Statistics 206

Homework 8 Solution

Due : Nov. 27, 2019, In Class

1. Tell true or false of the following statements.

   (a) With the R_p^2 criterion, we aim to select the model with the largest R_p^2.

   FALSE. The model with the largest R_p^2 is the full model. We should choose the model(s) starting from where adding additional X variables won't increase R_p^2 much anymore.

   (b) For models of the same size, their $Press_p$ values are monotonically decreasing with the decreasing of SSE_p.

   FALSE. $Press_p$ values are not monotone with respect to SSE_p.

   (c) For models of the same size, their C_p, AIC_p, BIC_p values are monotonically decreasing with the decreasing of SSE_p.

   TRUE.

   (d) For a given model, its SSE_p is always no greater than its $Press_p$.

   TRUE. The fitted value for the ith case when this case is deleted while fitting the regression model can never be better than the fitted value when the ith case is included in regression model fitting.

   (e) Compared with AIC_p, BIC_p criterion tends to select smaller models because it puts more penalty on model size.

   TRUE. $AIC_p = n \log \frac{SSE_p}{n} + 2p$, $BIC_p = n \log \frac{SSE_p}{n} + \log(n)p$. And when $n \geq 8$, then $\log(n) > 2$.

   (f) The stepwise procedures are guaranteed to find the best model according to a given criterion.

   FALSE. They may end up with suboptimal models rather than the global optimal.

Problems 2 to 4. Model Building and model selection case study in R. *Diabetes data. This data consist of* 19 *variables on* 403 *subjects from* 1046 *subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. We will consider building regression models with* `glyhb` *as the response variable as Glycosolated Hemoglobin* > 70 *is often taken as a positive diagnostics of diabetes. The data set and description are under Files/Homework. Please attach your R codes and plots.*

2. **Processing of the data.**

   (a) Read the data into R. Replace the missing values in the variable `frame` (indicated by an empty string ‘’) by ’NA’ and drop the old class ‘’.

```
> diabetes = read.table('diabetes.txt', header=TRUE)  #read data
> is.na(diabetes$frame)=which(diabetes$frame=='')  #repalce '' with NA
> diabetes$frame=droplevels(diabetes$frame) #takes away the old class ''
> summary(diabetes$frame)
large medium  small   NA's
103    184    104     12
```

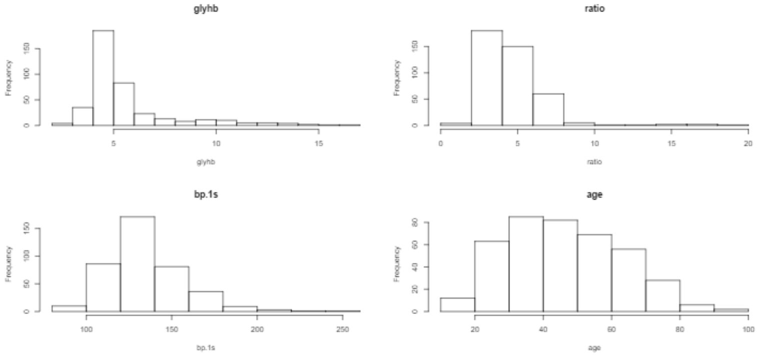
(b) Drop `id, bp.2s, bp.2d` from the data. The column `id` are patient IDs and thus is not a meaningful predictor. The variables `bp.2s, bp.2d` have many missing values. You may use the code:

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

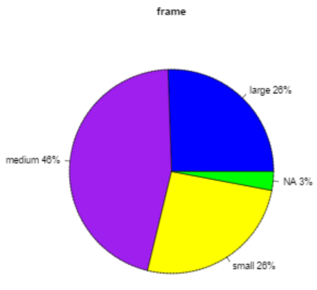
(c) Which of the (remaining) variables are quantitative variables and which are qualitative variables? Draw histogram for `glyhb` and comment on its distribution. Draw histograms for the rest quantitative variables and draw pie charts for qualitative variables.

```
> sapply(diabetes,class)
id       chol   stab.glu       hdl      ratio      glyhb  location
"integer" "integer" "integer" "integer" "numeric" "numeric"  "factor"
age      gender     height     weight      frame      bp.1s      bp.1d
"integer"  "factor" "integer" "integer"  "factor" "integer" "integer"
bp.2s      bp.2d      waist       hip  time.ppn
"integer" "integer" "integer" "integer" "integer"
```

`glyhb`,`ratio`,`bp.1s` and `age` are quantitative variables (either `numeric` or `integer`). `gender` and `frame` are qualitative variables (`factor`).

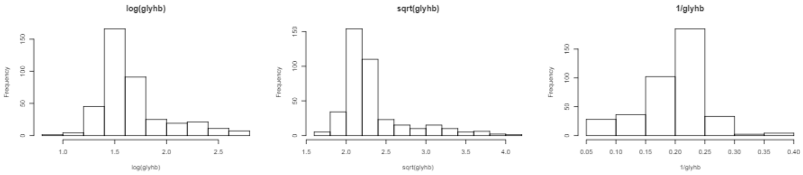


All four variables appear more or less right skewed.



`frame` has nearly half (46%) medium, one quarter (26%) large and one quarter (26%) small. `gender` has 58% female and 42% male.

(d) It turns out that the distribution of `glyhb` is severely right-skewed. Thus, you want to consider some transformations. Draw histogram for $\log(glyhb)$, $\sqrt{glyhb}$ and $\frac{1}{glyhb}$, respectively. Which distribution appears to be the most Normal like among the three? Denote it by `glyhb*`.



The third one, $\frac{1}{glyhb}$ appears to be the most normal like. Denote `glyhb*` $= \frac{1}{glyhb}$.

```
> glyhbs=1/diabetes$glyhb
> diabetes=cbind(glyhbs,diabetes)
```

(e) Replace the column `glyhb` in `data` by `glyhb*` and refer to `glyhb*` as `glyhb` hereafter and use it as the response variable.

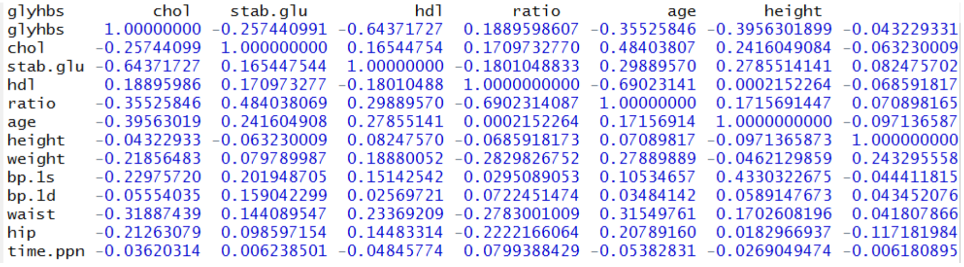
(f) Drop all the cases having missing value. You may use the code:

```
> index.na=apply(is.na(data), 1, any)
                                   ## identify cases with missing value.
> data.s=data[index.na==FALSE,]   ##drop cases with missing value.
> any(is.na(data.s)) ## this should return FALSE -- no NA in data.s
> dim(data.s)  ##this should return 366 16: 366 cases, 16 variables.
> table(data.s$frame)  ## this should show three classes.
```

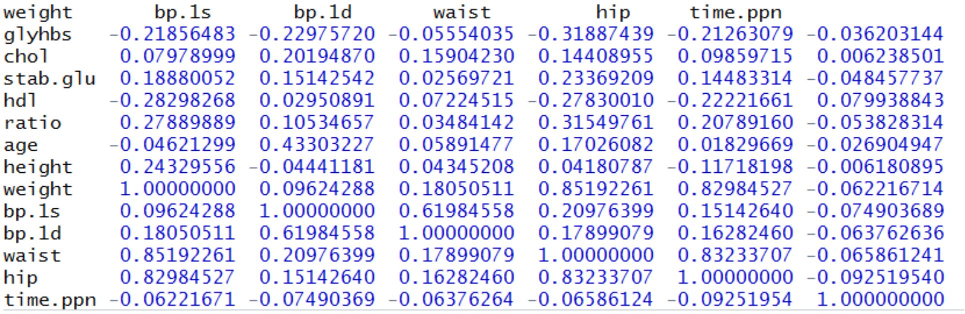
(g) Draw scatterplot matrix and obtain the pairwise correlation matrix for all quantitative variables. Do you observe nonlinearity?

The pairwise correlations are as follows:

Figure 1: Pairwise Correlations



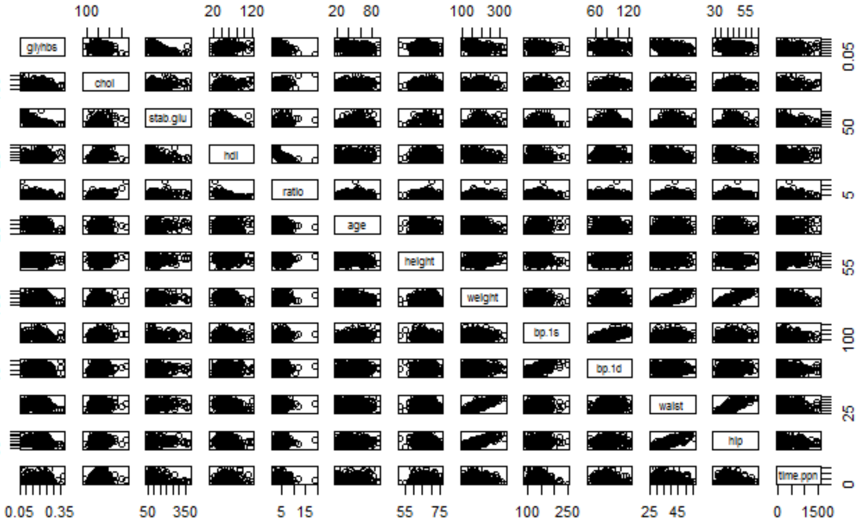
```
glyhbs          chol    stab.glu          hdl        ratio          age        height
glyhbs    1.00000000 -0.257440991 -0.64371727  0.1889598607 -0.35525846 -0.3956301899 -0.043229331
chol     -0.25744099  1.000000000  0.16544754  0.1709732770  0.48403807  0.2416049084 -0.063230009
stab.glu -0.64371727  0.165447544  1.00000000 -0.1801048833  0.29889570  0.2785514141  0.082475702
hdl       0.18895986  0.170973277 -0.18010488  1.0000000000 -0.69023141  0.0002152264 -0.068591817
ratio    -0.35525846  0.484038069  0.29889570 -0.6902314087  1.00000000  0.1715691447  0.070898165
age      -0.39563019  0.241604908  0.27855141  0.0002152264  0.17156914  1.0000000000 -0.097136587
height   -0.04322933 -0.063230009  0.08247570 -0.0685918173  0.07089817 -0.0971365873  1.000000000
weight   -0.21856483  0.079789987  0.18880052 -0.2829826752  0.27889889 -0.0462129859  0.243295558
bp.1s    -0.22975720  0.201948705  0.15142542  0.0295089053  0.10534657  0.4330322675 -0.044411815
bp.1d    -0.05554035  0.159042299  0.02569721  0.0722451474  0.03484142  0.0589147673  0.043452076
waist    -0.31887439  0.144089547  0.23369209 -0.2783001009  0.31549761  0.1702608196  0.041807866
hip      -0.21263079  0.098597154  0.14483314 -0.2222166064  0.20789160  0.0182966937 -0.117181984
time.ppn -0.03620314  0.006238501 -0.04845774  0.0799388429 -0.05382831 -0.0269049474 -0.006180895
```



```
weight        bp.1s       bp.1d       waist         hip      time.ppn
glyhbs   -0.21856483 -0.22975720 -0.05554035 -0.31887439 -0.21263079 -0.036203144
chol      0.07978999  0.20194870  0.15904230  0.14408955  0.09859715  0.006238501
stab.glu  0.18880052  0.15142542  0.02569721  0.23369209  0.14483314 -0.048457737
hdl      -0.28298268  0.02950891  0.07224515 -0.27830010 -0.22221661  0.079938843
ratio     0.27889889  0.10534657  0.03484142  0.31549761  0.20789160 -0.053828314
age      -0.04621299  0.43303227  0.05891477  0.17026082  0.01829669 -0.026904947
height    0.24329556 -0.04441181  0.04345208  0.04180787 -0.11718198 -0.006180895
weight    1.00000000  0.09624288  0.18050511  0.85192261  0.82984527 -0.062216714
bp.1s     0.09624288  1.00000000  0.61984558  0.20976399  0.15142640 -0.074903689
bp.1d     0.18050511  0.61984558  1.00000000  0.17899079  0.16282460 -0.063762636
waist     0.85192261  0.20976399  0.17899079  1.00000000  0.83233707 -0.065861241
hip       0.82984527  0.15142640  0.16282460  0.83233707  1.00000000 -0.092519540
time.ppn -0.06221671 -0.07490369 -0.06376264 -0.06586124 -0.09251954  1.000000000
```

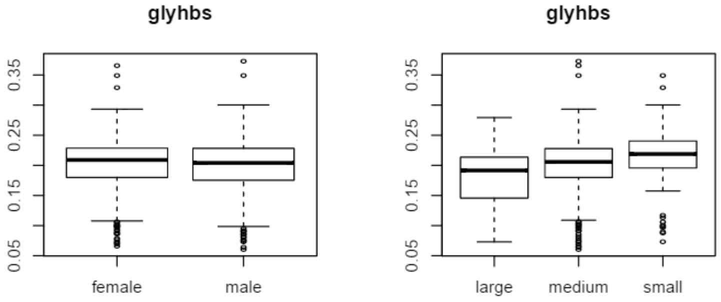
The scatterplot matrix is as follows:

Figure 2: Scatter Plot Matrix for Quantitative Variables



There is no obvious nonlinearity between `glyhb` with the other variables. There are positive linear relationships between weight and waist, weight and hip, bp.1s and bp.1d, waist and hip. We can see the correlation between these pairs are high.

(h) Draw side-by-side box plots to show how `glyhb` is distributed in male and female, and how it is distributed in the three `frame` classes.



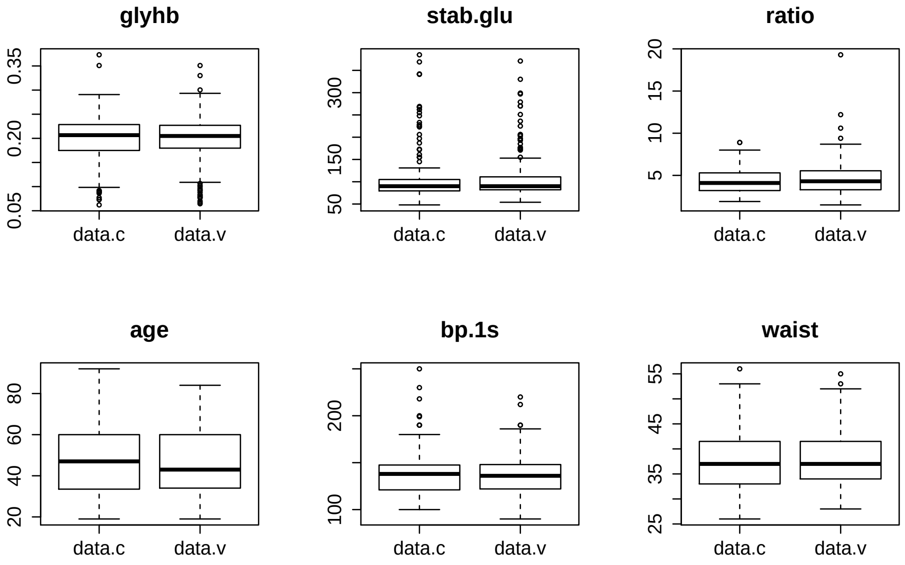
The distribution of `glyhb` is more symmetric in within each class. Also `glyhb` appears to decrease from small to large frame.

(i) Randomly split data into two equal halves: a training data set and a validation data set. You may use the code:

```
> set.seed(10) ## set seed for random number generator
               ##so everyone gets the same split of the data.
> n.s=nrow(data.s) ## number of cases in data.s (366)
> index.s=sample(1: n.s, size=366/2, replace=FALSE)
               ## randomly sample 183 cases to form the training data.
> data.c=data.s[index.s,]  ## get the training data set.
> data.v=data.s[-index.s,]
               ## the remaining 183 cases form the validation set.
```

(j) Examine whether the training data and validation data look alike. Draw side-by-side boxplots for `glyhb`, `stab.glu`, `ratio`, `age`, `bp.1s` and `waist`, in training data and validation data, respectively. Are these variables having similar distributions in these two sets?

```
> par(mfrow=c(2,3))
> boxplot(data.c$glyhb,data.v$glyhb,main=’glyhb’,names=c(’data.c’,’data.v’))
> boxplot(data.c$stab.glu,data.v$stab.glu,main=’stab.glu’,names=c(’data.c’,
+                                                                  ’data.v’))
> boxplot(data.c$ratio,data.v$ratio,main=’ratio’,names=c(’data.c’,’data.v’))
> boxplot(data.c$age,data.v$age,main=’age’,names=c(’data.c’,’data.v’))
> boxplot(data.c$bp.1s,data.v$bp.1s,main=’bp.1s’,names=c(’data.c’,’data.v’))
> boxplot(data.c$waist,data.v$waist,main=’waist’,names=c(’data.c’,’data.v’))
```



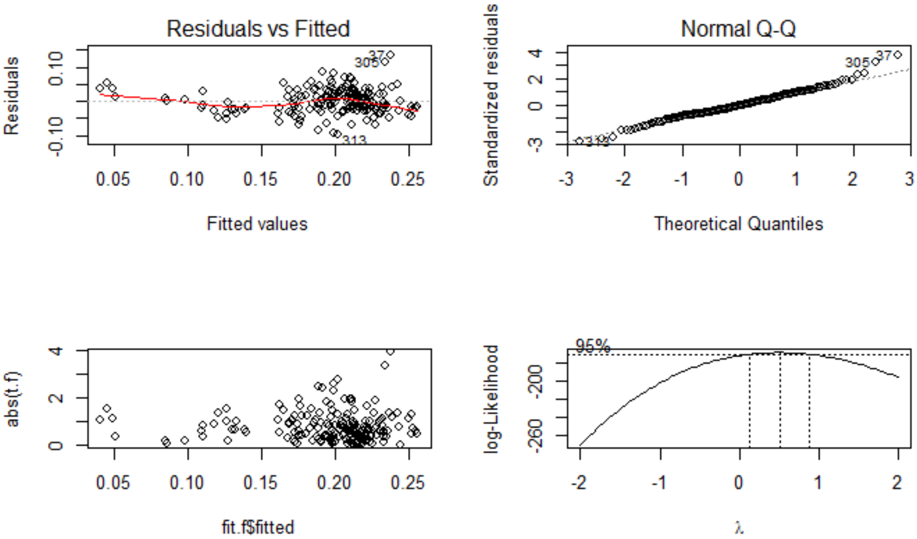
Yes, they have similar distributions.

3. **Selection of first-order effects.** We now consider subsets selection from the pool of all first-order effects of the 15 predictors.

   (a) Fit a model with all first-order effects (Model 1). How many regression coefficients are there in this model? What is the *MSE* from this model? Apply box-cox procedure on this model. Does it appear that any transformation of the response variable is still needed?

```
lm(glyhb ˜., data=data.c) ## data.c denotes the training data
Call:
lm(formula = glyhbs ˜ ., 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


> fit1=lm(glyhb˜.,data=data.c)
> length(fit1$coefficients)   #17 regression coefficients
[1] 17
> anova(fit1)['Residuals',3]   #MSE
[1] 0.001383855
```

Figure 3: Model Diagnostics



The box-cox plot suggests no further transformation on the response variable is needed.

(b) Consider best subsets selection using the R function `regsubsets()` from the `leaps` library with Model 1 as the full model. Return the top 1 best subset of all subset sizes up to 16. Get SSE_p,R^2_p, $R^2_{a,p}$, C_p, AIC_p,BIC_p for each of these models. Identify the best model according to each criterion. For the best model according to C_p criterion, what do you observe about its C_p value? Do you have a possible explanation?

```
> library(leaps)
> sub_set=regsubsets(glyhb~.,data=data.c,nbest=1,nvmax=16,method="exhaustive")
> sum_sub=summary(sub_set)
> n=nrow(data.c)
> ## number of coefficients in each model: p
> p.m=as.integer(as.numeric(rownames(sum_sub$which))+1)
> sse=sum_sub$rss
> 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,
+               aic, bic)
> fit0=lm(glyhb~1,data=data.c) ##fit the model with only intercept
> sse1=sum(fit0$residuals^2)
> p=1
> c1=sse1/0.001384-(n-2*p)
> aic1=n*log(sse1/n)+2*p
> bic1=n*log(sse1/n)+log(n)*p
> none=c(1,rep(0,16),sse1,0,0,c1,bic1,aic1)
> res_sub=rbind(none,res_sub) ##combine the results with other models
> colnames(res_sub)=c(colnames(sum_sub$which),"sse", "R^2", "R^2_a", "Cp",
+                     "aic", "bic")
> res_sub
(Intercept) chol stab.glu hdl ratio locationLouisa age gendermale height
none           1    0        0   0     0              0   0          0      0
1              1    0        1   0     0              0   0          0      0
2              1    0        1   0     0              0   1          0      0
3              1    0        1   0     0              0   1          0      0
4              1    0        1   0     1              0   1          0      0
5              1    0        1   0     1              0   1          0      0
6              1    0        1   0     1              0   1          0      0
7              1    0        1   0     1              0   1          0      0
8              1    0        1   0     1              0   1          0      1
9              1    0        1   0     1              0   1          0      1
10             1    0        1   0     1              0   1          0      1
11             1    1        1   0     1              0   1          0      1
12             1    1        1   0     1              1   1          0      1
13             1    1        1   1     1              1   1          0      1
14             1    1        1   1     1              1   1          0      1
15             1    1        1   1     1              1   1          0      1
16             1    1        1   1     1              1   1          1      1
weight framemedium framesmall bp.1s bp.1d waist hip time.ppn       sse
none      0           0          0     0     0     0   0        0 0.5158646
1         0           0          0     0     0     0   0        0 0.2864076
2         0           0          0     0     0     0   0        0 0.2574112
3         0           0          0     0     0     1   0        0 0.2428890
4         0           0          0     0     0     1   0        0 0.2401432
5         0           0          1     0     0     1   0        0 0.2367131
6         0           0          1     0     0     1   0        1 0.2343460
7         0           0          1     1     0     1   0        1 0.2331725
8         0           0          1     1     0     1   0        1 0.2326634
9         1           0          1     0     0     1   1        1 0.2314193
10        1           0          1     1     0     1   1        1 0.2303187
11        1           0          1     1     0     1   1        1 0.2300477
```

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

Best model:

SSE, R^2: Model 16 (full model)

R_a^2: Model 6 (glu, ratio, age, frame, waist, time)

C_p, AIC: Model 5 (glu, ratio, age, frame, waist)

BIC: Model 3 (glu, age, waist)

For the model with the smallest C_p statistic (Model 5), its C_p value is 0.053 which is much smaller than $p(= 6)$ of this model. Here all the models being considered are submodels of the full model, so their $SSE \geq SSE_f$ and thus the C_p statistic of a submodel satisfies $C_p \geq (n - P) - (n - 2p) = 2p - P$. If SSE_f is not much smaller than SSE of a submodel (i.e., the additional variables in the full model have not much additional contribution in explaining Y), then the C_p of the submodel could be quite small.

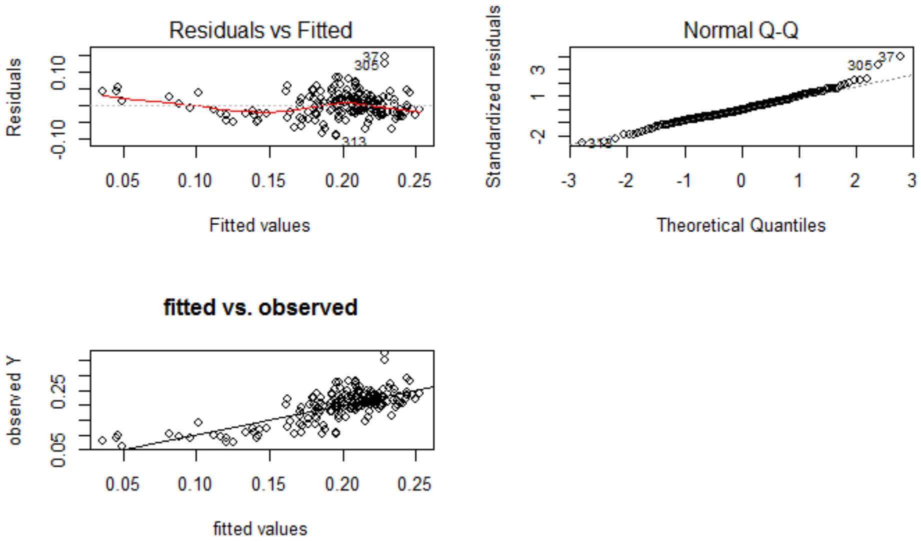
(c) We now explore stepwise procedures. Apply the `forward stepwise procedure` using R function `stepAIC()`, starting from the null-model and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs1. Is it the "best" model according to AIC_p criterion identified in the previous question? If not, how its AIC value compare with AIC of the "best" model?

```
> library(MASS)
> step.f=stepAIC(fit0,scope=list(upper=fit1, lower=~1), direction="both",
+                  k=2)
> step.f$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
glyhb ~ 1

Final Model:
glyhb ~ stab.glu + age + waist + ratio


Step Df    Deviance Resid. Df Resid. Dev       AIC
1                                 182  0.5158646 -1072.466
2 + stab.glu  1 0.229457010       181  0.2864076 -1178.148
3      + age  1 0.028996427       180  0.2574112 -1195.682
4    + waist  1 0.014522110       179  0.2428890 -1204.309
5    + ratio  1 0.002745821       178  0.2401432 -1204.389
```

The final model contains glu, age, waist and ratio. It's not the best model according to AIC_p criterion identified in part (i) and it's AIC of -1204.389 is slightly larger, indicating a slightly suboptimal model.

(d) Comment on the residual vs. fitted value plot and the residual Q-Q plot of Model fs1. Does this model appear to be adequate?

Figure 4: Model Diagnostics for Model fs1



The residual vs. fitted plot shows non-constant error variance. The Q-Q plot indicates slight right skewness. Otherwise, the model seems reasonable.

4. **Selection of first- and second- order effects.** We now consider subsets selection from the pool of first-order effects as well as 2-way interaction effects of the 15 predictors.

   (a) Fit a model with all first-order and 2-way interaction effects (Model 2). How many regression coefficients are there in this model? What is the MSE from this model? Do you have any concern about the fitting of this model and why?

   ```
   > fit2=lm(glyhb~.^2,data=data.c)
   > length(fit2$coefficients) #number of coefficients
   [1] 136
   > anova(fit2)["Residuals",3] #MSE
   [1] 0.001036088
   ```

   Relative to the sample size, there are too many X variables (136) in the model.

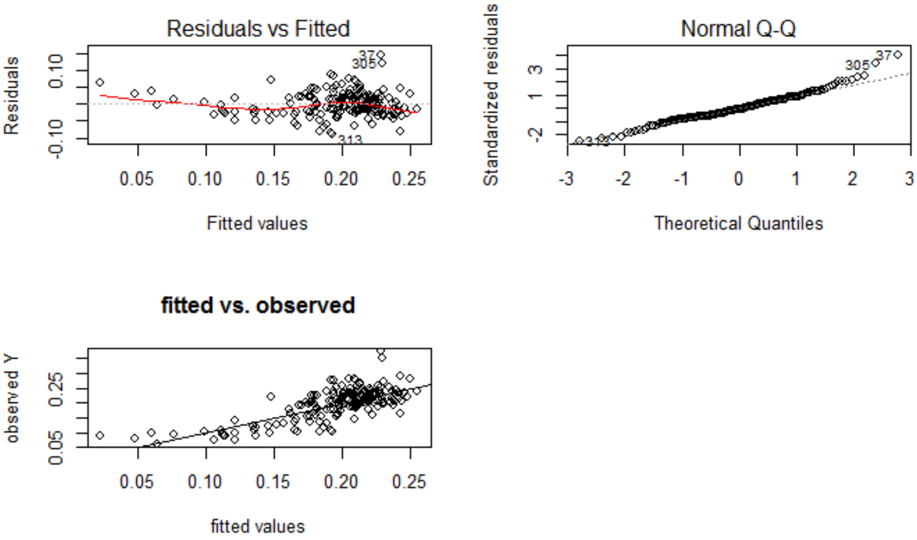
   (b) Apply the `forward stepwise procedure` using R function `stepAIC()`, starting from the null-model and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs2. Compare its AIC value with that of Model fs1. What do you find?

   ```
   > step.f2=stepAIC(fit0,scope=list(upper=fit2, lower=~1), direction="both",
   + k=2)
   > step.f2$anova
   Stepwise Model Path
   Analysis of Deviance Table

   Initial Model:
   glyhb ~ 1

   Final Model:
   glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio


   Step Df    Deviance Resid. Df Resid. Dev       AIC
   1                                       182  0.5158646 -1072.466
   2       + stab.glu  1 0.229457010       181  0.2864076 -1178.148
   3            + age  1 0.028996427       180  0.2574112 -1195.682
   4          + waist  1 0.014522110       179  0.2428890 -1204.309
   5          + ratio  1 0.002745821       178  0.2401432 -1204.389
   6 + stab.glu:ratio  1 0.003550630       177  0.2365926 -1205.115
   7      + ratio:age  1 0.002608308       176  0.2339843 -1205.144
   ```

   The final model contains glu, age, waist, ratio, glu:ratio and age:ratio. Its AIC is -1205.144, which is slightly smaller than that of Model.fs1.

(c) Comment on the residual vs. fitted value plot and the residual Q-Q plot of Model fs2. Does this model appear to be adequate?

Figure 5: Model Diagnostics for Model fs2



The residual vs. fitted plot still shows non-constant error variance. The Q-Q plot indicates slight right skewness. Otherwise, the model seems reasonable.

(d) Apply the `forward selection procedure`. What model do you end up with?

Notes: You could try best subsets selection using the R function `regsubsets()` from the `leaps` library, e.g. return the top 1 best subset of all subset sizes up to 16 with the full model being Model 2. However, be careful, you may have to stop the R session due to the slowness of this procedure! So save all that you want to save before you try this.

```
> step.f3=stepAIC(fit0,scope=list(upper=fit2, lower=~1), direction="forward",
+                     k=2)
> step.f3$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
glyhb ~ 1

Final Model:
glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio + age:ratio
```

```
Step Df    Deviance Resid. Df Resid. Dev       AIC
1                                       182  0.5158646 -1072.466
2       + stab.glu  1 0.229457010       181  0.2864076 -1178.148
3            + age  1 0.028996427       180  0.2574112 -1195.682
4          + waist  1 0.014522110       179  0.2428890 -1204.309
5          + ratio  1 0.002745821       178  0.2401432 -1204.389
6 + stab.glu:ratio  1 0.003550630       177  0.2365926 -1205.115
7      + ratio:age  1 0.002608308       176  0.2339843 -1205.144
```

We end up with Model.fs2, the same model obtained by forward stepwise procedure.