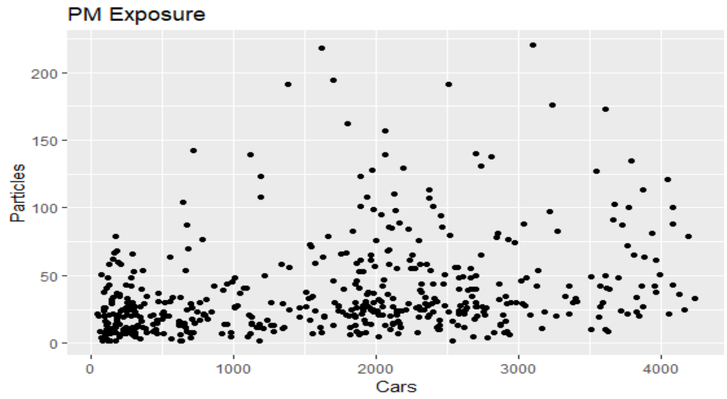
Jake Szendre

Exam 1

1. Section 1: Introduction and Problem Background

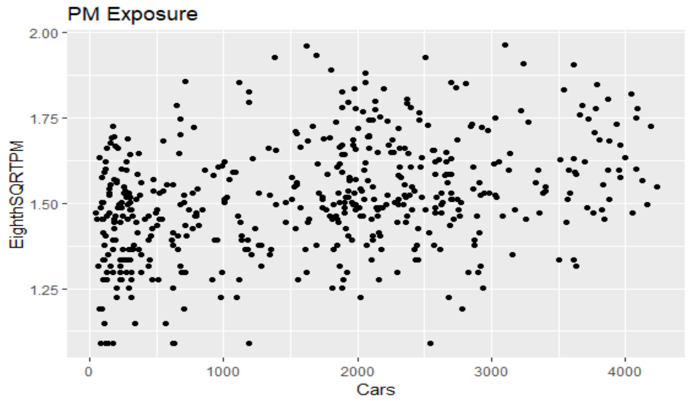
(a) PM particles can be dangerous, and sometimes even fatal, to humans. If the amount of PM particles can be limited in any way, or if we figure out where there is a high concentration of them, we can raise awareness and potentially save lives and prevent diseases from occurring. In this study, we seek to find a relationship between the amount of PM particles in an area and the traffic congestion in the same area. If a relationship can be found, we can then predict the amount of dangerous PM particles will be in a traffic area, and officials can give proper warning/preventative suggestions.

(b) The data in use considers various different traffic intersections (the amount of cars passing through them in a day) and the amount of PM particles in those intersections. On the next page is a graph depicting the data: each dot represents one traffic intersection, the x-value represents the amount of cars that pass through it, and the y-value represents the amount of PM particles found at it. The median average amount of PM particles in this set of data is 27.0, while the mean average is 37.88, telling us that the data is skewed with outlying intersections that have a large amount of PM particles.



(c) Judging off the data as it is, it does not look like an SLR model would fit because there are many outlying data points. The correlation between Cars and Particles for this data set is r = 0.300898, which tells us that the relationship is not very linear. In order to try fix these issues with the data, we will transform it by taking the eighth root of the amount of particles, but leaving the amount of cars to stay the same.

Transforming the data now makes our graph able to use an SLR model, as you can see at the top of the next page. Notice the y-axis is now titled, "EighthSQRTPM," signifying that the y-values have now been taken the eighth root of. The reasons why an SLR model will now be suitable will be explained in the next section, specifically refer to 3(a).



2. Section 2: Statistical Modeling

(a) Statistical model: $(\hat{y}_i)$^(1/8) = $\beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i~ N(0,σ^2)

(b) $(\hat{y}_i)$^(1/8) = Eighth root of the amount of PM particles of the $_i$th intersection

x_i = Amount of cars passing through the $_i$th intersection

$_i$= 1,…,n

β_0= If the amount of cars in the $_i$th intersection is 0, this is the estimated eighth root amount of its PM particles, on average.

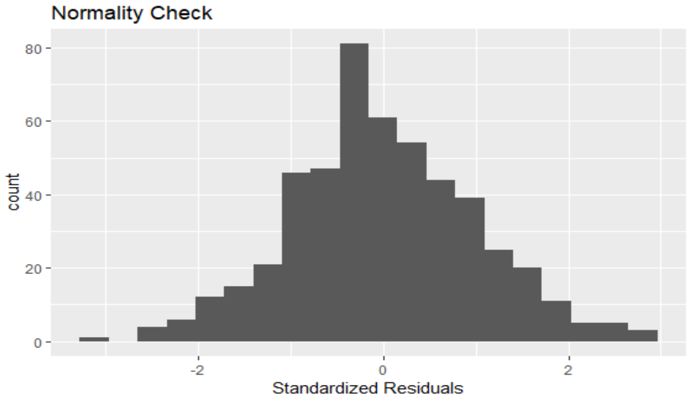
β_1= As the amount of cars in the $_i$th intersection increases by 1 unit, the eighth root PM particles on average goes up by β_1 on average.

ε_i = distance away from predicted measure for the $_i$th value

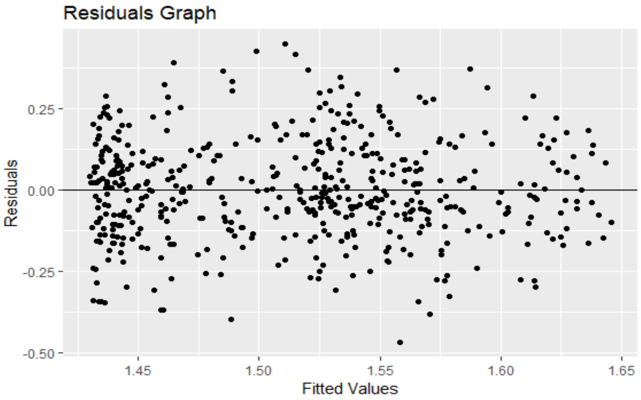
(c) The assumptions used follow the simple acronym, “LINE:” Linearity, Independence, Normality, Equal variance. We assume linearity in the relationship between number of cars in an intersection and the eighth root of PM particles. We assume independent results, i.e. we assume the amount of eighth root of PM particles in one intersection is not influenced by the amount of eighth root of PM particles in another. We assume normality of our data, i.e. we assume that if we were to take repeated samples at intersections each with the same amount of cars, the eighth root of the PM particles found would center at a value and gradually spread out. And finally, we assume equal variance about the best-fitting line. This means that we assume the deviances of the data points away from the best-fitting line are approximately equal across all points.

3. Section 3: Model Verification

(a)

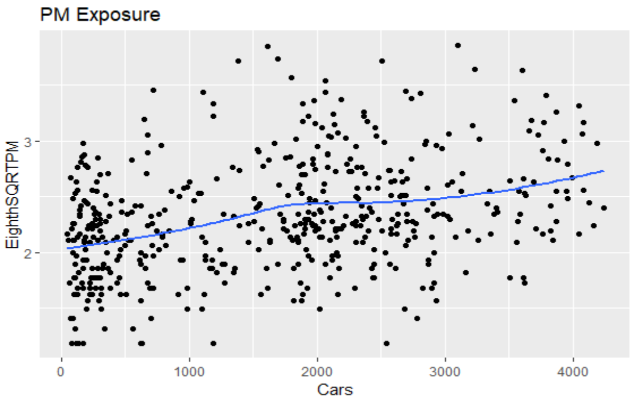


The data points are approximately normally distributed across the line, they center around 0 and there are no extreme outliers. This graph shows no skewness and is shaped like a bell curve. This data passes a KS test for normality, with a p-value of 0.672 (it's normal).



The eighth root of the amount of PM particles are independent, they do not affect the eighth root of PM particles of another intersection. Equal variance is now met, as you can see on the left. The dots deviate approximately the same amount throughout the graph. Our data satisfies a BP test for equal variance with a p-value of 0.8285 (i.e. it passes).

(b) Our adjusted R^2 value came out to be 0.1271, which tells us that only 12.71% of the variability in the amount of the eighth root of PM particles is due to the amount of cars passing an intersection. However, a relatively low R^2 value (ideally it would be above 70%) does not mean the end of the world. It tells us that there are other factors contributing to the eighth root of the amount of PM particles. All in all, we cannot claim that the amount of cars passing through an intersection alone is a good determiner of the eighth root of the amount of PM particles there. There are other factors that need to be considered. Our model fits the data well to an extent.



Now with a least-squares line through the data, we can see the high-variability of our model, explaining the low R^2 level.

Line equation: ŷ^(1/8)=(1.427+ (x)*(5.155*10^-05))

(c) Doing repeated cross-validation exercises to our data, my proposed model consistently reports mean biases of around -8.710925 PM particles. This tells us that my model usually under-predicts the amount of PM particles in an intersection by about 9 PM particles. Doing the same exercises, we also get mean coverages of 0.9396, which tells us that 93.96% of our prediction intervals carry the true mean amount of PM particles. We get mean RPMSE's of about 28.60605, telling us that our predictions were off on average of about 28.61 PM particles in either direction from the best-fitting line. We consistently get mean widths of about 118.21 PM particles, telling us the average mean widths of our prediction intervals. That number is a point of concern because it is a huge margin of error. Overall, our model simply is not adequate at predicting a precise amount of PM particles in a given intersection.

4. Section 4: Results

(a) Because our R^2 value is relatively low at 0.1271, this tells us that the relationship between PM particles and cars is weak. It can still be considered linear, but the relationship is weak. Performing a confidence interval on the true slope (the value where if the amount of cars increases by 1, the amount of eighth root of PM particles that will increase) comes out to be: (3.975145e-05, 6.334853e-05). Again, that is a wide difference, illustrating the unpredictability of using our model.

(b) According to our simple linear regression fitted model, ŷ=(1.427+ (x=1800)*(5.155*10^-05))^8 = 28.46221 PM particles. Creating a prediction interval comes to: (4.771343, 122.9854 PM particles). This interval shows the wide range and uncertainty of amount of particles that could be found at an intersection with 1800 cars passing through. If only 4.77 particles are actually there, the risk for disease is not anywhere near as large as if our prediction was the other side of this range, 122.99 particles. If it's actually 122.99 particles, we may have a large reason for concern.

5. Section 5: Conclusions

(a) In conclusion, the number of cars in an intersection is a factor that needs to be considered when it comes to the amount of PM particles present, however, cars cannot be the only factor considered to accurately predict the amount. To be statistically honest and accurate, the model I proposed earlier should not be used to estimate the amount of PM particles in an intersection. The margin of error is so large that if the prediction swings to one way or the other, and that prediction is accepted by local state officials, mass hysteria could transpire over a high amount of PM particles, when in reality a smaller amount could actually be present.

(b) I would consider other factors that contribute to the amount of PM particles in an intersection, not just number of cars passing through. One might consider temperature in an intersection, population within a certain distance of the intersection, or energy generators nearby, to name a few. Or, if just wanting to strictly study the relationship between PM particles and presence of cars, I would recommend studying different types of cars and their effects on PM particles. Not all are equally detrimental to our planet.

R Code Appendix

```
PM <- read.table("~/Stat 123/Stat 330/Stat 330/PMData.txt", header=TRUE)
scatter.smooth(PM)
SLR2 <- lm(data=PM, sqrt(Particles)~Cars)
library(ggplot2)
ggplot(data=PM,aes(x=Cars,y=Particles))+ggtitle("PM Exposure")+geom_point()
summary(PM)
cor(PM$Particles,PM$Cars)
EighthSQRTPM <- sqrt(sqrt(sqrt(y)))
y <- PM$Particles
ggplot(data=PM,aes(x=Cars,y=EighthSQRTPM))+ggtitle("PM Exposure")+geom_point()+geom_smooth(se=FALSE)
library(MASS)
library(lmtest)
cor(EighthSQRTPM,PM$Cars)
SLR <- lm(data=PM, EighthSQRTPM ~ Cars)
bptest(SLR)
ks.test(Resid, "pnorm")
Resid <- stdres(SLR)
ggplot()+geom_histogram(aes(Resid), bins=20)+xlab("Standardized Residuals")+ggtitle("Normality Check")
ggplot()+geom_point(aes(SLR$fitted.values,SLR$residuals))+geom_hline(yintercept = 0)+xlab("Fitted Values")+ylab("Residuals")+ggtitle("Residuals Graph")
summary(SLR)
#cross validation#
n.cv <- 500
bias <- rep(NA,n.cv)
rpmse <- rep(NA,n.cv)
cvg <- rep(NA,n.cv)
wid <- rep(NA,n.cv)
n.test <- 5
for(cv in 1:n.cv){
#Split into test and training sets#
  test.obs <- sample(1:nrow(PM),n.test)
  test.set <- PM[test.obs,]
  train.set <- PM[-test.obs,]
  #Fit an LM using Training Data only#
  train.lm <- lm(sqrt(sqrt(sqrt(Particles)))~Cars,data=train.set)
  #Fit model, get predicitons and prediciton intervals#
  pred <- predict.lm(train.lm,newdata=test.set,interval="prediction")^8
  #Calculate Results#
  bias[cv] <- mean(pred[,'fit']-test.set[,'Particles'])
  rpmse[cv] <- sqrt(mean((pred[,'fit']-test.set[,'Particles'])^2))
  cvg[cv] <- mean(pred[,'lwr'] < test.set[,'Particles'] & pred[,'upr'] > test.set[,'Particles'])
  wid[cv] <- mean(pred[,'upr']-pred[,'lwr'])
}
mean(bias)
mean(cvg)
mean(rpmse)
mean(wid)
#Model y=1.427e+00 + 5.155e-05x
df <- data.frame(Cars=1800)
predict.lm(SLR, newdata = df, interval = "prediction", level = .95)
confint(SLR)
```