DSC 424 Group Project

Kate Burns, Abigail Keller, Ronaldlee Ejalu

Executive Summary

Diabetes is a chronic condition in the United States and has significantly increased in recent years, especially among younger generations and racial minorities. One in three adults in the US have prediabetes and more than one in ten have been diagnosed with diabetes. Type II diabetes found in youth aged 10 to 19 increased for most racial categories and especially so in non-Hispanic blacks.[1] Type II diabetes is the most common among all subgroups as it accounts for more than 90% of all cases. There are an estimated 46.5% of those with diabetes who have not yet been diagnosed. Those with prediabetes can avoid diabetes diagnosis with weight loss and regular physical activity but if left unchecked, 15 to 30 percent will develop type II diabetes within five years.[2] Diabetes risk can be attributed to many factors which can be used to predict whether someone has or is likely to develop diabetes in their lifetime. Some of these factors include family history, race, age, socioeconomic status, and various health measures. There are additionally many health complications that are often related to diabetes including blindness, kidney failure, heart disease, and stroke.[2]

Various studies and prior research have been completed to better understand and more efficiently diagnose diabetes which can help to lessen negative outcomes among high risk individuals. One study found that physical inactivity, smoking and low birth weight are linked to the likelihood of developing diabetes. Additionally, the prevalence of certain diseases including diabetes aligned with an increase in body weight, and for certain racial and socioeconomic groups.[7] The national health and nutrition examination survey is a questionnaire containing a multitude of survey questions ranging from demographic, medical history and lifestyle inquiries. This survey was utilized to continue the research into diabetes and what variables are predictive of diabetes risk. The research and analysis methods used throughout this summary will attempt to identify those attributes found to be significant when assessing diabetes risk. Through this analysis, we hoped to find which factors were most influential to an individual's vulnerability to diabetes and whether or not these could be used to predict overall risk.

Various methods were used throughout our research to find which best fit our data and yielded meaningful insights or results. The methods used in our research include principal component analysis (PCA), canonical correlation analysis (CCA), and logistic regression. We found logistic regression is the most applicable as it allowed us to identify which variables are most predictive of diabetes risk. The goal of these methods overall includes useful interpretation of our data as well as finding the most important features when assessing diabetes risk. Principal component analysis and canonical correlation analysis provided beneficial details for our data and research. We used a version of PCA which allowed us to include all variables as these were significant to our overall results. Canonical correlation analysis was used to determine how certain demographic features are related to diabetes risk and the characteristics of that relationship.

The results throughout our analysis consistently demonstrated the impact of various demographic and socioeconomic factors when evaluating an individual's vulnerability to diabetes. We additionally found that family history (relatives with diabetes) or an individual's own risk assessment were useful to predicting their likelihood of either having or eventually developing diabetes. These results and

continued research could help to better predict diabetes risk early on and allow for treatments prior diabetes or prediabetes diagnosis. If an individual can determine their risk, they can better and more efficiently receive the necessary treatment to prevent the most harmful outcomes. Principal component analysis allowed for better interpretation of our data and provided useful groupings and we found similar associations when using canonical correlation. Some of the most important variable categories included demographic, diabetes risk and activity level and these were demonstrated throughout our analysis.

Logistic regression was also applied to the NHANES survey results to predict the likelihood that a person may be at risk for diabetes. In logistic regression, the predictive variable included responses to the question “has diabetes risk or does not have diabetes risk”. The logistic regression model proved to be a strong model of prediction for diabetes risk given the variables included in the survey results. The results showed that having a relative with diabetes as well as feeling at risk for diabetes or prediabetes were both strong predictors for being at risk of diabetes. Additionally, socioeconomic variables provided strong results around the prediction of diabetes risk. Salary, ethnicity, and education all played large roles in predicting diabetes risk. Not surprisingly, higher BMI and weight levels also provided strong results. Losing weight, eating healthy, and exercising can all help prevent the onset of diabetes.

The responses from the survey were sparsely populated as many respondents either left sections blank or refused to answer certain questions. This caused difficulties when searching for a sizable sample of complete data and affected measures of sampling adequacy. We were able to rectify this using various strategies, but we acknowledge the likely impacts to our results overall. The results are significant but may benefit from future testing with a more rigorous sampled data. Moreover, it may be beneficial to include additional variables to the dataset from the NHANES survey results. Certain variables from the examination and labs section of the questionnaire may allow for a more complete picture of the population health of the respondents. A richer dataset may yield better results through logistic regression analysis, as well as the relationship between groupings through principal component and canonical correlation analysis.

The overall objective of our research was to predict the vulnerability of an individual to diabetes and predict the probability that they could have or will develop diabetes in the future. Our research allowed us to find the most meaningful factors when assessing diabetes risk. Related research found that there are many risk factors that may contribute to the onset of diabetes including demographic and socioeconomic aspects. Our analysis provided meaningful insights into the data and was often consistent with these results. Our models provided predictors for diabetes risk given the variables included in the survey. Overall, we found that family history, socioeconomic and demographic factors as well as lifestyle all impact the likelihood an individual will develop the onset of diabetes. When a person felt at risk of diabetes, was classified as overweight, and had a family history of diabetes, they were more likely to be categorized as at risk for diabetes. Income and various demographic factors were also significant and predictive to how at risk an individual is for developing diabetes. These results are important to understanding diabetes and the identifiable factors which often accompany a higher risk for diabetes. The ability to identify these risk factors which are often predictive of diabetes can significantly benefit an individual and their overall diabetes risk.

Enhanced Logistic Regression for Diabetes Risk Prediction with Principal Component Analysis and Canonical Correlation

Kate Burns, Abigail Keller, Ronaldlee Ejalu

Abstract

Various factors influence an individual's vulnerability to diabetes and many of these conditions or features can be used to predict overall risk. Diabetes was among the top ten causes of deaths as of 2018 even though a large number of people living with diabetes remain undiagnosed. The relationship between diabetes and various risk factors has been widely studied and many results have identified factors which align with a higher risk of diabetes including body weight, activity level, and various ethnic and socioeconomic subgroups. In this study, three techniques were used to analyze diabetes risk using the National Health and Nutrition Examination Survey results. Statistical methods such as principal component analysis (PCA) and canonical correlation analysis (CCA) were utilized to gain a better understanding of the relationships between different variables within the NHANES survey results. Logistic regression was also applied to predict the likelihood that a person may be at risk for diabetes. PCA analysis as well as canonical correlation provided meaningful interpretations of our data and insights into the correlation among activity level, diabetes risk, body type, demographics and health conditions. Logistic regression proved to be a useful tool for identifying the variables most predictive of diabetes risk. Losing weight, eating well and exercising can all help prevent the onset of type II diabetes but other categories out of our control are equally as important to prediction including socioeconomic and demographic features.

Introduction

The prevalence of diabetes has been steadily increasing in the United States, including 8.2% of the US Population in 2018. Even more alarming, an estimated 88 million adults 18 years or older had prediabetes in 2018. Diabetes now stands among the top ten causes of deaths as of 2018.[1] By 2040, one in ten adults are expected to have diabetes, totaling 642 million adults. One of the main issues today is that 46.5% of adults with diabetes have not yet been diagnosed[2]. Because a large number of deaths in diabetic patients are due to late diagnosis, it is even more important that methods and techniques are created to assist with early diagnosis of diabetes.

It is estimated 382 million people around the world suffer from diabetes and its associated complications, which decreases people's quality of life. To alleviate the number of diabetic incidents, the population is always advised to integrate exercise, diet, blood sugar testing in its day to day activities. Clinicians and researchers continue to investigate how to enhance current prevention approaches, but diabetes self-management remains a critical element for improving self-care against the diabetic susceptible population.[4]

Literature Review

The relationship between a disease and their risk factors have been well documented throughout research. Physical inactivity, smoking and low birth weight are risk factors that have been linked particularly to the development of diabetes. Additionally, obesity has been found to be one of the main

risk factors associated with onset of type II diabetes. Paeratakul, Lovejoy, Ryan and Bray created a study around obesity-related chronic diseases in the US population according to gender, race and socioeconomic status. Multiple logistic regression analysis was conducted to estimate the odds ratio of each of the four chronic diseases studied – diabetes, high blood pressure, gallbladder disease and osteoarthritis – according to BMI category. The results showed that there was a significant increase in the prevalence of the four chronic diseases with increasing body weight in all gender, racial and socioeconomic groups[7].

Diabetes and their risk factors have also been known to be associated with low socioeconomic status. A population based study was completed by Connolly, Unwin, Sherriff, Bilous and Kelly, analyzing the hypothesis around the relationship between diabetics and their socioeconomic status. Their key findings suggested that the prevalence of type II diabetes increased in locations of low socioeconomic status. They also found that several of the risk factors associated with diabetes were much more common in deprived areas[5]. Another study focused on depression, obesity and diabetes, analyzing the psychological characteristics, social factors, and behaviors in health and disease risk. In particular, the research hypothesized the role risk factors have in relation to socioeconomic factors and chronic diseases. The findings suggested that type II diabetes had an inverse relationship with education, occupation and income across all age groups. Interestingly enough, men and women with less than a high school degree had nearly three times greater prevalence of diabetes than those with at least a high school degree[6].

This research work focuses on three techniques used to analyze diabetes risk: 1) Principal Component Analysis, 2) Canonical Correlation and 3) Logistic Regression. Data from the National Health and Nutrition Examination Survey (NHANES) was analyzed, focusing on one question around diabetes risk: Have you ever been told by a doctor or other health professional that you have health conditions or a medical or family history that increases your risk for diabetes? The goal is to find what factors influence an individual's vulnerability to diabetes and can these be utilized to assess and predict overall risk.

Methods

The National Health and Nutrition Examination Survey (NHANES) is a program through the National Center for Health Statistics (NCHS) that was created to measure the health and nutritional status of adults and children in the United States[3]. The dataset used for this research work contains thirty-eight variables, with 8,955 total samples (respondents to the survey). The variables include information around demographics, diabetes risk, body type, health conditions and activity level. There are twenty numerical variables, including fifteen continuous and five discrete. There are additionally eighteen categorical variables largely made up of binary and demographic attributes. Due to a lack of overall responses, the dataset contains many missing values throughout. The mice package in RStudio was applied and the imputation process was used to fill in missing data.

Various statistical methods such as principal component analysis (PCA) and canonical correlation were utilized within this research work around diabetes risk to gain a better understanding of the relationships between different variables within the NHANES survey results. Logistic regression was also applied to the NHANES survey results to predict the likelihood that a person may be at risk for diabetes.

Principal Component Analysis

Principal component analysis was initially performed using standard PCA methods but due to the importance of categorical attributes in our dataset, the final method using mixed PCA allowed us to include all variables. Prior to using principal component analysis, we checked for normality, linearity, multicollinearity, and homogeneity of variance where possible. When checking normality of numerical variables through skewness and kurtosis using the Jarque-Bera test, each variable had a significant p-value. Two instances of multicollinearity were found when using spearman's correlation. Some correlations were notably high including most weighed and weight with a correlation of 0.93, and weight 10 years ago and weight had a correlation of 0.83. We removed both most weighed and weight 10 years ago, as the information was redundant when also including weight. Sampling adequacy was checked and had an overall MSA (measure of sampling adequacy) of 0.66. This is not greater than the 0.7 threshold so while close, we cannot be fully confident in the sample size. This sampling adequacy is much higher than before data imputation. Bartlett's test of sphericity signified that there is enough shared variance in the data as the p-values were substantially low. When using Cronbach's Alpha, we found the overall raw alpha is 0.47 which is not an ideal measure of consistency. The dataset was split between qualitative and quantitative variables and the observations, levels, numerical variables and all variables are included in figure 1.

Canonical Correlation

Canonical correlation analysis was used to determine how the best linear combinations of demographics are related to the best linear combination of diabetic risk. Running Canonical Correlation Analysis, we needed to consider the assumption of linear regression: whether things are linear vs non-linear whether they are normal or they are sensitive to influential outliers, correlation. With Canonical correlation analysis being a descriptive procedure, we are only talking about relationships and association. Also, there is a need to consider the sample size, which typically the larger the sample size, the more reliable we can consider our information and the loadings that we get. Using the Yacca package in R, we built a Canonical Correlation Analysis model as indicated in figure 3.

Logistic Regression

In logistic regression, the dependent variable is binary, which means that it only contains classified data as 1 or 0. Within this research study, the dependent variable was binary in nature in regard to a yes or no question: has diabetes risk or does not have diabetes risk. The purpose of this logistic regression analysis was to find the best fit that describes the relationship between diabetes risk and the predictor variables around demographics, health conditions, body type and activity level.

Before beginning the logistic regression analysis, an exploratory analysis on correlation was completed. The variables were grouped into similar buckets around demographics, disease state and weight / activity levels. High correlation was found between the height and weight variables as well as BMI, which is a calculation of height over weight. Three weight variables and one height variable were removed while BMI remained in the dataset. BMI has been known to be a significant risk factor for diabetes, so the significance of the variable needed to be tested.

Once the dataset was tested for multicollinearity, a transformation of the data was needed. All of the continuous variables were scaled, and all of the categorical variables were converted into factors. A full model was run in order to see the significant variables within each factor. Once those variables were found, they were removed from the dataset.

Discussion and Results

Principal Component Analysis

Various iterations of principal component analysis were tested, and the ideal number of components were found using varimax rotation. The scree plot including the percentage of variance explained by each dimension is included as figure 2. The variance explained is consistently low after the elbow and additional dimensions did not yield meaningful groups. Five dimensions proved to be optimal as the groups were logical and variables below a 0.15 threshold were removed. A test sample containing 30% of the test data was used to compare to the 70% training sample. These values have been plotted in figure 4 and show that our samples were consistent between the training and test data.

The dimensions found when analyzing the squared loading after varimax-type rotation were closely related to the components in the original principal component analysis. Only five components were rotated as the groups became less meaningful in either direction. The loadings for these dimensions are included in figure 5 and they account for 17.2% of the total variance in our data. The first dimension included significant loadings for education level, annual income, smoking, chest pain, and ethnicity. The second dimension was largely related to overall diabetes risk, including diabetes relatives, whether the respondent feels at risk and overweight. The third dimension includes body type measurements (weight/height) and gender. The fourth dimension includes variables related to health conditions including age, blood pressure and blood cholesterol. The final component is related to activity level and consists of measures like walking/biking per day, moderate and vigorous activity, and time outdoors. Some variables did not have a notable significance to any dimension. While the variance explained was low overall, this is to be expected with a high proportion of categorical and binomial data. While we did not use these components in further regression analysis, they provided important insights into our data and which categories were significant to our research. These categories support our later results and identify some of the most important variables when assessing diabetes risk. These are consistent with prior research and our continued analysis using canonical correlation and logistic regression.

Dimension Names:

Dim1 = “Demographic”

Dim2 = “Diabetes Risk”

Dim3 = “Body Type”

Dim4 = “Health Conditions”

Dim5 = “Activity Level”

Canonical Correlation

We ended up with three varieties, sorted in the order of importance. Using Bartlett's Chi-Squared test, we were able to confirm that CV1, CV2 and CV3 as highly significant canonical variates. We renamed the variates as family history, diabetes risk and feel at risk. Speaking of the structural coefficients, these are placed in between the loadings of the variate and the variables. The structural loading's intercorrelations to interpret the variables to the variates, for example in variate 1 weight 10 years ago is the most important variable which has a strong positive relationship between variate 1 and itself , meaning that as one's weight they had 10 years ago increases, variate 1 also increases in the same direction, this means that people who remain overweight have a high diabetic risk, which raises the need for families with diabetic history to seek attention and get continuous follow-ups starting from childhood.

On the other hand, feel at risk diabetes is the most important variable in variate 1 and this demonstrates a strong negative correlation between variate 1 and feel at risk diabetes meaning that as feel at risk diabetes increases, variate 1 increases but in the opposite direction. When you look at the Canonical variate coefficients, in figure 3(a), those are the beta coefficients used to build the linear combinations of X's (Demographic) and linear combinations of Y's (Diabetic Risk).

Additionally, in figure 6 we can see the cancor is 0.18 meaning that there is 18 percent of overlapping variance in how the linear combinations of demographic variables predict or explain the linear combination of diabetes risk variables. If you think of this as Pseudo R squared: 100 – 18 = 82%, which represents error in the model meaning that we might not have everything that explains the demographics or diabetes variables.

In figure 7, there is a Graphical display of the structure correlation of the demographic scales on the first predictor canonical variate and of structural loadings of the Diabetes Risk scales on the first criterion canonical variate. Black bars correspond to positive correlations and white bars correspond to negative correlations. This is known as the Helios plot, which healthcare professionals would use to come up with strategic decisions on how to combat diabetes risk. This graph clears shows that individuals who are overweight, aged and with low levels of education are at risk of getting diabetes.

Logistic Regression

The overall dataset consisted of 8,955 total samples with 74% of respondents answering "yes" to feeling at risk for diabetes. Nineteen variables in the NHANES survey results made it to the final model through logistic regression. Figure 8 provides a view of the variables that were included in the final model.

Respondents that answered "yes" to having a relative in their family with diabetes was the variable that had the highest impact of increasing the log odds of being at risk for diabetes. Each unit increase of having a relative with diabetes increased the log odds of being at risk for diabetes by 1.53 as well as a high significant p-value. Respondents that answered "yes" to feeling at risk for diabetes or prediabetes were the next variable that significantly impacted the log odds of being at risk for diabetes. Each unit increase of feeling at risk for diabetes or prediabetes increased the log odds of actually being at risk for diabetes by 1.50 along with a highly significant p-value . In other words, respondents that felt their lifestyle and eating habits may be on the unhealthier side were correct in believing that they are at risk of having diabetes.

Respondents that had an annual income between $55K to $65K had a negative impact on diabetes risk. Each unit decrease in salary between $55K and $65K increases the log odds of being at risk for

diabetes by 0.53. Additionally, Non-Hispanic Black respondents had a significant impact on diabetes risk as well as respondents with a high school degree or GED. All of these conclusions support the research mentioned earlier around socioeconomic factors. Diabetes risk can be found in locations that may have a larger minority population that may be less educated.

Not surprisingly, respondents that answered "yes" to smoking every day have an impact on diabetes risk. For each unit increasing in smoking every day increased the log odds of being at risk for diabetes by 0.29. Being overweight as well as BMI were also significant factors for diabetes risk. Losing weight, eating well and exercising all can help prevent the onset of type II diabetes8.

The logistic model is also a good fit based on the difference between the Null deviance and the Residual deviance. The Null deviance was 7305 while the Residual deviance was 5286 with a difference of 2018.

Commonalities Between Methods

Throughout our analysis we found that demographic and socioeconomic factors impact an individual's overall diabetes risk. Some of these include education level, annual income, and ethnicity. We additionally found that respondents were often able to assess their own risk of diabetes and that it was predictive of their actual risk of diabetes. Family history or relatives with diabetes was also found to be significant when assessing overall risk. Both PCA and canonical correlation used similar groupings and shared cumulative variance. We found that principal component analysis clearly resulted in a diabetes risk component as well as a demographic dimension and found similar relationships when analyzing canonical correlations. These findings support the results from logistic regression and prior research that various demographic and risk factors including family history and weight are imperative to predicting diabetes risk.

Limitations

The National Health and Nutrition Examination survey is considered a preeminent source of data on diet and health; however, the data has some limitations. The sample allows for calculating estimates for people with Mexican ancestry, but the sample is too small to provide data for other Latino subgroups. Nationwide data from Mexican-origin Americans, African Americans, and European-origin Americans of all ages is included. Additionally, NHANES collects dietary and behavioral data through in-person interviews in English or Spanish where respondents recall the amount and type of food and beverages consumed over a 24 hour period and may be subjective.

All of the respondents have also visited a physician. In regard to predicting diabetes risk and analyzing common variables within the survey data, there is a large population that is being left out of the survey data. Because of the fact that 46.5% of adults with diabetes have not even been diagnosed[2], it is safe to assume that the majority of these adults have not seen a physician and are therefore not represented in the survey population.

The total NHANES dataset consists of 1,829 variables among six different categories: 1) demographics, 2) diet, 3) examination, 4) labs, 5) medications and 6) questionnaire. For purposes of simplicity, the main focus of this research came from the questionnaire dataset with the exception of some demographic and diet variables. Because of this, many of the answers to the questions may be subjective and not completely honest. Additionally, many respondents did not answer questions or did

not know the answer to a question. This caused low sampling measures in several sensitivity analyses, especially with regard to PCA. We rectified this through imputation of the data which substantially increased the overall sampling adequacy. Additionally, the research completed involving PCA included a smaller subset of variables as we only included numeric variables. Another PCA method was used and compared to our original output which allowed us to include a combination of both numeric and categorical data to reduce dimensions.

Future work

After working with the questionnaire dataset from the NHANES survey results, additional variables from the examination and labs section of the questionnaire can be added to get a more complete picture of the population health of the respondents. Even though a respondent may feel they are not at risk for diabetes, the lab and examination results may infer otherwise. By working with a richer dataset, the logistic regression analysis may yield better results as well as the relationship between variables through principal component and canonical correlation analysis. A more thorough questionnaire would allow for limitations to PCA to be mitigated and additional variables to be found that would capture a larger proportion of variance. In addition, more recent data collection or improved sampling could provide a more robust dataset to better verify the results found here. We also tend to perform cluster analysis once we are able to attain a richer dataset so that we can identify groups with similar and dissimilar characteristics.

Conclusion

The results from our research determined that family history, prediabetes, poverty, ethnicity, lack of education, unhealthy lifestyle and eating habits were independent predictors of diabetes risk. Additionally, diabetes can lead to loss of vision, blood pressure, heart disease, kidney disease, and nerve damage, so there is an urgent need for ethnically appropriate diabetes education and screening programs targeted towards the underserved large minority populations and health services planning that aims to reduce the risk of diabetes in these large minority populations. In addition to appropriate diabetes education, there is a greater need to sensitize the masses on the dangers of consuming white potatoes, high-fat diary, red/processed meats, high fat foods and sugary soda beverages since these items are associated with a high risk of diabetes, they should instead focus on increasing intake of several key food groups including whole grains, low-fat diary, nuts/seed, fruits and vegetables. It is also recommended that families with diabetic history seek attention and get continuous follow-ups starting from childhood. This research serves to better understand the risk factors related to diabetes and reaffirm the most important features when evaluating an individual's overall risk. Careful attention to these elements can help predict and treat diabetes symptoms before diagnosis and more detrimental consequences.

References

[1] US Department of Health and Human Services (2020). National Diabetes Statistics Report 2020. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

[2] Randall MD, Tamara(2016). How Many People Have Diabetes, https://www.diabetesdaily.com/learn-about-diabetes/basics/what-is-diabetes/how-many-people-have-diabetes/

[3] National Centers for Disease Control (2020). About the National Health and Nutrition Examination Survey, https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm

[4] Goyal, A., Gupta, Y., Singla, R., Kalra, S., & Tandon, N. (2020). American Diabetes Association "Standards of Medical Care—2020 for Gestational Diabetes Mellitus": A Critical Appraisal. Diabetes Therapy, 11(8), 1639-1644

[5] Connolly, V., Unwin, N., Sherriff, P., Bilous, R., & Kelly, W. (2000). Diabetes prevalence and socioeconomic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas. Journal of Epidemiology & Community Health, 54(3), 173-177.

[6] Everson, S. A., Maty, S. C., Lynch, J. W., & Kaplan, G. A. (2002). Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. Journal of psychosomatic research, 53(4), 891-895.

[7] Paeratakul, S., Lovejoy, J. C., Ryan, D. H., & Bray, G. A. (2002). The relation of gender, race and socioeconomic status to obesity and obesity comorbidities in a sample of US adults. International journal of obesity, 26(9), 1205-1210.

8 Mayo Clinic (2021). Type 2 Diabetes https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Appendix

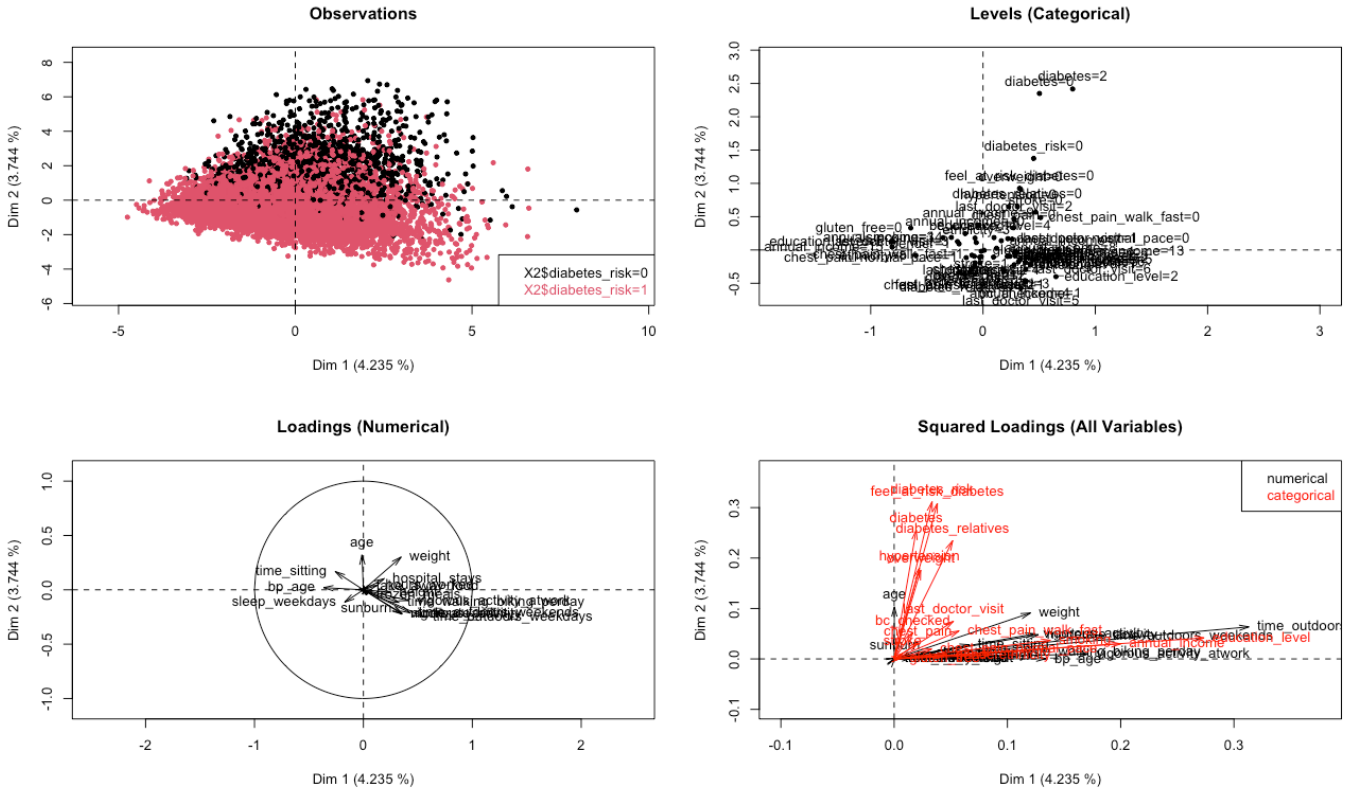

Figure 1: Observations, Levels, Numerical and All Variables

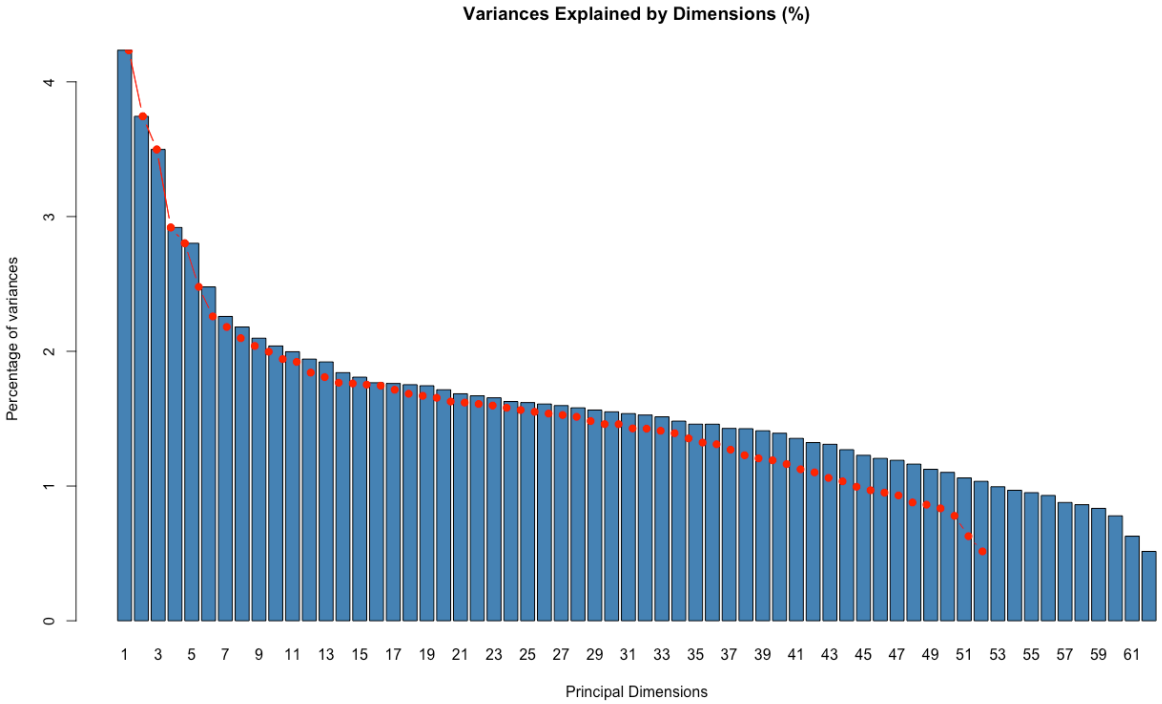

Figure 2: Scree Plot of variance explained by each dimension

Figure 3: Results of the Yacca Package

Includes parts a, b and c:

```
Canonical Correlation Analysis - Summary


Canonical Correlations:

      CV 1       CV 2       CV 3
0.18204234 0.09432901 0.03789932

Shared Variance on Each Canonical Variate:

       CV 1        CV 2        CV 3
0.033139413 0.008897962 0.001436358

Bartlett's Chi-Squared Test:

          rho^2      Chisq df    Pr(>X)
CV 1 3.3139e-02 3.9450e+02 21 < 2.2e-16 ***
CV 2 8.8980e-03 9.2863e+01 12 1.377e-14 ***
CV 3 1.4364e-03 1.2865e+01  5   0.02467 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Canonical Variate Coefficients:

        X Vars:
                         CV 1        CV 2        CV 3
gender             0.89429526 -0.18912433  1.63490395
education_level    0.09505831 -0.59726236 -0.03834936
ethnicity         -0.13795591  0.23242889  0.03576304
age                0.23828545  0.37250473 -0.35548042
annual_income      0.01385510 -0.34091056 -0.36409279
weight_10yearsago  0.90977477  0.08891969 -0.11358254
BMI                0.00221782  0.06546638 -0.01369131

        Y Vars:
                            CV 1      CV 2       CV 3
diabetes_relatives    -0.3761727 -2.161198 -0.7345859
diabetes_risk         -1.3779193  1.803018 -2.2305854
feel_at_risk_diabetes -1.3901202  0.278641  2.1102029
```

a) Canonical Variate Coefficients

```
Structural Correlations (Loadings):

        X Vars:
                          CV 1       CV 2        CV 3
gender              0.276898222 -0.1165562  0.83770809
education_level     0.112425737 -0.7899088 -0.15980203
ethnicity          -0.194129205  0.1650367 -0.02711457
age                 0.279225595  0.3753515 -0.34758188
annual_income       0.049453308 -0.5634578 -0.41083001
weight_10yearsago   0.838505959  0.0948980 -0.29943483
BMI                 0.001387613  0.1975230 -0.04556314

        Y Vars:
                            CV 1        CV 2       CV 3
diabetes_relatives    -0.5704860 -0.77441086 -0.2735573
diabetes_risk         -0.7708068  0.32117186 -0.5501868
feel_at_risk_diabetes -0.8591431 -0.02420265  0.5111627


Fractional Variance Deposition on Canonical Variates:

        X Vars:
                            CV 1         CV 2         CV 3
gender              7.667263e-02 0.013585351 0.7017548422
education_level     1.263955e-02 0.623955982 0.0255366901
ethnicity           3.768615e-02 0.027237111 0.0007351998
age                 7.796693e-02 0.140888759 0.1208131638
annual_income       2.445630e-03 0.317484705 0.1687812970
weight_10yearsago   7.030922e-01 0.009005631 0.0896612195
BMI                 1.925469e-06 0.039015354 0.0020759995

        Y Vars:
                            CV 1         CV 2       CV 3
diabetes_relatives    0.3254542 0.5997121726 0.07483361
diabetes_risk         0.5941431 0.1031513610 0.30270550
feel_at_risk_diabetes 0.7381269 0.0005857681 0.26128732


Canonical Communalities (Fraction of Total Variance
Explained for Each Variable, Within Sets):

        X Vars:
           gender   education_level         ethnicity               age     annual_income weight_10yearsago               BMI
       0.79201282        0.66213222        0.06565846        0.33966886        0.48871163        0.80175909        0.04109328

        Y Vars:
   diabetes_relatives         diabetes_risk feel_at_risk_diabetes
                    1                     1                     1
```

b) Structural Correlations

```
Canonical Variate Adequacies (Fraction of Total Variance
Explained by Each CV, Within Sets):


        X Vars:
     CV 1      CV 2      CV 3
0.1300722 0.1673104 0.1584798

        Y Vars:
     CV 1      CV 2      CV 3
0.5525748 0.2344831 0.2129421


Redundancy Coefficients (Fraction of Total Variance
Explained by Each CV, Across Sets):


        X | Y:
        CV 1         CV 2         CV 3
0.0043105147 0.0014887216 0.0002276337

        Y | X:
        CV 1         CV 2         CV 3
0.0183120031 0.0020864216 0.0003058612


Aggregate Redundancy Coefficients (Total Variance
Explained by All CVs, Across Sets):

        X | Y: 0.00602687
        Y | X: 0.02070429
```

c) Variate Adequacies

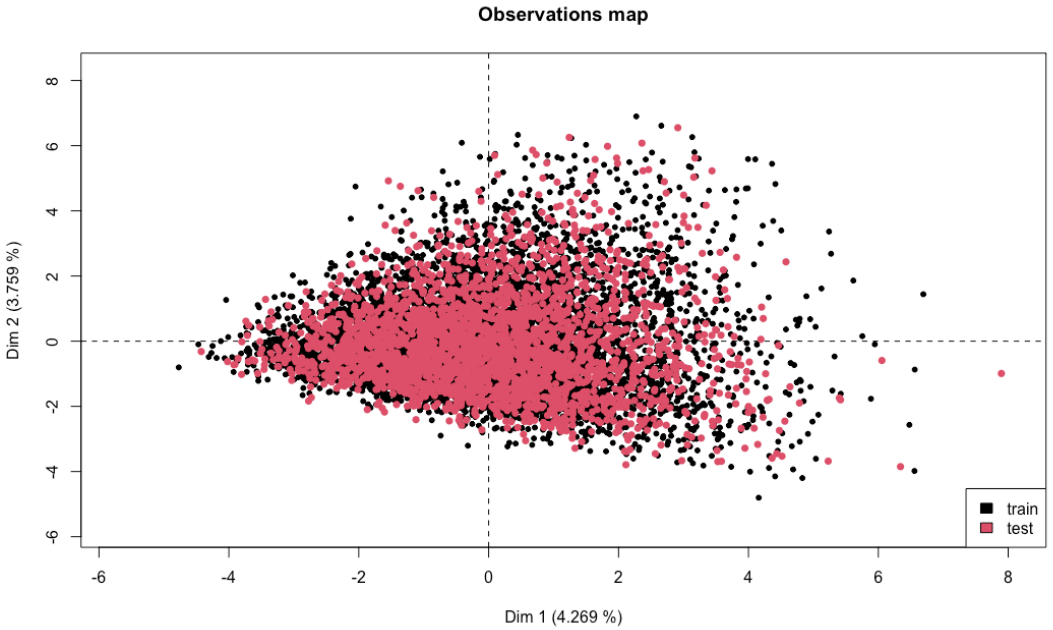

Figure 4: PCA Train/Test Sample Observations

	dim1.rot	dim2.rot	dim3.rot	dim4.rot	dim5.rot
time_walking_biking_perday	0.00	0.00	0.00	0.00	0.21
moderate_activity	0.00	0.00	0.02	0.00	0.25
bp_age	0.04	0.05	0.09	0.20	0.00
height	0.02	0.00	0.50	0.00	0.02
weight	0.00	0.06	0.42	0.01	0.00
time_outdoors_weekdays	0.06	0.01	0.04	0.00	0.32
time_outdoors_weekends	0.00	0.00	0.04	0.00	0.26
vigorous_activity_atwork	0.02	0.02	0.00	0.01	0.20
vigorous_activity	0.00	0.00	0.01	0.00	0.25
age	0.01	0.01	0.00	0.30	0.01
chest_pain_walk_fast	0.23	0.07	0.02	0.02	0.03
hypertension	0.04	0.04	0.03	0.30	0.00
overweight	0.01	0.24	0.02	0.05	0.04
diabetes_relatives	0.01	0.32	0.00	0.00	0.00
diabetes	0.00	0.14	0.01	0.22	0.00
diabetes_risk	0.00	0.41	0.00	0.00	0.01
feel_at_risk_diabetes	0.00	0.37	0.01	0.00	0.01
smoking	0.24	0.01	0.00	0.21	0.00
education_level	0.44	0.04	0.03	0.02	0.08
annual_income	0.51	0.03	0.04	0.04	0.07
bc_checked	0.01	0.00	0.00	0.23	0.02
gender	0.00	0.01	0.25	0.00	0.03
ethnicity	0.18	0.00	0.08	0.03	0.02

Figure 5: Dimensions after Varimax-type Rotation

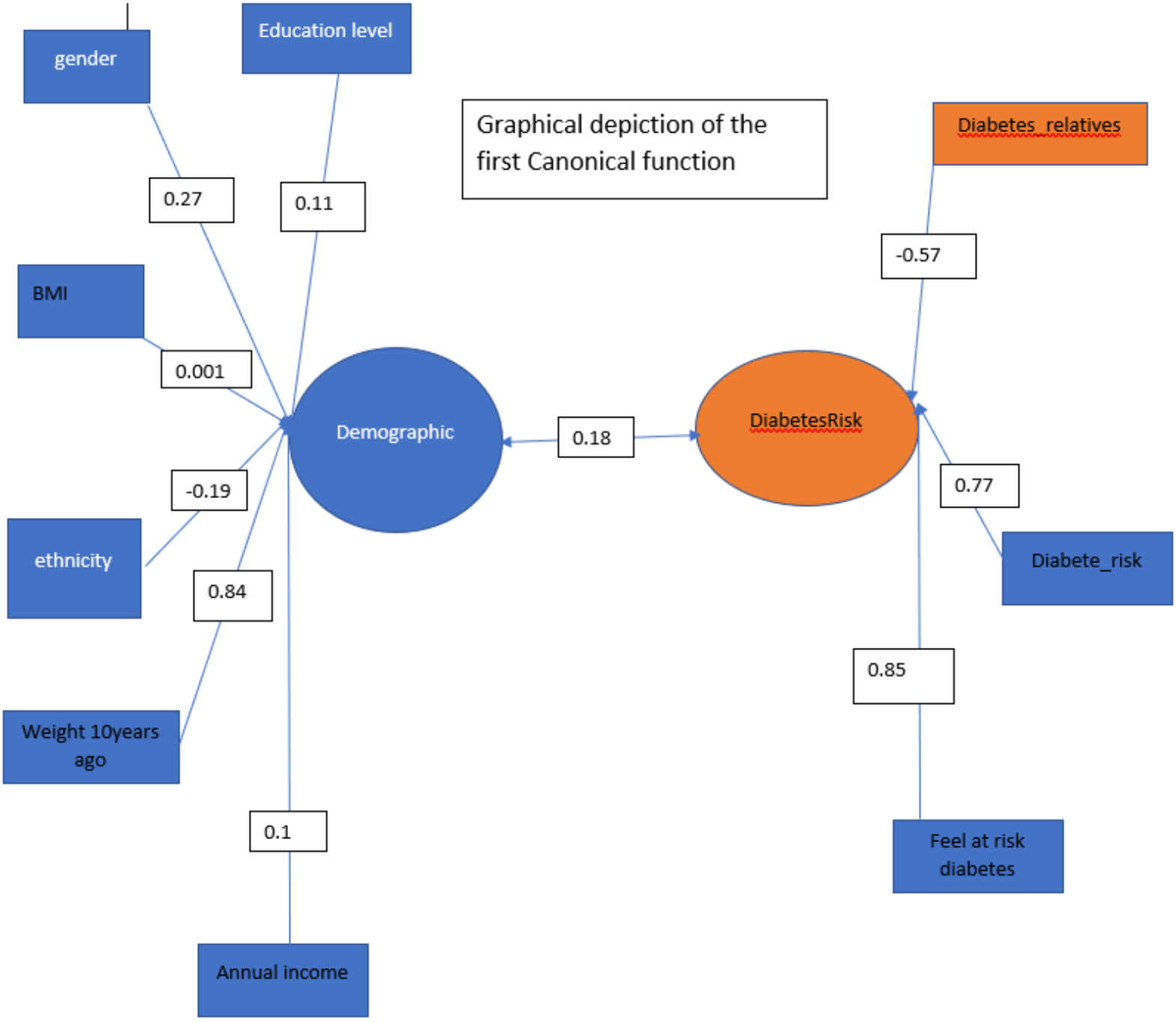

Figure 6: graphical depiction of the first canonical function

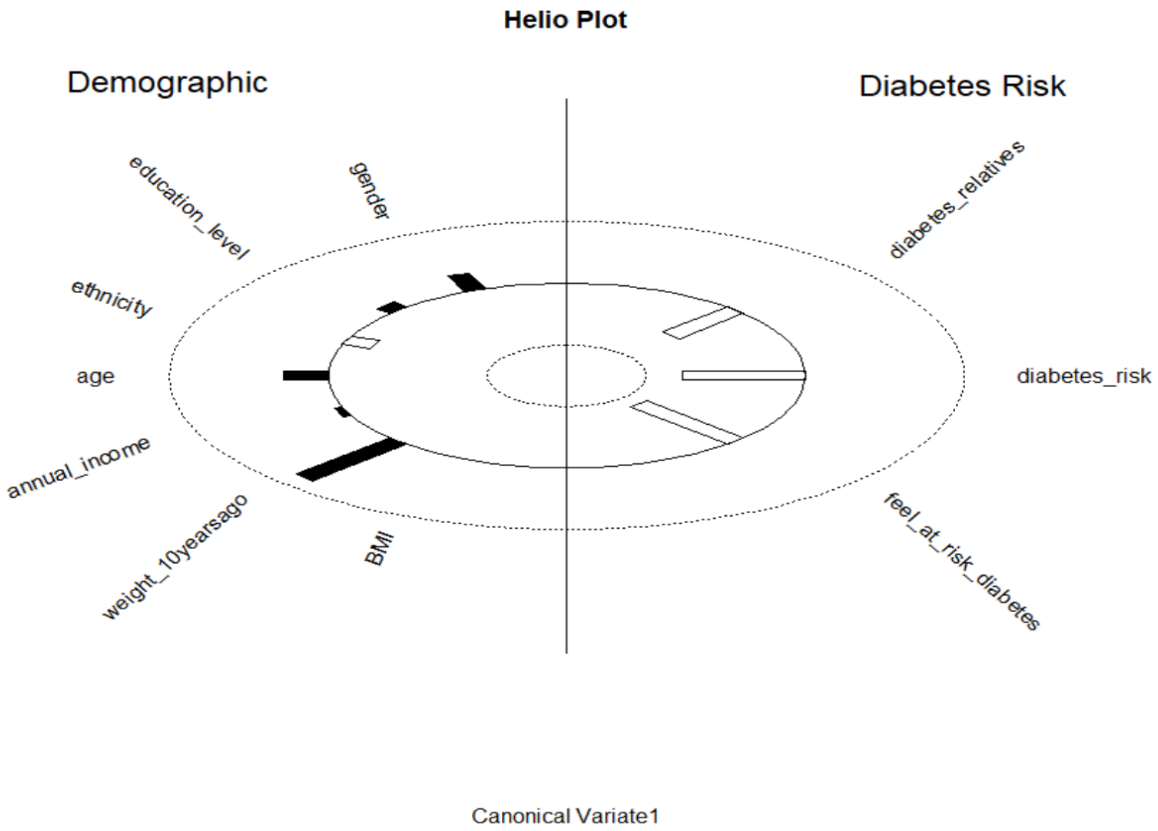

Figure 7: Helios plot

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.45420    0.33732  -7.276 3.45e-13 ***
annual_income55to65      -0.53097    0.14248  -3.727 0.000194 ***
smokingeveryday           0.28820    0.07790   3.699 0.000216 ***
chest_pain_walk_fastNo    0.79264    0.19157   4.138 3.51e-05 ***
chest_pain_normalYes      0.39853    0.08251   4.830 1.36e-06 ***
overweightYes             0.50244    0.09925   5.062 4.14e-07 ***
glutenfreeYes             0.65684    0.24683   2.661 0.007789 **
strokeYes                 0.41392    0.19308   2.144 0.032047 *
diabetesrelativesYes      1.53071    0.07790  19.651  < 2e-16 ***
diabetesNo                0.73627    0.15418   4.775 1.80e-06 ***
feelatriskdiabetesYes     1.49631    0.07489  19.979  < 2e-16 ***
NonHispBlack              0.33324    0.08802   3.786 0.000153 ***
hsgraduateorged           0.25986    0.08915   2.915 0.003560 **
frozen_meals             -0.11456    0.03240  -3.536 0.000407 ***
age                       0.10879    0.03879   2.805 0.005033 **
genderF                   0.42231    0.07240   5.833 5.44e-09 ***
vigorous_activity_atwork  0.08450    0.03647   2.317 0.020509 *
BMI                       0.17344    0.03687   4.705 2.54e-06 ***
bp_age                    0.33387    0.03940   8.474  < 2e-16 ***
last_doctor_visit         0.36067    0.03871   9.317  < 2e-16 ***
---
```

Figure 8: Significant Variables for Logistic Regression