SDGB 7840 HW 4: Speed Dating and Isoflavones

Darshan Patel

3/28/2019

## Part 1: Logistic Regression

In speed dating, participants meet many people, each for a few minutes, and then decide who they would like to see again. The dataset you will be working with contains information on speed dating experiments conducted on graduate and professional students. Each person in the experiment met with to randomly selected people of the opposite sex (only heterosexual pairings) for four minutes. After each speed date, each participant filled out a questionnaire about the other person.

Your goal is to build a model to predict which pairs of daters want to meet each other again (i.e., have a second date). The list of variables are:

## Warning: package 'knitr' was built under R version 3.4.4

|  |  |
| --- | --- |
| x | y |
| Decision | 1 = Yes (want to see the date again), 0 = No (do not want to see date again) |
| Like | Overall, how much do you like this person? (1 = don’t like at all, 10 = like a lot) |
| PartnerYes | How probable do you think it is that this person will say yes for you? (1 = not probable, 10 = extremely probable) |
| Age | Age |
| Race | Race (Caucasian, Asian, Black, Latino, or Other) |
| Attractive | Rate attractiveness of partner on scale of 1-10 (1 = awful, 10 = great) |
| Sincere | Rate sincerity of partner on a scale of 1-10 (1 = awful, 10 = great) |
| Intelligent | Rate intelligence of partner on a scale of 1-10 (1 = awful, 10 = great) |
| Fun | Rate how fun partner is on a scale of 1-10 (1 = awful, 10 = great) |
| Ambitious | Rate ambition of partner on a scale of 1-10 (1 = awful, 10 = great) |
| Shared Interests | Rate the extent to which you share interests/hobbies with partner on a scale of 1-10 (1 = awful, 10 = great) |

We will be using a reduced version of this experimental data will be used with unique male-female date pairs. In the file SpeedDating.csv, the variables have either M for male or F for female. For example, LikeM refers to the Like variable as answered by the male participant (about the female participant). Treat the rating scale variables (such as PartnerYes, Attractive, etc.) as *numerical variables* instead of categorical ones for the analysis.

library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.1 ✔ purrr 0.3.1   
## ✔ tibble 2.1.1 ✔ dplyr 0.8.0.1  
## ✔ tidyr 0.8.3 ✔ stringr 1.3.0   
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

# Read in data  
df = read\_delim("SpeedDating.csv", delim = ',')

## 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.

# Look at class of all variables  
sapply(df, class)

## DecisionM DecisionF LikeM LikeF   
## "integer" "integer" "numeric" "numeric"   
## PartnerYesM PartnerYesF AgeM AgeF   
## "integer" "integer" "integer" "integer"   
## RaceM RaceF AttractiveM AttractiveF   
## "character" "character" "numeric" "numeric"   
## SincereM SincereF IntelligentM IntelligentF   
## "integer" "integer" "numeric" "integer"   
## FunM FunF AmbitiousM AmbitiousF   
## "integer" "integer" "integer" "integer"   
## SharedInterestsM SharedInterestsF   
## "numeric" "numeric"

All variables are of the proper types.

### Question 1

Based on the variable Decision, fill out the contingency table. What percentage of dates ended with both people wanting a second date?

Answer:

# Display contingency table  
table("Decision Made by Men" = ifelse(df$DecisionM == 0, "No", "Yes"),   
 "Decision Made by Women" = ifelse(df$DecisionF == 0, "No", "Yes"))

## Decision Made by Women  
## Decision Made by Men No Yes  
## No 66 64  
## Yes 83 63

The percentage of dates that ended with both people wanting a second date is

# Get percentage of second dates  
paste(round(100 \* 63 / nrow(df), 2), '%', sep = '')

## [1] "22.83%"

### Question 2

A second date is planned only if both people within the matched pair want to see each other again. Make a new column in your dataset and call it second.date. Values in this column should be if there will be no second date, if there will be a second date. Construct a scatterplot for each numerical variable where the male values are on the -axis and the female values are on the -axis. Observations in your scatterplot should have a different color (or pch value) based on whether or not there will be a second date. Describe what you see. (Note: Jitter your points just for making these plots.)

Answer:

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(RColorBrewer)  
# Create second.date variable column  
df$second.date = ifelse((df$DecisionM == 1) & (df$DecisionF == 1), 1, 0)  
# A function to plot numerical variables  
# Params: Male rating, Female rating, second date indicator, title of variable  
plot\_features = function(x, y, c, t){  
 return(as.data.frame(cbind(x,y,c)) %>%   
 ggplot(aes(x=x, y=y, color=as.factor(c))) +   
 geom\_jitter() +   
 scale\_color\_manual(name = "Decision Made",   
 values = brewer.pal(3, "Dark2")) +   
 scale\_x\_continuous(limits = c(0, 10), breaks = seq(0, 10, 2)) +   
 scale\_y\_continuous(limits = c(0,10), breaks = seq(0,10,2.5)) +   
 labs(x = "Male Decision", y = "Female Decision",   
 title = paste0("Scatterplot of ", t),   
 subtitle = "by Gender") +   
 theme\_minimal() + theme(legend.position = "bottom",   
 plot.title = element\_text(size = 8),  
 plot.subtitle = element\_text(size = 6),   
 axis.text = element\_text(size = 6)))  
}  
  
# Create plots for all numerical variables   
g1 = plot\_features(df$LikeM, df$LikeF, df$second.date, "Likeness Rating")  
g2 = plot\_features(df$PartnerYesM, df$PartnerYesF, df$second.date, "Probable Rating")  
g3 = plot\_features(df$AgeM, df$AgeF, df$second.date, "Age")  
g4 = plot\_features(df$AttractiveM, df$AttractiveF, df$second.date, "Attractiveness Rating")  
g5 = plot\_features(df$SincereM, df$SincereF, df$second.date, "Sincerity Rating")  
g6 = plot\_features(df$IntelligentM, df$IntelligentF, df$second.date, "Intelligence Rating")  
g7 = plot\_features(df$FunM, df$FunF, df$second.date, "Fun Rating")  
g8 = plot\_features(df$AmbitiousM, df$AmbitiousF, df$second.date, "Ambitious Rating")  
g9 = plot\_features(df$SharedInterestsM, df$SharedInterestsF, df$second.date, "Shared Interests Rating")

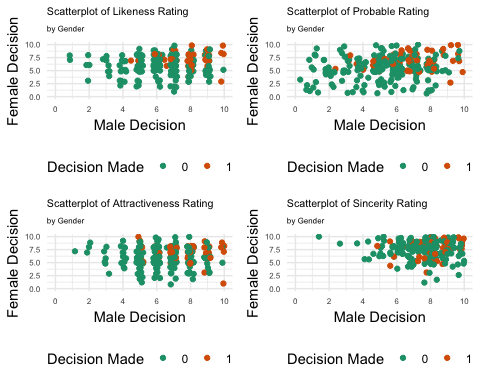
# Display 4 ratings plots  
grid.arrange(g1, g2, g4, g5, nrow = 2)

## Warning: Removed 17 rows containing missing values (geom\_point).

## Warning: Removed 18 rows containing missing values (geom\_point).

## Warning: Removed 22 rows containing missing values (geom\_point).

## Warning: Removed 45 rows containing missing values (geom\_point).



It is noted that whenever a second date did occur, i.e. decision made is 1, the rating given by both genders are ranked high. The sincerity rating are very close to each other unlike the probable ratings which are more disperse. Going to a second date is correlated with high likeness rating by female and high probable rating by female. Attractiveness rating by men is also correlated with having second dates.

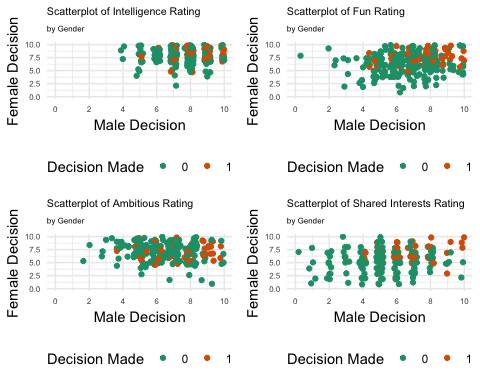
# Display other 4 rating plots  
grid.arrange(g6, g7, g8, g9, nrow = 2)

## Warning: Removed 41 rows containing missing values (geom\_point).

## Warning: Removed 24 rows containing missing values (geom\_point).

## Warning: Removed 54 rows containing missing values (geom\_point).

## Warning: Removed 63 rows containing missing values (geom\_point).



The people who had second dates usually fell in the top half in the rating scale for intelligence and fun ratings. People rarely rated below a on intelligence, resulting in huge crowdiness in the high ratings. The fun ratings and shared interests ratings are more disperse. Having a second date appears to be correlated with high ambition rating by men as well as high shared interests rating by men.

# Display age plot with replaced scales and title   
# and y = x line   
g3 + scale\_x\_continuous(limits = c(15, 45), breaks = seq(10, 45, 5)) +   
 scale\_y\_continuous(limits = c(15,45), breaks = seq(15, 45, 5)) +   
 geom\_abline(slope = 1, intercept = 0, linetype = 2) +   
 labs(x = "Male Age", y = "Female Age") + coord\_cartesian()

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.

## Scale for 'y' is already present. Adding another scale for 'y', which  
## will replace the existing scale.

## Warning: Removed 9 rows containing missing values (geom\_point).



The dotted line indicates a common age between male and female. According to the age scatterplot, there does not appear to be any association between age for either gender and going on a second date. There is one occurrence of a old man getting a second date with a woman who is at least years younger than him, and vice versa.

### Question 3

Many of the numerical variables are on rating scales from to . Are the responses within these ranges? If not, what should we do with these responses? Is there any missing data? If so, how many observations and for which variables?

Answer:

# Display summary statistics of numerical variables  
summary(df[,!colnames(df) %in% c("DecisionM", "DecisionF", "second.date", "RaceM", "RaceF")])

## LikeM LikeF PartnerYesM PartnerYesF   
## Min. : 1.000 Min. : 1.000 Min. : 0.000 Min. : 1.000   
## 1st Qu.: 6.000 1st Qu.: 5.000 1st Qu.: 5.000 1st Qu.: 5.000   
## Median : 7.000 Median : 7.000 Median : 6.000 Median : 6.000   
## Mean : 6.682 Mean : 6.366 Mean : 5.757 Mean : 5.835   
## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 7.000 3rd Qu.: 7.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :2 NA's :4 NA's :4 NA's :4   
## AgeM AgeF AttractiveM AttractiveF   
## Min. :18.0 Min. :19.00 Min. : 1.000 Min. : 1.000   
## 1st Qu.:24.0 1st Qu.:23.00 1st Qu.: 5.000 1st Qu.: 5.000   
## Median :27.0 Median :26.00 Median : 7.000 Median : 6.000   
## Mean :26.6 Mean :26.19 Mean : 6.687 Mean : 6.274   
## 3rd Qu.:29.0 3rd Qu.:28.00 3rd Qu.: 8.000 3rd Qu.: 8.000   
## Max. :42.0 Max. :55.00 Max. :10.000 Max. :10.000   
## NA's :3 NA's :5 NA's :3 NA's :2   
## SincereM SincereF IntelligentM IntelligentF   
## Min. : 1.000 Min. : 1.000 Min. : 4.000 Min. : 2.000   
## 1st Qu.: 7.000 1st Qu.: 7.000 1st Qu.: 7.000 1st Qu.: 7.000   
## Median : 8.000 Median : 8.000 Median : 8.000 Median : 8.000   
## Mean : 7.856 Mean : 7.784 Mean : 7.621 Mean : 7.923   
## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 8.250 3rd Qu.: 9.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :5 NA's :3 NA's :8 NA's :3   
## FunM FunF AmbitiousM AmbitiousF   
## Min. : 0.000 Min. : 1.000 Min. : 2.000 Min. : 1.000   
## 1st Qu.: 6.000 1st Qu.: 5.000 1st Qu.: 5.000 1st Qu.: 6.000   
## Median : 7.000 Median : 7.000 Median : 7.000 Median : 8.000   
## Mean : 6.863 Mean : 6.563 Mean : 6.768 Mean : 7.429   
## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 9.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :6 NA's :6 NA's :17 NA's :10   
## SharedInterestsM SharedInterestsF  
## Min. : 0.000 Min. : 0.00   
## 1st Qu.: 4.000 1st Qu.: 4.00   
## Median : 5.000 Median : 6.00   
## Mean : 5.588 Mean : 5.47   
## 3rd Qu.: 7.000 3rd Qu.: 7.00   
## Max. :10.000 Max. :10.00   
## NA's :27 NA's :30

All of the rating responses are well within the range from to . However, there is missing data is all of the numerical columns. The shared interests columns have a high number of missing data ( to ), whereas the ambitious ratings have moderate missing data ( to ). The other columns have little missing data (less than ).

# Find number of rows with null data  
nrow(df) - nrow(df[complete.cases(df),])

## [1] 76

A total of observations have missing data.

### Question 4

What are the possible race categories in your dataset? Is there any missing data? If so, how many observations and what should we do with them? Make a mosaic plot with female and male race. Describe what you see.

Answer:

# Find distinct races in both genders  
unique(c(df$RaceF, df$RaceM))

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

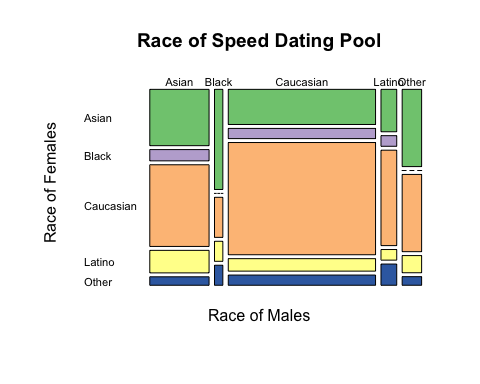
There are a total of race categories: Caucasian, Asian, Black, Latino and Other.

# Find locations where race is null for either  
# male or female participant  
union(which(is.na(df$RaceM)), which(is.na(df$RaceF)))

## [1] 166 167 29 30 66 169

Furthermore, observations have missing data (male or female). This is a considerably low number compared to the total number of observations. The missing race observations will be dropped.

# Create mosaic plot of races by gender  
mosaicplot(table(df$RaceM, df$RaceF), main = "Race of Speed Dating Pool",  
 xlab = "Race of Males", ylab = "Race of Females",   
 las=TRUE, cex.axis=0.7, color = brewer.pal(5, "Accent"))



It is apparent that the Caucasian race makes up most of the population for both male and female participants. There are no Black males nor Other males. The Latino, Black and Other race make up the minority groups in the speed dating pool whereas the Asian and Caucasian races make up most of the pool.

### Question 5

Use logistic regression to construct a model for second.date (i.e., second.date should be the response variable). Incorporate the discoveries and decisions made in question , and . Explain the steps you used to determine the best model. Include the summary output for your *final* model only, check the model assumptions, and evaluate the model by running the relevant hypothesis tests. Do *not* use Decision as an explanatory variable.

Answer:

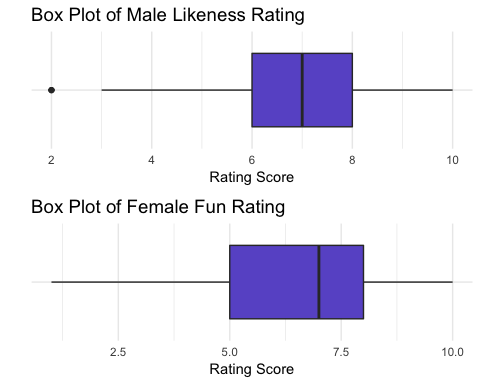
## Remove observations with missing values  
## Don't remove DecisionM and Decision (useful later on)  
df = df[complete.cases(df),]  
  
## Create full model using age, likeness rating and fun rating  
base\_model = glm(second.date ~ AgeM + AgeF + LikeM + LikeF + FunM + FunF - DecisionM - DecisionF, df, family = "binomial")  
  
## Perform backward stepwise selection  
best\_model = step(base\_model, direction = "backward", trace = 0)  
summary(best\_model)

##   
## Call:  
## glm(formula = second.date ~ LikeM + FunF, family = "binomial",   
## data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0459 -0.7075 -0.3900 -0.1062 2.4792   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.3307 1.3359 -6.236 4.49e-10 \*\*\*  
## LikeM 0.5802 0.1370 4.233 2.30e-05 \*\*\*  
## FunF 0.4490 0.1164 3.856 0.000115 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 220.43 on 199 degrees of freedom  
## Residual deviance: 170.15 on 197 degrees of freedom  
## AIC: 176.15  
##   
## Number of Fisher Scoring iterations: 5

A logistic regression model was first trained using six explanatory variables: AgeM, AgeF, LikeM, LikeF, FunM and FunF. These variables were selected because they appeared related to success of having a second date. Using this collection of variables, backward stepwise variable selection was performed to select the best subset of variables that resulted in the lowest AIC on the dataset with no missing values for any variables. There are a few assumptions made about this model. Namely, the explanatory variables are measured without error, the model is correctly specified, the outcomes are not linearly separable, no outliers or leverage points, observations are independent, no collinearity / multicollinearity, and that the sample size is at least for each outcome per predictor in the model. Now, the dataset is composed of records by people who have partaken in a speed dating round. Thus the explanatory variables are measured without error assuming the participants gave answers that reflected their true beliefs. Furthermore, the model is correctly specified; the stepwise selection process helped to narrow down on the important variables needed to explain second.date. Also, the observations are not linearly separable, since the glm algorithm did converge and fitted probabilities did not equal nor . However, the observations are not independent of each other however, since each person took part in multiple speed dating rounds (according to the data description given).

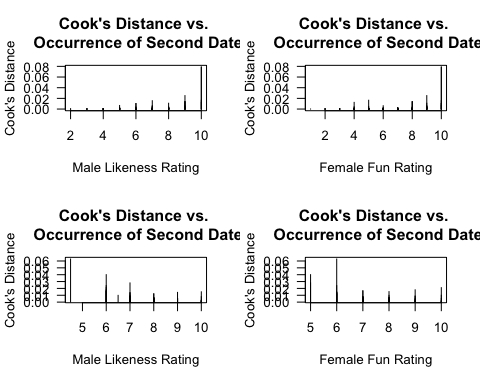
To check for outliers, box plots are made for the two numerical variables.

# A function to create box plots given variable to summarize and title  
box\_plotter = function(x, t){  
 return(as.data.frame(x) %>% ggplot(aes("", x)) +   
 geom\_boxplot(fill = "slateblue") +   
 ggtitle(paste0("Box Plot of ", t)) +   
 labs(x = "", y = "Rating Score") + coord\_flip() +   
 theme\_minimal() + theme(plot.title = element\_text(size = 14)))  
}  
  
# Create and display boxplots for both explanatory variables   
bp1 = box\_plotter(df$LikeM, "Male Likeness Rating")  
bp2 = box\_plotter(df$FunF, "Female Fun Rating")  
grid.arrange(bp1, bp2, nrow = 2)



After looking at the box plot of the numerical variables, it is apparent that there are outliers in one of the two numerical variables columns.

# Display Cook's distance for combinations of response and explanatory variables  
par(mfcol=c(2,2))  
plot(df$LikeM[df$second.date == 0],   
 cooks.distance(best\_model)[df$second.date == 0],   
 type="h", las=TRUE,   
 main="Cook's Distance vs. \n Occurrence of Second Date",  
 xlab="Male Likeness Rating", ylab="Cook's Distance")  
plot(df$LikeM[df$second.date == 1],   
 cooks.distance(best\_model)[df$second.date == 1], type="h", las=TRUE,   
 main="Cook's Distance vs. \n Occurrence of Second Date",   
 xlab="Male Likeness Rating", ylab="Cook's Distance")  
plot(df$FunF[df$second.date == 0],   
 cooks.distance(best\_model)[df$second.date == 0],   
 type="h", las=TRUE,   
 main="Cook's Distance vs. \n Occurrence of Second Date",   
 xlab="Female Fun Rating", ylab="Cook's Distance")  
plot(df$FunF[df$second.date == 1],   
 cooks.distance(best\_model)[df$second.date == 1], type="h", las=TRUE,   
 main="Cook's Distance vs. \n Occurrence of Second Date",   
 xlab="Female Fun Rating", ylab="Cook's Distance")



Since no point has a Cook’s distance of greater than , it is ruled that there are no leverage points in this model. However, since there outliers in one of the explanatory variables, the outlier/leverage point assumption is failed. To determine if there is collinearity or multicollinearity issues, check the variance inflation factors for the variables in the model.

# Import package for vif  
library(usdm)

## Loading required package: sp

## Warning: package 'sp' was built under R version 3.4.4

## Loading required package: raster

## Warning: package 'raster' was built under R version 3.4.4

##   
## Attaching package: 'raster'

## The following object is masked from 'package:dplyr':  
##   
## select

## The following object is masked from 'package:tidyr':  
##   
## extract

# Compute vif  
vif(as.data.frame(df[,c("LikeM", "FunF")]))

## Variables VIF  
## 1 LikeM 1.070754  
## 2 FunF 1.070754

The variables do not have a high inflation factor values (greater than ) and so it can be assumed that the two variables are not correlated with each other. Lastly, the sample size needs to be checked. There are predictors in the model, and outcomes, thus there must be observations for when second.date and likewise for second.date .

# Display frequency of second dates  
table("frequency" = df$second.date)

## frequency  
## 0 1   
## 152 48

The sample size assumption is passed since there are more than observations for both outcomes.

Now that the assumptions are checked, where only two failed, relevant hypothesis tests can be performed with caution. Let the null hypothesis be that

and that the alternative hypothesis be that at least one of the slopes is not . Then at the level of , , the test statistic, is

# Calculate test statistic  
summary(best\_model)$null.deviance - summary(best\_model)$deviance

## [1] 50.28225

and the degrees of freedom is

# Calculate degrees of freedom  
summary(best\_model)$df.null - summary(best\_model)$df.residual

## [1] 2

Then the -value, or the probability that , is

# Find probability of test statistic being greater than G  
pchisq(summary(best\_model)$null.deviance - summary(best\_model)$deviance,  
 df=summary(best\_model)$df.null - summary(best\_model)$df.residual,  
 lower.tail=FALSE)

## [1] 1.206003e-11

which is strictly less than . This means that the null hypothesis is rejected and that at least one of the two variables is nonzero and statistically significant. In addition to the log likelihood test for the overall model, a log-likelihood test for nested models can also be conducted using the base model before variable selection was done. This can be done because the same observations were used to create both models and the explanatory variables in the best model is a subset of the explanatory variables in the base model. Now, let the null hypothesis be that

(These are the variables not in the best model.) Let the alternative hypothesis be that at least one of the above slopes is not . The test statistic () and associated -value (or the probability that ) can be found from the ANOVA table with both of these models.

# Perform log-likelihood test  
anova(best\_model, base\_model, test="LRT")

## Analysis of Deviance Table  
##   
## Model 1: second.date ~ LikeM + FunF  
## Model 2: second.date ~ AgeM + AgeF + LikeM + LikeF + FunM + FunF - DecisionM -   
## DecisionF  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 197 170.15   
## 2 193 166.91 4 3.2375 0.5189

According to this, with a test statistic of and degrees of freedom of , the -value is which is strictly greater than . This means that at the significance level of , the null hypothesis cannot be rejected; this means that one of these four variables could be useful for the model. In addition to the log-likelihood test, additional -tests can be done for the slopes in the best model. For , let the null hypothesis be that and the alternative hypothesis be that . Likewise, for , the null and alternative hypotheses are and respectively. Then the -statistics and -values can be found from the summary output of the model.

# Display test statistics and p values   
summary(best\_model)$coef

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -8.3306620 1.3358934 -6.236023 4.488351e-10  
## LikeM 0.5801858 0.1370474 4.233466 2.301166e-05  
## FunF 0.4489956 0.1164455 3.855845 1.153306e-04

With a -value of and associated -value of , the coefficient estimate for LikeM is statistically significant at the significance level since the null hypothesis is rejected. Likewise for FunF, the coefficient estimate has a value of and an associated -value of . Thus at the significance level of , the null hypothesis is rejected and the coefficient estimate is statistically significant.

This is a pretty good model.

### Question 6

Redo question (1) using only the observations used to fit the final logistic regression model. What is your sample size? Does the number of explanatory variables in the model follow our rule of thumb? Justify your answer.

Answer:

# Display counts of decisions by both gender  
table("Decision Made by Men" = ifelse(df$DecisionM == 0, "No", "Yes"),   
 "Decision Made by Women" = ifelse(df$DecisionF == 0, "No", "Yes"))

## Decision Made by Women  
## Decision Made by Men No Yes  
## No 48 43  
## Yes 61 48

# Display sample size  
paste("The sample size is ", nrow(df), ".", sep = '')

## [1] "The sample size is 200."

According to the rule of thumb, there should be explanatory variables, or observations in each of the outcomes. This rule is satisfied because, thanks to variable subset selection, a low number of explanatory variables was found that could explain second.date. Only two variables were needed, or observations per outcome. It is clear from the contingency table above that there are more than observations per outcome.

### Question 7

Interpret the slopes in your model. Which explanatory variables increase the probability of a second date? Which ones decrease it? Is this what you expected to find? Justify.

Answer:

# Display coefficients of model  
summary(best\_model)$coef

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -8.3306620 1.3358934 -6.236023 4.488351e-10  
## LikeM 0.5801858 0.1370474 4.233466 2.301166e-05  
## FunF 0.4489956 0.1164455 3.855845 1.153306e-04

If the male likeness rating increases by , holding the fun rating by female constant, then the odds of having a second date increases by

# Calculate interpretation of coefficient of LikeM  
100 \* (exp(0.5801858) - 1)

## [1] 78.63703

. Likewise, if the fun rating by female increases by , holding the male likeness rating constant, then the odds of having a second date increases by

# Calculate interpretation of coefficient of FunF  
100 \* (exp(0.4489956) - 1)

## [1] 56.67378

.

Since LikeM and FunF cannot be , due to the structure of the experiment, the -intercept does not have a valid interpretation here.

Both of the explanatory variables increase the probability of a second date occurring; none decreased the probability. This was expected because most people like people who are fun and if a man does like a woman, a date is likely to occur.

### Question 8

Construct an ROC curve and compute the AUC. Determine the best threshold for classifying observations(i.e., second date or no second date) based on the ROC curve. Justify your choice of threshold. For your chosen threshold, compute (a) accuracy, (b) sensitivity, and (c) specificity.

Answer:

# Create ROC curve   
library(pROC)

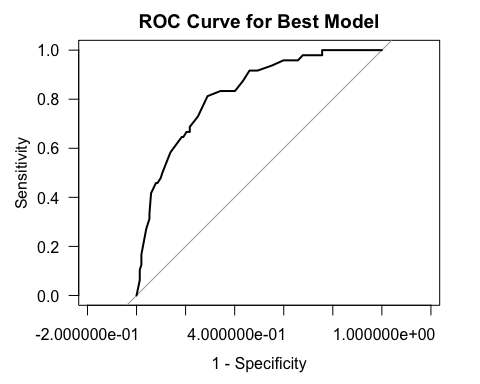
## Warning: package 'pROC' was built under R version 3.4.4

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

roc\_curve = roc(response = df$second.date,   
 predictor = best\_model$fitted.values,   
 plot = TRUE, las = TRUE, legacy.axes = TRUE,   
 main = "ROC Curve for Best Model")



The AUC, or area under the curve, is

# Calculate AUC  
auc(response = df$second.date, predictor = best\_model$fitted.values)

## Area under the curve: 0.8239

The best threshold for classifying observations is found by summing up the sensitivity and specificity and getting the highest value. It is the sum of the probability that a second date did occur when predicted it would and the probability that a second date would not occur when predicted it won’t. This is what most people would wish for after a first date.

# Find best threshold from ROC curve  
coords(roc\_curve, x="best",   
 ret=c("threshold", "specificity", "sensitivity"))

## threshold specificity sensitivity   
## 0.2331089 0.7105263 0.8125000

The best threshold for classifying observations is .

To compute the accuracy, sensitivity, and specificity, first predictions must be made using the model and classified using the threshold.

# Create predictions of second dates  
predictions = predict(best\_model, df, type = "response")  
df$chances = ifelse(predictions > 0.2331089, 1, 0)  
# Display predictions vs actual outcomes  
table("yhat" = df$chances,   
 "y" = df$second.date)

## y  
## yhat 0 1  
## 0 108 9  
## 1 44 39

According to the predictions made and actual occurrences of second dates, the accuracy is

# Accuracy calculation  
(108 + 39) / nrow(df)

## [1] 0.735

.

The sensitivity is

# Sensitivity calculation  
39 / (39 + 9)

## [1] 0.8125

.

Lastly, the specificity is

# Specificity calculation  
108 / (108 + 44)

## [1] 0.7105263

.

## Part 2: One-Way ANOVA

Kudzu is a plant that was imported to the United States from Japan and now covers over seven million acres in the South. The plant contains chemicals called isoflavones that have been shown to have beneficial effects on bones. One study used three groups of rats to compare a control group with rats that were fed either a low dose or a high dose of isoflavones from kudzu. One of the oucomes examined was bone mineral density in the femur (in grams per square centimeter). Rats were randomly assigned to one of the three groups. The data can be found in kudzu.jmp.

### Question 9

Identify the response variable.

Answer:

# Read in data from excel file  
library(readxl)

## Warning: package 'readxl' was built under R version 3.4.4

df = read\_excel("kudzu.xls", col\_names = TRUE)  
summary(df)

## Treatment BMD   
## Length:45 Min. :0.1980   
## Class :character 1st Qu.:0.2100   
## Mode :character Median :0.2200   
## Mean :0.2233   
## 3rd Qu.:0.2320   
## Max. :0.2670

The response variable is BMD, or bone mineral density, of rats.

### Question 10

Identify the factors (and levels) in the experiment.

Answer:

# Display factor and levels of variables  
levels(factor(df$Treatment))

## [1] "Control" "HighDose" "LowDose"

The variable Treatment is a a factor which has three levels: control, low dosage and high dosage.

### Question 11

How many treatments are included in the experiment?

Answer:

# Display frequency table of treatment variable  
table(df$Treatment)

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

There are observations in total, each in the control group, high dosage group and low dosage group.

### Question 12

What type of experimental design is employed?

Answer: In this study, the type of experimental design is completely randomized where the treatment groups are balanced.

### Question 13

Compute the mean, standard deviation and sample size for each treatment group and put the results into a table. Remember to include the units of measurement.

Answer:

# Create table of mean, sd, and sample size for each treatment group  
df %>% group\_by(Treatment) %>% summarize("mean, in g/cm^2" = mean(BMD),   
 "sd, in g/cm^2" = sd(BMD),   
 "sample size" = n())

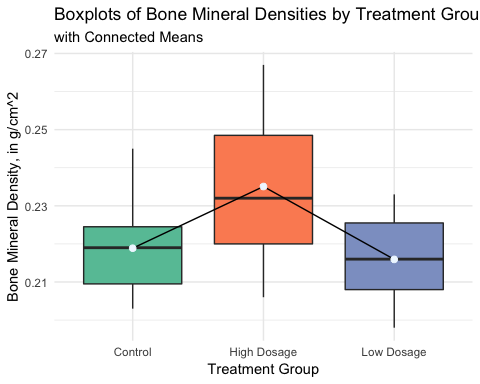
## # A tibble: 3 x 4  
## Treatment `mean, in g/cm^2` `sd, in g/cm^2` `sample size`  
## <chr> <dbl> <dbl> <int>  
## 1 Control 0.219 0.0116 15  
## 2 HighDose 0.235 0.0188 15  
## 3 LowDose 0.216 0.0115 15

### Question 14

Construct side-by-side box plots with connected means. Describe what you see.

Answer:

# Create side-by-size box plots with connected means  
library(RColorBrewer)  
ggplot(df, aes(x = Treatment, y = BMD, fill = Treatment)) +   
 geom\_boxplot() +   
 scale\_fill\_brewer(palette = "Set2") +   
 scale\_x\_discrete(labels = c("Control", "High Dosage", "Low Dosage")) +   
 stat\_summary(fun.y = mean, geom = "line", aes(group = 1),   
 lwd = 0.5, col = "black") +   
 stat\_summary(fun.y = mean, geom = "point",   
 pch = 19, size = 2, col = "aliceblue") +   
 labs(x = "Treatment Group", y = "Bone Mineral Density, in g/cm^2",  
 title = "Boxplots of Bone Mineral Densities by Treatment Group",  
 subtitle = "with Connected Means") +   
 theme\_minimal() + theme(legend.position = "none")



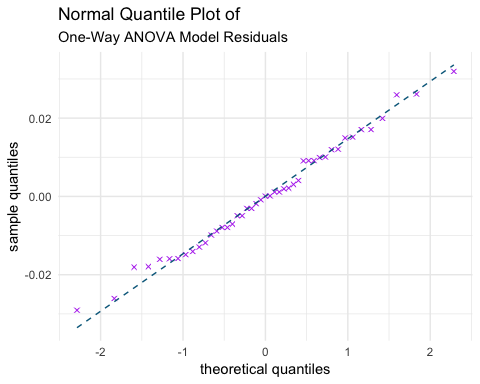
The mean bone mineral density is highest for the rats that received a high dosage of isoflavone. The mean bone mineral density is lowest for the rats that received a low dosage of isoflavones. Furthermore, the spread of the densities is greatest for the high dosage group and smallest for the low dosage group. The control group appears to come in the middle for both the mean density and spread of values.

### Question 15

Are the one-way ANOVA model assumptions satisfied? Justify your answer.

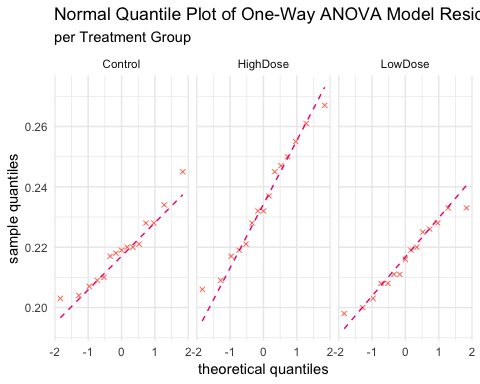
Answer: The one-way ANOVA model assumptions are: independent observations, errors are distributed with mean and standard deviation , and constant variance amongst treatment groups. First, the observations are independent from each other because the bone mineral density measurements come from different rats. Next, to check if errors are , check the normal quantile plot of residuals.

# Display normal quantile plot of residuals  
df %>% group\_by(Treatment) %>% summarize(mean(BMD)) %>%   
 left\_join(df, ., by = "Treatment") %>%   
 mutate("resids" = BMD - `mean(BMD)`) %>%  
 ggplot(aes(sample = resids)) +   
 stat\_qq(pch = 4, color = "darkorchid2") +  
 stat\_qq\_line(linetype = "dashed", color = "deepskyblue4") +   
 labs(x = "theoretical quantiles",   
 y = "sample quantiles",   
 title = "Normal Quantile Plot of",   
 subtitle = "One-Way ANOVA Model Residuals") +   
 theme\_minimal()



According to the plot above, the errors, or residuals, are distributed with mean and constant variance. This assumption is satisfied. Lastly, to check if there’s constant variance amongst the treatment groups, create normal quantile plot of residuals for each individual treatment group.

# Display normal quantile plot of residuals by treatment group  
df %>% ggplot(aes(sample=BMD)) +  
 facet\_grid(~ Treatment) +  
 stat\_qq(pch = 4, color = "coral1") +  
 stat\_qq\_line(linetype = "dashed", color = "deeppink2") +   
 labs(x = "theoretical quantiles",   
 y = "sample quantiles",   
 title = "Normal Quantile Plot of One-Way ANOVA Model Residuals",   
 subtitle = "per Treatment Group") +   
 theme\_minimal()

 According to the plot above, the distribution for all three treatment groups’ normal quantile is slightly heavy tailed. This last assumption is fulfilled with some leeway.

### Question 16

Run a one-way ANOVA model and discuss your results. (Let ; remember to include your hypotheses and identify the test statistic, degrees of freedom and -value.)

Answer:

# Perform one-way ANOVA model  
summary(aov(BMD ~ Treatment, df))

## 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

Let the null hypothesis be that

In other words, that the mean bone mineral density for all three treatment groups are the same. Let the alternative hypothesis be that at least two of the three means are different. Then at the significance level of , the test statistic calculated is , with degrees of freedom and . Using this test statistic, the -value, or probability that , is , which is strictly less than . Therefore the null hypothesis is rejected and it is true that at least two means are different.

### Question 17

Use Tukey’s multiple-comparisons method to compare the three groups. Include the visual results for the Tukey method. Which groups (if any) have significantly different means?

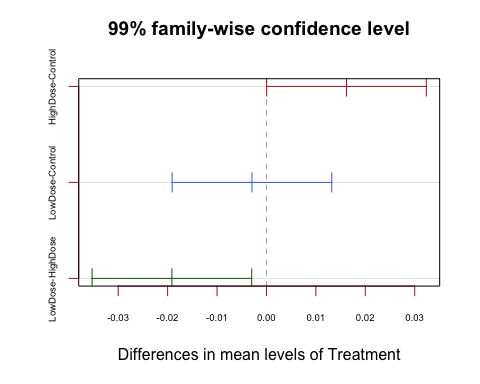
Answer:

# Save ANOVA model  
anova\_model = aov(BMD ~ Treatment, df)  
# Perform Tukey's muliple-comparisons procedure  
TukeyHSD(anova\_model, conf.level = 0.99)

## Tukey multiple comparisons of means  
## 99% family-wise confidence level  
##   
## Fit: aov(formula = BMD ~ Treatment, data = df)  
##   
## $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

The confidence intervals are shown below for each of these comparisons.

# Plot confidence intervals  
plot(TukeyHSD(anova\_model, conf.level = 0.99), cex.axis = 0.6,   
 col = c("firebrick", "royalblue2", "darkgreen"))



It appears to be that the high dosage treatment group has significantly different means than the other two treatment groups. A zero *does not* appear in the confidence interval for the pair high dosage/control but let’s plot this using ggplot for better visualizations just so it is clear to the eye.

(Credit: <https://stackoverflow.com/questions/33644034/how-to-visualize-pairwise-comparisons-with-ggplot2?rq=1>)

# Store TukeyHSD output into dataframe  
tky = as.data.frame(TukeyHSD(anova\_model)$Treatment)  
# Create column for x labels  
tky$pair = rownames(tky)  
# Plot pairwise TukeyHSD comparisons and color by significance level  
ggplot(tky, aes(colour=cut(`p adj`, c(0, 0.01, 0.05, 1),   
 label=c("p < 0.01","p < 0.05",  
 "Not significant")))) +  
 geom\_hline(yintercept=0, lty="11", colour="black") +  
 geom\_errorbar(aes(pair, ymin=lwr, ymax=upr), width=0.2) +  
 geom\_point(aes(pair, diff), col = "black") +  
 scale\_color\_brewer(palette = "Dark2") +   
 labs(x = "",  
 y = "Differences in mean labels of Treatment",  
 title = "99% family-wise confidence level",   
 colour= "") +  
 coord\_flip() +   
 theme\_minimal() +   
 theme(legend.position = "bottom")



The above observation is now clearer than before. Since does not appear in the confidence intervals for the comparisons with the high dosage treatment group, it can be said that this treatment group has a significantly different mean for bone mineral density than the other two treatment group. In fact, the low dosage / high dosage group and high dosage / control group have significantly different means.