How does public behavior affect the spread of COVID-19?

Quang Vuong

Contents

1 Introduction

We wish to investigate how public behavior affects the spread of COVID-19 within a community. In particular, our goal is to identify which aspects of public behavior, such as visiting restaurants, visiting bars and using public transit, predict case counts. Since many other factors including population demographics, population density and policy decisions also affect the spread of COVID-19, it is best to focus on datasets for a sufficiently small geographical area (like a city or a county) so that the confounding factors unrelated to the study are controlled.

The present dataset, collected from the `{covidcast}` API, contains daily COVID-19 case counts for 88 days from June 5th, 2021 to September 5th, 2021, along with several indicators of public behavior, in Manhanttan, New York. The indicators of public behavior are:

- `distancing` = Percentage of survey respondents reporting that people maintained a distance of at least 6ft (%)
- `public_transit` = Percentage of survey respondents reporting that they used public transit in the last day (%)
- `worked_outside` = Percentage of survey respondents who was indoors (excluding home) in the last day (%)
- `large_events` = Percentage of survey respondents who attended a crowded event in the last day (%)
- `mask_prop` = Percentage of survey respondents who mostly wore a mask outside in the last week (%)
- `other_mask_prop` = Percentage of survey respondents saying that other people mostly wore a mask outside (%)
- `bar_visit` = Number of bar visits per 100000 people (visits per 100000 people)
- `resto_visit` = Number of restaurant visits per 100000 people (visits per 100000 people)

Most of these indicators are obtained from Facebook surveys. Taking into account the issues of survey data such as response bias and subjectivity of answers, it is extremely unlikely that the variables in the dataset track their true respective quantities. Instead, we opt to interpret these indicators as proxies of public behavior, which are potentially useful for prediction.

Now we will summarize our two objectives for this study:

1. Identify which of the above indicators are important in predicting COVID-19 case counts.

2. Identify a model which can be used to predict future COVID-19 case counts.

Tupper et al. analyzed a model of COVID-19 transmission based on the specifics of social behavior such as reducing transmission rates through masks, social distancing and "bubbling," i.e. limiting social contact [1]. They have concluded that distancing is the most powerful method to reduce transmission, while the effects of masking and bubbling are more situational but still significant. Therefore, based on this study, we would directly expect that `distancing`, `mask_prop` and `other_mask_prop` are important variables that predict lower COVID-19 case counts when at higher levels. It is feasible that the other predictors are important in predicting higher COVID-19 case counts when at higher levels as well, since an argument can be made that they do not pertain to bubbling. We then hypothesize that all predictors will appear in the final selected model.

2 Exploratory data analysis

It is the most instructive to first look at how predictors are related to case counts and each other. We plot all predictors against the number of cases in Figure 1. `distancing`, `bar_visit` and `large_events` appear to have two clusters with different mean case counts. On the other hand, `mask_prop`, `other_mask_prop`, `resto_visit` and `worked_outside` appear to have two trend lines. `public_transit` also looks like there are two trend lines, but both are quite flat. This strongly suggests that there are two clusters within the data that exhibit different relationships between case counts and public behavior.

To continue the examination of the dataset, we will now look at the covariance matrix of the predictors.

```
##      distancing    large_events other_mask_prop
##       1.0000000      -0.7612822       0.7463194
##  bar_visit  mask_prop
##  1.0000000 -0.6964321
##      distancing    large_events other_mask_prop
##      -0.7612822       1.0000000      -0.6218936
##       bar_visit       mask_prop other_mask_prop
##      -0.6964321       1.0000000       0.6837441
##      distancing    large_events       mask_prop other_mask_prop
##       0.7463194      -0.6218936       0.6837441       1.0000000
## [1] 1
## [1] 1
## [1] 1
```

There is some substantial correlation between some predictors. In particular, `bar_visit`, `mask_prop`, `large_events` and `other_mask_prop` seem correlated with each other. Of these predictors, `large_events` appear to have two clusters with different mean case counts, while the remaining predictors seem to have two trend lines against case counts.

The most striking observation made so far is that there might be two clusters within the data that shows different relations between case counts and public behavior indicators, so we will now try to identify how the clusters are split. To do this, we attempt to fit a regression tree and look at the split at the root. Please note that we are only using this tree to expedite the exploration rather than seriously considering as a model. The fitted tree is plotted in Figure 2. Initial inspection of this tree shows that there is clustering by date. After inspecting the plots of the predictors against `cases` within each cluster identified by the treem, it appears that splitting by date reduces the clustering behavior so that trends are now much clearer.

Now, we will outline potential approaches to the analysis of this dataset. As mentioned above, we will analyze the both the original dataset and the dataset with time as an encoded categorical variable instead. Firstly, a full linear model will be fitted and interpreted from the inferential point of view to answer the first objective. Secondly, this model will be compared against ones obtained via a regularization approach and a variable selection approach. As interpretation is one of the goals of the analysis, we opt to avoid non-parametric approaches.

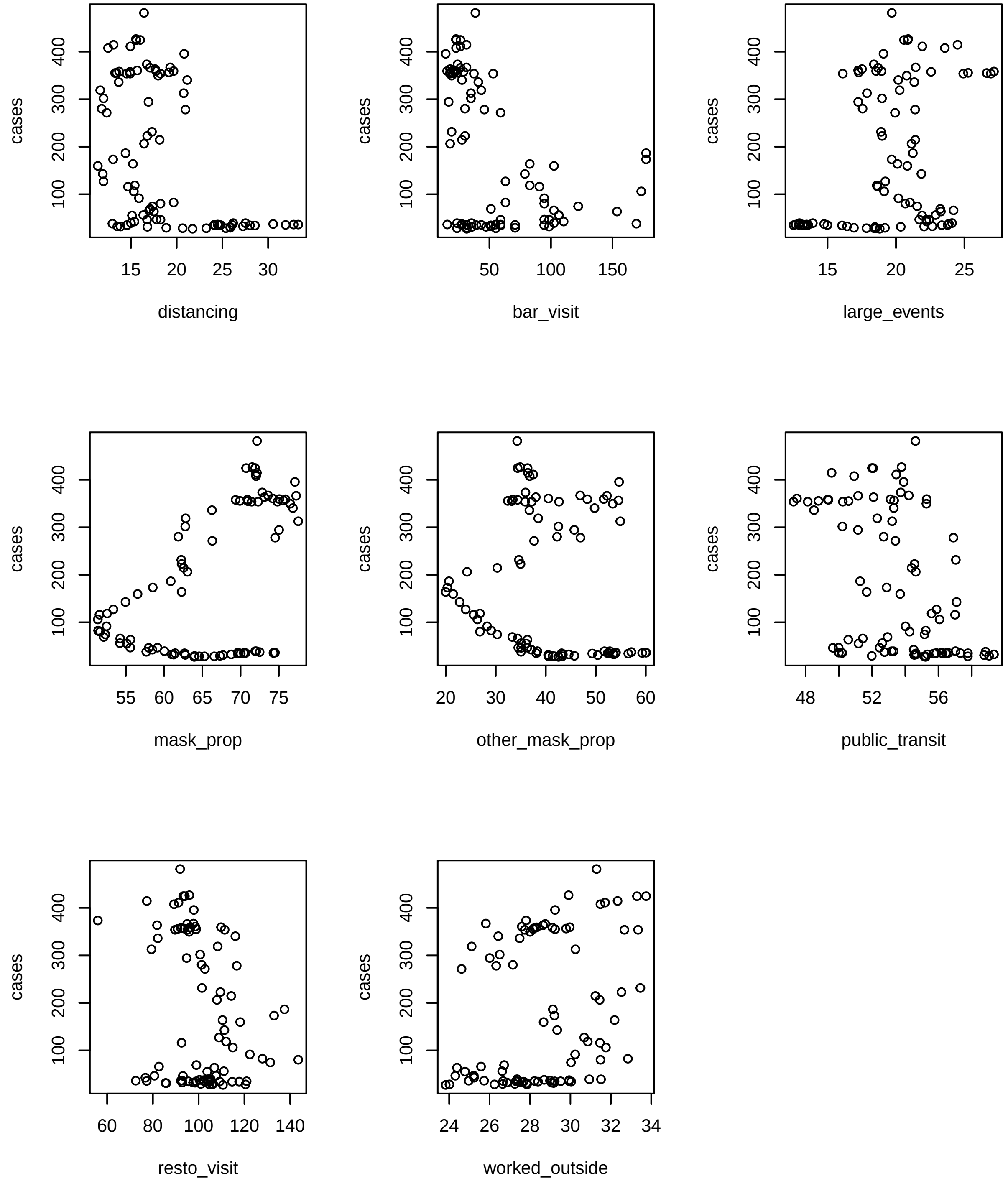

Figure 1: Scatterplots of predictors against case counts.

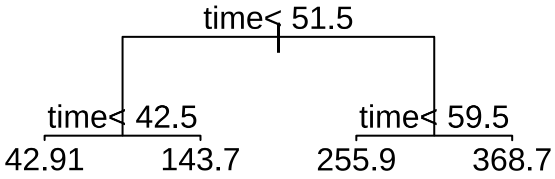

Figure 2: Regression tree to determine clusters in data.

3 Analysis and results

Let Y be a random variable that represents `cases` on a particular day, and let $x_1, ..., x_8$ denote the same for `distancing`, `bar_visit`,`large_events`,`mask_prop`,`other_mask_prop`,`public_transit`, `resto_visit`, and `worked_outside`. We posit the model

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i x_i + \varepsilon$$

where ε is normally distributed with mean 0 and standard deviation σ^2. We examine this model first because it is the simplest. The output of a least squares estimation procedure on the data with this model is as follows. The p-values are those calculated when testing the hypotheses $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ for each β_i included the model, using the t-statistic as the test statistic.

Table 1: Summary of full model, no time split

	Coefficients	SE	p
(Intercept)	-150.6229769	232.7747337	0.5194585
distancing	-16.2220120	2.4130566	0.0000000
bar_visit	-0.5391711	0.2423384	0.0289428
large_events	-1.6959917	3.0637467	0.5814380
mask_prop	15.4685690	1.6681405	0.0000000
other_mask_prop	-4.2513587	1.4810277	0.0052566
public_transit	-6.5669228	2.7934747	0.0212269
resto_visit	0.3100096	0.5481097	0.5732703
worked_outside	5.7810080	3.2073186	0.0752898

At the 5% significant level, only `distancing`, `bar_visit`, `mask_prop`, `other_mask_prop` and `public_transit` have an association with `cases`. Now, the existence of the time split identified in the Exploratory data analysis section informs us to posit the model, with the same notation,

$$Y = \beta_0 + \beta_1 t + \sum_{i=1}^{8} \beta_{i+1} x_i + \sum_{i=1}^{8} \beta_{i+9} t x_i + \varepsilon$$

where ε is as before and $t = 0$ if the observation is on or before July 25th, 2021 and 1 otherwise. Hence, the time split is encoded as a categorical variable to identify observations as before or after the time split. The output of a least squares estimation procedure on the data for this model is shown below. The p-values are those calculated when testing the hypotheses $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ for each β_i included in the model, using the t-statistic as the test statistic.

Table 2: Summary of full model with time split

	Coefficients	SE	p
(Intercept)	259.2717053	167.3746776	0.1258793
tsTRUE	-940.7630971	265.8381582	0.0007187
distancing	-0.1731718	1.8853138	0.9270772
bar_visit	0.1788521	0.1333859	0.1843002
large_events	-5.0799127	2.5062736	0.0464852
mask_prop	2.3038642	1.3195041	0.0851969
other_mask_prop	-4.7683338	0.9678219	0.0000054
public_transit	-2.0700962	1.7732226	0.2470019
resto_visit	-0.0927630	0.3318225	0.7806424
worked_outside	2.0770343	2.2961760	0.3688000
tsTRUE:distancing	-16.1734277	3.8239986	0.0000697
tsTRUE:bar_visit	0.0269364	0.5366448	0.9601106
tsTRUE:large_events	2.9867990	3.4811451	0.3938244
tsTRUE:mask_prop	11.9196511	2.3665376	0.0000035
tsTRUE:other_mask_prop	3.2141728	1.4618083	0.0312015
tsTRUE:public_transit	7.1590850	3.1994565	0.0284324
tsTRUE:resto_visit	-0.4443143	0.5990812	0.4607748
tsTRUE:worked_outside	3.6065949	3.1884057	0.2618486

At the 5% significant level, only `large_events` and `other_mask_prop` have significant relationships with `cases` at all times, while `distancing`, `mask_prop` and `public_transit` are only significantly related to `cases` after the split point. Even though it appears that the main effect of the latter variables are insignificant, t is used to express the clusters in the data where there might be different relationships between `cases` and the predictors, so the conclusions made are sound.

Currently, for the first objective of the study, we have two competing models which provide different answers to the question of which predictors are significantly associated with `cases`. To decide which models to adopt conclusions from, we will check if the error assumptions hold for the two full models by looking at their residual and QQ plots.

The full model that ignores the time split has a residual plot that has a very apparent line for observations with lower predicted cases, but it has a QQ plot which suggests only a small violation of the normal errors assumption. The full model that recognizes the time split has a more patternless residual plot, but there seems to be some heterocedasticity as well as a line on the left. The QQ plot for this model suggests a stronger violation of the normal errors assumption than the first model, but it is still minor. Therefore, we prefer the model that recognizes the time split for the first objective.

So far, we have pursued the first objective from an inferential standpoint, which slightly deviates from the objective as stated. We will now assume a predictive standpoint and compare models by their estimated mean squared prediction error. We will estimate this quantity with a leave-one-out cross-validation procedure that computes mean squared errors because it is flexible enough to adapt to the other model-fitting approaches

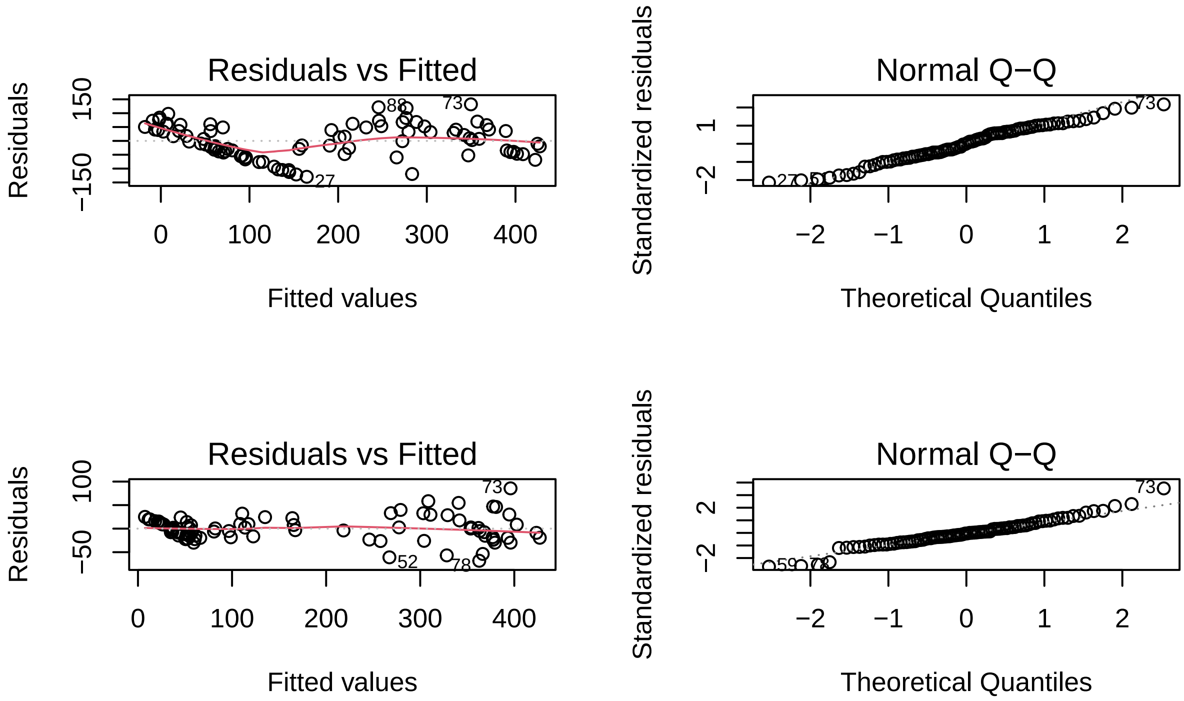

Figure 3: Diagnostic plots for full model without time split (top) and with time split (bottom)

that will be used in this study. The CV scores of the previous two full models are as follows. It is clear that the model that recognizes the time split performs much better.

```
## [1] "CV score of full model, no time split"

## [1] 4385.056

## [1] "CV standard error of full model, no time split"

## [1] 536.258

## [1] "CV score of full model with time split"

## [1] 1151.979

## [1] "CV standard deviation of full model with time split"

## [1] 241.5521
```

Next, we will attempt regularization approaches, owing to the correlation between predictors found in the Exploratory data analysis section. The two models are fitted again using LASSO. The first of them does not split recognize the time split, while the second does. These models have larger CV scores than the unregularized models, indicating that it is likely that all social behavior indicators are important in predicting case counts and informing that a ridge regression procedure might be helpful. However, the latter point does not seem to be the case. The outputs of the R code that fits these models are shown below.

```
## [1] "CV score and standard deviation of LASSO model, no time split"

##      s35
## 4298.099

##      s35
## 570.7778
```

```
## [1] "CV score and standard deviation of LASSO model with time split"
```

```
##      s70
## 1780.747
```

```
##      s70
## 358.5716
```

```
## [1] "CV score and standard deviation of ridge model, no time split"
```

```
##      s99
## 4423.339
```

```
##      s99
## 594.0015
```

```
## [1] "CV score and standard deviation of ridge model with time split"
```

```
##      s99
## 1887.076
```

```
##      s99
## 322.8658
```

Since both regularization approaches performed worse than the full model, it is concluded that they involved too much bias in the present setting. We still would like to see if the model can be simplified, so we now attempt stepwise variable selection on the full model that recognizes the time split. The function used, `stepAIC`, uses a likelihood-based criterion to add and remove predictors in steps. It is hoped that the properties of the likelihood make this stepwise selection procedure introduce less bias to the model, but we are not sure. The results of the procedure are shown below.

Table 3: Summary of stepwise selected model with time split

	Coefficients	SE	p
(Intercept)	114.7555881	108.8918523	0.2952909
tsTRUE	-557.2660137	133.3412813	0.0000774
other_mask_prop	-4.9958180	0.8780735	0.0000002
mask_prop	2.5151718	1.2651135	0.0504021
large_events	-4.1309769	1.6229131	0.0129407
distancing	0.6664742	1.5748791	0.6733513
worked_outside	1.4021543	1.9601688	0.4766012
bar_visit	0.2075449	0.1247990	0.1004247
tsTRUE:mask_prop	11.4270705	1.9705581	0.0000001
tsTRUE:distancing	-15.0214130	2.7127741	0.0000004
tsTRUE:other_mask_prop	3.2086727	1.3549165	0.0204193
tsTRUE:worked_outside	4.7663218	2.9677239	0.1124096

The stepwise selection procedure ended up selecting a model which omits `resto_visit` and `public_transit` as well as interaction terms of the time split with `bar_visit` and `large_events`. Notably, the time split is kept. To correctly assess the quality of this model, we must perform model fitting and stepwise selection from the beginning on each training set, which is accounted for in the calculations. The CV score of the stepwise-selected procedure is as below.

```
## [1] "CV score and standard deviation of stepwise-selected model with time split"
```

```
## [1] 1017.095
## [1] 195.0475
```

We can see that the stepwise selection procedure gives a model with a CV score that is slightly better than the full model that recognizes the time split, so the preferred model is stepwise selected one from a predictive standpoint.

4 Discussion

The predictive standpoint ignores concerns about hypothesis testing and focuses on which predictors appear in the selected model instead, so the stepwise selected model concludes that `other_mask_prop`, `mask_prop`, `large_events`, `distancing`, `worked_outside` and `bar_visit` are important predictors of COVID-19 case counts, where the relationships of `mask_prop`, `other_mask_prop`, `distancing` and `worked_outside` with `cases` change after July 25th, 2021. This is a superset of the variables concluded to be important from the inferential standpoint in the Analysis and results section. Since the first objective prioritizes the prediction of COVID-19 case counts, we will adopt the conclusion of the stepwise selected model instead of the full one; moreover, the inferential conclusion suffers from multiple comparison issues. To answer the second objective, we give the following estimate of the regression function of `cases`. Let Y be a random variable that represents case counts on a particular day, and let $x_1, x_2, x_3, x_4, x_5, x_6$ be `other_mask_prop`, `mask_prop`, `large_events`, `distancing`, `worked_outside` and `bar_visit` on the same day. Then

$$\mathbb{E}[Y|\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)] = 114.7556 - 4.9958x_1 + 2.5152x_2 - 4.1310x_3 + 0.6665x_4 + 1.4022x_5 + 0.2075x_6 = g_1(\boldsymbol{x})$$

if the day is on or before July 25th, 2021 and

$$\mathbb{E}[Y|\boldsymbol{x}] = g_1(\boldsymbol{x}) - 557.2660 + 3.2087x_1 + 11.4271x_2 - 15.0124x_4 + 4.7663x_5$$

if the day is after July 25th, 2021. Care must be taken in interpreting that coefficients in this model, since they only reflect how expected case counts vary when exactly one indicator of social behavior changes and other indicators are held constant. This does not lend to a natural interpretation of what the model says about how each indicator predicts case counts as a whole; to pursue this objective, the current methods are not suitable.

The current approach assumes that case counts and the included measures of social behavior are only linearly related, which is quite strong. Hence it can be argued that the analysis would be more complete if non-linear functions of the measures were included in the posited models, but it is decided that this is unnecessary. Firstly, the residual plot does not suggest that there are major uncaptured variations in the data unexplained by the model; there is only some heterocedasticity as well as an unusual line towards the left of the plot (i.e. for lower predicted case counts). The heterocedasticity is likely to be resolved by using log case counts instead of plain case counts, but they are monotonic functions of each other, so the answer to the first objective of the study is not likely to change. A method to eliminate the line in the residual plot is currently not known. Secondly, the QQ plot of the residuals shows that the normal errors assumption is not strongly violated. Altogether, the model assumptions hold up well, so the analysis may conclude here.

It is reasonable to doubt the validity of the introduction of the time split in order to improve the predictive performance of the model. In essence, such a decision forces the model to say that COVID-19 case counts are associated differently to the aspects of social behavior studied after a certain point in time, so the time split is unnatural in a sense. There is only dubious evidence of this in the present data set in the univariate scatterplots and the results of the naive the regression tree, the latter of which was initially based on the assumption that COVID-19 case counts fluctuate in waves and that the tree would identify if there was indeed a transition of waves observed in the dataset. A more compelling approach for judging the validity of the time split as well as potentially identifying a better split in the dataset would be to cluster the observations in the dataset based on the measures of social behavior, effectively leading to a semi-supervised analysis approach. This was not pursued due to time constraints.

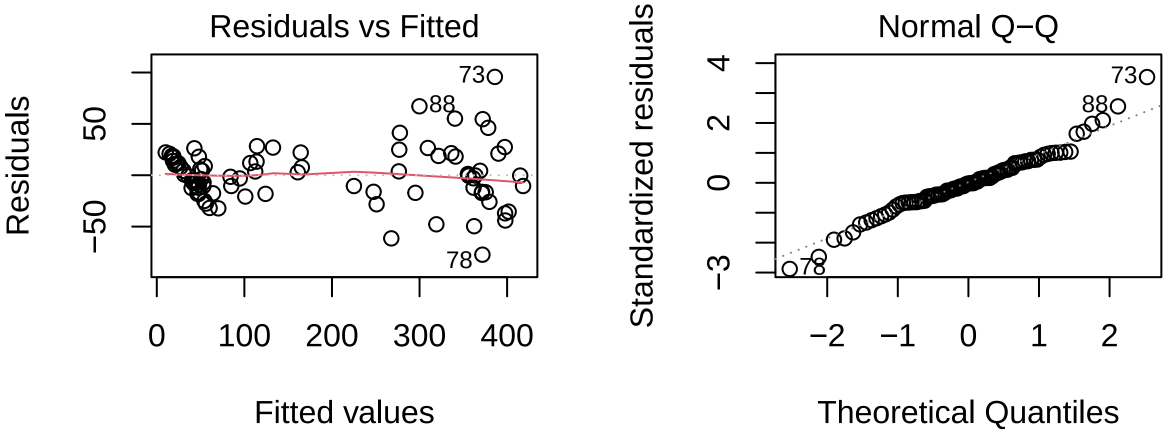

Figure 4: Diagnostic plots of stepwise-selected model.

References

[1] P. Tupper, H. Boury, M. Yerlanov, and C. Colijn, "Event-specific interventions to minimize COVID-19 transmission," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 50, pp. 32038–32045, 2020.