**MSDS 6372 Project 3**

**Team members:**

**(Scott Anderwald, Mavin Scott, Olufemi Adesanya)**

**MYOPIC STUDENT RESEARCH**

**Introduction**

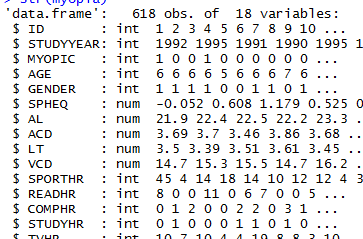
The research is about classifying myopic and non-myopic student. The thought of needing eye glasses came with the frightful thought of the constant kidding that would come from my classmates. The condition I came to deal with is what is called myopia which is better known as near sightedness. This condition causes one to have blurry vision at long distances and the need for corrective lenses. The data for the study was provided by Hosmer, D.W. Lemeshow, S. and Sturdivant, R.X. thru the University of Mass and is copyrighted by John Wiley and Sons publishers.

Myopia is a condition that generally occurs when the structure of the eyes is elongated and the focus point of the eye is shifted in front of the retina. The myopia study was of course where either the subject had the condition or not. This type of data requires the use of logistic regression and its ability to handle a binary variable.

**Descriptive Statistics about the Data**

Our dataset include 18 variables and 618 observations. The continuous variables in our dataset include: id, studyyear, age, spheq, al, acd, lt, vcd, sporthr, readhr, comphr, studyhr, tvhr, diopterh while the categorical variables include: myopic, gender, mommy, and daddy. Fig 1.0 below shows the output

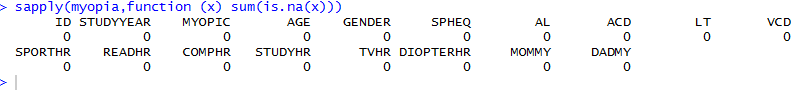
Fig 1.0 Myopia Dataset Summary



**Exploratory Data Analysis**

Based on the nature of our dependent variables which are categorical, we decided to utilize the logistic regression method to carry out our analysis. Logistic regression is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables. For this study, we wanted to see which variables would best determine the fit for determining which subjects would get myopia. The data set contains 618 observations and 18 variables. Certain variables were removed since their inputs did not increase the efficiency of the fit. Four variable were used for both the train and test subsets. The first step was to determine the number of missing variables. As the output in fig 1.2 below shows, no variable has a missing value.

Fig 1.2 below shows the number of variables with missing values



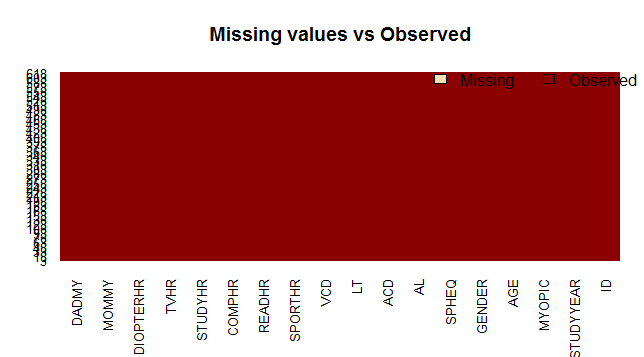
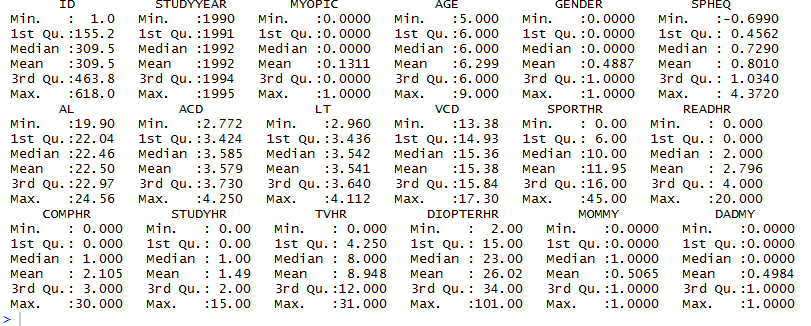


Fig 1.3 below shows the descriptive statistics for each variable

Fig 1.3 descriptive statistics



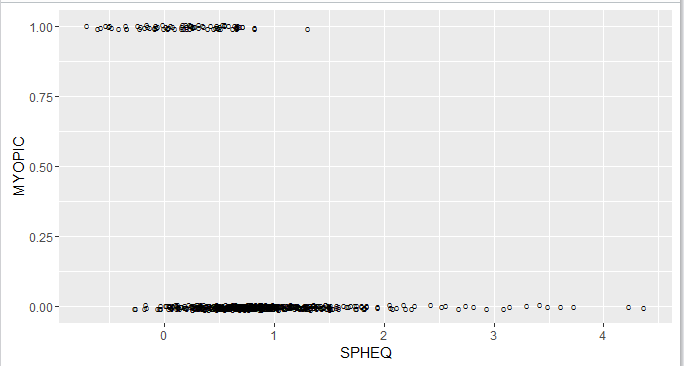
Since no observations appeared to be missing the data analysis continued with the determination of what is the clinical definition of myopic (myopia) which is when the spherical equivalent refraction (SPHEQ) is <= -75 diopter. For the analysis, MYOPIC versus SPHEQ see figure 1.4 below

Fig1.4

|  |  |
| --- | --- |
|  |  |

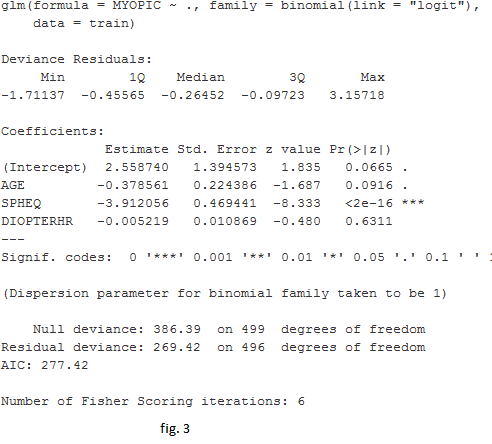
We then did a glm procedure using the MYOPIC and SPHEQ variables. Fig 1.5 below shows the outpu

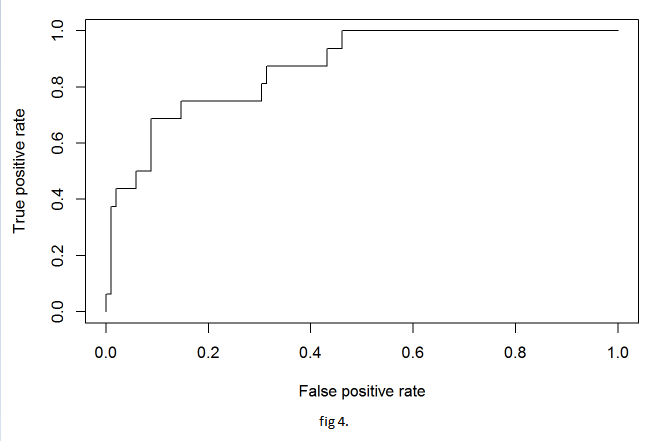
Fig 1.5



One can determine from the graph myopia does occur when the diopter reading for SPHEQ is less than or equal to -75.

Along with SPHEQ, the variables of age and a composite of near work activities defined as DIOPTERHR were inputted into the glm function of R. see figure 1.5 below. From the glm model it appears that variable of SPHEQ was the highest significance of determining whether subjects develop myopia. Also, noted from the study the near work amount has the least impact on the subjects.

Fig 1.5

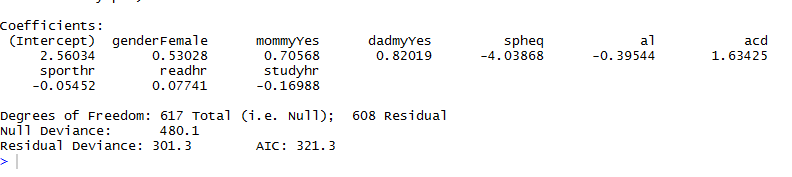
  
To determine where or not the model is fitted one would like to see how the model is does with a new data set. For this, we plot True positive rate versus false positive rates. Along with the plot we can calculate the accuracy of the model with the predict function of R for this study the model accuracy determination returned a value of 0.889830508 which is a good fit for the test set.

**Fitting the Model**

**Model1**

We decide to search for our best model for this dataset in order to get answers to the main effect of Myopia. Fig 1.6 below shows our first model which include multiple predictors ( 9 in total). Fig 1.6 shows the coefficients for intercept and the multiple predictors which is the slope for the multiple predictors.

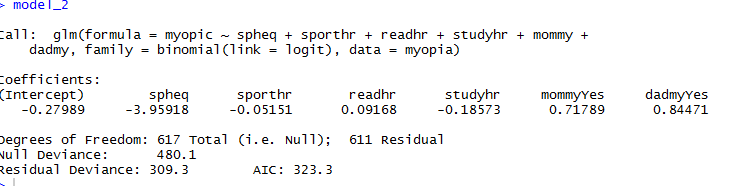
Fig 1.6



**Model2**

As we can see in the fig 1.7 below, our Model 2 has a higher residual deviance compare to our model 1, also the AIC for model 2 is higher than model1. Unlike our model1, only six predictor variables were utilized.

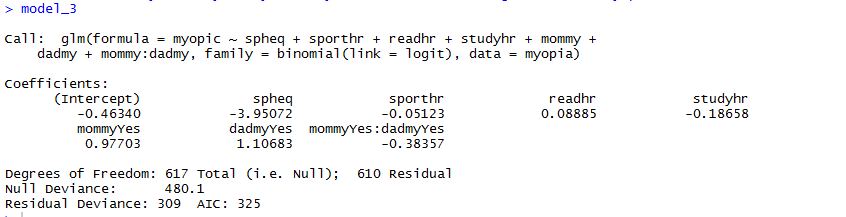
Fig 1.7



**Model3**

Our model3 include a combination of our model 2 with a mommy:daddmy variables interaction. Fig 1.8 below shows the coefficient output for our model 3. As seen in the output below, our model 3’s AIC is higher than our previous models.

Fig 1.8

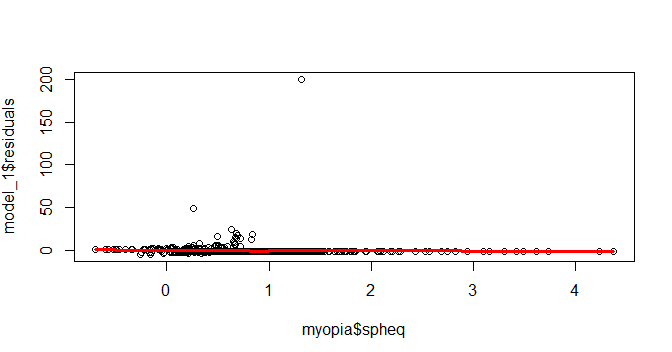


**Residual Plots**

**Model1 vs Spheq**

We did a residual plot of our model1 against the spheq variable in our myopia dataset. Fig 1.10 below shows the output. Also the below output shows the standard error for the Spheq Variable.

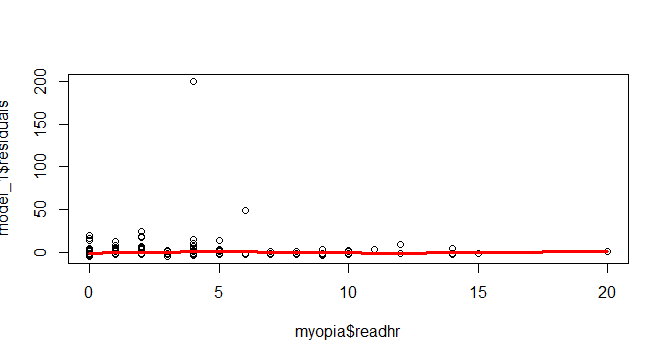
Fig 1.10 Model1 vs Spheq Plot





**Model1 VS Readhr**

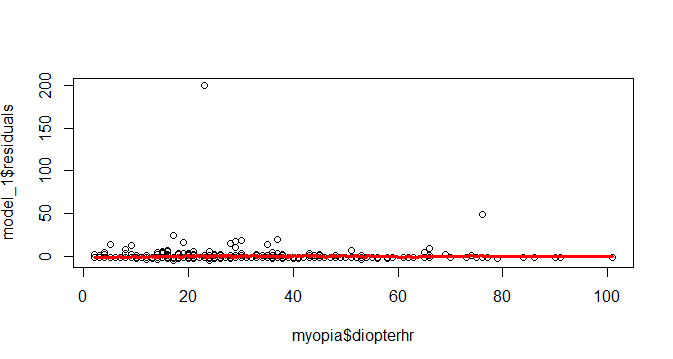
Fig 1.11 below shows our model1 against the readhr variable in our myopia dataset. Here we use the “READHR” variable to plot against out model1.





**Model1 VS Diopterhr**

Fig 1.12 below shows the plot using the “Diopterhr variables against the model1. The output below shows the result and the residual standard error.



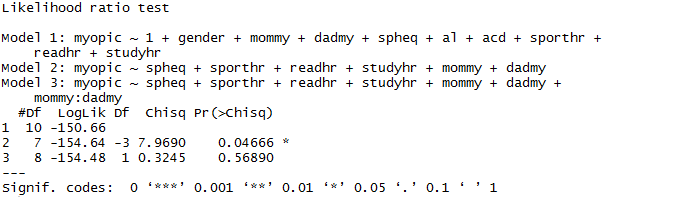


**Models Comparison/ Conclusion 2**

Looking at the output of our three models, we considered the Residual deviance, AIC to determine which among the three models is the best for our analysis. With logistic regression, instead of R squared as the statistic for overall fir of our model, we have deviance instead. The bigger the difference or deviance of the observed values from the expected values, the poorer the fit of the model so we want the smallest deviance if possible. As we discovered, adding more predictor variables to our equation made the deviance gets smaller, indicating an improvement in fit.

Based on our Likelihood ratio test, our model 3 seems to be the best model for our research because of the high chi-square and small deviance numbers. From the glm analysis is appears that the diopter of the eye has the biggest impact in determining if a subject will have myopia. With age and near work activities having a less of a significant effect on subject developing myopia. With the test set an accuracy of 0.889830508 was obtained which signifies a good fit for the model. Fig 1.9 below shows the chi-square output:

Fig 1.9



**R-CODE**

myopia <- read.csv(file ="myopia.csv",head= TRUE, na.strings = "c") # Bring in data#

summary(myopia) #data summary#

nrow(myopia) #to find number of rows#

sapply(myopia,function (x) sum(is.na(x))) #to check for missing values#

library (Amelia)

missmap(myopia, main = "Missing values vs Observed")

library(missmap)

install.packages("missmap")

install.packages("Amelia")

logdATA <- glm(MYOPIC ~ SPHEQ, family = binomial(link="logit"), data = myopia) #lOGISTIC rEGRESSION#

library(ggplot2)

ggplot(myopia, aes(x = SPHEQ, y = MYOPIC)) + geom\_jitter(shape = "O", position = position\_jitter(height = .02))

+ stat\_smooth(method="glm",se= FALSE)

required(gmlmulit)

model\_1 <- glmulti(myopic~., data=myopia, fitfunction="glm", level=1, method="h", crit="aicc", family=binomial(link="logit"))

melist=weightable(search.me)

head(melist)

Searching for interactions (subset of previous M.E. model)

search.int = glmulti(myopic ~ spheq + sporthr + readhr + studyhr + mommy + dadmy + mommy:dadmy, family=binomial(link=logit), data=myopia, fitfunction="glm", level=2, method="h", crit="aicc")

intlist=weightable(search.int)

head(intlist)

library(gmlmulit)

install.packages("glmulti")

model\_1 <-glm(formula = myopic ~ 1 + gender + mommy + dadmy + spheq + al +

acd + sporthr + readhr + studyhr, family = binomial(link = logit),

data = myopia)

library(aplore3)

attach(myopia)

library(rJava) #required library

library(glmulti)

install.packages("rJava")

model\_2<- glm(formula = myopic ~ spheq + sporthr + readhr + studyhr + mommy +

dadmy, family = binomial(link = logit), data = myopia)

model\_3 <- glm(formula = myopic ~ spheq + sporthr + readhr + studyhr + mommy +

dadmy + mommy:dadmy, family = binomial(link = logit), data = myopia)

AICc=function(object){

n=length(object$y)

r=length(object$coef)

AICc=AIC(object)+2\*r\*(r+1)/(n-r-1)

list(AIC=AIC(object), AICc=AICc, BIC=BIC(object))

}

require(lmtest)

ltest<- lrtest(model\_1,model\_2,model\_3)

install.packages("zoo")

plotmodel1 <- loess(model\_1$residuals ~ myopia$spheq,data=myopia)

plot(myopia$spheq, model\_1$residuals)

j <- order(myopia$spheq)

lines(myopia$spheq[j],plotmodel1$fitted[j],col="red",lwd=3)

plotmodel1\_readhr <- loess(model\_1$residuals ~ myopia$readhr,data=myopia)

plot(myopia$readhr, model\_1$residuals)

l <- order(myopia$readhr)

lines(myopia$readhr[l],plotmodel1\_readhr$fitted[l],col="red",lwd=3)

plotmodel1\_diopterhr <- loess(model\_1$residuals ~ myopia$diopterhr,data=myopia)

plot(myopia$diopterhr, model\_1$residuals)

l <- order(myopia$diopterhr)

lines(myopia$diopterhr[l],plotmodel1\_readhr$fitted[l],col="red",lwd=3)