Portfolio: Linear Regression Model

Marissa Llamas

2024-10-24

Executive Summary:

In this project, we analyze trends in the daily Air Quality Index in two combined major Texas regions from 2022 – Austin and the McAllen-Mission-Edinburg area– to gain a deeper understanding on how this metric is calculated from environmental factors to assess air cleanliness. Here, we construct a multiple regression model that predicts the daily air quality index using the daily mean concentration of PM 2.5 and an area's poverty rate for the year. This data was complied from the United States Environmental Protection Agency OutDoor Air Quality Data and the United States Census Poverty Data.

We constructed two multiple regression models based on a threshold for PM 2.5 concentration and both performed well with an R^2 values of virtually 0.99. However, we found the poverty rate of a region to be not significant, with still the PM 2.5 concentration as extremely significant.

Model Construction:

The multiple regression model uses daily mean concentration of PM 2.5 and the yearly poverty rate estimated for the area to predict the daily air quality index. PM 2.5, or particulate matter 2.5, are particles that are 2.5 micrometers less in diameter. These particles can include organic chemicals, dust, soot and metals, and their main source is from cars, trucks, wood burning, and other activities. These fine particles have been shown to cause detrimental health effects, most specifically heart and lung disease. These are especially harmful to disadvantaged groups, such as children, the elderly, individuals with asthma, and those who with preexisting heart and lung disease. *Source

Hence, the Daily Mean Concentration of PM 2.5 in the air is admissible to predict the daily air quality since it adds the unhealthiness of the air.

Additionally, using poverty as a predictor for air quality data can be relevant due to socioeconomic factors. For one, people living in low income households are typically more vulnerable to higher levels of air pollution since these individuals may lack access to alternative green transportation and have a higher proximity to pollution sources like industrial zones and highways. By incorporating poverty as a predictor, we can aim to better understand the impact on the environment and discuss environmental inequalities. *Source

Model Assessment:

In order to assess if the predictors meet the requirements used in the Multiple Regression model, we must validate the model assumptions. These data points are independent since the monitors that are used to measure the PM 2.5 concentration in one region are not influenced or affected by the results of another monitor. Each monitor independently measures PM 2.5 concentration independently in a respective area.

To assess linearity, we observe the fitted vs. residuals plot (Figure 1). At a glance, these values meet the constant variance condition, however, it is evident that this model does not meet linearity as the points do not follow a straight line centered at residuals = 0. When plotting the air quality index variable and the PM

2.5 predictor variable however, we see what looks like two separate linear equations that govern the behavior of air quality (Figure 2). This suggests we construct two linear models with a threshold of $B = 9.5$, where B is a value in the PM 2.5 predictor variable.

Model 1:

When levels of PM 2.5 are less or equal to than 9.5, our multiple linear regression model is:

$$\hat{AQI}_1 = 5.53 \cdot \text{PM}_{2.5} + 0.0003678 \cdot \text{Poverty} + 0.0938112$$

This model adheres to linearity as the residuals vs. fits plot for this model center and follow the residuals = 0 line. Additionally, constant variance is also met as points are equally surrounding the residuals = 0 line. Also, looking at the QQ Residuals plot, the data points seem to follow a straight line, indicating normality is reasonably well. Though, there do seem to be some small deviations of the line at the tail of the line. As an extra precaution, we look at the histogram of the residuals of the model. The residuals are centered at 0, which is expected, and the residuals are roughly bell-shaped. Overall, the histogram reveals the residuals are close to normally distributed, which support our model assumptions.

The F statistic for this model is around 776600 on 2 and 1093 degrees of freedom, which signifies a p-value that is close to 0. Meaning, this data provides extremely strong evidence that we can conclude that at least one of the slope coefficients is non-zero in our model. Now, looking at this a bit closer, we can investigate the meaning of each slope coefficient.

The coefficient for poverty rate had a t-value of 0.313, the coefficient for PM 2.5 concentration had a t-value of 1231.735, and the coefficient for the intercept had a t-value of 2.507. The only significant p-values that were significant were the coefficients for intercept and PM 2.5 concentration, meaning we estimate them to be are non-zero. This is to say, the coefficient for poverty rate was not significant and we can estimate that it is zero.

For context, the significance of the intercept reveals that when the PM 2.5 concentration and the poverty rate are both 0, this is to say there is the air quality index is 0 as well. This is a bit unsettling and unlikely as there are other factors in the air other than PM 2.5 that contribute to air quality. In which case, we would say this intercept does not mean much in context. Also, the likelihood of having a perfect 0% poverty rate for an entire region is quite slim. For the slope coefficient of PM 2.5 concentration, this is to say that when we control for changes in the `pov` variable, a one unit increase in PM 2.5 concentration results in an increase of 5.5 in the air quality index.

Model 2

Now, let us investigate our second model.

When levels of PM 2.5 are greater than 9.5, our multiple linear regression model is:

$$\hat{AQI}_2 = 1.90 \cdot \text{PM}_{2.5} + 0.002154 \cdot \text{Poverty} + 33.524602$$

This model *almost* adheres to our assumptions. For one, the residuals vs. fits plot follow a steady line at the residuals = 0 line when the fitted AQI is around 60-75 units. There is some deviation from this as the air quality index increases however, potentially due to the availability of data and outliers with residuals as high as 4 units.

These outliers are also further highlighted in Figure Set 4 in the Residuals vs. Leverage plot.

There are at least 3 outliers, one of which goes far beyond Cook's Distance, the point 388. We can calculate the Studentized Residuals to also choose those points that might seem unusual. We compute that 7 points are all moderately unusual points since their studentized residuals are greater than 2. There are also 30

points with high leverage - greater than twice the mean! Notably, these row values with high leverage values have a high AQI value as well.

The general behavior at the start of this plot leans toward a linear relationship. Still though, we see the QQ Residuals plot which follows a straight line, indicating the residuals are normally distributed. Yet, we cannot ignore the deviation at the right tail of the plot as it is disruptive. Finally, seeing the histogram of the residuals, we see that they do not look normally distributed and are skewed with residuals extending to the right, due potentially to the outliers and unusual points mentioned above.

Remarkably, the F-statistic for this model is around 109100 with 2 and 520 degrees of freedom, which equals to a p-value that is virtually 0. So, the data provides very strong evidence to conclude that at least one of the slope coefficients is non-zero in our model.

The coefficient for poverty rate had a t-value of 0.724, the coefficient for PM 2.5 concentration had a t-value of 461.129, and the coefficient for the intercept had a t-value of 449.090. The only p-values that were significant were the coefficients for intercept and PM 2.5 concentration. Meaning, the coefficient for poverty rate was not significant and we can estimate that it is zero.

We can delve in a bit further into the meaning of these results with the context of the model. The intercept of the model indicates that when there 0 mean concentration of PM 2.5 in the air, there is an air quality index of 33. This may make sense as PM 2.5 might not be the only predictor used to calculate this metric, but also this might just be the case due to the high leverage points that skew the model. This model predicts that after accounting for poverty rate, a 1 unit increase in the mean concentration of PM 2.5 will increase the Air Quality Index by 1.90 units.

Combining these two models together, we yield the following mega model:

$$\hat{AQI} = \hat{AQI}_1 \cdot I(PM_{2.5} < 9.5) + \hat{AQI}_2 \cdot I(PM_{2.5} \geq 9.5)$$

where I is an indicator variable that allows us to switch between our two multiple regression models based on the PM 2.5 concentration value.

Model Usage and Conclusion

This model may be used to predict the daily air quality index for Austin and the McAllen-Mission-Edinburg areas in 2022. To provide a comprehensive understanding of the predictors and our overall model, we can construct a 95% confidence interval to predict the mean daily air quality index for Austin and the McAllen-Mission-Edinburg areas in 2022. The mean daily PM 2.5 concentration for both areas is 8.79 units, and the mean poverty rate is 16.75. With these values, we can construct our interval or we can use R. Since the mean PM 2.5 value is below our threshold of 9.5, we will use our first model. We calculate the 95% confidence interval in Figure Set 5 and yield the following interval: $(48.728, 48.786)$

This is to say that we are 95% confident that the true mean air quality index for Austin and the McAllen-Mission-Edinburg areas with an average PM 2.5 concentration of 8.79 and average poverty rate of %16.75 is in between 48.72824 and 48.78628 units.

We can also construct 95% confidence intervals for the true coefficient slopes for our predictor variables. Based on the Figure Set 6 for example, we can say that we are 95% confident that the true population slope for mean daily PM 2.5 concentration less than 9.5 units, from our first model, is in between 5.526 and 5.544. For our second model, we say we are 95% confident the true population slope for mean daily PM 2.5 concentration greater than or equal to 9.5, is in between 1.890 and 1.906.

To conclude, the development of these two distinct models enabled us to account for a threshold in the air quality index variable. This way we have a nuanced predictive tool to further understand how air quality is measured. In the future, I would like to add weather data like daily mean temperature to assess if this variable would add to predicting the air quality index. For now, the two models seem reliable on these two regions based on their respective R^2 values, though we should not be too confident in this since we still have

to watch for over fitting and apply this model to untrained data. Still, this model offers the result that PM 2.5 concentration may be a better determinant of air quality than poverty.

Appendix

```
library(ggplot2)
library(dplyr)
```

Cleaning Data

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Data from U.S. Census Poverty Data
poverty_data = data.frame(
  region = c("Travis County", "McAllen-Mission-Edinburg"),
  rate = c(11.30, 27.70) )

# Data from U.S. Environmental Protection Agency
rgv_data = read.csv("C:\\Users\\mllam\\Downloads\\rgv_pm25_data_2022.csv")

austin_data = read.csv("C:\\Users\\mllam\\OneDrive\\Desktop\\austin-round-rock_pm25_data_2022.csv")

# Manually adding Poverty data for each region
rgv_data["Poverty Rate"] <- 27.70
```

```r
austin_data["Poverty Rate"]<- 11.30

# Cleaning Data: Removing Negative Values of Mean PM 2.5 Concentration

rgv_data <- rgv_data |>
              filter(Daily.Mean.PM2.5.Concentration >= 0)|>
              rename(pm = Daily.Mean.PM2.5.Concentration)|>
              rename(pov = `Poverty Rate`) |>
              rename(aqi = Daily.AQI.Value) |>
              select(Date, pm, Units, aqi, pov, County)

austin_data <- austin_data |>
              filter(Daily.Mean.PM2.5.Concentration >= 0) |>
              rename(pm = Daily.Mean.PM2.5.Concentration)|>
              rename(pov = `Poverty Rate`)|>
              rename(aqi = Daily.AQI.Value) |>
              select(Date, pm, Units, aqi, pov, County)

# Combining Data from Austin and Rio Grande Valley
data = rbind(austin_data, rgv_data)

sample_n(data, 3)
```

```
##        Date   pm    Units aqi  pov  County
## 1 9/14/2022 15.0 ug/m3 LC  62 11.3  Travis
## 2 2/18/2022  5.0 ug/m3 LC  28 11.3  Travis
## 3 7/29/2022  7.2 ug/m3 LC  40 27.7 Hidalgo
```

```r
# Attempt at first linear model with no threshold for PM 2.5
linear_model = lm(aqi ~ pm + pov, data = data)

plot(linear_model$fitted.values ,linear_model$residuals, col = '#dd9bb1', pch = 1, ylab = "Residuals", x
     main = "Assessing Linearity: Residuals vs. Fits")
```

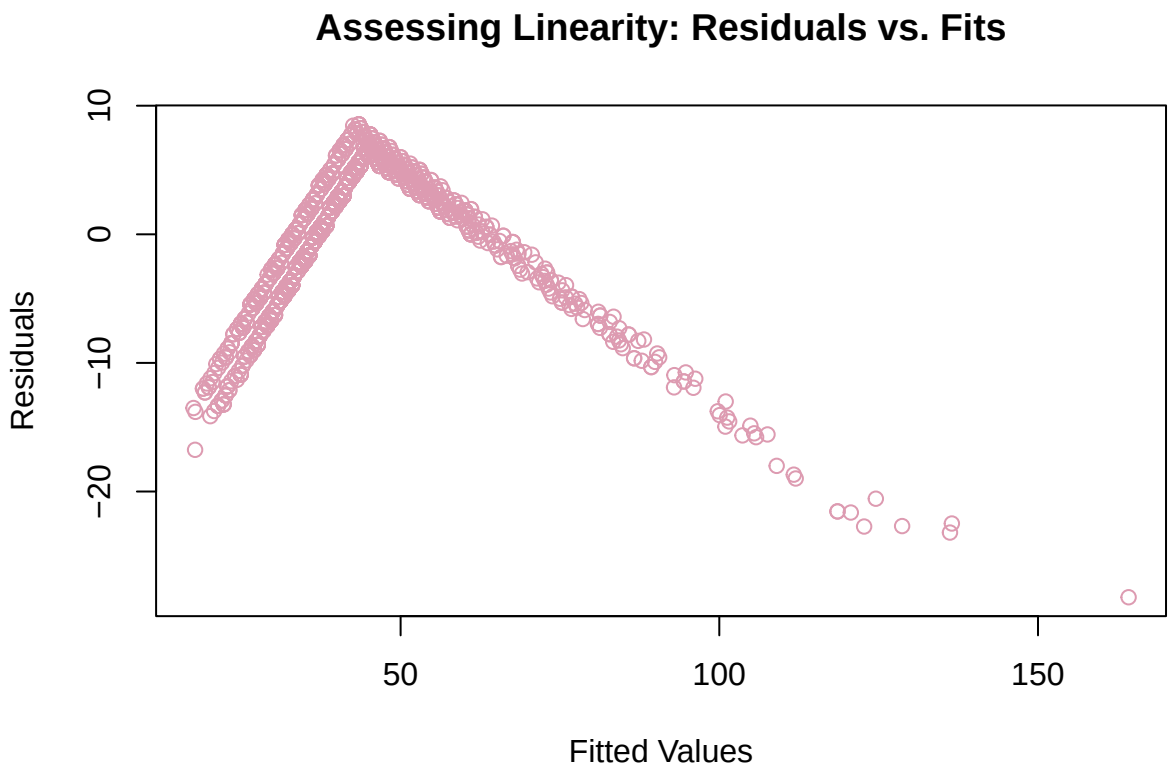


Figure 1:

```
plot(data$pm, data$aqi, col = '#dd9bb1', pch = 1, ylab = "Air Quality Index", xlab = "Daily Mean PM 2.5
abline(v=9.5, col = "maroon", lwd = 2)
```

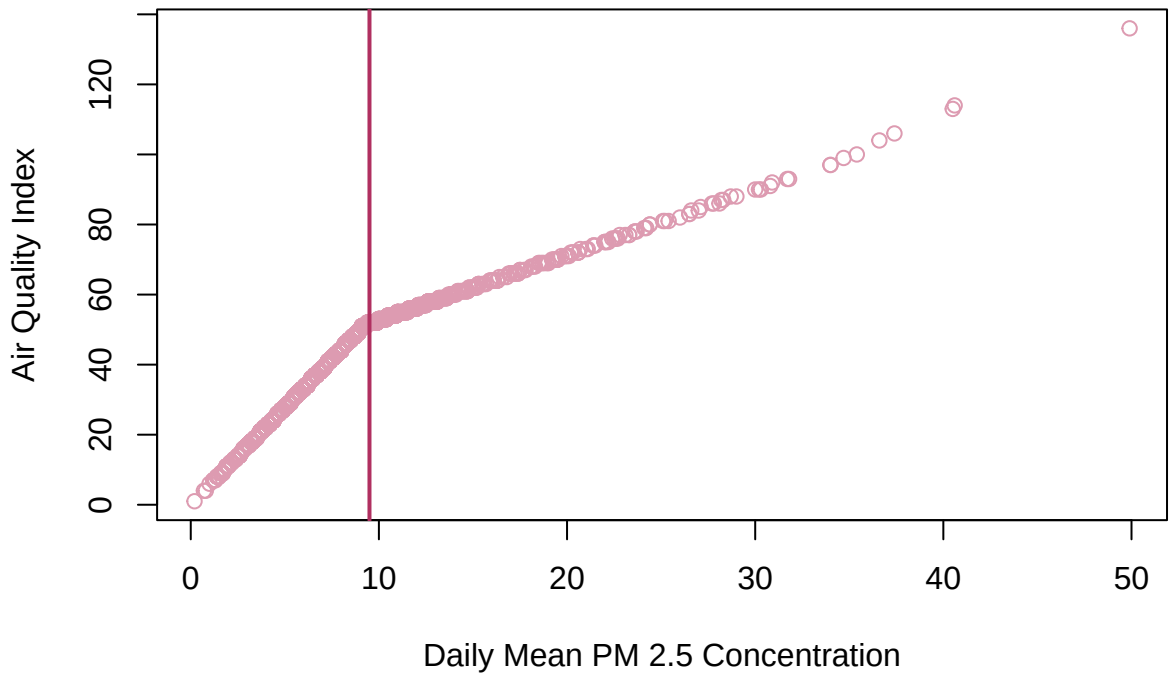


Figure 2:

```
data_less <- data |>
  filter(pm <= 9.5)

lm_less = lm(aqi ~ pm + pov, data = data_less)
plot(lm_less)
```

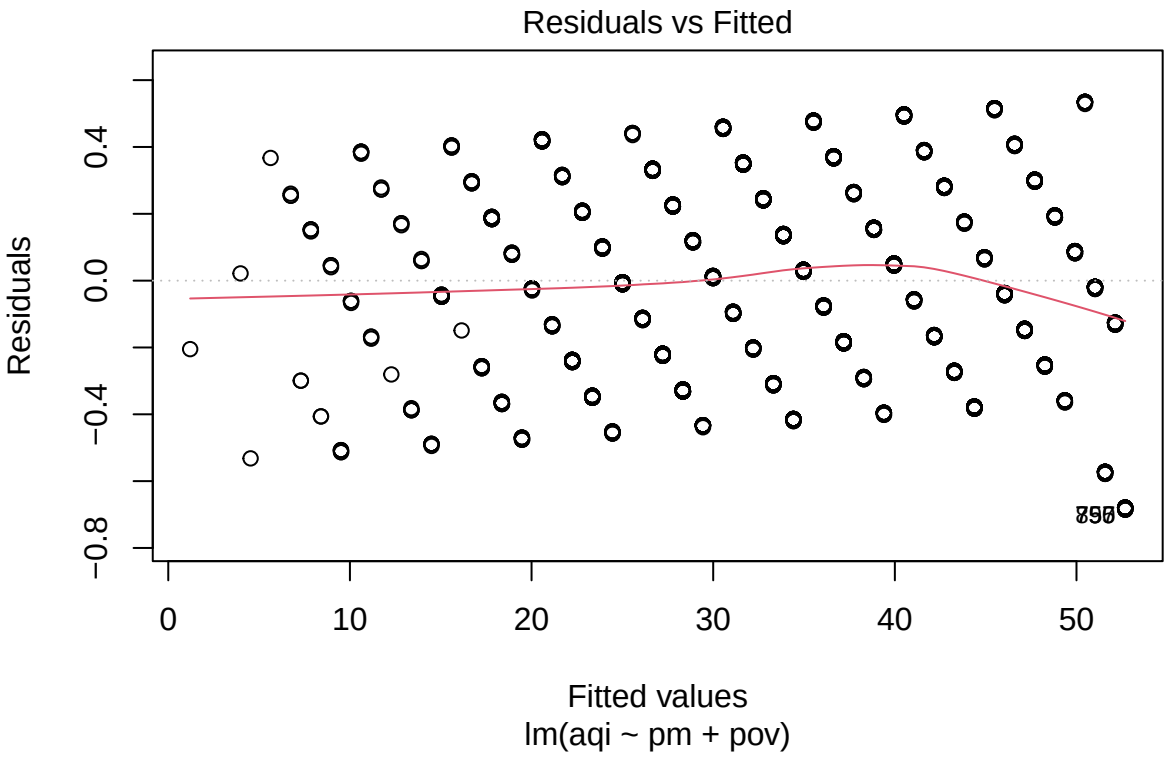
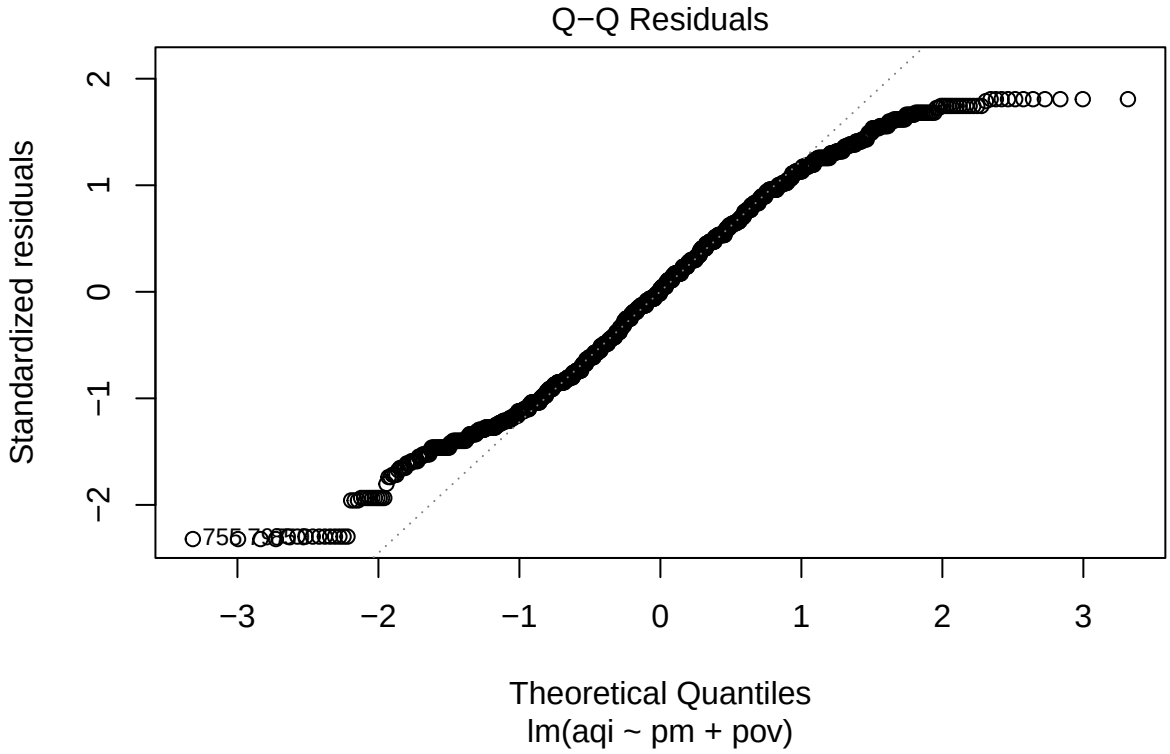
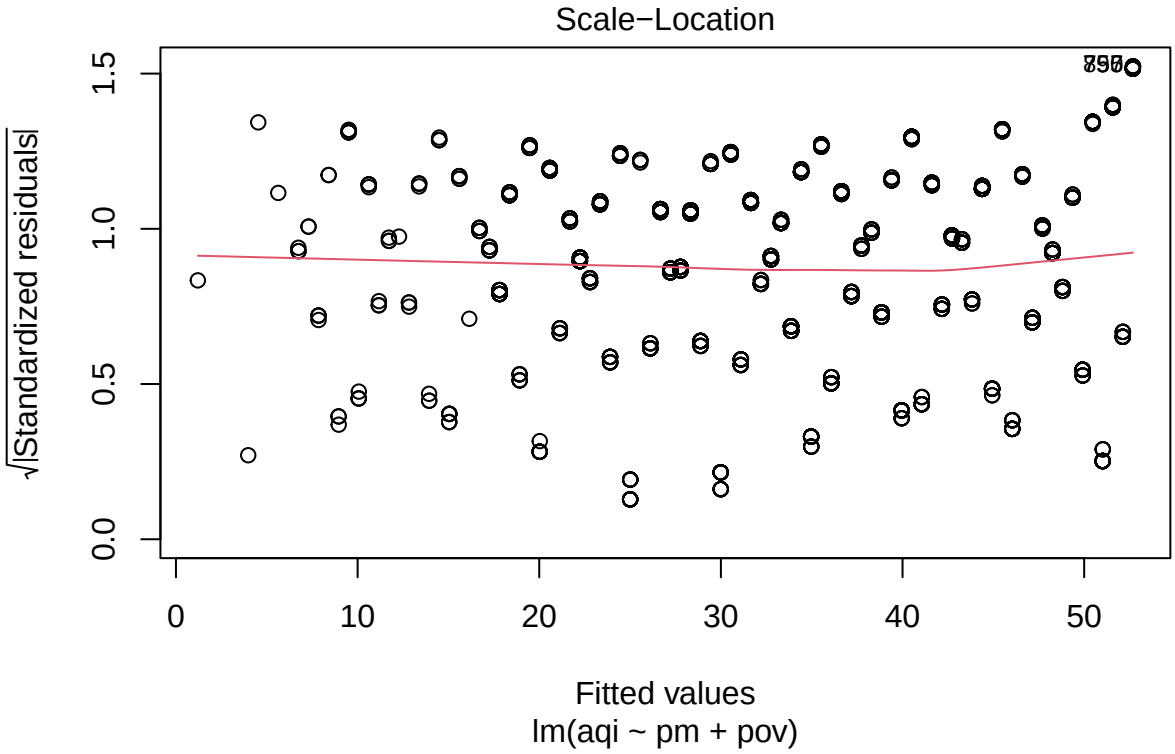
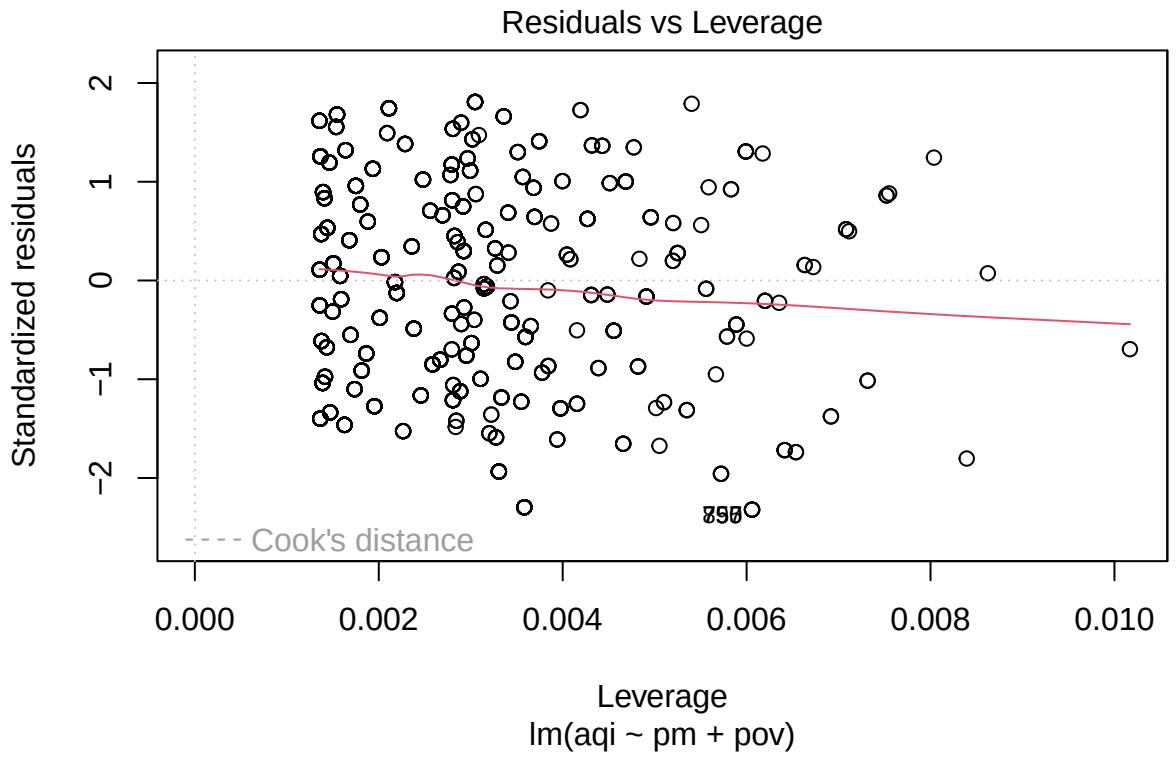

Residuals vs Fitted
Residuals
0.4
0.0
−0.4
−0.8
0
10
20
30
40
50
Fitted values
lm(aqi ~ pm + pov)

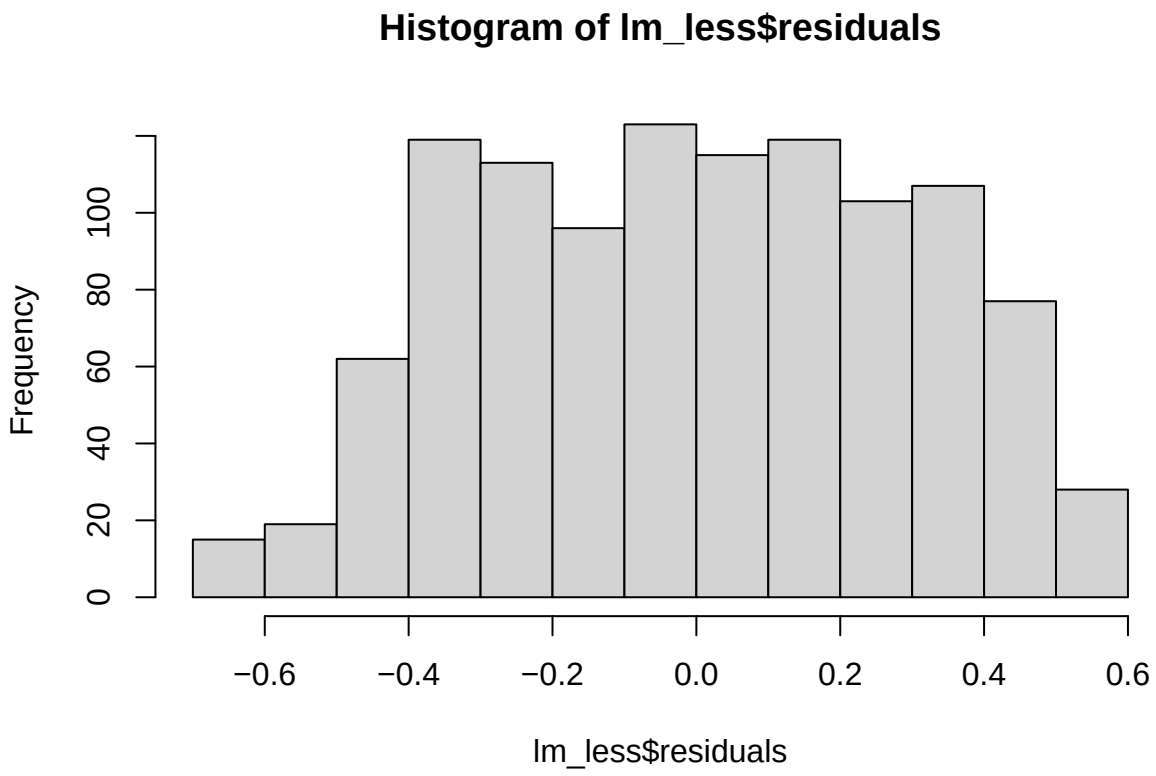
Figure Set 3:


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(aqi ~ pm + pov)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
10
20
30
40
50
Fitted values
lm(aqi ~ pm + pov)




```
hist(lm_less$residuals)
```




```
summary(lm_less)
```

```
##
## Call:
## lm(formula = aqi ~ pm + pov, data = data_less)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6853 -0.2435  0.0077  0.2463  0.5347
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 0.0938112  0.0374189    2.507   0.0123 *
## pm          5.5348704  0.0044936 1231.735   <2e-16 ***
## pov         0.0003678  0.0011770    0.313   0.7547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2962 on 1093 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9993
## F-statistic: 7.766e+05 on 2 and 1093 DF,  p-value: < 2.2e-16
```

```
require(MASS)
```

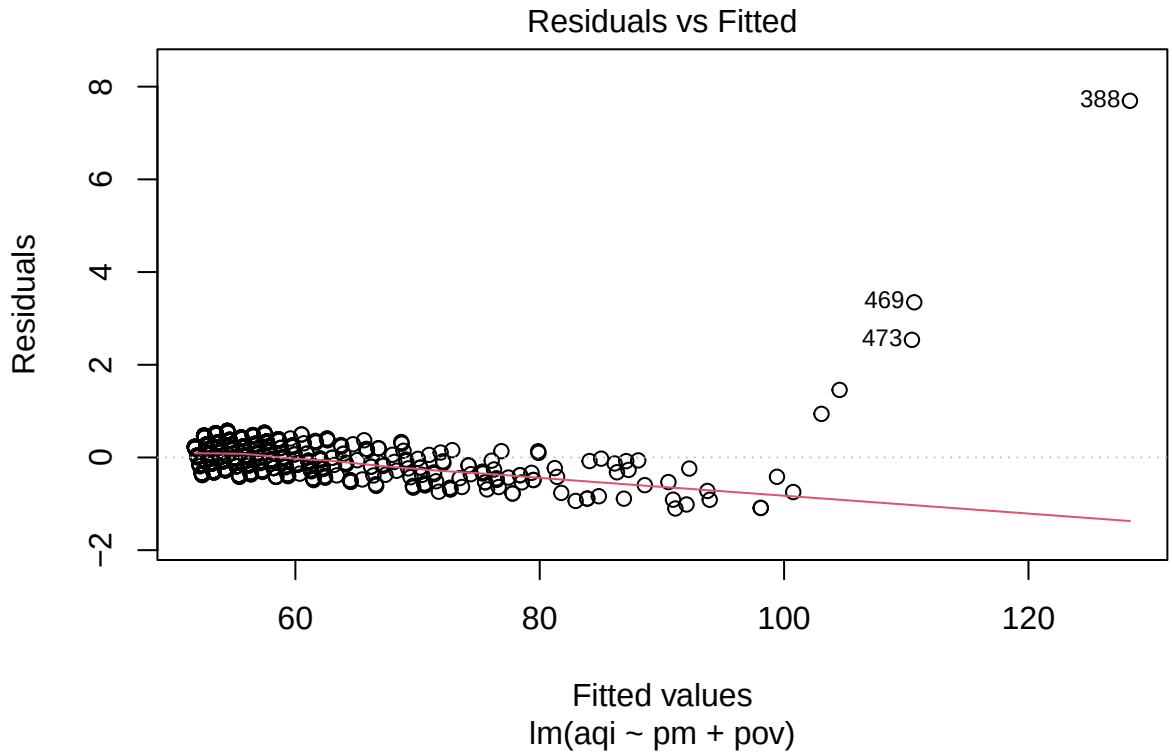
Figure Set 4:

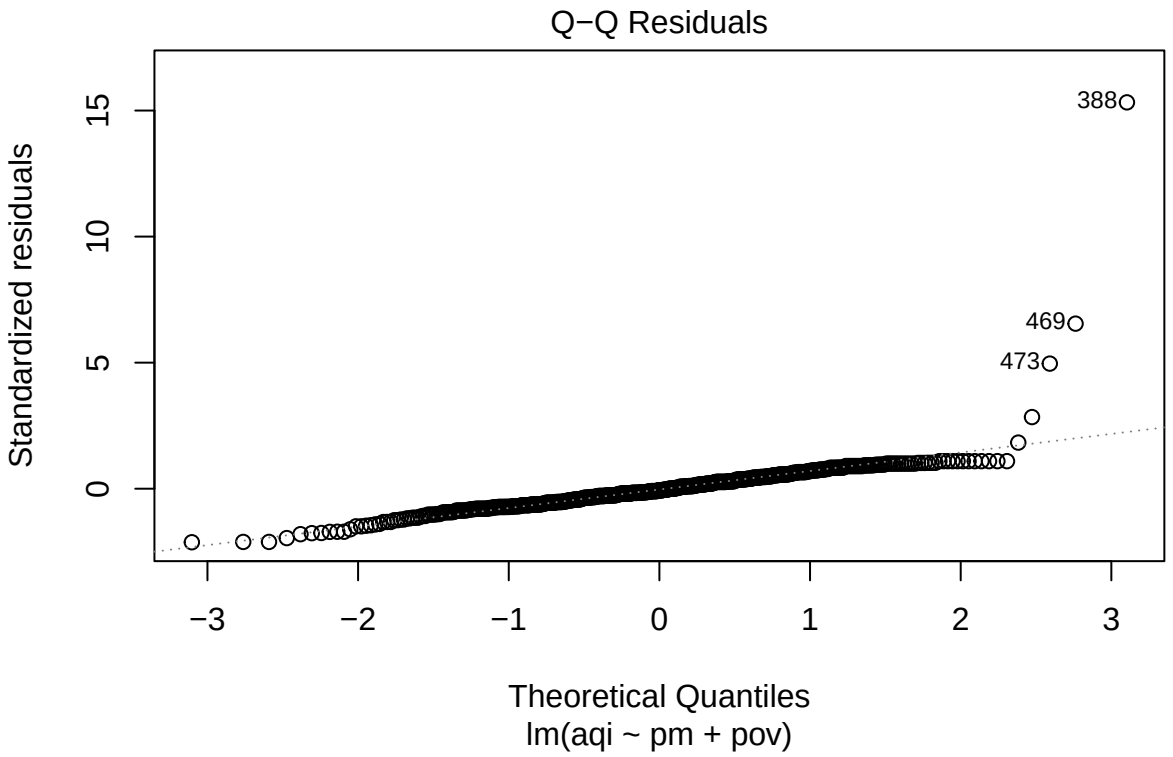
```
## Loading required package: MASS
```

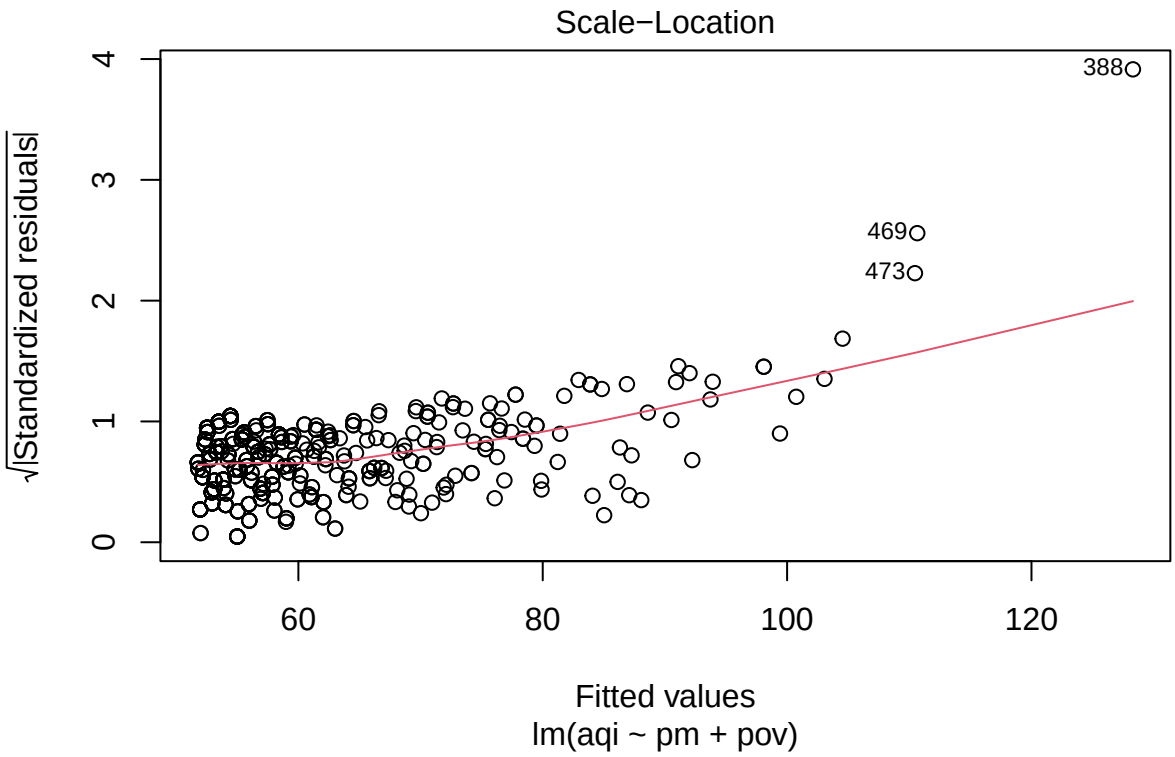
```
##
## Attaching package: 'MASS'
```

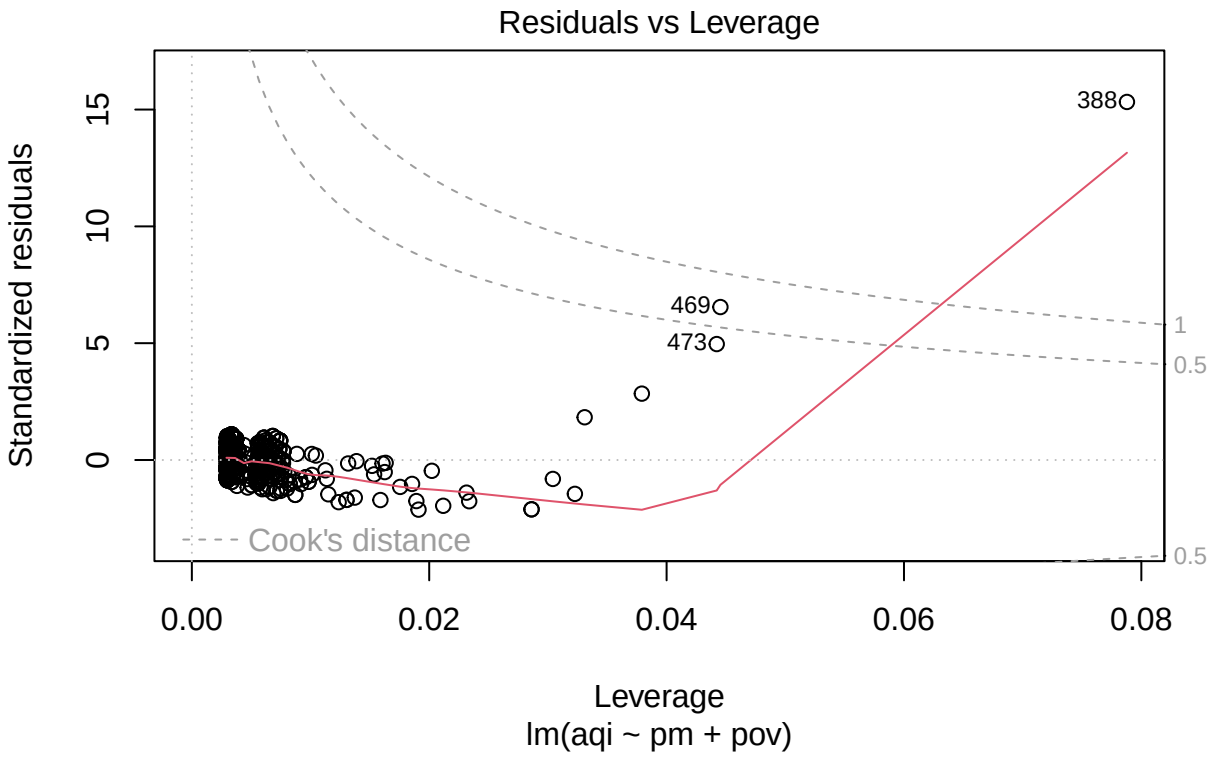
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
data_greater <- data |>
  filter(pm > 9.5)

lm_greater = lm(aqi ~ pm + pov, data = data_greater)

plot(lm_greater)
```

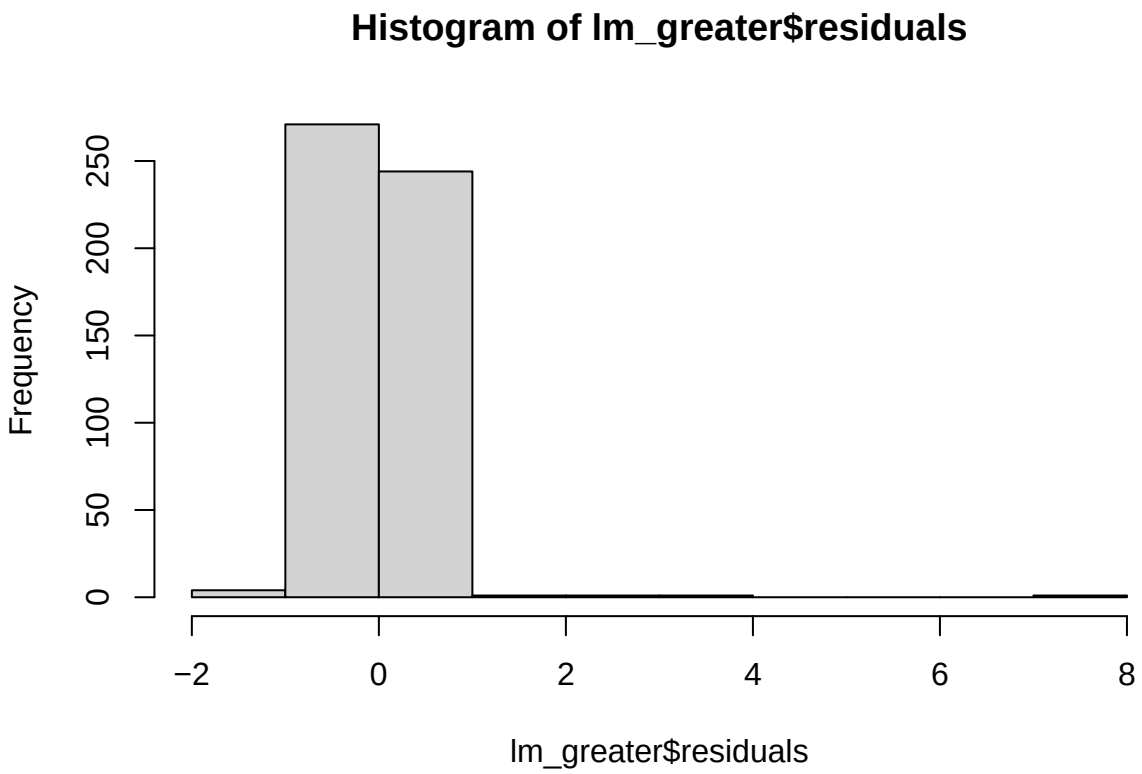



Q−Q Residuals
Standardized residuals
0
5
10
15
388
469
473
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(aqi ~ pm + pov)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
388
469
473
60
80
100
120
Fitted values
lm(aqi ~ pm + pov)




```
hist(lm_greater$residuals)
```




```
summary(lm_greater)
```

```
##
## Call:
## lm(formula = aqi ~ pm + pov, data = data_greater)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1003 -0.2758 -0.0342  0.2419  7.6945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.524602   0.074650 449.090   <2e-16 ***
## pm           1.898220   0.004116 461.129   <2e-16 ***
## pov          0.002154   0.002974   0.724    0.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5231 on 520 degrees of freedom
## Multiple R-squared:  0.9976, Adjusted R-squared:  0.9976
## F-statistic: 1.091e+05 on 2 and 520 DF,  p-value: < 2.2e-16
```

```
Studentized = studres(lm_greater)
which(abs(Studentized)>2)
```

```
##  48 272 275 388 407 469 473
##  48 272 275 388 407 469 473
```

```
lev_values = hatvalues(lm_greater)
which(lev_values > 2*mean(lev_values))
```

```
##  29  38  41  48 141 151 154 155 167 260 272 273 275 276 293 387 388 407 408 413
##  29  38  41  48 141 151 154 155 167 260 272 273 275 276 293 387 388 407 408 413
## 421 422 450 461 469 470 472 473 488 489
## 421 422 450 461 469 470 472 473 488 489
```

```
# Rows with High Leverage
row_numbers <- c(29, 38, 41, 48, 141, 151, 154, 155, 167, 260, 272, 273, 275, 276,
                 293, 387, 388, 407, 408, 413, 421, 422, 450, 461, 469, 470, 472,
                 473, 488, 489)

high_lev <- data_greater%>%
  slice(row_numbers)

high_lev
```

```
##         Date   pm    Units aqi  pov  County
## 1  5/20/2022 27.7 ug/m3 LC  86 11.3  Travis
## 2  6/13/2022 31.8 ug/m3 LC  93 11.3  Travis
## 3  6/16/2022 35.4 ug/m3 LC 100 11.3  Travis
## 4  7/17/2022 34.0 ug/m3 LC  97 11.3  Travis
## 5  5/20/2022 27.8 ug/m3 LC  86 11.3  Travis
## 6  6/13/2022 30.8 ug/m3 LC  91 11.3  Travis
## 7  6/16/2022 34.7 ug/m3 LC  99 11.3  Travis
## 8  6/17/2022 25.4 ug/m3 LC  81 11.3  Travis
## 9  7/17/2022 29.0 ug/m3 LC  88 11.3  Travis
## 10 5/20/2022 28.2 ug/m3 LC  87 11.3  Travis
## 11 6/13/2022 34.0 ug/m3 LC  97 11.3  Travis
## 12 6/14/2022 28.3 ug/m3 LC  87 11.3  Travis
## 13 6/16/2022 37.4 ug/m3 LC 106 11.3  Travis
## 14 6/17/2022 28.1 ug/m3 LC  86 11.3  Travis
## 15 7/17/2022 31.7 ug/m3 LC  93 11.3  Travis
## 16  4/4/2022 30.2 ug/m3 LC  90 27.7 Hidalgo
## 17  4/5/2022 49.9 ug/m3 LC 136 27.7 Hidalgo
## 18  5/6/2022 30.3 ug/m3 LC  90 27.7 Hidalgo
## 19  5/7/2022 26.6 ug/m3 LC  84 27.7 Hidalgo
## 20 5/20/2022 27.1 ug/m3 LC  85 27.7 Hidalgo
## 21 7/16/2022 27.0 ug/m3 LC  84 27.7 Hidalgo
## 22 7/17/2022 26.0 ug/m3 LC  82 27.7 Hidalgo
## 23  4/5/2022 36.6 ug/m3 LC 104 27.7 Hidalgo
## 24 5/20/2022 30.0 ug/m3 LC  90 27.7 Hidalgo
## 25 6/12/2022 40.6 ug/m3 LC 114 27.7 Hidalgo
## 26 6/13/2022 26.5 ug/m3 LC  83 27.7 Hidalgo
## 27 6/15/2022 30.9 ug/m3 LC  92 27.7 Hidalgo
## 28 6/16/2022 40.5 ug/m3 LC 113 27.7 Hidalgo
## 29 7/16/2022 26.5 ug/m3 LC  83 27.7 Hidalgo
## 30 7/17/2022 28.7 ug/m3 LC  88 27.7 Hidalgo
```

```r
mean_pm = mean(data$pm)
mean_pov = mean(data$pov)

new_data = data.frame(pm = c(mean_pm), pov = c(mean_pov))

predict(lm_less, newdata = new_data, interval="confidence", level = 0.95)
```

Figure Set 5:

```
##        fit      lwr      upr
## 1 48.75726 48.72824 48.78628
```

```r
library(broom)
tidy(lm_less, conf.int = T, conf.level = 0.95)
```

Figure Set 6:

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept) 0.0938     0.0374      2.51   0.0123  0.0204    0.167
## 2 pm          5.53       0.00449  1232.     0       5.53      5.54
## 3 pov         0.000368   0.00118     0.313  0.755  -0.00194   0.00268
```

```r
tidy(lm_greater, conf.int = T, conf.level = 0.95)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept) 33.5       0.0747    449.      0     33.4      33.7
## 2 pm           1.90      0.00412   461.      0      1.89      1.91
## 3 pov          0.00215   0.00297     0.724   0.469 -0.00369   0.00800
```

```r
# add more cities ANOVA
# dallas county - 14%,
# houston county - 15.8%
# houston city - 19%
# Amarillo city - 15.5%
# Lubbock city - 19.3%
# el paso city - 18.8%
# el paso county - 19.5%

aov_model = aov(aqi ~ County, data = data)
```