UNIVERSITÀ DI PISA

Sant'Anna
Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891910776

Analyzing the waiting time of academic publications: a survival model

Un modello di sopravvivenza per i tempi di accettazione delle pubblicazioni accademiche

Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi, Francesco Porro

Abstract In this paper a survival model is used to perform an analysis of the waiting time to publication for academic articles. The model is a multilevel excess hazard model and it allows to include non-linear and non-proportional effects of the covariates. The analysis is performed by considering covariates at two levels: the first one is the article level, the second one is the journal level.

Abstract *In questo articolo viene utilizzato un modello di sopravvivenza per effettuare un'analisi del tempo di attesa per la pubblicazione di articoli accademici. Il modello utilizzato é un modello multilivello con excess hazard che permette di includere effetti non lineari e non proporzionali delle covariate. L'analisi é condotta considerando covariate a due differenti livelli: articolo e rivista.*

Key words: Peer review, Waiting times, Net survival, Excess hazard model

Francesca De Battisti
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: francesca.debattisti@unimi.it

Giuseppe Gerardi
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: giuseppe.gerardi@unimi.it

Giancarlo Manzi
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: giancarlo.manzi@unimi.it

Francesco Porro
Dipartimento di Statistica e Metodi Quantitativi - Universitá degli Studi di Milano-Bicocca - piazza dell'Ateneo Nuovo, 1, 20126 MILANO e-mail: francesco.porro1@unimib.it

1 Introduction

The topic of waiting time in academic publication decisions is very relevant and interesting. The overall process of submission, especially for top-level journals, and any required revisions is such that many months, if not years, must pass between the submission and the acceptation of an article, if ever there will be one. De Battisti and Manzi [4] put forward some considerations on such issue, and suggested to apply multilevel models to find determinants affecting the waiting time until acceptance. The aim of this paper is to extend these considerations by carrying out a multilevel excess hazard model to analyze the waiting time to publication, working on the hierarchical data. We propose to consider the waiting time for academic publication as survival time, with article as units of interest, and model the effects of potential explanatory factors.

2 The methodology

Survival analysis typically considers the time until an event occurs. We usually refer to the time variable as *survival time*, because it measures the time that an individual has survived over a certain follow up period. We also usually refer to the event as a failure, because the event of interest usually is death, disease incidence, or any other negative individual experience. However, survival time may be for example the time to return to work after an elective surgical procedure, in which case failure is a positive event. As argued in [2], in the context of survival analysis, "the main survival indicator when comparing populations is net survival, that is, the hypothetical survival that patients would experience could they die only from their cancer ([3], [6])". In the general case, beyond the medical framework, the net survival refers to the occurrence of the event under study only because of specific causes. Moreover, the net survival is estimated comparing the all-cause hazard of death experienced by the patients to the general population from which the individuals come. One of the approaches proposed in literature to estimate net survival is modelling the excess hazard (see [2], [5], [7]). Starting from these proposals, in [2] a methodology to estimate an excess hazard regression model with non-linear and non-proportional effects due to the covariates and including a random effect is developed. The excess hazard approach is based on the assumption that the total hazard of the event occurrence, denoted by $\lambda(t,\mathbf{x},\mathbf{z})$, can be decomposed into the sum of a cause-specific hazard, denoted by $\lambda_+(t,\mathbf{x})$, and a hazard due to all the other causes, denoted by $\lambda_P(a+t,\mathbf{z})$ (the latter being usually estimated from general population life tables in the case of death). This means that:

$$\lambda(t,\mathbf{x},\mathbf{z}) = \lambda_+(t,\mathbf{x}) + \lambda_P(a+t,\mathbf{z}),$$

where $\mathbf{x}$ is a vector of prognostic variables, $\mathbf{z}$ is a vector of demographic characteristics, and a is the age at the diagnosis, so that $a+t$ denotes the age at death or at

censoring. The excess hazard is associated with the net survival through the classical relationship between hazard and survival function:

$$S(t) = \exp\left(-\int_0^t \lambda(v)dv\right).$$

Moreover, in [2] a multilevel excess hazard model is proposed. For each individual (or unit) j (with $j = 1,...,n_i$) from cluster (or group) i (with $i=1,...,D$), let t_{ij} denote the observed time-to-event and δ_{ij} be an indicator variable taking the value 1 in case of the event occurrence and 0 in case of censoring. Then the total hazard is:

$$\lambda(t,\mathbf{x}_{ij},\mathbf{z}_{ij},w_i) = \lambda_+(t,\mathbf{x}_{ij})\exp(w_i) + \lambda_P(a+t,\mathbf{z}_{ij}), \quad (1)$$

where w_i denotes a random effect at the cluster level. This model allows to introduce multiple covariate effects. For example, by expressing $\mathbf{x}_{ij}$ by its components, i.e. $\mathbf{x}_{ij} = (x_{1,ij}\, x_{2,ij}\, x_{3,ij})$, the following formula:

$$\log(\lambda_+(t,\mathbf{x}_{ij})) = \log(\lambda_0(t)) + \beta \cdot x_{1,ij} + f(x_{2,ij}) + g(t) \cdot x_{3,ij}$$

represents the logarithm excess hazard, with a linear and proportional effect related to $x_{1,ij}$, a non-linear and proportional effect due to a continuous function f of $x_{2,ij}$, and a non-proportional (time-depending) effect due to $x_{3,ij}$.

3 The dataset

The dataset used in this application is formed by 3489 published papers between 2011 and 2016 in the following top-level statistical journals: *Journal of Statistical Software* (J Stat Softw), *Fuzzy Sets and Systems* (Fussy Set Syst), the *Journal of the Royal Statistical Society - Series A - Statistics in Society* (JRSSA), the *Journal of the American Statistical Association* (JASA), the *Annals of Probability* (Ann Prob), the *Journal of Business Economic Statistics* (JBES), *Advanced in Data Analysis and Classification* (ADAC), and *Biostatistics*. The distribution of the articles is described in Table 1. We apply survival analysis on the waiting time of academic

Table 1 Distribution of articles across journals

Journal	*Publisher*	*Country*	*Eds' country*	*Number of Articles*
JRSSA	Royal Stat.Soc./Wiley	UK	UK	173 (4.96%)
ADAC	Springer	GER	ITA, GER, JAP	129 (3.70%)
JBES	Am.Stat.Ass./ Taylor&Francis	USA	USA	258 (7.39%)
Ann Prob	Inst.Math.Stats./ Bernoulli Society	USA	USA	493 (14.13%)
Biostatistics	Oxford Uni. Press	UK	NED, USA	330 (9.46%)
Fuzzy Set Syst	Elsevier	NED	BEL, FRA, GER, SPA	979 (28.06%)
JASA	Am.Stat.Ass./ Taylor&Francis	USA	USA	741 (21.24%)
J Stat Softw	UCLA Dept.Stats	USA	AUT, SWI, GER	386 (11.06%)

publications, namely the time that elapses between the submission and the publication of an article in a journal. More in details, for each article (uniquely identified by the Document Object Identifier - DOI) the waiting time (variable *Age*, in days) is calculated as the difference between the date of the acceptance when this was available (otherwise the date of the online publication or the date of the final revision) and the date of the submission (always available). The covariates considered in this application and their meanings are:

- *Bayes*: dichotomous variable (1=Bayesian article, 0=otherwise);
- *Month Scopus cit*: the average monthly number of Scopus citations per article;
- *Avg h Index*: the average *h* index of the authors;
- *Junior less 5*: dichotomous variable (1=if the Scopus *h* index of one of the authors is lower than 5, 0=otherwise);
- *Senior more 20*: dichotomous variable (1=if the Scopus *h* index of one of the authors is greater than 20, 0=otherwise);
- *Number author*: number of the authors;
- *USA all*: dichotomous variable (1=if all the authors are affiliated to US institutes, 0=otherwise);
- *USA*: dichotomous variable (1=if at least one author is affiliated to an US institute, 0=otherwise);
- *Same country*: dichotomous variable (1=if all the authors are affiliated to institutes in the same country, 0=otherwise);
- *Same nationality eds*: dichotomous variable (1=if the institutions of the most important author and of the journal editor have the same nationality, 0=otherwise);
- *Age journal*: age of the journal since its first issue;
- *IF*: the 2017 Thomson Reuters impact factor;
- *AI*: the AI index (Article Influence index) that measures the average influence of an article over the first five years after publication.

It is worth remarking that the last three variables refer to the journal level, that is they are second-level variables, meaning that they assume the same value for all the articles published in the same journal. The application of the survival analysis to this particular context requires some adjustments. First, our data refer only to published paper, not to all the submitted papers. For this reason, the event of interest (publication) occurs after a reasonable period of time for all our observed units. This is different from the usual case, in which individuals can be dead or alive at the end of the study period. In order to have a situation similar to the classical one in survival analysis models, we have to censor the data: we have to choose a time interval (which corresponds to the follow up period) to evaluate whether the article is published or not. In this way, it is possible to model the article waiting time. All the articles with a waiting time to publication (variable *Age*) greater than 3 years are censored. For these articles the variable δ (the indicator variable mentioned in Section 2) is set equal to 0. We selected the value of 1095 days (3 years) in order to have a restrained percentage (less than 5%) of censored articles. In this way, the censored articles represent the 4.39% of the total number of articles. The average of the waiting time (with censored data) is 451.49 days, the median is equal to 397.

4 Results and discussion

We applied the mixed effect excess hazard model described in (2) to the aforementioned dataset. We consider the 3489 articles as units, clustering them by the corresponding journal. In this work, following [2], we focus on the excess hazard function $\lambda_+(t,\mathbf{x}_{ij})\exp(w_i)$ of formula (2), and we assume a normal distribution of the random effect w, with zero mean and standard deviation σ.
The aim of this analysis is to identify which variables have a relevant impact on the waiting time to publication. We ran more than one hundred models, by setting different kinds of effect for each covariate. The calculations have been performed by using the R package *mexhaz* (Mixed Effect Excess Hazard Models). Such package provides estimates by MLE method, throught the implementation of numerical methods. We considered the following types of effects: linear and proportional, linear and non-proportional (that is, time-depending), non-linear and proportional, and non-linear and non-proportional. As suggested in [2], the logarithm of the baseline excess hazard and the functions (of time) for the time-dependent effects are modelled by cubic B-splines with two knots at 365 and 1094 days, respectively. For explicative purposes, in the following we report three models with a hierarchical level of complexity:

- Model 1: the effects of all the covariates are linear and proportional;
- Model 2: the effects of the covariates *Avg h Index*, *Number authors*, *IF*, *Age journal* and *AI* are linear and non-proportional. The effect of all the other covariates are linear and proportional;
- Model 3: the effects of the covariates *Avg h Index*, *Number authors*, *IF Age journal* and *AI* are non-linear and non-proportional.The effect of all the other covariates are linear and proportional.

The fittings of the three models are compared by using the Akaike Information Criterion (AIC), (see [1] for details). Table 2 reports the parameter estimates (and their standard errors) for the covariates with linear and proportional effects. For each model also the AIC index is reported. The model with the best fitting is Model 2, and, for this model, the highest Hazard Ratio (HR) is the one related to the covariate *Same nationality eds* and it is given by

$$HR_{Same\ nationality\ eds} = e^{0.064} = 1.066.$$

This value shows that if the nationality of the main author of an article is the same of the journal editor, the hazard rate is, *ceteris paribus*, higher than otherwise: this implies that in this case the waiting time to publication is smaller. Similarly, all other conditions being equal, articles with all US authors present a smaller waiting time than the others, since the corresponding HR is equal to

$$HR_{USA\ all} = e^{0.03} = 1.03.$$

Conversely, the model shows that articles with at least one author affiliated to an US institution, experience a higher waiting time than those with no author affilied to an

Table 2 Parameter estimates and the corresponding AIC for the three models

Variable	Model 1 AIC=45237.52	Model 2 AIC=45002.07	Model 3 AIC=45266.02
Bayes	-0.010 (0.060)	0.002 (0.060)	-0.011 (0.060)
Month Scopus cit	0.005 (0.005)	0.007 (0.005)	0.008 (0.005)
Avg h Index	0.004 (0.002)	LIN-NPH	NLIN-NPH
Junior less 5	0.010 (0.041)	-0.008 (0.041)	0.024(0.044)
Senior more 20	0.012 (0.049)	0.019 (0.050)	0.013 (0.058)
Number authors	-0.042 (0.017)	LIN-NPH	NLIN-NPH
USA all	0.028 (0.076)	0.030 (0.076)	-0.005 (0.076)
USA	-0.067 (0.061)	-0.064 (0.061)	-0.037 (0.060)
Same country	-0.044 (0.051)	-0.045 (0.051)	-0.055 (0.051)
Same nationality eds	0.077 (0.046)	0.064 (0.046)	0.039 (0.046)
Age journal	0.091 (0.012)	LIN-NPH	NLIN-NPH
IF	-0.063 (0.022)	LIN-NPH	NLIN-NPH
AI	-0.277 (0.051)	LIN-NPH	NLIN-NPH
Standard deviation σ	1.698	1.723	2.758

US institution, everything else being equal, as the HR for the variable *USA* is

$$HR_{USA} = e^{-0.064} = 0.938,$$

therefore teams with all US members seem to be more successful than mixed ones.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974)
2. Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A.: A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. Statistics in Medicine **35**, 3066–3084 (2016)
3. Danieli, C., Remontet, L., Bossard, N., Roche, I., Belot, A.: Estimating net survival: the importance of allowing for informative censoring. Statistics in Medicine **31**(8), 775–786 (2012)
4. De Battisti, F., Manzi, G.: Smart Tools for Academic Submission Decisions: Waiting Times Modeling. in Smart Statistics for Smart Applications Book of Short Papers SIS 2019, [edited by] Arbia, G., Peluso, S., Pini, A., Rivellini, G., Ed. Pearson, June 2019, 787–792 (2019)
5. Estve, J., Benhamou E, Croasdale M, Raymond L.: Relative survival and estimation of net survival: elements for further discussion. Statistics in Medicine **9**(5), 526–538 (1990)
6. Perme, MP., Stare, J., Estve, J.: On estimation in relative survival. Biometrics **68**(1), 113–120 (2012)
7. Remontet, L., Bossard, N., Belot, A., Estve, J., and the French network of cancer registries FRANCIM: An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Statistics in Medicine **26**, 2214-2228 (2007)