Particulate Matter Exposure

Matthew Lindeman

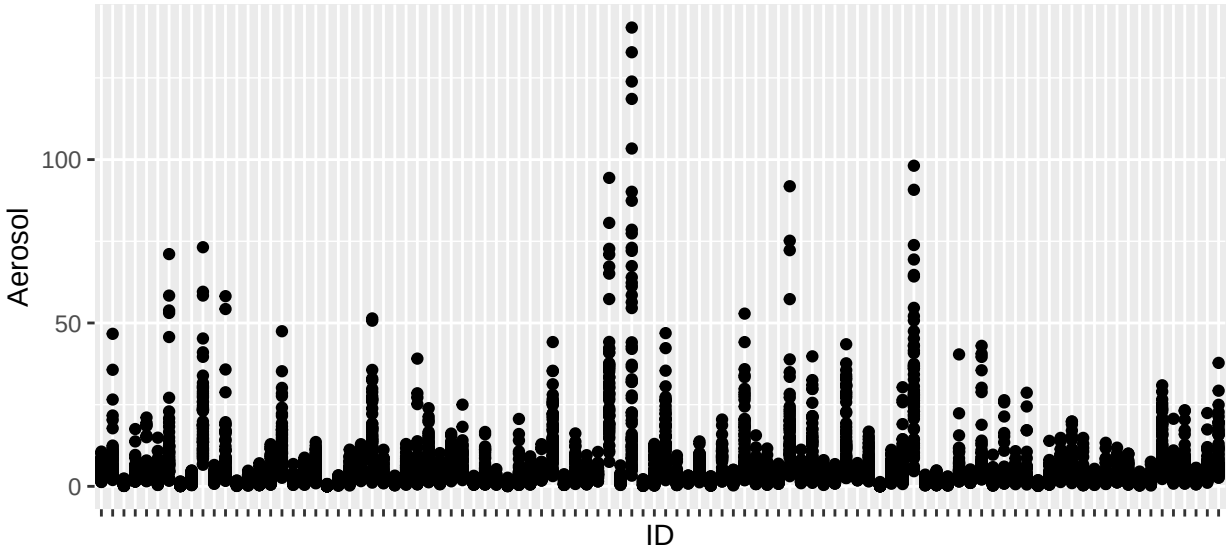
Section 0

From this analysis of trying to evaluate the true particulate matter exposure in children and how different activities affect the amount of PM a child is exposed to, I found a few interesting things regarding the impact various activities have and how specific children can be affected differently. A simple stationary PM measurement in a child's house does not give a very accurate representation of the actual amount of PM a child encounters as they go about various activities. Knowing what activity a child is engaged in can give a better idea of the amount of PM they may be exposed to. I also found that there were a lot of child-specific differences in how much of an effect activities had on their PM exposure. Lastly, I found that playing on the floor and watching TV were the two activities that, on average, led to the highest amount of PM exposure. Overall, I learned that to better understand PM exposure, a lot of factors (such as common activities) should be taken into account.

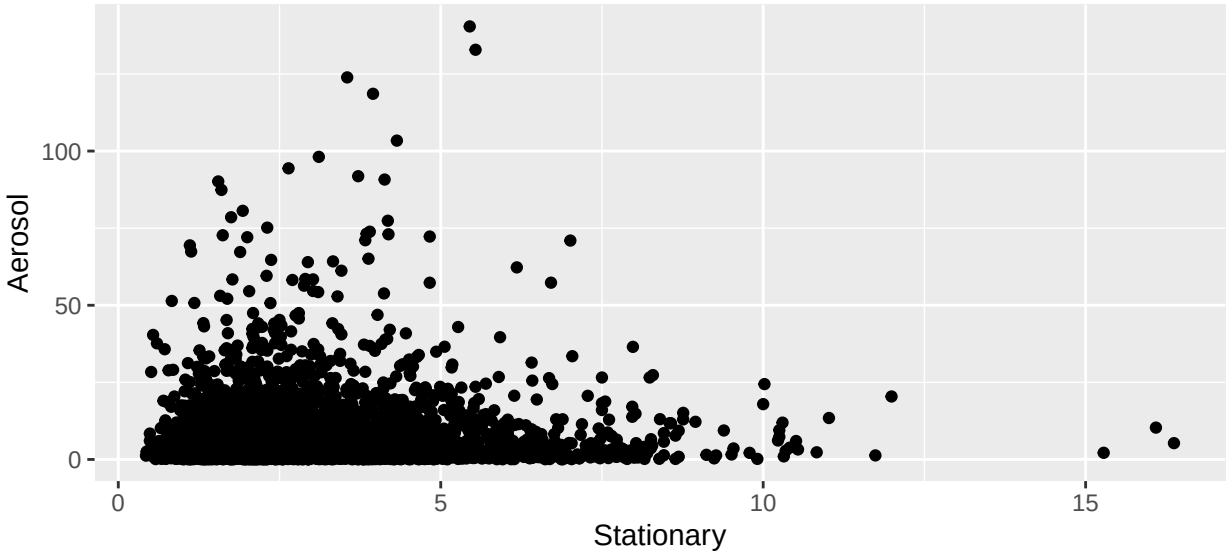
Section 1

The Environmental Protection Agency collected data from children wearing vests with a pollution monitor installed on the shoulder. The intent of the vests is to measure the amount of exposure that children have to PM to better study the health effects of pollution. Particulate matter (abbreviated as PM) is a mixture of particles that can be found in the air, and recent discoveries have shown that PM exposure can affect both heart and lung health. The goals of this study are to learn if certain activities lead to different levels of aerosol intake, how different the effects of activities are from child to child, and which activities lead to a higher PM exposure.

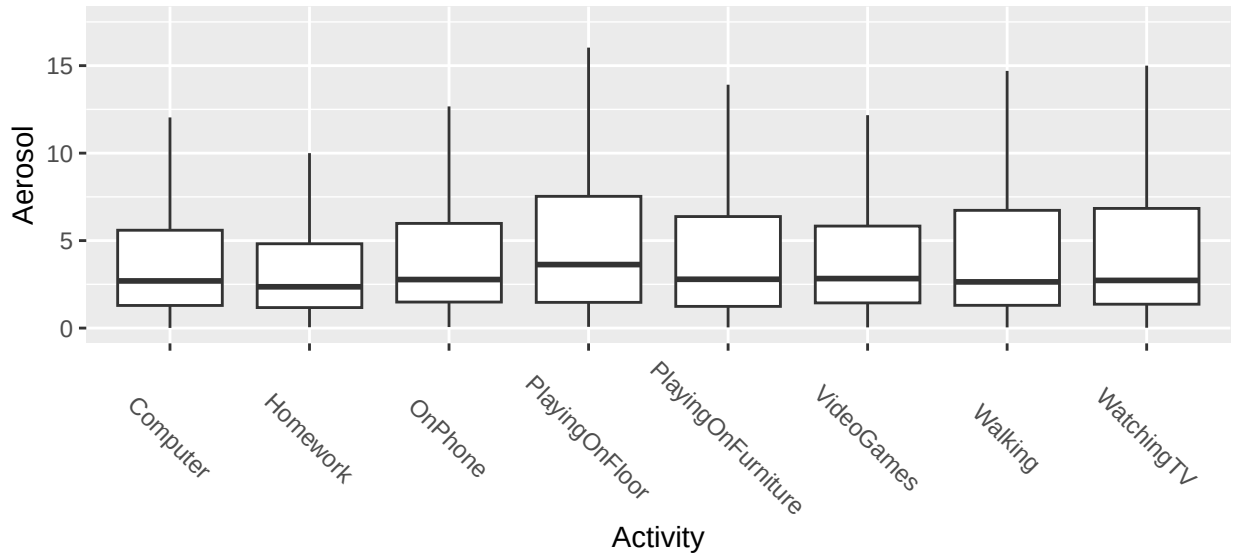
The data provides information on the child ID number (100 children were recruited), the PM measurement on the child's vest, the PM measurement of a stationary monitor placed in the home, the activity the child was engaged in (all three points just mentioned are recorded in 1 minute intervals for 1 hour), and the minute within the hour the child was wearing the vest.



This graphic shows the 59 different PM measurements from each child's vest (or aerosol measurement). It can be seen from the plot that there is a lot of variability between the aerosol ratings of each child that will need to be accounted for.



This graphic shows the general relationship between the Aerosol rating and the PM measurement of a stationary monitor placed in the home (or stationary measurement). It appears that the aerosol rating tends to be higher than the stationary rating, or in other words, the two different devices tend to give different PM measurements.



This graphic shows the relationship between the aerosol rating and different types of activities the kids engaged in (it does not show the outliers to better see the boxes). While there is a lot of overlap between the ratings of each activity, it appears that the average PM measurement tends to be higher when the child is playing on the floor.

A couple potential issues associated with the data that I want to address are the variability in PM measurements between children and the wide range (caused by a small amount of outliers) in PM measurements. The consequence for ignoring the variability in PM measurements between children is that the standard errors would be incorrect, leading to a model that fits the data poorly and doesn't predict accurately. By accounting for this issue, one of the main goals of the study (wondering if the effects of certain activities are child specific) will be able to be accomplished and the model will fit the data much better. The consequence for ignoring the wide range in PM measurements is that the assumptions made to allow the model to be valid may be violated. By accounting for this issue, inferences can be made regarding the data that will allow questions about the data to be answered.

The method I will use to analyze the data started above by making a few exploratory plots to determine what type of statistical model I want to fit to the data. From those plots and from looking at the data, I determined that I want to fit a longitudinal multivariate linear regression model. Then I will transform the aerosol measurements to the log scale for the wide-range issue mentioned in the paragraph above. The next step is to use iterative optimization to get the maximum likelihood estimates of the model parameters from the longitudinal MLR. Then I will verify the model assumptions so I can make inference results to address the goals of the study.

Section 2

Statistical Model: $y \quad N(X\beta, \sigma^2 B)$ $y = X\beta + \epsilon$ $\epsilon \quad \mathrm{N}(0, \sigma^2\mathrm{B})$

y is the response variable or target vector (the aerosol rating or PM measurement on a child's vest).

X is the design matrix containing a column for the intercept and columns to include the data from the data set (Child ID, Stationary Rating, and Activity).

Beta is the vector of parameters where when all x's are zero, y is beta 0 (intercept) on average and as the pth x goes up by 1, y goes up by beta p (coefficient for ID, stationary, and activity) on average.

B is a diagonal matrix with R down the diagonal and zeros everywhere else.

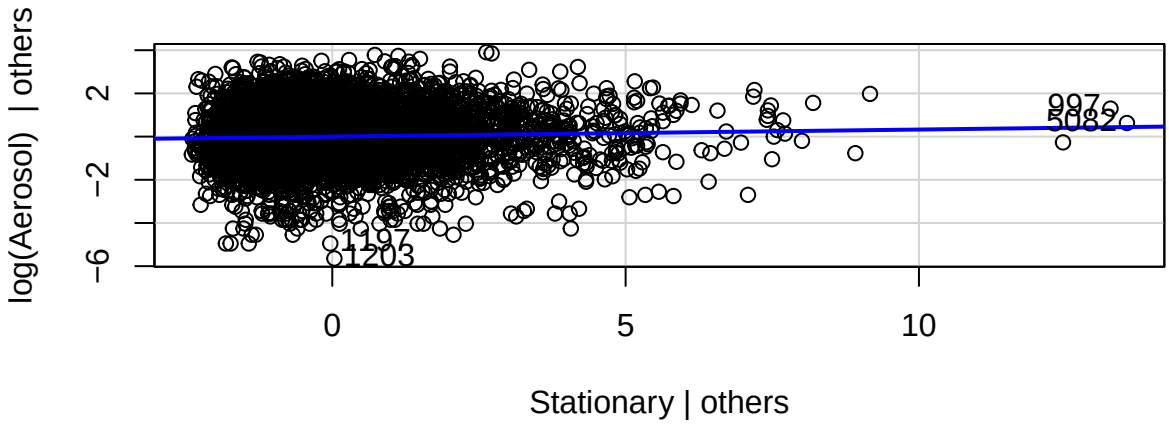
R is a 3x3 correlation matrix for each child with 1's down the diagonal and the other points being the correlation between time points, or in this case between minutes of measurement from the child's vest. For

example, the value in the 1st row and 2nd column is the correlation between the first minute and the second minute.

In this case the correlation function used in the model is the AR1MA1 correlation structure.

This model uses the LINE assumptions or the linearity, independence, normality, and equal variance assumptions. For linearity there needs to be a linear relationship between the numeric explanatory variables and the response variable. For independence the residuals need to be independent of each other, or in other words there can't be correlation between the data points. For normality the standardized residuals of the model need to be normally distributed. Lastly, for the equal variance assumption the standardized residuals need to have constant variance for all the levels of the explanatory variables.

Section 3

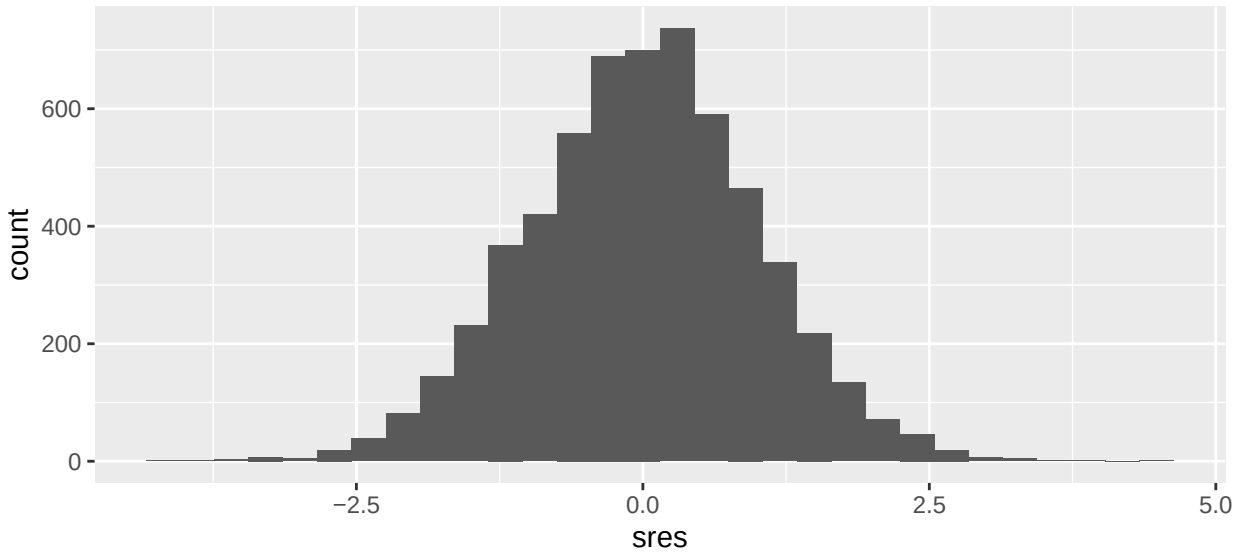


The linear assumption holds here because the added variable plot with the continuous variable shows a fairly linear shape.

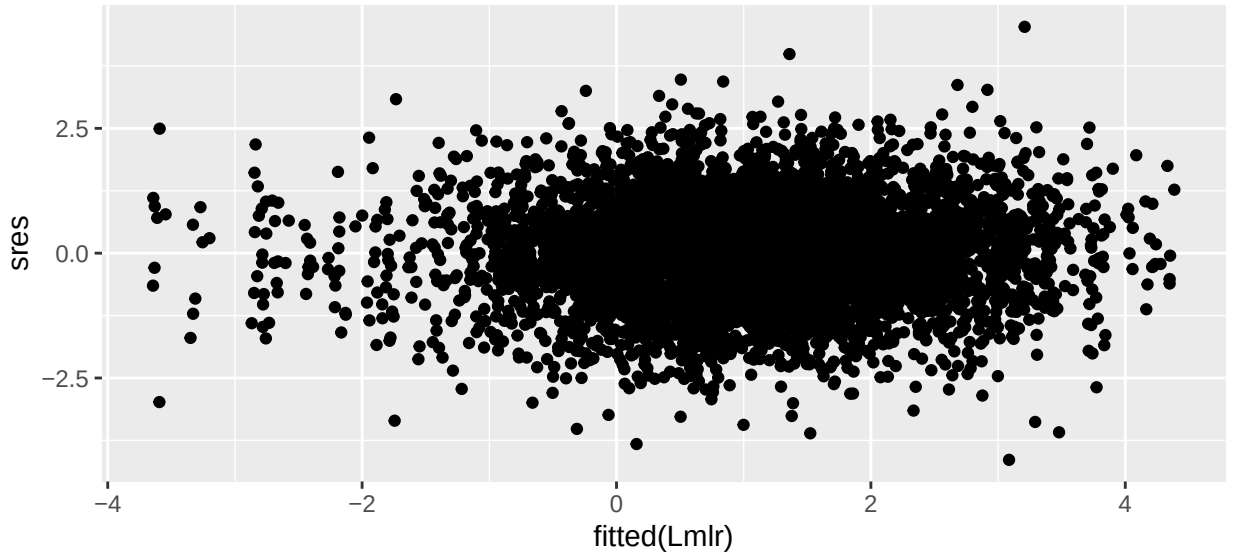
```
## [1] 0.6827369
```

```
## [1] 0.0004046054
```

A 60x60 matrix is too big to display and look at the correlation of the residuals, so I decided to compare the mean of the correlation matrix from a model with only stationary included to the mean of the correlation matrix from the longitudinal model that accounts for correlation. The first number above is the mean for the simple model, and the second number is the mean for the longitudinal model. The correlation is a lot smaller in the longitudinal model, so the independence assumption holds with this model.



The histogram of the residuals looks fairly normal, so the normality assumption holds here.



The plot of the residuals vs. fitted values is centered around zero and there is no concerning shape, meaning the equal variance assumption holds here.

After calculating a couple summary statistics, I found that the model fits the data very well. The first number is the RMSE (0.356) when looking at aerosol rating on the log scale. An RMSE of 0.356 seems pretty small compared to the range (9.55) and standard deviation (1.21) of the aerosol rating. The second number is the pseudo r-squared (0.913), which shows the percentage of the overall variance in aerosol that was captured by the model.

Section 4

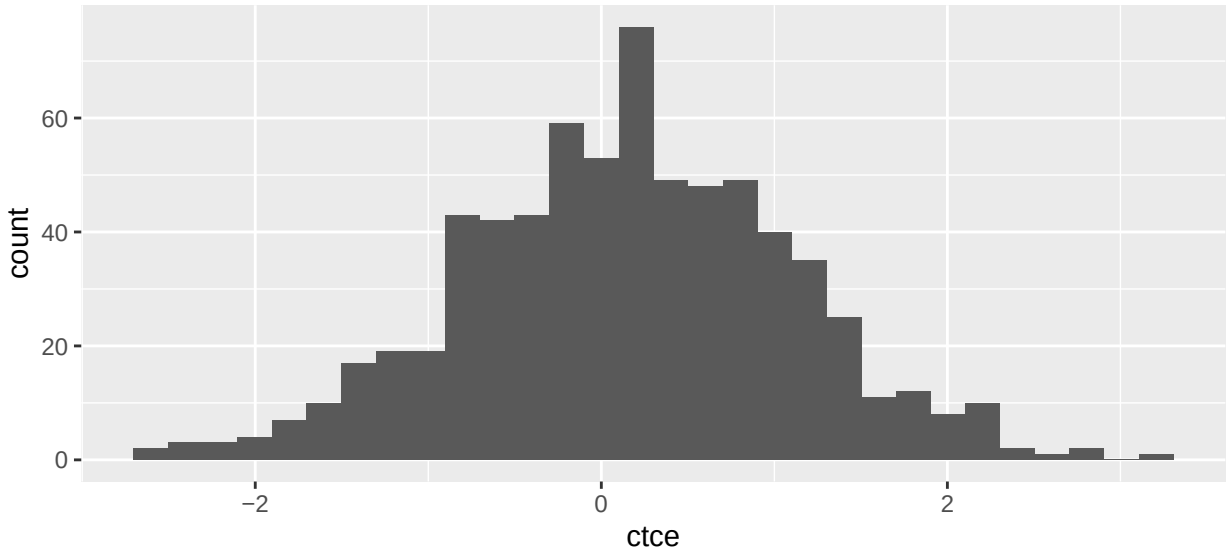
The r-squared value for a model that only uses the stationary measurement to explain PM exposure is 0.0017. This value is very small, especially compared to the r-squared calculated in the previous section (0.913). For that reason, I think that the stationary measurement alone does not do a good job explaining PM exposure.

```
## Analysis of Variance Table
##
## Model 1: log(Aerosol) ~ Stationary
## Model 2: log(Aerosol) ~ Stationary + Activity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   5898 8586.6
## 2   5891 8529.0  7    57.642 5.6877 1.439e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above is the result of a test that I ran to see if there is a difference in a model that only contains the stationary measurement versus a model that utilizes the stationary measurement and the activity. Since the p-value from this test is essentially 0, this means that there is a significant difference between a model that just uses stationary and a model that includes activity as well. This also means that activities help to explain more of the aerosol intake than just stationary alone.

```
##         Model  df      AIC      BIC   logLik   Test  L.Ratio p-value
## reduced     1 111 5445.639 6187.420 -2611.82
## Lmlr        2 804 3945.500 9318.397 -1168.75 1 vs 2 2886.139  <.0001
```

Above is the result of another test I ran to see if there is a child-specific difference in how activities affect PM exposure. Since the p-value is essentially 0, this means the difference between the two models is significant. In context of the study, this means that how much PM exposure a child gets from a certain activity is specific to that child.



This histogram shows all the different child-specific effects that activities have on the aerosol measurement, in the log scale. More specifically, this graph can be used to see how different the effects are of activities from child to child, broadly. To better understand the difference from child to child, I looked at the range of the histogram (5.81). When compared to the range of aerosol rating on the log scale (9.55), I would say that child specific PM exposure varies quite a bit depending on the activity they are doing.

```
##     computer   homework     onPhone   onFloor onFurniture videoGames    walking
## 1 -0.3703123 -0.7662439 -0.04248283 0.4653015  -0.3574568 -0.1232703 -0.8648086
##   watchingTV
## 1  0.1389643
```

The data frame above shows the average PM exposure for each activity averaged across children. Playing on the floor and watching TV are the two activities that led to higher PM exposure, on average.

Section 5

From this analysis of how different activities affect the amount of PM a child is exposed to, I found a few interesting things regarding the impact various activities have and how specific children can be affected differently. A simple stationary PM measurement in a child's house does not give a very accurate representation of the actual amount of PM a child encounters as they go about various activities. Knowing what activity a child is engaged in can give a better idea of the amount of PM they may be exposed to. I also found that there were a lot of child-specific differences in how much of an effect activities had on their PM exposure. Lastly, I found that playing on the floor and watching TV were the two activities that, on average, led to the highest amount of PM exposure. Some of the best "next steps" to take with this would be to look into the activities that led to higher PM exposure and see what it is about those activities that lead to that, and to look into more factors of children's homes (such as cleanliness) to see if those can better explain the child-specific differences seen in PM exposure.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(GGally)
library(forecast)
library(car)
library(multcomp)
library(nlme)
library(mvtnorm)
source("stdres.gls.R")
PM <- read.csv("~/STAT 469/BreathingZonePM.txt", sep=" ")
# Changing some variables to factors
PM$Activity <- as.factor(PM$Activity)
PM$ID <- as.factor(PM$ID)
ggplot(data = PM, mapping = aes(x = ID, y = Aerosol)) +
  geom_point() +
  theme(axis.text.x=element_blank())
ggplot(data = PM, mapping = aes(x = Stationary, y = Aerosol)) +
  geom_point()
# Boxplot for activity, without showing outliers
ggplot(data = PM, mapping = aes(x = Activity, y = Aerosol)) +
  geom_boxplot(outlier.alpha = 0) +
  coord_cartesian(ylim = c(0, 17.5)) +
  theme(axis.text.x = element_text(angle = -45, vjust = 0.5))
# Fitting chosen model and validating assumptions
Lmlr <- gls(model=log(Aerosol)~.-Minute+ID:Activity, data=PM,
            correlation=corARMA(form=~Minute|ID, q=1, p=1), method='ML')

# Decorrelating residuals
sres <- stdres.gls(Lmlr)
# Fitting just stationary model
mlr <- lm(formula=log(Aerosol)~Stationary, data=PM)
```

```r
# AVPLots
avPlots(mlr)
# Correlation matrix of decorrelated residuals
cmat <- cor(matrix(sres, 100, 59, byrow=T))

# Comparing the means of the correlation matrices for the mlr with only stationary
mean(cor(matrix(resid(mlr), 100, 59, byrow=T)))
# versus the model with correlation accounted for
mean(cmat)
# Histogram of the residuals
ggplot() +
  geom_histogram(mapping = aes(x = sres))
# Scatter plot of the fitted values vs. residuals
ggplot() +
  geom_point(mapping = aes(x = fitted(Lmlr), y = sres))
# RMSE
rmse <- sqrt((1/5900) * sum((log(PM$Aerosol) - fitted(Lmlr))^2))
# Pseudo R Squared
psRs <- (cor(log(PM$Aerosol), fitted(Lmlr)))^2

# Range of log aerosol
rn <- range(log(PM$Aerosol))
# Std dev of log aerosol
std <- sd(log(PM$Aerosol))
# Looking at the r-squared of the model with only stationary
summ <- summary(mlr)
rsqr <- summ[8]
# Anova test to compare stationary model to stationary + activity model
stact <- lm(formula=log(Aerosol)~Stationary+Activity, data=PM)
anova(mlr, stact)
# Anova test to see if the interaction term is significant
reduced <- gls(model=log(Aerosol)~.-Minute, data=PM,
            correlation=corARMA(form=~Minute|ID, q=1, p=1), method='ML')
anova(reduced, Lmlr)
# Histogram of the coefficients when looking at the effect on children
ctce <- coef(Lmlr)[109:801]
rnge <- range(ctce)
ggplot() +
  geom_histogram(mapping = aes(x = ctce))
# Looking at the average PM exposure across children for each of the activities
avgPM <- data.frame(
"computer" = sum(coef(Lmlr)[1:100])/99,
"homework" = sum(coef(Lmlr)[c(1:100, 102, 109:207)])/99,
"onPhone" = sum(coef(Lmlr)[c(1:100, 103, 208:306)])/99,
"onFloor" = sum(coef(Lmlr)[c(1:100, 104, 307:405)])/99,
"onFurniture" = sum(coef(Lmlr)[c(1:100, 105, 406:504)])/99,
"videoGames" = sum(coef(Lmlr)[c(1:100, 106, 505:603)])/99,
"walking" = sum(coef(Lmlr)[c(1:100, 107, 604:702)])/99,
"watchingTV" = sum(coef(Lmlr)[c(1:100, 108, 703:801)])/99
)
# Displaying the dataframe of the averages
avgPM
```