Final Project Data Analysis Rough Draft

William Ranick, Nibedita Bal, Hima Spandana Barla

Data: National Health and Nutrition Examination Survey (NHANES)

**Research Questions**

1.Does Systolic Blood pressure explain BMI?

2.The intake of which vitamins and organic compounds reduce the risk of cancer diagnosis?

3.How well can diabetes diagnosis be predicted from the diet of the patients?

**Hypotheses**

1.Does BMI affect blood pressure?

2.Some nutrients may decrease the risk of cancer diagnosis while increased alcohol and total caloric intake will increase the risk.

3.External NIH dietary supplements such as vitamins , minerals, herbs, or spices may help control diabetes.1 If you can't get enough vitamins and minerals from foods, you will need supplements.

**What variables will you be using to answer your research questions?**

1.For knowing the relationship between Systolic blood pressure and BMI,below are the variables used are BPXSY1: Systolic: Blood pressure (first reading) mm Hg - Dependent Variable, BMDAVSAD: Average Sagittal Abdominal Diameter (cm), BMXHT:Standing Height, BMXBMI: Body Mass Index (kg/m\*\*2), BMXWAIST:Waist Circumference (cm), BMXWT: Weight (kg).

2.The variables used for the nutrition and cancer analyses were general cancer diagnosis, age of diagnosis for various gastrointestinal cancers, and 64 quantitative variables which tracked the amount of different vitamins and organic compounds during only the day that data was collected.

3.Some of the variables which play a main role in answering diabetes diagnosis question are

DR1TP182 PFA 18:2 (Octadecadienoic) (gm), DR1TM161 MFA 16:1 (Hexadecenoic) (gm), DR1TP205 PFA 20:5 (Eicosapentaenoic) (gm), DR1TVARA- Vitamin A as retinol activity equivalents (mcg).There are certain vitamins and minerals that could be useful for your disease, depending on the type of medication regimen you use to manage your diabetes.

**Statistical methods used**

1.For Systolic blood pressure and BMI, exploratory data analysis is done with data cleaning, data preprocessing. Some statistical methods are used such as normalization, selecting significant variables by stepwise regression, multicollinearity checked by VIF and visualize the relationship among significant variables.

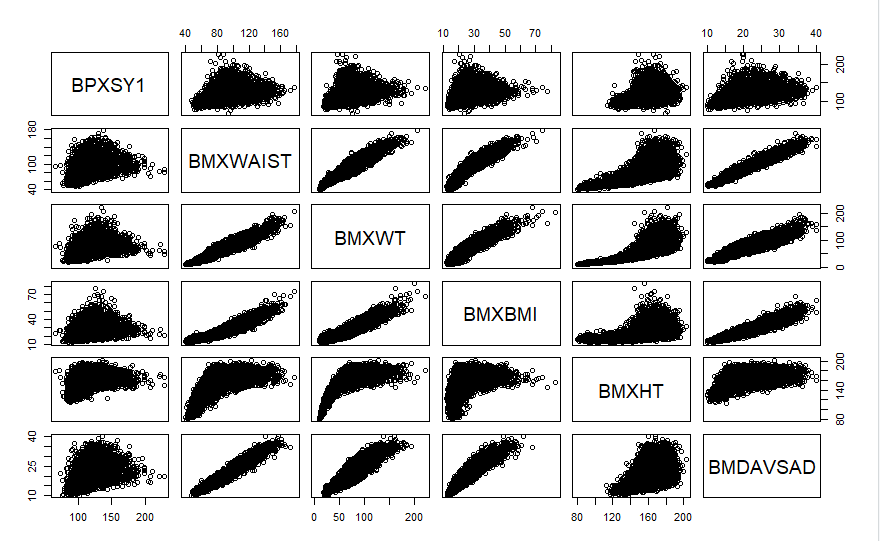
2,Exploratory analysis was initially performed to observe if there were any correlations between what nutrients were consumed and the age at which any gastrointestinal cancer was diagnosed.

A logistic regression was performed to measure if the intake of any of the listed organic compounds increased or decreased the rate of cancer diagnosis. A model was created from this regression and this model’s accuracy was validated with test data.

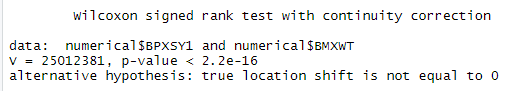
3.The methods which is used to significantly reduce the diabetes levels of the patients by controlling their diet or taking vitamin supplements along with food is multiple regression.

**Results in terms of statistics and application**

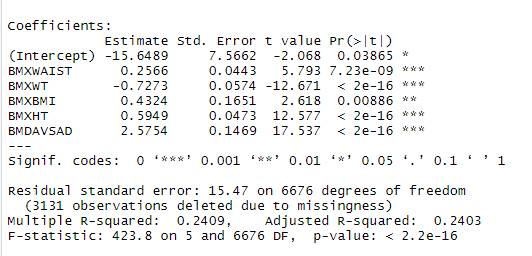
1.Collinearity among numerical variables is checked and it is shown that most of the variables are positive and linearly correlated;however, the vif values are greater than 5,so the variables are highly correlated.



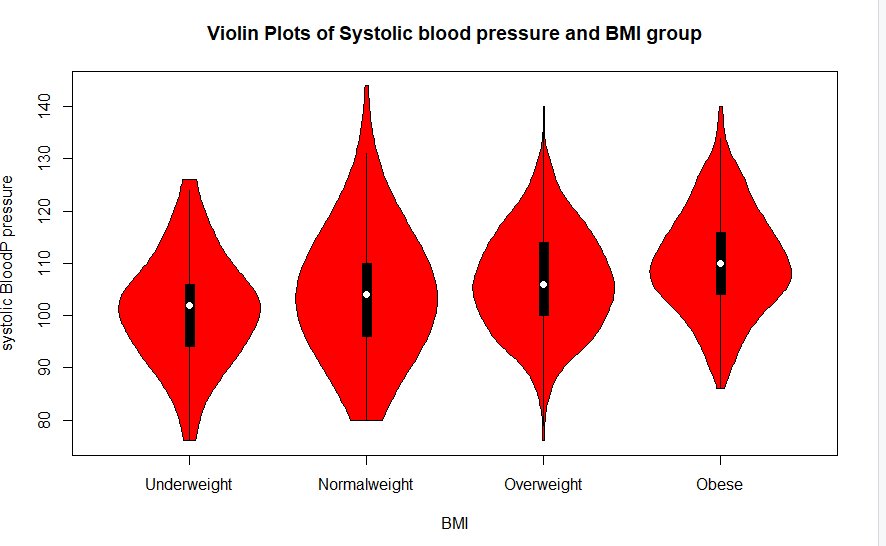
Normality of Systolic blood pressure and other independent variables are checked by Shapiro-Wilk test. The P value is less than .05 which means the data is not normal which means it cant reject the null hypothesis. As the data is not normal and the sample size is big , wilcox test was chosen for this variables.



After removing highly correlated variables, we build a linear regression model. Below are the findings we have gotten. It is shown that Waist Circumference,Weight (kg), Height,Sagittal Abdominal Diameter, and Body Mass Index are highly significant which means these variables impact on systolic blood pressure. From this result, it is assumed that if a person has larger belly, higher BMI which means the person is overweight or obese , he is more prone to systolic blood pressure.The P value is significant which means the null hypothesis is rejected and alternative hypothesis is accepted which tells BMDAVSAD,BMXHT, BMXWAIST, BMXWT and BMXBMI have surely impact on blood pressure.

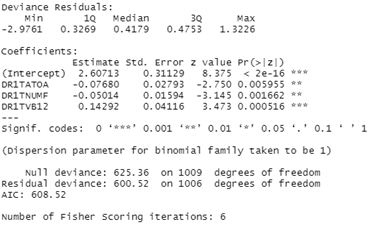


Below is a visualization among blood pressure and age group which tells that blood pressure increases as the BMI increases. It is clearly visible that obese plot has higher median and skewd towards high that means obese people suffer systolic blood pressure.



**2.**Various plots (not shown) were created during exploratory analysis to observe if there were any trends between organic compounds consumed and the age at which patients were diagnosed with any type of gastrointestinal cancer. This analysis was limited by the very few participants who had any of these cancers (27 out of over 10,000) and no such correlation was discovered.

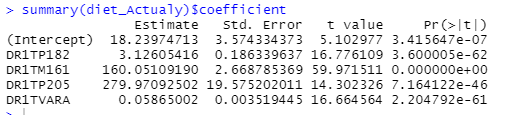
Possible correlations were discovered from the logistic regression performed to predict cancer diagnosis from consumed organic compounds. The most significant of these findings was that consuming 1 unit (1 mcg) of vitamin B-12 decreased the log odds of a cancer diagnosis by approximately 0.143, displaying a residual deviance of 600.52 on 1006 degrees of freedom.



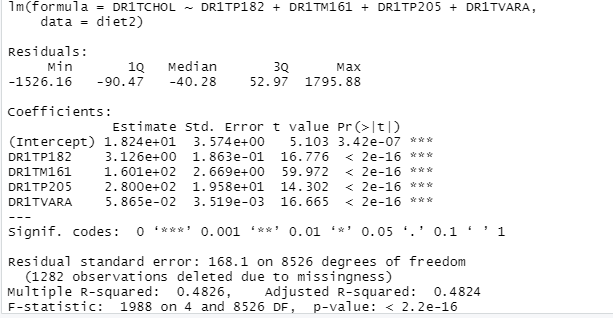
The model that yielded these results was validated with test data. Based on the few variables that were included in the model, the trained model accurately predicted only 0.03% of cancer diagnoses.



3.



For a given predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero. It can be seen that changes in DR1TP82 and DR1TM161, DR1TP205,DR1TVARA are significantly associated with changes in cholesterol. The multiple r square and the adjusted r square are almost same that is 48% and the p value is <2.2e-16



Performed multiple regression to figure out how the vitamins used by the patients in their diet has helped them to reduce the cholesterol levels in their body.Hence we can reject the null hypothesis and accept the alternative.

**Limitations of the analysis**

For the research question, Systolic blood pressure and BMI, the dataset found from the kaggle is a survey dataset which has overall 2 years of data but the dataset is incomplete which means the data covers only one or two conditions which we are trying to model-as when a model is built for knowing the relationship between BP and BMI, we need a long term data but here the small duration data might not give us accurate model. In addition to this, the data attributes collected are very ambiguous which takes a lot of time to understand and analyze. Also, there are many outliers that needed to be removed.

The primary limitation of the diet/cancer analysis was the available diet data and how it was reported. The online dictionary of all variables include many variables that were not actually present in the data set, such as measures of the amount of supplements taken monthly. The only data available to use for this analysis were measurements of how much of these compounds were consumed for only 1 day of study participation. This single day of dietary intake is very unlikely to be representative of the patient’s overall diet. Additionally, the data lacked sufficient numbers of participants with gastrointestinal cancers. It is possible that these types of cancers are the most likely to be affected by a patient’s diet.

The dataset chosen for diabetes diagnosis, the main limitation was cleaning and removing the outliers.The way National Health and Nutrition Examination Survey (NHANES) collected the data was not good enough to build the models. Dietary day one and two samples were only used in the study to examine the cholesterol levels

**Based upon the limitations, what would we do differently in future work**

1.In the future, we would choose the right dataset where every parameter is relevant for our analysis. In addition, we will probably select some research data rather than survey data that don’t include accurate information.

**2.**Based on the limitations for the diet/cancer analysis, future work would include the incorporation of different variables to represent the patient’s overall diet. There are some variables within the dictionary for the data set, such as the amount of different types of fish eaten for one month, which would likely be more representative of the patient’s diet. These variables, however, were absent from the data set itself. In the future, it would be helpful to find a version of the data that includes all of these variables, if not more. Additionally, the inclusion of more patients would be helpful in creating more accurate models, particularly when considering the small percentage of participants that had been diagnosed with different cancers.

**3.**In the next proposal, we would like to add the variables which would lead to the increase in the cholesterol levels in the patients.In the given dataset caffeine and alcohol numbers are given which make a big impact on the cholesterol levels and help us determine the impact on the patient on his daily routine.