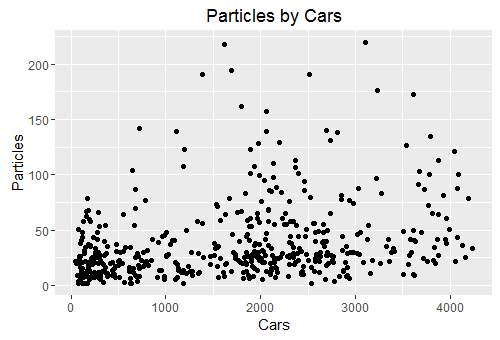
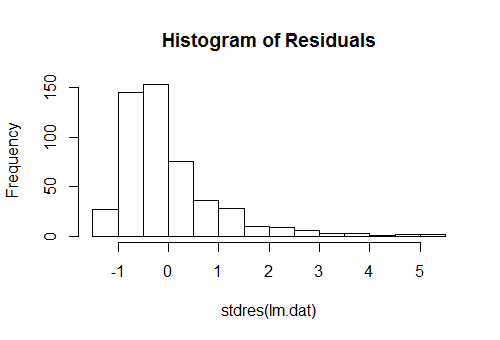
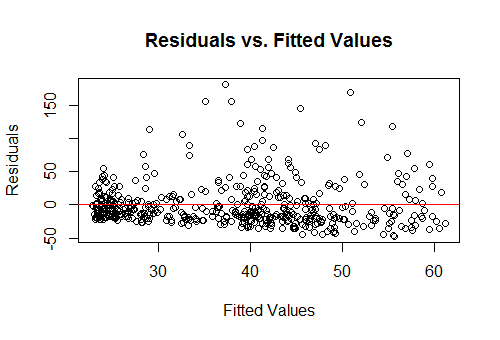
**Introduction and Problem Background**

There are many potential negative health impacts of exposure to "particulate matter," more commonly known as pollution or PM. These negative impacts can be experienced by either short-term or long-term exposure to particles. We have a dataset that contains measurements on the number of PM particles, as well as the number of cars passing through an intersection. We will analyze these data in hopes to define a relationship between the number of PM particles and the number of cars, as well as to see if traffic can be used to predict PM levels?

As previously stated, the dataset we have contains measurements on the number of PM particles at an intersection and the number of cars passing through that intersection. Below is a scatterplot of the data.

The correlation between Cars and Particles is .30. This doesn't indicate a strong relationship between the two variables. We can still say the data is linear, even though that relationship isn't strong.

We can still assume linearity in the data, but let's also take a look at some of the other assumptions that must be made in order to run a simple linear regression model. Below we have a "Histogram of Residuals" plot and a "Residuals vs. Fitted Values" plot. We must be able to assume that the residuals are normally distributed, but there seems to be a right skew in our residuals. The normality assumption is violated. Also, we must be able to assume that there is equal variance throughout the data, but looking at the residuals versus the fitted values, there seems to be unequal variance with there being fewer points and larger spread above the line. The assumption of equal variance is also violated. The last assumption, independence. We will go forward assuming independence recognizing that if these points are collected at the same intersection then there could be some dependency.



All in all, a simple linear regression (SLR) model would not be appropriate on these data. However, we can transform the data in order to make a SLR appropriate to use in this analysis. We will transform the data by taking the log of the response variable, in this case, Particles.

**Statistical Modeling**

log(yi)= β0 + β1i)+ εi where εi ~ N(0,σ2)

yi = observed value of the ith number of PM particles

β0 = average log number of particles if the number of cars was 0 mph.

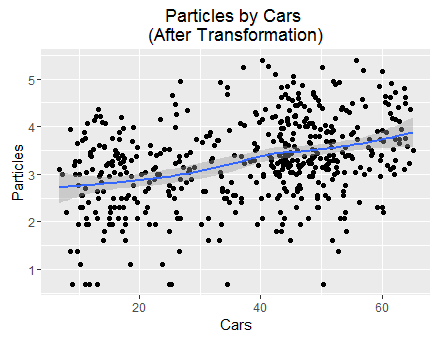
β1 = the average increase in the log number of particles as the number of cars goes up by 1 mph

xi = number of cars you are using to predict the ith number of particles

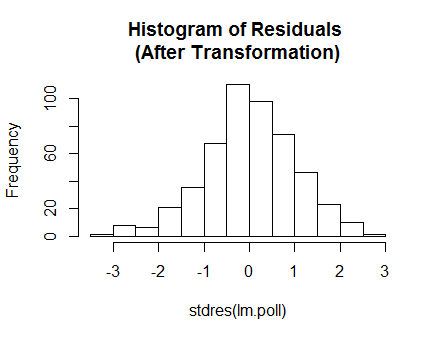
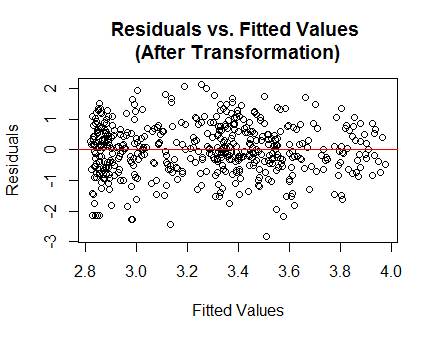
εi = the error associated with the ith observation

This model assumes linearity, normality, independence, and equal variance.

**Model Verification**

The scatter plot on the left shows the transformed data. This helps us to see that the linearity assumption is met. Also after transformation the correlation between Cars and the log of Particles is .36, which is slightly higher than the correlation of the untransformed data.

Below, we will look at the assumptions for normality, equal variance, and independence. The histogram of residuals shows that the errors are normally distributed, with maybe 1 possible outlier. The residuals vs. fitted values plot shows that there is equal variance across the data. This is distinctly different than with the untransformed data. We will also assume that each of the particle measurements are independent from one another.



I don't believe that the model fits the data very well. With a multiple R-squared value of .1296, we can say that only about 13% of the variation in the log number of Particles can be explained by the number of Cars.

In order to gauge how well this model is doing at predicting, we ran a cross-validation simulation. We obtained a bias of -10.13, which means that our predictions for the PM level are a little low on average. Our average RPMSE was 34.23, meaning that our predictions are about 34.23 units off the true value on average. Both of these values are fairly large on the scale for number of Particles suggesting that our predictions may not be as reliable as we would hope. 95% of the prediction intervals we created contained the true value, however the average width of those intervals was 136.2. In other words, we can give accurate prediction intervals, but those intervals might not be very helpful because they are so wide.

**Results**

With the hypothesis that there is no relationship between PM and the number of cars, we obtained a p-value of .00000000000000022. This, of course, is plenty of evidence to reject the null hypothesis and conclude that there is indeed a relationship between the number of cars and PM. Our estimate for that relationship is that as the number of cars increases by 1 the PM level increases by 1.000277. We are 95% confident that the true increase in PM level will lie between 1.000213 and 1.00034.

Using this model, we would predict that when there are 1800 cars passing through there will be a PM level of 27.2. We are 95% confident that the true PM level lies between 5.336 and 138.662.

**Conclusions**

In this study we have found that there is definitely a relationship between the number of cars and the PM level. However, there is a lot more that goes into explaining the variation in PM level besides the number of cars. We are able to make predictions on PM level using the number of cars passing through, but those predictions are very wide and likely unhelpful.

Next steps would involve determining a few more factors that you feel are contributing to the variation in PM levels. We could include those factors in a multiple linear regression model that would explain a higher percentage of the variation and would hopefully lead to narrower prediction intervals, and therefore more accurate predictions.