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.

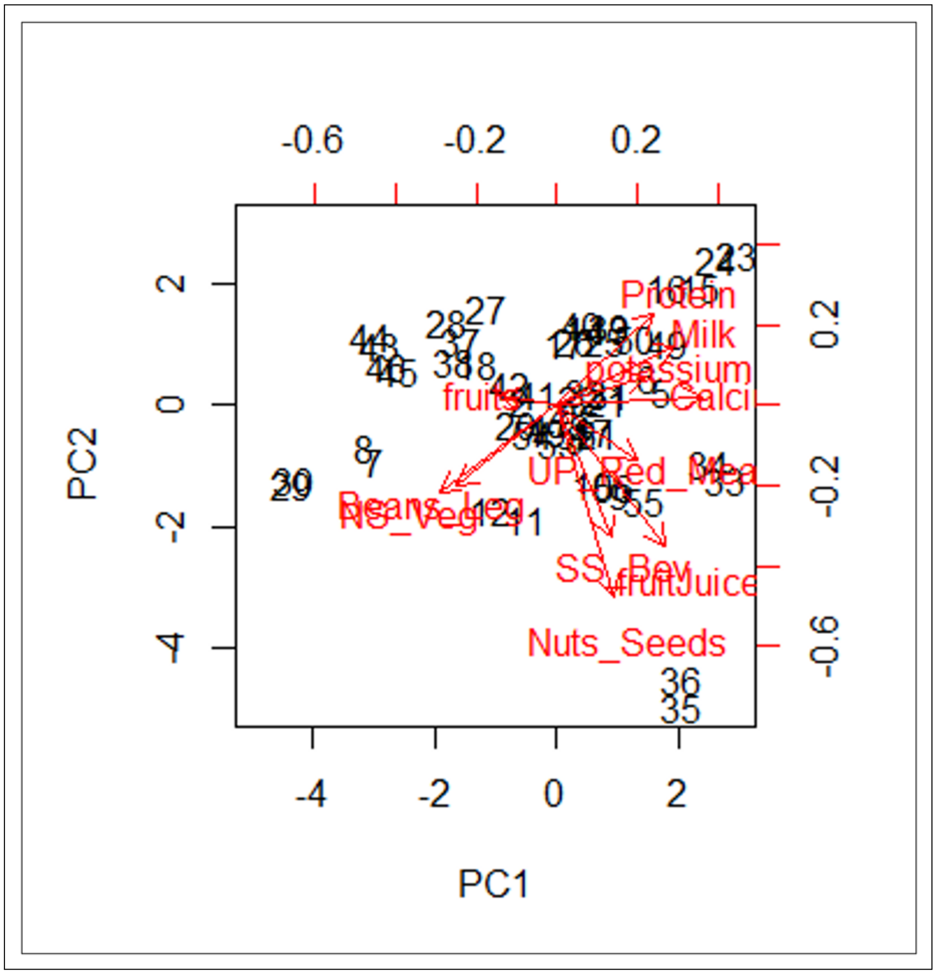

Figure 1: PCA Biplot for Year 1990.

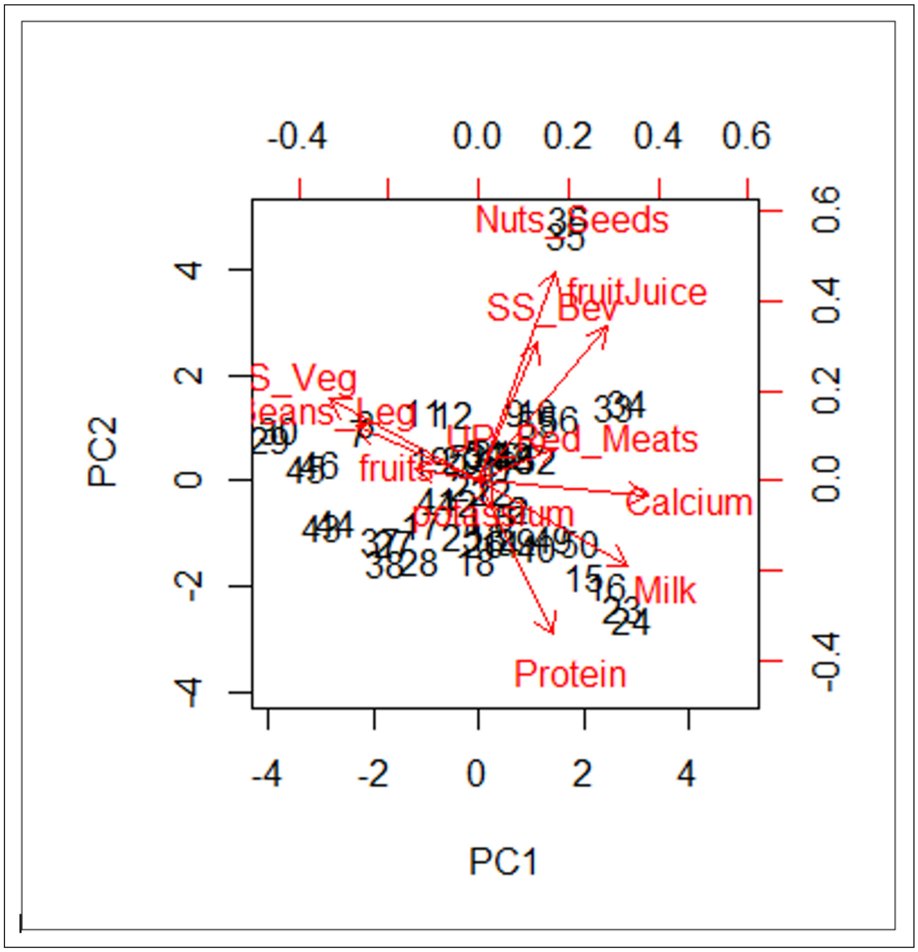

Figure 2: PCA Biplot for Year 2000.

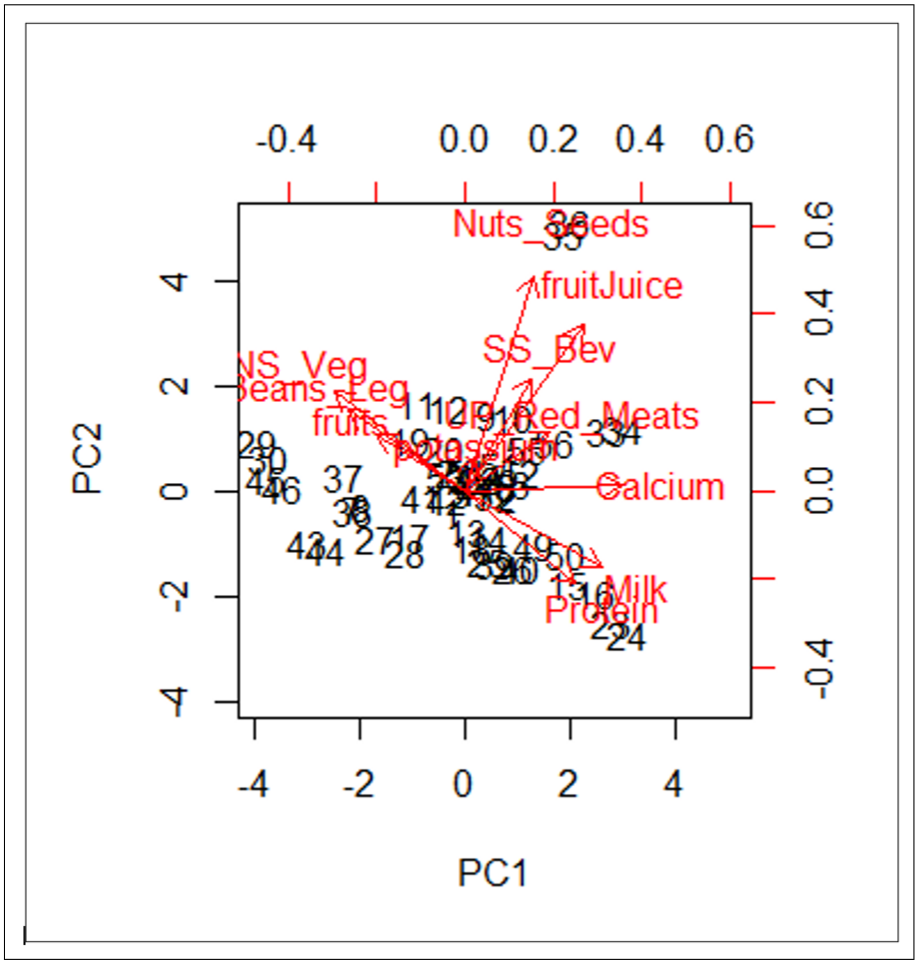

Figure 3: PCA Biplot for Year 2010.

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 analyzed 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%

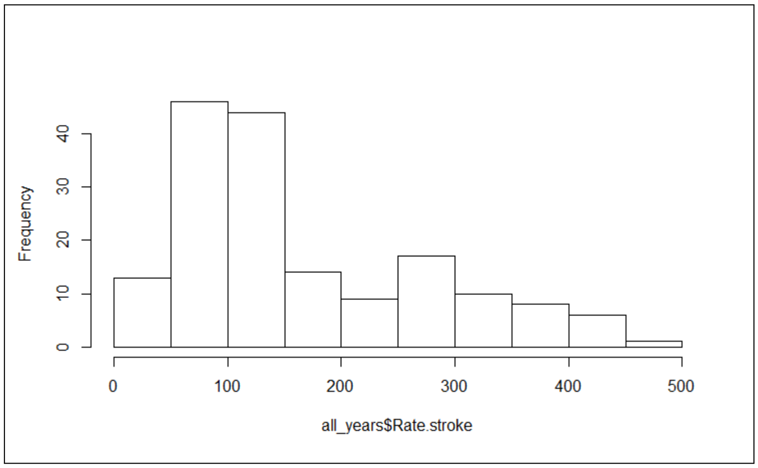

Figure 4: Histogram for Stroke rate

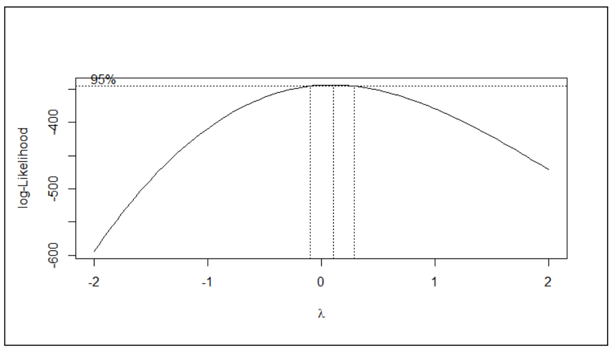

Figure 5: 95% Confidence Interval for Lamda, Stroke Rate.

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.

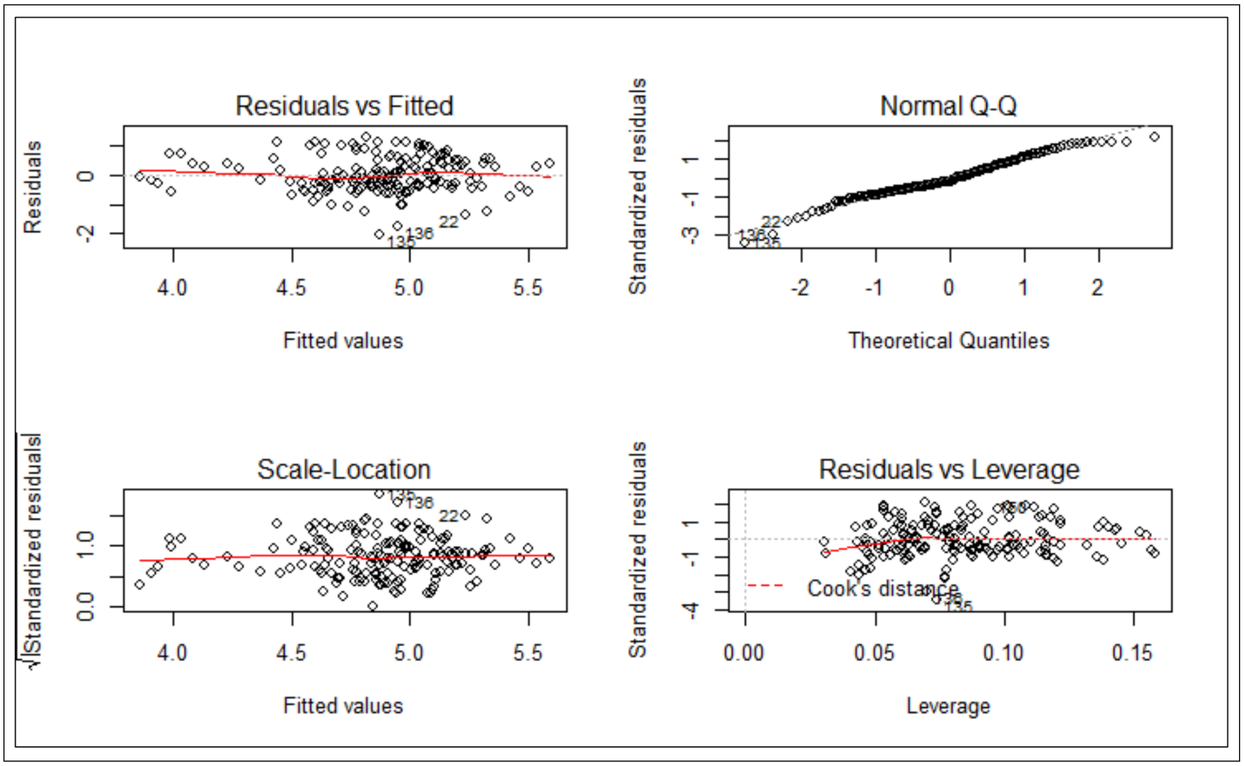

Figure 6: Diagnostic Plot for Log Model, Response Variable = Stroke.

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.

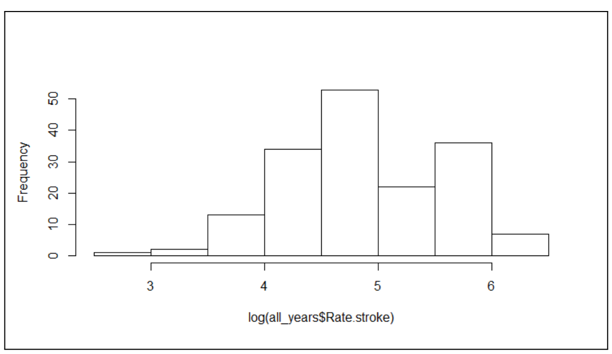

Figure 7: Histogram for Stroke Rate after *log* transformation.

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.

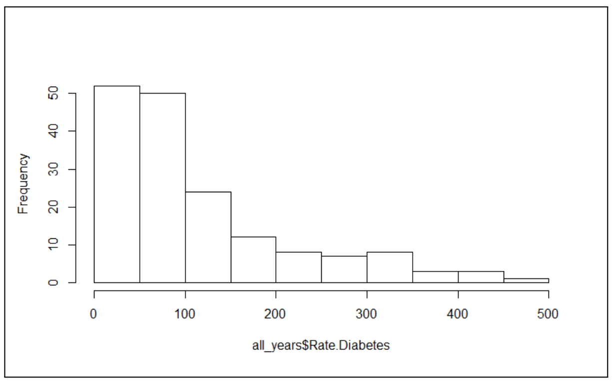

Figure 8: Histogram for diabetes rate.

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.

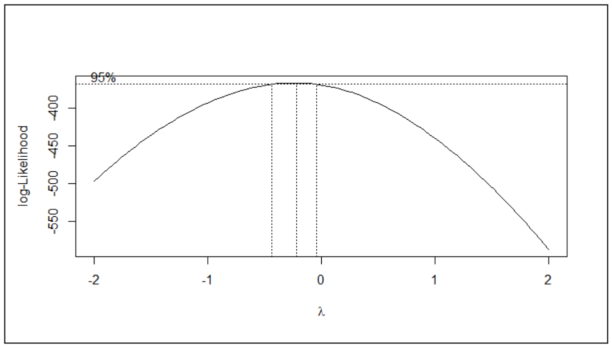

Figure 9: 95% Confidence Interval for Lamda, Diabetes Rate.

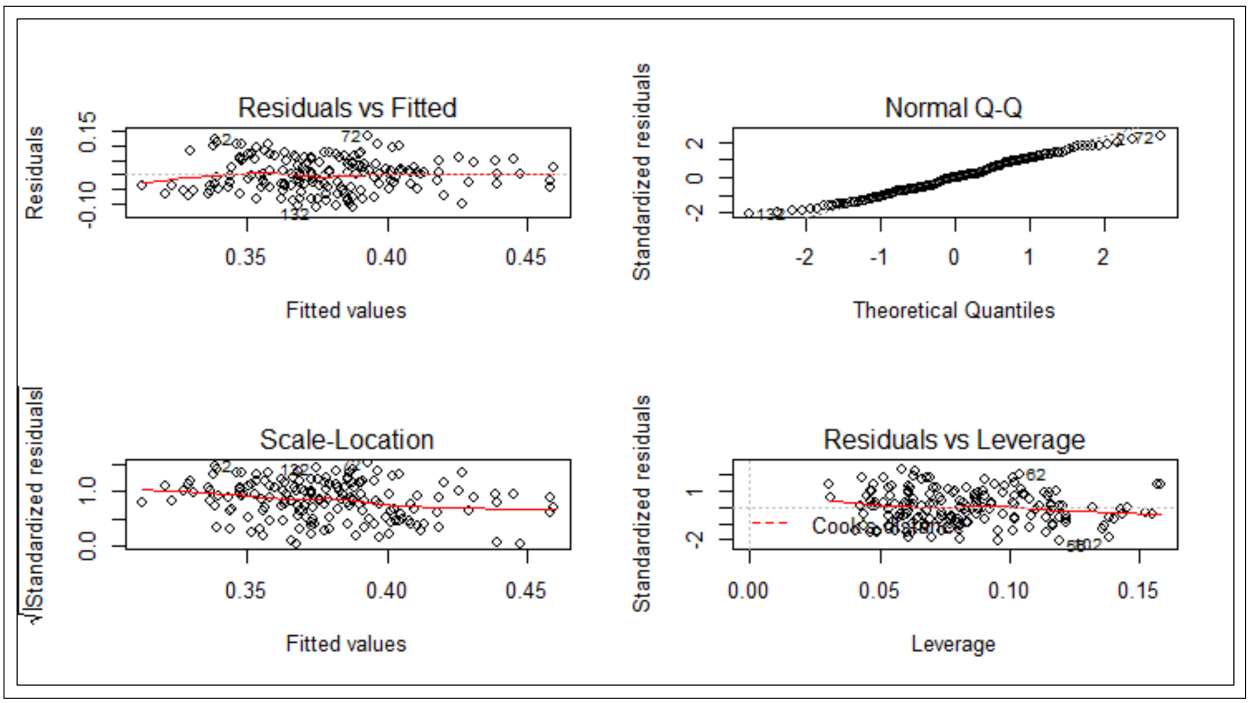

Figure 10: Diagnostic Plot for Box-Cox Model, Response Variable = Diabetes.

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.

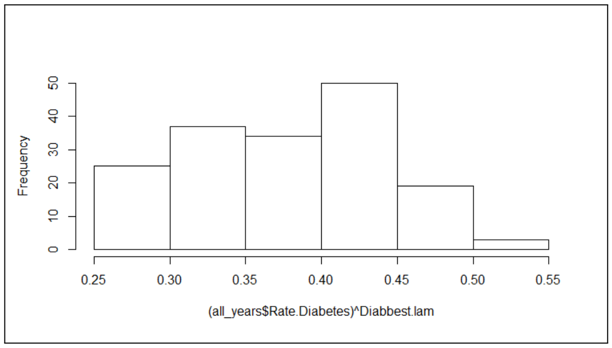

Figure 11: Histogram for Diabetes Rate after *Box-Cox* transformation.

	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.

Appendix A: Full Data Set

Year	High Income countries including HIC in Asia	Country Label	Country Name	Gender (female 1, Male 0)	Fruits, grams per day	Non-starchy vegetables, grams per day	Beans and legumes, grams per day	Nuts and seeds, grams per day	Unprocessed red meats, grams per day	Sugar-sweetened beverages, grams per day	Fruit juices, grams per day	Total protein, grams per day	Calcium, milligrams (mg) per day	Potassium, milligrams (mg) per day	Total milk, grams per day	Cause of death, Stroke, per 100k	Cause of death, Diabetes, per 100k	Cause of death, Cancer, per 100k
1990	HIC	AUS	Australia	1	159.31151	150.49599	15.939858	3.1044965	65.648636	110.39997	49.192272	85.916626	749.19263	2747.9321	170.35217	110.5469533	35.14046125	223.9384147
1990	HIC	AUS	Australia	0	140.58208	139.63774	18.12635	3.6225703	106.43897	156.58426	43.886585	92.454269	763.20398	2778.8025	188.52759	82.99883236	31.61263907	304.9853905
1990	HIC	AUT	Austria	1	142.14287	133.45335	9.1065178	4.0949373	102.93442	93.090706	59.895756	67.451645	737.94214	2810.1514	117.99114	144.6294749	372.7850327	30.88219447
1990	HIC	AUT	Austria	0	116.72826	116.27825	9.5277596	4.6677866	128.04825	124.90359	54.030766	72.493256	688.99634	2832.2078	116.34063	218.5355215	322.3100979	47.26288964
1990	HIC	BEL	Belgium	1	154.44269	148.2363	4.978188	3.0052705	76.479645	156.85667	60.583244	96.847198	802.00458	4208.9834	98.078117	186.2897654	332.151185	53.80870776
1990	HIC	BEL	Belgium	0	129.57446	131.63817	5.3503666	3.5739744	82.3862	202.14958	54.336403	100.58058	701.75836	4139.7983	94.575615	126.8571193	473.5059451	32.21826594
1990	Asia	BRN	Brunei Darussalam	1	92.637444	166.19374	7.8864026	4.3079109	18.684093	182.9133	6.6887312	68.125137	480.82251	1742.6666	27.901379	81.36804483	141.3815499	82.03344863
1990	Asia	BRN	Brunei Darussalam	0	81.005867	159.10403	7.1396027	4.1945276	18.750452	218.29222	6.3026342	68.177971	442.32208	1697.9625	27.153236	85.81213626	141.5343469	80.59348883
1990	HIC	CAN	Canada	1	149.77589	173.51277	14.888444	5.814899	57.704567	125.40845	113.62634	74.112221	876.97925	2838.0876	193.03419	85.77922094	36.48229886	233.952078
1990	HIC	CAN	Canada	0	125.86967	152.5697	15.665134	6.7413349	70.741798	164.18155	102.07849	79.677795	798.76971	2860.3403	186.93143	64.47667964	34.45755256	310.1178762
1990	HIC	CYP	Cyprus	1	193.49896	219.93274	69.882217	6.1379318	128.18497	100.04411	60.163132	81.211693	826.2702	2503.519	170.62866	161.6577128	137.8075959	163.1952326
1990	HIC	CYP	Cyprus	0	162.84715	194.85051	74.188736	6.9270687	141.64493	129.36247	54.441647	87.33876	772.82031	2519.4734	170.48256	109.5757082	101.8639551	193.9368954
1990	HIC	DNK	Denmark	1	132.31978	129.69385	31.080357	2.4302604	63.012394	59.524784	45.789452	81.217339	941.45563	2788.4971	159.17484	363.2394776	182.9027015	31.09857235
1990	HIC	DNK	Denmark	0	110.06332	114.92109	33.2967	2.7974608	70.384628	83.198914	41.254265	87.314697	871.63501	2813.6294	156.0202	407.917783	144.1890135	33.56760498
1990	HIC	FIN	Finland	1	135.48993	125.75286	7.8366876	2.4079967	55.491825	50.992348	58.933197	85.024368	1117.4998	2542.1951	319.85489	250.75526	191.3584389	25.31529482
1990	HIC	FIN	Finland	0	108.36874	109.46956	8.3972712	2.7361574	72.645805	75.450768	53.135303	91.423538	1044.5688	2555.4165	319.23703	296.3244102	125.9928919	16.77726931
1990	HIC	FRA	France, Monaco	1	114.96436	200.68903	18.402555	2.6559777	44.472233	74.039932	23.491199	80.656746	806.68726	2929.7466	158.46062	148.2218127	260.1729965	38.01627802
1990	HIC	FRA	France, Monaco	0	104.55181	186.86267	19.779915	3.0677223	58.878448	102.12857	21.216118	100.10294	879.42297	3403.5105	155.88129	112.7896115	446.3381293	31.58186594
1990	HIC	DEU	Germany	1	205.88083	221.08168	11.432217	2.4357324	52.780968	89.08918	78.332245	75.364861	1021.333	3431.9038	89.988457	213.4856576	61.71786397	298.2117361
1990	HIC	DEU	Germany	0	167.60831	189.73515	12.504461	2.8755314	67.213448	132.15805	69.259834	78.160034	949.29346	3330.6543	90.159378	137.8756664	36.03715234	343.0851244
1990	HIC	GRC	Greece	1	161.95056	165.04317	25.325388	5.7366557	101.76545	71.497383	27.922808	80.343323	951.56122	3975.9375	104.14078	58.69082204	312.9250049	236.3690614
1990	HIC	GRC	Greece	0	138.70767	147.27151	26.834435	6.2817788	107.95437	93.337769	25.096693	86.355988	891.57965	4015.2949	101.08551	49.09214735	227.3800099	385.4364202
1990	HIC	ISL	Iceland	1	105.65095	90.401009	5.192903	0.4955461	63.2145	102.15916	61.972347	92.503426	1044.8717	2344.3608	336.1077	241.0430157	110.2412137	16.34069072
1990	HIC	ISL	Iceland	0	85.100311	79.848534	5.5842638	0.55479509	71.198212	127.63717	55.461643	99.234215	965.21991	2364.0696	331.54584	269.5963559	90.37161784	17.31555298
1990	HIC	IRL	Ireland	1	111.6491	174.50885	39.341621	1.2548673	43.903561	108.63662	49.471588	85.864693	833.08337	3049.2185	219.54808	286.7809788	155.249822	37.0675936
1990	HIC	IRL	Ireland	0	95.877159	157.08861	42.016762	1.43165	50.391045	138.78421	44.570477	92.258636	810.46411	3078.1543	217.196	350.1343726	121.2214346	37.4313684
1990	HIC	ITA	Italy	1	222.10255	166.62062	22.692709	0.81436813	62.514748	48.194157	19.297182	76.472603	838.22504	2688.6426	126.7723	69.70774718	187.2215708	271.546507
1990	HIC	ITA	Italy	0	199.17184	148.45224	23.790369	0.90611005	69.745476	61.049908	17.334097	84.018501	821.38068	2937.6484	116.83884	50.84434268	153.3633403	425.6637881
1990	Asia	JPN	Japan	1	144.55228	264.8472	81.330879	2.2926466	53.515095	90.136383	23.148926	69.524361	576.47546	2774.6497	113.27097	142.3037694	204.091117	32.59231802
1990	Asia	JPN	Japan	0	108.78735	242.58707	71.306587	1.9242971	56.286198	117.54391	22.058584	65.601036	470.68033	2354.4951	97.743011	132.8325032	318.6338257	31.98079407
1990	HIC	LUX	Luxembourg	1	119.13248	141.62083	12.215356	3.5702846	95.585258	103.49871	67.308907	66.910301	870.11877	3276.2073	108.67527	300.5039534	232.6309167	44.39328323
1990	HIC	LUX	Luxembourg	0	100.51255	126.28574	13.032359	4.0825024	103.8773	136.64531	60.644188	72.079422	814.24664	3298.0132	106.84019	405.410232	170.9741857	30.89844745
1990	HIC	MLT	Malta	1	148.21938	128.03325	25.925381	8.1420155	123.61172	130.75967	102.49653	77.823555	1068.3932	2250.5452	253.13715	138.3912947	65.6080854	204.3869992
1990	HIC	MLT	Malta	0	126.20165	112.37483	27.974649	9.2002974	133.40355	166.36156	92.439926	82.707176	966.81909	2264.2288	250.07898	119.2120758	50.91215369	271.9682304
1990	HIC	NLD	Netherlands	1	168.65582	172.63239	46.96714	12.439658	85.827393	168.32877	194.53049	66.675926	955.84821	2770.4138	163.142	138.1856437	58.09927991	267.4459986
1990	HIC	NLD	Netherlands	0	145.16576	160.22549	53.699947	14.155262	93.261543	197.34187	175.69302	67.804375	883.30524	2722.731	158.26497	97.48312181	37.37624852	382.769191
1990	HIC	NZL	New Zealand	1	261.75635	170.40376	83.292503	1.5869818	70.793404	54.38353	49.510426	94.773636	715.68317	2875.283	193.0836	124.3430179	31.28807023	262.3652613
1990	HIC	NZL	New Zealand	0	204.99631	148.7225	81.669334	1.7228501	59.017643	63.053654	44.422302	101.83575	541.46179	2905.5098	180.08012	88.69067433	32.09431741	312.2011072
1990	HIC	NOR	Norway, Svalbard and Jan Mayen	1	115.48391	115.7535	3.485769	2.3564482	97.540077	71.19194	40.325745	79.160393	794.15875	2568.8999	183.36372	202.410189	293.1489594	27.23773228
1990	HIC	NOR	Norway, Svalbard and Jan Mayen	0	95.211037	101.2635	3.7233541	2.7910926	105.40411	96.584694	36.161125	85.101128	751.64691	2587.7366	179.80382	156.6131725	363.9308139	29.26598459
1990	HIC	PRT	Portugal	1	148.22816	158.43887	44.418266	2.9519796	107.1513	96.472931	9.5310802	73.558792	837.24664	3521.0405	145.51485	386.6025086	64.6913029	249.8211236
1990	HIC	PRT	Portugal	0	127.17575	141.14912	46.822361	3.3615072	118.96131	124.94087	8.5983763	79.040375	769.07178	3548.4119	144.84517	341.5324813	58.69985637	352.3111672
1990	Asia	KOR	Republic of Korea	1	146.37234	164.59535	72.050262	0.57572073	81.908554	26.420126	19.495726	74.834068	465.50217	2715.2434	117.42743	178.6540091	130.3108224	26.29604963
1990	Asia	KOR	Republic of Korea	0	107.87592	160.81282	64.499657	0.5511151	88.043091	41.35413	19.489893	74.754173	443.55917	2551.3154	99.735306	155.3351278	202.8179374	31.24411562
1990	Asia	SGP	Singapore	1	169.23154	168.42163	48.366718	3.1226339	37.38768	97.072838	7.791151	68.793411	474.39111	2619.7888	51.106758	81.83478081	130.5069454	36.42483614
1990	Asia	SGP	Singapore	0	142.83807	158.47678	43.004539	3.1033921	38.971302	121.0134	7.4823904	68.894638	424.12317	2550.8516	46.777908	70.98112613	182.8850847	27.04809205
1990	HIC	ESP	Spain	1	130.89755	129.3286	35.975891	7.9604626	73.190536	99.709251	25.057211	76.925148	846.508	3513.2173	184.49921	192.1619232	72.38189698	217.7680699
1990	HIC	ESP	Spain	0	115.57912	115.91761	38.108578	9.1152754	82.47747	128.84766	22.602625	82.673058	765.55975	3541.4204	182.50151	145.5112908	51.95833946	379.0985148
1990	HIC	SWE	Sweden	1	140.9924	141.32358	11.793897	2.5711465	83.810837	70.250031	70.055984	78.412903	865.55841	2849.2632	338.36707	308.1679258	174.7321513	31.99576986
1990	HIC	SWE	Sweden	0	108.76622	121.79988	13.146868	2.8985648	105.86221	93.084892	63.138577	83.173424	833.33661	2761.4944	331.35205	355.4465049	136.1715584	33.64092733
1990	HIC	CHE	Switzerland, Liechtenstein	1	158.64944	140.28749	6.0474172	4.4161873	59.278107	96.751625	87.624962	71.760468	1003.2766	2727.436	143.56708	152.8980529	47.38322017	239.2508177
1990	HIC	CHE	Switzerland, Liechtenstein	0	130.14381	124.05106	6.3736196	4.9911528	65.442749	123.73541	78.3274	75.791641	942.86267	2750.479	141.2681	108.2680061	34.65011637	338.6819592
1990	HIC	GBR	United Kingdom	1	121.22377	149.83224	45.199345	4.4890604	42.320396	149.96219	40.170666	69.08831	862.82397	3242.7217	207.31943	204.4979975	31.54492021	360.5623333
1990	HIC	GBR	United Kingdom	0	99.828865	131.59648	47.046387	5.1456985	46.56868	183.50287	36.03326	72.19091	807.54443	3449.6738	196.50546	135.4848063	28.09494174	430.5239469
1990	HIC	USA	USA, Puerto Rico and US Virgin Islands	1	91.59034	141.59204	24.638857	3.6891048	44.533421	255.1413	98.898193	82.591835	804.12561	2707.9653	175.69659	98.92616685	265.5123857	47.21529154
1990	HIC	USA	USA, Puerto Rico and US Virgin Islands	0	82.069633	130.02467	29.514435	4.2735281	54.942394	304.28864	89.807762	87.827232	788.9516	2738.8032	188.2646	70.31567222	326.9580503	41.62657791
2000	HIC	AUS	Australia	1	156.93558	140.87564	25.394457	3.4587562	59.977264	113.08083	44.382534	88.362076	774.39563	2761.6936	141.4053	99.08980825	38.92229661	231.4570259
2000	HIC	AUS	Australia	0	139.05083	131.08856	28.984497	4.0372586	97.200378	159.60054	39.966999	95.085266	788.95068	2794.6577	156.34441	74.24931796	41.28216489	306.2768703
2000	HIC	AUT	Austria	1	135.74251	124.82738	8.6576595	3.6854455	94.041382	94.232254	54.662483	67.276268	761.10437	2973.7349	101.09904	156.0734155	48.99565223	295.3756573
2000	HIC	AUT	Austria	0	112.52457	109.61582	9.1030989	4.2216549	117.13522	125.10043	49.08712	72.300339	710.97168	2999.2932	99.398003	97.90480565	36.10563245	351.5719924
2000	HIC	BEL	Belgium	1	133.71361	139.57938	4.1109142	3.3174531	71.919029	157.39064	55.795696	93.424423	829.38818	4503.3428	86.222519	150.5538975	54.18887139	308.4196382
2000	HIC	BEL	Belgium	0	112.65421	124.32944	4.4242029	3.9491925	77.513466	202.21303	49.940624	97.037079	725.79486	4432.8247	83.011314	102.759242	35.88012708	438.874893
2000	Asia	BRN	Brunei Darussalam	1	93.255791	162.02965	7.9055419	4.2786779	17.416063	184.18207	6.5921154	73.15107	490.51385	1908.8055	26.36484	120.1687452	61.8143391	66.34622216
2000	Asia	BRN	Brunei Darussalam	0	81.516647	155.00371	7.2445016	4.2414904	17.538639	218.08769	6.240077	73.263397	451.89902	1864.545	25.672522	113.6835738	53.2538607	67.34412945
2000	HIC	CAN	Canada	1	153.15549	165.58287	9.0262394	5.9419432	49.979919	124.32528	104.47198	68.692108	901.23657	3167.7688	166.7795	86.35602805	49.54854092	252.9996763
2000	HIC	CAN	Canada	0	128.9592	145.87479	9.5033998	6.8948326	61.286419	162.68036	93.726578	73.98101	821.00854	3198.7422	161.26653	62.73937182	50.53897904	310.874087
2000	HIC	CYP	Cyprus	1	176.8679	201.18304	55.898224	5.8970523	105.23022	103.84737	54.305317	80.811172	852.57922	2497.0767	143.65512	126.0473428	158.6225631	115.9413065
2000	HIC	CYP	Cyprus	0	149.5894	179.05966	59.501579	6.6632309	116.54733	133.64996	48.944839	86.848618	798.05823	2521.2527	143.09808	106.7457939	204.6934093	114.4647228

Appendix A: Full Data Set

2000	HIC	DNK	Denmark	1	137.13255	120.50188	23.369337	2.2778065	65.019264	60.344112	41.849892	75.572739	972.22675	3106.6472	136.37762	153.8390974	44.07652789	387.839504
2000	HIC	DNK	Denmark	0	114.87169	107.38815	24.792774	2.6323669	73.197098	83.519646	37.816227	81.432465	901.1687	3127.2288	134.5641	116.7617915	51.71746058	407.6371838
2000	HIC	FIN	Finland	1	134.83607	119.36099	6.2678781	2.3450027	48.904675	50.244759	54.931602	84.823036	1153.1167	2597.7612	282.24368	164.7261965	21.95371292	259.1480043
2000	HIC	FIN	Finland	0	108.85964	104.61882	6.7452226	2.6722312	64.083481	73.69899	49.376942	91.411972	1078.4974	2614.2607	281.04684	104.9933223	17.9903166	289.2044733
2000	HIC	FRA	France, Monaco	1	118.25168	190.27296	14.073345	2.7509842	39.275307	71.635826	22.135496	79.237938	835.85052	2855.0383	139.41356	270.3462026	113.6834122	44.6769527
2000	HIC	FRA	France, Monaco	0	107.97547	177.52371	15.152725	3.1817622	52.02327	98.419083	19.991577	98.34861	911.56238	3319.1694	137.14151	445.2743083	86.97322735	41.12656538
2000	HIC	DEU	Germany	1	202.32556	212.42178	8.8725176	2.4223804	50.996273	86.644485	73.789871	73.384361	1053.739	3549.0554	80.627731	351.3101282	148.7096735	62.32029212
2000	HIC	DEU	Germany	0	166.98608	184.18611	9.7788439	2.8706615	65.059708	126.41478	64.89991	76.156631	980.20508	3454.1055	80.576866	392.8281366	95.1749206	42.07160567
2000	HIC	GRC	Greece	1	156.94447	153.26363	21.689384	5.4906931	79.633827	74.905106	25.502964	82.284889	973.43542	4343.5977	92.874718	284.8685702	46.58440255	260.2242339
2000	HIC	GRC	Greece	0	133.93076	136.40349	22.941969	6.0053458	84.45372	98.269821	22.94742	88.264236	911.91309	4384.3184	90.207253	212.9553673	42.79111888	418.4055765
2000	HIC	ISL	Iceland	1	109.7618	87.140556	4.9691563	0.48862928	51.900578	101.85502	58.068756	96.85878	1067.0776	2630.8057	286.11246	249.218502	100.3444182	17.10624431
2000	HIC	ISL	Iceland	0	88.698891	77.162392	5.3503151	0.54659313	58.505409	126.7852	51.946598	103.89218	986.1734	2656.4941	282.05591	266.1867741	74.76210797	19.26004482
2000	HIC	IRL	Ireland	1	100.32801	149.22032	25.93013	1.1429518	34.537098	119.6334	41.89539	98.738945	887.65271	3570.6689	164.86142	264.6181251	116.2718072	32.24011579
2000	HIC	IRL	Ireland	0	86.283165	134.44777	27.765566	1.3069788	39.677517	152.44305	37.712353	106.13203	864.12744	3599.5757	163.08932	316.6431236	86.49031457	35.32552583
2000	HIC	ITA	Italy	1	224.95158	157.35268	17.542099	0.80710828	55.448383	48.001633	17.877642	76.409225	863.68823	2831.8843	111.54497	295.4159617	66.89741839	166.5190258
2000	HIC	ITA	Italy	0	202.23233	140.53569	18.414806	0.9002974	61.894539	60.816288	16.03463	83.995117	846.43866	3097.3887	102.46165	441.4398211	53.41087179	126.0305243
2000	Asia	JPN	Japan	1	144.44579	248.51274	71.121567	2.4562759	45.46479	89.806671	22.240952	69.688049	596.6275	2800.3015	105.50082	123.7311899	34.28639075	242.6123182
2000	Asia	JPN	Japan	0	108.77936	227.93599	62.133942	2.0622079	47.889767	117.56184	21.21431	65.807991	487.07666	2379.3999	90.746361	120.8230856	34.90215103	386.7464149
2000	HIC	LUX	Luxembourg	1	129.33229	125.97251	17.085117	2.1237285	77.729324	110.22777	58.635342	73.81179	909.70813	3294.7524	90.780373	167.7080648	46.72501492	290.0531312
2000	HIC	LUX	Luxembourg	0	110.1596	113.03977	18.276754	2.4395239	84.541412	144.0769	52.738739	79.435425	850.7381	3323.605	89.167709	109.3825858	32.60522166	368.698562
2000	HIC	MLT	Malta	1	146.15157	112.23044	17.326529	8.2437496	94.146416	140.01285	86.557259	85.555573	1126.4799	3116.4209	199.60292	123.1392167	63.15901238	219.1830212
2000	HIC	MLT	Malta	0	124.51641	98.586006	18.673586	9.3423347	101.62215	178.52522	77.943062	90.797775	1020.1349	3145.1926	196.69159	92.78423219	51.4539371	267.575776
2000	HIC	NLD	Netherlands	1	159.44373	159.89632	34.633797	14.912676	72.877235	172.63936	173.78696	74.755547	989.69928	3155.4744	137.97008	142.7128509	58.6136902	295.1275291
2000	HIC	NLD	Netherlands	0	138.13588	149.12082	39.729763	17.00531	79.305771	201.24881	156.60229	76.037735	914.78168	3105.1768	133.63802	92.33879549	41.11104492	372.4066457
2000	HIC	NZL	New Zealand	1	272.64822	161.52966	58.157494	1.5501611	66.783241	55.417229	45.151615	83.190598	724.62012	3092.6438	141.80777	253.1818323	113.5167618	39.32462781
2000	HIC	NZL	New Zealand	0	214.57637	141.36971	57.224464	1.686506	55.689926	63.855618	40.474728	89.483795	548.36914	3129.2424	132.28484	309.5834808	79.64607772	45.25869909
2000	HIC	NOR	Norway, Svalbard and Jan Mayen	1	111.08414	101.94965	2.3652024	2.2702041	73.690475	75.855202	34.811741	76.796295	830.44098	2723.6248	141.87833	35.47707371	308.5826178	172.1672709
2000	HIC	NOR	Norway, Svalbard and Jan Mayen	0	92.172432	89.600136	2.533752	2.6912515	79.74408	102.30985	31.093767	82.547157	786.13049	2754.2212	138.77773	37.98024242	360.3600827	124.4135271
2000	HIC	PRT	Portugal	1	148.01425	145.79445	36.844879	3.5467758	80.372063	98.583771	8.7940388	75.123535	870.0451	3375.3435	130.57526	297.2255061	81.94237745	246.8317014
2000	HIC	PRT	Portugal	0	126.97682	129.95859	38.768681	4.0271187	89.300606	127.98326	7.921093	80.59671	798.82837	3401.4138	129.8493	256.2272459	73.58777775	383.4590324
2000	Asia	KOR	Republic of Korea	1	154.00606	141.60341	50.431366	0.6478104	49.293682	27.05831	16.599808	73.043137	495.24023	3067.917	95.019936	141.6145328	46.14997571	150.4831118
2000	Asia	KOR	Republic of Korea	0	113.48779	138.4635	45.155766	0.6232236	52.948009	42.220142	16.574518	72.970657	472.26044	2886.0369	80.429169	115.0388961	45.7171956	253.770251
2000	Asia	SGP	Singapore	1	208.34708	153.89455	36.890083	2.9004135	19.432501	89.782166	7.6196752	61.061516	507.55408	3757.1062	54.563442	54.2114719	119.128525	23.53831515
2000	Asia	SGP	Singapore	0	175.7309	145.01447	32.782055	2.884057	20.244188	111.94642	7.3693447	61.289558	454.15424	3646.7913	50.028236	44.74222184	153.4988351	18.83099896
2000	HIC	ESP	Spain	1	124.41893	119.46782	26.234034	7.1423864	58.541115	99.919197	23.242201	76.334908	876.79346	3212.552	162.62575	145.6717882	68.54707926	213.7133585
2000	HIC	ESP	Spain	0	109.99393	107.3229	27.774431	8.1687603	65.971016	129.17416	20.951403	82.012474	793.0094	3237.252	160.30312	107.5731346	51.30999434	374.8992862
2000	HIC	SWE	Sweden	1	137.94211	132.68839	9.9513531	2.3461249	70.147667	70.988266	65.023178	77.73558	888.84308	2893.3821	298.54892	321.3963728	179.3799129	40.07808654
2000	HIC	SWE	Sweden	0	106.86545	114.81633	11.119232	2.6483653	88.73056	93.741737	58.45615	82.493095	855.93732	2807.4971	291.52789	358.436362	132.230567	41.78908612
2000	HIC	CHE	Switzerland, Liechtenstein	1	144.81905	134.86296	4.8277645	4.5149937	59.585728	98.249626	81.637741	68.665718	1016.6036	2744.916	126.98283	76.41386306	39.26896763	334.4711534
2000	HIC	CHE	Switzerland, Liechtenstein	0	119.49265	119.8265	5.1066332	5.2020564	65.887093	124.78244	72.760597	72.584869	956.10632	2769.0156	124.74404	113.4857109	52.78780991	239.2421551
2000	HIC	GBR	United Kingdom	1	120.68939	139.30351	34.943378	4.3475399	36.446835	153.44957	36.51432	68.516869	887.56586	3534.6196	174.43863	172.0273409	26.04674614	345.7956347
2000	HIC	GBR	United Kingdom	0	100.15926	122.93792	36.513992	4.9960084	40.103592	185.81911	32.689461	71.611267	831.48212	3768.4441	165.47891	109.6978217	25.00451856	394.2641182
2000	HIC	USA	USA, Puerto Rico and US Virgin Islands	1	90.856201	132.38176	18.141495	3.7196779	38.52512	264.21091	89.290665	81.013123	825.93347	2995.9187	149.42509	105.9460495	65.66429682	277.9705239
2000	HIC	USA	USA, Puerto Rico and US Virgin Islands	0	81.8899	122.19653	21.805296	4.3188114	47.584606	313.03064	80.749855	86.205826	810.6153	3034.658	159.74077	70.29134253	59.44772334	316.4887486
2010	HIC	AUS	Australia	0	138.82275	120.8653	22.1348	4.4053764	87.143646	169.54768	35.506111	85.237022	800.44629	2919.1914	135.28766	81.59886403	47.59029622	231.6618119
2010	HIC	AUS	Australia	1	156.75948	130.10384	19.378067	3.7803507	53.836079	120.16269	39.760921	79.2789	785.47571	2882.7881	121.96072	57.96501779	47.81671951	312.469752
2010	HIC	AUT	Austria	0	112.29446	103.89571	6.6103644	4.0921454	109.66609	124.87753	45.179657	68.044769	727.41223	3117.5986	81.521965	278.4852779	91.65769803	78.60912664
2010	HIC	AUT	Austria	1	134.07495	117.62232	6.2503548	3.5690949	87.993935	95.449593	50.372704	63.252792	777.98779	3084.0129	82.628593	343.5863986	62.73309779	62.67347708
2010	HIC	BEL	Belgium	0	109.60388	116.84699	3.6786447	3.6388621	75.930275	204.43196	46.350525	98.541626	743.1217	4321.1167	72.912041	123.9816467	56.01098199	307.9456284
2010	HIC	BEL	Belgium	1	129.45607	130.76671	3.4093521	3.0592546	70.393951	160.17274	51.887321	94.796906	848.88361	4386.5562	75.689102	86.7593889	41.52136745	427.7791604
2010	Asia	BRN	Brunei Darussalam	0	87.397346	149.27086	7.0795689	4.6099582	14.834725	216.38695	6.0908847	69.130196	460.83209	1844.098	24.225843	43.89293013	68.6562275	140.7735162
2010	Asia	BRN	Brunei Darussalam	1	99.911453	156.21239	7.7686539	4.7720242	14.52899	175.26044	6.2998629	68.923637	501.81952	1897.7913	24.463322	41.27571941	77.30957521	135.0749735
2010	HIC	CAN	Canada	0	129.30704	139.78021	7.0166655	7.3255506	58.401768	156.87065	89.625923	77.2052	848.00806	3270.7629	143.9976	69.7841577	46.75667394	269.7434081
2010	HIC	CAN	Canada	1	153.05696	158.34947	6.6501212	6.3131089	47.611103	120.29539	99.921761	71.647514	930.90149	3246.2351	148.80614	48.56330887	48.46269309	317.5812184
2010	HIC	CYP	Cyprus	0	137.11044	164.90126	43.838234	6.8106556	108.81567	137.17429	44.972866	83.452736	820.06836	2402.7351	120.3242	85.58853695	84.82315417	176.5982802
2010	HIC	CYP	Cyprus	1	162.40067	185.77074	41.236355	6.0447121	98.318932	106.4399	49.772495	77.677956	875.7663	2383.824	120.6095	91.87862472	71.72743356	243.2824529
2010	HIC	DNK	Denmark	0	112.58035	100.97889	18.978548	2.5285239	62.830345	84.371971	34.369675	90.062965	930.82861	3444.6563	116.02061	395.1455438	121.521275	52.15323599
2010	HIC	DNK	Denmark	1	133.47656	112.9466	17.705496	2.1877427	56.10442	61.610359	38.407429	83.950027	1003.3715	3410.5505	118.02145	439.351859	94.1391337	62.24757492
2010	HIC	FIN	Finland	0	109.4678	98.370537	5.4333019	2.8969276	54.323368	73.933449	45.797718	95.800026	1108.0352	2862.0327	251.1532	141.6257396	23.46479506	278.1801152
2010	HIC	FIN	Finland	1	134.87822	111.95744	5.0321984	2.540709	41.445534	50.704315	50.990818	88.982346	1183.7695	2827.3594	251.82544	94.71769996	21.76015483	320.2798852
2010	HIC	FRA	France, Monaco	0	110.59152	169.24455	13.125037	3.2213998	49.775745	97.037376	18.877249	96.989616	935.91113	3281.4382	120.42651	93.89985202	46.01931289	290.4538535
2010	HIC	FRA	France, Monaco	1	120.90807	181.30128	12.178112	2.7830491	37.578102	70.681084	20.90148	78.138153	857.97485	2822.1057	122.48158	72.59438758	46.41898194	444.8874394
2010	HIC	DEU	Germany	0	151.77888	175.89888	9.3324137	3.0301588	56.779251	119.38448	62.023224	77.095924	1017.9942	3326.1897	75.722328	113.5270951	344.5457844	78.78684544
2010	HIC	DEU	Germany	1	182.15224	201.93269	8.4244413	2.5560436	44.512634	83.018646	70.570549	74.307487	1093.4371	3412.9595	75.494614	78.82865082	427.514628	58.47750726
2010	HIC	GRC	Greece	0	124.12564	129.59557	18.433517	6.3043313	83.50824	96.712852	21.325071	86.878983	938.50854	3927.9653	82.478218	209.6632086	269.035253	45.82301811
2010	HIC	GRC	Greece	1	144.96205	145.06993	17.407503	5.7484603	78.661598	73.897705	23.770212	80.821297	1001.9888	3887.1689	85.077301	161.474593	432.0268808	43.61066047
2010	HIC	ISL	Iceland	0	92.403442	74.490395	4.4705577	0.64352256	54.0448	124.99861	49.656445	104.88444	1010.6261	2971.3704	261.75064	16.87892068	221.4814037	65.26339031
2010	HIC	ISL	Iceland	1	114.20163	84.107109	4.1522045	0.57517284	47.938648	100.51331	55.506626	97.339569	1093.4061	2943.1333	265.38312	24.11621585	291.7884862	59.3126054
2010	HIC	IRL	Ireland	0	98.995247	127.94511	17.164097	1.6373296	38.454445	151.4272	35.212151	92.327682	878.25061	3750.385	136.64633	71.82223094	237.6989396	27.75089877
2010	HIC	IRL	Ireland	1	114.11954	141.31181	15.962269	1.4294124	33.455822	120.15189	39.233578	85.894608	902.17743	3709.7063	137.99678	50.76879689	278.0813342	30.11617582
2010	HIC	ITA	Italy	0	204.8378	136.36047	15.71965	0.92574114	57.455776	59.037506	15.169219	84.550903	863.24982	3022.2859	93.904793	71.8183208	143.6714626	302.2160063
2010	HIC	ITA	Italy	1	227.11319	152.2932	14.946572	0.82897741	51.449787	46.678482	16.935787	76.889648	880.68579	2761.0613	102.38533	60.65996311	103.7443589	426.1949314

Appendix A: Full Data Set

2010	Asia	JPN	Japan	0	108.22189	216.14207	56.899757	2.160171	43.360737	112.18379	20.323427	64.751068	499.41745	2196.3184	83.622955	287.3950425	42.30940205	120.1866883
2010	Asia	JPN	Japan	1	143.60327	235.29352	65.309776	2.5685346	41.110466	85.407616	21.319603	68.538383	611.79315	2579.1563	97.622421	451.7099969	44.58434253	118.2368252
2010	HIC	LUX	Luxembourg	0	101.71802	105.29293	11.524368	2.49248	74.674973	146.05989	48.702518	80.890305	875.37512	3526.5974	75.436661	112.1511778	277.9723169	47.24559132
2010	HIC	LUX	Luxembourg	1	118.39726	116.85345	10.72389	2.1659908	68.69635	113.10076	54.232204	75.087036	934.83392	3495.3875	76.579056	74.64477413	328.8754419	34.53612412
2010	HIC	MLT	Malta	0	115.87161	92.907547	16.629944	8.740593	87.793488	182.77658	70.833893	90.154648	1040.6005	3156.5212	167.79219	213.2543034	96.20004283	58.69921152
2010	HIC	MLT	Malta	1	135.76419	105.42561	15.414862	7.7011871	81.192978	143.24155	78.826424	84.887886	1149.2172	3124.4683	170.75182	302.331483	81.84273236	58.33947441
2010	HIC	NLD	Netherlands	0	136.01802	142.03874	37.880409	15.996328	77.129761	200.50316	146.56447	81.430672	942.92761	3246.446	120.34307	102.6218649	52.30481608	338.6032818
2010	HIC	NLD	Netherlands	1	156.00677	151.85458	32.883099	14.013341	70.872177	173.45276	162.71309	80.071655	1019.7216	3296.1702	123.96318	65.24083678	40.4624599	412.6060599
2010	HIC	NZL	New Zealand	0	203.08235	131.3688	42.712078	2.054219	49.662956	67.405983	36.405117	95.942673	562.57703	3199.4907	114.38547	256.0282459	93.37265516	38.22803639
2010	HIC	NZL	New Zealand	1	258.20181	150.47093	43.421249	1.8943386	59.617767	58.52277	40.62941	89.208183	742.83801	3173.9241	121.84873	307.1045626	62.77019779	44.34704244
2010	HIC	NOR	Norway, Svalbard and Jan Mayen	0	92.75251	83.008415	0.94205385	2.6522877	68.035492	105.28139	28.291494	83.93605	809.51569	2971.2676	117.7985	39.69462718	121.5922768	304.4790117
2010	HIC	NOR	Norway, Svalbard and Jan Mayen	1	111.07014	94.128227	0.87635136	2.2352827	62.921833	78.775162	31.722685	78.052193	854.83472	2935.4724	120.30702	39.60920066	81.98829365	348.8231187
2010	HIC	PRT	Portugal	0	123.60711	123.20934	33.584518	3.473598	78.956413	130.75531	7.1638064	79.995911	815.50995	3407.3677	112.09917	254.1549746	200.4731894	95.58113
2010	HIC	PRT	Portugal	1	143.55954	137.78847	31.838121	3.0489988	70.898842	100.88745	7.978941	74.500534	888.02997	3377.074	112.91473	421.1943536	169.7288755	84.90402736
2010	Asia	KOR	Republic of Korea	0	108.98799	124.996	36.068375	0.67077637	40.552727	40.239349	14.344538	70.160065	493.87964	2859.1116	61.867115	84.76386878	44.74327608	157.1828063
2010	Asia	KOR	Republic of Korea	1	147.81392	127.79362	40.311871	0.69536829	37.71653	25.65185	14.353333	70.206551	517.81421	3037.8237	73.182602	77.74548684	46.93110849	271.5005448
2010	Asia	SGP	Singapore	0	151.5537	131.36411	22.617411	3.9632952	23.393072	114.15365	6.5692329	86.25061	475.16336	2647.4302	33.101646	39.25946677	123.5867374	20.50034701
2010	Asia	SGP	Singapore	1	179.78038	139.37749	25.481451	3.9812694	22.458994	91.421913	6.8724475	85.890518	531.50629	2721.3833	36.169983	30.69054172	134.7186318	15.7089256
2010	HIC	ESP	Spain	0	101.53709	102.00794	22.848114	8.5801945	63.706253	126.98962	19.332083	82.319885	816.99817	3113.8892	138.64375	103.3562343	64.58832597	218.3745179
2010	HIC	ESP	Spain	1	114.5921	113.21651	21.584639	7.4984307	56.473423	98.242676	21.484127	76.447685	903.48273	3083.9258	141.09637	76.36073073	50.33379971	368.6210923
2010	HIC	SWE	Sweden	0	108.98852	107.61472	8.1307516	3.0309141	75.007378	96.403137	54.22554	84.408211	874.81073	3127.8467	256.15945	144.8246449	46.93556259	311.320987
2010	HIC	SWE	Sweden	1	140.03665	124.12965	7.2556591	2.6810267	59.311497	73.538139	60.30545	79.560883	908.10583	3223.2559	261.70264	101.1650923	49.18606207	343.5133018
2010	HIC	CHE	Switzerland, Liechtenstein	0	121.26773	113.861	3.4249625	5.0352001	56.948917	117.40124	69.720734	70.829819	994.0155	2761.8665	112.69245	57.85298391	41.49486586	317.2121276
2010	HIC	CHE	Switzerland, Liechtenstein	1	146.25237	127.86882	3.2348995	4.4145555	51.507477	93.042747	78.235161	66.961388	1056.3242	2726.6533	114.49075	87.99679534	53.81309671	244.6883638
2010	HIC	GBR	United Kingdom	0	106.79031	118.46258	39.104225	5.2621541	35.330257	172.55777	32.57494	70.890526	862.84344	3770.0779	158.39336	121.5022718	25.40311796	325.8160369
2010	HIC	GBR	United Kingdom	1	128.05817	134.04881	37.272766	4.5814557	32.164341	143.84138	36.3451	67.86953	920.23492	3533.7256	165.94646	80.91800163	23.16727928	373.987593
2010	HIC	USA	USA, Puerto Rico and US Virgin Islands	0	77.611382	117.36357	18.014244	4.4758754	44.456772	308.32169	76.652885	86.656631	829.84277	2880.5049	145.44727	265.230369	83.04789107	67.16097462
2010	HIC	USA	USA, Puerto Rico and US Virgin Islands	1	85.610626	126.75581	14.936326	3.8523872	35.993305	262.2959	84.853745	81.422737	845.14313	2841.5598	135.76813	301.8841784	57.2017377	68.08433031

Appendix B: RCode

```
setwd("H:/CPC STA6950")
install.packages("xtable")
install.packages("ggplot")
library(ggplot2)
library(xtable)
library(MASS)
library(plyr)

#combine three years into one large data set
y1990<-read.csv("H:/CPC STA6950/STA6950 Project Data Year 1990.csv",
        header = T)
y1990$year<-c(1)
y2000<-read.csv("H:/CPC STA6950/STA6950 Project Data Year 2000.csv",
        header = T)
y2000$year<-c(2)
y2010<-read.csv("H:/CPC STA6950/STA6950 Project Data Year 2010.csv",
        header = T)
y2010$year<-c(3)

#rename columns Year 1990
colnames(y1990)[colnames(y1990) %in% c("v01_wt_median","v02_wt_median")]<-c("fruits","NS_Veg")
colnames(y1990)[colnames(y1990) %in% c("v05_wt_median","v06_wt_median")]<-c("Beans_Leg","Nuts_Seeds")
colnames(y1990)[colnames(y1990) %in% c("v10_wt_median","v15_wt_median")]<-c("UP_Red_Meats","SS_Bev")
colnames(y1990)[colnames(y1990) %in% c("v16_wt_median","v23_wt_median")]<-c("fruitJuice","Protein")
colnames(y1990)[colnames(y1990) %in% c("v36_wt_median","v41_wt_median")]<-c("Calcium","potassium")
colnames(y1990)[colnames(y1990) %in% c("v57_wt_median")]<-c("Milk")
colnames(y1990)[colnames(y1990) %in% c("female......1.")]<-c("Gender")
colnames(y1990)[colnames(y1990) %in% c("country.name")]<-c("countryname")

#rename columns Year 2000
colnames(y2000)[colnames(y2000) %in% c("v01_wt_median","v02_wt_median")]<-c("fruits","NS_Veg")
colnames(y2000)[colnames(y2000) %in% c("v05_wt_median","v06_wt_median")]<-c("Beans_Leg","Nuts_Seeds")
```

```
colnames(y2000)[colnames(y2000) %in% c("v10_wt_median","v15_wt_median")]<-c("UP_Red_Meats","SS_Bev")

colnames(y2000)[colnames(y2000) %in% c("v16_wt_median","v23_wt_median")]<-c("fruitJuice","Protein")

colnames(y2000)[colnames(y2000) %in% c("v36_wt_median","v41_wt_median")]<-c("Calcium","potassium")

colnames(y2000)[colnames(y2000) %in% c("v57_wt_median")]<-c("Milk")

colnames(y2000)[colnames(y2000) %in% c("female......1.")]<-c("Gender")


#rename columns Year 2010

colnames(y2010)[colnames(y2010) %in% c("v01_wt_median","v02_wt_median")]<-c("fruits","NS_Veg")

colnames(y2010)[colnames(y2010) %in% c("v05_wt_median","v06_wt_median")]<-c("Beans_Leg","Nuts_Seeds")

colnames(y2010)[colnames(y2010) %in% c("v10_wt_median","v15_wt_median")]<-c("UP_Red_Meats","SS_Bev")

colnames(y2010)[colnames(y2010) %in% c("v16_wt_median","v23_wt_median")]<-c("fruitJuice","Protein")

colnames(y2010)[colnames(y2010) %in% c("v36_wt_median","v41_wt_median")]<-c("Calcium","potassium")

colnames(y2010)[colnames(y2010) %in% c("v57_wt_median")]<-c("Milk")

colnames(y2010)[colnames(y2010) %in% c("female......1.")]<-c("Gender")


all_years<-rbind(y1990,y2000,y2010)


#histogram for response variables

#hist(all_years$Rate.Cancer)

hist(all_years$Rate.stroke, main=" ")

hist(all_years$Rate.Diabetes, main=" ")

##Summary stats for stroke and diabetes rate

s=summary(all_years$Rate.stroke)

d=summary(all_years$Rate.Diabetes)


#####response var treated as a continuous random var.

#####STROKE

##LM model for stroke

modelStr<-lm(formula = Rate.stroke ~ Gender+year+fruits+NS_Veg+Beans_Leg+

           Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice

          +Protein+Calcium+potassium+Milk,

          data = all_years)

summary(modelStr)
```

```
par(mfrow=c(2,2))
plot(modelStr)
extractAIC(modelStr)

#BC transformation
bcStr=boxcox(modelStr,lambda = seq(-2,2))
Strbest.lam=bcStr$x[which(bcStr$y==max(bcStr$y))]

##LM model with BC trans
bcmodelstr<-lm((Rate.stroke)^Strbest.lam ~ Gender+year+fruits+NS_Veg+Beans_Leg+
         Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
        +Protein+Calcium+potassium+Milk,
        data = all_years)
summary(bcmodelstr)
par(mfrow=c(2,2))
plot(bcmodelstr)
extractAIC(bcmodelstr)
hist((all_years$Rate.stroke)^Strbest.lam,main = " ")
BIC(bcmodelstr)

#LM log model
log_modelStr<-lm(formula = log(Rate.stroke) ~ Gender+year+fruits+NS_Veg+Beans_Leg+
          Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
         +Protein+Calcium+potassium+Milk,
         data = all_years)
summary(log_modelStr)
par(mfrow=c(2,2))
plot(log_modelStr)
extractAIC(log_modelStr)
hist(log(all_years$Rate.stroke),main = " ")
BIC(log_modelStr)

#gamma model link inverse, log trans
```

```
gamma_modelstr=glm(log(Rate.stroke) ~ Gender+year+fruits+NS_Veg+Beans_Leg+
          Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
         +Protein+Calcium+potassium+Milk,
         data = all_years, family = Gamma(link="inverse"))
summary(gamma_modelstr)
par(mfrow=c(2,2))
plot(gamma_modelstr)
extractAIC(gamma_modelstr)
##############End Stroke

#####Diabetes
#LM model no trans
modelDiab<-lm(formula = Rate.Diabetes ~ Gender+year+fruits+NS_Veg+Beans_Leg+
        Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
       +Protein+Calcium+potassium+Milk,
       data = all_years)
summary(modelDiab)
par(mfrow=c(2,2))
plot(modelDiab)
extractAIC(modelDiab)

#BC TRans
bcDiab=boxcox(modelDiab,lambda = seq(-2,2))
Diabbest.lam=bcDiab$x[which(bcDiab$y==max(bcDiab$y))]

#LM model with BC trans
bcmodeldiab<-lm((Rate.Diabetes)^Diabbest.lam ~ Gender+year+fruits+NS_Veg+Beans_Leg+
         Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
        +Protein+Calcium+potassium+Milk,
        data = all_years)
summary(bcmodeldiab)
par(mfrow=c(1,1))
plot(bcmodeldiab)
```

```
extractAIC(bcmodeldiab)
hist((all_years$Rate.Diabetes)^Diabbest.lam)
BIC(bcmodeldiab)

#LM Log Model
log_modeldiab<-lm(formula = log(Rate.Diabetes) ~ Gender+year+fruits+NS_Veg+Beans_Leg+
          Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
         +Protein+Calcium+potassium+Milk,
         data = all_years)
summary(log_modeldiab)
par(mfrow=c(2,2))
plot(log_modeldiab)
extractAIC(log_modeldiab)
BIC(log_modeldiab)

#Gamma model link inverse, log trans
gamma_modeldiab=glm(log(Rate.Diabetes) ~ Gender+year+fruits+NS_Veg+Beans_Leg+
           Nuts_Seeds+UP_Red_Meats+SS_Bev+fruitJuice
          +Protein+Calcium+potassium+Milk,
          data = all_years, family = Gamma(link="inverse"))
summary(gamma_modeldiab)
par(mfrow=c(2,2))
plot(gamma_modeldiab)
extractAIC(gamma_modeldiab)

##########END Diabetes

##best models
summary(bcmodelstr)
summary(bcmodeldiab)
summary(log_modelCan)

##########new Models with significant Explanatory Variables only
```

```
####STROKE####
beststroke<-lm((Rate.stroke)^Strbest.lam ~ Nuts_Seeds+Calcium
             +Beans_Leg+UP_Red_Meats,
             data = all_years)
summary(beststroke)
par(mfrow=c(2,2))
plot(beststroke)
extractAIC(beststroke)
BIC(beststroke)

####Diab####
bestdiab<-lm((Rate.Diabetes)^Diabbest.lam ~ year+Nuts_Seeds
                    +Milk+Beans_Leg,
              data = all_years)
summary(bestdiab)
par(mfrow=c(2,2))
plot(bestdiab)
extractAIC(bestdiab)
BIC(bestdiab)

###export model summaries to Latex###
xtable(beststroke)
xtable(bestdiab)

#########PCA Analysis #####################
#PCA for 1990
pr1990<-prcomp(y1990[,5:15],scale=TRUE)
pr1990
summary(pr1990)
par(mfrow=c(1,1))
plot(pr1990,type="l")  #Scree plot
biplot(pr1990,scale=0)
```

```
#PCA for 2000
pr2000<-prcomp(y2000[,5:15],scale=TRUE)
pr2000
summary(pr2000)
par(mfrow=c(1,1))
plot(pr2000,type="l")  #Scree plot
biplot(pr2000,scale=0)

#PCA for 2010
pr2010<-prcomp(y2010[,5:15],scale=TRUE)
pr2010
summary(pr2010)
par(mfrow=c(1,1))
plot(pr2010,type="l")  #Scree plot
biplot(pr2010,scale=0)



plot(modelCan)#PCA for 2010
pr2010<-prcomp(y2010[,5:15],scale=TRUE)
pr2010
summary(pr2010)
par(mfrow=c(1,1))
plot(pr2010,type="l")  #Scree plot
biplot(pr2010,scale=0)
##All Years
prAll<-prcomp(all_years[,5:15],scale=TRUE)
prAll
summary(prAll)
par(mfrow=c(1,1))
plot(prAll,type="l")  #Scree plot
biplot(prAll,scale=0)

###########End
```