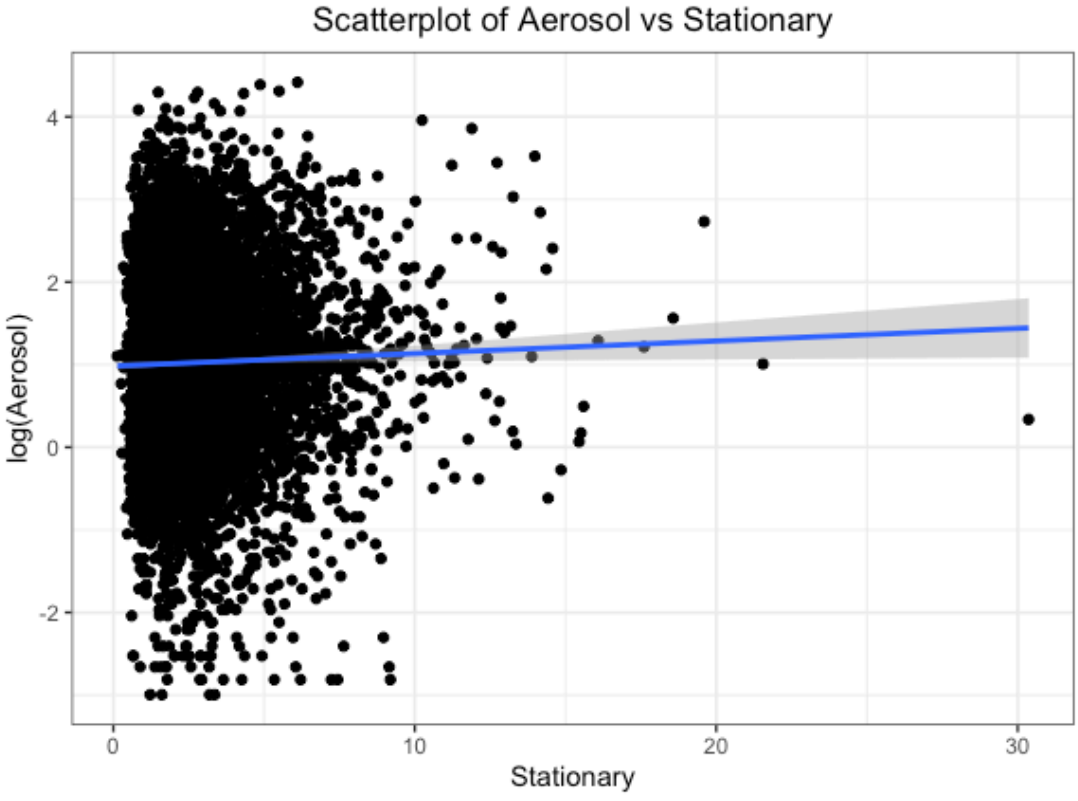
Michael Pixton

Midterm 2 – PM exposure

Section 1: Introduction and Problem Background

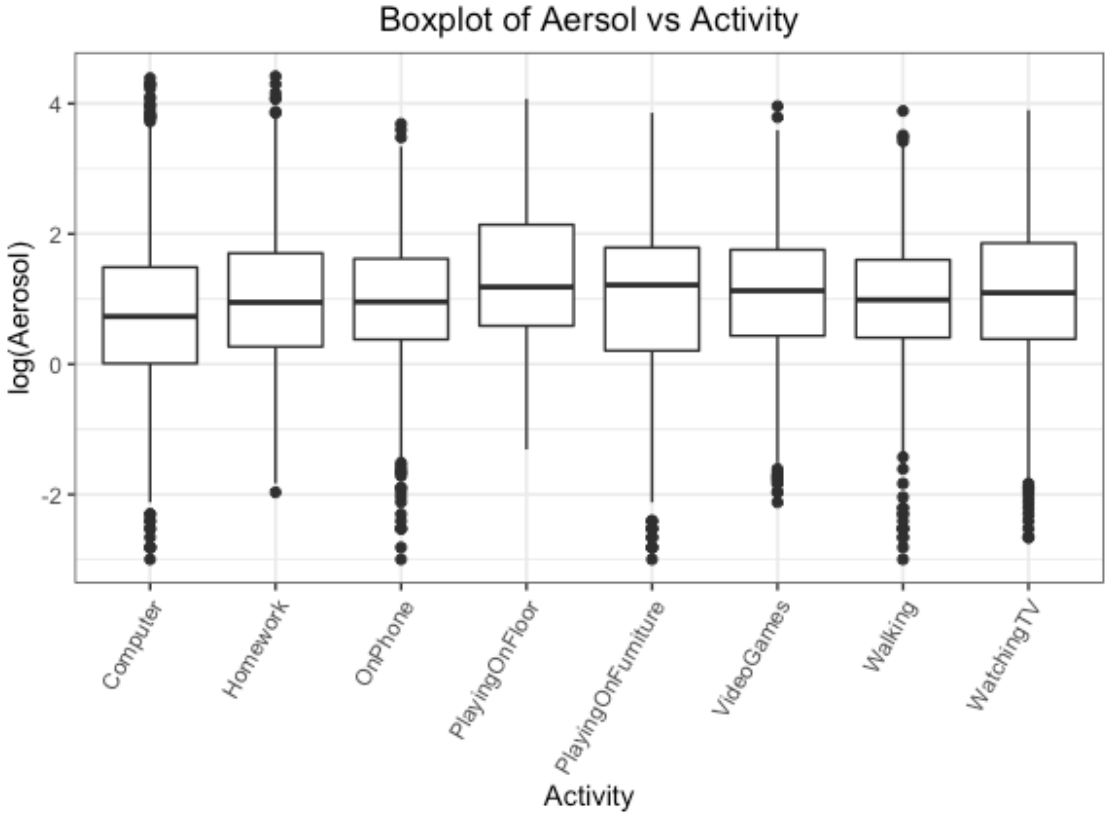
Particulate Matter or PM has been used as a measure of pollution for many years. It has been found to be correlated to many health risks and long-term health ailments. Therefore, many studies continue to be used to measure PM and look at risk factors. A dataset was collected to look at three important effects. The first, is how effective is PM being read using a stationary device, the second is what are the effects of PM on different activities a child does, and thirdly, do these effects change from child to child. The dataset contains readings for every minute with values for: child's ID, Aerosol (PM readings from a mobile device on the child), Stationary(PM readings from the stationary device) and Activity(type of activity the child was engaged in during that minute).We wish to use statistical modeling in order to analyze this dataset and fit a linear model. This model with help us answer the research questions relating to effectiveness of the stationary PM reader and the effects of children's activities.

An exploratory data analysis was done to look at the general appearance of our dataset. The plot below is a scatterplot of Aerosol PM readings vs Stationary PM reading.



Michael Pixton

The original data was found to not fit well, the range of Aerosol PM was 83 with a standard deviation of 6.5, so a transformation was performed to get the scatterplot above. A few different types of transformations were attempted, but the natural log of Aerosol was found to be the most effective. Log(Aerosol) will be used for better model fitting from now on. The scatterplot also shows us that Aerosol and Stationary do look to be linear, but their correlation doesn't look too strong. Another plot was created to show the spread of log(Aerosol) over the different activities.



Based on this plot, it seems that most of the activities are about the same, we are mostly just looking at relative differences here in the activities. Playing on the floor, and playing on the furniture seems to be the highest. We also see a lot of outliers here, hopefully our transformation will accommodate them when we create our linear model.

Another issue that was determined in this exploratory analysis is the issue of independence. A linear model assumes each of the data points is independent of any other data point, but we know that our dataset was 118 points (each minute) taken for each child.

This clearly breaks our independence assumption. To look at the correlations within each child, a 118 by 118 correlation structure was calculated. The mean of this structure was .81, a very high correlation, implying that our issue is a real problem and must be dealt with when we create our model.

Based on this exploratory data analysis, a Longitudinal Multiple Linear Regression Model will be our model of choice. This model will not only deal with our independence assumption, but will take advantage of the way our data is set up over a series of time to create a hopefully more accurate model to answer our research questions.

Section 2: Statistical Model

The Longitudinal Multiple Linear Regression Model first needs a correlation structure so that we can account for the independence assumption issue. Our correlation structure will be built around that we have 118 datapoints for each child, so a child will be correlated to itself but have no other correlation. We also need a method, various methods were tried from ARAM to a symmetric correlation based on the mean, but AR1 (auto regressive order 1) structure was found to have the lowest AIC value implying best fit. Therefore, the MLR has the following form:

$$\log(\text{Aerosol}) = XB + e$$
$$e \sim N(0, \sigma^2 * B)$$
$$B = AR1(\text{Minute}|\text{ID})$$

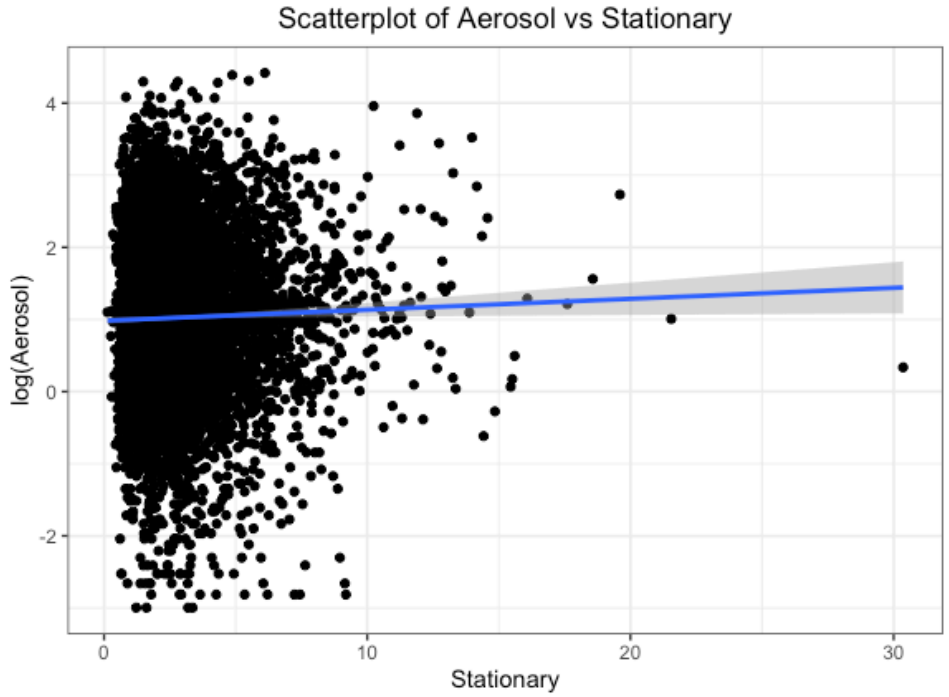
y or log(Aerosol) = the explanatory variable
X = the dependent variable matrix
B = the Beta coefficient that shows the effect of each variable from the X matrix
e = the residuals of our model
N = residuals are normal, mean at 0, variance of sigma^2 * B
AR1(Minute|ID) = our Betas are based off this function that finds the correlations by minute within each ID

This model requires that we meet some assumptions. First, that our data show linearity between independent and dependent variables. Second, that each data point is independent of another (has no effect). Thirdly, that our residuals are normal, showing normality of the data

based on the model. Finally, our data must show equal variance where no sections of the data spreads out or in too much.

Section 3: Model Validation

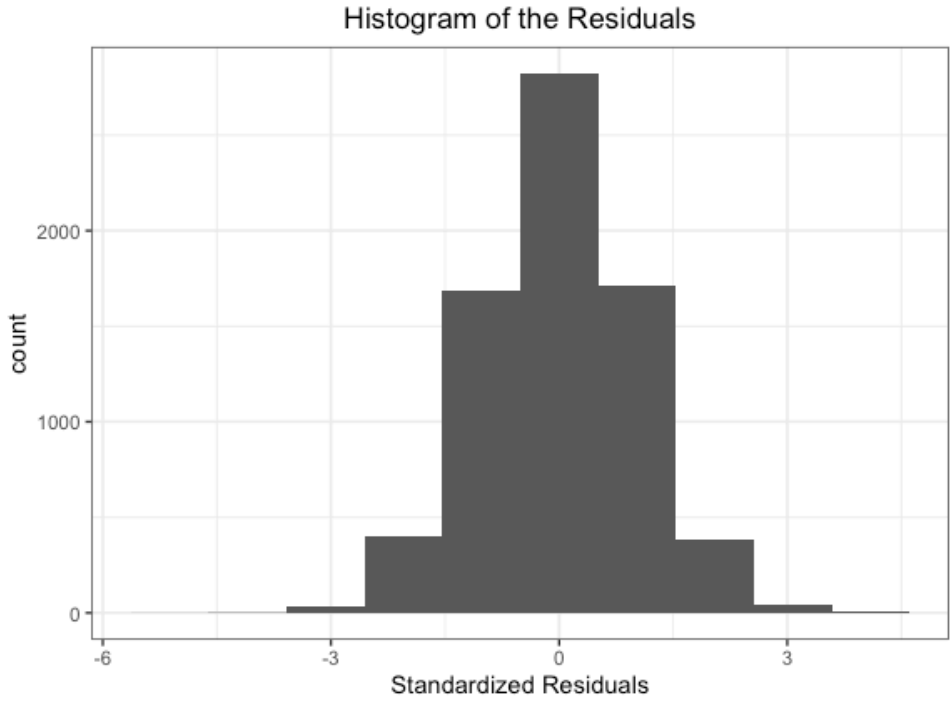
Now we must justify each of the assumptions stated above. Normality is easy enough, only one of the independent variables is used as quantitative variable and needs to be checked so the scatterplot again of Log(Aerosol) vs Stationary. This plot shows a linear relationship so we can accept that our Linearity assumption holds.



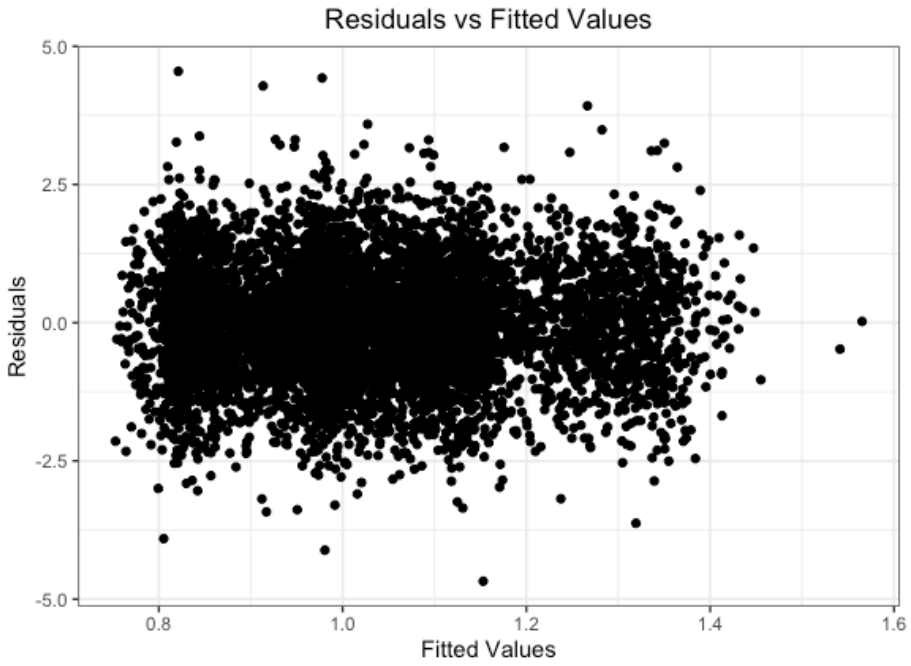
Next, our independence assumption that we built our model all around. First, the new model's correlations structure was calculated and the mean found based on the decorrelated residuals. Previously it had been a high correlation mean of .81 but now our model has a correlation mean of .01, clearly showing that our independence issues have been dealt with and out data now appears independent. We are still assuming that we have independence from child to child, but we believe that there was no effect from on child to another within this study.

Normality can be checked with a histogram of the standardized residuals from our model. To hold this assumption, was one of the main reasons we transformed out data

originally. The plot below shows just what we are looking for, a very normal looking curve with no outliers.



Equal Variance is proved with the plot below, a standardized residuals vs fitted values scatter plot, showing that our dataset shows no drastic changes in variance across all the data points.



Some other validation is needed to show how well our model fits the data. The Root Mean Squared Error RMSE was calculated and found to be 2.95. This means that on average our model is 2.95 PM reading values off from any single point. Based upon Aerosol having a standard deviation of 6.54 and a range of about to 83, we believe this shows our model is fitting the dataset quite well and much better than using summary statics such as standard deviation, mean, and range. A pseudo R-squared based upon the correlation of the predictions and the actuals was found to be .99, an almost perfect score for R-squared, although it is only a pseudo number.

Section 4: Analysis Results

Based upon the success of our model fitting the data, we are now ready to answer the research questions. The first question was: Is the stationary measurement, alone, a good predictor of actual PM exposure (aerosol)? In order to test this question a general linear hypothesis test was created to test the difference. We found that stationary measurement has a significant effect on log(Aerosol PM) based on a p-value of .02 and the null hypothesis being that Stationary PM has no effect on log(Aerosol PM). The true effect of Stationary PM was found to be between 1.01 and 1.02 on (after transforming it back), which shows what we would expect, a 1 to 1 change.

The second research question asks: Do activities change the pollution exposure? If so, which ones? To test this question, we looked at the effects of all the activities and found that all were significant except *Playing on Furniture* and *Walking*. The chart below shows all the Beta coefficients (specifically the activities for this question) from out model and how they affect log(Aerosol) with a 95% confidence interval. *Playing on the Floor* is noted to have the greatest effect on increasing PM levels.

Chart of all the Beta Effects on log(Aerosol)			
Betas	lower	est.	upper
(Intercept)	0.51419763	0.79386949	1.07354134
Stationary	0.01115037	0.01472975	0.01830913
ActivityHomework	0.03127408	0.23594519	0.4406163
ActivityOnPhone	0.09224672	0.29176137	0.49127601

ActivityPlayingOnFloor	0.3282747	0.53679058	0.74530645
ActivityPlayingOnFurniture	-0.2474766	-0.054869	0.13773854
ActivityVideoGames	0.14276709	0.34520919	0.54765128
ActivityWalking	-0.0490189	0.15917053	0.36735995
ActivityWatchingTV	0.21712721	0.41681258	0.61649795
ActivityComputer:ID	-0.0078449	0.00010782	0.00806058
ActivityHomework:ID	-0.0064138	0.00159206	0.00959796
ActivityOnPhone:ID	-0.0114411	-0.0034608	0.00451959
ActivityPlayingOnFloor:ID	-0.0102143	-0.0021799	0.00585441
ActivityPlayingOnFurniture:ID	-0.0024051	0.00557004	0.01354519
ActivityVideoGames:ID	-0.0133657	-0.0053628	0.00264014
ActivityWalking:ID	-0.0082121	-0.0001915	0.00782911
ActivityWatchingTV:ID	-0.0157652	-0.0078249	0.00011536

Last research question: Do the activities/stationary have different effects on PM exposure for different children? We can again use the chart above to see the interaction affect between each activity and ID. None were found to be significant so no, activities/stationary do not have different PM exposure for different children.

Section 5: Conclusions

The analysis today looked at a dataset analyzing PM levels on children. We found that PM levels can be effectively measure by a stationary reader. We also found that what activities a person engages in significantly affects their PM level intake and that it doesn't matter who you are.

This analysis can be continued or expanded by adding in a couple of important independent variables that could help explain PM levels such as the child's location and if their parent smokes or not. This would allow our model to be more accurate and possible predict PM levels for a child based upon a set of circumstances. Another way to better understand, would be to run the whole analysis again with a new set of children and compare the results.