The COVID-19 pandemic result in negative changes to usage of electronic resources

Keqi Yu #1004244150
Collaborator : Klara Maidenberg
TA : Steven Campbell

04/03/2021

Abstract

This study investigated whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area. The data is provided from University of Toronto's Libraries Assessment Librarian Klara Maidenberg. The usage data is obtained from Counting Online Usage of Networked Electronic Resources (COUNTER) usage report, and it is the number of usages of a scholarly journal or e-book in each month. The subject data is obtained using API key from Elsevier Developers. We will use the negative binomial model to investigate the research question. The outcome showed that there was a negative impact on the usages of e-resources in the school due to the closure of the libraries on the COVID-19 pandemic. Also, especially the subject Physical Science and Engineering has the biggest effect compared to other three subjects: Health Science, Life Science, and Social Science and Humanities. In addition, the usage of year 2020 is higher than year 2019 in the same period Jan-Apr.

Introduction

Currently, many industries are experiencing the COVID-19 pandemic and some of them have to be closed due to safety issues, for example, libraries. In recent decades, scholarly materials have increasingly moved to an electronic format, and are accessible online to subscribers. Before the investigations, we may think that since the libraries are closed, more people will access the e-resources, and more professionals will start to research on more health related projects leading to increased usage in the subject Health Science because of the current Covid-19 pandemic.

In this case, I am going to analyze the impact of usages of e-resources at the University of Toronto libraries to see whether my hypotheses satisfied or not due to the COVID-19 pandemic. The data I used for analysis is from the period Jan-Apr,2019 and Jan-Apr,2020, which I consider that March and April in 2020 as the month of Covid, since the school was closed on March,2020. The reason that I am using the monthly data instead of the total item requests for each journal is that I want to see the effect of the specific month. For the subject level, due to the fact that since they subdivided very detail, there are too much subject levels which are more than 300 different subjects, we decide to use the subject abbreviation which has 27 different segments. According to these 27 levels, we divide them into four big subject levels named Top Level, including Physical Sciences and Engineering, Health Sciences, Life Sciences, and Social Sciences and Humanities.

The purpose of this study is to find out whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area. In particular, I am going to investigate whether there is an increase or a decrease on the UofT libraries' electronic resources due to the Covid-19. After that, I will continue to investigate if specific subject areas are affected.

For the rest of my report, I am going to describe the usage data, the subject data, my procedure of data cleaning in the Date Summary section, the design of the scientific study and the statistical method I used to analyze the effect in the Methods section, and the result of my investigations in the Results section. In addition, summary of what was learned from Result section regarding the research question will be included in the Conclusion and Discussion section, as well some possible limitations will be discussed in that section. In the end, relevant figures and model outputs will be included in the Appendices section.

Data summary

We will use the COUNTER usage reports that show use of licensed content by UofT affiliated users. We are going to use the monthly data for each journal for statistical analysis. Also, we will get Elsevier Subject Classifications using API key from Elsevier Developers.

In order to make the data more suitable for analysis, the following data cleaning procedures were made.

1. After using API key from Elsevier Developers to conduct the subject level, I stored the subject level in a separate data set for each "unique" journal after querying by ISSN, since the process is too timer-consuming to conduct every time.

2. Due to the fact that there are too much subject levels which are more than 300 different subjects, we decide to use the subject abbreviation which has 27 different segments. It is because the database collects detailed subject level information and we had to simplify the subject level in order to build a reasonable model, since when you detailed too many subjects, you may not be able to detect the difference. According to these 27 levels, we divide them into four big subject levels named Top Level, including Physical Sciences and Engineering, Health Sciences, Life Sciences, and Social Sciences and Humanities.

3. Since there are a lot of missing data in 2017 and 2018, I only keep the data of 2019 and 2020. Since only the data of Jan-Apr of 2020 is available, in order to avoid the academic annual trends, we only keep the data of Jan-Apr for 2019 as well.

4. In order to distinguish Covid or not in our dataset, I add a new variable named covid, which has two levels including 0 (Not Covid) and 1 (Covid).

5. In order to match the subject level to each journal, I use the identifier print_ISSN, which is a unique code to identify journals.

6. After the data cleaning, I stored the final data in a usage dataset.

The overall usage data plots are provided below in Figure 1 and Figure 2, which is separated by year 2019 and 2020. In addition, the overall subject data plots are provided below in Figure 3.

Figure 1: Monthly Usage by Month Jan−Apr for 2019

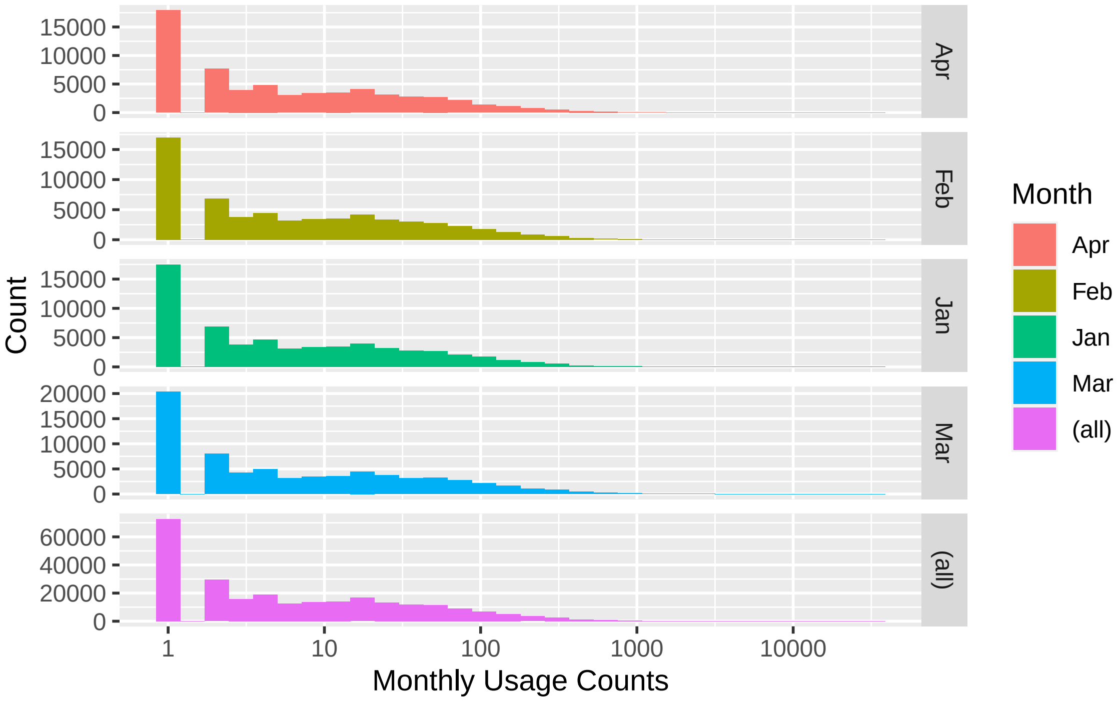
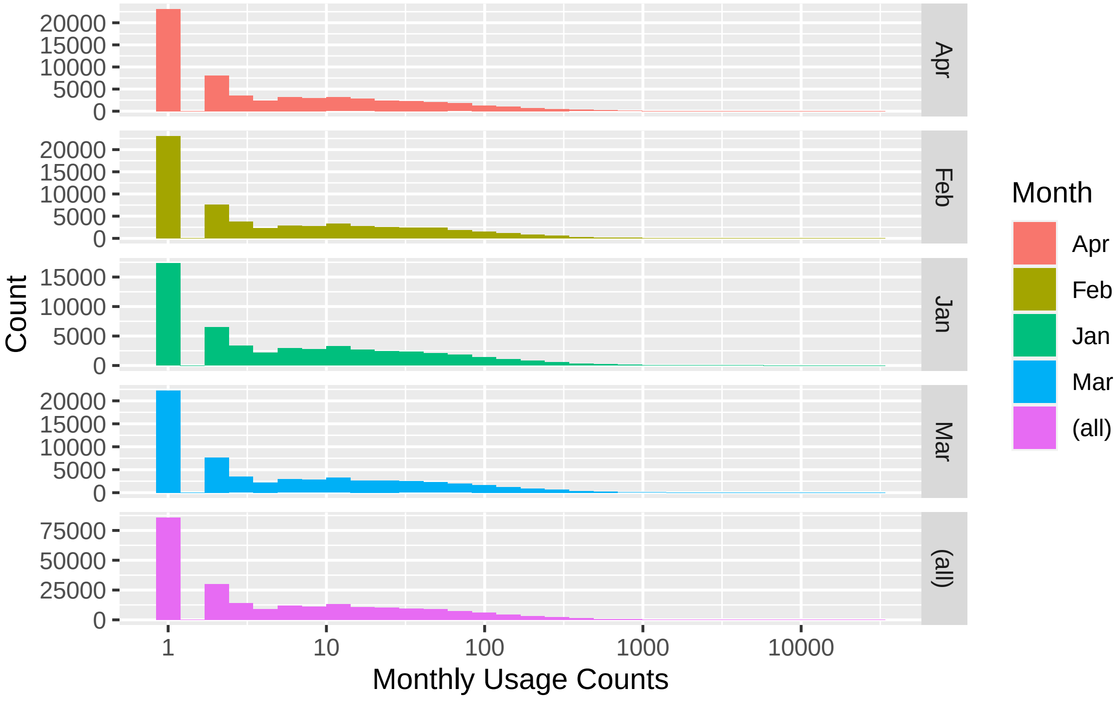
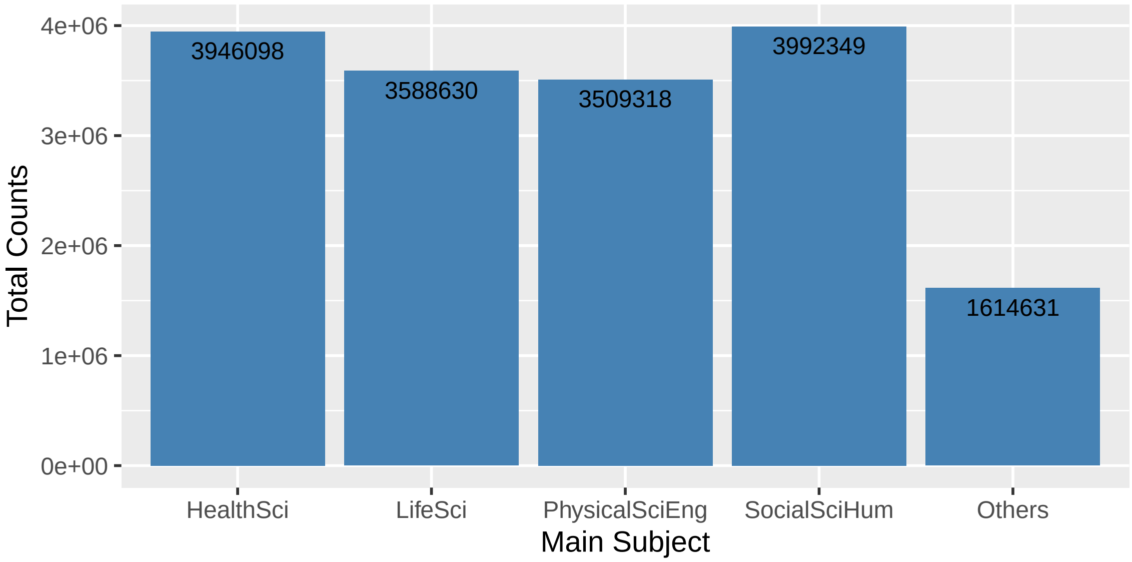


Figure 2: Monthly Usage by Month Jan−Apr for 2020



As we can observed from Figure 1 and Figure 2, the overall pattern for these two years are the same. Since the plot has been log transformed (normalized), there are spikes at value 1 for every year, which means that there are many zeroes in the dataset.

Figure 3: Barplot of total counts by main subject areas



HealthSci: Health Science,
LifeSci: Life Science
PhysicalSciEng: Physical Science and Engineering
SocialSciHum: Social Sciences and Humanities

From Figure 3, every subject seems have fairly similar numbers of usages over two years.

Methods

At first, there are a few analysis methods we might consider: OLS Regression, Ordinary Count Models (Poisson or Negative Binomial) and Zero-inflated Regressions (Poisson or Negative Binomial).

OLS(Ordinary least squares) regression is more commonly named linear regression, which estimates the relationship between one or more independent variables and a dependent variable. This can be writed as

$$Y = \beta_0 + \sum_{j=1 \ldots p} \beta_j X_j + \varepsilon$$

Ordinary Count Models(Poisson or Negative Binomial)

1. Poisson

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The Poisson distribution is used to model the number of events occurring within a given time interval.

2. Negative Binomial

$$f(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

The negative binomial is used to model the number of events with over-dispersion, which means that the data is dispersed and variance of the data is large.

Last but not least, Zero-inflated Regressions attempt to account for excess zeros.

However, since my data are monthly counts data which are highly non-normal (you can see from the previous section Data Summary plots), there are two models we used for the count data, which is Poisson and Negative Binomial model. The difference between them is that negative binomial has an extra parameter to model the over-dispersion. Since my count data are highly non-normal and over-dispersed (you can see from the previous section Data Summary plots), so we will use negative binomial model.

From the previous section Data Summary plots, there may be excess zeros, Zero-inflated Negative Binomial Regression might be more appropriate.

- Model 1 (Zero-inflated Negative Binomial Regression):

At first, we try to use the Zero-Inflated Negative Binomial Regression(ZINB) to fit the data, since it is for modeling count variables with excessive zeros and it is usually for over-dispersed count variables.

The purpose of ZINB is to test whether this model can explained count variable(monthly counts) with over-dispersed and excessive zeros, and what variables have impact to the count variable(monthly counts)

The probability distribution of the ZINB random variable y_i can be written

$$Pr(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)g(y_i = 0), if j = 0 \\ (1-\pi_i)g(y_i), if j > 0 \end{cases}$$

Then $g(y_i)$ is the negative binomial distribution given by

$$g(y_i) = Pr(Y = y_i \mid \mu, \alpha) = \frac{\gamma(y_i + \alpha^{-1})}{\gamma(\alpha^{-1})\gamma(y_i + 1)}(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i}$$

Then,

$$\mu_i = exp(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$

In order to answer the research question whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area, we will use Top.Level(main subject levels), covid(covid or not), Month(Jan.-Apr.), Year(2019,2020), and vendor to model the monthly count in the part of negative binomial model and the logit part of the model.

A zero-inflated model assumes that zero outcome is due to two different processes. In this case, the two processes are that a journal has been used vs. never been used. If not been used, the only possible outcome is zero. If been used, it is then a count process. The expected count is expressed as a combination of the two processes:

$$E(n_{usage} = k) = P(notbeenused) \times 0 + P(beenused) \times E(y = k \mid beenused)$$

Some variables are very highly correlated. So, we would use Chi-Squared test of independence to check if two categorical variables are independent.

The result shows that we can reject the null hypothesis and conclude that the variables Top.Level, covid and Year are, indeed, independent. So, we are going to use these two variables Top.Level (main subject levels), covid (covid or not) and Year (2019, 2020) to model the monthly count in the zero-inflated negative binomial model.

- Model 2 (Negative Binomial Model):

Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables, that is when the conditional variance exceeds the conditional mean.

Negative binomial regression can be written:

$$f(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

In order to answer the research question whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area, we will use Top.Level(main subject levels), covid(covid or not), Month(Jan.-Apr.), Year(2019, 2020), vendor and the interaction term of Top.Level(main subject levels) and covid(covid or not) to model the monthly count in the negative binomial model.

The log of the expected outcome is predicted with a linear combination of the predictors:

$$f(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

Therefore,

$$\widehat{monthly_counts_i} = e^{Intercept+b_1Month_i+b_2I(Year=2020)+b_3Top.Level_i+b_4vendor_i+b_5I(covid=1)+b_6Top.Level_i*I(covid=1)}$$

Negative Binomial Regression Assumptions :

1. Linearity in model parameters
2. Independence of individual observations
3. Multiplicative effects of independent variables
4. The conditional variance of the outcome variable exceeds the conditional mean

As we mentioned earlier, negative binomial models assume the conditional variance exceeds the conditional mean. the conditional variance of the outcome variable is 20987.47, which exceeds the conditional mean 12.8008, which satisfied the assumption of negative binomial model.

However, the data fails to satisfy the assumption of independence, because some data may come from the same book, however, it is still a good model to use since the other three assumptions satisfied, and basic conditions of the model conform to the data with over-dispersion.

Results

In order to compare simple Negative Binomial Regression with Zero-Inflated Negative Binomial Regression, we will use AIC and BIC, which a lower AIC score and will be the better-fit model and a lower BIC means that a model is considered to be more likely to be the true model. Using both AIC and BIC, the simple Negative Binomial Regression would fit better than the Zero-Inflated Regression. (AIC and BIC for negative binomial model are 3343178 and 3343368, then zero-inflated are 3417745, 3417890.) This indicates that Negative Binomial Regression could explain the excessive zeros well.

Therefore, we choose Negative Binomial Regression as our final model.

The summary table (see Appendix 1) shows that he predictors Month, Year, Top.Level, vendor, covid and the interaction term Top.Level and covid except Social Sciences and Humanities and covid in the negative binomial regression model can predict the number of monthly usage (count) significantly with p-value lower than 0.05.

From Table 1, the expected log count for Year 2020 is 0.47 higher than the expected log count for Year 2019, so the incident rate for Year 2020 is 1.61 times the incident rate for Year 2019. As a result, Year 2020 has higher journal use than Year 2019.

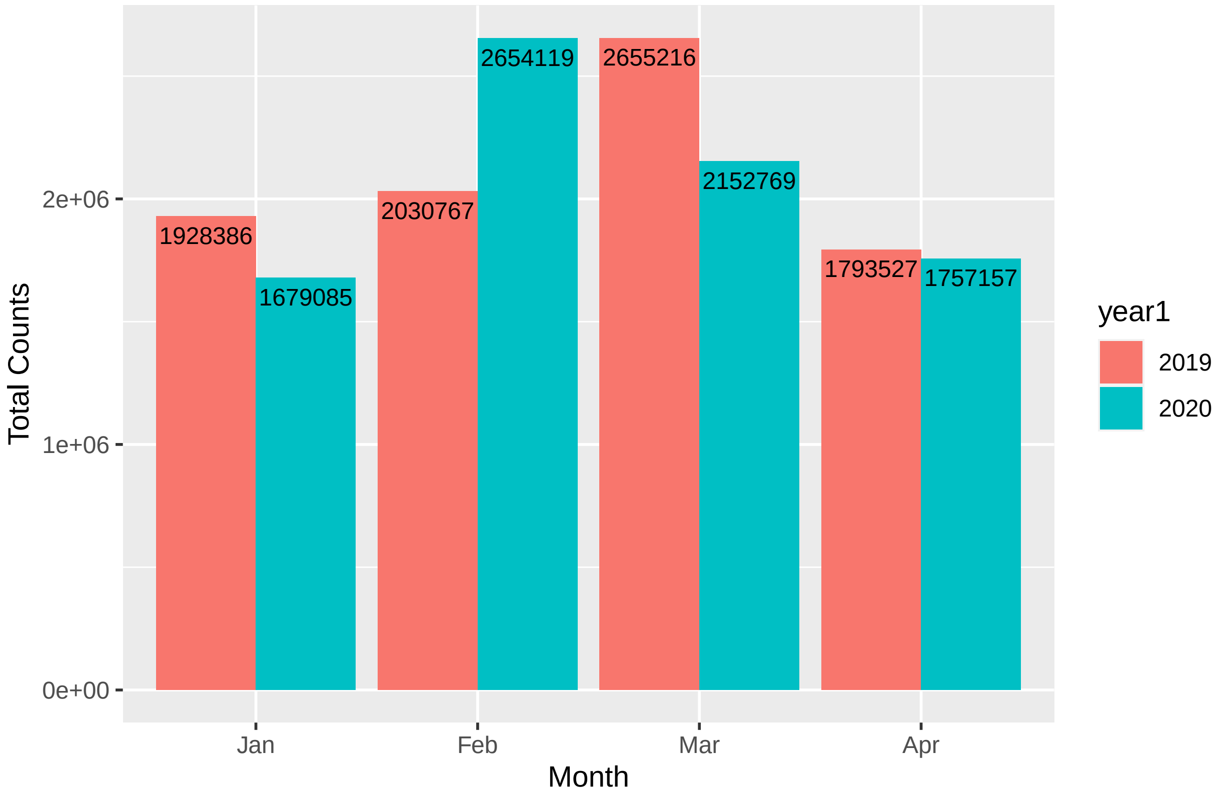
The period of Covid (Covid = 1) has an expected log count of 0.19 lower than that of non-Covid (covid = 0) holding other variables constant, which indicates that the period of Covid has 0.91 times the incident rate for non-Covid. In other words, COVID-19 pandemic has negative effect on journal use.

Table 1: Confindence Interval

	Estimate	ExpEstimate	2.5 %	97.5 %
(Intercept)	4.21	67.23	65.65	68.86
MonthFeb	0.16	1.18	1.15	1.20
MonthJan	-0.06	0.94	0.92	0.96
MonthMar	0.29	1.33	1.31	1.36
Year2020	0.47	1.61	1.58	1.64
Top.LevelLife Sciences	-0.26	0.77	0.75	0.79
Top.LevelPhysical Sciences and Engineering	-1.21	0.30	0.29	0.30
Top.LevelSocial Sciences and Humanities	-0.19	0.83	0.81	0.85
vendorSage	-0.31	0.73	0.71	0.76
vendorSpringer	-1.79	0.17	0.16	0.17
vendorTaylor and Francis	-1.15	0.32	0.31	0.32
vendorWiley	0.28	1.32	1.29	1.35
covid1	-0.19	0.83	0.80	0.86
Top.LevelLife Sciences:covid1	-0.09	0.91	0.87	0.95
Top.LevelPhysical Sciences and Engineering:covid1	0.21	1.24	1.19	1.29
Top.LevelSocial Sciences and Humanities:covid1	0.00	1.00	0.95	1.05

In addition, this result also coincides the bar plot of total counts by Month and Year. We can see from the Figure 4, there is a great increase in the month February of non-Covid, but a great drop for the month March of Covid period, then the changes become much flatter in April. The increase of February still overlap the decrease of March, which result in overall increase in Year 2020, but negative changes in Covid period.

Figure 4: Barplot of total counts by Month and Year



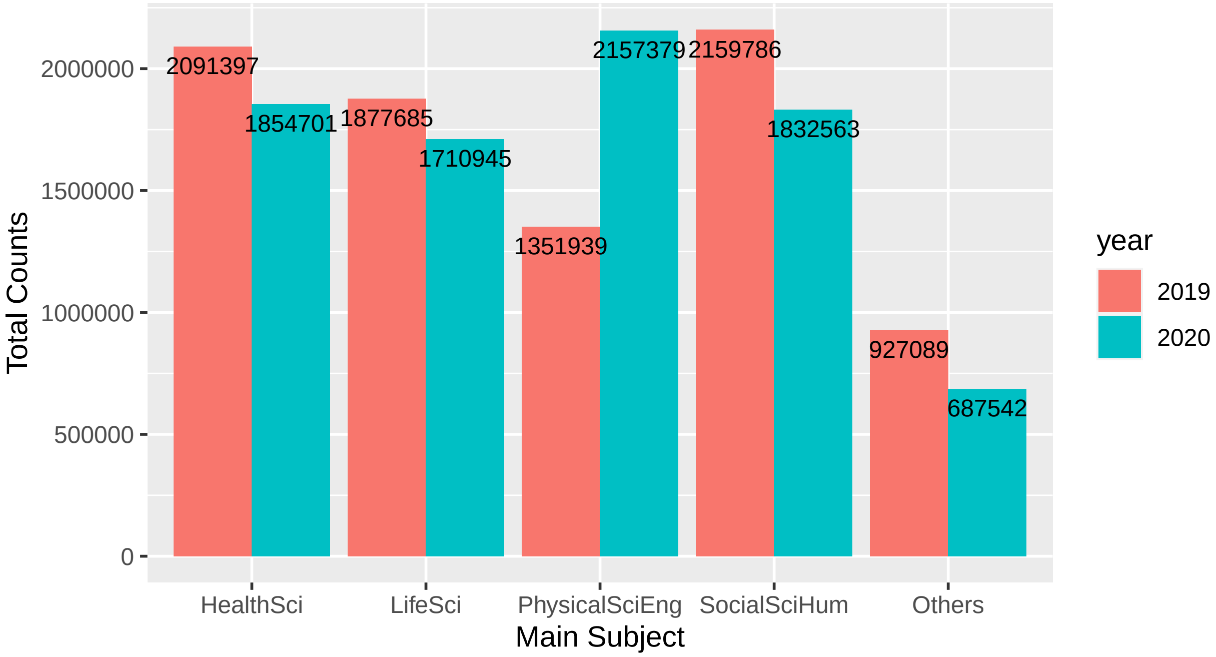
The interaction of covid and subject are significant which means that covid has significant effect in changing counts of subjects, especially the subject Physical Science and Engineering has the highest estimate, which may indicate that year would affect most in the subject Physical Science and Engineering.

This result also coincides the bar plot of total counts by main subject areas and Year. We can see from Figure 5, there is a large increase in the subject Physical Science and Engineering, but a decrease is experienced for other three subject.

Furthermore, the main subject Physical Science and Engineering includes many detailed subjects, for example, Chemical Engineering, Computer Science, Chemistry, Physical and Theoretical Chemistry, Materials Chemistry, Geology, Energy, Biomedical Engineering, Mathematics, and so on.

Therefore, Covid-19 has resulted in increase usages in these subject areas, which actually contradicts the hypotheses we made in the introduction that health science may be impacted most.

Figure 5: Barplot of total counts by Subject areas and Year



HealthSci: Health Science,
LifeSci: Life Science
PhysicalSciEng: Physical Science and Engineering
SocialSciHum: Social Sciences and Humanities

In short, overall year 2020 has a higher count of usages of e-resources than year 2020, because the increase in Feb 2021 covers the decrease in Mar 2021. The closure of the libraries due to the COVID-19 pandemic result in changes negatively to usage of electronic resources due to a dramatic drop in Mar 2021. However, especially the subject of Physical Science and Engineering, the usages experience a great increase compared to other three subject, nearly 60% increase.

Conclusion and Discussion

- Conclusion

The purpose of this study is to find out whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area. Through employing both the scientific model Negative Binomial Model and bar plots, overall, there is evidence to suggest that the closure of the libraries due to the COVID-19 pandemic reduces the usage of electronic resources. To answer the research question, the closure of the libraries due to the COVID-19 pandemic appears to result in changes to usage of electronic resources in specific subject area, which is Physical Science and Engineering. The number of usage increases dramatically.

- Limitations

There are many limitations in our dataset:

1. There are excessive zeros, approximately 60% of the data and some very large data in the dataset. Zero-inflated Negative Binomial can explain excessive zeros, but the variables Top.Level (main subject levels) and covid (covid or not) are highly correlated, so we may not be able to see the relationship between Top.Level and covid. Negative Binomial can explore the relationship, but it may not explain explain excessive zeros. So, these two models still have limitations.

2. Since our dataset only contains the first 4 months of 2020, the conclusion we made are only based on these eight months, which is very limited. It would be more appropriate to use additional months of data to represent the COVID-19 period. There would be interest in exploring if there were additional effects stemming from the start of the pandemic as people became more concerned. More data is needed for better study of the effect of pandemic.
3. We didn't consider the academics annual trends in our analysis. There is a finding suggesting high engagement particularly after inductions and at submission deadlines, and less usage in vacation periods. We should de-trend academics effects from our dataset.
4. There are many missing values in our original data, as well the subject levels.
5. Our model did not fulfill all the assumptions of negative binomial regressions, which I mentioned in the Method session.

Next, if we have more data from later months of 2021, we will do the further investigation by the methods we mentioned NB and ZINB model. Moreover, we could try Univariate Time Series analysis approach. Transforming the data into time series data frame and summing up the counts across all journals in a subject area and conducting time series analysis on the aggregate count. However, this method still has limitations, because our data is multivariate. Therefore, we could try Multivariate Time Series analysis from acp package, which can deal with discreteness, over-dispersion and cross correlation.

Appendices

Appendix 1: Summary table of the Negative Binomial Model

```
##
## Call:
## MASS::glm.nb(formula = monthly_counts ~ Month + Year + Top.Level +
##     vendor + covid + Top.Level * covid, data = usage, init.theta = 0.1956054439,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6254  -1.1491  -0.7826  -0.2607  17.0340
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       4.208e+00  1.153e-02
## MonthFeb                                          1.622e-01  1.019e-02
## MonthJan                                         -6.457e-02  1.020e-02
## MonthMar                                          2.868e-01  8.778e-03
## Year2020                                          4.740e-01  8.956e-03
## Top.LevelLife Sciences                           -2.624e-01  1.125e-02
## Top.LevelPhysical Sciences and Engineering       -1.210e+00  1.039e-02
## Top.LevelSocial Sciences and Humanities          -1.907e-01  1.174e-02
## vendorSage                                       -3.088e-01  1.788e-02
## vendorSpringer                                   -1.791e+00  7.847e-03
## vendorTaylor and Francis                         -1.152e+00  1.194e-02
## vendorWiley                                       2.757e-01  1.239e-02
## covid1                                           -1.875e-01  2.062e-02
## Top.LevelLife Sciences:covid1                    -9.332e-02  2.409e-02
## Top.LevelPhysical Sciences and Engineering:covid1  2.150e-01  2.192e-02
## Top.LevelSocial Sciences and Humanities:covid1     1.183e-05  2.493e-02
##                                                   z value Pr(>|z|)
## (Intercept)                                       364.990  < 2e-16 ***
## MonthFeb                                           15.921  < 2e-16 ***
## MonthJan                                           -6.331 2.44e-10 ***
## MonthMar                                           32.678  < 2e-16 ***
## Year2020                                           52.929  < 2e-16 ***
## Top.LevelLife Sciences                            -23.311  < 2e-16 ***
## Top.LevelPhysical Sciences and Engineering       -116.435  < 2e-16 ***
## Top.LevelSocial Sciences and Humanities           -16.253  < 2e-16 ***
## vendorSage                                        -17.274  < 2e-16 ***
## vendorSpringer                                   -228.240  < 2e-16 ***
## vendorTaylor and Francis                          -96.508  < 2e-16 ***
## vendorWiley                                        22.249  < 2e-16 ***
## covid1                                             -9.096  < 2e-16 ***
## Top.LevelLife Sciences:covid1                      -3.873 0.000107 ***
## Top.LevelPhysical Sciences and Engineering:covid1    9.807  < 2e-16 ***
## Top.LevelSocial Sciences and Humanities:covid1      0.000 0.999622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1956) family taken to be 1)
##
##     Null deviance: 641859  on 540727  degrees of freedom
## Residual deviance: 540619  on 540712  degrees of freedom
##   (781372 observations deleted due to missingness)
## AIC: 3344492
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  0.195605
##          Std. Err.:  0.000412
##
##  2 x log-likelihood:  -3344457.735000
```

Appendix 2: Summary table of the Zero-Inflated Negative Binomial Model

```
##
## Call:
## zeroinfl(formula = monthly_counts ~ Top.Level + covid + Year, data = usage,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
##  -0.4106  -0.4084  -0.3795  -0.2121 528.1341
##
## Count model coefficients (negbin with log link):
##                                             Estimate Std. Error z value
## (Intercept)                                 3.538480   0.008339  424.33
## Top.LevelLife Sciences                     -0.352010   0.010633  -33.10
## Top.LevelPhysical Sciences and Engineering -1.014024   0.009711 -104.42
## Top.LevelSocial Sciences and Humanities    -0.162459   0.010809  -15.03
## covid1                                     -0.171082   0.010778  -15.87
## Year2020                                    0.591496   0.008821   67.05
## Log(theta)                                 -1.777458   0.002194 -810.02
##                                            Pr(>|z|)
## (Intercept)                                  <2e-16 ***
## Top.LevelLife Sciences                       <2e-16 ***
## Top.LevelPhysical Sciences and Engineering   <2e-16 ***
## Top.LevelSocial Sciences and Humanities      <2e-16 ***
## covid1                                       <2e-16 ***
## Year2020                                     <2e-16 ***
## Log(theta)                                   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -24.2492    19.3207  -1.255   0.2094
## Top.LevelLife Sciences                       2.4148    21.6906   0.111   0.9114
## Top.LevelPhysical Sciences and Engineering  12.4681    19.3965   0.643   0.5204
## Top.LevelSocial Sciences and Humanities     -4.6221   152.4679  -0.030   0.9758
## covid1                                      -0.2932     0.1501  -1.954   0.0507
## Year2020                                     8.8737         NA      NA       NA
##
## (Intercept)
## Top.LevelLife Sciences
## Top.LevelPhysical Sciences and Engineering
## Top.LevelSocial Sciences and Humanities
## covid1                                     .
## Year2020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.1691
## Number of iterations in BFGS optimization: 46
## Log-likelihood: -1.709e+06 on 13 Df
```