SDGB 7840 Homework 4

Name: Jiayin Hu Class Time: 3:30-5:30pm

2019/4/25

Before the homework, we list the library we need first.

```
require(tidyverse)
require(knitr)
require(ggplot2)
require(gridExtra)
require(GGally)
require(readxl)
require(pROC)
require(usdm)
```

Logistic Regression

Read data.

```
date <- read_csv("SpeedDating.csv", col_names=TRUE)
```

Q1

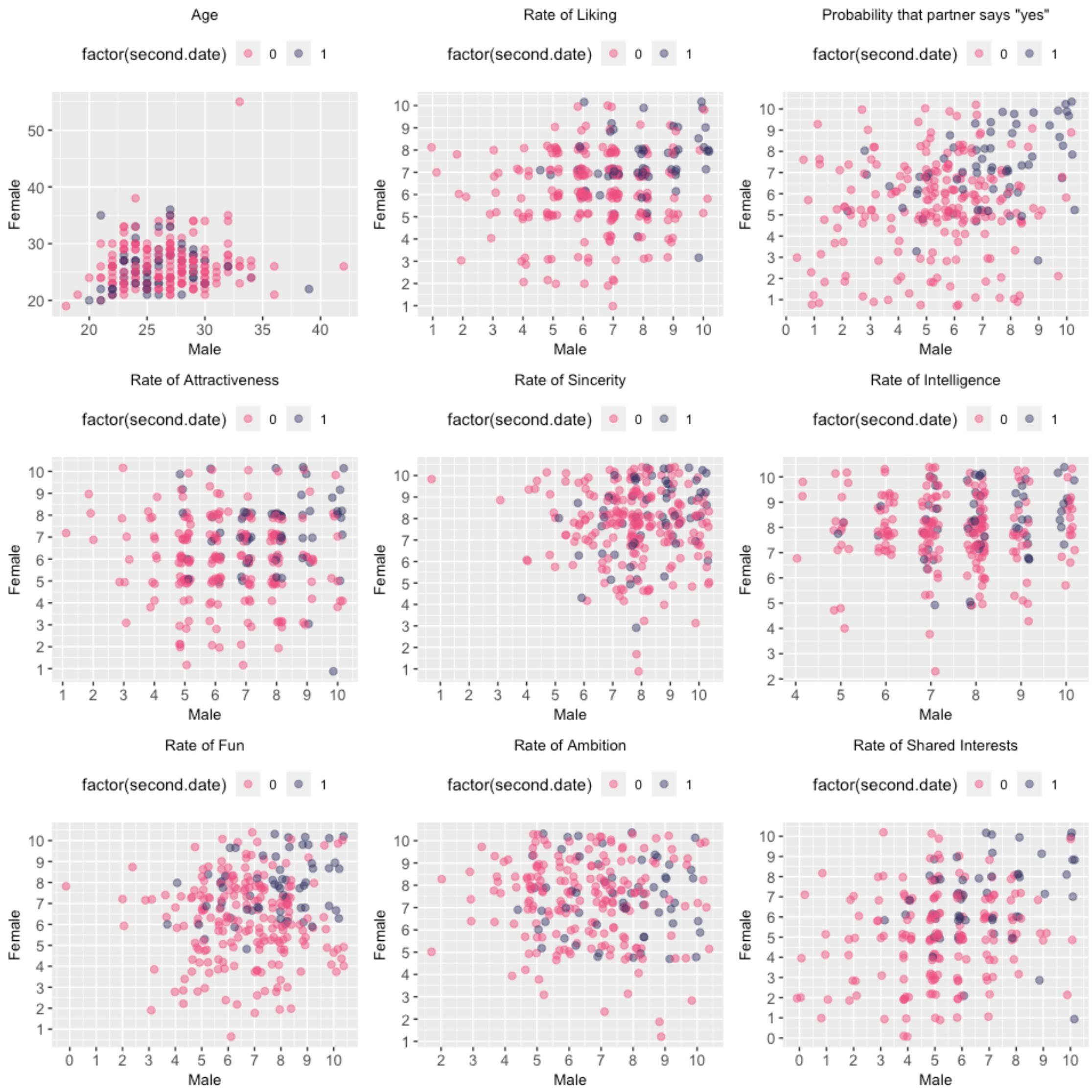
```
# Contingency table
xtabs(~DecisionM + DecisionF, data = date)

##          DecisionF
## DecisionM  0  1
##         0 66 64
##         1 83 63
```

From the contingency table, we can see that the two people in 63 dates want a second date. The percentage is

$$\frac{63}{276} = 22.83\%.$$

Q2

```
# Make a new column 'second.date'
date['second.date'] <- 0
date[(date$DecisionM == 1) & (date$DecisionF == 1), 'second.date'] <- 1

theme.info <- theme(plot.title = element_text(size=10, hjust=0.5),
                    axis.title = element_text(size=10),
                    axis.text = element_text(size=10),
                    legend.position = 'top')
# Jitter plot
j_like <- date %>% ggplot(aes(x = LikeM, y=LikeF, col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Liking") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_partneryes <- date %>% ggplot(aes(x = PartnerYesM, y=PartnerYesF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Probability that partner
says \"yes\"") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

s_age <- date %>% ggplot(aes(x = AgeM, y=AgeF, col=factor(second.date))) +
  geom_point(size=2) + ggtitle("Age") +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_attractive <- date %>% ggplot(aes(x = AttractiveM, y=AttractiveF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Attractiveness")
+
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_sincere <- date %>% ggplot(aes(x = SincereM, y=SincereF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Sincerity") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_intelligent <- date %>% ggplot(aes(x = IntelligentM, y=IntelligentF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Intelligence") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_fun <- date %>% ggplot(aes(x = FunM, y = FunF, col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Fun") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_ambitious <- date %>% ggplot(aes(x = AmbitiousM, y=AmbitiousF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Ambition") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

j_interest <- date %>% ggplot(aes(x = SharedInterestsM, y=SharedInterestsF,
col=factor(second.date))) +
  geom_point(size=2, position = "jitter") + ggtitle("Rate of Shared
Interests") +
  scale_x_continuous(breaks=seq(0, 10, 1)) +
  scale_y_continuous(breaks=seq(0, 10, 1)) +
  scale_color_manual(values=c("#ef528580", "#35386680")) +
  labs(x="Male", y="Female") + theme.info

jitter <- grid.arrange(s_age, j_like, j_partneryes, j_attractive, j_sincere,
j_intelligent,
                        j_fun, j_ambitious, j_interest, nrow=3)
```



The outcomes (second date or no second date) are not completely separable for each numerical variable. It is obvious that the observations of second date tend to rate higher for liking and attractiveness.

Q3

```
# check reasonableness of numerical variables
for (col in colnames(date[, c(3:6, 11:22)])){
  r <- any(date[col]<1 | date[col] > 10, na.rm = TRUE)
  if (r){
    print(paste("Not all responses to ", col, " are within range from 1 to 
10."))
    row <- which(date[col]<1 | date[col] > 10)
    # print(row)
    print(date[row, col])
  }else{
    print(paste("All responses to ", col, " are within range from 1 to 10."))
  }
}
```

```
## [1] "All responses to  LikeM  are within range from 1 to 10."
## [1] "All responses to  LikeF  are within range from 1 to 10."
## [1] "Not all responses to  PartnerYesM  are within range from 1 to 10."
## # A tibble: 1 x 1
##   PartnerYesM
##         <int>
## 1           0
## [1] "All responses to  PartnerYesF  are within range from 1 to 10."
## [1] "All responses to  AttractiveM  are within range from 1 to 10."
## [1] "All responses to  AttractiveF  are within range from 1 to 10."
## [1] "All responses to  SincereM  are within range from 1 to 10."
## [1] "All responses to  SincereF  are within range from 1 to 10."
## [1] "All responses to  IntelligentM  are within range from 1 to 10."
## [1] "All responses to  IntelligentF  are within range from 1 to 10."
## [1] "Not all responses to  FunM  are within range from 1 to 10."
## # A tibble: 1 x 1
##    FunM
##   <int>
## 1     0
## [1] "All responses to  FunF  are within range from 1 to 10."
## [1] "All responses to  AmbitiousM  are within range from 1 to 10."
## [1] "All responses to  AmbitiousF  are within range from 1 to 10."
## [1] "Not all responses to  SharedInterestsM  are within range from 1 to
10."
## # A tibble: 4 x 1
##   SharedInterestsM
##              <dbl>
## 1                0
## 2                0
## 3                0
## 4                0
## [1] "Not all responses to  SharedInterestsF  are within range from 1 to
10."
## # A tibble: 2 x 1
##   SharedInterestsF
##              <dbl>
## 1                0
## 2                0
```

According to the output, the responses to PartnerYesM, FunM, SharedInterestsM, SharedInterestsF are not within the range 1-10. After inspecting each variable, all the values without the range are 0. These should be treated as wrong records need correcting. These maybe caused by the respondents' cursoriness ignoring the range is from 1. So we correct the 0 answer to 1.

```
datec = date %>% mutate(PartnerYesM = replace(PartnerYesM, PartnerYesM == 0,
1)) %>%
  mutate(FunM = replace(FunM, FunM == 0, 1)) %>%
  mutate(SharedInterestsM = replace(SharedInterestsM, SharedInterestsM == 0,
1)) %>%
  mutate(SharedInterestsF = replace(SharedInterestsF, SharedInterestsF == 0,
1))

# check missing data in numerical variables

# datec[, c(3:6, 11:22)] %>% summarize_all(funs(sum(is.na(.))))
# able to use, does not show well in rmarkdown

for (col in colnames(datec[, c(3:6, 11:22)])){
  #exist_na <- any(is.na(datec[col])==TRUE)
  count_na <- sum(is.na(datec[col])==TRUE)
  if (count_na>0){
    print(paste(as.character(count_na), " missing observation(s) in responses
to ", col))
  }else{
    print(paste("No missing observations in responses to ", col))
  }
}

## [1] "2  missing observation(s) in responses to  LikeM"
## [1] "4  missing observation(s) in responses to  LikeF"
## [1] "4  missing observation(s) in responses to  PartnerYesM"
## [1] "4  missing observation(s) in responses to  PartnerYesF"
## [1] "3  missing observation(s) in responses to  AttractiveM"
## [1] "2  missing observation(s) in responses to  AttractiveF"
## [1] "5  missing observation(s) in responses to  SincereM"
## [1] "3  missing observation(s) in responses to  SincereF"
## [1] "8  missing observation(s) in responses to  IntelligentM"
## [1] "3  missing observation(s) in responses to  IntelligentF"
## [1] "6  missing observation(s) in responses to  FunM"
## [1] "6  missing observation(s) in responses to  FunF"
## [1] "17  missing observation(s) in responses to  AmbitiousM"
## [1] "10  missing observation(s) in responses to  AmbitiousF"
## [1] "27  missing observation(s) in responses to  SharedInterestsM"
## [1] "30  missing observation(s) in responses to  SharedInterestsF"
```

Missing data exist in the responses and the output above shows the specific number of missing data.

Q4

```
# check race categories
unique(c(datec$RaceM, datec$RaceF))

## [1] "Caucasian" "Asian"     "Latino"    "Black"     "Other"     NA
```

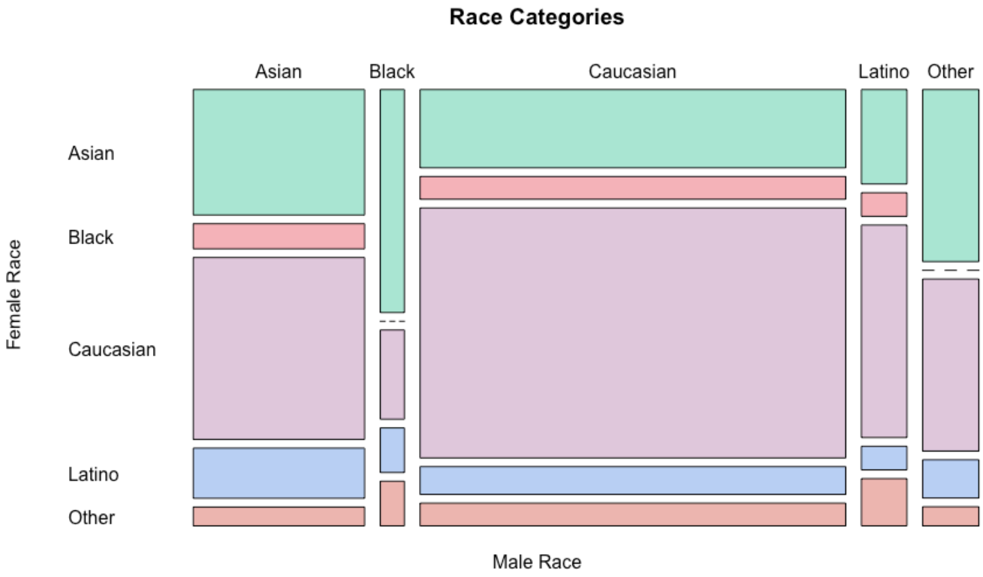
```
unique(datec$RaceM)

## [1] "Caucasian" "Asian"     "Latino"    "Black"     "Other"     NA

unique(datec$RaceF)

## [1] "Caucasian" "Asian"     "Other"     "Black"     "Latino"    NA
```

The possible race categories are Caucasian, Asian, Latino, Black and Other, for both male and famale.

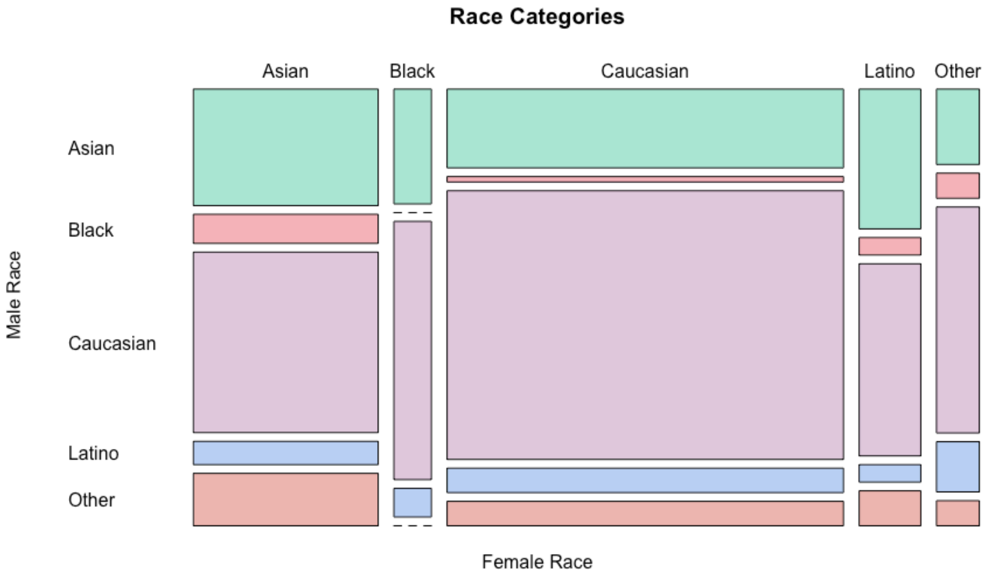
```
# check missing data in race
racem_count_na <- sum(is.na(datec["RaceM"]) == TRUE) # 2
racef_count_na <- sum(is.na(datec["RaceF"]) == TRUE) # 4
racemf_count_na <- sum((is.na(datec["RaceM"])) | (is.na(datec["RaceF"])) ==
TRUE) # 6
racem_count_na

## [1] 2

racef_count_na

## [1] 4

racemf_count_na

## [1] 6
```

2 missing data in race of male and 4 missing data in race of famale. There are 6 records in total missing race information in the datset. It is hard to amend the missing race information. If `RaceF`, `RaceM` are included in the model, the missing value should be removed.

```
mosaicplot(table(datec$RaceM, datec$RaceF), main = "Race Categories",
           xlab = "Male Race", ylab = "Female Race",
           las = TRUE, cex.axis =1, color=c("#67D5B595", "#EE778595",
"#C89EC495", "#84B1ED95", "#DE7E7395"))
```



```
mosaicplot(table(datec$RaceF, datec$RaceM), main = "Race Categories",
           xlab = "Female Race", ylab = "Male Race",
           las = TRUE, cex.axis =1, color=c("#67D5B595", "#EE778595",
"#C89EC495", "#84B1ED95", "#DE7E7395"))
```



Most participants are Caucasian and Asian males and females. The most dates are between Caucasian male and Caucasian female. No Black females date with Black males and other race males.

Q5

```
# Model selection
complete_date <- datec[complete.cases(datec),-c(1, 2)]

model.full <- glm(second.date ~ ., data = complete_date, family =
binomial(link="logit"))

step <- step(model.full, direction="backward", trace = 1, test="Chisq")

model.attemp1 <- glm(second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF +
AgeM +
                    RaceF + AttractiveF + AmbitiousF, data=datec, family =
binomial(link="logit"))
summary(model.attemp1)

model.attemp13 <- glm(second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF
+
                       AttractiveF + AmbitiousF, data=datec, family =
binomial(link="logit"))
summary(model.attemp13)
```

To select the model, we first only keep complete cases in the dataset and use backward stepwise method to find possible explanatory variables in the final model. However, when we establish the model with those possible explanatory variables, we only exclude those observations that have missing values in the variables that are actually included in that final model.[1]

After backward stepwise method based on AIC, we obtain `model.attemp1` with `LikeM`, `FunF`, `PartnerYesM`, `PartnerYesF`, `AgeM`, `RaceF`, `AttractiveF`, `AmbitiousF`. After that we remove the insiginificant variables and obtain `model.attemp13` as `model.final` with explanatory variables `LikeM`, `FunF`, `PartnerYesM`, `PartnerYesF`, `AttractiveF`, `AmbitiousF`.

```
# final model
model.final <- glm(second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF +
                    AttractiveF + AmbitiousF, data = datec, family =
binomial(link="logit"))
summary(model.final)

##
## Call:
## glm(formula = second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF +
##     AttractiveF + AmbitiousF, family = binomial(link = "logit"),
##     data = datec)
##
## Deviance Residuals:
```

[1] The process refers https://rcompanion.org/rcompanion/e_07.html

```
##     Min       1Q   Median       3Q      Max  
## -2.1475  -0.5828  -0.2897  -0.0281   2.6552  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -10.5161     1.6410  -6.408 1.47e-10 ***
## LikeM         0.4940     0.1345   3.673 0.000239 ***
## FunF          0.3486     0.1449   2.406 0.016140 *  
## PartnerYesM   0.3416     0.1029   3.321 0.000897 ***
## PartnerYesF   0.2693     0.1039   2.592 0.009537 ** 
## AttractiveF   0.2860     0.1211   2.361 0.018206 *  
## AmbitiousF   -0.3047     0.1284  -2.374 0.017618 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 282.88  on 254  degrees of freedom
## Residual deviance: 187.88  on 248  degrees of freedom
##   (21 observations deleted due to missingness)
## AIC: 201.88
## 
## Number of Fisher Scoring iterations: 6
```

In this dataset, the responses of explanatory variables are from respondents and we have corrected the irrational values, so that they should be measured without error.

The model seems corretly specified since the 6 variables includes rating questions for both male and female and the coefficients are reasonable according to our intuition.

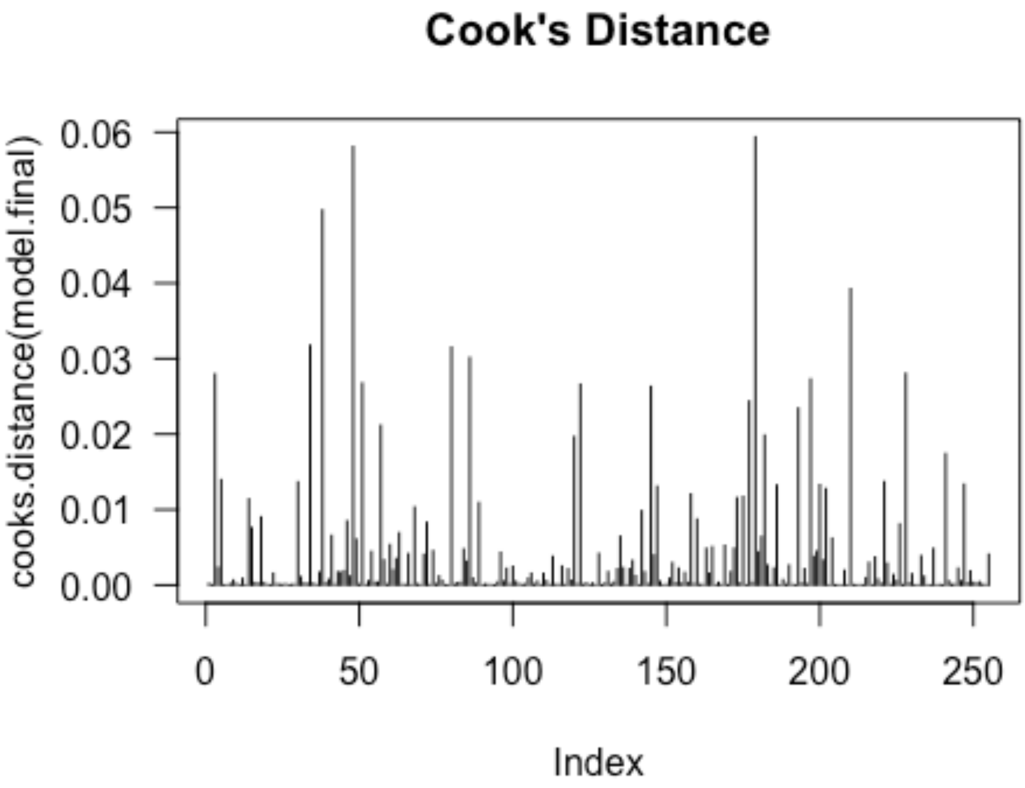
From the scatterplots in Q2, we can see that outcomes are not completely separable for each explanatory variables.

The plot of Cook's distance does not show obvious outliers in the dataset.

Each record is collected from each pair of participants, so the independence of observations should be ok.

Multicollinearity problem can be excluded, since VIF values for all explanatory variables in our final model are less than 10.

```
date.final <- datec %>% dplyr::select(second.date, LikeM, FunF, PartnerYesM, 
PartnerYesF, AttractiveF, AmbitiousF) %>% na.omit()
plot(cooks.distance(model.final), , type="h", las=TRUE, main="Cook's 
Distance")
```



```
usdm::vif(as.data.frame(date.final[-c(1)]))

##     Variables      VIF
## 1       LikeM 1.211254
## 2        FunF 1.878969
## 3 PartnerYesM 1.266736
## 4 PartnerYesF 1.244216
## 5 AttractiveF 1.457386
## 6  AmbitiousF 1.186762

# Can also be done by calculating from summary(model.final)
model.null <- glm(second.date ~ 1, data=date.final, family =
binomial(link="logit"))
#anova(model.final, model.null, test="Chisq")
anova(model.final, model.null, test="LRT")

## Analysis of Deviance Table
##
## Model 1: second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF +
AttractiveF +
##     AmbitiousF
## Model 2: second.date ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       248     187.88
## 2       254     282.88 -6      -95 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio test for overall model:

$$H_0: \beta_{LikeM} = \beta_{FunF} = \beta_{PartnerYesM} = \beta_{PartnerYesF} = \beta_{AttractiveF} = \beta_{AmbitiousF} = 0$$

$$H_a\text{: At least one of the slopes is not 0.}$$

The test statistic $G = 95$ which has a χ^2 distribution with $df = 6$. The p-value is 2.2×10^{-16}, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and the overall model is significant.

z-test for slopes:

$$H_0: \beta_{LikeM} = 0 \quad \text{v.s.}\, H_a: \beta_{LikeM} \neq 0$$

The test statistic $z = 3.673$ which has a standard normal distribution. The p-value is 0.000239, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and β_{LikeM} is statistically significant.

$$H_0: \beta_{FunF} = 0 \quad \text{v.s.}\, H_a: \beta_{FunF} \neq 0$$

The test statistic $z = 2.406$ which has a standard normal distribution. The p-value is 0.0161, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and β_{FunF} is statistically significant.

$$H_0: \beta_{PartnerYesM} = 0 \quad \text{v.s.}\, H_a: \beta_{PartnerYesM} \neq 0$$

The test statistic $z = 3.321$ which has a standard normal distribution. The p-value is 0.0008, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and $\beta_{PartnerYesM}$ is statistically significant.

$$H_0: \beta_{PartnerYesF} = 0 \quad \text{v.s.}\, H_a: \beta_{PartnerYesF} \neq 0$$

The test statistic $z = 2.592$ which has a standard normal distribution. The p-value is 0.0095, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and β_{LikeM} is statistically significant.

$$H_0: \beta_{AttractiveF} = 0 \quad \text{v.s.}\, H_a: \beta_{AttractiveF} \neq 0$$

The test statistic $z = 2.361$ which has a standard normal distribution. The p-value is 0.0182, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and $\beta_{AttractiveF}$ is statistically significant.

$$H_0: \beta_{AmbitiousF} = 0 \quad \text{v.s.}\, H_a: \beta_{AmbitiousF} \neq 0$$

The test statistic $z = -2.374$ which has a standard normal distribution. The p-value is 0.0176, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis and $\beta_{AmbitiousF}$ is statistically significant.

Q6

```
date.new <- datec[complete.cases(datec %>% dplyr::select(second.date, LikeM,
FunF, PartnerYesM, PartnerYesF, AttractiveF, AmbitiousF)),]
nrow(date.new)

## [1] 255

xtabs(~DecisionM + DecisionF, data = date.new)

##          DecisionF
## DecisionM  0  1
##         0 59 57
##         1 77 62

kable(table(date.new$second.date), col.names = c("second.date", "Freq"))
```

second.date	Freq
0	193
1	62

We have 6 explanatory variables in the final model. According to the rule of thumb, at least 10 observations for each outcome per predictor. Therefore we need at least 60 samples for each outcomes. In our model, we have 193 no second date and 62 second date. So the number of explanatory variables follows rule of thumb.

Q7

Holding other variables constant, if rate of LikeM increases by 1, the odds of second date increases by $(e^{0.4940} - 1) \times 100\% = 63.8858561\%$.

Holding other variables constant, if rate of FunF increases by 1, the odds of second date increases by $(e^{0.3486} - 1) \times 100\% = 41.7082244\%$.

Holding other variables constant, if rate of PartnerYesM increases by 1, the odds of second date increases by $(e^{0.3416} - 1) \times 100\% = 40.7197306\%$.

Holding other variables constant, if rate of PartnerYesF increases by 1, the odds of second date increases by $(e^{0.2693} - 1) \times 100\% = 30.9047796\%$.

Holding other variables constant, if rate of AttractiveF increases by 1, the odds of second date increases by $(e^{0.2860} - 1) \times 100\% = 33.1092455\%$.

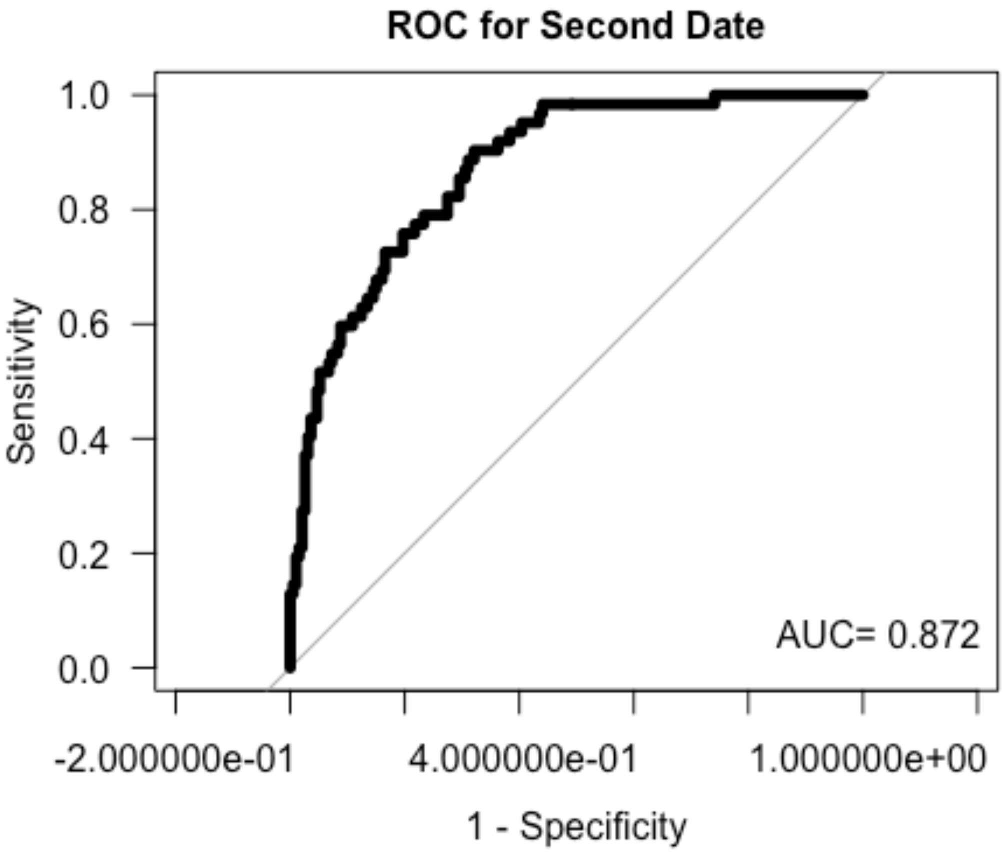
Holding other variables constant, if rate of AmbitiousF increases by 1, the odds of second date decreases by $(e^{0.3047} - 1) \times 100\% = 35.6218077\%$.

`LikeM`, `FunF`, `PartnerYesM`, `PartnerYesF`, `AttractiveF` increase the probability of a second date. This corresponds to our intuition that if you appreciate, understand your partner and the feeling between both you is good, a second date is more probable.

AmbitiousF decreases the probability of a second date. It is hard to tell ambition is good or not in an relationship. But an ambitious man maybe a little aggressive and proud in a speed dating, which may causes females feeling not so well.

Q8

```
AUC <- auc(response=date.final$second.date,
predictor=model.final$fitted.values)
ROC <- roc(response=date.final$second.date,
predictor=model.final$fitted.values,
    plot=TRUE, las=TRUE,    legacy.axes=TRUE, lwd=5,
    main="ROC for Second Date", cex.main=1, cex.axis=1, cex.lab=1)
legend("bottomright",legend=paste("AUC=", round(AUC, digits=3), sep=" "),
       bty="n", cex=1)
```



```
temp <- as.data.frame(t(coords(ROC, x="best", ret=c("threshold",
"specificity", "sensitivity"))))
kable(coords(ROC, x="best", ret=c("threshold", "specificity",
"sensitivity")))
```

	x
threshold	0.1613693

specificity	0.6787565
sensitivity	0.9032258

```
temp_df <- data.frame(date.final,
"fitted.values"=round(model.final$fitted.values, digits=3))
classify.best <- rep(1, times=nrow(temp_df))
classify.best[temp_df$fitted.values < coords(ROC, x="best", ret="threshold")]
<- 0

table(classify.best, temp_df$second.date)

##
## classify.best   0   1
##             0 131   6
##             1  62  56
```

The AUC is 0.872. And the best threshold for classifying observations is 0.1614, which means when estimated probability is equal or great than 0.1614, we will classify the observation as 1 and the observation with estimated probability less than 0.1614 will be classified as 0.

```
(131 + 56)/(131+56+62+6) # Accuracy

## [1] 0.7333333

56/(56+6) # Sensitivity

## [1] 0.9032258

131/(131+62) # Specificity

## [1] 0.6787565
```

$$\text{Accuracy:} \quad \frac{131+56}{131+56+62+6} = 0.7333$$

$$\text{Sensitivity:} \quad \frac{56}{56+6} = 0.9032$$

$$\text{Specificity:} \quad \frac{131}{131+62} = 0.6788$$

One-Way ANOVA

Read data

```
kudzu_data <- read_excel("kudzu.xls", col_names = TRUE)
kudzu_data$Treatment <- factor(kudzu_data$Treatment,
                               levels = c('Control', 'LowDose', 'HighDose'),
                               ordered = TRUE)
```

Q9

Response variable is bone mineral density in the femur (g/cm^2)

Q10

Factors is the dosage of isoflavones. Levels are control group (no usage of isoflavones), a low dose of isoflavones and a high dose of isoflavones.

Q11

3 treatments are included, since only one factor is explored.

Q12

Since rats were randomly assigned and each group has 15 observations, the experimental design is a comletely randomized, balanced design.

```
table(kudzu_data$Treatment)

## 
##  Control  LowDose HighDose 
##       15       15       15
```

Q13

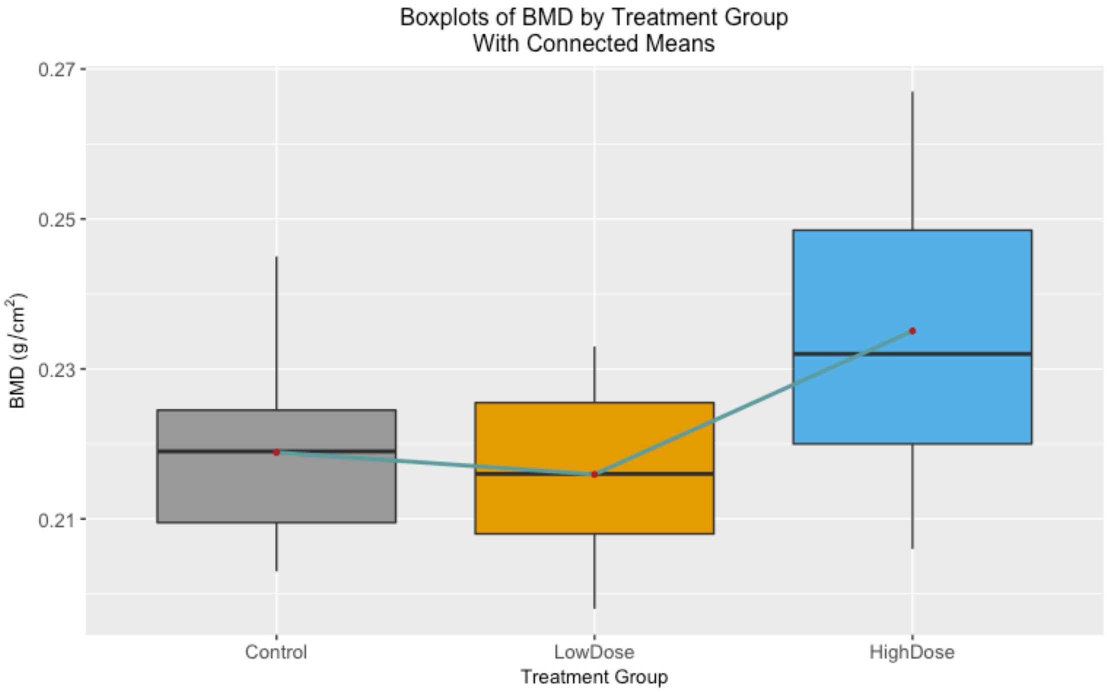
```
count = kudzu_data %>% group_by(Treatment) %>% count()
mean_sd = kudzu_data %>% group_by(Treatment) %>% summarize(mean(BMD),
sd(BMD))
table = data.frame(full_join(count, mean_sd, by = "Treatment"))
kable(table, col.names = c( 'Treatment', "Sample Size", "BMD Mean
($\\text{g/cm}^2$)", "BMD Std ($\\text{g/cm}^2$)"), caption = 'Summary
Table')
```

Summary Table

Treatment	Sample Size	BMD Mean (g/cm^2)	BMD Std (g/cm^2)
Control	15	0.2188667	0.0115873
LowDose	15	0.2159333	0.0115107
HighDose	15	0.2350667	0.0187711

Q14

```
# 14. side-by-side box plots
theme.info <- theme(plot.title = element_text(size=12, hjust=0.5),
                    axis.title = element_text(size=10),
                    axis.text = element_text(size=10))
kudzu_data %>%
  ggplot(aes(Treatment, BMD, fill = Treatment)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="line", aes(group=1), lwd=1, col="cadetblue")
+
  stat_summary(fun.y=mean, geom="point", pch=19, size=1, col="firebrick") +
  ggtitle("Boxplots of BMD by Treatment Group\nWith Connected Means") +
  scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9")) +
  labs(x="Treatment Group",
       y=expression(paste("BMD (", g/cm^{2},")"))) +
  theme.info + theme(legend.position="none")
```



From the box plot, the 1st quartile, 3rd quartile, median, maximum and mean BMD in HighDose group are obviously larger than those in Control group and LowDose group.

The 1st quartile, 3rd quartile, median and mean BMD in Control group and LowDose group are not much different. The mean and median BMD in LowDose group are a little lower than Control group.

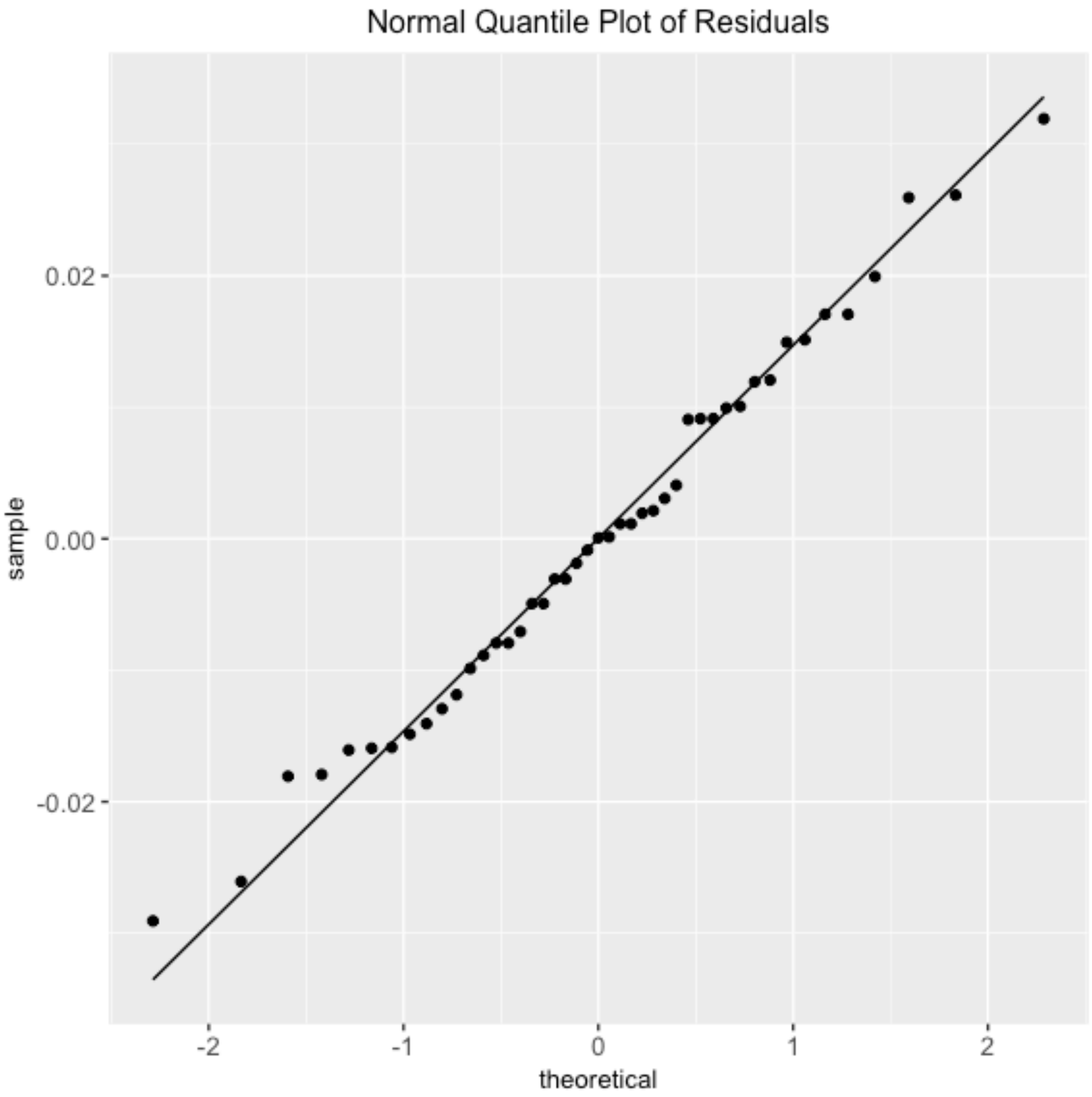
Q15

The experiment is a completely randomized design, so it satisfies independence. And we have check the sample size in each treatment. This is a balanced design.

By the normal quantile plot of residuals, the points basically coincide with the theoratical line. So the residuals follow normal distribution.

The largest standard deviation is 0.0188 (HighDose) and the smallest standard deviation is 0.0115 (LowDose).According to the rule of thumb, since the largest standard deviation is less than twice the smallest standard deviation, it can be regarded as constant variance.

```r
# residuals
mean <- kudzu_data %>%
  group_by(Treatment) %>%
  summarize(mean(BMD))
left_join(kudzu_data, mean, by = "Treatment") %>%
  mutate(residuals = BMD - `mean(BMD)`) %>%
  ggplot(aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Normal Quantile Plot of Residuals") +
  theme.info
```



Q16

One-way ANOVA hypotheses are:

$$H_0 : \mu_{\text{Control}} = \mu_{\text{LowDose}} = \mu_{\text{HighDose}} \quad \text{v. s.} \quad H_a : \text{At least two means are different}$$

The test statistic $F = 7.718$ with 2 and 42 degrees of freedom. $p - \text{value} = 0.0014 < 0.01 = \alpha$. Therefore we reject null hypothesis, which means at least two means are different.

```
# one-way ANOVA model
kudzu_anova <- aov(BMD ~ Treatment, data = kudzu_data)
summary(kudzu_anova)
```

```
##             Df   Sum Sq   Mean Sq F value Pr(>F)   
## Treatment    2 0.003186 0.0015928   7.718 0.0014 **
## Residuals   42 0.008668 0.0002064                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

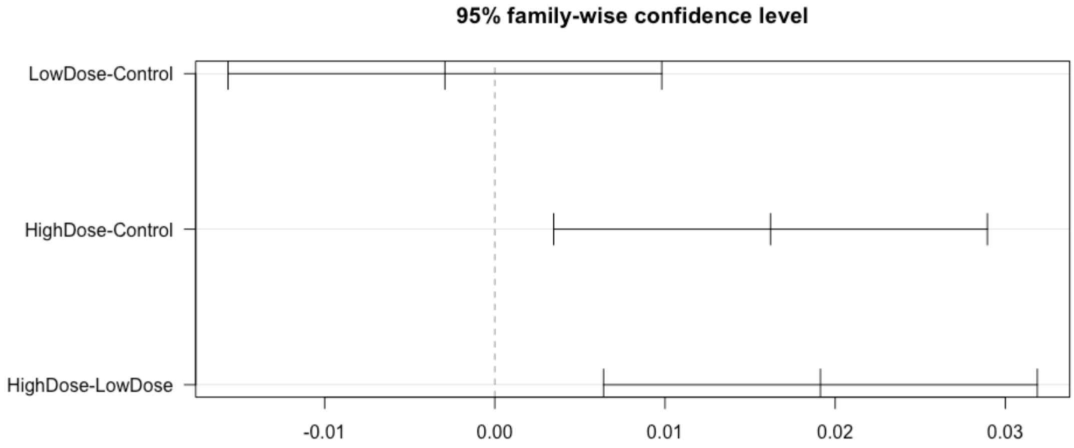
Q17

p-values of the difference in HighDose-Control, HighDose-LowDose are less than critical value 0.05. So HighDose group is significantly different from others. Also in the plot we can see that 0 is not in the confidence interval of those two pairwise differences.

```
# Tukey’s multiple-comparisons method
kudzu_Tukey <- TukeyHSD(kudzu_anova, conf.level = 0.95)
kudzu_Tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BMD ~ Treatment, data = kudzu_data)
##
## $Treatment
##                           diff          lwr         upr     p adj
## LowDose-Control  -0.002933333 -0.015677456 0.009810789 0.8423308
## HighDose-Control  0.016200000  0.003455877 0.028944123 0.0097645
## HighDose-LowDose  0.019133333  0.006389211 0.031877456 0.0020537
```

```
par(mar=c(2, 9, 3, 2))
plot(kudzu_Tukey, las=TRUE)
```



95% family-wise confidence level