Final Project

Shuhao Liu, Bingjie Shen, Yuhan Zhao, Aurora Zhao

14/04/2021

Summary

The objective of this report is to investigate the relationship between leukocyte telomere length and organic pollutants along with other variables and build an interpretable model that fits the data properly and produces accurate predictions.

In order to reduce stand error and produce stable estimates, we need to first eliminate variables with high multicollinearity by using VIF. We removed all variables whose VIF $>$ 8 which indicates high multicollinearity. Next, we use LASSO to do the automatic selection to get the reduced model.

After getting the reduced model, we take the goodness of fit, assumptions, intersection, outliers, and influence measure into consideration and analyze how these factors might affect our results. We then modify our model accordingly to improve fitness and prediction accuracy.

Objective

This report will mainly focus on creating a fitted model which will more precisely and efficiently argue the relationship between the mean leukocyte telomere length and the exposures to persistent organic pollutants and other potentially correlated factors.

Furthermore, analyzing different variates in this dataset separately which will provide us a knowledge of the marker of cellular aging that may be related to certain cancers. Also, what factors might cause the mean leukocyte telomere length to increase and what may cause this to decrease.

Apart from that, instead of investigating the pollutants and other features separately, we will also focus on whether there are any interactions between covariates, and how will this effect our final fitted model if such interactions exist.

Exploratory Data Analysis

This dataset provides information of 864 patients. There are 27 covariates included in the dataset with 23 numerical covariates and 4 categorical covariates.

We can start with analyzing the response variable. Mean leukocyte telomere length in our observations range from 0.5265724 to 2.3512373, with a mean length of 1.0543127. The first quartile of the telomere length is 0.8754077, and the third quartile is 1.2095474.

Also, we want to investigate categorical covariates, and here are some findings:

There are 374 males and 490 females in this study. About 31% of adults investigated have less than 9th grade or 9-11th grade; 23% of them have high school diploma; 26% of them have collage or AA degree; 19% of them have collage graduate degree. Only 23% of adults investigated currently smoke. 22% adults investigated are Mexican American, 18% are non-Hispanic black, 52% are non-Hispanic white, and 8% are other races. About a half of all the adults investigated are non-Hispanic white.

Now, let's choose some covariates in interest to analyze their summary statistics. We choose ageyrs (age in years), yrssmoke(years smoke) and a pollutant variable POP_PCB1.

For age in years (ageyrs), this study investigated 864 adults aging from 20 to 85, where the mean is 48.36. The first quantile is 34 and the third quantile is 63, so this study mainly investigated middle-aged adults.

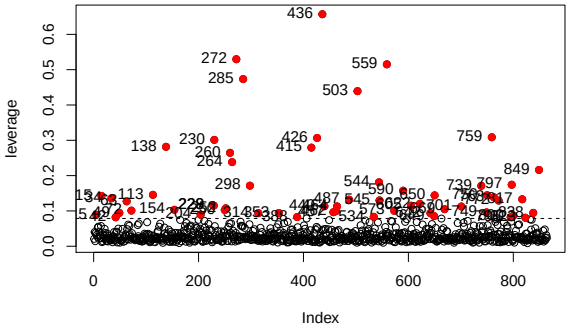
For years of smoke (yrssmoke), there are 472 non-smokers and 392 smokers. Apart from non-smokers, the mean of years of smoke of smokers is 23.37 and the maximum years is 69 years.

For POP_PCB1, the mean concentration of this pollutant is 3.8082176×10^4 range from 2000 to 5.72×10^5.

Outliers analysis

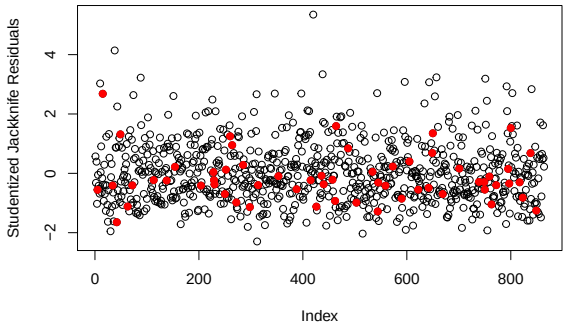
Outliers are unusual or extreme observations. They are typically caused by data entry error or Unusual observation. It is crucial to identify outliers as they may have a big impact on our results.

Leverage is a statistics that can help us identify x-outliers, it tells us whether y_i is close to y_i. The "rule of Thumb" of identifying points with large leverage is if the leverage is bigger than the threshold $\bar{h}$: $\bar{h} = 2 \times \frac{1}{n} \times (p+1)$, where p is the number of covariates. We will plot the leverage against the index below and hight the points with high leverage in red.



There are 59 data values exceed the leverage threshold, and we can tell from the plot that there are some points with extremely high leverage. However, but this statistics may not be

conclusive if the corresponding y values are also outliers. We can solve this issue by checking the jacknike residuals.



We plotted the studentized Jackknife Residuals against their index and highlighted the points with large leverage in red. From the plot, the can see that the points with large leverage have fairly small jacknike residuals, which means the observations are probably just very extreme, but are not observation errors and will not cause big issue in model building

Methods

Our model

The reduced model we finally get is:

```
Call:
lm(formula = length ~ . + POP_furan3 * POP_PCB11, data = pollutants_reduced[-omit_in
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.46766 -0.14768 -0.02205  0.12591  0.73724

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.429e+00  4.951e-02  28.867  < 2e-16 ***
POP_PCB11                   1.334e-03  4.727e-04   2.822  0.00488 **
POP_furan3                  6.809e-03  2.331e-03   2.921  0.00359 **
monocyte_pct               -5.146e-03  3.829e-03  -1.344  0.17926
BMI                        -2.457e-03  1.254e-03  -1.960  0.05031 .
ageyrs                     -6.993e-03  5.970e-04 -11.714  < 2e-16 ***
yrssmoke                   -4.992e-04  6.184e-04  -0.807  0.41972
ln_lbxcot                   3.461e-03  2.361e-03   1.466  0.14305
male2$male                 -2.581e-02  1.541e-02  -1.675  0.09434 .
edu_cat2$college_students   3.012e-02  1.658e-02   1.816  0.06965 .
race_cat2$Mexican_American -3.804e-03  1.870e-02  -0.203  0.83890
race_cat2$Non_Hispanic_Black 5.009e-02  1.975e-02   2.535  0.01141 *
POP_PCB11:POP_furan3       -6.058e-05  2.861e-05  -2.118  0.03449 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2103 on 840 degrees of freedom
Multiple R-squared:  0.2566, Adjusted R-squared:  0.246
F-statistic: 24.17 on 12 and 840 DF,  p-value: < 2.2e-16
```

Transformations or extensions of the basic multiple linear regression model

Now, let us turn our attention on the categorical variables.

Intuitively, instead of individually effect the mean leukocyte telomere length, there might be some correlation between the them. For instance, only one pollutant may not effect the

mean leukocyte telomere length dramatically, but if another pollutant is also found, there might be some chemical interaction which will probably make the condition more severe than individually do.

Notice that in our final reduced model, we eliminate the affects of all organic pollutants except POP_PCB11 and POP_furan3. We conjecture that whether there is an interaction between POP_PCB11 and POP_furan3?

We constructed a new model which adds an interaction between POP_PCB11 and POP_furan3. The analysis of goodness of fit will be covered in `Result` section.

Process of finding the model

First, we removed covariates with VIF greater than 8 and converted numerical covariates to categorical covariates. Then we used LASSO to derive a reduced model that fits the data better.

We also used cross validation to measure prediction accuracy with mean squared prediction error. After that, we estimated betas for minimum lambda and merged categories with 0 coefficients with their referent group.

After all those steps, we construct a reduced model with only 11 variates as shown above.

Goodness of fit of the model

To prove that our model performs better as we expected, we compared the four criteria with backward selection, forward selection and stepwise selection. All four criteria shows evidence that our reduced model actually fits the data pretty well.

Assumptions

We conducted an assumption analysis proving that the linearity assumption and the independence assumption are not broken, while the normality assumption and the heteroskedasticity assumption are slightly broken.

Results

VIF

According to our model, we choose to remove the covariates with VIF > 8.

Note that we did this before converting categorical covariates to numerical covariates, as the output cannot be categorical.

After removing the variates with large multicollinearity, we are left with: 27 variables, which are:

```
 [1] "POP_PCB3"        "POP_PCB7"        "POP_PCB8"        "POP_PCB9"
 [5] "POP_PCB10"       "POP_PCB11"       "POP_dioxin1"     "POP_dioxin2"
 [9] "POP_dioxin3"     "POP_furan1"      "POP_furan2"      "POP_furan3"
[13] "POP_furan4"      "whitecell_count" "lymphocyte_pct"  "monocyte_pct"
[17] "basophils_pct"   "neutrophils_pct" "BMI"             "edu_cat"
[21] "race_cat"        "male"            "ageyrs"          "yrssmoke"
[25] "smokenow"        "ln_lbxcot"       "length"
```

Convert the categorical variables

In the dataset given, the categorical covariates are recorded by numerical values, to actually illustrate the relationship between those categories and the mean leukocyte telomere length, first we need to convert those numerical values into meaningful categories.

The covariates we get after converting the categorical variable are:

```
"(Intercept)"                 "POP_PCB3"                    "POP_PCB7"
"POP_PCB8"                    "POP_PCB9"                    "POP_PCB10"
"POP_PCB11"                   "POP_dioxin1"                 "POP_dioxin2"
"POP_dioxin3"                 "POP_furan1"                  "POP_furan2"
"POP_furan3"                  "POP_furan4"                  "whitecell_count"
"lymphocyte_pct"              "monocyte_pct"                "basophils_pct"
"neutrophils_pct"             "BMI"                         "edu_cat$high_school_grads"
"edu_cat$college_students"    "edu_cat$college_grads"       "race_cat$Mexican_American"
"race_cat$Non_Hispanic_Black" "race_cat$Non_Hispanic_White" "male$male"
"ageyrs"                      "yrssmoke"                    "smokenow$Smoke"
"ln_lbxcot"
```

LASSO

LASSO stands for least absolute shrinkage and selection operator. It is a very popular and modern variable selection method. We use LASSO to do the automatic selection to derive our reduced model.

After applying LASSO, we get a set of covariates with 0 coefficients:

```
"POP_PCB3"                  "POP_PCB7"                  "POP_PCB8"
"POP_PCB9"                  "POP_PCB10"                 "POP_dioxin1"
"POP_dioxin2"               "POP_dioxin3"               "POP_furan1"
"POP_furan2"                "POP_furan4"                "whitecell_count"
"lymphocyte_pct"            "basophils_pct"             "neutrophils_pct"
"edu_cat$high_school_grads" "edu_cat$college_grads"     "race_cat$Non_Hispanic_White"
"smokenow$Smoke"
```

This indicates the fact that these covariates have little relationship between the mean outcome, therefore, based on the result of LASSO, we have enough evidence to remove those variates.

Merge the categories with 0 coefficients

However, apart from that, categorical variables are still a concern. As stated in the lecture, when fitting the model, we choose one category as the referent group. Referent group helps us better understand the effect of the non-referent group on the response variable. More precisely, the estimated coefficient of a non-referent group can be interpreted as the mean difference in the mean outcome between the referent group and that non-referent group holding other factors constant. Intuitively, if the coefficient of a non-referent group is 0, this will draw the fact that there is no difference between the referent group and the non-referent group in interest.

Based on this knowledge, to deal with the non-referent categories with 0 coefficients, we could merge those categories with their referent group to get a more precise on the estimation of our response.

1. Patients with education level: Less Than 9th Grade or 9-11th Grade (Includes 12th grade with no diploma), High School Grad/GED or Equivalent and College Graduate are merged into a group called `edu_cat2$Others`, Therefore, we only divide the education level into two categories: Some College or AA degree and Others.

2. Patients with race: Non-Hispanic White and Other Race (Including Multi-Racial) are merged into a group called `race_cat2$Others`. Therefore, we only divide the race into 3 categories: Mexican_American, Non_Hispanic_Black and others.

Final Reduced Model

To sum up everything stated so far, our final reduced model can be shown as:

```
Call:
lm(formula = length ~ ., data = pollutants_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4908 -0.1504 -0.0295  0.1285  1.1998

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.4234731  0.0507823  28.031  < 2e-16 ***
POP_PCB11                   0.0001461  0.0001656   0.882  0.37805
POP_furan3                  0.0048931  0.0017593   2.781  0.00553 **
monocyte_pct               -0.0042886  0.0037496  -1.144  0.25306
BMI                        -0.0016194  0.0012801  -1.265  0.20619
ageyrs                     -0.0065326  0.0005865 -11.138  < 2e-16 ***
yrssmoke                   -0.0005918  0.0006362  -0.930  0.35254
ln_lbxcot                   0.0041139  0.0024615   1.671  0.09503 .
male2$male                 -0.0319989  0.0158885  -2.014  0.04433 *
edu_cat2$college_students   0.0297646  0.0172469   1.726  0.08475 .
race_cat2$Mexican_American -0.0110103  0.0195266  -0.564  0.57300
race_cat2$Non_Hispanic_Black 0.0380976 0.0205916   1.850  0.06464 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2207 on 852 degrees of freedom
Multiple R-squared:  0.2324, Adjusted R-squared:  0.2225
F-statistic: 23.45 on 11 and 852 DF,  p-value: < 2.2e-16
```

Goodness of fit

Getting the final reduced model, we want to actually check whether this performs better than the original data (after removing variables with VIF > 8) and also the models fitted using Backward Selection, Forward Selection and Stepwise Selection.

We've created a table comparing the four criteria (AIC, BIC, Adjusted R^2 and MSE) in terms of goodness of fit.

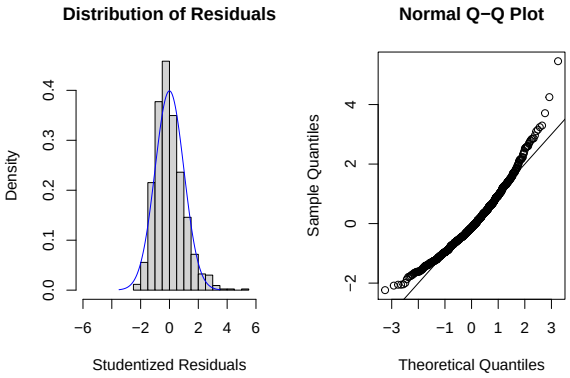
```
##                            AIC       BIC Adj R Sqr        MSE
## Original Data      -118.4352  33.93508 0.2147863 0.04916910
## Forward Selection  -118.4352  33.93508 0.2147863 0.04916910
## Backward Selection -118.4352  33.93508 0.2147863 0.04916910
## Stepwise Selection -118.4352  33.93508 0.2147863 0.04916910
## LASSO              -145.4452 -83.54478 0.2224694 0.04868799
```

Notice that, compared with original data and the models we got after applying forward selection, backward selection and stepwise selection. We can observe a dramatic decrease of both AIC and BIC, a slightly increase in the adjusted R^2 and a slightly decrease in the MSE.

Recall that we prefer model with lower AIC and BIC, higher adjusted R^2 and lower MSE. Based on the analysis of those four criteria of goodness of fit, this demonstrates the fact that our final reduced model fits the data better.

Assumptions of the reduced model

Normality

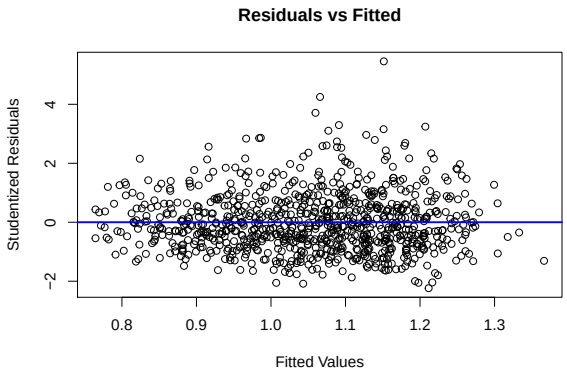


There are 864 observations in this dataset. By Central Limit Theorem, beta hat is approximately Normally distributed in a large sample. The two plots above suggest that Normality assumption is slightly broken:

(1) the histogram of studentized residuals have the basic shape of a standard normal, but is obviously skewed

(2) dots in the qq-plot mostly fall on 45 degree line, but have have a clear curved shape

Heteroskedasticity

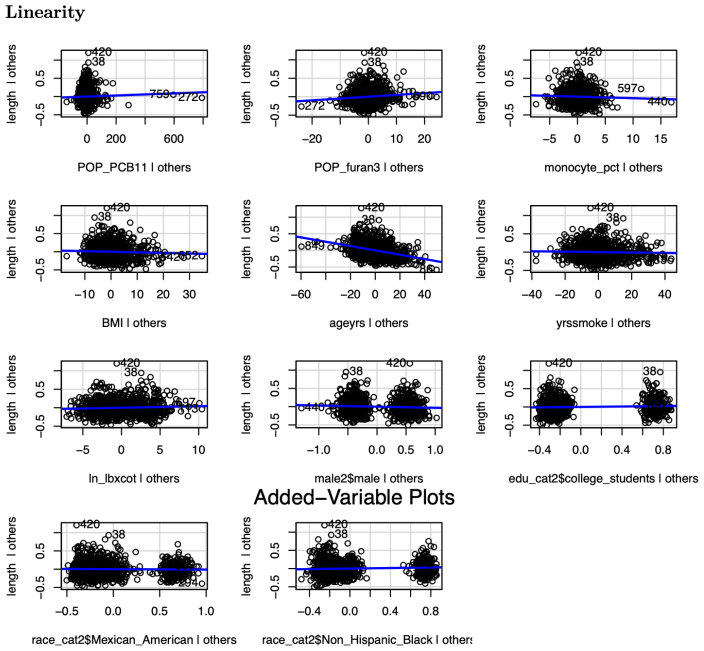


This plot suggests that Heteroskedasticity assumption is somewhat broken, since there isn't a constantly amount of variability across different levels of fitted values. Specifically, the variability is less at the lower end and higher at the end of the fitted values.

The third assumption we are supposed to focus on is linearity. To assess the linearity in x^* (x^* is an arbitrary covariate), first regress the response on other covariates except x^*,

next regress x^* on other covariates. Finally plot the residual of the first model fit against the latter one. Through investigating the plots, we are able to analyse the linearity of our reduced model.

Linearity



Notice that we do not observe any obvious pattern of each scatterplot, therefore, there is no evidence to reject the fact that the linearity is not broken.

Independence

This dataset contains a sample of n=864 adults. We can assume the sample selection is completely random, then Independence assumption is not broken.

To sum up, the linearity assumption and the independence assumption are not broken, while the normality assumption and the heteroskedasticity assumption are slightly broken.

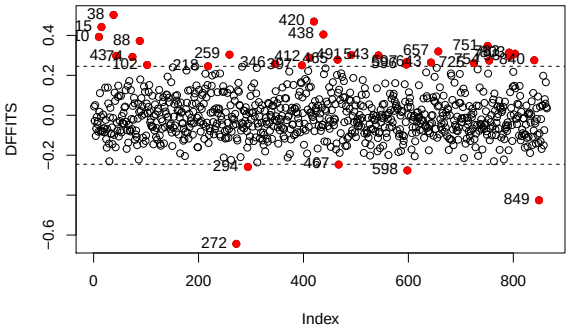
Influence Measures

Influential observations are observations that strongly impact our regression. There are mainly 3 ways to quantify Influence: DFFITS, Cook's distance and DFBetas. We will discuss them indivisually.

DFFITS

DFFITS measures the scaled difference between the fitted value for y_i, and what we would have gotten if we hadn't observed y_i. The rule of thumb when considering observations with

large impact on fitted values is $|DFFITS_i > 2\sqrt{\frac{p+1}{n}}$. We plotted the DFFITS values against the index of the observation, and highlighted the influential points in red.



As the plot illustrated, 15, 38, 272, 420, 438, 849 have very large impact on the fitted values

Cook's Distance

Cook's Distance is the scaled measure of average squared distance between fitted values with and without y_i. We will compare it to $F_{p+1,N-p-1}$ at 0.1 level and no observations seems to have a huge impact.

DFBETAS

DFBETAS measures the i-th observation's impact on coefficient estimates. For each variable, we plotted the output values againt covariate values, and highlighted the observations with large DFBetas. After observing the plots, we concluded that

-observation 272, 849 have a huge impact on the coefficient of POP_PCB11,

-observation 15 have a huge impact on the coefficient of POP_furan3,

-observation 597 have a huge impact on the coefficient of monocyte_pct,

-observation 793 have a huge impact on the coefficient of BMI,

-observation 420 have a huge impact on the coefficient of ageyrs,

-observations 38, 465 have a huge impact on the coefficient of yrssmoke,

-observations 38, 438 have a huge impact on the coefficient of yrssmoke,

To sum up, Observations 15, 38, 272, 420, 438, 849 have very large impact on the fitted values, and Observation 15, 38, 272, 438, 465, 597, 793, 849 have very impact on coefficient estimates

Let's have a look at the model without these observations:

```
##                                        AIC        BIC Adj R Sqr        MSE
## reduced model                  -145.4452  -83.54478 0.2224694 0.04868799
## reduced model without outliers -221.6060 -159.87217 0.2429021 0.04442353
```

We can see that after removing the obvious outliers, AIC and BIC decreases, Adjusted R-square increases, and MSE decreased.

This means that the refitted model not only fits the data better, but produces more accurate predictions.

Possible Interaction

First, let's build a model which involves the interaction between POP_PCB11 and POP_furan3.

To analyse whether this interaction makes our fitted model better, we may want to again analyse the AIC, BIC, Adjusted R squared and MSE of the new model.

```
##                                          AIC       BIC Adj R Sqr        MSE
## reduced model without interaction -221.6060 -159.8722 0.2429021 0.04442353
## reduced model with interaction    -224.1484 -157.6657 0.2460265 0.04424021
```

Compared with our final reduced model (outliers removed), we can clearly observe a decrease in AIC, BIC and MSE, an increase in adjusted R squared. This indicates that involving the interaction between POP_PCB11 and POP_furan3 actually makes our model better fitted the given data. Therefore, there is evidence showing that there is some interaction between the organic pollutantsPOP_PCB11 and POP_furan3.

Discussion

To sum up, we derive a well fitted model of our data. This is indicated by AIC, BIC, Adjusted R-square and mean squared error.

In terms of the effect of specific covariates ion leukocyte telomere length:

For the organic pollutants variables, the only two variable that have significant impact on mean leukocyte telomere length are POP_PCB11 and POP_furan3, when the concentration of the organic pollutants increases, the mean leukocyte telomere length also tends to increase.

For the age variable, older the patients, shorter the mean leukocyte telomere length. Also, longer the years the patients smoke for, shorter the mean leukocyte telomere length. Higher the BMI, shorter the mean leukocyte telomere length. Higher the log of cotinine in ng/mL, shorter the mean leukocyte telomere length

For the categorical data, college students tend to have longer telomere. Females tend to have longer telomere. Mexican American generally have shorter telomere and non-hispanic black usually have longer telomere.

When we were investigating the four assumptions, the normality assumption and the heteroskedasticity assumption are slightly broken. To solve the problem with the normality assumption, we could use the Central Limit Theorem which indicates normality when we are dealing with a large number of samples. However, the heteroskedasticity assumption is still broken. This indicates that we may have different residuals between different fitted values. This could be solved by using weighted least square.

The limitation would be the dataset is not comprehensive enough. There might be other covariate that are not recorded in the data.

Appendix

#Load the data

```
library(readr)

suppressWarnings(pollutants_Raw <- read_csv("pollutants.csv"))
pollutants_Raw <- pollutants_Raw[-1]   # remove the index column


# VIF
pollutants <- pollutants_Raw[-1]   # remove the response
VIF_boundry <- 8

while(TRUE) {
  VIFs <- rep(NA, ncol(pollutants))  # vector to store the VIF of each variable
  for(i in 1:ncol(pollutants)) {  # find the VIF of every covariates
    rm <- lm(as.formula(paste(names(pollutants)[i], "~", ".")), data = pollutants)  # fi
    VIFs[i] <- 1 / (1 - summary(rm)$r.squared)  # find the VIF of Xi, and store it in V
  }
  # message("VIFs: ", VIFs)
  if(max(VIFs) > VIF_boundry) {  # if the largest VIF is larger than 10
    j = which.max(VIFs)  # find the index of the covariate with the largest VIF
    pollutants <- pollutants[-j]  # remove that covariate from the dataset
  } else {
    break  # stop the loop if the VIFs for all covariates are <= 10
  }
} # end of while loop

pollutants$length <- pollutants_Raw$length  # add the output column to the dataset


# change categorical formats
maleLevels <- c("$female", "$male")
pollutants$male <- factor(pollutants$male, levels=c(0,1), labels = maleLevels)

eduLevels <- c("$high_school_students", "$high_school_grads", "$college_students", "$col
pollutants$edu_cat <- factor(pollutants$edu_cat, levels=c(1,2,3,4), labels = eduLevels)

raceLevels <- c("$Other", "$Mexican_American", "$Non_Hispanic_Black", "$Non_Hispanic_Whi
pollutants$race_cat <- factor(pollutants$race_cat, levels=c(1,2,3,4), labels = raceLevel

smokenowLevels <- c("$Not", "$Smoke")
pollutants$smokenow <- factor(pollutants$smokenow, levels=c(0,1), labels = smokenowLevel
```

```r
M_full <- lm(length ~., data = pollutants)


# Use LASSO to do automatic selection
library(glmnet)
M_lasso <- glmnet(x=model.matrix(M_full)[,-1], y=pollutants$length, alpha = 1)


# fit with cross validation
# default is 10 fold
cvfit_lasso <- cv.glmnet(x=model.matrix(M_full)[,-1], y=pollutants$length, alpha = 1)

## estimated betas for minimum lambda
coef_lasso <- coef(cvfit_lasso, s = "lambda.min")
covs_Names <- rownames(coef_lasso)[which(coef_lasso != 0)]  # get the names of the cova
removed_Names <- rownames(coef_lasso)[which(coef_lasso == 0)]  # get the names of the c

## Merge the catgories with 0 coefficients with their referent group
library(stringi)
pollutants2 <- pollutants
# Note: The idea behind the below code is to project the new factor levels to the old
if(is.element("male", substring(covs_Names, 1, stri_length("male")))) {  # we know that
  maleLevels2 <- rep("$Other", length(maleLevels))  # maleLevels2: all factors are "$Ot
  male_remained_Index <- which(substring(covs_Names, 1, stri_length("male")) == "male")
  category_index <- match(substring(covs_Names[male_remained_Index], stri_length("male")
  maleLevels2[category_index] <- maleLevels[category_index]  # set the factor values of
  pollutants2$male2 <- factor(pollutants2$male, levels=maleLevels, labels=maleLevels2)
  covs_Names <- covs_Names[-male_remained_Index]  # remove the individual categories fr
  covs_Names[length(covs_Names)+1] <- "male2"  # add the new variable to the remaining
}
if(is.element("edu_cat", substring(covs_Names, 1, stri_length("edu_cat")))) {  # we knou
  eduLevels2 <- rep("$Other", length(eduLevels))  # eduLevels2: all factors are "$Other
  edu_remained_Index <- which(substring(covs_Names, 1, stri_length("edu_cat")) == "edu_c
  category_index <- match(substring(covs_Names[edu_remained_Index], stri_length("edu_cat
  eduLevels2[category_index] <- eduLevels[category_index]  # set the factor values of t
  pollutants2$edu_cat2 <- factor(pollutants2$edu_cat, levels=eduLevels, labels=eduLevels
  covs_Names <- covs_Names[-edu_remained_Index]  # remove the individual categories fro
  covs_Names[length(covs_Names)+1] <- "edu_cat2"  # add the new variable to the remaini
}
if(is.element("race_cat", substring(covs_Names, 1, stri_length("race_cat")))) {  # we kn
  raceLevels2 <- rep("$Other", length(raceLevels))  # raceLevels22: all factors are "$O
  race_remained_Index <- which(substring(covs_Names, 1, stri_length("race_cat")) == "rac
  category_index <- match(substring(covs_Names[race_remained_Index], stri_length("race_c
  raceLevels2[category_index] <- raceLevels[category_index]  # set the factor values of
```

```
  pollutants2$race_cat2 <- factor(pollutants2$race_cat, levels=raceLevels, labels=raceLe
  covs_Names <- covs_Names[-race_remained_Index]  # remove the individual categories fr
  covs_Names[length(covs_Names)+1] <- "race_cat2"  # add the new variable to the remain
}
if(is.element("male", substring(covs_Names, 1, stri_length("smokenow")))) {  # we know t
  smokeLevels2 <- rep("$Other", length(smokeLevels))  # smokeLevels2: all factors are "
  smoke_remained_Index <- which(substring(covs_Names, 1, stri_length("smokenow")) == "sm
  category_index <- match(substring(covs_Names[smoke_remained_Index], stri_length("smoke
  smokeLevels2[category_index] <- smokeLevels[category_index]  # set the factor values
  pollutants2$smokenow2 <- factor(pollutants2$smokenow, levels=smokeLevels, labels=smoke
  covs_Names <- covs_Names[-smoke_remained_Index]  # remove the individual categories f
  covs_Names[length(covs_Names)+1] <- "smokenow2"  # add the new variable to the remain
}

covs_Names <- covs_Names[-1]  # remove the intercept
pollutants_reduced <- subset(pollutants2, select = covs_Names)
pollutants_reduced$length <- pollutants2$length  # add the outcome

M_reduced <- lm(length ~ ., data=pollutants_reduced)


# Try fitting the model using Backward Selection, Forward Selection and Stepwise Selec
# The model with only intercept
M0 <- lm(length~., data=pollutants)
n <- 864

# Forward Selection
Mfwd <- step(object = M0, # base model
             scope = list(lower = M0, upper = M_full), # smallest and largest model
             direction = "forward",
             trace = 0,
             k = 2)

# Backward Selection
Mback <- step(object = M_full, # base model
              scope = list(lower = M0, upper = M_full),
              direction = "backward",
              trace = 0,
              k = 2)

# Stepwise Selection
Mstepwise <- step(object = M_full, # base model
                  scope = list(lower = M0, upper = M_full),
                  direction = "both",
```

```
                  trace = 0,
                  k = 2)

aic_fwd <- AIC(Mfwd)
aic_back <- AIC(Mback)
aic_stepwise <- AIC(Mstepwise)
aic_lasso <- AIC(M_reduced)
aic_ori <- AIC(M_full)

bic_fwd <- BIC(Mfwd)
bic_back <- BIC(Mback)
bic_stepwise <- BIC(Mstepwise)
bic_lasso <- BIC(M_reduced)
bic_ori <- BIC(M_full)

adj_fwd <- summary(Mfwd)$adj.r.square
adj_back <- summary(Mback)$adj.r.square
adj_stepwise <- summary(Mstepwise)$adj.r.square
adj_lasso <- summary(M_reduced)$adj.r.square
adj_ori <- summary(M_full)$adj.r.square

mse_fwd <- sigma(Mfwd)^2
mse_back <- sigma(Mback)^2
mse_stepwise <- sigma(Mstepwise)^2
mse_lasso <- sigma(M_reduced)^2
mse_ori <- sigma(M_full)^2

x <- c(aic_ori, aic_fwd, aic_back, aic_stepwise, aic_lasso, bic_ori, bic_fwd, bic_back,b
       adj_ori, adj_fwd, adj_back, adj_stepwise, adj_lasso, mse_ori, mse_fwd, mse_back,
M <- matrix(x, nrow=5, ncol=4)
M <- data.frame(M)
rownames(M) =c("Original Data", "Forward Selection", "Backward Selection", "Stepwise Sel
colnames(M) =c("AIC", "BIC", "Adj R Sqr", "MSE")
M


par(mfrow=c(1,2))
## Analyze assumptions: Normality, Heteroskedasticity, Linearity, Independence
res <- resid(M_reduced)
student_res <- res/(sigma(M_reduced)*sqrt(1-hatvalues(M_reduced)))

## plot distribution of studentized residuals
hist(student_res,
     breaks=12,
```

```
     probability=TRUE,
     xlim=c(-6,6),
     xlab="Studentized Residuals",
     main="Distribution of Residuals")
grid <- seq(-3.5,3.5,by=0.05)
lines(x=grid,y=dnorm(grid),col="blue")


# qq-plot of studentized residuals
qqnorm(student_res)
abline(0,1)


# Heteroskedasticity
plot(student_res~fitted(M_reduced),
     xlab="Fitted Values",
     ylab="Studentized Residuals",
     main="Residuals vs Fitted")
abline(0,0, col="blue", lwd=2)


# Linearity
library(car)
avPlots(M_reduced)


# Influence Measures

#DFFITS
dffits_m <- dffits(M_reduced)


## plot DFFITS
plot(dffits_m,ylab="DFFITS")
abline(h = 2*sqrt((p+1)/n),lty=2)  ## add thresholds
abline(h = -2*sqrt((p+1)/n),lty=2)

## highlight influential points
dff_ind <- which(abs(dffits_m)>2*sqrt((p+1)/n))
points(dffits_m[dff_ind]~dff_ind,col="red",pch=19) ## add red points
text(y=dffits_m[dff_ind],x=dff_ind, labels=dff_ind, pos=2) ## label high influence poin

#Cook's Distance
D <- cooks.distance(M_reduced)
```

```r
# influential points
inf_ind <- which(pf(D,p+1,n-p-1,lower.tail=TRUE) > 0.1)

# plot cook's Distance
plot(D,ylab="Cook's Distance")
points(D[inf_ind]~inf_ind,col="red",pch=19) ## add red points
if(length(inf_ind) > 0) text(y=D[inf_ind],x=inf_ind, labels=inf_ind, pos=4) ## label hig


#DFBETAS
DFBETAS <- dfbetas(M_reduced)

## betas
for(ii in 2:8) {
  df_ind <- which(abs(DFBETAS[,ii]) > 2/sqrt(n))
  show_points <- order( -abs(DFBETAS[, ii]))[1:min(length(df_ind),5)]  # 5 or length(df_
  plot(pollutants_reduced$length ~ as.numeric(unlist(c(pollutants_reduced[colnames(DFBET
  points(x=as.numeric(unlist(c(pollutants_reduced[colnames(DFBETAS)[ii]])))[show_points]
  text(x=as.numeric(unlist(c(pollutants_reduced[colnames(DFBETAS)[ii]])))[show_points],
}


### Let's have a look at the model without these observations:
omit_ind <- c(15, 38, 272, 420, 438, 440, 465, 597, 759, 793, 849)

M_omit <- update(M_reduced, data = pollutants_reduced[-omit_ind,])

aic_reduced <- AIC(M_reduced)
aic_omit <- AIC(M_omit)

bic_reduced <- BIC(M_reduced)
bic_omit <- BIC(M_omit)

adj_reduced <- summary(M_reduced)$adj.r.squared
adj_omit <- summary(M_omit)$adj.r.squared

mse_reduced <- sigma(M_reduced)^2
mse_omit <- sigma(M_omit)^2

x <- c(aic_reduced, aic_omit, bic_reduced, bic_omit, adj_reduced, adj_omit, mse_reduced,
M2 <- matrix(x, nrow = 2, ncol = 4)
M2 <- data.frame(M2)
rownames(M2) =c("reduced model", "reduced model without outliers")
colnames(M2) =c("AIC", "BIC", "Adj R Sqr", "MSE")
```

```r
M2


#interaction
M_interaction <- lm(length~. + POP_furan3*POP_PCB11, data=pollutants_reduced[-omit_ind,]

aic_inter <- AIC(M_interaction)
bic_inter <- BIC(M_interaction)
adj_inter <- summary(M_interaction)$adj.r.squared
mse_inter <- sigma(M_interaction)^2
y <- c(aic_omit, aic_inter, bic_omit, bic_inter, adj_omit, adj_inter, mse_omit, mse_inte
matrix_inter <- matrix(y, nrow=2, ncol=4)
matrix_inter<- data.frame(matrix_inter)
rownames(matrix_inter) =c("reduced model without interaction",
                          "reduced model with interaction")
colnames(matrix_inter) =c("AIC", "BIC", "Adj R Sqr", "MSE")
matrix_inter
```