STA302/1001HF: Methods of Data Analysis I **Fall 2016**

Assignment 3 : Determinants of Plasma Level

Out: Nov. 15, 2016 *Due: Dec. 05, 2016*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 11 :00pm, Dec. 05, 2016. Submit your solution as instructed by Crowdmark, namely, one pdf file for each question.

Most problems on this assignment require using R. Your turned in solutions should not include all of the R output and graphs that you will produce. Write your solutions and include only sparingly R output or graphs when necessary to support a point you are making in response to the problem question.

Late assignments will be subject to a deduction of 4% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. For this assignment, no template for the solution. You are free to have your own style and be creative on presenting your solution. About 10% of the marks for this assignment will be for presentation.

Data

Observational studies have suggested that low dietary intake or low plasma concentration of retinol beta-carotene and other carotenoids might be associated with increased risk of developing certain types of cancer. A cross-sectional study was designed to investigate the relationship between personal characteristics and dietary factors, and resulting plasma concentrations of beta-carotene. Study subjects were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.

This data set consists of 315 observations on 12 variable. The variables are :

- Age : age of the 315 subject (years).
- Gender : M=Male, F=Female.
- Smoking status : 1=smoker, 0=non-smoker.
- Quetelet index : weight/height^2
- Vitamin use : 1= regular user, 0=not regular user.
- Calories : number of calories consumed per day.
- Fat : grams of fat consumed per day.
- Fiber : grams of fiber consumed per day.
- alcohol : number of alcoholic drinks consumed per week.
- CHL : cholesterol consumed (mg per day).
- DBC : dietary beta-carotene consumed (mcg per day).
- Plasma : plasma beta-carotene (mg/ml).

In **R**, to include a variable as dummy variable in your regression, you can use **factor()** or you put the variable into factor variable before using it. For the data posted for this assignment,

once your load in the data in **R**, you could check the type of variable in the data set by using **str()** and check if a variable is factor variable or not by **is.factor()**.

Run the following code :

```
a3 <- read.table("a3data.txt",sep="",header=T)
str(a3)
is.factor(a3$gender)                    # should see TRUE
is.factor(a3$smoke)                     # It is not a factor variable
a3$smoke = as.factor(a3$smoke)          # convert smoke to a factor variable
is.factor(a3$smoke)                     # what do you see?
```

Listing 1 – R code: create a factor variable from a categorical variable

It is also necessary to take the log transformation of plasma variable in the data. You do not need to verify that this transformation is necessary, just work with the transformed value as your response variable. The following code will add one more column in the data to have the transformed response variable.

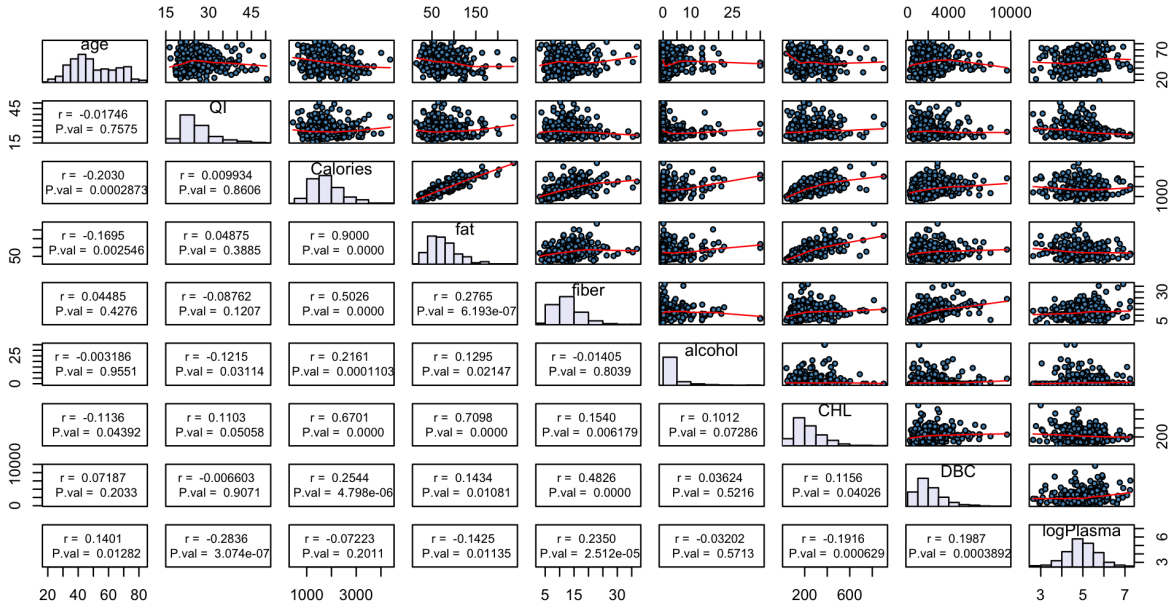
```
a3$logPlasma = log(a3$plasma)
head(a3)
```

Listing 2 – R code: creating transformed response variable

Questions (full mark : 20 points)

1. (5 points) Look at the pairwise correlations and scatterplots. For which paris 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 analysis between logPlasma and all predictors, and the pairwise correlation analysis between any two predictors)

Solution :



From above pairwise scatter and correlation plot, we have

(1) The following predictors have **strong** evidence of relationship with response (log of plasma) variable :

— Quetelet Index : r = -0.2836, $p < 0.0001$.

— fiber : r= 0.2350, $p < 0.0001$
— cholesterol : r=-0.1916, p=0.0006
— dietary beta-carotene : r=0.1987, p=0.0004

(2) The following predictors have **moderate** evidence of relationship with response (log of plasma) variable :
— age : r = 0.1401, $p = 0.0128$.
— fat : r= -0.1425, p=0.0114

(3) The following predictors have **strong** evidence of relationship :
— age and calories : r = -0.2030, p=0.0003
— age and fat : r=-0.1695, p=0.0025
— calories and fat : r=0.90, $p < 0.0001$
— calories and fiber : r=0.5026, $p < 0.0001$
— calories and alcohol : r=0.2161, $p = 0.0001$
— calories and cholesterol : r=0.6701, $p < 0.0001$
— calories and dietary beta-carotene : r=0.2544, $p < 0.0001$
— fat and fiber consumed : r=0.2765, $p < 0.0001$
— fat and cholesterol consumed : r=0.7098, $p < 0.0001$
— fiber and cholesterol consumed : r=0.1540, p=0.0062.
— fiber and dietary beta-carotene : r=0.4826, $p < 0.0001$

(4) The following predictors have **moderate** evidence of relationship :
— age and cholesterol consumed : r= -0.1136, p=0.0439
— Quetelet index and alcohol : r = -0.1215, p=0.0311
— fat and alcohol : r=0.1295, $p = 0.0215$
— fat and dietary beta-carotene : r=0.1434, $p = 0.0108$
— cholesterol and dietary beta-carotene : r=0.1156, $p = 0.0403$

Note that 6 of the predictor variables appear to have a linear relationship with the response, while many pairs of predictor variables are correlated.

Many of the linear relationships noted above do not seem so strong in the scatterplots, particularly the negative correlations. Even though there is a great deal of scatter, many of the correlations are statistically significantly different from 0 because of the reasonably large sample size. This is reflected in the small p-values for correlations that are fairly small (eg., less than 3 in absolute value). From the scatterplots, a particularly strong relationship exists between calories and fat consumed.

2. (5 points) 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 :

TABLE 1 – Fit three regression equations

Predictors in regression	**Coefficient of calories**	P-value
(1) calories	-0.000086	0.2011
(2) calories and fat	0.000351	0.0214
(3) calories and Quetelet index	-0.000083	0.2010

In regression equations (1) and (3), the coefficients of calories are similar and neither are significantly different from 0. That is, the conclusion regarding calories in regression (1) and (3) are consistent. This is because we don't have evidence to indicate that calories consumed and Quetelet index are correlated ($r = 0.010$, $p = 0.8606$). So adding Quetelet index to the model does not affect how calories consumed predicts log of plasma. However in the regression (2) where we have predictors calories and fat consumed, the coefficient of calories is in a different sign and is statistically significantly different from 0. Calories and fat consumed are strongly correlated ($r = 0.900$, $p < 0.0001$) ; having fat in the model affects how calories predicts log of plasma over and above the effect of fat consumed.

3. (3 points) 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 :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3760	0.2758	19.49	0.0000
age	0.0051	0.0030	1.74	0.0832
genderM	-0.2585	0.1262	-2.05	0.0414
smoke1	-0.2501	0.1162	-2.15	0.0322
QI	-0.0318	0.0066	-4.84	0.0000
Vitamin1	0.1616	0.0808	2.00	0.0463
Calories	-0.0001	0.0002	-0.37	0.7094
fat	-0.0006	0.0031	-0.18	0.8559
fiber	0.0264	0.0110	2.41	0.0165
alcohol	0.0010	0.0087	0.11	0.9119
CHL	-0.0005	0.0004	-1.22	0.2220
DBC	0.0001	0.0000	1.79	0.0747

From the fitted model, there is **strong evidence** that the coefficient of Quetelet index is non-zero for a model including all of the other predictor variables. There is **moderate evidence** that the coefficients of male, smoking status, vitamin use, and fiber consumed are non-zero for a model including all of the other predictor variables. There is **weak evidence** that the coefficients of age and dietary beta-carotene are non-zero for a model including all of the other predictor variables. There is no evidence that the coefficients of calories, fat, alcohol and cholesterol consumed are different from 0.

So from this model it seems that **important predictors of plasma** are Quetelet index, male, smoking status, vitamin use, and fiber consumed and possibly age and dietary beta-carotene.

4. (3 points) 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 ques-

tion ?

Solution :

$$\widehat{logPlasma} = 5.4052 - 0.0329\ QI + 0.0303\ fiber - 0.0002\ calories - 0.2560\ smoke\ 1 + 0.1625\ vitamin\ 1 + 0.0001\ DBC - 0.2786\ gender\ M + 0.0049\ age$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.4052	0.2721	19.87	0.0000
QI	-0.0329	0.0064	-5.12	0.0000
fiber	0.0303	0.0094	3.22	0.0014
Calories	-0.0002	0.0001	-2.45	0.0150
smoke1	-0.2560	0.1158	-2.21	0.0278
Vitamin1	0.1625	0.0796	2.04	0.0419
DBC	0.0001	0.0000	1.74	0.0822
genderM	-0.2786	0.1220	-2.28	0.0231
age	0.0049	0.0029	1.69	0.0912

Note that this model includes the variables (including the possibly age and dietary beta-carotene) identified as important in the regression in the previous question, but also includes calories which was not identified as important.

5. (4 points : 2 points for R code and 2 points for presentation of the whole solution) Clear R source code for this assignment. (Write brief and clear comment in the between of your source code to ensure your R code is readable.)

```r
a3 <- read.table("a3data.txt",sep="",header=T)
str(a3)
is.factor(a3$gender);
is.factor(a3$smoke)
a3$smoke = as.factor(a3$smoke)
a3$Vitamin=as.factor(a3$Vitamin) ;
a3$logPlasma=log(a3$plasma)

# Q1: conduct all pairwise cor.test() between quantitative variables
##================ mycor function =================
## mycor( )
## input: data in matrix form
## ouput: pairwise correlation with p-value and plot of data
##=================================================

mycor <- function(data){

#------------ put histograms on the diagonal -----------------

panel.hist <- function(x, ...){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col="lavender", ...)
}
```

Listing 3 – Source R code

```

#----------- put correlations & P-value & 95% CIs on the lower panels ----------------
#
panel.cor <- function(x, y, digits=4, prefix="", cex.cor, ...){
    usr <- par("usr");
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))

    txt1 <- format( cor(x,y), digits=digits )
    txt2 <- format(cor.test(x,y)$p.value, digits=digits)
    text(0.5,0.5, paste("r=",txt1, "\n P.val=",txt2), cex=0.8)
   }
   pairs(data, lower.panel=panel.cor, cex =0.7, pch = 21, bg="steelblue",
   diag.panel=panel.hist, cex.labels = 1.1,
   font.labels=0.9, upper.panel=panel.smooth)
}

#----------- put correlations & P-value & 0.95 CIs on the lower panels ----------------
#


# Q1: find all quantitative variables in a3 data set
#
a3cor =a3[,c(1,4,6:11,13)]
mycor(a3cor)

# 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 Y=logPlasma

m1 <- lm(logPlasma~Calories,data=a3)
m2 <- lm(logPlasma~Calories+fat,data=a3)
m3 <- lm(logPlasma~Calories+QI,data=a3)
summary(m1)
summary(m2)
summary(m3)

# Q3: Fit  logPlasma with all 11 predictor variables and catch important X's
m11 <- lm(logPlasma~age+gender+smoke+QI+Vitamin+Calories+fat+fiber+alcohol
+CHL+DBC,data=a3)
summary(m11)

# Q4: What model does stepwise regression produce for this data

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

# with all predictors in the model
fullmod <- lm(logPlasma~age+gender+smoke+QI+Vitamin+Calories+fat+fiber+alcohol
+CHL+DBC,data=a3)

# stepwise method: apply both directions method
mboth = step(nullmod ,scope=list(lower=formula(nullmod),upper=formula(fullmod)),
    direction="both")
formula(mboth)
summary(mboth)

## generate latex code for the summary
library(xtable)
xtable(summary(mboth))
```

Listing 4 – Source R code