Accelerated failure time models in analyzing duration of employment

L Basha* and E Gjika

Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania
*E-mail: lule.hallaci@fshn.edu.al

Abstract. Parametric accelerated failure time models (AFT), that assess the relationship between the event times and dependent variables constitute an essential class of regression models. In this research we have used the accelerated failure time models: in particular, we have taken into consideration exponential distribution, Weibull distribution, log-logistic model and lognormal distribution, to analyze the duration of employment for employees that have been in their current job or otherwise known as job tenure. A variety of commercial and public companies in Albania were surveyed for this study. Initially the goal of the research is to find the optimal probability distribution to suit the data. Then, the accelerated failure time model is used to assess the impact of the predictors, such as: the employee's age, wages, the employee's age when he has started the job, sex, profession, academic degree, marital status and years of experience prior to this position, in job termination. The log-normal AFT model was the most accurate model for predicting the outcomes of the study, and from this model we can conclude that the employee's age, the employee's age when he has started the job, wages, academic degree and position in the company can affect how long someone stays employed.

1. Introduction

The most common type of model used to predict the duration of socio-economic phenomena is the survival model [1]. There are many researches into the matter of duration of employment and unemployment using survival models [2], [3]. In [4] authors focus on job tenure as a measure of whether a job is a good match for someone after unemployment. The relationship between career flexibility and professional plateau is examined, with focus on how long employees have been employed and their own self-awareness is examined in [5]. Job duration is of uppermost importance to workers, as it can be interpreted as a measure of job-for-life. The median length of time that employees have been with their current employer was 4.2 years in January 2018, according to the Bureau of Labor Statistics. [6] adopt the semi-parametric Cox proportional hazards model, in and [7] determine whether work ability played a decisive role in keeping a job among hospital staff in São Paulo, Brazil, who eventually were laid off within 4 years. By using multiple-record Cox proportional hazards model [8] examine risk variables for health-related job loss over a two-year follow-up period. In this research, we have surveyed of the feature of 887 employees, in a variety of commercial and public companies, in Albania. In the first stage of the study, we analyze the duration of job tenure and find the parametric probability distribution that best fits the data. Given that the time is a continuous variable we have taken into consideration some continuous distributions such as: lognormal; Weibull; exponential and log-logistic distribution which is mostly used in the literature. Then, variables, which are thought to have a great impact of length of employment are studied. The factors studied are: the employee's age, wages, the employee's age when he has started the job, sex, profession, academic
degree, marital status and years of experience prior to this position. The accelerated failure time (AFT) model, is used to estimate how these factors may effect on job tenure. The procedure of parameter estimation, evaluation and covariate selection was done through R software packages.

2. Materials and methods

When the length of time or survival time, is assumed to follow a known distribution, then we can use a parametric survival model for the data. In the case of parametric models, full maximum likelihood method can be used to estimate parameters and the probability distribution of the variable can be estimated. The probability distributions we have considered are continuous distributions: lognormal; Weibull; exponential and log-logistic distribution. In survival analysis models, explanatory variables are considered to be related with the time it takes for the event to occur. To explore this relationship, researchers proposed the Cox Proportional Hazards model with right censoring [9]. A useful alternative to the Cox PH model is the accelerated failure time (AFT) model [10], [11]. This model supposes that the error term of this linear regression has a specific distribution. This model has the advantage of an intuitive physical interpretation [12] and gives more efficient estimates compare with Cox model [13].

We have denotes with \(X_1, \ldots, X_n\) the variable of interest with probability density function \(f\) and cumulative distribution function \(F\). The censoring variable is \(C_1, \ldots, C_n\) with continuous distribution function \(G\). We suppose that the variable \(X\) is independent of \(C\). When taking into account the random right censoring, the variable is not fully observed. We can only notice \((T_i, \delta_i)\) where \(T_i = \min(X_i, C_i)\) and \(\delta_i = I(X_i \leq C_i)\), for \(i = 1, \ldots, n\).

The relationship of logarithm of survival time variable and dependent variables, is linear according to the accelerated failure time model, as

\[
\log T = a_0 + \sum_{i=1}^{p} a_i z_i + \sigma \varepsilon
\]  

(1)

where, \(z = (z_1, \ldots, z_p)\) are the covariates, \(a = (a_1, \ldots, a_p)\) are the coefficients, \(\sigma (\sigma > 0)\) is an unknown scale parameter, and \(\varepsilon\) are random errors that does not depend on \(z_i\).

The AFT model presume that factors interrelate with time directly in the baseline hazard function. The hazard function at time \(t\) can be given as:

\[
h(t \mid z) = h_0(t \cdot e^{-z^T a}) \cdot e^{-z^T a}
\]  

(2)

where \(h_0(t)\) is the baseline hazard function. The influence of covariates on hazard in equation (2) depends on time and at the same time is disproportionate. A negative sign on a coefficient for a covariate indicates that it reduces the hazard value, but may also increase the duration of the exposure when the covariate increases [14]. The survivorship, the probability density, the hazard and cumulative hazard function are mathematically equivalent and all the relevant formulas are summarized as below:

\[
H(t \mid z) = \int_0^t h(T \mid z) dT = \int_0^t h_0(T \cdot e^{-z^T a})e^{-z^T a} dT = H_0(t \cdot e^{-z^T a}) = -\ln \left[ 1 - F_0(t \cdot e^{-z^T a}) \right]
\]

\[
S(t \mid z) = \exp[-H(t \mid z)] = \exp[-H_0(t \cdot e^{-z^T a})] = S_0(t \cdot e^{-z^T a}) = 1 - F_0(t \cdot e^{-z^T a})
\]  

(3)

\[
f(t \mid z) = h(t \mid z) \cdot S(t \mid z) = h_0(t \cdot e^{-z^T a}) \cdot e^{-z^T a} \cdot S_0(t \cdot e^{-z^T a}) = f_0(t \cdot e^{-z^T a}) \cdot e^{-z^T a}
\]

The general Weibull regression model is \(\log T = a_0 + a_1 z_1 + \ldots + a_p z_p + \sigma X\), where \(X\) has standard extreme value distribution: \(f_X(x) = \exp(x - e^x)\). The survivor function can be written as:
$S(t; z_1, ..., z_p) = \exp\left\{ -\exp[\alpha(\log(t) - a_0 - a_1z_1 - ... - a_pz_p)] \right\}$ \hspace{1cm} (4)

The log-logistic and lognormal regression models are examples of accelerated failure time models as well as the exponential model, which is a special case of Weibull distribution, for $\alpha = 1$. If the event time $T$ has lognormal distribution, we model its dependency from explanatory variables as equation (3), where the variable has standard normal distribution: $X \sim N(0,1)$. The survivor function has the form

$$S(t; z_1, ..., z_p) = 1 - \Phi\left( \frac{1}{\sigma}(\log(t) - a_0 - a_1z_1 - ... - a_pz_p) \right)$$ \hspace{1cm} (5)

If the event time $T$ has log-logistic distribution, we model its dependency from explanatory variables as equation (3), where $X$ has standard logistic distribution: $f(x) = \frac{e^x}{(1+e^x)^2}$. The survivor function has the form:

$$S(t; z_1, ..., z_p) = \frac{1}{1 + \exp[\alpha(\log(t) - a_0 - a_1z_1 - ... - a_pz_p)]}$$ \hspace{1cm} (6)

3. Study population

In our research, we have looked at how long 887 employees, have been employed in their current job, in a variety of commercial and public companies, in Albania. This study looks at the effect of a given period of time, from January 1991 to December 2018. The time when the employee begins working for the employer, is considered as the start time, while the end time is the time when financial relations are interrupted. At the time the data was received from the company some employees are currently working, while some are no longer employees of the current company. For this purpose, we have established a censoring variable. This variable has a value of 0 for the "still working" variable has a value of 0 if the data was last updated within the past week and a value of 1 otherwise. "still working" variable and a value of 1 otherwise. We have tried to fit a probability model for duration of job tenure. Given that the time is a continuous variable, in our case in months, we have taken into consideration some continuous distributions such as: lognormal; gamma; Weibull; exponential and log-logistic distribution. Some factors that are thought to be important in the time of staying employed with the same employer and the same job position are studied. The accelerated failure time (AFT) model, is used to estimate how these factors may effect on job tenure.

There are in total 887 employees, 534 of them were still in financial relationship with the company, in December 2018, the last month when the data has been update. The other part of the employees has had their employment relationship with the company interrupted. The shortest time that an employee has been in a current job position is one month and the longest time in the same job position is 26 years. In the first stage of this work we have proposed four probability distributions to fit the job tenure time: the lognormal, exponential, Weibull and log-logistic distribution. To estimate and evaluate the fitting we have used R software as a tool. To compare the fitting performance of the probability distributions we have used QQ plot graphical test and the information criteria statistics (AIC and BIC).
Figure 1. Goodness-of-fit plots for various distributions fitted to job tenure

Plots of empirical quantiles of residuals versus theoretical residuals expected by error distribution (in our case, Weibull, exponential, lognormal and log-logistic distribution) are shown in Figure 1. Q-Q plot show a straight line for lognormal distribution, which describe satisfactory the most of the empirical data, and the other proposed distributions such as the exponential, Weibull and log-logistic distribution seem too not describe the real observations.

Table 1. Goodness of fit statistics

|            | Exponential | Weibull | Lognormal | Log-logistic |
|------------|-------------|---------|-----------|--------------|
| AIC        | 5059        | 4515    | 4331      | 4351         |
| BIC        | 5064        | 4521    | 4336      | 4358         |

As it is clearly seen from Table 1, also the goodness-of-fit statistics for the job tenure time are in favor of the lognormal distribution, with parameters: meanlog: 5.2; sdlog: 2.44. From a given set of candidate models, we take the one that minimizes the AIC (BIC).

The Kaplan-Meier estimator [15] is a non-parametric maximum likelihood estimator for survival function. This estimator does not make an assumption on the distribution of $T$, but from the curve we can see which parametric distribution is appropriate for the data.

Figure 2 shows the estimated survival function using Kaplan-Meier method and various distributions fitted to job tenure, without taking into consideration the covariates. Also, we can see that exponential curve is very different from the Kaplan-Meier curve and again here we can confirm that lognormal distribution “best” fits the data.

In the next step we estimate the relationship between job tenure and all the covariates mention above. A stepwise regression method is conducted, to select the explanatory variables. A forward-selection
rule starts with no explanatory variables and then adds variables, one by one, based on which variable is the most statistically significant, until there are no remaining statistically significant variables. The selection procedure has been done with the principle of minimizing the indicators of goodness-of-fit, in this case we have used the criteria of minimizing Akaike Information Criteria (AIC). This procedure is done using accelerated failure time (AFT), exponential, Weibull, lognormal and log-logistic models. The statistical significance of the explanatory variables was observed at the 5% level after the procedure was completed. The final estimated model yielded the following results, given in Table 2. The first column of the table shows the parameters estimation, with the log-likelihood method. A log likelihood test indicates that the model on the whole is significantly better than one which does not include any covariates. According to the AFT model, factors: the employee's age, wages, the employee's age when he has started the job, sex, profession, and years of experience prior to this position in the company are statistically significant with a p-value less than 0.05. The marital status and academic degree are eliminated from the model based on AIC. The parameter estimates from AFT models are interpreted as effects on the time scale, which can either speed up or slow down the survival time. If the acceleration factor is greater than one, then the exposure is beneficial to survival. Male significantly increases the survival time by approximately 30%. One unit change in the employee's age when he has started the job at the current position in the company shortens survival time by 16%. The results suggest that the median survival time is decreased by a factor of 2.9 if the job position of an employee is a supervisory, compare with a financier. Profession, sex and the employee's age extends the duration of financial relationship.

|   | Value  | Std. Error | z     | p       |
|---|--------|------------|-------|---------|
| (Intercept) | 4.45   | 2.55e-01   | 17.48 | < 2e-16 |
| age        | 0.138  | 8.95e-03   | 15.44 | < 2e-16 |
| agebi      | -0.165 | 1.04e-02   | -15.78| < 2e-16 |
| ywcp       | 4.14e-03| 5.94e-04   | 6.97  | 3.3e-12 |
| salary     | -1.16e-05| 1.68e-06   | -6.87 | 6.3e-12 |
| genderM    | 0.265  | 1.20e-01   | 2.20  | 0.027   |
| posfinancier | 0.0337 | 1.53e-01   | 0.22  | 0.826   |
| posother   | 1.03   | 1.54e-01   | 6.73  | 1.7e-11 |
| posspecialist | 1.14   | 1.47e-01   | 7.80  | 6.3e-15 |
| possupervisory | 1.59e+01| 1.67e+03   | 0.01  | 0.992   |
| Log(scale) | -0.256 | 5.57e-02   | -4.59 | 4.5e-06 |

Factors in Final Model
[1] age _ agebi _ ywcp _ salary _ gender _ pos

| Deleted | Chi-Sq | d.f. | P    | AIC  |
|---------|--------|------|------|------|
| status  | 1.16   | 1    | 0.2806 | -0.84 |
| educ    | 4.70   | 2    | 0.0951 | -0.13 |

When using a model to fit survival data, the survival curves can be adjusted to account for all of the covariates used as predictors. The Kaplan-Meier method is unsuitable for this problem because it is not adjusting for covariates. Figure 3 shows adjusted survival curves for job tenure, where the average value chosen for a covariate being adjusted is used.
Figure 3. Survival curves adjusted for covariates.

The probability that an employee will continue to work with the current employer or company for at least one year, is 96.2%. The lower quartile of time is 71 months; this means that, nearly one-quarter of employees leave their company within 71 months from their start date, while half leave within 114 months and 75% of them interrupt their relationship with the company within 160 months from their beginning.

4. Conclusion
The objective of this study was to analyse the length of time the employee has worked in his or her current position or with the current employer, to find the model that best fits the data and to assess the impact of the employee's age, wages, the employee's age when he has started the job, sex, profession, academic degree, marital status and years of experience prior to this position, in job termination. As a result of preliminary analysis on how to model the phenomenon in question, four models were proposed: exponential, Weibull, lognormal and log-logistic. Based on graphical and statistical tests, the lognormal distribution was found to be the best fit for the empirical data, with parameters: meanlog: 5.2; sdlog: 2.44.

The results showed that the average work experience of an employee in a company is approximately 187 months, 25% of employees have been in their current position for more than 10 years, and 30% for less than 2 years. According to the AFT analysis, profession, sex and age extends the time of financial relationship, while the employee's age when he has started the job and wages affect positively in the interruption of the relationship, thus in the job tenure. Also, half of the employees leave within 114 months and 75% of them interrupt their relationship with the company within 160 months from their beginning.

5. References
[1] Lawless J L 2003 Statistical Models and Methods for Lifetime Data (New Jersey: Wiley Interscience, Hoboken)
[2] LancasterT 1979 Econometric methods for the duration of unemployment. Econometrica vol. 47, no. 4 pp 939–956.
[3] Landmesser J 2013 The use of methods of analysis of the duration of the labour force survey in Poland, Warsaw University of Life Sciences, Warsaw.
[4] Kohara M, Sasaki M and Machikita T 2013 Is longer unemployment rewarded with longer job tenure? Journal of the Japanese and International Economies Volume 29 Pages 44-56 https://doi.org/10.1016/j.jjie.2013.06.002
[5] Jiang Z, Hu X and Wang Z 2018 Career adaptability and plateaus: The moderating effects of tenure and job self-efficacy Journal of Vocational Behavior Volume 104 Pages 59-71 https://doi.org/10.1016/j.jvb.2017.10.006
[6] Basha L, Gjika E and Puka Ll 2020 Analyzing Job Tenure Factors in Private and Public Companies in Albania by Using Cox Proportional Hazards Model Advances in Science, Technology and Engineering Systems Journal vol. 5 no. 2 pp. 254-260 https://dx.doi.org/10.25046/aj050233 DOI: 10.25046/aj050233
[7] Martinez M C and Fischer F M 2019 Work Ability and Job Survival: Four-Year Follow-Up. *International journal of environmental research and public health, 16*(17), 3143. https://doi.org/10.3390/ijerph16173143

[8] Syddall H E, D'Angelo S, Ntani G, Stevens M, Harris E C, Linaker C H and Walker-Bone K 2020 Work participation and risk factors for health-related job loss among older workers in the Health and Employment after Fifty (HEAF) study: Evidence from a 2-year follow-up period. *PLoS One*, 15(9), e0239383. https://doi.org/10.1371/journal.pone.0239383

[9] Cox D 1972 Regression models and life – tables (with discussion) *Journal of the Royal Statistical Society* Series B 34 187–220.

[10] Kalbfleish J D and Prentice R L 2002 *The Statistical Analysis of Failure Time Data* 2nd ed. (New York: Wiley MR1924807)

[11] Yu X and Zhao Y 2018 Jackknife empirical likelihood inference for the accelerated failure time model. *SPRINGER* https://doi.org/10.1007/s11749-018-0601-7

[12] Wei L J 1992 The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879.

[13] Cox D and Oakes D 1984 *Analysis of Survival Data* (London: Chapman and Hall)

[14] Wang K, Ye X and Jie M 2018 An empirical analysis of post-work grocery shopping activity duration using modified accelerated failure time model to differentiate time-dependent and time-independent covariates. *PLoS ONE*, 13(11): e0207810. https://doi.org/10.1371/journal.pone.0207810

[15] Kaplan E and Meier P 1958 Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol. 53 pp. 457–481