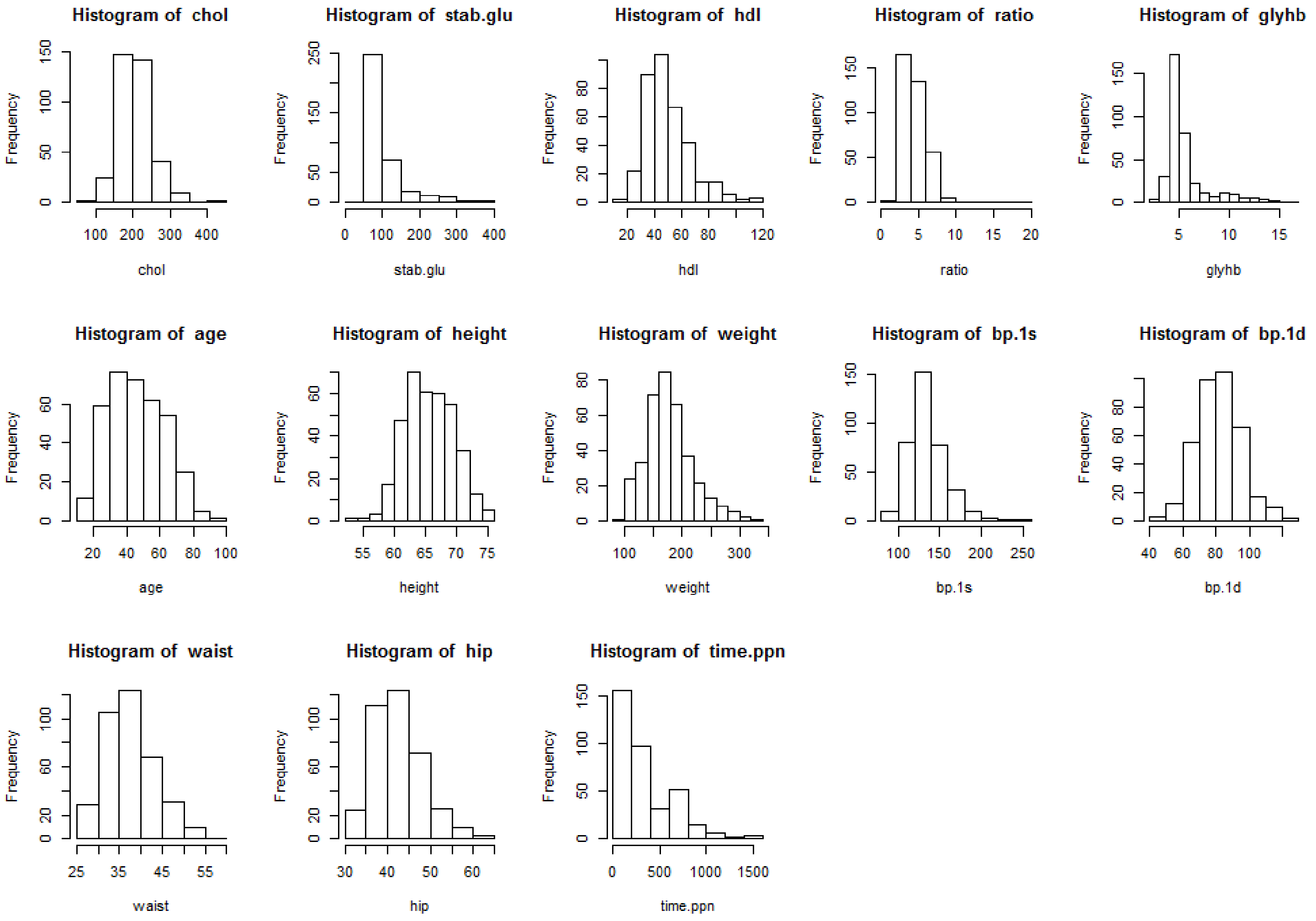
Zack Liu 915026153 11/30/17

Statistics 108, Project 1

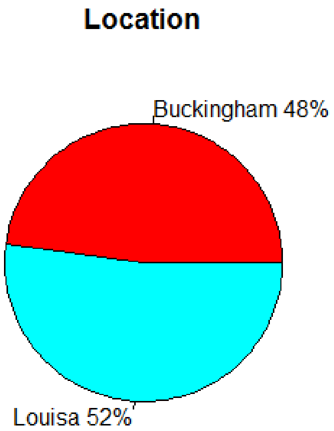
Data exploration and split data for validation later on:

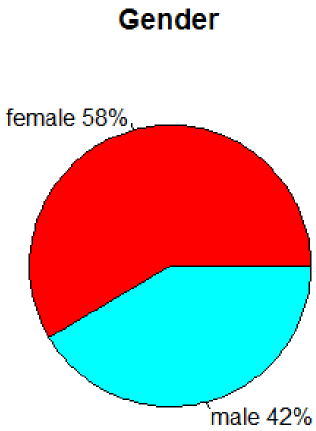
1.

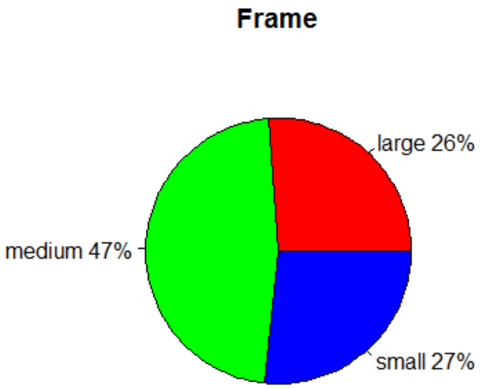
Among all variables, location, gender and frame are qualitative variables. The rest of the variables are quantitative variables.



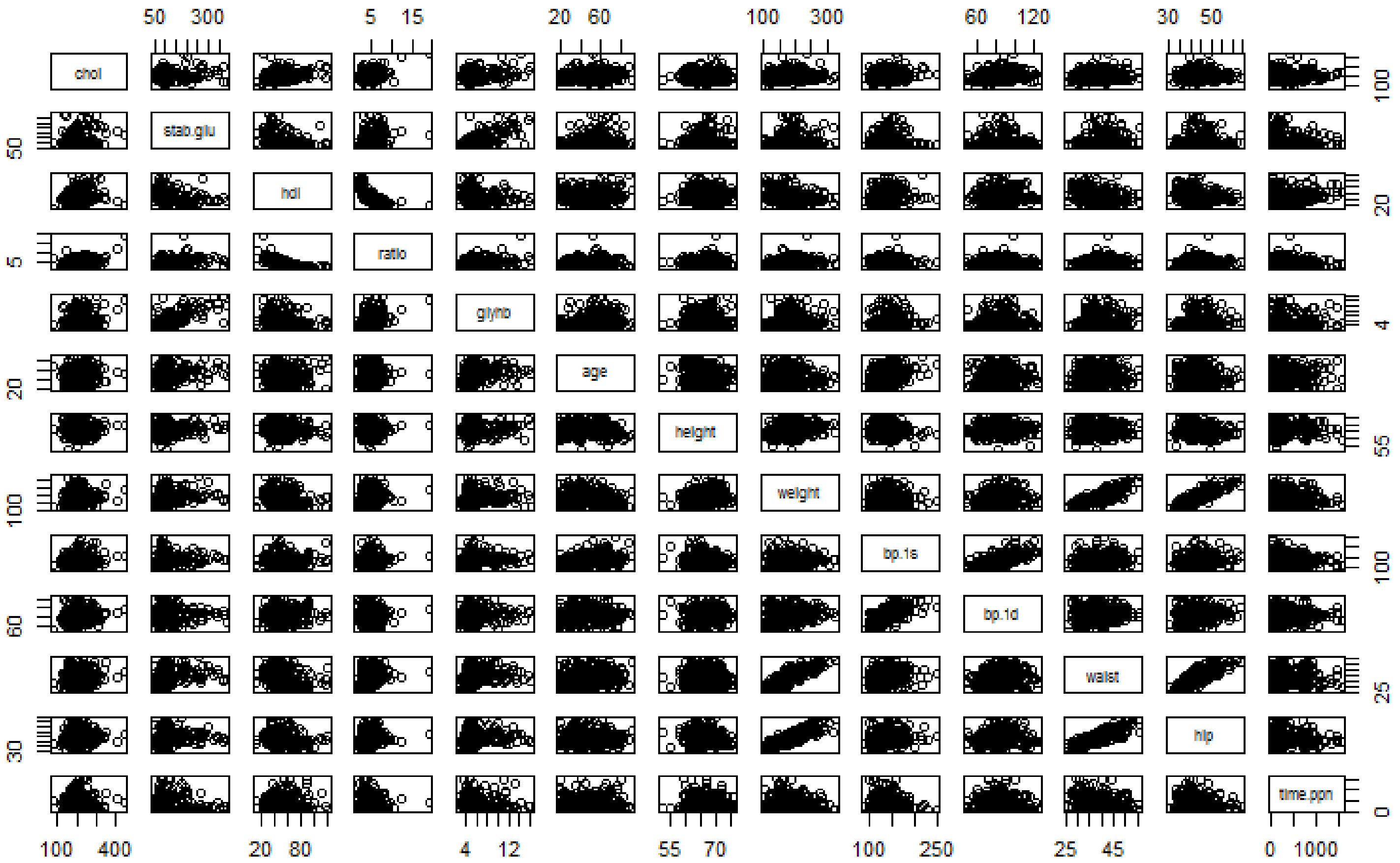
The distribution of the histogram of chol has heavy tails. The distribution of the histogram of stab.glu, hdl, ratio, glyhd, weight, bp.1s, and time.ppn are right-skewed. The distribuion of the histogram of age, height, bp.1d, waist and hip are approximately normal.





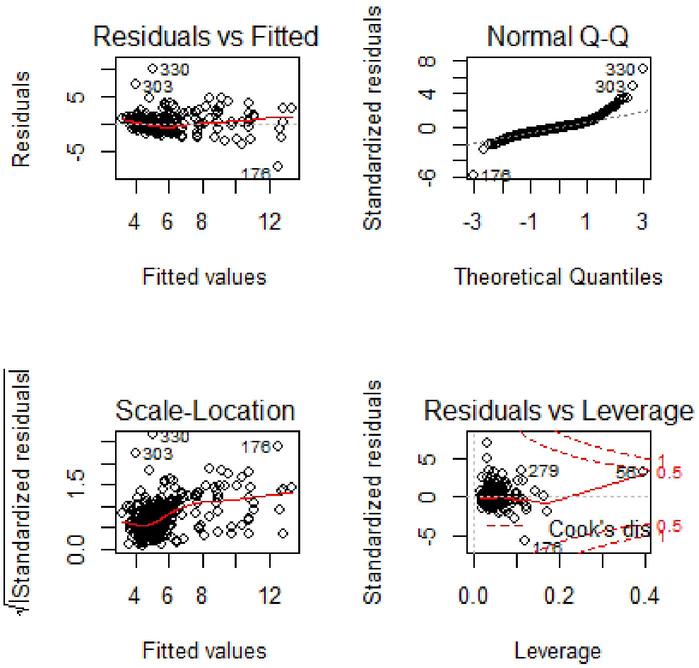


The pie chart of the location shows that the distribution of Buckingham and Louisa are about the same. The pie chart of the gender shows that the there are more females than male in the dataset. The pie chart of the frame shows that about half of the data are collected from people have a medium frame.



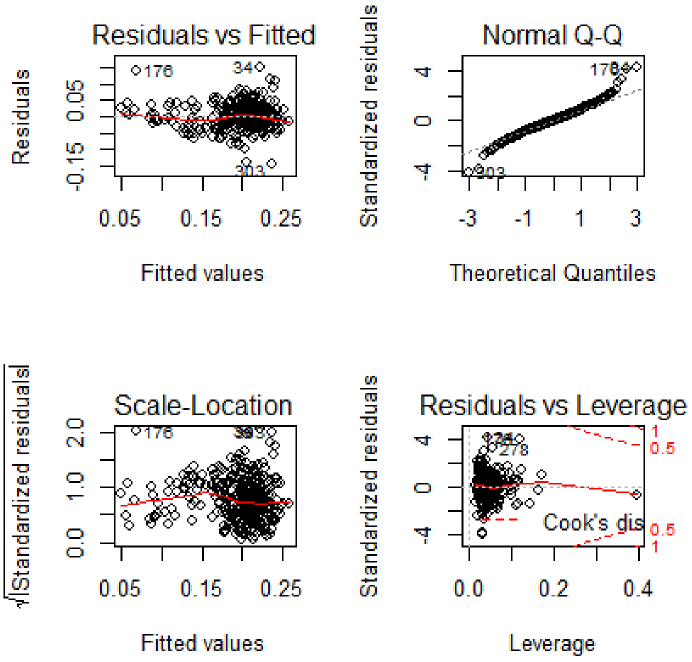
From the pairwise correlation matrix for all quantitative variables, we can see that some of the variables are correlated. Because some of the correlation plots show linear patterns or particular patterns.

2.



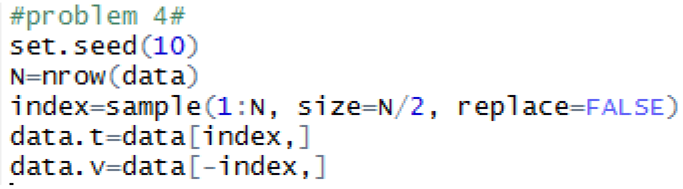
In Residuals vs Fitted plot, the reference line is approximately horizontal at x=0. Thus, the linearity assumption is hold. In the Normal Q-Q plot, we can see that points are spread below the refence line the left end and above the reference line at the right end, so the residuals are not normally distributed, and it has heavy tails. In the Scale-Location plot, we can see that the reference line is not horizonal, so the equal variance assumption doesn't hold. The Residuals vs Leverage plot shows no outlier.

3.



First applied boxcox function to the model1, the boxcox plot indicates that $(glyhd)^{-1}$ is a necessary transformation on glyhd. After applied to transformation, new diagnostic plots were plotted. We can tell that the equal variance assumption approximately holds on the new model; however, the normality assumption is still not hold. By applied the boxcox function to model2 again, the boxcox plot suggest that no further transformation on glyhd is needed.

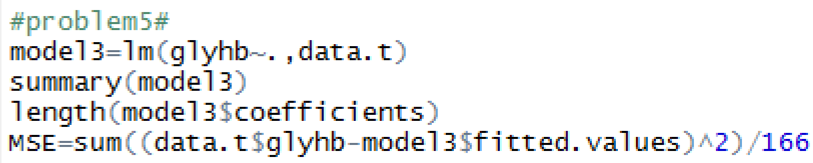
4.



```
#problem 4#
set.seed(10)
N=nrow(data)
index=sample(1:N, size=N/2, replace=FALSE)
data.t=data[index,]
data.v=data[-index,]
```

Selection of first-order effects

5.



```
#problem5#
model3=lm(glyhb~.,data.t)
summary(model3)
length(model3$coefficients)
MSE=sum((data.t$glyhb-model3$fitted.values)^2)/166
```

There are 17 regression coefficients in model 3, and the $\mathrm{MSE} \approx 0.00138$

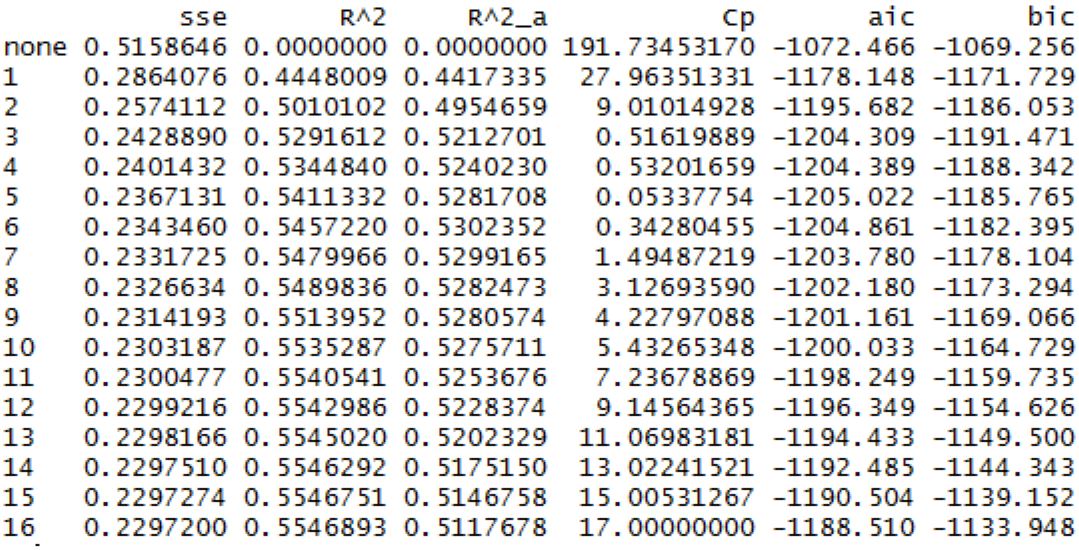
6.

Return of top 1 best subset of each subset size :

Subset sizes	**Best subsets**
null	null
1	Stab.glu
2	Stab.glu, age
3	Stab.glu, age, waist,
4	Stab.glu, age, waist, ratio
5	Stab.glu, age, waist, ratio, frame$small
6	Stab.glu, age, waist, ratio, frame$small, time.ppn
7	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s
8	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height
9	Stab.glu, age, waist, ratio, frame$small, time.ppn, heap, height, weight,
10	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap
11	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol
12	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol, location$louisa,
13	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol, location$louisa, hdl
14	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol, location$louisa, hdl, frame$medium

15	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol, location$louisa, hdl, frame$medium, bp.1d
16	Stab.glu, age, waist, ratio, frame$small, time.ppn, bp.1s, height, weight, heap, chol, location$louisa, hdl, frame$medium, bp.1d, gender$male

SSEp,R_p^2 , $R_{a,p}^2$, Cp, AICp, BICp for the best model of each subset:

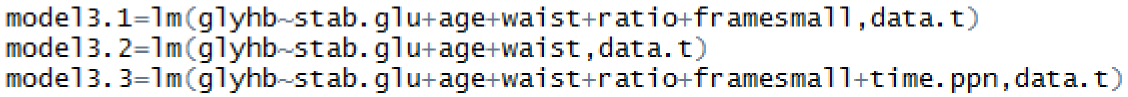


```
           sse       R^2     R^2_a           cp       aic       bic
none 0.5158646 0.0000000 0.0000000 191.73453170 -1072.466 -1069.256
1    0.2864076 0.4448009 0.4417335  27.96351331 -1178.148 -1171.729
2    0.2574112 0.5010102 0.4954659   9.01014928 -1195.682 -1186.053
3    0.2428890 0.5291612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

Best models According to different criterions:

Criterions	Best model
SSEp	Full model
R_p^2	Full model
$R_{a,p}^2$	Best Model of subset size 6
C_p	Best model of subset size 5
AIC_p	Best model of subset size 5
BIC_p	Best model of subset size 3

The C_p value of the best model according C_p criterion is 0.05337754, and the p is 6. The C_p value is quite smaller than the p. One possible explanation is that we use MSE of full model as an unbiased estimator of sigma in measuring C_p, and the MSE is overestimated.

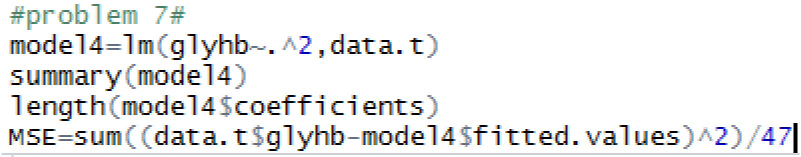


```
model3.1=lm(glyhb~stab.glu+age+waist+ratio+framesmall,data.t)
model3.2=lm(glyhb~stab.glu+age+waist,data.t)
model3.3=lm(glyhb~stab.glu+age+waist+ratio+framesmall+time.ppn,data.t)
```

Selection of first- and second- order effects

7.

There are 136 regression coefficients in this model. The MSE ≈0.001036. Model 4 has too many regression coefficients and we would lose lots of degree of freedom by fitting the data into this model.



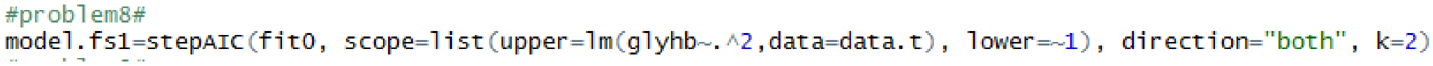
```
#problem 7#
model4=lm(glyhb~.^2,data.t)
summary(model4)
length(model4$coefficients)
MSE=sum((data.t$glyhb-model4$fitted.values)^2)/47
```

8.

The model being selected:

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

The AIC of model.fs1 is -1205.14, it is slightly smaller than the AIC of model3.1.



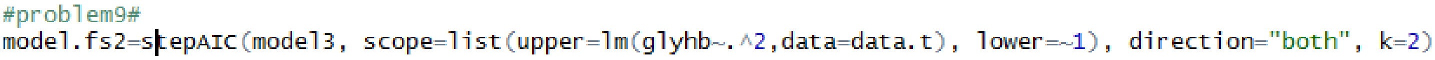
```
#problem8#
model.fs1=stepAIC(fit0, scope=list(upper=lm(glyhb~.^2,data=data.t), lower=~1), direction="both", k=2)
```

9.

The model being selected:

```
glyhb ~ chol + stab.glu + hdl + ratio + age + gender + height + 
    weight + bp.1s + bp.1d + waist + hip + time.ppn + stab.glu:gender + 
    hdl:ratio + age:bp.1d + weight:bp.1s + age:hip + hip:time.ppn + 
    gender:height + stab.glu:bp.1s + stab.glu:time.ppn + stab.glu:waist + 
    age:waist + chol:time.ppn + hdl:weight + bp.1d:waist + weight:hip
```

The AIC of model.fs2 is -1230.61 and it is significantly smaller than the AIC of model.fs1.

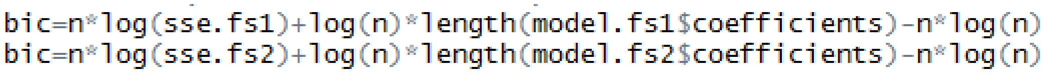


```
#problem9#
model.fs2=stepAIC(model3, scope=list(upper=lm(glyhb~.^2,data=data.t), lower=~1), direction="both", k=2)
```

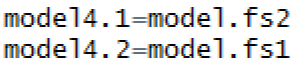
10.

```
BIC of model.fs1= -1182.677
BIC of model.fs2= -1137.536
```

The BIC of model.fs1 has a smaller value than the BIC of model.fs2. BIC and AIC choose different models.



```
bic=n*log(sse.fs1)+log(n)*length(model.fs1$coefficients)-n*log(n)
bic=n*log(sse.fs2)+log(n)*length(model.fs2$coefficients)-n*log(n)
```



```
model4.1=model.fs2
model4.2=model.fs1
```

Model validation

11.

PRESS of model3.1=0.25278

PRESS of model3.2=0.25398

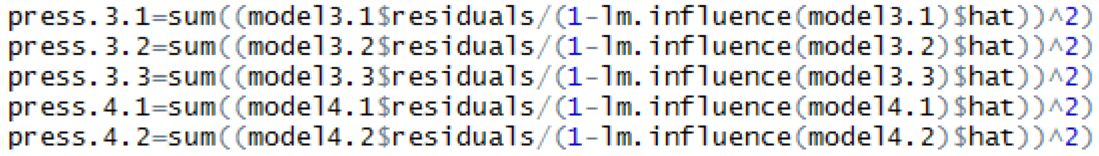
PRESS of model3.3=0.25257

PRESS of model4.1=0.21719

PRESS of model4.2=0.25348

From the result of PRESS, we can see that the value of PRESS is close to the value of SSE respectively.

By comparing the PRESS of each model, the model4.1 has a smallest PRESS values.



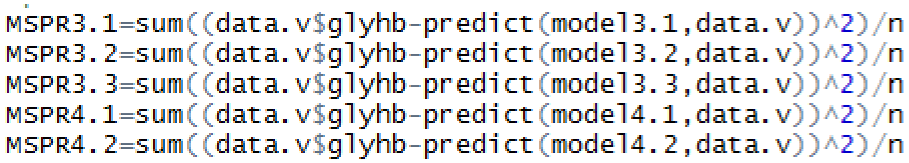
```
press.3.1=sum((model3.1$residuals/(1-lm.influence(model3.1)$hat))^2)
press.3.2=sum((model3.2$residuals/(1-lm.influence(model3.2)$hat))^2)
press.3.3=sum((model3.3$residuals/(1-lm.influence(model3.3)$hat))^2)
press.4.1=sum((model4.1$residuals/(1-lm.influence(model4.1)$hat))^2)
press.4.2=sum((model4.2$residuals/(1-lm.influence(model4.2)$hat))^2)
```

12.

MSPR of model3.1=0.001368	PRESS/n= 0.001381295
MSPR of model3.2=0.001377	PRESS/n= 0.001387887
MSPR of model3.3=0.001341	PRESS/n= 0.001380191
MSPR of model4.1=0.001798	PRESS/n= 0.001186856
MSPR of model4.2=0.001526	PRESS/n= 0.001385155

By comparing the MSPR of the models to the respective PRESS/n, the MSPR of model3.1, model3.2, and model3.3 are slightly smaller than their respective PRESS/n, and the MSPR of model4.1 and model4.2 are respectively much larger than their respective PRESS/n.

Model3.1 has the smallest MSPR.

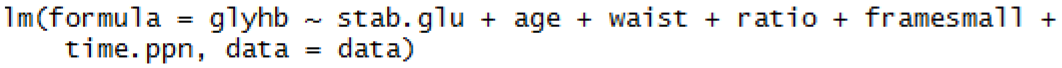


```
MSPR3.1=sum((data.v$glyhb-predict(model3.1,data.v))^2)/n
MSPR3.2=sum((data.v$glyhb-predict(model3.2,data.v))^2)/n
MSPR3.3=sum((data.v$glyhb-predict(model3.3,data.v))^2)/n
MSPR4.1=sum((data.v$glyhb-predict(model4.1,data.v))^2)/n
MSPR4.2=sum((data.v$glyhb-predict(model4.2,data.v))^2)/n
```

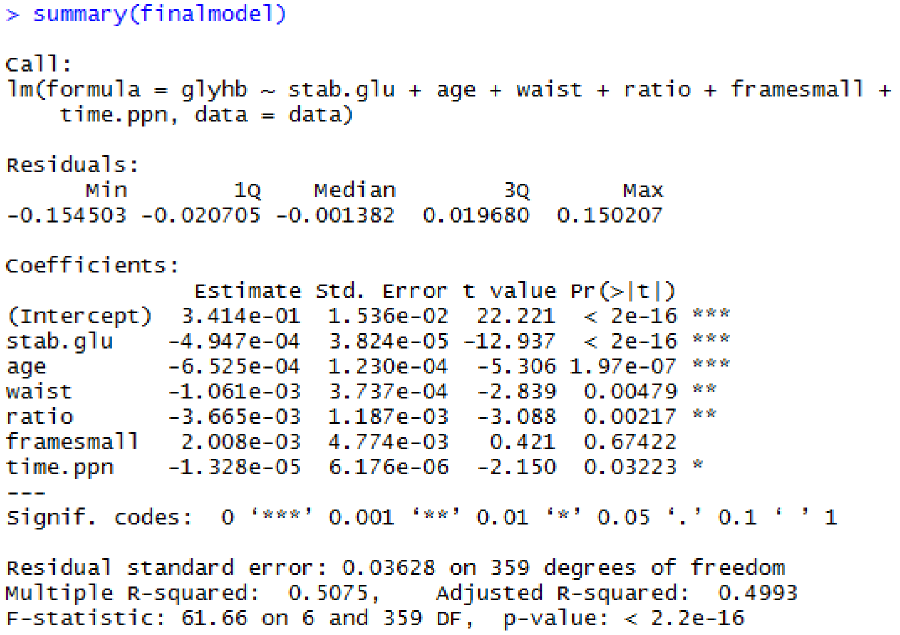
13.

The final model I would choose is model3.3, because it has a smallest MSPR. It indicates that model3.3 has a better predictive capacity. Also, model3.3 has significantly less predictor variables than model 4.1, which has the smallest PRESS.

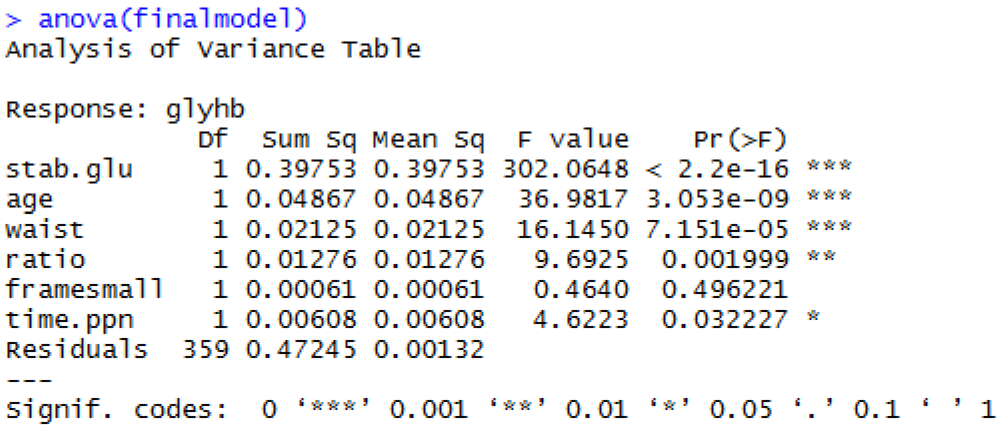
Fitted Regression function:



```
lm(formula = glyhb ~ stab.glu + age + waist + ratio + framesmall +
    time.ppn, data = data)
```

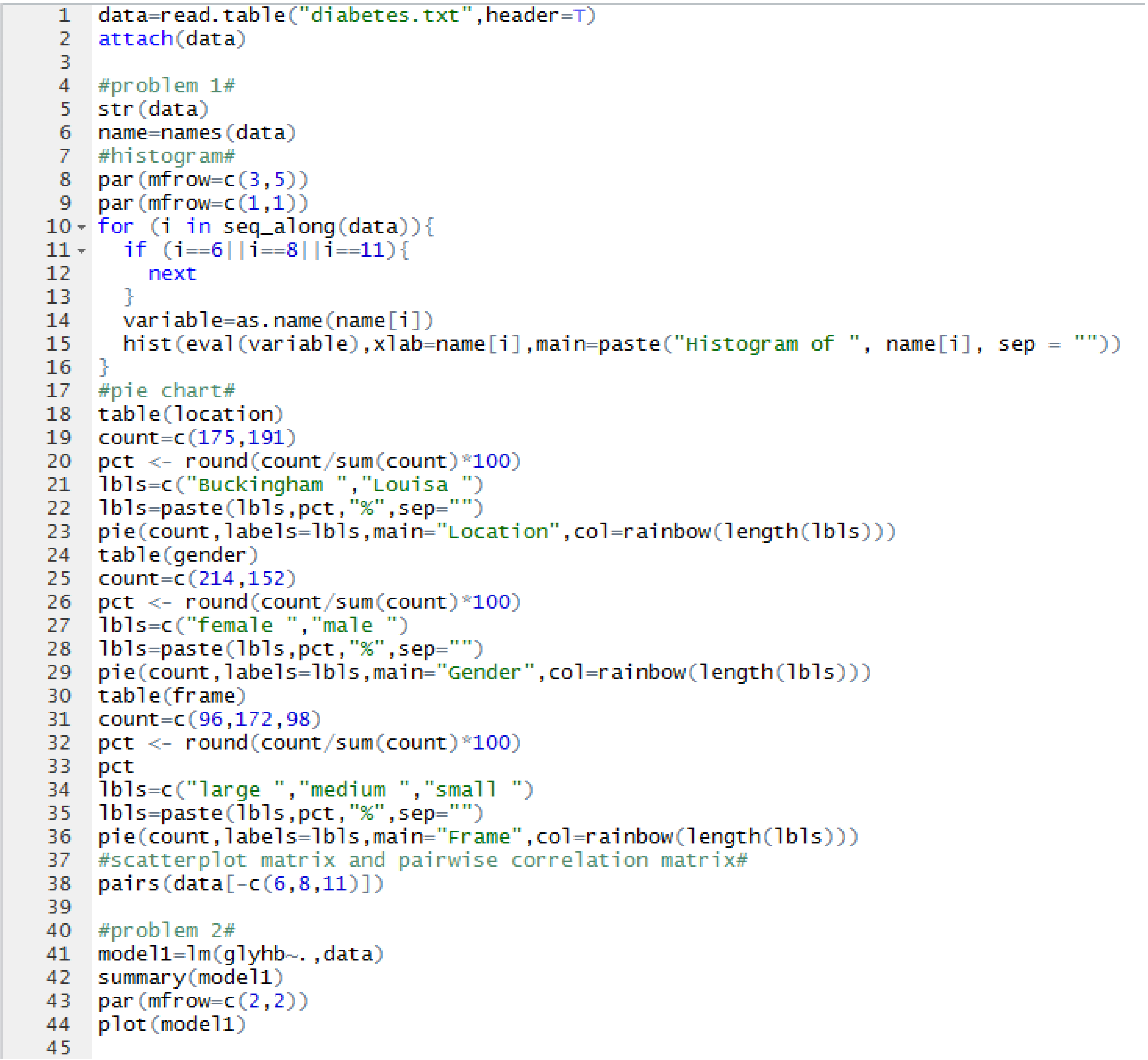


```
> summary(finalmodel)

Call:
lm(formula = glyhb ~ stab.glu + age + waist + ratio + framesmall + 
    time.ppn, data = data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.154503 -0.020705 -0.001382  0.019680  0.150207 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.414e-01  1.536e-02  22.221  < 2e-16 ***
stab.glu    -4.947e-04  3.824e-05 -12.937  < 2e-16 ***
age         -6.525e-04  1.230e-04  -5.306 1.97e-07 ***
waist       -1.061e-03  3.737e-04  -2.839  0.00479 ** 
ratio       -3.665e-03  1.187e-03  -3.088  0.00217 ** 
framesmall   2.008e-03  4.774e-03   0.421  0.67422    
time.ppn    -1.328e-05  6.176e-06  -2.150  0.03223 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03628 on 359 degrees of freedom
Multiple R-squared:  0.5075,	Adjusted R-squared:  0.4993 
F-statistic: 61.66 on 6 and 359 DF,  p-value: < 2.2e-16
```

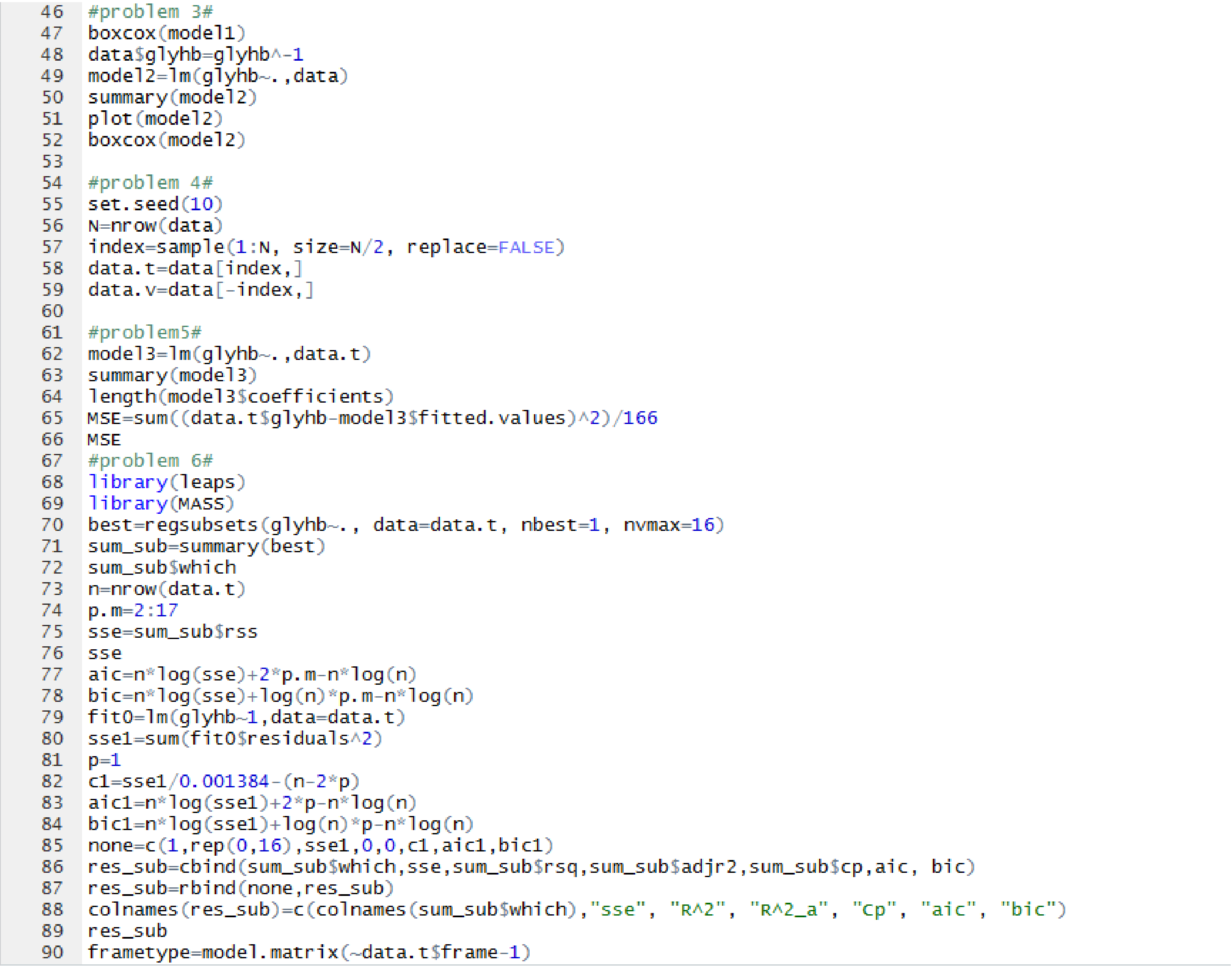


```
> anova(finalmodel)
Analysis of Variance Table

Response: glyhb
            Df  Sum Sq Mean Sq  F value    Pr(>F)    
stab.glu     1 0.39753 0.39753 302.0648 < 2.2e-16 ***
age          1 0.04867 0.04867  36.9817 3.053e-09 ***
waist        1 0.02125 0.02125  16.1450 7.151e-05 ***
ratio        1 0.01276 0.01276   9.6925  0.001999 ** 
framesmall   1 0.00061 0.00061   0.4640  0.496221    
time.ppn     1 0.00608 0.00608   4.6223  0.032227 *  
Residuals  359 0.47245 0.00132                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

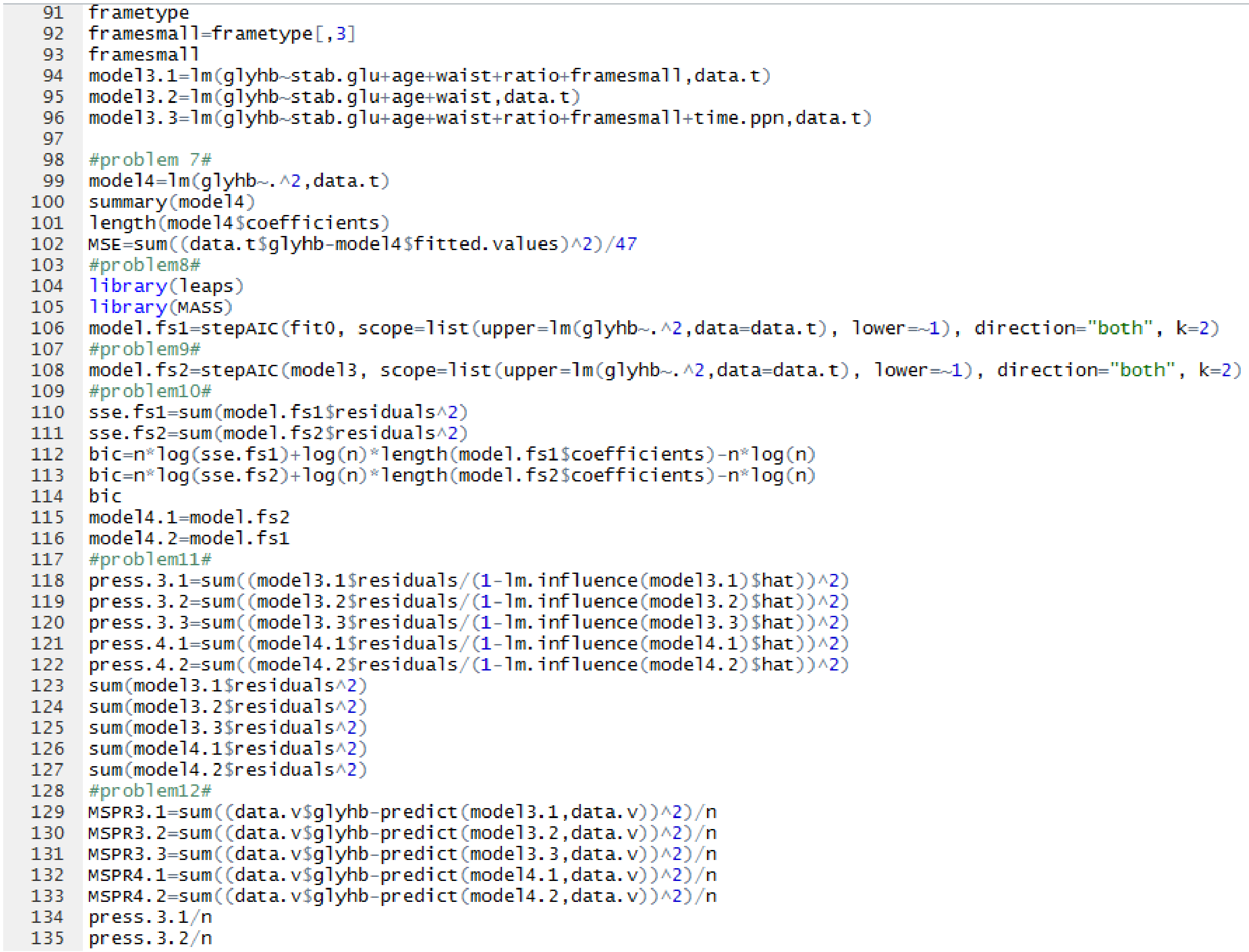
R code:



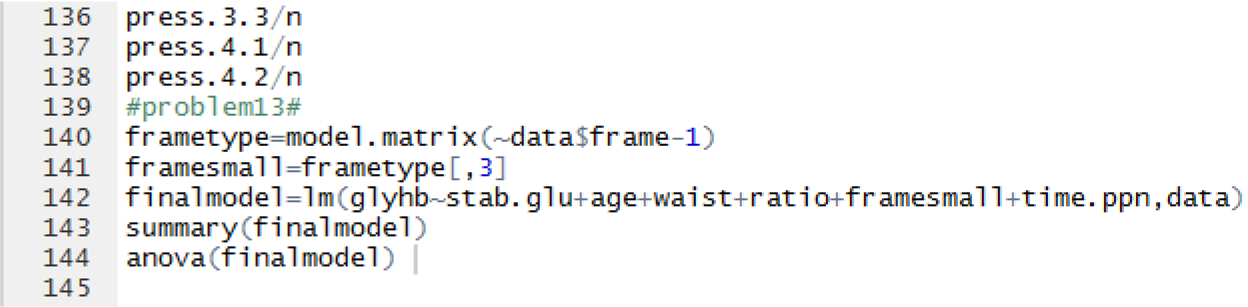
```
data=read.table("diabetes.txt",header=T)
attach(data)

#problem 1#
str(data)
name=names(data)
#histogram#
par(mfrow=c(3,5))
par(mfrow=c(1,1))
for (i in seq_along(data)){
  if (i==6||i==8||i==11){
    next
  }
  variable=as.name(name[i])
  hist(eval(variable),xlab=name[i],main=paste("Histogram of ", name[i], sep = ""))
}
#pie chart#
table(location)
count=c(175,191)
pct <- round(count/sum(count)*100)
lbls=c("Buckingham ","Louisa ")
lbls=paste(lbls,pct,"%",sep="")
pie(count,labels=lbls,main="Location",col=rainbow(length(lbls)))
table(gender)
count=c(214,152)
pct <- round(count/sum(count)*100)
lbls=c("female ","male ")
lbls=paste(lbls,pct,"%",sep="")
pie(count,labels=lbls,main="Gender",col=rainbow(length(lbls)))
table(frame)
count=c(96,172,98)
pct <- round(count/sum(count)*100)
pct
lbls=c("large ","medium ","small ")
lbls=paste(lbls,pct,"%",sep="")
pie(count,labels=lbls,main="Frame",col=rainbow(length(lbls)))
#scatterplot matrix and pairwise correlation matrix#
pairs(data[-c(6,8,11)])

#problem 2#
model1=lm(glyhb~.,data)
summary(model1)
par(mfrow=c(2,2))
plot(model1)

```



```
#problem 3#
boxcox(model1)
data$glyhb=glyhb^-1
model2=lm(glyhb~.,data)
summary(model2)
plot(model2)
boxcox(model2)

#problem 4#
set.seed(10)
N=nrow(data)
index=sample(1:N, size=N/2, replace=FALSE)
data.t=data[index,]
data.v=data[-index,]

#problem5#
model3=lm(glyhb~.,data.t)
summary(model3)
length(model3$coefficients)
MSE=sum((data.t$glyhb-model3$fitted.values)^2)/166
MSE
#problem 6#
library(leaps)
library(MASS)
best=regsubsets(glyhb~., data=data.t, nbest=1, nvmax=16)
sum_sub=summary(best)
sum_sub$which
n=nrow(data.t)
p.m=2:17
sse=sum_sub$rss
sse
aic=n*log(sse)+2*p.m-n*log(n)
bic=n*log(sse)+log(n)*p.m-n*log(n)
fit0=lm(glyhb~1,data=data.t)
sse1=sum(fit0$residuals^2)
p=1
c1=sse1/0.001384-(n-2*p)
aic1=n*log(sse1)+2*p-n*log(n)
bic1=n*log(sse1)+log(n)*p-n*log(n)
none=c(1,rep(0,16),sse1,0,0,c1,aic1,bic1)
res_sub=cbind(sum_sub$which,sse,sum_sub$rsq,sum_sub$adjr2,sum_sub$cp,aic, bic)
res_sub=rbind(none,res_sub)
colnames(res_sub)=c(colnames(sum_sub$which),"sse", "R^2", "R^2_a", "Cp", "aic", "bic")
res_sub
frametype=model.matrix(~data.t$frame-1)
```



```
frametype
framesmall=frametype[,3]
framesmall
model3.1=lm(glyhb~stab.glu+age+waist+ratio+framesmall,data.t)
model3.2=lm(glyhb~stab.glu+age+waist,data.t)
model3.3=lm(glyhb~stab.glu+age+waist+ratio+framesmall+time.ppn,data.t)

#problem 7#
model4=lm(glyhb~.^2,data.t)
summary(model4)
length(model4$coefficients)
MSE=sum((data.t$glyhb-model4$fitted.values)^2)/47
#problem8#
library(leaps)
library(MASS)
model.fs1=stepAIC(fit0, scope=list(upper=lm(glyhb~.^2,data=data.t), lower=~1), direction="both", k=2)
#problem9#
model.fs2=stepAIC(model3, scope=list(upper=lm(glyhb~.^2,data=data.t), lower=~1), direction="both", k=2)
#problem10#
sse.fs1=sum(model.fs1$residuals^2)
sse.fs2=sum(model.fs2$residuals^2)
bic=n*log(sse.fs1)+log(n)*length(model.fs1$coefficients)-n*log(n)
bic=n*log(sse.fs2)+log(n)*length(model.fs2$coefficients)-n*log(n)
bic
model4.1=model.fs2
model4.2=model.fs1
#problem11#
press.3.1=sum((model3.1$residuals/(1-lm.influence(model3.1)$hat))^2)
press.3.2=sum((model3.2$residuals/(1-lm.influence(model3.2)$hat))^2)
press.3.3=sum((model3.3$residuals/(1-lm.influence(model3.3)$hat))^2)
press.4.1=sum((model4.1$residuals/(1-lm.influence(model4.1)$hat))^2)
press.4.2=sum((model4.2$residuals/(1-lm.influence(model4.2)$hat))^2)
sum(model3.1$residuals^2)
sum(model3.2$residuals^2)
sum(model3.3$residuals^2)
sum(model4.1$residuals^2)
sum(model4.2$residuals^2)
#problem12#
MSPR3.1=sum((data.v$glyhb-predict(model3.1,data.v))^2)/n
MSPR3.2=sum((data.v$glyhb-predict(model3.2,data.v))^2)/n
MSPR3.3=sum((data.v$glyhb-predict(model3.3,data.v))^2)/n
MSPR4.1=sum((data.v$glyhb-predict(model4.1,data.v))^2)/n
MSPR4.2=sum((data.v$glyhb-predict(model4.2,data.v))^2)/n
press.3.1/n
press.3.2/n
```



```
press.3.3/n
press.4.1/n
press.4.2/n
#problem13#
frametype=model.matrix(~data$frame-1)
framesmall=frametype[,3]
finalmodel=lm(glyhb~stab.glu+age+waist+ratio+framesmall+time.ppn,data)
summary(finalmodel)
anova(finalmodel)

```