Exploring Predictors Related to Diabetes

Group 2: Seth Galluzzi (vzw6yk), Haley Egan (vkb6bn), JD Pinto (jp5ph), and Sydney Masterson (bmp6tz)

Executive Summary

According to the CDC, over 34 million US adults have diabetes. As the seventh leading cause of death in the US, diabetes is a public health problem that has reached epidemic proportions. This disease can lead to other health problems including kidney failure, cardiovascular disease, and nerve damage. The most common test to determine if someone has diabetes is an A1C test. An A1C test is used to measure the glycosylated hemoglobin levels within the bloodstream. When sugar builds up in the bloodstream and attaches itself to the hemoglobin within the red blood cells, there is a resulting increase in glycosylated hemoglobin. According to the data, a glycosylated hemoglobin value greater than 7% is considered a positive diabetes test.

The aim of our exploration is to gain an understanding of the predictors associated with diabetes. Identifying predictors of this condition can increase awareness, encourage preventative treatment and lower the number of people at risk. To determine the most accurate predictors, we first visualized the data and then performed model selection techniques (stepwise regression and forward selection).

The dataset explored is the *diabetes* dataset from the faraway package in R. It contains 403 objects and 19 variables and was obtained by interviewing African Americans to "understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia." We specifically chose this dataset for its large number of categorical and numerical variables and our interest in educating central Virginian communities about diabetes. One drawback that had to be accounted for was the small size of the data set. Some predictors within the data, such as cholesterol and age, shared some relationship with our response variable glycosylated hemoglobin. However, the predictor most correlated with glycosylated hemoglobin proved to be stabilized glucose. Adding additional predictors did little to increase the r-squared of our regressions. Therefore, more complex models were not preferred.

The visualizations that we explored included a histogram of glycosylated hemoglobin with a positive skew, and multiple scatter plots and box plots that further indicated the relationship between glycosylated hemoglobin and stabilized glucose. The linear regression between the two variables was $y = 2.2351 + 0.0312x$ with an R^2 of 0.560. Approximately 14% of African Americans within our dataset are considered at risk of diabetes.

Description of Data

To obtain the information in the dataset such as total cholesterol, weight, age, and gender, interviews were conducted with the 403 African American subjects from central Virginia. We chose this data set to make comparisons with variables that impact diabetes. These comparisons helped us to gain a better understanding of the prevalence of diabetes among African American communities in central Virginia. An explanation of the variables is below:

id	The subject ID of the observation. This column will not be needed for our exploration, and will be eliminated from the data set.
chol	The total cholesterol of the observation. Cholesterol is measured in milligrams per deciliter of blood. A healthy level of total cholesterol is between 125 to 200 mg/dL. The minimum cholesterol of our data set is 78, the maximum is 443, and the median is 204 mg/dL.
stab.glu	The stabilized glucose of the observation. Stabilized glucose is measured in milligrams per deciliter of blood.
hdl	The high density lipoprotein of the observation. HDL aides in the removal of cholesterol from the arteries. HDL is also measured in milligrams per deciliter of blood. A healthy level of HDL in males and females is above 40 mg/dL and 50 mg/dL respectively. The minimum HDL in our data set is 12, the maximum is 120, and the median is 46.
ratio	The ratio between total cholesterol and HDL. This ratio is often used to identify coronary heart disease. For perspective, a ratio of 3.5 is considered very good, while a ratio of 4.5 is considered at risk of coronary heart disease.

	Furthermore, most healthcare providers prefer a ratio less than 5. The minimum ratio in our data is 1.5, the maximum is 19.3, and the median is 4.2.
glyhb	The glycosylated hemoglobin value of the observation. This measures the amount of sugar in your blood. A glycosylated hemoglobin value greater than 7% is taken as a positive diabetes test. The minimum glycosylated hemoglobin value is 2.68, the maximum is 16.11, and the median is 4.84. There are also 13 missing values.
location	The location value of the observation. For this study, location was categorical with values Buckingham and Louisa with 200 and 203 total observations respectively.
age	The age of the observation. For this study, the minimum age was 19, the maximum age was 92, and the median was 44.
gender	The gender of the observation. For this study, gender is categorical with values male or female. There are 160 males and 228 females.
height	The height of the observation. Height was measured in inches. The minimum was 52 inches, the maximum was 76 inches, and the median was 66 inches.
weight	The weight of the observation. Weight was measured in pounds. The minimum was 99 lbs, the maximum was 325 lbs, and the median was 173.
frame	The frame of the observation. Frame is a categorical variable with values small, medium, and large. There are 101 small, 177 medium, and 99 large observations. There are also 11 missing values.
bp.1s	The first systolic blood pressure of the observation, measured in millimeters of mercury (mmHg).
bp.1d	The first diastolic blood pressure of the observation, measured in millimeters of mercury (mmHg).

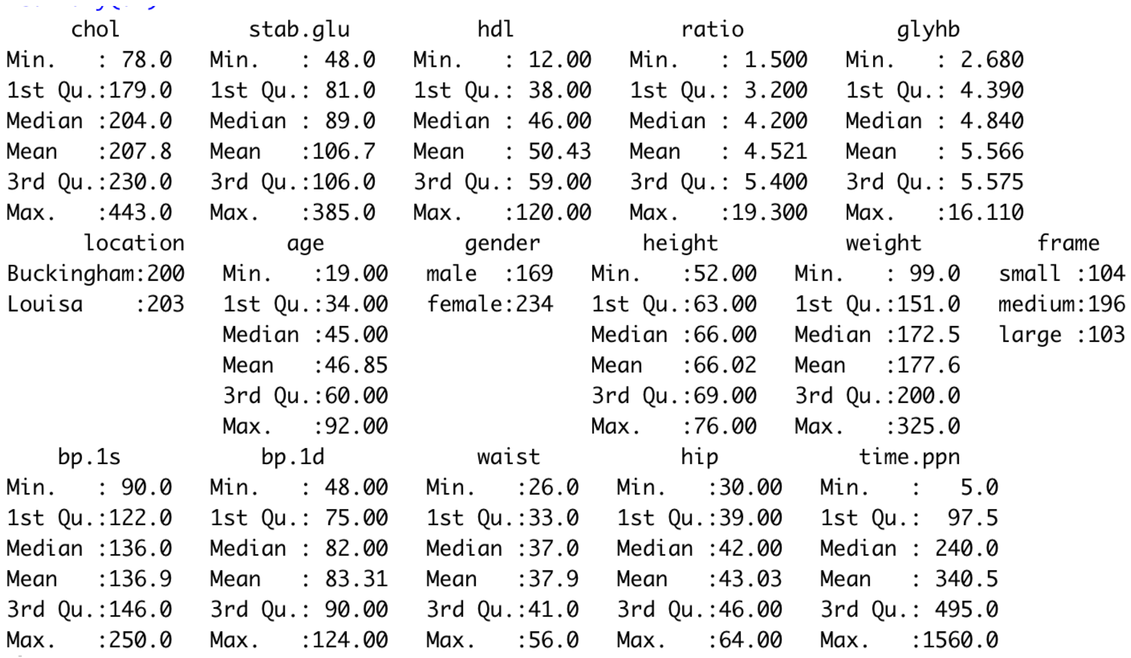
bp.2s	The second systolic blood pressure of the observation. This variable contains a multitude of missing values.
bp.2d	The second diastolic blood pressure of the observation. This variable contains a multitude of missing values.
waist	The waist measurement of the observation. Waist was measured in inches. The minimum was 26 inches, the maximum was 56 inches and the median was 37 inches.
hip	The hip length of the observation. Hip was measured in inches. The minimum was 30 inches, the maximum was 64 inches, and the median was 42 inches.
time.ppn	The postprandial time when the observation was conducted. This variable describes the amount of time since the last meal of the observation and is measured in minutes. The minimum of our data set is 5, the maximum is 1560, and the median is 240 minutes.

Some minor data cleaning was necessary before exploring the visualizations. First, the id column was deleted, as it did not provide value to the exploration of variables impacting diabetes risk. Second, the second systolic blood pressure and the second diastolic blood pressure columns were deleted from the data set. Unfortunately, both columns were missing a multitude of values and were deemed too incomplete to be considered. Finally, the missing values from the rest of the columns were replaced with their prospective medians. Medians were chosen over the prospective means because of the size of the data set and the possibility of a few outliers within each column drastically impacting the mean of the column. Most of these columns only contained one or two missing values. However, the columns glycosylated hemoglobin and frame had 13 and 12 missing values respectively. During the exploration phase, another variable was added named *diabetes*. This variable is categorical with values TRUE, when the subject's glycosylated hemoglobin levels are above 7, and FALSE, when the subject's glycosylated hemoglobin levels are equal to or less than 7.

One of the major challenges with this data set was the size. Since there were only 403 objects within the data frame, the size of the data set was a key factor when making decisions regarding the data cleaning, visualizations, and modeling. For example, we decided to replace the missing values rather than drop them because dropping them would create an even smaller data set. Furthermore, splitting the data for visualizations and modeling purposes was difficult because of the small sizes of the split data sets.

Besides missing values, it should be noted that smoking status was not collected during the interviews. According to the FDA, smokers are 30-40% more likely to develop type 2 diabetes. Thus, smoking status would be an interesting categorical variable to explore. However, since it was not included in the data set, we cannot account for its impact.

Below is the summary of the variables within the cleaned data. In the visualization stage, we also introduced the categorical variable *diabetes (not shown here),* which was described earlier in this section. Figure 1 depicts general information about our data. The median value of glycosylated hemoglobin is 4.840, with a minimum of 2.680 and a maximum of 16.110. The median value of stabilized glucose is 89, with a minimum value of 48 and a maximum of 385. The maximum of glycosylated hemoglobin is much higher than the median, indicating the possible presence of outliers in our dataset.



```
      chol          stab.glu          hdl             ratio            glyhb       
 Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.500   Min.   : 2.680  
 1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.200   1st Qu.: 4.390  
 Median :204.0   Median : 89.0   Median : 46.00   Median : 4.200   Median : 4.840  
 Mean   :207.8   Mean   :106.7   Mean   : 50.43   Mean   : 4.521   Mean   : 5.566  
 3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00   3rd Qu.: 5.400   3rd Qu.: 5.575  
 Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.300   Max.   :16.110  
       location        age            gender        height          weight          frame    
 Buckingham:200   Min.   :19.00   male  :169   Min.   :52.00   Min.   : 99.0   small :104  
 Louisa    :203   1st Qu.:34.00   female:234   1st Qu.:63.00   1st Qu.:151.0   medium:196  
                  Median :45.00                Median :66.00   Median :172.5   large :103  
                  Mean   :46.85                Mean   :66.02   Mean   :177.6               
                  3rd Qu.:60.00                3rd Qu.:69.00   3rd Qu.:200.0               
                  Max.   :92.00                Max.   :76.00   Max.   :325.0               
     bp.1s           bp.1d            waist          hip           time.ppn     
 Min.   : 90.0   Min.   : 48.00   Min.   :26.0   Min.   :30.00   Min.   :   5.0  
 1st Qu.:122.0   1st Qu.: 75.00   1st Qu.:33.0   1st Qu.:39.00   1st Qu.:  97.5  
 Median :136.0   Median : 82.00   Median :37.0   Median :42.00   Median : 240.0  
 Mean   :136.9   Mean   : 83.31   Mean   :37.9   Mean   :43.03   Mean   : 340.5  
 3rd Qu.:146.0   3rd Qu.: 90.00   3rd Qu.:41.0   3rd Qu.:46.00   3rd Qu.: 495.0  
 Max.   :250.0   Max.   :124.00   Max.   :56.0   Max.   :64.00   Max.   :1560.0  
```

Figure 1: General Information Regarding our Data

Using the cleaned data, we explored visualizations with different predictor variables. Glycosylated hemoglobin was used as the response variable to determine diabetes risk. A large number of the visualizations focused on glycosylated hemoglobin and the predictors that impact it.

Data Exploration and Visualizations

The histogram in figure 2 illustrates the frequency of glycosylated hemoglobin in the bloodstream within the subjects contained in the dataset . The large number of observations between four and six is consistent with the previous analysis of the data from figure 1. The histogram has a positive skew, indicating a curve that is not normally distributed.

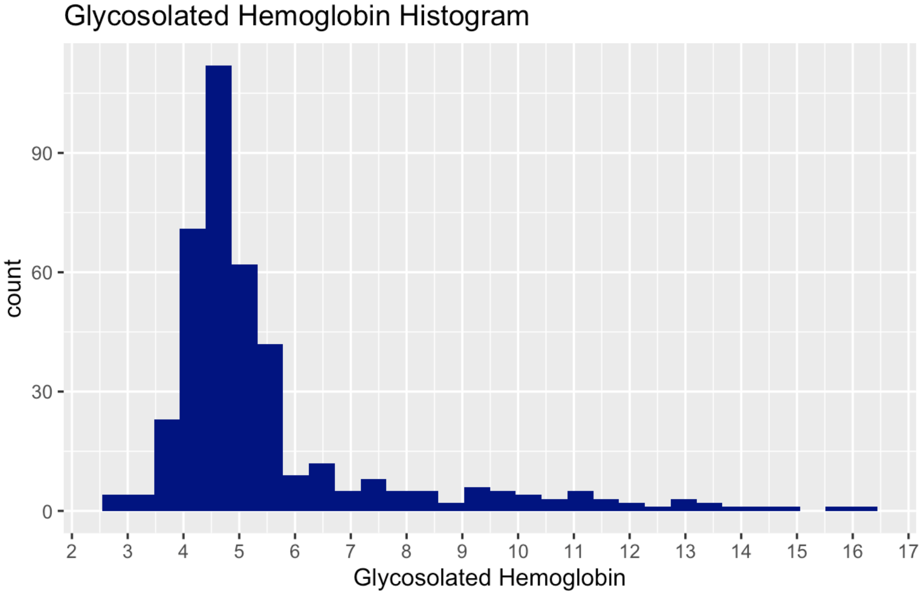


Figure 2: Glycosylated Hemoglobin Histogram

The   q-q plot below also demonstrates the presence of positive skew.

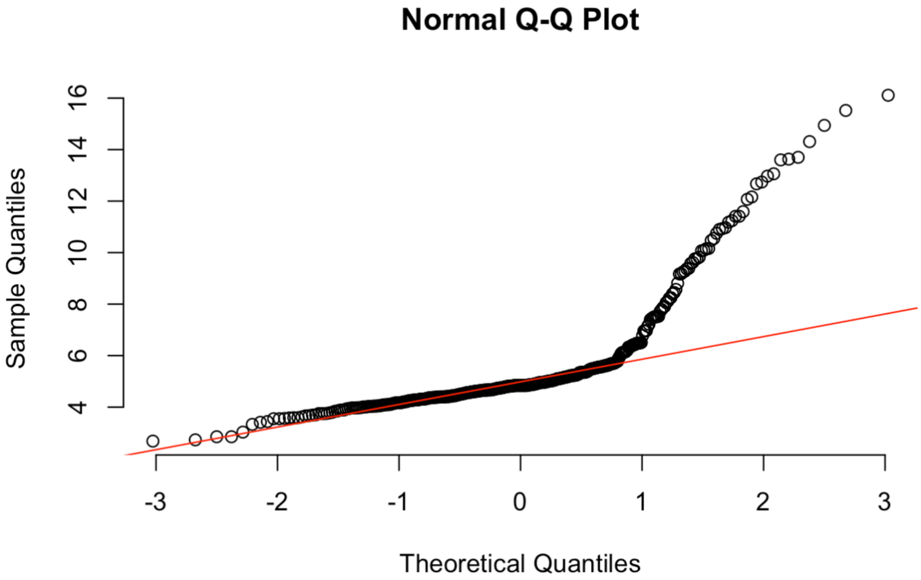


Figure 3: Glycosylated Hemoglobin Q-Q Plot

The scatter plot in figure 4 illustrates glycosylated hemoglobin and stabilized glucose between genders. The visualization indicates a positive linear relationship between the variables.. Additionally, the data indicates that only males have stabilized glucose values greater than 300.

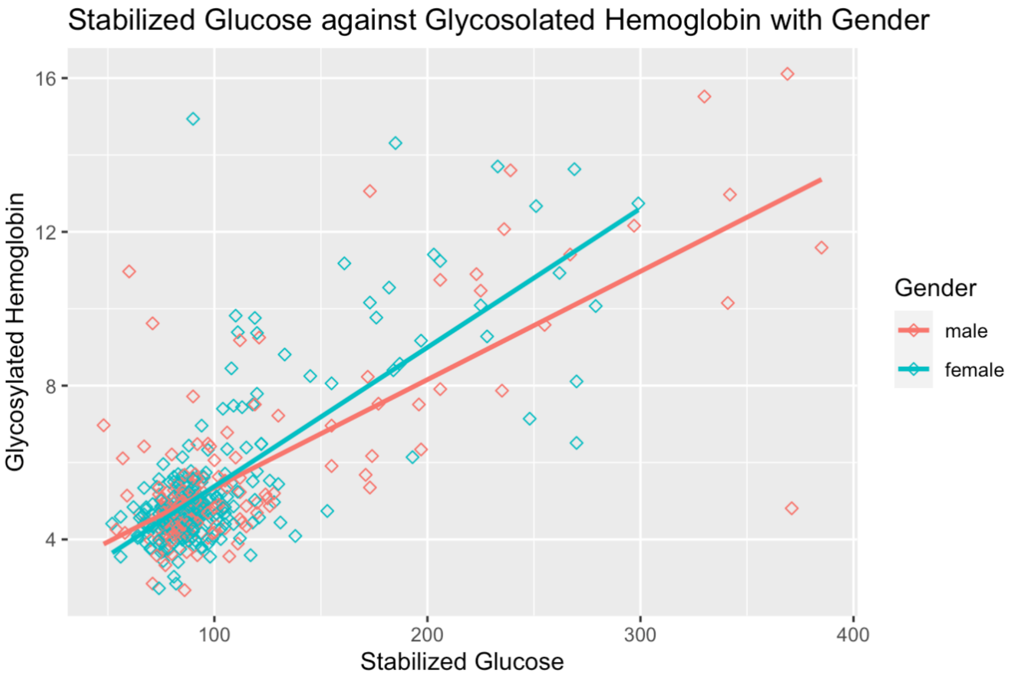


Figure 4: Stabilized Glucose Against Glycosylated Hemoglobin with Gender

As stated previously, a glycosylated hemoglobin value greater than 7 indicates diabetes risk. Using the diabetes variable, the bar graph in figure 5 depicts the total observations of diabetes risk between genders. The split between genders does not appear significant as there are roughly the same number of males and females. A total of 60 subjects were at risk of diabetes, while 338 were not. Nearly 15% of the subjects were categorized as a diabetes risk.

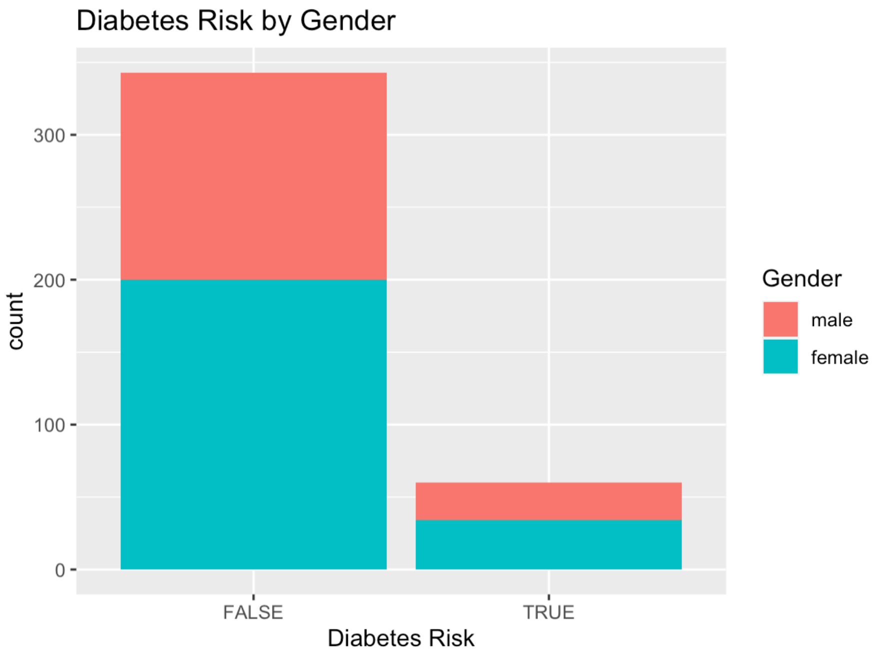


Figure 5: Diabetes Risk by Gender

The box plots in figure 6 compare the stabilized glucose and the cholesterol values of subjects at risk and not at risk of diabetes. The drastic differences in the plots indicates a major difference in the stabilized glucose values for subjects with and without diabetes risk. The difference in the cholesterol values of subjects with and without diabetes risk is minimal.

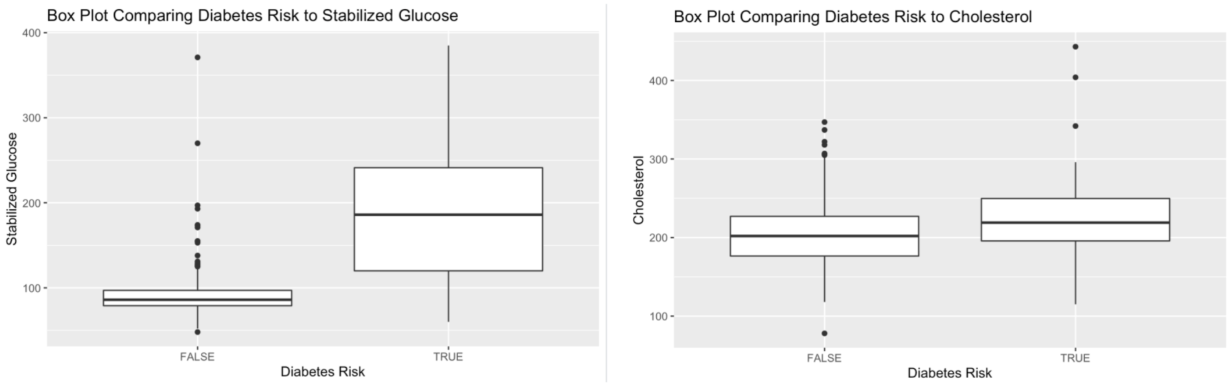


Figure 6: Box Plot Comparison of Stabilized Glucose and Cholesterol

Based on the exploration above, analyzing a simple linear regression between stabilized glucose and glycosylated hemoglobin made sense.

Simple Linear Regression (SLR)

Our model diagnostics showed that the response variable ‘glycosylated hemoglobin’ was most closely related to the predictor ‘stabilized glucose’. As a result, we conducted a Simple Linear Regression analysis on these response and predictor variables to gain insight. . Figure 7 below shows a scatterplot of Glycosylated Hemoglobin against Stabilized Glucose.

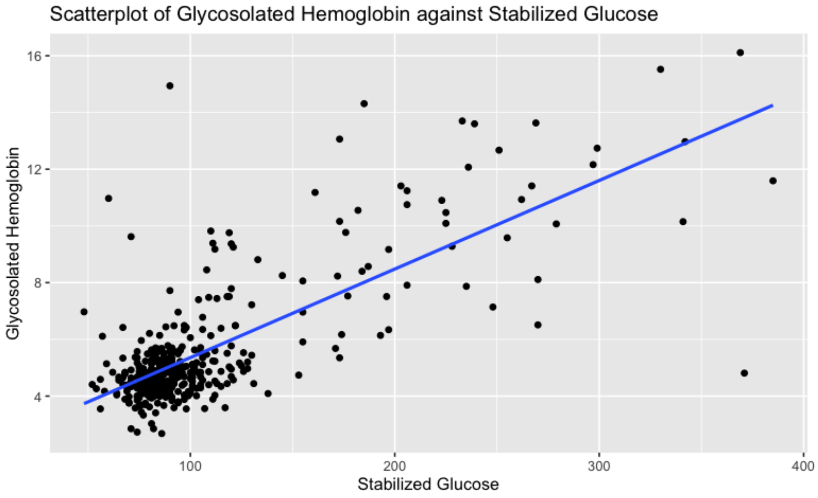
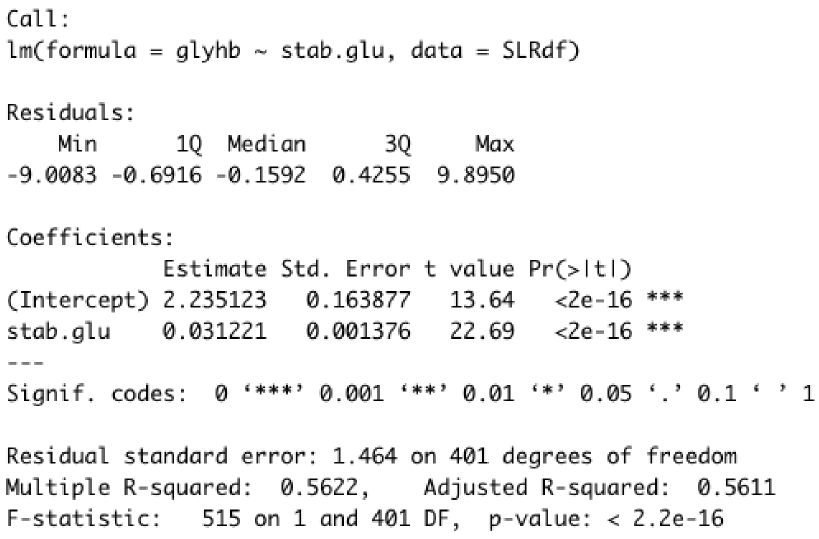


Figure 7: Scatter Plot of Glycosylated Hemoglobin against Stabilized Glucose

The initial scatterplot shows a positive linear relationship between glycosylated hemoglobin and stabilized glucose. However, there is clustering in the bottom left of the scatter plot. This indicates that not all of the regression assumptions are met, requiring further analysis and transformations of the data.



```
Call:
lm(formula = glyhb ~ stab.glu, data = SLRdf)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0083 -0.6916 -0.1592  0.4255  9.8950 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.235123   0.163877   13.64   <2e-16 ***
stab.glu    0.031221   0.001376   22.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.464 on 401 degrees of freedom
Multiple R-squared:  0.5622,	Adjusted R-squared:  0.5611 
F-statistic:   515 on 1 and 401 DF,  p-value: < 2.2e-16
```

Figure 8: Simple Linear Regression

From the Simple Linear Regression model in figure 8, the estimated regression equation is y = 2.2351 + 0.0312x. In other words, for every unit increase in stabilized glucose, glycosylated hemoglobin levels in the bloodstream will increase by 0.0312 percent holding all else constant. The SLR model confirms that there is a positive linear relationship between the response and predictor.

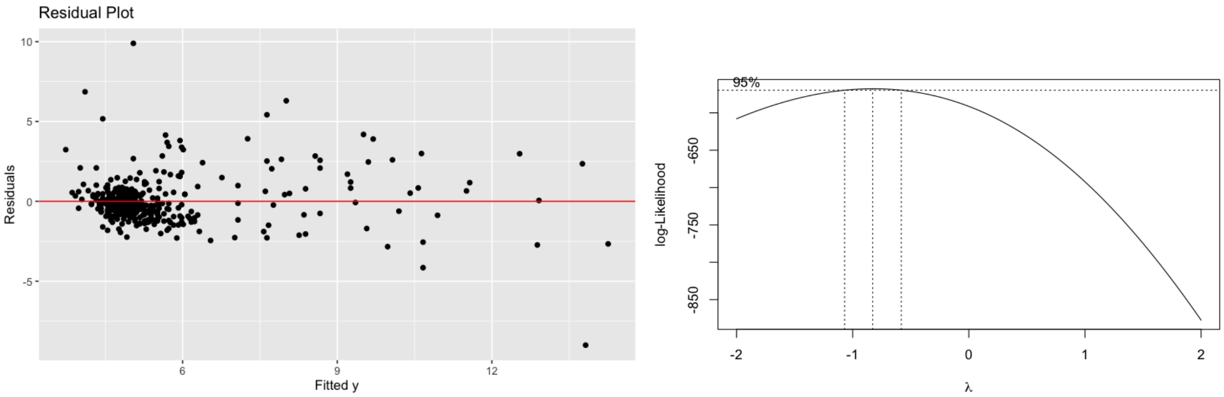


Figure 9: Residual and Box Cox Plots

Model diagnostics of the SLR were conducted to check the regression assumptions. Although there is a positive linear relationship between stabilized glucose and glycosylated hemoglobin and the error terms do appear to have a mean of zero, the residual plot in figure 9 shows that the error terms are clustered on the left side of the plot (also seen in the scatter plot). This could indicate that the error terms do not have a constant variance. From the scatter plot there also appears to be a linear pattern, but there is clustering on the left side, showing that the residuals are not evenly scattered around the line without an apparent pattern. Given these issues, some of the linear regression assumptions were not met. Therefore, a transformation of the data was necessary to meet the linear regression assumptions.

We used a box cox plot to help determine what type of transformation was needed to meet the linear regression assumptions. We determined that a log transformation of the t response variable would help.

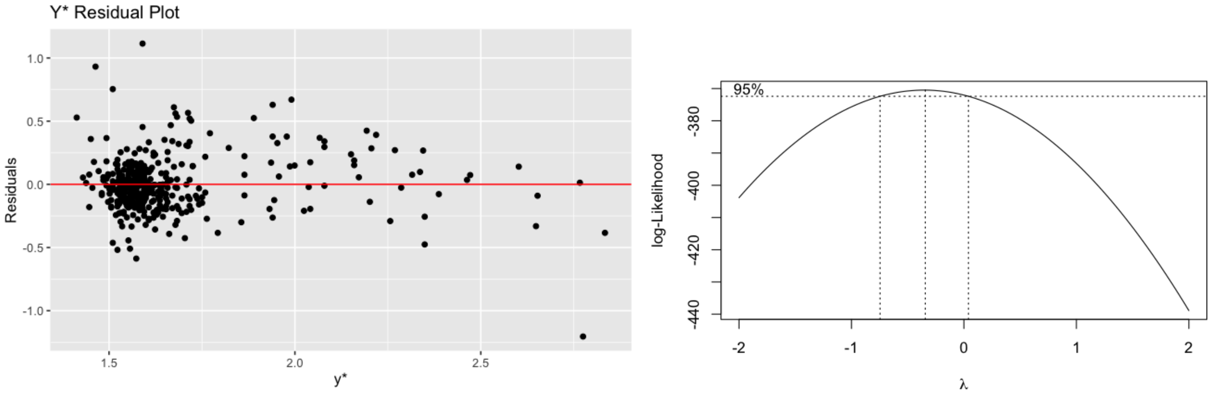


Figure 10: Transformed Residual and Box Cox Plots

The transformation against y had some improvements for the error terms, but there was still clustering. Because of this, we also used a log transformation on the predictor variable to help spread out the data along the x axis.. The result of the transformation is seen in figure 11.

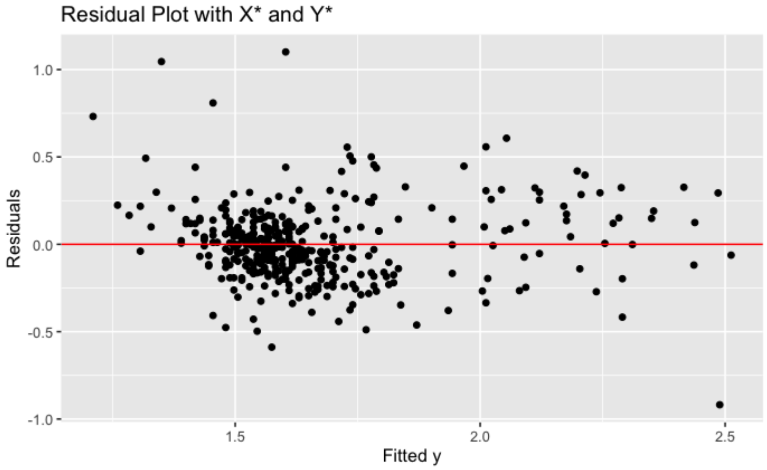


Figure 11: Transformed Residual

A transformation of both x and y did improve the model. However, while the error terms are more spread out, there was still some clustering on the left side of the residual plot.

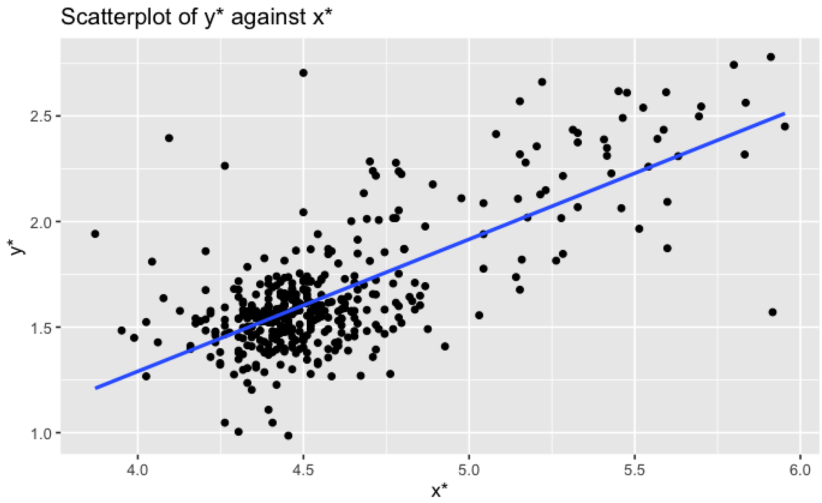


Figure 12: Scatter plot after Transformations

The scatter plot above (figure 12) shows improvement due to the log transformations of both the x and y variables. Clustering is still visible, but a linear relationship is more apparent.

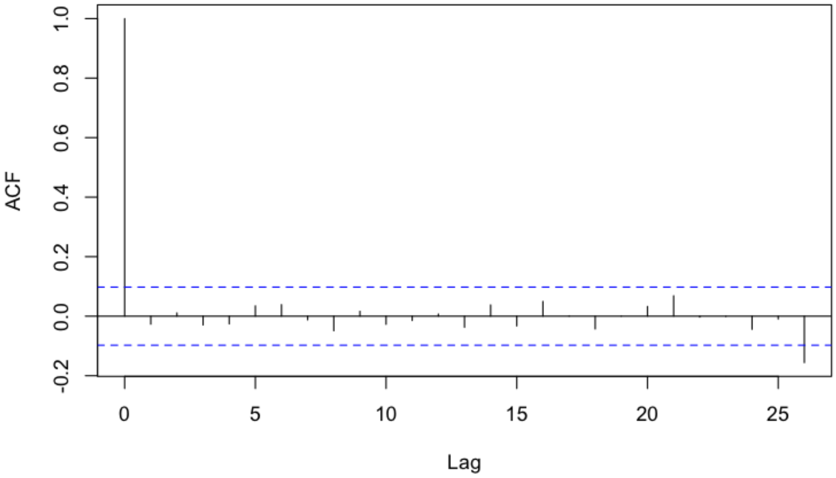


Figure 13: ACF Plot

An ACF plot, depicted in figure 13, was created to check assumption 4 - that the error terms are uncorrelated and independent. The assumption is largely met aside from the far left and right where the vertical lines cross the horizontal blue lines. However, the majority of the data is within the blue bands.

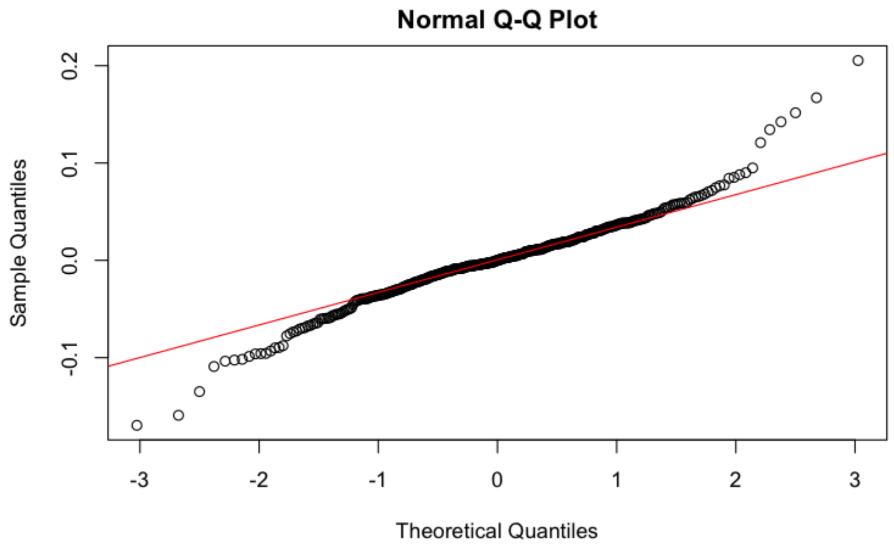


Figure 14: Q-Q Plot after Transformations

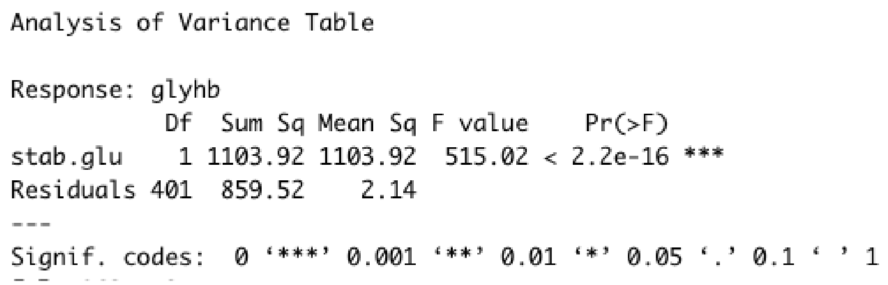
The QQ plot, seen in figure 14, checks assumption 5 - that the errors follow a normal distribution. A normal probability plot should fall close to the red line, representing expected values under normality. The errors almost entirely fall on the red line, except on the far left and right side where the errors curve downward and upward respectively. This assumption is regarded as the least important of the five and is close to being met.

After checking the assumptions, the correlation and residuals of glycosylated hemoglobin and stabilized glucose were analyzed. The correlation between the predictor and response is 0.749824. Thus, there is a strong positive relationship between glycosylated hemoglobin and stabilized glucose. The average value of all the residuals is 1.399475e-17. This suggests that the average data point is above, but very close to the fitted line, because 1.399475e-17 is very close to 0. On average, the line appears to fit the data well.

A hypothesis test was conducted on the Simple Linear Regression of glycosylated hemoglobin and stabilized glucose to determine whether or not there is evidence of a linear relationship.

H_0: $B_1 = 0$; H_a: $B_1 \neq 0$

H_0, the null hypothesis, suggests that there is no evidence of a linear relationship between glycosylated hemoglobin and stabilized glucose, (regression line would be flat). H_a, the alternative hypothesis, suggests that there is evidence of a linear relationship between glycosylated hemoglobin and stabilized glucose, (regression line would not be flat).



```
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq F value    Pr(>F)
stab.glu    1 1103.92 1103.92  515.02 < 2.2e-16 ***
Residuals 401  859.52    2.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 15: Anova Table

Based on the Anova F test in figure 15 for this SLR, we rejected the null hypothesis and concluded that there is a linear relationship between glycosylated hemoglobin and stabilized glucose. The p-value is very low, at 2.2e-16, less than any alpha, so it is unlikely that the slope of 0.0312 occurred by chance.

Outliers/Residuals

Before running a MLR we checked for outliers in the data by looking at the standardized residuals, studentized residuals, and externally studentized residuals. We also used the Bonferroni procedure to identify 3 potential outliers.

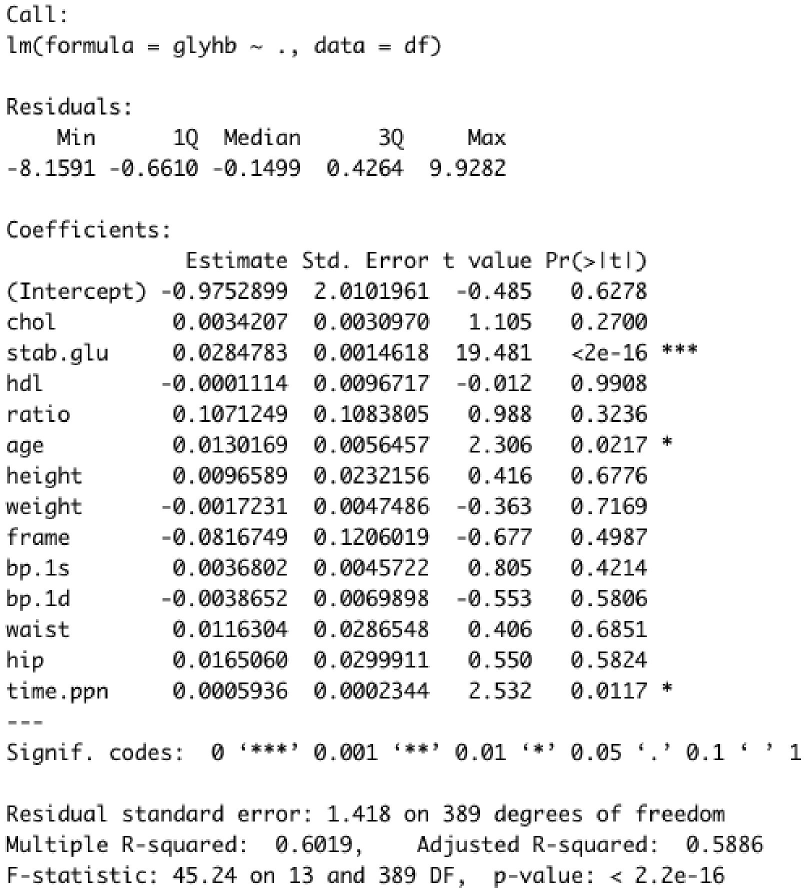
We calculated leverages, h_{ii}, to identify how far observation i is from the centroid of the predictor space. If h_{ii} >2pn, then observation i has high leverage and is an outlier in the predictor space. High leverage observations are data points that are most likely to be influential. All 403 of the data observations were identified as having a high leverage.

Cook’s distance, D_i, can be interpreted as the squared Euclidean distance that the vector of fitted values moves when observation i is removed from the regression model. A cutoff rule for an influential observation is D_i > F0.5,p.n−p. Based on Cook’s distance, no observations in the data set were identified as influential. $DFFITS_i$ measures how much the fitted value of observation i changes when it is removed from the regression model. Observation i is influential if $|DFFITS_i| > 2\sqrt{p}/n$. 32 observations were identified as influential based on $DFFITS_i$. For DFBETAS, 21 observations were identified.

Based on the multiple tests for residuals and outlying observations, there were no consistencies that could point toward any clear outlier in the data set. Therefore, no data points were removed from the analysis.

Multiple Linear Regression (MLR)

For Multiple Linear Regression, a new data frame was created with the following predictors: 'chol', 'stab.glu', 'hdl', 'ratio', 'age', 'height', 'weight', 'frame', 'bp.1s', 'bp.1d', 'waist', 'hip', 'time.ppn'. ‘Frame’ was factored into 1 for small, 2 for medium, and 3 for large. Categorical variables and variables with large amounts of missing data were removed. ‘Glyhb’ was left as the response variable. The result is shown in figure 16 below.



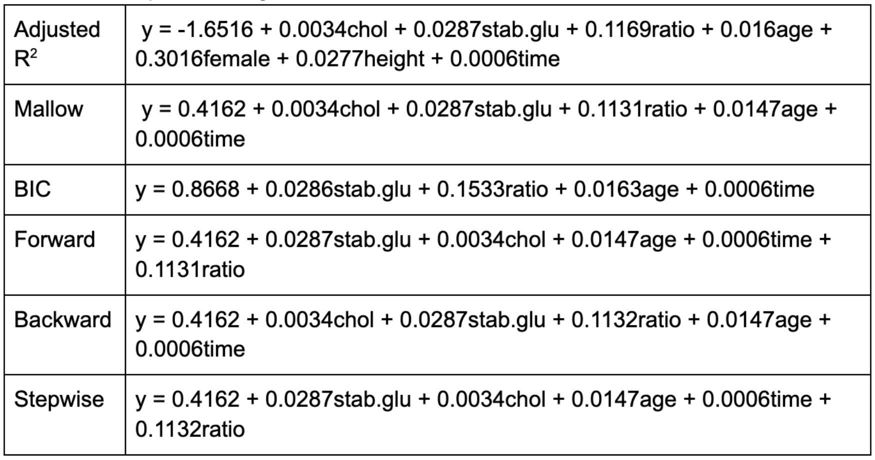
```
Call:
lm(formula = glyhb ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1591 -0.6610 -0.1499  0.4264  9.9282 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.9752899  2.0101961  -0.485   0.6278    
chol         0.0034207  0.0030970   1.105   0.2700    
stab.glu     0.0284783  0.0014618  19.481   <2e-16 ***
hdl         -0.0001114  0.0096717  -0.012   0.9908    
ratio        0.1071249  0.1083805   0.988   0.3236    
age          0.0130169  0.0056457   2.306   0.0217 *  
height       0.0096589  0.0232156   0.416   0.6776    
weight      -0.0017231  0.0047486  -0.363   0.7169    
frame       -0.0816749  0.1206019  -0.677   0.4987    
bp.1s        0.0036802  0.0045722   0.805   0.4214    
bp.1d       -0.0038652  0.0069898  -0.553   0.5806    
waist        0.0116304  0.0286548   0.406   0.6851    
hip          0.0165060  0.0299911   0.550   0.5824    
time.ppn     0.0005936  0.0002344   2.532   0.0117 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.418 on 389 degrees of freedom
Multiple R-squared:  0.6019,	Adjusted R-squared:  0.5886 
F-statistic: 45.24 on 13 and 389 DF,  p-value: < 2.2e-16
```

Figure 16: MLR with All Predictors

A linear model was run for all the predictors, against the response. Looking at the t test for all the coefficients, the only coefficients with significant p-values are stabilized glucose (stab.glu), age in years (age), and the postprandial time (in minutes) from when labs were drawn (time.ppn). All other 11 predictors have insignificant p-values.

When comparing all of the variables, there were some signs of multicollinearity. There was a moderately high correlation of -0.69 between hdl and ratio. There were also very strong correlations between waist and weight, and hip and waist, which does make sense. When examining the VIFs for all predictors, none of the VIFs exceeded 10, which indicated that there is not a high correlation between predictors.

In order to determine the best model for Multiple Linear Regression, model diagnostics were performed, including all possible regressions, adjusted R^2, CP, BIC, and Forward Selection, Backward Elimination, and Stepwise Regression. The results are shown in figure 17.



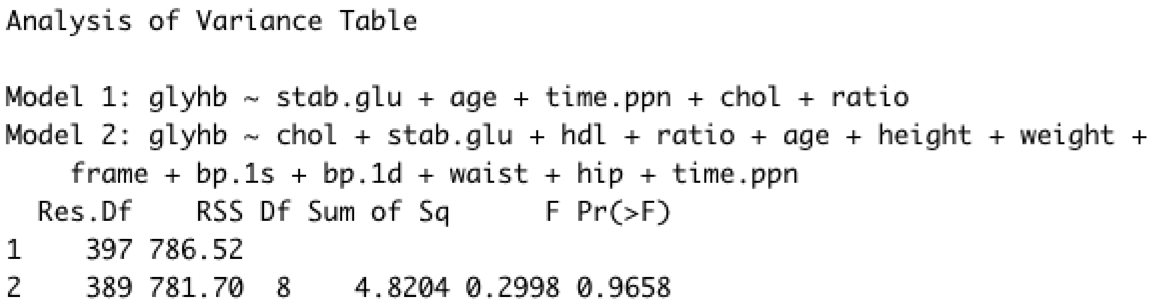
Adjusted R^2	y = -1.6516 + 0.0034chol + 0.0287stab.glu + 0.1169ratio + 0.016age + 0.3016female + 0.0277height + 0.0006time
Mallow	y = 0.4162 + 0.0034chol + 0.0287stab.glu + 0.1131ratio + 0.0147age + 0.0006time
BIC	y = 0.8668 + 0.0286stab.glu + 0.1533ratio + 0.0163age + 0.0006time
Forward	y = 0.4162 + 0.0287stab.glu + 0.0034chol + 0.0147age + 0.0006time + 0.1131ratio
Backward	y = 0.4162 + 0.0034chol + 0.0287stab.glu + 0.1132ratio + 0.0147age + 0.0006time
Stepwise	y = 0.4162 + 0.0287stab.glu + 0.0034chol + 0.0147age + 0.0006time + 0.1132ratio

Figure 17: Predictive Models Table

With the adjusted R^2, CP, BIC, there were different candidate models. Mallow's model contained the predictors of 'chol', 'stab.glu', 'ratio', 'age', and 'time.ppn'. BIC's model contained 'stab.glu', 'ratio', 'age', and 'time.ppn'. For Forward Selection, Backward Elimination, and Stepwise Regression, the suggested model was the same for all three tests, with the predictors 'stab.glu', 'chol', 'age', 'time.ppn', and 'ratio'.

A hypothesis test was conducted to determine if the reduced model could be used over the full model. The null hypothesis supports going with the reduced model by dropping the insignificant predictors, whereas the alternative hypothesis supports the full model by not dropping any predictors.

H_0 : $B_1 = B_3 = B_4 = B_6 = B_7 = B_8 = B_9 = B_{10} = B_{11} = B_{12} = B_{13} = 0$
H_a : at least one of the coefficients in H_0 is not 0.

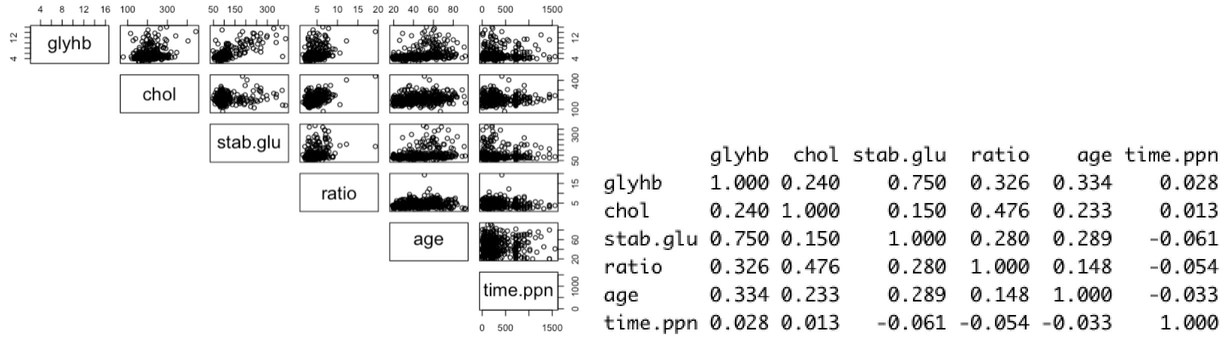


```
Analysis of Variance Table

Model 1: glyhb ~ stab.glu + age + time.ppn + chol + ratio
Model 2: glyhb ~ chol + stab.glu + hdl + ratio + age + height + weight +
    frame + bp.1s + bp.1d + waist + hip + time.ppn
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    397 786.52
2    389 781.70  8    4.8204 0.2998 0.9658
```

Figure 18: ANOVA Table

Based on the Partial F Test and the ANOVA table in figure 18, the F statistic is 0.2998, and the p-value is 0.9658. Due to the high p-value, we fail to reject the null hypothesis. There is little evidence for supporting the full model. Therefore, the reduced model can be used.

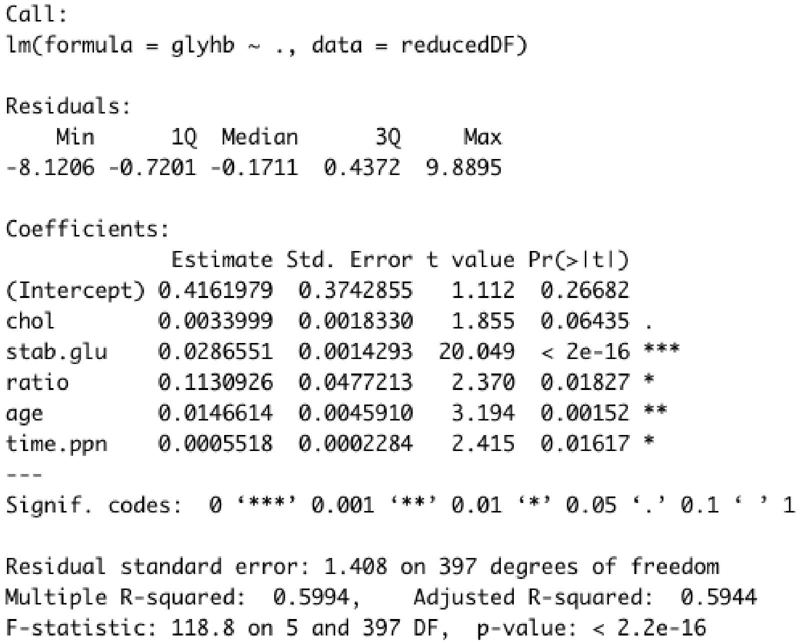
Following the guidance from the multiple model diagnostics and the hypothesis test performed, a reduced data frame was created to run the MLR, with the response of ‘glyhb’, and predictors ‘stab.glu’, ‘chol’, ‘age’, ‘time.ppn’, and ‘ratio’.



```
         glyhb  chol stab.glu  ratio    age time.ppn
glyhb    1.000 0.240    0.750  0.326  0.334    0.028
chol     0.240 1.000    0.150  0.476  0.233    0.013
stab.glu 0.750 0.150    1.000  0.280  0.289   -0.061
ratio    0.326 0.476    0.280  1.000  0.148   -0.054
age      0.334 0.233    0.289  0.148  1.000   -0.033
time.ppn 0.028 0.013   -0.061 -0.054 -0.033    1.000
```

Figure 19: Scatter Plots and Correlation Matrix

A correlation matrix and pairwise correlation shown in figure 19 indicate a strong correlation between glycosylated hemoglobin (glyhb) and stabilized glucose (stab.glu), at 0.75. There does not appear to be a strong correlation between the remaining variables.



```
Call:
lm(formula = glyhb ~ ., data = reducedDF)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1206 -0.7201 -0.1711  0.4372  9.8895 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.4161979  0.3742855   1.112  0.26682    
chol        0.0033999  0.0018330   1.855  0.06435 .  
stab.glu    0.0286551  0.0014293  20.049  < 2e-16 ***
ratio       0.1130926  0.0477213   2.370  0.01827 *  
age         0.0146614  0.0045910   3.194  0.00152 ** 
time.ppn    0.0005518  0.0002284   2.415  0.01617 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.408 on 397 degrees of freedom
Multiple R-squared:  0.5994,	Adjusted R-squared:  0.5944 
F-statistic: 118.8 on 5 and 397 DF,  p-value: < 2.2e-16
```

Figure 20: MLR

A Multiple Linear Regression model in figure 20 was run for the reduced model. The estimated regression equation is y = 0.4162 + 0.0034(chol) + 0.0287(stab.glu) + 0.1131(ratio) + 0.0147(age) + 0.0006(time.ppn). All the predictors have significant p-values, except 'chol'.

It was then necessary to check the regression assumptions and make sure they are met as much as possible before analysis.

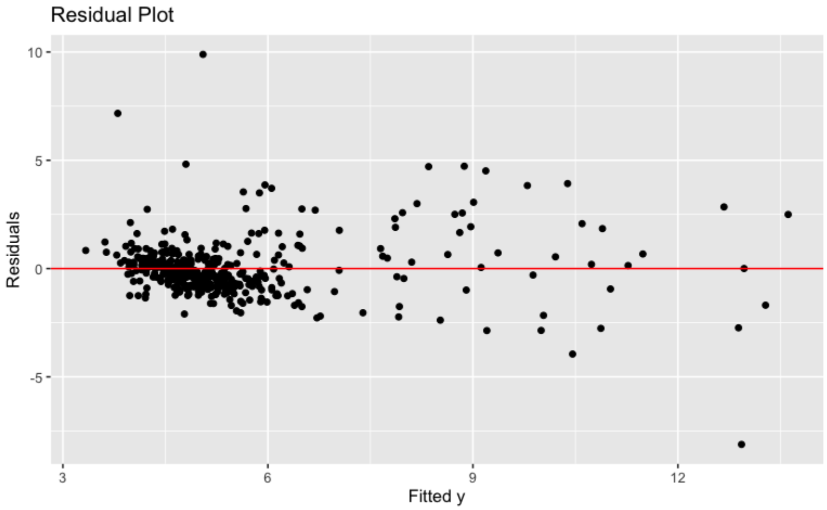


Figure 21: Residual Plot

The residual plot in figure 21 shows that the error terms are clustered on the far left of the residual plot. Therefore, there is not a constant variance and assumption 3 is not entirely met. However, assumption 2 is met, because the error terms appear to have a mean of 0.

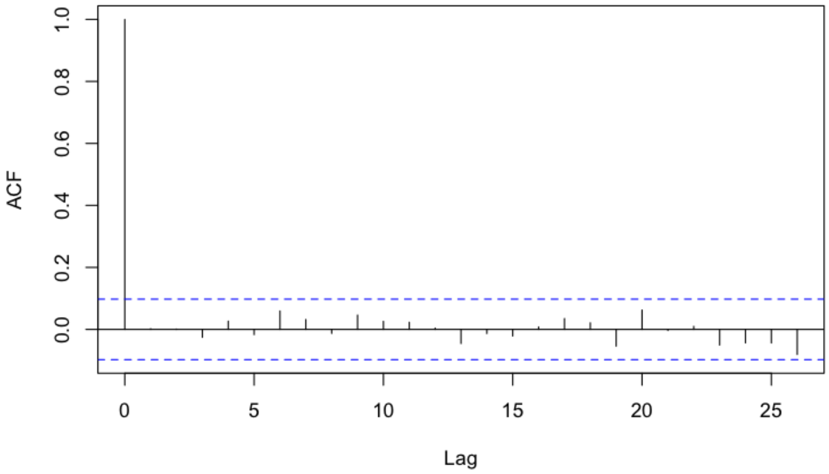


Figure 22: ACF Plot

Looking at the ACF plot in figure 22, assumption 4 appears to be met, because the error terms appear to be uncorrelated and independent. There are no vertical lines outside of the horizontal blue zone.

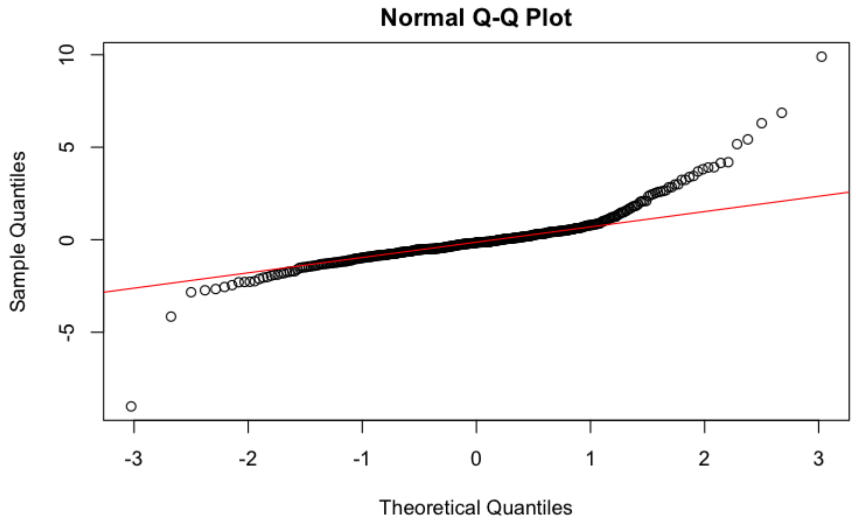
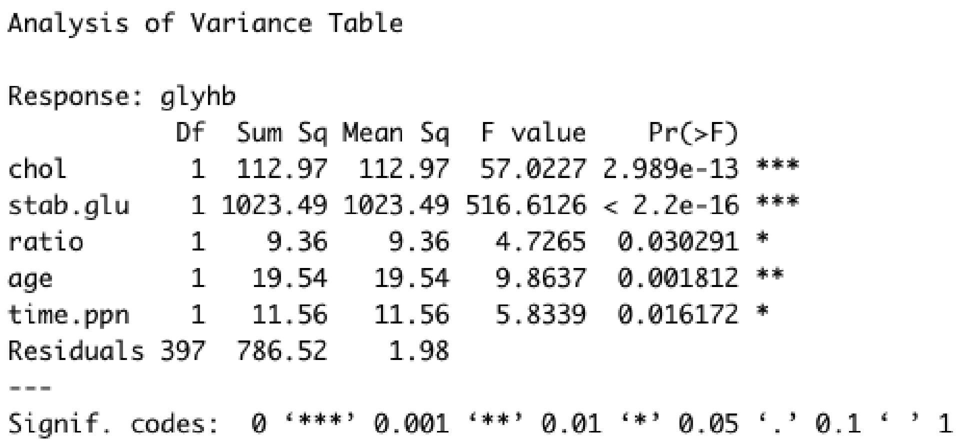


Figure 23: Q-Q Plot

Based on the Q-Q plot in figure 23, assumption 5 is largely met, indicating that the error terms follow normal distribution. The error terms fall along the red line, except for at the far right of the plot, which indicates a positive skew.

A hypothesis test was conducted to determine if there is a relationship between the response variable and predictor variables. H0 suggested that there is no evidence of a linear relationship, and Ha suggests that there is a linear relationship.

H_0: $B_1 = 0$; H_a: $B_1 \neq 0$



```
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
chol        1  112.97  112.97  57.0227 2.989e-13 ***
stab.glu    1 1023.49 1023.49 516.6126 < 2.2e-16 ***
ratio       1    9.36    9.36   4.7265  0.030291 *  
age         1   19.54   19.54   9.8637  0.001812 ** 
time.ppn    1   11.56   11.56   5.8339  0.016172 *  
Residuals 397  786.52    1.98                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

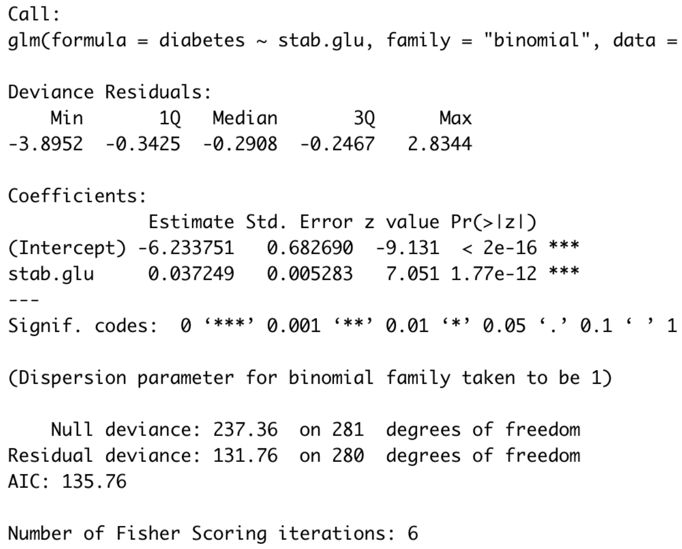
Figure 24: Anova Table

Based on the ANOVA F test (figure 24) for this MLR, all p-values are smaller than alpha, 0.05. Thus, we reject the null hypothesis and conclude that there is a relationship between the response variable, glycosylated hemoglobin, and the predictor variables.

Testing the Model

Logistic Regression 1 Predictor Model

To test our model, we ran two logistic regressions. We have both a 1 predictor model and a 5 predictor model. The output below is for the one predictor model.



```
Call:
glm(formula = diabetes ~ stab.glu, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8952  -0.3425  -0.2908  -0.2467   2.8344

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.233751   0.682690  -9.131  < 2e-16 ***
stab.glu     0.037249   0.005283   7.051 1.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 237.36  on 281  degrees of freedom
Residual deviance: 131.76  on 280  degrees of freedom
AIC: 135.76

Number of Fisher Scoring iterations: 6
```

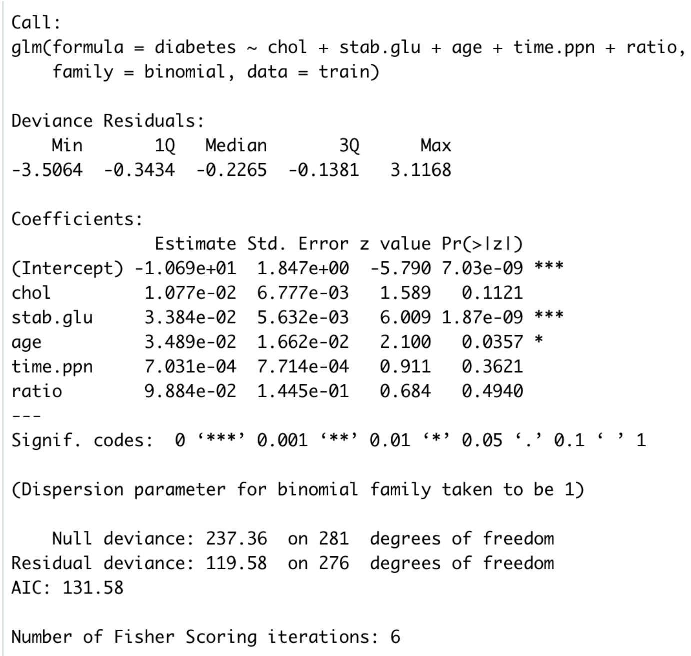
Figure 25: Logistic Regression 1 Predictor

The logistic regression in figure 25 above has diabetes as the response variable and stabilized glucose as the predictor variable. As you can recall from earlier, the diabetes column in the data set is a categorical variable with a value of 1 if glycosylated hemoglobin is above 7% in the bloodstream and 0 otherwise. This model is testing the likelihood of having diabetes. The results of this model indicate that the log odds of having diabetes increase by 0.037249 for an additional milligram of stabilized glucose per deciliter of blood.

Logistic Regression 5 Predictor Model

The output below is for the logistic regression with 5 predictors. The response variable is the likelihood of having diabetes. The predictor variables are cholesterol (chol) , stabilized glucose (stab.glu), age, time (time.ppn) and ratio. As we can see, besides the intercept, only the predictors of stabilized glucose and age are statistically significant.

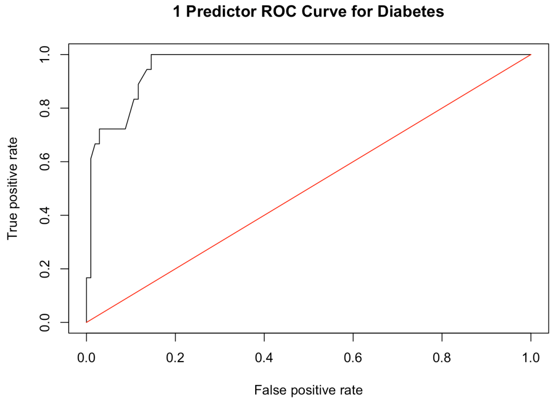
The results of this model indicate that the log odds of having diabetes increase by 0.0384 for an additional milligram of stabilized glucose per deciliter of blood holding all else constant. Additionally, for each additional year of age the log odds of having diabetes increase by 0.01077 holding all else constant.

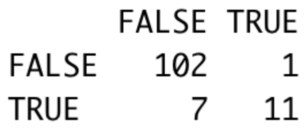


```
Call:
glm(formula = diabetes ~ chol + stab.glu + age + time.ppn + ratio,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5064  -0.3434  -0.2265  -0.1381   3.1168

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.069e+01  1.847e+00  -5.790 7.03e-09 ***
chol         1.077e-02  6.777e-03   1.589   0.1121
stab.glu     3.384e-02  5.632e-03   6.009 1.87e-09 ***
age          3.489e-02  1.662e-02   2.100   0.0357 *
time.ppn     7.031e-04  7.714e-04   0.911   0.3621
ratio        9.884e-02  1.445e-01   0.684   0.4940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 237.36  on 281  degrees of freedom
Residual deviance: 119.58  on 276  degrees of freedom
AIC: 131.58

Number of Fisher Scoring iterations: 6
```

Figure 26: Logistic Regression 5 Predictor

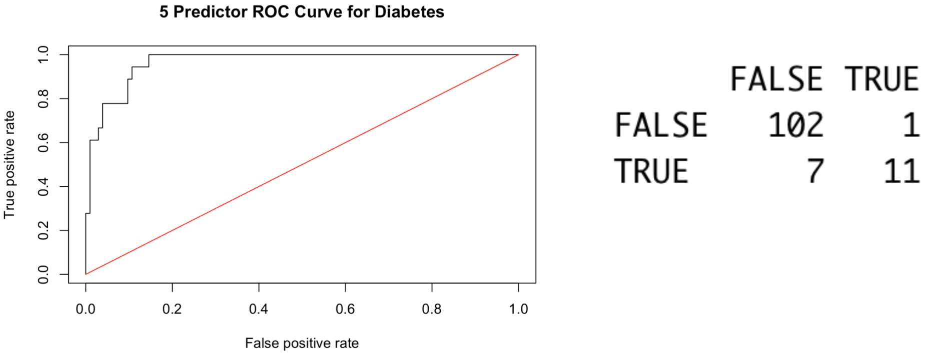
To examine what model was best when considering diabetes risk, these two models were compared based on our previous analysis. The first model, a simple linear regression between stabilized glucose and diabetes risk, is illustrated in figure 25. The second model, illustrated in figure 26, contains the predictors of cholesterol, stabilized glucose, age, postprandial time, and ratio. Both models indicate they are more accurate than random guessing. The model with five predictors has a marginally better ROC plot and an AUC score of 0.966 (compared to 0.960). This indicates that the more robust model is a better predictor of diabetes risk than the simple linear regression model. However, given how similar the AUC of both models are, the 5 predictor model is not worth the extra complexity.





```
      FALSE TRUE
FALSE   102    1
TRUE      7   11
```

Figure 26: ROC Curve of Simple Regression and Confusion Matrix (95%)



```
       FALSE TRUE
FALSE    102    1
TRUE       7   11
```

Figure 27: Roc Curve of Multiple Linear Regression and Confusion Matrix (95%)

Conclusion

Diabetes is a global epidemic that impacts millions of individuals on a day to day basis. The goal of our project was to contribute to an increased awareness of diabetes risks, and influence others to be active in the prevention of diabetes. Our analysis helped us to gain an understanding of the predictors related to this disease. We created visualizations with variables such as glycosylated hemoglobin, stabilized glucose, cholesterol, ratio, age weight etc. These visualizations showed a strong relationship between glycosylated hemoglobin and stabilized glucose.

To assess the regression assumptions and determine if transforming the response variable was necessary, we conducted a SLR between glycosylated hemoglobin and stabilized glucose. A transformation for y and x were needed, and while not perfect, the transformations did improve the error terms, better meeting the regression assumptions.

After the transformations, we analyzed the outlier and residuals, checked for multicollinearity, and performed model diagnostics to determine the best model for MLR. The model diagnostics revealed that the reduced model was more appropriate than the full model.

Through this analysis, we learned that predictors such as age, total cholesterol, and stabilized glucose are the most impactful variables to the risk of diabetes in our dataset. Our findings aligned with other research regarding the percentage of African Americans at risk of diabetes. We hope that these findings help bring awareness regarding predictors of this disease but want to be clear that our results are not exhaustive. Beyond our analysis, further exploration is important to evaluate what else can cause diabetes. A variable which was not considered is body mass index (BMI), which might be used to determine the risk of obesity, and its correlation to diabetes in African American communities. For a separate study it may be valuable to explore the variable ‘ratio’ to determine the risk of coronary heart disease in African American communities. Thank you for your time in reading our study.

References

Centers for Disease Control and Prevention. (2020, January 31). *LDL & HDL: Good & Bad Cholesterol*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/cholesterol/ldl_hdl.htm.

Centers for Disease Control and Prevention. (2021, August 10). *All about your A1C*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html.

Centers for Disease Control and Prevention. (2021, September 27). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/heartdisease/facts.htm.

Diabetes: Diabetes and obesity, cardiovascular risk factors. RDocumentation. (n.d.). Retrieved November 15, 2021, from https://www.rdocumentation.org/packages/faraway/versions/1.0.7/topics/diabetes.

Lipid panel with total cholesterol: HDL ratio. Lipid Panel with Total Cholesterol: HDL Ratio - Health Encyclopedia - University of Rochester Medical Center. (n.d.). Retrieved November 15, 2021, from https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=167&ContentID=lipid_panel_hdl_ratio.

Mayo Foundation for Medical Education and Research. (2020, January 29). *Cholesterol ratio or non-HDL cholesterol: Which is most important?* Mayo Clinic. Retrieved November 15, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/expert-answers/cholesterol-ratio/faq-20058006.

U.S. National Library of Medicine. (2020, October 2). *Cholesterol levels: What you need to know*. MedlinePlus. Retrieved November 15, 2021, from https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html.

WebMD. (n.d.). *Cardiovascular (heart) diseases: Types and treatments*. WebMD. Retrieved November 15, 2021, from https://www.webmd.com/heart-disease/guide/diseases-cardiovascular.