Increase in the usage of electronic resources after the closure of the libraries due to the COVID-19 pandemic

Liangjiayi Wang

2021-04-09

Abstract

This study investigated whether the closure of the libraries due to the COVID-19 pandemic has an impact on the usage of electronic resources. The university closed its physical libraries at the beginning of January 2020. The resources are collected from Counting Online Usage of Networked Electronic Resources (COUNTER) which records the use of licensed content by UofT affiliated users in specific periods. I will seperate the analysis into two parts: Preliminary Insights and Negative Binomial Model. Combined all findings, we can conclude that the online electronic resources have a sharp short-term growth in February 2020 and gradually fall back to the average level (i.e., approximately have 8.3 requests for each book on average) after March 2020. The total requests in 2020 is higher than the total requests in 2019. Overall, there is an increase in electronic resource usage after the libraries' closure due to the COVID-19 pandemic.

Contents

1 Introduction

The COVID-19 rapidly affects our daily life and slows down the global economy. Major cities moved into lock-down and closed most non-essential businesses and services for safety and health purpose. At the same time, UofT decided to close its library in Jan.2020, potentially changing students'/faculty members' information-seeking behavior. We expected most users would shift to remote learning through online resources. The purpose of this study is to examine whether the closure of the libraries due to the COVID 19 pandemic has an impact on the usage of electronic resources. In particular, we are going to investigate whether the resource usage or total requests trend had changed from Jan-Apr 2019 to Jan-Apr 2020.

The resources were collected from Counting Online Usage of Networked Electronic Resources (COUNTER), which records the use of licensed content by UofT affiliated users in specific periods. We use R to analyze the potential shifts of Electronic Content Usage under the COVID-19 pandemic. There are 587742 observations in this dataset with 33 variables. Define that the university closed its physical library in January 2020. To approach the project, I will investigate the change in the usage from January 2019 to April 2019 and from January 2020 to April 2020 through different statistical analysis methods.

For the rest of my analysis, first, I will describe the data cleaning procedures in the "Data Overview" section. Secondly I will state the research design and related statistical methods in the "Methods" section; Thirdly, I will interpret the relevant results and insights in "Results." Also, I will summarize the findings and address the potential challenges for future analysis in "Conclusion and Discussion." Lastly, some supplementary graphs and tables will be included in "Appendix".

(note: All information was provide by Klara Maidenberg, the Assessment Librarian at the University of Toronto Libraries.)

2 Data Overview

Vendor, Metric type and Reporting period are key variables I mainly used in analysis. There are 5 different vendors which are our online information provider and 3 different metric types: 'Unique_Title_Requests' & 'Unique_Item_Requests'& 'Total_Item_Requests'. The Reporting Period: Jan.2019 - Apr. 2019 & Jan.2020-Apr.2020 record the number of the monthly requests from each book.

Then, I will use the following data cleansing process to improve our data quality, increase overall productivity and accuracy.

For the first step, we only included the "Total_Item_Requests" metric type. We dropped other metric types (i.e.Unique_Item_Requests and Unique_Title_Requests) which occupy about 57% of origional data. The reason for only keeping "Total_Item_Requests" is because that the counts recoded in R4 closely matches total item requests.

The second step was to remove the missing (NA) data. The assessment librarian provided multiple Excel reports which are categorized by year/vendor/series. And we combined all the reports into our "combined.csv" file. In this process, a lot of "NA" values were created for Jan.2019, Jan.2020, Feb.2019, etc. Those "NA"s appear because we combined 2019 and 2020 reports without changing the column names. Thus, I can replace the "NA"s with 0's. In this way, I keep the essential data, and it will not influence the calculation of total requests numbers.

The last step was to create a subset. Since the original dataset has 33 variables, I only select the key variables for analysis and modelling. There have lots of books with the same title, but from different sources, so I recombine and match the book by their unique DOI, ISSN, ISBN.

3 Methods

The following analysis is conducted into two parts: Preliminary insights and Modelling.

3.1 Preliminary Insights

In preliminary insights, I will use the basic summary statistics, and line graph to have an overview of the dataset and analysis on how electronic usage changes by different vendors. And I will answer whether the average total requests have a huge difference between 2019 and 2020? Moreover, compared to 2019, whether each vendor's total requests have a different fluctuation trend in 2020?

3.2 Modeling

In the modeling part, I will choose the most appropriate statistical model to look for a relationship between variables, observe data patterns, draw conclusions, and ultimately answer the reseach questions. The count (or usage reports/total requests) is the outcome variable and indicates each book's monthly total requests. 'Month' is a categorical predictor variable with four levels: Jan, Feb, Mar, and Apr and represents the month of usage reports. 'Year' is also a categorical predictor variable with two levels indicating the year of usage reports, and it is coded as 2020 and 2019.

3.2.1 Model Selection (Poisson vs Negative Binomal)

The Poisson distribution is widely used for modelling the number of occurrences of an event occurs in an interval of time, distance, or volume. From the Poissoness graph (in appendix), we noticed that the point are followed by 45 degree line which indicate that the response variable followed poisson distribion. So Poisson regression could be one of our choices.

Meanwhile, negative binomial regression is a generalization of Poisson regression because it weakens the Poisson Model's restrictive assumption that the variance is equal to the mean. This inequality (mean $\neq$ variance) is captured by estimating the dispersion parameter. In other words, Negative binomial regression is also for modelling count variables, and it can be used for over-dispersed (variance $>$ mean) count data.

To decide use which model (poisson vs negative binomial), I fit a basic poisson model (no interaction included) and test whether there exists the over-dispersion in our data. From the over-dispersion test result in appendix, we clearly see that there is evidence of overdispersion. Thus, we will use the negative binomial regression in following analysis.

Furthermore, we have two type of negative binomial regression: with interaction and without interaction term. The explanatory variables: year and month may have interaction effects with each other, so we need use the model with interaction term to test the effects. To compare the model with interaction and without interaction, we will consider finding the model with the lowest value of the Akaike Information Criterion (AIC). Based on the AIC test result in appendix, we will select the Negative binomial model with interaction.

For this model, Year, month and $month_i$: $year2019$(interaction term) are predictor variables with March and year 2020 as reference level.

$$Y \sim NegativeBinomial(r, p)$$

$$log(p) = \beta_o + \beta_1 x_{Jan} + \beta_2 x_{Feb} + \beta_3 x_{Apr} + \beta_4 x_{year2019} + \beta_{5i} x_{month_i:year2019}$$

4 Result

4.1 Preliminary Insights

Firstly, we use the summary statistics to provide an overview of total requests for each year. From Table 1, We find that the gap between max and min is enormous in both years. The average total requests(mean) in 2020 (i.e. 36) is much larger compared to 2019 (i.e. 27). Based on this observation, we need to inspect the

reason that causes this difference. For instance, whether we have a dramatic increase in usages in a specific month of 2020 or there is an overall growth trend in 2020.

Table 1: Summary Table of Total Requests by Year

YEAR	mean	min	q1	med	q3	max
2019	27.56307	0	1	1	4	106891
2020	36.00842	0	1	2	5	108815

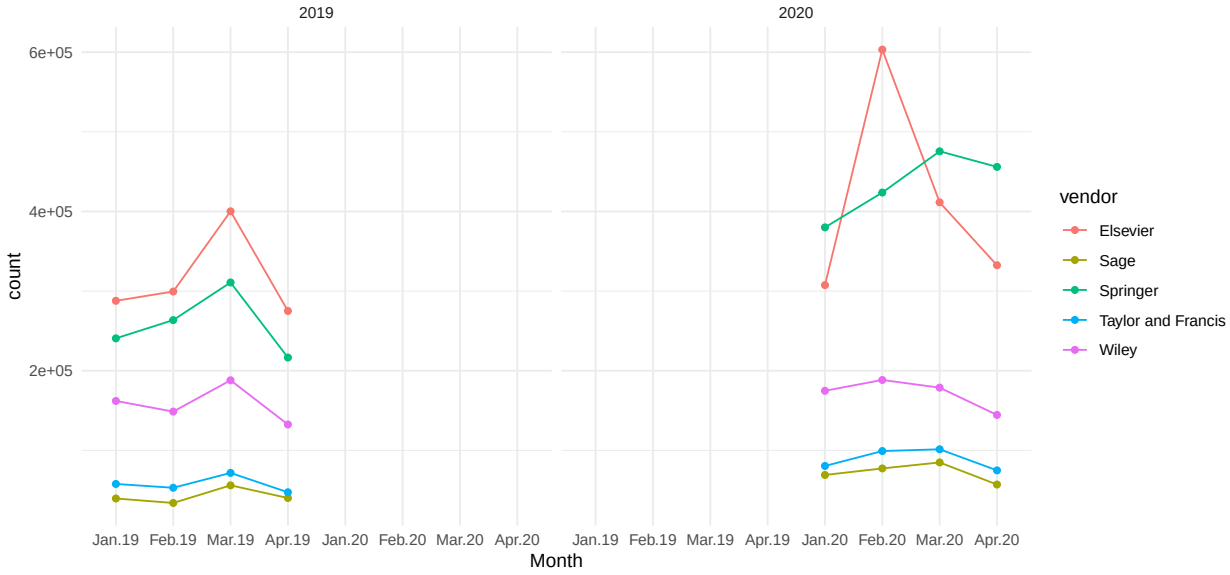
Next, we look at changes on the usage of electronic resources based on vendors. The Figure 1 indicates the total requests from each vendor by month.

From 2019 to 2020, Elsevier and Springer had the most outstanding overall total requests; Meanwhile, Taylor and Francis, and Sage had the lowest requests. Compared with the same period last year, no matter which vendor, the overall total requests from Jan.20-Apr.2020 were much higher than in 2019.

In 2019, the request for electronic resources reached a peak in March for each vendor. Comparatively, in 2020, the requestss reached a peak in February. The growth in requests for Wiley, Taylor and Francis, and Sage were slightly small in 2020. By contrast, the requests fluctuation from Elsevier and Springer were much significant. The Elsevier's total requests had a sharp increase from Jan.2020 to Feb.2020. After that, it decreased dramatically until April. From Jan.2020 to Apr.2020, Springer's total requests continuously climbed to a remarkable amount.

From vendors' perspective, the general usage of electronic resources from students/staff has significant growth in 2020 compared to 2019, especially in 2020 February.

Figure 1: Total Requests from each vendor by month



4.2 Modeling

We will set the requests for March 2020 as reference level. After we fit the negative binomial model with the interaction term, the transformed coefficient estimation results in Table 2 indicate that each preditor's confidence intervals do not include 0. Thus, all predictors are statistically significant at 5% level. Continue to focus on the results in table 2, after the mathematical calculations, we are able to interpret the usage change in a numerical way.

In March 2020, each book has about 10.22 requests (with range from 9.99 to 10.47) on average. Compared with March 2020, in January 2020 and April 2020, each book's usage reports approximately decreased by (0.80-1)*100% = 20% (with range from 16.5% to 21.8%) and 15% (with range from 12.1% to 17.7%), seperately. However, in Feburary 2020, each book's usage reports increased by 11% (with range from 7.6% to 14.9%) compared to March 2020 on average.

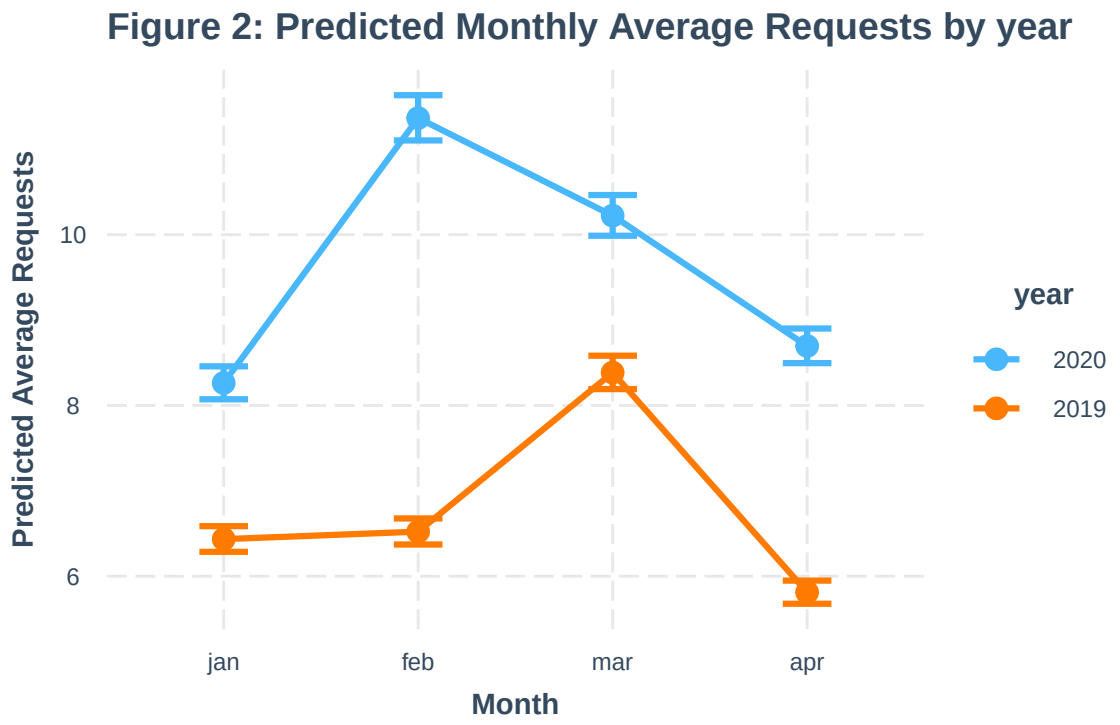
Differently, the usage reports in March 2019 are around 18% (with range from 15.2% to 20.6%) less compared to March 2020 on average. Also, from the interaction term, we noticed that each book's usage reports in Janurary 2019 are (0.949*0.820-1)x100%=22% less than January 2020 on average. In the same way, on average, the each book's usage reports in Feburary 2019 are 42% less than Feburary 2020 and April 2019 are 33% less than April 2020.

Table 2: Transformed coefficient of usage reports

		2.5 %	97.5 %
(Intercept)	10.223	9.988	10.465
monthjan	0.808	0.782	0.835
monthfeb	1.112	1.076	1.149
monthapr	0.851	0.823	0.879
year2019	0.820	0.794	0.848
monthjan:year2019	0.949	0.906	0.995
monthfeb:year2019	0.700	0.668	0.733
monthapr:year2019	0.815	0.778	0.854

Moreover, we transform the numerical results into a visualized graph in Figure 2. Figure 2 also indicates that the general monthly usage trend is uniform from 2019 to 2020; meanwhile, the average monthly usage in 2020 is more than the average monthly usage in 2019.

It is worth noting that after UofT close its offline library in January 2020, there is a transitory boost in usages from January 2020 to Feb 2021 and then fall into certain average level (i.e. around 8.3 requests for each book on average). This finding is consistent to our model numerical analysis and preliminary insights.



5 Conclusion/Discussion

In the analysis part, we research how vendors influence the total requests by year and month; moreover, we select the Negative binomial regression model and analyze the factors that influence the usage report. The outcome variable is each book's usage report in each month, and the predictive variable is the year, month, and the interaction term with year and month. From the vendors' perspective, the 'Elsevier' and 'Taylor and Francis' were the top two vendors that had the most number of usage in both 2020 and 2019. Also, the overall total usage is higher in 2020 than in 2019 and 'Elsevier' has a more significant increase in usage in February compared to other vendors. From our model's result, we find that the usage trend is similar in 2020 and 2019. However, there is a significant boost in February 2020. On average, the online resource's usage in February 2020 and March 2020 is higher than the usage reports in April 2020 and January 2020.

Combined the preceding results, after the library closed in January 2020, we can conclude that the online electronic resources had a sharp short-term growth in February 2020 and gradually fall back to the average level (i.e., approximately have 8.3 requests for each book on average) after March 2020. The increase in February is mainly contributed by the use of Elsevier's resources. Overall, there is an increase in the usage of electronic resources after the libraries' closure due to the COVID-19 pandemic.

Due to our dataset's limited information, we cannot identify the causal effect between the fluctuation of usage reports and library closure since there also have lots of unobservable factors. For instance, the usage may be affected by students' movement from one country to another. We need to use a more complicated model like two-stage least squares to eliminate the influence of potential and unobserved characteristics.

Moreover, because of the limitation of the computer's operational capability and the large dataset, we cannot match each book's subject area by using API. If we have the information for the subject area, we can put it into the model and find the usage changes in specific subject areas. Meanwhile, we may use the DOI and ISSN to identify the type of electronic resources, for example, journals, periodicals, newspapers, annuals, and non-textual resources (e.g., image, audio, video, etc.). After recombining the data, we can research what kind of electronic resources have the most significant usage change. If possible, we should limit the raw data

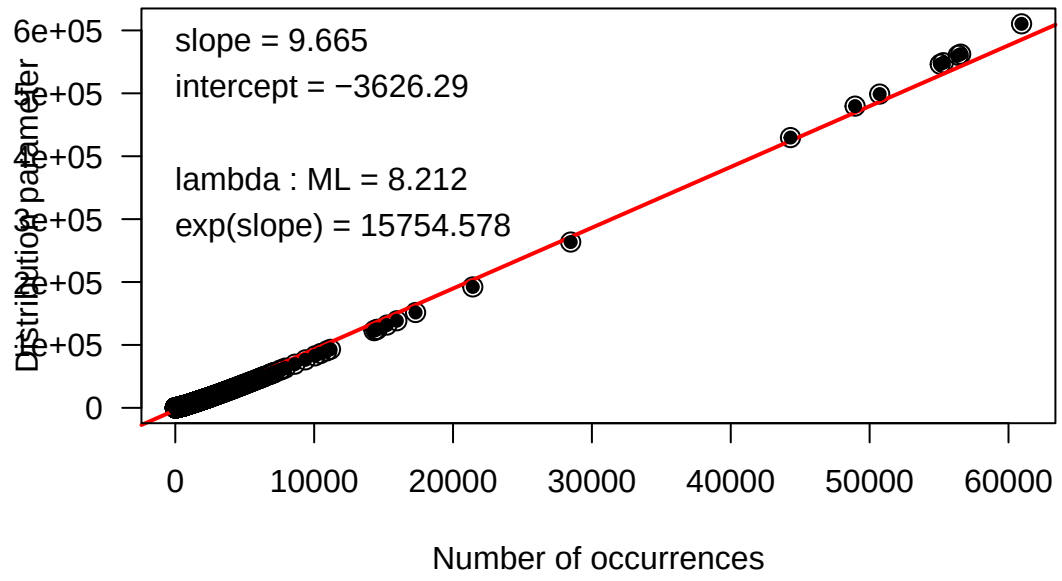
to the long-term electronic users for eliminating the potential bias due to the change of electronic resources users. With the control of the other possible influences, the model results must be more accurate, and we can interpret the estimation of predictors as the effect of library closure.

For further research, we can look at whether the amount of time students spend using electronic resources has changed; Students/staff prefer to use these resources during the day or at night? We may create a survey to collect the information that we need and do additional analysis.

6 Appendix

1. Poissoness plot

Figure 3: Poissoness plot



2. The dispersion test

```
##
##  Overdispersion test
##
## data:  model
## z = 3.6776, p-value = 0.0001177
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 3910.635
```

3. Model comparison by using AIC (with interaction and without interaction)

Model-1:Negative binomial model- Without interaction

In this model, Year and month is predictor variables with Mar and year2020 as reference level.

$$Y \sim NegativeBinomial(r, p)$$

$$log(p) = \beta_o + \beta_1 x_{Jan} + \beta_2 x_{Feb} + \beta_3 x_{Apr} + \beta_4 x_{year2019}$$

Model-2:Negative binomial model- With interaction

In this model, Year, month and $month_i : year2019$(interaction term) are predictor variables with March and year 2020 as reference level.

$$Y \sim NegativeBinomial(r, p)$$

$$log(p) = \beta_o + \beta_1 x_{Jan} + \beta_2 x_{Feb} + \beta_3 x_{Apr} + \beta_4 x_{year2019} + \beta_{5i} x_{month_i:year2019}$$

Compare the AIC for both models, AIC for model 1 is 2931264 and for model 2 is 2930995. Since 2930995 $<$ 2931264, we can conclude that model 2 (with interaction tem) is better.

Table 3: AIC for model 1

x
2931264

Table 4: AIC for model 2

x
2930995

4. Selected Model Result

```
##
## Call:  glm.nb(formula = count ~ month + year + month:year, data = df_clean,
##     link = "log", init.theta = 0.05802715618)
##
## Coefficients:
##       (Intercept)           monthjan           monthfeb           monthapr
##           2.32464           -0.21289            0.10596           -0.16181
##          year2019  monthjan:year2019  monthfeb:year2019  monthapr:year2019
##          -0.19823           -0.05192           -0.35717           -0.20482
##
## Degrees of Freedom: 979711 Total (i.e. Null);  979704 Residual
## Null Deviance:       513400
## Residual Deviance: 510700    AIC: 2931000
```