PM Exposure in Children

March 2019

Introduction and Problem Background

Pollution is a concern to environmentalists and the general public everywhere. One form of pollution is particulate matter (PM), which is a mixture of particles that can be found in the air. PM comes from power plants, automobiles, and other sources. Some PM particles are visible, while others are not. Negative health effects associated with PM include nonfatal heart attacks, irregular heartbeat, worsened asthma, decreased lung function, coughing, difficulty breathing, irritation of airways, and even premature death in people with heart or lung disease. Children, older adults, and those with heart and lung conditions are particularly susceptible to the negative effects of PM. In this analysis, we are interested in investigating how PM exposure is measured for children. Currently, scientists frequently measure atmospheric PM to assess exposure. Unfortunately, this could vary widely from true exposure when children spend time indoors.

To investigate how to more accurately assess PM exposure in children, researchers conducted a study where 60 children were fit with vests with PM sensors. In addition to the vest-mounted sensor, a central sensor was placed in the child's home. The child was then monitored for a period of two hours. Every minute, the PM measured at the vest (Aerosol), the PM measured at the central station (Stationary), and the current activity of the child (Activity) were recorded. The following table summarizes the variables, their descriptions, and corresponding summary statistics.

Variable	**Description**	**Mean**	**Min**	**Max**
ID	Child ID number	NA	1	60
Aerosol	The PM measurement on the child's vest	4.95	0.05	82.91
Stationary	The PM measurement of the stationary monitor	2.96	0.13	30.36
Activity	The activity the child was engaged in	NA	NA	NA
Minute	The minute the child was wearing the vest	NA	1	118

Figure 1 contains a scatterplot showing the relationship between Aerosol and Stationary. It is rather difficult to see a clear relationship between the two variables. A transformation may help clarify the relationship. In fact, to justify our assumptions later, a log transformation on Aerosol will be necessary. Ignoring the issue and not performing the log transformation will cause several assumptions necessary for our model to be unjustified. Normality and equal variance will each suffer, which means our hypothesis tests regarding the effects of different variables will no longer be valid. We will ultimately see that Stationary alone is a very poor indicator of Aerosol. That is, without additional information, Stationary does not tell us very much about Aerosol.

Figure 1 also contains a series of boxplots showing the relationship between activity type and log(Aerosol). Activity does not appear to have a huge effect on log(Aerosol), but there are certainly differences among the activities. For example, playing on the floor and playing on furniture seems to have a slight association with increased log(Aerosol).

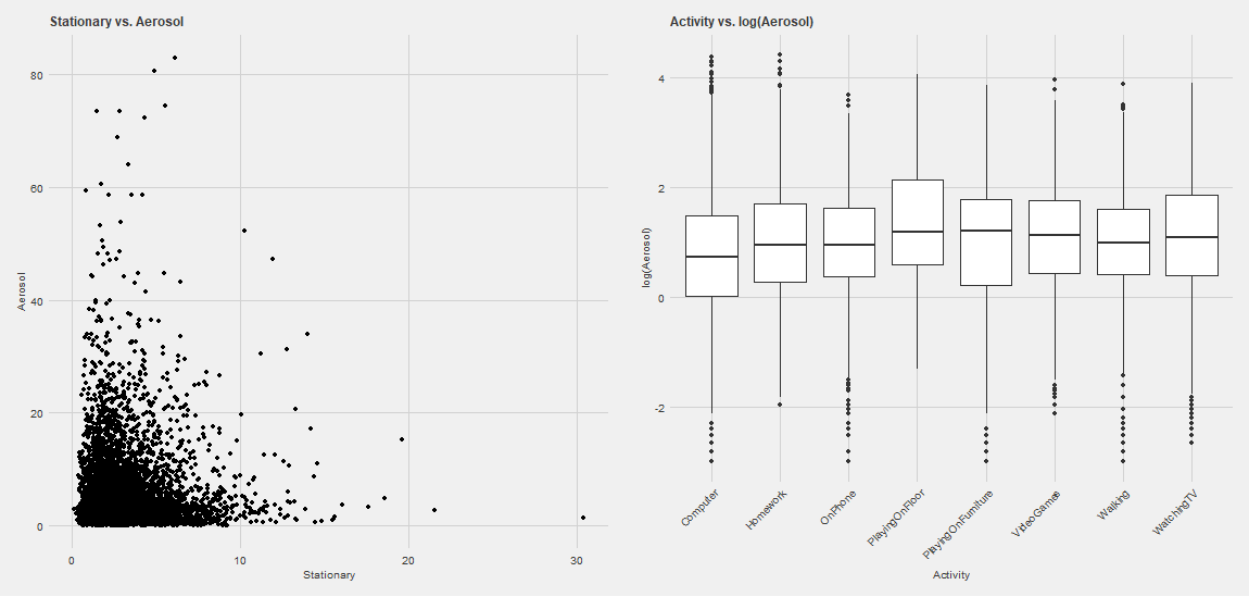

Figure 1: Exploratory Graphics

There are few other nuances and complications with our data. First, our observations are correlated in time. Each child was recorded for 118 minutes. A subset (the first five minutes) of the correlation matrix between observations of the same child is shown below. We can see that there is high correlation from minute to minute. For example, the correlation between observations at minute 3 and observations at minute 4 is 0.9513. Similar high amounts of correlation can be seen between observations for all 118 minutes. Our model will need to take this correlation into account. If correlation is not properly accounted for, our standard errors (and consequently our confidence and prediction intervals) will be incorrect.

	1st Minute	2nd Minute	3rd Minute	4th Minute	5th Minute	$\cdots$
1st Minute	1.0000	0.9517	0.8784	0.8794	0.6880	$\cdots$
2nd Minute	0.9517	1.0000	0.9199	0.9049	0.7797	$\cdots$
3rd Minute	0.8784	0.9199	1.0000	0.9513	0.8615	$\cdots$
4th Minute	0.8794	0.9049	0.9513	1.0000	0.8453	$\cdots$
5th Minute	0.6880	0.7797	0.8615	0.8453	1.0000	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

In addition to the correlation present in the data, another complication with regards to our data is the likely presence of interactions between our explanatory variables. For example, it is possible that the effect of certain activities depends on which house the child is in. Further along in the analysis, we will see that interactions are present in our data. They will need to be accounted for in our model. If we do not account for interactions, much of our model will be incorrect, including our estimates for the effects of the explanatory variables.

In this analysis, we will use a regression model with interaction terms and a AR(1) correlation structure. This model will allow us to see what truly affects log(Aerosol). It will account for the interactions between variables and will help account for the correlation present in the data.

Using our model, we hope to accomplish three primary research goals. We first want to determine whether the stationary measurement alone is a good predictor of actual PM exposure. We also want to find out if activities change pollution exposure. If so, we are interested in determining which specific activities do so. Our last research goal is to determine whether activities and stationary have different effects on PM exposure for different children. Our model will help us receive answers to these questions.

Statistical Model

In order to account for correlation and to properly answer our research questions, we fit the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{B})$$

where $\boldsymbol{B}$ is a block diagonal matrix with blocks $\boldsymbol{R}$ that follow an 118×118 AR(1) structure. Thus,

$$\boldsymbol{R} = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{117} \\ \phi & 1 & \phi & \vdots & \phi^{116} \\ \vdots & \vdots & \ddots & & \vdots \\ \phi^{117} & \cdots & & & 1 \end{pmatrix}$$

$\boldsymbol{Y}$ is a vector of the response variable (log(Aerosol)). $\boldsymbol{X}$ is a design matrix containing information from our explanatory variables, including the necessary interaction terms. The first column is a column of 1's, corresponding to the intercept. The next columns include linear terms for the explanatory variables (Stationary, Activity, ID) as well as the interactions (Activity $\times$ ID and Stationary $\times$ ID). $\boldsymbol{\beta}$ is a vector of coefficients. The first element of $\boldsymbol{\beta}$ is the intercept. The rest of the elements of $\boldsymbol{\beta}$ correspond to the effects of the explanatory variables and interactions. σ^2 is a measure of the overall variance of the model.

Because our observations come from a time series, the residuals of observations within an individual (the 118 observations corresponding to a single child) will be correlated. To account for this correlation, we determine that they follow an auto-regressive process of order 1. Essentially, this means that residuals that are close in time will be highly correlated with each other. As residuals go farther apart, this correlation will lessen. We can further describe this relationship as:

$$\epsilon_t = \phi \epsilon_{t-1} + \omega_t$$

where

$$\omega_t \sim \mathcal{N}(0, \sigma^2)$$

ϵ_t is the residual for each observation, while ϵ_{t-1} is the residual for the previous observation. ϕ is the correlation (or effect) of the previous observation. ω_t is random noise.

Essentially, observations within an individual (an individual child) are correlated, but observations between individuals (different children) are independent. Using a block matrix helps to formally define this.

Using this model will help us answer our research questions and will properly account for the correlation and interactions present in the data. By building correlation into the structure of the model itself, our standard errors and intervals will be more accurate. This model will allow us to determine the effects of the various explanatory variables on log(Aerosol). We will be able to determine if the stationary measurement alone is a good predictor of actual PM exposure (as measured by log(Aerosol)). We will also be able to determine if certain activities change pollution exposure and quantify those changes. Finally, we will be able to use our model to determine if activities/stationary have different effects on log(Aerosol) for different children by looking at the interaction effects.

To properly use this model, several assumptions will need to be justified. Our model should properly account for the correlation in the data. After accounting for correlation, residuals should be normally distributed and a Residuals vs. Fitted Values plot should show equal variance about the line. We will also need to check if there are linear relationships between log(Aerosol) and our quantitative explanatory variables. We will verify these assumptions in the following section.

Model Validation

We now verify the assumptions necessary for our model to work properly. In Figure 2, we can see added variable plots. We included the added variable plot for the stationary variable, as well as a small selection of the interaction terms (Stationary $\times$ ID). We can see linearity between our explanatory variables and the response, so our assumption of linearity holds.

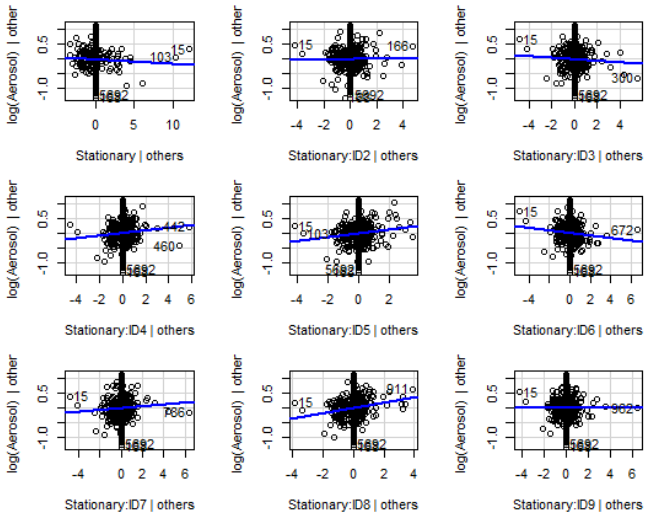

Figure 2: Added Variable Plots

In Figure 3 below, we can see a histogram of standardized (and decorrelated) residuals as well as a fitted values vs. residuals plot. The histogram shows us that the residuals are normally distributed, so our assumption of normality holds. The plot of fitted values vs. residuals shows us that the spread of residuals about the fitted line is fairly even as the fitted values increase. Thus, our assumption of equal variance holds.

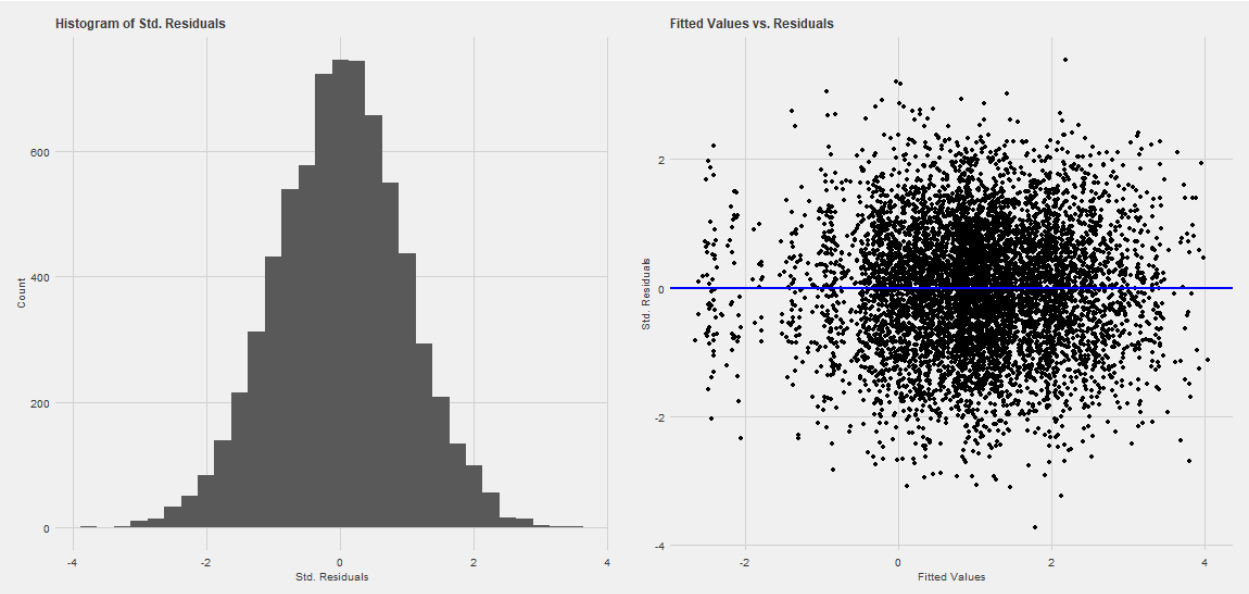

Figure 3: Model Verification Plots

We cannot assume true independence in our model because of the correlation present. However, by

accounting for the correlation through the AR(1) correlation structure, we can assume independence.

To measure how well our model fits the data, we calculated a value for pseudo-R^2 using the following formula.

$$R^2_{pseudo} = (cor(Y_{obs}, Y_{pred}))^2$$

Using this formula, we calculate R^2_{pseudo} to be 0.9142. Roughly, this means that approximately 91.42% of the variation in log(Aerosol) is explained by our model. This is pretty good.

Our model assumptions hold and our model fits the data well. We are now ready to continue with our analysis and answer our research questions.

Analysis Results

Our first research question concerned whether stationary measurement alone was a good predictor of actual PM exposure. We find that it is not. We show below a subset of the coefficient table from our model for the primary and interaction effects for stationary measurement.

Effect	**Estimate**	**Standard Error**	**Test Statistic**	**p-Value**
Stationary	-0.015	0.012	-1.263	0.207
Stationary:factor(ID)2	0.011	0.019	0.573	0.567
Stationary:factor(ID)3	-0.012	0.018	-0.689	0.491
Stationary:factor(ID)4	0.049	0.020	2.537	0.011
Stationary:factor(ID)5	0.068	0.019	3.488	0.000
Stationary:factor(ID)6	-0.038	0.018	-2.123	0.034
⋮	⋮	⋮	⋮	⋮
Stationary:factor(ID)60	-0.016	0.021	-0.764	0.445

As we can see in the table above, the p-value for the primary effect of Stationary on log(Aerosol) is 0.207, which is greater than our chosen alpha of 0.05. This means that, when controlling for the other factors in our model, stationary does not have a statistically significant relationship with our response variable. We need more information in order to accurately predict log(Aerosol). We note that many of the interaction terms between Stationary and ID are statistically significant. We discuss this further later in the analysis.

Another way that we evaluate this question is by comparing our full model to an otherwise identical model which used only Stationary as a single explanatory variable to predict log(Aerosol). We find that the RMSE for the reduced model is 3.4 times greater than the RMSE for the full model. In other words, a model with stationary measurement alone has considerably larger errors than our full model which included other terms. In addition, the reduced model only had a R^2_{pseudo} of less than 0.01. Finally, we compare this reduced model to our full model using an ANOVA test. The p-value from this test is less than 0.0001. Taken together, this evidence leads us to conclude that Stationary alone does not effectively explain log(Aerosol).

The second research question relates to the activity groups used to categorize the children's behavior. We find that, holding individuals and time constant, only one of the activities had a statistically significant relationship with log(Aerosol). We show below a subset of the coefficient table from our model for the primary effects for activity.

Effect	Estimate	Standard Error	Test Statistic	p-Value
ActivityHomework	-0.012	0.179	-0.070	0.944
ActivityOnPhone	-0.227	0.156	-1.456	0.145
ActivityPlayingOnFloor	-0.075	0.166	-0.454	0.650
ActivityPlayingOnFurniture	-0.401	0.184	-2.181	0.029
ActivityVideoGames	0.107	0.185	0.580	0.562
ActivityWalking	0.097	0.201	0.484	0.628
ActivityWatchingTV	0.030	0.167	0.177	0.860

On average, when other variables are included in our model, different levels of Activity do not explain changes in log(Aerosol). The one exception to this finding is PlayingOnFurniture. Holding individual, stationary, time, and interaction effects constant, PlayingOnFurniture corresponded with log(Aerosol) 0.401 units lower than the reference level, Computer, with a 95% confidence interval from -0.762 to -0.041. Given this information alone, it may seem that Activity is not generally a very important variable for understanding PM exposure for children. To investigate that idea, we construct a model that is identical to our full model but which includes only Stationary, ID, and Stationary:ID as explanatory variables. We conduct an ANOVA test on our full model and this reduced model and find a significant difference at $\alpha = 0.0001$. Clearly, we need to dig deeper to understand how different activities make an impact. We show below, taken from the coefficient table from our full model, some of the most noteworthy interaction effects for Activity.

Effect	Estimate	Standard Error	Test Statistic	p-Value
ActivityHomework:factor(ID)22	2.012	0.253	7.946	0.000
ActivityHomework:factor(ID)5	1.859	0.235	7.897	0.000
ActivityPlayingOnFurniture:factor(ID)49	1.820	0.266	6.836	0.000
ActivityPlayingOnFloor:factor(ID)7	1.755	0.259	6.771	0.000
ActivityHomework:factor(ID)16	1.694	0.245	6.922	0.000
⋮	⋮	⋮	⋮	⋮
ActivityWalking:factor(ID)2	-1.309	0.287	-4.567	0.000
ActivityPlayingOnFloor:factor(ID)59	-1.392	0.228	-6.110	0.000
ActivityHomework:factor(ID)27	-1.423	0.261	-5.447	0.000

We find that there are dozens of statistically significant interaction terms for ID and Activity. This leads us to the conclusion that the conditions in individual homes are highly important to determining whether a given activity corresponds with higher PM exposure. For example, when holding other primary and interaction effects constant, Aerosol measurements were $e^{2.012} = 7.478$ units higher for ID22 while he or she did homework relative to other individuals doing homework, with a 95% confidence interval from $e^{1.516} = 4.554$ to $e^{2.509} = 12.293$. This indicates that ID22 likely did homework in a part of his or her home such that he or she was exposed to additional PM pollution in a way that was unique to him or her.

We evaluate the question regarding which activities correspond with higher log(Aerosol) levels by examining the 25 largest positive interaction effects for Activity and ID. We find that of those effects, 11 were for PlayingOnFloor, 6 were for PlayingOnFurniture, 4 were for Homework, 2 were for OnPhone, 1 was for WatchingTV, 1 was for VideoGames, and 0 were for Walking. We conclude informally that while PlayingOnFloor and PlayingOnFurniture do not have statistically significant positive effects on average when controlling for other variables, they seem to be more commonly associated with higher PM levels in individual homes relative to other activities. In other words, individual homes tend to have conditions such that PlayingOnFloor exposes children to higher PM levels more often than other activities, even if there isn't a significant effect for everyone on average.

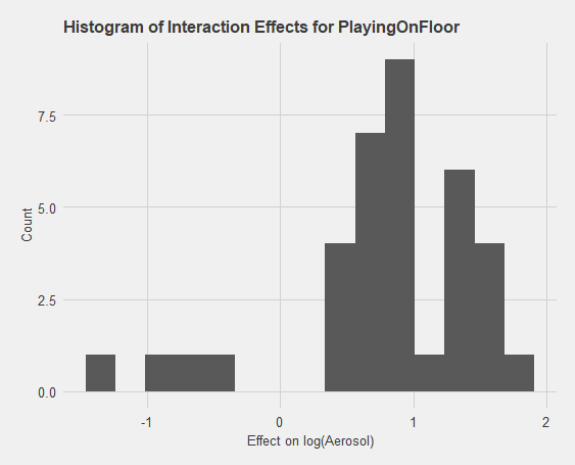

Figure 4: Histogram of Int. Effects

Finally, we consider the research question regarding whether the effects of Activity and Stationary are child-specific. In short, the effects of Activity and Stationary are child-specific. We find that of the 472 interaction terms in our model, 208 are statistically significant. The conditions of individual homes clearly have a large impact on whether Stationary or different levels of Activity have an effect (and what effect they have) on log(Aerosol). The largest positive statistically significant interaction effect in our model was ActivityHomework:factor(ID)22, which was referenced anecdotally earlier. The largest negative statistically significant interaction term in our model was ActivityHomework:factor(ID)27. The range for interaction effects is (-1.423, 2.012). This means that activities in different homes had average effects resulting in increases in log(Aerosol) by as much as 2.012 or decreases by as much as 1.423. As an example, a histogram of the statistically significant interaction effects between PlayingOnFloor and ID is shown in Figure 4. We can clearly see that for a number of homes, the effect of playing on the floor is negative after holding other factors constant, but for a majority of homes in which PlayingOnFloor was significant, the effect was positive. Like the other Activity levels in our model, the direction of the PlayingOnFloor effect is child-specific.

Conclusions

In our analysis, we were able to use a regression model with interaction terms and an AR(1) correlation structure to answer the research questions posed in the introduction. We discovered that Stationary alone does not explain log(Aerosol) well. Other information is needed. By including interaction terms between ID and Activity, as well as ID and Stationary, we were able to quantify the differences in log(Aerosol) between houses. We discovered that, on average, only one of the activities, playing on the furniture, had a statistically significant effect on log(Aerosol). By analyzing some of the interaction effects, however, we were also able to learn that playing on the floor and on furniture increases the amount of PM exposure for a great number of children. Overall, we learned that PM exposure is highly individualized. It is difficult to make blanket statements when each house is so different. However, by accounting for the differences between houses, our model works fairly effectively at explaining different levels of PM exposure.

As research into PM exposure continues, it would be helpful to have a greater number of subjects. It would also be helpful to collect more variables, such as distance from the stationary monitor or what room a child is in. These additional variables could be helpful in making our model more effective.