Lab2 STAT515

Mykola Signayevskyy

2023-04-13

#Question:

#Consider the "Body Fat.csv" data set. Consider "% Body Fat" as the response variable. #Using this data set fit an appropriate model to predict "% Body Fat". #Needs to Justify your actions and interpret your results.

#Hints: #Remove the `Density` variable before the analysis #Include an interaction term (if possible and if make sense) #Include a non-linear term (if possible and if make sense) #Use any variable selection technique #Use cross validation method to select an appropriate model # Use residual analysis. #Also use your imagination

```
library(ISLR)
library(MASS)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
source("/Users/mykola/Desktop/STAT515/third_lesson/hw.R")
```

```
body_fat <- read.csv("/Users/mykola/Desktop/STAT515/lab2/Body Fat(5).csv")
head(body_fat)
```

```
##   Subject Density X.Body.Fat Age Weight Height Neck Chest Abdomen   Hip Thigh
## 1       1  1.0708       12.3  23 154.25  67.75 36.2  93.1    85.2  94.5  59.0
## 2       2  1.0853        6.1  22 173.25  72.25 38.5  93.6    83.0  98.7  58.7
## 3       3  1.0414       25.3  22 154.00  66.25 34.0  95.8    87.9  99.2  59.6
## 4       4  1.0751       10.4  26 184.75  72.25 37.4 101.8    86.4 101.2  60.1
## 5       5  1.0340       28.7  24 184.25  71.25 34.4  97.3   100.0 101.9  63.2
## 6       6  1.0512       20.9  24 210.25  74.75 39.0 104.5    94.4 107.8  66.0
##   Knee Ankle Biceps Forearm Wrist
## 1 37.3  21.9   32.0    27.4  17.1
## 2 37.3  23.4   30.5    28.9  18.2
## 3 38.9  24.0   28.8    25.2  16.6
## 4 37.3  22.8   32.4    29.4  18.2
## 5 42.2  24.0   32.2    27.7  17.7
## 6 42.0  25.6   35.7    30.6  18.8
```

```
summary(body_fat)
```

```
##     Subject          Density        X.Body.Fat         Age        
##  Min.   :  1.00   Min.   :0.995   Min.   : 0.00   Min.   :22.00  
##  1st Qu.: 63.75   1st Qu.:1.041   1st Qu.:12.47   1st Qu.:35.75  
##  Median :126.50   Median :1.055   Median :19.20   Median :43.00  
##  Mean   :126.50   Mean   :1.055   Mean   :19.16   Mean   :44.88  
##  3rd Qu.:189.25   3rd Qu.:1.070   3rd Qu.:25.30   3rd Qu.:54.00  
##  Max.   :252.00   Max.   :1.109   Max.   :47.50   Max.   :81.00  
##      Weight          Height          Neck           Chest      
##  Min.   :118.5   Min.   :64.00   Min.   :31.10   Min.   : 79.30  
##  1st Qu.:159.0   1st Qu.:68.25   1st Qu.:36.40   1st Qu.: 94.35  
##  Median :176.5   Median :70.00   Median :38.00   Median : 99.65  
##  Mean   :178.9   Mean   :70.31   Mean   :37.99   Mean   :100.82  
##  3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:39.42   3rd Qu.:105.38  
##  Max.   :363.1   Max.   :77.75   Max.   :51.20   Max.   :136.20  
##     Abdomen            Hip            Thigh           Knee      
##  Min.   : 69.40   Min.   : 85.0   Min.   :47.20   Min.   :33.00  
##  1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98  
##  Median : 90.95   Median : 99.3   Median :59.00   Median :38.50  
##  Mean   : 92.56   Mean   : 99.9   Mean   :59.41   Mean   :38.59  
##  3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92  
##  Max.   :148.10   Max.   :147.7   Max.   :87.30   Max.   :49.10  
##      Ankle           Biceps         Forearm          Wrist      
##  Min.   :19.10   Min.   :24.80   Min.   :21.00   Min.   :15.80  
##  1st Qu.:22.00   1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60  
##  Median :22.80   Median :32.05   Median :28.70   Median :18.30  
##  Mean   :23.02   Mean   :32.27   Mean   :28.66   Mean   :18.23  
##  3rd Qu.:24.00   3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80  
##  Max.   :29.60   Max.   :45.00   Max.   :34.90   Max.   :21.40
```

```
table(is.na(body_fat))
```

```
##
## FALSE
##  4032
```

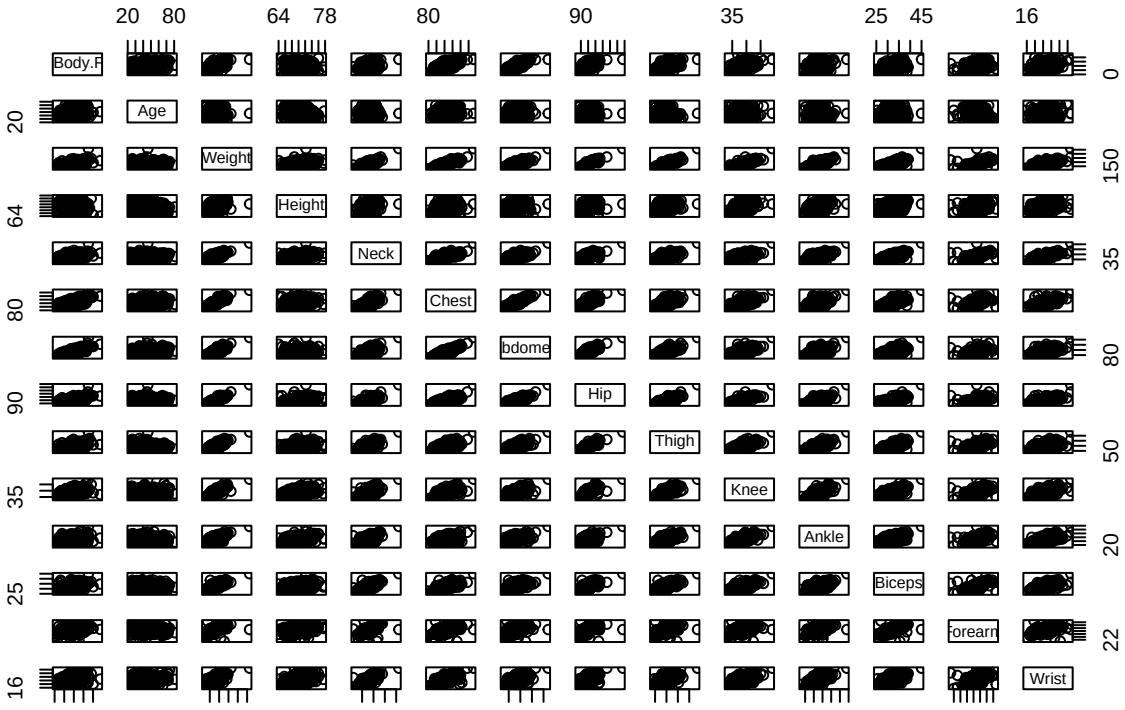
```
body_fat <- subset(body_fat, select = -c(Density))
```

```
model <- lm(X.Body.Fat ~ ., data = body_fat)
summary(model)
```

```
##
## Call:
## lm(formula = X.Body.Fat ~ ., data = body_fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.7048  -2.8819  -0.1974   3.0754  10.1943 
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -17.180743  22.698964  -0.757  0.44986  
## Subject      -0.002999   0.004060  -0.739  0.46081  
## Age           0.066075   0.033143   1.994  0.04734 *
```

```
## Weight       -0.084198   0.063288  -1.330  0.18467
## Height       -0.053832   0.180224  -0.299  0.76544
## Neck         -0.505502   0.237981  -2.124  0.03470 *
## Chest        -0.018672   0.103689  -0.180  0.85725
## Abdomen       0.954349   0.090681  10.524  < 2e-16 ***
## Hip          -0.215623   0.146547  -1.471  0.14252
## Thigh         0.255613   0.146659   1.743  0.08265 .
## Knee          0.050934   0.253532   0.201  0.84095
## Ankle        -0.017048   0.362784  -0.047  0.96256
## Biceps        0.164011   0.174622   0.939  0.34857
## Forearm       0.460991   0.199680   2.309  0.02182 *
## Wrist        -1.555907   0.557810  -2.789  0.00571 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.314 on 237 degrees of freedom
## Multiple R-squared:   0.75,  Adjusted R-squared:  0.7352
## F-statistic: 50.78 on 14 and 237 DF,  p-value: < 2.2e-16
```

```
pairs(~ X.Body.Fat + Age + Weight + Height + Neck + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Bicep
```



as I can see

```
model_i1 <- lm(`X.Body.Fat` ~ . + Abdomen*Hip*Age, data = body_fat) #I am trying to do logical in my op
summary(model_i1)
```

```
##
## Call:
## lm(formula = X.Body.Fat ~ . + Abdomen * Hip * Age, data = body_fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.6912 -3.0743 -0.2351  2.8249  9.3174
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.538e+02  1.347e+02   1.142  0.25462
## Subject         -3.369e-03  3.955e-03  -0.852  0.39519
## Age             -4.677e+00  2.974e+00  -1.573  0.11709
## Weight           4.600e-02  6.899e-02   0.667  0.50556
## Height          -4.539e-01  2.000e-01  -2.270  0.02414 *
## Neck            -4.904e-01  2.311e-01  -2.122  0.03491 *
## Chest           -1.474e-01  1.057e-01  -1.394  0.16453
## Abdomen          6.327e-02  1.432e+00   0.044  0.96480
## Hip             -1.583e+00  1.349e+00  -1.174  0.24174
## Thigh            1.371e-01  1.527e-01   0.898  0.37010
## Knee            -1.841e-01  2.533e-01  -0.727  0.46821
## Ankle            1.016e-01  3.530e-01   0.288  0.77376
## Biceps           1.453e-01  1.704e-01   0.853  0.39478
## Forearm          1.490e-01  2.065e-01   0.722  0.47123
## Wrist           -1.732e+00  5.479e-01  -3.161  0.00178 **
## Abdomen:Hip      9.787e-03  1.411e-02   0.694  0.48868
## Age:Abdomen      3.716e-02  3.131e-02   1.187  0.23651
## Age:Hip          5.185e-02  3.036e-02   1.708  0.08902 .
## Age:Abdomen:Hip -4.149e-04  3.111e-04  -1.334  0.18362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.18 on 233 degrees of freedom
## Multiple R-squared:  0.7693, Adjusted R-squared:  0.7514
## F-statistic: 43.15 on 18 and 233 DF,  p-value: < 2.2e-16
```

```
model_non_l <- lm(`X.Body.Fat` ~ . + Abdomen*Hip + I(Abdomen^2), data = body_fat)
summary(model_non_l)
```

```
##
## Call:
## lm(formula = X.Body.Fat ~ . + Abdomen * Hip + I(Abdomen^2), data = body_fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2358  -3.0031  -0.1262   2.8562   9.5162
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -62.194597  27.865109  -2.232 0.026559 *
## Subject       -0.003177   0.003962  -0.802 0.423490
## Age            0.073459   0.032431   2.265 0.024421 *
## Weight         0.023378   0.069015   0.339 0.735110
## Height        -0.391973   0.200483  -1.955 0.051752 .
## Neck          -0.467038   0.233177  -2.003 0.046333 *
## Chest         -0.139928   0.106446  -1.315 0.189946
## Abdomen        1.413103   0.364788   3.874 0.000139 ***
## Hip            0.986163   0.585671   1.684 0.093545 .
## Thigh          0.122571   0.149237   0.821 0.412298
## Knee          -0.085880   0.251459  -0.342 0.733013
## Ankle          0.054410   0.354609   0.153 0.878186
```

```
## Biceps         0.140313   0.171509   0.818 0.414122
## Forearm        0.228270   0.204730   1.115 0.265998
## Wrist         -1.781652   0.547916  -3.252 0.001316 **
## I(Abdomen^2)   0.003406   0.004596   0.741 0.459358
## Abdomen:Hip   -0.011704   0.005925  -1.975 0.049407 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.211 on 235 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7478
## F-statistic:  47.5 on 16 and 235 DF,  p-value: < 2.2e-16
```

```
library(leaps)


set.seed (1) # reproducibility

#data splitting 50%,50%
train=sample(c(TRUE,FALSE), nrow(body_fat),rep=TRUE)
test=(!train)
```

```
regfit.best=regsubsets(X.Body.Fat~.,data=body_fat[train,], nvmax =14)
```

```
test.mat=model.matrix(X.Body.Fat~.,data=body_fat[test,])
```

```
(val.errors=rep(NA,14))
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
for(i in 1:14){
  coefi=coef(regfit.best,id=i)
  pred=test.mat[,names(coefi)]%*%coefi
  val.errors[i]=mean((body_fat$X.Body.Fat[test]-pred)^2)
}

#MSE values for all 19 models.
val.errors
```

```
##  [1] 27.06001 24.02319 25.54736 25.57291 25.96166 24.49260 24.57001 24.84181
##  [9] 24.28397 24.46956 24.18193 24.10081 24.11999 24.09251
```

```
cor(body_fat[, -1])
```

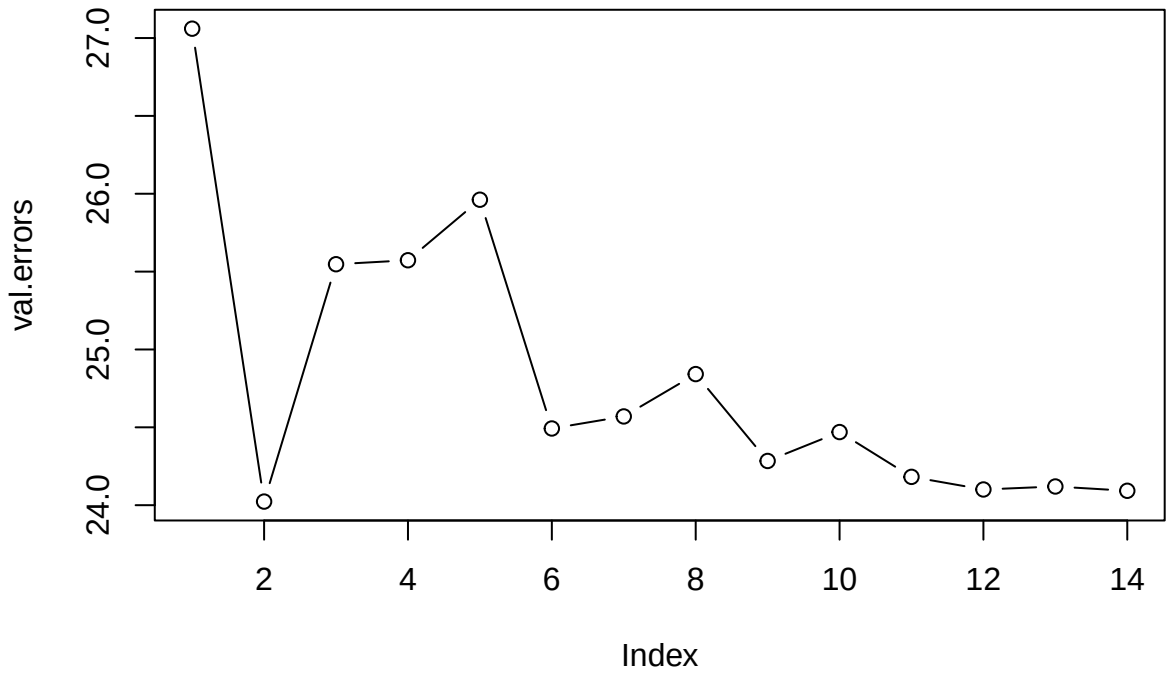
```
##             X.Body.Fat         Age      Weight      Height      Neck     Chest
## X.Body.Fat  1.00000000  0.29020273  0.61291560 -0.02555223 0.4907156 0.7029470
## Age         0.29020273  1.00000000 -0.01274609 -0.24521233 0.1135052 0.1764497
## Weight      0.61291560 -0.01274609  1.00000000  0.48688800 0.8307162 0.8941905
## Height     -0.02555223 -0.24521233  0.48688800  1.00000000 0.3211409 0.2268286
## Neck        0.49071562  0.11350519  0.83071622  0.32114085 1.0000000 0.7848350
## Chest       0.70294699  0.17644968  0.89419052  0.22682861 0.7848350 1.0000000
## Abdomen     0.81403228  0.23040942  0.88799494  0.18976623 0.7540774 0.9158277
## Hip         0.62566891 -0.05033212  0.94088412  0.37210602 0.7349579 0.8294199
## Thigh       0.56131873 -0.20009576  0.86869354  0.33855758 0.6956973 0.7298586
## Knee        0.50853827  0.01751569  0.85316739  0.50050052 0.6724050 0.7194964
## Ankle       0.32155180 -0.14809290  0.75183533  0.46921335 0.5879814 0.5925623
## Biceps      0.49358919 -0.04116212  0.80041593  0.31850749 0.7311459 0.7279075
## Forearm     0.36223617 -0.08505555  0.63030143  0.32202734 0.6236603 0.5801727
```

```
## Wrist       0.34655882  0.21353062  0.72977489  0.39777960 0.7448264 0.6601623
##               Abdomen         Hip       Thigh       Knee      Ankle      Biceps
## X.Body.Fat 0.8140323  0.62566891  0.5613187 0.50853827  0.3215518  0.49358919
## Age        0.2304094 -0.05033212 -0.2000958 0.01751569 -0.1480929 -0.04116212
## Weight     0.8879949  0.94088412  0.8686935 0.85316739  0.7518353  0.80041593
## Height     0.1897662  0.37210602  0.3385576 0.50050052  0.4692133  0.31850749
## Neck       0.7540774  0.73495788  0.6956973 0.67240498  0.5879814  0.73114592
## Chest      0.9158277  0.82941992  0.7298586 0.71949640  0.5925623  0.72790748
## Abdomen    1.0000000  0.87406618  0.7666239 0.73717888  0.5666402  0.68498272
## Hip        0.8740662  1.00000000  0.8964098 0.82347262  0.6914275  0.73927252
## Thigh      0.7666239  0.89640979  1.0000000 0.79917030  0.6896289  0.76147745
## Knee       0.7371789  0.82347262  0.7991703 1.00000000  0.7487857  0.67870883
## Ankle      0.5666402  0.69142749  0.6896289 0.74878572  1.0000000  0.58390226
## Biceps     0.6849827  0.73927252  0.7614774 0.67870883  0.5839023  1.00000000
## Forearm    0.5033161  0.54501412  0.5668422 0.55589819  0.5346917  0.67825513
## Wrist      0.6198324  0.63008954  0.5586848 0.66450729  0.6814612  0.63212642
##               Forearm     Wrist
## X.Body.Fat  0.36223617 0.3465588
## Age        -0.08505555 0.2135306
## Weight      0.63030143 0.7297749
## Height      0.32202734 0.3977796
## Neck        0.62366027 0.7448264
## Chest       0.58017273 0.6601623
## Abdomen     0.50331609 0.6198324
## Hip         0.54501412 0.6300895
## Thigh       0.56684218 0.5586848
## Knee        0.55589819 0.6645073
## Ankle       0.53469174 0.6814612
## Biceps      0.67825513 0.6321264
## Forearm     1.00000000 0.5855883
## Wrist       0.58558825 1.0000000
```

```
which.min(val.errors)
```

```
## [1] 2
```

```
plot(val.errors,type = 'b')
```




```
coef(regfit.best ,4)
```

```
## (Intercept)         Age      Height     Abdomen       Wrist
##   9.2660590   0.0612324  -0.5686116   0.7741765  -1.3612614
```

10-Fold Cross Validation

```
k <- 10 # 10-fold cross-validation
set.seed(1)
```

```
folds <- sample(1:k, nrow(body_fat), replace = TRUE)
```

```
(cv.errors <- matrix(NA, k, 14, dimnames = list(NULL, paste(1:14))))
```

```
##       1  2  3  4  5  6  7  8  9 10 11 12 13 14
##  [1,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [2,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [3,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [4,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [5,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [6,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [7,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [8,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [9,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [10,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
predict.regsubsets = function(object, newdata, id, ...){
  form=as.formula(object$call[[2]])
  mat=model.matrix(form, newdata)
  coefi=coef(object, id=id)
  xvars=names(coefi)
  mat[,xvars]%*%coefi
}
```

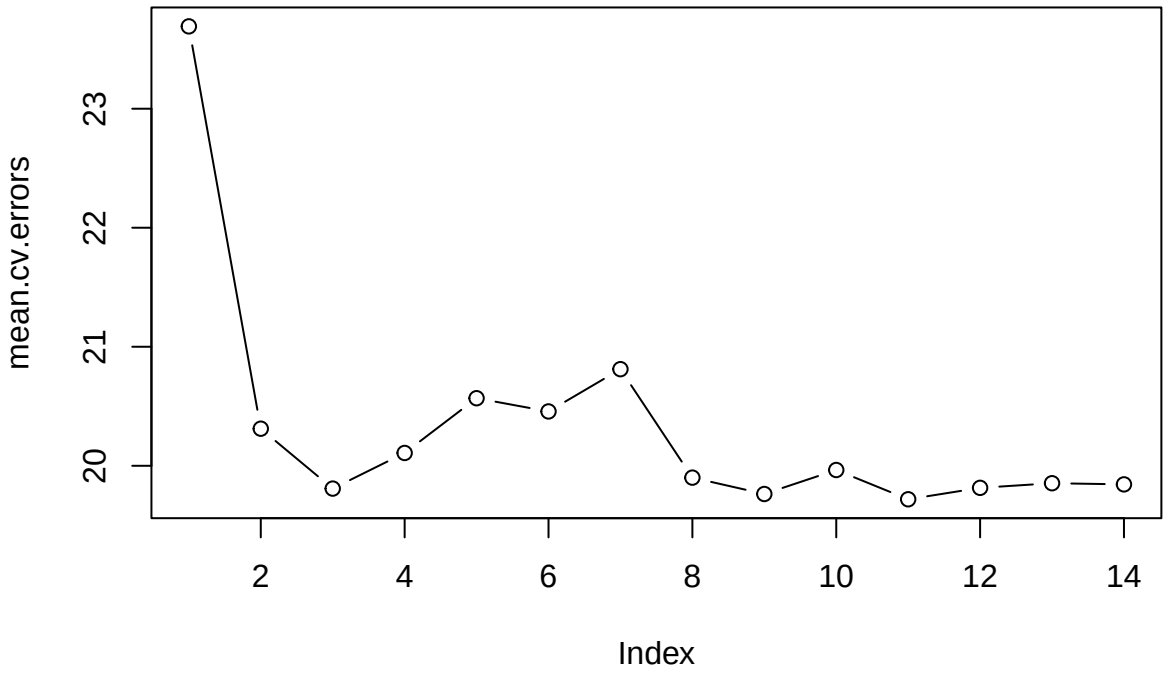
```r
for(j in 1:k){
  best.fit=regsubsets(X.Body.Fat~., data=body_fat[folds!=j,], nvmax=14)

  for(i in 1:14){
    pred = predict(best.fit, body_fat[folds==j,], id=i)
    cv.errors[j,i] = mean( (body_fat$X.Body.Fat[folds==j]-pred)^2 )
  }
}
```

```r
# Column average
mean.cv.errors=apply(cv.errors, 2, mean)
which.min(mean.cv.errors)
```

```
## 11
## 11
```

```r
plot(mean.cv.errors ,type="b")
```




```r
reg.best=regsubsets (X.Body.Fat~.,data=body_fat , nvmax=14)
coef(reg.best ,11) ## full data set.
```

```
##   (Intercept)       Subject           Age        Weight        Height
## -19.109082654  -0.002986554   0.067648852  -0.088570116  -0.033857319
##          Neck       Abdomen           Hip         Thigh        Biceps
##  -0.509398517   0.948498771  -0.207805736   0.269520034   0.161784051
##       Forearm         Wrist
##   0.459306969  -1.548933211
```

```r
reg.best=regsubsets (X.Body.Fat~.,data=body_fat , nvmax=14)
coef(reg.best ,3)
```

```
## (Intercept)      Weight     Abdomen       Wrist
## -27.9516417  -0.1145224   0.9773890  -1.2541221
```

First, I checked the logical permutations in my opinion, such as abdomen circumference and hip circumference. I also checked the permutations for abdomen circumference, hip circumference, and age. Because as people age, their overall waist circumference starts to increase due to less mobility. I saw that these crossovers have no serious effect on the model, while the Abdomen value itself has a huge weight for the body fat percentage dependent value. So I decided to move on to selecting variables for the model and finding the best model. I decided to start with the Validation Set Approach to find the best model. The results were good, because it turned out that for the model it's best to use 4 variables: Weight Abdomen Biceps Wrist, which means that this model can be easily explained.

After this, I used k-Fold Cross-Validation, which showed completely different results. This time the best number of variables for the model is 11. This is quite a lot and difficult to explain, even though the mean squared error is below 20. I noticed on the graph that the model with 3 variables has almost the same error value, which in my opinion is the best possible variant.

Also please note that when you try to knit file in phd the graph for Validation Set Approach is different from what I show in the code. When I run the code I get a graph with 4 best values, but when the file is knited it shows 2. I attach a screenshot to this file.