Homework 8

2024-10-24

##1.

a. We do not see significance in the leg predictor on the response.

```
data(seatpos, package = "faraway")

lmod <- lm(hipcenter ~ ., seatpos)

summary(lmod)
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

b.

```
x <- model.matrix(lmod)

x0 <- apply(x, 2, mean)

predict(lmod, new = data.frame(t(x0)), interval = "prediction", level=0.99)
```

```
##          fit       lwr       upr
## 1 -164.8849 -270.2157 -59.55403
```

c.

Part 1. Backwards elimination

```r
lmod <- lm(hipcenter ~ ., seatpos)

lmod2 <- update(lmod, . ~ . -Ht)

lmod3 <- update(lmod2, . ~ . -Weight)

lmod4 <- update(lmod3, . ~ . -Seated)

lmod5 <- update(lmod4, . ~ . -Arm)

lmod6 <- update(lmod5, . ~ . -Thigh)

lmod7 <- update(lmod6, . ~ . -Age)

lmod8 <- update(lmod7, . ~ . -Leg)


# Used in the process of finding the best model:
# summary(lmod)
# summary(lmod2)
# summary(lmod3)
# summary(lmod4)
# summary(lmod5)
# summary(lmod6)
# summary(lmod7)

summary(lmod8)
```

```
##
## Call:
## lm(formula = hipcenter ~ HtShoes, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.981 -27.150   2.983  22.637  73.731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 565.5927    92.5794   6.109 4.97e-07 ***
## HtShoes      -4.2621     0.5391  -7.907 2.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.55 on 36 degrees of freedom
## Multiple R-squared:  0.6346, Adjusted R-squared:  0.6244
## F-statistic: 62.51 on 1 and 36 DF,  p-value: 2.207e-09
```

Part 2. AIC: We find the optimal model seemingly to be Age + Ht + Leg

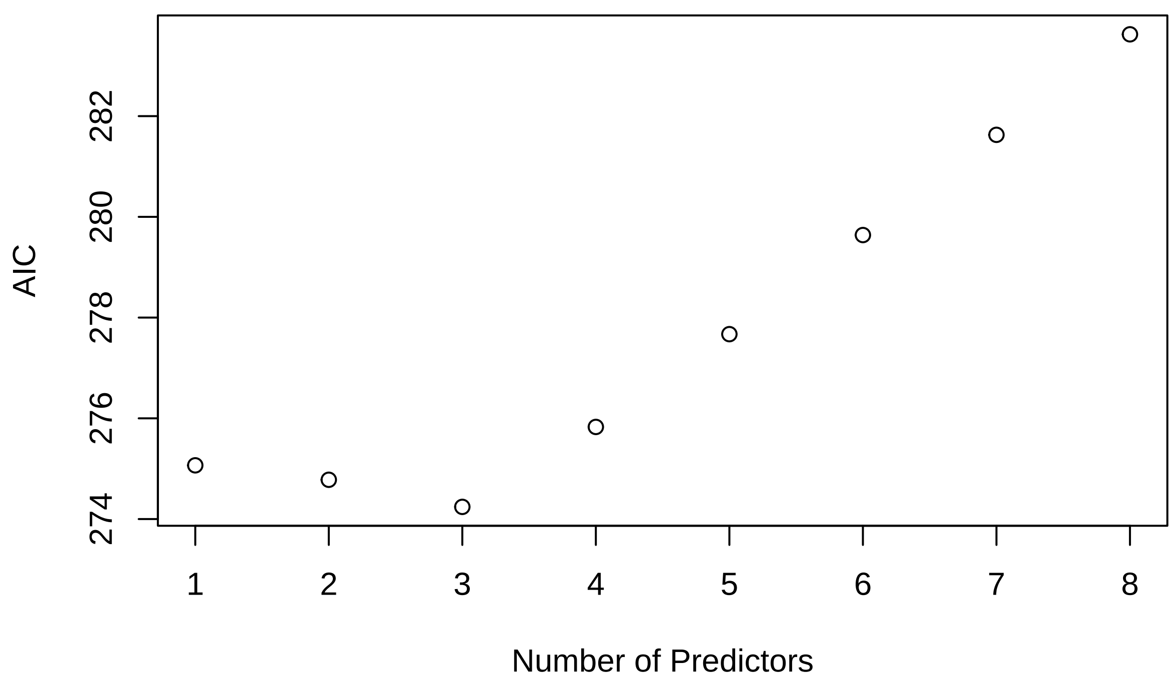
```r
require(leaps)
```

```
## Loading required package: leaps
```

```r
b <- regsubsets(hipcenter ~ ., seatpos)

rs <- summary(b)

rs$which
```

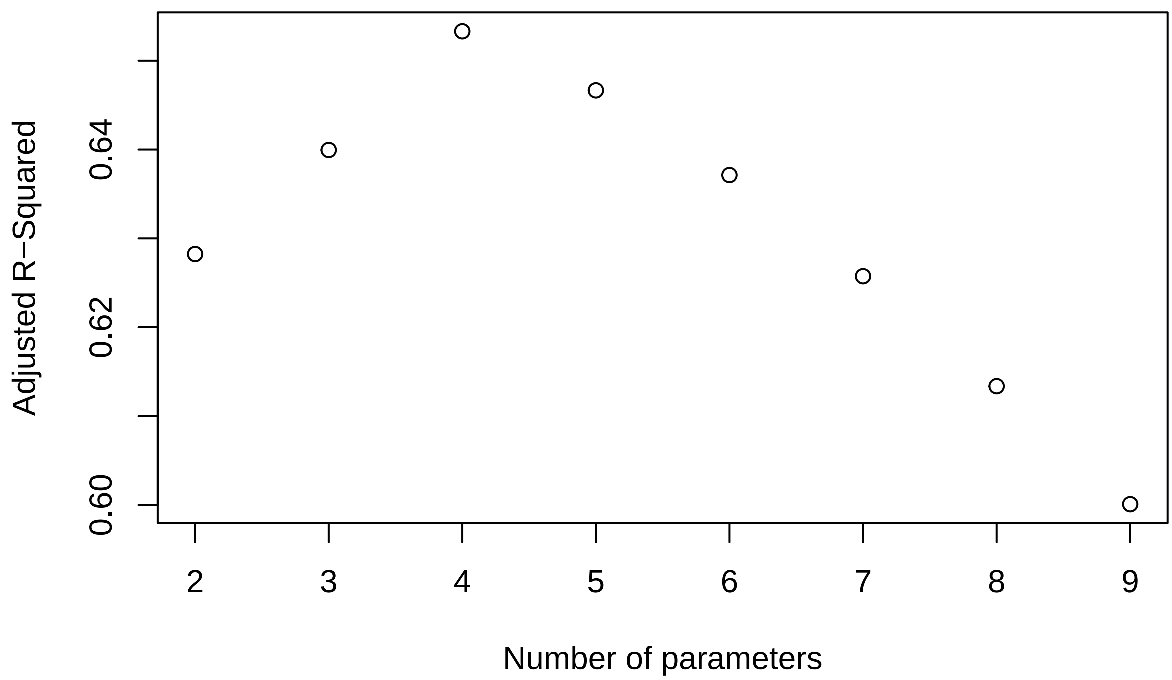
```
##   (Intercept)   Age Weight HtShoes    Ht Seated   Arm Thigh   Leg
## 1        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE FALSE
## 2        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 3        TRUE  TRUE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 4        TRUE  TRUE  FALSE    TRUE FALSE  FALSE FALSE  TRUE  TRUE
## 5        TRUE  TRUE  FALSE    TRUE FALSE  FALSE  TRUE  TRUE  TRUE
## 6        TRUE  TRUE  FALSE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 7        TRUE  TRUE   TRUE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 8        TRUE  TRUE   TRUE    TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
```

```r
AIC <- 38 * log(rs$rss/38) + (2:9) * 2

plot(AIC ~ I(1:8), ylab = "AIC", xlab = "Number of Predictors")
```



Part 3. Adjusted R^2: again, we find the best model to be Age + Ht + Leg

```r
plot(2:9, rs$adjr2, xlab = "Number of parameters", ylab = "Adjusted R-Squared")
```




```r
which.max(rs$adjr2)
```

```
## [1] 3
```

Part 4. Stepwise selection: we find Age + HtShoes + Leg appears to be the best model.

```r
lmod <- lm(hipcenter ~ ., data=seatpos)
step(lmod)
```

```
## Start:  AIC=283.62
## hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##
##           Df Sum of Sq   RSS    AIC
## - Ht       1      5.01 41267 281.63
## - Weight   1      8.99 41271 281.63
## - Seated   1     28.64 41290 281.65
## - HtShoes  1    108.43 41370 281.72
## - Arm      1    164.97 41427 281.78
## - Thigh    1    262.76 41525 281.87
```

```
## <none>                  41262 283.62
## - Age      1   2632.12 43894 283.97
## - Leg      1   2654.85 43917 283.99
##
## Step:  AIC=281.63
## hipcenter ~ Age + Weight + HtShoes + Seated + Arm + Thigh + Leg
##
##           Df Sum of Sq   RSS    AIC
## - Weight   1     11.10 41278 279.64
## - Seated   1     30.52 41297 279.66
## - Arm      1    160.50 41427 279.78
## - Thigh    1    269.08 41536 279.88
## - HtShoes  1    971.84 42239 280.51
## <none>                  41267 281.63
## - Leg      1   2664.65 43931 282.01
## - Age      1   2808.52 44075 282.13
##
## Step:  AIC=279.64
## hipcenter ~ Age + HtShoes + Seated + Arm + Thigh + Leg
##
##           Df Sum of Sq   RSS    AIC
## - Seated   1     35.10 41313 277.67
## - Arm      1    156.47 41434 277.78
## - Thigh    1    285.16 41563 277.90
## - HtShoes  1    975.48 42253 278.53
## <none>                  41278 279.64
## - Leg      1   2661.39 43939 280.01
## - Age      1   3011.86 44290 280.31
##
## Step:  AIC=277.67
## hipcenter ~ Age + HtShoes + Arm + Thigh + Leg
##
##           Df Sum of Sq   RSS    AIC
## - Arm      1    172.02 41485 275.83
## - Thigh    1    344.61 41658 275.99
## - HtShoes  1   1853.43 43166 277.34
## <none>                  41313 277.67
## - Leg      1   2871.07 44184 278.22
## - Age      1   2976.77 44290 278.31
##
## Step:  AIC=275.83
## hipcenter ~ Age + HtShoes + Thigh + Leg
##
##           Df Sum of Sq   RSS    AIC
## - Thigh    1     472.8 41958 274.26
## <none>                  41485 275.83
## - HtShoes  1    2340.7 43826 275.92
## - Age      1    3501.0 44986 276.91
## - Leg      1    3591.7 45077 276.98
##
## Step:  AIC=274.26
## hipcenter ~ Age + HtShoes + Leg
##
##           Df Sum of Sq   RSS    AIC
```

```
## <none>                   41958 274.26
## - Age      1    3108.8 45067 274.98
## - Leg      1    3476.3 45434 275.28
## - HtShoes  1    4218.6 46176 275.90
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + HtShoes + Leg, data = seatpos)
##
## Coefficients:
## (Intercept)          Age      HtShoes          Leg
##    456.2137       0.5998      -2.3023      -6.8297
```

d. Using the model chosen by AIC, it appears that for every increase of a unit of leg length, we can expect to see hipcenter decrease by -6.739. The model has a very similar multiple r-squared value to the original, though it has a much higher adjusted r-squared, 0.6533 to the original model's 0.6001.

```r
lmodAIC <- lm(hipcenter ~ Age + Ht + Leg, seatpos)
summary(lmodAIC)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Ht + Leg, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.715 -22.758  -4.102  21.394  60.576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 452.1976   100.9482   4.480 8.04e-05 ***
## Age           0.5807     0.3790   1.532   0.1347
## Ht           -2.3254     1.2545  -1.854   0.0725 .
## Leg          -6.7390     4.1050  -1.642   0.1099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.12 on 34 degrees of freedom
## Multiple R-squared:  0.6814, Adjusted R-squared:  0.6533
## F-statistic: 24.24 on 3 and 34 DF,  p-value: 1.426e-08
```

```r
z <- model.matrix(lmodAIC)

z0 <- apply(z, 2, mean)

predict(lmodAIC, new = data.frame(t(z0)), interval = "prediction", level=0.99)
```

```
##         fit      lwr       upr
## 1 -164.8849 -261.961 -67.80873
```

2.

a. Linear Regression with all predictors:

```r
data(fat, package="faraway")

##Values to remove every 10th observation starting at 1

fatseq <- seq(1, 252, by = 10)

fatTrain <- fat[-fatseq,]
fat10 <- fat[fatseq,]

rmse <- function(x,y){
  sqrt(mean((x-y)^2))
}

lmodF1 <- lm(siri ~ . -brozek -density, fatTrain)

rmse(lmodF1$fit, fatTrain$siri)
```

```
## [1] 1.406899
```

```r
pred <- predict(lmodF1, fat10)
y1 <- fat10$siri

rmse(pred, y1)
```

```
## [1] 1.946023
```

b. Linear regression with stepwise variable selection:

```r
lmStep <- step(lmodF1)
```

```
## Start:  AIC=186.31
## 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  447.4 184.32
## - neck     1       0.2  447.5 184.39
## - knee     1       0.2  447.5 184.39
## - age      1       0.3  447.6 184.45
## - wrist    1       1.4  448.7 185.02
## - height   1       1.6  449.0 185.13
## - ankle    1       2.9  450.2 185.76
## <none>                  447.3 186.31
## - biceps   1      10.7  458.1 189.66
## - abdom    1      16.1  463.5 192.31
## - forearm  1      18.5  465.8 193.47
```

```
## - chest    1      23.3  470.6 195.76
## - thigh    1      25.4  472.7 196.78
## - adipos   1      42.1  489.4 204.62
## - weight   1     576.0 1023.4 371.33
## - free     1    3385.3 3832.6 669.75
##
## Step:  AIC=184.32
## siri ~ age + weight + height + adipos + free + neck + chest +
##     abdom + thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - neck     1       0.2  447.5 182.39
## - knee     1       0.2  447.5 182.39
## - age      1       0.3  447.7 182.47
## - wrist    1       1.4  448.8 183.03
## - height   1       1.7  449.1 183.19
## - ankle    1       3.0  450.4 183.83
## <none>                  447.4 184.32
## - biceps   1      10.8  458.2 187.72
## - abdom    1      16.4  463.7 190.44
## - forearm  1      18.8  466.2 191.63
## - chest    1      24.8  472.1 194.50
## - thigh    1      27.1  474.4 195.59
## - adipos   1      43.6  491.0 203.34
## - weight   1     683.5 1130.8 391.90
## - free     1    3415.7 3863.0 669.54
##
## Step:  AIC=182.39
## siri ~ age + weight + height + adipos + free + chest + abdom +
##     thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - knee     1       0.2  447.7 180.50
## - age      1       0.2  447.8 180.52
## - wrist    1       1.3  448.8 181.03
## - height   1       1.7  449.2 181.23
## - ankle    1       3.3  450.8 182.07
## <none>                  447.5 182.39
## - biceps   1      10.7  458.2 185.74
## - abdom    1      16.4  463.9 188.54
## - forearm  1      18.7  466.2 189.66
## - chest    1      24.7  472.2 192.55
## - thigh    1      26.9  474.4 193.60
## - adipos   1      45.7  493.2 202.38
## - weight   1     688.4 1135.9 390.90
## - free     1    3464.1 3911.6 670.37
##
## Step:  AIC=180.5
## siri ~ age + weight + height + adipos + free + chest + abdom +
##     thigh + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - age      1       0.4  448.1 178.68
## - wrist    1       1.3  449.1 179.17
```

```
## - height   1       1.6  449.3 179.30
## - ankle    1       4.0  451.7 180.49
## <none>                  447.7 180.50
## - biceps   1      10.6  458.3 183.76
## - abdom    1      16.6  464.3 186.72
## - forearm  1      19.1  466.8 187.94
## - chest    1      24.7  472.4 190.62
## - thigh    1      32.1  479.8 194.15
## - adipos   1      48.9  496.6 201.94
## - weight   1     731.7 1179.4 397.41
## - free     1    3464.0 3911.7 668.37
##
## Step:  AIC=178.68
## siri ~ weight + height + adipos + free + chest + abdom + thigh +
##     ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - height   1       1.4  449.5 177.41
## - wrist    1       2.4  450.5 177.89
## - ankle    1       3.9  452.0 178.63
## <none>                  448.1 178.68
## - biceps   1      10.8  458.9 182.08
## - forearm  1      18.7  466.8 185.94
## - abdom    1      20.1  468.2 186.59
## - chest    1      25.1  473.2 188.99
## - thigh    1      33.4  481.5 192.95
## - adipos   1      49.4  497.5 200.31
## - weight   1     738.0 1186.1 396.68
## - free     1    3491.5 3939.6 667.97
##
## Step:  AIC=177.41
## siri ~ weight + adipos + free + chest + abdom + thigh + ankle +
##     biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - wrist    1       2.6  452.1 176.72
## - ankle    1       3.9  453.5 177.38
## <none>                  449.5 177.41
## - biceps   1      11.2  460.7 180.98
## - forearm  1      19.0  468.6 184.79
## - abdom    1      20.4  469.9 185.44
## - chest    1      25.3  474.9 187.81
## - thigh    1      32.1  481.6 190.99
## - adipos   1      79.2  528.7 212.09
## - weight   1     847.9 1297.4 414.96
## - free     1    3492.9 3942.4 666.14
##
## Step:  AIC=176.72
## siri ~ weight + adipos + free + chest + abdom + thigh + ankle +
##     biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## <none>                  452.1 176.72
## - ankle    1       6.1  458.2 177.74
```

```
## - biceps   1      12.9  465.1 181.09
## - forearm  1      22.1  474.2 185.50
## - abdom    1      23.4  475.5 186.12
## - chest    1      25.3  477.4 187.01
## - thigh    1      29.5  481.7 189.02
## - adipos   1      79.2  531.3 211.20
## - weight   1     847.4 1299.6 413.33
## - free     1    3709.0 4161.1 676.34
```

```r
rmse(lmStep$fitted.values, fatTrain$siri)
```

```
## [1] 1.414443
```

```r
predStep <- predict(lmStep, fat10)
rmse(predStep, y1)
```

```
## [1] 1.98911
```

c. Principal component regression:

```r
require(pls)
```

```
## Loading required package: pls

##
## Attaching package: 'pls'

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

```r
pcrmod <- pcr(siri ~ . -brozek -density, data = fatTrain)

pcrsme <- RMSEP(pcrmod, newdata = fat10)

## plot(pcrsme)

## Apropriate no. of components
which.min(pcrsme$val)
```

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

```r
pcrmod11 <- pcr(siri ~ . -brozek -density, data = fatTrain, ncomp = 11)

rmse(pcrmod11$fitted.values, fatTrain$siri)
```

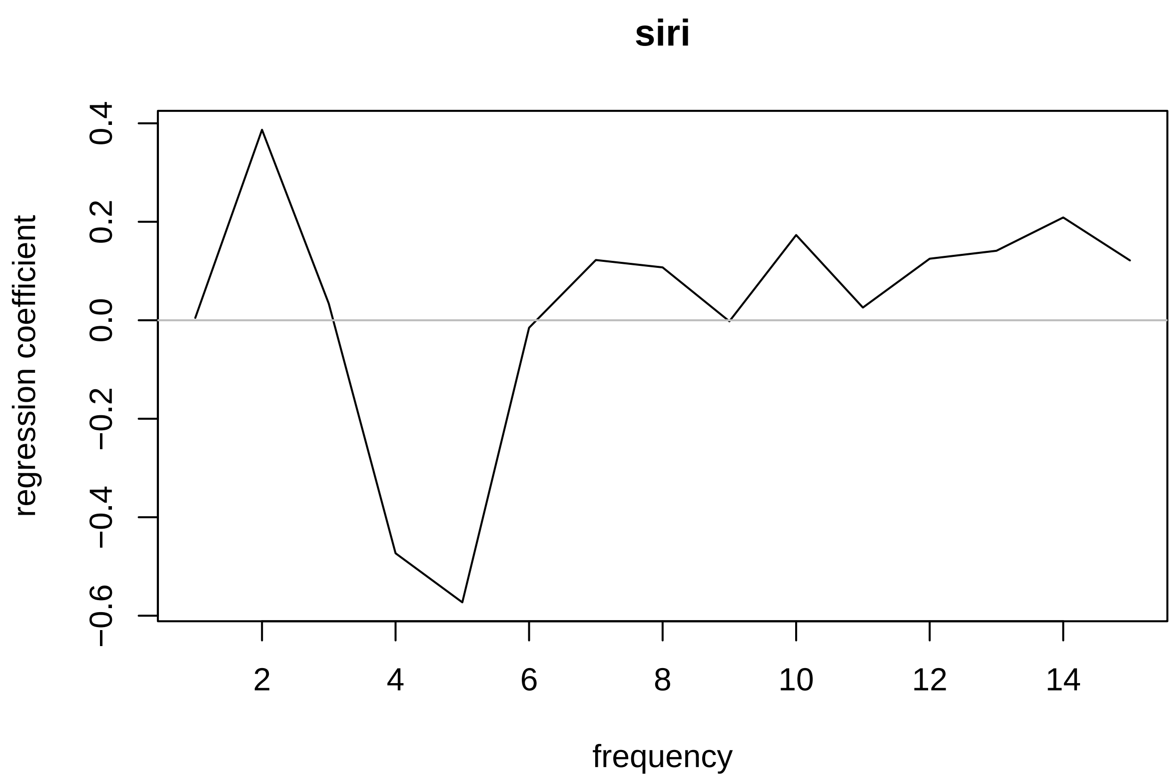
```
## [1] 2.971977
```

```r
rmse(predict(pcrmod11, fat10), fat10$siri)
```

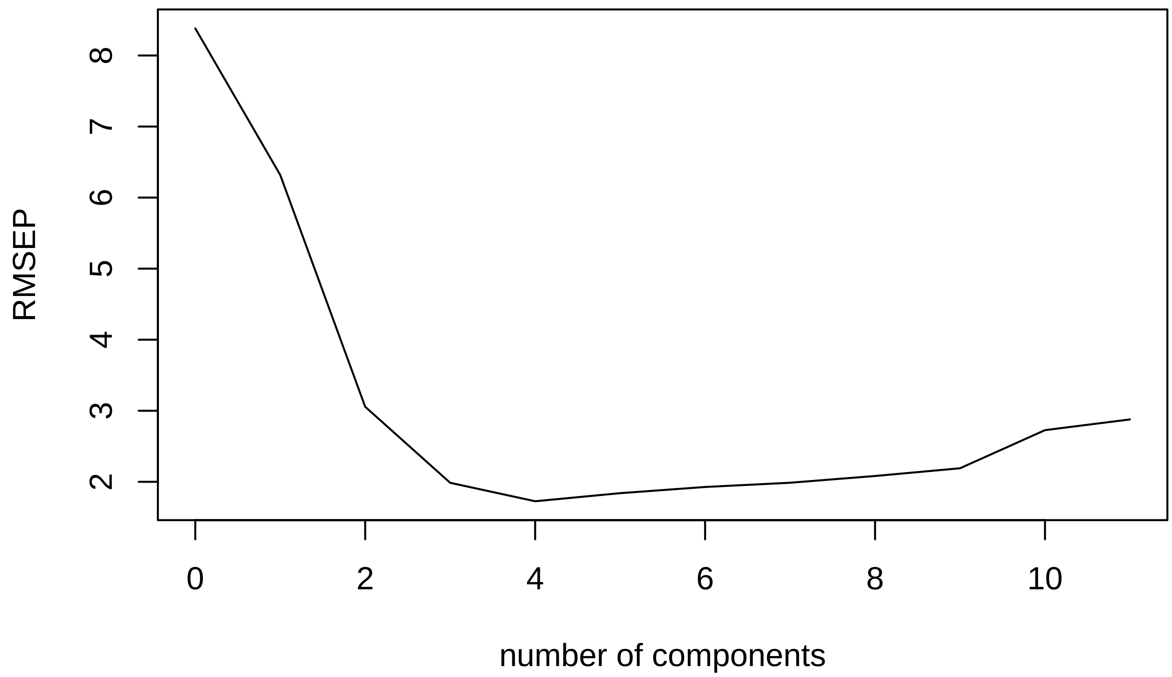
```
## [1] 2.973433
```

d. Partial least squares:

```r
set.seed(123)
plsmod <- plsr(siri ~ . -brozek -density, data = fatTrain, ncomp = 11, validation = "CV")

coefplot(plsmod, ncomp = 11, xlab = "frequency")
```




```r
plsCV <- RMSEP(plsmod, estimate = "CV")

plot(plsCV, main = "")
```




```r
which.min(plsCV$val) ## it appears that 5 is an appropriate ncomp value
```

```
## [1] 5
```

```r
ypred <- predict(plsmod, ncomp = 5)
rmse(ypred, fatTrain$siri)
```

```
## [1] 1.45939
```

```r
ytpred <- predict(plsmod, fat10, ncomp=5)
rmse(ytpred, fat10$siri)
```

```
## [1] 2.028371
```

e. Ridge regression:

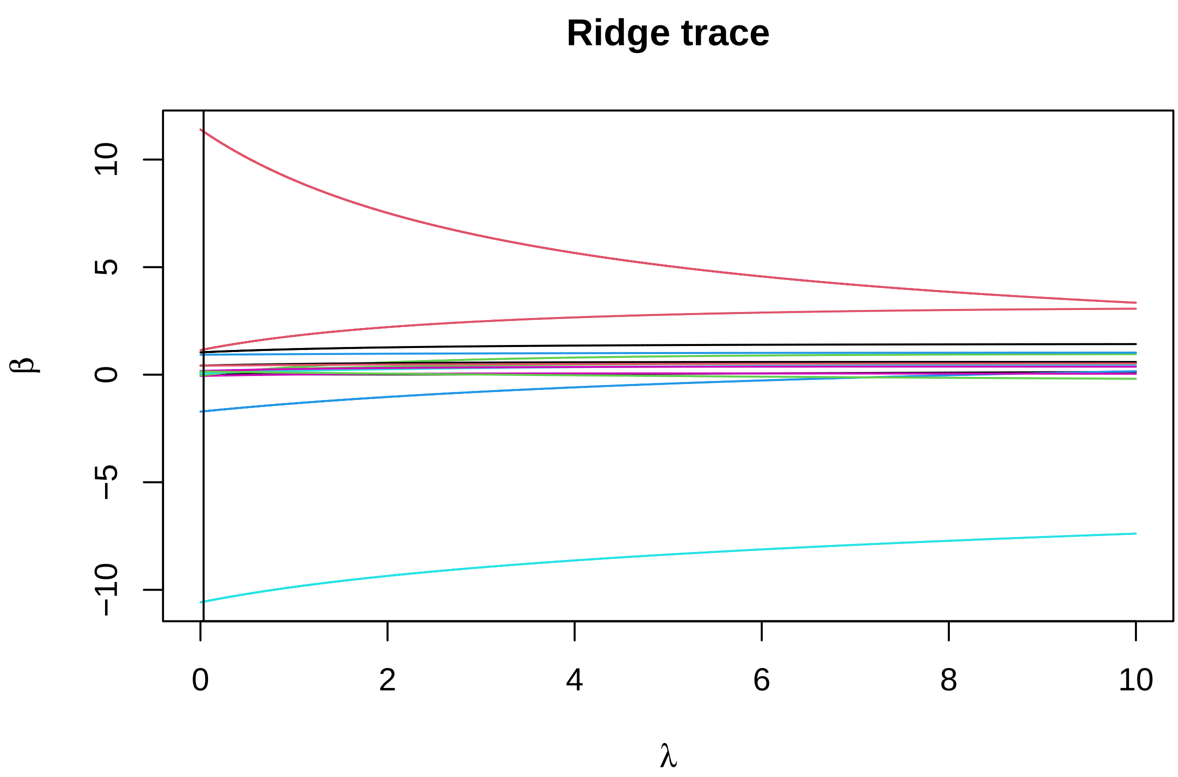
```r
require(MASS)
```

```
## Loading required package: MASS
```

```r
means <- apply(fatTrain[,4:18],2,mean)

fatMatrix <- as.matrix(sweep(fatTrain[,4:18],2,means))

test10 <- as.matrix(sweep(fat10[,4:18],2,means))

par(mfrow= c(1,1))

ysiri <- fatTrain$siri - mean(fatTrain$siri)

rgmod <- lm.ridge(ysiri ~ fatMatrix, lambda = seq(0, 10, 1e-4))

matplot(rgmod$lambda, t(rgmod$coef), type = "l", lty = 1, xlab = expression(lambda), ylab = expression(

select(rgmod)
```

```
## modified HKB estimator is 0.1113946
## modified L-W estimator is 0.4093012
## smallest value of GCV  at 0.0339
```

```r
abline(v=0.0339)
```




```r
rgyfit <- scale(fatMatrix, center=F, scale=rgmod$scales) %*% rgmod$coef[,468] + mean(fatTrain$siri)

rmse(rgyfit, fatTrain$siri)
```

```
## [1] 1.407043
```

```r
rgypred <-  scale(test10, center=F, scale=rgmod$scales) %*% rgmod$coef[,468] + mean(fatTrain$siri)

rmse(rgypred,fat10$siri)
```

```
## [1] 1.933964
```

Conclusion: We get solid results with both methods of linear regression, all predictors and stepwise variable selection. However, we do not get favorable results with the principal component regression. However, this could be further explored. We also have promising results from the partial least squares and ridge regression methods. It is unclear why the principal component regression performed returned poor results, but perhaps a different testing sample size would change the outcome.