University of California,Davis

plasma concentrations prediction

December 6, 2022

STAT 206

Statistical Methods for Research I

Professor: Jie Peng

Longbei Xu

Gu Gong

Yichu Chen

Contents

1 abstract **3**

2 Introduction **3**

3 Methodology and Results **3**

3.1 EDA . . . 3

3.1.1 Dataset . . . 3

3.1.2 Variables . . . 4

3.2 Assumption Checking . . . 4

3.3 Model Selection . . . 4

3.4 Model Summary . . . 5

3.5 Model Validation . . . 5

4 Conclusion and Discussion **6**

A Appendix **9**

1 abstract

Various types of plasma in blood have their own roles, among which retinol and beta plasma are considered to be important indicators of some important diseases. This study aims to find out how other important pathogenic indicators in the blood and some other indicators of lifestyle habits affect the levels of these two plasma. And according to the established model, the research suggests that the age, alcohol use, Cholesterol and then smoke status is important when predicting retinol plasma levels. Moreover, the Quetelet indicator, Cholesterol level, Dietary beta-carotene consumed , Age, the interaction between vitamin use and dietary beta-carotene intake, Fiber consumption and Vitamin Use attach more influence on the level of beta plasma.

2 Introduction

Vitamin A has been recognized to play an important role in the regulation of cell differentiation. It is commonly found in the human diet in the form of preformed Vitamin A (retinol), and provitamin A, which are a variety of carotenoids including beta-carotene, which can be converted to retinol. While retinol is usually contained in animal products like meat, liver, egg yolk and oily fish; beta-carotene is primarily found in green leafy vegetables and fruits. Despite the difference, both are very important substances in maintaining the level of Vitamin A in the human body . Based on previous studies, it has been established that, because loss of cell differentiation is a basic feature of malignancy, deficiency in Vitamin A, or low level of retinol and beta-carotene, may be associated with increased risk of certain types of cancers. However, what causes such deficiency remains unknown. In 1989, a study was carried out to investigate the association of retinol and beta-carotene plasma concentrations with several possible factors. [1] In this project we will conduct a regression analysis based on the data collected in the study and try to identify some important measures that significantly impact retinol and beta-carotene plasma concentrations.

3 Methodology and Results

3.1 EDA

Exploratory data analysis was first conducted, in which we examined missing values, variable types, their corresponding distributions and pairwise associations based on Pearson's correlation. Several visualization plots were created to help us understand the characteristics of our data and identify the general relationship between our explanatory variables and the two response variables. At the end, we also checked for the assumptions of our regression model.

3.1.1 Dataset

the dataset contains 315 observations with 14 variables, including plasma concentrations of beta-carotene and retinol as response variables and Age, Sex, Smoking Status, Quetelet Index, Vitamin Use, daily consumption of Calories, Fat, Fiber, Cholesterol and Dietary beta-carotene and weekly

consimption of Alcohol. More detailed description can be found in table 11. There is no missing value or duplicated entries in our dataset. More details can be found in table 22.

3.1.2 Variables

We first visually checked the distribution for all variables. As for continuous variables, from the histogram 3and boxplots 4 of all continuous variables we see that alcohol, dietary retinol and beta-carotene plasma concentration are right skewed. As for categorical variables, from pie chart of categorical variables 5, we see that sex variable there are about 87% Female and 13% Male, in Smoking status there are 37% former smoker, 50% non-smoker and 13% current smoker. In Vitamin use variable, there is 35% observations never use vitamin, 26% not often use, 39% often use.

Then we performed a visual analysis on the relationship between response variables and explanatory variables. A scatter plot matrix 8 gave us pairwise associations between our variables. Pairwise scatter plots of each variable present at the lower left corner of the matrix and the correlation value between each pair of variables is shown at the upper right corner. The plot below shows that there are high linear relationships between Calories and Fat, Calories and Fiber, Calories and cholesterol, Fat and cholesterol, indicating the possibility of multicollinearity. Also the plot shows that most of the variables are approximately linearly distributed with the response variables so that the linearity assumption holds.

3.2 Assumption Checking

We make 4 fundamental assumptions on our model: linearity, homoscedasticity, independence of error and normality of error. To check if all assumptions are met, we first fitted a full model for both response variables. For both responses, residual plots initially showed non-constant variance and the Q-Q plots showed strong deviation to normality. We applied Box-Cox procedure and implemented log-transformation for both response variables. Figure 0A 2 and 0B 2 shows the residuals distribution against fitted values and theoretical Normal quantiles. No obvious pattern is found on the residuals vs. fitted value plots and no strong deviation is detected on the Q-Q plots. We therefore conclude that all assumptions were met after log transformation was conducted.

3.3 Model Selection

After gaining a preliminary understanding of the variables of interest, we proceeded to statistical modeling, in which we applied methods of multiple regression analysis to further explore the statistical relationships between our responses and potential predictors. Because we are interested in identifying influential factors for the plasma concentrations of both beta-carotene and retinol, we built two separate regression models using the same set of predictors, which are Age, Sex, Smoking Status, Quetelet Index, Vitamin Use, consumption of Calories, Fat, Fiber, Cholesterol and Dietary beta-carotene per day, as well as the number of Alcohol drinks consumed per week. We also make 4 assumptions on our model: linearity, homoscedasticity, independence of error and normality of error, all of which were met.

In model selection, we adopted stepwise regression with AIC as our model selection criterion. We do not use the best subset selection procedure due to computation complexity. Because some of the explanatory variables are mutually correlated, we conducted one round of forward selection followed by another round of backward selection to avoid missing important variables. We tried to fit both of our models using first order main effect terms and second order interaction terms, and examined their corresponding AIC scores as well as (adjusted) R-squared values.

For retinol plasma concentration, the second order model (with interaction), which produced an AIC of 616.8, outperformed the model with only main effects, which produced an AIC of 613. For beta-carotene plasma concentration, both the second order model (with interaction) and the model with only main effects produced an AIC of roughly 168. We still prefer the model with interaction since its multiple R-squared and adjusted R-squared both exceeded that produced by the main effect model. In conclusion, the final models we chose both had interaction terms in them. We may also check the true predictive ability of our models by the mean squared prediction error (MSPE), which will be discussed in the model validation section. By calculating the variance inflation factor, we see no noticeable multicollinearity among the explanatory variables in both models.

3.4 Model Summary

The final model for beta-carotene plasma concentration contains 6 main effect terms and 1 interaction term, including Quetelet, Vitamin Use, Cholesterol, Dietary beta-carotene, Fiber, Age and the interaction between Vitamin Use and Dietary beta-carotene. Based on F-test, coefficients for Quetelet, Vitamin Use, Cholesterol and Dietary beta-carotene, Fiber and the interaction between Vitamin Use and Dietary beta-carotene are significant at alpha=0.05; the coefficient for Age is not significant at any level of significance. The final model produced a multiple R-squared of 0.2533 and an adjusted R-squared of 0.2288. Roughly speaking, the model is capable of explaining about 25 percent of the total variability of log-beta carotene plasma concentration in the sample.

The final model for retinol plasma concentration contains 4 main effect terms and 1 interaction term, including Alcohol, Age, Cholesterol, Sex and the interaction between Cholesterol and Sex. Based on F-test, coefficients for Alcohol, Age and Cholesterol is significant at alpha=0.05; coefficients for Sex and the interaction between Cholesterol and Sex are not significant at any level of significance. The final model produced a multiple R-squared of 0.1168 and an adjusted R-squared of 0.1009. Roughly speaking, the model is capable of explaining about 12 percent of the total variability of log-retinol plasma concentration in the sample.

3.5 Model Validation

To check the validity of the models in this research, firstly the overfitting problem should also be included in our concerns for these two models. To investigate these specific problems, this paper mainly compares the difference between the mean square prediction estimation of external validation, which refer to MSPE and the sum of square error divided by sample size, which refer to SSE/n. In the Retinol model, we can see that the MSPE for this model is 523095.1 is slightly larger than the SSE/n is 50125, considering the size of those two numbers. Moreover, among the Beta

plasma model, the MSPE is 0.447, which is smaller than the SSE/n, 0.606 in this case. And that means the overfitting problem does not exist among both of the models.

The other important diagnostic for the regression model is to find out the outliers and influential cases among the data based on two different response variables. First of all, this paper uses Bonferroni's threshold to check the outlying Response variables. The Bonferroni threshold is a family-wise error threshold, which treats a set of tests as one family, and the threshold is designed to control the probability of type I error. [2] By computation, we get that Bonferroni's threshold for beta plasma is 3.56. And compare it with the studentized deleted residuals for each observation, we can find out that the 257th observation is considered to be the outlying in response variable. Moreover, this paper uses the same method to identify the outlying in response variable, and the Bonferroni's threshold for retinol plasma is 3.56, and the studentized deleted residuals for the 81st observation is slightly larger than the Bonferroni's threshold for retinol plasma, which indicate that this observation can be considered to be the outlying in response variables.

Besides the outliers among the response variable, the outliers among the explanatory variables are also considered to be crucial to detect. In this paper, the method of Cook's distance is applied to the Retinol and Beta plasma model to check the outliers. The Cook's distance is a way to identify points that negatively affect your regression model. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance[3]. In the Retinol plasma model, the 276th and 49th are considered to be outliers among explanatory variables. And among the Beta plasma model, the 257th observations are considered to be outliers.

However, the reason for these outliers are still vague. If the outliers are caused by the lab record error or other types of typos. Then the outliers can be removed. Otherwise, the outliers are caused by some extreme occasional case. Since the research mainly focuses on the relationship between different types of plasma and other body indicators. For some extreme cases, the outliers may represent some specific situation, which can not be removed since that would eventually lead to bias of our data and that would cause the failure of finding out the actual conclusion of the research.

4 Conclusion and Discussion

Based on the final model the research achieved, the research concludes that the beta-carotene plasma is affected by Quetelet, Vitamin Use, Cholesterol, Dietary beta-carotene, Fiber, Age and the interaction between Vitamin Use and Dietary beta-carotene. And the research also suggest that the retinol plasma is affected by the Alcohol, Age, Cholesterol, Sex and the interaction between Cholesterol and Sex. Moreover, the universality of two different models perform very well in this case, which means the model would not vary a lot when facing more similar data from the same resources. However, the performance of each model is not good enough to be a solid model since the adjusted R-square is relatively low, which means the selected variables only can interpret a quite low amount of variation of two response variables. Based on the histogram of each variable, the right-skewed patterns among plentiful variables are very obvious. Although transform for the response variables for better fitting ability is applied, the model performance does not improve

to an acceptable level. Due to the sampling bias, the models are greatly disturbed. To fix such a problem, a larger and more unbiased dataset would help to achieve a better model. Besides the mathematical interpretation of each model, the realistic meaning behind that is also important. For the beta-carotene plasma, the higher Quetelet, less vitamin use, higher Cholesterol would lead to decrease in such plasma, which these variables are considered to be an indicator of unhealthy status. And having more fibers would help the body to have a higher beta plasma. Counter-intuitively, having more beta-carotene food would also decrease the beta-plasma in the tester's body. This could be attributed to data bias or a more complicated mystery of the human body. For the retinol plasma, having more alcohol and aged male intend to have a higher retinol plasma. However, having a higher Cholesterol could lead to a lower retinol plasma. In conclusion, these two response variables are generally affected by body status, which suggests having a good body status would help to decrease the risk of having disease caused by those two different plasma.

References

[1] David W Nierenberg, Therese A Stukel, John A Baron, Bradley J Dain, E Robert Greenberg, and Skin Cancer Prevention Study Group. Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130(3):511–521, 1989.

[2] Matthew Brett. Notes on the bonferroni threshold. matthew-brett.github.io/teaching/bonferroni_correction.html.

[3] S Glen. Cook's distance/cook's d: Definition, interpretation, 2016.

A Appendix

Number	Variables	Meaning
1	AGE	Age(years).
2	SEX	Sex of observation(1=Male, 2=Female).
3	SMOKESTAT	Smoking status (1=Never, 2=Former, 3=Current).
4	QUETELET	Quetelet is weight divided by height squared.
5	VITUSE	Vitamin Use (1=Yes, fairly often, 2=Yes, not often , 3=no).
6	CALORIES	Number of calories consume per day.
7	FAT	Grams of fat consumed per day.
8	FIBER	Grams of fiber consumed per day.
9	ALCOHOL	Number of alcoholic drinks consumed per week..
10	CHOLESTER	Cholesterol consumed (mg per day).
11	BETADIET	Dietary beta-carotene consumed (mcg per day).
12	RETDIET	Dietary retinol consumed (mcg per day).
13	BETAPLASMA	Plasma beta-carotene (ng/ml).
14	RETPLASMA	Plasma Retinol (ng/ml).

Table 1: The Explanation of each variable in the dataset

Number	Variables	Types	Number of Null
1	AGE	integer	0
2	SEX	character	0
3	SMOKESTAT	character	0
4	QUETELET	numeric	0
5	VITUSE	charater	0
6	CALORIES	numeric	0
7	FAT	integer	0
8	FIBER	numeric	0
9	ALCOHOL	numeric	0
10	CHOLESTER	numeric	0
11	BETADIET	integer	0
12	RETDIET	integer	0

Table 2: Preliminary exploratory analysis of variables

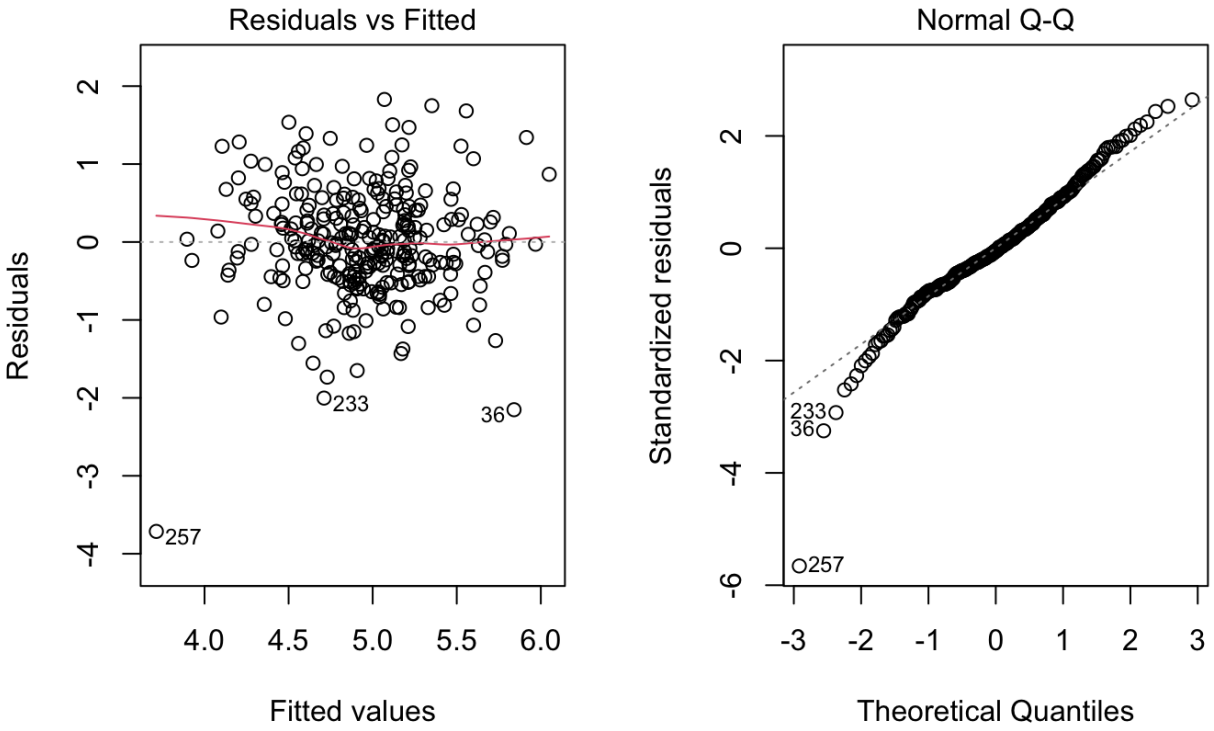

Figure 1: The residual plot and Q-Q plot for beta plasma

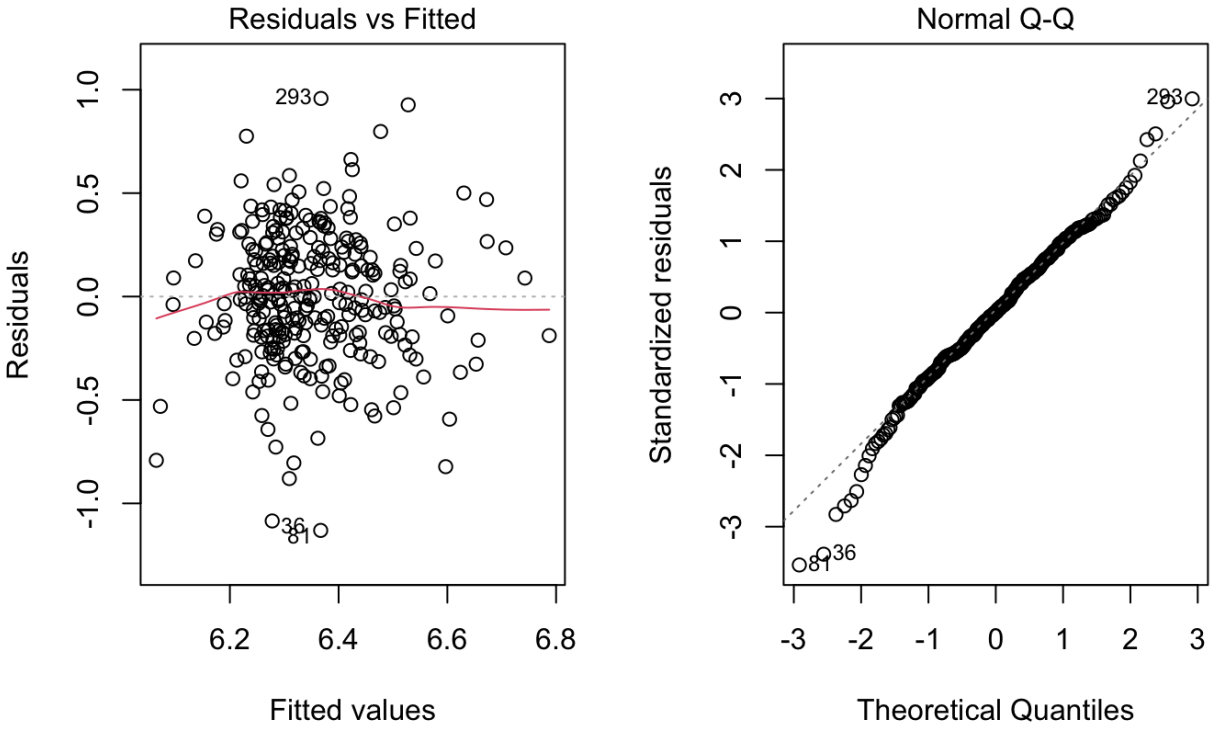

Figure 2: The residual plot and Q-Q plot for Retinol plasma

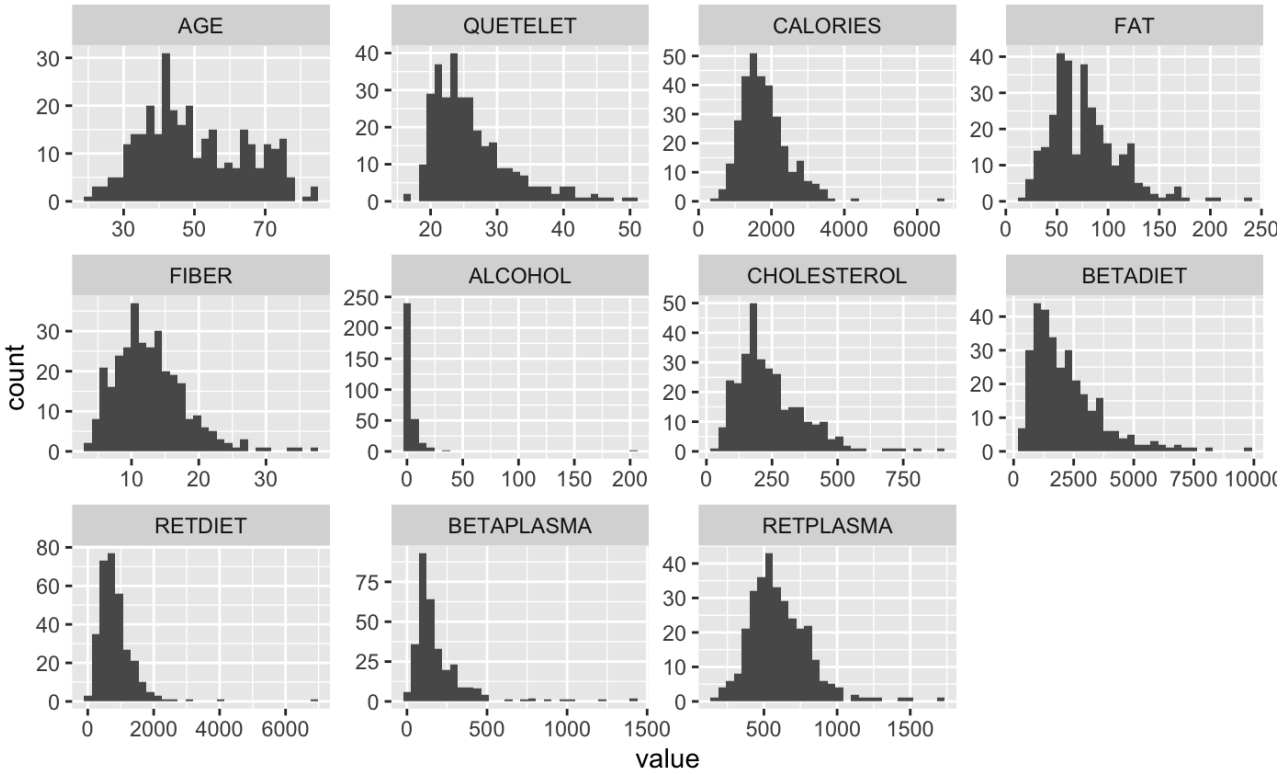

Figure 3: The histogram of continuous variables

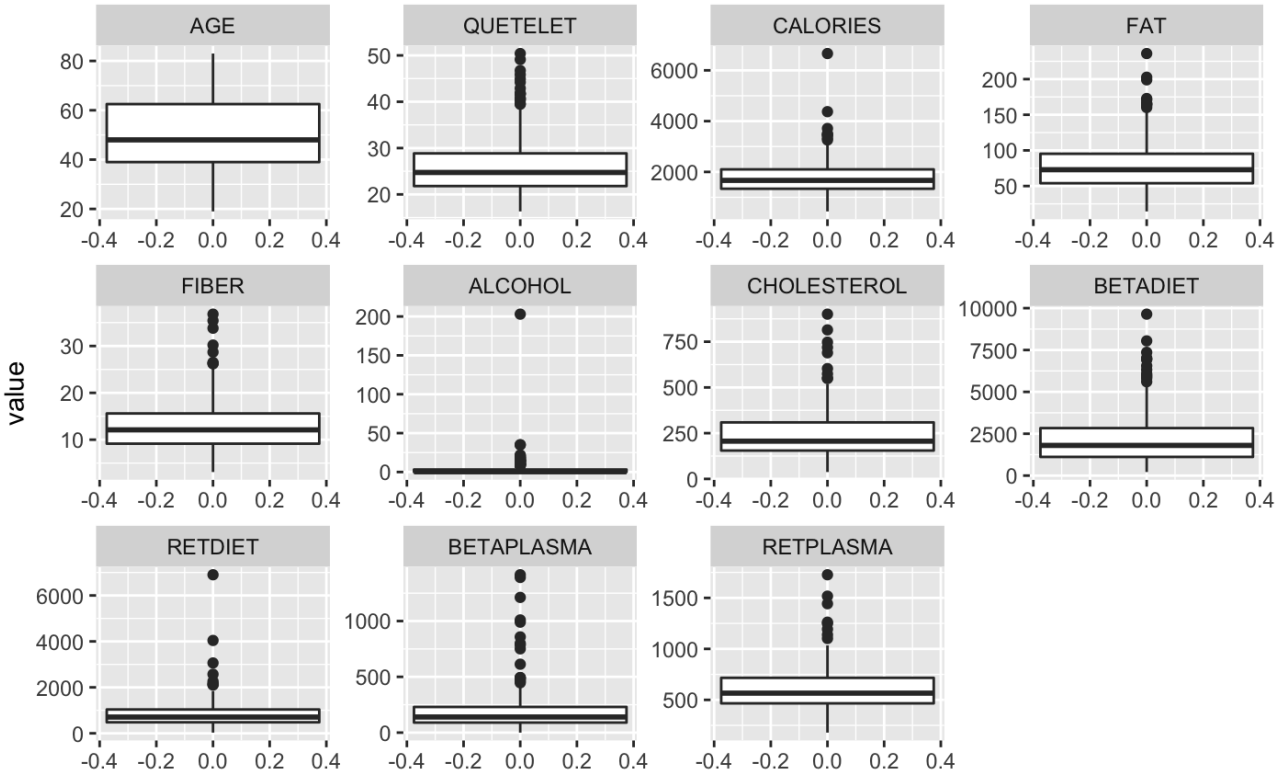

Figure 4: The boxplot of continuous variables

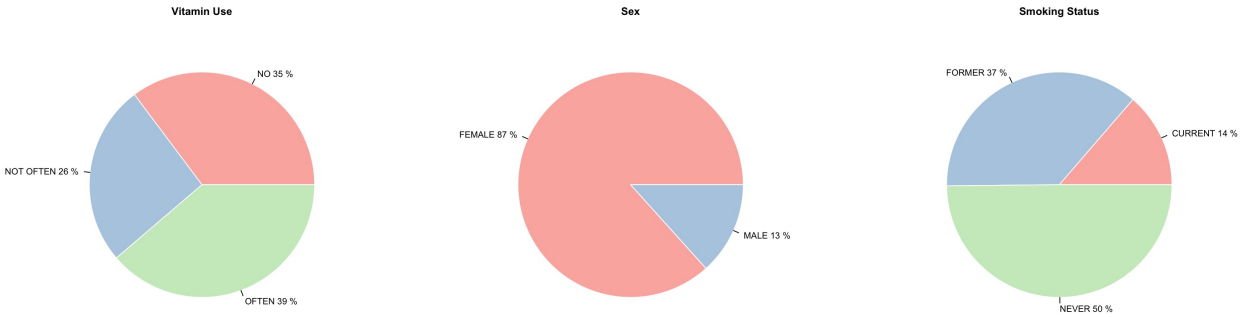

Figure 5: The pie chart of categorical variables

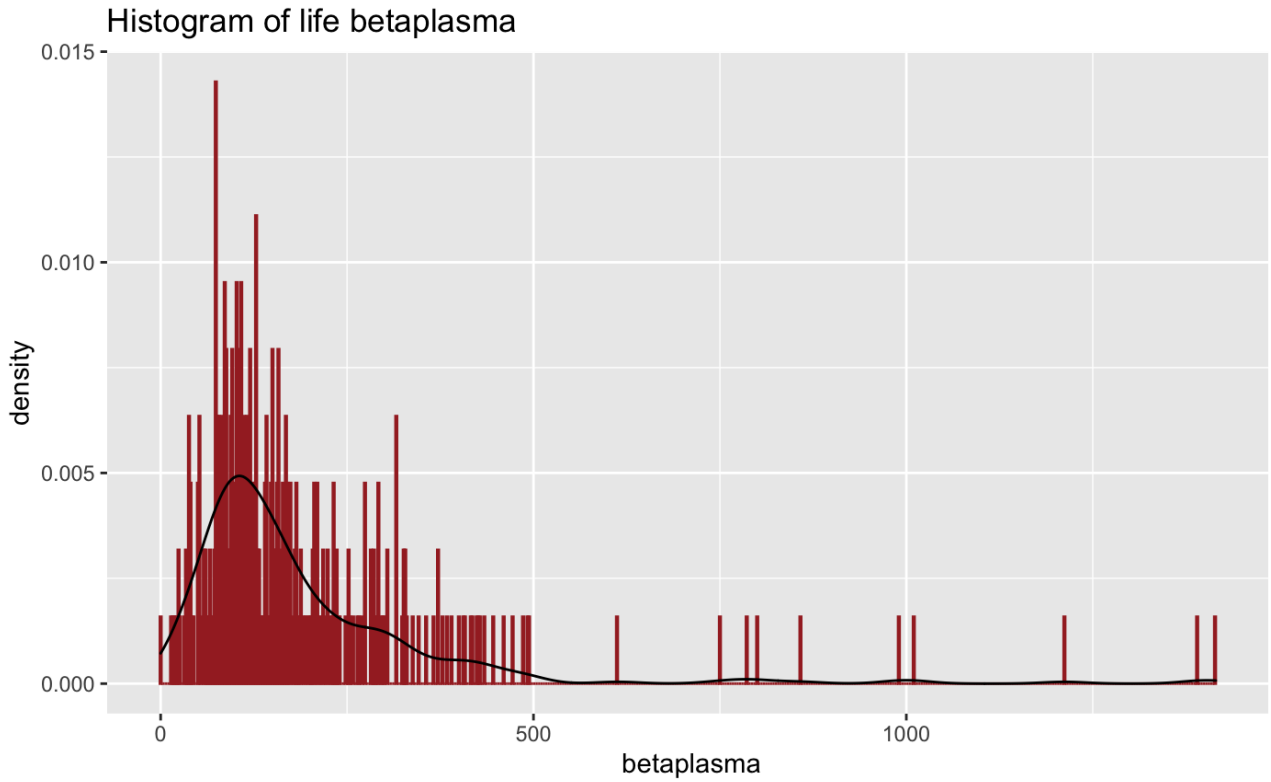

Figure 6: The histogram of beta plasma

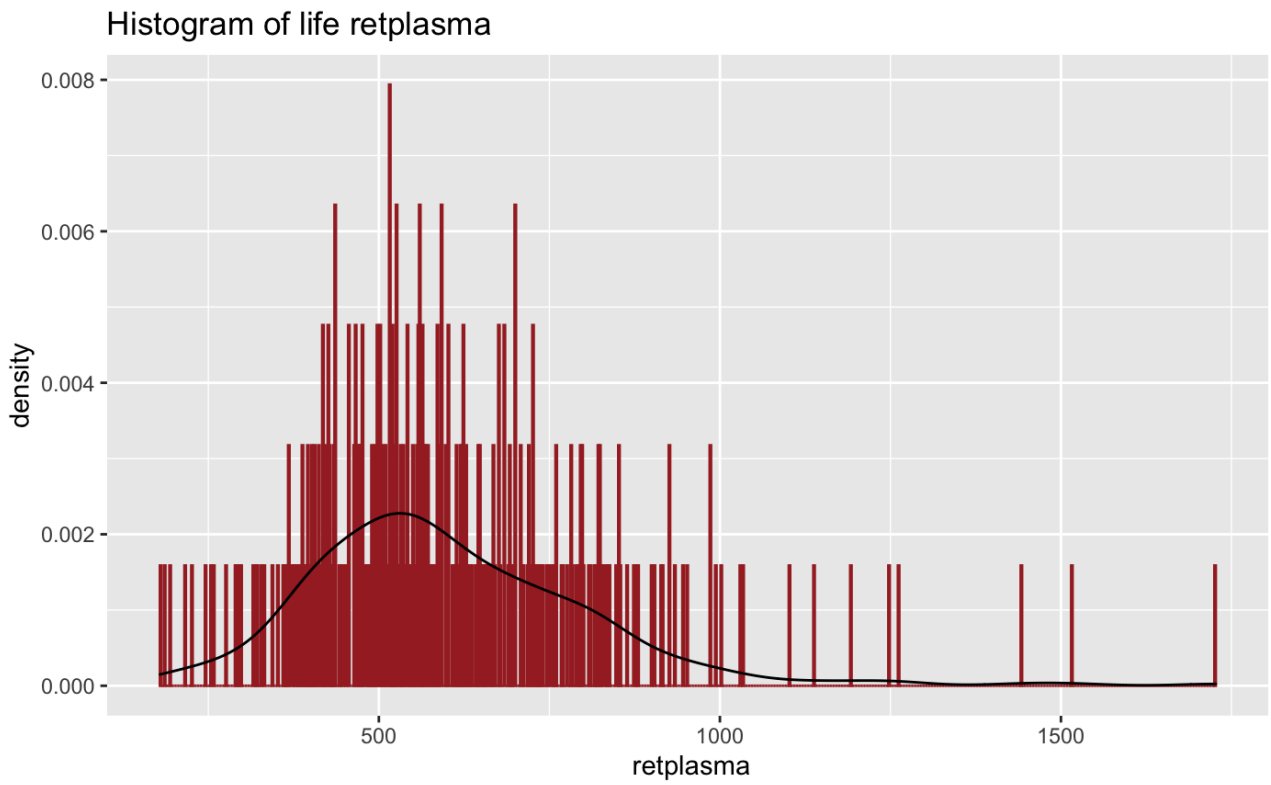

Figure 7: The histogram of retinol plasma

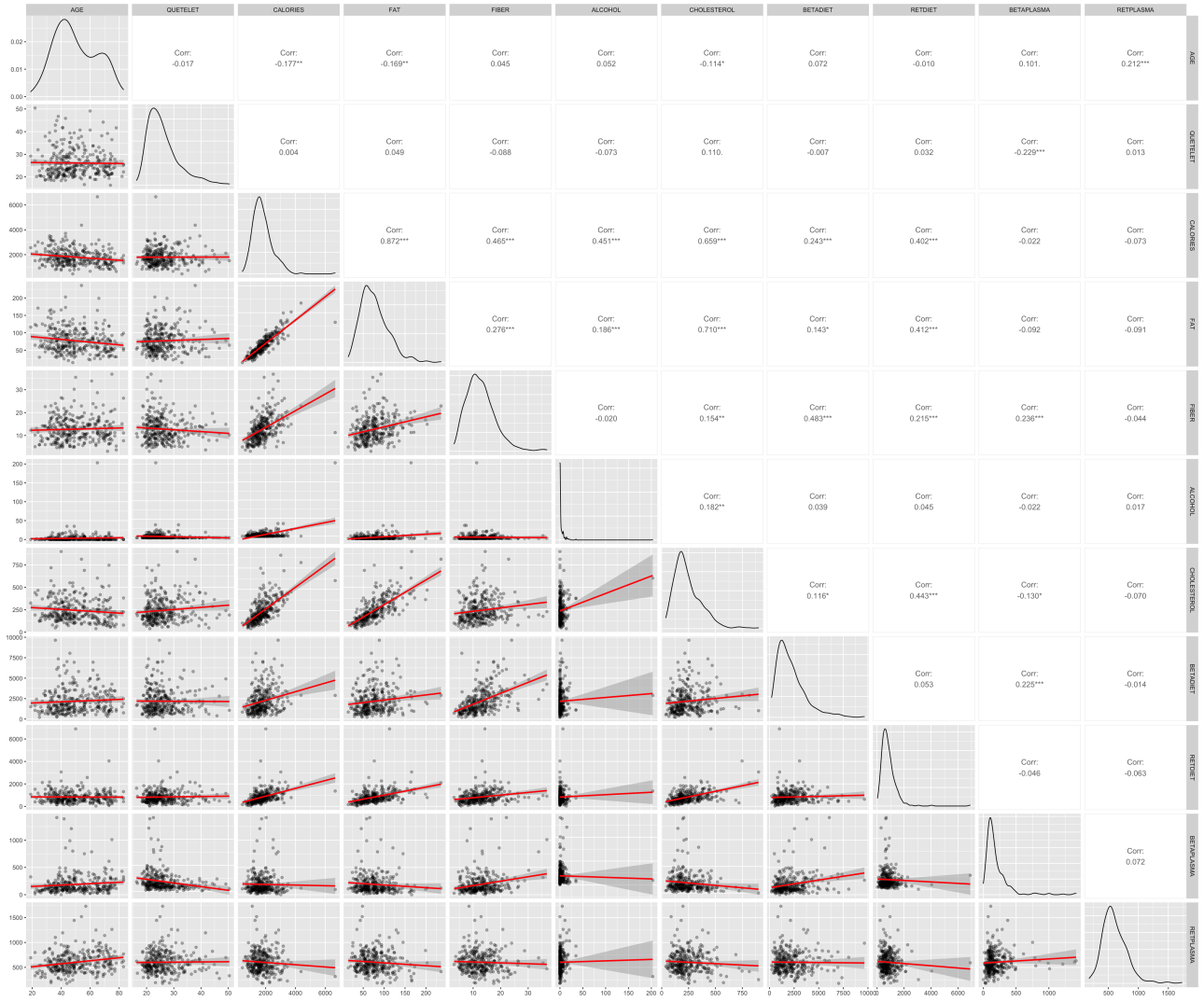

Figure 8: The correlation matrix of variables

206 Project Appendix B

Longbei Xu, Gu Gong, Yichu Chen

12/2/2022

Appendix B

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## Loading required package: carData
```

```
# import data
plasma <- read.delim("Plasma.txt", stringsAsFactors=TRUE)

# check missing value
colSums(is.na.data.frame(plasma))

# check data type
lapply(plasma, class)
```

```
# numerical and categorical sub.dataframe
num_data <- plasma[,unlist(lapply(plasma, is.numeric))]
cat_data <- plasma[,!unlist(lapply(plasma, is.numeric))]
cat.response <- cbind(plasma[,c('BETAPLASMA','RETPLASMA')], cat_data)

# column names
num_names <- names(num_data)
cat_names <- names(cat_data)
```

```
lowerF <- function(data, mapping, method = "lm") {
  p <- ggplot(data = data, mapping = mapping) +
    # scatter plot
    geom_point(alpha=0.3) +
    # smoothed regression line
    geom_smooth(method = method, color = "red")
  p
}
ggpairs(num_data, lower = list(continuous = wrap(lowerF, method= "lm")))
```

```
# correlation heatmap
ggcorr(num_data, method = c("everything", "pearson"), cex=2)+labs(title="Correlation Heatmap for Numeri
```

```r
# histogram and boxplots for continuous variables
num_long <- melt(num_data)
```

```
## No id variables; using all as measure variables
```

```r
ggplot(num_long,aes(x = value)) +
    facet_wrap(~variable,scales = "free") +
    geom_histogram(aes(y = ..density..), color = 2, fill = "white") +
    geom_density(lwd = 0.6,
                linetype = 2,
                color = 1,
                fill=2,
                alpha=0.1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
ggplot(num_long)+
  facet_wrap(~variable, scales='free')+
  geom_boxplot(aes(y= value))
```

```r
# pie charts for categorical variables

par(mfrow=c(1,3))
tab_vit <- table(cat_data$VITUSE)
pct_vit <- round(100*tab_vit/nrow(cat_data))
label_vit <- paste(levels(cat_data$VITUSE), pct_vit)
label_vit <- paste(label_vit, "%", sep=" ")
pie(tab_vit, border="white", labels=label_vit, col= brewer.pal(length(levels(cat_data$VITUSE)), "Pastel

tab_sex <- table(cat_data$SEX)
pct_sex <- round(100*tab_sex/nrow(cat_data))
label_sex <- paste(levels(cat_data$SEX), pct_sex)
label_sex <- paste(label_sex, "%", sep=" ")
pie(tab_sex, border="white", labels=label_sex, col= brewer.pal(length(levels(cat_data$SEX)), "Pastel1")

tab_smoke <- table(cat_data$SMOKSTAT)
pct_smoke <- round(100*tab_smoke/nrow(cat_data))
label_smoke <- paste(levels(cat_data$SMOKSTAT), pct_smoke)
label_smoke <- paste(label_smoke, "%", sep=" ")
pie(tab_smoke, border="white", labels=label_smoke, col= brewer.pal(length(levels(cat_data$SMOKSTAT)), "
```

```r
# boxplot for categorical variables vs. response variables

library(patchwork)
```

```
##
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```r
par(mfrow=c(1,2))

p1 <- ggplot(cat.response, aes(x=VITUSE)) +
      geom_boxplot(position = "dodge2", aes(y=BETAPLASMA)) +
      labs(title="Beta-Plasma VS. Vitamin Use")


p2 <- ggplot(cat.response, aes(x=VITUSE)) +
      geom_boxplot(position = "dodge2", aes(y=RETPLASMA)) +
      labs(title="Retinol-Plasma VS. Vitamin Use")

p3 <- ggplot(cat.response, aes(x=SEX)) +
      geom_boxplot(position = "dodge2", aes(y=BETAPLASMA)) +
      labs(title="Beta-Plasma VS. Sex")

p4 <- ggplot(cat.response, aes(x=SEX)) +
      geom_boxplot(position = "dodge2", aes(y=RETPLASMA)) +
      labs(title="Retinol-Plasma VS. Sex")

p5 <- ggplot(cat.response, aes(x=SMOKSTAT)) +
      geom_boxplot(position = "dodge2", aes(y=BETAPLASMA)) +
      labs(title="Beta-Plasma VS. Smoking Status")

p6 <- ggplot(cat.response, aes(x=SMOKSTAT)) +
      geom_boxplot(position = "dodge2", aes(y=RETPLASMA)) +
      labs(title="Retinol-Plasma VS. Smoking Status")


p1+p3+p5+p2+p4+p6+plot_layout(ncol = 3)
```

```r
set.seed(123)

# train-test split
N <- nrow(plasma)
pool <- c(1:N)

train_idx <- sample(pool, size=ceiling(N*0.9))
valid_idx <- pool[-train_idx]

# split for 2 responses
betaplasma <- plasma[, !names(plasma)%in%c("RETPLASMA")]
retplasma <- plasma[, !names(plasma)%in%c("BETAPLASMA")]

beta_train <- betaplasma[train_idx,]
beta_valid <- betaplasma[valid_idx,]

ret_train <- retplasma[train_idx,]
ret_valid <- retplasma[valid_idx,]
```

Beta-carotene plasma concentration

Initial Checking

```
#First-order full model before transformation (used for Box-Cox and assumption checking)
beta.reg <- lm(BETAPLASMA +1~ ., data=beta_train)
# summary(beta.reg)
```

```
# assumption checking for beta-carotene
par(mfrow=c(1,2))
plot(beta.reg, which=1)
plot(beta.reg, which=2)
```

```
# Box-Cox procedure for beta-carotene
library(MASS)
bc = boxcox(beta.reg)
```

```
# First-order full model with transformed Y
beta.reg <- lm(log(BETAPLASMA+1) ~ ., data=beta_train)
```

```
# initial assumption checking for beta carotene
par(mfrow=c(1,2))
plot(beta.reg, which=1)
plot(beta.reg, which=2)
```

Candidate Model (without interaction) for beta-carotene

```
#first-order model stepwise selection using AIC
step.reg.beta <- stepAIC(beta.reg, direction="both", trace=0)
summary(step.reg.beta)

# model AIC
AIC(step.reg.beta)
```

```
#residuals for first-order model (based on stepwise selection)
{plot(step.reg.beta, which=1)
plot(step.reg.beta, which=2)}
```

Final Model (with interaction) for beta-carotene

```
# interaction model for beta-carotene (stepwise AIC)
step.reg <- stepAIC(lm(log(BETAPLASMA+1)~1, data=beta_train), scope = ~(AGE+SEX+SMOKSTAT+QUETELET+VITUS

summary(step.reg)

# model AIC
AIC(step.reg)
```

```r
# anova table
step.reg$anova
```

Retinol plasma concentration

Initial Checking

```r
#First-order before transformation
ret.reg <- lm(RETPLASMA ~ ., data=ret_train)
# summary(ret.reg)

# before transformation
par(mfrow=c(1,2))
plot(ret.reg, which=1)
plot(ret.reg, which=2)
```

```r
# Box-Cox transformation
library(MASS)
bc1 =boxcox(ret.reg, data=ret_train)
```

```r
#First-order with transformation
ret.reg <- lm(log(RETPLASMA) ~ ., data=ret_train)
# summary(ret.reg)
```

```r
# after transformation: assumption checking
par(mfrow=c(1,2))
plot(ret.reg, which=1)
plot(ret.reg, which=2)
```

Candidate Model (without interaction) for retinol

```r
#first-order stepwise using AIC

ret.reg0 <- lm(log(RETPLASMA) ~ 1, data=ret_train)
ret.reg <- stepAIC(ret.reg0,
                   scope = ~(AGE+SEX+SMOKSTAT+QUETELET+VITUSE+CALORIES+FAT+FIBER+ALCOHOL+CHOLESTEROL+BE
summary(ret.reg)

AIC(ret.reg)
```

```r
#residuals for assumption checking
{plot(ret.reg, which=1)
plot(ret.reg, which=2)}
```

Final Model (with interaction) for retinol

```r
none_mod<- lm(RETPLASMA~1, data=ret_train) ##model with only intercept

full_mod<-lm(RETPLASMA~(.)^2,data=ret_train)
# summary(full_mod)

library(MASS)
#forward selection based on AIC:
stepRET.aic<-stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1), direction="both", k=2, trace = F
summary(stepRET.aic)

stepRET.aic$anova

AIC(stepRET.aic)
```

Model validation

Beta-Carotene

```r
step.reg.beta <- stepAIC(lm(log(BETAPLASMA+1)~1, data=beta_train), scope = ~(AGE+SEX+SMOKSTAT+QUETELET+

# train set SSE
sse_t <- sum(step.reg.beta$residuals^2)
```

```r
# VIF
vif(step.reg.beta)
```

Variance Inflation Factor

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```r
# MSPE
y.hat.beta <- predict(step.reg.beta, beta_valid)
MSPE.beta <- mean((log(beta_valid$BETAPLASMA+1) - y.hat.beta)^2)
MSPE.beta
```

```r
# SSE/n
n=N*0.7
sse.beta_train <- sum(step.reg.beta$residuals^2)
sse.beta_train/n
```

MSPE

```
# final model for beta-carotene, fit1
fit1<- lm(log(BETAPLASMA+1)~ QUETELET + VITUSE + CHOLESTEROL + BETADIET + FIBER + AGE + VITUSE:BETADIET
e1<-fit1$residuals  ##ordinary residuals
h1<-influence(fit1)$hat ##diagonals of the hat matrix: a.k.a. leverage values
de1<-e1/(1-h1) ##deleted residuals
{plot(e1,de1, xlab="residuals", ylab="deleted residuals")
abline(0,1)}
```

```
summary(h1)
```

```
library(MASS)
stu.res.del1 <- studres(fit1)
head(sort(abs(stu.res.del1), decreasing=TRUE))
```

```
# Bonferroni-corrected threshold for outlying Y
qt(1-.1/(2*n), n-10-1)
```

Outlying Y

```
## [1] 3.563261
```

```
h1 <- influence(fit1)$hat
p1 <- 10
sort(h1[which(h1>2*p1/n)], decreasing = TRUE)
```

```
# Cook's distance
res1 <- fit1$residuals
mse1 <- anova(fit1)["Residuals", 3]
cook.d1 <- res1^2*h1/(p1*mse1*(1-h1)^2)
```

Cook's Distance

```
# ranking of each observation based on Cook's Distance
sort(cook.d1[which(cook.d1>4/(n-p1))], decreasing = TRUE)
```

```
plot(fit1, which=4)
```

```
plot(fit1, which=5)
```

Influencial Points based on Cook's Distance

```
fit.257<-lm(log(BETAPLASMA+1)~  QUETELET + VITUSE + CHOLESTEROL + BETADIET + FIBER + AGE + VITUSE:BETADI
rbind(fit1$coefficients,fit.257$coefficients) ##compare fitted regression coefficients
```

```
{plot(fit1$fitted.value, predict(fit.257, beta_train[,c("QUETELET","CHOLESTEROL","BETADIET","FIBER","VI
abline(0,1)}
```

Evaluation on effect of influencial point

Retinol

```
# the final model
stepRET.aic
# SSE
sse.ret_train <- sum(stepRET.aic$residuals^2)
```

```
# VIF
vif(ret.reg)
```

Variance Inflation Factor

```
# MSPE
newdata2 <- ret_valid
y.hat.ret <- predict(stepRET.aic, newdata2)
MSPE2 <- mean((log(ret_valid$RETPLASMA+1) - y.hat.ret)^2)
MSPE2
```

```
# SSE/n
n=N*0.7
sse.ret_train/n
```

MSPE

```
# final model for retinol, fit2
fit2<- lm(log(RETPLASMA+1)~ ALCOHOL + AGE +  CHOLESTEROL + SEX + CHOLESTEROL:SEX, data = ret_train)
e2<-fit2$residuals ##ordinary residuals
h2<-influence(fit2)$hat ##diagonals of the hat matrix: a.k.a. leverage values
de2<-e2/(1-h2) ##deleted residuals
{plot(e2,de2, xlab="residuals", ylab="deleted residuals")
abline(0,1)}
```

```
summary(h2)
```

```
library(MASS)
stu.res.del2 <- studres(fit2)
head(sort(abs(stu.res.del2), decreasing=TRUE))
```

```
# Bonferroni-corrected threshold
qt(1-.1/(2*n), n-6-1)
```

```
p2 <- 6
sort(h2[which(h2>2*p2/n)], decreasing = TRUE)
```

Outlying Y

```
# Cook's Distance
res2 <- fit2$residuals
mse2 <- anova(fit2)["Residuals", 3]
cook.d2 <- res2^2*h2/(p2*mse2*(1-h2)^2)
```

Cook's Distance

```
# ranking of each observation based on Cook's Distance
sort(cook.d2[which(cook.d2>4/(n-p2))], decreasing = TRUE)
```

```
plot(fit2, which=4)
```

```
plot(fit2, which=5)
```

Influencial Points based on Cook's Distance

```
fit.27649<-lm(log(RETPLASMA+1)~ ALCOHOL + AGE +  CHOLESTEROL + SEX + CHOLESTEROL:SEX, data = ret_train,
rbind(fit2$coefficients,fit.27649$coefficients) ##compare fitted regression coefficients
```

```
{plot(fit2$fitted.value, predict(fit.27649, ret_train[,c("ALCOHOL" , "AGE" , "SEX" , "CHOLESTEROL")]), x
abline(0,1)}
```

Evaluation on effect of influencial point

Final Evaluation on Goodness-of-fit

```
# goodness of fit for final models
par(mfrow=c(1,2))
plot(fit1, which=1)
plot(fit1, which=2)
```

```
par(mfrow=c(1,2))
plot(fit2, which=1)
plot(fit2, which=2)
```

Model Summary

```
# beta-carotene
summary(fit1)
```

```
##
## Call:
## lm(formula = log(BETAPLASMA + 1) ~ QUETELET + VITUSE + CHOLESTEROL +
##     BETADIET + FIBER + AGE + VITUSE:BETADIET, data = beta_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7133 -0.3952 -0.0300  0.4019  1.8289
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.516e+00  2.853e-01  19.332  < 2e-16 ***
## QUETELET                 -3.091e-02  6.854e-03  -4.509 9.67e-06 ***
## VITUSENOT OFTEN           1.580e-01  1.866e-01   0.847  0.39780
```

```
## VITUSEOFTEN              -4.875e-02  1.819e-01  -0.268  0.78891
## CHOLESTEROL              -1.519e-03  3.287e-04  -4.621 5.89e-06 ***
## BETADIET                 -4.024e-05  5.841e-05  -0.689  0.49142
## FIBER                     1.606e-02  9.139e-03   1.757  0.07996 .
## AGE                       4.813e-03  2.941e-03   1.637  0.10287
## VITUSENOT OFTEN:BETADIET  9.668e-05  7.680e-05   1.259  0.20916
## VITUSEOFTEN:BETADIET      1.918e-04  7.267e-05   2.639  0.00879 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6983 on 274 degrees of freedom
## Multiple R-squared:  0.2533, Adjusted R-squared:  0.2288
## F-statistic: 10.33 on 9 and 274 DF,  p-value: 1.011e-13
```

```
# retinol
summary(fit2)
```

```
##
## Call:
## lm(formula = log(RETPLASMA + 1) ~ ALCOHOL + AGE + CHOLESTEROL +
##     SEX + CHOLESTEROL:SEX, data = ret_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13057 -0.18890  0.00087  0.21558  0.95761
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.1700814  0.0880028  70.112  < 2e-16 ***
## ALCOHOL              0.0143279  0.0039556   3.622 0.000347 ***
## AGE                  0.0041300  0.0014125   2.924 0.003740 **
## CHOLESTEROL         -0.0002921  0.0001682  -1.737 0.083577 .
## SEXMALE              0.2038189  0.1391861   1.464 0.144225
## CHOLESTEROL:SEXMALE -0.0004837  0.0004007  -1.207 0.228406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3213 on 278 degrees of freedom
## Multiple R-squared:  0.1168, Adjusted R-squared:  0.1009
## F-statistic: 7.353 on 5 and 278 DF,  p-value: 1.718e-06
```

ANOVA table

```
# beta-carotene
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: log(BETAPLASMA + 1)
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## QUETELET          1  14.331 14.3313 29.3917 1.298e-07 ***
```

```
## VITUSE            2   9.777  4.8886 10.0259 6.277e-05 ***
## CHOLESTEROL       1   8.662  8.6623 17.7653 3.397e-05 ***
## BETADIET          1   6.184  6.1837 12.6820 0.0004352 ***
## FIBER             1   1.944  1.9443  3.9874 0.0468301 *
## AGE               1   0.959  0.9589  1.9665 0.1619490
## VITUSE:BETADIET   2   3.462  1.7311  3.5502 0.0300460 *
## Residuals       274 133.602  0.4876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# retinol
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: log(RETPLASMA + 1)
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## ALCOHOL            1  1.5058 1.50583 14.5890 0.0001649 ***
## AGE                1  1.5343 1.53429 14.8646 0.0001436 ***
## CHOLESTEROL        1  0.5293 0.52927  5.1278 0.0243163 *
## SEX                1  0.0748 0.07483  0.7249 0.3952638
## CHOLESTEROL:SEX    1  0.1504 0.15041  1.4572 0.2284064
## Residuals        278 28.6944 0.10322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```