Concussion Management Study for Miami Athletes

PJ Mu
Blake Ballard

26th, April 2018

Summary

A concussion is defined as a complex pathophysiological process affecting the brain, induced by biomechanical forces, or more simply a temporary unconsciousness caused by a blow to the head. Research has shown that the best way to detect a concussion and to oversee the progress of recovery is through neurocognitive testing and symptoms reported by the individual who was injured. Some of the symptoms include: headache, fatigue, reduced coordination/balance problems, irritability, hypersomnia or insomnia, dizziness, confusion, nausea, vomiting, sensitivity to light, sensitivity to noise, etc. Concussions are a very sensitive subject in sports today and can have very negative long lasting effects on an athletes health.

The purpose of this analysis is to investigate any differences in recovery time based on an athletes concussion history and to determine whether there is a significant difference between those who have not had previous concussions and those who have had more than one concussion in their lifetime. This study will also point out some specific diagnostic measures that are more reactive to detecting neurocognitive change after a concussion occurs. Lastly, this study will point out some unique characteristics about the individuals in this concussion database who have sustained multiple concussions throughout their collegiate career.

Introduction

The Speech Pathology and Audiology department at Miami University developed a Concussion Management Program in 1999. This program has been collecting baseline and post-concussion symptom and neurocognitive data from 1999 up until today. The neurocognitive information was evaluated through paper based testing up until 2005, since then an ImPACT computer based testing method has been implemented. The data contains information such as concussion and medical history, personal information, testing results, symptom ratings, and the number of days since the injury occurred. The database includes 242 athletes seen for testing and these athletes make up 266 total concussions. Although there are only 266 concussions in the database, each row represents each time an athlete visits for a check up following a concussion. Because of this, there are actually 830 rows of data in the database. This means that on average an athlete will visit 3.12 times following a concussion before being cleared. An important objective of this

study is discover some statistically significant factors that affect the amount of time it takes an athlete to recover from a sustained concussion.

The predictors used to do the research consist of each athlete's: visual memory composite scores, verbal memory composite scores, visual motor speed scores, reaction time, impulse control scores, symptom scores, dominant hand Grooved Pegboard scores, non-dominant hand Grooved Pegboard scores, the number of concussions in the athlete's lifetime, gender, age, height, weight, year of education, and sport played. The response variables used include the number of days it took the athlete to return to neurocognitive baseline scores following a concussion and whether an athlete had multiple concussions.

Some athletes never returned to neurocognitive baseline, but were cleared to play by doctors; these athletes were removed from the analysis because the major interest of the study involved recovery patterns in neurocognitive testing.

Research Questions

This paper focuses on three main tasks:

1. Is there a difference in recovery patterns for those athletes with a history of multiple concussions?
2. Is there a diagnostic measure that is more sensitive in detecting neurocognitive change following a suspected concussion?
3. During varsity athletes' tenure at Miami University, how many individuals sustained multiple concussions? Is there anything unique about this population?

Methods and Analysis

1. Preliminary Analysis

Before any analysis could be done, it was important that the data be cleaned up and mapped in a way that allows efficient work to be done. To answer our questions of interest regarding recovery time and recovery patterns, it became much more efficient to clean the data down to only each athlete's last visit for each concussion that they sustained.

The objective of the preliminary analysis was to create graphical displays that represent how recovery time, in days, is affected by the number of concussions an athlete has endured. Figure 1 below, represents the time taken to recover from a concussion based on the number of concussions an athlete has endured in his or her lifetime; whereas figure 2, represents the time taken to recover from a concussion based only on the number of concussions an athlete has endured in his or her collegiate career. Other factors that may have an impact on causing an athlete to suffer from multiple concussions during their time at Miami is analyzed more in depth in the full scale analysis.

2. Developmental Methodology

Task 1:

The first task of the analysis is to investigate the difference in recovery time for athletes who have a history of multiple concussions. To explore this the Concussion Management Database was cleaned down to only each subjects code, current concussion total, and time from injury for their last visit of each concussion sustained at Miami University. This means that if a subject had three concussions at Miami, they would account for exactly three rows of data. Because a subject can attribute to multiple concussions in the database, the analysis had to account for some error caused by dependence within subjects. To account for the dependence, a mixed effects model was used to fit the data with the log of time from injury as the response, current concussion total as a predictor, and the subject as a random effect. The log of time from injury must be used as the response because its values are skewed to the right, meaning in most cases the athlete could return to play in a short amount of time, but there are some cases that take much longer.

Results:

Once the database was cleaned down and the normality and independence assumptions could be met, a model could be fit and tested. An ANOVA F-test and a 95 percent confidence interval was used to determine whether the average number of days to recover from a sustained concussion differs for the athletes who have a history of concussions compared to those who have no history of concussions. The F-test resulted with a test statistic equal to 6.51 and a p-value of .0114 meaning it can be concluded that the mean number of days it takes for an athlete with a history of concussions to recover differs from the mean number of days it takes for an athlete with no history of concussions. To follow this up, a 95 percent confidence interval was implemented. The resulting interval discovered that one can be 95 percent confident that it takes an athlete who has sustained multiple concussions in their lifetime an average of 1.07 to 1.71 more days to recover than an athlete who has no history of previous concussions.

Task 2:

The second task is to determine which diagnostic measurement is most sensitive to detect neurocognitive change. There are two main categories of diagnostic testing: *Impact* and Grooved Pegboard. The *Impact* measurement includes 6 different factors including: visual memory composite, verbal memory composite, visual motor speed, reaction time, impulse control, and symptom scores. With dominant Grooved Pegboard and non-dominant Grooved Pegboard, there are the 8 potential measurements in total that are sensitive in detecting neurocognitive change.

Neurocognitive change, in the study of concussions, can be measured by the number of days it takes athletes to return to their neurocognitive baseline scores. This factor is used as a response variable and the seven measurements from the Grooved Pegboard and *Impact* test are used as predictors to build a multiple linear regression model.

To obtain, the response variance, the total number of days until an athlete returned to neurocognitive testing baseline, only the last visit of each concussion is necessary, which has the accumulated days that deviated from baseline. In addition, the dataset is divided by gender and would be further analyzed in these two different groups.

Model Fitting:
A multiple regression is built with all seven variables from *Impact* and Grooved Pegboard to predict the time that it takes athletes' neurocognitive level to return to baseline. the relaimpo package (Ulrike Grömping, 2006) is used to build a relative importance matrix. Here, the matrix used is *lmg*, which can demonstrate partitioned R^2 of each predictors by averaging over orders.The variable with the highest score in *lmg* matrix means that it is the most important measure in explaining neurocognitive change. The same procedure is conducted to analyze the concussion dataset of both male and female athletes.

Table 1. Relative importance metrics for diagnostic measurements (Overall)

Relative Importance Metrics	
Significant Predictors	R^2
Visual motor speed	0.47
Visual memory composite	0.27
Grooved pegboard (dominant)	0.14
Impulse control	0.05
Grooved Pegboard (nondominant)	0.03
Reaction Time	0.02
Verbal memory composite	0.01
Total symptom score	0.01

Figure 1. Relative importance for diagnostic measurements (Overall)

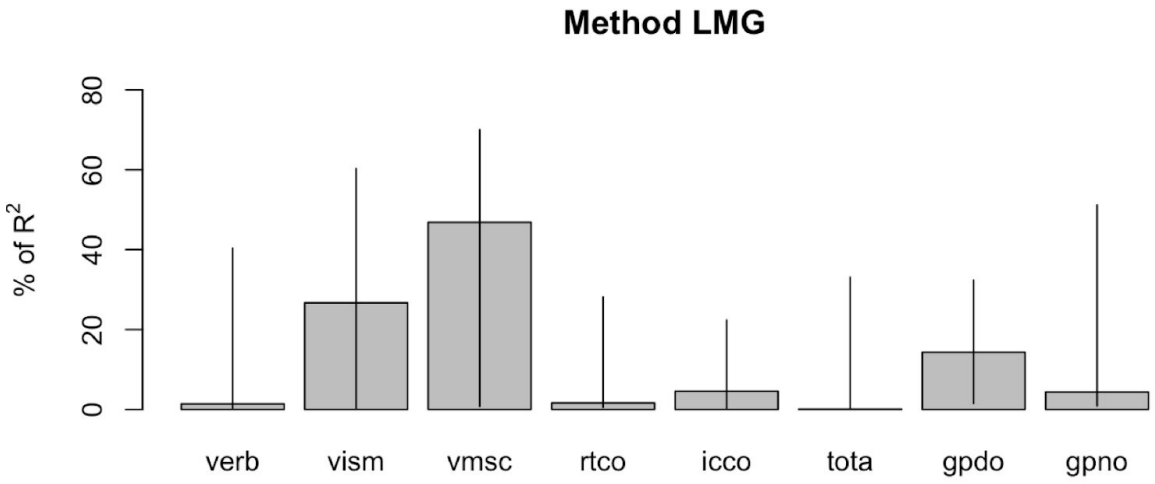


Table 2. Relative importance metrics for diagnostic measurements (Male)

Relative Importance Metrics	
Significant Predictors	R^2
Visual memory composite	0.62
Grooved pegboard (dominant)	0.19
Visual motor speed	0.07
Impulse control	0.06
Grooved pegboard (non-dominant)	0.03
Verbal memory composite	0.01
Total symptom score	0.01
Reaction Time	0.01

Figure 2. Relative importance for diagnostic measurements (Male)

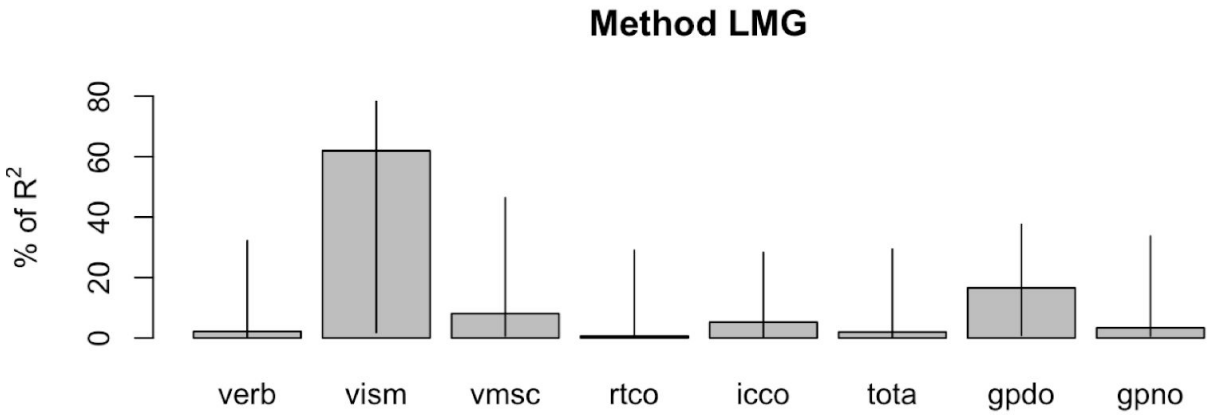
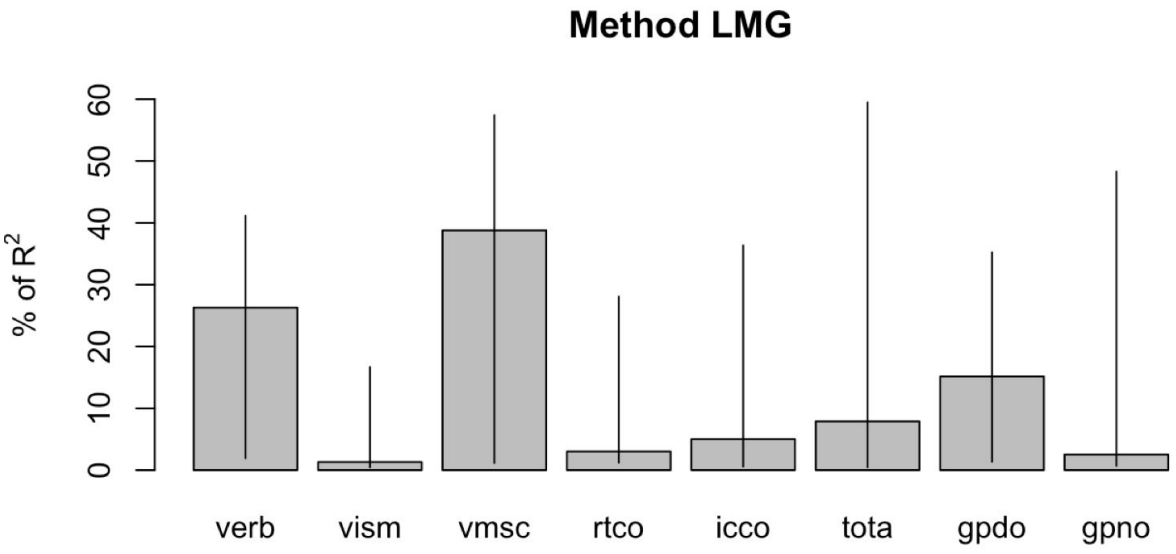


Table 3. Relative importance metrics for diagnostic measurements (Female)

Relative Importance Metrics	
Significant Predictors	R^2
Visual motor speed	0.39
Verbal memory composite	0.27
Grooved pegboard (dominant)	0.14
Total symptom score	0.07
Impulse control	0.06
Reaction time	0.04
Grooved Pegboard (non-dominant)	0.02
Visual memory composite	0.01

Figure 3. Relative importance metrics for diagnostic measurements (Female)



Model Result:

The whole data is subset into female athletes and male athletes because there may be neurocognitive differences in these two groups. For male athletes, Visual Memory Composite is the most sensitive measurement. For female athletes, Visual motor speed is the most sensitive one. For all athletes, visual motor speed is the most sensitive measurement in detecting neurocognitive change.

Task 3:

For task 3, an indicator of whether of athlete has sustained multiple concussions while at Miami University was used as the response variable. This was a perfect case for fitting a logistic regression model to predict the relationship between each individual who has sustained multiple concussions.

Model fitting:

Using R to fit a logistic model based on the input variables and response shown below is the starting point. Then, the permanence of the model needs to be tested by using validation testing, which is a method that splits the data into two groups (training data and test data). The training data is used to build the model. After a model is fit to the training data, the model is used to predict each of the cases in the testing dataset. The performance of the model is determined based on a misclassification error rate; the lower misclassification error the better the model.

Model Form:

$$log(\frac{P(multiple\ concussions)}{1-P(multiple\ concussions)}) = \beta 0 + \beta 1(year\ of\ education) + \beta 2(I\ basketball) + \beta 3(I\ football)$$

Where:

- Year of Education = Number of years enrolled in Miami University for each athlete.
- I basketball = Basketball indicator, 1 if sport = basketball, 0 if not.
- I football = Football indicator, 1 if sport = football, 0 if not.
- β_1 = Effect on log odds of years of education.
- β_2 = Effect on log odds of multiple concussion, when changing from other sports to basketball.
- β_3 = Effect on log odds of multiple concussion, when changing from other sports to football.

Table:

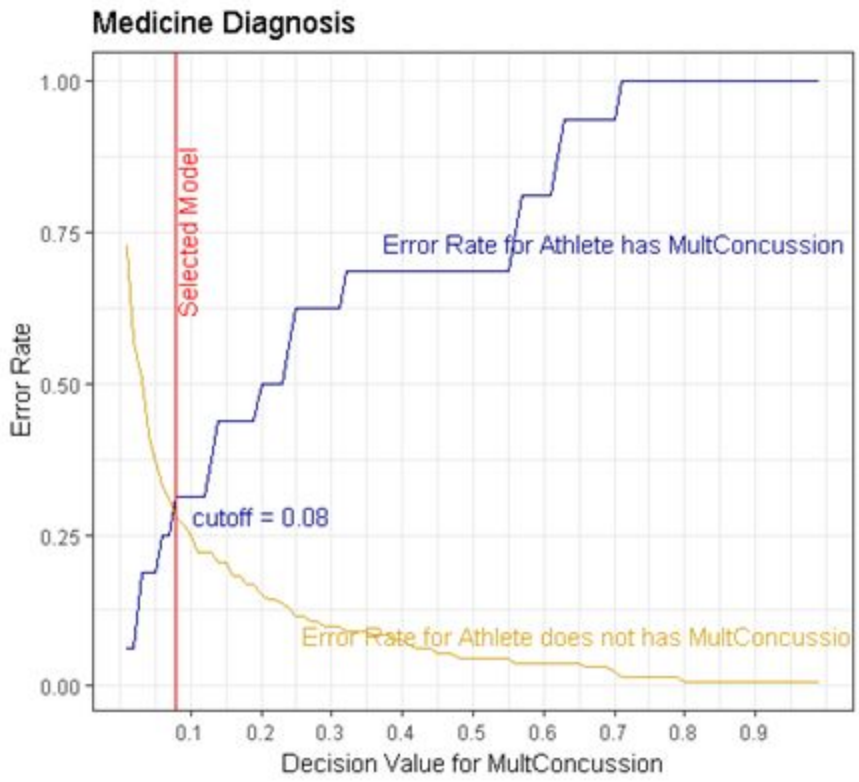
Predictors	Estimate	Standard Error	P-value	95% Log Odds Ratio Confidence Interval	95% Odds Ratio Confidence Interval
Intercept	-16.2925	4.2582	0.00013	(-25.49, -8.54)	(<0.001, <0.01)
Years of Education	0.9660	0.2996	0.00126	(0.41, 1.60)	(1.51, 4.96)
Sports (basketball)	2.5894	0.8194	0.00158	(1.05, 4.34)	(2.87, 77.05)
Sports (football)	1.0881	0.7643	0.15454	(-0.36, 2.74)	(0.70, 15.46)

The table shows the results of the logistic model. It includes the predictors used in this model, the effect of log odds ratio for each predictor, the standard error, the p-value, and the 95% confidence interval. The p-value is used to test the significance of each predictor. In this case, the p-value for Sports (football) is 0.15454, which is greater than 0.05. This means that the variable should be considered only slightly significant. Also, the 95% confidence interval shows that one could be 95% confident that the estimate for each predictors are inside the range shown.

Cut Off Value Introduced:

0.5 is commonly used as the cutoff value, which is used to decide the outcome for logistic regression (if $p > 0.5$, then yes, otherwise, no). In other words, given the input into the model, an output of a probability "p" will be returned. This output has a range from 0 to 1; if $p > 0.5$, then this person has multiple concussions, otherwise, the athlete does not. In this study, since it is related to the health of athletes, people are concerned about the accuracy of the true positive rate error and false negative error rate of the model. Therefore, there are many different cutoff values that could be used for the outcome of the model. In this case, trying all the values and selecting the best cut off value for the model is necessary.

Figure 4: Medicine Diagnosis (Finding Cut Off Value):



In the health field, this method can be used to optimize the decision threshold for medicine diagnosis. By looking at the plot above, the x-axis is the decision value for multiple concussions, and the y-axis is the misclassification error rate for the two different groups. One is the error rate for an athlete who has multiple concussions, the other is the error rate for an athlete who does not have multiple concussions. The intersection of the two lines in figure 4 show the decision threshold for multiple concussions is 0.08, which means that if the output from the model returns a probability greater than 0.08, then this person should be considered as having multiple concussions, otherwise, they should be not.

Model Performance:

Table 3:

Confusion Matrix

	Predicted Number of Athlete	
Actual Number	**No Multiple Concussion**	**Multiple Concussion**
No Multiple Concussion	**99**	**33**
Multiple Concussion	**4**	**12**

The accuracy of this logistic regression model can be tested based by the confusion matrix, which is a table used to describe the assessment of classification. In this case, it can be assessed for a logistic regression model. It compares the outcome from the model to the actual value in the data. The model accuracy can be computed using 1 – the misclassification error rate.

Model Accuracy: $1 - \frac{33+4}{99+33+4+12} = 76\%$

Conclusion Result

In conclusion, based on the model information, 76% confidence for the model accuracy has been proved based on testing. Therefore, the model showed that the year of education, sport of basketball and football are the final variables should be chosen. Based on the model, after the variables input, if the outcome from the model is greater than 0.08, the athlete should be considered multiple concussion, otherwise, it is not a multiple concussion.

General Discussion

The conclusion that as more concussions are sustained the longer it takes for an athlete to recover seems to be consistent with what is believed by many concussion specialists. One thing that led to problems in the analysis was the fact that there were so few a number of athletes who have had multiple concussions while at Miami in this database. This made it very difficult to use statistical tools to try and describe the type of athlete who is most vulnerable to sustaining multiple concussions while in college. This problem will go away as more data is collected and more cases of multiple concussions while playing sports at Miami are observed.

Reference

Bates, D., Mächler, M., & Bolker, B. M. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,67*(1). 1-48. doi:10.18637/jss.v067.i01

How To Find Relationship Between Variables, Multiple Regression. (n.d.). Retrieved April 13, 2018, from http://www.statsoft.com/Textbook/Multiple-Regression

Shaikh, F., Dar, P., Srivastava, T., & Analytics Vidhya Content Team. (2016, July 08). Simple Guide to Logistic Regression in R. Retrieved April 13, 2018, from https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <URL: https://www.R-project.org>.

Sachs, M. C. (2017, August 09). Generate ROC Curve Charts for Print and Interactive Use. Retrieved April 13, 2018, from https://cran.r-project.org/web/packages/plotROC/vignettes/examples.html

Saldanha, R. A., & Hjorth, J. S. (1995). Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap. *The Statistician,44*(2), 288. doi:10.2307/2348459

Townsend, J. T. (1971). Erratum to: Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics,10*(4), 256-256. doi:10.3758/bf03212817

Will. "FORMULAE IN R: ANOVA AND OTHER MODELS, MIXED AND FIXED."*Conjugateprior*, May 2013, conjugateprior.org/2013/01/formulae-in-r-anova/.

Appendix

	Numerator df	Denominator df	Sum Squared	Mean Squared	F Value	P-Value
Concussion Total	**1**	**215**	**2.40**	**2.40**	**6.512**	**.0114**

```
##################
library(haven)
data <- read_sav("C:/Users/Blake/Downloads/NEW Concussion
Database 2017.sav")
new <- data[,c(2,3,5,6,7,8,13,29,30,33)]
data1 <- data[,c(1,2,3,33,46,150,151,152,153,154)]
attach(data1)
names(data1)
length(daysuntilsymptomfree[daysuntilsymptomfree>0])
data2 <- data1[daysuntilsymptomfree>0,]
data3 <- data1[daysuntilcleared>0,]
data4 <- data1[daysuntilneurocogfree>0,]
data3 <- data3[-c(27,81,239:274),]
length(unique(data1$Code))
length(unique(data3$Code))
data5 <- data3[data3$CurrentConcussionTotal>0,]

##Task 1
library(lme4)

##Normalize
shapiro.test(log(data5$tfi))

##Break into one concussion (1) vs. multiple concusion (2)
data5$CurrentConcussionTotal <-
ifelse(data5$CurrentConcussionTotal > 1, 2, 1)
data5$CurrentConcussionTotal <-
as.factor(data5$CurrentConcussionTotal)
```

```
#get rid of 1 extreme outlier
data5 <- data5[data5$tfi < 300,]

##mixed effects model and anova test
mix <- lmer(log(tfi) ~ factor(CurrentConcussionTotal) +
(1|Code), data=data5)
anova(mix)
summary(mix)

##create confidence interval
#interval with unlogged days
exp(confint(mix))

##Task 2
library(leaps)
library(MASS)
library(car)
library(tidyverse)
library(relaimpo)
library(dplyr)
library(broom)

library(haven)
Concussion2017 <- read_sav("~/Desktop/STA
475/Concussion2017.sav")



clean <- Concussion2017
%>%select(Gender,daysuntilneurocogfree,daysuntilcleared,verbmc,v
ismc,vmscomp,rtcomp,iccomp,totalsymtscore,gpdom,gpnondom)

total<- clean%>% filter(daysuntilcleared > 0)
total<-na.omit(total)
total<-total[total$daysuntilneurocogfree!=-99,]
total<-total[total$gpdom !=-99,]

male<-total[total$Gender==1,]
```

```
female<-total[total$Gender==2,]

# backward stepwise for TOTAL

backreg <-
regsubsets(daysuntilneurocogfree~.-Gender-daysuntilcleared,data
= total,
                        nvmax=8, method="backward")


# capture all input subsets for forward selection into list for
use in CV
inputlists.back <- list(NULL)
for(p in 1:8){
  inputlists.back[[p]] <- names(coef( backreg, id=p))[-1]
}


# fit backward selected models (with LOOCV)
cvmse.back <- rep(NA,8)
for(p in 1:8){
  preds <- rep(NA, nrow(total))
  for( i in 1:nrow(total)){
    datasub <- total[-i,
c("daysuntilneurocogfree",inputlists.back[[p]]) ]
    submod <- lm(daysuntilneurocogfree ~ . , datasub)
    preds[i] <- predict(submod, total[i, ])
  }
  cvmse.back[p] <- mean((preds-total$daysuntilneurocogfree)^2)
}

# choose model with lowest MSE with its variable list
nvars.back <- which.min(cvmse.back)
c("daysuntilneurocogfree",inputlists.back[[nvars.back]])

cvmse.back[nvars.back]
# 211.6502
```

```
total.model<-lm(daysuntilneurocogfree~ vismc+ vmscomp, data =
total)

calc.relimp(total.model,type=c("lmg"),
            rela=TRUE)

boot <- boot.relimp(total.model , b = 500, type = c("lmg"), rank
= TRUE,
                    diff = TRUE, rela = TRUE)

plot(booteval.relimp(boot,sort=F))


# backward stepwise for MALE

backreg <-
regsubsets(daysuntilneurocogfree~.-Gender-daysuntilcleared,data
= male,
                      nvmax=8, method="backward")


# capture all input subsets for forward selection into list for
use in CV
inputlists.back <- list(NULL)
for(p in 1:8){
  inputlists.back[[p]] <- names(coef( backreg, id=p))[-1]
}


# fit backward selected models (with LOOCV)
cvmse.back <- rep(NA,8)
for(p in 1:8){
  preds <- rep(NA, nrow(male))
  for( i in 1:nrow(male)){
    datasub <- male[-i,
c("daysuntilneurocogfree",inputlists.back[[p]]) ]
    submod <- lm(daysuntilneurocogfree ~ . , datasub)
    preds[i] <- predict(submod, male[i, ])
  }
```

```
  cvmse.back[p] <- mean((preds-male$daysuntilneurocogfree)^2)
}

# choose model with lowest MSE with its variable list
nvars.back <- which.min(cvmse.back)
c("daysuntilneurocogfree",inputlists.back[[nvars.back]])

cvmse.back[nvars.back]
# 204.0003


# build the model
male.model<-lm(daysuntilneurocogfree~ vismc+ gpdom, data = male)

calc.relimp(male.model,type=c("lmg"),
            rela=TRUE)

boot <- boot.relimp(male.model , b = 500, type = c("lmg"), rank
= TRUE,
                    diff = TRUE, rela = TRUE)

plot(booteval.relimp(boot,sort=F))


# backward stepwise for FEMALE

backreg <-
regsubsets(daysuntilneurocogfree~.-Gender-daysuntilcleared,data
= female,
                      nvmax=8, method="backward")


# capture all input subsets for forward selection into list for
use in CV
inputlists.back <- list(NULL)
for(p in 1:8){
  inputlists.back[[p]] <- names(coef( backreg, id=p))[-1]
}
```

```
# fit backward selected models (with LOOCV)
cvmse.back <- rep(NA,8)
for(p in 1:8){
  preds <- rep(NA, nrow(female))
  for( i in 1:nrow(female)){
    datasub <- female[-i,
c("daysuntilneurocogfree",inputlists.back[[p]]) ]
    submod <- lm(daysuntilneurocogfree ~ . , datasub)
    preds[i] <- predict(submod, female[i, ])
  }
  cvmse.back[p] <- mean((preds-female$daysuntilneurocogfree)^2)
}

# choose model with lowest MSE with its variable list
nvars.back <- which.min(cvmse.back)
c("daysuntilneurocogfree",inputlists.back[[nvars.back]])

cvmse.back[nvars.back]
# 204.0003


# build the model
female.model<-lm(daysuntilneurocogfree~ verbmc+ gpdom + vmscomp,
data = female)

calc.relimp(female.model,type=c("lmg"),
            rela=TRUE)

boot <- boot.relimp(female.model , b = 500, type = c("lmg"),
rank = TRUE,
                    diff = TRUE, rela = TRUE)

plot(booteval.relimp(boot,sort=F))
```

```
##logistic regression model
#data import
install.packages("readxl")
install.packages("dplyr")
library(dplyr)
library(readxl)
concussion <- read_excel("logistic.xlsx")

#data clean

#variable selection
concussion <- concussion[,3:11]
str(concussion)

#remove the missing data
concussion <- na.omit(concussion)

#check str
str(concussion)
unique(concussion$Sport)

#code sports not in 1 5 8 to be basketball
for(i in 1:length(concussion$Sport)){
  if(!(concussion$Sport[i] %in% c(1,5,8))){
    concussion$Sport[i] = 0
  }
}

concussion$Sport <- as.character(concussion$Sport)
str(concussion$Sport)
for(i in 1:length(concussion$Sport)){
  if(concussion$Sport[i] == "1"){
    concussion$Sport[i]  = "football"
  }
  else if(concussion$Sport[i] == "5"){
    concussion$Sport[i]  = "hockey"
  }
  else if(concussion$Sport[i] == "8"){
    concussion$Sport[i]  = "basketball"
```

```
  }
  else{
    concussion$Sport[i]  = "others"
  }
}

#set column to factors
str(concussion)
concussion$Gender <- factor(concussion$Gender)
concussion$Sport <- factor(concussion$Sport)
concussion$Position <- factor(concussion$Position)
concussion$PrevConc <- as.numeric(concussion$PrevConc)
concussion$MultConcussion <- factor(concussion$MultConcussion)

concussion <- na.omit(concussion)


######################back to logistic
str(concussion)
concussion$binary <- ifelse(concussion$MultConcussion == "yes",
1, 0)

#generate the LOOCV
pred_prob <- c()

for(i in 1:nrow(concussion)){
  train <- concussion[-i,]
  test <- concussion[i,]
  log_mod <- glm(formula = binary~.-MultConcussion,
                 family=binomial(link=logit), data = train)
  pred_prob[i] <- ifelse(predict(log_mod, test, type =
"response"),
                                 1, 0)
}

for(i in 1:nrow(concussion)){
  train <- concussion[-i,]
  test <- concussion[i,]
  log_mod <- glm(formula = binary~.-MultConcussion,
```

```
                  family=binomial(link=logit), data = train)
  pred_prob[i] <- ifelse(predict(log_mod, test, type =
"response") > 0.08,
                         1, 0)
}
table(concussion$binary, pred_prob)


1 - mean(factor(concussion$binary)!= factor(pred_prob))

# Compare to each threshold between 0.01 and .99 by steps of .01
tune_eta <- data.frame(eta = seq(0.01, 0.99, by = 0.01),
                       loocv_misclass = NA,
                       loocv_yes_misclass = NA,
                       loocv_no_misclass = NA)
for(eta in 1:nrow(tune_eta)){
  # assign class based on new threshold
  pred_class <- ifelse(pred_prob >= tune_eta$eta[eta], 1, 0)

  #overall error rate
  tune_eta$loocv_misclass[eta] <- 1 - mean(pred_class ==
concussion$binary)

  #calculate conditional error rates
  conf_mat <- as.matrix(table(pred_class, concussion$binary))
  nyes <- sum(concussion$binary == 1)
  nno <- sum(concussion$binary == 0)

  tune_eta$loocv_yes_misclass[eta] <- conf_mat[1,2]/nyes
  tune_eta$loocv_no_misclass[eta] <- conf_mat[2,1]/nno
}

difference <- abs(tune_eta$loocv_yes_misclass -
tune_eta$loocv_no_misclass)
which.min(difference)

selected_mod <- tune_eta[which.min(difference),]
selected_mod
```

```
install.packages("ggplot2")
library(ggplot2)
ggplot()+
  geom_line(aes(x=eta,y=loocv_yes_misclass),
            color="darkblue", data=tune_eta)+
  geom_line(aes(x=eta,y=loocv_no_misclass),
            color="goldenrod", data=tune_eta) +
  geom_vline(xintercept = selected_mod$eta, color="red") +
  theme_bw() +
  scale_x_continuous(breaks=seq(0.1,0.9,by=0.1))+
  annotate(geom="text",x=selected_mod$eta,
           y=.75,angle=90,label="Selected Model",
           vjust=1,color="red") +
  annotate(geom="text",x=selected_mod$eta,
           y=.75,angle=90,label="Selected Model",
           vjust=1,color="red")+
  annotate(geom="text",x=c(.7,.7),
           y=c(.75,.10),angle=c(0,0),
           label=c("Error Rate for Athlete has MultConcussion",
                   "Error Rate for Athlete does not has
MultConcussion Error"),
           vjust=1,color=c("darkblue","goldenrod"))+
  annotate(geom="text",x=0.2,
           y = 0.30, angle=0,
           label="cutoff = 0.08",
           vjust=1,color="darkblue")+
  theme_bw() +
  labs(x="Decision Value for MultConcussion", y="Error Rate") +
  ggtitle("Medicine Diagnosis")

#model form:
log_mod <- glm(formula = binary~.-MultConcussion,
               family=binomial(link=logit), data = concussion)
summary(log_mod)
```