Term Project:

COVID-19 Impact On The Economy of USA and Canada

Dec 2, 2020

1 Introduction

1.1 Background

This year, the Covid-19 pandemic has caused vast challenges in our lives. It has impacted the way we socialize and behave, affected the health of millions, and also led to a severe economic recession that has put hundreds of thousands of people and businesses in financial distress. Recent announcements from research institutions about new vaccines and their high effectiveness, such as from Moderna, Pfizer and Oxford, has caused plenty of optimism amongst the population. Despite great optimism for the release of vaccines that will solve the majority of the problems, it is worth reflecting upon the lessons that can be learned for the future, so the impact that a pandemic can have in the economy and in our lives can be minimized.

1.2 Objectives

We are interested in how the economy, measured by GDP, is affected differently across North American states/provinces due to the ongoing pandemic caused by the Covid-19 virus outbreak, and specifically, what factors might explain the difference in the decline of these economies. Thus, the main objective of this project is to seek for significant relationships between covariates and the extent of the economic downturn (denoted as *GDP drop*) in Canadian provinces and American states.

1.3 Covariates Selection

The covariates chosen for the analysis were the number of cases per 100,000 people(*cases*), the Human Development Index (*HDI*), Unemployment Rate in 2019 (*unemployment*), Urban population as a percentage of total population(*urban*), and the contribution to the GDP from the tourism sector in percentage(*tourism*). It was also considered appropriate to add a categorical variable (*country*) to acknowledge the probable differences between the two countries (USA = 0, Canada = 1).

Given the economic downturn is caused by a public health crisis, *cases* and *HDI* are included to seek for the possible connection. *Urban* population percentage is included as it reflects both the economic structure of a state/province and how people interact socially. We suspect *Tourism* would have high explanatory power as mainstream media paints a gloomy picture of the tourism sector in this pandemic.

Due to the complex nature of this economic problem, we acknowledge that our set of explanatory variables may not adequately explain the variance in the response variable, but is nonetheless useful for shedding some light on the relations among the factors, despite possibly not achieving a high R^2.

1.4 Data Collection

All data we used are retrieved online. Detailed descriptions are as follows:

- *GDP drop*: The percentage of decrease of GDP of American states and Canadian provinces are retrieved respectively from the U.S. Department of Commerce and Statista, which measures the regression percentage in GDP of a state/province from Dec 31, 2019 to Jun 30, 2020. The percentage change is calculated as:

$$GDP\ drop = 1 - \frac{GDP\ of\ Jun\ 30}{GDP\ of\ Dec\ 31, 2019}$$

- *cases*: Cumulative count of Covid-19 cases until June 30, 2020 for American states and Canadian provinces are retrieved from the website of CDC and CTV news respectively, which is measured in cases per 100,000 people. The data for Canada is directly available, while the statistics for the US is calculated by the formula:

$$cases = \frac{Cumulative\ covid19\ cases}{total\ population(state/province)} * 100,000$$

- *HDI*: Sub-national Human Development Index data for American states and Canadian provinces are collected from Global Data Lab, with the most recent data being the one from year 2018. The Human Development Index covers three data dimensions, which are life expectancy, education, and per capita income. HDI is calculated as the geometric mean of the normalized Life Expectancy Index (LEI), Education Index (EI) and Income Index (II):

$$HDI = \sqrt[3]{(LEI * EI * II)}$$

- *unemployment*: The Unemployment rate of each economy (provinces and states) in the year 2019 was collected from Statista. Someone unemployed implies that they are willing to work and have been actively looking for a job without success. We are using it as a measurement of how the state of the economy was and its efficiency prior to the pandemic and the lockdown. The more the unemployed, the lower the output and hence growth in the economy. It is calculated as:

$$Unemployment\ rate = \frac{People\ unemployed}{People\ in\ the\ labour\ force}$$

- *tourism*: The tourism contribution to the GDP of each economy as a proportion of total GDP (provinces and states) for Canada in 2014 and the US in 2018 was collected from the SP Global web page for the states, and Statistics Canada for the provinces. It is worth mentioning that even though the years differ, since it is a measure of the structure of the economy as a proportion, we expect that it stays fairly similar for some years.

$$Tourism = \frac{GDP\ from\ Tourism}{Total\ GDP}$$

- *urban*:

  US: Urban Population Percentage in the US is retrieved from the website of Iowa State University. It measures the percentage of a state's population that lived in an urban area in 2010. Collected every decade, 2010 is the most recent data available at this source. We suspect urbanization rate would be higher in 2020 across the states, but the rate of change should be moderate and similar across the states. Hence we adopt the data.

  Canada: Urban Population Percentage in Canada is not directly available. We used related data from Statistic Canada collected in 2019 to calculate the percentages:

$$urban = \frac{population\ in\ census\ metropolitan\ areas}{total\ population}$$

1.5 Caveats

It is worth noting that our data gathered from various sources may not agree in their time being collected. For example, the "urban population percentage" data from the 50 American states are from 2010, while those of Canadian provinces are from 2019. Also, some discrepancy may exist between the two countries in how certain statistics are defined/measured, such as *urban* and *tourism*.

Nevertheless, the data of one country for each variable are collected from the same source, (i.e. all data 50 American states' *unemployment* rates are retrieved from one source, while that of Canadian provinces from another). Thus, including the *country* variable in the model is an effective measure to mitigate some of the effects of having data from different sources.

Yukon, Nunavut, Northwest Territories and PEI of Canada are not included in the analysis given their small size in population and in GDP.

2 Data Analysis

2.1 Data Overview

2.1.1 Boxplots

As of June 30, 2020, Canadian provinces in general report fewer cases than American states.

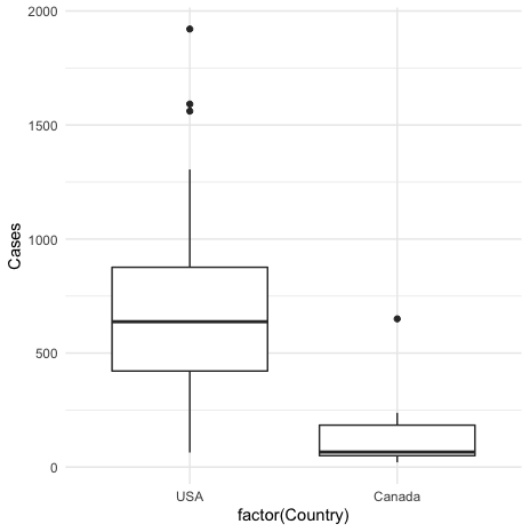

Figure 1: cases by country

Canada provinces and US states experienced similar drops in GDP:

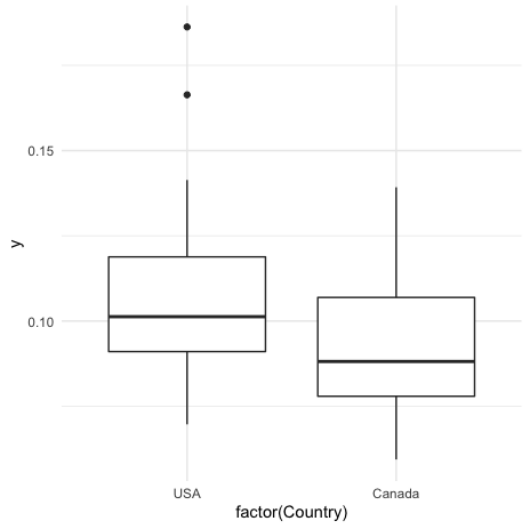

Figure 2: GDP drop by country

2.1.2 Response variable vs covariates

Scatter-plots of *GDP drop* against each predictive variable:

Some outliers are present, and can potentially have high leverage in the multi-covariate model.

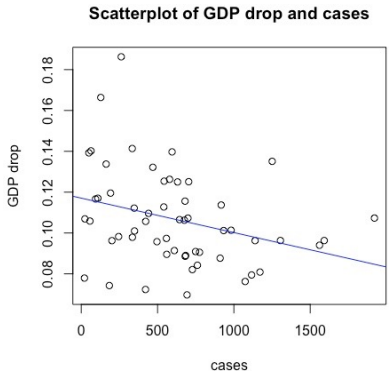

Figure 3: Scatterplot of GDP drop vs cases

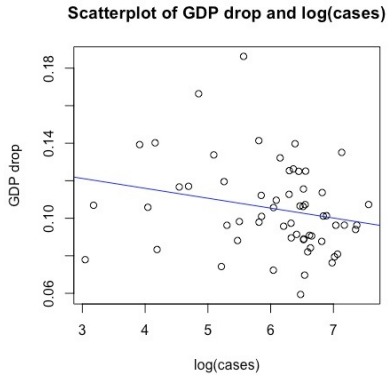

Figure 4: Scatterplot of GDP drop vs log(cases)

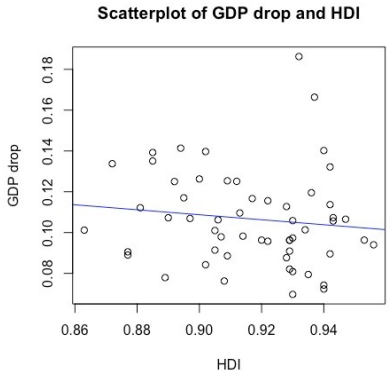

Figure 5: Scatterplot of GDP drop and HDI

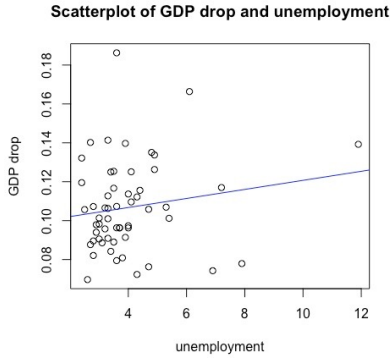

Figure 6: Scatterplot of GDP drop and unemployment

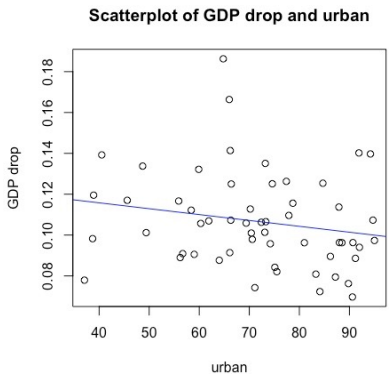

Figure 7: Scatterplot of GDP drop and urban

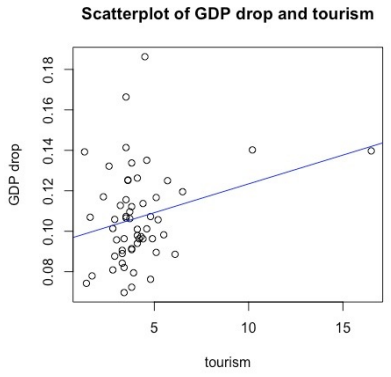

Figure 8: Scatterplot of GDP drop and tourism

Furthermore, the correlation values between variables are calculated:

- correlation between GDP drop and cases: -0.2593148
- correlation between GDP drop and urban: -0.1690294
- correlation between GDP drop and tourism: 0.3113015
- correlation between GDP drop and unemployment: 0.09736781
- correlation between GDP drop and HDI: -0.1044564

As shown above, the correlation between the response variable and each covariate is quite moderate. *Tourism* has the highest correlation with the response variable, at 0.3113015.

2.1.3 Correlation between covariates

Cases, *urban* and *HDI* show relatively high correlation:

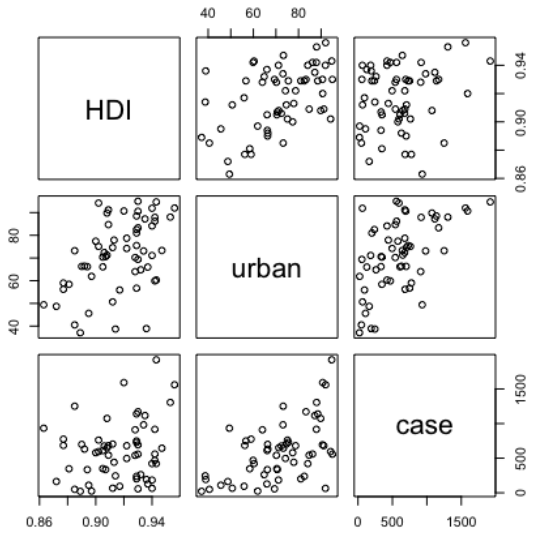

Figure 9: Matrix of scatter-plots

2.2 Model Selection

Full model

We start with a model with all predictive variables:

```
mod.full <- lm(y~urban + tourism + case + HDI + unemploy + country)
```

```
Call:
lm(formula = y ~ urban + tourism + case + HDI + unemploy + country)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043276 -0.013582 -0.002963  0.008684  0.067107

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.318e-02  1.320e-01  -0.251  0.80255
urban       -2.644e-04  2.435e-04  -1.086  0.28263
tourism      3.178e-03  1.455e-03   2.185  0.03343 *
case        -1.756e-05  8.382e-06  -2.095  0.04104 *
HDI          1.415e-01  1.461e-01   0.968  0.33736
unemploy     7.754e-03  2.569e-03   3.019  0.00393 **
country1    -4.000e-02  1.177e-02  -3.399  0.00130 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02009 on 52 degrees of freedom
Multiple R-squared:  0.3543,    Adjusted R-squared:  0.2798
F-statistic: 4.755 on 6 and 52 DF,  p-value: 0.0006244
```

Figure 10: Summary for full model

The coefficients for *case*, *tourism*, *unemployment* and *country* are significant. *Urban* and *HDI* do not add much explanatory power in presence of *cases*, which coincides with their relatively high correlation with *cases*.

Alternative model

```
mod.alt <- lm(y~urban + tourism + country)
```

```
Call:
lm(formula = y ~ urban + tourism + country)

Residuals:
      Min        1Q    Median        3Q       Max
-0.030238 -0.016633 -0.002739  0.013807  0.074513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1269009  0.0152764   8.307 2.76e-11 ***
urban       -0.0004914  0.0001993  -2.465   0.0168 *
tourism      0.0037124  0.0014845   2.501   0.0154 *
country1    -0.0100496  0.0090552  -1.110   0.2719
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02187 on 55 degrees of freedom
Multiple R-squared:  0.1903,    Adjusted R-squared:  0.1461
F-statistic: 4.308 on 3 and 55 DF,  p-value: 0.008441
```

Figure 11: Summary for alternative model

The **alternative model** doesn't include *cases*. Although it indicates significance for *urban*, it yields a lower R^2. There is not enough evidence to support the inclusion of *urban*.

```
mod.alt2 <- lm(y~HDI + tourism + country)
```

```
Call:
lm(formula = y ~ HDI + tourism + country)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036302 -0.014420 -0.004566  0.010662  0.080732

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.211862   0.126285   1.678   0.0991 .
HDI         -0.128509   0.136994  -0.938   0.3523
tourism      0.002989   0.001529   1.955   0.0557 .
country1    -0.005765   0.009268  -0.622   0.5365
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02287 on 55 degrees of freedom
Multiple R-squared:  0.115,     Adjusted R-squared:  0.06669
F-statistic: 2.381 on 3 and 55 DF,  p-value: 0.07936
```

Figure 12: Summary for alt2 model

Alt2 Model also fails to show enough evidence to support the inclusion of *HDI*. Results from **regsubset** from **leaps** confirms the findings above, and we have a **base model**:

```
mod.base <- lm(y~tourism + case + unemploy +country)
```

```
Call:
lm(formula = y ~ tourism + case + unemploy + country)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043957 -0.014315 -0.002686  0.008991  0.069832

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.418e-02  1.261e-02   6.676 1.38e-08 ***
tourism      2.600e-03  1.358e-03   1.915  0.06078 .
case        -2.159e-05  7.081e-06  -3.049  0.00355 **
unemploy     7.283e-03  2.389e-03   3.049  0.00355 **
country1    -3.944e-02  1.150e-02  -3.430  0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01999 on 54 degrees of freedom
Multiple R-squared:  0.3361,    Adjusted R-squared:  0.2869
F-statistic: 6.834 on 4 and 54 DF,  p-value: 0.0001586
```

Figure 13: Summary for base model

Interaction model

We further explore the possibility of including interaction terms. Given the difference in the social and economical structures between the two countries, and the discrepancies in how certain statistics are measured, it may be plausible to fit a model as below:

```
mod.interact <- lm(y~((tourism + case + unemploy) * country)
```

```
Call:
lm(formula = y ~ ((tourism + case + unemploy) * country))

Residuals:
      Min        1Q    Median        3Q       Max
-0.044881 -0.014196 -0.000499  0.008111  0.070206

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.837e-02  1.589e-02   4.931 9.09e-06 ***
tourism            2.511e-03  1.373e-03   1.829  0.07323 .
case              -2.065e-05  7.277e-06  -2.838  0.00651 **
unemploy           8.836e-03  3.666e-03   2.410  0.01960 *
country1          -5.229e-02  4.819e-02  -1.085  0.28302
tourism:country1   1.563e-02  1.507e-02   1.037  0.30450
case:country1     -3.862e-05  3.820e-05  -1.011  0.31677
unemploy:country1 -2.125e-03  5.200e-03  -0.409  0.68448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02011 on 51 degrees of freedom
Multiple R-squared:  0.3655,    Adjusted R-squared:  0.2785
F-statistic: 4.198 on 7 and 51 DF,  p-value: 0.001016
```

Figure 14: Summary for interaction model

It yields a lower adjusted R^2 than the **base model**. Hence this is not adopted.

Log model

Covid *cases* tend to grow exponentially. We suspect $\log(case)$ could be a better measure of the severity of the pandemic.

```
mod.log <- lm(y~(tourism + log(case) + unemploy + country))
```

```
Call:
lm(formula = y ~ (tourism + log(case) + unemploy + country))

Residuals:
      Min        1Q    Median        3Q       Max
-0.041495 -0.012250 -0.000505  0.009327  0.069904

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.149602   0.026108   5.730  4.6e-07 ***
tourism      0.002318   0.001335   1.736 0.088211 .
log(case)   -0.012000   0.003399  -3.531 0.000856 ***
unemploy     0.006432   0.002349   2.738 0.008349 **
country1    -0.048407   0.012117  -3.995 0.000197 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01951 on 54 degrees of freedom
Multiple R-squared:  0.3678,    Adjusted R-squared:  0.3209
F-statistic: 7.853 on 4 and 54 DF,  p-value: 4.594e-05
```

Figure 15: Summary for log model

Compared to the **base model**, it yields a higher R^2.

Hence, the **log model** is selected: `lm(y~(tourism + log(case) + unemploy + country))`

2.3 Diagnostic

We examine the diagnostic plots of the selected model: `lm(y~(tourism + log(case) + unemploy + country))`

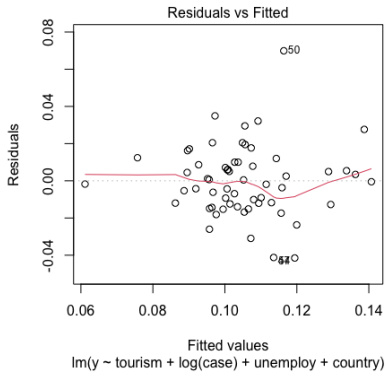

Figure 16: Residual vs $\hat{y}$

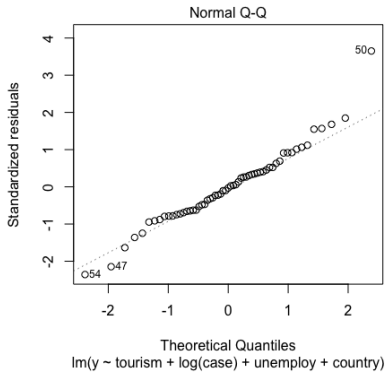

Figure 17: QQ-plot for the selected model

The residual plot shows the points are roughly randomly scattered around the line of $y = 0$. The errors are mostly normally distributed with no obvious patterns.

The Q-Q plot falls around the dashed line, an indication that the assumption of normality of residuals is not violated. However, there exist a few residual values in the upper right of the plot are slightly higher than the estimated quantile values.

We fit the residual plots of each individual variable to explore if polynomial terms are necessary to include:

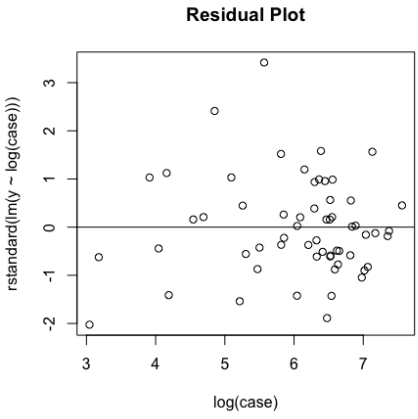

Figure 18: Residual vs log(cases)

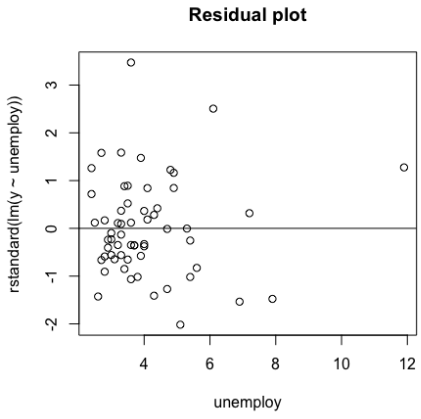

Figure 19: Residual vs unemployment

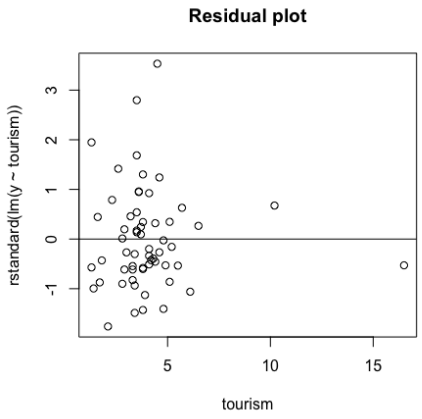

Figure 20: Residual vs tourism

We observe roughly symmetric distributions around $y = 0$ for all residual plots, and no obvious pattern is present. Thus, there is not enough evidence to include polynomial terms. However, we can see some outliers in x in certain graphs, so a leverage plot is created for further exploration:

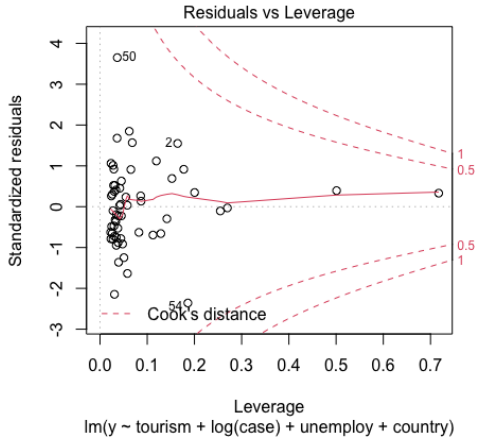

Figure 21: Leverage plot for the selected model

All of the points have low to moderate leverage, given all the points lie within the 0.5 dashed line. Diagnostic plots show no obvious fallacy in the chosen model.

3 Conclusion

Statistics, including Tourism, log(Cases), Unemployment and the categorical variable, Country, were significant in explaining the decline in the output of the economy, measured by GDP. In contrast, HDI, an indicator of the development of the region prior to the outbreak, showed no important relationship with the extent of the economic impact. Urban population percentage does not contribute much explanatory power when Cases are included.

States/Provinces that were suffering from inefficiency due to higher unemployment rates prior to the lock-downs were those who struggled from greater economic declines. States/Provinces with an economic structure that has high reliance on tourism also were more severely impacted. Moreover, states/provinces with less severe drop in economy are also the ones with more cases, which emphasizes the distinction between coping with the crisis from the public health standpoint and from the economic standpoint. Furthermore, the significance of the categorical variable, Country, acknowledges the differences in policies, economic structures, or even culture between these two countries, which can be further studied and discussed.

Finally, given the complex nature of this economic problem, a relatively low value of 0.36 for the R^2 is within our expectation. Nonetheless, we believe that including statistics that captures the broader economic fundamentals of a state/province may improve the R^2. For example, the percentage of GDP from the service sector can replace or use together with the percentage of GDP from Tourism. Other useful economic measurements may include median saving rate, an average measure for the digitization of companies and number of starting companies in 2019.

4 Appendix & Reference

4.1 Data Sources:

Change in GDP between Dec 31, 2019 to Jun 30, 2020:

U.S. Department of Commerce:

https://www.bea.gov/sites/default/files/2020-10/qgdpstate1020_0.pdf

Covid-19 cases until June 30, 2020:

CDC:

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

CTV News:

https://www.ctvnews.ca/health/coronavirus/tracking-every-case-of-covid-19-in-canada-1.4852102

The number of cases per 100,000 people:

New York Times: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

https://www.nytimes.com/interactive/2020/world/canada/canada-coronavirus-cases.html

Human Development Index:

Global Data Lab: https://globaldatalab.org/shdi/2018/indices/CAN+USA/?levels=4&interpolation=0&extrapolation=0&nearest_real=0

Unemployment rate:

Statista: https://www.statista.com/statistics/223675/state-unemployment-rate-in-the-us/

https://www.statista.com/statistics/442316/canada-unemployment-rate-by-provinces/

Tourism contribution:

SP Global: https://www.spglobal.com/ratings/en/research/articles/200427-tourism-dependent-u-s-states-cou

Statistics Canada: https://www150.statcan.gc.ca/n1/daily-quotidien/181010/t001b-eng.htm

Urban population Percentage:

"Urban Percentage of the Population for States, Historical" by *Iowa State University*:

https://www.icip.iastate.edu/tables/population/urban-pct-states

Statistic Canada: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710013501

4.2 Data as CSV

Data file: https://docs.google.com/spreadsheets/d/1am06wfZij0vhpKDgi1kZnng2LpcY1DfYtATPkm9VSqw/edit?usp=sharing