Predict Toronto Covid-19 recovery by using GLM

yiwen feng 1004890932

2020/12/22

Github Repo

https://github.com/otisfeng/sta304-final-project

Abstract

In order to calculate the probability of active patients recovery from the Covid-19, I use the GLM model to find the relationship between various variables and the outcome in the training group then use the result to predict the test group. During the process of building the model, I find out some important aspects of well-healed people and imformation of Toronto epidemic.

Keywords

Toronto, Covid-19, GLM, AIC, BIC, Recovery rate, Death rate.

Instruction

Covid-19 pandemic as the global number one issue right now, Toronto is getting worse everyday and there is no sign that the situation is getting better. The first case in Toronto happened on January 25, 2020 which was a man traveled back from Wuhan, China. On March 16, city of Toronto asked bars and theatres to close and restaurant change to take out or dilivery service only. June 18, the city recorded 1000 death and Canada recorded 100,000 cases. On November 16 Canada counted to 300,000 cases which less than a month pass 200,000.(A timeline of events in Canada's fight against COVID-19 2020) In the table blow showing the top 3 and last 3 Toronto neibourhood of people diagnosis Covid-19, some of the neighhoods has a lot higher risk compared to others. The pandemic is getting out control as the cases rocket up in shorter time, so it is a big concern that how dangerous can Covid-19 threat human life. The data I use is from the city of Toronto's open data portal called COVID-19 Cases in Toronto which records the information of each confirmed cases and it will be used to calculate the recovery probability of Covid-19 diagnosed people. In the methodology section, I describe the data and the model that was used to perform the analysis.

Rank	Neighbourhood	Value
1	Mount Olive-Silverstone-Jamestown	978
2	West Humber-Clairville	880
3	Rouge	854
139	Blake-Jones	35
140	Woodbine-Lumsden	35
141	Runnymede-Bloor West Village	34

Methodology

Data

I put a lot work into cleaning the data set, first I removed all the data with unknown gender, unknown/missing source of inflection and all probable covid-19 cases. Secondly, I built three age groups people under 19 or 19 and 20 to 29 years became young group, 30-39 years, 40 to 49 years and 50-59 years went to middle age group and others in group elder. Then I combined transgender and other and made female, male as F and M. Lastly, I add two new variables hospitalized and situation to replace variable currently and ever hospitalized, current and ever in ICU and current and ever intubated. For variable situation if the person has ever been to or current in the ICU or intubated then it would be urgent, normal if only hospitalized or never been to hospital. After cleaning the data, I separate the data into 30,108 training data which is either healed or dead patients for permorming models and 2,128 test data which only containes active patients ready for making prediction. The full model is Outcome~Outbreak associated+Age+Source of inflection +Gender+Hospitolized+Situation.

Model

```
##
## Call:
## glm(formula = factor(Outcome) ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1524   0.0155   0.0468   0.1339   2.0927
##
## Coefficients: (1 not defined because of singularities)
##                                             Estimate Std. Error z value
## (Intercept)                                  1.64516    0.04576  35.951
## Outbreak.AssociatedSporadic                  2.10782    0.12200  17.278
## AgeMiddle age                                3.32605    0.14565  22.836
## AgeYoung                                     5.71941    0.70901   8.067
## Source.of.InfectionCommunity                -0.18591    0.16750  -1.110
## Source.of.InfectionHealthcare               -0.71019    0.18005  -3.944
## Source.of.InfectionInstitutional            -0.85156    0.27241  -3.126
## Source.of.InfectionN/A - Outbreak associated       NA         NA      NA
## Source.of.InfectionPending                  10.02709  216.84483   0.046
## Source.of.InfectionTravel                    0.02845    0.26424   0.108
## GenderM                                     -0.26075    0.06198  -4.207
## GenderOther                                 10.44886  361.27277   0.029
## HospitalizedYes                             -1.51330    0.06942 -21.799
## SituationUrgent                             -1.94213    0.13882 -13.991
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## Outbreak.AssociatedSporadic                 < 2e-16 ***
## AgeMiddle age                               < 2e-16 ***
## AgeYoung                                   7.22e-16 ***
## Source.of.InfectionCommunity                0.26704
## Source.of.InfectionHealthcare              8.00e-05 ***
## Source.of.InfectionInstitutional            0.00177 **
## Source.of.InfectionN/A - Outbreak associated      NA
## Source.of.InfectionPending                  0.96312
## Source.of.InfectionTravel                   0.91427
## GenderM                                    2.59e-05 ***
## GenderOther                                 0.97693
```

```
## HospitalizedYes                                < 2e-16 ***
## SituationUrgent                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12470.2  on 30107  degrees of freedom
## Residual deviance:  7276.1  on 30095  degrees of freedom
## AIC: 7302.1
##
## Number of Fisher Scoring iterations: 14
```

The GLM model parameters tell us the strength of associations and the target is on estimating the model parameters. I select outbreak associated, age, source of inflection, gender, hospitalized and situation as the explanatory variables x and outcome as response variable y. Since in the training data set, outcome has resolved and fatal which sets resolved as 1 and fatal as 0. The number of dummy variable depends on the number of your code under this categorical variables minus one. The one varialbe that was minus is the base line, all the left dummy variables of this categorical variable will take the base line variable as the reference. If the P-value of variable is smaller than 0.05, we say this variable is significant to our predition. Sporadic outbreak associated, middle age, young age, healthcare source of inflection, institutional source of inflection, male gender, hospitalized and urgent are significant parameters. To interpret the model, keep all other variables unchanged, if the patient is male, 0.26075 of survival chance will drop. Keep all other variables unchanged, if the patient hospitalized or in urgent situation, 1.51330 and 1.94213 of survival chance will drop. Keep all other variables unchanged, if the patient is in young age or middle age group, the survival chance increases by 5.71941 and 3.32605.

#AIC

```
## Start:  AIC=7302.14
## factor(Outcome) ~ Outbreak.Associated + Age + Source.of.Infection +
##     Gender + Hospitalized + Situation

##
## Call:
## glm(formula = factor(Outcome) ~ Outbreak.Associated + Age + Source.of.Infection +
##     Gender + Hospitalized + Situation, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1524   0.0155   0.0468   0.1339   2.0927
##
## Coefficients: (1 not defined because of singularities)
##                                             Estimate Std. Error z value
## (Intercept)                                  1.64516    0.04576  35.951
## Outbreak.AssociatedSporadic                  2.10782    0.12200  17.278
## AgeMiddle age                                3.32605    0.14565  22.836
## AgeYoung                                     5.71941    0.70901   8.067
## Source.of.InfectionCommunity                -0.18591    0.16750  -1.110
## Source.of.InfectionHealthcare               -0.71019    0.18005  -3.944
## Source.of.InfectionInstitutional            -0.85156    0.27241  -3.126
## Source.of.InfectionN/A - Outbreak associated       NA         NA      NA
## Source.of.InfectionPending                  10.02709  216.84483   0.046
## Source.of.InfectionTravel                    0.02845    0.26424   0.108
## GenderM                                     -0.26075    0.06198  -4.207
## GenderOther                                 10.44886  361.27277   0.029
```

```
## HospitalizedYes                                -1.51330    0.06942 -21.799
## SituationUrgent                                -1.94213    0.13882 -13.991
##                                                Pr(>|z|)
## (Intercept)                                     < 2e-16 ***
## Outbreak.AssociatedSporadic                     < 2e-16 ***
## AgeMiddle age                                   < 2e-16 ***
## AgeYoung                                       7.22e-16 ***
## Source.of.InfectionCommunity                    0.26704
## Source.of.InfectionHealthcare                  8.00e-05 ***
## Source.of.InfectionInstitutional                0.00177 **
## Source.of.InfectionN/A - Outbreak associated         NA
## Source.of.InfectionPending                      0.96312
## Source.of.InfectionTravel                       0.91427
## GenderM                                        2.59e-05 ***
## GenderOther                                     0.97693
## HospitalizedYes                                 < 2e-16 ***
## SituationUrgent                                 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12470.2  on 30107  degrees of freedom
## Residual deviance:  7276.1  on 30095  degrees of freedom
## AIC: 7302.1
##
## Number of Fisher Scoring iterations: 14
```

#BIC

```
## Start:  AIC=7411.09
## factor(Outcome) ~ Outbreak.Associated + Age + Source.of.Infection +
##     Gender + Hospitalized + Situation
##
## Call:
## glm(formula = factor(Outcome) ~ Outbreak.Associated + Age + Source.of.Infection +
##     Gender + Hospitalized + Situation, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1524   0.0155   0.0468   0.1339   2.0927
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error z value
## (Intercept)                                     1.64516    0.04576  35.951
## Outbreak.AssociatedSporadic                     2.10782    0.12200  17.278
## AgeMiddle age                                   3.32605    0.14565  22.836
## AgeYoung                                        5.71941    0.70901   8.067
## Source.of.InfectionCommunity                   -0.18591    0.16750  -1.110
## Source.of.InfectionHealthcare                  -0.71019    0.18005  -3.944
## Source.of.InfectionInstitutional               -0.85156    0.27241  -3.126
## Source.of.InfectionN/A - Outbreak associated        NA         NA      NA
## Source.of.InfectionPending                     10.02709  216.84483   0.046
## Source.of.InfectionTravel                       0.02845    0.26424   0.108
```

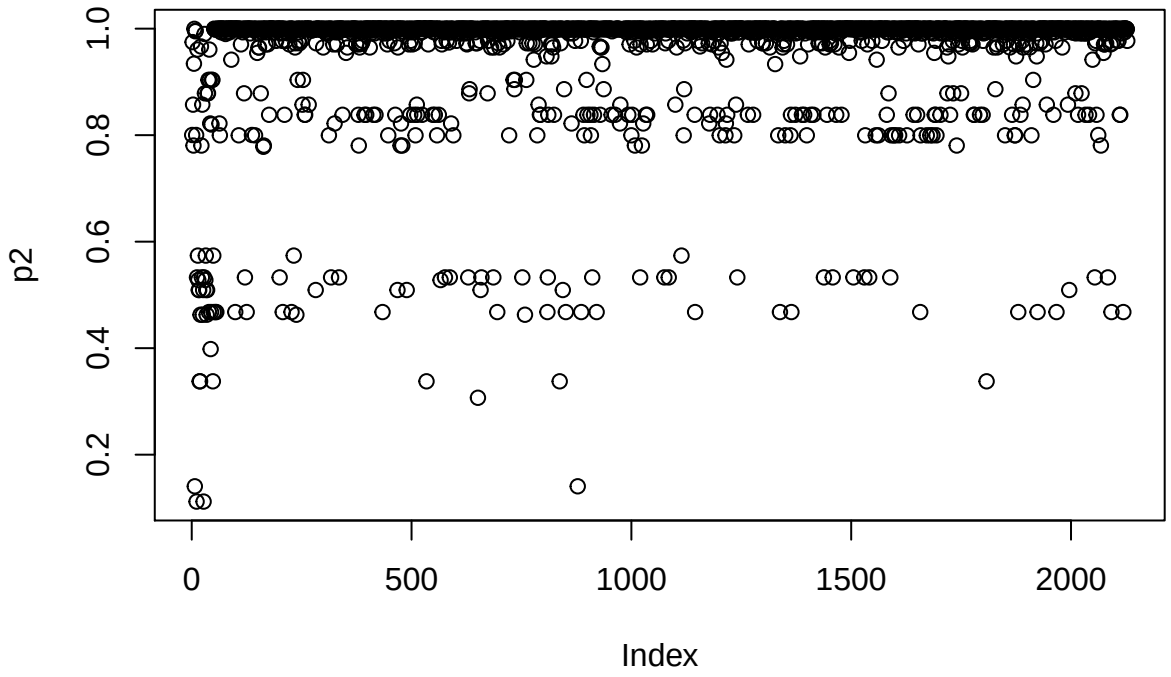
```
## GenderM                                          -0.26075    0.06198  -4.207
## GenderOther                                      10.44886  361.27277   0.029
## HospitalizedYes                                  -1.51330    0.06942 -21.799
## SituationUrgent                                  -1.94213    0.13882 -13.991
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## Outbreak.AssociatedSporadic                      < 2e-16 ***
## AgeMiddle age                                    < 2e-16 ***
## AgeYoung                                        7.22e-16 ***
## Source.of.InfectionCommunity                     0.26704
## Source.of.InfectionHealthcare                   8.00e-05 ***
## Source.of.InfectionInstitutional                 0.00177 **
## Source.of.InfectionN/A - Outbreak associated          NA
## Source.of.InfectionPending                       0.96312
## Source.of.InfectionTravel                        0.91427
## GenderM                                         2.59e-05 ***
## GenderOther                                      0.97693
## HospitalizedYes                                  < 2e-16 ***
## SituationUrgent                                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12470.2  on 30107  degrees of freedom
## Residual deviance:  7276.1  on 30095  degrees of freedom
## AIC: 7302.1
##
## Number of Fisher Scoring iterations: 14
```

By using AIC and BIC, we find out the final models for them are consistent with GLM model. By selecting lower AIC, we know AIC has a better fit.

Results

```
Based on the GLM model we have, we make prediction on test group and we get a scatterplot of the percen
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```




```
##   Resolved_prop
## 1     0.8970865
```

```
##    Fatal_prop
## 1 0.001879699
```

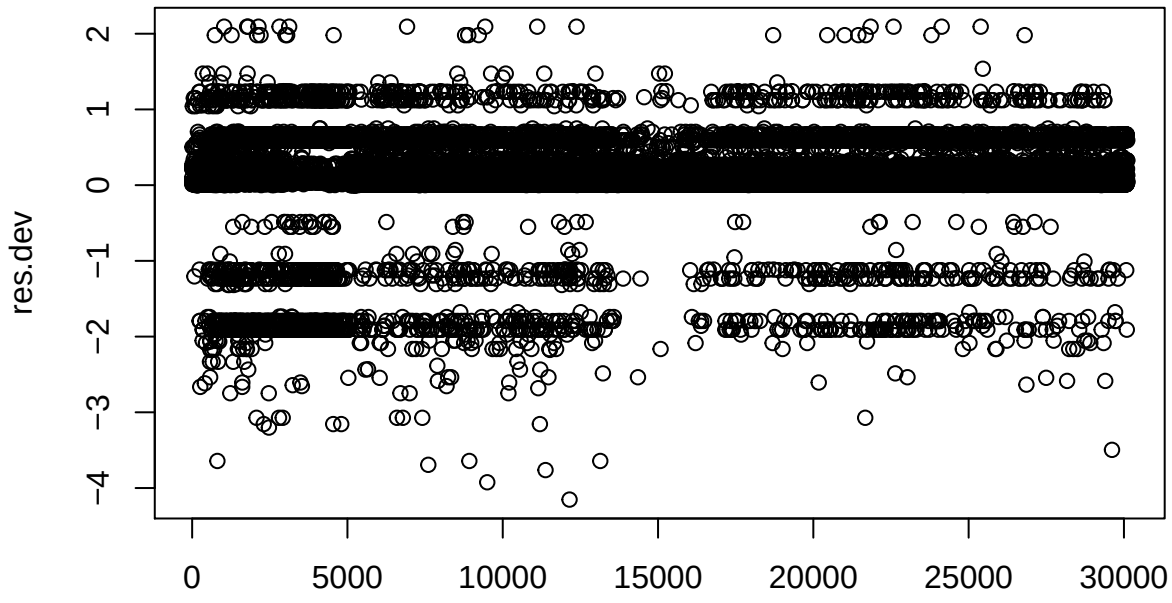
Discussion

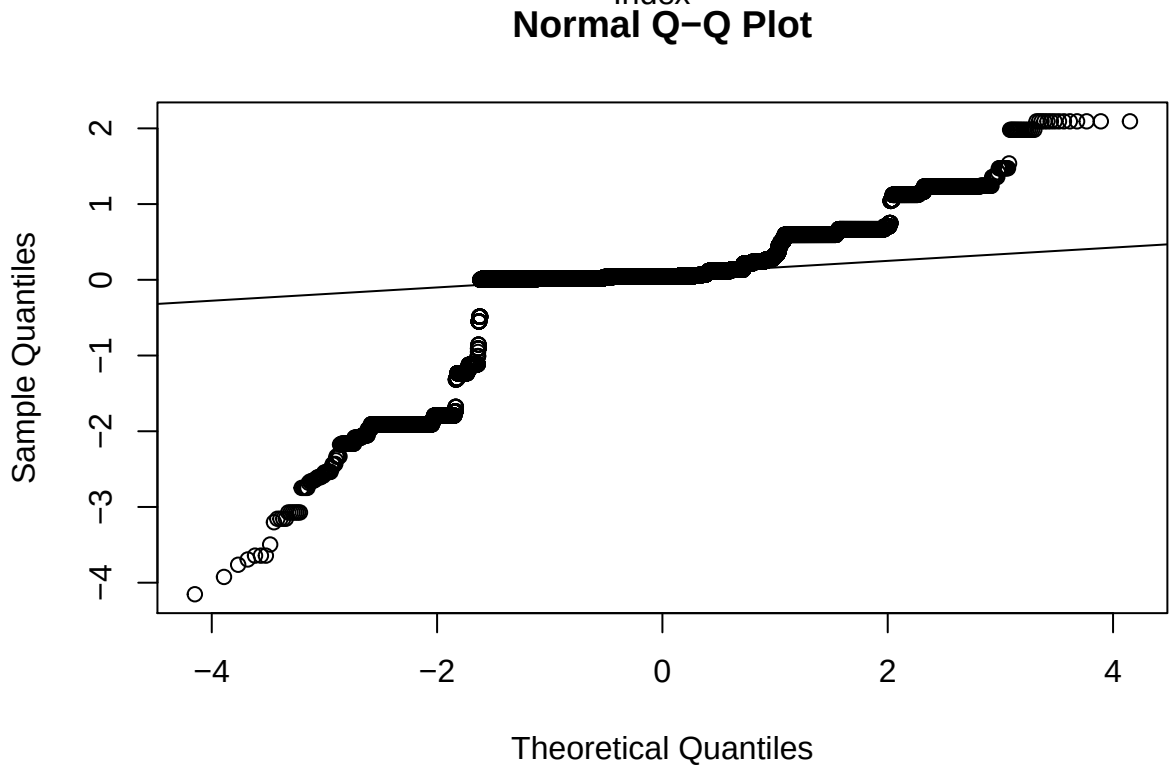
Summary

First, I built a GLM model to find which variables were significant to the outcome and checked the model with AIC, BIC which helped choosing best predictors. Then I used the model to predict the test data set and got the result of the percentage of safe and dangerous patients.

Conclusions

People with older age are more dangerous than others because their bodies immunity power is low that Covid-19 is easier to do more damage so they should be more careful. Working in institutional and healthcare, people might have higher chance to expose under other Covid-19 diagnosed people. People hospitalized or urgent means Covid-19 hurt them a lot. Since the test group only has 2,128 observations our death rate is overrated and the recovery rate is underrated. If the data become millions and billions, the death rate will be a lot lower since most people get Covid-19 have no symptom and recover really soon.





Weakness & Next Steps

From the normal QQ plot, we can clearly see that only middle part follows the normal line, the lower and higher parts deviate the normal line so that the standardized residual violates the normal assumption. The residual plot shows most of residuals are around zero, as there are no pattern exists so it does not violate the assumption of the constant variance. For the process of the variable selection, I regrouped the variables by my own thoughts, the group interval might influcnced our model results.

Reference

1. A timeline of events in Canada's fight against COVID-19. (2020, December 15). Retrieved December 22, 2020, from https://www.cp24.com/news/a-timeline-of-events-in-canada-s-fight-against-covid-19-1.5231865