Xue, Mi

STAT 586

Final Project

Statistical Analysis of Plasma Retinol Levels

May. 05, 2017

Introduction

It is well known that dietary intake, beta-carotene, retinol are strongly associated with the risk of developing cancer, however, due to the insufficient or conflicting evidences, the relationship between, personal factors, carotenoids and diseases are still inconclusive. In order to find out the determinants of plasma concentration of these micronutrients, we use the data acquired from the "patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous." The purpose of this project is to derive an appropriate regression models to summarize the characteristics between personally factors, micronutrients and plasma Retinol level.

Data

The dataset can be referenced at American Journal of Epidemiology 1989; 130:511- 521. It has not been published yet but a related reference is Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. This dataset has 13 variables, consists of 3 categorical variables and 9 quantitative variables.

Quantitative Variables:

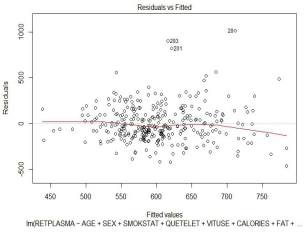
AGE: age (years);
QUETELET: weight height 2;
CALORIE: number of calories consumed per day;
FAT: grams of fat consumed per day;
FIBER: grams of fiber consumed per day;
ALCOHOL: number of alcoholic drinks consumed per day;
CHOLESTEROL: Cholesterol consumed (mg per day);
BETADIET: Dietary beta-carotene consumed (mcg per day);.
RETDIET: Dietary retinol consumed (mcg per day);
RETPLASMA: Plasma Retinol (ng/ml), the response variable;

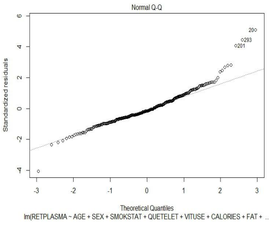
Categorical Variables:

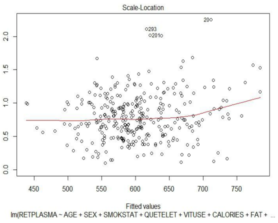
SEX: sex (1=male, 2=female);
SMOKESTAT: smoking status (1=never, 2=former, 3=current);
VITUSE: Vitamin Use (0= No, 1=Yes, fairly often, 2=Yes, not often).

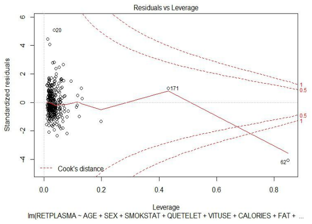
Interpretation

Plotting the original dataset can show a much more indicative view of the data's consistence and normality. Four graphs are shown below, suggesting that the QQ plot (2) is heavy tailed, suggesting how the data depart from normality. Residuals VS leverage (4) graph has a high leverage and large standardized residuals point out of Cook's distance. Therefore, transformation is needed so as to make the data approximately normal as well as to delete the outlier out the special point #62.

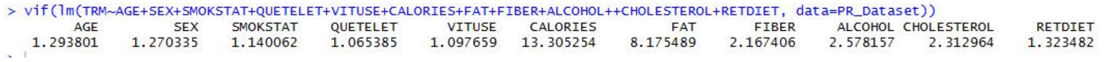






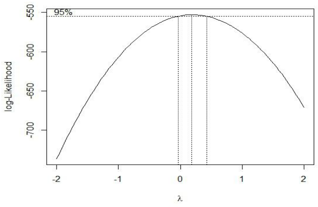


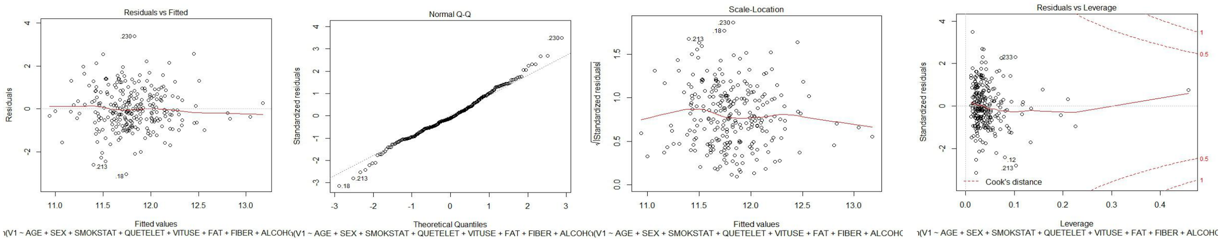
Besides boxcox transformation and deleting outliers, there is one more step need to be done. In order to build linear regression models, it is necessary to check collinearity among variables because it can interfere the correctness of the final models. Hence I did the VIF test to check the data and found that variable CALORIES shows a suspicious value 13.305254(>10), indicating CALORIES truly can cause the collinearity problem with other variables, so that I need to delete CALORIES from the full model.



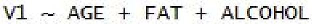
```
> vif(lm(TRM~AGE+SEX+SMOKSTAT+QUETELET+VITUSE+CALORIES+FAT+FIBER+ALCOHOL++CHOLESTEROL+RETDIET, data=PR_Dataset))
      AGE        SEX   SMOKSTAT   QUETELET     VITUSE   CALORIES        FAT      FIBER    ALCOHOL CHOLESTEROL    RETDIET
 1.293801   1.270335   1.140062   1.065385   1.097659  13.305254   8.175489   2.167406   2.578157    2.312964   1.323482
```

After assigned the value of λ =0.18(the peak of log-likelihood graph), we can successfully do the transformation, After removing point 62#, the four graphs are shown below， which can be seen that the data is almost follow a normal distribution and no point has a outlier influence now:



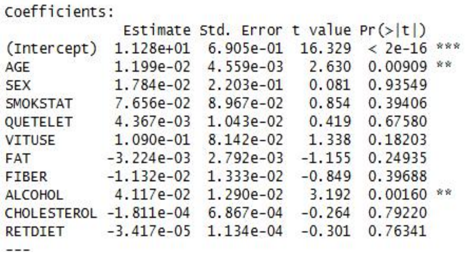


The next step is checking the significant of all the variables in the full model (after deleting the variable CALORIES), we use stepwise selection to allow moves in each direction. The result is

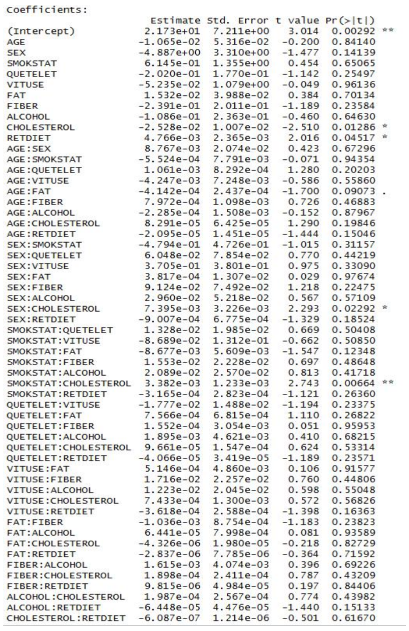


```
V1 ~ AGE + FAT + ALCOHOL
```

However, some variables may be removed from the model, when they are deemed important to be included. So I use another method to select the variables. We know that two independent variables interact if the effect of one of the variables differs depending on the level of the other variable. So I use the normal ANOVA analysis and then add all interaction effects into the model, and the outcome is shown below:

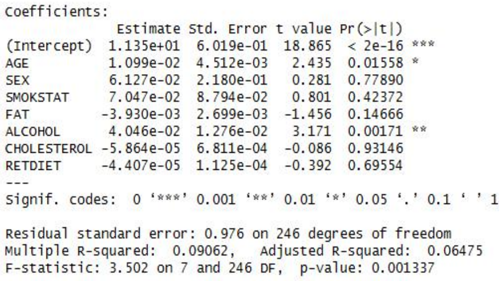


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.128e+01  6.905e-01  16.329  < 2e-16 ***
AGE          1.199e-02  4.559e-03   2.630  0.00909 ** 
SEX          1.784e-02  2.203e-01   0.081  0.93549    
SMOKSTAT     7.656e-02  8.967e-02   0.854  0.39406    
QUETELET     4.367e-03  1.043e-02   0.419  0.67580    
VITUSE       1.090e-01  8.142e-02   1.338  0.18203    
FAT         -3.224e-03  2.792e-03  -1.155  0.24935    
FIBER       -1.132e-02  1.333e-02  -0.849  0.39688    
ALCOHOL      4.117e-02  1.290e-02   3.192  0.00160 ** 
CHOLESTEROL -1.811e-04  6.867e-04  -0.264  0.79220    
RETDIET     -3.417e-05  1.134e-04  -0.301  0.76341    
---
```



```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)   
(Intercept)           2.173e+01  7.211e+00   3.014  0.00292 **
AGE                  -1.065e-02  5.316e-02  -0.200  0.84140   
SEX                  -4.887e+00  3.310e+00  -1.477  0.14139   
SMOKSTAT              6.145e-01  1.355e+00   0.454  0.65065   
QUETELET             -2.020e-01  1.770e-01  -1.142  0.25497   
VITUSE               -5.235e-02  1.079e+00  -0.049  0.96136   
FAT                   1.532e-02  3.988e-02   0.384  0.70134   
FIBER                -2.391e-01  2.011e-01  -1.189  0.23584   
ALCOHOL              -1.086e-01  2.363e-01  -0.460  0.64630   
CHOLESTEROL          -2.528e-02  1.007e-02  -2.510  0.01286 * 
RETDIET               4.766e-03  2.365e-03   2.016  0.04517 * 
AGE:SEX               8.767e-03  2.074e-02   0.423  0.67296   
AGE:SMOKSTAT         -5.524e-04  7.791e-03  -0.071  0.94354   
AGE:QUETELET          1.061e-03  8.292e-04   1.280  0.20203   
AGE:VITUSE           -4.247e-03  7.248e-03  -0.586  0.55860   
AGE:FAT              -4.142e-04  2.437e-04  -1.700  0.09073 . 
AGE:FIBER             7.972e-04  1.098e-03   0.726  0.46883   
AGE:ALCOHOL          -2.285e-04  1.508e-03  -0.152  0.87967   
AGE:CHOLESTEROL       8.291e-05  6.425e-05   1.290  0.19846   
AGE:RETDIET          -2.095e-05  1.451e-05  -1.444  0.15046   
SEX:SMOKSTAT         -4.794e-01  4.726e-01  -1.015  0.31157   
SEX:QUETELET          6.048e-02  7.854e-02   0.770  0.44219   
SEX:VITUSE            3.705e-01  3.801e-01   0.975  0.33090   
SEX:FAT               3.817e-04  1.307e-02   0.029  0.97674   
SEX:FIBER             9.124e-02  7.492e-02   1.218  0.22475   
SEX:ALCOHOL           2.960e-02  5.218e-02   0.567  0.57109   
SEX:CHOLESTEROL       7.395e-03  3.226e-03   2.293  0.02292 * 
SEX:RETDIET          -9.007e-04  6.775e-04  -1.329  0.18524   
SMOKSTAT:QUETELET     1.328e-02  1.985e-02   0.669  0.50408   
SMOKSTAT:VITUSE      -8.689e-02  1.312e-01  -0.662  0.50850   
SMOKSTAT:FAT         -8.677e-03  5.609e-03  -1.547  0.12348   
SMOKSTAT:FIBER        1.553e-02  2.228e-02   0.697  0.48648   
SMOKSTAT:ALCOHOL      2.089e-02  2.570e-02   0.813  0.41718   
SMOKSTAT:CHOLESTEROL  3.382e-03  1.233e-03   2.743  0.00664 **
SMOKSTAT:RETDIET     -3.165e-04  2.823e-04  -1.121  0.26360   
QUETELET:VITUSE      -1.777e-02  1.488e-02  -1.194  0.23375   
QUETELET:FAT          7.566e-04  6.815e-04   1.110  0.26822   
QUETELET:FIBER        1.552e-04  3.054e-03   0.051  0.95953   
QUETELET:ALCOHOL      1.895e-03  4.621e-03   0.410  0.68215   
QUETELET:CHOLESTEROL  9.661e-05  1.547e-04   0.624  0.53314   
QUETELET:RETDIET     -4.066e-05  3.419e-05  -1.189  0.23571   
VITUSE:FAT            5.146e-04  4.860e-03   0.106  0.91577   
VITUSE:FIBER          1.716e-02  2.257e-02   0.760  0.44806   
VITUSE:ALCOHOL        1.223e-02  2.045e-02   0.598  0.55048   
VITUSE:CHOLESTEROL    7.433e-04  1.300e-03   0.572  0.56826   
VITUSE:RETDIET       -3.618e-04  2.588e-04  -1.398  0.16363   
FAT:FIBER            -1.036e-03  8.754e-04  -1.183  0.23823   
FAT:ALCOHOL           6.441e-05  7.998e-04   0.081  0.93589   
FAT:CHOLESTEROL      -4.326e-06  1.980e-05  -0.218  0.82729   
FAT:RETDIET          -2.837e-06  7.785e-06  -0.364  0.71592   
FIBER:ALCOHOL         1.615e-03  4.074e-03   0.396  0.69226   
FIBER:CHOLESTEROL     1.898e-04  2.411e-04   0.787  0.43209   
FIBER:RETDIET         9.815e-06  4.984e-05   0.197  0.84406   
ALCOHOL:CHOLESTEROL   1.987e-04  2.567e-04   0.774  0.43982   
ALCOHOL:RETDIET      -6.448e-05  4.476e-05  -1.440  0.15133   
CHOLESTEROL:RETDIET  -6.087e-07  1.214e-06  -0.501  0.61670   
```

As seen from the normal ANOVA table on the left, we can keep the variables of AGE and ALCOHOL, and this outcome is already being included in the stepwise selection conclusion. Then, we analysis the ANOVA table contained the interaction effects on the right side. As a result, CHOLESTEROL, RETDIET, AGE*FAT, SEX*CHOLESTEROL, SMOKSTAT*CHOLESTEROL are significant (in order to be more conservative, we assign the significant level =0.1). As a common sense, if main effect of A is not significant while its interaction effect is significant, we cannot deny its impact on the dependent variable. Based on the combined conclusion, we can keep the variable AGE, SEX, SMOKSTAT, FAT, ALCOHOL, CHOLESTEROL and RETDIET and run the ANOVA table to determine its regression model.



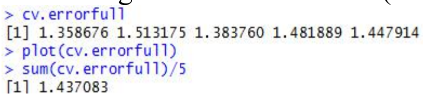
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.135e+01  6.019e-01  18.865  < 2e-16 ***
AGE          1.099e-02  4.512e-03   2.435  0.01558 *  
SEX          6.127e-02  2.180e-01   0.281  0.77890    
SMOKSTAT     7.047e-02  8.794e-02   0.801  0.42372    
FAT         -3.930e-03  2.699e-03  -1.456  0.14666    
ALCOHOL      4.046e-02  1.276e-02   3.171  0.00171 ** 
CHOLESTEROL -5.864e-05  6.811e-04  -0.086  0.93146    
RETDIET     -4.407e-05  1.125e-04  -0.392  0.69554    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.976 on 246 degrees of freedom
Multiple R-squared:  0.09062,	Adjusted R-squared:  0.06475 
F-statistic: 3.502 on 7 and 246 DF,  p-value: 0.001337
```

Hence, the final model is RETDIET=11.35+0.011AGE+0.0613SEX+0.07SMOKSTAT-0.0039FAT+0.0405ALCOHOL-0.00006CHOLESTEROL-0.000044RETDIET.

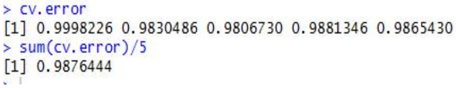
To validate this conclusion, we can use 5-fold cross validation method to compare the full model (after boxcox transformation) and final model.

The average MSE of full model (after boxcox transformation) is:



```
> cv.errorfull
[1] 1.358676 1.513175 1.383760 1.481889 1.447914
> plot(cv.errorfull)
> sum(cv.errorfull)/5
[1] 1.437083
```

The average MSE of final model is:



```
> cv.error
[1] 0.9998226 0.9830486 0.9806730 0.9881346 0.9865430
> sum(cv.error)/5
[1] 0.9876444
```

As a result, the final model have a smaller MSE compared with the full model. Moreover, the adjusted R square of the final model is 0.06475, bigger than the adjusted R square of full model (after boxcox transformation) 0.04406, which means the final model delete some useless terms and can make more accurate prediction.

Conclusion

There is wide variability in plasma concentration of the micronutrients in human, and the variability is tightly associated with person's characteristics and habits. As seen from the final model, we can find out that AGE, SEX, SMOKSTAT, FAT, ALCOHOL,CHOLESTEROL and RETDIET are determine the level of Plasma Retinol Level. However, it is important to note that some information was inconsistent and over reported. For example, male may have a higher smoking rate than female, whether variable QUETELET may have a relationship with calories, fat and cholesterol, males in this database have a higher Plasma Retinol Level (700.7381 ng/ml) than female (579.618 ng/ml). In addition, even the adjusted R square of our final model is bigger than the full model's, but the value is still very small, which means so more detailed checking should be made.

Reference

Plasma Retinol Dataset
staff.pubhealth.ku.dk/~tag/Teaching/share/data/Plasma.html

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