including the variables of Table 1 (except MOMMY and DADMY that were combined in the new variable of MYOPIC_PARENTS). The following issues can be raised here:

- Underdispersion is observed for this model meaning that data exhibit less variation than we would expect based on this binomial distribution (for defectives) since the ratio $\frac{Residual\ deviance}{Degrees\ of\ freedom} = 0.5175 < 1$.
- A possible reason for this underdispersion problem could be the existence of subgroups that are correlated to one another, the so-called phenomenon of autocorrelation. After examining the potential correlations (that are already pointed out in the third section of Descriptive Analysis) the following relations are verified:
    - Between the composite of near-work activities (DIOPTERHR) and time spent reading for pleasure (READHR), playing video games/working on the pc (COMPHR), reading/studying for school assignments (STUDYHR) and watching television (TVHR).
    - Between the axial length (AL) and anterior chamber depth (ACD), vitreous chamber depth (VCD) and lens thickness (LT).
- A great number of variables are included which are not statistically significant (since the null hypothesis is H_0: $b_1 = b_2 = \ldots = b_m = 0$ and $P_r > 0.05$)

Therefore, the stepwise method is subsequently applied for the variable selection in the logistic model (AIC and BIC criterion) after having removed the axial length (AL) and composite of near-work activities (DIOPTERHR) variables because of which the multicollinearity problem existed. After stepwise methods are run, the remaining variables are the following ones per criterion (See *Table 3 – Logistic Model (AIC)* and *Table 4 – Logistic Model (BIC*) for more details*)*:

Model type	Variables selected
Logistic model (AIC)	Gender + Spherical equivalent refraction + Anterior chamber depth + Time spent engaging in sports/outdoor activities + Time spent reading for pleasure + Time spent reading/studying for school assignments + Parents' myopia
Logistic model (BIC)	Spherical equivalent refraction + Time spent engaging in sports/outdoor activities + Parents' myopia

3.2. LASSO MODEL

Another model is created by using the lasso regression method via which a relevant variable selection is also performed. The tuning parameter lambda (λ) used in this method needs to be firstly defined so as to limit the intense of regularization in our model, i.e. how many

variables will be included. Since as λ increases, more coefficients are set to zero (See *Diagram 5 – Lasso Coefficient Shrinkage*) the λ value selected is this which limits the error within 1 standard error of the minimum.

After Lasso method is applied, only the following 4 variables are selected as significant as all the others are shrunk to zero (See *Table 5 – Lasso Model*):

Model type	Variables selected
Lasso model	Gender + Spherical equivalent refraction + Time spent engaging in sports/outdoor activities + Parents' myopia

This model is much simpler in terms of number of coefficients but to be noticed that in case such a model is created including the above Lasso coefficients, then this still suffers from underdispersion with the AIC to be 337.21 indicating not the best model compared to those of the stepwise method. In fact, this is a model better for predictions, although this is out of scope and thus useless for the current study.

4. FINAL MODEL & INTERPRETATION

The goal of this section is to compare all the models described in sections 4.1 and 4.2. so as to define which are finally the variables that influence the creation of myopia and to what degree. The factors that may be taken into consideration as selection criteria are the findings of the descriptive analysis, the interpretation difficulty level of the models, the AIC criterion (i.e. lower AIC values are considered better), the multicollinearity degree (i.e. VIF values <10 are considered better) and last but not least the dispersion ratio (i.e. the closer to 1 is $\frac{Residual\ deviance}{Degrees\ of\ freedom}$ the better the model is). In the following table the models and their crucial information are summarized:

Model type	Variables selected	AIC	$\frac{Residual\ deviance}{Degrees\ of\ freedom}$	Collinearity
Full model	Year subject entered the study + Age at first visit + Gender + Spherical equivalent refraction + Anterior chamber depth + Lens thickness + Vitreous chamber depth + Time spent engaging in sports/outdoor activities + Time spent playing video games/working on the pc + Time spent playing video games/working on the pc + Time spent reading/studying for school assignments + Time spent watching television + Parents' myopia [Number of variables: 12]	343.15	0.5168	X
Logistic model (AIC)	Gender + Spherical equivalent refraction + Anterior chamber depth +	335.33	0.5186	X

	Time spent engaging in sports/outdoor activities + Time spent reading/studying for school assignments + Parents' myopia [Number of variables: 6]			
Logistic model (BIC)	Spherical equivalent refraction + Time spent engaging in sports/outdoor activities + Parents' myopia [Number of variables: 3]	339.10	0.5344	X
Lasso model	Gender + Spherical equivalent refraction + Time spent engaging in sports/outdoor activities + Parents' myopia [Number of variables: 4]	337.21	0.5289	X

The analysis initiated with the full model which is too complicated as it includes 12 variables the majority of which are not statistically important. After this, the full model is improved by running the stepwise methods with AIC and BIC; BIC and Lasso set a higher penalty, and this is why much fewer variables are included in the selected model.

What it is important to be highlighted here is that spherical equivalent refraction (SPHEQ) and parents' myopia (MYOPIC_PARENTS) exist in all models indicating that they consist significant variables. On top of that, it is observed that the dispersion ratio $\frac{Residual\ deviance}{Degrees\ of\ freedom}$ is approximately the same in all models and less than 1, something which signals that underdispesion characterizes all models.

Therefore, taken into account the above points, the AIC logistic model is the most preferable model since this presents the lowest AIC value and has not any crucial difference in comparison to the others in terms of dispersion. The p-values of anterior chamber depth (ACD) and Time spent reading/studying for school assignments (STUDYHR) variables are not significant ($P_r > 0.05$) so these variables will not be included in the final selected model which have the following format:

logit(p) = -1.11 + 0.62 * GENDER1 - 0.01 * SPHEQ - 0.05 * SPORTHR + 0.07 * READHR + 1.02 * MYOPIC_PARENTS1 + 1.24 * MYOPIC_PARENTS2 + 1.69 * MYOPIC_PARENTS3

When interpreting this model, it is concluded that:

- The log odds of developing myopia is -0.49 (-1.11 + 0.62) when the object's gender is female and all the other variables are equal to zero.
- 1 diopter increase in spherical equivalent refraction (SPHEQ) decreases the log odds of developing myopia by 0.01, assuming all the other variables are constant.
- 1 hour increase in time spent engaging in sports/outdoor activities (SPORTHR) decreases the log odds of developing myopia by 0.05, assuming all the other variables are constant.

- 1 hour increase in time spent reading/studying for school assignments (READHR) decreases the log odds of developing myopia by 0.07, assuming all the other variables are constant.
- The log odds of developing myopia is -0.09 (-1.11 + 1.02) when the object has myopic mother and all the other variables are equal to zero.
- The log odds of developing myopia is 0.13 (-1.11 + 1.24) when the object has myopic father and all the other variables are equal to zero.
- The log odds of developing myopia is 0.58 (-1.11 + 1.69) when the object both myopic parents and all the other variables are equal to zero.

5. CONCLUSIONS

A great number of environmental, hereditary and physiological variables were examined to conduct a study about the important factors that contribute to the development of myopia.

As analyzed, the most determinant factor is the myopia that is developed in both parents; when it exists, the probability of inheriting the disease to the object-child is the biggest possible while secondarily the time spending by the object in outdoor/sport activities is something that prevents to the maximum the myopia development. Last but not least is that there is difference between genders; females have higher chances of having myopia than these of males.

6. APPENDIX I (LIST OF TABLES)

Table 2 – Initial Logistic Model

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.461e+02  2.156e+02  -1.142  0.25353
STUDYYEAR        1.227e-01  1.078e-01   1.138  0.25493
AGE              2.347e-01  2.403e-01   0.977  0.32881
GENDER1          5.964e-01  3.342e-01   1.785  0.07434 .
SPHEQ           -1.332e-02  1.604e-03  -8.302  < 2e-16 ***
ACD              4.566e-03  3.404e-03   1.341  0.17985
LT              -5.216e-03  5.848e-03  -0.892  0.37240
VCD             -4.513e-03  3.288e-03  -1.373  0.16988
SPORTHR         -4.672e-02  2.070e-02  -2.257  0.02400 *
READHR           7.470e-02  4.835e-02   1.545  0.12234
COMPHR           3.687e-02  4.587e-02   0.804  0.42151
STUDYHR         -1.595e-01  9.321e-02  -1.711  0.08709 .
TVHR            -3.434e-03  2.790e-02  -0.123  0.90204
MYOPIC_PARENTS1  9.607e-01  5.618e-01   1.710  0.08728 .
MYOPIC_PARENTS2  1.118e+00  5.649e-01   1.980  0.04773 *
MYOPIC_PARENTS3  1.650e+00  5.455e-01   3.024  0.00249 **

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: 311.15 on 602 degrees of freedom
AIC: 343.15

Number of Fisher Scoring iterations: 7
```

Table 3 – Logistic Model (AIC)

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.109642   0.716819  -1.548  0.12162
GENDER1          0.621458   0.303538   2.047  0.04062 *
SPHEQ           -0.012890   0.001564  -8.243  < 2e-16 ***
ACD              0.005442   0.003127   1.741  0.08176 .
SPORTHR         -0.050834   0.019757  -2.573  0.01008 *
READHR           0.066584   0.046242   1.440  0.14989
STUDYHR         -0.136951   0.081755  -1.675  0.09391 .
MYOPIC_PARENTS1  1.024641   0.555454   1.845  0.06508 .
MYOPIC_PARENTS2  1.242296   0.554305   2.241  0.02501 *
MYOPIC_PARENTS3  1.688349   0.536015   3.150  0.00163 **

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: 315.33  on 608  degrees of freedom
AIC: 335.33

Number of Fisher Scoring iterations: 7
```

Table 4 – Logistic Model (BIC)

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.257071   0.561766  -0.458  0.64723
SPHEQ           -0.012741   0.001521  -8.378  < 2e-16 ***
SPORTHR         -0.046409   0.018947  -2.449  0.01431 *
MYOPIC_PARENTS1  1.043085   0.549902   1.897  0.05785 .
MYOPIC_PARENTS2  1.221106   0.547887   2.229  0.02583 *
MYOPIC_PARENTS3  1.713236   0.530411   3.230  0.00124 **

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: 327.10 on 612 degrees of freedom
AIC: 339.1

Number of Fisher Scoring iterations: 6
```

Table 5 – Lasso Model

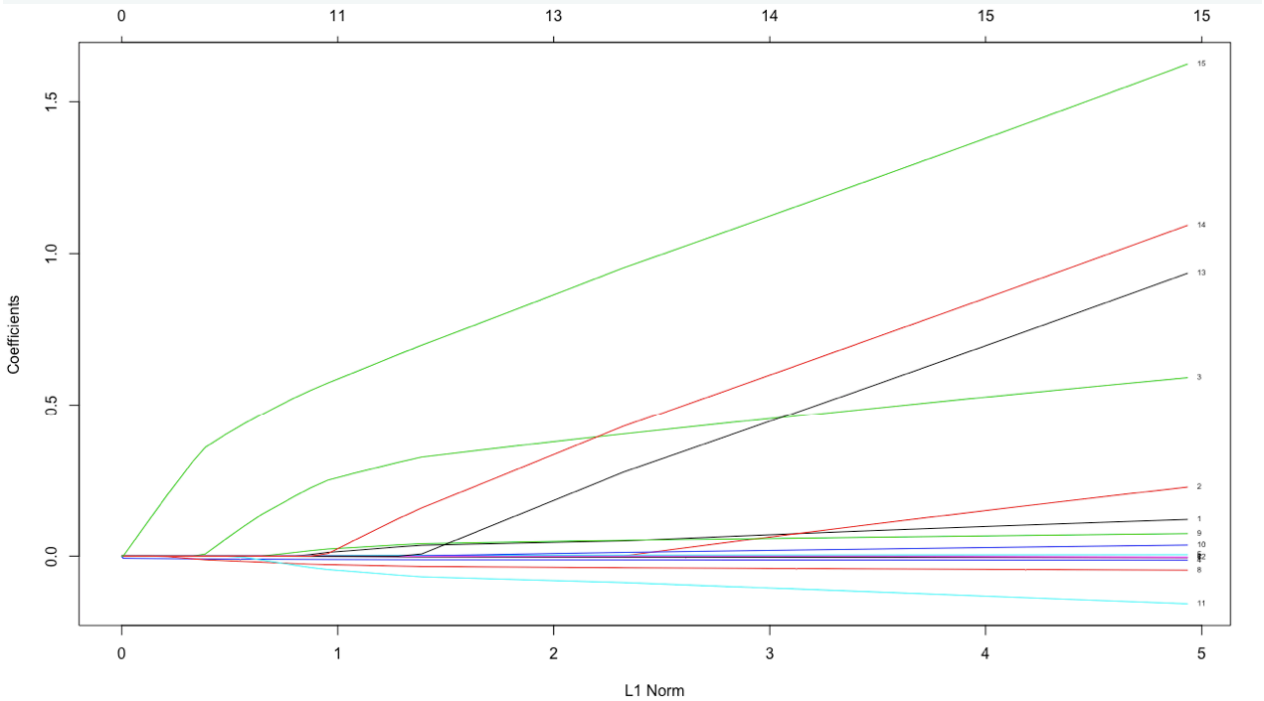
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.563010   0.587202  -0.959  0.33766
GENDER1          0.558858   0.285098   1.960  0.04997 *
SPHEQ           -0.012800   0.001518  -8.431  < 2e-16 ***
SPORTHR         -0.045623   0.019067  -2.393  0.01672 *
MYOPIC_PARENTS1  1.068180   0.551814   1.936  0.05290 .
MYOPIC_PARENTS2  1.247536   0.549918   2.269  0.02329 *
MYOPIC_PARENTS3  1.742197   0.531637   3.277  0.00105 **
```

```
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: 323.21 on 611 degrees of freedom
AIC: 337.21

Number of Fisher Scoring iterations: 6
```

7. APPENDIX II (ADDITIONAL DIAGRAMS)



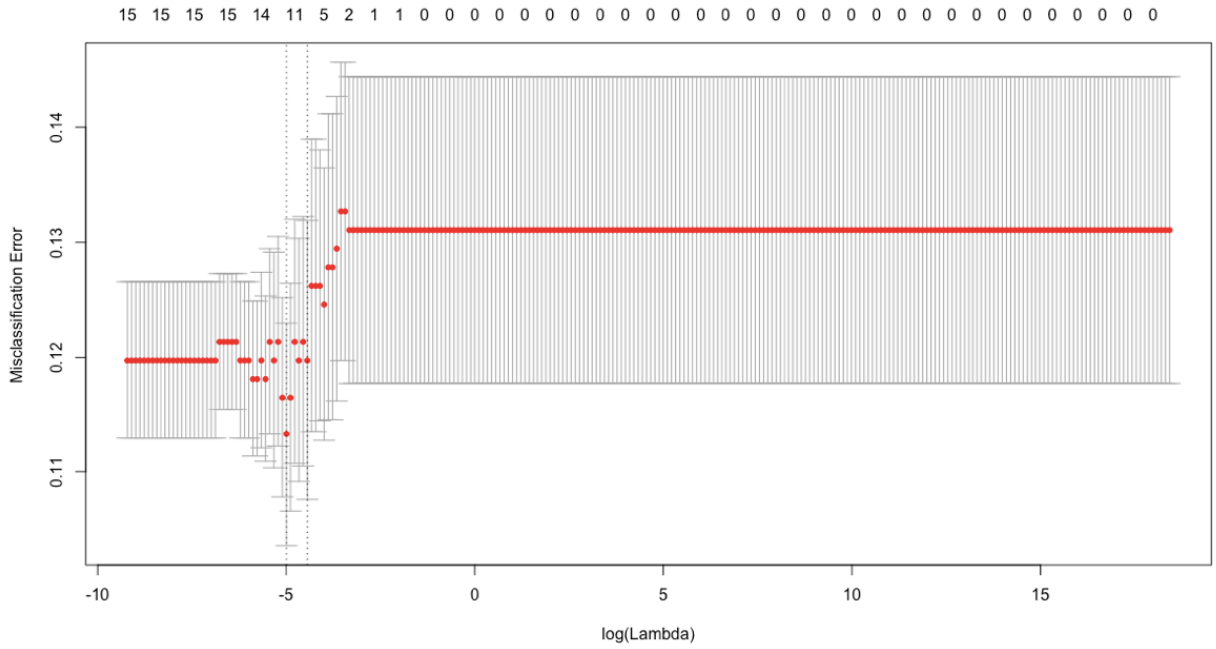


Diagram 5 – Lasso Coefficient Shrinkage