Determinants of Plasma Retinol and Beta-Carotene Levels

Huang Fang (913439658) *email* : $hgfang@ucdavis.edu$
Meng Li (913470824) *email* : $moeli@ucdavis.edu$
Yuan Xu (913489828) *email* : $yxyxu@udavis.edu$

Abstract

This project is designed to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids using linear regression model. We use AIC criterion to select the best models and model validation based on C_p, $Press_p$ and $MSPE_v$ to choose the first-order models instead of second-order model, which indicate that *age*, *smokstat*, fat, *betadiet* are statistically significant factors on plasma beta-carotene levels and *quetelet*, *vituse*, *cholesterol*, *betadiet*, *age* and *sex* are significant factors on plasma beta-carotene. The results basically support previous study however some variables used to be thought significant turned out not significant in our model and the situation is the same the other way around.

Introuction

Several researches have supported that low human plasma concentrations of beta-carotene, retinol, or other carotenoids may have strong association with an increased risk of developing cancer [1-3]. Due to the important role of beta-carotene and retinol, more attention is giving to the determinants of plasma levels of beta-carotene and retinol. Nierenberg et al. [4] claimed that dietary carotene were positively related to beta-carotene levels, while cigarette smoking and Quetelet index were negatively related to that. Men were reported to have lower plasma levels of retinol, beta-carotene than women[5-6].

A cross-sectional study was designed to investigate the relationship between personal characteristics and dietary micronutrients, and plasma concentrations of retinol, beta-carotene and other carotenoids.

The dataset contains these measurements (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 day, cholesterol consumed, dietary beta-carotene consumed, dietary retinol consumed, plasma beta-carotene, plasma retinol) from 315 patients. Among these variables, $sex, smokstat, vituse$ are categorical variables, and the rests are quantitative variables.

Methods and Results

• Preliminary investigation

From the dataset, we can find that *sex* (1 for male and 2 for female), *smokstat* (1 for never, 2 for former and 3 for current smoker) and *vituse* (1 for fairly often using vitamin, 2 for not often using and 3 for not using) are categorical variables. The rest $age, quetelet, calories, fat, fiber, alcohol, cholest$ $retdiet, betaplasmaandretplasma$ are quantitative variables. No severe correlation between X variables.

No significant statistical evidence is shown to support a high correlation between *retplasma* and *betaplasma* up to 3nd order (see Table 1). So it is safe to fit regression model separately

for *retplasma* and *betaplasma*.

The histograms of the original Y data are right-skewed, and the boxcox plot implies *log* transformation for *retplasma*, $\frac{1}{\sqrt{betaplasma+90}}$ transformation for *betaplasma* (by applying two parameters boxcox method) since there is a zero value from *betaplasma*. The new response variables are denoted by $retplasma^*$ and $betaplasma^*$.

After finishing transformation, the residual plots and Q-Q plots according to initial fits satisfy model assumptions: linearity, normality and constant error variance.

• Model selection

AIC is a powerful criterion in model selection, with both model complexity and fit-goodness considered, so it is chosen as our selection criterion. The pool of our potential variables are: $age, sex, smokstat, quetelet, vituse, calories, fat, fiber, alcohol, cholesterol, betadiet, retdiet$, *betaplasma* and *retplasma*.

Since we have many variables in our pool, so it is easier and more efficient for us to find a local optimazation instead of a global optimatization, which implies forward selection stepwise an useful method for us, by applying the forward stepwise procedure, we find that the best 1st order model for *retplasma* and *betaplasma* are:

$$retplasma^* = \beta_0 + \beta_1 age + \beta_2 former + \beta_3 never + \beta_4 fat + \beta_5 betadiet + \epsilon$$

$$\begin{aligned} betaplasma^* =& \beta_0 + \beta_1 quetelet + \beta_2 notoften + \beta_3 often + \beta_4 cholesterol + \beta_5 betadiet + \beta_6 age + \\ & \beta_7 male + \epsilon \end{aligned}$$

And the best model with 2 way interactions for *retplasma* and *betaplasma* are:

$$retplasma^* = \beta_0 + \beta_1 age + \beta_2 former + \beta_3 never + \beta_4 fat + \beta_5 age \times former + \beta_6 age \times never + \epsilon$$

$$\begin{aligned} betaplasma^* = & \beta_0 + \beta_1 quetelet + \beta_2 notoften + \beta_3 often + \beta_4 cholesterol + \beta_5 betadiet + \beta_6 age + \\ & \beta_7 male + \beta_8 former + \beta_9 never + \beta_{10} notoften \times betadiet + \beta_{11} often \times betadiet \\ & + \beta_{12} betadiet \times former + \beta_{13} betadiet \times never + \epsilon \end{aligned}$$

• Model validation

Comparing the best two models for *retplasma* and *betaplasma*:(See table 2, 3, 4, 5, 6, 7)

(1) Comparing the signs of regression coefficients between training data and validation data. For *retplasma* the best model with 1st order, the sign of *smokstatnever* and *betadiet* changes, and for the best model with 2 way interactions, the sign of *age* changes, there is no major difference between the best 1st order model and the best model with 2 way interactions in this point.

For *betaplasma* the best 1st order model, the sign of all predictors keeps the same, and the best model with 2 way interactions also keeps the same sign. So both these two model are perfect in this point.

(2) Comparing the $Cp, MSE, \frac{Press_p}{n}, MSPE_v$:(See table 6 and table 7), for both *retplasma* and *betaplasma*, the value of $Cp, MSE, \frac{Press_p}{n}, MSPE_v$ are both very similar for 1st order

model and model with interactions. Which means that the two models for *retplasma* and *betaplasma* have approximately the same degree of bias and the same predict ability, and the value of $\frac{Press_p}{n}$ is close to MSE, so overfitting is not a concern to both of our models.

Based on the principle of **parsimony**, we choose the best 1st order model as our final model for both *retplasma* and *betaplasma*.

Final model equation for *retplasma*:
$retplasma^* = \beta_0 + \beta_1 age + \beta_2 former + \beta_3 never + \beta_4 fat + \beta_5 betadiet + \epsilon$

Final model equation for *betaplasma*:
$$betaplasma^* = \beta_0 + \beta_1 quetelet + \beta_2 notoften + \beta_3 often + \beta_4 cholesterol + \beta_5 betadiet + \beta_6 age + \beta_7 male + \epsilon$$

The mode assumptions:
(1) ϵ_i are uncorrelated, $Cov(\epsilon_i, \epsilon_j) = 0 \quad if \quad i \neq j$ (2) $E(\epsilon_i) = 0$ (3) $Var(\epsilon_i) = \sigma^2$

Our final fitted regression functions:
$retplasma^* = 6.17 + 0.0047 age + 0.116 former + 0.01 never - 9 \times 10^{-4} fat - 1.3 \times 10^5 betadiet$
$$betaplasma^* = 6 \times 10^{-2} + 5.9 \times 10^{-4} quetelet - 5.5 \times 10^{-3} notoften - 5.9 \times 10^{-3} often + 1.5 \times 10^5 cholsterol - 1.8 \times 10^{-6} betadiet - 1.5 \times 10^{-4} age + 4 \times 10^{-3} male$$
(Relevant statistics can be found in Table 8 and 9)

From the residual plot and the Q-Q plot of our final model, we find no nonlinear pattern at the residual plot and the residuals are a little bit heavy tailed.(From figure 8)

• Model diagnostic: Outlying and influential cases

Outlying in Y and X observations are identified through examining residuals and leverage values, respectively. For response variables, their corresponding studentized deleted residuals are used for testing the null hypothesis H_0: The model is correct and all cases follow the model. No Y outlier is found by Bonferronis procedure, given significant level α.

For predictor variables, we calculate leverage of each case and large leverage value is an indication of potential outlying in X. In practice, we compare it with twice the value of mean leverage and identify the following X cases: for Plasma Retinol, 36 42 62 73 89 95 124 152 225 253 255 266 274 309 313 cases have X outliers; for Plasma beta-carotene, 25 94 96 112 144 152 190 225 226 236 253 255 257 276 286 296 309 cases have X outliers.

In order to further determine whether the outlying cases (in Y and/or in X) are influential in regression function, we compute Cook's distance D_i and $p_i = P(F_{p,n-p} < D_i)$ for each outlying cases. As a result, There is no case whose p_i is more than 20% thus each case has little influence on the fitted values.

Moreover, residual vs. fitted value plots supports constancy of the error variance, and Q-Q plots implies a strong linear pattern, thus normality assumption holds well.

Conclusion and Discussion

We conclude that plasma concentrations of these two micronutrients are associated with di-

etary habits and personal characteristics. Firstly, plasma retinol and plasma beta-carotene have no obvious association. secondly, plasma retinol is associated with age, $smokstat$, fat, $betadiet$ where plasma retinol is positively related with age, $smokstatformer$ and $smokstatnever$ while negatively related with fat, $betadiet$, which means older people and former smoker and those who consume less fat and beta-carotene tend to have higher plasma retinol level. Last, plasma beta-carotene is associated with $quetelet$, $vituse$, $cholesterol$, $betadiet$, age and sex where plasma beta-carotene is positively related with $betadiet$, age, $vitusenotoften$ and $vituseoften$ while negatively related with $quetelet$, $cholesterol$ and $sexmale$ which reflects older and female people with lower quetelet and those who consume more beta-carotene and vitamin and less cholesterol tend to have higher plasma beta-carotene level. Based on the result of standardization, age and $quetelet$ play the most influential role in determining levels of plasma retinol and plasma beta-carotene among all those useful predictor variables, respectively.

With respect to limitations, after implementation of boxcox to choose the best transformation for response variables, it still appears some heavy tailed pattern in Q-Q plot. Another limitation is our use of stepwise procedure could end up with a suboptimal model rather than the global best model.

In conclusion, the results of our analysis are basically consistent with previous study. For example, it is both agreed that dietary beta-carotene and female sex are positive related to beta-carotene levels while the former study shows negative relation between beta-carotene and smoking status but our analysis shows no significant relation between those two factors. Different potential pools of predictor variables and sizes of data maybe the reason of that. A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study.

Appendices

• Appendix 1: Figures and tables

Table 1: F test of different orders

Order	MSR	MSE	F^*	$F(p-1, n-p)$	$Conclusion$
1	70200	43555.69	1.612	3.87134	not reject H_0
2	91219	43347.24	2.105	3.024681	not reject H_0
3	72194.67	43347.24	1.665	2.63364	not reject H_0

Table 2: 1st order for RETPLASMA

	Intercept	AGE	SMOKSTAT FORMER	SMOKSTAT NEVER	FAT	BETADIET
training	6.256149	0.003636	0.20185156	0.06672407	-0.001658	-2.738098e-05
validation	6.091019	0.005397	0.03630081	-0.04239080	-0.000394	2.895886e-06

Table 3: 2nd order for RETPLASMA

	Intercept	AGE	SMOKSTAT FORMER	SMOKSTAT NEVER	FAT	AGE: SMOKSTAT FORMER	AGE: SMOKSTAT NEVER
training	6.59371	-0.0050	-0.43843	-0.2727511	-0.00163	0.01351	0.00746
validation	6.18163	0.0034	-0.06510	-0.1577390	-0.00036	0.00230	0.00258

Table 4: 1st order for BETAPLASMA

	Intercept	QUETELET	VITUSE NOTOFTEN	VITUSE OFTEN	CHOLESTEROL	BETADIET	AGE	SEXMALE
training	0.05858	0.000631	-0.00729	-0.00765	1.8882e-05	-1.1305e-06	-0.000173	0.0051
validation	0.06101	0.000602	-0.00353	-0.00375	1.2192e-05	-2.6168e-06	-0.000112	0.0031

Table 5: 2nd order for BETAPLASMA

	Intercept	QUETELET	VITUSE NOTOFTEN	VITUSE OFTEN	CHOLESTEROL	BETADIET	AGE
training	0.05148376	0.000644	-0.000541	0.001668	1.205707e-05	6.232810e-06	-1.764190e-04
validation	0.05570450	0.000620	-0.002539	0.003336	1.388229e-05	1.325601e-06	-8.715286e-05
	SEXMALE	SMOKSTAT FORMER	SMOKSTAT NEVER	VITUSE NOT OFTEN :BETADIET	VITUSE OFTEN: BETADIET	BETADIET: SMOKSTAT FORMER	BETADIET: SMOKSTAT NEVER
training	0.004679	0.001421	0.003767	-3.056289e-06	-4.058262e-06	-4.265811e-06	-5.619232e-06
validation	0.001802	0.000999	0.000864	-2.131652e-07	-2.950730e-06	-2.717517e-06	-3.731058e-06

Table 6: Comparison of candidate models for RETPLASMA

	Cp	MSE	Pressp/n	MSPEv
order1	56.30696	0.1098752	0.1045810	0.1155047
order2	56.88520	0.1092649	0.1045346	0.1177755

Table 7: Comparison of candidate models for BETAPLASMA

	Cp	MSE	Pressp/n	MSPEv
order1	-4.810602	0.000164	0.000156	0.000153
order2	-6.398431	0.000165	0.000147	0.000149

Table 8: Final model for RETPLASMA

	Intercept	AGE	SMOKSTAT FORMER	SMOKSTAT NEVER	FAT	BETADIET
coefficient	6.165	0.0047	0.116	0.012	-0.00093	-1.343e-05
Std. Error	9.538e-02	1.309e-03	5.988e-02	5.791e-02	5.692e-04	1.291e-05
p-value	¡ 2e-16	0.000433	0.052925	0.860615	0.103616	0.298856

Table 9: Final model for BETAPLASMA

	Intercept	QUETELET	VITUSE NOTOFTEN	VITUSE OFTEN	CHOLESTEROL	BETADIET	AGE	SEXMALE
coefficient	6.074e-02	5.859e-04	-5.459e-03	-5.852e-03	1.508e-05	-1.816e-06	-1.458e-04	4.034e-03
Std.error	4.412e-03	1.165e-04	1.829e-03	1.643e-03	5.635e-06	4.789e-07	5.131e-05	2.260e-03
p-value	¡ 2e-16	8.39e-07	0.003064	0.000426	0.007856	0.000181	0.004785	0.075314

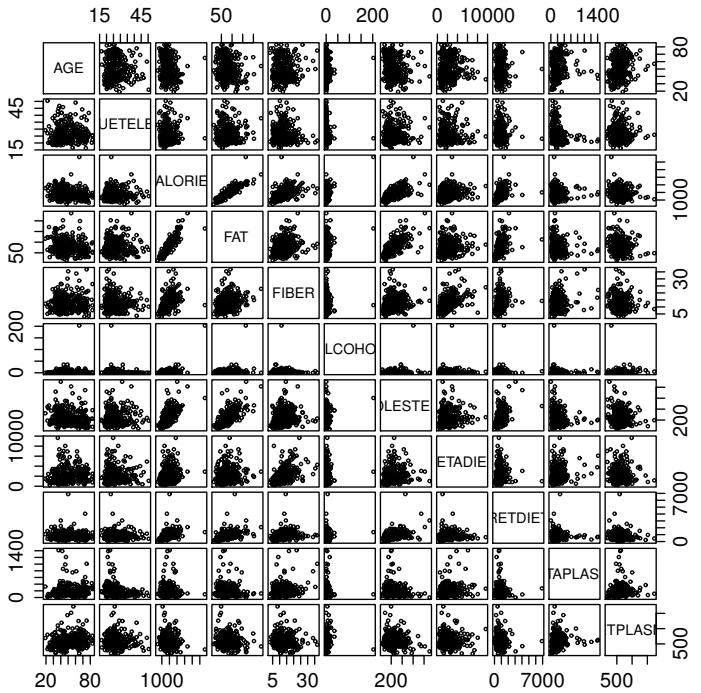

Figure 1: pairs plot

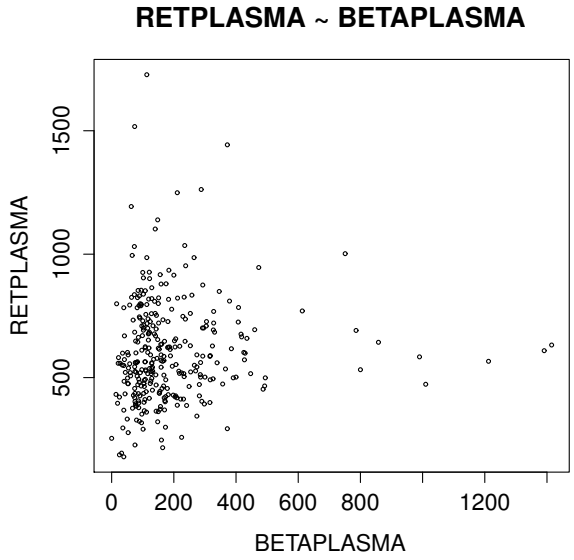

Figure 2: scatter plot

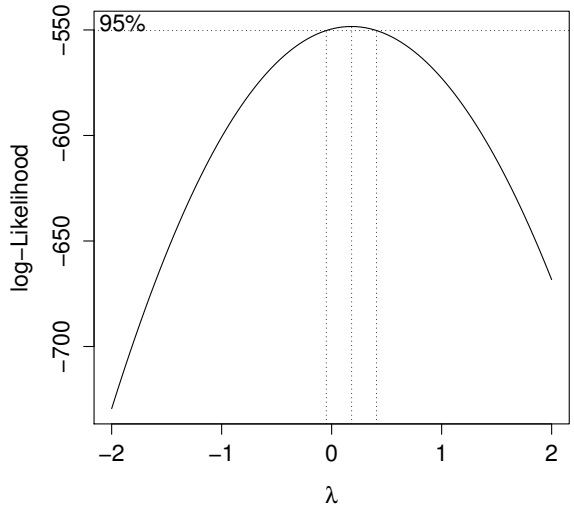

Figure 3: retplasma boxcox

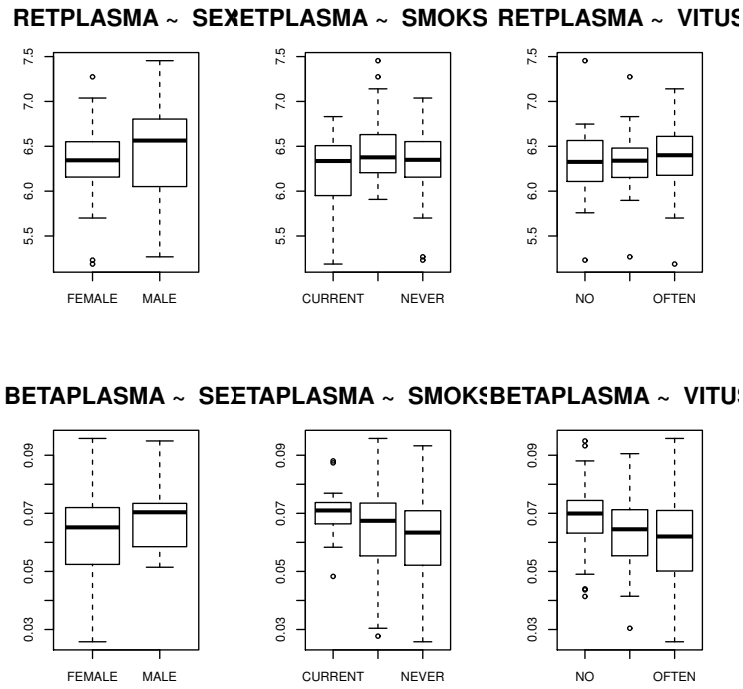

Figure 4: boxplot training and validation

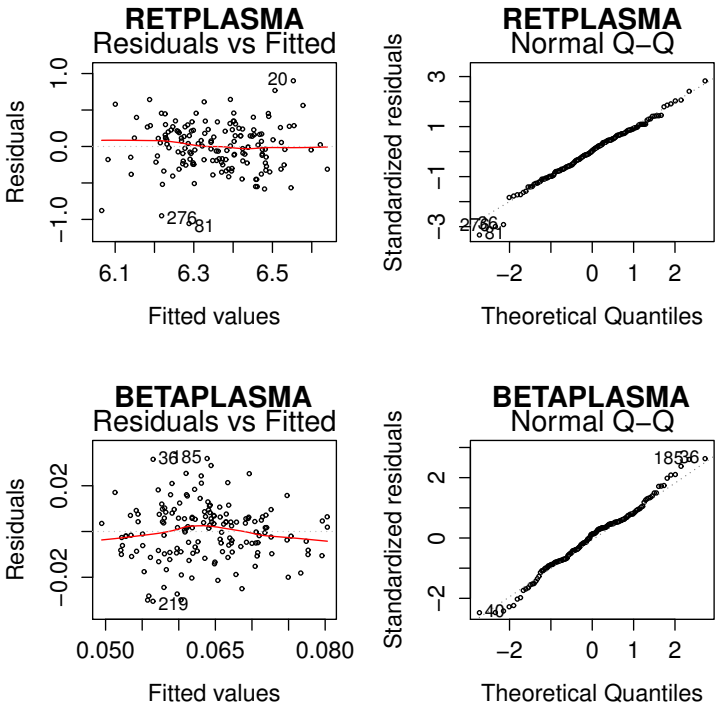

Figure 5: Residual plot and QQ plot

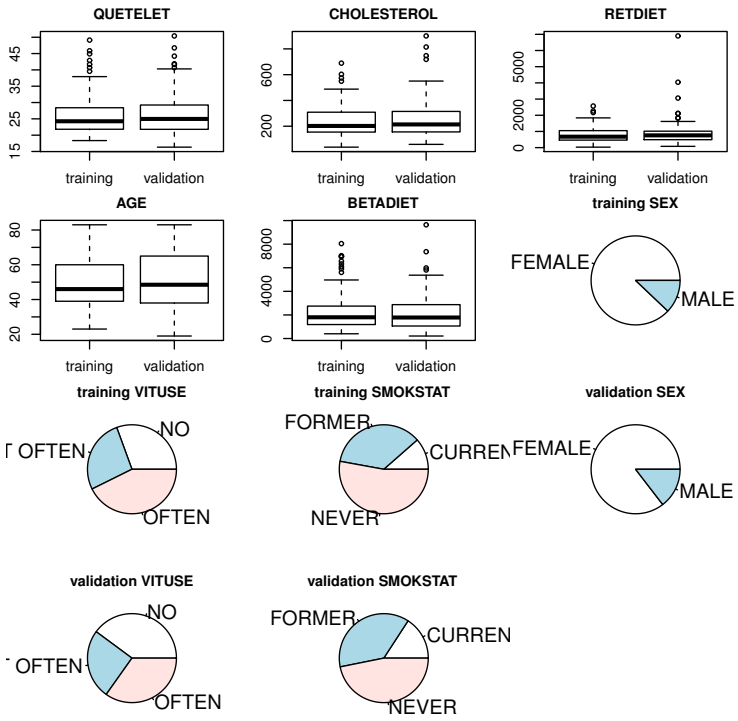

Figure 6: training validation

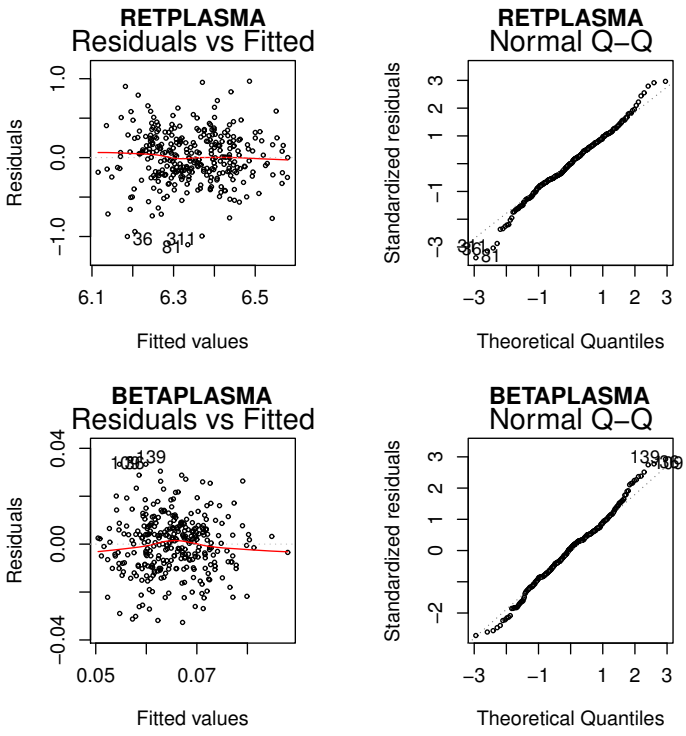

Figure 7: Residual plot and QQ plot

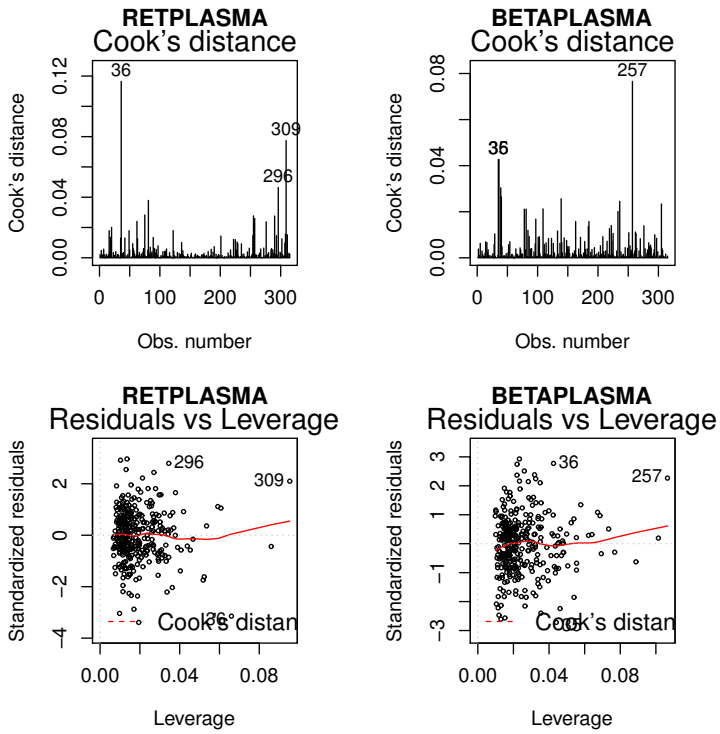

Figure 8: Cook's distance and Residual leverage

• Appendix 2: R codes and outputs

```
> library(MASS)
> library(geoR)
> options(width = 80)
> plasma = read.table("~/academic/Sta206/Plasma.txt", header = TRUE)
> rownames(plasma) = c(1: nrow(plasma))
> dim(plasma)
[1] 315  14
> sapply(plasma, class)
        AGE         SEX    SMOKSTAT    QUETELET      VITUSE    CALORIES
  "integer"    "factor"    "factor"   "numeric"    "factor"   "numeric"
        FAT       FIBER     ALCOHOL CHOLESTEROL    BETADIET     RETDIET
  "numeric"   "numeric"   "numeric"   "numeric"   "integer"   "integer"
 BETAPLASMA   RETPLASMA
  "integer"   "integer"
> var = colnames(plasma)
> qual_var = c("SEX", "SMOKSTAT", "VITUSE")
> quant_var = var[!var \%in\% qual_var]
> b_var = var[-which(var == "RETPLASMA")]
> r_var = var[-which(var == "BETAPLASMA")]
> pred_var =  var[-which(var == "BETAPLASMA" | var == "RETPLASMA")]
```

```
> round(cor(plasma[, quant_var]), 3)
               AGE QUETELET CALORIES    FAT  FIBER ALCOHOL CHOLESTEROL BETADIET
AGE          1.000   -0.017   -0.177 -0.169  0.045   0.052      -0.114    0.072
QUETELET    -0.017    1.000    0.004  0.049 -0.088  -0.073       0.110   -0.007
CALORIES    -0.177    0.004    1.000  0.872  0.465   0.451       0.659    0.243
FAT         -0.169    0.049    0.872  1.000  0.276   0.186       0.710    0.143
FIBER        0.045   -0.088    0.465  0.276  1.000  -0.020       0.154    0.483
ALCOHOL      0.052   -0.073    0.451  0.186 -0.020   1.000       0.182    0.039
CHOLESTEROL -0.114    0.110    0.659  0.710  0.154   0.182       1.000    0.116
BETADIET     0.072   -0.007    0.243  0.143  0.483   0.039       0.116    1.000
RETDIET     -0.010    0.032    0.402  0.412  0.215   0.045       0.443    0.053
BETAPLASMA   0.101   -0.229   -0.022 -0.092  0.236  -0.022      -0.130    0.225
RETPLASMA    0.212    0.013   -0.073 -0.091 -0.044   0.017      -0.070   -0.014
            RETDIET BETAPLASMA RETPLASMA
AGE          -0.010      0.101     0.212
QUETELET      0.032     -0.229     0.013
CALORIES      0.402     -0.022    -0.073
FAT           0.412     -0.092    -0.091
FIBER         0.215      0.236    -0.044
ALCOHOL       0.045     -0.022     0.017
CHOLESTEROL   0.443     -0.130    -0.070
BETADIET      0.053      0.225    -0.014
RETDIET       1.000     -0.046    -0.063
BETAPLASMA   -0.046      1.000     0.072
RETPLASMA    -0.063      0.072     1.000
> pairs(plasma[, quant_var], gap = 0.3, cex = 0.5)

> par(mfrow = c(1,1))
> plot(plasma$BETAPLASMA, plasma$RETPLASMA, main = "RETPLASMA ~ BETAPLASMA", cex
= 0.5,
+ xlab = "BETAPLASMA", ylab = "RETPLASMA")

fit1 = lm(RETPLASMA ~ BETAPLASMA, data = plasma)
fit2 = lm(RETPLASMA ~ BETAPLASMA + I(BETAPLASMA^2), data = plasma)
fit3 = lm(RETPLASMA ~ BETAPLASMA + I(BETAPLASMA^2) + I(BETAPLASMA^3), data = plasma)

fit_order1_r = lm(RETPLASMA ~ ., data = plasma[, r_var])
#In oder to apply boxcox transformation, we have to exclude the case where BETAPLASMA
= 0.
boxcox(fit_order1_r)

#We should apply log transformation to RETPLASMA.
#However BETAPLASMA includes 0, so we should use double parameter boxcox.
> boxcoxfit(plasma$BETAPLASMA, lambda2 = TRUE)
Fitted parameters:
```

```
       lambda       lambda2          beta       sigmasq
-0.6286177497 89.5735798168  1.5388697521  0.0001843618

Convergence code returned by optim: 0
#So the transformation for RETPLASMA is log, and the transformation for BETAPLASMA
is 1/sqrt(x + 90)
> plasma$RETPLASMA = log(plasma$RETPLASMA)
> plasma$BETAPLASMA = 1/sqrt(plasma$BETAPLASMA + 90)
> #split the data into training set and validation set. the ratio we set is 3:1
> set.seed(10)
> data = plasma[sample(1:nrow(plasma)), ]
> data_v = data[1:(round(nrow(data)/2)), ]
> data_t = data[-(1:(round(nrow(data)/2))), ]
> par(mfrow = c(2,3))
> invisible(
+ sapply(qual_var, function(x)
+ boxplot(data_t$RETPLASMA ~ data_t[[x]], main = paste("RETPLASMA ~ ", x)))
+ )
> invisible(
+ sapply(qual_var, function(x)
+ boxplot(data_t$BETAPLASMA ~ data_t[[x]], main = paste("BETAPLASMA ~ ", x)))
+ )

> fit_null_r = lm(RETPLASMA ~ 1, data = data_t[, r_var])
> fit_order1_r = lm(RETPLASMA ~ ., data = data_t[, r_var])
> fit_null_b = lm(BETAPLASMA ~ 1, data = data_t[, b_var])
> fit_order1_b = lm(BETAPLASMA ~ ., data = data_t[, b_var])

> stepAIC(fit_null_r, scope = list(upper = fit_order1_r), direction = "both", k
= 2)
> fit_order1_best_r = lm(formula = RETPLASMA ~ AGE + SMOKSTAT + FAT + BETADIET,
data = data_t[, r_var])
> stepAIC(fit_null_b, scope = list(upper = fit_order1_b), direction = "both", k
= 2)
> fit_order1_best_b = lm(formula = BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL
+ BETADIET + AGE + SEX, data = data_t[, b_var])
> par(mfrow = c(2,2))
> #Plots for the best 1st order model
> plot(fit_order1_best_r, which = 1, cex = 0.5, main = "RETPLASMA")
> plot(fit_order1_best_r, which = 2, cex = 0.5, main = "RETPLASMA")
> plot(fit_order1_best_b, which = 1, cex = 0.5, main = "BETAPLASMA")
> plot(fit_order1_best_b, which = 2, cex = 0.5, main = "BETAPLASMA")

> #Assumptions hold
> #With 2 way interaction.
```

```
> fit_order2_r = lm(RETPLASMA ~ . + .^2, data = data_t[, r_var])
> fit_order2_b = lm(BETAPLASMA ~ . + .^2, data = data_t[, b_var])
> stepAIC(fit_null_r, scope = list(upper = fit_order2_r), direction = "both", k
= 2)
> fit_order2_best_r = lm(formula = RETPLASMA ~ AGE + SMOKSTAT + FAT + AGE:SMOKSTAT,
data = data_t[, r_var])

> stepAIC(fit_null_b, scope = list(upper = fit_order2_b), direction = "both", k
= 2)
> fit_order2_best_b = lm(formula = BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL
+ BETADIET +
                         AGE + SEX + SMOKSTAT + VITUSE:BETADIET + BETADIET:SMOKSTAT,
                       data = data_t[, b_var])
> plot(fit_order2_best_r, which = 1, cex = 0.5, main = "RETPLASMA")
> plot(fit_order2_best_r, which = 2, cex = 0.5, main = "RETPLASMA")
> plot(fit_order2_best_b, which = 1, cex = 0.5, main = "BETAPLASMA")
> plot(fit_order2_best_b, which = 2, cex = 0.5, main = "BETAPLASMA")
#Approximately the same to previous.

#Check if there is difference between the distribution of variables in training
and validation sets.
> par(mfrow = c(4,3), mar = c(2, 2, 2, 2))
> var_quant_check = c("QUETELET", "CHOLESTEROL", "RETDIET", "AGE", "BETADIET")
> var_qual_check = c("SEX", "VITUSE", "SMOKSTAT")
> invisible(
  sapply(var_quant_check, function(x)
    boxplot(data_t[[x]], data_v[[x]], names = c("training", "validation"), main
= x))
)
> invisible(
  sapply(var_qual_check, function(x)
    pie(table(data_t[[x]]), main = paste("training", x)))
)

> invisible(
  sapply(var_qual_check, function(x)
    pie(table(data_v[[x]]), main = paste("validation", x)))
)

#Compare 1st and 2nd order model
#Sign, Cp, Pressp, MSPE
#Sign
> newdata = data_v[, pred_var]
> fit_order1_r_v = lm(formula = RETPLASMA ~ AGE + SMOKSTAT + FAT + BETADIET, data
= data_v[, r_var])
```

```
> fit_order2_r_v = lm(formula = RETPLASMA ~ AGE + SMOKSTAT + FAT + AGE:SMOKSTAT,
data = data_v[, r_var])
> fit_order1_b_v = lm(formula = BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL + BETADIET
+ AGE + SEX, data = data_v[, b_var])
> fit_order2_b_v = lm(formula = BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL + BETADIET
+
+ AGE + SEX + SMOKSTAT + VITUSE:BETADIET + BETADIET:SMOKSTAT,
+ data = data_v[, b_var])
> sign_order1_r = rbind(fit_order1_best_r$coefficients, fit_order1_r_v$coefficients)
> rownames(sign_order1_r) = c("training", "validation")
> sign_order2_r = rbind(fit_order2_best_r$coefficients, fit_order2_r_v$coefficients)
> rownames(sign_order2_r) = c("training", "validation")
> sign_order1_b = rbind(fit_order1_best_b$coefficients, fit_order1_b_v$coefficients)
> rownames(sign_order1_b) = c("training", "validation")
> sign_order2_b = rbind(fit_order2_best_b$coefficients, fit_order2_b_v$coefficients)
> rownames(sign_order2_b) = c("training", "validation")
> sign_order1_r
           (Intercept)         AGE SMOKSTATFORMER SMOKSTATNEVER           FAT
training      6.256149 0.003635579     0.20185156    0.06672407 -0.0016581294
validation    6.091019 0.005397297     0.03630081   -0.04239080 -0.0003942241
                BETADIET
training   -2.738098e-05
validation  2.895886e-06
> sign_order2_r
           (Intercept)          AGE SMOKSTATFORMER SMOKSTATNEVER           FAT
training      6.593710 -0.005006310    -0.43842523    -0.2727511 -0.0016289936
validation    6.181629  0.003375977    -0.06509568    -0.1577390 -0.0003584559
           AGE:SMOKSTATFORMER AGE:SMOKSTATNEVER
training          0.013509118       0.007464006
validation        0.002301113       0.002581345
> sign_order1_b
           (Intercept)     QUETELET VITUSENOT OFTEN  VITUSEOFTEN  CHOLESTEROL
training    0.05858431 0.0006310318    -0.007292306 -0.007651452 1.888199e-05
validation  0.06100500 0.0006019101    -0.003532618 -0.003748463 1.219154e-05
                BETADIET           AGE     SEXMALE
training   -1.130462e-06 -0.0001729479 0.005123517
validation -2.616761e-06 -0.0001120571 0.003096017
> sign_order2_b
           (Intercept)     QUETELET VITUSENOT OFTEN VITUSEOFTEN  CHOLESTEROL
training    0.05148376 0.0006442790    -0.000540827 0.001668205 1.205707e-05
validation  0.05570450 0.0006199946    -0.002538669 0.003336168 1.388229e-05
              BETADIET           AGE     SEXMALE SMOKSTATFORMER SMOKSTATNEVER
training   6.232810e-06 -1.764190e-04 0.004678885   0.0014210159  0.0037673433
validation 1.325601e-06 -8.715286e-05 0.001802016   0.0009987548  0.0008641907
           VITUSENOT OFTEN:BETADIET VITUSEOFTEN:BETADIET
```

```
training              -3.056289e-06        -4.058262e-06
validation            -2.131652e-07        -2.950730e-06
           BETADIET:SMOKSTATFORMER BETADIET:SMOKSTATNEVER
training             -4.265811e-06          -5.619232e-06
validation           -2.717517e-06          -3.731058e-06
#For RETPLASMA
#Using Full model to estimate sigma2
> n = nrow(data_t)
> sigma2_r = sum(fit_order2_r$residuals^2)/(n - length(fit_order2_r$coefficients))
> sigma2_b = sum(fit_order2_b$residuals^2)/(n - length(fit_order2_b$coefficients))
> compare = function(fit1, fit2, n, sigma2, response, newdata){
+ p1 = length(fit1$coefficients)
+ p2 = length(fit2$coefficients)
+ MSE1 = sum(residuals(fit1)^2)/(n - p1)
+ MSE2 = sum(residuals(fit2)^2)/(n - p2)
+ Cp1 = sum(residuals(fit1)^2)/sigma2 - (n - 2*p1)
+ Cp2 = sum(residuals(fit2)^2)/sigma2 - (n - 2*p2)
+ Press_p1 = sum(fit1$residuals^2/(1-influence(fit1)$hat)^2)
+ Press_p2 = sum(fit2$residuals^2/(1-influence(fit2)$hat)^2)
+ MSPEv1 = mean((data_v[[response]] - predict(fit1, newdata))^2)
+ MSPEv2 = mean((data_v[[response]] - predict(fit2, newdata))^2)
+ row1 = c(Cp1, Press_p1/n, MSE1, MSPEv1)
+ row2 = c(Cp2, Press_p2/n, MSE2, MSPEv2)
+ result = rbind(row1, row2)
+ colnames(result) = c("Cp", "MSE", "Pressp/n", "MSPEv")
+ rownames(result) = c("order1", "order2")
+ return(result)
+ }
> compare(fit_order1_best_r, fit_order2_best_r, n, sigma2_r, "RETPLASMA", newdata)
             Cp       MSE  Pressp/n     MSPEv
order1 56.30696 0.1098752 0.1045810 0.1155047
order2 56.88520 0.1092649 0.1045346 0.1177755
> compare(fit_order1_best_b, fit_order2_best_b, n, sigma2_b, "BETAPLASMA", newdata)
              Cp          MSE     Pressp/n        MSPEv
order1 -4.810602 0.0001638030 0.0001563827 0.0001527844
order2 -6.398431 0.0001654556 0.0001466870 0.0001488832
#According to parsimony principle, we choose the first order models as our final
models
> fit_final_r = lm(RETPLASMA ~ AGE + SMOKSTAT + FAT + BETADIET, data = plasma[,
r_var])
> fit_final_b = lm(BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL + BETADIET + AGE
+ SEX, data = plasma[, b_var])

par(mfrow = c(2,2), mar = rep(4, 4))
plot(fit_final_r, which = 1, cex = 0.5, main = "RETPLASMA")
```

```
plot(fit_final_r, which = 2, cex = 0.5, main = "RETPLASMA")
plot(fit_final_b, which = 1, cex = 0.5, main = "BETAPLASMA")
plot(fit_final_b, which = 2, cex = 0.5, main = "BETAPLASMA")

> outlyingY = function(fit, alpha, n = nrow(plasma), p = length(coef(fit)))
+ {
+ stu_del_res = studres(fit)
+ bfrn_thld = qt(1 - alpha/(2*n), n - p - 1)
+ sort(stu_del_res[which(stu_del_res > bfrn_thld)], decreasing = TRUE)
+ }
> outlyingY(fit_final_r, .1)
named numeric(0)
> outlyingY(fit_final_b, .1)  #no outlying Y
named numeric(0)

> outlyingX = function(fit, n = nrow(plasma), p = length(coef(fit)))
+ {
+ h = as.vector(influence(fit)$hat)
+ which(h > (2*p/n))
+ }
> outlyingX(fit_final_r)
 [1]  36  42  62  73  89  95 124 152 225 253 255 266 274 309 313
> outlyingX(fit_final_b)
 [1]  25  94  96 112 144 152 190 225 226 236 253 255 257 276 286 296 309

> cook_dist = function(fit, n = nrow(plasma), p = length(coef(fit)))
+ {
+ h = as.vector(influence(fit)$hat)
+ res = fit$residuals
+ mse = anova(fit)["Residuals", 3]
+ cook_d = sort(res^2*h/(p*mse*(1-h)^2), decreasing = TRUE)
+ p_i = round(pf(cook_d, p, n - p), 3)
+ tbl = rbind(round(cook_d, 3), p_i)
+ rownames(tbl) = c("cook's distance", "p_i")
+ tbl_some_large = tbl[, tbl[2, ] > 0.2]
+ return(tbl_some_large)
+ }
> cook_dist(fit_final_r)

cook's distance
p_i
> cook_dist(fit_final_b)    #no influential cases

cook's distance
p_i
```

```
> #Let's see the case with top3 cook's distance and its p_i
> cookDistTop3 = function(fit, n = nrow(plasma), p = length(coef(fit)))
+ {
+ h = as.vector(influence(fit)$hat)
+ res = fit$residuals
+ mse = anova(fit)["Residuals", 3]
+ cook_d = sort(res^2*h/(p*mse*(1-h)^2), decreasing = TRUE)[1:3]
+ p_i = round(pf(cook_d, p, n - p), 3)
+ tbl = rbind(round(cook_d, 3), p_i)
+ rownames(tbl) = c("cook's distance", "p_i")
+ return(tbl)
+ }

> #ret
> cookDistTop3(fit_final_r)
                   36   309   296
cook's distance 0.116 0.077 0.046
p_i             0.006 0.002 0.000
> #beta
> cookDistTop3(fit_final_b)
                  257    36    35
cook's distance 0.077 0.043 0.043
p_i             0.000 0.000 0.000

> #########Standardization (centering and rescale)###################
> plasma_star = cbind(scale(plasma[, quant_var])/(nrow(plasma) - 1), plasma[, qual_var])
> fit_final_r_s = lm(RETPLASMA ~ AGE + SMOKSTAT + FAT + BETADIET, data = plasma_star[,
r_var])
> fit_final_b_s = lm(BETAPLASMA ~ QUETELET + VITUSE + CHOLESTEROL + BETADIET + AGE
+ SEX, data = plasma_star[, b_var])
> coef(fit_final_r_s)
   (Intercept)            AGE SMOKSTATFORMER  SMOKSTATNEVER            FAT
 -4.451669e-04   1.995890e-01   1.089298e-03   9.527583e-05  -9.240866e-02
      BETADIET
 -5.819451e-02
> coef(fit_final_b_s)
    (Intercept)        QUETELET VITUSENOT OFTEN     VITUSEOFTEN     CHOLESTEROL
   0.0007267621    0.2552668249   -0.0012595613   -0.0013502925    0.1441717273
       BETADIET             AGE         SEXMALE
  -0.1938581313   -0.1539851305    0.0009307059
```

Reference

[1] 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.

[2] Wald NJ. Retinol, beta-carotene and cancer. Cancer Surv1987; 6: 635-651.

[3] Russell-Briefel R, Bates MW, Kuller LH. The relationship of plasma carotenoids to health and biochemical factors in middle-aged men. Am J Epidemiol 1985; 122: 741-749.

[4] Stryker WS, Kaplan LA, Stein EA, Stampfer MJ, Sober A, Willett WC. The relation of diet, cigarette smoking, and alcohol consumption to plasma beta-carotene and alpha-tocopherol levels. Am J Epidemiol1988; 127: 283-296.

[5] Dimitrov NV, Boone CW, Hay MB. Plasma beta-carotene levels: kinetic patterns during administration of various doses of beta-carotene. J Nutr Growth Cancer1987; 3: 227-238.