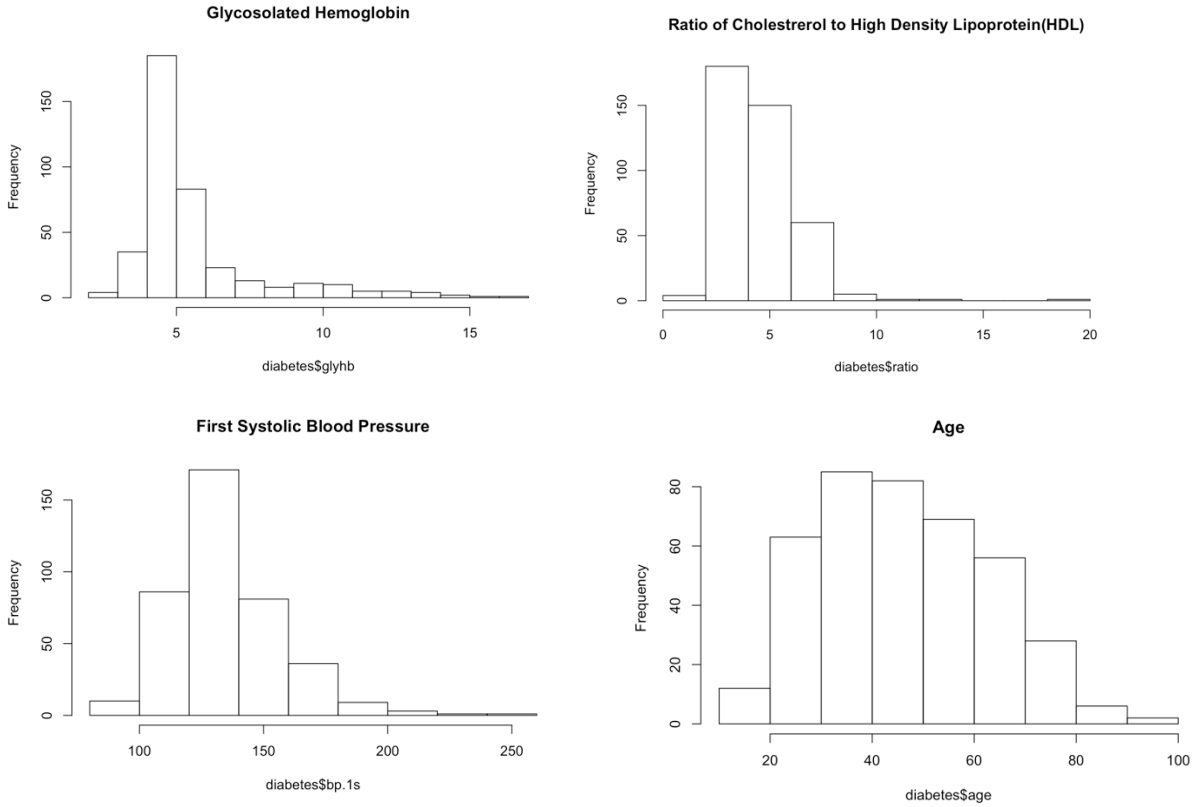
Diabetes Data Project

1. The missing values in the variable frame were replaced by “NA”. An empty categorical value that had no label was removed from frame as well.

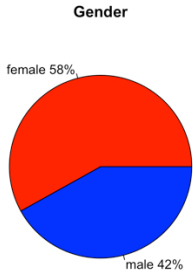
2. The variables Glycosolated Hemoglobin (glyhb), cholesterol/High Density Lipoprotein ratio (ratio), first systolic blood pressure (bp.1s), and age are quantitative variables. The variables gender and frame size (frame) are qualitative variables.

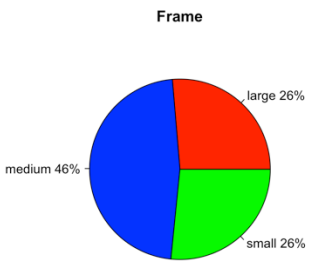
   The histogram for Glycosolated Hemoglobin (glyhb) of study participants displays a right-skewed bell curve.
   The histogram for the ratio of cholesterol to High Density Lipoprotein (HDL) of study participants displays a right-skewed bell curve.
   The histogram for the First Systolic Blood Pressure (bp.1s) of study participants displays a bell curve.
   The histogram for age of study participants displays a wide bell curve. Most of the participants were between the ages of about 20 years old to 60 years old.



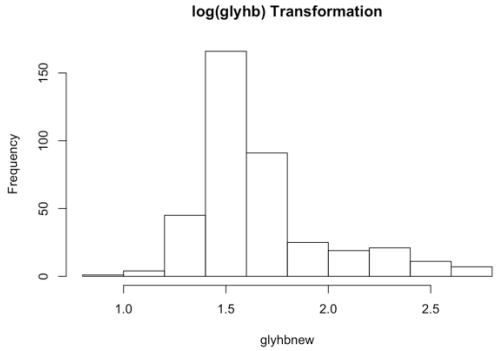
The pie chart for gender of study participants shows that a little more than half of participants are female with the remaining fraction consisting of male participants.

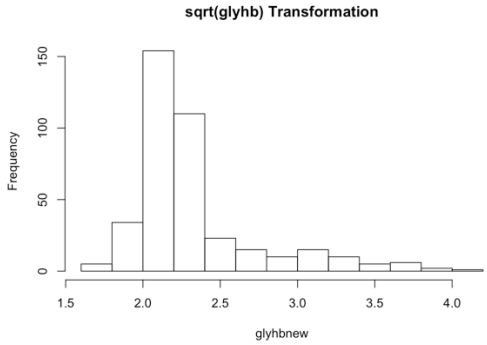
The pie chart for body frame size of study participants shows that almost half of participants have a medium frame size, about a quarter of participants have a large frame size, and a quarter of participants have a small frame size.

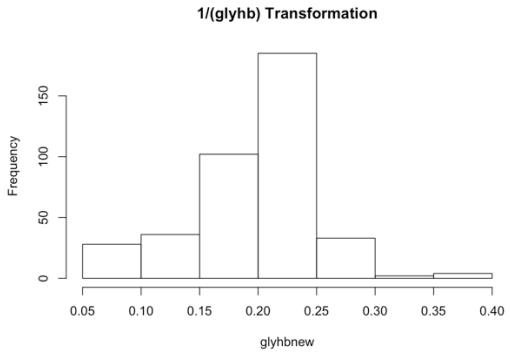




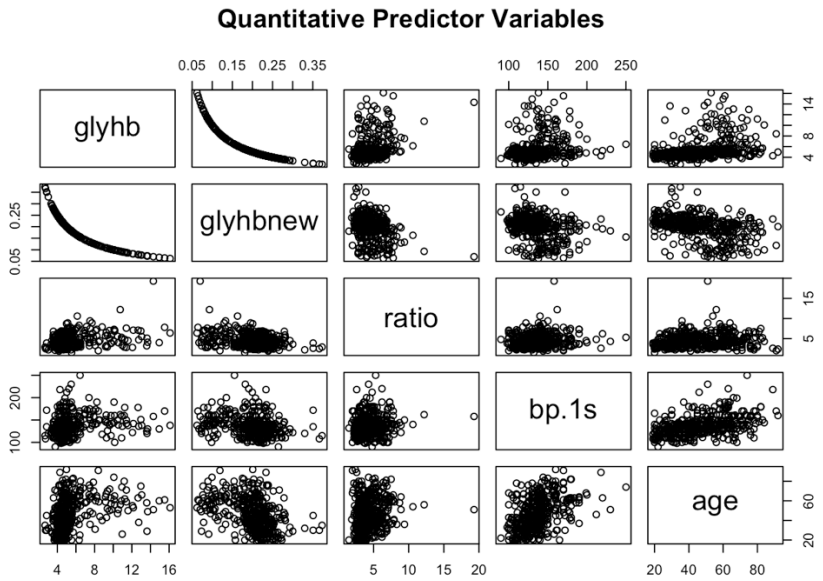
3. Looking at the severity of the right-skewed curve shown in the histogram for Glycosolated Hemoglobin (*glyhb*) of study participants, concern about the Normal error assumption arises. Among possible transformations to the data for glyhb – log(glyhb), sqrt(glyhb), and 1/glyhb – the distribution of the transformation 1/glyhb appears to be the most normal and will for the rest of the report be denoted by glyhb* (glyhbnew in R code).



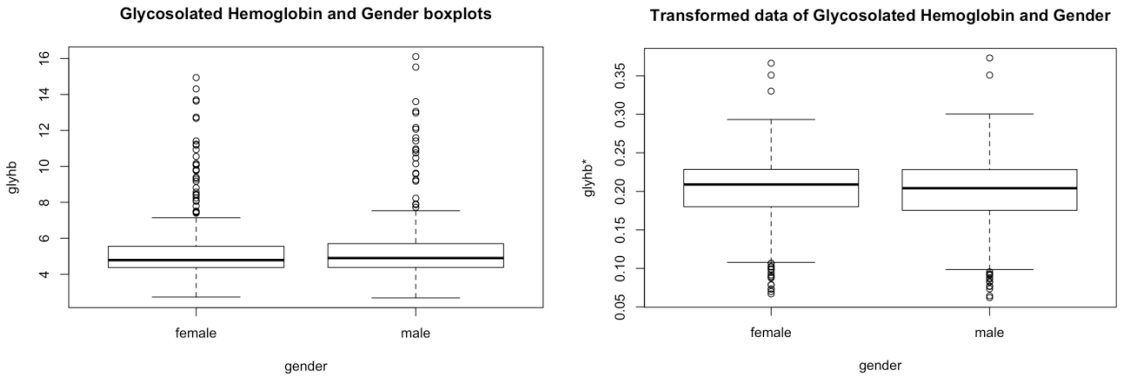




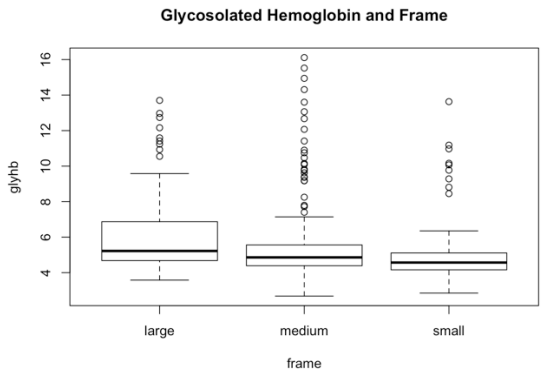
4. The scatterplot matrix shows there is non-linearity between glyhb and glyhb* which can be expected since the variables hold the inverse data of each other. There is no obvious non-linearity between the rest of the variables.

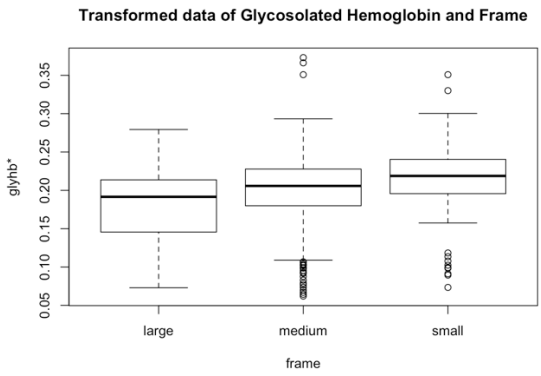


5. Below are boxplots to show the frequencies of glyhb and glyhb* levels in male and female participants. These boxplots show that there is no obvious difference in glyhb between male and female participants. There is a difference in the interquartile ranges, quartiles, and outliers between the data of glyhb and glyhb*.

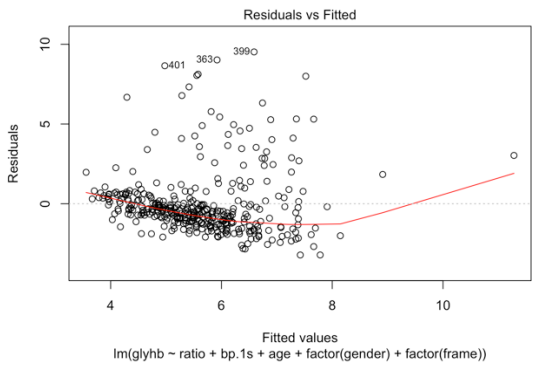


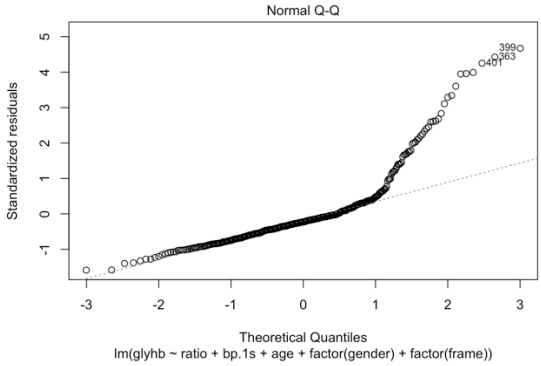
Below are boxplots to show the frequencies of glyhb and glyhb* levels for participants of different frame sizes. In the boxplot comparing glyhb levels in large, medium, and small frames, we see that the interquartile range increases as the frame size increases, and as the frame size decreases the median of the glyhb levels decreases. In the boxplot comping glyhb* levels in large, medium, and small frames, we see that the interquartile range increases as the frame size increases, but the median of the glyhb* levels increases as the frame size decreases. Also, the two boxplots differ in the interquartile ranges and where outliers can be found.



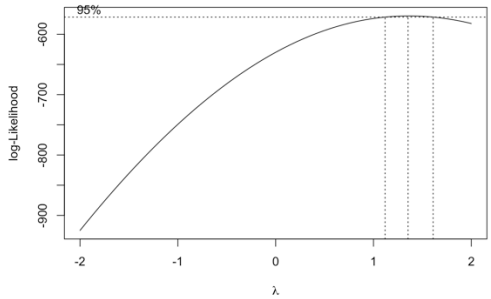


6. From our residuals vs. fitted values plot we can assume there is nonlinearity in the model because there is a clear nonlinear pattern. The Q-Q plot shows a heavy left tail so we can assume the model distribution is non-symmetrical. Also, the left tail goes off the diagonal line of the Q-Q plot so the errors of the model are not normally distributed.



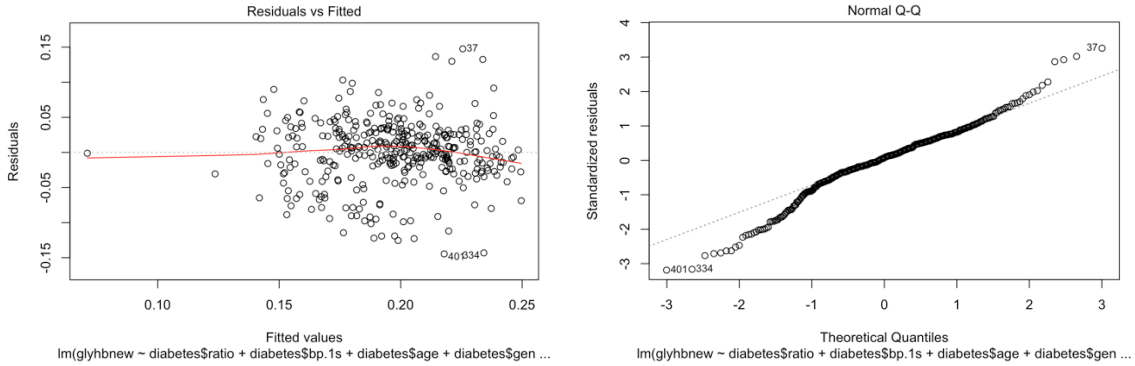


7. Using boxcox() from library MASS, we see that the transformed glyhb data (1/glyhb) is a good choice because it has the smallest $\hat{\lambda}$ that is closest to zero compared to the other transformations log(glyhb) and sqrt(glyhb).



8. Looking at the residuals vs. fitted values for the regresion glyhb* to ratio, bp.1s, age, gender, and frame, we can assume nonlinearity. The Q-Q plot has tails are heavy, but much more symmetrical meaning that the model distribution is symmetrical and the

errors of the model are normally distributed compared to the previous model made with the original glyhb data.



9. In model 2 where glyhb* is regressed to the X variables ratio, bp.1s, age, gender, and frame, the X variable gender has the greatest p-value which means it is the least significant. We will make a new model without the X variable, gender, and check the p-values of the X variables again.

10. In model 3, glyhb* is regressed with the X variables ratio, bp.1s, age, and frame. The X variable, bp.1s has the largest p-value among the other variables, and therefore is the least significant and will also be dropped from the model.

11. Model 4 consists of glyhb* regressed to X variables ratio, age, and frame.

    The model equation for model 4 is (Y=glyhb*):

    $$Y = B_0 + B_1 ratio + B_2 age + B_3 frame.medium + B_4 frame.small + \varepsilon$$

    The fitted regression function for each class of frame is:

    Frame.large: $\hat{Y} = 0.2768197 + (-0.0079466) ratio + (-0.0010782) age$

    Frame.medium: $\hat{Y} = 0.2768197 + (-0.0079466) ratio + (-0.0010782) age + 0.0075518$

    Frame.small: $\hat{Y} = 0.2768197 + (-0.0079466) ratio + (-0.0010782) age + 0.0130291$

    The regression coefficients related to frame would be:
    frame. medium B_3= 0.0075518 and frame.small B_3 = -0.0130291.

12. Model 5 consists of glyhb* regressed to X variables ratio, age, and frame and the interaction between age and frame.

    The model equation is (Y=glyhb*) :

    $$Y = B_0 + B_1 ratio + B_2 age + B_3 frame.medium + B_4 frame.small + B_5 age{:}\, frame.medium + B_6 age{:}\, frame.small + \varepsilon$$

Frame.large: $\hat{Y} = 0.2574393 + (-0.0077235)ratio + (-0.0007324)age$
Frame.medium: $\hat{Y} = 0.2574393 + (-0.0077235)ratio + (-0.0007324)age + 0.04033323frame.medium + (-0.0006623)age{:}\,frame.medium$
Frame.small: $\hat{Y} = 0.2574393 + (-0.0077235)ratio + (-0.0007324)age + 0.024039frame.small + (-0.0001672)age{:}\,frame.small$

Regression coefficients related to frame would be:
Frame.medium B_3 = 0.04033323, frame.small B_4 = 0.0240396,
age:frame.medium B_5 = -0.0006623, age:frame.small B_6 = -0.0001672

13.

	R_p^2	$R_{a,p}^2$	AIC_p	BIC_p
Model 2	0.2502	0.238	-2301.28	-2024.991
Model 3	0.2502	0.24	-2302.25	-2101.977
Model 4	0.2444	0.2363	-2327.17	-2188.798
Model 5	0.252	0.2399	-2327.033	-2050.223

According the R_p^2 criteria, model 5 is the best. According to the $R_{a,p}^2$ criteria, model 3 is the best. According to the AIC_p criteria, model 4 is the best. According to the BIC_p criteria, model 4 is the best.

For the remaining parts:

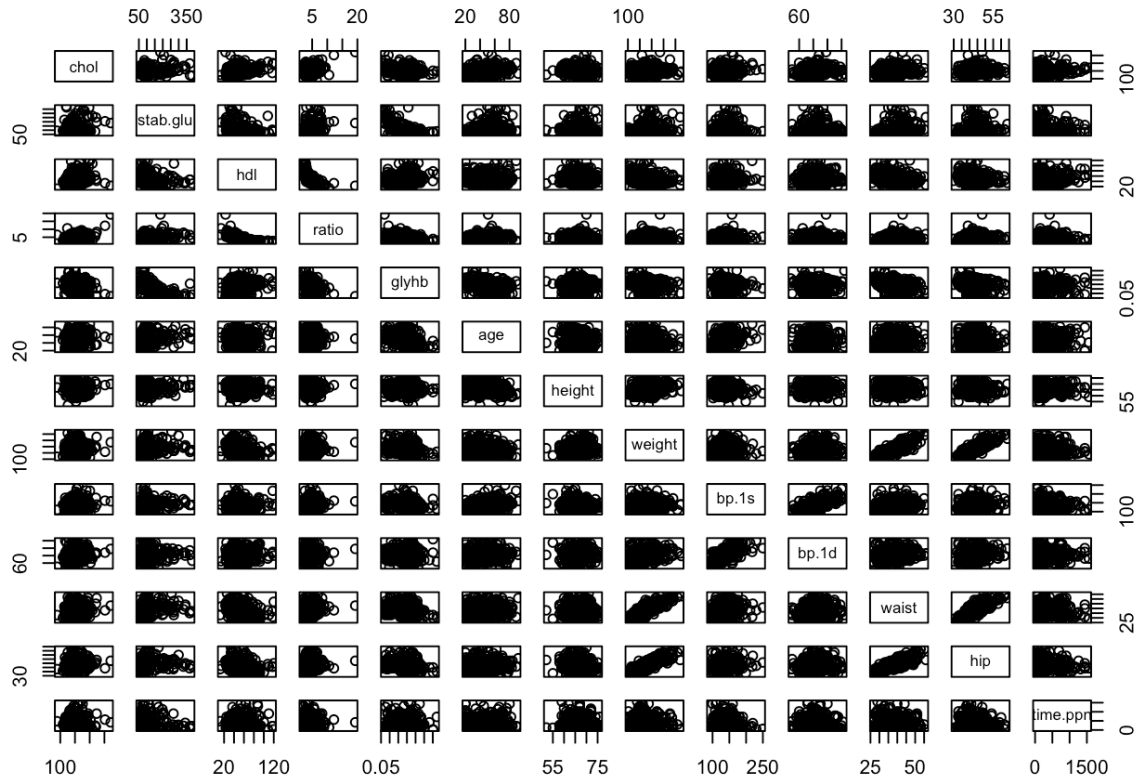
- bp.2s = Second Systolic Blood Pressure
- bp.2d = Second Diastolic Blood Pressure
- chol = Cholesterol
- stab.glu = Stabilized Glucose
- height = height in inches
- weight = weight in pounds
- waist = waist in inches
- hip = hip in inches
- time.ppn = Postprandial Time when Labs were Drawn in minutes
- location = Buckingham or Louisa

14. The column of the variable id has been dropped because it is not an explanatory variable. Variables bp.2s and bp.2d were dropped because they have too many missing values

15. For the rest of the report glyhb has been transformed into its inverse.

16. All cases that had missing values were dropped from the data.

17. Scatterplot matrix and pairwise correlation matrix for quantitative variables in data shows that there is no obvious non-linearity between variables.



18. Model 6 consists of all first-order effects. There are 17 regression coefficients in model. The MSE of the Model 6 is 0.00133.

19. According the R_p^2 criteria, model 16 that has of all X variables is the best. According to the $R_{a,p}^2$ criteria, model 7 that has of X variables stab.glu, ratio, location, age, waist, time.ppn, and chol is the best. According to the BIC_p criteria, model 4 that has of X variables stab.glu, ratio, age, and waist is the best. According to the AIC_p criteria, model 7 is the best.

20. The first-order model to be selected was the model that has the X variables stab.glu, ratio, location, age, waist, time.ppn, and chol. This "best" model according to AIC_p criterion is the same "best" model identified in part 19.
Model equation: $Y = B_0 + B_1 stab.glu + B_2 age + B_3 ratio + B_4 waist + B_5 time.ppn + B_6 location.louisa + B_7 chol + \varepsilon$
Fitted regression function: $\hat{Y} = (3.494\text{e}-01) + (-4.952\text{e}-04) stab.glu + (-6.216e-04) age + (-2.910e-03) ratio + (-1.093e-03) waist + (-1.177e-05) time.ppn + (6.523e-03) location.louisa + (-7.209e-05) chol$

The second-order model to be selected has X variables stab.glu, age, ratio, waist, time.ppn, location, interaction stab.gly:time.ppn, interaction stab.glu:age, and interaction age:ratio.

Model equation: $Y = B_0 + B_1 stab.glu + B_2 age + B_3 ratio + B_4 waist + B_5 time.ppn + B_6 location.louisa + B_7 stab.glu{:}\, time.ppn + B_8 stab.glu{:}\, age + B_9 age{:}\, ratio + \varepsilon$

Fitted regression function: $\hat{Y} = (3.355\text{e}-01) + (-7.876\text{e}-04) stab.glu + (-8.332e-04) age + (2.77e-03) ratio + (-1.034e-03) waist + (3.580e-05) time.ppn + (6.642e-03) location.louisa + (-4.619e-07) stab.glu{:}\, time.ppn + (7.341e-06) stab.glu{:}\, age + (-1.209e-04) age{:}\, ratio$