Jared Yu

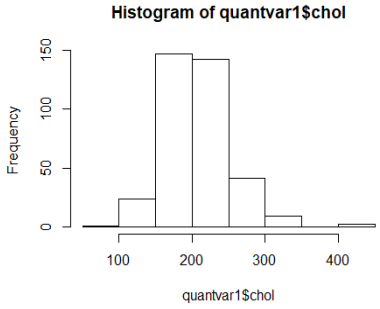
STA 108

Project 2

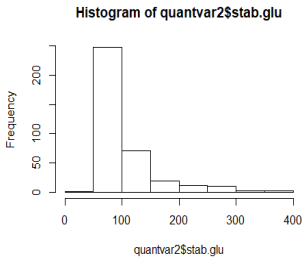
12/5/17

Project 2

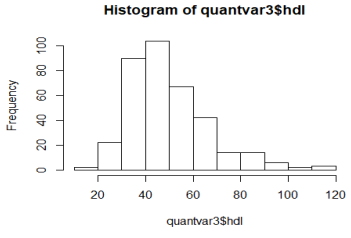
1. Among all the variables, the following are quantitative variables: chol, stab.glu, hdl, ratio, glyhb, age, height, weight, bp.1s, bp.1d, waist, hip, and time.ppn. The qualitative variables are: location, gender, and frame.
    a. Histogram for 'chol': The distribution for the histogram of the variable 'chol' appears unimodal, with most of the data around 200. It also appears somewhat skewed to the right. There appears to be some more outlying data around 400 also.



    i.
    b. Histogram for 'stab.glu': The distribution appears unimodal, and mostly below or near 100. It is also heavily skewed towards the right side. The spread towards the right goes to 400, while on the left it stops at 0.



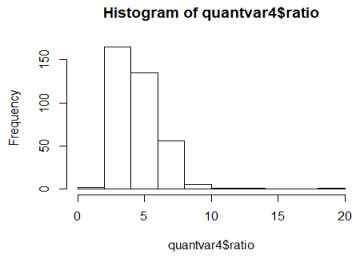
    i.
    c. Histogram for 'hdl': The distribution appears unimodal, with most of the data collecting around 40-50. There is also a skew to the right. It stretches much further to the right than it does to the left. On the left it goes to 0, but to the right it goes to 120.



    i.

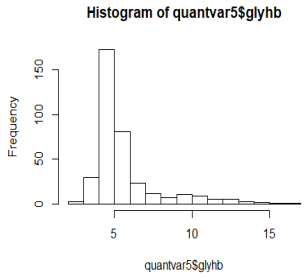
d. Histogram for ‘ratio’: The distribution appears unimodal, with most of the data collecting around 5 or slightly below. The distribution is also skewed to the right. It stretches much further to the right than it does left.

i.



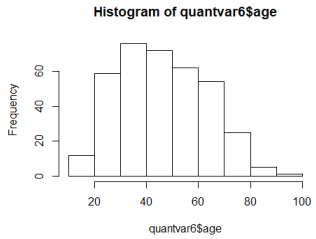
e. Histogram for ‘glyhb’: The distribution is unimodal; however, it seems to be quite sharp at around 5 or slightly below that area. There is also a skew to the right. The shape is also very sharp, and the spread to the right is much more than it is to the left. The shape is also much thinner in comparison to most other histograms.

i.



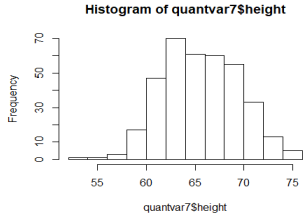
f. Histogram for ‘age’: The distribution appears unimodal; however, it is a very large gathering of data around a rather wide range. Therefore, the curve looks quite round. It appears almost uniform within a certain range, about 20-70, before the data begins to skew right.

i.



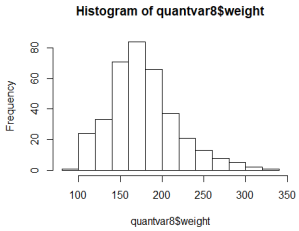
g. Histogram for ‘height’: The distribution is unimodal; however, it is also quite round. Therefore, it seems that the data gathers around 60-70 in an almost equal distribution. It appears almost uniform within a certain area. This time the skew is towards the left, but it is also not obvious or extreme.

i.



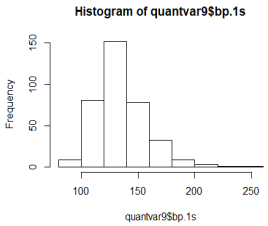
h. Histogram for ‘weight’: The distribution is unimodal, and the skew Is towards the right. Most of the data collects around 150-200, and then stretches towards 350. It is

must much spread to the right side than the left side. The peak in the shape is also somewhat sharp.



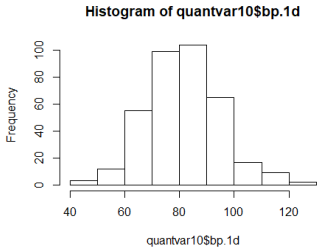
i.

i. Histogram for 'bp.1s': The distribution of bp.1s is unimodal, and skewed to the right. Most of the data is collecting around 100-150. There are some more extreme values around 250. The peak in the shape is also quite sharp.



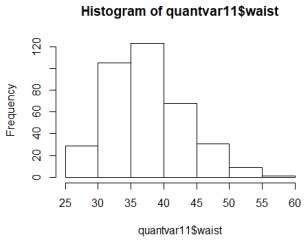
i.

j. Histogram for 'bp.1d': The distribution is unimodal, with most of the data collecting around 80. It also seems quite even with no noticeable skew in the data. It seems to spread quite evenly to the left and right.



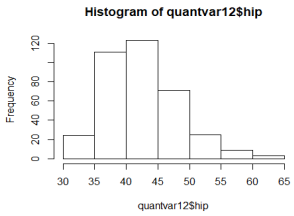
i.

k. Histogram for 'waist': The distribution is unimodal, with most of the data collecting around 30-40. The data is somewhat skewed to the right, but it is not too extreme. On the right it spreads to 60, but on the left it goes to 25.



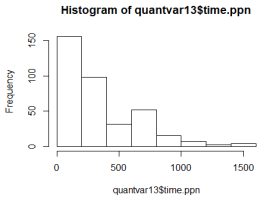
i.

l. Histogram for 'hip': The distribution for the data is unimodal, with most of the data collecting around 35-45. The data is also skewed towards the right side. On the right it spreads to 65, but on the left it spreads to 30.



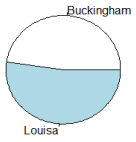
i.

m. Histogram for 'time.ppn': The distribution appears mostly unimodal, but also slightly bimodal due to the slight spike in the data around 750. The data is mostly collecting below 500. There is also a heavy skew towards the right side.



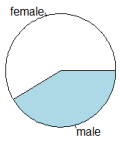
i.

n. Pie chart for 'location': The distribution for the location is roughly equal, with about half of the people in the data coming from Buckingham, and slightly more coming from Louisa.



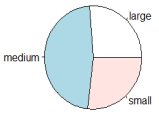
i.

o. Pie chart for 'gender': The distribution of the gender is heavier for females. There is also less males in the data, however the difference is not too severe. It is something that is noticeable however.



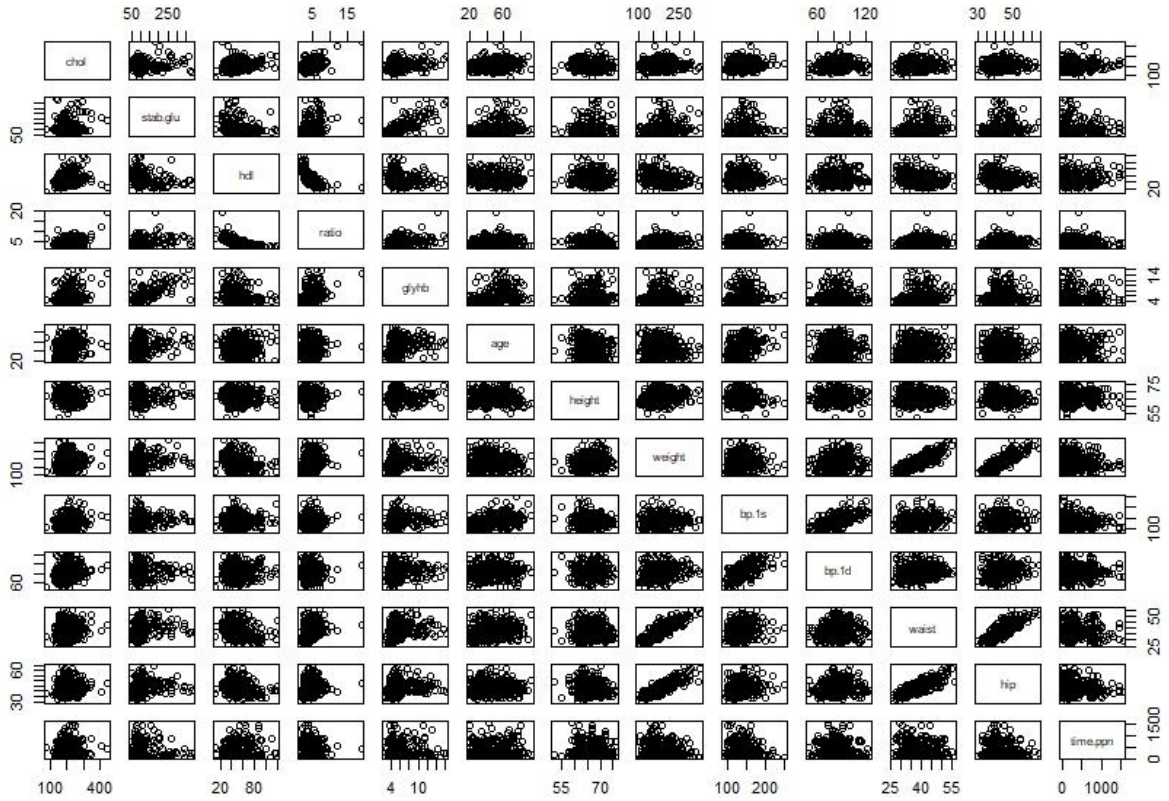
i.

p. Pie chart for 'frame': The distribution is over three types of frames: large, medium, and small. Many of the data is collected around the medium type. This makes up approximately half of the data in the entire sample. The rest of the data is split roughly evenly in the other half of the data between large and small.



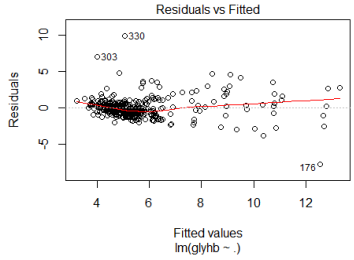
i.

q. After creating a scatterplot matrix: The clearest linear relationship between different quantitative variables are those that are between the physical characteristics of weight, waist, and hip. It makes sense that in a study, these are the most obvious connections to identify. It makes sense to think that people who are heavier will have larger waist and hip sizes. The variables 'hdl' and 'ratio' also seem to have a correlation that is negative. Therefore, multicollinearity exists between 'weight,' 'waist,' and 'hip' along with 'hdl' and 'ratio.' There is some slight relationship between 'bp.1s' and 'bp.1d,' however there is a large spread for the data.

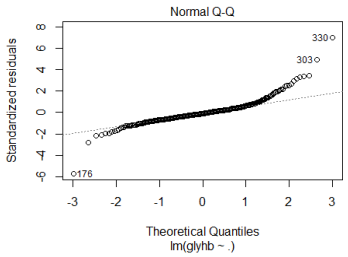


2. The Residuals vs Fitted plot suggests that the linear relationship between 'glyhb' and the other variables is neither positive nor negative. Certain outliers do exist in the plot, with the data from rows 303 and 330 existing much higher than the other data in the same region. Also, the data point for row 176 also exists at an extreme position, further below the other data points in the same region. The variation from the 0 line is fanning outward to the right

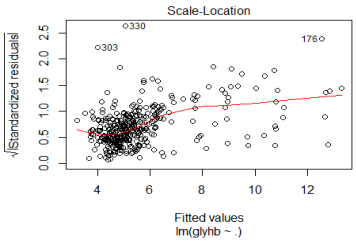
and much of the data is gathered between 4-6 on the fitted values axis. This is known as heteroscedasticity, and is not ideal for regression. At this concentrated area, the variance is much narrower. The assumption of equal variances is not possible due to the lack of randomness around the 0 line.



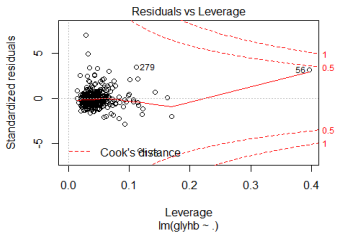
a. The Normal Q-Q plot shows that from around -1 to 1, most of the data falls directly on the line. The S-shaped curve from the bottom and then up towards the end suggests that the data has heavy-tailed residuals. The reason is that there are too many extreme positive and negative residuals in the data. Due to the data for this plot not being linear, the condition for normally distributed error terms isn't met.



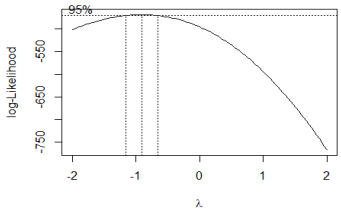
b. The Scale-Location plot shows how there is a lack of randomness in terms of how the residuals are being spread among the predictors. Most of the data is around 4-6, and in that area the data is dense around a narrower area. To the right of this section, the residuals are much more dispersed. There is a lack of equal variance, and so there is heteroscedasticity in this plot.



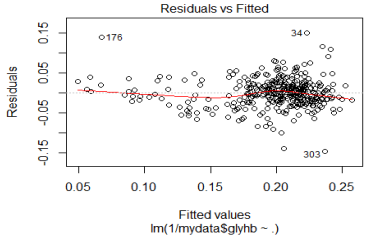
c. The Residuals vs Leverage plot shows that there is an extreme outlier which may be influential. The data from row 56 shows that there is an outlier near to the Cook's distance of 0.5 to the upper right. However, this point is not yet passed the dashed line of Cook's distance, and so the Cook's distance score is not yet too high. The next point that comes into question is from row 176, and at the bottom left, it seems to be getting close to the Cook's distance. However, this point has also not crossed the line, and so it may not be influential enough to the results of regression.



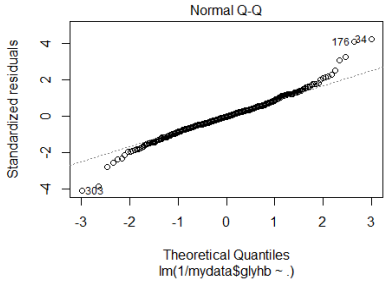
3. After performing 'boxcox' on the data, it becomes clear that the value for lambda should be equal to -1. This is done by replacing 'glyhb' with '1/mydata$glyhb.' Here 'mydata' is the name of the text file which contains the diabetes information.



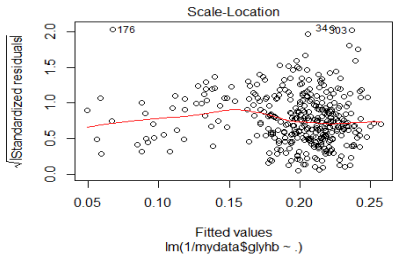
a. The Residuals vs Fitted plot still shows a horizontal relationship between Residuals and Fitted Values. On the left side, there are few data points, and there is an outlier from row 176 which is much higher than the other points. The points also spread out in a fan or cone, indicating that there is heteroscedasticity in the data. There is a lack of randomness for the data, and most of the data is densely gathered around 0.20 on the Fitted Values axis. The assumption of equal variance is not possible given the data.



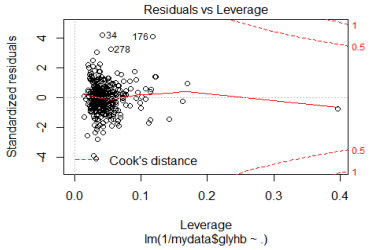
b. The Normal Q-Q plot shows that from around -1 to almost 2, most all the data falls exactly on the line. However, the data still shows an S-curve which indicates that the model is heavy-tailed. Still, there aren't too many points which are on these heavy-tailed regions, but the plot is still not linear at the edges. So, the condition for normally distributed error terms isn't met again.



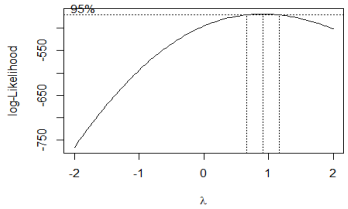
c. The Scale-Location plot shows that there is a lack of randomness, so the residuals aren't spread equally around the predictors. Although the line looks nearly horizontal, the points are concentrated on the right side. There are also some points which are stretched further upward on the right side than the other data in the region.



d. The Residuals vs Leverage plot shows that there aren't any points which exist too far out which leads towards a high Cook distance score. There are outliers however, and points 34, 176, and 278 stand out from the others in the region. They seem to be those that are closest to approaching Cook's distance line. There is also a point which is far to the right, however it is near to the 0 line, and so it is still distance from Cook's distance line as well.



e. After Using '1/mydata$glyhb to regress on the other variables to create Model 2, 'boxcox' was performed again on this updated Model. Now the 'boxcox' indicates that the recommended value for lambda is 1, or in other words, 'Y' has become ideally transformed after a single transformation.



5. After creating Model 3 by fitting all first-order effects, the number of regression coefficients is at 17. This was determined by using 'length(Model3$coeffcieints)' which determines the number of coefficients in RStudio. Similarly, the MSE was determined by using 'summary(Model3)$sigma^2' which can find out the MSE. The value for the MSE is 0.001383855.

6. Table of criterion values

Row or Model size	"best" 1 subset for each size
None	1/data.t$glyhb ~ 1
1	1/data.t$glyhb ~ 1 + stab.glu
2	1/data.t$glyhb ~ 1 + stab.glu + age
3	1/data.t$glyhb ~ 1 + stab.glu + age + waist
4	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + waist
5	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + framesmall + waist
6	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + framesmall + waist + time.ppn
7	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + framesmall + bp.1s + waist + time.ppn
8	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + height + framesmall + bp.1s + waist + time.ppn
9	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + height + weight + framesmall + waist + hip + time.ppn
10	1/data.t$glyhb ~ 1 + stab.glu + ratio + age + height + weight + framesmall + bp.1s + waist + hip + time.ppn
11	1/data.t$glyhb ~ 1 + chol + stab.glu + ratio + age + height + weight + framesmall + bp.1s + waist + hip + time.ppn
12	1/data.t$glyhb ~ 1 + chol + stab.glu + ratio + locationLouisa + age + height + weight + framesmall + bp.1s + waist + hip + time.ppn
13	1/data.t$glyhb ~ 1 + chol + stab.glu + hdl + ratio + locationLouisa + age + height + weight + framesmall + bp.1s + waist + hip + time.ppn
14	1/data.t$glyhb ~ 1 + chol + stab.glu + hdl + ratio + locationLouisa + age + height + weight + framemedium + framesmall + bp.1s + waist + hip + time.ppn
15	1/data.t$glyhb ~ 1 + chol + stab.glu + hdl + ratio + locationLouisa + age + height + weight + framemedium + framesmall + bp.1s + bp.1d + waist + hip + time.ppn
16	1/data.t$glyhb ~ 1 + chol + stab.glu + hdl + ratio + locationLouisa + age + gendermale + height + weight + framemedium + framesmall + bp.1s + bp.1d + waist + hip + time.ppn

Below is a table with the SSE_p, R_p^2, $R_{a,p}^2$, C_p, AIC_p, and BIC_p for all subset sizes.

```
##             sse       R^2     R^2_a           Cp       aic       bic
## none 0.5158646 0.0000000 0.0000000 191.73453170 -1072.466 -1069.256
## 1    0.2864076 0.4448009 0.4417335  27.96351331 -1178.148 -1171.729
## 2    0.2574112 0.5010102 0.4954659   9.01014928 -1195.682 -1186.053
## 3    0.2428890 0.5291612 0.5212701   0.51619889 -1204.309 -1191.471
## 4    0.2401432 0.5344840 0.5240230   0.53201659 -1204.389 -1188.342
## 5    0.2367131 0.5411332 0.5281708   0.05337754 -1205.022 -1185.765
## 6    0.2343460 0.5457220 0.5302352   0.34280455 -1204.861 -1182.395
## 7    0.2331725 0.5479966 0.5299165   1.49487219 -1203.780 -1178.104
## 8    0.2326634 0.5489836 0.5282473   3.12693590 -1202.180 -1173.294
## 9    0.2314193 0.5513952 0.5280574   4.22797088 -1201.161 -1169.066
## 10   0.2303187 0.5535287 0.5275711   5.43265348 -1200.033 -1164.729
## 11   0.2300477 0.5540541 0.5253676   7.23678869 -1198.249 -1159.735
## 12   0.2299216 0.5542986 0.5228374   9.14564365 -1196.349 -1154.626
## 13   0.2298166 0.5545020 0.5202329  11.06983181 -1194.433 -1149.500
## 14   0.2297510 0.5546292 0.5175150  13.02241521 -1192.485 -1144.343
## 15   0.2297274 0.5546751 0.5146758  15.00531267 -1190.504 -1139.152
## 16   0.2297200 0.5546893 0.5117678  17.00000000 -1188.510 -1133.948
```

Below is a table with the best models for each criterion.

SSE_p	R_p^2	$R_{a,p}^2$	C_p	AIC_p	BIC_p
0.2297200 (row 16)	0.5546893 (row 16)	0.5302352 (row 6)	0.05337754 (row 5)	-1205.022 (row 5)	-1191.471 (row 3)

Model 3.1, or the best AIC_p would be row 5. (1/data.t$glyhb ~ 1 + stab.glu + ratio + age + framesmall + waist) with a value of -1205.022.

Model 3.2., or the best BIC_p would be row 3. (1/data.t$glyhb ~ 1 + stab.glu + age + waist) with a value of -1191.471.

Model 3.3, or the best $R_{a,p}^2$ would be row 6. (1/data.t$glyhb ~ 1 + stab.glu + ratio + age + framesmall + waist + time.ppn) with a value of 0.5302352.

The Mallows's C_p is said to be ideal when it is near to the 'p' that the statistic is referring to. In other words, if p = 6, then the ideal C_p value would be the one nearest to 6. However, when p becomes large, as in this case where there are 16 predictors, then there are values which are much lower than 'p.' In this example, the lowest value is 0.05337754, which is preferable to choosing values near to the 'p.' Although there are values near 'p,' in the practical application of Mallows's C_p, then it would be ideal to choose the smallest value.

7. After fitting Model 4 using 'lm(1/data.t$glyhb ~.^2,data=data.t),' I was able to use 'length(Model4$coefficients)' to determine that the number of regression coefficients for this model is 136. The MSE was determined using 'summary(Model4)$sigma^2,' and the value is 0.001036088. The concern is that with so many predictors for a given model, it doesn't make sense since the original model had only 16 predictors. The result is that overfitting may occur, and this is due to the additional predictor variables. The regression model becomes overcomplicated from all the 2-way interaction effects. What happens from the statistician's point of view is that there is an increase in the uncertainty of the predictive capability of the predictor variables.
8. Using the forward stepwise procedure that starts with a none-model in addition to the AIC_p as a criterion, the model that is chosen is 1/data.t$glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio. The AIC value of Model fs1 is -1205.14 in comparison to Model 3.1's AIC value which is -1205.022. The new Model fs1 has a less favorable AIC value, and this is interesting due to the additional complexity involved in having a model that is fitted with first-order and 2-way interaction effects. However, they both involve a similar number of predictors, so it is not too worrisome.
9. By applying the forward stepwise procedure again, however this time with a starting point of the full model, or Model 3, the model that is chosen is 1/data.t$glyhb ~ chol + stab.glu + hdl + ratio + age + gender + height + weight + bp.1s + bp.1d + waist + hip + time.ppn + stab.glu:gender + hdl:ratio + age:bp.1d + weight:bp.1s + age:hip + hip:time.ppn + gender:height + stab.glu:bp.1s + stab.glu:time.ppn + stab.glu:waist + age:waist + chol:time.ppn + hdl:weight + bp.1d:waist + weight:hip. The AIC value for this new Model

named fs2 is -1230.61. This time the AIC value is lower, and therefore it is preferable to the previous Model fs1 which has an AIC value of -1205.14. It is not certain why the AIC is higher for fs2 than fs1, but the forward stepwise procedure may have found itself into a different local minimum than fs1. The forward stepwise procedure will not guarantee to find the "best" subset due to how it's like a greedy algorithm which finds the next best option rather than the actual best option. In addition, fs2 is much longer, so the numerous predictors may have lead the criterion to imply that it is a better model to choose from.

10. The BIC value of Model fs1 is -1182.677, and the BIC value of Model fs2 is -1137.536. The BIC value for Model fs1 is lower than fs2, so it is preferable. According to AIC, the better Model is fs2, and according to BIC, the better model is fs1. So, Model 4.1 is fs2, and Model 4.2 is fs1.
11. After calculating the PRESS value for Models 3.1, 3.2, 3.3, 4.1, and 4.2, and Model 4.1 is lowest at 0.2171946. So, Model 4.1 is preferable in comparison to Model 3.1, 3.2, 3.3, and Model 4.2. Model 4.1 and 4.2 are based off regression models that have both first and second order effects. Due to the high number of predictors in Model 4.1, it makes sense that the PRESS value for the Model 4.1 has a better value in comparison to the other models which only have first-order effects, or not a high amount of predictor variables.

PRESS Model 3.1	*PRESS* Model 3.2	*PRESS* Model 3.3	*PRESS* Model 4.1	*PRESS* Model 4.2
0.252777	0.2539834	0.252575	0. 2171946	0.2534834

12. After calculating the MSPR using the validation set, the Model 3.1, 3.2, and 3.3 MSPR values are lower than that of their respective PRESS/n values. However, the MSPR values for Model 4.1 and 4.2 is lower than its respective PRESS/n value. The smallest MSPR value is with Model 3.3, and the value is 0.00134099.

MSPR Model 3.1	*MSPR Model* 3.2	*MSPR Model* 3.3	*MSPR Model* 4.1	*MSPR Model* 4.2
0.001368448	0.001377312	0.00134099	0.001797609	0.00152642
$\frac{PRESS}{n}$ *Model* 3.1	$\frac{PRESS}{n}$ *Model* 3.2	$\frac{PRESS}{n}$ *Model* 3.3	$\frac{PRESS}{n}$ *Model* 4.1	$\frac{PRESS}{N}$ *Model* 4.2
0.001381295	0.001387887	0.001380191	0. 001186856	0.001385155

13. According to MSPR, I would choose Model 3.3, since it has the lowest value, and therefore is preferable over the other models according to the MSPR method. I wouldn't use Model 4.1 even though it has the best outcome for PRESS/n and BIC. The model may appear to have good prediction power; however, the model is complicated since it uses so many predictors. Therefore, it's hard to interpret the model. Also, when it goes through cross validation, the validation data gives it the highest (and therefore, the worst) MSPR out of all the models. My goal is for prediction, so MSPR is also a good validation method to use for determining how good is the regression model. Also, MSPR makes it so that the best model is chosen without worrying about overfitting. The reason is that MSPR utilizes validation data from a different data set, which would cancel out the effects of a model overestimating the value of prediction. Additionally, Model 3.3 also did well in comparison to others on the PRESS/n test.

Below are reports of the summary() and anova() for Model 5:

```
summary(Model5)
## Call:
## lm(formula = 1/mydata$glyhb ~ stab.glu + age + ratio + waist +
##     time.ppn + X6L + stab.glu:time.ppn + stab.glu:age + age:ratio,
##     data = mydata)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.153964 -0.021172 -0.001289  0.020548  0.151535
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.355e-01  2.349e-02  14.281  < 2e-16 ***
## stab.glu           -7.876e-04  1.630e-04  -4.832 2.01e-06 ***
## age                -8.332e-04  4.096e-04  -2.034 0.042680 *
## ratio               2.777e-03  4.237e-03   0.655 0.512645
## waist              -1.034e-03  3.450e-04  -2.997 0.002920 **
## time.ppn            3.580e-05  1.572e-05   2.277 0.023349 *
## X6L                 6.642e-03  3.772e-03   1.761 0.079123 .
## stab.glu:time.ppn -4.619e-07  1.336e-07  -3.457 0.000612 ***
## stab.glu:age        7.341e-06  2.770e-06   2.650 0.008406 **
## age:ratio          -1.209e-04  8.145e-05  -1.484 0.138739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03538 on 356 degrees of freedom
## Multiple R-squared:  0.5354, Adjusted R-squared:  0.5236
## F-statistic: 45.58 on 9 and 356 DF,  p-value: < 2.2e-16
```

```
anova(Model5)
```

```
## Analysis of Variance Table
##
## Response: 1/mydata$glyhb
##                    Df  Sum Sq Mean Sq  F value    Pr(>F)
## stab.glu            1 0.39753 0.39753 317.4948 < 2.2e-16 ***
## age                 1 0.04867 0.04867  38.8708 1.282e-09 ***
## ratio               1 0.02148 0.02148  17.1591 4.298e-05 ***
## waist               1 0.01252 0.01252   9.9982 0.0017015 **
## time.ppn            1 0.00646 0.00646   5.1600 0.0237090 *
## X6L                 1 0.00315 0.00315   2.5120 0.1138731
## stab.glu:time.ppn   1 0.01383 0.01383  11.0471 0.0009806 ***
## stab.glu:age        1 0.00722 0.00722   5.7636 0.0168739 *
## age:ratio           1 0.00276 0.00276   2.2018 0.1387387
## Residuals         356 0.44574 0.00125
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Proj 2 markdown final Jared Yu December 4, 2017

```
#Set the working directory
setwd("C:/Users/jyqq9/Desktop/STA 108/Project 2")

#Read in the data from the text file
mydata = read.table("diabetes.txt", header=T)

#Examine the data
head(mydata)
```

```
##   chol stab.glu hdl ratio glyhb   location age gender height weight  frame
## 1  203       82  56   3.6  4.31 Buckingham  46 female     62    121 medium
## 2  165       97  24   6.9  4.44 Buckingham  29 female     64    218  large
## 3  228       92  37   6.2  4.64 Buckingham  58 female     61    256  large
## 4   78       93  12   6.5  4.63 Buckingham  67   male     67    119  large
## 5  249       90  28   8.9  7.72 Buckingham  64   male     68    183 medium
## 6  248       94  69   3.6  4.81 Buckingham  34   male     71    190  large
##   bp.1s bp.1d waist hip time.ppn
## 1   118    59    29  38      720
## 2   112    68    46  48      360
## 3   190    92    49  57      180
## 4   110    50    33  38      480
## 5   138    80    44  41      300
## 6   132    86    36  42      195
```

```
dim(mydata)
```

```
## [1] 366  16
```

```
#1
#The quantiative variables are chol, stab.glu, hdl, ratio, glyhb, age, height
, weight,
#bp.1s, bp.1d, waist, hip, time.ppn. The qualitative variables are location,
gender, and frame.

#histogram for each quantiative variable
#chol's distribution
quantvar1 = subset(mydata, select = c(chol))
hist(quantvar1$chol)

#stab.glu's distribution
quantvar2 = subset(mydata, select = c(stab.glu))
hist(quantvar2$stab.glu)
```

```r
#hdl's distribution
quantvar3 = subset(mydata, select = c(hdl))
hist(quantvar3$hdl)

#ratio's distribution
quantvar4 = subset(mydata, select = c(ratio))
hist(quantvar4$ratio)

#glyhb's distribution
quantvar5 = subset(mydata, select = c(glyhb))
hist(quantvar5$glyhb)

#age's distribution
quantvar6 = subset(mydata, select = c(age))
hist(quantvar6$age)

#height's distribution
quantvar7 = subset(mydata, select = c(height))
hist(quantvar7$height)

#weight's distribution
quantvar8 = subset(mydata, select = c(weight))
hist(quantvar8$weight)

#bp.1s' distribution
quantvar9 = subset(mydata, select = c(bp.1s))
hist(quantvar9$bp.1s)

#bp.1d's distribution
quantvar10 = subset(mydata, select = c(bp.1d))
hist(quantvar10$bp.1d)

#waist's distribution
quantvar11 = subset(mydata, select = c(waist))
hist(quantvar11$waist)

#hip's distribution
quantvar12 = subset(mydata, select = c(hip))
hist(quantvar12$hip)

#time.ppn's distribution
quantvar13 = subset(mydata, select = c(time.ppn))
hist(quantvar13$time.ppn)

#pie chart for each qualitative variable
#Load MASS package for the table function
library(MASS)

#location's distribution
qualvar1 = subset(mydata, select = c(location))
location = qualvar1$location
```

```r
location.pie = table(location)
pie(location.pie)

#gender's distribution
qualvar2 = subset(mydata, select = c(gender))
gender = qualvar2$gender
gender.pie = table(gender)
pie(gender.pie)

#frame's distribution
qualvar3 = subset(mydata, select = c(frame))
frame = qualvar3$frame
frame.pie = table(frame)
pie(frame.pie)

#draw a scatterplot matrix and obtain the pairwise correlation matrix for all
#quantiative variables in the data
quantvarmatrix = mydata[c(1:5,7,9:10,12:16)]
pairs(quantvarmatrix)

#2
#regress glybh on all predictor variables
Model1 = lm(glyhb~.,data=mydata)
plot(Model1)

#3
library(MASS)
boxcox(mydata$glyhb~.,data=mydata)

Model2 = lm(1/mydata$glyhb~., data=mydata)
plot(Model2)

boxcox(Model2)

#4
set.seed(10) #set seed for random number generator
            #so everyone gets the same split of the data.
N=nrow(mydata) #number of observations in the data
index=sample(1:N, size=N/2, replace=FALSE) #randomly sample
                #N/2 observation to form the training data.
data.t=mydata[index,] #get the training data set
data.v=mydata[-index,] #the remaining N/2 observations form the validation se
t


#5
Model3 = lm(1/data.t$glyhb~., data=data.t)
summary(Model3)

##
## Call:
## lm(formula = 1/data.t$glyhb ~ ., data = data.t)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.097813 -0.022472 -0.002034  0.021097  0.134611
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.819e-01  8.499e-02   5.670 6.19e-08 ***
## chol           -6.857e-05  1.695e-04  -0.405   0.6863
## stab.glu       -5.314e-04  5.418e-05  -9.807  < 2e-16 ***
## hdl             1.211e-04  5.492e-04   0.220   0.8258
## ratio          -2.414e-03  6.588e-03  -0.366   0.7145
## locationLouisa -1.808e-03  5.969e-03  -0.303   0.7623
## age            -5.487e-04  2.199e-04  -2.495   0.0136 *
## gendermale     -7.422e-04  1.018e-02  -0.073   0.9420
## height         -1.212e-03  1.123e-03  -1.079   0.2820
## weight          2.210e-04  2.034e-04   1.087   0.2788
## framemedium     1.417e-03  7.861e-03   0.180   0.8572
## framesmall     -1.062e-02  9.596e-03  -1.107   0.2699
## bp.1s          -1.214e-04  1.708e-04  -0.711   0.4782
## bp.1d           3.198e-05  2.505e-04   0.128   0.8986
## waist          -1.893e-03  1.148e-03  -1.649   0.1010
## hip            -1.177e-03  1.352e-03  -0.870   0.3854
## time.ppn       -1.444e-05  9.881e-06  -1.461   0.1459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0372 on 166 degrees of freedom
## Multiple R-squared:  0.5547, Adjusted R-squared:  0.5118
## F-statistic: 12.92 on 16 and 166 DF,  p-value: < 2.2e-16
```

```
length(Model3$coefficients) #of regression coefficients
```

```
## [1] 17
```

```
summary(Model3)$sigma^2 #MSE from this model
```

```
## [1] 0.001383855
```

```
#6
library(leaps)
best = regsubsets(1/data.t$glyhb~., data=data.t, nbest=1, nvmax=16)
sum_sub=summary(best)
sum_sub$which
```

```
##    (Intercept)  chol stab.glu   hdl ratio locationLouisa   age gendermale
## 1         TRUE FALSE     TRUE FALSE FALSE          FALSE FALSE      FALSE
## 2         TRUE FALSE     TRUE FALSE FALSE          FALSE  TRUE      FALSE
## 3         TRUE FALSE     TRUE FALSE FALSE          FALSE  TRUE      FALSE
## 4         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 5         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 6         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
```

```
## 7         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 8         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 9         TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 10        TRUE FALSE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 11        TRUE  TRUE     TRUE FALSE  TRUE          FALSE  TRUE      FALSE
## 12        TRUE  TRUE     TRUE FALSE  TRUE           TRUE  TRUE      FALSE
## 13        TRUE  TRUE     TRUE  TRUE  TRUE           TRUE  TRUE      FALSE
## 14        TRUE  TRUE     TRUE  TRUE  TRUE           TRUE  TRUE      FALSE
## 15        TRUE  TRUE     TRUE  TRUE  TRUE           TRUE  TRUE      FALSE
## 16        TRUE  TRUE     TRUE  TRUE  TRUE           TRUE  TRUE       TRUE
##    height weight framemedium framesmall bp.1s bp.1d waist   hip time.ppn
## 1   FALSE  FALSE       FALSE      FALSE FALSE FALSE FALSE FALSE    FALSE
## 2   FALSE  FALSE       FALSE      FALSE FALSE FALSE FALSE FALSE    FALSE
## 3   FALSE  FALSE       FALSE      FALSE FALSE FALSE  TRUE FALSE    FALSE
## 4   FALSE  FALSE       FALSE      FALSE FALSE FALSE  TRUE FALSE    FALSE
## 5   FALSE  FALSE       FALSE       TRUE FALSE FALSE  TRUE FALSE    FALSE
## 6   FALSE  FALSE       FALSE       TRUE FALSE FALSE  TRUE FALSE     TRUE
## 7   FALSE  FALSE       FALSE       TRUE  TRUE FALSE  TRUE FALSE     TRUE
## 8    TRUE  FALSE       FALSE       TRUE  TRUE FALSE  TRUE FALSE     TRUE
## 9    TRUE   TRUE       FALSE       TRUE FALSE FALSE  TRUE  TRUE     TRUE
## 10   TRUE   TRUE       FALSE       TRUE  TRUE FALSE  TRUE  TRUE     TRUE
## 11   TRUE   TRUE       FALSE       TRUE  TRUE FALSE  TRUE  TRUE     TRUE
## 12   TRUE   TRUE       FALSE       TRUE  TRUE FALSE  TRUE  TRUE     TRUE
## 13   TRUE   TRUE       FALSE       TRUE  TRUE FALSE  TRUE  TRUE     TRUE
## 14   TRUE   TRUE        TRUE       TRUE  TRUE FALSE  TRUE  TRUE     TRUE
## 15   TRUE   TRUE        TRUE       TRUE  TRUE  TRUE  TRUE  TRUE     TRUE
## 16   TRUE   TRUE        TRUE       TRUE  TRUE  TRUE  TRUE  TRUE     TRUE
```

```
n=nrow(data.t)
n
```

```
## [1] 183
```

```
p.m=2:17
sse=sum_sub$rss
sse
```

```
##  [1] 0.2864076 0.2574112 0.2428890 0.2401432 0.2367131 0.2343460 0.2331725
##  [8] 0.2326634 0.2314193 0.2303187 0.2300477 0.2299216 0.2298166 0.2297510
## [15] 0.2297274 0.2297200
```

```
aic=n*log(sse)+2*p.m-n*log(n)
aic
```

```
##  [1] -1178.148 -1195.682 -1204.309 -1204.389 -1205.022 -1204.861 -1203.780
##  [8] -1202.180 -1201.161 -1200.033 -1198.249 -1196.349 -1194.433 -1192.485
## [15] -1190.504 -1188.510
```

```
bic=n*log(sse)+log(n)*p.m-n*log(n)
bic
```

```
##  [1] -1171.729 -1186.053 -1191.471 -1188.342 -1185.765 -1182.395 -1178.104
##  [8] -1173.294 -1169.066 -1164.729 -1159.735 -1154.626 -1149.500 -1144.343
## [15] -1139.152 -1133.948
```

```
res_sub=cbind(sum_sub$which,sse,sum_sub$rsq,sum_sub$adjr2,sum_sub$cp,aic, bic
)
fit0=lm(1/data.t$glyhb~1,data=data.t) # fit the model with only intercept
sse1=sum(fit0$residuals^2)
p=1
c1=sse1/0.001384-(n-2*p)
aic1=n*log(sse1)+2*p-n*log(n)
bic1=n*log(sse1)+log(n)*p-n*log(n)
none=c(1,rep(0,16),sse1,0,0,c1,aic1,bic1)
res_sub=rbind(none,res_sub) # combine the results with other models
colnames(res_sub)=c(colnames(sum_sub$which),"sse", "R^2", "R^2_a", "Cp", "aic
", "bic")
res_sub
```

```
##      (Intercept) chol stab.glu hdl ratio locationLouisa age gendermale
## none           1    0        0   0     0              0   0          0
## 1              1    0        1   0     0              0   0          0
## 2              1    0        1   0     0              0   1          0
## 3              1    0        1   0     0              0   1          0
## 4              1    0        1   0     1              0   1          0
## 5              1    0        1   0     1              0   1          0
## 6              1    0        1   0     1              0   1          0
## 7              1    0        1   0     1              0   1          0
## 8              1    0        1   0     1              0   1          0
## 9              1    0        1   0     1              0   1          0
## 10             1    0        1   0     1              0   1          0
## 11             1    1        1   0     1              0   1          0
## 12             1    1        1   0     1              1   1          0
## 13             1    1        1   1     1              1   1          0
## 14             1    1        1   1     1              1   1          0
## 15             1    1        1   1     1              1   1          0
## 16             1    1        1   1     1              1   1          1
##      height weight framemedium framesmall bp.1s bp.1d waist hip time.ppn
## none      0      0           0          0     0     0     0   0        0
## 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     1   0        0
## 4         0      0           0          0     0     0     1   0        0
## 5         0      0           0          1     0     0     1   0        0
## 6         0      0           0          1     0     0     1   0        1
## 7         0      0           0          1     1     0     1   0        1
## 8         1      0           0          1     1     0     1   0        1
## 9         1      1           0          1     0     0     1   1        1
## 10        1      1           0          1     1     0     1   1        1
## 11        1      1           0          1     1     0     1   1        1
## 12        1      1           0          1     1     0     1   1        1
```

```
## 13        1      1           0          1     1     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
##             sse       R^2     R^2_a           Cp       aic       bic
## none 0.5158646 0.0000000 0.0000000 191.73453170 -1072.466 -1069.256
## 1    0.2864076 0.4448009 0.4417335  27.96351331 -1178.148 -1171.729
## 2    0.2574112 0.5010102 0.4954659   9.01014928 -1195.682 -1186.053
## 3    0.2428890 0.5291612 0.5212701   0.51619889 -1204.309 -1191.471
## 4    0.2401432 0.5344840 0.5240230   0.53201659 -1204.389 -1188.342
## 5    0.2367131 0.5411332 0.5281708   0.05337754 -1205.022 -1185.765
## 6    0.2343460 0.5457220 0.5302352   0.34280455 -1204.861 -1182.395
## 7    0.2331725 0.5479966 0.5299165   1.49487219 -1203.780 -1178.104
## 8    0.2326634 0.5489836 0.5282473   3.12693590 -1202.180 -1173.294
## 9    0.2314193 0.5513952 0.5280574   4.22797088 -1201.161 -1169.066
## 10   0.2303187 0.5535287 0.5275711   5.43265348 -1200.033 -1164.729
## 11   0.2300477 0.5540541 0.5253676   7.23678869 -1198.249 -1159.735
## 12   0.2299216 0.5542986 0.5228374   9.14564365 -1196.349 -1154.626
## 13   0.2298166 0.5545020 0.5202329  11.06983181 -1194.433 -1149.500
## 14   0.2297510 0.5546292 0.5175150  13.02241521 -1192.485 -1144.343
## 15   0.2297274 0.5546751 0.5146758  15.00531267 -1190.504 -1139.152
## 16   0.2297200 0.5546893 0.5117678  17.00000000 -1188.510 -1133.948
```

```
#Model 3.1, 3.2, 3.3
#Create qualitative variable framesmall
is.factor(data.t[,11]) #TRUE
```

```
## [1] TRUE
```

```
summary(data.t$frame)
```

```
##  large medium  small
##     46     89     48
```

```
small.index=which(data.t$frame=="small")
small.index #observations of frame = small
```

```
##  [1]   1   4   6   7  15  16  21  30  33  35  36  40  51  52  64  68  73
## [18]  77  79  81  84  87  92  95  99 100 117 119 120 125 127 132 133 137
## [35] 139 141 146 148 159 162 166 167 168 170 172 177 180 182
```

```
#data.t[1:11]
n=dim(data.t)[1]
X11s=rep(0,n)
X11s[small.index]=1 #small = 1, rest are 0

Model3.1=lm(1/data.t$glyhb ~ stab.glu + ratio + age + X11s + waist,data=data.
t) #AIC selected
Model3.2=lm(1/data.t$glyhb~.,data=data.t[c(2,7,14)]) #BIC selected
Model3.3=lm(1/data.t$glyhb~ stab.glu + ratio + age + X11s + waist + time.ppn,
data=data.t) #AdjR^2 selected
```

```r
#7
Model4=lm(1/data.t$glyhb~.^2,data=data.t)
length(Model4$coefficients) #136
```

```
## [1] 136
```

```r
summary(Model4)$sigma^2 #0.001036088
```

```
## [1] 0.001036088
```

```r
#8
best = regsubsets(1/data.t$glyhb~., data=data.t, nbest=1, nvmax=16)
fit0=lm(1/mydata$glyhb~1,data=mydata)
fs1=stepAIC(fit0,scope=list(upper=Model4,lower=~1),direction="both",k=2)
## Step:  AIC=-2436.1
## 1/mydata$glyhb ~ stab.glu + age + ratio + waist + time.ppn +
##     location + stab.glu:time.ppn + stab.glu:age + age:ratio
# Step:  AIC=-2436.1
# 1/mydata$glyhb ~ stab.glu + age + ratio + waist + time.ppn +
#   location + stab.glu:time.ppn + stab.glu:age + age:ratio
sse.fs1=sum(fs1$residuals^2)
p.fs1=length(fs1$coefficients)

#9
fs2=stepAIC(Model3,scope=list(upper=Model4,lower=~1),direction="both",k=2)

## Step:  AIC=-1230.61
## 1/data.t$glyhb ~ chol + stab.glu + hdl + ratio + age + gender +
##     height + weight + bp.1s + bp.1d + waist + hip + time.ppn +
##     stab.glu:gender + hdl:ratio + age:bp.1d + weight:bp.1s +
##     age:hip + hip:time.ppn + gender:height + stab.glu:bp.1s +
##     stab.glu:time.ppn + stab.glu:waist + age:waist + chol:time.ppn +
##     hdl:weight + bp.1d:waist + weight:hip
# Step:  AIC=-1230.61
# 1/data.t$glyhb ~ chol + stab.glu + hdl + ratio + age + gender +
#   height + weight + bp.1s + bp.1d + waist + hip + time.ppn +
#   stab.glu:gender + hdl:ratio + age:bp.1d + weight:bp.1s +
#   age:hip + hip:time.ppn + gender:height + stab.glu:bp.1s +
#   stab.glu:time.ppn + stab.glu:waist + age:waist + chol:time.ppn +
#   hdl:weight + bp.1d:waist + weight:hip

#10
n = nrow(data.t)
bic.fs1=n*log(sse.fs1)+p.fs1*log(n)-n*log(n)
bic.fs1 #-1049.11
```

```
## [1] -1049.11
```

```r
sse.fs2=sum(fs2$residuals^2)
p.fs2=length(fs2$coefficients)
```

```r
bic.fs2=n*log(sse.fs2)+p.fs2*log(n)-n*log(n)
bic.fs2 #-1137.536
```

```
## [1] -1137.536
```

```r
#set Models 4.1 and 4.2
Model4.1 = fs1
Model4.2 = fs2

#11
PRESSModel3.1 = sum(Model3.1$residuals^2/(1-influence(Model3.1)$hat)^2)
PRESSModel3.2 = sum(Model3.2$residuals^2/(1-influence(Model3.2)$hat)^2)
PRESSModel3.3 = sum(Model3.3$residuals^2/(1-influence(Model3.3)$hat)^2)
PRESSModel4.1 = sum(Model4.1$residuals^2/(1-influence(Model4.1)$hat)^2)
PRESSModel4.2 = sum(Model4.2$residuals^2/(1-influence(Model4.2)$hat)^2)
PRESSModel3.1 #0.252777
```

```
## [1] 0.252777
```

```r
PRESSModel3.2 #0.2539834
```

```
## [1] 0.2539834
```

```r
PRESSModel3.3 #0.252575
```

```
## [1] 0.252575
```

```r
PRESSModel4.1 #0.4742118
```

```
## [1] 0.4742118
```

```r
PRESSModel4.2 #0.2171946
```

```
## [1] 0.2171946
```

```r
#12
Yhat.3.1 = predict(Model3.1, newdata=data.v)
Y.3.1 = 1/data.v[,5]
m = nrow(data.v)
MSPR.3.1 = sum((Yhat.3.1-Y.3.1)^2)/m
MSPR.3.1 #0.001368448
```

```
## [1] 0.001368448
```

```r
Yhat.3.2 = predict(Model3.2, newdata=data.v)
Y.3.2 = 1/data.v[,5]
MSPR.3.2 = sum((Yhat.3.2-Y.3.2)^2)/m
MSPR.3.2 #0.001377312
```

```
## [1] 0.001377312
```

```r
Yhat.3.3 = predict(Model3.3, newdata=data.v)
Y.3.3 = 1/data.v[,5]
```

```r
MSPR.3.3 = sum((Yhat.3.3-Y.3.3)^2)/m
MSPR.3.3 #0.00134099
```

```
## [1] 0.00134099
```

```r
Yhat.4.1 = predict(Model4.1, newdata=data.v)
Y.4.1 = 1/data.v[,5]
MSPR.4.1 = sum((Yhat.4.1-Y.4.1)^2)/m
MSPR.4.1 #0.001159244
```

```
## [1] 0.001159244
```

```r
Yhat.4.2 = predict(Model4.2, newdata=data.v)
Y.4.2 = 1/data.v[,5]
MSPR.4.2 = sum((Yhat.4.2-Y.4.2)^2)/m
MSPR.4.2 #0.001797609
```

```
## [1] 0.001797609
```

```r
m = nrow(data.t)
PRESSModel3.1/m #0.001381295
```

```
## [1] 0.001381295
```

```r
PRESSModel3.2/m #0.001387887
```

```
## [1] 0.001387887
```

```r
PRESSModel3.3/m #0.001380191
```

```
## [1] 0.001380191
```

```r
PRESSModel4.1/m #0.002591321
```

```
## [1] 0.002591321
```

```r
PRESSModel4.2/m #0.001186856
```

```
## [1] 0.001186856
```

```r
#13
#Create qualitative variable for Louisa
is.factor(mydata[,6]) #TRUE
```

```
## [1] TRUE
```

```r
summary(mydata$location)
```

```
## Buckingham     Louisa
##        175        191
```

```r
Louisa.index=which(mydata$location=="Louisa")
Louisa.index #observations of frame = Louisa
```

```
##   [1]  10  11  12  13  14  15  16  17  27  72  73  74  75  76  77  78  79
##  [18]  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96
##  [35]  97  98  99 100 101 105 106 107 108 111 115 116 117 118 124 125 137
##  [52] 142 143 144 145 146 147 154 155 156 157 160 161 162 163 164 165 166
##  [69] 167 168 169 170 171 172 173 174 175 176 183 184 185 209 210 211 212
##  [86] 213 251 252 253 254 255 257 258 259 260 261 262 263 264 265 266 267
## [103] 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284
## [120] 285 286 287 288 289 290 293 294 295 304 305 306 307 308 309 310 311
## [137] 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328
## [154] 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345
## [171] 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362
## [188] 363 364 365 366
```

```
#mydata[1:6]
n=dim(mydata)[1]
X6L=rep(0,n)
X6L[Louisa.index]=1 #Louisa = 1, rest are 0
#Create Model 5
Model5 = lm(1/mydata$glyhb ~ stab.glu + age + ratio + waist + time.ppn + X6L
+ stab.glu:time.ppn + stab.glu:age + age:ratio, data=mydata)
summary(Model5)
```

```
##
## Call:
## lm(formula = 1/mydata$glyhb ~ stab.glu + age + ratio + waist +
##     time.ppn + X6L + stab.glu:time.ppn + stab.glu:age + age:ratio,
##     data = mydata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.153964 -0.021172 -0.001289  0.020548  0.151535
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.355e-01  2.349e-02  14.281  < 2e-16 ***
## stab.glu          -7.876e-04  1.630e-04  -4.832 2.01e-06 ***
## age               -8.332e-04  4.096e-04  -2.034 0.042680 *
## ratio              2.777e-03  4.237e-03   0.655 0.512645
## waist             -1.034e-03  3.450e-04  -2.997 0.002920 **
## time.ppn           3.580e-05  1.572e-05   2.277 0.023349 *
## X6L                6.642e-03  3.772e-03   1.761 0.079123 .
## stab.glu:time.ppn -4.619e-07  1.336e-07  -3.457 0.000612 ***
## stab.glu:age       7.341e-06  2.770e-06   2.650 0.008406 **
## age:ratio         -1.209e-04  8.145e-05  -1.484 0.138739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03538 on 356 degrees of freedom
## Multiple R-squared:  0.5354, Adjusted R-squared:  0.5236
## F-statistic: 45.58 on 9 and 356 DF,  p-value: < 2.2e-16
```

```r
anova(Model5)
```

```
## Analysis of Variance Table
##
## Response: 1/mydata$glyhb
##                     Df  Sum Sq Mean Sq  F value    Pr(>F)
## stab.glu             1 0.39753 0.39753 317.4948 < 2.2e-16 ***
## age                  1 0.04867 0.04867  38.8708 1.282e-09 ***
## ratio                1 0.02148 0.02148  17.1591 4.298e-05 ***
## waist                1 0.01252 0.01252   9.9982 0.0017015 **
## time.ppn             1 0.00646 0.00646   5.1600 0.0237090 *
## X6L                  1 0.00315 0.00315   2.5120 0.1138731
## stab.glu:time.ppn    1 0.01383 0.01383  11.0471 0.0009806 ***
## stab.glu:age         1 0.00722 0.00722   5.7636 0.0168739 *
## age:ratio            1 0.00276 0.00276   2.2018 0.1387387
## Residuals          356 0.44574 0.00125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```