An Analysis of COVID-19 Incidence in New York City

Ian Finn

In Fulfillment of the IBM Data Science Professional Certificate

May 30, 2020

Outline

- Motivation
  - Ramifications of COVID-19 and state-wide lockdown
- Data
  - COVID-19 incidence
  - Foursquare API data on NYC social venues
- Results and discussion
- Concluding remarks

Impact of COVID-19 in NYC

- In the early months of 2020 NYC was the global epicenter of the virus
    - 4.5% of COVID-19 cases worldwide
    - 13.4% of cases in the U.S.
- Yet there is a clear downward trend in new cases reported
    - 425 new cases identified on May 8th
    - Peak of 6,213 on April 6th
- Ebb is likely due to the state-wide shutdown of non-essential businesses

Impact of the Lockdown

- March 22nd Governor Andrew Cuomo instituted a state-wide lockdown
- All businesses deemed non-essential temporarily shuttered
  - All non-solitary outside activities also banned
- Macroeconomic effects:
  - 1.4 million New Yorkers filed new jobless claims over a 5 week period
  - New York State unemployment rate at 13%
  - Breaks the post-Great Depression, seasonally-adjusted record of 10.3%

How to Proceed?

- Given the severe economic consequences, and the trend in new cases reported, is the lockdown still necessary?
- Re-opening businesses could lead to a re-emergence of the virus
  - Overwhelm healthcare infrastructure of the city
- Understanding of the social venues most strongly correlated with COVID-19 could improve policymaking
  - Focus preventative measures on highest-risk areas

COVID-19 Data

- Earliest data available is utilized to minimize confounding effect of lockdown
- Dependent variables (by zip code):
    - Total number of positive COVID-19 cases
    - Ratio of positive tests to total tests administered
    - Growth rate in positive cases between 4/1/2020 and 5/12/2020 (date of writing)
- Compiled by NYC Department of Health and Mental Hygiene

Population Data

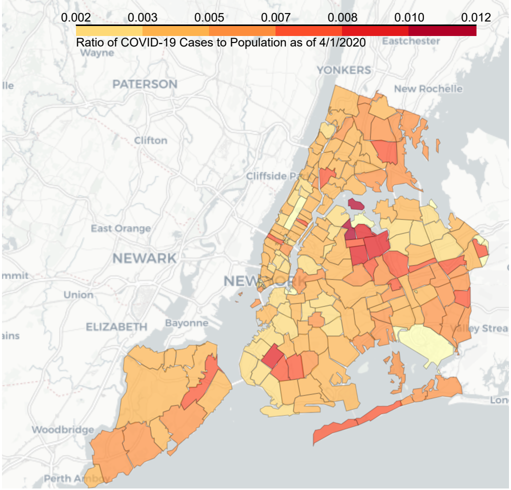
- Population is an important predictor of disease transmission generally
  - Exploratory analysis below confirms effect in this instance
  - Utilized as a control throughout the analysis
- Taken from the U.S. Census Bureau
- Scraped and merged onto zip code-level COVID-19 data

Foursquare Venues Data

- Independent variables of interest:
  - The total number of venues categorized as a hotel, restaurant, transportation terminal, store, market or recreational venue
- Segmented by zip code
  - Assigned geographic coordinates using geocoder
- Final variables generated through string searches of the venue categories returned by queries to the Foursquare API
- Radius of search is required input parameter
  - Duplicates removed in the event radii overlap

Choropleth Map

Figure: Ratio of Positive Cases to Total Population by NYC Zip Code

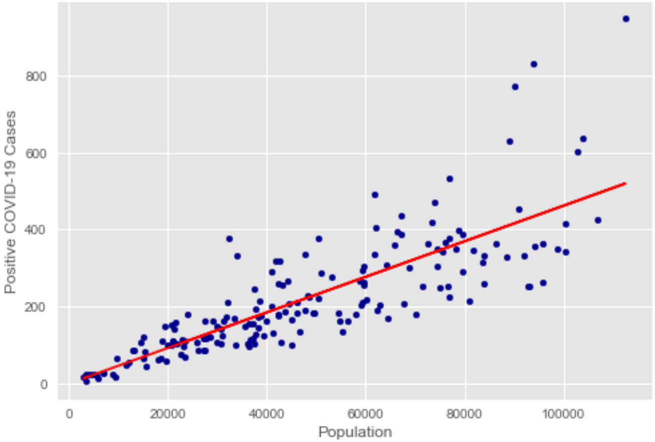


Choropleth Map Discussion

- Data is as of 4/1/2020
- Large degree of variation in COVID-19 incidence across zip codes
- Staten Island and Manhattan have similar incidence rates
  - Despite much greater population density in Manhattan
- Queens is most impacted borough
- Relationship between population and COVID-19 needs clarification
  - Solution: Scatter plot by zip code

COVID-19 and Population

Figure: Scatter Plot of Positive COVID-19 Cases and Population



Scatter Plot Interpretation

- Simple regression line is overlaid
- Two important observations:
  - Slope demonstrates a clear, linear relationship between population and COVID-19
  - Variation in positive cases is correlated with population
    - As population increases, so does the variance in COVID-19 incidence
    - Regression models need heteroskedasticity-robust standard errors
    - Reduces the potentiality of biased inference

Setup

- Models 1 and 2 estimated with Ordinary Least Squares (OLS) and LASSO
  - Grid search utilized to find optimal tuning parameter in LASSO
- Dependent variable is total positive COVID-19 cases as of 4/1/2020
  - Earliest data available utilized to minimize confounding effect of lockdown
- Heteroskedasticity-robust standard errors
- VIF calculated to assess multicollinearity
  - Rule of thumb: VIF $<$ 10 is preferred
- Significance convention: * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Results

Table: Total Positive COVID-19 Cases as of 4/1/2020

	Model 1: OLS			Model 2: LASSO			
	Coefficient	Standard Error	T-Statistic	Coefficient	Standard Error	T-Statistic	VIF
Population	0.005***	0.00	11.786	0.004***	0.000	37.509	1.26
Transportation	9.559*	5.239	1.824	6.560	5.813	1.129	1.10
Market	-1.588	3.870	-0.410	0.000	3.818	1.000	1.45
Store	1.051	1.209	0.869	0.000	1.414	1.000	2.03
Restaurant	-0.591	0.850	-0.695	-0.126	0.516	-0.243	3.74
Bar	-1.608	2.694	-0.597	-2.162	2.406	-0.899	2.44
Recreation	-2.408**	1.020	-2.362	-1.236	1.106	-1.118	2.65
Hotel	10.518*	5.595	1.880	0.000	8.654	1.000	2.03

Discussion

- **Population** is the single best predictor of how many positive COVID-19 cases are observed
  - As expected, is significant at the 1% level
- **Transportation** is significant at the 6.8% level in Model 1
  - Likely due to dense concentration of riders in poorly ventilated confines
- Interestingly, **Market**, **Store**, and **Hotel** are not statistically significant predictors
  - Coefficients are shrunk to 0 in LASSO regression
  - Following models estimated with/without these covariates

Potential Problems with Models 1 and 2

- Inference may be biased by the fact that access to testing was greater in certain zip codes
- Testing resources have been scarce
  - As of May 12th, 1,182,998 people in the entire state of New York have been tested
  - Roughly 61 tests per 1,000 people
- Solution: Models 3 and 4
  - OLS regression with ratio of positive tests to total tests as dependent variable

Results

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

	Model 3: OLS			Model 4: OLS		
	Coefficient	Standard Error	T-Statistic	Coefficient	Standard Error	T-Statistic
Population	$8.386e^{-7}$***	$2.65e^{-7}$	3.159	$8.386e^{-7}$***	$2.58e^{-7}$	3.450
Transportation	0.009*	0.005	1.956	0.0096**	0.005	2.094
Market	0.002*	0.003	0.506			
Store	0.001	0.001	0.908			
Restaurant	-0.002**	0.001	-2.443	-0.001**	0.001	-2.288
Bar	0.002	0.002	0.958	0.002	0.002	0.979
Recreation	-0.005***	0.001	-3.981	-0.005***	0.001	-4.537
Hotel	0.003	0.007	0.436			

Discussion

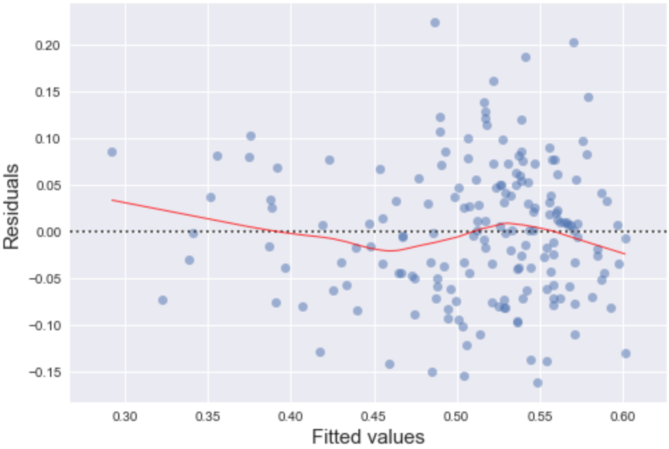
- Inferences from this model are qualitatively different
- Importance of **Hotels** is dramatically reduced in Model 3
- **Restaurant** is negative, statistically significant predictor in both models
- **Transportation** is significant at the 5% level in Model 4
  - Just barely below threshold in Model 3

Potential Problems With Models 3 and 4

- OLS regression with a proportion as the dependent variable can render misleading results
  - Predicted values may not be in [0,1] interval
- May violate normality and linearity assumptions of OLS
  - Latter is necessary for Gauss-Markov theorem to apply
- Diagnostic plots are employed to test whether these assumptions hold
  - Linearity: studentized residuals plotted against fitted values should look like white noise
  - Normality: normal Q-Q plot of studentized residuals should fall on 45° line

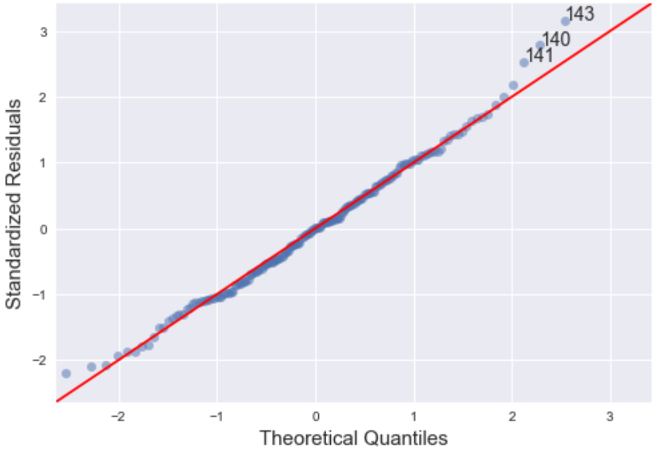
Diagnostic Plot

Figure: Residuals vs. Fitted Values from Model 3



Diagnostic Plot

Figure: Normal Q-Q Plot from Model 3



Diagnostic Plot Discussion

- Both plots cast doubt on the appropriateness of OLS
  - Normality assumption does not hold in tails of distribution
  - Residual vs. Fitted Values plot illustrates clear non-linearity
- A non-linear model is needed
  - Solution: method proposed by Papke and Wooldridge (1996)
  - Generalized linear model (GLM) with Logit link function and Binomial family
  - Relies on the fact that testing for COVID-19 is a sequence of Bernoulli trials

Results

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

	Model 5: GLM			Model 6: GLM		
	Coefficient	Standard Error	T-Statistic	Coefficient	Standard Error	T-Statistic
Population	$3.368e^{-6}$***	1.05^{-6}	3.197	$3.58e^{-6}$***	1.03^{-6}	3.475
Transportation	0.0365**	0.018	2.019	0.039**	0.018	2.144
Market	0.007	0.013	0.536			
Store	0.004	0.004	0.941			
Restaurant	-0.007**	0.003	-2.488	-.006**	0.002	-2.296
Bar	0.009	0.009	1.009	0.0087	0.008	1.022
Recreation	-0.022***	0.005	-4.021	-0.0202***	0.004	-4.521
Hotel	0.012	0.027	0.454			

Discussion

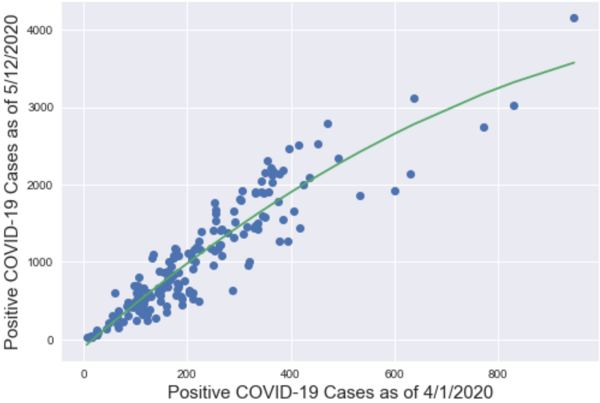
- Results are similar to those in Models 3 and 4
  - **Population**, **Recreation**, and **Restaurant** continue to be statistically significant predictors
- Yet **Transportation** is also now statistically significant at the 5%
  - Important result for policymakers
  - Trenchant measures should be taken to ensure the safety of these areas

Growth Rate in COVID-19 Incidence

- What factors contributed to the growth rate in positive COVID-19 cases over time?
- Models 7 and 8 take the growth rate as dependent variable
  - Calculated over the period 4/1/2020 - 5/12/2020
  - Earliest date data available to time of writing
- No transformation applied to data
- In scatter plot below:
  - Growth is not exponential
  - No logarithmic transformation necessary

Scatter Plot

Figure: Scatter Plot of Positive Cases Over Time with Quadratic Fit



Results

Table: Growth Rate in COVID-19 Cases

	Model 7: OLS			Model 8: OLS		
	Coefficient	Standard Error	T-Statistic	Coefficient	Standard Error	T-Statistic
Population	$7.222e^{-6}$**	$3.37e^{-6}$	(2.143)	$8.148e^{-6}$**	$3.19e^{-6}$	(2.555)
Transportation	-0.023	0.086	(-0.264)	-0.021	0.087	(-0.140)
Market	0.0380	0.045	(0.880)			
Store	0.005	0.023	(0.204)			
Restaurant	-0.020**	0.008	(-2.378)	-0.018***	0.006	(-2.917)
Bar	-0.046*	0.024	(-1.926)	-0.040*	0.021	(-1.923)
Recreation	-0.053***	0.017	(-3.194)	-0.061***	0.012	(-5.157)
Hotel	-0.087	0.091	(-0.953)			

Discussion

- **Population** of a given zip code remains a statistically significant covariate
  - Both Model 6 and 7
- **Restaurant** and **Recreation** are negatively correlated with the growth rate of COVID-19
  - Significant at the 1% level in Model 7
  - Underscores the efficacy of the lockdown

COVID-19 Impact on NYC

- NYC Department of Health and Mental Hygiene (DOHMH):
  - The number of confirmed deaths attributed to the virus is at least 15,253
  - Another 5,051 deaths "probably" due to the same cause
  - Represents 17.3% - 23.0% of all deaths in the United States due to the virus
- Half of all hotels in NYC are not operating
- 186,000 shops employing fewer than 10 people could fail

Insights for Policymakers

- Shuttered businesses are beginning to re-open
  - Will alleviate macroeconomic effects of the lockdown
  - But understanding which activities/venues contribute to virus transmission is crucial
  - Can inform preventative measures
- Number of transportation terminals is a positive, statistically significant predictor
  - In preferred GLM model and Model 4
  - Should be areas of focus
- Restaurants, bars and recreational venues do not have positive, significant impact
  - Regardless of dependent variable