Analysis of Particulate Matter Exposure in Children

Carly Lundgreen and Logan Perry

March 28, 2019

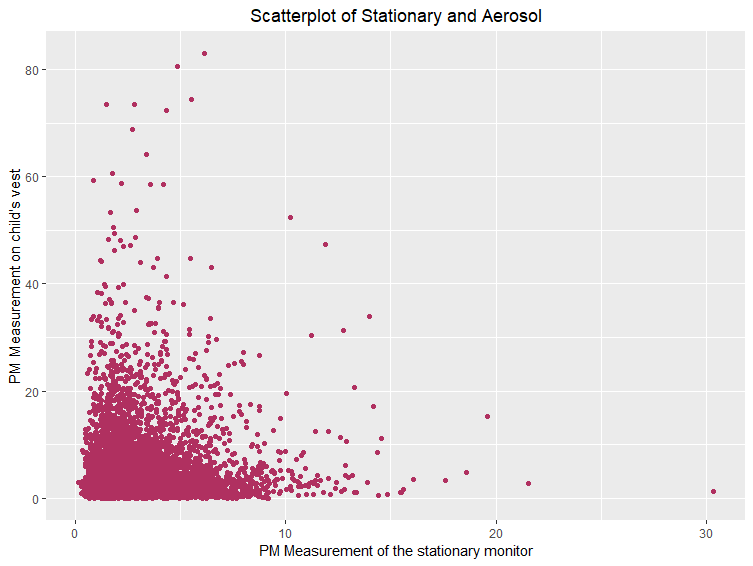
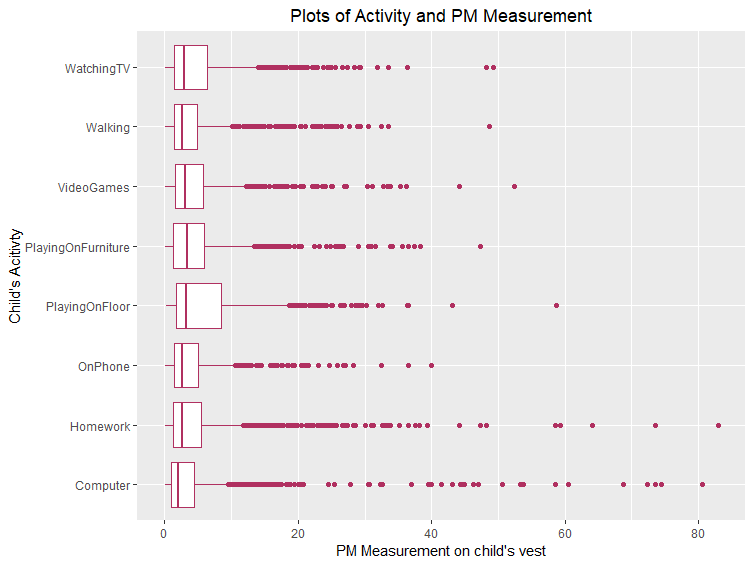
### Executive Summary

Our findings concerning the topic of particulate matter (PM) exposure in children were that the stationary monitor PM readings and activity were significant factors in estimating the PM measured on the childrens’ vests (Aerosol), although only using the stationary PM measurement still provides a good estimate of Aerosol. There is also a stationary child-to-child affect which makes sense considering that the children likely came from different households.

### Introduction and Background

PM (particulate matter) is a mixture of particles in the air that can either be seen with the naked eye or are microscopically small. Measuring PM can be particularly tricky because it can vary drastically from area to area and even from outdoors to indoors. There have been studies that suggest PM exposure is linked to various adverse health effects such as nonfatal heart attacks and decreased lung function. To get a better idea of how much PM exposure children really get, the EPA attached a PM measuring apparatus to a vest and put them on children along with a GoPro camera to observe what the children doing. They also installed a stationary PM monitor in the home of each child. The measurement of the vest monitor is the ‘Aerosol’ variable, while the stationary monitor’s readings are found in the Stationary variable. The ‘Minute’ variable counts the number of minutes each child has been observed. The last variable in the dataset is ID which is the number assigned to each child.

The scatterplot below illustrates the relationship between the stationary PM measurement and the PM measurement from the child’s vest. A log transformation will be applied to Aerosol when we fit our model. We can also see that the spread of PM measurements for children playing on the computer is comparable to the spread of PM measurements for children doing homework, and these two activities seem to have the largest spread in PM measurements.



The goal of the study is to evaluate different effects such as Stationary and Activity on Aerosol PM measurements. One challenge we foresaw in this analysis was the correlation being the same within subject ID since the same subject was recorded 118 times. If we do not take this into account, our standard errors will be incorrect which will affect our confidence and prediction intervals. The advantage of accounting for this correlation is having more accurate standard errors allowing for correct interval estimates. The scatterplot of Aerosol and Stationary shows heteroskedasticity, or unequal variance along the regression line, which means that the linearity assumption for our regression analysis is not met. To account for this correlation, we will fit a GLS model with an Autoregressive correlation structure of order 1 on ID and we plan on applying a log transformation to the Aerosol variable to deal with the heteroskedasticity in the response variable. Adding this correlation structure will decorrelate the residuals and allow us to perform inference procedures on various parameters.

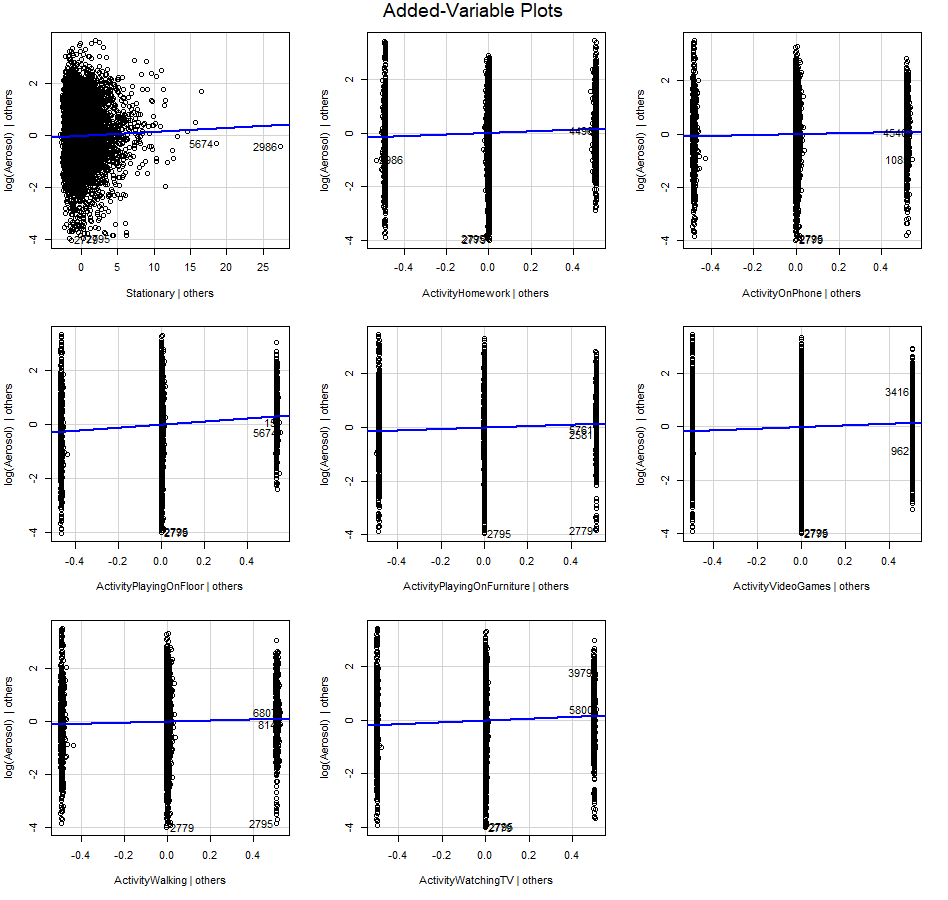
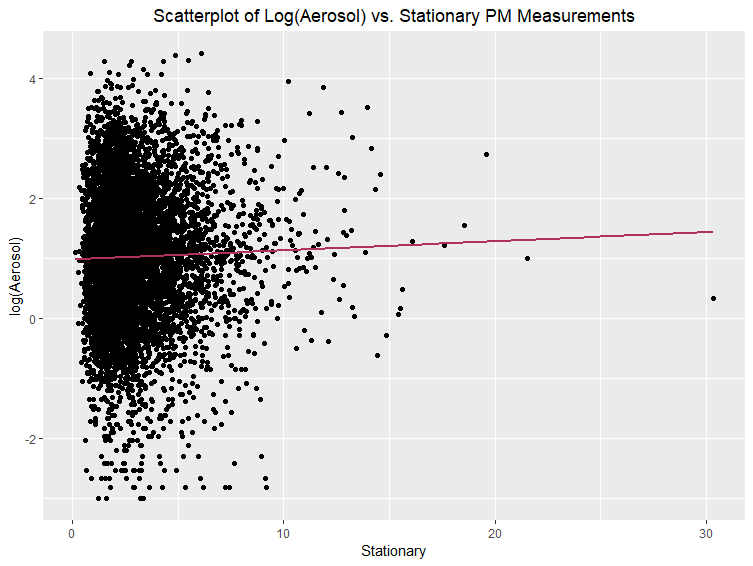
### Statistical Model

We define our model as follows: log(Y) ~ N(Xβ, σ2B) Where Y is the 1 x n vector of Aerosol measurements for each of the 60 people in this study, as recorded 118 times (each measurement being per minute). In this case, n=7080. X is defined as the design matrix consisting of a column of 1’s, as well as a linear term for ‘Stationary’ and ‘Activity’. We left out the child’s ID and the Minute at which the PM exposure was measured from the model—the effects of these are accounted for with a general symmetric correlation structure. β is a 1 x P vector of the beta coefficients, which are the effects of each of the explanatory variables on the children’s log(Aerosol) measurement. σ is the variance of the log(Aerosol) for the subjects, and B is the n x n block diagonal matrix with zeros on the off-diagonals, and 3x3 correlation matrices (R) along the diagonal—indicating a correlation structure occurring longitudinally among patient. Each correlation matrix, R, has 1’s along the diagonal and correlation values along the off diagonals—which are the correlations between PM measurements of a single child over the course of the 118 minutes.

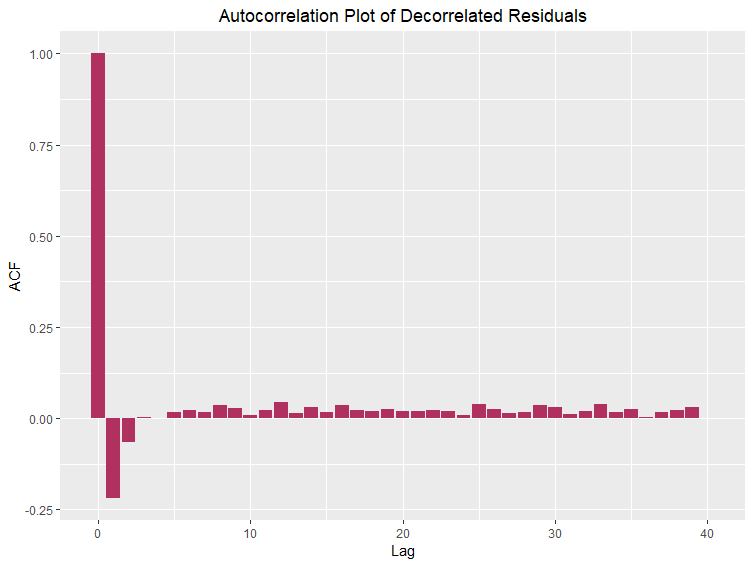
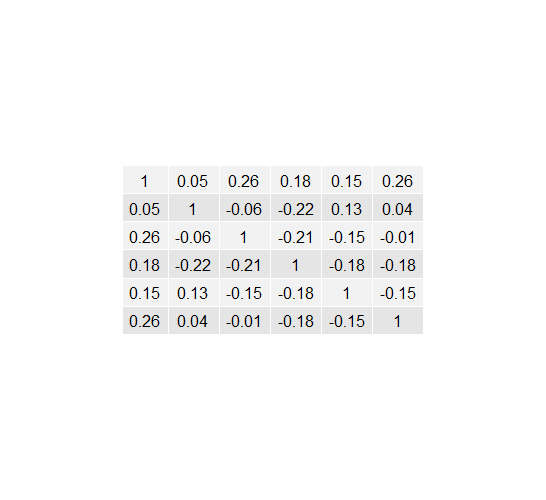
This model assumes linearity among the responses with respect to the stationary PM measurement--which was the only quantitative variable used in this model. It also assumes independence among responses. This is not the case, as there is clear correlation between within-child measurements, but the model will account for this and the residuals should be decorrelated. Finally, the model assumes normality and equal variance among the decorrelated residuals.

### Model Validation

Based on the added-variable plots shown below, we can see that the data is sufficiently linear enough to proceed. There appears to be a slightly positive, linear relationship between stationary and log(Aerosol).



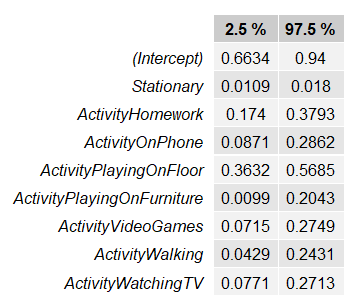
There is not independence between all log(Aerosol) measurements because the subjects are correlated with themselves. This is accounted for in our AR1 correlation structure. Shown below are the first 6 rows and columns of the correlation matrix after the AR1 correlation correction was applied. The correlation values are reasonable after compensation with the AR1 correction. We can also see from the ACF plot below that the correlations have (mostly) stabilized near zero.



Based on the histogram below, we can see that the decorrelated and standardized residuals are indeed normally distributed. The final assumption of equal variance is affirmed by a plot of the fitted values and decorrelated standardized residuals below--which seem to have an even spread about the zero line.

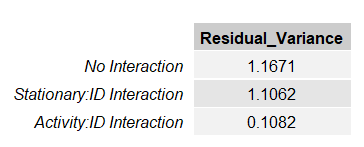
### Analysis Results

The p-value for the effect of the stationary measurement on the PM measurement taken from the children’s vests is 0.0000. A confidence interval for this effect is found to be (0.663, 0.940). Because the stationary measurement is significantly positive, it appears that the stationary effect does well at explaining PM exposure without including the effects of activities. As a further analysis, we performed an ANOVA test on our model and compared it to a model that only contained the stationary variable. We obtained a p-value < 0.0001 showing that there were statistically significant differences between the two. Further investigation showed that our model was better by AIC and BIC, but there was not a large difference. The AIC was smaller by 74 with a score of 6947 and the BIC was 30 smaller with a score of 7022. Therefore, we can conclude that while having stationary alone isn’t the best model, it is still a good model.



In addition to the effect of the stationary measurement, all of the activities were significantly associated with higher PM exposures. A table of confidence intervals for these effects is given. Each of the activities--homework, on phone, playing on the floor and on furniture, video games, walking, and watching TV--all lead to higher PM exposure.

We performed two ANOVA tests to see whether the effects of activity/stationary measurement are equal across ID--which would mean these effects are not child-specific. We compared our model, which did not include interactions between stationary and ID or activity and ID, with two different models. The first included an interaction between ID and stationary, and the second included an interaction between ID and activity. The p-values for both of these tests were < 0.0001, which means that the effects of both stationary and activity are in fact child-specific. When looking at the variances of the residuals from each of these models, the variance for the model that includes the interaction between activity and ID is about 0.11, which is much smaller than the variances of our model and the model that includes the interaction between stationary and ID. A table below summarizes these variances.



It seems that there is less variability in the effects of activity from child to child than there is in the effect of stationary PM measures from child to child.

### Conclusions

This analysis of particulate matter exposure in children found that the effects of activities such as homework or watching TV will affect the amount of particulate matter that children are exposed to. Specifically, these activities are associated with increased PM exposure. These effects differ from child to child, however. The stationary measurement was also found to be child-specific, and this can be attributed to the high likelihood that the 60 children came from different households. We also found that Stationary is a good predictor on its own, but adding Activity creates a better GLS model using an AR1 correlation structure given ID.

Moving forward, we recommend observing different parts of the home to see if there is more pollution in different areas. This could be achieved by adding multiple stationary PM monitors in different parts of the house. It would also be interesting to note the floor surface in the house since this is most likely where the PM ends up. I would anticipate that since hardwood and tile surfaces are easier to clean that those homes would have lower PM levels compared to homes with mostly carpet.

Appendix: R Code

pollution <- read.csv("https://mheaton.byu.edu/Courses/Stat469/Topics/2%20-%20TemporalCorrelation/3%20-%20Project/Data/BreathingZonePM.txt", sep = " ", header = TRUE)

ggplot(data = pollution) +  
 geom\_point(mapping = aes(x = Stationary, y = Aerosol), col = "maroon") +  
 labs(x = "PM Measurement of the stationary monitor", y = "PM Measurement on child's vest", title = "Scatterplot of Stationary and Aerosol") +  
 theme(plot.title = element\_text(hjust=0.5))

##Validating a gls model  
  
# create a linear model under the independence assumption and check the correlation matrix structure  
stationary.lm <- lm(formula = log(Aerosol) ~ ., data = pollution)  
resids.lm <- stdres(stationary.lm)  
corr.matrix <- matrix(data = resids.lm, ncol = 118, byrow = TRUE) %>% cor() #all numbers above 0.5 from the looks of it, so there is definitely some temporal correlaion within the resuls for each child  
View(corr.matrix)  
  
# find the best model for our correlation  
  
AR1.gls <- gls(model = log(Aerosol)~.-ID-Minute, data = pollution, correlation = corAR1(form = ~Minute|ID), method = "ML")  
  
AIC(AR1.gls)

## [1] 6947.316

####################################  
## Validate GLS Model Assumptions ##  
####################################  
  
#Linearity   
ggplot(data=pollution,mapping=aes(x=Stationary,y=log(Aerosol))) + geom\_point() +   
 geom\_smooth(method="lm",se=FALSE,col="maroon") + ggtitle("Scatterplot of Log(Aerosol) vs. Stationary PM Measurements") +  
 theme(plot.title = element\_text(hjust=0.5))

#Independence   
# There isn't independence, but the correlation is accounted for in our gls model  
##we need to check that the residuals have actually been decorrelated  
new.corr.matrix <- matrix(data=stdres.gls(AR1.gls),ncol=118,byrow = TRUE) %>% cor()

acf.plot <- acf(stdres.gls(AR1.gls),lag.max=40) ##the ACF plot also shows the decorrelation (AR1 captures most of the correlation)

ACF.dframe <- data.frame(Lag=acf.plot$lag, ACF=acf.plot$acf)  
ggplot(data=ACF.dframe, aes(x=Lag, y=ACF)) + geom\_col(fill="maroon") +   
 ggtitle("Autocorrelation Plot of Decorrelated Residuals") +   
 theme(plot.title = element\_text(hjust=0.5))

library(gridExtra)

library(grid)  
grid.newpage()  
grid.table(round(head(new.corr.matrix)[,1:6],2))

##Normality of standardized/decorrelated residuals  
  
ggplot()+geom\_histogram(mapping=aes(x=stdres.gls(AR1.gls)),color="black",size=1,fill="maroon",binwidth=.4) + ggtitle("Histogram of Standardized Residuals") +   
 labs(x="Standardized & Decorrelated Residuals", y="Frequency") + theme(plot.title = element\_text(hjust = 0.5))

##Plot of fitted values vs. residuals   
ggplot(data=pollution, mapping=aes(x=fitted(AR1.gls),y=stdres.gls(AR1.gls))) + geom\_point(color="maroon") + geom\_abline(slope=0, intercept=0,size=1) +  
 ggtitle("Fitted Values vs. Residuals") + labs(x="Fitted Values",y="Residuals") + theme(plot.title = element\_text(hjust = 0.5))

###########################  
## Statistical Inference ##  
###########################  
  
##1   
other.model <- gls(model=log(Aerosol)~.-Stationary-Minute,data=pollution, correlation=corAR1(form = ~Minute|ID),method="ML")  
anova(other.model,AR1.gls)

## Model df AIC BIC logLik Test L.Ratio p-value  
## other.model 1 69 5971.763 6445.450 -2916.882   
## AR1.gls 2 11 6947.316 7022.832 -3462.658 1 vs 2 1091.553 <.0001

confint(AR1.gls)

##2   
summary(AR1.gls) ## it looks like all activities are significant at the .05 level..

intervals <- tableGrob(round(confint(AR1.gls),4))  
grid.newpage()  
grid.arrange(intervals)

##3   
  
stationary.int.gls <- gls(model = log(Aerosol)~Stationary + Activity + ID:Stationary, data=pollution,correlation = corAR1(form = ~Minute|ID), method = "ML")  
  
activity.int.gls <- gls(model = log(Aerosol)~Stationary + Activity + ID:Activity, data=pollution,correlation = corAR1(form = ~Minute|ID), method = "ML")  
  
anova(AR1.gls,stationary.int.gls)

## Model df AIC BIC logLik Test L.Ratio  
## AR1.gls 1 11 6947.316 7022.832 -3462.658   
## stationary.int.gls 2 70 6579.704 7060.256 -3219.852 1 vs 2 485.6124  
## p-value  
## AR1.gls   
## stationary.int.gls <.0001

anova(AR1.gls,activity.int.gls)

## Model df AIC BIC logLik Test L.Ratio  
## AR1.gls 1 11 6947.316 7022.832 -3462.658   
## activity.int.gls 2 483 4455.279 7771.088 -1744.640 1 vs 2 3436.037  
## p-value  
## AR1.gls   
## activity.int.gls <.0001

s1 <- (summary(AR1.gls)$sigma)^2  
s2 <- (summary(stationary.int.gls)$sigma)^2  
s3 <- (summary(activity.int.gls)$sigma)^2  
  
Residual\_Variance <- round(c(s1,s2,s3),4)  
Model <- c("No Interaction", "Stationary:ID Interaction", "Activity:ID Interaction")  
table <- tableGrob(as.data.frame(Residual\_Variance, Model))  
grid.newpage()  
grid.arrange(table)