The Impact of Pollution on Asthmatic Patients of California

Sharanya Sharma

March 23, 2023

Abstract

The purpose of this report is to foster the curiosity of how pollution affects the health of people in California. The California Office of Environmental Health Hazard Assessment (OEHHA) dataset includes levels of certain pollutants prevalent in the state, as well as figures of diseases commonly associated with pollution. Few such pollutants include pm2.5, diesel emission, and toxic release, with this study analyzing the effects on asthma. With a series of tests, we determine that the level of pollutants can significantly predict the log number of asthmatic people in California, as well as the idea that the log spread of asthmatic people throughout California is significantly different.

Introduction

Over the years, the increasing use of fossil fuels, burning of greenhouse gasses, and mass deforestation are credited to being some of the main contributors to air pollution. When pollutants such as carbon dioxide, methane, and nitrous oxide are formed from the burning of fossil fuels, they are trapped within the Earth's atmosphere and absorb solar radiation, thereby increasing the Earth's temperature (Turrentine et al. 2021). This is the process commonly referred to as Global Warming. Although there are many initiatives that aim to protect and preserve the Earth from further destruction, many fail to realize that certain communities are more prone to air pollution because of the disproportionate balance of greenhouse emitting manufacturing plants that populate such industrial-oriented communities. Thus, in 2021 the OEHHA, designed a tool that graphed various air pollutants and their effects on communities across California. By using demographics information from the 2010 and 2020 census, as well the 2019 ACS (American Community Surveys), the CalEnviroScreen 4.0 tool depicts Californian communities that are disproportionately "burdened" by pollution (OEHHA, 2022). In addition to pollutants, the tool provides data on areas affected with diseases most commonly associated with pollution, such as asthma and cardiovascular disease. A few such pollutants that are measured by OEHHA include pm2.5 levels, toxic chemical release from factories, and diesel emissions. According to the New York State Department of Health, fine chemicals, dust, soot, and metals that compose pm2.5 may worsen lung function or medical conditions (McDonald et al. 2022). In this study, we will look at the effects of these pollutants on on a random sample of 1000 data points in California, and if one of these pollutants (or a combination of them) can be used to predict the amount of asthmatic patients in California. In addition to this, we will compare the 4 regions of California (northern, southern, inland, and coastal), to determine if there is a significant difference in the amount of asthmatic patients amongst different parts of California.

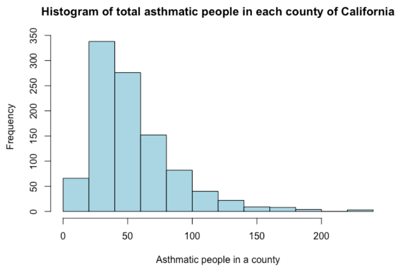
Hypothesis 1: The number of asthmatic people in each region (northern, southern, inland, coastal) is significantly different.

Hypothesis 2: The pm2.5 level, or a combination of this pollutant with toxic chemicals and diesel emissions in a given county, can significantly predict the number of asthmatic people in California.

Exploratory Data Analysis

Histogram

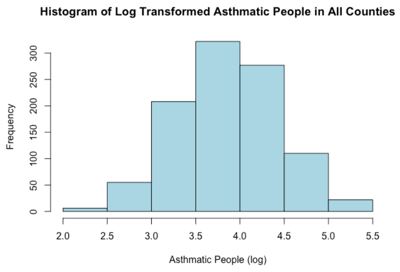
The first graph to consider is the histogram of asthmatic people. For each vertical bar on the histogram below, the value on the x-axis represents the number of asthmatic people in a county and the frequency represents the amount of counties that have "x" number of asthmatic people. In this manner, we are able to visualize the spread of the frequency of asthmatic people in California. Notice that this histogram contains most of its data points clumped to the left side, resulting in a right-skewed histogram. Therefore, before proceeding with any analysis we must transform the data to look normal.



Applying a logarithmic transformation to the histogram results in a bell-curve shape, which is ideal for approximating normality, as shown below. Thus, we will update our hypotheses as the log asthmatic values will be used for all the remaining graphs as well.

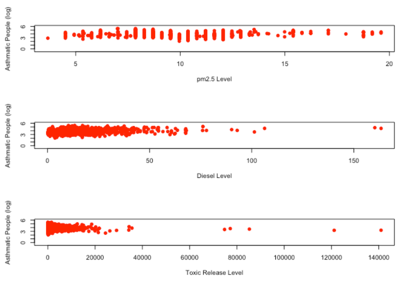
New Hypothesis 1: The log number of asthmatic people in each region (northern, southern, inland, coastal) is significantly different.

New Hypothesis 2: The pm2.5 level, or a combination of this pollutant with toxic chemicals and diesel emissions in a given county, can significantly predict the log number of asthmatic people in California.



Scatter Plot

Since the goal of the first hypothesis is to determine a prediction, we must consider the relationship between each of the predictor variables (pm2.5, toxic release, and diesel) and the log asthmatic people. As all the variables are numeric, a scatter plot would best represent a visualization of the data.

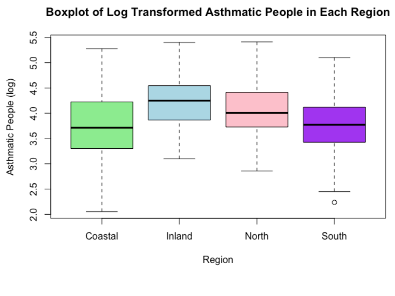


The first scatter plot of log asthmatic number of people to pm2.5 level shows the most comparatively linear pattern. However, since the data is mostly horizontal, this suggests that any input value of x would result in a similar value of y, implying that the log asthmatic number of people may not have much correlation with the pm2.5 level. The second and third plots representing log asthmatic number of people to diesel level, and log asthmatic people to toxic release level (respectively) shows much of the data clumped to the left side of each graph. This suggests that there is not much variance among most of the data points as all the x and corresponding y values are around the same area on the graph.

Boxplot

Finally, the boxplot allows for a comparison between the distributions of the log number of asthmatic people within each of the 4 regions of California. The thick black line in the middle of each distribution represents the median log asthmatic number of people. Thus 50% of the data lies above this line, and 50% of the data lies below this line. Similarly, 50% of the data lie within each colored box, this area is known as the interquartile range. Finally the top 25% of the data lie above the colored "box" and the lowest 25% of the data lie below the colored "box", in an area known as the "whiskers". Values even farther than this boundary are represented by hollow circles, and are considered outliers.

Looking at the boxplots below, Inland and North seem to have a relatively similar median and range of log asthmatic people, with the exception of an outlier in the North. The southern region has a slightly larger range, implying that there is more variation of the log quantity of people who have or don't have asthma. Since the southern boxplot has 3 outliers on the lower end of values, this suggests that certain counties in the southern section have incredibly low number of asthmatic patients compared to the other counties in the region. Finally, the coastal region has the largest range with the lowest log asthmatic median value, implying that the counties in the coastal range have the largest variation of the log quantity of people of who have or don't have asthma.



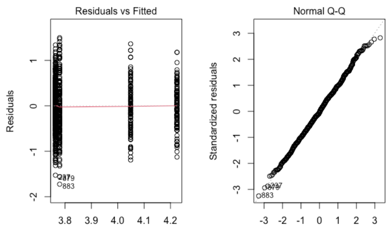
Statistical Methods

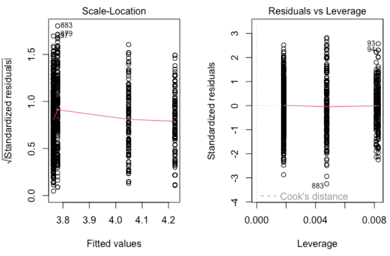
ANOVA

The ANOVA test is used to compare the means of multiple samples all affected by asthma. For the next part of the statistical analysis, we will use a one-way ANOVA test to determine if the mean log number of asthmatic people between the northern, southern, inland, and coastal regions are statistically different. This study will therefore determine whether there are statistically more asthmatic people in a certain region of California compared to another.

This test assumes that the data is randomly sampled, has normally distributed residuals, and equal variance of the residuals. Proceed with verifying these assumptions before conducting the test. Since the data was collected using a random sample of census data, the first assumption is valid. For the next assumption refer to the Normal Q-Q plot below. Notice that a majority of the data points follow the normality line, apart from a few deviations on either ends from outliers. Therefore, normality can be assumed.

Finally, consider the Residual vs Fitted plot below. Notice that the variance is not randomly scattered above and below the residual line, but rather clumped in three distinct linear patterns. This suggests that the variances are not equal. Further confirm this assumption of unequal variance with a Levene's Test.





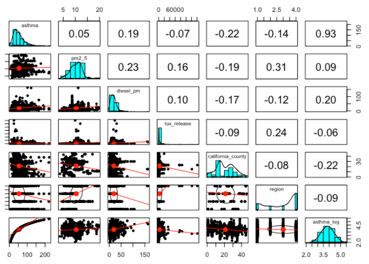
Since the p-value of the Levene's Test is 0.0000006004, which is much less than the significance level 0.05, there is further evidence to assume that the variances are in fact not equal.

The criteria for ANOVA assumption has failed, thus proceed with a non-parametric test. Implied by the name, non-parametric tests do not require any parameters or assumptions. However because of this, these tests are not as powerful and should only be used if an assumption for a regular test fails.

Therefore proceed with a non-parametric Welch's ANOVA Test. Although typically the Welch's non-parametric test is seen to be used with two samples, the Welch's ANOVA test is a non-parametric method of comparing three or more sample means (Frost, 2022). Therefore, maintain the same hypothesis that the log number of asthmatic people in each region (northern, southern, inland, coastal) is significantly different.

Linear Regression

The linear regression model can be used to predict the log number of asthmatic people depending on the level of pm2.5, toxic release, and or diesel emission. Before generating the possible models, we must test for co-linearity and a linear relationship between the predictors and the response variable using the graph below.



Checking for co-linearity, the absolute correlation values between the predictors (pm2.5 level, diesel emission, and toxic release) and the response variable (log number of asthmatic people) are all much less than 1, suggesting that there is no co-linearity between the predictor variables. This passes the co-linearity assumption.

Checking the relationship of the predictors with the log number of asthmatic people, no plots appear to be distinctly non-linear, thus passing the linearity assumption. The plots for toxic release have a few points that deviate from the majority, implying those may be outliers, however since there is an overall linear pattern we can assume there is some linear relationship between the predictors and the response variable.

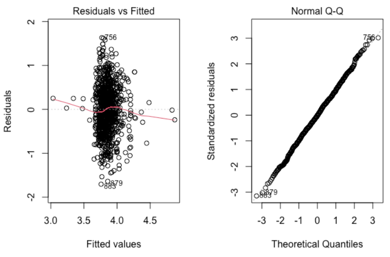
Proceed with developing 3 models. Diesel has the highest correlation value with log asthma of 0.2, the first model will fit only diesel. The second model will use both diesel and pm2.5, and the third model will use all three predictors. In order to determine which of these three models will create the strongest prediction, consider the Akaike's Information Criterion (AIC) and Bayes' Information Criterion (BIC). Both the AIC and BIC values are a score that measure how well the model fits the data with a penalty depending on the number of parameters being considered.

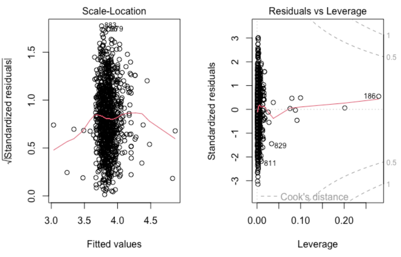
	df	AIC	BIC	rsq	adj_rsq
fit_1var	3	1627.07	1641.79	0.04	0.04
fit_2var	4	1627.05	1646.68	0.04	0.04
fit_3var	5	1620.28	1644.82	0.05	0.05

Looking at the AIC and BIC values, it appears that the third model with all three predictors has the lowest AIC value of 1620.28 with a close second lowest BIC value of 1644.82. In addition to this, the adjusted r-squared value 0.05 is highest, implying that the percentage of variation in the data explained by this model is comparatively the best.

In order to proceed with testing if the linear model can determine a significant prediction, three assumptions must be tested. Similar to the ANOVA assumptions, linear regression also assumes a random sample, normally distributed residuals and equality of variance with the residuals.

Earlier in this report, it was determined that the data was randomly selected since the data was collected using a random sample of census data. Looking at the Normal Q-Q plot below, a majority of the data follows the linear pattern, apart from a few deviations at both ends such as 756, 879, 883. Therefore normality can be assumed. The variation is evenly scattered above and below the residual line, suggesting equal variance. Although the data points are heavily dense in the center as opposed to an even scatter, the residual line is relatively maintained at 0 with equal scatter above and below the line, therefore we can arguably assume equal variance. Therefore proceed with the hypothesis that the combination of pm2.5, toxic chemicals, and diesel emissions in a given county, can significantly predict the number of asthmatic people in California.





Results

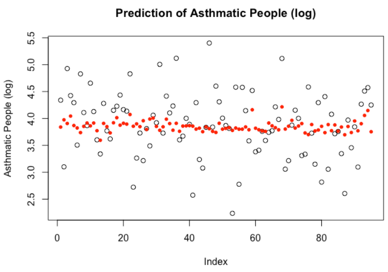
Welch's ANOVA

In order to interpret the outcome of this Welch's ANOVA test, consider the significance level alpha of 0.05. The significance level is the probability of rejecting the null hypothesis (the idea that the mean log number of asthmatic people within each region is equal), when it is in fact true. The p-value from the test above is compared with the significance level alpha to determine the result of the Welch's ANOVA test.

The p-value of 2.2e-16 is much less than the significance level 0.05, implying that we reject the null hypothesis. There is evidence that suggests the mean log number of asthmatic people within each region is in fact, not equal. Therefore we have evidence to support our original hypothesis.

Linear Regression

In order to generate a prediction, split the data into training data and test data. The training data will be fit with the linear model, thus considered "trained". The trained data is then compared to the test data, which is the data points selected from the actual data.



The red points represent the training data, whereas the white points represent the actual data. Hence, the red points represent how well the model can predict the actual data points.

Since all the predicted values lie within the scatter of the actual data, the regression model demonstrated a fairly good prediction of the log number of asthmatic people in each county. Therefore we reject the null hypothesis that the slope of the regression line is 0. In other words, there is evidence that suggests that a combination of pm2.5 level, diesel emissions, and toxic release can significantly predict the log number of asthmatic people in California.

Discussion

Both the Welch's ANOVA test and the prediction generated from the regression model resulted in significant outcomes. The first test shows that the log mean number of asthmatic people are different within each of the four regions of California. Therefore, there is evidence that the northern, southern, inland, and coastal regions of California have significantly different log amounts of people diagnosed with this disease. This further supports the idea that various communities in California are disproportionately affected by diseases commonly associated with pollution, asthma in particular. With this information, biologists and environmental activists can promote better health and well-being by focussing on initiatives in regions more prone to asthma.

The second test shows that pollutants such as pm2.5, toxic release, and diesel emission can significantly predict the log number of asthmatic people in California. Therefore, there is evidence that communities with higher levels of these pollutants present in the air suffer from asthma significantly more than communities with cleaner air quality. This is additional information biologists can use to encourage the use of green energy, recycling, and other environmentally friendly initiatives for the betterment of human health and the ecosystem as a whole.

Since the ANOVA test was conducted using a non-parametric Welch's test, limitations lie with the power of how significant these results may actually be. Another limitation is the access to new data. Since the data was collected using census information from 2010 and 2020, it may not be as reliable to give the most accurate representation of the current ecosystem status.

If given more time, we could analyze ways to make the variance equal and use the standard ANOVA test for results instead. Perhaps this may result in a more powerful, thus more convincing test result. Since this data is comparing the pollutant levels with asthma levels of only Californian citizens, this may create bias as other states or countries with different levels of pollution may have different results. If given the time and resources to conduct this test on a global scale, perhaps the results would be more applicable to anyone reading the report.

References

Amanda MacMillan, Jeff Turrentine. “Global Warming 101.” Definition, Facts, Causes and Effects of Global Warming, 7 Apr. 2021, https://www.nrdc.org/stories/global-warming-101#causes.

James V. McDonald et al. “Department of Health.” Fine Particles (PM 2.5) Questions and Answers, September 2022 https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm.

“CalEnviroScreen 4.0.” Oehha.ca.gov, https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40.

Frost, Jim. “Benefits of Welch’s ANOVA Compared to the Classic One-Way ANOVA.” Statistics By Jim, 29 Sept. 2022, https://statisticsbyjim.com/anova/welchs-anova-compared-to-classic-one-way-anova/.

Wickham H, Hester J, Bryan J (2022). *readr: Read Rectangular Text Data*. R package version 2.1.3, https://CRAN.R-project.org/package=readr.

Revelle, W. (2022) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.9.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL:https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Qiu Y (2021). *prettydoc: Creating Pretty Documents from R Markdown*. R package version 0.4.1, https://CRAN.R-project.org/package=prettydoc.

Yihui Xie (2022). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.41.

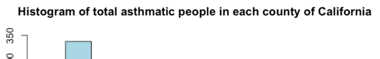
Appendix

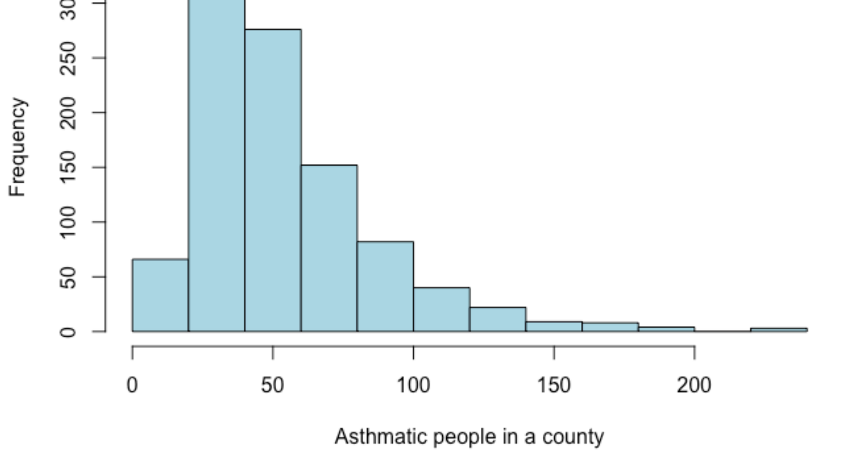
```
# load in data
library(readr)
library(psych)
library(tidyverse)
library(car)
library(prettydoc)
library(knitr)

# read data
pollution_full <- read_csv("dataset_20.csv")
```

```
# Extract columns of interest including x and y variables only
asthma_y <- pollution_full["asthma"]
pm2.5_xnum <- pollution_full["pm2_5"]
diesel_xnum <- pollution_full["diesel_pm"]
toxrel_xnum <- pollution_full["tox_release"]
county_xcat <- pollution_full["california_county"]

# Create a new dataframe with the columns of interest
pollution <- cbind(asthma_y, pm2.5_xnum, diesel_xnum, toxrel_xnum, county_xcat)
```

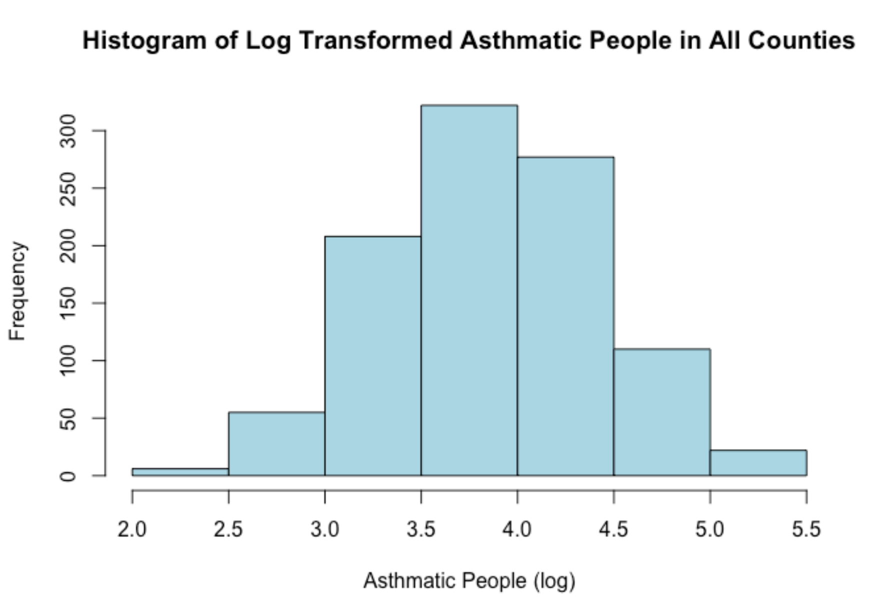
```
# Classify northern region counties
north <- pollution %>%
  filter(california_county == "Sacramento"
        |california_county == "Butte"
        |california_county == "El Dorado"
        |california_county == "Humboldt"
        |california_county == "Lake"
        |california_county == "Lassen"
        |california_county == "Mendocino"
        |california_county == "Napa"
        |california_county == "Nevada"
        |california_county == "Placer"
        |california_county == "Shasta"
        |california_county == "Solano"
        |california_county == "Sonoma"
        |california_county == "Tehama"
        |california_county == "Trinity"
        |california_county == "Sutter"
        |california_county == "Yuba") %>%
  mutate(region = "North")

# Classify southern region counties
south <- pollution %>%
  filter(california_county == "Los Angeles"
         |california_county == "Imperial"
         |california_county == "Kern"
         |california_county == "Orange"
         |california_county == "Riverside"
         |california_county == "San Diego"
         |california_county == "Ventura") %>%
  mutate(region = "South")

# Classify inland region counties
inland <- pollution %>%
  filter(california_county == "Fresno"
        |california_county == "Amador"
        |california_county == "Calaveras"
        |california_county == "Inyo"
        |california_county == "Kings"
        |california_county == "Madera"
        |california_county == "Merced"
        |california_county == "San Benito"
        |california_county == "San Bernardino"
        |california_county == "San Joaquin"
        |california_county == "Stanislaus"
        |california_county == "Tulare" ) %>%
  mutate(region = "Inland")

# Classify coastal region counties
coastal <- pollution %>%
  filter(california_county == "Alameda"
         |california_county == "Contra Costa"
         |california_county == "Marin"
         |california_county == "Monterey"
         |california_county == "San Francisco"
         |california_county == "San Luis Obispo"
         |california_county == "San Mateo"
         |california_county == "Santa Barbara"
         |california_county == "Santa Clara"
         |california_county == "Santa Cruz") %>%
  mutate(region = "Coastal")

# create a dataframe of pollution data with added column indicating county region
all_regions <- rbind(north, south, inland, coastal)
View(all_regions)
```

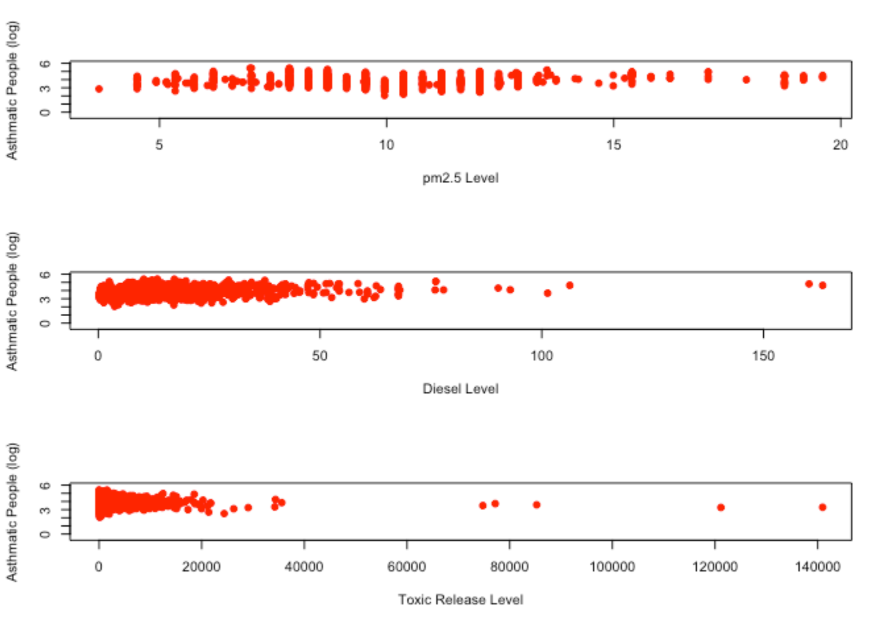
```
# Histogram of asthmatic people in all counties
hist(all_regions$asthma,
     main = "Histogram of total asthmatic people in each county of California",
     xlab = "Asthmatic people in a county",
     col = 'lightblue')
```



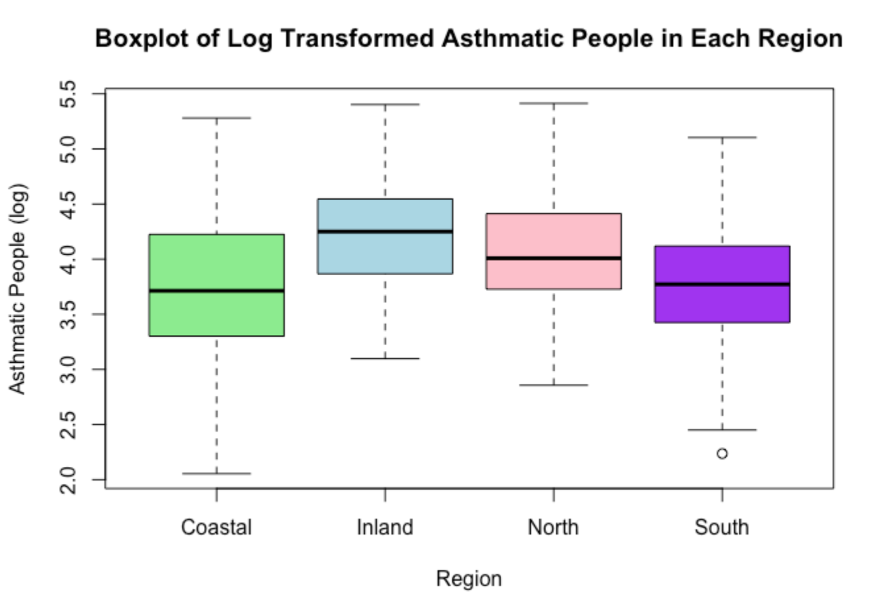


```r
# Log transformation
asthma_log <- log(all_regions$asthma+1)

# Histogram of log transformed asthmatic people in all counties
hist(asthma_log,
     main = "Histogram of Log Transformed Asthmatic People in All Counties",
     xlab = "Asthmatic People (log)",
     col = 'lightblue')
```

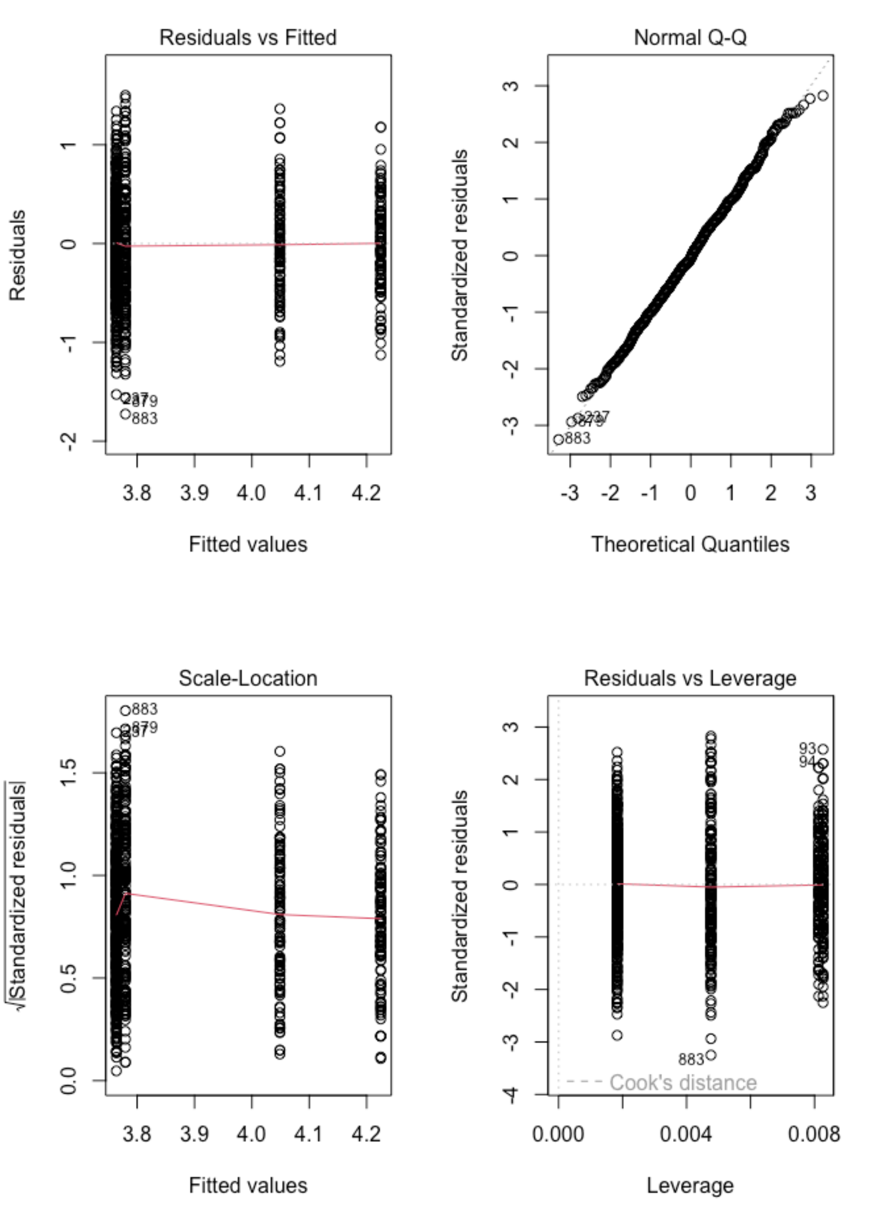




```r
# boxplot of log transformed asthmatic people in each region
boxplot(asthma_log~region,
       main = "Boxplot of Log Transformed Asthmatic People in Each Region",
       data=all_regions,
       xlab="Region",
       ylab="Asthmatic People (log)",
       col= c('lightgreen', 'lightblue','pink','purple'))
```



```r
# Run the anova model
asthma_log_aov <- aov(asthma_log~(all_regions$region))

par(mfrow=c(1,2))

# plot the anova model
plot(asthma_log_aov)
```

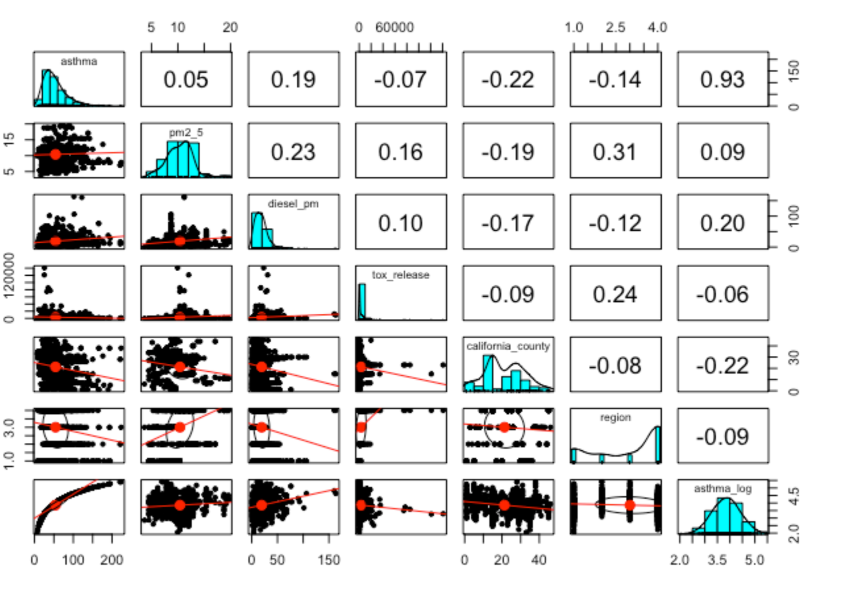


```r
# Find the residuals to find estimate of error
res = asthma_log_aov$residuals
```

```r
# Further confirm unequal variance with Levene's Test
leveneTest(asthma_log~all_regions$region)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   3  7.3567 7.043e-05 ***
##       996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Replace asthma column with log transformed asthma
all_regions['asthma_log'] <- asthma_log

# Make categorical variables in factor format
all_regions$region <- as.factor(all_regions$region)
all_regions$california_county <- as.factor(all_regions$california_county)

# Check data for co-linearity
pairs.panels(all_regions, lm=TRUE, cor=T)
```

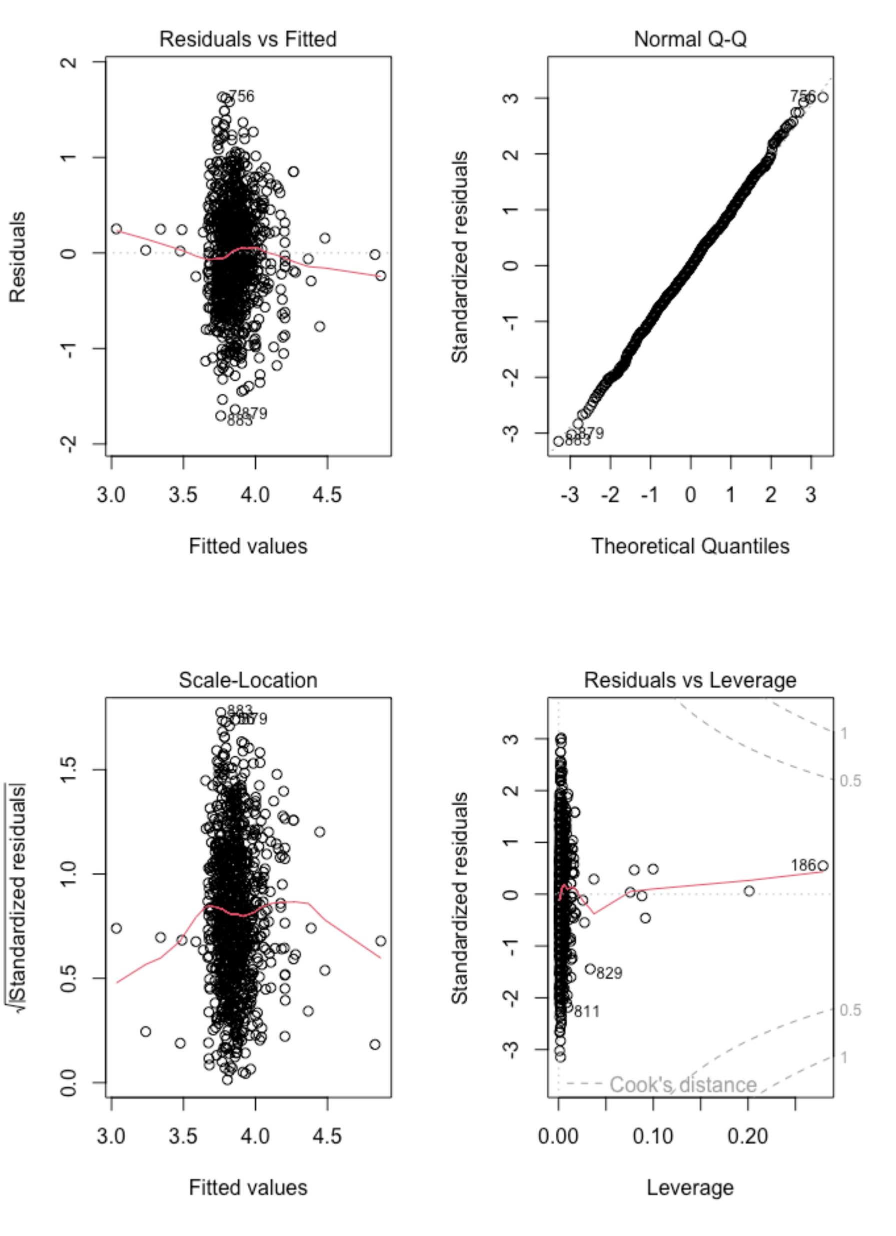


```r
# Fit a model with one predictor, diesel_pm
fit_1var <- lm(asthma_log~diesel_pm, data=all_regions)

# fit model with diesel_pm and pm2_5 as predictors
fit_2var <- lm(asthma_log~diesel_pm+pm2_5, data=all_regions)

# fit model with solar level, wind, and temp as predictors
fit_3var <- lm(asthma_log~diesel_pm+pm2_5+tox_release, data=all_regions)
```

```r
# Calculate AIC of each model, where each row represents a model, and columns represent df a
result <- AIC(fit_1var,fit_2var,fit_3var)

# Create list by adding other metrics to your table
models <- list(fit_1var,fit_2var,fit_3var)

# Append BIC column to result
result$BIC <- sapply(models, BIC)

# returns a list of length "models", and applies "summary" to each corresponding element of
model_summary <- lapply(models, summary)

# Create a for loop to iterate through the table and assign rsq and adjusted rsq value model
for(i in 1:length(models)){
  result$rsq[i] <- model_summary[[i]]$r.squared
  result$adj_rsq[i] <- model_summary[[i]]$adj.r.squared
}

# display the table, round decimals to 2 digits, center alignment
kable(result, digits = 2, align = "c")
```

	df	AIC	BIC	rsq	adj_rsq
fit_1var	3	1627.07	1641.79	0.04	0.04
fit_2var	4	1627.05	1646.68	0.04	0.04
fit_3var	5	1620.28	1644.82	0.05	0.05

```r
# plot the residual vs fitted and normal q-q plot
par(mfrow=c(1,2))
plot(fit_3var)
```



```r
# Conduct Welch's ANOVA test
oneway.test(asthma_log~region, data=all_regions)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  asthma_log and region
## F = 37.912, num df = 3.00, denom df = 308.74, p-value < 2.2e-16
```

```r
# Separate the data into training set containing most of the data, and test set containing l

# Select 95 rows from all_regions data for test data
splitter <- sample(1:nrow(all_regions), 95, replace = F)

# Exclude rows of training data set
psub_train <- all_regions[-splitter,]

# Create a set of test data
psub_test <- all_regions[splitter,]

# Fit model using training set
fit_3var_split <- lm(asthma_log ~ diesel_pm+pm2_5+tox_release, data=psub_train)

# Use model to make a prediction of test data
prediction <- predict(fit_3var_split, psub_test)

# Plot test data values
plot(psub_test$asthma_log,
     main="Prediction of Asthmatic People (log)",
     ylab="Asthmatic People (log)",
     pch=1)
# plot model prediction for given pi
points(prediction, pch=20, col = "red") #plot the model predictions for those points
```