Analysis of the Multiple Linear Regression (MLR) model for positive rate of COVID-19 based on the mobility data in Ontario

Siqi Zheng 1005065830

2020-12-21

Contents

Code and data supporting this analysis is available at: https://github.com/siqi-zheng/STA304_FINAL_PROJECT

Abstract

The increasing positive rate of COVID-19 is a major concern in Ontario. The Ontario government encourages people to reduce travel and stay at home as much as possible, while some people question whether such policy is useful. This report therefore discusses whether the number of visits to certain locations (represented by Google Mobility Index) affect the positive rate. The report first explores how Google Mobility Index and the number of direction requests (Apple Mobility Index) react to changes in the social environment, such as first reported death due to COVID-19 and the shutdown policies. Next, the analysis focuses on building a multiple linear regression model for the positive rate using Google Mobility Index based on both p-value/AIC/BIC criterion and practical significance. We conclude by arguing that the number of visits to workplaces and residence is useful for the model and the positive rate is positively associated with these variables.

Key words: COVID-19 Positive Rate, Mobility, Ontario, Multiple Linear Regression, Time Series Data, AIC, BIC

Introduction

COVID-19 pandemic is a huge challenge in Ontario in 2020. In order to stop the spread of COVID-19, Ontario government decided to close public services and businesses and encourage people to work from home (Nielsen, 2020). However, even though the Ontario government has made efforts to reduce the unnecessary travel, the proportion of positive COVID-19 cases seemed to increase until present. Furthermore, the government closed many public locations to avoid physical contact, but the cost is also high because the restriction impacts the economy to a great extent.

Hence one objective of this report is to investigate whether the government's current policies were effective in reducing the positive rate by lowering the travel in Ontario. The visualizations illustrate whether there is a significant change in the number of visits to a location (Google Mobility Index) and/or the number of pull requests (Apple Mobility Index) when certain policies/issues are announced.

Another objective, the main objective, of this report is to build a model that explores the relationship between positive rate of COVID-19 and Google Mobility Index. Specifically, this article will discuss how visits to

different locations contribute to the proportion of Ontario residents who are tested positive for COVID-19 every day using multiple linear regression model (MLR).

An additive model with statistically significant predictors will be produced. Then selection criteria including AIC and BIC will be used, and the analysis will compare the resultant models from AIC/BIC with the first model. Then the report will argue that some variables are not appropriate for a model that predicts positive rate.

Results section will explore the details the final model, the inferences of this dataset along with conclusions are presented in Conclusion section. The Limitations section and Next Steps section will provide a comprehensive discussion from the datasets to the selection of a model, and suggest some possible options for further research.

It is usually hard to establish the causal link between variables in an observational study that utilizes observations as data. In this case, however, we can safely assume that causal relationship exists in the final multiple linear regression model (MLR) i.e. when the number of visits (Google Mobility Index) changes, the positive rate will change. There are two reasons for this. First, in practice, the reduction in number of visits will inevitably lead to decrease in physical contacts with other people and stop the spread of virus. Second, when the model is constructed, the report will omit certain predictors when the predictors have uncertain causal effects on positive rate. The details can be found in the Model section.

Methodology

Data

Links to download the datasets are attached in the References section.

Mobility is the key concept in this analysis, but Google and Apple has a slightly different way to define it. Google Mobility Index (2020) is defined by the percentage change in number of visits to a type of location compared to a baseline value, while Apple (2020) defines Mobility Index as a relative volume of directions requests on Apple Map.

In our analysis, the Apple Dataset will only be used to understand the trends in Exploratory Data Analysis sub-section. The reasons behind will be discussed in the next section. Hence the focus of this analysis will be on the Google mobility dataset and the official dataset for COVID-19 cases in Ontario. In the following paragraphs, a brief introduction of all datasets are provided, but the readers are welcome to acquire more information via the links in the References section.

The Google mobility dataset (2020) helps understand the impact of COVID-19 on the change of number of visits in different locations around the world. In particular, the change of visits in a type of location is calculated by comparing with a baseline value. The baseline is the median value, for the corresponding day of the week, during January 3 to February 6, 2020.

This Google Dataset (2020) contains 7 columns, one date column and six columns for number of visits to six types of locations. Hence each case (row) represents the number of visits to six types of locations on a given date. The dataset contains 0 non-NA cases for each type on a given date. The date range is from 2020-02-17 to 2020-11-24.

The six types of locations in this dataset are Grocery & pharmacy, Parks, Transit (stations), Retail & Recreation, Workplaces and Residential. The types are very self-explanatory maybe except for the difference between grocery and retail in this case. In this dataset, the variable Grocery & pharmacy includes data about mobility trends for markets, food warehouses, food shops and drug stores, while the variable Retail & Recreation includes data about mobility trends for shopping centers (Google LLC., 2020).

Since the Google dataset shows the relative percentage change compared to a baseline value, we need to convert the Mobility Index to normalized absolute term for our model so that we can explain the model in a sensible way. Therefore, the most convenient way is to define the mobility index of the baseline value (x_0) as 100. Consequently, the (normalized) Mobility Index for the day i (x_i) will be:

$$x_i = 100 \text{ (base value)} + r_i \text{ (percentage change/original Mobility Index)} \tag{1}$$

where r_i is the percentage change i.e. the original Mobility Index in the Google Mobility Dataset.

The official dataset from the Ontario government consists of dates and positive rate. However, the corresponding column for the positive rate of COVID-19 in the official dataset from the Ontario government (2020) records the positive rate in the last day. Therefore the column is modified by shifting all numbers one day ahead.

In this dataset, there is one date column and one column for positive rate. In particular, there are 237 non-NA cases for positive rate of COVID-19, and the date range is from 2020-04-18 to 2020-04-18.

We are effectively making inference to the whole population in Ontario using Google Mobility Dataset since we build our model based on Google Dataset and the official COVID-19 dataset from Ontario, so the Google Dataset serves as a frame in our analysis. All data points about Mobility Index collected from Google users is a sample. In our analysis, we need to assume that the sample is representative to the change in mobility of the Ontario population, but this may not be true even though this dataset is the most comprehensive dataset for mobility that is available publicly. Therefore, the limitations will be addressed in the Discussion section.

The Apple Mobility Dataset contains 4 columns, one date column and three columns for different types of direction requests, namely "Driving", "Walking" and "Transit". Hence each case (row) represents the number of direction requests for each type on a given date. The dataset contains 319 non-NA cases for each type given any date. The date range is from 2020-01-14 to 2020-11-27.

Exploratory Data Analysis - Trend Overview

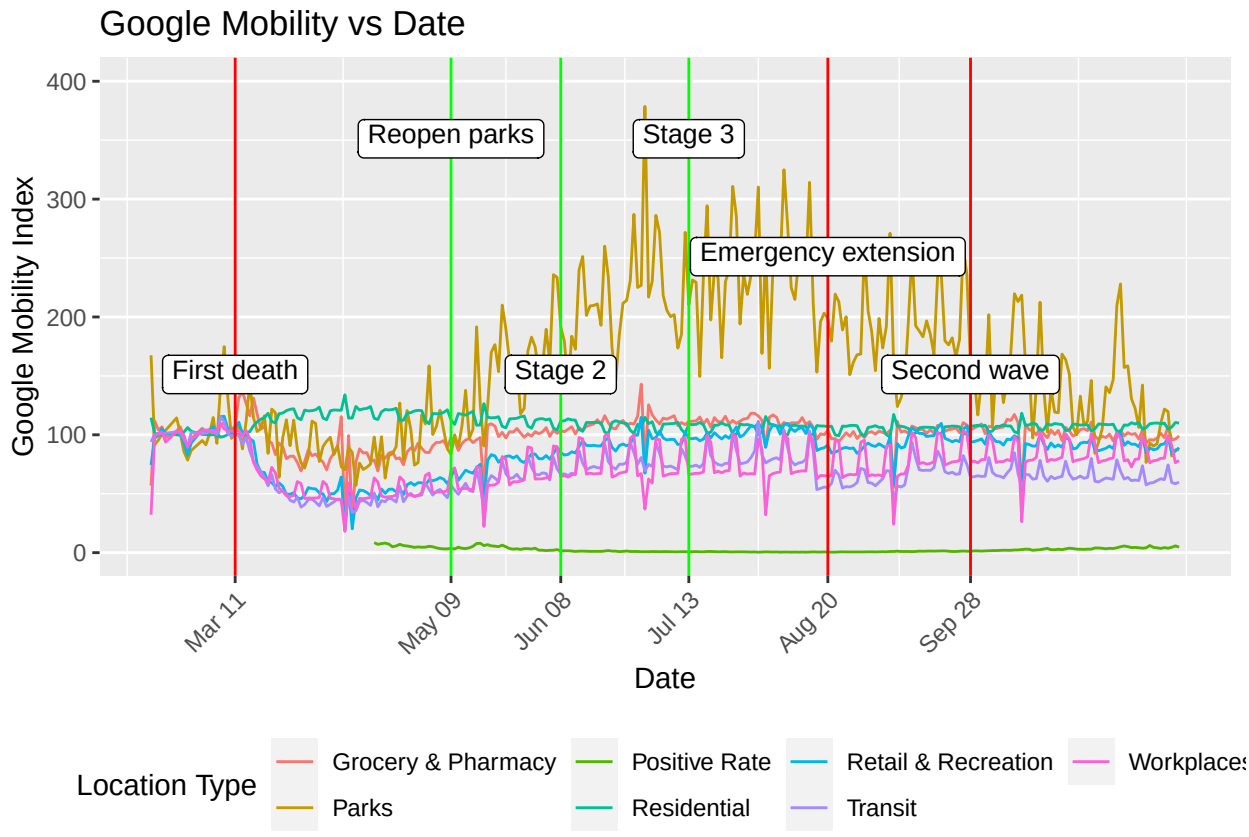

Figure 1. Google Mobility Index vs Date

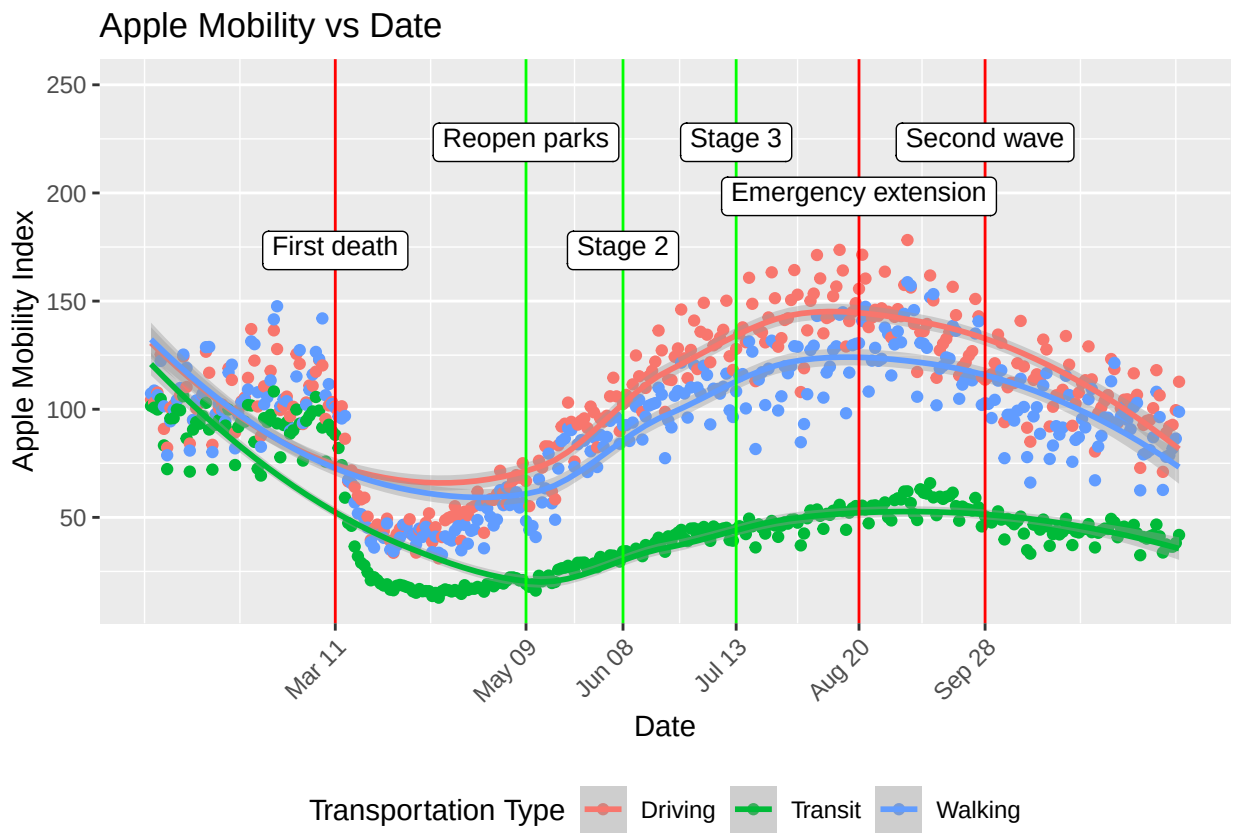

Figure 2. Apple Mobility Index vs Date

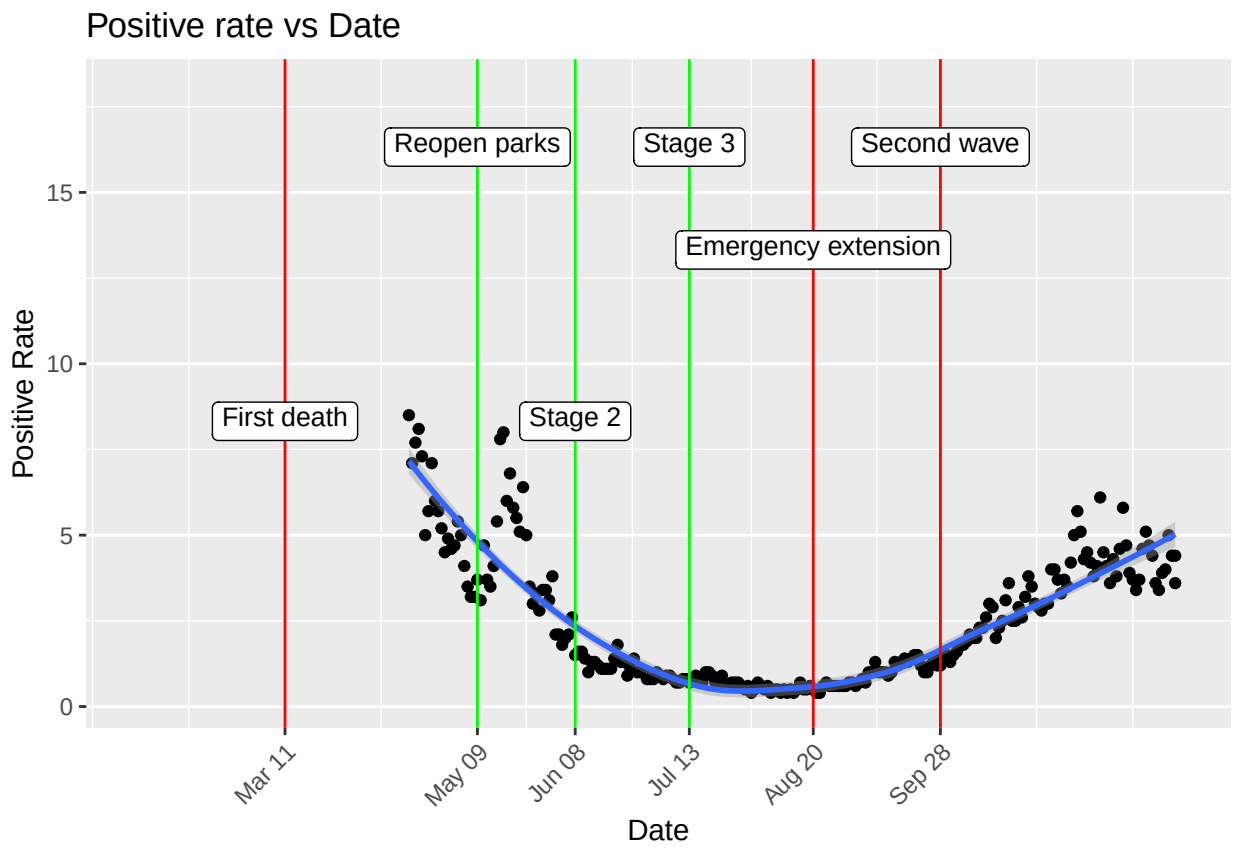

Figure 3. Positive rate of COVID-19 vs Date

The above visualizations demonstrate the trends for mobility with Google Mobility Dataset, Apple Mobility Dataset and the official dataset about COVID-19 cases in Ontario. In all three figures, green lines represent reopening policy while red lines represent serious issues about COVID-19/policy change from the Ontario government. All visualizations in this section are generated with package tidyverse, xts and zoo.

In the following analysis, the important dates will be discussed since the first death due to COVID-19, along with the changes in mobility in Figure 1 and 2 and the change in positive rate. Then, I will provide explanation about why Google Dataset is preferred over Apple Dataset for the model in the next section.

When the first death due to COVID-19 occurred in Ontario on Mar 11 shown by the red line, it is clear that people significantly reduced the number of visits to all locations shown in Figure 1 and the number of pull requests shown in Figure 2. At this moment, the public reacted efficiently the numbers and concrete facts.

The government also announced new policies and closed many public locations for a short period of time. However, the parks were reopened on May 9. Then, the number of visits shown and the number of direction requests on Apple Map started to increase shown on Figure 1 and 2.

Almost at the same time, the positive rate dropped to nearly 0 at the end of July shown on Figure 3. Some cities were allowed to enter stage 2 and stage 3. At stage 2 and stage 3, certain businesses could finally reopen (Nielsen, 2020). As shown by the green lines, stage 2 started on June 8 and stage 3 started on July 13.

Nonetheless, Ford government still extended the emergency pandemic orders until Sept. 22. The Mobility Index for Google and Apple decreased again when the announcement was made. This might be a wise choice given that it aimed to protect residents from virus by reducing possible physical contact. However, the government did not fully anticipate, if not underestimated, the consequences of the reopening policies, so the emergency order extension did not stop the virus at all. The positive rate began to rebound a few days after the announcement and continued to increase until present.

I have reasons to believe that the government should rather be more cautious at the very beginning. This is because if one considered the 14-day incubation period, it was possible that many residents already caught COVID-19 when the restriction was relaxed. They might developed symptoms later and tested positive after the emergency order extension. At the end of the day, the Ontario government had no choice but officially announced that Ontario entered second wave of COVID-19 on 2020-09-28.

The government policies seemed to be effective on the Google Mobility Index of Parks. Based on Figure 1, the number of visits to parks changed significantly when certain issues and/or policies were released. In contrast, the visit to residential was not affected by the CVOID-19 issues and policies. This is trivial given that people would always go home regardless of the restrictions. Nonetheless, after an initial decrease in the visits to the rest of the public locations, the Google Mobility Index for these locations was relatively constant regardless of the restrictions.

If we compare the changes in mobility due to policy change with changes due to the occurrence of the first death, we can conclude that the Mobility Index is affected much by the concrete fact such as first death but not the changes in policy. Therefore, in order to reduce the number of visits/number of direction requests, the government should either modify their policies and/or raise the public's awareness towards COVID-19.

A more rigorous way to understand the trends is to decompose the time series data into multiple components. A time series is a list of number (Mobility Index of a type of visits in Google Mobility Dataset/positive rate of COVID-19), along with some information about what dates those information was recorded. This information can be stored as a ts object in R (Hyndman, 2018). There are many methods to decompose the time series, and one traditional way is to apply additive method on the time series by breaking the time series data y_t into trend $\hat{T}_t$ with MA method, seasonality $\hat{S}_t$ with averaging the detrended values of a weekday and the remainder $\hat{R}_t$. The equation of the model is shown below.

$$y_t = \hat{T}_t + \hat{S}_t + \hat{R}_t \tag{2}$$

If one is interested in the specific trend, one might run the codes in the original .Rmd. The codes decomposed the time series data of Mobility Index using additive method discussed above. The validity of this method in this context is beyond the scope of this report and is not the focus of the analysis, so I will only show a sample of decomposition below using the time series data of Mobility Index of Workplaces. As you can see in Figure 4, the trend for Google Mobility Index of Workplaces is relatively constant after a huge decrease at the beginning.

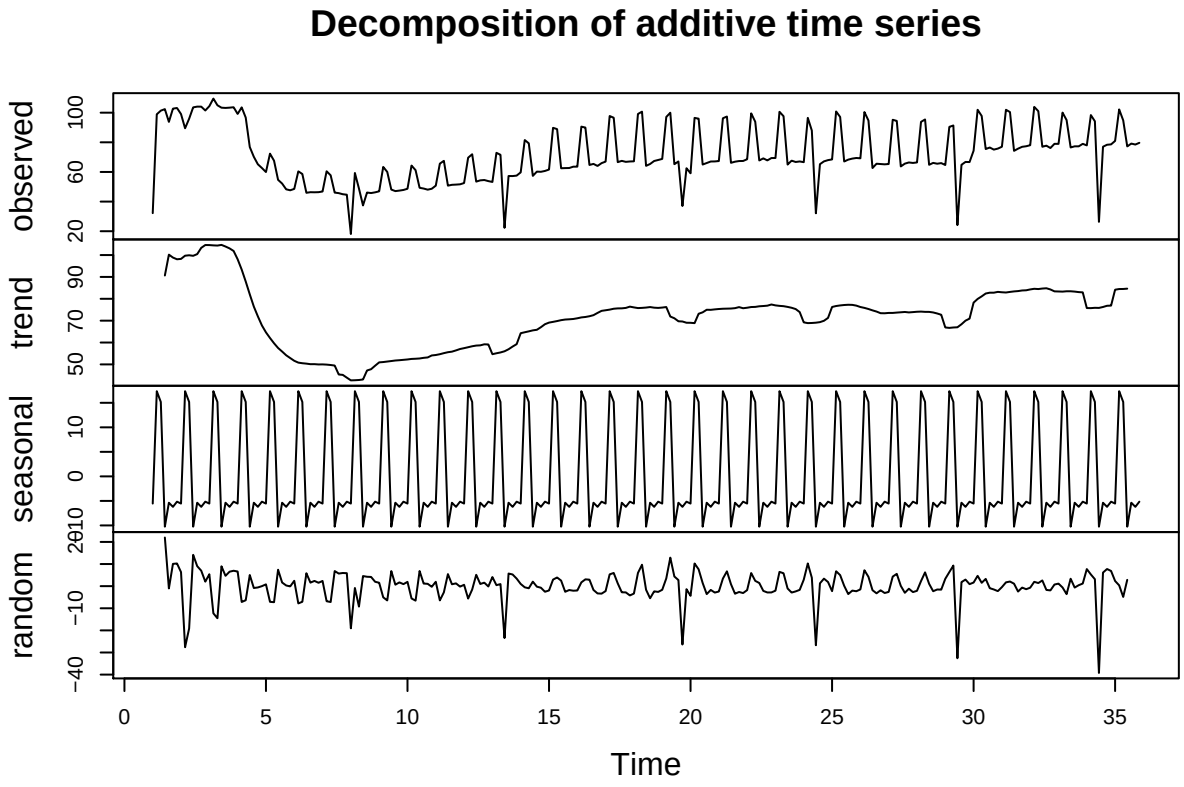

Figure 4. Decompose time series data of Mobility Index of Workplaces with additive method

Finally, I will build up a model for positive rate of COVID-19 based on Google Dataset in the next section. There are numerous reasons why Apple Dataset is not useful for the analysis. First, even though a trend was shown for Apple Mobility Index in Figure 2, the data points spread widely in the Apple Dataset compared with Google Dataset. It would be a difficult task to fully understand the fluctuations in the Apple Dataset.

Furthermore, there is not observed weekly or monthly patterns in the Apple Dataset compared to the Google Dataset. Consequently, if researchers would like to apply time series analysis on the topic of Mobility, Google Dataset is a more appropriate option. Moreover, there are more NA values in this Apple Dataset. Hence, there are fewer available cases to build our model. Moreover, NAs could be important since the reasons behind the missing information can be useful to our analysis.

All these reasons support why Google Dataset is more suitable for our model. Nonetheless, both datasets have similar limitations in this context. These limitations will be further addressed in the Limitation section, but I will focus on the Google Mobility Dataset in the following analysis.

Model

All models are generated with Rstudio and base R. No additional packages are required.

Multiple Linear Regression Model with Significant Predictors

In this section, I will attempt to build a model for predicting the positive rate from Mobility Index. A linear regression model may be useful given that both positive rate and Mobility Index are numerical variables. Apart from this, multiple linear regression model (MLR) is selected for four more reasons. First, linear regression model is commonly used and has standard practice to diagnose whether the model satisfies the assumptions, so the report is more statistically rigorous. Second, multiple linear regression incorporates

different Mobility Index into consideration and allows transformation of the variables to build alternative models. Third, the model can be simple yet useful in explaining the association. Fourth, there are many criteria to select the most appropriate multiple linear regression model in this context, so it is more flexible when we need to compare multiple models.

The full multiple linear regression model if we incorporate all six predictors (Mobility Index of different locations) from the Google Mobility Dataset is:

$$\begin{aligned} y = {} & \beta_0 + \beta_1 Retail\&Recreation \\ & + \beta_2 Grocery\&Pharmacy + \beta_3 Parks \\ & + \beta_4 Transit + \beta_5 Workplaces + \beta_6 Residential \end{aligned} \tag{3}$$

y is the response variable positive rate (the proportion who are tested positive for COVID-19).

β_0 is the y-intercept of this function, which may not have practical meaning since it is very unlikely that the Mobility Index does not change from the baseline value at all.

β_1 is the coefficient of the variable (Mobility Index of) Retail & Recreation.

β_2 is the coefficient of the variable (Mobility Index of) Grocery & Pharmacy.

β_3 is the coefficient of the variable (Mobility Index of) Parks.

β_4 is the coefficient of the variable (Mobility Index of) Transit.

β_5 is the coefficient of the variable (Mobility Index of) Workplaces.

β_6 is the coefficient of the variable (Mobility Index of) Residential.

If we use least squares method (LS method) to fit a full multiple linear regression model, we can obtain the estimates of the coefficients, the corresponding p-values of t-test for individual variables and the F-statistic for the whole model. Indeed, the F-statistic of the full model is significant at 5% level, so we investigate the statistical significance of each variable via t-test. After selecting the significant predictors from the full model at 5% level, a model with only significant predictors is obtained:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 Retail\&Recreation + \hat{\beta}_5 Workplaces + \hat{\beta}_6 Residential \tag{4}$$

The number of variables is reduced to 3, and the detailed explanation of the estimates of the coefficients will be provided in the Results section.

Alternative Models – AIC/BIC Criterion

Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Sheather, 2009) are common criteria to address a model. Both criteria are likelihood-based criteria that balance goodness of fit and a penalty for model complexity. Also, a model with smaller AIC/BIC is considered better in general. The difference is that AIC may overfit a model given that the penalty of AIC for model complexity is relatively weak, while BIC penalizes complex model more heavily, favoring simpler models than AIC (Sheather, 2009).

Backward elimination is used with AIC/BIC. This elimination starts with all six predictors (i.e. full model) and removes the predictor with the largest p-value each time to obtain a smaller AIC/BIC. Backward-elimination method is chosen because the initial estimate of the standard error of the model will usually be smaller since all predictors are used at the beginning (Sheather, 2009).

As the number of predictors of the full model is not large i.e. the model is not complex, AIC and BIC are expected to produce a similar result. Unsurprisingly, the linear model obtained by backward-elimination AIC method is the same as the model from backward-elimination BIC method:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 Retail\&Recreation + \widehat{\beta}_3 Parks + \widehat{\beta}_5 Workplaces + \widehat{\beta}_6 Residential \tag{5}$$

The consequence of different criteria (AIC&BIC vs P-value) for the multiple linear regression model is that the selection of variables are different. In particular, we need to justify whether the additional variable from AIC/BIC, Mobility Index of Parks, should be incorporated in the final model, because the AIC/BIC model with this additional variable attains the lowest AIC/BIC value. It is hard to tell whether we should strictly follow the 0.05 significance level to reject the additional variable or not, so the selection should consider the practical significance of the additional variable in the model as well.

There are at least three practical reasons why the third variable is not necessary for the multiple linear regression model. First, the variable is not statistically significant at 0.05 significance level using t-test. Even though this should not be the deciding factor of whether we should discard the Mobility Index of Parks, it nonetheless provides a mathematical justification to reject this variable. Second, visits to a park do not really contribute to more COVID-19 cases, since a park is usually an open space with sunlight and air circulation. Virus is hard to survive and spread when it is exposed to such environment. Moreover, people in the parks can easily follow the social distancing requirement and avoid physical contact. Hence it is unlikely for people to catch COVID-19 because of a visit to a park.

If we remove the additional Mobility Index of Parks, the results from AIC/BIC are consistent with the model with significant predictors. Not only the choice of variables is consistent, but also the models with different criteria show consistency in the model significance and predictor significance. In other words, the p-values of the global F-test for AIC/BIC model and the predictors in AIC/BIC model are smaller than 0.05, so the global F-test and incorporated variables are statistically significant in AIC/BIC model.

The other predictor Retail & Recreation should also be omitted even if it is significant. That is because we are only interested in how mobility affects the positive rate. This variable, however, is largely affected by the positive rate. When the positive rate is high, the government first closed Recreation centers including gyms and shopping malls because they were less essential compared to other public services (Nielsen, 2020). As a result, the number of visits was reduced due to increase in positive rate and change in restrictions. That is, the variable does not really have any causal effect on the positive rate. Hence this variable is not appropriate for a model that aims to predict the positive rate.

Hence the final model will only contain 2 significant predictors. In the next sub-section, we will evaluate the main four assumptions of the final MLR model by diagnostic plots.

Diagnostic Plots and Interpretation for the final Model

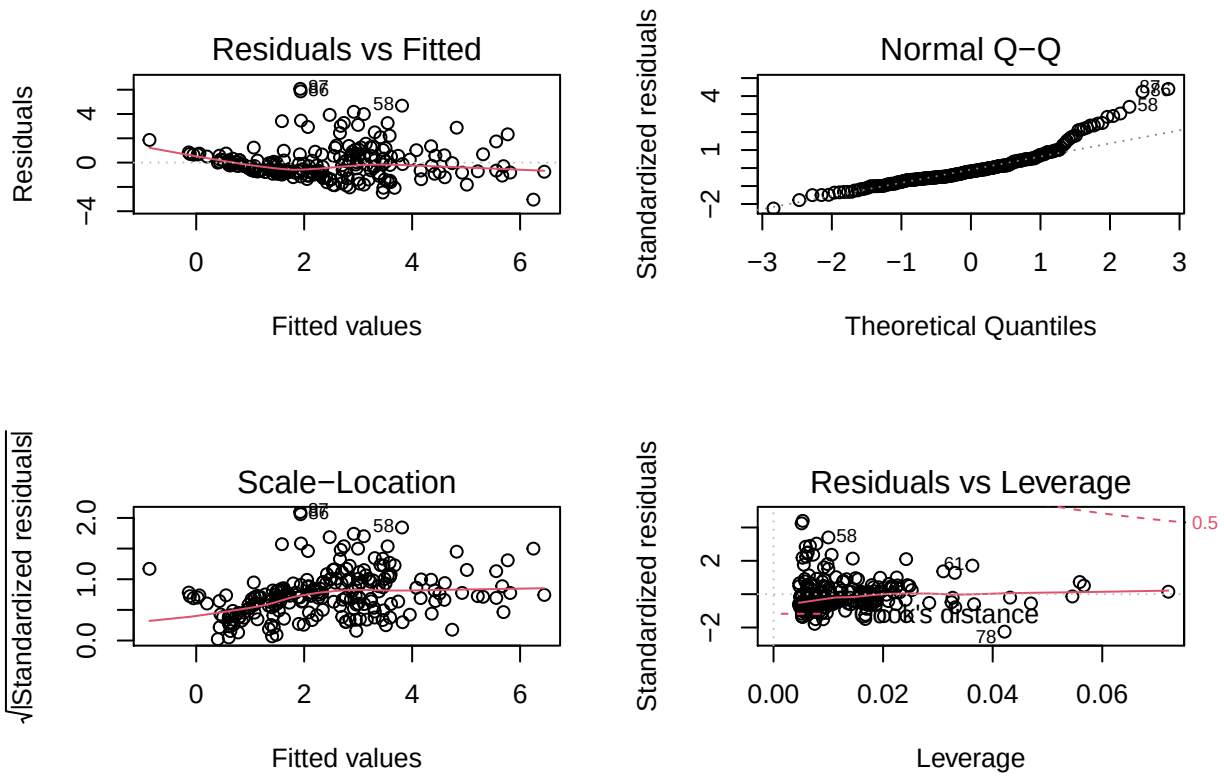

Figure 5. Diagnostic plots for the final model

The first plot Residuals vs Fitted in Figure 5 shows the association between residuals and fitted values of the model. The overall trend is close to a horizontal line i.e. no association between these two variables. From this graph, we may say **the first assumption of linear relationship** is mainly satisfied as the trend is relatively constant.

The second plot Normal Q-Q in Figure 5 can be used to examine whether **the errors are normally distributed**. If the standardized residuals are normally distributed, the points should be close to the theoretical quantile-quantile line for normal distribution (straight dotted line in Figure 5). Since the points on the tail deviate from the theoretical q-q line, we can claim that the second normal error MLR assumption is violated to a small extent. A remedy may be needed and discussed in the Limitation section.

The third graph is Scale-Location plot in Figure 5, which shows the relationship between $\sqrt{\text{Standardized Residual}}$ vs fitted values. It can be used to check **the third assumption of constant variance**. The assumption is satisfied if no pattern is found (horizontal line) and the points are equally spread. Since there is only a slightly increasing trend in this graph, the assumption of constant variance is mainly satisfied.

The fourth graph is Residuals vs Leverage. This graph can be used to identify influential points. There are not leverage points or outliers in this case using the Cook Distance Criterion.

The last assumption is that **the errors are uncorrelated**. This assumption can only be checked by the design of the study, which includes the data collection process that is not publicly available. As a result, we cannot determine whether the assumption is satisfied unless Google is willing to share the insights of the dataset.

Results

The policies may not be effective in decreasing the number of visits based on the Exploratory Data Analysis section. Facts are more powerful than the fines. Given that models are built upon facts, it is worth exploring how the positive rate correlates with the number of visits to certain places.

As we discuss in the last section, since the F-statistic of the final model is 105.4 and the p-value of the global F-test on this model is smaller than 0.05, the global F-test is statistically significant. Therefore, we can further discuss the p-values of the t-test in the final model (partial F-test on each variable).

The final MLR model includes two significant variables from the full MLR model and these variables Workplaces (β_5) and Residential (β_6) are significant as expected, so we can provide explanation for the coefficients for this model.

$$\hat{\beta}_5 = 0.1199 \tag{6}$$

which means when Mobility Index of Workplaces increases by 1 unit (increases by 1% compared to the baseline value), the positive rate, on average, changes by 0.1199% keeping all other predictors constant.

$$\hat{\beta}_6 = 0.5143 \tag{7}$$

which means when Mobility Index of Residential increases by 1 unit (increases by 1% compared to the baseline value), the positive rate, on average, changes by 0.5143% keeping all other predictors constant.

The positive rate is positively associated with the Mobility Index of Workplaces and Residential i.e. the number of visits to workplaces and residence as expected. The more visits people make on a given day, the higher the positive rate of COVID-19 will be.

Discussion

Summary

This article explores the impact of COVID-19 and related policies/social issues on mobility during COVID-19 using line plots and states that the policies may not be effective in reducing the number of visits to most locations. In contrast, the mobility decreases significantly when the first death due to COVID-19 was released.

Furthermore, based on the final multiple linear regression model based on p-value/AIC/BIC criterion, the analysis produces a final model with two variables, the Mobility Index of Workplaces and Residential. To reduce the spread of COVID-19, we need to change the policies and decrease the number of visits to workplaces and residence.

Conclusion

Our first conclusion from the Exploratory Data Analysis is that policies are not very useful in terms of discouraging unnecessary travel, so the government have to reconsider the strategies against COVID-19. Also, the public is more reactive to the solid facts, such as the first death due to COVID-19. Therefore, the government should present more concrete facts to the public and raise people's awareness of COVID-19.

From our final model, we conclude that the positive rate of COVID-19 is positively associated with number of visits to workplaces and residence. When Mobility Index of Workplaces increases by 1 unit, the positive rate, on average, changes by 0.1199% keeping all other predictors constant. When Mobility Index of Residential increases by 1 unit, the positive rate, on average, changes by 0.5143% keeping all other predictors constant. If we assume the existence of causal relationship as described at the beginning of the article, then residents should work from home and stay at home as much as possible.

Hence we urge the government to rethink about the current policies. The public has rights to know how dangerous COVID-19 can be and more information to make choices. Specifically, the government can incorporate more facts to appeal to people's emotion and logic.

The residents in Ontario, on the other hand, have to find ways to reduce visits to workplaces. Indeed, there may be communication cost for software and cloud services if workers have to stay home, but working from home also saves time and transportation costs. People should also reduce the number of visits to other people's residence, and may want to communicate via internet instead of face-to-face. I acknowledge that this suggestion may not be the best option especially when Christmas is coming, but we all know that this pandemic cannot be stopped without collaboration among Ontario residents.

Finally, I also want to argue that the protection & prevention of COVID-19 during travel is still not enough given that the positive rate is positively associated with the number of visits to workplaces and residence. Even though many services require wearing masks, people have to be aware that contact of contaminated surface may cause the transmission of COVID-19 as well (Ontario government, n.d.). Therefore, if one has to travel to work or travel back home, one must not only wear masks but also reduce physical contact with human beings and clean the surface of objects as much as possible.

Limitations

Dataset In our analysis, Mobility Index is defined as the number of visits or number of request to a direction in Ontario, but the concept of mobility and mobility index may be redefined depending on the research questions. Therefore, for researchers who are interested in the general social behaviors of Ontario residents under COVID-19, the datasets many not be useful.

One limitation of the Google Mobility dataset is that there is a time lag i.e. the data will represent the changes in mobility around 2 to 3 days ago. This is because it takes time to update the results. Therefore, the dataset may be more appropriate to show trends over months (Google LLC., 2020).

This limitation will unavoidably affect the values of estimates of the parameters in the MLR model, but it may not have a huge impact on whether there is a correlation between the positive rate and the Mobility Index. Hence the model itself is still informative given the time lag in the predictors variables.

Furthermore, Google claims that it updates the way to collect the data this year to ensure the consistency of the data (Google LLC., 2020). The starting point is good, but the change in the method of data collection may substantially affect the data themselves. If data are altered in an unexpected way and the new method removes some important characteristics, then it is very difficult to tell whether the dataset can even be used in academia.

Even though Google Map is popular, it only represents people who use smart phones or other digital devices. These users may share some common identities, so we cannot conclude how representative the Google Mobility Dataset is due to this concern.

The Apple Mobility dataset is even more problematic. First, compared to Google Map users, the Apple Map users are even less representative of the Ontario population, even without learning about the demographic information. This is because the data can only represent Apple Map users who own an Apple device i.e. who may be presumably from a certain socioeconomic class. However, this does not mean the Apple Mobility dataset is completely useless. In other words, the absolute values may not have practical meanings, but the visualization indeed demonstrates how the trends vary with time.

Model Selection Even though the report explores the use of multiple linear regression model for positive rate using Mobility Index, it is still possible that there does not exist any linear relationship between the variables. In the Next Step section, I will recommend some solutions for the issue of model usage.

Furthermore, the model is built using all available data given that there are not many data points in the original Google Mobility Dataset. Therefore there are not testing datasets. The drawback is that it may be hard to determine how useful the multiple linear regression model will be in practice.

Among multiple linear regression models based on different criteria, the selection of the final model is built upon the context of COVID-19 pandemic and the daily visits of the locations. This may be subjective because the author has limited knowledge in epidemiology and social science. As a result, other scholars may have diverse opinions on which criterion should be utilized.

In particular, the omission of variables can be subjective as well. One consequence of the different criteria is that the model will contain additional variable. As discussed in the Model section, the additional variables are not included since they have little practical meaning in my opinion. Nevertheless, experts in other disciplines may object this idea and propose a new criterion for variable selection.

Weaknesses & Next Steps

Data In the sub-section Dataset under the Limitations section, the article discusses the main limitations of the datasets and the model selection, so this section will further examines the possible solutions to these issues.

To fully understand the definition of mobility, researchers may explore the changes in frequency of visits of individuals to a location instead of the total change during a period. A case study may be used to reveal the impact of changes in mobility on human's daily life. Furthermore, these researches may need a more comprehensive and formal definition of mobility with other available data.

Speaking of the Google Mobility dataset, the time-lag issue has two solutions. First, scholars can contact Google directly for assistance, and try to understand how we can reorganize the dataset so that the Mobility Index corresponds to the given date. Second, the scholars may collaborate with the Ontario government for the data, and generate a new dataset. However, the time cost and the money cost will be very high for such a project, and this project may require a team of experts from multiple areas.

In order to understand how the method of data generation evolves over time, data scientists should communicate with Google team as suggested above. Nonetheless, Google may not be able to share the information due to privacy reasons, so it is always better to have first-hand data in this case.

Model If MLR is still believed the most appropriate option, the Box-cox transformation may be performed on the final model since the final model violates the assumption of normality. However, the validity of the model does not improve much after transformation based on the diagnostic plots and adjusted R-squared value (see the original Rmd file for details), so we may need to reconsider the choice of model.

Multiple linear regression model is used for the analysis, but statisticians can be more creative in this respect. Because of the author's limited knowledge in the inter-disciplinary research of epidemiology and statistics, there may be many more useful models that do not require linear relationship between variables. Specifically, data scientists should consider other models, for example, non-parametric models, to investigate the association between positive rate and Mobility Index.

Scholars should also establish a more consistent standard for variable selection/model selection in the inter-disciplinary research of epidemiology and statistics. This requires more empirical researches in this discipline. Furthermore, researchers can test the model on other data in different regions and report how useful the model is in explaining the relationship.

References

Apple Inc. (2020, November 24). *Mobility Trends Reports.* https://www.apple.com/covid19/mobility/

Google LLC. (2020, November 24). *Google COVID-19 Community Mobility Reports.* https://www.google.com/covid19/mobility/

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice.* (2nd ed.) OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on December 19th, 2020.

Nielsen, K. (2020, December 8). *A timeline of the novel coronavirus in Ontario.* Globalnews.ca. https://globalnews.ca/news/6859636/ontario-coronavirus-timeline/

Ontario government. (n.d.). *COVID-19 response framework: keeping Ontario safe and open.* Ontario government. https://data.ontario.ca/dataset/status-of-covid-19-cases-in-ontario

Ontario government. (2020, November 24). *Status of COVID-19 cases in Ontario.* https://www.ontario.ca/page/covid-19-response-framework-keeping-ontario-safe-and-open

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Ryan, J. A., & Ulrich, M. J. (2020). *xts: eXtensible Time Series.* R package version 0.12.1. https://CRAN.R-project.org/package=xts

Sheather, S.J. (2009). *A Modern Approach to Regression with R.* Springer. DOI: 10.1007/978-0-387-09608-7

Wickham H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'.* R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Zeileis, A., & Grothendieck, G. (2005). *zoo: S3 Infrastructure for Regular and Irregular Time Series.* Journal of Statistical Software, 14(6), 1-27. DOI: 10.18637/jss.v014.i06