STA206 Assignment 7

Zhen Zhang

November 18, 2015

Problem 2

(a)

```r
diabetes <- read.table("diabetes.txt", na.strings = c("NA", ""), header = T)
```

(b)

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

(c)

```r
str(diabetes)
```

```
## 'data.frame':    403 obs. of  16 variables:
##  $ chol    : int  203 165 228 78 249 248 195 227 177 263 ...
##  $ stab.glu: int  82 97 92 93 90 94 92 75 87 89 ...
##  $ hdl     : int  56 24 37 12 28 69 41 44 49 40 ...
##  $ ratio   : num  3.6 6.9 6.2 6.5 8.9 ...
##  $ glyhb   : num  4.31 4.44 4.64 4.63 7.72 ...
##  $ location: Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  46 29 58 67 64 34 30 37 45 55 ...
##  $ gender  : Factor w/ 2 levels "female","male": 1 1 1 2 2 2 2 2 2 1 ...
##  $ height  : int  62 64 61 67 68 71 69 59 69 63 ...
##  $ weight  : int  121 218 256 119 183 190 191 170 166 202 ...
##  $ frame   : Factor w/ 3 levels "large","medium",..: 2 1 1 1 2 1 2 2 1 3 ...
##  $ bp.1s   : int  118 112 190 110 138 132 161 NA 160 108 ...
##  $ bp.1d   : int  59 68 92 50 80 86 112 NA 80 72 ...
##  $ waist   : int  29 46 49 33 44 36 46 34 34 45 ...
##  $ hip     : int  38 48 57 38 41 42 49 39 40 50 ...
##  $ time.ppn: int  720 360 180 480 300 195 720 1020 300 240 ...
```

```r
(quantitative_vars <- unlist(sapply(1:length(diabetes), function(i) {
    if (!is.factor(diabetes[[i]]))
        names(diabetes[i])
})))
```

```
##  [1] "chol"     "stab.glu" "hdl"      "ratio"    "glyhb"    "age"
##  [7] "height"   "weight"   "bp.1s"    "bp.1d"    "waist"    "hip"
## [13] "time.ppn"
```

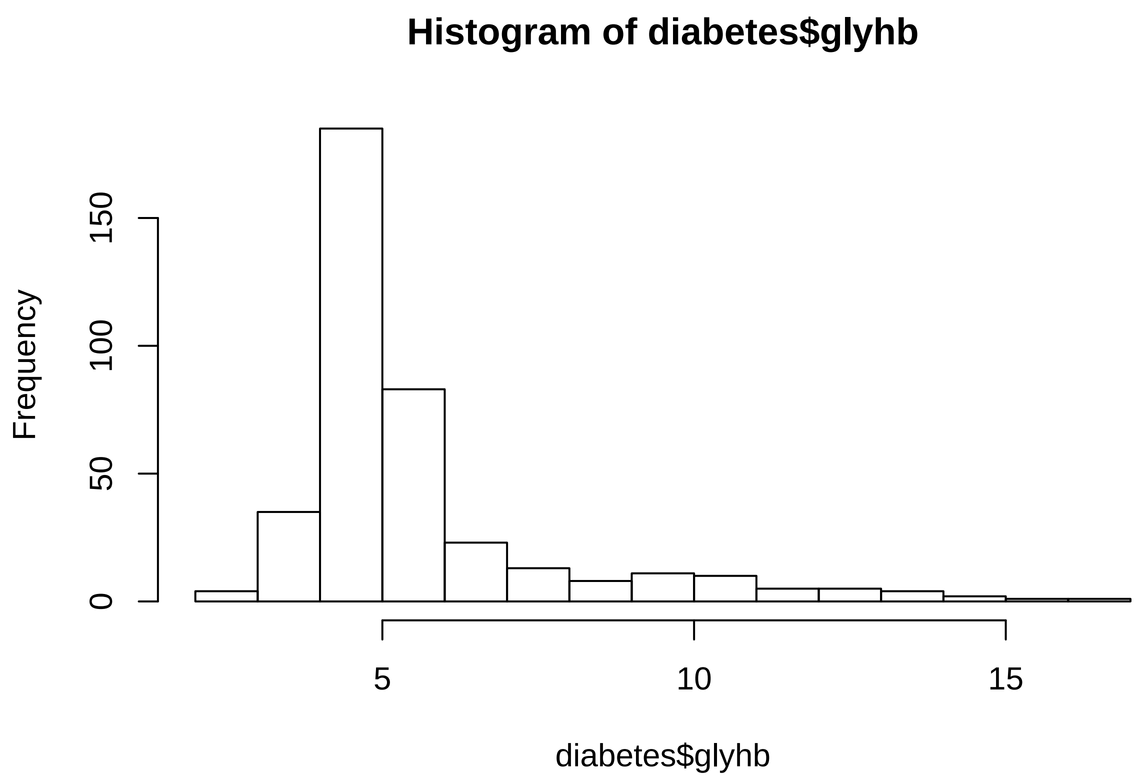
```r
(qualitative_vars <- unlist(sapply(1:length(diabetes), function(i) {
    if (is.factor(diabetes[[i]]))
        names(diabetes[i])
})))
```

```
## [1] "location" "gender"   "frame"
```

So quantitative variables: chol, stab.glu, hdl, atio, glyhb, age, height, weight, bp.1s, bp.1d, waist, hip, time.ppn

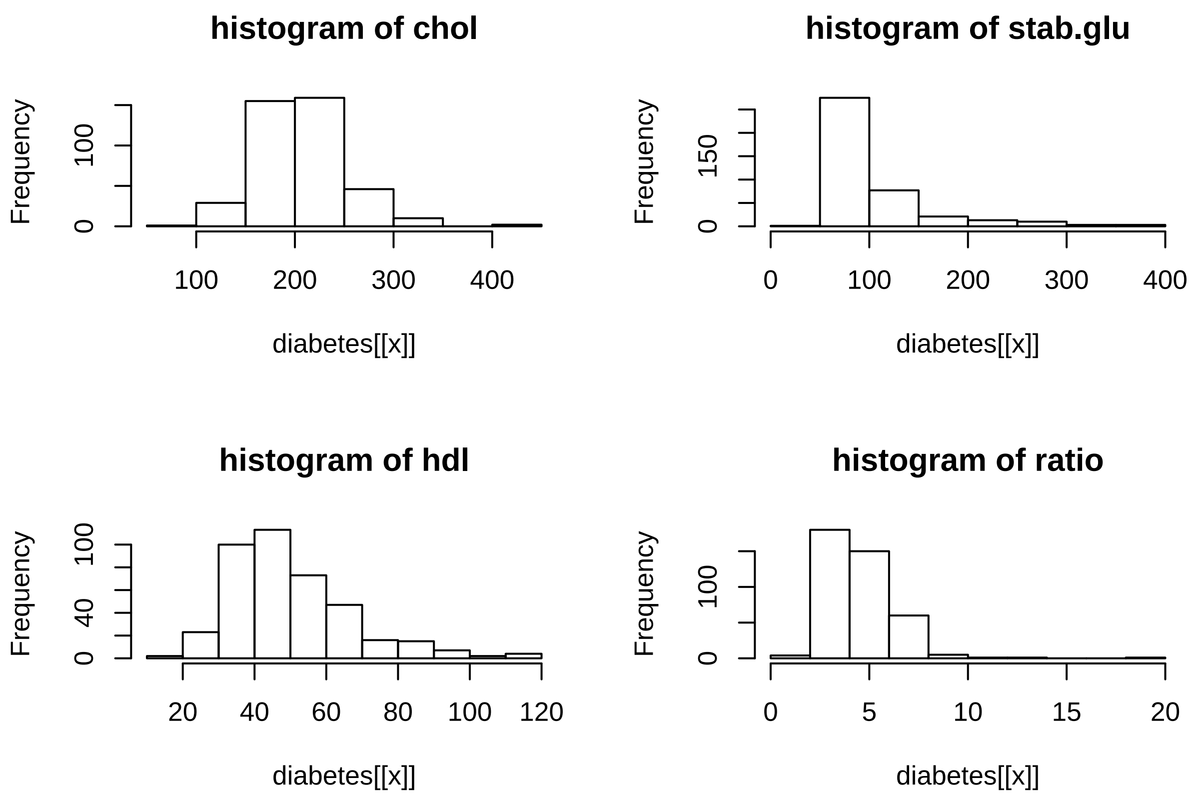
qualitative variables: location, gender, frame

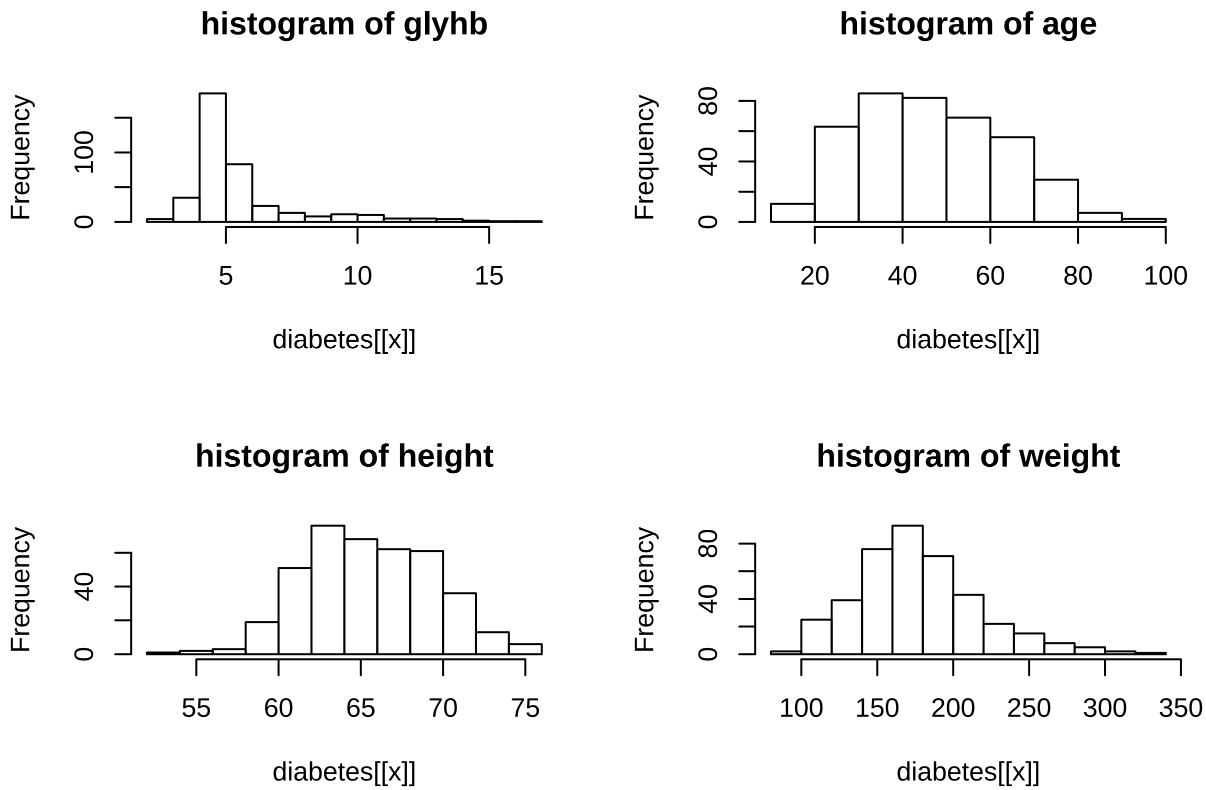
```r
hist(diabetes$glyhb)
```

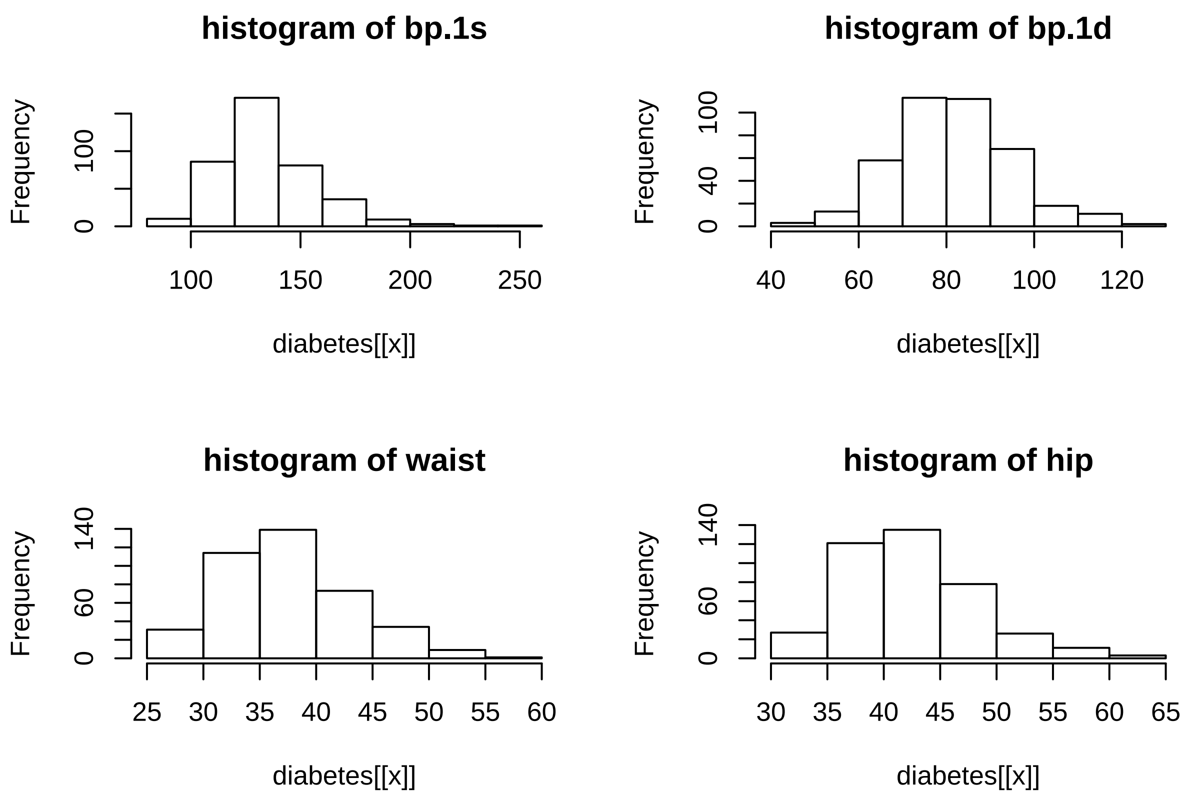


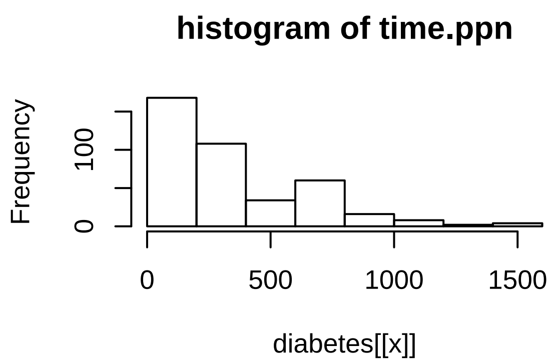
It is a little right skewed.

```r
par(mfrow = c(2, 2))
invisible(lapply(quantitative_vars, function(x) hist(diabetes[[x]], main = paste("histogram of",
    x))))
```

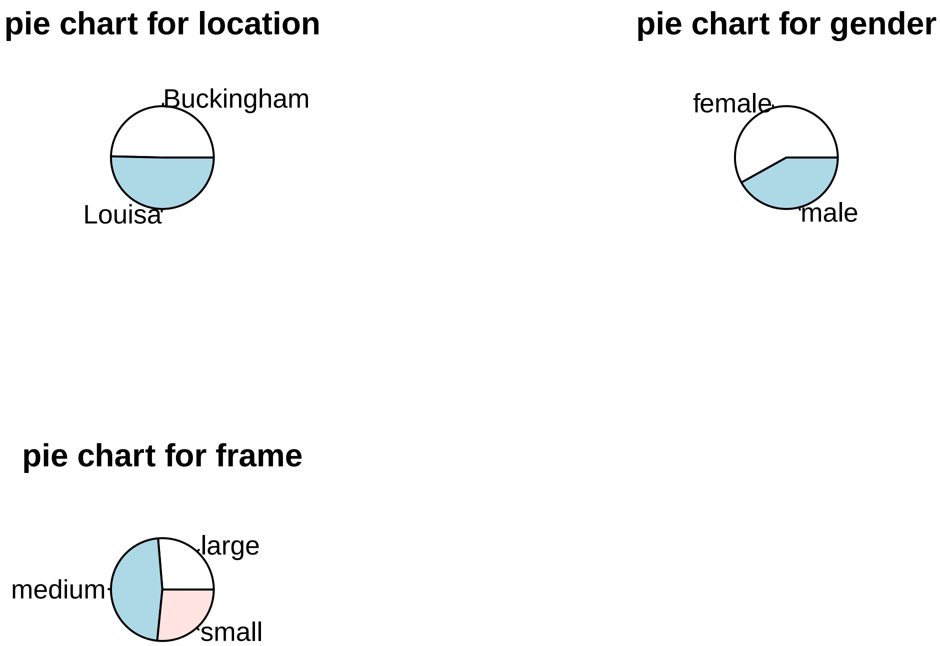

histogram of chol
Frequency
0
100
100
200
300
400
diabetes[[x]]
histogram of stab.glu
Frequency
0
150
0
100
200
300
400
diabetes[[x]]
histogram of hdl
Frequency
0
40
100
20
40
60
80
100
120
diabetes[[x]]
histogram of ratio
Frequency
0
100
0
5
10
15
20
diabetes[[x]]


histogram of glyhb
Frequency
0
100
5
10
15
diabetes[[x]]
histogram of age
Frequency
0
40
80
20
40
60
80
100
diabetes[[x]]
histogram of height
Frequency
0
40
55
60
65
70
75
diabetes[[x]]
histogram of weight
Frequency
0
40
80
100
150
200
250
300
350
diabetes[[x]]


histogram of bp.1s
Frequency
0
100
100
150
200
250
diabetes[[x]]
histogram of bp.1d
Frequency
0
40
100
40
60
80
100
120
diabetes[[x]]
histogram of waist
Frequency
0
60
140
25
30
35
40
45
50
55
60
diabetes[[x]]
histogram of hip
Frequency
0
60
140
30
35
40
45
50
55
60
65
diabetes[[x]]

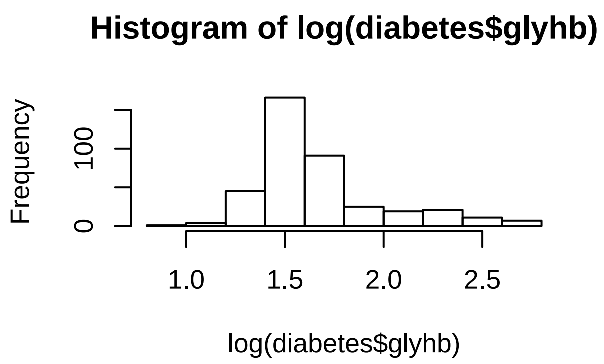


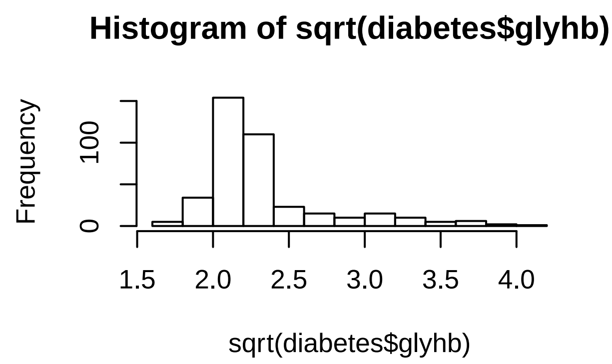

```r
par(mfrow = c(2, 2))
invisible(lapply(qualitative_vars, function(x) pie(table(diabetes[[x]]),
    main = paste("pie chart for", x))))
```

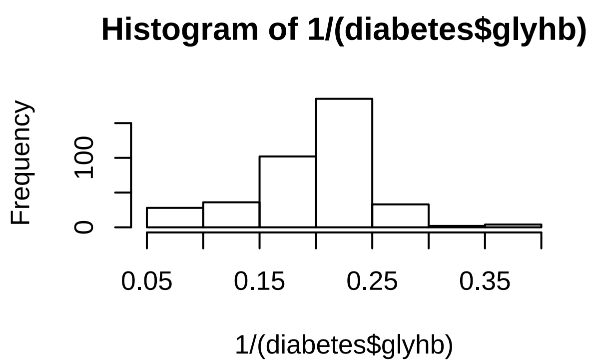


(d)

```r
par(mfrow = c(2, 2))
hist(log(diabetes$glyhb))
hist(sqrt(diabetes$glyhb))
hist(1/(diabetes$glyhb))
```







The last transformation, $\frac{1}{glyhb}$ appears to be the most Normal like among the three.

```
glyhb_trans <- 1/diabetes$glyhb
```

(e)

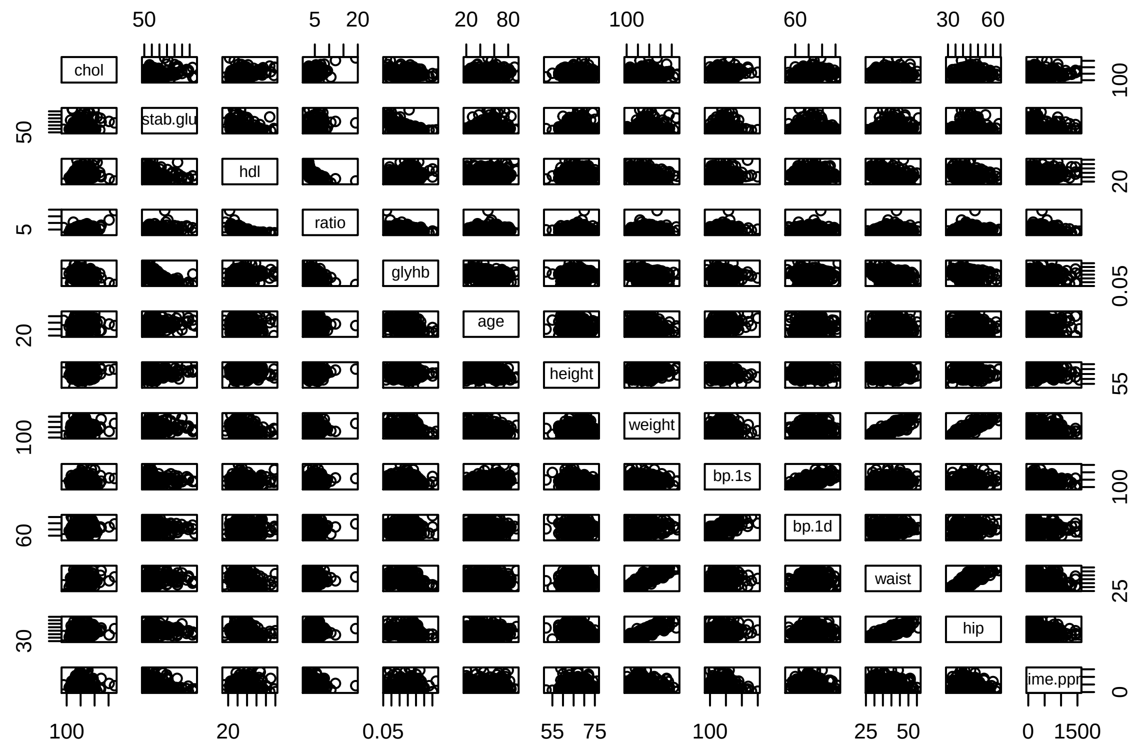
```
diabetes$glyhb <- glyhb_trans
```

(f)

```
index.na = apply(is.na(diabetes), 1, any)
diabetes.s = diabetes[index.na == FALSE, ]
```

(g)

```
pairs(diabetes.s[, quantitative_vars])
```




```r
cor(diabetes.s[, quantitative_vars])
```

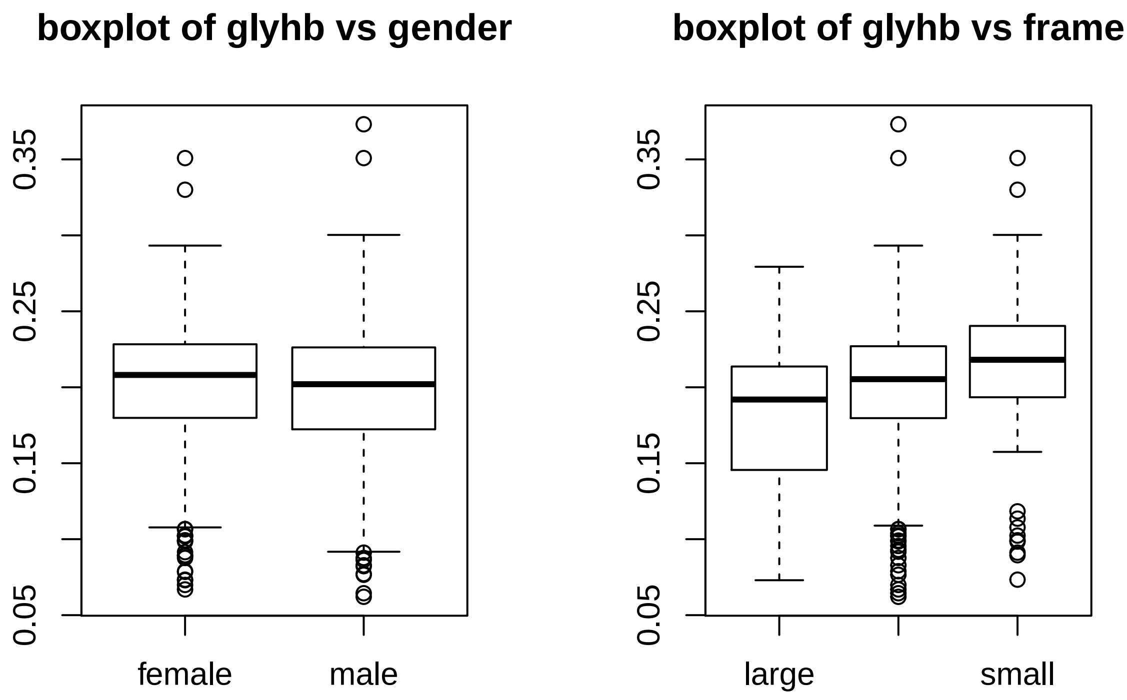
```
##                  chol    stab.glu           hdl       ratio
## chol      1.000000000  0.16544754  0.1709732770  0.48403807
## stab.glu  0.165447544  1.00000000 -0.1801048833  0.29889570
## hdl       0.170973277 -0.18010488  1.0000000000 -0.69023141
## ratio     0.484038069  0.29889570 -0.6902314087  1.00000000
## glyhb    -0.257440991 -0.64371727  0.1889598607 -0.35525846
## age       0.241604908  0.27855141  0.0002152264  0.17156914
## height   -0.063230009  0.08247570 -0.0685918173  0.07089817
## weight    0.079789987  0.18880052 -0.2829826752  0.27889889
## bp.1s     0.201948705  0.15142542  0.0295089053  0.10534657
## bp.1d     0.159042299  0.02569721  0.0722451474  0.03484142
## waist     0.144089547  0.23369209 -0.2783001009  0.31549761
## hip       0.098597154  0.14483314 -0.2222166064  0.20789160
## time.ppn  0.006238501 -0.04845774  0.0799388429 -0.05382831
##                glyhb          age       height      weight
## chol     -0.25744099  0.2416049084 -0.063230009  0.07978999
## stab.glu -0.64371727  0.2785514141  0.082475702  0.18880052
## hdl       0.18895986  0.0002152264 -0.068591817 -0.28298268
## ratio    -0.35525846  0.1715691447  0.070898165  0.27889889
## glyhb     1.00000000 -0.3956301899 -0.043229331 -0.21856483
## age      -0.39563019  1.0000000000 -0.097136587 -0.04621299
## height   -0.04322933 -0.0971365873  1.000000000  0.24329556
## weight   -0.21856483 -0.0462129859  0.243295558  1.00000000
```

```
## bp.1s    -0.22975720  0.4330322675 -0.044411815  0.09624288
## bp.1d    -0.05554035  0.0589147673  0.043452076  0.18050511
## waist    -0.31887439  0.1702608196  0.041807866  0.85192261
## hip      -0.21263079  0.0182966937 -0.117181984  0.82984527
## time.ppn -0.03620314 -0.0269049474 -0.006180895 -0.06221671
##                bp.1s       bp.1d       waist         hip     time.ppn
## chol      0.20194870  0.15904230  0.14408955  0.09859715  0.006238501
## stab.glu  0.15142542  0.02569721  0.23369209  0.14483314 -0.048457737
## hdl       0.02950891  0.07224515 -0.27830010 -0.22221661  0.079938843
## ratio     0.10534657  0.03484142  0.31549761  0.20789160 -0.053828314
## glyhb    -0.22975720 -0.05554035 -0.31887439 -0.21263079 -0.036203144
## age       0.43303227  0.05891477  0.17026082  0.01829669 -0.026904947
## height   -0.04441181  0.04345208  0.04180787 -0.11718198 -0.006180895
## weight    0.09624288  0.18050511  0.85192261  0.82984527 -0.062216714
## bp.1s     1.00000000  0.61984558  0.20976399  0.15142640 -0.074903689
## bp.1d     0.61984558  1.00000000  0.17899079  0.16282460 -0.063762636
## waist     0.20976399  0.17899079  1.00000000  0.83233707 -0.065861241
## hip       0.15142640  0.16282460  0.83233707  1.00000000 -0.092519540
## time.ppn -0.07490369 -0.06376264 -0.06586124 -0.09251954  1.000000000
```

Yes, I observe nonlinearity.

(h)

```r
par(mfrow = c(1, 2))
boxplot(glyhb ~ gender, diabetes.s, main = "boxplot of glyhb vs gender")
boxplot(glyhb ~ frame, diabetes.s, main = "boxplot of glyhb vs frame")
```



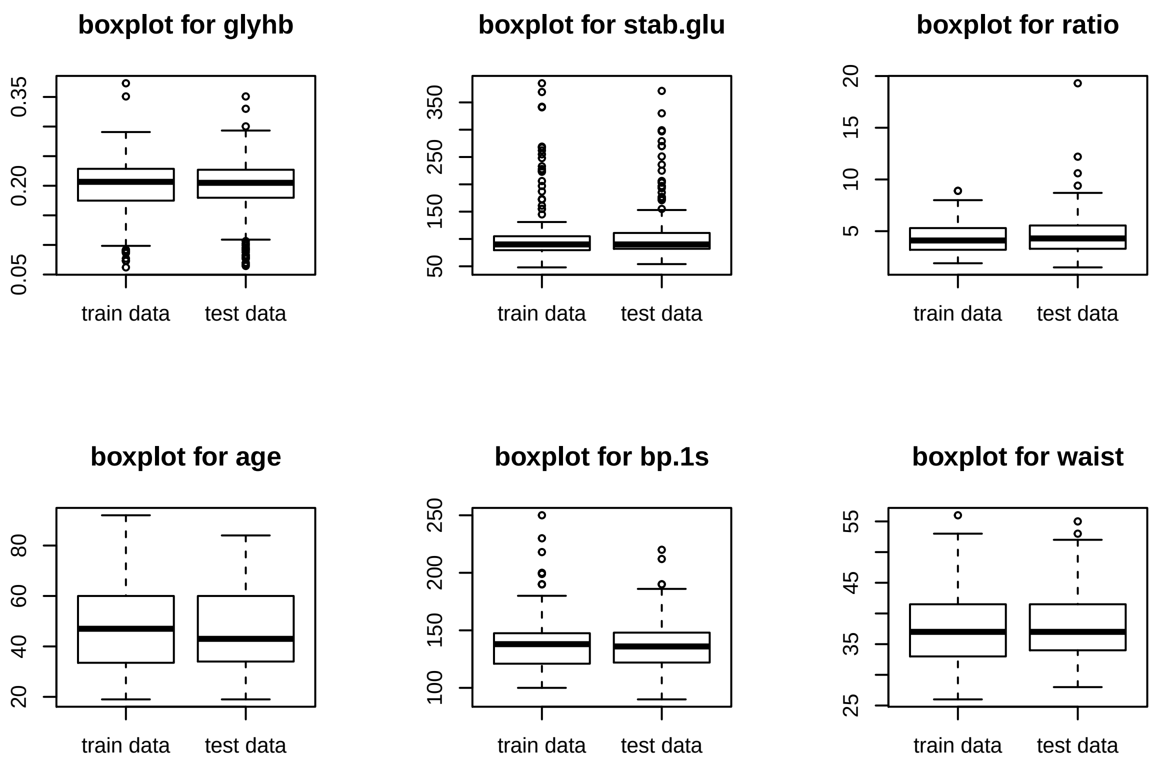
There is little difference related to female and male, but for frame, the mean of glyhb is increasing with respect to levels of large, medium and small.

(i)

```
set.seed(10)
n_samples <- nrow(diabetes.s)
sample_index <- sample(1:n_samples, n_samples/2)
diabetes.c = diabetes.s[sample_index, ]
diabetes.v = diabetes.s[-sample_index, ]
```

(j)

```
par(mfrow = c(2, 3))
invisible(lapply(c("glyhb", "stab.glu", "ratio", "age", "bp.1s", "waist"),
    function(x) {
        boxplot(diabetes.c[[x]], diabetes.v[[x]], names = c("train data",
            "test data"), main = paste("boxplot for", x))
    }))
```



Problem 3

(a)

```r
fit1 <- lm(glyhb ~ ., data = diabetes.c)
summary(fit1)
```

```
##
## Call:
## lm(formula = glyhb ~ ., data = diabetes.c)
##
## 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 *
```

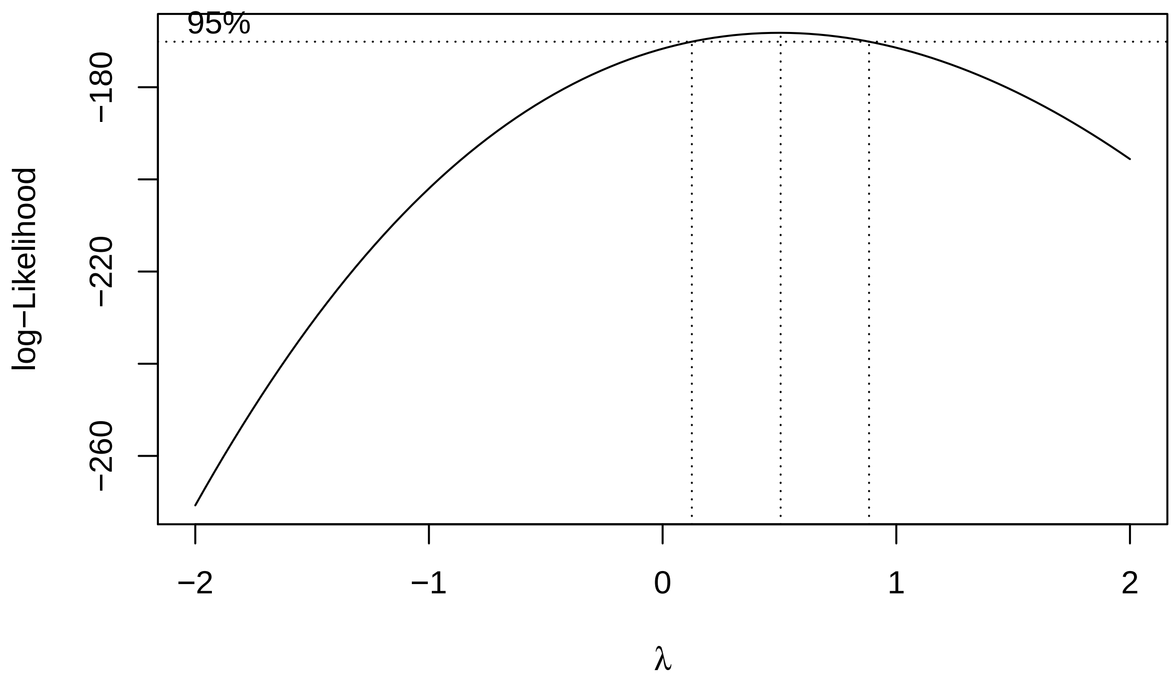
```
## gendermale     -7.422e-04  1.018e-02  -0.073   0.9420
## height         -1.212e-03  1.123e-03  -1.079   0.2820
## weight          2.210e-04  2.034e-04   1.087   0.2788
## framemedium     1.417e-03  7.861e-03   0.180   0.8572
## framesmall     -1.062e-02  9.596e-03  -1.107   0.2699
## bp.1s          -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
```

```r
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##            Df   Sum Sq  Mean Sq  F value    Pr(>F)
## chol        1 0.020540 0.020540  14.8423 0.0001667 ***
## stab.glu    1 0.216364 0.216364 156.3487 < 2.2e-16 ***
## hdl         1 0.005738 0.005738   4.1462 0.0433172 *
## ratio       1 0.000736 0.000736   0.5316 0.4669696
## location    1 0.000253 0.000253   0.1826 0.6696823
## age         1 0.022071 0.022071  15.9493 9.756e-05 ***
## gender      1 0.000131 0.000131   0.0949 0.7584701
## height      1 0.002088 0.002088   1.5091 0.2210117
## weight      1 0.003855 0.003855   2.7858 0.0969857 .
## frame       2 0.003052 0.001526   1.1028 0.3343538
## bp.1s       1 0.001462 0.001462   1.0563 0.3055641
## bp.1d       1 0.000169 0.000169   0.1225 0.7268181
## waist       1 0.005810 0.005810   4.1988 0.0420276 *
## hip         1 0.000920 0.000920   0.6651 0.4159399
## time.ppn    1 0.002954 0.002954   2.1347 0.1458854
## Residuals 166 0.229720 0.001384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are 17 regression coefficients. The MSE is 0.001384

```r
library(MASS)
boxcox(fit1)
```



A square root transformation is needed.

(b)

```
library(leaps)

sum_sub <- summary(regsubsets(glyhb ~ ., data = diabetes.c, nbest = 1,
    nvmax = 16))

n = nrow(diabetes.c)
p.m = as.integer(rownames(sum_sub$which)) + 1
ssto = sum((diabetes.c$glyhb - mean(diabetes.c$glyhb))^2)
sum_sub$sse = (1 - sum_sub$rsq) * ssto
sum_sub$aic = n * log(sum_sub$sse/n) + 2 * p.m


cri_sum_sub <- sapply(c("sse", "rsq", "adjr2", "cp", "aic", "bic"), function(x) {
    sum_sub[[x]]
})
cri_sum_sub
```

```
##              sse       rsq     adjr2          cp       aic        bic
##   [1,] 0.2864076 0.4448009 0.4417335 27.96351331 -1178.148  -97.26343
##   [2,] 0.2574112 0.5010102 0.4954659  9.01014928 -1195.682 -111.58759
##   [3,] 0.2428890 0.5291612 0.5212701  0.51619889 -1204.309 -117.00490
```

```
##  [4,] 0.2401432 0.5344840 0.5240230  0.53201659 -1204.389 -113.87598
##  [5,] 0.2367131 0.5411332 0.5281708  0.05337754 -1205.022 -111.29922
##  [6,] 0.2343460 0.5457220 0.5302352  0.34280455 -1204.861 -107.92898
##  [7,] 0.2331725 0.5479966 0.5299165  1.49487219 -1203.780 -103.63812
##  [8,] 0.2326634 0.5489836 0.5282473  3.12693590 -1202.180  -98.82868
##  [9,] 0.2314193 0.5513952 0.5280574  4.22797088 -1201.161  -94.60031
## [10,] 0.2303187 0.5535287 0.5275711  5.43265348 -1200.033  -90.26322
## [11,] 0.2300477 0.5540541 0.5253676  7.23678869 -1198.249  -85.26923
## [12,] 0.2299216 0.5542986 0.5228374  9.14564365 -1196.349  -80.16010
## [13,] 0.2298166 0.5545020 0.5202329 11.06983181 -1194.433  -75.03414
## [14,] 0.2297510 0.5546292 0.5175150 13.02241521 -1192.485  -69.87691
## [15,] 0.2297274 0.5546751 0.5146758 15.00531267 -1190.504  -64.68628
## [16,] 0.2297200 0.5546893 0.5117678 17.00000000 -1188.510  -59.48265
```

The best model according to each criterion:

```r
apply(cri_sum_sub[, c(1, 4, 5, 6)], 2, which.min)
```

```
## sse  cp aic bic
##  16   5   5   3
```

```r
apply(cri_sum_sub[, c(2, 3)], 2, which.max)
```

```
##   rsq adjr2
##    16     6
```

So the best model with predictors: sse: 16, cp: 5, aic: 5, bic: 3, rsq: 16, adjusted rsq: 6.

The best C_p value:

```r
cri_sum_sub[5, 4]
```

```
##         cp
## 0.05337754
```

The C_p value of the best model according C_p criterion is 0.05337754. It is the smallest value of C_p, and no overfitting due to smaller than p.

(c)

```r
stepAIC(lm(glyhb ~ 1, data = diabetes.c), scope = list(upper = lm(glyhb ~
    ., data = diabetes.c)), direction = "both")
```

```
## Start:  AIC=-1072.47
## glyhb ~ 1
##
##            Df Sum of Sq     RSS     AIC
## + stab.glu  1  0.229457 0.28641 -1178.2
## + age       1  0.080171 0.43569 -1101.4
## + ratio     1  0.062778 0.45309 -1094.2
## + waist     1  0.055768 0.46010 -1091.4
```

```
## + hdl       1  0.026343 0.48952 -1080.1
## + bp.1s     1  0.026201 0.48966 -1080.0
## + hip       1  0.022197 0.49367 -1078.5
## + chol      1  0.020540 0.49533 -1077.9
## + weight    1  0.019826 0.49604 -1077.6
## + frame     2  0.024818 0.49105 -1077.5
## <none>                 0.51586 -1072.5
## + bp.1d     1  0.001406 0.51446 -1071.0
## + gender    1  0.001168 0.51470 -1070.9
## + height    1  0.000406 0.51546 -1070.6
## + time.ppn  1  0.000253 0.51561 -1070.6
## + location  1  0.000223 0.51564 -1070.5
##
## Step:  AIC=-1178.15
## glyhb ~ stab.glu
##
##            Df Sum of Sq     RSS     AIC
## + age       1  0.028996 0.25741 -1195.7
## + waist     1  0.022234 0.26417 -1190.9
## + ratio     1  0.012237 0.27417 -1184.1
## + hip       1  0.010172 0.27624 -1182.8
## + bp.1s     1  0.010011 0.27640 -1182.7
## + chol      1  0.007447 0.27896 -1181.0
## + weight    1  0.005955 0.28045 -1180.0
## + hdl       1  0.003151 0.28326 -1178.2
## <none>                 0.28641 -1178.2
## + time.ppn  1  0.001218 0.28519 -1176.9
## + bp.1d     1  0.001132 0.28528 -1176.9
## + frame     2  0.003582 0.28283 -1176.5
## + location  1  0.000158 0.28625 -1176.2
## + height    1  0.000008 0.28640 -1176.2
## + gender    1  0.000005 0.28640 -1176.2
## - stab.glu  1  0.229457 0.51586 -1072.5
##
## Step:  AIC=-1195.68
## glyhb ~ stab.glu + age
##
##            Df Sum of Sq     RSS     AIC
## + waist     1  0.014522 0.24289 -1204.3
## + hip       1  0.010402 0.24701 -1201.2
## + weight    1  0.008376 0.24904 -1199.7
## + ratio     1  0.007022 0.25039 -1198.7
## + hdl       1  0.003946 0.25347 -1196.5
## <none>                 0.25741 -1195.7
## + chol      1  0.001726 0.25568 -1194.9
## + bp.1s     1  0.000870 0.25654 -1194.3
## + time.ppn  1  0.000797 0.25661 -1194.2
## + bp.1d     1  0.000563 0.25685 -1194.1
## + height    1  0.000525 0.25689 -1194.1
## + gender    1  0.000041 0.25737 -1193.7
## + location  1  0.000012 0.25740 -1193.7
## + frame     2  0.001194 0.25622 -1192.5
## - age       1  0.028996 0.28641 -1178.2
## - stab.glu  1  0.178283 0.43569 -1101.4
```

```
##
## Step:  AIC=-1204.31
## glyhb ~ stab.glu + age + waist
##
##            Df Sum of Sq     RSS     AIC
## + ratio     1  0.002746 0.24014 -1204.4
## <none>                  0.24289 -1204.3
## + time.ppn  1  0.001329 0.24156 -1203.3
## + chol      1  0.001173 0.24172 -1203.2
## + hdl       1  0.000947 0.24194 -1203.0
## + weight    1  0.000556 0.24233 -1202.7
## + bp.1s     1  0.000492 0.24240 -1202.7
## + height    1  0.000432 0.24246 -1202.6
## + bp.1d     1  0.000046 0.24284 -1202.3
## + gender    1  0.000037 0.24285 -1202.3
## + hip       1  0.000009 0.24288 -1202.3
## + location  1  0.000007 0.24288 -1202.3
## + frame     2  0.002491 0.24040 -1202.2
## - waist     1  0.014522 0.25741 -1195.7
## - age       1  0.021285 0.26417 -1190.9
## - stab.glu  1  0.160773 0.40366 -1113.3
##
## Step:  AIC=-1204.39
## glyhb ~ stab.glu + age + waist + ratio
##
##            Df Sum of Sq     RSS     AIC
## <none>                  0.24014 -1204.4
## - ratio     1  0.002746 0.24289 -1204.3
## + time.ppn  1  0.001658 0.23849 -1203.7
## + frame     2  0.003514 0.23663 -1203.1
## + weight    1  0.000726 0.23942 -1202.9
## + bp.1s     1  0.000666 0.23948 -1202.9
## + height    1  0.000443 0.23970 -1202.7
## + chol      1  0.000173 0.23997 -1202.5
## + hdl       1  0.000104 0.24004 -1202.5
## + hip       1  0.000052 0.24009 -1202.4
## + bp.1d     1  0.000038 0.24011 -1202.4
## + location  1  0.000005 0.24014 -1202.4
## + gender    1  0.000000 0.24014 -1202.4
## - waist     1  0.010246 0.25039 -1198.7
## - age       1  0.019240 0.25938 -1192.3
## - stab.glu  1  0.142762 0.38291 -1121.0
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = diabetes.c)
##
## Coefficients:
## (Intercept)     stab.glu          age        waist        ratio
##   0.3489987   -0.0005368   -0.0006412   -0.0013985   -0.0028483
```

```
fs1 <- lm(glyhb ~ stab.glu + age + waist + ratio, data = diabetes.c)
```

So the best model is: $glyhb = \beta_0 + \beta_1 stab.glu + \beta_2 age + \beta_3 waist + \beta_4 ratio$, and the corresponding AIC is

-1204.39

The best model's AIC in regsubsets is

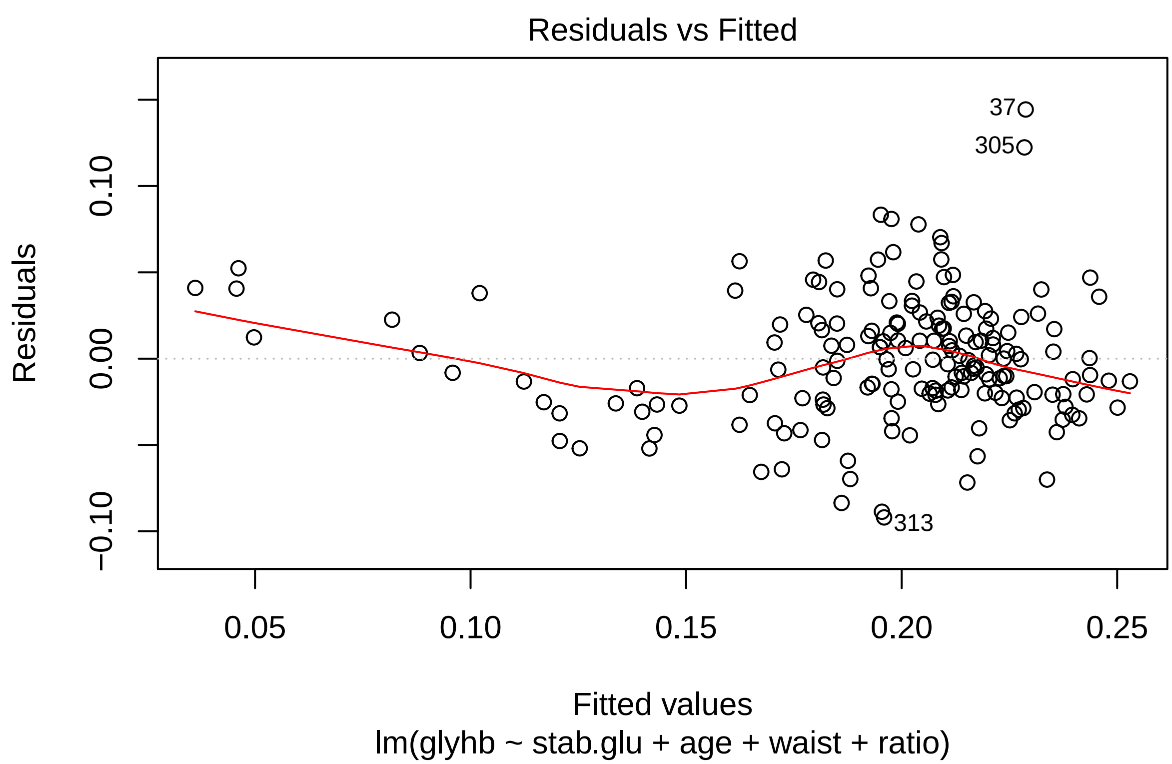
```r
cri_sum_sub[5, 5]
```

```
##       aic
## -1205.022
```

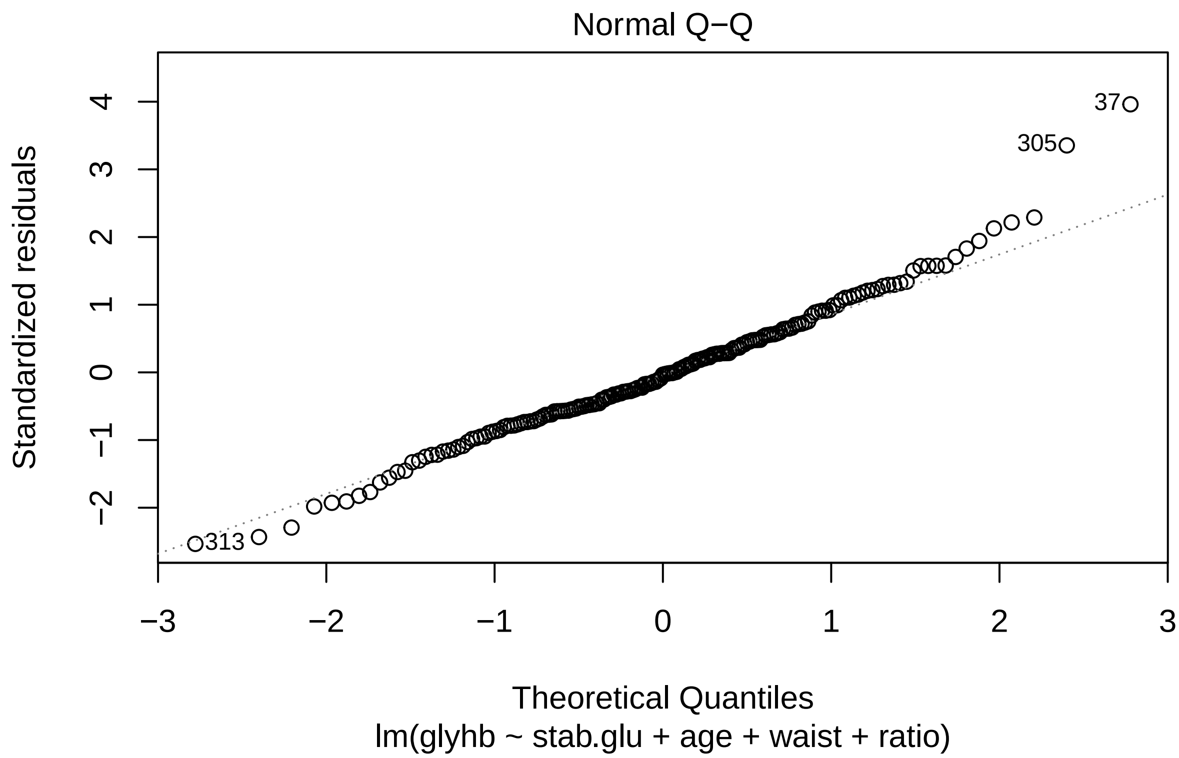
They are not identical, and have a large distance.

(d)

```r
plot(fs1, which = 1)
```




```r
plot(fs1, which = 2)
```



The residual vs. fitted value plot shows a square tendancy. The residual Q-Q plot is a little heavy tailed. It seems to be adequate.

Problem 4

(a)

```r
fit2 <- lm(glyhb ~ .^2, data = diabetes.c)
summary(fit2)
```

```
##
## Call:
## lm(formula = glyhb ~ .^2, data = diabetes.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.066819 -0.009752 -0.001635  0.009280  0.042555
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.133e+00  3.228e+00   0.971  0.33672
## chol                      -5.600e-03  1.225e-02  -0.457  0.64977
## stab.glu                  -1.371e-02  6.817e-03  -2.012  0.05000 .
## hdl                       -3.071e-02  3.928e-02  -0.782  0.43818
## ratio                     -2.814e-01  4.942e-01  -0.569  0.57181
```

```
## locationLouisa           -4.790e-02  3.061e-01  -0.156  0.87634
## age                       1.686e-03  1.540e-02   0.109  0.91329
## gendermale                6.657e-02  4.405e-01   0.151  0.88052
## height                   -2.639e-02  4.756e-02  -0.555  0.58160
## weight                    1.166e-02  9.517e-03   1.225  0.22670
## framemedium               3.967e-01  5.944e-01   0.667  0.50782
## framesmall                1.792e-01  7.073e-01   0.253  0.80110
## bp.1s                     1.742e-02  1.326e-02   1.314  0.19519
## bp.1d                    -3.251e-02  1.709e-02  -1.903  0.06322 .
## waist                    -9.879e-02  7.367e-02  -1.341  0.18635
## hip                       2.880e-02  8.062e-02   0.357  0.72249
## time.ppn                  1.749e-05  7.376e-04   0.024  0.98118
## chol:stab.glu            -1.651e-05  7.115e-06  -2.320  0.02473 *
## chol:hdl                  1.223e-05  1.497e-05   0.817  0.41810
## chol:ratio                2.885e-05  1.419e-04   0.203  0.83974
## chol:locationLouisa       1.363e-04  6.579e-04   0.207  0.83677
## chol:age                  1.249e-06  3.252e-05   0.038  0.96953
## chol:gendermale          -1.745e-03  1.635e-03  -1.067  0.29127
## chol:height               2.877e-05  1.987e-04   0.145  0.88548
## chol:weight               1.342e-05  2.954e-05   0.454  0.65164
## chol:framemedium          1.649e-03  1.034e-03   1.594  0.11761
## chol:framesmall          -1.068e-03  1.493e-03  -0.715  0.47791
## chol:bp.1s                3.760e-05  2.805e-05   1.340  0.18661
## chol:bp.1d               -7.914e-05  3.697e-05  -2.141  0.03753 *
## chol:waist               -5.654e-05  1.448e-04  -0.391  0.69788
## chol:hip                 -5.507e-05  1.989e-04  -0.277  0.78312
## chol:time.ppn             2.964e-06  1.286e-06   2.305  0.02560 *
## stab.glu:hdl              1.060e-04  3.309e-05   3.203  0.00244 **
## stab.glu:ratio            5.812e-04  2.725e-04   2.133  0.03820 *
## stab.glu:locationLouisa  -6.171e-04  3.770e-04  -1.637  0.10840
## stab.glu:age              1.548e-05  1.503e-05   1.030  0.30825
## stab.glu:gendermale       2.875e-04  5.597e-04   0.514  0.60983
## stab.glu:height           1.359e-04  8.474e-05   1.603  0.11553
## stab.glu:weight           2.369e-06  1.416e-05   0.167  0.86791
## stab.glu:framemedium      6.882e-04  2.616e-04   2.631  0.01147 *
## stab.glu:framesmall      -4.643e-04  5.573e-04  -0.833  0.40897
## stab.glu:bp.1s            1.965e-05  1.013e-05   1.940  0.05837 .
## stab.glu:bp.1d           -2.713e-05  1.780e-05  -1.524  0.13416
## stab.glu:waist           -1.149e-04  7.863e-05  -1.461  0.15079
## stab.glu:hip              6.871e-05  9.218e-05   0.745  0.45975
## stab.glu:time.ppn        -7.964e-07  5.862e-07  -1.359  0.18078
## hdl:ratio                 6.803e-03  1.567e-03   4.342 7.48e-05 ***
## hdl:locationLouisa       -7.939e-04  2.411e-03  -0.329  0.74338
## hdl:age                  -1.519e-06  1.041e-04  -0.015  0.98842
## hdl:gendermale            6.565e-03  5.079e-03   1.292  0.20253
## hdl:height                7.588e-05  6.148e-04   0.123  0.90229
## hdl:weight               -1.112e-04  9.666e-05  -1.150  0.25596
## hdl:framemedium          -5.792e-03  4.096e-03  -1.414  0.16396
## hdl:framesmall            3.091e-03  4.978e-03   0.621  0.53761
## hdl:bp.1s                -1.300e-04  8.506e-05  -1.529  0.13301
## hdl:bp.1d                 2.471e-04  1.234e-04   2.003  0.05101 .
## hdl:waist                 1.273e-04  4.764e-04   0.267  0.79045
## hdl:hip                   6.354e-04  7.490e-04   0.848  0.40056
## hdl:time.ppn             -8.101e-08  3.975e-06  -0.020  0.98383
```

```
## ratio:locationLouisa        9.351e-03  2.688e-02   0.348  0.72944
## ratio:age                  -2.823e-05  1.272e-03  -0.022  0.98238
## ratio:gendermale            7.991e-02  6.310e-02   1.266  0.21160
## ratio:height                6.270e-04  7.839e-03   0.080  0.93659
## ratio:weight               -1.127e-03  1.051e-03  -1.072  0.28903
## ratio:framemedium          -6.235e-02  4.170e-02  -1.495  0.14151
## ratio:framesmall            6.402e-02  6.173e-02   1.037  0.30496
## ratio:bp.1s                -1.353e-03  1.086e-03  -1.247  0.21875
## ratio:bp.1d                 3.285e-03  1.587e-03   2.070  0.04395 *
## ratio:waist                 3.240e-03  5.374e-03   0.603  0.54949
## ratio:hip                   3.986e-03  8.802e-03   0.453  0.65274
## ratio:time.ppn             -2.417e-05  5.270e-05  -0.459  0.64857
## locationLouisa:age         -8.612e-04  8.731e-04  -0.986  0.32900
## locationLouisa:gendermale  -2.863e-02  4.080e-02  -0.702  0.48636
## locationLouisa:height       3.606e-03  4.287e-03   0.841  0.40455
## locationLouisa:weight       5.835e-04  8.539e-04   0.683  0.49777
## locationLouisa:framemedium -4.932e-02  2.260e-02  -2.183  0.03407 *
## locationLouisa:framesmall  -4.514e-02  3.267e-02  -1.382  0.17359
## locationLouisa:bp.1s        5.018e-04  8.417e-04   0.596  0.55390
## locationLouisa:bp.1d        1.213e-04  9.881e-04   0.123  0.90281
## locationLouisa:waist       -2.624e-03  5.096e-03  -0.515  0.60896
## locationLouisa:hip         -3.089e-03  5.578e-03  -0.554  0.58237
## locationLouisa:time.ppn    -4.859e-05  3.847e-05  -1.263  0.21287
## age:gendermale              8.144e-04  1.505e-03   0.541  0.59101
## age:height                 -1.480e-04  1.936e-04  -0.764  0.44845
## age:weight                  2.218e-05  2.516e-05   0.882  0.38246
## age:framemedium             2.234e-03  1.038e-03   2.151  0.03661 *
## age:framesmall              1.439e-03  1.281e-03   1.123  0.26709
## age:bp.1s                   1.296e-05  2.180e-05   0.594  0.55514
## age:bp.1d                  -7.092e-05  3.378e-05  -2.100  0.04116 *
## age:waist                  -7.216e-05  1.535e-04  -0.470  0.64049
## age:hip                     1.884e-04  2.079e-04   0.906  0.36945
## age:time.ppn               -1.821e-06  1.318e-06  -1.381  0.17371
## gendermale:height          -2.170e-03  7.105e-03  -0.305  0.76140
## gendermale:weight           2.621e-03  1.316e-03   1.992  0.05221 .
## gendermale:framemedium     -6.525e-02  5.769e-02  -1.131  0.26375
## gendermale:framesmall      -8.914e-02  8.423e-02  -1.058  0.29533
## gendermale:bp.1s            1.243e-03  1.201e-03   1.035  0.30593
## gendermale:bp.1d           -9.349e-04  1.657e-03  -0.564  0.57524
## gendermale:waist           -1.999e-02  7.670e-03  -2.606  0.01223 *
## gendermale:hip             -2.174e-03  1.118e-02  -0.194  0.84668
## gendermale:time.ppn         8.226e-05  8.700e-05   0.945  0.34925
## height:weight              -1.166e-04  1.488e-04  -0.783  0.43728
## height:framemedium          6.689e-04  6.758e-03   0.099  0.92158
## height:framesmall           1.854e-03  8.244e-03   0.225  0.82301
## height:bp.1s               -1.200e-04  1.706e-04  -0.704  0.48499
## height:bp.1d                1.843e-04  2.158e-04   0.854  0.39739
## height:waist                1.957e-03  9.857e-04   1.985  0.05296 .
## height:hip                 -9.591e-04  1.142e-03  -0.840  0.40523
## height:time.ppn            -3.500e-06  9.849e-06  -0.355  0.72390
## weight:framemedium          2.866e-03  1.263e-03   2.269  0.02792 *
## weight:framesmall           2.665e-03  1.645e-03   1.621  0.11177
## weight:bp.1s                7.445e-06  2.155e-05   0.345  0.73132
## weight:bp.1d               -8.075e-05  3.548e-05  -2.276  0.02744 *
```

```
## weight:waist               2.229e-05  8.586e-05   0.260  0.79627
## weight:hip                 9.717e-05  6.235e-05   1.558  0.12586
## weight:time.ppn           -1.861e-08  1.731e-06  -0.011  0.99147
## framemedium:bp.1s         -3.618e-04  8.793e-04  -0.411  0.68258
## framesmall:bp.1s          -7.127e-04  1.196e-03  -0.596  0.55405
## framemedium:bp.1d         -8.529e-04  1.192e-03  -0.716  0.47775
## framesmall:bp.1d          -8.671e-04  1.410e-03  -0.615  0.54158
## framemedium:waist         -1.415e-02  6.071e-03  -2.331  0.02410 *
## framesmall:waist          -1.275e-02  8.027e-03  -1.589  0.11881
## framemedium:hip           -5.645e-03  8.292e-03  -0.681  0.49933
## framesmall:hip            -7.836e-03  1.190e-02  -0.659  0.51329
## framemedium:time.ppn       8.785e-05  4.999e-05   1.757  0.08535 .
## framesmall:time.ppn        1.704e-05  5.969e-05   0.285  0.77660
## bp.1s:bp.1d                7.517e-06  1.850e-05   0.406  0.68641
## bp.1s:waist               -1.802e-04  1.412e-04  -1.276  0.20816
## bp.1s:hip                 -7.468e-05  1.485e-04  -0.503  0.61748
## bp.1s:time.ppn            -1.016e-07  1.677e-06  -0.061  0.95195
## bp.1d:waist                5.677e-04  2.473e-04   2.296  0.02621 *
## bp.1d:hip                  2.241e-04  2.460e-04   0.911  0.36690
## bp.1d:time.ppn            -2.496e-06  1.991e-06  -1.254  0.21620
## waist:hip                 -9.113e-04  6.661e-04  -1.368  0.17775
## waist:time.ppn             1.247e-05  6.340e-06   1.967  0.05513 .
## hip:time.ppn              -1.024e-05  1.244e-05  -0.823  0.41440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03219 on 47 degrees of freedom
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.6345
## F-statistic:  3.34 on 135 and 47 DF,  p-value: 4.097e-06
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                    Df   Sum Sq  Mean Sq  F value    Pr(>F)
## chol                1 0.020540 0.020540  19.8242 5.210e-05 ***
## stab.glu            1 0.216364 0.216364 208.8278 < 2.2e-16 ***
## hdl                 1 0.005738 0.005738   5.5379  0.022844 *
## ratio               1 0.000736 0.000736   0.7100  0.403708
## location            1 0.000253 0.000253   0.2439  0.623688
## age                 1 0.022071 0.022071  21.3027 3.046e-05 ***
## gender              1 0.000131 0.000131   0.1267  0.723467
## height              1 0.002088 0.002088   2.0157  0.162282
## weight              1 0.003855 0.003855   3.7209  0.059786 .
## frame               2 0.003052 0.001526   1.4730  0.239629
## bp.1s               1 0.001462 0.001462   1.4108  0.240891
## bp.1d               1 0.000169 0.000169   0.1636  0.687723
## waist               1 0.005810 0.005810   5.6081  0.022043 *
## hip                 1 0.000920 0.000920   0.8883  0.350753
## time.ppn            1 0.002954 0.002954   2.8513  0.097927 .
## chol:stab.glu       1 0.000000 0.000000   0.0001  0.992975
## chol:hdl            1 0.002009 0.002009   1.9390  0.170328
## chol:ratio          1 0.000544 0.000544   0.5246  0.472460
```

```
## chol:location      1 0.000669 0.000669   0.6457  0.425704
## chol:age           1 0.000185 0.000185   0.1786  0.674536
## chol:gender        1 0.000606 0.000606   0.5844  0.448405
## chol:height        1 0.000000 0.000000   0.0000  0.996057
## chol:weight        1 0.000497 0.000497   0.4797  0.491981
## chol:frame         2 0.000364 0.000182   0.1754  0.839640
## chol:bp.1s         1 0.002011 0.002011   1.9409  0.170130
## chol:bp.1d         1 0.000432 0.000432   0.4170  0.521599
## chol:waist         1 0.000282 0.000282   0.2717  0.604610
## chol:hip           1 0.000558 0.000558   0.5387  0.466626
## chol:time.ppn      1 0.000020 0.000020   0.0193  0.890048
## stab.glu:hdl       1 0.003234 0.003234   3.1212  0.083772 .
## stab.glu:ratio     1 0.001253 0.001253   1.2094  0.277059
## stab.glu:location  1 0.000021 0.000021   0.0206  0.886467
## stab.glu:age       1 0.003429 0.003429   3.3092  0.075268 .
## stab.glu:gender    1 0.006981 0.006981   6.7379  0.012555 *
## stab.glu:height    1 0.000601 0.000601   0.5796  0.450256
## stab.glu:weight    1 0.005588 0.005588   5.3935  0.024593 *
## stab.glu:frame     2 0.002164 0.001082   1.0442  0.359999
## stab.glu:bp.1s     1 0.000088 0.000088   0.0851  0.771837
## stab.glu:bp.1d     1 0.011502 0.011502  11.1014  0.001687 **
## stab.glu:waist     1 0.001273 0.001273   1.2282  0.273394
## stab.glu:hip       1 0.001531 0.001531   1.4779  0.230177
## stab.glu:time.ppn  1 0.007040 0.007040   6.7945  0.012213 *
## hdl:ratio          1 0.009196 0.009196   8.8756  0.004562 **
## hdl:location       1 0.001756 0.001756   1.6944  0.199367
## hdl:age            1 0.000031 0.000031   0.0301  0.862907
## hdl:gender         1 0.000611 0.000611   0.5902  0.446195
## hdl:height         1 0.000531 0.000531   0.5126  0.477575
## hdl:weight         1 0.000830 0.000830   0.8009  0.375375
## hdl:frame          2 0.001538 0.000769   0.7421  0.481632
## hdl:bp.1s          1 0.001250 0.001250   1.2063  0.277651
## hdl:bp.1d          1 0.000555 0.000555   0.5355  0.467933
## hdl:waist          1 0.000083 0.000083   0.0801  0.778362
## hdl:hip            1 0.001531 0.001531   1.4776  0.230228
## hdl:time.ppn       1 0.000778 0.000778   0.7511  0.390529
## ratio:location     1 0.000159 0.000159   0.1530  0.697446
## ratio:age          1 0.000979 0.000979   0.9447  0.336054
## ratio:gender       1 0.000010 0.000010   0.0093  0.923400
## ratio:height       1 0.002068 0.002068   1.9963  0.164276
## ratio:weight       1 0.000296 0.000296   0.2854  0.595715
## ratio:frame        2 0.002709 0.001355   1.3075  0.280151
## ratio:bp.1s        1 0.000995 0.000995   0.9603  0.332141
## ratio:bp.1d        1 0.001288 0.001288   1.2433  0.270500
## ratio:waist        1 0.000193 0.000193   0.1860  0.668231
## ratio:hip          1 0.000336 0.000336   0.3242  0.571806
## ratio:time.ppn     1 0.000810 0.000810   0.7814  0.381199
## location:age       1 0.001914 0.001914   1.8472  0.180596
## location:gender    1 0.002855 0.002855   2.7554  0.103585
## location:height    1 0.000252 0.000252   0.2429  0.624443
## location:weight    1 0.000287 0.000287   0.2769  0.601194
## location:frame     2 0.005886 0.002943   2.8404  0.068462 .
## location:bp.1s     1 0.002612 0.002612   2.5213  0.119027
## location:bp.1d     1 0.004248 0.004248   4.0999  0.048588 *
```

```
## location:waist     1 0.000008 0.000008   0.0081  0.928760
## location:hip       1 0.000311 0.000311   0.3000  0.586479
## location:time.ppn  1 0.003002 0.003002   2.8975  0.095323 .
## age:gender         1 0.004171 0.004171   4.0256  0.050589 .
## age:height         1 0.000051 0.000051   0.0495  0.824942
## age:weight         1 0.004140 0.004140   3.9959  0.051411 .
## age:frame          2 0.003321 0.001660   1.6026  0.212187
## age:bp.1s          1 0.004507 0.004507   4.3504  0.042454 *
## age:bp.1d          1 0.000296 0.000296   0.2854  0.595683
## age:waist          1 0.000028 0.000028   0.0270  0.870278
## age:hip            1 0.003644 0.003644   3.5168  0.066971 .
## age:time.ppn       1 0.001656 0.001656   1.5983  0.212376
## gender:height      1 0.002553 0.002553   2.4645  0.123151
## gender:weight      1 0.001568 0.001568   1.5131  0.224786
## gender:frame       2 0.003899 0.001949   1.8815  0.163666
## gender:bp.1s       1 0.000540 0.000540   0.5213  0.473867
## gender:bp.1d       1 0.000522 0.000522   0.5042  0.481156
## gender:waist       1 0.000373 0.000373   0.3596  0.551590
## gender:hip         1 0.000235 0.000235   0.2270  0.635949
## gender:time.ppn    1 0.000809 0.000809   0.7805  0.381484
## height:weight      1 0.002392 0.002392   2.3088  0.135340
## height:frame       2 0.000535 0.000267   0.2582  0.773558
## height:bp.1s       1 0.000660 0.000660   0.6373  0.428693
## height:bp.1d       1 0.000595 0.000595   0.5743  0.452333
## height:waist       1 0.000028 0.000028   0.0271  0.869881
## height:hip         1 0.000530 0.000530   0.5115  0.478015
## height:time.ppn    1 0.001833 0.001833   1.7696  0.189847
## weight:frame       2 0.004258 0.002129   2.0550  0.139445
## weight:bp.1s       1 0.002210 0.002210   2.1332  0.150796
## weight:bp.1d       1 0.000231 0.000231   0.2228  0.639120
## weight:waist       1 0.000315 0.000315   0.3041  0.583920
## weight:hip         1 0.000263 0.000263   0.2542  0.616491
## weight:time.ppn    1 0.000003 0.000003   0.0033  0.954428
## frame:bp.1s        2 0.000181 0.000090   0.0873  0.916570
## frame:bp.1d        2 0.004412 0.002206   2.1290  0.130291
## frame:waist        2 0.005818 0.002909   2.8078  0.070490 .
## frame:hip          2 0.000053 0.000027   0.0258  0.974530
## frame:time.ppn     2 0.003876 0.001938   1.8706  0.165321
## bp.1s:bp.1d        1 0.001695 0.001695   1.6364  0.207101
## bp.1s:waist        1 0.000049 0.000049   0.0475  0.828502
## bp.1s:hip          1 0.000324 0.000324   0.3131  0.578426
## bp.1s:time.ppn     1 0.001176 0.001176   1.1353  0.292090
## bp.1d:waist        1 0.006527 0.006527   6.2996  0.015578 *
## bp.1d:hip          1 0.000183 0.000183   0.1767  0.676105
## bp.1d:time.ppn     1 0.000829 0.000829   0.8003  0.375553
## waist:hip          1 0.002689 0.002689   2.5958  0.113844
## waist:time.ppn     1 0.003554 0.003554   3.4300  0.070308 .
## hip:time.ppn       1 0.000703 0.000703   0.6781  0.414396
## Residuals         47 0.048696 0.001036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MSE is 0.001036. There are 136 regression coefficients.

I have concern of overfitting.

(b)

```r
stepAIC(lm(glyhb ~ 1, data = diabetes.c), scope = list(upper = lm(glyhb ~
    .^2, data = diabetes.c)), direction = "both")
```

```
## Start:  AIC=-1072.47
## glyhb ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + stab.glu   1  0.229457 0.28641 -1178.2
## + age        1  0.080171 0.43569 -1101.4
## + ratio      1  0.062778 0.45309 -1094.2
## + waist      1  0.055768 0.46010 -1091.4
## + hdl        1  0.026343 0.48952 -1080.1
## + bp.1s      1  0.026201 0.48966 -1080.0
## + hip        1  0.022197 0.49367 -1078.5
## + chol       1  0.020540 0.49533 -1077.9
## + weight     1  0.019826 0.49604 -1077.6
## + frame      2  0.024818 0.49105 -1077.5
## <none>                  0.51586 -1072.5
## + bp.1d      1  0.001406 0.51446 -1071.0
## + gender     1  0.001168 0.51470 -1070.9
## + height     1  0.000406 0.51546 -1070.6
## + time.ppn   1  0.000253 0.51561 -1070.6
## + location   1  0.000223 0.51564 -1070.5
##
## Step:  AIC=-1178.15
## glyhb ~ stab.glu
##
##              Df Sum of Sq     RSS     AIC
## + age        1  0.028996 0.25741 -1195.7
## + waist      1  0.022234 0.26417 -1190.9
## + ratio      1  0.012237 0.27417 -1184.1
## + hip        1  0.010172 0.27624 -1182.8
## + bp.1s      1  0.010011 0.27640 -1182.7
## + chol       1  0.007447 0.27896 -1181.0
## + weight     1  0.005955 0.28045 -1180.0
## + hdl        1  0.003151 0.28326 -1178.2
## <none>                  0.28641 -1178.2
## + time.ppn   1  0.001218 0.28519 -1176.9
## + bp.1d      1  0.001132 0.28528 -1176.9
## + frame      2  0.003582 0.28283 -1176.5
## + location   1  0.000158 0.28625 -1176.2
## + height     1  0.000008 0.28640 -1176.2
## + gender     1  0.000005 0.28640 -1176.2
## - stab.glu   1  0.229457 0.51586 -1072.5
##
## Step:  AIC=-1195.68
## glyhb ~ stab.glu + age
##
##                  Df Sum of Sq     RSS     AIC
## + waist          1  0.014522 0.24289 -1204.3
## + hip            1  0.010402 0.24701 -1201.2
## + weight         1  0.008376 0.24904 -1199.7
```

```
## + ratio         1  0.007022 0.25039 -1198.7
## + hdl           1  0.003946 0.25347 -1196.5
## + stab.glu:age  1  0.002815 0.25460 -1195.7
## <none>                      0.25741 -1195.7
## + chol          1  0.001726 0.25568 -1194.9
## + bp.1s         1  0.000870 0.25654 -1194.3
## + time.ppn      1  0.000797 0.25661 -1194.2
## + bp.1d         1  0.000563 0.25685 -1194.1
## + height        1  0.000525 0.25689 -1194.1
## + gender        1  0.000041 0.25737 -1193.7
## + location      1  0.000012 0.25740 -1193.7
## + frame         2  0.001194 0.25622 -1192.5
## - age           1  0.028996 0.28641 -1178.2
## - stab.glu      1  0.178283 0.43569 -1101.4
##
## Step:  AIC=-1204.31
## glyhb ~ stab.glu + age + waist
##
##                   Df Sum of Sq     RSS     AIC
## + ratio            1  0.002746 0.24014 -1204.4
## <none>                         0.24289 -1204.3
## + stab.glu:age     1  0.002150 0.24074 -1203.9
## + time.ppn         1  0.001329 0.24156 -1203.3
## + chol             1  0.001173 0.24172 -1203.2
## + hdl              1  0.000947 0.24194 -1203.0
## + weight           1  0.000556 0.24233 -1202.7
## + bp.1s            1  0.000492 0.24240 -1202.7
## + height           1  0.000432 0.24246 -1202.6
## + age:waist        1  0.000080 0.24281 -1202.4
## + stab.glu:waist   1  0.000070 0.24282 -1202.4
## + bp.1d            1  0.000046 0.24284 -1202.3
## + gender           1  0.000037 0.24285 -1202.3
## + hip              1  0.000009 0.24288 -1202.3
## + location         1  0.000007 0.24288 -1202.3
## + frame            2  0.002491 0.24040 -1202.2
## - waist            1  0.014522 0.25741 -1195.7
## - age              1  0.021285 0.26417 -1190.9
## - stab.glu         1  0.160773 0.40366 -1113.3
##
## Step:  AIC=-1204.39
## glyhb ~ stab.glu + age + waist + ratio
##
##                   Df Sum of Sq     RSS     AIC
## + stab.glu:ratio   1  0.003551 0.23659 -1205.1
## <none>                         0.24014 -1204.4
## - ratio            1  0.002746 0.24289 -1204.3
## + stab.glu:age     1  0.002386 0.23776 -1204.2
## + time.ppn         1  0.001658 0.23849 -1203.7
## + frame            2  0.003514 0.23663 -1203.1
## + ratio:age        1  0.000902 0.23924 -1203.1
## + weight           1  0.000726 0.23942 -1202.9
## + bp.1s            1  0.000666 0.23948 -1202.9
## + height           1  0.000443 0.23970 -1202.7
## + chol             1  0.000173 0.23997 -1202.5
```

```
## + stab.glu:waist  1  0.000149 0.23999 -1202.5
## + hdl             1  0.000104 0.24004 -1202.5
## + age:waist       1  0.000079 0.24006 -1202.5
## + hip             1  0.000052 0.24009 -1202.4
## + bp.1d           1  0.000038 0.24011 -1202.4
## + location        1  0.000005 0.24014 -1202.4
## + ratio:waist     1  0.000001 0.24014 -1202.4
## + gender          1  0.000000 0.24014 -1202.4
## - waist           1  0.010246 0.25039 -1198.7
## - age             1  0.019240 0.25938 -1192.3
## - stab.glu        1  0.142762 0.38291 -1121.0
##
## Step:  AIC=-1205.12
## glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio
##
##                  Df Sum of Sq     RSS     AIC
## + ratio:age       1 0.0026083 0.23398 -1205.1
## <none>                        0.23659 -1205.1
## + time.ppn        1 0.0017079 0.23489 -1204.4
## - stab.glu:ratio  1 0.0035506 0.24014 -1204.4
## + height          1 0.0009334 0.23566 -1203.8
## + frame           2 0.0033195 0.23327 -1203.7
## + bp.1s           1 0.0007466 0.23585 -1203.7
## + stab.glu:age    1 0.0006609 0.23593 -1203.6
## + stab.glu:waist  1 0.0005916 0.23600 -1203.6
## + weight          1 0.0003696 0.23622 -1203.4
## + hdl             1 0.0003115 0.23628 -1203.4
## + age:waist       1 0.0002539 0.23634 -1203.3
## + chol            1 0.0001931 0.23640 -1203.3
## + ratio:waist     1 0.0001590 0.23643 -1203.2
## + bp.1d           1 0.0000799 0.23651 -1203.2
## + gender          1 0.0000593 0.23653 -1203.2
## + hip             1 0.0000130 0.23658 -1203.1
## + location        1 0.0000113 0.23658 -1203.1
## - waist           1 0.0086327 0.24522 -1200.6
## - age             1 0.0184053 0.25500 -1193.4
##
## Step:  AIC=-1205.14
## glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio
##
##                  Df Sum of Sq     RSS     AIC
## <none>                        0.23398 -1205.1
## - age:ratio       1 0.0026083 0.23659 -1205.1
## + time.ppn        1 0.0019693 0.23201 -1204.7
## + stab.glu:age    1 0.0011229 0.23286 -1204.0
## + height          1 0.0010043 0.23298 -1203.9
## + bp.1s           1 0.0005834 0.23340 -1203.6
## + hdl             1 0.0004351 0.23355 -1203.5
## + stab.glu:waist  1 0.0004169 0.23357 -1203.5
## + chol            1 0.0002772 0.23371 -1203.4
## + weight          1 0.0002713 0.23371 -1203.4
## + frame           2 0.0027231 0.23126 -1203.3
## + gender          1 0.0001272 0.23386 -1203.2
## + bp.1d           1 0.0000804 0.23390 -1203.2
```

```
## + age:waist       1 0.0000781 0.23391 -1203.2
## + location        1 0.0000033 0.23398 -1203.2
## + hip             1 0.0000016 0.23398 -1203.1
## + ratio:waist     1 0.0000012 0.23398 -1203.1
## - stab.glu:ratio  1 0.0052565 0.23924 -1203.1
## - waist           1 0.0087815 0.24277 -1200.4
```

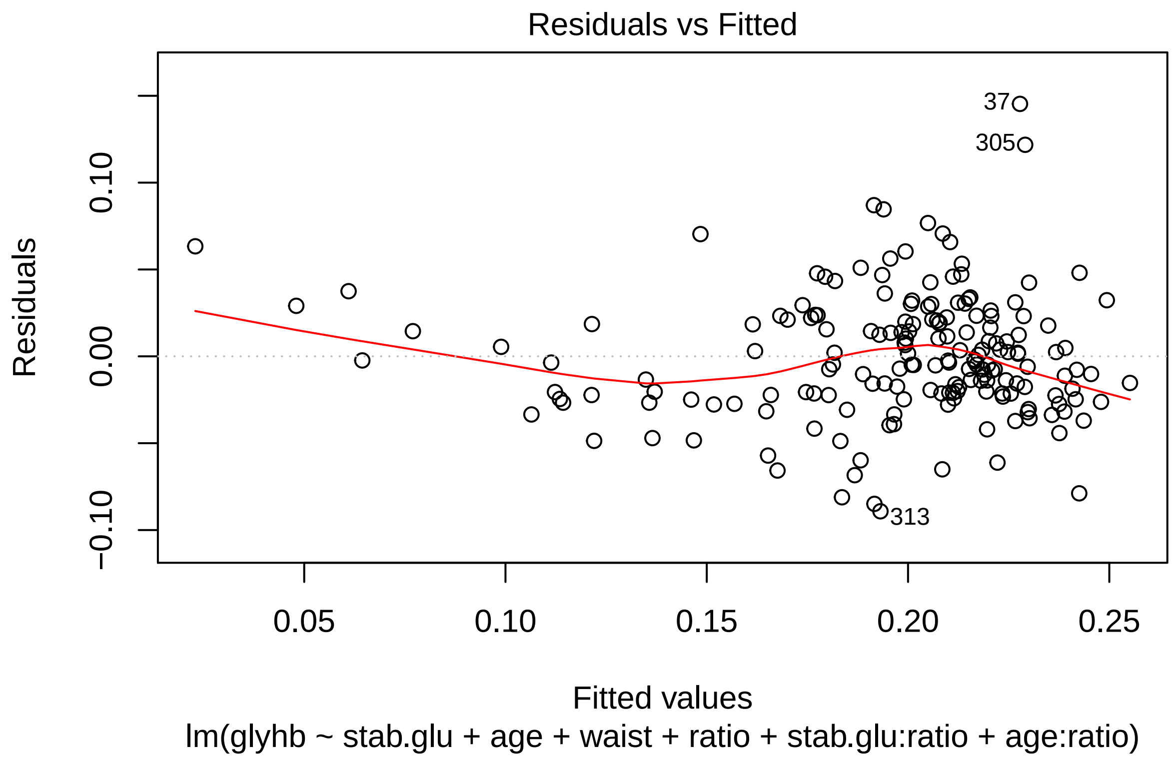
```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
##     age:ratio, data = diabetes.c)
##
## Coefficients:
##    (Intercept)        stab.glu             age           waist
##      3.527e-01      -9.522e-04       7.247e-05      -1.305e-03
##          ratio  stab.glu:ratio       age:ratio
##     -2.158e-03       7.507e-05      -1.724e-04
```

```r
fs2 <- lm(glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio,
    data = diabetes.c)
```

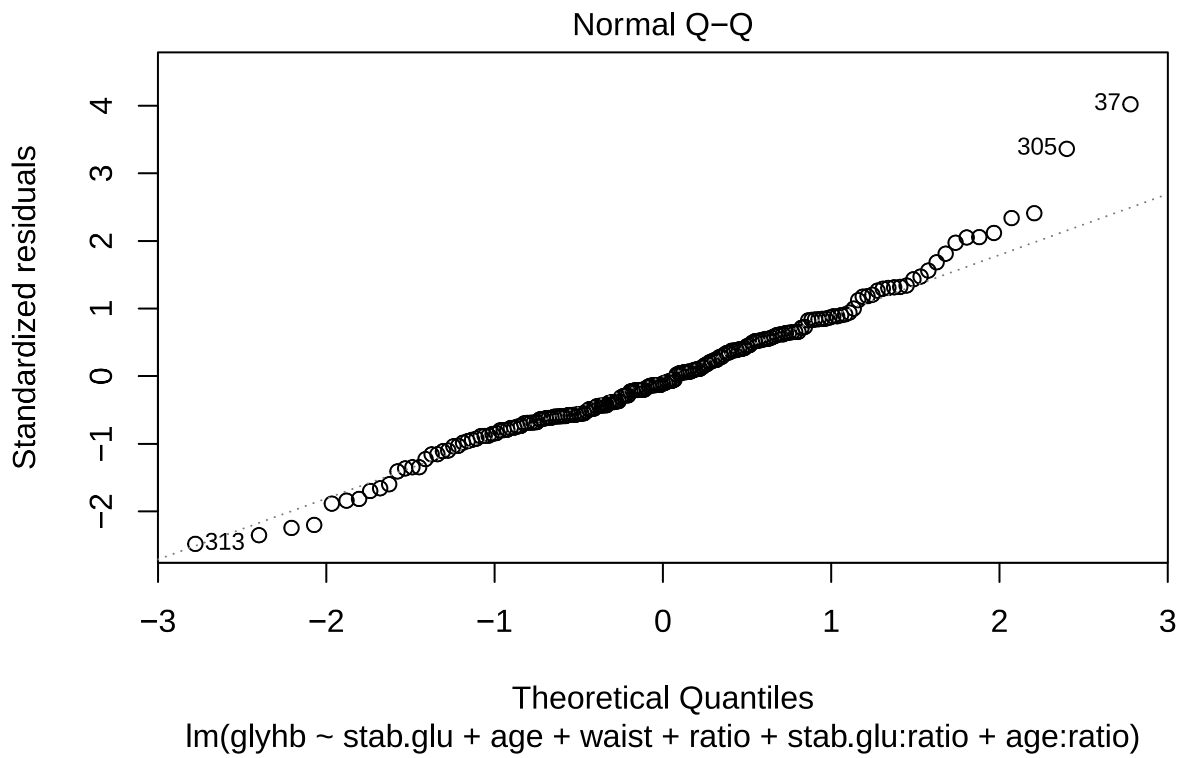
The model is $glyhb\ \beta_0 + \beta_1 stab.glu + \beta_2 age + \beta_3 waist + \beta_4 ration + \beta_5 age * ration + \beta_6 age * ratio$ now. The AIC is -1205.14, a little improvement on model fs1.

(c)

```r
plot(fs2, which = 1)
```




```
plot(fs2, which = 2)
```



The residual vs. fitted value plot still has a pattern of square. The residual Q-Q plot is a little heavy tailed. It seems not to be adequate.

(d)

```
stepAIC(lm(glyhb ~ 1, data = diabetes.c), scope = list(upper = lm(glyhb ~
    .^2, data = diabetes.c)), direction = "forward")
```

```
## Start:  AIC=-1072.47
## glyhb ~ 1
##
##            Df Sum of Sq     RSS     AIC
## + stab.glu  1  0.229457 0.28641 -1178.2
## + age       1  0.080171 0.43569 -1101.4
## + ratio     1  0.062778 0.45309 -1094.2
## + waist     1  0.055768 0.46010 -1091.4
## + hdl       1  0.026343 0.48952 -1080.1
## + bp.1s     1  0.026201 0.48966 -1080.0
## + hip       1  0.022197 0.49367 -1078.5
## + chol      1  0.020540 0.49533 -1077.9
## + weight    1  0.019826 0.49604 -1077.6
## + frame     2  0.024818 0.49105 -1077.5
## <none>                 0.51586 -1072.5
## + bp.1d     1  0.001406 0.51446 -1071.0
## + gender    1  0.001168 0.51470 -1070.9
```

```
## + height    1  0.000406 0.51546 -1070.6
## + time.ppn  1  0.000253 0.51561 -1070.6
## + location  1  0.000223 0.51564 -1070.5
##
## Step:  AIC=-1178.15
## glyhb ~ stab.glu
##
##            Df Sum of Sq     RSS     AIC
## + age       1 0.0289964 0.25741 -1195.7
## + waist     1 0.0222336 0.26417 -1190.9
## + ratio     1 0.0122372 0.27417 -1184.1
## + hip       1 0.0101724 0.27624 -1182.8
## + bp.1s     1 0.0100112 0.27640 -1182.7
## + chol      1 0.0074466 0.27896 -1181.0
## + weight    1 0.0059545 0.28045 -1180.0
## + hdl       1 0.0031506 0.28326 -1178.2
## <none>                  0.28641 -1178.2
## + time.ppn  1 0.0012180 0.28519 -1176.9
## + bp.1d     1 0.0011321 0.28528 -1176.9
## + frame     2 0.0035822 0.28282 -1176.5
## + location  1 0.0001580 0.28625 -1176.2
## + height    1 0.0000079 0.28640 -1176.2
## + gender    1 0.0000047 0.28640 -1176.2
##
## Step:  AIC=-1195.68
## glyhb ~ stab.glu + age
##
##                Df Sum of Sq     RSS     AIC
## + waist         1 0.0145221 0.24289 -1204.3
## + hip           1 0.0104019 0.24701 -1201.2
## + weight        1 0.0083758 0.24904 -1199.7
## + ratio         1 0.0070223 0.25039 -1198.7
## + hdl           1 0.0039458 0.25346 -1196.5
## + stab.glu:age  1 0.0028146 0.25460 -1195.7
## <none>                      0.25741 -1195.7
## + chol          1 0.0017263 0.25568 -1194.9
## + bp.1s         1 0.0008704 0.25654 -1194.3
## + time.ppn      1 0.0007973 0.25661 -1194.2
## + bp.1d         1 0.0005627 0.25685 -1194.1
## + height        1 0.0005250 0.25689 -1194.1
## + gender        1 0.0000412 0.25737 -1193.7
## + location      1 0.0000122 0.25740 -1193.7
## + frame         2 0.0011941 0.25622 -1192.5
##
## Step:  AIC=-1204.31
## glyhb ~ stab.glu + age + waist
##
##                  Df  Sum of Sq     RSS     AIC
## + ratio           1 0.00274582 0.24014 -1204.4
## <none>                         0.24289 -1204.3
## + stab.glu:age    1 0.00214962 0.24074 -1203.9
## + time.ppn        1 0.00132861 0.24156 -1203.3
## + chol            1 0.00117284 0.24172 -1203.2
## + hdl             1 0.00094731 0.24194 -1203.0
```

```
## + weight          1 0.00055551 0.24233 -1202.7
## + bp.1s           1 0.00049179 0.24240 -1202.7
## + height          1 0.00043161 0.24246 -1202.6
## + age:waist       1 0.00008002 0.24281 -1202.4
## + stab.glu:waist  1 0.00007014 0.24282 -1202.4
## + bp.1d           1 0.00004616 0.24284 -1202.3
## + gender          1 0.00003677 0.24285 -1202.3
## + hip             1 0.00000922 0.24288 -1202.3
## + location        1 0.00000748 0.24288 -1202.3
## + frame           2 0.00249149 0.24040 -1202.2
##
## Step:  AIC=-1204.39
## glyhb ~ stab.glu + age + waist + ratio
##
##                  Df Sum of Sq     RSS     AIC
## + stab.glu:ratio  1 0.0035506 0.23659 -1205.1
## <none>                        0.24014 -1204.4
## + stab.glu:age    1 0.0023863 0.23776 -1204.2
## + time.ppn        1 0.0016578 0.23849 -1203.7
## + frame           2 0.0035143 0.23663 -1203.1
## + ratio:age       1 0.0009024 0.23924 -1203.1
## + weight          1 0.0007262 0.23942 -1202.9
## + bp.1s           1 0.0006657 0.23948 -1202.9
## + height          1 0.0004432 0.23970 -1202.7
## + chol            1 0.0001733 0.23997 -1202.5
## + stab.glu:waist  1 0.0001486 0.24000 -1202.5
## + hdl             1 0.0001042 0.24004 -1202.5
## + age:waist       1 0.0000793 0.24006 -1202.5
## + hip             1 0.0000519 0.24009 -1202.4
## + bp.1d           1 0.0000376 0.24011 -1202.4
## + location        1 0.0000050 0.24014 -1202.4
## + ratio:waist     1 0.0000009 0.24014 -1202.4
## + gender          1 0.0000000 0.24014 -1202.4
##
## Step:  AIC=-1205.12
## glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio
##
##                  Df Sum of Sq     RSS     AIC
## + ratio:age       1 0.0026083 0.23398 -1205.1
## <none>                        0.23659 -1205.1
## + time.ppn        1 0.0017079 0.23489 -1204.4
## + height          1 0.0009334 0.23566 -1203.8
## + frame           2 0.0033195 0.23327 -1203.7
## + bp.1s           1 0.0007466 0.23585 -1203.7
## + stab.glu:age    1 0.0006609 0.23593 -1203.6
## + stab.glu:waist  1 0.0005916 0.23600 -1203.6
## + weight          1 0.0003696 0.23622 -1203.4
## + hdl             1 0.0003115 0.23628 -1203.4
## + age:waist       1 0.0002539 0.23634 -1203.3
## + chol            1 0.0001931 0.23640 -1203.3
## + ratio:waist     1 0.0001590 0.23643 -1203.2
## + bp.1d           1 0.0000799 0.23651 -1203.2
## + gender          1 0.0000593 0.23653 -1203.2
## + hip             1 0.0000130 0.23658 -1203.1
```

```
## + location        1 0.0000113 0.23658 -1203.1
##
## Step:  AIC=-1205.14
## glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio
##
##                   Df  Sum of Sq     RSS     AIC
## <none>                          0.23398 -1205.1
## + time.ppn        1 0.00196928 0.23201 -1204.7
## + stab.glu:age    1 0.00112288 0.23286 -1204.0
## + height          1 0.00100427 0.23298 -1203.9
## + bp.1s           1 0.00058338 0.23340 -1203.6
## + hdl             1 0.00043510 0.23355 -1203.5
## + stab.glu:waist  1 0.00041688 0.23357 -1203.5
## + chol            1 0.00027720 0.23371 -1203.4
## + weight          1 0.00027134 0.23371 -1203.4
## + frame           2 0.00272313 0.23126 -1203.3
## + gender          1 0.00012720 0.23386 -1203.2
## + bp.1d           1 0.00008037 0.23390 -1203.2
## + age:waist       1 0.00007809 0.23391 -1203.2
## + location        1 0.00000326 0.23398 -1203.2
## + hip             1 0.00000155 0.23398 -1203.1
## + ratio:waist     1 0.00000120 0.23398 -1203.1
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
##     age:ratio, data = diabetes.c)
##
## Coefficients:
##    (Intercept)        stab.glu             age           waist
##      3.527e-01      -9.522e-04       7.247e-05      -1.305e-03
##          ratio  stab.glu:ratio       age:ratio
##     -2.158e-03       7.507e-05      -1.724e-04
```

The model is the same as forward stepwise procedure.

Problem 5

(a)

```r
fit3 <- lm(glyhb ~ (stab.glu + age + waist + ratio)^2, data = diabetes.c)
summary(fit3)
```

```
##
## Call:
## lm(formula = glyhb ~ (stab.glu + age + waist + ratio)^2, data = diabetes.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.087028 -0.021663 -0.004788  0.022456  0.145784
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.376e-01  6.758e-02   4.995 1.44e-06 ***
## stab.glu       -7.941e-04  4.458e-04  -1.781   0.0766 .
## age            -6.979e-05  1.172e-03  -0.060   0.9526
## waist          -4.686e-04  1.920e-03  -0.244   0.8075
## ratio          -5.113e-03  1.463e-02  -0.350   0.7271
## stab.glu:age    5.255e-06  4.560e-06   1.152   0.2507
## stab.glu:waist -9.323e-06  1.076e-05  -0.866   0.3875
## stab.glu:ratio  6.139e-05  4.133e-05   1.486   0.1392
## age:waist      -8.184e-06  2.808e-05  -0.291   0.7710
## age:ratio      -1.908e-04  1.292e-04  -1.477   0.1415
## waist:ratio     1.314e-04  3.751e-04   0.350   0.7265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0367 on 172 degrees of freedom
## Multiple R-squared:  0.5509, Adjusted R-squared:  0.5248
## F-statistic:  21.1 on 10 and 172 DF,  p-value: < 2.2e-16
```

```
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                 Df   Sum Sq  Mean Sq  F value    Pr(>F)
## stab.glu         1 0.229457 0.229457 170.3514 < 2.2e-16 ***
## age              1 0.028996 0.028996  21.5273 6.886e-06 ***
## waist            1 0.014522 0.014522  10.7814  0.001242 **
## ratio            1 0.002746 0.002746   2.0385  0.155171
## stab.glu:age     1 0.002386 0.002386   1.7716  0.184942
## stab.glu:waist   1 0.000893 0.000893   0.6633  0.416538
## stab.glu:ratio   1 0.002000 0.002000   1.4847  0.224702
## age:waist        1 0.000245 0.000245   0.1822  0.670030
## age:ratio        1 0.002775 0.002775   2.0604  0.152989
## waist:ratio      1 0.000165 0.000165   0.1228  0.726494
## Residuals      172 0.231678 0.001347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are 11 regression coefficients. MSE is 0.001347

For model fs1:

```
anova(fs1)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##            Df   Sum Sq  Mean Sq  F value    Pr(>F)
## stab.glu    1 0.229457 0.229457 170.0791 < 2.2e-16 ***
## age         1 0.028996 0.028996  21.4929  6.85e-06 ***
## waist       1 0.014522 0.014522  10.7641  0.001245 **
## ratio       1 0.002746 0.002746   2.0353  0.155439
```

```
## Residuals 178 0.240143 0.001349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(fs1_press_p = sum((residuals(fs1))^2/(1 - influence(fs1)$hat)^2))
```

```
## [1] 0.2535404
```

$SSE_p = 0.240143$

$MSE_p = 0.001349$

$C_p = \frac{0.240143}{0.001347} - (183 - 2 * 5) = 5.279881$

$Press_p = 0.2535404$

For model fs2:

```
anova(fs2)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                 Df   Sum Sq  Mean Sq  F value    Pr(>F)
## stab.glu         1 0.229457 0.229457 172.5946 < 2.2e-16 ***
## age              1 0.028996 0.028996  21.8107 5.951e-06 ***
## waist            1 0.014522 0.014522  10.9233  0.001151 **
## ratio            1 0.002746 0.002746   2.0654  0.152454
## stab.glu:ratio   1 0.003551 0.003551   2.6707  0.103996
## age:ratio        1 0.002608 0.002608   1.9619  0.163067
## Residuals      176 0.233984 0.001329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(fs2_press_p = sum((residuals(fs2))^2/(1 - influence(fs2)$hat)^2))
```

```
## [1] 0.2534834
```

$SSE_p = 0.233984$

$MSE_p = 0.001329$

$C_p = \frac{0.233984}{0.001347} - (183 - 2 * 7) = 4.707498$

$Press_p = 0.2534834$

There is little difference between $Press_p$ and SSE_p, so there is no evidence of overfitting for fs2. And there is a little bias in fs1, since C_p of fs2 is small than that of fs1.

(b)

```
fs1_v = lm(glyhb ~ stab.glu + age + waist + ratio, data = diabetes.v)
fs2_v = lm(glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio,
    data = diabetes.v)
```

```r
fs1$coefficients
```

```
##   (Intercept)      stab.glu           age         waist         ratio 
##  0.3489987020 -0.0005368259 -0.0006411598 -0.0013984694 -0.0028482611
```

```r
fs1_v$coefficients
```

```
##   (Intercept)      stab.glu           age         waist         ratio 
##  0.3287126110 -0.0004436401 -0.0006693833 -0.0008450557 -0.0042812293
```

```r
anova(fs1)
```

```
## Analysis of Variance Table
## 
## Response: glyhb
##            Df   Sum Sq  Mean Sq  F value    Pr(>F)    
## stab.glu    1 0.229457 0.229457 170.0791 < 2.2e-16 ***
## age         1 0.028996 0.028996  21.4929  6.85e-06 ***
## waist       1 0.014522 0.014522  10.7641  0.001245 ** 
## ratio       1 0.002746 0.002746   2.0353  0.155439    
## Residuals 178 0.240143 0.001349                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fs1_v)
```

```
## Analysis of Variance Table
## 
## Response: glyhb
##            Df   Sum Sq  Mean Sq  F value    Pr(>F)    
## stab.glu    1 0.169298 0.169298 128.2353 < 2.2e-16 ***
## age         1 0.019939 0.019939  15.1027 0.0001435 ***
## waist       1 0.008000 0.008000   6.0599 0.0147812 *  
## ratio       1 0.011171 0.011171   8.4619 0.0040892 ** 
## Residuals 178 0.234999 0.001320                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients of fs1: values change a little bit, while signs do not change. MSE change from 0.001349 to 0.001320. fs1 is consistent on train and validation data.

```r
fs2$coefficients
```

```
##    (Intercept)       stab.glu            age          waist 
##   3.526665e-01  -9.522192e-04   7.246565e-05  -1.305246e-03 
##          ratio stab.glu:ratio      age:ratio 
##  -2.158350e-03   7.507002e-05  -1.724382e-04
```

```r
fs2_v$coefficients
```

```
##    (Intercept)       stab.glu            age          waist
##   3.122342e-01  -2.435421e-04  -8.409084e-04  -9.389972e-04
##          ratio stab.glu:ratio      age:ratio
##   7.797037e-05  -3.984021e-05   3.365585e-05
```

```r
anova(fs2)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                 Df   Sum Sq  Mean Sq  F value    Pr(>F)
## stab.glu         1 0.229457 0.229457 172.5946 < 2.2e-16 ***
## age              1 0.028996 0.028996  21.8107 5.951e-06 ***
## waist            1 0.014522 0.014522  10.9233  0.001151 **
## ratio            1 0.002746 0.002746   2.0654  0.152454
## stab.glu:ratio   1 0.003551 0.003551   2.6707  0.103996
## age:ratio        1 0.002608 0.002608   1.9619  0.163067
## Residuals      176 0.233984 0.001329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fs2_v)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                 Df   Sum Sq  Mean Sq  F value    Pr(>F)
## stab.glu         1 0.169298 0.169298 128.5411 < 2.2e-16 ***
## age              1 0.019939 0.019939  15.1387 0.0001415 ***
## waist            1 0.008000 0.008000   6.0743 0.0146757 *
## ratio            1 0.011171 0.011171   8.4820 0.0040515 **
## stab.glu:ratio   1 0.003090 0.003090   2.3460 0.1274038
## age:ratio        1 0.000103 0.000103   0.0785 0.7796463
## Residuals      176 0.231805 0.001317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients of fs2: values change a lot, some of them even change their signs. MSE change from 0.001329 to 0.001317. fs2 is not consistent on train and validation data.

$MSPE$:

```r
(fs1_mspe <- mean((predict(fs1, diabetes.v[-5]) - diabetes.v[5])^2))
```

```
## [1] 0.001329283
```

```r
(fs2_mspe <- mean((predict(fs2, diabetes.v[-5]) - diabetes.v[5])^2))
```

```
## [1] 0.00152642
```

For model fs1, $Press_p/n$ is $0.2535404/183 = 0.001385467$ and SSE_p/n is $0.240143/183 = 0.001312257$. $MSPE_v = 0.001329283$ is a little bit higher than SSE_p/n.

For model fs2, $Press_p/n$ is $0.2534834/183 = 0.001385155$ and SSE_p/n is $0.233984/183 = 0.001278601$. But $MSPE_v = 0.00152642$ is much higher than them.

The first model has a smaller $MSPE_v$.

(c)

I will use fs2 on internal validation and fs1 on external validation.

So fs1 will be selected as the final model, since external validation is more reasonable.

```
fs1_f <- lm(glyhb ~ stab.glu + age + waist + ratio, data = diabetes.s)
summary(fs1_f)
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = diabetes.s)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.152555 -0.020528 -0.000382  0.019560  0.148412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.380e-01  1.306e-02  25.881  < 2e-16 ***
## stab.glu    -4.922e-04  3.838e-05 -12.825  < 2e-16 ***
## age         -6.561e-04  1.229e-04  -5.338 1.67e-07 ***
## waist       -1.080e-03  3.516e-04  -3.071  0.00229 **
## ratio       -3.661e-03  1.181e-03  -3.100  0.00209 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03643 on 361 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.495
## F-statistic: 90.45 on 4 and 361 DF,  p-value: < 2.2e-16
```

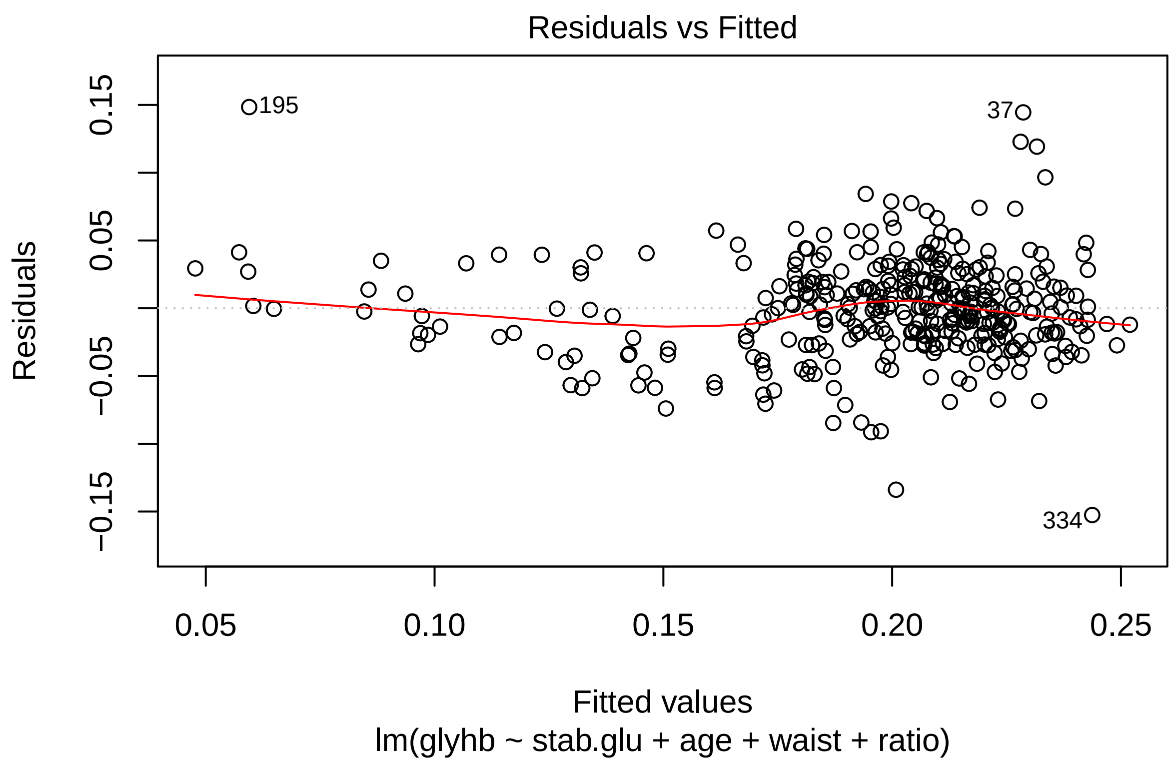
```
anova(fs1_f)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## stab.glu    1 0.39753 0.39753 299.5043 < 2.2e-16 ***
## age         1 0.04867 0.04867  36.6682 3.515e-09 ***
## waist       1 0.02125 0.02125  16.0081 7.655e-05 ***
## ratio       1 0.01276 0.01276   9.6103  0.002087 **
## Residuals 361 0.47915 0.00133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

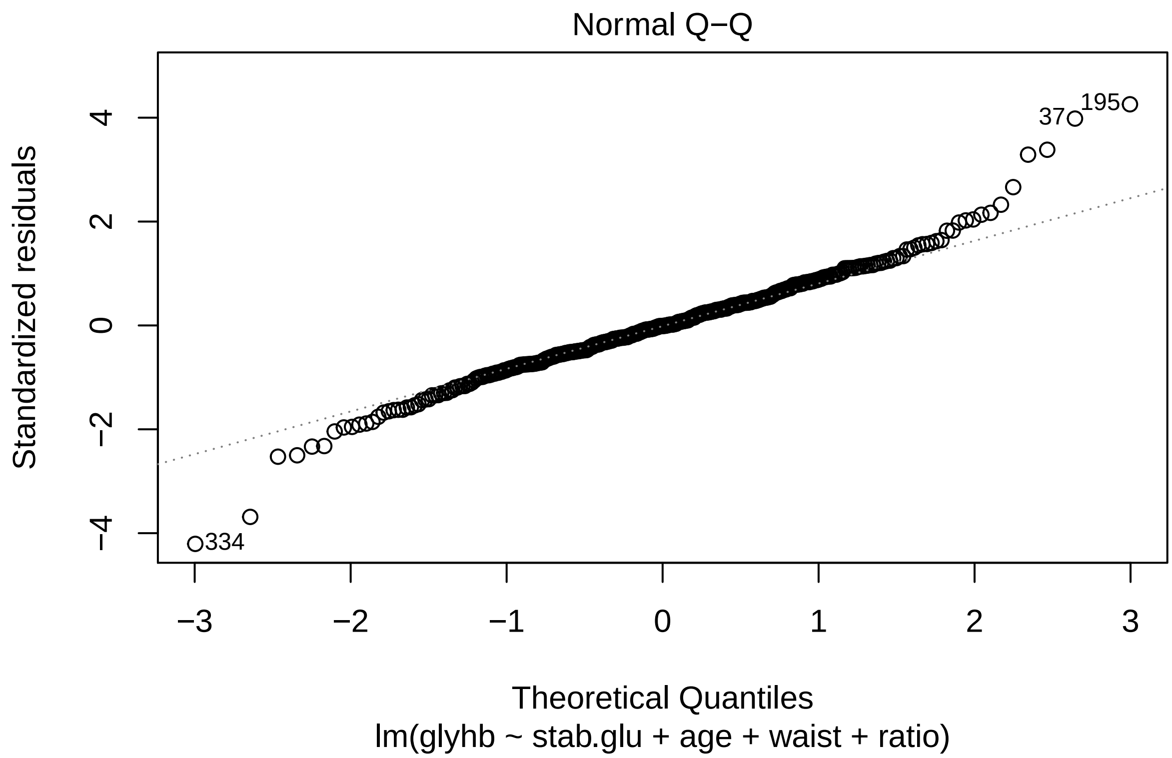
Problem 6

(a)

```r
plot(fs1_f, which = 1)
```




```r
plot(fs1_f, which = 2)
```



The residual plot is good, and qqplot shows a little tendancy of heavy tailed.

(b)

```
res_del = residuals(fs1_f)/(1 - influence(fs1_f)$hat)
deleted_res_del = studres(fs1_f)
alpha = 0.1
n = nrow(diabetes.s)
p = 5
bon_thre = qt(1 - alpha/(2 * n), n - p - 1)
names(which(abs(deleted_res_del) > bon_thre))
```

```
## [1] "37"  "195" "334" "363"
```

So the 37th, 195th, 334th, 363th observations seem to be outliers.

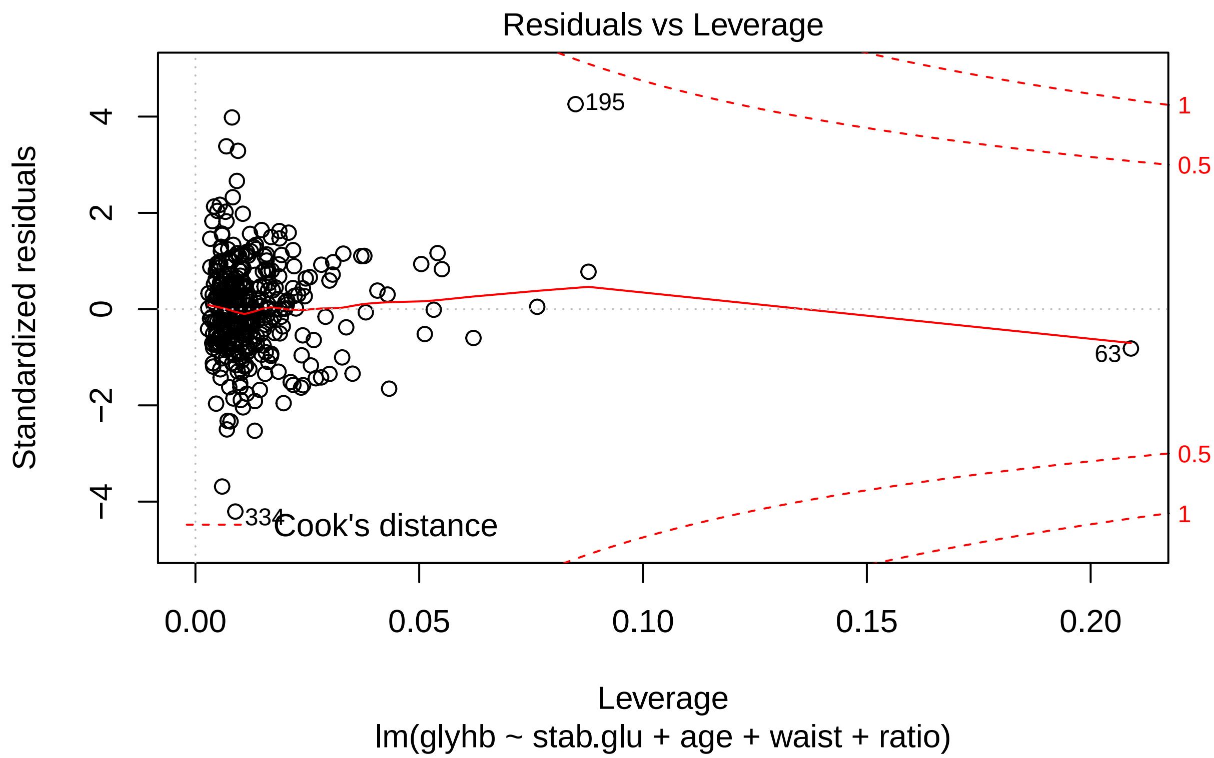
(c)

```
hh = influence(fs1_f)$hat
hh_mean = mean(hh)
which(hh > 2 * p/n)
```

```
##  23  33  47  56  58  61  63 100 134 148 151 156 161 174 177 195 257
##  21  30  42  50  52  54  56  89 118 132 135 139 144 156 159 176 233
## 295 315 329 359 365 382 388 398 399 401
## 268 288 299 326 332 348 354 362 363 365
```

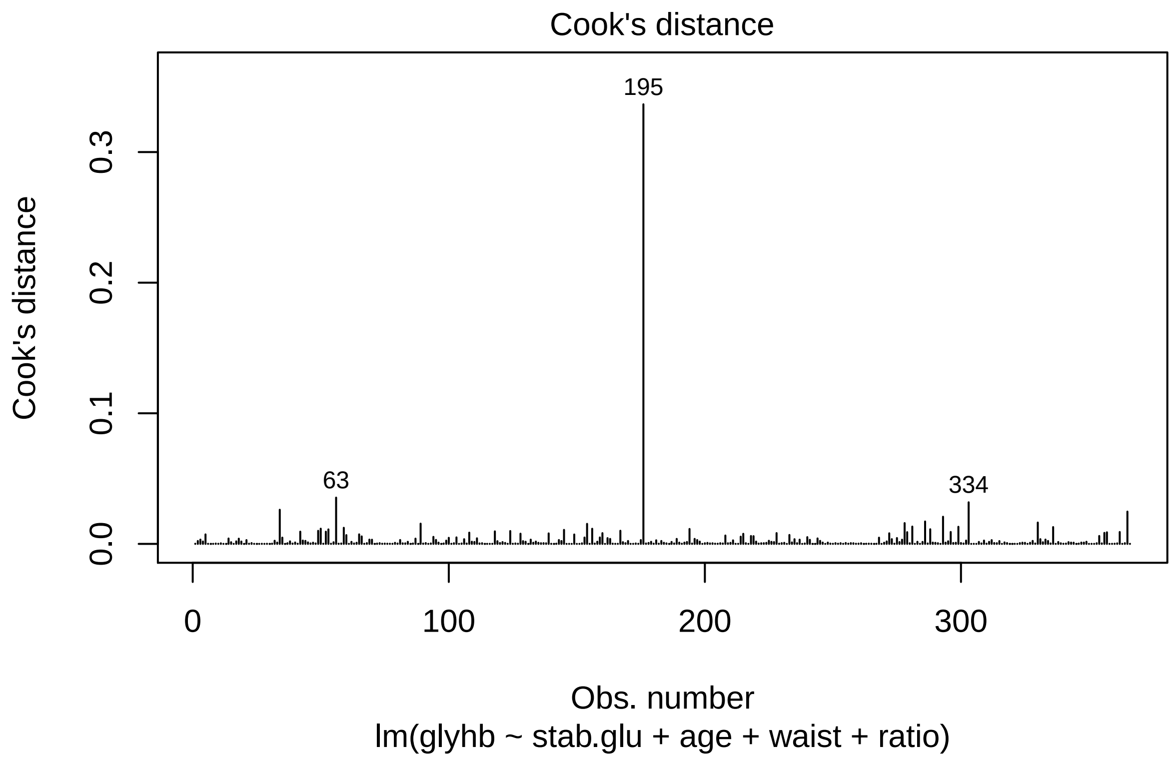
These are identified as outlying X observations.

```r
plot(fs1_f, which = 5)
```



(d)

```r
plot(fs1_f, which = 4)
```



Yes, observation 63, 195 and 334 are influential.

(e)

```
indices <- which(row.names(diabetes.s) == 195)
with_influential <- fs1_f$fitted.values[-indices]
without_influential <- lm(glyhb ~ stab.glu + age + waist + ratio, data = diabetes.s[-indices,
    ])$fitted.values
```

Now I have the values with and without high influential values. Now calculate the average percent difference:

```
mean(abs((with_influential - without_influential)/with_influential))
```

```
## [1] 0.01075285
```

The average absolute percent difference is 1.075285%. So these influential observation indeed makes an influence on the model.