Jonah Meherg

Oktaviano Harsono (Vino)

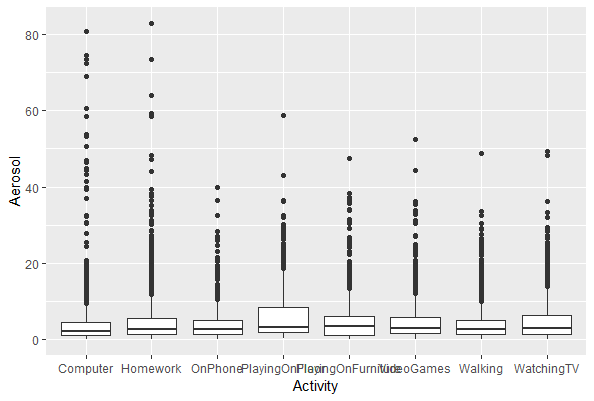
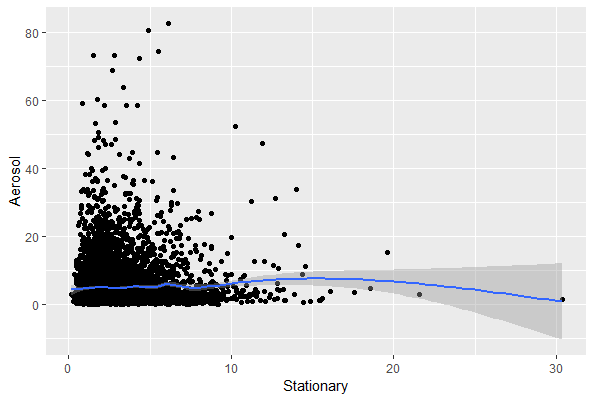
Stat 469

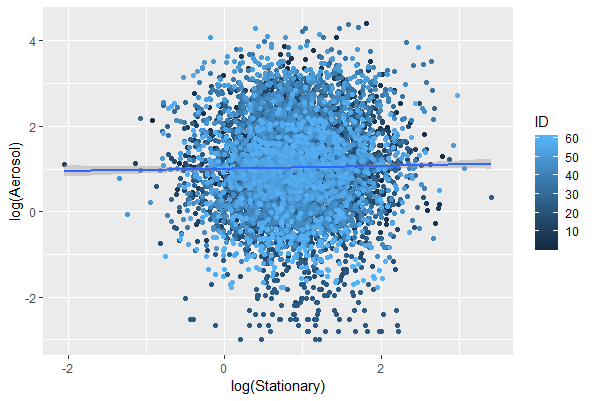
Professor Matthew Heaton

PM Exposure Analysis

Section 1: Introduction and Problem Background

    The problem that we are facing at this time is determining true PM exposure in children. This is to determine if there is a high PM exposure that may be linked to health related problems that affect the heart and lungs. We are interested to see if a stationary measurement gives a good estimate of PM exposure, as well as if different activities have different effects on PM exposure. We are also interested on if the activities and stationary measurement have different effects on PM exposure for different children. After conducting this experiment we collected data for the child ID number, the PM measurement on each child given a measurement vest, the PM measurement of a stationary monitor, the activity the child was participating in, and the minute when the child was wearing the vest for each measurement.





The data set may have many problems. The main problem that needs to be taken into account is any possible lurking variables that come as a result of the differences between each of the children being tested. These may include fitness level of the child, the child’s age, any past health conditions, and dietary habits. Another issue seen in the data is the locations of the houses and possible differences in PM exposure in different rooms of the differing houses. For example a child that may be changing activities and moving across rooms may give a PM exposure measurement very different from the stationary measure if there is not good airflow around the house. The overall data set appears to be heteroskedastic as well.

    The first consequence is that any predictions made with a model that doesn’t take these in to account will be invalid because they will fail to account for modifying their model to create independence among the correlation between the overall times for each child measured, as well as not looking at each child independently of one another, but dependent on themselves. A consequence to not accounting for the, before described, lurking variables is that the predicted values for the PM exposure will be inaccurate, due to not accounting for these possible variables. Accounting for each child individually gives an overall more accurate and justified prediction. It also allows us to determine if activities and stationary measurements have different effects on PM exposure for different children.

To account for these issues we will do a log-transformed longitudinal multiple linear regression analysis to account for temporal correlation for each of the children as well as heteroskedasticity as observed in the original dataset. Adjusting for temporal correlation allows us to assume dependence within each child, but independence overall after the correlation is taken into account, this is able to help with fixing any lurking variables as well as locations, as this will take into account the relationship among the same child over a period of time. Transforming the data allows us to assume our dataset to be linear at the start of the analysis and thus use linear modeling to obtain results from this analysis.

Section 2: Statistical Model

The model used is a log-transformed longitudinal multiple linear regression model using an AR(1) correlation structure. This is represented by

y = Xß + ε

ε ~ *N*(0, σ2B)

where y is an n x 1 vector giving the log response variable of aerosol measurement, X is a n x (p+1) design matrix of explanatory data log-transforming stationary measurement, ß is a (p+1) x 1 vector of coefficients for each explanatory variable and intercept, and ε is the residuals normally distributed with mean 0, and variance using a covariance matrix of σ2B where σ2 is the variance of the error and B is block diagonal with blocks R that follow a general 118 x 118 correlation structure.

    Our model is assuming linearity on account of linearity of data after log-transform of dataset. Independence is accounted for after correlating the residuals and adjusting using a correlation structure. Normality is  assumed because of the normally distributed residuals. Equal variance is assumed because of the variance of residuals being the same across all residuals.

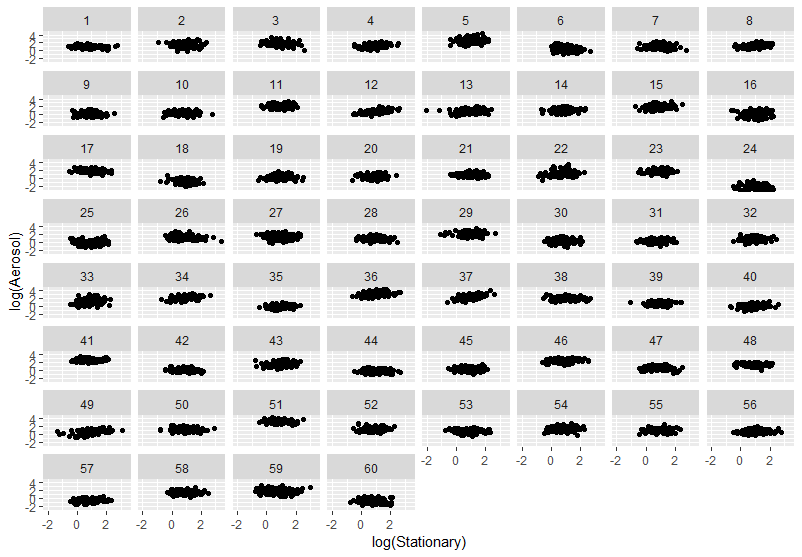
Section 3: Model Validation

             Model df AIC      BIC logLik Test L.Ratio p-value

pm.gls.no.int     1 70 5895.448 6376.001 -2877.724

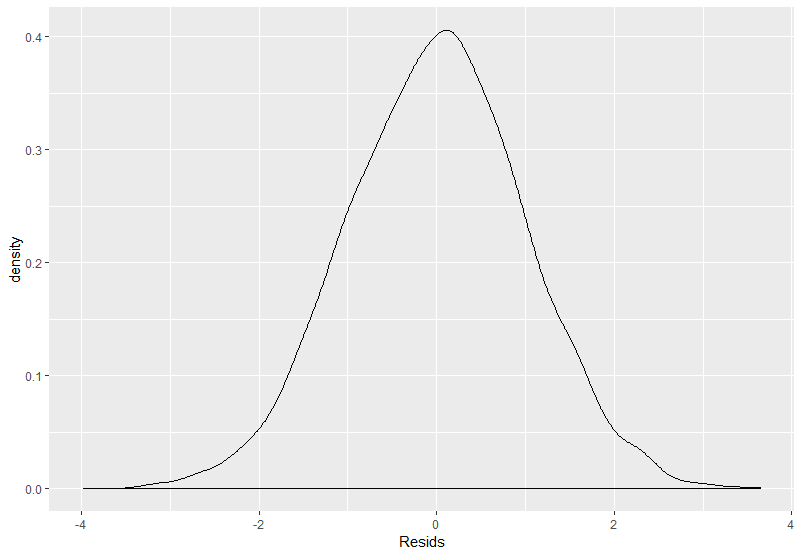
pm.gls            2 542 4056.303 7777.149 -1486.151 1 vs 2 2783.146  <.0001

To account for correlation, we used the AR(1) correlation structure as it captures most of the correlation. The linearity assumption is fulfilled by looking at the relationship between the log stationary and log aerosol by each individual, which shows that the data is linear across each individual tested.



The independence assumption is fulfilled as we decorrelate the residuals making the correlation independent. The matrix is a 60 by 60 matrix with 1’s along the diagonals and correlations on the outside that are close to 0.

The normality assumption is fulfilled by the density plot of the decorrelated residuals and fails the K.S. test with a p-value of 0.5454



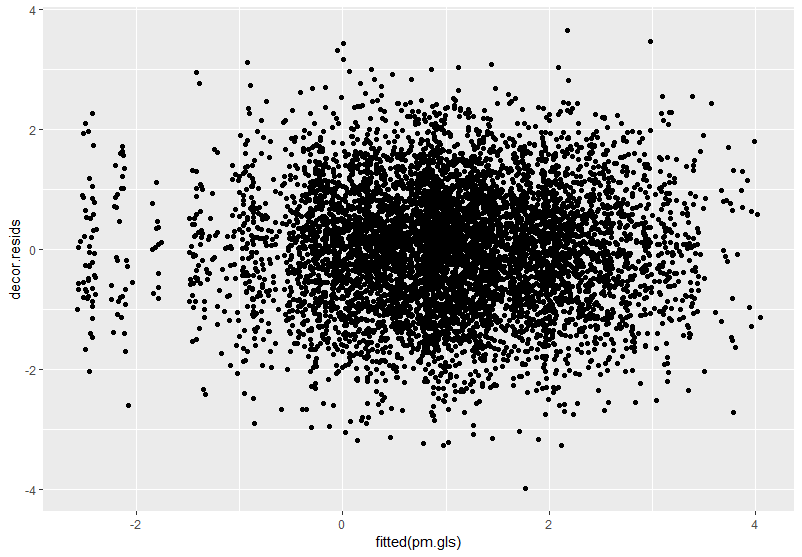
One-sample Kolmogorov-Smirnov test

data:  decor.resids

D = 0.0094985, p-value = 0.5454

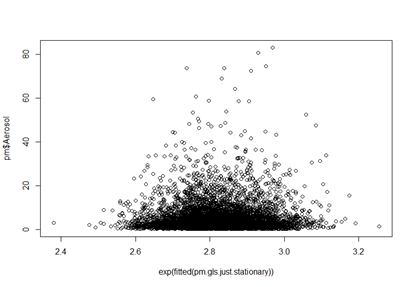
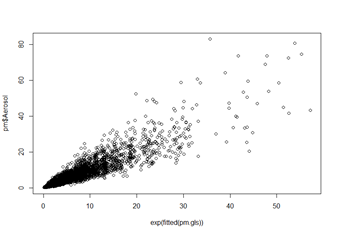
alternative hypothesis: two-sided

The equal variance assumption is fulfilled by the fitted values vs. decorrelated residuals scatter plot.



Section 4: Analysis Results

After analysis of the data, stationary measurement alone does not do a good job explaining PM exposure because the r-squared is very low (0.001121276) in comparison to a full model (0.8458215 ), and the RPMSE is much larger for a model using just the stationary variable (6.87326) as opposed to using a full model including all variables (2.604617)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Df | AIC | BIC | logLik | Test | L.Ratio | p-value |
| Just.stationary (1) | 4 | 6995.334 | 7022.794 | -3493.667 |  |  |  |
| Full.model (2) | 542 | 4056.303 | 7777.149 | -1489.151 | 1 vs 2 | 4015.031 | <.0001 |

Activities are seen, in addition to the stationary measurement, to explain PM exposure. With being on the phone leading to in general higher PM exposure as evidenced by simultaneous general linear hypotheses tests.

Simultaneous Tests for General Linear Hypotheses

Fit: gls(model = log(Aerosol) ~ (log(Stationary) + Activity) \* as.factor(ID),

   data = pm, correlation = corAR1(form = ~Minute | ID), method = "ML")

Linear Hypotheses:

      Estimate Std. Error z value Pr(>|z|)

1 == 0  0.14552    0.02262   6.434 1.24e-10 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Adjusted p values reported -- single-step method)

There are child-specific effects on activities and stationary measurements

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Df | AIC | BIC | logLik | Test | L.Ratio | p-value |
| No.interaction (1) | 70 | 5895.448 | 6376.001 | -2877.724 |  |  |  |
| Interaction (2) | 542 | 4056.303 | 777.149 | -1486.151 | 1 vs 2 | 2783.146 | <.0001 |

Section 5: Conclusions

After this analysis we have concluded that PM exposure is a result of the stationary exposure levels, as well as the activities and person involved. We have concluded that it does not solely depend on the stationary PM exposure levels, but also depends on the child and the activity they are participating in. It was found that being on the phone resulted in generally higher PM exposure levels. A suggested course of action is reducing the child’s time on the phone as a result.

Appendix

library(tidyverse)

library(multcomp)

library(nlme)

library(car)

## Read in the data

pm <- read.delim("https://mheaton.byu.edu/Courses/Stat469/Topics/2%20-%20TemporalCorrelation/3%20-%20Project/Data/BreathingZonePM.txt",

                header = TRUE, sep = "")

## Exploratory plots of relationship between stationary and aerosol

ggplot(pm, aes(x=Stationary, y=Aerosol)) + geom\_point() + geom\_smooth()

ggplot(pm, aes(x=Activity, y=Aerosol)) + geom\_boxplot()

ggplot(pm, aes(x=log(Stationary), y=log(Aerosol))) + geom\_point() + geom\_smooth()

ggplot(pm,aes(x=log(Stationary),y=log(Aerosol)))+geom\_point() +

 facet\_wrap(~ID)#Use log transform

## Fit a regular lm and show there is correlation

pm.lm <- lm(log(Aerosol)~(log(Stationary)+Activity)\*as.factor(ID),data=pm)

summary(pm.lm)

ar1.coefs <- sapply(split(resid(pm.lm), f=pm$ID),

                   function(x){acf(x,plot=FALSE)$acf[2]})

ggplot() + geom\_histogram(aes(x=ar1.coefs))

## Fit a gls model with ar correlation

pm.gls <- gls(log(Aerosol)~(log(Stationary)+Activity)\*as.factor(ID),data=pm,correlation=corAR1(form=~Minute|ID),

             method="ML")

summary(pm.gls)

pm.gls.no.int <-

 gls(log(Aerosol)~log(Stationary)+Activity+as.factor(ID),data=pm,

     correlation=corAR1(form=~Minute|ID), method="ML")

anova(pm.gls.no.int, pm.gls)

## Assumptions Justification

V <- getVarCov(pm.gls)

decor.resids <- c(solve(t(chol(V)))%\*%matrix(resid(pm.gls),ncol=60))

cor(matrix(decor.resids, ncol = 60, byrow=TRUE))

qplot(fitted(pm.gls), decor.resids, geom="point")

ggplot() + geom\_density(aes(x=decor.resids)) + xlab("Resids")

ks.test(decor.resids, "pnorm")

pm.gls.raw <- gls(Aerosol~(Stationary+Activity)\*as.factor(ID),data=pm,correlation=corAR1(form=~Minute|ID),

                 method="ML")

decor.resids <- c(solve(t(chol(V)))%\*%matrix(resid(pm.gls.raw),ncol=60))

qplot(fitted(pm.gls.raw), decor.resids, geom="point")

ar1.coefs <- sapply(split(decor.resids,

                         f=pm$ID),function(x){acf(x,plot=FALSE)$acf[2]})

ggplot() + geom\_histogram(aes(x=ar1.coefs))

## Does stationary alone explain PM exposure?

pm.gls.just.stationary <-

 gls(log(Aerosol)~log(Stationary),data=pm,correlation=corAR1(form=~Minute|ID),

     method="ML")

(pm$Aerosol - exp(fitted(pm.gls)))^2 %>% mean() %>% sqrt()

(pm$Aerosol - exp(fitted(pm.gls.just.stationary)))^2 %>% mean() %>% sqrt()

cor(pm$Aerosol, exp(fitted(pm.gls)))^2

cor(pm$Aerosol, exp(fitted(pm.gls.just.stationary)))^2

plot(pm$Aerosol~exp(fitted(pm.gls.just.stationary)))

plot(pm$Aerosol~exp(fitted(pm.gls)))

## Does any activities, in addition to the stationary measurement, explain PM exposure?

 anova(pm.gls.just.stationary, pm.gls)

act <- "OnPhone"

a <- matrix(0,nrow=length(coef(pm.gls)), ncol=1)

a[str\_detect(names(coef(pm.gls)), act),] <- c(1,rep(1/60,59))

summary(glht(pm.gls, linfct=t(a)))

## Interaction effect?

anova(pm.gls.no.int, pm.gls)