BUSINESS
ANALYTICS
Master of Science

Department of Management Science
and Technology (AUEB)

Project I
Myopia Study

Statistics for business analytics II

Eva Giannatou

February 2017

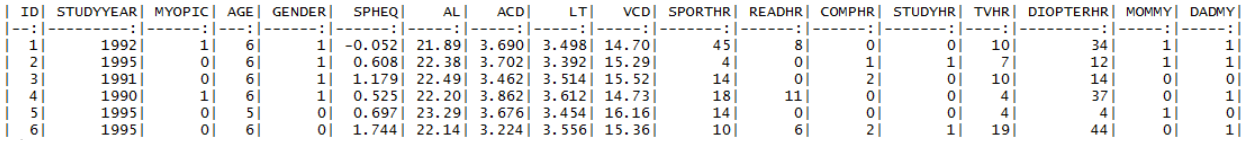
Contents

Figures

1. Introduction

The dataset 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. 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 the dataset 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.



```
| ID| STUDYYEAR| MYOPIC| AGE| GENDER|  SPHEQ|    AL|   ACD|    LT|   VCD| SPORTHR| READHR| COMPHR| STUDYHR| TVHR| DIOPTERHR| MOMMY| DADMY|
|--:|---------:|------:|---:|------:|------:|-----:|-----:|-----:|-----:|-------:|------:|------:|-------:|----:|---------:|-----:|-----:|
|  1|      1992|      1|   6|      1| -0.052| 21.89| 3.690| 3.498| 14.70|      45|      8|      0|       0|   10|        34|     1|     1|
|  2|      1995|      0|   6|      1|  0.608| 22.38| 3.702| 3.392| 15.29|       4|      0|      1|       1|    7|        12|     1|     1|
|  3|      1991|      0|   6|      1|  1.179| 22.49| 3.462| 3.514| 15.52|      14|      0|      2|       0|   10|        14|     0|     0|
|  4|      1990|      1|   6|      1|  0.525| 22.20| 3.862| 3.612| 14.73|      18|     11|      0|       0|    4|        37|     0|     1|
|  5|      1995|      0|   5|      0|  0.697| 23.29| 3.676| 3.454| 16.16|      14|      0|      0|       0|    4|         4|     1|     0|
|  6|      1995|      0|   6|      0|  1.744| 22.14| 3.224| 3.556| 15.36|      10|      6|      2|       1|   19|        44|     0|     1|
```

Figure 1 Dataset preview

Figure 1 represents the 6 first lines of the dataset. The table represented below describes the dataset's variable names, descriptions and their values.

Variable Name	Variable Description	Values
ID	Subject identifier	1-618
STUDYYEAR	Year subject entered the study	year
MYOPIC	Myopia within the first five years of follow up. MYOPIC is defined as SPHEQ <= −0.75 D.	0 = No 1 = Yes
AGE	Age at first visit	years
GENDER	Gender	0 = Male 1= Female
SPHEQ	Spherical Equivalent Refraction. : A measure of the eye's effective focusing power.	diopter
AL	Axial Length. The length of eye from front to back.	mm.
ACD	Anterior Chamber Depth. The length from front to back of the aqueous-containing space of the eye between the cornea and the iris.	mm.
LT	Lens Thickness. The length from front to back of the crystalline lens.	mm.
VCD	Vitreous Chamber Depth. The length from front to back of the aqueous-containing space of the eye in front of the retina.	mm.
SPORTHR	How many hours per week outside of school the child spent engaging in sports/outdoor activities	Hours per week.
READHR	How many hours per week outside of school the child spent reading for pleasure	Hours per week.
COMPHR	How many hours per week outside of school the child spent playing video/computer games or working on the computer	Hours per week.
STUDYHR	How many hours per week outside of school the child spent reading or studying for school assignments	Hours per week.

TVHR	How many hours per week outside of school the child spent watching television	Hours per week.
DIOPTERHR	Composite of near-work activities defined as DIOPTERHR = 3× (READHR + STUDYHR) + 2 × COMPHR + TVHR	Hours per week.
MOMMY	Was the subject's mother myopic?	0 = No 1 = Yes
DADMY	Was the subject's father myopic?	0 = No 1 = Yes

MYOPIC is defined as SPHEQ <= −0.75 diopter. Therefore "SPHEQ" is highly associated to myopia but is it enough to predict the existence of myopia in children? The relationship between "SPHEQ" and "MYOPIC" is clear when examining the following plot. From the plot we can see that low "SPHEQ" values are associated with the existence of myopia. However, there are more variables influencing the existence of myopia in children. In the next steps, I will evaluate the dataset's variables and I will choose the ones which are statistically significant and therefore can be used as input in the model.

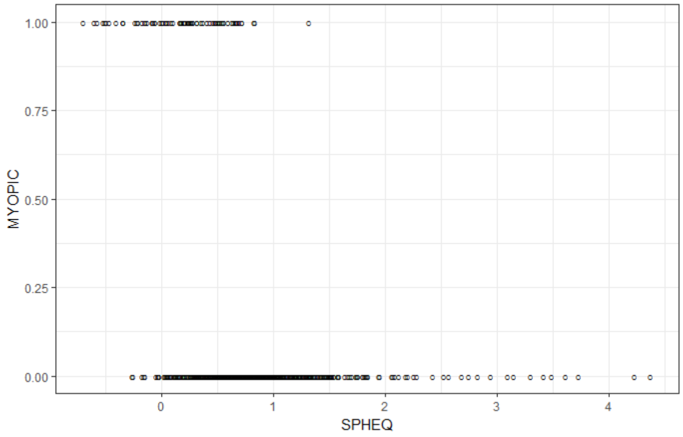


Figure 2 Use SPHEQ to predict MYOPIA.

Ultimately in figure 2 we should not see 2 points on the MYOPIC axis for the same SPHEQ value. In this case it is obvious that "SPHEQ" influences the existence of myopia but it is not enough to accurately predict it. We will need to add more attributes to the model in order to improve the prediction. To do so we need to examine the correlation between each attribute and the existence of myopia.

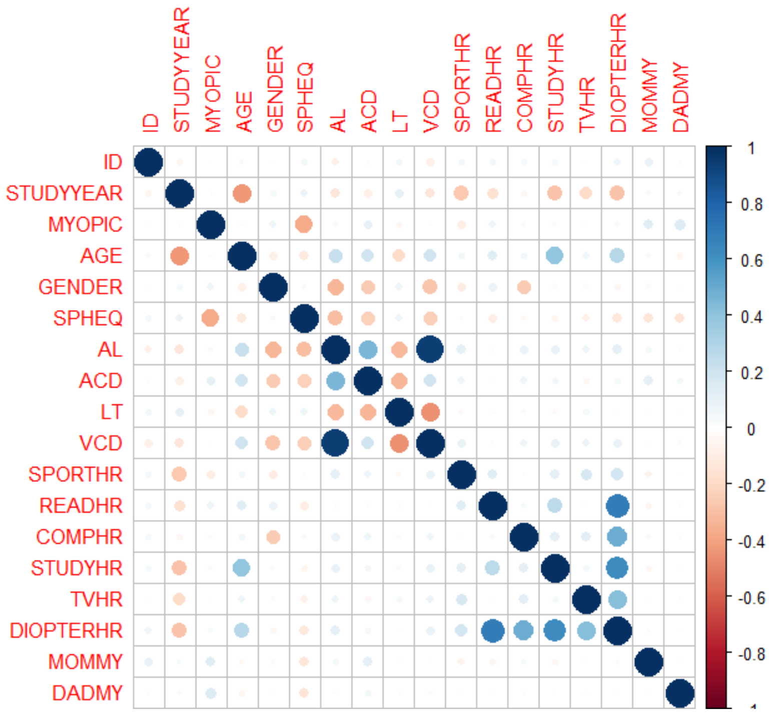
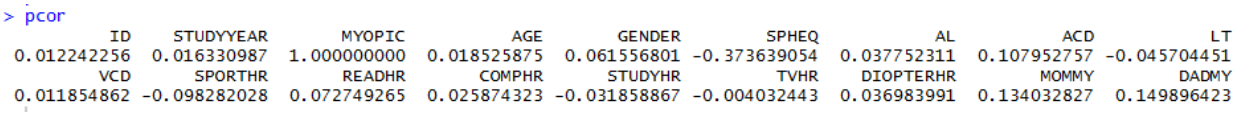


Figure 3 Correlation between variables

Figure 3 represents the correlations between all variables. High positive correlation is represented with dark blue color while high negative correlation with red. Lightly colored cells represent low correlation and white cells no correlation at all.

It is obvious for example that "DIOPTERHR" is highly correlated to "SPORTHR", "TVHR", "STUDYHR", "COMPHR" and "READHR". Therefore, the "DIOPTERHR" variable will not be included in the prediction model in order to prevent collinearity problems.

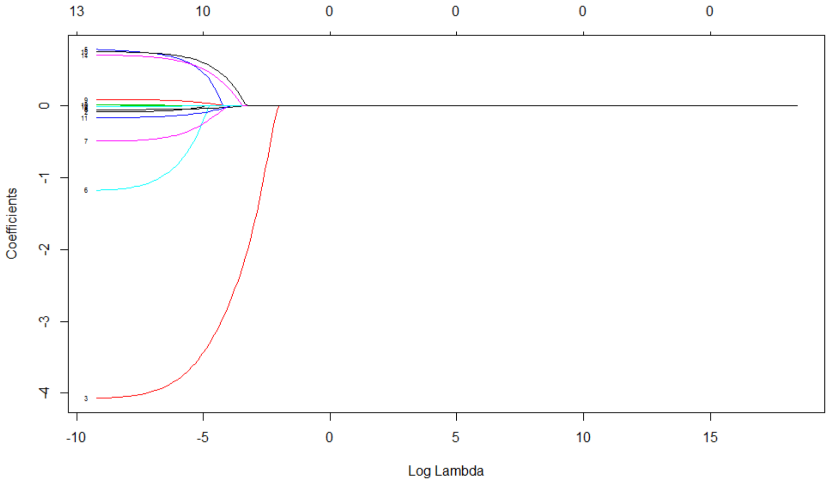
I will now focus on how each attribute is correlated to the existence of myopia. Figure4 represents the correlation between each attribute and the existence of myopia. According to figure the attributes which are highly correlated to price are "SPHEQ", "ACD", ‘'MOMMY", "DADMY", "SPORTHR" , "READHR","GENDER".



```
> pcor
         ID    STUDYYEAR       MYOPIC          AGE       GENDER        SPHEQ           AL          ACD           LT 
0.012242256  0.016330987  1.000000000  0.018525875  0.061556801 -0.373639054  0.037752311  0.107952757 -0.045704451 
        VCD      SPORTHR       READHR       COMPHR      STUDYHR         TVHR    DIOPTERHR        MOMMY        DADMY 
0.011854862 -0.098282028  0.072749265  0.025874323 -0.031858867 -0.004032443  0.036983991  0.134032827  0.149896423 
```

Figure 4 Variable correlation to myopia

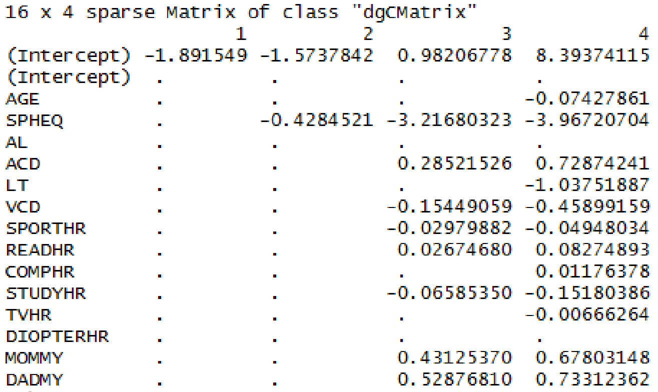
2. Attribute reduction



I will now use Glmnet Lasso in order to further examine which variables should be included in the model model. Having a set of input measurements x1, x2 ...xp and an outcome measurement of myopia, the lasso aims to minimize sum((y-yhat)^2). The Cp statistic is an estimate of the mean-square error in a model based on a selected subset of predictors, corrected for the number of predictors.

Figure 5 Lasso: attribute selection

Each colored line of figure 5 represents the value of a different coefficient in the model. Attribute number 3 is SPHEQ which as it was mentioned before is strongly associated to myopia. Lambda is the weight given to the regularization term (the L1 norm). When lambda approaches zero, the loss function of the model approaches the ordinary least squares (OLS) loss function. OLS is the method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed and predicted values. Therefore, when lambda is very small, the LASSO solution should be very close to the OLS solution, and all of the coefficients will be included in the model. L1 norm is the regularization term for LASSO. So when L1 norm is small, the regularization is high. Therefore, an L1 norm of zero gives an empty model, and increased L1 norm, result more variables to be characterized as significant. Lambda.min is the value of λ that gives minimum mean cross-validated error. However lamda.min was too complex and over fitted. lambda.1se, returns the most regularized model such that error is within one standard error of the minimum.



```
16 x 4 sparse Matrix of class "dgCMatrix"
                    1          2           3           4
(Intercept) -1.891549 -1.5737842  0.98206778  8.39374115
(Intercept)   .         .           .           .
AGE           .         .           .         -0.07427861
SPHEQ         .        -0.4284521 -3.21680323 -3.96720704
AL            .         .           .           .
ACD           .         .          0.28521526  0.72874241
LT            .         .           .         -1.03751887
VCD           .         .         -0.15449059 -0.45899159
SPORTHR       .         .         -0.02979882 -0.04948034
READHR        .         .          0.02674680  0.08274893
COMPHR        .         .           .          0.01176378
STUDYHR       .         .         -0.06585350 -0.15180386
TVHR          .         .           .         -0.00666264
DIOPTERHR     .         .           .           .
MOMMY         .         .          0.43125370  0.67803148
DADMY         .         .          0.52876810  0.73312362
```

Figure 6 Lasso: selected attributes

According to lasso attributes “SPHEQ”, ‘’ACD”, “VCD”, “SPORTHR”, “READHR”, “STUDYHR”, “MOMMY”, “DADMY” are important for predicting the existence of myopia. On the contrary, “AL” and “DIOPTERHR” should be excluded from the prediction model. DIOPTERHR equals to 3× (READHR + STUDYHR) + 2 × COMPHR + TVHR and in order to avoid collinearity issues it will be excluded from the model.

3. Attribute visualization

I added a new variable called “PARENTS” which equals to MOMMY + DADMY. Visualizing the dataset’s variables will help us to better comprehend them.

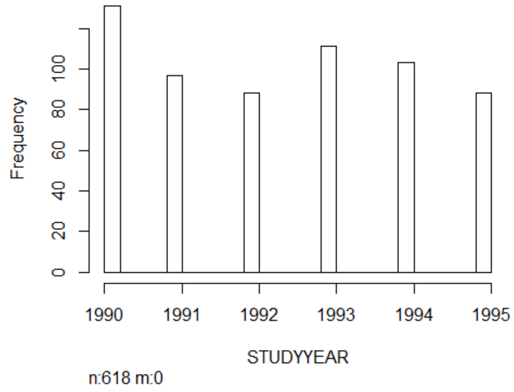


Figure 7 Frequency histogram: STUDYYEAR

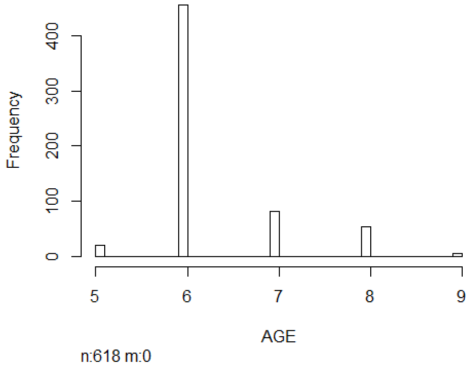


Figure 8 Frequency histogram: AGE

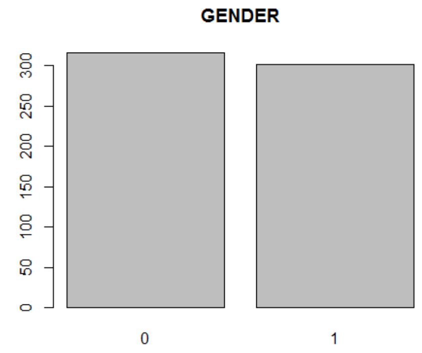


Figure 9 Frequency bar plot: PARENTS

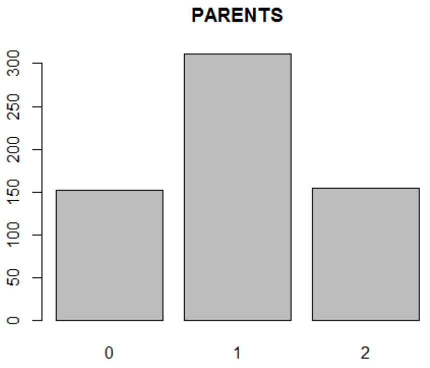


Figure 10 Frequency bar plot: GENDER

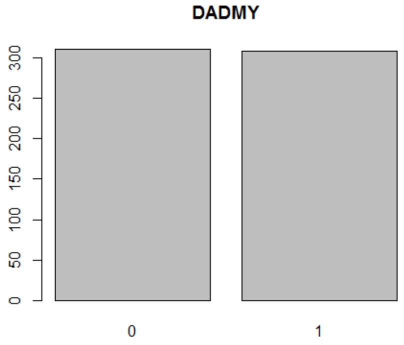


Figure 11 Frequency bar plot: MOMMY

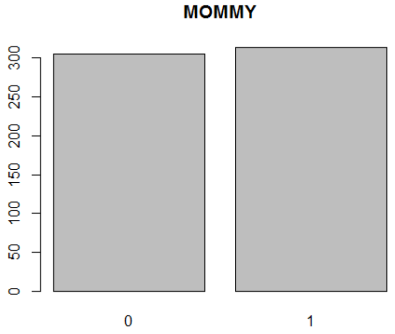


Figure 12 Frequency bar plot: DADMY

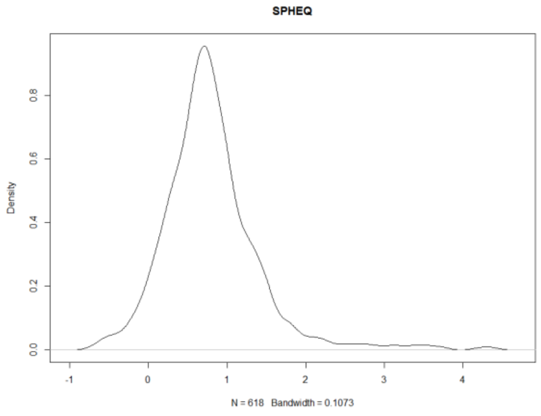


Figure 13 Box plot: SPHEQ

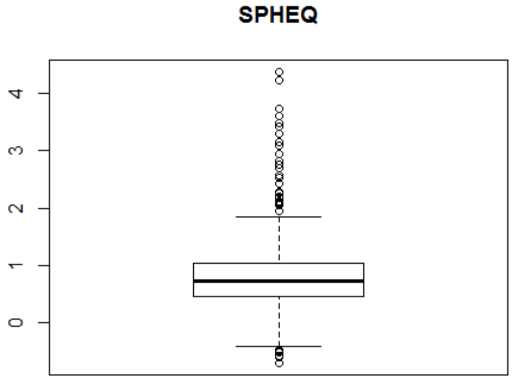


Figure 14 Density plot: SPHEQ

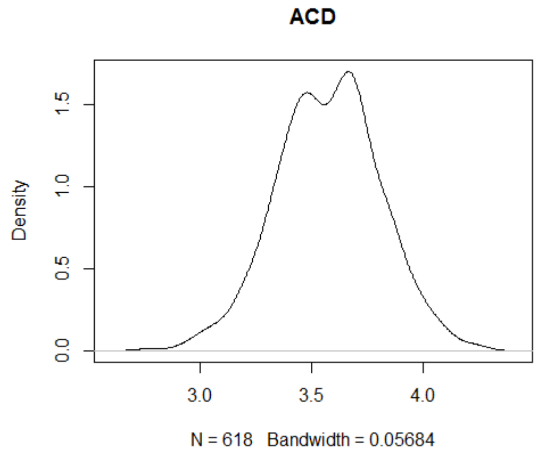


Figure 15 Density plot: ACD

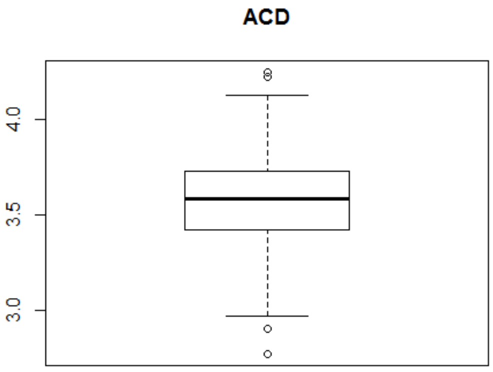


Figure 16 Box plot: ACD

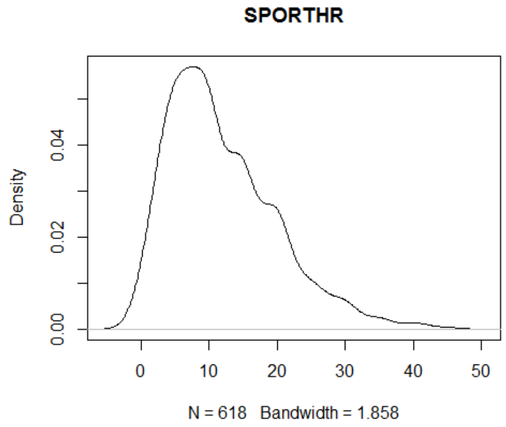


Figure 17 Density plot: SPORTHR

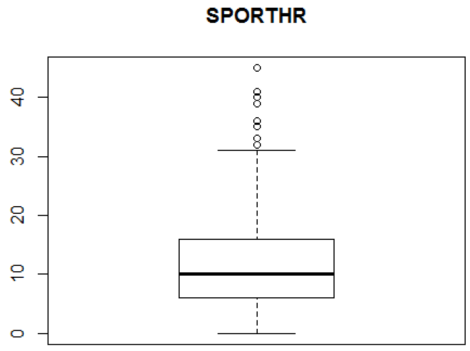


Figure 18 Box plot: SPORTHR

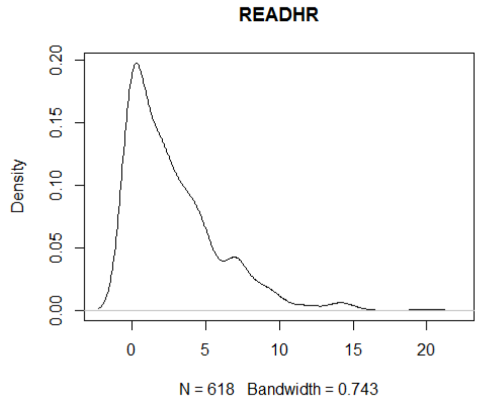


Figure 19 Density plot: READHR

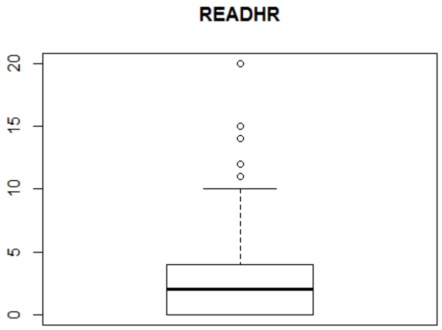
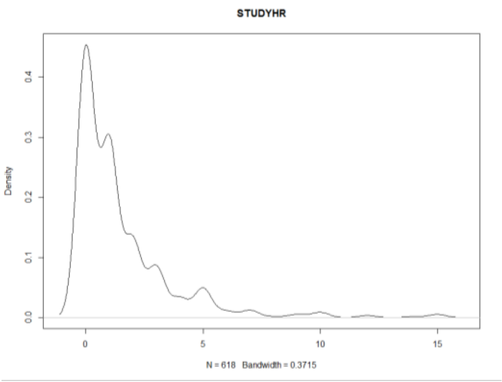


Figure 20 Box plot: READHR



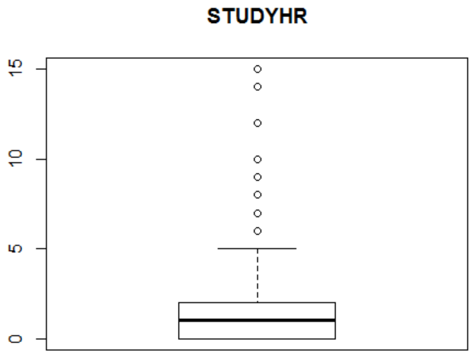


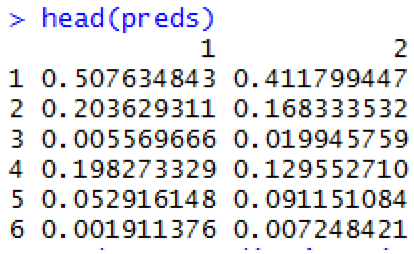
Figure 21 Box plot: STUDYHR

Figure 22 Density plot: STUDYHR

4. Lasso Model

4.1. Train Lasso model

Least Absolute Shrinkage and Selection Operator (LASSO) was used for attribute selection but it can also be used for prediction. More specifically it creates a regression model that is penalized with the L1-norm which is the sum of the absolute coefficients. This has the effect of shrinking coefficient values (and the complexity of the model), allowing some with a minor effect to the response to become zero



```
> head(preds)
            1           2
1 0.507634843 0.411799447
2 0.203629311 0.168333532
3 0.005569666 0.019945759
4 0.198273329 0.129552710
5 0.052916148 0.091151084
6 0.001911376 0.007248421
```

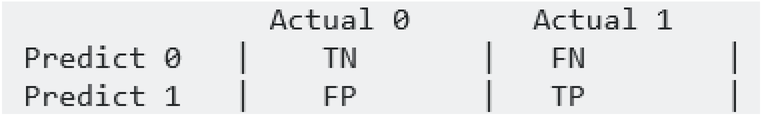
Figure 23 Lasso model: Choose lamda

The first and second column in figure represents the predictions made using lamda.min and lamda.1se accordingly. I chose to use lamda.min because it led to a better prediction.

4.2. Evaluate Lasso model

I will now examine how well did the model perform. To do so I need the compare the predicted values with the actual values. I will use a confusion matrix and three measurements, precision, recall and F-score.

In this case where the predicted value is either 1 or 0, the confusion matrix is:



```
             Actual 0         Actual 1
Predict 0   |    TN       |    FN       |
Predict 1   |    FP       |    TP       |
```

Figure 24 Confusion matrix

Where:

- TP = true positive (declare H1 when, in truth, H1)
- FN = false negative (declare H0 when, in truth, H1)
- FP = false positive
- TN = true negative

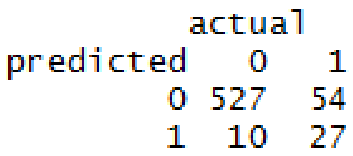
Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on understanding and measuring relevance.

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn} \qquad F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Figure 25 Precision, Recall, F-Score

F-score is a measure that combines precision and recall. F-score is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean.

The lasso fitted model resulted the following confusion matrix.



```
         actual
predicted   0   1
        0 527  54
        1  10  27
```

Figure 26 Confusion matrix: Lasso model

Where:

- accuracy = 554 / 618 = 0.90
- precision = 27/(27+10) = 0.73
- recall = 27/(27+55) = 0.33

5. GLM Model

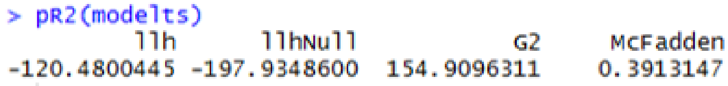
5.1. Logistic regression assumptions

We make three assumptions when using a logistic regression model. We will now examine whether our data set meets these assumptions.

5.1.1. Goodness of fit

The first assumption is that the model fits the data. R-squared is a commonly used goodness of fit test. However, when analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds have been developed. These are "pseudo" R-squareds because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squareds can arrive at very different values.

Logistic regression models are fitted using the method of maximum likelihood - i.e. the parameter estimates are those values which maximize the likelihood of the data which have been observed. McFadden's R squared measure is defined as 1-l_mod/l_null, where l_mod is the log likelihood value for the fitted model and l_null is the log likelihood for the null model which includes only an intercept as predictor. Pseudo R2 does not take values close to 1 like the linear regression R2 does. In logistic regression McFadden values from 0.2-0.4 indicate excellent model fit. McFadden equals to 0.39 which indicates goodness of fit.



```
> pR2(modelts)
         llh      llhNull           G2     McFadden
-120.4800445 -197.9348600  154.9096311    0.3913147
```

Figure 27 McFadden R-squared goodness of fit test

Furthermore, the Pearson χ^2 is similar to the residual sum of squares used in linear models.

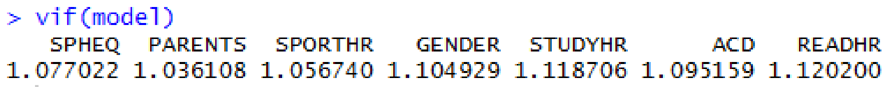
```
> sum(residuals(model, type = "pearson")^2)
[1] 574.3662
> deviance(model)
[1] 303.761
> 1 - pchisq(deviance(model), df.residual(model))
[1] 1
```

The p-value is large indicating no evidence of lack of fit. Concluding, the model meets the goodness of fit assumption.

5.1.2. Independence of observations

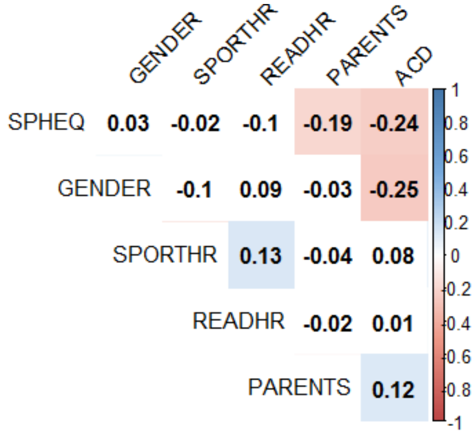
The observations are all independent. Each measurement was taken by different children. There is a small probability measuring two children from the same family. However in the majority of the cases, study subjects were not related. The age of the subjects was not independent since most of the children were in the same class. However, attribute "AGE" will not be included in the model. Therefore, we can suggest that we have independent observations.

5.1.3. Multicollinearity



```
> vif(model)
   SPHEQ  PARENTS  SPORTHR   GENDER  STUDYHR      ACD   READHR 
1.077022 1.036108 1.056740 1.104929 1.118706 1.095159 1.120200 
```

Figure 28 Vif: Test of collinearity test



	GENDER	SPORTHR	READHR	PARENTS	ACD
SPHEQ	0.03	-0.02	-0.1	-0.19	-0.24
GENDER		-0.1	0.09	-0.03	-0.25
SPORTHR			0.13	-0.04	0.08
READHR				-0.02	0.01
PARENTS					0.12

Figure 29 Correlation matrix

There is no multicollinearity between attributes which were used in the model. Lasso was used for attribute selection. Lasso considers collinearity during the attribute selection process. Moreover, if we plot the correlation between the variables which were used as input in the model (figure 28), we can see that there no significant collinearity. Finally, A VIF value >= 10 indicates high collinearity and inflated standard errors. In this case all vif values are close to 1 (figure 29).

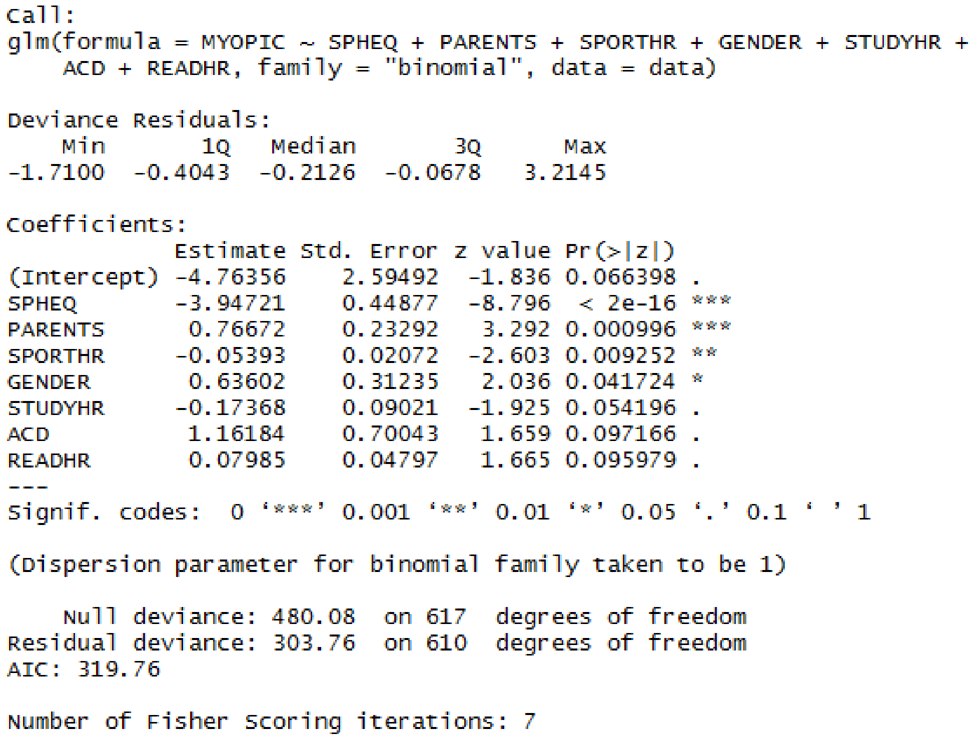
5.2.Train GLM model

I will also use the glm() function for fitting a logistic regression model. Logistic regression is used for fitting a model y=f(x), when y>0. In this scenario y is the binary attribute "MYOPIC" which means that its values are either 1 if the kid has myopia or 0 if not. Since the predicted variable is binary , I chose to use a model called "binomial logistic regression".

Automatic methods are useful when the number of explanatory variables is large and it is not easy to fit all possible models. In this case, it is more efficient to use a search algorithm (e.g., Forward selection, Backward elimination and Stepwise regression) to find the best model. The R function step() can also be used to perform variable selection.

```
mnull <- glm(MYOPIC ~1, data = data, family = "binomial")

summary(mnull)

mfull <-glm(MYOPIC~.-ID-STUDYYEAR-AL-DIOPTERHR, data = data,

family = "binomial")

summary(mfull)

summary(step(mnull, scope=list(lower=mnull,upper=mfull), direction='both' ))

model <- glm(formula = MYOPIC ~ SPHEQ + PARENTS + SPORTHR + GENDER + STUDYHR
+ ACD + READHR, family = "binomial", data = data)
```

To perform both ways selection we need to begin by specifying the null model which is the constant model and the full model which contains all the attributes affecting the existence of myopia. A range of all possible models found between the null and the full model will be examined using search(). This tells R to start with the null model and search through models lying in the range between the null and full model using the both ways selection algorithm. It gives rise to the following output. According to this procedure, the best model is the one that includes the variables SPHEQ, PARENTS, SPORTHR, GENDER, STUDYHR and READHR.

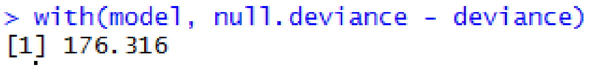


```
Call:
glm(formula = MYOPIC ~ SPHEQ + PARENTS + SPORTHR + GENDER + STUDYHR +
    ACD + READHR, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7100  -0.4043  -0.2126  -0.0678   3.2145

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.76356    2.59492  -1.836 0.066398 .
SPHEQ       -3.94721    0.44877  -8.796  < 2e-16 ***
PARENTS      0.76672    0.23292   3.292 0.000996 ***
SPORTHR     -0.05393    0.02072  -2.603 0.009252 **
GENDER       0.63602    0.31235   2.036 0.041724 *
STUDYHR     -0.17368    0.09021  -1.925 0.054196 .
ACD          1.16184    0.70043   1.659 0.097166 .
READHR       0.07985    0.04797   1.665 0.095979 .
---
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: 303.76  on 610  degrees of freedom
AIC: 319.76

Number of Fisher Scoring iterations: 7
```

Figure 30 GLM model summary

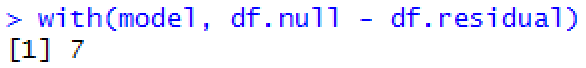
On the top of figure 30 we can see the model that was called. Next, we see the deviance residuals, which are a measure of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model. The next part of the output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. Attributes with stars on the right of the table such as SPHEQ and PARETNS are statistically significant. Low P-values suggest strong association to the predicted attribute, which means that “SPHEQ” is strongly associated to existence or absence of myopia to children. The negative coefficient for this predictor suggests that all other variables being equal, the higher the SHPEQ value is the less likely to have myopia. More specifically, for every one diopter change in spheq, the log odds of a child to have myopia decreases by 3.95. In the logit model the response variable is log odds: ln(odds) = ln(p/(1-p)) = a*x1 + b*x2 + … + z*xn. Since GENDER is a dummy variable, in female (female=1) kids the log odds of having myopia are increased by 0.64 while an hour per week increase in “SPORTHR” reduces the log odds by 0.17. Moreover, for a one unit increase in parents, the log odds of their child to have myopia increases by 0.77.

Below the table of coefficients there are fit indices, including the null and deviance residuals and the AIC. The lower the AIC and deviance residuals, the better. We want to see measures of how well our model fits. The output produced by summary (model) included indices of fit (shown below the coefficients), including the null and deviance residuals and the AIC. One measure of model fit is the significance of the overall model. This test asks whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of predictor variables in the model). To find the difference in deviance for the two models (i.e., the test statistic) we can use the command:



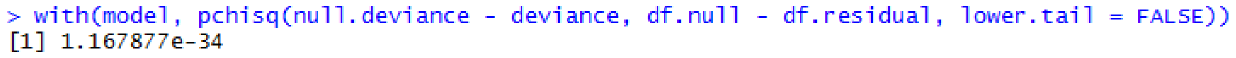
```
> with(model, null.deviance - deviance)
[1] 176.316
```

The degrees of freedom for the difference between the two models is equal to the number of predictor variables in the mode, and can be obtained using:



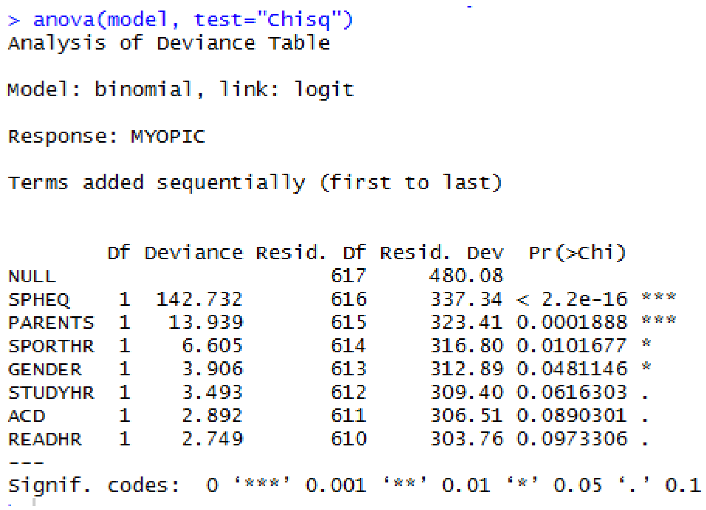
```
> with(model, df.null - df.residual)
[1] 7
```

Finally, the p-value can be obtained using:



```
> with(model, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 1.167877e-34
```

The chi-square of 176.316 with 7 degrees of freedom and an associated p-value of less than 0.001 tells us that our model as a whole fits significantly better than an empty model.

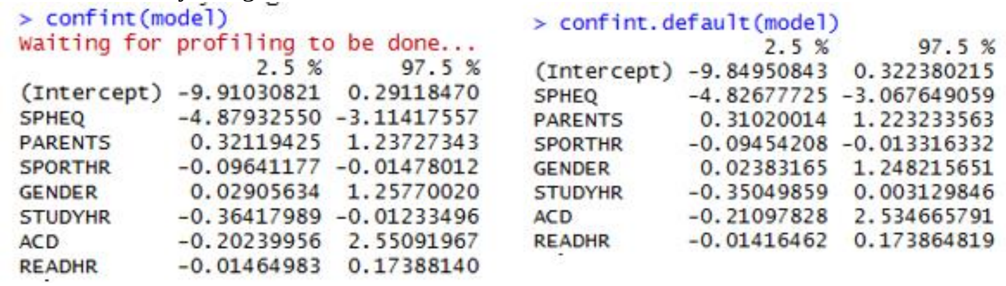


```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: MYOPIC

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                      617     480.08              
SPHEQ    1  142.732       616     337.34 < 2.2e-16 ***
PARENTS  1   13.939       615     323.41 0.0001888 ***
SPORTHR  1    6.605       614     316.80 0.0101677 *  
GENDER   1    3.906       613     312.89 0.0481146 *  
STUDYHR  1    3.493       612     309.40 0.0616303 .  
ACD      1    2.892       611     306.51 0.0890301 .  
READHR   1    2.749       610     303.76 0.0973306 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Figure 31 Anova: Analyze table of deviance

I will now use the anova (figure 31) function on the model to analyze the table of deviance. The difference between the null deviance and the residual deviance suggest how much better our model is comparing to the null model, the model with only the intercept. By adding variables to the model we are reducing the residual deviance. Our aim is to minimize the residual deviance. When adding important attributes to the model such as SPHEQ and PARENTS we get a greater reduction of the residual deviance. Attributes READHR, ACD and STUDYHR do improve the model but result lower residual deviance reductions.

We can use the confint function to obtain confidence intervals for the coefficient estimates. Note that for logistic models, confidence intervals are based on the profiled log-likelihood function. We can also get CIs based on just the standard errors by using the default method.



```
> confint(model)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -9.91030821  0.29118470
SPHEQ       -4.87932550 -3.11417557
PARENTS      0.32119425  1.23727343
SPORTHR     -0.09641177 -0.01478012
GENDER       0.02905634  1.25770020
STUDYHR     -0.36417989 -0.01233496
ACD         -0.20239956  2.55091967
READHR      -0.01464983  0.17388140
```

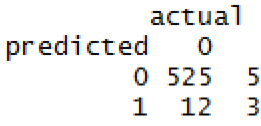
```
> confint.default(model)
                  2.5 %       97.5 %
(Intercept) -9.84950843  0.322380215
SPHEQ       -4.82677725 -3.067649059
PARENTS      0.31020014  1.223233563
SPORTHR     -0.09454208 -0.013316332
GENDER       0.02383165  1.248215651
STUDYHR     -0.35049859  0.003129846
ACD         -0.21097828  2.534665791
READHR      -0.01416462  0.173864819
```

Figure 32 Confint: Coefficient confidence intervals

The confint.default (figure 32) function in the MASS library generates the Wald confidence limits, while the confint() function produces the profile-likelihood limits. The profile-likelihood method is thought to be superior, especially for small sample sizes like this one.

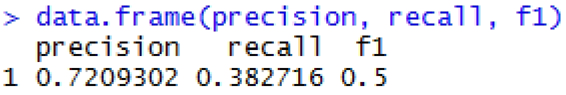
5.3.Evaluate GLM model

In order to examine how well did the model perform. To do so I need the compare the predicted values with the actual values. I will use a confusion matrix and three measurements, precision, recall and F-score.



```
         actual
predicted   0   1
        0 525  50
        1  12  31
```

Figure 33 GLM model: confusion matrix



```
> data.frame(precision, recall, f1)
  precision   recall  f1
1 0.7209302 0.382716 0.5
```

Figure 34 GLM model: precision, recall & F-score

Where:

- accuracy = 554 / 618 = 0.90
- precision = 31/(31+12) = 0.72
- recall = 31/(31+50) = 0.38

However, by setting the parameter type='response', R will output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y=0. In this application it is possible that different threshold could be a better option. The accuracy of the model is 88% which is quite good. However, this results depends on the manual split (0.5) that I did.

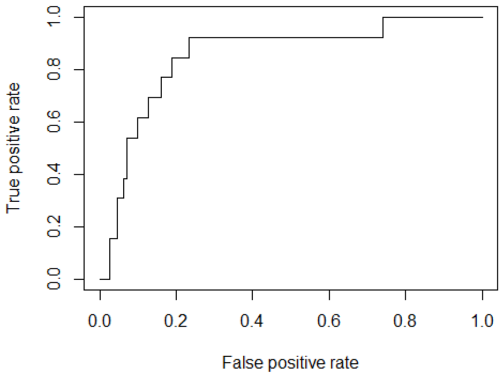
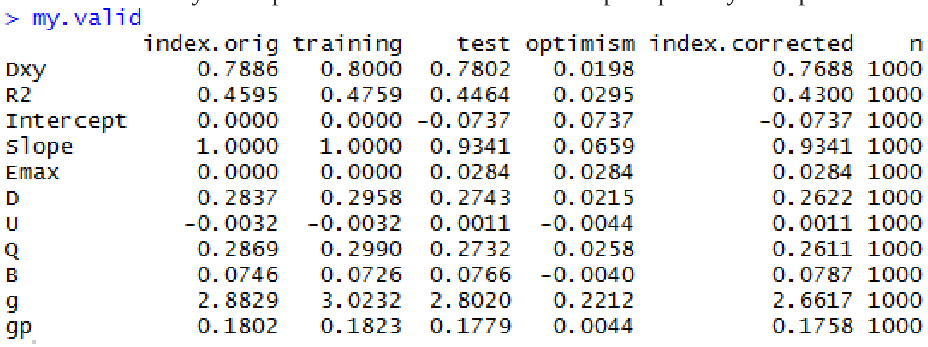


Figure 35 ROC curve

I will now plot the ROC curve (figure 35) and calculate the “Under the curve” (AUC) in order to measure the performance of the classifier. ROC plots the true positive rate against the false positive rate. AUC is the area under the ROC curve. A model with good predictive ability should have an AUC closer to 1 than to 0.5. AUC equals to 0.85 which is close to 1 (optimum) and therefore suggests that the model has good predictive ability.

The model is likely overoptimistic. We now use bootstrap to quantify the optimism:



```
> my.valid
          index.orig training    test optimism index.corrected    n
Dxy           0.7886   0.8000  0.7802   0.0198          0.7688 1000
R2            0.4595   0.4759  0.4464   0.0295          0.4300 1000
Intercept     0.0000   0.0000 -0.0737   0.0737         -0.0737 1000
Slope         1.0000   1.0000  0.9341   0.0659          0.9341 1000
Emax          0.0000   0.0000  0.0284   0.0284          0.0284 1000
D             0.2837   0.2958  0.2743   0.0215          0.2622 1000
U            -0.0032  -0.0032  0.0011  -0.0044          0.0011 1000
Q             0.2869   0.2990  0.2732   0.0258          0.2611 1000
B             0.0746   0.0726  0.0766  -0.0040          0.0787 1000
g             2.8829   3.0232  2.8020   0.2212          2.6617 1000
gp            0.1802   0.1823  0.1779   0.0044          0.1758 1000
```

Figure 36 Bootstrap

On the top of figure 36, Dxy equals to 0.7688. The column called optimism denotes the amount of estimated overestimation by the model. The column index.corrected is the original estimate minus the optimism. In this case, the bias-corrected Dxy is a bit smaller than the original. The bias-corrected c-index (AUC) is c=(1+Dxy)/2=0.8949. We can also calculate a calibration curve using resampling (figure 37).

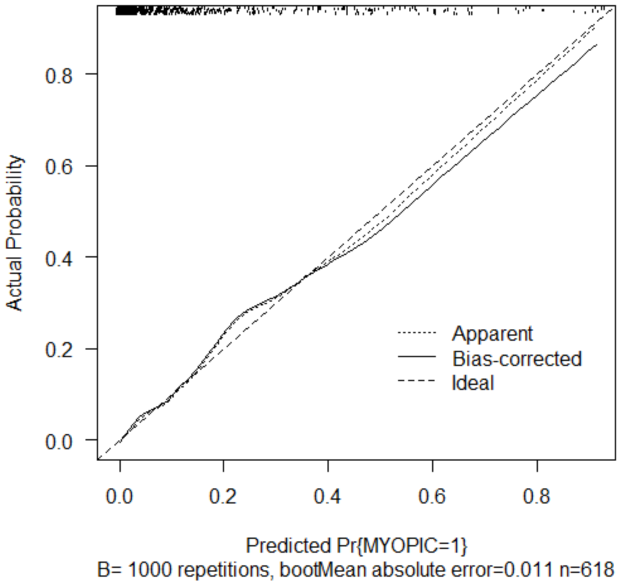


Figure 37 Calibration curve

The plot provides some evidence that our models is overfitting: the model underestimates low probabilities and overestimates high probabilities. There is also a systematic underestimation around 0.2.

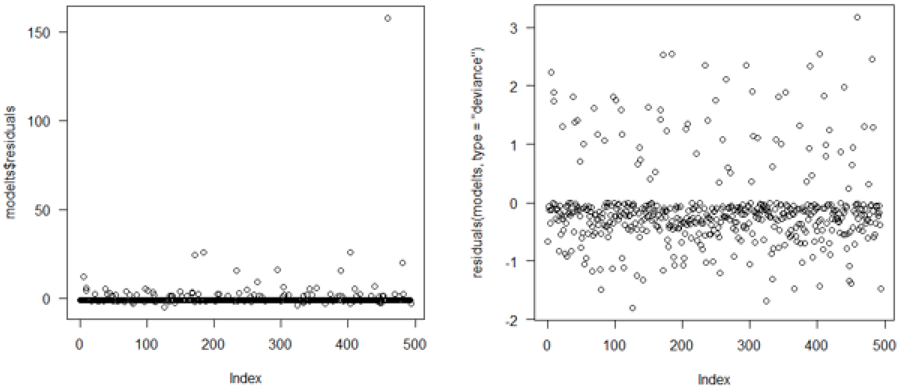
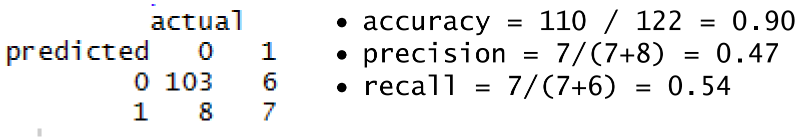


Figure 38 GLM model: residuals

The fact the same data was used as training and test set may led to overfitting. Therefore it is useful to split the data set into two parts, training and testing set. The training set will be used to fit our model which I will be testing over the testing set.

Where:



```
          actual
predicted   0   1
        0 103   6
        1   8   7
```

- accuracy = 110 / 122 = 0.90
- precision = 7/(7+8) = 0.47
- recall = 7/(7+6) = 0.54

Figure 39 GLM model, split train and test set: confusion matrix

6. Discussion

This dataset is an imbalanced dataset**.** Only 15% of the dataset's subjects are myopic students. According to the confusion matrix, the model's accuracy is high, approximately 90%. The model predicts 527/537 cases for non-myopic students correctly which is the reason for high accuracy. However, the model is unable to successfully predict myopic students. This is caused due to the class imbalance between myopic and non-myopic children. In such cases, models with lower accuracy levels may have better predictive power, better precision, recall and F score.

Moreover, the size of the dataset is small and the subjects of this study came from the same school and have approximately the same age. Concluding, this statistic analysis provides insights concerning the correlation between the studied variables and the existence of myopia in children, but the fitted model does not have strong predictive power.

7. Bibliography

Precision, Recall & F-score

https://en.wikipedia.org/wiki/Precision_and_recall

Lasso

http://www.ats.ucla.edu/stat/r/dae/logit.htm

http://machinelearningmastery.com/penalized-regression-in-r/

GLM model assumptions

http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram_______________

http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/

http://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation

https://www.r-bloggers.com/veterinary-epidemiologic-research-glm-evaluating-logistic-regression-models-part-3/

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

https://rpubs.com/ryankelly/ml_logistic

Train GLM model

http://rstudio-pubs-static.s3.amazonaws.com/82432_cb82f2f0111d4e01b8e74edc5720eb7a.html

https://datascienceplus.com/perform-logistic-regression-in-r/

https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

https://rpubs.com/gin/209945

http://pubs.thetazero.com/view/peperomia-203313

http://www.shizukalab.com/toolkits/plotting-logistic-regression-in-r

http://rcompanion.org/rcompanion/e_07.html

Evaluate GLM model

http://nlp.stanford.edu/manning/courses/ling289/logistic.pdf

http://stats.stackexchange.com/questions/64788/interpreting-a-logistic-regression-model-with-multiple-predictors

https://datascienceplus.com/perform-logistic-regression-in-r/

https://www.r-bloggers.com/optimizing-probability-thresholds-for-class-imbalances/

http://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html

8. R Code

```
#print first 6 rows with kable function from knitr
knitr::kable(head(data))

#use SPHEQ to estimate myopia
library(ggplot2)
ggplot(myopia, aes(x = SPHEQ, y = MYOPIC))
+ geom_jitter(shape = "O", position = position_jitter(height = 0))
+theme_bw()

#plot correlation between variables
library(corrplot)
corrplot(cor(data))

#plot correlation between attributes and myopia
correlations <- cor(myopia)
pcor <- correlations[,3]
library(corrplot)
corrplot(cor(myopia),method='e')

#Lasso: Attribute selection
library(glmnet)
myopia_mat <- model.matrix(MYOPIC~.-ID-STUDYYEAR, myopia)[,-3]
lambdas <- 10 ^ seq(8,-4,length=250)
myopia_models_lasso <- glmnet(myopia_mat,myopia$MYOPIC,alpha=1,
lambda=lambdas, family="binomial")
plot(myopia_models_lasso, xvar = "lambda", label = TRUE)
lasso.cv <- cv.glmnet(myopia_mat,myopia$MYOPIC, alpha=1,
lambda=lambdas, family="binomial")
lasso.cv <- cv.glmnet(myopia_mat,myopia$MYOPIC,alpha=1,
lambda=lambdas, family="binomial", type.measure = "auc")
coef(lasso.cv , s = c(1,0.1,0.01,0.001))
lasso.cv$lambda.min
#[1] 0.004861239

#Lasso: Prediction
library(glmnet)
myopia_mat <- model.matrix(MYOPIC~.-ID-STUDYYEAR, data)[,-3]
lambdas <- 10 ^ seq(8,-4,length=250)
myopia_models_lasso <- glmnet(myopia_mat,myopia$MYOPIC,alpha=1,
lambda=lambdas, family="binomial")
predict(myopia_models_lasso, type="coefficients",
s = lasso.cv$lambda.min)
preds <- predict(myopia_models_lasso, myopia_mat, type = "response", s = c(las
so.cv$lambda.min, lasso.cv$lambda.1se))
head(preds)
preds <- predict(myopia_models_lasso, myopia_mat, type = "class", s = lasso.cv
$lambda.min)

#Evaluate lasso model performance
table(predicted = preds, actual = myopia$MYOPIC)
#actual vs predicted
mean(preds  == myopia$MYOPIC)
#lasso model accuracy
#[1] 0.8980583
```

```
#Evaluate logistic regression assumptions
#Goodness of fit
pR2(model)
#
sum(residuals(model, type = "pearson")^2)
#[1] 574.3662
deviance(model)
#[1] 303.761
1 - pchisq(deviance(model), df.residual(model))
#[1] 1

#collinearity
collin <- cor(subset(myopia, select=c(SPHEQ, PARENTS, SPORTHR, GENDER ,ACD ,
READHR)))
dev.off()
library(corrplot)
corrplot(collin , type="upper")
M <- collin
cor.mtest <- function(mat, ...) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat<- matrix(NA, n, n)
  diag(p.mat) <- 0
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], ...)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
    }
  }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
# matrix of the p-value of the correlation
p.mat <- cor.mtest(subset(myopia, select=c(SPHEQ, PARENTS,
SPORTHR, GENDER ,ACD , READHR)))
head(p.mat[, 1:6])
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF",
"#77AADD", "#4477AA"))

#create correlation plot
corrplot(M, method="color", col=col(200),
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, #Text label color and rotation
         # Combine with significance
         p.mat = p.mat, sig.level = 0.01, insig = "blank",
         # hide correlation coefficient on the principal diagonal
         diag=FALSE
)

#Fit GLM model
myopia$READING <- myopia$STUDYHR + myopia$READHR
mnull <- glm(MYOPIC ~1, data = data, family = "binomial")
summary(mnull)
mfull <-glm(MYOPIC~.-ID-STUDYYEAR-AL-DIOPTERHR, data = data,
 family = "binomial")
summary(mfull)
summary(step(mnull, scope=list(lower=mnull,upper=mfull),
direction='both' ))
model1 <- glm(formula = MYOPIC ~ SPHEQ + PARENTS + SPORTHR + GENDER
+ STUDYHR + ACD + READHR, family = "binomial", data = data)

#evaluate GLM model
with(model, null.deviance - deviance)
with(model, df.null - df.residual)
with (model, pchisq(null.deviance - deviance, df.null - df.residual,
lower.tail = FALSE))

#analyze table of deviance
anova(model, test="Chisq")

#evaluate GLM model accuracy
#set threshold 0.5
fitted.results <- predict(model,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$MYOPIC)
print(paste('Accuracy',1-misClasificError))
#[1] "Accuracy 0.879032258064516"

#plot ROC curve
library(ROCR)
p <- predict(model, newdata=test, type="response")
pr <- prediction(p, test$MYOPIC)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
#[1] 0.8537769

#split into different train and test set
# 80% of the sample size
smp_size <- floor(0.80 * nrow(data))
# set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)
train <- data[train_ind, ]
test <- data[-train_ind, ]
##threshold
fitted.results <- predict(modelts,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.3,1,0)
misClasificError <- mean(fitted.results != test$MYOPIC)
print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.887096774193548"
table(predicted = fitted.results , actual = test$MYOPIC)

modelts <- glm(formula = MYOPIC ~ SPHEQ + PARENTS + SPORTHR + GENDER
+ STUDYHR + ACD + READHR, family = "binomial", data = train )

#LOGISTIC
#Performance
table(predicted = fitted.results, actual = test$MYOPIC)
#actual predicted
mean(fitted.results  == test$MYOPIC)
#[1] 0.8790323
```

```
#residuals
plot(modelts$residuals)
plot(residuals(modelts, type="deviance") )

#quanify optimism
my.valid <- validate(model, method="boot", B=1000)
my.valid
my.calib <- calibrate(model, method="boot", B=1000)
par(bg="white", las=1)
plot(my.calib, las=1)
```