An Examination of Pandemic Air Quality and Spread

Akshay Sridharan, Andy Ackerman, Matt Johnson
Department of Statistics and Operations Research
UNC Chapel Hill

November 3, 2020

Abstract

Localized air pollution and global pandemic are two of the most tragically salient realities of the day. The latter challenges the well-being, livelihood, and accepted societal norms of the current generation while the former threatens the very existence of future generations. Unfortunately for some, these trends coincide in what must be the nexus of non-ideal respiratory conditions. Yet it seems intuitively plausible that insights gleaned from one trend are capable of informing proper precautionary measures for the other. That is to say, using Particulate Matter 2.5 as a proxy for "quarantine abidance", it may be possible to forecast spread, lend credence to otherwise unpopular restrictions of personal autonomy, and mitigate rampant contagion. More pointedly, we examine the efficacy of preventative stay at home orders and bar closures as well as the effect of adherence to such preventative measures on Covid-19 cases in Greenville, SC and Raleigh, NC.

1 Introduction

Particulate Air Matter (PM) is a fairly broad category of air pollution encapsulating anything from minuscule yet solid particles to liquid droplets suspended in the atmosphere. As such, PM is derived from a variety of sources ranging from automobile emissions to industry exhaust and even active construction debris. In short, it encompasses much of the modern industrialized economy. Given the ubiquity of particulate matter in everyday life, it would seem that such a measure could provide a fairly comprehensive insight into overall activity during times of state-encouraged, or even mandated, quarantine. In particular, a positive correlation between lagged, per-day Covid-19 cases and pollutant counts is not only plausible but intuitively probable. This intuition is primarily driven by the assumption that communicable disease is predicated on human interaction to facilitate spread, and when excursions outside of the house are curtailed, the corresponding decrease in activity should help to reduce contagion.

This pandemic has laid bare more than just shortcomings in infrastructure and diagnosis; it has also exaggerated disparities in regulation abidance. That is to say, in the wake of sweeping outbreak, numerous precautionary regulations, including mask mandates and compulsory (or pseudo-compulsory) stay at home orders, were put in place across the United States, yet adherence to such policies vary widely. This observation was motivated by authorial personal experience, yet such testimonial motivation seems to be reinforced by concrete political action. More pointedly, historically conservative geographic precincts are conventionally less stringent in their application of regulation, and this was again the case when citizens in the deep south continued attending public school, frequenting bars or eating establishments, and generally operating according to the pre-pandemic status-quo well after national outcry for quarantine. By contrast, progressive leaning municipalities were far quicker and more systematic in their reduction of human interaction. In short, liberal leaning townships err on the side of caution and adherence while conservative learning townships prioritize exercising discretionary autonomy. Thus, it seems that any analysis of pandemic spread would be incomplete without consideration of geography and political leaning more generally.

To that end, this study will compare the historically conservative city of Greenville, SC against the conventionally progressive (at least relatively so) Raleigh, NC. These two cities were chosen quite intentionally so as to juxtapose political leanings while also controlling for other possible confounding variables. That is to say, another potential source of variation in pandemic cases is overall population. Quite intuitively, a larger population is far more likely to demonstrate markedly higher rates of spread. Thus, Greenville (population density 2343/ square mile) and Raleigh (population density 3323/ square mile [1]) are chosen as fairly comparable metropolitan areas. While these are certainly not cities of exactly the same size, there is not an insurmountable discrepancy in population density. Moreover, any remaining influence attributable to distinctions in population will be considered explicitly as a factor of the models below. As a final consideration, it is favorable to analyze cities with comparable weather patterns or climate more generally. Insofar as meteorological variation can result in dissipation of air pollutants, climate could be another confounding variable preventing particulate matter from being a viable proxy to compare across geographic expanses. Quite conveniently, Raleigh North Carolina and Greenville South Carolina are separated by a mere 200 miles. Thus, they experience largely the same weather patterns – certainly more so than Greenville and say, Los-Angeles California.

[1] https://www.opendatanetwork.com/

1.1 Data Description

The meteorological data used in this analysis comes from OpenAQ.org[2]. OpenAQ is an organization that collects and provides data for air quality measurements in 95 countries. Measurements include O3 (Ozone), CO (Carbon Monoxide), NO2 (Nitrogen Dioxide), PM10 (Particulate Matter 10), PM2.5 (Particulate Matter 2.5), SO2 (Sulfur Dioxide), and BC (Black Carbon) in specific locations. By building this open source database of air quality, OpenAQ is encouraging collaboration between scientists across the globe in an attempt to fix air inequalities.

The second of our two data sets is a CDC tabulation of Covid-19 cases per day at a county resolution made available at USAFacts.org[3]. Specifically, with respect to the two cities of interest, we have per day infection counts for Wake County, NC and Greenville County, SC. Also, since the data is timestamped, it is possible to simply join the air quality data with the lagged air pollutant data to assess trends between the two. More pointedly, these two data sets are aimed at providing a holistic conception of fluctuations in air quality and infectious disease in tandem. A graphical depiction of the raw (untransformed) but cleaned data appears in Figure 1. Note, there is a large amount of redundancy along the left boundary of both plots. This represents particulate matter fluctuation preceding the first confirmed case of Covid-19.

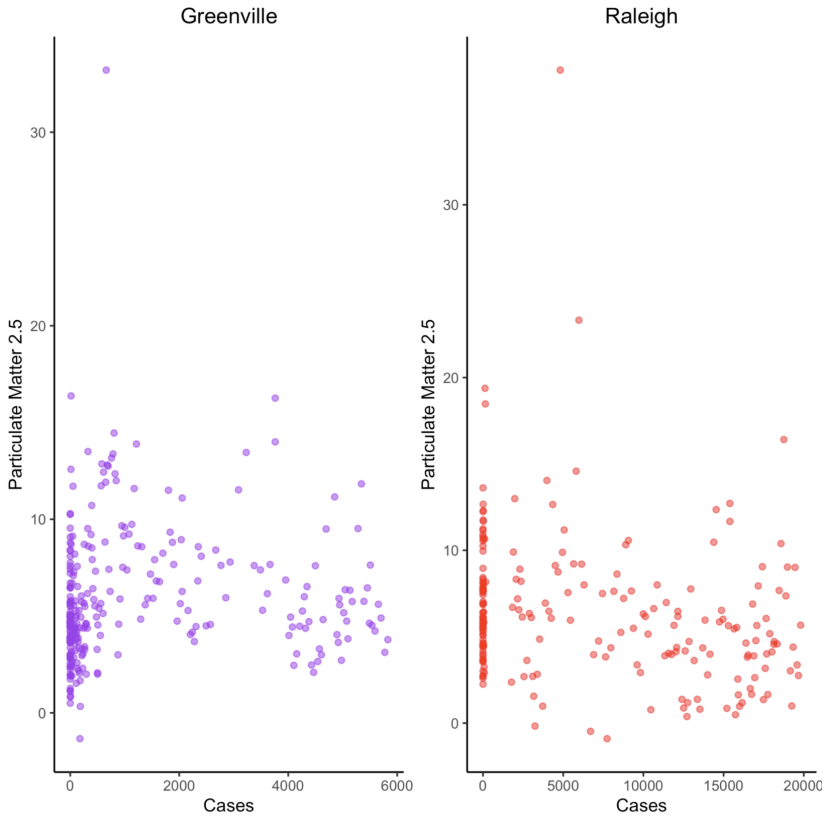


Figure 1: Scatterplot of Raw but Cleaned Data

The initial cleaning and merging of the data was quite manageable. We simply created columns for each type of emission and used averages of the emission values for the day as a single observation. This way, each day has its own row with average emission values. Since the data is all timestamped, we were able to merge the coronavirus cases with "Stay at Home" and "Bar Close" binary variables as new columns. A new column titled "newcasesper" was then appended to the data. This column contains the number of new cases each day divided by the estimated population density of the respective city for normalization. Said variable was then led by six days. The explicit motivation for such a choice will be delayed until the section 2 (Methodology). Finally, we stacked the Greenville and Raleigh data row-wise with an added "Location" variable to denote the relative location. Often during analysis this

[2]https://openaq.org
[3]https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/south-carolina

set was split again to separate the Greenville and Raleigh observations for modeling. The final dataset contained 418 observations with 198 from Raleigh and 220 from Greenville. While this slight asymmetry is non-ideal, such a minimal distinction should not be overtly detrimental to the model.

Note that the Raleigh data begins on January 10, 2020 and is continuous until March 24th. After March 24th there is a short gap where no data was collected; we believe due to the stay at home order, until June 1st. The collection then picks up again and continues until October 12th. Values in this time period have been imputed, however given this relative hole in the data comes at an integral point of the pandemic, it could still pose significant problems for analysis and particularly interpretation. By contrast, the Greenville data begins March 2, 2020 and is continuous until October 12th. Each of the untransformed yet cleaned trends are plotted against time as part of Figure 2.

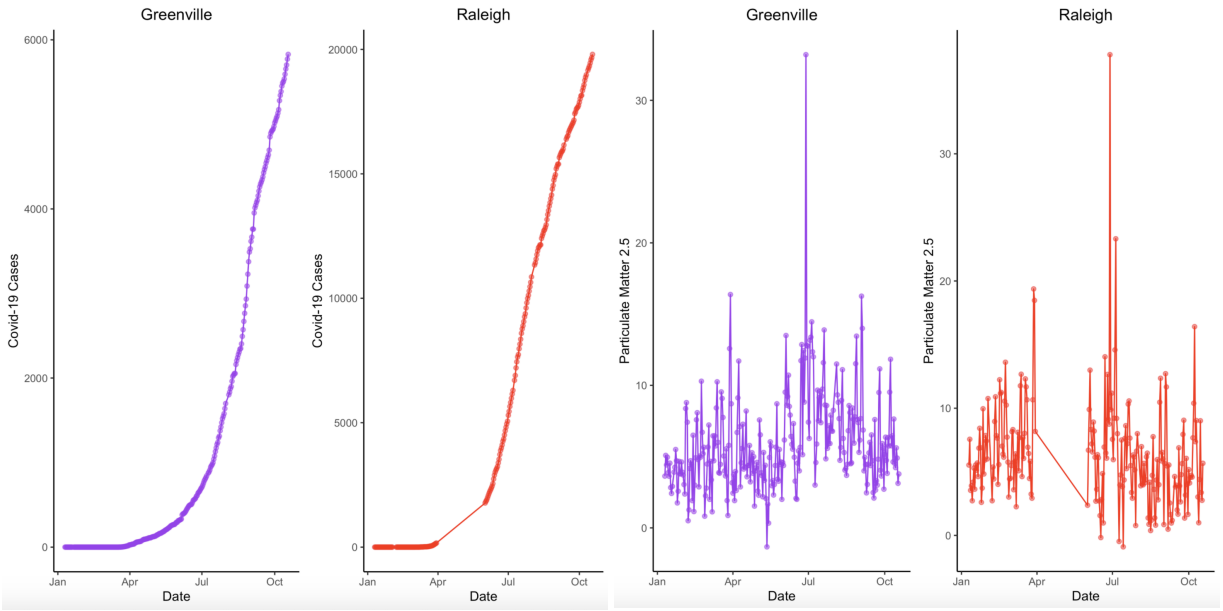


Figure 2: Time Dependent Plot of Air Pollution and Aggregate Covid-19 Cases

2 Methodology

The most pressing preliminary question is one of scale. That is to say, aggregate cases, in both Raleigh and Greenville far outpace any per day trend for particulate air matter in the corresponding city. This is the case to such an extent that failing to transform the response variable could holistically skew the analysis. Therefore, it was necessary to ponder multiple response variable transformations in an attempt to ameliorate any such scale discrepancies. Historically, log and square root transforms are the most common, as they succeed in refining scale without demonstrably distorting the trend itself. Moreoever, these are reasonably straight forward to implement. However, the draw-back is a lessened capacity for interpretation. That is, it seems natural to speak of "total number of cases in a given county", but the log total or square root total, while legitimate, seems much more nebulous.

With this in mind, we put forth an alternative transformation. Specifically, a per day trend derived from the original aggregate trend simultaneously alleviates scale discrepancy concerns (by compressing the cases by an order of magnitude; this is comparable to the shrinkage seen in a more conventional square root transform) and provides an interpretable quantity. In fact, when reporting on pandemic spread, many newscasts default to per day spread. Hence such a metric already seems entrenched in the vernacular surrounding Covid-19. One final caveat completes this transform: scaling by population density. In an effort to ensure that per day counts are as directly comparable as possible, we attempt to control for the influence (intuitively a sizable one) population size/density has in determining viral contagion. Therefore, case counts given in the coming results will be relative to population density and thus be directly comparable to distinct geographic locations without fear of confounding variability (at least as far as this confounding variable is population).

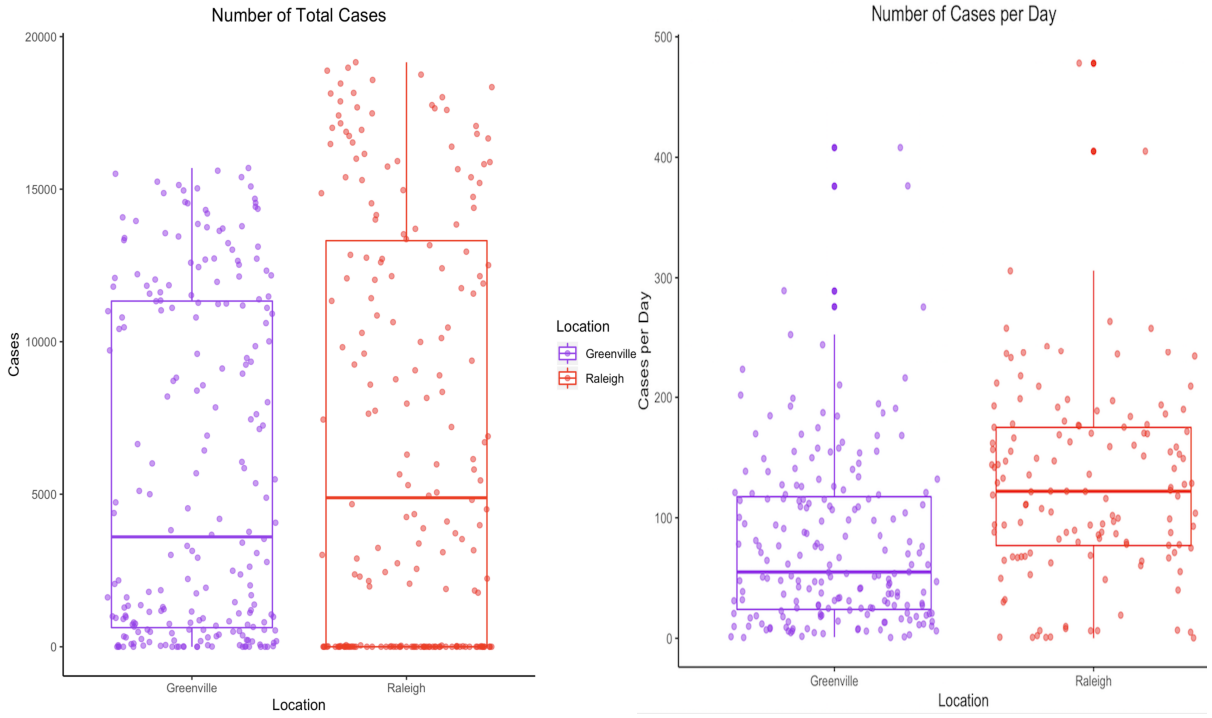


Figure 3: Box Plots of Variable Tranformation

Figure 3 illuminates the role of the response variable transformation. By considering per day, population controlled, Covid-19 cases, as opposed to aggregate Covid-19 cases, the scale has, in effect, been compressed into the hundreds of cases per day rather than the tens of thousands of cases in total. This is much more comparable (albeit still not perfectly analogous) to the scale of particulate matter which is given in 10's of micrograms per cubic centimeter. A second boon of the variable transform is mitigating skewness or distortion due redundancy at zero. It would seem that there are far fewer days lacking in new confirmed cases than there were days preceding the first confirmed case. Thus, using the per day count rather than the aggregate count has a centralizing effect in that it seems to tug the median towards the spatial center of the distribution. All this aside, the most notable deleterious effect of the variable transformation seems to be the introduction of some outliers. The are several points ranging outside of the whiskers of the box plot with two points along the upper boundary of the Raleigh plot doing so in dramatic fashion. Therefore, it will be imperative to consider influential points in conjunction with the model itself. Specifically, high leverage outlying points could be significantly distorting results and thus viable candidates for omission. For now, explicit discussion of this will be reserved for section 3.3 (Model Validity).

Scale, and by extension variable transformations now being considered, the next consideration is model derivation and selection. To that end, note that both the data on Covid-19 cases and air pollution are given with respect to time. More specifically, the data sets are constructed so as to give readings on consecutive days. Therefore, it is likely that any tenable model will include time-series techniques. To underscore such an intuition, consider the autocorrelation plots in Figure 4. Such a graphic elucidates the significant autocorrelation within the first two lags of all variables and within the first 10 lags of all but Raleigh Particulate Matter.

Hence, in this segment of the analysis we hope to predict the future development of new cases by the presence of air contaminants in each location. Simultaneously city-specific trends will be controlled for by a "fixed-effects" term. Specifically, the general time series model to be implemented is

$$y_t = m_t + \epsilon_t \tag{1}$$

where m_t is the general linear regression component and ϵ_t is the first order autoregressive component given by

$$\epsilon_t = \phi_1 \epsilon_{t-1} + z_t. \tag{2}$$

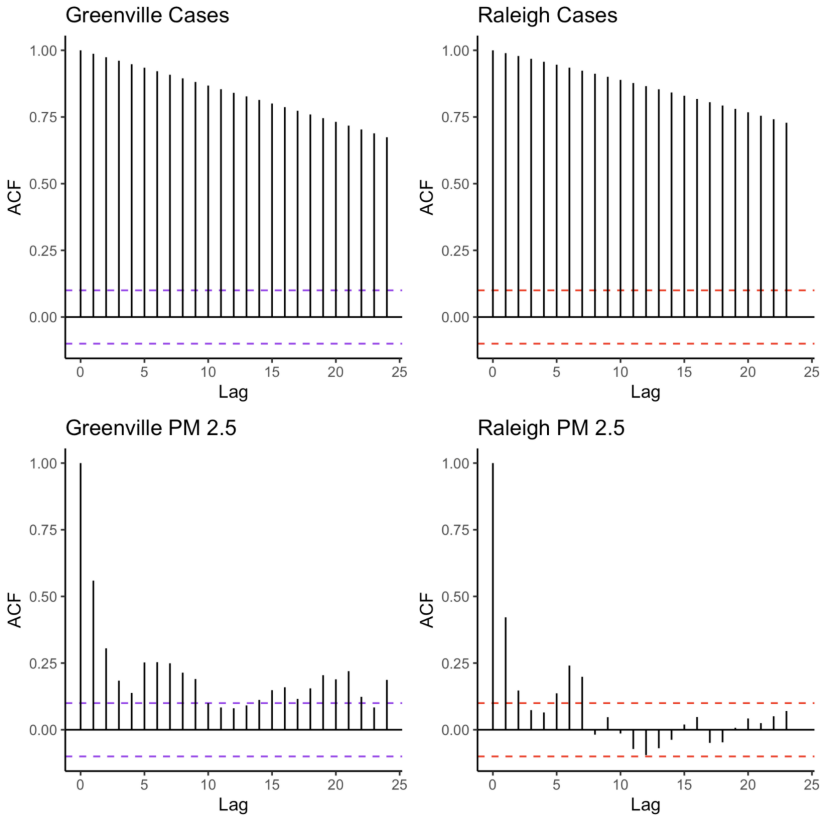


Figure 4: Geographic Specific Autocorrelation Plots

Furthermore, the the covariates within the time-series will be lagged by six days. The motivation for such a lag time is empirical and epidemiological in nature, but admittedly subject to fluctuation over the course of the pandemic. More specifically, the approximate time between infection and case reporting is six days, implying that transmission may be estimated to have occurred nearly a week prior to the reporting date. Again such a report time has fluctuated over the course of the pandemic, but this seems a reasonable and fairly comprehensive lag for our purposes.

2.1 Model Selection

It is important to note several constraints that informed our variable selection procedure. The first and most poignant of which is a relatively small number of explanatory variables. Quite plainly, in the original data set there were merely two (common) explanatory variables – PM 2.5 and O3. That said, in later models it was possible to construct binary categorical variables from public records on stay at home orders and bar closure mandates. However, even with these additions included, the total number of regressors was still well below ten. As a result, contemporary and highly touted models for shrinkage and selection, such as Ridge and particularly Lasso, are not preferable. While these methods are uniquely adept at refining models in high dimension, they do not offer a demonstrable benefit, and indeed are arguably sub-optimal, for comparatively low dimensional data. Hence, these models are considered and even implemented, but not included in the report itself. [4]

At a high level, two basic classes of models – fixed-effect non-zero intercept and fixed-effect zero intercept panel regression – were considered and subsequently refined through model selection. The

[4] This report is meant to be a concise summary of our most resounding results and accompanying procedures. As such, it is certainly not a comprehensive documentation of our much more exhaustive efforts. That said, we have compiled the totality of our work into a github page for public record and general reference. Here, initial scratch musing, various graphical depictions, and more tangential models can be found. For example, much consideration was given to a non-zero intercept general linear model, a zero-inflated poisson regression, and even GLMnet procedures. Such work can be found at https://github.com/mattymo18/STOR-664-Project.

motivation for considering the effective elimination of an intercept term was primarily a pragmatic predictive consideration, but it also has a reasonable interpretation. That is to say, the zero-intercept model increased predictive power, as indicated by a spike in adjusted R^2 of nearly 40%, yet there is also a clean, albeit theoretical, interpretation of said model. It is quite impractical to assert that air pollution would ever truly reach negligible levels, and they certainly were not negligible at the onset of the pandemic. Thus, it seems rather counterintuitive predict daily Covid-19 cases given zero particulate air matter. In all actuality, it is evident that air pollution was present prior to the onset of the pandemic. Yet if such a vacuum of air pollution is ever actually achieved, and importantly air pollution is a reasonably proxy for activity level, we would expect spread to also be nearly annihilated. Thus such a theoretical genesis of emissions will also be treated as a zero point of case count, and the model will have a zero intercept. Thus, the full model corresponding to 1 is

$$m_t = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 t_5. \tag{3}$$

Here t_1 corresponds to the numerical ozone count at time t; t_2 corresponds to the numerical particulate matter 2.5 count at time t; t_3 corresponds to the binary location specifier (where 1 indicates Raleigh) ; t_4 corresponds to the binary stay at home mandate at time t (where 1 indicates a mandate in place); and t_5 corresponds to the binary bar closure mandate at time t (following the same convention as the stay at home mandate variable).

With regard to model selection of this now specified full model, consideration will be given to AIC, adjusted R^2 and Mallow's Cp in determining the optimal number of parameters.

3 Results

The full model boasts an in-sample adjusted R^2 of 0.6582 and statistically significant estimates for all included covariates. Taking a look at the signage of the estimates, both ozone and particulate matter yield positive effects on cases per day. This coincides with the hypothesis inasmuch as increases in air pollution are indicative (significantly so) of high per day case counts. Put plainly, when activity level remains high, confirmed cases rise in accordance with the corresponding elevated exposure. Secondly, consider the signage of the location binary variable. Recall that a 1 indicates location Raleigh. The interpretation here being, Raleigh demonstrates a markedly lower number of population controlled confirmed cases per day. Again, this is consistent with the original intuition that more laissez-fare regulation would allow for more rapid spread. Finally, note the signage of the two binary mandate-related variables. The stay at home order appears to have a demonstrable effect on reducing spread, whereas bar closure would, at first glance, appear to do quite the opposite.

This is quite a counterintuitive result, as one would hardly expect that the closure of a public environment conducive to close proximity interaction would actually increase, rather than decrease, spread. However, upon reflection, the stay at home binary and the bar closure are capturing much of the same variability; moreoever, their respective time-frames exhibit substantial overlap. More specifically the stay at home order is actually a subset of the time that bar closures were in effect. That is to say, when a stay at home order goes into place, it seems a mere afterthought that someone would be concerned with bar closures; people aren't supposed to be leaving the house, much less frequenting bars. As one final piece of evidence that even such significant results may not be so directly indicative of bars closures inducing Coronavirus outbreaks, note that the signage of the sum of the two effects is net negative. Therefore, collectively, even with the results at hand, it would appear that mandates are reducing overall contagion. Thus, it would seem that the positive sign is more an artifact of the construction of the binary variables than of a disconcerting trend. Figure 5 illuminates such results.

3.1 Refinement

Now the full model will be refined via AIC. Recall that stepwise AIC reduction will iterate through various versions of reduced models and suggest the one that minimizes $AIC = nlog\frac{SSE_p}{n} + 2p$. Figure 6 demonstrates that removal of any variable from the full model will increase AIC. Thus, it would seem that the reduced model via AIC is in accordance with the full model.

	Estimate	Std. Error	t value	Pr(>|t|)
O3	0.002885	0.0005081	5.677	3.16e-08
PM25	8.521e-06	1.963e-06	4.34	1.936e-05
loc	-7.59e-05	1.974e-05	-3.844	0.000147
Stay_At_Home	-0.0003094	3.611e-05	-8.569	5.127e-16
Bar_Close	0.0002045	2.225e-05	9.189	6.067e-18

Table 6: Fitting linear model: newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close - 1

Observations	Residual Std. Error	R^2	Adjusted R^2
313	0.0001308	0.6636	0.6582

Figure 5: Zero-Intercept Panel Regression Output

```
Start:  AIC=-5592.68
newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close + loc - 1

               Df  Sum of Sq        RSS     AIC
<none>                       5.2687e-06 -5592.7
- loc           1 2.5278e-07 5.5215e-06 -5580.0
- PM25          1 3.2219e-07 5.5909e-06 -5576.1
- O3            1 5.5140e-07 5.8201e-06 -5563.5
- Stay_At_Home  1 1.2560e-06 6.5247e-06 -5527.8
- Bar_Close     1 1.4443e-06 6.7131e-06 -5518.8
```

Figure 6: Stepwise AIC

As a second metric to judge model reduction, consider optimization of adjusted R^2. It is imperative to note the distinction between multiple R^2 and adjusted R^2. Insofar as the latter penalizes for additional covariates, optimizing this metric becomes a trade-off between model complexity and fit. In essence this quantity represents the amount of variation in the response variable explained by a given model, adjusted for number of covariates. Thus, the model with the largest adjusted R^2 likely wields the most predictive power. Figure 7 depicts such an optimization procedure.

Finally, consider Mallow's Cp. In so doing, one aims to minimize $\gamma_p = \frac{1}{\sigma^2}(\sum_{i=1}^{n} b_i^2 + \sum_{i=1}^{n} var(\hat{Y}_i))$. That said, strict minimization is not the only criteria for such a procedure; fit must be considered as well. Thus, when evaluating the Cp plot, the model which minimizes the Cp statistic under the constraint that it be close to the line depicting a "good fit" is optimal. Figure 8 provides the results. Notice all three methods (AIC, adjusted R^2 and Mallow's Cp) agree that the five parameter model is to be preferred.

3.2 Prediction Fit

Analysis of model fit will be done via the root mean squared error. Conceptually, this is a measure of residual standard deviation, or more plainly, the spread of the prediction errors. Hence, as with most errors, smaller errors are indicative of better performance. In this case, a small error would correspond to concentration of the residuals along the line of best fit. Secondly, note that as is often the case with observational data (as opposed to a designed experiment), the model can only be cross validated by splitting up the sample into a training and test set respectively. The training set will be composed of randomly selected observations of 75% of the sample size, and the test data set, by default, will be what remains of this subset of the total sample. Figure 9 demonstrates remarkable predictive fit and very minimal discrepancy in fit between training and test sets.

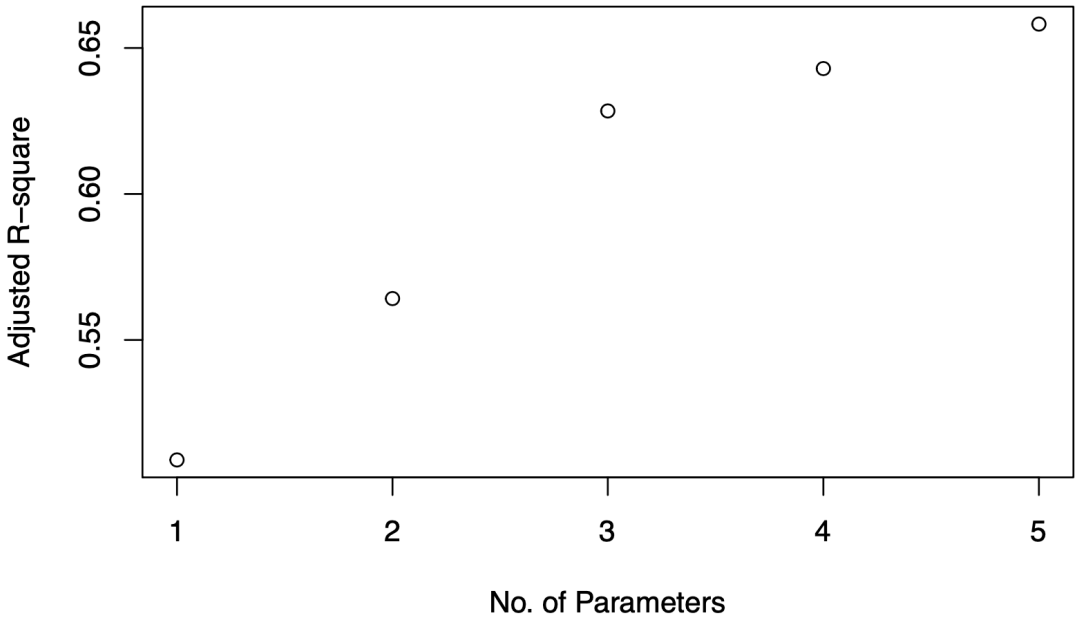


Figure 7: Optimization via Adjusted R^2

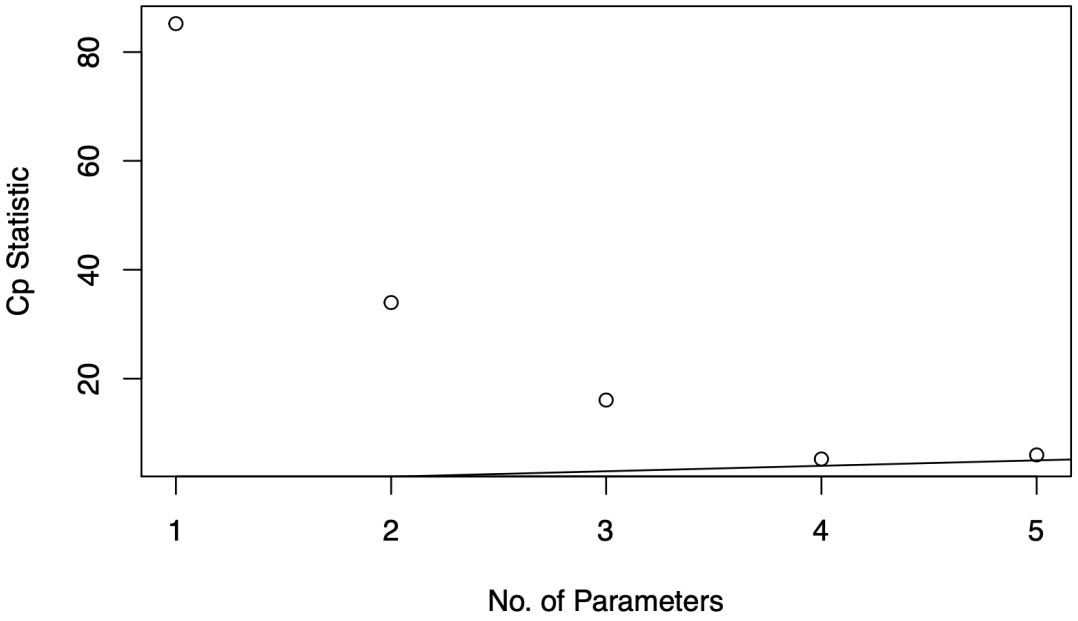


Figure 8: Optimization via Mallow's Cp

Training RMSE	Test RMSE
0.0001297	0.0001664

Figure 9: Analysis of Fit via RSME

3.3 Model Validity

This section is dedicated to assessing the validity of model assumptions – such as normality and the effect of high leverage outliers – that undergird appropriate usage of general linear regression. We will take each in turn. Recall that one of the pillars upon which general linear regression is predicated is an assumption that residuals are normally distributed. Intuitively there is some concern in modeling a trend (cases per day) that is very likely growing at an exponential rate, or at the very least more than a linear trend, with a linear model. However, when one examines the normal Q-Q plot given in Figure 10, it becomes evident that the residuals are relatively well behaved outside of the tails (particularly the right tail). This would seem to indicate that there is no fundamental flaw in the application of our model, only that it may be prudent to temper the extent to which one generalizes results or forecasts well into the future. That said, in the short to medium term, our model can, and indeed does, perform well in both fit and forecast.

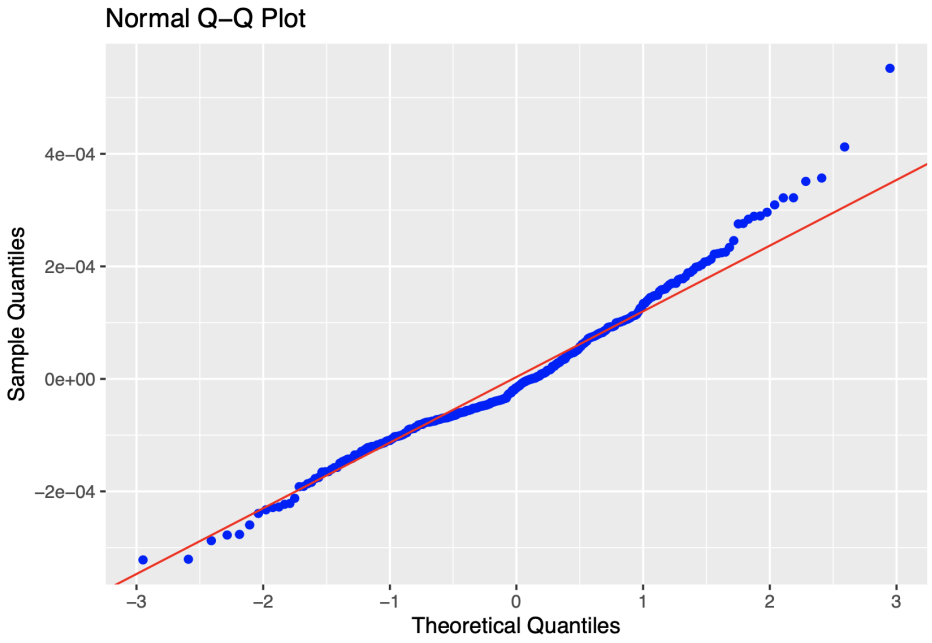


Figure 10: Check of Normality Assumption

Recall there are a number of outliers in our data set. It is now time to assess their leverage on the model itself. We will juxtapose the influence of outliers on two non-zero intercept models as compared to the chosen zero intercept model. This will highlight another relative strength of the model, namely that it is robust to outliers. Through a review of the Cook's distance bar plots for the two non-zero intercept models, it is immediately apparent that there are several values that exceed the threshold calculated by the formula $4/(n-k-1)$. We may thereby assess the impact of these 18 observations through a sensitivity analysis on the beta coefficients and model fit. Note that there is significant overlap between the omission sets of each of these models, with 15 observations shared between the two. After isolating these respective outliers and retraining the two non-zero intercept models on the remaining training data, the most notable changes in model fit occur in the beta coefficient corresponding to the "PM25" variable. In both models, the omission of their respective outlier sets lead to a decrease in the beta coefficient corresponding to "PM25" by an approximate factor of 2. Furthermore, the omission leads to the outright insignificance of the "PM25" variable in both models, despite the pre-omission models showing the variable to be highly significant. This is further coupled with a large decrease in the adjusted R^2 values of the models. Thus, from this information, in conjunction with the very minor changes observed in the remaining beta coefficients, we see that the non-zero intercept models are highly sensitive to outlying observations, particularly when calculating the coefficient for the "PM25" variable. It is important, however, to note that the changes observed in this beta coefficient were only

of magnitude, with the non-negativity of the estimated "PM25" coefficient remaining across all models. In any case, the sensitivity analysis shows the impact of these observations and brings into question the viability of such models in both an explanatory and predictive capacity. Coefficient output for both non-zero intercept models and the zero-intercept model (depicted high to low) can be seen in Figure 11.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001005	1.589e-05	6.325	9.469e-10
PM25	8.751e-06	1.954e-06	4.478	1.081e-05
factor(Location)Raleigh	-0.0001418	1.884e-05	-7.528	6.454e-13
Stay_At_Home	-0.0003706	3.297e-05	-11.24	1.339e-24
Bar_Close	0.0002666	2.085e-05	12.79	4.998e-30

Table 8: Fitting linear model: newcasesper ~ PM25 + factor(Location) + Stay_At_Home + Bar_Close

Observations	Residual Std. Error	R^2	Adjusted R^2
297	0.0001124	0.4063	0.3982

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.243e-05	2.884e-05	2.512	0.01256
O3	0.001048	0.0008499	1.234	0.2184
PM25	8.187e-06	1.943e-06	4.214	3.352e-05
factor(Location)Raleigh	-0.0001385	1.884e-05	-7.349	2.042e-12
Stay_At_Home	-0.000373	3.305e-05	-11.29	9.829e-25
Bar_Close	0.0002594	2.084e-05	12.45	8.67e-29

Table 6: Fitting linear model: newcasesper ~ O3 + PM25 + factor(Location) + Stay_At_Home + Bar_Close

Observations	Residual Std. Error	R^2	Adjusted R^2
296	0.0001113	0.4067	0.3964

	Estimate	Std. Error	t value	Pr(>|t|)
O3	0.002767	0.0004528	6.112	3.139e-09
PM25	1.031e-05	1.766e-06	5.836	1.417e-08
loc	-0.0001277	1.827e-05	-6.992	1.849e-11
Stay_At_Home	-0.0003776	3.316e-05	-11.39	4.165e-25
Bar_Close	0.0002592	2.077e-05	12.48	6.176e-29

Table 14: Fitting linear model: newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close - 1

Observations	Residual Std. Error	R^2	Adjusted R^2
297	0.0001118	0.7331	0.7286

Figure 11: Coefficient Tables for Reference in Leverage Discussion

By contrast, consider the influence of outliers in the zero-intercept model. Once again, when reviewing the Cook's distance bar plot for the zero-intercept model (Figure 12), it is apparent that there exists a subset that exceeds the typically cited threshold for the statistic. We proceed by assessing the impacts of the omission of these 14 observations on the zero-intercept model through a sensitivity analysis. Note that of these 14 observations within the omission set, only 1 observation lies in common with the preceding two omission sets. After isolating the observations and retraining the models, one can observe a stark distinction relative to the same process conducted upon the non-zero intercept models. Namely, following the omission of these observations, consistently minor changes across the regression beta coefficients are noted, with all values maintaining a high level of significance following the omission. In fact, we see that the beta coefficient associated with the location binary "loc" increases in magnitude and becomes more statistically significant as well. Furthermore, note a non-trivial increase

in the adjusted R^2 value as well. This finding indicates the robustness of the model coefficients and deems it a strong candidate for exploratory analysis.

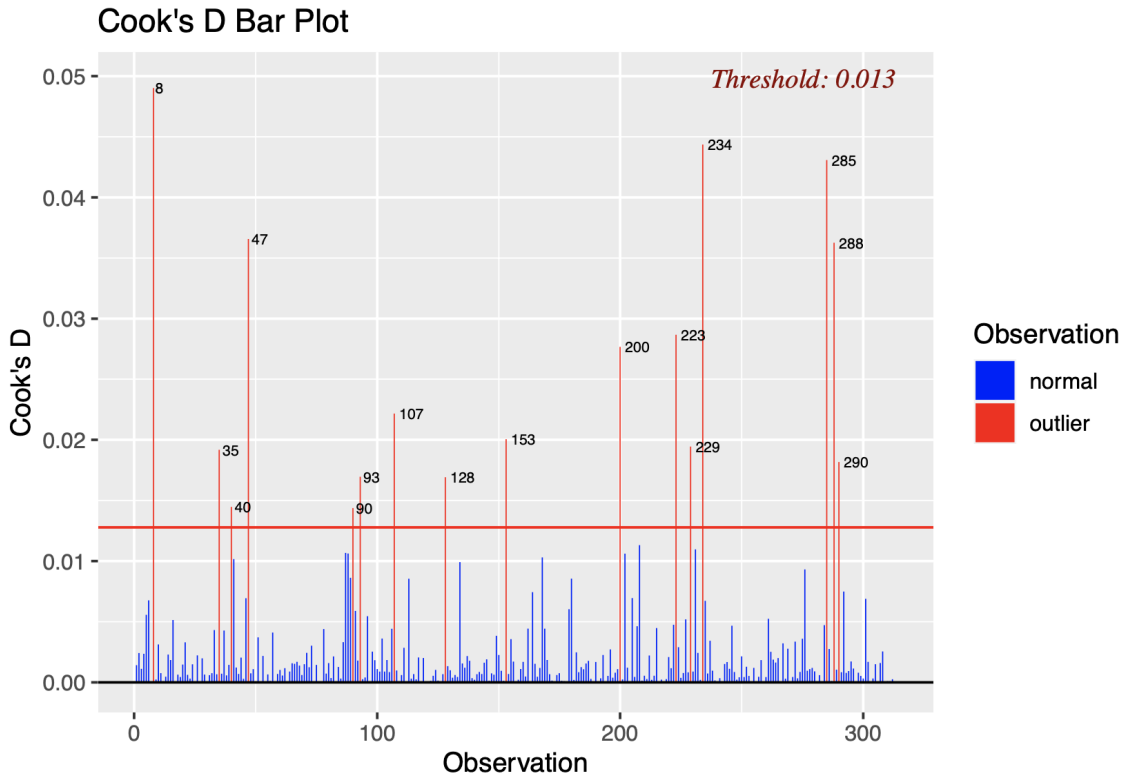


Figure 12: Analysis of Leverage

4 Conclusion

In short, the insights of this analysis, while limited, should not be discounted. Specifically, a non-trivial amount of variation in per day, population controlled Covid-19 cases is explained through a model predominantly predicated on air pollution and mandate abidance. Furthermore, coefficients pertaining to ozone and particulate matter 2.5 are significantly positive. Thus, insofar as measures of air pollution are effective proxies for general human activity (given the composition of particulate matter this does not seem like a farfetched assumption), stubbornly adhering to a pre-pandemic routine is imprudent.

Such findings may seem intuitive, yet that does not diminish their relevance or gravity. That said, the true novelty of this analysis lies in its political commentary. A highly statistically significant negative coefficient on the binary location variable indicates that Greenville experiences a much larger per day case count than Raleigh. Quite plainly, Greenville was slower to enact a stay at home order, more reticent to close bars, and, at least anecdotally, far less stringent in enforcing what mask mandates were in place. The evidence seems clear: caution is warranted and heightened autonomy comes at a cost.

In parting, and along this vein, note that stay at home orders had a significant dampening effect on contagion. While strict quarantine may be far less than ideal (and conversation may still yet to be had regarding the deleterious effects on mental health of prolonged isolation) it is evident that refraining from social interaction helps combat contagion. Here again, non-ideal and, at times, unpopular restrictions of autonomy may well be advisable and even commendable.

Merits of this analysis aside, there is undoubtedly room for improvement. Most notably, the scope of analysis could be expanded and the data more consistent. By that I mean the current model considers data on two diametrically opposed political municipalities, one of which contains a significant omission in data. Generalizing these intuitions and results to a more expansive list of cities could help provide

national or even global insight into pandemic protocol. Further, scraping data that provides an uninterrupted synopsis of the quarantine time period would assuage difficulties in interpretation. Finally, including more regressors, potentially even more contrived variables such as the binaries included in the regression above, could allow for more nuanced model selection.

1. Guan, Wei-jie, et al. "Clinical Characteristics of Coronavirus Disease 2019 in China: NEJM." New England Journal of Medicine, 7 May 2020, www.nejm.org/doi/full/10.1056/NEJMoa2002032.

2. Smith, Richard, and K.D.S Young. Linear Regression.

3. Faraway, Julian R. Linear Models with R. 2nd ed., CRC Press.

Pandemic Air Quality Study R Code

11/18/20

```
library(tidyverse)
library(lubridate)
library(gplots)
library(glmnet)
library(pscl)
library(dplyr)
library(knitr)
library(kableExtra)
library(broom)
library(leaps)
library(olsrr)
library(memisc)
library(pander)
library(car)
```

```
DF <- read_csv("derived_data/DF.Final.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

Our final dataframe utilized within the analysis contains 418 total observations, with 198 being data on Raleigh, NC, and 220 being data on Greenville, SC. For Raleigh, NC, we have consecutive observations for the dates from January 10th, 2020 to March 24th, 2020 and additionally from June 1st, 2020 to October 12th, 2020. For Greenville, SC, we have consecutive observations from March 2nd, 2020 to October 12th, 2020.

Initializing function for use in evaluating RMSE

```
rmse=function(x,y){sqrt(mean((x-y)^2))}
```

Initial Panel regression model with non-zero intercept

```
fixed_effect1 <- lm(newcasesper ~ PM25 + O3 + factor(Location) + Stay_At_Home + Bar_Close, DF)
summary(fixed_effect1)
```

```
##
## Call:
## lm(formula = newcasesper ~ PM25 + O3 + factor(Location) + Stay_At_Home +
##     Bar_Close, data = DF)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.330e-04 -7.935e-05 -3.202e-05  7.603e-05  7.394e-04
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8.780e-05  2.998e-05   2.929  0.00359 **
## PM25                       5.543e-06  1.779e-06   3.116  0.00196 **
## O3                         7.774e-04  8.967e-04   0.867  0.38650
## factor(Location)Raleigh -8.510e-05  1.850e-05  -4.600 5.64e-06 ***
## Stay_At_Home            -3.014e-04  3.374e-05  -8.932  < 2e-16 ***
## Bar_Close                  2.051e-04  1.956e-05  10.486  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.0001393 on 412 degrees of freedom
## Multiple R-squared:  0.2661, Adjusted R-squared:  0.2572
## F-statistic: 29.88 on 5 and 412 DF,  p-value: < 2.2e-16
```

Selection methods applied to non-zero intercept model

```
#Create numeric binary variable for location with Raleigh = 1 and Greenville = 0.
DF$loc <- ifelse(DF$Location == "Raleigh", 1, 0)

#Initialize train and test sets. These will be randomized with 75% of observations composing the traini

#75% of the sample size
smp_size <- floor(0.75 * nrow(DF))

#Setting a seed for the sake of reproducibility.
set.seed(123)
train_ind <- sample(seq_len(nrow(DF)), size = smp_size)

#Form test and train sets
train <- DF[train_ind,]
test <- DF[-train_ind,]
```

```
reglm1 <- regsubsets(newcasesper ~ PM25 + O3 + factor(Location) + Stay_At_Home + Bar_Close, train)
lm1s <- summary(reglm1)
lm1 <- lm(newcasesper ~ PM25 + O3 + factor(Location) + Stay_At_Home + Bar_Close, train)

#AIC
fixed_effect1.1 <- step(lm1)
```

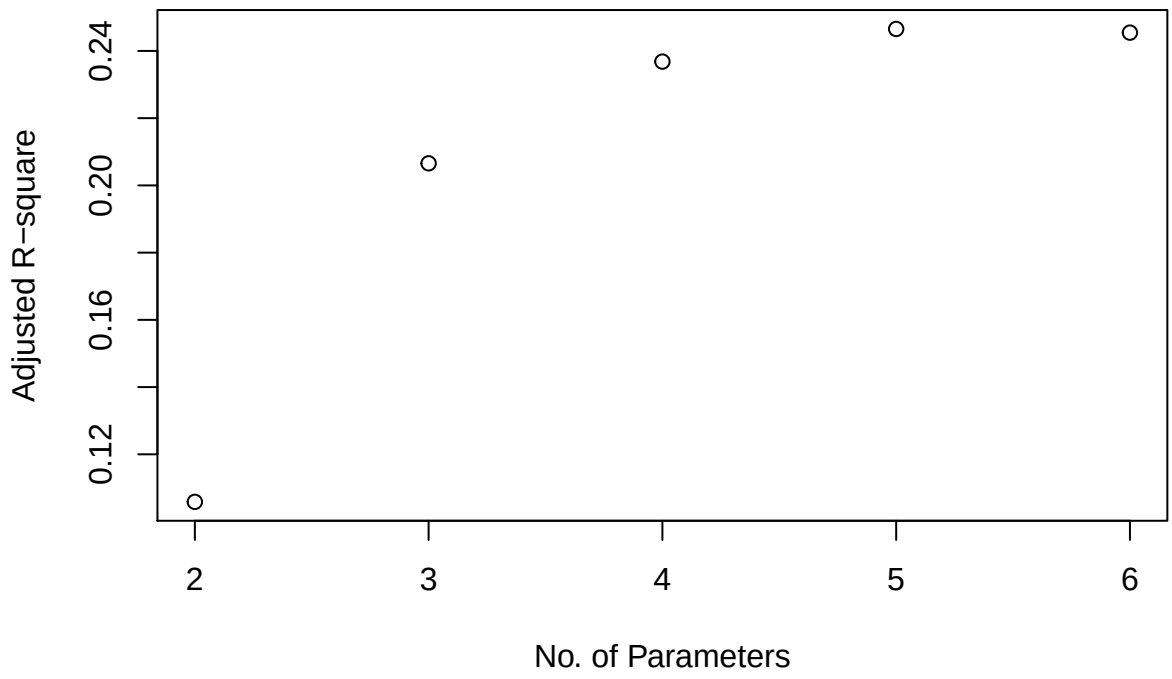
```
## Start:  AIC=-5525.05
## newcasesper ~ PM25 + O3 + factor(Location) + Stay_At_Home + Bar_Close
##
##                     Df  Sum of Sq        RSS     AIC
## - O3                 1 1.1760e-08 6.5096e-06 -5526.5
## <none>                             6.4978e-06 -5525.0
## - PM25               1 9.6250e-08 6.5941e-06 -5522.4
## - factor(Location)   1 2.7506e-07 6.7729e-06 -5514.1
## - Stay_At_Home       1 1.1443e-06 7.6421e-06 -5476.3
## - Bar_Close          1 1.6281e-06 8.1259e-06 -5457.1
##
## Step:  AIC=-5526.48
## newcasesper ~ PM25 + factor(Location) + Stay_At_Home + Bar_Close
##
##                     Df  Sum of Sq        RSS     AIC
## <none>                             6.5096e-06 -5526.5
## - PM25               1 1.0528e-07 6.6149e-06 -5523.5
## - factor(Location)   1 3.0038e-07 6.8100e-06 -5514.4
## - Stay_At_Home       1 1.1365e-06 7.6461e-06 -5478.1
## - Bar_Close          1 1.7052e-06 8.2148e-06 -5455.7
```

```
#Adjusted R^2
lm1s$which[which.max(lm1s$adjr2),]
```

```
##             (Intercept)                    PM25                      O3
##                    TRUE                    TRUE                   FALSE
```

```
## factor(Location)Raleigh            Stay_At_Home               Bar_Close
##                    TRUE                    TRUE                    TRUE
```

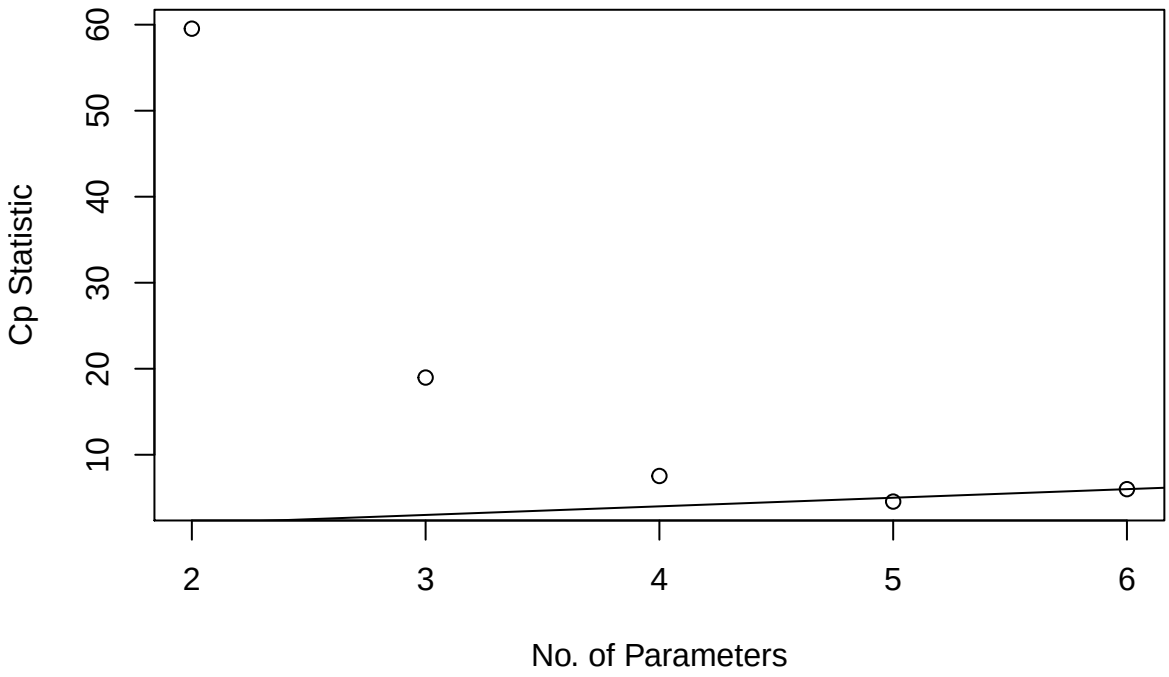
```
plot(2:6,lm1s$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
```




```
#Mallow's Cp
lm1s$which[which.min(lm1s$cp),]
```

```
##             (Intercept)                    PM25                      O3
##                    TRUE                    TRUE                   FALSE
## factor(Location)Raleigh            Stay_At_Home               Bar_Close
##                    TRUE                    TRUE                    TRUE
```

```
plot(2:6,lm1s$cp,xlab="No. of Parameters",ylab="Cp Statistic")
abline(0,1)
```




```
#From the above variable selection methods we see that the model that optimizes
#the Adjusted R^2 metric is the one predicting `newcasesper` by an intercept
#term and the variables `PM25`, `O3`, `factor(Location)Raleigh`, `Stay_At_Home`,
#and `Bar_Close`. The model that optimizes the AIC and Mallow's Cp
#statistic is the one predicting `newcasesper` by an intercept term and
#the variables `PM25`, `factor(Location)Raleigh`, `Stay_At_Home`, and
#`Bar_Close`. However, the statistic for this model is slightly above
#the Cp = p line when considering the Mallow's Cp metric. We will
#thereby proceed by considering both of these models.
```

```
fixed_effect_intercept1 <- lm(newcasesper ~ O3 + PM25 + factor(Location) + Stay_At_Home + Bar_Close, tra
fixed_effect_intercept2 <- lm(newcasesper ~ PM25 + factor(Location) + Stay_At_Home + Bar_Close, train)

#Calculating the prediction RMSE on the test set
rmse(fitted(fixed_effect_intercept1), train$newcasesper)
```

```
## [1] 0.0001440828
```

```
rmse(predict(fixed_effect_intercept1,test),test$newcasesper)
```

```
## [1] 0.0001197185
```

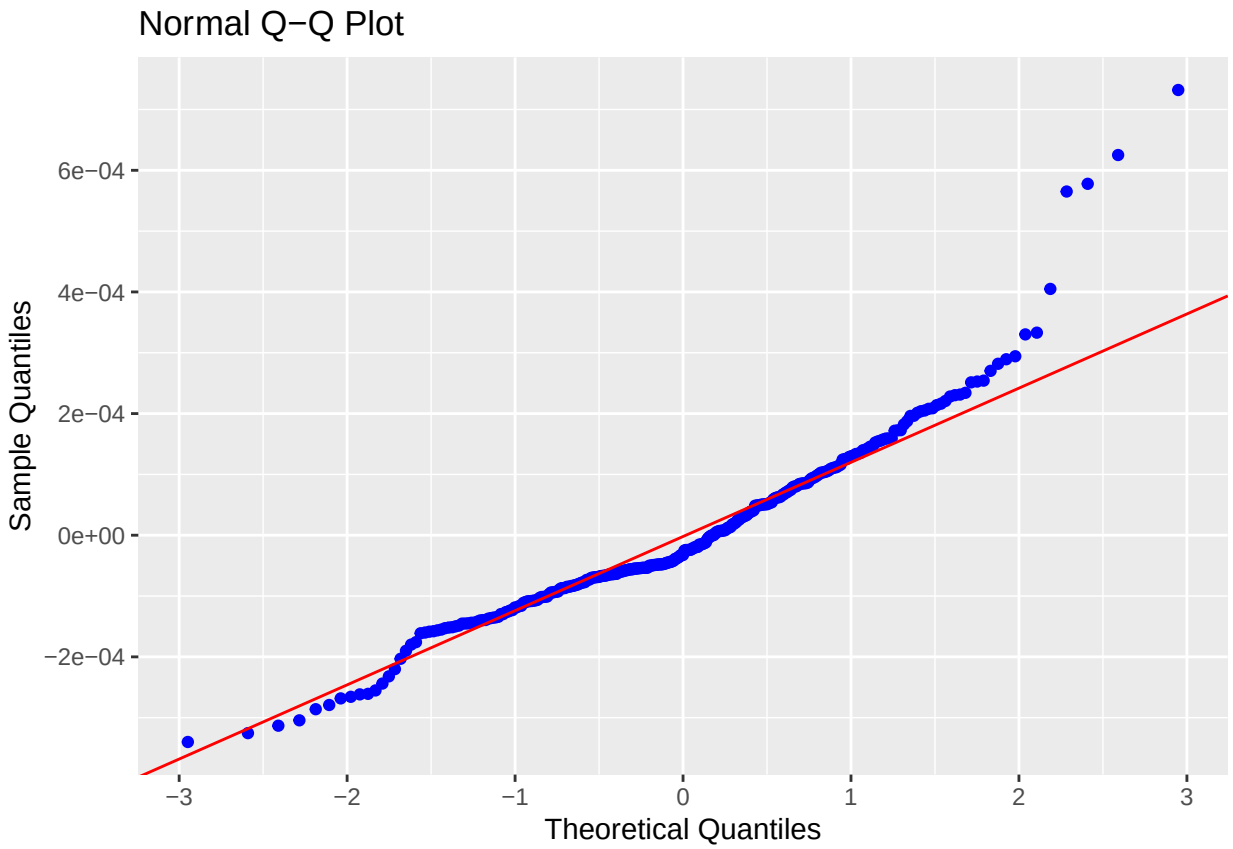
```
rmse(fitted(fixed_effect_intercept2), train$newcasesper)
```

```
## [1] 0.000144213
```

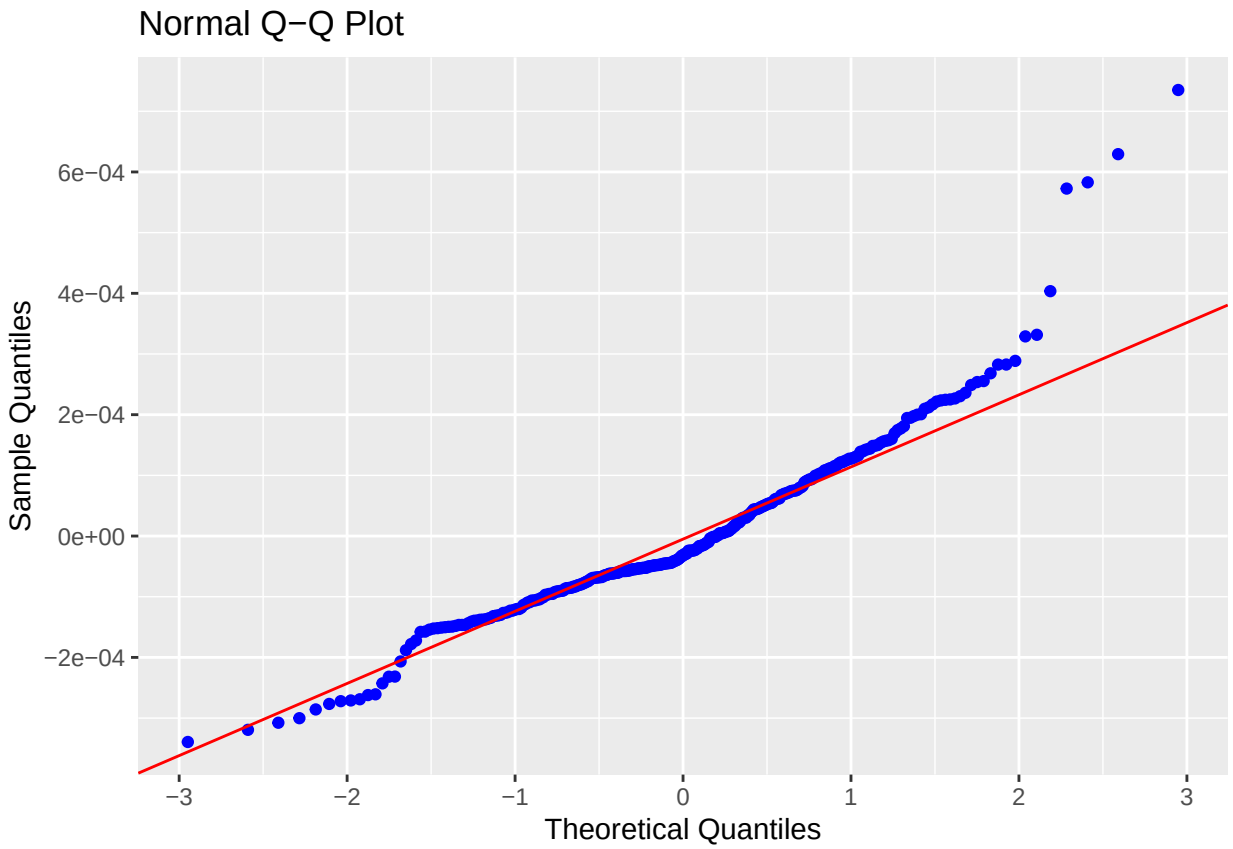
```
rmse(predict(fixed_effect_intercept2,test),test$newcasesper)
```

```
## [1] 0.0001198446
```

```
#Considering the residual Q-Q plot to visualize any violations of the normality assumption.
ols_plot_resid_qq(fixed_effect_intercept1)
```

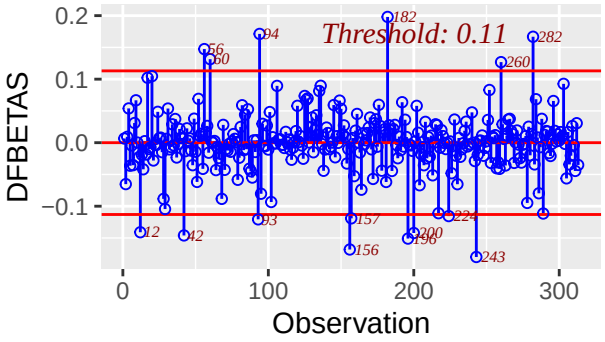



```
ols_plot_resid_qq(fixed_effect_intercept2)
```

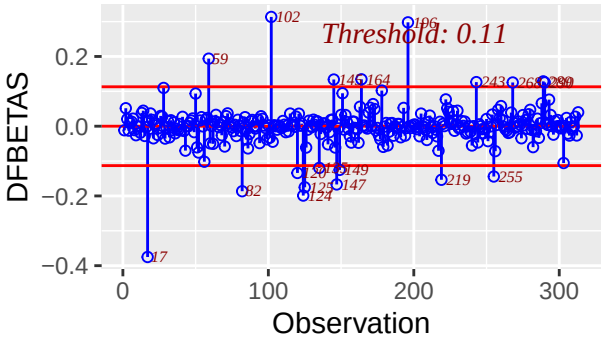



```
#Examining DFBETA plots to identify points particularly influential in estimating each parameter.
ols_plot_dfbetas(fixed_effect_intercept1)
```

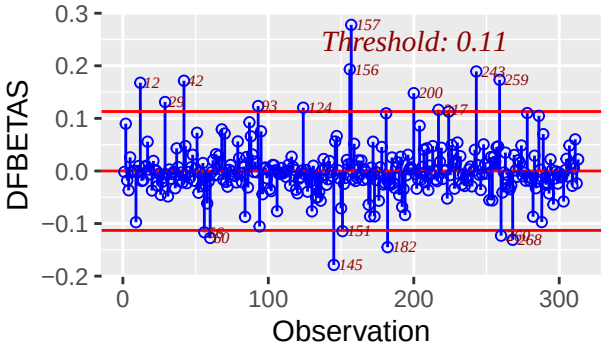
Influence Diagnostics for (Intercep



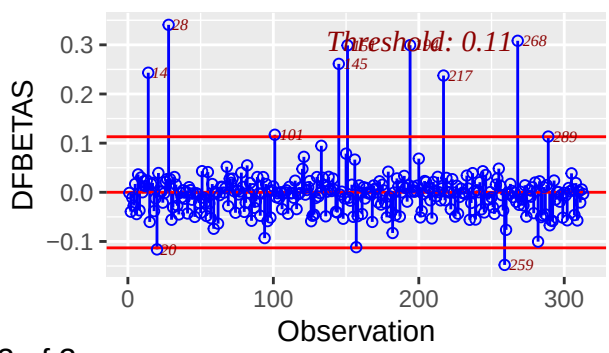
Influence Diagnostics for PM25



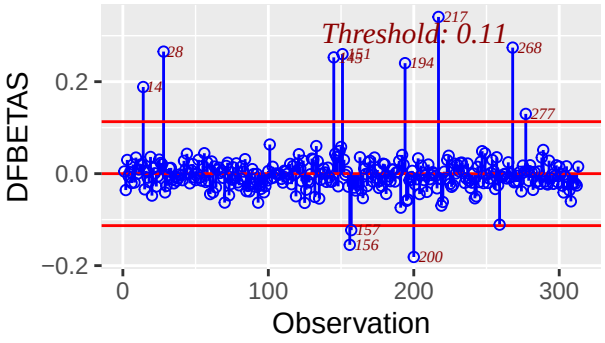
Influence Diagnostics for O3



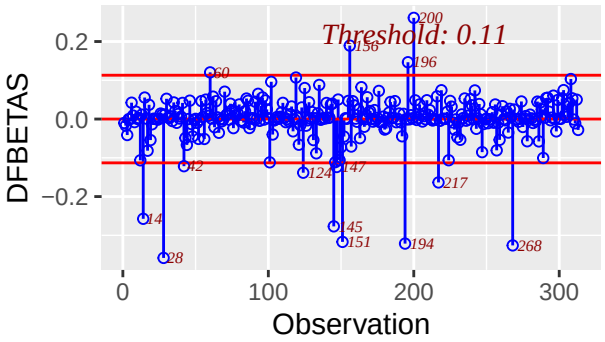
Influence Diagnostics for factor(Lc



Influence Diagnostics for Stay_At_Home

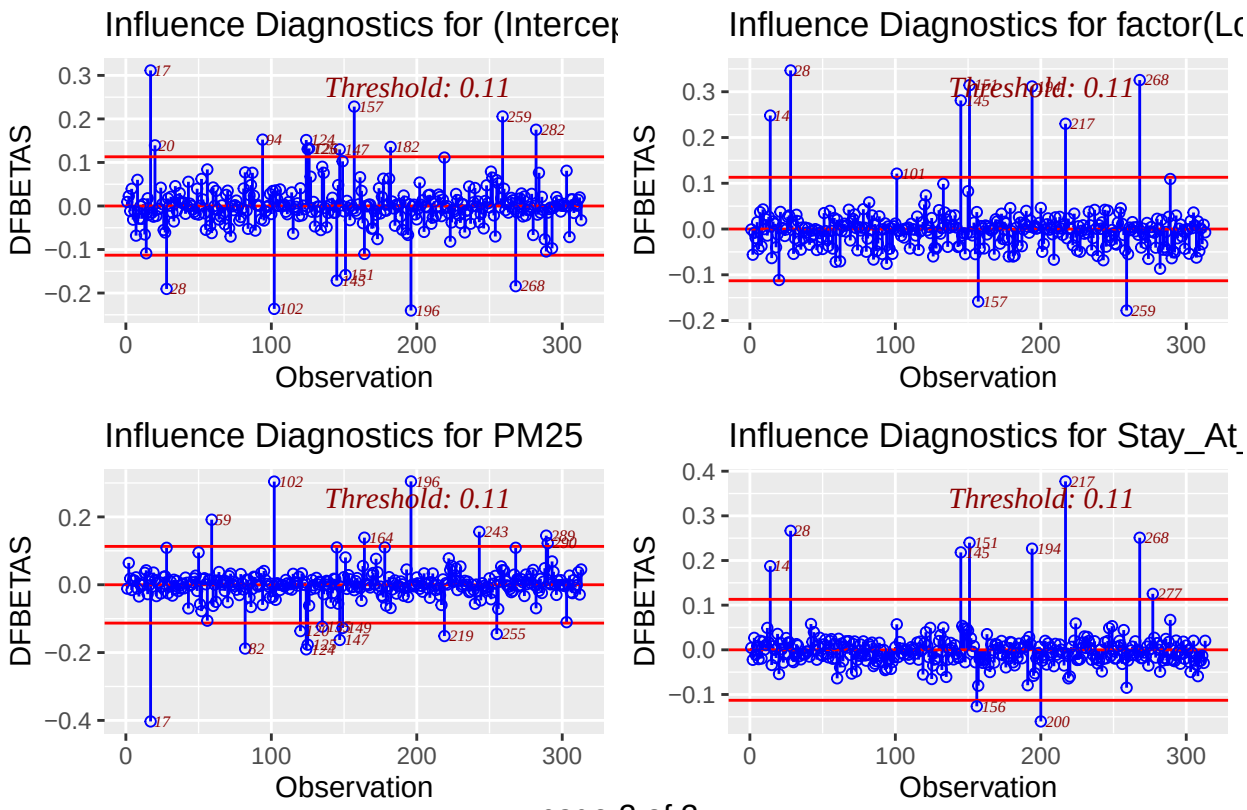


Influence Diagnostics for Bar_Close

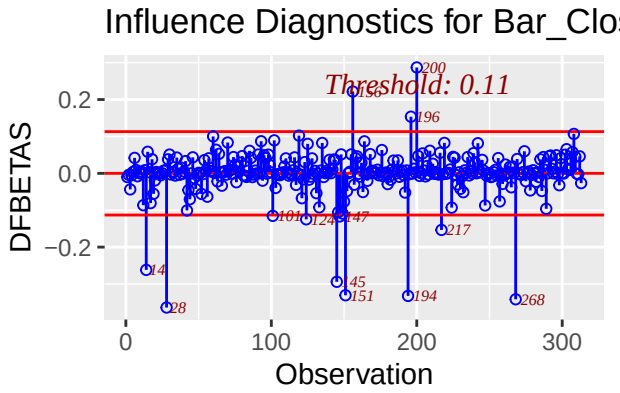


```
ols_plot_dfbetas(fixed_effect_intercept2)
```

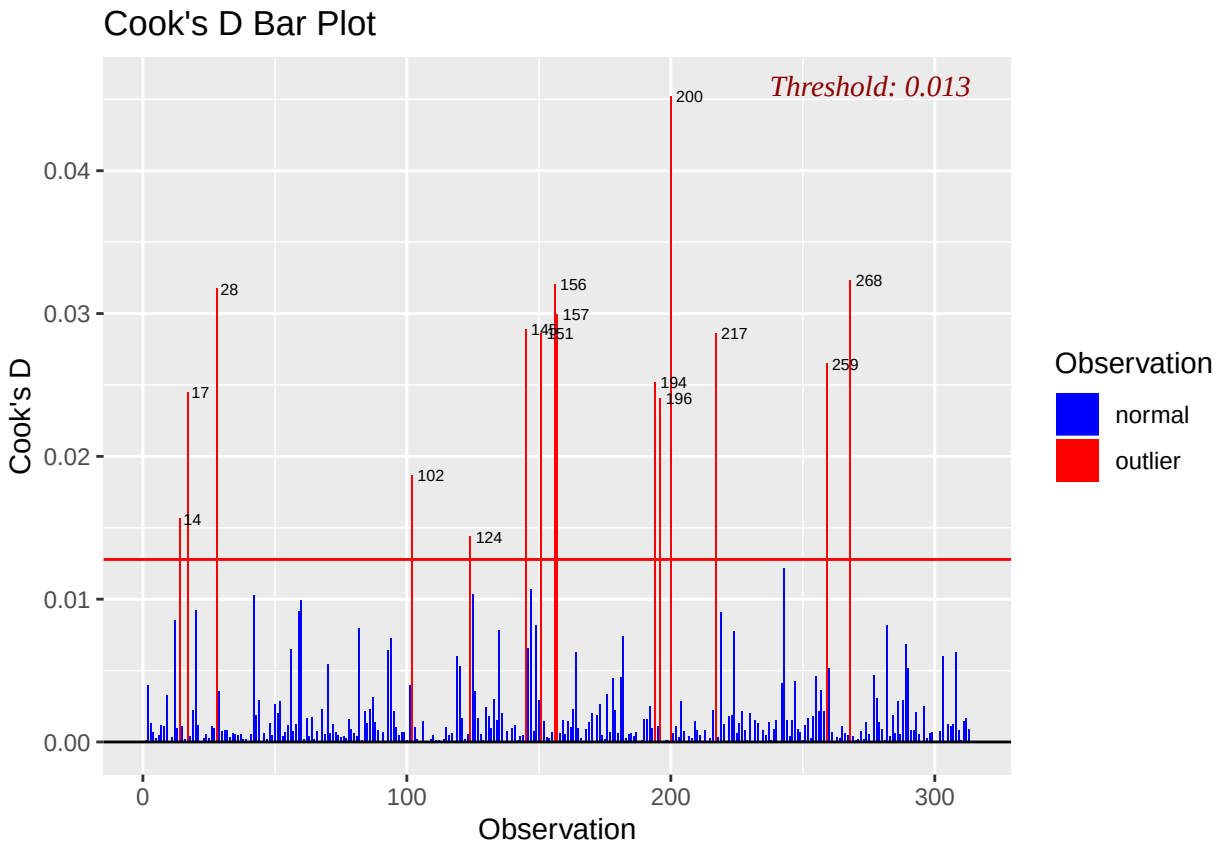
page 1 of 2



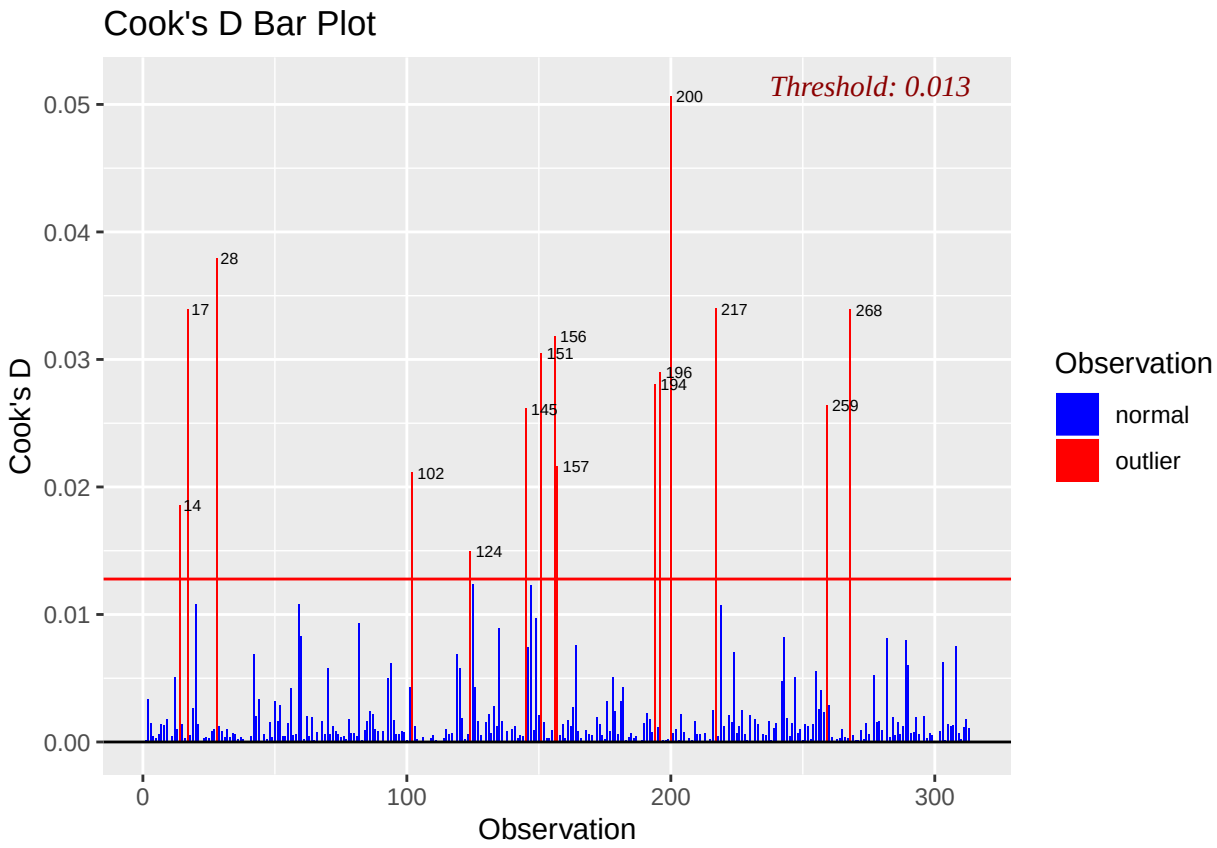
page 2 of 2



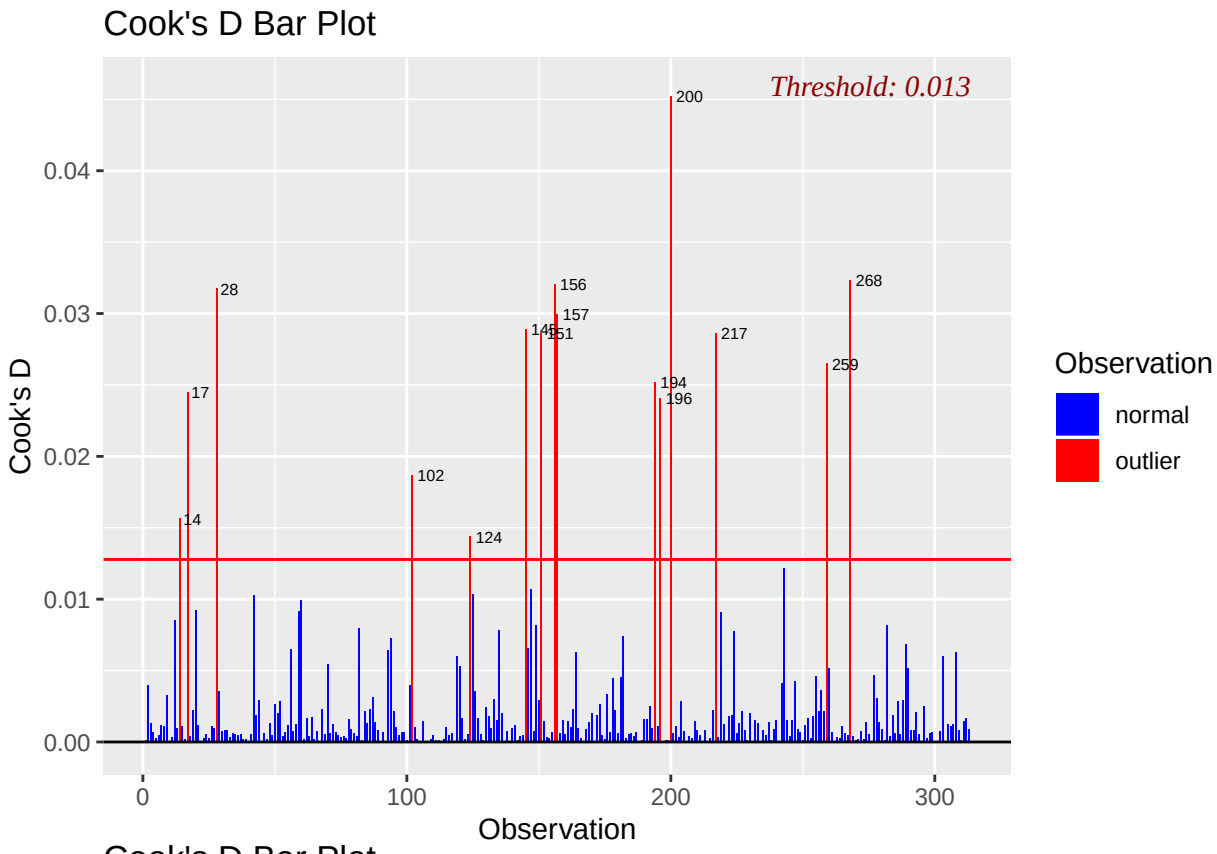

```
#Determining high leverage points using Cook's Distance. Threshold established as 4/(n-k-1),
#where n is the number of observations and k denotes the number of independent variables.
ols_plot_cooksd_bar(fixed_effect_intercept1)
```

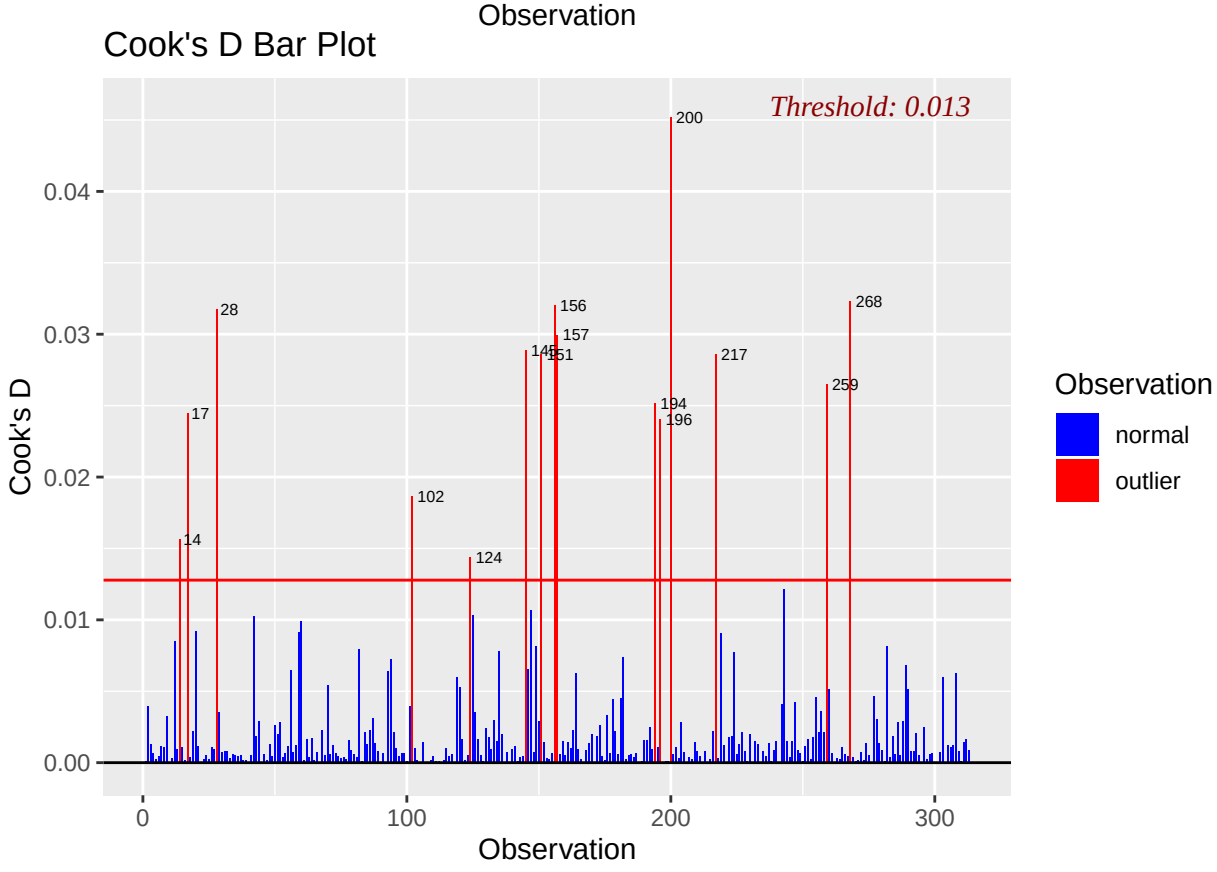



```
ols_plot_cooksd_bar(fixed_effect_intercept2)
```

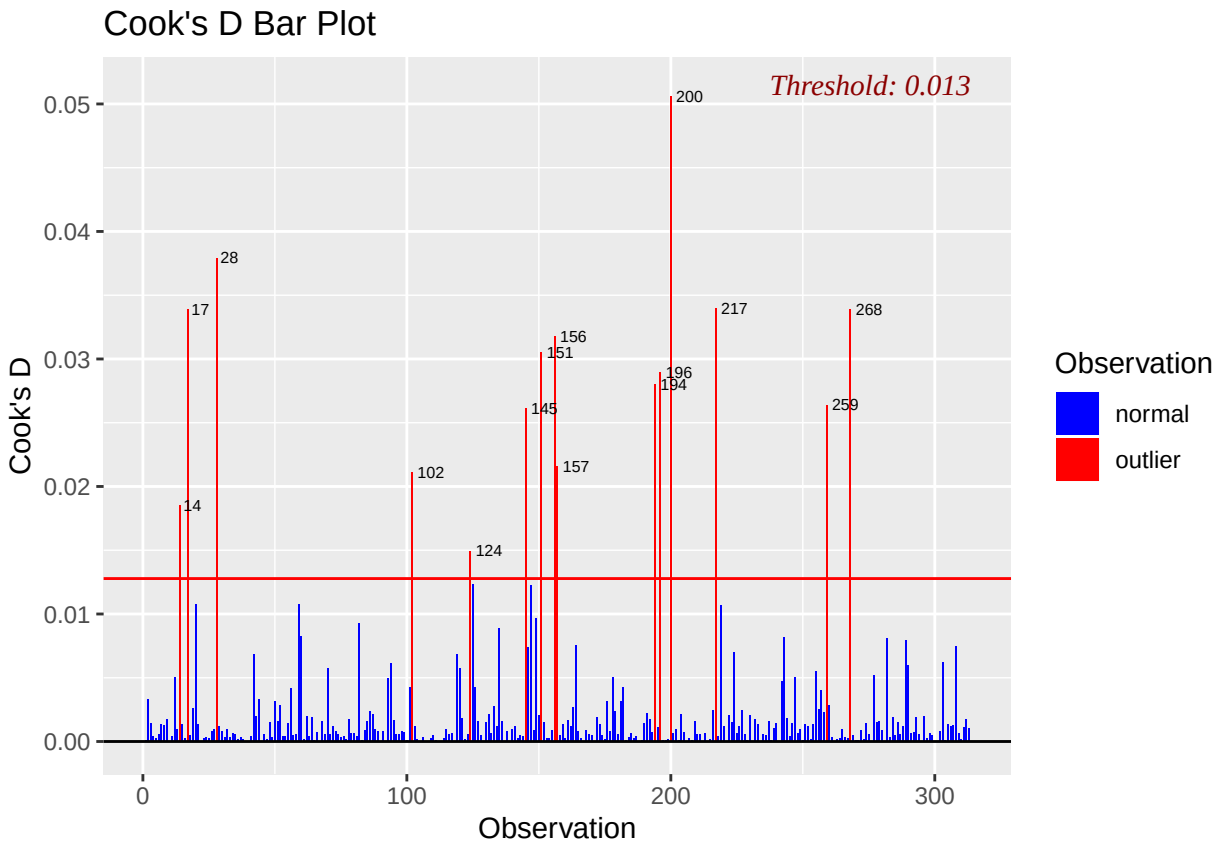


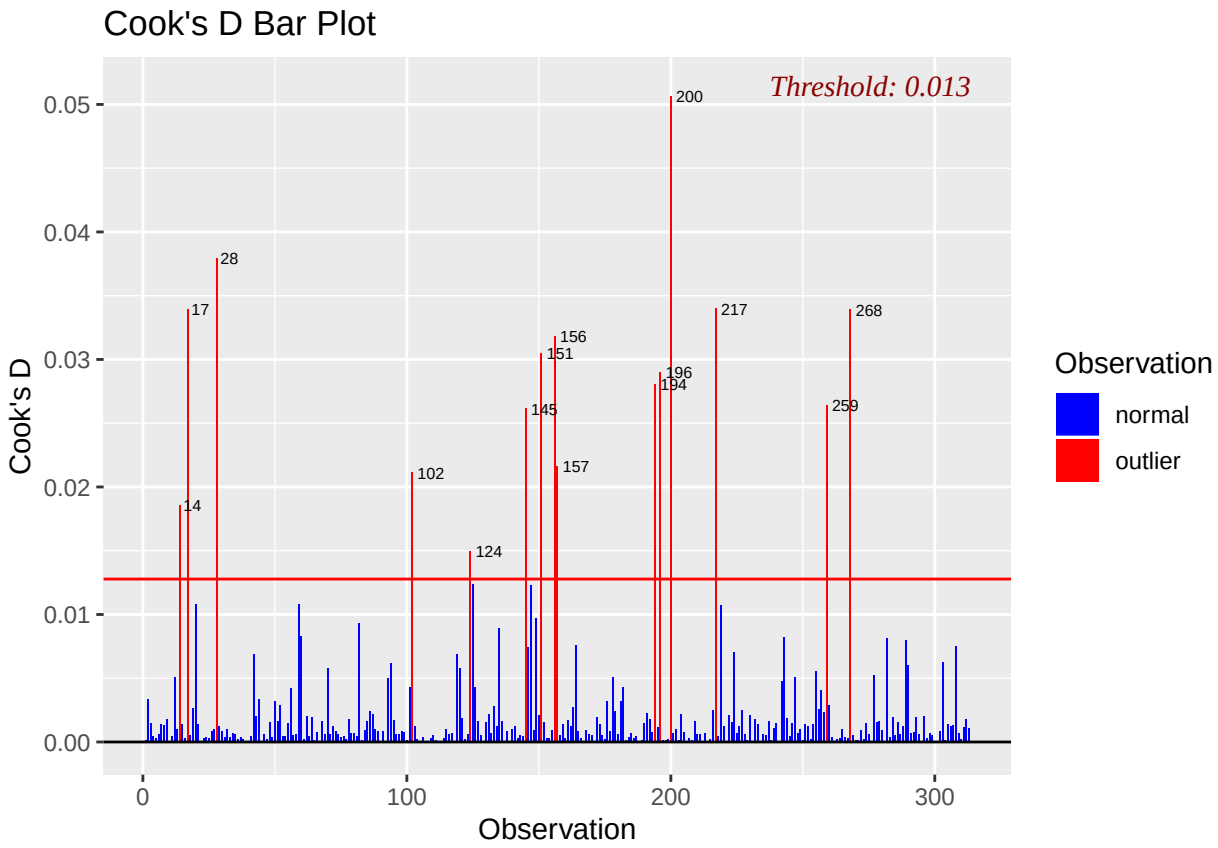

```
#Identifying Cook's D statistic outliers with threshold 4/(n-k-1)
mod1outliers <- ols_plot_cooksd_bar(fixed_effect_intercept1)$data[ols_plot_cooksd_bar(fixed_effect_inte
```


Cook's D Bar Plot
Threshold: 0.013
Cook's D
Observation
0.00
0.01
0.02
0.03
0.04
0
100
200
300
200
28
156
157
145
151
268
217
259
17
194
196
102
14
124
Observation
normal
outlier


Cook's D Bar Plot
Threshold: 0.013
Cook's D
Observation
0.00
0.01
0.02
0.03
0.04
0
100
200
300
200
28
156
157
145
151
268
217
259
17
194
196
102
14
124
Observation
normal
outlier

```
mod2outliers <- ols_plot_cooksd_bar(fixed_effect_intercept2)$data[ols_plot_cooksd_bar(fixed_effect_inte
```






```
#train[mod1outliers$obs,]
```

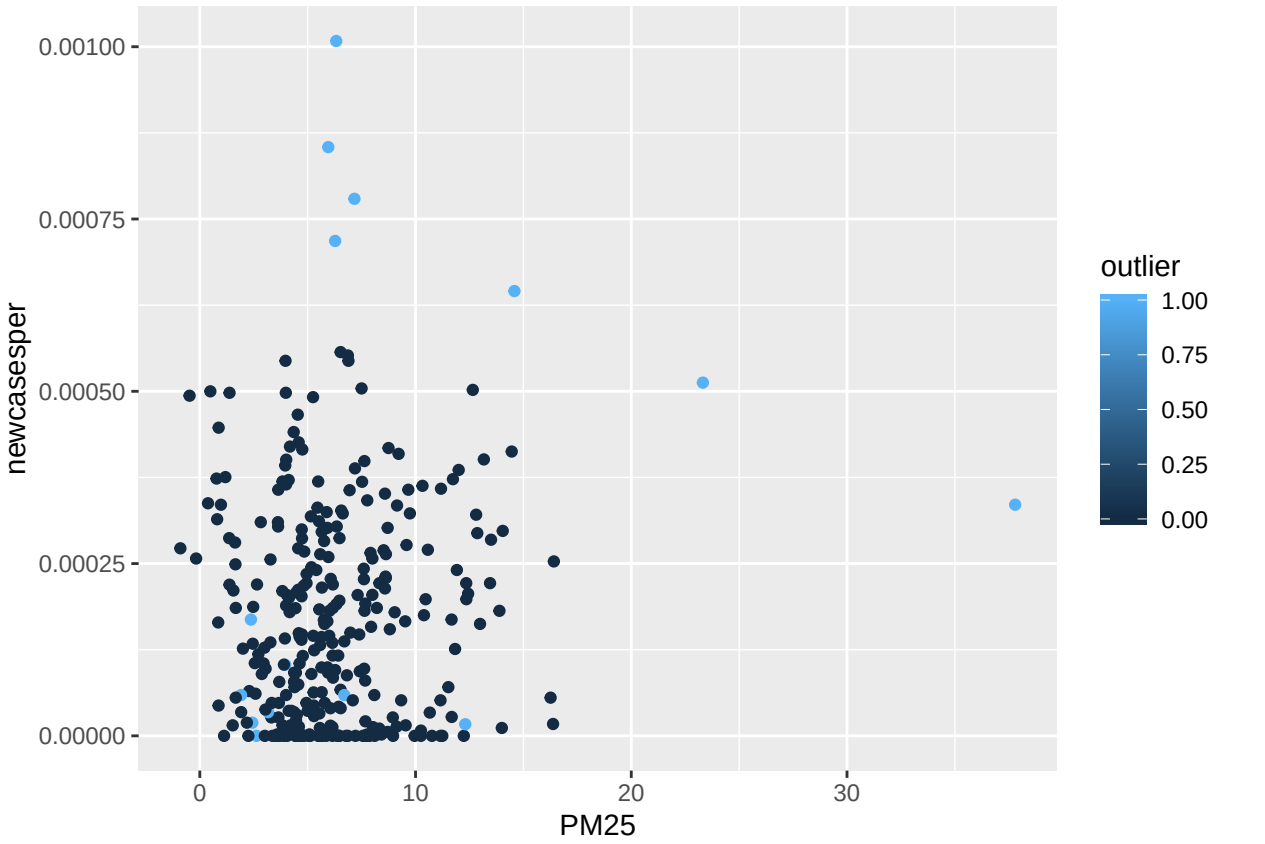
```
#train[mod2outliers$obs,]
```

```
#train[union(mod1outliers$obs, mod2outliers$obs),]
```

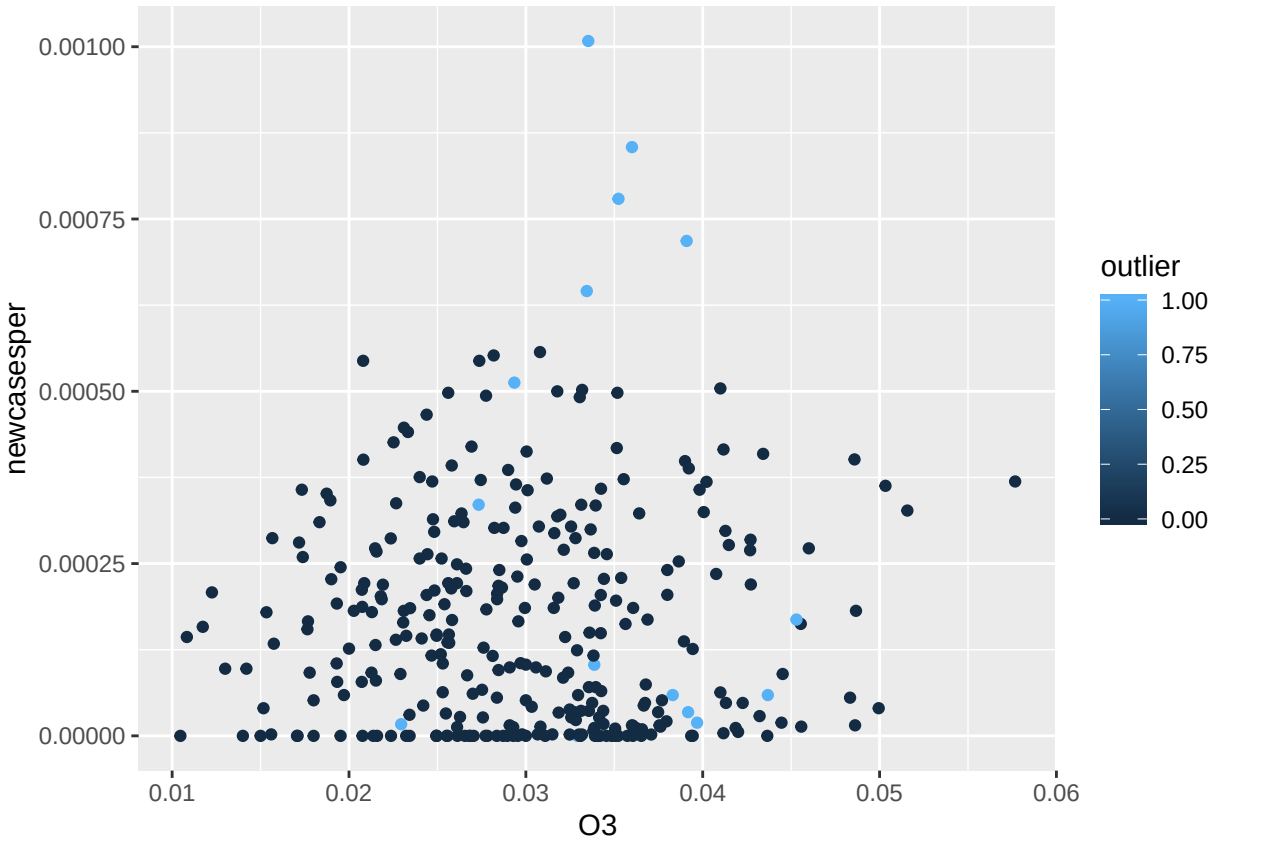
```
#train[intersect(mod1outliers$obs, mod2outliers$obs),]
```

```
train3 <- train
train3$outlier <- ifelse(train3$X1 %in% train[union(mod1outliers$obs,mod2outliers$obs),]$X1,1,0)
```

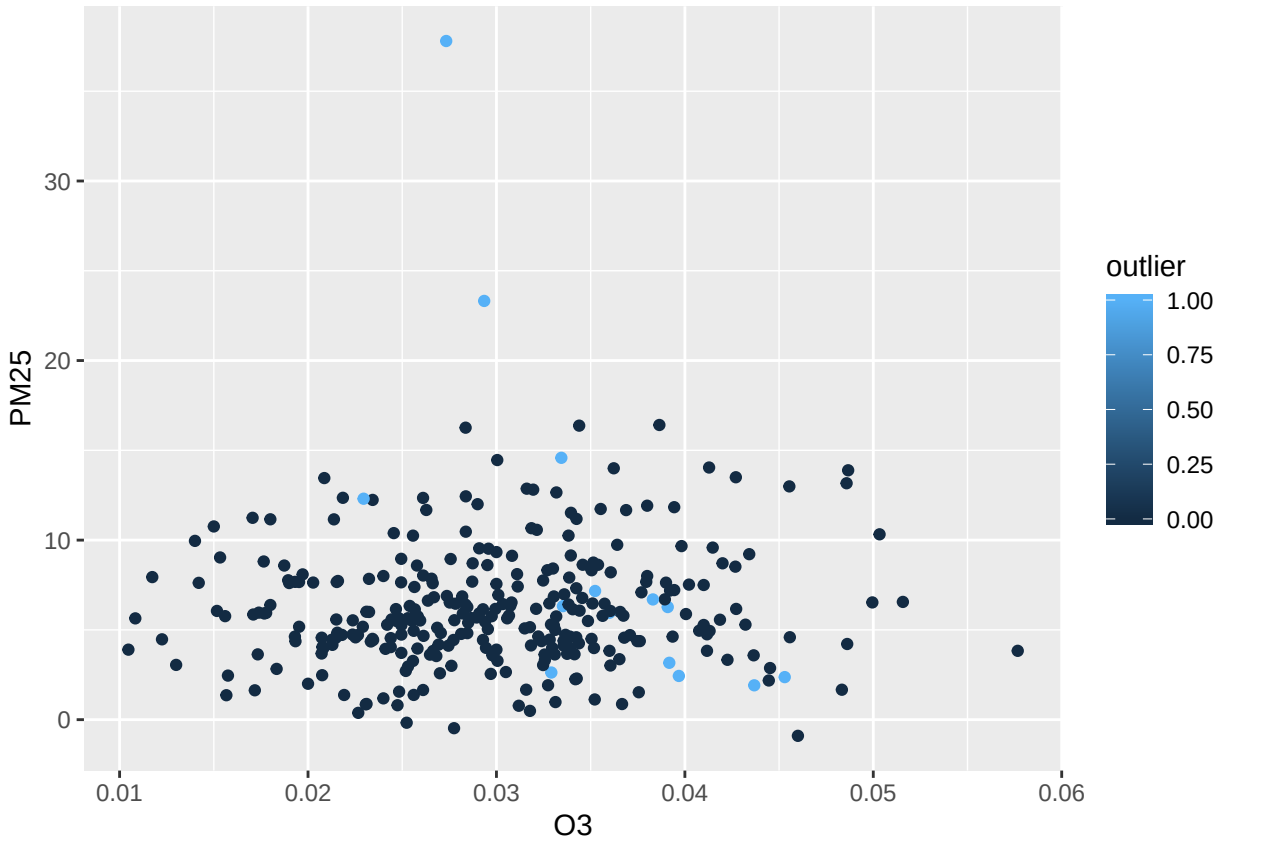
```
#Visualizing outliers
ggplot(train3, aes(y= newcasesper, x = PM25, color = outlier)) + geom_point()
```



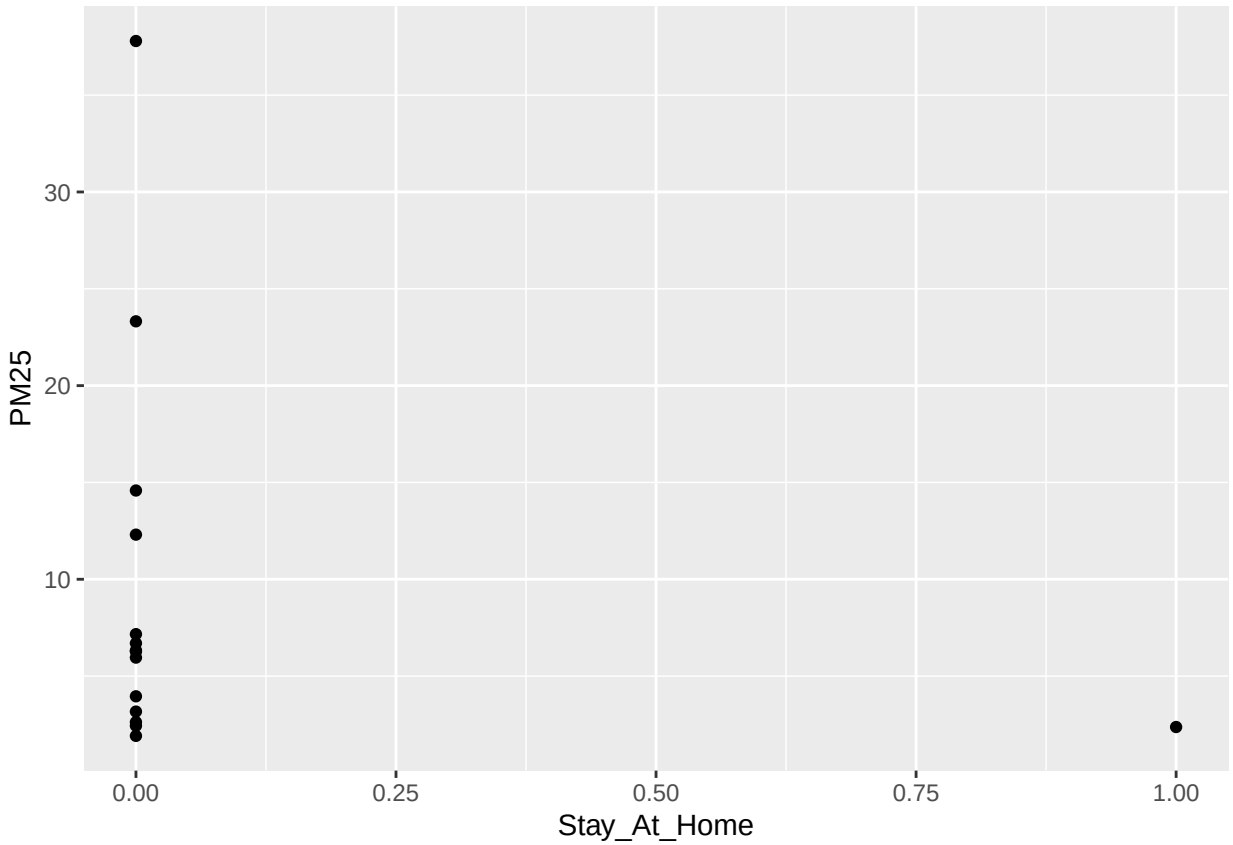

```
ggplot(train3, aes(y= newcasesper, x = O3, color = outlier)) + geom_point()
```



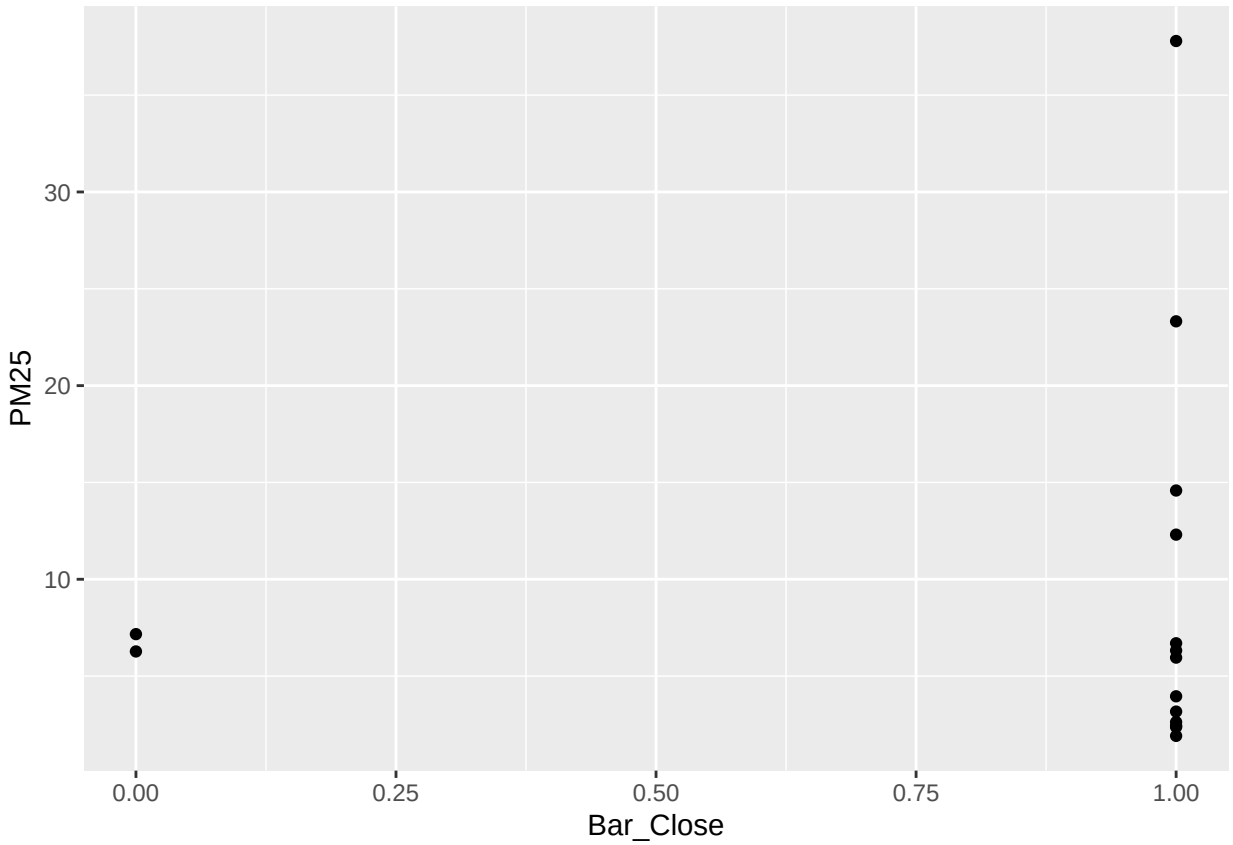

```
ggplot(train3, aes(y= PM25, x = O3, color = outlier)) + geom_point()
```



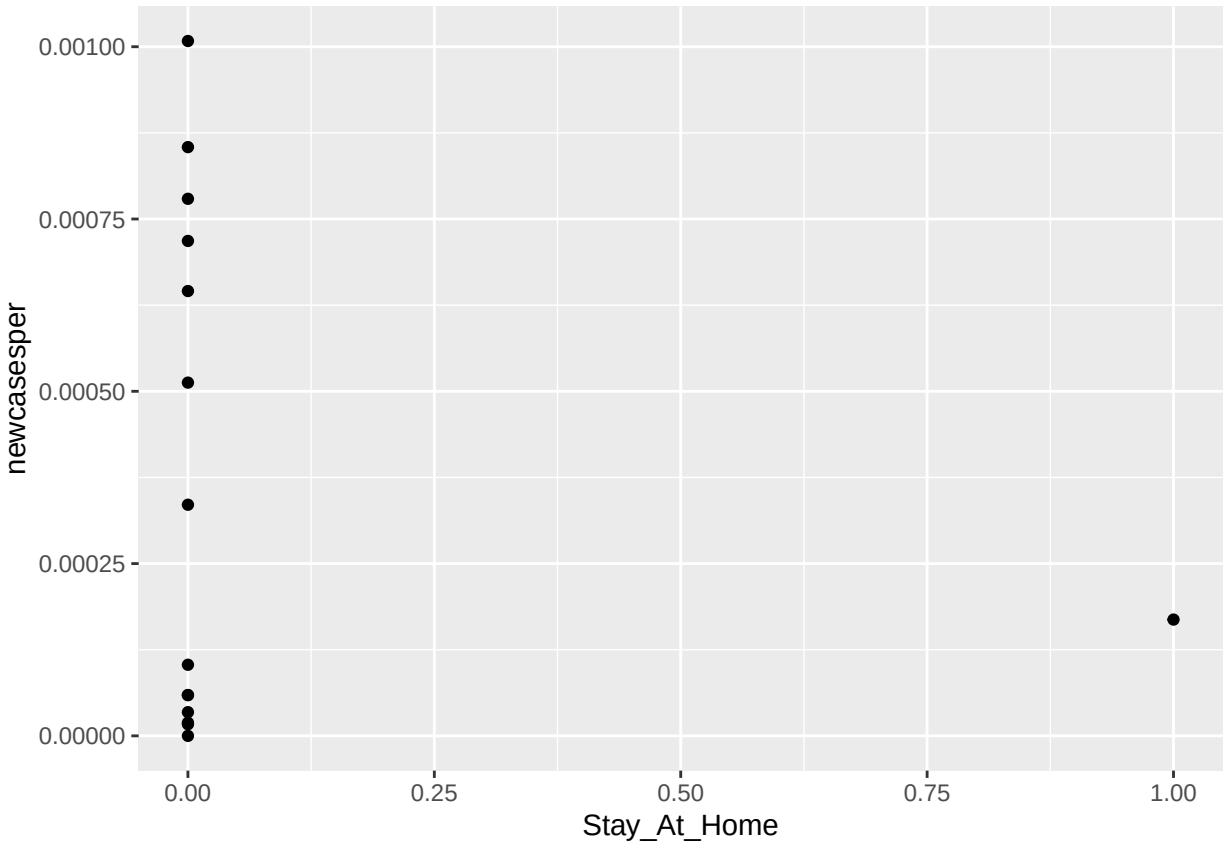

```
ggplot(train3[train3$outlier == 1,], aes(y= PM25, x = Stay_At_Home)) + geom_point()
```



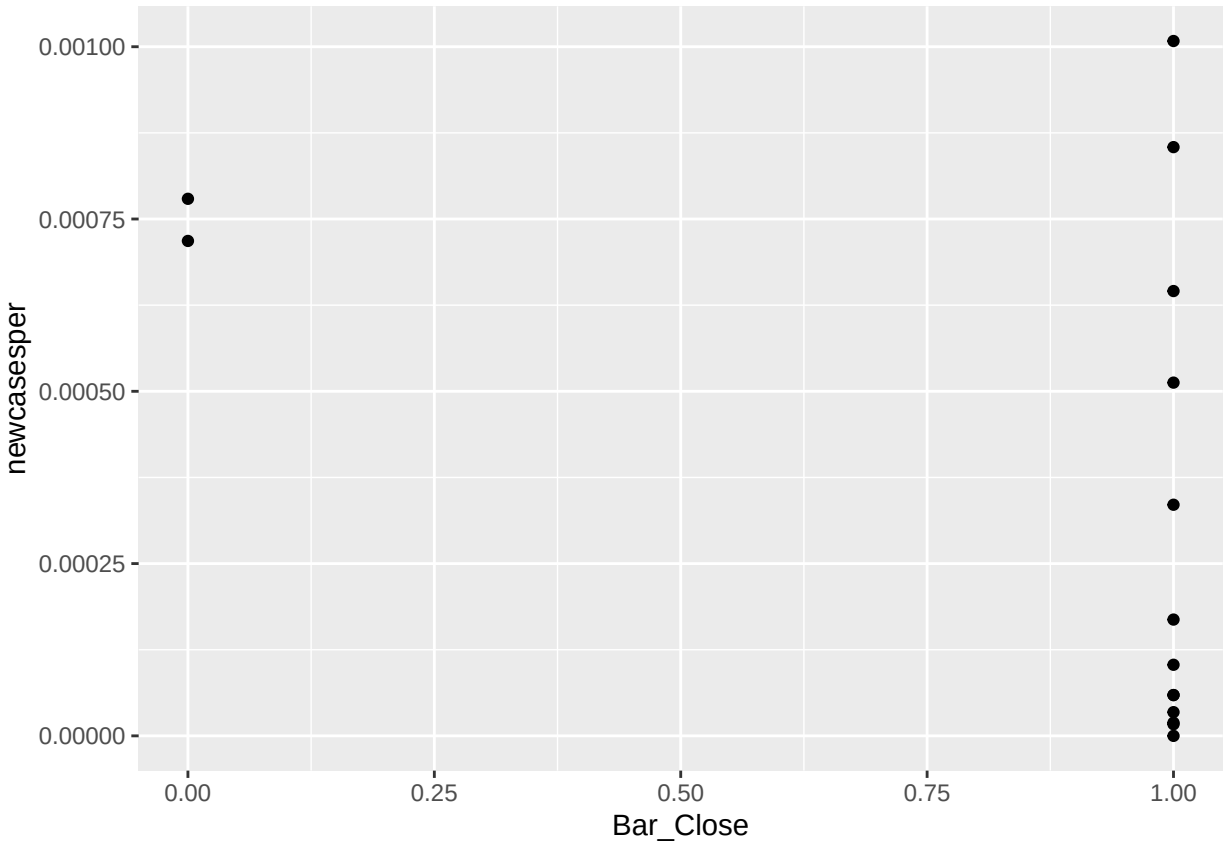

```
ggplot(train3[train3$outlier == 1,], aes(y= PM25, x = Bar_Close)) + geom_point()
```




```
ggplot(train3[train3$outlier == 1,], aes(y= newcasesper, x = Stay_At_Home)) + geom_point()
```




```
ggplot(train3[train3$outlier == 1,], aes(y= newcasesper, x = Bar_Close)) + geom_point()
```




```
#Refitting models without outlier observations as sensitivity analysis
train1 <- train[-mod1outliers$obs,]
train2 <- train[-mod2outliers$obs,]

fixed_effect_intercept1.no <- lm(newcasesper ~ O3 + PM25 + factor(Location) + Stay_At_Home + Bar_Close,
fixed_effect_intercept2.no <- lm(newcasesper ~ PM25 + factor(Location) + Stay_At_Home + Bar_Close, trai
```

```
#Model fits without the omission of outliers
summary(fixed_effect_intercept1)
```

```
##
## Call:
## lm(formula = newcasesper ~ O3 + PM25 + factor(Location) + Stay_At_Home +
##     Bar_Close, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.399e-04 -8.439e-05 -3.246e-05  8.014e-05  7.320e-04
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                9.186e-05  3.599e-05   2.552 0.011194 *
## O3                         8.067e-04  1.082e-03   0.745 0.456696
## PM25                       4.731e-06  2.218e-06   2.132 0.033760 *
## factor(Location)Raleigh -8.168e-05  2.266e-05  -3.605 0.000364 ***
## Stay_At_Home            -2.996e-04  4.075e-05  -7.353 1.78e-12 ***
## Bar_Close                2.091e-04  2.384e-05   8.770  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001455 on 307 degrees of freedom
## Multiple R-squared:  0.2575, Adjusted R-squared:  0.2454
## F-statistic:  21.3 on 5 and 307 DF,  p-value: < 2.2e-16
```

```
summary(fixed_effect_intercept2)
```

```
##
## Call:
## lm(formula = newcasesper ~ PM25 + factor(Location) + Stay_At_Home +
##     Bar_Close, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.394e-04 -8.528e-05 -3.233e-05  7.515e-05  7.351e-04
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.147e-04  1.877e-05   6.115 2.92e-09 ***
## PM25                     4.916e-06  2.203e-06   2.232 0.026340 *
## factor(Location)Raleigh -8.431e-05  2.236e-05  -3.770 0.000196 ***
## Stay_At_Home            -2.947e-04  4.019e-05  -7.333 2.00e-12 ***
## Bar_Close                2.117e-04  2.357e-05   8.982  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001454 on 308 degrees of freedom
## Multiple R-squared:  0.2562, Adjusted R-squared:  0.2465
## F-statistic: 26.52 on 4 and 308 DF,  p-value: < 2.2e-16
```

```
#Model fits with omission of Cook's D statistic outliers
summary(fixed_effect_intercept1.no)
```

```
##
## Call:
## lm(formula = newcasesper ~ O3 + PM25 + factor(Location) + Stay_At_Home +
##     Bar_Close, data = train1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.699e-04 -8.132e-05 -1.512e-05  6.799e-05  3.987e-04
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.219e-04  2.973e-05   4.101 5.35e-05 ***
## O3                       3.537e-04  8.921e-04   0.396    0.692
## PM25                     3.117e-06  2.219e-06   1.405    0.161
## factor(Location)Raleigh -1.304e-04  1.975e-05  -6.605 1.88e-10 ***
## Stay_At_Home            -3.600e-04  3.544e-05 -10.160  < 2e-16 ***
## Bar_Close                2.541e-04  2.117e-05  12.002  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000117 on 292 degrees of freedom
```

```
## Multiple R-squared:  0.3747, Adjusted R-squared:  0.364
## F-statistic:    35 on 5 and 292 DF,  p-value: < 2.2e-16
```

```
summary(fixed_effect_intercept2.no)
```

```
##
## Call:
## lm(formula = newcasesper ~ PM25 + factor(Location) + Stay_At_Home +
##     Bar_Close, data = train2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.743e-04 -7.922e-05 -1.468e-05  6.926e-05  3.982e-04
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.314e-04  1.753e-05   7.496 7.87e-13 ***
## PM25                       3.263e-06  2.186e-06   1.493    0.136
## factor(Location)Raleigh -1.313e-04  1.959e-05  -6.705 1.04e-10 ***
## Stay_At_Home            -3.577e-04  3.489e-05 -10.252  < 2e-16 ***
## Bar_Close                  2.551e-04  2.100e-05  12.146  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001169 on 293 degrees of freedom
## Multiple R-squared:  0.3744, Adjusted R-squared:  0.3659
## F-statistic: 43.84 on 4 and 293 DF,  p-value: < 2.2e-16
```

```
#Examination of multicollinearity in the model using the variance inflation factor
vif(fixed_effect_intercept1) %>%
  pander(caption = "Variance Inflation Factor values for 5 parameter model")
```

O3	PM25	factor(Location)	Stay_At_Home	Bar_Close
1.129	1.039	1.891	1.537	2.017

```
vif(fixed_effect_intercept2) %>%
  pander(caption = "Variance Inflation Factor values for 4 parameter model")
```

PM25	factor(Location)	Stay_At_Home	Bar_Close
1.026	1.845	1.498	1.974

Panel regression models with zero intercept

```
fixed_effect2 <- lm(newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close + loc - 1, train)
```

Selection methods applied to zero-intercept model

```
reglm2 <- regsubsets(newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close + loc, train, intercept = F)
lm2s <- summary(reglm2)
lm2 <- lm(newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close + loc - 1, train)

#AIC
fixed_effect2.1 <- step(lm2)
```

```
## Start:  AIC=-5520.48
## newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close + loc - 1
##
##                Df  Sum of Sq        RSS     AIC
## <none>                       6.6357e-06 -5520.5
## - loc           1 1.9388e-07 6.8296e-06 -5513.5
## - PM25          1 1.9621e-07 6.8319e-06 -5513.4
## - O3            1 6.6418e-07 7.2999e-06 -5492.6
## - Stay_At_Home  1 1.1724e-06 7.8080e-06 -5471.6
## - Bar_Close     1 1.5770e-06 8.2127e-06 -5455.7
```

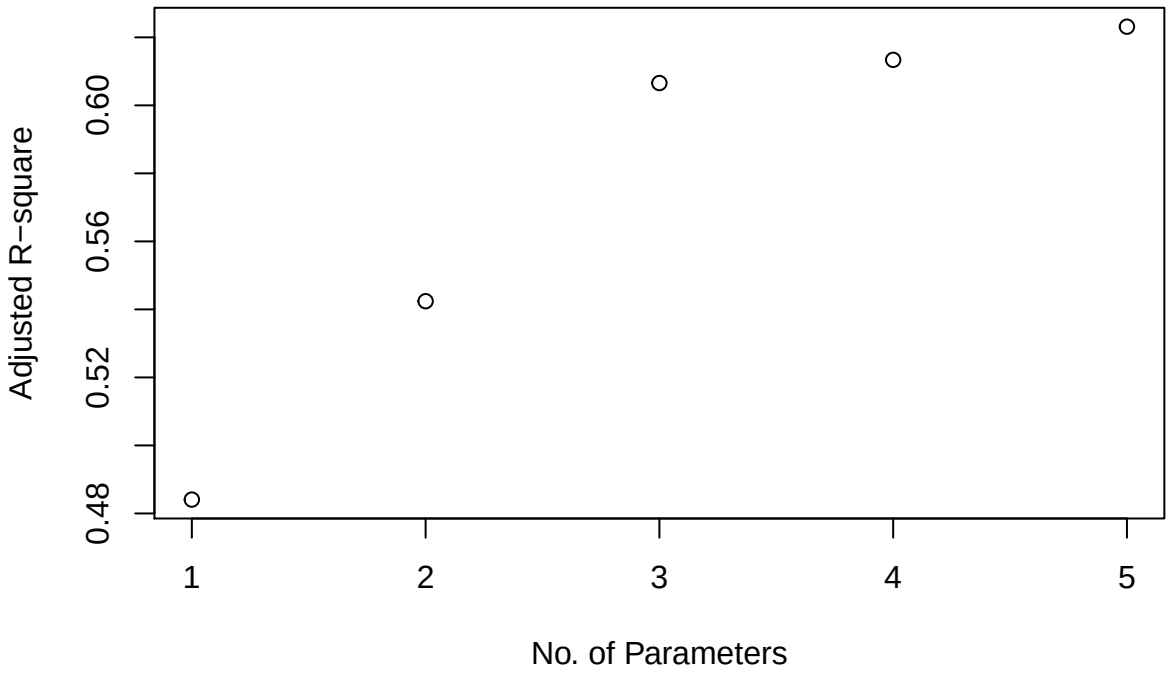
```
summary(fixed_effect2.1)
```

```
##
## Call:
## lm(formula = newcasesper ~ PM25 + O3 + Stay_At_Home + Bar_Close +
##     loc - 1, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.276e-04 -8.153e-05 -1.572e-05  8.535e-05  7.221e-04
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## PM25          6.439e-06  2.134e-06   3.018  0.00276 **
## O3            3.164e-03  5.698e-04   5.552 6.09e-08 ***
## Stay_At_Home -3.031e-04  4.109e-05  -7.377 1.52e-12 ***
## Bar_Close     2.054e-04  2.401e-05   8.556 5.62e-16 ***
## loc          -6.601e-05  2.200e-05  -3.000  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001468 on 308 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6231
## F-statistic: 104.5 on 5 and 308 DF,  p-value: < 2.2e-16
```

```
#Adjusted R^2
lm2s$which[which.max(lm2s$adjr2),]
```

```
##         PM25           O3 Stay_At_Home    Bar_Close          loc
##         TRUE         TRUE         TRUE         TRUE         TRUE
```

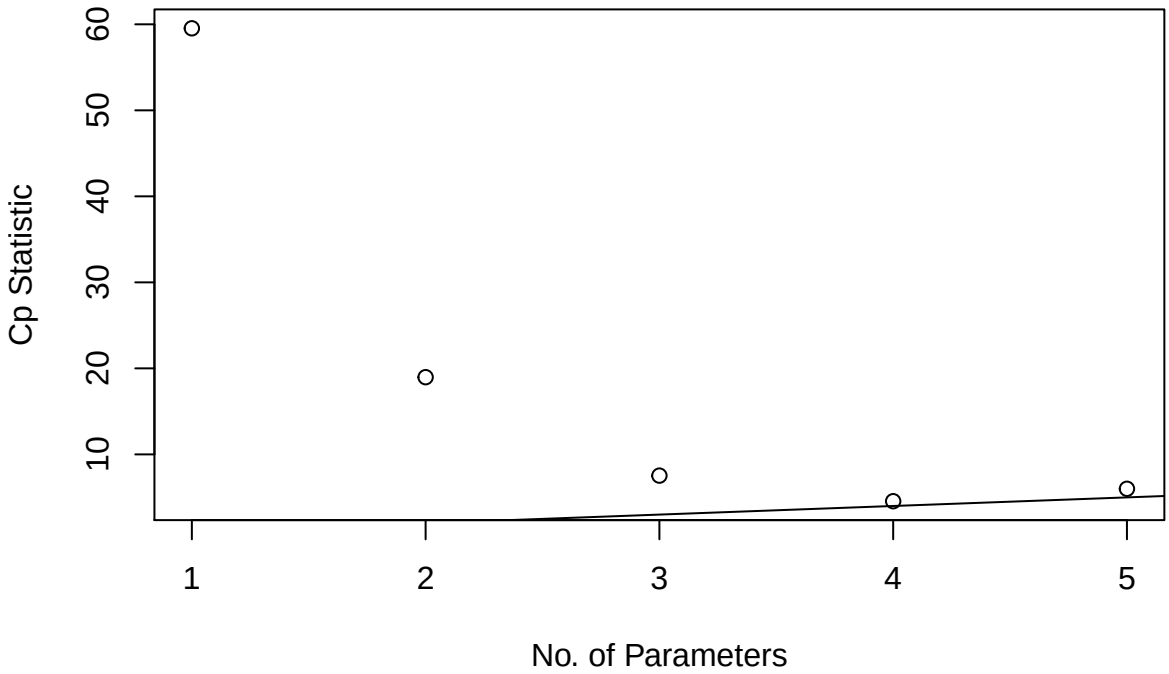
```
plot(1:5,lm2s$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
```




```
#Mallow's Cp
lm2s$which[which.min(lm2s$cp), ]
```

```
##         PM25           O3 Stay_At_Home    Bar_Close          loc
##         TRUE         TRUE         TRUE         TRUE         TRUE
```

```
plot(1:5,lm1s$cp,xlab="No. of Parameters",ylab="Cp Statistic")
abline(0,1)
```




```
#From the above variable selection methods we see that the zero-intercept
#model that optimizes the Adjusted R^2, AIC, and Mallow's Cp metrics
#is the one predicting `newcasesper` by the variables `PM25`, 'O3', `loc`,
```

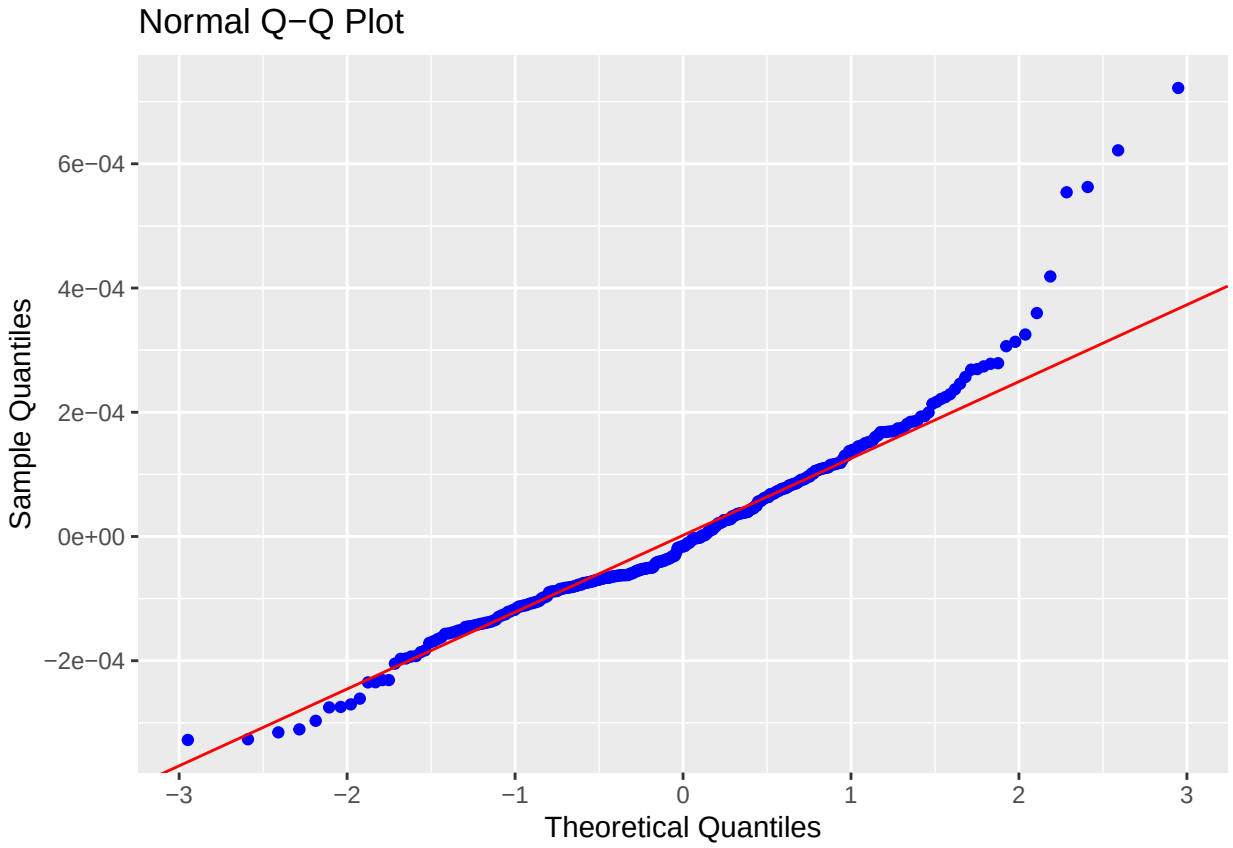
```
#`Stay_At_Home`, and `Bar_Close`. In other words, the complete model
#is the one that optimizes the values of all three variable selection
#methods. We will thereby proceed by considering this model.

fixed_effect_ni <- lm(newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close - 1, train)

#Calculating the prediction RMSE on the test set
rmse(fitted(fixed_effect_ni), train$newcasesper)
```

```
## [1] 0.0001456031
```

```
rmse(predict(fixed_effect_ni,test),test$newcasesper)
```

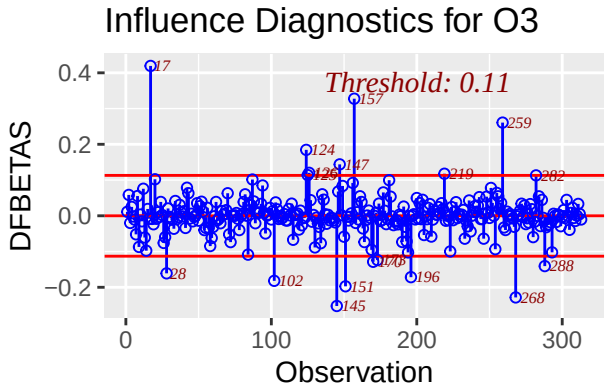
```
## [1] 0.000120756
```

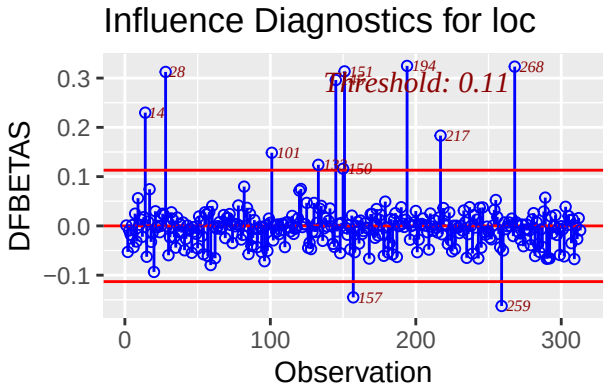
```
#Considering the residual Q-Q plot to visualize any violations of the normality assumption.
ols_plot_resid_qq(fixed_effect_ni)
```

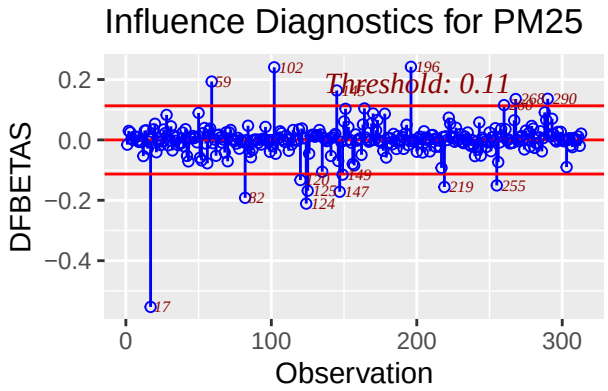


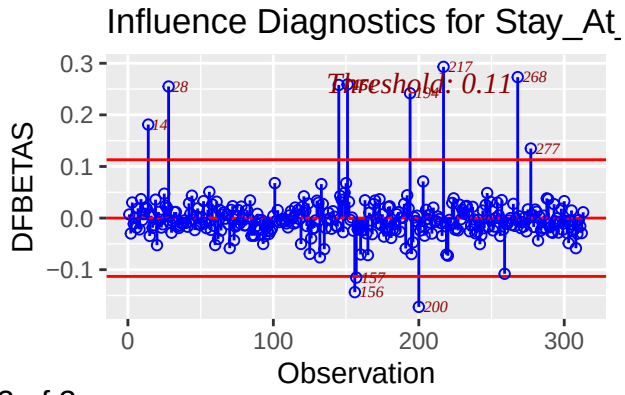

```
#Examining DFBETA plots to identify points particularly influential in estimating each parameter.
ols_plot_dfbetas(fixed_effect_ni)
```

page 1 of 2

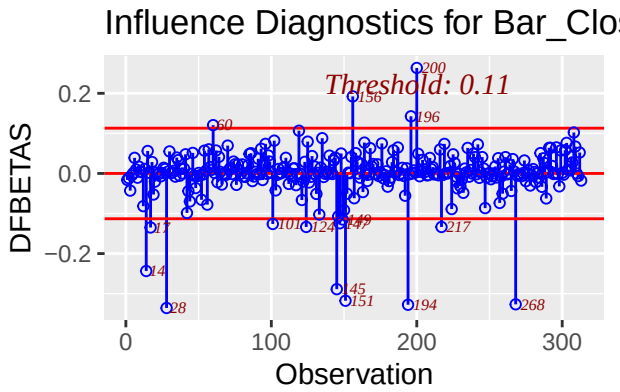




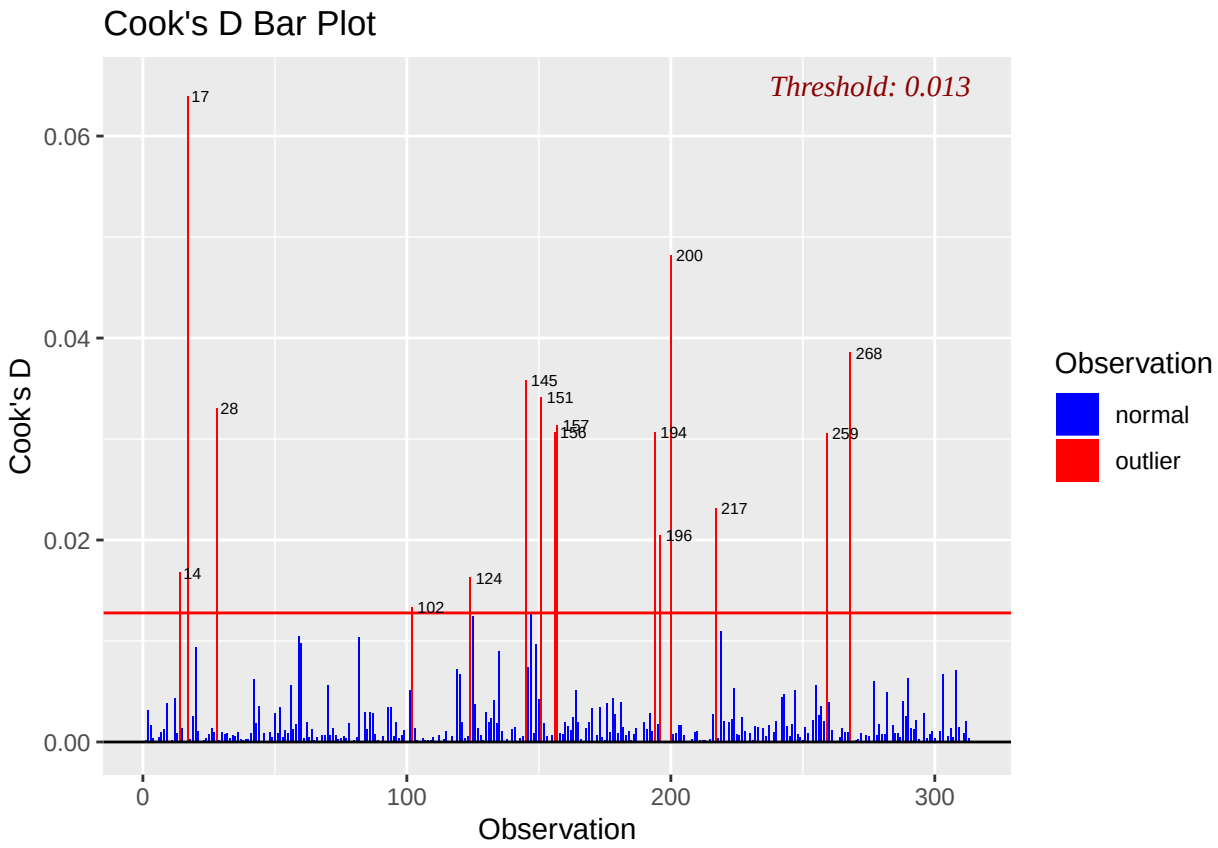




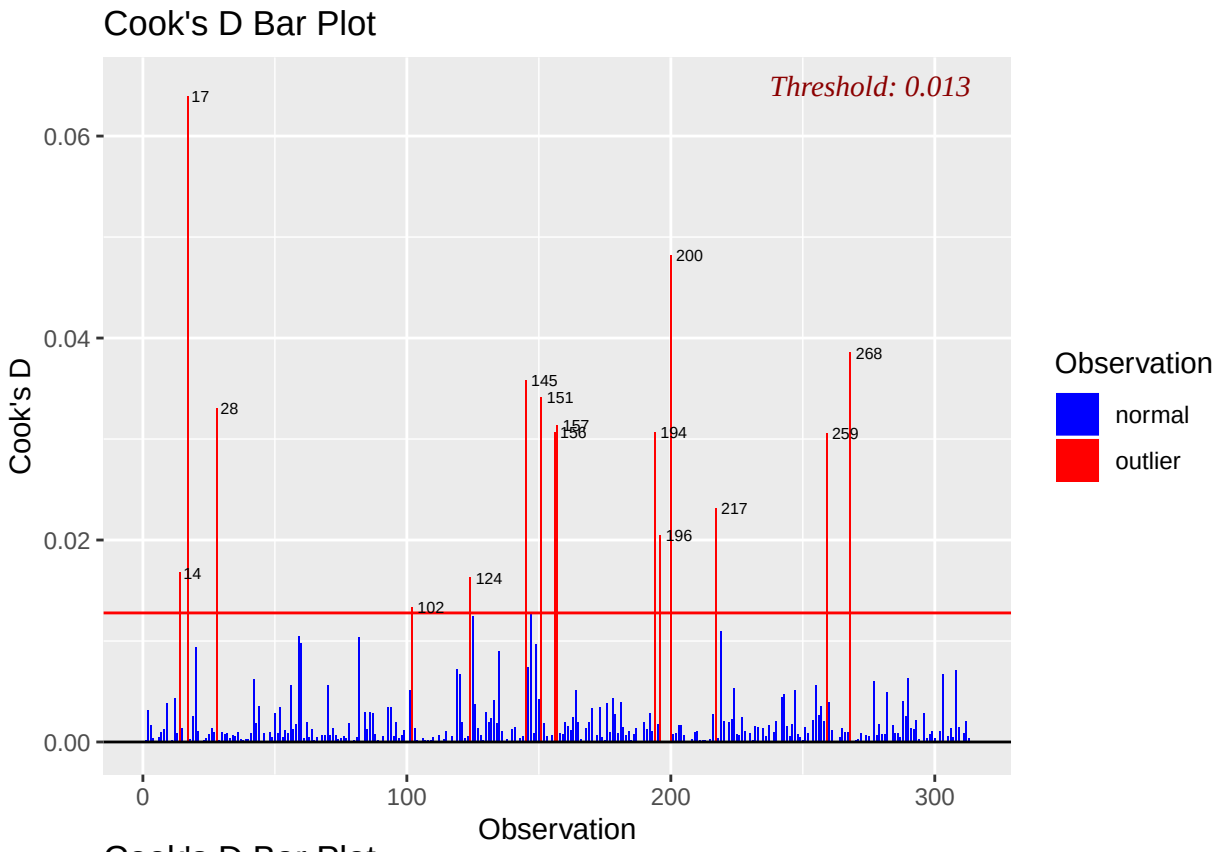
page 2 of 2

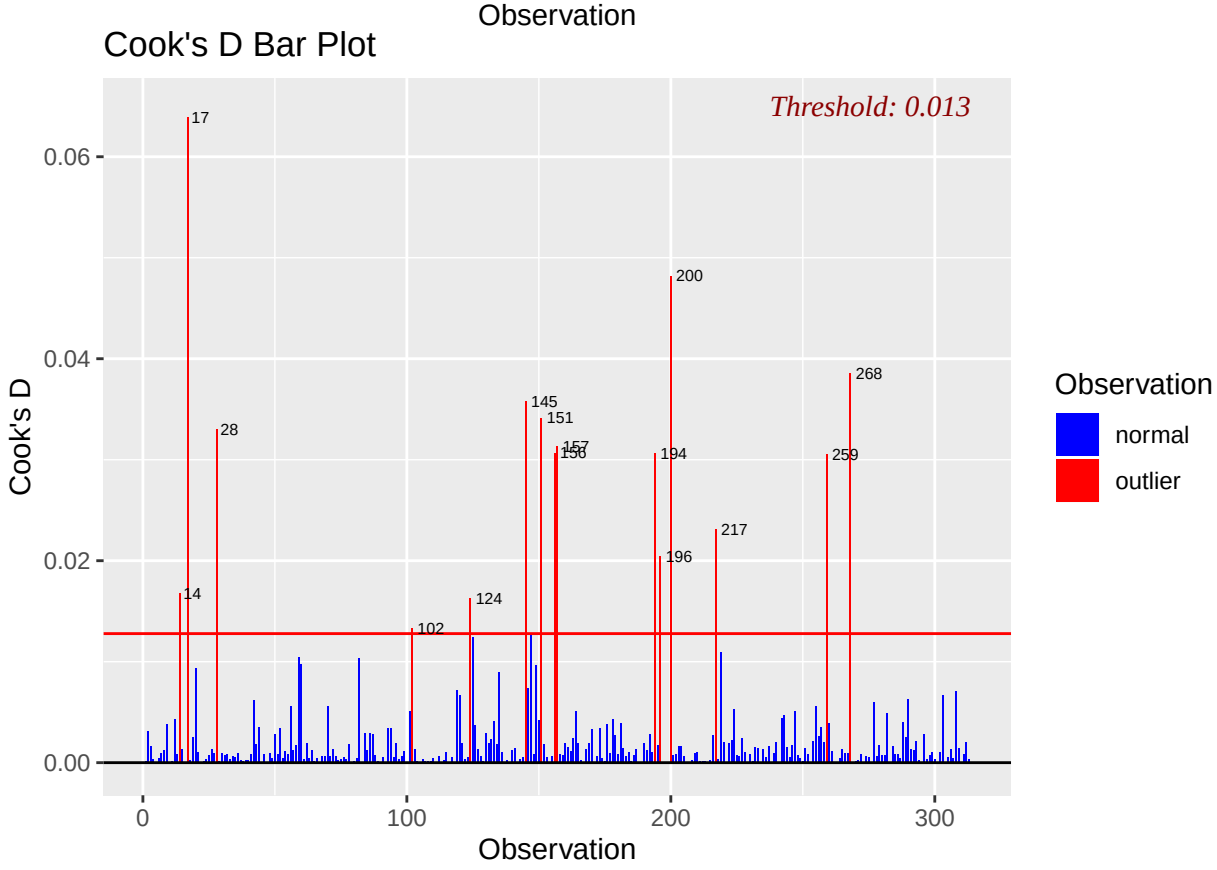



```
#Determining high leverage points using Cook's Distance.
ols_plot_cooksd_bar(fixed_effect_ni)
```




```
#Identifying Cook's D statistic outliers with threshold 4/(n-k-1)
mod3outliers <- ols_plot_cooksd_bar(fixed_effect_ni)$data[ols_plot_cooksd_bar(fixed_effect_ni)$data$fct_
```


Cook's D Bar Plot
Threshold: 0.013
Cook's D
0.06
0.04
0.02
0.00
17
28
14
102
124
145
151
157
156
200
194
196
217
259
268
0
100
200
300
Observation
Observation
normal
outlier


Cook's D Bar Plot
Threshold: 0.013
Cook's D
0.06
0.04
0.02
0.00
17
28
14
102
124
145
151
157
156
200
194
196
217
259
268
0
100
200
300
Observation
Observation
normal
outlier

```r
train4 <- train[-mod3outliers$obs,]

intersect(mod1outliers$obs,mod3outliers$obs)
```

```
##  [1]  14  17  28 102 124 145 151 156 157 194 196 200 217 259 268
```

```r
fixed_effect_ni.no <- lm(newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close - 1, train4)

#Model fit without the omission of outliers
summary(fixed_effect_ni)
```

```
##
## Call:
## lm(formula = newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close -
##     1, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.276e-04 -8.153e-05 -1.572e-05  8.535e-05  7.221e-04
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## O3            3.164e-03  5.698e-04   5.552 6.09e-08 ***
## PM25          6.439e-06  2.134e-06   3.018  0.00276 **
## loc          -6.601e-05  2.200e-05  -3.000  0.00292 **
## Stay_At_Home -3.031e-04  4.109e-05  -7.377 1.52e-12 ***
## Bar_Close     2.054e-04  2.401e-05   8.556 5.62e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001468 on 308 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6231
## F-statistic: 104.5 on 5 and 308 DF,  p-value: < 2.2e-16
```

```r
#Model fit with omission of Cook's D statistic outliers
summary(fixed_effect_ni.no)
```

```
##
## Call:
## lm(formula = newcasesper ~ O3 + PM25 + loc + Stay_At_Home + Bar_Close -
##     1, data = train4)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.800e-04 -7.048e-05 -6.740e-06  7.922e-05  4.159e-04
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## O3            3.306e-03  5.409e-04   6.112 3.14e-09 ***
## PM25          6.264e-06  2.138e-06   2.930  0.00366 **
## loc          -1.100e-04  1.962e-05  -5.609 4.71e-08 ***
## Stay_At_Home -3.626e-04  3.637e-05  -9.970  < 2e-16 ***
## Bar_Close     2.517e-04  2.172e-05  11.585  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.0001201 on 293 degrees of freedom
## Multiple R-squared:  0.7019, Adjusted R-squared:  0.6968
## F-statistic: 137.9 on 5 and 293 DF,  p-value: < 2.2e-16
```

```
vif(fixed_effect_ni) %>%
  pander(caption = "Variance Inflation Factor values for zero-intercept, 5 parameter model")
```

```
## Warning in vif.default(fixed_effect_ni): No intercept: vifs may not be
## sensible.
```

O3	PM25	loc	Stay_At_Home	Bar_Close
4.545	3.57	3.304	1.645	3.345

```
cor(train[,c("O3","PM25","loc","Stay_At_Home","Bar_Close")]) %>%
  pander
```

	O3	PM25	loc	Stay_At_Home	Bar_Close
O3	1	0.05775	-0.1396	0.2823	0.138
PM25	0.05775	1	0.01509	-0.131	-0.1154
loc	-0.1396	0.01509	1	-0.2268	0.5267
Stay_At_Home	0.2823	-0.131	-0.2268	1	0.3289
Bar_Close	0.138	-0.1154	0.5267	0.3289	1

Multicollinearity Discussion:

We next begin consideration of potential issues of multicollinearity within our models. To do so, we compute the variance inflation factors for each of our predictors in each regression. Note that the notion of the variance inflation factor

Non-zero intercept models: As shown by the Variance Inflation Factor tables, the computed values are consistently below 2.2, indicating low to moderate multicollinearity per the oft-cited threshold values of 5 and 10.

Zero-intercept model: As the estimation of the predictor R^2 values could potentially be negative given the constraint on the intercept term, the Variance Inflation Factor metrics used in the preceding discussion cannot be directly applied with strong interpretation. However, applying the same calculation methods we still see that the Variance Inflation Factors computed are consistently below the threshold value of 5 indicating moderate multicollinearity. Further inspecting a correlation matrix of the predictors we see that the Pearson correlation between the pairs (`O3`,`Stay_At_Home`), (`Bar_Close`, `loc`), and (`Bar_Close`, `Stay_At_Home`) are particularly large in magnitude and are further all positive. This correlation is logical for the pair (`Bar_Close`, `Stay_At_Home`) and provides insight into the positive coefficient for the `Bar_Close` variable. In the correlation for the pair (`Bar_Close`, `loc`) we see a reflection of the extended duration of bar closures in Raleigh relative to Greenville, and so this once again follows logically. Finally, the positive correlation in the pair (`O3`, `Stay_At_Home`) is of some interest and provides evidence against our guiding hypothesis. With respect to discussion of multicollinearity, however, the values do not provide strong evidence for concern, especially given the Variance Inflation Factors calculated for the same variables in the Non-zero intercept models.

```
rmse(fitted(fixed_effect_intercept1), train$newcasesper)
```

```
## [1] 0.0001440828
```

```
rmse(predict(fixed_effect_intercept1,test),test$newcasesper)
```

```
## [1] 0.0001197185
```

```
RMSEm1 <- data.frame(matrix(ncol = 2, nrow = 1))
colnames(RMSEm1) <- c("Training RMSE", "Test RMSE")
RMSEm1[1,1] <- rmse(fitted(fixed_effect_intercept1), train$newcasesper)
RMSEm1[1,2] <- rmse(predict(fixed_effect_intercept1,test),test$newcasesper)

RMSEm1 %>%
  pander(caption = "Fixed-Effects Model with non-zero intercept and 6 parameters")
```

Table 5: Fixed-Effects Model with non-zero intercept and 6 parameters

Training RMSE	Test RMSE
0.0001441	0.0001197

```
rmse(fitted(fixed_effect_intercept2), train$newcasesper)
```

```
## [1] 0.000144213
```

```
rmse(predict(fixed_effect_intercept2,test),test$newcasesper)
```

```
## [1] 0.0001198446
```

```
RMSEm2 <- data.frame(matrix(ncol = 2, nrow = 1))
colnames(RMSEm2) <- c("Training RMSE", "Test RMSE")
RMSEm2[1,1] <- rmse(fitted(fixed_effect_intercept2), train$newcasesper)
RMSEm2[1,2] <- rmse(predict(fixed_effect_intercept2,test),test$newcasesper)

RMSEm2 %>%
  pander(caption = "Fixed-Effects Model with non-zero intercept and 5 parameters")
```

Table 6: Fixed-Effects Model with non-zero intercept and 5 parameters

Training RMSE	Test RMSE
0.0001442	0.0001198

```
rmse(fitted(fixed_effect_ni), train$newcasesper)
```

```
## [1] 0.0001456031
```

```
rmse(predict(fixed_effect_ni,test),test$newcasesper)
```

```
## [1] 0.000120756
```

```
RMSEm3 <- data.frame(matrix(ncol = 2, nrow = 1))
colnames(RMSEm3) <- c("Training RMSE", "Test RMSE")
RMSEm3[1,1] <- rmse(fitted(fixed_effect_ni), train$newcasesper)
RMSEm3[1,2] <- rmse(predict(fixed_effect_ni,test),test$newcasesper)

RMSEm3 %>%
  pander(caption = "Fixed-Effects Model with zero-intercept and 5 parameters")
```

Table 7: Fixed-Effects Model with zero-intercept and 5 parameters

Training RMSE	Test RMSE
0.0001456	0.0001208