Homework 7

Huang Fang ID:913439658

November 21, 2015

Question 2

```
options(width = 60)
diabetes = read.table("~/academic/Sta206/diabetes.txt", header = TRUE)
```

(a)

```
summary(diabetes$frame)
```

```
##         large medium  small
##     12    103    184    104
```

```
diabetes$frame[diabetes$frame == ""] = NA
diabetes$frame = factor(diabetes$frame)
levels(diabetes$frame)
```

```
## [1] "large"  "medium" "small"
```

(b)

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

(c)

```
data_class = sapply(data, class)
data_class
```

```
##      chol  stab.glu       hdl     ratio     glyhb  location
## "integer" "integer" "integer" "numeric" "numeric"  "factor"
##       age    gender    height    weight     frame     bp.1s
## "integer"  "factor" "integer" "integer"  "factor" "integer"
##     bp.1d     waist       hip  time.ppn
## "integer" "integer" "integer" "integer"
```

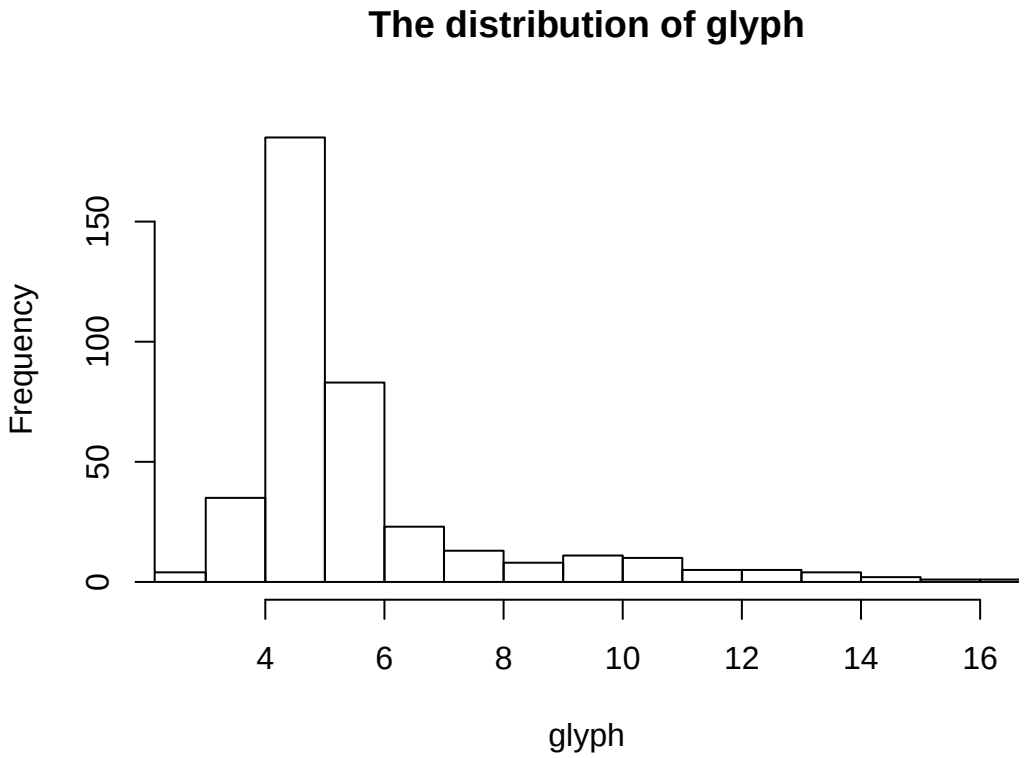
```
#Quantitative variables
names(data_class[data_class != "factor"])
```

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

```r
#Qulitative variables
names(data_class[data_class == "factor"])
```

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

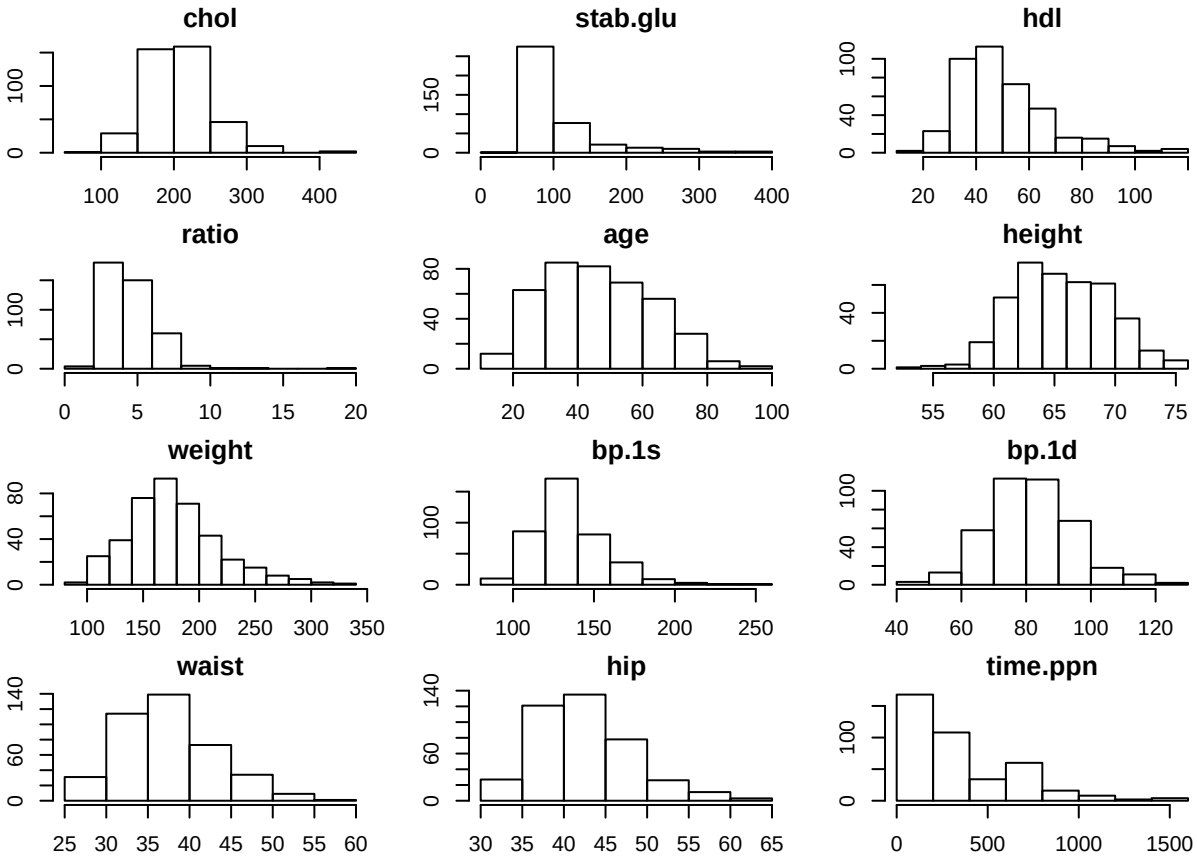
```r
par(mfrow = c(1,1), mar = c(5,5,5,5), cex.lab = 1)
hist(data$glyhb, main = "The distribution of glyph", xlim = range(na.omit(data$glyhb)), xlab = "glyph")
```



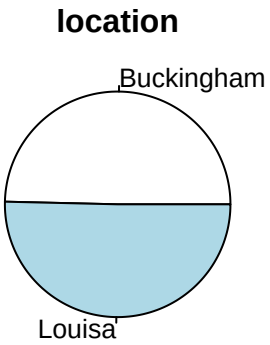

```r
summary(data$glyhb)
```

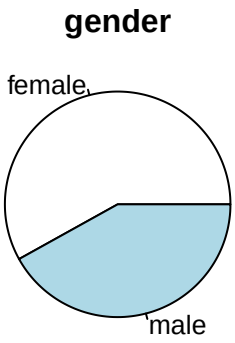
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.68    4.38    4.84    5.59    5.60   16.11      13
```

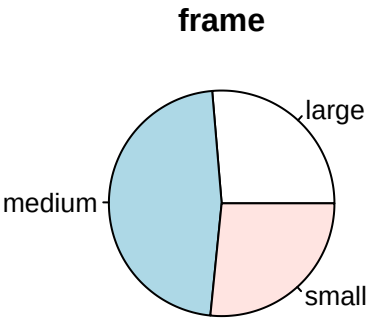
```r
#The distribution of "glyhb" is right-skewed, and mainly distributed on [4, 6]
quant_var = names(data_class[data_class != "factor"])
rest_quant_var = quant_var[quant_var != "glyhb"]
par(mfrow = c(4,3), mar = c(2,2,2,2))
invisible(
  sapply(rest_quant_var, function(x) hist(data[[x]], main = x))
)
```




```
qual_var = names(data_class[data_class == "factor"])
par(mfrow = c(2,2))
invisible(
  sapply(qual_var, function(x) pie(table(data[[x]]), main = x))
)
```

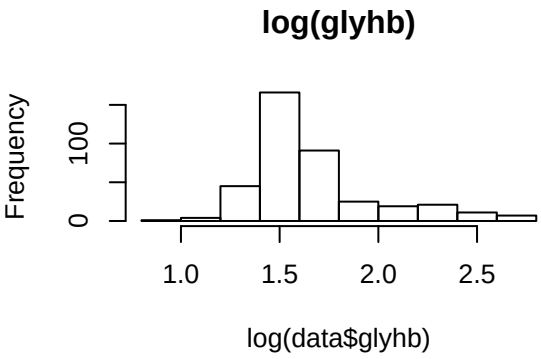


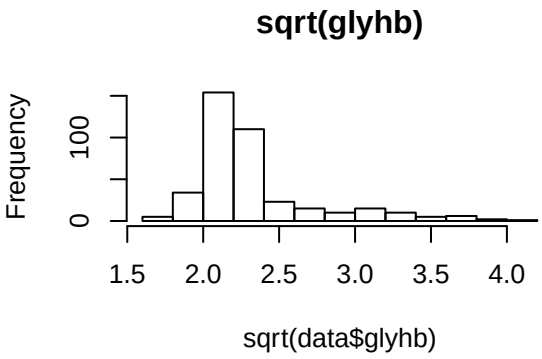


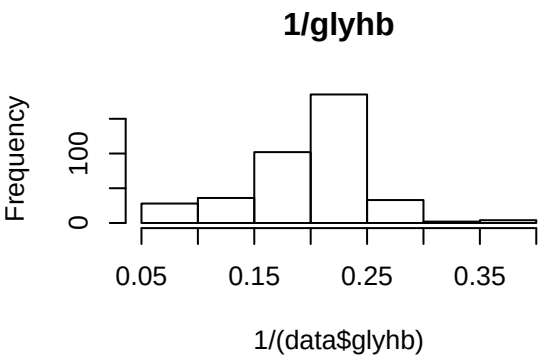


(d)

```
par(mfrow = c(2,2))
hist(log(data$glyhb), main = "log(glyhb)")
hist(sqrt(data$glyhb), main = "sqrt(glyhb)")
hist(1/(data$glyhb), main = "1/glyhb")
```







1/glyhb appears to be most Normal like among the three.

(e)

```
data$glyhb = 1/data$glyhb
```

(f)

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

```
## [1] FALSE
```

```
dim(data.s)
```

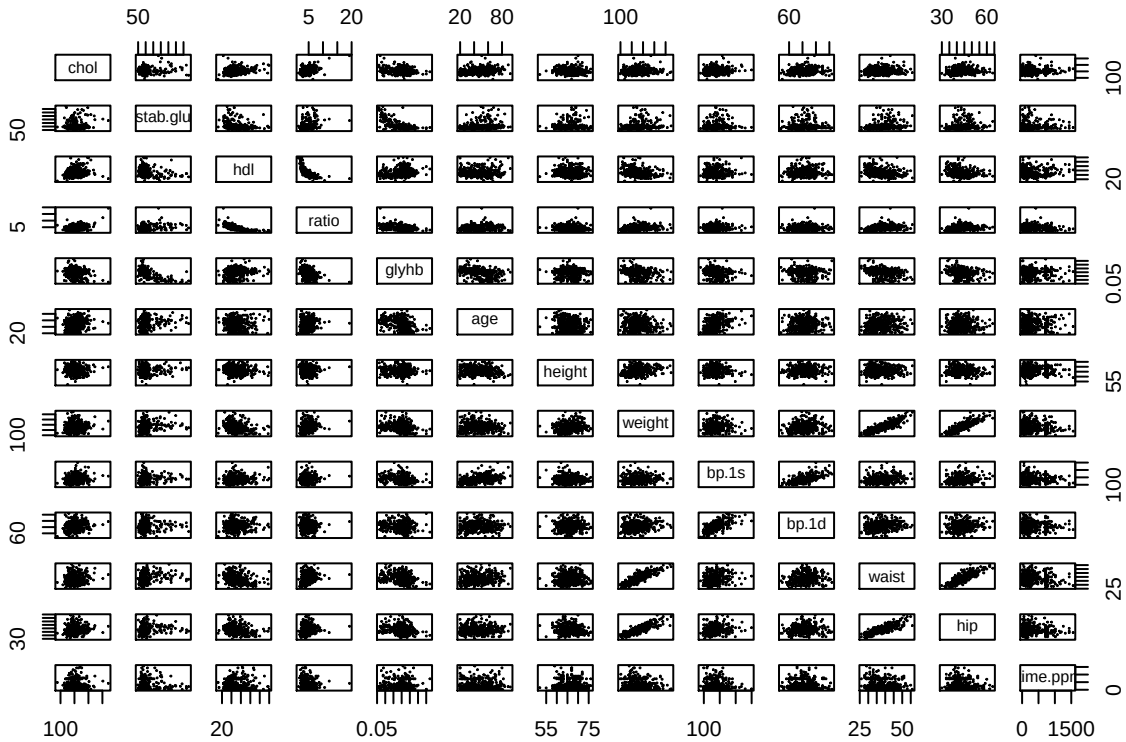
```
## [1] 366  16
```

```
table(data.s$frame)
```

```
##
##  large medium  small
##     96    172     98
```

(g)

```
par(mfrow = c(1,1), mar = c(0,0,0,0))
pairs_var = paste(quant_var, collapse = "+")
pairs_var = paste0("~", pairs_var)
pairs(~., data = data.s[, quant_var], cex = 0.1)
```



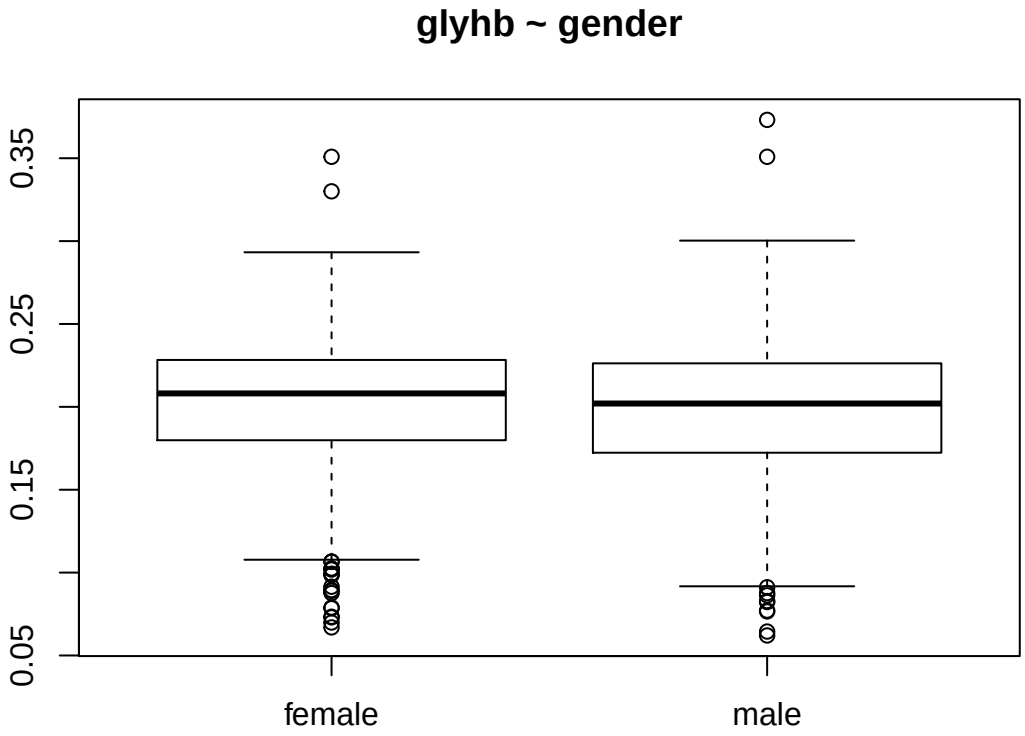
```r
round(cor(data.s[, quant_var]),3)
```

```
##            chol stab.glu    hdl  ratio  glyhb    age height
## chol      1.000    0.165  0.171  0.484 -0.257  0.242 -0.063
## stab.glu  0.165    1.000 -0.180  0.299 -0.644  0.279  0.082
## hdl       0.171   -0.180  1.000 -0.690  0.189  0.000 -0.069
## ratio     0.484    0.299 -0.690  1.000 -0.355  0.172  0.071
## glyhb    -0.257   -0.644  0.189 -0.355  1.000 -0.396 -0.043
## age       0.242    0.279  0.000  0.172 -0.396  1.000 -0.097
## height   -0.063    0.082 -0.069  0.071 -0.043 -0.097  1.000
## weight    0.080    0.189 -0.283  0.279 -0.219 -0.046  0.243
## bp.1s     0.202    0.151  0.030  0.105 -0.230  0.433 -0.044
## bp.1d     0.159    0.026  0.072  0.035 -0.056  0.059  0.043
## waist     0.144    0.234 -0.278  0.315 -0.319  0.170  0.042
## hip       0.099    0.145 -0.222  0.208 -0.213  0.018 -0.117
## time.ppn  0.006   -0.048  0.080 -0.054 -0.036 -0.027 -0.006
##          weight  bp.1s  bp.1d  waist    hip time.ppn
## chol      0.080  0.202  0.159  0.144  0.099    0.006
## stab.glu  0.189  0.151  0.026  0.234  0.145   -0.048
## hdl      -0.283  0.030  0.072 -0.278 -0.222    0.080
## ratio     0.279  0.105  0.035  0.315  0.208   -0.054
## glyhb    -0.219 -0.230 -0.056 -0.319 -0.213   -0.036
## age      -0.046  0.433  0.059  0.170  0.018   -0.027
## height    0.243 -0.044  0.043  0.042 -0.117   -0.006
## weight    1.000  0.096  0.181  0.852  0.830   -0.062
## bp.1s     0.096  1.000  0.620  0.210  0.151   -0.075
## bp.1d     0.181  0.620  1.000  0.179  0.163   -0.064
## waist     0.852  0.210  0.179  1.000  0.832   -0.066
## hip       0.830  0.151  0.163  0.832  1.000   -0.093
## time.ppn -0.062 -0.075 -0.064 -0.066 -0.093    1.000
```

I observe nonlinearity.

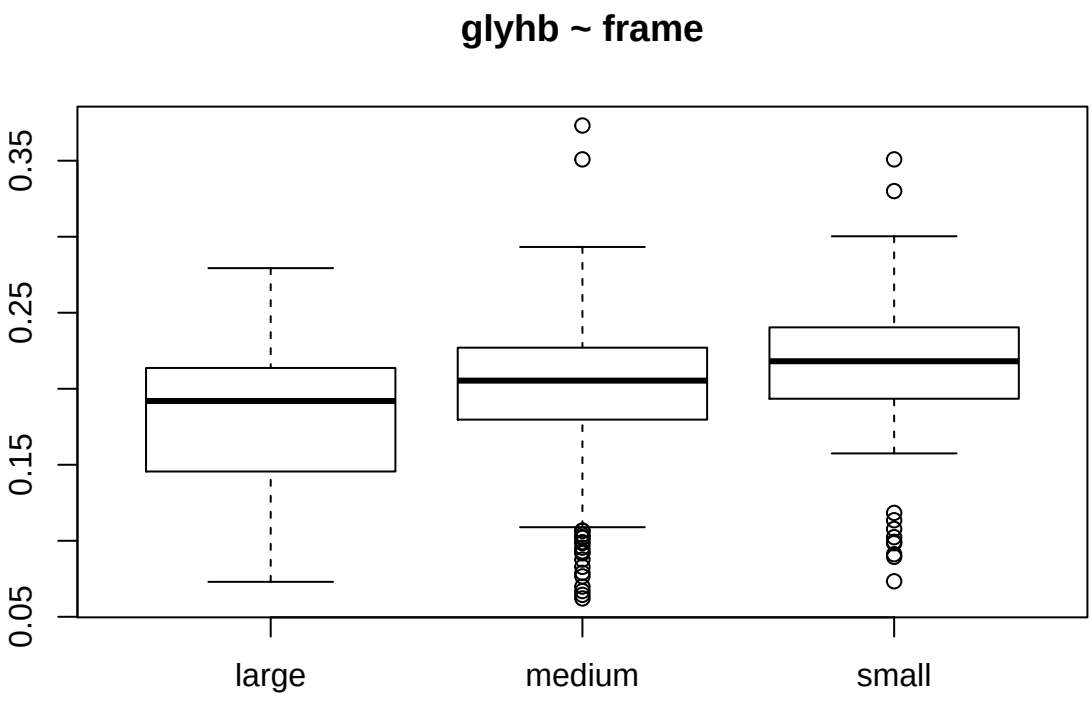
(h)

```r
par(mfrow = c(1,1), mar = c(4,4,4,4))
boxplot(glyhb ~ gender, data = data.s, main = "glyhb ~ gender")
```



There is no obvious relationship between "glyhb" and "gender".

```
boxplot(glyhb ~ frame, data = data.s, main = "glyhb ~ frame")
```



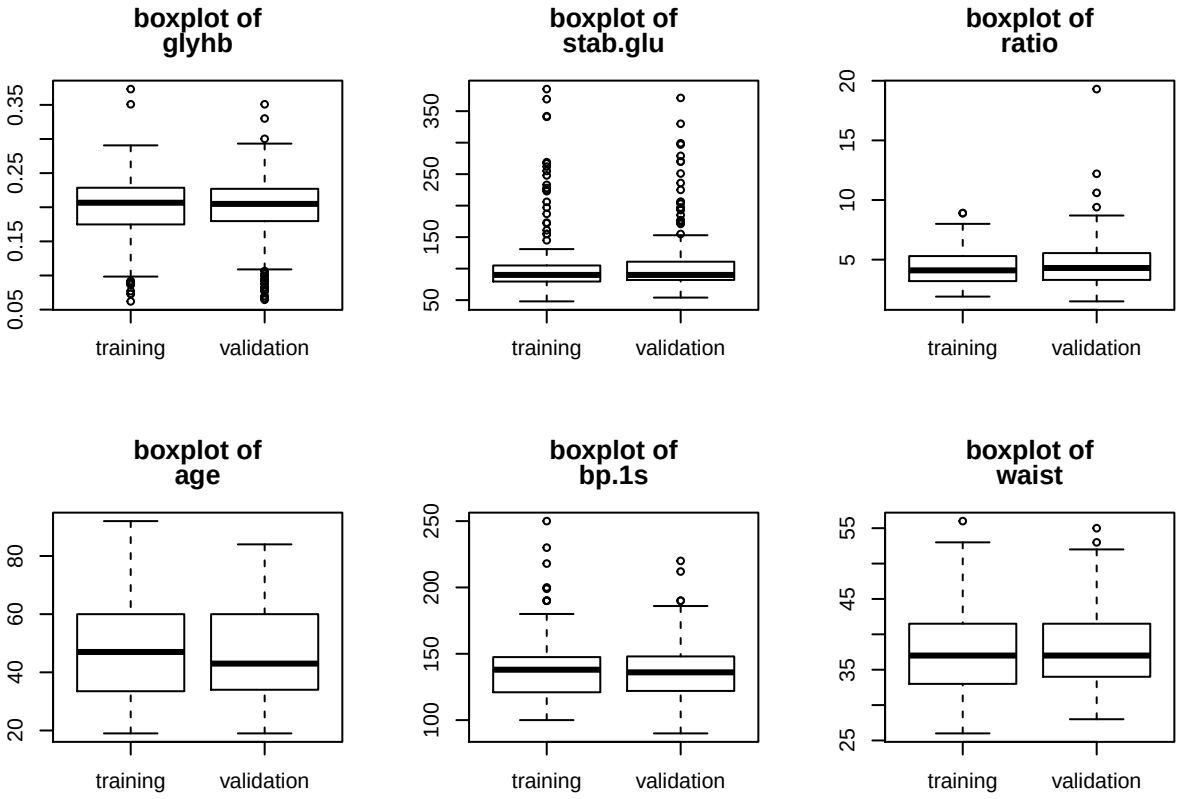
There is a relationship between "glyhb" and "frame".

(i)

```r
set.seed(10)
n.s=nrow(data.s)
index.s=sample(1: n.s, size=366/2, replace=FALSE)
data.c=data.s[index.s,]
data.v=data.s[-index.s,]
```

(j)

```r
par(mfrow = c(2,3), mar = c(4,4,4,1), cex.lab = 0.2)
var_comp = c("glyhb", "stab.glu", "ratio", "age", "bp.1s", "waist")
invisible(
  sapply(var_comp, function(x)
                       boxplot(data.c[,x], data.v[,x], names = c("training","validation"), main = c("boxp
)
```



They approximately have the same distribution.

Question 3

(a)

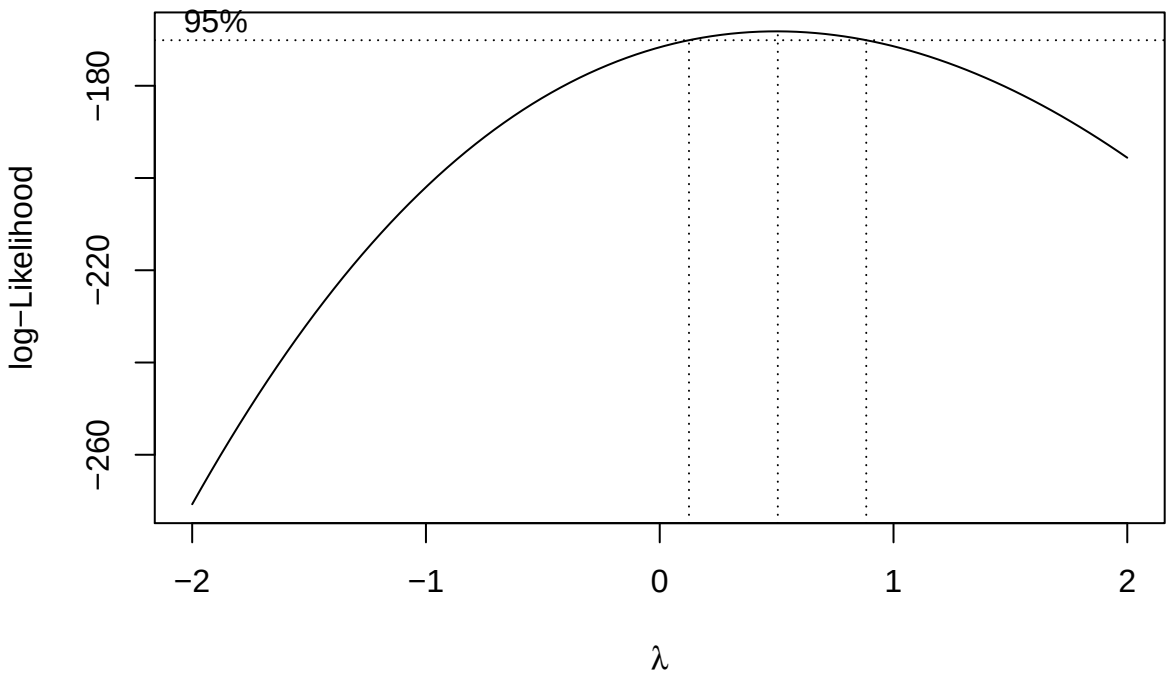
```r
fit.full = lm(glyhb~., data = data.c)
summary(fit.full)
```

```
##
## Call:
## lm(formula = glyhb ~ ., data = data.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
```

There are 16 regression coefficients.
MSE = 0.0372^2 = 0.00138384

```
par(mfrow = c(1,1))
library(MASS)
boxcox(fit.full)
```



Because the value of log_Likelihood is already high when x=1, and we have done transformation before, so we don't need to do a transformation any more.

(b)

```
library("leaps")

sub_set = regsubsets(glyhb~., data = data.c, nbest = 1, nvmax = 16, method = "exhaustive")
sum_sub = summary(sub_set)
sum_sub
```

```
## Subset selection object
## Call: regsubsets.formula(glyhb ~ ., data = data.c, nbest = 1, nvmax = 16,
##     method = "exhaustive")
## 16 Variables  (and intercept)
##                 Forced in Forced out
## chol                FALSE      FALSE
## stab.glu            FALSE      FALSE
## hdl                 FALSE      FALSE
## ratio               FALSE      FALSE
## locationLouisa      FALSE      FALSE
## age                 FALSE      FALSE
## gendermale          FALSE      FALSE
## height              FALSE      FALSE
## weight              FALSE      FALSE
## framemedium         FALSE      FALSE
## framesmall          FALSE      FALSE
## bp.1s               FALSE      FALSE
## bp.1d               FALSE      FALSE
## waist               FALSE      FALSE
## hip                 FALSE      FALSE
## time.ppn            FALSE      FALSE
## 1 subsets of each size up to 16
```

```
## Selection Algorithm: exhaustive
##           chol stab.glu hdl ratio locationLouisa age
## 1  ( 1 )  " "  "*"      " " " "   " "            " "
## 2  ( 1 )  " "  "*"      " " " "   " "            "*"
## 3  ( 1 )  " "  "*"      " " " "   " "            "*"
## 4  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 5  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 6  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 7  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 8  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 9  ( 1 )  " "  "*"      " " "*"   " "            "*"
## 10  ( 1 ) " "  "*"      " " "*"   " "            "*"
## 11  ( 1 ) "*"  "*"      " " "*"   " "            "*"
## 12  ( 1 ) "*"  "*"      " " "*"   "*"            "*"
## 13  ( 1 ) "*"  "*"      "*" "*"   "*"            "*"
## 14  ( 1 ) "*"  "*"      "*" "*"   "*"            "*"
## 15  ( 1 ) "*"  "*"      "*" "*"   "*"            "*"
## 16  ( 1 ) "*"  "*"      "*" "*"   "*"            "*"
##           gendermale height weight framemedium framesmall
## 1  ( 1 )  " "        " "    " "    " "         " "
## 2  ( 1 )  " "        " "    " "    " "         " "
## 3  ( 1 )  " "        " "    " "    " "         " "
## 4  ( 1 )  " "        " "    " "    " "         " "
## 5  ( 1 )  " "        " "    " "    " "         "*"
## 6  ( 1 )  " "        " "    " "    " "         "*"
## 7  ( 1 )  " "        " "    " "    " "         "*"
## 8  ( 1 )  " "        "*"    " "    " "         "*"
## 9  ( 1 )  " "        "*"    "*"    " "         "*"
## 10  ( 1 ) " "        "*"    "*"    " "         "*"
## 11  ( 1 ) " "        "*"    "*"    " "         "*"
## 12  ( 1 ) " "        "*"    "*"    " "         "*"
## 13  ( 1 ) " "        "*"    "*"    " "         "*"
## 14  ( 1 ) " "        "*"    "*"    "*"         "*"
## 15  ( 1 ) " "        "*"    "*"    "*"         "*"
## 16  ( 1 ) "*"        "*"    "*"    "*"         "*"
##           bp.1s bp.1d waist hip time.ppn
## 1  ( 1 )  " "   " "   " "   " " " "
## 2  ( 1 )  " "   " "   " "   " " " "
## 3  ( 1 )  " "   " "   "*"   " " " "
## 4  ( 1 )  " "   " "   "*"   " " " "
## 5  ( 1 )  " "   " "   "*"   " " " "
## 6  ( 1 )  " "   " "   "*"   " " "*"
## 7  ( 1 )  "*"   " "   "*"   " " "*"
## 8  ( 1 )  "*"   " "   "*"   " " "*"
## 9  ( 1 )  " "   " "   "*"   "*" "*"
## 10  ( 1 ) "*"   " "   "*"   "*" "*"
## 11  ( 1 ) "*"   " "   "*"   "*" "*"
## 12  ( 1 ) "*"   " "   "*"   "*" "*"
## 13  ( 1 ) "*"   " "   "*"   "*" "*"
## 14  ( 1 ) "*"   " "   "*"   "*" "*"
## 15  ( 1 ) "*"   "*"   "*"   "*" "*"
## 16  ( 1 ) "*"   "*"   "*"   "*" "*"
```

```r
p.m = as.integer(rownames(sum_sub$which)) + 1
n = nrow(data.c)
ssto = sum((data.c$glyhb-mean(data.c$glyhb))^2)
sse = (1-sum_sub$rsq)*ssto
aic = n*log(sse/n)+2*p.m
bic = n*log(sse/n)+log(n)*p.m
res_sub = cbind(sum_sub$which, sse, sum_sub$rsq, sum_sub$adjr2, sum_sub$cp, bic, aic)

#Model with only intercept
fit1 = lm(glyhb~1, data = data.c)
full = lm(glyhb~., data = data.c)
sse1 = sum(fit1$residuals^2)
p = 1 #only one parameter
c1 = sse1/summary(full)$sigma^2 - (n-2*p)
aic1 = n*log(sse1/n)+2*p
bic1 = n*log(sse1/n)+log(n)*p
none = c(1,rep(0,16),sse1,0,0,c1,bic1,aic1)

#Combine together
res_sub = rbind(none,res_sub)
colnames(res_sub)[19:21] = c("r2p", "r2ap", "cp")
round(res_sub,2)
```

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

```
## 11            0      1      1           0          1     1
## 12            0      1      1           0          1     1
## 13            0      1      1           0          1     1
## 14            0      1      1           1          1     1
## 15            0      1      1           1          1     1
## 16            1      1      1           1          1     1
##      bp.1d waist hip time.ppn  sse  r2p r2ap     cp
## none     0     0   0        0 0.52 0.00 0.00 191.77
## 1        0     0   0        0 0.29 0.44 0.44  27.96
## 2        0     0   0        0 0.26 0.50 0.50   9.01
## 3        0     1   0        0 0.24 0.53 0.52   0.52
## 4        0     1   0        0 0.24 0.53 0.52   0.53
## 5        0     1   0        0 0.24 0.54 0.53   0.05
## 6        0     1   0        1 0.23 0.55 0.53   0.34
## 7        0     1   0        1 0.23 0.55 0.53   1.49
## 8        0     1   0        1 0.23 0.55 0.53   3.13
## 9        0     1   1        1 0.23 0.55 0.53   4.23
## 10       0     1   1        1 0.23 0.55 0.53   5.43
## 11       0     1   1        1 0.23 0.55 0.53   7.24
## 12       0     1   1        1 0.23 0.55 0.52   9.15
## 13       0     1   1        1 0.23 0.55 0.52  11.07
## 14       0     1   1        1 0.23 0.55 0.52  13.02
## 15       1     1   1        1 0.23 0.55 0.51  15.01
## 16       1     1   1        1 0.23 0.55 0.51  17.00
##           bic      aic
## none -1069.26 -1072.47
## 1    -1171.73 -1178.15
## 2    -1186.05 -1195.68
## 3    -1191.47 -1204.31
## 4    -1188.34 -1204.39
## 5    -1185.76 -1205.02
## 6    -1182.39 -1204.86
## 7    -1178.10 -1203.78
## 8    -1173.29 -1202.18
## 9    -1169.07 -1201.16
## 10   -1164.73 -1200.03
## 11   -1159.73 -1198.25
## 12   -1154.63 -1196.35
## 13   -1149.50 -1194.43
## 14   -1144.34 -1192.48
## 15   -1139.15 -1190.50
## 16   -1133.95 -1188.51
```

```
best_for_each1 = lapply(c("r2p", "r2ap"), function(x) round(res_sub,2)[order(-res_sub[, x]),][1,])
best_for_each2 = lapply(c("sse", "cp", "bic", "aic"), function(x) round(res_sub,2)[order(res_sub[, x]),]
best_for_each = append(best_for_each1, best_for_each2)
names(best_for_each) = c("r2p", "r2ap", "sse", "cp", "bic", "aic")
best_for_each
```

```
## $r2p
##    (Intercept)           chol       stab.glu            hdl
##           1.00           1.00           1.00           1.00
##          ratio locationLouisa            age     gendermale
##           1.00           1.00           1.00           1.00
```

```
##         height          weight     framemedium      framesmall
##           1.00            1.00            1.00            1.00
##          bp.1s           bp.1d           waist             hip
##           1.00            1.00            1.00            1.00
##       time.ppn             sse             r2p            r2ap
##           1.00            0.23            0.55            0.51
##             cp             bic             aic
##          17.00       -1133.95       -1188.51
##
## $r2ap
##    (Intercept)            chol       stab.glu             hdl
##           1.00            0.00            1.00            0.00
##          ratio locationLouisa             age      gendermale
##           1.00            0.00            1.00            0.00
##         height          weight     framemedium      framesmall
##           0.00            0.00            0.00            1.00
##          bp.1s           bp.1d           waist             hip
##           0.00            0.00            1.00            0.00
##       time.ppn             sse             r2p            r2ap
##           1.00            0.23            0.55            0.53
##             cp             bic             aic
##           0.34       -1182.39       -1204.86
##
## $sse
##    (Intercept)            chol       stab.glu             hdl
##           1.00            1.00            1.00            1.00
##          ratio locationLouisa             age      gendermale
##           1.00            1.00            1.00            1.00
##         height          weight     framemedium      framesmall
##           1.00            1.00            1.00            1.00
##          bp.1s           bp.1d           waist             hip
##           1.00            1.00            1.00            1.00
##       time.ppn             sse             r2p            r2ap
##           1.00            0.23            0.55            0.51
##             cp             bic             aic
##          17.00       -1133.95       -1188.51
##
## $cp
##    (Intercept)            chol       stab.glu             hdl
##           1.00            0.00            1.00            0.00
##          ratio locationLouisa             age      gendermale
##           1.00            0.00            1.00            0.00
##         height          weight     framemedium      framesmall
##           0.00            0.00            0.00            1.00
##          bp.1s           bp.1d           waist             hip
##           0.00            0.00            1.00            0.00
##       time.ppn             sse             r2p            r2ap
##           0.00            0.24            0.54            0.53
##             cp             bic             aic
##           0.05       -1185.76       -1205.02
##
## $bic
##    (Intercept)            chol       stab.glu             hdl
##           1.00            0.00            1.00            0.00
```

```
##          ratio locationLouisa            age     gendermale
##           0.00           0.00           1.00           0.00
##         height         weight    framemedium     framesmall
##           0.00           0.00           0.00           0.00
##          bp.1s          bp.1d          waist            hip
##           0.00           0.00           1.00           0.00
##       time.ppn            sse            r2p           r2ap
##           0.00           0.24           0.53           0.52
##             cp            bic            aic
##           0.52       -1191.47       -1204.31
##
## $aic
##    (Intercept)           chol       stab.glu            hdl
##           1.00           0.00           1.00           0.00
##          ratio locationLouisa            age     gendermale
##           1.00           0.00           1.00           0.00
##         height         weight    framemedium     framesmall
##           0.00           0.00           0.00           1.00
##          bp.1s          bp.1d          waist            hip
##           0.00           0.00           1.00           0.00
##       time.ppn            sse            r2p           r2ap
##           0.00           0.24           0.54           0.53
##             cp            bic            aic
##           0.05       -1185.76       -1205.02
```

Under SSE_p criterion, the best model is the model with all variables, the "full model".
Under R_p^2 criterion, the best model is also the full model.
Under $R_{a,p}^2$ criterion, the best model is the model with variables "chol", "stab.glu", "ratio", "age", "framesmall", "waist" and "time.ppn"
Under C_p criterion, the best model is the model with variables "stab.glu", "ratio", "age", "framesmall" and "waist".
Under AIC_p criterion, the best model is the model with variables "stab.glu", "ratio", "age", "framesmall" and "waist".
Under BIC_p criterion, the best model is the model with variables "stab.glu", "ratio", "age", "framesmall" and "waist".

The minmum value of C_p is 0.05, it means that the overall (in-sample) mean-squared-estimation-error divided by the error variance is 0.05

The value of C_p is strange, this is because our estimated sigma is big, we can find that when p >= 7, the SSE of these models are about the same, because n is not very big, so $C_p = SSE_p/(SSE/(n-p_{full})) - (n-2p) \approx 2p - p_{full}$, and now the value of C_p is not very meaningful anymore.

(c)

```
fit.null = lm(glyhb~1, data = data.c)
fit.full = lm(glyhb~., data = data.c)
stepAIC(fit.null, scope=list(upper=fit.full), direction="both", k=2)
```

```
## 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
```

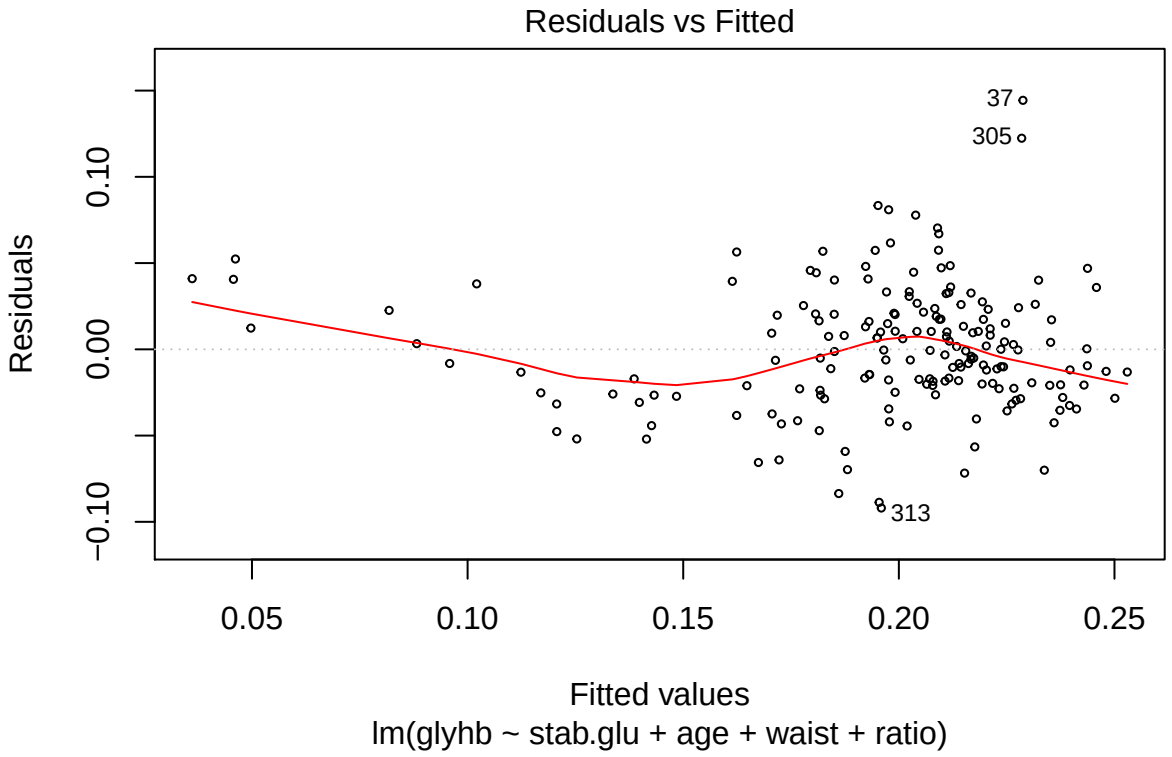
```
## - age       1  0.028996 0.28641 -1178.2
## - stab.glu  1  0.178283 0.43569 -1101.4
##
## Step:  AIC=-1204.31
## glyhb ~ stab.glu + age + waist
##
##            Df Sum of Sq     RSS     AIC
## + ratio     1  0.002746 0.24014 -1204.4
## <none>                  0.24289 -1204.3
## + time.ppn  1  0.001329 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 = data.c)
##
## Coefficients:
## (Intercept)     stab.glu          age        waist
##   0.3489987   -0.0005368   -0.0006412   -0.0013985
##       ratio
##  -0.0028483
```

According to "stepAIC""", the model being selected is the model with variables"stab.glu","age","ratio","waist".

It is not the best model in the previous question. its "AIC" value is a little bit larger than the model in (b), because it doesn't include "frame small"

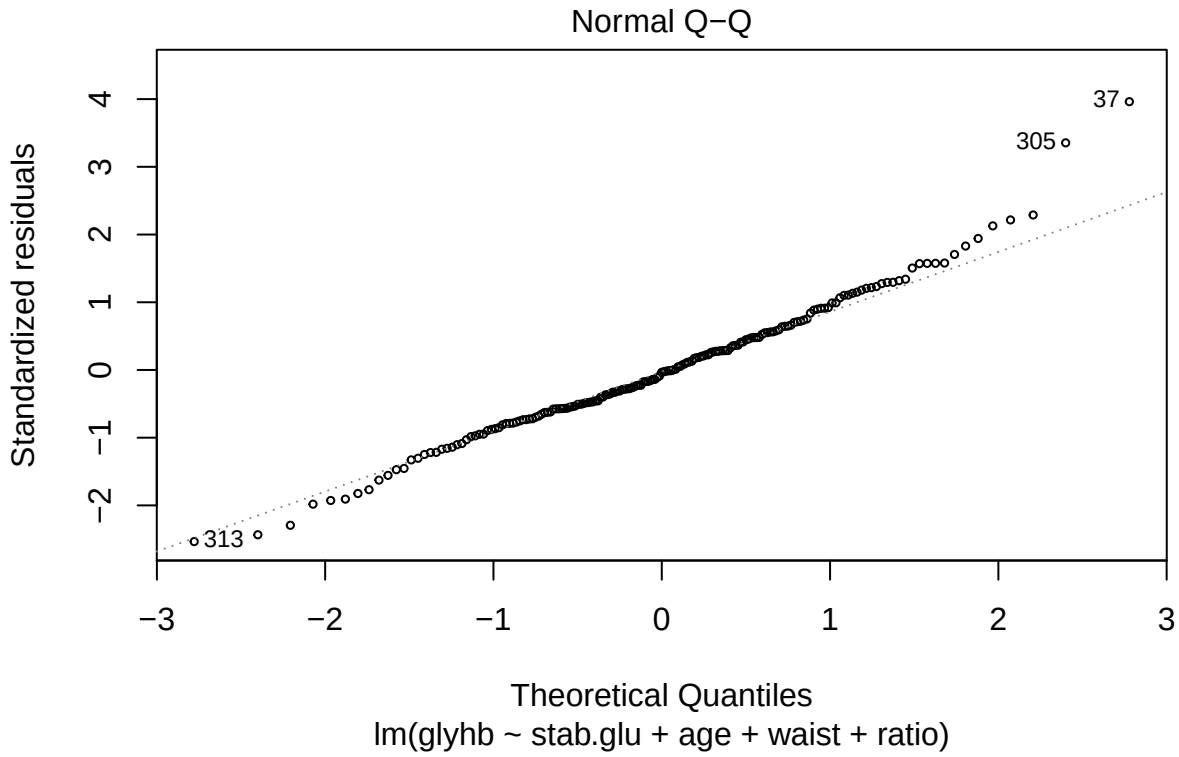
(d)

```
fs1 = lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.c)
par(mfrow = c(1,1))
plot(fs1, which = 1, cex = 0.5)
```



It shows a little nonlinear pattern, and we also observe a little heteroscedasticity.

```
plot(fs1, which = 2, cex = 0.5)
```



The distribution of residuals is a little bit heavy-tailed.
The model seems to be inadequate.

Question 4
(a)

```
fit.full = lm(glyhb~. + .^2, data = data.c)
sum_full = summary(fit.full)
nrow(sum_full$coefficients)
```

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

There are 136 regression coefficients.
The MSE is 0.03219^2 = 0.001036196
There are many insignificant variables included in this model, and the variances of estimations are generally high, also the total in-sample variance is high.

(b)

```
stepAIC(fit.null, scope=list(upper=fit.full), direction="both", k=2)
```

```
## 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 = data.c)
##
## Coefficients:
```

```
##    (Intercept)        stab.glu             age
##      3.527e-01      -9.522e-04       7.247e-05
##          waist           ratio  stab.glu:ratio
##     -1.305e-03      -2.158e-03       7.507e-05
##      age:ratio
##     -1.724e-04
```

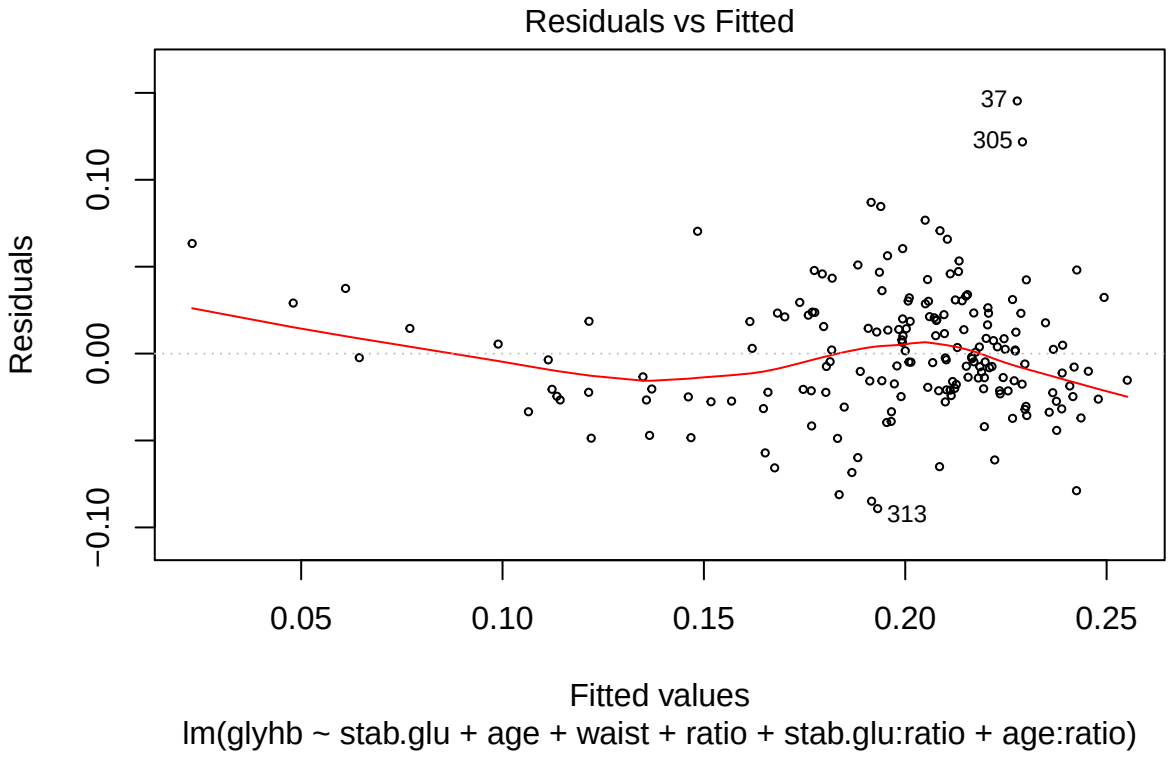
The model being selected is the model with varibles "stab.glu", "age", "ratio", "waist", "stab.glu:ratio" and "age:ratio".

```
fs2 = lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
    age:ratio, data = data.c)
```

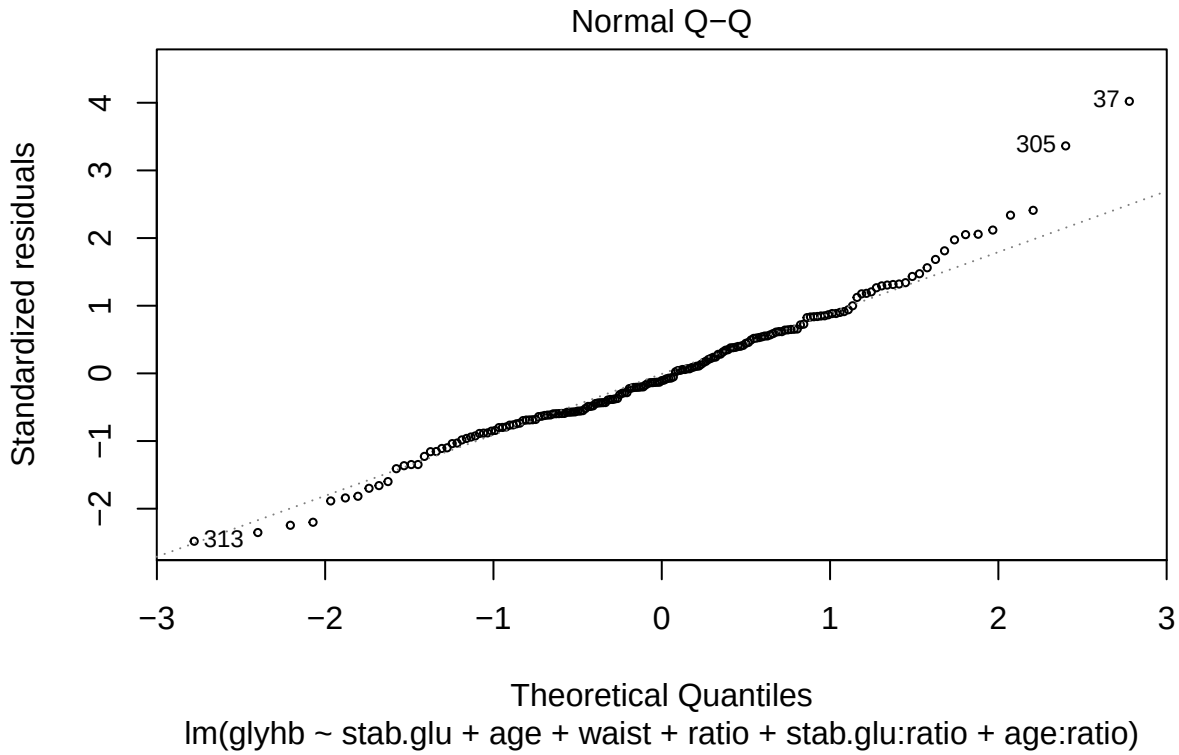
The AIC value in this model is just slightly samller than model fs1.

(c)

```
plot(fs2, which = 1, cex = 0.5)
```



```r
plot(fs2, which = 2, cex = 0.5)
```



No obvious help, it still seems to be not very adequate.

(d)

```r
stepAIC(fit.null, scope=list(upper=fit.full), direction="forward", k=2)
```

```
## 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 = data.c)
##
## Coefficients:
##    (Intercept)        stab.glu             age
##      3.527e-01      -9.522e-04       7.247e-05
##          waist           ratio  stab.glu:ratio
##     -1.305e-03      -2.158e-03       7.507e-05
##      age:ratio
##     -1.724e-04
```

```
#sub_set = regsubsets(glyhb~. + .^2, data = data.s, nbest = 1, nvmax = 16, method = "exhaustive", really
#sum_sub = summary(sub_set)
#sum_sub
```

We end up with the same result as the answer in (b).

Question 5
(a)

```
#Model3
var = c("glyhb", "stab.glu", "age", "ratio", "waist")
#To get a more precise estimation of sigma2, we will use the full dataset "data.s"
data_model3 = data.s[,var]
fs3 = lm(glyhb~. + .^2, data = data_model3)
summary(fs3)
```

```
##
## Call:
## lm(formula = glyhb ~ . + .^2, data = data_model3)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.155347 -0.021285 -0.000838  0.020720  0.145701
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.399e-01  5.112e-02   6.649 1.12e-10 ***
## stab.glu       -7.619e-04  3.533e-04  -2.157   0.0317 *
## age            -2.094e-04  8.056e-04  -0.260   0.7951
## ratio          -5.607e-03  1.001e-02  -0.560   0.5759
## waist          -7.736e-04  1.434e-03  -0.540   0.5898
## stab.glu:age    6.572e-06  2.855e-06   2.302   0.0219 *
## stab.glu:ratio -1.167e-05  2.074e-05  -0.563   0.5739
## stab.glu:waist -9.143e-07  7.702e-06  -0.119   0.9056
## age:ratio      -7.334e-05  8.870e-05  -0.827   0.4089
## age:waist      -2.116e-05  2.102e-05  -1.006   0.3149
## ratio:waist     1.850e-04  2.439e-04   0.758   0.4487
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03637 on 355 degrees of freedom
## Multiple R-squared:  0.5104, Adjusted R-squared:  0.4966
## F-statistic: 37.01 on 10 and 355 DF,  p-value: < 2.2e-16
```

```
p = nrow(summary(fs3)$coefficients)
p
```

```
## [1] 11
```

```
sigma2 = summary(fs3)$sigma^2
```

There are 11 regression coefficients.
MSE = 0.03637^2 = 0.001322777

```
#For Model2
summary(fs2)
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
##     age:ratio, data = data.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.089202 -0.022258 -0.003599  0.021182  0.145324
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.527e-01  3.162e-02  11.152  < 2e-16 ***
## stab.glu       -9.522e-04  2.186e-04  -4.355 2.25e-05 ***
## age             7.247e-05  5.277e-04   0.137   0.8909
## waist          -1.305e-03  5.079e-04  -2.570   0.0110 *
```

```
## ratio          -2.158e-03  6.565e-03  -0.329   0.7427
## stab.glu:ratio  7.507e-05  3.775e-05   1.988   0.0483 *
## age:ratio      -1.724e-04  1.231e-04  -1.401   0.1631
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03646 on 176 degrees of freedom
## Multiple R-squared:  0.5464, Adjusted R-squared:  0.531
## F-statistic: 35.34 on 6 and 176 DF,  p-value: < 2.2e-16
```

```
#SSE = MSE * df(SSE) = 0.03646^2*176 = 0.2339624
#MSE = 0.03646^2 = 0.001329332
MSEp_2 = 0.03646^2
SSEp_2 = sum(residuals(fs2)^2)
SSEp_2
```

```
## [1] 0.2339843
```

```
#Cp for Model2
Cp_2 = SSEp_2/sigma2 - (nrow(data.c)-2*length(fs2$coef))
Cp_2
```

```
## [1] 7.845373
```

```
#Calculat Pressp
Pressp_2 = sum(fs2$residuals^2/(1-influence(fs2)$hat)^2)
Pressp_2
```

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

```
#For Model1
summary(fs1)
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.091989 -0.022720 -0.001251  0.020707  0.144356
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.490e-01  1.843e-02  18.932  < 2e-16 ***
## stab.glu    -5.368e-04  5.219e-05 -10.287  < 2e-16 ***
## age         -6.412e-04  1.698e-04  -3.776 0.000217 ***
## waist       -1.398e-03  5.075e-04  -2.756 0.006465 **
## ratio       -2.848e-03  1.997e-03  -1.427 0.155439
## ---
## Signif. codes:
```

```
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03673 on 178 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.524
## F-statistic: 51.09 on 4 and 178 DF,  p-value: < 2.2e-16
```

```
#SSE = MSE * df(SSE) = 0.03673^2*178 = 0.2401385
#MSE = 0.03673^2 = 0.001349093
MSEp_1 = 0.03673^2
SSEp_1 = sum(residuals(fs1)^2)
SSEp_1
```

```
## [1] 0.2401432
```

```
#Cp for Model2
Cp_1 = SSEp_1/sigma2 - (nrow(data.c)-2*length(fs1$coef))
Cp_1
```

```
## [1] 8.5003
```

```
#Calculat Pressp
Pressp_1 = sum(fs1$residuals^2/(1-influence(fs1)$hat)^2)
Pressp_1
```

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

The value of SSE, MSE, Cp, Press for fs1 and fs2 are close to each other, and fs2 is a little better than fs1 since all these values are smaller in fs2.

Overfitting is not a big concern since the value of Pressp/n is relatively small.

(b)

```
fs1_v = lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.v)
sign(summary(fs1_v)$coefficients[,1]) == sign(summary(fs1)$coefficients[,1])
```

```
## (Intercept)    stab.glu         age       waist       ratio
##        TRUE        TRUE        TRUE        TRUE        TRUE
```

For Model1, All the sign of estimators are the same.

```
fs2_v = lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
    age:ratio, data = data.v)
sign(summary(fs2_v)$coefficients[,1]) == sign(summary(fs2)$coefficients[,1])
```

```
##    (Intercept)       stab.glu            age          waist
##           TRUE           TRUE          FALSE           TRUE
##          ratio stab.glu:ratio      age:ratio
##          FALSE          FALSE          FALSE
```

For Model2, the sign of “age”, “ratio”, “stab.glu:ratio” and “age:ratio” are different. Mainly because these variables are not significant themselves.

```
mod_sum1 = cbind(coef(summary(fs1))[,1], coef(summary(fs1_v))[,1],
                 coef(summary(fs1))[,2], coef(summary(fs1_v))[,2])
colnames(mod_sum1) = c("Train Est","Valid Est","Train s.e.","Valid s.e.")
mod_sum2 = cbind(coef(summary(fs2))[,1], coef(summary(fs2_v))[,1],
                 coef(summary(fs2))[,2], coef(summary(fs2_v))[,2])
colnames(mod_sum2) = c("Train Est","Valid Est","Train s.e.","Valid s.e.")
list_mod_sum = list(round(mod_sum1,5), round(mod_sum2,5))
names(list_mod_sum) = c("fs1", "fs2")
list_mod_sum
```

```
## $fs1
##             Train Est Valid Est Train s.e. Valid s.e.
## (Intercept)   0.34900   0.32871    0.01843    0.01878
## stab.glu     -0.00054  -0.00044    0.00005    0.00006
## age          -0.00064  -0.00067    0.00017    0.00018
## waist        -0.00140  -0.00085    0.00051    0.00049
## ratio        -0.00285  -0.00428    0.00200    0.00147
##
## $fs2
##                 Train Est Valid Est Train s.e. Valid s.e.
## (Intercept)       0.35267   0.31223    0.03162    0.03031
## stab.glu         -0.00095  -0.00024    0.00022    0.00014
## age               0.00007  -0.00084    0.00053    0.00055
## waist            -0.00131  -0.00094    0.00051    0.00050
## ratio            -0.00216   0.00008    0.00657    0.00622
## stab.glu:ratio    0.00008  -0.00004    0.00004    0.00003
## age:ratio        -0.00017   0.00003    0.00012    0.00012
```

For fs1, the value and standard error of estimators are similar.
For fs2, the value and standard error of some estimators are significantly different to each other.

It appears that fs1 has consistent estimates on the training data and validation data, but fs does not.

```
#MSPE, SSEp and Pressp
newdata = data.v[, -5]
MSPE_1 = mean((data.v$glyhb - predict(fs1, newdata))^2)
MSPE_2 = mean((data.v$glyhb - predict(fs2, newdata))^2)
MSPE = c(MSPE_1, MSPE_2)
SSEp = c(SSEp_1/n, SSEp_2/n)
Pressp = c(Pressp_1/n, Pressp_2/n)
crit = cbind(MSPE, SSEp, Pressp)
rownames(crit) = c("fs1", "fs2")
crit
```

```
##            MSPE        SSEp      Pressp
## fs1 0.001329283 0.001312258 0.001385467
## fs2 0.001526420 0.001278603 0.001385155
```

The value of "MSPE", "SSEp" and "Pressp" are close to each other.
Model1 has smaller MSPE.

(c)
Internal validation is based on Pressp, and external validation is base on MSPE, fs1 performs better on external validation and fs2 performs better on internal validation.

I would choose fs1 as my final model. It shows more consistency on training and validation data.

```
fit.final = lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.s)
summary(fit.final)
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.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
```

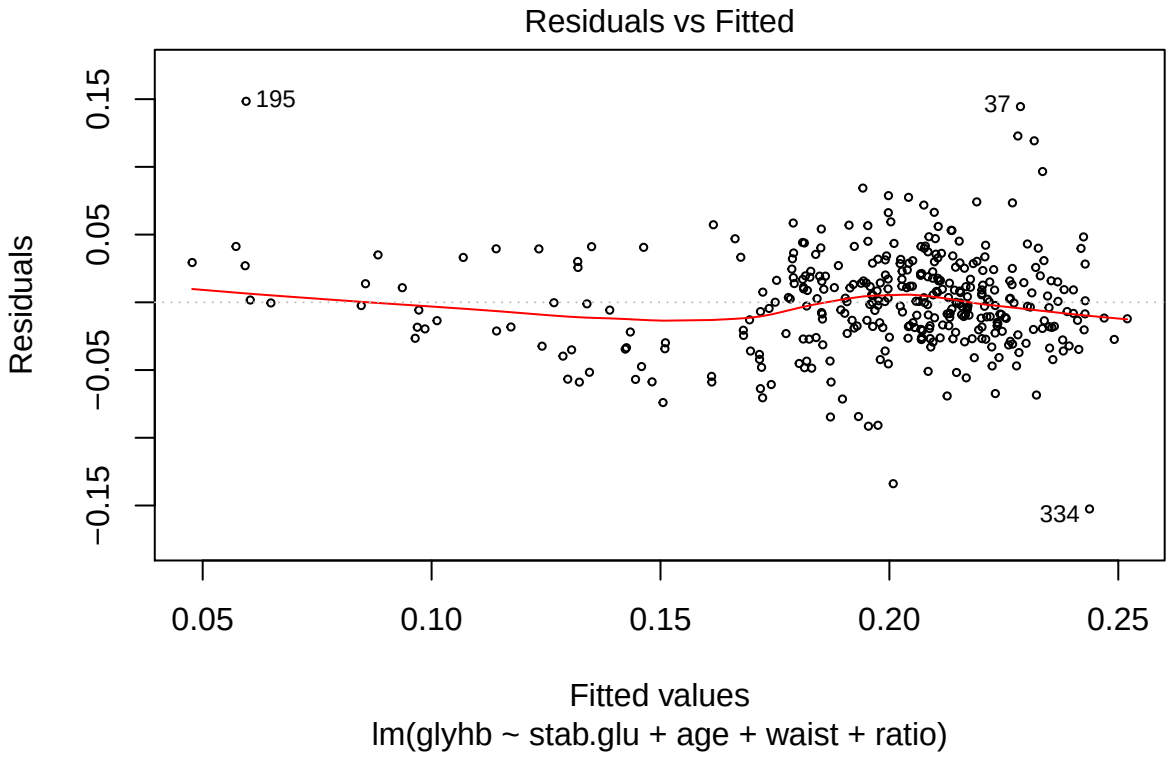
```
anova(fit.final)
```

```
## 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
```

The fitted regression function is $glyhb = 0.34 - 0.0005stab.glu - 0.00066age - 0.0011waist - 0.0037ratio$

Question 6 (a)

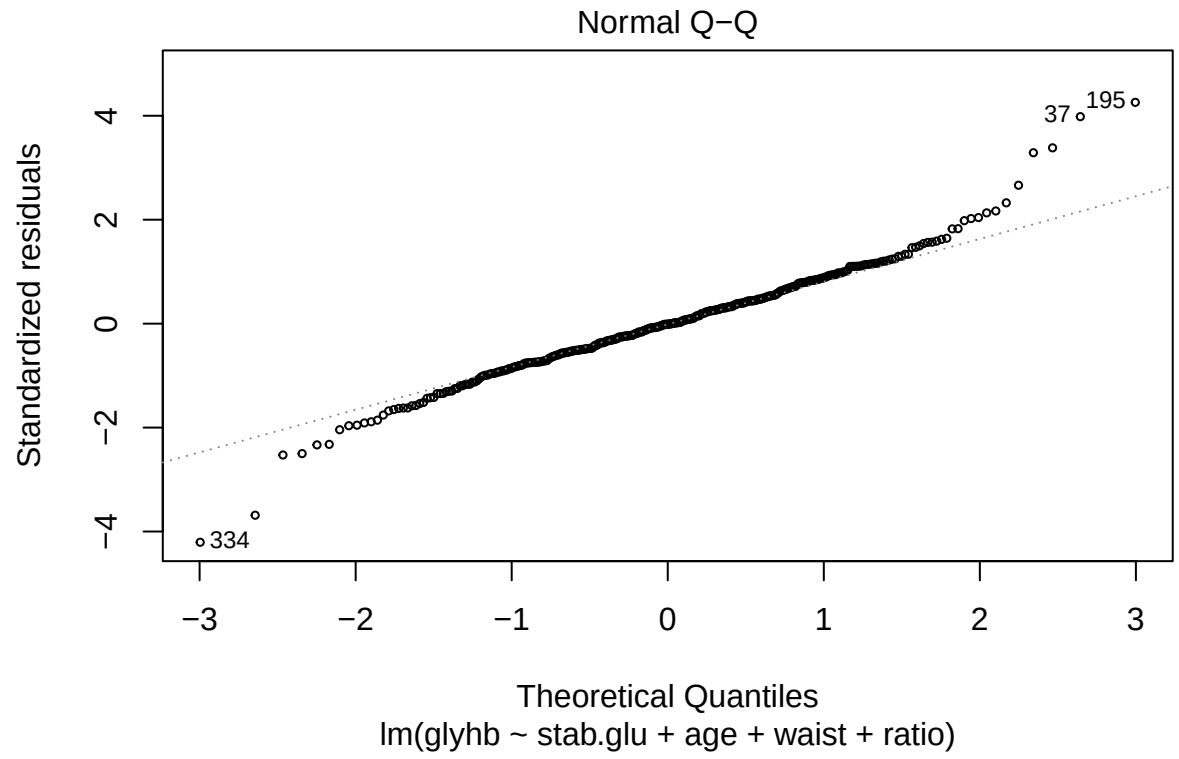
```
plot(fit.final, which = 1, cex =0.5)
```



The nonlinear pattern looks better than previous, but its not meaningful, because the only thing we do is changing our data from training data to the whole data.

We can still observe heteroscedasticity, of course this is the case, because we didn't do any remedy.

```
plot(fit.final, which = 2, cex = 0.5)
```



As good as previous, although there it is a little bit heavy-tailed.

(b)

```
n = nrow(data.s)
stu.res.del = studres(fit.final)
test = qt(1-0.1/(2*n), n-10-1)
stu.res.del[abs(stu.res.del) > test]
```

```
##        37       195       334       363
##  4.067460  4.363632 -4.307267 -3.752719
```

There are 4 outlying Y observations, case 37, 195, 334 and 363.

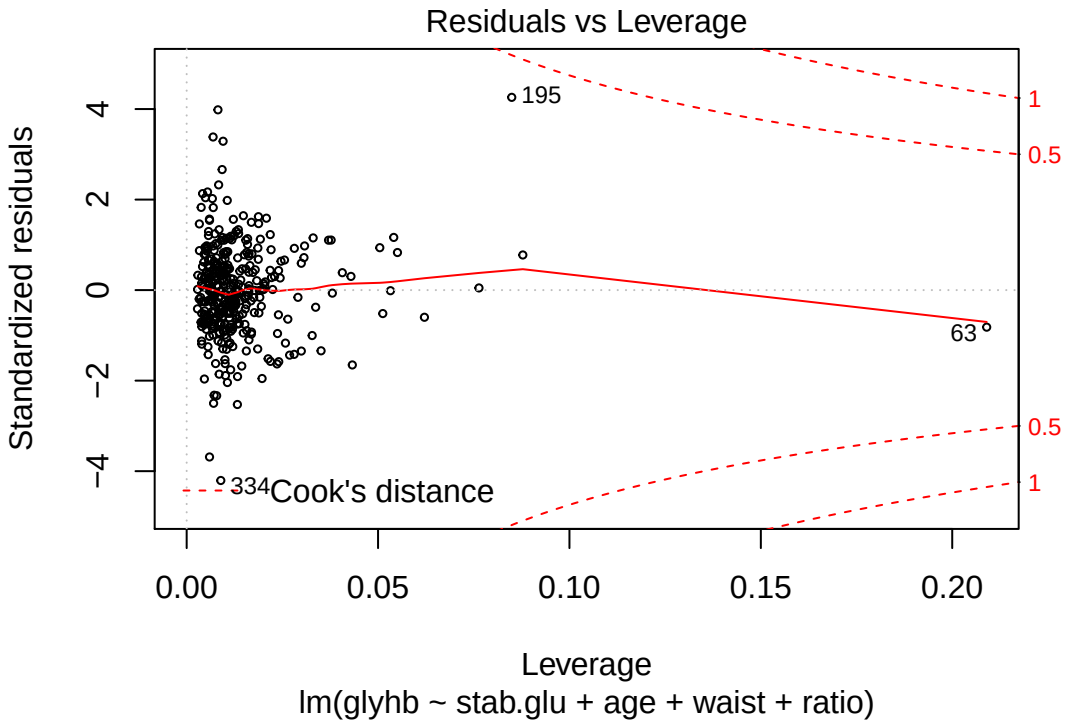
(c)

```
h = as.vector(influence(fit.final)$hat)
index.X = which(h>(2*mean(h)))
index.X
```

```
##  [1]  21  30  42  50  52  54  56  89 118 132 135 139 144 156
## [15] 159 176 233 268 288 299 326 332 348 354 362 363 365
```

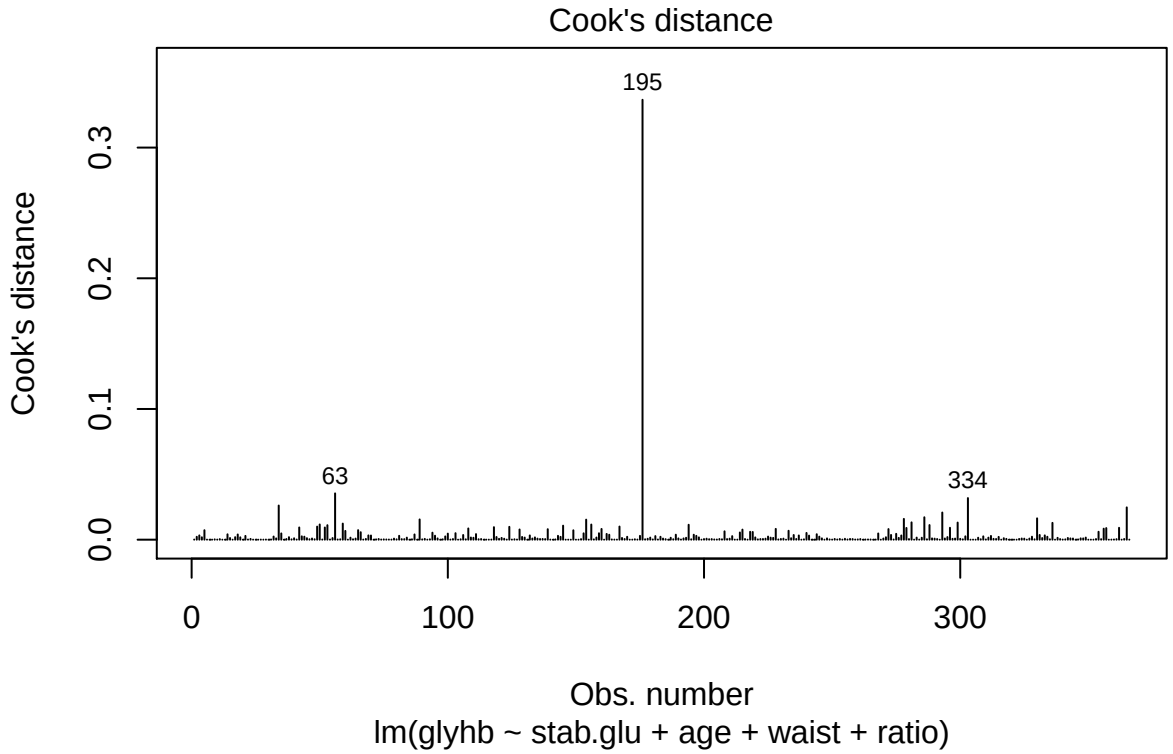
These are the index of outlying X observations. totally 28 cases.

```
par(mfrow = c(1,1), mar = c(5,5,5,5))
plot(fit.final, which = 5, cex = 0.5)
```



(d)

```
res = residuals(fit.final)
mse = anova(fit.final)["Residuals", 3]
p = nrow(coef(summary(fit.final)))
cook.d = res^2*h/(p*mse*(1-h)^2)
plot(fit.final, which = 4)
```




```
head(round(sort(pf(cook.d, p, n-p), decreasing = TRUE),5))
```

```
##     195      63     334      37     401     321
## 0.10935 0.00067 0.00051 0.00032 0.00028 0.00018
```

The maximum of pi is about 10.9%, the 195 case has some impact on the fitted value, but not significant.

(e)

```
#With all cases.
fv_all = predict(fit.final)
fv_all_drop = fv_all[names(fv_all) != "195"]  #Drop the case 195 in order to compare with model without

data_drop = data.s[rownames(data.s) != "195",]
fit.final_drop = lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data_drop)
fv_all_without = predict(fit.final_drop)
avg_abs_per_diff = mean(abs(fv_all_drop/fv_all_without-1))
avg_abs_per_diff
```

```
## [1] 0.01159376
```

So the average absolue percentage difference is 1.16%, so again, the 195 case has some impact on the fitted value, but not significant.