*Daniel Hufnagle ‘23*

*Spring 2023*

*Ethical Data Science Capstone*

Variability of Plasma Retinol Concentration Based on Personal Health

# **Project Description**

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. Retinol deficiency can lead to a range of health problems, including night blindness, increased susceptibility to infections, dry eyes, dry skin, or spots in the eyeball. This data can be used to develop and test statistical models that can predict plasma retinol levels based on other variables, such as age, sex, and various dietary intakes.

Research questions:

* What correlations/relationships can be found between plasma retinol levels and variables within this data set?
* Are there any unusual trends/relationships with these variables?

# **Data and Challenges**

This data set was found on [this website](http://lib.stat.cmu.edu/datasets/Plasma_Retinol). It is titled “Determinants of Plasma Retinol and Beta-Carotene Levels.” Subjects were tested in a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol. Study subjects were 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 data was collected on 315 subjects over 14 variables including: age, sex, smoking status, Quetlet (weight/(height^2)), vitamin use, calorie intake, fat intake, fiber intake, alcohol intake, cholesterol intake, dietary beta-carotene intake, dietary retinol intake, plasma beta-carotene concentration, and plasma retinol concentration. There were no missing values or N/As in my data, the only issue I ran into was formatting the text file before importing into RStudio.

# **Modeling and Analysis Methods**

An exploratory data analysis was carried out for all variables within the study in order to detect some characteristics of the variables or subjects to address the study of the relations in which I was interested. Multiple regression analysis was used to study the relationships between plasma retinol and age, sex, smoking status, vitamin use, calories, fat, fiber, cholesterol, dietary beta-carotene intake, alcohol intake, and Quetlet. Exploratory factor analysis was used to test the importance of each independent variable on the dependent variable plasma retinol. Pearson Correlation was used on numerical data within the set to gauge its relationship to retinol plasma concentrations. Linear discriminant analysis was tested on all independent variables to analyze their influence on plasma retinol concentrations. Density analysis was tested on the relationship between subject sex and retinol plasma concentration.

# **Modeling & Analysis**

The first analysis method used was scatterplot matrix of all numerical variables paired with the Pearson Correlation coefficients of each relationship.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

While there are strong correlations between some variables such as fat and calories (.87) these are not the relationships I am looking for. I am attempting to analyze the relationships of retinol plasma (rplasma) with all other variables. When focusing on the bottom row it is evident that all of the correlations with retinol plasma are very weak. The strongest correlation visible is between retinol and age (.2116). While this number is interesting the likelihood that there is a true correlation between these two variables, while still weak, is low. It is most likely just a skewed number due to outliers. To analyze this possibility, I created a scatterplot of subject age by retinol concentration. Below is this scatterplot which has also been colored to show male and female subjects. There are 42 male subjects and 273 female subjects.

Chart, scatter chart

Description automatically generated

This graph more accurately shows the distributions of the subject ages and the range of retinol concentration. Again, the weak correlation is noticeable as the value of the retinol regression line at 20 years old is about 500 µg/ml while subjects nearly 60 years older have values that have not even increased by 250 µg/ml. This chart shows that there is virtually no correlation between these two variables and the most likely reasoning for any correlation value is probably due to outliers in the data set as evident in the above scatterplot since some older subjects have very high retinol levels. Since there is no significant correlation between age and retinol levels, could there be a relationship between retinol and something which has an impact on other variables such as calorie intake? For example, if subject A has a higher caloric intake than subject B subject A’s fat, cholesterol, and fiber could increase as well. While this is assuming that those three variables depend on calorie intake it is a reasonable assumption. The visualization below depicts retinol concentration by calories consumed per day which is then separated by the 3 smoking statuses recorded: current smoker, former smoker, and never smoked.

A screenshot of a computer

Description automatically generated

This chart shows there is a weak negative correlation between the calorie count and retinol levels of all smokers. This is evident from the correlation matrix in the section above showing a coefficient of -.073 between these 2 variables. Although this correlation is almost insignificant due to its value of -.073 it is still a relatively noticeable difference within the chart above.

Regarding the importance of the effect of each of the independent variables on retinol concentration, there are a few variables that have some influence on the dependent variable. I used an LDA variable importance plot on the data set to visualize what variables have a significant influence on the retinol concentration.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The LDA plot shows how much influence an independent variable has on the dependent variable. The larger the size of the box the more important it is. It also compares an independent variable to other independent variables. For example, the graph shows a connection between Age and the variables of Alcohol, Vitamin use, and Smoking status. This makes sense due to the tendency for drinking, smoking and vitamin use to be more prevalent in older subjects. It is clear that the most significant variable by far is sex in dark blue. Not only does it largely affect the dependent variable but also other independent variables. However, these charts are intended to reach from 0-100 meaning that even though it is significant it still has very little effect.

To further analyze the amount of influence on this variable I ran an ANOVA analysis on the table. This is depicted below. Before doing that, I needed to alter the categorical variables of sex, smoking status, and vitamin use to analyze its relationship more accurately on the dependent variable. I used the code below to change these variables to binary and make a second variable where necessary.

* P.S1 <- ifelse(sex == 1, 1, 0)
* P.SM1 <- ifelse(smoke == 1, 1, 0)
* P.SM2 <- ifelse(smoke == 2, 1, 0)
* P.V1 <- ifelse(VitUse == 1, 1, 0)
* P.V2 <- ifelse(VitUse == 2, 1, 0)

This code allows for me to break up the categorical columns as such:

* if sex = 1 (male) it stays as 1, if not then it becomes 0 (now female)
* if smoke status = 1 (never) it becomes a 1 and the 2 others (former and current) become 0
* if smoke status = 2 (former) it becomes a 1 and the same happens to the other 2 as above
* the same happens to the vitamin use as smoke status 1 = often, 2 = not often, 3 = no

A screenshot of a computer

Description automatically generated

The p-value in the ANOVA table shows that there is a significant correlation between subject sex and is statistically significant in predicting the value of retinol levels with a p-value of .000426. To look further into this, I created a density plot of retinol levels based on subject sex depicted below.

A screen shot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The density plot shows that females have a lower distribution of retinol levels compared to males. This leads to questions such as; could there be differences in the male and female subjects’ biology and dietary intakes which leads to this happening? The numbers below the plot show the averages of female and male numerical columns. These averages show that the male subjects have a higher calorie, fat, fiber, alcohol, and cholesterol intake compared to the female subjects. When analyzing the categorical data, I found that 52% of women never smoked and 39% regularly took vitamins. As for the men, 52% were former smokers and 54% never took vitamins.

# **Insights & Results**

My data analysis has brought me to the conclusion that there is a significant variability in retinol plasma concentration in this data set. Although there are some variables that have correlations and relationships with the dependent variable, such as sex, the variability is too wide, and the subjects are unequally split between male and female which could very easily skew the data set. There are over 6x as many females (273) as males (42) in the data set. While the sex variable was identified as being a significant predictor of retinol concentration, the correlation between the two was still very weak. The sampled males had a higher distribution of retinol levels compared to those of female. While this may be unique it is also possible this is evident due to the uneven number of samples. A better understanding of this subject would require further study with more equal levels of male and female subject to test.