DATA 603 Report - Multiple Regression Analysis of Canadian COVID-19 Data

Devan Constance, Devin Norris, Norma-Jean Rocky and Nathan Tell

12/06/2020

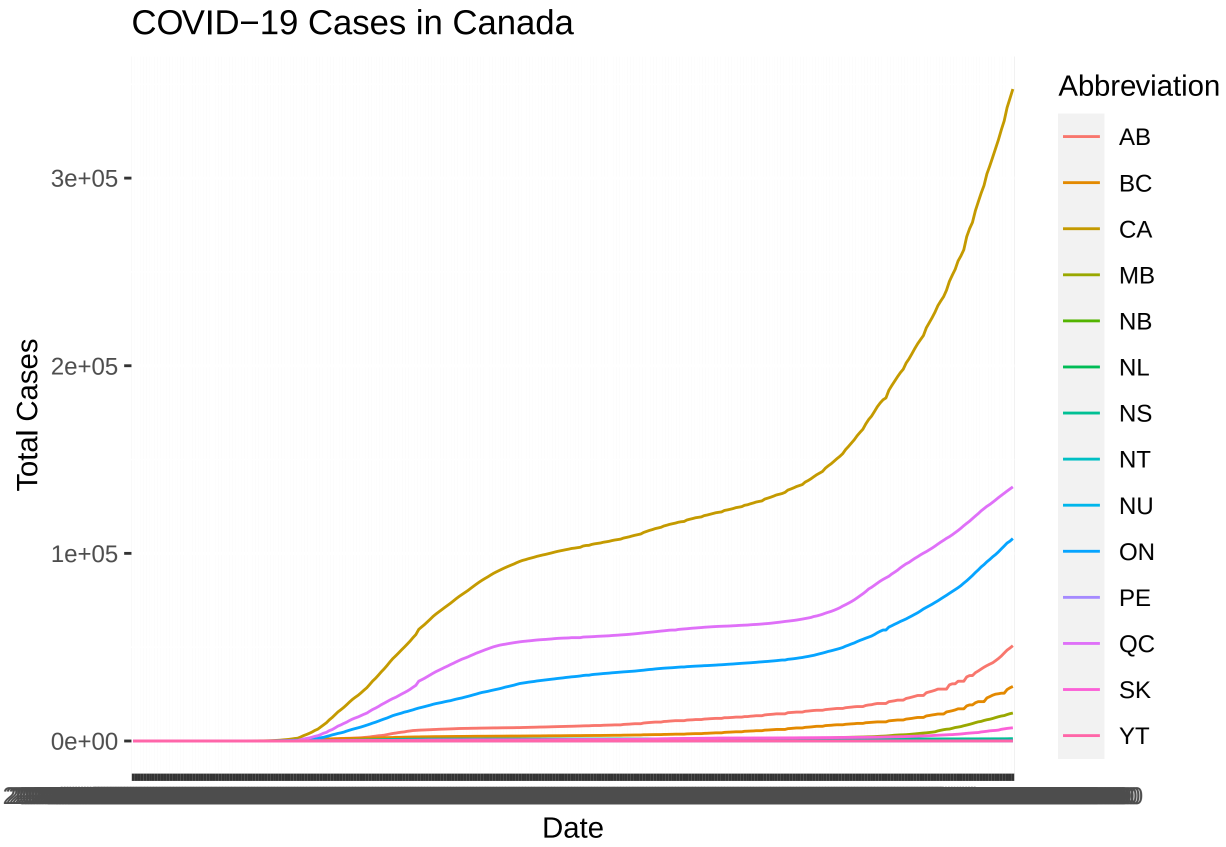
Contents

Introduction

COVID-19 (commonly referred to as the Coronavirus, or just COVID) is an infectious respiratory disease that has affected just about everyone in the world in some way or another. There have been thousands of deaths, widespread lockdowns, travel restrictions, large economic aid programs, and countless other major effects caused by this virus. As seen in the figure below, coronavirus cases are steadily increasing across Canada. This disease has been able to change a whole world culture towards wearing face masks and halted a number of different industries due to restrictions. At the time of writing this paper - December 2020 - our group's exploration into COVID-19 is extremely relevant, however, there is already a lot of research being done on this topic. Therefore, the question must be answered: why is our exploration important? This exploration is important because we are looking at Canada specifically and which unique Canadian demographic variables contribute to infections from the COVID-19 disease.

```
measures=read.csv("/home/jovyan/603/Contextual_Health_Measures_by_Province_100k.csv")
totals=read.csv("/home/jovyan/603/Provincial_Daily_Totals.csv")
```

```
# plotting the provinces and country totals over time
ggplot(data=prov, aes(x=SummaryDate, y=TotalCases, group=Abbreviation,color=Abbreviation)) +
  ggtitle("COVID-19 Cases in Canada") + xlab("Date") + ylab("Total Cases") + geom_line()
```



We will conduct this analysis by looking at provincial-level data from the COVID-19 Canada Open Data website. Additionally, there will be a discussion how our findings relate to some of the other COVID-19 research in Canada. With this ever-changing global pandemic, this report attempts to find a little clarity about the path forward.

Methodology

Datasets

Two data sets from the COVID-19 Canada Open Data website will be used in this modelling exercise. Both contain information collected in 2020 and are from open data sources available for use with attribution.

The first dataset is the Provincial Daily Totals, which will be used as the main data set (ESRI Canada COVID-19 Data Repository, 2020). This first dataset has COVID-19 infection data including daily and cumulative (total) figures. The data is broken down by province, which is particularly relevant because health systems and subsequent social restrictions are under provincial jurisdiction in Canada. Daily and cumulative COVID-19 figures are broken down by active cases, deaths, hospitalization, intensive care unit (ICU) admittance, recovery cases, and number of people tested. While there is a significant amount of data available, our model will focus on the latest reported cumulative cases, which was November 25th, 2020, when the model was created.

The second dataset, Contextual Health Measures by Province, contains several demographic and health factors that we will treat as independent variables and analyze for the (potential) effects they have on COVID-19 infection figures from the first dataset (COVID-19 Related Public Safety Data Content, 2020). The contextual health factors dataset is grouped by province and includes: percentage of population that is rural, percentage of population that are seniors, number of physicians per 100,000, unemployment, percentage of population who are immigrants, percentage of population that are Aboriginal, percentage of population that are post-secondary educated, percentage of children in low income families, percentage of households experiencing food insecurity, percentage of population with diabetes, percentage of population with chronic obstructive pulmonary disease (COPD), percentage of population with high blood pressure, percentage of population with mood disorders, and average number of patient days.

Modelling Plan

The model will test the independent variables located in the Contextual Health Factors table with infections per one hundred thousand people as the dependent variable. The first step will analyze which variables are statistically significant to keep in the model using a stepwise regression test. After that, we will test for higher order variables as well as any interaction terms that may be in the model. We will first begin with an additive model in the form:

$$Cases100k = \beta_0 + \beta_1 RPOP + \beta_2 SEN + \beta_3 PHYS + \beta_4 UNEM + \beta_5 IMM + \beta_6 ABG + \beta_7 PSE$$

$$+\beta_8 LIC + \beta_9 FISC + \beta_1 0 DIA + \beta_{11} COPD + \beta_{12} HBP + \beta_{13} MDO + \beta_{14} PTD + \epsilon$$

Where:

1. RPOP = Percentage of population that is rural
2. SEN = Percentage of population that are seniors
3. PHYS = Family Doctors per 100,000 people
4. UNEM = Unemployment
5. IMM = Percentage of population who are immigrants
6. ABG = Percentage of population that are Aboriginal
7. PSE = Percentage of population that are post-secondary educated
8. LIC = Percentage of children in low income families
9. FISC = Percentage of households experiencing food insecurity
10. DIA = Percentage of population with diabetes
11. COPD = Percentage of population with COPD
12. HBP = Percentage of population with high blood pressure
13. MDO = Percentage of population with mood disorders
14. PTD = Average number of patient days

To build a valid multiple regression model, several assumptions must be met, and will be tested for using the methods below:

1. Linearity - Using residual plots
2. Homoscedasticity (Equal Variance) - Using the Breusch-Pagan Test and a residual plot
3. Multicollinearity - Using variance inflation factor (VIF) analysis
4. Normality - Using the Shapiro-Wilks Test, and normal Q-Q plot
5. Outliers - Using Cook's distance test
6. Independence - Using the residual plots

Results

Variable Selection and Model Creation

The first step that was completed for this regression model was to create the additive base model. To establish this we first test all of the associated variables to determine which were significant. We'll test the following statistical hypothesis, using the individual t-test:

$$Ho: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$
$$(i = 1, 2, ..., 14)$$

```r
# modelling cases per 100k using all possible predictor variables
results=measures
model=lm(cases_100 ~ RuralPop + Seniors + FamilyDrPer100k +Unemployment + Immigrantss
         + Aboriginal + PostSecEd + LowIncome +FoodInsecurity + Diabetes + COPD + HighBP
         + MoodDisorders + PatientDays, data = results)
summary(model)
```

```
##
## Call:
## lm(formula = cases_100 ~ RuralPop + Seniors + FamilyDrPer100k +
##     Unemployment + Immigrantss + Aboriginal + PostSecEd + LowIncome +
##     FoodInsecurity + Diabetes + COPD + HighBP + MoodDisorders +
##     PatientDays, data = results)
##
## Residuals:
## ALL 9 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (6 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2301.55         NA      NA       NA
## RuralPop           28.24         NA      NA       NA
## Seniors          -197.42         NA      NA       NA
## FamilyDrPer100k    15.41         NA      NA       NA
## Unemployment       27.03         NA      NA       NA
## Immigrantss        44.94         NA      NA       NA
## Aboriginal         22.57         NA      NA       NA
## PostSecEd          20.69         NA      NA       NA
## LowIncome          76.88         NA      NA       NA
## FoodInsecurity        NA         NA      NA       NA
## Diabetes              NA         NA      NA       NA
## COPD                  NA         NA      NA       NA
## HighBP                NA         NA      NA       NA
## MoodDisorders         NA         NA      NA       NA
## PatientDays           NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 8 and 0 DF,  p-value: NA
```

Based on the individual t-test summary, there were multiple errors due to singularities when dealing with the entire data set and every possible predictor variable. To narrow down our model to only contain significant predictors, we used stepwise regression. This process returned the best variables to be used in the model based on a p-value of 0.1 or less:

```
# running stepwise regression
stepwise=ols_step_both_p(model,pent=0.1,prem=0.1,details = FALSE)
summary(stepwise$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.40 -164.01   28.64  104.85  482.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2228.644    418.636   5.324 0.000479 ***
## RuralPop       -34.039      5.967  -5.704 0.000293 ***
## MoodDisorders -213.008     58.156  -3.663 0.005214 **
## COPD           288.724    111.689   2.585 0.029451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 256.5 on 9 degrees of freedom
## Multiple R-squared:  0.8199, Adjusted R-squared:  0.7598
## F-statistic: 13.65 on 3 and 9 DF,  p-value: 0.001067
```

The model provided by the stepwise selection method showed reasonable significance at an overall p-value of 0.001067, with each predictor variable having a p-value of less than 0.05. So the first order model resulting from the stepwise regression is as follows:

```
additive=lm(cases_100~RuralPop+MoodDisorders+COPD,data=results)
```

$$\hat{Cases}100k = 2228.64 - 34.039RuralPop - 213.008MoodDisorders + 288.724COPD$$

This means that only the rural population percentage, prevalence of mood disorders and the percentage of the population with COPD were significant factors in predicting the number of COVID cases per 100,000 people. Additionally the significance of each variable is within a reasonable evaluation that a further test of significance using the partial f-test would not be needed. This additive model provided a starting point, with a residual mean squared error (RMSE) of 256, and an adjusted R-squared value of 0.7598 or 75.98%.

Multicollinearity Assumption

```
imcdiag(additive,method="VIF")
```

```
##
## Call:
## imcdiag(mod = additive, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                  VIF detection
## RuralPop      1.3987         0
## MoodDisorders 1.6891         0
## COPD          2.1315         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

Before further analysis was conducted, a test for multicollinearity using the VIF function was completed. The results from this function show that there was no multicollinearity in the model. Each of the values had a value between 1.39 and 2.13. This is not ideal, as the values suggest mild multicollinearity is present, however, they have also not reached the critical value of 5 meaning there is no need for correction for any of the variables. Therefore the model passes the multicollinearity assumption.

Possible Interaction Terms

The next step involved testing for possible interaction between the predictor variables. The model we tested and the statistical hypotheses used to test for interaction terms are given as follows:

$$\hat{Cases}100k = \beta_0 + \beta_1 RPOP + \beta_2 MDO + \beta_3 COPD + \beta_4 RPOP * MDO + \beta_5 MDO * COPD + \beta_6 RPOP * COPD$$

$$Ho : \beta_i = 0$$
$$Ha : \beta_i \neq 0$$
$$(i = 1, 2, ..., 6)$$

```
inter=lm(cases_100~(RuralPop+MoodDisorders+COPD)^2,data=results)
summary(inter)
```

```
##
## Call:
## lm(formula = cases_100 ~ (RuralPop + MoodDisorders + COPD)^2,
##     data = results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -312.47 -159.61  -26.42  125.74  446.12
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -248.651   3150.138  -0.079    0.940
## RuralPop                 -18.178     68.879  -0.264    0.801
## MoodDisorders           -116.825    243.609  -0.480    0.649
## COPD                    1081.299    734.180   1.473    0.191
## RuralPop:MoodDisorders     3.711      4.669   0.795    0.457
## RuralPop:COPD            -10.340     13.486  -0.767    0.472
## MoodDisorders:COPD       -47.809     59.320  -0.806    0.451
##
## Residual standard error: 279.3 on 6 degrees of freedom
## Multiple R-squared:  0.8576, Adjusted R-squared:  0.7153
## F-statistic: 6.024 on 6 and 6 DF,  p-value: 0.02304
```

From this test we found that no interaction terms were significant, as all variables had a non-significant p-value greater than 0.05. From this, we can infer that our intial additive model is still the best model up until this point.

Possible Higher Order Terms

After it was determined that there were no interactions present, an additional improvement check was made to see if there were any significant higher order terms that improved our base additive model. From those variables we would check if the higher order terms of those variables as well as any interactions were necessary to keep in the model using individual t-tests. We checked the following three models:

$$\hat{Cases}100k = \beta_0 + \beta_1 RPOP + \beta_2 MDO + \beta_3 COPD + \beta_4 RPOP^2$$

$$\hat{Cases}100k = \beta_0 + \beta_1 RPOP + \beta_2 MDO + \beta_3 COPD + \beta_4 MDO^2$$

$$\hat{Cases}100k = \beta_0 + \beta_1 RPOP + \beta_2 MDO + \beta_3 COPD + \beta_4 COPD^2$$

The individual t-test method was again used to test the following statistical hypotheses about each of the models above:

$$Ho: \beta_4 = 0$$

$$Ha: \beta_4 \neq 0$$

```r
highermodel=lm(cases_100~RuralPop+I(RuralPop^2)+MoodDisorders+COPD,data=results)
summary(highermodel)
```

```
##
## Call:
## lm(formula = cases_100 ~ RuralPop + I(RuralPop^2) + MoodDisorders +
##     COPD, data = results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -336.08 -115.05   -0.10   98.91  390.44
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1637.6577   574.6904   2.850  0.02149 *
## RuralPop         8.6563    30.5647   0.283  0.78420
## I(RuralPop^2)   -0.6782     0.4772  -1.421  0.19298
## MoodDisorders -245.0702    59.5523  -4.115  0.00337 **
## COPD           365.1752   118.7300   3.076  0.01522 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 243.1 on 8 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.7843
## F-statistic: 11.91 on 4 and 8 DF,  p-value: 0.001893
```

```
highermodel=lm(cases_100~RuralPop+MoodDisorders+I(MoodDisorders^2)+COPD,data=results)
summary(highermodel)
```

```
## 
## Call:
## lm(formula = cases_100 ~ RuralPop + MoodDisorders + I(MoodDisorders^2) + 
##     COPD, data = results)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -353.0 -145.0   30.4  119.0  459.1 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1372.23    2439.03   0.563 0.589112    
## RuralPop             -34.03       6.28  -5.419 0.000632 ***
## MoodDisorders         -9.96     572.05  -0.017 0.986536    
## I(MoodDisorders^2)   -12.88      36.09  -0.357 0.730327    
## COPD                 310.29     132.15   2.348 0.046826 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 269.9 on 8 degrees of freedom
## Multiple R-squared:  0.8227, Adjusted R-squared:  0.734 
## F-statistic:  9.28 on 4 and 8 DF,  p-value: 0.004241
```

```
highermodel=lm(cases_100~RuralPop+MoodDisorders+COPD +I(COPD^2),data=results)
summary(highermodel)
```

```
## 
## Call:
## lm(formula = cases_100 ~ RuralPop + MoodDisorders + COPD + I(COPD^2), 
##     data = results)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -343.3 -129.2   30.7  108.4  448.1 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    738.660   1338.635   0.552 0.596158    
## RuralPop       -31.590      6.213  -5.084 0.000948 ***
## MoodDisorders -200.625     57.980  -3.460 0.008564 ** 
## COPD           921.686    552.250   1.669 0.133677    
## I(COPD^2)      -73.335     62.714  -1.169 0.275918    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 251.4 on 8 degrees of freedom
## Multiple R-squared:  0.8462, Adjusted R-squared:  0.7692 
## F-statistic:    11 on 4 and 8 DF,  p-value: 0.002456
```

Using the individual t-test on higher order models resulted in p-values greater than 0.05 for every higher order term, suggesting they should not be added to our model.

Best Model: Additive

No significant interaction or higher order terms were found, leaving us with the following model:

```r
bestmodel=lm(cases_100~RuralPop+MoodDisorders+COPD,data=results)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = cases_100 ~ RuralPop + MoodDisorders + COPD, data = results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.40 -164.01   28.64  104.85  482.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2228.644    418.636   5.324 0.000479 ***
## RuralPop       -34.039      5.967  -5.704 0.000293 ***
## MoodDisorders -213.008     58.156  -3.663 0.005214 **
## COPD           288.724    111.689   2.585 0.029451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 256.5 on 9 degrees of freedom
## Multiple R-squared:  0.8199, Adjusted R-squared:  0.7598
## F-statistic: 13.65 on 3 and 9 DF,  p-value: 0.001067
```

$$\hat{Cases}100k = 2228.64 - 34.039 RuralPop - 213.008 MoodDisorders + 288.724 COPD$$

Now that the final model has been set, the regression assumptions can be tested to see whether this model is valid. As seen above in the model summary, the RMSE was 256.5 on 9 degrees of freedom and the model had an adjusted R-squared value of 75.98% meaning that, before any tests have been done, this model does look promising.

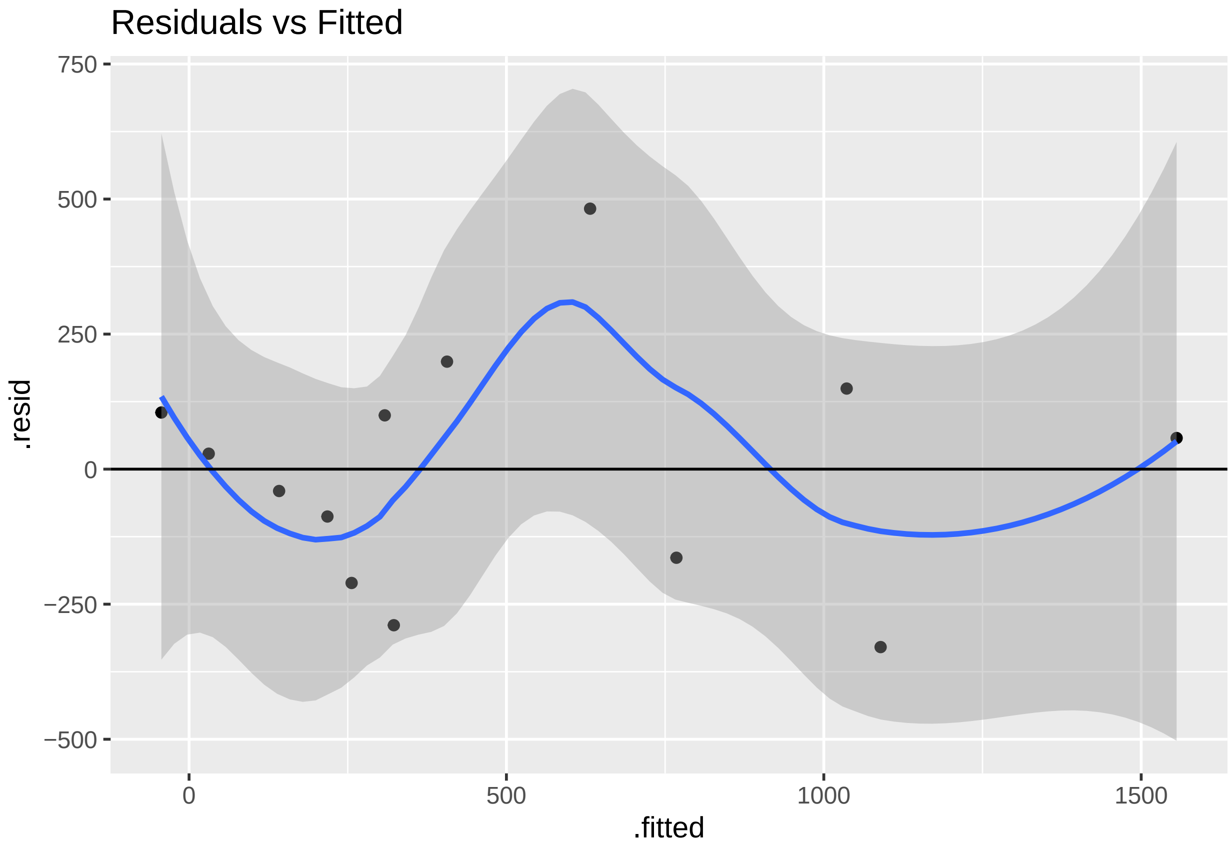
Linearity Assumption

The first test checks for linearity between the dependent and independent variables, using residual plots (as shown in the figure below). Our statistical hypothesis for this check are as follows:

$$Ho : \text{Linearity assumption holds true}$$
$$Ha : \text{Linearity assumption is NOT true}$$

If any patterns or grouping within the residuals were found, then the model would be considered non-linear and we could consider applying a non-linear transformation of the predictors and/or integrating higher-order terms.

```r
ggplot(bestmodel, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth(method= 'loess',formula = y~x)+
  geom_hline(yintercept = 0) +ggtitle('Residuals vs Fitted')
```



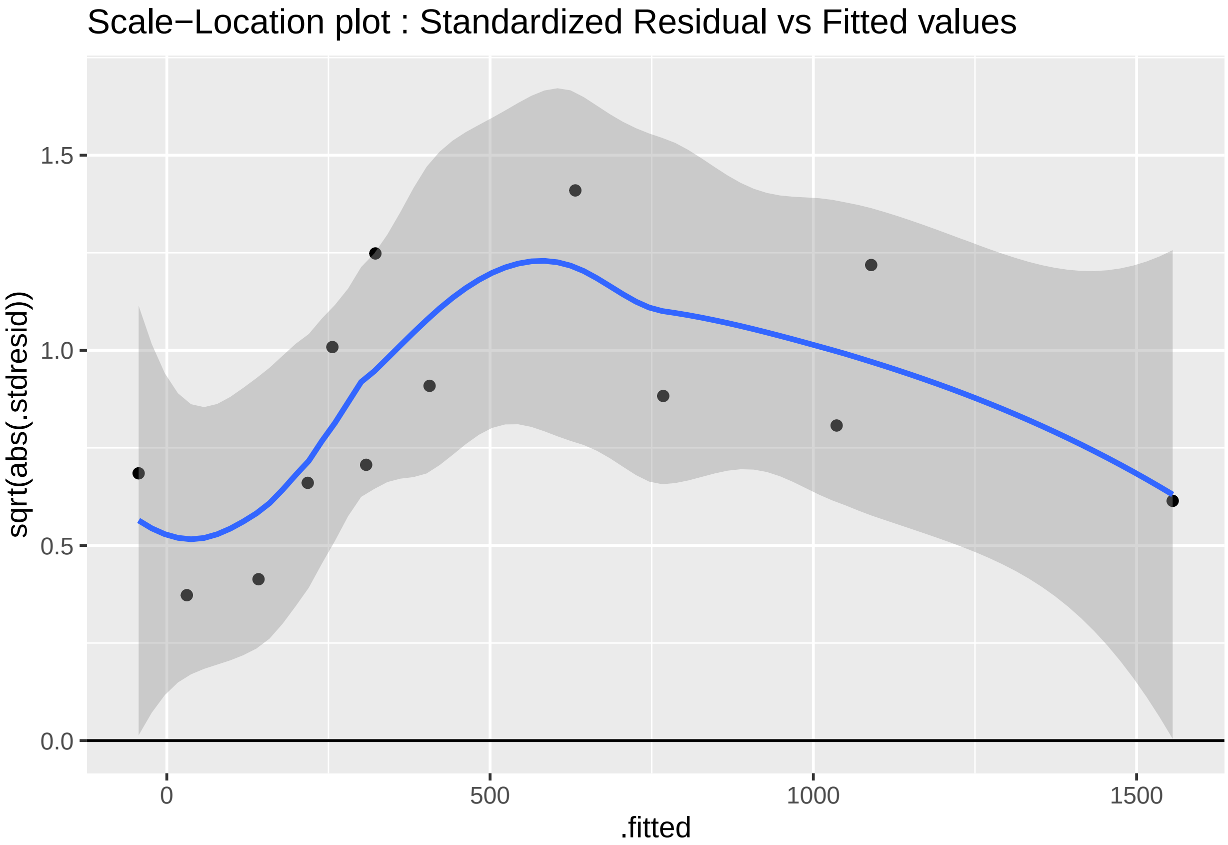
The residual plot does not show any discernible pattern to suggest it is non-linear, but given we only have 13 data points, the plot also does not provide much insight into whether the relationship is truly linear. For our model, we will assume that the relationship is approximately linear and thus the linearity assumption has been met.

Equal Variance Assumption

To assess if heteroscedasticity is present in our data, we prepared a Breusch-Pagan test as well as a residual plot to check if there is constant variance across the data. The hypotheses for this test were:

$$Ho: \text{Heteroscedasticity is not present}$$
$$Ha: \text{Heteroscedasticity is present}$$

```
ggplot(bestmodel, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(method= 'loess',formula = y~x)+
   ggtitle("Scale-Location plot : Standardized Residual vs Fitted values")
```



```
bptest(bestmodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bestmodel
## BP = 1.9528, df = 3, p-value = 0.5823
```

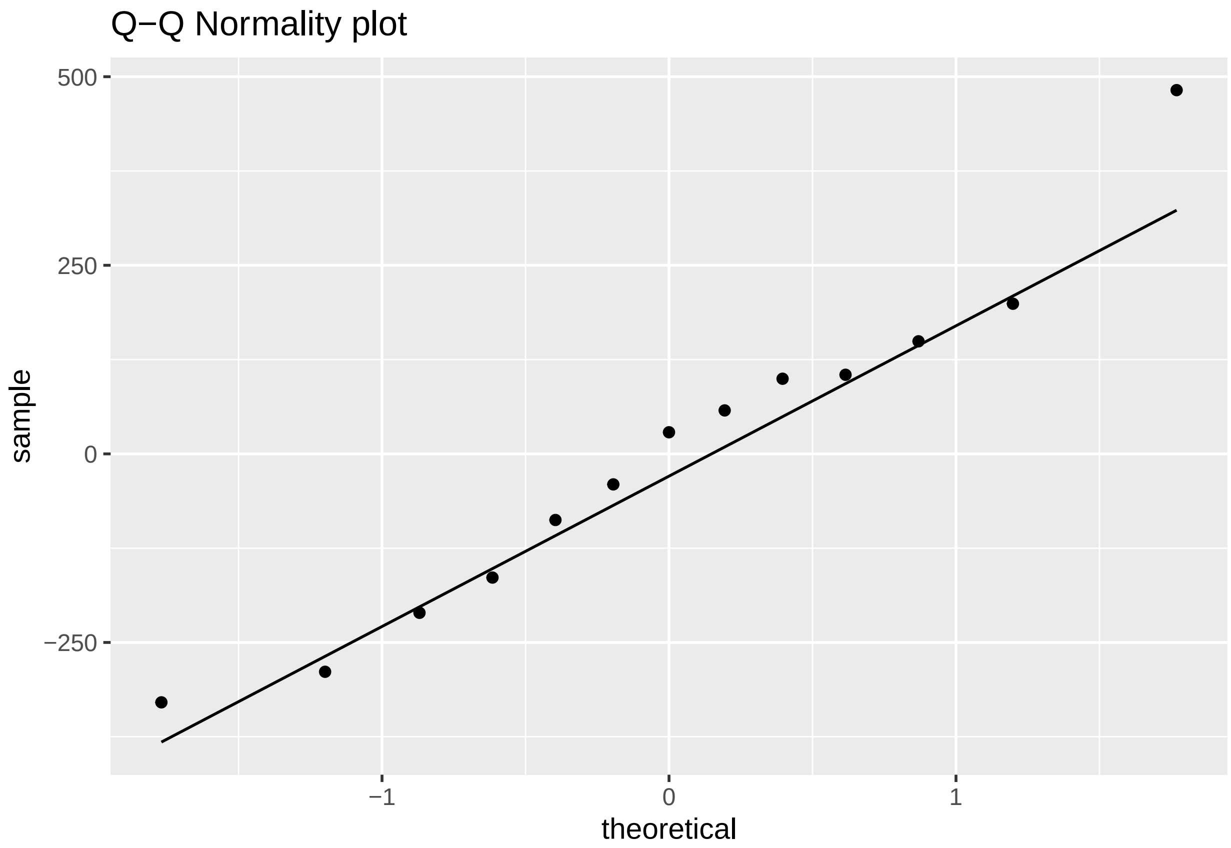
The first part of this test was to look at the residual plot. The plot seems to have fairly equal scatter and therefore, does not clearly show any signs of the residuals being homoscedastic. To assess the homoscedasticity of our data mathematically, we conducted the Breusch-Pagan (BP) test and it returned a p-value of 0.5823. Using an alpha value of 0.05, this p-value shows that there is not enough evidence to reject the null hypothesis. While the graph does not present a clear conclusion, the BP test suggests that we are safe in assuming that heteroscedasticity is not present here.

Normality Assumption

The fourth assumption to verify was normality, with the following hypotheses:

$$Ho: \text{ The sample data is normally distributed}$$
$$Ha: \text{ The sample data is NOT normally distributed}$$

```
#Normality plot
ggplot(results, aes(sample=bestmodel$residuals)) +
  stat_qq() +labs(title= 'Q-Q Normality plot')+
  stat_qq_line()
```



```
shapiro.test(residuals(bestmodel))
```

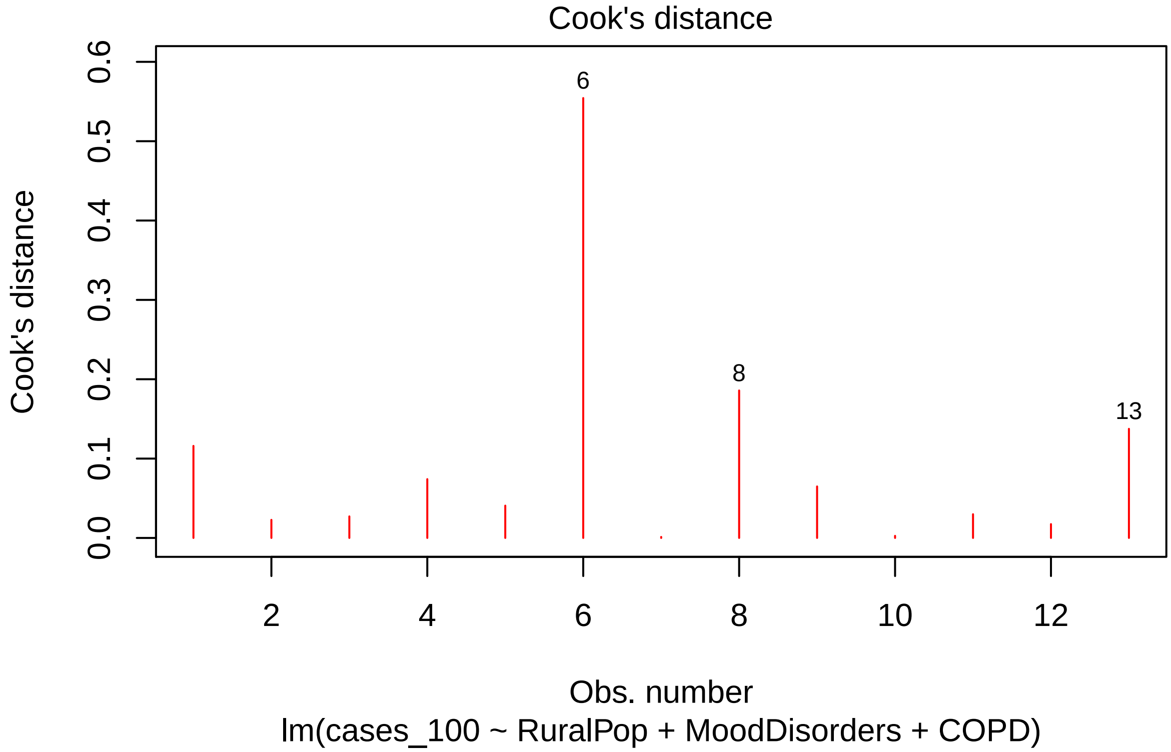
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bestmodel)
## W = 0.96467, p-value = 0.8236
```

A Shapiro-Wilks test for normality was performed and resulted in a p-value of 0.8236. Using an alpha value of 0.05, we fail to reject the null hypothesis, suggesting that the residuals are normally distributed. This can be visualized with a normal Q-Q plot, and can be seen in the figure above. The data falls closely in-line with the normality line through the middle. There are some values that have diverted from the normal line through the middle of the plot, as well as in the lower left and upper right corners, but they are not enough to skew the data significantly. Thus, given the Shapiro-Wilks test results and the Q-Q plot, the data can be considered normally distributed.

Outliers

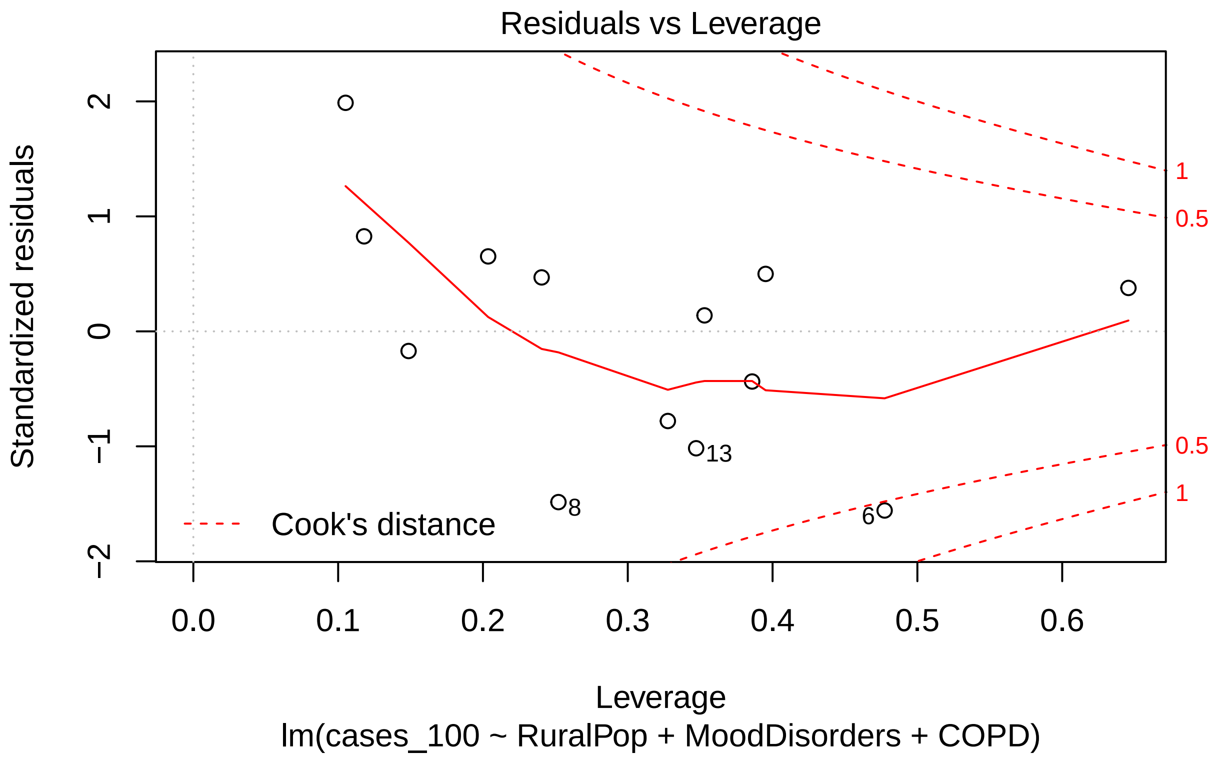
Cook's Distance can help us identify any significant outliers in the data. As seen in the Residuals vs. Leverage plot below, we first tested for values with Cook's distance scores greater than 1 and based on the observation values none were considered influential.

```
plot(bestmodel,pch=4,col="red",which=c(4))
```



As seen in the figure above, the closest value to being considered an outlier pertains to value #6 representing the Northwest Territories, which is marginally above 0.5 in Cook's distance but is less than 1 and so may be considered influential. This is also another significant population in Canada that could be skewing our results. We can see this more clearly when we look at the leverage plot below.

```
plot(bestmodel,which=5)
```

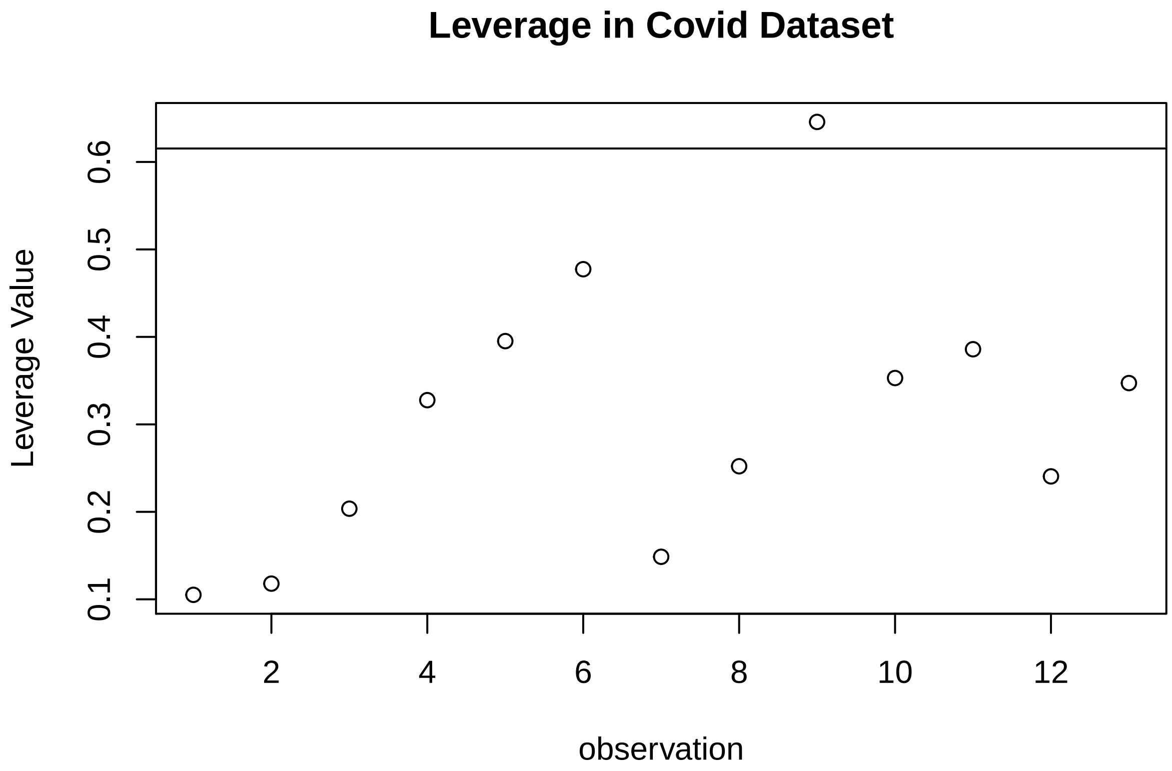


The Northwest Territories observation was marginally considered an outlier according to the leverage plot (exceeding a Cook's distance of 0.5) and so we will test the impact this data point has on our model.

```
lev=hatvalues(bestmodel)
p = length(coef(bestmodel))
n = nrow(results)
outlier = lev[lev>(2*p/n)]
print(outlier)
```

```
##         9
## 0.6457534
```

```
plot(rownames(results),lev, main = "Leverage in Covid Dataset", xlab="observation",
    ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
```



```
outliers=c(9)
data2=results[-outliers,]
newbestmodel=lm(cases~RuralPop+MoodDisorders+COPD,data=data2)
summary(newbestmodel)
```

```
##
## Call:
## lm(formula = cases ~ RuralPop + MoodDisorders + COPD, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16962 -11282  -1473   3245  38468
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   125979.3    39945.3   3.154  0.01352 *
## RuralPop       -2708.5      553.9  -4.890  0.00121 **
## MoodDisorders -14577.7     6417.8  -2.271  0.05277 .
## COPD           26253.5    10500.6   2.500  0.03693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18410 on 8 degrees of freedom
## Multiple R-squared:  0.7661, Adjusted R-squared:  0.6784
## F-statistic: 8.734 on 3 and 8 DF,  p-value: 0.006639
```

From the above summary, removing the outlier produced a model that was far less effective with a RMSE of 18,410 and adjusted R-square of 67.84%. Furthermore the overall model contains a potentially non-significant variable which would make the model itself invalid. We will leave our original best fit (additive) model unchanged with the Northwest Territories data in the dataset.

Independence Assumption

The error terms in the multiple regression model should not be correlated, and we did not identify any discernible patterns or trends in the residual plots above to indicate otherwise. Additionally, while there is time-series data available in the datasets, our model used cumulative figures, not daily numbers, for the analysis and so there are no concerns about correlation of the errors (and therefore non-independence) related to time.

Summary of Multiple Linear Regression Assumption Tests

The model has passed all of the multiple linear regression assumptions and can therefore be considered valid and useful for predictive purposes. We have accounted for a considerable proportion of the fluctuation in cases by having an adjusted R-squared value of 75.98% and therefore it does not seem as if anything is missing in order for us to compare our findings to other models, which will be discussed next (see Discussion section).

Discussion

The next step in our exploration is to talk about the resulting best fit model and the variables present in the model. We are doing this so we can discuss the nature of our variables and why they may relate to COVID-19 cases. The final component of the discussion will consider how our best fit model compares to other models that have been created for Canadian data. The reason that this will be done is to see which variables are important to the other models, allowing us to see what information may be missing from our own model. With that information it may be possible to improve our final model.

Model Discussion

While this report presents a succinct discussion of a fitted model for COVID-19 cases per 100,000 people in Canada, we previously conducted various iterations of the model to identify the appropriate dependent and independent variables. For example, we attempted using total COVID-19 counts as the response variable, while including population counts in the predictive variables, however, the resulting model essentially suggested that a province with a higher population would have higher total case counts. This seemed quite obvious and didn't provide many meaningful insights, particularly for the health and demographic measures that we were actually interested in studying.

We also wanted to remove the effect of time (which would have encroached on the independence assumption), so used the latest COVID-19 cumulative case counts at the time of model development rather than looking at the change in COVID-19 counts over time. It was useful to have comparative figures across provinces, so the provincial cumulative case counts and population figures were used to compute a case count per 100,000 people to use as our dependent variable.

The final independent variables in the fitted model included:

1. RuralPop - Percentage of the population that is rural
2. MoodDisorders - Percentage of the population with mood disorders
3. COPD - Percentage of the population with chronic obstructive pulmonary disease

Below is a brief discussion about each variable and it's related coefficient in our model, in order to better understand the real-life impacts of these variables on COVID-19 cases in Canada.

RuralPop: Our model suggests that for every 1% increase in a province's rural population, COVID-19 cases would decrease by 34 cases per 100,000. This is a logical interpretation as communicable diseases, such as COVID-19, spread more rapidly through denser populations, whereas rural populations with lower population density would see a slower spread of the disease. This has been observed in large cities around the world such as Toronto, New York, London, where the disease spreads rapidly and infects many people who are in close proximity to each other.

Mood Disorders: The second variable that we have in our model is mood disorders. For every 1% increase in the population that has mood disorders, COVID-19 cases would decrease by 214 cases per 100,000. The (negative) relationship between mood disorders and COVID-19 cases was the most interesting insight from our model as it is counter-intuitive and does not align with recent research by Wang et al. about the increased COVID-19 risk associated with mental health disorders. According to Statistics Canada, "mood disorder refers to whether or not, in some specified period, the person classified as meeting the criteria for either of the mood disorders measured, specifically: major depressive episode or bipolar disorder. Mental conditions or problems were identified through a set of questions pertaining to the feelings, the symptoms, severity, the intensity and the impact relative to each of the disorders." Note that the mood disorder data was collected in 2018, well before the first case of COVID-19 was discovered, so we are only modelling the provincial population with pre-existing mood disorders.

COPD: The final variable in our model is the percentage of the population with COPD, suggesting that for every 1% increase in the population with COPD, COVID-19 cases will increase by 289 cases per 100,000. COPD is an obstructive lung disease that affects ~4% of the population nationally, and having COPD greatly increases your risk of severe illness from COVID-19. However, while it is expected that a higher prevalence of

COPD would result in more hospitalizations and deaths, it is interesting that our model suggests that COPD prevalence is considered a predictive variable for case counts. According to Leung et al, estimating the excess risk for COPD patients to contract COVID-19 is an ongoing and "challenging exercise."

Model Comparison

There are two major studies that we will compare the results from our model with. The first was produced by the Government of Alberta's (now referred to as the Alberta model), and is a COVID-19 modelling paper similar to our own (Government of Alberta, 2020). The paper itself has only a small section on COVID-19 cases, because most of the paper is meant to show the capacity of Alberta hospitals and the action that will be taken by the Albertan Government to prevent the spread of COVID, but it still does have modelling on COVID-19 cases.

The results from this model are split into three different scenarios: probable, elevated and extreme. These predictions provided in April 2020 show that for all events the prediction is that the peak of the disease in may 2020, in November we know that this is not the case, cases have continued to rise since April and will most likely continue to do so as the disease is able to reach more people. The report itself does not talk about the variables used to present this information, but they all show quick peaks and then long drops over time in cases. From this very limited analysis, it is already easy to tell that our model performs very positively when compared to this one. This is because our information is much more current and about the variables that go into COVID cases in Canada, as opposed to the very early estimates that would skew the data.

The second model that we will be looking at is more recent and more relevant to our own exploration. This second modelling report was authored by the Public Health Agency of Canada on November 20th, 2020 (now referred to as the Canada model), making the timeframe of the data almost exactly to the data our group obtained (Public Health Authority of Canada, 2020). The modelling in this report shows that the likelihood of a resurgence in cases is likely going into the holiday season. This follows our data as well as can be seen in the figure on page 2 where the cases start to rise again at the end of each respective curve.

The modelling process done in this model is more based on other countries. The number of cases in Canada is compared to other countries such as Italy, Belgium, Portugal, the United States and a few other countries that have already had resurgences, pushing the number of cases at the end of the curve up. This is similar to our findings but applies its model with very different information it seems, basing the curve solely on the number of cases so far, and the trends of other countries. Their estimates also show the first peak of the disease in May, confirming the main idea of the Alberta model.

Where our model and the Canada model also differ is the section on response, and the modelling that occurs for the different scenarios that could occur with the disease moving forward. The Canada model shows three separate scenarios, similar to the Alberta model, which shows an increase, maintain, and decrease in contact between people that will drastically shift the curve.

In this sense, this Public Health Agency model is much better than ours and may have better capability to predict what may happen to COVID-19 cases moving forward. On the other hand, our model is much more clear about which variables contribute the most to COVID-19. That said, our fitted model may not be as powerful for prediction as there are many other factors (including public health restrictions and potential vaccinations) that our model does not account for.

Conclusion

Using our selected datasets from the Canada COVID-19 Open Data website, we were able to fit a linear model to predict COVID-19 cases per 100,000 people given various health measures using the stepwise regression procedure. This procedure resulted in a first order model as follows:

$$\hat{Cases}100k = 2228.64 - 34.039RPop - 213.008MDO + 288.724COPD$$

The first order model was first tested for multicollinearity to ensure that there was no correlation between the independent variables that would impact the model. Once the model passed the multicollinearity assumption, we tested for interaction and higher-order terms, however, none of these terms were considered significant (i.e., all p-values $> \alpha = 0.05$) in our model and we reverted to the additive model as the best fit model.

To determine the usability of this model, we then assessed it for the various multiple linear regression assumptions, including linearity, homoscedasticity, independence, normality, and outliers, and found that all assumptions were met. That said, we only had 13 data points in our model, so we verified our findings against some of the current research and models on COVID-19 cases. Ultimately, our model provided interesting insights on health measures that potentially influence COVID-19 case counts (per 100,000) and identified some potential topics for future study.

The final step in the process of designing a model is to look at what sort of issues have occurred and improve upon the flaws that have been found. In the case of this model, the best way to improve upon it is by adding more data. The health measures that are included in this dataset such as COPD, and mood disorders, may not be the only significant predictors of COVID cases, as our model's adjusted R-squared value was around 76%. This means that 24% of the variance in COVID cases are unexplained by our model, and other possible predictors would need to be looked into. Another course of action, in our opinion, is to increase the range of the experiment. The skewed results that were present in the model were most often linked to how there were only 14 data points in the model, one for each province and territory. Although this report was meant to find results about Canada specifically, the addition of more regions and more geographic variability would most likely help to stabilize the model and lead to better results from the assumption tests performed in the results section. Then once this information is found it could be compared to different statistics in regards to COVID-19 such as deaths, hospitalizations, and any number of other COVID related statistics.

Bibliography

COVID-19 Related Public Safety Data Content (Updated: April 28, 2020), *Contextual Health Measures By Province* [Online] Available at: https://resources-covid19canada.hub.arcgis.com/datasets/exchange::contextual-health-measures-by-province?selectedAttribute=PatientDays (Accessed: November 23, 2020)

ESRI Canada COVID-19 Data Repository (Updated: November 23, 2020), *Provincial Daily Totals* [Online] Available at: https://resources-covid19canada.hub.arcgis.com/datasets/provincial-daily-totals (Accessed: November 23, 2020)

Government of Alberta (Updated: April 8, 2020), *COVID-19 Modelling* [Online] Available at: https://www.alberta.ca/assets/documents/covid-19-case-modelling-projection.pdf (Accessed: November 23, 2020)

Government of Alberta, *Enhanced Public Health Measures* [Online] Available at: https://www.alberta.ca/enhanced-public-health-measures.aspx (Accessed: November 26, 2020)

Leung JM, Niikura M, Yang CWT, et al. (2020) *COVID-19 and COPD.* European Respiratory Journal, 56: 2002108. DOI: https://doi.org/10.1183/13993003.02108-2020 (Accessed: December 4, 2020)

Public Health Agency of Canada (Updated: November 20, 2020) *Update on COVID-19 Modelling in Canada: Epidemiology and Modelling* [Online] Available at: https://www.canada.ca/content/dam/phac-aspc/documents/services/diseases-maladies/coronavirus-disease-covid-19/epidemiological-economic-research-data/update-covid-19-canada-epidemiology-modelling-20201120-eng.pdf (Accessed: December 2, 2020)

Statistics Canada (Updated: September 10, 2013) *Canadian Community Health Survey - Mental Health.* [Online] Available at: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=assembleDESurv&DECId=463355&RepClass=583&Id=119789&DFId=180520 (Accessed: December 4, 2020)

Wang, Q. et al. (October 2020) *Increased risk of COVID-19 infection and mortality in people with mental disorders: analysis from electronic health records in the United States.* World Psychiatry 2021. DOI: https://onlinelibrary.wiley.com/doi/10.1002/wps.20806