ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY
MSC IN BUSINESS ANALYTICS

ASSIGNMENT IN THE COURSE
«STATISTICS FOR BUSINESS ANALYTICS II»

PROJECT 1 - MYOPIA STUDY

ERASMIA KORNELATOU

ATHENS, 2020

Table of Contents

Introduction

The aim of this 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. The 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 is below:

Column	Description	Value/Unit	Name
1	Year subject entered the study	year	STUDYYEAR
2	Myopia within the first five years of follow up	0 = No; 1 = Yes	MYOPIC
3	Age at first visit	years	AGE
4	Gender	0 = Male; 1 = Female	GENDER
5	Spherical Equivalent Refraction	diopter	SPHEQ
6	Axial Length	mm	AL
7	Anterior Chamber Depth	mm	ACD

8	Lens Thickness	mm	LT
9	Vitreous Chamber Depth	mm	VCD
10	Time spent engaging in sports/outdoor activities	hours per week	SPORTHR
11	Time spent reading for pleasure	hours per week	READHR
12	Time spent playing video/computer games or working on the computer	hours per week	COMPHR
13	Time spent reading or studying for school assignments	hours per week	STUDYHR
14	Time spent watching television	hours per week	TVHR

Column	Description	Value/Unit	Name
15	Composite of near-work activities	hours per week	DIOPTERHR
16	Was the subject's mother myopic?	0 = No; 1 = Yes	MOMMY
17	Was the subject's father myopic?	0 = No; 1 = Yes	DADMY

Column 2: MYOPIC is defined as SPHEQ <= -0.75 D.
Column 5: A measure of the eye's effective focusing power. Eyes that are "normal" (don't require glasses or contact lenses) have spherical equivalents between -0.25 diopters (D) and +1.00 D. The more negative the spherical equivalent, the more myopic the subject.
Column 6: The length of eye from front to back.
Column 7: The length from front to back of the aqueous-containing space of the eye between the cornea and the iris.
Column 8: The length from front to back of the crystalline lens.
Column 9: The length from front to back of the aqueous-containing space of the eye in front of the retina.
Column 15: the composite is defined as DIOPTERHR = 3 × (READHR + STUDYHR) + 2 × COMPHR + TVHR.

Descriptive analysis and exploratory data analysis

In the first place, we are going to merge “MOMMY” and “DADMY” variables into a new variable “PARENTSMY” . This variable is going to take the value 1 when at least one of “MOMMY” or “DADMY” has value 1 (has myopia).

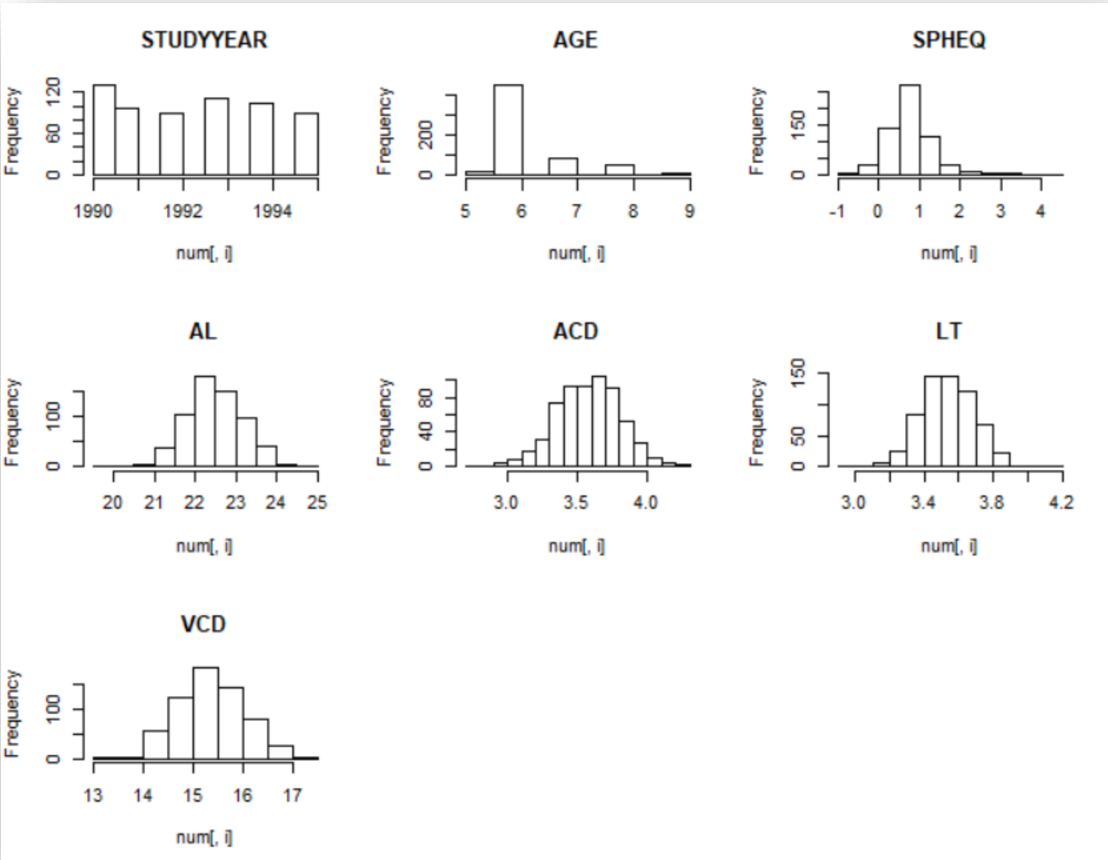
After having deleted ID column, as we are not going to use it, the structure of the variables is the below:

```
> str(myopia)
'data.frame':   618 obs. of  18 variables:
 $ 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 ...
 $ PARENTSMY: num  1 1 0 1 1 1 1 0 0 0 ...
```

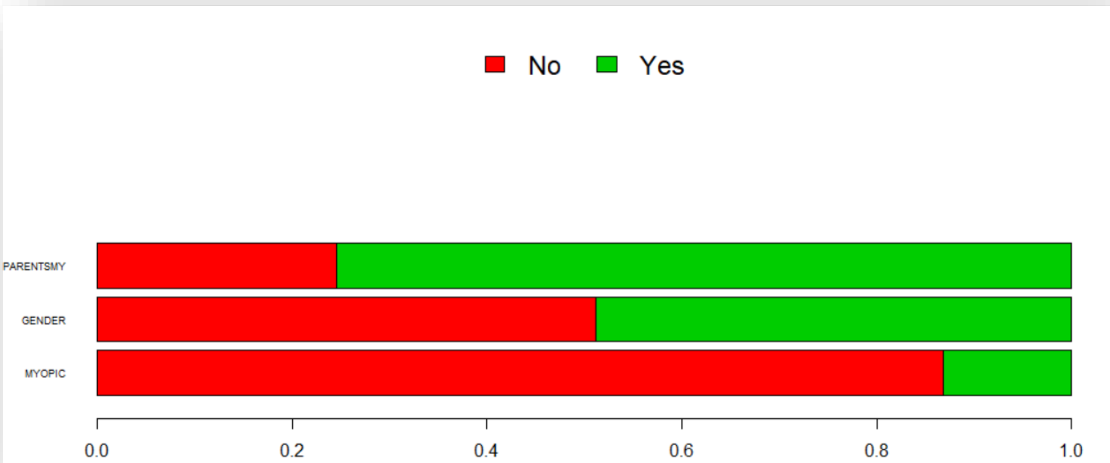
As we observe, the variables “MYOPIC”, “GENDER”, “PARENTSMY” have only 0 and 1 values. So, we should convert them to factor variables. We should also convert variables “STUDYYEAR”, “AGE”, “SPORTHR”, “READHR”, “COMPHR”, “STUDYHR”, “TVHR”, “DIOPTERHR” into numeric variables. After converting them, the types of the variables are as below:

```
STUDYYEAR    MYOPIC       AGE    GENDER     SPHEQ        AL       ACD        LT
"numeric"  "factor" "numeric"  "factor" "numeric" "numeric" "numeric" "numeric"
      VCD   SPORTHR    READHR    COMPHR   STUDYHR      TVHR DIOPTERHR PARENTSMY
"numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"  "factor"
```

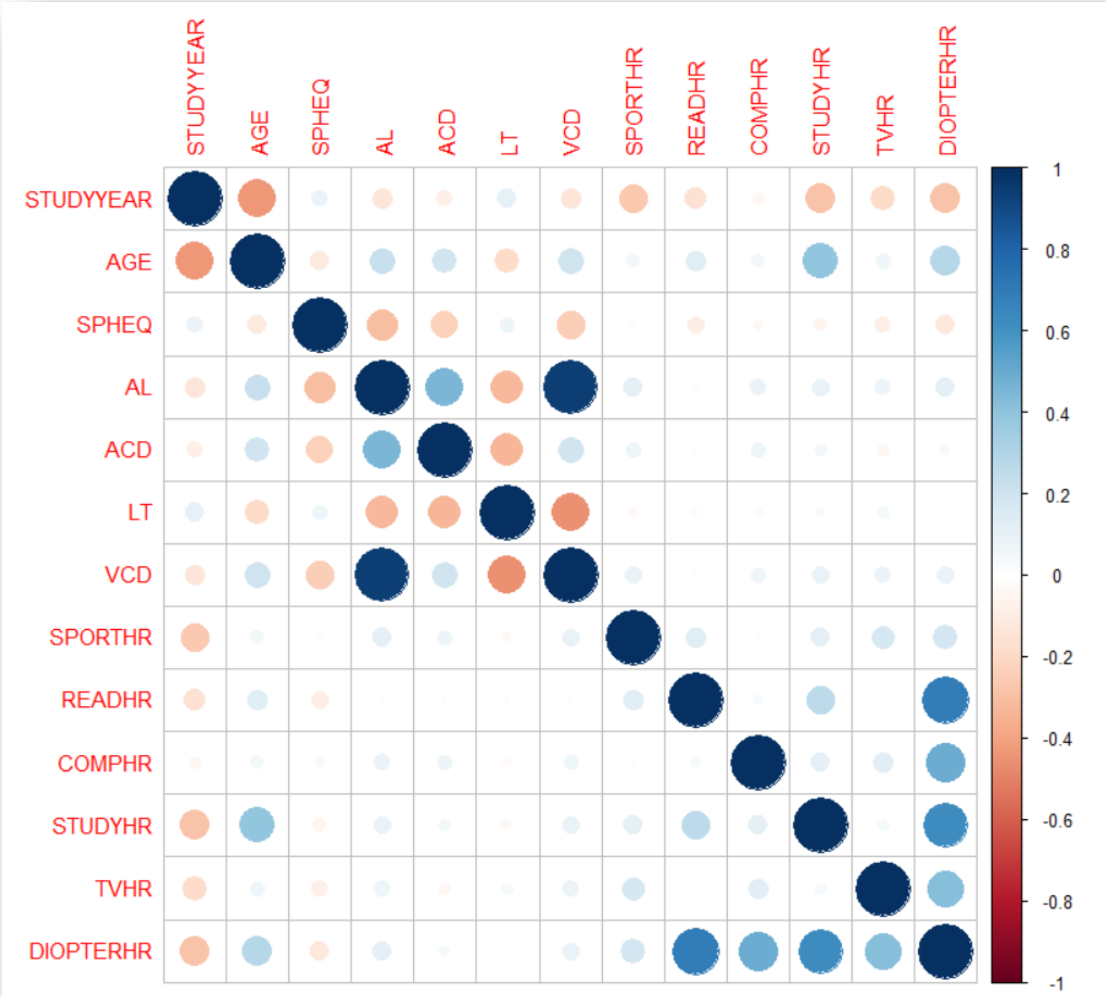
Analysis for numerical variables

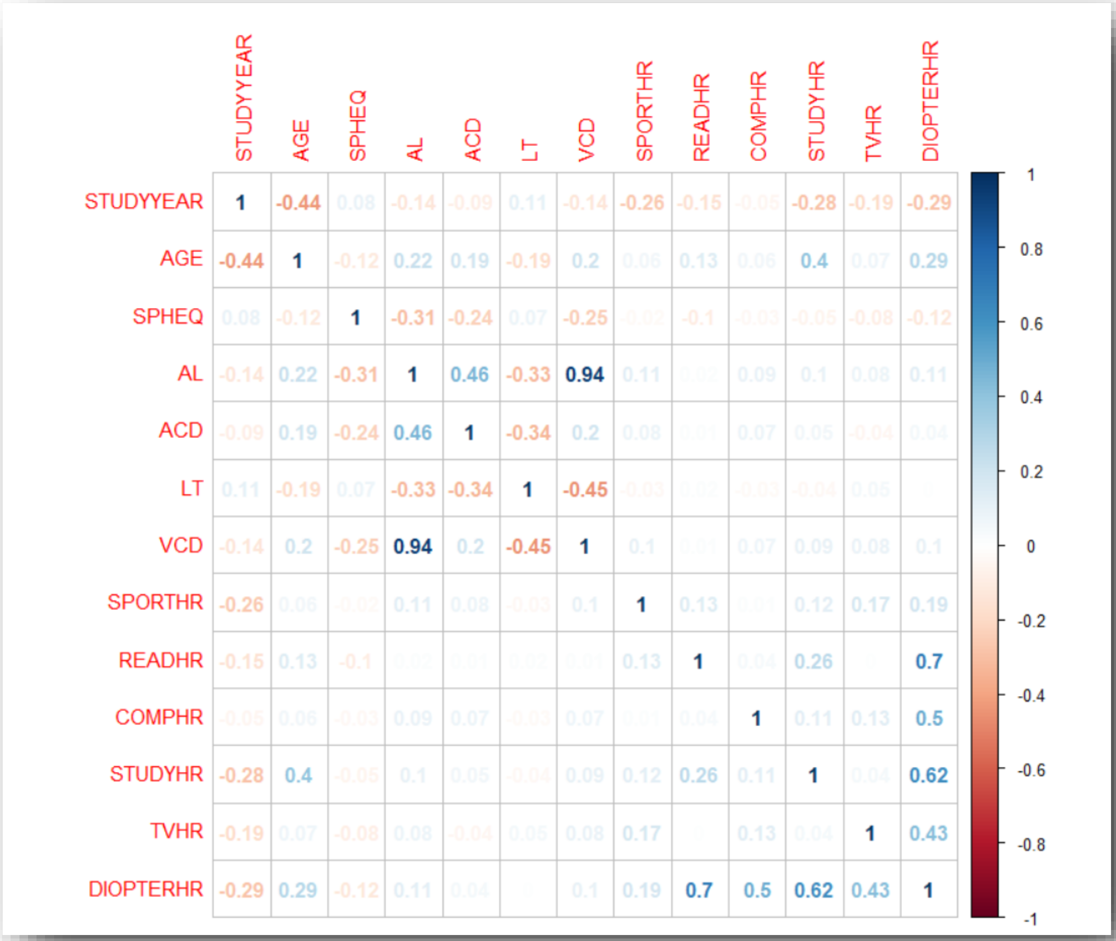


Analysis for factor variables



Vizualization of bivariate assosiations





At first glance, we notice that there is a multicollinearity issue in our data. Specifically, there is high correlation between “DIOPTERHR” and “READHR”, “COMPHR”, “STUDYHR”, “TVHR” variables. Also, there is high correlation between “AL” and “VCD” variable. To face this problem, there is high possibility of deleting the variable “DIOPTHER” and “AL” from the dataset , but let’s find more about the importance of knowledge of these variable for the knowledge of the variable “MYOPIC” .

Generally, we are going to use “Lasso” and “Stepwise” methods to find out for the importance of all variables for our response variable “MYOPIC” .

Methods for variable selection

Backward stepwise method for variable selection

Backward stepwise method is a stepwise regression approach that begins with a full model and at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data.

```
Step:  AIC=345.39
MYOPIC ~ SPHEQ + SPORTHR + PARENTSMY

            Df Deviance    AIC
<none>           320.53 345.39
- SPORTHR    1   327.95 346.60
- PARENTSMY  1   329.54 348.18
- SPHEQ      1   452.69 471.34

Call:  glm(formula = MYOPIC ~ SPHEQ + SPORTHR + PARENTSMY, family = "binomial",
    data = myopia)

Coefficients:
(Intercept)        SPHEQ      SPORTHR    PARENTSMY
   -0.54529     -3.83186     -0.05045      1.34430

Degrees of Freedom: 617 Total (i.e. Null);  614 Residual
Null Deviance:      480.1
Residual Deviance: 320.5        AIC: 328.5
```

According to the results of Backward stepwise method, the important variables of the dataset are “SPORTHR”, “PARENTSMY”, “SPHEQ” .

LASSO method for variable selection

The LASSO (Least Absolute Shrinkage and Selection Operator) is a method of automatic variable selection which can be used to select predictors X* of a target variable Y from a larger set of potential or candidate predictors X.

The LASSO formulates curve fitting as a quadratic programming problem, where the objective function penalizes the absolute size of the regression co-efficients, based on the value of a tuning parameter λ. In doing so, the LASSO can drive the coefficients of irrelevant variables to zero, thus performing auto-matic variable selection.

```
16 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -0.60270972
STUDYYEAR    .
AGE          .
GENDER       .
SPHEQ       -2.45886991
AL           .
ACD          .
LT           .
VCD          .
SPORTHR     -0.00851992
READHR       .
COMPHR       .
STUDYHR      .
TVHR         .
DIOPTERHR    .
PARENTSMY    0.26590132
```

According to the results of Lasso method, the important variables of the dataset are “SPORTHR”, “PARENTSMY”, “SPHEQ” .

Based on both results of the 2 methods mentioned , we come to the conclusion that the important variables for our model are “SPORTHR”, “PARENTSMY”, “SPHEQ”.

So, here is the final model:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.54529    0.56262  -0.969  0.33244
SPHEQ       -3.83186    0.43398  -8.830  < 2e-16 ***
PARENTSMY1   1.34430    0.51039   2.634  0.00844 **
SPORTHR     -0.05045    0.01973  -2.557  0.01056 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression assumptions

Before interpreting the final model, we should check the 3 logistic regression assumptions.

<u>Multicollinearity</u>

```
> vif(final_model)
    SPHEQ PARENTSMY   SPORTHR
 1.017105  1.000624  1.017263
```

As it seems , all the "VIF" values are not bigger than 10 . So, we do not face any multicollinearity issues.

Goodness of fit

```
> with(pchisq(deviance, df.residual), data = final_model)
[1] 5.423914e-25
```

The value of the pchisq test's result is quite small , so the model fits well.

Independence of observations

According to the description of the dataset , all observations are independent.

Interpretation of our Model

$$p(MYOPIC = 1) = \frac{e^{-0.54 - 3.83 * SPHEQ + 1.34 * PARENTSMY1 - 0.05 * SPORTHR}}{1 + e^{-0.54 - 3.83 * SPHEQ + 1.34 * PARENTSMY1 - 0.05 * SPORTHR}}$$

The odds of being myopic when the all the other variables are equal to zero, is

$e^{-0.54}$ =0.58 .

One D increase in spherical equivalent refraction, brings a decrease of 3.83 in the log odds of having myopia; equivalently, the odds ratio decreases by $e^{-3.83}$= 0.02 .

One hour per week increase in time spending in sports, brings a decrease of 0.05 in the log odds of having myopia; equivalently, the odds ratio decreases by $e^{-0.05}$ = 0.95.

If one of both parents has myopia, then the log odds of his children to have myopia is $e^{+1.34}$ = 3.80.

Conclusion

According to our model , the existence of myopia depends on the D value of spherical equivalent refraction, the time spending in sports and the heredity. Specifically, low prices in spherical equivalent refraction during childhood can be an evidence of suffering from myopia in the long run. Time spending doing sports is a good way of reducing the probability of developing myopia. Last but not least, if one or both parents have myopia it is very possible that the child will have myopia as well.