Research Paper
Academic Platform Journal of Engineering and Smart Systems (APJESS) 10(3), 157-167, 2022
ISSN: 2822-2385
Received: 18-Feb-2022 Accepted: 02-Sep-2022
https://doi.org/10.21541/apjess.1075099

Modelling of Factors Influencing the Citation Counts in Statistics

*[1]Olcay ALPAY, [2]Nazan DANACIOĞLU, [3]Emel ÇANKAYA

[1] Department of Statistics, Sinop University, Sinop, Turkey, olcayb@sinop.edu.tr
[2] Department of Statistics, Sinop University, Sinop, Turkey, nazand@sinop.edu.tr
[3] Department of Statistics, Sinop University, Sinop, Turkey, ecankaya@sinop.edu.tr

Abstract

Citation is considered as the most popular quality assessment metric for scientific papers, and it is thus important to determine what factors promote the citation count of a paper in comparison to the others in the same field. The main aim of this study is to model the citation counts of the research published in SCI or SCI-Expanded journals of Statistics field which has a growing number of scientific works in Turkey. Modeling aspect is here highlighted to represent the right-skewed nature of the citations. Due to the fact that distribution of citation counts involves a great number of zeros, this study serves for an additional aim that is to model the counts with advanced discrete regression models for a more precise prediction. Data collected for this study consist of the citation counts of the papers produced by 261 Statisticians in between 2005-2017. Discrete models varying from Poisson to Zero-Inflated or Hurdle were constructed by possible influential factors, such as the publication age, the number of references, the journal category etc. Predictive performances of alternative discrete models were compared via AIC and Vuong test. Results suggested that Zero Inflated Negative Binomial and Hurdle Negative Binomial mixture models are the best forms to predict the zero inflation of citation counts. In addition, the influential factors of the final model were interpreted to make some suggestions to Statisticians to increase the citation counts of their work.

Keywords: Citation Count; Count Data Regression; Negative Binomial Regression; Zero-Inflated Regression

1. INTRODUCTION

Citation counts and related indicators are known to be used as a vital performance criterion in the academic evaluation of scientific articles, journals, researchers, and universities. Scientific journals are usually classified according to the journal impact factor, which is a scale depending on the citation counts of the articles they publish. Therefore, it has recently been questioned why some articles are more cited than the others and which factors affect the citation counts. There are many studies in the literature about citation, especially about estimating citation counts. An explanatory analysis of citation can inform us both about how conductive to citation success the personal characteristics of the authors are (such as their research experience, academic title, gender, etc.) and about the importance of the role of bibliometric features in raising a study's citation rate (such as the length of an article and its number of co-authors) [1]. Putting aside the basic descriptive analyses, the modeling citation count for predictive purposes attracts great attention. Attempts to achieve this aim are based on two separate approaches. The first considers regression modelling that can evaluate the skewness of citation counts with zero inflation. In this respect, generalized linear models (GLMs) have been found particularly useful to model such properties (e.g., Onodera and Yoshikane [2]). The second approach concentrates on classification of publications based on the magnitude of their citation counts. Defined also as machine learning methods, decision trees [3], support vector machines [4], and neural network [5] have been the most frequently used methods for this purpose. However, such methods serving for classification have the drawback of using vague information to define classification boundaries. Besides, classification approach is a simplified form of citation analysis without concentrating on citation patterns or features. To the best of our knowledge, there is only one recent study trying to ease such shortcomings [5].

In this study, we prefer the approach of regression modeling. In this respect, a series of linear modeling were applied by Vaio et al. [1] where the dependent variable is the number of times an author was cited, and the independent variables were the bibliometric variables collected for the basic sample. The question of to what extent the future number of citations that a paper will receive was addressed by Mingers et al. [6] if it is known how many citations it has received so

* Corresponding Author

Attribution (CC BY) license https://creativecommons.org/licenses/by/4.0/

far. Based on retrospective cohort study, Lokker et al. [7] compared 20 articles and journals in terms of citation counts determined by McMaster online ranking system for two years. The potential of a publication to create a scientific change was studied by Chen [8] and they proposed the structural variation model for estimating citation counts. An empirical pilot analysis to the time-dependent distribution of the percentages of never-cited papers was performed by Hu and Wu [9] in a series of different, consecutive citation time windows following their publication in selected six sample journals. They study the influence of paper length on the chance of papers' getting cited. Multiple linear regression was also proposed as a suitable method for the log-transformed data (citation count+1) based on the simulated discrete log-normal data [10]. However, it is well known that the citation count data is right-skewed with a potential number of zeros and log transformation is not the best strategy for modeling such data. Thus, there have been attempts to utilize the generalized linear modeling like in Maliniak et al. [11] who reported a significant influence of gender variable amongst the many others by means of Negative Binomial Model. Besides they highlighted huge gap between the genders. For this conclusion, Zigerell [12] commented that more data is necessary to come through this result and stated that the gender gap is more prevalent in elite journals. Applicability of different models by right-censoring the data was also assessed in Santos and Irizo [13] so as to deal with the skewness of citation counts.

Motivated by extracting what features are responsible for particularly zero counts, we here prefer the GLM approach in modeling counts rather than classification approach. It is also of our interest to compare the predictive performances of existing discrete regression models. Understanding of the intriguing factors that influence citations can be the goal of different scientific disciplines. However, literature review we performed revealed that such modeling strategies have been applied to limited number of scientific disciplines. For example, Qian et al. [14] applied negative binomial regression models to study the effect of various factors on the citation rates in Computer science. Politics, Economy, and Business appears to be attractive fields ([1], [6], [11]). Ahlgren et al. [15] used a very large publication set, however, across all disciplines regardless of which field it is.

Various studies indicate that the citation behavior of a paper differentiates according to the scientific field, or even sub-fields. For example, the number of citations per paper is detected to be much higher in social sciences than natural sciences [16]. In some disciplines, there may be "hot" topics or sub-fields that can influence the paper to be highly-cited. For example, papers published on analytic chemistry, organic chemistry and physical chemistry receive more citations than those on biochemistry [17]. Therefore, disciplines have different citation manners and factors affecting the citation counts of the papers vary accordingly [18]. In order to give an insight about the factors affecting the citation counts, we present a collection of studies belonging to different disciplines with varying methodological aspects in Table A1. It can be seen that the considered factors can be categorized as paper-related (e.g., characteristics of title, abstract, references of the study topic), author-related (e.g., number of authors, authors' academic rank or gender) and journal-related (e.g., the scope or impact factor of the journal). As far as Statistics science is concerned, however, such a study of citation behavior for the papers of this discipline appears lacking in the literature.

Therefore, this study aims to fulfill this gap by modeling the citation counts of publications in the SCI or SCI-Exp. journal lists, belonging to Turkish academicians in the field of Statistics. It offers a modest attempt to identify some of the factors that determine the citation counts of authors who published their work in this field. Besides, on the contrary to the work listed above, we put here more focus on the modeling the articles with zero citation counts. Discrete regression models like basically Poisson or Negative Binomial could be adequate for moderate size of zero counts; however, advanced discrete regression strategy is necessary to achieve the excess of zeros. Amongst the alternative zero-inflated models, we are here motivated to perform a comparison and choose the best one.

2. DISCRETE REGRESSION MODELS

The number of occurrences of any event as a result of the trials carried out in a specified process can be expressed as count data. The count data is the type of data that the observations can only take non-negative integer values (0, 1, 2, 3…). It is well known that the application of linear regression modeling (suitable for continuous response) for count data reveal inefficient and inconsistent predictions. General Linear Models (GLM), however, is a combination of linear and nonlinear regression models that take into account the non-normality of count data [19]. For example, Poisson Regression (PR) is the first natural choice [20] and other discrete models commonly preferred to describe the relationship between variables are Negative Binomial Regression (NBR), Zero-Inflated Regression (ZIP-ZINB) and Hurdle Regression (HP-HNB).

2.1. Poisson Regression (PR)

Let Y be a non-negative integer valued random variable and has a Poisson distribution with mean (μ) parameter set as $\mu = \lambda$. Then the probability function of Y is given as

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad , \quad y = 0,1,2,.... \text{ and } \lambda > 0 \tag{1}$$

with the expected value (E) and variance (V) of the function in (1) equal to

$$E(Y) = V(Y) = \lambda \tag{2}$$

Regression modelling can be constructed through the natural log link function as

$$log(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{3}$$

where β_i are unknown regressors and x_i are the predictors. Conditional expected value is then obtained by exponentiating both sides of (3) as below:

$$E(Y_i|x_i) = exp(x_i^{'}\beta) \tag{4}$$

2.2. Negative Binomial Regression (NB)

NB regression provides a facility to slacken the assumption that the mean is equal to the variance, essential for the Poisson model. The classical NBR model is a mixture of Poisson and Gamma distribution.

$$P(Y=y_i)=\frac{\Gamma(y_i+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_i+1)}\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i}\left(\frac{1}{1+\theta\lambda_i}\right)^{\theta^{-1}} \quad y_i=0,1,\ldots. \tag{5}$$

The expected value and variance of the distributional form of (5) are

$$E(Y)=\mu \text{ and } V(Y)=\mu(1+\mu\theta) \tag{6}$$

where θ is a dispersion parameter and μ is the mean parameter.

For a standard counting model, if the data contains more zeros than expected, it is called as zero-inflation. In this case, two-part mixed models will be preferred to fit the data. These

- Zero-inflated models
- Hurdle models

2.3. Zero-Inflated Regression Models (ZIP-ZINB)

Let π is the structural zero ratio, then zero-inflated regression models can be expressed as follows:

$$P(Y=y_i)=\begin{cases}\pi_i+(1-\pi_i)P(S_i=0) & y_i=0\\ (1-\pi_i)P(S_i=y_i) & y_i>0\end{cases} \tag{7}$$

where $P(S)$ is the probability function of random variable S for which any discrete distribution can be selected. Generally, Poisson or Negative Binomial distribution is preferred, and inserting the probability function of these in (7) results in ZIP and ZINB models respectively.

2.4. Hurdle Regression Models (HP-HNB)

This model is also the mixture of the two components, the first of which includes the binary responses showing the positive counts (1) against the zero counts (0); the second includes only positive counts.

(i) Hurdle Poisson (HNB) Model:

$$P(Y=y_i)=\begin{cases}w_i & y_i=0\\ (1-w_i)\frac{e^{-\mu_i}\mu^{y_i}}{(1-e^{-\mu_i})y_i!} & y_i>0\end{cases} \tag{8}$$

(ii) Hurdle Negative Binomial (HNB) Model:

$$P(Y=y_i)=\begin{cases}w_i & y_i=0\\ (1-w_i)\frac{\Gamma(y_i+\theta^{-1})}{y_i!\Gamma(\theta^{-1})}\frac{(1+\theta^{-1}\mu_i)^{-(y_i+\theta^{-1})}\theta^{-y_i}\mu_i^{y_i}}{1-(1+\theta^{-1}\mu_i)^{\theta^{-1}}} & y_i>0\end{cases} \tag{9}$$

In both models defined in (8) and (9), μ is the mean parameter, θ is the dispersion parameter and $w_i=P(Y_i=0)$.

3. MODEL SELECTION CRITERIA

In this study, three information criteria based on log-likelihood, and Vuong statistic were used in deciding the appropriate model.

Information Criteria:

- $AIC=-2LL+2k$
- $AICc=AIC+2k(k+1)/(n-k-1)$
- $BIC=-2LL+k(ln(n))$

where AIC = Akaike information criterion, AIC_c = Corrected AIC, BIC = Bayesian information criterion, LL = log-likelihood, k = Number of parameters in the estimated model and n = Sample size.

Vuong Statistics:

Assessment of fitting performance of Model 1 vs Model 2 by means of their corresponding probability functions $P_1(.)$ and $P_2(.)$ respectively can be achieved by Voung statistic (V) as below:

$$V=\frac{\overline{m}\sqrt{n}}{S_m}\sim N(0,1) \text{ and } m_i=log\left(\frac{P_1(Y_i|X_i)}{P_2(Y_i|X_i)}\right) \tag{10}$$

- $V>1.96\Rightarrow 1^{st}$ model is preferred.
- $V<-1.96\Rightarrow 2^{nd}$ model is preferred.

where $\overline{m}$ = mean and S_m= standard deviation of m_i values.

4. APPLICATION

There are 32 Statistics Departments in state and private universities in Turkey. Additionally, we searched for Statisticians employed in other departments like Econometrics, Actuary, Biostatistics, etc. and found 261 academicians in total. SCI & SCI-Exp. publications (2005-2017) of all those Statisticians were searched through Scopus, and all the analyses were performed in SPSS and R.

Response variable: Citation counts of 1529 papers published in the period of 2005-2017.

Factors:

- Age (Age of paper)
- Ref (Number of references in the paper)
- Ath (Number of authors in the paper)
- FCA (First citation age)
- Pg (Number of pages in the paper)
- TL (Title length)

Journal Categories: Artificial Intelligence (AI), Biology (Bio), Chemistry (Chm), Computer (Comp), Econometric (Eco), Education (Edu), Energy (Engy), Engineering (Eng), Environment (Env), Fuzzy (Fuz), Management (Man), Mathematics (Math), Medicine (Med), Operational Research (OR), Other (Oth), Physics (Phy), Social (Soc), Statistics (Stat)

Data were first analyzed descriptively using the demographic features. It is observed that authors of the papers are almost

equally distributed in gender (Figure 1). Distribution of authors in terms of academic title is 37% assistant professors, 23% associate professors and 40% professors respectively (Figure 2). Besides, only 28% of those work for private universities (Figure 3).

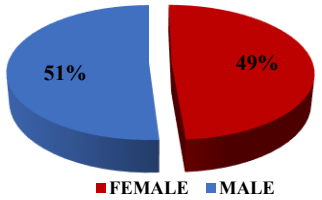

Figure 1. Distribution of authors by gender

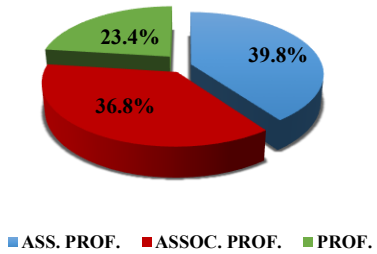

Figure 2. Distribution of authors by title

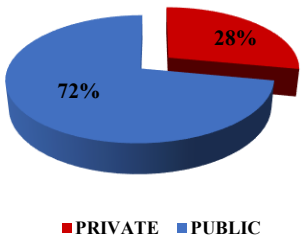

Figure 3. Distribution of authors by type of workplace

When the academicians are evaluated according to their experiences, Figure 4 reveals that majority of those (85.5%) have less than 20 years of experience.

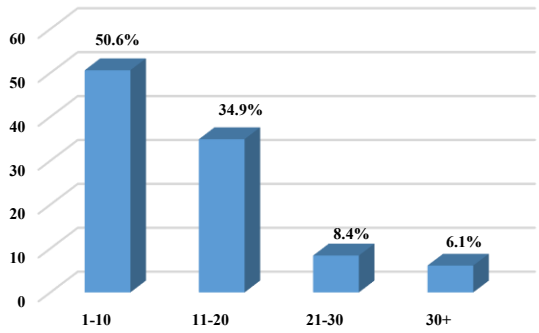

Figure 4. Distribution of authors by experience (years)

It can be observed that 79.7% of the authors have published less than 10 articles (Figure 5).

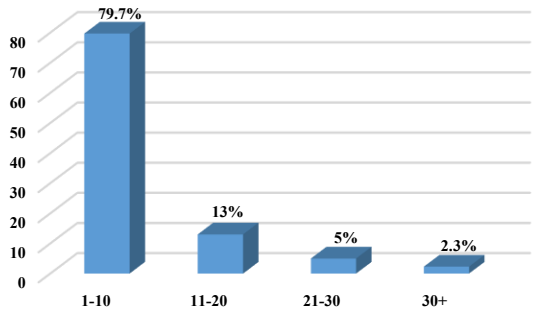

Figure 5. Distribution of authors by the number of publications

Visual presentation of how the total amount of publications is distributed amongst the authors of different titles is presented in Figure 6. To exemplify, 39.8% of our sample consists of assistant professors (Figure 2), 38.2% of whom owns 1-10 papers; 0.8% publish 11-20 papers; 0.8% publish 21-30 papers and none (0%) publishes 30+ papers as seen with blue bars of Figure 6. Publications with an amount of 1-10 seem to be highly produced by assistant professors. Although the percentages of more than 11, 21 or 30 publications are low in total, it can be seen that the number of publications naturally increases in line with the title of authors (Figure 6).

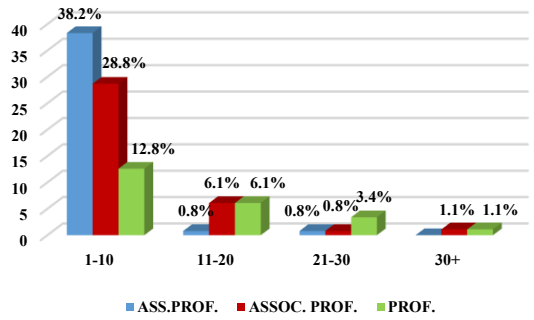

Figure 6. Number of publications by authors' title

A simultaneous evaluation of the authors with respect to their working experiences and the number of publications they realized is performed by the crosstabulation produced as in Table 1. For interpretation of such a table, let's concentrate on the authors who have 1-10 years of experience (highlighted in grey) for the moment. The total of this row shows that 132 authors have such an experience, which corresponds to 50.6% of the grand total (that is 261 as seen in the right bottom corner). Cell frequencies in this row show that how 132 authors are distributed amongst the number of publication categories. For example, 118 authors of those own 1-10 number of papers. This table also presents the percentages of cell frequencies within the experience (the row total), within the number of SCI publication (the column total) and also within the grand total. For the above example, we can say that 89.4% of the authors with 1-10 years of experience (row total=132) publish 1-10 number of papers. We can additionally say that 56.7% of all authors owning 1-10 number of papers (column total=208) have 1-10 years of experience. Besides, we can also say that 45.2% of the whole sample (grand total=261) has 1-10 years of experience and

also 1-10 number of papers. Therefore, such a presentation enables one to make interpretation within the row or column category as well as within the whole sample. According to totals within the whole sample in Table 1, it appears that half of authors (50.6%) have an experience less than 10 years and also the majority of the authors (79.7%) produced 1-10 number of papers. Amongst the authors experienced up to 10 years, the percentage of having 1-10 SCI publications appears to be 89.4%. Although it is not as high as this percentage, a similar pattern is observed for each experience category. That is the majority of each experience group owns 1-10 number of papers (see %within experience). If we look at "% within SCI" values, more than half of the authors with 1-10 number of papers (56.7%) have the experience less than 10 years. However, productivity corresponding to more than 10 papers (i.e., 11-20, 21-30, 30+) belongs mainly to the authors having 11-20 years (see the percentages of 47.1%, 46.2% and 83.3% for this experience group). Surprisingly, the productivity of the academicians decreases as they get more experienced.

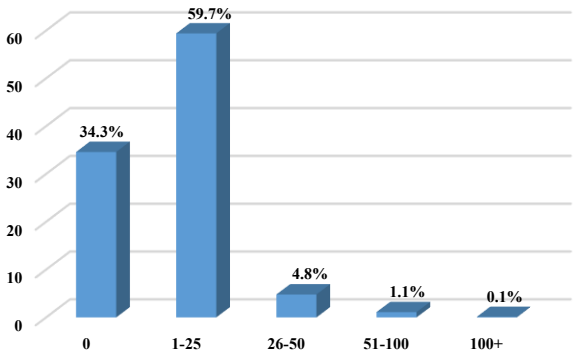

Figure 7. Distribution of citation counts

Table 1. Number of SCI publications according to experiences of authors

			# of SCI Publications				
			1-10	**11-20**	**21-30**	**30+**	**Total**
Experience	**1-10**	Count	118	10	3	1	132
		% within experience	89.4%	7.6%	2.3%	.8%	
		% within SCI	56.7%	29.4%	23.1%	16.7%	
		% of Total	45.2%	3.8%	1.1%	.4%	50.6%
	11-20	Count	64	16	6	5	91
		% within experience	70.3%	17.6%	6.6%	5.5%	
		% within SCI	30.8%	47.1%	46.2%	83.3%	
		% of Total	24.5%	6.1%	2.3%	1.9%	34.9%
	21-30	Count	14	5	3	0	22
		% within experience	63.6%	22.7%	13.6%	.0%	
		% within SCI	6.7%	14.7%	23.1%	.0%	
		% of Total	5.4%	1.9%	1.1%	.0%	8.4%
	30+	Count	12	3	1	0	16
		% within experience	75.0%	18.8%	6.3%	.0%	
		% within SCI	5.8%	8.8%	7.7%	.0%	
		% of Total	4.6%	1.1%	.4%	.0%	6.1%
	Total	Count	208	34	13	6	261
		% of Total	79.7%	13.0%	5.0%	2.3%	100.0%

Table 2. Citation counts according to experiences of faculty member

			Citation Counts						
			0-25	**26-50**	**51-100**	**101-150**	**151-200**	**200+**	**Total**
Experience	**1-10**	Count	100	16	6	6	2	2	132
		% within experience	75.8%	12.1%	4.5%	4.5%	1.5%	1.5%	
		% within citation	57.5%	51.6%	26.1%	60.0%	22.2%	14.3%	
		% of Total	38.3%	6.1%	2.3%	2.3%	.8%	.8%	50.6%
	11-20	Count	51	12	12	3	3	10	91
		% within experience	56.0%	13.2%	13.2%	3.3%	3.3%	11.0%	
		% within citation	29.3%	38.7%	52.2%	30.0%	33.3%	71.4%	
		% of Total	19.5%	4.6%	4.6%	1.1%	1.1%	3.8%	34.9%
	21-30	Count	13	1	5	1	0	2	22
		% within experience	59.1%	4.5%	22.7%	4.5%	.0%	9.1%	
		% within citation	7.5%	3.2%	21.7%	10.0%	.0%	14.3%	
		% of Total	5.0%	.4%	1.9%	.4%	.0%	.8%	8.4%
	30+	Count	10	2	0	0	4	0	16
		% within experience	62.5%	12.5%	.0%	.0%	25.0%	.0%	
		% within citation	5.7%	6.5%	.0%	.0%	44.4%	.0%	
		% of Total	3.8%	.8%	.0%	.0%	1.5%	.0%	6.1%
	Total	Count	174	31	23	10	9	14	261
		% of Total	66.7%	11.9%	8.8%	3.8%	3.4%	5.4%	100.0%

As far as citation counts are concerned, the distribution is obtained as in Figure 7, highly right-skewed as expected. It can be seen that there is a high percentage of publications with zero citation counts (n=525 corresponding to 34.3%).

When academicians are evaluated according to the number of citations they receive for their publications, it can be seen that the least number of citations in total belong to professors (Table 3). On the contrary, papers having more than 100 citations appear to be produced by associate professors and professors, presumably due to the most aged papers belong to those academicians.

Table 3. Citation counts by title of authors

Citation Counts	Ass. Prof.	Assoc. Prof.	Prof.
0-25	33.7%	23.4%	9.6%
26-50	4.2%	5.4%	2.3%
51-100	0.8%	3.4%	4.6%
101-150	1.1%	0.8%	1.9%
151-200		1.1%	2.3%
200+		2.7%	2.7%
Total	39.8%	36.8%	23.4%

Crosstabulation of citation count groupings against the experiences of authors is presented in Table 2. It is observed that majority of the authors own papers having less than 50 citations regardless of their experiences. Astonishingly, although only 5.4% of the authors manage to receive more than 200 citations, majority of this percent belongs to those having 11-20 years of experience. It can be concluded that the highly influential papers are more likely to be produced within 11-20 years of experience.

Table 4. Expected percentages and zeros according to all fitted models

	PR	NB	ZIP	ZINB	HP	HNB
Expected counts	133	520	525	541	525	525
Expected percentages (%)	8	34	34.3	35.4	34.3	34.3

(Observed number of zeros is 525 corresponding to 34.3%)

As mentioned above the distribution of citation counts is right-skewed in nature and generalized linear models are more suitable to represent this skewness. Besides, the high percentage of zero citation counts requires particular attention in modeling. Motivated by determining the influential factors on citation counts, PR, NB, ZIP, ZINB, HP, and HNB models were fitted the data and estimated parameters are listed in Table A2. Significant factors ($p<0.05$) are here indicated as bold.

Performance assessment in predicting zero counts for all the considered models is presented in Table 4. Noting that the observed percentage of zeros is 34.3%, Zero Inflated Poisson and Hurdle Models seem to be better than others for this purpose. As expected, PR model is not appropriate for this type of skewed data.

Table 5. Information criteria results

	PR	NB	ZIP	ZINB	HP	HNB
AIC	15586.0	7474.4	13245.5	7365.8	13251.2	7452.6
AICc	15586.8	7475.2	13248.7	7369.1	13254.4	7455.9
BIC	15615.6	7504.0	13229.2	7347.4	13234.8	7436.2
-LL	7768.0	3712.2	6574.8	3633.9	6577.6	3678.3

Model comparison in terms of information criteria as given in Table 5 suggested that ZINB and HNB models are the most suitable ones. As ZINB is slightly better than HNB, Vuong statistics would be more decisive to select the final model. This test is based on the principle that the differences between the likelihoods indicate which model fits the data better. Therefore, only significant pairwise comparisons of models achieved via Vuong test were presented in Table 6. Note that the estimated values (V) give the indication of model preferences based on the comparisons as described in Section 3.

Table 6. Estimated Vuong Statistics (v) with model preferences

Model 1 \ Model 2	NB	ZIP	ZINB	HNB
PR	-12.96 *Model 2> Model 1	-9.36 Model 2> Model 1	-13.27 Model 2> Model 1	-13.12 Model 2> Model 1
NB		10.61 Model 1> Model 2		-3.31 Model 2> Model 1
ZIP				-10.73 Model 2> Model 1
ZINB				3.93 Model 1> Model 2

For interpretation, *e.g., estimated $V = -12.96$ implies Model 2 (NB) is significantly better than Model 1 (PR).

Results indicate that ZINB is better than HNB, thus considered as the final model for the data at hand. Mathematical form of the final model can be presented as

$$\begin{aligned}\mu = exp(&-1.353 + \mathbf{0.247}Age + \mathbf{0.017}Ref + 0.007Ath - 0.014FCA + \mathbf{2.062}Pg - \mathbf{0.002}TL + \mathbf{1.911}AI \\ &+ \mathbf{1.05}Bio + 1.221Chm + \mathbf{2.062}Comp + \mathbf{1.043}Eco + \mathbf{1.314}Edu + \mathbf{2.402}Engy + \mathbf{2.114}Eng \\ &+ \mathbf{1.316}Env + \mathbf{1.829}Fuz + 0.827Man + \mathbf{1.334}Math + \mathbf{1.676}Med + \mathbf{2.276}OR + \mathbf{1.242}Oth \\ &+ 1.062Phy + \mathbf{0.989}Soc + \mathbf{1.274}Stat)\end{aligned} \tag{10}$$

for the count part, and

$$\begin{aligned}(\pi/(1-\pi)) = exp(&7.046 - \mathbf{2.378}Age - 0.03Ref - 0.034Ath + 0.029FCA - 3.485Pg + 0.004TL - 19.7AI \\ &- 0.494Bio - 0.378Chm - 3.485Comp - 2.134Eco - 3.297Edu - \mathbf{4.884}Engy - \mathbf{4.367}Eng \\ &- 3.48Env - 2.113Fuz - 3.861Man - 3.36Math - \mathbf{3.408}Med - 1.648OR - 4.107Oth \\ &- 0.954Phy - 3.957Soc - 2.903Stat)\end{aligned} \tag{11}$$

for the binary part. The regression coefficients given as bold are the significant factors at 0.05 significance level.

5. CONCLUSION

There is ever growing interest for the quality assessment of research papers and citation count has been considered as the best indicator for this purpose. Although this measure is frequently used in variety of disciplines, there is a lack of interest for Statistics science. In this respect, our study fulfilled this gap and presented influential factors that affect the citation counts of the papers in Statistics. This study also investigates the suitability of the models within GLM framework. Due to the high percentage of zero citation counts and skewness of its distribution ZINB model is concluded to be the best amongst the others to model the citation counts.

In the final model, the citation counts are observed to be positively related to the age of the paper, the number of references and the number of pages in the paper. It is natural to observe the amount of citation to increase as the paper gets more aged. The variety of references as a reflection of the authors' knowledge also increases the frequency of the citation. The number of references in a paper has been stated as a good predictor for the citation behavior in many studies (e.g., [2], [8], [17], [21]). Besides, the length of the paper seems to rise up the citation counts as also stated by Stremersch et al. [22]. However, the question of what the optimal number is requires a special methodological attention and the answer differs from discipline to discipline. Vaio et al. [1] detected the optimal number of pages as 36 for Economic History, however, Robson and Mousquès [23] stated this number as 11 for Environmental Modeling discipline.

As reported in many studies, we here also detected that the study topics or sub-fields of Statistics have high impact on the citation of the papers. Fields of Artificial Intelligence, Biology, Computer, Econometric, Education, Energy, Engineering, Environment, Fuzzy, Mathematics, Medicine, Operational Research, and Social studies are observed to be the most influential subjects.

On the other hand, our model suggests that title length affects the citation count negatively. That is, the papers with longer titles receive significantly less citation than those with shorter titles. Such an effect has also been stated by a variety of studies in the literature (e.g., [24], [25]). However, it must be noted that this influence can depend on the discipline of the study as the papers of Medicine with longer titles were detected to receive more citations [26].

In addition, binary part of our model suggests that the probability of observing zero citation can be reduced by publishing the papers in the disciplines of Energy, Engineering, and Medicine. It can also be concluded that the higher the age of paper is, the lesser zero counts is. This is a natural result as the paper gets older, the probability of receiving citation increases. This effect is also reflected in positive count part of the model as the coefficient for the age of the paper variable is significant and positive.

In order to increase the citation counts, academicians studying in the field of Statistics are recommended to keep the title of the paper short, increase the number of references, and produce immense papers with more pages. Additionally, they can lead their interest to the statistical applications of particularly Operational Research, Computer Science, Artificial Intelligence, Fuzzy as well as Energy, Engineering, and Medicine.

For those who are interested, the study of the optimal number of pages for the papers of Statistics discipline can be suggested as a future work. Besides, models constructed here can be reproduced by enlarging the variety of the factors that presumably have impact on the citation behavior in Statistics.

Author contributions: The authors contributed equally to the study.
Conflict of Interest: No conflict of interest was declared by the authors.
Financial Disclosure: This paper is a part of Sinop University Scientific Research Project: FEF-1901-16-07.
Acknowledgments: The authors would like to thank the Editors and anonymous reviewers for their valuable comments.

REFERENCES

[1] G. Di Vaio, D. Waldenström and J. Weisdorf, Citation success: Evidence from economic history journal publications. Explorations in Economic History, vol. 49, no. 1, pp. 92-104, 2012.

[2] N. Onodera and F. Yoshikane, Factors affecting citation rates of research articles. Journal of the Association for Information Science and Technology, vol. 66, no. 4, pp.739–764, 2015

[3] M. Y. Wang, G. Yu, and D.R. Yu, Mining typical features for highly cited papers. Scientometrics, vol. 87 no. 3, pp. 695–706, 2011.

[4] L. Fu and C. Aliferis, Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics, vol. 85, no.1, pp. 257-270, 2010.

[5] X. Ruan, Y. Zhu, J. Li and Y. Cheng, Predicting the citation counts of individual papers via a BPneural network. Journal of Informetrics, 14, 2020.

[6] J. Mingers, F. Macri and D. Petrovici, Using the h-index to measure the quality of journals in the field of Business and Management. Information Processing & Management, vol. 48, no. 2, pp. 234-241, 2012.

[7] C. Lokker, K.A. McKibbon, R.J. McKinlay, N.L. Wilczynski and R.B. Haynes, Prediction of Citation Counts for Clinical Articles at Two Years Using Data Available Within Three Weeks of Publication: Retrospective Cohort Study. Bmj, vol. 336, no. 7645, pp. 655-657, 2008.

[8] C. Chen, Predictive effects of structural variation on citation counts. Journal of the Association for Information Science and Technology, vol. 63, no. 3, pp. 431-449, 2012.

[9] Z. Hu and Y. Wu, Regularity in the time-dependent distribution of the percentage of never- cited papers: An empirical pilot study based on the six journals. Journal of Informetrics, vol. 8, no. 1, pp. 136–146, 2014.

[10] M. Thelwall and P. Wilson, Regression for Citation Data: An Evaluation of Different Methods. Journal of Informetrics, vol. 8, no. 4, pp. 963–971, 2014.

[11] D. Maliniak, R. Powers and B.F. Walter, The gender citation gap in international relations. International Organization, vol. 67, no. 4, pp. 889-922, 2013.

[12] L.J. Zigerell, Is The Gender Citation Gap in International Relations Driven by Elite Papers?. Research & Politics, April-June, 1-7, 2015.

[13] J.B. Santos and F.J.O. Irizo, Modelling Citation Age Data with Right Censoring. Scientometrics, vol. 62, no. 3, pp. 329-342, 2005.

[14] Y. Qian, W. Rong, N. Jiang, J. Tang and Z. Xiong, Citation regression analysis of computer science publications in different ranking categories and subfields. Scientometrics, vol. 108, pp. 1–24, 2017.

[15] P. Ahlgren, C. Colliander and P. Sjogarde, Exploring the relation between referencing practices and citation impact: A large-scale study based on Web of Science data. Journal of the Association for Information Science and Technology, vol. 69, no.5, pp. 728–743, 2018.

[16] P.F. Skilton, Does the human capital of teams of natural science authors predict citation frequency?. Scientometrics, vol. 78, no. 3, pp. 525–542, 2009.

[17] L. Bornmann, H. Schier, W. Marx, and H.D. Daniel, What factors determine citation counts of publications in chemistry besides their quality?. Journal of Informetrics, vol. 6, no. 1, pp. 11–18, 2012.

[18] I. Tahamtan, A.S. Afshar, and K. Ahamdzadeh, Factors affecting number of citations: a comprehensive review of the literature. Scientometrics, vol. 107, no. 3, pp. 1195-1225, 2016.

[19] J. Hardin and J. Hilbe, Generalized linear models and extensions. Texas, USA: Stata Corporation, 2012.

[20] E. Arıcan, Nitel yanıt değişkene sahip regresyon modellerinde tahmin yöntemleri. Master Thesis. Cukurova University, Institute of Science, 2010. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp.

[21] F. Didegah, and M. Thelwall, Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, vol. 64, no. 5, pp. 1055–1064, 2013.

[22] S. Stremersch, I. Verniers, and P.C. Verhoef, The quest for citations: Drivers of article impact. Journal of Marketing, vol. 71, no. 3, pp. 171-193, 2007.

[23] B.J. Robson, and A. Mousquès, Predicting citation counts of environmental modelling papers. 7th International Congress on Environmental Modelling and Software - San Diego, California, USA - June 2014.

[24] I. Ayres, and F.E. Vars, Determinants of citations to articles in elite law reviews. The Journal of Legal Studies, vol. 29(S1), pp. 427-450, 2000.

[25] S. Stremersch, N. Camacho, S. Vanneste, and I. Verniers, Unraveling scientific impact: Citation types in marketing journals. International Journal of Research in Marketing, vol. 32, no. 1, pp. 64-77, 2015.

[26] M. E. Falagas, A. Zarkali, D.E. Karageorgopoulos, V. Bardakas, and M.N. Mavros, The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. PLoS One, 8(2), e49476, 2013.

[27] X. Bai, F. Zhang and I. Lee, Predicting the citations of scholarly paper. Journal of Informetrics, vol. 13, pp. 407–418, 2019.

[28] H. Beydokhti, N. Riahinia, H.R. Jamali, S. Asadi and S.M. Riahi, Factors Affecting the Number of Citations to Clinical Therapeutic Articles Mentioning Level of Evidence. Mod Care J. vol. 17, no. 2, 2020.

[29] Z. Su, Prediction of future citation count with machine learning and neural network. 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 14-16 April 2020.

Appendix A

Table A1. A collection of study examples on citation

Author	Discipline	Method	Influential factors
Santos and Irizo [13]	• Applied Economics	• Maximum likelihood estimation with data censored to the right (Log-normal, Weibull and Log-logistic)	• Log-logistic model is the best fit
Lokker et al. [7]	• Medicine	• Multiple regression	• The number of authors • Selection for abstraction in a synoptic journal • Clinical relevance score • Number of pages • Structured abstract • Number of cited references • Original papers • Multicentered study • Study about therapy
Fu and Aliferis [4]	• Biomedical publications • Cardiology • Endocrinology • Gastroenterology • Hematology • Medical Oncology • Nephrology • Pulmonary disease • Rheumatology	• Machine learning methods (SVM models) • Logistic regression-based classifier	• Citation history of authors • Certain topics • High impact journal
Wang et al. [3]	• Astronomy • Astrophysics	• Data mining	• The number of citations that papers obtain • Authors with high reputations receive disproportionately more citations than authors with low reputations • The reputation of a journal
Vaio et al. [1]	Economic history	• Poisson and negative binomial regression modeling	• Professors at economics and history departments • Length of papers • Co-authors • Conference and workshops presentations • Anglo-Saxon and German history • Gender
Chen [8]	• Terrorism • Mass extinction • Complex network analysis • Knowledge domain visualization	• Negative binomial models of complex network analysis • Zero-inflated negative binomial regression models	• The number of coauthors • The number of references
Maliniak et al. [11]	• Teaching, • Research, • International Policy	• Network analysis • Negative-binomial model	• Women tend to cite themselves less than men • Men tend to cite men more than women • Age of paper • Co-authorship employed by research university
Hu and Wu [9]	• Information Science • Multi-disciplinary Science	• Negative exponential model	• The length of a paper

Table A1. (Continues) A collection of study examples on citation

Author	Discipline	Method	Influential factors
Onodera and Yoshikane [2]	• Condensed matter physics, • Inorganic and nuclear chemistry • Electric and electronic engineering • Biochemistry and molecular biology, • Physiology, • Gastroenterology)	• Negative binomial multiple regression	• The price index • Number of references
Oian et al. [14]	• Computer science	• Negative binominal regression model	• Classification of a publication • Number of authors • Maximum h-index of all authors of a paper • Average number of papers published by a publication
Ahlgren et al. [15]	• Bibliometric	• Quantile regression	• Number of cited references • References to more recent publications
Ruan et al. [5]	• Library • Information • Documentation	• Four-layer back propagation (BP) neural network model	• Citations in the first two years • First-cited age • Paper length • Month of publication • Self-citations of journals
Bai et al. [27]	• Physics	• PPI model • Multi-feature model	• Inherent quality of scholarly paper • Scholarly paper impact decaying over time • Early citations • Early citers' impact
Beydokhti et al [28]	• Medicine	• Basic statistical methods	• Journals' impact factor • Level of evidence • Number of references • Number of authors • Number of title words • Length of article • Subject • Type of study design • Geographical area of corresponding author • Journal and publisher
Su [29]	• Physiology	• SVM • Decision Tree • Random Forest • Neural Network	• The sum of citing countries • The number of citing organizations • Total number of citing journals • The amount of citing subjects • The sum of citing languages • Average citation counts obtained • Average increment of citation counts obtained • The sum of funding organizations

Table A2. Estimated parameters for PR, NB, ZIP, ZINB, HP, and HNB Models.Significant factors (p<0.05) are highlighted bold

	PR		NB		ZIP				ZINB				HP				HNB			
	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Estimate/Pr(>|z|) Count Part		Estimate/Pr(>|z|) Binary Part		Estimate/Pr(>|z|) Count Part		Estimate/Pr(>|z|) Binary Part		Estimate/Pr(>|z|) Count Part		Estimate/Pr(>|z|) Binary Part		Estimate/Pr(>|z|) Count Part		Estimate/Pr(>|z|) Binary Part	
Int	**-1.639**	0.000	**-2.288**	0.000	**-1.144**	0.000	1.166	0.088	**-1.353**	0.000	**7.046**	0.000	**-1.195**	0.000	**-2.214**	0.000	**-2.136**	0.000	**-2.214**	0.000
Age	**0.248**	< 2e-16	**0.323**	< 2e-16	**0.189**	< 2e-16	**-0.388**	< 2e-16	**0.247**	< 2e-16	**-2.378**	0.000	**0.188**	< 2e-16	**0.409**	< 2e-16	**0.268**	< 2e-16	**0.409**	< 2e-16
Ref	**0.017**	< 2e-16	**0.018**	0.000	**0.015**	< 2e-16	**-0.019**	0.000	**0.017**	0.000	-0.030	0.079	**0.015**	< 2e-16	**0.022**	0.000	**0.016**	0.000	**0.022**	0.000
Ath	**0.009**	0.000	0.011	0.085	**0.004**	0.003	-0.034	0.132	0.007	0.257	-0.034	0.279	**0.004**	0.003	0.033	0.126	0.005	0.454	0.033	0.126
FCA	**-0.062**	< 2e-16	**0.089**	0.000	**-0.014**	0.000	0.007	0.608	-0.014	0.066	0.029	0.458	**-0.014**	0.000	-0.011	0.432	-0.016	0.085	-0.011	0.432
Pg	**-0.015**	0.000	**-0.019**	0.017	**1.982**	0.000	-0.869	0.271	**2.062**	0.000	-3.485	0.066	**2.039**	0.000	**1.782**	0.002	**2.458**	0.000	**1.782**	0.002
TL	**-0.002**	0.000	**-0.002**	0.034	0.000	0.292	**0.007**	0.000	**-0.002**	0.035	0.004	0.538	0.000	0.315	**-0.007**	0.000	0.000	0.981	**-0.007**	0.000
AI	**0.944**	0.001	**2.627**	0.000	**1.963**	0.000	0.673	0.500	**1.911**	0.000	-19.700	0.988	**2.011**	0.000	0.241	0.779	**3.010**	0.000	0.241	0.779
Bio	**1.164**	0.000	**1.167**	0.003	**0.951**	0.000	-0.595	0.469	**1.050**	0.010	-0.494	0.811	**1.013**	0.001	**1.250**	0.030	**1.253**	0.019	**1.250**	0.030
Chm	**1.291**	0.000	1.218	0.052	**1.724**	0.000	0.947	0.371	1.221	0.067	-0.387	0.876	**1.778**	0.000	-0.062	0.947	**1.907**	0.028	-0.062	0.947
Comp	**2.128**	< 2e-16	**2.266**	0.000	**1.982**	0.000	-0.869	0.271	**2.062**	0.000	-3.485	0.066	**2.039**	0.000	**1.782**	0.002	**2.458**	0.000	**1.782**	0.002
Eco	**1.214**	0.000	**1.202**	0.002	**1.213**	0.000	0.307	0.686	**1.043**	0.010	-2.134	0.401	**1.271**	0.000	0.513	0.345	**1.552**	0.004	0.513	0.345
Edu	**1.839**	0.000	**1.567**	0.000	**1.617**	0.000	-0.898	0.279	**1.314**	0.003	-3.297	0.201	**1.675**	0.000	**1.762**	0.005	**1.627**	0.004	**1.762**	0.005
Engy	**2.776**	< 2e-16	**2.721**	0.000	**2.551**	< 2e-16	-1.492	0.070	**2.402**	0.000	**-4.884**	0.024	**2.608**	< 2e-16	**2.415**	0.000	**2.828**	0.000	**2.415**	0.000
Eng	**2.360**	< 2e-16	**2.384**	0.000	**2.180**	< 2e-16	-1.153	0.093	**2.114**	0.000	**-4.367**	0.014	**2.238**	0.000	**2.075**	0.000	**2.535**	0.000	**2.075**	0.000
Env	**1.712**	0.000	**1.440**	0.000	**1.546**	0.000	-0.269	0.713	**1.316**	0.001	-3.480	0.086	**1.605**	0.000	**1.113**	0.028	**1.701**	0.001	**1.113**	0.028
Fuz	**2.071**	< 2e-16	**1.939**	0.000	**1.980**	0.000	-0.205	0.795	**1.829**	0.000	-2.113	0.251	**2.038**	0.000	1.128	0.055	**2.268**	0.000	1.128	0.055
Man	**0.727**	0.021	**1.178**	0.032	**0.940**	0.010	-0.483	0.623	0.827	0.138	-3.861	0.147	**0.921**	0.022	1.253	0.104	1.086	0.131	1.253	0.104
Math	**1.695**	0.000	**1.518**	0.000	**1.668**	0.000	-0.096	0.885	**1.334**	0.000	-3.362	0.054	**1.725**	0.000	**0.986**	0.016	**1.819**	0.000	**0.986**	0.016
Med	**1.972**	< 2e-16	**1.868**	0.000	**1.814**	0.000	-0.639	0.339	**1.676**	0.000	**-3.408**	0.045	**1.872**	0.000	**1.523**	0.000	**2.065**	0.000	**1.523**	0.000
OR	**2.751**	< 2e-16	**2.532**	0.000	**2.616**	< 2e-16	-0.504	0.600	**2.276**	0.000	-1.648	0.489	**2.672**	< 2e-16	1.429	0.077	**2.817**	0.000	1.429	0.077
Oth	**1.668**	0.000	**1.496**	0.001	**1.658**	0.000	0.353	0.664	**1.242**	0.004	-4.107	0.094	**1.716**	0.000	0.523	0.398	**1.871**	0.001	0.523	0.398
Phy	**1.460**	0.000	**1.131**	0.037	**1.247**	0.000	-0.419	0.665	1.062	0.060	-0.954	0.697	**1.313**	0.000	1.149	0.121	1.223	0.073	1.149	0.121
Soc	**1.274**	0.000	**1.232**	0.001	**1.164**	0.000	-0.287	0.700	**0.989**	0.012	-3.957	0.052	**1.219**	0.000	**1.058**	0.038	**1.365**	0.009	**1.058**	0.038
Stat	**1.563**	0.000	**1.397**	0.000	**1.563**	0.000	-0.009	0.989	**1.274**	0.000	-2.903	0.086	**1.621**	0.000	**0.879**	0.028	**1.727**	0.000	**0.879**	0.028