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

Quang Vuong

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

We will first briefly look at basic descriptive statistics of all variables involved.

Table 1: Descriptive statistics of predictors and case counts.

	Mean	Standard.deviation
distancing	18.43	5.31
bar_visit	55.30	40.10
large_events	19.63	3.51
mask_prop	65.84	7.81
other_mask_prop	40.26	10.25
public_transit	53.55	2.81
resto_visit	101.36	14.48
worked_outside	28.63	2.49
cases	179.94	150.07

It is clear that all variables have very different scales from each other, so models with unstandardized and standardized variables will both be considered.

It is the most instructive to now 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`, `public_transit` and `worked_outside` appear to have two trend lines. 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.

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, which 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 tree, 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, full linear models will be fitted for their simplicity and ease of interpretation. Secondly, these models will

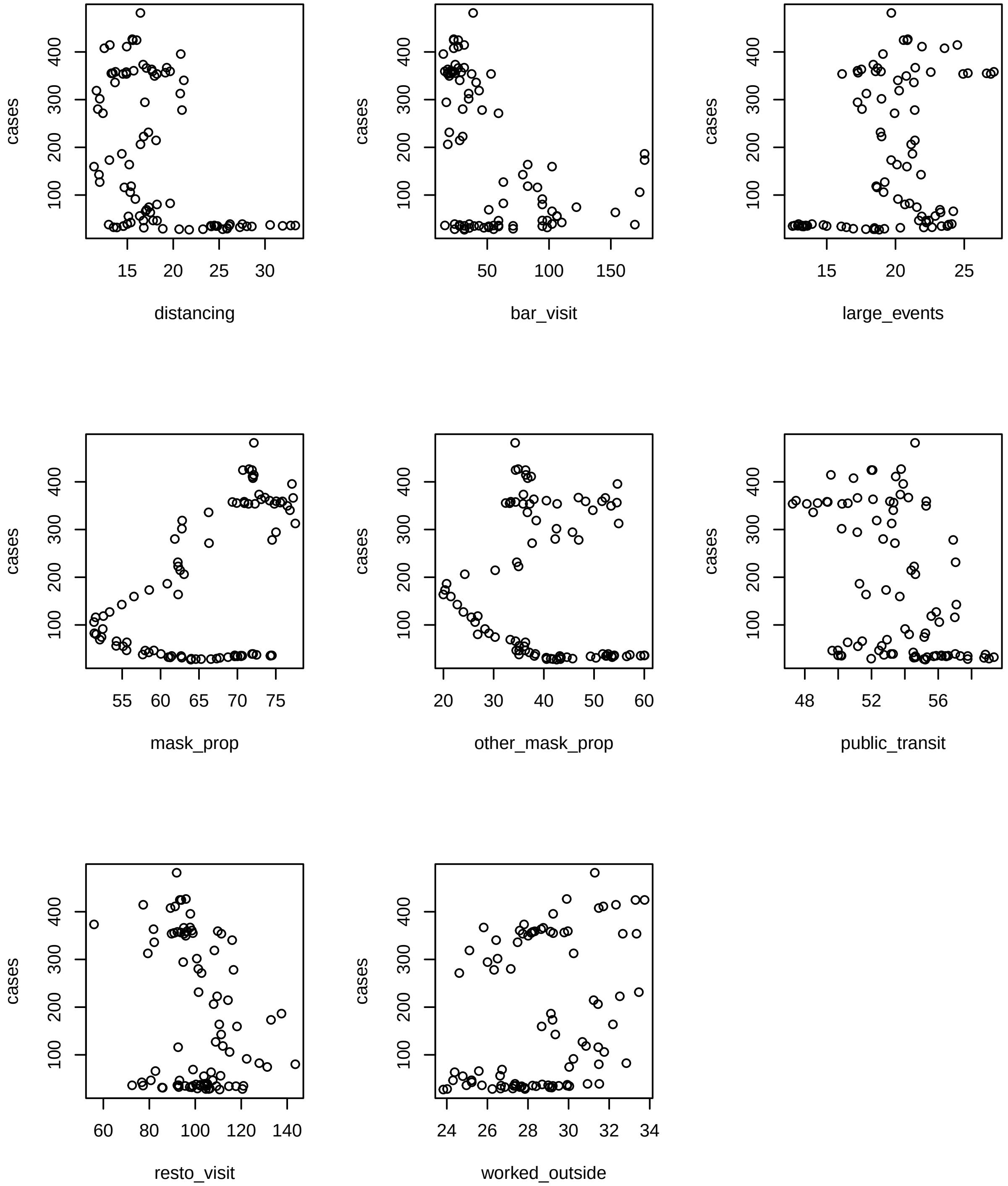

Figure 1: Scatterplots of predictors against case counts.

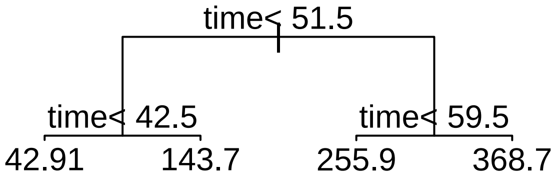

Figure 2: Regression tree to determine clusters in data.

be compared against regularized and stepwise-slected models. As interpretation is one of the goals of the analysis, we opt to avoid non-parametric approaches. We will compare the models by their estimated mean squared prediction error calculated from a leave-one-out cross-validation procedure that uses mean squared errors. This estimate is chosen because it is flexible enough to adapt to the other model-fitting approaches that will be used in this study.

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 first 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. Least squares estimation is done on the data to estimate the coefficients in the model; we refrain from showing the results of the estimation before a preferred model has been selected.

Now, the existence of the time split identified in the Exploratory data analysis section informs us to consider 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. Again, least squares estimation is used to estimate the coefficients of this model.

To check if these model assumptions hold well, we will look at their residual and QQ plots in Figure 3. Both residual plots have an unusual line for lower fitted case counts, but this is much less apparent for the model with the time split. A method to resolve this is currently unknown. There also seems to be some heterocedasticity for both models, which is possibly remedied by fitting log case counts instead. However, it is decided that this is not necessary because homocesdasticity is only involved in calculating standard errors to be used for hypothesis testing of the model coefficients, which is not of importance currently. Otherwise, both QQ plots suggest that the normal errors assumption is upheld quite well, and the residual plot of the full model with the time split is patternless. Therefore, it is concluded that the assumptions of the model with the time split are decently upheld.

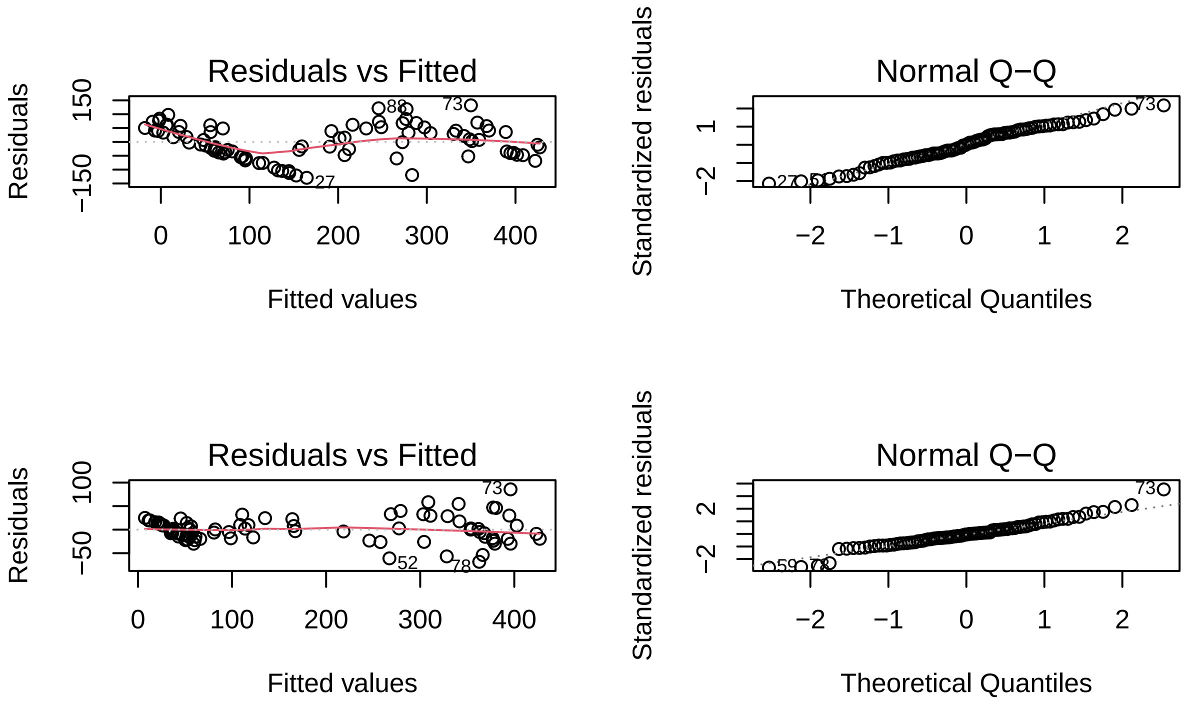

Figure 3: Diagnostic plots for full model without time split (top) and with time split (bottom)

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.

Table 2: CV scores and standard errors of the full linear models with and without the time split.

Model	Score	SE
Full linear, no time split	4385.056	536.2580
Full linear with time split	1151.979	241.5521

Note that the standard errors of the CV scores are calculated by dividing the standard deviation of the computed scores by the square root by the sample size. This is because the standard deviation of the computed scores is an unbiased estimator of the standard deviation of the out-of-sample squared prediction error, and the division is required to obtain an unbiased estimate of the standard deviation of the mean out-of-sample squared prediction error.

Next, we will attempt regularization approaches, owing to the correlation between predictors found in the Exploratory data analysis section. The two posited models are fitted again using LASSO. 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 CV scores and standard errors of these models are shown below.

Table 3: CV scores and standard errors of regularized models.

Model	Score	SE
LASSO, no time split	4298.099	570.7778
LASSO with time split	1780.747	358.5716
Ridge, no time split	4423.339	594.0015
Ridge with time split	1887.076	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. To see if a more parsimonious model performs better, We now attempt stepwise variable selection on the model with the time split. It is hoped that the function used for this procedure, `stepAIC`, introduces less bias into the variable selection due to the properties of the likelihood, but we are not sure. The CV score of the stepwise-selection procedure is as below, where variable selection is performed on each training set.

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

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

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

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

To judge if non-linear transformations of data are necessary, we will consult the residual and QQ plots of the stepwise-selected model in Figure 4. They look like those of the full linear model with the time split, which have been discussed previously. Altogether, the model assumptions still hold up well, so the analysis may conclude here, with the stepwise-selected model being the most preferred one. There is no need to reconsider models with standardized variables because the all of the linear models fitted are scale-invariant, and `cv.glmnet` has built-in scaling.

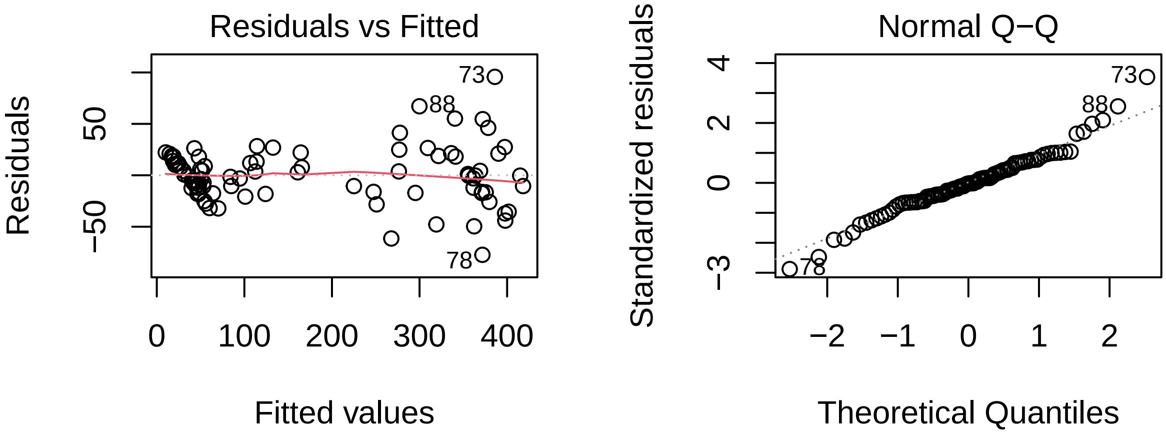

Figure 4: Diagnostic plots of stepwise-selected model.

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. The coefficient estimates, their standard errors and p-values are shown in the following table.

Table 4: 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

4 Discussion

The predictive standpoint focuses on which predictors appear in the selected model without regards for hypothesis testing concerns, 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. Since the first objective prioritizes the prediction of COVID-19 case counts, this is a satisfactory conclusion. 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. The coefficients in this model should not be interpreted directly because they 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.

It is reasonable to doubt the validity of the introduction of the time split in order to improve the predictive performance of the model, which essentially states that the relationship between case counts and public behavior change after a certain point in time. 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 would essentially be a semi-supervised learning procedure, which was not pursued due to time constraints.

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.