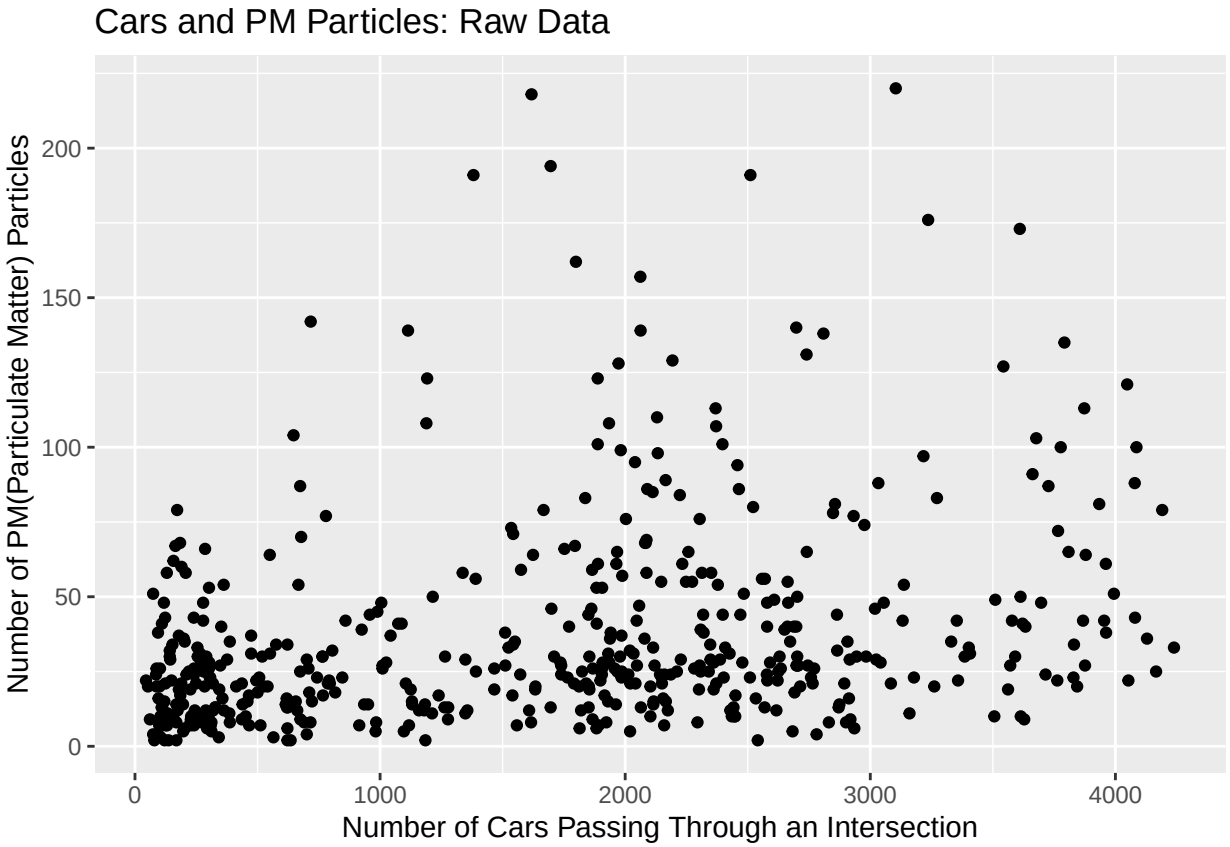
Midterm Analysis #1 - Pollution

Jillian Maw

2/18/2022

Section 1: Introduction and Problem Background

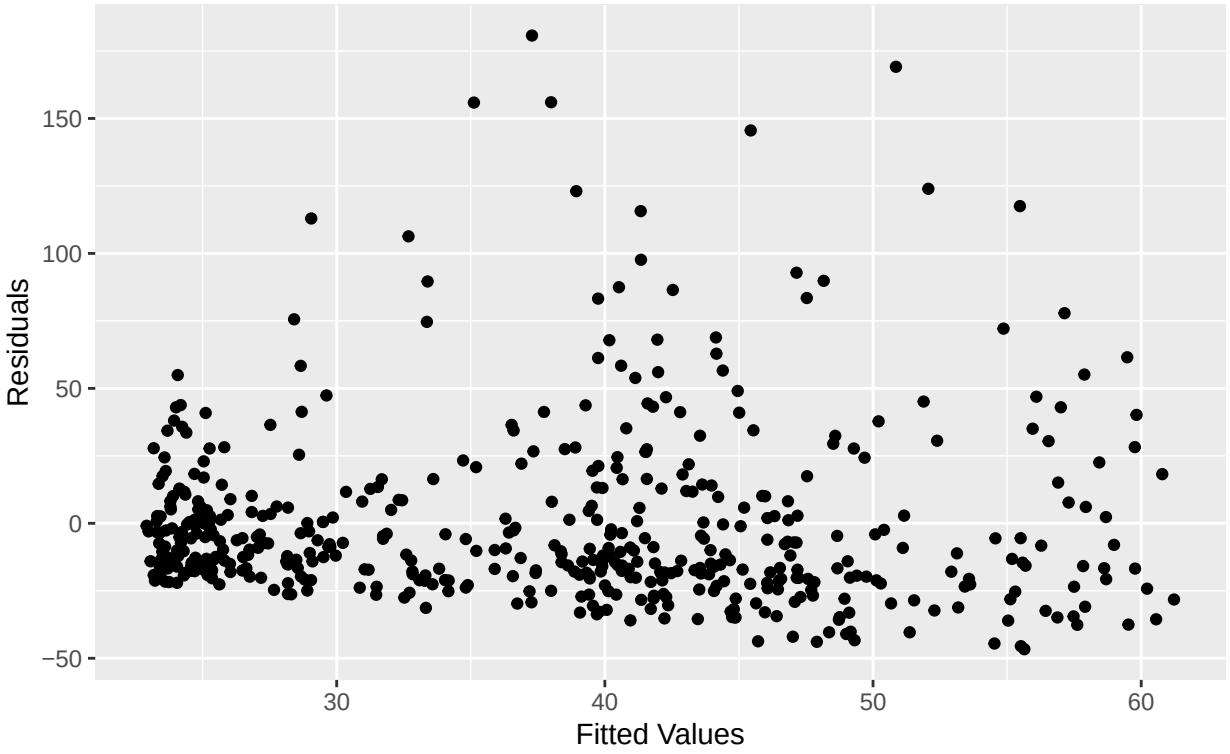
In this report, I will be studying a facet of the relationship between human activity and particulate matter particles, focusing on the number of cars passing an intersection and the number of particulate matter particles in the air near said intersection. Particulate matter, also known as particle pollution or PM, can lead to a variety of ill health effects in both the long and short term, so watching this variable can be useful for monitoring possible health issues in the community. Determining if there is a relationship between the number of cars in an intersection and the number of particulate matter particles would be beneficial to know if the traffic on an intersection can be used to predict the number of particulate matter particles. I will determine this by using simple linear regression modeling on the data.
To do a simple linear regression model, I will be checking for a linear equation that can model the relationship between the number of cars passing through an intersection and the number of particulate matter particles in the air near the intersection. I will be checking the correlation between the two variables and show a scatterplot of the observations from the data. The data has two variables, number of cars and number of particulate matter particles. The number of cars ranges from 45 to 4239 cars, with an average of 1683.222 cars and a median of 1851.5 cars. The number of particulate matter particles ranges from 2 to 220 particles, with an average of 37.878 particles and a median of 27 particles. The correlation between the number of cars and the particulate matter particles is 0.3009 out of 10, which is very weak. Below is a graph of the data. It can be seen that there is a lot of variation in the observations, which suggests a simple linear regression model on the raw data is not best suited.

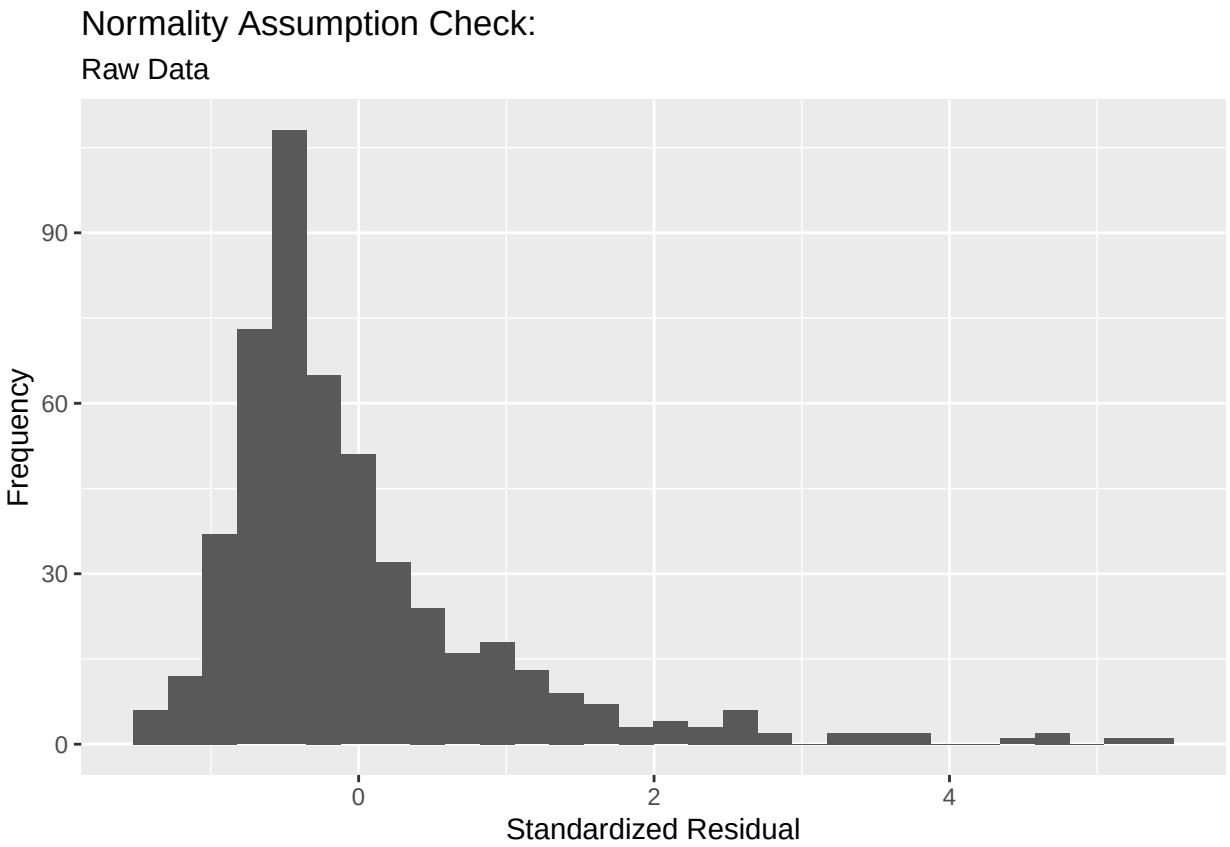


As will be seen shortly, a simple linear regression model would not be appropriate for this data, because the data is not independent, following a Normal distribution, nor have an equal variance of observations from a linear regression model. The independence of the data cannot be assumed because the observation of particulate matter particles is clearly affected by the number of cars from previous days, and tends to linger between observations, therefore, we cannot assume each observation is independent of the other. Data was also not collected consecutively or randomly, which could affect the strength of the relationship. Below are two graphs, the first showing the lack of equal variance around a mean of 0, and the second showing the skew of the data to the left, making it ineligible for the Normal distribution.

Linearity and Equal Variance Assumption Checks:

Raw Data





Below, I have the results for the One-sample Kolmogorov-Smirnov test, also called a KS-test, and the Jarque-Bera test for normality, also called a JB-test, which conducts hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution is skewed. I set the p-value to be 0.05, to prove significance. Because the One-sample Kolmogorov-Smirnov test rejects the null hypothesis and the studentized Breusch-Pagan test rejects the null hypothesis, as seen in the test results below, I accept that I have not met the Normal assumption for the data yet.

I then have a statistical test for checking the Equal Variance assumption for the model. Checking the scatterplot of fitted values versus residuals above, it appears that I have a model with linearity but mostly unequal variance, and can only proceed with the Breusch-Pagan test, or BP-test, with caution to confirm equal variance. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis assumes that the data has homoskedasticity, while the alternative hypothesis assumes that the data has heteroskedasticity. I set the p-value to be 0.05, to prove significance. The code results below shows the Breusch-Pagan test produces a p-value that rejects the null hypothesis, so I accept that I do not have the Equal Variance assumption met for the data.

Finally, I will check the data for outliers that could affect the Normal distribution assumption further. I used Cook's formula to check how much the data set's simple linear regression is affected by each individual point. The use of this formula reveals 22 possible outliers. I could flag those points for future consideration, but due to the failing of the earlier KS-test and JB-test, I will leave the points alone for this report and instead focus on how to meet the assumptions mentioned previously.
Instead, I will provide a suitable transformation of the data in the next section of the report, that will address the Normal distribution and the equal variance assumptions needed for a linear model equation. I do not know how to address the lack of independence is a data set yet, but I will note that it can affect the strength of the relationship in the data.

Section 2: Statistical Modeling

Below is the simple linear regression model I will use to analyze the relationship between the number of cars that pass through an intersection and the number of particulate matter particles in the air. I will transform the data as shown to achieve a linear regression:

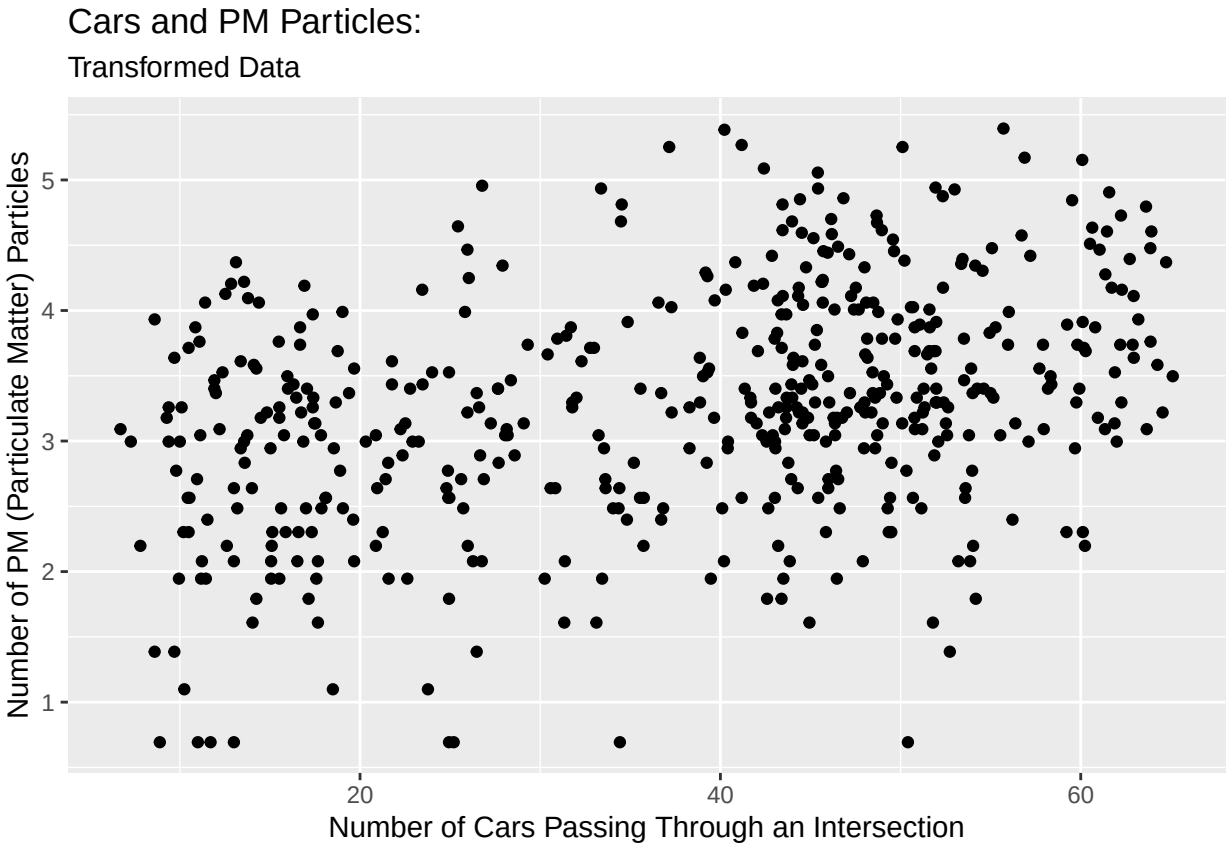
$$\log(y_i) = \beta_0 + \beta_1\sqrt{x_i} + \epsilon_i, \text{ where } \epsilon_i \overset{iid}{\sim} N(0,\ \sigma^2)$$

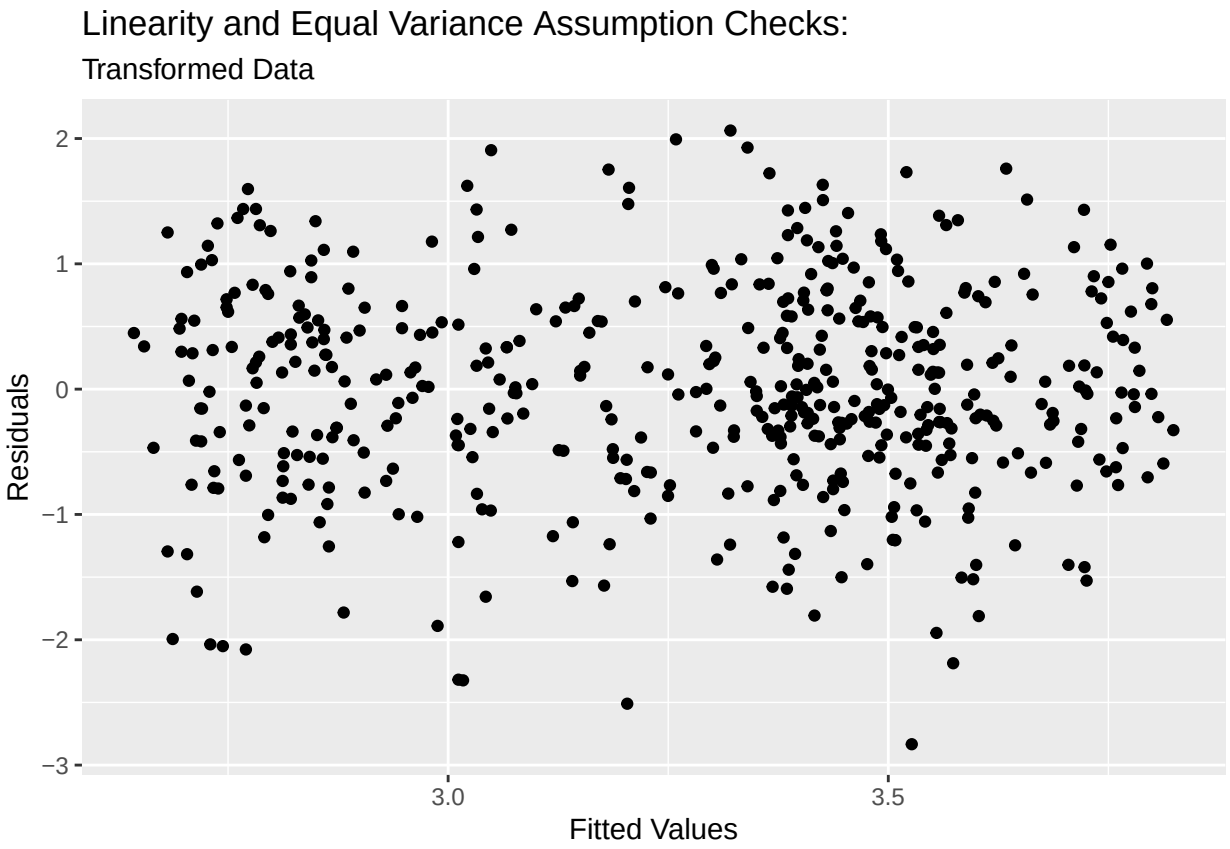
Above, $\log(y_i)$ represents the response variable, the number of particulate matter (PM) particles, at a given observation i. There are a total of 500 observations used to create this simple linear regression model. The $\sqrt{x_i}$ represents the explanatory variable number of cars passing through an intersection, at a given observation i, which we are using to explain the response variable of the number of particulate matter particles (using statistical modeling). β_0 represents the intercept coefficient, which says when $\sqrt{x_i}$, the number of cars passing through an intersection, is 0, the average $\log(y_i)$, the number of particulate matter particles, is the intercept coefficient. Another way to think of this is the base amount of pollution present at the intersection is $beta_0$ particles of particulate matter. The slope coefficient β_1 states that as the explanatory variable $\sqrt{x_i}$ increases, the average response variable $\log(y_i)$ increases by the slope coefficient. In this model's case, it means that as the square root of the number of cars passing through an intersection increases by 1, the average log of the number of particulate matter particles also increases by β_1 particles. The ϵ_i represents the residual errors, or the difference from the true average of particulate matter. The symbol $\overset{iid}{\sim}$ means "independent and identically distributed", which means I assume the model meets two of the assumptions needed for simple linear. The N is short for Normal distribution, meaning the simple linear regression model's residuals (or, the difference between the real number of particulate matter particles and the predicted number of particulate matter particles) follow a Normal distribution's shape and behaviors, standardized at a mean of 0 and a standard deviation of σ^2. The symbol σ^2 represents the variance of the data around the regression line fitted to the data by this model. Another way to think of the variance is that it is the square of the standard deviation. The standard deviation shows that for any x_i, 99.7% of the response variables will be within 3 standard deviations of the regression line made by $\beta_0 + \beta_1\sqrt{x_i}$, the intercept coefficient plus the product of the slope intercept and the square root of the explanatory variable. The above model gives us the following numerical representations of the model: $\log(y_i) = 2.5074 + 0.0202\ \sqrt{x_i} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,\ \ 0.6824)$.

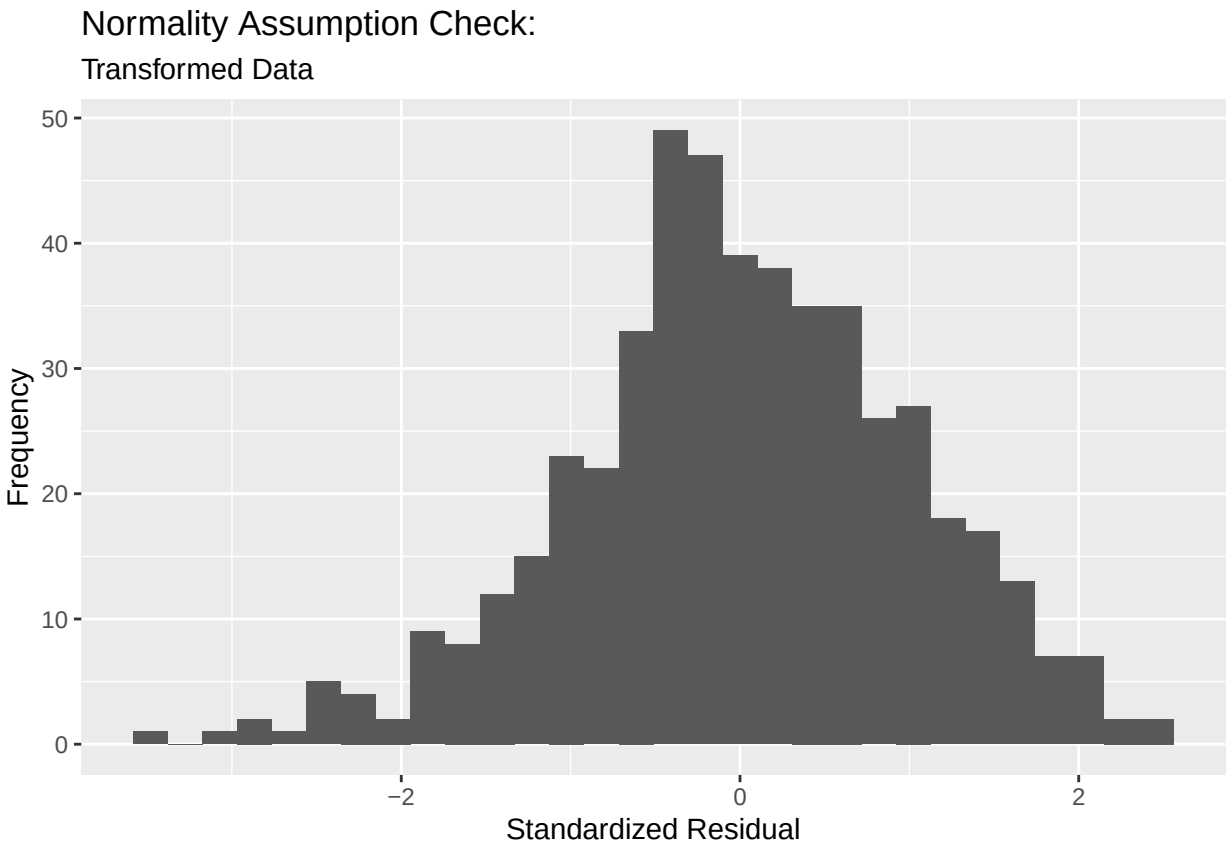
In the above model, we assume we are able to meet the assumptions of linearity, independence, equal variance, and normality to create a simple linear regression, which we will explain, with the exception of independence due to aforementioned reasons, in the section below.

Section 3: Model Verification

Below are two graphs, the first showing the new scatterplot of the transformed data, the second showing the equal variance of the observations' standard residuals (the difference between an observed and a predicted value) around a mean of 0, and the third showing the spread of residuals in a Normal distribution. These graphs show that the data is now eligible for the equal variance and Normal distribution assumptions, and already met the linearity assumption.


Cars and PM Particles:
Transformed Data
Number of PM (Particulate Matter) Particles
5
4
3
2
1
20
40
60
Number of Cars Passing Through an Intersection


Linearity and Equal Variance Assumption Checks:
Transformed Data
Residuals
2
1
0
−1
−2
−3
3.0
3.5
Fitted Values

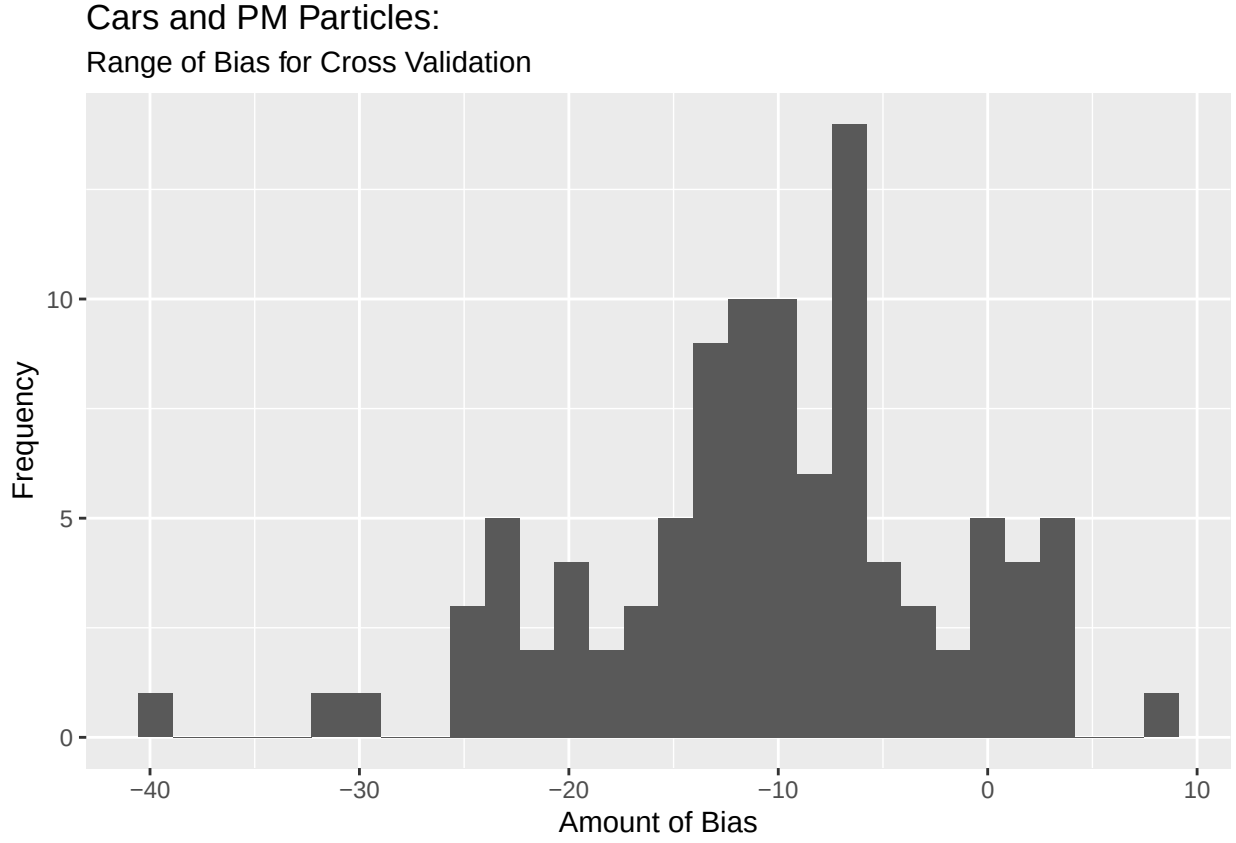


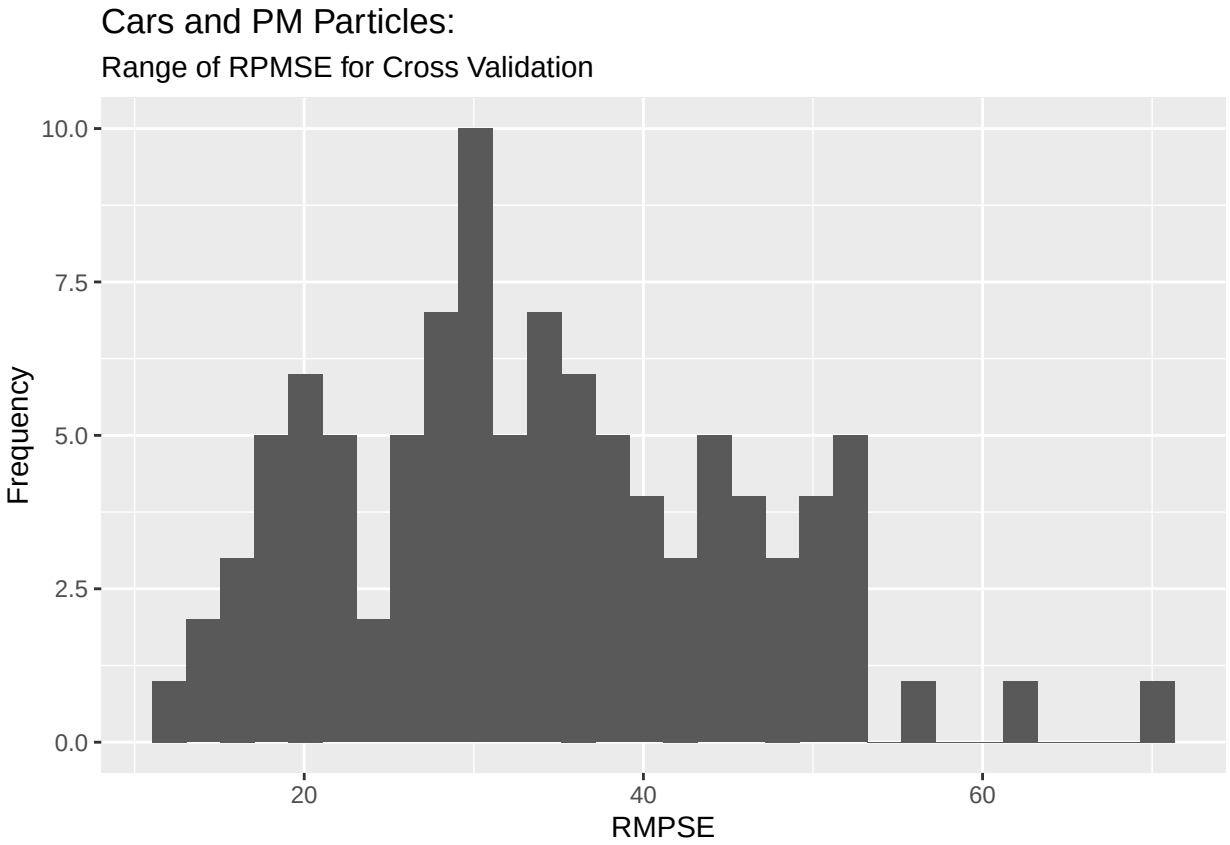
Below, we will see the results of the statistical tests for normality. As explained earlier in Section 2, the One-sample Kolmogorov-Smirnov test and the Jarque-Bera test for normality conduct hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution is skewed. I set the p-value to be 0.05, to prove significance. Because the One-sample Kolmogorov-Smirnov test fails to reject the null hypothesis and the studentized Breusch-Pagan test fails to reject the null hypothesis, as seen in the test results below, I accept that I have met the Normal assumption for the data.

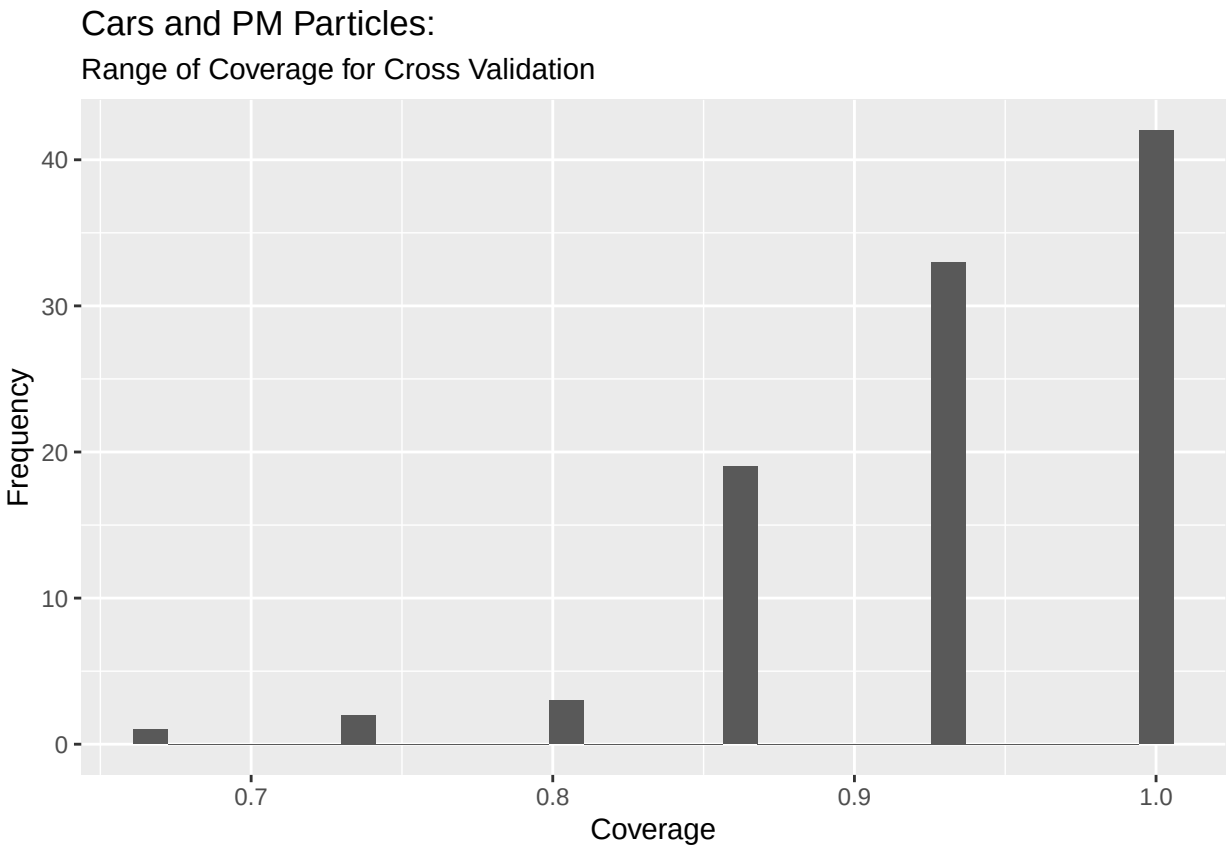
I then have a statistical test for checking the Equal Variance assumption for our model. Checking the scatterplot of fitted values versus residuals above, it appears that we have a model with linearity and mostly equal variance now and can proceed with the Breusch-Pagan test to confirm equal variance. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis assumes that the data has homoskedasticity, while the alternative hypothesis assumes that the data has heteroskedasticity. We set the p-value to be 0.05, to prove significance. The code results below shows the Breusch-Pagan test produces a p-value that fails to reject the null hypothesis, so we accept that we have the Equal Variance assumption met for our data.

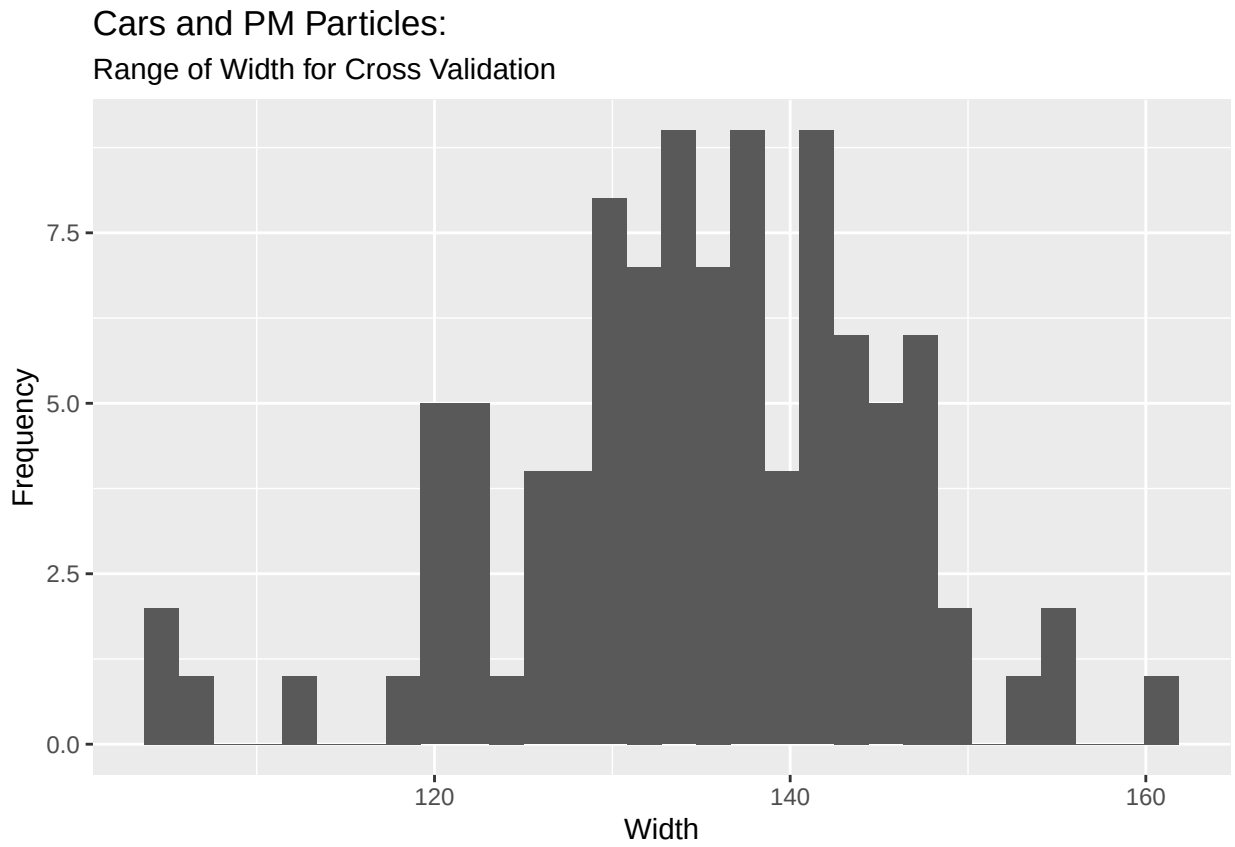
Finally, I can check the data for outliers that could affect the Normal assumption of the data set. Using Cook's formula to check how much the data set's simple linear regression is affected by each individual point, I have flagged the 29 observations of outliers for future consideration, but due to the passing of the earlier KS-test and JB-test, I will leave the points alone for this report.
Based on the transformed data meeting the Normal distribution assumption and the equal variance assumption now, I think my data fits the model well now. I also have measurements of the percent of variation in the number of particulate matter particles that is explained away by number of cars, called R^2. For the raw data, the R^2 is 0.0905. For the transformed data, the R^2 is 0.1341, a small improvement from the first graph.

I conducted a cross validation procedure 100 times on 15 randomly selected observations of the data, newly selected each time, and calculated an average bias of -10.5425. This means our predictions are, on average, lower than the true average number of particulate matter particles. Considering how skewed to the left the original data was on the first scatterplot, a higher bias in these cross validation procedures are understandable. I also calculated an average Root Predictive Mean Square Error of 33.8644, which means the predictions are off, on average, 33.8644 particulate matter particles. Considering the range of the number of particulate matter particles present in observations, this seems a reasonable amount of error. To see how far the predictions ranged, I calculated the width to be 134.83 particulate matter particles, on average. In addition, the coverage, or the percentage of prediction intervals that contain the true average number of particulate matter particles, to be 0.938. Below are graphs showing in greater detail the results of the cross validation procedures.




Cars and PM Particles:
Range of RPMSE for Cross Validation
Frequency
10.0
7.5
5.0
2.5
0.0
20
40
60
RMPSE


Cars and PM Particles:
Range of Coverage for Cross Validation
Frequency
40
30
20
10
0
0.7
0.8
0.9
1.0
Coverage



Section 4: Results

With simple linear regression models, we want to perform a hypothesis test on β_1 (explained earlier as the square root of the number of cars passing through an intersection increasing by 1 as the average log of the number of particulate matter particles also increases by 0.0202 particles) to test whether or not there is a linear relationship between our explanatory and response variables. To see if there is a relationship between the number of cars passing through an intersection and the number of particulate matter particles, we create two opposing hypotheses to test. The null hypothesis is that β_1 is 0, and therefore proves there is not a linear relationship. The alternative hypothesis is that $beta_1$ is *not* 0, and therefore proves there *is* a linear relationship. We will set the p-value (the probability of obtaining hypothesis test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct) at 0.05, a widely accepted academic standard. If the p-value is less than 0.05, we can conclude that the number of cars passing through an intersection has a statistically significant, linear effect on the number of particulate matter particles. The null and alternate hypothesis are mathematically:

$$H_0 : \ \beta_1 = 0 H_a : \ \beta_1 \neq 0$$

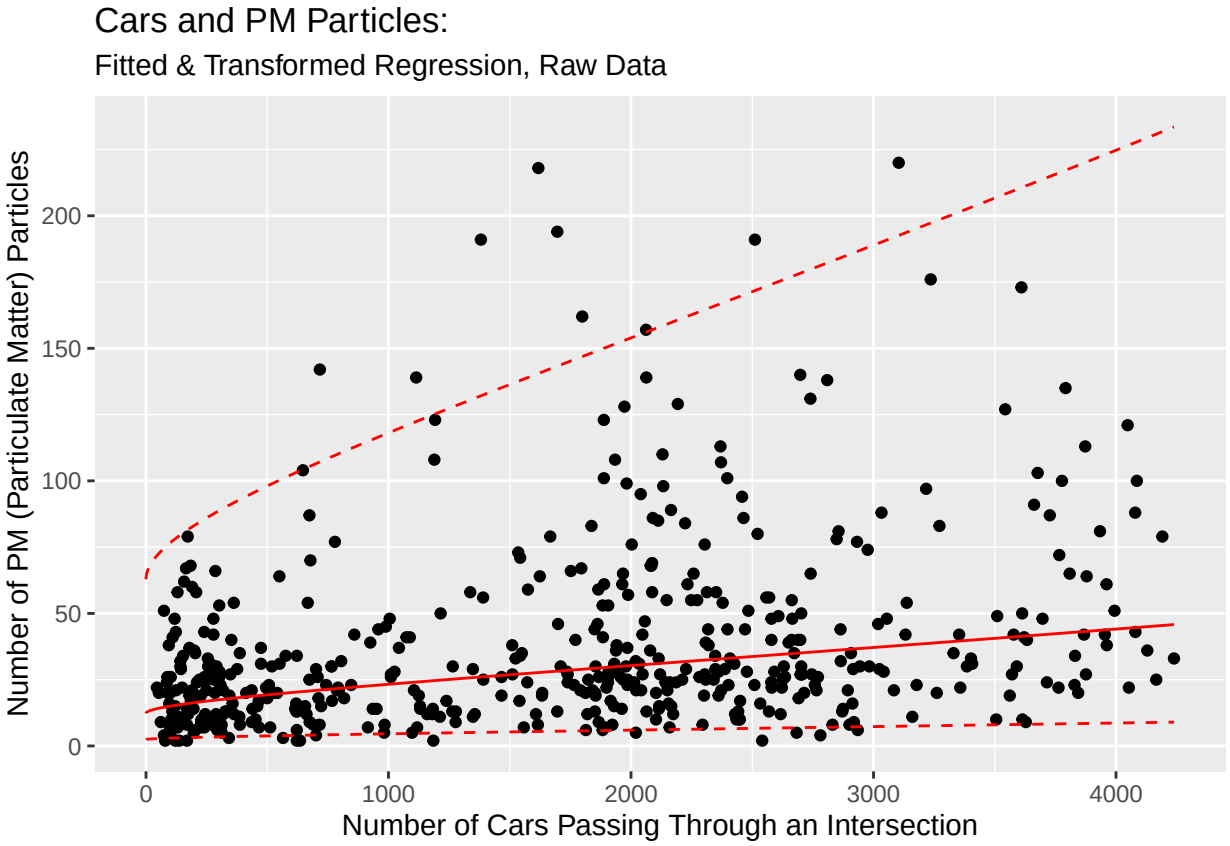
Below, we see the results of the test:

Assuming the null hypothesis is true, the probability of observing a slope of 0.01, or more extreme, is essentially 0. Therefore, we conclude there is a statistically significant, linear effect that the number of cars has on the number of particulate matter particles.

We are 95% confident that if we sampled repeatedly from our model, the true value of $beta_0$ would be between 2.3218 and 2.693. In context, that means if the number of cars was 0, the average amount of particulate matter particles would be between 2.3218 particles and 2.693 particles. We are 95% confident that if we

sampled repeatedly from our model, the true value of $beta_1$ would be between 0.0157 and 0.0247. In context, this means that for every 1 car increase of the number of cars, the average particulate matter particles increases by between 0.0157 particles and 0.0247 particles.

I am 95% confident that if the number of cars passing through an intersection was 1800, the associated number of particulate matter particles would be between 5.7005 particles and 146.9377 particles, on average, with a best fit of 28.9415 particles. Being 95% confident means that if I sampled from my model repeatedly I would have the true average of particulate matter particles contained in the prediction interval I just gave, about 95% of the time. Below is a scatterplot of that range of the prediction interval for all observations.



Section 5: Conclusions

In conclusion, I found that the relationship between the number of cars passing through an intersection and the number of particulate matter particles is generally positive, once the data had been transformed and proven with tests and cross validation. This relationship could be affected by the assumptions made, namely that the data (once transformed) is linear, independent, follows a Normal distribution, and has equal variance about the regression model, but I did what I could to adjust for these transformations to the data to be true, and accepted that I don't know how to adjust for dependence yet. I created a scatterplot of the relationship's prediction interval, showing the traffic on an intersection can be used to predict the number of particulate matter particles, but the range of possible values of particulate matter particles is large.

Some next steps the environmental scientists can consider next include finding the effect dependence has on their collected data, with the help of a more practiced statistician, and repeating the study for validation and meta-analysis at another location. Another less cautious suggestion could be to create an experiment to determine methods to reduce particulate matter particles, despite nearby traffic.

Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(tinytex)
library(ggplot2)
library(MASS)
library(normtest)
library(lmtest)
library(SciViews)
options(scipen = 999)
#rewrite the text below the code chunk.
#Make each section included in the code chunk a new paragraph.
# (a) In your own words, describe the background of the problem and the goals of the study.
# (b) Summarize what data you are going to use to fulfill the goals mentioned above. Explore the data b
# (c) Comment on whether or not a simple linear regression (SLR) model would be appropriate for this da
pollute <- read.table("~/R programming/STAT_330/Past_Data/CarPollutionData.txt",
                      sep = ' ', header = TRUE)
#x <- pollute$Cars
#y <- pollute$Particles
#Calculate the correlation and covariance between Cars and Particles
pollute.cov <- round(cov(pollute$Cars, pollute$Particles), digits = 4)
pollute.cor <- round(cor(pollute$Cars, pollute$Particles), digits = 4)
C1 <- as.numeric(round(summary(pollute$Cars)["Min."], digits = 4))
C2 <- as.numeric(round(summary(pollute$Cars)["Median"], digits = 4))
C3 <- as.numeric(round(summary(pollute$Cars)["Mean"], digits = 4))
C4 <- as.numeric(round(summary(pollute$Cars)["Max."], digits = 4))
P1 <- as.numeric(round(summary(pollute$Particles)["Min."], digits = 4))
P2 <- as.numeric(round(summary(pollute$Particles)["Median"], digits = 4))
P3 <- as.numeric(round(summary(pollute$Particles)["Mean"], digits = 4))
P4 <- as.numeric(round(summary(pollute$Particles)["Max."], digits = 4))
pollute.scatter <- ggplot(data = pollute, mapping=aes(x=Cars, y=Particles)) +
  geom_point() +
  xlab('Number of Cars Passing Through an Intersection') +
  ylab('Number of PM(Particulate Matter) Particles') +
  ggtitle('Cars and PM Particles: Raw Data')
suppressMessages(print(pollute.scatter))
pollute.raw.slr <- lm(formula=Particles~Cars, data=pollute)
pollute.raw.r2 <- round(summary(pollute.raw.slr)$r.squared, digits = 4)
# Draw a fitted values vs. residuals plot to check the L and E assumption.
fit.vs.resids.2 <- ggplot(pollute, aes(x=pollute.raw.slr$fitted.values,
                                       y=pollute.raw.slr$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Linearity and Equal Variance Assumption Checks:',
          subtitle ='Raw Data')
suppressMessages(print(fit.vs.resids.2))

# Draw a histogram (or density plot) of standardized residuals to check the N assumption.
standardized.residuals2 <- stdres(pollute.raw.slr)
pollute.freq2 <- ggplot() +
  geom_histogram(mapping=aes(x=standardized.residuals2)) +
  xlab('Standardized Residual') +
```

```
  ylab('Frequency') +
  ggtitle('Normality Assumption Check:', subtitle = 'Raw Data')
suppressMessages(print(pollute.freq2))
#Create a kable table
# Conduct a KS and JB test for normality.
ks.test(standardized.residuals2, "pnorm")
jb.norm.test(standardized.residuals2)
#Create a kable table
# Conduct a BP test for equal variance.
bptest(pollute.raw.slr)
#Create a kable table
# Identify any outlying observations using Cook's distance.
cooks2 <- cooks.distance(pollute.raw.slr)
outlier.where2 <- 4/length(pollute$Particles)
outliers2 <- pollute[which(cooks2>outlier.where2),]
#Make each section included in the code chunk a new paragraph.
# (a) Mathematically write out a justifiable (perhaps after a transformation) simple linear regression
# (b) Interpret any parameters (i.e. any Greek letters) in the context of the problem and on the level
# (c) Explain (you'll justify these later on in the report) any assumptions you may have used.
#below was r^2=0.1341 and had better variance it seems than other transformations
pollute.slr <- lm(formula=I(log(Particles))~I(sqrt(Cars)), data=pollute)
pollute.r2 <- round(summary(pollute.slr)$r.squared, digits = 4)
pollute.beta.0 <- round(as.numeric(coef(pollute.slr)["(Intercept)"]), digits = 4)
pollute.beta.1 <- round(as.numeric(coef(pollute.slr)["I(sqrt(Cars))"]), digits = 4)
pollute.var <- round(sigma(pollute.slr)^2, digits = 4)
#Make each section included in the code chunk a new paragraph.
# (a) Justify any assumptions you made in your model using graphics, summary statistics or hypothesis t
# (b) Do you think that your model fits the data well? Explain why or why not based on appropriate summ
# (c) How well does your model do at predicting? Describe how you know and support your conclusions wit
pollute.scatter <- ggplot(data = pollute,
                          mapping=aes(x=I(sqrt(Cars)), y=I(log(Particles)))) +
  geom_point() +
  xlab('Number of Cars Passing Through an Intersection') +
  ylab('Number of PM (Particulate Matter) Particles') +
  ggtitle('Cars and PM Particles:', subtitle = 'Transformed Data')
suppressMessages(print(pollute.scatter))

# Draw a fitted values vs. residuals plot to check the L and E assumption.
fit.vs.resids.1 <- ggplot(pollute, aes(x=pollute.slr$fitted.values,
                                       y=pollute.slr$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Linearity and Equal Variance Assumption Checks:',
          subtitle ='Transformed Data')
suppressMessages(print(fit.vs.resids.1))

# Draw a histogram (or density plot) of standardized residuals to check the N assumption.
standardized.residuals <- stdres(pollute.slr)
pollute.freq <- ggplot() +
  geom_histogram(mapping=aes(x=standardized.residuals)) +
  xlab('Standardized Residual') +
  ylab('Frequency') +
```

```
  ggtitle('Normality Assumption Check:', subtitle = 'Transformed Data')
suppressMessages(print(pollute.freq))
#Create a kbale table
# Conduct a KS and JB test for normality.
ks.test(standardized.residuals, "pnorm")
jb.norm.test(standardized.residuals)
#Create a kable table
# Conduct a BP test for equal variance.
bptest(pollute.slr)
# Identify any outlying observations using Cook's distance.
cooks <- cooks.distance(pollute.slr)
outlier.where <- 4/length(pollute$Particles)
outliers <- pollute[which(cooks>outlier.where),]
#cross validation
#set seed for reproducibility
set.seed(87)
n.cv <- 100 #Number of CV studies we'll run
bias <- rep(NA, n.cv) #n.cv empty biases (one for each CV)
RPMSE <- rep(NA, n.cv) #n.cv empty RPMSE (one for each CV)
coverage <- rep(NA, n.cv) #n.cv empty coverage (one for each CV)
width <- rep(NA, n.cv) #n.cv empty width (one for each CV)
n.test <- 15 #How big my test set is
for(i in 1:n.cv){
  # Choose which obs. to put in test set
  test.obs <- sample(1:nrow(pollute), n.test)

  # Split data into test and training sets
  test.set <- pollute[test.obs,]
  train.set <- pollute[-test.obs,]

  # Using training data to fit a (possibly transformed) model
  train.lm <- lm(I(log(Particles))~I(sqrt(Cars)),data=train.set)

  # Predict test set
  test.preds <- exp(predict.lm(train.lm, newdata=test.set, interval="prediction"))

  # Calculate bias
  bias[i] <- mean(test.preds[,1]- test.set$Particles)

  # Calculate RPMSE
  RPMSE[i] <- sqrt(mean((test.preds[,1] - test.set$Particles)^2))

  #coverage
  coverage[i] <- mean((test.preds[,2] < test.set$Particles) &
                        (test.preds[,3] > test.set$Particles))

  #width
  width[i] <- mean(test.preds[,3] - test.preds[,2])
}

pollute.stddev <- round(sigma(pollute.slr), digits = 4)
mean.bias <- round(mean(bias), digits = 4)
mean.RPMSE <- round(mean(RPMSE), digits = 4)
```

```
mean.coverage <- round(mean(coverage), digits = 4)
mean.width <- round(mean(width), digits = 4)
pollute.CV.bias <- ggplot() +
  geom_histogram(mapping=aes(x=bias)) +
  xlab('Amount of Bias') +
  ylab('Frequency') +
  ggtitle('Cars and PM Particles:', subtitle = 'Range of Bias for Cross Validation')
suppressMessages(print(pollute.CV.bias))

pollute.CV.RPMSE <- ggplot() +
  geom_histogram(mapping=aes(x=RPMSE)) +
  xlab('RMPSE') +
  ylab('Frequency') +
  ggtitle('Cars and PM Particles:', subtitle = 'Range of RPMSE for Cross Validation')
suppressMessages(print(pollute.CV.RPMSE))

CV.coverage <- ggplot() +
  geom_histogram(mapping=aes(x=coverage)) +
  xlab('Coverage') +
  ylab('Frequency') +
  ggtitle('Cars and PM Particles:', subtitle = 'Range of Coverage for Cross Validation')
suppressMessages(print(CV.coverage))

CV.width <- ggplot() +
  geom_histogram(mapping=aes(x=width)) +
  xlab('Width') +
  ylab('Frequency') +
  ggtitle('Cars and PM Particles:', subtitle = 'Range of Width for Cross Validation')
suppressMessages(print(CV.width))
#Make each section included in the code chunk a new paragraph.
# (a) Based on your fitted model, test (even though testing is "vague") if there a relationship between
# (b) If there is a relationship in (a), provide an estimate (along with a measure of uncertainty) of w
# (c) Regardless of your answer above, the environmental scientists are making you use your model to pr
#Create a kable table
#print("Hypothesis Tests:")
summary(pollute.slr)
pollute.CI <- confint(pollute.slr, level = 0.95)
#gives prediction on transformed scale
pollute.predict <- data.frame(Cars = 1800)
prediction <- exp(predict.lm(pollute.slr, newdata=pollute.predict,
                         interval="prediction",
                         level=0.95))
#Provide a plot of the data with a fitted regression line
# **on the original scale of the data**
cars.seq <- seq(0, 4239, length=10000) #max cars: 4239
#particles.seq <- seq(0, 5.203055e+97, length=10000) #max particles: 220
preds <- data.frame(Cars=cars.seq)
#preds$Cars <- (predict.lm(pollute.slr, newdata = preds))^2
preds <- exp(predict.lm(pollute.slr, newdata=preds, interval="prediction"))
#ggplot
pollute.est.reg <- ggplot() +
  geom_point(data=pollute, mapping=aes(x=Cars,y=Particles)) +
  geom_line(mapping=aes(x=cars.seq, y= preds[,'fit']), color="red") +
```

```
  geom_line(mapping=aes(x=cars.seq, y= preds[,'lwr']), color="red", linetype=2) +
  geom_line(mapping=aes(x=cars.seq, y= preds[,'upr']), color="red", linetype=2) +
  xlab('Number of Cars Passing Through an Intersection') +
  ylab('Number of PM (Particulate Matter) Particles') +
  ggtitle('Cars and PM Particles:',
          subtitle = 'Fitted & Transformed Regression, Raw Data')
suppressMessages(print(pollute.est.reg))
#Make each section included in the code chunk a new paragraph.
# (a) Briefly summarize the main findings of your analysis in 1 paragraph and without using statistical
# (b) Identify 1-2 "next steps" that the environmental scientists should consider in studying the relat
#End of midterm's code
```