

STAT 331 Final Project

Group 24

Muhammad Raza Nasser

Chijioke Pascal Ehirim

Vansh Joshi

Summary:

With a sample size of 864 adults and using exposure and other covariates we built a predictive model. The exploratory data analysis led us to quite interesting observations like, we found a high correlation amongst the exposures and a negative correlation between age and telomere length. Furthermore, it should be noted that the diagnostic plots of the data do not follow the normality assumption, hence, to correct this we used log transformation on the length variable. Under this transformation, it was found that all 4 of the assumptions were met. Additionally, we utilized LASSO regression for variable selection to build the most suitable predictive model. Also, we removed the education category in our analysis to improve the predictive accuracy as well as get the lowest mean squared error. Since there was a high amount of correlation amongst certain exposures, it was observed that only furan3 was used in our predictive model. Therefore, it can be concluded that we only need the value of the concentration of furan3 in an individual coupled with other covariates to predict the mean leukocyte telomere length in an individual.

Objective:

To find a predictive model, using the data of 864 adults, that finds the relationship between the leukocyte telomere length and 11 PCBs, 3 dioxins, 4 furans, and other covariates, including categorical like male, ageyrs, smokenow, and race and non-categorical variables like whitecell_count, BMI etc.

Exploratory data analysis:




























Data summary:



Number of rows:	864	Categorical variables:	4
Number of columns:	33	Numeric variables:	29

Summary statistics for the categorical variables:

variable	missing	categories	distribution
1 edu_cat	0	4	1: 270, 3: 228, 2: 199, 4: 167
2 race_cat	0	4	4: 448, 2: 191, 3: 154, 1: 71
3 male	0	2	0: 490, 1: 374
4 smokenow	0	2	0: 664, 1: 200

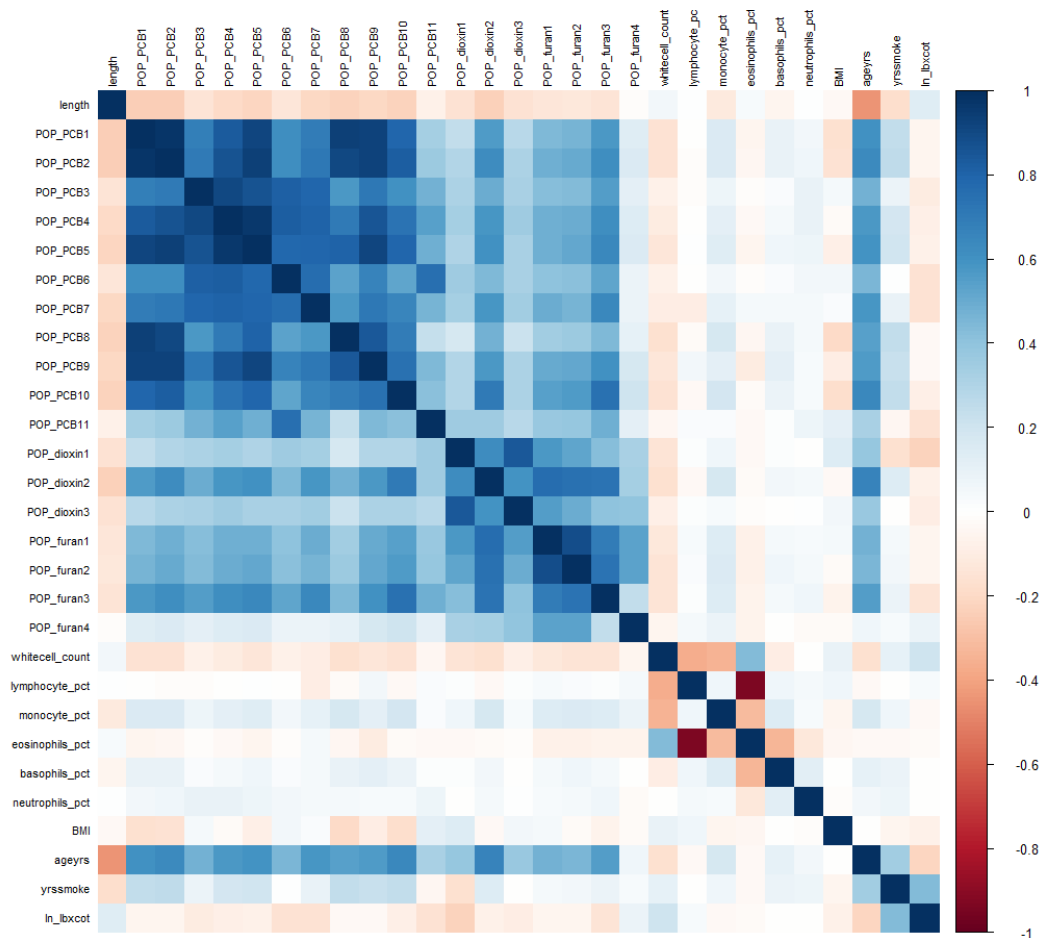
Summary statistics for the numeric variables:

Variable	missing	Mean	SD	p0	p25	p50	p75	p100	histogram
1 length	0	1.05	0.250	0.527	0.875	1.03	1.21	2.35	
2 POP_PCB1	0	38082.	40428.	2000	9975	27600	53325	572000	
3 POP_PCB2	0	15637.	14647.	2000	4800	11500	21825	165000	
4 POP_PCB3	0	10158.	10719.	2000	3700	6200	12000	123000	
5 POP_PCB4	0	38456.	41518.	2100	11475	25550	50650	487000	
6 POP_PCB5	0	52650.	54677.	2100	15600	36300	68625	708000	
7 POP_PCB6	0	16820.	24097.	2000	4400	9400	19500	319000	
8 POP_PCB7	0	12682.	13828.	1100	4000	7450	15625	144000	
9 POP_PCB8	0	10530.	11133.	1100	3800	6950	14425	187000	
10 POP_PCB9	0	12220.	12613.	1100	3900	8050	16025	144000	
11 POP_PCB10	0	24.5	20.8	1.7	9.1	18.4	34.9	172	
12 POP_PCB11	0	38.2	53.4	1.3	14.8	24.5	42.9	845	
13 POP_dioxin1	0	57.7	56.8	1.9	23.9	41.3	71.6	760	
14 POP_dioxin2	0	47.8	40.0	1.4	21.3	37.8	62.4	281	
15 POP_dioxin3	0	494.	524.	36.8	197.	342.	603	8190	
16 POP_furan1	0	6.37	4.88	1	3.2	5.2	7.7	44.4	
17 POP_furan2	0	5.39	4.18	0.8	2.6	4.2	6.82	33.5	
18 POP_furan3	0	6.67	5.79	0.7	2.2	5.05	9.3	38.3	
19 POP_furan4	0	11.5	11.1	0.9	6.4	9.65	14	234	
20 whitecell_count	0	7.19	2.13	2.3	5.6	6.9	8.3	20.1	
21 lymphocyte_pct	0	29.9	8.63	5.8	24	29.0	35.4	73.4	
22 monocyte_pct	0	7.94	2.10	1.6	6.6	7.7	9.1	23.8	
23 eosinophils_pct	0	58.6	9.57	21.6	52.4	59.3	65.2	88.1	
24 basophils_pct	0	2.90	2.29	0	1.5	2.3	3.7	28.2	
25 neutrophils_pct	0	0.667	0.475	0	0.4	0.6	0.8	5.5	
26 BMI	0	28.1	6.02	16.2	23.9	27.4	31.2	63.0	
27 ageyrs	0	48.4	18.3	20	34	46	63	85	

28 yrssmoke	0	10.6	15.4	0	0	0	20	69	
29 ln_lbxcot	0	-0.980	3.83	-4.51	-4.07	-2.73	2.80	6.58	

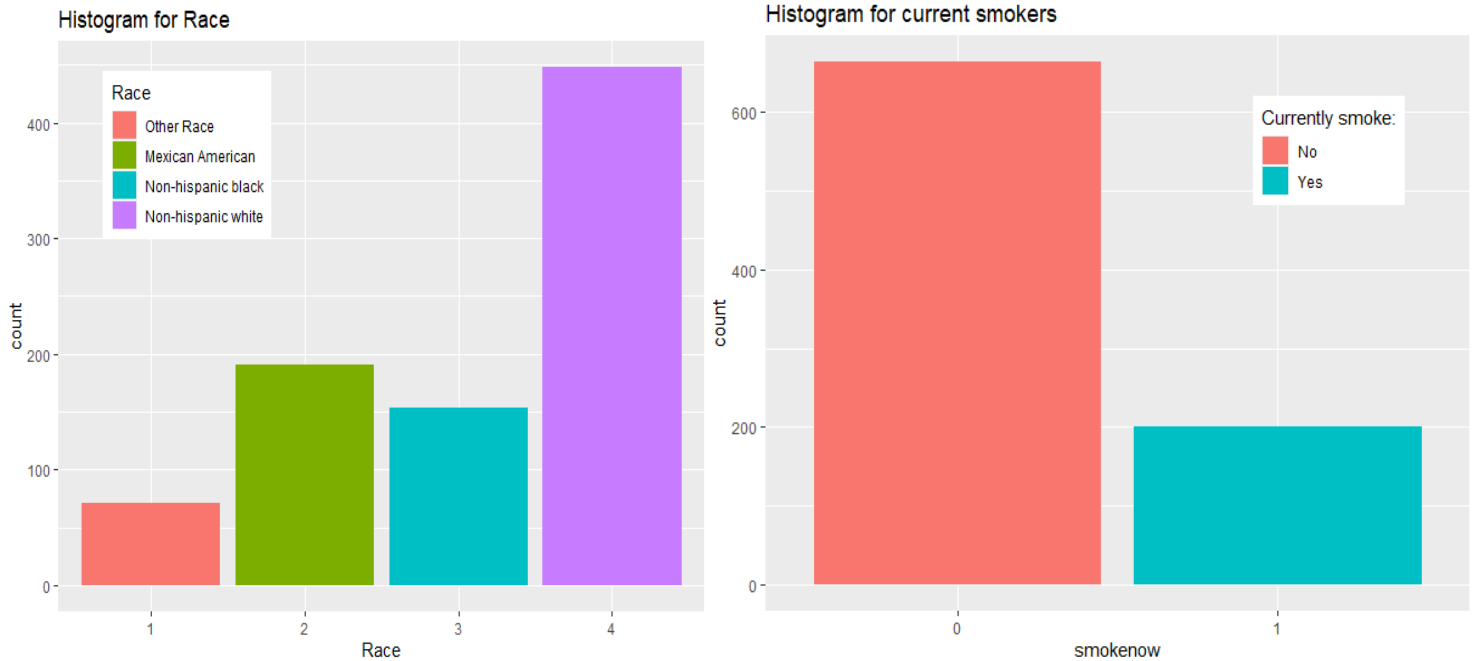
Observations: It can be clearly seen from the histograms that all the exposures are right skewed. Additionally, it was observed that the concentration of the first 9 PCBs is extremely large as compared to PCB 10 and 11 and an opposite pattern was seen in dioxins as the first two dioxins were quite small in concentration and dioxin 4 was quite large in concentration. Finally, all the exposures also have a quite large standard deviation.

Correlation Heat Map:



Observations: The heat map also offers some critical observations such as that the correlation between PCB1 and PCB2 is 0.9710387, which signifies that these two exposures are quite closely linked and move in the same direction. Similar correlation levels can also be seen amongst other PCB exposures. On the other hand, eosinophils_pct and lymphocyte_pct have a very high negative correlation of -0.9346368, meaning that these two covariates almost always move in the opposite direction. An interesting correlation pattern was observed regarding Age years, as this covariate is positively correlated with all the exposures but negatively correlated (-0.4454242) with length.

We can now observe the categorical variables `race_cat`, which has four levels (“Other Race”, “Mexican American”, “Non-Hispanic black” and “Non-Hispanic white”), which tell us about the race of our observed adults for this study, and the categorical variable `smokenow` which has 2 levels (“Does not currently smoke” and “Currently smoke”) which depicts whether the observed adult currently smoke or not. The following histograms depicts the distribution of the race and current smokers in the observed data:



Observations:

These histograms offer an important glimpse into the dataset and as such it can be observed that the sample size is dominated by adults of the Non-Hispanic white race, and their number is almost twice as more than any other race. Moreover, more than two-thirds of the adults in the sample are non-smokers, 664, to be more precise.

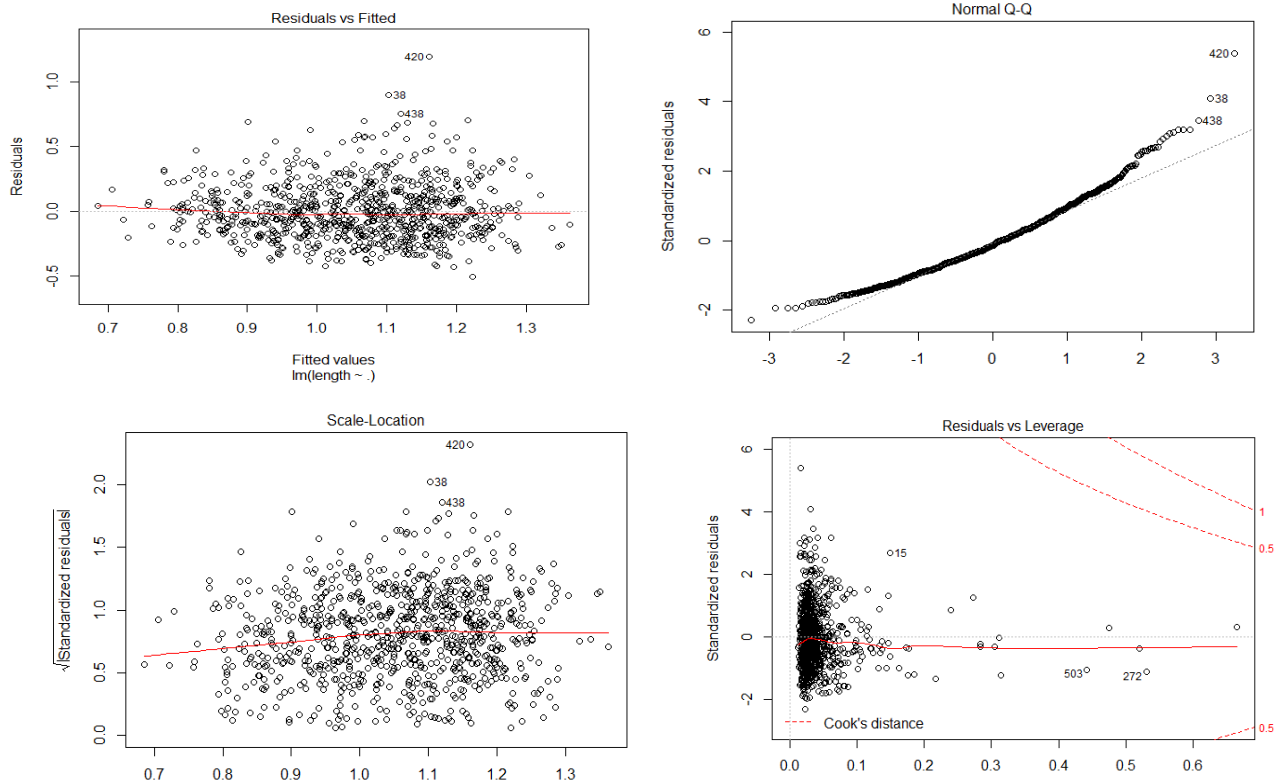
Method:

Log(length) = $\beta_0 + \beta_1 \text{POP_furan3} + \beta_2 \text{Whitecell_count} + \beta_3 \text{lymphocyte_pct} + \beta_4 \text{monocyte_pct} + \beta_5 \text{BMI} + \beta_6 \text{I}(\text{race_cat} = 3) + \beta_7 \text{I}(\text{race_cat} = 4) + \beta_8 \text{I}(\text{male} = 1) + \beta_9 \text{ageyrs} + \beta_{10} \ln_l\text{bxcot} + \epsilon_i$,

Variable	Estimate
Intercept	0.3409
POP_furan3	0.0017
whitecell_count	-0.0016
lymphocyte_pct	0.0001
monocyte_pct	-0.0013
BMI	-0.0011

race_cat3	0.0425
race_cat4	-0.0100
male1	-0.0182
ageyrs	-0.0055
ln_lbxcot	0.0018

When we use all the covariates and no transformation on our model, we get the following diagnostic plots:

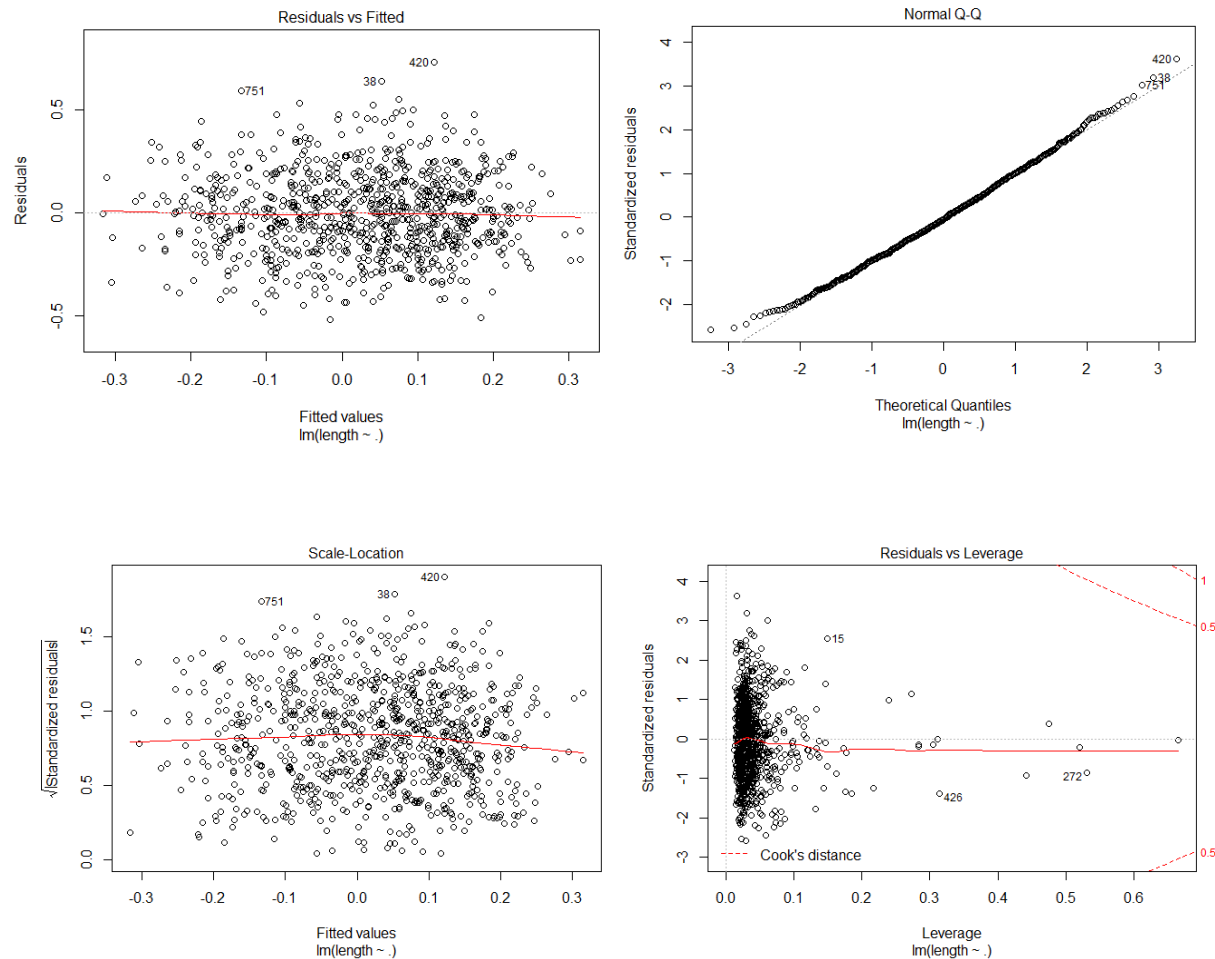


Observations (diagnostic for full model):

- **Residual vs fitted:** If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you do not have non-linear relationships. In our chart the residuals are evenly spread, therefore it shows the relationship is linear.
- **Normal QQ:** It is desirable if the residuals are lined well along the straight dashed line. Since our graph does not have the residuals aligned along the straight dashed line, therefore it is not normal.
- **Scale-location:** This shows that if the points are evenly scattered then the relationship is homoscedastic. Since in our graph the points are not evenly scattered, this shows that is heteroscedastic.

- Residual vs leverage: Within this graph we look for cases that are outside of the dashed lines, also known as Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. Since all the points on our plot are within Cook's distance, this means the data is not skewed.

After Log transformation:



- Residual vs fitted: After the log transformation, in our chart the residuals remain evenly spread, therefore it shows the relationship remains linear.
- Normal QQ: After the log transformation our graph shows that the residuals have aligned along the straight dashed line, therefore it has become normal.
- Scale-location: After the log transformation our graph shows that the points have become evenly scattered, this shows that the relationship has now become homoscedastic.
- Residual vs leverage: After the log transformation all the points on our plot remain within Cook's distance, this means the data is still not skewed.

To find the most suitable combination of lasso and ridge regression for our model we used elastic regression and splitting our dataset into two parts, training set (with first 600 observations) and

test set (with last 264 observations). Next, using elastic regression we found an alpha which gives us the most suitable combination of lasso and ridge regressions and the respective mean-squared error.

Alpha	MSE
0	0.04439493
0.1	0.04361567
0.2	0.04316252
0.3	0.04295783
0.4	0.04287323
0.5	0.04272931
0.6	0.04278797
0.7	0.04263170
0.8	0.04266738
0.9	0.04250608
1	0.04259855

From the table, it can be observed that, with alpha 0.9 we have the minimum mean squared error value of 0.04250608. Hence, we used the combination of lasso and ridge regression, where alpha = 0.9, to get our reduced model:

Variable	Estimate
Intercept	0.3580
POP_PCB1	-1.868629e-08
POP_dioxin3	-2.313782e-07
POP_furan3	0.0024
whitecell_count	-0.0032
lymphocyte_pct	0.0002
monocyte_pct	-0.0026
neutrophils_pct	0.0024
BMI	-0.0012
edu_cat3	0.0185
edu_cat4	0.0303
race_cat3	0.0380
race_cat4	-0.0218
male1	-0.0224
ageyrs	-0.0055
ln_lbxcot	0.0033

After removing the edu_cat variable we get the following table after applying the elastic regression:

Alpha	MSE
0	0.04358134
0.1	0.04289218
0.2	0.04253731
0.3	0.0432565
0.4	0.04219883
0.5	0.04216646
0.6	0.04205294
0.7	0.04211622
0.8	0.04202033
0.9	0.04211469
1	0.04200420

Reason for removing education: After removing the covariate of education and using elastic regression, we see that $\alpha = 1$ gives us the smallest value of the mean-squared error (0.04200420). Hence, we get our new reduced model using lasso regression.

Therefore, this is our final prediction model that gives us the least mean squared error and hence, will predict the mean leukocyte telomere length in an individual. Moreover, our data, after performing the log transformation, satisfy the four important assumptions of normality, homoscedasticity, linearity and independence (as shown in the diagnostic plots of data after log transformation).

Results:

Final model estimates:

Variable	Estimate
Intercept	0.3409
POP_furan3	0.0017
whitecell_count	-0.0016
lymphocyte_pct	0.0001
monocyte_pct	-0.0013
BMI	-0.0011
race_cat3	0.0425
race_cat4	-0.0100
male1	-0.0182
ageyrs	-0.0055
ln_lbxcot	0.0018

After running our model on a test data out of our original data of 264 observations (training data of 600 observations) we got a mean squared error of 0.04200420. We reached this MSPE after implementing the log transformation and removing the categorical variable, edu_cat. It should be noted that only 1 out of 18 exposures i.e., no PCBs and no dioxins and out of all the furans, furan3 was the only exposure that was used in our prediction model. It is reasonable to include furan3 as the only exposure in our model as furan3 is highly correlated with other exposures

Relation between the Age of an adult and log of telomere length:

Age range	Mean log length
20 to 35	0.1414
35 to 60	0.0443
60 and above	-0.1194

Hence, we observe that as the age increases the mean log length decreases.

Relation between the Race of an adult and log of telomere length:

Race	Mean log length
Other races	0.0848
Mexican American	0.0353
Non-Hispanic Black	0.0771
Non-Hispanic White	-0.0057

Therefore, it can be seen that non-Hispanic whites, on average, have the shortest mean log length. Whilst on the other hand, those not belonging to non-Hispanic White, Non-Hispanic Black and Mexican American races have, on average, the longest mean log length.

Relation between the Race of an adult and log of telomere length:

Currently Smoke	Mean log length
Yes	0.0685
No	0.0126

Consequently, the table demonstrates that the mean log length for adults who currently smoke is higher as compared to adults who do not currently smoke.

Discussion:

Since we have achieved a relatively low mean squared error through our model, this shows that we have found a model that is quite effective at discovering the relationship between the mean telomere length in an individual and 18 different exposures coupled with other covariates. Using the original data, we got 3 exposures (PCB1, dioxin3, furan3) in our model, but we also got a relatively high MSPE, and our data was right skewed. Therefore, to meet all the assumptions and

to resolve the issues of MSPE and skewness we used the log transformation on the length as well as removed the categorical variable of edu_cat. This normalized the data, made the data homoscedastic and gave us a relatively low MSPE.

Limitations:

- Majority of the adults in our data belong to the Non-Hispanic White category, which means our data might be biased.
- Similarly, a majority of the adults within the data do not currently smoke when compared with the people who currently smoke, this also poses the risk of the data being not representative.
- Also, using LASSO we lose quite a lot of the exposures because they do not fit into the model. This is a downside of using lasso as we might have lost some exposures and other covariates that might, otherwise, be quite beneficial to our model.

Appendix:

We read in our data file and remove the first column, index

```
dat <- read.csv("pollutants.csv")
```

```
dat = dat[,-1] ##removing the first coloumn of index from data
```

We convert the categorical variables as factors

```
dat$male = as.factor(dat$male)
```

```
dat$edu_cat = as.factor(dat$edu_cat)
```

```
dat$smokenow = as.factor(dat$smokenow)
```

```
dat$race_cat = as.factor(dat$race_cat)
```

Exploratory data analysis:

we use the Skim function to skim through and look at a summary

statistics of the data

```
library(skimr)
```

```
skim(dat)
```

We then build a correlation matrix/plot to show the correlation

between the numeric variables in the data

```
library(corrplot)
```

```
df = dat[-c(27,28,29,32)]
```

```
corr_mat = cor(df)
```

```
corrplot(corr_mat,method = "color",tl.col="black",tl.cex = 0.7)
```

We build a histogram for the race category

and the Currently smoking catgeory

```
library(ggplot2)
```

```

ggplot(dat) + geom_bar(aes(x=race_cat,fill=race_cat)) +
  xlab("Race") + ggtitle("Histogram for Race") +
  scale_fill_discrete(name = "Race", labels = c("Other Race",
                                                "Mexican American",
                                                "Non-hispanic black",
                                                "Non-hispanic white")) +
  theme(legend.position = c(0.2,0.8))

ggplot(dat) + geom_bar(aes(x=smokenow,fill=smokenow)) +
  ggtitle("Histogram for current smokers") +
  scale_fill_discrete(name = "Currently smoke:", labels = c("No", "Yes")) +
  theme(legend.position = c(0.8,0.8))

```

Full model and model diagnostics

```

mfull = lm(length~.,data = dat)
plot(mfull)

```

Log transformation on length and model diagnostics

```

dat$length = log(dat$length)
mfull_log = lm(length~.,data = dat)
plot(mfull_log) ## model diagnostic plots

```

Method:

we run elastic net regression to find best model

```

set.seed(24)
library(glmnet)
mo = lm(length~., data = dat) #full model

```

```

X = model.matrix(mo)[-1] #covariate matrix for our data
y = dat$length #response variable (length)
train_id = 1:600 #splitting train and test data set
X_train = X[train_id,] #first 600 rows for the train dataset
X_test = X[-train_id,] #next 264 rows for the test dataset
y_train = y[train_id] #first 600 response variable to train dataset
y_test = y[-train_id] #next 264 response variable to test dataset
list.fit = list()

# it calculates the mspe of our reduced model for alphas equal to
# 0,0.1,...,1.
for(i in 0:10){
  fit.name = paste0("alpha", i/10)
  list.fit[[fit.name]]=
    cv.glmnet(x=X_train,y=y_train,alpha = i/10)}

results = data.frame()
for(i in 0:10){
  fit.name = paste0("alpha", i/10)
  pred=
    predict(list.fit[[fit.name]],s="lambda.min",newx = X_test)
  mse = mean((y_test-pred)^2)
  store = data.frame(alpha = i/10, mse,fit.name)
  results = rbind(results,store)
}
print(results)

```

```

alpha0.9.fit = cv.glmnet(x=X_train,y=y_train,alpha = 0.9)
alpha0.9.pred = predict(alpha0.9.fit,s="lambda.min",newx = X_test)

mean((y_test-alpha0.9.pred)^2) #mspe of alpha = 0.9

round(coef(alpha0.9.fit,s='lambda.min'),4) #coefficients of reduced model with alpha =0.9

## We remove the education category and re-run elastic-net regression
## to find a better model
set.seed(24)
library(glmnet)
mo = lm(length~.-edu_cat, data = dat)
X = model.matrix(mo)[-1]
y = dat$length
train_id = 1:600
X_train = X[train_id,]
X_test = X[-train_id,]
y_train = y[train_id]
y_test = y[-train_id]
list.fit = list()

for(i in 0:10){
  fit.name = paste0("alpha", i/10)
  list.fit[[fit.name]]=
    cv.glmnet(x=X_train,y=y_train,alpha = i/10)}

results = data.frame()

```

```

for(i in 0:10){
  fit.name = paste0("alpha", i/10)
  pred=
    predict(list.fit[[fit.name]],s="lambda.min",newx = X_test)
  mse = mean((y_test-pred)^2)
  store = data.frame(alpha = i/10, mse,fit.name)
  results = rbind(results,store)
}
print(results)

```

```

alpha.fit = cv.glmnet(x=X_train,y=y_train,alpha = 1)
alpha.pred = predict(alpha.fit,s="lambda.min",newx = X_test)

```

```

mean((y_test-alpha.pred)^2) #mspe for alpha = 1

```

```

round(coef(alpha.fit,s='lambda.min'),4) #coefficients of reduced model with alpha = 1

```

Results:

We create a table showing mean log length for different ranges in Ageyrs

```

s1 = mean(dat$length[which(dat$ageyrs>=20 & dat$ageyrs<=35)])

```

```

s2 = mean(dat$length[which(dat$ageyrs>35 & dat$ageyrs<=60)])

```

```

s3 = mean(dat$length[which(dat$ageyrs>60)])

```

```

c = c(s1,s2,s3)

```

```

x = matrix(nrow = 3,ncol = 2)

```

```

x[,2] = round(c,4)

```

```

x[,1] = c("20 to 35", "35 to 60", "Greater than 60")

```

```

colnames(x) = c("Age Range", "Mean of log length")

```

```
as.table(x)
```

```
## Table of mean log length for different race categories
```

```
r1 = mean(dat$length[which(dat$race_cat==1)])
```

```
r2 = mean(dat$length[which(dat$race_cat==2)])
```

```
r3 = mean(dat$length[which(dat$race_cat==3)])
```

```
r4=mean(dat$length[which(dat$race_cat==4)])
```

```
r = c(r1,r2,r3,r4)
```

```
x = matrix(nrow = 4,ncol = 2)
```

```
x[,2] = round(r,4)
```

```
x[,1] = c("Other race", "Mexican American",  
          "Non-Hispanic Black", "Non-Hispanic White")
```

```
as.table(x)
```

```
## mean log length for different currently smoking categories
```

```
smoke0 = mean(dat$length[which(dat$smokenow==0)])
```

```
smoke1 = mean(dat$length[which(dat$smokenow==1)])
```