Reasearch on factors influence new death by Covid-19 using multiple linear regression model

Canyang Wang

2022/12/16

Introduction

My Question of Proposed Research

Studying the factors that may influence the number of new deaths from COVID-19 is critical for people in Canada, as they want to avoid dying from the disease. Therefore, conducting research to investigate the factors that impact the number of new COVID-19 deaths is helpful for scientists to come up with ways to save more lives. Many people are still facing the issue of COVID-19, and understanding the factors that can reduce the number of new COVID-19 deaths can improve their quality of life to some extent.

Background and Literature

One article mentioned that country type influences the death rate and they got the result that in some highly developed countries, people have less death rate and positive rate due to the cause of Covid-19.(WSerhii Kozlovskyi et al.,2021). Their research indicates that positive rate may be a factor, thus, I also use it as my variable. But for the difference, it is also clear that they argued the determinants of Covid-19 death rate will be affected by different types of countries.

A second article found that clinical conditions and vaccination are important for the death rate of COVID-19 patients. They concluded that certain clinical conditions, particularly asthma, increase the death rate, while a high vaccination rate decreases it. Their research indicates that vaccination influences the number of new deaths, so I have also included it as a variable in my study. However, I did not consider clinical conditions in my research because my dataset does not have relevant data.

A third article found that older age and early hospitalization also impact the risk of death from COVID-19. They showed that younger patients and those who sought hospital care early have a lower risk of death from COVID-19. Like their research, I also looked at the impact of hospitalization on reducing the number of new COVID-19 deaths. However, we studied cases in different countries, with their study examining cases in China and mine looking at cases in Canada.

Methods section

Model selection

Data Description and Data Chosen

The variables I chose are relatively consistent with my research question and my background research based on my cleaned dataset:

- reproduction_rate: The rate people who got Covid-19 and then get again.
- new_deaths: The new number of people who died because of Covid-19 in Canada
- new_tests_per_thousand: The people per thousand who take a test for Covid-19 in Canada
- new_vaccinations: The new people who get vaccinations in Canada
- positive_rate: The rate for the tested people who are positive in Canada
- icu_patients_per_million: The patients per million who are in ICU in Canada
- hosp_patients_per_million: People per million who go to a hospital in Canada

I divided the final cleaned data into training data and testing data. Training data is 400 rows and testing data is 136 rows.

Use the training data to fit a multiple linear regression (MLR) model. I remove variables with p-values greater than 0.05 to create a model and refit the model until all variables are significant. I then use the variance inflation factor (VIF) to test for dependencies between variables and remove any with VIF values greater than 5 to create a new model. I use a partial F test to compare the new model to a subset model with a randomly removed variable(compared with the new model) and choose the model based on reject H0 or not(p-value less than 0.05 means I should reject H0 to choose new model) otherwise, choose the subset of new model). Finally, I use an F test on the chosen model to ensure all variables have p-values less than 0.05 and create a final model.

Model validation

I basically validate my model in two steps:

Firstly, using a function in r code to calculate adjusted R square, AIC, AICc, BIC and SSres compared with a subset of fitted model. If the fitted model has the biggest R^2_{adj} or the smallest R^2, AIC, AICc, BIC, I can conclude that my fitted model is the best and nearly validated.

Secondly, using the test data to fit the training model. If I find my test model's variables are not all significant(check p-value less than 0.05 or not) use the same variable as the training model, then it happens limitations in fitted training model. Moreover, also check adjusted R square, AIC, AICc, BIC and SSres have similar values as my train model, if not, add this into limitation.

Model Violations and Diagnostics

Using residual plots and QQ plots to check the four assumptions of my model: normality, linearity, homoscedasticity, and independence. I use QQ plots to check for normality, and if the points do not lie on a straight line, I may need to use a transformation like Box-Cox to modify my model. I use residual plots to check for linearity and independence, and if there are patterns or clustering, I may need to transform the model or remove variables. I also check for homoscedasticity by looking at the variance(expand outward or shrink inward) in the residual plot and may use a variance stabilizing transformation like taking the log if needed.

Then check for leverage, outlier and influence points.

If standard residual points r_i follow standard normal and don't belong to [-2,2] in small datasets or [-4,4] in large datasets, they are outliers so remove these points.

If $D_i > 50th$ percentile of F(p + 1, n - p - 1) $|DFFITS_i| > 2 * \sqrt{\frac{p+1}{n}}$ and $|DFBETA_{j(i)}| > \frac{2}{\sqrt{n}}$, these are influential points and remove these points.

Get leverage points by judging whether $h_{ii} > \frac{p+1}{n}$.

Result

Description of Data

For finding my dataset, I use "kaggle" and then choose the most convenient and suitable dataset called "owid-covid-data.csv". The original dataset I found online had 219175 observations and 67 variables in total.

Firstly, filtering out the original data according to the observations have "location=Canada", then removing the observations which have NA and selecting the variables present in method section to make a new dataset.

The Table6 shows that the chosen variables in my dataset have several upper outliers need to remove. It also shows that the new deaths have a large range, which may indicate that fewer new deaths will occur in the future as people become more knowledgeable about Covid-19.

Scatter plots show that increased positive rate, ICU patients, hospital visits, and testing lead to an increase in new deaths. In contrast, increased vaccination, and reproduction rate lead to a decrease in new deaths.

Presenting the Analysis Process and the Results

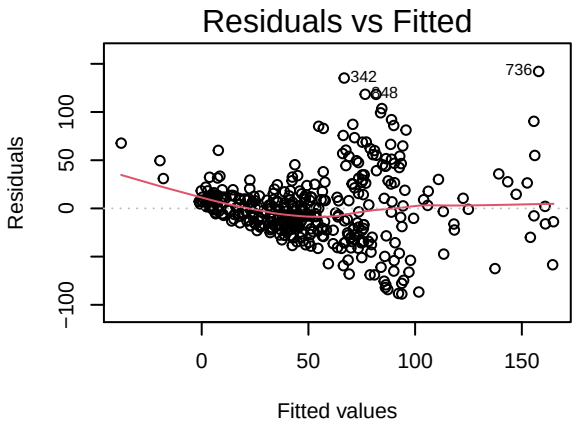
Table 1: Summary of three versions of model

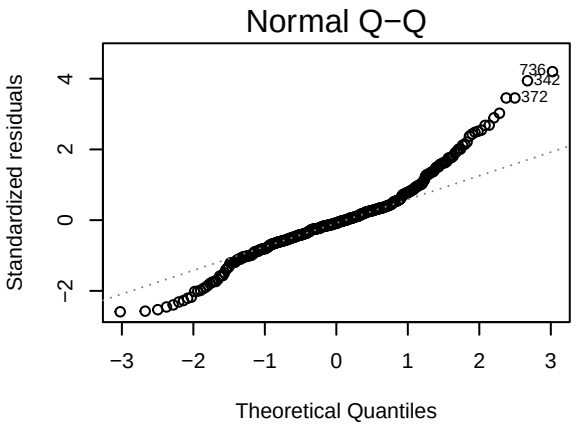
term	estimate	std.error	statistic	p.value	VIF
(Intercept1)	-0.323025	12.240775	-0.026389	0.978960	NA
- new_vaccinations_1	-0.000035	0.000013	-2.780620	0.005687	NA
- hosp_patients_per_million_1	0.696312	0.061545	11.313867	0.000000	NA
- new_tests_per_thousand_1	5.495283	1.574346	3.490519	0.000537	NA
- reproduction_rate_1	-2.087593	10.289211	-0.202891	0.839325	NA
- icu_patients_per_million_1	0.190497	0.296369	0.642770	0.520748	NA
- positive_rate_1	-228.983981	42.818274	-5.347810	0.000000	NA
(Intercept2)	-2.597473	4.910180	-0.528997	0.597105	NA
- new_vaccinations_2	-0.000034	0.000012	-2.813641	0.005144	NA
- hosp_patients_per_million_2	0.702415	0.053628	13.097898	0.000000	NA
- new_tests_per_thousand_2	5.376280	1.459229	3.684331	0.000261	NA
- icu_patients_per_million_2	0.196130	0.294707	0.665510	0.506114	NA
- positive_rate_2	-232.390290	39.341633	-5.906981	0.000000	NA
(Intercept3)	-1.468709	4.604694	-0.318959	0.749926	NA
- new_vaccinations3	-0.000032	0.000012	-2.737214	0.006476	1.073086
- hosp_patients_per_million3	0.723503	0.043235	16.734296	0.000000	2.255694
- new_tests_per_thousand3	5.738216	1.353142	4.240661	0.000028	1.053668
- positive_rate3	-242.379164	36.340311	-6.669705	0.000000	2.241964
(Intercept4)	3.066906	0.294431	10.416376	0.000000	NA
- new_vaccinations4	-0.000002	0.000001	-3.162494	0.001685	1.073086
- hosp_patients_per_million4	0.044835	0.002764	16.218147	0.000000	2.255694
- new_tests_per_thousand4	0.422862	0.086522	4.887343	0.000001	1.053668
- positive_rate4	-14.369482	2.323655	-6.184001	0.000000	2.241964

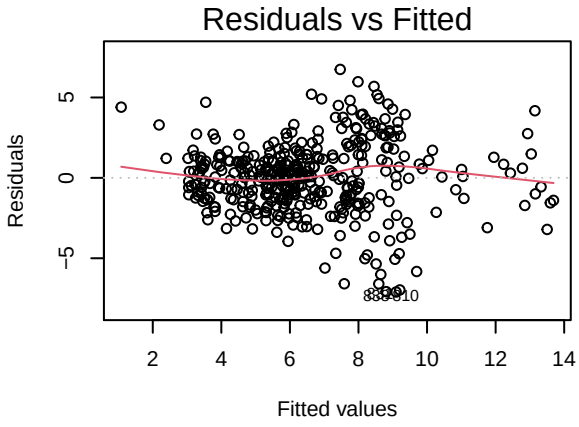
I fit an MLR model using initial variables, including "new_death" as the dependent variable. I removed the insignificant variable "reproduction_rate" and refit the model until I got a third model with all significant variables (excluding "icu_patients_per_million"). I performed a VIF test on the third model and found

that all variables had a VIF less than 5, indicating no dependency between them. Also after I checked the goodness of my final model, since I found my third model did not match homoscedasticity, thus, I refit the new model by taking a square root transformation and doing the same check as above, to get the new fitted model.

Goodness of the Final Model







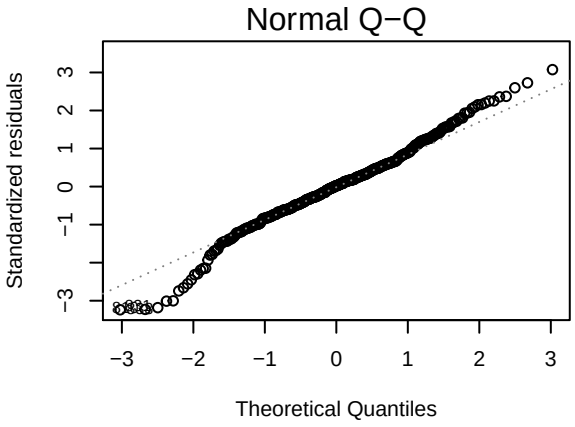


Table 2: ANOVA table for partial F test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
396	2096.342	NA	NA	NA	NA
395	1911.300	1	185.0422	38.24187	0

Table 3: ANOVA table for F test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
new_vaccinations	1	110.8176	110.817633	22.90220	2.4e-06
hosp_patients_per_million	1	1469.8329	1469.832852	303.76398	0.0e+00
new_tests_per_thousand	1	141.1387	141.138745	29.16853	1.0e-07
positive_rate	1	185.0422	185.042190	38.24187	0.0e+00
Residuals	395	1911.2996	4.838733	NA	NA

Then I make a QQ plot and a residual plot to check the four assumptions for third model, as the first 2 graphs, all points do not fit a straight line, thus normality does not fit

The residual plot has no points clustered together and curved patterns, thus it meets the linearity and independence assumptions, but the homoscedasticity does not fit since there happens outward expansion.

Thus, after I refit the model as a new third model, I redo the QQ plot and residual plot as the second 2 graphs. Now, all points fit a straight line, thus normality fits.

The residual plot has no points clustered together and curved patterns and there is no shrink inward or expansion outward, thus, the linearity, independence and homoscedasticity assumptions all fit.

For leverage points, outliers, and influence points, after calculation in R based on the new third model, I get zero outlier, 36 leverage points and no influence points.

I then did a partial F test between the fourth model (new third model excluding "positive_rate") and the new third model, resulting in a p-value of 0 and indicating that the new third model should be chosen since H0(fourth model should be chosen) is rejected. Finally, I applied an F test to the new third model and found that all variables had p-values < 0.05, making it the final model.

Table 4: Summary of characteristics for different model and test model

variables	Model4	Model3New	TestModel
SSres	2096.3417761	1911.2995860	564.1915591
Rsq	0.4509508	0.4994149	0.5665017
Rsq_adj	0.4467913	0.4943457	0.5532651
AIC	668.5938185	633.6296616	201.4924975
AICc	668.7453336	633.8423199	202.1337188
BIC	692.5511412	661.5784489	222.9684268

In Table4, I use two versions of training model(new third model and the subset of it) to compare adjust R square, AIC, AICc, BIC and SSres. My new third model has the bigger R^2_{adj} and the smaller R^2, AIC, AICc, BIC among the two models, also, the Rsq and Rsq_adj are like my test model.

Then, I also make a MLR with the same variables based on my test data, all variables are still significant(p-value > 0.05) except for "new_vaccinations", thus, this is my limitation. But, overall, my fitted model is good.

Discussion Section

Final Model Interpretation and Importance

My final model is $y_{newDeaths} = 3.066906 + (-0.000002)x_{newVaccination} + 0.044835x_{hostPatients} + 0.422862x_{newTest} + (-14.369482)x_{positiveRate}$ The intercept of my model is 3.066906. Holding other variables unchange, with one unit increase in $x_{newVaccination}$, $y_{newDeaths}$ will decrease in 0.000002, one unit increase in $x_{hostPatients}$, $\hat{y}_{newDeaths}$ will increase in 0.044835, one unit increase in $x_{bewTest}$, $y_{newDeaths}$ will increase in 0.422862, and one unit increase in $x_{positiveRate}$, $y_{newDeaths}$ will decrease in 14.369482, This model indicates that the positive rate has the most influences on new deaths and the new vaccination matters the least. Other two factors also have influences but not too much compared with positive rate. Thus, the final model clearly emphasized my research questions that new vaccination, people who go to hospitals, people who take tests and positive rate do have influence on the new deaths caused by Covid-19 in Canada.

Limitations of the Analysis

My test model does not well fetch with my training model since there is a variable insignificant in test data, this may be solved either by getting more samples or working with the correctness of my data.

Compared with my literature research, I used quite a few variables and data sets (just Canada's data), thus, it may happen inaccuracy in the process of concluding a result. I should add more data or find more useful variables in the future.

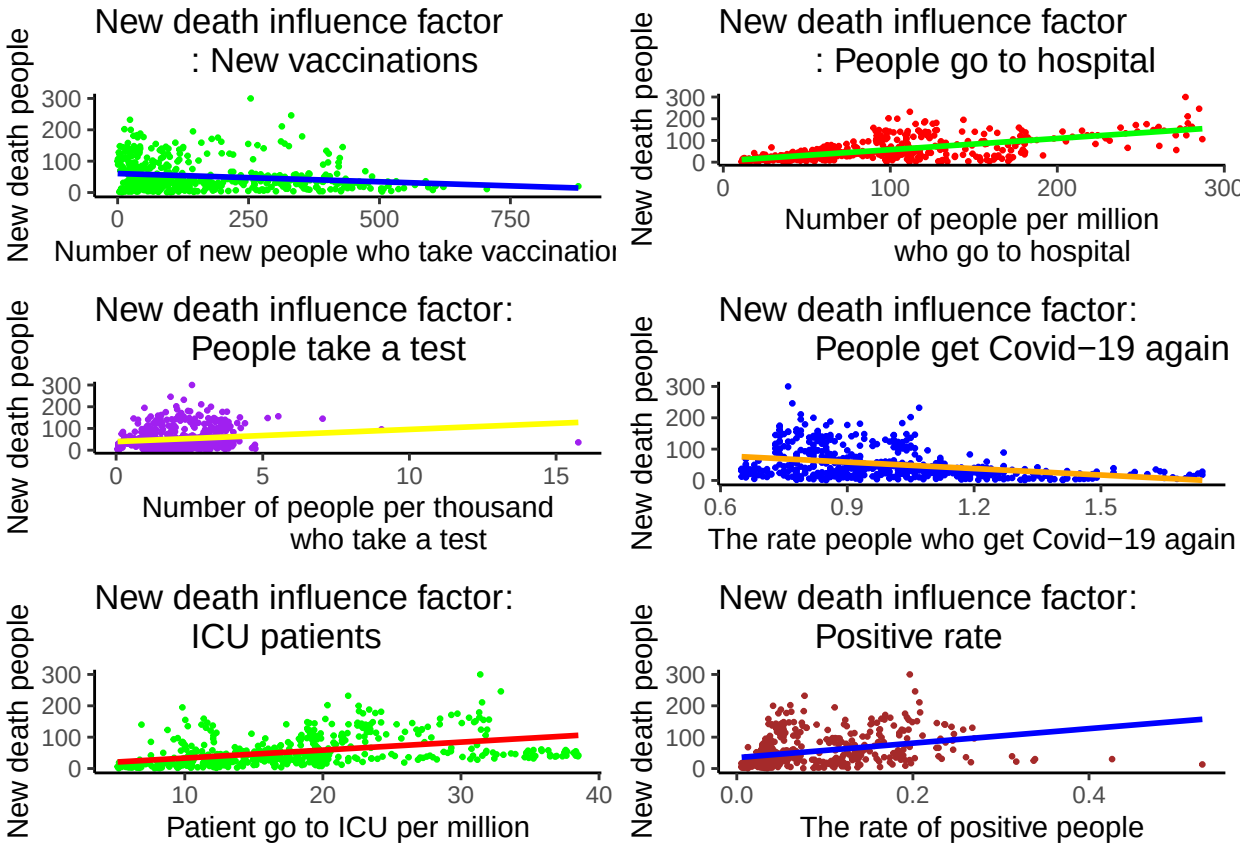
Appendix

Table 5: Table 1: Summary of all variables

variables	minimum	maximum	mean	median	IQR	Lower_outlier	Upper_outlier
new deaths	1.000	300.000	52.00	36.000	49.25	0	35
reproduction rate	0.650	1.740	0.99	0.960	0.28	0	15
new tests per thousand	0.053	15.759	2.25	2.210	1.53	0	4
new vaccinations	15.000	879740.000	159418.34	98003.000	218408.00	0	7
positive rate	0.005	0.529	0.08	0.047	0.08	0	18
icu patients per million	5.137	38.501	17.42	15.935	9.63	0	17
hosp patients per million	10.982	286.856	89.43	70.750	70.62	0	26

Table 6: Summary Table of First Version of Test Model

term	estimate	std.error	statistic	p.value
(Intercept)	2.102585	0.525651	3.999961	0.000105
new_vaccinations	-0.000002	0.000001	-1.706298	0.090322
hosp_patients_per_million	0.043311	0.004597	9.422111	0.000000
new_tests_per_thousand	0.717049	0.165985	4.319949	0.000031
positive_rate	-10.747342	4.556101	-2.358890	0.019808



References

Esai Selvan, M. (2020, May 27). Risk factors for death from covid-19. Nature News. Retrieved October 22, 2022.

Fu, L., Fei, J., Xiang, H.-X., Xiang, Y., Tan, Z.-X., Li, M.-D., Liu, F.-F., Liu, H.-Y., Zheng, L., Li, Y., Zhao, H., & Xu, D.-X. (2020, January 1). Influence factors of death risk among COVID-19 patients in Wuhan, China: A hospital-based case-cohort study. medRxiv. Retrieved October 22, 2022.

Kozlovskyi, S., Bilenko, D., Dluhopolskyi, O., Vitvitskyi, S., Bondarenko, O., & Korniichuk, O. (2021). Determinants of COVID-19 Death Rate in Europe: Empirical Analysis. Problemy Ekorozwoju, 16(1), 17–28.

Letsoalo, R. (2022, October 3). Covid analysis. Kaggle. Retrieved October 22, 2022, from https://www.kaggle.com/datasets/reikagileletsoalo/covid-analysis?select=owid-covid-data.csv