An Analysis of COVID-19 Incidence in New York City

Ian Finn

May 2020

Abstract

The COVID-19 pandemic has had a dramatic impact on the economic, social and personal lives of Americans across the country, and this effect has been perhaps most keenly felt in New York City. The ongoing national dialogue regarding when and how to re-open shuttered businesses underscores the need to understand the types of social activities and venues most strongly correlated with the initial spread of the virus. To that end, this report investigates the incidence of COVID-19 cases in New York city by leveraging Foursquare data on social venues throughout the city. Regression and machine learning techniques are utilized, with population data from the U.S. Census Bureau as a control. I find that the number of transportation terminals in a given zip code is a statistically significant predictor of COVID-19 cases, and this result is robust to a number of model specifications. I also find that the number of restaurants and recreational venues is not a positive predictor of COVID-19 cases in the early stages of the outbreak, nor a positive predictor of the growth rate in cases between 4/1/2020 and 5/12/2020, the time of writing.

Introduction

This report explains the incidence of COVID-19 cases in New York City (NYC) by leveraging Foursquare mobile application data on social venues across the 177 zip codes contained within the city's borders. In the early months of 2020 NYC was the global epicenter of the virus, and at the time of writing represents 4.5% of all COVID-19 cases worldwide, and 13.4% of cases in the United States. Nevertheless, the city is witnessing a clear downward trend in daily new cases reported: on May 8th only 425 new cases were identified, down from a peak of 6,213 on April 6th. This ebb is likely due to aggressive measures enacted by state and local policymakers. On March 22nd Governor Andrew Cuomo instituted a state-wide lockdown which mandated the closure of all nonessential businesses; leaving only grocery stores, pharmacies and other essential operations open. All non-solitary outside activities, including team sports, were also banned.

While these stringent restrictions have doubtlessly played an important role in stemming the tide of new cases, their macroeconomic effects cannot be overlooked. Over the course of the last five weeks, data from the New York Department of Labor reveal that the total number of New Yorkers filing new jobless claims has reached 1.4 million. This is roughly two-thirds the total jobless claims filed during the entire span of the Great Recession, which lasted from June 2007-November 2009. Adding these 1.4 million jobless filers to the ranks who had previously filed in the months before the COVID-19 crisis began puts New York state's unemployment rate at approximately 13%.[1] This level of unemployment breaks the previous seasonally adjusted post-Great Depression record of 10.3% in February 1976, and does not take into account the sizable population of workers who are underemployed. To put this in perspective, during the Great Depression New York state's peak unemployment rate was about 23%.

Given the severe economic consequences of the mandated lockdown, and the encouraging trend in new cases reported, there is currently a robust debate among New York state policymakers about whether such drastic measures are still necessary. Health officials continually warn that a headlong rush to reopen businesses, if not done in a cautious and sensible manner, could lead to a re-emergence of the virus that once again overwhelms the healthcare infrastructure of the city. Therefore, an understanding of the types of activities and social venues most strongly correlated with the initial spread of COVID-19 could help inform policymakers' decision-making about which businesses can be re-opened, while still maintaining a tolerable level of risk of a subsequent outbreak. The following analysis does exactly this and, using Foursquare data on social venues across the 177 zip codes of NYC, explains which venues are the best predictors of COVID-19 incidence.

Data

The dependent variables in the analysis are the total number of positive COVID-19 cases, the ratio of positive tests to total tests administered and the growth rate in positive cases between 4/1/2020 (the earliest date available) and 5/12/2020 (date of writing), segmented by NYC zip code. These data are assembled by the NYC Department of Health and Mental Hygiene (DOHMH) Incident Command System for COVID-19 Response (Surveillance and Epidemiology Branch in collaboration with Public Information Office Branch), and are uploaded to GitHub daily. It should be noted that, due to the fact that this information is being collected in real-time, it is by definition preliminary and subject to change as the COVID-19

[1] Mayra Rodriguez Valladares. "New York State Unemployment Rate Is At Highest Level Since the Great Depression." Forbes, April 26, 2020.

response continues. The analysis focuses on reported COVID-19 cases as of 4/1/2020, the earliest date available, in order to investigate which venues most strongly contributed to the initial spread of the virus while minimizing the confounding effect of the lockdown. It seems reasonable to conclude that this effect is minimal, as the state-wide closure of businesses had been implemented only 10 days prior to the date the data was compiled.

Given the fact that the population of an area is an important predictor of disease transmission generally, this variable is used as a control throughout the analysis. These data are taken from U.S. Census Bureau. Including this variable in predictive models ensures that any inferences made about the types of venues associated with positive COVID-19 cases are not simply an artifact of the size of the population in a given zip code. This variable is also included in regressions which take the ratio of positive to total tests as the dependent variable in order to account for heterogeneity in public testing resources available, which are likely correlated with population.

Lastly, the independent variables of interest throughout the analysis are the total number of social venues located in each NYC zip code, as categorized by the author. These data are retrieved from the venues group of the Foursquare City Guide[2] API by first assigning each zip code centroid in NYC a latitude and longitude using the geocoder library, and then inputting these coordinates into formatted queries. The results are limited to a radius of 500 meters from the coordinates inputted, and duplicates are removed whenever radii overlap. The final variables are the total number of venues in a given zip code categorized as a hotel, restaurant, transportation terminal, store, market or recreational venue. These variables are generated through string searches of the venue categories returned by the API. The variable **Market**, for example, is the total number of venues in a given zip code with "market", "grocery", "supermarket", etc. listed in the venue category. It should be noted that these data suffer from coverage issues. However, missing observations can reasonably be characterized as missing at random, and correlated with population in a given zip code, which is controlled for throughout the analysis.[3] Summary statistics for the independent variables are given in the Appendix, while summary statistics for the dependent variables are shown in Table 1.

Results and Discussion

Exploratory Analysis

In order to visualize the variation in incidence of COVID-19 across New York City's zip codes, I first generate a choropleth map of the ratio of positive cases as of 4/1/2020 to total population.

[2] Foursquare City Guide is a mobile application developed by Foursquare Labs Inc., which provides personal recommendations of local venues based on users' browsing and check-in history.

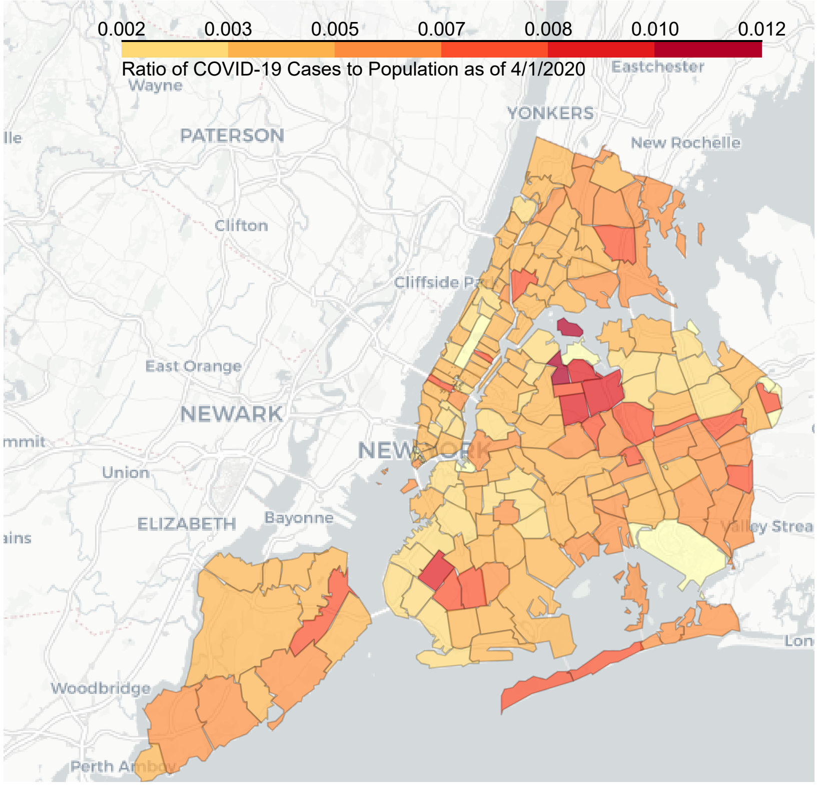
[3] For an excellent treatment see Donald B. Rubin, "Inference and Missing Data," *Biometrika* 63, no. 3 (December 1976): 581-592.

Table 1: Summary Statistics of Dependent Variables

	Cases as of 5/12/2020	Cases as of 4/1/2020	Ratio of Positive Tests to Total Tests	Growth Rate (%)
Count	177.00	177.00	177.00	177.00
Mean	1030.53	219.800	0.51	359.58
S.D.	733.00	152.00	0.095	135.90
Min.	22.00	6.00	0.25	75.00
25%	486.00	112.00	0.44	255.56
50%	831.00	182.00	0.51	361.08
75%	1457.00	306.00	0.58	457.74
Max.	4152.00	947.00	0.77	888.52

Dependent variables are calculated at the zip code level. *N*% refers to the *Nth* percentile. As such, 50% is the median. *Ratio of Positive Tests to Total Tests* is as of 4/1/2020. *Growth Rate* is calculated between 4/1/2020 and 5/12/2020. *Growth Rate* is expressed as a percentage.

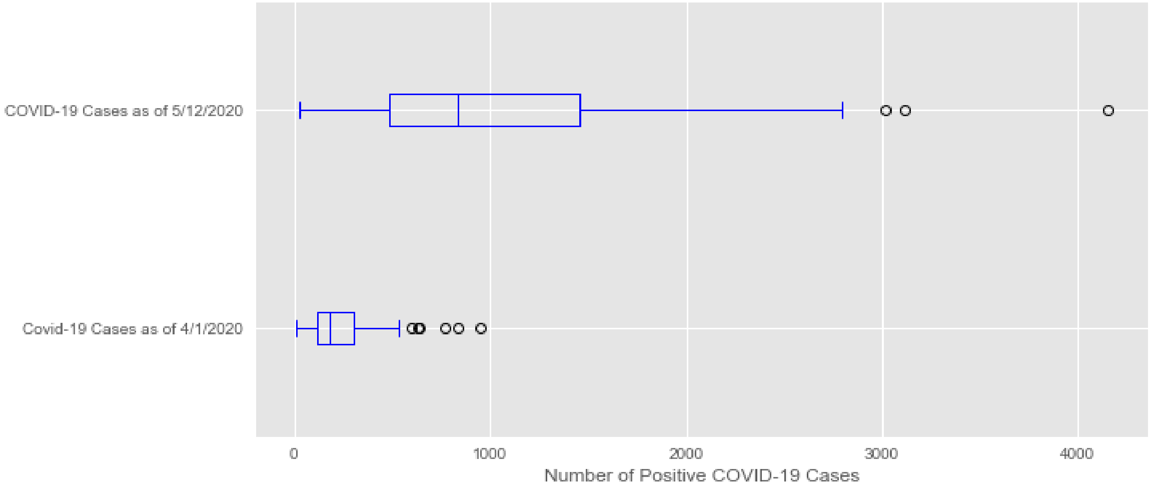
Figure 1: Ratio of Positive Cases to Total Population as of 4/1/2020 by NYC Zip Code



It is immediately obvious from Figure 1 that there exists a large degree of variation in COVID-19 cases across zip codes. Interestingly, Staten Island and Manhattan have similar incidence rates despite the fact that population density is demonstrably higher in the latter.

To more incisively analyze the distribution of positive cases as of 4/1/2020 and 5/12/2020, I generate box plots which are displayed in Figure 2.

Figure 2: Box Plots of Total Positive COVID-19 Cases by Zip Code: April vs. May



A number of observations can be gleaned from this plot. First of all, the rapid transmission of the virus between April and May is immediately apparent. On April 1st the median number of positive cases was 182, while in May it had swelled to 831. This underscores the importance of the New York state lockdown, as the disparity between these numbers would likely be much greater had it never been implemented. Secondly, both distributions have a number of outliers which likely represent the more densely populated zip codes of NYC. To further investigate this relationship I generate a scatter plot of positive COVID-19 cases in April and population, and overlay a simple regression line.

Figure 3: Scatter Plot of Positive COVID-19 Cases as of 4/1/2020 and Population

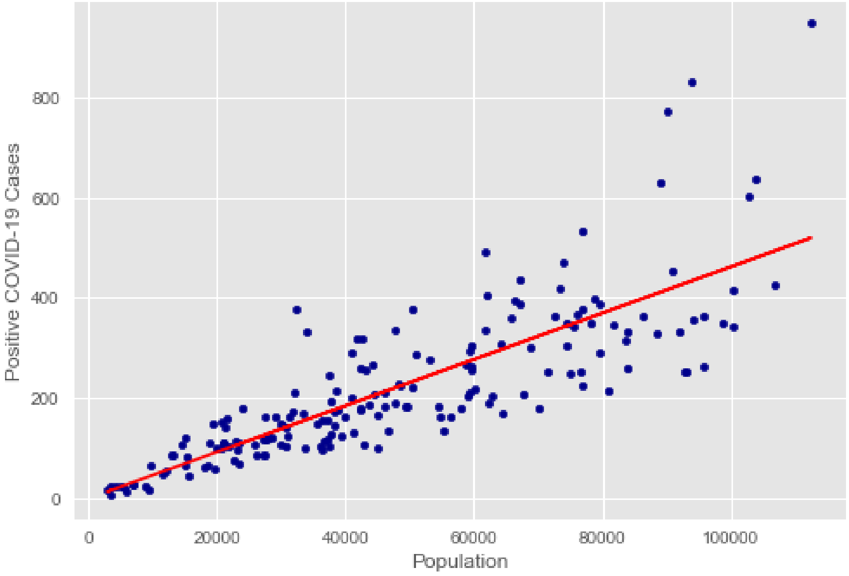


Figure 3 illustrates 2 important points. First of all, there is in fact a clear, linear relationship between COVID-19 cases and population which should be incorporated in the regression analysis to follow. Secondly, the variation in positive cases appears to increase as the population increases, and this heteroskedasticity should also be accounted for in the models derived below.

Regression Analysis

As an initial step in understanding the factors that explain COVID-19 transmission in NYC I run an OLS regression with total positive COVID-19 cases as of 4/1/2020, the earliest date available by zip code, as the dependent variable. As mentioned in the introduction, the earliest data is utilized in order to curtail the confounding effect of the lockdown, which was implemented in New York state on March 22nd. The independent variables are the total population and the total number of different types of venues, as categorized above. Heteroskedasticity robust standard errors are employed to minimize the risk of biased inference, as indicated above. To assess potentiality of multicollinearity, I also calculate the variance inflation factor (VIF) for each of the independent variables. A widely utilized rule of thumb is that the VIF for each variable should be below 10, which is indeed the case. The results of this model (Model 1) are illustrated in Table 2.

Table 2: Total Positive COVID-19 Cases as of 4/1/2020 (OLS)

Model 1

	Coefficient	Standard Error	P-Value	VIF
Constant	18.359 (1.26)	14.53	0.206	7.10
Population	0.005*** (11.786)	0.00	0.000	1.26
Transportation	9.559* (1.824)	5.239	0.068	1.10
Market	-1.588 (-0.410)	3.870	0.682	1.45
Store	1.051 (0.869)	1.209	0.385	2.03
Restaurant	-0.591 (-0.695)	0.850	0.487	3.74
Bar	-1.608 (-0.597)	2.694	0.551	2.44
Recreation	-2.408** (-2.362)	1.020	0.018	2.65
Hotel	10.518* (1.880)	5.595	0.060	2.03

Model estimated using Ordinary Least Squares with heteroskedasticity-robust standard errors. T-statistics reported in parentheses. Significance levels reported are as follows: * p<0.10, ** p<0.05, and *** p<0.01. VIF is the variance inflation factor.

It is evident from the t-statistics (reported in parantheses) that the population of a given zip code is the single best predictor of how many positive COVID-19 cases are observed. The generally accepted norm is that a z statistic greater than 1.96 in absolute value implies statistical significance. The z-statistic for **Population** is 11.786, and the corresponding p-value is .000; which means that, with repeated sampling from the data generating process, the probability of observing an effect at least as large, given that there is no relationship between

COVID-19 cases and population, is 0.0%! **Transportation** is significant at the 6.8% level, which seems reasonable given that transportation hubs often witness dense concentrations of people in minimally ventilated confines, and are therefore favorable environments for disease transmission. Moreover, many residents of the city continued to rely on public transportation for their work commute in April, despite the lockdown being instituted on March 22nd. It should be noted, however, that statistical significance is distinct from practical significance, and the latter may in fact be of more interest to policymakers. For example, the coefficient on **Store** implies that an additional store in a zip code, ceteris paribus, will lead to roughly 1 additional positive case of COVID-19, on average. While this effect is not statistically significant, it should nevertheless be taken into account when designing preventative measures to be implemented within the retail industry.

Interestingly, **Market** does not appear to be important predictor of the spread of the virus, and this is empirically verified by the results of a LASSO regression, shown below. LASSO is a type of penalized regression in which the sum of the absolute value of the regression coefficients is constrained to be less than a constant, which is equivalent to imposing a double exponential prior, centered at zero, on the coefficients. This often results in the "shrinking" of coefficient estimates. As can be seen from the results below, when this technique is implemented with 10-fold cross-validation, and the tuning parameter is optimized using a grid search, the coefficients on **Market** and **Store** are shrunk to 0. It should be noted that the inferences made in the following models below are robust to the omission of these variables.

Table 3: Total Positive COVID-19 Cases as of 4/1/2020 (LASSO)

Model 2

	Coefficient	Standard Error	P-Value
Constant	21.216	16.995	0.214
	(1.248)		
Population	0.004***	0.000	0.000
	(37.509)		
Transportation	6.560	5.813	0.261
	(1.129)		
Market	0.000	3.818	1.000
	(0.000)		
Store	0.000	1.414	1.000
	(0.000)		
Restaurant	-0.126	0.516	0.808
	(-0.243)		
Bar	-2.162	2.406	0.370
	(-0.899)		
Recreation	-1.236	1.106	0.265
	(-1.118)		
Hotel	0.000	8.654	1.000
	(0.000)		

Dependent variable is total COVID-19 cases as of 4/1/2020, by zip code. Model estimated using Least Absolute Shrinkage and Selection Operator (LASSO). Heteroskedasticity-robust standard errors employed. T-statistics reported in parentheses. Significance levels reported are as follows: * p<0.10, ** p<0.05, and *** p<0.01.

A legitimate concern for OLS in this instance is that inference may be biased by the fact that, for various reasons, access to testing was greater in some zip codes as opposed to others. As of May 12th, 1,182,998 people in the entire state of New York have been tested for the virus, which is roughly 61 tests per 1,000 people. It seems reasonable to think that, given the dearth of resources available for testing, some areas in NYC would have greater access than others. To alleviate this concern, I also run an OLS regression in which the dependent variable is the proportion of people testing positive for the virus, out of the total number of tests administered in a given zip code, while retaining a control for total population. Model 3 includes all covariates utilized in the regressions above, while Model 4 removes those covariates that were eliminated in the LASSO regression.

Table 4: Ratio of Positive COVID-19 Tests to Total Tests as of 4/1/2020 (OLS)

	Model 3			**Model 4**		
		Standard			Standard	
	Coefficient	Error	P-Value	Coefficient	Error	P-Value
Constant	0.513***	0.017	0.000	0.513***	0.017	0.000
	(29.789)			(30.511)		
Population	$8.386e^{-7}$***	$2.65e^{-7}$	0.002	$8.885e^{-7}$***	$2.58e^{-7}$	0.001
	(3.159)			(3.450)		
Transportation	0.009*	0.005	0.051	0.0096**	0.005	0.036
	(1.956)			(2.094)		
Market	0.002*	0.003	0.613			
	(0.506)					
Store	0.001	0.001	0.364			
	(0.908)					
Restaurant	-0.002**	0.001	0.015	-0.001**	0.001	0.022
	(-2.443)			(-2.288)		
Bar	0.002	0.002	0.338	0.002	0.002	0.328
	(0.958)			(0.979)		
Recreation	-0.005***	0.001	0.000	-0.005***	0.001	0.000
	(-3.981)			(-4.537)		
Hotel	0.003	0.007	0.663			
	(0.436)					

Dependent variable is the ratio of positive COVID-19 cases to total cases as of 4/1/2020, by zip code. Model estimated using Ordinary Least Squares and heteroskedasticity-robust standard errors. T-statistics reported in parentheses. Significance levels reported are as follows: * $p<0.10$, ** $p<0.05$, and *** $p<0.01$.

As illustrated by the table above, the inferences garnered from this model are radically different, and cast significant doubt on the results obtained in Model 1. While **Population** has retained its statistical significance, the importance of the number of hotels in a given zip code is dramatically reduced, which accords with the results of the LASSO regression above. Moreover, the number of restaurants in a given zip code is now a significant, negative predictor in the model. Finally, the coefficient on **Transportation** is now statistically significant at the 5% level in Model 4, and is barely below this threshold in Model 3.

It should be noted, however, that an OLS regression with a proportion as the dependent variable can potentially render misleading results. First of all, predicted values from such models may not necessarily lie within the interval [0,1]. Secondly, such models may violate the normality assumption, as well as the linearity assumption necessary for the Gauss-Markov theorem to be applicable. Nevertheless, such models can be instructive if handled with care. As a beginning step to determine whether the model is appropriate, I examine the true and predicted values for the proportion of positive COVID-19 cases obtained. Fortunately the true proportion varies along a very restricted interval, with a minimum of 0.25 and a maximum of 0.77, and due to this clustering of the data none of the predicted values from Model 3 fall outside the [0,1] interval, as can be seen in the table below:

Table 5: Comparison of True and Predicted Values from Model 3

	True Proportion	Predicted Proportion
Count	177.00	177.00
Mean	0.512	0.512
S.D.	0.095	0.060
Min.	0.250	0.292
25%	0.441	0.488
50%	0.514	0.527
75%	0.576	0.556
Max	0.772	0.602

N % refers to the *N*th percentile. As such, 50% is the median. Predicted proportion derived from an OLS regression with ratio of positive cases to total tests of 4/1/2020 as the dependent variable and heteroskedasticity-robust standard errors (Model 3).

To further assess the appropriateness of OLS in this instance, I analyze 2 commonly utilized diagnostic plots. In order to gauge the linearity assumption I plot the studentized residuals from Model 3 against the fitted values. If the model is well specified, this plot will look like random noise, with a fitted red line that is roughly horizontal; otherwise, we are underfitting the model and not capturing the non-linearity of the data. Secondly, in order to determine the plausibility of the normality assumption I employ a normal Q-Q plot, which graphs the studentized residuals against the expected order statistics of the standard normal distribution. If the normality assumption is appropriate, the studentized residuals will fall on the line y=x (in red).

Figure 4: Residuals vs. Fitted Values from Model 3

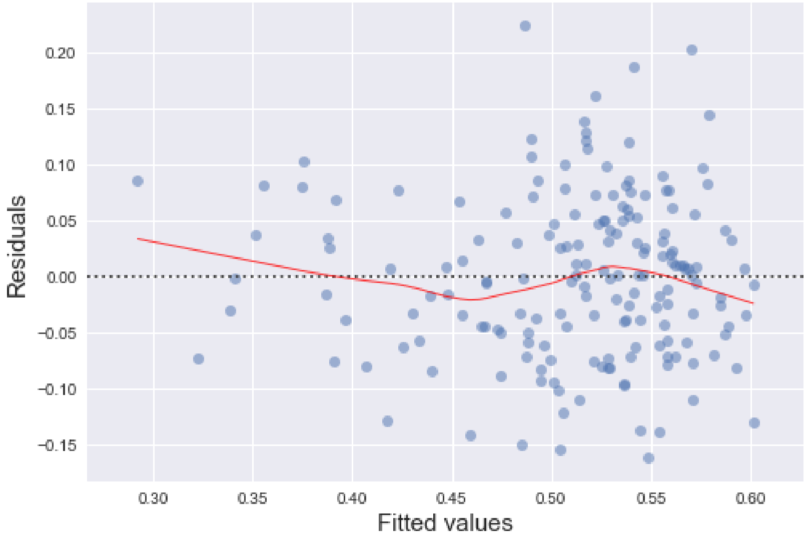
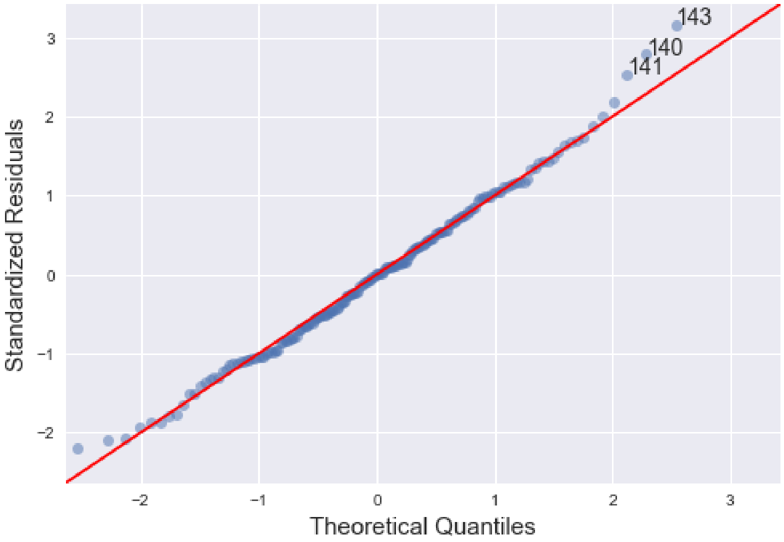


Figure 5: Normal Q-Q Plot from Model 3



Both of these plots cast serious doubt on the appropriateness of OLS in this instance. As we can see from Figure 5, the normality assumption does not appear to hold in the tails of the distribution. Moreover, Figure 4 indicates that there is in fact non-linearity in the data which are not being captured in Model 3. Once again, and as is often the case in such settings, the tails of the distribution are most problematic. Taken together, this suggests that a non-linear model is more appropriate when analyzing the proportion of positive COVID-19 tests. As an improvement I employ a method proposed by Papke and Wooldridge (1996) in which a generalized linear model is utilized with robust standard errors. This approach makes use of the logit link function (that is, the logit transformation of the response variable) and the binomial distribution, which is the proper choice of family given this particular setting, as testing for COVID-19 can be thought of as a sequence of Bernoulli trials. The results of this technique are illustrated in Table 6.

As is evident from the table, the use of the approach advocated by Papke and Wooldridge (1996) does not qualitatively affect most of the the inference. The variables **Population**,

Recreation, and **Restaurant** continue to be statistically significant predictors in Model 4 and Model 5. However, an important result for policymakers is that **Transportation** is also now statistically significant at the 5% level in both models. The robustness of this result suggests that it is in fact an important predictor of COVID-19 transmission in NYC. According to Model 5, in repeated sampling we would observe an effect at least as large, if there were no relationship between transportation terminals and the proportion of positive COVID-19 tests, only 3.2% of the time. These results lend important insights into how and which venues in NYC can be re-opened while still maintaining a tolerable risk of virus transmission. In the case of transportation hubs, it also suggests that trenchant measures should be taken to ensure the safety of these areas.

Table 6: Ratio of Positive COVID-19 Tests to Total Tests as of 4/1/2020 (GLM)

	Model 4			**Model 5**		
		Standard			Standard	
	Coefficient	Error	P-Value	Coefficient	Error	P-Value
Constant	0.0524**	0.068	0.440	0.0541	0.067	0.418
	(0.772)			(0.810)		
Population	$3.368e^{-6}$***	1.05^{-6}	0.001	$3.58e^{-6}$***	$1.03e^{-6}$	0.001
	(3.197)			(3.475)		
Transportation	0.0365**	0.018	0.044	0.039**	0.018	0.032
	(2.019)			(2.144)		
Market	0.007	0.013	0.592			
	(0.536)					
Store	0.004	0.004	0.347			
	(0.941)					
Restaurant	-0.007**	0.003	0.013	-0.006**	0.002	0.022
	-(2.488)			(-2.296)		
Bar	0.009	0.009	0.313	0.0087	0.008	0.307
	(1.009)			(1.022)		
Recreation	-0.022***	0.005	0.000	-0.0202***	0.004	0.000
	(-4.021)			(-4.521)		
Hotel	0.012	0.027	0.650			
	(0.454)					

Dependent variable is the ratio of positive COVID-19 cases to total cases as of 4/1/2020, by zip code. Model estimated using Generalized Linear Model with Logit link function and Binomial distribution, with heteroskedasticity-robust standard errors employed. T-statistics reported in parantheses. Significance levels reported are as follows: * $p<0.10$, ** $p<0.05$, and *** $p<0.01$.

As a final step in the analysis I investigate the factors that explain the growth rate in positive COVID-19 cases over time. Models 6 and 7 shown below displays the results of an OLS regression in which this growth rate, calculated over the period 4/1/2020 - 5/12/2020, is taken as the dependent variable and the same set of independent variables are utilized. No transformation is applied to the data, as the growth in positive cases appears to be largely linear, and is certainly not exponential in a manner that would require logarithmic scaling. This can be seen in Figure 6 below, which plots positive COVID-19 cases as of 5/12/2020 against positive COVID-19 cases as of 4/1/2020, with a quadratic polynomial fitted to the data.

Figure 6: Scatter Plot of Positive COVID-19 Cases Over Time with Quadratic Fit

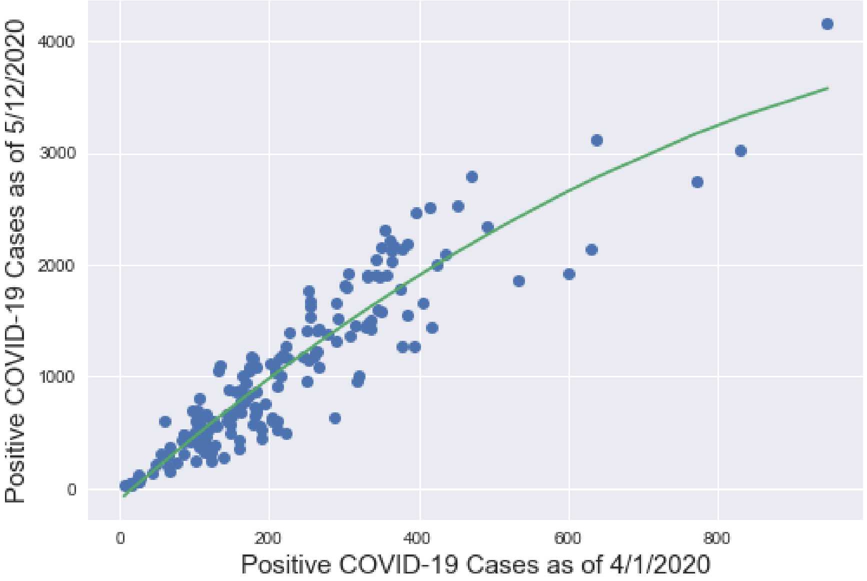


Table 7: Growth Rate in COVID-19 Cases Between 4/1/2020 and 5/12/2020

	Model 6			**Model 7**		
		Standard			Standard	
	Coefficient	Error	P-Value	Coefficient	Error	P-Value
Constant	4.004***	0.235	0.000	4.036***	0.229	0.000
	(17.029)			(17.658)		
Population	$7.222e^{-6}$**	$3.37e^{-6}$	0.032	$8.148e^{-6}$**	$3.19e^{-6}$	0.011
	(2.143)			(2.555)		
Transportation	-0.023	0.086	0.791	-0.021	0.087	0.889
	(-0.264)			(-0.140)		
Market	0.0380	0.045	0.379			
	(0.880)					
Store	0.005	0.023	0.839			
	(0.204)					
Restaurant	-0.020**	0.008	0.017	-0.018***	0.006	0.004
	(-2.378)			(-2.917)		
Bar	-0.046*	0.024	0.054	-0.040*	0.021	0.054
	(-1.926)			(-1.923)		
Recreation	-0.053***	0.017	0.001	-0.061***	0.012	0.000
	(-3.194)			(-5.157)		
Hotel	-0.087	0.091	0.341			
	(-0.953)					

Dependent variable is the growth rate in total COVID-19 cases from 4/1/2020-5/12/2020, by zip code. Model estimated OLS with heteroskedasticity-robust standard errors employed. T-statistics reported in parantheses. Significance levels reported are as follows: * p<0.10, ** p<0.05, and *** p<0.01.

As one would expect, the population of a given zip code remains a statistically significant covariate in both models. In addition, the results indicate that **Restaurant** and **Recreation** are negatively correlated with the growth rate of COVID-19 cases over this time period, which underscores the efficacy of the shutdown measures enacted in the city.

Conclusion

The COVID-19 pandemic has had a dramatic impact on the economic, social and personal lives of Americans across the country. Due to a variety of factors including population density and economic scale, this effect has perhaps been most keenly felt in New York City. According to the NYC Department of Health and Mental Hygiene (DOHMH), the number of confirmed deaths attributed to the virus is at least 15,253, with another 5,051 deaths "probably" due to the same cause; this represents 17.3% - 23.0% of all deaths in the United States due to the virus. The state-wide closure of all businesses deemed to be non-essential, while clearly efficacious in curbing the spread of the disease, has also had a devastating effect on the livelihoods of many New Yorkers. According to the New York Times, city officials estimate that half of the hotels in the city are not operating, and some 186,000 shops employing fewer than 10 people could fail.[4] This has spurned an intense debate over whether shuttered businesses should be allowed to re-open and, if so, how to accomplish this in a manner that does not unnecessarily put additional lives at risk. This discussion highlights the importance of understanding which factors contributed to the transmission of the virus in the first place, so that these insights can be incorporated into a reasoned plan for resuming normal economic and social activities.

To that end, the foregoing analysis employed a regression framework to investigate the categories of business venues most strongly associated with positive COVID-19 cases as of 4/1/2020, the earliest date for which zip code-level data were available. The number of transportation hubs in a given zip code was identified as a statistically significant predictor of the proportion of positive tests using the preferred GLM model; in addition, it was statistically significant at the 3.6% level using OLS. Interestingly, the number of restaurants, bars, and recreational venues in a zip code did not have a positive, statistically significant impact on the proportion of COVID-19 positive cases. Moreover, these venues did not have a positive, statistically significant impact on the growth rates in cases. While future research should utilize counterfactual observations or instrumental variables to isolate causal effects, these insights are nevertheless useful to policymakers tasked with managing the timing and form of the resumption of normal economic activity within the city.

[4] J. David Goodman. “I Don’t Think the New York That We Left Will Be Back For Some Years.” New York Times, May 6, 2020.

Appendix

Table 8: Summary Statistics of Independent Variables by NYC Zip Code

	Population	Hotel	Trans.	Store	Market	Restaurant	Bar	Recr.
Count	177.00	177.00	177.00	177.00	177.00	177.00	177.00	177.00
Mean	47645.06	0.44	0.67	6.65	2.26	18.46	2.68	6.11
S.D.	26698.40	1.05	1.12	6.23	1.99	17.14	4.01	7.45
Min.	22.00	1.05	1.12	6.26	1.99	17.14	4.01	7.45
25%	486.00	27403.00	0.00	2.00	1.00	5.00	0.00	1.00
50%	42653.00	0.00	0.00	5.00	2.00	12.00	1.00	3.00
75%	67094.00	0.00	1.00	10.00	4.00	28.00	4.00	9.00
Max.	112425.00	7.00	6.00	26.00	9.00	65.00	23.00	44.00

Independent variables are calculated at the zip code level. $N\%$ refers to the Nth percentile. As such, 50% is the median. $Trans.$ and *Rec.* refer to the total number of transportation terminals and recreational venues, respectively, in a given zip code.

References

[1] Donald B. Rubin, “Inference and Missing Data,” *Biometrika* 63, no. 3 (December 1976): 581-592.

[2] J. David Goodman. “I Don’t Think the New York That We Left Will Be Back For Some Years.” New York Times, May 6, 2020.

[3] Leslie E. Papke and Jeffrey M. Wooldridge, "Econometric Methods for Fractional Response Variables with an Application to 410(k) Plan Participation Rates," *Journal of Applied Econometrics* 11, no. 6 (November 1996): 619-632.

[4] Mayra Rodriguez Valladares. “New York State Unemployment Rate Is At Highest Level Since the Great Depression.” Forbes, April 26, 2020.