STAT 306: Final Group Project

Yicheng Huang, Jinyi Lu, Jiaqi Teng, Zetong Wu

I. Introduction

Diabetes is a long-term condition that occurs when blood sugar levels are too high, which can affect health and even cause complications. Therefore, detecting diabetes in the early stage is important for effective treatment and preventing complications. Researchers have found that factors like blood glucose levels, medication use, HbA1c levels, and family history can influence a diabetes diagnosis.

1. Motivation & Research Question

Based on the World Health Organization's 2024 report, diabetes ranks among the top ten leading causes of death worldwide. As Asian students currently living in Canada, we have observed that high sugar consumption is common in Canadian diets. In addition, Asians may have different genetic susceptibility to diabetes, which increases our focus on how lifestyle and genetic factors work together to influence diabetes risk. This observation, along with the rising probability of getting diabetes prompted us to consider a possible connection between dietary habits and the risk of developing diabetes.

In this study, we aim to investigate the following research question:
"What is the effect of Hemoglobin A1c level on the probability of being diagnosed with diabetes?"

By analyzing these variables, we hope to gain a better understanding of the factors that influence diabetes diagnosis and contribute to more informed health guidance for individuals navigating environments with high-sugar diets.

2. Data Description

The dataset we used is called the Diabetes Dataset, which is from Kaggle and can be found here. It was collected from a clinical screening program where adult participants underwent routine diagnostic tests for diabetes. This dataset contains 9,538 medical records focusing on diabetes diagnosis and associated risk factors, and it is the observational. There are 17 features in the dataset, including 16 essential attributes and 1 outcome.

Response Variable:

- Outcome: Diabetes diagnosis result (1 = Diabetes, 0 = No Diabetes).

Explanatory variable:

- Age: Individual's age, ranging from 18 to 90.
- Pregnances: The number of times the patient has been pregnant.
- BMI: Body Mass Index, based on a person's height and weight (kg/m²).
- Glucose: Blood glucose concentration (mg/dL).
- LDL, HDL: Low and High-Density Lipoprotein, respectively (mg/dL).
- Triglycerides: Fat levels in the blood (mg/dL).
- BloodPressure: Systolic blood pressure (mmHg).
- HbA1c: Hemoglobin A1c level (%) is the average blood sugar over months.

- WaistCircumference, HipCircumference: Waist and Hip measurement, respectively (cm).
- WHR: Waist circumference divided by hip circumference.
- FamilyHistory: Whether the individual has a family history of diabetes (1 = Yes, 0 = No).
- DietType: Dietary habit of an individual (0 = Unbalanced, 1 = Balanced, 2 = Vegan).
- Hypertension: Whether the individual has high blood pressure (1 = Yes, 0 = No).
- MedicationUse: Whether the individual is taking medicine (1 = Yes, 0 = No).

3. Data Cleaning

Before analyzing the data, we clean the data, which includes checking the missing data, removing the wrong data, and converting the data. We find that there is no missing value. Then, we view the data summary and notice that the minimum value for LDL and HDL is negative. As the types of lipoproteins that carry cholesterol in the human body (Bhatt, 2020), LDL and HDL must be greater than 0, so we remove all of the incorrect data, where LDL<= 0 and HDL<= 0. Last, we convert all categorical covariates into factors, which will be more convenient in the analysis.

II. Exploratory Analysis

In this part, we use heatmaps, box plots, and contingency tables to explore the relationships between variables.

1. Heatmap - Correlation Between Covariates

We notice that some of the covariates seem to be correlated, so we use heatmaps to examine the relationship between them.

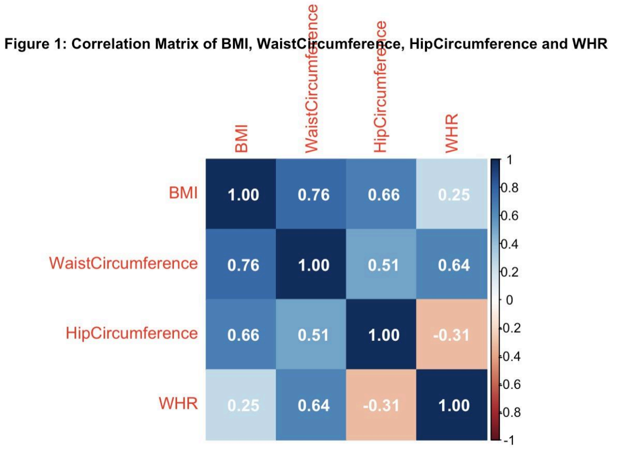


Figure 1: Correlation Matrix of BMI, WaistCircumference, HipCircumference and WHR

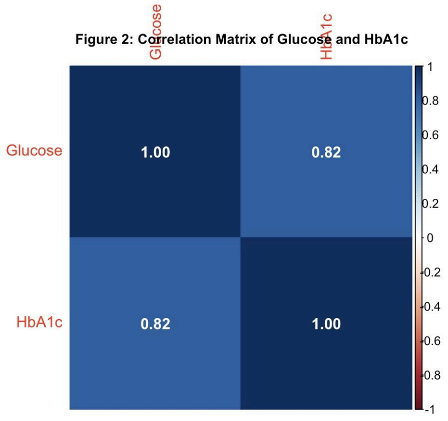


Figure 2: Correlation Matrix of Glucose and HbA1c

In the covariates, BMI, WaistCircumference, HipCircumference and WHR all appear to measure obesity, so we compare them together. In Figure 1, we can see that WaistCircumference and HipCircumference have a high correlation with the other two. Also, WHR is the ratio of WaistCircumference and HipCircumference. BMI and WHR have a relatively low correlation, and they have different ways of measuring health risk (Fogoros, 2025), so we decide to remove WaistCircumference and HipCircumference here.

Additionally, glucose and HbA1c are both indicators of blood sugar level. HbA1c measures the average blood sugar level in the past three months, while the current glucose level may be affected by many factors such as after meals. The figure 2 shows a high correlation between glucose and HbA1c. In order to avoid multicollinearity, we use only one of them. Since HbA1c provides more reliable data, we remove glucose.

2. Boxplots and contingency tables

For continuous variables:

Figure 3: Age by Diabetes Status

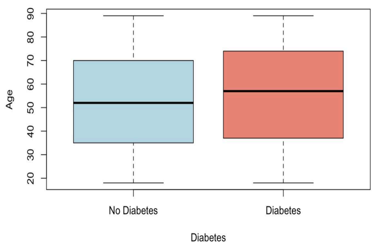


Figure 4: Pregnancies by Diabetes Status

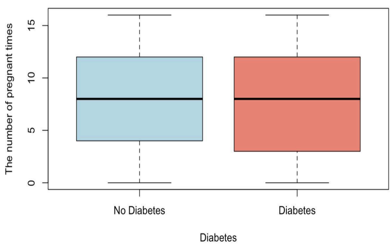


Figure 5: BMI by Diabetes Status

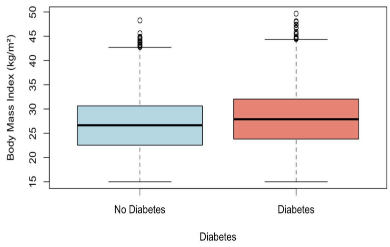


Figure 6: BloodPressure by Diabetes Status

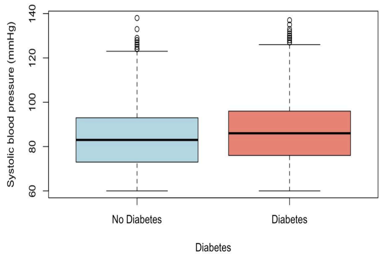


Figure 7: HbA1c Level by Diabetes Status

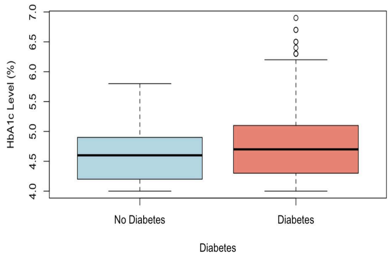


Figure 8: LDL by Diabetes Status

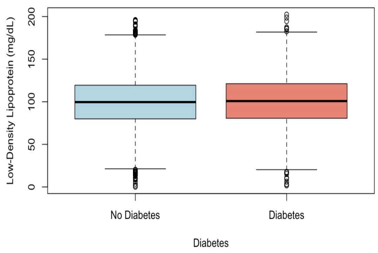


Figure 9: HDL by Diabetes Status

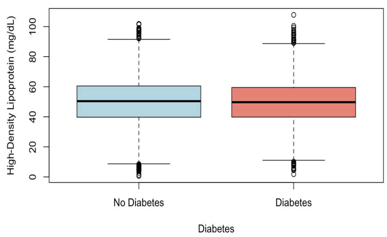


Figure 10: Triglycerides by Diabetes Status

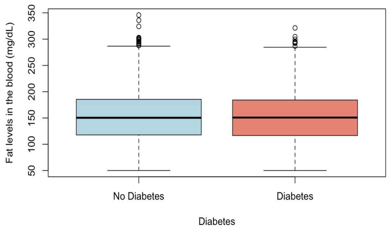
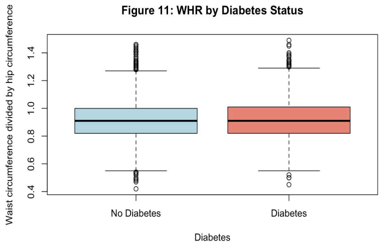
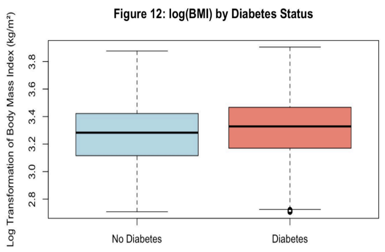


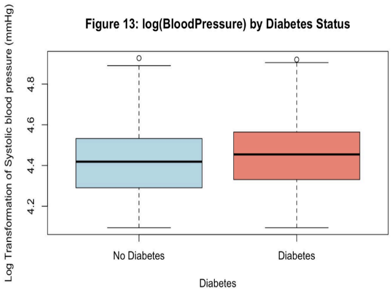
Figure 11: WHR by Diabetes Status

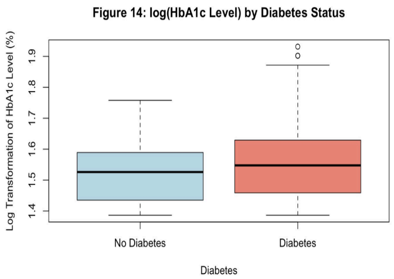


The nine boxplots (Figure 3 - 11) above show how each continuous covariate relates to diabetes status. Most variables, such as Pregnancies, LDL, HDL, triglycerides, and WHR, show a large overlap between the "No Diabetes" and "Diabetes" groups, with similar medians. However, Age, BMI, BloodPressure, and HbA1c level shows some differences in medians between the two groups. Among these, HbA1c shows the most distinct shift, which makes sense as it's a key variable for diagnosing diabetes. Overall, while many features overlap, these four variables may provide a stronger relation with diabetes.

Log Transformation:







After visualizing the boxplots of nine continuous variables, we observe that three out of four covariates we selected contain outliers above the upper bound, suggesting that the data may be right-skewed. Therefore, we try various transformation methods to address this issue, such as log transformation and square root transformation. We find the log transformation produces the most effective results in reducing the influence of outliers and improving the overall distribution of the data, because log transformation compresses large values more aggressively than other methods. Therefore, we apply log transformation to BMI, BloodPressure and HbA1c before further analysis (Figure 12 - 14).

For categorical variables:

```
             No Diabetes History Has Diabetes History
 No Diabetes                6246                    0
 Diabetes                    397                 2883
[1] "Tabel 1: Contingency Table of Outcome and FamilyHistory"
```

```
             Unbalanced Balanced Vegan
 No Diabetes       3761     1887   598
 Diabetes          2026      959   295
[1] "Tabel 2: Contingency Table of Outcome and DietType"
```

```
             No Hypertension Hypertension
 No Diabetes            6244            2
 Diabetes               3272            8
[1] "Tabel 3: Contingency Table of Outcome and Hypertension"
```

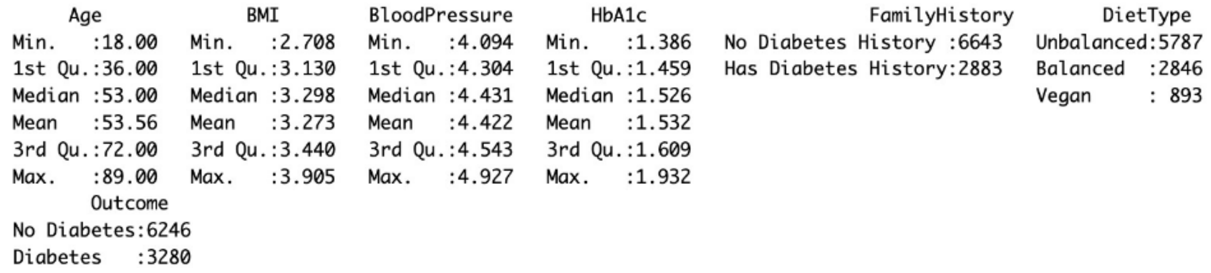
```
             Not Taking Medicine Taking Medicine
 No Diabetes                3864            2382
 Diabetes                   1807            1473
[1] "Tabel 4: Contingency Table of Outcome and MedicationUse"
```

Among the four remaining categorical variables, we find that FamilyHistory, DietType, and MedicationUse show clear distinctions between the diabetes and non-diabetes groups.

For example, in Table 1, almost all patients with a family history of diabetes are in the diabetes group, while those without such history rarely do. Similarly, in Table 2, a higher proportion of patients with an unbalanced diet have diabetes compared to those following balanced or vegan diets.

MedicationUse also differs noticeably between groups, with more individuals in the diabetes group reporting that they are taking medication. However, this variable likely reflects treatment after diagnosis rather than a predictor of diabetes itself, and including it in the model may not be appropriate. Although Hypertension also shows differences, the effect may be less useful for inference because the number of hypertensive individuals is extremely small (only 10 in total), which has a very weak statistical power. Thus, we can remove those two variables.

Summary of Selected Variables



```
      Age             BMI         BloodPressure        HbA1c                       FamilyHistory         DietType   
 Min.   :18.00   Min.   :2.708   Min.   :4.094   Min.   :1.386   No Diabetes History :6643   Unbalanced:5787  
 1st Qu.:36.00   1st Qu.:3.130   1st Qu.:4.304   1st Qu.:1.459   Has Diabetes History:2883   Balanced  :2846  
 Median :53.00   Median :3.298   Median :4.431   Median :1.526                               Vegan     : 893  
 Mean   :53.56   Mean   :3.273   Mean   :4.422   Mean   :1.532                                                
 3rd Qu.:72.00   3rd Qu.:3.440   3rd Qu.:4.543   3rd Qu.:1.609                                                
 Max.   :89.00   Max.   :3.905   Max.   :4.927   Max.   :1.932                                                
         Outcome    
 No Diabetes:6246  
 Diabetes   :3280  
```

In conclusion, we drop Pregnancies, LDL, HDL, Triglycerides, WHR, Hypertension and MedicationUse, and 12 medical records with incorrect values. We select Age, BMI, BloodPressure, HbA1c, FamilyHistory and DieType as our covariates, and have 9526 medical records.

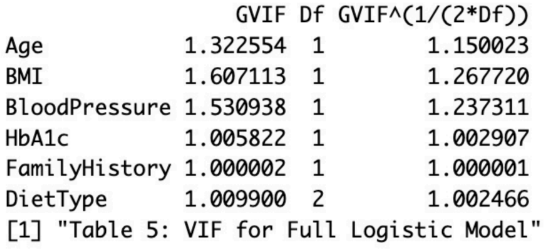
III. Statistical Analysis

We choose a logistic regression model to address our research question, since our response variable diabetes diagnosis result is binary. Thus, logistic regression is suitable for estimating the effect of Hemoglobin A1c level on the probability of being diagnosed with diabetes.

1. Model Diagonstics

Variance Inflation Factor (VIF)

To further investigate multicollinearity, we calculate the Variance Inflation Factors (VIFs) for the full model, which is the additive model containing all six covariates we select.



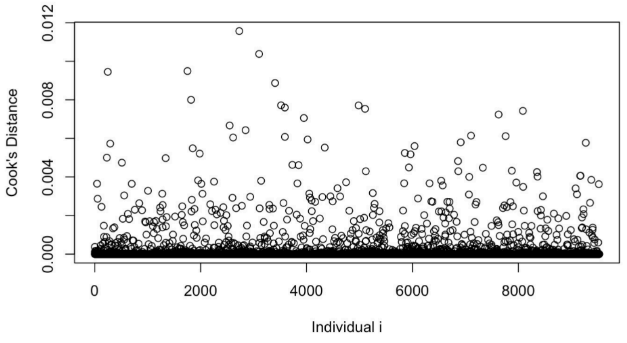
```
                  GVIF Df GVIF^(1/(2*Df))
Age           1.322554  1        1.150023
BMI           1.607113  1        1.267720
BloodPressure 1.530938  1        1.237311
HbA1c         1.005822  1        1.002907
FamilyHistory 1.000002  1        1.000001
DietType      1.009900  2        1.002466
[1] "Table 5: VIF for Full Logistic Model"
```

We find that all values of VIF are smaller than 5 in Table 5, which means multicollinearity is not the issue for the model.

Cook's Distance

We calculate the Cook's distance to measure the influence of a data point.

Figure 15: Cook's Distance for rule-of-thumb (>1)



From Figure 15, we can notice that Cook's distance of most individuals is concentrated between 0.0000 and 0.0004, and all of them are smaller than 0.012. According to the rule-of-thumb, where D_i > 1 is considered influential, we conclude that there is no influential point in the full model.

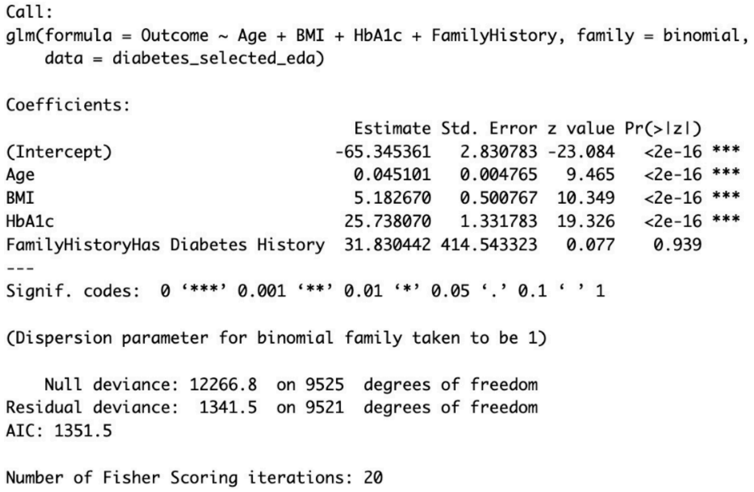
2. Model Selection (Backward)

Model	**AIC**
Age + BMI + BloodPressure + HbA1c + FamilyHistory + DietType	1357.47
Age + BMI + BloodPressure + HbA1c + FamilyHistory	1353.49
Age + BMI + HbA1c + FamilyHistory	1351.54

Table 6: AIC for Different Models

To identify the model with the best performance, we apply backward selection using the Akaike Information Criterion(AIC) as the selection metric. As shown in Table 6, the full model that contains all six covariates has an AIC of 1357.47. Removing DietType slightly reduces the AIC to 1353.49, so the DietType does not contribute substantially to model performance. With further simplification of the model by excluding BloodPressure, the resulting model has the lowest AIC (1351.54). After proper model selection, our final model contains covariates of Age, BMI, HbA1c, and FamilyHistory with no diabetes history as the baseline.

3. Output for Selected Model



```
Call:
glm(formula = Outcome ~ Age + BMI + HbA1c + FamilyHistory, family = binomial,
    data = diabetes_selected_eda)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -65.345361   2.830783 -23.084   <2e-16 ***
Age                              0.045101   0.004765   9.465   <2e-16 ***
BMI                              5.182670   0.500767  10.349   <2e-16 ***
HbA1c                           25.738070   1.331783  19.326   <2e-16 ***
FamilyHistoryHas Diabetes History 31.830442 414.543323   0.077    0.939
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12266.8  on 9525  degrees of freedom
Residual deviance:  1341.5  on 9521  degrees of freedom
AIC: 1351.5

Number of Fisher Scoring iterations: 20
```

Intercept: When all predictors are set to 0, and the individual has no family history of diabetes, the baseline log-odds of having diabetes is -65.345.

Controlling for all other variables, the following variables are statistically significant predictors of diabetes (p < 0.05):

Age: Each additional year in age is associated with a 0.045 increase in the log-odds of having diabetes.

BMI: A one-unit (in kg/m^2) increase in log-transformed BMI is associated with a 5.183 increase in the log-odds of diabetes.

HbA1c: A one percentage increase in average blood sugar over months is associated with a 25.738 increase in the log-odds of having diabetes.

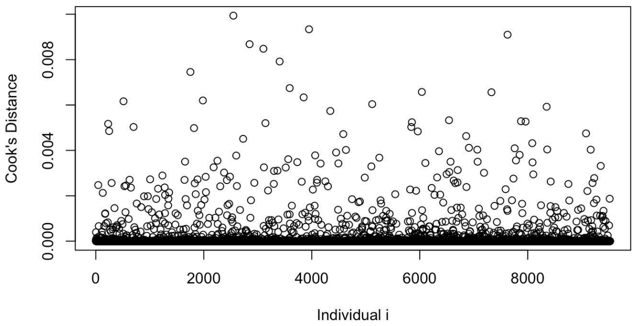
The variables are not statistically significant predictors of diabetes (p > 0.05):

FamilyHistory: The high standard error of 414.543 and the p-value of 0.939, indicating high variability and no statistically significant association between family history and diabetes in this model.

Although it is non-significant, it is meaningful to reduce the AIC, suggesting that it contributes to the fitted model.

Cook's Distance for selected model

Figure 16: Cook's Distance for Backward Selected Model



From Figure 16, we do not find any influential points, since all data points have Cook's distance that is smaller than 1 (for rule-of-thumb).

IV. Conclusion

1. Discussion

This study explores the factors influencing diabetes diagnosis, particularly emphasizing the effect of Hemoglobin A1c (HbA1c) levels. Through rigorous data cleaning, we ensure the reliability of the data. After conducting exploratory analysis and applying backward selection, the optimal logistic regression model contains four covariates, which are Age, BMI, HbA1c and FamilyHistory.

Our exploratory analysis involves heatmaps, boxplots, and contingency tables, which provide essential reasons to remove covariates and apply log transformation. Heatmaps reveal relatively high correlations among some covariates, which helps us to remove covariates avoiding redundancy and multicollinearity in the model. Boxplots indicate distinct median differences of covariates, which highlight the importance of parts of covariates in inference. Contingency tables also demonstrate clear distinctions, prompting usability of the covariates. Log transformations are applied to address data skewness and outliers, enhancing the normality and stability of distributions for covariates. These transformations contribute to improving model performance.

Our refined logistic regression model is chosen by using the backward method based on minimizing AIC. This model with four covariates performs well, with low VIF and low Cook's distance. HbA1c emerges as a particularly strong and highly significant covariate, with a positive relationship indicating that a higher HbA1c level is associated with an increased probability of being diagnosed with diabetes. This supports our research question. BMI also shows a clear positive association with the probability of being diagnosed with diabetes, consistent with existing clinical evidence (Gupta & Bansal, 2020). Age similarly indicates an increasing probability of being diagnosed with diabetes with advancing years. Although FamilyHistory is not significant when controlling for other covariates, it is still included based on model selection criteria, where removing it causes a sharp increase in AIC.

2. Limitation

- **Limited to binary outcomes:** Logistic regression is a model for binary classification, meaning the outcome must be either 0 or 1. This makes it less suitable for modeling continuous outcomes, such as different stages or severity levels of diabetes. In addition, if the outcome variable were more complicated, a different type of model would be more appropriate.

- **Sensitive to outliers and influential points:** Logistic regression is sensitive to extreme or influential observations. These extreme data points can skew the estimated coefficients and make the model less reliable when applied to new data. Although we use Cook's Distance to check for influential points, the results are highly dependent on the chosen threshold (Werth, 2022), which adds uncertainty into the analysis.

- **Covariates:** Many of the covariates in our dataset show little impact between the diabetes and non-diabetes groups, which reduces the model's overall performance. We notice that only Hb1Ac and FamilyHistory have relatively larger coefficients than others.

3. Feature work

If possible, we will incorporate additional features that are more strongly associated with diabetes, such as genetic markers and lifestyle data. We will try to study and discuss the imbalanced coefficients of the fitted model. For future work, it may be worthwhile to investigate other methods like LASSO or Ridge regression in variable selection. This may also help address the issue with the non-significant covariate and improve the statistical inference.

References

Asinow Tech. (2025). *Diabetes Dataset*. Kaggle. https://www.kaggle.com/datasets/asinow/diabetes-dataset

Bhatt, A. (2020, October 27). *Cholesterol: Understanding HDL vs. LDL*. Harvard Health. https://www.health.harvard.edu/blog/understanding-cholesterol-hdl-vs-ldl-2018041213608

Fogoros, R. N. (2025, March 13). *BMI, waist circumference, or waist-to-hip ratio?*. Verywell Health. https://www.verywellhealth.com/bmi-waist-circumference-waist-to-hip-ratio-1745981

Gupta, S., & Bansal, S. (2020). Does a rise in BMI cause an increased risk of diabetes?: Evidence from India. *PloS one*, 15(4), e0229716. https://doi.org/10.1371/journal.pone.0229716

Werth, R. (2022). *Categorical regression in Stata and R*. https://bookdown.org/sarahwerth2024/CategoricalBook/handling-influential-observations-stata.html

World Health Organization. (2024, August, 07). *The top 10 causes of death.* https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death