STA 108 –Diabetes Project Report

Dayanara Lebron-Aldea

Dataset and Purpose:

The diabetes.txt file contains 19 variables giving information about: _._,_ on 403 subjects out of 1046 subjects that were interviewed with the purpose of understanding the prevalence of diabetes and related diseases. This project focuses on building regression models with Glycosolated hemoglobin (glybh) as its response variable. This dataset was priorly subsetted to contain the information of those subjects who had a glybh > 70 as this is a sign of diabetes.

Methodology:

I started by doing some data exploration by looking at the types of variables for glyhb,ratio, bp.1s, age, gender and frame. Histograms and pie-charts were produced to observe the distributions of this variables and their levels. Distributions of several transformations of the response variable were inspected.

A few regression models were constructed by dropping the least significant variables, starting with the models:

Model1: $glyhb_i = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 frame_i + \beta_4 bp.1s_i + \beta_4 ratio_i$

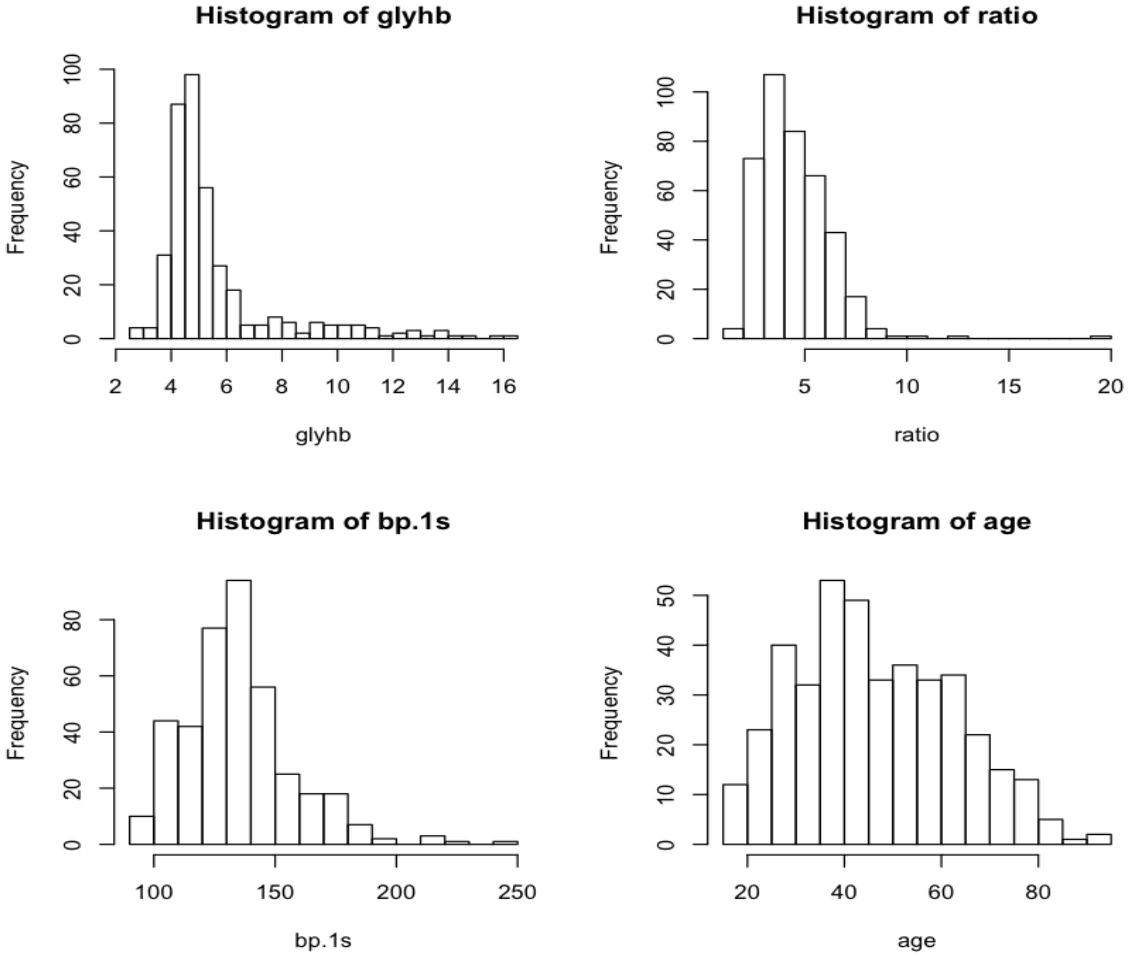
and Model 2 follows includes the same variables as model 1 but with the transformed glyhb as the response variable. Values of $R_i^2, R_{a,p}^2, AIC\ and\ BIC$ were used to decide which model was the best.

Another model (Model6) took into account all variables within the data.frame except: id, bp.2s and bp.2d, with response variable being the selected transformation of glyhb. Model 6 is the model with all first order effects, having 18 predictor variables. Relationships between variables were inspected by using scatterplots and a correlation matrix.

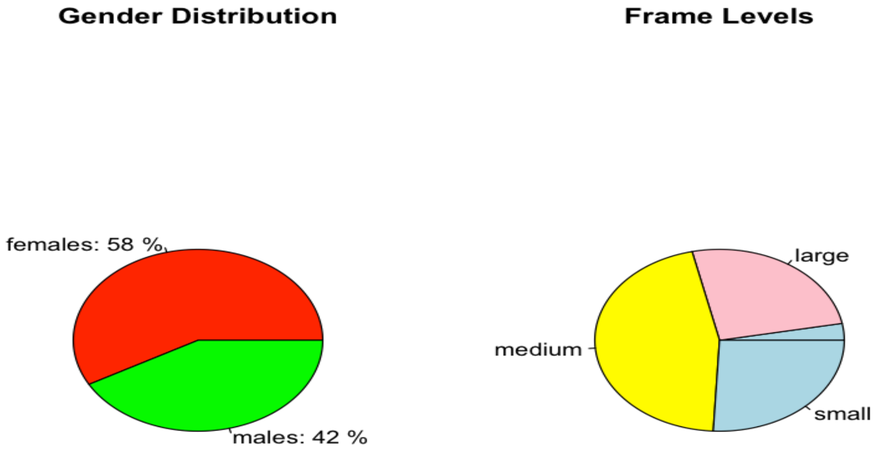
To select the best regression model made our of different combinations of the 18 variables, the regsubsets method from the leaps library was used, and values of R^2 , $R_{a,p}^2, AIC\ and\ BIC$ were inspected.

Results:

The first model was constructed using the quantitative variables: glyhb, ratio, bp.1s and age, and the qualitative variables: gender, and frame. **Histogram1** shows the histogram for the quantitative variables. Age is has roughly a normal distribution, but we see that the age is concentrated between 20 and 65 y/o. Variables ratio, bp1.s (systolic pressure) are skewed to the right as well as the response variable glyhb, as you may remember the response variable should be a normal distribution so that we accomplish the normal error assumption (shape bell) which we are not achieving. here.

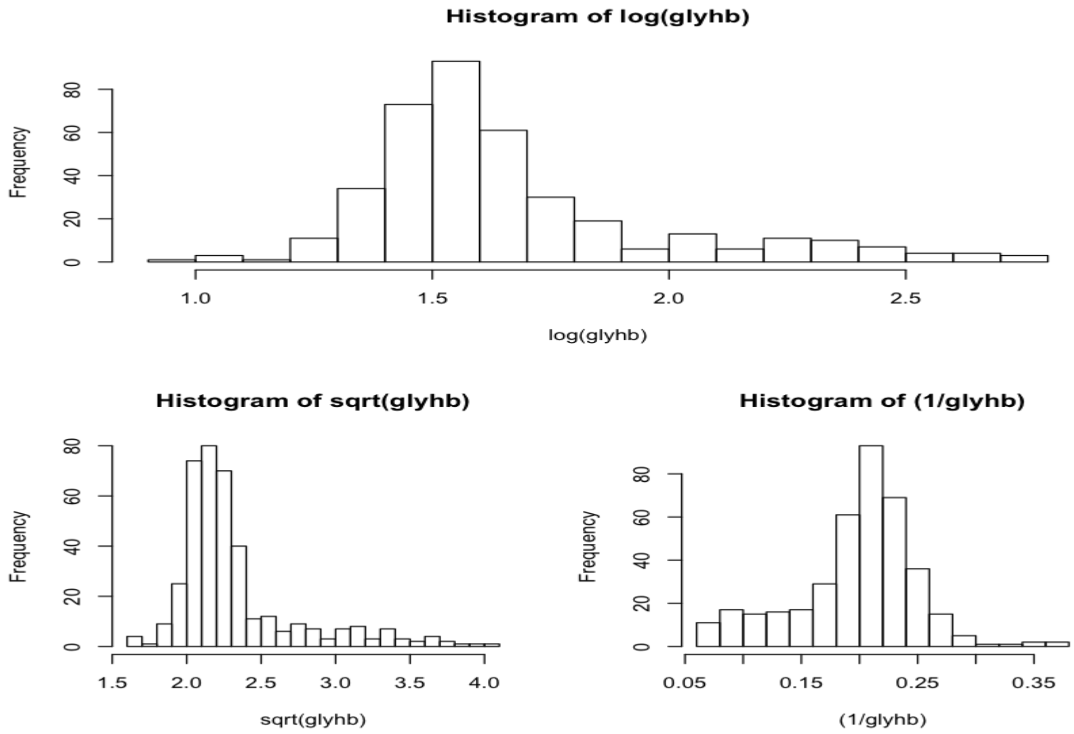


Histogram1: Distribution for quantitative variables

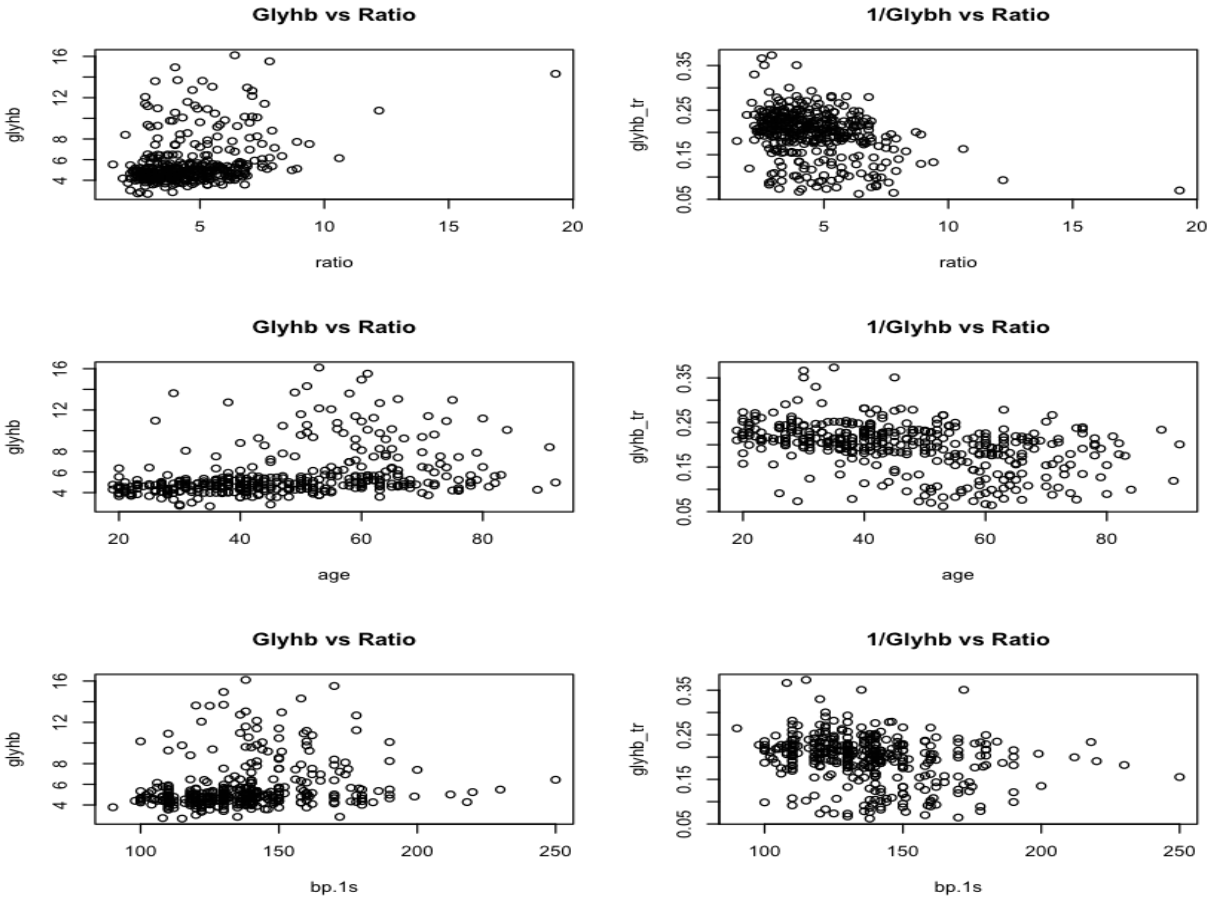


Graph1: Distribution of Qualitative Variables

Through the pie-charts in **Graph1,** we can see that majority of subjects in this dataset are females(58%) and a great portion of subjects have a medium frame. Retaking, on the distribution of glyhb **Histogram2** shows the transformations we took in consideration. Among all, we can see that (1/glyhb) achieves the normal-like distribution. So far this is the primary candidate as the new response variable.

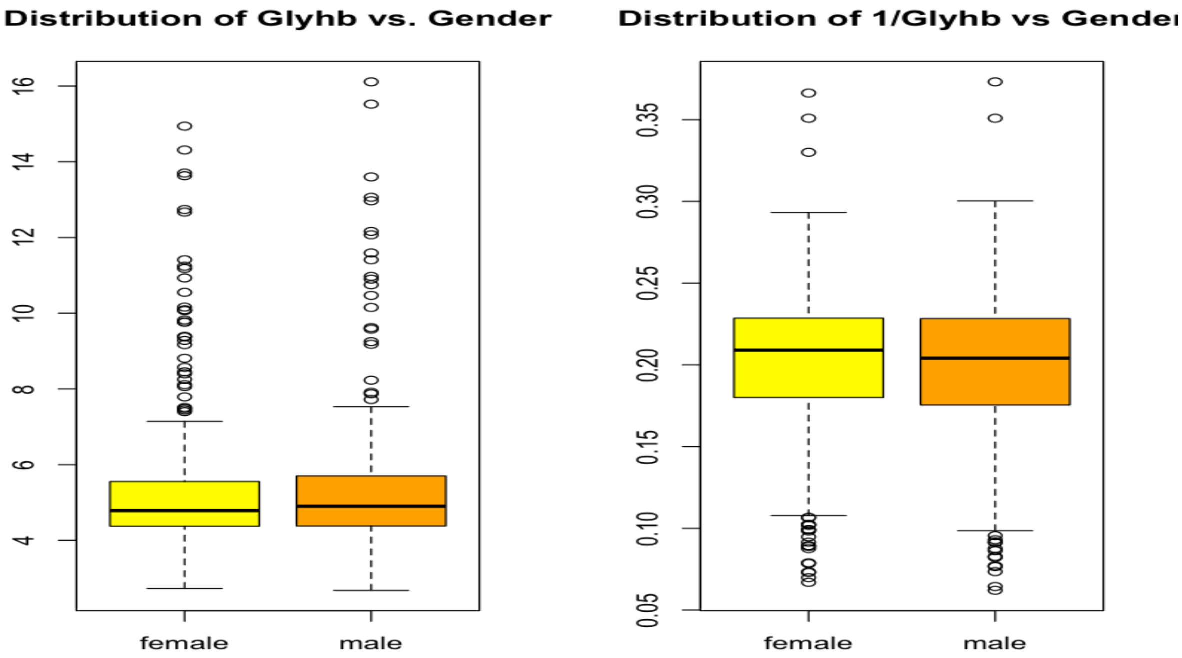


Histogram2: Distributions of transformations

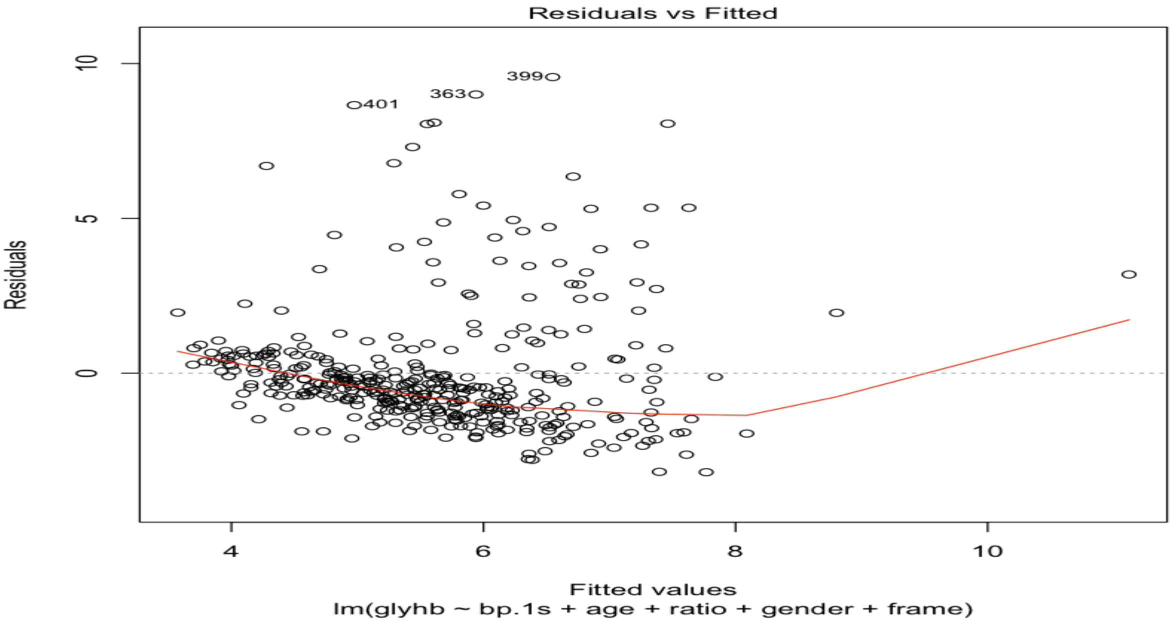


Graphic 2: Shows the scatterplots of glyhb and glybh* against predictor variables.

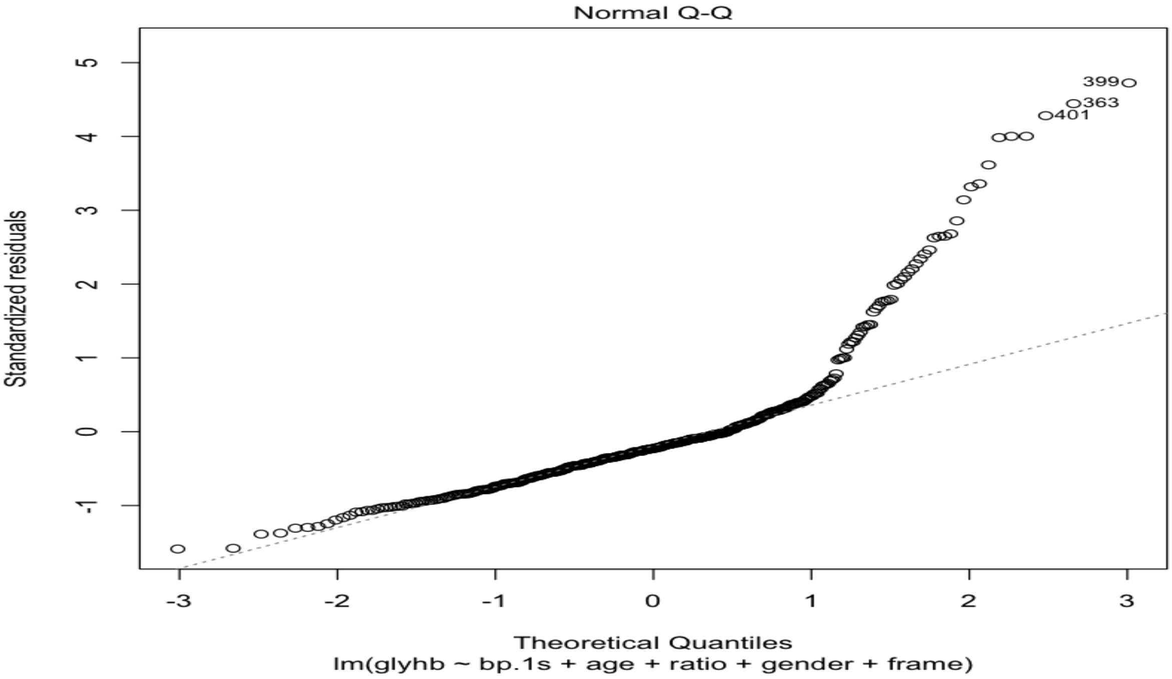
We can see some extreme variables in the histograms agains ratio, nevertheless graphics against 1/glyhb gives a more centered and evenly spread scatters. In the distribution against gender it appears that male had slightly higher levels of glyhb than females. This is still true when inspected against glyhb*.



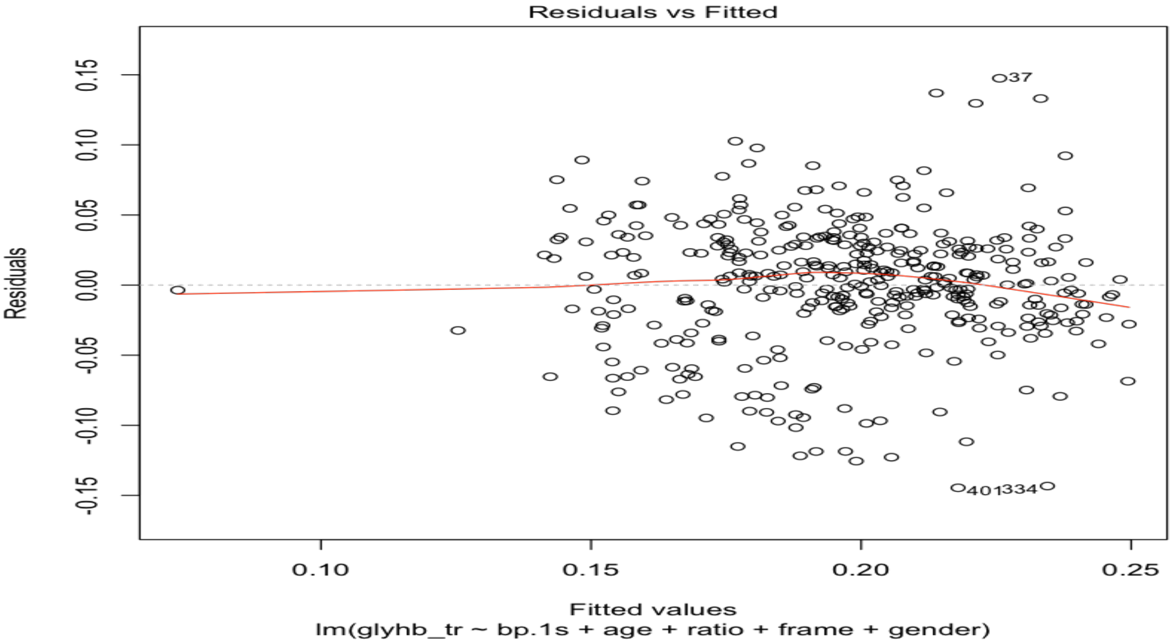
After regressing model1, we can see that the qqplot and residuals vs fitted values plot is not evenly spreaded around 0, not holding the assumptions for the regression model. We see a very heavy tailed qqplot in **Graphic 4**.



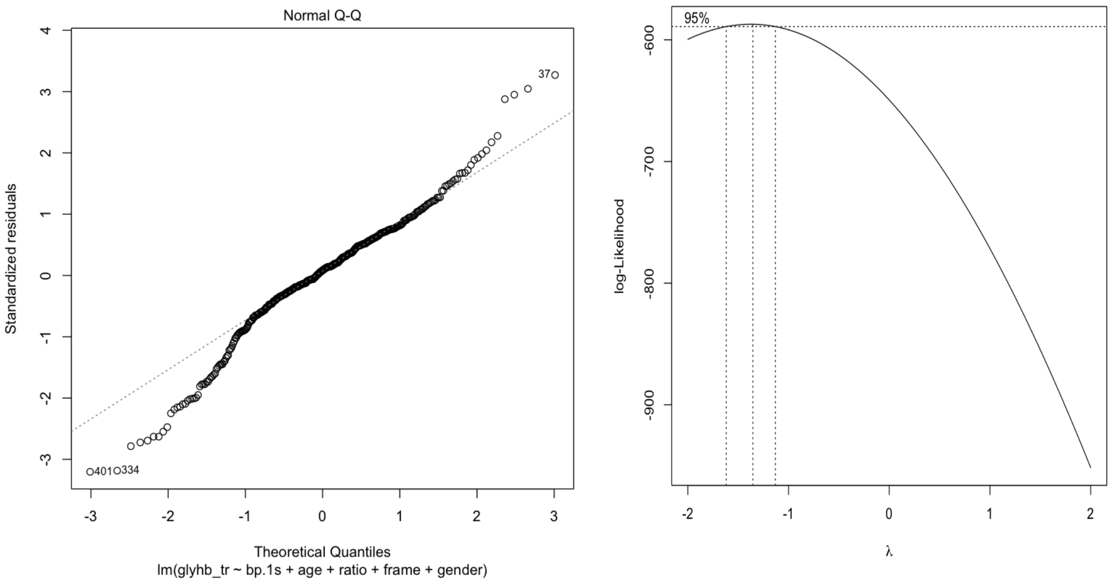
Graphic 3: Residuals vs Fitted values in Model1



Graphic4: QQplot for Model1



Graphic5: Residuals vs Fitted for Model 2.



Graphic 5,6: QQplot for Model2 and BoxCox Transformation

Comparing results of Model 2 and Model 1 we see that the qqplot for Model 2 shows a more linear pattern going from corner to corner of the graph and that residuals are spreaded and centered around 0. Assumptions do hold when using (1/glyhb) as our response variable, from now on we will denote it **glyhb***. The *Boxcox* method was use to corroborate that this was a good choice of transformation. **Graphic 6** shows that the 95% confidence interval for lambda (λ) is centered around -1.2, which we can round to -1. Following the equation of $transf = var^{\lambda}$, we have that this is equivalent to 1/glyhb, so, we made a good choice.

Models:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3068053  0.0215863  14.213  < 2e-16 ***
bp.1s       -0.0001295  0.0001142  -1.134    0.258
age         -0.0009952  0.0001619  -6.146 2.03e-09 ***
ratio       -0.0078796  0.0013852  -5.688 2.58e-08 ***
framelarge  -0.0172390  0.0151605  -1.137    0.256
framemedium -0.0100327  0.0148770  -0.674    0.500
framesmall  -0.0042682  0.0152572  -0.280    0.780
gendermale   0.0014384  0.0048411   0.297    0.767
```

Table 1: **Coefficient Results of Model 2**

From the results of model 2 we can see that there are some variables which are not significant this are frame and gender. Instead of just dropping the least significant variable, I ran two models 1 without gender and the other without frame and looked at their respective values of SSE, the SSE for the model without gender was 0.77 and the model without frame was 0.785, which means that according to this, I should drop frame instead of gender. Nevertheless, I will drop gender, as instructions state and see what happens.

Model 3 is now Model 2 without gender, below is the summary for the model:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.3072474  0.0215090  14.285  < 2e-16 ***
age         -0.0009937  0.0001616  -6.148 2.01e-09 ***
ratio       -0.0078542  0.0013809  -5.688 2.59e-08 ***
framelarge  -0.0169719  0.0151155  -1.123    0.262    
framemedium -0.0100873  0.0148579  -0.679    0.498    
framesmall  -0.0043435  0.0152366  -0.285    0.776    
bp.1s       -0.0001299  0.0001141  -1.138    0.256    
```

Table 2: Coefficient Results of Model 3

Once again, frame and bp.1s are the least significant values, SSE's for a model dropping frame is 0.785, and SSE for model dropping bp.1s is 0.7929. According to this, I dropping frame aids in lowering the residual errors, but instructions state that Model 4 should contain ratio,frame and age, so I will proceed to drop bp.1s instead.

Model 4 follows the following equation:

$$glyhb_i^* = \beta_{0i} + \beta_{1i}age + \beta_{2i}ratio + B_{3i}frame_{large} + B_{4i}frame_{med} + B_{5i}frame_{small} + \varepsilon_i$$

When the subject belongs to the "" frame then the equation is: $glyhb_i^* = \beta_{0i} + \beta_{1i}age + \beta_{2i}ratio + B_{3i}frame_0 + \varepsilon_i$; when the subject belongs to the "large" frame then the equation is: $glyhb_i^* = (\beta_{0i} + B_{3i}) + \beta_{1i}age + \beta_{2i}ratio + \varepsilon_i$;

model for subjets belonging to the medium frame level is :

$$glyhb_i^* = (\beta_{0i} + B_{4i}) + \beta_{1i}age + \beta_{2i}ratio + \varepsilon_i;$$

Subjects that belong to the small frame have a model:

$$glyhb_i^* = (\beta_{0i} + B_{5i}) + \beta_{1i}age + \beta_{2i}ratio + \varepsilon_i;$$

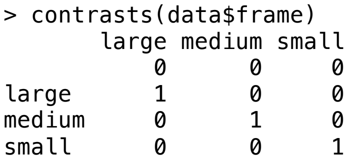
Being 0 the frame to which it does not belong. The coefficients $B_{3i}\ to\ B_{5i}$ refers to how much bigger or smaller it is to the reference one which is "".

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.2960631  0.0166287  17.804  < 2e-16 ***
age         -0.0010848  0.0001461  -7.423 7.47e-13 ***
ratio       -0.0077330  0.0013811  -5.599 4.11e-08 ***
framelarge  -0.0199590  0.0144877  -1.378    0.169    
framemedium -0.0123822  0.0142013  -0.872    0.384    
framesmall  -0.0067563  0.0145654  -0.464    0.643    
```

Table 3: Coefficient Results of Model 4

This means that the intercept for the large frame is ˜ 0.2 times smaller than the “” level. The intercept for medium is -0.012 times smaller. This would probably mean that the category “” takes in individual who have a smaller frame than the small section.

The coding for the frames is here:



```
> contrasts(data$frame)
       large medium small
           0      0     0
large      1      0     0
medium     0      1     0
small      0      0     1
```

For the model with interactions of age against frame, things change a little, equation is now:

$$glyhb_i^* = \beta_{0i} + \beta_{1i}age + \beta_{2i}ratio + B_{3i}frame_{large} + B_{4i}frame_{med} + B_{5i}frame_{small} + B_{6i}age{:}\,frame_{large} + B_{7i}age{:}\,frame_{med} + B_{8i}age{:}\,frame_{small} + \varepsilon_i$$

Depending to which group it belongs then only that coefficient is included in the model, for example for a subject in the large frame the equation is:

$$glyhb_i^* = (\beta_{0i} + B_{3i} + B_{6i}) + \beta_{1i}age + \beta_{2i}ratio + \varepsilon_i$$

* * Results

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.2980768  0.0338565   8.804   <2e-16 ***
age             -0.0011394  0.0005606  -2.032   0.0428 *  
ratio           -0.0075171  0.0013842  -5.431    1e-07 ***
framelarge      -0.0416434  0.0369671  -1.126   0.2607    
framemedium     -0.0010547  0.0350389  -0.030   0.9760    
framesmall      -0.0173176  0.0358082  -0.484   0.6289    
```

```
age:framelarge   0.0004065  0.0006302   0.645   0.5192
age:framemedium -0.0002599  0.0006033  -0.431   0.6668
age:framesmall   0.0002381  0.0006306   0.378   0.7060
```

Table 4: Coefficient Results of Model 5

Model Selection criteria:

Model	R_p^2	$R_{a,p}^2$	AIC_p	BIC_p
Model 2	0.2521	0.2381	-1274.01948	-1238.4636
Model 3	0.2519	0.24	-1275.929	-1244.324
Model 4	0.2469	0.237	-1275.9293	-1244.32419
Model 5	0.2545	0.2388	-1292.14864	-1264.40359

Table 5: AIC, BIC and goodness of fit for Models 2-5

According to **Table 5** I would select Model 5 as my model since it is the one in which R_p^2 is maximixed and AIC is minimized.

Stepwise-Models:

For this part a model with 18 variables was modeled, all results can be seen in the RegSubsets Results secion and the Forward-Stepwise Results section. Here only the final results will be discussed. The MSE for this model is 0.00131. We proceeded to consider the best subset selection using the R function regsubsets() returning the top 1 model subset of all subset sized up to 16. The **Table of criteria values** in the Regsubsets Results section gives the AIC, BIC and goodness of fit for the models. According to this, we could select model #7 which is one of the model that has max(R^2) and min(AIC). Model #7 contains the variables: chol, stab.glu, ratio, locatioLoiusa, age, frame-medium, waist and time. Its R^2 adjusted was 0.509 and its AIC: -2489.448.

Using the forward selection model we started with an empty model and started seeing the difference in AIC when adding more variables. We tried this with and without interaction of all values. The first order model that achieves the highest AIC is glyhb ~ stab.glu + age + ratio + waist + time.ppn + location + chol with a Step: AIC=-2489.45. the model that achives the highest AIC with a second-order model (interaction) is: glyhb ~ stab.glu + age + ratio + waist + time.ppn + location + stab.glu:time.ppn + stab.glu:age with a Step: AIC=-2500.93.

Conclusion:

This project focused on the ethics of model selection via the inspection of validation of assumptions, selection of variables in the model, interpretation of coefficients for models with and w/o interaction, and how to select a good model though the observation of goodness of fit and AIC values. The stepwise procedures were also discussed and all applications were done in R.

Appendix:

RegSubsets Results:

```
 sumsub$rsq
 [1] 0.4208172 0.4699838 0.4921365 0.5050711 0.5122942 0.5154262 0.5183197
 [8] 0.5194573 0.5199636 0.5203697 0.5210870 0.5214980 0.5220113 0.5224668
[15] 0.5226984 0.5227536
```

```
 sumsub$adjr2
 [1] 0.4192644 0.4671342 0.4880298 0.4997206 0.5056857 0.5075255 0.5091324
 [8] 0.5089536 0.5081271 0.5071931 0.5065745 0.5056360 0.5047984 0.5038961
[15] 0.5027554 0.5014241
```

Table of criteria values

```
   (Intercept) chol stab.glu hdl ratio locationLouisa age gendermale height
1            1    0        1   0     0              0   0          0      0
2            1    0        1   0     0              0   1          0      0
3            1    0        1   0     1              0   1          0      0
4            1    0        1   0     1              0   1          0      0
5            1    0        1   0     1              0   1          0      0
6            1    0        1   0     1              1   1          0      0
7            1    1        1   0     1              1   1          0      0
8            1    1        1   0     1              1   1          0      0
9            1    1        1   0     1              1   1          0      0
10           1    1        1   0     1              1   1          0      0
11           1    1        1   0     1              1   1          0      1
12           1    1        1   0     1              1   1          0      1
13           1    1        1   0     1              1   1          0      1
14           1    1        1   0     1              1   1          0      1
15           1    1        1   0     1              1   1          0      1
16           1    1        1   1     1              1   1          0      1

weight framelarge framemedium framesmall bp.1s bp.1d waist hip time.ppn
1       0          0           0          0     0     0     0   0        0
```

```
2       0           0           0          0     0     0     0   0        0
3       0           0           0          0     0     0     0   0        0
4       0           0           0          0     0     0     1   0        0
5       0           0           0          0     0     0     1   0        1
6       0           0           0          0     0     0     1   0        1
7       0           0           0          0     0     0     1   0        1
8       0           0           1          0     0     0     1   0        1
9       0           0           1          0     1     0     1   0        1
10      1           0           1          0     0     0     1   1        1
11      1           0           1          0     0     0     1   1        1
12      1           0           1          0     1     0     1   1        1
13      1           1           1          1     0     0     1   1        1
14      1           1           1          1     1     0     1   1        1
15      1           1           1          1     1     1     1   1        1
16      1           1           1          1     1     1     1   1        1
     R^2 R^2_a       BIC       AIC
1  0.421 0.419 -2424.468 -2432.322
2  0.470 0.467 -2451.808 -2463.588
3  0.492 0.488 -2461.891 -2477.599
4  0.505 0.500 -2465.639 -2485.273
5  0.512 0.506 -2465.225 -2488.787
6  0.515 0.508 -2461.714 -2489.202
7  0.518 0.509 -2458.033 -2489.448
8  0.519 0.509 -2452.993 -2488.335
9  0.520 0.508 -2447.461 -2486.730
10 0.520 0.507 -2441.852 -2485.048
11 0.521 0.507 -2436.486 -2483.609
12 0.521 0.506 -2430.881 -2481.931
13 0.522 0.505 -2425.357 -2480.333
14 0.522 0.504 -2419.787 -2478.691
15 0.523 0.503 -2414.042 -2476.873
16 0.523 0.501 -2408.159 -2474.916

index.best=c(which.max(sumsub$rsq),which.max(sumsub$adjr2),which.min(bic),which.min(aic))
> index.best
[1] 16  7  4  7
```

Forward Stepwise Selection Results

```
> step = stepAIC(fit,scope=list(upper=formula(fita), lower=~1), direction="forward", k=2)
Start:  AIC=-2229.52
glyhb ~ 1

           Df Sum of Sq     RSS     AIC
+ stab.glu  1   0.41092 0.56556 -2432.3
+ age       1   0.15568 0.82080 -2292.7
+ ratio     1   0.12082 0.85567 -2277.1
+ waist     1   0.09411 0.88238 -2265.5
+ chol      1   0.06777 0.90872 -2254.5
+ bp.1s     1   0.05572 0.92077 -2249.6
+ frame     3   0.05360 0.92289 -2244.7
+ weight    1   0.04274 0.93375 -2244.3
+ hip       1   0.03909 0.93739 -2242.8
+ hdl       1   0.02778 0.94871 -2238.3
+ location  1   0.00751 0.96898 -2230.4
<none>                  0.97649 -2229.5
+ bp.1d     1   0.00323 0.97326 -2228.8
+ gender    1   0.00321 0.97328 -2228.8
+ height    1   0.00185 0.97464 -2228.2
+ time.ppn  1   0.00127 0.97522 -2228.0
```

```
Step:  AIC=-2432.32
glyhb ~ stab.glu

           Df Sum of Sq     RSS     AIC
+ age       1  0.048011 0.51755 -2463.6
+ ratio     1  0.027813 0.53775 -2449.2
+ waist     1  0.027375 0.53819 -2448.9
+ chol      1  0.022440 0.54312 -2445.5
+ bp.1s     1  0.017978 0.54759 -2442.4
+ hip       1  0.012829 0.55274 -2438.9
+ frame     3  0.016626 0.54894 -2437.5
+ weight    1  0.008719 0.55684 -2436.2
+ location  1  0.005918 0.55965 -2434.3
+ time.ppn  1  0.004603 0.56096 -2433.4
+ hdl       1  0.004347 0.56122 -2433.2
<none>                  0.56556 -2432.3
+ bp.1d     1  0.001479 0.56409 -2431.3
+ height    1  0.000155 0.56541 -2430.4
+ gender    1  0.000004 0.56556 -2430.3

Step:  AIC=-2463.59
glyhb ~ stab.glu + age

           Df Sum of Sq     RSS     AIC
+ ratio     1 0.0216318 0.49592 -2477.6
+ waist     1 0.0213292 0.49622 -2477.4
+ hip       1 0.0150624 0.50249 -2472.7
+ weight    1 0.0147622 0.50279 -2472.4
+ chol      1 0.0108483 0.50671 -2469.5
+ hdl       1 0.0068618 0.51069 -2466.6
+ time.ppn  1 0.0054227 0.51213 -2465.5
+ location  1 0.0050870 0.51247 -2465.3
+ frame     3 0.0099879 0.50757 -2464.9
<none>                  0.51755 -2463.6
+ bp.1s     1 0.0019617 0.51559 -2463.0
+ bp.1d     1 0.0006436 0.51691 -2462.1
+ height    1 0.0002335 0.51732 -2461.8
+ gender    1 0.0002150 0.51734 -2461.7

Step:  AIC=-2477.6
glyhb ~ stab.glu + age + ratio

           Df Sum of Sq     RSS     AIC
+ waist     1 0.0126305 0.48329 -2485.3
+ hip       1 0.0096454 0.48628 -2483.0
+ weight    1 0.0075893 0.48833 -2481.4
+ time.ppn  1 0.0062093 0.48971 -2480.3
+ location  1 0.0053771 0.49054 -2479.7
<none>                  0.49592 -2477.6
+ chol      1 0.0018782 0.49404 -2477.0
+ bp.1s     1 0.0015866 0.49433 -2476.8
+ gender    1 0.0006790 0.49524 -2476.1
+ hdl       1 0.0006232 0.49530 -2476.1
+ bp.1d     1 0.0004604 0.49546 -2475.9
+ height    1 0.0000279 0.49589 -2475.6
+ frame     3 0.0048179 0.49110 -2475.3

Step:  AIC=-2485.27
glyhb ~ stab.glu + age + ratio + waist

           Df Sum of Sq     RSS     AIC
+ time.ppn  1 0.0070532 0.47624 -2488.8
+ location  1 0.0044096 0.47888 -2486.7
<none>                  0.48329 -2485.3
```

```
+ chol      1 0.0024685 0.48082 -2485.2
+ hdl       1 0.0016207 0.48167 -2484.5
+ bp.1s     1 0.0005589 0.48273 -2483.7
+ weight    1 0.0003401 0.48295 -2483.5
+ gender    1 0.0002404 0.48305 -2483.5
+ hip       1 0.0000654 0.48323 -2483.3
+ bp.1d     1 0.0000104 0.48328 -2483.3
+ height    1 0.0000081 0.48328 -2483.3
+ frame     3 0.0021537 0.48114 -2480.9

Step:  AIC=-2488.79
glyhb ~ stab.glu + age + ratio + waist + time.ppn

           Df  Sum of Sq     RSS     AIC
+ location  1 0.00305830 0.47318 -2489.2
<none>                   0.47624 -2488.8
+ chol      1 0.00220483 0.47403 -2488.5
+ hdl       1 0.00131073 0.47493 -2487.8
+ bp.1s     1 0.00084035 0.47540 -2487.4
+ gender    1 0.00035063 0.47589 -2487.1
+ weight    1 0.00029865 0.47594 -2487.0
+ hip       1 0.00023533 0.47600 -2487.0
+ bp.1d     1 0.00006398 0.47617 -2486.8
+ height    1 0.00000623 0.47623 -2486.8
+ frame     3 0.00164350 0.47459 -2484.1

Step:  AIC=-2489.2
glyhb ~ stab.glu + age + ratio + waist + time.ppn + location

         Df  Sum of Sq     RSS     AIC
+ chol    1 0.00282554 0.47035 -2489.4
<none>                 0.47318 -2489.2
+ hdl     1 0.00137535 0.47180 -2488.3
+ bp.1s   1 0.00091832 0.47226 -2487.9
+ hip     1 0.00057535 0.47260 -2487.7
+ gender  1 0.00035251 0.47283 -2487.5
+ weight  1 0.00024723 0.47293 -2487.4
+ bp.1d   1 0.00016584 0.47301 -2487.3
+ height  1 0.00001341 0.47317 -2487.2
+ frame   3 0.00228351 0.47090 -2485.0

Step:  AIC=-2489.45
glyhb ~ stab.glu + age + ratio + waist + time.ppn + location + 
    chol

         Df  Sum of Sq     RSS     AIC
<none>                 0.47035 -2489.4
+ bp.1s   1 0.00060561 0.46975 -2487.9
+ hip     1 0.00048422 0.46987 -2487.8
+ weight  1 0.00017961 0.47017 -2487.6
+ gender  1 0.00016179 0.47019 -2487.6
+ hdl     1 0.00007810 0.47028 -2487.5
+ bp.1d   1 0.00001438 0.47034 -2487.5
+ height  1 0.00000178 0.47035 -2487.4
+ frame   3 0.00187652 0.46848 -2484.9
>
```

Forward Stepwise Selection with Interactions

```
> step = stepAIC(fit.0,scope=list(upper=formula(fit.2), lower=~1),direction="forward", k=2)
Start:  AIC=-2229.52
glyhb ~ 1

           Df Sum of Sq     RSS     AIC
+ stab.glu  1   0.41092 0.56556 -2432.3
+ age       1   0.15568 0.82080 -2292.7
+ ratio     1   0.12082 0.85567 -2277.1
+ waist     1   0.09411 0.88238 -2265.5
+ chol      1   0.06777 0.90872 -2254.5
+ bp.1s     1   0.05572 0.92077 -2249.6
+ frame     3   0.05360 0.92289 -2244.7
+ weight    1   0.04274 0.93375 -2244.3
+ hip       1   0.03909 0.93739 -2242.8
+ hdl       1   0.02778 0.94871 -2238.3
+ location  1   0.00751 0.96898 -2230.4
<none>                  0.97649 -2229.5
+ bp.1d     1   0.00323 0.97326 -2228.8
+ gender    1   0.00321 0.97328 -2228.8
+ height    1   0.00185 0.97464 -2228.2
+ time.ppn  1   0.00127 0.97522 -2228.0

Step:  AIC=-2432.32
glyhb ~ stab.glu

           Df Sum of Sq     RSS     AIC
+ age       1  0.048011 0.51755 -2463.6
+ ratio     1  0.027813 0.53775 -2449.2
+ waist     1  0.027375 0.53819 -2448.9
+ chol      1  0.022440 0.54312 -2445.5
+ bp.1s     1  0.017978 0.54759 -2442.4
+ hip       1  0.012829 0.55274 -2438.9
+ frame     3  0.016626 0.54894 -2437.5
+ weight    1  0.008719 0.55684 -2436.2
+ location  1  0.005918 0.55965 -2434.3
+ time.ppn  1  0.004603 0.56096 -2433.4
+ hdl       1  0.004347 0.56122 -2433.2
<none>                  0.56556 -2432.3
+ bp.1d     1  0.001479 0.56409 -2431.3
+ height    1  0.000155 0.56541 -2430.4
+ gender    1  0.000004 0.56556 -2430.3

Step:  AIC=-2463.59
glyhb ~ stab.glu + age

               Df Sum of Sq     RSS     AIC
+ ratio         1 0.0216318 0.49592 -2477.6
+ waist         1 0.0213292 0.49622 -2477.4
+ hip           1 0.0150624 0.50249 -2472.7
+ weight        1 0.0147622 0.50279 -2472.4
+ chol          1 0.0108483 0.50671 -2469.5
+ stab.glu:age  1 0.0085927 0.50896 -2467.9
+ hdl           1 0.0068618 0.51069 -2466.6
+ time.ppn      1 0.0054227 0.51213 -2465.5
+ location      1 0.0050870 0.51247 -2465.3
+ frame         3 0.0099879 0.50757 -2464.9
<none>                      0.51755 -2463.6
+ bp.1s         1 0.0019617 0.51559 -2463.0
+ bp.1d         1 0.0006436 0.51691 -2462.1
+ height        1 0.0002335 0.51732 -2461.8
+ gender        1 0.0002150 0.51734 -2461.7

Step:  AIC=-2477.6
glyhb ~ stab.glu + age + ratio
```

```
                 Df Sum of Sq     RSS     AIC
+ waist           1 0.0126305 0.48329 -2485.3
+ hip             1 0.0096454 0.48628 -2483.0
+ weight          1 0.0075893 0.48833 -2481.4
+ time.ppn        1 0.0062093 0.48971 -2480.3
+ stab.glu:age    1 0.0060417 0.48988 -2480.2
+ location        1 0.0053771 0.49054 -2479.7
<none>                        0.49592 -2477.6
+ chol            1 0.0018782 0.49404 -2477.0
+ bp.1s           1 0.0015866 0.49433 -2476.8
+ gender          1 0.0006790 0.49524 -2476.1
+ hdl             1 0.0006232 0.49530 -2476.1
+ bp.1d           1 0.0004604 0.49546 -2475.9
+ ratio:age       1 0.0001121 0.49581 -2475.7
+ height          1 0.0000279 0.49589 -2475.6
+ stab.glu:ratio  1 0.0000012 0.49592 -2475.6
+ frame           3 0.0048179 0.49110 -2475.3

Step:  AIC=-2485.27
glyhb ~ stab.glu + age + ratio + waist

                 Df Sum of Sq     RSS     AIC
+ time.ppn        1 0.0070532 0.47624 -2488.8
+ location        1 0.0044096 0.47888 -2486.7
+ stab.glu:age    1 0.0043839 0.47891 -2486.7
<none>                        0.48329 -2485.3
+ chol            1 0.0024685 0.48082 -2485.2
+ hdl             1 0.0016207 0.48167 -2484.5
+ age:waist       1 0.0010309 0.48226 -2484.1
+ bp.1s           1 0.0005589 0.48273 -2483.7
+ ratio:age       1 0.0003543 0.48294 -2483.6
+ weight          1 0.0003401 0.48295 -2483.5
+ stab.glu:ratio  1 0.0002685 0.48302 -2483.5
+ gender          1 0.0002404 0.48305 -2483.5
+ ratio:waist     1 0.0001829 0.48311 -2483.4
+ hip             1 0.0000654 0.48323 -2483.3
+ stab.glu:waist  1 0.0000122 0.48328 -2483.3
+ bp.1d           1 0.0000104 0.48328 -2483.3
+ height          1 0.0000081 0.48328 -2483.3
+ frame           3 0.0021537 0.48114 -2480.9

Step:  AIC=-2488.79
glyhb ~ stab.glu + age + ratio + waist + time.ppn

                    Df Sum of Sq     RSS     AIC
+ stab.glu:time.ppn  1 0.0123200 0.46392 -2496.6
+ stab.glu:age       1 0.0054193 0.47082 -2491.1
+ age:time.ppn       1 0.0044917 0.47175 -2490.3
+ location           1 0.0030583 0.47318 -2489.2
<none>                           0.47624 -2488.8
+ chol               1 0.0022048 0.47403 -2488.5
+ ratio:time.ppn     1 0.0015812 0.47466 -2488.0
+ waist:time.ppn     1 0.0014917 0.47475 -2488.0
+ hdl                1 0.0013107 0.47493 -2487.8
+ bp.1s              1 0.0008403 0.47540 -2487.4
+ age:waist          1 0.0006240 0.47561 -2487.3
+ ratio:age          1 0.0005877 0.47565 -2487.2
+ ratio:waist        1 0.0004896 0.47575 -2487.2
+ gender             1 0.0003506 0.47589 -2487.1
+ weight             1 0.0002986 0.47594 -2487.0
+ hip                1 0.0002353 0.47600 -2487.0
+ stab.glu:ratio     1 0.0001482 0.47609 -2486.9
+ bp.1d              1 0.0000640 0.47617 -2486.8
+ stab.glu:waist     1 0.0000180 0.47622 -2486.8
+ height             1 0.0000062 0.47623 -2486.8
+ frame              3 0.0016435 0.47459 -2484.1

Step:  AIC=-2496.62
glyhb ~ stab.glu + age + ratio + waist + time.ppn + stab.glu:time.ppn

                 Df Sum of Sq     RSS     AIC
+ stab.glu:age    1 0.0056013 0.45832 -2499.2
+ location        1 0.0041142 0.45980 -2498.0
<none>                        0.46392 -2496.6
+ chol            1 0.0014457 0.46247 -2495.8
+ hdl             1 0.0010969 0.46282 -2495.5
+ age:waist       1 0.0009984 0.46292 -2495.4
+ ratio:age       1 0.0008391 0.46308 -2495.3
+ bp.1s           1 0.0005213 0.46340 -2495.0
+ gender          1 0.0003528 0.46356 -2494.9
+ age:time.ppn    1 0.0003491 0.46357 -2494.9
+ stab.glu:waist  1 0.0002772 0.46364 -2494.8
+ weight          1 0.0002020 0.46372 -2494.8
+ waist:time.ppn  1 0.0001853 0.46373 -2494.8
+ ratio:waist     1 0.0001433 0.46377 -2494.7
+ hip             1 0.0001221 0.46380 -2494.7
+ height          1 0.0000660 0.46385 -2494.7
+ stab.glu:ratio  1 0.0000352 0.46388 -2494.6
+ ratio:time.ppn  1 0.0000057 0.46391 -2494.6
+ bp.1d           1 0.0000010 0.46392 -2494.6
+ frame           3 0.0014600 0.46246 -2491.8

Step:  AIC=-2499.17
glyhb ~ stab.glu + age + ratio + waist + time.ppn + stab.glu:time.ppn + 
    stab.glu:age

                 Df Sum of Sq     RSS     AIC
+ location        1 0.0045693 0.45375 -2500.9
<none>                        0.45832 -2499.2
+ age:waist       1 0.0018627 0.45645 -2498.7
+ ratio:age       1 0.0017316 0.45658 -2498.6
+ chol            1 0.0016677 0.45665 -2498.5
+ hdl             1 0.0011538 0.45716 -2498.1
+ weight          1 0.0004355 0.45788 -2497.5
+ bp.1s           1 0.0004344 0.45788 -2497.5
+ gender          1 0.0003922 0.45792 -2497.5
+ stab.glu:waist  1 0.0003545 0.45796 -2497.5
+ waist:time.ppn  1 0.0001998 0.45812 -2497.3
+ age:time.ppn    1 0.0001740 0.45814 -2497.3
+ hip             1 0.0001395 0.45818 -2497.3
+ ratio:waist     1 0.0000921 0.45822 -2497.2
+ stab.glu:ratio  1 0.0000834 0.45823 -2497.2
+ ratio:time.ppn  1 0.0000589 0.45826 -2497.2
+ height          1 0.0000199 0.45830 -2497.2
+ bp.1d           1 0.0000025 0.45831 -2497.2
+ frame           3 0.0012466 0.45707 -2494.2

Step:  AIC=-2500.93
glyhb ~ stab.glu + age + ratio + waist + time.ppn + location + 
    stab.glu:time.ppn + stab.glu:age

                    Df  Sum of Sq     RSS     AIC
<none>                            0.45375 -2500.9
+ chol               1 0.00230776 0.45144 -2500.8
+ age:waist          1 0.00187008 0.45188 -2500.5
+ ratio:age          1 0.00141153 0.45234 -2500.1
+ stab.glu:location  1 0.00136504 0.45238 -2500.1
+ location:age       1 0.00128397 0.45246 -2500.0
+ hdl                1 0.00122044 0.45253 -2499.9
+ hip                1 0.00049129 0.45326 -2499.3
+ bp.1s              1 0.00048733 0.45326 -2499.3
+ gender             1 0.00039636 0.45335 -2499.3
+ location:time.ppn  1 0.00037483 0.45337 -2499.2
+ weight             1 0.00036366 0.45338 -2499.2
+ ratio:waist        1 0.00033615 0.45341 -2499.2
+ ratio:location     1 0.00028394 0.45346 -2499.2
+ age:time.ppn       1 0.00026259 0.45348 -2499.2
+ waist:time.ppn     1 0.00012255 0.45362 -2499.0
+ stab.glu:waist     1 0.00006659 0.45368 -2499.0
+ bp.1d              1 0.00005233 0.45369 -2499.0
+ ratio:time.ppn     1 0.00002132 0.45373 -2498.9
+ stab.glu:ratio     1 0.00001300 0.45373 -2498.9
+ height             1 0.00000896 0.45374 -2498.9
+ location:waist     1 0.00000035 0.45375 -2498.9
+ frame              3 0.00196157 0.45179 -2496.6
```

Code

```
#1) Read data
diab<-read.table("diabetes.txt", header=T)
attach(diab)
diab<-droplevels(diab)

#2) Type of variables: glyhb, ratio, bp.1s, age, gender, frame
#glyhb, ratio, bp.1s, age are quantitative variables, and gender and frame are qualitative variables.

str(diab)

par(mfrow=c(1,2))
n=dim(diab)[1]
pct=round(100*table(diab$gender)/n)
labels=c("females:","males:")
lbls=paste(labels,pct)
lbls=paste(lbls,'%',sep=' ')
pie(table(diab$gender),labels=lbls, col=c("red","green"), main="Gender Distribution")
pie(table(droplevels(diab$frame)),col=c("light blue","pink","yellow"), main="Frame Levels")

###COMMENT ON DISTRIBUTION


par(mfrow=c(2,2))
hist(glyhb, breaks=20)
hist(ratio, breaks=20)
hist(bp.1s, breaks=20)
hist(age,breaks=20)

#COMMENT ON DISTRIBUTION

#3)Draw histograms for different transformations
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(log(glyhb), breaks=20)
hist(sqrt(glyhb), breaks=20)
hist((1/glyhb),breaks=20)

#1/glybh appears to have a normal distribution
glyhb_tr=(1/glyhb)

#4)Scatter plot matrix of predictor variables: ratio, bp.1s, age against glybh and glybh_tr
par(mfrow=c(3,2))
```

```
qvar=c(ratio,bp.1s,age)
plot(ratio, glyhb, main="Glyhb vs Ratio")
plot(ratio, glyhb_tr, main="1/Glybh vs Ratio")
plot(age, glyhb, main="Glyhb vs Ratio")
plot(age, glyhb_tr, main="1/Glyhb vs Ratio")
plot(bp.1s, glyhb, main="Glyhb vs Ratio")
plot(bp.1s, glyhb_tr, main="1/Glyhb vs Ratio")


#5) Distribution of glyhb for males and females
par(mfrow=c(1,2))
boxplot(glyhb~gender, col=c("yellow","orange"),main="Distribution of Glyhb vs. Gender")
boxplot(glyhb_tr~gender,col=c("yellow","orange"), main="Distribution of 1/Glyhb vs Gender")

#What do you observe in th boxplots?


#6) Regression of glyhb to ratio, bp.1s, gender and frame
mod1<-lm(glyhb~bp.1s+age+ratio+gender+frame)
summary(mod1)

#Residuals qqplot and residuals vs fitted values plots
plot(mod1)

#Do model assumptions hold?
No.

#7 Using box-cox transformation to check what transformation is a good choice
library(MASS)
boxcox(glyhb~bp.1s+age+ratio+gender+frame)

#In this plot the lambda that maximixes the log-likelihood function is about -1.3 we can
round this to -1
#our choice of glyhb is correct to be 1/glyhb.

#9 From model 2 dropped gender
mod2<-lm(glyhb_tr~bp.1s+age+ratio+frame+gender)
#Anova with this model kept SSE at 0.777 excludng gender
#Anova with model without frame has SSE at 0.785 so according to anova I will drop gender
anova(mod2)
summary(mod2)

plot(mod2)


#9) mod3<-lm(glyhb_tr~age+ratio+frame+bp.1s)
anova(mod3)
summary(mod3)

AIC(mod3)
BIC(mod3)
#Dropping frame I get a SSe of 0.785
#Dropping bp.1s I get an SSE of 0.7929, It would of been better to drop frame

mod4<-lm(glyhb_tr~age+ratio+frame)
summary(mod4)
AIC(mod4)
BIC(mod4)

mod5<-lm(glyhb_tr~age+ratio+frame+age:frame)
summary(mod5)
AIC(mod4)
BIC(mod4)

#Drop id, bp.2s, bp.2d from data
```

```
drops=c("id","bp.2s","bp.2d")
diab<-diab[,!(names(diab)%in%drops)]

#glyhb values are the inverse of the function
diab$glyhb=(1/glyhb)

#Drop cases having NA's
index.na=apply(is.na(diab),1,any)
diab_2=diab[index.na==FALSE,]
any(is.na(diab_2))
table(diab_2$frame)

#17. Draw scatterplot matrix and obtain the pairwise correlation matrix for all
quantitative variables in the data.
#Comment on their relationships

#drop the factor variables
diabq=diab_2[,!(sapply(diab_2,class)%in%'factor')]

#compute correlation matrix
cor(diabq)


#Draw scatterplot matrix
plot(diab_2)

mod6<-lm(glyhb~., data=diab_2)
summary(mod6)

#install.packages(leaps)
library(leaps)
subset=regsubsets(glyhb~.,data=diab_2, nbest=1, nvmax=16)
sumsub=summary(subset)
sumsub$rsq
sumsub$adjr2

## BIC
## sample size
n=nrow(diab_2)
## number of coefficients in each model: p
p.m=as.integer(as.numeric(rownames(sumsub$which))+1)
bic = n*log(sumsub$rss/n) + (log(n))*p.m

## AIC
aic = n*log(sumsub$rss/n) + 2*p.m

## table of all criteria values
sumtable = cbind(sumsub$which,sumsub$rsq,sumsub$adjr2,bic,aic)
colnames(sumtable)=c(colnames(sumsub$which), "R^2", "R^2_a", "BIC", "AIC")
round(sumtable,3) ## round the results to three decimals

## get the index for the best model within each size group (the first one in the group)
index.best=c(which.max(sumsub$rsq),which.max(sumsub$adjr2),which.min(bic),which.min(aic))


#stepwise procedure-forward selection
##### Forward selection  ###################
fit = lm(glyhb~1, data=diab_2)  ## initial model: none-model with only intercept term
fita = lm(glyhb~., data=diab_2)
step = stepAIC(fit,scope=list(upper=formula(fita), lower=~1), direction="forward", k=2)

##### Forward selection with interactions ##########
fit.0 =lm(glyhb~1, data=diab_2) ## none-model without X variable
fit.2 = lm(glyhb~.^2, data=diab_2)  ## full model with all 2-way interactions
step = stepAIC(fit.0,scope=list(upper=formula(fit.2), lower=~1),direction="forward", k=2)
```