Lab 2 - NHANES A1C Descriptive Analysis

Github Repository: https://github.com/fardaevm/Diabetes-NHANES

Mukhammadali Fardaev, Umair Habib, Vishnu Gorur, Samuel Dominguez

April 14, 2025

1. Introduction

2. Data Loading

3. Data Splitting

4. Model Specification

coeftest(model_2, vcov. = vcovHC(model_2, type = "HC1"))

4.1 Modeling

In order to examine the relationship between A1C and different demographic and health indicators, we developed two linear regression on our exploratory dataset (30% of the sample: 603 observations). After, we applied the final features to the confirmatory dataset (70% of the sample, 1402 observations) to guarantee consistent description patterns.

Initial exploratory analysis of the important variables, we found that the raw waist-to-height ratio points are moderately normal while the A1C values were positively skewed (figure 1). Since, this skewness can impact a model's normality assumptions and inflate residual variance, we decided to replace A1C with its logarithm.

Model 1:

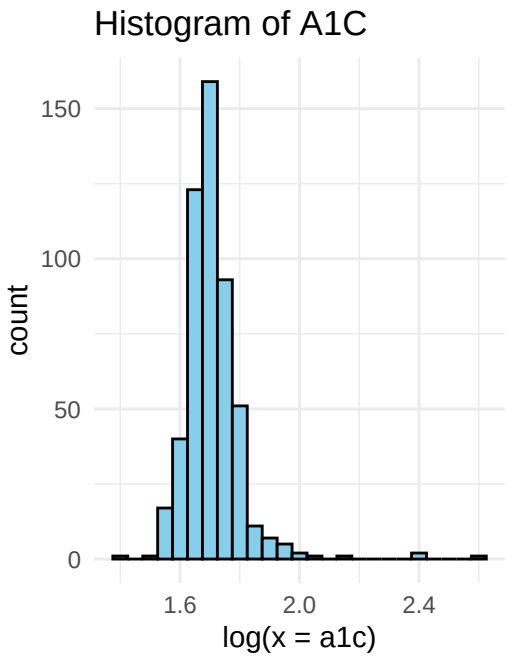
$$\log(\text{A1C}_i) = \beta_0 + \beta_1 \cdot \text{WeightHeightRatio}_i + \varepsilon_i$$

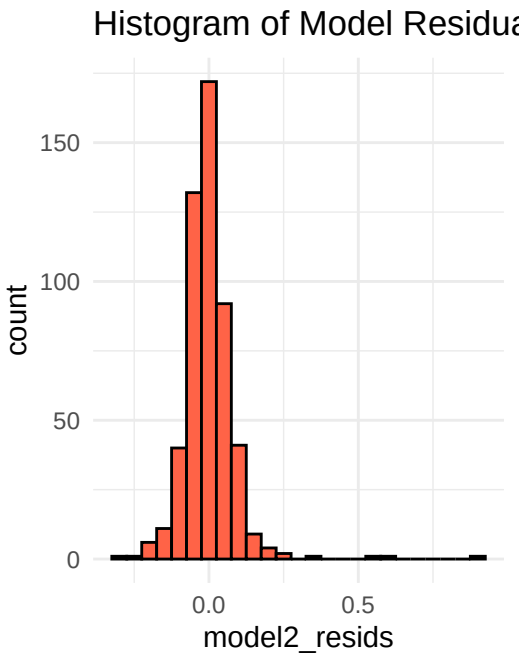
gan by fitting the univariate model where we can quantify the relationship between the waist-to-height ratio and log(A1C) without any other covariates. It addresses the following question: "What is the relationship between A1C levels and Waist_Height_Ratio?"

Model 2:

$$\begin{aligned}\log(\text{A1C}_i) = \beta_0 &+ \beta_1 \cdot \text{WaistHeightRatio}_i + \beta_2 \cdot \text{Age}_i + \beta_3 \cdot \text{Gender}_i \\ &+ \beta_4 \cdot \text{Ethnicity}_i + \beta_5 \cdot \text{IncomePovertyRatio}_i + \beta_6 \cdot \text{FamilyHistory}_i \\ &+ \beta_7 \cdot log(\text{Triglycerides}_i) + \beta_8 \cdot log(\text{Insulin}_i) + \varepsilon_i\end{aligned}$$

The expanded second model includes the joint distribution of demographic (age, gender, ethnicity), socioeconomic (income-to-poverty ratio), and health-related values (family history, triglycerides, insulin) together with the waist-to-height ratio.





We can see that they are positively skewed

Final Regression Models of A1C (Robust Standard Errors)

	Dependent variable:	
	log(A1C)	
	(1)	(2)
Waist-Height Ratio	0.234***	0.040
	(0.026)	(0.029)
Age		0.002***
		(0.0001)
Male		0.007
		(0.005)
Mexican American		0.027**
		(0.011)
Other Hispanic		0.021**

```
                                                (0.008)

Non-Hispanic Black                             0.043***
                                                (0.007)

Non-Hispanic Asian                             0.035***
                                                (0.006)

Other/Multi-Racial                              0.026**
                                                (0.011)

Income-to-Poverty Ratio                        -0.001
                                                (0.002)

Family History: Yes                            0.025***
                                                (0.008)

Triglycerides (log)                             0.012**
                                                (0.006)

Insulin (log)                                  0.026***
                                                (0.004)

-------------------------------------------------------
Observations                    1,196           1,196
R2                              0.062           0.242
Adjusted R2                     0.062           0.235
=======================================================
Note:                      *p<0.1; **p<0.05; ***p<0.01
```

```r
bp_result <- bptest(model_2)

bp_stat <- round(bp_result$statistic, 3)
bp_df   <- bp_result$parameter
bp_pval <- signif(bp_result$p.value, 4)
```

5. Model Assumptions

Our model meets most of the key assumptions for our regression model. The component plus residual plots for features - including log-transformed A1C, triglycerides, and insulin - , to some extent, confirm the linearity assumption between the outcome and the predictor

variables. Additionally, the residual versus fitted plot shows that the residuals are randomly distributed around zero and supports our second assumption of zero expectation conditional mean. The variance of residuals is constant based on the Breusch-Pagan test (BP = 13.561, df = 12, p-value = 0.3296), revealing the homoscedasticity. Subsequently, the Q-Q plot and residual histogram demonstrate that the residuals are approximately normally distributed with minor deviations at the tail.

When it comes to Independence and Identical Distribution assumption, it is important to acknowledge the fact that NHANES data are collected through sophisticated sampling design that can cause dependencies among observations - specifically within geographic clusters, demographic data, and oversampling of subgroups (Centers for Disease Control and Prevention, 2023). Survey weights and design variables should be incorporated to develop unbiased population estimates.

The same set of assumptions were met in our confirmatory dataset, except for homoscedasticity. We rejected the null hypothesis in the Breusch-Pagan test due to the p-value 3.725e-05 $<=$ 0.05. The outcome of this implied that our confirmatory model contained heteroscedasticity.

Moving forward, models can benefit from more additional improvements whether by adding more independent variables, exploring various interaction effects, addressing the outliers, or performing nonlinear transformations. These additional specifications, indeed, can potentially enhance the model's R^2 value (0.22) and boost overall model performance, by acquiring more meaningful relationships and improve the model's explanatory power.

6. Model Results and Interpretation

Appendix

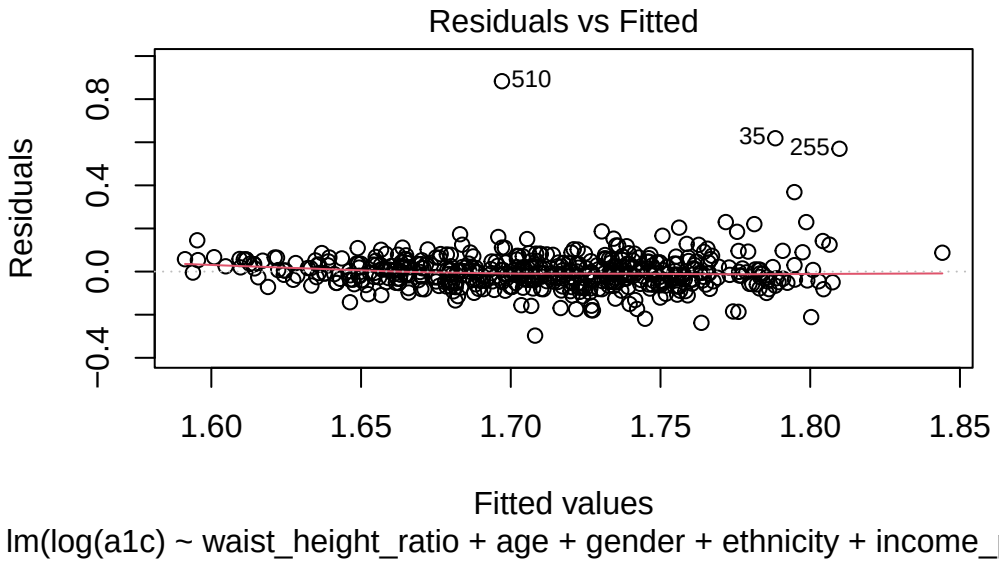
Appendix A

Specific Dataset Links:

1. NHANES datasets:
    1. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017
2. DIQ_J - Diabetes questionnaire: (Family-History)
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DIQ_J.htm
3. DEMO_J - Demographics (Age, Gender, Ethnicity, Income-to-Poverty Ratio,):
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DEMO_J.htm
4. GHB_J - Glycemic indicators (A1C):
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/GHB_J.htm
5. BMX_J - Body measurements (Waist and Height Measurements):
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/BMX_J.htm
6. TRIGLY_J - Triglyceride (Triglycerides):
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/TRIGLY_J.htm
7. INS_J - Insulin (Insulin):
    1. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/INS_J.htm

Appendix B

2. Linear Conditional Expectation



Residuals vs Fitted Plot demonstrates that residuals are randomly scattered across point zero, meeting our assumption for Zero Conditional Expectation Error.

3. **No Perfect Collinearity**

```
                        GVIF Df GVIF^(1/(2*Df))
waist_height_ratio   1.730323  1        1.315417
age                  1.145945  1        1.070488
gender               1.119255  1        1.057948
ethnicity            1.302089  5        1.026748
income_poverty_ratio 1.115795  1        1.056312
family_history       1.085194  1        1.041726
log(triglycerides)   1.339832  1        1.157511
log(insulin)         1.709608  1        1.307520

                      GVIF    Df GVIF^(1/(2*Df))
waist_height_ratio   FALSE FALSE           FALSE
age                  FALSE FALSE           FALSE
gender               FALSE FALSE           FALSE
ethnicity            FALSE  TRUE           FALSE
```

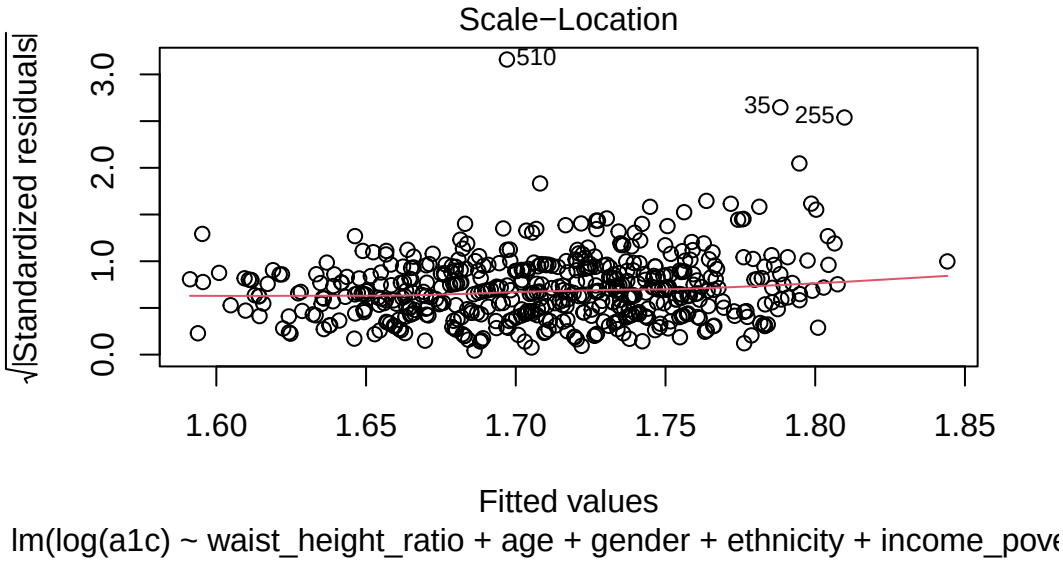
```
income_poverty_ratio FALSE FALSE           FALSE
family_history       FALSE FALSE           FALSE
log(triglycerides)   FALSE FALSE           FALSE
log(insulin)         FALSE FALSE           FALSE
```

The variance inflation factor for all features are below the threshold and have no evidence of multicollinearity.

4. Homoscedastic Errors

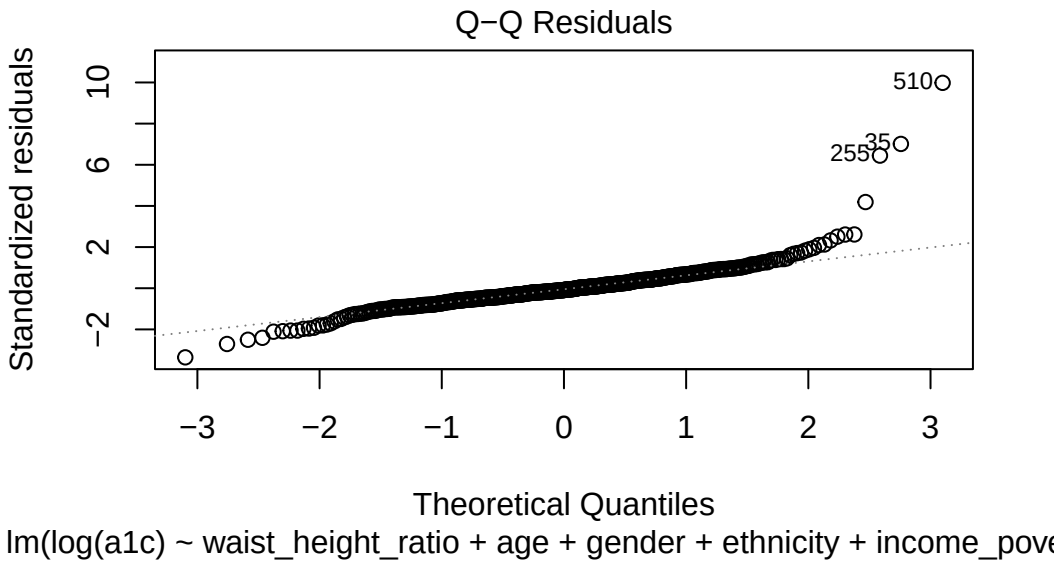
```
    studentized Breusch-Pagan test

data:  model_2
BP = 13.561, df = 12, p-value = 0.3296
```

Breusch-Pagan test with the p-value=0.33>0.05 informs us that we fail to reject the null hypothesis for homoscedasticity. In another words, model's residuals have constant variance.



When we plot standardized errors vs fitted values, we can see slight heteroscedasticity as our fitted values increase.

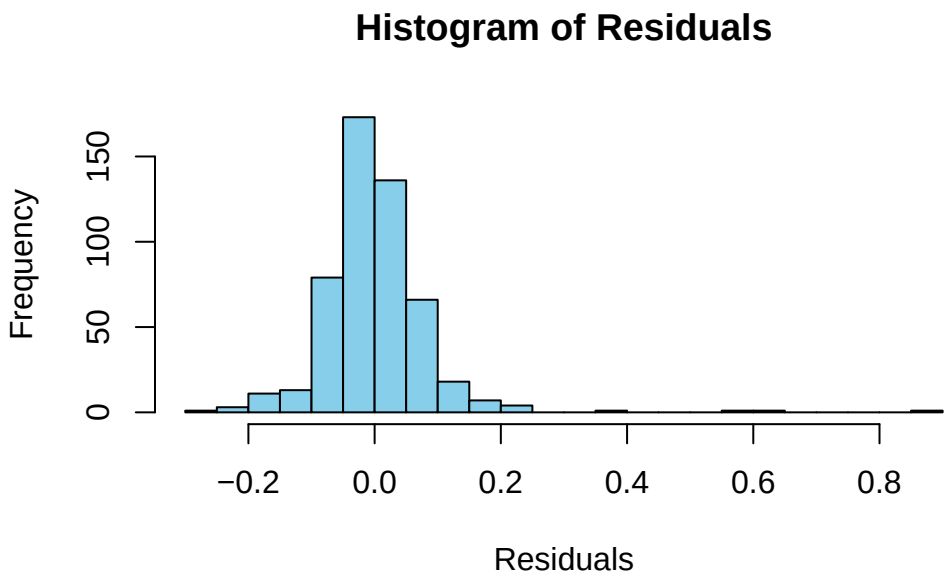
5. Normality



QQ plot demonstrates that

```
    Shapiro-Wilk normality test

data:  resid(model_2)
W = 0.79091, p-value < 2.2e-16
```

Shapiro test suggests that residual distribution is not normal. When investigating the QQ plot, we can still assume there is normality since most points fall along the line; However, there are few outliers that cause skewness on the right side of the plot that can violate our Normality assumption.



The residual distribution above also shows that it is positively skewed, violating the Normality assumption.

```
skewness(model_2$residuals) + c(-1,1)*(6/dim(exploratory_df[1]))^0.5
```

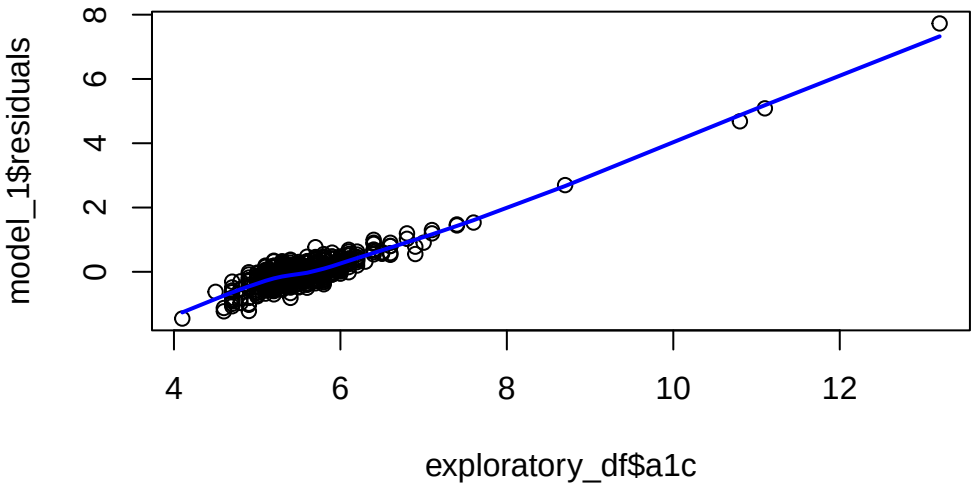
```
[1] 3.082649 5.640076
```

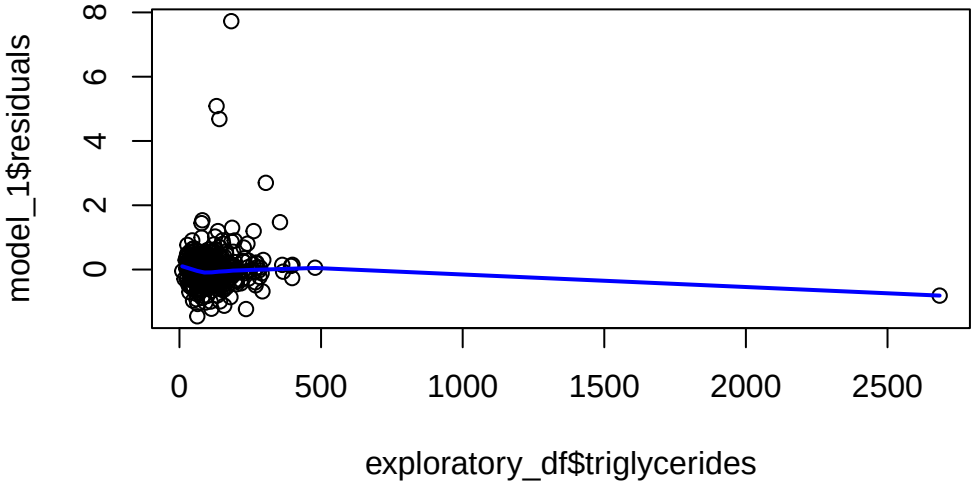
```
kurtosis(model_2$residuals) + c(-1,1)*(6/dim(exploratory_df[1]))^0.5
```

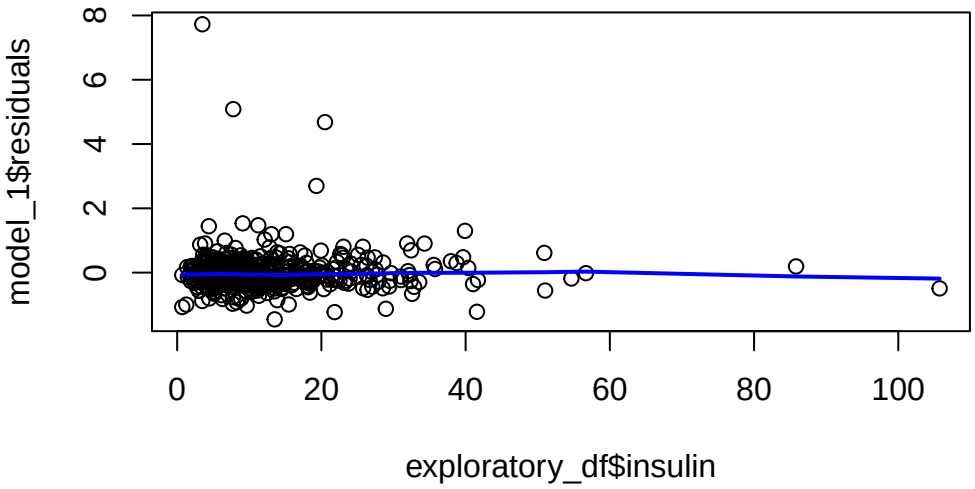
```
[1] 29.88141 32.43884
```

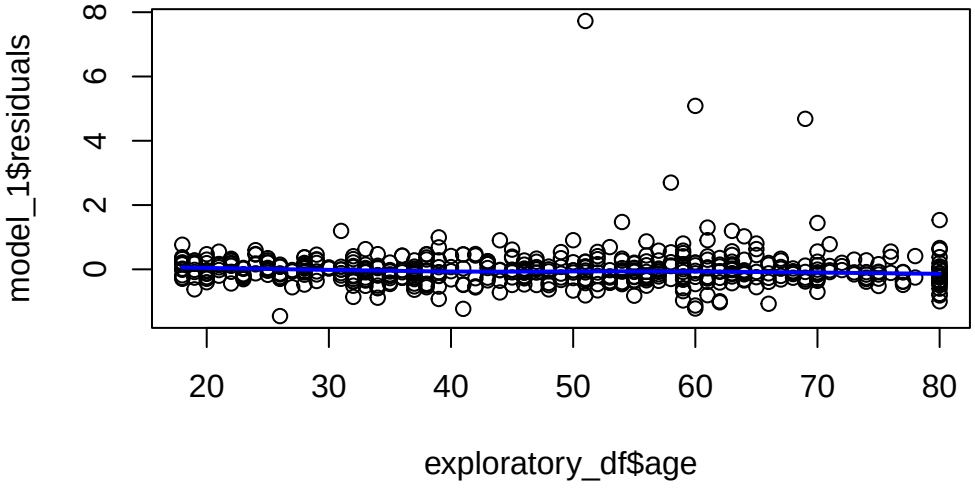
Based on skewness and kurtosis, there is still violation of Normality assumption, even after transforming variables like A1C, triglycerides, and insulin.

Appendix C

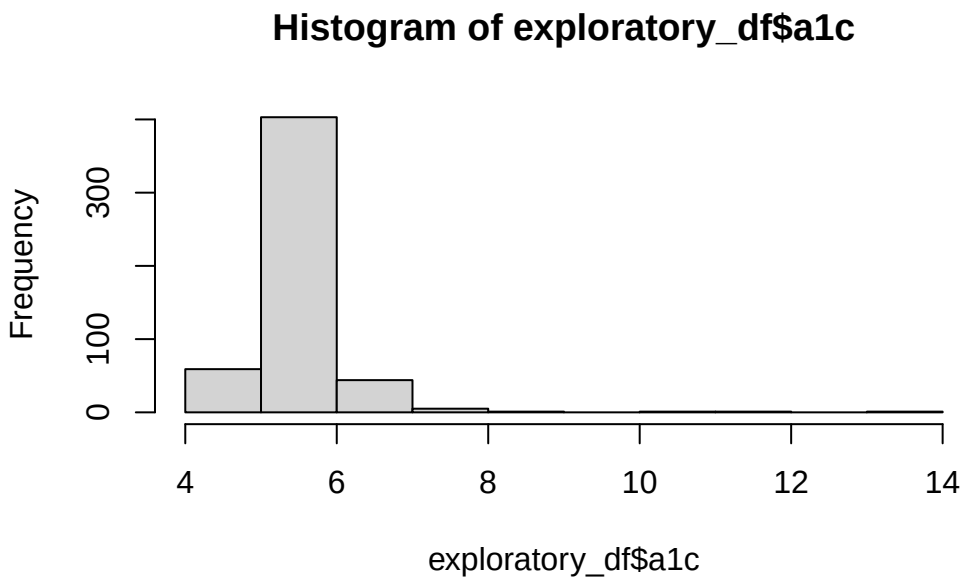

model_1$residuals
0
2
4
6
8
4
6
8
10
12
exploratory_df$a1c

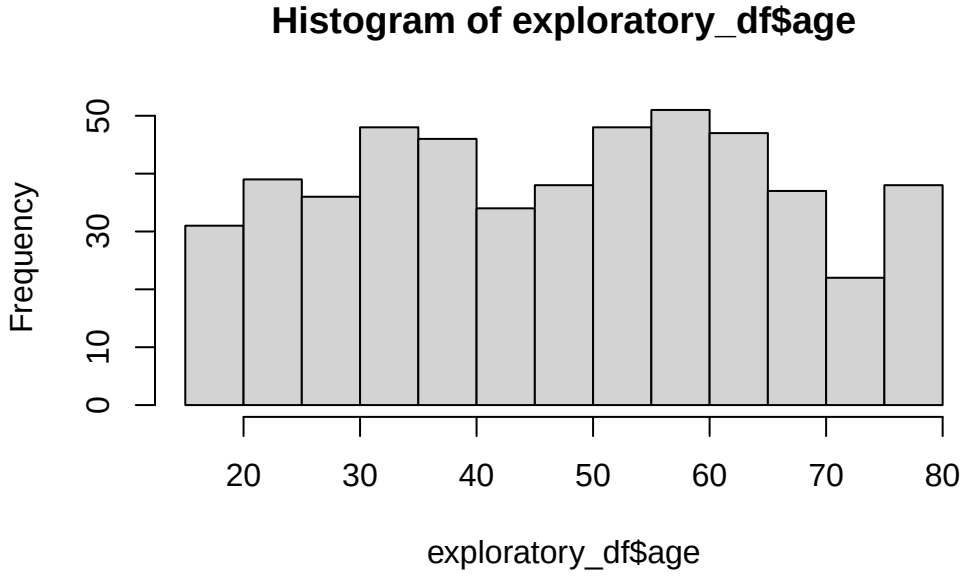

model_1$residuals
0
2
4
6
8
0
500
1000
1500
2000
2500
exploratory_df$triglycerides

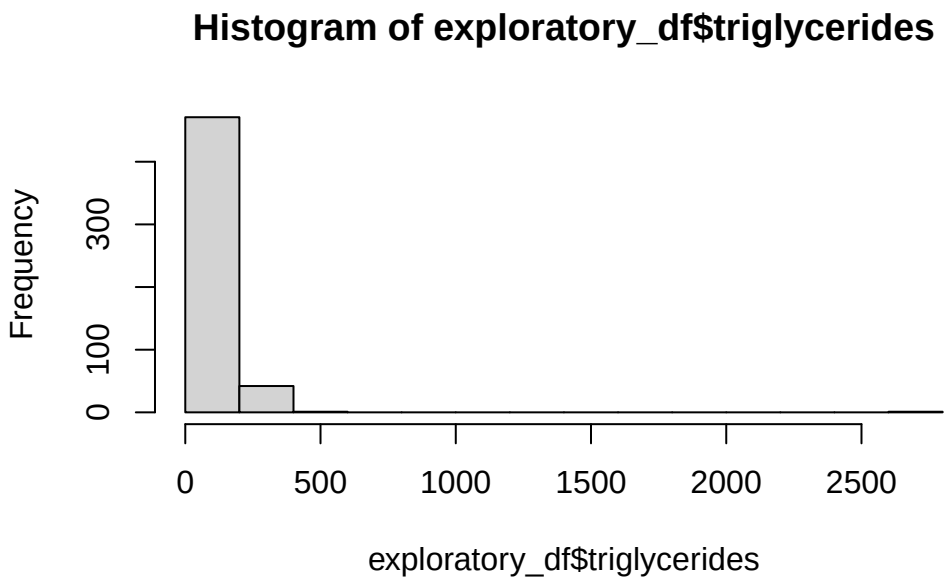


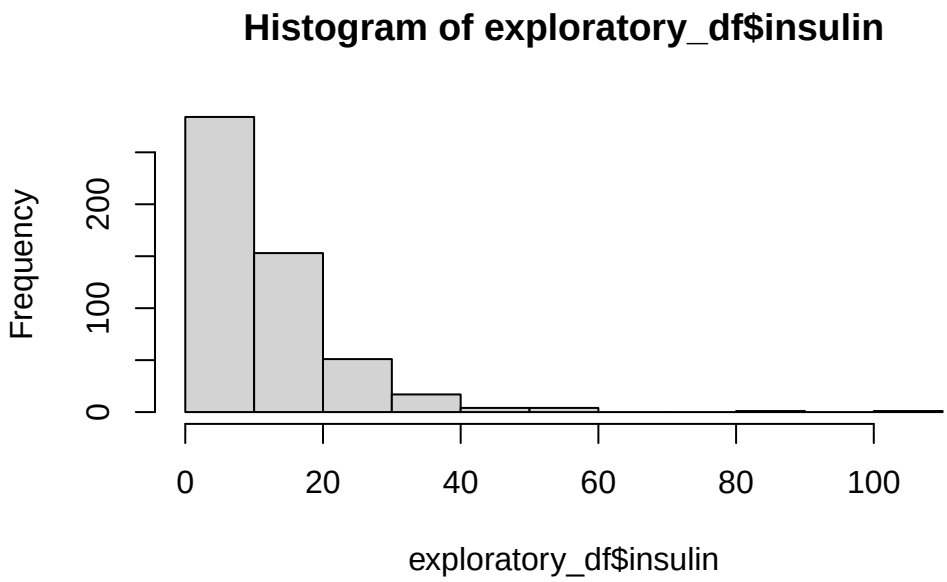


At first glance, all relationships seems approximately linear and no need to transform variables. However, for A1C, Triglycerides and Insulin the observations mostly occur on the left side, potentially causing skewness and kurtosis. Thus, we want to see their distribution and decide whether we want to transform them below.


Histogram of exploratory_df$a1c
Frequency
0
100
300
4
6
8
10
12
14
exploratory_df$a1c


Histogram of exploratory_df$age
Frequency
0
10
30
50
20
30
40
50
60
70
80
exploratory_df$age





Since A1C, Triglycerides, and Insulin are heavily right skewed, we decided to transform them using logarithm. This resulted in improved kurtosis and skewness scores.