A3: Determinants of Plasma Level

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA302H1F-L0101

Nov. 30, 2016

Q1: Look at the pairwise correlations and scatterplots. For which pairs of variables is there strong evidence of a linear relationship ? For which pairs of variables is there moderate evidence of a linear relationship ? Note that the untransformed response and non-quantitative variables are not considered. (Consider the pairwise correlation between logPlasma and all predictors, and the pairwise correlation between any two predictors)

Solution:

After observing the pairwise corrleations and scatterplots I can conclude the following things:

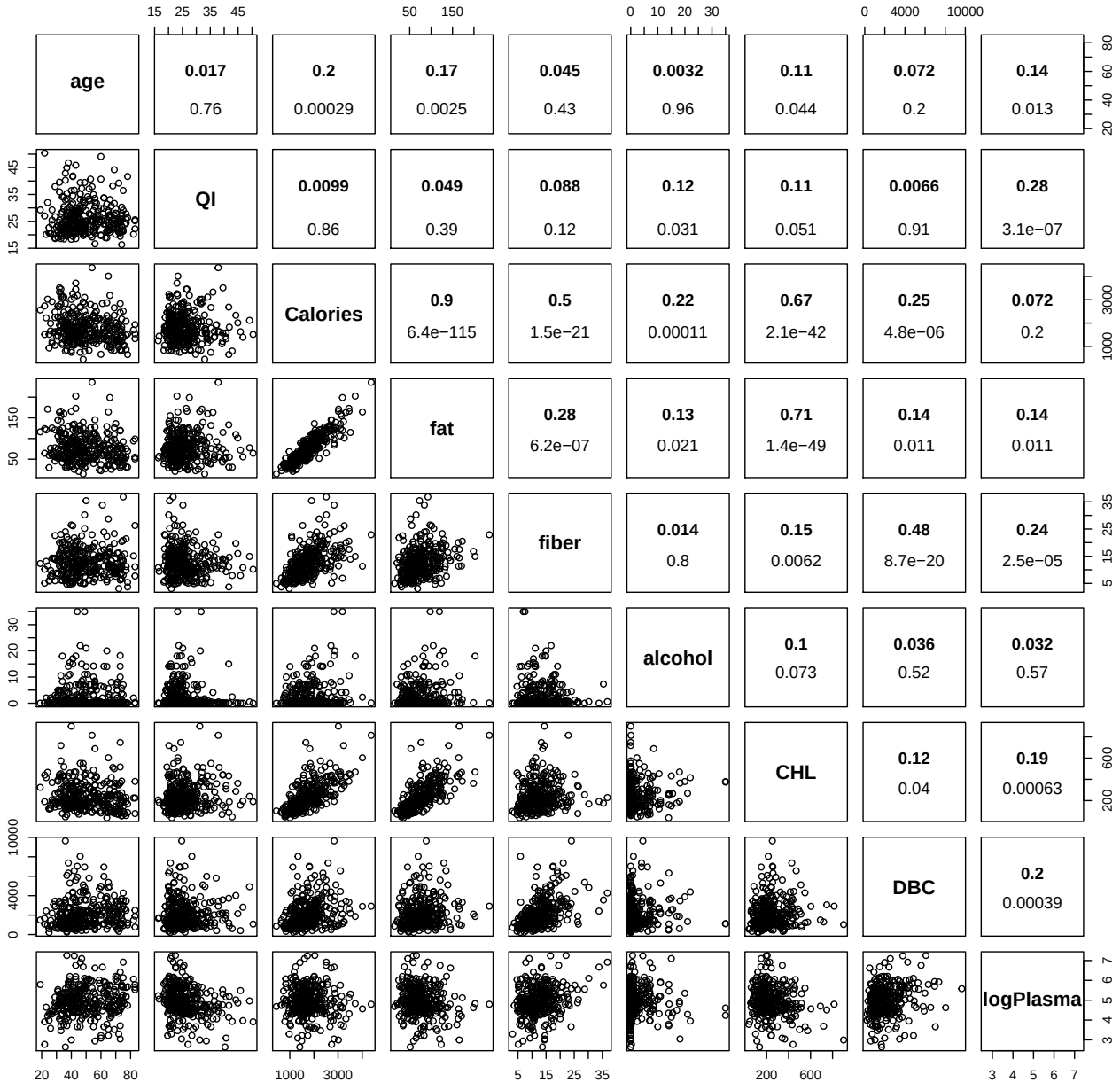
Strong Evidence:

The pairs which stood out to portray strong evidence of a linear relationship were the pairs: (Calories, alcohol); (age, Calories); (age, fat); (QI, log(Plasma)); (Calories, fat); (Calories, fiber); (fat, fiber); (Calories, DBC); (Calories, CHL); (DBC, log(Plasma)); (CHL, log(Plasma)); (fiber, log(Plasma)); (fat, CHL); (fiber, CHL); (fiber, DBC). These pairs produced p-values from the correlation tests in which each of the pairs resulted with values less than 1%, hence therefore strong evidence of a linear relationship.

Moderate Evidence:

The pairs which stood out to portray moderate evidence of a linear relationship were the pairs: (fat, log(Plasma)); (age, CHL); (QI, alcohol); (fat, DBC); (fat, alcohol); (age, log(Plasma)); (fat, log(plasma)); (CHL, DBC). These pairs produced p-values from the correlation tests in which each of the pairs resulted with values less than 5% and greater than 1%.

PS: The pairwise correlation and scatter plots have been placed on the next page since it appears well on its own.


All Predictors
age
QI
Calories
fat
fiber
alcohol
CHL
DBC
logPlasma
0.017
0.76
0.2
0.00029
0.17
0.0025
0.045
0.43
0.0032
0.96
0.11
0.044
0.072
0.2
0.14
0.013
0.0099
0.86
0.049
0.39
0.088
0.12
0.12
0.031
0.11
0.051
0.0066
0.91
0.28
3.1e−07
0.9
6.4e−115
0.5
1.5e−21
0.22
0.00011
0.67
2.1e−42
0.25
4.8e−06
0.072
0.2
0.28
6.2e−07
0.13
0.021
0.71
1.4e−49
0.14
0.011
0.14
0.011
0.014
0.8
0.15
0.0062
0.48
8.7e−20
0.24
2.5e−05
0.1
0.073
0.036
0.52
0.032
0.57
0.12
0.04
0.19
0.00063
0.2
0.00039

Q2: Fit the three regression equations with (1) calories only, (2) calories with fat, and (3) calories and Quetelet index as the predictor variable(s) and log of plasma as the dependent variable. For these regressions compare the coefficient of calories and the p-value for the two-sided test with null hypothesis that this coefficient is 0. What is the difference between regressions (2) and (3) that results in different coefficients and p-values for calories?

Solution:

(1) Calories.

$$\widehat{log(plasma)} = 5.10684631 - 0.00008586\text{Calories}$$

(2) Calories with fat.

$$\widehat{log(plasma)} = 5.0273755 + 0.0003506\text{Calories} - 0.0090997\text{fat}$$

(3) Calories and Quetelet index.

$$\widehat{log(plasma)} = 6.02953937 - 0.00008252\text{Calories} - 0.03550306\text{QI}$$

From the three computed regression equations (1 to 3) we note that the coefficient of calories is -0.00008586, 0.0003506 and - 0.00008252 respectively.

The coefficient of (2) is positive, whilst (1) and (3) have negative coefficients which are also very close to each other. If we compare the p-values, we found (1) 0.201, (2) 0.02136 and (3) 0.201. We also notice that the p-values of (1) and (3) are basically the same while the p-value for (2) is much smaller.

Given the null hypothesis that the coefficient of calories is zero at 5% signifance level. We can conclude this:

- (1) $0.201 > 0.05$, we fail to reject the null hypothesis.
- (2) $0.02136 < 0.05$, we reject the null hypothesis.
- (3) $0.201 > 0.05$, we fail to reject the null hypothesis.

The difference in the coefficient value and p-values for calories in (2) and (3) are mainly resulted by the multicollinearity. We can also note that the correlation differences calculated in Question 1, between (2) and (3) have also caused the differences in the coefficients and p-values of calories. Correlation between Calories and Fat was 0.9 whilst correlation between Calories and QI was only 0.0099, the higher the correlation means it's more likely that one predictor can be expressed with another and also more likely have a different correlation and p-value.

Q3: A commonly asked question is which variables are important in predicting the response, log of plasma. Fit the regression with all 11 possible predictor variables. From the R output, which variables seem to be important predictors of the log of plasma?

Solution:

After fitting the regression with the full model of all 11 possible predictor variables. We notice that the p-values for the coefficients Gender, Smoke, QI, Vitamin and Fiber are 0.0414, 0.0322, 2.03e-06 (0.00000203), 0.0463, 0.0165 respectively. These are the only values which are less than 0.05, which is at the 5% significance level.

We therefore proceed to reject the null hypothesis stating that the coefficient of these variables equal to 0 and fail to reject the others, by this we can also conclude that Gender, Smoke, QI, Vitamin and Fiber are variables which seem ot be important predictors of the log of plasma.

```
##
## Call:
## lm(formula = logPlasma ~ age + factor(gender) + smoke + QI +
##     Vitamin + Calories + fat + fiber + alcohol + CHL + DBC, data = a3data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91040 -0.36341  0.02106  0.40162  1.98162
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.376e+00  2.758e-01  19.490  < 2e-16 ***
## age              5.131e-03  2.952e-03   1.738   0.0832 .
## factor(gender)M -2.585e-01  1.262e-01  -2.048   0.0414 *
## smoke           -2.501e-01  1.162e-01  -2.152   0.0322 *
## QI              -3.179e-02  6.562e-03  -4.844 2.03e-06 ***
## Vitamin          1.616e-01  8.080e-02   2.000   0.0463 *
## Calories        -7.220e-05  1.935e-04  -0.373   0.7094
## fat             -5.683e-04  3.126e-03  -0.182   0.8559
## fiber            2.643e-02  1.097e-02   2.410   0.0165 *
## alcohol          9.604e-04  8.671e-03   0.111   0.9119
## CHL             -5.221e-04  4.267e-04  -1.224   0.2220
## DBC              5.327e-05  2.978e-05   1.789   0.0747 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6738 on 303 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2029
## F-statistic: 8.265 on 11 and 303 DF,  p-value: 1.101e-12
```

Q4: One widely-used method to find a parsimonious model is to apply stepwise procedure. In backward elimination, it starts with all the predictors in the model, remove one predictor at a time to give a smaller AIC. In forward selection, it just reverses the backward method, it starts with no variables in the model, adding one predictor at time by AIC as criteria until no more predictor can be added to produce smaller AIC value. Stepwise regression alternates forwards steps with backwards steps. The idea is to end up with a model where no variables are redundant given the other variables in the model.

Question : What model does stepwise regression produce for this data ? Are the independent variables in the final model that seemed to be important in the previous question?

Solution:

After the stepwise regression process the final model results to be:

$$\widehat{log(plasma)} = 5.40517671 - 0.03291517\text{QI} + 0.03026223\text{fiber} - 0.00018437\text{calories} - 0.25596945\text{smoke} +$$

$$0.16254534\text{Vitamin} + 0.00005162\text{DBC} - 0.27856441\text{gender} + 0.00494748\text{age}$$

Which is logPlasma ~ QI + fiber + Calories + smoke + Vitamin + DBC + gender + age.

We also notice that Gender, Smoke, QI, Vitamin and fiber are included as independent variables in the final model which seemed to be important in the previous question, however Calories, DBC and age were not but still included in this final model.

Q5: Source R code

```
# R code for STA302 or STA1001H1F Assignment 3
# copyright by Sunil M Khambaita
# date: November 30th, 2016
#

# Loading in the a3 data set
a3data = read.table("a3data.txt", sep = "", header = T)

# Performing log transformation on the plasma column
a3data$logPlasma = log(a3data$plasma)
str(a3data)


## Q1: Pairwise Correlation and Scatterplotting

# Finding a subset of logPlasma and all the predictors
all_predictors = (a3data[,c(1, 4, 6:11, 13)])

# change the upper pannel with with Peareson 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), pos = 3, font = 2, cex = 1.2)
  vertical <- (par("usr")[3] + par("usr")[4]) / 3;
  text(horizontal, vertical, format(cor.test(x,y)$p.value, digits=2), cex = 1.2)
}

# pairwise plot of variables in all_predictors data
pairs(all_predictors, main = "All Predictors", pch = 21, upper.panel=panel.pearson, font.labels = 2)

## Q2: Finding three regression equations

# Fitting a linear model mod1 which predicts log(Plasma) from Calories.
mod1 = lm(logPlasma ~ Calories, data = a3data)
mod1
summary(mod1)

# Fitting a linear model mod2 which predicts log(Plasma) from Calories with fat.
mod2 = lm(logPlasma ~ Calories + fat, data = a3data)
mod2
summary(mod2)

# Fitting a linear model mod3 which predicts log(Plasma) from Calories with QI.
mod3 = lm(logPlasma ~ Calories + QI, data = a3data)
mod3
summary(mod3)

## Q3: Regression equation with all predictors.

# Fitting a linear model mod4 which predicts log(Plasma) with all the predictors.
mod4 = lm(logPlasma ~ age + factor(gender) + smoke + QI + Vitamin + Calories + fat + fiber +
            alcohol + CHL + DBC, data = a3data)
mod4
summary(mod4)

# Fitting a linear model nullmod which predicts log(Plasma) by itself.
```

```r
nullmod = lm(logPlasma ~ 1, data = a3data)

# Fitting a linear model fullmod which predicts log(Plasma) with all predictors.
fullmod = lm(logPlasma ~ age + factor(gender) + smoke + QI + Vitamin + Calories + fat + fiber +
               alcohol + CHL + DBC, data = a3data)

# stepwise selection method using AIC
bothways = step(nullmod ,scope = list(lower = formula(nullmod), upper = formula(fullmod)),
                direction="both")
formula(bothways)
```