Analyze COVID-19 Cases in Toronto —— Is it useful to close customs?

Xueqi Wang

2020, Dec, 22nd

Abstract

A virus is a disease that spreads quickly. The global outbreak of COVID19 this year has made everyone nervous.Countries all over the world must strictly require their own countries to curb the spread of the virus. From the initial spread to the later spread, this is completely new data for all of us.To this end, we need statistics to analyze the data of the existing cases and analyze all kinds of important information.Toronto was closed to varying degrees in October and November, and customs was once closed. But now, compared with October, the number of infections is still rising. This report analyzes that Toronto needs to continue to close customs from the data of the infected population, age, gender, and infection method by using the logistic linear regression.

Keywords

COVID19, Gender, Age Group, Hospitalize, Source of Infection, Toronto, data analyze, logistic

Introduction

The project is based on the data of COVID-19 cases in Toronto and grouped by Toronto Public Health and making the model to analyze my topic. the first cases was reported in January 2020 and the data last refreshed on Dec,2020. The data collect the person lots of information as variables, such as gender, age group. source of infection, etc.

I worked on this data before during October 2 as my STA304 problem set 1, which the data was reported in January 2020 and the data last refreshed on Sep 30, 2020. Based on my old data analyze report, I was focus on that the male between age 20-29 is the most value that has COVID-19, and close contact and outbreak associated became two most source of infection. After 2 months, the situation is getting worse. The city is lockdown again during Nov 20. This is a good opportunity to re-improve my last inference.

My project is defined as the extent to describe one of the relationships in the dataset. We will discuss the topic of “Is it useful to close customs?”. This report is actually helping me to get understanding the detail information of COVID-19 cases, I can use this data result to explain to my parents and friends. The data has clear and detail variables for an analysis, it is a real-time data during this year which the result is useful in real-time. I will be using logistic linear regression to get the relationship of Ever Hospitalized variable and dummy variables(age group, gender, Source of Infection) and making the analyze of that.

Data

By checking the data set, it has 18 different kind of variables. In my report, I would like to narrow my scope of data set. The target population will be focus on the 20-29 years old Toronto living people that has been hospitalized by Covid-19 and the source of infection is ‘Travel’. For this analysis, I clean the whole data by using filter() and

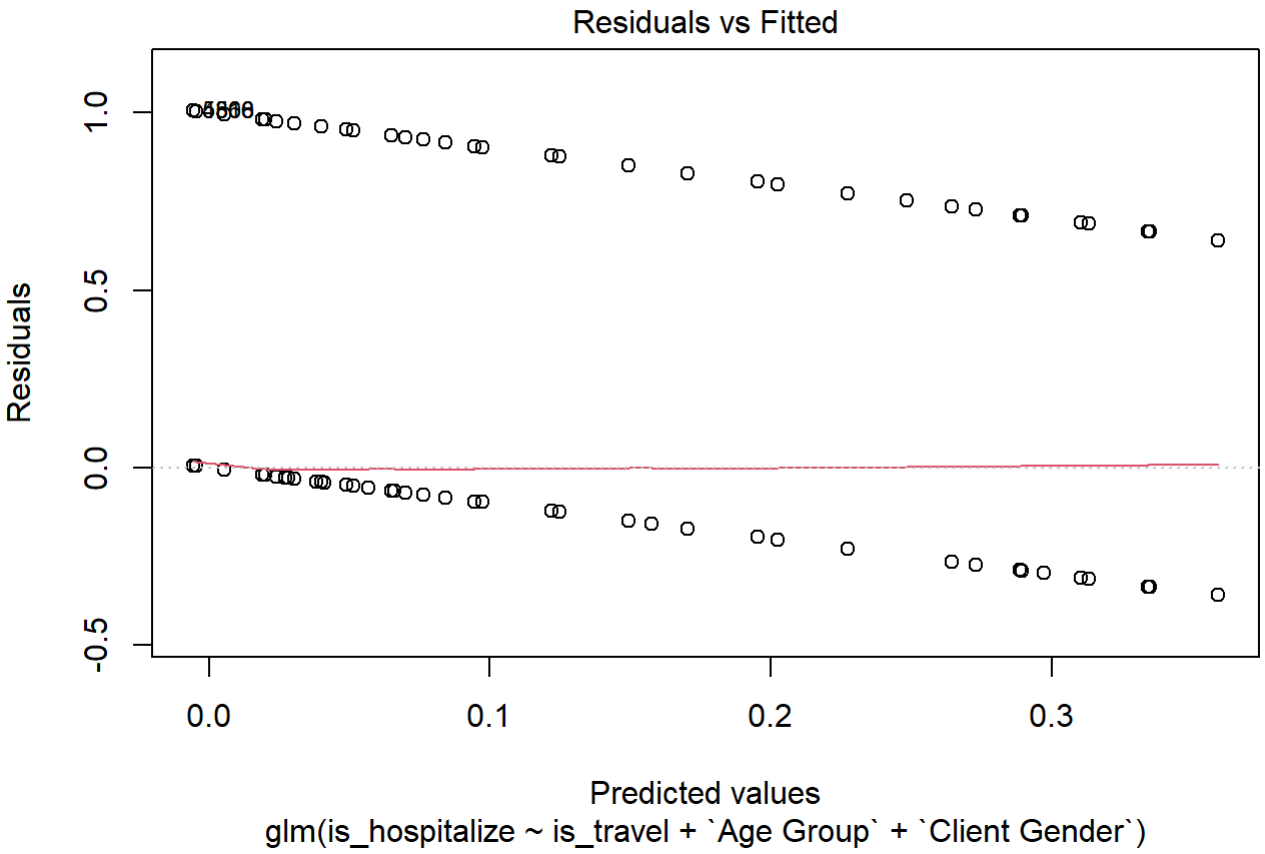
mutate() in r-code. For detail, I mutate the variable ‘source of infection’ by only choosing the ‘travel’, then named as “is_travel”. And repeat the step for ‘ever hospitalized’ variable, yes=1 and no=0, then named as “is_hospitalize”.

```
##
## Call:
## glm(formula = is_hospitalize ~ is_travel + `Age Group` + `Client Gender`,
##     data = reduced_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.35921  -0.07623  -0.02410  -0.00565   1.00558
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.005583   0.003358  -1.663  0.09637 .
## is_travel                  0.045797   0.007301   6.273 3.58e-10 ***
## `Age Group`20 to 29 Years  0.001031   0.003972   0.260  0.79522
## `Age Group`30 to 39 Years  0.011237   0.004138   2.715  0.00663 **
## `Age Group`40 to 49 Years  0.029679   0.004269   6.952 3.63e-12 ***
## `Age Group`50 to 59 Years  0.057057   0.004242  13.451  < 2e-16 ***
## `Age Group`60 to 69 Years  0.130364   0.004743  27.486  < 2e-16 ***
## `Age Group`70 to 79 Years  0.270101   0.005770  46.814  < 2e-16 ***
## `Age Group`80 to 89 Years  0.294212   0.005644  52.125  < 2e-16 ***
## `Age Group`90 and older    0.208175   0.006583  31.625  < 2e-16 ***
## `Client Gender`MALE        0.024784   0.002178  11.380  < 2e-16 ***
## `Client Gender`OTHER       0.032768   0.066724   0.491  0.62336
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.05784141)
##
##     Null deviance: 3260.2  on 49491  degrees of freedom
## Residual deviance: 2862.0  on 49480  degrees of freedom
##   (34 observations deleted due to missingness)
## AIC: -588.48
##
## Number of Fisher Scoring iterations: 2
```

Model&Results

In this part, I used glm() in R to create a logistic regression model to find the probability of how is admission related. I use is_travel, Age group and client gender to be the dummy variable also called explanatory variables. Those two main personal information might be an important factor to affect the probability of whether hospitalized or not. Let P be the probability of people has hospitalized if they got covid-19. And let the Age group to be X_1 to X_8 since it has 8 different groups, the female to be X_9 and male to be X_{10}. The coefficient of X_1 is $\hat{\beta}_1$, coefficient of X_2 is $\hat{\beta}_2$ and so on. When we see the p-value for the null hypothesis that all $\hat{\beta}$ value are lower than 0.3. So we can reject these null hypothesis value.

Figure 1



For model Figure1, the red line is actually stay in the range of residuals, and the line is very close to straight line. Which also represents that all of them are actually related to the proportion of hospitalization.

Figure 2

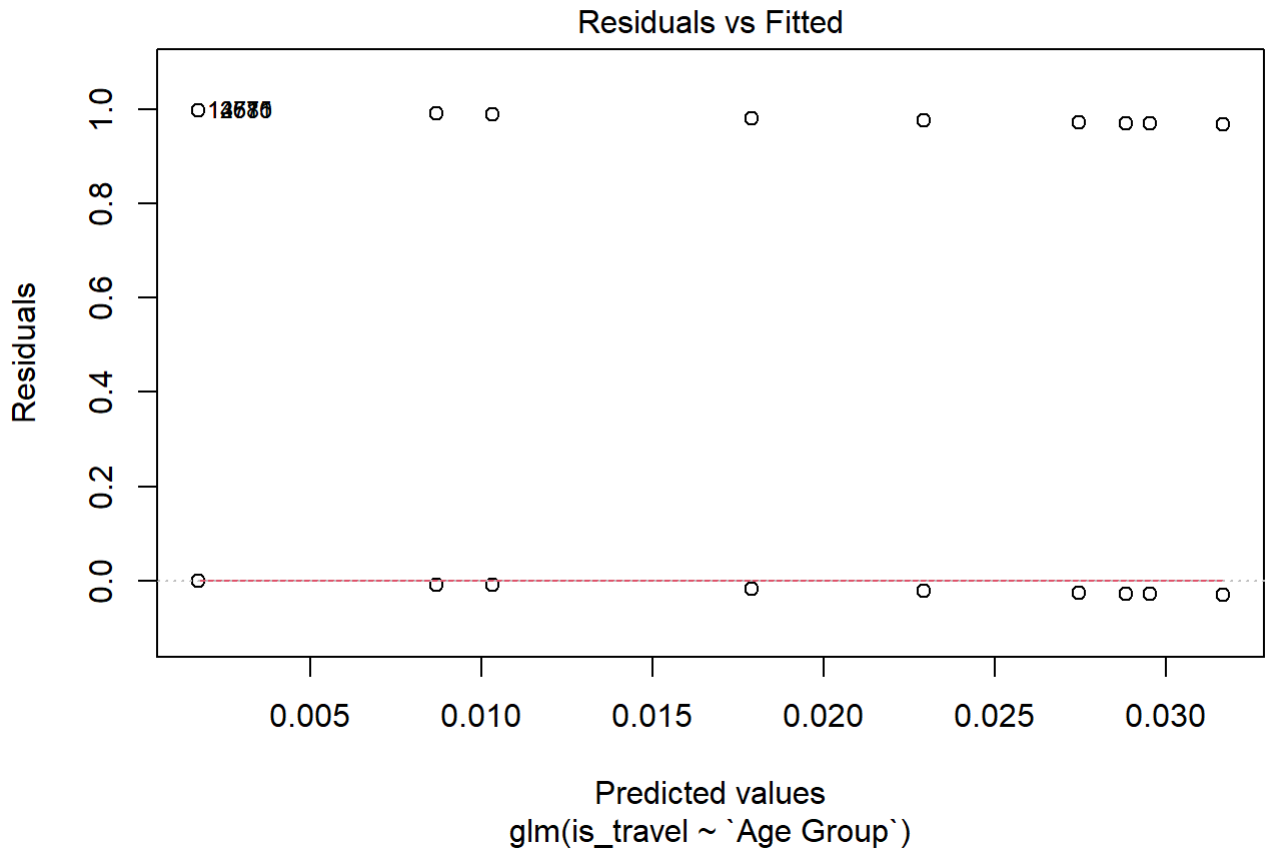
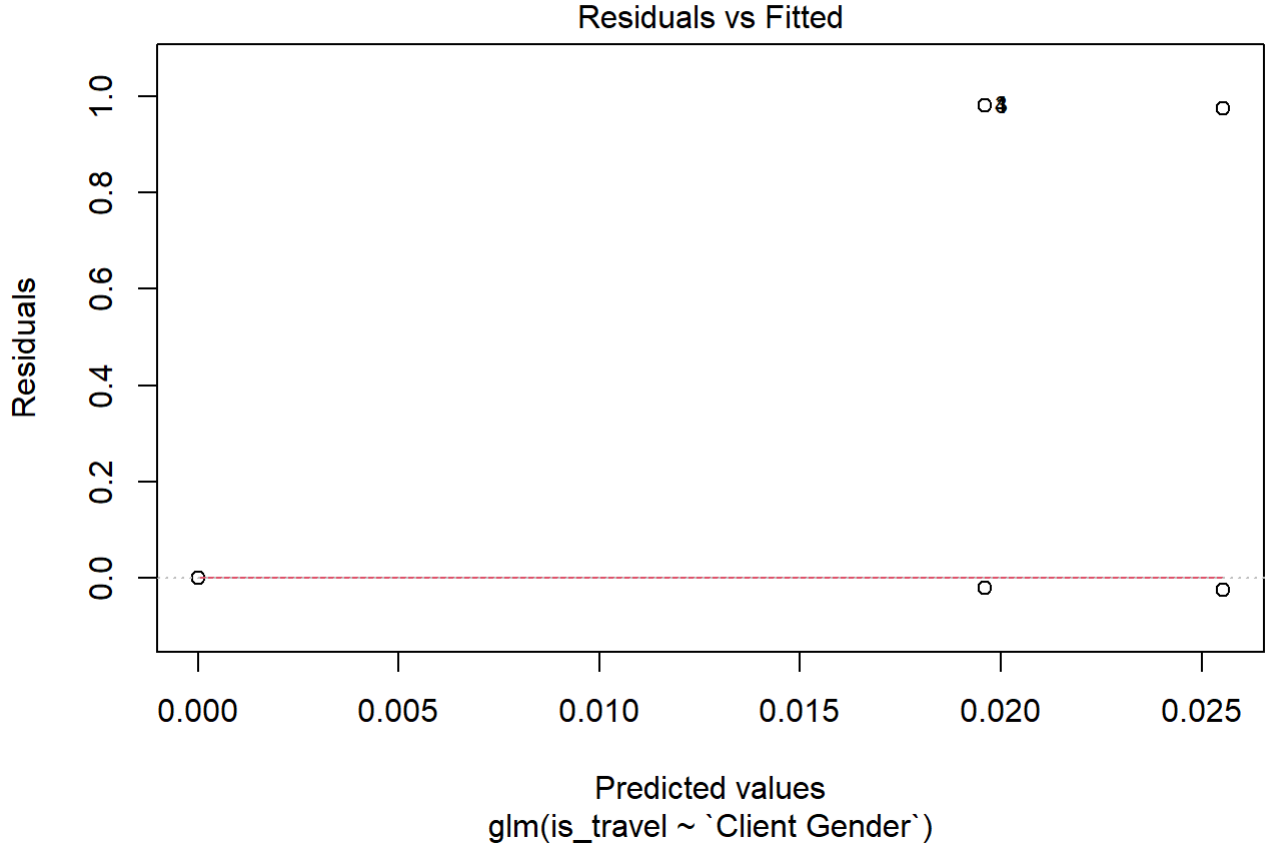


Figure 3



For the alternative model, I create two linear regression model for dummy variable as Figure2&3, first is the model for is_travel and Age Group(Figure2) and second is for is_travel ans Client Gender(Figure3). Both of two QQ plot does not follow the normal distribution, which means that these dummy variables are not related to each other. Because they were dummy variable for is_hospitalize, so even if they are not related, it is reasonable and expected.

Discussion

The result of my report by the statistic, r-code and logistic linear regression model, targeting the population of Toronto people who got COVID19 in 20-29 years old, has been hospitalized cause traveling. The result shows that it does has relationship between them. The government should close the customs, it is useful. The model as a small world can be analyze in my report, while it also can be represent the large world, which means not only Toronto, the whole world should be close the customs, which can more effectively prevent the spread of the virus.

-Weaknesses

This report is based on the result of my own problem set 1, the model is limited by the quality of data, which means the data might be not enough for me to get the result. Also in the gender variable, I delete those individuals who answer“unknown” and“transgender”, these can not contribute to the analysis.

-Next Steps

The data might be not clear enough, I should find more data to support my report in the future. Based on current data and results, I can make a follow-up survey, which can helps the report to be more improve on the probability of the results obtained. For example, people who traveling between countries or cities. Did they travel alone or with other people. These will give me a clearer understanding and so on.

References

1. Wu, Changbao, and Mary E. Thompson. “Basic Concepts in Survey Sampling.” Sampling Theory and Practice. Springer, Cham, 2020. 3-15.

2.Open dataset,“About COVID-19 Cases in Toronto”, https://open.toronto.ca/dataset/covid-19-cases-in-toronto/,2020 (https://open.toronto.ca/dataset/covid-19-cases-in-toronto/,2020)

3. COVID-19 Pandemic Seen From the Margins -Stefania Milan, Emiliano Treré, 2020. (2020). Retrieved 9 December 2020, https://journals.sagepub.com/doi/full/10.1177/2056305120948233Toronto (https://journals.sagepub.com/doi/full/10.1177/2056305120948233Toronto), Peel move into COVID-19

4. lockdown Monday as Ontario tries to stop ‘worst-case scenario’|CBC News. (2020). Retrieved 9 December 2020, https://www.cbc.ca/news/canada/toronto/covid-19-coronavirus-ontario-november-20-toronto-peel-1.5809575 (https://www.cbc.ca/news/canada/toronto/covid-19-coronavirus-ontario-november-20-toronto-peel-1.5809575)

5. Wang, X. (2020). Crowdmark. Retrieved 22 December 2020, from https://app.crowdmark.com/score/ddc6a7ca-b48e-40a2-9233-c36e7e95be5d (https://app.crowdmark.com/score/ddc6a7ca-b48e-40a2-9233-c36e7e95be5d)