Stat331 final project

Setup

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-1
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
pollutants_original <- read.csv("pollutants.csv", header = TRUE)
pollutants <- pollutants_original[-1]
```

```
gender_list <- c('female', 'male')
edu_list <- c('before_high_school', 'high_school', 'college', 'college_grad')
eth_list <- c('other', 'mexi_us', 'nonhisp_black', 'nonhisp_white')
smoke_list <- c('no', 'yes')

gender <- gender_list[pollutants$male + 1]
education <- edu_list[pollutants$edu_cat]
race <- eth_list[pollutants$race_cat]
smoke_now <- smoke_list[pollutants$smokenow + 1]
```

```
pollutants$male <- factor(gender)
pollutants$edu_cat <- factor(education)
pollutants$race_cat <- factor(race)
pollutants$smokenow <- factor(smoke_now)

pollutants$male <- factor(gender, levels=gender_list)
pollutants$edu_cat <- factor(education, levels=edu_list)
pollutants$race_cat <- factor(race, levels=eth_list)
pollutants$smokenow <- factor(smoke_now, levels=smoke_list)
```

```
## Factors removed
pollutants_factorsRemoved <- pollutants[-c(27,28,29,32)]

pollutants_factorsRemovedCor <- cor(pollutants_factorsRemoved)
```

Train and test data

```r
set.seed(57)

N <- nrow(pollutants)
sampleTrain <- sample(1:N, round(N*0.8,0), replace = FALSE)

dataTrain <- pollutants[sampleTrain,]
dataTest <- pollutants[-sampleTrain,]
```

Bounds for model selection

```r
M0 <- lm(length ~ 1, data = dataTrain)
Mfull <- lm(length ~ ., data = dataTrain)
```

```r
## Forward selection using AIC
MfwdAIC <- step(object = M0, scope = list(lower = M0, upper = Mfull),
             trace = FALSE, direction = "forward", k = 2)
summary(MfwdAIC)
```

```
##
## Call:
## lm(formula = length ~ ageyrs + POP_furan3 + male + edu_cat +
##     ln_lbxcot, data = dataTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49518 -0.15410 -0.02362  0.11880  1.17749
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.3488676  0.0290273  46.469  < 2e-16 ***
## ageyrs              -0.0070550  0.0005705 -12.366  < 2e-16 ***
## POP_furan3           0.0065961  0.0017063   3.866 0.000121 ***
## malemale            -0.0438412  0.0174596  -2.511 0.012269 *
## edu_cathigh_school   0.0109026  0.0236026   0.462 0.644283
## edu_catcollege       0.0584597  0.0224159   2.608 0.009307 **
## edu_catcollege_grad  0.0443406  0.0245573   1.806 0.071422 .
## ln_lbxcot            0.0042414  0.0023275   1.822 0.068843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2208 on 683 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.2312
## F-statistic: 30.64 on 7 and 683 DF,  p-value: < 2.2e-16
```

```r
## Prediction accuracy
MfwdAIC.res <- pollutants$length[-sampleTrain] - predict(MfwdAIC, newdata = dataTest)
mspeMfwdAIC <- mean(MfwdAIC.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMfwdAIC))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0497111774047571"
```

```r
## Forward selection using BIC
MfwdBIC <- step(object = M0, scope = list(lower = M0, upper = Mfull),
             trace = FALSE, direction = "forward", k = log(nrow(dataTrain)))
summary(MfwdBIC)
```

```
##
## Call:
## lm(formula = length ~ ageyrs + POP_furan3, data = dataTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52126 -0.15948 -0.02463  0.12493  1.16255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3721521  0.0242857  56.500  < 2e-16 ***
## ageyrs      -0.0074886  0.0005613 -13.342  < 2e-16 ***
## POP_furan3   0.0070227  0.0017097   4.108 4.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2225 on 688 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2192
## F-statistic: 97.84 on 2 and 688 DF,  p-value: < 2.2e-16
```

```r
## Prediction accuracy
MfwdBIC.res <- pollutants$length[-sampleTrain] - predict(MfwdBIC, newdata = dataTest)
mspeMfwdBIC <- mean(MfwdBIC.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMfwdBIC))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0487280632244949"
```

```r
## Backward selection using AIC
MbckAIC <- step(object = Mfull, scope = list(lower = M0, upper = Mfull),
                trace = FALSE, direction = "backward", k = 2)
summary(MbckAIC)
```

```
##
## Call:
## lm(formula = length ~ POP_PCB1 + POP_PCB3 + POP_furan3 + BMI +
##     edu_cat + race_cat + male + ageyrs + smokenow, data = dataTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48574 -0.15262 -0.02903  0.12360  1.18525
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.416e+00  5.894e-02  24.023  < 2e-16 ***
## POP_PCB1               -6.249e-07  3.452e-07  -1.810  0.07071 .
## POP_PCB3                2.140e-06  1.199e-06   1.784  0.07482 .
```

```
## POP_furan3             5.971e-03  1.935e-03   3.086  0.00211 **
## BMI                   -2.068e-03  1.471e-03  -1.406  0.16021
## edu_cathigh_school     1.208e-02  2.460e-02   0.491  0.62356
## edu_catcollege         5.928e-02  2.332e-02   2.542  0.01124 *
## edu_catcollege_grad    4.094e-02  2.674e-02   1.531  0.12619
## race_catmexi_us       -5.201e-02  3.628e-02  -1.433  0.15221
## race_catnonhisp_black  2.142e-03  3.708e-02   0.058  0.95395
## race_catnonhisp_white -4.497e-02  3.332e-02  -1.349  0.17766
## malemale              -3.438e-02  1.753e-02  -1.961  0.05023 .
## ageyrs                -6.715e-03  6.347e-04 -10.580  < 2e-16 ***
## smokenowyes            3.442e-02  2.092e-02   1.645  0.10043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2202 on 677 degrees of freedom
## Multiple R-squared:  0.2499, Adjusted R-squared:  0.2355
## F-statistic: 17.35 on 13 and 677 DF,  p-value: < 2.2e-16
```

```
## Prediction accuracy
MbckAIC.res <- pollutants$length[-sampleTrain] - predict(MbckAIC, newdata = dataTest)
mspeMbckAIC <- mean(MbckAIC.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMbckAIC))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0504995925627993"
```

```
## Backward selection using BIC
MbckBIC <- step(object = Mfull, scope = list(lower = M0, upper = Mfull),
                trace = FALSE, direction = "backward", k = log(nrow(dataTrain)))
summary(MbckBIC)
```

```
##
## Call:
## lm(formula = length ~ POP_furan3 + ageyrs, data = dataTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52126 -0.15948 -0.02463  0.12493  1.16255
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3721521  0.0242857  56.500  < 2e-16 ***
## POP_furan3   0.0070227  0.0017097   4.108 4.48e-05 ***
## ageyrs      -0.0074886  0.0005613 -13.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2225 on 688 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2192
## F-statistic: 97.84 on 2 and 688 DF,  p-value: < 2.2e-16
```

```
## Prediction accuracy
MbckBIC.res <- pollutants$length[-sampleTrain] - predict(MbckBIC, newdata = dataTest)
mspeMbckBIC <- mean(MbckBIC.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMbckBIC))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0487280632244948"
```

Correlation testing for PCB 1,2,4,5

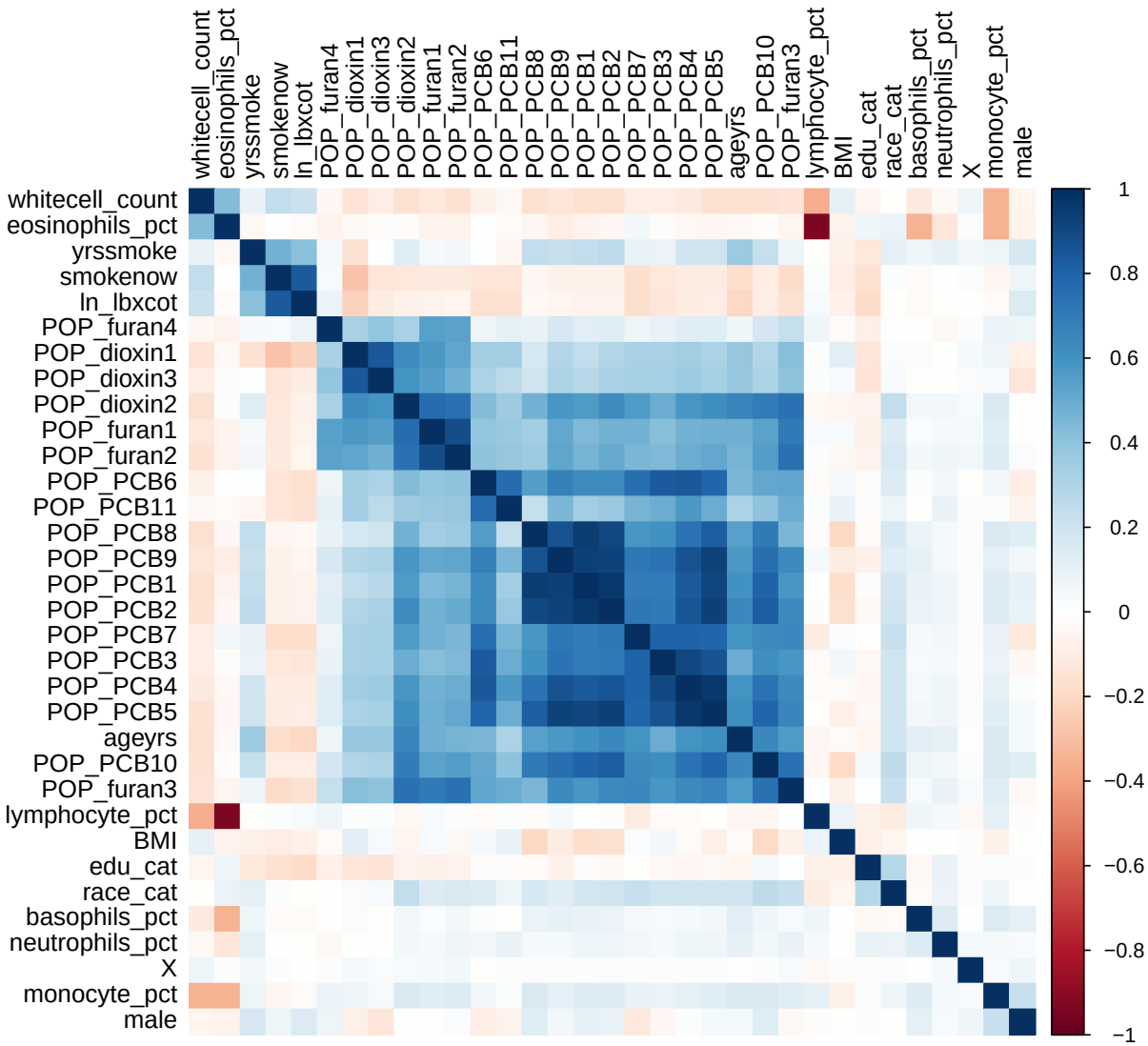
```
pollutants_pcbs <- dataTrain[,-c(2,3,5,6)]
cor(pollutants_pcbs[,2:8])
```

```
##            POP_PCB3  POP_PCB6  POP_PCB7  POP_PCB8  POP_PCB9 POP_PCB10 POP_PCB11
## POP_PCB3  1.0000000 0.8324176 0.8036781 0.6000042 0.7380250 0.6118652 0.4952676
## POP_PCB6  0.8324176 1.0000000 0.7501541 0.5527229 0.6733482 0.5142743 0.7615891
## POP_PCB7  0.8036781 0.7501541 1.0000000 0.5847329 0.7126358 0.6393772 0.4673341
## POP_PCB8  0.6000042 0.5527229 0.5847329 1.0000000 0.8628308 0.7098066 0.2389620
## POP_PCB9  0.7380250 0.6733482 0.7126358 0.8628308 1.0000000 0.7543283 0.4534208
## POP_PCB10 0.6118652 0.5142743 0.6393772 0.7098066 0.7543283 1.0000000 0.4046687
## POP_PCB11 0.4952676 0.7615891 0.4673341 0.2389620 0.4534208 0.4046687 1.0000000
```

Manual analysis of variate correlation. This is done solely against the training set to avoid leakage of information from the holdout set, which would negatively impact the accuracy of model performance evaluation.

```
# Correlation matrix
dataTrain_num <- pollutants_original[row.names(dataTrain),]
dataTrain_cor <- cor(dataTrain_num[!colnames(dataTrain_num) %in% c('length')])

# Correlation plot
corrplot(dataTrain_cor, method="color", type="full", order="hclust",  tl.col = "black")
```



Based on the correlation plot, the following variates are correlated with other variates:PCB 1,3,4,5,8,9,10, furan 1, dioxin 1,2, Lymphocyte_pct. We remove them from the dataset and analyze the resulting correlation plot.

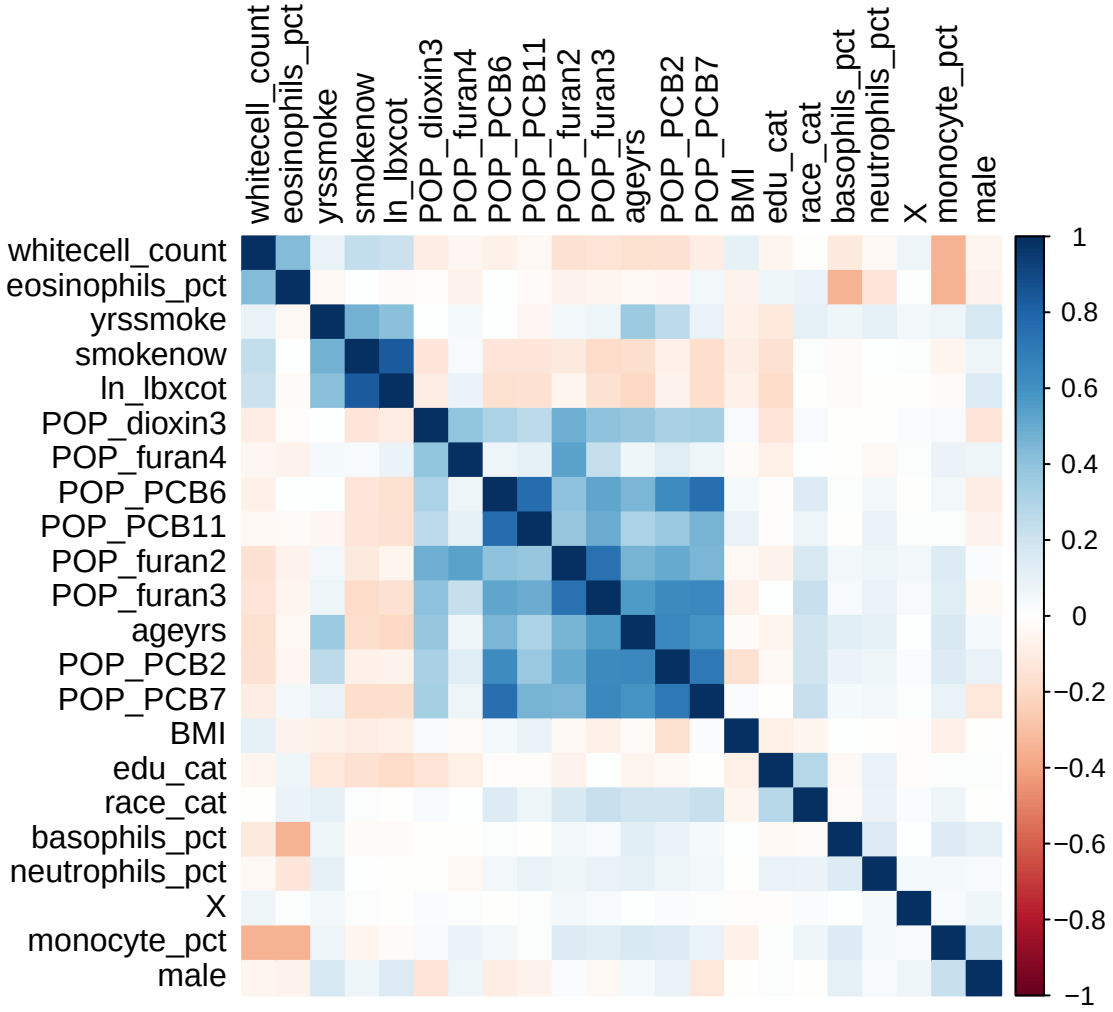
```
# New datasets
cols_to_exclude <- c('POP_PCB1','POP_PCB3','POP_PCB4','POP_PCB5','POP_PCB8',
                     'POP_PCB9','POP_PCB10','POP_furan1','POP_dioxin1',
                     'POP_dioxin2','lymphocyte_pct')
pollutants_corRemoved <- pollutants[!colnames(pollutants) %in% cols_to_exclude]
```

```
dataTrain_corRemoved <- dataTrain[!colnames(dataTrain) %in% cols_to_exclude]

# Correlation matrix
dataTrain_corRemoved_num <- pollutants_original[
  sampleTrain,
  !colnames(pollutants_original) %in% cols_to_exclude]

dataTrain_corRemoved_cor <- cor(
  dataTrain_corRemoved_num[
    !colnames(dataTrain_corRemoved_num) %in% c('length')])

# Correlation plot
corrplot(dataTrain_corRemoved_cor, method="color", type="full", order="hclust",
         tl.col = "black")
```



Stepwise model selection with PCB, Furan, Dioxin, and other correlated features removed

```r
# From Discord:
# Doing this [setting dataTrain2<-dataTrain, etc.] because we just did the manual variable selection ba
# For now, will just set dataTrain2 <- dataTrain and dataTest2 <- dataTest
# Later we can go through and remove all the 2's. I just don't want any conflicts rn in case anyone was

# Originally, this section defined a second training/test set. Changed this
set.seed(57)

N2 <- N
sampleTrain2 <- sampleTrain

dataTrain2 <- dataTrain
dataTest2 <- dataTest
```

Bounds for model selection

```r
M0_corRemoved <- lm(length ~ 1, data = dataTrain2)
Mfull_corRemoved <- lm(length ~ ., data = dataTrain2)
```

```r
## Forward selection using AIC
MfwdAIC_corRemoved <- step(object = M0_corRemoved, scope = list(lower = M0_corRemoved, upper = Mfull_co
             trace = FALSE, direction = "forward", k = 2)
summary(MfwdAIC_corRemoved)
```

```
##
## Call:
## lm(formula = length ~ ageyrs + POP_furan3 + male + edu_cat +
##     ln_lbxcot, data = dataTrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49518 -0.15410 -0.02362  0.11880  1.17749
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.3488676  0.0290273  46.469  < 2e-16 ***
## ageyrs              -0.0070550  0.0005705 -12.366  < 2e-16 ***
## POP_furan3           0.0065961  0.0017063   3.866 0.000121 ***
## malemale            -0.0438412  0.0174596  -2.511 0.012269 *
## edu_cathigh_school   0.0109026  0.0236026   0.462 0.644283
## edu_catcollege       0.0584597  0.0224159   2.608 0.009307 **
## edu_catcollege_grad  0.0443406  0.0245573   1.806 0.071422 .
## ln_lbxcot            0.0042414  0.0023275   1.822 0.068843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2208 on 683 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.2312
## F-statistic: 30.64 on 7 and 683 DF,  p-value: < 2.2e-16
```

```r
## Prediction accuracy
MfwdAIC_corRemoved.res <- pollutants$length[-sampleTrain2] - predict(MfwdAIC_corRemoved, newdata = data
```

```r
mspeMfwdAIC_corRemoved <- mean(MfwdAIC_corRemoved.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMfwdAIC_corRemoved))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0497111774047571"
```

```r
## Forward selection using BIC
MfwdBIC_corRemoved <- step(object = M0_corRemoved, scope = list(lower = M0_corRemoved, upper = Mfull_co
            trace = FALSE, direction = "forward", k = log(nrow(dataTrain2)))
summary(MfwdBIC_corRemoved)
```

```
##
## Call:
## lm(formula = length ~ ageyrs + POP_furan3, data = dataTrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52126 -0.15948 -0.02463  0.12493  1.16255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3721521  0.0242857  56.500  < 2e-16 ***
## ageyrs      -0.0074886  0.0005613 -13.342  < 2e-16 ***
## POP_furan3   0.0070227  0.0017097   4.108 4.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2225 on 688 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2192
## F-statistic: 97.84 on 2 and 688 DF,  p-value: < 2.2e-16
```

```r
## Prediction accuracy
MfwdBIC_corRemoved.res <- pollutants$length[-sampleTrain2] - predict(MfwdBIC_corRemoved, newdata = data
mspeMfwdBIC_corRemoved <- mean(MfwdBIC_corRemoved.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMfwdBIC_corRemoved))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0487280632244949"
```

```r
## Backward selection using AIC
MbckAIC_corRemoved <- step(object = Mfull_corRemoved, scope = list(lower = M0_corRemoved, upper = Mfull
              trace = FALSE, direction = "backward", k = 2)
summary(MbckAIC_corRemoved)
```

```
##
## Call:
## lm(formula = length ~ POP_PCB1 + POP_PCB3 + POP_furan3 + BMI +
##     edu_cat + race_cat + male + ageyrs + smokenow, data = dataTrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.48574 -0.15262 -0.02903  0.12360  1.18525
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.416e+00  5.894e-02  24.023  < 2e-16 ***
## POP_PCB1               -6.249e-07  3.452e-07  -1.810  0.07071 .
## POP_PCB3                2.140e-06  1.199e-06   1.784  0.07482 .
## POP_furan3              5.971e-03  1.935e-03   3.086  0.00211 **
## BMI                    -2.068e-03  1.471e-03  -1.406  0.16021
## edu_cathigh_school      1.208e-02  2.460e-02   0.491  0.62356
## edu_catcollege          5.928e-02  2.332e-02   2.542  0.01124 *
## edu_catcollege_grad     4.094e-02  2.674e-02   1.531  0.12619
## race_catmexi_us        -5.201e-02  3.628e-02  -1.433  0.15221
## race_catnonhisp_black   2.142e-03  3.708e-02   0.058  0.95395
## race_catnonhisp_white  -4.497e-02  3.332e-02  -1.349  0.17766
## malemale               -3.438e-02  1.753e-02  -1.961  0.05023 .
## ageyrs                 -6.715e-03  6.347e-04 -10.580  < 2e-16 ***
## smokenowyes             3.442e-02  2.092e-02   1.645  0.10043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2202 on 677 degrees of freedom
## Multiple R-squared:  0.2499, Adjusted R-squared:  0.2355
## F-statistic: 17.35 on 13 and 677 DF,  p-value: < 2.2e-16
```

```
## Prediction accuracy
MbckAIC_corRemoved.res <- pollutants$length[-sampleTrain2] - predict(MbckAIC_corRemoved, newdata = data
mspeMbckAIC_corRemoved <- mean(MbckAIC_corRemoved.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMbckAIC_corRemoved))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0504995925627993"
```

```
## Backward selection using BIC
MbckBIC_corRemoved <- step(object = Mfull_corRemoved, scope = list(lower = M0_corRemoved, upper = Mfull
               trace = FALSE, direction = "backward", k = log(nrow(dataTrain2)))
summary(MbckBIC_corRemoved)
```

```
##
## Call:
## lm(formula = length ~ POP_furan3 + ageyrs, data = dataTrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52126 -0.15948 -0.02463  0.12493  1.16255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3721521  0.0242857  56.500  < 2e-16 ***
## POP_furan3   0.0070227  0.0017097   4.108 4.48e-05 ***
## ageyrs      -0.0074886  0.0005613 -13.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2225 on 688 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2192
## F-statistic: 97.84 on 2 and 688 DF,  p-value: < 2.2e-16
```

```
## Prediction accuracy
MbckBIC_corRemoved.res <- pollutants$length[-sampleTrain2] - predict(MbckBIC_corRemoved, newdata = data
mspeMbckBIC_corRemoved <- mean(MbckBIC_corRemoved.res^2)

print(paste("MSPE of foward selection model based on AIC:",mspeMbckBIC_corRemoved))
```

```
## [1] "MSPE of foward selection model based on AIC: 0.0487280632244948"
```