

**Linear Regression analysis of potential factors correlated with COVID-19
death rate in Toronto neighborhoods**

Yuxuan Yang
December 16, 2021

Introduction

Since the beginning of COVID-19 pandemic, there have been many deaths, among which are our colleagues, friends, or even instructors. The purpose of my project is to investigate potential linear relationship between COVID-19 death rate in Toronto neighborhoods and proportion of elderly COVID patients, proportion of patients that had close contact with confirmed cases, proportion of patients that have ever been in hospital, and percent of patients that have ever been in the ICU for each neighborhood. Through this potential linear correlation, we could hopefully gain some intuitions in how to facilitate our neighborhoods to best combat the coronavirus in the coming days.

To provide some backgrounds, we would look at a paper recording the clinical observations of 7 elderly COVID-19 patients in Zhongnan Hospital of Wuhan University; this paper showed that death rate among elderly people (aged above 80) doesn't vary much from that of younger people (Rui, Sirui, Xuebei, Xujun, & Yanggan, 2020). Another paper, which investigated the prognosis of elderly patients in the ICU during the first wave, demonstrated that both being in the ICU and being elderly might be associated with higher COVID-19 death rate (Dres et al., 2021). The last paper we will look at studies characteristics of COVID-19 patients with respiratory difficulties and shows the potential of early-stage respiratory symptoms to predict the survival rate of COVID patients (Formenti et al., 2021). All these findings would facilitate the predictor selection process in our project, as discussed later in the methods section.

Methods

Before we start to build our models, we will first do a 50/50 split and separate our neighborhood-level data into 2 parts: the train and test sets. We will run regressions and select models using only the train set, while leaving the test set for final model validation.

Variable Selection

We would include `prop_elderly` (proportion of elderly patients), `prop_close_contact` (proportion of patients that had close contact with confirmed COVID cases), `prop_hospital` (percent of patients ever in hospital), and `prop_ICU` (percent of people ever in ICU) as predictors, and regress `death_rate` (COVID-19 death rate in each neighborhood) on these predictors on train data and look at the p-values. This model is the full model we will start with in our variable selection process. The inclusion of these 4 variables is a decision affected by the paper findings we discussed earlier.

Model Violations and Diagnostics

Meanwhile, we will look at the response VS. fitted value plot and pairwise scatterplots between these 4 predictors to check conditions 1 and 2, along with residuals VS. fitted value plot, residuals VS. predictors plot, and normal QQ plot to determine whether all linear regression assumptions satisfy in the full model. If we observe weird patterns in any of these plots, we would need to transform either the response, the predictors, or both, depending on the results of Box-Cox transformation on the response and the predictors.

Then we will regress the transformed response on the 4 variables, and call the new model `full2`. We will recheck all the assumptions and conditions 1 and 2. If all assumptions satisfy,

we'll continue to check VIF (variance inflation factor) to see whether there're severe multicollinearity issues in model full2. Otherwise, we will repeat the process of transforming variables and running regressions, until all assumptions are satisfied.

We will check VIF of full2 and be aware of the existence of multicollinearity if VIF is above 5. We will also check the Cook's Distance, DFFITS, and DFBETAS to determine influential points, together with checking hat values and standard residuals for outliers and leverage points.

Variable Selection

Determining on the p-values in model full2, we will run a partial F-test to determine whether we could remove a subset of predictors all at once. But we'll be cautious and also perform hypothesis tests regressing death_rate on each variable. We'll compare the individual t-test results with our partial F-test results to determine whether we're satisfied with full2, or we want to try a few more models. Again, we will check assumptions, conditions, influential points, leverages, outliers, and VIF for each new model. Lastly, we will compare all these models and choose a potential final model.

Model Validation

We'll transform the test set the same way we transformed the train set, and apply the final model on the test set. We'll check assumptions, conditions, VIF, influential points, leverages, and outliers for the test set and compare the performance of the final model on test set with that on the train set.

Results Section

Description of Data

We'll first look at numerical summaries of all variables for both the train and test sets.

Table 1: Summary statistics for all variables

Variable	mean (s.d.) in training	mean (s.d.) in test
COVID-19 death rate	0.023 (0.018)	0.02 (0.014)
Proportion of patients who had close contact with confirmed cases	0.075 (0.015)	0.078 (0.015)
Proportion of patients ever in hospital	0.07 (0.019)	0.064 (0.017)
Proportion of patients ever in ICU	0.013 (0.006)	0.011 (0.005)
Proportion of elderly COVID-19 patients	0.197 (0.066)	0.195 (0.067)

From the means and standard deviations of each variable in Table 1, we didn't observe much difference between train and test sets, since all variable means in test are within one standard deviation of those in train.

Process of Obtaining Final Model

As mentioned in Methods section, we started with the full model, included all variables in Table 1, and checked for conditions and assumptions. We observed slight weird patterns in the

death rate VS. fitted value plot, residuals VS. fitted value plot, and residuals VS. proportion elderly patients plot, so we decided to run Box-Cox transformation on both the response and all 4 predictors.

Table 2: Box-Cox Power Transformations to Multi-normality

	Est Power	Rounded Power
death_rate	0.1039	0.0
prop_close_contact	0.5629	1.0
prop_hospital	0.2421	1.0
prop_ICU	0.5166	0.5
prop_elderly	-0.4787	0.0

Since rounded power is 0.0 for death_rate in Table 2, we will try log transformation on death_rate. We did so, and built a new model called full2, regressing log death rate on all 4 predictors and rechecked the assumptions and conditions. Now all the assumptions and conditions are satisfied. We also checked that all VIFs of full2 are smaller than 5, so we know there's no severe multicollinearity. We then calculated COOK's Distance, DFFITS, and DFBETAS, together with hat values and standard residuals; we're aware of the existence of influential points, outliers and leverage points in our train observations.

Next, we ran a partial F-test to see whether we could remove prop_close_contact, prop_hospital, and prop_ICU all at once. We reached a p-value larger than 0.05, indicating we fail to reject the null hypothesis that the models with and without these 3 predictors have any difference. It seems we could just remove them all immediately, but let's be cautious and regress death rate on each predictor.

Table 3: Individual t-tests results

	t-test 1	t-test 2	t-test 3	t-test 4
Prop_close_contact	-10.8487 (6.0811)			
Prop_hospital		23.2321*** (4.2057)		
Prop_ICU			42.5175* (16.4231)	
Prop_elderly				8.3324*** (0.8887)

Note: The response is log death rate.

Standard errors are in parentheses.

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

These individual t-tests reveal we can only remove prop_close_contact, since its coefficient, -10.8487, is not significant. This is different from the partial F-test results, so we guess there are interactions between some predictors. Let's look at a few more models and compare them with full2.

Table 4: Compare all potential models

	full2	mod1	mod2	mod3
Prop_close_contact	-5.5458 (4.1223)			
Prop_hospital	3.7005 (4.7152)	3.8827 (4.7430)	46.903** (15.968)	6.7566 (4.1409)
Prop_ICU	12.6176 (13.2217)	16.0288 (13.0581)	-2.438 (44.646)	
Prop_elderly	7.4039*** (1.1062)	7.4390*** (1.1129)	17.971*** (4.852)	7.2372*** (1.1050)
Prop_hospital*prop_elderly			-175.309** (51.202)	
Prop_ICU*prop_elderly			277.398 (205.161)	
Prop_hospital*prop_ICU			-567.149 (721.820)	
Adjusted R-squared	0.5839	0.5786	0.6335	0.5754
Observations	68	68	68	68

Note: The response is log death rate

Standard errors are in parentheses.

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

From Table 4, we should pick prop_hospital, prop_elderly, and the interaction term between prop_hospital and prop_elderly in mod2 as predictors due to the significant p-values and highest adjusted R-squared. However, VIFs in mod2 are much larger than 5, indicating severe multicollinearity. Also, considering prop_close_contact and prop_ICU are not significant in any model, we would choose mod3 as the final model and include prop_hospital and prop_elderly as predictors.

Note that we've confirmed all the assumptions and conditions are satisfied in the above models, and are aware of the influential points, leverage points, and outliers in our train set (though we can't necessarily remove them). We can now apply this final model on our test set and see how it performs.

Goodness of Final Model

We'll transform death rate by log and apply the final model on test set.

Table 5: mod3_test results

	Estimated Coefficients (s.d.)	VIF
Prop_hospital	12.9494*** (3.1229)	1.656452
Prop_elderly	6.1751*** (0.7722)	1.656452

Note: The response is log death rate.

Standard errors are in parentheses.

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

We've applied the final model on test set and confirmed all assumptions, conditions and VIFs. We're also aware of influential points, leverage points, and outliers in the test set. From Table 5, both prop_hospital and prop_elderly are significantly linearly correlated with log COVID-19 death rate in the test set, though the exact estimated coefficients differ from those in the train set (probably due to different influential observations). In conclusion, we have evidence our final model is validated, and it actually performs better on the test set than the train set.

Discussion Section

Final Model Interpretation and Importance

The final model interprets that COVID-19 death rate in a neighborhood is significantly linearly correlated with both proportion of elderly COVID patients and proportion of COVID patients ever in hospital, yet has nothing to do with proportion of patients ever in ICU or proportion of patients who had close contact with confirmed cases. That is, our neighborhoods might need to focus on protection for the elderly population and population with underlying diseases in the future days.

Limitations of Analysis

Though we are aware of the existence of different influential observations in the train and test sets, they are part of our sample; we couldn't remove them and correct the regression lines without contextual reasons. This causes our final model to perform differently in the train and test sets.

Another limitation is that we've been considering only 4 predictors since the beginning of this project. There might be other factors related to COVID-19 death rate and our predictors left undiscussed, causing our final model to have omitted variable bias. But we couldn't correct our model at this stage due to our inability to collect more variables.

References

- Dres, M., Hajage, D., Lebbah, S., Kimmoun, A., Pham, T., Béduneau, G., Combes, A., Mercat, A., Guidet, B., Demoule, A., & Schmidt, M. (2021). Characteristics, management, and prognosis of elderly patients with COVID-19 admitted in the ICU during the first wave: insights from the COVID-ICU study : Prognosis of COVID-19 elderly critically ill patients in the ICU. *Annals of Intensive Care*, 11(1), 77–77. <https://doi.org/10.1186/s13613-021-00861-1>
- Formenti, P., Umbrello, M., Castagna, V., Cenci, S., Bichi, F., Pozzi, T., Bonifazi, M., Coppola, S., & Chiumello, D. (2021). Respiratory and peripheral muscular ultrasound characteristics in ICU COVID 19 ARDS patients. *Journal of Critical Care*, 67, 14–20. <https://doi.org/10.1016/j.jcrc.2021.09.007>
- Rui, L., Sirui, L., Xuebei, D., Xujun, Y., & Yanggan, W. (2020). Clinical observations in very elderly patients with COVID-19 in Wuhan. *Geriatrics & Gerontology International*, 20(7), 709–714. <https://doi.org/10.1111/ggi.13974>
- Data source: <https://open.toronto.ca/dataset/covid-19-cases-in-toronto/>
This dataset contains all the COVID-19 cases (confirmed or probable) in 140 neighborhoods of Toronto reported to Toronto Public Health, together with their demographic, geographic, and severity information since the first case was reported in January 2020. (This is the *covid_data* in our rmd file)

Appendix

Figure 1: assumptions and conditions' checks for initial full model

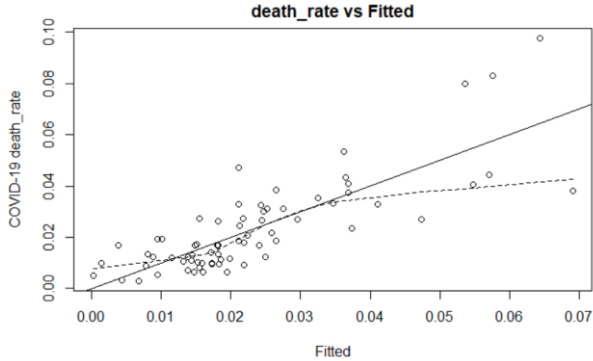
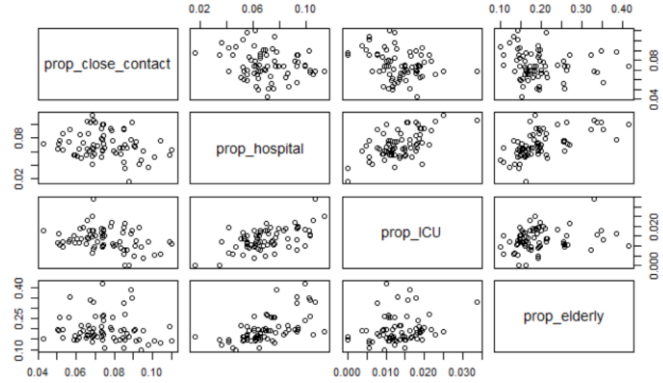
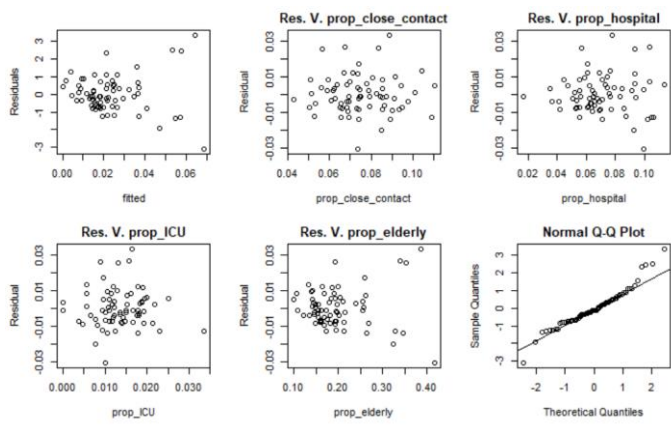
<p>Condition1: response VS. fitted value plot</p> 	<p>Interpretation: Scatterplots are a bit deviated from the identity function, which may require us to transform the response.</p>
<p>Condition 2: pairwise scatterplots</p> 	<p>Interpretation: No weird patterns.</p>
<p>Assumptions: Residual VS. fitted value plot, Residual VS. predictors plots, and normal QQ plot</p> 	<p>Interpretation: Residual plots don't look too bad, except that the residual VS. fitted plot & the residual VS. prop_elderly plot have a bit fanning patterns. Also, the normal QQ-plot looks good overall, except a bit non-normality and few outliers.</p>

Figure 2: assumptions and conditions' checks for transformed model, full2

<p>Condition1: response VS. fitted value plot</p>	<p>Interpretation: Condition 1 seems to satisfy.</p>
---	--

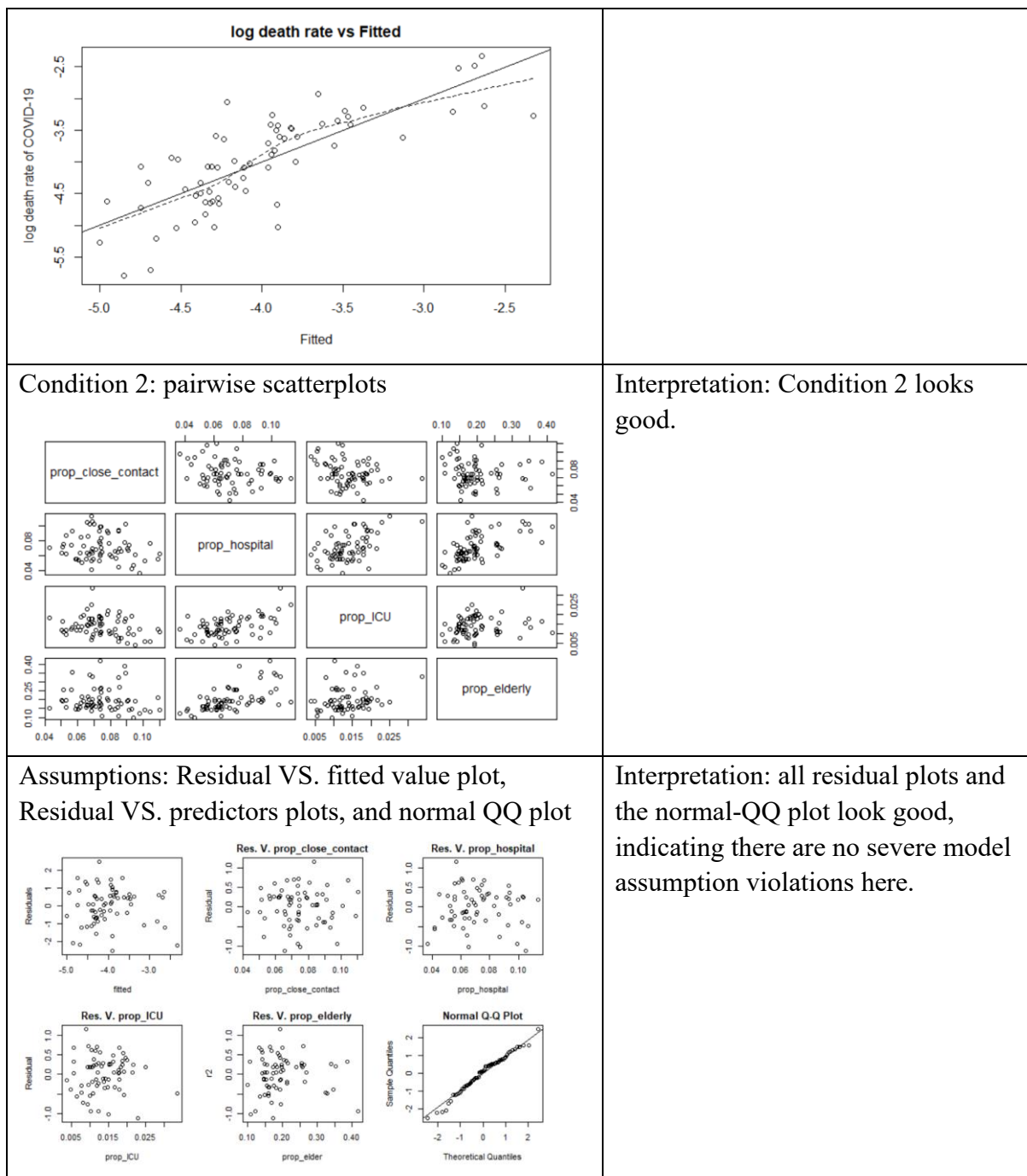
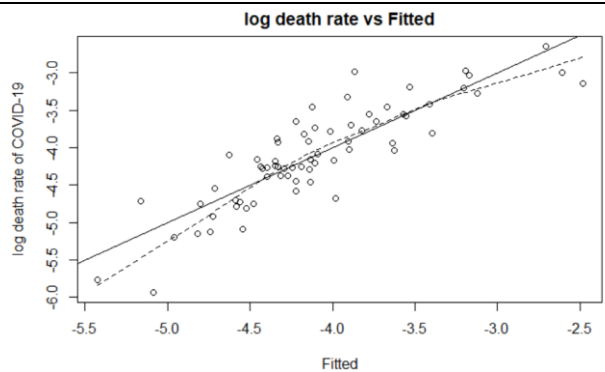
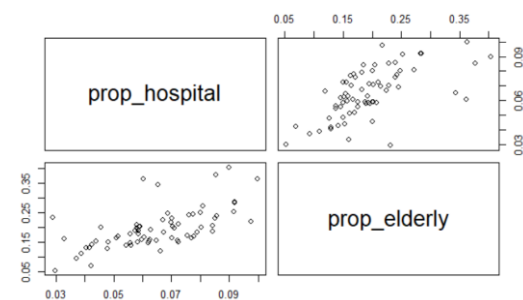


Figure 3: assumptions and conditions' checks for final model on the test set, mod3_test

Condition1: response VS. fitted value plot	Interpretation: Condition 1 satisfies.
--	--

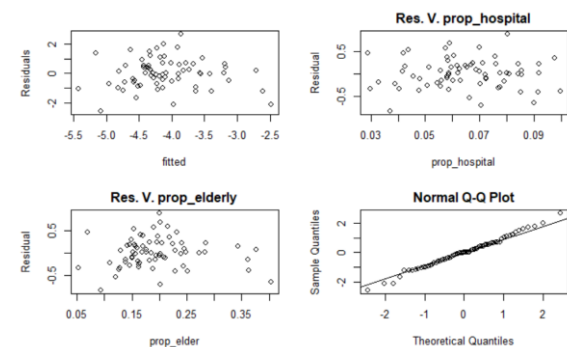


Condition 2: pairwise scatterplots



Interpretation: Condition 2 satisfies.

Assumptions: Residual VS. fitted value plot, Residual VS. predictors plots, and normal QQ plot



Interpretation: All assumptions seem to have no severe violations.