Particulate Matter Exposure Project

Naomi Zubeldia & Jacob Johnson

2023-03-13

Section 0: Executive Summary

Air pollution causes health issues all over the world. In this analysis, we will be investigating the effectiveness of current (stationary) air pollution monitors and if the activities that a person is engaged with has an effect on pollution levels experienced. We used a model that included how pollution levels around a specific person change over time. We found that using the activity a person is engaged in, we were able to better model and predict pollution levels than by simply using the traditional stationary measurement. The exact effect of each activity varies from person to person, but on average, playing on the floor and watching TV were the activities that were related to the highest increase in pollution.

Section 1: Introduction and Problem Background

Air pollution is a problem in our modern society due to the increase of chemicals being emitted into the air by power plants, industries, and automobiles. In some areas of the world, the air can be so bad that it is advised not to go outside. Scientists measure the air quality through quantifying the amount of particulate matter (PM), which is a mixture of particles that can be found in the air and are only perceptible by microscope. People who have lung or heart diseases, elderly, or children are most affected by PM.

The main goal of this analysis is to understand the amount of PM exposure that children have, which is difficult due to varying times spent outdoors or in other activities. To do this, we want to know if a stationary measurement of where a child lives does a good job at explaining the PM exposure. Since children have a lot of activities indoors, we also wanted to test if these activities explain more of the PM exposure that children have rather than a stationary measurement. With the previous two objectives in mind, we wanted to make sure that the effects of activities/stationary measurements are child specific and find any variability of that. Lastly, we want to learn what activities lead to higher PM exposure on average.

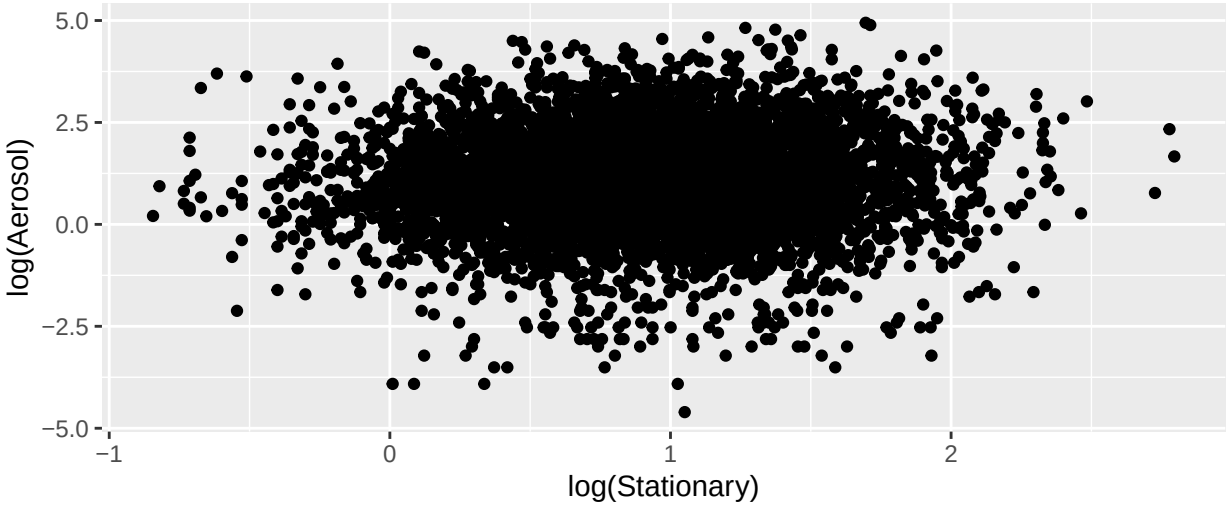
To accomplish these goals, we will be analyzing data from 100 children who wore vests with pollution monitors on the shoulder for about one hour. Each child was also given a GoPro to observe which activities they were engaged in. For the stationary measurement, a PM monitor was installed at each home to measure the atmospheric PM. So, the variable we are observing is the PM measurement from each child's vest, which is labeled as Aerosol. The variables that we are comparing to see which is most influential on PM levels are: the stationary measurement, the activities that child engaged in, minute the observation was taken, and the child specific ID. To get a quick summary of the data, we included the summary statistics of the data in Table 1. below.

Table 1: Summary Table

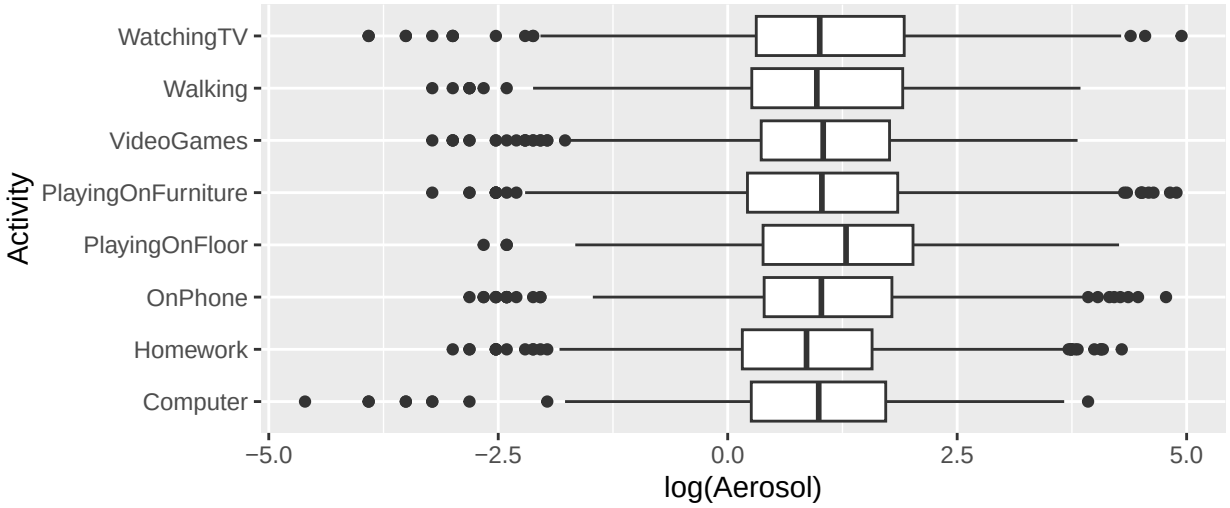
ID	Stationary	Activity	Minute	Aerosol
1 : 59	Min. : 0.430	VideoGames : 749	Min. : 1	Min. : 0.010
2 : 59	1st Qu.: 1.750	WatchingTV : 748	1st Qu.:15	1st Qu.: 1.350
3 : 59	Median : 2.490	OnPhone : 744	Median :30	Median : 2.750
4 : 59	Mean : 2.824	PlayingOnFurniture: 743	Mean :30	Mean : 5.725
5 : 59	3rd Qu.: 3.540	Walking : 741	3rd Qu.:45	3rd Qu.: 6.100
6 : 59	Max. :16.370	PlayingOnFloor : 739	Max. :59	Max. :140.410
(Other):5546	NA	(Other) :1436	NA	NA

It appears from the observed variable of aerosol that the PM vest measurements have a huge range (larger than 100) yet the median is quite small, meaning that most of the children were not exposed to high levels of PM. Since the max was so large, we chose to examine the aerosol measurements more closely in a histogram and saw that they were pretty skewed. We determined that in order to have normality in our observations we would log transform the aerosol response. The stationary measurements of PM levels have a smaller range will a low median, so we also decided to log the stationary measurement. The other variables are categorical, so we did not transform them.

In the figure below, we created a scatter plot to observe the relationship between the two numerical variables of log(aerosol) and log(stationary) PM measurements. The two appear to have a weak positive linear relationship, meaning that on average, as log(stationary) increases, so does log(aerosol). This is quantified by a correlation value of 0.0442774, which is quite close to 0, showing once again how weak the relationship is.

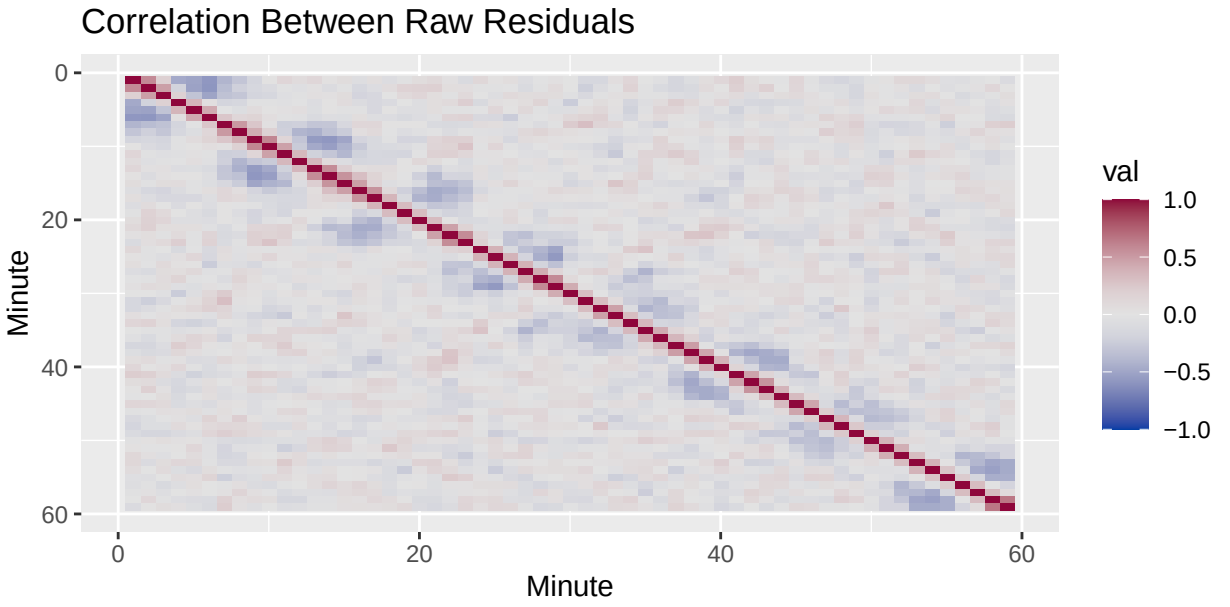


In order to explore the different activities that the children did, so we created the boxplots below.



Overall, the box plots all look similar between activities, which is interesting because the activities can be quite different and we expect the PM measurements to be different for different indoor or outdoor conditions. Even in these box plots, we can see that there are quite a few outliers or log(aerosol) measurements that differ from the majority of the other measurements. There are more with lesser levels of PM but still a few on the higher end.

To further explore the data, we first fit a simple linear regression model to the data. We were aware that there were multiple observations per child and wanted to check how that would affect the correlation in the linear model. We created a correlation matrix, which is visualized below, using the residuals of the model. It appears that there is a lot of correlation between the observations that were taken close to each other and some negative correlation with activities taken further apart, which means this basic model, which assumes independence between observations, is not valid.



If we go forward with a model that does not account for the correlation in the observations, we risk making incorrect conclusions about the data and not answering the questions we are trying to answer. We risk the consequences of incorrect confidence intervals and p-values for all of the tests we do, as well as the results being biased. If we account for the correlation, then we will be able to run a regression analysis on the data to find what influences the PM levels and have correct measures of the uncertainty of those effects.

Since we want to conduct this analysis correctly, we will be using a linear model that accounts for correlation between observations called a Longitudinal Multiple Linear Regression Model. It follows a normal distribution with a mean determined by a combination of the log(stationary) measurement, activity, and child ID. How these explanatory variables are combined is determined by β coefficients which are estimated from our data. The variance of this model is where we will account for the correlation between observations. The variance is made of a B matrix that has smaller R matrices on the diagonal that contain the correlation between measurements from each child called a block, which has ones along the diagonal so that we assume that each child is independent of the others. This structure accounts for both the correlation for measurements by the same child and independence between children, allowing us to make inference.

Section 2: Statistical Model

The specifics of the model that we will be using is detailed below. This model was selected after comparing many possible models using the AIC information criteria. We considered multiple different models that each had a unique way to incorporate the correlation between observations for the same child. We tried an AR(1) structure, which captures slowly decaying temporal correlation, an MA(1) structure, which captures quickly decaying temporal correlation, and an ARMA(1,1) structure that combines both quickly and slowly decaying correlation. We did not consider a general symmetric correlation structure (which allows for a unique coefficient describing the pairwise correlation between every measurement over time) because estimating that type of structure was computationally impractical with a data set this large.

After fitting all 3 of the models listed above, we chose the model that gave us the lowest AIC value, meaning that it fit the data best. From the AIC comparisons, we found that a ARMA(1,1) model that captures both quickly decaying and slowly decaying temporal correlation was best. We will use this correlation structure to represent the correlation between the log(aerosol) measurements for the same child across time. The full model, using this method, is shown and explained below.

$$log(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{B})$$
$$B = \text{diag}(\boldsymbol{R}, ..., \boldsymbol{R})$$
$$\hat{\epsilon}_{j,T+1} \sim \text{ARMA}(p = 1, q = 1)$$

$\boldsymbol{y}$: A vector of the aerosol levels that we are modeling. We are modeling the log(aerosol) for each child at each minute.

$\boldsymbol{X}$: A matrix of our explanatory variables. We have a column of 1's to allow for an intercept as well as a column for each of the following: log(stationary) measurement, activity, and Child ID. We are also allowing for an interaction between these explanatory variables and Child ID, meaning that the effect of log(stationary) and activity can be child specific.

$\boldsymbol{\beta}$: A vector of coefficients representing the effect of each of our explanatory variables from $\boldsymbol{X}$ on log(aerosol). For example, the coefficient for Watching TV represents the expected change in log(aerosol) when, holding all else constant, a child goes from using the computer to watching TV.

$\boldsymbol{\epsilon}$: A vector of errors. This represents the list of all the differences between our modeled or expected log(aerosol) value and the true recorded log(aerosol) value. These errors are distributed as a draw from a multivariate normal distribution with mean 0, meaning that we expect that on average, our modeled values are correct. The covariance matrix that determines how these errors vary is described below.

$\hat{\epsilon}_{j,T+1} \sim \text{ARMA}(p = 1, q = 1)$: We are modeling that the log(aerosol) measurements for the j^{th} specific child will be related to the other measurements taken across time for that same child. Specifically, we are saying that the errors across time are correlated using a combination of an AR(1) (which captures the slowly decaying temporal correlation) and MA(1) (which captures the quickly decaying temporal correlation) correlation structure.

σ^2: The overall variance of our model. This represents the variability around our modeled log(aerosol) value compared to the log of the true recorded aerosol value. σ, the square-root of this variance, represents how far, on average, our modeled log(aerosol) value is from the recorded value.

$\boldsymbol{B}$: A correlation matrix that allows us to model correlation between the log(aerosol) observations from the same child and also assume independence between children. This matrix has a block design with the 59×59 $\boldsymbol{R}$ matrix on the diagonal. This R Matrix is explained below.

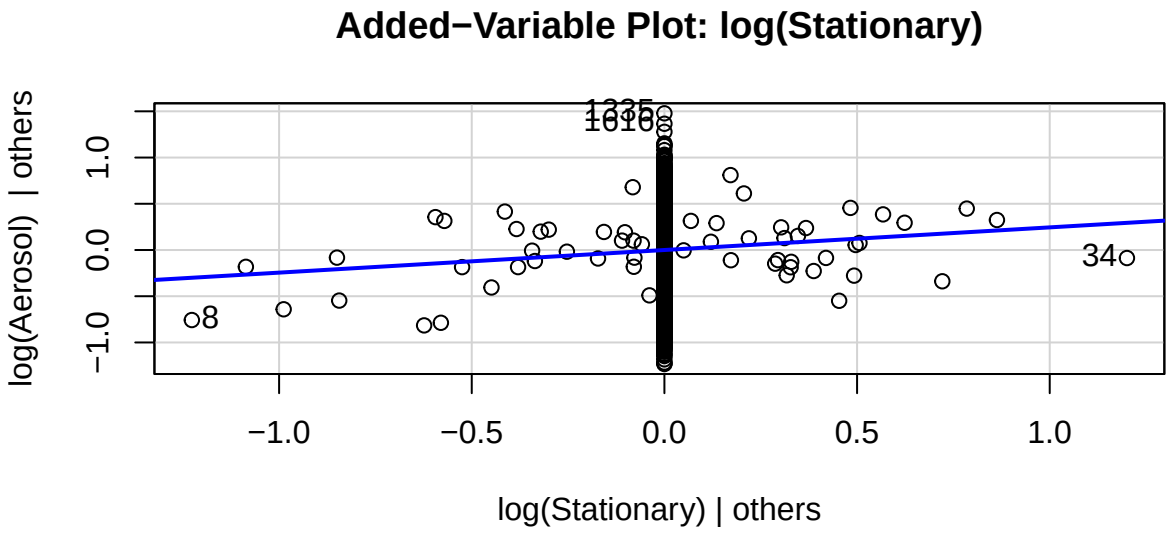
$\boldsymbol{R}$: A matrix that is repeated down the diagonal of the covariance matrix. This matrix shows us how correlated each measurement is across time for each child. This is where the ARMA($p = 1, q = 1$) correlation structure is incorporated into our model. There are 1's along the diagonal, meaning that the measurements from one child are considered independent from the measurements of a different child.

The longitudinal model detailed above will be very useful for us as we answer the questions of interest in this analysis. We can use the coefficient that represents the effect of the log(stationary) aerosol measure to understand how the stationary measurement explains the actual aerosol values. We can also see the effect of activity by testing all of the activity coefficients, and we can see if these coefficients are child-specific or not. Lastly, we can look at the individual activities to see which has the highest increase in PM exposure.

The model detailed above requires us to make four assumptions. These assumptions are: linearity (assuming that the relationship between each explanatory variable and log(aerosol) level is linear), independence (assuming that we capture all of the correlation between observations), normality (our residual errors are normally distributed), and equal variance (there should be equal variance in our standardized residuals, no matter what the fitted value is). In order to prove the accuracy of our model, we test these assumptions below.

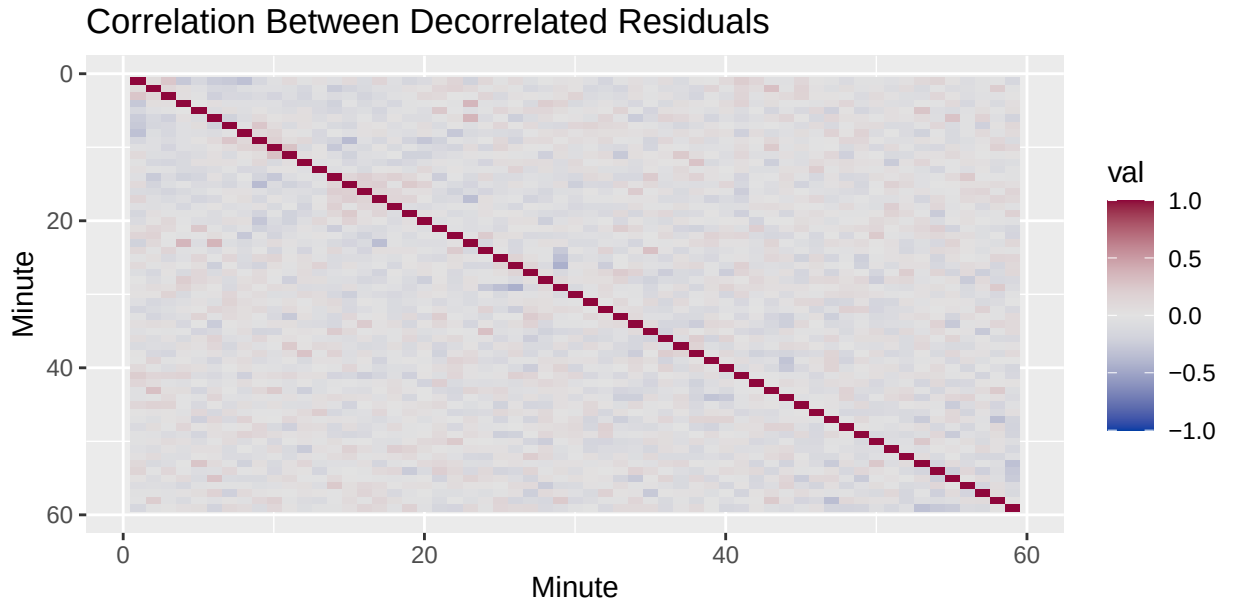
Section 3: Model Validation

After fitting the longitudinal model with the best correlation parameters, we need to check the assumptions listed above to make sure that our model is valid. For the linearity assumption, we created the added variable plot below and found that the relationship between log(aerosol) and log(Stationary) (which is the only numeric/continuous explanatory variable) is linear, showing that the model meets the linearity assumption.

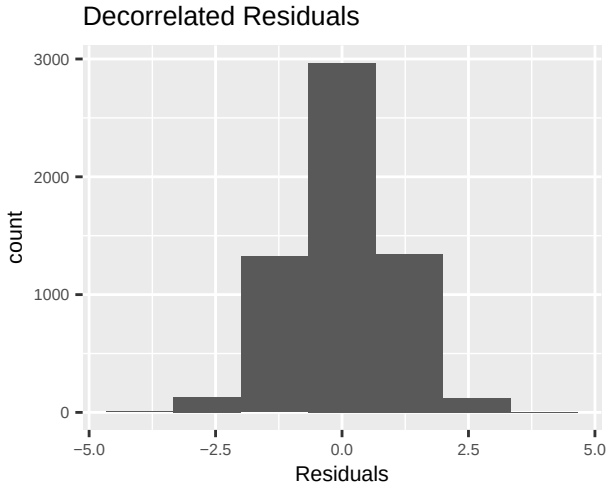


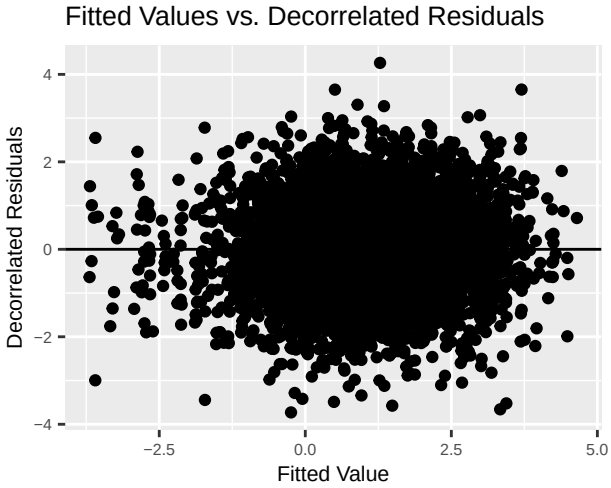
To make sure that the model correctly accounts for all correlation between the observations, we made a matrix of the standardized residuals. The matrix is very large, so we once again created a colored plot to

easily spot any correlation. From the plot below, we can see that there is little to no correlation compared to the plot we did earlier of the basic regression model, as evidenced by the lack of color apart from the diagonal line. With little to no correlation in the standardized residuals, the model meets the independence assumption, which allows for us to make valid inference from it.



For testing the normality assumption, we created a histogram of the standardized residuals below. The standardized residuals appear to follow a bell shape and be distributed normally, which meets what is needed for the normality assumption.





For the last assumption of equal variance, we plotted the fitted values of the model versus the standardized residuals. We can observe that the points appear to be randomly scattered without a pattern, meaning that we can safely make the equal variance assumption.

After making sure the model fits the assumptions, we tested if the model fits and explains the data well. We used an estimate of the statistical measure R^2 to explain how well our model fits the data. We got an R^2 value of 0.91514, which is very good. The model explains 92% of the variation in the output can be explained by the input variables of our model.

Now that we verified the validity of the model, we will use it to reach the goal of the analysis in understanding

how PM exposure influences kids through the different activities. We also hope to answer the research questions of learning what variables are important to keep in the model and how different children and activities influence PM exposure as well.

Section 4: Analysis Results

Now that we have a valid longitudinal model, we can answer our questions of interest. First, we can see how well the stationary PM measurement can represent the actual recorded PM aerosol level.

To do this, we fit a model (similar to the one detailed above) that only included a coefficient for log(stationary). This model has an R^2 value of 0.0018, meaning that only using the stationary measurement, we could only explain 0.18% of the variability in PM level. So, it appear that the stationary measure on its own is not a very good predictor of the true aerosol measure recorded on the vests that each child wore.

Next, we can test if including an effect for the activity that the child is participating in creates a better model. To do this, we create a model with both activity and stationary and compared it to the model using only stationary. The formal hypotheses for this comparison are detailed below.

H_0 : All $\beta_{Activity} = 0$: There is no effect for Activity

H_0 : Any $\beta_{Activity} \neq 0$: There is some effect for Activity

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## stat.gls     1  5 5758.473 5791.886 -2874.236
## act.gls      2 12 5679.022 5759.214 -2827.511 1 vs 2 93.45097  <.0001
```

Conducting this hypothesis test, we get a p-value of practically 0, which means that a difference in the performance of the two models this extreme is unexpected under our null hypothesis, so we reject the null hypothesis that there is no effect for activity and we conclude that Activity helps predict PM level. So using activities explains more than just the stationary measurement.

Next, we can see if the effect of each activity and the stationary measurement are child specific. To test this, we create a new model where we allow there to be an interaction between child ID and activity and child ID and log(stationary). This will allow us to test if the effect of a given activity is different for a different child or if the effect of the stationary measurement is different for each child. The formal hypotheses for this test are detailed below.

H_0 : All $\beta_{Interaction} = 0$: There is no interaction

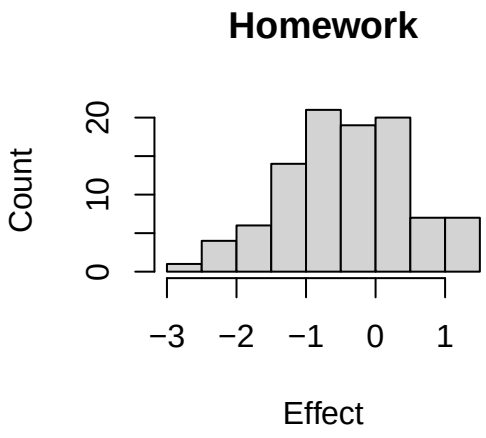
H_0 : Any $\beta_{Interaction} \neq 0$: There is an interaction

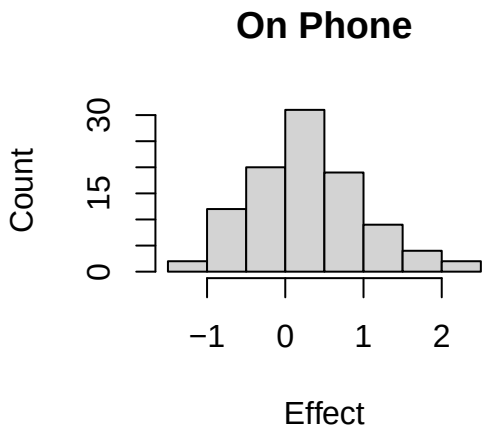
```
##          Model  df      AIC      BIC     logLik   Test  L.Ratio p-value
## stat.gls     1   5 5758.473 5791.886 -2874.2364
## pm.gls       2 903 3654.317 9688.802  -924.1586 1 vs 2 3900.156  <.0001
```

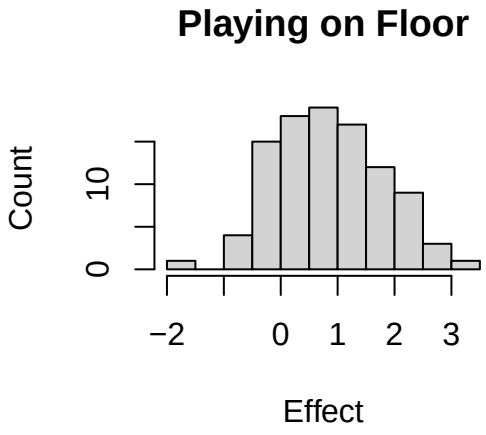
Conducting this test, we get a p-value of practically 0, meaning that at least one of the interaction coefficients is not 0. We conclude that the effect of log(stationary) and the effect of activity are child specific. We can also quantify the variability of these effects from child to child. The standard deviation of the log(stationary) effects across all children is 0.1575203, meaning that on average, the effect of log(stationary) on log(aerosol) for one child is 0.1575203 different from that same effect for another child. The standard deviation of the activity effects across all children and all activities is 0.9364833, meaning that on average, the effect of any activity on log(aerosol) for one child is 0.9364833 different from that same effect for another child.

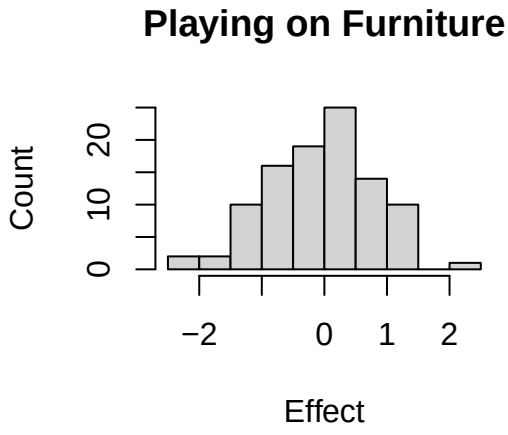
Even though we found that the effect of activity is child specific, we can still look at which activities, on average, lead to higher PM levels. Our model uses log(aerosol) as its response, so it is important to note

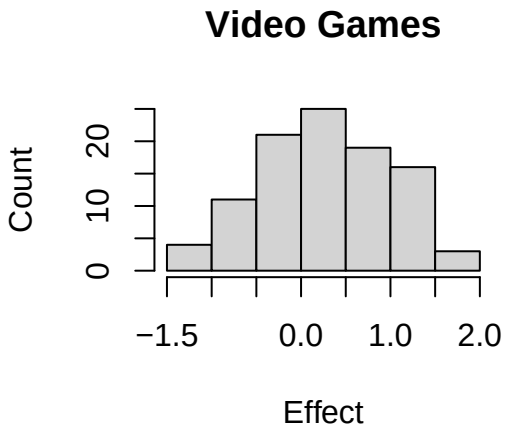
that an increase of log(aerosol) will also mean an increase in aerosol, meaning that we can easily interpret the coefficients in our model to see which activities lead to higher PM levels. Below we show a histogram of the coefficients for each activity. The numbers shown show the effect of each activity compared to the baseline activity of using the computer.

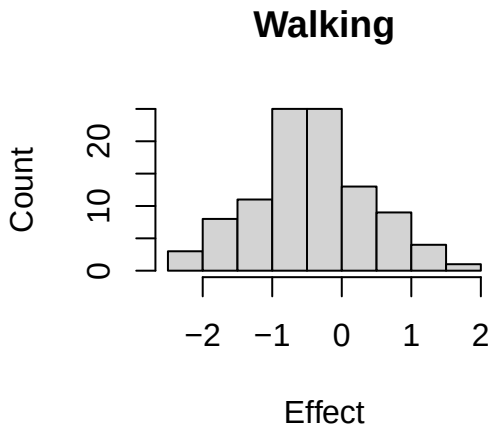












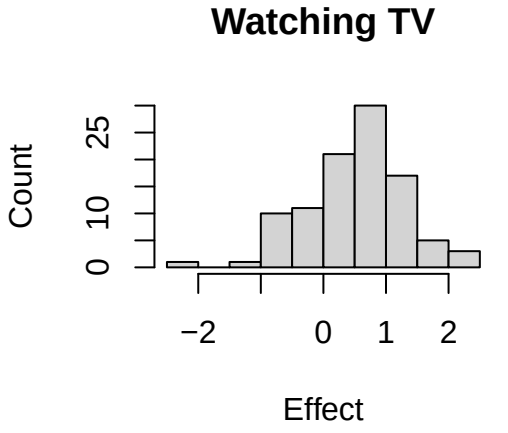


Table 2: Table 2: Average Activity Effects (compared to Computer)

Homework	Phone	Playing on Floor	Playing on Furniture	Video Games	Walking	Wathcing TV
-0.4019713	0.2936872	0.8617819	-0.0554213	0.2445836	-0.4352669	0.5182003

From the histograms and table above, we can see that being on the phone, playing on the floor, playing video games, and watching TV all led to higher PM levels than the baseline of computer. The activity that had the highest average effect on PM level was playing on the floor.

Section 5: Conclusion

From this analysis, it appears that just a stationary measurement of atmospheric PM does not truly represent the exposure that children have to PM due to the different activities they participate in. We were able to conclude that the stationary measurements did not explain the PM exposure of the 100 children participants, which indicated how important it is to take into account different activities that they do. Also, another variability that simulated PM levels could not account for is how exposure is different from child to child. Lastly, since we determined that PM levels depend on activity, we found that activities like playing on the floor, watching TV, or being on a phone have a higher effect on exposure.

The next steps researchers should consider for measuring air pollution in the home is what activities are usually done by the people. This will all depend on each person's age, socioeconomic background, and area trends. For children specifically, they should check if they are playing on the floor frequently because that activity leads to the most exposure. Another angle that they could study is why the floor has more PM exposure and what pollution particles tend to be there and where they come from.

Apendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
library(tidyverse)
library(kableExtra)
library(knitr)
library(nlme)
library(colorspace)
library(car)
pmdat <- read.table('BreathingZonePM.txt', header=TRUE) %>%
  select(ID,Stationary,Activity,Minute,Aerosol) %>%
  mutate(Activity=as.factor(Activity)) %>%
  mutate(ID=as.factor(ID))
knitr::kable(summary(pmdat), caption='Summary Table', format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options=c("HOLD_position","scale_down"))
hist(pmdat$Aerosol)
hist(log(pmdat$Aerosol))
ggplot(mapping=aes(y=log(Aerosol),x=log(Stationary)),data=pmdat)+
  geom_point()

cor1 <- cor(log(pmdat$Stationary), log(pmdat$Aerosol))
## Boxplot of Aerosol by Activity
ggplot(mapping=aes(x=log(Aerosol),y=Activity),data=pmdat)+geom_boxplot()
base<-lm(log(Aerosol)~log(Stationary)+Activity*ID + log(Stationary):ID,
         data=pmdat)

## Plot the Raw Residual Correlation
mat <- round(cor(matrix(resid(base),100,59,byrow=T)),2)
ndat <- cbind(as.numeric(mat),expand.grid(59:1,1:59))
colnames(ndat) <- c("val","x","y")
ndat <- as.data.frame(ndat)
ggplot(ndat)+
  geom_tile(aes(x=x,y=y,fill=val)) +
  scale_fill_continuous_diverging(limits=c(-1,1)) +
  ggtitle("Correlation Between Raw Residuals") +
  xlab("Minute") + ylab("Minute") +
  scale_y_continuous(breaks=c(0,20,40,60),
                     labels=c("60","40","20","0"))
## Fit our GLS model with AR(1), MA(1) Correlation Structure
pm.gls <- gls(model=log(Aerosol)~log(Stationary) + Activity*ID + log(Stationary):ID,
              data=pmdat, correlation=corARMA(form=~1|ID, p=1, q=1), method="ML")
## AvPlots to test for linearity
#### log(Stationary) is the only numeric variable, so that is the one we test
avPlot(base, "log(Stationary)")

#source("../STAT469/glstools-master/glstools-master/stdres.gls.R")
```

```
source("stdres.gls.R")

## Plot the Decorrelated Residual Correlation
sres <- stdres.gls(pm.gls)
newmat <- round(cor(matrix(sres,100,59,byrow=T)),2)
newdat <- cbind(as.numeric(newmat),expand.grid(59:1,1:59))
colnames(newdat) <- c("val","x","y")
newdat <- as.data.frame(newdat)
ggplot(newdat)+
  geom_tile(aes(x=x,y=y,fill=val))+
  scale_fill_continuous_diverging(limits=c(-1,1)) +
  ggtitle("Correlation Between Decorrelated Residuals") +
  xlab("Minute") + ylab("Minute") +
  scale_y_continuous(breaks=c(0,20,40,60),
                     labels=c("60","40","20","0"))
#### Hist of Standardized Residuals
ggplot()+
  geom_histogram(mapping=aes(x=sres),bins=7)+
  xlab("Residuals")+
  ggtitle("Decorrelated Residuals")+
  theme(text = element_text(size = 8))

#### plotting fitted values vs std residuals
ggplot(data = pmdat) +
  geom_point(mapping = aes(x = fitted(pm.gls), y = sres))+
  ggtitle("Fitted Values vs. Decorrelated Residuals") +
  xlab("Fitted Value") + ylab("Decorrelated Residuals") +
  geom_hline(yintercept = 0)+
  theme(text = element_text(size = 8))
## Calculated R^2 for model to see how well it fits the data
#source("../STAT469/glstools-master/glstools-master/predictgls.R")
source("predictgls.R")
fit.vals <- predictgls(pm.gls)
SSR <- sum((log(pmdat[['Aerosol']]) - fit.vals)^2)
SST <- sum((log(pmdat[['Aerosol']]) - mean(log(pmdat[['Aerosol']])))^2)
R2 <- 1-SSR/SST
## Does Stationary alone do a good job
#### Create a model with just stationary
stat.gls <- gls(model=log(Aerosol)~log(Stationary), data=pmdat,
                correlation=corARMA(form=~1|ID,p=1,q=1), method="ML")

#### Calculated R^2 for model to see how well it fits the data
fit.vals.stat <- predictgls(stat.gls)
SSR.stat <- sum((log(pmdat[['Aerosol']]) - fit.vals.stat)^2)
SST.stat <- sum((log(pmdat[['Aerosol']]) - mean(log(pmdat[['Aerosol']])))^2)
R2.stat <- 1-SSR.stat/SST.stat

## Does Activity add anything
#### Create a model with stationary and activity
act.gls <- gls(model=log(Aerosol)~log(Stationary) + Activity, data=pmdat,
                correlation=corARMA(form=~1|ID,p=1,q=1), method="ML")

##Hypothesis test for using activity vs. using stationary alone
```

```
anova(stat.gls, act.gls)
## Child Specific Activities
#### Create a model with an interaction (pm.gls)
## Anova to compare them
anova(stat.gls, pm.gls)

#### Look at how variable these effects are
## SD for activity interactions
sa <- sd(coef(pm.gls)[109:801])
## SD for log(Stationary) interactions
sl <- sd(coef(pm.gls)[802:900])

## Which activities lead to higher PM levels?
newnames <- names(coef(pm.gls))[109:900] %>%
  str_remove_all("[:ID0123456789]")

meantable <- numeric(7)

#### Homework
hist(coef(pm.gls)[which(newnames=="ActivityHomework")+108],
     main="Homework",xlab="Effect", ylab="Count")
meantable[1] <- mean(coef(pm.gls)[which(newnames=="ActivityHomework")+108])

#### OnPhone
hist(coef(pm.gls)[which(newnames=="ActivityOnPhone")+108],
     main="On Phone",xlab="Effect", ylab="Count")
meantable[2] <- mean(coef(pm.gls)[which(newnames=="ActivityOnPhone")+108])

#### PlayingOnFloor
hist(coef(pm.gls)[which(newnames=="ActivityPlayingOnFloor")+108],
     main="Playing on Floor",xlab="Effect", ylab="Count")
meantable[3] <- mean(coef(pm.gls)[which(newnames=="ActivityPlayingOnFloor")+108])

#### PlayingOnFurniture
hist(coef(pm.gls)[which(newnames=="ActivityPlayingOnFurniture")+108],
     main="Playing on Furniture",xlab="Effect", ylab="Count")
meantable[4] <- mean(coef(pm.gls)[which(newnames=="ActivityPlayingOnFurniture")+108])

#### VideoGames
hist(coef(pm.gls)[which(newnames=="ActivityVideoGames")+108],
     main="Video Games",xlab="Effect", ylab="Count")
meantable[5] <- mean(coef(pm.gls)[which(newnames=="ActivityVideoGames")+108])

#### Walking
hist(coef(pm.gls)[which(newnames=="ActivityWalking")+108],
     main="Walking",xlab="Effect", ylab="Count")
meantable[6] <- mean(coef(pm.gls)[which(newnames=="ActivityWalking")+108])

#### WatchingTV
hist(coef(pm.gls)[which(newnames=="ActivityWatchingTV")+108],
     main="Watching TV",xlab="Effect", ylab="Count")
meantable[7] <- mean(coef(pm.gls)[which(newnames=="ActivityWatchingTV")+108])
```

```
#### Table of mean activity effects
meantable <- matrix(meantable, nrow=1)
colnames(meantable) <- c("Homework", "Phone", "Playing on Floor",
                         "Playing on Furniture", "Video Games", "Walking", "Wathcing TV")
knitr::kable(meantable, caption='Table 2: Average Activity Effects (compared to Computer)', format="lat
  kable_styling(latex_options=c("HOLD_position","scale_down"))
```