STAT 331 Final Project Report

Daniel Mao
Yufei Wang
Caroline Xu
Yuchen Zhang

Contents

Summary

This STAT 331 Final project mainly analyzes how the pollutants, cells, smoking conditions, and general peronal info affect the telomere length. In our project, we set up a number of questions in the objectives which could help us analyze the relationship between pollutant and the person health. Our report is based on an analysis of the `pollutant.csv` dataset. The covariates were divided into four groups and we analyze how each group was related to the outcome, telomere length. We removed explanatory variables with large VIF one by one, until there are no more with "high" multicollinearity. We evaluated the prediction accuracies of the models with k-fold cross validation and then selected model which have smallest error. Then we used different criteria to evaluate the goodness of fit of our final model. After the model selection process, we also checked if the four assumptions of a linear model, namely linearity, independence, normality, and homoskedasticity, were all met. In conclusion, according to the results from various models, we found out that all the correlation coefficient between length and different pollutants are negative, and the telomere length does not have strong relationship with any type of cells.

Objective

For the issues of relationship between pollutants with personl health, we have some questions to ask.

- What is the mean of the differences of telomere length between the person currently smokes with the person currently not smokes?
- Does female or male have higher percentage of elder cellular aging?
- Which kind of pulltant has the most significant effect on the mean leukocyte telomere length, PCBs, dioxins or furans?
- Is there a positive linear relationship or a negative linear relationship between BMI and the mean telomere length?

The main purpose of our analysis is to find out how the persistant organic pollutants could affect the telomere length. Our dataset is based on a population of $n = 864$ adults in a study invertigating. To be specific, we are going to explore the relationships between the personal basic conditions, the smoking related conditions, the percent of cells index and four different pollutants with cellular aging, which might be related to certain cancers. The telomere length is exactly a marker of cellular aging.

This report summarizes the statistical modeling and analysis result associated with the relationship between exposure to persistant organic pollutants and telomere length. We divide the covariates into four groups and analyze how each group is related to the outcome, telomere length. More specifically, the goal of the analysis is to find out

- among the 18 pollutants recorded, which 3 are most closely related to telomere length,
- how are the concentrations of the 6 kinds of cells related to telomere length,
- whether telomere length is related to smoking, and
- which group of people, characterized by sex, age, BMI, race, and/or education, is most likely to have long telomere length.

After the statement of the goal of report, the next procedure is to state the method that will be used in the analysis step. First for the given model, we use VIF to test and eliminate the multicollinearity. We use VIF > 10 as the standard and remove all other explanatory variables, we take this step one by one until there is no more multicollinearity. Then for the model building part, we do automatic selection. Two different selection method was used in our analysis procedure. They are Forward Selection and Backward Elimination respectively. We use AIC and BIC in each method to get our model, and make comparison between each other. After all of these steps, we almost already have our most fitted model, but we need to check Model Predictive Accuracy. And that is exactly the reason of why we are seeking for the Mean Squared Prediction Error(MSPE) as small as possible. Ultimately, the interests questions could have their answers. In the summary of our results, we could find the figures of corresponding covariates, and the plots could give us the answer of relationships.

Exploratory Data Analysis

The recorded explanatory covariates can be divided into 4 groups:

- the first group contains all 18 kinds of pollutants,
- the second group contains all 6 kinds of cells,
- the third group contains all covariates related to “smoking”,
- and the last group contains all others remaining.

Report Summary Statistics:

We found that the mean of the telomere length variable is 1.0543127 and the range of telomere length variable is 1.8246648. The mean and range of the concentrations of the pollutants are summarized below:

```
##        POP_PCB1  POP_PCB2  POP_PCB3  POP_PCB4  POP_PCB5  POP_PCB6  POP_PCB7
## mean   38082.18  15636.81  10157.75  38455.79  52650.23  16820.02  12681.94
## range 570000.00 163000.00 121000.00 484900.00 705900.00 317000.00 142900.00
##        POP_PCB8  POP_PCB9 POP_PCB10 POP_PCB11 POP_dioxin1 POP_dioxin2
## mean   10529.75  12220.25   24.4917   38.1537     57.6535     47.8125
## range 185900.00 142900.00  170.3000  843.7000    758.1000    279.6000
##       POP_dioxin3 POP_furan1 POP_furan2 POP_furan3 POP_furan4
## mean     494.4167     6.3714     5.3896     6.6688    11.5448
## range   8153.2000    43.4000    32.7000    37.6000   233.1000
```

The PCB with largest mean is POP_PCB5. The PCB with largest range is POP_PCB5. The dioxin with largest mean is POP_dioxin3. The dioxin with largest range is POP_dioxin3. The furan with largest mean is POP_furan4. The furan with largest range is POP_furan4.

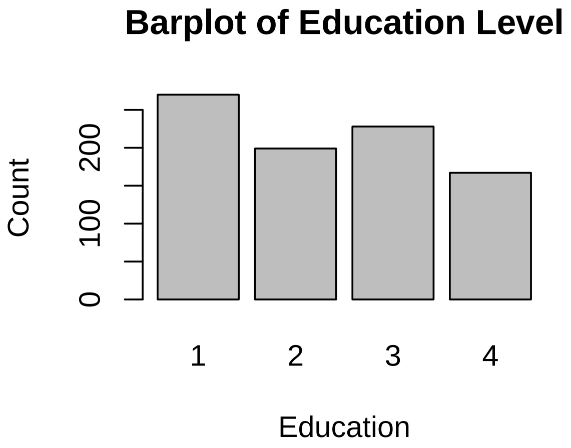
The mean of the cells counts are summarized below:

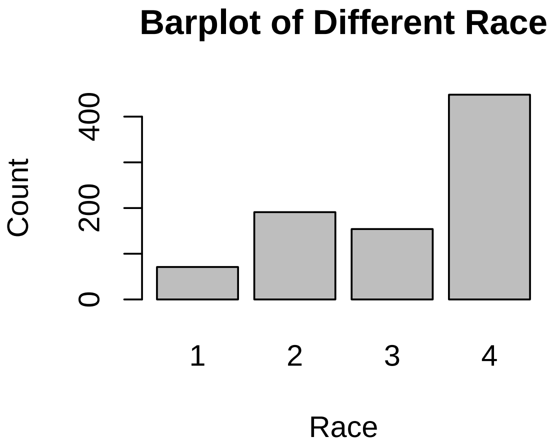
```
## whitecell  lymphocy    monocy eosinophi   basophi neutrophi
##    7.1910    2.0833    0.5552    4.3033    0.2040    0.0479
```

The type of cell with the largest mean is whitecell. The type of cell with the smallest mean is neutrophi.

For the smoke subdataframe, the lowest year smoked cigarettes is 0 year and we find that there are 472 people never smoked. Dividing by the sample size, we found that there are approximately 55% people who do not smoke. The longest years smoked is 69. There was only one person who smoked for this many years. From the variable “smokenow”, we found that there are 664 poeple who do not currently smoke and there are 200 people who currently smoke. This means that there are currently more than three times as many non-smokers as there are smokers.

```
##     BMI mean    BMI range    edu_cat 1    edu_cat 2    edu_cat 3    edu_cat 4
##      28.0942      46.8300     270.0000     199.0000     228.0000     167.0000
##   race_cat 1   race_cat 2   race_cat 3   race_cat 4       female         male
##      71.0000     191.0000     154.0000     448.0000     490.0000     374.0000
##  ageyrs mean ageyrs range
##      48.3553      65.0000
```



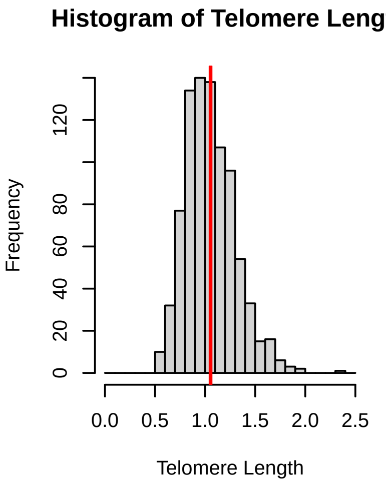


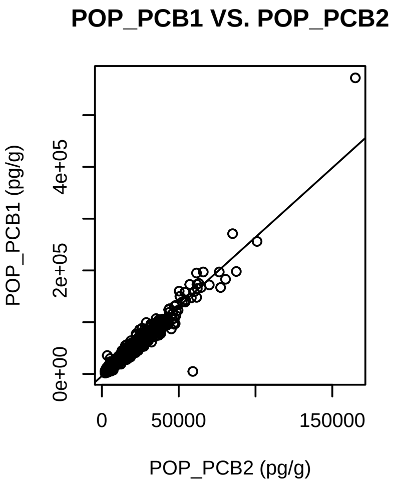
```
##    max cor (df)    min cor (df) max cor (cells) min cor (cells)
##          0.9710         -0.4454          0.9295         -0.0495
```

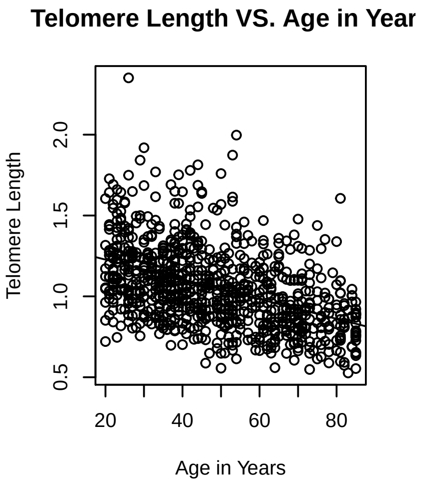
We can see the maximum number of correlation coefficient is 1 which is the relarionship between POP_PCB1 and POP_PCB2. This means POP_PCB1 and POP_PCB2 have a strong positive relationship. Also, the minimum number of correlation coefficient is -0.9346368, which is the strong negative realtionship between eosinophils_pct and lymphocyte_pct. We also found that length has the smallest correlation coefficient which close to -1 with ageyrs which is -0.4454. The strongest relationship in the data fram cells is between eosinophils_pct and whitecell_count which is 0.92951238. This means that eosinophils_pct and whitecell_count has a strong positibe relationship. The correlation coefficient closest to 1 or -1 between length and other variables in cells dataframe is -0.04949 which is between length and monocyte_pct. We found that the telomere length does not have strong relationship with any type of cells.

We can also get that the strongest relationship between length and pollutant is length and POP_PCB2. However, the correlation coefficient closest to 1 or -1 is -0.2457 which is closer than 0. This means that the telomere length and POP_PCB2 does not have strong relationship and the telomere length does not have strong relationship with any pollutant types.

From the dataframe smoke and others, we can find that all of them do not have a correlation coefficient which is close to 1 or -1. This means neither of them have a strong relarionship between each other.







From the first histogram of the telomere length, we found that it is not really symmetric because it has a longer tail on the right and there is an outlier which is larger than 2.0. In the second plot, we can find they have a really strong positive linear relationship. In the third plot, we can find they have a negative linear relationship between the mean telomere length and age in year.

Interesting Findings:

- The PCB/dioxin/furan with largest mean and range are the same one.
- We found that the telomere length does not have strong relationship with any of the other variables.
- The relationship between different types of organic pollutants was very strong. We found there are strong relationships between different type of organic pollutant PCBs. Also, all the correlation coefficient between length and different pollutants are negetive.
- The telomere length and age in year has a negative linear relationship. This indicates as people grow up, their telomere length tend to be shorter However, they do not have a really strong linear relationship.

How these inform the rest of our analysis:

We are going to use the variable which has a correlation coefficient that close to 1 or -1 to get a model. Also, the variable which has an absolute correlation coefficient close to 1 will have a lower value of VIF when we do the model selection. This is because having the absolute correlation coefficient closer to 1 means that the variables have stronger linear relationship. Thus having lower VIF value also means the variables have stronger relationship.

Methods

Our Models

Model 1 is the model that takes all explanatory variables into account.

Model 2 is the model that has only the variates with "large" correlation with the outcome. The variates selected were: POP_PCB1, POP_PCB2, POP_PCB3, POP_PCB4, POP_PCB5, POP_PCB8, POP_PCB9, eosinophils_pct, and lymphocyte_pct. A brief summary of model 2 is printed below:

```
##
## Call:
## lm(formula = length ~ POP_PCB1 + POP_PCB2 + POP_PCB3 + POP_PCB4 +
##     POP_PCB5 + POP_PCB8 + POP_PCB9 + eosinophils_pct + lymphocyte_pct,
##     data = df)
##
## Coefficients:
##     (Intercept)         POP_PCB1         POP_PCB2         POP_PCB3
##       1.105e+00       -1.433e-06       -4.579e-06        1.134e-06
##        POP_PCB4         POP_PCB5         POP_PCB8         POP_PCB9
##      -1.292e-06        1.194e-06        1.357e-07        3.257e-06
## eosinophils_pct   lymphocyte_pct
##       4.271e-03       -4.167e-03
```

Models 3 ~ 6 are models built by automatic selection from a reduced model. To get the reduced model, we removed the explanatory variables with large VIF's one by one, until there are no more with "high" multicollinearity. As a rule of thumb, we considered a VIF > 10 to indicate high multicollinearity.

The explanatory variables removed were:

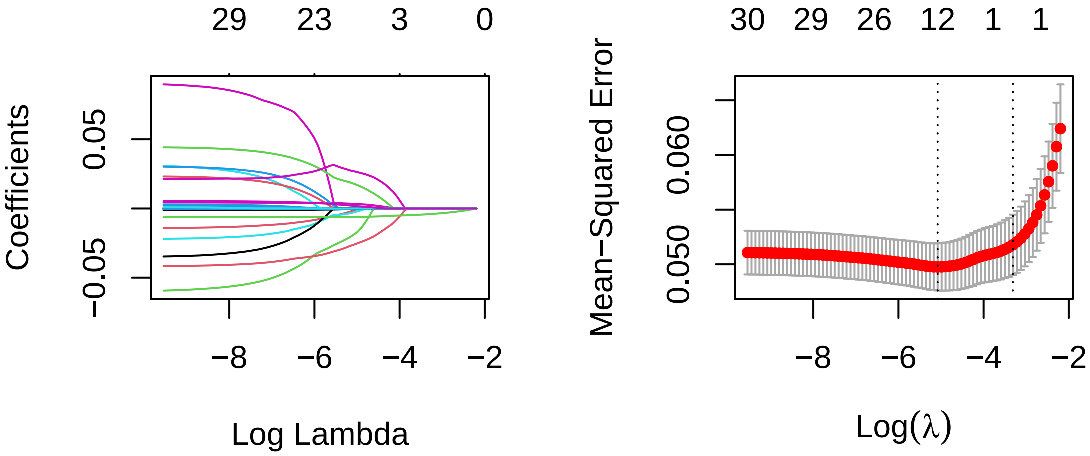
```
## [1] "POP_PCB1"        "POP_PCB2"        "POP_PCB4"        "POP_PCB5"
## [5] "whitecell_count"
```

After removing these variables, we got a reduced model. Then we built models 3 ~ 6 out of the reduced model. Model 3 was built by forward selection, based on AIC. Model 4 was built by forward selection, based on BIC. Model 5 was built by backward selection, based on AIC. Model 6 was built by backward selection, based on BIC.

The numbers of the predictor variables of each model are printed below:

```
## [1] "Model 3 has 31 predictors"
## [1] "Model 4 has 31 predictors"
## [1] "Model 5 has 5 predictors"
## [1] "Model 6 has 2 predictors"
```

Model 7 was built by LASSO.
Let's first have a look at the paths and the MSPEs by lambda plot:



The number of predictor variables in model 7 is: 32.

Model Selection

To select the best model from models 1 ~ 7, we computed the k-fold cross validation of each. The prediction errors are printed below:

```
##     M1     M2     M3     M4     M5     M6     M7
## 0.0519 0.0605 0.0512 0.0512 0.0493 0.0496 0.0627
```

From the MSPE's we saw that model 5 had the least MSPE and hence was considered to be the "best" model.

How did we select a model?

We first calculated the VIF's of each covariate and eliminated the covariates with VIF larger than 10. Then we used forward selection and backward selection, based on AIC and BIC, to get four fitted models. Then we evaluated the prediction accuracies of the four models with k-fold cross validation and seleted the one with smallest error.

Parsimony and Interpretability

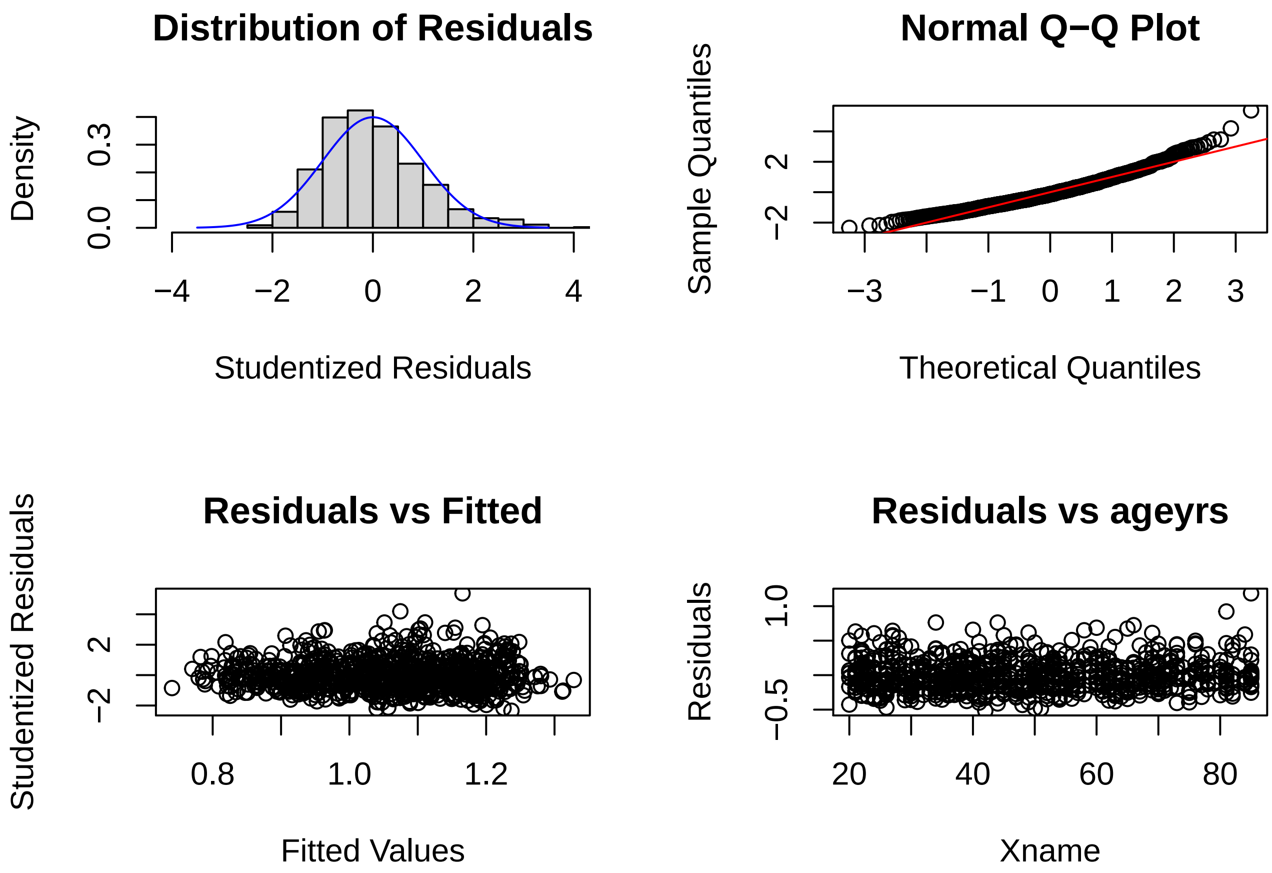
From our calculations, we saw that the models built by forward selection were more complex the models built by backward selection were relatively simple.

Goodness of Fit

```
##        R2 Adj R2    MSE       AIC       BIC
## M1 0.2450 0.2121 0.0472 -109.7454   71.1943
## M2 0.0689 0.0591 0.0582   17.3520   69.7293
## M3 0.2414 0.2131 0.0474 -115.6205   41.5114
## M4 0.2414 0.2131 0.0474 -115.6205   41.5114
## M5 0.2240 0.2195 0.0485 -148.1111 -114.7801
## M6 0.2141 0.2123 0.0492 -143.0867 -124.0404
## M7     NA     NA     NA        NA        NA
```

From the table above we observed that model 5 had relatively high R^2, adjusted R^2, and MSE and relatively low AIC and BIC values. This meant that model 5 fit the data well.

Are the necessary assumptions met?



- The residuals distribution plots and the normal Q-Q plots showed that the normality assumption was met.
- From the residuals vs. fitted values plots we saw that the variance of the studentized residuals were approximately the same for all levels of fitted values. So the homoskedasticity assumption was met.
- From the residuals vs. X plots we saw that the residuals were fairly “random”. So the linearity assumption was met.

Results

In this section of the report, we are going to state some outcomes and findings of our data analysis. After we checked the goodness of fit in the last part, we concluded that model 5 fitted the given data the best. So this was the ultimate model of our report.

There ware 5 explanatory variables in model 5, namely POP_furan3: concentration of a furans, male1: the gender (0=female, 1=male), ageyrs: age in years, ln_lbxcot: log of cotinine in ng/mL, and lymphocyte_pct: percentage of lymphocytes (out of white blood cells).

The p-values of the variables in the final model could give us the informations about significance. The p-value of POP_furan3 was 7.61e-05, which was smaller than 0.05. So POP_furan3 was statistically significant. The p-value of lymphocyte_pct was 0.13071. This is greater than 0.05 So lymphocyte_pct was not statistically significant. The p-value of male1 was 0.00598, which was less than 0.05. So male1 was significant. The p-value of ageyrs was "< 2e-16", which was less than 0.05. So ageyrs was statistically significant. Finally, the variable of ln_lbxcot has the p-value of 0.03974 < 0.05, which is also a significant variable.

Then for the analysis of the final model, there should be more explanations about the covariates that were left in the final model5. From the summary table of model 5, we could find that there are positive linear relationships between the variables POP_furan3 and ln_lbxcot with the telomere length. More specifically, if there is a unit change in POP_furan3, the change of the mean outcome will be 0.0062438; if ln_lbxcot has a unit change (ng/mL), the change of the mean outcome will be 0.0042712. Besides that, the other three variables male1, ageyrs and lymphocyte_pct are all have negative linear relationship with the telomere length. Similarly, if male1 has a unit change, the change of the mean outcome will be -0.0428367; if ageyrs has a unit change (years), the change of the mean outcome will be -0.0070265; if lymphocyte_pct has a unit change (percentage), the change of the mean outcome will be -0.0001737.

Attach the summary of the final model: model 5 here, and present it in the form of coefficients table:

```
##
## Call:
## lm(formula = length ~ POP_furan3 + lymphocyte_pct + male + ageyrs +
##     ln_lbxcot, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51625 -0.15193 -0.02809  0.12208  1.18579
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.4113679  0.0348798  40.464  < 2e-16 ***
## POP_furan3      0.0062438  0.0015705   3.976 7.61e-05 ***
## lymphocyte_pct -0.0173733  0.0114847  -1.513  0.13071
## male1          -0.0428367  0.0155442  -2.756  0.00598 **
## ageyrs         -0.0070265  0.0005058 -13.892  < 2e-16 ***
## ln_lbxcot       0.0042712  0.0020738   2.060  0.03974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2211 on 858 degrees of freedom
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.2195
## F-statistic: 49.55 on 5 and 858 DF,  p-value: < 2.2e-16
```

Ultimately, there is an examine part of the data. There is a X-outliers that exists in the Histogram of Telomere Length, that is about 2.4 of index. That X-outliers could have some effections on the results of the final model. The X-outliers could make the mean of the telomere length smaller than the true value. Compare to the case that if there is no X-outliers, the X-outliers could make the estimates that have positive linear relationship smaller and could make the estimates that have negative linear relationship larger.

Discussion

We used VIF to get a reduced model. Then, using forward with AIC and BIC, backward with AIC and BIC, and LASSO to get model 3 to model 7. After that we used prediction accuracy to find the best model and evaluated the parsimony and goodness of fit of the final model. From the rsiduals vs. fitted values plot we noticed that it showed approximately but not perfect equal variance. We also verified other model assumptions like linearity and normality. They all turned out to be approximately satisfied. We also found that there existed a strong relationship between pollutant PCBs. POP_PCB2 had the most significant effect on the mean leukocyte telomere length. There was a weak negative relationship between BMI and the mean telomere length.

Ultimatley, we have some limitations of the analysis to share: 1. We did not assess independence of the data and we think that the outcomes might be dependent. For example, people living near to each other might have similar concentrations of the 18 pollutants. 2. The dataset might be too small. 3. We did not handle heteroskedasticity.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = FALSE,
                      results = "hold",
                      message = FALSE,
                      warning = FALSE,
                      fig.width=3,
                      fig.height=3)
n <- 864
df <- read.csv(file="pollutants.csv")
# the X variables is the same as the index so we can delete it
df = df[,-1]
# since the five kinds of cells other than white blood cells
# are all recorded in percentage, out of white blood cells,
# we first multiply all five by the count of white blood cells
# to get the count of each.
df[21] = df[20] * df[21] / 100
df[22] = df[20] * df[22] / 100
df[23] = df[20] * df[23] / 100
df[24] = df[20] * df[24] / 100
df[25] = df[20] * df[25] / 100
# names(df)[which(names(df)=="lymphocyte_pct")]  = "lymphocyte_count"
# names(df)[which(names(df)=="monocyte_pct")]    = "monocyte_count"
# names(df)[which(names(df)=="eosinophils_pct")] = "eosinophils_count"
# names(df)[which(names(df)=="basophils_pct")]   = "basophils_count"
# names(df)[which(names(df)=="neutrophils_pct")] = "neutrophils_count"
# categorical variates should be treated as factor.
df[,"edu_cat"]  = factor(df[,"edu_cat"])#,levels=c("< High School","High School","Colle
df[,"race_cat"] = factor(df[,"race_cat"])
df[,"male"]     = factor(df[,"male"])
df[,"smokenow"] = factor(df[,"smokenow"])
# sub-dataframes.
# sub-dataframes include length and all the pollutants elements
pollutants <- df[,c(1, 2:19)]
# sub-dataframes include length and all the cell elements
cells      <- df[,c(1,20:25)]
# sub-dataframes include length, smoke years, smoke now ,and cotinine
smoke      <- df[,c(1,31:33)]
# ub-dataframes include length, age, gender, education level, race level and BMI
others     <- df[,c(1,26:30)]
if (!(length(pollutants) == 1 + 18)) message("ERROR")
if (!(length(cells)      == 1 + 6))  message("ERROR")
if (!(length(smoke)      == 1 + 3))  message("ERROR")
if (!(length(others)     == 1 + 5))  message("ERROR")
```

```r
y.smr = summary(df$length)
# get the mean of all the variables in pollutants
p.mean.vec = round(apply(pollutants, 2, FUN = mean)[-1], 4)
# get the range of all the variables in pollutants
p.range.vec = round(apply(pollutants, 2, FUN = function(col) range(col)[2]-range(col)[1
p.table = rbind(p.mean.vec, p.range.vec)
rownames(p.table) = c("mean", "range")
p.table
# get the mean of all the variables in cells
c.mean.vec = round(apply(cells, 2, FUN = mean)[-1], 4)
# get the range of all the variables in cells
names(c.mean.vec) = sapply(names(df)[20:25], FUN = function(str) substr(str, 1, nchar(st
c.mean.vec
num_nosmoke <- which(smoke$yrssmoke == min(smoke$yrssmoke))
num_largesmoke <- which(smoke$yrssmoke == max(smoke$yrssmoke))
round(c("BMI mean" = mean(others$BMI), "BMI range" = range(others$BMI)[2] - range(others
  "edu_cat 1" = sum(others$edu_cat==1), "edu_cat 2" = sum(others$edu_cat==2),
  "edu_cat 3" = sum(others$edu_cat==3), "edu_cat 4" = sum(others$edu_cat==4),
  "race_cat 1" = sum(others$race_cat==1), "race_cat 2" = sum(others$race_cat==2),
  "race_cat 3" = sum(others$race_cat==3), "race_cat 4" = sum(others$race_cat==4),
  "female" = sum(others$male==0), "male" = sum(others$male==1), "ageyrs mean" = mean(oth
  "ageyrs range" = range(others$ageyrs)[2] - range(others$ageyrs)[1]), 4)
par(mfrow = c(1, 2))
plot(others$edu_cat, xlab = "Education", ylab = "Count", main = "Barplot of Education Le
plot(others$race_cat, xlab = "Race", ylab = "Count", main = "Barplot of Different Race")
df_cor = df[,-c(27,28,29,32)]
df_cor <- cor(df_cor)
max_cor <- max(df_cor[row(df_cor) != col(df_cor)])
min_cor <- min(df_cor[row(df_cor) != col(df_cor)])
cells_cor <- cor(cells)
cmax_cor <- max(cells_cor[row(cells_cor) != col(cells_cor)])
cmin_cor <- min(cells_cor[row(cells_cor) != col(cells_cor)])
round(c("max cor (df)" = max_cor, "min cor (df)" = min_cor, "max cor (cells)" = cmax_cor
"min cor (cells)" = cmin_cor), 4)
pollutants_cor <- cor(pollutants)
pollutants_cor_len <- pollutants_cor[1,]
min_pollutants <- min(pollutants_cor_len[pollutants_cor_len != 1])
##round(c("length and yrssmoke" = cor(smoke[,c(1:2)])[1,2], "length and ln_lbxcot" = c
##  "yrssmoke and ln_lbxcot" = cor(smoke[,c(2,4)])[1,2]), 4)
##round(c("length and BMI" = cor(others[,c(1:2)])[1,2], "length and ageyrs" = cor(othe
##  "BMI and ageyrs" = cor(others[,c(2,6)])[1,2]), 4)
par(mfrow = c(1, 3))

hist(df$length, breaks = seq(0, 2.5, 0.1),
     xlab = "Telomere Length",
     main = "Histogram of Telomere Length")
abline(v = mean(df$length), col = "red", lwd = 2)

plot(pollutants$POP_PCB1 ~ pollutants$POP_PCB2, main = "POP_PCB1 VS. POP_PCB2",
     xlab = "POP_PCB2 (pg/g)", ylab = "POP_PCB1 (pg/g)")
abline(lm(pollutants$POP_PCB1 ~ pollutants$POP_PCB2))

plot(df$length ~ df$ageyrs, main = "Telomere Length VS. Age in Years",
     xlab = "Age in Years", ylab = "Telomere Length")
abline(lm(df$length ~ df$ageyrs))
library(MASS)
library(regclass)
# This is the model with all the variables
M1 <- lm(length ~ ., data = df)
# These are the variables with correlation coefficients larger than 0.9 or smaller tha
M2 <- lm(length ~ POP_PCB1 + POP_PCB2 + POP_PCB3 + POP_PCB4 + POP_PCB5 +
           POP_PCB8 + POP_PCB9 + eosinophils_pct + lymphocyte_pct, data = df)
M2
# Remove variables with high VIF.
stepwiseVIF <- function(model, threshold = 10)
{
    max_VIF = max(VIF(model))
    max_VIF_var = rownames(VIF(model))[which.max(VIF(model))]
    temp_model = model
    while (max_VIF > threshold)
    {
        update_formula = paste0(".~.-", max_VIF_var)
        temp_model = update(object = temp_model, formula = update_formula)
        max_VIF = max(VIF(temp_model))
        max_VIF_var = rownames(VIF(temp_model))[which.max(VIF(temp_model))]
    }
    final_model = temp_model
    return(final_model)
}
model.full = M1
model.reduced = stepwiseVIF(model.full)
rownames(VIF(model.full))[!(rownames(VIF(model.full)) %in% rownames(VIF(model.reduced)))
# Build model 3.
M3 = stepAIC(model.reduced, k = 2,
             trace = FALSE, direction = "forward")
# Build model 4.
M4 = stepAIC(model.reduced, k = log(nrow(df)),
             trace = FALSE, direction = "forward")
```

```r
# Build model 5.
M5 = stepAIC(model.reduced, k = 2,
             trace = FALSE, direction = "backward")
# Build model 6.
M6 = stepAIC(model.reduced, k = log(nrow(df)),
             trace = FALSE, direction = "backward")
l = list(M3,M4,M5,M6)
for (i in 3:6)
{
    model = l[[i-2]]
    var_names = names(model$coef)[-1]
    print(sprintf("Model %i has %i predictors", i, length(var_names)))
}
library(glmnet)
reduced.X = model.matrix(model.reduced)[,-1]
# Build model 7.
M7 = glmnet(x = reduced.X, y = df$length, alpha = 1) # LASSO
plot(M7, xvar = "lambda", label = TRUE)
M7.cv = cv.glmnet(x = reduced.X, y = df$length, alpha = 1)
plot(M7.cv)
coefs = coef(M7.cv, s = "lambda.min")
models_list = list(M1,M2,M3,M4,M5,M6,M7)
# Compute Prediction Accuracy
ntot <- nrow(df) # total number of observations
# number of cross-validation replications
Kfolds <- 12
df <- df[sample(ntot),] # permute rows
df$index <- rep(1:Kfolds,each=ntot/Kfolds)

# storage space
mspe1 <- rep(NA, Kfolds) # mspe for M1
mspe2 <- rep(NA, Kfolds) # mspe for M2
mspe3 <- rep(NA, Kfolds) # mspe for M3
mspe4 <- rep(NA, Kfolds) # mspe for M4
mspe5 <- rep(NA, Kfolds) # mspe for M5
mspe6 <- rep(NA, Kfolds) # mspe for M6
mspe7 <- rep(NA, Kfolds) # mspe for M7

for(ii in 1:Kfolds)
{
    if(ii%%100 == 0) message("ii = ", ii)
    train.ind <- which(df$index!=ii) # training observations

    # using R functions
    M1.cv <- update(M1, subset = train.ind)
    M2.cv <- update(M2, subset = train.ind)
    M3.cv <- update(M3, subset = train.ind)
    M4.cv <- update(M4, subset = train.ind)
    M5.cv <- update(M5, subset = train.ind)
    M6.cv <- update(M6, subset = train.ind)
    M7.cv <- cv.glmnet(x = reduced.X[train.ind,], y = df$length[train.ind], alpha = 1)
    # cross-validation residuals
    M1.res <- df$length[-train.ind] - predict(M1.cv, newdata = df[-train.ind,])
    M2.res <- df$length[-train.ind] - predict(M2.cv, newdata = df[-train.ind,])
    M3.res <- df$length[-train.ind] - predict(M3.cv, newdata = df[-train.ind,])
    M4.res <- df$length[-train.ind] - predict(M4.cv, newdata = df[-train.ind,])
    M5.res <- df$length[-train.ind] - predict(M5.cv, newdata = df[-train.ind,])
    M6.res <- df$length[-train.ind] - predict(M6.cv, newdata = df[-train.ind,])
    M7.res <- df$length[-train.ind] - predict(M7.cv, newx = reduced.X[-train.ind,], s =
    # mspe for each model
    mspe1[ii] <- mean(M1.res^2)
    mspe2[ii] <- mean(M2.res^2)
    mspe3[ii] <- mean(M3.res^2)
    mspe4[ii] <- mean(M4.res^2)
    mspe5[ii] <- mean(M5.res^2)
    mspe6[ii] <- mean(M6.res^2)
    mspe7[ii] <- mean(M7.res^2)
}

mspe_1 <- mean(mspe1)
mspe_2 <- mean(mspe2)
mspe_3 <- mean(mspe3)
mspe_4 <- mean(mspe4)
mspe_5 <- mean(mspe5)
mspe_6 <- mean(mspe6)
mspe_7 <- mean(mspe7)
mspe <- c("M1" = mspe_1, "M2" = mspe_2, "M3" = mspe_3,
          "M4" = mspe_4, "M5" = mspe_5, "M6" = mspe_6,
          "M7" = mspe_7 )
round(mspe, 4)
best_model = models_list[[which.min(mspe)]]
# summarize goodness of fit.
table = matrix(nrow = 7, ncol = 5)
for (i in 1:6)
{
    model = models_list[[i]]
    table[i,1] = summary(model)$r.squared
    table[i,2] = summary(model)$adj.r.squared
    table[i,3] = mean(summary(model)$residuals^2)
    table[i,4] = AIC(model)
    table[i,5] = BIC(model)
}
rownames(table) = paste0("M", 1:7)
colnames(table) = c("R2", "Adj R2", "MSE", "AIC", "BIC")
round(table, 4)
do_plots <- function(model, X, Xname)
{
    resid = resid(model)
    stdresid = resid/(sigma(model)*sqrt(1-hatvalues(model)))
    # distribution of studentized residuals.
    hist(stdresid, breaks=12, probability = TRUE,
         xlim = c(-4,4),
         xlab = "Studentized Residuals",
         main = "Distribution of Residuals")
    grid <- seq(-3.5, 3.5, length.out = 1000)
    lines(grid, dnorm(grid), col="blue")
    qqnorm(stdresid)
    abline(0, 1, col = "red")
    # residuals vs fitted
    plot(stdresid ~ fitted(model),
         xlab = "Fitted Values",
         ylab = "Studentized Residuals",
         main = "Residuals vs Fitted")
    # residuals vs X.
    for (i in length(X))
    {
        plot(resid ~ X[[i]],
             xlab = "Xname",
             ylab = "Residuals",
             main = paste0("Residuals vs ", Xname[i]))
    }
}
do_plots(M5, list(df$POP_furan3, df$ageyrs), c("POP_furan3", "ageyrs"))
summary(best_model)
```