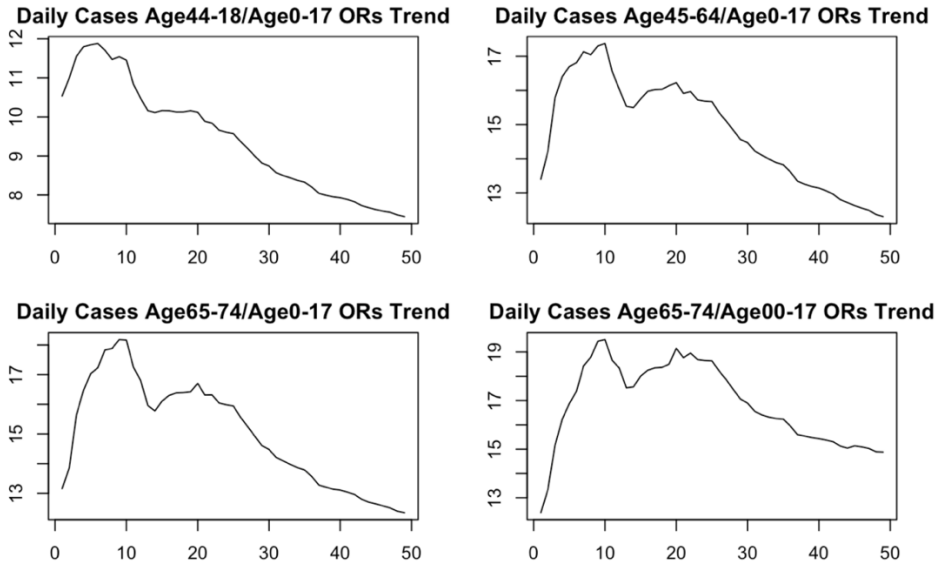
---
title: "writeup"
author: "Xinyuan Zheng xz2906"
date: "5/16/2020"
output: pdf_document

---

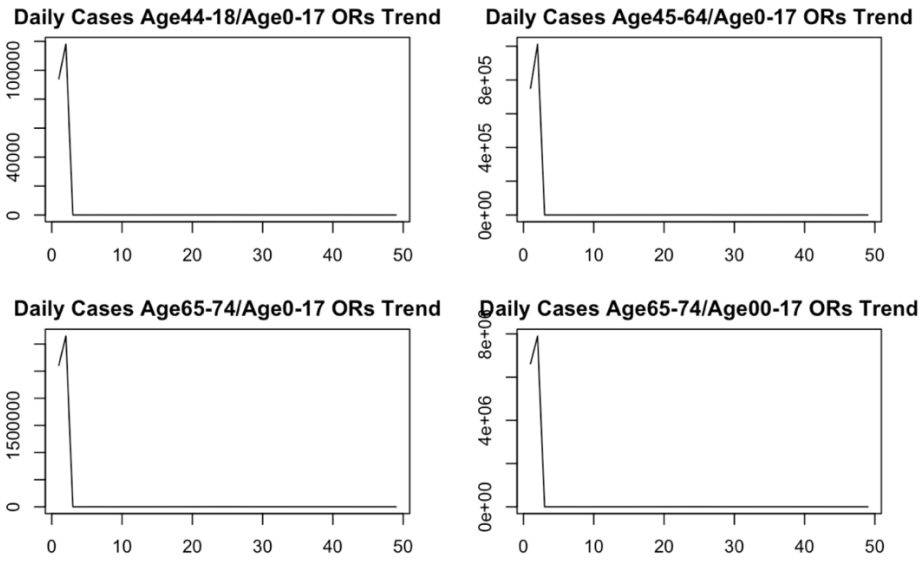
https://github.com/XinyuanZHENG/glm-covid19-project

Part I

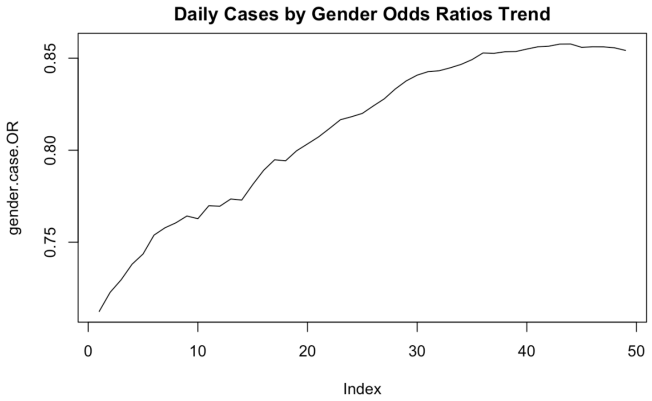
In this project, I first use odds ratios to see the characteristics of the patients of COVID-19 in New York City. Data is from NYC Department of Health and Mental Hygiene [1].

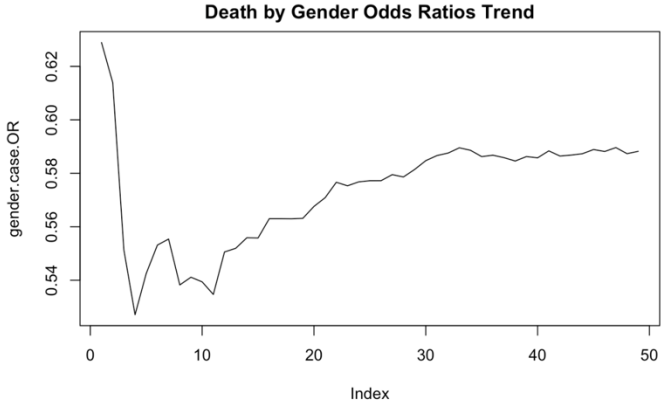


As we can see from the plots, older people are more affected by COVID-19, the odds ratio can be as high as 19 on certain dates. The differences in age group gradually decrease in the later phrase.



There are nearly no deaths in age group 0-17.





The number of male cases is significantly greater than female cases, the number of male deaths is significantly greater than female deaths as well. The trends also gradually decrease.

Part II

I use a few GLMs to fit the covid-19 data of the city. Previously a similar work was done by Kraemer [2] to predict COVID-19 in China.

Predictors

I create some predictors to fit the model.

First, an indicator for stay-at-home order. The order is effective at 8PM on Sunday, March 22 [3]. I also adjust the date by adding the estimated incubation period (5.1 days) [4]. I round up this date to be conservative. The indicator is set to be 0 before the order date plus incubation period, i.e. March 28, and is set to be 1 after March 28.

Second, an indicator for testing criteria. New York expanded testing criteria for COVID-19 on April 25 [5]. This event could affect the number of cases. The indicator is set to be 0 before April 25 and 1 after April 25.

I also create some lagged daily cases and deaths.

Dataset

The dataset is divided into train and test (9:1).

Variable and Model Selection

I run some time series analysis to get familiar with the structure of the dataset. Then a simple Poisson GLM including all predictors is built to see if some predictors are not significant. I also build several GLMs including subsets of the predictors and choose appropriate candidates based on AIC. In this step I consider first-order interaction terms.

After feature selection, I use package “glmnet” to fit the model.

While adding rigid penalty terms, I use 10-fold cross validation to choose lambda and fit the GLM. The optimal lambda gives a CV error of 704375.3. The sum of squared residuals on test set is 2694339.

While adding LASSO penalty terms, again I use 10-fold cross validation to choose lambda and fit the model. The optimal lambda gives a CV error of 674864.3. The sum of squared residuals on test set is 1275282.

LASSO gives a better result in this case.

Summary

Odds ratios can be used to reveal the epidemiological characteristics of COVID-19 in NYC.

Using age group 0-17 and female gender group as baselines, we can clearly see the significant differences between demographic groups.

GLMs are useful in predicting the cases in NYC.

At any reasonable significance level, the stay-at-home order largely reduces the number of cases in NYC. Expanding test criteria significantly increase the number of confirmed cases in the city.

Reference

[1] https://github.com/nychealth

[2] M. U. G. Kraemer *et al.*, "The effect of human mobility and control measures on the COVID-19 epidemic in China," *Science*, Vol. 368, Issue 6490, pp. 493-497, doi:10.1126/science.abb4218 (2020).

[3] https://www.governor.ny.gov/sites/governor.ny.gov/files/atoms/files/EO202.6.pdf

[4] S. A. Lauer, K. H. Grantz *et al.*, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application," *Annals of Internal Medicine*, 2020;172:577-582. doi:10.7326/M20-0504

[5] https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-announces-expansion-diagnostic-testing-criteria