Predicting Diabetes Status Using Logistic Regression on NHANES Public Health Data

Caleb Frankenberger
Radford University
cfrankenberger@radford.edu
December 3, 2024

Abstract

Diabetes is one of the most prevalent diseases in the United States, with an estimated 38% of U.S. adults having prediabetes. According to the Centers for Disease Control and Prevention (CDC), 8.7 million adults have undiagnosed diabetes, while 29.7 million are diagnosed with the condition. This paper utilizes data from the National Health and Nutrition Examination Survey (NHANES) to develop a logistic regression model aimed at identifying key determinants of diabetes status. Understanding these determinants is essential for raising awareness about diagnosed, undiagnosed, and prediabetes conditions. This study focuses on using non-invasive predictors, such as demographic and lifestyle information, to provide an accessible alternative to traditional bloodwork screenings.

1. Introduction

Regression analysis is a widely used statistical method for analyzing the relationship between a dependent (response) variable and various independent variables, known as predictors. Logistic regression, a specialized form of regression analysis, is used to predict binary outcomes with two possible values, such as “positive” or “negative”. This makes it particularly useful in public health, where critical outcomes such as survival after surgery, HIV disease status, or patient relapse, can be represented using binary variables. These outcomes are often influenced by patient demographics, medical history, and lifestyle characteristics. By estimating the probability of a binary outcome, logistic regression is well-suited for this study, which focuses on predicting diabetes status.

Diabetes is traditionally diagnosed through blood work, which can be resource-intensive, invasive, and inaccessible for some populations. This study aims to develop a model for predicting diabetes status using non-invasive predictors, such as demographic and lifestyle factors, as an alternative method for early risk assessment. If effective, the model could support the development of a remote, low-cost diabetes screener that reduces both cost and reliance on laboratory tests.

In their 2002 study, Tabaei and Herman developed a multivariate logistic regression model to screen for undiagnosed diabetes, incorporating demographic factors (age, sex, BMI) and laboratory measures (random capillary plasma glucose, postprandial time) (Tabaei & Herman, 2002). Their model demonstrated strong predictive power, with an area under the receiver operating curve (AUC) of 0.88. However, their model’s reliance on laboratory data restricts its use to clinical settings. In contrast, the present study aims to evaluate diabetes risk using only non-invasive predictors, such as demographic and lifestyle factors. This approach seeks to determine whether a fully non-invasive model can achieve comparable performance to Tabaei and Herman’s method, while enhancing accessibility and practicality.

The data used for this analysis comes from the National Health and Nutrition Examination Survey (NHANES). NHANES is a comprehensive survey conducted by the National Center for Health Statistics that collects data on the health and nutrition of the U.S. population. It is one of the most trusted and widely used datasets for public health research. The survey combines interviews, laboratory tests, physical examinations, and questionnaires to provide a representative sample of the population. The data are publicly available and widely used in public health research and analysis. The specific dataset for this analysis is a subset of the *NHANES 2017–March 2020* survey data.

The subset contains the following variables identified as possible determinants of a positive diabetes status: *age*, *body mass index*, *diabetes status*, *education level*, *hypertension*, and *race*. These variables, along with an explanation, variable type, and calculation procedure, are shown with detail in Appendix A.

To ensure that the findings are representative of the U.S. population, the analysis incorporates NHANES sample weights, strata, and cluster variables. These design elements account for the complex survey design, to provide unbiased estimates and standard errors.

This report begins with a discussion of the methodology and model development (Section 2). Section 3 presents early data analysis, including cleaning procedures, descriptive statistics, and early modeling results. Finally, Section 4 summarizes the major findings, discusses limitations, and explores potential extensions of this work.

2. Methodology

In this study, logistic regression is used to model the relationship between the binary dependent variable diabetes status, and the independent predictor variables listed in Section 1. Diabetes status will be represented as Y, where:

$$Y = \begin{Bmatrix} 1, if\ the\ individual\ has\ diabetes, \\ 0, \quad otherwise \end{Bmatrix}$$

The goal is to model the probability of $Y = 1$, denoted by $P(Y = 1 \mid \boldsymbol{X})$. Here, $\boldsymbol{X}$ is a row vector $(x_1, x_2, \dots, x_k)$, where each element is a predictor variable. This probability can be expressed using the logistic regression function:

$$P(Y = 1 \mid \boldsymbol{X}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}$$

Five assumptions are made for the logistic regression model:

1. **Binary Outcome Variable**: The dependent variable Y is binary, taking either 1 or 0 as a possible value.
2. **Independent Observations**: Each observation in the dataset is independent from the others. Note: The NHANES dataset, being a complex survey, does not provide independent observations. Instead, NHANES provides strata, cluster, and weight data to make inferences about the general U.S. population.
3. **Little/No Multicollinearity**: Predictor variables are not highly correlated with each other, ideally having no multicollinearity at all.
4. **Linear Relationship to Log-Odds**: Independent variables should be linearly related to the log-odds of the dependent variable.
5. **Sufficient Sample Size**: Logistic regression assumes sufficiently large sample size for reliable estimations.

Parameters were estimated using the PROC SURVEYLOGISTIC procedure in SAS. This technique employs Maximum Likelihood Estimation (MLE) while incorporating the complex survey design of the NHANES dataset. Maximum Likelihood Estimation identifies the parameter values that maximize the likelihood of observing the given data, in this case, diabetes status.

Each parameter estimate (β_j) in the logistic regression model represents the change in log-odds of diabetes status, for each one-unit increase in the corresponding predictor variable (x_j), when all other predictors are held constant. The log-odds can be exponentiated to obtain an odds ratio:

$$Odds\ Ratio = e^{\beta_j}$$

For example, if $\beta_1\ (age) = 0.5$, the odds ratio is $e^{0.5} \approx 1.65$, meaning for every one-unit increase in age, the odds of having diabetes are multiplied by 1.65, corresponding to a 65% increase in the odds.

The output of the model, expressed as log-odds, provides a quantitative measure for the likelihood of diabetes status for an individual given, their specific predictor values. For example, if the model predicts a probability greater than 0.5 (i.e., $P(Y = 1 \mid \boldsymbol{X}) > 0.5$), then we can classify anyone with a predicted probability above this threshold as having diabetes. This threshold can be adjusted depending on the desired balance between specificity (false positives) and sensitivity (false negatives).

To compare against the Tabaei and Herman model, we will look at the c-statistic and concordance percentage, which are widely used in logistic regression to assess the discriminatory power of a model. The c-statistic, equivalent to the area under the receiver operating characteristic (ROC) curve, measures the model's ability to correctly classify positive and negative outcomes. The concordance percentage reflects the proportion of correctly ranked pairs of observations.

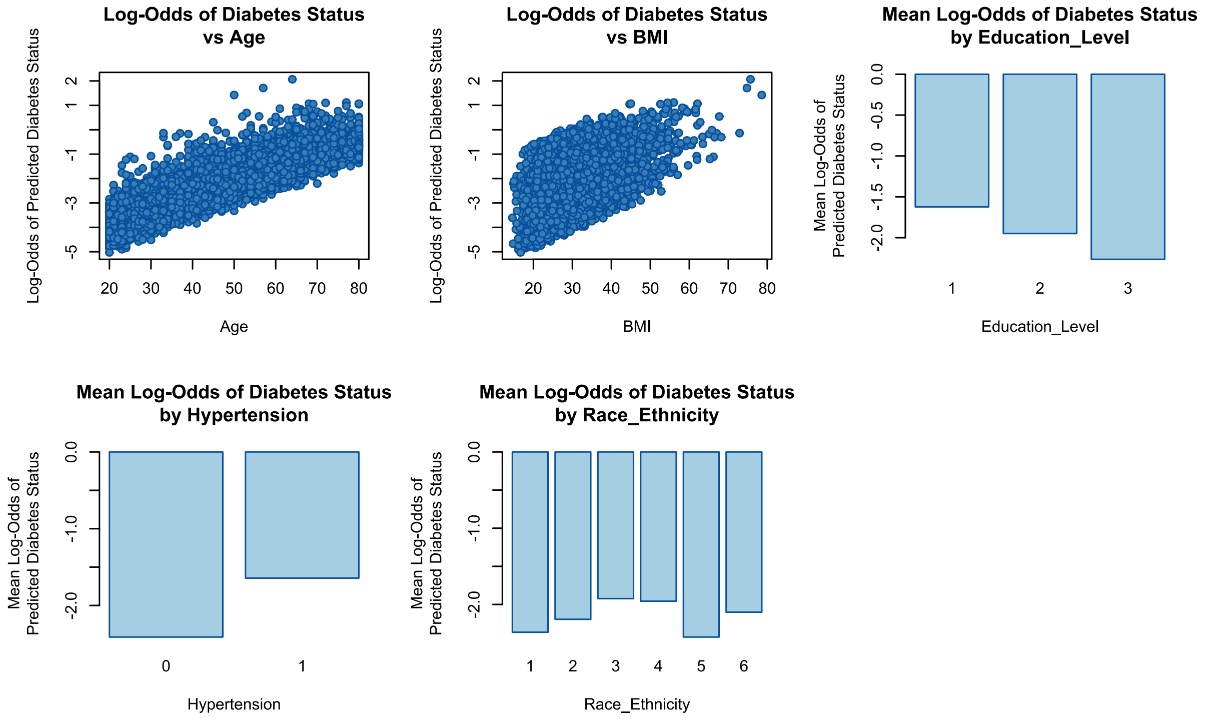
3. Data Analysis

To verify the low multicollinearity assumption, the General Variance Inflation Factor (GVIF) values for each predictor were computed using the *car* package in R (Fox, Weisberg, & Price, 2024). The GVIF values are shown below:

General Variance Inflation Factors					
Predictor:	Age	BMI	Education Level	Hypertension	Race
GVIF:	1.100174	1.069665	1.050699	1.042695	1.051179

All predictors had a GVIF of ≤ 3, indicating no significant multicollinearity among the predictors in the model.

To assess the linearity assumption between the predictors and the log-odds of the response, a matrix of plots was created. Continuous variables were visualized with scatterplots of each variable against the log-odds of the response (diabetes status). Categorical variables were represented using barplots of the mean log-odds of diabetes status across their respective levels.



The plots indicate clear linear relationships between diabetes status and both continuous variables, age and BMI. For the categorical variables, education level and hypertension show distinct linear-like trends in the mean log-odds. However, race does not display a clear pattern. The plot suggests that diabetes status varies across racial categories.

Originally, a full model was fitted including all first-order predictors. From this model, we examined the contribution of each predictor individually. At a significance level of 0.05, all predictors were significant, with $p \leq 0.05$. The full diagnostic results from this model can be seen in Appendix B. However, the inclusion of race and education level increased the complexity of the model, as both were coded into multiple dummy variables. To achieve a more parsimonious model, race and education level were excluded, and the model was re-fit with only age, BMI, and hypertension as predictors of diabetes status.

The new model produced the following output:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.4	**Somers' D**	0.533
Percent Discordant	23.2	**Gamma**	0.535
Percent Tied	0.4	**Tau-a**	0.142
Pairs	7198970	**c**	0.766

Analysis of Maximum Likelihood Estimates				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
NOTE: The degrees of freedom for the t tests is 25.				
Intercept	-7.8064	0.2867	-27.23	<.0001
Age	0.0592	0.00297	19.95	<.0001
BMI	0.0790	0.00577	13.68	<.0001
Hypertension	0.2474	0.0731	3.39	0.0023

With the goal of comparing against the Tabaei and Herman model in mind, the c-statistic is 0.766, and the percent concordant is 76.4%. The parameter coefficient estimates are shown in the accompanying table, where all predictors remain significant ($p < 0.05$). This indicates that age, BMI, and hypertension all contribute significantly to a positive diabetes status. Using the same table, we can derive the estimated regression function for the new model, as follows:

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \widehat{\beta_3}x_3 \,,$$

where $x_1 = Age,\ x_2 = BMI,\ x_3 = Hypertension.$

After plugging in the parameter estimates, we get the following:

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -7.8064 + 0.0592x_1 + 0.0790x_2 + 0.2474x_3$$

Expressed in probability form (as seen in Section 1):

$$\widehat{P}(Y = 1 \mid x_1, x_2, x_3) = \frac{e^{-7.8064+0.0592x_1+0.0790x_2+0.2474x_3}}{1+e^{-7.8064+0.0592x_1+0.0790x_2+0.2474x_3}}$$

By excluding race and education level, the model is significantly simpler while retaining reasonably strong predictive power.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Confidence Limits**	
NOTE: The degrees of freedom in computing the confidence limits is 25.			
Age	1.061	1.054	1.067
BMI	1.082	1.069	1.095
Hypertension	1.281	1.102	1.489

The odds ratio, as explained in Section 2, provides a clear interpretation of how each predictor influences the likelihood of diabetes. The table above, obtained from the SAS output, shows these estimates along with their 95% confidence limits.

- **Age**: For every additional year of age, the odds of having diabetes increase by 6.1%.
- **BMI**: For every one-unit increase in BMI, the odds of having diabetes rise by 8.2%.
- **Hypertension**: Having hypertension raises the odds of having diabetes by 28.1%.

Interestingly, while hypertension shows the largest increase in the odds compared to the other predictors, it also has the highest p-value when testing for predictor significance.

4. Discussion

Compared to the model by Tabaei and Herman (2002), which incorporated both demographic and laboratory data, the current model offers a more accessible (and scalable) alternative by relying solely on non-invasive predictors such as age, BMI, and hypertension. While this approach sacrifices some predictive power, it improves in practicality and in the feasibility for implementation. Age, BMI, and hypertension are simple and inexpensive to measure and do not require trained medical professionals, unlike laboratory-based measurements such as blood work. The model achieved a c-statistic of 0.766 and a concordant percentage of 76.4%, indicating reasonably strong discriminatory power. Although this performance falls slightly short of Tabaei and Herman's model (c-statistic: 0.88), the tradeoff reflects the balance between simplicity and accuracy.

The non-invasive nature of this model represents an important step toward creating a cost-effective and accessible diabetes risk screener. By using easily obtainable inputs, such as age, BMI, and hypertension status, the model could be used to create a simple screening tool for early risk detection of diabetes. Users could input their own values, and the screener could then inform them to seek further evaluation from a healthcare provider depending on their results, acting as an early warning system for diabetes.

References

[1] Centers for Disease Control and Prevention. (2023). National diabetes statistics report, 2023. Retrieved [November 19, 2024], from https://www.cdc.gov/diabetes/php/data-research/index.html

[2] Fox, J., Weisberg, S., & Price, B. (2024). *Companion to Applied Regression (R package version 3.1-3)*. Retrieved from https://CRAN.R-project.org/package=car

[3] National Health and Nutrition Examination Survey (NHANES). (n.d.). Retrieved [November 19, 2024], from https://www.cdc.gov/nchs/nhanes/index.htm

[4] R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

[5] SAS Institute Inc. (2024). SAS Software Version 9.4. SAS Institute Inc., Cary, NC.

[6] Tabaei, B. P., & Herman, W. H. (2002). A multivariate logistic regression equation to screen for diabetes: Development and validation. Diabetes Care, 25(11), 1999–2003. https://doi.org/10.2337/diacare.25.11.1999

Appendix A – Detailed Variable Descriptions

This appendix provides detailed descriptions of the variables used in the analysis, including their types, calculation methods, and the files from which they were derived. The data used in this analysis were retrieved from the following NHANES XPT files*: P_BMX*, *P_BPXO*, *P_DEMO*, and *P_DIQ*.

Variable	**Type**	**Description**	**Calculation**
Age	Continuous	The age of the individual, in years.	Taken directly from the NHANES variable RIDAGEYR.
Body Mass Index	Continuous	A measure of body fat, calculated using height and weight.	Taken directly from the NHANES variable BMXBMI.
Diabetes Status	Binary Categorical	Identifies whether the individual has been diagnosed with diabetes.	Recoded from the NHANES variable DIQ010. Assigned as 1 for positive diabetes status and zero for negative.
Education Level	Categorical	The highest level of education attainted by the individual	Recoded from the NHANES variable DMDEDUC2 into three groups: less than high school (1), high school graduate (2), and college graduate or above (3).
Hypertension	Binary Categorical	Whether the individual is hypertensive, calculated through blood pressure measurements.	Derived from systolic (BPXOSY1, BPXOSY2, BPXOSY3) and diastolic (BPXODI1, BPXODI2, BPXODI3) blood pressure measurements. Measurements were averaged across up to three readings, ignoring missing values. Hypertension was defined as systolic ≥ 130 mmHg or diastolic ≥ 80 mmHg.
Race	Categorical	The race of the individual	Recoded from the NHANES variable RIDRETH3 into six categories: 1 = Mexican American 2 = Other Hispanic 3 = Non-Hispanic White 4 = Non-Hispanic Black 5 = Non-Hispanic Asian 6 = Other Race (including multi-racial).

Appendix B – Full Model Diagnostics

A full model including all first-order predictors was originally fitted using SAS. From the output, several diagnostic plots are generated which can provide useful insight into the model:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.9	**Somers' D**	0.543
Percent Discordant	22.7	**Gamma**	0.545
Percent Tied	0.4	**Tau-a**	0.145
Pairs	7198970	**c**	0.771

As seen, the concordance percentage of 76.9% indicates that the model correctly distinguishes between diabetic and nondiabetic individuals in 76.9% of the observed pairs. Similarly, a c-statistic of 0.771 demonstrates moderately strong ability to correctly classify diabetes status.

Analysis of Maximum Likelihood Estimates				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
NOTE: The degrees of freedom for the t tests is 25.				
Intercept	-7.3638	0.3776	-19.50	<.0001
Age	0.0587	0.00285	20.59	<.0001
BMI	0.0801	0.00592	13.53	<.0001
Race_Ethnicity	0.1122	0.0359	3.12	0.0045
Education_Level	-0.3178	0.0625	-5.08	<.0001
Hypertension	0.1912	0.0764	2.50	0.0192

After consideration, race and ethnicity ended up being dropped as predictors in order to reduce the complexity of the model.

Appendix C – Code

The code used for the logistic regression model creation is provided below. The full code for the project, including data cleaning, visualization, and more, can be found on the GitHub repo: https://github.com/calebfrankenberger/nhanes-risk-factor-analysis

```
/*===============================================================================================
===
  Project   : NHANES Diabetes Risk Factor Analysis
  Purpose   : Data analysis
  Programmer: Caleb Frankenberger
  Date      : 11/11/2024
=================================================================================================
=*/


/*-----------------------------------------------------------------------------------------------
---
  Import the data that was cleaned in Data_Cleaning.sas
-------------------------------------------------------------------------------------------------
-*/

libname cln_dta
"/home/u64010893/sasuser.v94/Projects/NHANES_Diabetes_Risk_Analysis/Data/Processed";

data analysis_data;
    set cln_dta.clean_data;
run;


/*-----------------------------------------------------------------------------------------------
---
  Fit the logistic regression model (initially using all predictors)
-------------------------------------------------------------------------------------------------
-*/
proc surveylogistic data=cln_dta.clean_data;
    weight WTMECPRP;
    strata SDMVSTRA;
    cluster SDMVPSU;

    model Diabetes_Status(event='1') = Age BMI Education_Level Hypertension Race_Ethnicity;
    output out=diag predicted=pred;
run;


/*-----------------------------------------------------------------------------------------------
---
  Fit the logistic regression model (dropping race and education level)
-------------------------------------------------------------------------------------------------
-*/
proc surveylogistic data=cln_dta.clean_data;
    weight WTMECPRP;
    strata SDMVSTRA;
    cluster SDMVPSU;

    model Diabetes_Status(event='1') = Age BMI Hypertension;
    output out=diag predicted=pred;
run;
```