

SUMMARY

In this report, we analyze the pollutant.csv data which contains a sample of $n = 864$ adults included in a study investigating the relationship between exposure to persistent organic pollutants and leukocyte telomere length—a marker of cellular aging that may be related to certain cancers. Our goal is to determine whether/which pollutants improve prediction of the outcome (i.e., length).

Our methods for achieving this goal is by making a model for prediction where we consider various types of model selection: forward, backward, stepwise where we use both criteria AIC and BIC, and LASSO. Next, we will verify that the LINE assumption is met for the chosen model and answer the question to our objective. In the end, we understood that 1 particular pollutant type assisted in the improvement of prediction of the outcome.

Finally, we will discuss what we did differently in our methods which could be improved on.

OBJECTIVE

The main objective of this project is to determine whether/which pollutants improve prediction of the outcome (i.e., length) by finding the best predictive model.

DATA PREPROCESSING

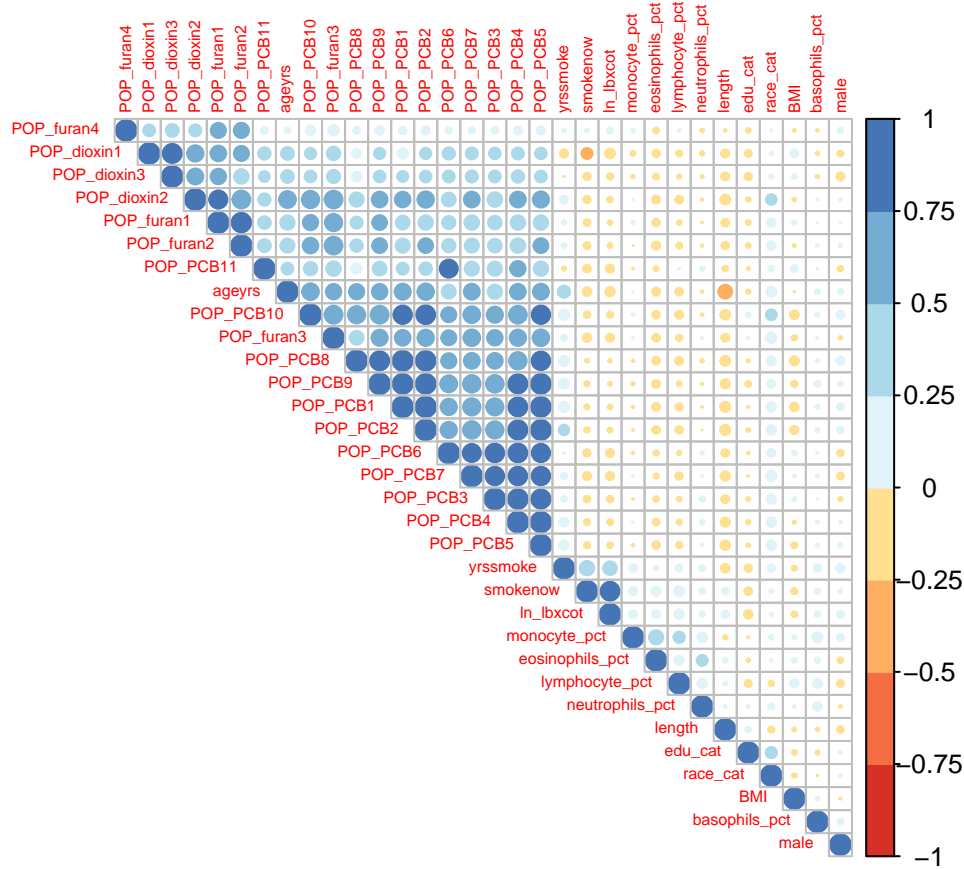
Our group did the regular data-preprocessing step: change categorical data to the factor data type. Also, we removed the whitecell_count (= number of white blood cells) covariate since we multiplied the covariates which represent the percentage of these cells with this value. The main reason we did this was for the model matrix to have independent columns (i.e., percentage of white blood cell does not rely on the number of white blood cell). Lastly, we acknowledge how the ‘pollutant’ dataset is a complete dataset without missing values.

EDA

In this section, we made plots of different kinds for all the covariates of the dataset in hopes of finding some interesting results.

First, here is a correlation plot showing the correlation of all the covariates and outcome in the dataset.

```
## corrplot 0.84 loaded
```



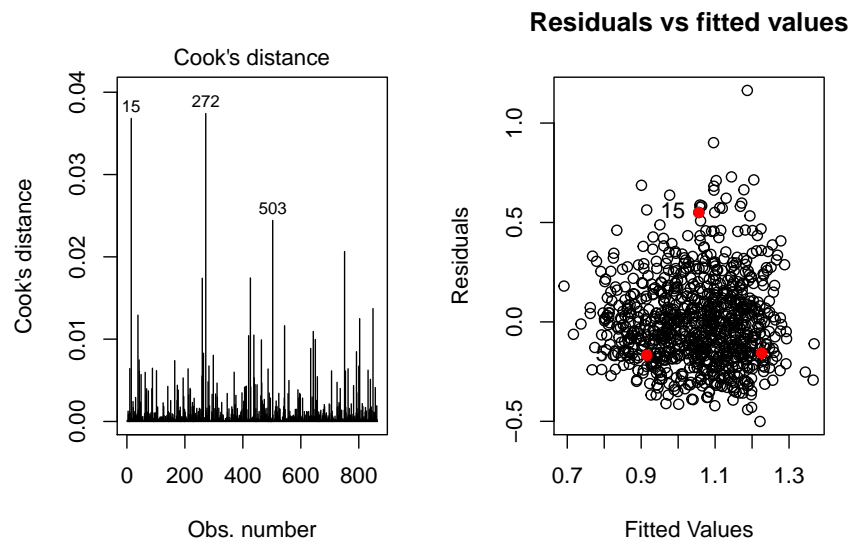
From this correlation plot, interesting findings include age and the outcome (telomere length) being negatively-correlated, and the pollutant dioxin1 also being quite negatively-correlated to the covariate smoke_now.

Also, looking at the vertical column (length), we can deduce that there is almost no correlation between length and the covariates POP_furan4 (first row), and all the white blood cell components. So, from this, in our work of finding a prediction model, we will consider removing the insignificant covariates which almost have no correlation with length. We will do this using hypothesis testing (Appendix: line 359).

Now, we summarize the prominent findings of the code in the Appendix: Section Explanatory Data Analysis.

Focusing on the adults taking part in this study, from the plots generated, we can summarize the majority of them are non-smokers of the non-Hispanic White race. Also, there are more females taking part in this study than males. The average BMI indicates that the participants taking part in this study are on average overweight. Next, they are between the age-group of around 34-63.

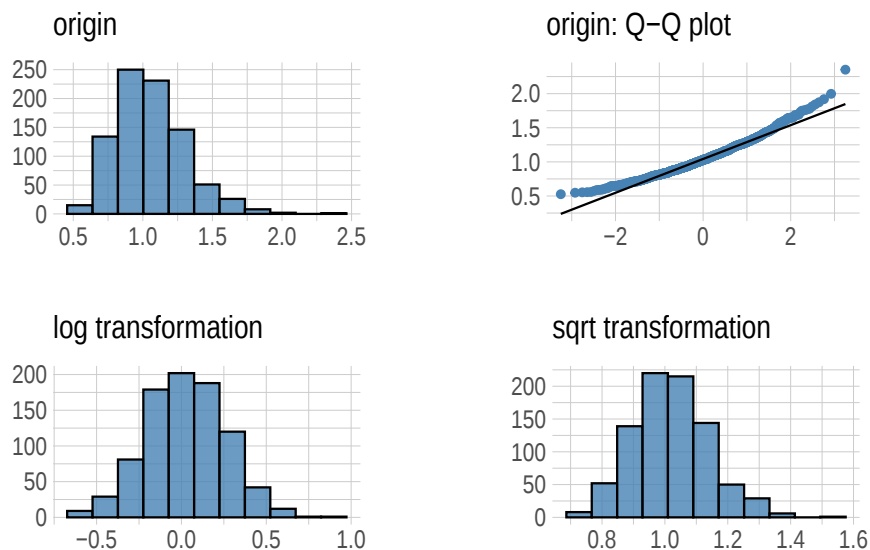
Now, we focus on the identification of outliers in the dataset if present. This is done by using Cook's distance and plotting the residuals vs. fitted values.



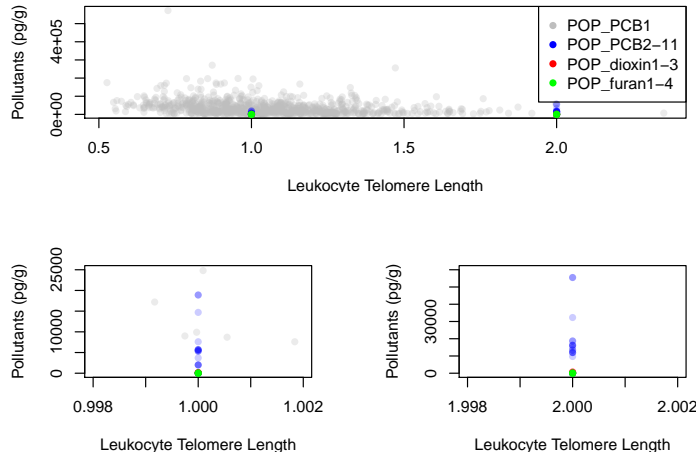
From the first plot, it is visible that in all the data, there are 3 very influential datapoints at index 15,272,503. There is also one more at around index 770. Now, we highlight these influential points in the residual vs. fitted values. Here, in the second plot, we see that the points fit the overall model and will not have much effect on the model with or without them. This will be further explained in the limitations section of the discussions section.

The final interesting observation we noticed from the appendix code was that all the variables in the dataset were right-skewed. This is seen from all the normality diagnosis plots in the Appendix: Section EDA, plot origin. However, in all the normality diagnosis plots, it can be seen that the log transformation fixes this skew, since the histogram of the log transformed data is bell-shaped. Here is an example with the variable length:

Normality Diagnosis Plot (length)



Since the study aimed to investigate the relationship between length and the pollutants, a scatterplot of pollutants(pg/g) vs. telomere length has been plotted.



The first plot considers all the telomere length, whereas the bottom two plots are enlarged versions of the plot focusing on length 1 and 2.

From here, we can see that pollutants of type PCB1 seem to have occurrences over all the telomere lengths, whereas pollutants of type PCB2-11, dioxin1-3, and furan1-4 only have occurrences for leukocyte telomere length of length 1 and 2.

METHODS

We divide our observations into two separate datasets, training set (traindat) and a test set (testdat). Traindat contains 664 of the observations in pollutant.csv and testdat contains the other 200 observations. We let testdat be the holdout set that we will later use to evaluate the MSPE of our selected models on different data with the goal of providing unbiased results.

We decided to use three different model selection methods to select three different best models according to the criteria of each respective model selection method. We start with the full model which contains all the potential covariates and the intercept. Next, we will get our candidate models using some selection algorithm. The criteria for our candidates are based on AIC, BIC, and LASSO.

Recall that to do a selection algorithm, we need a starting model. The starting model we decided to go with contains all the potential covariates and intercept (i.e., full model). Here, we chose the stepwise selection because it is the most efficient compared to forward and backward selection. In addition to that, using forward or backward selection may cause the wrong removal of some covariates. Next, we also consider the more modern approach, LASSO, to also be a candidate for the best fitting model.

After obtaining the three candidate models, we wanted to take interactions between the covariates and itself into consideration. Therefore, we also applied the stepwise selection with AIC and BIC, as well as LASSO, taking into account all possible one-to-one interactions. Our starting model contains all the potential covariates squared and the intercept. Looking at the covariates present in each model, we realized that the model with interactions obtained from stepwise selection using BIC, is the same with the model obtained using the same method but without interactions. Thus, we have 5 models to consider.

From EDA, based on the correlation plot, we also notice that POP_furan4, and all the white blood cell components seems to be uncorrelated to length. Furthermore, as discussed in the EDA, we observed that the p-values of lymphocyte, eosinophils, basophils, neutrophils, BMI, edu_cat (1= less than 9th grade of 9-11th grade), edu_cat (2= highschool grad/ GED or equivalent) and all the race_cat assessed individually

are not significant at the $\alpha = 0.05$ level. Hence, we removed those covariates from our data and repeated our above procedure. We split the data into training and testing datasets, we conducted stepwise selection based on AIC and BIC as well as LASSO on the training set, considering both all possible one-to-one interactions and no interactions at all.

Thereafter, we realized that the models (with and without interactions) we obtained applying stepwise selection based on BIC are identical to the stepwise selection based on BIC that we did on the data that we did not manipulate. For that reason, we only take the model (no data manipulations and no interactions) obtained by stepwise using BIC into account.

Onward, we checked all nine models for the assumptions of linear regression: linearity, independence, normality and heteroskedasticity. We also calculated their respective mean squared prediction error and its square root to measure the prediction accuracy of each model. Below is the summary of the RMSPE of each models:

	From model with all covariates		From model with only significant covariates	
Without Interactions i.e. .	AIC	0.228444*	AIC	0.227353*
	BIC	0.221453	LASSO	0.223235
	LASSO	0.222867		
With Interactions i.e. .^2	AIC	0.242596*	AIC	0.243453*
	LASSO	0.228189*	LASSO	0.227725*

Top 3 models based on RMSPE:

1. Model selected based on stepwise selection (criteria: BIC).
2. Model selected based on LASSO with no interactions and no data manipulation.
3. Model selected based on LASSO with no interactions, done on the data containing only significant covariates.

We choose the BIC, since:

1. It has the lowest RMSPE,
2. Parsimony -> model with the least covariate
3. It is very interpretable: Since it has no interactions and has the least covariate it is very easy to interpret the data.

RESULTS

We have

$$\text{Length} \sim 1.4351718 + 0.0070121\text{POP_furan3} - 0.1453943\text{monocyte_count} - 0.0070514\text{ageyrs}$$

as our chosen model.

	Estimate	Standard Error	P-values
Intercept	1.4351718	0.0363128	< 2e-16
POP_furan3	0.0070121	0.0018724	0.000196
monocyte_count	-0.1453943	0.0494699	0.003407
ageyrs	- 0.0070514	0.0005721	< 2e-16

We ended up having 3 variables in the final model. The table above can be interpreted as follows:

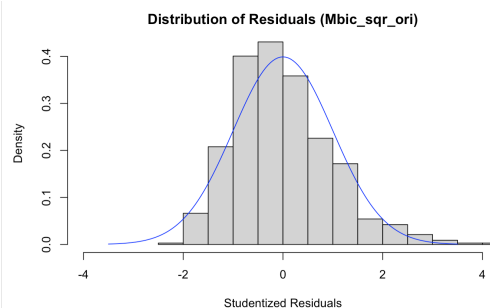
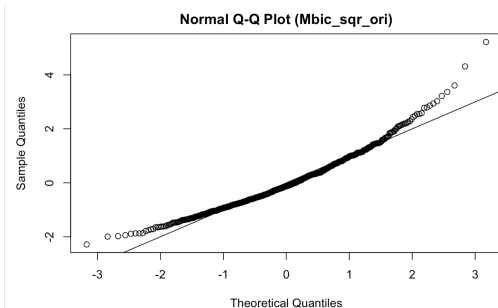
1. The mean outcome when the covariates POP_furan3, monocyte_count, and ageyrs are zero is 1.4351718.
2. The mean difference in the outcome for every 1 unit increase in POP_furan3, holding other covariates fixed is 0.0070121.
3. The mean difference in the outcome for every 1 unit increase in monocyte_count, holding other covariates fixed is -0.1453943.
4. The mean difference in the outcome for every 1 unit increase in ageyrs, holding other covariates fixed is -0.0070514. Out of all three covariates, the largest mean change in the outcome for some 1 unit increase is the monocyte_count.

All the p-values for our covariates are very small, and hence, they are all very statistically significant, with the most statistically significant being ageyrs.

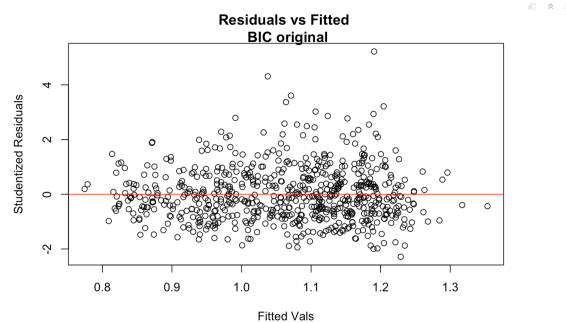
Overall, based on the residual plot and RMSPE, our final prediction model predicts the test data the best. We have assumed independence for our data. Then, here is how we prove the assumptions linearity, normality, and homoskedasticity.

Assumptions

###Normality: From the QQPlot and plot distribution of studentized residuals we can see that the data is slightly skewed to the right. Besides the slight skew, we can say that the normality assumption is fairly reasonable.

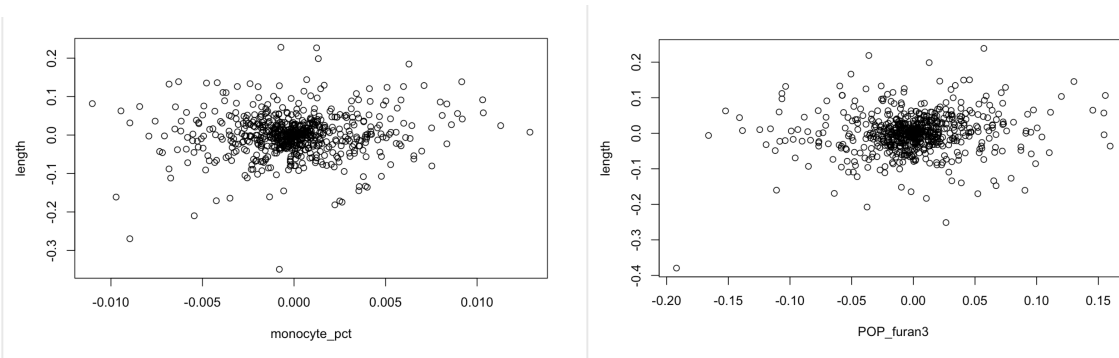


Heteroskedasticity: From the studentized residuals vs fitted values scatterplot, we can see that the points lie more or less horizontally within a constant band around the horizontal line where residual is zero. Therefore, we may conclude that our data has constant variability.



Independence: Looking at our data, none of their observations seem to depend on each other. Moreover, we removed the covariate whiteblood cell and changed all its components into their own amounts. Therefore, it seems reasonable to assume that our data is independent.

Linearity: by plotting the residuals of y to residuals of each of the covariates, notice that the points on the normal qqplot lie more or less along a straight line. Thus, linearity assumption may be assumed.



Hence, all of our regression assumptions have been reasonably satisfied by the final model. There are a couple of outliers, but not too much that it affects the assumptions.

One interesting finding is that we thought we would end up with the LASSO model, because it is the newest and more complicated algorithm out of them all. However, we ended up with the BIC model, which was quite unexpected.

DISCUSSIONS

Using the model found in results, we can answer the objective question: Determine whether/which pollutants improve prediction of the outcome (i.e., length). Since the model only includes pollutant POP_furan3, this is the pollutant which improves prediction of the outcome, which is quite interesting given how this type of pollutant was only a component of telomere lengths 1 and 2. Other telomere lengths didn't have any of the furan3 pollutants recorded for it. However, this result is true due to some decision we made in the methods.

1. Recall that in the EDA section, we observed that most of the variables in the data are right-skewed, and that using a log-transformation on the data fixes this skew and makes the variable normally-distributed. So, from here, it would have been a good decision to use log-transformed variables instead to find the best predictive model. However, we chose to not do that as it would be difficult to interpret the meaning behind the log-transformed data, although using the transformed data might yield a more accurate model and possibly a different answer to our objective.
2. We did not remove the outliers(influential units) we found in the EDA section before dividing the data into a test and training set, since we could not find any legitimate reason to remove this data other than improving the trends and possibly the model picked. It is not shown in the appendix, but when we checked the individual units in hopes of finding an explanation to why it is so influential, we found that the covariate values of these influential units mingled well with the non-influential units. So, we couldn't find a reason why the units should be removed, and we believed that it wouldn't impact the analysis done. However, it is worth a shot doing our methods with the outliers removed to check for a better result.
3. We only used RMSPE as our indicator of deciding which model is the best. However, there are many more methods of deciding this such as the log accuracy ratio(Tofallis, 2014), which could give different results on what model is chosen. Of course, we can always check how good this model is by comparing its predicted outcome value and the actual outcome value in the testing set.

Now, we address the sampling error in the dataset. As mentioned in EDA, the dataset has the following biases: The majority of the adults in the study are non-smokers of the non-Hispanic White race. Also, there are more females taking part in this study than males. Next, the average BMI indicates that the participants taking part in this study are on average overweight. Finally, they are between the age-group of around 34-63. More variation might increase the credibility of the results found.

References

Tofallis, C. (2014, November 12). A better measure of relative prediction accuracy for model selection and model estimation. Retrieved April 14, 2021, from <https://link.springer.com/article/10.1057/jors.2014.103>