**What factors influence an individual's risk of developing diabetes?**

Venkata siva ram pande Kalivela

25 November 2023

Applied Econometrics, ECON 5670.

1. RESEARCH QUESTION:  
    Every year, millions of people worldwide are diagnosed with diabetes, a health condition where the body’s ability to process sugar is compromised. This not only affects the person’s health, but it also places a burden on healthcare systems and economies. Diabetes is a costly disease to manage, and it can cause several other health problems that can also be expensive to treat. According to the CDC, the cost of diabetes in the United States is estimated to be over $327 billion per year. This includes the cost of medical care, medications, and lost productivity. Diabetes is also a leading cause of death in the United States. In 2017, an estimated 83,564 people died from diabetes-related causes. If we can identify the key factors that make someone more likely to develop diabetes, we could help people understand their risks and potentially make changes in their lives to prevent it. This is particularly important when considering lifestyle habits, like what we eat, how much we exercise, and even how much sleep we get. The purpose of this research is to explore these factors in-depth, to provide clear information to the public about what might increase their risk. The benefits of understanding these factors are vast: people could lead healthier lives, healthcare costs could be reduced, and we could see a decline in new diabetes cases. By examining these elements, we aim to provide knowledge that will be valuable to individuals, healthcare providers, and policymakers alike.
2. NON-ECONOMETRIC BACKGROUND:

Alcohol intake is commonly practiced in various societies, but its effects on health remain a topic of active research and discussion. As a chronic condition marked by high blood sugar levels, diabetes presents serious health challenges and places a substantial strain on healthcare systems around the world. Grasping how alcohol might impact the risk of diabetes is essential for formulating effective prevention methods and health policies.

The study by Baliunas et al. (2009) in the journal "Diabetes Care" provides a comprehensive analysis of the dose-response relationship between alcohol consumption and the risk of incident diabetes. The authors conducted a systematic review and meta-analysis of 20 cohort studies, offering insights into how different levels of alcohol consumption may differentially impact diabetes risk. Their findings underscore the importance of understanding the nuanced effects of alcohol consumption patterns on diabetes risk. The link between junk food consumption and diabetes is often attributed to the rapid increase in obesity rates worldwide, as obesity is a significant risk factor for type 2 diabetes. The study by Malik et al. (2010) in "Diabetes Care" explores this relationship, emphasizing how diets rich in high-glycemic index foods and saturated fats can lead to insulin resistance, a precursor to diabetes. This research aims to delve deeper into the direct and indirect impacts of junk food on diabetes risk, examining the mechanisms through which diet influences glucose metabolism and insulin sensitivity.

The growing recognition of stress, in both its psychological and physiological forms, as a potential contributing factor in the onset of type 2 diabetes is noteworthy. The influence of stress on factors like blood sugar levels and insulin resistance can potentially initiate the development of diabetes in those predisposed to the condition. The intricate link between stress and diabetes involves various biological processes, including the activation of the hypothalamic-pituitary-adrenal axis due to stress, as well as behavioral aspects such as overeating when stressed. This study is set to explore the different facets of stress, encompassing chronic, work-related, and emotional stress, to determine their individual impacts on the likelihood of developing diabetes. The overarching aim is to thoroughly understand how effective management of stress a key component in strategies can be aimed at preventing diabetes.

In addition to this, regular physical activity is commonly acknowledged as a crucial element in both the prevention and management of type 2 diabetes. Consistent exercise contributes to maintaining a healthy weight, enhancing insulin sensitivity, and reducing blood sugar levels. The study conducted by Knowler and colleagues in 2002, published in the "New England Journal of Medicine," highlights the effectiveness of lifestyle changes, particularly in terms of increased physical activity, in markedly lowering the risk of diabetes among individuals with high susceptibility. This research intends to delve deeper into how different types and levels of physical activity influence diabetes prevention, with the goal of developing specific recommendations for various groups prone to diabetes.

1. DEPENDENT VARIABLE:  
    Diabetes is a health condition where the body cannot handle sugar properly. We want to know what things in a person's life might make them more likely to have this condition. To figure this out, we are going to look at whether people have diabetes or not. This will be measured using the dependent variable, Yi, which will take on a value of 1, if person is clinically diagnosed with diabetes. Conversely, Yi will take on a value of 0 if individual "i" is not diagnosed with diabetes. Data on the dependent variable was collected from a survey collected as a part of research paper (Diabetes Dataset 2019).   
   <https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019>.

3. POTENTIAL INDEPENDENT VARIABLES:

There are various independent variables that will be considered in this analysis. As of now, no regressions have been run, so it is uncertain which of these variables will make it to the final model. Decisions regarding variable inclusion will be made later. For now, all potential independent variables will be defined, and a prediction on their expected sign will be provided.

Data on these variables was sourced from: https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019.

The first independent variable is **Age\_LT40i** variable takes a value of 1 if individual “i” is under 40 years old, and 0 otherwise. Younger individuals might not have encountered many of the lifestyle or biological factors that increase diabetes risk over time. Given this, a negative sign is anticipated for this variable, suggesting a lower risk in this age group.

Moving forward in age, the variable **Age\_40T49i** will assume a value of 1 if individual “i” falls within the 40 to 49 age bracket, and 0 otherwise. As people age, they may be more exposed to factors that increase diabetes risk, such as weight gain or sedentary lifestyles. Thus, this age group might present a higher risk compared to the under-40 bracket, leading us to predict a positive sign.

For the variable **Age\_50T59i**, which indicates individuals aged between 50 and 59, a value of 1 is assigned to those within this age range, and 0 otherwise. With increasing age, the cumulative effect of potential risk factors might be more pronounced. Consequently, we would anticipate a positive sign for this variable, possibly even stronger than the 40-49 age bracket.

The oldest age group in our dataset is represented by **Age\_MT60i**. If individual “i” is 60 years or older, this variable will be 1, and 0 otherwise. Given the long-term exposure to potential diabetes risk factors in this age group, a positive sign is expected, indicating a higher likelihood of diabetes occurrence.

The next potential regressor is "**Gender\_dumi**." This is a dummy variable representing the gender of individual “i”. Specifically, "**Gender\_dumi**" will take a value of 1 if the individual is male and 0 if not male. Research has shown that gender can play a role in the risk of developing certain diseases, including diabetes. For instance, in some studies, women have shown a slightly higher susceptibility to type 2 diabetes compared to men, possibly due to hormonal differences or other biological factors. As such, if being female is associated with a higher likelihood of having diabetes, we would expect a positive sign on this variable. Conversely, if being male is associated with a higher risk, the coefficient would be negative. However, this relationship is not straightforward and will be investigated in the research.

High blood pressure or hypertension has been linked to diabetes. The variable **highBP\_dumi** will hold a value of 1 if individual “i” has been diagnosed with high blood pressure, and 0 if not. Given this established connection, we expect a positive coefficient for this variable.

Physical activity plays a vital role in maintaining health and preventing diseases. Therefore, the **physicallyActive\_dumi** variable, which will be 1 if individual “i” engages in regular physical activity, and 0 if not. Regular exercise can aid in blood sugar regulation and weight management, both crucial for diabetes prevention. Thus, we anticipate a negative sign for this variable.

The next independent variable to discuss is **Smoking\_dumi**. It assumes a value of 1 if individual “i” smokes and 0 if not. Given the known health risks associated with smoking, including its potential link to diabetes, individuals who smoke might be at a higher risk of developing diabetes. Therefore, we anticipate a positive sign for this variable.

Our next variable, **Alcohol\_dumi**, is designated as 1 if individual “i” regularly consumes alcohol and 0 if they do not. While occasional alcohol consumption might not pose significant risks, excessive or regular consumption can influence diabetes onset. Consequently, we could predict a positive coefficient for this variable.

The next variable **Regularmedicine\_dumi** will take a value of 1 if individual “i” is on any regular medication and 0 otherwise. Regular intake of certain medications could hint at underlying health conditions which might correlate with diabetes. Hence, this variable might exhibit a positive relationship with our dependent variable.

Diet plays a crucial role in health outcomes. Therefore, **junkfood\_dumi**is another important variable, assuming a value of 1 if individual “i” frequently consumes junk or processed foods and 0 otherwise. Given that such foods can contribute to obesity and blood sugar spikes, a positive sign is expected for this variable.

Another potential contributor is **stress\_dumi**, which will be 1 if individual “i” often faces high stress levels and 0 otherwise. Persistent stress can interfere with glucose metabolism, and thus, we anticipate a positive coefficient for this variable.

Pregnancy can affect glucose levels, making **Preg\_dumi** significant. This variable takes a value of 1 for individuals who have ever been pregnant, and 0 for those who have never experienced pregnancy. Given the possibility of gestational diabetes during pregnancy, we would predict a positive sign.

Lastly, frequent urination is a common symptom of elevated blood sugar. Thus, the variable **Urinationfreq\_dumi** is defined as 1 if individual “i” reports frequent urination and 0 otherwise. A positive coefficient is expected, aligning with the symptomatology of diabetes.

The next variable, **BMIi** represents the Body Mass Index of individual “i”, a continuous measure of body weight adjusted for height. Higher BMI values can indicate overweight or obesity, both major risk factors for diabetes. Given the direct relationship between obesity and diabetes, a positive coefficient is expected for this variable.

**Sleep\_hri**represents the average amount of sleep individual "i" gets each night. Adequate sleep is important for good health, particularly in preventing diabetes, as it helps maintain the body's ability to regulate blood sugar. We expect a negative sign for **sleep\_hri**, indicating that fewer hours of sleep may be associated with a higher risk of developing diabetes. More sleep, on the other hand, should be linked to a lower risk.

Building on this, **Pdiabetics\_dumi**, which signifies gestational diabetes, holds a value of 1 if individual “i” was diagnosed with gestational diabetes and 0 if not. Clearly representing a form of diabetes, it would have a straightforward positive relationship with our outcome variable.

|  |  |  |
| --- | --- | --- |
| **Table 1: Variable Definitions and Expected Signs** | | |
| **Variable** | **Definition** | **Expected sign** |
| Yi | 1, if person is clinically diagnosed with diabetes, and 0 otherwise | Not applicable |
| Age\_LT40i | 1 if the individual 'i''s age is less than 40, and 0 otherwise. | - |
| Age\_40T49i | 1 if the individual 'i''s age is between 40 and 49, inclusive, and 0 otherwise. | + |
| Age\_50T59i | 1 if the individual 'i''s age is between 50 and 59, inclusive, and 0 otherwise. | + |
| Age\_MT60i | 1 if the individual 'i''s age is 60 or above, and 0 otherwise. | + |
| Gender\_dumi | 1 if the individual 'i' is male, and 0 otherwise. | ? |
| highBP\_dumi | 1 if the individual 'i' has high blood pressure, and 0 otherwise. | + |
| PhysicallyActive\_dumi | 1 if the individual 'i' is physically active, and 0 otherwise. | - |
| smoking\_dumi | 1 if the individual 'i' is a smoker, and 0 otherwise. | + |
| Alcohol\_dumi | 1 if the individual 'i' consumes alcohol regularly, and 0 otherwise. | + |
| RegularMedicine\_dumi | 1 if the individual 'i' takes medicine regularly, and 0 otherwise. | + |
| junkfood\_dumi | 1 if the individual 'i' regularly consumes junk food, and 0 otherwise. | + |
| stress\_dumi | 1 if the individual 'i' frequently experiences stress, and 0 otherwise. | + |
| Preg\_dumi | 1 if the individual 'i' is ever pregnant, and 0 otherwise. | + |
| Urinationfreq\_dumi | 1 if the individual 'i' frequently urinates, and 0 otherwise. | + |
| BMIi | Represents the individual 'i''s Body Mass Index, a measure of body fat based on height and weight. | + |
| Sleep\_hri | Indicates the average number of hours the individual 'i' sleeps per night. | - |
| Pdiabetics\_dumi | 1, if individual 'i' is diagnosed with gestational diabetics, and 0 therwise. | + |

4. SUMMARY STATISTICS AND DATA ISSUES:

Summary Statistics for all observations can be found in table 2. It is essential to review these statistics to ensure that there are no discrepancies in the data that could compromise the results of estimation. As shown in the table, there are no missing observations for any variables. All means appear to be of a reasonable value. Specifically, for the dependent variable to have sufficient weight in the regression model, the mean of Y must lie within the range (0.20, 0.80), and our data confirms that it does, which is 0.28. Similarly, for the dummy variables to be impactful, their means should lie within the range (0.10, 0.90). All dummy variables in our dataset satisfy this criterion, except for `pdiabetics\_dum`, which has a very low mean (0.01) and will be excluded from further analysis. Additionally, the difference between the mean of Y and the means of all regressors is within acceptable limits, with none exceeding 90 units. There are no zero standard deviations, and the maximum and minimum values for all variables are within expected bounds. Given these observations, we can confidently state that there are no apparent issues with the data.

One critical data consideration for a model where the observable dependent variable only assumes values of 0 and 1 is the "largest-smallest" rule. This rule has two essential parts. The first, known as the "largest-smallest rule inside," dictates that for every continuous regressor Xji​, the maximum value of Xji​ when Yi​=0 must exceed the smallest value of Xji​ when Yi​=1. The rule's second part, termed the "largest-smallest rule outside," prescribes that the largest value of Xji​ when Yi​=1 must be greater than the smallest value of Xji​ when Yi​=0. Tables 3 and 4 showcase summary statistics for all variables, segmented by the category of Y. Specifically, Table 3 offers statistics when Yi​=0, and Table 4 provides insights for the case when Yi​=1.

A review of tables 3 and 4 confirms that the "largest-smallest" rules hold true for all continuous regressors, which are BMI, and Sleep\_hr in this dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TABLE 2. SUMMARY STATISTICS FOR ALL OBSERVATIONS IN THE SAMPLE | | | | | |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| **Y** | 948 | 0.28 | 0.45 | 0 | 1 |
| **Age\_LT40** | 948 | 0.51 | 0.5 | 0 | 1 |
| **Age\_40T49** | 948 | 0.17 | 0.38 | 0 | 1 |
| **Age\_50T59** | 948 | 0.16 | 0.37 | 0 | 1 |
| **Age\_MT60** | 948 | 0.15 | 0.36 | 0 | 1 |
| **Gender\_dum** | 948 | 0.61 | 0.49 | 0 | 1 |
| **highBP\_dum** | 948 | 0.24 | 0.43 | 0 | 1 |
| **PhysicallyActive\_dum** | 948 | 0.51 | 0.5 | 0 | 1 |
| **BMI** | 948 | 25.76 | 5.4 | 15 | 45 |
| **smoking\_dum** | 948 | 0.11 | 0.32 | 0 | 1 |
| **Alcohol\_dum** | 948 | 0.2 | 0.4 | 0 | 1 |
| **Sleep\_hr** | 948 | 6.95 | 1.27 | 4 | 11 |
| **RegularMedicine\_dum** | 948 | 0.35 | 0.48 | 0 | 1 |
| **junkfood\_dum** | 948 | 0.1 | 0.3 | 0 | 1 |
| **stress\_dum** | 948 | 0.26 | 0.44 | 0 | 1 |
| **Preg\_dum** | 948 | 0.16 | 0.37 | 0 | 1 |
| **pdiabetics\_dum** | 948 | 0.01 | 0.12 | 0 | 1 |
| **Urinationfreq\_dum** | 948 | 0.3 | 0.46 | 0 | 1 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE 3. SUMMARY STATISTICS FOR OBSERVATIONS FOR WHICH Yi=0 | | | | | | | |
| **Y** | **N Obs** | **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| **0** | **682** | **Age\_LT40** | 682 | 0.66 | 0.47 | 0 | 1 |
| **Age\_40T49** | 682 | 0.17 | 0.37 | 0 | 1 |
| **Age\_50T59** | 682 | 0.12 | 0.33 | 0 | 1 |
| **Age\_MT60** | 682 | 0.05 | 0.21 | 0 | 1 |
| **Gender\_dum** | 682 | 0.62 | 0.48 | 0 | 1 |
| **highBP\_dum** | 682 | 0.14 | 0.35 | 0 | 1 |
| **PhysicallyActive\_dum** | 682 | 0.52 | 0.5 | 0 | 1 |
| **BMI** | 682 | 25.35 | 5.37 | 15 | 42 |
| **smoking\_dum** | 682 | 0.12 | 0.32 | 0 | 1 |
| **Alcohol\_dum** | 682 | 0.19 | 0.39 | 0 | 1 |
| **Sleep\_hr** | 682 | 7.01 | 1.21 | 4 | 11 |
| **RegularMedicine\_dum** | 682 | 0.18 | 0.38 | 0 | 1 |
| **junkfood\_dum** | 682 | 0.11 | 0.31 | 0 | 1 |
| **stress\_dum** | 682 | 0.2 | 0.4 | 0 | 1 |
| **Preg\_dum** | 682 | 0.12 | 0.32 | 0 | 1 |
| **Urinationfreq\_dum** | 682 | 0.27 | 0.45 | 0 | 1 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE 4. SUMMARY STATISTICS FOR OBSERVATIONS FOR WHICH Yi=1 | | | | | | | |
| **Y** | **N Obs** | **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| **1** | **266** | **Age\_LT40** | 266 | 0.12 | 0.33 | 0 | 1 |
| **Age\_40T49** | 266 | 0.19 | 0.39 | 0 | 1 |
| **Age\_50T59** | 266 | 0.27 | 0.44 | 0 | 1 |
| **Age\_MT60** | 266 | 0.42 | 0.49 | 0 | 1 |
| **Gender\_dum** | 266 | 0.58 | 0.49 | 0 | 1 |
| **highBP\_dum** | 266 | 0.49 | 0.5 | 0 | 1 |
| **PhysicallyActive\_dum** | 266 | 0.48 | 0.5 | 0 | 1 |
| **BMI** | 266 | 26.83 | 5.36 | 18 | 45 |
| **smoking\_dum** | 266 | 0.11 | 0.31 | 0 | 1 |
| **Alcohol\_dum** | 266 | 0.24 | 0.43 | 0 | 1 |
| **Sleep\_hr** | 266 | 6.82 | 1.41 | 4 | 10 |
| **RegularMedicine\_dum** | 266 | 0.81 | 0.39 | 0 | 1 |
| **junkfood\_dum** | 266 | 0.09 | 0.28 | 0 | 1 |
| **stress\_dum** | 266 | 0.42 | 0.5 | 0 | 1 |
| **Preg\_dum** | 266 | 0.28 | 0.45 | 0 | 1 |
| **Urinationfreq\_dum** | 266 | 0.38 | 0.49 | 0 | 1 |

Tables 3 and 4 show some interesting details from the data. Most of the people under 40 years old in our study are active, which might help them have a healthier weight. But, even though many people in our study are men, more of them say they have family members with diabetes or high blood pressure. This could mean these health issues run in their families. When we look at sleep, most people sleep about 7 hours, there is possibility for diabetics. Also, people who say they urinate a lot also show signs of diabetes. This makes sense because urination a lot can be a sign of this health issue. These details give us a clearer picture of the health and habits of the people in our study.

The average Body Mass Index (BMI) increases from 25.35 in individuals without diabetes to 26.83 in those with the condition, suggesting that a higher BMI may be linked to a greater likelihood of developing diabetes. Physical activity levels, as indicated by the 'PhysicallyActive\_dum' variable, show a minor decrease from 0.52 to 0.48, which suggests that the amount of physical activity is relatively similar between the two groups in this dataset. There is, however, a significant difference in the use of regular medication; the average for 'RegularMedicine\_dum' jumps from 0.18 in those without diabetes to 0.81 in those with it, which could reflect the treatment of diabetes or other related medical conditions.

A final crucial data aspect to address is the dummy rule, pertinent to all models where the observable dependent variable only assumes values of either zero or one, as seen in probit model. The dummy rule is bifurcated into two key principles. The initial rule asserts that, for every dummy variable in the analysis, when the dummy is marked as 1, the observable dependent variable must exhibit both 0 and 1 values. The second rule states that when the dummy is marked as 0, the observable dependent variable should still represent both 0 and 1 values. Table 5 catalogs the range (minimum and maximum) of 'Y' for each dummy variable in our analysis, encompassing 'Age\_LT40', 'Age\_40T49', 'Age\_50T59', 'Age\_MT60', 'Gender\_dum', 'Family\_Diabetics\_dum', 'highBP\_dum', 'PhysicallyActive\_dum', 'smoking\_dum', 'Alcohol\_dum', 'RegularMedicine\_dum', 'junkfood\_dum', 'stress\_dum', 'Preg\_dum', and 'Urinationfreq\_dum'. A close inspection of Table 5 (contains panels) confirms that each dummy variable adheres to the dummy rule, as 'Y' consistently assumes both 0 and 1 values across the board.

Table 5. Information for checking the Dummy Rule

Analysis variable: Y

|  |  |  |  |
| --- | --- | --- | --- |
| **variable** |  | **Minimum** | **Maximum** |
| **Age\_LT40** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Age\_40T49** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Age\_50T59** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Age\_MT60** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Gender\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Family\_Diabetics\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **PhysicallyActive\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **smoking\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **highBP\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Alcohol\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **RegularMedicine\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **junkfood\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **stress\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Preg\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |
| **Urinationfreq\_dum** | **0** | 0 | 1 |
| **1** | 0 | 1 |

THEORETICAL MODEL AND ESTIAMTION PROCEDURE:

In this study, the observable dependent variable is a binary variable that takes on a value of 1 if a person is diagnosed with diabetics, and 0 if not. Using Ordinary Least Squares (OLS) regression for binary dependent variables is fundamentally flawed due to several violations of the OLS model's core assumptions. Firstly, OLS presupposes a linear relationship between regressors and the dependent variable, which does not hold true for binary dependent variables that operate within a non-linear. Secondly, the assumption of homoscedasticity is violated since the variance of errors in a binary setup is intrinsically linked to the probability of the outcome, leading to heteroscedasticity. Lastly, OLS can yield predictions beyond the binary range of 0 and 1, which is nonsensical for probabilities. These issues necessitate the use of alternative models like probit or logit that correctly accommodate the binary nature of the dependent variable and produce bounded, probabilistic predictions. Therefore, the appropriate model to use is the probit model, and the appropriate estimation procedure for the probit model is maximum likelihood estimation.  
In the probit model, we use Y\* instead of Yi in a probit model because Y\* is a continuous variable that allows us to use a normal distribution to model the error term and the CDF to model the probability of Y being equal to 1. This makes it easier to work with the probit model and to estimate its parameters.   
The latent variable Y\* and the binary dependent variable Yi is given by:  
Y=0, if Y\*>0, and Y=1, if Y\*0  
We can calculate the estimated marginal effects from the formula given below, it is important to note that estimated marginal effects does not equal to just parameter estimates.

Marginal Effect of Xji = f () X [∂(/ Xji]. ………(1)

which is conditioned on some values of x and f () is pdf of a standard normal  
distribution evaluated at i : f () = (2Π)-1/2 exp[-1/2 i2]

If the relationship between Yi and Xji is linear, then: f () x j.  .…….(2)

j , represents the estimated coefficient of regressor.

Predicted probabilities for individual observations can be obtained by inputting the values of the independent variables into the estimated probit model.  
The formula for predicted Outcome, or, Probability that the Event Occurs is given by

= Prob (Yi=1/X) = F[]. ……….(3)

F[] is CDF of a standard normal distribution evaluated at :

F[] =

These predicted outcomes can be used for various purposes, such as evaluating the model's performance, understanding the impact of individual predictors, and making decisions based on the probability of the event occurring.

In this probit model, now we are discussing the absolute measures of fit. It is basically provided the statistics that determine how good the fit of a single model is in absolute terms. It involves four measures of fit in which 3 involve tests.

The Pseudo General F-test using the Likelihood Ratio (LR) Test Statistic in the context of a probit model is a method used to compare the goodness of fit. The formula for the LR test statistic is: LR = - 2 x (ln(L0) – (ln(L1))

Where the ln(L0) is the natural log of the likelihood of the restricted model under H0 ,ln(L1is the natural log of the likelihood of the unrestricted model under H1.

The null hypothesis of Pseudo General F-test is model does not fit the data well, with all slope coefficients being jointly zero. The alternative hypothesis suggests that the model is a good fit, with at least one coefficient significantly different from zero. If the p-value is less than 0.10 (at typical significance level), we can reject the null hypothesis, concluding that the model fits the data well.

Similarly, the Wald Test is used to examine the significance of individual parameters. The null hypothesis H0: βj = 0 (j= 1,2,3…), and the alternative hypothesis Ha : Ho is not true.

With a p-value below 0.10 indicating the rejection of the null hypothesis that a specific coefficient is zero, and hence, that the model is a good fit.

Hosmer and Lemeshow goodness of fit test have hypothesis statements different from previous ones H0 : The fit of the model is relatively good; Ha : The fit of the model is relatively poor. Considering p value if we fail to reject null hypothesis then conclude is fit of the model is relatively good.

The last one is percentage of correct predictions. A correct prediction for outcome Yi=1 occurs if Prob (Yi=1) > 0.5 and Yi= 1, and a correct prediction for outcome Yi = 0 occurs if Prob (Yi=1) < 0.5 and Yi= 0. We convert these above predictions into percentage. If both outcomes greater than or equal to 80%, then the fit is good. The model is not good fit if one is 80% and other is not. So, both must be greater than or equal to 80 % to be good fit.

When assessing how well our model fits the data, we employ relative measures of fit which is used to compare two or more competing models, but they cannot be used themselves. Among these measures are the information criterion and pseudo-R-squared.  
In the information criterion, we have Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

The formula for AIC = - 2 ln(L) + 2k. ………… (4)

, where *k* is the number of parameters in the model, and ln(*L*) is the log-likelihood of the model. This criterion is applied to each model under consideration, and the resulting AIC values guide us in determining the model's fit. Generally, the model with the lower AIC value is seen as having a relatively better fit. A model with a lower AIC value is preferred because it indicates a better balance between accurately fitting the data and maintaining simplicity, thereby reducing the risk of overfitting.

Similarly, Bayesian Information Criterion (BIC), the lower the BIC of the model is relative goodness of fit. The formula for BIC = *k* ln(n) - 2 ln(L) …………….(5)

where *n* is the number of observations, *k* is the number of parameters in the model, and ln(*L*) is the log-likelihood of the model.

Finally, as a part of relative measures of fit, we use Pseudo R-squared. In which McKelvey and Zavoina's R-squared and the Aldrich-Nelson R-squared used, a higher value indicates the better the model is deemed to be at explaining the variability of the binary response, thus suggesting a stronger relative fit among the models being compared.

EMPIRICAL ANALYSIS:   
In this study, we aim to systematically investigate the various factors that may significantly impact an individual's susceptibility to developing diabetes. Given below population regression model,

Yi\* = β1 + β2 Age\_40T49i + β3 Age\_50T59i +β4 Age\_MT60i + β5 junkfood\_dum + β6 Alcohol\_dum+ β7 RegularMedicine\_dum + β8 stress\_dum + β9 Preg\_dum + Εi. ……….(1)

Where β1 is intercept and β2, β3, β4, β5 ,β6,β7, β8, and β9 are the population parameters. Yi\* is latent variable. If *Yi*∗​>0, then *Yi*​=1 (the individual is diagnosed with diabetes). If *Yi*∗​≤0, then *Yi*​=0 (the individual is not diagnosed with diabetes). We estimate the population regression model through the maximum likelihood estimates.  
 = + Age\_40T49i + Age\_50T59i + Age\_MT60i + junkfood\_dum +

Alcohol\_dum+ RegularMedicine\_dum + stress\_dum + Preg\_dum …...…. (2)  
The observed binary outcome *Yi*​, which indicates whether an individual is diagnosed with diabetes (1) or not (0), is determined by whether the latent variable *Yi*∗​ crosses a certain threshold, typically zero.

= + 0.882048 **Age\_40T49i** + 0.861250 **Age\_50T59i** +1.738636 **Age\_MT60i** +

(<.0001 ) (<.0001) (<.0001) (<.0001)

0.390475 **junkfood\_dum** +0.239628 **Alcohol\_dum**+.266737 **RegularMedicine\_dum** +

(<.0418) (<.0868) (<.0001)

0.290078 **stress\_dum** + 0.391861 **Preg\_dum**  …...…. (3)

(<0.0204). (<0.0052)

In our probit model, we obtained parameter estimates through the QLIM procedure. The numbers in parentheses next to each regressor represent their respective p-values. The null hypothesis for each parameter is that it is not significant (βj=0 for j=1,2, ...7,8,9), while the alternative hypothesis posits that the parameter is indeed significant. If a p-value falls below the threshold of 0.10, it leads to the rejection of the null hypothesis, thereby indicating that the parameter is significant. From the estimated regression model presented, it is evident that all the parameters are significant when evaluated at a 90% confidence level. Overall, the signs of the coefficients appear to conform to the expected directions. Older age, junk food consumption, alcohol consumption, regular medication usage, stress, and pregnancy status all show a positive relationship with the risk of developing diabetes. The statistical significance of these variables (as determined by their p-values) will help in understanding the robustness of these relationships. This model provides valuable insights into the factors that could potentially influence an individual's risk of developing diabetes.

In terms of absolute measures of fit, the Likelihood Ratio Test (LR Test) is particularly compelling. With a test statistic of 479.8022 and a p-value less than 0.0001, it strongly rejects the null hypothesis of the Pseudo General F-test. This result indicates that the model with covariates offers a significantly better fit than a model without them, affirming its effectiveness.  
 In the context of the Wald test results from our regression analysis, for the age categories (40-49, 50-59, and over 60 years), junk food consumption ('junkfood\_dum'), alcohol consumption ('Alcohol\_dum'), regular medication use ('RegularMedicine\_dum'), stress ('stress\_dum'), and pregnancy ('Preg\_dum'), the p-values are all below the conventional significance levels (such as 0.05 or 0.10). A p-value below these thresholds indicates that there is a statistically significant difference between the estimated coefficient and zero, leading to the rejection of the null hypothesis for each of these variables. Therefore, since the null hypothesis is rejected for all these parameters, it indicates that each of these factors has a statistically significant effect on the probability of an individual being diagnosed with diabetes. This collective significance of all parameters underscores the conclusion that the model is a good fit for analyzing the risk factors associated with diabetes.

The Hosmer and Lemeshow Test further supports the model's adequacy. With a test statistic of 7.2453 and a p-value of 0.4038, the test fails to reject its null hypothesis, suggesting that the model fits the data well. This is particularly reassuring as it shows the model's reliability in representing the observed data. The percentage of correct predictions reveals specificity of 89.1%, exceeding the 80%. However, it also highlights an area for improvement, with a sensitivity of 69.2%, which is below the 80% mark. This suggests that while the model is generally robust in predicting outcomes, its ability to correctly identify diabetes cases could be enhanced.   
 In our regression analysis, the Goodness-of-Fit Measures, specifically McFadden's LRI (0.4264) and McKelvey-Zavoina (0.5672), play a crucial role in assessing the model's performance. For both of these measures, a higher value is indicative of a better model fit. McFadden's LRI, with a value of 0.4264, suggests a moderate level of explanatory power. This measure is a relative index of the model's goodness-of-fit and is particularly useful in the context of models with binary outcomes. It compares the likelihood of the model with the covariates to a null model, with higher values indicating a better fit. Similarly, the McKelvey-Zavoina R-squared value of 0.5672 also demonstrates a good level of explanatory power. This measure is a pseudo R-squared statistic tailored for binary choice models like the probit model you are using. It provides an estimate of the proportion of the variance in the dependent variable that is accounted for by the model. In your case, a value of 0.5672 indicates that the model does a reasonably good job of explaining the variability in the binary response of diabetes diagnosis. Together, these measures suggest that your model has a good capacity to explain the factors influencing the risk of diabetes, capturing a significant portion of the variability in the outcome. This reinforces the model's effectiveness in identifying key risk factors associated with diabetes.

Regarding relative measures of fit, the Information Criteria, comprising AIC (663.49349) and BIC (707.18268), are crucial for model comparison. Although these criteria are more suited for evaluating relative fit among different models, the lower values in this case suggest a balanced model that accurately fits the data while avoiding overfitting.

In conclusion, the combination of these various tests and measures converges to suggest that the model is a good fit for the data. It effectively captures the essential factors influencing diabetes risk, with room for refinement, particularly in its sensitivity. This comprehensive evaluation assures the model's reliability in explaining the risk factors for diabetes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 6 | | | | |
|  | Marginal effects | | | |
| Variables | Obs no. 1 | Obs no. 180 | Obs no.539 | Obs no. 948 |
| Age\_40T49 | **0.15906** | **0.03708** | **-0.35188** | **-0.15594** |
| Age\_50T59 | **0.15531** | **0.03621** | -0.34358 | -0.15226 |
| Age\_MT60 | **0.31353** | 0.07309 | **-0.6936** | -0.30737 |
| Alcohol\_dum | **0.04321** | **0.010074** | -0.095596 | -0.042364 |
| RegularMedicine\_dum | **0.22843** | **0.05325** | **-0.50534** | **-0.22395** |
| junkfood\_dum | **0.07042** | **0.01642** | -0.15577 | -0.06903 |
| stress\_dum | **0.05231** | 0.01219 | **-0.11572** | -0.05128 |
| Preg\_dum | **0.07067** | **0.01647** | -0.15633 | -0.06928 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE 7 | | | | | | | | | |
| Observation no. | Age\_LT40 | Age\_40T49 | Age\_50T59 | Age\_MT60 | Alcohol\_dum | RegularMedicine\_dum | junkfood\_dum | stress\_dum | Preg\_dum |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 180 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 539 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 948 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |

In a probit model, parameter estimates do not directly provide the estimated marginal effects. These effects are based on the size of the estimated coefficients from the model and how the factors are spread out or vary. Table 6, which contains the estimated marginal effects of the regressors for observations 1, 180, 539, and 948. Table 7 provides the values of regressors for observations 1,180,539, and 948.

The marginal effects for variables for observation #1 where Yi=1 (person is diagnosed with diabetes). To give you the interpretation of person like #1 : an Individual falls within the 50 to 59-year-old age range. They have not experienced pregnancy, do not consume alcohol, and refrain from eating junk food. Furthermore, they do not take any medications on a regular basis. ( Age\_40T49= 0 , Age\_50T59 = 1, Age\_MT60= 0, Alcohol\_dum= 0, RegularMedicine\_dum = 0, junkfood\_dum= 0, stress\_dum Preg\_dum=0)

* **Age\_40T49 (0.15906)**: An individual like #1, who is not in the 40 to 49-year-old age range, does not have a 15.906% higher probability of being diagnosed with diabetes that is associated with this age group, all else held constant.
* **Age\_50T59 (0.15531)**: An individual like #1, who is within the 50 to 59-year-old age range, has a 15.531% higher probability of being diagnosed with diabetes that is associated with this age group, all else held constant.
* **Age\_MT60 (0.31353)**: An individual like #1, who is not 60 years or older, does not have a 31.353% higher probability of being diagnosed with diabetes that is typically associated with this older age group, all else held constant.
* **Alcohol\_dum (0.043213)**: An individual like #1, who does not consume alcohol, does not have the 4.3213% higher probability of being diagnosed with diabetes that is associated with alcohol consumption, all else held constant.
* **RegularMedicine\_dum (0.22843)**: An individual like #1, who does not take any medications regularly, does not have the 22.843% higher probability of being diagnosed with diabetes that is associated with regular medication use, all else held constant.
* **junkfood\_dum (0.07042)**: An individual like #1, who refrains from consuming junk food, does not have the 7.042% higher probability of being diagnosed with diabetes that is associated with this factor, all else held constant.
* **stress\_dum (0.05231)**: An individual like #1, who experiences stress, has a 5.231% higher probability of being diagnosed with diabetes that is associated with experiencing stress, all else held constant.
* **Preg\_dum (0.07067)**: An individual like #1, who has not been pregnant, does not have the 7.067% higher probability of being diagnosed with diabetes that is associated with pregnancy, all else held constant.

The marginal effects for variables for observation #180 where Yi=1 (person is diagnosed with diabetes) To give you the interpretation of person like #180 : The individual falls within the over 60 age category. They have a history of alcohol consumption and are on regular medication. Stress is also a factor in their profile, but they do not consume junk food and have not been pregnant. ( Age\_40T49= 0 , Age\_50T59 = 0, Age\_MT60= 1, Alcohol\_dum= 1, RegularMedicine\_dum = 1, junkfood\_dum= 0, stress\_dum= 1, Preg\_dum=0)

* **Age\_MT60 (0.07309)**: An individual like #180, who is over 60 years old, has a 7.309% higher probability of being diagnosed with diabetes, all else held constant.
* **Alcohol\_dum (0.010074)**: An individual like #180, who has a history of alcohol consumption, has a 1.0074% higher probability of being diagnosed with diabetes, all else held constant.
* **RegularMedicine\_dum (0.05325)**: An individual like #180, who is on regular medication, has a 5.325% higher probability of being diagnosed with diabetes, all else held constant.
* **stress\_dum (0.01219)**: An individual like #180, who experiences stress, has a 1.219% higher probability of being diagnosed with diabetes, all else held constant.

For the other variables where the individual does not possess the traits, the marginal effects indicate that there is no additional probability of being diagnosed with diabetes related to those specific factors, as they are not present in the individual's profile:

* **Age\_40T49 (0.03708)**: Since the individual like #180 is not in the 40 to 49 age group, this factor does not increase their probability of being diagnosed with diabetes.
* **Age\_50T59 (0.03621)**: As the individual like #180 is not within the 50 to 59 age bracket, this factor does not contribute to an increased probability of being diagnosed with diabetes.
* **junkfood\_dum (0.01642)**: Since the individual like #180 does not consume junk food, this factor does not raise their probability of being diagnosed with diabetes.
* **Preg\_dum (0.01647)**: As the individual like #180 has not been pregnant, this condition does not increase their probability of being diagnosed with diabetes.

TABLE 7 shows the table for marginal effects for variables for observation #539where Yi=0 (person is not diagnosed with diabetes) To give you the interpretation of person like #539 : The individual falls within the 50 to 59 age category and is on regular medication. They do not consume alcohol, junk food, and have not experienced stress or been pregnant. (Age\_40T49= 0 , Age\_50T59 = 1, Age\_MT60= 0, Alcohol\_dum= 0, RegularMedicine\_dum = 1, junkfood\_dum= 0, stress\_dum= 0, Preg\_dum=0)

* **Age\_40T49 (-0.35188)**: A person like #539, who is not in the 40 to 49 age range, is 35.188% less likely to be diagnosed with diabetes compared to those in this age range, all else held constant.
* **Age\_50T59 (-0.34358)**: A person like #539, who is in the 50 to 59 age range, is 34.358% less likely to be diagnosed with diabetes compared to those not in this age range, all else held constant.
* **Age\_MT60 (-0.6936)**: A person like #539, who is not over 60 years old, is 69.36% less likely to be diagnosed with diabetes compared to those who are over 60, all else held constant.
* **junkfood\_dum (-0.15577)**: A person like #539, who does not consume junk food, is 15.577% less likely to be diagnosed with diabetes compared to those who consume junk food, all else being equal.
* **Alcohol\_dum (-0.095596)**: A person like #539, who does not consume alcohol, is 9.5596% less likely to be diagnosed with diabetes compared to those who do consume alcohol, with no changes in other factors.
* **RegularMedicine\_dum (-0.50534)**: A person like #539, who takes regular medication, is 50.534% less likely to be diagnosed with diabetes compared to those who do not take regular medication, assuming other variables remain unchanged.
* **stress\_dum (-0.11572)**: A person like #539, who does not experience stress, is 11.572% less likely to be diagnosed with diabetes compared to those who experience stress, with all other factors held constant.
* **Preg\_dum (-0.15633)**: A person like #539, who has not been pregnant, is 15.633% less likely to be diagnosed with diabetes compared to those who have been pregnant, given all other variables are held constant.

The marginal effects for variables for observation #948 where Yi=0 (Individual is not diagnosed with diabetes) To give you the interpretation of person like #948 : The individual falls within the 50 to 59 age category and is on regular medication. They do not consume alcohol, junk food, and have not experienced stress or been pregnant. (Age\_40T49= 0 , Age\_50T59 = 1, Age\_MT60= 0, Alcohol\_dum= 0, RegularMedicine\_dum = 1, junkfood\_dum= 0, stress\_dum= 0, Preg\_dum=0)

* **Age\_40T49 (-0.15594)**: A person like #948, who is not in the 40 to 49 age range, is 15.594% less likely to be diagnosed with diabetes compared to those who are in this age range, all else held constant.
* **Age\_50T59 (-0.15226)**: A person like #948, who is in the 50 to 59 age range, is 15.226% less likely to be diagnosed with diabetes compared to those who are not in this age range, all else held constant.
* **Age\_MT60 (-0.30737)**: A person like #948, who is not over 60 years old, is 30.737% less likely to be diagnosed with diabetes compared to those who are over 60, all else held constant.
* **junkfood\_dum (-0.06903)**: A person like #948, who does not consume junk food, is 6.903% less likely to be diagnosed with diabetes compared to those who consume junk food, all else being equal.
* **Alcohol\_dum (-0.042364)**: A person like #948, who does not consume alcohol, is 4.2364% less likely to be diagnosed with diabetes compared to those who do consume alcohol, with no changes in other factors.
* **RegularMedicine\_dum (-0.22395)**: A person like #948, who takes regular medication, is 22.395% less likely to be diagnosed with diabetes compared to those who do not take regular medication, assuming other variables remain unchanged.
* **stress\_dum (-0.05128)**: A person like #948, who does not experience stress, is 5.128% less likely to be diagnosed with diabetes compared to those who do experience stress, with all other factors held constant.
* **Preg\_dum (-0.06928)**: A person like #948, who has not been pregnant, is 6.928% less likely to be diagnosed with diabetes compared to those who have been pregnant, given all other variables are held constant.

|  |  |  |
| --- | --- | --- |
| TABLE .8 | | |
| **Observation** | **Y** | **Predicted probability** |
| 710 | 0 | 0.10101 |
| 405 | 0 | 0.0169 |
| 942 | 1 | 0.7307 |
| 943 | 1 | 0.489 |

TABLE 8 shows the predicted of the observations 710, 405, 942, and 943.  
 Interpretation of the predicted outcome of individual #710 who is not diagnosed with diabetes is 60 years old, with a history of medication use but no alcohol or junk food consumption, not experiencing stress, and having been pregnant (Age\_40T49=0, Age\_50T59= 0, Age\_MT60 = 1, Alcohol\_dum = 0, RegularMedicine\_dum = 1, junkfood\_dum = 0, stress\_dum = 0, Preg\_dum = 1.) the probability that this person would be diagnosed with diabetes is 10.10%.

Interpretation of the predicted outcome of individual #405 who is not diagnosed with diabetes is below 40 with no history of medication use , no alcohol or junk food consumption, not experiencing stress, and have never been pregnant (Age\_40T49=0, Age\_50T59= 0, Age\_MT60 = 0, Alcohol\_dum = 0, RegularMedicine\_dum = 0, junkfood\_dum = 0, stress\_dum = 0, Preg\_dum = 0.) the probability that this person would be diagnosed with diabetes is 1.69 %.

Interpretation of the predicted outcome of individual # 942 who is diagnosed with diabetes and is not within the age categories of 40-49, 50-59, or over 60. They have a history of alcohol consumption and are on regular medication, but do not consume junk food, have not experienced stress, and have not been pregnant. (Age\_40T49=0, Age\_50T59= 0, Age\_MT60 = 1, Alcohol\_dum = 1, RegularMedicine\_dum = 0, junkfood\_dum = 0, stress\_dum = 0, Preg\_dum = 1.) the probability that this person would be diagnosed with diabetes is 73.07%.

Interpretation of the predicted outcome of individual #943 who is diagnosed with diabetes and falls within the 40 to 49 age category. They consume alcohol but do not use regular medication, consume junk food, experience stress, or have been pregnant. (Age\_40T49=0, Age\_50T59= 0, Age\_MT60 = 1, Alcohol\_dum = 1, RegularMedicine\_dum = 0, junkfood\_dum = 0, stress\_dum = 0, Preg\_dum = 1.) the probability that this person would be diagnosed with diabetes is 48.90%.

These interpretations of predicted probabilities serve to highlight the nuanced relationship between demographic factors, lifestyle choices, and the likelihood of a diabetes diagnosis. The varying probabilities across different profiles suggest that age, medication use, alcohol consumption, and pregnancy history are influential factors in diabetes risk. For example, the lower probability for a younger individual with no adverse health behaviors implies that early intervention and lifestyle management may significantly reduce diabetes risk. Conversely, the higher probability associated with older age groups and certain behaviors like alcohol consumption points to the need for targeted healthcare strategies for these populations. These insights can inform healthcare providers and policymakers in developing tailored diabetes prevention and management programs, ultimately aiming to reduce the prevalence of diabetes.

HETEROSKEDASTICITY:  
Heteroskedasticity refers to the circumstance in which the variance of the errors in a regression model is not constant across all levels of the independent variables. The presence of heteroskedasticity can lead to inefficient estimations and can bias the standard errors, thereby compromising the integrity of hypothesis tests. A common method to test for heteroskedasticity is the Breusch-Pagan test, which assesses the squared residuals against the independent variables.   
Null hypothesis H0: = 0 from which var(εi)= σ2(1+ 1).  
There is no heteroskedasticity, implying that the variance of the errors is constant across all levels of the independent variables.  
Alternative Hypothesis H1: H0 is not true, there is heteroskedasticity.  
If the p-value from the Breusch-Pagan test is less than the significance level (commonly 0.10, 0.05, or 0.01), the null hypothesis is rejected, indicating the presence of heteroskedasticity. This suggests that the variability of the errors in the model is influenced by the values of the independent variables. If the p-value is greater than the significance level, the null hypothesis is not rejected, and it is concluded that there is no evidence of heteroskedasticity.  
 In the empirical analysis of our probit model, to check heteroskedasticity we conducted the Breusch-Pagan test, particularly focusing on the variables `alcohol\_dum` and `junkfood\_dum`. The results yielded p-values of 0.9994 for `\_H.alcohol\_dum` and 0.9997 for `\_H.junkfood\_dum`, far exceeding the 0.10 significance level. These high p-values suggest that there is no statistical evidence to reject the null hypothesis of homoskedastic errors in the context of these variables. Given this lack of evidence for heteroskedasticity, it is appropriate to rely on the original uncorrected probit estimates. These findings underpin the robustness of our model's assumptions and the reliability of the estimated coefficients and their standard errors, which are based on the premise of constant error variance across observations.

TABLE 9 provides the parameters and its uncorrelated standard errors.

|  |  |
| --- | --- |
| TABLE 9 | |
| **Parameter** | **Standard Error** |
| Intercept | 0.496761 |
| Age\_40T49 | 0.170872 |
| Age\_50T59 | 0.175892 |
| Age\_MT60 | 0.206097 |
| Alcohol\_dum | 0.173883 |
| RegularMedicine\_dum | 0.198812 |
| junkfood\_dum | 0.134649 |
| stress\_dum | 0.13026 |
| Preg\_dum | 0.178264 |

CONCLUSIONS AND POLICY RECOMMENDATIONS:  
Using a probit regression model, our econometric analysis of the diabetes risk factors has revealed several important variables. People become more susceptible to diabetes as they become older, especially beyond the age of 40. Our model supports this tendency, showing positive coefficients for 40–49, 50–59, and 60+ age groups. Lifestyle decisions that greatly increase the risk include consuming junk food and alcohol on a regular basis. Furthermore, using medications daily, which is frequently a stand-in for underlying medical issues, is a powerful indicator of diabetes. Stress has also been found to be a positive risk factor, and a history of pregnancy, possibly including episodes of gestational diabetes, is linked to a higher chance of being diagnosed with diabetes.

Given these insights, we put forth a suite of policy recommendations to mitigate these risk factors. We propose the establishment of routine diabetes screenings, especially for those in the 40-plus age demographic. It is imperative for public health initiatives to foster education on the importance of moderating alcohol intake and reducing consumption of unhealthy foods, while also promoting healthier dietary practices and regular physical activity. It is equally important for those on chronic medication regimens to receive ongoing evaluation and management to lessen diabetes risks. Incorporating stress management techniques into overall public health policies is crucial, considering the impact of stress on diabetes. Pregnant individuals should be provided with comprehensive information regarding the risks and management of gestational diabetes to prevent future diabetes risk. Provision of nutrition counseling and meal planning assistance can play a vital role for those at heightened risk of diabetes. Furthermore, advocating for corporate wellness programs can create supportive environments that encourage stress alleviation and endorse healthier lifestyles.

These policy suggestions aim to not just curb the rise of diabetes but also promote overall well-being, which could translate into lower healthcare costs and a healthier populace. Tailoring these strategies to fit the specific needs of different communities and monitoring their effectiveness is critical to reducing the prevalence of diabetes.

SHORTCOMINGS:  
While this study provides valuable insights into the risk factors associated with diabetes, there are some limitations that should be acknowledged. First, the data used in the analysis comes from a single survey, which may not be fully representative of the broader population. Expanding the dataset to include other surveys or sources could improve generalizability. The model incorporates a limited set of independent variables concentrated largely on lifestyle and demographic factors. Incorporating clinical information, family medical histories, genetic data, and more extensive dietary details could allow for a more comprehensive analysis. Related factors like income level, education, ethnicity, and psychosocial elements may also impact susceptibility.

REFERENCES:

* Baliunas, D. O., Taylor, B. J., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). "Alcohol as a Risk Factor for Type 2 Diabetes: A systematic review and meta-analysis," Diabetes Care, Vol. 32, No. 11, pp. 2123-2132. DOI: 10.2337/dc09-0227.
* Bellamy, L., Casas, J. P., Hingorani, A. D., & Williams, D. (2009). "Type 2 diabetes mellitus after gestational diabetes: A systematic review and meta-analysis," The Lancet, Vol. 373, No. 9677, pp. 1773-1779. DOI: 10.1016/S0140-6736(09)60731-5.
* Bliss, C. I. (1934). "The Method of Probits," Science, Vol. 79, No. 2037, pp. 38-39. DOI: 10.1126/science.79.2037.38.
* Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). "Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin," New England Journal of Medicine, Vol. 346, No. 6, pp. 393-403. DOI: 10.1056/NEJMoa012512.
* Tigganeha4. (2019). "Diabetes Dataset 2019," Kaggle. Available at: <https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019>.
* Wooldridge, Jeffrey M. (2015). Introductory Econometrics: A Modern Approach. 6th ed., Cengage Learning.