Predicting Contributing Factors to Diabetes.

Sejin Kim

Kenyon College

Author Note

Sejin Kim, Kenyon College.

Contact: kim3@kenyon.edu

Keywords: diabetes, nhanes, regression

Predicting Contributing Factors to Diabetes.

I. Introduction

Diabetes has long plagued the world, but within the past 60 years, the number of people affected by this disease has increased dramatically. The CDC estimated that in 2015, just under 8% of the United States population was diagnosed with diabetes. According to a brief time series analysis of the historical data that the CDC publishes from 1980 to 2017, the rate of diabetes is continuing to grow. An ACF plot found that at four lags in, the present values for the total percentage of the population with diabetes is well related with the past total percentage reports. If I were to continue on this path, I could see changes within the healthcare and pharmaceutical industries that may not be for the best.

To further explore this, I am using a dataset from the National Health and Nutrition Examination Survey. The "survey" is actually a group of surveys, each designed to "assess the health..of adults and childre in the United States." While some surveys are very much qualitative information, the data that I will be analyzing is purely quantitative.

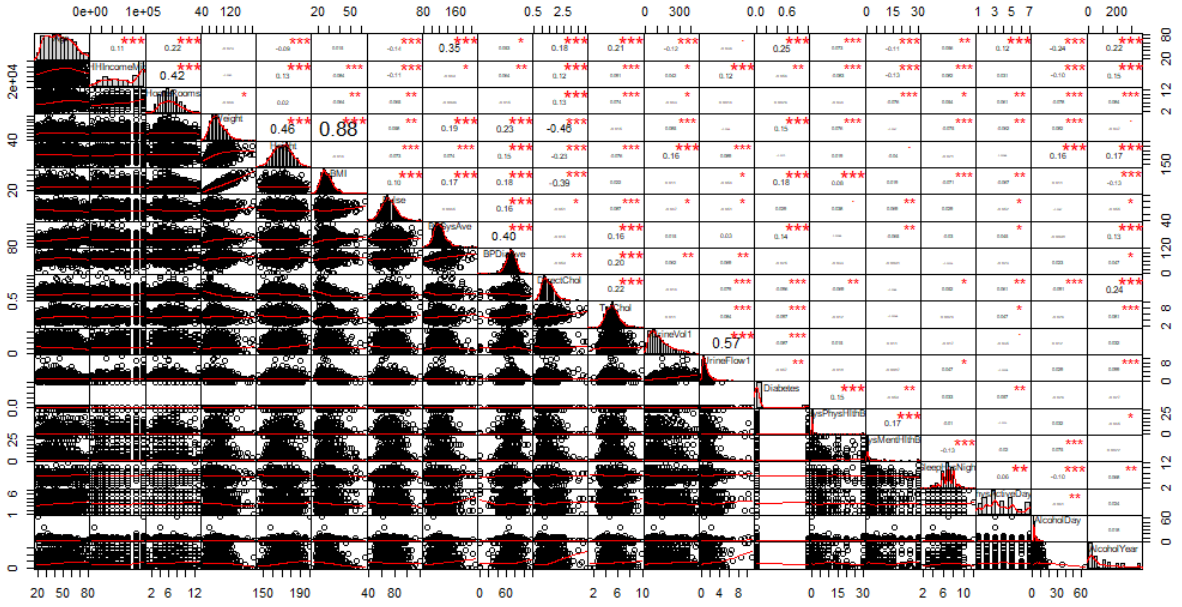
The NHANES program began in the 1960's, and is a major program section at the Centers for Disease Control and Prevention. It is used by both government and institutional analysts to find patterns in the approaches to and general sentiment towards healthcare and nutrition domestically. Because of its comprehensive nature, it is widely applicable for modeling and predicting patterns in many diifferent fields. The data has been used to create growth charts for children, develop the policy to remove lead from common household items, influence immunization schedules, and track diabetes. All of the data is collected and anonymized to protect the healthcare rights afforded by HIPPA.

While some of this data can be used in isolation, say to determine the rates of change in diabetes diagnoses, more comprehensive data can be used to holistically determine if there are leading causes in diabetes, including changes which may not have been foreseen at the beginning of the data collection, but which is due to more than just chance.

The NHANES dataset were retrieved in a cleaned form from the Kenyon College department of statistics. The dataset originally included about 10,000 observations across 70 different variables. The applicable dataset was cleaned to only include 2167 observations across 28 different variables. The data includes observations gathered from 2009 to 2012's biannual surveys, and the dataset was filled on a rolling schedule.

III. Methods

Analysis began by exploring the data graphically. To determine if there were any variables that were heavily correlated with each other, a numerical correlation matrix and the accompanying p-value matrix were created. In analyzing this correlation matrix, I tagged correlations which had a correlation $cor > 0.25$ and $p > 0.05 = \alpha$ to find if any of these variables were heavily correlated with each other, and whether those correlations were due to more than just chance. Given a dataset with this many variables, some overlap and high correlations between variables is to be expected.

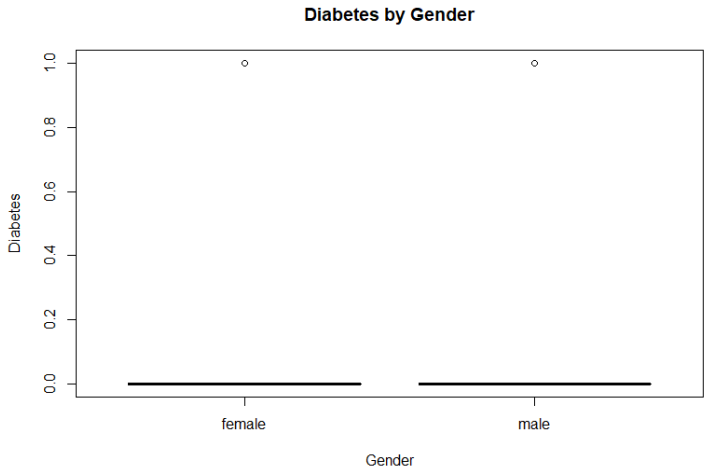


It should be noted that in the histograms, shown on the diagonal where correlation would just equal one, some of these histograms are far from normal. However, because we are running a logistic regression model, not a linear regression model, the normality condition

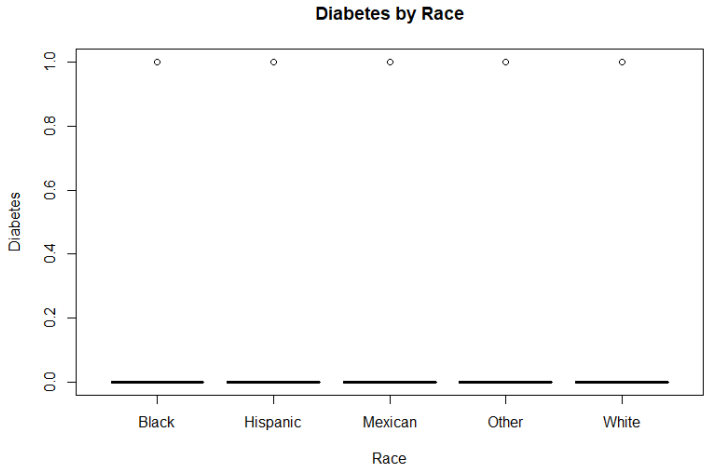
does not matter. Additionally, there is no reasonable way to have the normality condition be satisfied for a binary variable.

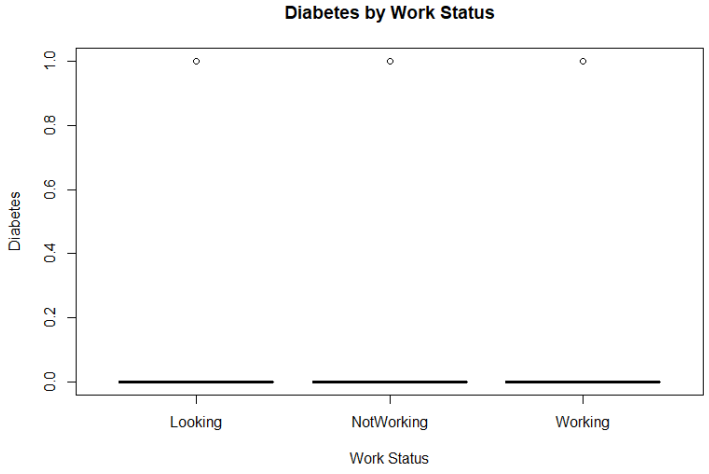
The correlations that were found did make sense. The variable *HomeRooms* was correlated to *HHIncomeMid*, which makes sense because as income increases, the number of rooms in your home would also increase. *Height*, *Weight*, and *BMI* are all highly correlated, which makes sense, since BMI is derived from both height and weight, by the CDC's definition[1]. As expected, average diastolic blood pressure (*BPDiaAve*) and average systolic blood pressure (*BPSysAve*) are also related. Of note is that only the average systolic blood pressure is somewhat highly correlated with age, not average diastolic blood pressure. Finally, the first urine flow rate measurement (*UrineFlow1*) is somewhat highly correlated with the volume of urine (*UrineVol1*). There were no other variables which had a high correlation that were also significant.

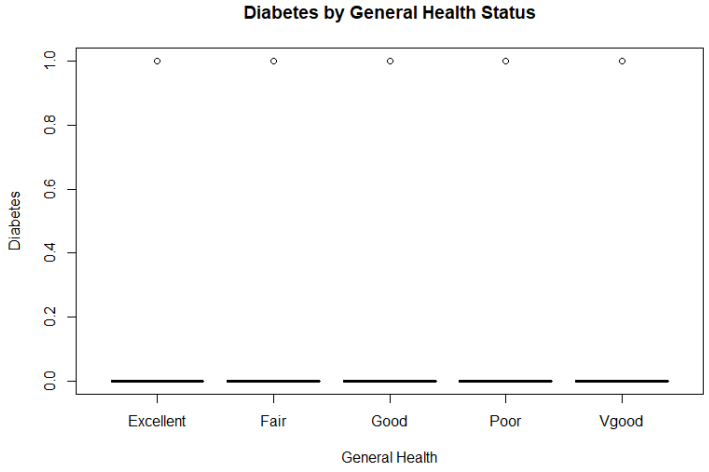
Plots were made for most of the variables, against the binary variable *Diabetes*. The expected behavior here was that for any given boxplot, the majority of every group would be within the box, which should have been clustered around 0. Some groups may have had an outlier of diabetics, if the variable were to be significant or heavily correlated, which would show up in any boxplot as a group of outlier points where $Diabetes = 1$. Several of these plots are given as follows.



[1]Of note is that the CDC does NOT take into account the gender, but different BMI models may take this into account. For the purposes of this study, gender is not used.







As shown, the behavior of each of these plots is to be expected. The majority of the surveyed individuals do not have diabetes. This is backed up with summary statistics, which report that only 145 out of the 2167 observations report having diabetes.

The multiple logistic regression models were determined using a stepwise procedure using the `step` function in R. This function selects models to minimize the AIC value, rather than by only using the smallest p-value. Like any model building procedure, including those for linear regression model building, one should not blindly accept the results based on the output from the computer, but rather should double check the model or build a second model from available variabls that are biologically or scientifically sensible. That is, it would not make sense for the observation number to be a significant variable. If it were to show up significantly (it does not, but for the sake of this example, let us continue to use it), I should strongly consider throwing it out, since it is an artificial variable that only affects the regression by chance.

In this analysis, I will be using multiple correlation to investigate the relationship between potential independent variables. For example, if two independent variables have some uncanny relationship to one another, I would likely drop at least one of them in the final model, which ever one is less significant. There may be reasons why I would choose one variable over another one for the final model, though. In this dataset, let us examine the *BMI* variable. According to the CDC, BMI is derived from height and from weight, both of which are supposedly independent variables. As a result, in our final analysis, I probably would *not* want to use all three of these variables, but rather which ever variable accounts for the most variability.

After creating the initial model using stepwise logistic regression, the model was examined more closely to check for abnormal behavior, especially in the z-values and the p-values. In the initial model generated by the stepwise regression method, the average systolic blood pressure showed itself to not be a significant and meaningful variable in the regression. As a result, I chose to remove it from the regression. To ensure that the change to the model was truly appropriate, I also examined the Wald statistics using the type II tests in ANOVA, and found that the p-value, the probability that the average systolic blood pressure was in the model due to more than just chance, was above my model's cutoff value of $\alpha = 0.1$.

Through this analysis, the final model is as follows:

$$\begin{aligned} log(Odds) = \beta_0 + \beta_1 \times (Age) + \\ \beta_2 \times (BMI) + \\ \beta_3 \times (TotChol) + \\ \beta_4 \times (DaysPhysHlthBad) + \\ \beta_5 \times (HealthGen(Fair)) + \\ \beta_6 \times (HealthGen(Good)) + \\ \beta_7 \times UrineVol1 + \\ \beta_8 \times (DirectChol) + \\ \beta_9 \times (DaysMentHlthBad) + \\ \beta_{10} \times (Work(NotWorking)) + \epsilon \overset{iid}{\sim} N(0, \sigma) \end{aligned} \tag{1}$$

When I fit this model to the data, I got the following logistic regression line for the data:

$$\begin{aligned} log(\hat{O}dds) = -5.350001 + 0.077244 \times (Age) + \\ 0.078084 \times (BMI) - \\ 0.442099 \times (TotChol) + \\ 0.045104 \times (DaysPhysHlthBad) + \\ 1.236488 \times (HealthGen(Fair)) + \\ 0.658304 \times (HealthGen(Good)) - \\ 0.003053 \times UrineVol1 - \\ 0.611166 \times (DirectChol) - \\ 0.034084 \times (DaysMentHlthBad) - \\ 1.040257 \times (Work(NotWorking)) + \epsilon \overset{iid}{\sim} N(0, \sigma) \end{aligned} \tag{2}$$

It is worth noting that there are quite a few variables that have been used here. However,

I believe that this *is* an appropriate number of variables, given the complexity of the dataset and the number of observations that I am using. The primary concern when trying to cut back on the number of variables that I am using is to avoid over-fitting the data to the point where the analysis is no longer adaptable. However, given the number of observations, I believe this model to be an appropriate model. I confirmed this by using using a training and holdout model validation strategy. In this, I got a correlation between the holdout and the predictor of 0.3367. While this is not perfect, I believe it to be a perfectly reasonable result.

To continue my analysis, I also generated an overall p-value for the model. To do this, I defined the null model and compared it to the full model using an analysis of variance, or ANOVA. I used the χ^2 distribution to run this test. In this test, I found that with a residual deviance of 1111.18 on 2166 degrees of freedom and a p-value of ≈ 0 (reported as 2.2×10^{-16}, I rejected the null hypothesis that the final model generated is significantly different from the null model. In the same vein as the ANOVA, I also ran a likelyhood ratio test of nested models. In this analysis, I found that with a log-likelyhood of -555.59 on one degree of freedom, and a p-value of ≈ 0 (reported as 2.2×10^{-16}, I rejected the null hypothesis that the models are similar. The primary difference between this test and the ANOVA from above is that this test is an asymptotic likelyhood test, and uses the Wald test as its base.

All of the analyses were carried out using R version 3.6.1.

III. Results

In this project, I aimed to find which variables could most effectively predict if a person had diabetes by using the NHANES dataset by running a multiple logistic stepwise regression analysis. Ultimately, the variables that pertained to weight and overall wellbeing were the best predictors of diabetes. Because of the nature of the analysis, no transformations were used to reshape any of the variables. Additionally, because of the ultimate model that was chosen, no interaction terms were employed. Only one of the variables in any possible

interaction would have been used, since I am looking at a multiple logistic regression, as opposed to a linear regression.

The continued significance of the average systolic blood pressure, but notably not the average diastolic blood pressure, stuck out to me as unusual. Given that there was such a correlation between average systolic blood pressure and average diastolic blood pressure, I would have expected that if one had shown up in the final model, then both would have shown similar behavior. Although one of them was not outed by the stepwise regression process, only one ever showed up, not both.

Determining what R^2 is for a logistic regression model is inherently harder than for, say, a linear regression model. Researchers cannot agree on what the calculation for R^2 should be, since they don't even know what it is defined by: explained variation or goodness of fit. As such, it is important to report all three of the pseudo-R^2 values. Ultimately, the model that I generated and chose yielded a $R^2_{McFadden}$[2] value of 0.2677, a $R^2_{CoxSnell}$[3] value of 0.1283, and a $R^2_{Nagelkerge}$[4] value of 0.3198.

IV. Further Discussion

It goes without saying that this is not the only definitive way to analyze the NHANES dataset. Using *Diabetes* as a predictor could prove useful in reverse-engineering the usefulness of diabetes to determine some other response variable, especially in a linear regression model, such as weight, even though weight was not nearly as significant here.

It is worth noting that the full dataset was not analyzed. Only 28 variables, believed before analysis to be relevant, were used. In a later and deeper study, the full dataset should be reviewed and analyzed, to see if there are any other hiding variables that are significant for predicting diabetes diagnoses.

[2]McFadden's pseudo-R squared value uses the method of maximum likelyhood.

[3]Cox and Snell's pseudo-R squared value uses the ratio of the likelyhoods, which reflects the improvement of the full model over the intercept model, meaning that the smaller the ratio, the greater the improvement over the null model.

[4]Nagelkerke/Gragg and Uhler's pseudo-R squared value divides the Cox and Snell model by its maximum possible value, and tells the ability of the full model to perfectly predict the outcome.

It is also worth noting that the time-range of this study is relatively limited. The NHANES data collection began in the 1960's, and yet, I have only analyzed four years from the late 2000's to the early 2010's. It would be of interest to see how the predictors have changed over time, perhaps using some sort of time-series, even with the logarithmic coefficients presented in the logistic model, to see how the odds have changed.

As part of exploring the world of the statistics datasets published by the Centers for Disease Control and Prevention, I was able to analyze the historical data for diabetes rates from 1980 to 2017. This dataset is also given in the GitHub repository, cleaned and refactored. In it, there are some variables, including those for total percentage, a low (best-case) percentage scenario, and a high (worst-case) percentage scenario. Given that this was historical data collected regularly from year to year, I elected to run a time-series analysis on this dataset.

As stated in Part I, the total percentage of the population with diabetes has gone up to just under 10% by 2017. It is worth noting that this dataset does not and cannot differentiate between Type I and Type II diabetes, so nothing can definitely be said about linking raising obesity rates to diabetes. Curiously, the detection rate for diabetes does not seem to have changed within the past forty years. When plotting the best-case scenario against the worst-case scenario, the range between these points seems to not have changed in any meaningful way. Further examining these time series is of interest, and I believe that this would be a valuable and interesting project to embark on down the line.

V. Conclusion

Diabetes is not a problem that will disappear any time soon. Rates of diabetes diagnoses are rising across the country. However, by using a stepwise regression model to analyze the NHANES dataset, I was able to find variables that, when known, increase the odds of diabetes, for better or for worse. I suppose it would be a bad day to find you have diabetes. Furthermore, I was able to show that by analyzing these variables, one could account, approximately, for between 12.8% and 32.0% of the United States population's

odds of having diabetes. While these odds do not sound stellar, a further analysis could pave the way for increasing the odds and utlimately decreasing the amount of diabetes by controlling variables and encouraging healthier habits.

Works Cited

Felipe de Mendiburu (2019). agricolae: Statistical Procedures for Agricultural Research. R package version 1.3-1. https://CRAN.R-project.org/package=agricolae

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2019). Hmisc: Harrell Miscellaneous. R package version 4.2-0. https://CRAN.R-project.org/package=Hmisc

John Fox and Sanford Weisberg (2019). An R Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Scott J. Long. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks: Sage Publications, 1997.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R. Pruim, D. T. Kaplan and N. J. Horton. The mosaic Package: Helping Students to 'Think with Data' Using R (2017). The R Journal, 9(1):77-102.

Richard M. Heiberger (2019). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package version 3.1-37. URL https://CRAN.R-project.org/package=HH

Stephane Champely (2018). pwr: Basic Functions for Power Analysis. R package version

1.2-2. https://CRAN.R-project.org/package=pwr

Thomas Lumley based on Fortran code by Alan Miller (2017). leaps: Regression Subset Selection. R package version 3.0. https://CRAN.R-project.org/package=leaps

Torsten Hothorn, Frank Bretz and Peter Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.

U.S. Department of Health and Human Services. (2017). National Health and Nutrition Examination Survey. National Center for Health Statistics, Centers for Disease Control and Prevention.

U.S. Department of Health and Human Services. (2017). Diagnosed Diabetes. Division of Diabetes Translation, United States Diabetes Surveillance System, Centers for Disease Control and Prevention.