"625 Final Project Paper"

Yeseul Ha, Joy Zhang, Jingyu Zhao

December 2023

1 Introduction

Diabetes is a chronic disease that impacts 38 million adults in the United States, and it is a condition where the blood glucose levels are too high [3]. Diabetes is the eighth leading cause of death in the US and is categorized as either Type 1 or Type 2 diabetes. 90-95% of diabetes are Type 2, where the body is not capable of using insulin in the blood [2]. On the other hand, Type 1 diabetes consists of 5-10% of diabetes cases, and in this condition the body cannot produce its own insulin to use the blood glucose for energy [2]. Research has shown that high blood glucose is linked to other severe health problems, such as heart disease, nerve damage, and kidney disease [2]. The risk of diabetes has also been shown to drastically increase with age. According to the CDC, Type 2 diabetes develops most often in people over the age of 45 [3]. Some studies have even shown that after accounting for other health factors such as BMI, the ability for insulin to adjust blood glucose level continues to decreases with age, contributing to an increased risk of developing diabetes [6, 1]. In recent years, more young adults and adolescents are more developing Type II diabetes. In our research specifically, we would like to further examine the relationship between age and diabetes.

To look at age as a predictor of diabetes, we decided to construct two models: the first being an unadjusted model with just age as a predictor, and the second being an adjusted model, adjusting for almost all other covariates in our data set.

2 Method

To begin examining our research question, we used a balanced dataset that was compiled and cleaned from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) [8]. The BRFSS is an annual health-related telephone survey conducted by the CDC that gathers information on participants' health-related risk behaviors, chronic health conditions, and their engagement with preventative services. Our data contains 70,692 observations, with several health and background variables. Our diabetes outcome was binary with a value of 1 for individuals who are either diabetic or pre-diabetic, and 0 otherwise.

2.1 Data Cleaning

Before proceeding with our model construction, we chose to re-categorize certain categorical covariates for ease of interpretation. Age was originally a categorical covariate with 13 levels, however we re-coded the variable into 4 categories: Under 35, 35-49, 50-64, and over 64. Income was re-categorized into low income (less than $25,000 a year), medium income ($25,000 to $75,000 a year), and high income (over $75,000 a year). Education was transformed into a binary indicator with a 1 representing having a college education and a 0 representing no college education. Finally, physical and mental health covariates were also transformed into binary indicators, with a 1 representing 5 or more days of poor health, and 0 representing under 5 days. We chose 5 as a cutoff because it was right around the mean of our dataset.

2.2 Model Selection

Due to consistent research on the limitations of BMI as an effective health and obesity indicator, we decided to exclude BMI variable from our model [4, 5, 7]. This decision was based on the fact that BMI does not directly measure the body fat, and many factors including muscle and bone mass can influence the index. We also chose to

exclude our General Health covariate, since our data set already included Physical Health and Mental Health and the General Health would be redundant and highly correlated to the previous two.

A generalized linear model (GLM) with a logit link function was fitted in R using the `glm` package since our `Diabetes_binary` response, Y_i, takes either 1 or 0 at the individual level. Observation pairs (x_i, Y_i) are independent. The GLM link function takes the following assumptions: for the systemic component, $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta$ and for the random component, $Y_i \sim \text{Binomial}(n_i, \pi_i)$. The mean function was $\mu_i = \pi_i$ and the variance function was $v(\pi_i) = \pi_i(1-\pi_i)$. The first glm model we fit with just the age variable was

$$logit(\pi_i) = \beta_0 + \beta_1 \cdot Age_{\{35-49\}} + \beta_2 \cdot Age_{\{50-65\}} + \beta_3 \cdot Age_{\{over65\}}$$

with the reference age category as the age group of under 35 years old. Our second glm model with the rest of the selected covariates was

$$\begin{aligned} logit(\pi_i) =& \beta_0 + \beta_1 \cdot Age_{\{35-49\}} + \beta_2 \cdot Age_{\{50-65\}} + \beta_3 \cdot Age_{\{over65\}} + \beta_4 \cdot HighBP_1 + \beta_5 \cdot HighChol_1 \\ &+\beta_6 \cdot CholCheck + \beta_7 \cdot Smoker_1 + \beta_8 \cdot Stroke_1 + \beta_9 \cdot HeartDiseaseorAttack_1 + \beta_{10} \cdot PhysActivity_1 \\ &+\beta_{11} \cdot Fruits_1 + \beta_{12} \cdot Veggies_1 + \beta_{13} \cdot HvyAlcoholConsump_1 + \beta_{14} \cdot AnyHealthcare_1 \\ &+\beta_{15} \cdot NoDocbcCost_1 + \beta_{16} \cdot High\_Physical\_Health + \beta_{17} \cdot High\_Mental\_Health \\ &+\beta_{18} \cdot DiffWalk_1 + \beta_{19} \cdot Male + \beta_{20} \cdot Medium\_Income + \beta_{21} \cdot High\_Income + \beta_{22} \cdot Edu\_College \end{aligned} \tag{1}$$

The reference category for income was the low income group, and the reference category for education was chosen to be the no college education group.

3 Results

3.1 Descriptive Statistics

Before examining our model results, we first created a couple of plots to look at the distribution of age among our covariates in our data set. We started by looking at our diabetes response variable categorized by age (Figure 1), and saw that individuals with diabetes tend to be older than individuals without diabetes.

We also looked at the distribution of age on its own and found that age appears to be left skewed, with more older individuals than younger.

Finally, we chose to look at our predictor covariates categorized by age. While certain covariates such as smoking status and amount of fruits and veggies consumed did not have large differences in age breakdown, we found that most of our health indicators such as high blood pressure, high cholesterol, individuals who had experienced a stroke, and heart disease did have large discrepancies in age categories. All four plots (Figure 2) suggested that groups of individuals suffering from the previously mentioned health disparities had higher percentages of older people compared to those who did not suffer from the health disparities.

This difference in age breakdown further interests us in our research question to see whether or not age is a significant predictor for diabetes. Other key factors that we noted were that less than 10% of our individuals had recorded a stroke in the past, the same being true for individuals with heavy alcohol consumption, not being able to see a doctor due to cost, and poor mental health. In addition, over 65% of total participants hold college degress and beyond and over half of the particpants consume fruits and vegetables at least once a day.

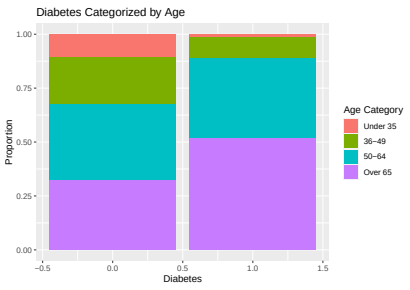

Figure 1: Diabetes categorized by age.

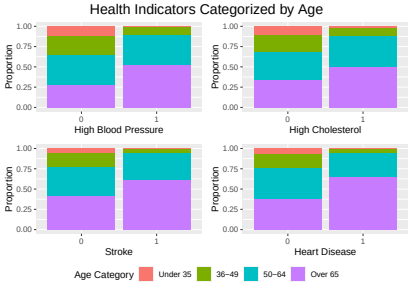

Figure 2: Health indicators by age.

3.2 Model Inference

Because our two models were nested, we were able to perform a likelihood ratio test to determine which one was a better fit. We found that our second model adjusting for all other covariates was superior to the one with just age as a predictor, and proceeded to interpret our results from the adjusted model summary.

Table 1: Regression Coefficients

	Estimate	SE	z value	Pr(>|z|)	
(Intercept)	-3.3036	0.1012	-32.63	0.0000	***
age_cat1	0.8376	0.0539	15.53	0.0000	***
age_cat2	1.2571	0.0513	24.49	0.0000	***
age_cat3	1.4118	0.0518	27.23	0.0000	***
HighBP1	1.0209	0.0185	55.11	0.0000	***
HighChol1	0.6463	0.0180	35.85	0.0000	***
CholCheck1	1.4266	0.0780	18.30	0.0000	***
Smoker1	-0.0064	0.0180	-0.36	0.7210	
Stroke1	0.1817	0.0398	4.57	0.0000	***
HeartDiseaseorAttack1	0.4186	0.0274	15.26	0.0000	***
PhysActivity1	-0.2212	0.0202	-10.93	0.0000	***
Fruits1	-0.0941	0.0187	-5.03	0.0000	***
Veggies1	-0.0848	0.0224	-3.78	0.0002	***
HvyAlcoholConsump1	-0.8978	0.0470	-19.11	0.0000	***
AnyHealthcare1	0.0410	0.0452	0.91	0.3644	
NoDocbcCost1	0.0861	0.0329	2.62	0.0089	**
DiffWalk1	0.5330	0.0240	22.20	0.0000	***
Sex1	0.2918	0.0183	15.95	0.0000	***
income_cat1	-0.2218	0.0246	-9.03	0.0000	***
income_cat2	-0.5033	0.0244	-20.66	0.0000	***
edu_cat1	-0.1218	0.0198	-6.15	0.0000	***
PhysHlth_cat1	0.2931	0.0227	12.91	0.0000	***
MentHlth_cat1	0.0178	0.0244	0.73	0.4648	

Table 1 contains the results of our model analysis, including the estimates of the regression coefficients, their standard errors, and associated p-values. Looking at the values, we found that all age categories were significant in their relation to our Diabetes outcome. Specifically, compared to individuals under the age of 35, adults aged 35-49 are $e^{0.8376} = 2.318$ times more likely ($p < 0.05$) to have diabetes compared to individuals ages 35 and under, adjusting for all covariates. Adults aged 50-64 are $e^{1.2571} = 3.515$ times more likely ($p < 0.05$) to have diabetes compared to individuals ages 35 and under, and finally adults aged 65 and older are $e^{1.4118} = 4.099$ times more likely ($p < 0.05$) to have diabetes compared to individuals ages 35 and under, again adjusting for all covariates. We use e^{β_i} because our coefficients represent the log odds values. These results support previous research that age is related to diabetes, and further confirms that older individuals are more likely to be diabetic.

3.3 Model Diagnostics

After completing variable selection, we conducted a pair-wise assessment of multicollinearity among the included covariates using Fisher's exact tests. Given the reference coding of the model, this made VIF statistics applicable. The goodness of fit for the full model was evaluated using the Pearson chi-square test instead of the Hosmer-Lemeshow test, providing simplicity and avoiding additional parameter specifications.

To address overdispersion concerns, we explored alternatives beyond the typical residuals vs. fitted values trend observation, such as (1) dividing the residual deviance by its degrees of freedom, and (2) estimating the scale parameter $a(\phi)$ using the Pearson chi-square statistic. Qualitative assessment of these values' proximity to 1 was performed. Additionally, Cook's distances and leverages were computed using the 'glm.diag' function in R to identify high influence and leverage points.

The Pearson chi-square statistic, measuring the goodness of fit for the interaction model, yielded a large value, indicating a significant difference between observed and expected frequencies. With 35,327 degrees of freedom, this suggests a substantial number of observations in the contingency table. The p-value (0.491785) exceeds the significance level (0.05), indicating that our model was a good fit. Fisher's exact test revealed significance among all relevant covariates ($p < 0.001$). While the plot of residuals vs. fitted values (Figure 3) did not show a random pattern, further investigation into overdispersion using the scale parameter estimate indicated a value close to 1 (1.10589), suggesting no overdispersion in our full model. Furthermore, influential and high-leverage

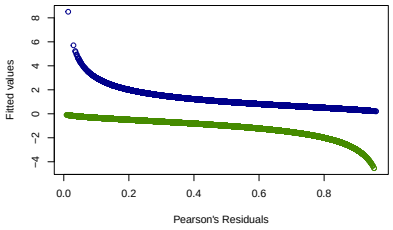

Figure 3: Pearson's residuals vs. Fitted values

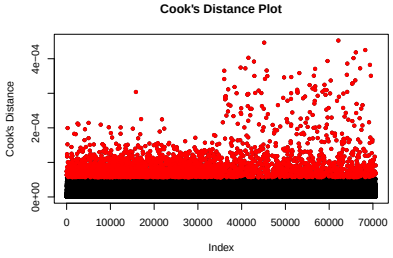

Figure 4: Cook's distance plot

points were detected (Figure 4 and Figure 5, respectively). Individual points were scrutinized, revealing no apparent issues in the data. However, additional insights from study conductors are essential to assess the potential impact on conclusions and whether any points should be removed.

The non-random scatter observed in the Pearson residual vs. fitted values, attributed to the binary nature of the data, may warrant further investigation to ensure its alignment with the study's objectives

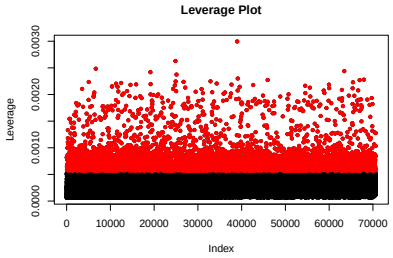

Figure 5: Leverage plot

3.4 RShiny Application

A secondary goal of our project was to create an application for not just our use, but for any researcher interested in the BRFSS data set. While we were specifically interested in diabetes as an outcome variable, the data set contains many other health indicators, such as high blood pressure, high cholesterol, heart disease, etc. In our RShiny application, users have the ability to select their specific covariate of interest, and build and analyze a model using the same methods we used.

4 Discussion

As previously mentioned, our results are consistent with previous research on the effect of age on diabetes [1, 2, 7, 9, 11]. Aging has shown to affect glucose sensitivity and impair insulin secretion by the pancreatic cells, increasing the risk for diabetes. Additionally, our study showed that higher prevalence of diabetes or pre-diabetes were shown in males. However, limitations of our study were that all data was extracted from telephone surveys, which could potentially exhibit self-report bias. Our research may also fail to fully capture non-modifiable diabetes risk factors such as genetic factors, gestational diabetes, and family history.

5 Contributions

Yeseul conducted the literature review and helped write the background, model selection, and conclusion for our paper. Joy developed the RShiny app, organized the github, and helped write the descriptive statistics and made plots for the paper. Jingyu coded the data filtering, performed the covariate association visualization, and helped with model building and writing the model diagnostics portion of the paper. All three helped with editing the paper, and writing the README for the github.

6 References

[1] N. Bahour et al. "Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers". In: Geroscience 44.1 (Feb. 2022), pp. 415–427.

[2] Boon How Chew et al. "Age 60years was an independent risk factor for diabetes-related complications despite good control of cardiovascular risk factors in patients with type 2 diabetes mellitus". In: Experimental Gerontology 48.5 (May 2013), pp. 485–491. issn: 0531-5565. doi: 10.1016/j.exger.2013.02.017. url: http://dx.doi.org/10.1016/j.exger.2013.02.017.

[3] National Institute of Diabetes, Digestive, and Kidney Diseases. Diabetes Statistics. Feb.2023. url: https://www.niddk.nih.gov/health- information/health- statistics/diabetes-statistics.

[4] Centers for Disease Control and Prevention. What is Diabetes. Sept. 5, 2023. url: https://www.cdc.gov/diabetes/basics/diabetes.html.

[5] Dong Hoon Lee et al. "Comparison of the association of predicted fat mass, body mass index,and other obesity indicators with type 2 diabetes risk: two large prospective studies in US men and women". In: European Journal

of Epidemiology 33.11 (Aug. 2018), pp. 1113–1123.issn: 1573-7284. doi: 10.1007/s10654-018-0433-5. url: http://dx.doi.org/10.1007/s10654-018-0433-5.

[6] Shriraam Mahadevan and Iftikhar Ali. "Is body mass index a good indicator of obesity?" In: International Journal of Diabetes in Developing Countries 36.2 (June 2016), pp. 140–142. issn: 1998-3832. doi: 10.1007/s13410-016-0506-5. url: http://dx.doi.org/10.1007/s13410-016-0506-5.

[7] K. Mordarska and M. Godziejewska-Zawada. "Diabetes in the elderly". In: Prz Menopauzalny 16.2 (June 2017), pp. 38–43.

[8] K. M. Narayan et al. "Effect of BMI on lifetime risk for diabetes in the U.S". In: Diabetes Care 30.6 (June 2007), pp. 1562–1566.3

[9] I O Okwechime, S Roberson, and A Odoi. "Prevalence and Predictors of Pre-Diabetes and Diabetes among Adults 18 Years or Older in Florida: A Multinomial Logistic Modeling Approach". In: PLoS One 10.12 (2015), e0145781.

[10] Alex Teboul. Diabetes Health Indicators Dataset. 2021. url: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data.

[11] Z. Yan et al. "The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study". In: Diabetes Metab Syndr Obes 16 (2023), pp. 85–93

7 Appendix

Table 1: Descriptive Statistics by age in 2015 BRFSS survey

	0: Under 35	1: 35-49	2: 50-64	3: 65 or Over
	(N=4424)	(N=10961)	(N=25587)	(N=29720)
Diabetes_binary				
0	3892 (88.0%)	7542 (68.8%)	12503 (48.9%)	11409 (38.4%)
1	532 (12.0%)	3419 (31.2%)	13084 (51.1%)	18311 (61.6%)
HighBP				
0	3801 (85.9%)	7303 (66.6%)	11075 (43.3%)	8681 (29.2%)
1	623 (14.1%)	3658 (33.4%)	14512 (56.7%)	21039 (70.8%)
HighChol				
0	3733 (84.4%)	7060 (64.4%)	11357 (44.4%)	11379 (38.3%)
1	691 (15.6%)	3901 (35.6%)	14230 (55.6%)	18341 (61.7%)
CholCheck				
0	270 (6.1%)	523 (4.8%)	656 (2.6%)	300 (1.0%)
1	4154 (93.9%)	10438 (95.2%)	24931 (97.4%)	29420 (99.0%)
Smoker				
0	3004 (67.9%)	6571 (59.9%)	13314 (52.0%)	14205 (47.8%)
1	1420 (32.1%)	4390 (40.1%)	12273 (48.0%)	15515 (52.2%)
Stroke				
0	4394 (99.3%)	10726 (97.9%)	24117 (94.3%)	27060 (91.0%)
1	30 (0.7%)	235 (2.1%)	1470 (5.7%)	2660 (9.0%)
HeartDiseaseorAttack				
0	4370 (98.8%)	10455 (95.4%)	22411 (87.6%)	23007 (77.4%)
1	54 (1.2%)	506 (4.6%)	3176 (12.4%)	6713 (22.6%)
PhysActivity				
0	727 (16.4%)	2714 (24.8%)	7717 (30.2%)	9835 (33.1%)
1	3697 (83.6%)	8247 (75.2%)	17870 (69.8%)	19885 (66.9%)
Fruits				
0	1771 (40.0%)	4600 (42.0%)	10573 (41.3%)	10499 (35.3%)
1	2653 (60.0%)	6361 (58.0%)	15014 (58.7%)	19221 (64.7%)
Veggies				
0	843 (19.1%)	2144 (19.6%)	5513 (21.5%)	6432 (21.6%)
1	3581 (80.9%)	8817 (80.4%)	20074 (78.5%)	23288 (78.4%)

	0: Under 35	1: 35-49	2: 50-64	3: 65 or Over
HvyAlcoholConsump				
0	4124 (93.2%)	10367 (94.6%)	24430 (95.5%)	28751 (96.7%)
1	300 (6.8%)	594 (5.4%)	1157 (4.5%)	969 (3.3%)
AnyHealthcare				
0	445 (10.1%)	862 (7.9%)	1534 (6.0%)	343 (1.2%)
1	3979 (89.9%)	10099 (92.1%)	24053 (94.0%)	29377 (98.8%)
NoDocbcCost				
0	3806 (86.0%)	9426 (86.0%)	22466 (87.8%)	28355 (95.4%)
1	618 (14.0%)	1535 (14.0%)	3121 (12.2%)	1365 (4.6%)
DiffWalk				
0	4224 (95.5%)	9447 (86.2%)	18839 (73.6%)	20316 (68.4%)
1	200 (4.5%)	1514 (13.8%)	6748 (26.4%)	9404 (31.6%)
Sex				
0	2330 (52.7%)	6082 (55.5%)	13942 (54.5%)	16032 (53.9%)
1	2094 (47.3%)	4879 (44.5%)	11645 (45.5%)	13688 (46.1%)
income_cat				
0	1136 (25.7%)	2487 (22.7%)	7013 (27.4%)	9688 (32.6%)
1	1182 (26.7%)	2051 (18.7%)	5584 (21.8%)	9480 (31.9%)
2	2106 (47.6%)	6423 (58.6%)	12990 (50.8%)	10552 (35.5%)
edu_cat				
0	1182 (26.7%)	3065 (28.0%)	8869 (34.7%)	11526 (38.8%)
1	3242 (73.3%)	7896 (72.0%)	16718 (65.3%)	18194 (61.2%)
PhysHlth_cat				
0	3783 (85.5%)	8429 (76.9%)	17619 (68.9%)	20853 (70.2%)
1	641 (14.5%)	2532 (23.1%)	7968 (31.1%)	8867 (29.8%)
MentHlth_cat				
0	3289 (74.3%)	8177 (74.6%)	19396 (75.8%)	25495 (85.8%)
1	1135 (25.7%)	2784 (25.4%)	6191 (24.2%)	4225 (14.2%)