On The Association Between Dietary Factors and Mortality Due to Stroke and Diabetes:

A Retrospective Examination of Western and Asian High-Income Countries.

By

Faaiza Castellanos, B.A. (Economics), University of California at Berkeley, 2000 and
Jill DeWitt, M.S (Secondary Math Education), Grand Valley State University, 2006

Advisors:
Dr. Jossy Uvah
Dr. Achraf Cohen

A Graduate Capstone Project
In Partial Fulfillment of the Degree of
Master of Science in Mathematical Sciences
University of West Florida

Contents

List of Figures

List of Tables

Abstract

A Retrospective Look at the Association Between Dietary Factors and Mortality due to Stroke and Diabetes in Western and Asian High-Income Countries

The purpose of this research paper was to explore the relationship between diet, specifically the consumption of fruit, non-starchy vegetables, bean and legumes, nuts and seeds, unprocessed red meats, sugar-sweetened beverages, fruit juices, milk, protein, calcium, and protein, and rate of stroke and diabetes in Western, high-income countries. Our study took a retrospective approach by looking at data from three years-1990, 2000, and 2010. We used Principal Component Analysis and Multiple Linear Regression to determine which foods or nutrients had a significant effect on stroke or diabetes rates. Gender and year were also considered as factors in the study. The study found that in general, only a few nutrients had a positive or negative impact on disease rates.

1 Introduction

1.1 Statement of Problem

Diet and how it pertains to the development and progression of chronic disease is a growing topic of interest. The increase in prevalence of deadly diseases, such as diabetes and stroke, has led many to examine their diet in a more profound manner, searching for clues that can prevent, halt the progression of, or even reverse such conditions. In response to the escalating interest in diet and disease, many doctors and self-help gurus are promoting a variety of diets, such as keto, paleo, low-fat, low-carb, and many more. Many ancient civilizations have also promoted a variety of foods based on culture and environment. It seems that nutritional intake and its impact on diseases has been a major concern for most people for thousands of years.

With the variety of recommendations and guidelines circulating, the question remains-What foods or nutrients are linked with some of the main killers in our society, namely diabetes and stroke? In order to study diet, it is important to examine a variety of foods and food groups, rather than just focusing on one or a few. In this paper, we investigate a variety of nutrients intake-fruit, non-starchy vegetables, bean and legumes, nuts and seeds, unprocessed red meats, sugar-sweetened beverages, fruit juices, milk, protein, calcium, and protein, and their link to stroke and diabetes/kidney disease. In addition, we restrict our study to Western and Asian high-income countries in an effort to control factors such as high poverty and low healthcare that contribute to lack of resources and education to buy and choose healthy foods.

1.2 Relevance of Problem

The study of food and its link to diabetes and stroke is of paramount importance to the World Health Organization (WHO) [2]. Every year societies spend a large amount of money on healthcare in order to treat diabetes and stroke. Unfortunately, many individuals also die from these diseases, leading to familial emotional turmoil as well as loss of economic contribution from the deceased individual. If such diseases can be prevented, reversed, and even treated with diet, individuals could decrease healthcare cost, benefit the economy, and be socially and emotionally healthier. Diet is a daily part of everyone's lives, so modifications to diet are available to almost all at a much lower price than medication or surgery.

1.3 Literature Review

The importance of diet in preventing death due to diabetes is a major concern of the World Health Organization (WHO) [2], with valid concern. According to the WHO, the number of people with diabetes rose from 108 million in 1980 to 422 million in 2014 and between 2000 and 2016, there was a 5% increase in premature mortality from diabetes. Diabetes is a disease that then can lead to a myriad of other deadly diseases, such as kidney failure, heart attacks, and stroke. Overall, the WHO estimates that diabetes was the seventh leading cause of death in 2016. However, they do stress the importance of a healthy diet in order to prevent and even reverse diabetes, recommending avoiding sugar and saturated fats. The American

Diabetes Association [3] recommends to fill half your plate with non-starchy vegetables and to also include fruits, lean meats and plant-based sources of protein, less added sugar, and less processed foods.

Strokes are another major cause of deaths worldwide, and the WHO [8] estimates that 15 million people suffer from a stroke worldwide annually, of which 5 million die and another 5 million remain permanently disabled. The major cause of stroke is high blood pressure, which is often associated with diet. The Centers for Disease Control and Prevention (CDC) [7] states that up to 80% of strokes can be prevented through healthy lifestyle changes, and diet is a key component of those changes. According to the CDC, eating plenty of fresh fruits and vegetables, foods low in saturated fats, trans fat, and cholesterol, and foods high in fiber can prevent high cholesterol, which in turn lowers your chances of suffering a stroke. Also, limiting sodium can lower blood pressure, which as indicated before, increases the likelihood of a stroke.

The WHO in Europe [1] provides even further research on the connection between diet and disease and offers recommendations for prevention. The importance of preventing disease with diet is a major issue of study and increasing prevalence of these diseases places a huge stress on the healthcare system, impacting national economies and health service budgets negatively. With correct knowledge and programs to education and support populations in implementing healthy dietary guidelines, many countries can increase longevity, improve their economies, de-stress their healthcare systems, and improve mental health. [15]

To study our data in terms of descriptive statistics, we used Principal Component Analysis, or PCA. PCA [9] is a technique for feature extraction, meaning that it combines our variables in a way that we can omit the least influential variables while retaining their most valuable parts. There are three key assumptions that are behind PCA [16] and can indicate when PCA would not provide a strong analysis. The assumptions are linearity, large variances have important structure, and the principal components are orthogonal. The linearity assumption organizes the analysis as a change of basis. As for variances, principal components with larger variances are indicative of structure, while those with smaller variances just represent noise, which is irrelevant to the strength of the signal being analyzed. The assumption that principal components are orthogonal is suggestive that PCA can be solved using linear algebra decomposition techniques.

We then used Multiple Linear Regression with a Box-Cox transformation to model the data. Halinski and Feldt [11] provide a framework for choosing the best procedure to pursue while keeping two goals in mind. First, the best model should produce an equation that yields the best predictions for the population. Second, the best model should contain an optimal number of explanatory variables. We began our model-building using a general linear model using a Gamma, then Poisson distribution, but ultimately settled on a normal multiple linear regression model with a transformation of the response variable. In general, the multiple regression models in this study strive to balance accuracy and parsimony, that is, the models should accurately describe both the systematic and random components and be as simple as possible. There are four basic assumptions that must be met to use multiple linear regression analysis [12]:

1. There must be a linear relationship between the response variable and the explanatory variables.

2. The model error term (referred to as the model "residuals") must be normally distributed. Residuals can be thought of as the information not explained by the model. A residual plot and QQ plot can be used to determine if this assumption is met.

3. There should not be any multicollinearity. This means that explanatory variables are not correlated with each other. This assumption can be tested by examining the Variance Inflation Factor.

4. Homoscedasticity–This means that the variance of the error terms are consistent across the explanatory variables. A Scale-Location plot can be used to determine if this assumption is met.

The regression analysis in this study addresses each of these assumptions for each proposed multiple linear regression model. Nested models were compared using the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Our goal was to consider these two measures together and to choose models that were were favored by both criterion, as argued by Kuha [13].

1.4 Limitations

There were many limitations of the various methods of data collection and analysis. We collected data from the Global Dietary Database, or GDD, and they state that they collect data on dietary habits via surveys, which then are based on the volunteer's bias. Some volunteers might not remember their exact diet, others might overstate or understate their food consumption, and yet others might just not wish to disclose honest information. Of course, the GDD uses a complex food description and classification system to address the issue of variation of description of diet, but this does not address the issue of volunteer dishonesty or lack of memory recall.

In addition, the data collected was based on observational studies, where the volunteers were not affected in any manner and were just questioned on their diet. The data that was studied could easily indicate correlation, but not necessarily causation. A more promising study would be an experimental study on diet, where volunteers would be divided into an experimental and control group. Each nutrient could be then studied separately to see how it affects development of diabetes and stroke. Of course, such a study would be very time-consuming, as human lifespan can last upwards of 100 years. In addition, to study a specific nutrient in such a manner would require that the experimental group consume that nutrient consistently for years on end. Certain foods being studied, such as sugar-sweetened beverages, are commonly considered to be unhealthy, and to require a group of volunteers to consume the beverages consistently for years and years would be unethical.

The data collected was international, indicating a large variety of genetics, spices, culinary techniques, exercise, and social behavior that play a part in disease. Genetics are a factor that can affect someone's chance of developing diabetes or a stroke, as well as influence how sensitive they are to certain food items. The use of spices may also affect how foods react in the body and lead to the development of disease. Culinary techniques, such as deep frying, grilling,

boiling, and sauteing, can also alter foods and contribute to diabetes and strokes. Other than dietary factors, exercise is a main agent in health, affecting diet metabolism and also disease. Lastly, social behavior can lead to change in emotion and eating habits which can greatly affect disease. There is much research on emotions, neurotransmitters, and disease, and even more to study.

2 Dietary Factors and Mortality

2.1 The Data Set

We obtained the data regarding nutrition from The Global Dietary Database (GDD) [4], which is a project of the Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy at Tufts University. The program is also supported by the Bill and Melinda Gates Foundation, which is actively engaged in programs to support public policy. The goal of the program is to understand and improve diet through data collection, analysis, and recommendation, thus leading to public policies that aim to prevent disease and improve healthcare. GDD collects its data from [6] items, there is a large variation on their descriptions. In order to address this issue, the GDD applied FoodEx2, which is a complex food description and classification system developed by the European Food Safety Authority. This system allows GDD to standardize the global dietary intake, thus leading to data which is more reliable.

We obtained the data regarding disease from the Global Health Data Exchange (GHDx) [5]. The GHDx is a data catalog created and supported by an independent global health research center at the University of Washington. The population figures used to calculate the disease rates are estimated based on World Population Prospects: 2015 Revision, from the United Nations Population Division and disease mortality figures are obtained from the WHO Human Mortality Database.

We are most interested in discovering which nutrients had a positive or negative impact on disease rates and whether gender or year was a significant factor. The data set consisted of the following variables:

- **Country**-Each row contained data from one of twenty-eight countries. All countries were Western or Asian high-income countries. A full list of the countries can be found in Appendix A.
- **Gender** was a coded continuous factor where Female = (1) and Male = (2)
- **Nutrient Intake**- Eleven nutrients were included in the data set- fruit, non-starchy vegetables, bean and legumes, nuts and seeds, unprocessed red meats, sugar-sweetened beverages, fruit juices, milk, protein, calcium, and protein. All measurements were in average grams per day, except calcium and potassium, which were measured in average milligrams per day.
- **Disease Rates**- Rates for stroke and diabetes for each country were measured as counts per 100,000 people.

- **Year**- Data from each country was included for three different years, 1990, 2000, and 2010. The year was a coded continuous factor where 1990= (1), 2000= (2), and 2010= (3).

In summary, each of the 28 countries has 6 rows of data, a row for each gender and for each of the three years. We gathered the data from two online sources. The Global Dietary Database [4] provided the nutrient data, and the Global Health Data Exchange [5] provided the disease rate data. We then used the statistical computing and programming language R to read the data, and for all statistical graphics and regression analysis. The R code can be viewed in Appendix B.

2.2 Principle Components Analysis

In an attempt to factor a variety of aspects of diet into our study, we obtained data for 11 nutrients. This is a substantial amount of data, so we utilized Principal Component Analysis (PCA) [16] to identify which variables impact disease the most. We utilized PCA as a purely descriptive statistics method, thus we used it to identify key foods and nutrients for the years 1990, 2000, and 2010, rather than to omit variables in our analysis. Biplots were created, using R, to help visualize the principle components and to see if any clusters of countries are revealed.

We first examine the PCA biplot for 1990. A PCA biplot [14] shows the PCA score plot and the loading plot, where the PCA score plot displays the PCA scores and the loading plot portrays how strongly each of the variables impacts a principal component. The PCA Biplots for all three years are shown at the end of this section (Figures 1-3).

As seen in Figure 1, the PCA score plots are the individual countries where the data was gathered from and the PCA loading plot vectors are the food variables. Examining the PCA scores, clusters of countries represent countries that impact the two PCAs in a similar manner, meaning they exhibit similar characteristics. The country represented by 43 and 44 is the Republic of Korea (male and female), so this country impacts PCA 1 strongly but PCA 2 negligibly. The countries represented by 15, 16, 23, and 24 are Finland and Iceland. These countries impact both PCA 1 and PCA 2. Both of these countries are high-income European nations, and they have similarities in their food intake. Examining the loading plot, vectors that are close to each other and have small angle between them are positively correlated. Non-starchy vegetables and beans and legumes are positively correlated, and so are sugar-sweetened beverages and fruit juice. If vectors meet at a 90°angle, they are not correlated, such as nuts and seeds and milk, as well as fruits juice and non-starchy vegetables. Lastly, if two vectors diverge at a large angle, such as 180°, they are negatively correlated. Non-starchy vegetables, beans and legumes, and fruit are negatively correlated with milk.

The PCA biplot for 2000 (Figure 2) shows us similar information. Again, the Republic of Korea was a cluster by itself, along with a cluster shown for Finland and Iceland. Non-starchy vegetables and beans and legumes are again positively correlated, and so are sugar-sweetened beverages and fruit juice. Nuts and seeds and milk, as well as fruits and non-starchy vegetables are not correlated. Again, non-starchy vegetables, beans and legumes, and fruit are negatively correlated with milk.

Lastly, the PCA biplot for 2010 (Figure 3) shows similarities to the plots for 1990 and 2000 but does vary. The country represented by 35 and 36, the Netherlands, is a cluster by itself. The countries 43, 44,and 45 are also clustered, referring to the Republic of Korea and Singapore. As for the loading plot, non-starchy vegetables and beans, as well as sugar-sweetened beverages, fruit juice, and nuts and seeds are positively correlated. Milk, protein, and potassium are also positively correlated. Protein and fruit juice exhibit no correlated, as well as fruit juice and non-starchy vegetables and beans. Protein and milk are negatively correlated with beans and non-starchy vegetables. The PCA biplots provide us invaluable information on the relationships between the variables being studied, specifically the countries and the food intake.

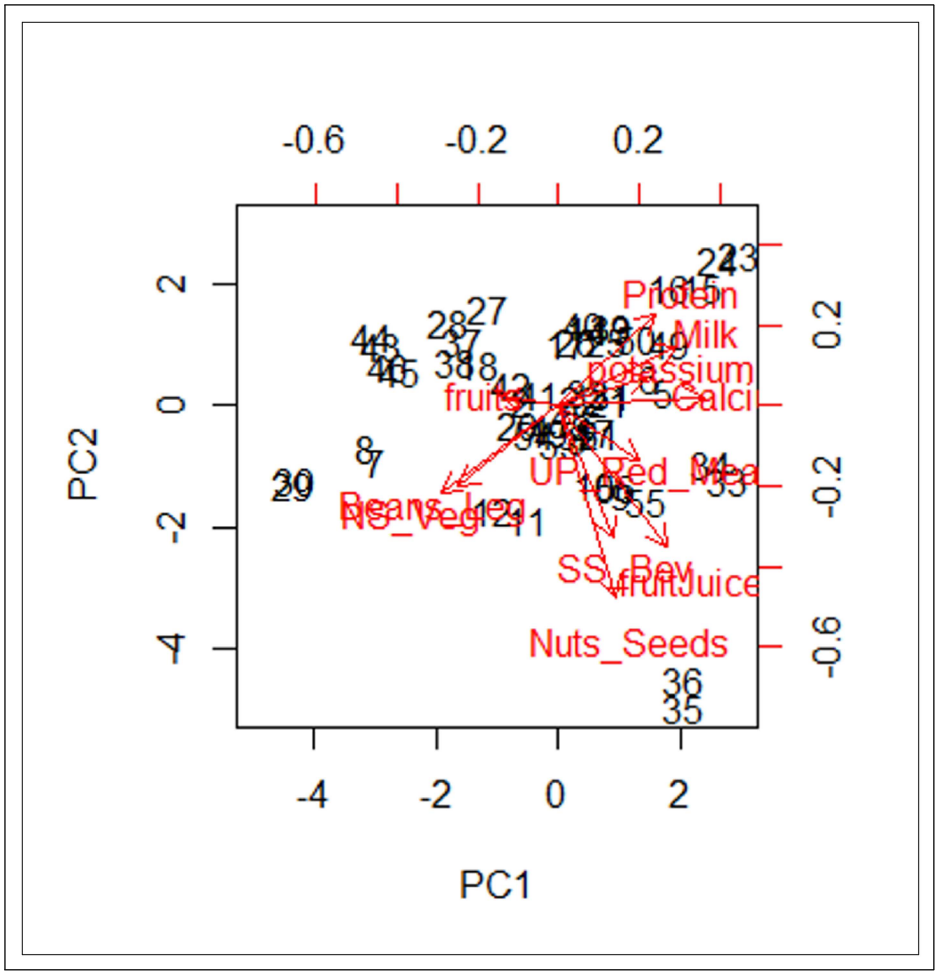

Figure 1: PCA Biplot for Year 1990.

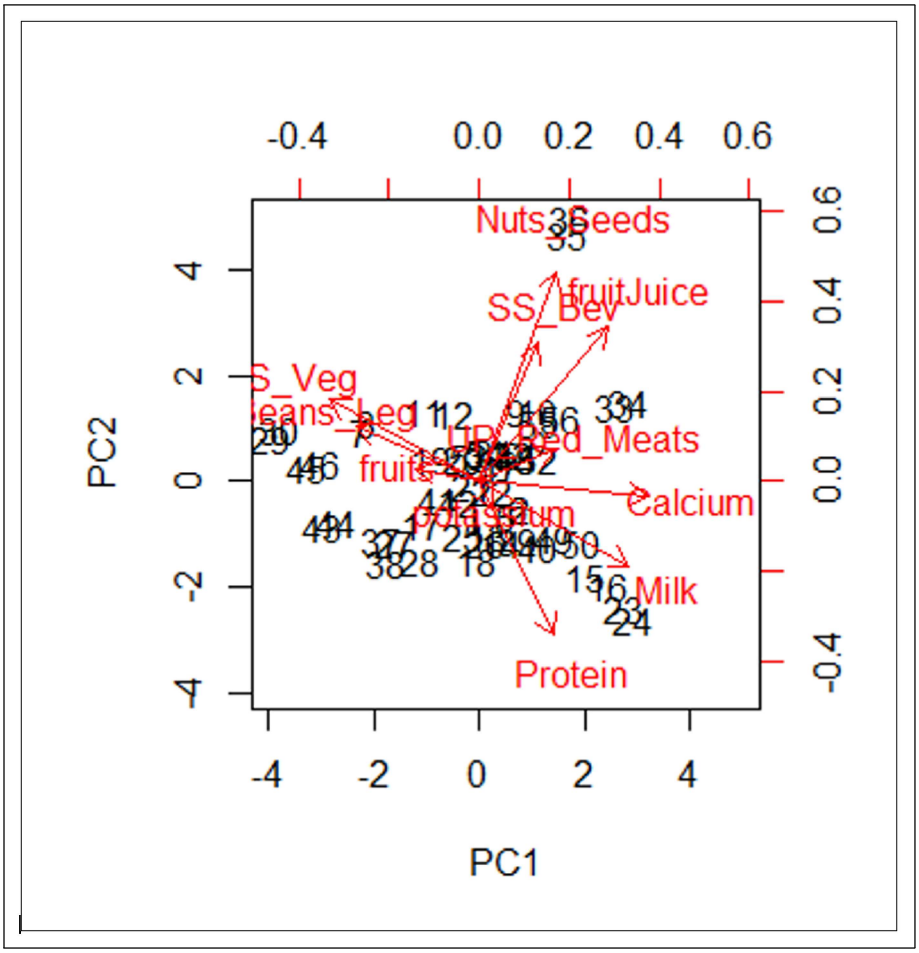

Figure 2: PCA Biplot for Year 2000.

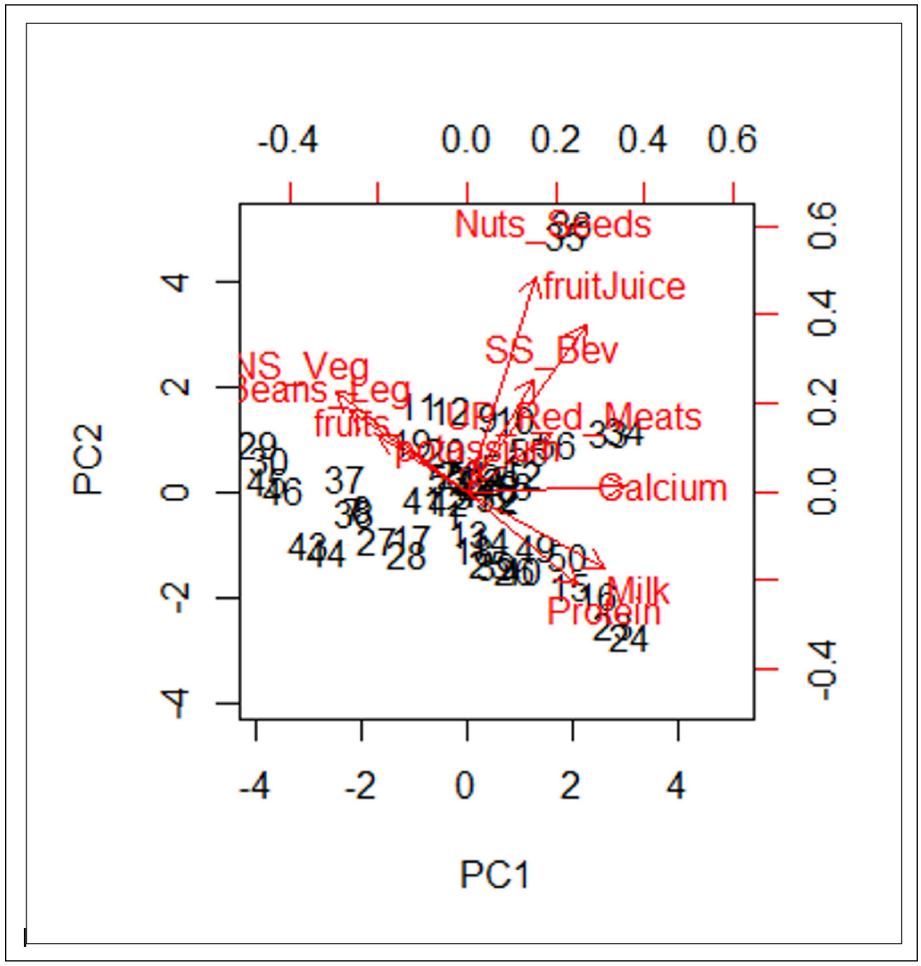

Figure 3: PCA Biplot for Year 2010.

2.3 Methodology: Justification, Assumptions and Development of the Model

The regression analysis focused on two response variables–Stroke rate and diabetes rate. Each response variable is recorded as a count per 100,000. Each model included thirteen potential explanatory variables–eleven nutrients, gender, and year. Several linearized and general linearized models were investigated, analyzed and compared to determine which model had the best fit. Since the response data was count data, two general linearized models were considered. A general linearized model is linear in its parameters. There are three components of a general linearized model. The systematic component is of the form

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_3 x_{pi} \tag{1}$$

The random component follows one of the distributions in the family of Exponential Dispersion Models (EDMs). The final component is the link function component, g, which links the mean, μ, to the linear predictor (1), such that

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_3 x_{pi}$$

The general linearized model assumes that the response comes from an EDM. Initially, for this study two EDM's were considered–Poisson and Gamma. The Poisson distribution [10] was considered because the data was recorded as counts (per 100,000), which could be considered a rate, but since population data was not available to be used as a meaningful offset it was determined that the Poisson distribution was not appropriate. Next, the Gamma distribution was considered. The Gamma distribution can be considered when the response data is positive and continuous which typically means that the data is skewed to the right. Although, the responses in the data under study was positive and skewed right, it was determined that a transformation of the response did a fairly good job of normalizing the response variable. Ultimately, a general linearized model using the Normal distribution as the EDM was used. This can be justified when the response comes from a normal distribution or a transformation of the response makes it approximately normal. A multiple linear regression model is a special case of the general linearized model. Multiple linear regression models with a response variable y, and p explanatory variables, $x_1, x_2, ...x_p$, consist of two components–a systematic component and a random component. The systematic component is the expected value of the response variable which is linearly related to the explanatory variables x_j such as:

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \tag{2}$$

In equation (2), $\mu_i = E[y_i]$ is the expected value of the response variable and the intercept is β_0. The regression parameters, β_j, x_{ji} represent each of the p explanatory variables. The response variable, y_i, is the random component in the model. it is assumed to have a constant

variance, σ^2, and to be normally distributed, that is, $y_i \sim N(\mu_i, \sigma^2)$. It should be noted that linear models are a special case of general linearized models. The goal of the statistical model is to mathematically represent the most important systematic and random components of the data. With a good model, one can understand how variables are related to each other and how the explanatory variables significantly affect the response variable. All model assumptions where checked and analyized using R. The "Residuals vs. Fitted" plot is used to determine constant variance. Ideally, the plot should show a random pattern around the horizontal line at zero. The "Normal Q-Q plot" can be used to show whether the random component is normally distributed. The "Residuals vs. Leverage" plot can be used to determine if there are outliers or influential observations in the data. The R code for all statistical analysis is located in Appendix B.

Model appropriateness was determined using an ANOVA F-test. This is important because it should be determined whether the explanatory variables are useful predictors of the response variable. This can be determined by testing whether the regression sum of squares (SSReg) is larger than the residual sum of squares (RSS). The ANOVA F-test produces the following results:

- F Statistic:

$$F = \frac{SSReg/(p)}{RSS/(n-p-1)} = \frac{MSReg}{MS_E} \sim F_{(p,n-p-1)}$$

- The Coefficient of Determination, R^2. This is the proportion of the total variation explained by regression:

$$R^2 = 1 - \frac{RSS}{SS_T}T$$

https://www.overleaf.com/project/60541973b5a9aa428084ba62

- Adjusted R^2. This is the proportion of the total variation explained by the regression adjusted for the number of explanatory variables:

$$R^2_{adjusted} = 1 - \frac{RSS/(n-p-1)}{SS_T/(n-1)}$$

In the spirit of parsimony, several nested models were considered. Also, a transformations of the response variables was considered. The goal was to get the simplest models with the best predictive and interpretive value. To compare nested models, Akaike's Information Criterion (AIC) was calculated:

$$AIC = nlog(RSS/n) + 2(p+1) \tag{3}$$

Smaller AIC values (closer to $-\infty$) represent better models. In equation (3), the term, $2(p+1)$, is called the penalty. The Bayesian Information Criterion (BIC) was also calculated:

$$BIC = nlog(RSS/n) + log(n)(p+1) \tag{4}$$

The BIC is inclined to select more parsimonious models than AIC. Smaller BIC values (closer to $-\infty$) represent better models. In equation (4), the term, $log(n)(p+1)$, is called the penalty.

If the assumption of normality of the random variable y_i is not satisfied, a transformation of y_i can be considered. Common transformations include a logarithmic or square root transformation. Another useful method is the Box-Cox transformation. The Box-Cox transformation is used to determine the best lambda, λ, that should be used to transform the response variable, y, where the transformed model becomes $y^\lambda = \beta x + \epsilon$. Often, a Box-Cox transformation can improve linearity and homoscedasticity so it can be a very useful tool. The Box-Cox transformation uses a maximum likelihood estimator for λ. If λ is equal to 1, then no transformation is needed. If λ is 0, then a *log* transformation should be used, that is $y^\lambda = log(y)$. If $\lambda \neq 0$, then

$$y^\lambda = \frac{y^\lambda - 1}{\lambda}.$$

Lambda should be relatively small, usually between -3 and 3. R was used to determine the best lambda for the Box-Cox transformation of the data.

2.4 Testing and Analysis

The data was regressed using the thirteen explanatory variables. All model assumptions were checked and addressed if necessary. After an acceptable model was produced, the significant explanatory variables were kept in the model and model adequacy and diagnostics were reevaluated. Ultimately, two models were developed for the response variables stroke rate and diabetes rate. All of the details for model development follow, starting with the model for stroke rate.

2.4.1 Model Formulation for Stroke Rate

Let's begin by examining some descriptive statistics for the response variable Stroke rate. From Table 1, we can see that the mean is greater than the median, so it appears that this data is skewed to the right. The histogram of the variable stroke rate also shows that the data is right skewed (see Figure 4). In order to satisfy the linearity assumption, a transformation of

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
16.88	84.32	123.44	163.55	249.60	451.71

Table 1: Summary Statistics for Stroke Rate

the response variable stroke rate was necessary. The Box-Cox method was used to determine the best transformation. R was used to determine the best lambda, λ, for the Box-Cox transformation. Figure 5 shows the 95% confidence interval for the maximum likelihood estimator, λ.

Using the *boxcox* function in R, the best lambda was calculated to be $\lambda = \frac{10}{99}$. Note that zero is withing the 95% confidence interval for Lambda in Figure 5. Since zero is within the 95%

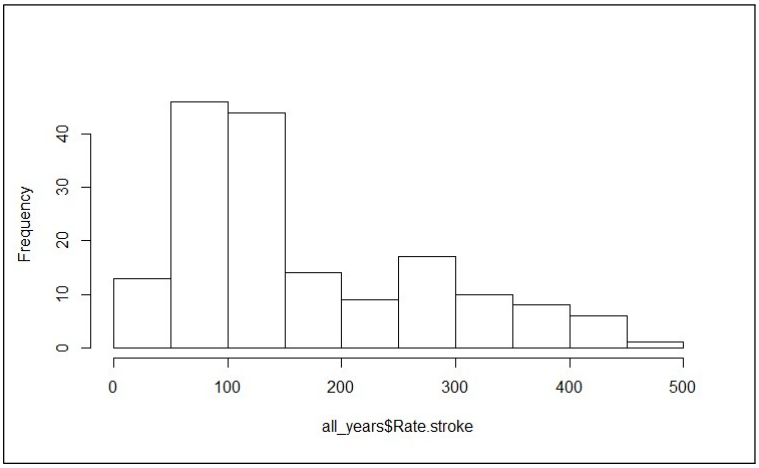

Figure 4: Histogram for Stroke rate

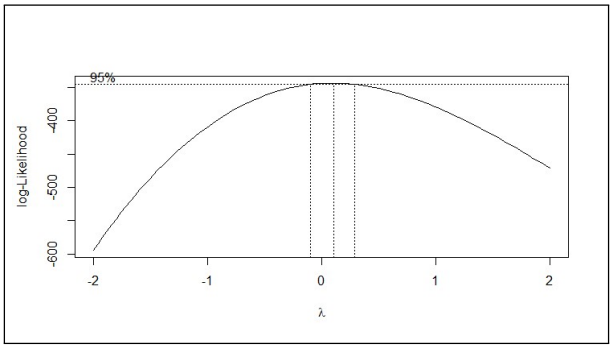

Figure 5: 95% Confidence Interval for Lamda, Stroke Rate.

confindence interval, A *log* transformation was used. The model diagnostics for the *log* model are shown on the next page in Figure 6. A visual analysis of the diagnostic plot in Figure 6 demonstrates that there are no major violations of the model assumptions for the transformed model. The *residuals vs fitted plot* shows a generally random pattern, the *normal QQ plot* shows some slight deviations from normality near the tails of the data, the *scale-location* plot shows that the data has homoscedasticity, and the *residuals vs. leverage plot* show that there are no outliers or influential data points.

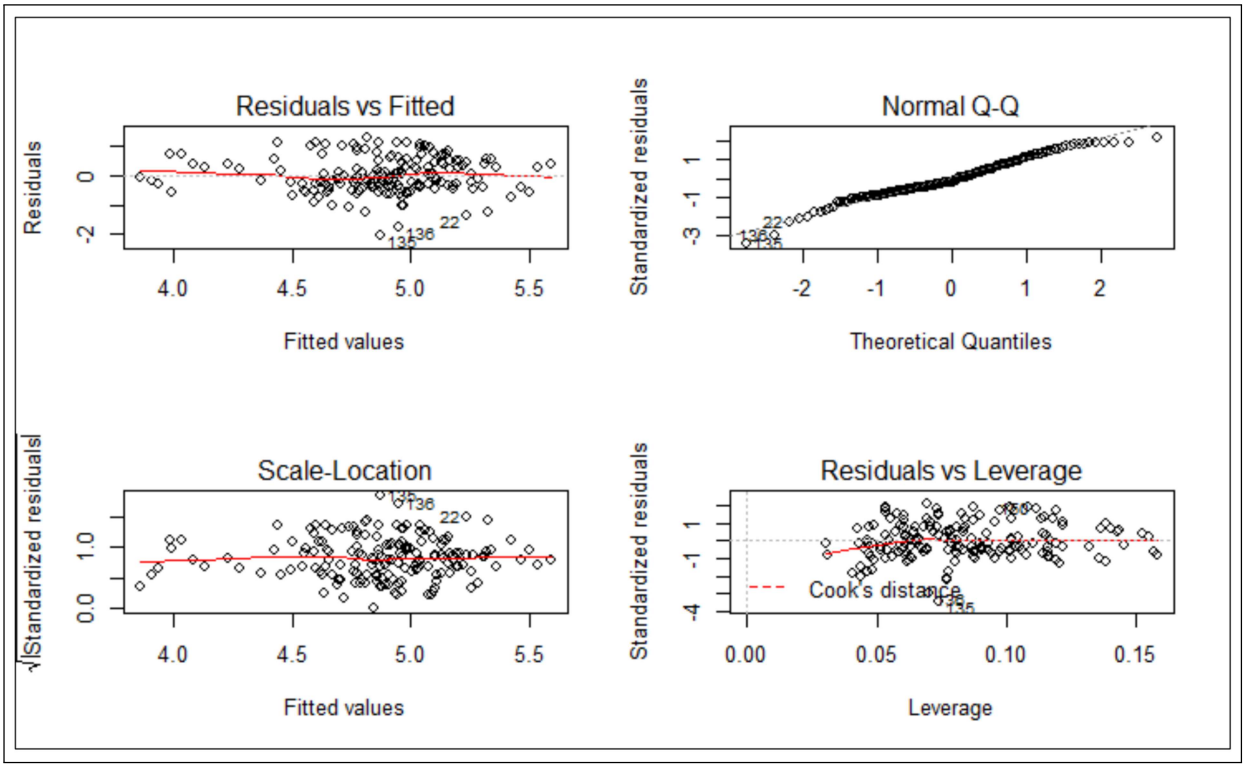

Figure 6: Diagnostic Plot for Log Model, Response Variable = Stroke.

To visualize the effect that a *log* transformation has on the response variable, we created a histogram of the transformed data. Figure 7 is a histogram of the stroke rate after a log transformation. We can see that the transformed data appears to have more normal distribution.

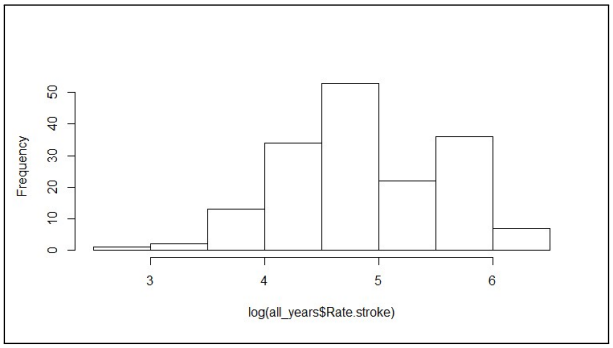

Figure 7: Histogram for Stroke Rate after *log* transformation.

Table 2 includes a full summary of the multiple linear regression model with a *log* transformation and all thirteen explanatory variables included in the model. The F-statistic for the model was 3.261, with a corresponding p-value of 0.0002194. The value of R^2 was 0.2155. Using a level of significance of 0.05, the regression model shows that the significant explanatory

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4537	0.1146	12.69	0.0000
Gender	0.0023	0.0193	0.12	0.9072
year	-0.0130	0.0122	-1.07	0.2860
fruits	-0.0004	0.0003	-1.29	0.2006
NS_Veg	0.0004	0.0003	1.26	0.2101
Beans_Leg	0.0011	0.0006	1.79	0.0752
Nuts_Seeds	-0.0109	0.0043	-2.55	0.0116
UP_Red_Meats	0.0008	0.0004	2.15	0.0334
SS_Bev	-0.0000	0.0002	-0.10	0.9210
fruitJuice	-0.0001	0.0004	-0.21	0.8369
Protein	-0.0010	0.0011	-0.96	0.3395
Calcium	0.0001	0.0001	1.72	0.0879
potassium	0.0000	0.0000	1.86	0.0641
Milk	0.0002	0.0002	1.03	0.3044

Table 2: Linear Regression results for Stroke Rate with Log transformation

variables were nuts & seeds, and unprocessed red meat. Beans & legumes, calcium, and potassium were mildly significant predictors with p-values of 0.0752, 0.0879, and 0.0641 respectively. Next, the model was recalculated using these five explanatory variables. Potassium was no longer significant and was dropped from the model. The final model had an AIC of -757.22 and a BIC of -259.7149, which is evidence that this model is better than the full model containing all thirteen explanatory variables. Comparison of the full and final model are summarized in Table 3.

Model	AIC	BIC	R^2	F-Statistic	p-value
Full Model	-144.9	380.7262	0.2155	3.261	0.0002194
Final Model	-752.22	-259.7149	0.1631	7.939	0.000007153

Table 3: Comparison of full and final models for Stroke Rate

Table 4 summarizes the new multiple regression model with parameter estimates, with a *log* transformation of the response variable, stroke rate, with the following explanatory variables: nuts seeds, unprocessed red meat, beans legumes, calcium. The sign of the coefficient on the explanatory variable indicates a positive or negative effect on Stroke rate. A positive effect would indicate stroke rate decreased so the sign on the explanatory variable would be negative, likewise, a negative effect would indicate that stroke rate increased so the sign on the explanatory variable would be positive. Nuts & seeds have a positive effect on stroke rate, but unprocessed red meat, beans & legumes, and calcium have a negative effect on stroke rate. According to the CDC [7], foods low in saturated fats can prevent stroke, so nuts and seeds can have high saturated fats thus contributing to a higher risk of stroke, which our study supports. In addition, the CDC states that foods high in in fiber can prevent high cholesterol which can lead to a stroke, and beans and legumes are high in fiber, which our study shows has a negative effect on stroke. Unprocessed red meats can actually have high levels of saturated fat

and cholesterol, which would imply that they would increase the risk of stroke according to the CDC, but we did not find that in our results.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4458	0.0493	29.32	0.0000
Nuts_Seeds	-0.0115	0.0029	-3.88	0.0002
Calcium	0.0002	0.0001	3.18	0.0018
Beans_Leg	0.0015	0.0005	3.07	0.0025
UP_Red_Meats	0.0009	0.0003	2.64	0.0090

Table 4: Final linear regression model results for Stroke Rate, with Log Transformation

2.4.2 Model Formulation for Diabetes Rate

Next, we investigated the analysis of the response variable, diabetes rate. The analysis followed the same process as the analysis for stroke rate, starting with a look at some descriptive statistics for the response variable, diabetes rate (Table 5).

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
17.99	46.75	72.05	116.67	149.87	473.51

Table 5: Summary Statistics for diabetes Rate

From Figure 8, we can see that, like stroke rate, the variable diabetes rate is skewed to the right. Similar to stroke rate, a transformation of the response variable, diabetes rate, was necessary.

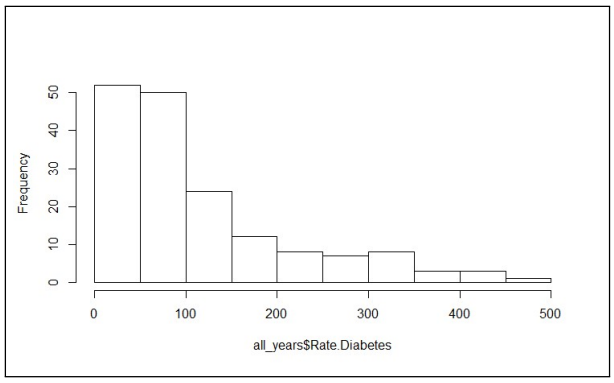

Figure 8: Histogram for diabetes rate.

We used R to determine the best lambda (λ) for the Box-Cox transformation. Figure 9 shows the 95% confidence interval for the maximum likelihood estimator, λ. Using the Box-Cox function in R, the best lambda was calculated to be $\lambda = -\frac{2}{9}$. The 95% confidence interval did not contain zero or another convenient estimate for Lambda, so $\lambda = -\frac{2}{9}$ was used for the transformation.

The model diagnostics for the Box-Cox model are shown in Figure 10. A visual analysis of the diagnostic plot shows that there are no major violations of the model assumptions for either model. The *residuals vs fitted plot* shows a generally random pattern, the *normal QQ plot* shows some slight deviations from normality near the tails of the data, the *scale-location* plot shows that the data has homoscedasticity, and the *residuals vs. leverage* plot shows that there are no outliers or influential data points.

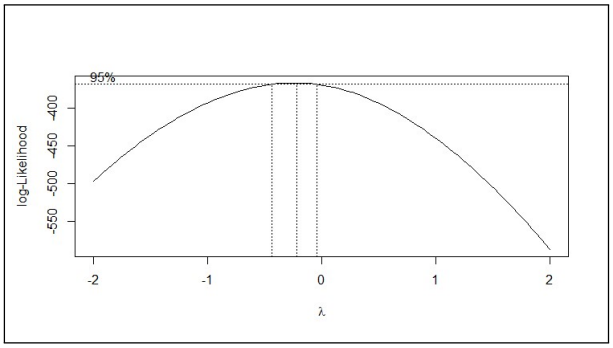

Figure 9: 95% Confidence Interval for Lamda, Diabetes Rate.

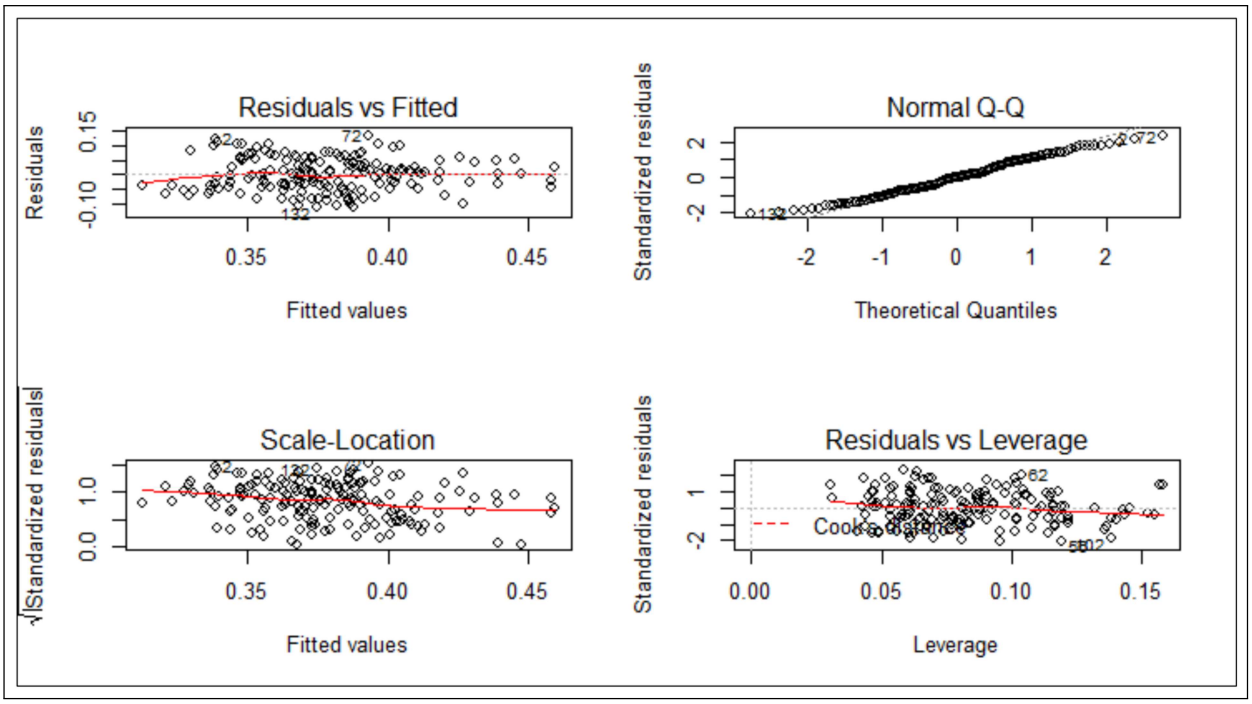

Figure 10: Diagnostic Plot for Box-Cox Model, Response Variable = Diabetes.

To visualize the effect that the Box-Cox transformation has on the response variable, we created a histogram of the transformed data. Figure 11 is histogram of the diabetes rate after a Box-Cox transformation (with $\lambda = -\frac{2}{9}$). We can see that the Box-Cox transformation does a good job "normalizing" the response variable.

Table 6 includes a full summary of the multiple linear regression model with a Box-Cox transformation and all thirteen explanatory variables included in the model. Using a level of significance of 0.05, the regression model shows that the significant explanatory variables were unprocessed red meat, protein, and milk. Year, beans & legumes, and nuts & seeds were mildly significant predictors with p-values of 0.07011, 0.09395, and 0.08111 respectively.

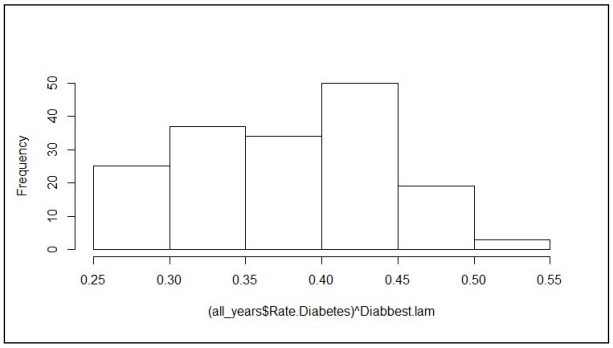

Figure 11: Histogram for Diabetes Rate after *Box-Cox* transformation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4098	0.0657	6.23	0.0000
Gender	-0.0175	0.0111	-1.58	0.1161
year	0.0127	0.0070	1.82	0.0701
fruits	0.0003	0.0002	1.48	0.1404
NS_Veg	-0.0003	0.0002	-1.30	0.1951
Beans_Leg	0.0006	0.0004	1.69	0.0940
Nuts_Seeds	0.0043	0.0024	1.76	0.0811
UP_Red_Meats	-0.0007	0.0002	-2.91	0.0042
SS_Bev	0.0001	0.0001	0.52	0.6006
fruitJuice	0.0000	0.0002	0.16	0.8762
Protein	-0.0014	0.0006	-2.32	0.0217
Calcium	0.0000	0.0000	0.38	0.7060
potassium	0.0000	0.0000	0.57	0.5676
Milk	0.0003	0.0001	2.10	0.0372

Table 6: Linear Regression results for Diabetes Rate with Box-Cox transformation

Next, the model was recalculated using these six explanatory variables. Unprocessed red meat and protein were no longer significant and were dropped from the final model. Table 7 summarizes other model statistics that can be used to determine which model is the best. A comparison of the AIC and BIC for both transformations shows that the model using a Box-Cox transformation is best. Based on the model diagnostics, histogram, AIC/BIC, and Coefficient of Determination R^2, the linearized model using a Box-Cox transformation was determined to be the best and was used to develop the final parsimonious model.

Model	AIC	BIC	R^2	F-Statistic	p-value
Full Model	-457.7402	-410.8807	0.2019	2.996	0.0006074
Final Model	-937.5627	-440.0555	0.1172	5.41	0.0004082

Table 7: Comparison of full and final models for Diabetes Rate

The final model had an AIC of -937.56 and a BIC of -440.0555, which is evidence that the

final model is better than the full model containing all thirteen explanatory variables. Table 8 summarizes the new multiple regression model, including parameter estimates, with a Box-Cox transformation with the following explanatory variables: year, nuts & seeds, milk, beans & legumes. Nuts & seeds, milk, and beans & legumes all had a negative effect on the response variable, diabetes rate. The explanatory variable "year" was a coded numerical factor where year 1990 was coded as "1", 2000 was coded as "2", and 2010 was coded as "3". In this model that means that for every decade, from 1990 to 2010, the rate of diabetes increased by about 0.28. According to the WHO [2], saturated fats increase the risk of diabetes and non-starchy vegetables, plant-based proteins, and lean meats decrease the risk of diabetes. Our study supports these recommendations as beans and legumes had a negative effect on diabetes and they are plant-based protein sources that the WHO recommends. Nuts and seeds are also also plant based protein sources, yet they can also have high levels of saturated fats. Our study shows that nuts and seeds have a negative relationship. In addition, the WHO estimates that the number of diabetes cases rose from 108 million in 1980 to 422 million in 2014, and our study agrees with the increase in diabetes rate from 1990 to 2010.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2873	0.0219	13.14	0.0000
year	0.0185	0.0061	3.03	0.0029
Nuts_Seeds	0.0046	0.0016	2.86	0.0047
Milk	0.0002	0.0001	2.30	0.0227
Beans_Leg	0.0005	0.0003	1.89	0.0602

Table 8: Final linear regression model results for Diabetes Rate, with Box-Cox Transformation

3 Conclusions

3.1 Interpretation and Summary

The aim of this project was to study if there is a relationship between diet and death by strokes and diabetes. We started by utilizing Principal Component Analysis to describe the data. We noticed that there were clusters of countries laying geographically close to each other who had similar food intake, such as Finland and Iceland. We also noticed that non-starchy vegetables and legumes were similar in characteristics, and so were sugar-sweetened beverages and fruit juice. After the descriptive statistics, we progressed to create a model for our variables by utilizing regression analysis. Multiple linear regression results provided parameter estimates for the multiple linear regression equations. Following are the final regression equations for the response variable stroke rate and diabetes rate, respectively

$$\begin{aligned} log(y_{Stroke}) = 1.4458 - 0.0115(NutsSeeds) + 0.0002(calcium) \\ + 0.0015(BeansLegumes) + 0.0009(UPRedMeats) \end{aligned}$$

$$\begin{aligned} (y_{Diabetes})^{\lambda} = 0.2873 - 0.0185(Year) + 0.0046(NutsSeeds) + 0.0002(milk) \\ + 0.0005(BeansLegumes) \end{aligned}$$

Note that in each case, the response variable has been transformed. For the response variable, stroke rate a *log* transformation was used, and for the response variable, diabetes rate, the Box-Cox transformation, with $\lambda = -\frac{2}{9}$, was used. In this context, interpretation of the intercept does not make sense since that would imply that consuming none of the eleven nutrients is within the parameters of the model. Both models show that consumption of beans and legumes is associated with slightly higher rates of both stroke and diabetes. Nuts and seeds have a positive effect on the rate of diabetes, but a negative effect on the rate of stroke. Consumption of unprocessed red meat has a slightly negative effect on the rate of stroke. Finally, the variable year is a significant factor in the model for diabetes. Because this effect is negative, we can conclude that over time, specifically each decade from 1990 to 2010, the rate of diabetes has decreased. Overall, these models shows some interesting associations between certain nutrients and disease rates for stroke and diabetes. These correlations do not imply causation and further research and experimentation is needed to state definitively that nutrition and consumption of certain foods causes higher or lower rates of stroke or diabetes.

3.2 Suggestions for Further Study

The conclusions made by the model presented demonstrate correlation, but in order to study if causation does in fact exist, an experimental study would need to be conducted. In

addition, nuts and seeds, unprocessed red meat, and beans and legumes were determined to influence diabetes and stroke, but the method in how these foods are prepared and seasoned can play a huge role on their effect on disease. It would be useful to study how such methods could affect the correlation, as well as how the addition of other food items, such as spices and condiments, not mentioned in our study could further influence the incidence of disease or stroke. Also, the food categories studied were extremely general, and studying more specific foods could uncover more accurate information on the link of food and disease. For example, when studying unprocessed red meat, specifying how the animals were raised, such as factory-farmed versus grass-fed, and the type of red meat, such as lamb, cattle, or bison, would be extremely relevant to study.

References

[1] Cancer: Overview, prevention, and management. https://www.who.int/health-topics/cancertab=tab$_2$, $March$2021.

[2] Diabetes: Key facts. https://www.who.int/news-room/fact-sheets/detail/diabetes, March 2021.

[3] Eat good to feel good. https://www.diabetes.org/nutrition/healthy-food-choices-made-easy, March 2021.

[4] The gdd 2015 beta-version. https://www.globaldietarydatabase.org/gdd-2015-beta-version, March 2021.

[5] Global health data exchange. http://ghdx.healthdata.org/, March 2021.

[6] Scope of current data collection. https://www.globaldietarydatabase.org/methods/scope-current-data-collection, March 2021.

[7] Stroke. https://www.cdc.gov/stroke/, March 2021.

[8] Stroke, cerebrovascular accident. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html, March 2021.

[9] M. Brems. A one-stop shop or principle components analysis. *Towards Data Science*, 2017.

[10] E. L. Frome. The analysis of rates using poisson regression models. *Biomettics*, Volume 39, No. 3, 1983.

[11] R. Halinski. The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, Volume 7:151-157, 1970.

[12] Osborne Jason. Four assumptions of multiple linear regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, Volume 8, 2003.

[13] J. Kuha. Aic and bic: Comparisons of assumptions and performance. *Sociological Methods and Research*, Volume 33(2):188-229, 2004.

[14] Linh Ngo. Principal component analysis explained simply. https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/, June 2021.

[15] A. Robertson. Diet and disease. *WHO Regional Publications, European Series*, 2021.

[16] J. Shlens. A tutorial on principal component analysis. *Google Research*, 2014.