Diabetes Risk - A Multilevel Analysis of Health, Lifestyle, Demographics, and Socioeconomic Factors

Porimol Chandro[*]
p.chandro@student.uw.edu.pl

Thi Hoang Nhu Ho[*]
t.ho2@student.uw.edu.pl

Faculty of Economics Sciences, University of Warsaw

June, 2024

Abstract

This research uses information from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) to examine the risk variables linked to type 2 diabetes mellitus (T2DM). The study finds significant predictors of T2DM, including demographic, lifestyle, clinical, and socioeconomic characteristics, through exploratory data analysis, logistic regression, and probit modeling. The results emphasize how important it is for diabetes risk to be influenced by variables like obesity, high cholesterol, age, gender, food, physical activity, mental health, and socioeconomic position. The necessity of focused public health initiatives to support healthy lifestyles, enhance healthcare service accessibility, tackle socioeconomic inequalities, and customize preventative measures based on demographic traits are among the policy implications. These results advance our knowledge of the risk factors for type 2 diabetes and help guide the creation of successful care and prevention plans.

1 Introduction

Diabetes is a big economic challenge due to its widespread prevalence and substantial financial implications. According to estimates from the Centers for Disease Control and Prevention (CDC), the total costs associated with diagnosed diabetes, un-diagnosed diabetes, and pre-diabetes approach $400 billion annually. This financial burden encompasses direct medical expenses, lost productivity, and reduced quality of life for individuals and their families.

Understanding the risk factors associated with diabetes is essential for several reasons. Firstly, proactive management and early intervention are crucial for mitigating the adverse effects of dia-

[*] Both authors contributed equally to this project work.

betes. By identifying at-risk individuals, healthcare resources can be allocated more efficiently to preventive measures such as lifestyle interventions and targeted screenings.

Secondly, analyzing diabetes risk factors aids in informing public health policies and interventions aimed at reducing its prevalence and impact. By identifying social determinants of health contributing to disparities in diabetes prevalence, policymakers can design targeted programs to address these disparities and promote health equity.

In summary, delving into diabetes risk factors using econometric techniques provides invaluable insights into the economic implications of this chronic disease. It facilitates evidence-based decision-making in healthcare policy, resource allocation, and public health interventions, ultimately aiming to alleviate the economic burden of diabetes on individuals and society.

The rest of this paper is organized as follows: In Section 2, we present a comprehensive literature review, examining previous research relevant to our study. Section 3 describes the data sources and the process of data collection. In Section 4, we detail the methodology employed to analyze the data. Section 5 presents the results of our analysis. Finally, Section 6 discusses the findings, highlighting their implications and suggesting avenues for future research.

Predictive analysis of diabetes risk is an important area of research due to the increasing prevalence of diabetes worldwide. Utilizing binary dependent variable models, such as the logit and probit models, provides robust tools for identifying factors associated with diabetes onset and for predicting individual risk. This literature review explores key studies employing these models, highlights their methodologies, and discusses their contributions to the field.

2 Experimental Studies on Diabetes Risk Factors

Understanding the complex risk factors of type 2 diabetes requires a thorough evaluation of clinical, demographic, genetic, and lifestyle variables. Several experimental investigations have been done to elucidate these connections, using a variety of approaches such as longitudinal cohort studies and randomized controlled trials. These research seek to quantify the impact of many predictors on the development of T2DM, giving significant information for preventive and treatment approaches.

2.1 Clinical Factors

One significant predictor of T2DM is Body Mass Index (BMI). ElSeddawy et al., 2022 conducted a longitudinal cohort study involving 10,000 participants over ten years. Their findings indicated a robust correlation between higher BMI and increased diabetes risk, with individuals exhibiting a BMI greater than 30 being three times more likely to develop T2DM compared to those with a normal BMI. The study emphasized that central obesity, characterized by a higher waist-to-hip ratio, plays a crucial role in insulin resistance, a precursor to T2DM.

Moreover, the risk of diabetes is closely correlated with cholesterol levels. Tsao et al., 2023's research indicates that a higher risk of diabetes is linked to elevated cholesterol levels. Research has indicated that an increased risk of cardiovascular illnesses, including type 2 diabetes, is associated with elevated levels of low-density lipoprotein (LDL) cholesterol.

Diabetes is also more likely to strike people who have had a stroke or coronary heart disease (CHD). According to a Tsao et al., 2023 retrospective cohort research involving 15,000 patients, people with a history of CHD or stroke were 30% more likely to develop type 2 diabetes. The connection between vascular and metabolic health is highlighted by this association, underscoring the necessity of coordinated care approaches.

Formulated Research Hypotheses:

- **H1a:** The risk of type 2 diabetes is greatly increased by *central obesity* and higher *BMI.*
- **H1b:** Type 2 diabetes is associated with a higher risk of high *cholesterol.*
- **H1c:** Having a history of *heart disease* or *stroke* raises the chance of getting type 2 diabetes.

2.2 Demographic Factors

One important demographic factor for T2DM prediction is age. According to , according toVyas et al., 2019, which comprised a cohort of 50,000 people in a range of age groups, showed that the risk of having type 2 diabetes rises with age. Interestingly, although people in the 30-44 age range were at a higher risk than younger adults, people in the 45+ age range had a marked increase in the incidence of T2DM. Age-related physiological alterations like reduced insulin sensitivity and pancreatic beta-cell activity are blamed for this tendency. In order to successfully treat and lower the risk of type 2 diabetes, these findings highlight the significance of age-specific preventative approaches and early interventions across all age groups.

Diabetes risk is also influenced by gender variations. According to ElSeddawy et al., 2022, men are 20% more likely than women to acquire diabetes. A meta-analysis of 30 research with over 100,000 participants revealed that variations in body fat distribution and hormonal factors account for some of this discrepancy.

Based on these findings, several hypotheses can be formulated:

- **H2a:** With age, there is a steady increase in the risk of type 2 diabetes.
- **H2b:** Men have a higher risk of developing type 2 diabetes compared to women due to differences in *body fat distribution* and *hormonal influences.*

2.3 Lifestyle Factors

Diet has a major impact on the risk of diabetes. According to ElSeddawy et al., 2022's dietary intervention trial, people who consumed diets high in refined sugars and saturated fats were 30% more likely to develop diabetes than people who had a diet high in fruits, vegetables, and whole grains. The importance of a healthy diet in preventing type 2 diabetes is highlighted by this study. Daily fruit and vegetable consumption is linked to a decreased risk of diabetes. According to Vyas (2019), people with a 25% lower chance of developing diabetes were those who ate fruits and vegetables at least once a day. These foods' high fiber and nutritional content, which enhance insulin sensitivity and lower inflammation, are responsible for their beneficial effect. Taken together, these results highlight how crucial dietary decisions are in reducing the risk of diabetes.

Exercise is yet another essential lifestyle component. An randomized controlled trial (RCT) conducted by the American Heart Association (2023) with 3,000 participants showed that diabetes incidence was reduced by 40% with regular physical activity, defined as 150 minutes of moderate-intensity exercise per week. On the other hand, a sedentary lifestyle greatly raised the risk of diabetes. The necessity of incorporating physical activity into everyday activities as a preventive intervention against diabetes is highlighted by the advantages of consistent exercise and the risks associated with inactivity.

The relationship between alcohol and tobacco use and the development of diabetes is complicated. Smokers had a 50% increased chance of acquiring diabetes, according to Razavian et al., 2015 analysis of data from a cohort trial with 10,000 participants. There was a complicated link between alcohol intake and the risk of diabetes; moderate alcohol use may provide some protection, but severe drinking elevated the risk by 20%. These results suggest a more balanced strategy is needed to address the impacts of alcohol intake, even while lifestyle changes like quitting smoking can considerably lower the risk of diabetes.

These results allow for the formulation of the following hypotheses:

H3a: A diet rich in *fruits*, and *vegetables* reduces the risk of developing type 2 diabetes.

H3b: Regular *physical activity* reduces the incidence of type 2 diabetes, while *sedentary behavior* increases the risk.

H3c: *Smoking* increases the risk of developing type 2 diabetes, while the relationship between *alcohol consumption* and diabetes risk varies with the level of consumption.

2.4 Socioeconomic Factors

Diabetes risk is greatly influenced by socioeconomic status, which includes variables including money, education, and access to healthcare treatments.

According to research by Razavian et al. (2015), those with lower incomes are 40% more likely to acquire type 2 diabetes mellitus (T2DM). The main cause of this elevated risk is the restricted availability of healthcare services, which postpones the detection and treatment of prediabetes and diabetes. Reduced access to nutrient-dense food and fewer possibilities for physical activity are two major risk factors for type 2 diabetes that are frequently associated with lower income. Financial limitations can also increase stress levels and make it more difficult to afford preventative healthcare, which raises the risk of diabetes.

Higher education levels are typically linked to improved health literacy, which empowers people to make knowledgeable decisions regarding their well-being. As reported by Razavian et al. (2015), those with less education are more likely to develop type 2 diabetes. A portion of this elevated risk can be attributed to ignorance on the significance of preventive healthcare and good lifestyle options. Individuals with lower levels of education could not completely comprehend the consequences of unhealthy eating habits, being sedentary, and the need for routine health screenings, which could delay the identification and treatment of diabetes.

Along with socioeconomic factors, self-reported health and mental health have a major impact on diabetes risk. In a 5,000-person longitudinal research, the Tsao et al., 2023 discovered that high levels of stress and sadness, as well as low levels of self-reported health, were important indicators of diabetes risk. These results emphasize that diabetes prevention initiatives should address both physical and mental health, with mental health assistance being just as important as medical therapies.

Moreover, the significance of health insurance coverage in relation to diabetes prevention and risk cannot be overstated. Based on a cohort study with 10,000 participants, Razavian et al., 2015 also discovered that people without health insurance had a 30% higher risk of acquiring diabetes. This demonstrates the vital role that health insurance plays in guaranteeing that people have access to early treatments and preventive care, as well as frequent screenings and prompt management of pre-diabetic diseases.

These results lend support to a number of theories, including:

H4a: The risk of type 2 diabetes is higher in those with lower *income levels* and in those with lower *educational attainment.*

H4b: *Mental health problems* and low *self-reported health* are associated with a higher risk of type 2 diabetes.

H4c: Lack of *health insurance* increases the risk of developing type 2 diabetes due to reduced access to preventive care and early interventions.

3 Dataset

3.1 Data source

The information used in this study comes from the Behavioral Risk Factor Surveillance System (BRFSS)[*] for the year 2015. BRFSS is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC) since 1984. Its goal is to collect data on health-related risk behaviors, chronic health issues, and the use of preventive services among the American population.

The 2015 BRFSS dataset, provided from Kaggle, includes responses from a large sample of 253,680 people across the United States. This dataset contains 22 variables, including direct questions addressed to participants and derived variables based on individual replies.

3.2 Feature Details

The Diabetes Health Indicators Dataset includes healthcare statistics and lifestyle survey information for individuals, as well as their diagnoses related to diabetes. It features 35 variables that encompass demographics, lab test results, and responses to survey questions for each participant. The target variable for classification indicates whether a participant is diabetic, pre-diabetic, or healthy.

1. **HighBP:** High blood pressure status (0 = no high BP, 1 = high BP)
2. **HighChol:** High cholesterol status (0 = no high cholesterol, 1 = high cholesterol)
3. **CholCheck:** Cholesterol check in past 5 years (0 = no, 1 = yes)
4. **BMI:** Body Mass Index
5. **Smoker:** Have smoked at least 100 cigarettes in entire life (0 = no, 1 = yes)
6. **Stroke:** Ever had a stroke (0 = no, 1 = yes)
7. **HeartDiseaseorAttack:** Coronary heart disease or myocardial infarction (0 = no, 1 = yes)
8. **PhysActivity:** Physical activity in past 30 days (0 = no, 1 = yes)
9. **Fruits:** Consume fruit 1 or more times per day (0 = no, 1 = yes)
10. **Veggies:** Consume vegetables 1 or more times per day (0 = no, 1 = yes)
11. **HvyAlcoholConsump:** Heavy alcohol consumption (0 = no, 1 = yes)
12. **AnyHealthcare:** Have any kind of healthcare coverage (0 = no, 1 = yes)

[*] https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

13. **NoDocbcCost:** Could not see a doctor due to cost in past 12 months (0 = no, 1 = yes)
14. **GenHlth:** General health status (1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor)
15. **MentHlth:** Number of days mental health not good in past 30 days (1-30)
16. **PhysHlth:** Number of days physical health not good in past 30 days (1-30)
17. **DiffWalk:** Difficulty walking or climbing stairs (0 = no, 1 = yes)
18. **Sex:** Sex (0 = female, 1 = male)
19. **Age:** Age category (1-13, where 1 = 18-24 years, ..., 13 = 80 or older)
20. **Education:** Education level (1-6, where 1 = no school/kindergarten, ..., 6 = college graduate)
21. **Income:** Income level (1-8, where 1 = less than $10,000, ..., 8 = $75,000 or more)
22. **Diabetes:** Diabetes status (0 = no, 1 = prediabetes or diabetes)

3.3 Exploratory Data Analysis

3.3.1 Summary of the Target Variable: Diabetes

The binary target variable ”Diabetes” indicates if a person has diabetes (1) or not (0). Out of the overall population, 84.24% of the persons in the dataset are not diabetics, and only 15.76% of the population has the disease.

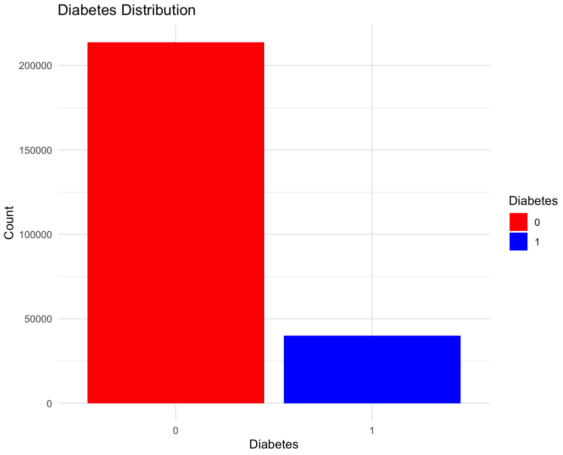

Figure 1: Distribution of the target variable Diabetes

This distribution reveals that, although diabetes is less prevalent in the dataset, it still impacts a notable 15.76% of the population. Understanding the distribution is crucial, as it highlights the importance of identifying factors that contribute to diabetes prevalence and the potential economic implications for healthcare systems.

3.3.2 Variable Selection

In our research, we use the London School of Economics (LSE) method, which is a reliable way to find and eliminate inconsequential variables that is better than the conventional specific-to-general strategy. By preserving the integrity of important predictors, the LSE technique guarantees a thorough and effective reduction of model complexity.

Steps to Identify and Drop Insignificant Variables:

1. **Step 1: Start with a General Model** *(Appendix 7.1)* Begin with a comprehensive model that includes all potential predictors.
2. **Step 2: Test Insignificant Variables Jointly** *(Appendix 7.6)*

   Perform an analysis of variance (ANOVA) to test the joint insignificance of all variables that appear insignificant in the initial model.
3. **Step 3: Apply the Step-by-Step Approach** We must eliminate each unimportant variable one at a time, beginning with the most insignificant one, if they are not jointly insignificant. We re-estimate the model and retest the variables that remain after eliminating the least significant variable from the overall model. Until every variable that remains is significant, we continue.

We can make sure that the variables in our final prediction model only significantly contribute to explaining the variation in the outcome variable by adhering to this systematic method using the linear hypothesis function.

Structure of a Typical Process (If There Were Insignificant Variables): Even though in this instance all variables are statistically significant, here's how you would approach it if there were insignificant variables:

1. **Identify Insignificant Variables:** As a general rule, variables with p-values greater than 0.05 are considered insignificant. These could be dropped.
2. **Remove Insignificant Variables:** The above-mentioned strategy should be used to exclude all inconsequential variables in order to simplify the model and lessen the issue of over-fitting.

4 Methodology

4.1 Modeling Binary Outcomes

The logit model and the probit model, two statistical models frequently used for assessing binary outcomes, are employed in this work. These models calculate the correlation between a group of independent factors X_1 to X_n and the likelihood that a particular event (like the beginning of diabetes) will occur.

4.1.1 Logit Model

The logit model is a type of regression where the dependent variable is binary. It estimates the probability of an event occurring (e.g., diabetes onset) by modeling the log-odds of the event as a linear combination of predictor variables.

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

4.1.2 Probit Model

Similar to the logit model, the probit model also handles binary outcomes. It assumes that the probability of the dependent variable being 1 follows a cumulative normal distribution.

$$\Phi^{-1}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{2}$$

Where Φ is the cumulative distribution function of the standard normal distribution.

The specifics of the data and the overall performance of the model are frequently what determine which of the logit and probit models to use. Though in many cases the results produced by both models are comparable, there may be minor variations based on the underlying distribution of the mistakes.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be utilized in this study to assess model performance and choose the best choice. The AIC takes goodness-of-fit into account and penalizes model complexity. The model with the lowest AIC value shall be deemed more suitable for the given data. BIC similar to AIC but includes a stronger penalty for models with more parameters. Lower BIC values also indicate a better model fit, with a more stringent penalty for complexity.

4.2 Interpretation and Model Selection

Given the provided AIC and BIC *(Appendix 7.5)* values for both the logit and probit models, we can determine which model to select based on these criteria.

The AIC for the probit model (174763.6) is lower than the AIC for the logit model (175220.1). This suggests that, according to AIC, the probit model provides a better fit to the data.

Similarly, the BIC for the probit model (175129.2) is lower than the BIC for the logit model (175585.6). This suggests that, according to BIC, the probit model is preferred.

4.2.1 Final Model Selection

Both AIC and BIC criteria indicate that the probit model is preferred over the logit model because it has lower values for both criteria. Therefore, based on the provided metrics, the probit model is the better choice for our data of this study.

4.2.2 Model Comparison

AIC: The probit model has a lower AIC value, indicating a better fit with a penalty for complexity.
BIC: The probit model also has a lower BIC value, indicating a better fit with a stronger penalty for the number of parameters.

Based on the AIC and BIC comparison, it's suggests that the probit model is the perfect choice for this study. The probit model is selected as the final model. We decided to to use probit as our final selected model because it has lower AIC and BIC values compared to the logit model, indicating it offers a better balance of fit and complexity for our dataset.

5 Results

A wide range of predictors was used to estimate the probit model, including demographic variables like age, sex, and education, lifestyle characteristics like smoking and physical activity, and health indicators like high blood pressure and cholesterol levels. Important insights into the factors influencing diabetes were offered by the complete model.

```
#probit_model
Coefficients:
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept)          -3.7092562  0.0361899 -102.494  < 2e-16 ***
HighBP                0.3835850  0.0075115   51.066  < 2e-16 ***
HighChol              0.3148489  0.0071278   44.172  < 2e-16 ***
BMI                   0.0345192  0.0004984   69.263  < 2e-16 ***
Smoker               -0.0211618  0.0070612   -2.997 0.002727 **
Stroke                0.0863669  0.0144900    5.960 2.52e-09 ***
HeartDiseaseorAttack  0.1472202  0.0102318   14.388  < 2e-16 ***
PhysActivity         -0.0359404  0.0077905   -4.613 3.96e-06 ***
Veggies              -0.0228983  0.0084168   -2.721 0.006518 **
HvyAlcoholConsump    -0.3568356  0.0181708  -19.638  < 2e-16 ***
AnyHealthcare         0.0597438  0.0169592    3.523 0.000427 ***
GenHlth               0.2942179  0.0042187   69.741  < 2e-16 ***
MentHlth             -0.0012837  0.0004664   -2.753 0.005914 **
PhysHlth             -0.0028518  0.0004211   -6.772 1.27e-11 ***
Sex                   0.1262399  0.0070949   17.793  < 2e-16 ***
elementary            0.0740827  0.0266737    2.777 0.005480 **
high_school_graduate -0.0582229  0.0163991   -3.550 0.000385 ***
college              -0.0366516  0.0166557   -2.201 0.027768 *
college_graduate     -0.0896380  0.0170836   -5.247 1.55e-07 ***
A35_39                0.2481620  0.0270482    9.175  < 2e-16 ***
A40_44                0.3370682  0.0250245   13.470  < 2e-16 ***
A45_49                0.4602634  0.0231698   19.865  < 2e-16 ***
A50_54                0.5616472  0.0217836   25.783  < 2e-16 ***
A55_59                0.6094937  0.0212647   28.662  < 2e-16 ***
A60_64                0.7337351  0.0209357   35.047  < 2e-16 ***
A65_69                0.8241847  0.0210004   39.246  < 2e-16 ***
A70_74                0.8669622  0.0216484   40.047  < 2e-16 ***
A75_79                0.8329331  0.0227056   36.684  < 2e-16 ***
A80_older             0.7383859  0.0227873   32.403  < 2e-16 ***
I20K                 -0.0275926  0.0155366   -1.776 0.075736 .
I25K                 -0.0501256  0.0149088   -3.362 0.000773 ***
I35K                 -0.0866153  0.0144187   -6.007 1.89e-09 ***
I50K                 -0.1309914  0.0139152   -9.414  < 2e-16 ***
I75K                 -0.1527396  0.0140529  -10.869  < 2e-16 ***
I75K_more            -0.2392482  0.0137647  -17.381  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 221031  on 253679  degrees of freedom
Residual deviance: 174694  on 253645  degrees of freedom
AIC: 174764
Number of Fisher Scoring iterations: 6
```

The research findings strongly supported the hypothesis regarding health indicators, revealing substantial positive correlations between diabetes risk and factors such as high blood pressure, high cholesterol, heart disease, and stroke. Specifically, high blood pressure (HighBP, β = 0.384) and high cholesterol (HighChol, β = 0.315) were significantly associated with an increased risk of developing diabetes. Additionally, body mass index (BMI, β=0.035) showed a strong positive link, indicating that higher BMI levels considerably increase the risk of diabetes. Moreover, individuals with a history of stroke (= 0.086) or heart disease/attack (= 0.147) were found to be at a higher risk for diabetes, highlighting the importance of cardiovascular disorders in diabetes prevalence. As a result, we **accept H1a, H1b, and H1c,** rejecting the null hypothesis that central obesity, higher BMI, high cholesterol, and a history of cardiovascular problems do not increase diabetes risk.

Contrary to expectations, the findings demonstrated that certain lifestyle factors, such as regular physical activity (PhysActivity = -0.036) and vegetable consumption (Veggies, β = -0.023), were indeed protective against diabetes, **supporting H3a and H3b**. However, the hypothesis regarding smoking (β=-0.021) was not supported, as the observed relationship was negligible. Therefore, **we reject H3c**. Interestingly, high alcohol consumption (HvyAlcoholConsump, β = -0.357) was also significantly negatively linked with diabetes, suggesting complex interactions between alcohol use and other lifestyle variables that require deeper exploration.

General health (GenHlth, β = 0.294) had a high positive correlation with diabetes, emphasizing its significance as a predictor. Both mental health (MentHlth, β = -0.001) and physical health (PhysHlth, β = -0.003) were negatively linked with diabetes, but with lesser effects, support H4b. Furthermore, lack of health insurance was found to significantly increase the risk of developing type 2 diabetes due to reduced access to preventive care and early interventions (AnyHealthcare, β = 0.0597438), we also accept H4c. The data also revealed substantial gender disparities, with sex (β=0.126) having a strong positive correlation with diabetes prevalence. Educational attainment yielded conflicting results: elementary education was positively related to diabetes, whereas higher educational levels (high school, college, and college graduate) were negatively associated, demonstrating a protective impact of higher education.

Age and sex were found to be significant determinants of diabetes risk, **supporting H2a and H2b** in line with the hypothesis. Older age groups (35-39 years to 80+ years) had increasingly higher diabetes rates, indicating greater vulnerability with age. Higher income levels (20K to 75K and above) were associated with lower diabetes risk, underscoring the importance of socioeconomic status in diabetes risk, so we accept H4a.

The likelihood ratio test *(Appendix 7.6)* strongly supported the conclusion that the full model, which includes all predictors, improves model fit more than the null model, which does not include any predictors (χ^2 = 46337, df = 34, $p < 0.001$). This highlights the significance of the variables in the full model. The individual effects of each independent variable on the risk of developing diabetes were further clarified by marginal effects analysis *(Appendix 7.7)*, which highlighted the negative effects of physical inactivity and low vegetable consumption and the positive associations with factors like high blood pressure, elevated cholesterol, higher BMI, incidence of stroke, and heart disease or heart attacks.

Moreover, the linktest *(Appendix 7.9)* that was performed confirmed that the model's link function was adequate, which further supported its dependability ($p < 0.05$) Finally, a holistic view of the model's explanatory power was given by the thorough evaluation of R-squared statistics *(Appendix 7.9)*, which showed that the probit model explains between 20 and 37 percent of the variance in the dependent variable, diabetes. All of these results highlight the importance of the risk variables that have been found as well as the reliability of the used probit model in explaining the intricacies of the development of diabetes.

6 Findings

This study makes important contributions to elucidating the risk factors for type 2 diabetes (T2DM). Based on empirical studies, it is clear that clinical, demographic, lifestyle, and socioeconomic factors all play important roles in diabetes risk.

One of the notable findings is the strong correlation between body mass index (BMI) and the risk of T2DM, confirming that obesity, especially abdominal obesity, is a major risk factor. This is consistent with previous studies, such as Elseddawy et al. (2022), which have shown the importance of weight control in preventing diabetes.

In addition, our study also highlights the role of high cholesterol levels in increasing the risk of T2DM. This is consistent with previous studies on the association between cholesterol and cardiovascular disease, highlighting the importance of maintaining healthy cholesterol levels.

Demographically, age and gender are important factors influencing disease risk. These findings provide a basis for developing prevention strategies based on age and gender, as discussed in Vyas (2019).

Lifestyle also plays an important role, with diet and physical activity clearly influencing disease risk. Our study confirms that a diet rich in fruits and vegetables, along with regular exercise, can reduce the risk of T2DM. This highlights the importance of lifestyle interventions in diabetes prevention.

Finally, socioeconomic factors such as income and education level have a strong influence on disease risk. People with low income and low education levels are at higher risk of T2DM, highlighting the need for public health policies to reduce social inequalities and improve access to health care.

7 Conclusion and Policy Implications

The risk factors for type 2 diabetes are made clearer by our study, and it also serves as a foundation for the creation of sensible regulations and preventative measures. To begin with, public health education initiatives are required to increase knowledge of the significance of upholding a nutritious diet and engaging in regular physical exercise.

Secondly, improving access to healthcare services for low-income and less-educated populations is essential. This includes providing regular health screenings and support for managing weight, cholesterol, and other risk factors.

Third, since stress and sadness can raise the risk of diabetes, health programs should prioritize promoting mental health. This danger can be reduced by incorporating psychological support services into comprehensive healthcare initiatives.

Lastly, our study suggests that in order to improve the efficacy of diabetes prevention and management, specific preventive strategies depending on age and gender are required. Programs for healthcare should be created with various target groups' unique requirements in mind.

References

[1] Ahmed I ElSeddawy et al. "Predictive Analysis of Diabetes-Risk with Class Imbalance". In: *Computational Intelligence and Neuroscience* 2022 (2022).

[2] Narges Razavian et al. "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors". In: *Big Data* 3.4 (2015), pp. 277–287.

[3] Connie W Tsao et al. "Heart disease and stroke statistics—2023 update: a report from the American Heart Association". In: *Circulation* 147.8 (2023), e93–e621.

[4] Sonali Vyas et al. "Review of predictive analysis techniques for analysis diabetes risk". In: *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE. 2019, pp. 626–631.

Appendix

Variables Selection

7.1 General variables

```
#Model with selected general variables
general<- lm(Diabetes ~ ., data = health_data[, !(names(health_data) %in% c
    ("Education","Age", "Income"))])
summary(general)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.00207 -0.20114 -0.08309  0.03545  1.18735 

Coefficients:
                        Estimate  Std. Error t value Pr(>|t|)    
(Intercept)          -0.26648174  0.02607069 -10.222  < 2e-16 ***
HighBP                0.08089472  0.00153398  52.735  < 2e-16 ***
HighChol              0.06262958  0.00144702  43.282  < 2e-16 ***
BMI                   0.00770611  0.00010672  72.212  < 2e-16 ***
Smoker               -0.00747473  0.00139494  -5.358 8.40e-08 ***
Stroke                0.03944223  0.00346542  11.382  < 2e-16 ***
HeartDiseaseorAttack  0.07003532  0.00244340  28.663  < 2e-16 ***
PhysActivity         -0.01127428  0.00164441  -6.856 7.09e-12 ***
Fruits               -0.00015607  0.00144454  -0.108 0.913960    
Veggies              -0.00400703  0.00177617  -2.256 0.024072 *  
HvyAlcoholConsump    -0.05358605  0.00289402 -18.516  < 2e-16 ***
AnyHealthcare         0.01945994  0.00315488   6.168 6.92e-10 ***
GenHlth               0.05408266  0.00081657  66.232  < 2e-16 ***
MentHlth             -0.00036502  0.00009851  -3.706 0.000211 ***
PhysHlth              0.00047573  0.00009259   5.138 2.78e-07 ***
Sex                   0.01456074  0.00137554  10.586  < 2e-16 ***
elementary            0.03170875  0.02568392   1.235 0.216989    
high_school          -0.00046338  0.02538120  -0.018 0.985434    
high_school_graduate -0.02250397  0.02519569  -0.893 0.371768    
college              -0.02006526  0.02519908  -0.796 0.425876    
college_graduate     -0.02418124  0.02520378  -0.959 0.337344    
A25_29               -0.00318065  0.00583366  -0.545 0.585600    
A30_34               -0.00613472  0.00544144  -1.127 0.259571    
A35_39                0.00505676  0.00527450   0.959 0.337701    
```

```
A40_44                0.01151556  0.00517237   2.226 0.025991 *  
A45_49                0.02376040  0.00505429   4.701 2.59e-06 ***
A50_54                0.03687635  0.00492475   7.488 7.02e-14 ***
A55_59                0.04427983  0.00487933   9.075  < 2e-16 ***
A60_64                0.06903793  0.00486791  14.182  < 2e-16 ***
A65_69                0.08885820  0.00490889  18.101  < 2e-16 ***
A70_74                0.09738879  0.00505717  19.258  < 2e-16 ***
A75_79                0.08572096  0.00527978  16.236  < 2e-16 ***
A80_older             0.05432596  0.00524504  10.358  < 2e-16 ***
I15K                  0.00217642  0.00454833   0.479 0.632288    
I20K                 -0.01022365  0.00428538  -2.386 0.017047 *  
I25K                 -0.01854313  0.00414471  -4.474 7.68e-06 ***
I35K                 -0.02918129  0.00403072  -7.240 4.51e-13 ***
I50K                 -0.04038899  0.00391507 -10.316  < 2e-16 ***
I75K                 -0.04426044  0.00389887 -11.352  < 2e-16 ***
I75K_more            -0.05195200  0.00381914 -13.603  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3317 on 253640 degrees of freedom
Multiple R-squared:  0.1715,Adjusted R-squared:  0.1713 
F-statistic:  1346 on 39 and 253640 DF,  p-value: < 2.2e-16
```

7.2 Step1

```
model1 <- lm(Diabetes ~ ., data = health_data[, 1:19])
summary(model1)
```

```

Call:
lm(formula = Diabetes ~ ., data = health_data[, 1:19])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.00207 -0.20114 -0.08309  0.03545  1.18735 

Coefficients:
                                 Estimate  Std. Error t value Pr(>|t|)    
(Intercept)                   -0.26648174  0.02607069 -10.222  < 2e-16 ***
HighBP                         0.08089472  0.00153398  52.735  < 2e-16 ***
HighChol                       0.06262958  0.00144702  43.282  < 2e-16 ***
BMI                            0.00770611  0.00010672  72.212  < 2e-16 ***
Smoker                        -0.00747473  0.00139494  -5.358 8.40e-08 ***
```

```
Stroke                         0.03944223 0.00346542  11.382  < 2e-16 ***
HeartDiseaseorAttack           0.07003532 0.00244340  28.663  < 2e-16 ***
PhysActivity                  -0.01127428 0.00164441  -6.856 7.09e-12 ***
Fruits                        -0.00015607 0.00144454  -0.108 0.913960    
Veggies                       -0.00400703 0.00177617  -2.256 0.024072 *  
HvyAlcoholConsump             -0.05358605 0.00289402 -18.516  < 2e-16 ***
AnyHealthcare                  0.01945994 0.00315488   6.168 6.92e-10 ***
GenHlth                        0.05408266 0.00081657  66.232  < 2e-16 ***
MentHlth                      -0.00036502 0.00009851  -3.706 0.000211 ***
PhysHlth                       0.00047573 0.00009259   5.138 2.78e-07 ***
Sex                            0.01456074 0.00137554  10.586  < 2e-16 ***
Age25_29                      -0.00318065 0.00583366  -0.545 0.585600    
Age30_34                      -0.00613472 0.00544144  -1.127 0.259571    
Age35_39                       0.00505676 0.00527450   0.959 0.337701    
Age40_44                       0.01151556 0.00517237   2.226 0.025991 *  
Age45_49                       0.02376040 0.00505429   4.701 2.59e-06 ***
Age50_54                       0.03687635 0.00492475   7.488 7.02e-14 ***
Age55_59                       0.04427983 0.00487933   9.075  < 2e-16 ***
Age60_64                       0.06903793 0.00486791  14.182  < 2e-16 ***
Age65_69                       0.08885820 0.00490889  18.101  < 2e-16 ***
Age70_74                       0.09738879 0.00505717  19.258  < 2e-16 ***
Age75_79                       0.08572096 0.00527978  16.236  < 2e-16 ***
Age80_older                    0.05432596 0.00524504  10.358  < 2e-16 ***
Educationelementary            0.03170875 0.02568392   1.235 0.216989    
Educationhigh_school          -0.00046338 0.02538120  -0.018 0.985434    
Educationhigh_school_graduate -0.02250397 0.02519569  -0.893 0.371768    
Educationcollege              -0.02006526 0.02519908  -0.796 0.425876    
Educationcollege_graduate     -0.02418124 0.02520378  -0.959 0.337344    
IncomeI15K                     0.00217642 0.00454833   0.479 0.632288    
IncomeI20K                    -0.01022365 0.00428538  -2.386 0.017047 *  
IncomeI25K                    -0.01854313 0.00414471  -4.474 7.68e-06 ***
IncomeI35K                    -0.02918129 0.00403072  -7.240 4.51e-13 ***
IncomeI50K                    -0.04038899 0.00391507 -10.316  < 2e-16 ***
IncomeI75K                    -0.04426044 0.00389887 -11.352  < 2e-16 ***
IncomeI75K_more               -0.05195200 0.00381914 -13.603  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3317 on 253640 degrees of freedom
Multiple R-squared:  0.1715,Adjusted R-squared:  0.1713 
F-statistic:  1346 on 39 and 253640 DF,  p-value: < 2.2e-16
```

```
# Perform ANOVA to test nested models
anova(model1, model1a)
```

```
Analysis of Variance Table

Model 1: Diabetes ~ HighBP + HighChol + BMI + Smoker + Stroke + HeartDiseaseorAttack +
    PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare +
    GenHlth + MentHlth + PhysHlth + Sex + Age + Education + Income
Model 2: Diabetes ~ HighBP + HighChol + BMI + Smoker + Stroke + HeartDiseaseorAttack +
    PhysActivity + Veggies + HvyAlcoholConsump + AnyHealthcare +
    GenHlth + MentHlth + PhysHlth + Sex + A40_44 + A45_49 + A50_54 +
    A55_59 + A65_69 + A70_74 + A75_79 + A80_older + I20K + I25K +
    I35K + I50K + I75K + I75K_more
  Res.Df   RSS  Df Sum of Sq     F    Pr(>F)
1 253640 27903
2 253651 27998 -11   -95.296 78.75 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.3 Step2 - same for next steps

```r
model2 <- lm(Diabetes ~     HighBP + HighChol + BMI + Smoker
              +  Stroke + HeartDiseaseorAttack + PhysActivity + Fruits + Veggies
              +  HvyAlcoholConsump + AnyHealthcare + GenHlth + MentHlth
              +  PhysHlth + Sex + Age +  Income
              +  elementary + high_school_graduate + college + college_graduate
               ,data = health_data)

# # let's drop "the most insignificant" variable from model2 that is Fruits
# # and test joint hypothesis: beta_high_school=beta_Fruits=0

linearHypothesis(general, c("high_school=0", "Fruits=0"))
```

```
Linear hypothesis test

Hypothesis:
high_school = 0
Fruits = 0

Model 1: restricted model
Model 2: Diabetes ~ HighBP + HighChol + BMI + Smoker + Stroke + HeartDiseaseorAttack +
    PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare +
    GenHlth + MentHlth + PhysHlth + Sex + elementary + high_school +
    high_school_graduate + college + college_graduate + A25_29 +
    A30_34 + A35_39 + A40_44 + A45_49 + A50_54 + A55_59 + A60_64 +
    A65_69 + A70_74 + A75_79 + A80_older + I15K + I20K + I25K +
    I35K + I50K + I75K + I75K_more

  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1 253642 27903
2 253640 27903  2 0.0013185 0.006  0.994
```

7.4 Step3 - Final step

```
model6 <- lm(Diabetes ~ HighBP + HighChol + BMI + Smoker
             +  Stroke + HeartDiseaseorAttack + PhysActivity + Veggies
             +  HvyAlcoholConsump + AnyHealthcare + GenHlth + MentHlth
             +  PhysHlth + Sex +  elementary + high_school_graduate
             + college + college_graduate + A35_39 + A40_44 + A45_49
             + A50_54 + A55_59 + A60_64 + A65_69 + A70_74 + A75_79
             +  A80_older + I20K + I25K + I35K + I50K + I75K + I75K_more
             ,data = health_data)


summary(model6)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.00294 -0.20116 -0.08310  0.03533  1.18727

Coefficients:
                        Estimate  Std. Error t value Pr(>|t|)
(Intercept)          -0.26946724  0.00643254 -41.891  < 2e-16 ***
HighBP                0.08087621  0.00153383  52.728  < 2e-16 ***
HighChol              0.06261163  0.00144648  43.286  < 2e-16 ***
BMI                   0.00770132  0.00010654  72.283  < 2e-16 ***
Smoker               -0.00754891  0.00139130  -5.426 5.78e-08 ***
Stroke                0.03943048  0.00346536  11.378  < 2e-16 ***
HeartDiseaseorAttack  0.07004242  0.00244334  28.667  < 2e-16 ***
PhysActivity         -0.01126291  0.00163748  -6.878 6.07e-12 ***
Veggies              -0.00406510  0.00173275  -2.346 0.018975 *
HvyAlcoholConsump    -0.05352780  0.00289172 -18.511  < 2e-16 ***
AnyHealthcare         0.01950658  0.00315390   6.185 6.22e-10 ***
GenHlth               0.05407744  0.00081615  66.259  < 2e-16 ***
MentHlth             -0.00036430  0.00009847  -3.700 0.000216 ***
PhysHlth              0.00047457  0.00009256   5.127 2.95e-07 ***
Sex                   0.01462402  0.00137083  10.668  < 2e-16 ***
```

```
elementary            0.03206309  0.00622655   5.149 2.61e-07 ***
high_school_graduate -0.02195039  0.00367809  -5.968 2.41e-09 ***
college              -0.01949505  0.00370632  -5.260 1.44e-07 ***
college_graduate     -0.02372671  0.00375302  -6.322 2.59e-10 ***
A35_39                0.00889004  0.00354752   2.506 0.012212 *
A40_44                0.01535604  0.00339186   4.527 5.98e-06 ***
A45_49                0.02759540  0.00321078   8.595  < 2e-16 ***
A50_54                0.04070350  0.00300473  13.546  < 2e-16 ***
A55_59                0.04811444  0.00292565  16.446  < 2e-16 ***
A60_64                0.07288190  0.00290161  25.118  < 2e-16 ***
A65_69                0.09271616  0.00296447  31.276  < 2e-16 ***
A70_74                0.10124209  0.00320727  31.566  < 2e-16 ***
A75_79                0.08956401  0.00355650  25.183  < 2e-16 ***
A80_older             0.05815087  0.00351213  16.557  < 2e-16 ***
I20K                 -0.01144922  0.00347814  -3.292 0.000996 ***
I25K                 -0.01977285  0.00329510  -6.001 1.97e-09 ***
I35K                 -0.03044242  0.00314397  -9.683  < 2e-16 ***
I50K                 -0.04168213  0.00299210 -13.931  < 2e-16 ***
I75K                 -0.04558797  0.00297306 -15.334  < 2e-16 ***
I75K_more            -0.05329441  0.00287391 -18.544  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3317 on 253645 degrees of freedom
Multiple R-squared:  0.1715,Adjusted R-squared:  0.1713
F-statistic:  1544 on 34 and 253645 DF,  p-value: < 2.2e-16
```

7.5 Table: Dependent variable

```r
library("stargazer")
stargazer(general, model6, type="text")
```

```
==================================================================================
                                          Dependent variable:
                     -------------------------------------------------------------
                                               Diabetes
                                (1)                            (2)
----------------------------------------------------------------------------------
HighBP                         0.081***                       0.081***
                               (0.002)                        (0.002)

HighChol                       0.063***                       0.063***
```

	(0.001)	(0.001)
BMI	0.008***	0.008***
	(0.0001)	(0.0001)
Smoker	-0.007***	-0.008***
	(0.001)	(0.001)
Stroke	0.039***	0.039***
	(0.003)	(0.003)
HeartDiseaseorAttack	0.070***	0.070***
	(0.002)	(0.002)
PhysActivity	-0.011***	-0.011***
	(0.002)	(0.002)
Fruits	-0.0002	
	(0.001)	
Veggies	-0.004**	-0.004**
	(0.002)	(0.002)
HvyAlcoholConsump	-0.054***	-0.054***
	(0.003)	(0.003)
AnyHealthcare	0.019***	0.020***
	(0.003)	(0.003)
GenHlth	0.054***	0.054***
	(0.001)	(0.001)
MentHlth	-0.0004***	-0.0004***
	(0.0001)	(0.0001)
PhysHlth	0.0005***	0.0005***
	(0.0001)	(0.0001)
Sex	0.015***	0.015***
	(0.001)	(0.001)
elementary	0.032	0.032***
	(0.026)	(0.006)
high_school	-0.0005	
	(0.025)	

high_school_graduate	-0.023	-0.022***
	(0.025)	(0.004)
college	-0.020	-0.019***
	(0.025)	(0.004)
college_graduate	-0.024	-0.024***
	(0.025)	(0.004)
A25_29	-0.003	
	(0.006)	
A30_34	-0.006	
	(0.005)	
A35_39	0.005	0.009**
	(0.005)	(0.004)
A40_44	0.012**	0.015***
	(0.005)	(0.003)
A45_49	0.024***	0.028***
	(0.005)	(0.003)
A50_54	0.037***	0.041***
	(0.005)	(0.003)
A55_59	0.044***	0.048***
	(0.005)	(0.003)
A60_64	0.069***	0.073***
	(0.005)	(0.003)
A65_69	0.089***	0.093***
	(0.005)	(0.003)
A70_74	0.097***	0.101***
	(0.005)	(0.003)
A75_79	0.086***	0.090***
	(0.005)	(0.004)
A80_older	0.054***	0.058***
	(0.005)	(0.004)

```
I15K                             0.002
                                (0.005)

I20K                            -0.010**                      -0.011***
                                (0.004)                        (0.003)

I25K                           -0.019***                      -0.020***
                                (0.004)                        (0.003)

I35K                           -0.029***                      -0.030***
                                (0.004)                        (0.003)

I50K                           -0.040***                      -0.042***
                                (0.004)                        (0.003)

I75K                           -0.044***                      -0.046***
                                (0.004)                        (0.003)

I75K_more                      -0.052***                      -0.053***
                                (0.004)                        (0.003)

Constant                       -0.266***                      -0.269***
                                (0.026)                        (0.006)

----------------------------------------------------------------------------------
Observations                    253,680                        253,680
R2                               0.171                          0.171
Adjusted R2                      0.171                          0.171
Residual Std. Error       0.332 (df = 253640)            0.332 (df = 253645)
F Statistic          1,345.854*** (df = 39; 253640) 1,543.750*** (df = 34; 253645)
==================================================================================
Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Choice between logit and probit based on information criteria

```
# probit model estimation
probit_model <- glm(Diabetes ~    HighBP + HighChol + BMI + Smoker
                + Stroke + HeartDiseaseorAttack + PhysActivity + Veggies
                + HvyAlcoholConsump + AnyHealthcare + GenHlth + MentHlth
                + PhysHlth + Sex
                +  elementary + high_school_graduate + college + college_graduate
                + A35_39 + A40_44 + A45_49 + A50_54
                + A55_59 + A60_64 + A65_69 + A70_74 + A75_79 +  A80_older
                + I20K + I25K + I35K + I50K + I75K + I75K_more
                , data=df,
                family=binomial(link="probit"))
summary(probit_model)
```

```
Coefficients:
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept)          -3.7092562  0.0361899 -102.494  < 2e-16 ***
HighBP                0.3835850  0.0075115   51.066  < 2e-16 ***
HighChol              0.3148489  0.0071278   44.172  < 2e-16 ***
BMI                   0.0345192  0.0004984   69.263  < 2e-16 ***
Smoker               -0.0211618  0.0070612   -2.997 0.002727 **
Stroke                0.0863669  0.0144900    5.960 2.52e-09 ***
HeartDiseaseorAttack  0.1472202  0.0102318   14.388  < 2e-16 ***
PhysActivity         -0.0359404  0.0077905   -4.613 3.96e-06 ***
Veggies              -0.0228983  0.0084168   -2.721 0.006518 **
HvyAlcoholConsump    -0.3568356  0.0181708  -19.638  < 2e-16 ***
AnyHealthcare         0.0597438  0.0169592    3.523 0.000427 ***
GenHlth               0.2942179  0.0042187   69.741  < 2e-16 ***
MentHlth             -0.0012837  0.0004664   -2.753 0.005914 **
PhysHlth             -0.0028518  0.0004211   -6.772 1.27e-11 ***
Sex                   0.1262399  0.0070949   17.793  < 2e-16 ***
elementary            0.0740827  0.0266737    2.777 0.005480 **
high_school_graduate -0.0582229  0.0163991   -3.550 0.000385 ***
college              -0.0366516  0.0166557   -2.201 0.027768 *
college_graduate     -0.0896380  0.0170836   -5.247 1.55e-07 ***
A35_39                0.2481620  0.0270482    9.175  < 2e-16 ***
A40_44                0.3370682  0.0250245   13.470  < 2e-16 ***
A45_49                0.4602634  0.0231698   19.865  < 2e-16 ***
A50_54                0.5616472  0.0217836   25.783  < 2e-16 ***
A55_59                0.6094937  0.0212647   28.662  < 2e-16 ***
A60_64                0.7337351  0.0209357   35.047  < 2e-16 ***
A65_69                0.8241847  0.0210004   39.246  < 2e-16 ***
```

```
A70_74                0.8669622  0.0216484   40.047  < 2e-16 ***
A75_79                0.8329331  0.0227056   36.684  < 2e-16 ***
A80_older             0.7383859  0.0227873   32.403  < 2e-16 ***
I20K                 -0.0275926  0.0155366   -1.776 0.075736 .
I25K                 -0.0501256  0.0149088   -3.362 0.000773 ***
I35K                 -0.0866153  0.0144187   -6.007 1.89e-09 ***
I50K                 -0.1309914  0.0139152   -9.414  < 2e-16 ***
I75K                 -0.1527396  0.0140529  -10.869  < 2e-16 ***
I75K_more            -0.2392482  0.0137647  -17.381  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221031  on 253679  degrees of freedom
Residual deviance: 174694  on 253645  degrees of freedom
AIC: 174764

Number of Fisher Scoring iterations: 6
```

```
# logit model estimation
logit_model <- glm(Diabetes ~      HighBP + HighChol + BMI + Smoker
                +  Stroke + HeartDiseaseorAttack + PhysActivity + Veggies
                +  HvyAlcoholConsump + AnyHealthcare + GenHlth + MentHlth
                +  PhysHlth + Sex
                +  elementary + high_school_graduate + college + college_graduate
                + A35_39 + A40_44 + A45_49 + A50_54
                +  A55_59 + A60_64 + A65_69 + A70_74 + A75_79 +  A80_older
                + I20K + I25K +  I35K + I50K + I75K + I75K_more
                , data=df, family=binomial(link = "logit"))
summary(logit_model)
```

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.7743171  0.0691417 -97.977  < 2e-16 ***
HighBP                0.7106283  0.0138440  51.331  < 2e-16 ***
HighChol              0.5740490  0.0129336  44.384  < 2e-16 ***
BMI                   0.0613014  0.0008788  69.755  < 2e-16 ***
Smoker               -0.0391972  0.0126705  -3.094 0.001978 **
Stroke                0.1343378  0.0245516   5.472 4.46e-08 ***
HeartDiseaseorAttack  0.2288440  0.0173643  13.179  < 2e-16 ***
PhysActivity         -0.0606686  0.0137342  -4.417 9.99e-06 ***
Veggies              -0.0413344  0.0148909  -2.776 0.005506 **
```

```
HvyAlcoholConsump    -0.6842444  0.0350265 -19.535  < 2e-16 ***
AnyHealthcare         0.0979130  0.0308528   3.174 0.001506 **
GenHlth               0.5265409  0.0076452  68.872  < 2e-16 ***
MentHlth             -0.0020678  0.0008164  -2.533 0.011313 *
PhysHlth             -0.0061568  0.0007326  -8.405  < 2e-16 ***
Sex                   0.2310843  0.0127831  18.077  < 2e-16 ***
elementary            0.1203088  0.0455174   2.643 0.008214 **
high_school_graduate -0.0959230  0.0283035  -3.389 0.000701 ***
college              -0.0534617  0.0288236  -1.855 0.063626 .
college_graduate     -0.1521528  0.0297237  -5.119 3.07e-07 ***
A35_39                0.5743204  0.0564514  10.174  < 2e-16 ***
A40_44                0.7713318  0.0518489  14.877  < 2e-16 ***
A45_49                1.0138716  0.0481076  21.075  < 2e-16 ***
A50_54                1.2009446  0.0455940  26.340  < 2e-16 ***
A55_59                1.2935617  0.0446687  28.959  < 2e-16 ***
A60_64                1.5162661  0.0440697  34.406  < 2e-16 ***
A65_69                1.6735475  0.0441198  37.932  < 2e-16 ***
A70_74                1.7427936  0.0450468  38.689  < 2e-16 ***
A75_79                1.6795706  0.0466186  36.028  < 2e-16 ***
A80_older             1.5090315  0.0469038  32.173  < 2e-16 ***
I20K                 -0.0442415  0.0268025  -1.651 0.098811 .
I25K                 -0.0771845  0.0258310  -2.988 0.002808 **
I35K                 -0.1424051  0.0251064  -5.672 1.41e-08 ***
I50K                 -0.2211469  0.0243564  -9.080  < 2e-16 ***
I75K                 -0.2572017  0.0247483 -10.393  < 2e-16 ***
I75K_more            -0.4237710  0.0244110 -17.360  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221031  on 253679  degrees of freedom
Residual deviance: 175150  on 253645  degrees of freedom
AIC: 175220

Number of Fisher Scoring iterations: 6
```

```
cat("AIC␣and␣BIC␣for␣logit␣model:", AIC(logit_model),",", BIC(logit_model),
     "\n")
cat("AIC␣and␣BIC␣for␣probit␣model:", AIC(probit_model),",", BIC(probit_
    model), "\n")
```

```
[1] AIC and BIC for logit model: 175220.1 , 175585.6
[1] AIC and BIC for probit model: 174763.6 , 175129.2
```

Testing Final model

7.6 Likelihood ratio test

```
# Joint insignificance of all variables test
null_probit = glm(Diabetes~1, data=df[,1:19], family=binomial(link="probit"))
lrtest(probit_model, null_probit)
```

```
Likelihood ratio test

Model 1: Diabetes ~ HighBP + HighChol + BMI + Smoker + Stroke + HeartDiseaseorAttack +
    PhysActivity + Veggies + HvyAlcoholConsump + AnyHealthcare +
    GenHlth + MentHlth + PhysHlth + Sex + elementary + high_school_graduate +
    college + college_graduate + A35_39 + A40_44 + A45_49 + A50_54 +
    A55_59 + A60_64 + A65_69 + A70_74 + A75_79 + A80_older +
    I20K + I25K + I35K + I50K + I75K + I75K_more

Model 2: Diabetes ~ 1

  #Df  LogLik  Df Chisq Pr(>Chisq)
1  35  -87347
2   1 -110515 -34 46337  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.7 marginal effects

```
# marginal effects for the average observation
(meff = probitmfx(formula=Diabetes ~ HighBP + HighChol + BMI + Smoker
                 + Stroke + HeartDiseaseorAttack + PhysActivity + Veggies
                 + HvyAlcoholConsump + AnyHealthcare + GenHlth + MentHlth
                 + PhysHlth + Sex +  elementary + high_school_graduate
                 + college + college_graduate + A35_39 + A40_44 + A45_49
                 + A50_54 +  A55_59 + A60_64 + A65_69 + A70_74 + A75_79
                 + A80_older + I20K + I25K + I35K + I50K + I75K + I75K_more
                 , data=df, , atmean=TRUE))
```

```
Marginal Effects:
                         dF/dx   Std. Err.        z     P>|z|    
HighBP               7.0705e-02  1.4343e-03  49.2944 < 2.2e-16 ***
HighChol             5.7734e-02  1.3467e-03  42.8703 < 2.2e-16 ***
BMI                  6.1279e-03  8.8891e-05  68.9375 < 2.2e-16 ***
Smoker              -3.7510e-03  1.2499e-03  -3.0010 0.0026908 ** 
Stroke               1.6116e-02  2.8387e-03   5.6772 1.369e-08 ***
HeartDiseaseorAttack  2.8161e-02  2.1085e-03  13.3560 < 2.2e-16 ***
PhysActivity        -6.4555e-03  1.4160e-03  -4.5588 5.144e-06 ***
Veggies             -4.1020e-03  1.5214e-03  -2.6962 0.0070138 ** 
HvyAlcoholConsump   -5.1421e-02  2.0578e-03 -24.9882 < 2.2e-16 ***
AnyHealthcare        1.0246e-02  2.8078e-03   3.6489 0.0002633 ***
GenHlth              5.2230e-02  7.4106e-04  70.4805 < 2.2e-16 ***
MentHlth            -2.2788e-04  8.2799e-05  -2.7522 0.0059190 ** 
PhysHlth            -5.0626e-04  7.4634e-05  -6.7833 1.175e-11 ***
Sex                  2.2635e-02  1.2816e-03  17.6606 < 2.2e-16 ***
elementary           1.3758e-02  5.1751e-03   2.6585 0.0078483 ** 
high_school_graduate -1.0144e-02  2.8042e-03  -3.6175 0.0002975 ***
college             -6.4387e-03  2.8958e-03  -2.2235 0.0261843 *  
college_graduate    -1.5777e-02  2.9820e-03  -5.2907 1.218e-07 ***
A35_39               5.0446e-02  6.1895e-03   8.1502 3.633e-16 ***
A40_44               7.1465e-02  6.1500e-03  11.6204 < 2.2e-16 ***
A45_49               1.0280e-01  6.1856e-03  16.6189 < 2.2e-16 ***
A50_54               1.2927e-01  6.0711e-03  21.2925 < 2.2e-16 ***
A55_59               1.4147e-01  5.9873e-03  23.6280 < 2.2e-16 ***
A60_64               1.7717e-01  6.2235e-03  28.4689 < 2.2e-16 ***
A65_69               2.0591e-01  6.5319e-03  31.5233 < 2.2e-16 ***
A70_74               2.2544e-01  7.1199e-03  31.6636 < 2.2e-16 ***
A75_79               2.1874e-01  7.6058e-03  28.7600 < 2.2e-16 ***
A80_older            1.8647e-01  7.2610e-03  25.6806 < 2.2e-16 ***
I20K                -4.8235e-03  2.6742e-03  -1.8037 0.0712724 .  
I25K                -8.6615e-03  2.5068e-03  -3.4552 0.0005499 ***
I35K                -1.4711e-02  2.3410e-03  -6.2841 3.299e-10 ***
I50K                -2.1901e-02  2.1885e-03 -10.0075 < 2.2e-16 ***
I75K                -2.5418e-02  2.1895e-03 -11.6090 < 2.2e-16 ***
I75K_more           -4.0698e-02  2.2457e-03 -18.1230 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

dF/dx is for discrete change for the following variables:
[1] "HighBP"         "HighChol"       "Smoker"      "Stroke"         "HeartDiseaseorAttack"
[6] "PhysActivity"   "Veggies"        "HvyAlcoholConsump"     "AnyHealthcare"     "Sex"
[11] "elementary"    "high_school_graduate"  "college"   "college_graduate"     "A35_39"
[16] "A40_44"        "A45_49"    "A50_54"    "A55_59"    "A60_64"
[21] "A65_69"        "A70_74"    "A75_79"    "A80_older" "I20K"
[26] "I25K"          "I35K"      "I50K"      "I75K"      "I75K_more"
```

7.8 Linktest

```
source("functions/linktest.R")
linktest_result <- linktest(probit_model)
summary(linktest_result)
```

```
Call:
glm(formula = y ~ yhat + yhat2, family = binomial(link = model$family$link))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.044619   0.006134  -7.274 3.49e-13 ***
yhat         0.555525   0.006894  80.583  < 2e-16 ***
yhat2        0.003721   0.001877   1.982   0.0475 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221031  on 253679  degrees of freedom
Residual deviance: 174672  on 253677  degrees of freedom
AIC: 174678

Number of Fisher Scoring iterations: 7
```

7.9 R-squared statistics

```
# R-squared statistics
PseudoR2(probit_model,c("McFadden","Tjur","McKelveyZavoina","
    VeallZimmermann","Nagelkerke"))
```

```
 McFadden            Tjur McKelveyZavoina VeallZimmermann      Nagelkerke
      0.2096415       0.1900308       0.3666762       0.3317112       0.2870549
```