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 between logPlasma and all predictors, and the pairwise correlation between any two predictors)

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 ?

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 ?

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 question ?

I used the data of A2 to show you how to use backward elimination, forward selection and stepwise regression. Run the following code to learn how to do variable selection by **step()** in R :

```r
a2 <- read.table("DataA2.txt",sep="",header=T)
str(a2)
names(a2)
is.factor(a2$sex)
is.factor(a3$smoke)

# no predictor in the model
nullmod= lm (log(fev)~1, data=a2)
# with all predictors in the model
fullmod = lm( log(fev)~log(age)+ht+sex+smoke,data=a2)

# forward selection method
forwards = step(nullmod,scope=list(lower=formula(nullmod),upper=formula(fullmod)),
     direction="forward")
formula(forwards)

# backward elimination method
backwards = step(fullmod,scope=list(lower=formula(nullmod),upper=formula(fullmod))
    , direction="backward")
formula(backwards)

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

Listing 3 – R example: How to do variable selection by stepwise

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.)