Effects of Social Distancing and Border Lockdowns on Pollution in 6 Different States in the USA

Sameeha Patel, Reem AlRowaili, Luca Martial - March 29th 2020

null

Contents

Data Extraction

Note: all extractions attempt to limit data to 6 US states for the purposes of the project: Washington, New York, Massachusetts, California, Illinois and Wisconsin.

Pollution Data

Data on PM 2.5 concentration (in g/m3), were retrieved from all atmospheric sensor stations in each state and aggregated to output a single value per day per state from January 2018 to March 2020. Data was made freely available by the United States Environmental Protection Agency.

Python code can be viewed here.

Qualitative Measures for Social Distancing Levels

Level of social distancing measures was classified using 4 levels of lockdown from lowest to highest: normal level (no restrictions on movement), emergency state, school shut down and shelter-in-place. Data was categorized manually using readily available websites from January 2018 to March 2020. These can be automated by use of web scraping in the future.

Raw data files can be viewed here.

COVID-19 Cases

When examining the effect of border lockdown level on pollution levels and of social isolation level on pollution levels, it appears that the number of incident cases of COVID-19 may be directly associated with both exposures (border lockdown level and social isolation level) and indirectly associated with the outcome (pollution levels). In effect, an increase in COVID-19 cases would suggest that stricter measures would be put in place to limit infection transmission. An increase in COVID-19 cases may also suggest a limit in workforce supply, thereby limiting pollution caused by transport and industrial activity for example. These associations imply that the variable fits the classical criteria for confounding. Number of new incident cases per state from January 2018 to March 2020 was determined using a publicly available dataset.

Python code can be viewed here.

Data Merging

All data were merged together in long format. Python code can be viewed here.

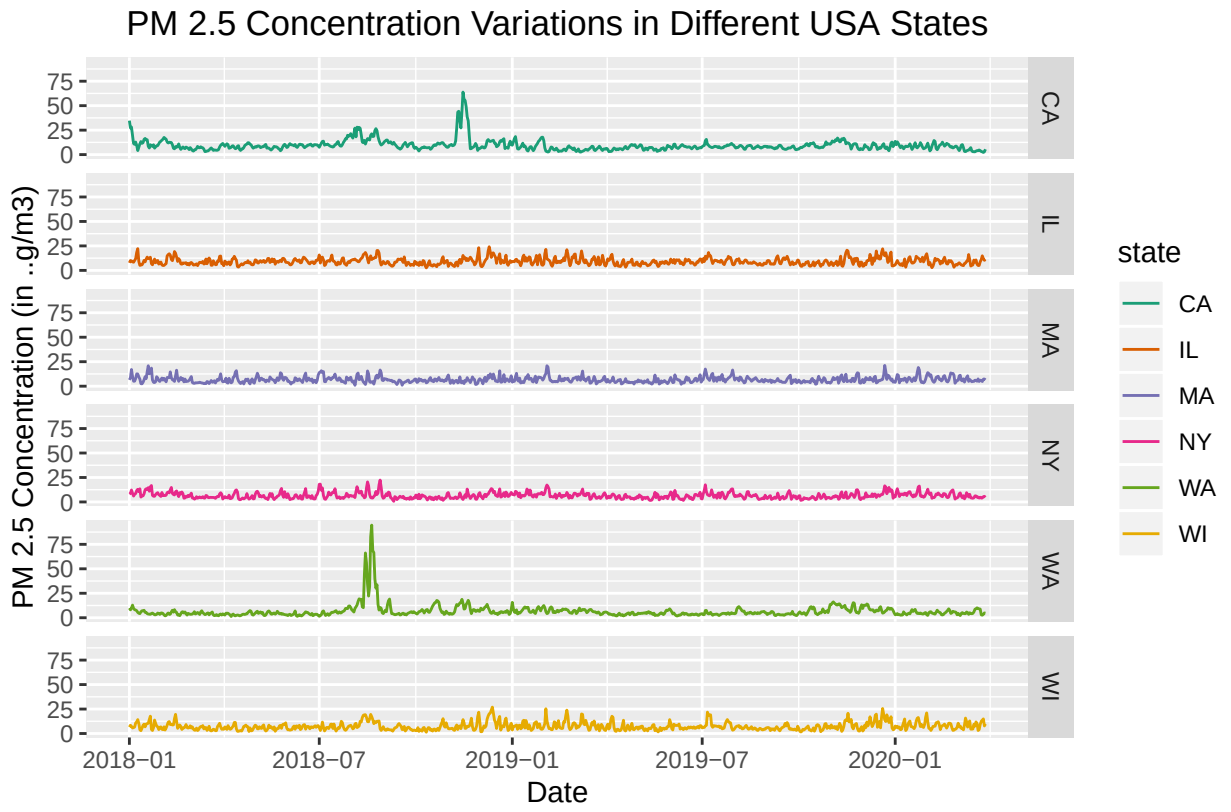
Statistical Analysis

Pollution Exploratory Data Analysis

```
# Reading in pollution data
pol <- read.csv('air_state_pollution_pm25.csv')

# Renaming columns
names(pol) <- c('date', 'pm', 'state')

# Converting date column to datetime format
pol$date <- anytime::anydate(pol$date)

# Checking summary of data
summary(pol)
```

```
##       date                  pm          state
##  Min.   :2018-01-01   Min.   : 0.6517   CA:817
##  1st Qu.:2018-07-24   1st Qu.: 4.7411   IL:817
##  Median :2019-02-13   Median : 6.7474   MA:817
##  Mean   :2019-02-12   Mean   : 7.6478   NY:817
##  3rd Qu.:2019-09-05   3rd Qu.: 9.3143   WA:817
##  Max.   :2020-03-27   Max.   :94.6136   WI:814
```

```
# Plotting facet plot of pm2.5 variations over time
ggplot(pol, aes(x = date, y = pm, color = state)) +
  geom_line() +
  facet_grid(state~.) +
  scale_color_brewer(palette="Dark2") +
  labs(title = 'PM 2.5 Concentration Variations in Different USA States', caption = 'Based on data from
  theme(plot.title = element_text(hjust = 0.5))
```



Based on data from: https://www.epa.gov/outdoor−air−quality−data/download−daily−data

There appear to be several outliers of PM 2.5 measurements, particularly in the states of California and Washington.

Merged data EDA

Describe numerically and graphically the relationships between variables

```
# Reading in merged data
df.merged <- read.csv('pol_iso_merged.csv')

# Renaming columns
names(df.merged) <- c('date', 'pm', 'state', 'iso_status', 'cases')

# Setting date column as datetime
df.merged$date <- anytime::anydate(df.merged$date)

# Releveling iso status in correct order
df.merged$iso_status <- factor(df.merged$iso_status, levels = c('Normal', 'Emergency', 'SchoolClosure',

# Exploring dataframe
head(df.merged)
```

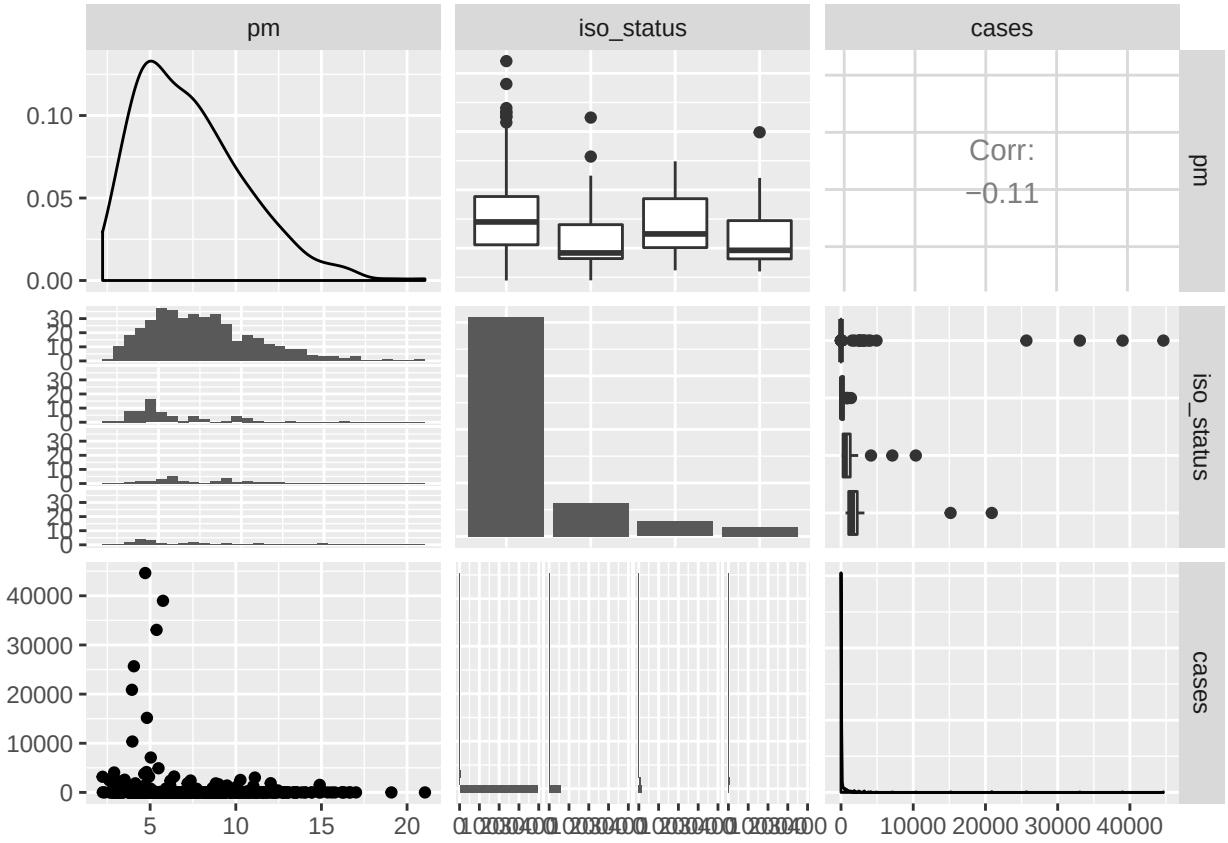
```
##         date       pm state iso_status cases
## 1 2018-01-01 34.69667    CA     Normal     0
## 2 2018-01-02 27.24974    CA     Normal     0
```

```
## 3 2018-01-03 28.13387    CA     Normal     0
## 4 2018-01-04 19.48033    CA     Normal     0
## 5 2018-01-05 11.33473    CA     Normal     0
## 6 2018-01-06 10.45164    CA     Normal     0
```

```
# Exploring dataframe
summary(df.merged)
```

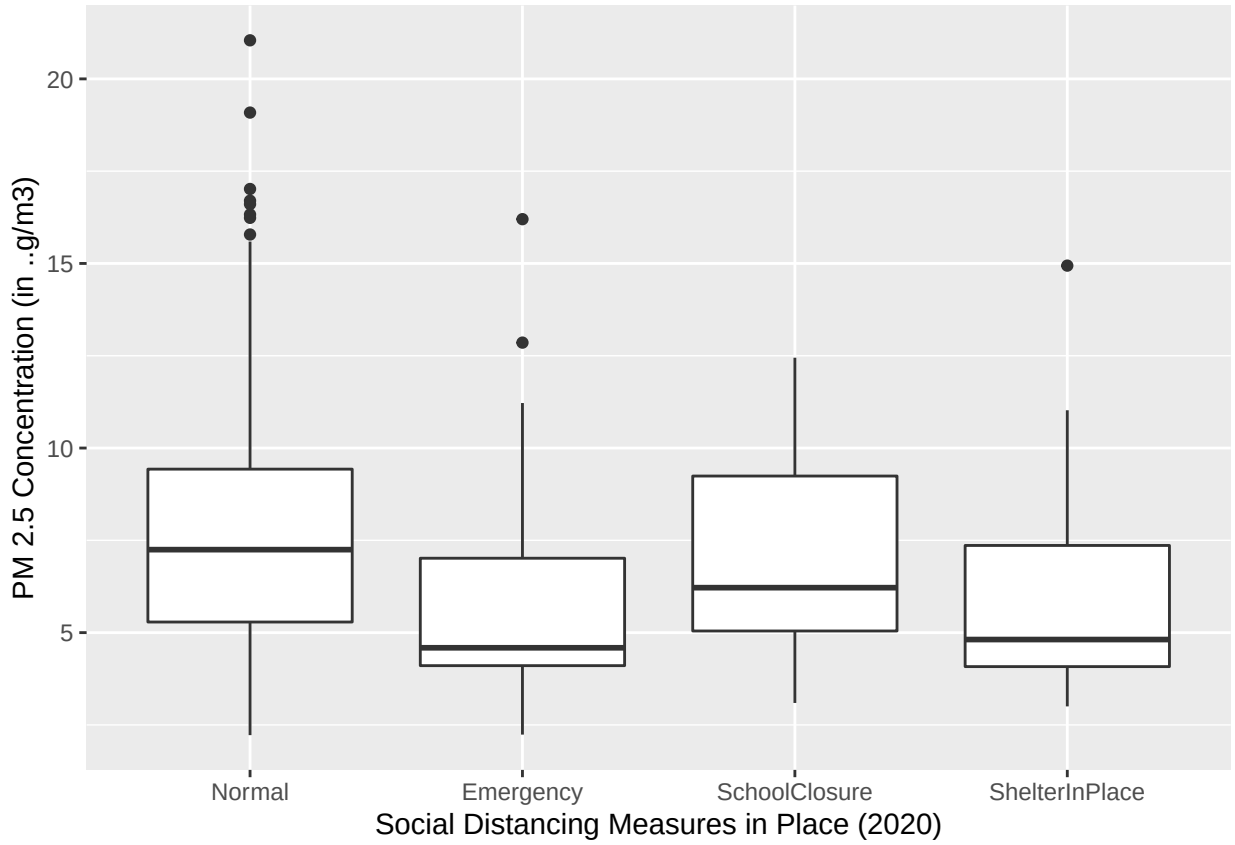
```
##       date                 pm               state             iso_status
##  Min.   :2018-01-01   Min.   : 0.6517   CA:817   Normal         :4790
##  1st Qu.:2018-07-24   1st Qu.: 4.7405   IL:817   Emergency      :  63
##  Median :2019-02-13   Median : 6.7459   MA:817   SchoolClosure :  28
##  Mean   :2019-02-12   Mean   : 7.6474   NY:817   ShelterInPlace:  17
##  3rd Qu.:2019-09-05   3rd Qu.: 9.3132   WA:817
##  Max.   :2020-03-27   Max.   :94.6136   WI:813
##      cases
##  Min.   :    0.00
##  1st Qu.:    0.00
##  Median :    0.00
##  Mean   :   59.63
##  3rd Qu.:    0.00
##  Max.   :44635.00
```

```
# Exploring relationships between pm 2.5 levels, state and isolation status for 2020
latest.df <- df.merged %>% filter(df.merged$date >= as.Date('2020-01-01'))
ggpairs(latest.df[ , c('pm', 'iso_status', 'cases')])
```



When subsetting to 2020, we first observe a weak negative correlation between incidence of COVID-19 cases and PM 2.5 concentration (-0.11). We also observe that there appears to be a difference in PM 2.5 levels between all social distancing measures in place. This plot can be further examined and plotted for the same year (2020).

```
# Plotting boxplot for iso status and pm 2.5
ggplot(latest.df, aes(x = iso_status, y = pm)) +
  geom_boxplot() +
  labs(x = 'Social Distancing Measures in Place (2020)', y = 'PM 2.5 Concentration (in  g/m3)')
```



Outliers seem to be slightly affecting the overall levels of PM 2.5 during periods without social distancing measures in place. We observe a slight difference in PM 2.5 concentrations between each increase in level of social distancing measure in the year 2020. It can be argued that the plot seems to be following a somewhat linear trend.

Performing simple linear regressions

```
# SLR with iso status and pm 2.5
slr.iso <- lm(pm ~ iso_status, data = df.merged)
summary(slr.iso)
```

```
##
## Call:
## lm(formula = pm ~ iso_status, data = df.merged)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.028 -2.878 -0.915  1.650 86.934
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.67984    0.07029 109.257  < 2e-16 ***
## iso_statusEmergency       -1.94199    0.61693  -3.148  0.00165 **
## iso_statusSchoolClosure   -0.40052    0.92206  -0.434  0.66403
## iso_statusShelterInPlace -1.49782    1.18200  -1.267  0.20515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.865 on 4894 degrees of freedom
## Multiple R-squared:  0.002369,   Adjusted R-squared:  0.001758
## F-statistic: 3.874 on 3 and 4894 DF,  p-value: 0.008848
```

```
# Displaying confidence intervals
confint(slr.iso)
```

```
##                              2.5 %     97.5 %
## (Intercept)               7.542033  7.8176397
## iso_statusEmergency      -3.151452 -0.7325179
## iso_statusSchoolClosure  -2.208171  1.4071235
## iso_statusShelterInPlace -3.815060  0.8194289
```

Interpretation and comments: When shifting from 'Normal' social distancing status to 'Emergency' social distancing status, we are 95% confident that average PM 2.5 concentrations decrease between 0.73 and 3.15 g/m3. Although the confidence interval suggests that this may be a significant decrease, we must bear in mind that these are unadjusted estimates and that the standard errors are moderately wide.

```
# SLR with cases and pm 2.5
slr.cases <- lm(pm ~ cases, data = df.merged)
summary(slr.cases)
```

```
##
## Call:
## lm(formula = pm ~ cases, data = df.merged)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.002 -2.906 -0.901  1.667 86.959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.654e+00  6.965e-02 109.889   <2e-16 ***
## cases       -1.128e-04  6.145e-05  -1.836   0.0664 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.868 on 4896 degrees of freedom
## Multiple R-squared:  0.0006883,  Adjusted R-squared:  0.0004842
## F-statistic: 3.372 on 1 and 4896 DF,  p-value: 0.06636
```

```r
# Displaying confidence intervals
confint(slr.cases)
```

```
##                     2.5 %       97.5 %
## (Intercept)  7.5175472301 7.790650e+00
## cases       -0.0002333012 7.623732e-06
```

Interpretation and comments: For every increase in 1 case, we are 95% confident that average PM 2.5 concentrations decrease between 0.000008 and 0.000233 g/m3. Once again, we must bear in mind that these are unadjusted estimates.

```r
# SLR with state and pm 2.5
slr.state <- lm(pm ~ state, data = df.merged)
summary(slr.state)
```

```
##
## Call:
## lm(formula = pm ~ state, data = df.merged)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.172 -2.606 -0.903  1.423 87.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.3951     0.1651  56.900   <2e-16 ***
## stateIL      -0.1879     0.2335  -0.805    0.421
## stateMA      -2.6911     0.2335 -11.524   <2e-16 ***
## stateNY      -2.8939     0.2335 -12.393   <2e-16 ***
## stateWA      -2.7444     0.2335 -11.753   <2e-16 ***
## stateWI      -1.9703     0.2338  -8.427   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.72 on 4892 degrees of freedom
## Multiple R-squared:  0.06146,    Adjusted R-squared:  0.0605
## F-statistic: 64.07 on 5 and 4892 DF,  p-value: < 2.2e-16
```

```r
# Displaying confidence intervals
confint(slr.state)
```

```
##                  2.5 %     97.5 %
## (Intercept)  9.0714348  9.7188401
## stateIL     -0.6457277  0.2698417
## stateMA     -3.1488582 -2.2332889
## stateNY     -3.3516946 -2.4361252
## stateWA     -3.2022325 -2.2866631
## stateWI     -2.4286697 -1.5119748
```

Interpretation and comments: The states of Massachusetts, New York, Washington and Wisconsin seem to observe significantly lower concentrations of PM 2.5 when compared with California. Once again, although the confidence interval suggests that this may be a significant decrease, we must bear in mind that these are unadjusted estimates and that the standard errors remain moderately wide.

```r
# SLR with date and pm 2.5
slr.date <- lm(pm ~ date, data = df.merged)
summary(slr.date)
```

```
##
## Call:
## lm(formula = pm ~ date, data = df.merged)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.129 -2.899 -0.896  1.662 86.815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.0123260  5.2917092   4.349  1.4e-05 ***
## date        -0.0008565  0.0002949  -2.904   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.865 on 4896 degrees of freedom
## Multiple R-squared:  0.001719,   Adjusted R-squared:  0.001515
## F-statistic: 8.432 on 1 and 4896 DF,  p-value: 0.003703
```

```r
# Displaying confidence intervals
confint(slr.date)
```

```
##                    2.5 %        97.5 %
## (Intercept) 12.638201885 33.3864501312
## date        -0.001434705 -0.0002782503
```

Interpretation and comments: For every increase in 1 day from January 1st 2018 to March 27th 2020, we are 95% confident that average PM 2.5 concentrations decrease between 0.0003 and 0.0014 g/m3. Once again, although the confidence interval suggests that this may be a significant decrease, we must bear in mind that these are unadjusted estimates.

Checking linearity and independence assumption

Before fitting a full model we can verify the following assumptions:

- Linearity: a case has been made that linearity may be a possibility.
- Independence: a case can be made that all residuals are independent of each other.

Fitting a multiple linear regression model including all covariates in the model

```r
# MLR with all covariates
mlr.pm <- lm(pm ~ iso_status + cases + state + date, data = df.merged)
summary(mlr.pm)
```

```
##
## Call:
```

```
## lm(formula = pm ~ iso_status + cases + state + date, data = df.merged)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.931 -2.604 -0.886  1.420 87.813
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.131e+01  5.306e+00   4.016    6e-05 ***
## iso_statusEmergency       -1.703e+00  6.098e-01  -2.793  0.00525 **
## iso_statusSchoolClosure    4.227e-01  9.052e-01   0.467  0.64055
## iso_statusShelterInPlace -1.384e+00  1.170e+00  -1.183  0.23688
## cases                     -5.449e-05  6.103e-05  -0.893  0.37206
## stateIL                   -2.089e-01  2.333e-01  -0.896  0.37054
## stateMA                   -2.714e+00  2.334e-01 -11.630  < 2e-16 ***
## stateNY                   -2.897e+00  2.337e-01 -12.400  < 2e-16 ***
## stateWA                   -2.753e+00  2.334e-01 -11.797  < 2e-16 ***
## stateWI                   -2.000e+00  2.337e-01  -8.559  < 2e-16 ***
## date                      -6.619e-04  2.958e-04  -2.238  0.02529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.713 on 4887 degrees of freedom
## Multiple R-squared:  0.0652, Adjusted R-squared:  0.06329
## F-statistic: 34.09 on 10 and 4887 DF,  p-value: < 2.2e-16
```

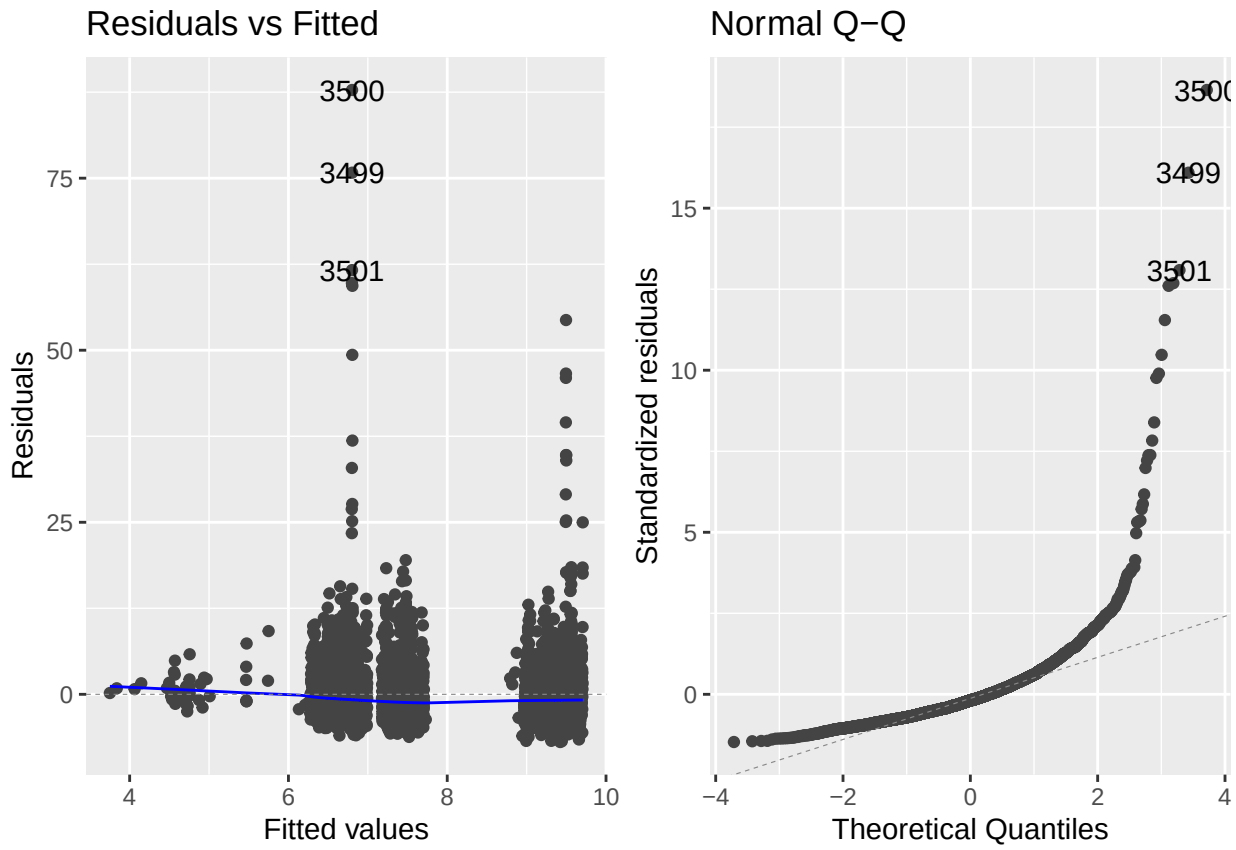
```
# Displaying confidence intervals
confint(mlr.pm)
```

```
##                                 2.5 %        97.5 %
## (Intercept)               10.9083766555  3.171318e+01
## iso_statusEmergency       -2.8984608814 -5.074959e-01
## iso_statusSchoolClosure   -1.3519440809  2.197363e+00
## iso_statusShelterInPlace -3.6771532643  9.095068e-01
## cases                     -0.0001741403  6.516981e-05
## stateIL                   -0.6662448146  2.484220e-01
## stateMA                   -3.1715721625 -2.256551e+00
## stateNY                   -3.3555073269 -2.439309e+00
## stateWA                   -3.2103274628 -2.295354e+00
## stateWI                   -2.4581576916 -1.541932e+00
## date                      -0.0012417617 -8.198961e-05
```

Interpretation and comments: When compared to 'Normal' social distancing status, we are 95% confident that 'Emergency' social distancing status is associated with an average PM 2.5 concentration decrease of between 0.51 and 2.90 g/m3, adjusting for incidence of COVID-19 cases, state and date of the year. We observe a very slight decrease in the standard error of the coefficient when compared to the simple linear regression model. The adjusted R-squared shows a poor fit of the model.

Checking residuals to verify model assumptions

```
# Verifying MLR model assumptions
autoplot(mlr.pm)[1:2]
```



We observe a large deviation from the 45 degree line in the QQplot. The normality assumption has not been met. The residual plot also shows us that the residuals do not observe a random pattern around the fitted regression line. This is enough evidence to show that the homoscedasticity assumption has not been met, giving us a strong reason to investigate other models. In order to do so, we need to pay closer attention to each covariate.

Exploring a log transformation may produce a model that meets the linearity and homoscedasticity assumptions.

Refitting a multiple linear regression model with log transformation

```
# MLR with all covariates
mlr.pmlog <- lm(log(pm) ~ iso_status + cases + state + date, data = df.merged)
summary(mlr.pmlog)
```

```
##
## Call:
## lm(formula = log(pm) ~ iso_status + cases + state + date, data = df.merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19784 -0.30787 -0.01236  0.29032  2.85659
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.108e+00  5.323e-01   3.960 7.59e-05 ***
```

```
## iso_statusEmergency      -2.439e-01  6.117e-02  -3.987 6.79e-05 ***
## iso_statusSchoolClosure   9.665e-02  9.081e-02   1.064   0.2872
## iso_statusShelterInPlace -2.085e-01  1.173e-01  -1.777   0.0757 .
## cases                    -8.009e-06  6.123e-06  -1.308   0.1909
## stateIL                   2.549e-02  2.340e-02   1.089   0.2761
## stateMA                  -3.201e-01  2.341e-02 -13.674  < 2e-16 ***
## stateNY                  -3.575e-01  2.344e-02 -15.250  < 2e-16 ***
## stateWA                  -4.340e-01  2.341e-02 -18.538  < 2e-16 ***
## stateWI                  -2.535e-01  2.344e-02 -10.815  < 2e-16 ***
## date                      1.079e-06  2.967e-05   0.036   0.9710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4727 on 4887 degrees of freedom
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.123
## F-statistic:  69.7 on 10 and 4887 DF,  p-value: < 2.2e-16
```

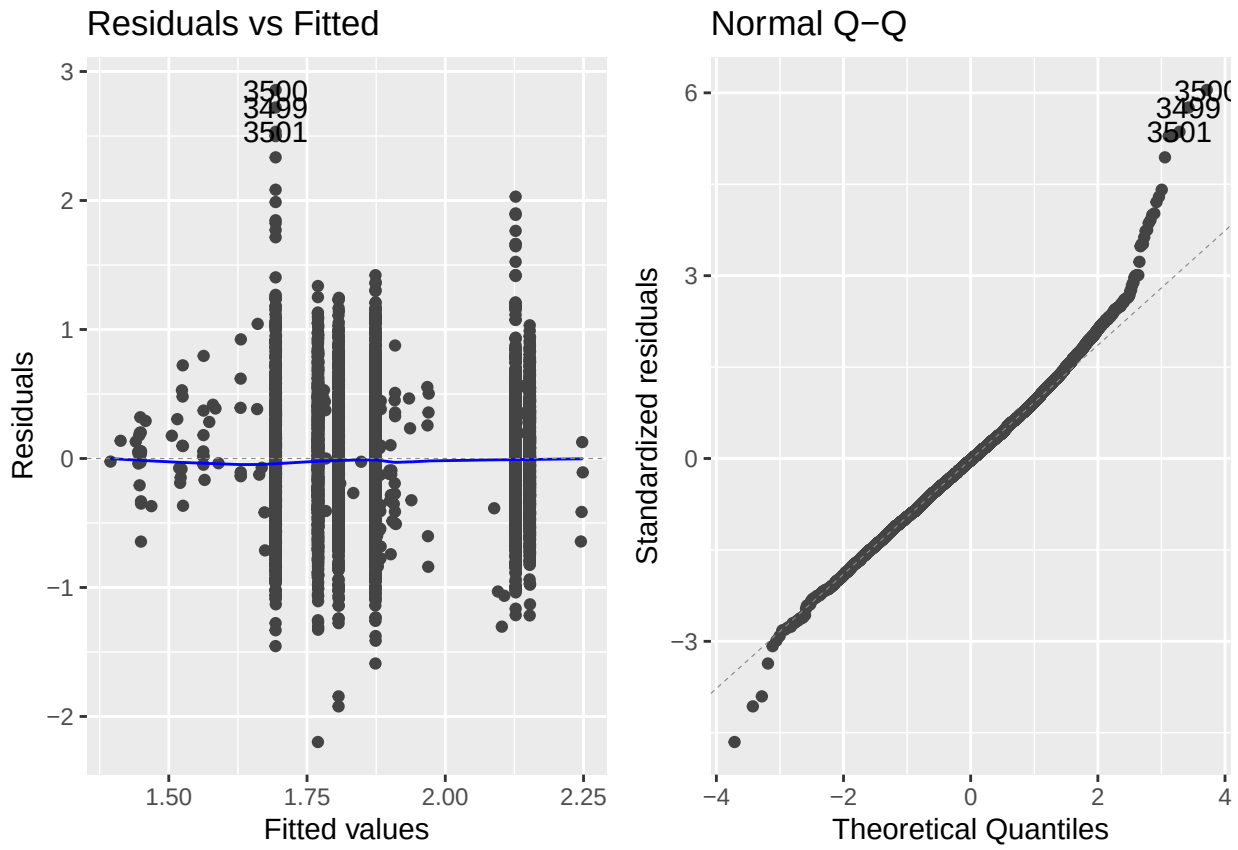
```
# Displaying exponentiated confidence intervals
exp(confint(mlr.pmlog))
```

```
##                              2.5 %     97.5 %
## (Intercept)              2.8992901 23.3715302
## iso_statusEmergency      0.6950191  0.8834116
## iso_statusSchoolClosure  0.9218453  1.3161017
## iso_statusShelterInPlace 0.6449587  1.0217799
## cases                    0.9999800  1.0000040
## stateIL                  0.9798212  1.0739791
## stateMA                  0.6934985  0.7601686
## stateNY                  0.6680249  0.7323326
## stateWA                  0.6188770  0.6783700
## stateWI                  0.7412091  0.8125641
## date                     0.9999429  1.0000593
```

Interpretation and comments: When compared to 'Normal' social distancing status, we are 95% confident that 'Emergency' social distancing status is associated with an average PM 2.5 concentration increase of between 0.7 and 0.9 g/m3, adjusting for incidence of COVID-19 cases, state and date of the year. We observe a very slight decrease in the standard error of the coefficient when compared to the simple linear regression model. The adjusted R-squared shows a poor fit of the model.

Checking residuals to verify model assumptions

```
# Verifying MLR model assumptions
autoplot(mlr.pmlog)[1:2]
```



We observe a slight deviation from the 45 degree line in the QQplot. The normality assumption doesn't seem to have been met. The residual plot also shows us that the residuals do not observe a random pattern around the fitted regression line. This is enough evidence to show that the homoscedasticity assumption has not been met. However, the fit of this model seems to be better than that of the previous model.