Analysing the Affects of Age Groups on the Outcomes of COVID-19 Using Propensity Score Matching

Peiyu Li

December 21, 2020

Code and data supporting this analysis is available at:

https://github.com/PeiyuBellaLi/STA304-Final-Report

Abstract

This report focuses on if older people have a high risk of mortality once they are infected by COVID-19. The cutoff of the age to be considered as older or not is suggested to be 60 by some evidence. The sample contains case details in Alberta and Ontario and is selected using propensity score matching. Since the outcomes, in this case, are either recovered or deceased, a logistic regression model is performed. The p-value of the age variable indicates that it has a great impact on the outcome. Other variables such as gender, exposure type, and province are also significant. The ROC curve helps to verify the accuracy of the model, and causal inference is made at the end.

Key words: COVID-19, Observational Study, Propensity Score, Causal Inference, Logistic Regression

Introduction

COVID-19 (coronavirus disease 2019) pandemic is the biggest challenge our society faced in 2020. It is an infectious disease transmitted by a newly found coronavirus. People will experience mild to moderate respiratory illness after infected with the coronavirus, but most people can recover without special treatment. However, older people and those with underlying medical conditions are more likely to experience serious illness or even die (World Health Organization, 2020). COVID-19 has reached Canada since January 2020 (Marchand-Senécal, 2020). It is important to know if age is significantly related with the outcomes after being infected so that people can be more prepared. The conclusion drawn from the observational data can only indicate the association, but the causal inference would be more practical.

In order to make a causal inference, propensity score matching is a popular method (Lanza et al., 2013). It is the probability of treatment assignment conditional on observed base features, and it becomes popular to minimize the confounding effects when dealing with observational data (Austin, 2011). In this report, I will use propensity score matching to decide if I can make a causal inference between whether or not a person is over 60 years old and whether this person is recovered or dead.

According to the Government of Canada (2020), older individuals especially those over 60 years old are at risk of more severe outcomes. There is also other evidence that supports that age increases the risk for severe illness, more specifically, for adults over the age of 60 (National Women's Health Network, 2020). CDC also

reports that 80% of death cases in the U.S. are adults 65 years old and older (Centers for Disease Control and Prevention, 2020). As a result, I choose 60 as a cutoff for the age group and it is used to calculate the propensity score as a treatment in this analysis.

The observational data I use is the COVID-19 case details in Canada collected from the COVID-19 Canada website (Berry et al., 2020). In the Methodology section (Section 2), I describe the data, the model I used to perform the propensity score matching, and the final logistic regression model. The result from propensity score matching and the statistics from the logistic model are presented in the Results section (Section 3). The conclusion and limitations are discussed in the Discussion section (Section 4).

Methodology

Data

The COVID-19 case details data is provided by Esri Canada COVID-19 Data Repository (Berry et al., 2020), and it combines multiple data sets provided by each province in Canada. The goal of this report is to see if age has a significant impact on the outcome of COVID-19. Since only the cases in Ontario and Alberta contains individuals ages, only the observations in these two provinces will be used to do the analysis. In order to draw a causal inference at the end, the propensity score matching is done by using the treatment group (age above or below 60) to calculate the score using other variables except for the outcome (case status, recovered or deceased). This process matches one observation who is above 60 with another observation who is younger than 60 with a similar propensity score.

After matching, there are 55862 observations (27931 pairs) left. The following table (Table 1) summarizes the percentages of each level in each variable based on the treatment group (age above 60 or not). It is obvious that the percent of observations in each level for the variables gender, exposure type, and province are almost the same in each treatment group. The reason behind this is because of the propensity score matching which was calculated based on these three variables. However, percent of people who died after being infected is different in the two age groups. More specifically, 0.3% of observations who are younger than 60 years old are deceased while 13% of observations who are older than 60 are deceased. Whether this difference is significant or not is needed to be verified by fitting a regression model.

	Gender		Exposure		Province		Case Status	
Age under 60 (100%)	Male	45.3%	Close Contact	20.7%	Alberta	14.9%	Recovered	99.7%
	Female	54.1%	Outbreak	38.7%	Ontario	85.1%	Deceased	0.3%
	Not Reported	0.6%	Travel-Related	3.4%				
			Not Reported	37.2%				
Age above 60 (100%)	Male	45.0%	Close Contact	20.7%	Alberta	14.9%	Recovered	87.0%
	Female	54.5%	Outbreak	38.7%	Ontario	85.1%	Deceased	13.0%
	Not Reported	0.5%	Travel-Related	3.4%				
			Not Reported	37.2%				

Table 1: Summary of variables by treatment groups

Model

In order to verify if older people who are above 60 are more likely to die due to COVID-19, logistic regression will be performed using R. The reason why a logistic model is more appropriate is that the dependent

variable, case status, is a categorical variable with only two categories, recovered or deceased. Also, all predictor variables (age, gender, exposure type, province) are all categorical variables in this data set. The logistic model that will be used is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{above60_1} + \beta_2 x_{gender_{Male}} + \beta_3 x_{gender_{NotReported}} + \beta_4 x_{exposure_{NotReported}} + \beta_5 x_{exposure_{Outbreak}} + \beta_6 x_{exposure_{TravelRelated}} + \beta_7 x_{province_{Ontario}} + \epsilon$$

In this equation, p is the proportion of individuals will be died after being infected by COVID-19 and $log(\frac{p}{1-p})$ is the log odds. It is easy to show that if the log odds increase, the probability of being deceased will also increase since $p = \frac{e^y}{1+e^y}$ if y is the value of the log odds.

The baseline in this model is the observations who are under 60 years old female, infected due to close contact, and live in Alberta. The β_0 represents the log odds of the probability that an individual with the baseline characteristic will die due to COVID-19. Similarly, β_1 to β_7 indicates the changes in the log odds if a person's age group, gender, exposure type, and province are different from the baseline.To evaluate the performance of the model, a ROC curve (Figure 1) will be shown in the next section. The area under the ROC curve indicates the accuracy of the model (Rawat, 2017).

Results

The summary of the model for each variables is presented in the following table (Table 2) after fitting the logistic regression model in R.

	(Intercept)	above60_1	gender_Male	gender_Not Reported
Estimate	-8.24170	4.19542	0.38746	0.09122
Std. Error	0.16671	0.12125	0.03769	0.21270
P-value	< 2e-16	< 2e-16	< 2e-16	0.668
Significance	***	***	***	
	exposure_Not Reported	exposure_ Outbreak	exposure_Travel Related	province_ Ontario
Estimate	0.62999	2.13811	0.56862	0.67226
Std. Error	0.08178	0.07024	0.14309	0.09273
P-value	1.33e-14	< 2e-16	7.07e-05	4.18e-13
Significance	***	***	***	***

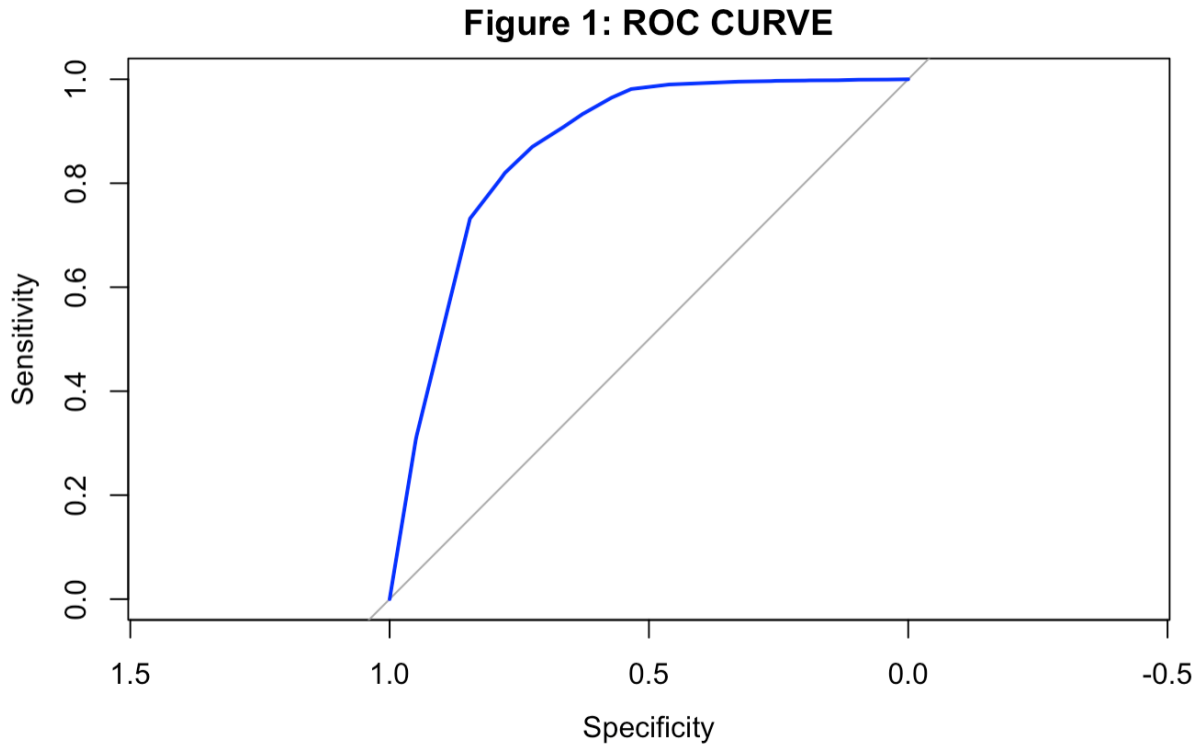
Table 2: Summary of the model

The equation of this model becomes:

$$log(\frac{p}{1-p}) = -8.24 + 4.19 x_{above60_1} + 0.38 x_{gender_{Male}} + -0.09 x_{gender_{NotReported}} + 0.63 x_{exposure_{NotReported}} + 2.14 x_{exposure_{Outbreak}} + 0.57 x_{exposure_{TravelRelated}} + 0.67 x_{province_{Ontario}}$$

The baseline is still a female who is younger than 60 and living in Alberta, and is infected due to close contact, then the log odds of the probability that she will die is -8.24. The remaining coefficients indicate the changes in the log odds. For example, if another observation is an over 60-year-old male living in Ontario and is infected because of the outbreak, then the log odds of the probability of he will die will be increased 4.19 by the variable age group, 0.38 by gender, 2.14 by exposure type, and increased 0.67 by province.

The accuracy of this model can be checked using the ROC curve (Figure 1). The ROC curve evaluates the performance of the logistic model. The area under the blue line indicates the accuracy of this model, and the area is 0.8676. This means the model is 86.76% accurate.



The main goal of this analysis is to see if older people are more likely to die because of COVID-19. From Table 2, if a person is over 60 years old, then the log odds increase 4.19 compare to an individual who is below 60 years old. The coefficient for this variable is very significant according to its p-value in Table 2. Also, except for the variable gender that was not reported, all other variables are significant and have a meaningful impact on the log odds. Therefore, a causal conclusion can be drawn from this model. If a person is older than 60, then the probability that this person will die after being infected by COVID-19 will increase significantly because the log odds will increase. Other variables include gender, exposure type, and which province the person lives in will also influence this probability.

Discussion

Summary

Based on the COVID-19 case details data in Canada, a regression model analysis was done driven by the question if people older than 60 are more likely to die due to COVID-19 as some pieces of evidence showing. After cleaning the data, only cases reported in Ontario and Alberta provided a specific age group and were used to do this analysis. In order to make a causal inference later, the propensity score matching was done. Then a logistic regression model was performed in R, and the accuracy of this model was checked by the ROC curve.

Conclusions

Before fitting the logistic model, one of the patterns in Table 1 shows that there is a big difference in the proportion of people who were infected by the coronavirus and died between people who are above or below

60 years old. This difference is proved to be significant after fitting the logistic model, and the results indicate the log odds of being deceased increase a lot if the observation is older than 60, versus younger than 60. In this model, the other three variables also show a significant impact on the result. They suggest that being a male, living in Ontario, and being infected due to the outbreak or travel will also increase the risk of mortality.

The ROC curve confirms that this logistic model is accurate enough to make a conclusion and prediction. In addition, because this is also a propensity score analysis, causal inference can be made even though the data is observational. Thus, the conclusion for the previous question is being older than 60 years old will significantly increase the probability of mortality.

Weakness & Next Steps

Although the sample size is large and contains the case details in two provinces, it is still not enough to make the conclusion only based on the data in two of the provinces in Canada. Also, other variables that are related to the individuals who were infected are limited. Thus, the propensity scores for many observations are the same since it is calculated only based on their provinces, exposure type, and gender. However, these problems are hard to avoid due to the confidentiality of COVID-19 cases. In addition, some observations did not provide their age, gender, and exposure type information. This may also result in a biased result. Another problem is that the propensity score matching is helpful to make causal inference based on observational data, but this is not perfect since it drops the observations that cannot be matched.

In future studies, the age variable can be divided into more age groups, versus simply using the cutoff of 60. This cutoff was suggested by some of the evidence mentioned in the Introduction section, but there might be other cutoff selections. However, it will make more sense if there are more age groups and doing the propensity score matching multiple times. As the pandemic continues, the sample size will be larger, and there will be more case detail information provided. It will be also interesting to look at the relationship between the length of the recovery period and age.

During the pandemic, some people are anxious due to some 'rumors', but there are also some people who do not realize the seriousness of this coronavirus. It is important and meaningful to use statistical methods to verify some of the results.

References

Alexander, R. (2020). Running Through a Propensity Score Matching Example.
https://q.utoronto.ca/courses/184060/modules/items/1998838

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. https://doi.org/10.1080/00273171.2011.568786

Berry, I., Soucy, J., Tuite, A.,& Fisman, D. (2020). Open access epidemiologic data and an interactive dashboard to monitor the COVID-19 outbreak in Canada.
https://doi.org/10.1503/cmaj.75262

Centers for Disease Control and Prevention. (2020). Older Adults. At greater risk of requiring hospitalization or dying if diagnosed with COVID-19.
https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html

Government of Canada. (2020). People who are at high risk for severe illness from COVID-19. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/people-high-r i sk-for-severe-illness-covid-19.html

Lanza, S. T., Moore, J. E.,& Butera, N. M. (2013). Drawing Causal Inferences Using Propensity Scores: A Practical Guide for Community Psychologists.
https://doi.org/10.1007/s10464-013-9604-4

Marchand-Senécal, X., Kozak, R., Mubareka, S., Salt, N., Gubbay, J. B., Eshaghi, A., Allen, V., Li, Y., Bastien, N., Gilmour, M., Ozaldin, O.,& Leis, J. A. (2020). Diagnosis and Management of First Case of COVID-19 in Canada: Lessons applied from SARS. Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciaa227

National Women's Health Network. (2020). How Does COVID-19 Affect Different Age Groups?. https://www.nwhn.org/how-does-covid-19-affect-different-age-groups/

Rawat, A. (2017, October 31). Binary Logistic Regression. An overview and implementation in R. https://towardsdatascience.com/implementing-binary-logistic-regression-in-r-7d802a9d98fe

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C.,& Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12,p. 77. http://www.biomedcentral.com/1471-2105/12/77/

Robinson, D., Hayes, A.,& Couch, S. (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.2. https://CRAN.R-project.org/package=broom

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

World Health Organization. (2020). Coronavirus. https://www.who.int/health-topics/coronavirus#tab=tab_1