Statistics 206

Homework 8 (Solution)

Due : Friday, Nov. 25, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. Tell true or false of the following statements. Provide a brief justification for you answer.

   (a) A correct model must be a good model.

ANS. FALSE. Consider the case where the true model actually includes a large number of predictors X. In practice (with a possibly not large sample size), it may be difficult to fit such a complicated model, i.e., to obtain good estimation for the coefficients. In other words, the true model, though unbiased, may have large model variance. Therefore, in terms of prediction, simpler models with fewer predictors may be preferred.

   (b) With many nuisance X variables, the model tends to have a large model bias.

ANS. FALSE. Will not necessarily be biased but will have large variance.

   (c) We should select the model with the largest R^2.

ANS. FALSE. It will always increase R^2 if we include more predictors. However, complicated models are not always preferred.

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

ANS. TRUE. This is straightforward according to the formula.

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

ANS. 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.

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

ANS. 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 $\log(n) > 2$.

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

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

2. **Bias-variance trade-off**. Consider the following simulation study. You can modify the codes in "bias-variance-trade-off-simulation2.R". You should read the codes carefully and use R help whenever necessary to understand the codes.

   - The true regression function is

   $$f(x) = \sin(x) + \sin(2x).$$

   - The sample size is $n = 30$ and the design points X_i are equally spaced on $[-3, 3]$.
   - The models to be considered are polynomial regression models with order $l = 1, 2, 3, 5, 7, 9$.
   - The observed data are generated according to:

   $$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad \epsilon_i \sim_{i.i.d.} N(0, \sigma^2).$$

   - Consider three different noise levels with $\sigma = 0.5, \quad 2, \quad 5$.
   - Generate 1000 independent sets (replicates) of observations under each noise level.

   Answer the following questions and include relevant plots along with your answers.

   (a) Is there a correct model among the models being considered?Explain your answer.

ANS. There is no correct model since $f(x)$ can not be represented by finite order polynomials

(b) What is the model variance for each of these models? Does the model variance change with the error variance?

ANS. The model variance is $p\sigma^2$. Here $p = l + 1$. It changes with the error variance σ^2.

(c) Comment on the model bias for each of these models. Does the model bias change with the error variance?

ANS. For $l = 1, 2, 3$, the model has large bias; for $l = 5$, the model has small bias; for $l = 7, 9$, there is little bias. The model bias does not change with the error variance.

(d) Which one, the model variance or the model bias, is the dominant component in the mean-squared-estimation-error? Does the answer depend on the error variance and why?

ANS. When $\sigma = 0.5$, model bias is dominant.
When $\sigma = 2$, bias and variance are on similar scale.
When $\sigma = 5$, model variance becomes dominant.

(e) Which model is the best model according to the mean-squared-estimation-error? Does the answer depend on the error variance and why?

ANS. When $\sigma = 0.5$, $l = 7$ gives the best model. Since model bias is dominant, models with small bias are preferred.
When $\sigma = 2$, $l = 5$ achieves the optimal bias-variance trade-off.
When $\sigma = 5$, the simplest model $l = 1$ is the best. Since model variance is dominant, models with small variance (simple models) are preferred.

(f) Comment on $E(SSE)$. Do you observe different patterns under different noise levels? Give an explanation.

ANS. When $\sigma = 0.5$, model bias is dominant. For underfit models $E(SSE)$ is much larger than $(n - p)\sigma^2$.
When $\sigma = 2$, for underfit models $E(SSE)$ is considerably larger than $(n - p)\sigma^2$
When $\sigma = 5$, model variance is dominant. For all models, $E(SSE)$ is only slightly larger (in terms of relative maginitude) than $(n - p)\sigma^2$.

Problems 3, 4, 5, 6, 7. 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. You are required to use R Markdown: Please submit a .rmd file* ***and*** *its corresponding .html file.*

3. **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 ‘’.

(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.

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

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

(f) 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.

(g) Split data into two equal halves: a training data set and a validation data set. Set seed for random number generator using "set.seed(10)", so the results are reproducible.

(h) 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?

4. **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 (further) transformation of the response variable is still 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?

(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?

(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?

5. **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?

   (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?

   (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?

   (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.

6. **Model validation.** We now consider validation of **Model fs1** and **Model fs2**.

   (a) **Internal validation of Models fs1 and fs2**. For this purpose, we need to compute C_p and $Press_p$ for these models. For C_p, we need an unbiased estimator of the error variance σ^2. The largest model we have considered so far is **Model 2**. However, this model has a very large number of regression coefficients (relative to the sample size), making its parameter estimation unreliable due to large sampling variability. Therefore, we decided to use a smaller model consisting of all predictors identified by **Model fs1** (the forward stepwise selected first-order model), as well all the 2-way interaction terms among these predictors. Denote this model by **Model 3**. Note that, **Model fs2** is also a sub-model of**Model 3**. How many regression coefficients are there in Model 3 ? What is MSE from **Model 3**? Calculate SSE_p, MSE_p, C_p and $Press_p$ for **Model fs1** and **Model fs2**, respectively, and briefly comment on the results, e.g., does it appear to be substantial model bias in these two models? Should over-fitting be a concern?

   (b) **External validation using the validation set**. We now fit **Model fs1** and **Model fs2** on the validation data set. Compare the fitted regression coefficients from the training data and those from the validation data. Are the two sets of estimated regression coefficients having the same sign? Are their values similar? How about the two sets of standard errors? Does it appear that **Model fs1** and **Model fs2** have consistent estimates on the training data and validation data? Calculate the

mean squared prediction error (MSPE) using the validation data for each of the two models. How do these $MSPE_v$ compare with the respective $Press_p/n$ and SSE_p/n. (Note here n is the sample size of the training data, i.e., 183)? Which model among the two has a smaller $MSPE_v$?

(c) Based on both internal and external validation, which model you would choose as the final model? Fit the final model using the entire data set (training and validation combined). Write down the fitted regression function and report the R summary and anova outputs.

7. **Model diagnostics: Outlying and influential cases.** Conduct model diagnostics for the final model from the previous problem (fitted on the entire data set).

   (a) Draw residuals vs. fitted values plot and residuals Q-Q plot and comment on these plots.

   (b) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure at $\alpha = 0.1$.

   (c) Obtain the leverage values and identify any outlying X observations. Draw residuals vs. leverage values plot.

   (d) Draw an influence index plot using Cook's distance. Are there any influential cases according to this measure?

   (e) Calculate the average absolute percent difference in the fitted values with and without the most influential case identified from the previous question. What does this measure indicate the influence of this case?

HW8_Question34567

Wookyeong Song (most of them from Yan-Yu Chen)

2022/11/25

3. 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, na.strings = c("NA",""))
diabetes$frame <- as.factor(diabetes$frame)
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` is patient IDs and thus is not a meaningful predictor. The variables `bp.2s`, `bp.2d` have many missing values.

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

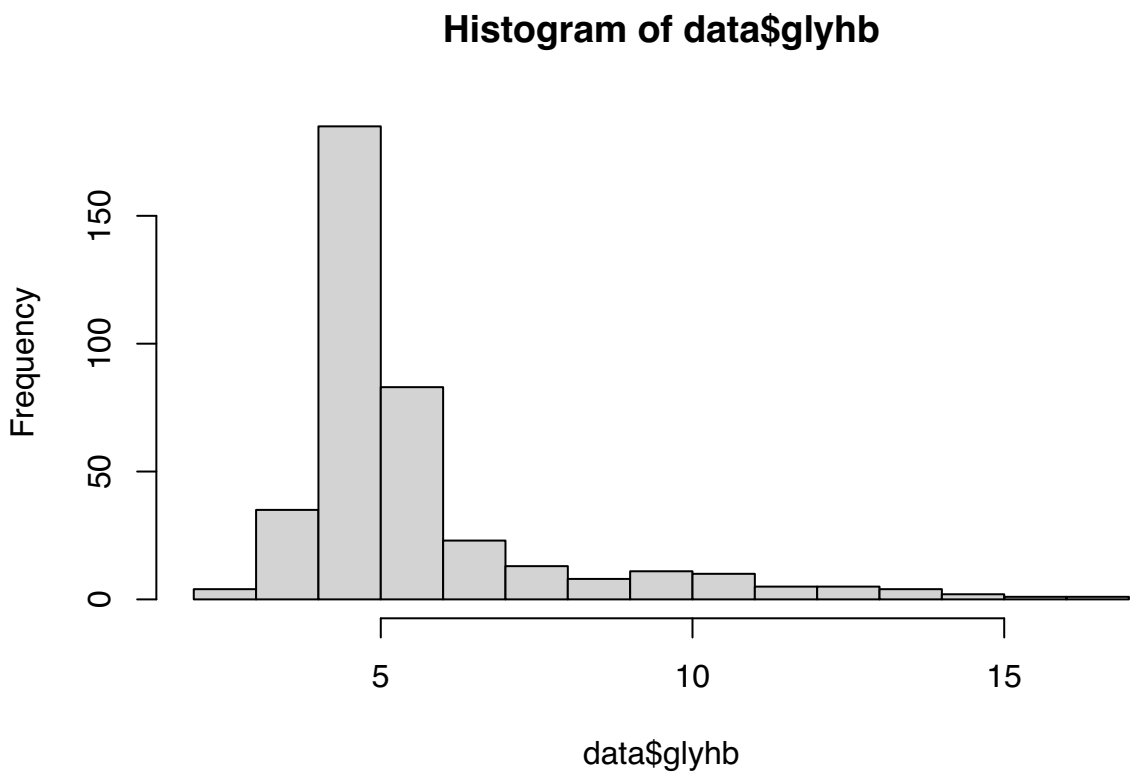
(c) Which of the (remaining) variables are quantitative variables and which are qualitative variables? Draw histogram for `glyhb` and comment on its distribution. Draw histograms for the rest quantitative variables and draw pie charts for qualitative variables.

```
data$location<-as.factor(data$location)
data$gender<-as.factor(data$gender)
sapply(data, class)
```

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

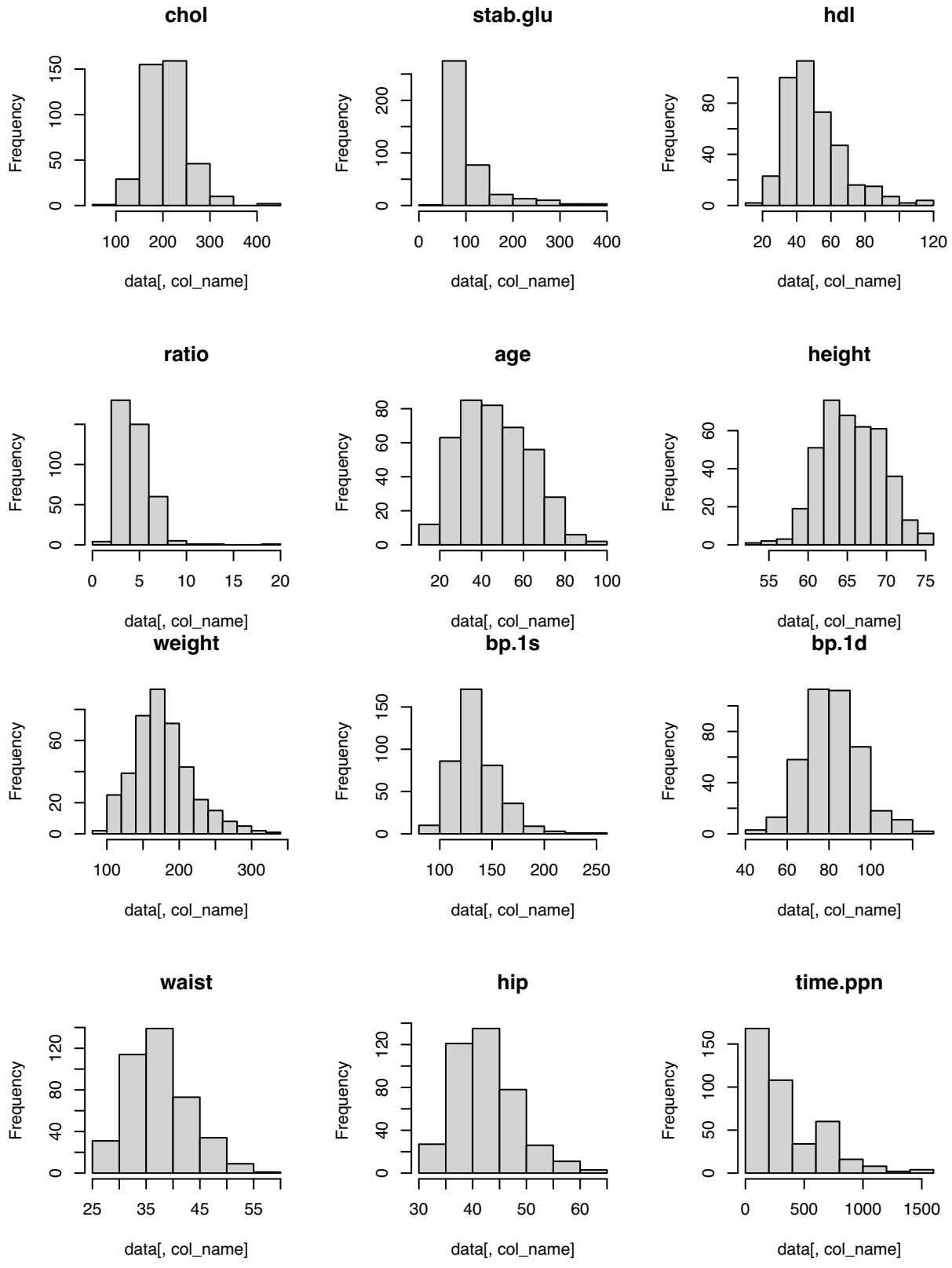
`frame`, `location`, and `gender` are qualitative variables, and the others are quantitative.

```
hist(data$glyhb)
```



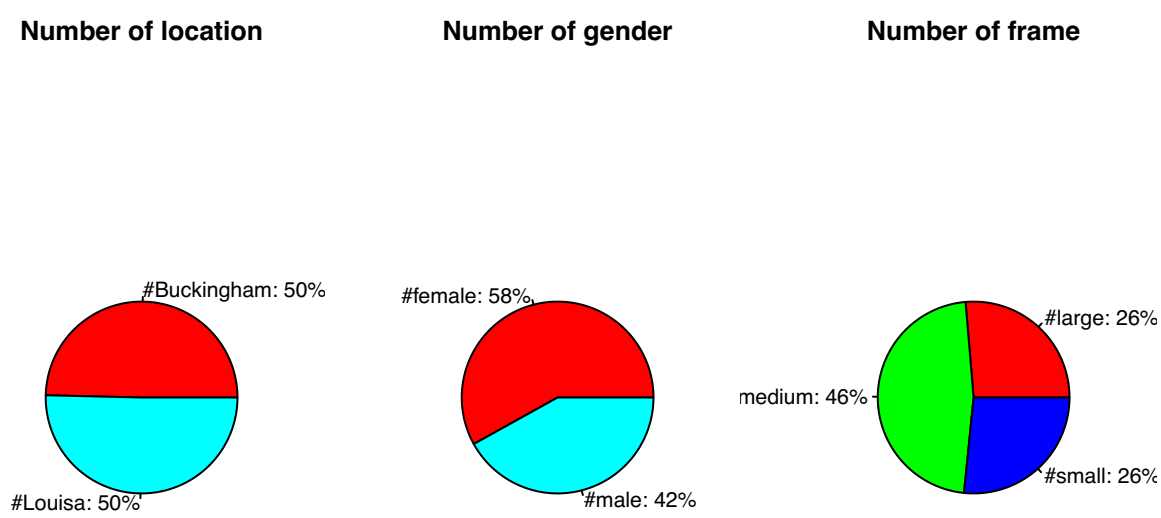
`glyhb` is severely right-skewed.

```
quantitative<-c('chol', 'stab.glu', 'hdl', 'ratio', 'age', 'height', 'weight', 'bp.1s', 'bp.1d', 'waist',
par(mfrow=c(2,3))
for (col_name in quantitative){
  hist(data[,col_name], main=col_name)
}
```



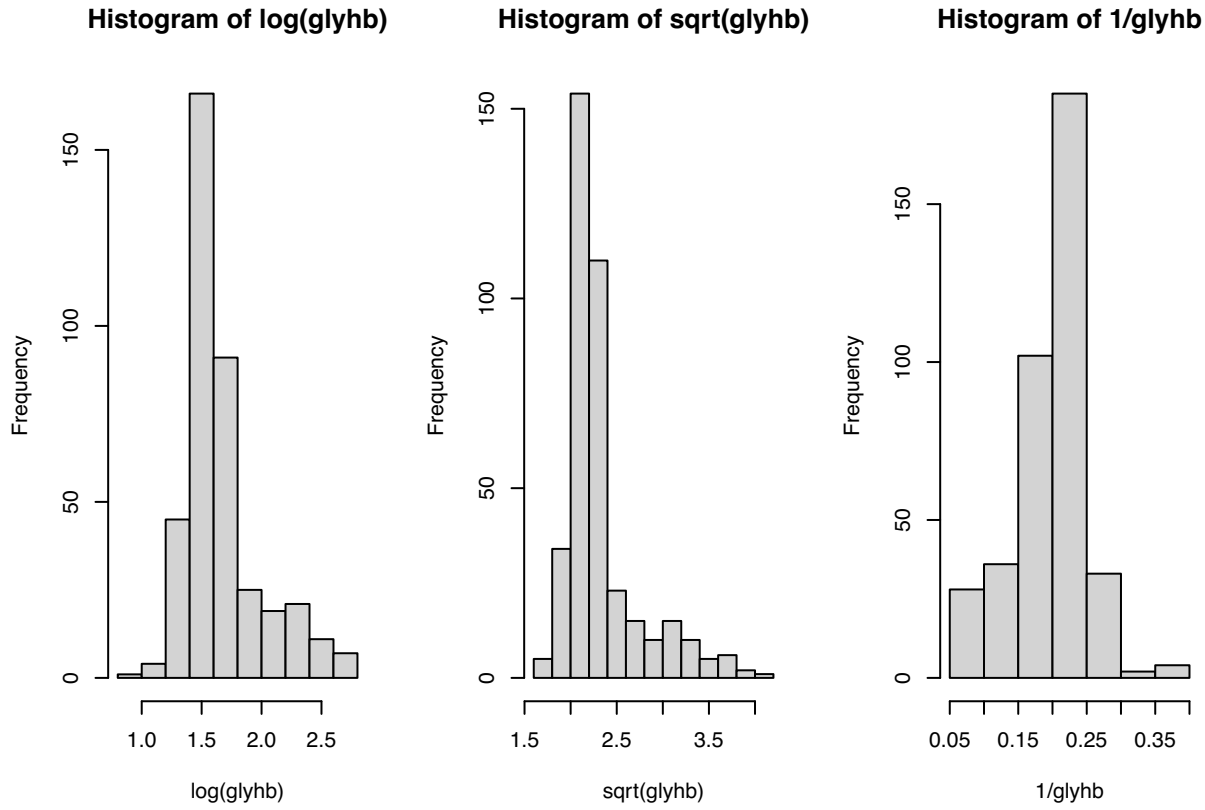

```
qualitative<-c('location', 'gender', 'frame')
par(mfrow=c(1,3))
for (col_name in qualitative){
  lbls <- names(table(data[,col_name]))
```

```
  pct <- round(100*table(data[,col_name])/nrow(data))
  lab <- paste("#",lbls,": ", pct, "%", sep='')
  pie(table(data[,col_name]),labels=lab,col=rainbow(length(lab)), main=paste('Number of', col_name))
}
```



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

```
par(mfrow=c(1,3))
with(data,{
  hist(log(glyhb))
  hist(sqrt(glyhb))
  hist(1/glyhb)
})
```



The third one, $\frac{1}{\text{glyhb}}$ appears to the most normal like.

```
data$glyhb<-1/data$glyhb
```

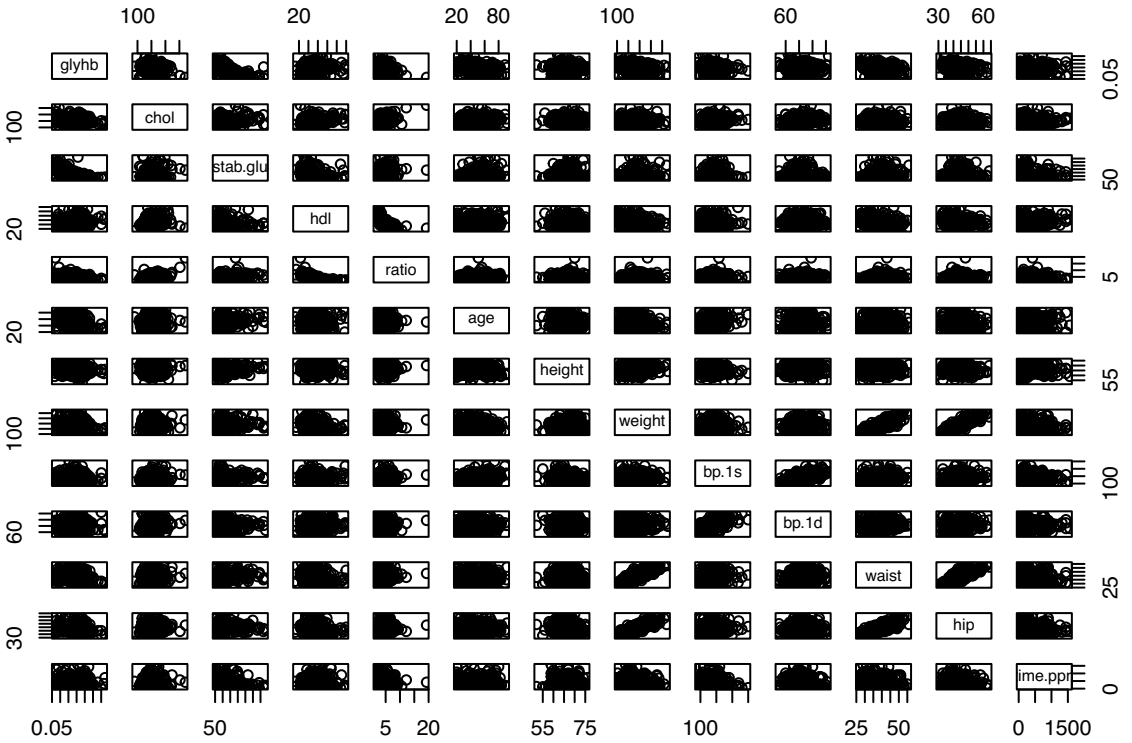
(e) Draw scatterplot matrix and obtain the pairwise correlation matrix for all quantitative variables. Do you observe nonlinearity?

```
cor(data[,c("glyhb",quantitative)], use = "pairwise.complete.obs")
```

```
##                glyhb        chol    stab.glu         hdl       ratio
## glyhb      1.00000000 -0.24839121 -0.65137623  0.16375215 -0.33555595
## chol      -0.24839121  1.00000000  0.15009181  0.18658089  0.47552150
## stab.glu  -0.65137623  0.15009181  1.00000000 -0.16189930  0.28034883
## hdl        0.16375215  0.18658089 -0.16189930  1.00000000 -0.68690710
## ratio     -0.33555595  0.47552150  0.28034883 -0.68690710  1.00000000
## age       -0.40312789  0.23311911  0.28925848  0.03808771  0.14851006
## height    -0.05330161 -0.05885800  0.09066898 -0.10141870  0.09433518
## weight    -0.22015373  0.06688851  0.18545340 -0.29098250  0.28164907
## bp.1s     -0.24456562  0.20334442  0.16646712  0.01980412  0.11938628
## bp.1d     -0.04781292  0.17160549  0.02201413  0.06573191  0.04819273
## waist     -0.29703122  0.12448920  0.21844606 -0.26836902  0.30416249
## hip       -0.20235099  0.07940240  0.13350183 -0.21006001  0.19462190
## time.ppn  -0.03316139  0.01436263 -0.06176748  0.08169882 -0.05306162
##                   age       height       weight       bp.1s       bp.1d
## glyhb    -0.403127890 -0.053301606 -0.22015373 -0.24456562 -0.04781292
## chol      0.233119114 -0.058858000  0.06688851  0.20334442  0.17160549
## stab.glu  0.289258476  0.090668981  0.18545340  0.16646712  0.02201413
```

```
## hdl       0.038087711 -0.101418696 -0.29098250  0.01980412  0.06573191
## ratio     0.148510057  0.094335177  0.28164907  0.11938628  0.04819273
## age       1.000000000 -0.090492934 -0.05596967  0.44304121  0.05822719
## height   -0.090492934  1.000000000  0.25125145 -0.04782717  0.03859780
## weight   -0.055969674  0.251251452  1.00000000  0.09087303  0.17595615
## bp.1s     0.443041214 -0.047827169  0.09087303  1.00000000  0.59655663
## bp.1d     0.058227187  0.038597804  0.17595615  0.59655663  1.00000000
## waist     0.149645490  0.051093923  0.84985462  0.19648887  0.16710958
## hip       0.008819445 -0.107832375  0.82911496  0.13665508  0.14580516
## time.ppn -0.034601537 -0.008357438 -0.06072703 -0.08945769 -0.09319329
##                waist          hip     time.ppn
## glyhb    -0.29703122 -0.202350994 -0.033161389
## chol      0.12448920  0.079402397  0.014362625
## stab.glu  0.21844606  0.133501829 -0.061767478
## hdl      -0.26836902 -0.210060006  0.081698821
## ratio     0.30416249  0.194621897 -0.053061619
## age       0.14964549  0.008819445 -0.034601537
## height    0.05109392 -0.107832375 -0.008357438
## weight    0.84985462  0.829114962 -0.060727026
## bp.1s     0.19648887  0.136655081 -0.089457686
## bp.1d     0.16710958  0.145805158 -0.093193288
## waist     1.00000000  0.837079941 -0.062063408
## hip       0.83707994  1.000000000 -0.085851819
## time.ppn -0.06206341 -0.085851819  1.000000000
```

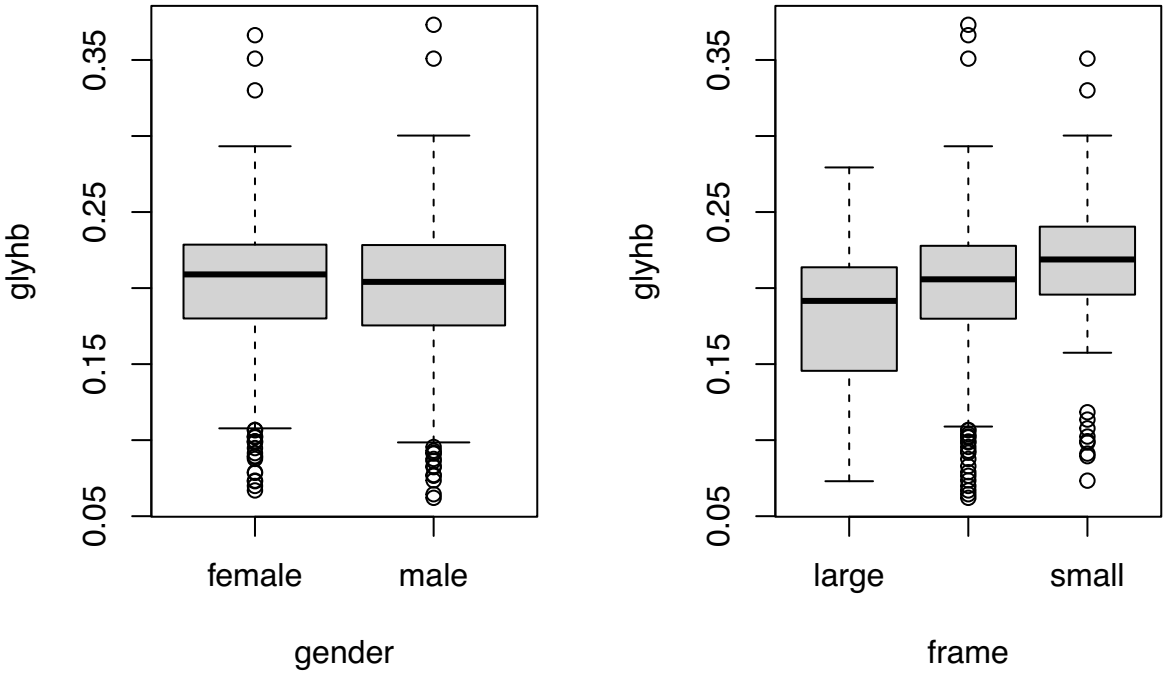
```
pairs(data[,c("glyhb",quantitative)])
```



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.

(f) 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.

```
par(mfrow=c(1,2))
with(data,{
  boxplot(glyhb~gender)
  boxplot(glyhb~frame)
})
```



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

(g) Split data into two equal halves: a training data set and a validation data set. Set seed for random number generator using `set.seed(10)`, so the results are reproducible.

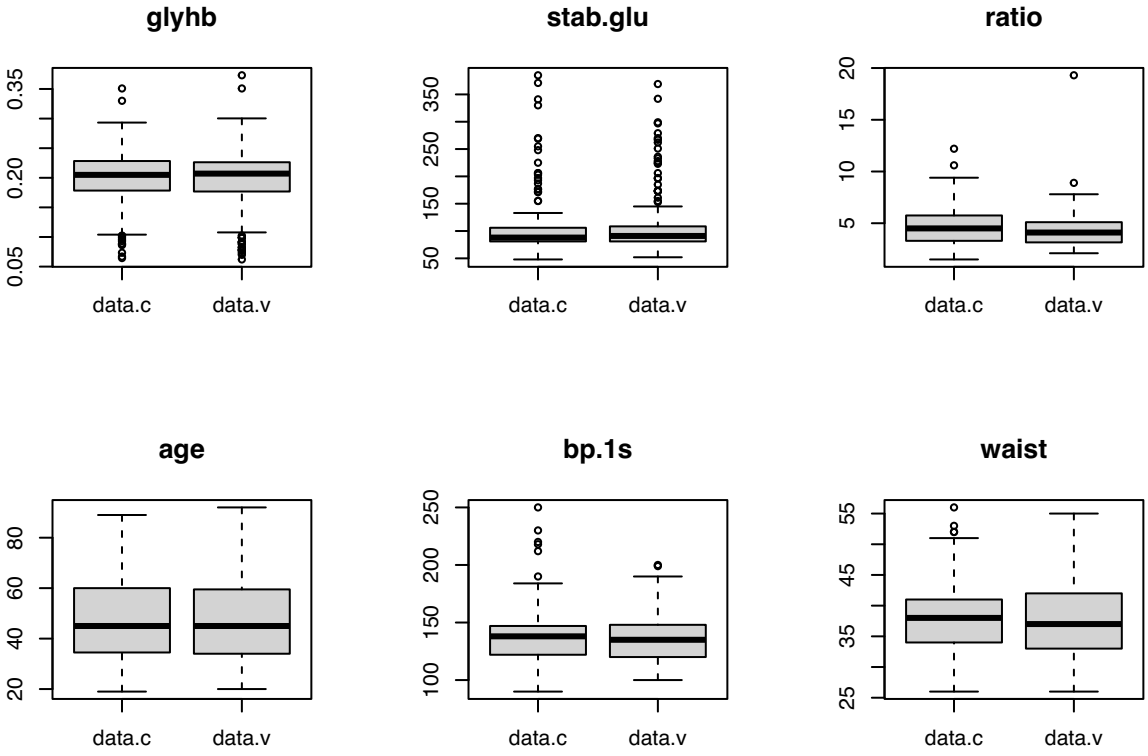
```
data.s <- data[complete.cases(data),]
```

(Of course you can still continue fitting models with the original data, and the default for `lm` is `na.action=na.omit`. However when you calculate AIC or BIC by yourself, n should be taken as the number of complete cases (observations without missing).)

```
set.seed(10)
n.s <- nrow(data.s) ## number of cases in data.s (366)
index.s <- sample(1: n.s, size=n.s/2, replace=FALSE)
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.
```

(h) 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))
for (col_name in c('glyhb', 'stab.glu', 'ratio',
                   'age', 'bp.1s', 'waist')){
  boxplot(data.c[, col_name],data.v[, col_name],main=col_name,names=c('data.c','data.v'))
}
```



Yes, they have similar distributions.

4. 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 (further) transformation of the response variable is still needed?

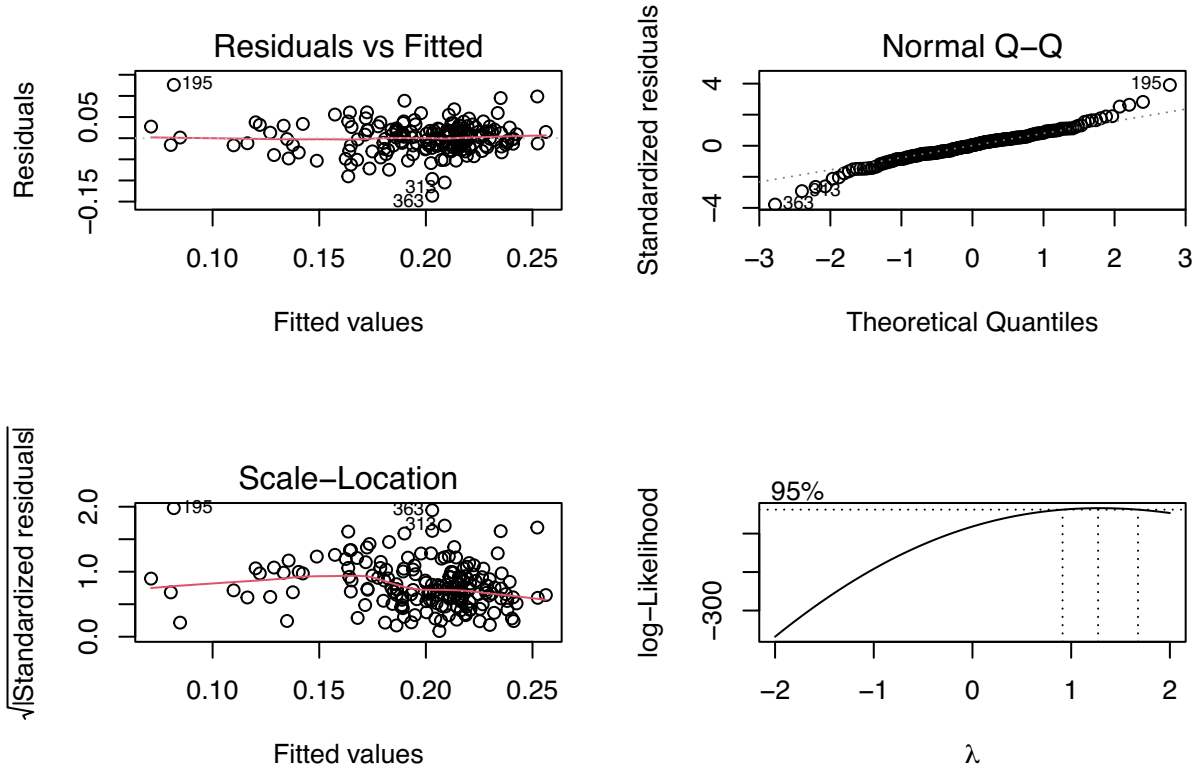
```
fit1<-lm(glyhb~.,data=data.c)
length(fit1$coef) #17 regression coefficients
```

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

```
anova(fit1)['Residuals',3] #MSE
```

```
## [1] 0.001375587
```

```
par(mfrow=c(2,2))
plot(fit1,which=1:3)
MASS::boxcox(fit1)
```



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
```

```r
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              1   1          0      0
## 7              1    0        1   1     1              0   1          0      0
## 8              1    0        1   1     1              1   1          0      0
## 9              1    0        1   1     1              1   1          1      0
## 10             1    0        1   1     1              1   1          1      1
## 11             1    0        1   1     1              1   1          1      1
## 12             1    0        1   1     1              1   1          1      1
## 13             1    0        1   1     1              1   1          1      1
## 14             1    0        1   1     1              1   1          1      1
## 15             1    0        1   1     1              1   1          1      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 0.4301895
## 1         0           0          0     0     0     0   0        0 0.2906079
## 2         0           0          0     0     0     0   0        0 0.2626576
## 3         0           0          0     0     0     1   0        0 0.2505603
## 4         0           0          0     0     0     1   0        0 0.2449681
## 5         0           0          0     0     0     1   0        1 0.2395285
## 6         0           0          0     0     0     1   0        1 0.2374648
## 7         0           0          0     0     1     1   0        1 0.2357689
## 8         0           0          0     0     1     1   0        1 0.2339016
## 9         0           0          0     0     1     0   1        1 0.2324666
## 10        0           0          0     0     1     0   1        1 0.2304848
## 11        0           0          0     1     1     0   1        1 0.2292228
## 12        0           0          0     1     1     1   1        1 0.2288829
## 13        1           0          0     1     1     1   1        1 0.2285546
## 14        1           0          1     1     1     1   1        1 0.2284851
## 15        1           1          1     1     1     1   1        1 0.2284015
## 16        1           1          1     1     1     1   1        1 0.2283475
##            R^2     R^2_a         Cp       aic       bic
## none 0.0000000 0.0000000 129.830539 -1102.492 -1105.702
## 1    0.3244654 0.3207332  32.260913 -1175.484 -1169.065
## 2    0.3894375 0.3826534  13.942102 -1191.989 -1182.361
## 3    0.4175582 0.4077966   7.147863 -1198.618 -1185.780
## 4    0.4305576 0.4177611   5.082557 -1200.749 -1184.701
## 5    0.4432024 0.4274736   3.128130 -1202.858 -1183.601
## 6    0.4479994 0.4291812   3.627934 -1202.442 -1179.975
```

```
## 7    0.4519416 0.4300193   4.395083 -1201.753 -1176.077
## 8    0.4562824 0.4312839   5.037580 -1201.208 -1172.323
## 9    0.4596182 0.4315059   5.994382 -1200.335 -1168.240
## 10   0.4642250 0.4330753   6.553692 -1199.901 -1164.597
## 11   0.4671584 0.4328820   7.636323 -1198.906 -1160.392
## 12   0.4679487 0.4303921   9.389185 -1197.178 -1155.454
## 13   0.4687117 0.4278434  11.150557 -1195.440 -1150.508
## 14   0.4688733 0.4246127  13.100028 -1193.496 -1145.354
## 15   0.4690676 0.4213791  15.039243 -1191.563 -1140.211
## 16   0.4691931 0.4180310  17.000000 -1189.606 -1135.045
```

Best model:

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

R^2_a: Model 10 ((Intercept), stab.glu, hdl, ratio, locationLouisa, age, gendermale, height, bp.1d, hip, time.ppn)

C_p, AIC: Model 5 ((Intercept), stab.glu, ratio, age, waist, time.ppn)

BIC: Model 3 ((Intercept), stab.glu, age, waist)

For the model with the smallest C_p statistic (Model 5), its C_p value is 3.128 which is 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?

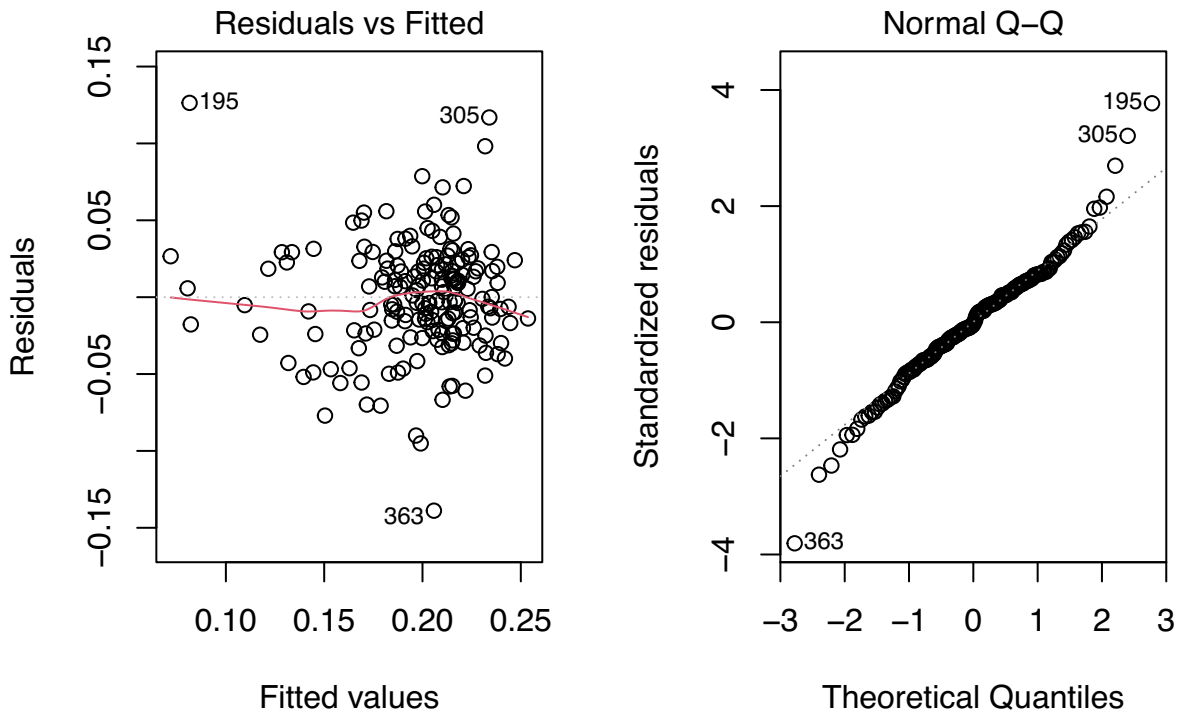
```
library(MASS)
step.f<-stepAIC(fit0,scope=list(upper=fit1, lower=~1), trace = 0, direction="both", k=2)
step.f$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## glyhb ~ 1
##
## Final Model:
## glyhb ~ stab.glu + age + waist + ratio + time.ppn
##
##
##         Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                                 182  0.4301895 -1105.702
## 2 + stab.glu  1 0.139581597       181  0.2906079 -1175.484
## 3      + age  1 0.027950303       180  0.2626576 -1191.989
## 4    + waist  1 0.012097245       179  0.2505603 -1198.618
## 5    + ratio  1 0.005592183       178  0.2449681 -1200.749
## 6 + time.ppn  1 0.005439661       177  0.2395285 -1202.858
```

The final model contains (Intercept), stab.glu, age, waist, ratio, time.ppn. It is the best model according to AIC_p criterion identified in part (b).

(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?

```
par(mfrow=c(1,2))
plot(step.f,which=1:2)
```



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

5. 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] 0.001203833
```

Relative to the sample size, there are too many predictors (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?

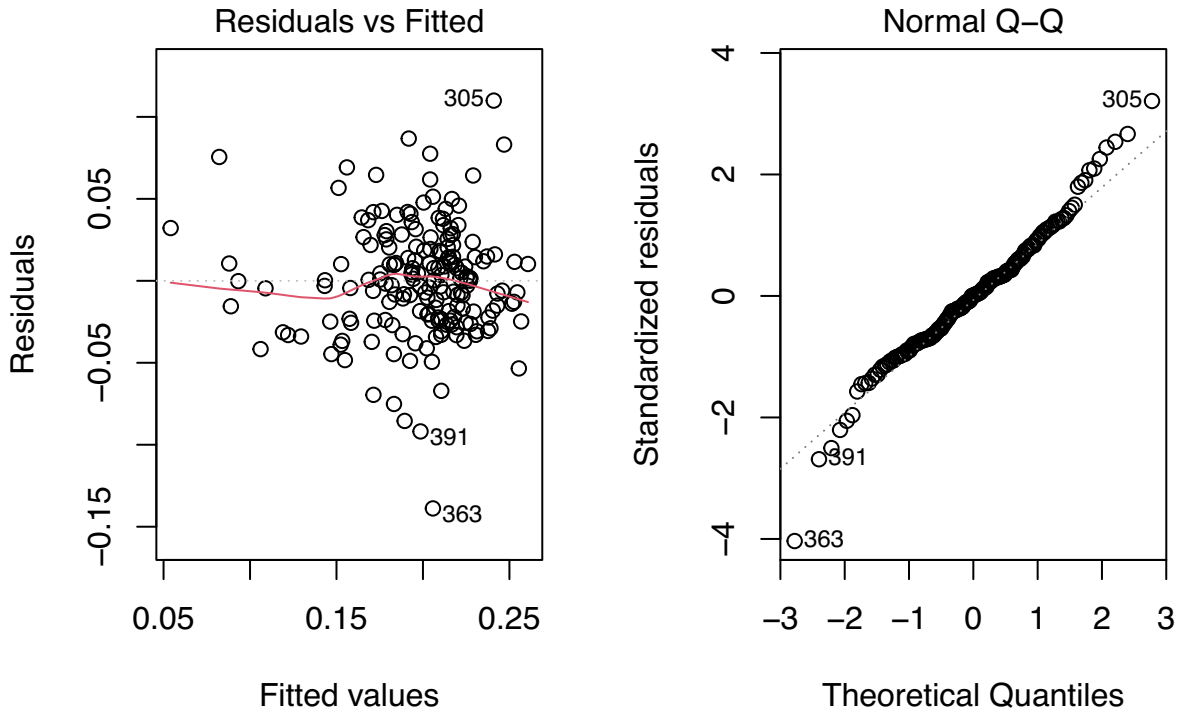
```
step.f2<-stepAIC(fit0,scope=list(upper=fit2, lower=~1), trace=0, direction="both",k=2)
step.f2$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## glyhb ~ 1
##
## Final Model:
## glyhb ~ stab.glu + age + waist + time.ppn + location + ratio +
##     stab.glu:age + stab.glu:time.ppn + location:ratio
##
##
##                   Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                                          182  0.4301895 -1105.702
## 2           + stab.glu  1 0.139581597       181  0.2906079 -1175.484
## 3                + age  1 0.027950303       180  0.2626576 -1191.989
## 4       + stab.glu:age  1 0.019773430       179  0.2428841 -1204.312
## 5              + waist  1 0.008715929       178  0.2341682 -1209.000
## 6           + time.ppn  1 0.007251709       177  0.2269165 -1212.757
## 7  + stab.glu:time.ppn  1 0.007044333       176  0.2198722 -1216.528
## 8           + location  1 0.003337334       175  0.2165348 -1217.327
## 9          + age:waist  1 0.002517431       174  0.2140174 -1217.467
## 10      + location:age  1 0.002460912       173  0.2115565 -1217.583
## 11             + ratio  1 0.002552167       172  0.2090043 -1217.804
## 12    + ratio:location  1 0.002971100       171  0.2060332 -1218.424
## 13      - age:location  1 0.001461041       172  0.2074943 -1219.131
## 14         - age:waist  1 0.001660360       173  0.2091546 -1219.673
```

The final model contains (Intercept), stab.glu, age, waist, time.ppn, locationLouisa, ratio, stab.glu:age, stab.glu:time.ppn, locationLouisa:ratio. Its AIC is -1219.672677, which is smaller than that of Model.fs1.

(c) Comment on the residual vs. fitted value plot and the residual Q-Q plot of Modelfs2. Does this model appear to be adequate?

```
par(mfrow=c(1,2))
plot(step.f2,which=1:2)
```



The residual vs. fitted plot still shows non-constant error variance. Otherwise, the model seems reasonable.

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

```
step.f3<-stepAIC(fit0,scope=list(upper=fit2, lower=~1), trace=0, 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 + time.ppn + location + ratio +
##     stab.glu:age + stab.glu:time.ppn + age:waist + age:location +
##     location:ratio
##
##
##                    Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                                           182  0.4301895 -1105.702
## 2           + stab.glu  1 0.139581597       181  0.2906079 -1175.484
## 3               + age  1 0.027950303       180  0.2626576 -1191.989
## 4       + stab.glu:age  1 0.019773430       179  0.2428841 -1204.312
## 5             + waist  1 0.008715929       178  0.2341682 -1209.000
## 6           + time.ppn  1 0.007251709       177  0.2269165 -1212.757
## 7  + stab.glu:time.ppn  1 0.007044333       176  0.2198722 -1216.528
## 8           + location  1 0.003337334       175  0.2165348 -1217.327
## 9          + age:waist  1 0.002517431       174  0.2140174 -1217.467
## 10      + location:age  1 0.002460912       173  0.2115565 -1217.583
## 11            + ratio  1 0.002552167       172  0.2090043 -1217.804
```

```
## 12    + ratio:location  1 0.002971100       171  0.2060332 -1218.424
```

The final model contains (Intercept), stab.glu, age, waist, time.ppn, locationLouisa, ratio, stab.glu:age, stab.glu:time.ppn, age:waist, age:locationLouisa, locationLouisa:ratio. Its AIC is -1218.4243336.

6. We now consider validation of the two models `fs1` and `fs2` selected by the forward stepwise procedure.

(a) Internal validation of Models fs1 and fs2

For this purpose, we need to compute C_p and $Press_p$ for these models. For C_p, we need an unbiased estimator of the error variance σ^2. The largest model we have considered so far is Model 2. However, this model has a very large number of regression coefficients (relative to the sample size), making its parameter estimation unreliable due to large sampling variability. Therefore, we decided to use a smaller model consisting of all predictors identified by Model fs1 (the forward stepwise selected first-order model), as well all the 2-way interaction terms among these predictors. Denote this model by Model 3. Note that, Model fs2 is also a sub-model of Model 3. How many regression coefficients are there in Model 3 ? What is MSE from Model 3? Calculate SSE_p, MSE_p, C_p and $Press_p$ for Models fs1 and fs2 and briefly comment on the results, e.g., does it appear to be substantial model bias in these two models? Should overfitting be a concern?

```
data.cc<-data.c[, c("glyhb",names(step.f$coefficients)[-1])]
fit3<- lm (glyhb~.^2, data = data.cc)
length(fit3$coef)  #number of coefficents in Model 3
```

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

```
mse3<-anova(fit3)["Residuals",3]
mse3#MSE for Model 3
```

```
## [1] 0.001256631
```

```
sse.fs1<-anova(step.f)["Residuals",2] #first order selected
sse.fs1
```

```
## [1] 0.2395285
```

```
sse.fs2<-anova(step.f2)["Residuals",2] #second order selected
sse.fs2
```

```
## [1] 0.2091546
```

```
mse.fs1<-anova(step.f)["Residuals",3] #MSE for Model fs1
mse.fs1
```

```
## [1] 0.001353268
```

```
mse.fs2<-anova(step.f2)["Residuals",3] #MSE for Model fs2
mse.fs2
```

```
## [1] 0.001208986
```

```
p.fs1<-length(step.f$coefficients) #5
p.fs1
```

```
## [1] 6
```

```
p.fs2<-length(step.f2$coefficients) #7
p.fs2
```

```
## [1] 10
```

```
cp.fs1<-sse.fs1/mse3-(n-2*p.fs1) #C_p for Model fs1
cp.fs1
```

```
## [1] 19.61159
```

```
cp.fs2<-sse.fs2/mse3-(n-2*p.fs2) #C_p for Model fs2
cp.fs2
```

```
## [1] 3.440732
```

```
press.fs1<-sum(step.f$residuals^2/(1-influence(step.f)$hat)^2)
press.fs1
```

```
## [1] 0.2614933
```

```
press.fs2<-sum(step.f2$residuals^2/(1-influence(step.f2)$hat)^2)
press.fs2
```

```
## [1] 0.2411787
```

For both Model fs1 and Model fs2, $C_p \approx p$ and $Press_p$ and SSE_p are reasonably close, supporting their validity: little bias and not much overfitting.

(b) External validation using the validation set

We now fit Models fs1 and fs2 on the validation data set. Compare the fitted regression coefficients from the training data and those from the validation data. Are the two sets of estimated regression coefficients having the same sign? Are their values similar? How about the two sets of standard errors? Does it appear that Models fs1 and fs2 have consistent estimates on the training data and validation data? Calculate the mean squared prediction error (MSPE) using the validation data for each of the two models. How do these $MSPE_v$ compare with the respective $Press_p/n$ and SSE_p/n. (Note here n is the sample size of the training data, i.e., 183)? Which model among the two has a smaller $MSPE_v$?

```
fit.fs1.v<-lm(step.f,data=data.v) #Model fs1 on validation data
summary(step.f) #summary on training data
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + ratio + time.ppn,
##     data = data.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.138871 -0.021429 -0.001112  0.021780  0.126366
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.408e-01  1.977e-02  17.236  < 2e-16 ***
## stab.glu    -4.284e-04  5.370e-05  -7.978 1.79e-13 ***
## age         -6.839e-04  1.735e-04  -3.942 0.000116 ***
## waist       -1.105e-03  5.229e-04  -2.114 0.035919 *
## ratio       -3.614e-03  1.718e-03  -2.104 0.036822 *
## time.ppn    -1.878e-05  9.365e-06  -2.005 0.046497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03679 on 177 degrees of freedom
## Multiple R-squared:  0.4432, Adjusted R-squared:  0.4275
## F-statistic: 28.18 on 5 and 177 DF,  p-value: < 2.2e-16
```

```
summary(fit.fs1.v) #summary on validation data
```

```
##
## Call:
## lm(formula = step.f, data = data.v)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.154135 -0.019784 -0.001598  0.018649  0.148175
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.430e-01  1.830e-02  18.744  < 2e-16 ***
## stab.glu    -5.637e-04  5.542e-05 -10.172  < 2e-16 ***
## age         -6.193e-04  1.739e-04  -3.562 0.000473 ***
## waist       -9.838e-04  4.791e-04  -2.054 0.041494 *
## ratio       -3.864e-03  1.651e-03  -2.340 0.020407 *
## time.ppn    -8.609e-06  8.168e-06  -1.054 0.293359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03581 on 177 degrees of freedom
## Multiple R-squared:  0.5707, Adjusted R-squared:  0.5585
## F-statistic: 47.05 on 5 and 177 DF,  p-value: < 2.2e-16
```

```
#percent change in parameter estimation
round(abs(coef(step.f)-coef(fit.fs1.v))/abs(coef(step.f))*100,3)
```

```
## (Intercept)    stab.glu         age       waist       ratio    time.ppn
##       0.657      31.582       9.442      11.006       6.901      54.150
```

```
sd.fs1<- summary(step.f)$coefficients[,"Std. Error"]
sd.fs1.v<- summary(fit.fs1.v)$coefficients[,"Std. Error"]
#percent change in standard errors
round(abs(sd.fs1-sd.fs1.v)/sd.fs1*100,3)
```

```
## (Intercept)    stab.glu         age       waist       ratio    time.ppn
##       7.442       3.203       0.238       8.386       3.889      12.777
```

Consistency for Model `fs1`: reasonable. Signs for parameter estimates are all the same, but percent change can be as big as 50%.

```
##mean squared prediction error
pred.fs1<-predict.lm(step.f,data.v[,-5])
mspe.fs1<-mean((pred.fs1-data.v[,5])^2)
mspe.fs1
```

```
## [1] 0.00130831
```

```
press.fs1/n
```

```
## [1] 0.001428925
```

```
mse.fs1
```

```
## [1] 0.001353268
```

```
fit.fs2.v<-lm(step.f2,data=data.v) #Model fs1 on validation data
summary(step.f2) #summary on training data
```

```
##
## Call:
## lm(formula = glyhb ~ stab.glu + age + waist + time.ppn + location +
##     ratio + stab.glu:age + stab.glu:time.ppn + location:ratio,
##     data = data.c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.138830 -0.023252  0.000347  0.018983  0.109920
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.883e-01  2.841e-02  13.667  < 2e-16 ***
## stab.glu             -1.248e-03  2.396e-04  -5.209 5.34e-07 ***
## age                  -2.299e-03  4.422e-04  -5.198 5.63e-07 ***
## waist                -1.051e-03  5.001e-04  -2.102 0.037044 *
## time.ppn              3.084e-05  2.442e-05   1.263 0.208301
## locationLouisa        3.885e-02  1.591e-02   2.442 0.015602 *
## ratio                 1.141e-03  2.389e-03   0.478 0.633405
## stab.glu:age          1.625e-05  4.155e-06   3.910 0.000132 ***
## stab.glu:time.ppn    -4.303e-07  2.142e-07  -2.009 0.046115 *
## locationLouisa:ratio -6.270e-03  3.106e-03  -2.019 0.045076 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03477 on 173 degrees of freedom
## Multiple R-squared:  0.5138, Adjusted R-squared:  0.4885
## F-statistic: 20.31 on 9 and 173 DF,  p-value: < 2.2e-16
```

```
summary(fit.fs2.v) #summary on validation data
```

```
##
## Call:
## lm(formula = step.f2, data = data.v)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151212 -0.016775 -0.001678  0.014444  0.146443
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.288e-01  2.775e-02  11.846  < 2e-16 ***
## stab.glu             -4.871e-04  2.190e-04  -2.224  0.02745 *
## age                  -5.829e-04  4.070e-04  -1.432  0.15396
## waist                -9.308e-04  4.761e-04  -1.955  0.05222 .
## time.ppn              4.109e-05  2.055e-05   2.000  0.04706 *
## locationLouisa       -1.589e-02  1.537e-02  -1.034  0.30265
## ratio                -4.975e-03  1.859e-03  -2.676  0.00817 **
## stab.glu:age          5.538e-07  3.587e-06   0.154  0.87751
## stab.glu:time.ppn    -4.482e-07  1.718e-07  -2.608  0.00990 **
## locationLouisa:ratio  5.147e-03  3.320e-03   1.550  0.12287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03527 on 173 degrees of freedom
## Multiple R-squared:  0.5929, Adjusted R-squared:  0.5718
## F-statistic:    28 on 9 and 173 DF,  p-value: < 2.2e-16
```

```
#percent change in parameter estimation
round(abs(coef(step.f2)-coef(fit.fs2.v))/abs(coef(step.f2))*100,3)
```

```
##          (Intercept)             stab.glu                  age
##               15.329               60.975               74.646
##                waist             time.ppn       locationLouisa
##               11.433               33.260              140.892
##                ratio         stab.glu:age    stab.glu:time.ppn
##              535.923               96.591                4.159
## locationLouisa:ratio
##              182.100
```

```
sd.fs2<- summary(step.f2)$coefficients[,"Std. Error"]
sd.fs2.v<- summary(fit.fs2.v)$coefficients[,"Std. Error"]
#percent change in standard errors
round(abs(sd.fs2-sd.fs2.v)/sd.fs2*100,3)
```

```
##          (Intercept)             stab.glu                  age
##                2.308                8.583                7.961
##                waist             time.ppn       locationLouisa
##                4.786               15.851                3.404
##                ratio         stab.glu:age    stab.glu:time.ppn
##               22.162               13.665               19.781
## locationLouisa:ratio
##                6.890
```

Consistency for Model `fs2`: Both sign and magnitude changed.

```
#mean squared prediction error
pred.fs2<-predict.lm(step.f2, data.v[,-5])
mspe.fs2<-mean((pred.fs2-data.v[,5])^2)
mspe.fs2 #larger than mspe.fs1
```

```
## [1] 0.001492486
```

```
press.fs2/n
```

```
## [1] 0.001317916
```

```
mse.fs2
```

```
## [1] 0.001208986
```

For both models, $MSPE_v$ is not much bigger than $Press_p/n$ and SSE_p/n, though $MSPE_v$ is closer to $Press_p/n$ and SSE_p/n in Model fs1. Moreover, Model fs1 has smaller $MSPE_v$.

(c) Based on both internal and external validation, which model you would choose as the final model? Fit the final model using the entire data set (training and validation combined). Write down the fitted regression function and report the R summary and anova outputs.

Model `fs1` is preferred based on smaller $MSPE_v$ and more consistent parameter estimation in training and validation data sets.

```
fit.fs1.final<-lm(step.f, data=data.s) #fit Model fs1 on whole data
summary(fit.fs1.final)
```

```
##
## Call:
## lm(formula = step.f, data = data.s)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.153480 -0.020857 -0.001696  0.020034  0.149250
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.446e-01  1.332e-02  25.862  < 2e-16 ***
## stab.glu    -4.943e-04  3.819e-05 -12.945  < 2e-16 ***
```

```
## age         -6.576e-04  1.223e-04  -5.379 1.35e-07 ***
## waist       -1.116e-03  3.500e-04  -3.187  0.00156 **
## ratio       -3.731e-03  1.175e-03  -3.176  0.00162 **
## time.ppn    -1.359e-05  6.125e-06  -2.218  0.02716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03624 on 360 degrees of freedom
## Multiple R-squared:  0.5073, Adjusted R-squared:  0.5004
## F-statistic: 74.13 on 5 and 360 DF,  p-value: < 2.2e-16
```

```
anova(fit.fs1.final)
```

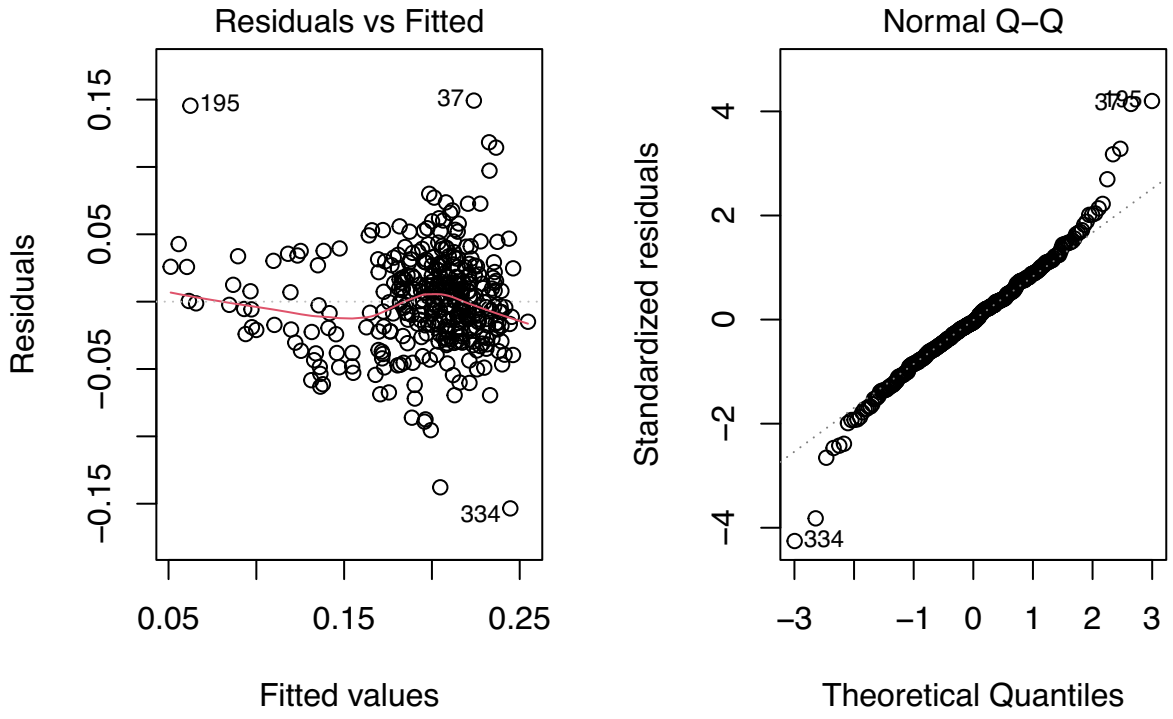
```
## Analysis of Variance Table
##
## Response: glyhb
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## stab.glu    1 0.39753 0.39753 302.7569 < 2.2e-16 ***
## age         1 0.04867 0.04867  37.0664 2.928e-09 ***
## waist       1 0.02125 0.02125  16.1820 7.015e-05 ***
## ratio       1 0.01276 0.01276   9.7147  0.001975 **
## time.ppn    1 0.00646 0.00646   4.9205  0.027163 *
## Residuals 360 0.47269 0.00131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Model diagnostics: Outlying and influential cases

Conduct model diagnostics for the final model from the previous problem (fitted on the entire data set).

(a) Draw residual vs. fitted value plot and residual Q-Q plot and comment on these plots.

```
par(mfrow=c(1,2))
plot(fit.fs1.final,which=1:2)
```



The residual plot shows non-constancy in error variance. The Normal QQ plot shows heavy tails probably due to outliers.

(b) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure at $\alpha = 0.1$.

```
## check outliers in Y
res<-residuals(fit.fs1.final)# residuals of the final model
p <- length(fit.fs1.final$coefficients)
h1 <- influence(fit.fs1.final)$hat
d.res.std<-studres(fit.fs1.final) #studentized deleted residuals
qt(1-0.1/(2*n.s),n.s-1-p) # bonferronis thresh hold
```

```
## [1] 3.675875
```

The studentized deleted residuals are calculated through this equation:

$$t_i = e_i \sqrt{\frac{n-p}{SSE(1-h_{ii}) - e_i^2}}, \text{ where } p = \text{Tr}(H)$$

To identify the outlying Y observations, we use the Bonferroni outlier test procedure at $\alpha = 0.1$. The Bonferroni's threshold is $t(1 - \frac{\alpha}{2n}; n-p) = 3.48823$. The Y observations corresponding to those studentized deleted residuals which are greater than the Bonferroni's threshold can be deemed as significant outlying observations. They are as follows:

```
idx.Y <- as.vector(which(abs(d.res.std)>=qt(1-0.1/(2*n.s),n.s-1-p)))
idx.Y ## outliers
```

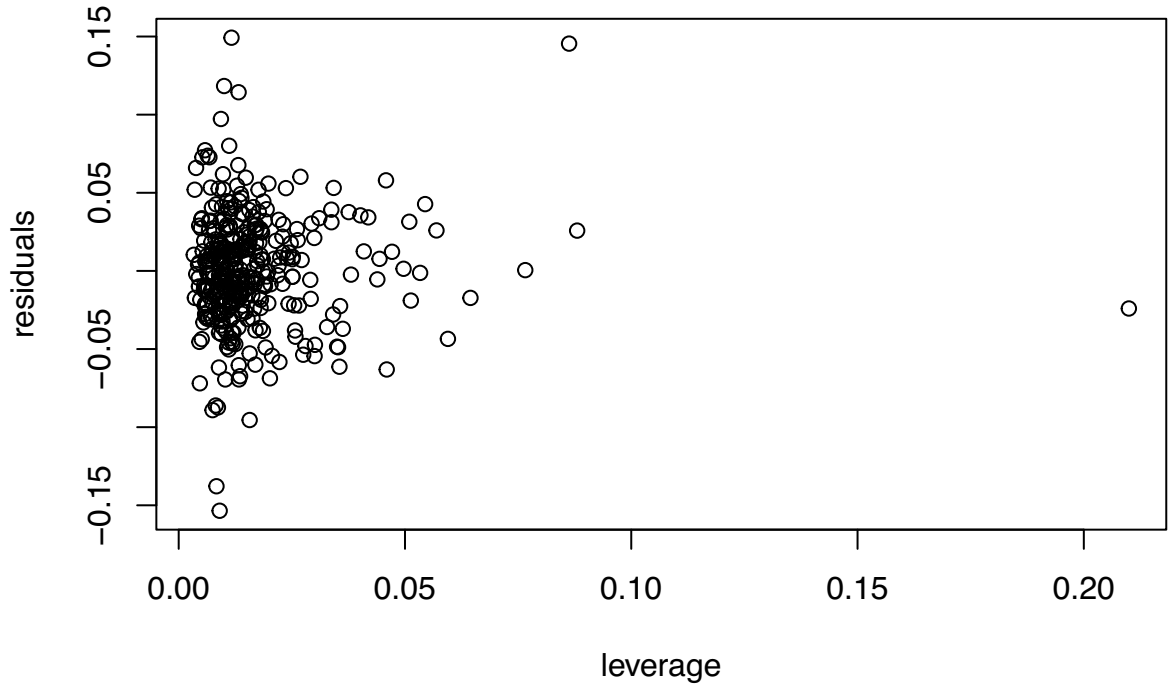
```
## [1]  34 176 303 330
```

(c) Obtain the leverage and identify any outlying X observations. Draw residual vs. leverage plot.

```
idx.X <- as.vector(which(h1>(2*p/n.s)))
idx.X ## two outliers
```

```
##  [1]  16  21  30  42  52  56  59  60  66  81  86  89 118 129 132 135 139 156 176
## [20] 245 268 279 299 326 332 333 336 348 362 363 365
```

```
plot(h1,res,xlab="leverage",ylab="residuals")
```



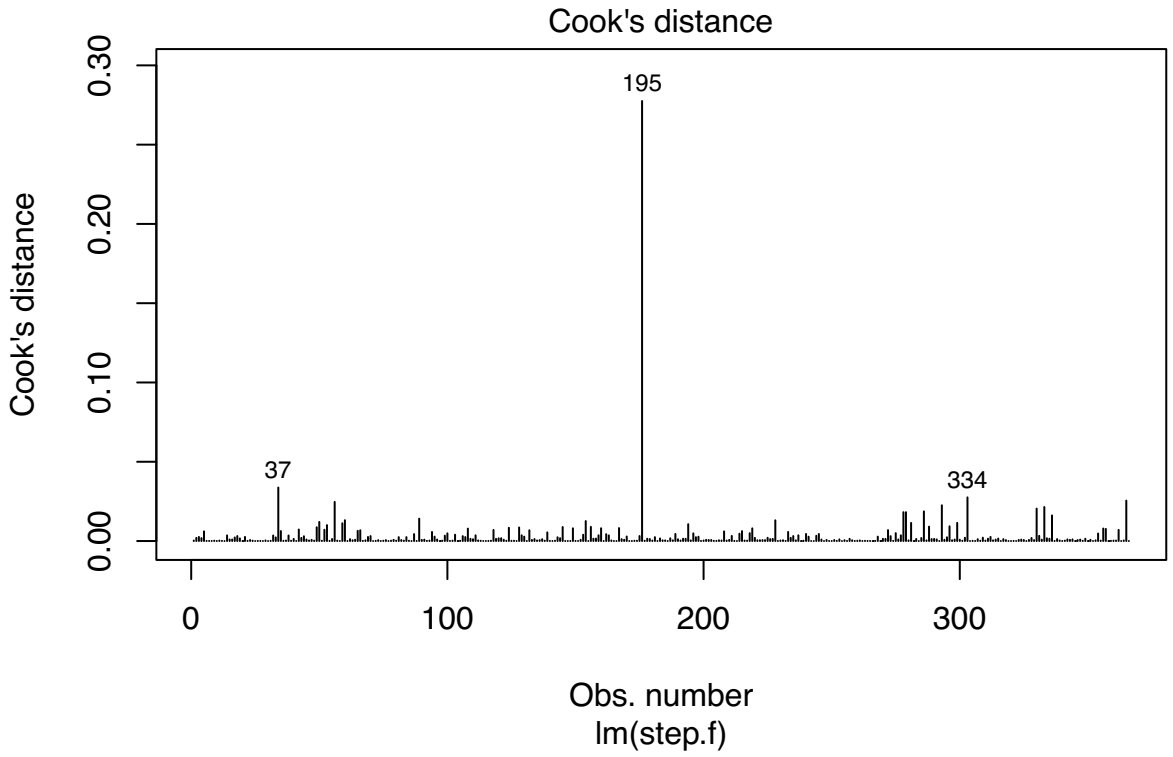

```
idx.X ## outliers
```

```
##  [1]  16  21  30  42  52  56  59  60  66  81  86  89 118 129 132 135 139 156 176
## [20] 245 268 279 299 326 332 333 336 348 362 363 365
```

The leverages are obtained and compared with the value of $\frac{2p}{n} = 0.0327869$. The cases with $h_{ii} > \frac{2p}{n}$ are defined as outlying X observations. There are 31 cases defined as outlying X observations, their indexes are shown above.

(d) Draw an influence index plot using Cook's distance. Are there any influential cases according to this measure?

```
plot(fit.fs1.final, which=4)
```



Case 195 is an influential case according to Cook's distance.

(e) Calculate the average absolute percent difference in the fitted values with and without the most influential case identified from the previous question. What does this measure indicate the influence of this case?

```
fit.fs1.final2<-lm(fit.fs1.final, data=data.s[-195,])
f1<-fitted(fit.fs1.final)
f2<-fitted(fit.fs1.final2)
SUM<-sum(abs((f1[-195]-f2)/f1[-195]))
SUM<-SUM+abs((f1[195]-predict(fit.fs1.final,newdata = data.s[195,]))/f1[195])
per.average<-SUM/n.s
per.average
```

```
##          215
## 0.0004076336
```

The potential influential case identified previously is the 195th case, we fit the model without 195th case and calculate the average absolute difference in the fitted values as 0.041%. For 195th case, the percentage change on the fitted value with or without the case is very small. Therefore, no case have an unduly large influence on prediction and thus all cases may be retained.