

Homework 5

Hyunjoon Rhee

4/11/2021

Problem 1

```
library(faraway)
data(fat)
```

```
dim(fat)
```

a)

```
## [1] 252 18
```

```
rm = seq(10,252,by=10)
train = fat[-rm,]
test = fat[rm,]
```

```
a <- lm(siri ~ . - brozek - density, data=train)
summary(a)
```

```
##
## Call:
## lm(formula = siri ~ . - brozek - density, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.8314 -0.6722  0.1828  0.9150  6.6619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.591885   6.448868  -1.953 0.052193 .
## age           0.007978   0.012320   0.648 0.517983
## weight       0.362999   0.023314  15.570 < 2e-16 ***
## height       0.049026   0.040315   1.216 0.225315
## adipos      -0.514032   0.114074  -4.506 1.09e-05 ***
## free        -0.564773   0.014889 -37.933 < 2e-16 ***
## neck         0.016525   0.089863   0.184 0.854272
## chest        0.120219   0.039590   3.037 0.002694 **
```

```
## abdom      0.140108    0.042186    3.321 0.001056 **
## hip        0.006197    0.056101    0.110 0.912148
## thigh      0.195057    0.054460    3.582 0.000424 ***
## knee       0.106637    0.093534    1.140 0.255542
## ankle      0.125118    0.081303    1.539 0.125325
## biceps     0.096199    0.064656    1.488 0.138278
## forearm    0.230775    0.073332    3.147 0.001888 **
## wrist      0.139279    0.206804    0.673 0.501378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.55 on 211 degrees of freedom
## Multiple R-squared:  0.9692, Adjusted R-squared:  0.967
## F-statistic: 442.5 on 15 and 211 DF,  p-value: < 2.2e-16
```

```
sqrt(mean((a$fit-train$siri)^2))
```

```
## [1] 1.494315
```

```
pred <- predict(a, test)
sqrt(mean((pred-test$siri)^2))
```

```
## [1] 1.131529
```

Using the test data to predict the root mean square error, which turned out to have a smaller value than just using the training data.

```
library(MASS)
b = step(a, direction = c("backward"), criterion = c("BIC"))
```

b)

```
## Start:  AIC=214.36
## siri ~ (brozek + density + age + weight + height + adipos + free +
##      neck + chest + abdom + hip + thigh + knee + ankle + biceps +
##      forearm + wrist) - brozek - density
##
##           Df Sum of Sq  RSS   AIC
## - hip      1      0.0 506.9 212.37
## - neck     1      0.1 507.0 212.39
## - age      1      1.0 507.9 212.81
## - wrist    1      1.1 508.0 212.84
## - knee     1      3.1 510.0 213.75
## - height   1      3.6 510.4 213.94
## <none>          506.9 214.36
## - biceps   1      5.3 512.2 214.73
## - ankle    1      5.7 512.6 214.89
## - chest    1     22.2 529.0 222.07
## - forearm  1     23.8 530.7 222.77
```

```

## - abdom    1      26.5  533.4 223.92
## - thigh    1      30.8  537.7 225.75
## - adipos   1      48.8  555.7 233.21
## - weight   1     582.4 1089.3 386.01
## - free     1    3456.8 3963.7 679.21
##
## Step: AIC=212.37
## siri ~ age + weight + height + adipos + free + neck + chest +
##       abdom + thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - neck      1         0.1  507.0 210.40
## - age        1         1.0  507.9 210.81
## - wrist      1         1.1  508.0 210.86
## - knee       1         3.2  510.1 211.80
## - height     1         3.5  510.4 211.95
## <none>                        506.9 212.37
## - biceps     1         5.3  512.2 212.73
## - ankle      1         5.7  512.6 212.89
## - chest      1        23.1  530.0 220.50
## - forearm    1        23.8  530.7 220.78
## - abdom      1        27.9  534.9 222.55
## - thigh      1        34.2  541.2 225.21
## - adipos     1        50.3  557.2 231.85
## - weight     1       683.9 1190.8 404.23
## - free       1    3488.9 3995.8 679.05
##
## Step: AIC=210.4
## siri ~ age + weight + height + adipos + free + chest + abdom +
##       thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - age        1         1.1  508.1 208.88
## - wrist      1         1.3  508.3 208.99
## - knee       1         3.1  510.1 209.80
## - height     1         3.6  510.6 210.02
## <none>                        507.0 210.40
## - biceps     1         5.4  512.4 210.80
## - ankle      1         5.6  512.6 210.89
## - chest      1        23.2  530.2 218.55
## - forearm    1        24.6  531.6 219.15
## - abdom      1        28.0  535.0 220.60
## - thigh      1        34.4  541.4 223.29
## - adipos     1        50.8  557.8 230.07
## - weight     1       689.6 1196.6 403.34
## - free       1    3532.0 4039.0 679.49
##
## Step: AIC=208.88
## siri ~ weight + height + adipos + free + chest + abdom + thigh +
##       knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - wrist      1         2.9  511.0 208.19
## - height     1         3.3  511.4 208.35

```

```

## - knee      1      4.5  512.5 208.87
## <none>                508.1 208.88
## - ankle     1      5.2  513.2 209.18
## - biceps    1      6.0  514.0 209.53
## - forearm   1     23.6  531.6 217.18
## - chest     1     24.2  532.3 217.46
## - abdom     1     33.7  541.8 221.48
## - thigh     1     35.3  543.3 222.12
## - adipos    1     51.1  559.1 228.63
## - weight    1    699.1 1207.2 403.34
## - free      1   3598.0 4106.0 681.23
##
## Step: AIC=208.19
## siri ~ weight + height + adipos + free + chest + abdom + thigh +
##       knee + ankle + biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## - height   1         3.8  514.8 207.89
## <none>                511.0 208.19
## - knee     1         5.7  516.7 208.72
## - ankle    1         6.9  517.9 209.24
## - biceps   1         7.0  518.0 209.30
## - chest    1        23.8  534.8 216.53
## - forearm  1        27.7  538.7 218.16
## - thigh    1        32.4  543.4 220.13
## - abdom    1        37.3  548.3 222.19
## - adipos   1        49.3  560.3 227.11
## - weight   1       696.5 1207.5 401.40
## - free     1      3798.4 4309.4 690.20
##
## Step: AIC=207.89
## siri ~ weight + adipos + free + chest + abdom + thigh + knee +
##       ankle + biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## <none>                514.8 207.89
## - knee     1         5.1  519.9 208.12
## - ankle    1         7.4  522.2 209.11
## - biceps   1         7.5  522.4 209.18
## - chest    1        24.0  538.9 216.25
## - forearm  1        28.8  543.6 218.23
## - thigh    1        30.0  544.8 218.73
## - abdom    1        39.1  553.9 222.49
## - adipos   1        86.6  601.4 241.18
## - weight   1       819.8 1334.7 422.13
## - free     1      3809.4 4324.2 688.98

```

```
library(pls)
```

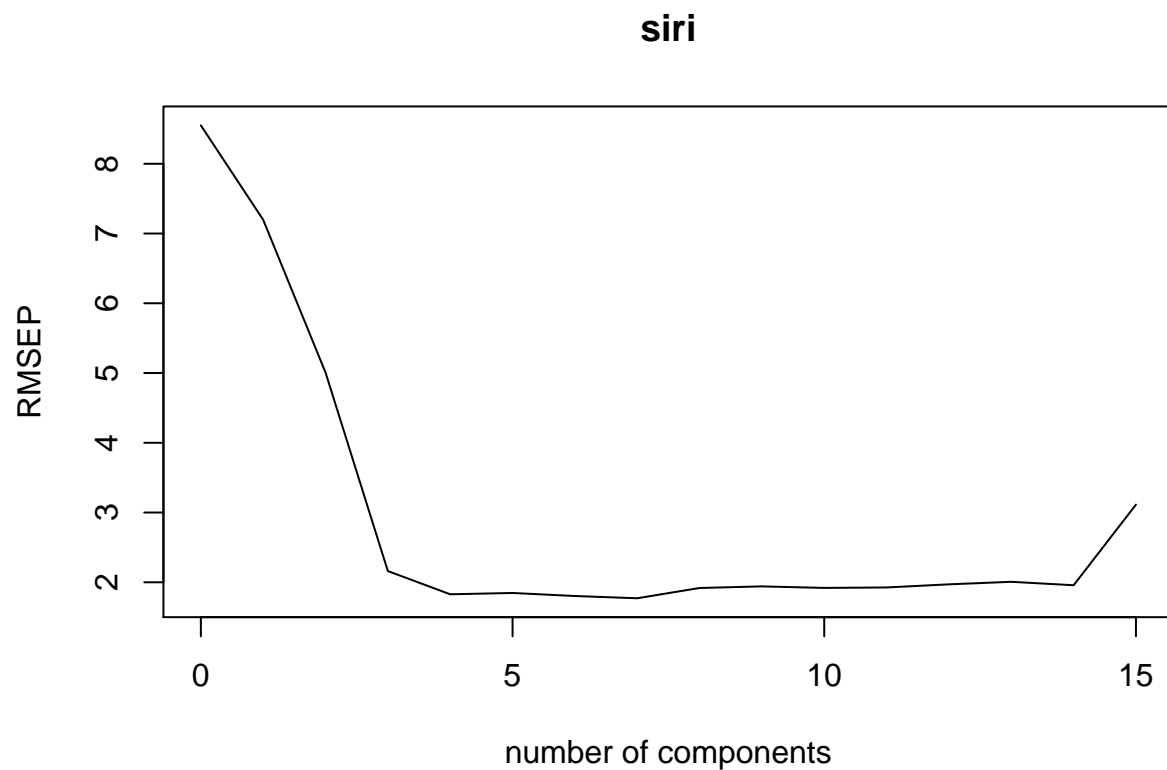
c)

```
## Warning: package 'pls' was built under R version 4.0.2
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##      loadings

pcrm = pcr(siri ~ . -brozek - density, data = train, validation = "CV", ncomp = 15)
pcrfat = RMSEP(pcrm, estimate = "CV")
plot(pcrfat)
```



```
library(lars)
```

d)

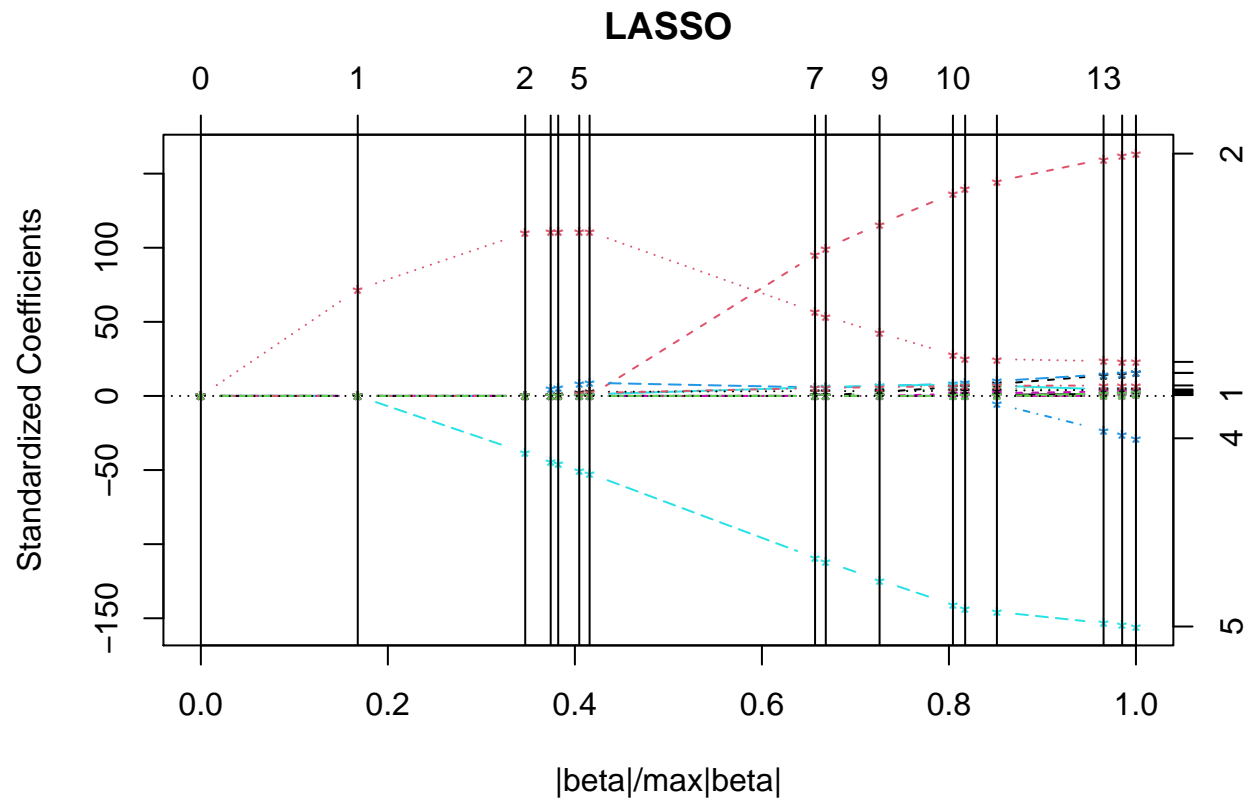
```
## Warning: package 'lars' was built under R version 4.0.2
```

```
## Loaded lars 1.2
```

```

traind <- train[ , -c(1:3)]
x <- as.matrix(traind)
lm.lasso<- lars(x,train$siri)
plot(lm.lasso)

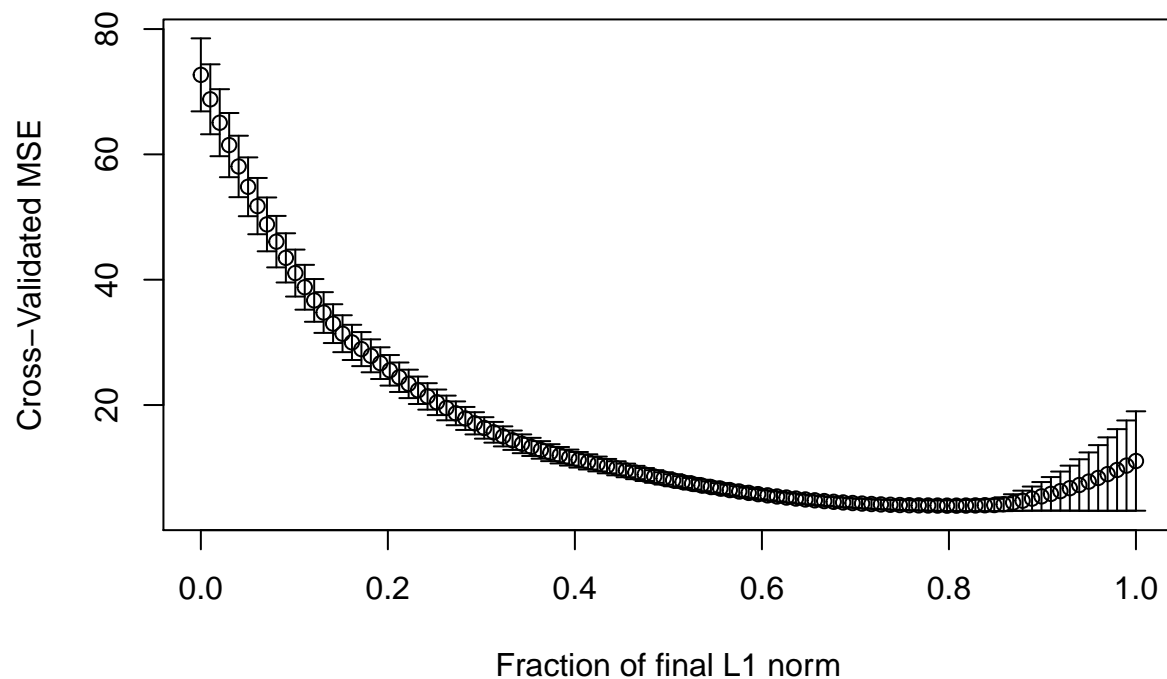
```



```

cvlass <- cv.lars(x,train$siri)

```



```
svm = cvlass$index[which.min(cvlass$cv)]
```

e) By comparing everything, pcr shows the lowest RMSE value.

Problem 2

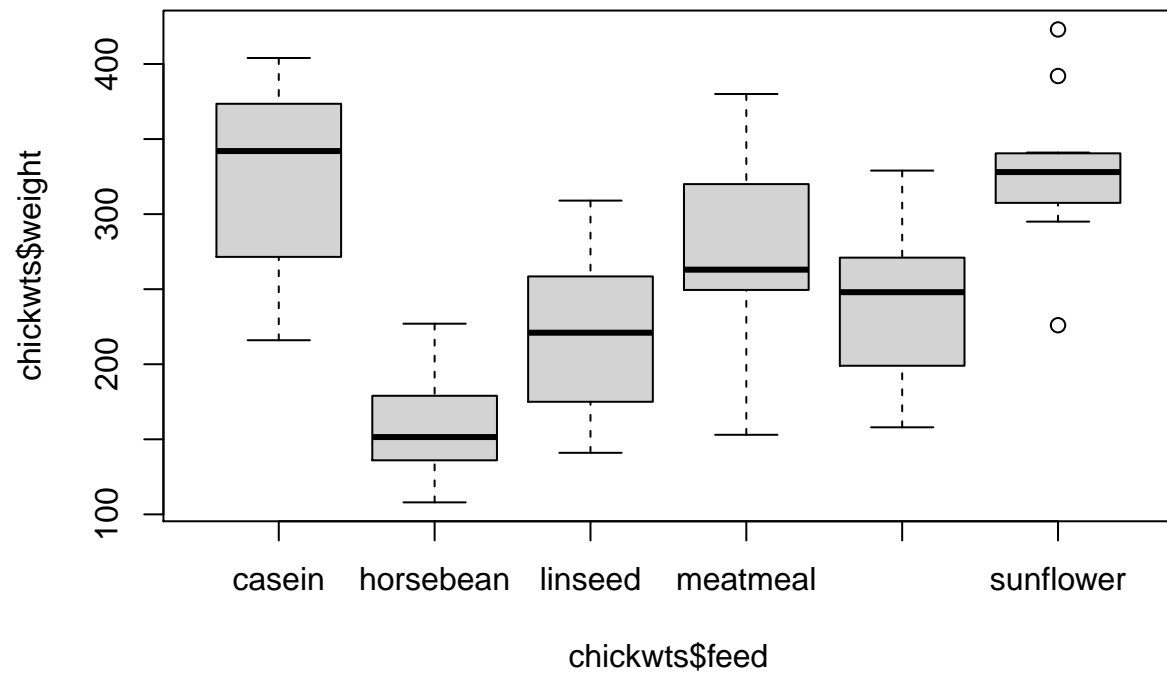
```
data(chickwts)

a2 = lm(weight ~ feed, data = chickwts)
anova(a2)
```

a)

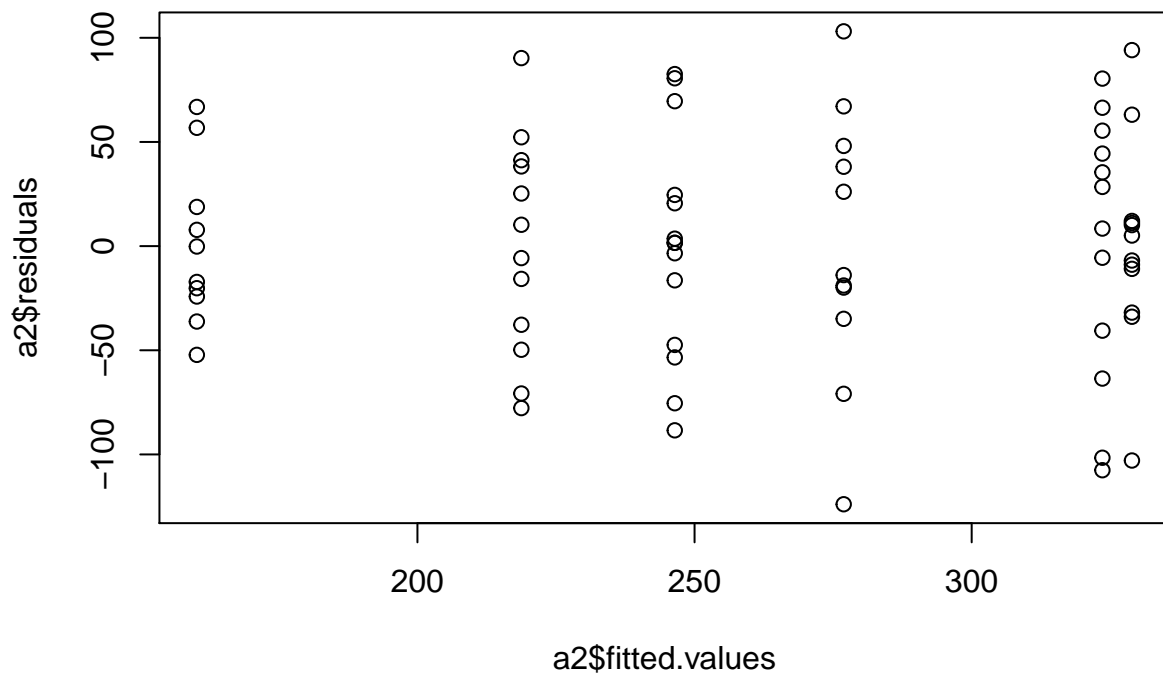
```
## Analysis of Variance Table
##
## Response: weight
##          Df Sum Sq Mean Sq F value    Pr(>F)
## feed      5 231129   46226   15.365 5.936e-10 ***
## Residuals 65 195556     3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#plot  
plot(chickwts$weight~chickwts$feed)
```



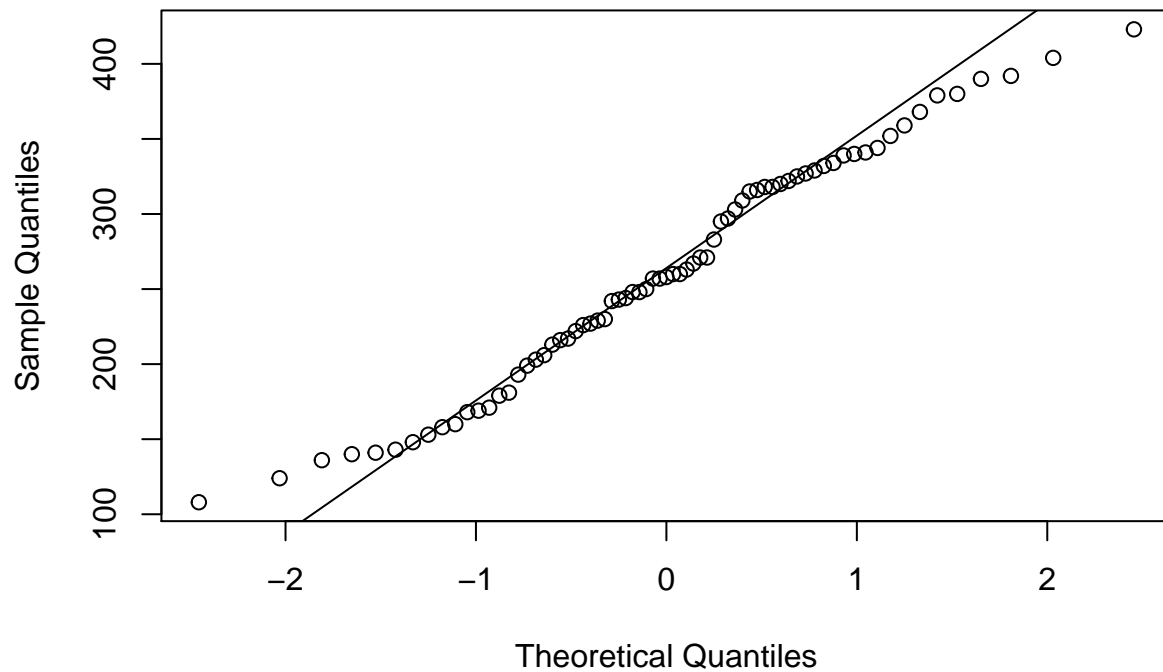
Assumptions

```
plot(a2$fitted.values,a2$residuals)
```

```
qqnorm(chickwts$weight)
qqline(chickwts$weight)
```

Normal Q-Q Plot



```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.2
```

```
## Registered S3 methods overwritten by 'car':
```

```
##   method                      from  
##   influence.merMod             lme4  
##   cooks.distance.influence.merMod lme4  
##   dfbeta.influence.merMod       lme4  
##   dfbetas.influence.merMod      lme4
```

```
##
```

```
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
```

```
##
```

```
##   logit, vif
```

```
leveneTest(weight ~ feed,data=chickwts)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group  5  0.7493 0.5896
##      65
```

```
a2b = lm(weight ~ feed-1, data = chickwts)
```

```
anova(a2b)
```

```
## Analysis of Variance Table
##
## Response: weight
##      Df Sum Sq Mean Sq F value    Pr(>F)
## feed    6 5079211  846535  281.38 < 2.2e-16 ***
## Residuals 65  195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
a2b$coefficients
```

```
##      feedcasein feedhorsebean  feedlinseed  feedmeatmeal  feedsoybean
##      323.5833      160.2000      218.7500      276.9091      246.4286
## feedsunflower
##      328.9167
```

```
attach(chickwts)
contrasts(feed)
```

```
##      horsebean linseed meatmeal soybean sunflower
## casein         0      0      0      0      0
## horsebean      1      0      0      0      0
## linseed        0      1      0      0      0
## meatmeal       0      0      1      0      0
## soybean        0      0      0      1      0
## sunflower      0      0      0      0      1
```

```
contrasts(feed) = contr.sum(6)
a2c=lm(weight~feed)
summary(a2c)
```

```
##
## Call:
## lm(formula = weight ~ feed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413   1.571   38.170  103.091
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  259.131      6.543   39.602 < 2e-16 ***
## feed1         64.452     14.490    4.448 3.47e-05 ***
## feed2        -98.931     15.601   -6.341 2.48e-08 ***
## feed3        -40.381     14.490   -2.787 0.00697 **
## feed4         17.778     15.005    1.185 0.24042
## feed5        -12.703     13.641   -0.931 0.35519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

```
a2c$coefficients
```

```
## (Intercept)      feed1      feed2      feed3      feed4      feed5
##   259.13128    64.45206   -98.93128   -40.38128    17.77781   -12.70271
```

```
259.13+64.45-98.93-40.38+17.78-12.70
```

```
## [1] 189.35
```

There are differences of weights among feed.

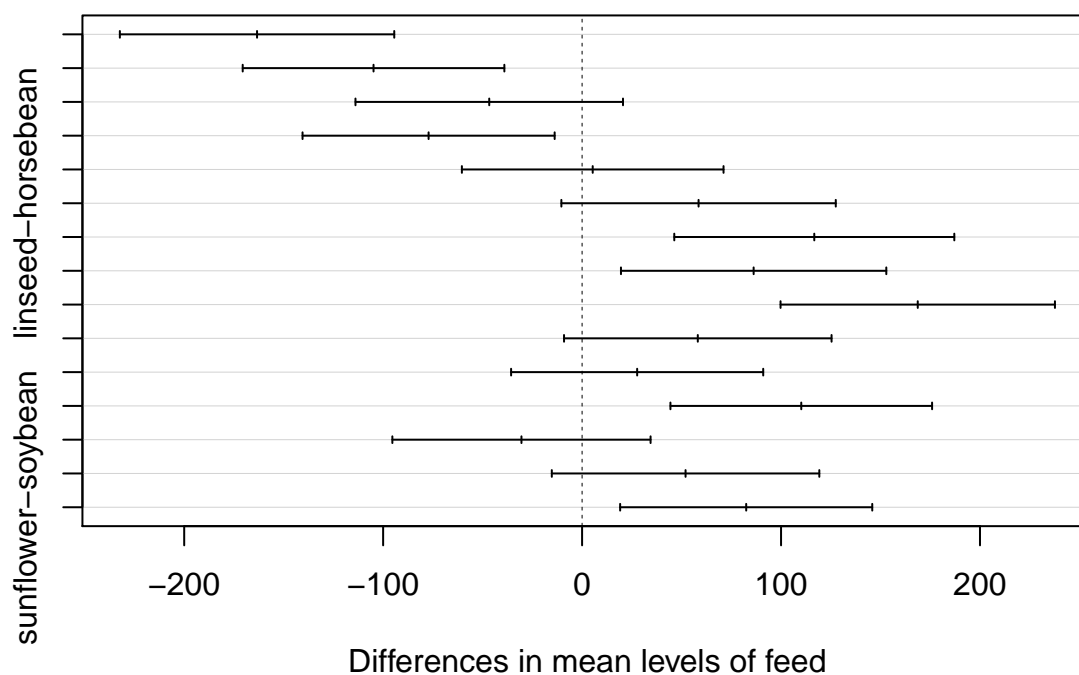
```
pairwise.t.test(weight,feed,p.adjust.method = "bonferroni")
```

b)

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  weight and feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 3.1e-08 -          -          -          -
## linseed   0.00022 0.22833 -          -          -
## meatmeal   0.68350 0.00011 0.20218 -          -
## soybean    0.00998 0.00487 1.00000 1.00000 -
## sunflower 1.00000 1.2e-08 9.3e-05 0.39653 0.00447
##
## P value adjustment method: bonferroni
```

```
tci<-TukeyHSD(aov(weight~feed),data=chickwts)
plot(tci)
```

95% family-wise confidence level



Problem 3

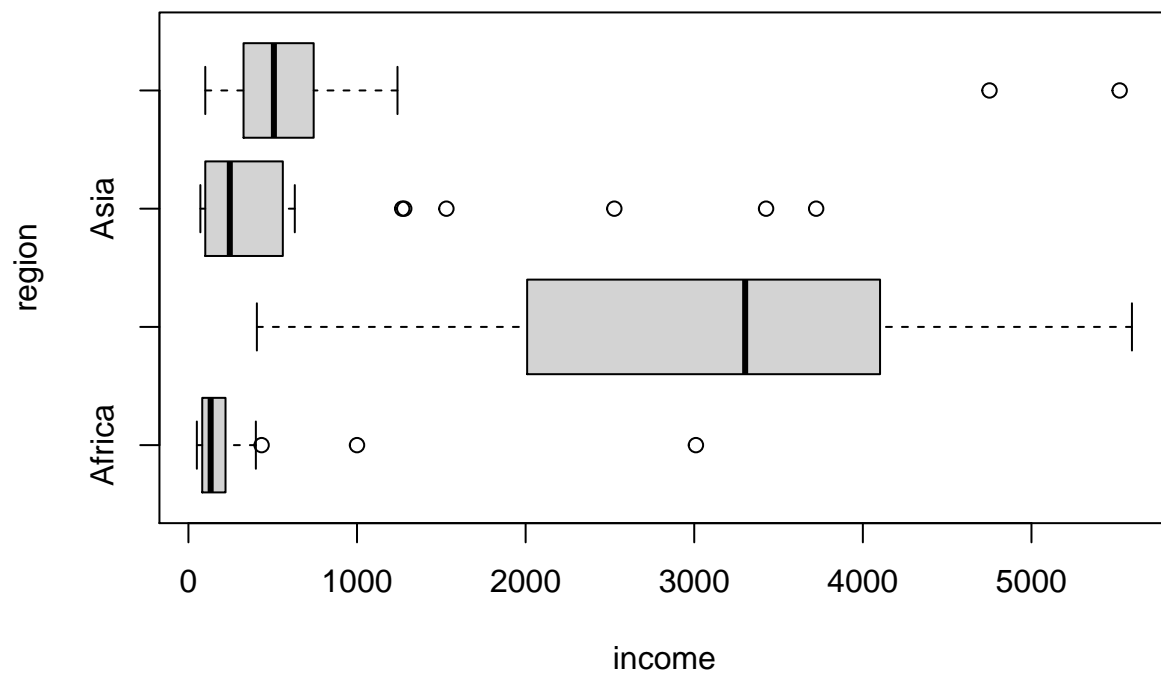
```
data(infmort)
```

```
a3<- aov(lm(income ~ region,data=infmort))
summary(a3)
```

a)

```
##           Df    Sum Sq Mean Sq F value    Pr(>F)
## region      3  96878763 32292921   29.16 1.16e-13 ***
## Residuals 101 111857493  1107500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(income ~ region,data=infmort,horizontal=TRUE)
```

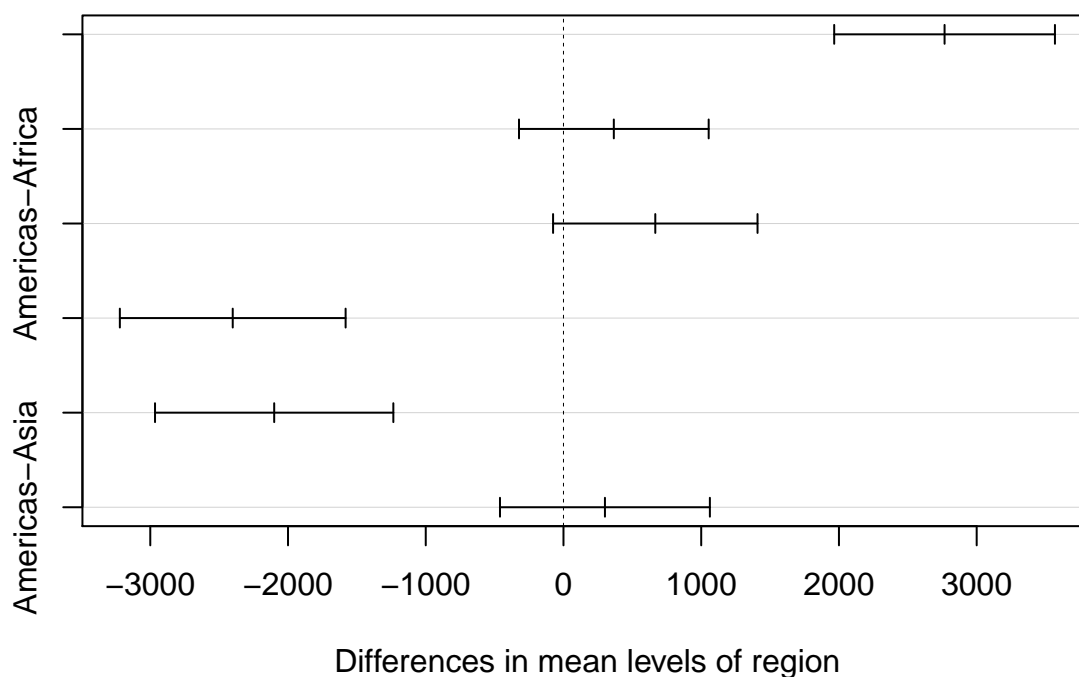


```
t<- TukeyHSD(a3)
t$region[,4]
```

```
## Europe-Africa Asia-Africa Americas-Africa Asia-Europe Americas-Europe
## 1.257883e-13 5.103789e-01 9.450001e-02 7.172019e-11 3.855869e-08
## Americas-Asia
## 7.311095e-01
```

```
plot(t)
```

95% family-wise confidence level



```
aov(formula = lm(income ~ region, data = infmort))
```

```
## Call:
## aov(formula = lm(income ~ region, data = infmort))
##
## Terms:
##           region Residuals
## Sum of Squares  96878763 111857493
## Deg. of Freedom      3      101
##
## Residual standard error: 1052.378
## Estimated effects may be unbalanced
```

Problem 3

a)