CSC 423 - Data Analysis Course Project

Jasmine Dumas, Samantha Stanley, Christopher Lee

November 9, 2015

Preliminary Analysis

1. **Data source**: copy/paste into a text editor to save as a .txt file and save in the same folder as current document, brain.Rmd  
   Source: <http://lib.stat.cmu.edu/DASL/Datafiles/Brainsize.html>

setwd("/Users/jasminedumas/Desktop/depaul/CSC423")  
brain <- read.table("braindata.txt", header = TRUE, sep = "", na.strings="¥")  
brain <- na.omit(brain) # remove 3 rows of missing data  
head(brain)

## Gender FSIQ VIQ PIQ Weight Height MRI\_Count  
## 1 Female 133 132 124 118 64.5 816932  
## 3 Male 139 123 150 143 73.3 1038437  
## 4 Male 133 129 128 172 68.8 965353  
## 5 Female 137 132 134 147 65.0 951545  
## 6 Female 99 90 110 146 69.0 928799  
## 7 Female 138 136 131 138 64.5 991305

ncol(brain) # number of columns

## [1] 7

names(brain)

## [1] "Gender" "FSIQ" "VIQ" "PIQ" "Weight" "Height"   
## [7] "MRI\_Count"

nrow(brain) # number of rows

## [1] 38

1. **Research Question**: Are the size and weight of your brain indicators of your mental capacity? In this study by Willerman et al. (1991) the researchers use Magnetic Resonance Imaging (MRI) to determine the brain size of the subjects. The researchers take into account gender and body size to draw conclusions about the connection between brain size and intelligence.
2. **Dependent Variable**: MRI\_Count
3. **Independent Variable**: Gender, FSIQ, VIQ, PIQ, Weight, Height
4. **Quantitative vs. Qualitative**: The quantitative variables are MRI\_Count, FSIQ, VIQ, PIQ, Weight, Height. The qualitative variable is Gender
5. **Data cleansing**: The NA strings were specified in the read.table function and were removed with the na.omit function. The Gender variable will needed to be coded as 0, 1 which R does automatically when assembling the model.
6. **Correlation and Scatter plots**: The only predictor variables that seem to be linear are height and weight. It is expected to see that Weight and Height is correlated.

# Correlation between the repsonse variable and the independent variables (Should be high)  
cor(brain$FSIQ, brain$MRI\_Count)

## [1] 0.3337137

cor(brain$VIQ, brain$MRI\_Count)

## [1] 0.3002791

cor(brain$PIQ, brain$MRI\_Count)

## [1] 0.3777816

cor(brain$Weight, brain$MRI\_Count)

## [1] 0.5133785

cor(brain$Height, brain$MRI\_Count)

## [1] 0.5883772

# Correlation between the independent variables (Should be low)  
cor(brain$FSIQ, brain$VIQ)

## [1] 0.9451143

cor(brain$FSIQ, brain$PIQ)

## [1] 0.9344266

cor(brain$FSIQ, brain$Weight)

## [1] -0.05148285

cor(brain$FSIQ, brain$Height)

## [1] -0.1184478

cor(brain$VIQ, brain$PIQ)

## [1] 0.7760202

cor(brain$VIQ, brain$Weight)

## [1] -0.07608804

cor(brain$VIQ, brain$Height)

## [1] -0.1189765

cor(brain$PIQ, brain$Weight)

## [1] 0.002512154

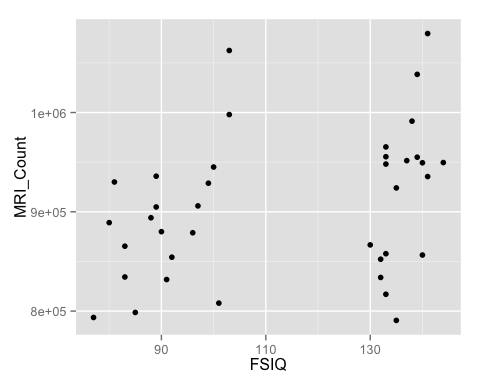
cor(brain$PIQ, brain$Height)

## [1] -0.09315559

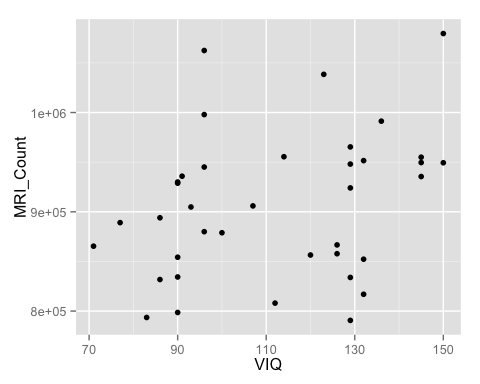
cor(brain$Height, brain$Weight)

## [1] 0.699614

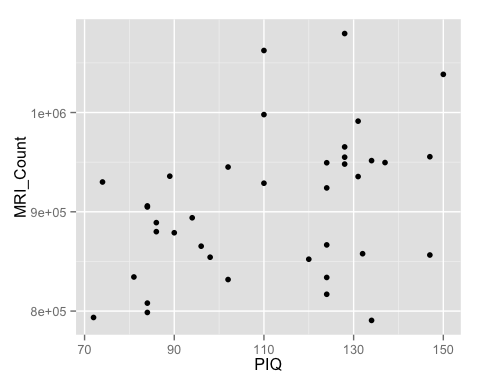
#############################################################################  
library(ggplot2)  
  
ggplot(brain, aes(y = MRI\_Count, x = FSIQ)) + geom\_point()



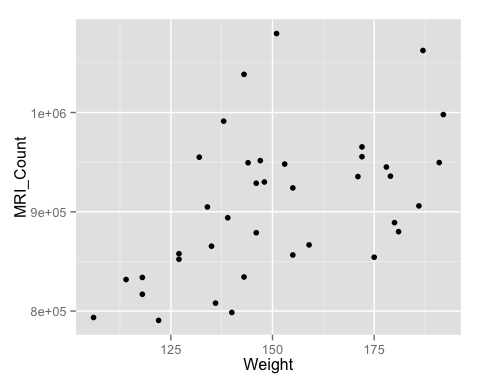
ggplot(brain, aes(y = MRI\_Count, x = VIQ)) + geom\_point()



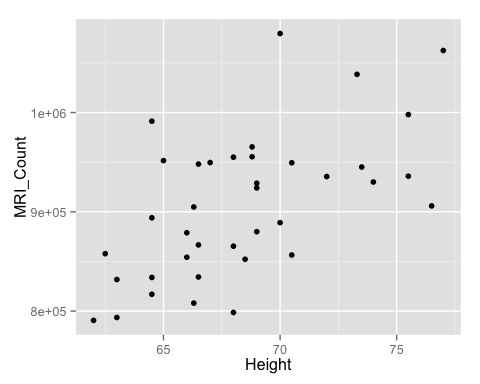
ggplot(brain, aes(y = MRI\_Count, x = PIQ)) + geom\_point()



ggplot(brain, aes(y = MRI\_Count, x = Weight)) + geom\_point()



ggplot(brain, aes(y = MRI\_Count, x = Height)) + geom\_point()



1. **Checking of assumptions**
   * Slight sign of heteroscedasticity was identified and transformation was conducted to address the issue
   * There was no sign of autocorrelation on both original and transformed model
   * No appararent outlier exists on original dataset
2. **Checking for interaction terms**
   * No interaction terms were found for both full and optimized model
   * Both full and optimized model with interaction terms did not have any parameter that was statistically significant
   * Also, adding interaction terms lowered adjusted R squared value
3. **Checking for higher order models (quadratic?)**
   * No higher order models were found for both full and optimized model
   * Full model had only 1 parameter (weight) with statistical significance at higher order which was shown as stasticially insignificant in linear model.
   * Optimized model did not have any parameters with higher order that was statistically significant
   * Further examination is needed on weight variable
4. **Examination of residuals**
   * 3 influencial data points were identified (#6, #8 , #13) and therefore identified outlier (record #6) was removed
   * No extreme outlier was identified on new model
   * There was no sign of autocorrelation based on Durbin Watson Test
5. **Transformations if appropriate**
   * Based on model result, both log and inverse transformation was attempted.
   * Inverse model improved model output with higher stasticial parameters’ significance and improved adjusted r squared value