Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 132
DOI: 10.11159/icsta24.132

Statistical Analysis on Factors Affecting Journal's Impact Factors

Forest Ho-Chen
[1]University of Pennsylvania
GUTM-0217, 3820 Locust Walk, Philadelphia, PA 19104-6134, USA
foresth@seas.upenn.edu

Abstract **-** With many new journals trying to optimize their popularity by changing certain parameters, the effect of the frequency of publication, measured in issues per year, is widely unknown. Previous works have examined predecessors to the impact factor as a measure of popularity, the advantages and disadvantages of open access, and the use of non-citable items in journals. This research paper uses data from twenty-nine high impact (impact factor greater than 10) journals to estimate relationships between different parameters and the impact factor or the total citations. An ordinary least squares regression was performed on the data. The statistical analysis on the relationship between the frequency of publications and the popularity shows that both the number of issues per year and the year the journal started are correlated with the impact factor and the total number of citations, while the field of the journal is not significant. Going from a monthly (12 per year) to a weekly (52 per year) publishing schedule would be an increase of 40 issues per year, which is correlated with an increase in the impact factor of 7.96, which is about half of a standard deviation of the data set. However, the number of issues per year and the age of the journal are positively correlated with each other in this study, leading to the statistical analysis stating that the frequency of publication is less significant than the age of the journal. Since the starting year of a journal cannot be changed, the frequency of publication should be considered to maximize the social and scientific influence.

Keywords: frequency of publication, journal popularity, impact factor, total citations

1. Introduction

Living in the information technology era, everybody reads something every day to gain all sorts of information. The format one reads can be social media, newspapers, magazines, journals, either printed copy or from the internet. While all sorts of social media have their convenience and immediacy, providing their biggest benefit and worst curse, the more serious readers continue to read periodicals for the authenticity, scientific ethics, and long-term impacts.

There are thousands of publications in the world. Some attract millions of readers while others are hardly read. What are the factors affecting the popularity of the publication? The contents of interest, the level of quality, the existence of a paywall, and the frequency of publication are all potential factors that could determine the popularity of a publication.

Another thing to consider is what to use to define the popularity of a publication. Quantifying the popularity of a publication is difficult. One method to do this is by analyzing the total number of subscribers that a publication has. This can have some issues because this can restrict the number of people interested in the publication.

Another method to quantify the popularity is to look at the impact factor. The impact factor of a publication is determined by the number of citations and the number of articles published in the previous two years (1). This allows readers to see how many times an article is cited within two years of its publication. This also adjusts for the number of articles published by the publication.

For young entrepreneurs who start their own publications, the relationship between the frequency of publication and the popularity of the publication is generally a big mystery.

Literary review:

In "Ranking Political Science Journals: Reputational and Citational Approaches" by Micheal W. Giles and James C. Garand, the Robust ISI Impact Score is defined by the Institute for Scientific Information (3). It has a similar definition to

the impact factor: "The RISI Impact score is the average number of citations an article published in the journal during the 1998 to 2003 time-period received in 2003 and 2004 from all journals reported by ISI."

In *The Future of Scholarly Publishing: Open Access and the Economics of Digitisation* by Peter Weingart and Neils C. Taubert (4), the authors talk about the advantages and disadvantages of open access articles for citations. The authors claim that because open access articles are available to a larger audience and are available earlier, open access articles can receive more citations. They also include data showing that different subjects are affected differently by open access, but most subjects had an increase in citations with open access. It follows that having a subscription paywall would be correlated with a decrease in the number of citations, which may lead to a decrease in the journal's impact factor.

In *Multidimensional Journal Evaluation: Analyzing Scientific Periodicals Beyond the Impact Factor* by Stefanie Haustein (5), Haustein discusses the "uncitedness" of different document types. These range from 4.0% of reviews being uncited to 100% of book reviews being uncited. Haustein also talks about how non-citable items can affect a journal's impact factor: "The publication of non-citable items has thus improved the outcome of Lancet's impact factor by 77%" (Haustein, 2012, page 233).

The current paper is similar in spirit to all of these papers because it focuses on the scientific journals and the factors that can affect the journal's impact factor.

2. Methods and data:

The data for the impact factors and the total number of citations comes from impactfactorforjournal.com (2). Publications were only included in the data set if they had an impact factor greater than 10, which is only about 2.2% of all journals in the data set. Additionally, the data set was required to have at least five publications in each of the following ranges: 10-15, 15-20, 20-25, and 25-30. For each of these publications, data about their frequency of publication was taken by looking at the number of publications in 2020. The number of editors was counted on the editorial board or the advisory board page. The "year started" data was taken from the year of the first edition, and the "year started frequency" data was taken from the first year when the number of editions per year was the same as the number of editions in 2020.

The cost of subscription and paywall data is potentially flawed due to some publications having both open access and subscription content. The data for these two may also be difficult to find. As a result, data on these factors were excluded from the regression analysis.

For some publications, the year started frequency is the same as the year started or one year later. As a result, the regression analysis never controls for both of these variables at the same time.

Summary statistics:

Table 1: Summary Statistics

	Impact factor	*Total Citations*	*Issues per year*	*Year Started*	*Year Started Frequency*	*Number of Editors*	*Topic*	*Engineering*	*Medical*	*Science*	*Technology*
Mean	26.7	128316.7	20.4	1955.4	1974.724	41.931	2.655	0.103	0.276	0.483	0.138
Standard Error	2.77	37694.07	3.60	11.50	10.57	8.48	0.16	0.06	0.08	0.09	0.07
Median	22.973	50151	12	1980	2002	26	3	0	0	0	0
Mode	#N/A	#N/A	12	1983	2010	10	3	0	0	0	0
Standard Deviation	14.8984	202988.7770	19.3954	61.9455	56.9366	45.6774	0.8567	0.3099	0.4549	0.5085	0.3509
Sample Variance	221.962	41204443571	376.180	3837.244	3241.778	2086.424	0.734	0.0961	0.207	0.259	0.123
Kurtosis	1.923	3.832	-1.041	0.017	2.101	6.389	-0.288	5.961	-0.950	-2.148	3.123
Skewness	1.476	2.141	0.784	-1.088	-1.777	2.317	-0.339	2.748	1.059	0.073	2.216
Range	59.894	744794	51	204	195	208	3	1	1	1	1
Minimum	10.776	898	1	1812	1824	6	1	0	0	0	0
Maximum	70.67	745692	52	2016	2019	214	4	1	1	1	1
Sum	773.358	3721184	592	56706	57267	1216	77	3	8	14	4
Count	29	29	29	29	29	29	29	29	29	29	29

Table 1 reports summary statistics for the variables. Table 2 reports more specifics regarding the issues per year. The average impact factor was 26.7, which is quite high. However, this sample is restricted to high impact factor journals.

Table 2: Summary Statistics for the frequency ranges

1		4-,6		12		24-36		>36	
Mean	0.137931034	Mean	0.206896552	Mean	0.275862069	Mean	0.137931034	Mean	0.24137931
Standard Error	0.065166288	Standard Error	0.076553056	Standard Error	0.084465164	Standard Error	0.065166288	Standard Error	0.080869237
Median	0	Median	0	Median	0	Median	0	Median	0
Mode	0	Mode	0	Mode	0	Mode	0	Mode	0
Standard Deviation	0.350931203	Standard Deviation	0.41225082	Standard Deviation	0.454858826	Standard Deviation	0.350931203	Standard Deviation	0.43549417
Sample Variance	0.123152709	Sample Variance	0.169950739	Sample Variance	0.206896552	Sample Variance	0.123152709	Sample Variance	0.189655172
Kurtosis	3.123076923	Kurtosis	0.352037656	Kurtosis	-0.95014245	Kurtosis	3.123076923	Kurtosis	-0.405594406
Skewness	2.21632551	Skewness	1.527297457	Skewness	1.05852949	Skewness	2.21632551	Skewness	1.275689994
Range	1	Range	1	Range	1	Range	1	Range	1
Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	0
Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	1
Sum	4	Sum	6	Sum	8	Sum	4	Sum	7
Count	29	Count	29	Count	29	Count	29	Count	29

The standard deviation of the impact factor is 14.9, which is relatively high compared to 26.7. The total citations had a standard deviation that was larger than the average. Almost half of the journals were science, with the others split between engineering, medical, and technology. The average issues per year was 20.4, but this is not smoothly continuous because most academic journals are published periodically, for example, weekly, biweekly, monthly, quarterly, and annually. Therefore, table 2 uses the bins for the frequency. Over a quarter were published on a monthly schedule, while almost a quarter of the journals were published on a weekly basis.

Theory and statistical analysis:

The popularity of a publication, measured by the impact factor, is expected to be positively correlated with the frequency of publication. This is because as the frequency of publications increases, readers have more information to cite from the publication. For this reason, the regression equation for explaining the variation in the Impact Factor variable (Y) may be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon$$

where the explanatory variables are defined as: Y is the impact factor, X_1 is the number of issues per year, X_2 is the year started, X_3 is the year started frequency, X_4 is the number of editors, and the rest of the variables are dummy variables for the topic with respect to technology.

The null hypothesis, in this case, is that $\beta_1 > 0$. As a result, the alternative hypothesis is that $\beta_1 < 0$. A 95% confidence interval will be used, which corresponds to a significance level of 0.05. This significance level balances the probability that the null hypothesis will be rejected when it should not be rejected and the probability that the null hypothesis will not be rejected when it should be rejected.

Using the dependent variable of total citations, the equation can also be written by using a similar functional form:

$$Y_1 = \beta_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \epsilon_1$$

In this equation, Y_1 is the total number of citations.

The null hypothesis, in this case, is as follows:

H_0: $\beta_1 > 0$ and $\beta_{11} > 0$. As a result, the alternative hypothesis is that $\beta_1 < 0$ or $\beta_{11} < 0$. A 95% confidence interval will be used, which corresponds to a significance level of 0.05.

Although data was collected about both the year the publication started and when the publication started publishing at the current frequency, only one of those was controlled for at a time.

3. Results and Discussion of data:

Figure 1 shows the relationship between the impact factor and the frequency of publication. There is still a significant variation in the impact factor at lower frequencies, but there is more variation at higher frequencies. It appears that there is a positive correlation between the issues per year and the impact factor, with a correlation coefficient of 0.479, indicating a relatively strong positive linear association.

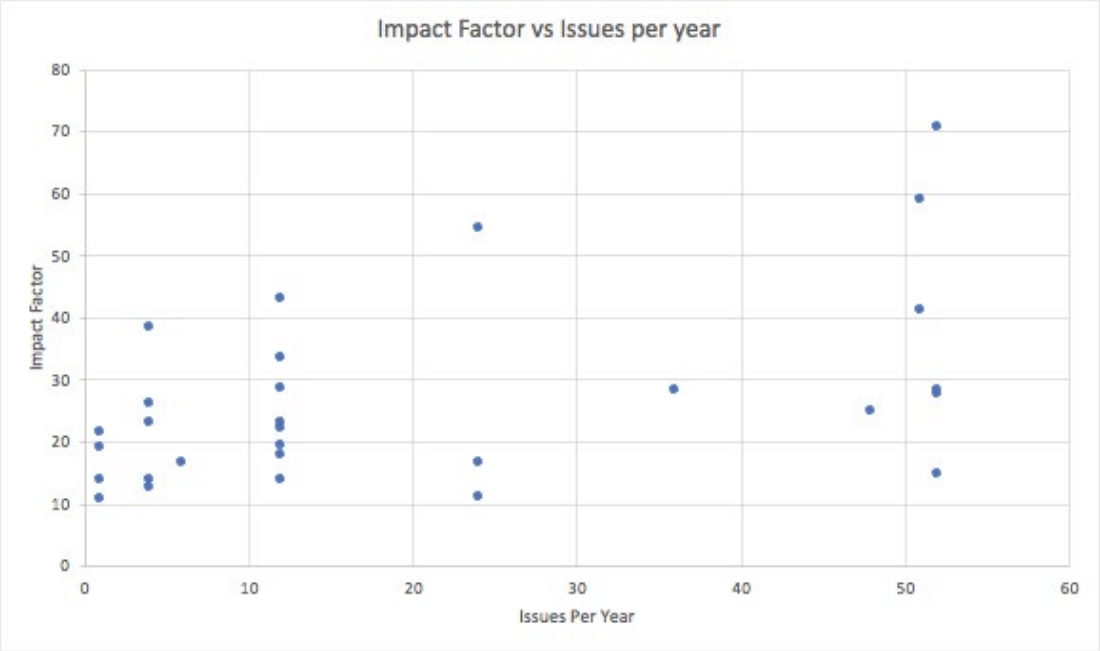

Figure 1: Impact Factor vs Issues Per Year (Unconditional Graph)

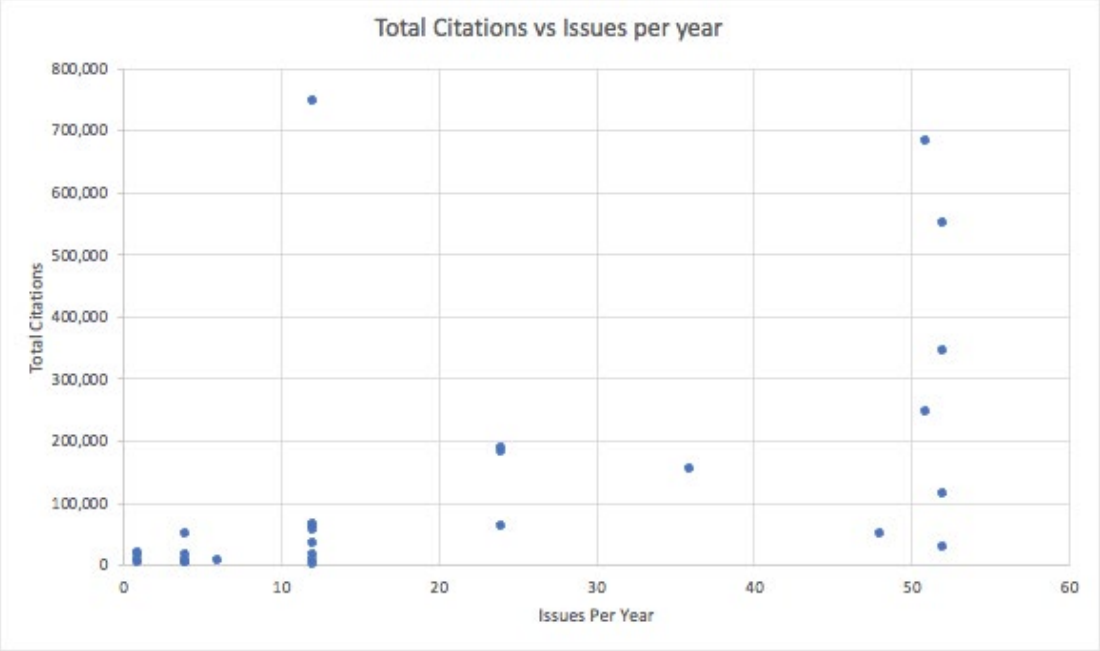

Figure 2: Total Citations vs Issues Per Year (Unconditional Graph)

Figure 2 shows an unconditional graph of the total citations vs the issues per year. The journals publishing fewer than ten times per year all have relatively low numbers for total citations. Although there are some journals that publish more than 30 times per year that have fewer than 100,000 citations, this is a much smaller fraction in comparison to the journals that

publish fewer than 30 times per year. The variation in citations around the frequency of 50 is relatively high, while the variation around the frequency of 12 is relatively small with one notable exception, which is the journal Nature. This makes sense because Nature is among the most prestigious journals. The correlation coefficient between the issues per year and the total number of citations is 0.507 when no other factor is controlled for.

Models 1, 3, 5, 7, and 9 used the dependent variable of impact factor and used continuous variables to describe the frequency of publication.

Table 3: Regressions for Models 1, 3, 5, 7, and 9

Dependent variable is...	Impact Factor	Impact Factor	Impact Factor	Impact Factor	Impact Factor
Independent variable	Model (1)	Model (3)	Model (5)	Model (7)	Model (9)
Issues Per Year	0.368 (0.130)	0.184 (0.134)	0.187 (0.137)	0.0532 (0.153)	0.199 (0.149)
Year Started				-0.183 *** (0.0540)	
Year Started Frequency		-0.127 ** (0.0456)	-0.126 ** (0.0465)		-0.134 ** (0.0535)
Number of Editors			-0.0124 (0.0505)	-0.0653 (0.0547)	-0.0148 (0.0569)
Engineering				-2.96 (9.29)	-3.51 (10.1)
Medical				-9.34 (8.64)	-3.88 (8.98)
Science				-6.21 (7.45)	0.0149 (7.79)
Intercept	19.1 (3.62)	273 (91.4)	272 (93.1)	392 (110)	290 (108)
R^2	0.230	0.406	0.407	0.511	0.422
Adjusted R^2	0.201	0.360	0.336	0.377	0.264
Number of Observations	29	29	29	29	29

Notes: Standard errors are included in parentheses. The topic variables are with respect to technology journals. One asterisk indicates that the coefficient is significant to the 0.1 significance level. Two asterisks indicate that the coefficient is significant to the 0.05 significance level. Three asterisks indicate that the coefficient is significant to the 0.01 significance level.

Models 7 and 9, which control for most factors, both have a positive relationship between the issues per year and the impact factor. These regression coefficients would be interpreted as one more issue per year being correlated with an increase of the impact factor of 0.0532 by model 7 and 0.199 by model 9. However, these are not statistically significant at the 5% significance level. However, this may be mostly due to the small sample size. For ten more issues per year, that would be an increase of the impact factor of 1.99. This is about 13% of one standard deviation in the impact factor. Going from a monthly (12 per year) to a weekly (52 per year) publishing schedule would have an increase of 40, increasing the impact factor by 7.96, which is about half of a standard deviation, a significant increase in the impact factor.

For the year started, the coefficient would be interpreted as starting the journal one year earlier leads to an increase in the impact factor of 0.183, which is at least a 1.8% increase for the journals analyzed. For the year started frequency, the coefficient would be interpreted as having the same frequency for one more year leads to an increase in the impact factor of 0.134, which is at least a 1.3% increase.

The correlation between the impact factor and the issues per year was not statistically significant in any of the models. However, the correlation between the impact factor and the year started or year started frequency was statistically significant in all models where they were included. No other factor was statistically significant in any of these models.

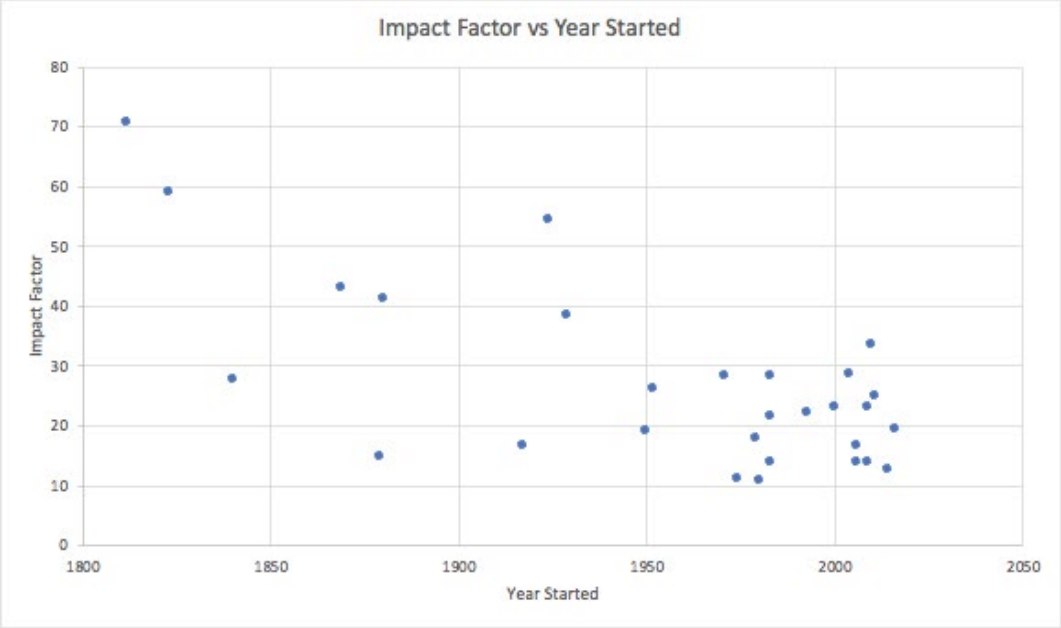

Figure 3: Impact Factor vs Year Started (Unconditional Graph)

Figure 3 is an unconditional graph between the impact factor and the year started. There appears to be a much clearer relationship between the impact factor and the year started in comparison to the impact factor and the issues per year. The R^2 value for the trendline is 0.4649, which means that the correlation coefficient is 0.682.

Models 2, 4, 6, 8, 10, and 11 used the dependent variable of total citations and used continuous variables to describe the frequency of publication.

Table 4: Regressions for Models 2, 4, 6, 8, 10, and 11

Dependent variable is...	Total Citations	Total Citations	Total Citations	Total Citations	Total Citations	Total Citations
Independent variable	Model (2)	Model (4)	Model (6)	Model (8)	Model (10)	Model (11)
Issues Per Year	5305 *** (1736)	4816 ** (2029)	4629 ** (2045)	1379 (1944)	4593 ** (2048)	1570. (1930)
Year Started				-1906 *** (613)		-2118 *** (680.)
Year Started Frequency		-335 (691)	-370 (694)		-487 (737)	
Number of Editors				432 (648)	1000. (782)	386 (688)
Engineering					-42999 (138962)	-17357 (116894)
Medical					16256 (123484)	-93618 (108774)
Science					138965 (107220)	61837 (93693)
Intercept	20024 (48488)	691883 (1386619)	734460 (1391048)	3809752 (1228472)	888087 (1487024)	4219572.384
R^2	0.257	0.264	0.288	0.481	0.410	0.583
Adjusted R^2	0.230	0.207	0.203	0.418	0.250	0.469
Number of Observations	29	29	29	29	29	29

Notes: Standard errors are included in parentheses. The topic variables are with respect to technology journals. One asterisk indicates that the coefficient is significant to the 0.1 significance level. Two asterisks indicate that the coefficient is significant to the 0.05 significance level. Three asterisks indicate that the coefficient is significant to the 0.01 significance level.

For the total citations test, the total number of citations was found to be correlated with the number of issues per year, and this was significant to the 0.05 significance level in models 4, 6, and 10 and significant to the 0.01 significance level in model 2. Following model 10, holding other factors constant, one additional issue per year on average correlates with an additional 4593 citations for journals with an impact factor above 10. However, in models 8 and 11, the total number of citations was correlated with the year started to the 0.01 significance level and the correlation coefficient for the issues per year was not statistically significant to the 0.1 significance level.

Table 5: Regressions for Models 12, 13, 14, 15, 16, and 17

Dependent variable is...	Impact Factor	Impact Factor	Impact Factor	Impact Factor	Impact Factor	Impact Factor
Independent variable	Model (12)	Model (13)	Model (14)	Model (15)	Model (16)	Model (17)
1	-21.802 ** (8.78)	-12.108 (8.188)	-13.066 (9.83)	-13.507 (10.12)	-6.941 (8.426)	-7.102 (10.262)
4-,6	-16.264 ** (7.79)	-5.609 (7.55)	-6.444 (8.63)	-6.3813 (8.8175)	-0.5519 (7.872)	1.0500 (9.133)
12	-12.883 * (7.25)	0.310 (7.58)	0.592 (8.26)	0.2118 (8.499)	2.485 (7.4492)	2.588 (8.0993)
24-36	-10.494 (8.78)	2.122 (8.60)	0.873 (9.79)	1.594 (10.182)	-0.5327 (7.8066)	2.379 (9.2416)
Year Started					-0.162 *** (0.0466)	-0.183 *** (0.0606)
Year Started Frequency		-0.154 *** (0.0501)	-0.148 ** (0.061)	-0.150 ** (0.0627)		
Number of Editors				-0.02369 (0.06170)		-0.07411 (0.06151)
Engineering			-3.512 (10.41)	-3.812 (10.67)		-2.182 (10.04)
Medical			1.353 (11.12)	0.0367 (11.87)		-0.4806 (11.53)
Science			2.050 (7.80)	0.9318 (8.486)		-5.031 (8.407)
Intercept	38.041 (5.29)	332.802 (96.1)	320.066 (119.48)	326.740 (123.34)	343.853 (88.07)	391.066 (119.29)
R^2	0.242	0.462	0.474	0.478	0.503	0.540
Adjusted R^2	0.116	0.346	0.263	0.230	0.395	0.322
Number of Observations	29	29	29	29	29	29

Notes: Standard errors are included in parentheses. The topic variables are with respect to technology journals. One asterisk indicates that the coefficient is significant to the 0.1 significance level. Two asterisks indicate that the coefficient is significant to the 0.05 significance level. Three asterisks indicate that the coefficient is significant to the 0.01 significance level.

Models 12 to 17 use buckets for the frequency with respect to >36. In model 12, the difference is statistically significant for yearly, quarterly, and bimonthly to the 0.05 significance level, and monthly to the 0.1 significance level. This could imply that the wrong functional form for the continuous variable tests was chosen. However, when controlling for the year started or the year started frequency, the difference is not statistically significant. Instead, the year started and the year started frequency are the only factors that are statistically significant for models 13 to 17. Once again, the minus sign indicates that the newer the journal, the lower the journal's impact factor will be.

To assess the robustness of the results and to capture the nonlinearity in the data, a log-linear regression was performed. Without controlling for any other variables, the regression coefficient on the frequency was found to be 0.0118 with a standard error of 0.00439, making it statistically significant at the 5% level of significance. This is generally consistent with the results from the linear model. Controlling for the number of editors and the field gave a regression coefficient of 0.0123 with a standard error of 0.00504, statistically significant at the 5% level of significance. This would be interpreted as one additional issue per year being correlated with a 1.23% increase in the impact factor.

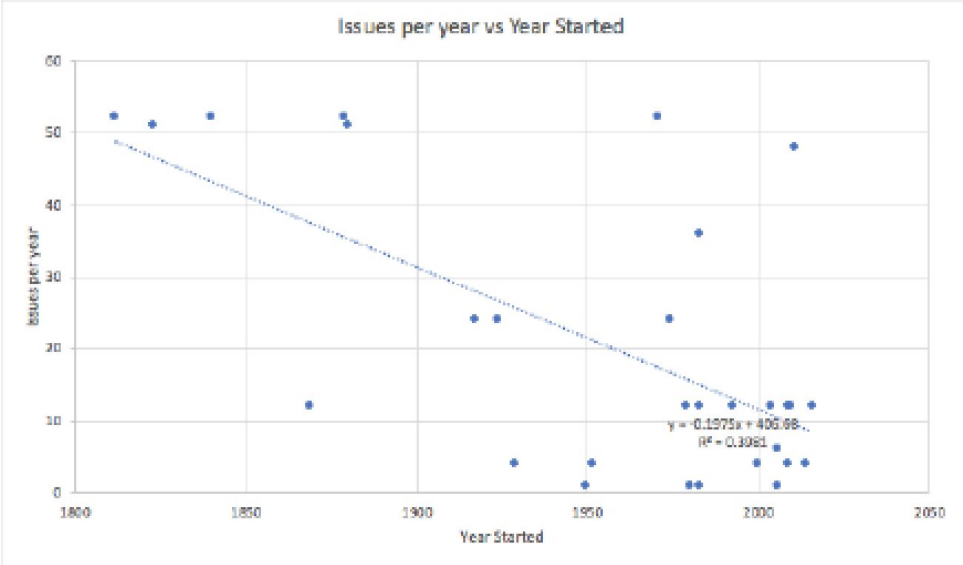

Figure 4: Issues Per Year vs Year Started

Figure 4 shows the relationship between the issues per year and the year started. The R^2 value is 0.3981, which is higher than the R^2 value for model 1, model 2, and model 12, which do not control for the year started or year started frequency. This correlation may cause bias in the regressions that omitted the year started or year started frequency.

Finally, a subset F-test was performed on the field dummy variables to test for the joint significance of the topic. Two regressions were performed with the y variable as the log of the impact factor in order to obtain the goodness of fit measures that go into the F statistic. Controlling for the issues per year, the number of editors, and the field, the R^2 value was 0.224. Without the field variables, the R^2 value was 0.215. This translates to an F statistic of 0.0871, which has a p-value of 0.966. Thus, the field variables are not jointly significant determinants of the impact factor. This is likely due to the small sample size of journals relative to the number of fields.

Limitations:

One major limitation of the statistical analysis was that the sample size was small. Since there were only 29 data points, many of the factors that may be statistically significant may have been labelled as insignificant. However, the data points were collected such that there was variation in the impact factor. Although there was variation in the impact factor, the variation in the fields and topics was relatively limited, as almost half of the journals were science journals. Another issue appeared with the frequency, as all four annual journals are part of the same publisher. In addition, there were non-citable items for some of the journals, so this may be an important variable that is not controlled for as some journals can have large increases in the impact factor due to non-citable items (5). Other factors, such as the average number of articles per issue, the existence of a paywall, the cost of subscription, and the size of the field, which was roughly controlled for by using the categories of engineering, science, medicine, and technology, are other factors that may have a significant influence on the impact factor of a journal.

Avenues for future research:

Future research can focus on some of the missing aspects. For example, the size of the field may have a significant effect on the impact factor. Similarly, non-citable items also need to be accounted for. Determining the relationship between how often teachers and professors tell their students about publications and the impact factor may also be useful.

4. Conclusion:

The null hypothesis, which stated that an increase in the frequency of publication is correlated with an increase in the impact factor, cannot be rejected. The statistical analysis on the relationship between the frequency of publications and the popularity shows that both the number of issues per year and the age of the journal (which is related to the negative of the year started)

are positively correlated with the journal's impact factor. However, the number of issues per year and the age of the journal are positively correlated with each other in this study, leading to the statistical analysis stating that the frequency of publication is less significant than the age of the journal. Since the starting year of a journal cannot be changed, the frequency of publication should be considered to maximize the social and scientific influence.

5. Acknowledgements:

Professor Mark Foley from Davidson College Department of Economics is greatly appreciated for mentorship on statistics, economics, and this research.

6. References:

[1] "The Clarivate Analytics Impact Factor." Web of Science Group, August 6, 2019. https://clarivate.com/webofsciencegroup/essays/impact-factor/.

[2] https://impactfactorforjournal.com/journal-impact-factor-list-2019/

[3] Giles, Micheal W., and James C. Garand. "Ranking Political Science Journals: Reputational and Citational Approaches." PS: Political Science and Politics 40, no. 4 (2007): 741-51. Accessed July 4, 2021. http://www.jstor.org/stable/20452059.

[4] Weingart, Peter, and Niels C. Taubert. 2017. The Future of Scholarly Publishing : Open Access and the Economics of Digitisation. Cape Town, South Africa: African Minds. http://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=1658553&site=e host-live.

[5] Stefanie Haustein. 2012. Multidimensional Journal Evaluation : Analyzing Scientific Periodicals Beyond the Impact Factor. Knowledge & Information. Berlin: De Gruyter Saur