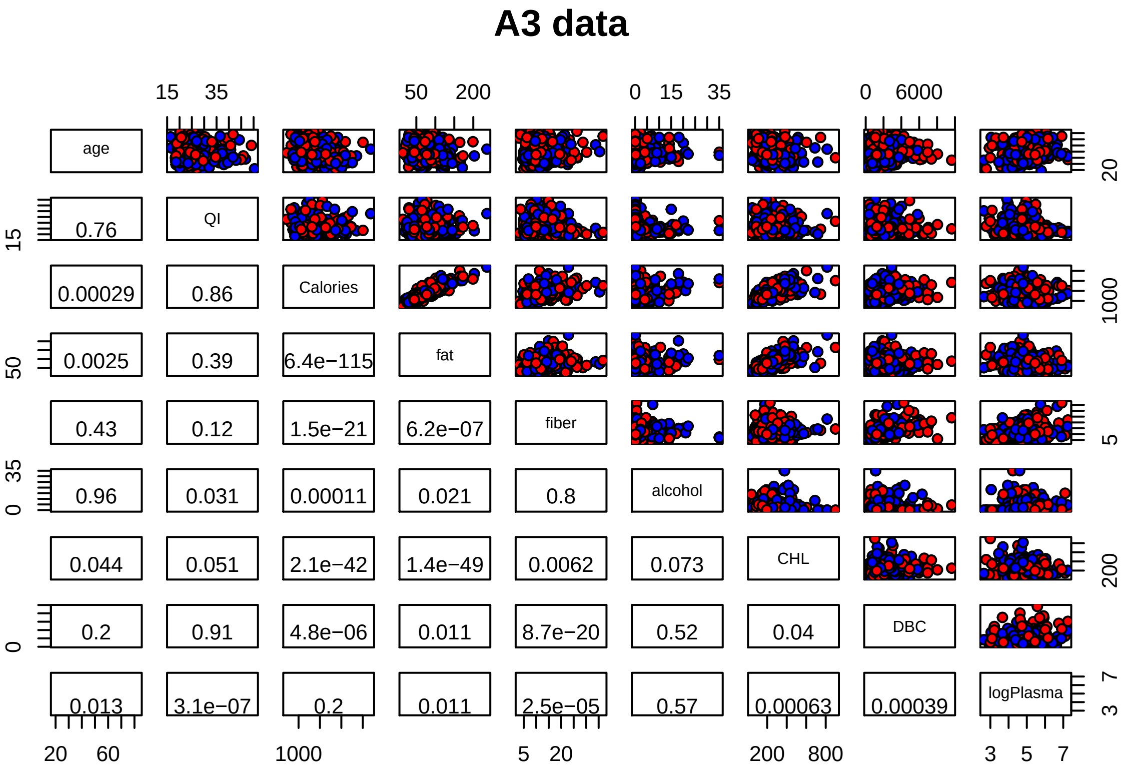
A3: Determinants of Plasma Level

Last name: Akad
First name: Dogan
Student ID: 1001386083
Course section: STA302H1F-L0101

Dec. 5, 2016

Q1: Which variables have linear relationship?



Comments on the plots: Lower part of the plot represents the p values and the upper part represents the scatter plots. Usually, if a p value is less than 0.01 we would say that we have strong evidence, and if the p value is in interval (0.01, 0.05) we say we have moderate evidence. Looking at their p values, pairs of variables with strong evidence of nonzero correlation are:

CHL and logPlasma (p-value = 0.0006)

DBC and logPlasma (p-value = 0.0004)

Fat and age(p-value = 0.0025)

Fat and fiber(p-value = 6.2e-07)

Fat and CHL(p-value = 1.4e-49)

Calories and fat(p-value = 6.4e-115)

Calories and fiber(p-value = 1.5e-21)

Calories and alcohol (p-value = 0.00011)

Calories and CHL(p-value = 2.1e-42)

Calories and DBC (p-value = 4.8e-06)

Calories and age (p-value = 0.00029)

logPlasma and QI (p-value = 3.1e-07)

Fiber and CHL (p-value = 0.0062)

Fiber and DBC (p-value = 8.7e-20)

Fiber and logPlasma (p-value = 2.5e-05)

Pairs of variables with moderate evidence of nonzero correlation are:

Fat and DBC (p-value = 0.011)

Fat and logPlasma (p-value = 0.011)

Fat and alcohol (p-value = 0.021)

Age and CHL (p-value = 0.044)

Age and logPlasma (p-value = 0.013)

QI and alcohol (p-value = 0.031)

CHL and DBC (p-value = 0.04)

Q2: Fit regression equations.

Coefficient of Calories for (1) : -8.586e-05

p-value for regression (1) : 0.201

Coefficient of Calories for (2) : 0.0004

p-value for regression (2) : 0.021

Coefficient of Calories for (3) : -8.252e-05

p-value for regression (3) : 0.201

Comparing the coefficients:

Coefficient for Regression (2) > Coefficient for Regression (3) > Coefficient for Regression (1)

Comparing the p-values:

p-value for regression (1) = p-value for regression (3) > p-value for regression (2)

Since calories and fat have a very strong evidence of nonzero correlation, they have a high degree of multicollinearity among (2) and (3) which results in different coefficients for calories. Looking at the plot in the first question, comparing their p values, we can say that there is a weak evidence of nonzero correlation between calories and QI which are the predictors for (3). The difference in p values can be explained by the different evidences of nonzero correlation. Since there is a weak evidence of nonzero correlation between calories and QI, the p-value for calories in (3) is similar to the one in (1). Since there is a strong evidence of nonzero correlation between calories and fat, its p value is different from (3).

Q3: Important Variables

We can determine by looking at the p values of the predictors. Lower p value means the better the predictor. If p value is less than 0.05 we reject the fact that there is no linear relationship between the predictor and response variable. So the p value of the predictors which are less than 0.05 can be considered important. According to that, QI(p value: 2.03e-06) is the most important predictor. The other important predictors are fiber(0.02), gender(p value: 0.04), smoke(p value: 0.03) and vitamin(p value: 0.05).

Q4: Stepwise Regression

After doing stepwise regression, final model includes QI, fiber, calories, smoke, vitamin, DBC, gender and age as predictor and logPlasma as response. No, the independent variables in the final model are not the important variables that we got in previous question. Some of them are the same, but not all of them.

Q5: Source R code

```r
# ---------> complete and run the following code for this assignment  <-------
#
#
# R code for STA302 or STA1001H1F assignment 3
# copyright by Dogan Akad
# date: Dec. 5, 2016
#

## Load in the data set
a3 = read.table("/Users/doganakad/Desktop/uoft/first semester/STA302/Assignments/A3/a3data.txt",sep="",header=T)
str(a3)
is.factor(a3$gender)
is.factor(a3$smoke)
a3$smoke = as.factor(a3$smoke)
is.factor(a3$smoke)
a3$logPlasma = log(a3$plasma)
## ==>Q1
## subtract gender,smoke,vitamin,plasma
a3_subset = a3[, -c(2,3,5,12)]
panel.pearson <- function(x, y,...) {
  horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
  vertical <- (par("usr")[3] + par("usr")[4]) / 2.8;
  text(horizontal, vertical, format(abs(cor.test(x,y)$p.value), digits=2))
}
pairs(a3_subset, main = "A3 data", pch = 21, bg = c("red", "blue"), lower.panel = panel.pearson)
## ==> Q2 fit regression lines
## Regression line with calories only
m0 = lm(a3$logPlasma ~ a3$Calories)
## Regression line with calories and fat
m1 = lm(a3$logPlasma ~ a3$Calories + a3$fat)
## Regression line with calories and QI
m2 = lm(a3$logPlasma ~ a3$Calories + a3$QI)
summary(m0)
summary(m1)
summary(m2)
##==> Q3
# Multiple regression line with all possible predictor variables
m3 = lm(a3$logPlasma ~ a3$age + a3$gender + a3$smoke + a3$QI + a3$Vitamin + a3$Calories + a3$fat
        + a3$fiber + a3$alcohol + a3$CHL + a3$DBC)
summary(m3)

## Q4
# Regression with no predictors
nullmod = lm (logPlasma~1, data =a3)
# Regression with all the predictors
fullmod = m3
# Stepwise regression applying both ways
bothways = step ( nullmod , scope = list ( lower = formula ( nullmod ),upper = formula ( fullmod )),
direction ="both")
formula (bothways)
```