A1C LEVELS IN DIABETICS

Danielle Deans, Isabel Rios, Pierce Hentosh, Logan Powell

THE "SILENT KILLER"

Diabetes is considered the 8th leading cause of death in the United States. It's known as the "Silent Killer" due to the lack of symptoms and diagnosis in most individuals.

According to the CDC,

- About 38 million people are affected but only 1 in 5 are aware.
  - The number of diagnosis continues to increase every year.
- Diabetes can be fatal as it can cause serious health complications to develop due to extended abnormal blood sugar levels
  - Kidney Failure
  - Lower-limb Amputations
  - Nerve Damage
  - Heart Attacks

DIABETES CONTINUED

Diabetes is affected by several different factors such as a person's age or cholesterol.

However, symptoms aren't always obvious.

- Frequent urination
- Fatigue
- Increase in appetite and thirst.
- Weight Loss
- Blurry Vision
- Dry Skin

The lack of obvious symptoms contributes to issue of unawareness and low number of diagnosis,

BASIS OF THE MAIN PROBLEM

Due to the severity of diabetes, being able to predict how a person's life and health affects diabetes can be life saving.

Of the two types of Diabetes,

- Type 1 Diabetes can be managed.
- Type 2 Diabetes can be managed and prevented.

Lifestyle changes can be made that prevent the development of type 2 diabetes and keep A1C levels in a healthy range

To manage diabetes, lifestyle changes and medications can be administered to improve a person's quality of life.

- The CDC has a lifestyle change program that encourages healthy eating, more physical activity, and stress management.

MAIN ISSUE TO ADDRESS

HOW DO DIFFERENT PHYSICAL FACTORS (BOTH PHYSIOLOGICAL AND ENVIRONMENTAL) IMPACT THE BLOOD SUGAR LEVELS OF DIABETICS?

DATASET CHANGE

Originally, we began our research with a dataset from a survey done yearly by the CDC. Specifically, the 2023 Behavioral

Risk Factor Surveillance System survey data.

The survey collected data from landlines and cellphones in 48 states.

https://www.cdc.gov/brfss/annual_data/annual_2023.html

However, this dataset consisted mainly of categorical variables and dummy variables. The response variable for this data set was also categorical. Therefore, it required logistic regression.

We wanted to apply concepts discussed in this class. So, we decided to opt for a new dataset with a numerical response variable and apply linear regression.

DATA RESOURCE

Our new dataset stems from a study done in central Virginia on the prevalence of obesity, diabetes, and other cardiovascular risk for African Americans.

The raw data consists of 19 variables and 403 observations.

The dataset is courtesy of Dr. John Schorling, Department of Medicine, University of Virginia School of Medicine.

Schorling, J. (n.d.). Diabetes dataset.

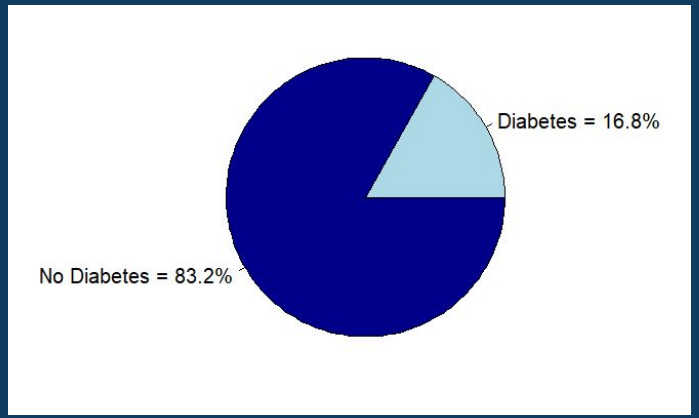
https://hbiostat.org/data/repo/diabetes

DATASET VARIABLES

- ID - study patient ID (numeric)
- chol - total cholesterol (numeric)
- stab.glu - Stabilized Glucose (fasting blood sugar - numeric)
- hdl - High Density Lipoprotein (“good cholesterol” - numeric)
- location - County patient resides in (Factor with two levels: Buckinham, Louisa)
- age - in years (numeric)
- gender - (factor with two levels: male, female)
- height - in inches (numeric)
- weight - in pounds (numeric)
- frame - (factor with three levels: small, medium, large)
- bp.1s - systolic blood pressure (numeric)
- bp.1d - diastolic blood pressure (numeric)
- bp.2s - second reading
- bp.2d - second reading
- waist - in inches (numeric)
- hip - in inches (numeric)
- time.ppn - Postprandial Time (in minutes)

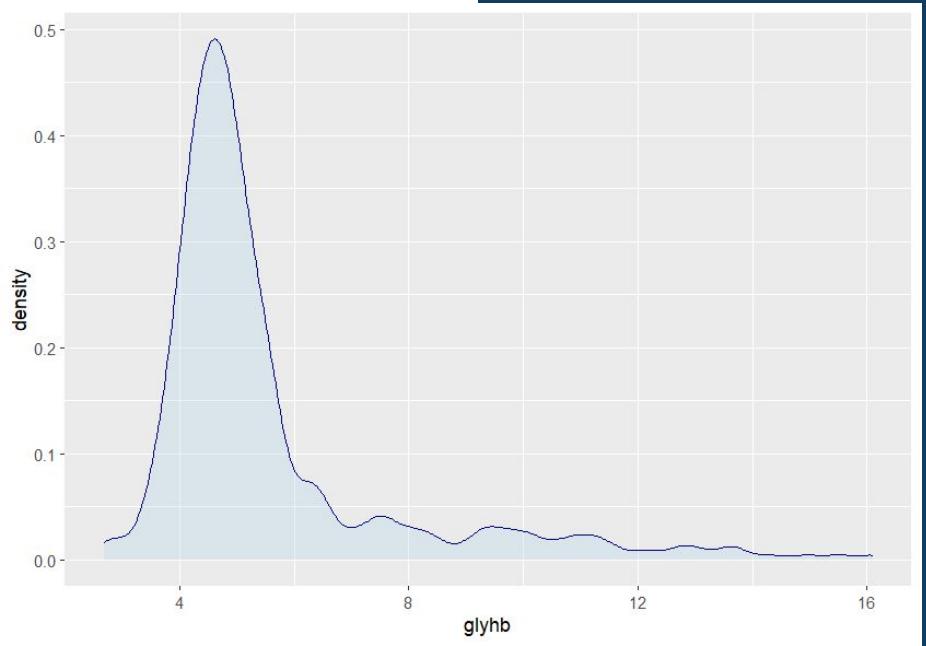
RESPONSE VARIABLE

- Response Variable: Glycosolated Hemoglobin (HbA1c)
- According to CDC 12.1% of African American adults in the US have diabetes.
- A person with HbA1c level greater than 6.5% is considered diabetic
- Following that condition, 16.8% of the observations in our data had diabetes. This is slightly higher than CDC, but not extreme



DISTRIBUTION OF HBA1C

```
The decimal point is at the |

 2 | 7788
 3 | 034455666667778888888899999
 4 | 000000000000000011111111111122222222222233333333333333333334444444444+102
 5 | 00000000000000001111111111112222222222222223333333333444445555555555+10
 6 | 011111223334444555558
 7 | 0001244555557899
 8 | 11234468
 9 | 22334466888
10 | 111256899
11 | 022446
12 | 1277
13 | 01667
14 | 39
15 | 5
16 | 1
```



CONCERNS ABOUT THE DATA

- We were unable to find an explanation of why there are missing data points (are the missing data points significant)?
- Observations were gathered through interviews, the exact methodology used for collecting the data is not known from the data source or documentation
  - There are two blood pressure readings bp.1 and bp.2
    - Is bp.2 follow up bp; is bp.2 done only when bp.1 is incorrect for patient; how much time between bp1 and bp2 readings?
  - How did they determine frame?
    - Used a different source for their Frame measurements
- There were observations that had missing values for the target variable (glyhb) so to solve this issue these observations were omitted from the analysis

DATA PREPROCESSING

Before fitting the model, several preprocessing steps had to be taken to clean and prepare the data:

- Removing or imputing missing/unuseable data
- Transforming skewed data
- Dealing with heavily correlated predictors (multicollinearity)

index	id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender
Missing Values	0.0	1.0	0.0	1.0	1.0	13.0	0.0	0.0	0.0
Percent Missing (%)	0.0	0.25	0.0	0.25	0.25	3.23	0.0	0.0	0.0

index	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
Missing Values	5.0	1.0	12.0	5.0	5.0	262.0	262.0	2.0	2.0	3.0
Percent Missing (%)	1.24	0.25	2.98	1.24	1.24	65.01	65.01	0.5	0.5	0.74

DEALING WITH MISSING DATA

- Removed ID and frame column: The ID column simply referred to the specific patient ID index and was not useful to our modeling. Based on the source studies, it was not clear how the categories of the frame column were calculated and with data for height and weight we deemed it not necessary.
- Dropped missing values for glyhb and time.ppn: These variables had a small amount of missing values (13 for glyhb and 3 for time.ppn), considering glyhb was our response, it did not seem appropriate to impute this value. For time.ppn, the value logically did not seem to follow any greater population distribution (such as height or weight).
- The final missing values were imputed based on average value of column grouped by gender.

```
diabetes_data_imputed = dataset_imputed %>%
  group_by(gender) %>%
  select(where(is.numeric)) %>%
  mutate(across(where(is.numeric), ~ replace(., is.na(.), mean(., na.rm = TRUE))))
```

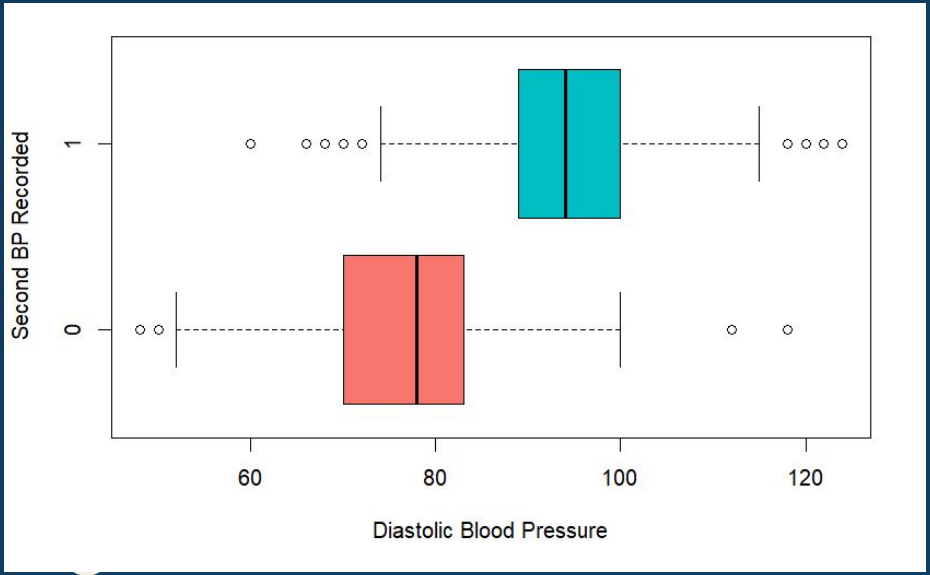
MISSING VALUES

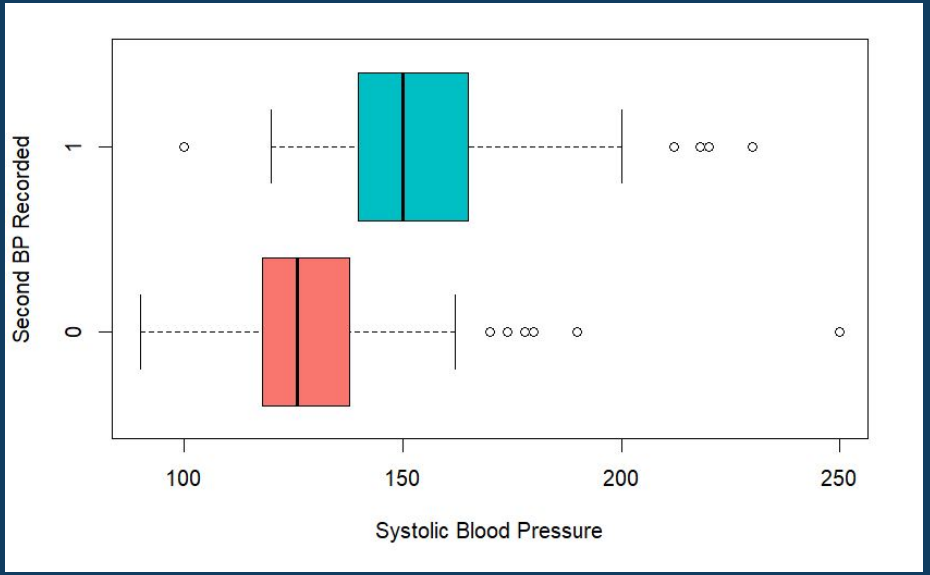
index	id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender
Missing Values	0.0	1.0	0.0	1.0	1.0	13.0	0.0	0.0	0.0
Percent Missing (%)	0.0	0.25	0.0	0.25	0.25	3.23	0.0	0.0	0.0

index	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
Missing Values	5.0	1.0	12.0	5.0	5.0	262.0	262.0	2.0	2.0	3.0
Percent Missing (%)	1.24	0.25	2.98	1.24	1.24	65.01	65.01	0.5	0.5	0.74

REMOVING SECOND BLOOD PRESSURE MEASURE

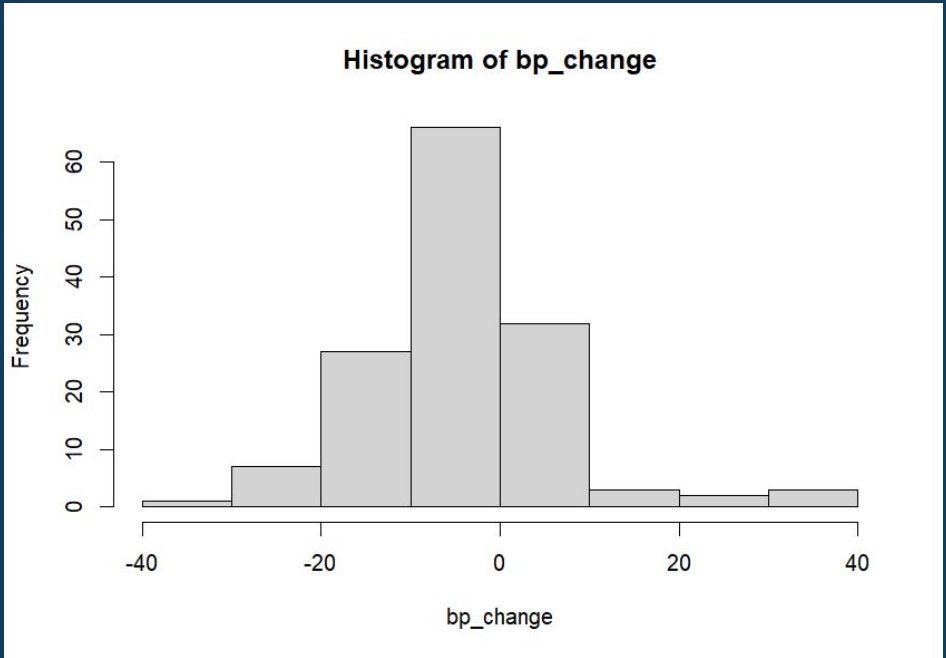
- Less than ½ of the data had a second blood pressure measurement taken.
- These measurements appear to correspond to individuals that had high BP in first measurement

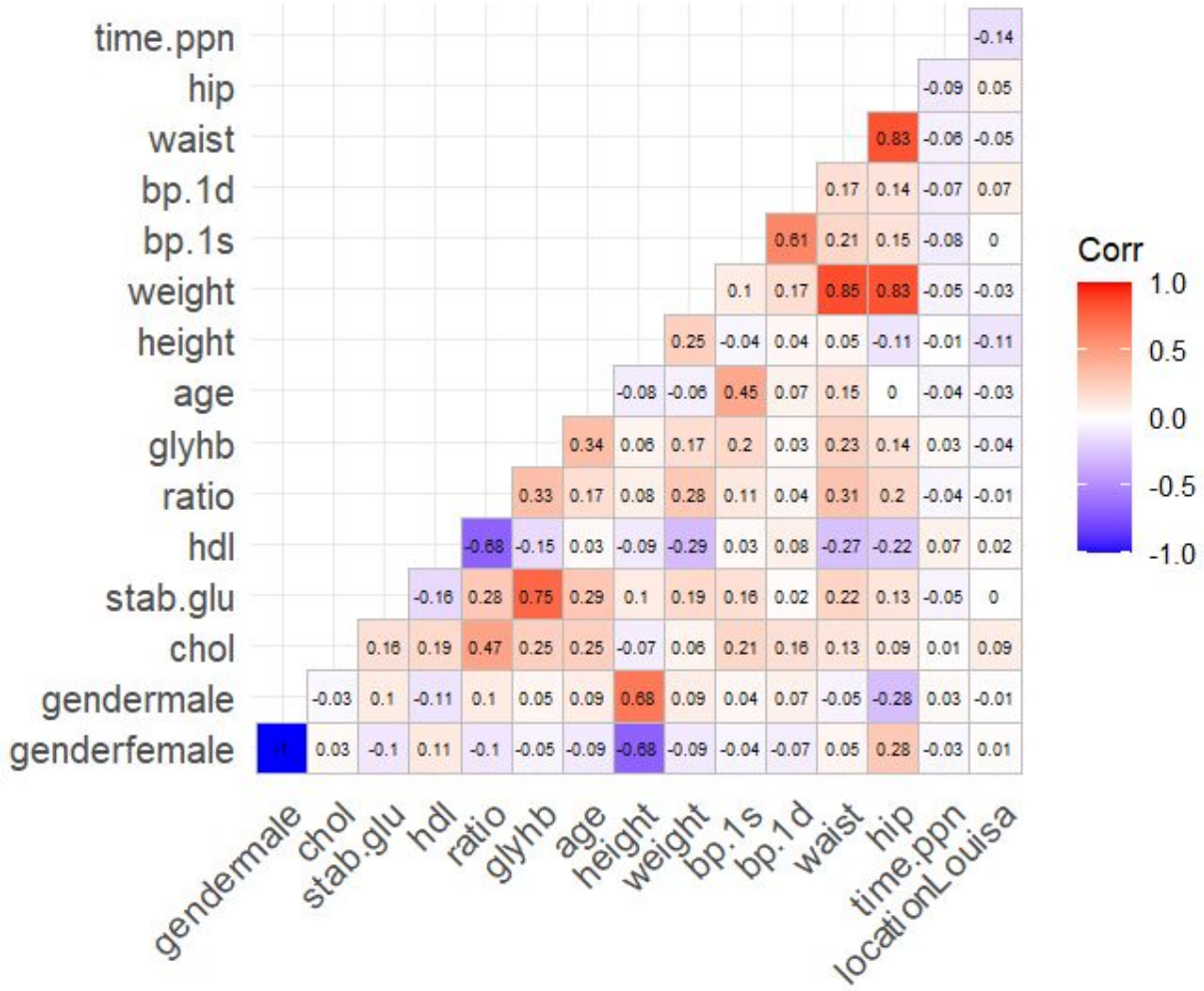




REMOVING SECOND BLOOD PRESSURE MEASURE

- Difference between first and second measurements are quite small.
- Mean & median of changes were below 3
- Not much additional information is gained, so dropping the variable seems appropriate.




time.ppn
hip
waist
bp.1d
bp.1s
weight
height
age
glyhb
ratio
hdl
stab.glu
chol
gendermale
genderfemale
-0.14
-0.09 0.05
0.83 -0.06 -0.05
0.17 0.14 -0.07 0.07
0.61 0.21 0.15 -0.08 0
0.1 0.17 0.85 0.83 -0.05 -0.03
0.25 -0.04 0.04 0.05 -0.11 -0.01 -0.11
-0.08 -0.06 0.45 0.07 0.15 0 -0.04 -0.03
0.34 0.06 0.17 0.2 0.03 0.23 0.14 0.03 -0.04
0.33 0.17 0.08 0.28 0.11 0.04 0.31 0.2 -0.04 -0.01
-0.68 -0.15 0.03 -0.09 -0.29 0.03 0.08 -0.27 -0.22 0.07 0.02
-0.16 0.28 0.75 0.29 0.1 0.19 0.16 0.02 0.22 0.13 -0.05 0
0.16 0.19 0.47 0.25 0.25 -0.07 0.06 0.21 0.16 0.13 0.09 0.01 0.09
-0.03 0.1 -0.11 0.1 0.05 0.09 0.68 0.09 0.04 0.07 -0.05 -0.28 0.03 -0.01
-1 0.03 -0.1 0.11 -0.1 -0.05 -0.09 -0.68 -0.09 -0.04 -0.07 0.05 0.28 -0.03 0.01
gendermale
chol
stab.glu
hdl
ratio
glyhb
age
height
weight
bp.1s
bp.1d
waist
hip
time.ppn
locationLouisa
Corr
1.0
0.5
0.0
-0.5
-1.0

ADDRESSING MULTICOLLINEARITY

Initial correlation matrices show that certain predictors are very heavily correlated with one another. Such as hip and waist and cholesterol and hdl. To address the high correlation of these variables, hip and waist were combined to form hip/waist ratio column. Ratio was removed from the data set, as cholesterol and hdl are already within the dataset and not heavily correlated with one another. bp.1s and bp.1d are also inherently correlated with each other, Mean arterial pressure was calculated from these columns to combine them.

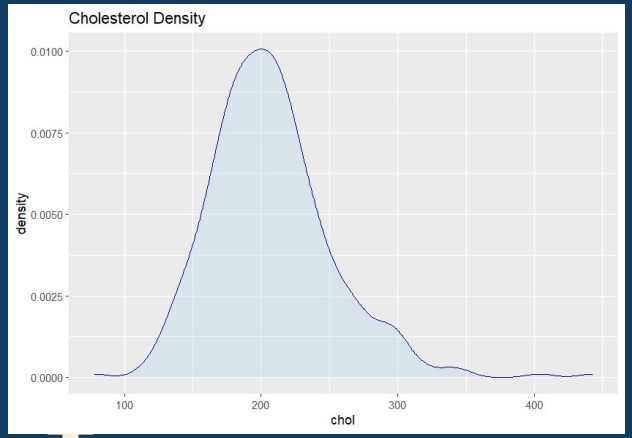
Variance inflation factor (VIF) was used to measure the initial effects of multicollinearity within the predictors and show the results of the corrections. VIF greater than 4 are known to be associated with multicollinearity.

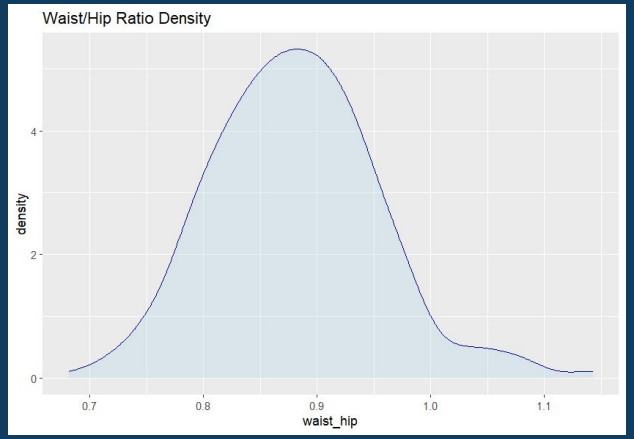
Before:

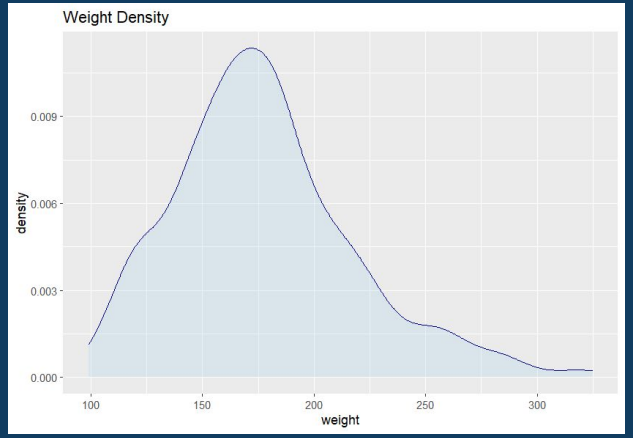
```
##   gender     chol stab.glu      hdl    ratio      age   height   weight
## 2.425217 3.920198 1.208780 5.647044 7.037430 1.714784 2.275876 7.395914
##    bp.1s    bp.1d    waist      hip time.ppn location
## 2.207348 1.817952 5.388071 6.798683 1.046834 1.111147
```

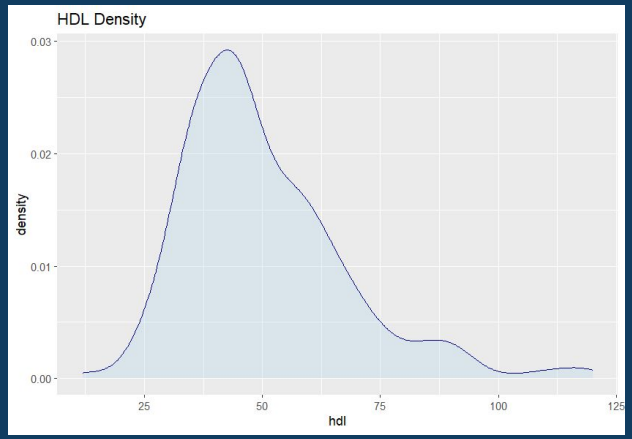
After:

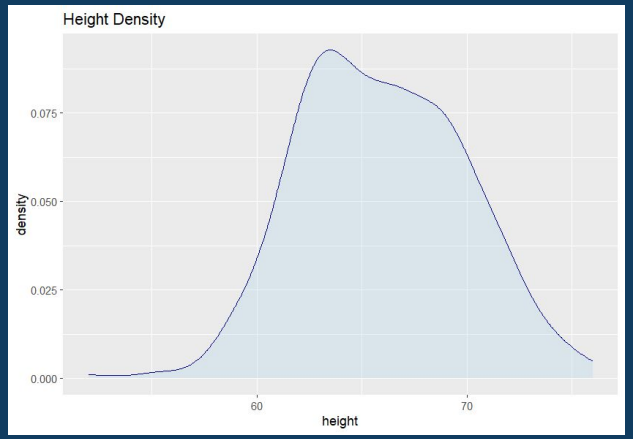
```
##              gender                chol                 hdl
##            2.124940            1.182038            1.213564
##                 age              height              weight
##            1.399840            2.125734            1.338284
##            time.ppn            location                 map
##            1.043327            1.091511            1.156062
##     waist_hip_ratio transformed_stab.glu
##            1.377605            1.223022
```


Cholesterol Density
density
0.0100
0.0075
0.0050
0.0025
0.0000
100
200
300
400
chol


Waist/Hip Ratio Density
density
4
2
0
0.7
0.8
0.9
1.0
1.1
waist_hip

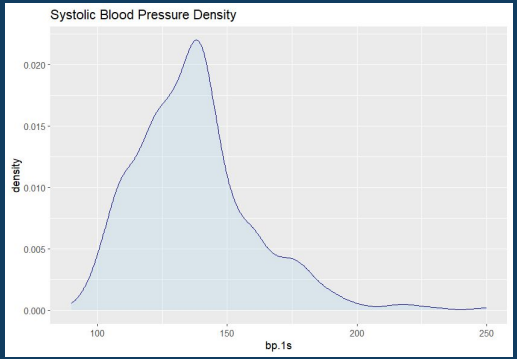

Weight Density
density
0.009
0.006
0.003
0.000
100
150
200
250
300
weight

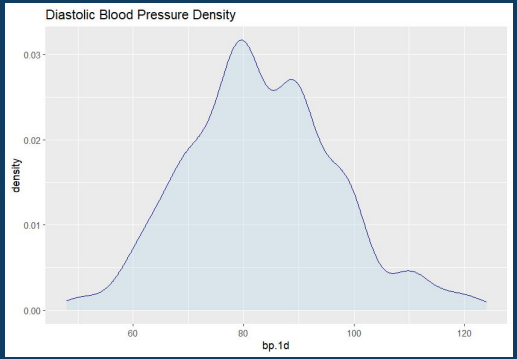

HDL Density
density
0.03
0.02
0.01
0.00
25
50
75
100
125
hdl

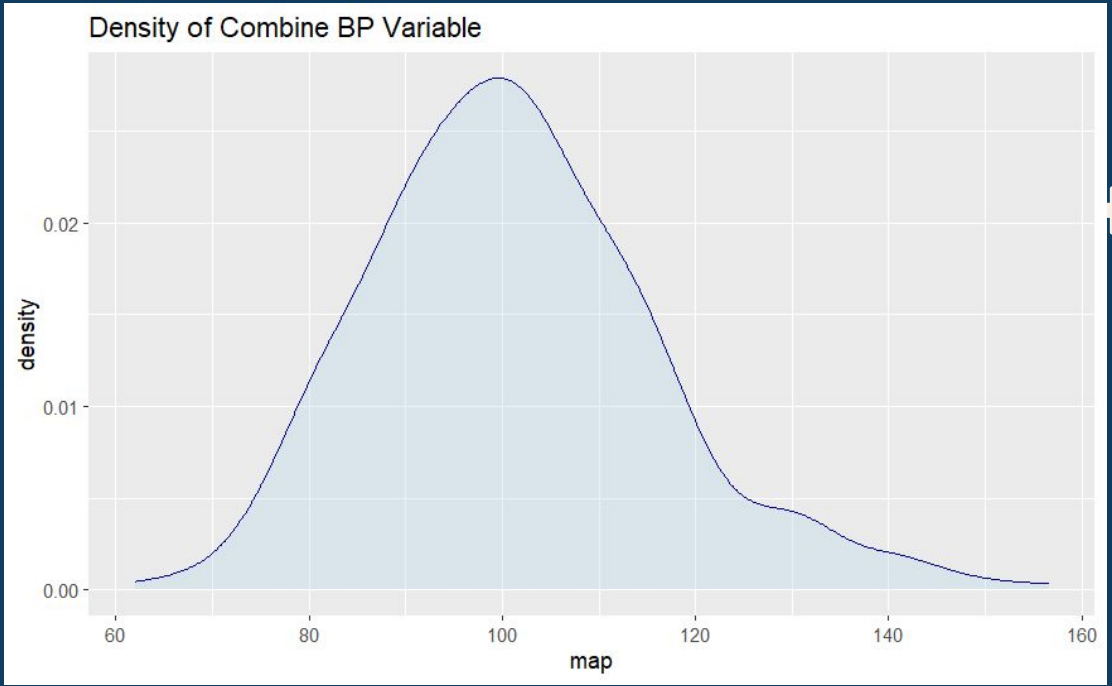

Height Density
density
0.075
0.050
0.025
0.000
60
70
height

SYSTOLIC/DIASTOLIC BP

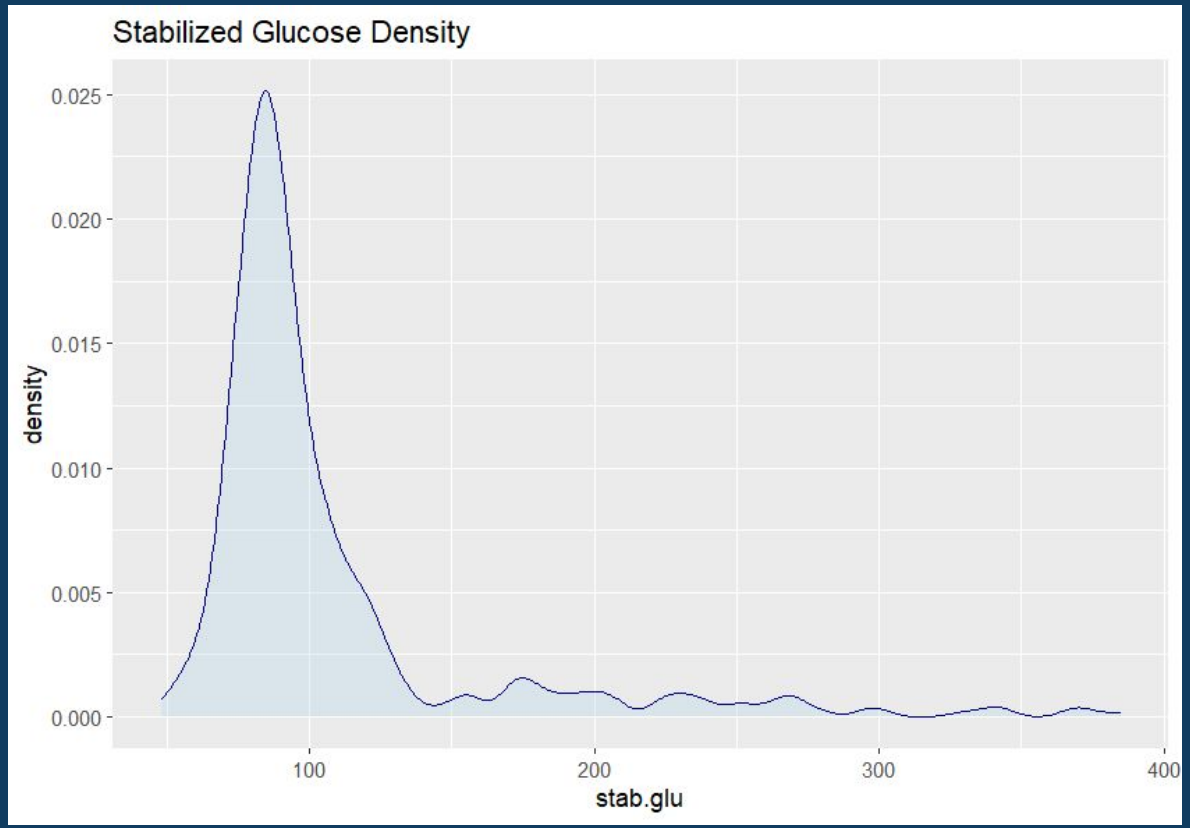
- Systolic and Diastolic Blood Pressure are correlated (.61)
- Combine into one variable (map)





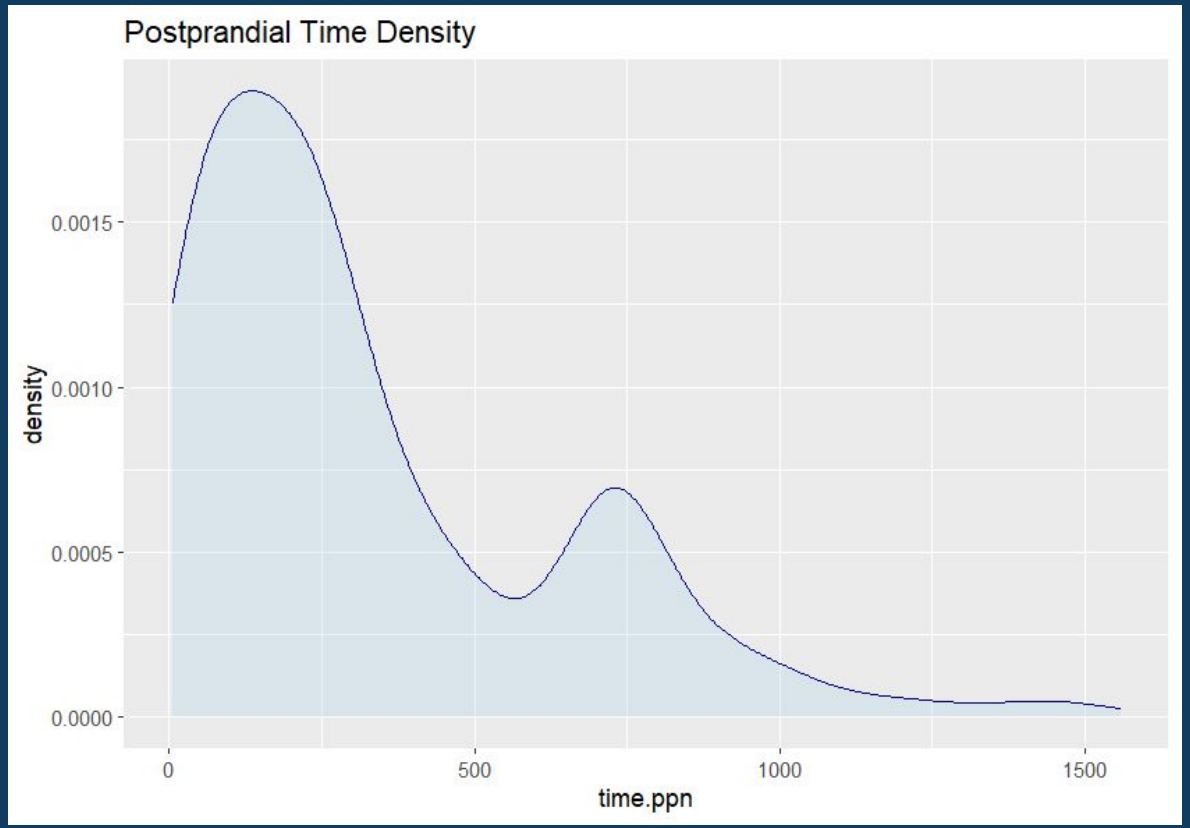


STABILIZED GLUCOSE

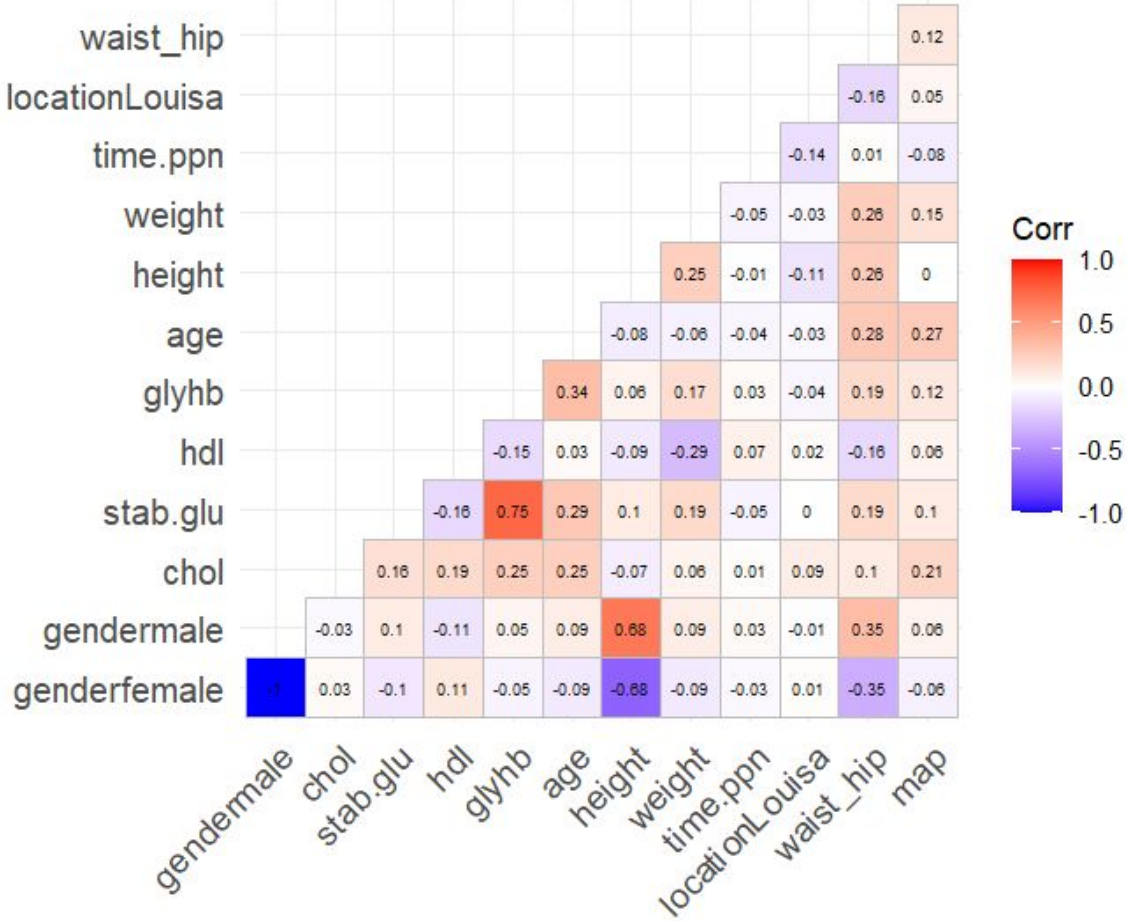


- Density is very right skewed
- Indicates we may need a transformation such as log.
- This is examined more later on

POSTPRANDIAL TIME



- Density is bimodal
- We may want to test if a higher degree polynomial is appropriate to fit this distribution
- This is tested later


waist_hip
locationLouisa
time.ppn
weight
height
age
glyhb
hdl
stab.glu
chol
gendermale
genderfemale
gendermale
chol
stab.glu
hdl
glyhb
age
height
weight
time.ppn
locationLouisa
waist_hip
map
Corr
1.0
0.5
0.0
-0.5
-1.0

HYPOTHESIS

- Would the regression model containing all 11 features be useful in predicting glycosylated hemoglobin amount?
- Which variables are the most important in predicting glycosylated hemoglobin amount? Does this vary based on Gender?
- Does the county a person lives in affect their glycosylated hemoglobin levels?
- Does the Postprandial time and gender interaction have a significant relationship to the level of glyhb blood in the population of Black Virginians with diabetes?

FULL MODEL

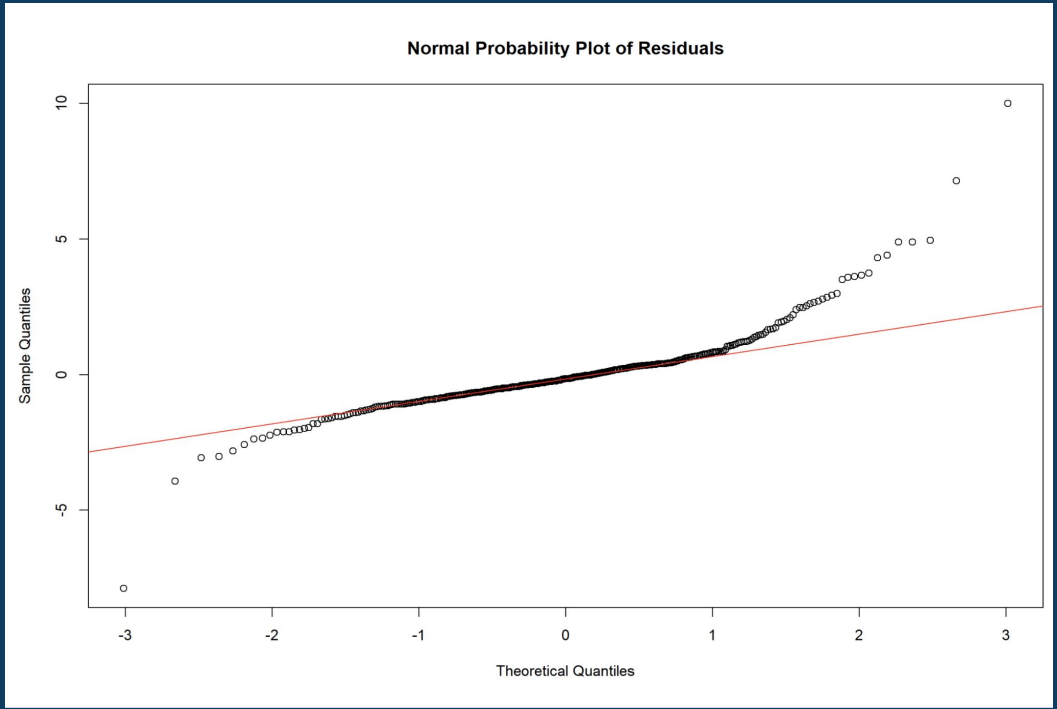
Full Model:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11}$$

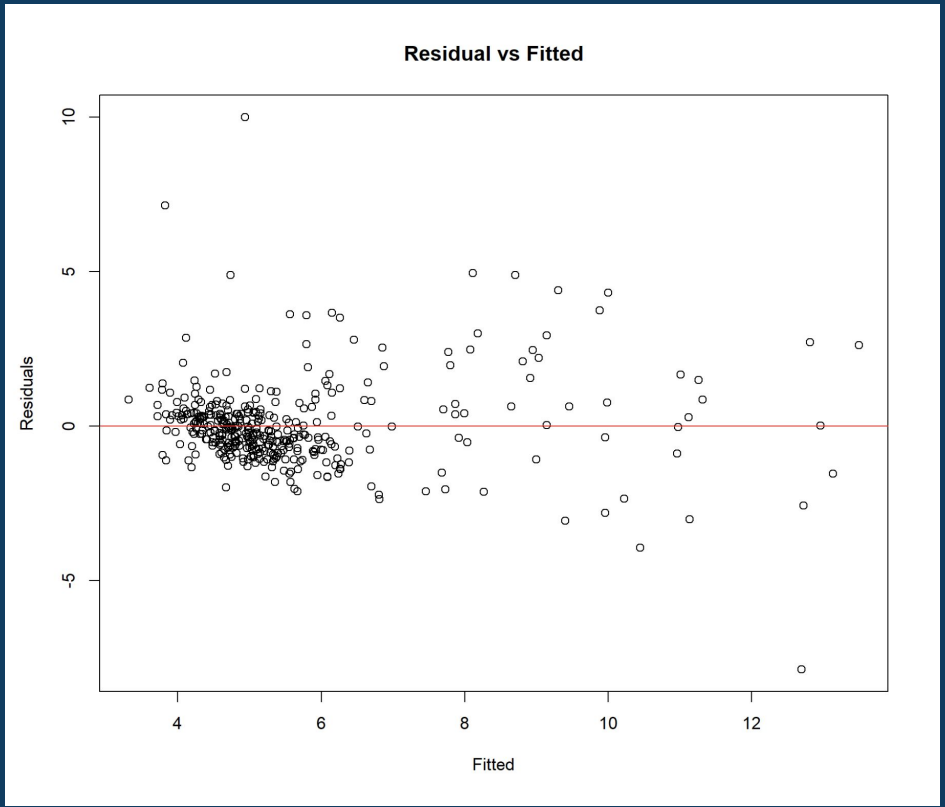
Estimate column shows b_i's

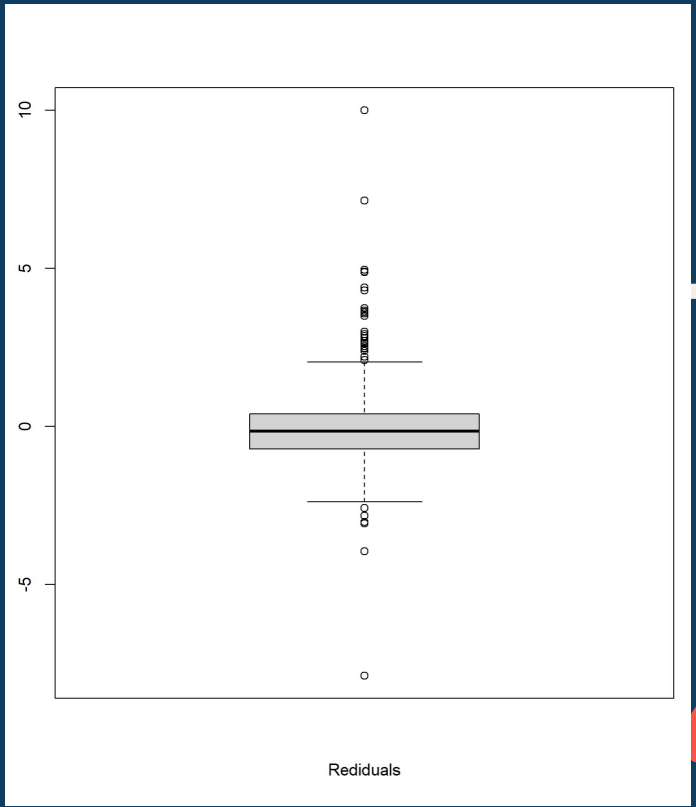
```
Call:
lm(formula = glyhb ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9119 -0.6992 -0.1512  0.4153 10.0022

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.1928277  2.0335587  -0.587 0.557844
gendermale     -0.2786126  0.2165541  -1.287 0.199037
chol            0.0062634  0.0017803   3.518 0.000488 ***
stab.glu        0.0285226  0.0014801  19.270  < 2e-16 ***
hdl            -0.0091963  0.0046520  -1.977 0.048793 *
age             0.0155364  0.0052223   2.975 0.003120 **
height          0.0245895  0.0272971   0.901 0.368267
weight          0.0008187  0.0020885   0.392 0.695272
time.ppn        0.0005572  0.0002424   2.299 0.022039 *
locationLouisa -0.1316798  0.1529897  -0.861 0.389948
waist_hip       0.2527819  1.1731665   0.215 0.829518
map             0.0016840  0.0052728   0.319 0.749623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.44 on 375 degrees of freedom
Multiple R-squared:  0.6017,    Adjusted R-squared:  0.5901
F-statistic: 51.51 on 11 and 375 DF,  p-value: < 2.2e-16
```

NORMALITY OF RESIDUALS

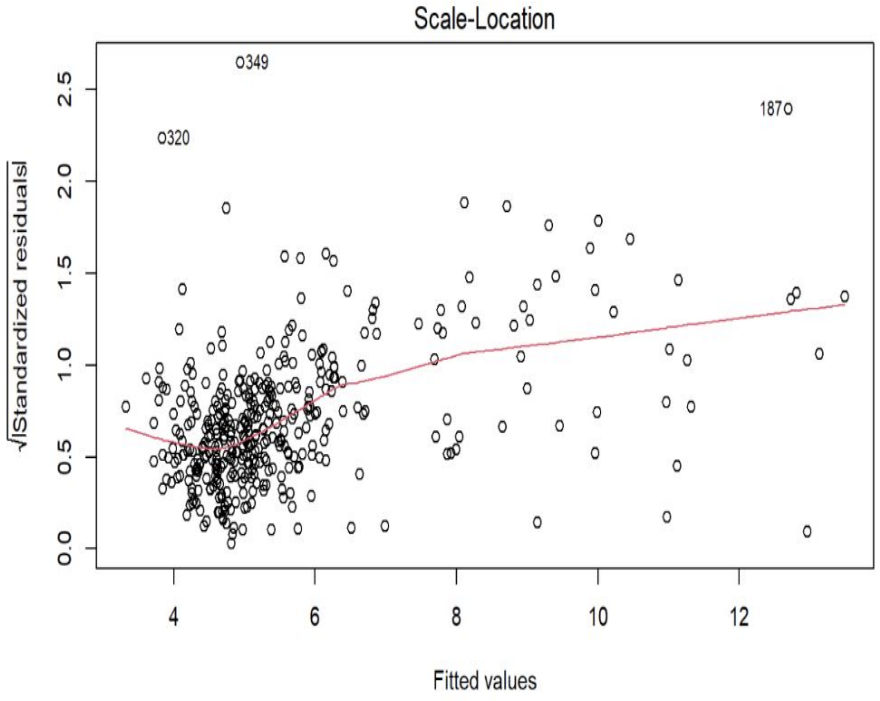


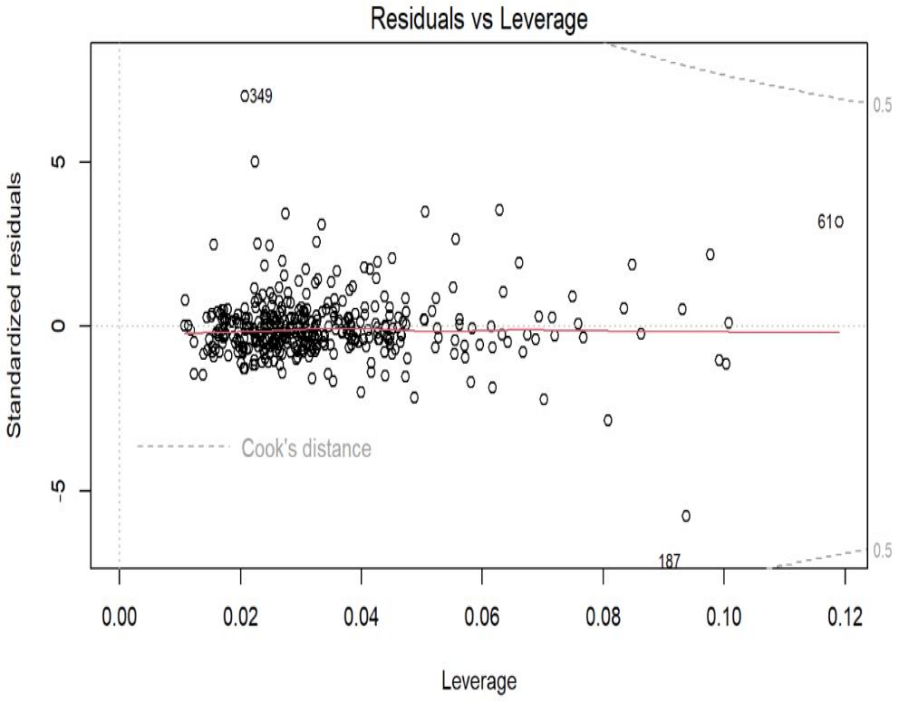
RESIDUAL V FITTED PLOT





EQUAL VAR/INFLUENTIAL





L.I.N.E ASSUMPTIONS

The original un-transformed data:

L: Using a multiple linear regression to fit the a full model for the dataset, [The mean response is a linear function of X]

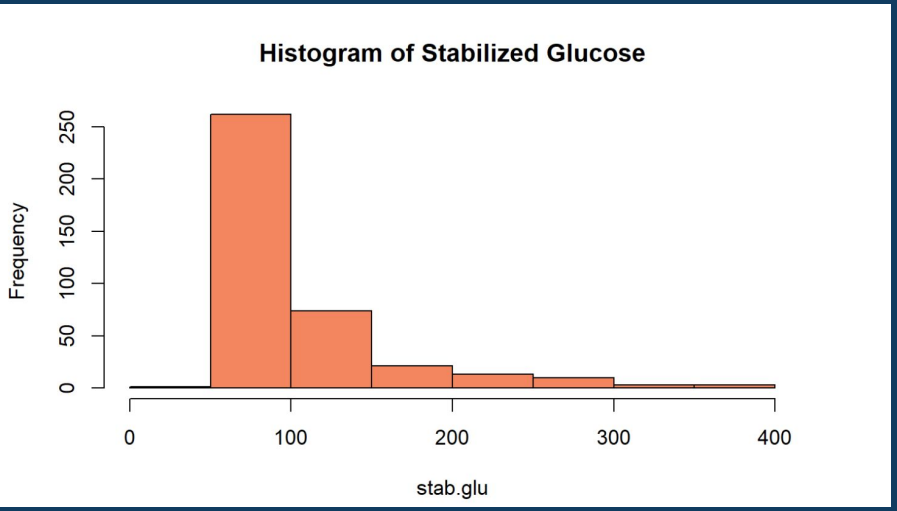
-Linear lack of fit test shows that the full model may not be the the best model to fit the data

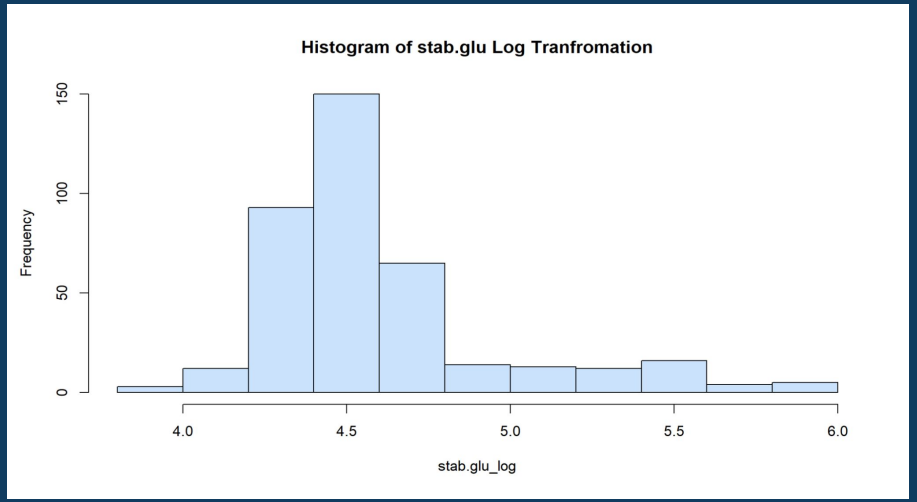
I: Durbin Watson test for randomness showed that there may be correlation within the data set

N: Shapiro-Wilkes test rejected the null that the residuals are not normally distributed

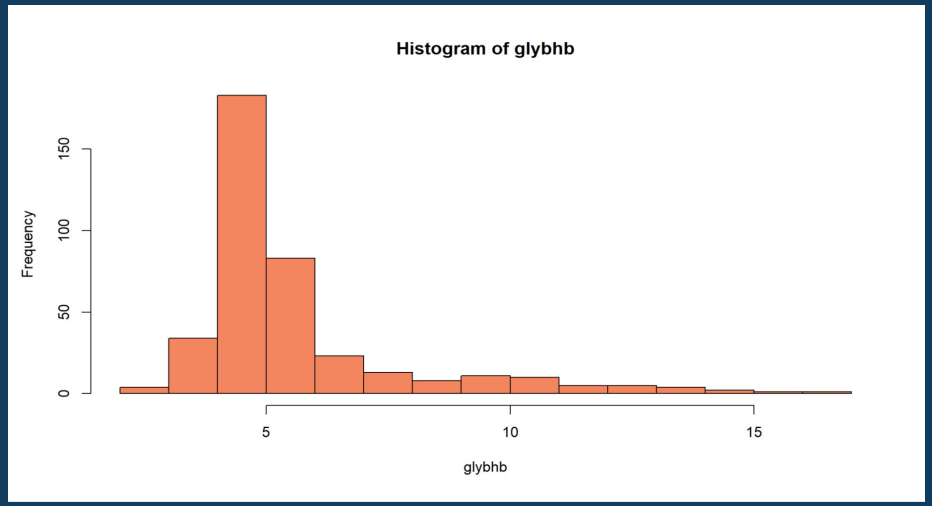
E: From the residual and scale-location plots it can be said that the equal variance assumption can be maintained

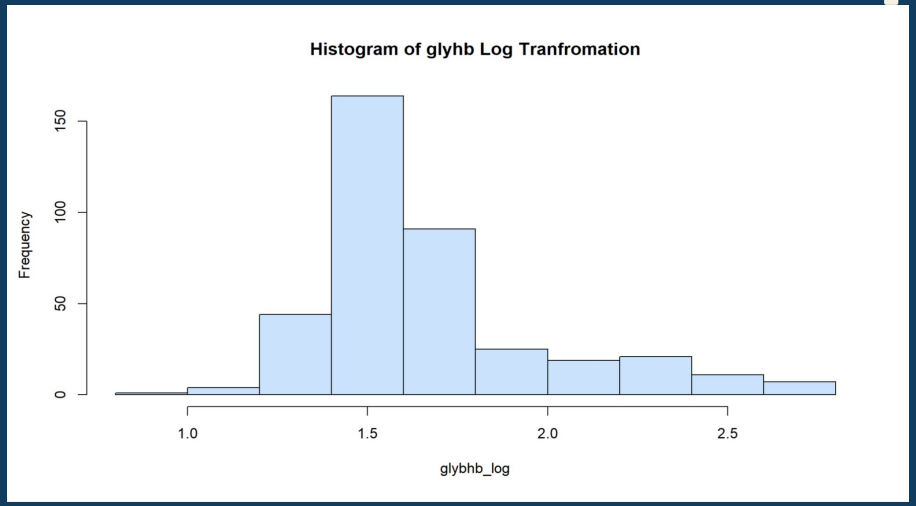
DATA TRANSFORMATIONS



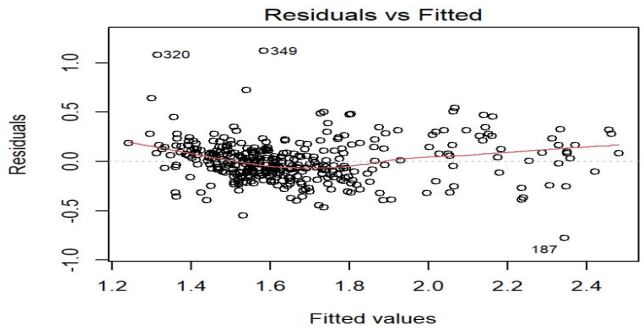


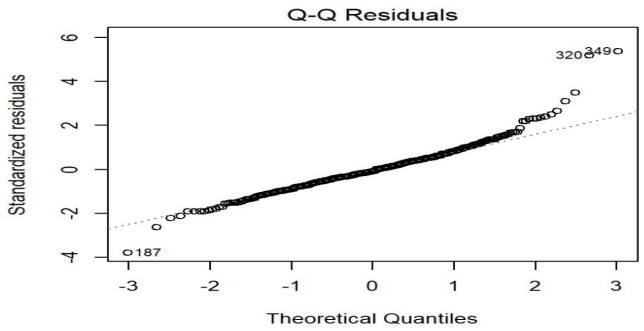
TRANSFORMATION OF DATA

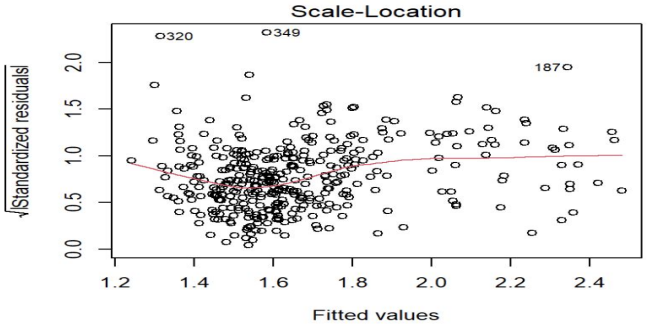


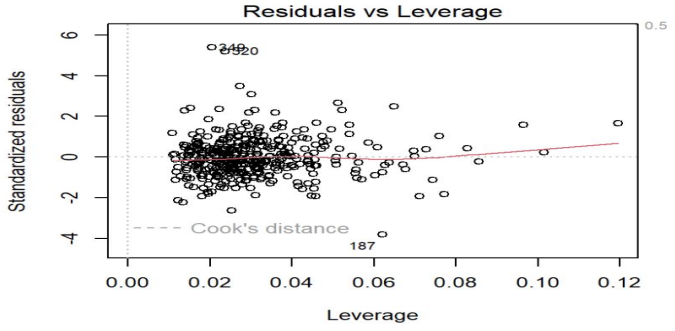


MODEL PLOTS WITH TRANSFORMATIONS



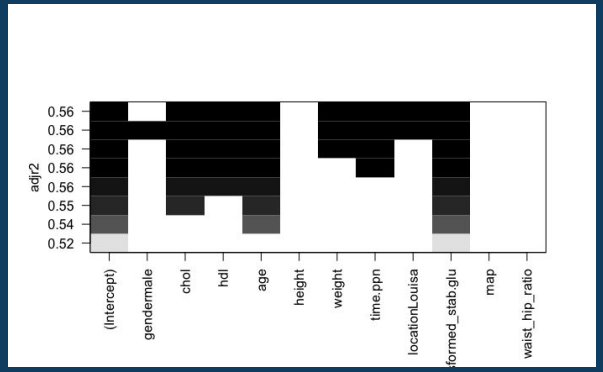


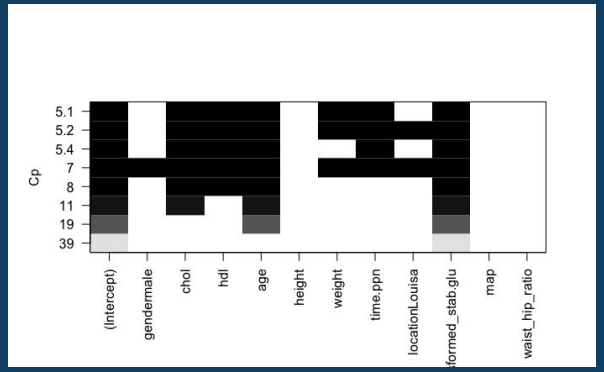


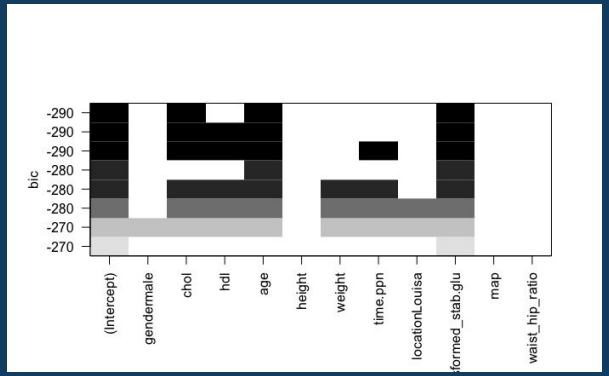


BEST SUBSET SELECTION

After fitting the model best subset selection was used to narrow down important predictors. Due to the number of predictors within the full model an exhaustive approach was possible (2^11 = 2048), and stepwise selection was not needed.







F TEST BASED ON BESTSUBSET

We will run F-test to check if this method agrees with variables selected in best subset. We will use 5% significance level.

- Ho: $\beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$
- Ha: At least one $\beta_i \neq 0$ for i = 8, 9, 10, 11

```
Analysis of Variance Table

Model 1: glyhb ~ chol + stab.glu + hdl + age + weight + time.ppn + location
Model 2: glyhb ~ chol + stab.glu + hdl + age + weight + time.ppn + location +
    gender + height + waist_hip + map
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    379 16.632
2    375 16.580  4  0.051434 0.2908 0.8839
```

P-value = .88 > .05 Fail to Reject Ho.

We have evidence that gender, height, waist-hip ratio, and combined blood pressure predictors are not significant to the regression and may be removed from the model.

LOCATION COMPARISON

We specifically considered the significance of location to the model.

There were two possible values for location:

- Louisa County, Virginia
- Buckingham County, Virginia

While these counties are geographically very similar, they have key lifestyle differences. As previously discussed, a person's lifestyle can directly impact diabetes. We aim to use location as a proxy variable for lifestyle differences.

- Louisa county is more urban due to its proximity to major cities like Richmond, Virginia while Buckingham county remains rural.
- Louisa has about 37,600 residents while Buckingham has less than half of that at 16,800.
- The economy in Louisa county benefits from tourism since the county is closer to Lake Anna while Buckingham focuses more on mining and other historical practices.
- Education in Buckingham county is less extensive than Louisa county.

LOCATION COMPARISON

We considered the possibility that these differences impacted a person's lifestyle enough to influence whether or not they have diabetes.

- Healthy eating decreases the risk of developing diabetes.
  - The urban nature of Louisa county may mean they offer more food options, but not necessarily more healthy options as they also have more tourism.
  - Buckingham county's dependence on agriculture may mean more healthy options available and more physical activity.
  - However, we also considered that the lack of educational opportunities in Buckingham could mean less educated decisions on healthier lifestyles and diabetes in general.

GENERAL LINEAR TEST - LOCATION

When considering location, it was added back into the previously determined significant variables for a general linear test.

Ho: $\beta_{location} = 0$

Ha: $\beta_{location} \neq 0$

The full model includes location while the reduced model excludes the location from the model.

DF_F: n - 8 = 382

DF_R: n - 7 = 383

```
> #Critical Value for the F-Test
> qf(1-0.05,1,382)
[1] 3.865917
```

```
> full.model <- lm(glyhb ~ chol + stab.glu + hdl+ age + weight + time.ppn + location, data = transformed)
> anova(full.model)
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq  F value    Pr(>F)
chol        1  2.4585  2.4585  56.1648 4.685e-13 ***
stab.glu    1 18.3694 18.3694 419.6490 < 2.2e-16 ***
hdl         1  0.1975  0.1975   4.5120   0.03430 *
age         1  0.7281  0.7281  16.6336 5.521e-05 ***
weight      1  0.0883  0.0883   2.0179   0.15627
time.ppn    1  0.2157  0.2157   4.9275   0.02702 *
location    1  0.0723  0.0723   1.6508   0.19963
Residuals 382 16.7213  0.0438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> reduced.model <- lm(glyhb ~ chol + stab.glu + hdl+ age + weight + time.ppn, data = transformed)
> anova(reduced.model)
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq  F value    Pr(>F)
chol        1  2.4585  2.4585  56.0696 4.864e-13 ***
stab.glu    1 18.3694 18.3694 418.9372 < 2.2e-16 ***
hdl         1  0.1975  0.1975   4.5043   0.03445 *
age         1  0.7281  0.7281  16.6053 5.597e-05 ***
weight      1  0.0883  0.0883   2.0145   0.15662
time.ppn    1  0.2157  0.2157   4.9191   0.02715 *
Residuals 383 16.7936  0.0438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GENERAL LINEAR TEST - LOCATION

The critical value used to compare the F-statistic to = 3.865917

If the F >= critical value, we reject Ho.

$$F = \left(\frac{SSE(R) - SSE(F)}{DF_R - DF_F} \right) / \left(\frac{SSE(F)}{DF_F} \right)$$

$$F = \left(\frac{16.7936 - 17.7213}{382 - 382} \right) / \left(\frac{16.7213}{382} \right)$$

F = 1.651702

Since F = 1.651702 < 3.865917, we fail to reject the null hypothesis and conclude that location is not significant to the model.

```
> full.model <- lm(glyhb ~ chol + stab.glu + hdl+ age + weight + time.ppn + location, data = transformed)
> anova(full.model)
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq  F value    Pr(>F)
chol        1  2.4585  2.4585  56.1648 4.685e-13 ***
stab.glu    1 18.3694 18.3694 419.6490 < 2.2e-16 ***
hdl         1  0.1975  0.1975   4.5120   0.03430 *
age         1  0.7281  0.7281  16.6336 5.521e-05 ***
weight      1  0.0883  0.0883   2.0179   0.15627
time.ppn    1  0.2157  0.2157   4.9275   0.02702 *
location    1  0.0723  0.0723   1.6508   0.19963
Residuals 382 16.7213  0.0438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> reduced.model <- lm(glyhb ~ chol + stab.glu + hdl+ age + weight + time.ppn, data = transformed)
> anova(reduced.model)
Analysis of Variance Table

Response: glyhb
           Df  Sum Sq Mean Sq  F value    Pr(>F)
chol        1  2.4585  2.4585  56.0696 4.864e-13 ***
stab.glu    1 18.3694 18.3694 418.9372 < 2.2e-16 ***
hdl         1  0.1975  0.1975   4.5043   0.03445 *
age         1  0.7281  0.7281  16.6053 5.597e-05 ***
weight      1  0.0883  0.0883   2.0145   0.15662
time.ppn    1  0.2157  0.2157   4.9191   0.02715 *
Residuals 383 16.7936  0.0438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #Critical Value for the F-Test
> qf(1-0.05,1,382)
[1] 3.865917
```

LOCATION COMPARISON CONCLUSION

Despite the differences in education, area, and economy, location does not significantly impact the model.

EXAMINING MODELS FROM BEST SUBSET

In examining best subset methodology, we saw that BIC indicated that only cholesterol, age, and stabilized glucose variables were in the ideal model. We will run an F test on this reduced model to see if the weight, time.ppn, and hdl variables may be dropped.

- Run test at 5% significance level.
- Ho: $\beta_{hdl} = \beta_{weight} = \beta_{time.ppn} = 0$
- Ha: At least one of β_{hdl} , β_{weight} , $\beta_{time.ppn} \neq 0$

```
Analysis of Variance Table

Model 1: glyhb ~ chol + stab.glu + hdl + age + weight + time.ppn
Model 2: glyhb ~ chol + stab.glu + age
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    380 16.717
2    383 17.263 -3  -0.54543 4.1327 0.006679 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- P-value = .006679 < .05
- Reject Ho
- We do have evidence to conclude that hdl, weight, and time.ppn should not be dropped from the model, given that chol, stab.glu, and age are retained.

POSTPRANDIAL TIME - POLYNOMIAL FIT

- First create a model with second degree polynomial and test if significant
- If F test shows, significance continue to third degree polynomial

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + age + weight + poly(time.ppn,
    2), data = scaled)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80294 -0.12784 -0.02628  0.10841  1.10874

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.1999028  0.1529534  -7.845 4.46e-14 ***
chol                0.0008576  0.0002553   3.359 0.000862 ***
stab.glu            0.5511848  0.0326607  16.876  < 2e-16 ***
hdl                -0.0012132  0.0006762  -1.794 0.073611 .
age                 0.0029923  0.0007089   4.221 3.05e-05 ***
weight              0.0004179  0.0002843   1.470 0.142469
poly(time.ppn, 2)1  0.4650770  0.2106699   2.208 0.027870 *
poly(time.ppn, 2)2 -0.1876883  0.2137757  -0.878 0.380517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2098 on 379 degrees of freedom
Multiple R-squared:  0.5689,    Adjusted R-squared:  0.5609
F-statistic: 71.44 on 7 and 379 DF,  p-value: < 2.2e-16
```

POSTPRANDIAL TIME - POLYNOMIAL FIT

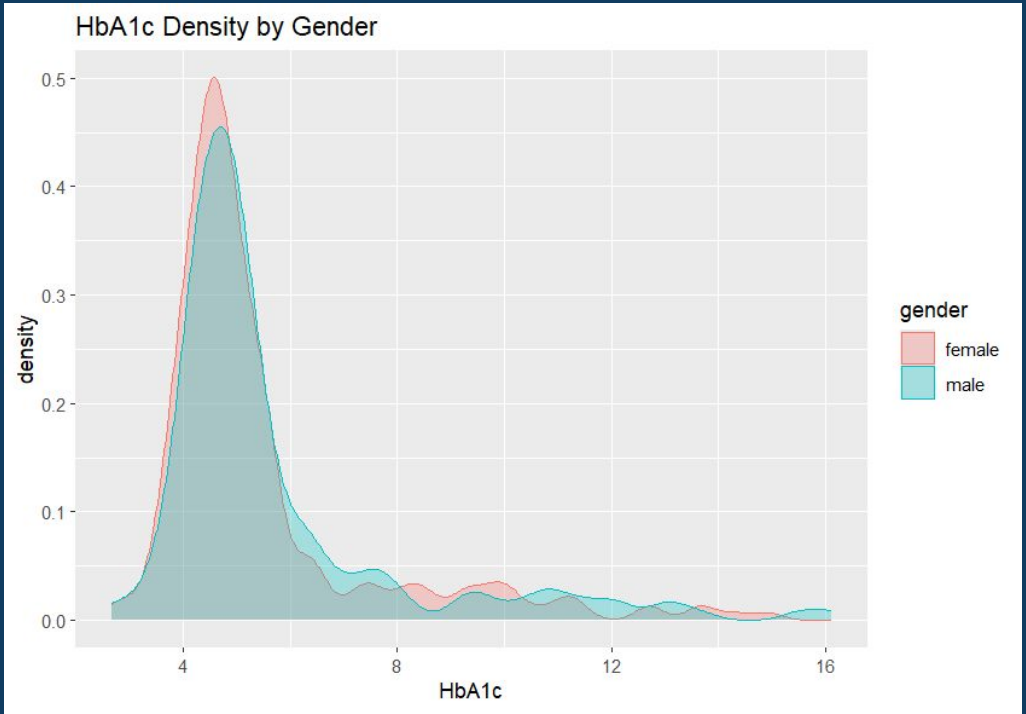
- Run test at 5% significance level.
- Ho: $\beta_{time.ppn\text{^}2} = 0$
- Ha: $\beta_{time.ppn\text{^}2} \neq 0$

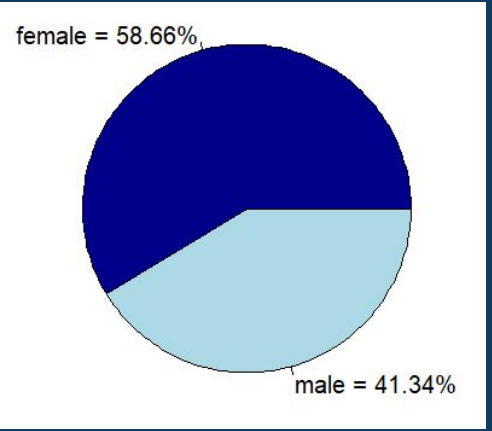
```
Analysis of Variance Table

Model 1: glyhb ~ chol + stab.glu + hdl + age + weight + time.ppn
Model 2: glyhb ~ chol + stab.glu + hdl + age + weight + poly(time.ppn,
    2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    380 16.717
2    379 16.683  1  0.033932 0.7708 0.3805
```

- P-value = .3805 > .05
- Fail to Reject Ho
- We do not have evidence to conclude that a higher degree predictor for postprandial time is significant to the regression.

EXAMINING DIFFERENCES BASED ON GENDER

- At first glance, gender does not appear to significantly impact HbA1c levels





EXAMINING DIFFERENCES BASED ON GENDER

- Could the impact of gender change based on the level of another predictor?
- We found different studies come to different conclusions on whether impact of postprandial time differs by gender
- Look into interaction with gender and Postprandial Time

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + age + weight + gender *
    time.ppn, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1333 -0.7269 -0.1723  0.4671  9.9402

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.4933403  0.5595135   0.882 0.378483
chol                 0.0062873  0.0017529   3.587 0.000379 ***
stab.glu             0.0285894  0.0014698  19.451  < 2e-16 ***
hdl                 -0.0092990  0.0045938  -2.024 0.043646 *
age                  0.0148342  0.0048433   3.063 0.002349 **
weight               0.0015446  0.0019365   0.798 0.425581
gendermale           0.1056497  0.2247256   0.470 0.638535
time.ppn             0.0008550  0.0003065   2.789 0.005550 **
gendermale:time.ppn -0.0007042  0.0004884  -1.442 0.150162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.434 on 378 degrees of freedom
Multiple R-squared:  0.6018,    Adjusted R-squared:  0.5933
F-statistic:  71.4 on 8 and 378 DF,  p-value: < 2.2e-16
```

EXAMINING DIFFERENCES BASED ON GENDER

- Run F test at 5% significance level.
- Ho: $\beta_{gender*time.ppn} = 0$
- Ha: $\beta_{gender*time.ppn} \neq 0$

```
Analysis of Variance Table

Model 1: glyhb ~ chol + stab.glu + hdl + age + weight + gender * time.ppn
Model 2: glyhb ~ chol + stab.glu + hdl + age + weight + gender + time.ppn
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    378 777.05
2    379 781.32 -1   -4.2739 2.079 0.1502
```

- P-value = .1491 > .05
- Fail to Reject Ho
- We do not have evidence to conclude that the interaction of gender and postprandial time is significant to the regression.

SIGNIFICANCE OF WEIGHT VARIABLE

- Based on investigation in previous slides, our current model includes cholesterol, stabilized glucose (log transformed), HDL, age, and postprandial time

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + age + weight + time.ppn,
    data = transformed)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8035 -0.1245 -0.0214  0.1103  1.1019

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.208e+00  1.523e-01  -7.935 2.38e-14 ***
chol         8.582e-04  2.553e-04   3.362 0.000852 ***
stab.glu     5.476e-01  3.240e-02  16.902  < 2e-16 ***
hdl         -1.297e-03  6.692e-04  -1.939 0.053258 .
age          3.010e-03  7.085e-04   4.248 2.72e-05 ***
weight       4.304e-04  2.839e-04   1.516 0.130313
time.ppn     7.688e-05  3.470e-05   2.216 0.027310 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2097 on 380 degrees of freedom
Multiple R-squared:  0.568,     Adjusted R-squared:  0.5612
F-statistic: 83.27 on 6 and 380 DF,  p-value: < 2.2e-16
```

- We see that weight variable has high p-value, this indicates that a t-test would fail to reject the null hypothesis and we do not have evidence that the variable is significant to the regression.

CURRENT MODEL

Full Model: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11}$

Current Model: $\log(\hat{Y}) = b_0 + b_2X_2 + b_3\log(X_3) + b_4X_4 + b_5X_5 + b_8X_8$

Estimate column shows b_i's

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.9119 -0.6992 -0.1512  0.4153 10.0022

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.1928277  2.0335587  -0.587 0.557844
gendermale     -0.2786126  0.2165541  -1.287 0.199037
chol            0.0062634  0.0017803   3.518 0.000488 ***
stab.glu        0.0285226  0.0014801  19.270  < 2e-16 ***
hdl            -0.0091963  0.0046520  -1.977 0.048793 *
age             0.0155364  0.0052223   2.975 0.003120 **
height          0.0245895  0.0272971   0.901 0.368267
weight          0.0008187  0.0020885   0.392 0.695272
time.ppn        0.0005572  0.0002424   2.299 0.022039 *
locationLouisa -0.1316798  0.1529897  -0.861 0.389948
waist_hip       0.2527819  1.1731665   0.215 0.829518
map             0.0016840  0.0052728   0.319 0.749623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.44 on 375 degrees of freedom
Multiple R-squared:  0.6017,    Adjusted R-squared:  0.5901
F-statistic: 51.51 on 11 and 375 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + age + time.ppn,
    data = transformed)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81908 -0.12591 -0.01017  0.10962  1.08551

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.161e+00  1.493e-01  -7.777 7.04e-14 ***
chol         9.049e-04  2.538e-04   3.565  0.00041 ***
stab.glu     5.565e-01  3.192e-02  17.434  < 2e-16 ***
hdl         -1.575e-03  6.447e-04  -2.444  0.01498 *
age          2.860e-03  7.028e-04   4.070 5.72e-05 ***
time.ppn     7.498e-05  3.474e-05   2.159  0.03151 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2101 on 381 degrees of freedom
Multiple R-squared:  0.5654,    Adjusted R-squared:  0.5597
F-statistic: 99.13 on 5 and 381 DF,  p-value: < 2.2e-16
```

PREDICTOR CONFIDENCE INTERVALS

- Below are the confidence intervals for the predictors kept in our final model

	2.5 %	97.5 %
(Intercept)	-1.454405e+00	-0.8673696747
chol	4.057638e-04	0.0014039391
stab.glu	4.937598e-01	0.6192904264
hdl	-2.843066e-03	-0.0003079229
age	1.478259e-03	0.0042417497
time.ppn	6.682373e-06	0.0001432836

- We see that 0 is not included within any of these confidence intervals, so all are significant

OUTLIERS AND INFLUENTIAL POINTS

Outliers and influential points were analyzed using studentized residuals, leverage values, BFFITS, BFBETAS, and Cook's distance.

$$e_i^* = \frac{e_i}{SE_{(e_i)}} = \frac{e_i}{\sqrt{MSE_{(1-h_{ii})}}}.$$

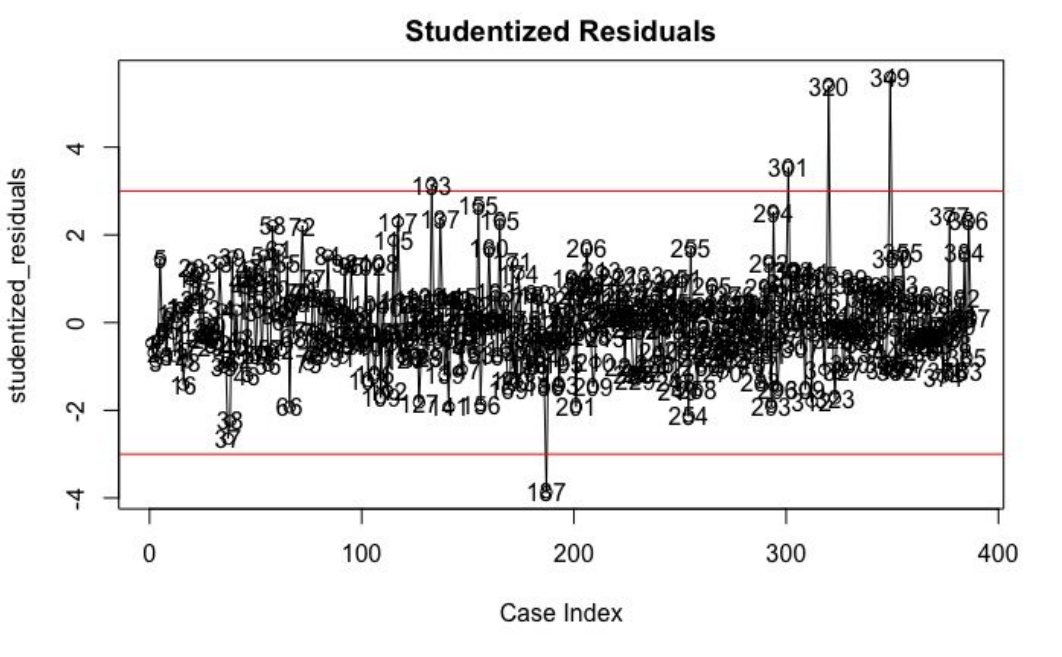
$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{p \times MSE} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]$$

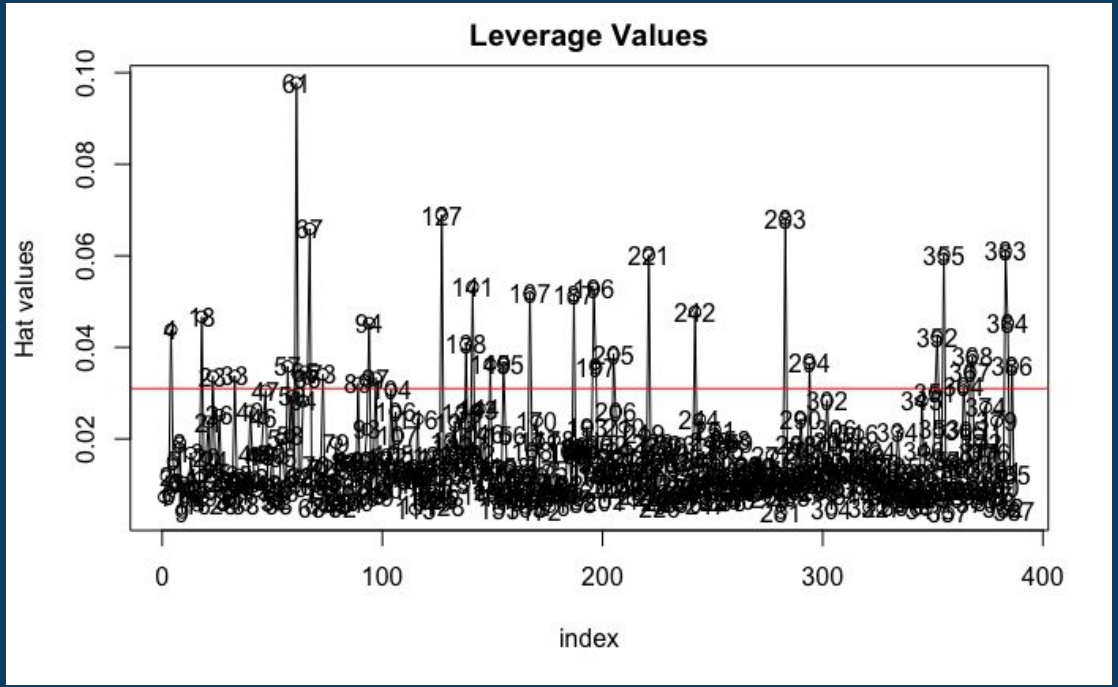
$$h_{ii} = X_i^T (X^T X)^{-1} X_i$$

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{MSE_{(i)} C_{kk}}, k = 0, 1, 2, \cdots, p-1, C_{kk} = diag((X'X)^{-1})$$

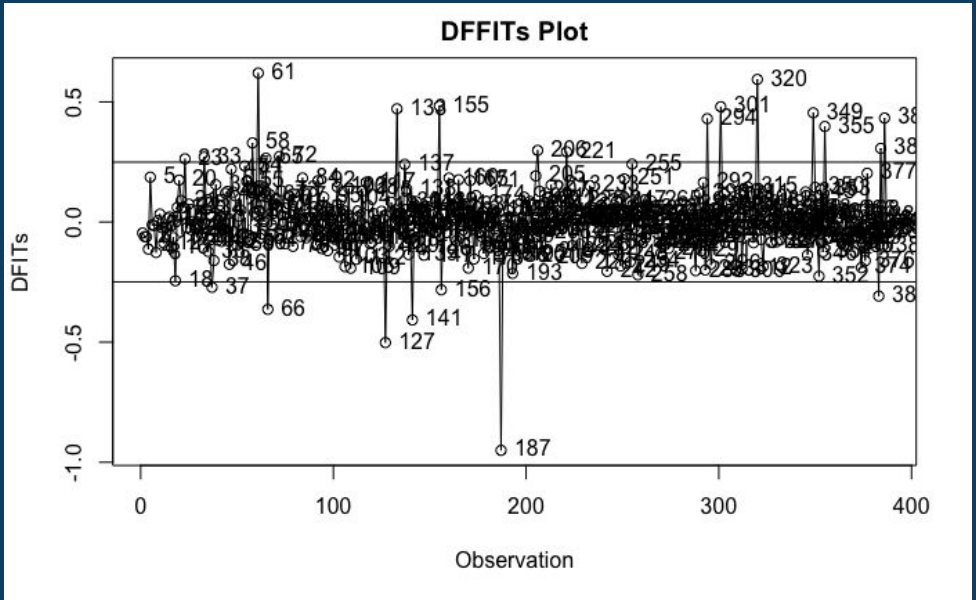
OUTLIERS AND INFLUENTIAL POINTS



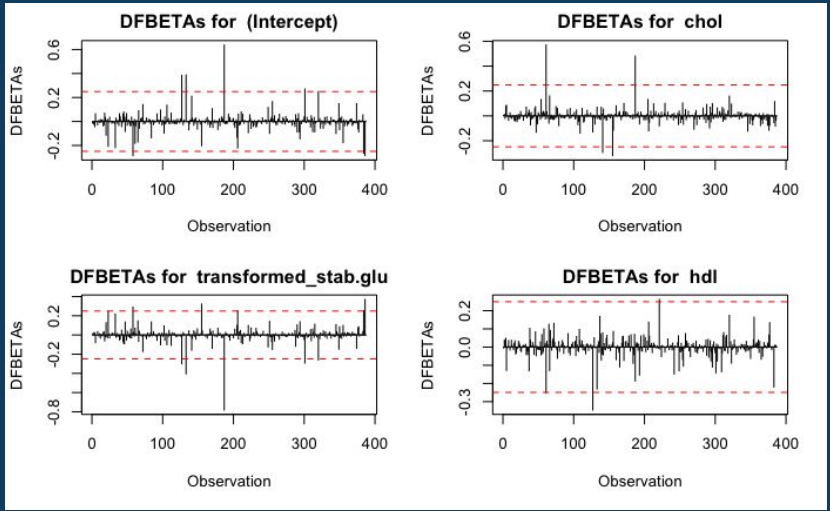
OUTLIERS AND INFLUENTIAL POINTS

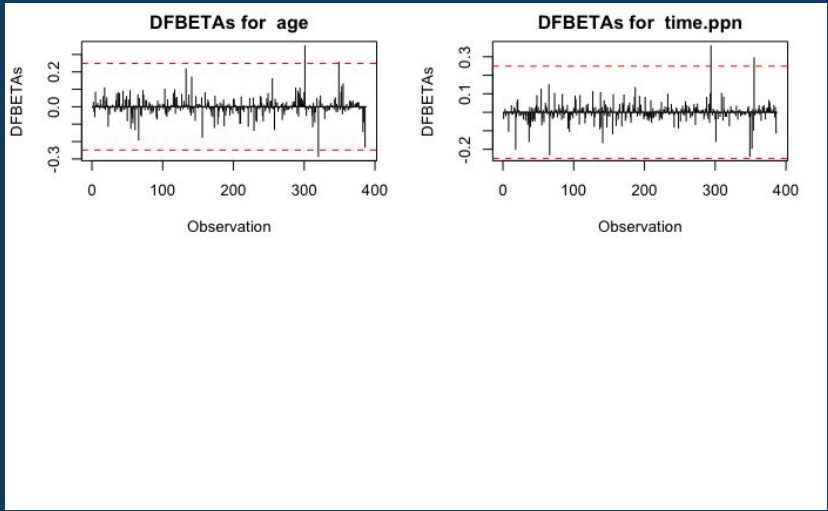


OUTLIERS AND INFLUENTIAL POINTS



OUTLIERS AND INFLUENTIAL POINTS



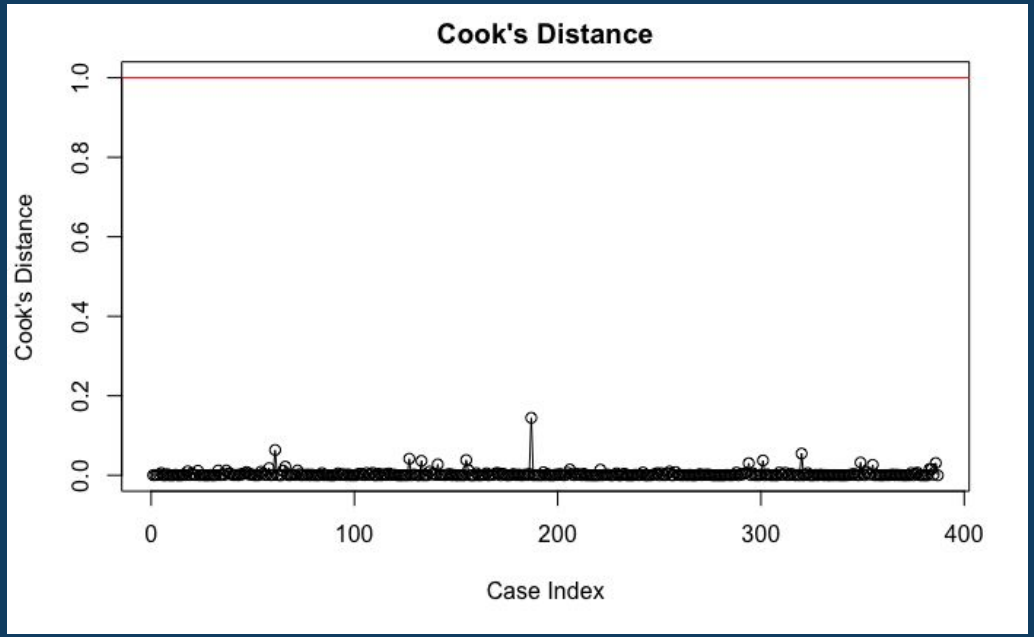


OUTLIERS AND INFLUENTIAL POINTS

Based on the plots and results points 61, 187, 320, and 349 are worthy of further investigation:

	gender	chol	hdl	age	height	weight	time.ppn	location	map	waist_hip_ratio	transformed_glyhb	transformed_stab.glu
1	female	443	23	51	70	235	420	Buckingham	118.00000	0.8958333	2.660959	5.220356
2	male	143	46	68	67	158	90	Louisa	100.66667	0.8604651	1.570697	5.916202
3	male	220	66	26	70	150	300	Buckingham	104.00000	0.8461538	2.395164	4.094345
4	female	203	51	60	59	123	60	Louisa	91.33333	0.8780488	2.704042	4.499810

OUTLIERS AND INFLUENTIAL POINTS



FURTHER CONSIDERATIONS - CHOL AND HDL

```
Analysis of Variance Table

Model 1: glyhb ~ chol + stab.glu + hdl + age + time.ppn
Model 2: glyhb ~ chol + stab.glu + hdl + age + time.ppn + hdl * chol
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    381 16.819
2    380 16.528  1   0.29016 6.6709 0.01017 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- HDL is considered ‘good’ cholesterol.
- It would make sense that the interaction between these two variables might be of some significance.
- Issues with correlation between ratio and hdl, we opted to test an interaction term instead.

CONCLUSIONS

- Which variables are the most important in predicting glycosylated hemoglobin amount?
  - Stab.glu is most significant. We see other variables chol, hdl, age, and time.ppn also appear significant

- Does the county a person lives in affect their glycosylated hemoglobin levels?
  - No significant effect could be concluded.

- Does the Postprandial time and gender interaction have a significant relationship to the level of glyhb blood in the population of Black Virginians with diabetes?
  - No significant difference / interaction could be concluded.

QUESTIONS?

CITATIONS

- Introduction CDC Information
  - https://www.cdc.gov/diabetes/prevention-type-2/prediabetes-prevent-type-2.html
  - https://www.cdc.gov/pcd/issues/2024/23_0189.htm
- Gender differences in fasting and postprandial metabolic traits
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC9546939/
- Diabetes as risk factor for incident coronary heart disease
  - https://pubmed.ncbi.nlm.nih.gov/24859435/
- Absence of a sexual dimorphism in postprandial glucose metabolism
  - https://www.nature.com/articles/s41387-022-00184-5
- Diabetes and Blood Pressure
  - https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

CITATIONS

- Buckingham County Info
  - https://www.buckinghamcountyva.org/community/about.php
- Buckingham Mapcarta
  - https://mapcarta.com/21844992
- Louisa Mapcarta
  - https://mapcarta.com/21887212
- Lake Anna Info
  - https://www.louisacounty.gov/868/Lake-Anna
- Louisa County Education
  - https://schoolquality.virginia.gov/divisions/louisa-county-public-schools
- Buckingham Education
  - https://schoolquality.virginia.gov/divisions/buckingham-county-public-schools