PM Exposure

Mary Curtis and Jillian Warburton

2023-03-24

Section 0: Executive Summary

Section 1: Introduction and Problem Background

Particulate matter (abbreviated as PM) is a mixture of particles that can be found in the air, the majority of which arise as a result of power plants, industries and automobiles. PM exposure can lead to health complications in the heart and lungs, especially for children, so it is important to measure PM exposure of children to reduce future exposures. The goals of this study are to more accurately measure true PM exposure in children. This study attempts to measure true PM exposure by taking the PM measurement of a stationary monitor, the minutes and activities a child was engaged in during an hour, a child's ID number, and compares it to the PM measurement of a vest installed with a monitor on the shoulder (near where a child's mouth would be). We hope to discover if a stationary PM measurement alone can explain a child's PM exposure, if activities explain more of a child's PM intake than a stationary measurement, if some activities or stationary measurements differ drastically between children, and which activities (on average) lead to a higher PM exposure.

Section 2: Statistical Model

The model for this data set is $\mathbf{y} \sim \mathcal{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{B})$, with those variables expanded below.

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

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

Specifically, this is a autoregressive correlation structure of Order 1 (AR1) constant, general symmetric correlation structure with the time covariate the minute that the child is wearing the vest (`minute`) and the child's ID as the grouping factor (`ID`):

Notation: $\mathbf{y} = \begin{bmatrix} \text{PM exposure}_1 \\ \text{PM exposure}_2 \\ \vdots \\ \text{PM exposure}_{5900} \end{bmatrix}$

where PM exposure_i is the PM measurement on the child's vest.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,900} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,900} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{5202,1} & x_{5202,2} & \dots & x_{5202,900} \end{bmatrix}$$

where $x_{i,1}$ to $x_{i,99}$ = ID of child 2 to 100

where $x_{i,100}$ = log(PM measurement of the stationary monitor)

where $x_{i,101} = \begin{cases} 1 & \text{if the child is doing homework} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,102} = \begin{cases} 1 & \text{if the child is on the phone} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,103} = \begin{cases} 1 & \text{if the child is playing on the floor} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,104} = \begin{cases} 1 & \text{if the child is playing on furniture} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,105} = \begin{cases} 1 & \text{if the child is playing video games} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,106} = \begin{cases} 1 & \text{if the child is walking} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,107} = \begin{cases} 1 & \text{if the child is watching TV} \\ 0 & \text{otherwise} \end{cases}$

where $x_{i,108}$ = The minute the child was wearing the vest (ranges from 0-60)

where $x_{i,109} = x_{i,1}x_{i,100}$, in other words, this term equals the PM measurement of the stationary monitor if the child's ID = 2.

where $x_{i,110} = x_{i,2}x_{i,100}$

...

where $x_{i,207} = x_{i,99}x_{i,100}$

where $x_{i,208} = x_{i,1}x_{i,101}$, in other words, this term equals 1 if child's ID = 2 and if the current activity is doing homework.

where $x_{i,209} = x_{i,2}x_{i,101}$

...

where $x_{i,306} = x_{i,99}x_{i,101}$

where $x_{i,307} = x_{i,1}x_{i,102}$, in other words, this term equals 1 if child's ID = 2 and if the current activity is being on the phone.

where $x_{i,308} = x_{i,2}x_{i,102}$

...

where $x_{i,405} = x_{i,99}x_{i,102}$

(This trend continues for $x_{i,103}$ to $x_{i,107}$)

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{\text{ID2}} \\ \beta_{\text{ID3}} \\ \vdots \\ \beta_{\text{ID100}} \\ \beta_{\text{log(Stationary)}} \\ \beta_{\text{Homework}} \\ \beta_{\text{OnPhone}} \\ \beta_{\text{PlayingOnFurniture}} \\ \beta_{\text{VideoGames}} \\ \beta_{\text{Walking}} \\ \beta_{\text{WatchingTV}} \\ \beta_{\text{Minute}} \\ \beta_{\text{ID2:log(Stationary)}} \\ \beta_{\text{ID3:log(Stationary)}} \\ \vdots \\ \beta_{\text{ID100:log(Stationary)}} \\ \beta_{\text{ID2:Homework}} \\ \beta_{\text{ID3:Homework}} \\ \vdots \\ \beta_{\text{ID100:Homework}} \\ \beta_{\text{ID2:OnPhone}} \\ \vdots \\ \beta_{\text{ID2:PlayingOnFloor}} \\ \vdots \\ \beta_{\text{ID2:PlayingOnFurniture}} \\ \vdots \\ \beta_{\text{ID2:VideoGames}} \\ \vdots \\ \beta_{\text{ID2:Walking}} \\ \vdots \\ \beta_{\text{ID2:WatchingTV}} \end{bmatrix}.$$

Section 3: Model Validation

Section 4: Analysis Results

Section 5: Conclusions

Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE, fig.align = 'center')
library(tidyverse) #for ggplot, dplyr, and magrittr
library(GGally) #for ggpairs
library(car) #for variance inflation factors & added-variable plots
library(multcomp) #for generalized linear hypothesis test
library(RColorBrewer) #to color graphs with more accessible color palette
```

```r
library(nlme) #for generalized least squares function gls()
library(DataExplorer) #for correlation plot
library(mvtnorm) #for dmvnorm function in iterative optimization
#source("/cloud/project/stdres.gls.R") #Dr. Heaton's package for gls prediction
source("~/R programming/STAT_469/stdres.gls.R")
options(scipen = 5) #for reducing scientific notation
set.seed(29) #for reproducibility
#load dataset
#data <- read.table("/cloud/project/BreathingZonePM.txt", header=TRUE) %>%
data <- read.table("~/R programming/STAT_469/Unit3/BreathingZonePM.txt", header=TRUE) %>%
  mutate(
    ID = as.factor(ID),
    Activity = as.factor(Activity)
  )
#End of project's code
ggplot(data = data, mapping = aes(x = Aerosol, y = Stationary)) +
  geom_point() +
  labs(title = "PM Measurements of the Child's Vest and the Stationary Monitor",
       y = "PM Measurement of the Stationary Monitor",
       x = "PM Measurement on the Child's Vest")

# interesting... the scale for stationary is much smaller than that of aerosol, implying that often the

ggplot(data = data, mapping = aes(x = log(Aerosol), y = log(Stationary))) +
  geom_point() +
  labs(title = "PM Measurements of the Child's Vest and the Stationary Monitor",
       y = "PM Measurement of the Stationary Monitor",
       x = "PM Measurement on the Child's Vest")

data %>%
  ggplot(mapping = aes(x = Activity, y = Aerosol)) +
  geom_boxplot() +
  labs(
    title = "PM Measurements on the Child's Vest by Activity",
    y = "PM Measurement on the Child's Vest",
    x = "Activity"
  ) + theme(axis.text = element_text(size = 5))
# doesn't show that there is a significant difference here... but i bet if you grouped it by the ID you

data %>%
  ggplot(mapping = aes(x = Activity, y = log(Aerosol))) +
  geom_boxplot() +
  labs(
    title = "PM Measurements on the Child's Vest by Activity",
    y = "PM Measurement on the Child's Vest",
    x = "Activity"
  ) + theme(axis.text = element_text(size = 5))
# doesn't show that there is a significant difference here... but i bet if you grouped it by the ID you
data %>%
  filter(
    ID == 50
  ) %>%
  ggplot(mapping = aes(x = Minute, y = Aerosol)) +
```

```
  geom_point()

# clear correlation between time and PM measurement
data %>%
  filter(
    ID == 25
  ) %>%
  ggplot(mapping = aes(x = Minute, y = log(Stationary))) +
  geom_point()

# fitting an independent linear model to check for correlation in the residuals
data_lm <- lm(data = data,
              formula = log(Aerosol) ~ ID + log(Stationary) + Activity + Minute) # I think we will need
#I think you are right. Trying sqrt, exp, and ^3 didn't fix the issues. I just wish the residuals cente

mean(abs(cor(matrix(data = data_lm$residuals, nrow = 100, byrow = T))))
# mean of the absolute values of the correlation by child from one minute to the next is too high to be
#mean is 0.2428177 with interaction vs. 0.2439345 without interaction

summary(data_lm)
avPlots(data_lm,  ~log(Stationary) + Minute + Activity + Minute:Activity, ask=FALSE)

#looks to be met now
ggplot() +
  geom_density(mapping = aes(x = log(data$Aerosol))) + #we could change log to base=>1, but it won't ce
  xlab('Standardized Residuals') +
  ylab('Frequency') +
  ggtitle('Normality Assumption Check') +
  geom_vline(xintercept = 0, col = "red", linewidth = 0.75)


ggplot() +
  geom_density(mapping = aes(x = log(data$Stationary))) +
  xlab('Standardized Residuals') +
  ylab('Frequency') +
  ggtitle('Normality Assumption Check') +
  geom_vline(xintercept = 0, col = "red", linewidth = 0.75)

#resid_std <- resid(data_lm)
shapiro.test(data_lm$residuals[1:4999])
shapiro.test(data_lm$residuals[901:5900])
#it kept saying "Error in shapiro.test(resid_std) : sample size must be between 3 and 5000"

#we could change log to include base=>5, but it won't center over 0
#we could change log to include base=>5, but it won't center over 0
ar1 <- gls(model = log(Aerosol) ~ ID + log(Stationary) + Activity + Minute,
    data = data,
    correlation = corAR1(form = ~Minute|ID),
    method = "ML")

AIC(ar1) # much better!
#5429.446 with interaction vs 5434.073 without interaction Minute:Activity
AIC(data_lm)
```

```r
#12259.09 with interaction vs 12259.09 without interaction
# MA1

ma1 <- gls(model = log(Aerosol) ~ ID + log(Stationary) + Activity + Minute,
    data = data,
    correlation = corARMA(q = 1, form = ~Minute|ID),
    method = "ML")

AIC(ma1)
#8380.247 with interaction vs 8402.891 without interaction
# DONT RUN, this takes FOREVER haha, I think this is the model he said you would have to run overnight
# but we should probably just use the AR1 model

# symm <- gls(model = log(Aerosol) ~ ID + log(Stationary) + Activity + Minute,
#     data = data,
#     correlation = corSymm(form=~1:60|ID),
#     method = "ML")
#
# AIC(symm)
```