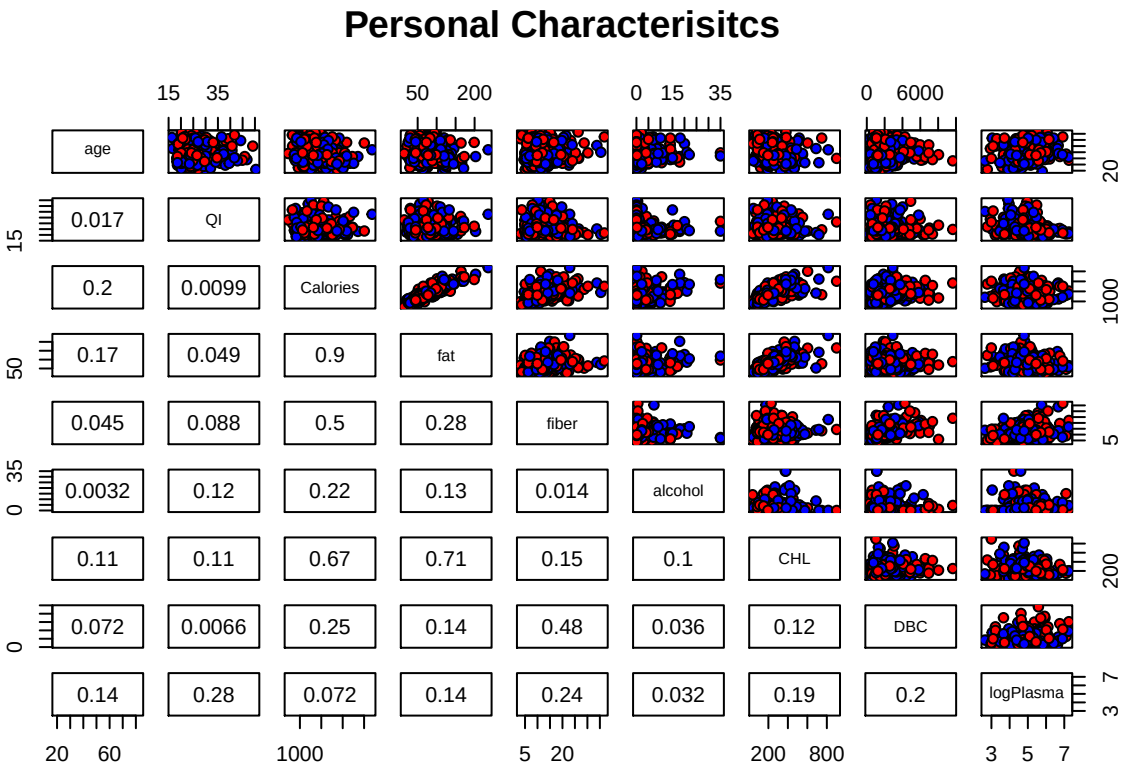
Assignment#3

Last name: Zheng
First name: Siyuan
Student ID: 1000726814
Course section: STA302H1F-L0101

Dec. 4, 2016

Q1



Comments:
Based on the Pearson correlation coefficient indicated on the plot, the Calories-fat(very strong),Calories-CHL,fat-CHL pairs have strong evidence of a linear relationship;the Calories-fiber,fiber-DBC pairs have moderate evidence of a linear relationship;

And to consider logPlasma and other predictors, all of the pairs(i.e. logPlasma and one of the others) have weak or even no evidence of a linear relationship.

Q2

Model	Model Name	Coefficient for Variable Calories	P-value
1st	logPlasma~Calories	-0.0000858597	0.201
2nd	logPlasma~Calories + Fat	0.0003505757	0.02136
3rd	logPlasma~Calories + QI	-8.251896e-05	0.201

Cor(fat,Calories) = 0.9000301
Cor(QI,Calories) = 0.009933543
Cor(QI,logPlasma) = -0.2835995
Cor(Calories, logPlasma) = -0.07222683
Cor(Fat,logPlasma) = -0.1424885

Comparison of β and p-value:
The 1st model has the similar coefficient of calories with the 3rd model ,as well as the corresponding p-values, while 2nd model does not share these two similarities with the former two. The p-value of 2nd model is less than 0.05(moderate evidence to reject H_0) while the 1st and 3rd's ones are greatter than 0.1(no evidence to reject H_0).

Difference that results in difference:
There is no correlation between calories and logPlasma but there is a weak correlation between QI and logPlasma,moreover, there is no correlation between calories and QI, thus 1st and 3rd model will have the similar coefficients and p-values for calories.

There is a very strong correlation between fat and calories, while either of them do not have correlation with logPlasma, when two predictor variables are perfectly correlated, many response functions will lead to the same fitted values for the observations.Since such many different functions provide the same good fit, the coefficient of predictor variables cannot be interpreted as the effect.

Q3

Predictor Variable	p-value	note
age	0.0832	0.05<p<0.1
genderM	0.0414	0.01<p<0.05
smoke1	0.0322	0.01<p<0.05
QI	2.03e-06	<0.01
Vitamin1	0.0463	p<0.05
Calories	0.7094	>0.1
fat	0.8559	>0.1
fiber	0.0165	p<0.05
alcohol	0.9119	>0.1
CHL	0.2220	>0.1
DBC	0.0747	0.05<p<0.1

There is a strong evidence that QI(p-value < 0.01) seems to be the very important variable in predicting the response(logPlasma) while there are some moderate evidences that gender,smoke, vitamin and fiber(0.01<p-value<0.05) are the relatively important variables in predicting the response.

Q4

logPlasma ~ -3.292e-02QI + 3.026e-02fiber + -1.844e-04Calories + -2.560e-01 smoke + 1.625e-01Vitamin + 5.162e-05DBC + -2.786e-01gender + 4.947e-03age

By stepwise regression:
Calories, DBE, age are the three predictor variables that are not included in question 3 while the other 5 are. These three predictor variables seem not to be important in question 3, since their p-value are relatively large(especially for Calories). The large p-values results in failing to reject the null hypothesis which is $\beta = 0$, i.e., coefficients of Calories,DBE, age equal to 0. But this model minimizes AIC, thus from this perspective, these independent variables seem to be important.

Q5

```
#Note: for this part, some line breaks are
#added(or adjusted) from the originial R codes that
#are to be executed, just for better display effect

#Q1
rm(list=ls())
# Also, use control + L to clear the console
# Read data from a3data.txt into R Studio(environment)
a3 = read.table("a3data.txt",sep="",header=T)
# str(a3)
# is.factor(a3$gender)  #TRUE
# is.factor(a3$smoke)   #FALSE
# is.factor(a3$Vitamin) #FALSE
a3$smoke = as.factor(a3$smoke)  # factor smoke
a3$Vitamin = as.factor(a3$Vitamin) # factor Vitamin
a3$logPlasma = log(a3$plasma) # take log of plasma and append to a3
# is.factor(a3$smoke) #TRUE
# is.factor(a3$gender) #TRUE
# is.factor(a3$Vitamin) #TRUE
# head(a3)

#2,3,5,12 represents column of gender, smoke,
#Vitamin and plasma respectively
#get rids of dummy variables and untransformed response
subset_data <- a3[,-c(2,3,5,12)]
# head(subset_data) # works

# let lower panel to indicate the Pearson correlation coefficient
panel.pearson <- function(x, y, ...) {
horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
vertical <- (par("usr")[3] + par("usr")[4]) / 2;
text(horizontal, vertical, format(abs(cor(x,y)), digits=2))
}

# generate the plot and let the lower part
#to indicate correlation coefficient
pairs(subset_data[1:9],main
      = "Personal Characterisitcs",
      pch=21,bg=c("red","blue"),
      lower.panel=panel.pearson)
```

```r
#Q2
rm(list=ls()) # clean the environment
a3 <- read.table("a3data.txt",sep="",header=T)
a3$smoke = as.factor(a3$smoke)  # factor smoke
a3$Vitamin = as.factor(a3$Vitamin) # factor Vitamin
a3$logPlasma = log(a3$plasma) # take log of plasma and append to a3
head(a3)   # works
caloriesFit <- lm(a3$logPlasma~a3$Calories)  # SLR
caloriesFatFit <- lm(a3$logPlasma~a3$Calories+a3$fat) # MLR
caloriesQIFit <- lm(a3$logPlasma~a3$Calories+a3$QI)  # Another MLR
caloriesFit$coef # get coefficients
caloriesFatFit$coef # get coefficients
caloriesQIFit$coef # get coefficients
# get the table to get p-value of coefficient t-test
summary(caloriesFit)
# get the table to get p-value of coefficient t-test
summary(caloriesFatFit)
 # get the table to get p-value of coefficient t-test
summary(caloriesQIFit)
cor(a3$fat,a3$Calories) # get correlation
cor(a3$QI,a3$Calories) # get correlation
cor(a3$QI,a3$logPlasma) # get correlation
cor(a3$Calories,a3$logPlasma) # get correlation
cor(a3$fat,a3$logPlasma) # get correlation
```

```r
#Q3
rm(list=ls()) # clean the environment
a3 <- read.table("a3data.txt",sep="",header=T)
a3$smoke = as.factor(a3$smoke)  # factor smoke
a3$Vitamin = as.factor(a3$Vitamin) # factor vitamin
# take log of plasma and append a column to a3
a3$logPlasma = log(a3$plasma)
head(a3) # check a3
# Produces the MLR
fit <- lm(logPlasma~age+gender+smoke+
QI+Vitamin+Calories+fat+fiber+alcohol+CHL+DBC,data=a3)
# get the p-value of t-test of coefficients from the table
summary(fit)
# Using anova test should give the same information
# just double check the answer
```

```r
install.packages("car")
library(car)
Anova(fit,type=3)




#Q4
rm(list=ls()) # clean environment
a3 <- read.table("a3data.txt",sep="",header=T)
a3$smoke = as.factor(a3$smoke)  # factor smoke
a3$Vitamin = as.factor(a3$Vitamin) # factor vitamin
# take log of plasma and add a column to a3
a3$logPlasma = log(a3$plasma)
head(a3) # just several rows of a3
str(a3)
names(a3) # get the columns headers of a3
is.factor(a3$smoke)  #TRUE
is.factor(a3$Vitamin) #TRUE
is.factor(a3$gender) #TRUE

# no predictor
nullmod <- lm(logPlasma~1,data=a3)

# full predictor
fullmod <- lm(logPlasma~age+gender+smoke+QI+
Vitamin+Calories+fat+fiber+alcohol+CHL+DBC,data=a3)

# stepwise method : apply both directions method
bothways = step ( nullmod , scope = list
( lower = formula ( nullmod ),upper = formula ( fullmod )),
direction ="both")
formula ( bothways )
# get the coefficients
bothways
```