STA302H1 Final Project

Mashaal Siddiqui

2023-05-31

R Markdown

Introduction

This paper aims to find a relationship between Systolic Blood Pressure (mm hg) and Glucose levels(Uu/mL). My interest lies in Blood Pressure and what can be affected by blood pressure or what blood pressure affects. Blood pressure can be affected by numerous factors such as sleep quality, exercise, genetics, body weight, cholesterol, medical conditions such as heart conditions, and more. I wanted to see if there was a relationship between glucose levels and systolic blood pressure. Some references I researched state that low blood sugar levels cause high blood pressure by increasing adrenaline which constricts the blood vessels. Other sources state that high blood sugar levels increase blood pressure by damaging blood vessels and causing plaque buildup. With these two sources, I wanted to see if in a survey of people if it was more likely that high blood sugar levels cause high blood pressure rather than the other way around. According to one paper, hyperglycemia (high glucose levels) is often found with hypertension, and other abnormalities (Henry P, 2002). In Pima Indians, systolic hypertension was found in men who had glucose intolerance, or high glucose levels. Another paper states that men with systolic blood pressure $\geq$ 160 mm hg have higher glucose levels regardless of Body Mass Index (Filipovský J, 1996). This relationship states the opposite however higher glucose concentrations were found in hypertension men.

Blood pressure is determined with two numbers. The first one is Systolic blood pressure, which is the blood pressure during heart muscle contractions. Diastolic Blood pressure refers to the pressure that blood is flowing against arteries while the heart is resting between contractions. I will be using mainly Systolic blood pressure as the explanatory variable. The systolic Blood Pressure is generally a better indicator for heart problems. Glucose levels while fasting are determined normal under 100 mg/dL. Greater than 126 mg/dL is diabetes. This relationship will be analyzed using survey results from the 2013-2014 National Health and Nutrition Examination Survey.

Methods

To determine the relationship, I performed a simple regression model with the explanatory variable as glucose, and the response variable as blood pressure. I used n=32 data points for each variable. This was obtained by matching the SEQN numbers as participant ID's. The SEQN numbers represents the number sequence of the surveyed individuals. The glucose data was obtained from the GLU_H data file, where there were different columns as well as the SEQN number. I used the LBXGLU column which had Fasting Glucose levels in mg/dL. From the PAXHD_G file, I used the columns BPXSYS3 and BPXDI3 which have the 3rd systolic and diastolic readings in mm hg. I took the first 32 points that had values from all columns without missing values. I also removed an outlier which had a blood pressure of 0 mm hg, since it cannot be possible. The National Health and Nutrition Examination Survey surveyed about 5000 Americans, where they used a 4-stage sample design. The first was selecting based on counties, the second was sampling based on area segments, the third being dwelling units, and the last being occupants within the dwelling units, such as households. I then used convenience sampling, by taking the first 32 values of the data set without missing values. I then applied a log transformation to improve normality, analyzed outliers, and fixed the model to decrease errors.

Results

When plotting the data, I noticed that the data was skewed to the left. To assume normality, I plotted the residuals of glucose which did not have a normal distribution. I decided to apply a log transformation on both variables to improve it's shape. Although the shape was not completely normal, it did improve. As a result, the new plot shifted values to the right and was less skewed. Additionally, the new modeled regression line had another point that was on the line as opposed to before.

```
##
## Call:
## lm(formula = health$SYS ~ health$GLU, data = health)
##
## Coefficients:
## (Intercept)   health$GLU
##     83.0954       0.3867
```

```
## [1] 0.3867027
```

```
##
## Call:
## lm(formula = health$SYS ~ health$GLU, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.633 -10.827  -2.606   9.654  21.793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.0954    11.3148   7.344 3.52e-08 ***
## health$GLU    0.3867     0.1066   3.626  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.65 on 30 degrees of freedom
## Multiple R-squared:  0.3047, Adjusted R-squared:  0.2815
## F-statistic: 13.15 on 1 and 30 DF,  p-value: 0.001055
```

```
## [1] 123.3125
```

```
## [1] 104
```

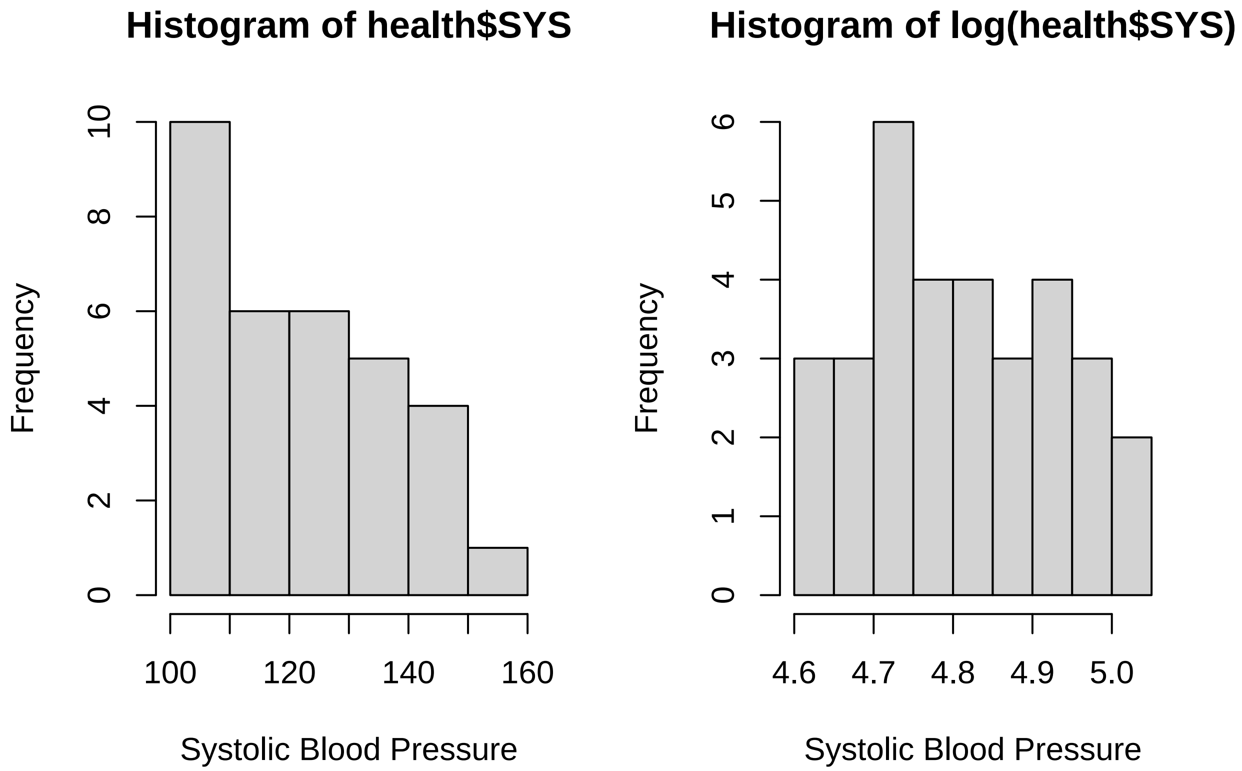

Figure 1. Sys. variable without and with lo

```
##
## Call:
## lm(formula = health$SYS ~ health$GLU, data = health)
##
## Coefficients:
## (Intercept)   health$GLU
##     83.0954       0.3867
```

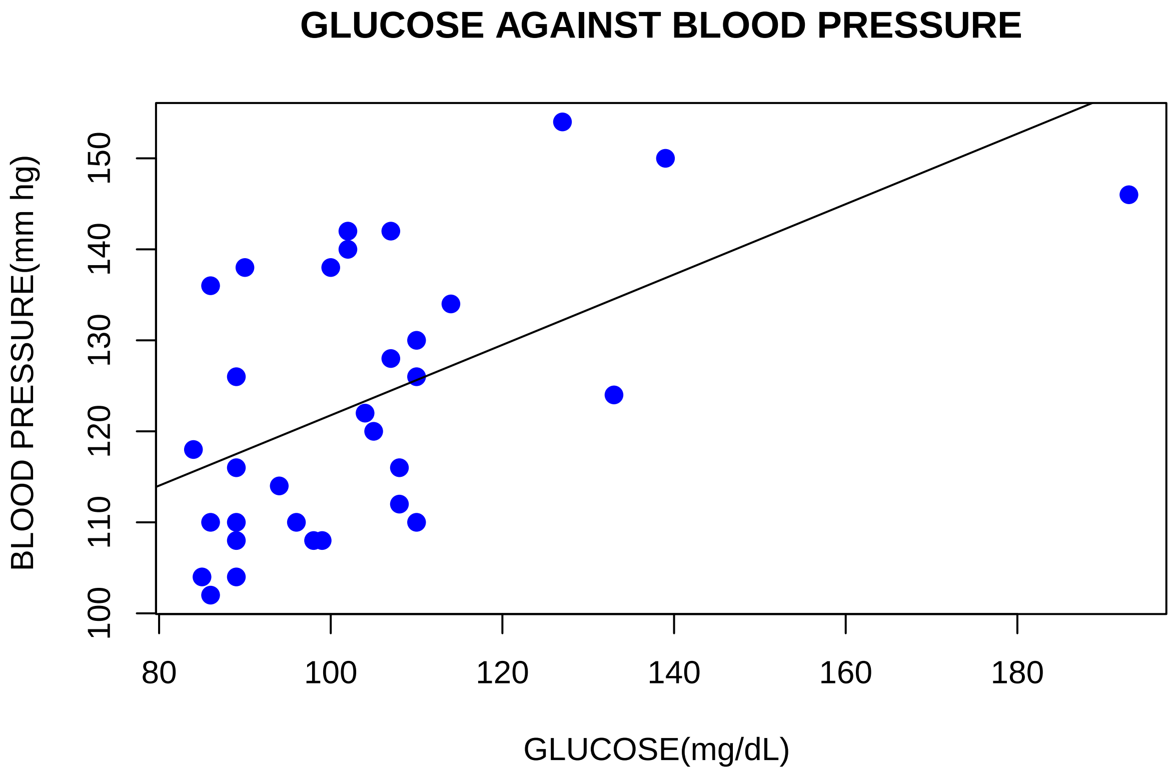

Figure 2. Glucose Against Blood Pressure without transformations

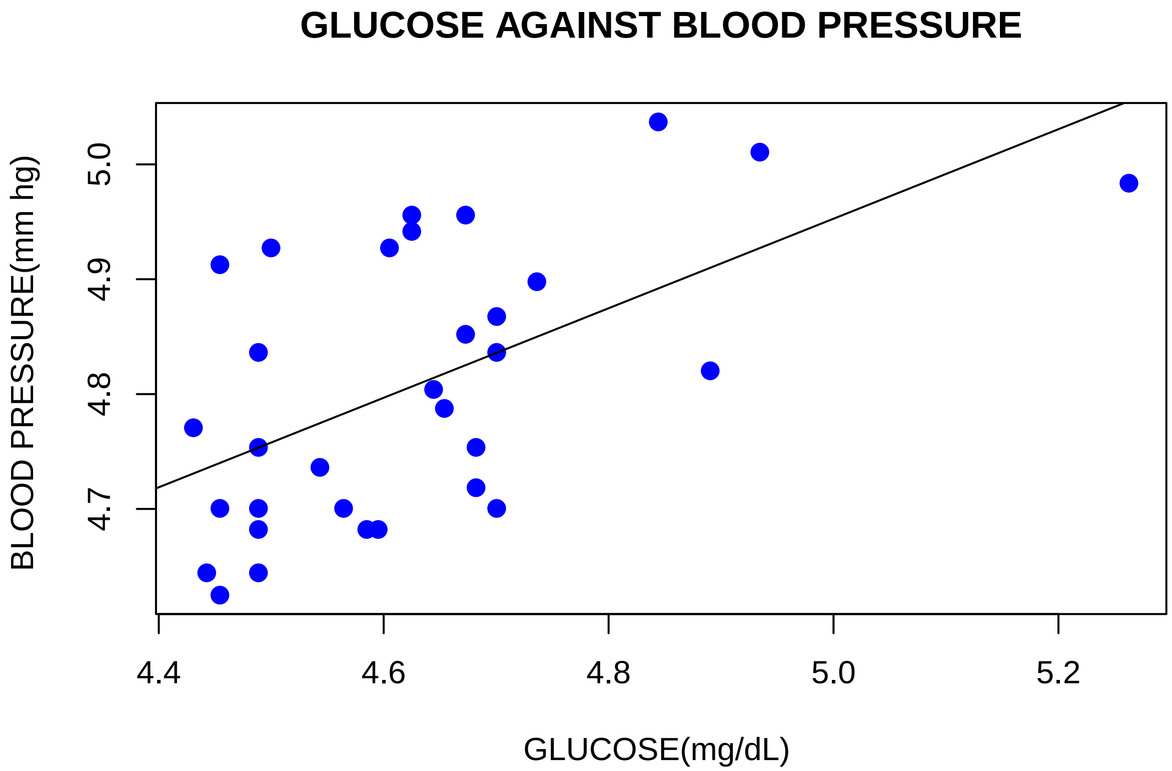

Figure 3. Glucose Against Blood Pressure with log transformations

```
##
## Call:
## lm(formula = log(health$SYS) ~ log(health$GLU), data = health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13547 -0.08423 -0.02020  0.08304  0.17263
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.0041     0.4764   6.306 5.94e-07 ***
## log(health$GLU)   0.3897     0.1029   3.789  0.00068 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09986 on 30 degrees of freedom
## Multiple R-squared:  0.3237, Adjusted R-squared:  0.3011
## F-statistic: 14.36 on 1 and 30 DF,  p-value: 0.0006795
```

```
## [1] -52.70989
```

```
## [1] -48.31268
```

After applying a transformation and simple linear regression, I obtained the model:

$$y_i = B_0 + B_1 x + e$$

$$y = 3.0041 + 0.3897x$$

where B_0 is 3.0041 and B_1 is 0.3897. This means for every one increase in blood pressure, there is about 0.3867 increase in glucose levels. The slope is not high, however, we can see that the line trends upwards instead of downwards. This agrees with sources that state that higher glucose levels can correlate to higher blood pressure, instead of lower glucose levels. The general mean glucose level is 104 and the mean systolic blood pressure is 123.3125 from the data given. The standard deviation of the intercept, $std(B_0) is 0.4764$ and the standard deviation of the slope, $std(B_1) is 0.1029$. The residual standard error, where $RSE = \sqrt{(}RSS/n-2)$, is 0.09986. This means that model predicts blood pressure from the given glucose with a deviation error of about 0.09986. The coefficient of determination, R^2 is 0.3237. This is an increase from the coefficient of determination of the original model, which was 0.3047, and increased in the adjusted value from 0.2815 to 0.3011. A significant coefficient would a number close to 1, or above 0.7. This coefficient is not significant enough, so there is not a significant enough relationship between glucose levels and blood pressure, only somewhat correlated. I then conducted a hypothesis test with 5% significance level for the slope, where the null hypothesis is that there is no relationship between the glucose and blood pressure variables. I obtained a T value of 3.787 and a p-value of 0.0007. Since a p-value of 0.05 or less is significant, and we reject the null hypothesis and accept the alternate hypothesis. We can consider it to be statistically significant. I then did a 95% confidence interval to determine if B_1 will be contained in the interval.

$$H_0 : B_1 = 0$$

$$H_a : B_1 \neq 0$$

$$T = \frac{(B_1 - 0)}{SE(B_1)}$$

$$T = \frac{0.3897}{0.1029}$$

$$T = 3.787$$

$$p = 0.0007$$

I also conducted a 95% confidence interval of containing the value of B_1.

$$B_1 = 0.3897$$

$$SE(B_1) = 0.1029$$

$$(B_1 - 2SE(B_1), B_1 + 2SE(B_1))$$

$$(0.3897 - 2(0.1029), 0.3897 + 2(0.1029))$$

$$(-0.0219, 0.5955)$$

This means there is a 95% chance that the interval [-0.0219, 0.5955] will contain B_1

Looking at the original plot, we can see that many of the values cluster near the left around the mean of the blood pressure of about 123. When plotting the regression, we can see the values more spread out and the line over most data points. There is a data point with blood pressure of around 180 which could be pulling the line above. I analyzed any residuals and errors with any significance. The residual standard error is 0.09986, so the values differ from the regression line by about 0.09986 units. I plotted a QQplot and QQnorm of the residuals. From the plot the values follow the line somewhat in the middle but there is deviation at the lower and higher quantiles. Because of this, we cannot be certain about normality. To analyze any deviations further, I calculated and leverage values, and Cook's distance.

```
##             1             2             3             4             5 
## -0.0714492232  0.1306529983  0.0003314451 -0.0712590268  0.0398279136 
##             6             7             8             9            10 
## -0.0529098881 -0.0100701333  0.0315839886 -0.0384933764  0.0268562046 
##            11            12            13            14            15 
## -0.1354700960  0.1449965093  0.0828916531  0.1695089533 -0.1088013529 
##            16            17            18            19            20 
## -0.0910779570  0.1284476712  0.1493035295 -0.0896648782  0.0479692253 
##            21            22            23            24            25 
##  0.1726278288  0.0001999372  0.0834924552 -0.0395466912 -0.1103005325 
##            26            27            28            29            30 
## -0.1127579526 -0.0752092127  0.1351188945 -0.1089993547 -0.0303288582 
##            31            32 
## -0.1150542437 -0.0824164300
```

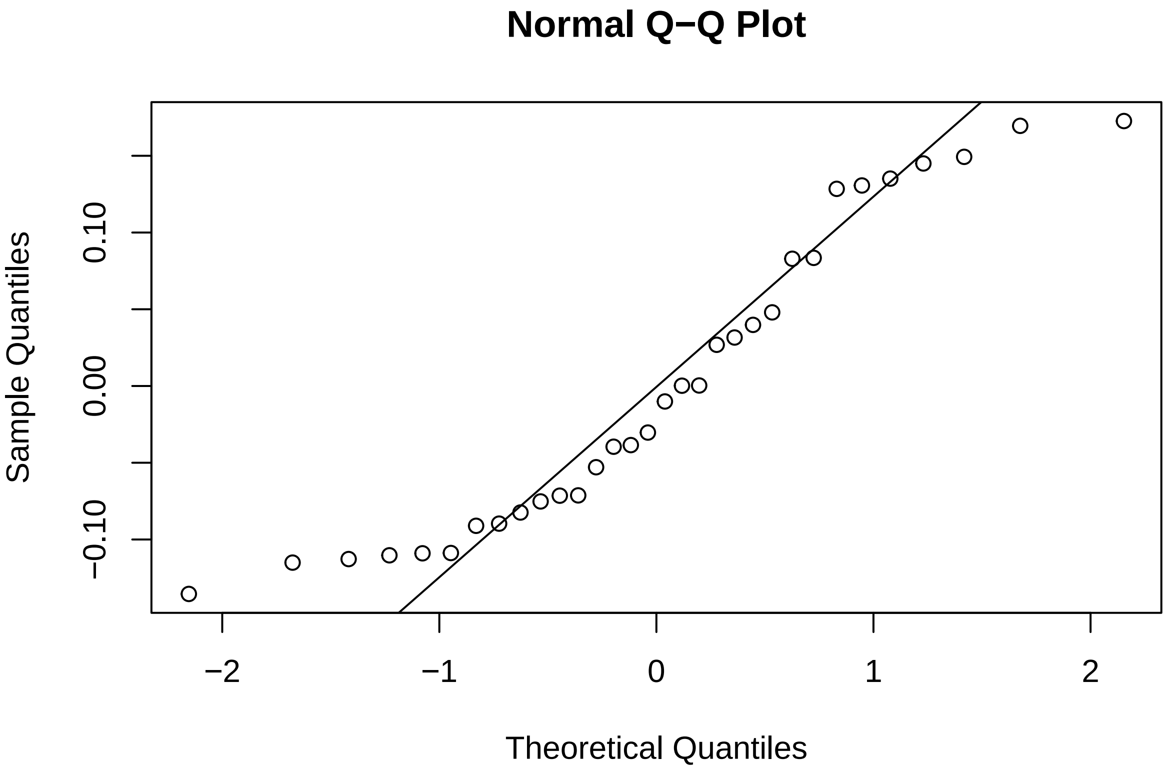

Figure 4. Q−Q Plot of Residuals

The bad leverage points are defined as being larger than $2\frac{p}{n}$ where p is the number of parameters and n is the number of observations. Since p is 2 and n is 32, we get a cutoff of 0.125. So, bad leverage values are greater than 0.125. From the following list, we see that the 1st observation is a leverage value with a hat value of 0.45846619. The 23rd observation also has a hat value of 0.1308201. I then calculated Cook's distance to validate the outlier. Cook's distance is $D_i > $ 50th percentile of F(p, n-p), or $D_i > F^{-}1_{p,n-p}(0.5)$. I then removed these values and observed the new model.

After removing the observation, we can see that the slope decreased and the intercept increased. However, the R^2 has decreased slightly from 0.3047 to 0.2718. The Adjusted R^2 has lowered slightly from 0.2815 to 0.2467. The difference in R^2 is not drastic but still lowers the correlation. The first observation had a glucose level of 193 and a systolic blood pressure of 146. This glucose level is too high and may indicate diabetes for the participant. It is much higher than the mean of 104. The systolic blood pressure is not too higher than it's mean of 123.3125. Since it farther because of it's glucose level, the observation is an outlier, however, it still agrees with the line since the determination coefficient lowers after removing it. The Residual Standard Error has also increased slightly from 0.09986 to 0.1002. There are reasons for removing and not removing the outlier which will be considered. I will explore not removing the outlier first.

When substituting it in the original regression line,

$$y = 83.0954 + 0.3867(193) = 157.73$$

In the new line, it gives a value of

$$y = 65.0085 + 0.5693(107) = 174.8834$$

The model estimates a blood pressure of 157.37 mm hg which is close to it's value of 146. For non-extreme values, the lines output similar lines to each other. However, for extreme values, the new line outputs higher blood pressure levels, for both high and low glucose levels. Even though in the plot it appears farther, it still agrees with the original line somewhat. It also does not need to be removed if it does not make sense. In this scenario, a glucose level of 193 and a blood pressure of 143 makes sense as it is possible for people who have diabetes or higher bloods pressure. It also does not need to be removed if it does not affect the regression line. For reasons removing the outlier, it should be done if it is a mistake and the data was inputted correctly. In this situation, it is likely not an error. It should also be removed if it is creating an association when there is none. In this case, it is barely increasing the association as R^2 is slightly higher, but it is not creating a relationship on its own. Additionally, decreasing sample size also naturally decreases R^2. The AIC and BIC also did not decrease by a significant amount, even though a lower value is optimal. Lastly, extreme values could help the model, and be more accurate between the variables when glucose levels are high. I will leave the outlier in the model for these reasons and to improve predictions. To improve the model further, I will correct variance using weighted least squares.

I will use the predicted values as our approximated weights where $w = 1/\sigma^2$

I will then fit through weighted least squares.

```r
methodweights=lm(log(health$SYS)~log(health$GLU),data=health, weights = w)
summary(methodweights)
```

```
##
## Call:
## lm(formula = log(health$SYS) ~ log(health$GLU), data = health,
##     weights = w)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.012479 -0.007843 -0.001901  0.007199  0.017142
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.9494     0.5118   5.763 2.71e-06 ***
## log(health$GLU)   0.4015     0.1109   3.622  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009377 on 30 degrees of freedom
## Multiple R-squared:  0.3042, Adjusted R-squared:  0.281
## F-statistic: 13.12 on 1 and 30 DF,  p-value: 0.001067
```

```
AIC(methodweights)
```

```
## [1] -52.4101
```

```
BIC(methodweights)
```

```
## [1] -48.01289
```

```
par(mfrow=c(1,2))
# predictor vs residuals
plot(log(health$GLU),rstandard(methodweights), sub = "Figure 5. Predictor vs Residuals", xlab="Num of re
#fitted values vs residuals
plot(methodweights$fitted.values, rstandard(methodweights), sub = "Figure 5.1. Fitted values vs Residual
```

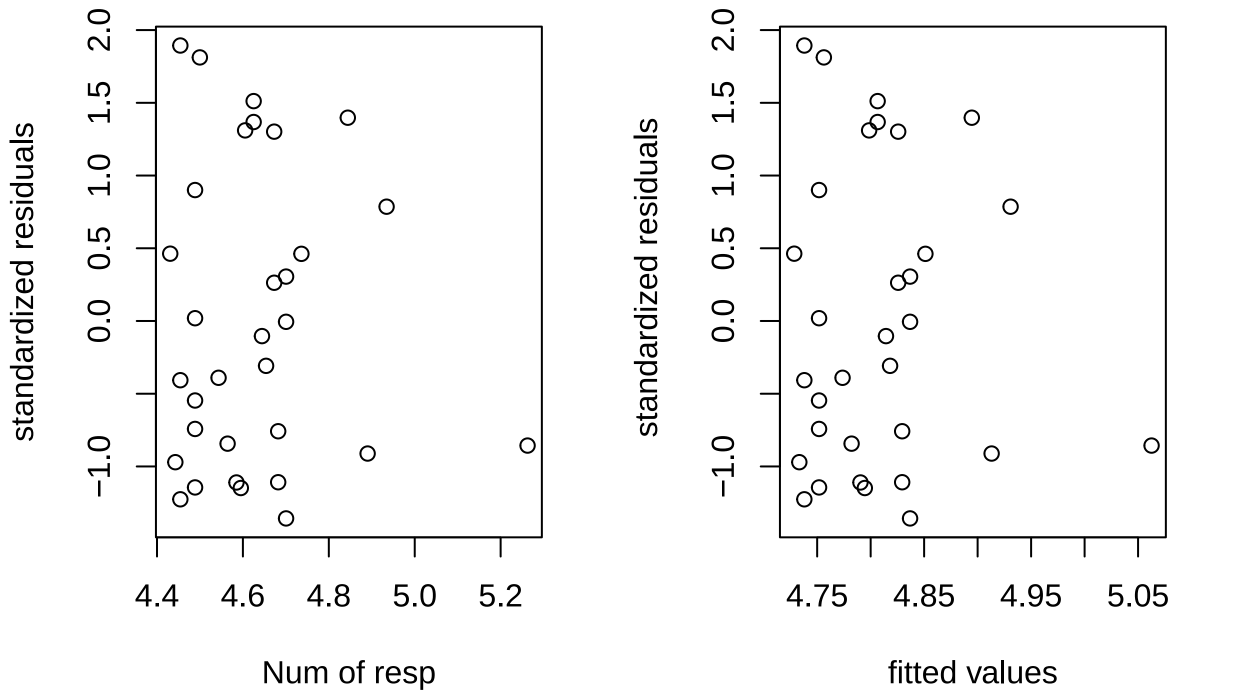

Figure 5. Predictor vs Residuals

Figure 5.1. Fitted values vs Residuals

```
plot(log(SYS)~log(GLU), health, sub = "Figure 6. Weighted Least Squares Regression line", xlab = "GLUCOS
abline(newmod, col="blue")
abline(methodweights, col="red")
```

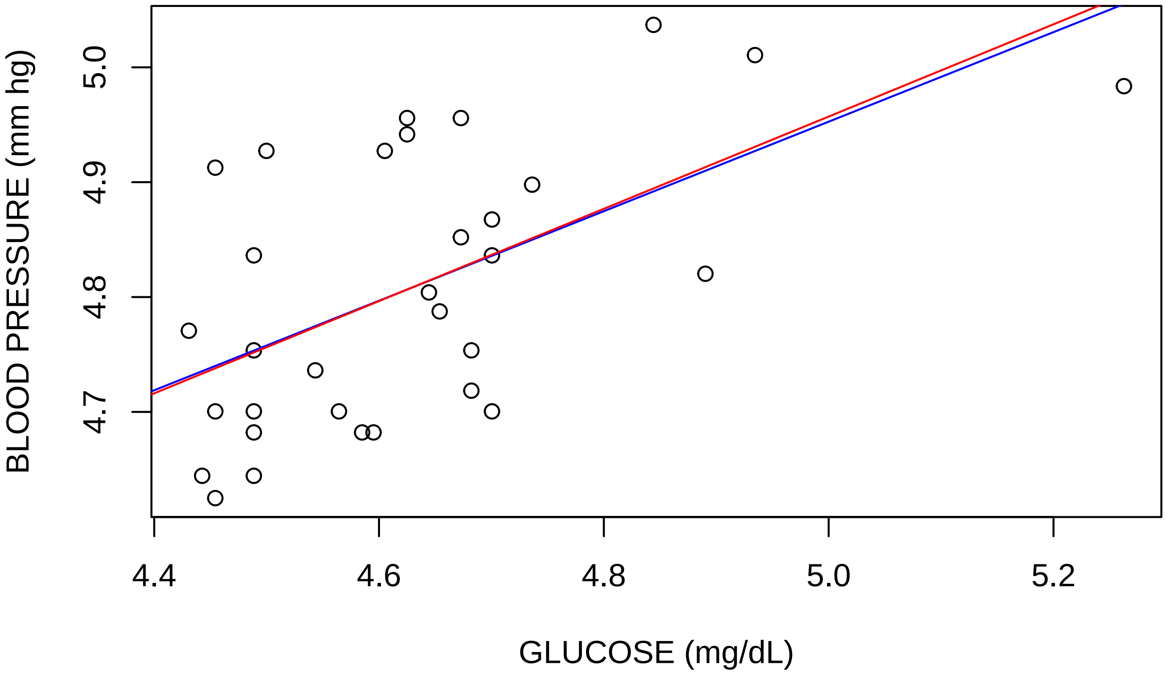

Figure 6. Weighted Least Squares Regression line

```
hist(methodweights$residuals, main = "Residual Histogram", sub = " Figure 7. WLS Residual Histogram", xl
```

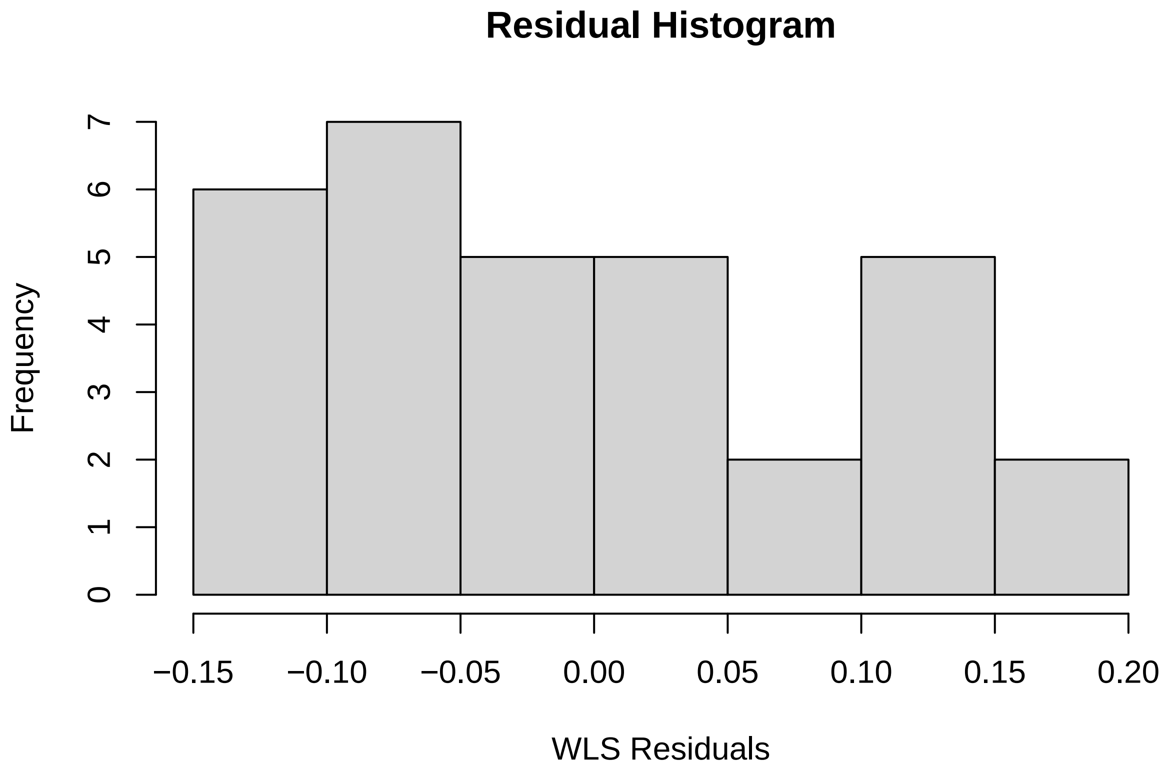

Figure 7. WLS Residual Histogram

After applying the weighted values onto the model, the residual standard error decreased from 0.09986 to 0.009377. The adjusted R^2 also decreased from 0.3011 to 0.281. There was only a slight decrease and from observing the plot, the line in increasing slightly more than the unweighted model. Additionally, the p-value in this model is 0.001067 which is significant for a 5% significance level hypothesis test where

$$H_0 : B_1 = 0$$

and

$$H_a : B_1 \neq 0$$

. The null hypothesis is accepted. The AIC and BIC of this model does not increase from the unweighted model which is ideal. Because the residual error decreased, I will use the methodweights model as my final model with regression line $y = 2.9494 + 0.4015x$.

Lastly, the model must be validated to see if the model can be used for predictions for specific data and can describe the population. This will be done by creating two datasets, a training dataset and a testing dataset. I will split my data set in a ratio of 70/30 where 70 is the training and 30 is the testing.

```
set.seed(1)
index <- sample(1:nrow(health), 0.7*nrow(health))
train_data <- health[index, ]
test_data <- health[-index, ]

trained_model <- lm(log(train_data$SYS)~log(train_data$GLU), data = train_data)
fitsigma = data.frame(x=log(train_data$GLU), y=abs(trained_model$residuals))
auxmodel = lm(y~x, fitsigma)
w = 1/auxmodel$fitted.values^2
```

```
weighted_mod=lm(log(train_data$SYS)~log(train_data$GLU),data=train_data, weights = w)
summary(weighted_mod)
```

```
##
## Call:
## lm(formula = log(train_data$SYS) ~ log(train_data$GLU), data = train_data,
##     weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4856 -0.9355 -0.3735  1.0703  1.8672
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.1998     0.4889   6.544 2.23e-06 ***
## log(train_data$GLU)   0.3472     0.1041   3.336  0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.176 on 20 degrees of freedom
## Multiple R-squared:  0.3575, Adjusted R-squared:  0.3254
## F-statistic: 11.13 on 1 and 20 DF,  p-value: 0.003294
```

```
anova(weighted_mod)
```

```
## Analysis of Variance Table
##
## Response: log(train_data$SYS)
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## log(train_data$GLU)  1 15.393 15.3931  11.128 0.003294 **
## Residuals           20 27.666  1.3833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mean((log(train_data$GLU) - log(train_data$SYS))^2)
```

```
## [1] 0.04996085
```

```
test_values <- predict(weighted_mod, newdata = test_data)
```

```
## Warning: 'newdata' had 10 rows but variables found have 22 rows
```

```
#test_values

actual_values <- log(test_data$SYS)
#actual_values

mse <- mean(test_values - actual_values)^2
```

```
## Warning in test_values - actual_values: longer object length is not a multiple
## of shorter object length
```

```
mse
```

```
## [1] 0.001017463
```

```
r_squared <- 1 - sum((actual_values - test_values)^2) / sum((test_values - mean(actual_values))^2)
```

```
## Warning in actual_values - test_values: longer object length is not a multiple
## of shorter object length
```

```
r_squared
```

```
## [1] -2.124918
```

After finding the MSE, we can see that the test data has a MSE of 0.001 and the MSE of the training data is about 0.005. The MSE is smaller for the test data. The low MSE indicates better accuracy and the model is validated.

Discussion

My final model is the regression line $y = 2.9494 + 0.4015x$

This model was obtained after transformations, analyzing outliers, and lowering variance. Firstly, I applied a transformation to decrease it's left skew, improve normality, and increase its accuracy. The model had an increase in R^2 and the residual standard error was decreased. I then analyzed the outliers that could significantly affect the model. According to the leverage in this model, values that were considered high were the 1st and 23rd observations. These values had only a small impact on the data and were kept in to improve statistical accuracy since it gives information to people who have higher glucose and blood pressure levels and they were not errors. I then used weighted least squares to improve the model and lower the residual standard error. It slightly changed the model from a slope of 0.3897 to 0.4015 and the intercept from 3.0041 to 2.9494. The residual standard error decreased from 0.09986 to 0.009377. Lastly, I validated the model to check if it's accuracy can hold for future predictions.

The correlation between glucose and high blood pressure is is not significant. This is explained by the weak adjusted R^2 value of 0.281. Additionally, the QQplot shows that the residuals are not completely linear and the histogram is not normal, so the model cannot explain any trends completely. These limitations would be removed if the residuals were normal. There is an upward trend of glucose and blood pressure. With higher glucose levels, it is more likely for there to be high blood pressure. With further analysis, knowing the blood glucose and blood pressure relationship can help keep people healthy and alter their lifestyle. If they have existing conditions such as diabetes, they can be at risk for other conditions such as hypertension. Since there is a range of normal blood pressure that is common, these values were found toward lower glucose levels and the left of the plot. This affected the randomness of the data and any trends. Even though the correlation is low and the residuals are not completely normal, the data agrees with some findings that the papers previously mentioned, where high blood pressure levels are associated with high glucose levels.

References

Centers for Disease Control and Prevention. (2017, January 26). National Health and Nutrition Examination Survey. Kaggle. https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey Filipovský, J., Ducimetiére, P., Eschwége, E., Richard, J. L., Rosselin, G., & Claude, J. R. (1996). The relationship of blood pressure with glucose, insulin, heart rate, free fatty acids and plasma cortisol levels according to degree of obesity in middle-aged men. Journal of hypertension, 14(2), 229–235. https://doi.org/10.1097/00004872-199602000-00012 Henry, P., Thomas, F., Benetos, A., & Guize, L. (2002). Impaired fasting glucose, blood pressure and cardiovascular disease mortality. Hypertension, 40(4), 458–463. https://doi.org/10.1161/01.hyp.0000032853.95690.26 Irmanie Hemphill, M. (2023, January 23). Low blood sugar with high blood pressure. K Health. https://khealth.com/learn/hypertension/low-blood-sugar-with-high-blood-pressure/#:~:text=Yes%2C%20low%20blood%20sugar%20can,can%20raise%20blood%20pressure%20levels Mayo Foundation for Medical Education and Research. (2023, May 3). Diabetes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451#:~:text=Fasting%20blood%20sugar%20test OmegaQuant. (2022, September 14). Are blood sugar and blood pressure related?. OmegaQuant. https://omegaquant.com/are-blood-sugar-and-blood-pressure-related/#:~:text=It%20turns%20out%20that%20hyperglycemia,%2C Understanding blood pressure readings. www.heart.org. (2023, May 30). https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings