Analyze Outbreak Association and Fatality of COVID-19 Cases in Toronto

Jingjing Zhan 1002898340

Dec 22, 2020

Link to github repository: https://github.com/isabelzjj/STA304-Final-Project

Abstract

COVID-19 is a global health concern which has taken more than a million lives and infected over 70 million worldwide ("COVID-19 CORONAVIRUS PANDEMIC,"n.d.). As a contagious disease which causes death (Ries, 2020), both scientists and everyday people are concerned about its mechanism. In this report, we explain our analysis in the relationship between whether a COVID-19 case is associated with a COVID-19 outbreak and whether the case has fatal outcome.

keywords

COVID-19, Outbreak, Fatal Cases, Logistic Regression, Propensity Score Matching, Toronto

Introduction

From the first case appearing on this planet to it being a global public health concern, it only took COVID-19 a few months (Muccari, Chow & Murphy, 2020). It changed everyone's lifestyle in a short amount of time. With its fast transmission and possible fatal outcome (Ries, 2020), this completely new virus throws human in a dark place where humans don't know how the virus actually works (Gallagher, 2020). And the possibility of COVID-19 taking the life of our loved ones also injected fear to almost everyone.

COVID-19 cases can be classified into two categories: outbreak associated or sporadic (Toronto Public Health, 2020). The outbreak associated cases are cases happened in "Toronto congregate settings" (Toronto Public Health, 2020), such as "hospitals", "long-term care homes", "homeless shelters" etc (Toronto Public Health, 2020). And the sporadic cases are not associated with COVID-19 outbreaks (Toronto Public Health, 2020). With this characteristic being different in different cases, it may help us study what is associated with fatality of COVID-19.

The dataset "COVID-19 Cases in Toronto" (Toronto Public Health, 2020) is used in this analysis. The dataset and its details will be explained in Data section. The model used for analysis and its details will be explained in Model section. The results will be conveyed in Results section. And the conclusion, weaknesses, next steps will be talked about in Discussion section.

Data

The dataset we used for this analysis is observational data of all confirmed or probable COVID-19 cases in Toronto from the start of the pandemic in Toronto (Toronto Public Health, 2020). It has the infected individual's biological information, geographic information and the case-specific information such as "reported date", "Source of Infection" etc (Toronto Public Health, 2020).

We chose and cleaned up the following variables from the dataset to study the correlation between COVID-19 outbreak associated cases and fatal outcome in Toronto.

Table 1: Table 1

Variable	Value	Code
Age Group	19 and younger	1
Age Group	20 to 29 Years	2
Age Group	30 to 39 Years	3
Age Group	40 to 49 Years	4
Age Group	50 to 59 Years	5
Age Group	60 to 69 Years	6
Age Group	70 to 79 Years	7
Age Group	80 to 89 Years	8
Age Group	90 and older	9
Client Gender	MALE	1
Client Gender	FEMALE	2
Client Gender	TRANSGENDER	3
Client Gender	OTHER	4
Ever Hospitalized	Yes	1
Ever Hospitalized	No	0
Outbreak Associated	Outbreak Associated	1
Outbreak Associated	Sporadic	0
Outcome	FATAL	1
Outcome	ACTIVE or RESOLVED	0

We chose variable "Outcome", assigned code "outcome_code" to it and used it as our outcome variable. If "Outcome" was "FATAL", we assigned outcome_code = 1 to it, otherwise, we assigned 0 to it. We chose variable "Outbreak Associated", assigned code "outbreak_associated_code" to it and used the code as our treatment variable. If "Outbreak Associated" variable had value "Outbreak Associated", we assigned outbreak_associated_code = 1 to it and it was considered as "treated". If "Outbreak Associated" was "Sporadic", we assigned outbreak_associated_code = 0 to it.

Since the dataset contains observational data, we couldn't assign people to treatment group ("Outbreak Associated") or control group ("Sporadic") as we could have done in an experiment. So we chose the predictor variables "Age Group", "Client Gender", "Ever Hospitalized" to help us match observations with similar settings. For the predictor variable "Age Group", if its value was '19 and younger", code"age_group_code = 1" was assigned to it; if its value was "20 to 29 Years", code 2 was assigned to it; if its value was "30 to 39 Years", code 3 was assigned to it; if its value was "40 to 49 Years", code 4 was assigned to it; if its value was "50 to 59 Years", code 5 was assigned to it; if its value was "60 to 69 Years", code 6 was assigned to it; if its value was "70 to 79 Years", code 7 was assigned to it; if its value was "80 to 89 Years", code 8 was assigned to it; if its value was "90 and older", code 9 was assigned to it.

For predictor variable "Client Gender", if its value was "MALE", code "gender_code = 1" was assigned to it; if its value was "FEMALE", code 2 was assigned to it; if its value was "TRANSGENDER", code 3 was assigned to it;if its value was "OTHER", code 4 was assigned to it. For predictor variable "Ever Hospitalized", which indicates whether the individual was ever or currently being hospitalized (Toronto Public Health, 2020),

if its value was "YES", code 1 was assigned to it; if its value was "NO", code 0 was assigned to it.

We looked at pairs of cases where the infected individuals had similar background information for those three predictor variables but one case in the pair was treated (the individual's case was associated with COVID-19 outbreak) and the other case in the pair was not treated (the individual's case was not associated with COVID-19 outbreak). And then we looked at the outcome, whether the case was reported fatal to study the relationship between COVID-19 outbreak associated cases and fatal outcome.

Model

We needed two models in this analysis, one was used for perform Propensity Score Matching and the other one was used to study the relationship of COVID-19 outbreak associated cases and fatal outcome. We chose logistic regression model for both of them.

For Propensity Score Matching, we had the following logistic regression model (denoted as "model 1") (Pruim, 2016):

$$\log(\frac{p_1}{1-p_1}) = \beta_0 + \beta_1 X_{ageGroupCode} + \beta_2 X_{genderCode} + \beta_3 X_{everHospitalizedCode}$$

In this model, p_1 is the probability of the case being associated with COVID-19 outbreak (Caetano, 2020), $X_{ageGroupCode}$ is the age code we assigned to different age groups. $X_{genderCode}$ is the code we assigned to the individual's gender. $\beta_3 X_{everHospitalizedCode}$ is the code we assigned to variable "Ever Hospitalized". For example, if a case with the infected individual being a 75 year-old male and was hospitalized due to COVID-19, then it has $X_{ageGroupCode} = 7$, $X_{genderCode} = 1$, $\beta_3 X_{everHospitalizedCode} = 1$ and the calculated p_1 will be the probability of this case being associated with COVID-19 outbreak.

Since "Outbreak Associated" variable is binary, and age, gender, hospitalized experience are useful information to identify and "explain" (Alexander, 2020) a COVID-19 infected individual, logistic regression model are the best choice to match observations here.

For studying the relationship between COVID-19 outbreak associated cases and fatal outcome, we had the following logistic regression model (denoted as "model 2"):

$$\log(\frac{p_2}{1-p_2}) = \beta_0 + \beta_1 X_{ageGroupCode} + \beta_2 X_{genderCode} + \beta_3 X_{everHospitalizedCode} + \beta_4 X_{outbreakAssociated}$$

In this model, p_2 is the probability of the case being reported as fatal outcome (Caetano, 2020). $X_{ageGroupCode}$, $X_{genderCode}$, $\beta_3 X_{everHospitalizedCode}$ are the same as the variables in model 1. $\beta_4 X_{outbreakAssociated}$ is the code we assigned to variable "Outbreak Associated". For example, if a case with the infected individual being a 75 year-old male, hospitalized due to COVID-19, associated with COVID-19 outbreak in a long-term care home, then it has $X_{ageGroupCode} = 7$, $X_{genderCode} = 1$, $X_{everHospitalizedCode} = 1$, $X_{outbreakAssociated} = 1$ and the calculated p_2 will be the probability of this case being reported having fatal outcome.

Since the outcome that the case reporting fatal is binary, and variables age, gender, hospitalized experience, outbreak associated are all characteristics of an individual instead of some organization (Caetano, 2020), logistic regression model are the best choice to study relationship between COVID-19 outbreak associated cases and fatal outcome.

Results

Figure 1

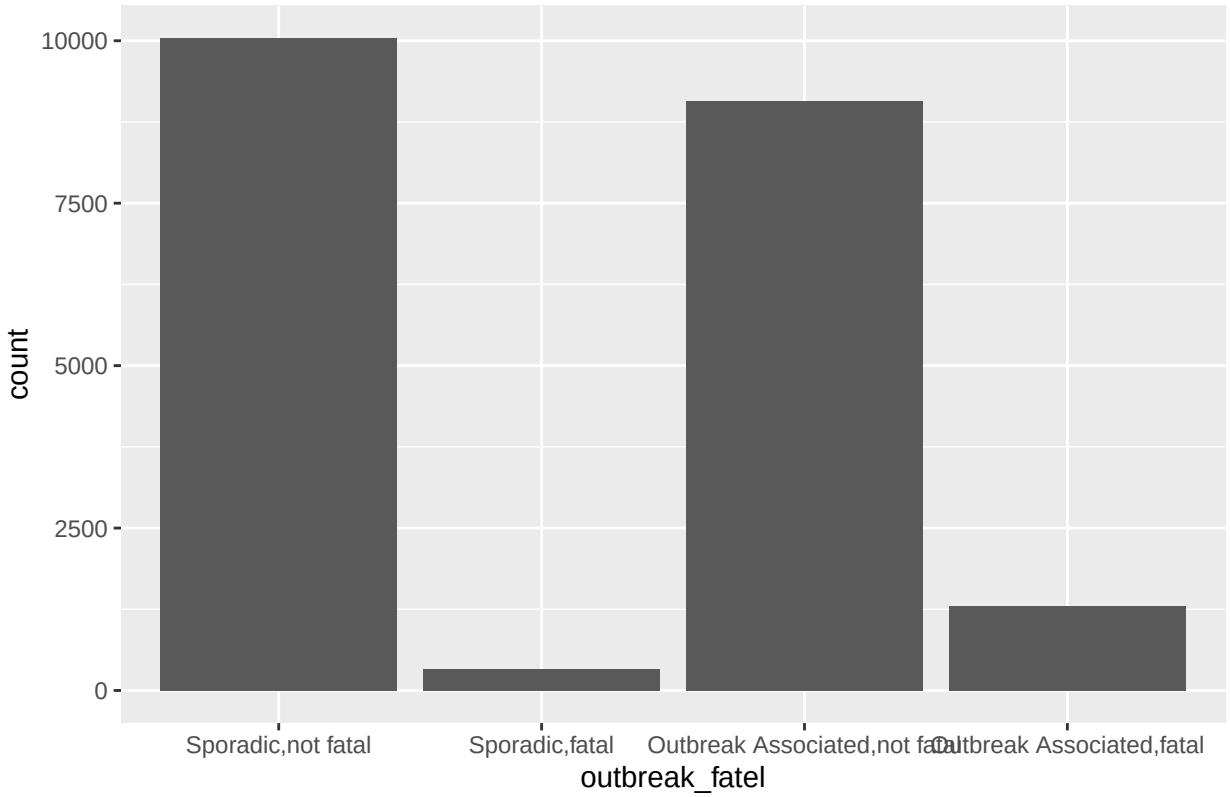

Table 2: Table 2: Summary Results

Variable	Estimate	p-value
outbreak_associated_code	1.37	< 2e-16

In Figure 1, the number of "Outbreak Associated, fatal" cases are around 1250 which is a lot higher than the number of "Sporadic, fatal" cases, which are around 300.

In Table 2, the estimated β_4 has value 1.37, the p-value of variable "outbreak_associated_code" is < 2e-16, which is way smaller than 0.5, using our logistic regression model (model 2) with predictor variables age_group_code, gender_code, ever_hospitalized_code, outbreak_associated_code. (Caetano, 2020).

Discussion

Based on the results of the logistic regression model used to study relationship between COVID-19 outbreak associated cases and fatal outcome, there is a positive correlation (since β_4 has value 1.37) between a COVID-19 infected case being outbreak associated and the outcome being reported fatal. If the COVID-19 case in Toronto is associated with a COVID-19 outbreak, it is more likely to have fatal outcome, compared to sporadic COVID-19 cases.

Weaknesses:

The dataset contains all the confirmed and probable cases and we included all of them in our analysis. According to Ontario Ministry of Health website, probable cases may not have a confirmed positive laboratory result for COVID-19 ("Case Definition – Coronavirus Disease (COVID-19),"n.d.). So we might have included cases where the outcome was reported fatal but it was not caused by COVID-19. Whether this case was associated with an COVID-19 outbreak or sporadic, we should not have included it in our analysis.

Since the number of "Resolved" cases, which have non-fatal outcome (Toronto Public Health, 2020) was "underreported due to a lag in data entry" (Toronto Public Health, 2020), then the number of cases with "outcome_code = 0" may not be accurate in our analysis. Also, we are still in the pandemic and the Toronto COVID-19 cases data are being updated every week (Toronto Public Health, 2020), we can not draw conclusion about the relationship of how deadly COVID-19 is as a disease in Toronto and whether the case is associated with an outbreak.

The dataset had 48500 observations after we cleaned up the code, but after performing Propensity Score Matching, there are only 20730 observations left, which is less than half of the original size. The reduction of the size might have caused incorrect results in our analysis.

Next Steps:

We can make our analysis work on a weekly or a monthly basis since COVID-19 is still an ongoing pandemic and we can compare the results of those analysis to draw valid conclusions. Also, we can include COVID-19 cases in a larger geographic region since COVID-19 is a global health concern. Cases may have its regional characteristics and cases and their outcome in different geographic regions may differ a lot. ("Same Virus, Different Countries,"n.d.).

References

Alexander, R. (2020, November 05). Difference in differences. Retrieved December 22, 2020, from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html

Bar charts. (n.d.). Retrieved December 22, 2020, from https://ggplot2.tidyverse.org/reference/geom_bar.html

Caetano, S. (2020). Final Project - Additional Instructions [Class handout]. Retrieved from University of Toronto, St. George campus STA304. Final Project, 2020.

Caetano, S. (2020). LogitRegression-Estimation-code.R [Class handout]. Retrieved from University of Toronto, St. George campus STA304. Week 4 - Ratio & Regression, 2020.

Caetano, S. (2020). Matching-PropensityScore-Amazon.R [Class handout]. Retrieved from University of Toronto, St. George campus STA304. Week 10 - Causality in Observational Studies, 2020.

Caetano, S. (2020). RTidyverse-Intro.R [Class handout]. Retrieved from University of Toronto, St. George campus STA304. Week 1 - Foundations, 2020.

Caetano, S. (2020). STA304 - Logistic Regression Intro.pdf [PDF Slides]. Retrieved from University of Toronto, St. George campus STA304. Week 4 - Ratio & Regression, 2020.

Caetano, S. (2020). STA304 - Multilevel.pdf [PDF Slides]. Retrieved from University of Toronto, St. George campus STA304. Week 6 - Multilevel Regression & Postratification, 2020.

Case Definition – Coronavirus Disease (COVID-19). (n.d.). Retrieved December 22, 2020, from http://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/2019_case_definition.pdf

Christopher H. Jackson (2011). Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software, 38(8), 1-29. URL http://www.jstatsoft.org/v38/i08/.

Convert a Numeric Object to Character in R Programming - as.character() Function. (n.d.). Retrieved December 22, 2020, from https://www.geeksforgeeks.org/convert-a-numeric-object-to-character-in-r-programming-as-character-function/

COVID-19 CORONAVIRUS PANDEMIC. (n.d.). Retrieved December 22, 2020, from https://www.worldometers.info/coronavirus/

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

Ebejer, J. (2013, July 16). Citing R packages in your Thesis/Paper/Assignments. Retrieved December 22, 2020, from https://www.blopig.com/blog/2013/07/citing-r-packages-in-your-thesispaperassignments/

Gallagher, J. (2020, October 22). Covid: Why is coronavirus such a threat? Retrieved December 22, 2020, from https://www.bbc.com/news/health-54648684

Gelfand, S., & City of Toronto. (n.d.). Opendatatoronto. Retrieved December 22, 2020, from https://sharlagelfand.github.io/opendatatoronto/

Ggplot2 axis ticks : A guide to customize tick marks and labels. (n.d.). Retrieved December 22, 2020, from http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels

Ggplot2 title : Main, axis and legend titles. (n.d.). Retrieved December 22, 2020, from http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

Grossetti, F. (n.d.). Augment. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/msmtools/versions/1.3/topics/augment

Hayes, A. (n.d.). Broom v0.7.3. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/broom/versions/0.7.3

Open Data Catalogue. (n.d.). Retrieved December 22, 2020, from https://open.toronto.ca/catalogue/?search=covid

Pruim, R. (2016, October 19). Mathematics in R Markdown. Retrieved December 22, 2020, from https://rpruim.github.io/s341/S19/from-class/MathinRmd.html

Matrix Construction. (n.d.). Retrieved December 22, 2020, from http://www.r-tutor.com/r-introduction/matrix/matrix-construction

Muccari, R., Chow, D., & Murphy, J. (2020, March 10). Coronavirus timeline: Tracking the critical moments of Covid-19. Retrieved December 22, 2020, from https://www.nbcnews.com/health/health-news/coronavirus-timeline-tracking-critical-moments-covid-19-n1154341

NA. (n.d.). Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/NA

Numeric. (n.d.). Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/numeric

Operators. (n.d.). Retrieved December 22, 2020, from https://www.statmethods.net/management/operators.html

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ries, J. (2020, March 12). Here's How COVID-19 Compares to Past Outbreaks. Retrieved December 22, 2020, from https://www.healthline.com/health-news/how-deadly-is-the-coronavirus-compared-to-past-outbreaks

Same Virus, Different Countries. (n.d.). Retrieved December 22, 2020, from https://www.uhnresearch.ca/news/same-virus-different-countries

Subset rows using column values - filter. (n.d.). Retrieved December 22, 2020, from https://dplyr.tidyverse.org/reference/filter.html

Toronto Public Health. (2020, December 16). COVID-19 Cases in Toronto. Retrieved December 22, 2020, from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, H. (n.d.). Case_when. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/case_when

Wickham, H. (n.d.). Count. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/plyr/versions/1.8.6/topics/count

Wickham, H. (n.d.). Filter. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter

Xie, Y. (n.d.). Kable. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/knitr/versions/1.30/topics/kable

Xie, Y., Dervieux, C., & Riederer, E. (n.d.). R Markdown Cookbook. Retrieved December 22, 2020, from https://bookdown.org/yihui/rmarkdown-cookbook/kable.html