HW 4

Matthew Fein

April 24, 2019

## Part one : Logistic Regression

require(tidyverse)  
require(expss)  
require(pROC)  
require(usdm)  
require(gdata)

data.df <- read\_csv("SpeedDating.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_integer(),  
## LikeM = col\_double(),  
## LikeF = col\_double(),  
## RaceM = col\_character(),  
## RaceF = col\_character(),  
## AttractiveM = col\_double(),  
## AttractiveF = col\_double(),  
## IntelligentM = col\_double(),  
## SharedInterestsM = col\_double(),  
## SharedInterestsF = col\_double()  
## )

## See spec(...) for full column specifications.

cro(data.df$DecisionM, data.df$DecisionF)

 data.df$DecisionF

 0

 1

 data.df$DecisionM

   0

66

64

   1

83

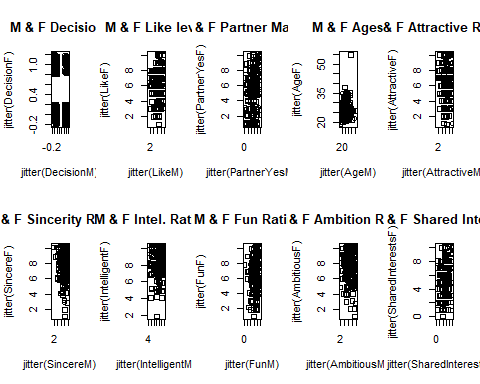
63

   #Total cases

149

127

data.df <- data.df %>% mutate(second.date = ifelse(data.df$DecisionM + data.df$DecisionF == 2, 1, 0))  
  
attach(data.df)  
par(mfrow= c(2,5))  
plot(jitter(DecisionM), jitter(DecisionF), main = "M & F Decisions", pch = second.date)  
plot(jitter(LikeM), jitter(LikeF), main = "M & F Like level", pch = second.date)  
plot(jitter(PartnerYesM), jitter(PartnerYesF), main = "M & F Partner Material", pch = second.date)  
plot(jitter(AgeM), jitter(AgeF), main = "M & F Ages", pch = second.date)  
plot(jitter(AttractiveM), jitter(AttractiveF), main = "M & F Attractive Ratings", pch = second.date)  
plot(jitter(SincereM), jitter(SincereF), main = "M & F Sincerity Ratings", pch = second.date)  
plot(jitter(IntelligentM), jitter(IntelligentF), main = "M & F Intel. Ratings", pch = second.date)  
plot(jitter(FunM), jitter(FunF), main = "M & F Fun Ratings", pch = second.date)  
plot(jitter(AmbitiousM), jitter(AmbitiousF), main = "M & F Ambition Ratings", pch = second.date)  
plot(jitter(SharedInterestsM), jitter(SharedInterestsF), main = "M & F Shared Interests", pch = second.date)



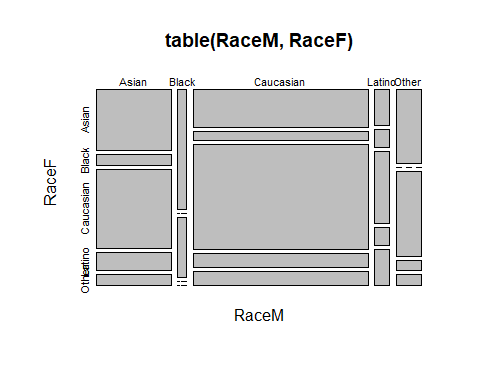
#dev.off()

Most variables appear uncorrelated except for age- this is also a positively skewed age distribution. There are more second dates for younger pairs. Fun ratings seem to be less important to woman than to men. Women also seem to be more willing to go on a second date with less attractive men than vice versa. In general the higher a couple is on both scales the more likely they are to dat again.3)Looking at the graphs no data apper to be out of range. There were missing values, they are:

summary(complete.cases(data.df))

1. The race categories are asian, Black, Caucasian, Hispanic and other. Missing values and mosaic plot:

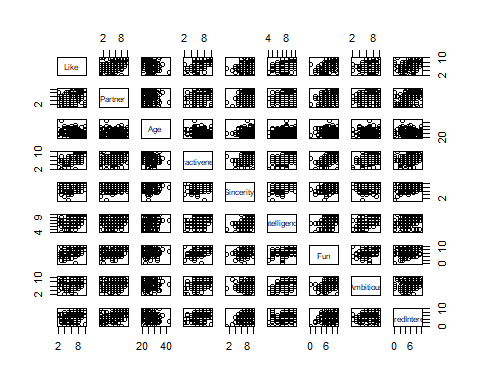
detach(data.df)  
  
data.trim <- data.df[complete.cases(data.df), ]  
  
  
attach(data.trim)  
  
table(RaceF)  
table(RaceM)  
as.factor(RaceF)  
as.factor(data.trim$RaceM)  
  
mosaicplot(table(RaceM, RaceF))



detach(data.trim)

5)Final model:

clean <- as.data.frame(read.csv("clean.csv"))  
clean.num <- as.data.frame(read.csv("cleancut.csv"))  
  
pairs(clean.num[ ,-1])



round(cor(clean.num), digits = 3)  
  
  
glm.1 <- glm(second.date ~ Like + Age + Partner, data=clean, family=binomial(link="logit"))  
summary(glm.1)

There is a strong positive correlation between like and attractiveness ratings, therefore, only like was used for all analyses. Each response was converted to a difference score for each pair. The female’s rating was subtracted from the Male’s rating meaning that positive values represent a higher rating by the male in the couple. VIF and G computation both support the validity of the model:

vif(as.data.frame(clean[,c("Like", "Age", "Partner")]))  
  
##log-likeliohood  
pchisq(summary(glm.1)$null.deviance - summary(glm.1)$deviance,   
 df=summary(glm.1)$df.null - summary(glm.1)$df.residual,  
 lower.tail=FALSE)

VIF’s are all close to 1 and the log likelihood test is statistically significant.6)

cro(data.trim$DecisionM, data.trim$DecisionF)

 data.trim$DecisionF

 0

 1

 data.trim$DecisionM

   0

48

43

   1

61

48

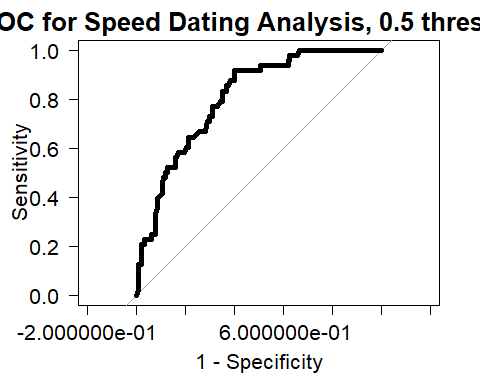
   #Total cases

109

91

The new sample size is 200 after removing any incomplete pairs of data. The model does follow the rule of thumb of at least 10 observations per possible outcome.7) Both like ratings and if one was seen as a potential partner have positive coefficients, which is an intuitive finding. Age shows a negative slope- while not significant being older does seem to put one at a disadvantage for a second date.8)

roc.figure1 <- roc(response=clean$second.date, predictor=glm.1$fitted.values, plot=TRUE, las=TRUE, legacy.axes=TRUE, lwd=5,  
 main="ROC for Speed Dating Analysis, 0.5 threshold", cex.main=1.6, cex.axis=1.3, cex.lab=1.3)



#dev.off()  
roc.info <- roc(response=clean$second.date, predictor=glm.1$fitted.values)  
# sensitivity and specificity for a specific threshold of 0.5  
coords(roc.info, x=0.5, input="threshold", ret=c("threshold", "sensitivity", "specificity"))  
coords(roc.info, x="best", input="threshold", ret=c("threshold", "sensitivity", "specificity"))  
# get AUC  
auc(response= clean$second.date, predictor=glm.1$fitted.values)

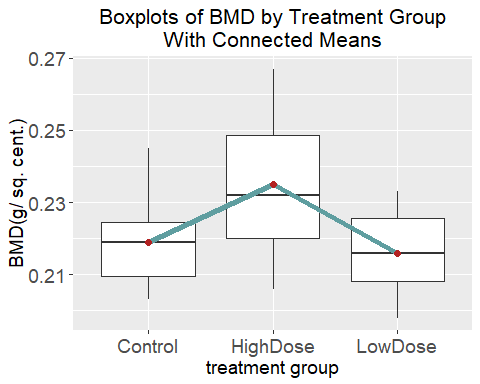
Threshild was left at .5. While the baseline rate of a second date was around 25% in the sample it does not seem to me that one has a better than random chance at getting a second date from a speed dating meetup.

## Part 2: One-Way ANOVA

1. Bone mineral density
2. Treatment level is the factor with 3 levels- control, low dose and high dose
3. There is one treatment- exposure to isoflavones
4. Randomized balance design13)Descriptives: Count, mean and standard deviations by group:

theme.info <- theme(plot.title = element\_text(size=16, hjust=0.5),  
 axis.title = element\_text(size=14),  
 axis.text = element\_text(size=14))  
  
kudzu <- read.csv("kudzu.csv")  
  
  
kudzu%>%   
 group\_by(Treatment) %>%  
 count()  
  
# compute mean for each treatment group  
aggregate(kudzu$BMD, by=list(kudzu$Treatment), mean)  
  
# compute standard deviation for each treatment group  
aggregate(kudzu$BMD, by=list(kudzu$Treatment), sd)

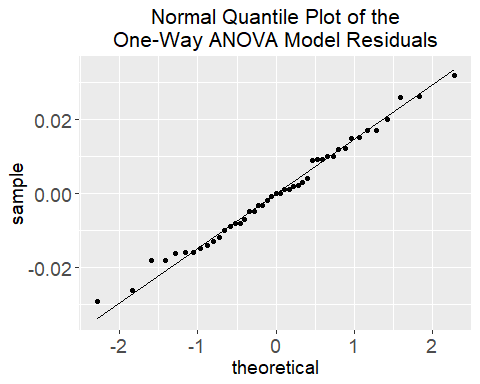
kudzu %>%   
 ggplot(aes(Treatment, BMD)) +   
 geom\_boxplot() +  
 stat\_summary(fun.y=mean, geom="line", aes(group=1), lwd=2, col="cadetblue") +  
 stat\_summary(fun.y=mean, geom="point", pch=19, size=2, col="firebrick") +  
 ggtitle("Boxplots of BMD by Treatment Group\nWith Connected Means") +  
 labs(x="treatment group",   
 y="BMD(g/ sq. cent.)") +  
 theme.info

 The high dose treatment group seems to have a higher mean tha nboth other groups.16) The slopes for all three treatment groups are not equivalent, F(2,42) = 7.72, p <.001.

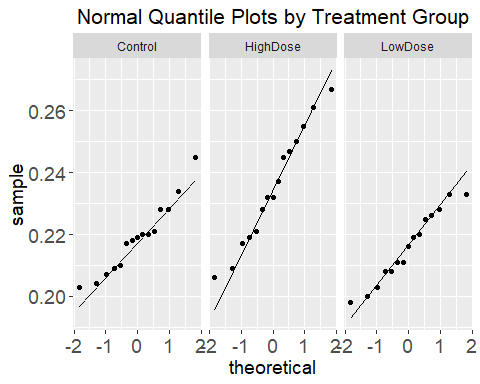
result <-aov(BMD ~ Treatment, data=kudzu)  
  
summary(result)

1. Assumptions are met as seen below :

# looking at residuals  
temp <- kudzu %>%  
 group\_by(Treatment) %>%  
 summarize(mean(BMD))  
left\_join(kudzu, temp) %>%  
 mutate(residuals = BMD - `mean(BMD)`) %>%  
 ggplot(aes(sample=residuals)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("Normal Quantile Plot of the\nOne-Way ANOVA Model Residuals") +  
 theme.info



# looking at response by treatment group  
kudzu%>%  
 ggplot(aes(sample=BMD)) +  
 facet\_grid(~ Treatment) +  
 stat\_qq() +   
 stat\_qq\_line() +  
 ggtitle("Normal Quantile Plots by Treatment Group") +  
 theme.info

 17) Group high dose is significantly different from the other two conditions:

TukeyHSD(result, conf.level=0.99)

## Tukey multiple comparisons of means  
## 99% family-wise confidence level  
##   
## Fit: aov(formula = BMD ~ Treatment, data = kudzu)  
##   
## $Treatment  
## diff lwr upr p adj  
## HighDose-Control 0.016200000 4.702669e-05 0.03235297 0.0097645  
## LowDose-Control -0.002933333 -1.908631e-02 0.01321964 0.8423308  
## LowDose-HighDose -0.019133333 -3.528631e-02 -0.00298036 0.0020537