Association Between Personal Characteristics and Levels of Plasma Retinol and Beta-Carotene

Zhengqian Cui (zhqcui@ucdavis.edu),
Zichun Hu (zichu@ucdavis.edu),
Mingqian Zhang (pazhang@ucdavis.edu)

Abstract

The study investigates the association between personal characterictics and carotene and retinol level in plasma. However, the predictors can hardly explain the variation. Basically, age and alcohol consumption affect more on the levels of plasma beta-carotene and retinol.

1 Introduction

1.1 About the Data

Our data contains 315 observations and 14 variables, none of which has missing values.

Variable	Explanation	Variable	Explanation
AGE	In years	FIBER	Grams of fiber consumed per day
SEX	"MALE" or "FEMALE"	ALCOHOL	Number of alcoholic drinks consumed per week
SMOKSTAT	"NEVER", "FORMER", "CURRENT"	CHOLESTEROL	Cholesterol consumed (mg per day)
QUETELET	$weight/height^2$	BETADIET	Dietary beta-carotene consumed (mcg per day)
VITUSE	Vitamin use, "NO", "NOT OFTEN" or "OFTEN"	RETDIET	Dietary retinol consumed (mcg per day)
CALORIES	Number of calories consumed per day	BETAPLASMA	Plasma beta-carotene (ng/ml)
FAT	Grams of fat consumed per day	RETPLASMA	Plasma Retinol (ng/ml)

Table 1: The explanation of each variable of the data

The given variables in our data are not identical to those shown in the reference report (1) list in the introduction of our data. This also reminds us not to stick to the conclusion in the reference report, but to draw some conclusions based on the data we have now. Generally speaking, *BETAPLASMA* and *RETPLASMA* are the quantities that we care about, i.e., as response variables. The final aim (beyond what we can do depending on the data itself) should be to evaluate the relative risk of suffering from the certain types of cancer, but we don't know the specific linkage between the levels of beta-carotene and retinol in the plasma and the risk of cancer, so our goal is to predict the two variables' values based on the remaining 12 variables (personal characteristics) in the data. Since the correlation between *RETPLASMA* and *BETAPLASMA* is only around 0.07, we treat them as two responses separately and develop linear models to predict them respectively.

1.2 Brief Summary of the Data

Three of the variables, *SEX*, *SMOKSTAT* and *VITUSE*, are categorical, while others are numerical. Our explanatory data analysis mainly focus on using the histograms, box plots (include the side-by-side box plots to compare the different distribution of the continuous variables given different categorical variables) and pie charts (for the categorical variables: *SEX*, *SMOKSTAT* and *VITUSE*) of the data to detect the distribution of each variable. The correlation matrix and scatter plots are used to observe the linear relationship among the continuous variables.

By examine the histograms of the two responses, we suspect that a logarithm (add 1 to each value before logarithm to keep the ranges non-negative after transformation) transformation may be good to eliminate the serious skewness. We denote the two corresponding transformed variables as *LOGBETAPLASMA* and *LOGRETPLASMA*. By do so, as Figure 1 shows, the distributions seem to be like normal ones, with an exception that *LOGBETAPLASMA* has a zero. Later we will put more effort on analyzing the association between these two variables and potential predictors. The rationality of doing so will be explained in the next section. Other transformation or processing on the data will also be discussed there.

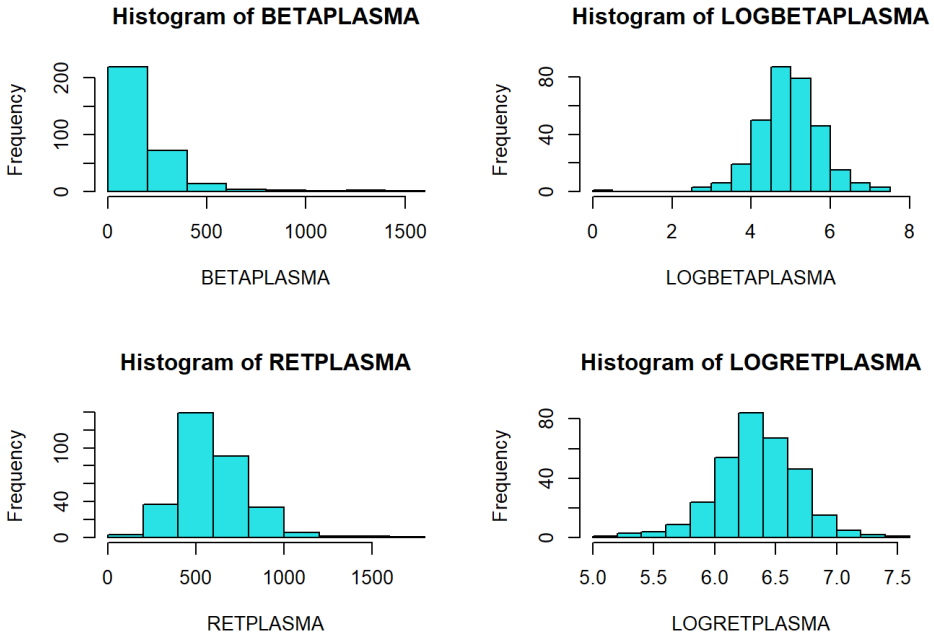


Figure 1: Histograms before and after logarithm transformation

Based on Figure 2, the box plots of 9 potential predictor variables, it can be observed that except *AGE*, the other 8 variables perform a skewed distribution with some potential outlying observations. From the histograms, it can be detected from the Figure 3 that the distribution of AGE has an obvious bi-modality which can not be observed easily in box plots. This finding further enhances our confidence that *AGE* may be categorized for a good performance of some specific model.

Figure 4 shows the three combinations of bi-variate distributions of the three categorical variables. At that point, we have noticed that some variables may be correlated, for example, the habit of smoking

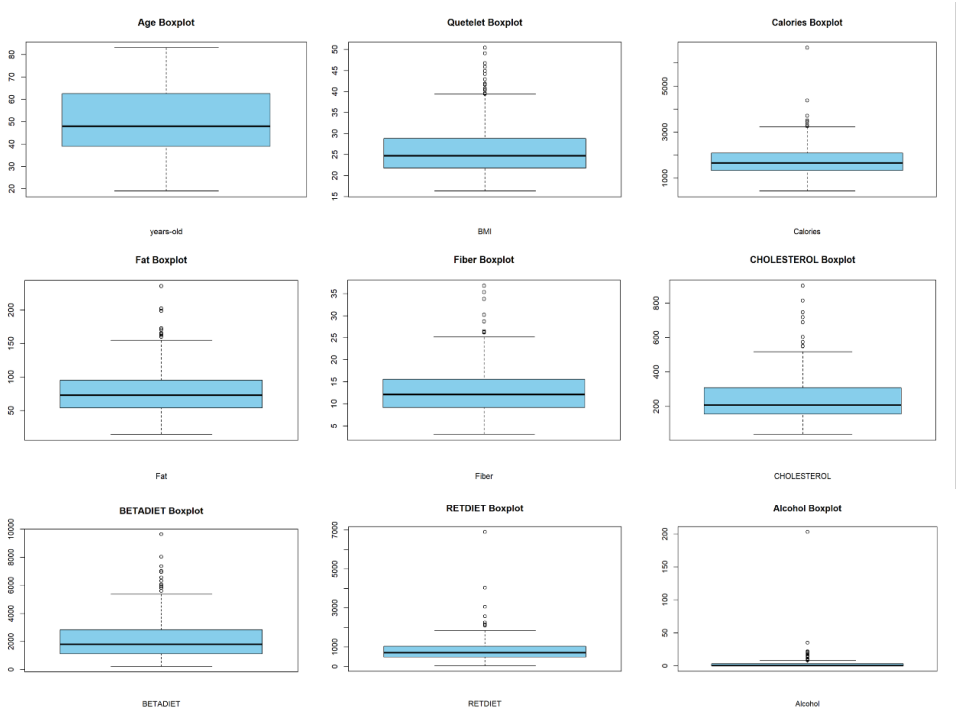


Figure 2: Box plots of continuous potential predictor variables

may be negatively correlated with routinely use of vitamin supplementary. Moreover, we can see some features about uni-variate distribution, such as there are much more females in the data than males.

The box plots of *BETAPLASMA* under three different type of smokers, Figure 5 shows there is a relatively obviously different distribution under these three cases. It can be inferred that the *SMOKSTAT* may be relevant with *BETAPLASMA*, and so the latter is likely to be included in the model predicting the former. And Figure 6 tells there is no obvious difference among the three types of distributions of *LOGRETPLASMA* under the three different frequencies of vitamin use.

The scatter plots Figure 18 show that there is an obvious linear relationship between *FAT* and *CALORIES*. At the same time, it's not as clear but it's still visible that there is an increasing tendency between the *CHOLESTEROL* and *CALORIES*, and *CHOLESTEROL* and *FAT*. After checking the correlation matrix Figure 7, *FAT* and *CALORIES* did have the strongest correlation among all the variables.

2 Methods and Results

To sum up, we developed linear models mainly through two routes: to increase the goodness of fit by introducing higher order terms and interactions or by categorizing some numerical variables into factors. Then we apply several procedures to select the models, and to assess the goodness and generality.

2.1 Models Involving Interactions

This part is mainly tried for predicting *LOGRETPLASMA*, and since the result is not so satisfying, we turn to another approach when predicting *LOGBETAPLASMA*. Firstly, the three categorical variables are transformed into factors. We primarily enter all the 14 predictor variables (including the dummy variables) into the model to fit each of the dependent variables (*BETAPLASMA* and *RETPLASMA*). As shown in Figures 8 and 9 , the regression coefficient determination of the *RETPLASMA* is only approximately 0.1 and the regression coefficient determination of the *BETAPLASMA* is approximately 0.2.

By plotting the plots of residuals versus fitted values and Q-Q plot of the residuals of each model Figure 10 and 11, as a result of the data gathering together caused by the skewness of most of the predictor variables, it is hard to tell whether the vertical distribution of residuals is close to normal directly base on the residuals plots. Therefore, by observing the Q-Q plots of each dependent variable, it is obvious that both of the residuals of the dependent variables have a severe right skewness.

Based on that, the Box-Cox method is used to transform each dependent variable to fix non-normality. By using the Box-Cox, it is observed in Figure 12 that the best λ for both two models are near 0, so we can use the log transformation to each of the dependent variables. After the log transformation, the Q-Q plot of each new residual, Figure 13 also perform more like a normal distribution. At the same time, the obvious skewness of the histogram of the original two dependent variables becomes less significant, and now the histogram is symmetrical after transformation, as shown previously. Therefore, using the logarithm transformed ones to replace the original dependent variables is reasonable.

We would like to detect the influential cases among the observations by using the Cook' distance. By plotting the Cook' distance plot of the regression function *LOGRETPLASMA* versus the 14 prediction variables, it can be detected that the Cook's distance of the 62th observation is larger than 1 and it can be observed that this observation consumed alcohol more than 200 drinks per week, which is well beyond the normal range. Therefore, we decided to delete this observation to avoid the influence on the regression. After deleting, it can be observed that there is an evident improvement in the regression coefficient determination of the *LOGRETPLASMA*, as Figure 14.

Before further exploring, to conduct the internal and external validation and external validation, it is preferred to split data into training data and validation data. Therefore, in the R coding, we use the set.seed () and sample() function to randomly choose 60 data as validation data.

As a result of the low value of the regression coefficient determination (lower than 0.15) in the model of *LOGRETPLASMA* regressing on the 14 original predictor variables, we decided to add the interaction terms to improve the regression coefficient determination. However, to test all three order interaction terms, the number of the terms is too complex to compute.

Therefore, we first use the AIC step wise to choose a reduced model *AGE*, *SMOKSTAT*, *CALORIES*, *FAT*, *FIBER*, and *ALCOHOL*, and we will add the interaction terms based on the reduced model. To avoid a sharp increment in the complexity of the model, one of our strategies is using the for-loop to enter the interaction terms with the criterion that any terms entered into the model should also cause a decrease in the BIC of the model. However, there is no term entering the model. So, the final model chosen on this selection route is shown as Figure 21.

2.2 Models Based on Categorizing Some Continuous Variables

We suspect that the reason why the models above don't work well is partly linked to the fact that some predictors has thick tail distribution or has too many similar values gathering around some point. Also, *AGE* has bi-modal distribution. These remind us that the model may be improved by introducing categorical variables based on the continuous variables. First, for all of the 9 continuous potential predictors, we categorize them respectively, each into two or three levels. Then, for each response, we fit a full model containing all the original predictors and all the new categorical ones, and select some of them using AIC criterion. After that, we try some choices of predictors manually based on our experience and background knowledge. To assess the generality of the models, we use 10-fold cross validation: the data is randomly sampled into 10 subsets, and each time the model is trained on 9 sets while tested on the remained one; then an overall assessment of the prediction is given. After this process, each model has three resulting values: RMSE, MAE and R^2. The RMSE is the square root of mean squared error estimated, and MAE $= \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$ is another indicator of the model's sum of residuals. When the model has homoscedasticity, RMSE and MAE should be near.

2.2.1 Categorization

Age is often categorized to three levels in observation studies (¡40, 40-65 and ¿65), and we follow the tradition. *QUETELET* is often referred to BMI now, and we categorize it by the common criterion. Calories intake should be categorized depending on sex, and since many of the people have too high intake, we categorize it by lower and upper quartile of each sex's calories intake in the data. Fat and fiber intake level are linked to the calories intake (2) (3), so we categorize *FIBER* and *FAT* based on $FIBER/CALORIES$ and $FAT/CALORIES$. The criterion for fiber is based on median since too many are too low compared to the suggestion in Dietary Guidelines for Americans. Fat normally contribute 20-35% of the calories intake (3) so it is categorized by 35% of the calories intake. Alcohol intake is categorized by non consumer, no more than 1 per week, or more. Some study suggests that cholesterol intake influence little about the plasma cholesterol and health (4), and we just categorize it by median in the data. According to other studies (5) (6), we categorize beta-carotene and retinol intake.

2.2.2 Predicting *LOGBETAPLASMA* and *LOGRETPLASMA*

The AIC step procedure suggest only QUETELET + FIBER + CALORIES + VITUSE to predict, but when we try a combination, things gets better. The R_a^2 and cross validation shows that our manual model shown in Figure 16 is the best to predict *LOGBETAPLASMA* among all the models tried. Similarly, the best model predicting *LOGRETPLASMA* is shown in Figure 17. It shows that age and alcohol consumption associate somehow more obvious with beta-carotene and retinol.

2.3 Ridge Regression

Ridge regression is a technique used to analyze multiple regression data that suffer from multicollinearity. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. In this report, we discuss the application of ridge regression to a dataset for predicting plasma retinol concentration, focusing on addressing multicollinearity among predictors. The dataset underwent a 70/30 split into training and testing subsets, followed by standardization to neutralize

scale disparities—crucial for the penalization process inherent to ridge regression. Optimal regularization was sought through a sequence of lambda values (0 to 5), determined via cross-validation techniques to minimize mean squared error, a process visualized in a plot showcasing the trade-off between complexity and fit. The selected lambda minimized the cross-validation error, as depicted by the mean squared error plot against various log(lambda) values, and a coefficient path plot elucidated the impact of regularization on the shrinkage of coefficient estimates.

The best-performing lambda, minimum value was employed to fit the final model, highlighting significant predictors including smoking status, age, sex, body mass index, and dietary factors, each variably affecting plasma retinol levels. The model's coefficients, as revealed by the glmnet output, suggest a nuanced relationship between these variables and the target biomarker, potentially informing further biomedical investigations. Performance evaluation on the test dataset yielded a standardized mean absolute error of 0.783092, indicative of the model's predictive adequacy.

Conclusively, ridge regression has demonstrated its utility in managing multicollinearity and providing robust predictions in the presence of numerous predictors. The resultant model not only underscores the significance of certain biological and lifestyle factors on plasma retinol but also offers a methodological framework for similar epidemiological studies. Future research might enhance precision by exploring more complex models or integrating additional predictive variables, thereby refining the model's utility in nutritional epidemiology and public health surveillance.

3 Conclultion

In conclusion, our analysis using ridge regression and various linear regression models, including first-order, second-order, third-order, and interaction models, has provided a comprehensive exploration of the relationship between the predictors and plasma retinol concentration. Despite the extensive modeling efforts within the linear framework and the elastic net family, the adjusted R-squared values remain relatively low, indicating a modest fit of the models to the data. This suggests that the variability in plasma retinol levels is only partially explained by the predictors included in our models.

References

[1] D. W. NIERENBERG, T. A. STUKEL, J. A. BARON, B. J. DAIN, E. R. GREENBERG, and T. S. C. P. S. GROUP, "DETERMINANTS OF PLASMA LEVELS OF BETA-CAROTENE AND RETINOL," *American Journal of Epidemiology*, vol. 130, pp. 511–521, 09 1989.

[2] L. Bazzano, J. He, L. Ogden, C. Loria, and P. Whelton, "Dietary fiber intake and reduced risk of coronary heart disease in us men and women the national health and nutrition examination survey i epidemiologic follow-up study," *Archives of internal medicine*, vol. 163, pp. 1897–904, 10 2003.

[3] A. Liu, N. Ford, F. Hu, K. Zelman, D. Mozaffarian, and P. Kris-Etherton, "A healthy approach to dietary fats: Understanding the science and taking action to reduce consumer confusion," *Nutrition Journal*, vol. 16, p. 53, 08 2017.

[4] M. Fernandez, "Dietary cholesterol provided by eggs and plasma lipoproteins in healthy populations," *Current opinion in clinical nutrition and metabolic care*, vol. 9, pp. 8–12, 02 2006.

[5] T. Grune, G. Lietz, A. Palou, A. Ross, W. Stahl, G. Tang, D. Thurnham, S.-a. Yin, and H. Biesalski, "Carotene is an important vitamin a source for humans," *The Journal of nutrition*, vol. 140, pp. 2268S–2285S, 10 2010.

[6] H. Gester, "Vitamin a - functions, dietary requirements and safety in humans," *International Journal for Vitamin and Nutrition Research*, vol. 67, pp. 71–90, 6 1996.

A Supplementary Tables and Figures

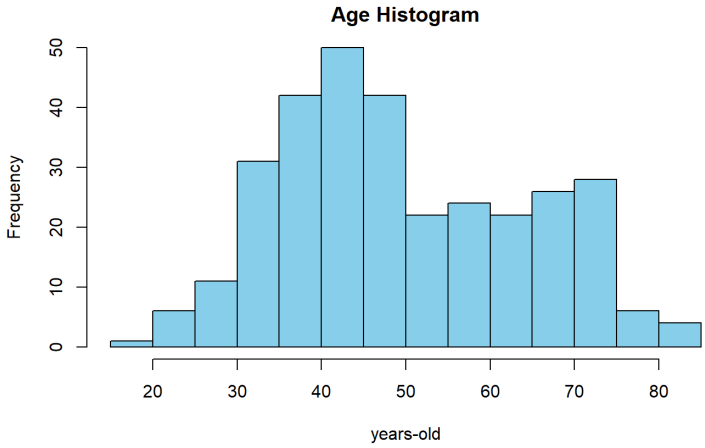


Figure 3: Histogram of *AGE*

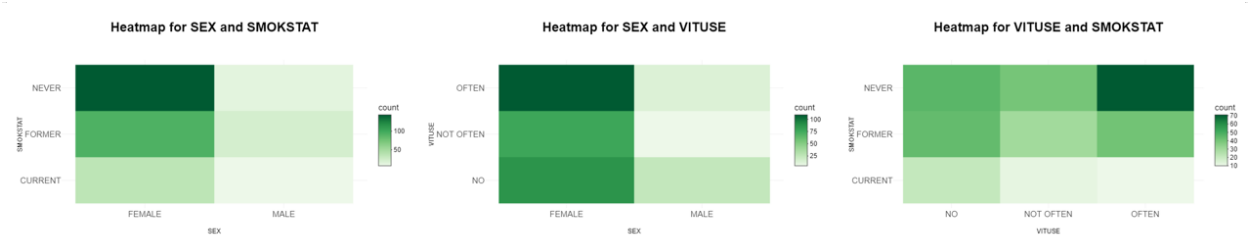


Figure 4: Bi-variate distributions of categorical variables

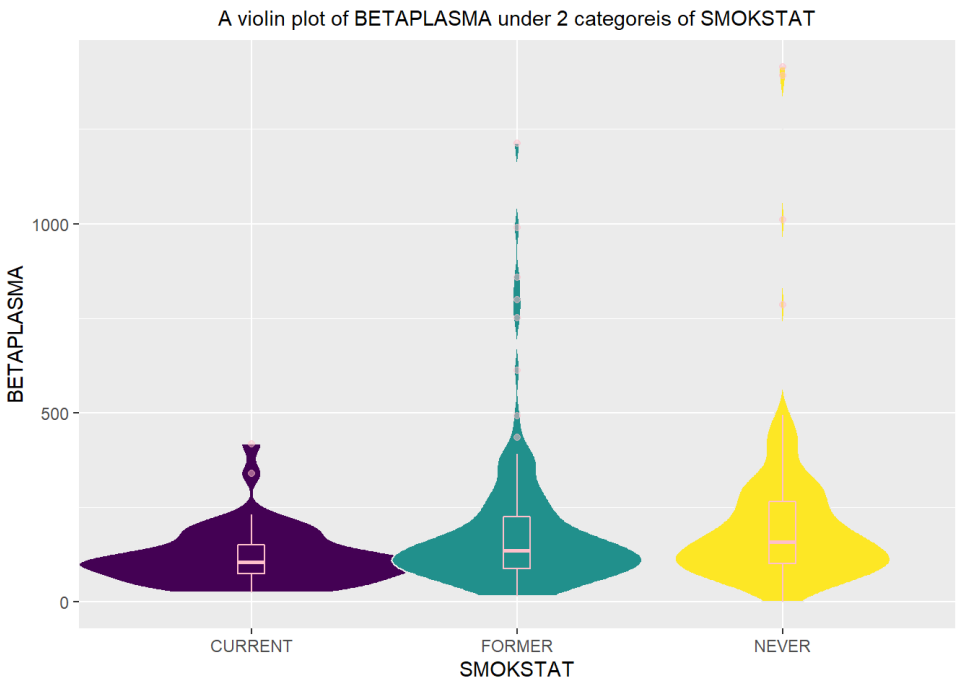


Figure 5: Grouped box plots of *BETAPLASMA* v.s. *SMOKSTAT*

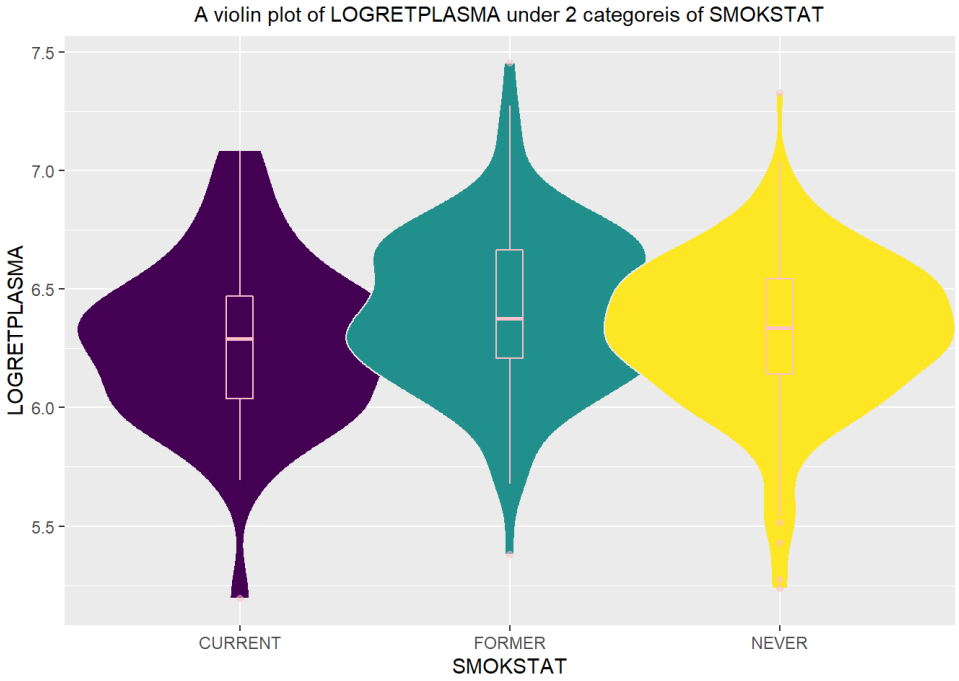


Figure 6: Grouped box plots of *LOGRETPLASMA* v.s. *SMOKSTAT*

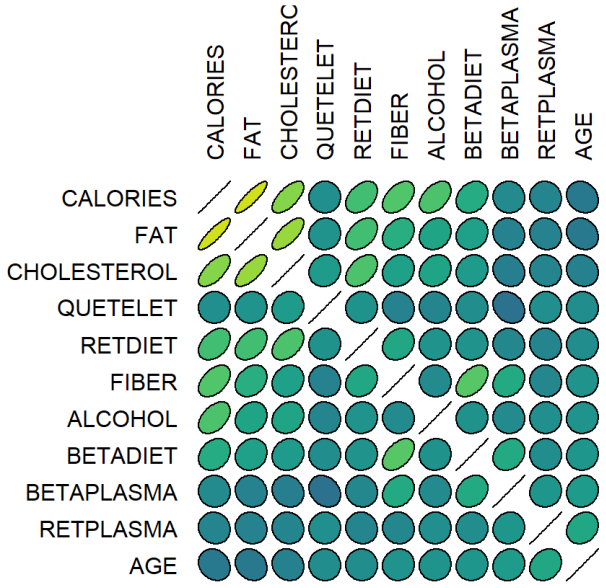


Figure 7: Figure representing the correlation matrix

```
Multiple R-squared:  0.1008,    Adjusted R-squared:  0.05886
```

Figure 8: R^2's for the model for *RETPLASMA*

```
Multiple R-squared:  0.1946,    Adjusted R-squared:  0.157
```

Figure 9: R^2's for the model for *BETAPLASMA*

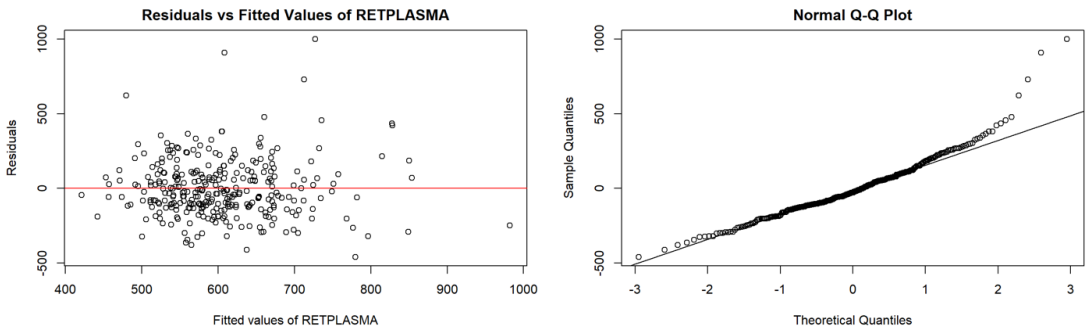


Figure 10: Plots of residuals for the model for *RETPLASMA*

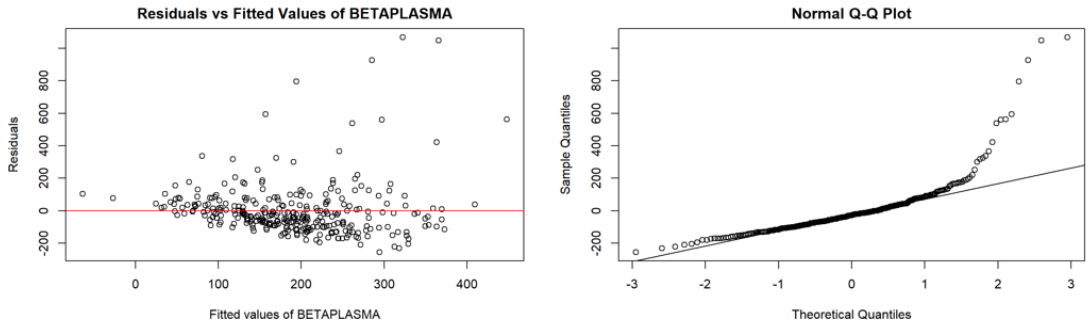


Figure 11: Plots of residuals for the model for *BETAPLASMA*

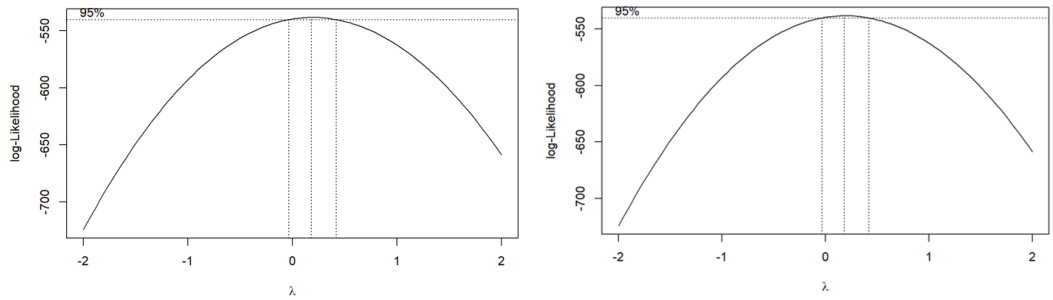


Figure 12: Box-Cox transformation to determine whether logarithm is appropriate

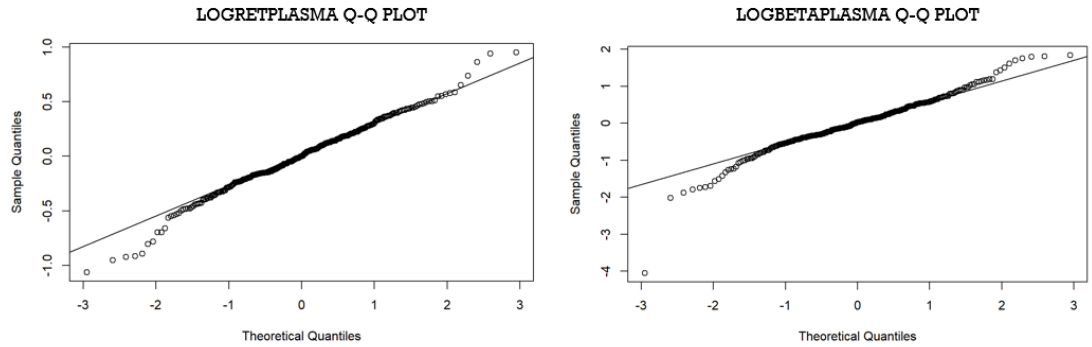


Figure 13: Q-Q Plot for the residuals for models predicting *LOGBETAPLASMA* and *LOGRETPLASMA*

```
Multiple R-squared:  0.1395,    Adjusted R-squared:  0.09925
```

Figure 14: R^2's for the model for *LOGRETPLASMA*

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.089e+00  9.493e-02  64.140  < 2e-16 ***
SMOKSTATFORMER  8.007e-02  3.834e-02   2.088  0.03761 *
AGE             5.458e-03  1.323e-03   4.125 4.79e-05 ***
CALORIES        1.713e-04  9.391e-05   1.824  0.06911 .
FAT            -3.471e-03  1.483e-03  -2.340  0.01991 *
FIBER          -8.764e-03  4.862e-03  -1.802  0.07247 .
ALCOHOL         1.083e-02  4.031e-03   2.688  0.00759 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3202 on 307 degrees of freedom
Multiple R-squared:  0.1253,    Adjusted R-squared:  0.1082
F-statistic: 7.327 on 6 and 307 DF,  p-value: 2.471e-07
```

Figure 15: The model predicting *LOGRETPLASMA*, after trying to add interactions

```
Call:
lm(formula = LOGBETAPLASMA ~ AGECAT + SMOKSTAT + QUETELET + VITUSE +
    FIBER + ALCOHOLCAT + CHOLESTEROL + BETADIET, data = plasma)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84437 -0.36581 -0.01082  0.39759  1.77760

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.912e+00  2.342e-01  20.977  < 2e-16 ***
AGECAT40-65             2.081e-01  8.967e-02   2.321  0.02097 *
AGECAT65+               2.497e-01  1.117e-01   2.235  0.02614 *
SMOKSTATFORMER          1.692e-01  1.236e-01   1.368  0.17229
SMOKSTATNEVER           2.958e-01  1.198e-01   2.468  0.01415 *
QUETELET               -2.993e-02  6.548e-03  -4.570 7.13e-06 ***
VITUSENOT OFTEN         3.064e-01  9.732e-02   3.149  0.00181 **
VITUSEOFTEN             3.162e-01  8.927e-02   3.542  0.00046 ***
FIBER                   1.995e-02  8.126e-03   2.455  0.01467 *
ALCOHOLCATNot frequent  1.586e-01  9.501e-02   1.669  0.09616 .
ALCOHOLCATFrequent      1.682e-01  9.420e-02   1.785  0.07524 .
CHOLESTEROL            -8.291e-04  3.076e-04  -2.695  0.00743 **
BETADIET                4.960e-05  2.913e-05   1.703  0.08963 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6544 on 300 degrees of freedom
Multiple R-squared:  0.2509, Adjusted R-squared:  0.2209
F-statistic: 8.373 on 12 and 300 DF,  p-value: 1.167e-13
```

Figure 16: The model predicting *LOGRETPLASMA*, after categorization and manual selection

```
Call:
lm(formula = LOGRETPLASMA ~ AGECAT + SEX + SMOKSTAT + FATCAT +
    FIBER + ALCOHOLCAT + BETADIETCAT, data = plasma)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9818 -0.2111 -0.0068  0.2037  0.9810

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.284658   0.081914  76.723  < 2e-16 ***
AGECAT40-65             0.104649   0.043234   2.421  0.01609 *
AGECAT65+               0.213358   0.056045   3.807  0.00017 ***
SEXMALE                 0.096033   0.057231   1.678  0.09439 .
SMOKSTATFORMER          0.095686   0.059673   1.604  0.10987
SMOKSTATNEVER           0.023413   0.057484   0.407  0.68408
FATCATToo much         -0.094038   0.040819  -2.304  0.02191 *
FIBER                  -0.005315   0.003953  -1.344  0.17983
ALCOHOLCATNot frequent  0.098534   0.046602   2.114  0.03530 *
ALCOHOLCATFrequent      0.098631   0.044696   2.207  0.02809 *
BETADIETCATNot too low -0.046631   0.041172  -1.133  0.25828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3205 on 302 degrees of freedom
Multiple R-squared:  0.1185, Adjusted R-squared:  0.08933
F-statistic:  4.06 on 10 and 302 DF,  p-value: 2.989e-05
```

Figure 17: The model predicting *LOGRETPLASMA*, after categorization and manual selection

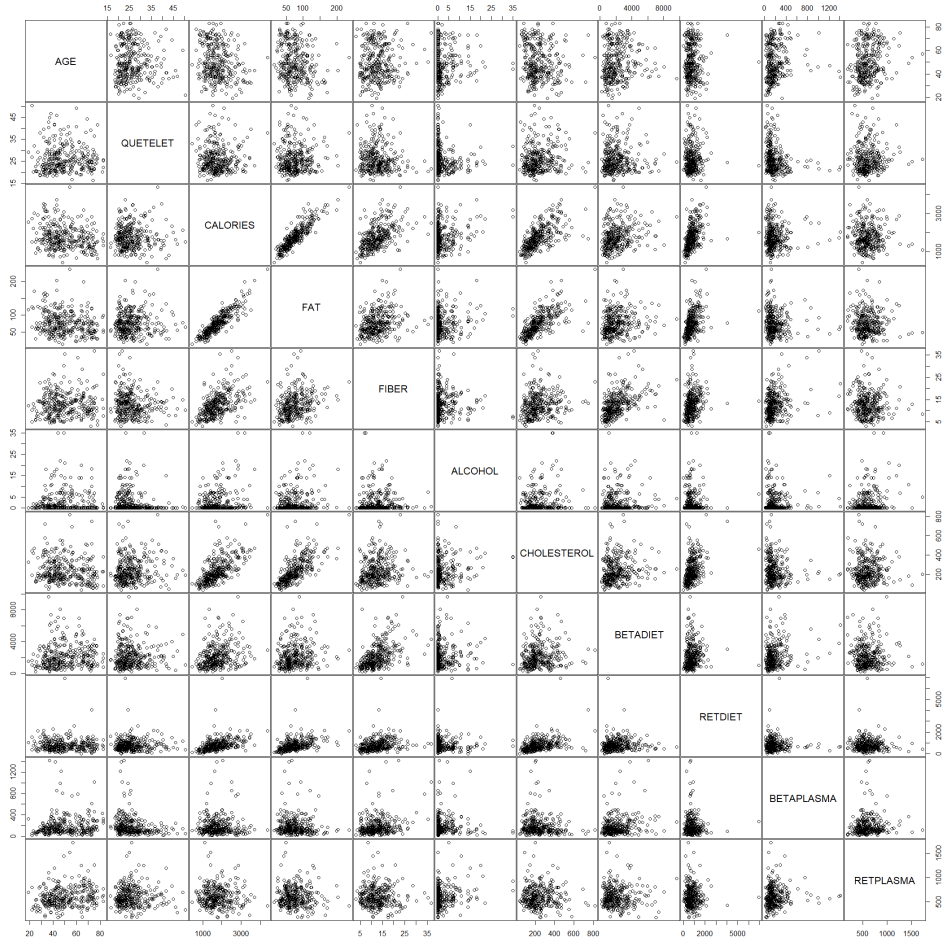


Figure 18: Paired scatter plots between all the continuous variables

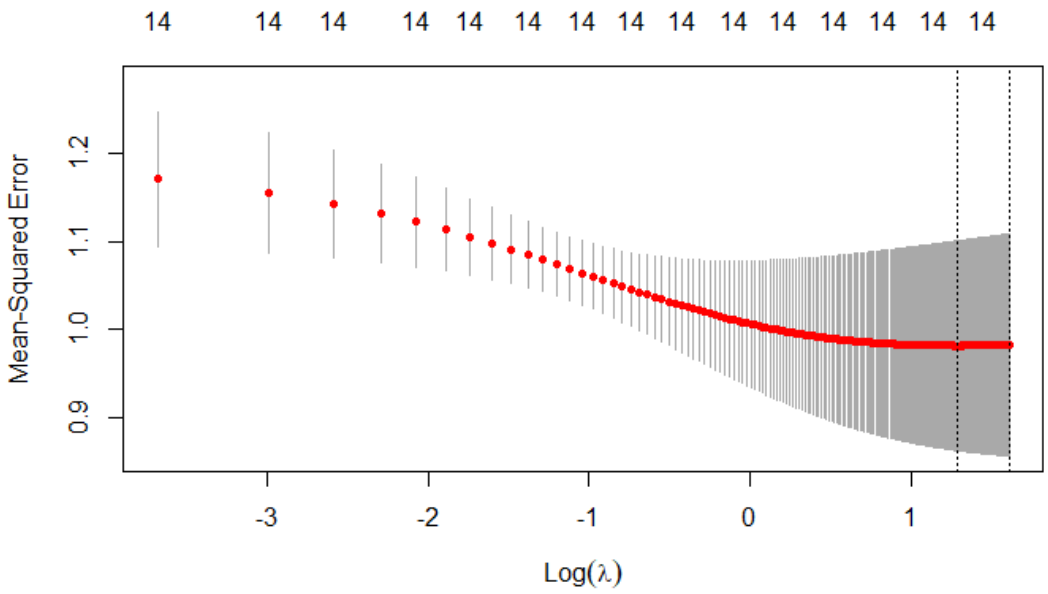


Figure 19: Mean Sqaure Error of Different λ

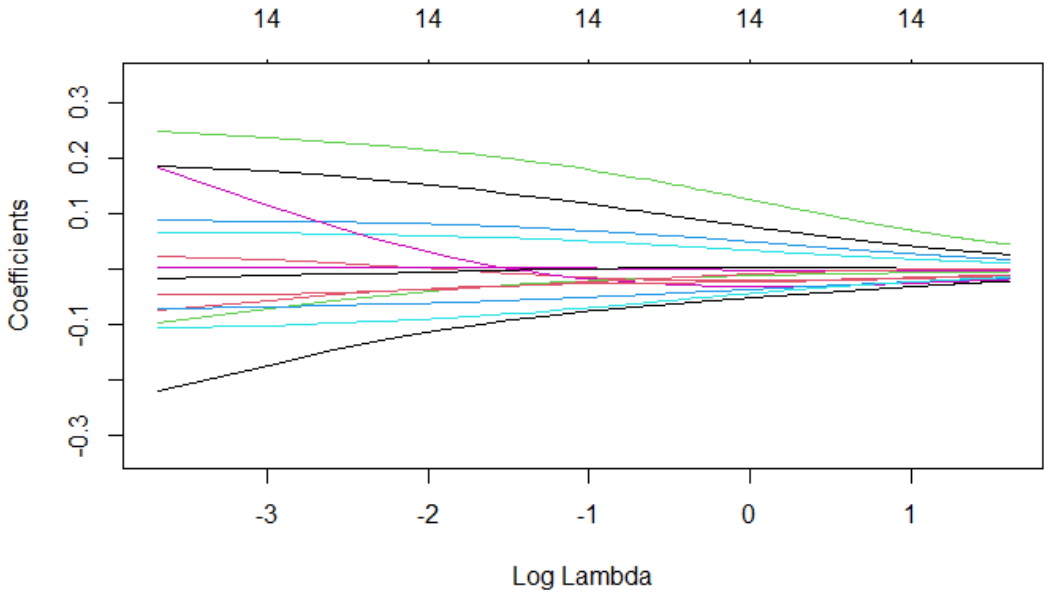


Figure 20: Coefficients vs for Ridge Regression

```
15 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)     1.713944e-16
SMOKSTATFORMER  3.321677e-02
SMOKSTATNEVER  -1.346340e-02
AGE             5.629206e-02
SEX             2.180310e-02
QUETELET        1.387502e-02
CALORIES       -2.281896e-02
FAT            -2.734142e-02
FIBER          -1.522317e-02
ALCOHOL        -5.966724e-03
CHOLESTEROL    -1.994772e-02
BETADIET       -1.938990e-02
RETDIET        -4.587777e-03
Vituse_Often    4.016285e-04
Vituse_No      -6.477459e-04
```

Figure 21: R Result of GLM

Startup

```
setwd("E:\\2023 fall\\206\\project")
plasma <- read.table("Plasma.txt", header = TRUE)
attach(plasma)

plasma$SEX <- as.factor(SEX)
plasma$SMOKSTAT <- as.factor(SMOKSTAT)
plasma$VITUSE <- as.factor(VITUSE)
plasma$LOGBETAPLASMA <- log(plasma$BETAPLASMA+1)
plasma$LOGRETPLASMA <- log(plasma$RETPLASMA+1)

plasma$AGECAT <- cut(plasma$AGE, breaks = c(0, 40, 65, 100), labels = c("40-",
                                                                        "40-65", "65+"))

plasma$QUETELETCAT <- cut(plasma$QUETELET, breaks = c(0, 18.5, 25, 30, 60),
                          right = FALSE,
                          labels = c("Underweight", "Healthy Weight", "Overweight", "Obesit
y"))

CAL_MALE_L <- quantile(CALORIES[which(plasma$SEX == "MALE")], 1/3)
CAL_MALE_H <- quantile(CALORIES[which(plasma$SEX == "MALE")], 2/3)
CAL_FEMALE_L <- quantile(CALORIES[which(plasma$SEX == "FEMALE")], 1/3)
CAL_FEMALE_H <- quantile(CALORIES[which(plasma$SEX == "FEMALE")], 2/3)
plasma$CALCAT <- NA


for (i in 1:nrow(plasma)) {
  if (plasma$SEX[i] == "MALE") {
    if (plasma$CALORIES[i] <= CAL_MALE_L) {
      plasma$CALCAT[i] <- "LOW"
    } else if (plasma$CALORIES[i] <= CAL_MALE_H) {
      plasma$CALCAT[i] <- "MED"
    } else {plasma$CALCAT[i] <- "HIGH"}
  } else {
    if (plasma$CALORIES[i] <= CAL_FEMALE_L) {
      plasma$CALCAT[i] <- "LOW"
    } else if (plasma$CALORIES[i] <= CAL_FEMALE_H) {
      plasma$CALCAT[i] <- "MED"
    } else {plasma$CALCAT[i] <- "HIGH"}
  }
}
plasma$CALCAT <- as.factor(plasma$CALCAT)

plasma$FATCATINIT <- cut(plasma$FAT, breaks = c(0, 72.9, 500), labels = c("Too low", "Not too
low"))
plasma$FAT.CAL <- plasma$FAT * 9 / plasma$CALORIES
plasma$FATCAT <- cut(plasma$FAT.CAL, breaks = c(0, 0.35, 1), labels = c("Not too much", "Too
much"))
plasma$FIBER.CAL <- plasma$FIBER / plasma$CALORIES * 1000
plasma$FIBERCAT <- cut(plasma$FIBER.CAL, breaks = c(0, 8.86, 50), labels = c("Too low", "Not
too low"))
plasma$FIBERCATINIT <- cut(plasma$FIBER, breaks = c(0, 12.1, 50), labels = c("Too low", "Not
too low"))
plasma$ALCOHOLCAT <- cut(plasma$ALCOHOL, breaks = c(0, 0.1, 1.1, 300), right = FALSE,
```

```
                        labels = c("Never", "Not frequent", "Frequent"))
plasma$BETADIETCAT <- cut(plasma$BETADIET, breaks = c(0, 2000, 10000), labels = c("Too low",
"Not too low"))

plasma$RETDIETCAT <- NA
for (i in 1:nrow(plasma)) {
  if (plasma$SEX[i] == "MALE") {
    if (plasma$RETDIET[i] > 1000) {
      plasma$RETDIETCAT[i] <- "Not too low"
    } else {plasma$RETDIETCAT[i] <- "Too low"}
  } else {
    if (plasma$RETDIET[i] > 800) {
      plasma$RETDIETCAT[i] <- "Not too low"
    } else {plasma$RETDIETCAT[i] <- "Too low"}
  }
}
plasma$RETDIETCAT <- as.factor(plasma$RETDIETCAT)
plasma$CHOLESTEROLCAT <- cut(plasma$CHOLESTEROL, breaks = c(0, 206.1, 1000), labels = c("Rela
tively low", "Relatively high"))
```

```
library(lubridate)
library(viridis)
library(ggplot2)
library(ggthemes)
```

```
## Warning: 程辑包'ggthemes'是用R版本4.1.2 来建造的
```

```
library(hrbrthemes)
```

```
## Warning: 程辑包'hrbrthemes'是用R版本4.1.2 来建造的
```

```
library(plotly)
```

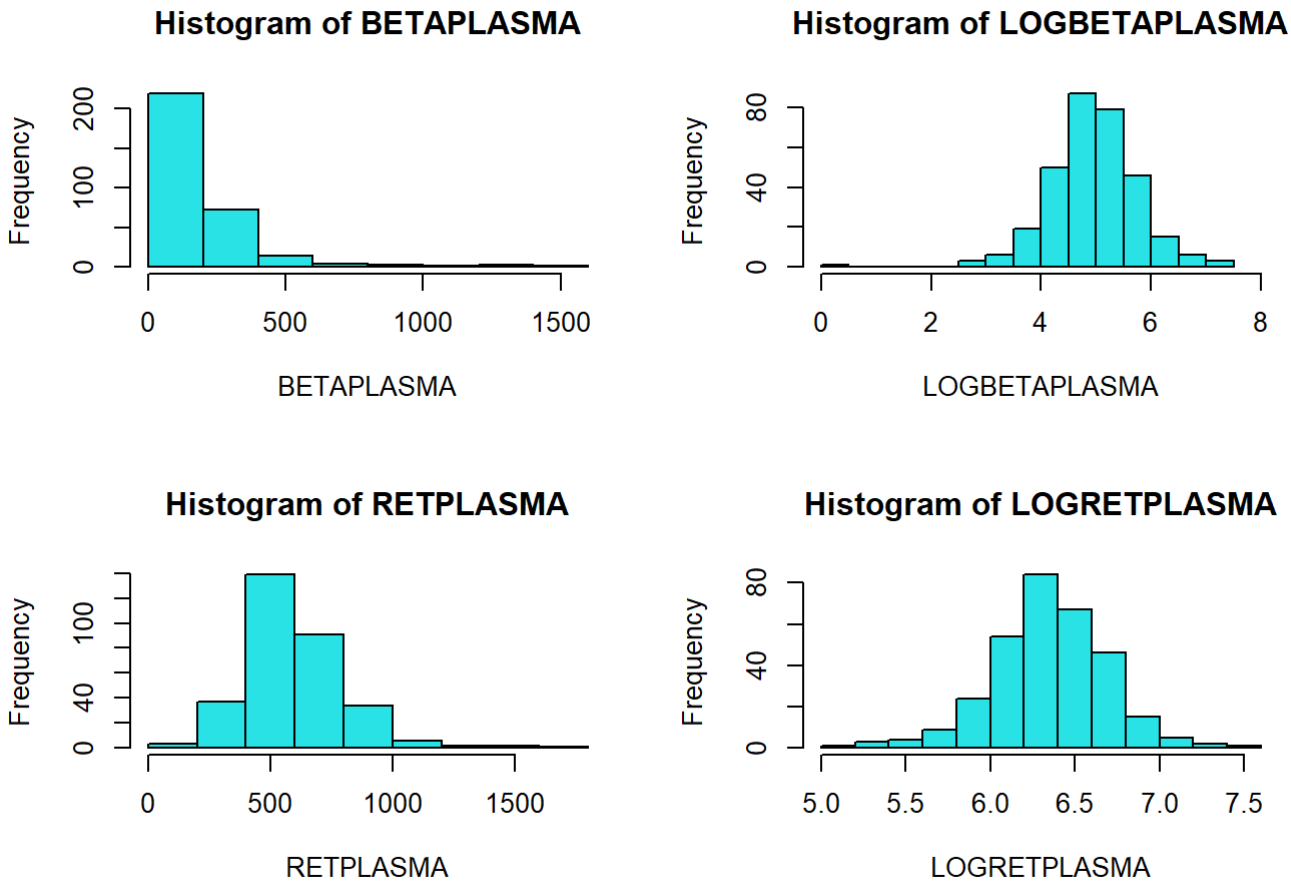
```
## Warning: 程辑包'plotly'是用R版本4.1.2 来建造的
```

```
library(car)
library(dplyr)
```

1. EDA

1.1 Distribution of the responses

```
par(mfrow=c(2,2))
hist(plasma$BETAPLASMA, main = "Histogram of BETAPLASMA", xlab = "BETAPLASMA", col = 5)
hist(plasma$LOGBETAPLASMA, main = "Histogram of LOGBETAPLASMA", xlab = "LOGBETAPLASMA", col =
5, breaks = 12, xlim = c(0,8))
hist(plasma$RETPLASMA, main = "Histogram of RETPLASMA", xlab = "RETPLASMA", col = 5)
hist(plasma$LOGRETPLASMA, main = "Histogram of LOGRETPLASMA", xlab = "LOGRETPLASMA", col = 5)
```




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

1.2 Bivariate Graphs

1.2.1 SEX and SMOKSTAT

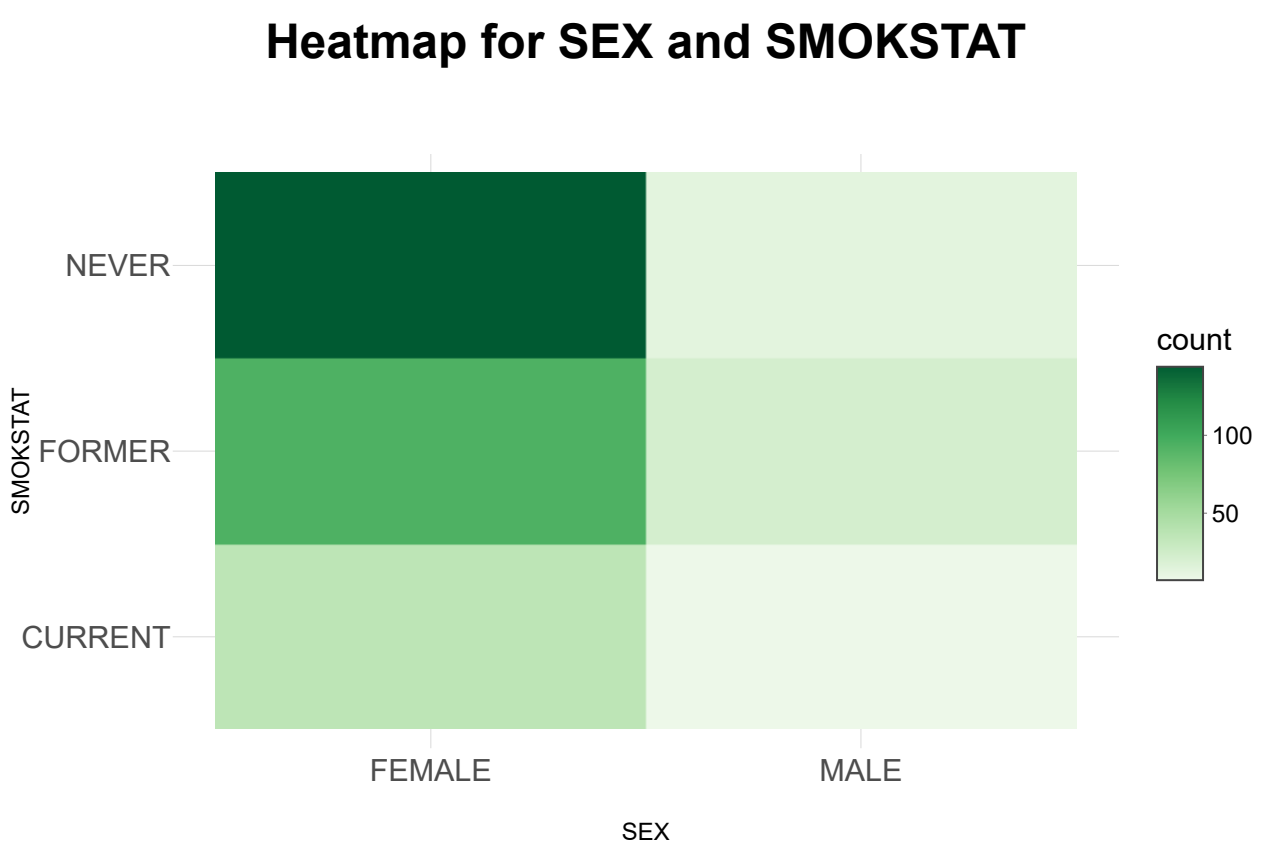
```
tpFrame <- plasma%>%
  group_by(SEX, SMOKSTAT)%>%
  summarize(count=n())
```

```
## `summarise()` has grouped output by 'SEX'. You can override using the `.groups` argument.
```

```
tpFrame <- tpFrame %>%
  mutate(text = paste0("SEX: ", SEX, "\n", "SMOKSTAT: ", SMOKSTAT, "\n", "count: ", count))

p <- ggplot(tpFrame, aes(SEX, SMOKSTAT, fill= count, text=text)) +
  geom_tile() +
  scale_fill_distiller(palette = "viridis", direction = 1) +
  theme_ipsum() +
  ggtitle("Heatmap for SEX and SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning in pal_name(palette, type): Unknown palette viridis
```

```
ggplotly(p, tooltip="text")
```



1.2.2 SEX and VITUSE

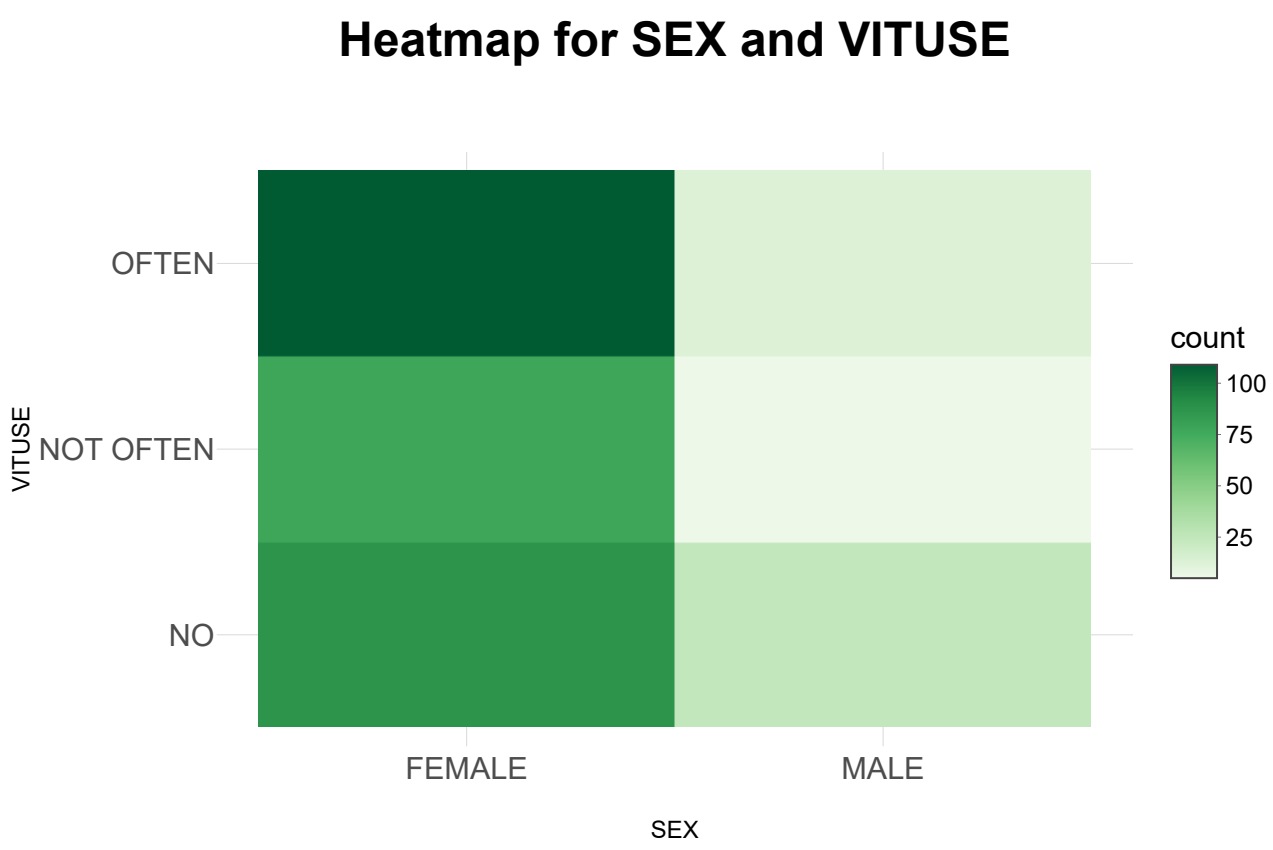
```
tpFrame <- plasma%>%
  group_by(SEX, VITUSE)%>%
  summarize(count=n())
```

```
## `summarise()` has grouped output by 'SEX'. You can override using the `.groups` argument.
```

```
tpFrame <- tpFrame %>%
  mutate(text = paste0("SEX: ", SEX, "\n", "VITUSE: ", VITUSE, "\n", "count: ", count))

p <- ggplot(tpFrame, aes(SEX, VITUSE, fill= count, text=text)) +
  geom_tile() +
  scale_fill_distiller(palette = "viridis", direction = 1) +
  theme_ipsum() +
  ggtitle("Heatmap for SEX and VITUSE") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning in pal_name(palette, type): Unknown palette viridis
```

```
ggplotly(p, tooltip="text")
```



1.2.3 VITUSE and SMOKSTAT

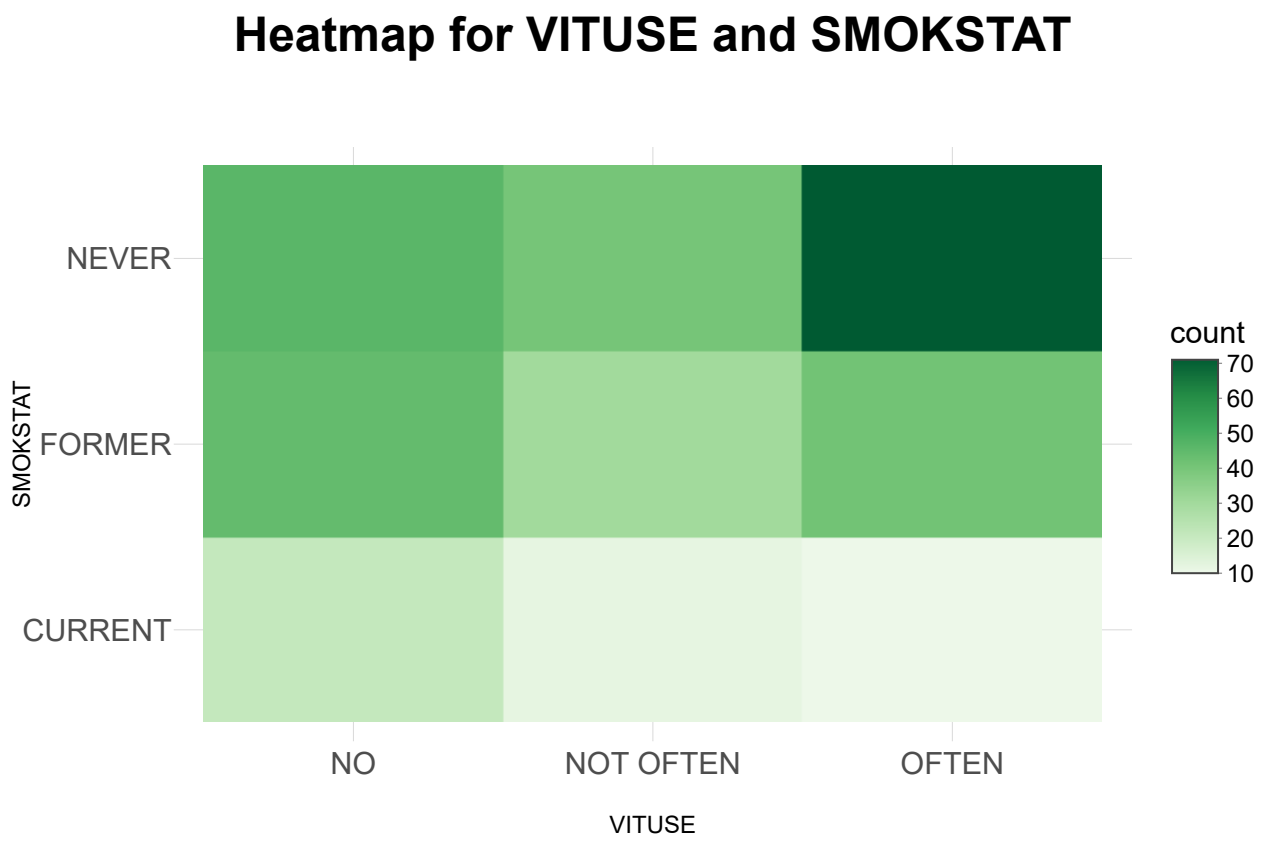
```
tpFrame <- plasma%>%
  group_by(VITUSE, SMOKSTAT)%>%
  summarize(count=n())
```

```
## `summarise()` has grouped output by 'VITUSE'. You can override using the `.groups` argumen
t.
```

```
tpFrame <- tpFrame %>%
  mutate(text = paste0("VITUSE: ", VITUSE, "\n", "SMOKSTAT: ", SMOKSTAT, "\n", "count: ", cou
nt))

p <- ggplot(tpFrame, aes(VITUSE, SMOKSTAT, fill= count, text=text)) +
  geom_tile() +
  scale_fill_distiller(palette = "viridis", direction = 1) +
  theme_ipsum() +
  ggtitle("Heatmap for VITUSE and SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning in pal_name(palette, type): Unknown palette viridis
```

```
ggplotly(p, tooltip="text")
```



1.2.3-2 Heatmap of correlations

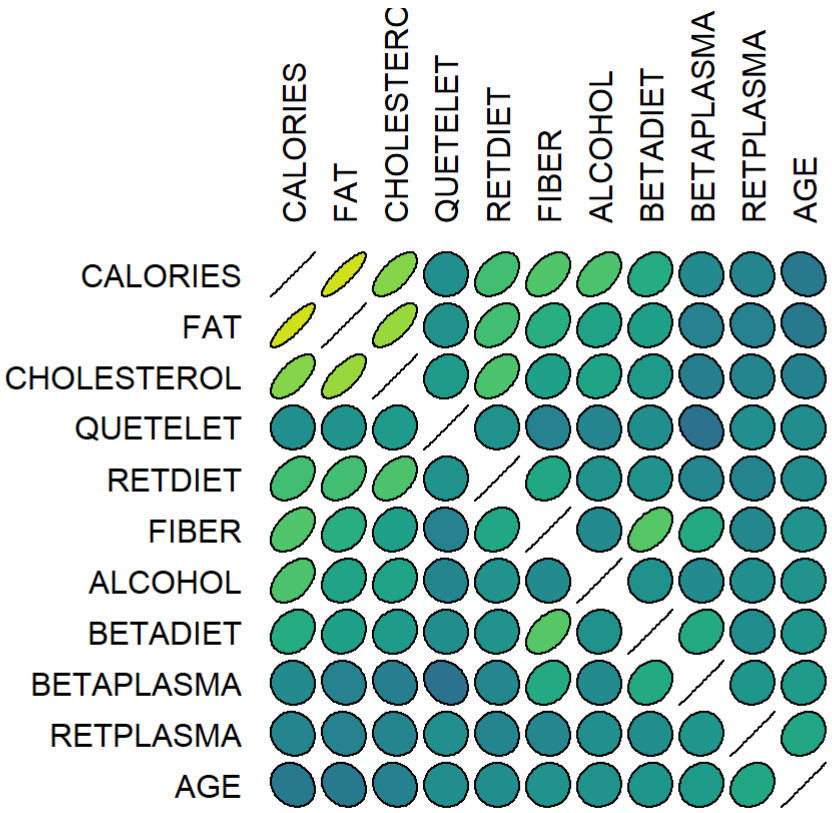
```
# heatmap(plasma[, c(1,4,6,7:14)], Colv = NA, Rowv = NA, scale="column", col = coul, xlab="va
riable", ylab="car", main="heatmap")
library(ellipse)
```

```
##
## 载入程辑包：'ellipse'
```

```
## The following object is masked from 'package:car':
##
##     ellipse
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

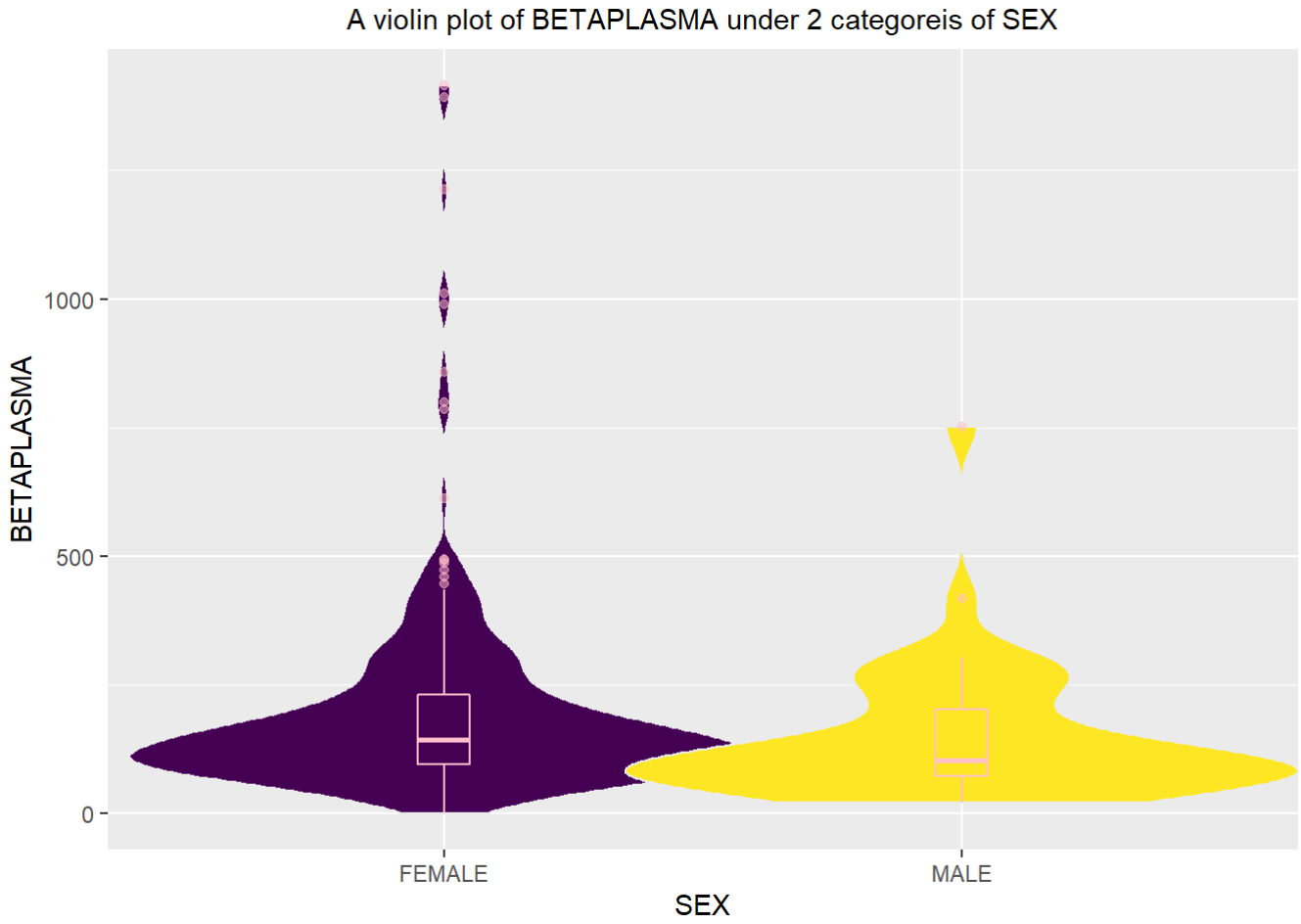
```
# Use of the mtcars data proposed by R
corr <- cor(plasma[, c(1,4,6,7:14)])

# Build a Pannel of 100 colors with Rcolor Brewer
# my_colors <- brewer.pal(5, "Spectral")
# my_colors <- colorRampPalette(my_colors)(100)
viridis_palatte <- viridis(100)

# Order the correlation matrix
order <- order(corr[1, ])
corr_order <- corr[order, order]
plotcorr(corr_order, col=viridis_palatte[corr_order*50+50] , mar=c(1,1,1,1))
```



1.2.4 SEX and BETAPLASMA

```
par(mfrow = c(1,2))
plasma %>%
  ggplot(aes(x=SEX, y=BETAPLASMA, fill=SEX)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of BETAPLASMA under 2 categoreis of SEX") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

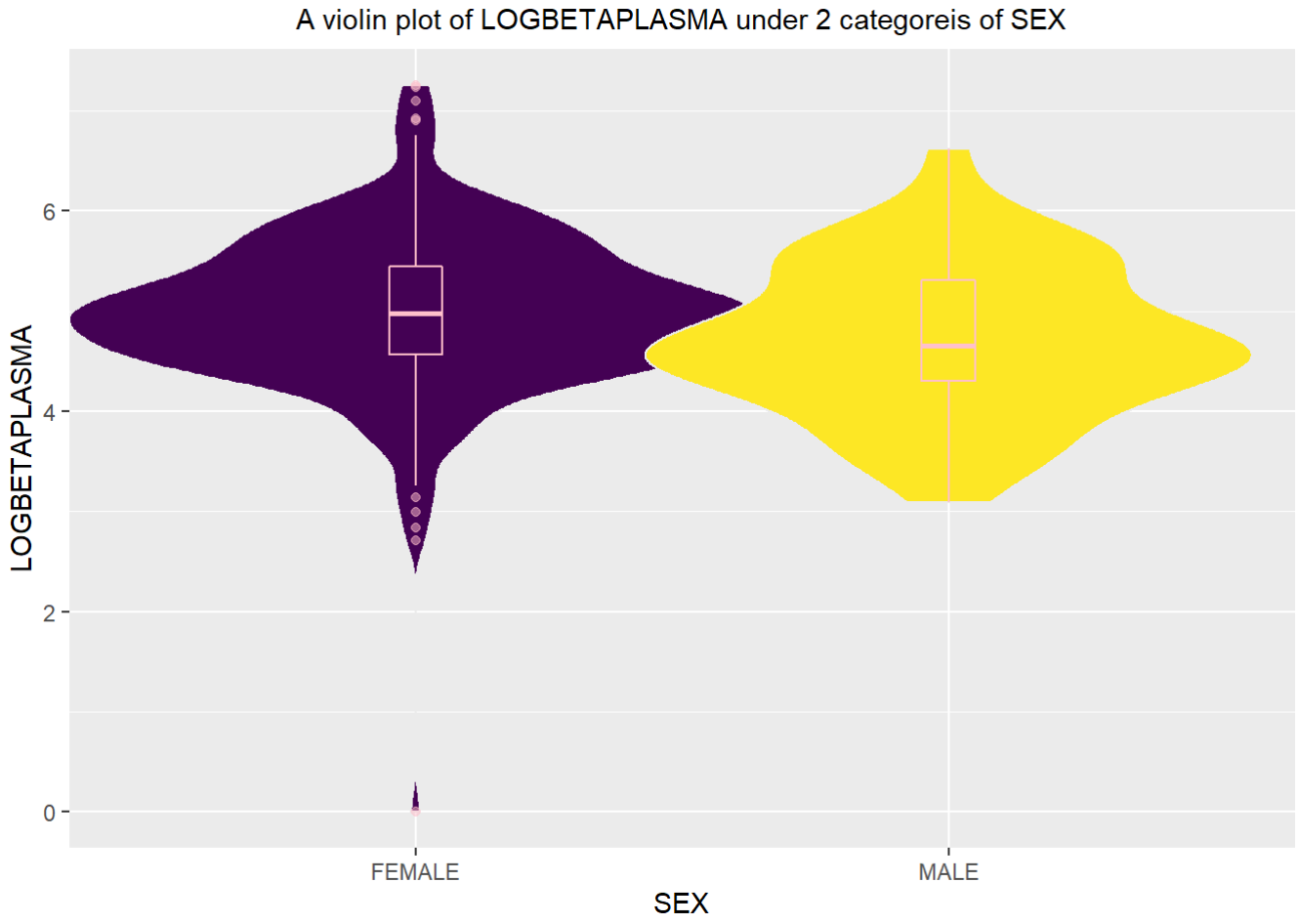



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=SEX, y=LOGBETAPLASMA, fill=SEX)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGBETAPLASMA under 2 categoreis of SEX") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```




```
  xlab("type")
```

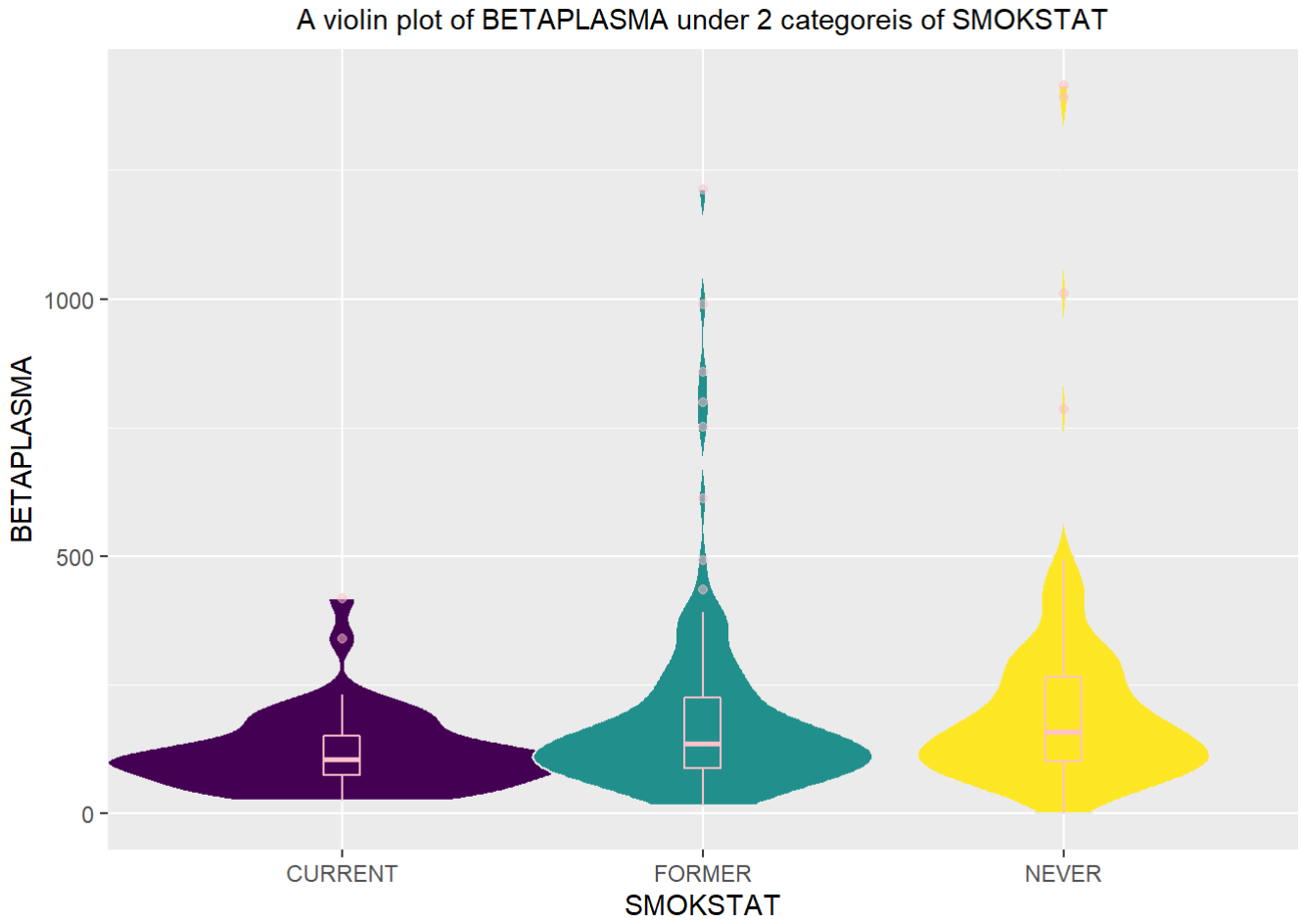
```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

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

1.2.5 SMOKSTAT and BETAPLASMA

```
plasma %>%
  ggplot(aes(x=SMOKSTAT, y=BETAPLASMA, fill=SMOKSTAT)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of BETAPLASMA under 2 categoreis of SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

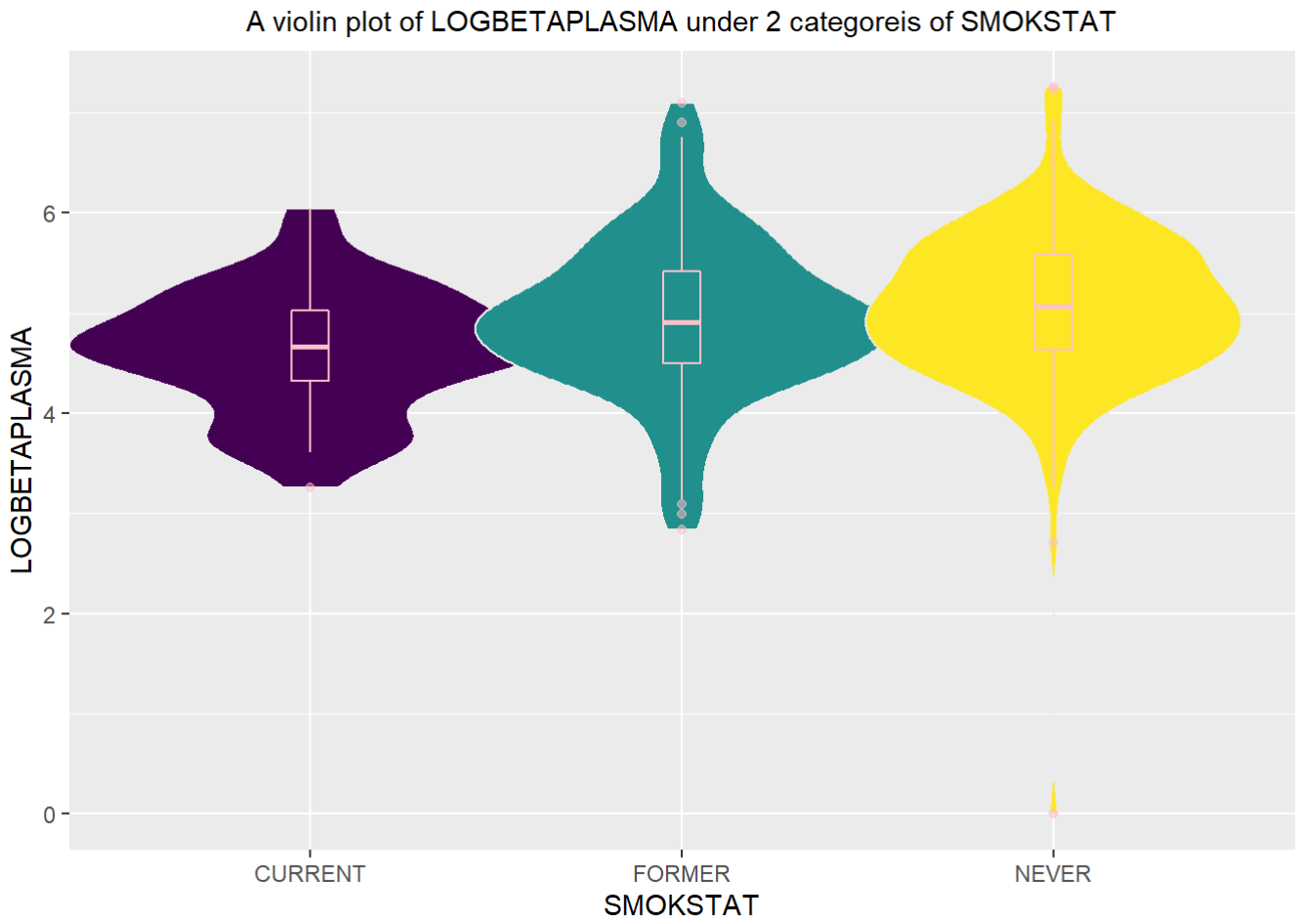



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=SMOKSTAT, y=LOGBETAPLASMA, fill=SMOKSTAT)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGBETAPLASMA under 2 categoreis of SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```



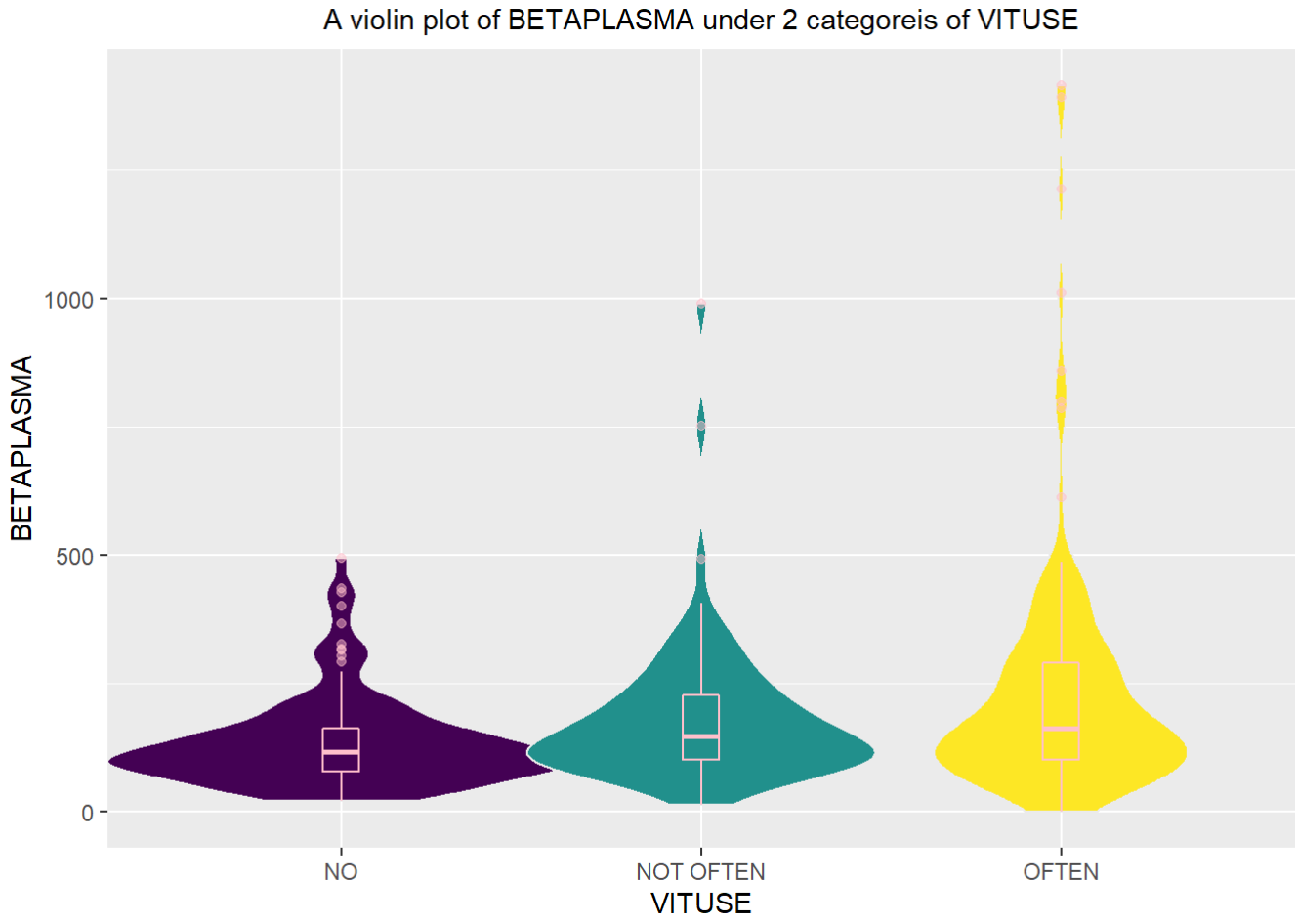

```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

1.2.6 VITUSE and BETAPLASMA

```
plasma %>%
  ggplot(aes(x=VITUSE, y=BETAPLASMA, fill=VITUSE)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of BETAPLASMA under 2 categoreis of VITUSE") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

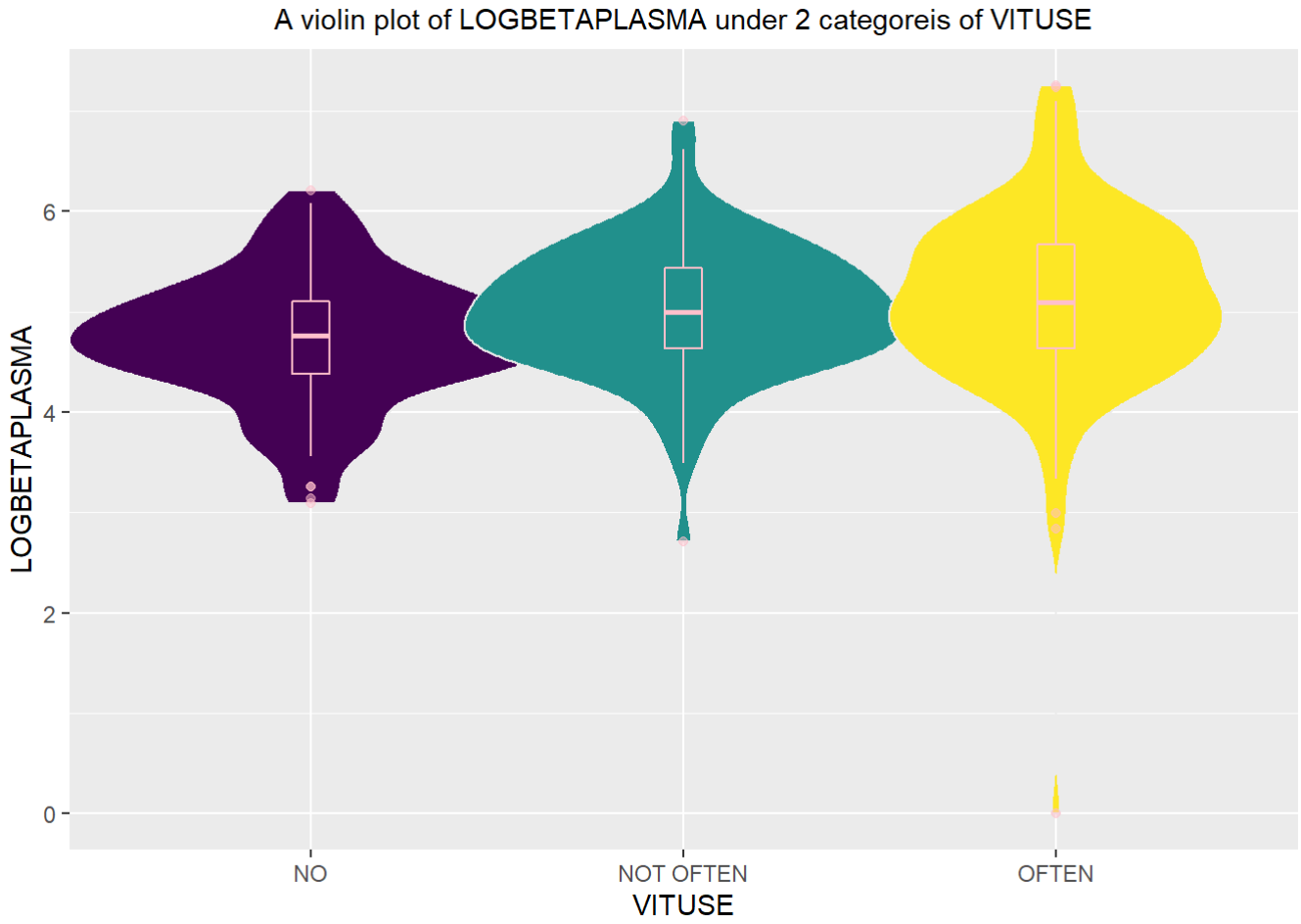



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=VITUSE, y=LOGBETAPLASMA, fill=VITUSE)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGBETAPLASMA under 2 categoreis of VITUSE") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```



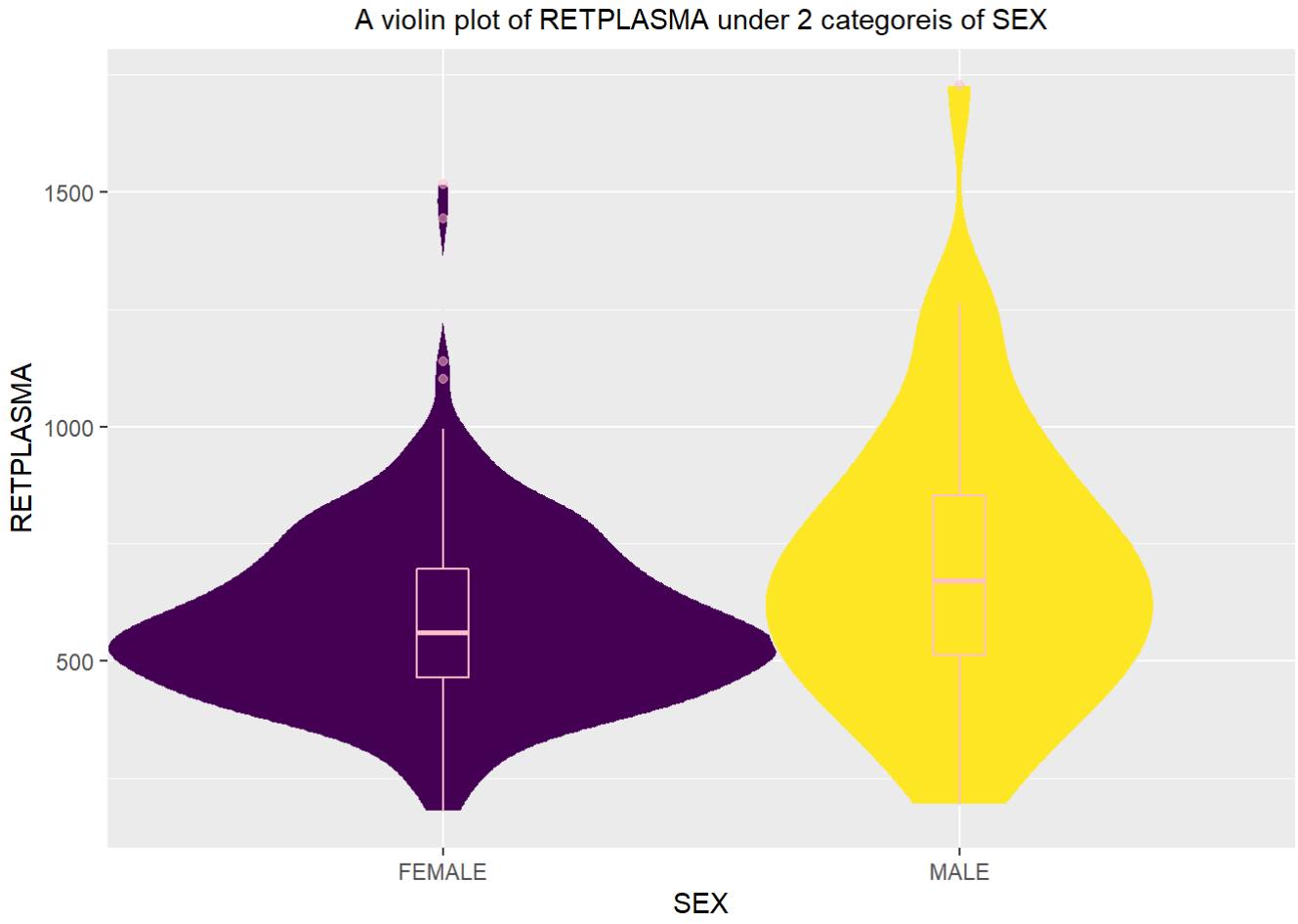

```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

1.2.7 Boxplots about RETPLASMA

```
plasma %>%
  ggplot(aes(x=SEX, y=RETPLASMA, fill=SEX)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of RETPLASMA under 2 categoreis of SEX") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

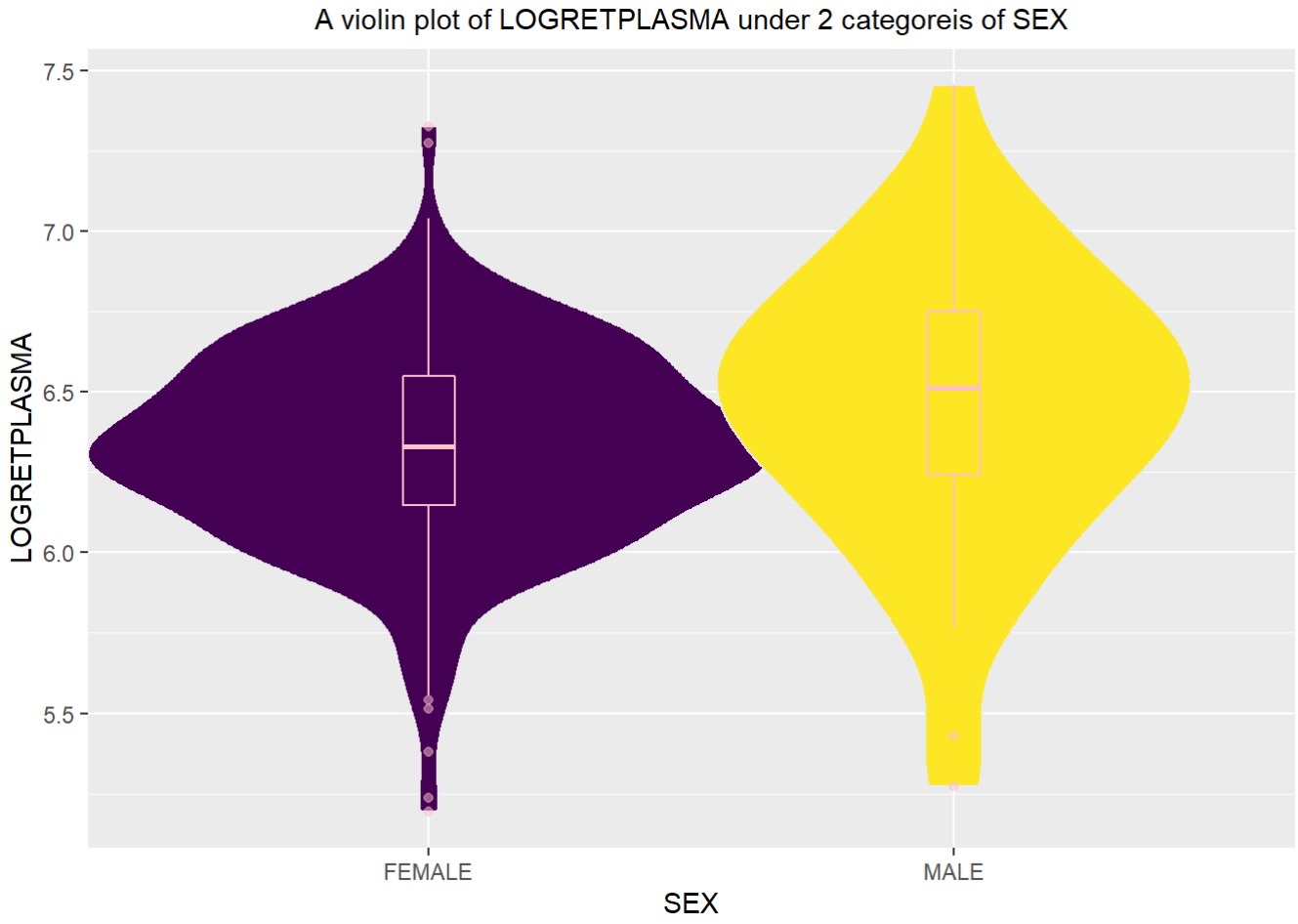



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=SEX, y=LOGRETPLASMA, fill=SEX)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGRETPLASMA under 2 categoreis of SEX") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

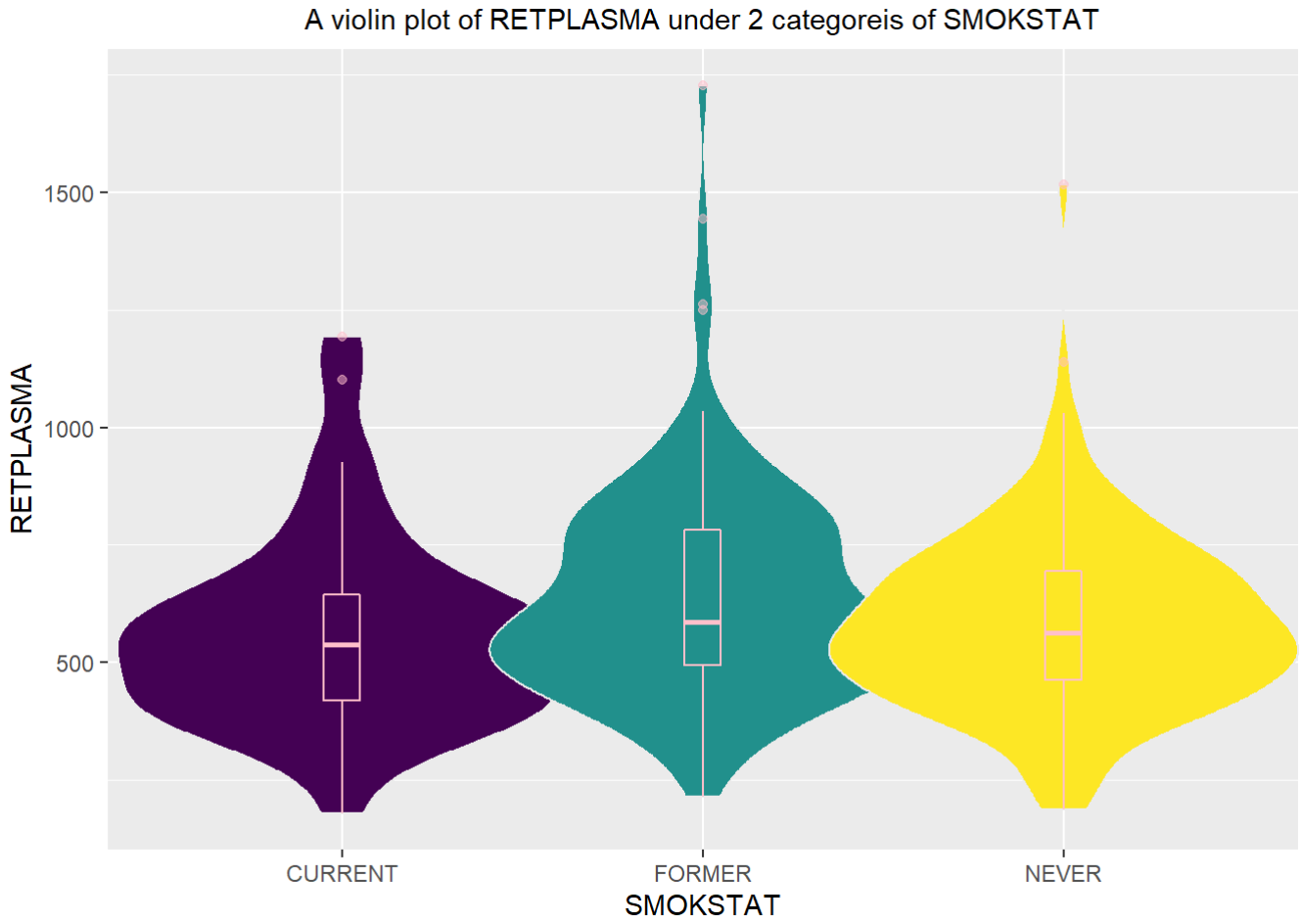



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=SMOKSTAT, y=RETPLASMA, fill=SMOKSTAT)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of RETPLASMA under 2 categoreis of SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

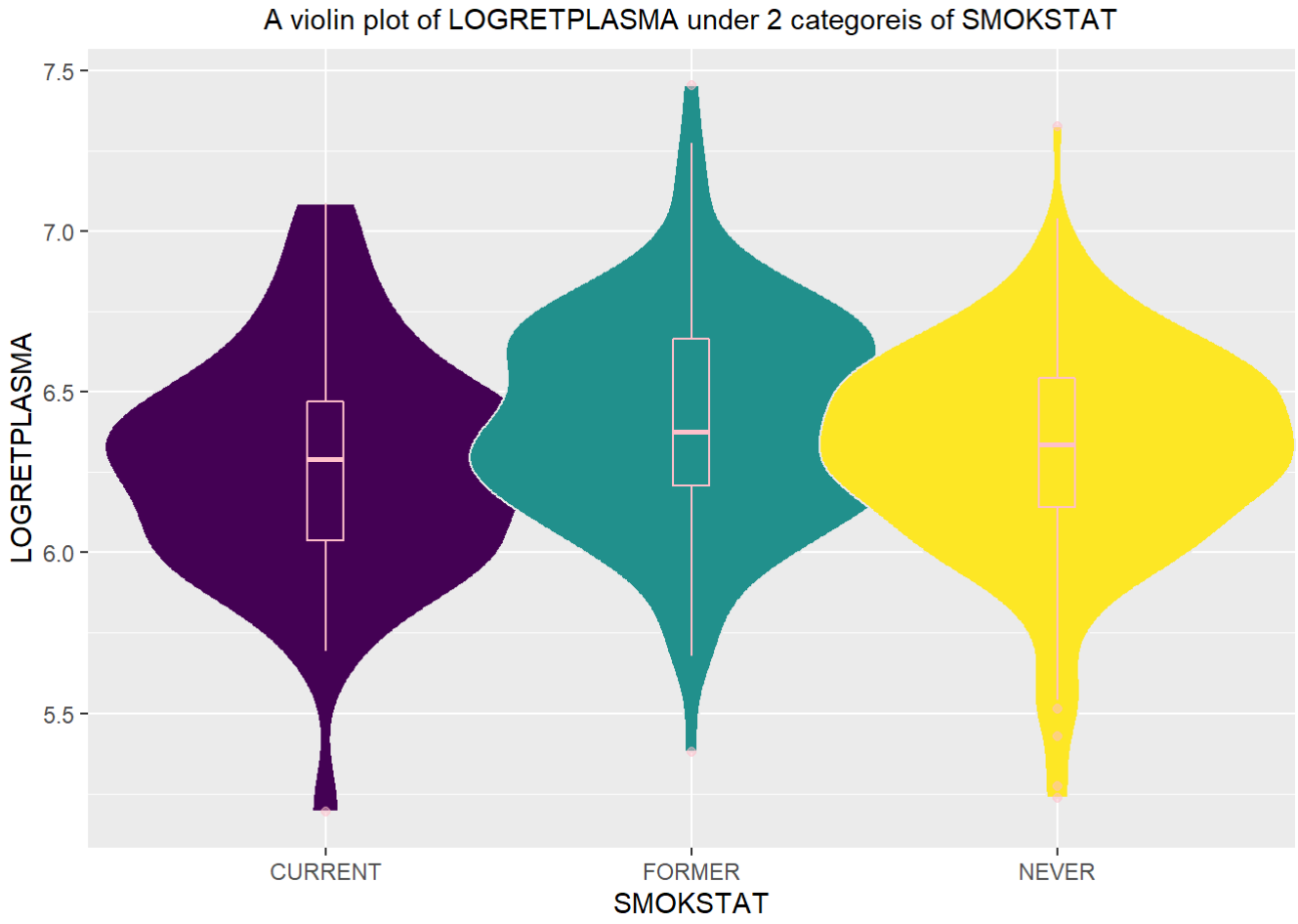



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=SMOKSTAT, y=LOGRETPLASMA, fill=SMOKSTAT)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGRETPLASMA under 2 categoreis of SMOKSTAT") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

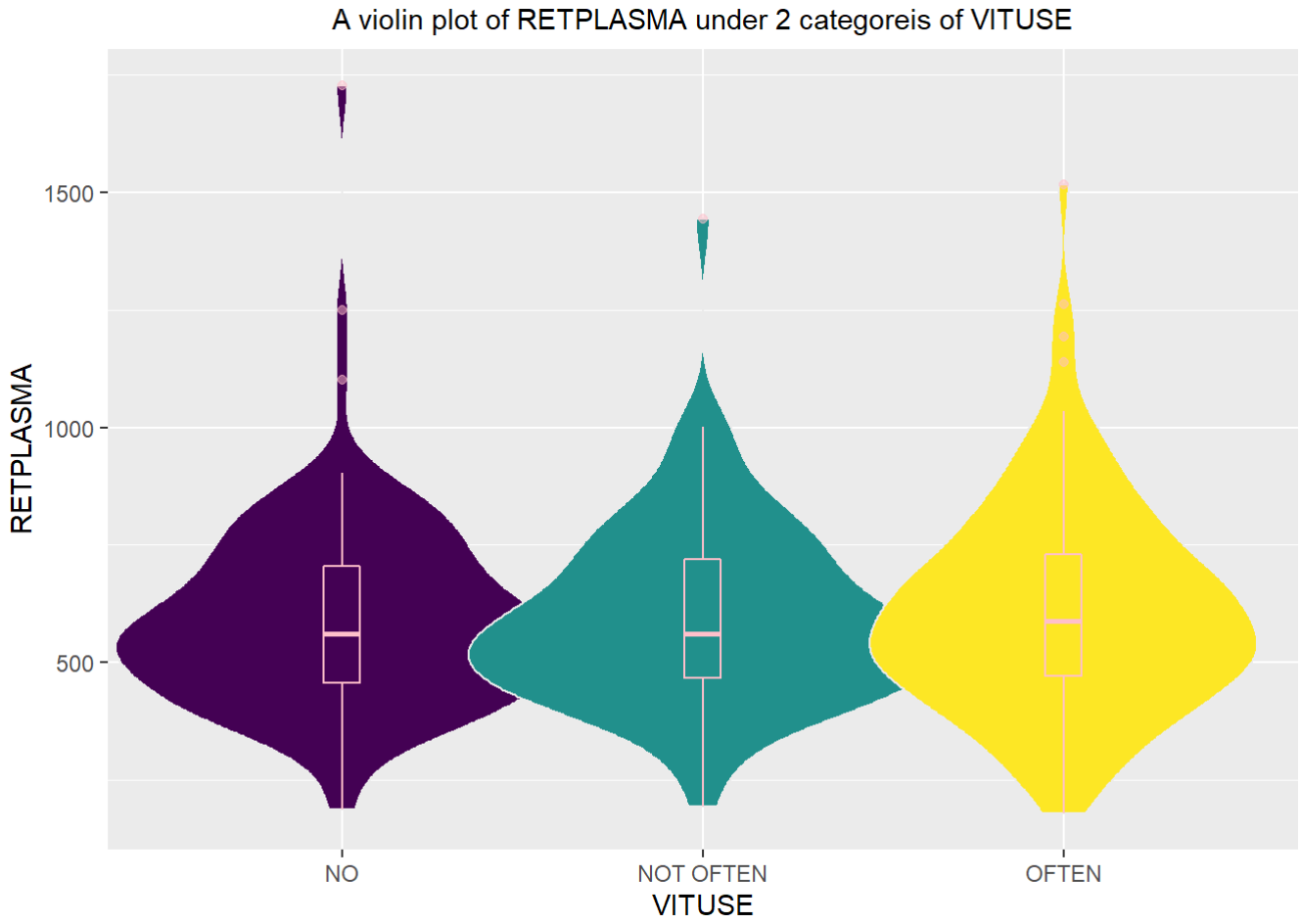



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=VITUSE, y=RETPLASMA, fill=VITUSE)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of RETPLASMA under 2 categoreis of VITUSE") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

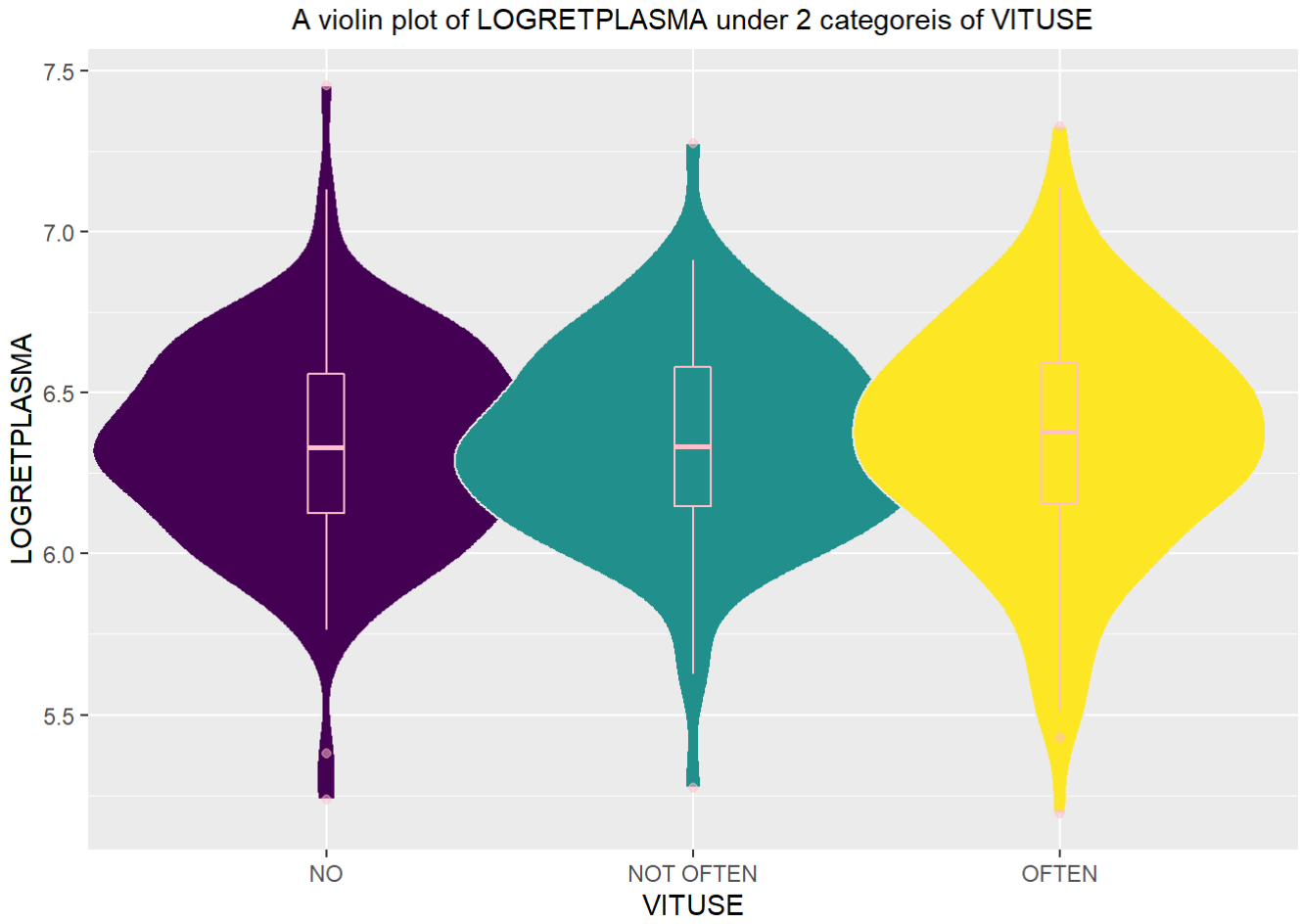



```
  xlab("type")
```

```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

```
plasma %>%
  ggplot(aes(x=VITUSE, y=LOGRETPLASMA, fill=VITUSE)) +
  geom_violin(width=1.3, color = "#EBEBEB") +
  geom_boxplot(width=0.1, color="pink", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A violin plot of LOGRETPLASMA under 2 categoreis of VITUSE") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: position_dodge requires non-overlapping x intervals
```




```
  xlab("type")
```

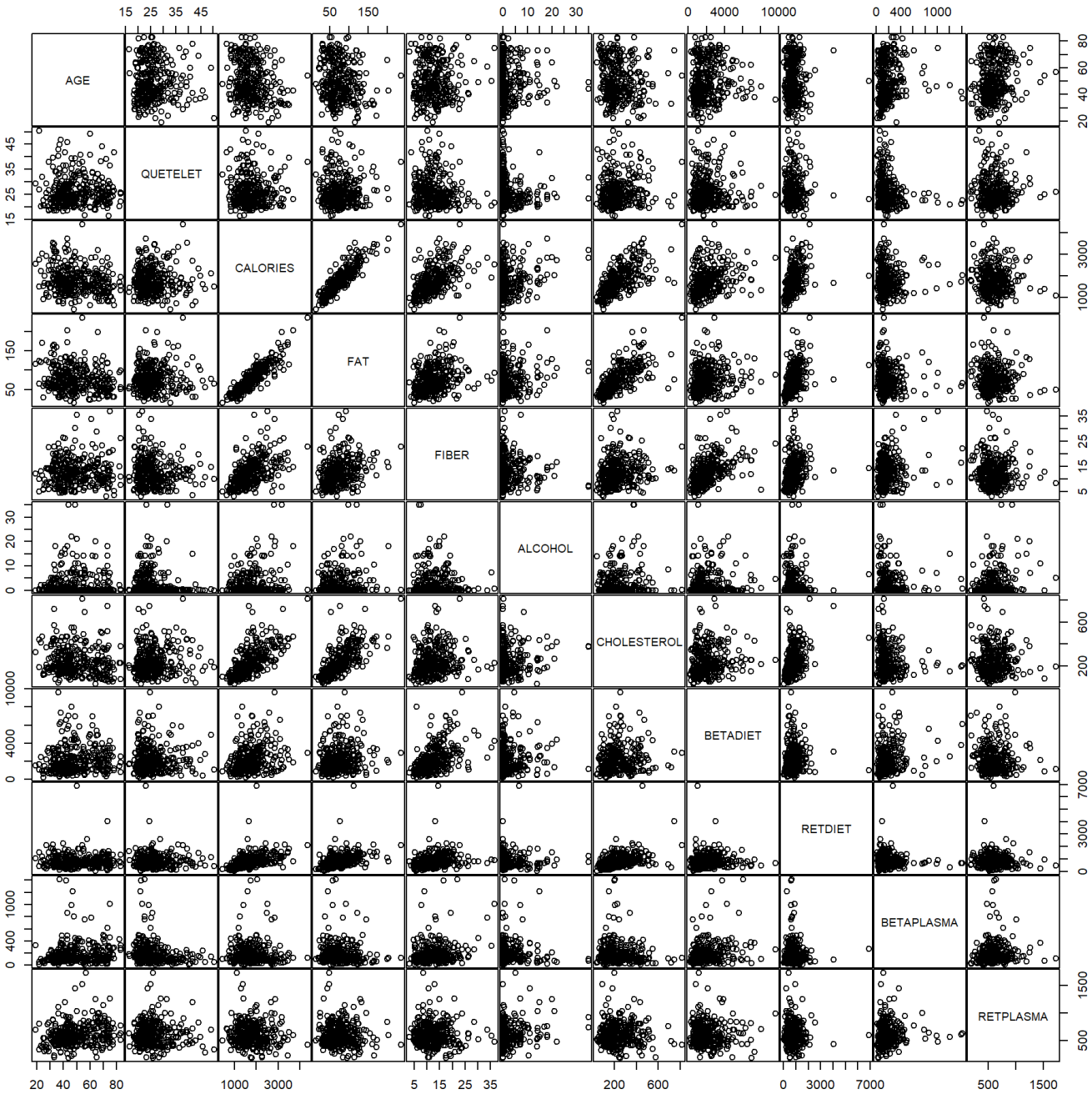
```
## $x
## [1] "type"
##
## attr(,"class")
## [1] "labels"
```

To delete unusual observations, and paired scatter plots for continuous variables and LOGBETAPLASMA and

LOGRETPLASMA

```
# drop the observations with LOGBETA 0 and too large ALCOHOL
if (length(which(plasma$LOGBETA == 0)) > 0 | length(which(plasma$ALCOHOL == 203)) > 0) {
  plasma <- plasma[-which(plasma$LOGBETA == 0), ]
  plasma <- plasma[-which(plasma$ALCOHOL == 203), ]
}
```

```
pairs(plasma[, c(1,4,6:14)], gap = 0.1)
```



2

2.1 Raw models

```
fit.raw.ret <- lm(LOGRETPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET + VITUSE
                  + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL
                  + BETADIET + RETDIET, data = plasma)
summary(fit.raw.ret)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET +
##     VITUSE + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL +
##     BETADIET + RETDIET, data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05729 -0.18089 -0.00098  0.20005  0.95152
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.041e+00  1.321e-01  45.725  < 2e-16 ***
## AGE              5.050e-03  1.432e-03   3.527 0.000486 ***
## SEXMALE          5.696e-02  6.142e-02   0.927 0.354447
## SMOKSTATFORMER   8.954e-02  5.996e-02   1.493 0.136397
## SMOKSTATNEVER    1.633e-02  5.913e-02   0.276 0.782664
## QUETELET         1.524e-03  3.137e-03   0.486 0.627360
## VITUSENOT OFTEN  4.114e-02  4.833e-02   0.851 0.395303
## VITUSEOFTEN      4.138e-02  4.451e-02   0.929 0.353385
## CALORIES         1.492e-04  9.795e-05   1.524 0.128627
## FIBER           -6.751e-03  5.402e-03  -1.250 0.212385
## FAT             -2.915e-03  1.540e-03  -1.892 0.059408 .
## ALCOHOL          1.102e-02  4.251e-03   2.591 0.010042 *
## CHOLESTEROL     -5.681e-05  2.167e-04  -0.262 0.793376
## BETADIET        -1.314e-05  1.434e-05  -0.916 0.360169
## RETDIET         -6.370e-07  3.575e-05  -0.018 0.985794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.321 on 298 degrees of freedom
## Multiple R-squared:  0.1275, Adjusted R-squared:  0.08655
## F-statistic: 3.112 on 14 and 298 DF,  p-value: 0.0001504
```

```
null = lm(LOGRETPLASMA ~ 1, data = plasma)
full = lm(LOGRETPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET + VITUSE
                  + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL
                  + BETADIET + RETDIET, data = plasma)
step(full, scope = list(lower=null,upper=full),
     direction="both", criterion = "BIC", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + ALCOHOL, data = plasma)
##
## Coefficients:
## (Intercept)          AGE      ALCOHOL
##     6.05789      0.00512      0.01382
```

```
step(null, scope = list(lower=null,upper=full),
     direction="both", criterion = "BIC", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + ALCOHOL, data = plasma)
##
## Coefficients:
## (Intercept)          AGE      ALCOHOL
##     6.05789      0.00512      0.01382
```

Best model for LOGRETPLASMA without extra categorizing

```
fit.raw.ret.final <- lm(formula = LOGRETPLASMA ~ AGE, data = plasma)
summary(fit.raw.ret.final)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE, data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19063 -0.19352 -0.00572  0.21013  1.06846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.095509   0.066465  91.711  < 2e-16 ***
## AGE         0.005101   0.001273   4.006 7.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.328 on 311 degrees of freedom
## Multiple R-squared:  0.04908,    Adjusted R-squared:  0.04602
## F-statistic: 16.05 on 1 and 311 DF,  p-value: 7.719e-05
```

Raw model for LOGBETAPLASMA

```
fit.raw.beta <- lm(LOGBETAPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET + VITUSE
                   + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL
                   + BETADIET + RETDIET, data = plasma)
summary(fit.raw.beta)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET +
##     VITUSE + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL +
##     BETADIET + RETDIET, data = plasma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0246 -0.3746 -0.0038  0.3954  1.8826
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.991e+00  2.704e-01  18.460  < 2e-16 ***
## AGE              5.051e-03  2.930e-03   1.724  0.08573 .
## SEXMALE         -2.278e-01  1.257e-01  -1.812  0.07100 .
## SMOKSTATFORMER   2.016e-01  1.227e-01   1.643  0.10151
## SMOKSTATNEVER    2.854e-01  1.210e-01   2.358  0.01901 *
## QUETELET        -3.120e-02  6.420e-03  -4.860  1.9e-06 ***
## VITUSENOT OFTEN  2.726e-01  9.892e-02   2.756  0.00621 **
## VITUSEOFTEN      2.979e-01  9.110e-02   3.270  0.00120 **
## CALORIES        -2.263e-04  2.005e-04  -1.129  0.25973
## FIBER            3.024e-02  1.106e-02   2.735  0.00661 **
## FAT              1.376e-03  3.152e-03   0.437  0.66275
## ALCOHOL          5.547e-03  8.701e-03   0.638  0.52428
## CHOLESTEROL     -3.236e-04  4.435e-04  -0.730  0.46624
## BETADIET         4.820e-05  2.935e-05   1.642  0.10164
## RETDIET          4.407e-05  7.316e-05   0.602  0.54739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6569 on 298 degrees of freedom
## Multiple R-squared:  0.2501, Adjusted R-squared:  0.2149
## F-statistic: 7.099 on 14 and 298 DF,  p-value: 1.152e-12
```

```
null = lm(LOGBETAPLASMA ~ 1, data = plasma)
full = lm(LOGBETAPLASMA ~ AGE + SEX + SMOKSTAT + QUETELET + VITUSE
                  + CALORIES + FIBER + FAT + ALCOHOL + CHOLESTEROL
                  + BETADIET + RETDIET, data = plasma)
step(null, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE,
##     data = plasma)
##
## Coefficients:
##     (Intercept)         QUETELET            FIBER         CALORIES
##       5.4368759       -0.0292541        0.0433816       -0.0002637
## VITUSENOT OFTEN      VITUSEOFTEN
##       0.2871437        0.3538229
```

```
step(full, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ QUETELET + VITUSE + CALORIES + FIBER,
##     data = plasma)
##
## Coefficients:
##     (Intercept)         QUETELET  VITUSENOT OFTEN      VITUSEOFTEN
##       5.4368759       -0.0292541        0.2871437        0.3538229
##        CALORIES            FIBER
##      -0.0002637        0.0433816
```

Optimal model without categorizing, for LOGBETAPLASMA

```
fit.raw.beta.final <- lm(formula = LOGBETAPLASMA ~ QUETELET + FIBER + CHOLESTEROL +
    BETADIET, data = plasma)
summary(fit.raw.beta.final)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ QUETELET + FIBER + CHOLESTEROL +
##     BETADIET, data = plasma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0701 -0.3814 -0.0413  0.3900  1.9776
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.541e+00  2.064e-01  26.844  < 2e-16 ***
## QUETELET    -3.020e-02  6.485e-03  -4.657 4.77e-06 ***
## FIBER        2.550e-02  8.357e-03   3.051  0.00248 **
## CHOLESTEROL -1.040e-03  3.147e-04  -3.304  0.00106 **
## BETADIET     6.488e-05  2.995e-05   2.166  0.03107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6815 on 308 degrees of freedom
## Multiple R-squared:  0.1658, Adjusted R-squared:  0.155
## F-statistic:  15.3 on 4 and 308 DF,  p-value: 1.99e-11
```

Cross validation

```
library(caret)
```

```
## 载入需要的程辑包：lattice
```

```
set.seed(1234)

train_control <- trainControl(method = "CV", number = 10)

model <-
  train(
    LOGBETAPLASMA ~ QUETELET + FIBER + CHOLESTEROL + BETADIET,
    data = plasma,
    method = "lm",
    trControl = train_control
  )

print(model)
```

```
## Linear Regression
##
## 313 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 281, 283, 281, 281, 282, 282, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.6821771  0.1702002  0.5185102
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
set.seed(1234)

train_control <- trainControl(method = "CV", number = 10)

model <-
  train(
    LOGRETPLASMA ~ AGE,
    data = plasma,
    method = "lm",
    trControl = train_control
  )

print(model)
```

```
## Linear Regression
##
## 313 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 281, 283, 281, 281, 283, 282, ...
## Resampling results:
##
##   RMSE       Rsquared    MAE
##   0.3229943  0.07670949  0.250968
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

2.2 To use our categorized data

```
fit.cat.ret.full <- lm(LOGRETPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT + QUETELET + QUETELETCAT
+ VITUSE
                    + CALORIES + CALCAT + FAT + FATCAT + FIBER + FIBERCAT + ALCOHOL + ALCOHOLC
AT + CHOLESTEROL
                    + CHOLESTEROLCAT + BETADIET + BETADIETCAT + RETDIET + RETDIETCAT, data = p
lasma)
summary(fit.cat.ret.full)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT + QUETELET +
##     QUETELETCAT + VITUSE + CALORIES + CALCAT + FAT + FATCAT +
##     FIBER + FIBERCAT + ALCOHOL + ALCOHOLCAT + CHOLESTEROL + CHOLESTEROLCAT +
##     BETADIET + BETADIETCAT + RETDIET + RETDIETCAT, data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99172 -0.19869  0.00482  0.19694  0.96555
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.054e+00  3.013e-01  20.097   <2e-16 ***
## AGE                            3.347e-03  3.173e-03   1.055   0.2924
## AGECAT40-65                    5.069e-02  6.804e-02   0.745   0.4570
## AGECAT65+                      1.093e-01  1.300e-01   0.841   0.4010
## SEXMALE                        8.300e-02  6.645e-02   1.249   0.2127
## SMOKSTATFORMER                 8.725e-02  6.255e-02   1.395   0.1641
## SMOKSTATNEVER                  2.207e-02  6.135e-02   0.360   0.7193
## QUETELET                       6.428e-03  7.204e-03   0.892   0.3730
## QUETELETCATHealthy Weight      2.965e-04  1.756e-01   0.002   0.9987
## QUETELETCATOverweight         -1.258e-02  1.875e-01  -0.067   0.9466
## QUETELETCATObesity            -8.217e-02  2.229e-01  -0.369   0.7126
## VITUSENOT OFTEN                5.291e-02  4.973e-02   1.064   0.2883
## VITUSEOFTEN                    4.833e-02  4.561e-02   1.060   0.2902
## CALORIES                       7.813e-05  1.240e-04   0.630   0.5291
## CALCATLOW                     -5.450e-02  9.519e-02  -0.573   0.5674
## CALCATMED                     -5.624e-02  6.810e-02  -0.826   0.4095
## FAT                           -2.314e-03  1.957e-03  -1.182   0.2381
## FATCATToo much                -3.138e-02  5.918e-02  -0.530   0.5963
## FIBER                         -6.285e-03  6.722e-03  -0.935   0.3506
## FIBERCATNot too low           -8.289e-03  6.167e-02  -0.134   0.8932
## ALCOHOL                        1.099e-02  5.931e-03   1.853   0.0649 .
## ALCOHOLCATNot frequent         1.075e-01  4.811e-02   2.233   0.0263 *
## ALCOHOLCATFrequent             4.708e-02  6.240e-02   0.755   0.4512
## CHOLESTEROL                    2.275e-05  2.753e-04   0.083   0.9342
## CHOLESTEROLCATRelatively high -8.500e-03  6.152e-02  -0.138   0.8902
## BETADIET                       9.595e-07  2.027e-05   0.047   0.9623
## BETADIETCATNot too low        -6.014e-02  5.858e-02  -1.027   0.3055
## RETDIET                       -1.476e-05  4.473e-05  -0.330   0.7417
## RETDIETCATToo low             -2.593e-02  5.384e-02  -0.482   0.6304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3238 on 284 degrees of freedom
## Multiple R-squared:  0.1537, Adjusted R-squared:  0.07021
## F-statistic: 1.841 on 28 and 284 DF,  p-value: 0.007353
```

```
null = lm(LOGRETPLASMA ~ 1, data = plasma)
full = lm(LOGRETPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT + QUETELET + QUETELETCAT + VITUSE
                   + CALORIES + CALCAT + FAT + FATCAT + FIBER + FIBERCAT + ALCOHOL + ALCOHOLC
AT + CHOLESTEROL
                   + CHOLESTEROLCAT + BETADIET + BETADIETCAT + RETDIET + RETDIETCAT, data = p
lasma)
step(null, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + ALCOHOL, data = plasma)
##
## Coefficients:
## (Intercept)          AGE      ALCOHOL
##     6.05789      0.00512      0.01382
```

```
step(full, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGE + ALCOHOL, data = plasma)
##
## Coefficients:
## (Intercept)          AGE      ALCOHOL
##     6.05789      0.00512      0.01382
```

```
# The result of BIC step wise selection from full or null model both suggest only AGE as a pr
edictor.
```

```
fit.cat.beta.full <- lm(LOGBETAPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT + QUETELET + QUETELETCA
T + VITUSE
                   + CALORIES + CALCAT + FAT + FATCAT + FIBER + FIBERCAT + ALCOHOL + ALCOHOLC
AT + CHOLESTEROL
                   + CHOLESTEROLCAT + BETADIET + BETADIETCAT + RETDIET + RETDIETCAT, data = p
lasma)
summary(fit.cat.beta.full)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT +
##     QUETELET + QUETELETCAT + VITUSE + CALORIES + CALCAT + FAT +
##     FATCAT + FIBER + FIBERCAT + ALCOHOL + ALCOHOLCAT + CHOLESTEROL +
##     CHOLESTEROLCAT + BETADIET + BETADIETCAT + RETDIET + RETDIETCAT,
##     data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83351 -0.38842 -0.01901  0.35934  1.90126
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.527e+00  6.115e-01   9.039  < 2e-16 ***
## AGE                           2.141e-03  6.440e-03   0.332 0.739841
## AGECAT40-65                   1.988e-01  1.381e-01   1.440 0.151068
## AGECAT65+                     2.169e-01  2.639e-01   0.822 0.411669
## SEXMALE                      -1.145e-01  1.349e-01  -0.849 0.396775
## SMOKSTATFORMER                1.652e-01  1.270e-01   1.301 0.194285
## SMOKSTATNEVER                 2.697e-01  1.245e-01   2.166 0.031174 *
## QUETELET                     -2.712e-02  1.462e-02  -1.854 0.064715 .
## QUETELETCATHealthy Weight    -2.859e-01  3.564e-01  -0.802 0.423098
## QUETELETCATOverweight        -2.679e-01  3.806e-01  -0.704 0.482024
## QUETELETCATObesity           -2.954e-01  4.524e-01  -0.653 0.514229
## VITUSENOT OFTEN               2.939e-01  1.009e-01   2.912 0.003877 **
## VITUSEOFTEN                   3.208e-01  9.259e-02   3.465 0.000612 ***
## CALORIES                     -4.274e-04  2.517e-04  -1.698 0.090557 .
## CALCATLOW                    -2.334e-01  1.932e-01  -1.208 0.227984
## CALCATMED                    -1.827e-01  1.382e-01  -1.322 0.187356
## FAT                           2.992e-03  3.973e-03   0.753 0.451941
## FATCATToo much               -1.433e-01  1.201e-01  -1.193 0.233826
## FIBER                         3.076e-02  1.365e-02   2.255 0.024923 *
## FIBERCATNot too low          -1.399e-01  1.252e-01  -1.118 0.264592
## ALCOHOL                      -9.979e-03  1.204e-02  -0.829 0.407903
## ALCOHOLCATNot frequent        1.762e-01  9.766e-02   1.804 0.072326 .
## ALCOHOLCATFrequent            2.638e-01  1.267e-01   2.083 0.038132 *
## CHOLESTEROL                   3.793e-05  5.588e-04   0.068 0.945929
## CHOLESTEROLCATRelatively high -1.340e-01  1.249e-01  -1.073 0.284297
## BETADIET                      7.726e-05  4.114e-05   1.878 0.061426 .
## BETADIETCATNot too low       -8.452e-02  1.189e-01  -0.711 0.477783
## RETDIET                       2.333e-05  9.079e-05   0.257 0.797387
## RETDIETCATToo low            -6.280e-03  1.093e-01  -0.057 0.954215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6573 on 284 degrees of freedom
## Multiple R-squared:  0.2844, Adjusted R-squared:  0.2139
## F-statistic: 4.032 on 28 and 284 DF,  p-value: 6.724e-10
```

```
null = lm(LOGBETAPLASMA ~ 1, data = plasma)
full = lm(LOGBETAPLASMA ~ AGE + AGECAT + SEX + SMOKSTAT + QUETELET + QUETELETCAT + VITUSE
                  + CALORIES + CALCAT + FAT + FATCAT + FIBER + FIBERCAT + ALCOHOL + ALCOHOLC
AT + CHOLESTEROL
                  + CHOLESTEROLCAT + BETADIET + BETADIETCAT + RETDIET + RETDIETCAT, data = p
lasma)
step(null, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE,
##     data = plasma)
##
## Coefficients:
##     (Intercept)         QUETELET            FIBER         CALORIES
##       5.4368759       -0.0292541        0.0433816       -0.0002637
## VITUSENOT OFTEN      VITUSEOFTEN
##       0.2871437        0.3538229
```

```
step(full, scope = list(lower=null,upper=full),
     direction="both", k = log(313), trace = 0)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ QUETELET + VITUSE + CALORIES + FIBER,
##     data = plasma)
##
## Coefficients:
##     (Intercept)         QUETELET  VITUSENOT OFTEN      VITUSEOFTEN
##       5.4368759       -0.0292541        0.2871437        0.3538229
##        CALORIES            FIBER
##      -0.0002637        0.0433816
```

```
# Still the result when candidate X variables don't include the categorized ones
```

Based on the interpretation and knowledge of the real world

```
fit.cat.beta.manual <- lm(LOGBETAPLASMA ~ AGECAT + SMOKSTAT + QUETELET + VITUSE
                  + FIBER + ALCOHOLCAT + CHOLESTEROL + BETADIET,
                  data = plasma)

summary(fit.cat.beta.manual)
```

```
##
## Call:
## lm(formula = LOGBETAPLASMA ~ AGECAT + SMOKSTAT + QUETELET + VITUSE +
##     FIBER + ALCOHOLCAT + CHOLESTEROL + BETADIET, data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84437 -0.36581 -0.01082  0.39759  1.77760
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.912e+00  2.342e-01  20.977  < 2e-16 ***
## AGECAT40-65             2.081e-01  8.967e-02   2.321  0.02097 *
## AGECAT65+               2.497e-01  1.117e-01   2.235  0.02614 *
## SMOKSTATFORMER          1.692e-01  1.236e-01   1.368  0.17229
## SMOKSTATNEVER           2.958e-01  1.198e-01   2.468  0.01415 *
## QUETELET               -2.993e-02  6.548e-03  -4.570 7.13e-06 ***
## VITUSENOT OFTEN         3.064e-01  9.732e-02   3.149  0.00181 **
## VITUSEOFTEN             3.162e-01  8.927e-02   3.542  0.00046 ***
## FIBER                   1.995e-02  8.126e-03   2.455  0.01467 *
## ALCOHOLCATNot frequent  1.586e-01  9.501e-02   1.669  0.09616 .
## ALCOHOLCATFrequent      1.682e-01  9.420e-02   1.785  0.07524 .
## CHOLESTEROL            -8.291e-04  3.076e-04  -2.695  0.00743 **
## BETADIET                4.960e-05  2.913e-05   1.703  0.08963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6544 on 300 degrees of freedom
## Multiple R-squared:  0.2509, Adjusted R-squared:  0.2209
## F-statistic: 8.373 on 12 and 300 DF,  p-value: 1.167e-13
```

```
set.seed(1234)

train_control <- trainControl(method = "CV", number = 10)

model <-
  train(
    LOGBETAPLASMA ~ AGECAT + SMOKSTAT + QUETELET + VITUSE +
    FIBER + ALCOHOLCAT + CHOLESTEROL + BETADIET,
    data = plasma,
    method = "lm",
    trControl = train_control
  )

print(model)
```

```
## Linear Regression
##
## 313 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 281, 283, 281, 281, 282, 282, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.6642418  0.2303499  0.5133462
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
fit.cat.ret.manual <- lm(LOGRETPLASMA ~ AGECAT + SEX + SMOKSTAT + FATCAT + FIBER + ALCOHOLCAT
+  BETADIETCAT, data = plasma)
summary(fit.cat.ret.manual)
```

```
##
## Call:
## lm(formula = LOGRETPLASMA ~ AGECAT + SEX + SMOKSTAT + FATCAT +
##     FIBER + ALCOHOLCAT + BETADIETCAT, data = plasma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9818 -0.2111 -0.0068  0.2037  0.9810
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.284658   0.081914  76.723  < 2e-16 ***
## AGECAT40-65              0.104649   0.043234   2.421  0.01609 *
## AGECAT65+                0.213358   0.056045   3.807  0.00017 ***
## SEXMALE                  0.096033   0.057231   1.678  0.09439 .
## SMOKSTATFORMER           0.095686   0.059673   1.604  0.10987
## SMOKSTATNEVER            0.023413   0.057484   0.407  0.68408
## FATCATToo much          -0.094038   0.040819  -2.304  0.02191 *
## FIBER                   -0.005315   0.003953  -1.344  0.17983
## ALCOHOLCATNot frequent   0.098534   0.046602   2.114  0.03530 *
## ALCOHOLCATFrequent       0.098631   0.044696   2.207  0.02809 *
## BETADIETCATNot too low -0.046631   0.041172  -1.133  0.25828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3205 on 302 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.08933
## F-statistic:  4.06 on 10 and 302 DF,  p-value: 2.989e-05
```

```
set.seed(1234)

train_control <- trainControl(method = "CV", number = 10)

model <-
  train(
    LOGRETPLASMA ~ AGECAT + SEX + SMOKSTAT + FATCAT +
    FIBER + ALCOHOLCAT + BETADIETCAT,
    data = plasma,
    method = "lm",
    trControl = train_control
  )

print(model)
```

```
## Linear Regression
##
## 313 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 281, 283, 281, 281, 283, 282, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.3224767  0.1068526  0.2557534
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

STA206_Elastic_Net_Model

Mingqian Zhang

2023-12-12

```
setwd("D:\\RProjects\\UCD_STA206_Pj")
library(glmnet)
```

```
##      Matrix
```

```
## Loaded glmnet 4.1-8
```

```
library(MASS)
library(ggplot2)
library(caret)
```

```
##      lattice
```

```
library(readxl)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(Metrics)
```

```
##
##     'Metrics'
```

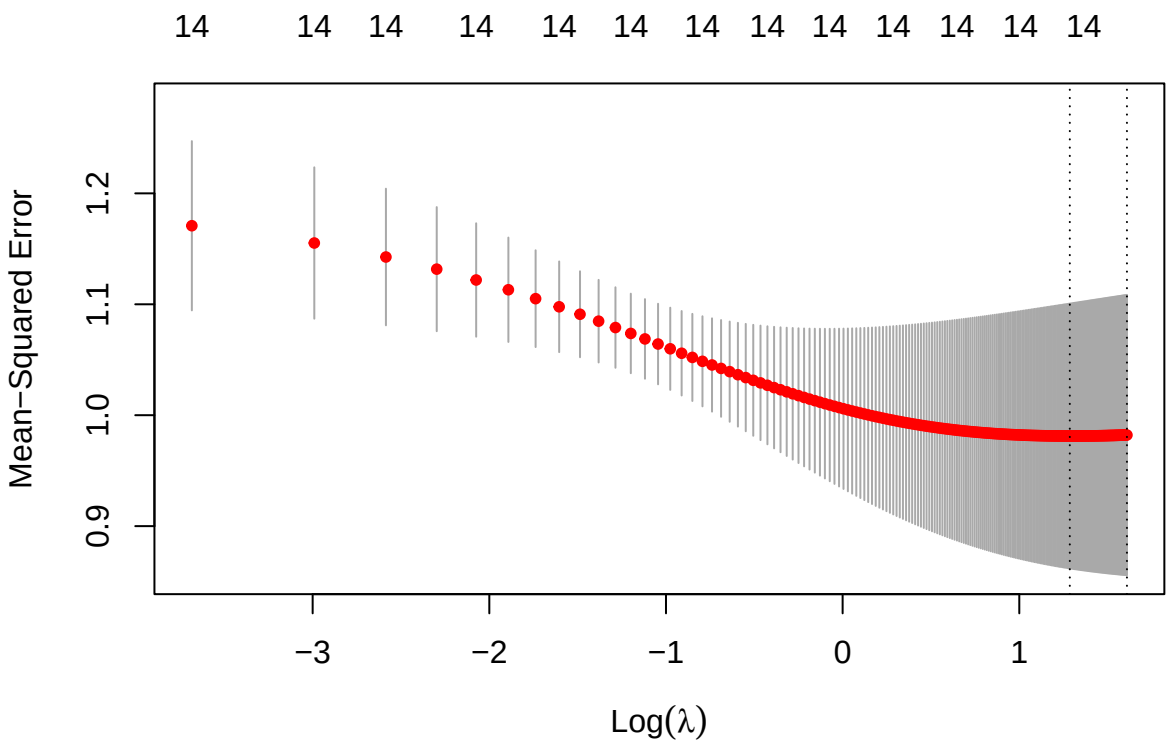
```
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
library(ggplot2)
data<-read.table('Plasma.txt',header=TRUE)


n=315
p=12
#model_o<-lm(RETPLASMA~.-BETAPLASMA ,data=data)
#summary(model_o)

dummy_variables <- model.matrix(~ SMOKSTAT - 1, data = data)  #   -1

# Converting Smoke into Dummy
data <- cbind( dummy_variables,data)
colnames(data)[colnames(data) %in% c("FORMER", "NEVER")] <- c("smoke1", "smoke2")
data <- data[, !names(data) %in% "SMOKSTAT"]
data <- data[, !names(data) %in% "SMOKSTATCURRENT"]

data$SEX <- ifelse(data$SEX == "FEMALE", 0, 1)
```

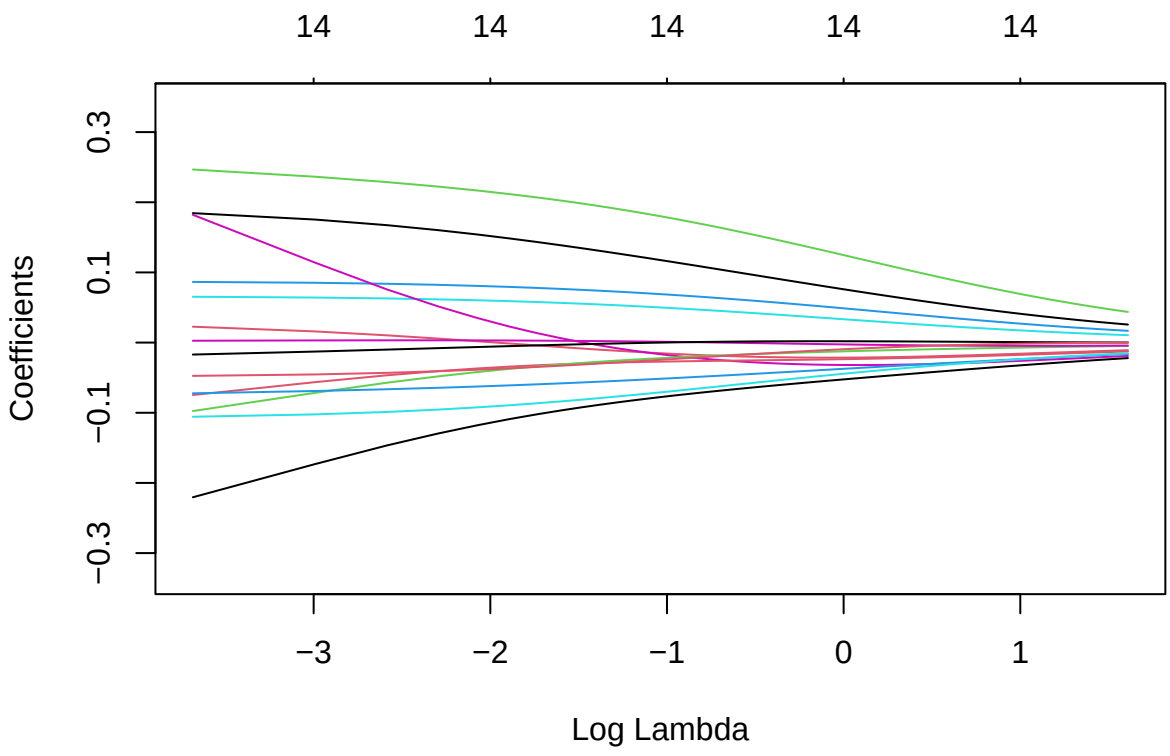
```r
# Creating the 'Vituse_Often' dummy variable
data$Vituse_Often <- ifelse(data$VITUSE == "OFTEN", 1, 0)

# Creating the 'Vituse_No' dummy variable
data$Vituse_No <- ifelse(data$VITUSE == "NO", 1, 0)
data <- data[, !names(data) %in% "VITUSE"]

# Reorder the columns to make RETPLASMA and BETAPLASMA the first two variables
data <- data[c("RETPLASMA", "BETAPLASMA", setdiff(names(data), c("RETPLASMA", "BETAPLASMA")))]

data[,1]<-log(data[,1]+1)
data[,2]<-log(data[,2]+1)

#Data without betaplasma
data2 <- data[, !names(data) %in% "BETAPLASMA"]



#write.csv(data2,"data_second.csv")
```

```r
##       70%  30%
set.seed(123)
d_index <- createDataPartition(data2$RETPLASMA,p = 0.7)
train_d <- data2[d_index$Resample1,]
test_d <- data2[-d_index$Resample1,]
##
scal <- preProcess(train_d,method = c("center","scale"))
train_ds <- predict(scal,train_d)
test_ds <- predict(scal,test_d)
```

```r
lambdas <- seq(0,5, length.out = 200)
X <- as.matrix(train_ds[,2:15])
Y <- train_ds[,1]
set.seed(1245)
ridge_model <- cv.glmnet(X,Y,alpha = 0,lambda = lambdas,nfolds =3)
plot(ridge_model)
```




```
plot(ridge_model$glmnet.fit, "lambda", label = T)
```




```
ridge_min <- ridge_model$lambda.min
##  ridge_min  ridge
ridge_best <- glmnet(X,Y,alpha = 0,lambda = ridge_min)
coef(ridge_best)
```

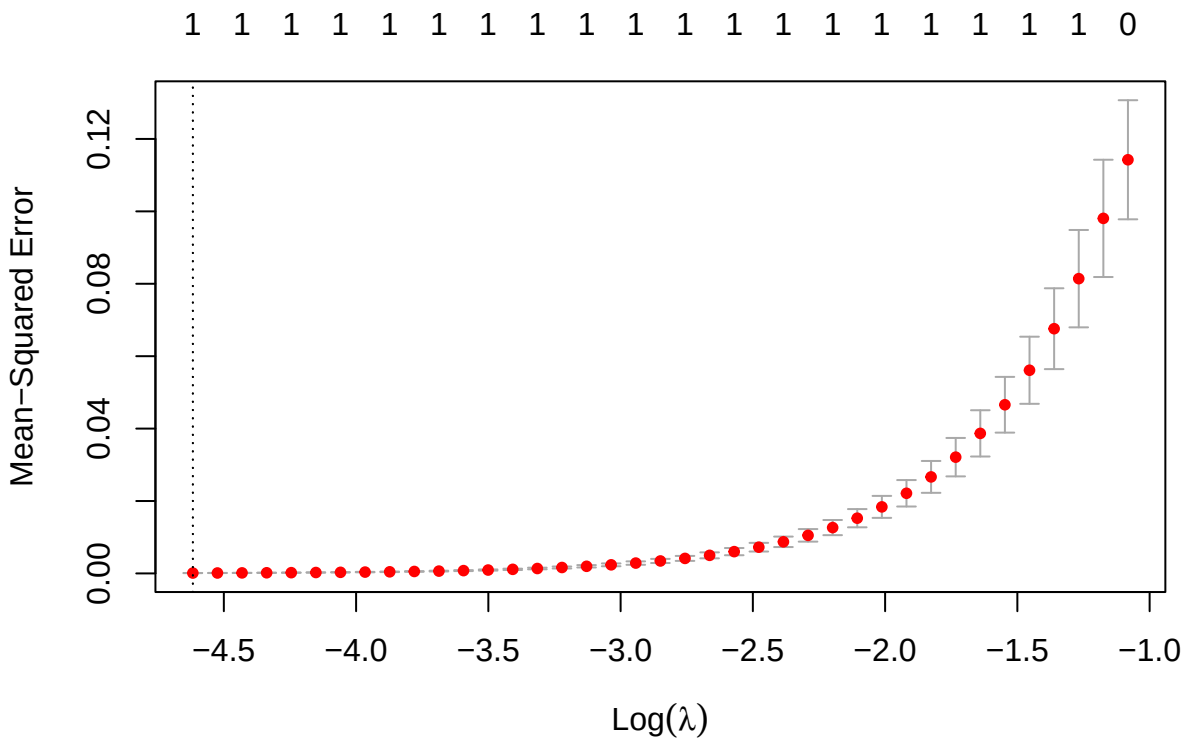
```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)     1.713944e-16
## SMOKSTATFORMER  3.321677e-02
## SMOKSTATNEVER  -1.346340e-02
## AGE             5.629206e-02
## SEX             2.180310e-02
## QUETELET        1.387502e-02
## CALORIES       -2.281896e-02
## FAT            -2.734142e-02
## FIBER          -1.522317e-02
## ALCOHOL        -5.966724e-03
## CHOLESTEROL    -1.994772e-02
## BETADIET       -1.938990e-02
## RETDIET        -4.587777e-03
## Vituse_Often    4.016285e-04
## Vituse_No      -6.477459e-04
```

```
test_pre <- predict(ridge_best,as.matrix(test_ds[,2:15]))
sprintf("      : %f",mae(test_ds$RETPLASMA,test_pre))
```

```
## [1] "      : 0.783092"
```

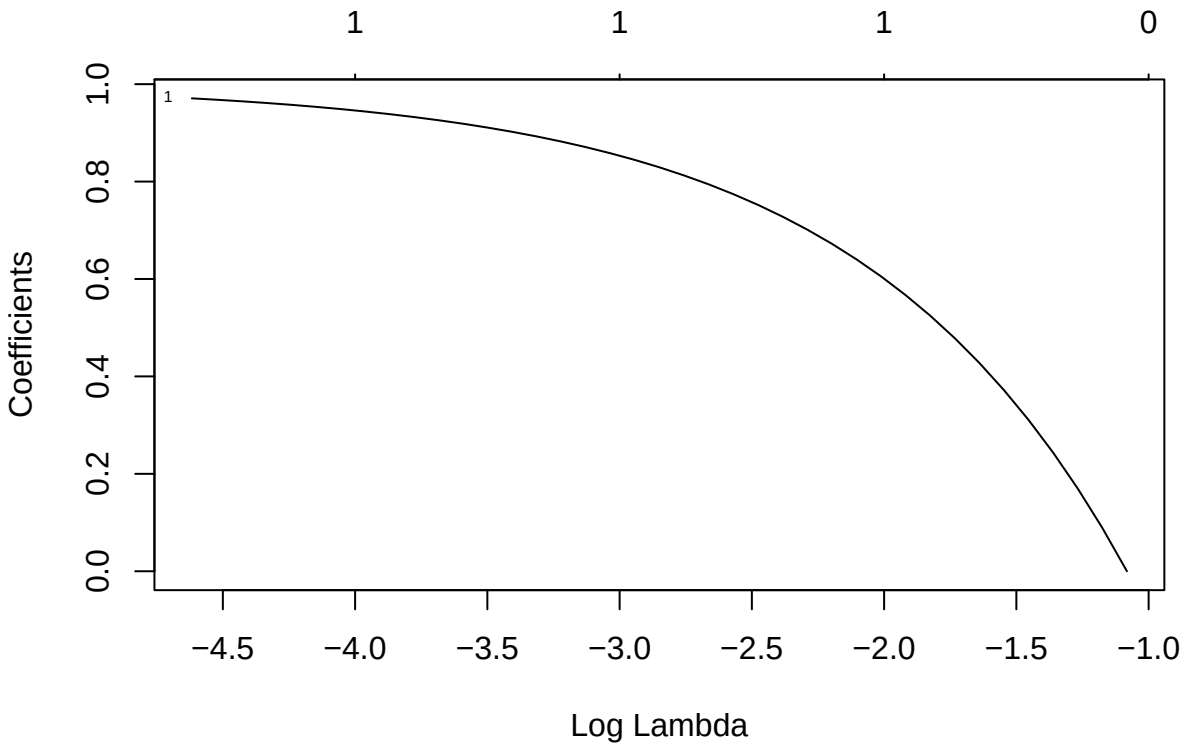
```
## [1] "      : 0.783092"
```

```
# Using lasso model
# Assuming all columns except 'RETPLASMA' are predictors
# Convert categorical variables if there are any


# Prepare the data
x <- as.matrix(data[, !names(data) %in% "BETAPLASMA"]) # Exclude 'Unnamed: 0' and response variable
y <- as.matrix(data$RETPLASMA)



# Fit the LASSO model
set.seed(123) # For reproducibility
cv_model <- cv.glmnet(x, y, alpha = 1) # alpha = 1 for LASSO
plot(cv_model)
```




```
plot(cv_model$glmnet.fit, "lambda", label = T)
```

```
## Warning in plotCoef(x$beta, lambda = x$lambda, df = x$df, dev = x$dev.ratio, : 1
## or less nonzero coefficients; glmnet plot is not meaningful
```




```r
# Optimal lambda
opt_lambda <- cv_model$lambda.min

# Fit model on selected lambda
lasso_model <- glmnet(x, y, alpha = 1, lambda = opt_lambda)

# View the coefficients
coef(lasso_model)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)    0.1850120
## RETPLASMA      0.9708495
## SMOKSTATFORMER .
## SMOKSTATNEVER  .
## AGE            .
## SEX            .
## QUETELET       .
## CALORIES       .
## FAT            .
## FIBER          .
## ALCOHOL        .
## CHOLESTEROL    .
## BETADIET       .
## RETDIET        .
## Vituse_Often   .
```

```
## Vituse_No      .
```

```
summary(lasso_model)
```

```
##           Length Class     Mode
## a0         1     -none-    numeric
## beta      15     dgCMatrix S4
## df         1     -none-    numeric
## dim        2     -none-    numeric
## lambda     1     -none-    numeric
## dev.ratio  1     -none-    numeric
## nulldev    1     -none-    numeric
## npasses    1     -none-    numeric
## jerr       1     -none-    numeric
## offset     1     -none-    logical
## call       5     -none-    call
## nobs       1     -none-    numeric
```