Determinants of Plasma Retinol Level and Plasma Beta-carotene Level

Graduate Project Report

Table of Contents

1. Introduction

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients, and plasma concentrations of the micronutrients also varied widely from subject to subject. All these facts motivate me to study the factors that play important roles in human plasma concentrations of retinol and beta-carotene.

In this study, a cross-sectional data was used to investigate the relationship between personal characteristics, dietary factors, and plasma concentrations of retinol and beta-carotene. Study subjects were 276 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.

Many epidemiologic studies have shown a strong association between plasma beta-carotene, plasma retinol and the risk of cancer. Given that plasma beta-carotene and plasma retinol are key measurements in other studies, it would be valuable to study the determinants of plasma beta-carotene level and plasma retinol level. In addition, the relationship between plasma beta-carotene or retinol levels and the age, sex, quetelet, calories, fat, fiber, alcohol consumed per week, cholesterol consumed per day, dietary beta-carotene consumed per day, and dietary retinol consume per day is less clear from the previous literatures. The main objective of this study is to find out what factors influence beta-carotene levels and retinol levels in human plasma respectively and maybe further inspire scientists to evaluate the internal factors that may have some effect or relationship with the beta-carotene and retinol in people plasma.

2. Data Analysis

The data is collected from an observational experiment with 276 patients. In this study, there are totally 12 independent variables, including age, sex, smoking status, quetelet, vitamin use, number of calories consumed per day, grams of fat consumed per day, grams of fiber consumed per day, number of alcoholic drinks consumed per week, cholesterol milligram consumed per day, dietary beta-carotene microgram consumed per day and dietary retinol microgram consumed per day; and 2 dependent variables, including plasma beta-carotene (ng/ml) and plasma Retinol (ng/ml). This study will examine whether and how any of these personal characters would have an effect on plasma beta-carotene and plasma retinol.

Specifically, there are two types of variables among 12 independent variables: quantitative variables and categorical variables.

2.1. Descriptive Statistics

Quantitative Variables:

AGE: Age (years)
QUETELET: Quetelet (weight/(height^2)
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
CHOLESTEROL: Cholesterol consumed (mg per day)
BETADIET: Dietary beta-carotene consumed (mcg per day)
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

Categorical Variables:

SEX: Sex (1=Male, 2=Female).
SMOKSTAT: Smoking status (1=Never, 2=Former Smoker, 3=Current Smoker)
VITAMIN: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)

Descriptive Analysis Table

Variable	Min.	1st Qt	Median	Mean	3rd Qt	Max.
AGE	19	39	48	50.28	62.25	83
SEX	1	2	2	1.873	2	2
SMOKING	1	1	1	1.609	2	3
QUETELET	16.33	22.01	24.77	26.34	29.04	50.4
VITAMIN	1	1	2	1.935	3	3
CALORIES	445.2	1318.6	1649.7	1772.9	2073.5	6662.2
FAT	14.4	52.6	72.2	75.23	94.12	235.9
FIBER	3.7	9.075	12.1	12.839	15.6	36.8
ALCOHOL	0	0	0.5	3.334	3.125	203
CHOLESTEROL	37.7	154.2	203.8	235.9	302.1	814.7
BETADIET	214	1109	1824	2177	2792	9642
RETDIET	30	475.2	700	822	1026.2	6901
BETAPLASMA	41	97	145	199.9	233.5	1415
RETPLASMA	216	471.8	569	611	725.5	1727

This descriptive analysis table above is calculated using R. After observation of the descriptive analysis table above, we find that most of the data is dependable since the mean is close to the median, except for alcohol, cholesterol, betadiet and retdiet. The reason that the means are much larger than its medians in case of those four variables might be the existence of outliers. For example, the average number of alcoholic drinks consumed per week is 3.334 given its median 0.5. We could suspect that some people may over drink alcohol too much while some people seldom drink. The existence of outliers reduces the quality of data to a great extent and leads the estimation results of data biased or inaccurate. Those outliers will be removed later after additional data analysis.

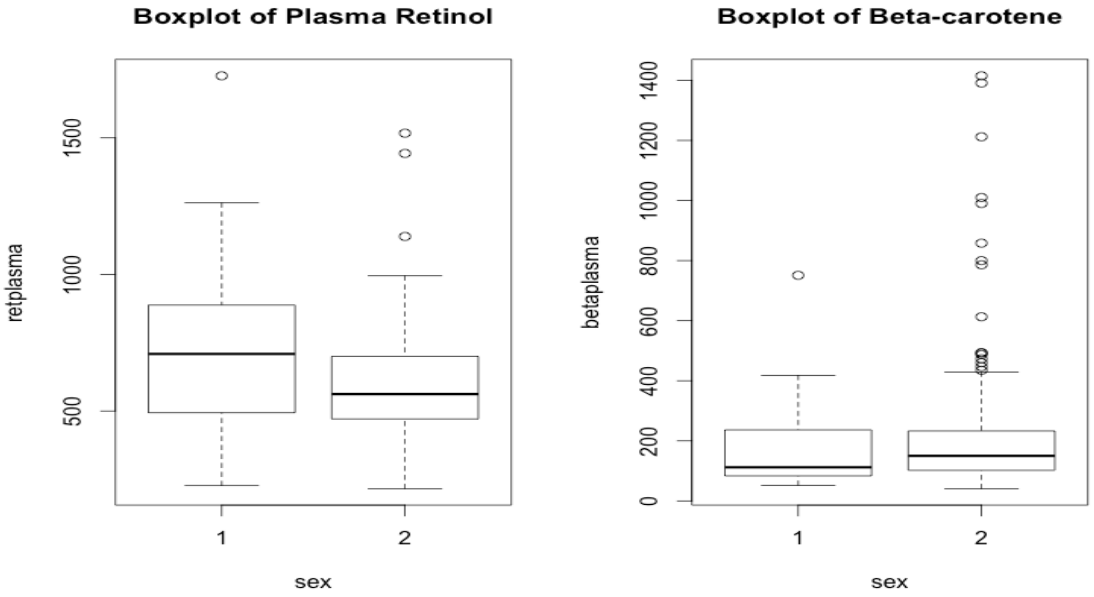
2.2. Category Variables

This study includes three category variables: sex, vitamin use habit and smoking status. We will examine the distribution of plasma retinol for each categorical variable, and the distribution of plasma beta-carotene for each categorical variable respectively.

Firstly, we will compare the plasma retinol distribution of male and female, and plasma beta-carotene distribution of male and female. In this study, male is represented

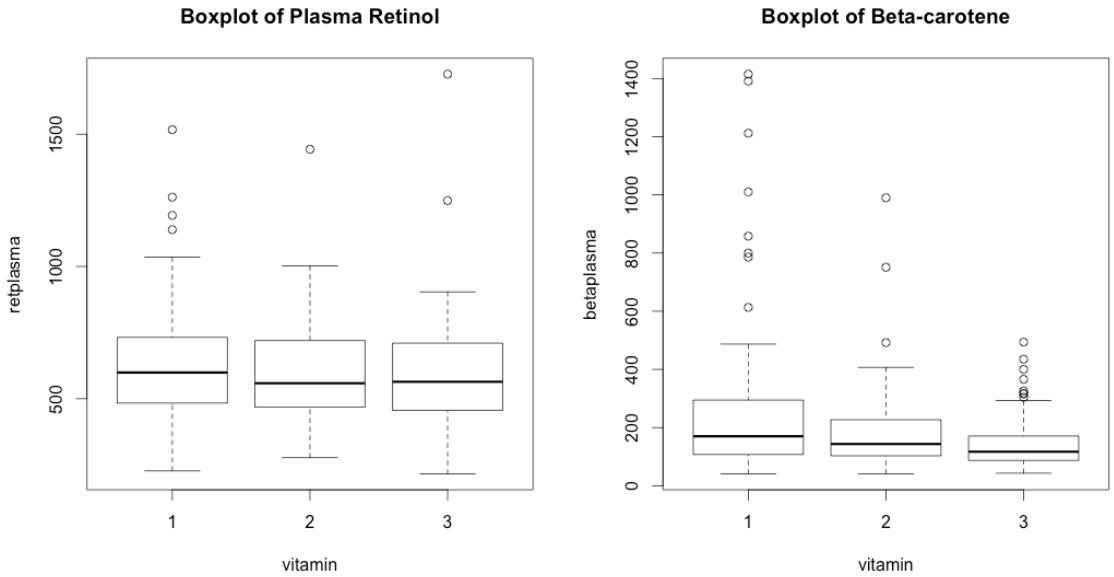
by 1, and female is by 2. The side-by-side boxplots on the left below shows the relationship between sex and plasma retinol. The median plasma retinol level for females is lower than for males. It should be noted that even the third quartile of the females' distribution is lower than the median plasma retinal level for males. Therefore, by looking at the center of the plasma retinol boxplot, we conclude that in general, the level of plasma retinol for males is higher than that for females. Judging by the IQR, which measures the variability only among the middle 50% of the distribution, there is more spread in the distribution of males than females. The middle 50% of the plasma retinol distribution of females is more homogeneous than the males' plasma retinol distribution. Although we see outliers in both distributions, there is only one high outlier in males' distribution compared with three high outliers in the females' distribution.

However, in the scenario of plasma beta-carotene (right side boxplot below), the median level for females is slightly higher than for males. From the range of the data, there is much more variability in the females' plasma beta-carotene distribution than in the males' distribution.

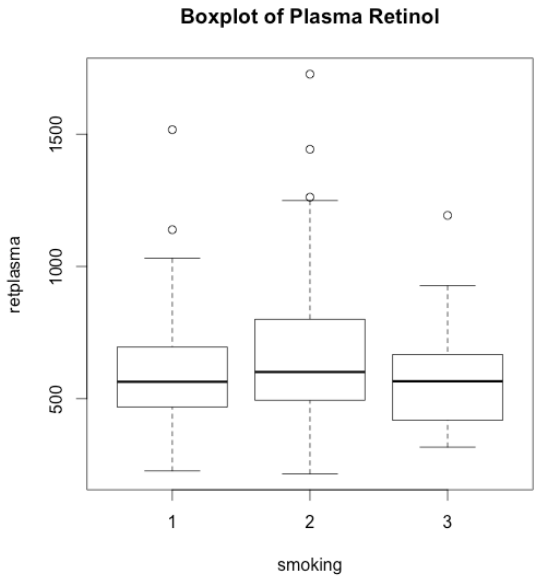


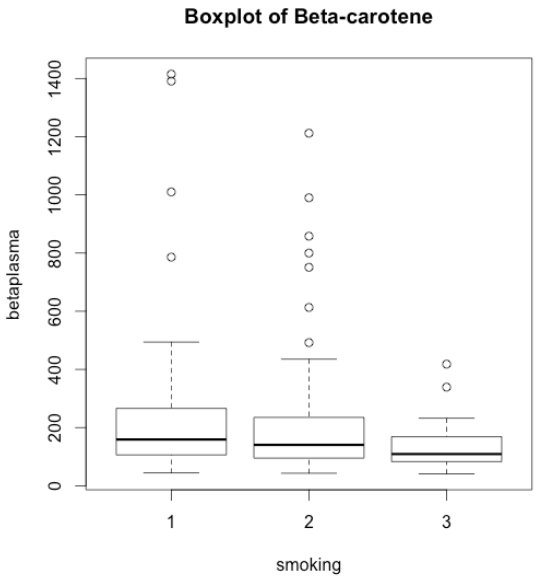
Applying the same method used in contrasting distributions from male and female, the comparative boxplots below reveal the distribution of plasma retinol by vitamin use habit and plasma beta-carotene by vitamin use habit. A 3-point scale is used to measure the vitamin use, that is 1=Yes, fairly often, 2=Yes, not often, and 3=No. From the boxplots of plasma retinol by vitamin use levels, we can see that the median plasma retinol level for people who use vitamin fairly often is slightly higher than people who do not often or never use vitamin. It might indicate that using vitamin fairly often has some positive effect on plasma retinol level but need further research. The distributions of plasma retinol by different vitamin groups have roughly the same range and IQR, which indicates there is not much variability in the plasma retinol among three groups. From the boxplots of plasma beta-carotene with vitamin below, we find that people who use vitamin very often have relatively higher plasma concentration of beta-carotene than the other two groups, and people who use vitamin but not often have higher level than non-

users. There are a lot of outliers in the group of people who use vitamin fairly often in plasma concentrations of both micronutrients.



Regarding the study of patients' smoking status, we use a 3-point scale as similar as vitamin use habit, where 1=Never Smoke, 2=Former Smoker, 3=Current Smoker. The boxplots below show the relationships between plasma retinol, plasma beta-carotene and different status of smoking. It is interesting to see that the plasma retinol distribution of former smokers is higher than the current smokers' plasma retinol distribution, which means the former smokers usually have a higher level of plasma retinol than others. The IQR of non-smokers is the smallest among the three groups, which means the middle 50% of the plasma retinol distribution of non-smokers have less variability. However, the range of current smokers is the smallest compared with the former smoker and non-smokers. Regarding the plasma beta-carotene condition, people who are in a never smoking group have a relatively higher plasma beta-carotene level when compared with the other two groups, and the former smoker group has much more outliers when compared with the other groups. The current smokers bear a relative lower level of plasma beta-carotene than nonsmokers and former smokers.



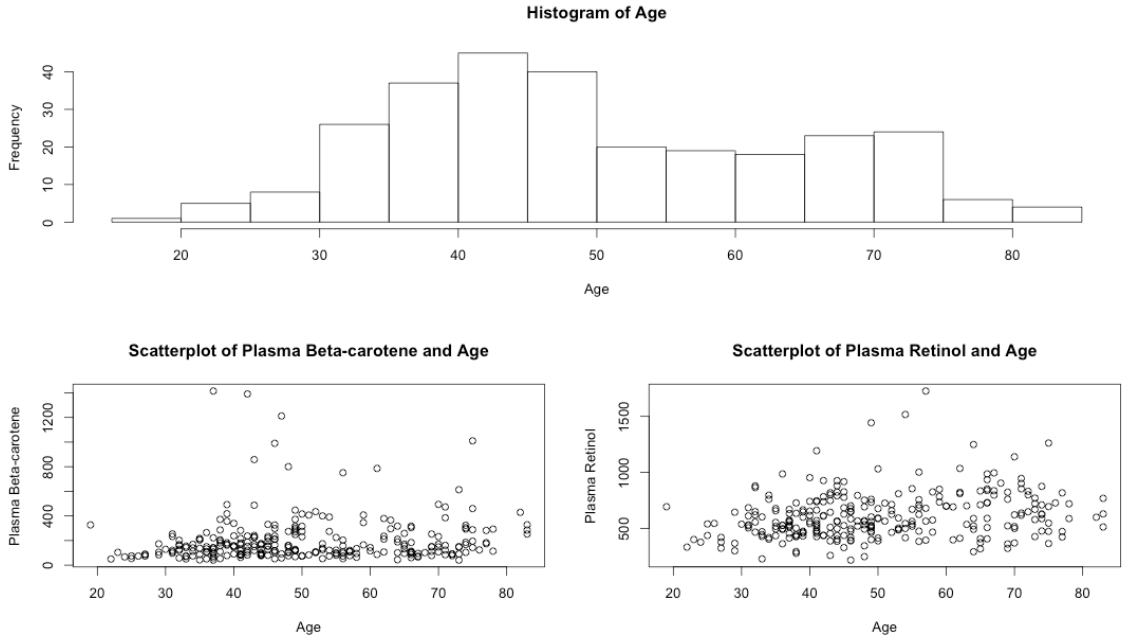


2.3. Quantitative Variables

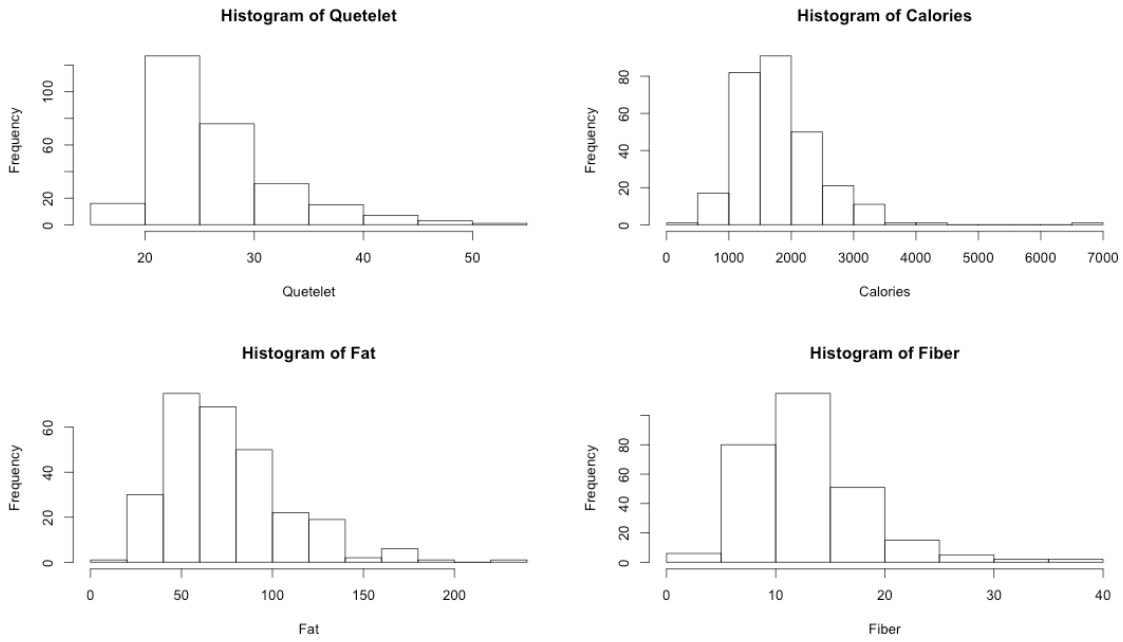
After comparing and contrasting the category variables using boxplots above, we move on to evaluate the distributions of quantitative variables, which are best represented graphically by histograms. The quantitative variables in this study including: age, quetelet, number of calories consumed per day, grams of fat consumed per day, grams of fiber consumed per day, number of alcoholic drinks consumed per week, cholesterol milligram consumed per day, dietary beta-carotene microgram consumed per day and dietary retinol microgram consumed per day.

From the descriptive statistics table, we know that patients' age in this research ranges from 19 to 83 and the mean of age is around 50. As shown in the histogram of age below, most of our patients' ages fall in the area between 30 and 75, which imply that most participants are middle-aged or elderly people. The distributions of age in terms of plasma retinol and plasma beta-carotene are randomly spread, which do not violate the

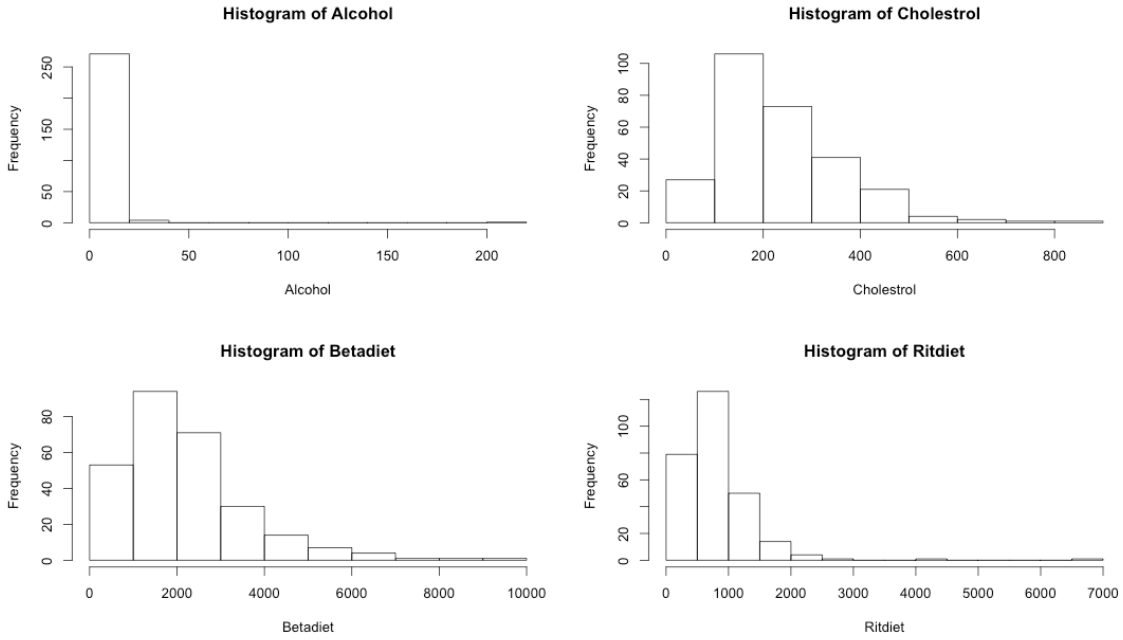
normality assumption.



The histograms of the quetelet, number of calories consumed per day, grams of fat consumed per day, grams of fiber consumed per day and are shown below. This study uses the formula quetelet=(weight/ (height^2) to measure patients' height and weight in order to estimate a healthy body weight based on how tall a person is. The lower the value, the thinness the person is. According to WHO, the quetelet level from 18.5 to 24.9 is considered normal, and any level below this range is considered underweight while any level above is considered overweight. The histogram of the quetelet shows that most of the data fall in the intervals from 20 to 30, which indicates that most patients are healthy though some are a little overweight. As shown in the descriptive statistics table and the histogram below, the mean of participant's calories consumed is 1773 and the most of the data falls into the area between 1000 and 2500. The FDA chose a 2000 calorie-per-day diet as a base in developing food labels so that consumers could easily calculate the daily values needed for their own diets. 2000 calories is also the amount of total calories per day that a moderately active adult female (weighing approximately 132 pounds) would need to maintain her weight. FDA also suggests that a less active person would need fewer calories while a more active person would need more. Without knowing the patients' weight and active level, we cannot determine how much calorie they need or whether the patients are on a healthy diet. However, one important thing we can tell from the histograms below is that the distributions of calories, fat and fiber are roughly in bell shapes, which follows the independent identical normal distribution assumption for a linear model.



The histograms below show the distribution of the number of alcoholic drinks consumed per week, cholesterol milligram consumed per day, dietary beta-carotene microgram consumed per day and dietary retinol microgram consumed per day. The histogram of alcohol is clearly skewed right, which indicates that the mean is greater than the median and the data has high outliers. Given the descriptive analysis, we notice that the mean number of alcohol consumed per week is 3.3 while the maximum number consumed is 203. After observing the data, except two data points of 203 and 35, the rest numbers of alcohol consumed per week are no larger than 20, therefore the two data points (203 and 35) are dropped out and the rest numbers smaller than 20 are kept. Similarly, we can find some outliers in the histogram of cholesterol, betadiet and retdiet on the right side of each figure. In general, most of the data distributes in the interval between 100 and 500 milligram for cholesterol consumed per day, in the interval between 1000 and 4000 microgram for dietary beta-carotene consumed per day, and less than 2000 microgram for dietary retinol consumed per day.

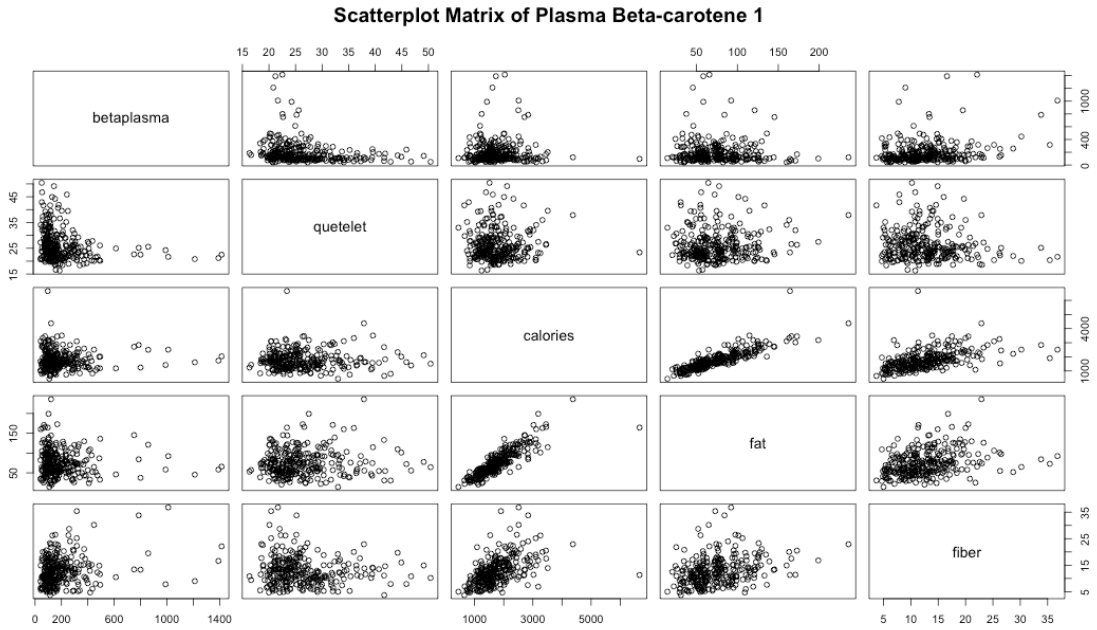


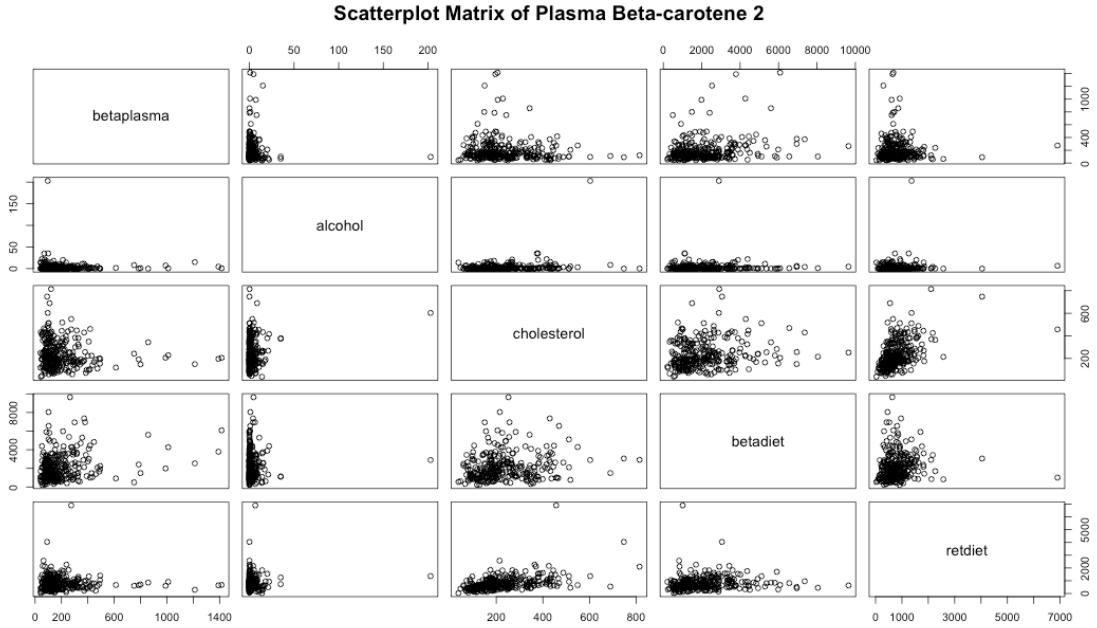
3. Regression Methods and Results

3.1. Data Cleaning - Plasma Beta-carotene Case

The main objective of this study is to find out the relationship between plasma beta-carotene or retinol levels and the age, sex, quetelet, calories, fat, fiber, alcohol consumed per week, cholesterol consumed per day, dietary beta-carotene consumed per day, and dietary retinol consume per day. From the scatterplot of age earlier, we can see that the data is randomly distributed, now let us draw the scatterplot of plasma beta-carotene against all the other eight predictors to check for outliers or data aggregations.

Scatterplot Matrix of Plasma Beta-carotene



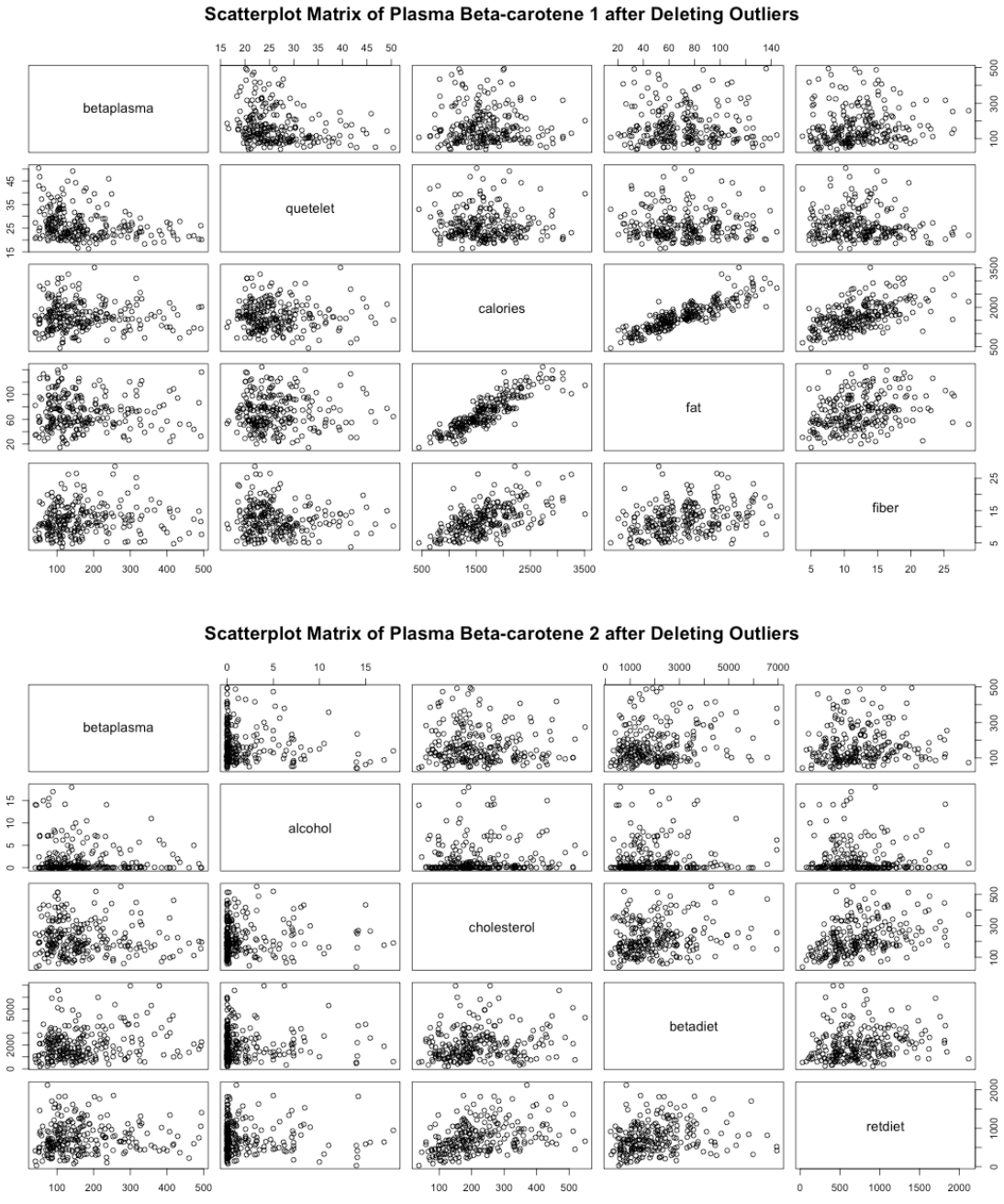


The scatterplots of plasma beta-carotene above apparently have some patterns, which could be caused by outliers. Therefore, the next step is removing the outliers and recreating the scatterplots. The outliers are removed by choosing number calories consumed per day less than 4000, grams of fat consumed per day less than 150, grams of fiber consumed per day less than 30, number of alcoholic drinks consumed per week less than 20, cholesterol milligram consumed per day less than 600, dietary beta-carotene

microgram consumed per day less than 7000 and dietary retinol microgram consumed per day less than 2200.

After removing all the outliers, another two scatterplots are created, whose distributions are more randomly distributed as shown below.

Scatterplot Matrix of Plasma Beta-carotene after Deleting Outliers



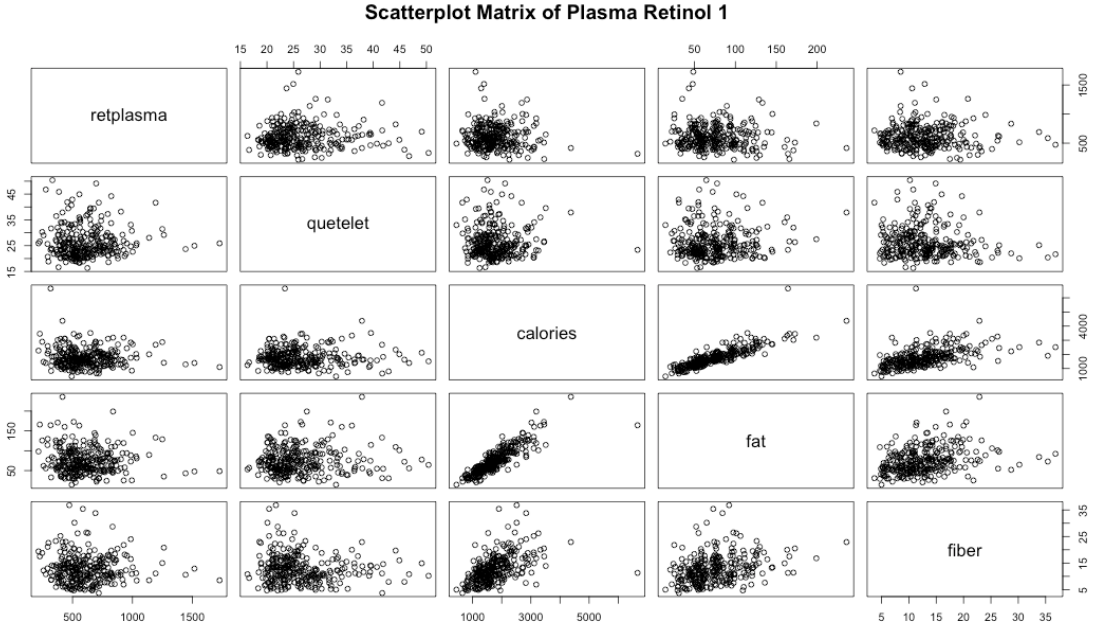
3.2. Data Cleaning - Plasma Retinol Case

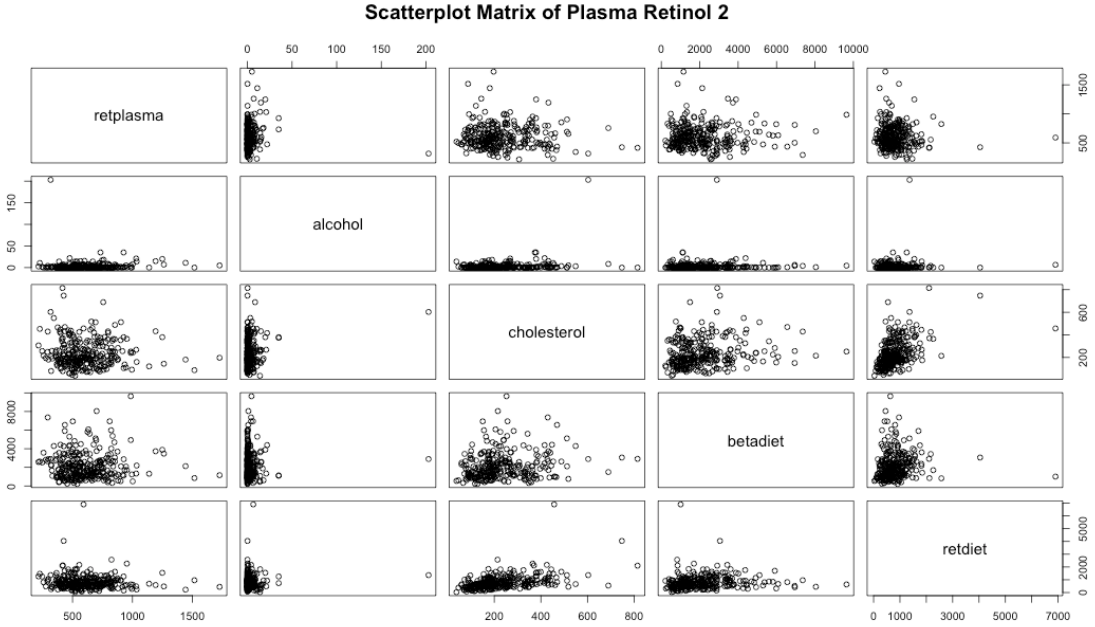
In addition to plasma beta-carotene, similar methods are applied to study the 12 factors that influence retinol levels in human plasma. Firstly, the scatterplot matrix of

retinol plasma with all the continuous predictors is generated to check outliers or data aggregations. Each scatterplot of retinol plasma with a predictor is analyzed for further observation of the outliers. Outliers are removed so that the scatterplot will be symmetrically distributed. The corresponding scatterplots for before and after data cleaning are shown below.

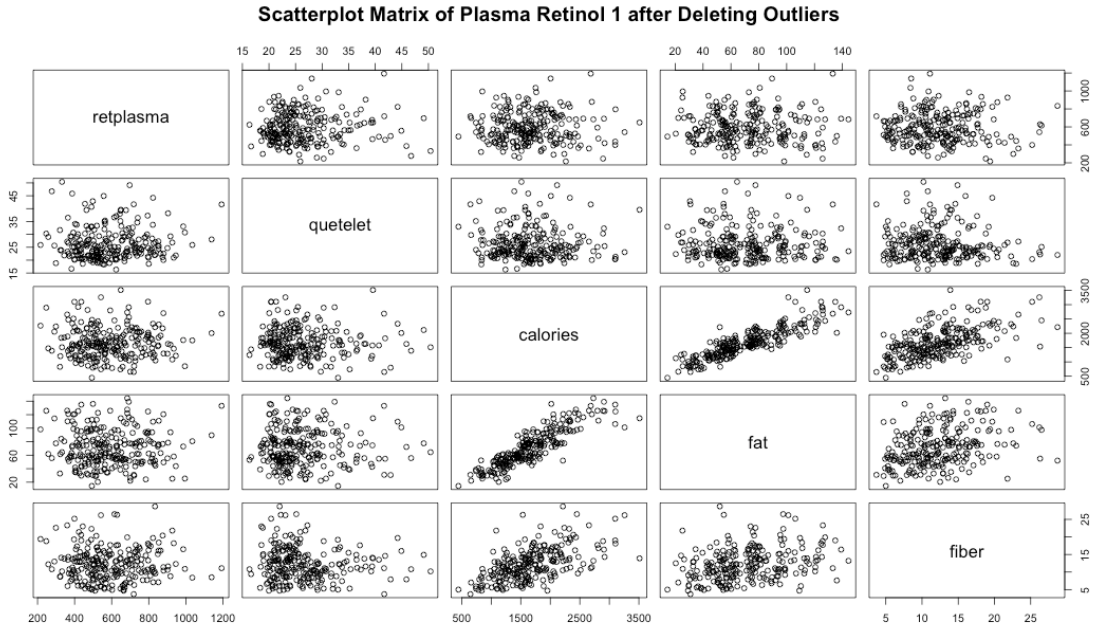
Similar to the results of plasma beta-carotene, after removing the outliers of each pair of plasma retinol and variables, the distributions become wider and more randomly spread.

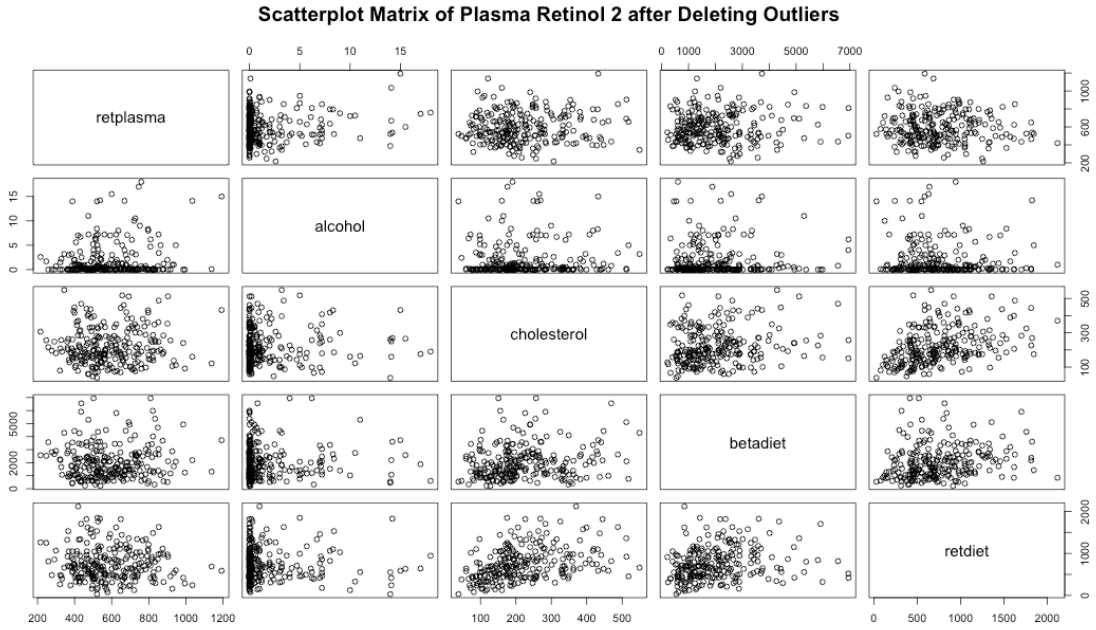
Scatterplot Matrix of Plasma Retinol





Scatterplot Matrix of Plasma Beta-carotene after Deleting Outliers





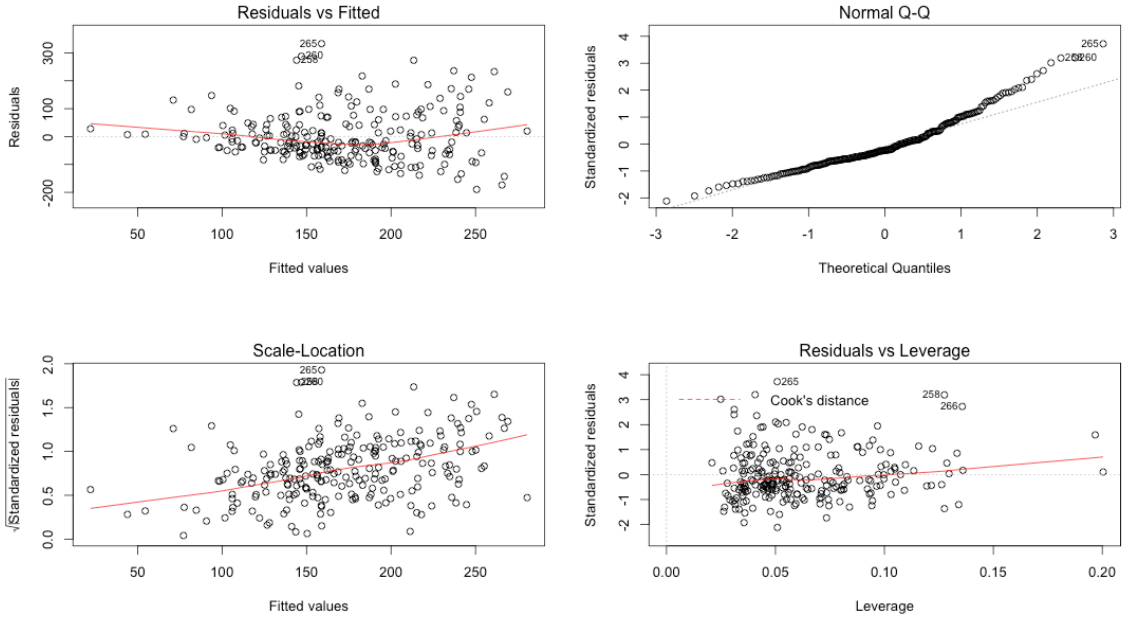
3.3. Best Subset Regression and Stepwise Regression to Select Model Variables - Plasma Beta-carotene Case

In the analysis process, factor variables are generated for the three categorical variables (sex, vitamin and smoking), and named as sex2, vitamin2, vitamin3, smoking2,

and smoking3 in R. This study utilizes a multiple linear regression with respect to the dependent variable plasma beta-carotene.

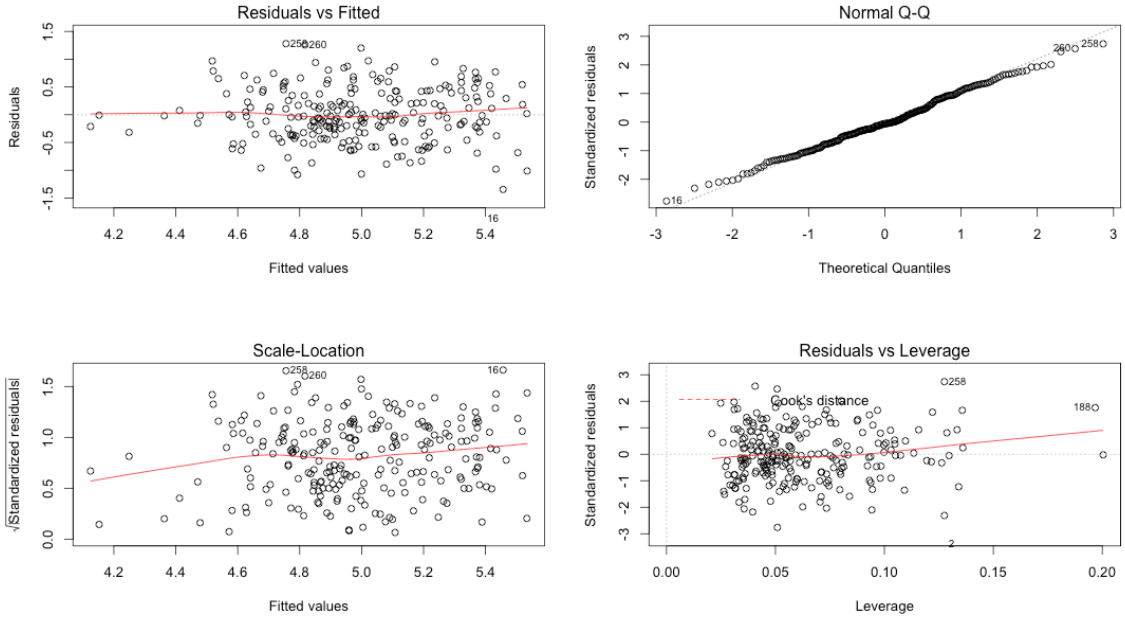
betaplasma~age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol +betadiet+retdiet

Diagnostic Plots for Plasma Beta-carotene



From the diagnostic plots, we can check the normality by viewing residuals plot and QQ plot. The points in QQ plot are away from the straight line which indicates the violation of normality. Therefore, we consider applying a transformation of dependent variables in order to fit the normality assumption. We use log function to do a translation, and use log (plasma beta-carotene) to replace plasma beta-carotene. From recreated diagnostic plots after log-transformation of plasma beta-carotene levels due to severe asymmetry of the residuals on the original scale, we got new diagnostic plots where data almost obeyed normal distribution as we expected (see below).

Diagnostic Plots for Log Plasma Beta-carotene



Log(betaplasma)~age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadiet+retdiet

Regression Result for Log Plasma Beta-carotene

Variables	Coefficients	Std.	t value	Pr(>|t|)	
(Intercept)	5.36E+00	2.99E-01	17.892	2.00E-16	***
age	5.03E-03	2.55E-03	1.973	0.0497	*
sex2	9.68E-02	1.20E-01	0.808	0.4197	
smoking2	-1.22E-01	7.42E-02	-1.647	0.1009	
smoking3	-2.58E-01	1.04E-01	-2.484	0.0137	*
quetelet	-2.61E-02	5.41E-03	-4.823	2.61E-06	***
vitamin2	-2.30E-02	8.37E-02	-0.274	0.7842	
vitamin3	-1.65E-01	7.98E-02	-2.061	0.0404	*
calories	-2.66E-04	1.94E-04	-1.369	0.1724	
fat	3.25E-03	3.10E-03	1.049	0.2951	
fiber	1.03E-02	1.13E-02	0.91	0.3636	
alcohol	-1.58E-02	9.88E-03	-1.603	0.1103	
cholesterol	1.94E-05	4.43E-04	0.044	0.9651	
betadiet	6.91E-05	2.93E-05	2.36	0.0191	*
retdiet	1.23E-04	1.02E-04	1.203	0.2303	

Multiple R-squared: 0.233	p-value: 7.938e-08
Adjusted R-squared: 0.1851	AIC: 363.4935

The best subset regression, stepwise regression, forward selection and backward elimination are conducted in this study in order to find the subset of variables in the data resulting in the best performing model. The variable selection results from different selection methods are listed below. As a result, we select age, smoking, quetelet, vitamin, alcohol and betadiet as variables in the final model.

Selection	Final Model
Forward	logbetaplasma ~ age + sex + smoking + quetelet + vitamin + calories + fat + fiber + alcohol + cholesterol + betadiet + retdiet
Backward	logbetaplasma ~ age + smoking + quetelet + vitamin + alcohol + betadiet
Stepwise	logbetaplasma ~ age + smoking + quetelet + vitamin + alcohol + betadiet

Result of Best Subset Selection

# of variables	age	sex2	smoking2	smoking3	quetelet	vitamin2	vitamin3
1					*		
2					*		
3					*		
4	*				*		
5	*				*		*
6	*			*	*		*
7	*		*	*	*		*
8	*	*	*	*	*		*
# of variables	calories	fat	fiber	alcohol	cholesterol	betadiet	retdiet
1							
2						*	
3				*		*	
4				*		*	
5				*		*	
6				*		*	
7				*		*	
8				*		*	

3.4. The Final Model - Plasma Beta-carotene Case

Final Model: Log(betaplasma) ~ age + smoking + quetelet + vitamin + alcohol + betadiet

Regression Result for Final Model for Log Plasma Beta-carotene

Variables	Coefficients	Std.	t value	Pr(>|t|)	
(Intercept)	5.45E+00	1.99E-01	27.379	2.00E-16	***
age	5.39E-03	2.26E-03	2.387	0.01778	*
smoking2	-1.25E-01	7.29E-02	-1.71	0.08863	.
smoking3	-2.58E-01	1.02E-01	-2.546	0.01157	*
quetelet	-2.69E-02	5.26E-03	-5.108	6.84E-07	***
vitamin2	-1.61E-02	8.24E-02	-0.195	0.8456	
vitamin3	-1.56E-01	7.82E-02	-1.992	0.04752	*
alcohol	-2.01E-02	9.50E-03	-2.119	0.0352	*
betadiet	7.45E-05	2.57E-05	2.905	0.00403	**

Multiple R-squared: 0.2217, p-value: 9.73e-10
Adjusted R-squared: 0.1946 AIC: 355.0007

Compared with the original model, most of the p values of the predictors in the final model are smaller than 0.05, which means the predictors are statistically significant related to the dependent variables. The adjusted R-squared is 19.46% in the final model, which is larger than the adjusted R-square in the original model 18.51%. The regression model fit the data well since 19.46% variability can be explained by the final linear regression. AIC also reduces from 363 to 355 after variables selection. In conclusion, a small p value with adjusted R-square=19.46% and R-square=22.17% indicate that the final model is persuasive and could forecast accurately.

Final Model: Log (plasma beta-carotene) = 5.54+ 0.005age – 0.125smoking1 - 0.258smoking2 – 0.027quetelet – 0.016vitamin2 -0.156vitamin3 -0.020 alcohol + 0.0007betadiet

The coefficient of quetelet is negative, which implies that with the increase of quetelet, there would be a decrease of plasma beta-carotene; that is if a person is overweight, his or her plasma beta-carotene would be low. Similarly as with quetelet, the coefficient of alcohol use is also negative, which indicates the more alcohol consumed, the lower level of the plasma beta-carotene.

The coefficients of age and betadiet are positive, which indicates that with the increase of average number of these variables, there would be an increase in the level of plasma beta-carotene. For example, if the person is older or dietary beta-carotene consumed per day is high, the plasma beta-carotene concentration would rise as well.

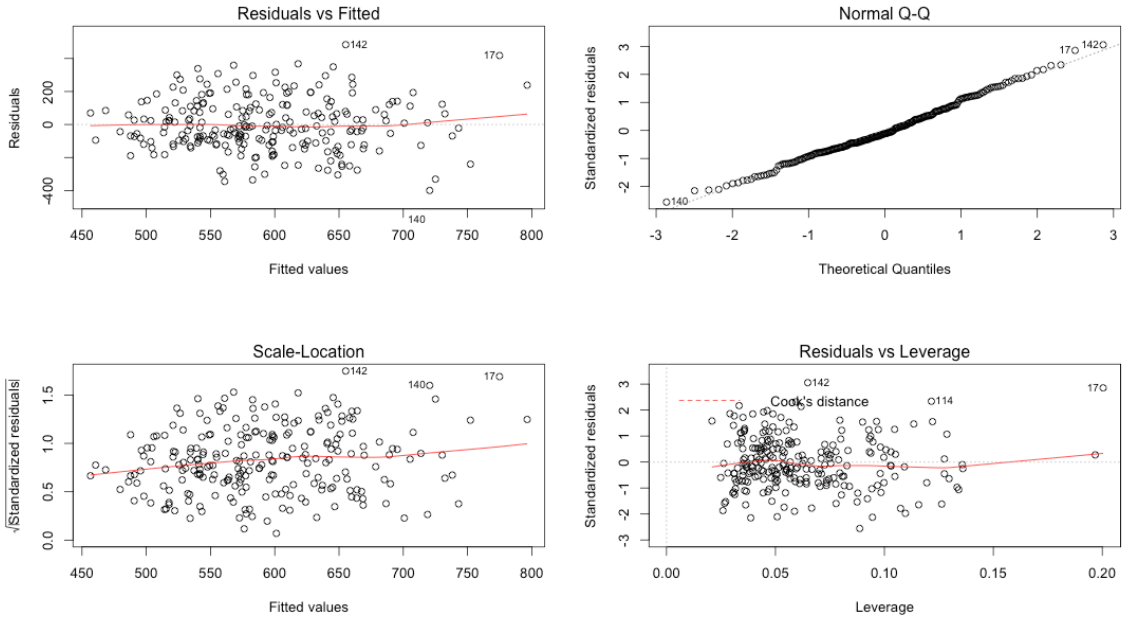
3.5. Best Subset Regression and Stepwise Regression to Select Model Variables - Plasma Retinol Case

To study what are the determinants of human plasma retinol and how those determinants impact on plasma retinol, a multiple linear regression of plasma retinol is performed as similar as the model with plasma beta-carotene. The independent variables includes age, sex, smoking status, quetelet, vitamin use, number of calories consumed per day, grams of fat consumed per day, grams of fiber consumed per day, number of alcoholic drinks consumed per week, cholesterol milligram consumed per day, dietary beta-carotene microgram consumed per day and dietary retinol microgram consumed per

day. From the diagnostic plot, the original regression satisfies the general assumptions of linear regression since residuals are basically randomly distributed and the points in QQ plot obeys normal distribution.

retplasma~age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadiet+retdiet

Diagnostic Plots for Plasma Retinol



Regression Result for Plasma Retinol

Variables	Coefficients	Std.	t value	Pr(>|t|)	
(Intercept)	5.16E+02	1.14E+02	4.538	9.06E-06	***
age	3.07E+00	9.70E-01	3.166	0.00175	**
sex2	-1.04E+02	4.42E+01	-2.353	0.01945	*
smoking2	4.97E+01	2.77E+01	1.793	0.07418	.
smoking3	7.02E+00	3.99E+01	0.176	0.86062	
quetelet	1.93E+00	2.07E+00	0.93	0.35321	
vitamin2	-3.32E+01	3.14E+01	-1.057	0.2914	
vitamin3	-5.02E+01	3.01E+01	-1.667	0.09688	.
calories	8.11E-02	7.42E-02	1.094	0.27522	
fat	-1.75E+00	1.17E+00	-1.497	0.13565	
fiber	-2.48E+00	4.28E+00	-0.579	0.56347	
alcohol	8.14E+00	3.66E+00	2.223	0.02713	*
cholesterol	1.42E-01	1.69E-01	0.841	0.40097	
betadiet	-5.10E-03	1.09E-02	-0.468	0.64037	
retdiet	-4.23E-02	3.89E-02	-1.087	0.27802	

Multiple R-squared: 0.1576 p-value: 0.0001488

Adjusted R-squared: 0.1077 AIC: 3370.255

From the results of the original model for plasma retinol above, it is obvious that the model does not fit the data well given the adjusted R-squared is low (0.1077), which means that only 10.77% variability can be explained by this linear regression. Besides the low R-squared and adjusted R-squared, the AIC is so large which is another indicator of poor goodness of fit. Therefore, best subset regression and stepwise, forward and backward regression are performed again to select model variables that are statistically significantly related to the response variable.

Selection	Final Model
Forward	retplasma ~ age + sex + smoking + quetelet + vitamin + calories + fat + fiber + alcohol + cholesterol + betadiet + retdiet
Backward	retplasma ~ age + sex + alcohol
Stepwise	retplasma ~ age + sex + alcohol

Result of Best Subset Selection

# of variables	age	sex2	smoking2	smoking3	quetelet	vitamin2	vitamin3
1					*		
2					*		
3					*		
4	*				*		
5	*				*		*
6	*			*	*		*
7	*		*	*	*		*
8	*	*	*	*	*		*
# of variables	calories	fat	fiber	alcohol	cholesterol	betadiet	retdiet
1							
2						*	
3				*		*	
4				*		*	
5				*		*	
6				*		*	
7				*		*	
8				*		*	

3.6. The Final Model - Plasma Retinol Case

Final Model: retplasma ~ age + sex + alcohol

Regression Result for Final Model for Plasma Retinol

Variables	Coefficients	Std.	t value	Pr(>|t|)	
(Intercept)	536.1515	68.3342	7.846	1.30E-13	***
age	2.8768	0.8783	3.275	0.00121	**
sex2	-102.5317	41.726	-2.457	0.01469	*
alcohol	8.9426	3.3823	2.644	0.00872	**

Multiple R-squared: 0.1195 p-value: 6.641e-07

Adjusted R-squared: 0.1088 AIC: 3359.355

Although the adjusted R-squared value in the final model does not significantly increase, the value of AIC decreases from 3370.25 to 3359.35 when compared with the original model. The most important thing is that the p value of all predictors in the final model are much smaller than 0.05, that is, all predictors are statistically significant related to the dependent variables. Even the goodness of fit of plasma retinol final model is not ideal, the final model has improved a lot in terms of AIC and p value compared with the original model.

Final Model: plasma retinol = 536.15+ 2.88age - 102.53sex2 + 8.94alcohol

The coefficients of age and alcohol are positive, which indicates with the increase of these variables, there would also be an increase in the plasma retinol level. In other words, if the person is older or the number of alcohol consumed per week is high, the plasma retinol concentration is likely to be high.

4. Conclusion and Discussion

Multiple linear regression models are frequently used in quantitative healthcare research to examine the effects that various factors may have on output variables of interest. This study examined 276 patients who had an elective surgical procedure during a three-year period to explore what factors would influence beta-carotene level and retinol level in human plasma. In this study, there are totally 12 independent variables, including age, sex, smoking status, quetelet, vitamin use, number of calories consumed per day, grams of fat consumed per day, grams of fiber consumed per day, number of alcoholic drinks consumed per week, cholesterol milligram consumed per day, dietary beta-carotene microgram consumed per day and dietary retinol microgram consumed per day; and 2 dependent variables, including plasma beta-carotene (ng/ml) and plasma retinol (ng/ml). The multiple regression models applied in this study show whether and how any of these personal characters would have an effect on plasma beta-carotene and plasma retinol.

There is wide variability in plasma concentrations of these micronutrients in humans, and some of this variability is associated with dietary habits and personal characteristics. As for plasma beta-carotene, we point out that the quetelet value and number of alcohol consumed are negatively related to beta-carotene in human plasma. In other words, if a person is overweight, his or her plasma beta-carotene is more likely to be low. Also the more use of alcohol, the lower the plasma beta-carotene. However, age and betadiet have a positive relationship with beta-carotene in human plasma, which

means that with the increase of the average number of these variables, there would also be an increase in plasma beta-carotene. The older the person is, the more likely his or her plasma beta-carotene would be high. Similarly, the higher his or her dietary beta-carotene consumed per day is, the more likely his or her plasma beta-carotene would be high. As for plasma retinol, age and alcohol have a positive relationship with human plasma retinol. Elderly people and/or alcohol lovers are more likely to be expected to have high plasma retinol.

A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study. Scientists may need to put more effort in evaluating the internal factors that may have an effect or relationship with the beta-carotene and retinol in people plasma.

Reference

These data have not been published yet but a related reference is
Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology* 1989;130:511-521.

Jun Wang, Wen Cui, Shikun Lv, Ningning Sun, A Study of Determinants of Plasma Retinol and Beta-Carotene, Industrial Engineering, Tsinghua University, Fall 2009.

R code

```
plasmaRetinol <- read.delim("~/Documents/6214/6214 final
project/plasmaRetinol.txt", header=FALSE)
View(plasmaRetinol)
colnames(plasmaRetinol) <-
c("age","sex","smoking","quetelet","vitamin","calories","fat","fiber","alcohol","choles
terol","betadiet","retdiet","betaplasma","retplasma")
summary(plasmaRetinol)

#Boxplot of Plasma Retinol and Plasma Beta-carotene by sex
par(mfrow=c(1,2))
boxplot(retplasma~sex,data=plasmaRetinol, main="Boxplot of Plasma Retinol",
      xlab="sex", ylab="retplasma")
boxplot(betaplasma~sex,data=plasmaRetinol, main="Boxplot of Beta-carotene",
      xlab="sex", ylab="betaplasma")

#Boxplot of Plasma Retinol and Plasma Beta-carotene by vitamin
boxplot(retplasma~vitamin,data=plasmaRetinol, main="Boxplot of Plasma Retinol",
      xlab="vitamin", ylab="retplasma")
boxplot(betaplasma~vitamin,data=plasmaRetinol, main="Boxplot of Beta-
carotene",
      xlab="vitamin", ylab="betaplasma")

#Boxplot of Plasma Retinol and Plasma Beta-carotene by smoking
boxplot(retplasma~smoking,data=plasmaRetinol, main="Boxplot of Plasma
Retinol",
      xlab="smoking", ylab="retplasma")
boxplot(betaplasma~smoking,data=plasmaRetinol, main="Boxplot of Beta-
carotene",
      xlab="smoking", ylab="betaplasma")

#Simple Histogram
hist(plasmaRetinol$age,main="Histogram of Age",xlab="Age")
```

```
par(mfrow=c(2,2))
hist(plasmaRetinol$quetelet,main="Histogram of Quetelet",xlab="Quetelet")
hist(plasmaRetinol$calories,main="Histogram of Calories",xlab="Calories")
hist(plasmaRetinol$fat,main="Histogram of Fat",xlab="Fat")
hist(plasmaRetinol$fiber,main="Histogram of Fiber",xlab="Fiber")

par(mfrow=c(2,2))
hist(plasmaRetinol$alcohol,main="Histogram of Alcohol",xlab="Alcohol")
hist(plasmaRetinol$cholesterol,main="Histogram of Cholestrol",xlab="Cholestrol")
hist(plasmaRetinol$betadiet,main="Histogram of Betadiet",xlab="Betadiet")
hist(plasmaRetinol$retdiet,main="Histogram of Ritdiet",xlab="Ritdiet")

#scatter plot of Plasma Beta-carotene
plot(plasmaRetinol$age,plasmaRetinol$betaplasma,main="Scatterplot of Plasma
Beta-carotene and Age",xlab="Age", ylab="Plasma Beta-carotene")
pairs(~betaplasma+quetelet+calories+fat+fiber,data=plasmaRetinol,
    main="Scatterplot Matrix of Plasma Beta-carotene 1")
pairs(~betaplasma+alcohol+cholesterol+betadiet+retdiet,data=plasmaRetinol,
    main="Scatterplot Matrix of Plasma Beta-carotene 2")

#scatter plot of Plasma Retinol
plot(plasmaRetinol$age,plasmaRetinol$retplasma,main="Scatterplot of Plasma
Retinol and Age",xlab="Age", ylab="Plasma Retinol")
pairs(~retplasma+quetelet+calories+fat+fiber,data=plasmaRetinol,
    main="Scatterplot Matrix of Plasma Retinol 1")
pairs(~retplasma+alcohol+cholesterol+betadiet+retdiet,data=plasmaRetinol,
    main="Scatterplot Matrix of Plasma Retinol 2")

#age
attach(mtcars)
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(plasmaRetinol$age,main="Histogram of Age",xlab="Age")
plot(plasmaRetinol$age,plasmaRetinol$betaplasma,main="Scatterplot of Plasma
Beta-carotene and Age",xlab="Age", ylab="Plasma Beta-carotene")
plot(plasmaRetinol$age,plasmaRetinol$retplasma,main="Scatterplot of Plasma
Retinol and Age",xlab="Age", ylab="Plasma Retinol")

#cleaning data
plasmaRetinol_sub<-
subset(plasmaRetinol,calories<4000&fat<150&fiber<30&alcohol<20&cholesterol<
600 &betadiet<7000 &retdiet<2200 &betaplasma<500&retplasma<1200)
View(plasmaRetinol_sub)

plasmaRetinol_sub<-
subset(plasmaRetinol,calories<4000&fat<150&fiber<30&alcohol<20&cholesterol<
600 &betadiet<7000 &retdiet<2200)
```

```
View(plasmaRetinol_sub)

#scatterplot after cleaning data
#scatter plot of Plasma Beta-carotene after deleting
pairs(~betaplasma+quetelet+calories+fat+fiber,data=plasmaRetinol_sub,
    main="Scatterplot Matrix of Plasma Beta-carotene 1 after Deleting Outliers")
pairs(~betaplasma+alcohol+cholesterol+betadiet+retdiet,data=plasmaRetinol_sub,
    main="Scatterplot Matrix of Plasma Beta-carotene 2 after Deleting Outliers")

#scatter plot of Plasma Retinol after deleting
pairs(~retplasma+quetelet+calories+fat+fiber,data=plasmaRetinol_sub,
    main="Scatterplot Matrix of Plasma Retinol 1 after Deleting Outliers")
pairs(~retplasma+alcohol+cholesterol+betadiet+retdiet,data=plasmaRetinol_sub,
    main="Scatterplot Matrix of Plasma Retinol 2 after Deleting Outliers")

#factor categorical variables
plasmaRetinol_sub$sex=as.factor(plasmaRetinol_sub$sex)
plasmaRetinol_sub$smoking=as.factor(plasmaRetinol_sub$smoking)
plasmaRetinol_sub$vitamin=as.factor(plasmaRetinol_sub$vitamin)

#################### Plasma Beta-carotene ##################
#regression analysis
lm1<-lm(betaplasma~
age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadi
et+retdiet, data=plasmaRetinol_sub)
summary(lm1)
#Diagnostic Plots for Plasma Beta-carotene
par(mfrow=c(2,2))
plot(lm1)

#log betaplasma
logbetaplasma=log(plasmaRetinol_sub$betaplasma)
lm2<-lm(logbetaplasma~
age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadi
et+retdiet, data=plasmaRetinol_sub)
summary(lm2)
#Diagnostic Plots for Log Plasma Beta-carotene
par(mfrow=c(2,2))
plot(lm2)

#model fit
library(MASS)
stepAIC(lm2)
AIC.choice <- lm2
summary(AIC.choice)
AIC(lm2)
```

```
#Forward,Backword,and Stepwise selection
library(MASS)
step.forward <- stepAIC(lm2, direction="forward")
step.forward$anova
step.backward <- stepAIC(lm2, direction="backward")
step.backward$anova
step.both <- stepAIC(lm2, direction="both")
step.both$anova

#Best Subset Selection
library(leaps)
attach(plasmaRetinol_sub)
leaps<-
regsubsets(logbetaplasma~age+sex+smoking+quetelet+vitamin+calories+fat+fiber
+alcohol+cholesterol+betadiet+retdiet,data=plasmaRetinol_sub,nbest=1)
summary(leaps)

#final model
lm5<-lm(logbetaplasma~ age + smoking + quetelet + vitamin + alcohol +  betadiet,
data=plasmaRetinol_sub)
summary(lm5)
stepAIC(lm5)
AIC.choice <- lm5
summary(AIC.choice)
AIC(lm5)

##################### Plasma Retinol ###################
#regression analysis
lm3<-lm(retplasma~
age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadi
et+retdiet, data=plasmaRetinol_sub)
summary(lm3)
#Diagnostic Plots for Plasma Retinol
par(mfrow=c(2,2))
plot(lm3)

#model fit
library(MASS)
stepAIC(lm3)
AIC.choice <- lm3
summary(AIC.choice)
AIC(lm3)

#Forward,Backword,and Stepwise selection
library(MASS)
```

```
step.forward <- stepAIC(lm3, direction="forward")
step.forward$anova
step.backward <- stepAIC(lm3, direction="backward")
step.backward$anova
step.both <- stepAIC(lm3, direction="both")
step.both$anova

#Best Subset Selection
library(leaps)
attach(plasmaRetinol_sub)
leaps<-
regsubsets(retaplasma~age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alc
ohol+cholesterol+betadiet+retdiet,data=plasmaRetinol_sub,nbest=1)
summary(leaps)

#final model
lm6<-lm(retplasma~ age + sex + alcohol, data=plasmaRetinol_sub)
summary(lm6)
stepAIC(lm6)
AIC.choice <- lm6
summary(AIC.choice)
AIC(lm6)

#################### Plasma Retinol by Plasma Beta-carotene
##################
#regression analysis
lm4<-lm(retplasma~
age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alcohol+cholesterol+betadi
et+retdiet+betaplasma, data=plasmaRetinol_sub)
summary(lm4)
#Diagnostic Plots for Plasma Retinol by Plasma Beta-carotene
par(mfrow=c(2,2))
plot(lm4)

#model fit
library(MASS)
stepAIC(lm4)
AIC.choice <- lm4
summary(AIC.choice)
AIC(lm4)

#Forward,Backword,and Stepwise selection
library(MASS)
step.forward <- stepAIC(lm4, direction="forward")
step.forward$anova
step.backward <- stepAIC(lm4, direction="backward")
```

```
step.backward$anova
step.both <- stepAIC(lm4, direction="both")
step.both$anova

#Best Subset Selection
library(leaps)
attach(plasmaRetinol_sub)
leaps<-
regsubsets(retaplasma~age+sex+smoking+quetelet+vitamin+calories+fat+fiber+alc
ohol+cholesterol+betadiet+retdiet+betaplasma,data=plasmaRetinol_sub,nbest=1)
summary(leaps)
```