PM Exposure

Joshua Christensen, Greg Paulukaitis and Riley Millar

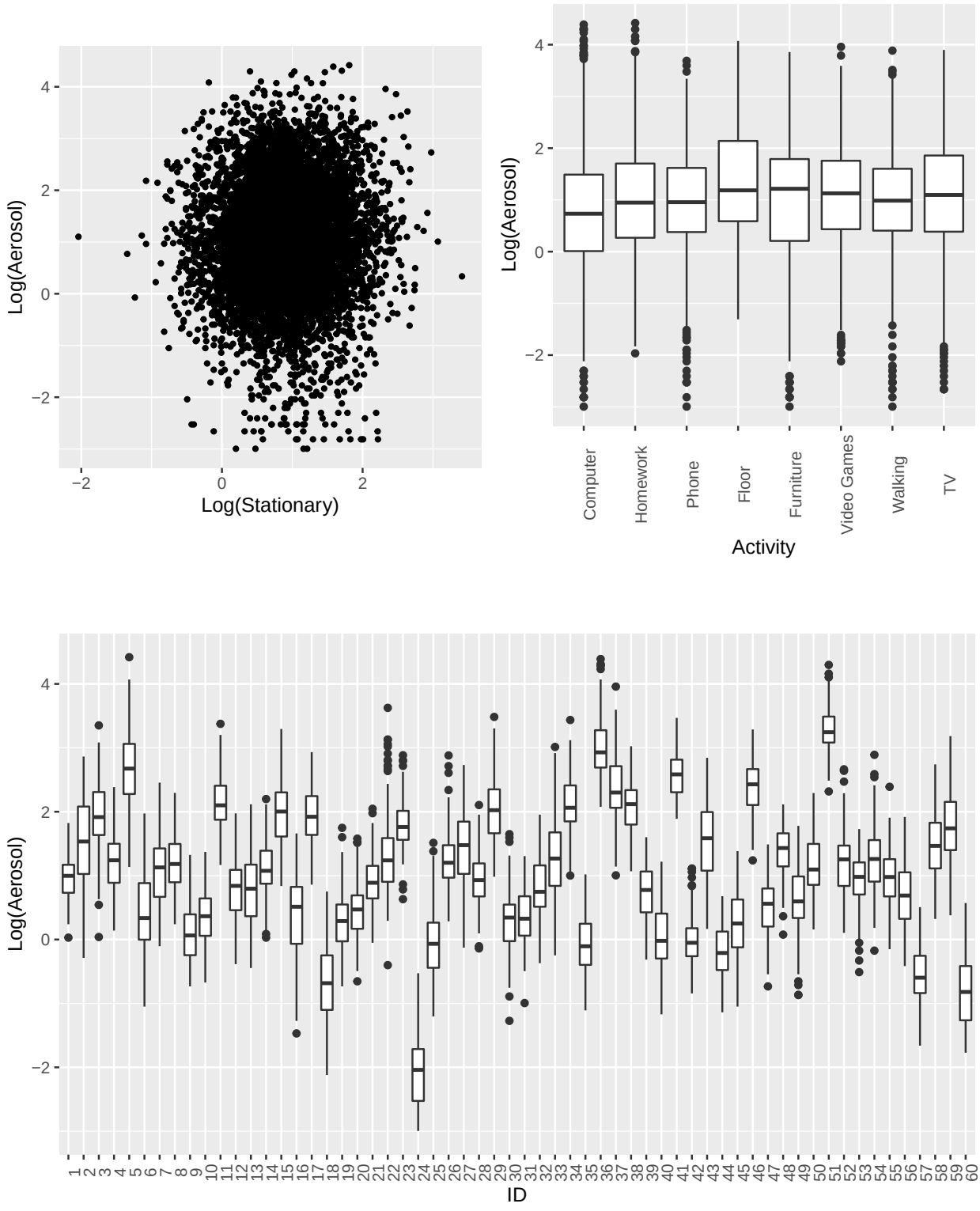
4/1/2021

Section 0: Executive Summary

Our analysis aimed to answer three general questions regarding child particulate matter exposure. The first was to determine whether a stationary atmosphere monitor provided a good indication of the actual particulate matter exposure that the child experienced. The second was whether the amount of PM exposure was related to the activities that the child was engaged in, and whether the activity effects varied by child. The third was whether we could accurately model a child's PM exposure when we included stationary measurements along with child-specific activity effects. The dataset we used contains child IDs, aerosol PM measurement, stationary PM measurement, activities and the minute each measurement was taken. From our analysis we determined that stationary PM measurements are not sufficient to explain actual PM exposure on their own. We also determined that activities did have a significant effect on PM exposure and that the effects of each activity were highly variable by child. We concluded that by using child-specific stationary PM measurement effects and child-specific activity effects together, we were able to accurately explain actual PM exposure.

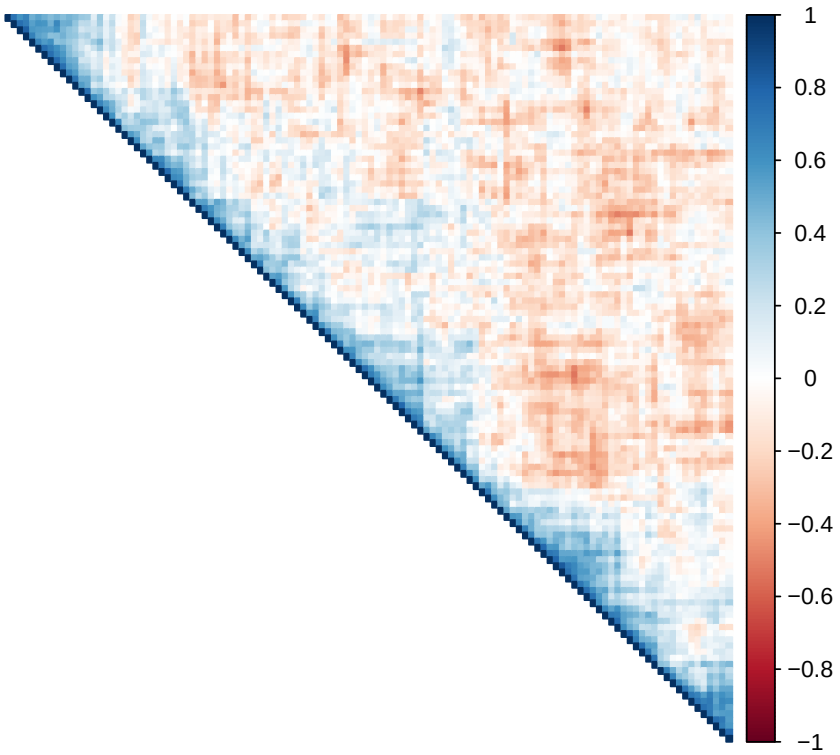
Section 1: Introduction and Problem Background

The air around us is filled with various types of particles. Certain types of "particulate matter" (PM) can be detrimental to the health of those who breathe it in, especially those who suffer with conditions such as asthma. This exposure can occur both indoors and outdoors. We have access to a dataset consisting of PM measurements in the homes of 60 different children. These measurements were collected over the course of two hours and in two separate ways: 1) via a stationary PM monitor placed inside each child's home, and 2) via a vest-mounted PM monitor that each child wore throughout the two hours. We also have a catalog of activities that each child was engaged in during the two-hour time span, and how long (in minutes) the respective child did each activity. By performing analyses on this dataset, we hope to determine if the stationary measurement alone is sufficient to understand a child's PM exposure, and if there are certain activities in the home that lead to higher average PM exposure than other activities. We also hope to learn if the effects of the stationary/activity measurements are specific to each child, and how much variability there is in the effects of each activity for the 60 children.



One potential issue in our dataset is the correlation between observations across minutes for each individual child. If we group observations by child, and treat the individual minutes as individual observations, it is clear to see that these observations (minutes) are correlated with each other. There is likely not much

correlation between children, since their living situations are variable. If we ignored this issue, our standard error calculations would be incorrect. This would in turn cause problems in our confidence intervals and prediction intervals. We account for this issue by using an ARMA correlation structure with 2 auto-regressive terms and 1 moving-average term. By doing this, we will improve our ability to make inference and to accurately portray the variation of the data.



Due to the correlation of observations across minutes for each child, we will fit a longitudinal ARMA 2-1 model with the logged stationary variable by itself and compare it to a similar ARMA 2-1 model that includes all variables and interactions. This will allow us to determine whether or not the stationary measurement alone is a good indicator for a child's amount of PM exposure. We will check our model assumptions and verify the validity of our model using metrics such as R^2 and root mean square error (RMSE). We will be able to use these models to determine if the effects of the various activities are child-specific, how much variability there is among the effects of the activities, and which activities lead to higher PM exposure.

Section 2: Statistical Model

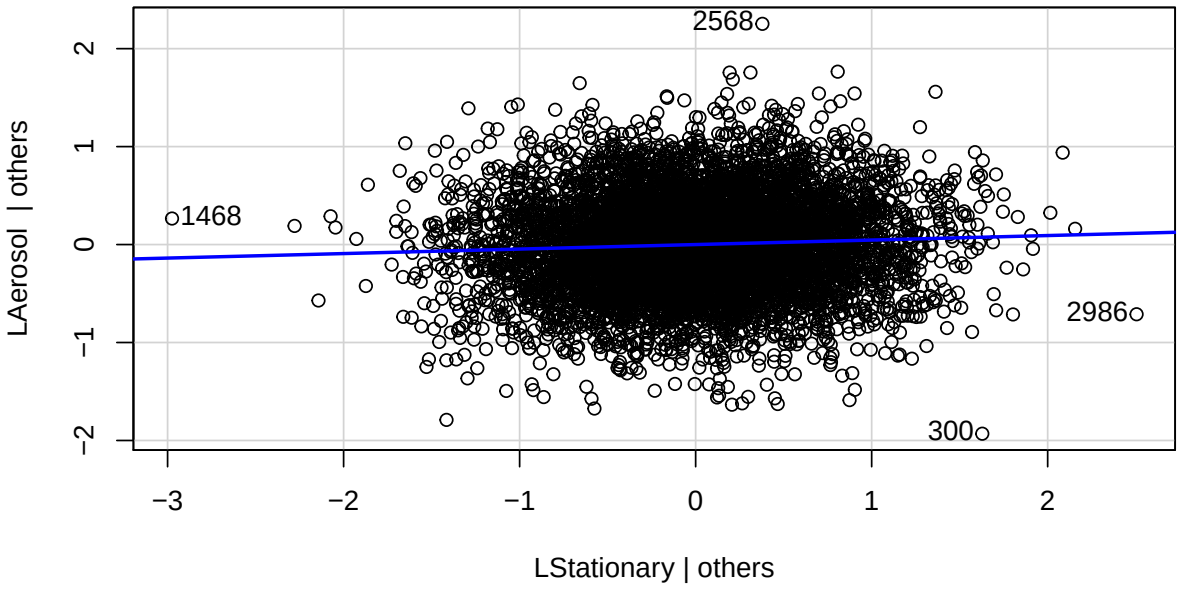
We will use a longitudinal multiple linear regression model to analyze the particulate matter data. We define the model for our data as $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{B})$. In our model, $\boldsymbol{y}$ represents the log of the aerosol particulate matter measurement taken from the vest of a given child. We are using $\boldsymbol{y}$ as our response variable because we wish to model the actual particulate matter exposure that a child experiences. $\boldsymbol{X}$ represents the design matrix which includes our explanatory variables as well as a column of ones, representing the intercept. We included the natural log of the stationary PM measurement, the activity, and the child ID as variables in our $\boldsymbol{X}$ matrix. We also included interaction terms for child ID with the natural log of stationary measurements and for child ID with activity after observing that the apparent effects of these two variables

seemed to vary drastically by child. We did not include minute as an effect in our model since we did not expect PM measurements to change purely as a function of time. Our parameters in this model are $\boldsymbol{\beta}$, σ^2, and the parameters of our correlation structure, ϕ_1, ϕ_2 and θ. $\boldsymbol{\beta}$ is a vector of coefficients that quantifies the relationship between each respective explanatory variable and the response variable, in addition to setting the intercept. σ^2 represents the variance of the residuals from our model. The $\boldsymbol{B}$ matrix is a block diagonal matrix that includes $\boldsymbol{R}$ matrices of size 118×118 for each individual child. The off-diagonal values of these matrices define the temporal correlation of each child's aerosol PM measurements. These values are determined by estimating ϕ_1, ϕ_2 and θ within our defined correlation function. These variables can be thought of as governing the amount of long-term (for the ϕ's) and short-term (for θ) temporal correlation we expect to see.

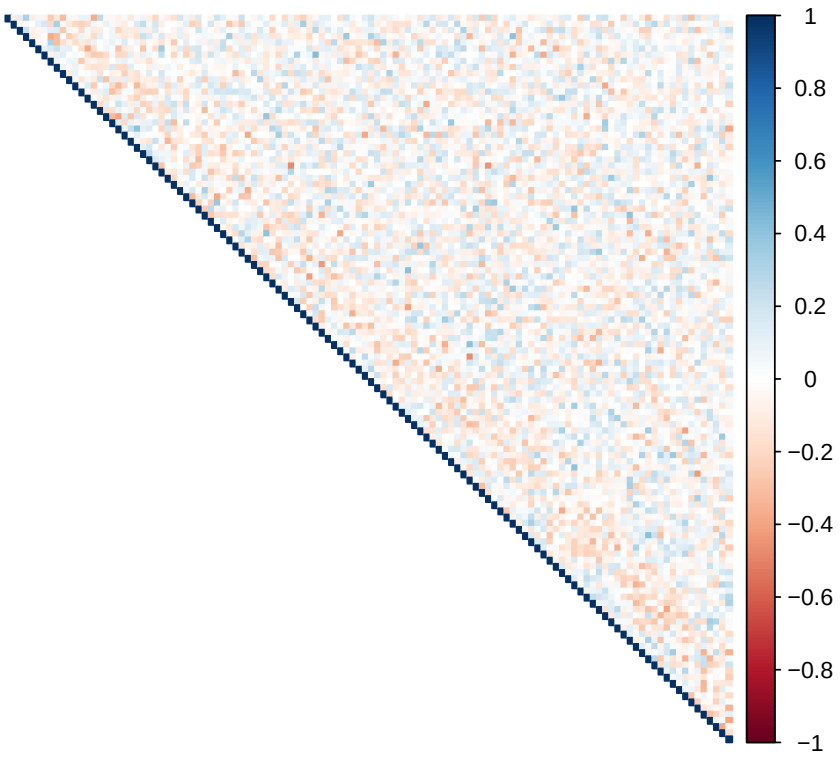
Our model depends on four assumptions. The first is that all quantitative variables have a linear relationship with the response variable. In this case this would mean that the log of the stationary PM has a linear relationship with the log of the aerosol PM. The second is that each child is independent of the other children. Also, while we do not assume independence between measurements on the same child, the validity of our model does depend on capturing the temporal correlation with our correlation function. The third assumption is that the residuals are normally distributed with mean 0 and variance σ^2. The last assumption is that our $\boldsymbol{B}$ matrix correlation structure combined with σ^2 accurately represents the variability of the data. In other words, while our response is not assumed to have constant variance, we would like our covariance matrix to account for the changes in variance so that the underlying σ^2 is still constant. This is equivalent to constant variance of the residuals.

Section 3: Model Validation

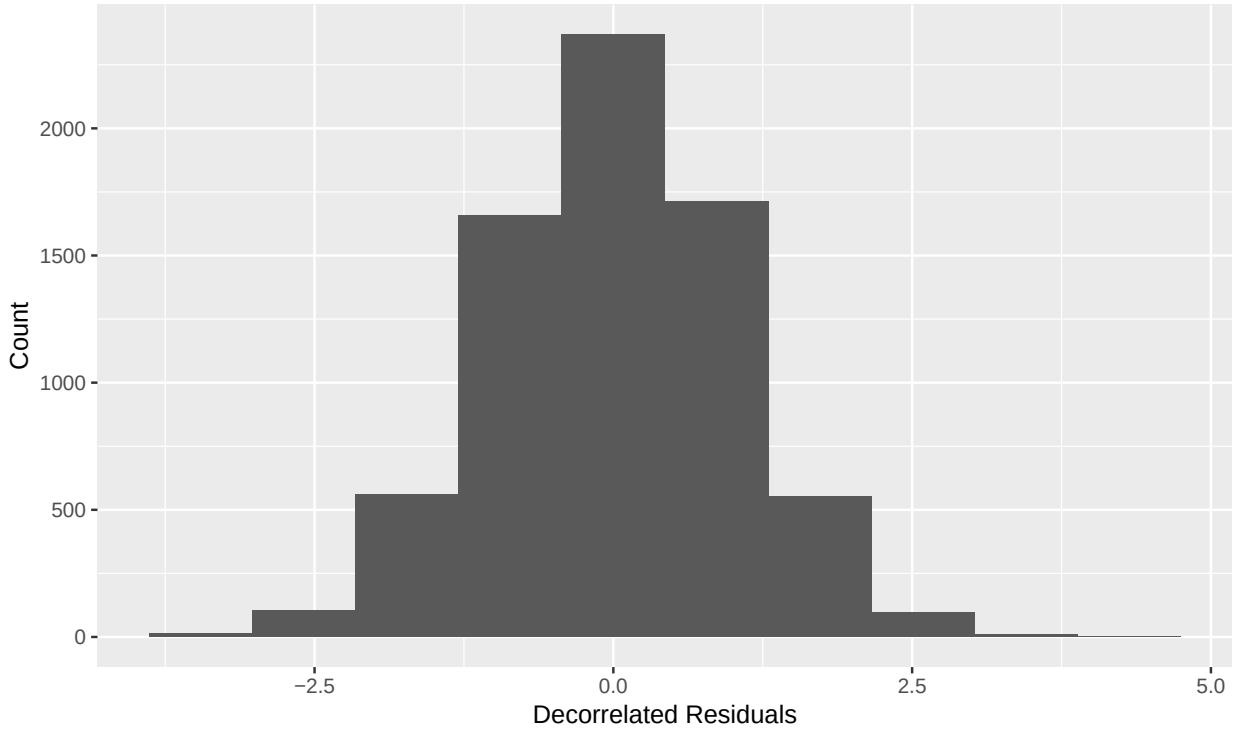
We tested the linearity assumption using added-variable plots, which regress both the response and the explanatory variable against all other variables in the model and then display the resulting fitted values plotted against one another. This allows us to evaluate the relationship between the explanatory variable and the response, while accounting for the effects of all other variables in the model. The added-variable plot displayed below for the log of the stationary measurements against the log of the aerosol measurements shows no clear deviation from a linear relationship.



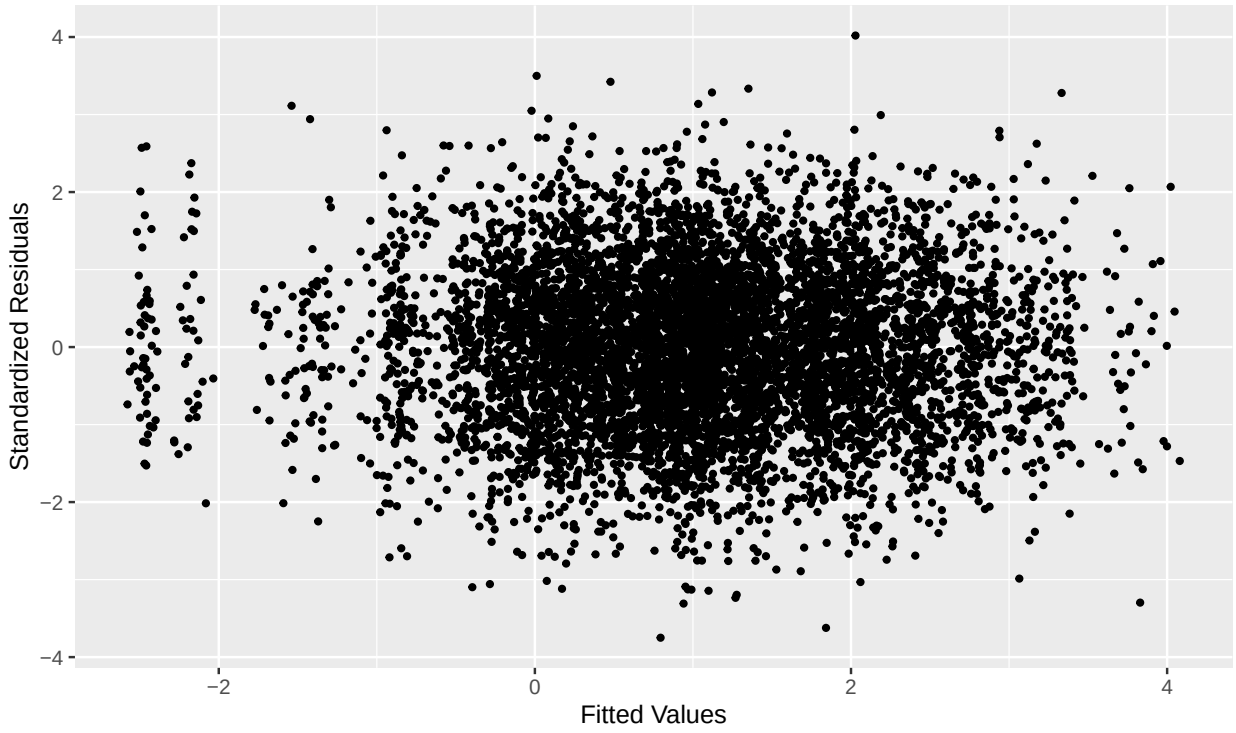
We believe the assumption of independence to be reasonable due to the lack of connection between children. While children living very near each other could potentially be correlated, the fact that our measurements take place indoors would largely negate that relationship. Children from the same household could certainly be correlated, but to our knowledge our dataset does not include any children from the same household. Additionally, while we do not assume independence between measurements of the same child we do claim to have accurately captured the correlation in our ARMA correlation structure. This is evident in the correlation plot of the decorrelated residuals shown below.



We checked the assumption of normally distributed residuals by creating a histogram of the standardized residuals. While the histogram is not perfectly normal in shape it is concentrated in the center and tapers at the edges. Additionally, there are no evident outliers, though there is slight concern for the long lower tail. We also performed a Kolmogorov-Smirnov hypothesis test for normality. Given our p-value of 0.1161 we failed to reject the null hypothesis, therefore concluding that the standardized residuals came from a normal distribution.



The final assumption of constant variance in our residuals was checked by plotting the standardized residuals against the standardized residuals against the fitted values from our model. The scatterplot of standardized residuals against the fitted values shows constant variance.



We calculated our coefficient of determination, R^2 as 0.91, indicating that 91% of the variation in mean log aerosol PM measurement can be explained using the log of the stationary PM measurement, activities, child IDs, and interactions included in our model. The root mean square error is 0.321. This is moderately small relative to our data, which confirms that our model fits the data well.

Section 4: Analysis Results

After creating a model with only the Stationary measurement as a predictor, the resulting R^2 value was 0.00002, which can be interpreted that about 0.002% of the variability in the log of the aerosol PM measurements can be explained by the log stationary PM measurements. This is exceptionally bad model performance so we would say that log of stationary PM does not do a good job of explaining log aerosol PM measurements on its own.

In order to test if the effects of activities were different for different children, we fit 2 models: 1 with interactions between child and activity and 1 without interactions. We then performed an ANOVA test for the difference between the models and found that the p-value was less than 0.0001, which meant that interactions had a significant effect for each child. (See histograms) From observing the distributions of the effects of each activity for every child, we found that most of the activities had a unimodal, roughly bell-shaped distribution. In terms of variance, the effect of the log(stationary) was the smallest with a standard deviation of 0.139, and the highest variability was from the effect of using a computer with a standard deviation of 1.156, and all other activities had standard deviations between 0.45 to 0.725.

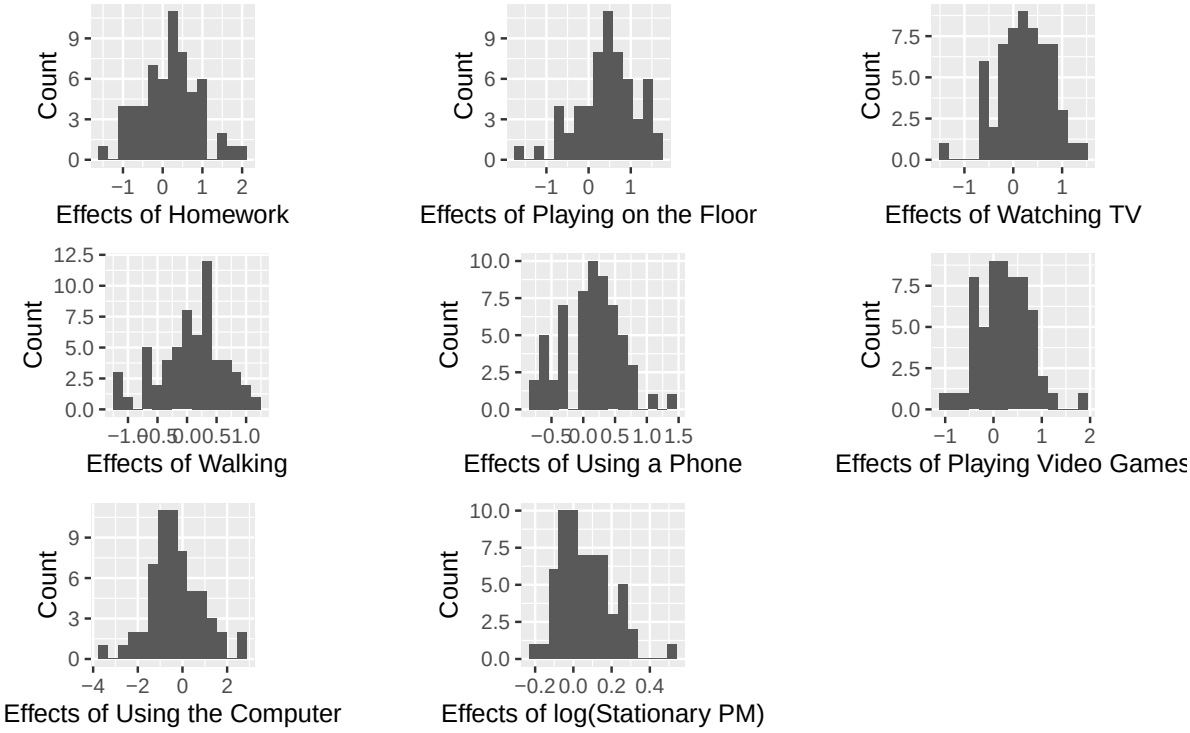


Table 1: Summaries of Effect Distributions

	Mean	Standard Deviation
Homework	0.180	0.724
Floor	0.438	0.700
TV	0.223	0.526
Walking	0.071	0.544
Phone	0.120	0.460
Furniture	0.141	0.577
Video Games	0.211	0.527
Computer	-0.331	1.156
Log(Stationary)	0.061	0.139

Although the stationary measurement alone does not explain PM exposure very well, we found that our model, which includes both log of the stationary measurement and child-specific activities, does a more effective job. We calculated an R^2 value of 0.91 which means that 91% of the variability in the log of the aerosol PM measurement is explained by the log of the stationary PM reading, the different activities, the different children, and the interactions of each child performing each activity. We found that on average, all activities except for time spent on a computer, led to a positive increase in PM exposure. Specifically, the effect of playing on the floor led to the highest vest exposure reading at a 43.8% increase or about a 1.55 PM higher Aerosol measurement on average. The next highest increases were from watching TV and playing video games which both had a greater than 20% increase (>1.22 PM). Contrarily, a child using a computer had around a 33.1% lower exposure or about 1.39 PM less than the average Aerosol count.

Section 5: Conclusions

From our analysis, we found that monitoring the stationary PM concentrations was inefficient in determining the exposure of a child to PM. Additionally, we found that the activities a child engages in impact their PM exposure, although the effects of each activity vary greatly by child. Specifically, children using a computer had widely varying PM exposures, unlike their stationary PM measurements which had much less variation. In our final model we used child ID, log of stationary PM measurement, activity, and interaction terms to account for child-specific variation in the effects of stationary measurements and activities. This model was able to accurately explain most of the variation in the true exposure of a child to particulate matter. We found that all activites except for computer usage led to an increase in PM exposure, with playing on the floor being the most deleterious activity.

We suggest that instead of using one stationary monitor, researchers place multiple stationary monitors in different rooms throughout the house. If this is not possible due to cost restrictions, we recommend that the stationary monitor be moved between rooms at some predetermined time interval.

Appendix A: Analysis Code

```r
# Include packages
library(ggplot2)
library(corrplot)
library(nlme)
library(car)
library(multcomp)
library(magrittr)
library(gridExtra)
library(knitr)

# Include standardized gls residuals function
source("stdres.gls.R")

# Read in the data
pm <- read.table("https://mheaton.byu.edu/docs/files/Stat469/Topics/2%20-%20TemporalCorrelation/3%20-%2
    header = TRUE)

# Convert factor variables to factor type
pm$Activity <- factor(pm$Activity)
pm$ID <- factor(pm$ID)

# Create variables for the log of stationary and the log of
# aerosol
pm$LAerosol <- log(pm$Aerosol)
pm$LStationary <- log(pm$Stationary)

# Scatterplot of log(aerosol) against log(stationary)
p1 <- ggplot(pm, aes(x = LStationary, y = LAerosol)) + geom_point(size = 1) +
    labs(x = "Log(Stationary)", y = "Log(Aerosol)") + theme(aspect.ratio = 1)

# Observe side-by-side boxplots split by activity
p2 <- ggplot(pm, aes(x = Activity, y = LAerosol)) + geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90), aspect.ratio = 1) +
    labs(y = "Log(Aerosol)") + scale_x_discrete(labels = c("Computer",
    "Homework", "Phone", "Floor", "Furniture", "Video Games", "Walking",
    "TV"))

# Combine graphs into a grid
grid.arrange(p1, p2, ncol = 2, nrow = 1)

# Check PM by ID
ggplot(pm, aes(x = ID, y = LAerosol)) + geom_boxplot() + theme(axis.text.x = element_text(angle = 90)) +
```

```
    labs(y = "Log(Aerosol)")

# Fit an MLR with all variables except Minute
pm1.lm <- lm(LAerosol ~ LStationary + Activity + ID, data = pm)

# Store the residuals
resids1 <- resid(pm1.lm)

# Restructure into a 60 X 118 matrix
resid1.mat <- matrix(resids1, 60, 118, byrow = TRUE)

# Evaluate the correlation
corrplot(cor(resid1.mat), method = "color", type = "upper", tl.pos = "n")

# Linearity Assumption
avPlots(pm1.lm, terms = "LStationary")

# Fit GLS model with only LStationary
pmarmastat.gls <- gls(model = LAerosol ~ LStationary, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 2, q = 1), method = "ML")

# Calculate R2
badR2 <- 1 - sum((pmarmastat.gls$fitted - pm$LAerosol)^2)/sum((pm$LAerosol -
    mean(pm$LAerosol))^2)

# Fit GLS model without interactions
pmarmanoint.gls <- gls(model = LAerosol ~ LStationary + Activity +
    ID, data = pm, correlation = corARMA(form = ~1 | ID, p = 2, q = 1),
    method = "ML")

# Fit GLS model
pmarma.gls <- gls(model = LAerosol ~ LStationary + Activity + ID +
    ID:Activity + ID:LStationary, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 2, q = 1), method = "ML")

# Calculate R2
R2 <- 1 - sum((pmarma.gls$fitted - pm$LAerosol)^2)/sum((pm$LAerosol -
    mean(pm$LAerosol))^2)

# Calculate the RMSE
rmse <- (pm$LAerosol - pmarma.gls$fitted)^2 %>% mean() %>% sqrt()

# ANOVA to determine significant difference
```

```
statind.anova <- anova(pmarmastat.gls, pmarma.gls)
noint.anova <- anova(pmarma.gls, pmarmanoint.gls)

# Store the decorrelated residuals
sres <- stdres.gls(pmarma.gls)

# Restructure into a 50 X 4 matrix
sresid.mat <- matrix(sres, 60, 118, byrow = TRUE)

# Evaluate the correlation
corrplot(cor(sresid.mat), method = "color", type = "upper", tl.pos = "n")

# Draw a histogram of the decorrelated residuals
ggplot() + geom_histogram(mapping = aes(x = sres), bins = 10) + xlab("Decorrelated Residuals") +
    ylab("Count")

# Run ks-test for normality
ks <- ks.test(sres, "pnorm")

# Scatterplot of the fitted values vs. decorrelated residuals
ggplot(mapping = aes(fitted(pmarma.gls), sres)) + geom_point(size = 1) +
    xlab("Fitted Values") + ylab("Standardized Residuals")

# Write a function to pull the true effects of each activity
true_effect <- function(ind, pattern) c(coef(pmarma.gls)[ind], coef(pmarma.gls)[ind] +
    coef(pmarma.gls)[grep(pattern, names(coef(pmarma.gls)))])

# Pull the true effects of all of the activities across children
# by combining their interactions with the baseline
homework <- true_effect(3, "ActivityHomework.*ID.*")
floor <- true_effect(5, "ActivityPlayingOnFloor.*ID.*")
watching <- true_effect(9, "ActivityWatching.*ID.*")
walking <- true_effect(8, "ActivityWalking.*ID.*")
phone <- true_effect(4, "ActivityOnPhone.*ID.*")
furniture <- true_effect(6, "ActivityPlayingOnFurniture.*ID.*")
videogames <- true_effect(7, "ActivityVideoGames.*ID.*")

# Pull the computer effect by using the baseline ID effects and no
# interactions since computer is the baseline activity
computer <- c(0, coef(pmarma.gls)[10:68])

# Pull the true effects of LStationary
lstationary <- true_effect(2, "LStationary.*ID.*")
```

```
# Write a function for a quick histogram in ggplot
gghist <- function(x, xlab, bins) ggplot() + geom_histogram(mapping = aes(x = x),
    bins = bins) + xlab(xlab) + ylab("Count") + theme(aspect.ratio = 1)

# Plot the effects for all the explanatory variables across
# children
p1 <- gghist(homework, "Effects of Homework", 15)
p2 <- gghist(floor, "Effects of Playing on the Floor", 15)
p3 <- gghist(watching, "Effects of Watching TV", 15)
p4 <- gghist(walking, "Effects of Walking", 15)
p5 <- gghist(phone, "Effects of Using a Phone", 15)
p6 <- gghist(videogames, "Effects of Playing Video Games", 15)
p7 <- gghist(computer, "Effects of Using the Computer", 15)
p8 <- gghist(lstationary, "Effects of log(Stationary PM)", 15)

# Combine graphs into a grid
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol = 3, nrow = 3)

# Calculate the means and standard deviations Create kable to
# display them with labels
kable(data.frame(c(mean(homework), mean(floor), mean(watching), mean(walking),
    mean(phone), mean(furniture), mean(videogames), mean(computer),
    mean(lstationary)), c(sd(homework), sd(floor), sd(watching), sd(walking),
    sd(phone), sd(furniture), sd(videogames), sd(computer), sd(lstationary)),
    row.names = c("Homework", "Floor", "TV", "Walking", "Phone", "Furniture",
        "Video Games", "Computer", "Log(Stationary)")), caption = "Summaries of Effect Distributions",
    digits = 3, col.names = c("Mean", "Standard Deviation"))
```

Appendix B: EDA and Model Evaluation Code

```
# Color the scatterplot by ID
p2 <- ggplot(pm[pm$ID < 16, ], aes(x = LStationary, y = LAerosol,
    color = factor(ID))) + geom_point(size = 1) + labs(x = "Log(Stationary)",
    y = "Log(Aerosol)", color = "ID") + theme(aspect.ratio = 1)
p3 <- ggplot(pm[pm$ID > 15 & pm$ID < 31, ], aes(x = LStationary, y = LAerosol,
    color = factor(ID))) + geom_point(size = 1) + labs(x = "Log(Stationary)",
    y = "Log(Aerosol)", color = "ID") + theme(aspect.ratio = 1)
p4 <- ggplot(pm[pm$ID > 30 & pm$ID < 46, ], aes(x = LStationary, y = LAerosol,
    color = factor(ID))) + geom_point(size = 1) + labs(x = "Log(Stationary)",
    y = "Log(Aerosol)", color = "ID") + theme(aspect.ratio = 1)
p5 <- ggplot(pm[pm$ID > 45, ], aes(x = LStationary, y = LAerosol,
    color = factor(ID))) + geom_point(size = 1) + labs(x = "Log(Stationary)",
```

```
    y = "Log(Aerosol)", color = "ID") + theme(aspect.ratio = 1)

# Confirmation of no relationship between either PM measurement
# and minute
ggplot(pm, aes(x = Minute, y = LAerosol, color = factor(ID))) + geom_point()
ggplot(pm, aes(x = Minute, y = LStationary, color = factor(ID))) +
    geom_point()

# Observe boxplots by activity for individual children to see
# variability in activity effects
ggplot(pm[pm$ID == 59, ], aes(x = Activity, y = Aerosol)) + geom_boxplot()

# MLR option
pm2.lm <- lm(Aerosol ~ Stationary + Activity, data = pm)

# Storing resids for alternative MLR
resids2 <- resid(pm2.lm)

# Structuring alternative residuals as a matrix
resid2.mat <- matrix(resids2, 60, 118, byrow = TRUE)

# Checking alternative MLR correlation
corrplot(cor(resid2.mat), method = "color", type = "upper", tl.pos = "n")

# Create models with three different correlation structures and
# compare AIC AR1 model
pmar.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corAR1(form = ~1 | ID), method = "ML")

# MA1 model
pmma.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 0, q = 1), method = "ML")

# AR2 model
pmar2.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 2, q = 0), method = "ML")

# MA2 model
pmma2.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 0, q = 2), method = "ML")

# AR3 model
pmar3.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
```

```
    correlation = corARMA(form = ~1 | ID, p = 3, q = 0), method = "ML")

# AR4 model
pmar4.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 4, q = 0), method = "ML")

# ARMA models
pmarma.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 1, q = 1), method = "ML")

pmar2ma.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 2, q = 1), method = "ML")

pmar2ma2.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 2, q = 2), method = "ML")

pmarma2.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 1, q = 2), method = "ML")

# General symmetric model pmgensym.gls <- gls(model = Aerosol ~
# Stationary + Activity, data = pm, correlation = corSymm(form =
# ~1|ID), method = 'ML')

# Compare the AICs
AIC(pmar.gls)
AIC(pmma.gls)
AIC(pmar2.gls)
AIC(pmma2.gls)
AIC(pmar3.gls)
AIC(pmar4.gls)
AIC(pmarma.gls)
AIC(pmar2ma.gls)
AIC(pmar2ma2.gls)
AIC(pmarma2.gls)

# Fit different models that include and exclude interactions to
# determine if they are significant
pmarma1.gls <- gls(model = LAerosol ~ LStationary + Activity, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 1, q = 1), method = "ML")

pmarma2.gls <- gls(model = LAerosol ~ LStationary + Activity + ID,
    data = pm, correlation = corARMA(form = ~1 | ID, p = 1, q = 1),
    method = "ML")
```

```
pmarma3.gls <- gls(model = LAerosol ~ LStationary + Activity + ID +
    ID:Activity + ID:LStationary, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 1, q = 1), method = "ML")

pmarma4.gls <- gls(model = LAerosol ~ LStationary + Activity + ID +
    ID:Activity, data = pm, correlation = corARMA(form = ~1 | ID,
    p = 1, q = 1), method = "ML")

pmarma5.gls <- gls(model = LAerosol ~ LStationary + Activity + ID +
    ID:LStationary, data = pm, correlation = corARMA(form = ~1 | ID,
    p = 1, q = 1), method = "ML")

pmarma6.gls <- gls(model = LAerosol ~ LStationary, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 1, q = 1), method = "ML")

pmarma7.gls <- gls(model = LAerosol ~ LStationary + ID, data = pm,
    correlation = corARMA(form = ~1 | ID, p = 1, q = 1), method = "ML")

pmarma8.gls <- gls(model = LAerosol ~ Activity + ID, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 1, q = 1), method = "ML")

pmarma9.gls <- gls(model = LAerosol ~ LStationary + Activity + ID +
    ID:Activity + ID:LStationary, data = pm, correlation = corARMA(form = ~1 |
    ID, p = 2, q = 1), method = "ML")

# Compare AICs
AIC(pmarma1.gls, pmarma2.gls, pmarma3.gls, pmarma4.gls, pmarma5.gls,
    pmarma6.gls, pmarma7.gls, pmarma8.gls, pmarma9.gls)

# Ensure that model differences are significant
anova(pmarma1.gls, pmarma2.gls)
anova(pmarma2.gls, pmarma3.gls)
anova(pmarma2.gls, pmarma4.gls)
anova(pmarma2.gls, pmarma5.gls)
anova(pmarma4.gls, pmarma3.gls)
anova(pmarma5.gls, pmarma3.gls)
anova(pmarma3.gls, pmarma9.gls)
```