finalproject

phantomOfLaMancha

3/26/2021

```r
get.reduced.model = function(model, i){
  # convenient helper to return the new model with ith feature removed
  # i can be vector or number

  # first column of data will be response variable, other columns are features of original
  # model, intercept wouldn't appear here as a feature
  data = model$model
  r = nrow(data)
  c = ncol(data)

  # special case if there is only 1 feature left
  if(c==2){
    return(lm(data[1:r,1]~1))
  }

  # we shouldn't receive a model with only intercept
  if(c==1){
    stop("get.reduced.model() recieved a model with intercept only")
  }

  # explanatory variable
  names = colnames(data)[2:c]
  # response variable
  yname = colnames(data)[1]
  formu = as.formula( paste(yname, "~", paste( names[-i], collapse = "+")))
  # new model
  m =  lm(formu , data=data)
  return(m)
}

## note: right now this function could only do 10 fold

get_col <- function(mat,i,j, breaks, cols=NULL, palette="Blues") {
    if (is.null(cols)) {
        cols <- brewer.pal(length(breaks)+1, palette)}
    val <- 1
    for (b in breaks) {
      if (is.na(mat [i,j])){
        val <- 0
      }
      else if (mat[i,j] > b) {
            val <- val + 1}
            }
    cols[val]
    }

require(RColorBrewer)
```

```
## Loading required package: RColorBrewer
```

```
col_areas <- function(matrix,
                                                 breaks=NULL,
                                                 cols=NULL,
                                                 palette="Blues",
                                                 xlab="West    <------------>   East",
                                                 ylab="South   <------------>  North",
                                                 ...){
    if (is.null(breaks)) {
            breaks <- unique(fivenum(matrix))}

  plot(c(0, 100*ncol(matrix)),
            c(0, 100*nrow(matrix)), frame.plot=TRUE,
            type="n",
            xlab=xlab,
            ylab=ylab, axes=FALSE, ...)

  nr <- nrow(matrix)
  nc <- ncol(matrix)
    for (i in 1:nr) {
        for (j in 1:nc) {
            rect((j-1)*100,
                 (nr-i+1)*100,
                 j*100,
                 (nr-i)*100,
                 border=NA,
                 col=get_col(matrix,i,j,breaks,cols,palette))
                 }
               }
}
```

understanding our polulation:

```
library("eikosograms")
```

```
## Warning: package 'eikosograms' was built under R version 4.0.4
```

```
library("venneuler")
```

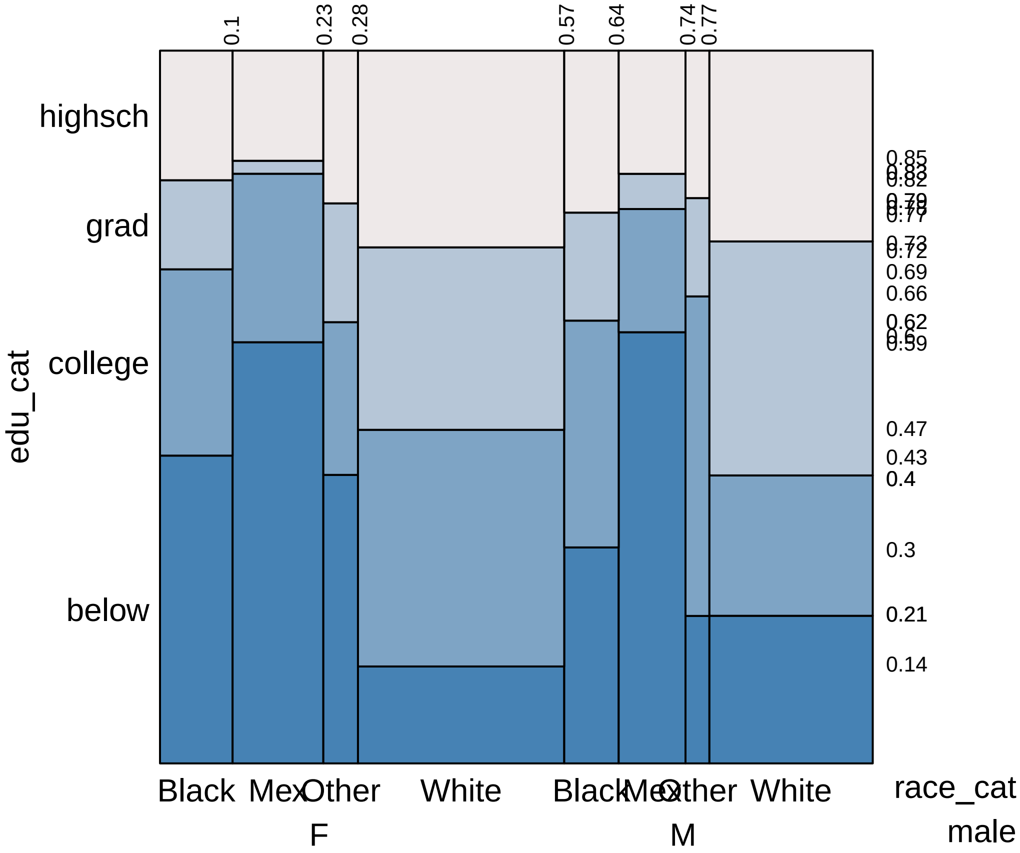
```
## Warning: package 'venneuler' was built under R version 4.0.3
```

```
## Loading required package: rJava
```

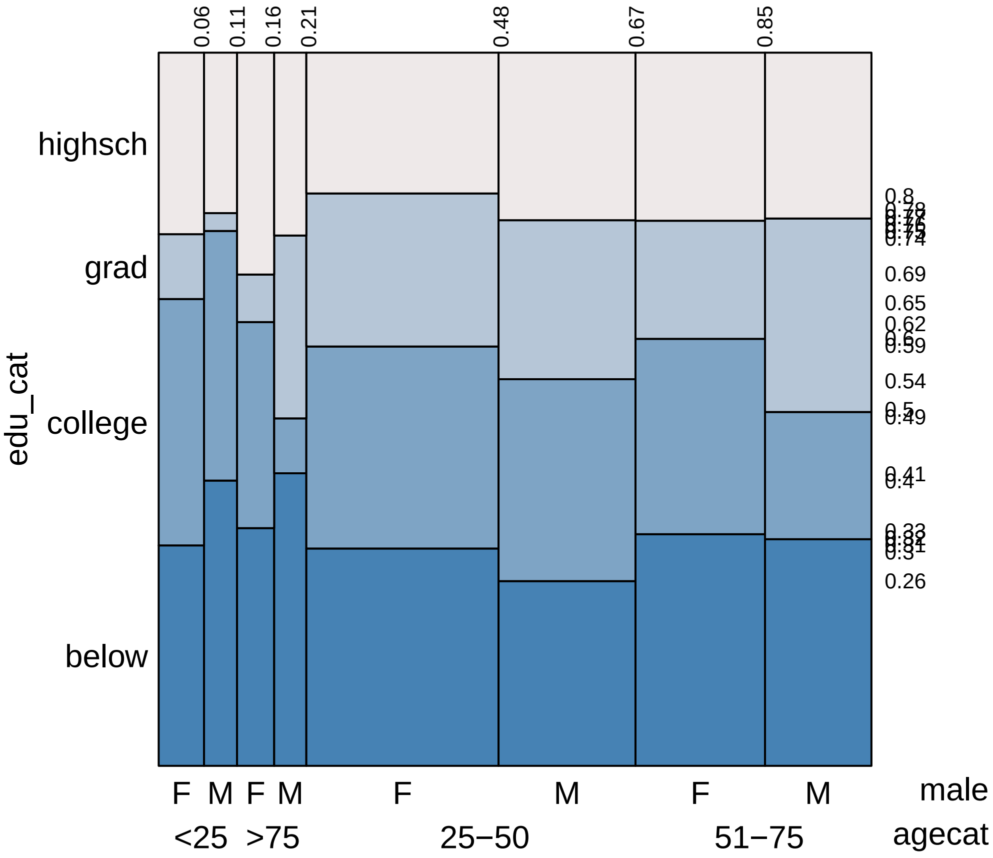
```
## Warning: package 'rJava' was built under R version 4.0.3
```

```
data = read.csv("pollutants.csv")

# change factor features to reasonable names

ind = data$male == 1
```

```r
data$male[ind] = "M"
data$male[!ind] = "F"
data$agecat = ceiling(data$ageyrs/25 )
agecat = c("<25","25-50","51-75",">75")

for (i in 1:4){
  ind = data$agecat == i
  data$agecat[ind] = agecat[i]
}


edu=c("below", "highsch", "college","grad")
for (i in 1:4){
  ind = data$edu_cat == i
  data$edu_cat[ind] = edu[i]
}

race=c("Other", "Mex", "Black","White")
for (i in 1:4){
  ind = data$race_cat == i
  data$race_cat[ind] = race[i]
}




eikos(edu_cat~ race_cat + male ,data=data)
```



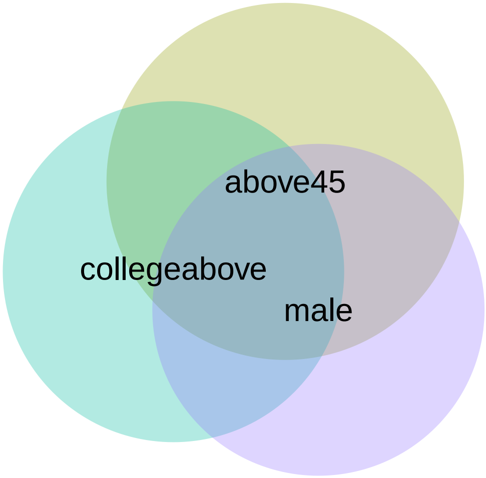
```
eikos(edu_cat~ male+agecat ,data=data)
```



```
# look at intersection

# note surface of above 45 should be approximately half of surface of total population

collegeabove = which( (data$edu_cat == "college") + (data$edu_cat == "grad") ==1 )
collegeabove.names = rep("collegeabove", length(collegeabove ))

white= which( data$race_cat == "White" )
white.names = rep("White", length(white))

median(data$ageyrs)
```

```
## [1] 46
```

```
above45 = which(data$ageyrs>45)
above45.names= rep("above45", length(above45))

male = which(data$male == "M")
male.names = rep("male", length(male))

female = which(data$male == "F")
female.names = rep("female", length(female))

subjectinfo = c(above45, collegeabove, male)
names = c(above45.names , collegeabove.names, male.names)
ven = venneuler(data.frame(elements = subjectinfo, sets=names))
plot(ven)
```



```r
# get rid of the agecat data we added
if (colnames(data)[ ncol(data)] == "agecat"){
  data = data[,-ncol(data)]
}
```

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.0.4

## Loading required package: Matrix

## Loaded glmnet 4.1-1
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.4

## Loading required package: carData
```

```r
data = read.csv("pollutants.csv")

# the index does not really mean anything
data = data[,-1]

nTotal = nrow(data)

#change some feature to factor type
data$race_cat = factor(data$race_cat)
data$edu_cat = factor(data$edu_cat)
```

```
data$male = factor(data$male)
data$smokenow= factor(data$smokenow)

data.train = data[1:700,]
data.test = data[701:nTotal,]
runif(1)
```

```
## [1] 0.9125422
```

correlation between features

```
model = lm(length~. , data=data)
#original vif

vif(model)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## POP_PCB1            33.044120  1        5.748401
## POP_PCB2            34.281125  1        5.855009
## POP_PCB3             9.351143  1        3.057964
## POP_PCB4            31.742239  1        5.634025
## POP_PCB5            59.896895  1        7.739308
## POP_PCB6            11.386658  1        3.374412
## POP_PCB7             4.870075  1        2.206825
## POP_PCB8            12.982575  1        3.603134
## POP_PCB9            12.441595  1        3.527264
## POP_PCB10            6.020678  1        2.453707
## POP_PCB11            4.725769  1        2.173883
## POP_dioxin1          5.276251  1        2.297009
## POP_dioxin2          5.413132  1        2.326614
## POP_dioxin3          4.398509  1        2.097262
## POP_furan1           6.154213  1        2.480769
## POP_furan2           6.195336  1        2.489043
## POP_furan3           4.464346  1        2.112900
## POP_furan4           1.821809  1        1.349744
## whitecell_count      1.548380  1        1.244339
## lymphocyte_pct   12250.336528  1      110.681238
## monocyte_pct       726.843372  1       26.960033
## eosinophils_pct 15071.561945  1      122.766290
## basophils_pct      867.412798  1       29.451873
## neutrophils_pct     37.984114  1        6.163125
## BMI                  1.263662  1        1.124127
## edu_cat              1.543109  3        1.074978
## race_cat             2.052848  3        1.127352
## male                 1.350324  1        1.162034
## ageyrs               3.238631  1        1.799620
## yrssmoke             2.204139  1        1.484634
## smokenow             4.006708  1        2.001676
## ln_lbxcot            3.963407  1        1.990831
```

```
t1=colnames( model$model)


while (TRUE) {
  score = vif(model)
  if (max(score) <10){
    break
  }
  ind = which.max(score)
  # this is safe with factor data type
  model = get.reduced.model(model, ind)
}
# reduced model vif
vif(model)
```

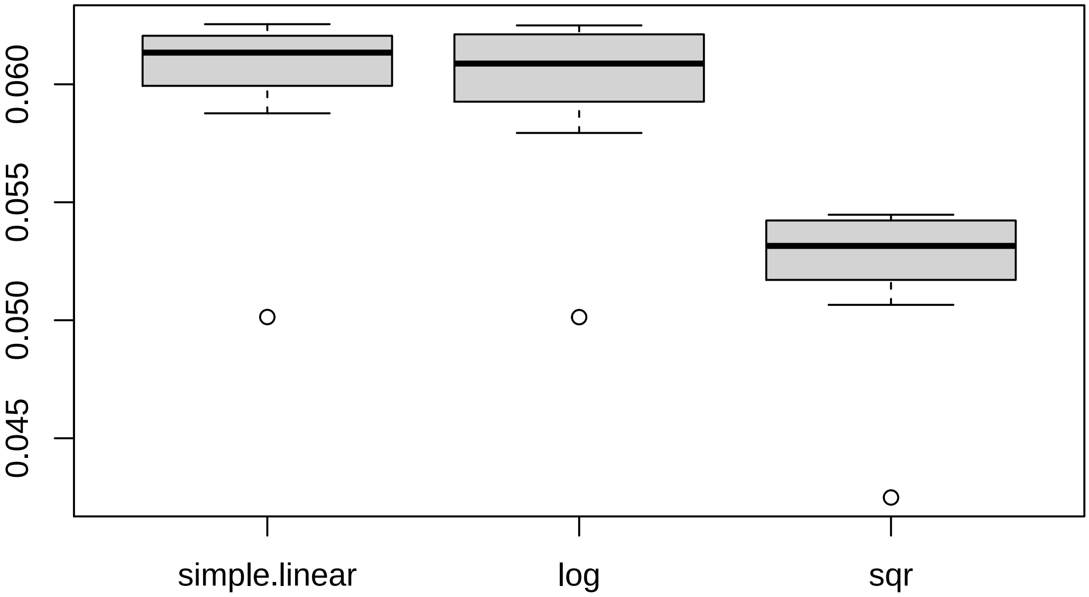
```
##                     GVIF Df GVIF^(1/(2*Df))
## POP_PCB3        5.310340  1        2.304417
## POP_PCB6        9.083828  1        3.013939
## POP_PCB7        4.686485  1        2.164829
## POP_PCB8        5.894052  1        2.427767
## POP_PCB9        7.640480  1        2.764142
## POP_PCB10       5.149483  1        2.269247
## POP_PCB11       4.210120  1        2.051858
## POP_dioxin1     5.184345  1        2.276916
## POP_dioxin2     5.275271  1        2.296796
## POP_dioxin3     4.311410  1        2.076394
## POP_furan1      6.000097  1        2.449509
## POP_furan2      6.154621  1        2.480851
## POP_furan3      4.412739  1        2.100652
## POP_furan4      1.812793  1        1.346400
## whitecell_count 1.533642  1        1.238403
## lymphocyte_pct  1.370966  1        1.170882
## monocyte_pct    1.255543  1        1.120510
## basophils_pct   1.097132  1        1.047441
## neutrophils_pct 1.083675  1        1.040997
## BMI             1.257562  1        1.121411
## edu_cat         1.498239  3        1.069704
## race_cat        2.012804  3        1.123657
## male            1.345703  1        1.160045
## ageyrs          3.224432  1        1.795670
## yrssmoke        2.147610  1        1.465473
## smokenow        3.967106  1        1.991759
## ln_lbxcot       3.946223  1        1.986510
```

```
t2=colnames( model$model)

setdiff(t1,t2)
```

```
## [1] "POP_PCB1"        "POP_PCB2"        "POP_PCB4"        "POP_PCB5"
## [5] "eosinophils_pct"
```

does one feature explain the model?

```
Xfull = lm(length~., data=data)$model

res = matrix(0, nrow = (ncol(Xfull)), ncol = 3)


for(c in 2:ncol(Xfull)){
  model = lm(data$length~Xfull[,c])
  res[c,1] = mean(model$residuals^2)
  if(! is.factor(Xfull[,c])){
    modelpower2 = lm(data$length~poly( Xfull[,c], 2))
    modellog = lm(log(data$length)~ Xfull[,c])
    res[c,2] = mean(modelpower2$residuals^2)
    res[c,3] = mean(modellog$residuals^2)
  }
  #res[c,3] = mean(modelpower2$residuals^2)
}

removezero = function(v){
  v[v==0] = NA
  v
}

box = list(simple.linear=removezero(res[,1]), log=removezero(res[,2]), sqr=removezero(res[,3]) )

boxplot(box)
```



```
which.min(removezero(res[,1]))
```

```
## [1] 30
```

```
which.min(removezero(res[,2]))
```

```
## [1] 30
```

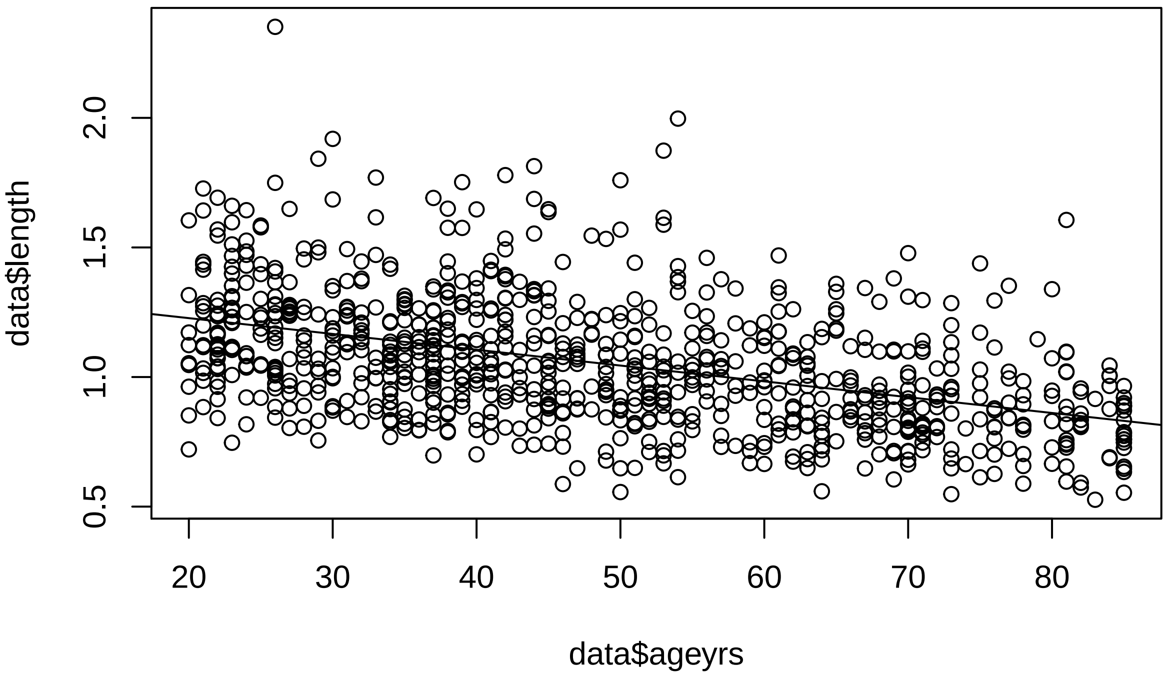
```r
which.min(removezero(res[,3]))
```

```
## [1] 30
```

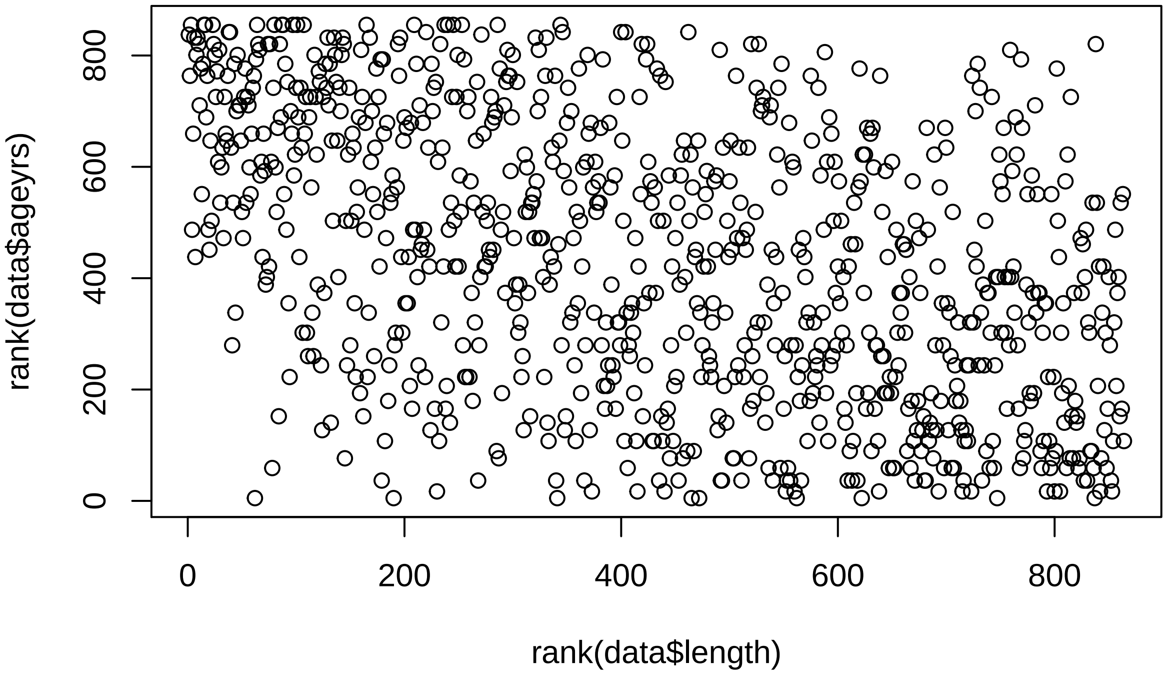
```r
colnames(Xfull)[30]
```

```
## [1] "ageyrs"
```

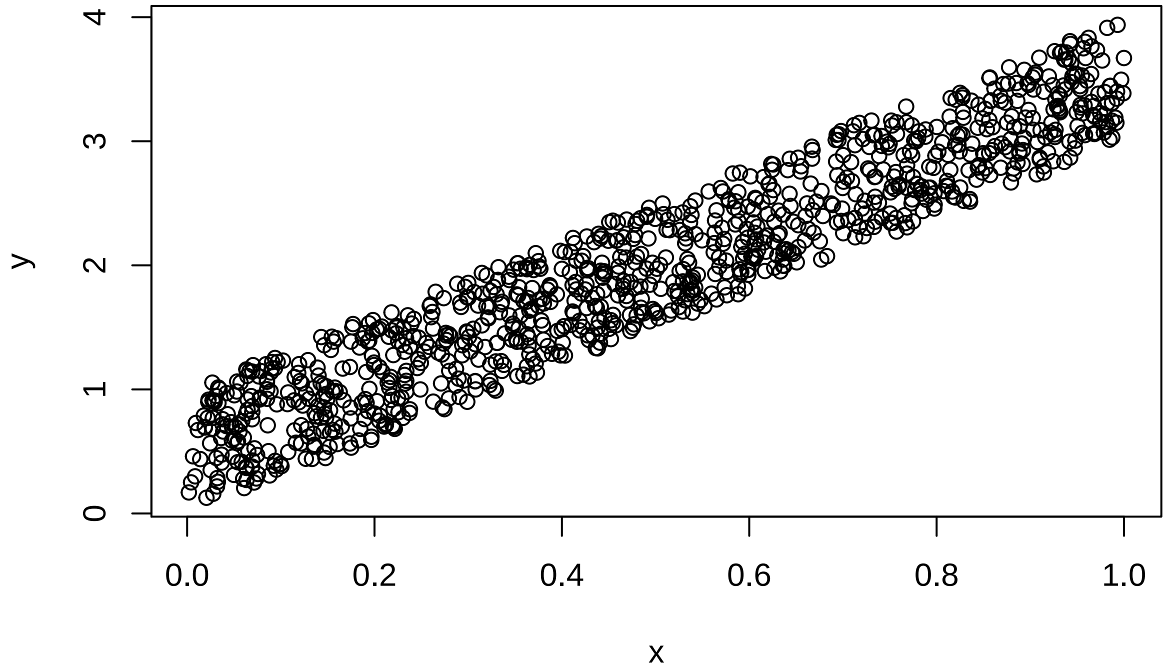
```r
simplelinear = lm(length~ageyrs, data=data)
plot(data$ageyrs, data$length)
abline(simplelinear$coefficients)
```



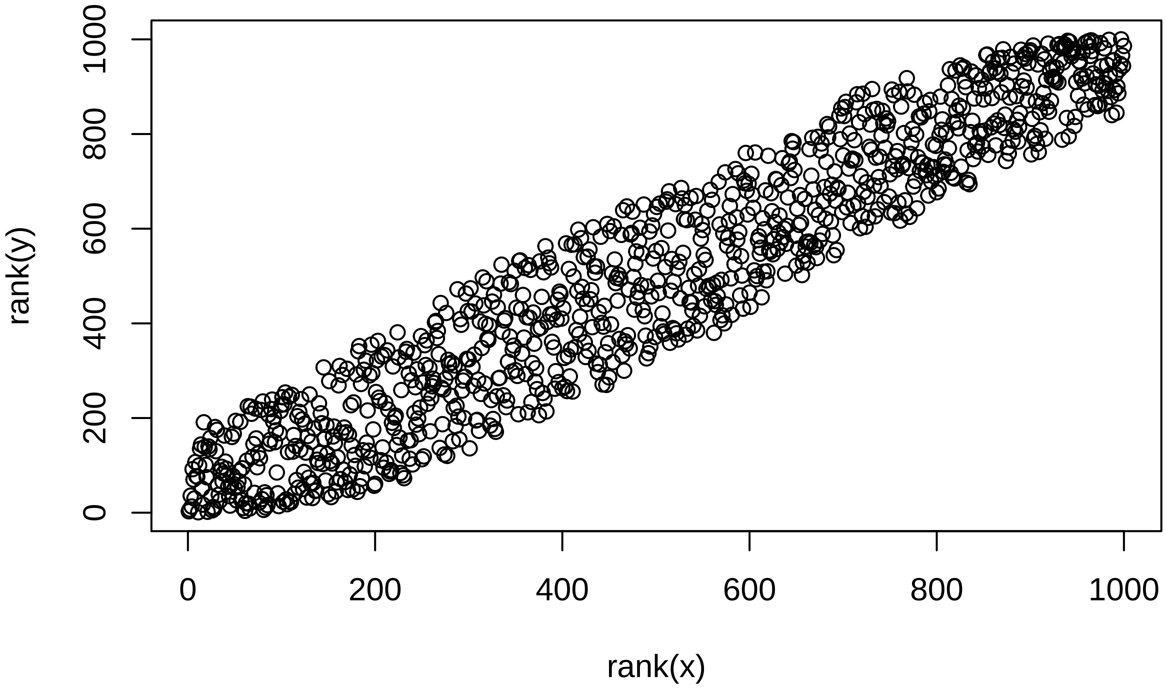
```r
plot(rank(data$length) , rank( data$ageyrs))
```



```
x=runif(1000)
y=3*x+runif(1000)
plot(y~x)
```



```
plot(rank(y)~rank(x))
```



```r
rank(c(15,1,3,6,4))
```

```
## [1] 5 1 2 4 3
```

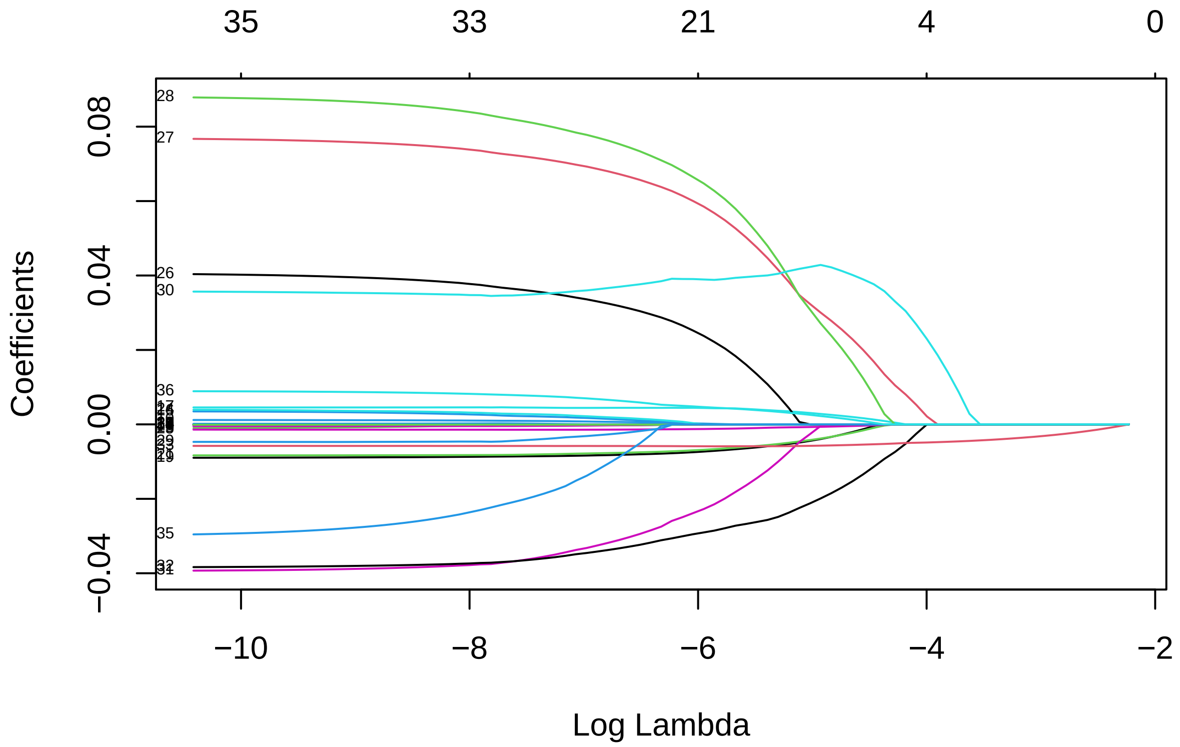
seems there is a linear relationship but looks insufficient.

Also seems sqr or log does not do exponentially better here

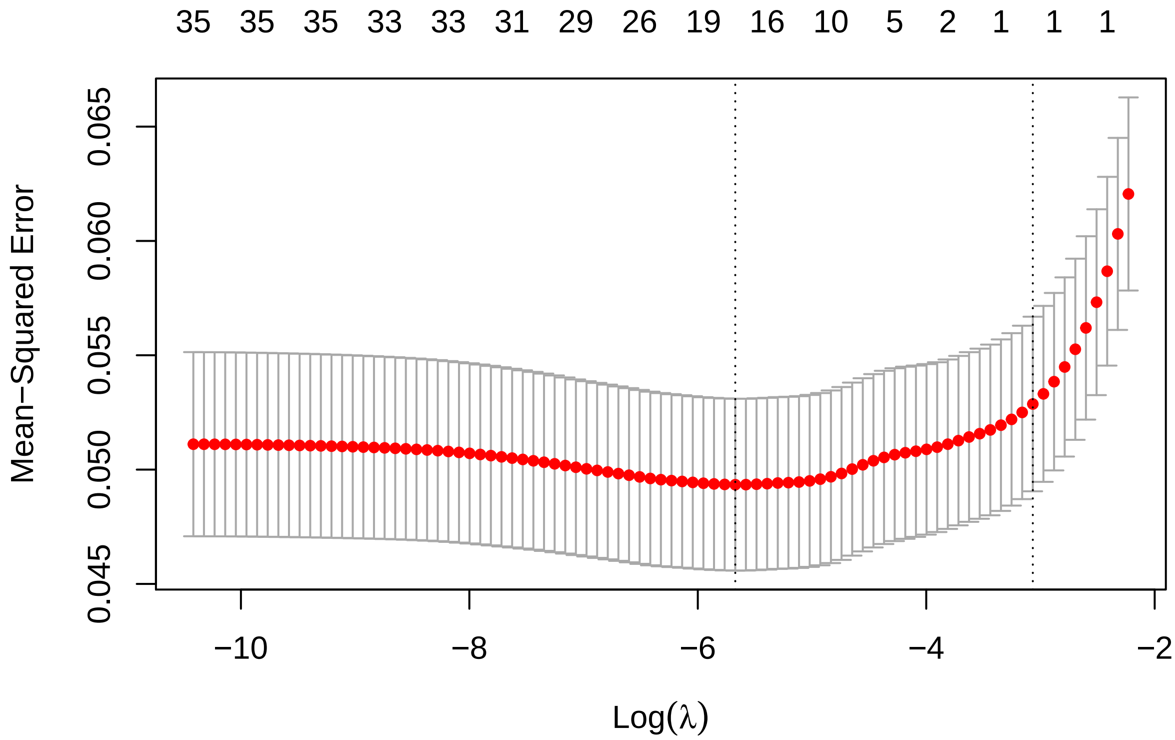
what about 2 features

best model based on lasso/ridge

```r
### LASSO
## fit models
M = model.matrix(lm(length~., data=data))
y_train = data$length[1:700]
X_train = M[1:700,-1]
y_test= data$length[701:nTotal]
X_test= M[701:nTotal,-1]

M_lasso <- glmnet(x=X_train,y=y_train,alpha = 1)
####

####
## plot paths
plot(M_lasso,xvar = "lambda",label=TRUE)
```



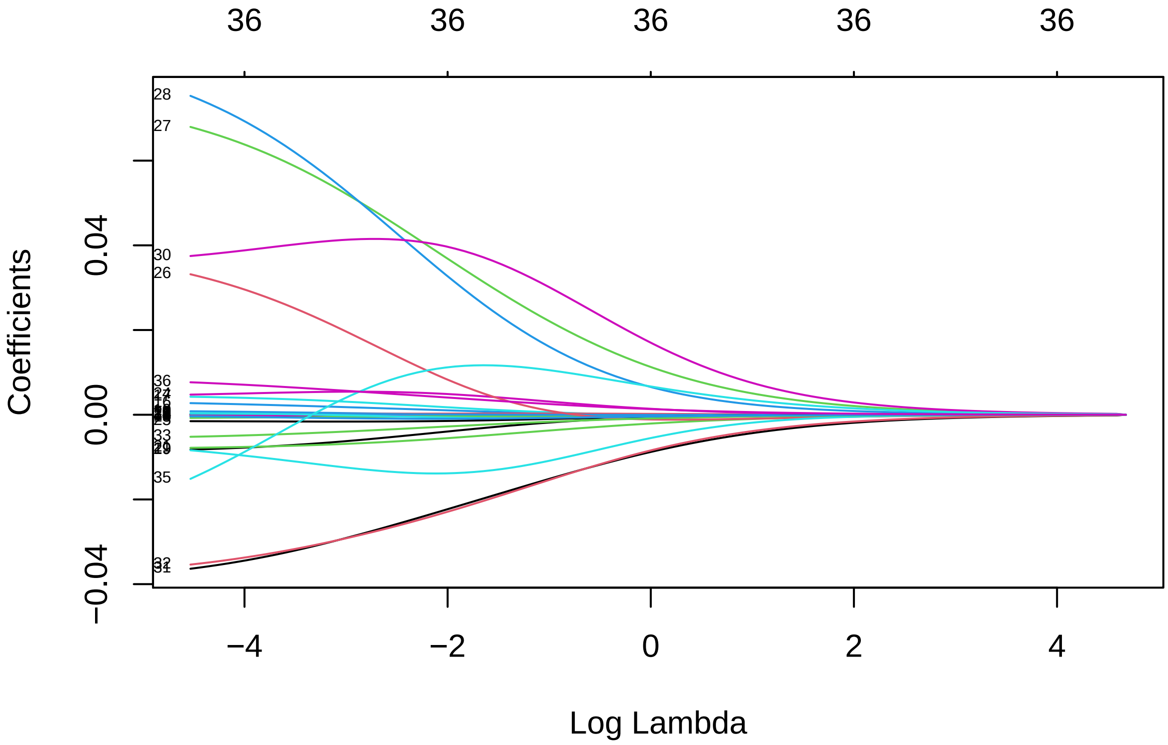
```
## fit with crossval
cvfit_lasso <-  cv.glmnet(x=X_train,y=y_train,alpha = 1)

## plot MSPEs by lambda
plot(cvfit_lasso)
```



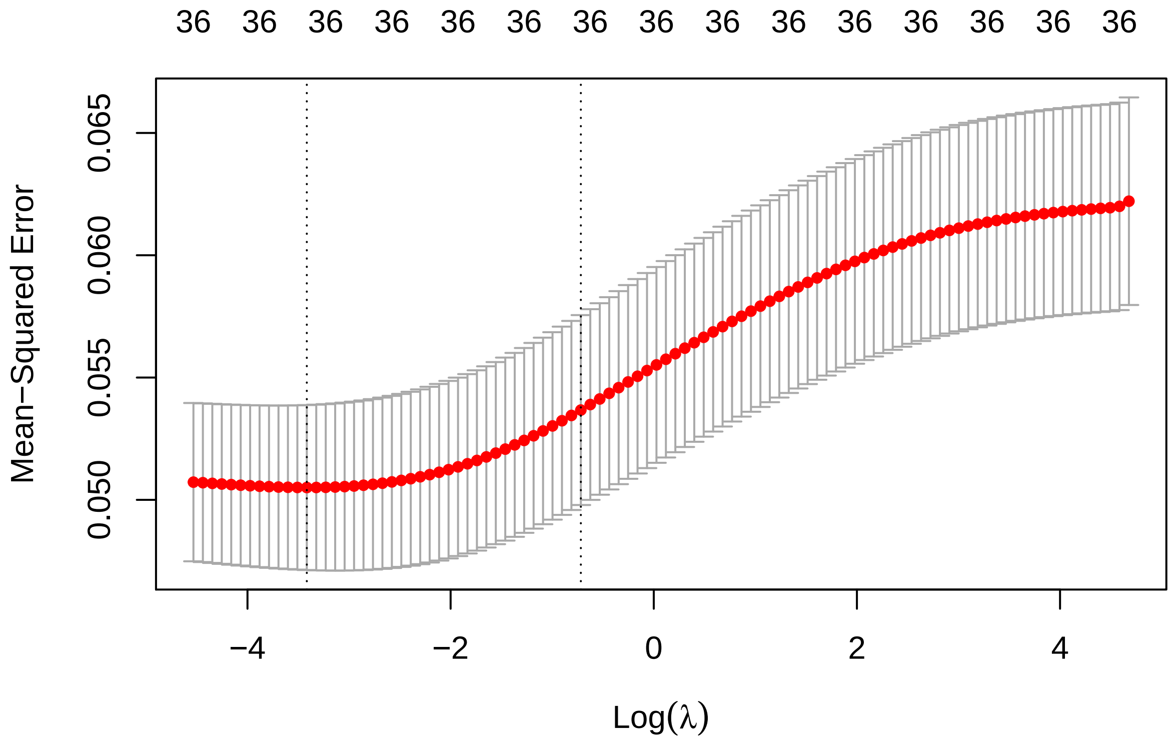
```r
## estimated betas for minimum lambda
coef(cvfit_lasso, s = "lambda.min")
```

```
## 37 x 1 sparse Matrix of class "dgCMatrix"
##                             1
## (Intercept)      1.435105e+00
## POP_PCB1        -1.390128e-07
## POP_PCB2         .
## POP_PCB3         5.421380e-07
## POP_PCB4         .
## POP_PCB5         .
## POP_PCB6         .
## POP_PCB7         .
## POP_PCB8         .
## POP_PCB9         .
## POP_PCB10        .
## POP_PCB11        2.746914e-05
## POP_dioxin1      .
## POP_dioxin2      .
## POP_dioxin3     -6.741939e-06
## POP_furan1       .
## POP_furan2       .
## POP_furan3       4.287018e-03
## POP_furan4       .
## whitecell_count -6.673098e-03
## lymphocyte_pct   .
## monocyte_pct    -6.261320e-03
## eosinophils_pct  .
## basophils_pct    .
## neutrophils_pct  .
## BMI             -1.126403e-03
## edu_cat2         1.832569e-02
## edu_cat3         5.263736e-02
## edu_cat4         5.790583e-02
## race_cat2        .
## race_cat3        3.937936e-02
## race_cat4       -1.813660e-02
## male1           -2.721445e-02
## ageyrs          -5.873940e-03
## yrssmoke         .
## smokenow1        .
## ln_lbxcot        4.200155e-03
```

```
## predictions
pred_lasso <- predict(cvfit_lasso,newx=X_test,  s="lambda.min")

## MSPE in test set
MSPE_lasso <- mean((pred_lasso-y_test)^2)






## RIDGE
## fit models
M_ridge <- glmnet(x=X_train,y=y_train,alpha = 0)

## plot paths
plot(M_ridge,xvar = "lambda",label=TRUE)
```



```
## fit with crossval
cvfit_ridge <-  cv.glmnet(x=X_train,y=y_train,alpha = 0)

## plot MSPEs by lambda
plot(cvfit_ridge)
```



```
## estimated betas for minimum lambda
coef(cvfit_ridge, s = "lambda.min")## alternatively could use "lambda.1se"
```

```
## 37 x 1 sparse Matrix of class "dgCMatrix"
##                             1
## (Intercept)      1.406165e+00
## POP_PCB1        -3.472767e-07
## POP_PCB2        -1.522438e-07
## POP_PCB3         1.272181e-06
## POP_PCB4        -3.919611e-08
## POP_PCB5        -4.171764e-08
## POP_PCB6         1.068579e-07
## POP_PCB7        -5.950297e-07
## POP_PCB8        -3.729365e-07
## POP_PCB9         1.179275e-07
## POP_PCB10        5.169337e-04
## POP_PCB11        6.236251e-05
## POP_dioxin1     -9.221877e-05
## POP_dioxin2     -3.255973e-04
## POP_dioxin3     -9.635121e-06
## POP_furan1      -5.681058e-04
## POP_furan2       2.087375e-03
## POP_furan3       3.523864e-03
## POP_furan4      -1.075695e-04
## whitecell_count -6.985406e-03
## lymphocyte_pct   1.726866e-04
## monocyte_pct    -7.243544e-03
## eosinophils_pct  1.901752e-04
## basophils_pct    6.749112e-05
## neutrophils_pct  5.325444e-03
## BMI             -1.604144e-03
## edu_cat2         2.423045e-02
## edu_cat3         5.762761e-02
## edu_cat4         6.047889e-02
## race_cat2       -1.128511e-02
## race_cat3        4.045355e-02
## race_cat4       -3.157856e-02
## male1           -3.129465e-02
## ageyrs          -4.395474e-03
## yrssmoke        -7.097452e-04
## smokenow1       -1.045812e-03
## ln_lbxcot        6.287474e-03
```

```
## predictions
pred_ridge <- predict(cvfit_ridge,newx=X_test,  s="lambda.min")

## MSPE in test set
MSPE_ridge <- mean((pred_ridge-y_test)^2)


## stepwise

M0 = lm(length~1, data=data.train)
Mfull = lm(length~., data=data.train)
Mstep <- step(object = M0,
              scope = list(lower = M0, upper = Mfull),
              direction = "both", trace = 1, k = 2)
```

```
## Start:  AIC=-1943.58
## length ~ 1
##
##                  Df Sum of Sq    RSS     AIC
## + ageyrs          1    8.1006 35.352 -2086.0
## + POP_dioxin2     1    2.5259 40.927 -1983.5
## + POP_PCB2        1    2.3184 41.135 -1980.0
## + POP_PCB1        1    2.2646 41.188 -1979.0
## + POP_PCB8        1    2.0272 41.426 -1975.0
## + POP_PCB7        1    1.9125 41.540 -1973.1
## + POP_PCB10       1    1.8958 41.557 -1972.8
## + POP_PCB5        1    1.7698 41.683 -1970.7
## + POP_PCB4        1    1.5900 41.863 -1967.7
## + POP_PCB9        1    1.5790 41.874 -1967.5
## + yrssmoke        1    1.2307 42.222 -1961.7
## + POP_dioxin1     1    1.1190 42.334 -1959.8
## + POP_dioxin3     1    0.9838 42.469 -1957.6
## + POP_furan1      1    0.9474 42.506 -1957.0
## + race_cat        3    1.1467 42.306 -1956.3
## + POP_furan3      1    0.8617 42.591 -1955.6
## + POP_PCB3        1    0.8509 42.602 -1955.4
## + POP_PCB6        1    0.8195 42.633 -1954.9
## + edu_cat         3    0.9666 42.486 -1953.3
## + ln_lbxcot        1    0.7157 42.737 -1953.2
## + monocyte_pct     1    0.6965 42.757 -1952.9
## + POP_furan2       1    0.6520 42.801 -1952.2
## + male             1    0.4558 42.997 -1949.0
## + smokenow         1    0.3435 43.109 -1947.1
## + POP_PCB11        1    0.3355 43.117 -1947.0
## + basophils_pct    1    0.1275 43.326 -1943.6
## <none>                         43.453 -1943.6
## + lymphocyte_pct   1    0.1189 43.334 -1943.5
## + BMI              1    0.1073 43.346 -1943.3
## + POP_furan4       1    0.0082 43.445 -1941.7
## + whitecell_count  1    0.0047 43.448 -1941.7
## + eosinophils_pct  1    0.0022 43.451 -1941.6
## + neutrophils_pct  1    0.0014 43.452 -1941.6
##
## Step:  AIC=-2086
## length ~ ageyrs
##
##                   Df Sum of Sq    RSS     AIC
## + POP_furan3       1    0.6348 34.718 -2096.7
## + race_cat         3    0.5707 34.782 -2091.4
## + POP_PCB10        1    0.3651 34.987 -2091.3
## + edu_cat          3    0.5171 34.835 -2090.3
## + POP_furan2       1    0.2625 35.090 -2089.2
## + POP_PCB3         1    0.2184 35.134 -2088.3
## + whitecell_count  1    0.1940 35.158 -2087.8
## + male             1    0.1935 35.159 -2087.8
## + POP_PCB5         1    0.1800 35.172 -2087.6
## + POP_PCB4         1    0.1769 35.176 -2087.5
## + POP_PCB11        1    0.1652 35.187 -2087.3
## + POP_PCB6         1    0.1534 35.199 -2087.0
## + POP_furan1       1    0.1528 35.200 -2087.0
## + POP_dioxin2      1    0.1495 35.203 -2087.0
## + POP_PCB9         1    0.1363 35.216 -2086.7
## + POP_PCB7         1    0.1181 35.234 -2086.3
## + BMI              1    0.1179 35.235 -2086.3
## <none>                         35.352 -2086.0
## + POP_PCB2         1    0.0989 35.254 -2086.0
## + monocyte_pct     1    0.0844 35.268 -2085.7
## + ln_lbxcot        1    0.0829 35.270 -2085.6
## + lymphocyte_pct   1    0.0645 35.288 -2085.3
## + POP_PCB1         1    0.0518 35.301 -2085.0
## + eosinophils_pct  1    0.0267 35.326 -2084.5
## + POP_PCB8         1    0.0166 35.336 -2084.3
## + neutrophils_pct  1    0.0142 35.338 -2084.3
## + POP_furan4       1    0.0111 35.341 -2084.2
## + yrssmoke         1    0.0110 35.341 -2084.2
## + smokenow         1    0.0062 35.346 -2084.1
## + POP_dioxin3      1    0.0028 35.350 -2084.1
## + basophils_pct    1    0.0011 35.351 -2084.0
## + POP_dioxin1      1    0.0003 35.352 -2084.0
## - ageyrs           1    8.1006 43.453 -1943.6
##
## Step:  AIC=-2096.68
## length ~ ageyrs + POP_furan3
##
##                   Df Sum of Sq    RSS     AIC
## + edu_cat          3    0.4625 34.255 -2100.1
## + race_cat         3    0.4447 34.273 -2099.7
## + whitecell_count  1    0.1585 34.559 -2097.9
## + male             1    0.1552 34.562 -2097.8
## + monocyte_pct     1    0.1038 34.614 -2096.8
## <none>                         34.718 -2096.7
## + ln_lbxcot        1    0.0916 34.626 -2096.5
## + BMI              1    0.0716 34.646 -2096.1
## + lymphocyte_pct   1    0.0579 34.660 -2095.8
## + POP_PCB3         1    0.0383 34.679 -2095.5
## + POP_dioxin1      1    0.0324 34.685 -2095.3
## + POP_PCB6         1    0.0211 34.697 -2095.1
## + eosinophils_pct  1    0.0204 34.697 -2095.1
## + POP_PCB10        1    0.0192 34.698 -2095.1
## + smokenow         1    0.0153 34.702 -2095.0
## + POP_PCB11        1    0.0140 34.704 -2095.0
## + POP_dioxin3      1    0.0133 34.704 -2094.9
## + POP_PCB4         1    0.0109 34.707 -2094.9
## + POP_dioxin2      1    0.0101 34.708 -2094.9
## + POP_furan4       1    0.0099 34.708 -2094.9
## + neutrophils_pct  1    0.0063 34.711 -2094.8
## + POP_PCB5         1    0.0059 34.712 -2094.8
## + POP_furan1       1    0.0057 34.712 -2094.8
## + POP_PCB1         1    0.0038 34.714 -2094.8
## + POP_PCB9         1    0.0021 34.715 -2094.7
## + POP_PCB8         1    0.0018 34.716 -2094.7
## + basophils_pct    1    0.0010 34.717 -2094.7
## + POP_PCB2         1    0.0007 34.717 -2094.7
## + POP_PCB7         1    0.0000 34.718 -2094.7
## + yrssmoke         1    0.0000 34.718 -2094.7
## + POP_furan2       1    0.0000 34.718 -2094.7
## - POP_furan3       1    0.6348 35.352 -2086.0
## - ageyrs           1    7.8737 42.591 -1955.6
##
## Step:  AIC=-2100.07
## length ~ ageyrs + POP_furan3 + edu_cat
##
##                   Df Sum of Sq    RSS     AIC
## + race_cat         3    0.5443 33.711 -2105.3
## + male             1    0.1706 34.084 -2101.6
## + ln_lbxcot        1    0.1657 34.089 -2101.5
## + whitecell_count  1    0.1331 34.122 -2100.8
## + monocyte_pct     1    0.1242 34.131 -2100.6
## <none>                         34.255 -2100.1
## + lymphocyte_pct   1    0.0941 34.161 -2100.0
## + POP_PCB3         1    0.0557 34.199 -2099.2
## + BMI              1    0.0556 34.199 -2099.2
## + smokenow         1    0.0408 34.214 -2098.9
## + eosinophils_pct  1    0.0384 34.217 -2098.9
## + POP_PCB6         1    0.0250 34.230 -2098.6
## + POP_PCB4         1    0.0197 34.235 -2098.5
## + POP_PCB11        1    0.0167 34.238 -2098.4
## + POP_PCB5         1    0.0097 34.245 -2098.3
## + POP_PCB9         1    0.0093 34.246 -2098.3
## + POP_dioxin1      1    0.0082 34.247 -2098.2
## + POP_PCB10        1    0.0059 34.249 -2098.2
## + POP_PCB1         1    0.0058 34.249 -2098.2
## + yrssmoke         1    0.0043 34.251 -2098.2
## + POP_furan2       1    0.0039 34.251 -2098.2
## + POP_dioxin2      1    0.0037 34.251 -2098.2
## + POP_PCB8         1    0.0025 34.253 -2098.1
## + POP_furan4       1    0.0018 34.253 -2098.1
## + neutrophils_pct  1    0.0017 34.253 -2098.1
## + POP_dioxin3      1    0.0005 34.255 -2098.1
## + basophils_pct    1    0.0004 34.255 -2098.1
## + POP_furan1       1    0.0002 34.255 -2098.1
## + POP_PCB2         1    0.0002 34.255 -2098.1
## + POP_PCB7         1    0.0001 34.255 -2098.1
## - edu_cat          3    0.4625 34.718 -2096.7
## - POP_furan3       1    0.5803 34.835 -2090.3
## - ageyrs           1    7.4000 41.655 -1965.2
##
## Step:  AIC=-2105.28
## length ~ ageyrs + POP_furan3 + edu_cat + race_cat
##
##                   Df Sum of Sq    RSS     AIC
## + male             1    0.1809 33.530 -2107.1
## + ln_lbxcot        1    0.1519 33.559 -2106.4
## + monocyte_pct     1    0.1507 33.560 -2106.4
## <none>                         33.711 -2105.3
## + smokenow         1    0.0677 33.643 -2104.7
## + BMI              1    0.0651 33.646 -2104.6
## + whitecell_count  1    0.0515 33.659 -2104.4
## + POP_PCB3         1    0.0316 33.679 -2103.9
## + POP_PCB1         1    0.0315 33.679 -2103.9
## + POP_dioxin2      1    0.0282 33.683 -2103.9
## + POP_furan4       1    0.0282 33.683 -2103.9
## + POP_dioxin1      1    0.0261 33.685 -2103.8
## + POP_furan1       1    0.0187 33.692 -2103.7
## + lymphocyte_pct   1    0.0161 33.695 -2103.6
## + POP_PCB8         1    0.0142 33.697 -2103.6
## + POP_PCB2         1    0.0138 33.697 -2103.6
## + POP_PCB6         1    0.0104 33.700 -2103.5
## + POP_dioxin3      1    0.0096 33.701 -2103.5
## + yrssmoke         1    0.0072 33.704 -2103.4
## + POP_PCB9         1    0.0052 33.706 -2103.4
## + POP_PCB11        1    0.0045 33.706 -2103.4
## + neutrophils_pct  1    0.0037 33.707 -2103.4
## + POP_furan2       1    0.0022 33.709 -2103.3
## + basophils_pct    1    0.0010 33.710 -2103.3
## + POP_PCB5         1    0.0009 33.710 -2103.3
## + POP_PCB4         1    0.0008 33.710 -2103.3
## + POP_PCB10        1    0.0006 33.710 -2103.3
## + eosinophils_pct  1    0.0006 33.710 -2103.3
## + POP_PCB7         1    0.0002 33.711 -2103.3
## - race_cat         3    0.5443 34.255 -2100.1
## - edu_cat          3    0.5621 34.273 -2099.7
## - POP_furan3       1    0.5014 34.212 -2096.9
## - ageyrs           1    6.5742 40.285 -1982.6
##
## Step:  AIC=-2107.05
## length ~ ageyrs + POP_furan3 + edu_cat + race_cat + male
##
##                   Df Sum of Sq    RSS     AIC
## + ln_lbxcot        1    0.2160 33.314 -2109.6
## <none>                         33.530 -2107.1
## + monocyte_pct     1    0.0947 33.435 -2107.0
## + smokenow         1    0.0809 33.449 -2106.7
## + BMI              1    0.0687 33.461 -2106.5
## + whitecell_count  1    0.0683 33.461 -2106.5
## + POP_dioxin1      1    0.0379 33.492 -2105.8
## + POP_dioxin3      1    0.0271 33.503 -2105.6
## + yrssmoke         1    0.0227 33.507 -2105.5
## + POP_PCB3         1    0.0223 33.508 -2105.5
## + POP_dioxin2      1    0.0212 33.509 -2105.5
## + POP_furan4       1    0.0152 33.515 -2105.4
## + POP_PCB1         1    0.0148 33.515 -2105.4
## + lymphocyte_pct   1    0.0144 33.515 -2105.3
## + POP_PCB10        1    0.0143 33.516 -2105.3
## - male             1    0.1809 33.711 -2105.3
## + POP_furan1       1    0.0110 33.519 -2105.3
## + POP_PCB7         1    0.0073 33.523 -2105.2
## + neutrophils_pct  1    0.0048 33.525 -2105.2
## + POP_PCB2         1    0.0039 33.526 -2105.1
## + POP_PCB6         1    0.0028 33.527 -2105.1
## + POP_PCB8         1    0.0025 33.527 -2105.1
## + eosinophils_pct  1    0.0024 33.527 -2105.1
## + POP_PCB9         1    0.0014 33.528 -2105.1
## + POP_PCB11        1    0.0012 33.529 -2105.1
## + POP_PCB4         1    0.0009 33.529 -2105.1
## + basophils_pct    1    0.0004 33.529 -2105.1
## + POP_furan2       1    0.0000 33.530 -2105.1
## + POP_PCB5         1    0.0000 33.530 -2105.1
## - race_cat         3    0.5546 34.084 -2101.6
## - edu_cat          3    0.5850 34.115 -2100.9
## - POP_furan3       1    0.4627 33.993 -2099.5
## - ageyrs           1    6.2900 39.820 -1988.7
##
## Step:  AIC=-2109.57
## length ~ ageyrs + POP_furan3 + edu_cat + race_cat + male + ln_lbxcot
##
##                   Df Sum of Sq    RSS     AIC
## + whitecell_count  1    0.1260 33.188 -2110.2
## <none>                         33.314 -2109.6
## + monocyte_pct     1    0.0908 33.223 -2109.5
## + BMI              1    0.0459 33.268 -2108.5
## + POP_dioxin2      1    0.0306 33.283 -2108.2
## + smokenow         1    0.0302 33.284 -2108.2
## + POP_PCB3         1    0.0262 33.288 -2108.1
## + POP_dioxin3      1    0.0244 33.289 -2108.1
## + POP_furan4       1    0.0224 33.291 -2108.0
## + POP_PCB1         1    0.0182 33.296 -2108.0
## + POP_dioxin1      1    0.0136 33.300 -2107.9
## + POP_furan1       1    0.0123 33.302 -2107.8
## + lymphocyte_pct   1    0.0112 33.303 -2107.8
## + POP_PCB10        1    0.0102 33.304 -2107.8
## + yrssmoke         1    0.0098 33.304 -2107.8
## + POP_PCB6         1    0.0069 33.307 -2107.7
## + POP_PCB2         1    0.0058 33.308 -2107.7
## + POP_PCB11        1    0.0052 33.309 -2107.7
## + POP_PCB7         1    0.0051 33.309 -2107.7
## + neutrophils_pct  1    0.0046 33.309 -2107.7
## + POP_PCB8         1    0.0046 33.309 -2107.7
## + POP_PCB9         1    0.0030 33.311 -2107.6
## + eosinophils_pct  1    0.0014 33.312 -2107.6
## + POP_PCB4         1    0.0010 33.313 -2107.6
## + basophils_pct    1    0.0004 33.313 -2107.6
## + POP_PCB5         1    0.0000 33.314 -2107.6
## + POP_furan2       1    0.0000 33.314 -2107.6
## - ln_lbxcot        1    0.2160 33.530 -2107.1
## - male             1    0.2450 33.559 -2106.4
## - race_cat         3    0.5435 33.857 -2104.2
## - POP_furan3       1    0.4918 33.806 -2101.3
## - edu_cat          3    0.7275 34.041 -2100.4
## - ageyrs           1    5.5940 38.908 -2002.9
##
## Step:  AIC=-2110.23
## length ~ ageyrs + POP_furan3 + edu_cat + race_cat + male + ln_lbxcot +
##     whitecell_count
##
##                   Df Sum of Sq    RSS     AIC
## + monocyte_pct     1    0.1843 33.004 -2112.1
## <none>                         33.188 -2110.2
## - whitecell_count  1    0.1260 33.314 -2109.6
## + POP_dioxin2      1    0.0339 33.154 -2108.9
## + BMI              1    0.0285 33.159 -2108.8
## + POP_PCB3         1    0.0279 33.160 -2108.8
## + POP_dioxin3      1    0.0240 33.164 -2108.7
## + POP_furan4       1    0.0232 33.165 -2108.7
## + smokenow         1    0.0227 33.165 -2108.7
## + POP_PCB1         1    0.0222 33.166 -2108.7
## + POP_dioxin1      1    0.0169 33.171 -2108.6
## + eosinophils_pct  1    0.0145 33.173 -2108.5
## + POP_furan1       1    0.0132 33.175 -2108.5
## + POP_PCB10        1    0.0097 33.178 -2108.4
## + POP_PCB6         1    0.0085 33.179 -2108.4
## + POP_PCB11        1    0.0080 33.180 -2108.4
## + POP_PCB8         1    0.0078 33.180 -2108.4
## + POP_PCB2         1    0.0077 33.180 -2108.4
## + neutrophils_pct  1    0.0057 33.182 -2108.3
## + POP_PCB7         1    0.0047 33.183 -2108.3
## + yrssmoke         1    0.0046 33.183 -2108.3
## + POP_PCB9         1    0.0043 33.184 -2108.3
## + POP_PCB4         1    0.0016 33.186 -2108.3
## + lymphocyte_pct   1    0.0007 33.187 -2108.2
## + POP_furan2       1    0.0004 33.187 -2108.2
## + POP_PCB5         1    0.0002 33.188 -2108.2
## + basophils_pct    1    0.0002 33.188 -2108.2
## - race_cat         3    0.4227 33.611 -2107.4
## - ln_lbxcot        1    0.2736 33.461 -2106.5
## - male             1    0.2819 33.470 -2106.3
## - POP_furan3       1    0.4723 33.660 -2102.3
## - edu_cat          3    0.6907 33.879 -2101.8
## - ageyrs           1    5.7106 38.898 -2001.1
##
## Step:  AIC=-2112.13
## length ~ ageyrs + POP_furan3 + edu_cat + race_cat + male + ln_lbxcot +
##     whitecell_count + monocyte_pct
##
##                    Df Sum of Sq    RSS     AIC
## <none>                          33.004 -2112.1
## + POP_dioxin2       1    0.0312 32.972 -2110.8
## + BMI               1    0.0311 32.972 -2110.8
## + POP_dioxin3       1    0.0266 32.977 -2110.7
## + POP_PCB3          1    0.0264 32.977 -2110.7
## + POP_PCB1          1    0.0195 32.984 -2110.5
## + POP_dioxin1       1    0.0186 32.985 -2110.5
## + POP_furan4        1    0.0184 32.985 -2110.5
## + smokenow          1    0.0184 32.985 -2110.5
## + POP_PCB10         1    0.0137 32.990 -2110.4
## + POP_furan1        1    0.0086 32.995 -2110.3
## + POP_PCB6          1    0.0084 32.995 -2110.3
## + POP_PCB11         1    0.0074 32.996 -2110.3
## + neutrophils_pct   1    0.0065 32.997 -2110.3
## + POP_PCB2          1    0.0061 32.997 -2110.3
## - monocyte_pct      1    0.1843 33.188 -2110.2
## + POP_PCB8          1    0.0048 32.999 -2110.2
## + POP_PCB9          1    0.0043 32.999 -2110.2
## + yrssmoke          1    0.0036 33.000 -2110.2
## + POP_PCB7          1    0.0033 33.000 -2110.2
## + POP_PCB4          1    0.0020 33.002 -2110.2
## + basophils_pct     1    0.0012 33.002 -2110.2
## + lymphocyte_pct    1    0.0009 33.003 -2110.1
## + eosinophils_pct   1    0.0002 33.003 -2110.1
## + POP_PCB5          1    0.0001 33.003 -2110.1
## + POP_furan2        1    0.0000 33.004 -2110.1
## - male              1    0.1983 33.202 -2109.9
## - race_cat          3    0.4099 33.413 -2109.5
## - whitecell_count   1    0.2195 33.223 -2109.5
## - ln_lbxcot         1    0.2938 33.297 -2107.9
## - POP_furan3        1    0.4891 33.493 -2103.8
## - edu_cat           3    0.7085 33.712 -2103.3
## - ageyrs            1    5.4747 38.478 -2006.7
```

```
MSPE_step = mean((  predict(Mstep, newdata=data.test) - y_test)^2)

p = predict(Mstep, newdata=data.test)
```

```
cvfit_lasso$del
```

```
## NULL
```

```
MSPE_lasso
```

```
## [1] 0.05169661
```

```
MSPE_ridge
```

```
## [1] 0.05290817
```

```
MSPE_step
```

```
## [1] 0.05387623
```

say we try to fit with only 2 features

we first see if lasso choose the same simple linear model

```
# lasso choose the same single variable
min(which((M_lasso$lambda)<=exp( -2.5)))
```

```
## [1] 4
```

```
coefs = M_lasso$beta[,4]
which(coefs!=0)
```

```
## ageyrs
##     33
```

```
library("plot3D")
```

```
## Warning: package 'plot3D' was built under R version 4.0.4
```

```
# 2 feature lasso choose
min(which((M_lasso$lambda)<=exp( -3.8)))
```

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

```
coefs = M_lasso$beta[,18]
choosen=which(coefs!=0)
coefs[choosen]
```

```
##    race_cat3       ageyrs
##  0.013839164 -0.004676005
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.4
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.4
```

```
## Warning: package 'tidyr' was built under R version 4.0.4
```

```
## Warning: package 'readr' was built under R version 4.0.4
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## x tidyr::unpack() masks Matrix::unpack()
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.4
```

```
## Loading required package: lattice
```

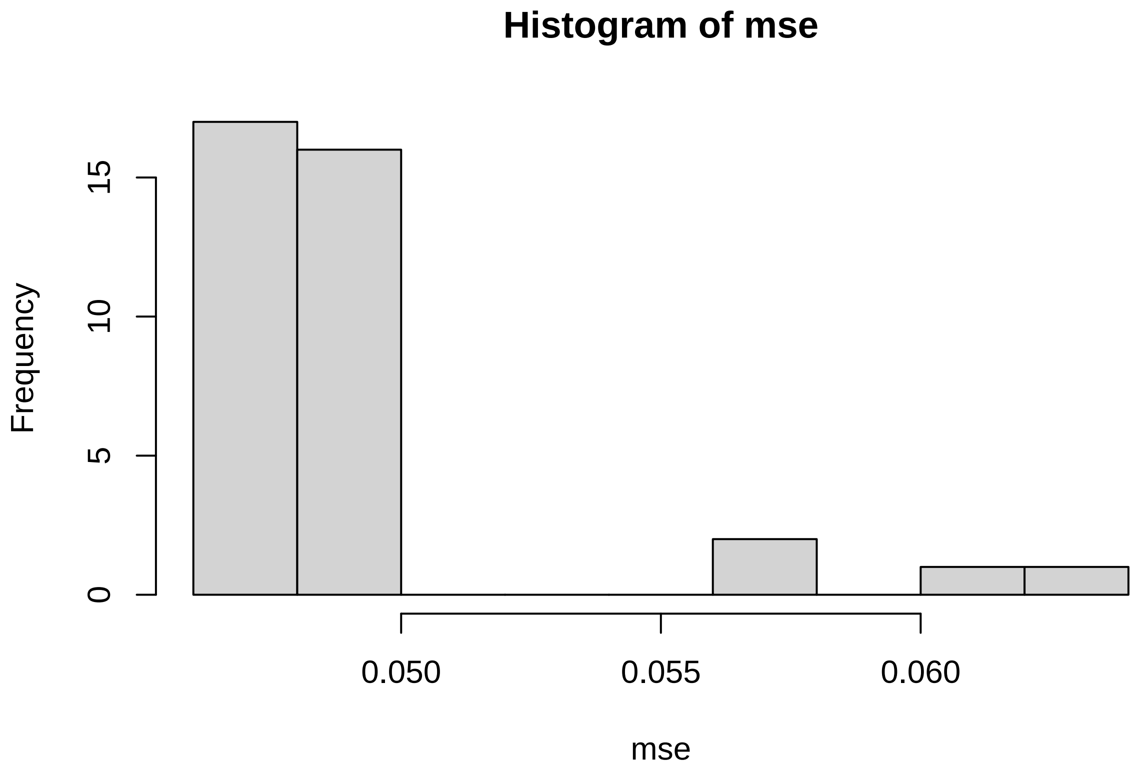
```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.4
```

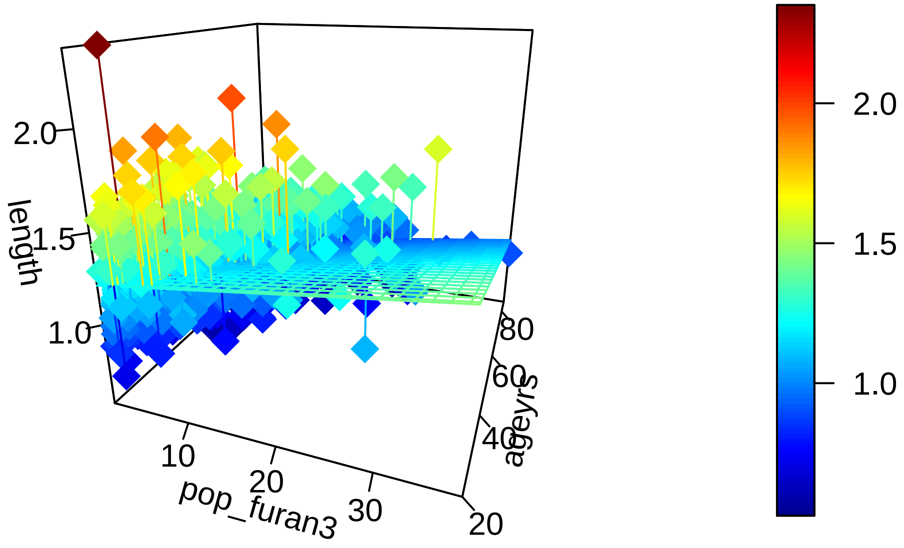
```r
models= regsubsets(length~., data=data, nvmax=2)
# rss of all 2 feature model, we see no magical model
mse = models$rss/nrow(data)
hist(mse)
```



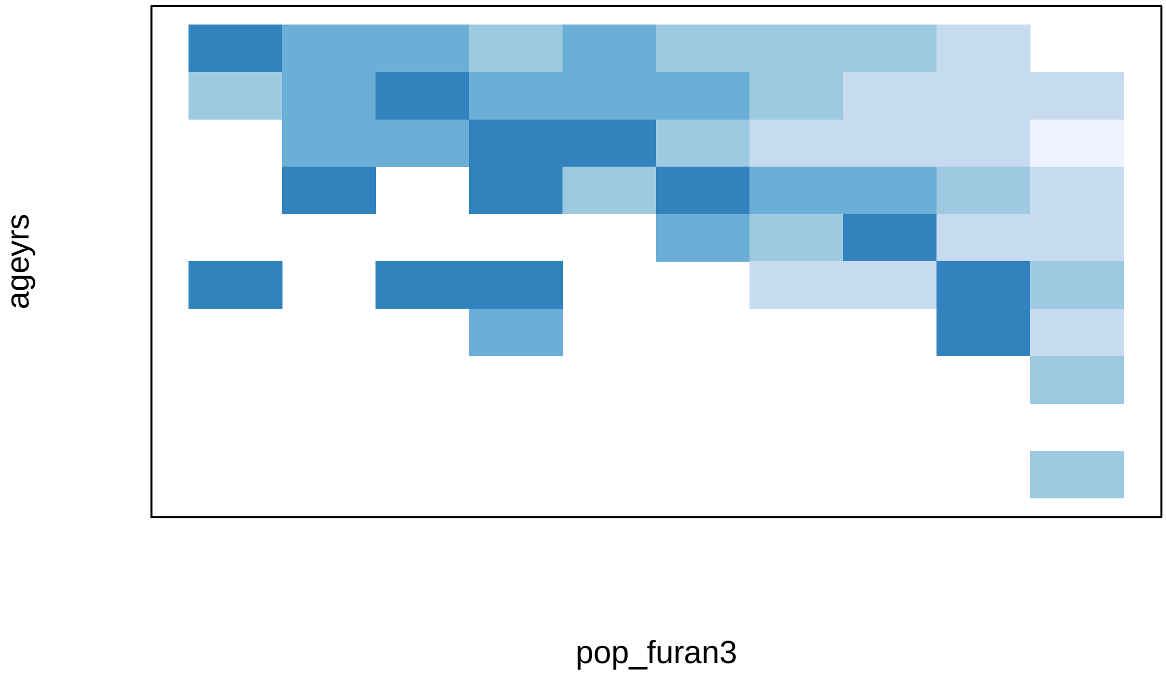
```
library("loon")
```

```
## Warning: package 'loon' was built under R version 4.0.4

## Loading required package: tcltk

## loon Version 1.3.4.
## To learn more about loon, see l_web().
```

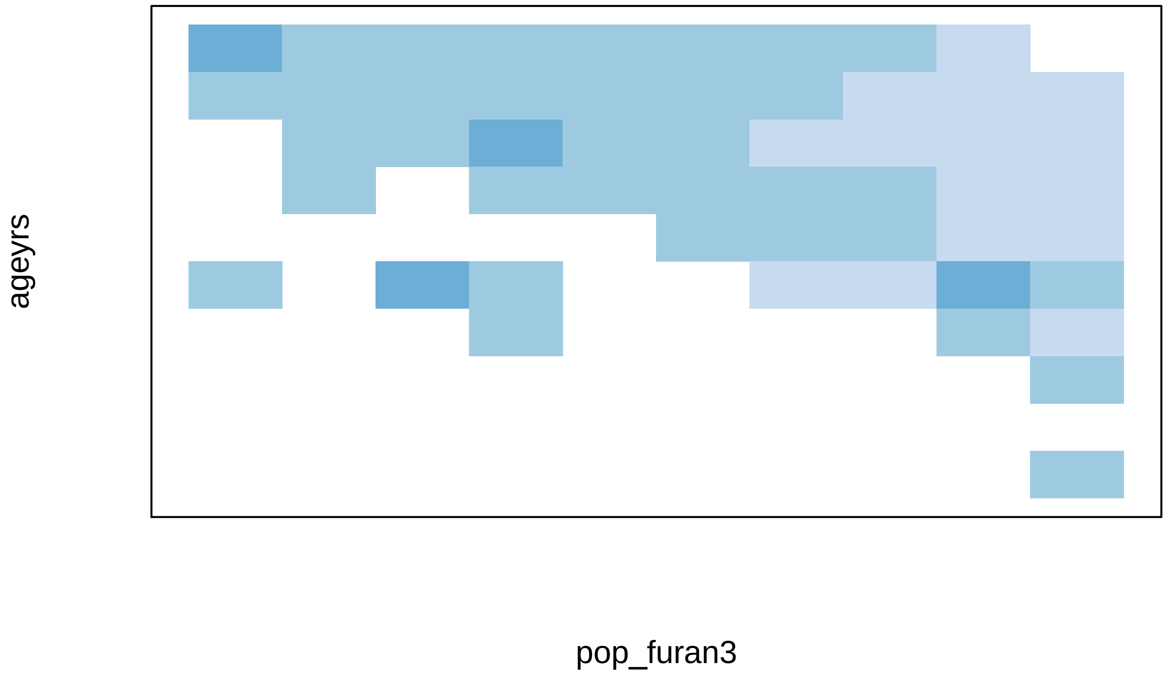
```
z=data$length
y=data$ageyrs
x=data$POP_furan3

fit <- lm(z ~ x + y)
# predict values on regular xy grid
grid.lines = 26
x.pred <- seq(min(x), max(x), length.out = grid.lines)
y.pred <- seq(min(y), max(y), length.out = grid.lines)
xy <- expand.grid( x = x.pred, y = y.pred)
z.pred <- matrix(predict(fit, newdata = xy),
                 nrow = grid.lines, ncol = grid.lines)
# fitted points for droplines to surface

fitpoints = predict(fit)
# scatter plot with regression plane
scatter3D(x, y, z, pch = 18, cex = 2,
    theta = 20, phi = 20, ticktype = "detailed",
    surf = list(x = x.pred, y = y.pred, z = z.pred,
    facets = NA, fit = fitpoints), xlab="pop_furan3", ylab="ageyrs",zlab="length")
```



```
#turn ageyrs and pop_furan into grids

miny=min(y)
intervaly = (max(y)-miny)/10
minx=min(x)
intervalx = (max(x)-minx)/10

xy = matrix(0, nrow = 10, ncol = 10)
count = matrix(0, nrow = 10, ncol = 10)
for (i in 1:nrow(data)){
  xgrid = (x[i]-minx)/intervalx
  ygrid = (y[i]-miny)/intervaly
  count[xgrid,ygrid] = 1 + count[xgrid,ygrid]
  xy[xgrid,ygrid] = xy[xgrid, ygrid] + z[i]
}
xygrid = xy/count
col_areas(xygrid,xlab="pop_furan3", ylab="ageyrs")
```



```
maxz=max(z)
minz=min(z)
breaks = seq( minz, maxz, by=(maxz-minz)/5 )
col_areas(xygrid,xlab="pop_furan3", ylab="ageyrs", breaks = breaks)
```

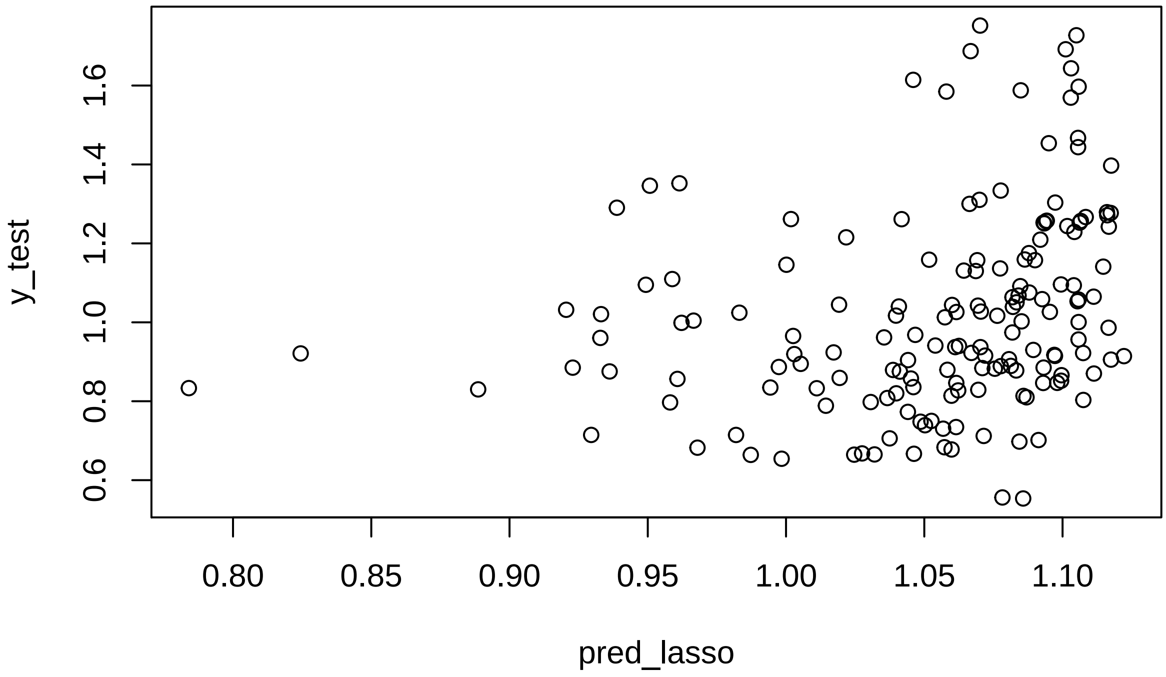


```
# anyway how does this compare to the best fit?
```

4/4

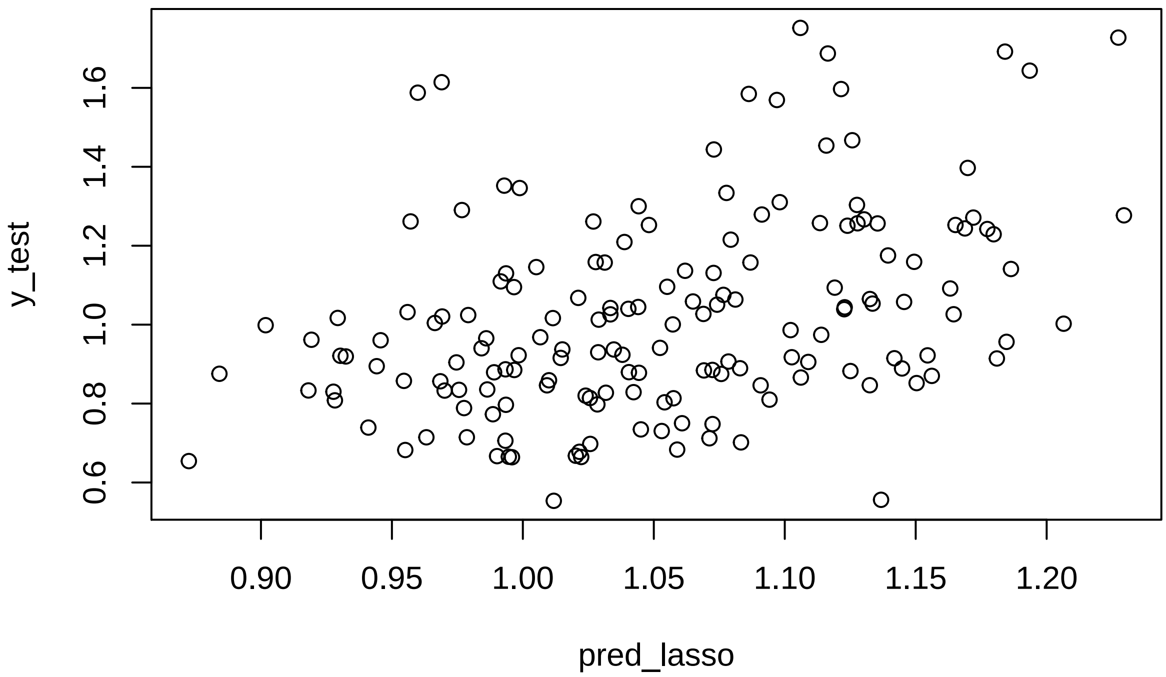
```
cols = colnames(data)
po.ind = str_detect(cols, "POP")

# this is to test tranforamtion of data's result on lasso result
lasso.on.pollutants =function(data){
  M = model.matrix(lm(length~., data=data))
  cols = colnames(M)
  po.ind = str_detect(cols, "POP")
  y_train = data$length[1:700]
  X_train = M[1:700,po.ind]
  y_test= data$length[701:nTotal]
  X_test= M[701:nTotal,(1:ncol(M))[po.ind]]


  M_lasso <- glmnet(x=X_train,y=y_train,alpha = 1)
  ## plot paths

  ## fit with crossval
  cvfit_lasso <-  cv.glmnet(x=X_train,y=y_train,alpha = 1)

  ## plot MSPEs by lambda

  ## estimated betas for minimum lambda
  ## predictions
  pred_lasso <- predict(cvfit_lasso,newx=X_test,  s="lambda.min")

  ## MSPE in test set
  MSPE_lasso <- mean((pred_lasso-y_test)^2)
  print(paste("mspe",MSPE_lasso) )

  plot(pred_lasso, y_test)

  return( coef(cvfit_lasso, s = "lambda.min"))

}



model = lasso.on.pollutants(data)
```

```
## [1] "mspe 0.0599442940117174"
```



```
#########################


# log transform
newdata = data
newdata[,po.ind] = log(data[,po.ind])
chosen.po.ind= which(lasso.on.pollutants(newdata)!=0)
```

```
## [1] "mspe 0.0549633507362908"
```



```
chosen.po.ind= chosen.po.ind[2:length(chosen.po.ind)]
```

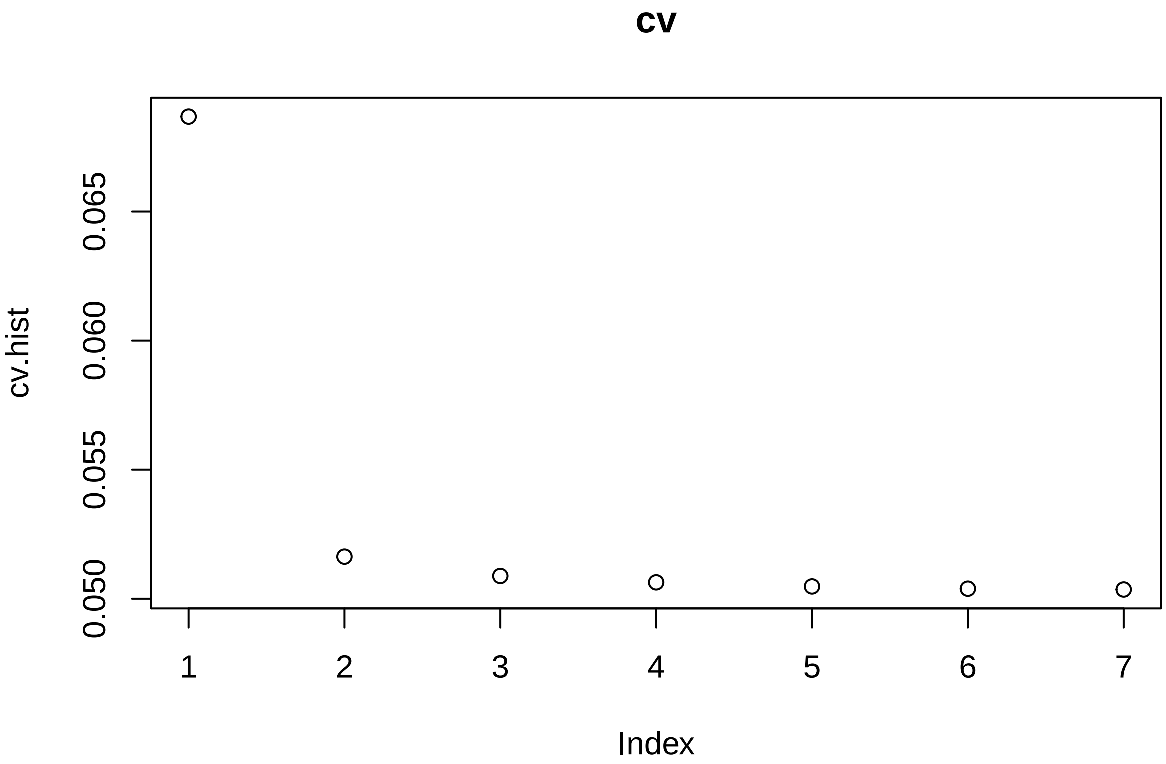
```
kfolds.cv <- function(dat, expr){
  kfolds=10
  mspe = rep(0, kfolds)
  ind = rep(1:kfolds, length=nrow(dat))
  for(ii in 1:kfolds) {
    train<- which(ind!=ii) # training observations
    M.cv <- lm(expr, data=data[train,])
    # cross-validation residuals
    M.res <- dat$length[-train] - # test observations
      predict(M.cv, newdat = dat[-train,]) # prediction with training dat
    # mspe
    mspe[ii] <- mean(M.res^2)
  }
  mean(mspe)
}



forward.change = function(data, expr, show=FALSE){
  model = lm(expr, data=newdata)
  initial.colname = names( model$coefficients)[-1]
  tempnames = colnames(data)
  cv.hist=c()
```

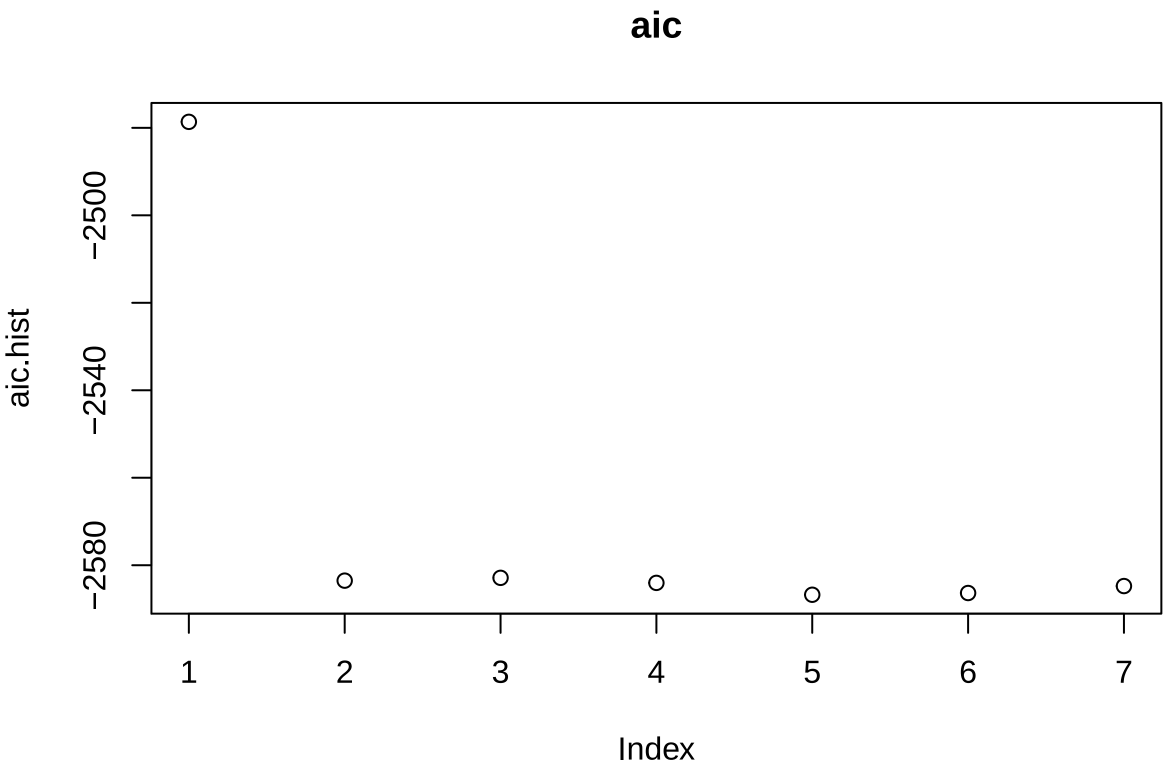
```
aic.hist = c()
coef.hist = list()
j=0
models = list()
while (TRUE) {
  j=j+1
  print(paste("step", j))
  cov.in.m = colnames(model$model)
  cov.all = colnames(newdata)
  names.to.try = cov.all[! cov.all %in% cov.in.m]
  nn = length(names.to.try)
  #update tracks
  cv.hist[j]=kfolds.cv(newdata, expr)
  aic.hist[j] = extractAIC(model)[2]
  coef.hist[[j]] = coef(model)

  cv.score = rep(0, nn)
  if(length(names.to.try) == 0){
    print("chose all ")
    break
  }
  for (i in 1:nn) {
    name = names.to.try[i]
    newexpr =   paste(expr,  "+", name )
    newmodel = lm(newexpr, data=newdata)
    cv.score[i] = kfolds.cv(newdata, newexpr)
  }
  ind = which.min(cv.score)
  if(cv.score[ind]>cv.hist[j]){
    print ("done choosing model")
    break
  }else{
    # update our model
    print(paste("added", names.to.try[ind]))
    expr = paste(expr,"+", names.to.try[ind])
    model =  lm(expr, data=newdata)
    models[[j]] = model
  }
}
plot(cv.hist, main = "cv")
plot(aic.hist, main = "aic")

i = length(initial.colname)
j = length(coef.hist)
M = matrix(0, nrow = i, ncol = j)
for (ii in 1:i){
  for (jj in 1:j) {
    M[ii,jj] =  coef.hist[[jj]][initial.colname[ii]]
  }
}
if(show==TRUE){
  par(cex=0.7)
  plot(M[1,], main=initial.colname[[1]], type = 'l', col=1, ylim = range(M))
```

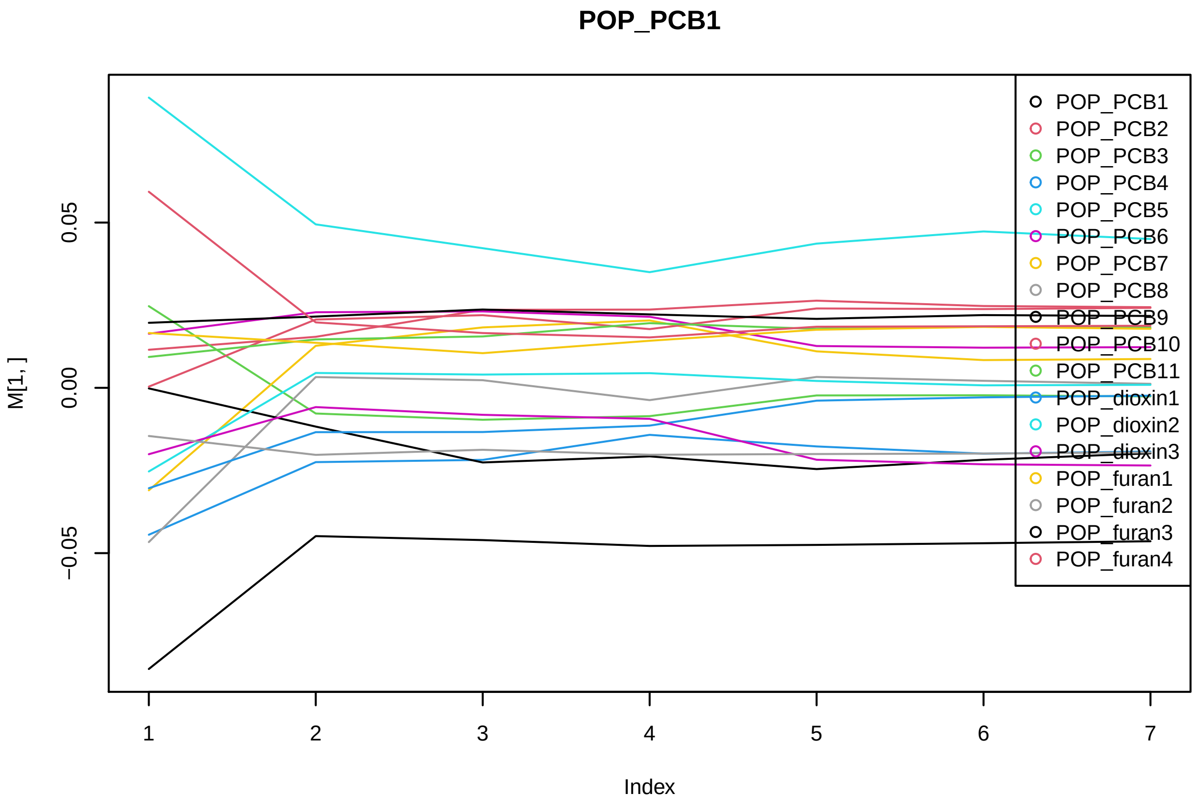
```r
    for (a in 2:i){
      lines(1:j, M[a,] ,col=a)
    }
    legend("topright",legend = initial.colname, col = 1:i, pch=1)
  }
  return(list(cv=cv.hist, coef=coef.hist, aic=aic.hist))

}
```

```r
# log transform
newdata=data
newdata[,po.ind] = log(newdata[,po.ind])

# forward start from length over pollutants
expr = paste("length~", paste(colnames(data)[po.ind] , collapse = "+"))

forward.change(newdata, expr,TRUE)
```

```
## [1] "step 1"
## [1] "added ageyrs"
## [1] "step 2"
## [1] "added race_cat"
## [1] "step 3"
## [1] "added BMI"
## [1] "step 4"
## [1] "added male"
## [1] "step 5"
## [1] "added eosinophils_pct"
## [1] "step 6"
## [1] "added neutrophils_pct"
## [1] "step 7"
## [1] "done choosing model"
```


cv
cv.hist
0.050
0.055
0.060
0.065
1
2
3
4
5
6
7
Index

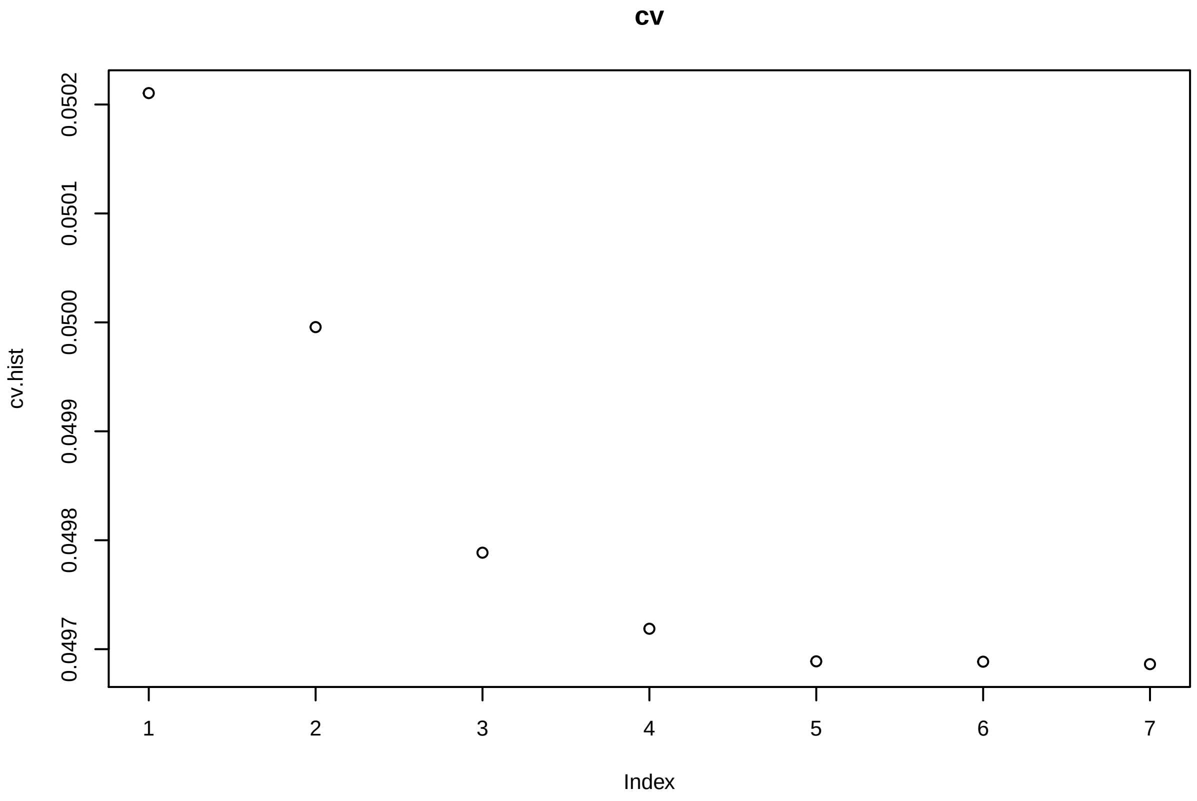

aic
aic.hist
−2500
−2540
−2580
1
2
3
4
5
6
7
Index

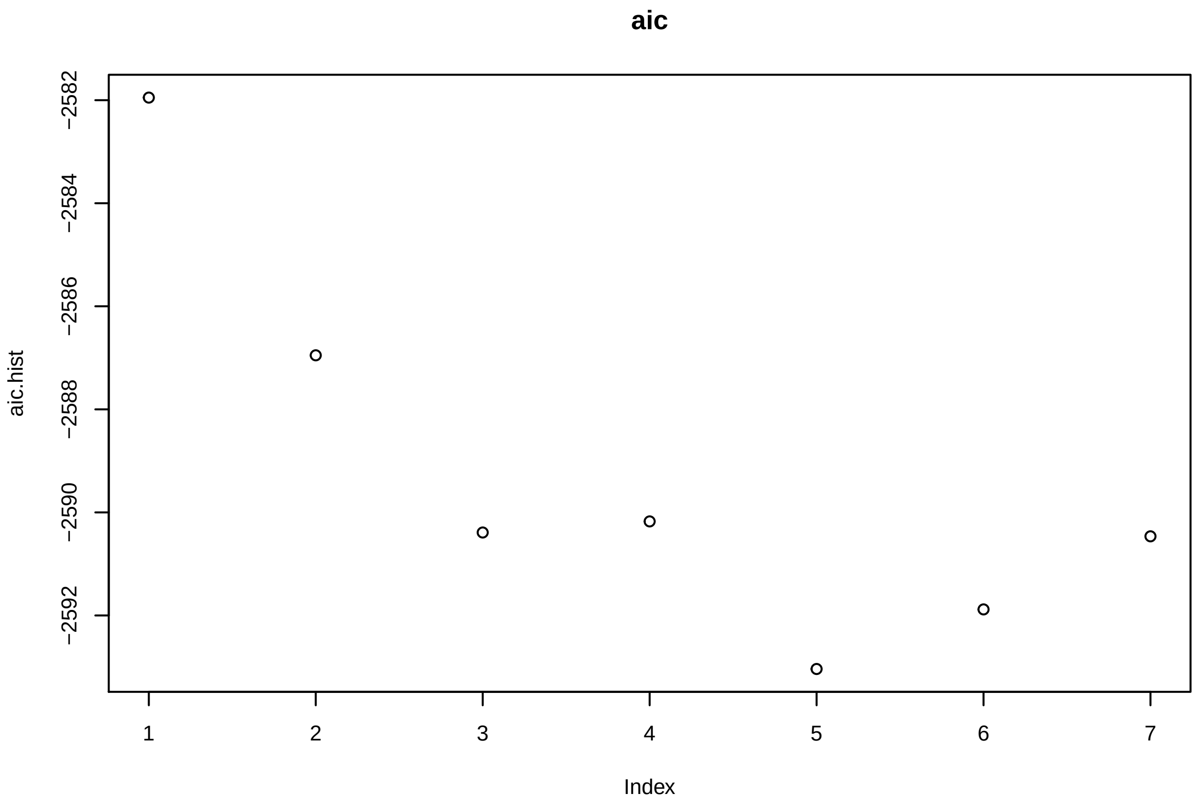


```
## $cv
## [1] 0.06867809 0.05162992 0.05088106 0.05063295 0.05047446 0.05038853 0.05035651
##
## $coef
## $coef[[1]]
##   (Intercept)      POP_PCB1      POP_PCB2      POP_PCB3      POP_PCB4
##  1.7867981029 -0.0850406913  0.0115181406  0.0247104925 -0.0444365546
##      POP_PCB5      POP_PCB6      POP_PCB7      POP_PCB8      POP_PCB9
##  0.0877917668  0.0163681749 -0.0310010872 -0.0466599572 -0.0001647580
##     POP_PCB10     POP_PCB11   POP_dioxin1   POP_dioxin2   POP_dioxin3
##  0.0003646602  0.0093426437 -0.0303672459 -0.0252980400 -0.0200752306
##    POP_furan1    POP_furan2    POP_furan3    POP_furan4
##  0.0165842859 -0.0145660804  0.0196767274  0.0593101111
##
## $coef[[2]]
##  (Intercept)     POP_PCB1     POP_PCB2     POP_PCB3     POP_PCB4     POP_PCB5
##  1.165516424 -0.044845918  0.015432407 -0.007758309 -0.022467556  0.049447686
##     POP_PCB6     POP_PCB7     POP_PCB8     POP_PCB9    POP_PCB10    POP_PCB11
##  0.022865272  0.012716798  0.003263650 -0.011726308  0.020668960  0.014645422
##  POP_dioxin1  POP_dioxin2  POP_dioxin3   POP_furan1   POP_furan2   POP_furan3
## -0.013404871  0.004496220 -0.005822726  0.013625555 -0.020268351  0.021558053
##   POP_furan4       ageyrs
##  0.019799252 -0.007832624
##
## $coef[[3]]
##  (Intercept)     POP_PCB1     POP_PCB2     POP_PCB3     POP_PCB4     POP_PCB5
##  1.256534541 -0.046052691  0.023606824 -0.009646651 -0.021793664  0.042238498
##     POP_PCB6     POP_PCB7     POP_PCB8     POP_PCB9    POP_PCB10    POP_PCB11
##  0.023118499  0.018279116  0.002312819 -0.022574044  0.021983960  0.015546194
##  POP_dioxin1  POP_dioxin2  POP_dioxin3   POP_furan1   POP_furan2   POP_furan3
## -0.013364635  0.004016498 -0.008159764  0.010486500 -0.018736669  0.023692141
##   POP_furan4       ageyrs    race_cat2    race_cat3    race_cat4
##  0.016574623 -0.007507695 -0.028626307  0.021177358 -0.026436233
##
## $coef[[4]]
##  (Intercept)     POP_PCB1     POP_PCB2     POP_PCB3     POP_PCB4     POP_PCB5
##  1.354423213 -0.047823271  0.023673274 -0.008559876 -0.014220431  0.034990155
##     POP_PCB6     POP_PCB7     POP_PCB8     POP_PCB9    POP_PCB10    POP_PCB11
##  0.021495005  0.020377085 -0.003705402 -0.020723474  0.017764636  0.019559909
##  POP_dioxin1  POP_dioxin2  POP_dioxin3   POP_furan1   POP_furan2   POP_furan3
## -0.011419100  0.004444383 -0.009376655  0.014240281 -0.020235731  0.022218486
##   POP_furan4       ageyrs    race_cat2    race_cat3    race_cat4          BMI
##  0.015246263 -0.007327993 -0.028691557  0.026734851 -0.024998713 -0.002485045
##
## $coef[[5]]
##  (Intercept)     POP_PCB1     POP_PCB2     POP_PCB3     POP_PCB4     POP_PCB5
##  1.381619591 -0.047513168  0.026375447 -0.002297587 -0.017720625  0.043628290
##     POP_PCB6     POP_PCB7     POP_PCB8     POP_PCB9    POP_PCB10    POP_PCB11
##  0.012653933  0.011038931  0.003298528 -0.024574374  0.024017870  0.017846634
##  POP_dioxin1  POP_dioxin2  POP_dioxin3   POP_furan1   POP_furan2   POP_furan3
## -0.003876096  0.002085829 -0.021744629  0.017577241 -0.020019743  0.020861501
##   POP_furan4       ageyrs    race_cat2    race_cat3    race_cat4          BMI
##  0.018490075 -0.007261017 -0.024352615  0.025335556 -0.023796346 -0.001943039
##        male1
## -0.039717156
##
## $coef[[6]]
##     (Intercept)        POP_PCB1        POP_PCB2        POP_PCB3        POP_PCB4
##    1.3261766150   -0.0470024798    0.0247459856   -0.0022452077   -0.0199095002
##        POP_PCB5        POP_PCB6        POP_PCB7        POP_PCB8        POP_PCB9
##    0.0473095611    0.0121491697    0.0083978425    0.0021312772   -0.0217768428
##       POP_PCB10       POP_PCB11     POP_dioxin1     POP_dioxin2     POP_dioxin3
##    0.0238233363    0.0185411626   -0.0028800112    0.0007579246   -0.0231258703
##      POP_furan1      POP_furan2      POP_furan3      POP_furan4          ageyrs
##    0.0184278077   -0.0198753018    0.0220055100    0.0186146454   -0.0072160480
##       race_cat2       race_cat3       race_cat4             BMI           male1
##   -0.0233761524    0.0310616678   -0.0239525050   -0.0018654178   -0.0399070076
## eosinophils_pct
##    0.0010370585
##
## $coef[[7]]
##     (Intercept)        POP_PCB1        POP_PCB2        POP_PCB3        POP_PCB4
##    1.3186274629   -0.0464038067    0.0243979197   -0.0025942475   -0.0192703659
##        POP_PCB5        POP_PCB6        POP_PCB7        POP_PCB8        POP_PCB9
##    0.0450274385    0.0123025524    0.0087200145    0.0012235676   -0.0198024901
##       POP_PCB10       POP_PCB11     POP_dioxin1     POP_dioxin2     POP_dioxin3
##    0.0242039144    0.0181979520   -0.0023584562    0.0009129338   -0.0234880441
##      POP_furan1      POP_furan2      POP_furan3      POP_furan4          ageyrs
##    0.0178541647   -0.0195354257    0.0217389005    0.0187745768   -0.0072161294
##       race_cat2       race_cat3       race_cat4             BMI           male1
##   -0.0231543705    0.0315350166   -0.0244302771   -0.0018600927   -0.0398372576
## eosinophils_pct neutrophils_pct
##    0.0011115929    0.0102447000
##
##
## $aic
## [1] -2478.630 -2583.523 -2582.883 -2584.038 -2586.749 -2586.356 -2584.766
```

```
# start from over ageyrs
expr = "length~ageyrs"
forward.change(newdata, expr)
```

```
## [1] "step 1"
## [1] "added male"
## [1] "step 2"
## [1] "added race_cat"
## [1] "step 3"
## [1] "added BMI"
## [1] "step 4"
## [1] "added POP_furan4"
## [1] "step 5"
## [1] "added POP_PCB8"
## [1] "step 6"
## [1] "added POP_dioxin3"
## [1] "step 7"
## [1] "done choosing model"
```

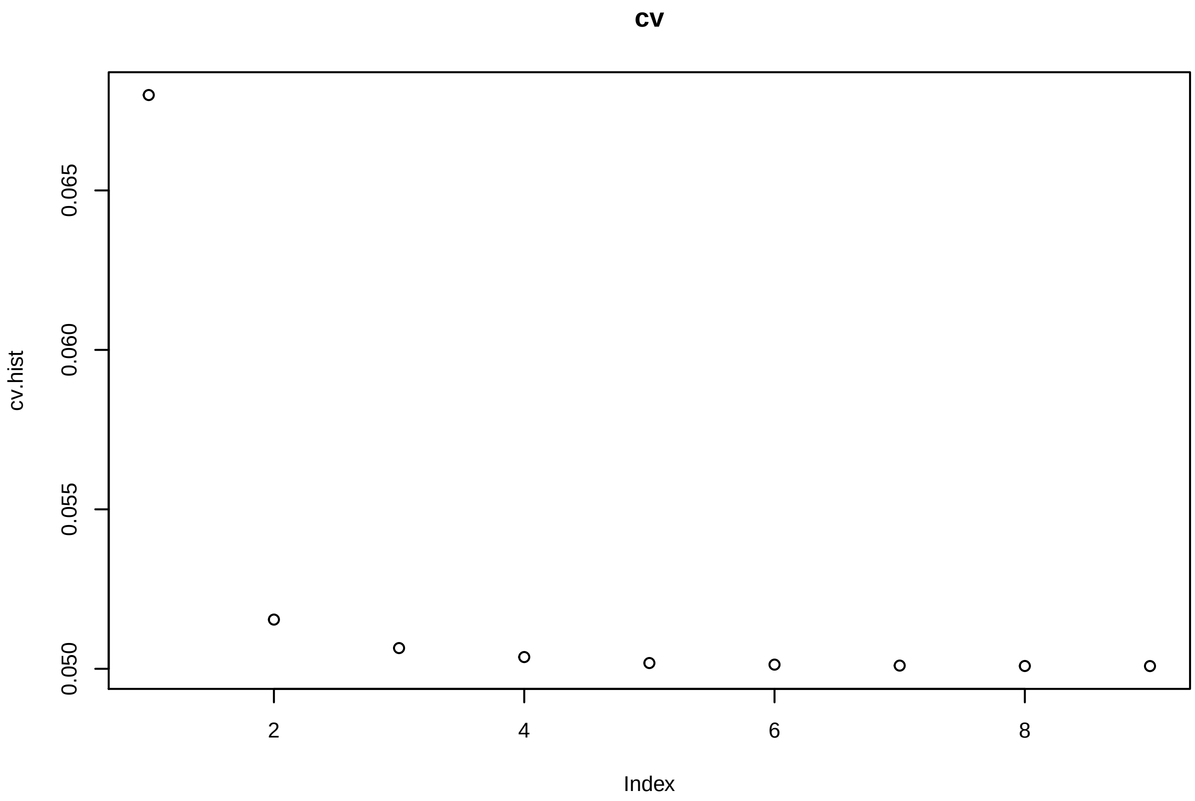


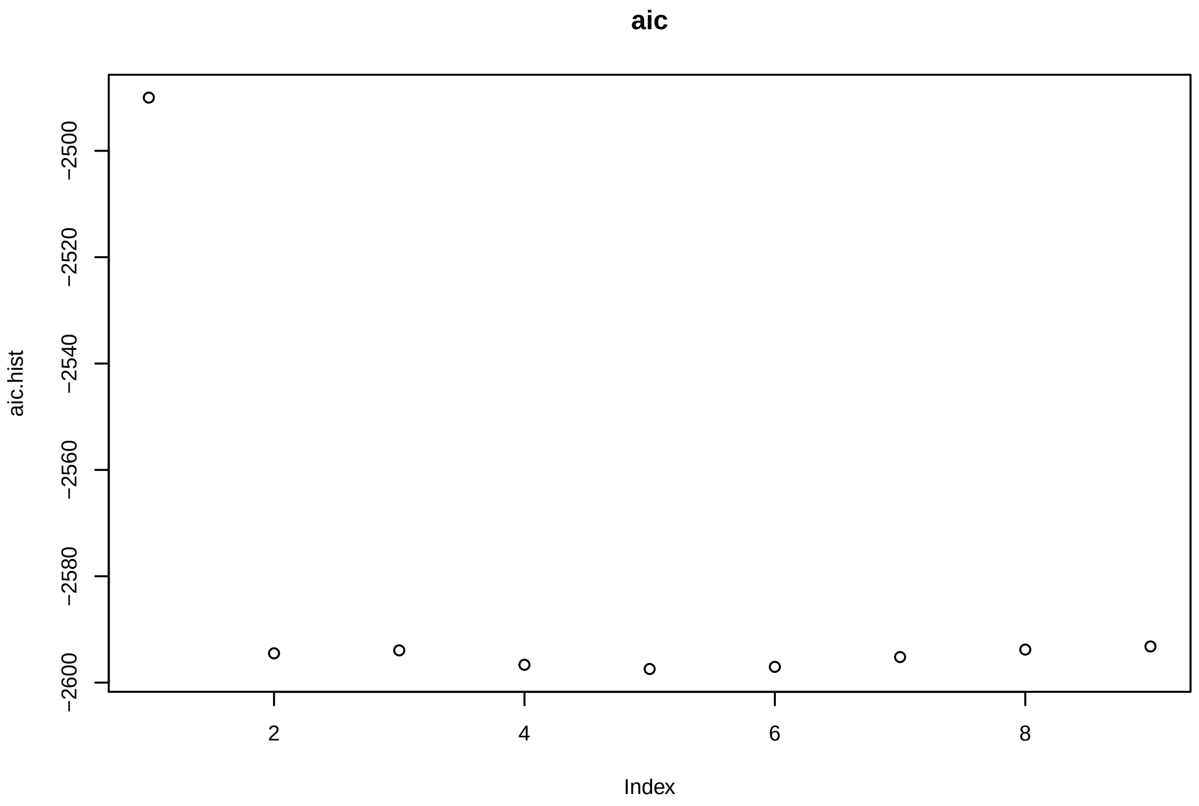


```
## $cv
## [1] 0.05021045 0.04999565 0.04978858 0.04971877 0.04968883 0.04968851 0.04968626
##
## $coef
## $coef[[1]]
##  (Intercept)       ageyrs
##  1.349257536 -0.006099533
##
## $coef[[2]]
##  (Intercept)       ageyrs        male1
##  1.363540811 -0.006030777 -0.040677241
##
## $coef[[3]]
##  (Intercept)       ageyrs        male1    race_cat2    race_cat3    race_cat4
##  1.380307147 -0.006029646 -0.040548778 -0.045993353  0.025344031 -0.021650989
##
## $coef[[4]]
##  (Intercept)       ageyrs        male1    race_cat2    race_cat3    race_cat4
##  1.426497113 -0.006022854 -0.040776359 -0.044144257  0.027709907 -0.021652346
##          BMI
## -0.001681833
##
## $coef[[5]]
##  (Intercept)       ageyrs        male1    race_cat2    race_cat3    race_cat4
##  1.373783837 -0.006106999 -0.043245950 -0.037049075  0.021646593 -0.018550762
##          BMI   POP_furan4
## -0.001605548  0.024353428
##
## $coef[[6]]
##  (Intercept)       ageyrs        male1    race_cat2    race_cat3    race_cat4
##  1.261044320 -0.006535249 -0.045887556 -0.035026224  0.016222773 -0.020619022
##          BMI   POP_furan4     POP_PCB8
## -0.001247678  0.022910621  0.014483577
##
## $coef[[7]]
##  (Intercept)       ageyrs        male1    race_cat2    race_cat3    race_cat4
##  1.294510636 -0.006345889 -0.049945789 -0.032946859  0.018756386 -0.019848530
##          BMI   POP_furan4     POP_PCB8  POP_dioxin3
## -0.001108495  0.027413632  0.014935184 -0.010208405
##
##
## $aic
## [1] -2581.950 -2586.951 -2590.390 -2590.174 -2593.039 -2591.882 -2590.464
```

```
# start from choosen pollutens
chosen.pos = colnames(data) [chosen.po.ind]
expr = paste("length~", paste(chosen.pos, collapse = "+"))
t=forward.change(newdata, expr)
```

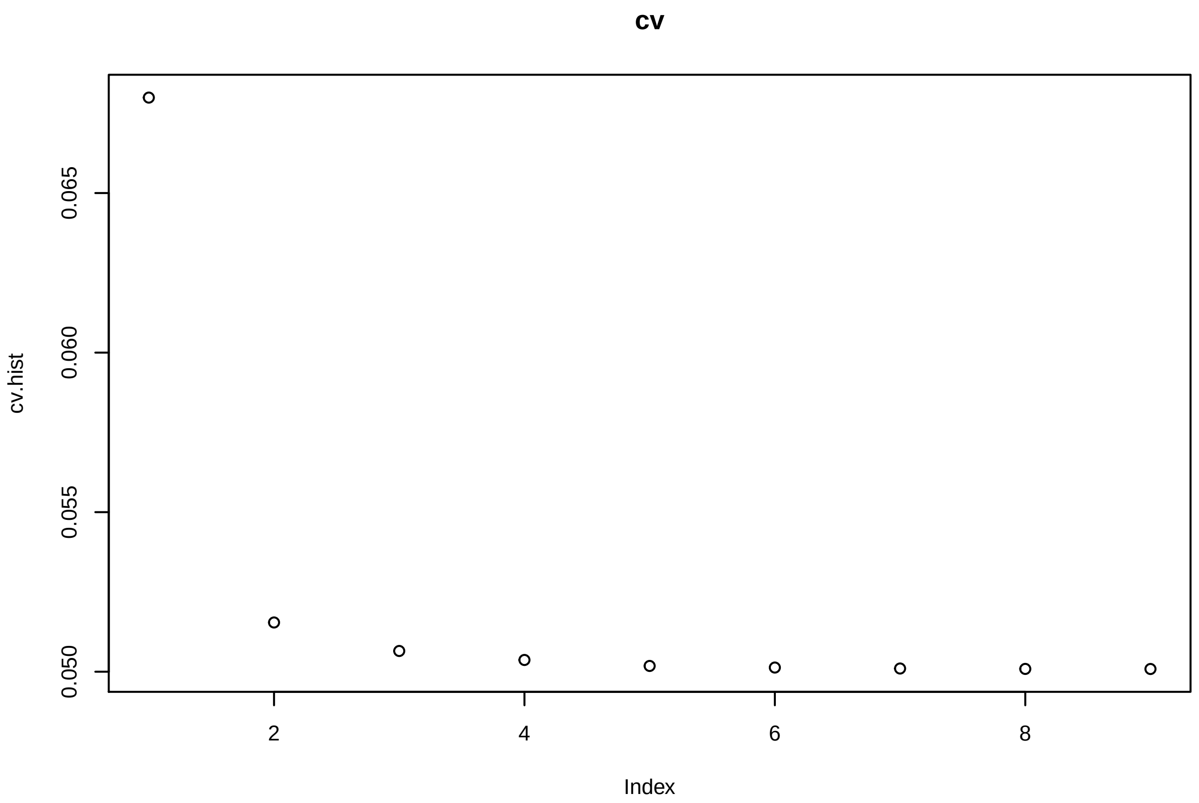
```
## [1] "step 1"
## [1] "added ageyrs"
## [1] "step 2"
## [1] "added race_cat"
## [1] "step 3"
## [1] "added male"
## [1] "step 4"
## [1] "added BMI"
## [1] "step 5"
## [1] "added eosinophils_pct"
## [1] "step 6"
## [1] "added POP_furan1"
## [1] "step 7"
## [1] "added neutrophils_pct"
## [1] "step 8"
## [1] "added POP_PCB5"
## [1] "step 9"
## [1] "done choosing model"
```

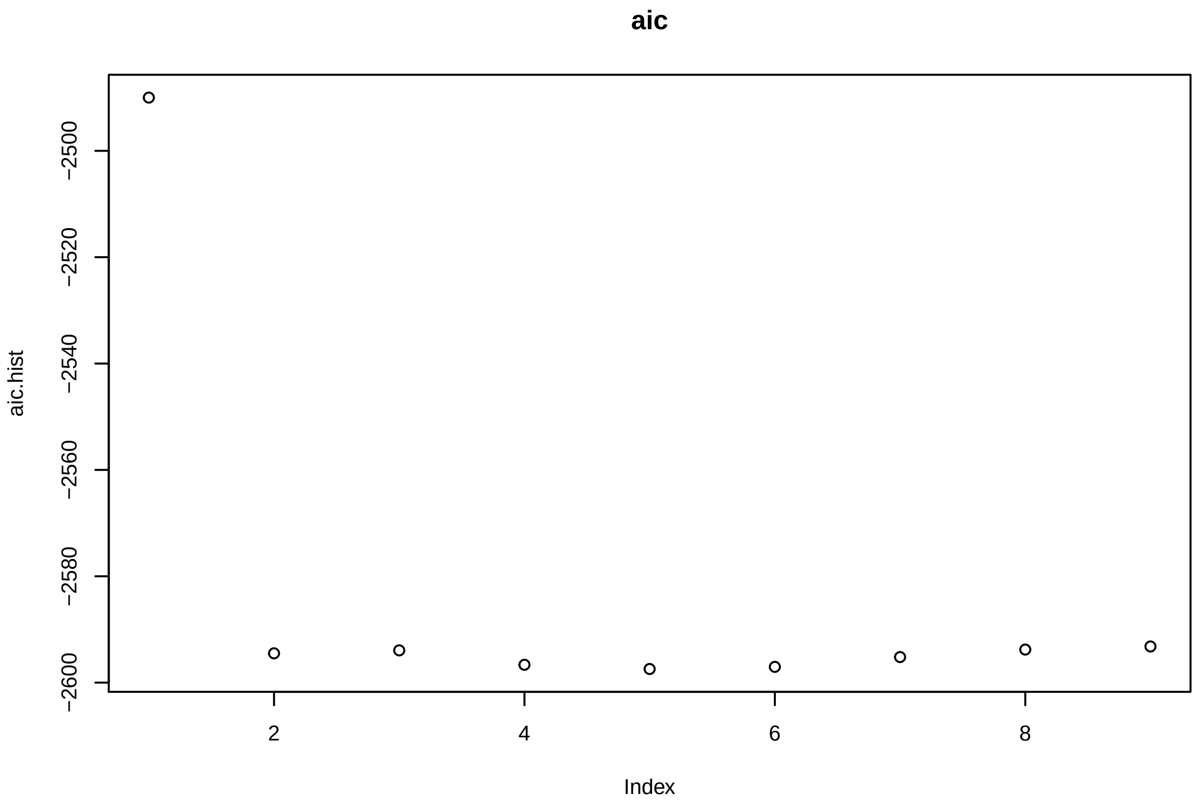

cv
cv.hist
0.050
0.055
0.060
0.065
2
4
6
8
Index

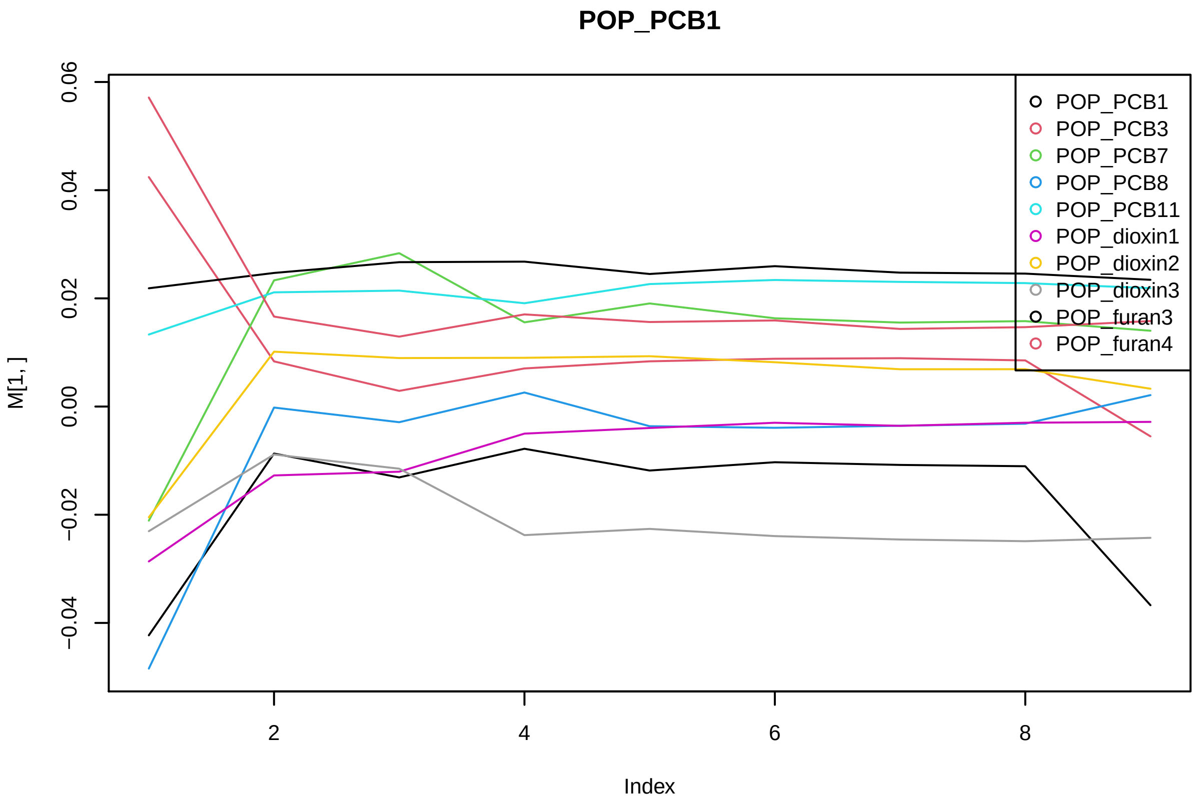


```
chosen.pos = colnames(data) [chosen.po.ind]
expr = paste("length~", paste(chosen.pos, collapse = "+"))
t=forward.change(newdata, expr, TRUE)
```

```
## [1] "step 1"
## [1] "added ageyrs"
## [1] "step 2"
## [1] "added race_cat"
## [1] "step 3"
## [1] "added male"
## [1] "step 4"
## [1] "added BMI"
## [1] "step 5"
## [1] "added eosinophils_pct"
## [1] "step 6"
## [1] "added POP_furan1"
## [1] "step 7"
## [1] "added neutrophils_pct"
## [1] "step 8"
## [1] "added POP_PCB5"
## [1] "step 9"
## [1] "done choosing model"
```


cv
cv.hist
0.050
0.055
0.060
0.065
Index
2
4
6
8


aic
aic.hist
−2500
−2520
−2540
−2560
−2580
−2600
2
4
6
8
Index



```
# forward start from lm(length~1) done
# chosen pollute + other by forward done
# error analysis
# visualize the smoke stuff
# how the coefficients vary
```