Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
4(330) 2017

DOI: http://dx.doi.org/10.18778/0208-6018.330.02

Wioletta Grzenda
Warsaw School of Economics, Institute of Statistics and Demography, Event History and Multilevel Analysis Unit, wgrzend@sgh.waw.pl

# Modelling the Duration of the First Job Using Bayesian Accelerated Failure Time Models[1]

**Abstract:** In this paper, the duration of the first job of young people aged 18–30 has been analyzed. The aim of the work is to find the distribution which best describes the investigated phenomenon. Bayesian accelerated failure time models have been used for modelling. The use of the Bayesian approach made it possible to extend past research. More precisely, prior information could be included in the study, which let us compare distributions of model parameters. Moreover, the comparison of explanatory power of competing models based on the Bayesian theory was possible. The duration of the first job for men and women was also compared using the abovementioned methods.

**Keywords:** parametric survival models, AFT models, the Bayesian approach, MCMC, employment

**JEL:** J630

[1] This study has been prepared as part of the project granted by the National Science Centre, Poland entitled “The modeling of parallel family and occupational careers with Bayesian methods” (2015/17/B/HS4/02064).

# 1. Introduction

There are many studies that investigate the duration of employment and unemployment in the classical approach (Lancaster, 1979; Drobnič, Frątczak, 2001; Landmesser, 2013). The Bayesian approach proposed in this paper is an attempt to expand existing knowledge in this field. The main advantage of the Bayesian approach (Gelman et al., 2000; Bolstad, 2007) is the ability to take into account additional information from outside the sample. Depending on the sample size – and thus the amount of information it provides – the effect of prior knowledge on the received posterior distribution may vary (Ibrahim et al., 2001; Kim, Ibrahim, 2000; Grzenda, 2013). In addition, in the Bayesian approach model parameters are random variables. Therefore, the proposed approach makes the comparison of their distributions possible. The Bayesian theory provides also other opportunities in terms of analyzing how well the model fits the data as well as comparing explanatory power of competing models (Spiegelhalter et al., 2002).

Survival models are most frequently used to model the duration of socio-economic phenomena (Lawless, 2003; Lee, Wang, 2003). If the analytical form of probability distribution function or survival function is known, parametric models are used. Accelerated failure time models (AFT) are among the latter (Kalbfleisch, Prentice, 2002; Wei, 1992). These models belong to regression models, in which logarithmic transformation of duration serves as a dependent variable. Moreover, it is assumed that exogenous variables have a linear impact on the logarithm of duration in these models. The AFT models are an alternative to the Cox models, which are used to analyze the impact of explanatory variables on the hazard function (Cox, 1972). In this paper, the parametric accelerated failure time models based on Bayesian approach are considered (Walker, Mallick, 1999).

The selection of a suitable model from many competing ones is a key to econometric modelling. In the case of models estimated using the maximum likelihood method, the tests based on the logarithm likelihood function play an important role. The most popular test is the likelihood ratio test (Wilks, 1935; 1938); however, it has significant limitations. This test can be used only in the case of nested models. Information criteria are more universal; the most popular is the Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Raftery, 1996). In the Bayesian approach, Bayesian versions of these criteria are also considered. These are expected Akaike information criterion (EAIC) and the expected Bayesian information criterion (EBIC) (Spiegelhalter et al., 2002). These criteria are based on the expected value of posterior distributions, not on the maximum of the likelihood function.

Another generalization of the AIC criterion in the Bayesian theory is the deviance information criterion (DIC) (Spiegelhalter et al., 2002; Congdon, 2006; Ando, 2010). This criterion is based on the posterior mean of the deviance. The approach proposed in this study enables the comparison of the explanatory power of competing models in the context of Bayesian theory. The Bayesian comparison of competing models has its origins in the Bayesian statistical hypothesis testing, which is based on the posterior odds ratio (Jeffreys, 1961). Model comparison is based on determining posterior probabilities for each model (Marzec, 2008; Osiewalski, 2001). The basic measure for these methods is the Bayesian factor (Kass, Raftery, 1995), which is defined as the ratio of the two marginal density functions of the observation vector.

In this study, accelerated failure time models (AFT) are used to analyze the duration of the first job among young people. Subjects aged 18–30 at the time of the study were investigated. Recently a much higher unemployment rate has been observed for young people than for other age groups (Central Statistical Office, 2014). Many studies also indicate that the situation of women on the labour market is worse than that of men (Drobnič, Frątczak, 2001; Landmesser, 2013). The social roles traditionally assigned to women limit their opportunities to pursue a professional career. This is particularly true among young women who often start their career and family at the same time. In this study, the duration of the first job for men and women was compared using Bayesian accelerated failure time models.

## 2. A Bayesian approach to the parametric survival models

The parametric accelerated failure time models (AFT) describe the relationship between survival functions for two units. Let $S_i(t)$ denote survival function for the $i$-th unit, and $S_j(t)$ for the $j$-th unit, then for AFT models the equality holds (Allison, 2010):

$$S_i(t) = S_j(\gamma t), \tag{1}$$

for all $t$ and $\gamma > 0$, where $\gamma$ is called an acceleration factor. The dependent variable in the survival model is the duration of the phenomenon $T$. The acceleration factor enables the assessment of the impact of exogenous variables on the survival time. Besides the survival function there are other ways to describe the distribution of $T$ such as probability density function ($f_i(t)$) and hazard function ($h_i(t)$).

The AFT models belong to a broad class of regression models, therefore, the AFT model can be written in the following form:

$$Y = \ln(T) = \mathbf{x}'\boldsymbol{\beta} + \sigma\epsilon. \tag{2}$$

Hence,

$$T = \exp(\mathbf{x'\beta}) \cdot \exp(\sigma\epsilon), \tag{3}$$

where **x** is the vector of explanatory variables, **β** is the regression vector, σ is the scaling factor and ϵ a random component. For various distributions of the random component ϵ, different distributions for event time *T* are obtained.

Exponential model, Weibull model, gamma model, log-logistic model and log-normal model are the most often investigated AFT models. The exponential model is the AFT model with σ = 1. Hence, it takes the form:

$$\ln(T) = \mathbf{x'\beta} + \epsilon, \tag{4}$$

where the random component ϵ has a Gumbel distribution with density function:

$$f(\epsilon) = \exp(\epsilon - \exp(\epsilon)) \text{ for } \epsilon \in \mathbf{R}. \tag{5}$$

The Weibull model is the AFT model with σ ≠ 1,

$$\ln(T) = \mathbf{x'\beta} + \sigma\epsilon, \tag{6}$$

where the random component ϵ also has a Gumbel distribution.

The gamma model is obtained when the random component ϵ in the AFT model has a log-gamma distribution (Lee, Wang, 2003):

$$f(\epsilon) = \begin{cases} \dfrac{|\delta|}{\Gamma(1/\delta^2)}[\exp(\delta\epsilon)/\delta^2]^{1/\delta^2}\exp[-\exp(\delta\epsilon)/\delta^2] & \delta \neq 0, \\ \left(1/\sqrt{2\pi}\right)\exp(-\epsilon/2) & \delta = 0. \end{cases} \tag{7}$$

The log-normal model is obtained when the random component ϵ in the AFT model has a normal distribution:

$$\epsilon \sim N(0, 1). \tag{8}$$

The log-logistic model is obtained when random component ϵ in the AFT model has a logistic distribution:

$$f(\epsilon) = \exp(\epsilon)/\left(1 + \exp(\epsilon)\right)^2 \text{ for } \epsilon \in \mathbf{R}. \tag{9}$$

The AFT models in the classical approach are estimated based on the likelihood function. Let $\mathbf{t} = (t_1, t_2, \ldots, t_n)'$ be the vector of survival times that are independent and identically distributed. Let $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$ denote the censor variables vector, where $\nu_i = 0$, if $t_i$ is right-censored and $\nu_i = 1$, if $t_i$ is the failure time for $i = 1, 2, \ldots, n$. Let $D = (n, \mathbf{t}, \mathbf{X}, \boldsymbol{\nu})$ denote the observed data, where $\mathbf{X}$ $(n \times p)$ is a matrix of independent random variables. Then the formula for the likelihood function takes the following form:

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} [f_i(t_i|\boldsymbol{\theta})]^{\nu_i} [S_i(t_i|\boldsymbol{\theta})]^{1-\nu_i}, \tag{10}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ is the vector of unknown parameters.

The likelihood function is also the basis for statistical inference in the Bayesian approach. However, in the Bayesian approach the inference about any element of the vector of parameters is based on the posterior distribution, calculated using Bayesian theorem (Gelman et al., 2000; Bolstad, 2007). In addition, the Bayesian approach requires prior distributions containing initial knowledge on the vector of parameters. Then the posterior distributions are calculated using the formula:

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}, \tag{11}$$

where $p(\boldsymbol{\theta})$ is the a priori joint probability distribution, and $p(D)$ the marginal density of observation.

Using the normality constant, the above formula can be written as:

$$p(\boldsymbol{\theta}|D) \propto L(\boldsymbol{\theta}|D)p(\boldsymbol{\theta}). \tag{12}$$

In our study, the Markov Chain Monte Carlo Method (MCMC) was used to estimate the vector of parameters (Congdon, 2006).

## 3. The model selection

There are many criteria (Ando, 2010) of the comparison of models and thus selecting the best one, depending on the modelling method. What is usually being verified is how well the model fits the data and what predictive power it has.

Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Raftery, 1996) are calculated in order to assess the fit of the models based on the likelihood function. These criteria are defined through the following formulas:

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + 2p, \tag{13}$$

$$BIC = -2\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + p\log(n), \tag{14}$$

where $\ell(\widehat{\boldsymbol{\theta}}|\mathrm{x})$ denotes the maximum of the logarithm of the likelihood function, $p$ the number of model parameters and $n$ the number of observations. In both formulas, the first component of the sum describes the model fit and the second component determines the degree of complexity of the model. The criteria differ in how they assess the simplicity of the model, which is especially important in the case of models estimated on small samples.

In Bayesian approach, criteria such as the expected Akaike information criterion (EAIC) and the expected Bayesian information criterion (EBIC) are the generalization of AIC and BIC criteria being so commonly used in the classical approach (Spiegelhalter et al., 2002). They are defined as follows:

$$EAIC = \overline{D(\boldsymbol{\theta})} + 2p, \tag{15}$$

$$EBIC = \overline{D(\boldsymbol{\theta})} + p\log(n), \tag{16}$$

where $p$ is the number of model parameters and $n$ is the number of observations.

In line with the Bayesian approach, the deviance information criterion (DIC) is used to evaluate how well the model fits the data (Spiegelhalter et al., 2002; Congdon, 2006). Let $p(\mathbf{y}|\boldsymbol{\theta})$ be the joint probability distribution of data $\mathbf{y}$ and the vector of parameters $\boldsymbol{\theta}$. Then, the Bayesian deviance is defined by (Gill, 2008):

$$D(\boldsymbol{\theta}) = -2\log[p(\mathbf{y}|\boldsymbol{\theta})] + 2\log[f(\mathbf{y})], \tag{17}$$

where $f(\mathbf{y})$ is a function of data, usually equal to 1. Then the posterior expectation of the deviance, which is the measure of Bayesian model fit, is given by:

$$\overline{D(\boldsymbol{\theta})} = \mathrm{E}_{\theta}\,[-2\log[p(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}] + 2\log[f(\mathbf{y})], \tag{18}$$

However, the effective number of parameters in the model is calculated by the formula:

$$\rho_D = \overline{D(\boldsymbol{\theta})} - D(\widetilde{\boldsymbol{\theta}}), \tag{19}$$

where $\widetilde{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y})$. Then the deviance information criterion (DIC) is given by:

$$DIC = \overline{D(\boldsymbol{\theta})} + \rho_D = 2\overline{D(\boldsymbol{\theta})} - D(\widetilde{\boldsymbol{\theta}}). \tag{20}$$

Another measure used to compare competing models in the Bayesian approach is the Bayes factor. In Bayesian statistics, model selection is equivalent to testing relevant statistical hypotheses. Let $D$ be the observed data. Let $H_1$ and $H_2$ be two mutually exclusive and complementary hypotheses, and $P(H_1)$ i $P(H_2) = 1 - P(H_1)$ be their prior probability. Then, based on Bayes' theorem, we have:

$$P(H_k|D) = \frac{P(D|H_k)P(H_k)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}, \tag{21}$$

for $k = 1, 2$. Then the odds ratio of posterior hypothesis $H_1$ with respect to the hypothesis $H_2$ is calculated according to the formula:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)}\frac{P(H_1)}{P(H_2)}, \tag{22}$$

where $P(H_1)/P(H_2)$ means the prior odds ratio, and the quotient $P(D|H_1)/P(D|H_2)$ is called the Bayes factor. With this factor it can be established how many times the hypothesis $H_1$ is more posterior likely than the hypothesis $H_2$.

In the terminology of the two competing models $M_1$ and $M_2$, the Bayes factor is given by:

$$BF = \frac{P(D|M_1)}{P(D|M_2)}. \tag{23}$$

In the Bayesian approach:

$$P(D|M_k) = \int P(\boldsymbol{\theta}_k|M_k)P(D|\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k, \tag{24}$$

for $k = 1, 2$. Given that the model is a fixed value we have:

$$P(D) = \int P(\boldsymbol{\theta})P(D|\boldsymbol{\theta})d\boldsymbol{\theta} \tag{25}$$

In order to approximate the marginal density of the observation vector ($D$), the estimate of Newton and Raftery is most commonly used (Newton, Raftery, 1994):

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{p(\mathbf{y}|\boldsymbol{\theta}^{(i)})}\right]^{-1}, \tag{26}$$

where $\boldsymbol{\theta}^{(i)}$ is a sample from the posterior distribution $p(\mathbf{y}|\boldsymbol{\theta}^{(i)})$.

## 4. The scope of research

For the purpose of this study, a data set from the panel survey Generations and Gender Survey (GGS) for Poland, conducted under the Generations and Gender Programme (GGP) was used. The collected data dates back to the second half of 2014. In addition, the missing information, which did not changed over time, was supplemented based on the previous round of research carried out over the years 2010–2011. This study was conducted among a random sample of respondents aged 18–79.

From the complete set of data, individuals who were aged 18–30 at the time of the study were singled out, which resulted in 1.210 observations. Only jobs undertaken after the age of 15 were taken into consideration. The analyzed feature is the duration of the first job in months (time). In our sample, the value of this feature ranges from 1 month to 165 months, with the average of 25.93. At the time of the study 540 respondents were still working, while 670 lost their first job. Based on this information a censoring variable (censor) was established for the purpose of modelling. This variable takes on the value 0 if the respondent still worked in their first job at the time of the study, and 1 otherwise. The available data do not provide information on the status of respondents who lost their first job. It is possible that they became unemployed, moved to another job or decided to continue learning as the study focuses on young people.

The study also aims to compare the situation of women and men on the labour market. In the studied sample there were 47.85% of men (sex = 1) and 52.15% of women (sex = 2).

## 5. The model estimation

In the first stage of the study, the estimation of all models in classical approach was performed. The hypothesis that the exponential model, which is a special case of the Weibull model, is suitable for describing the duration of first jobs, was verified. Based on the Lagrange Multiplier Statistics test, it was established that at any level of significance the hypothesis in question should be rejected. This means that the shape parameter of Weibull model is different from 1 in the estimated model, so the risk is not constant over time. Therefore, the exponential model was excluded from further analysis.

In the next stage, the Bayesian estimation of Weibull, gamma, log-normal and log-logistic models was carried out. The modelling was performed using the Gibbs sampler (Casella, George, 1992; Gilks, Wild, 1992). To minimize the impact of initial values on posterior estimation, 2000 burn-in iterations were carried out, while another 10.000 chain states were accepted for posterior reasoning.

For the models estimated for the entire sample, the non-informative prior distributions were used for all parameters of these models. In each model, normal prior distributions with the mean 0 and variance $10^6$ were used to estimate the intercept.

In the Weibull model, for the scale parameter, a gamma distribution with the shape parameter 0.001 and the inverse scale parameter also 0.001 was taken as prior distribution. The results for the posterior sample are shown in Table 1. Based on the highest probability density intervals (Bolstad, 2007), both parameters of the model are statistically significant. Based on Geweke test, it was found that there is no indication that Markov chains have converged at any level for all the parameters of the model.

Table 1. Posterior sample mean, interval statistics and Geweke convergence diagnostics for Weibull model

| Weibull Model | | | | | | |
|---|---|---|---|---|---|---|
| **Parameter** | **Mean** | **Standard Deviation** | **Highest Probability Density Interval ($\alpha = 0.05$)** | | **Geweke diagnostics** | |
| | | | | | **z** | **p-value** |
| Intercept | 3.9525 | 0.0637 | 3.8272 | 4.0762 | –1.3307 | 0.1833 |
| Scale | 1.6116 | 0.0512 | 1.5104 | 1.7103 | –1.4651 | 0.1429 |

Source: own calculations

Another estimated model is a generalized gamma model, in which the gamma distribution was used as a normal prior distribution for the scale parameter with the shape parameter 0.001 and the inverse scale parameter 0.001. For the shape parameter of gamma distribution, the normal prior distribution with the mean of 0 and variance of $10^6$ was applied. The resulting characteristics of the posterior sample are shown in Table 2. The conclusions are analogous to the ones drawn for the Weibull model.

The gamma distribution is a generalization of the Weibull distribution and the log-normal distribution. In the gamma model, we obtain the Weibull distribution for the shape parameter that equals 1, and the log-normal distribution for the value 0.

In the log-normal model, the gamma distribution was selected as a prior distribution for the scale parameter with the shape parameter 0.001 and the inverse scale parameter 0.001. The resulting characteristics of the posterior sample are shown in Table 3. For this model, the significance of parameters and convergence of the generated chains was established.

The last investigated model is the log-logistic model, in which the gamma distribution was selected as a prior distribution for the scale parameter with the shape parameter 0.001 and the inverse scale parameter 0.001. The resulting characteristics of the posterior sample are shown in Table 4. The results indicate the significance of model parameters and convergence of generated chains.

Table 2. Posterior sample mean, interval statistics and Geweke convergence diagnostics for gamma model

| Gamma Model | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.5339 | 0.1473 | 1.2480 | 1.8224 | –0.9476 | 0.3433 |
| Scale | 1.5662 | 0.0850 | 1.3947 | 1.7298 | –0.7229 | 0.4698 |
| Shape | –2.1765 | 0.2303 | –2.6291 | –1.7382 | –0.9293 | 0.3527 |

Source: own calculations

Table 3. Posterior sample mean, interval statistics and Geweke convergence diagnostics for log-normal model

| Lognormal Model | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.5339 | 0.1473 | 3.0713 | 3.3261 | –0.6578 | 0.5107 |
| Scale | 1.9461 | 0.0572 | 1.8338 | 2.0579 | –0.1145 | 0.9088 |

Source: own calculations

The comparisons of the estimated models were performed using DIC statistics, the Bayes factor and EAIC and EBIC criteria (Tab. 5). Based on DIC statistics, the gamma model turned out to be the best Bayesian model to fit the data. Next, the value of Bayes factor was calculated for each model in relation to the gamma model (Tab. 5). The values of EAIC and EBIC criteria also indicate the superiority of the gamma model. Therefore, the gamma model was determined to have the greatest explanatory power in modelling the phenomenon in question i.e. the duration of the first job among young people aged 18–30. Figures 1 and 2 show the assumed prior distributions and the obtained posterior distributions for the parameters of shape and scale for the gamma model.

The gamma model, for which the best fit statistics were obtained, was used to model the duration of the first job for men and women separately. The values of the estimated parameters and the highest probability density intervals are shown in Tables 6 and 7. The results indicate the significance of all model parameters and no grounds to reject the hypothesis of the convergence of generated Markov chains at any level of significance.

Table 4. Posterior sample mean, interval statistics and Geweke convergence diagnostics LLogistic model

| LLogistic Model | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval ($\alpha = 0.05$) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 3.1311 | 0.0656 | 3.0040 | 3.2612 | –0.9319 | 0.3514 |
| Scale | 1.1795 | 0.0376 | 1.1091 | 1.2560 | –1.0706 | 0.2844 |

Source: own calculations

Table 5. The Fit Statistics of Bayesian models

| | B. Weibull model | B. Gamma model | B. Lognormal model | B. LLogistic model |
|---|---|---|---|---|
| *DIC* | 3721.733 | 3406.767 | 3553.039 | 3601.146 |
| $\rho_D$ | 1.985 | 3.069 | 1.990 | 1.987 |
| $\log(p(y\|M_i))$ | –1861.4068 | –1704.4266 | –1777.2973 | –1800.8477 |
| $\log(BF)$ | 156.9802 | 0 | 72.8707 | 96.4211 |
| *EAIC* | 3723.748 | 3409.698 | 3555.049 | 3603.159 |
| *EBIC* | 3733.945 | 3424.993 | 3565.246 | 3613.356 |

Source: own calculations

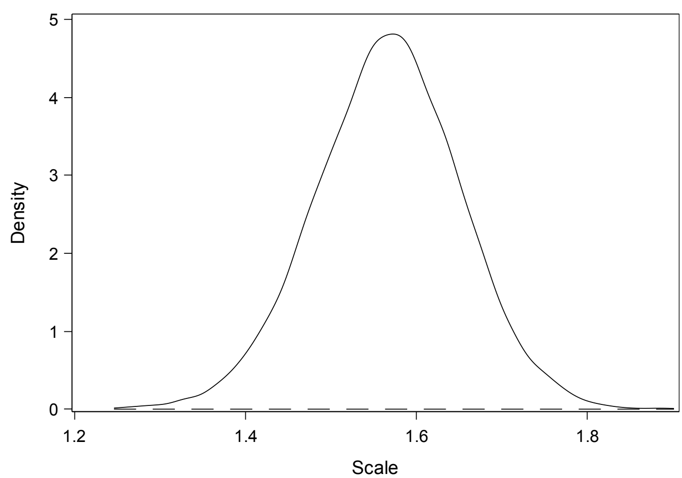


Figure 1. The prior and posterior density for the scale parameter for the gamma model

Source: own calculations

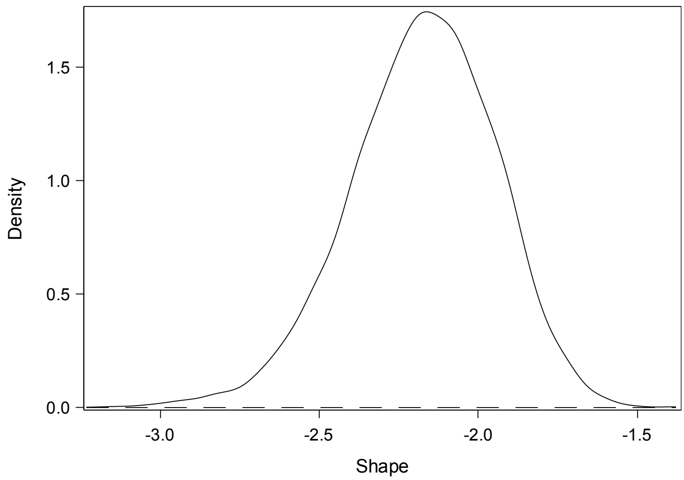


Figure 2. The prior and posterior density for the shape parameter for the gamma model

Source: own calculations

Table 6. Posterior sample mean, interval statistics and Geweke convergence diagnostics for the gamma model (Men)

| Gamma Model for Men | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.4859 | 0.2753 | 0.9349 | 2.0562 | 0.6347 | 0.5256 |
| Scale | 1.6222 | 0.1653 | 1.2815 | 1.9467 | 0.5906 | 0.5548 |
| Shape | –2.4398 | 0.4747 | –3.4160 | –1.5625 | 0.9282 | 0.3533 |

Source: own calculations

The Figures 3 and 4 show the posterior distributions for the scale and shape parameters for both models. The results clearly show the differences in distributions of the duration of the first job for men and women.

The Bayesian approach enables the inclusion of additional information from outside the sample in the analysis. In this study, the results for similar models applied to a cohort of up to five years older were chosen as prior information. The informative gamma distribution was selected for the scale parameter: for men its expected value was 1.7951 with a standard deviation of 0.0977; for women its expected value was 1.7477 with a standard deviation of 0.0742. Informative normal distributions were chosen for the shape parameter: for men its expected value was –1.2416 and standard deviation 0.2500, and for women its expected value was –1.0423 and the standard deviation 0.1936. The results indicate the significance of all model parameters and there are no grounds to reject the hypothesis of the convergence of generated Markov chains at any level of significance (Tables 8–9). Figures 5–8 show the selected prior distributions and resulting posterior.

Table 7. Posterior sample mean, interval statistics and Geweke convergence diagnostics for the gamma model (Women)

| Gamma Model for Women | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.5203 | 0.1820 | 1.1586 | 1.8842 | 0.1671 | 0.8673 |
| Scale | 1.4868 | 0.1035 | 1.2839 | 1.6960 | 0.2085 | 0.8349 |
| Shape | –2.0662 | 0.2913 | –2.6361 | –1.5050 | 0.1585 | 0.8741 |

Source: own calculations

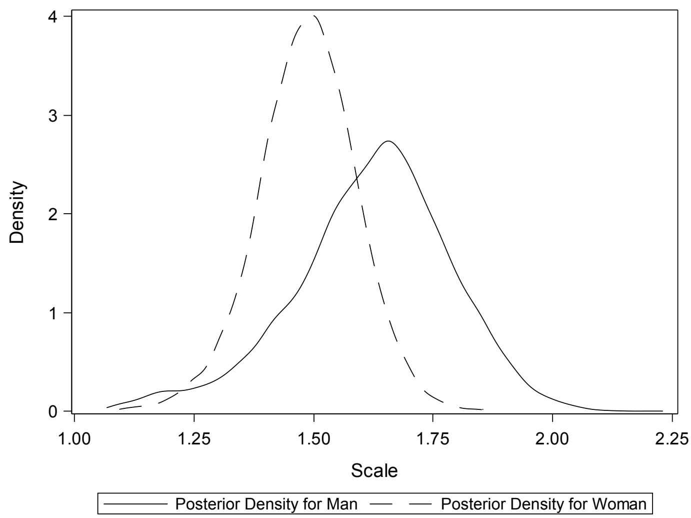


Figure 3. The posterior density for the scale parameter for men and women

Source: own calculations

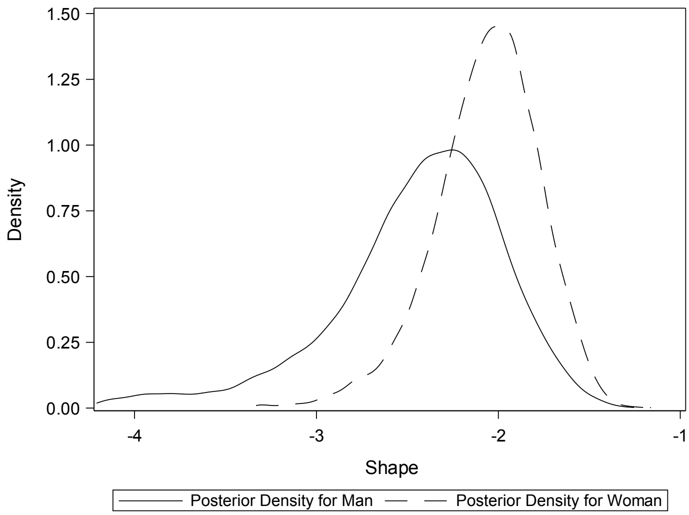


Figure 4. The posterior density for the shape parameter for men and women

Source: own calculations

Table 8. Posterior sample mean, interval statistics and Geweke convergence diagnostics

| Gamma Model for Men with informative prior | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.4859 | 0.2753 | 0.9349 | 2.0562 | 0.6347 | 0.5256 |
| Scale | 1.6222 | 0.1653 | 1.2815 | 1.9467 | 0.5906 | 0.5548 |
| Shape | –2.4398 | 0.4747 | –3.4160 | –1.5625 | 0.9282 | 0.3533 |

Source: own calculations

Table 9. Posterior sample mean, interval statistics and Geweke convergence diagnostics

| Gamma Model for Women with informative prior | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Mean | Standard Deviation | Highest Probability Density Interval (α = 0.05) | | Geweke diagnostics | |
| | | | | | z | p-value |
| Intercept | 1.5203 | 0.1820 | 1.1586 | 1.8842 | 0.1671 | 0.8673 |
| Scale | 1.4868 | 0.1035 | 1.2839 | 1.6960 | 0.2085 | 0.8349 |
| Shape | –2.0662 | 0.2913 | –2.6361 | –1.5050 | 0.1585 | 0.8741 |

Source: own calculations

In the last stage of the analysis, the models estimated for non-informative and informative prior distributions were compared. Model evaluation reveals that the models estimated for non-informative prior distributions are a better fit for the empirical data (Tab. 10).

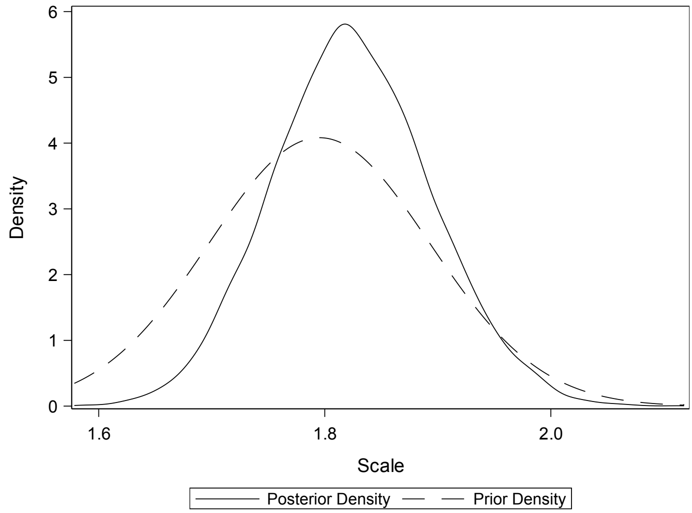


Figure 5. The prior and posterior density for the scale parameter for men

Source: own calculations

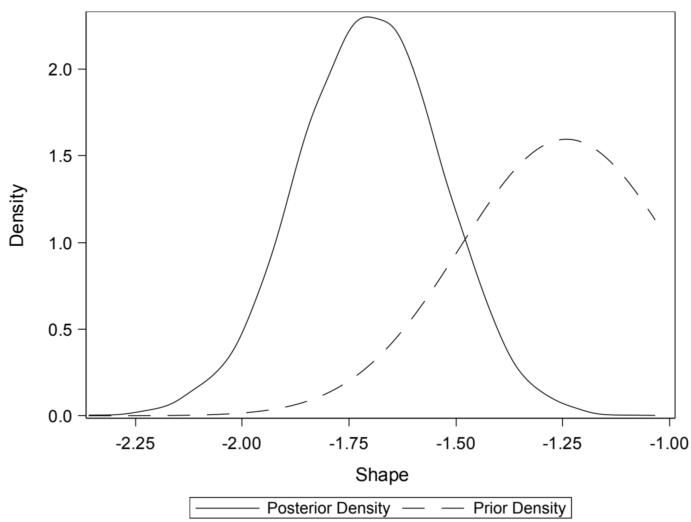


Figure 6. The prior and posterior density for the shape parameter for men

Source: own calculations

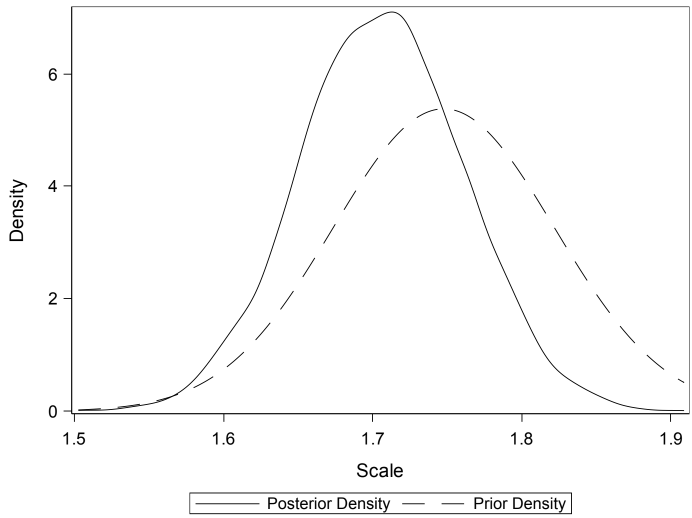


Figure 7. The prior and posterior density for the scale parameter for women

Source: own calculations

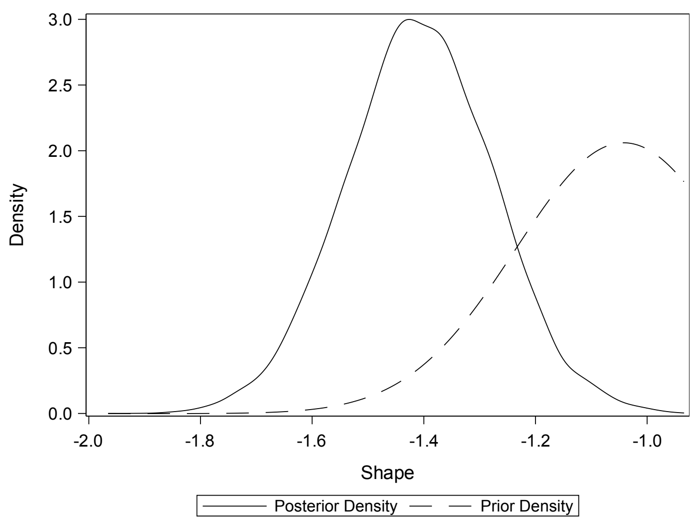


Figure 8. The prior and posterior density for the shape parameter for women

Source: own calculations

Table 10. The Fit Statistics of gamma models for men and women

| | Gamma model for men with noinformative prior | Gamma model for women with noinformative prior | Gamma model for men with informative prior | Gamma model for women with informative prior |
|---|---|---|---|---|
| DIC | 1610.743 | 1795.976 | 1612.190 | 1800.832 |
| $\rho_D$ | 2.990 | 3.050 | 2.020 | 1.969 |
| *EAIC* | 1613.753 | 1798.926 | 1616.17 | 1804.863 |
| *EBIC* | 1626.837 | 1812.268 | 1629.254 | 1818.2049 |
| $\log(p(y\|M_i))$ | –806.236 | –898.848 | –951.122 | –1148.963 |

Source: own calculations

# 6. Summary and conclusions

The objective of this study was to analyze the duration of the first job among young people and find the model that best fits the data. As a result of preliminary analysis on how to model the phenomenon in question, four models were proposed: Weibull, gamma, log-normal and log-logistic. The modelling was carried out in line with the Bayesian approach. Based on DIC statistics, the Bayes factor, EAIC and EBIC statistics, the gamma model was found to be the best fit for the empirical data. The gamma model, for which the best fit statistics were observed, was used to model the duration of the first job for men and women separately. The gamma models were estimated with the non-informative and informative prior distributions. In both cases the posterior distributions for men and women are different.

Modelling according to the classical approach usually results in the estimated values of unknown parameters and standard errors. The results enabled the development of figures of functions describing the process of existing work, such as a survival function. Based on these figures, the way the process looked at a given point in time can be understood. In the initial phase of this study, the models were estimated in line with the classical approach and the resulting graph of survival function for women was below the graph of survival function for men. This means that men are more likely than women not to lose their job throughout the studied period. The Bayesian approach makes it possible to develop the charts based on the posterior means of the model parameters. However, this is not the only benefit arising from the use of Bayesian modelling. Not only can a difference in the posterior means for men and women be observed, but we can also see how these parameters are distributed. The posterior means for the parameters of the models for men and women shown in Tables 6 and 7 are different. Furthermore, it was noticed that (Figures 3 and 4) distributions of parameters are much more concentrated for women. This may suggest that women's behaviour on the labour market is less diverse than men's in the context of the duration of their first job.

Another important advantage of the Bayesian approach is the possibility to include additional information from outside the sample in the process of estimation. The most reliable prior information is that derived from previous studies, so in this work the results of the analysis of the previous birth cohort were used. The impact of the prior information on the posterior distribution depends on the amount of information that can be obtained from the data, which is then related to the sample size. In this study, the inclusion of informative prior affected the posterior distributions of model parameters. Combining additional information with information from the sample is a way to accurately estimate model parameters. The posterior distributions are more concentrated than assumed prior distributions.

This study confirmed the results of other works that suggest that women are more exposed to job loss than men (Drobnič, Frątczak, 2001). Therefore, women in Poland not only have greater difficulty in finding employment (CSO 2014), but also, as shown in this study, are less likely to keep it.

## References

Akaike H. (1973), *Information theory and an extension of the maximum likelihood principle*, [in:] B.N. Petrov, F. Csaki (eds.), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest.

Allison P.D. (1995), *Survival Analysis Using the SAS®: A Practical Guide*, 2nd ed., SAS Institute Inc., Cary.

Ando T. (2010), *Bayesian Model Selection and Statistical Modeling*, CRC Press, Boca Raton.

Bolstad W.M. (2007), *Introduction to Bayesian Statistics*, Wiley & Sons, USA.

Casella G., George E.I. (1992), *Explaining the Gibbs sample*, "The American Statistician", no. 46, pp. 167–174.

Central Statistical Office (2014), *Labour Force Survey* (LFS).

Congdon P. (2006), *Bayesian Statistical Modelling*, 2nd ed., John Wiley & Sons Inc., United Kingdom.

Cox D.R. (1972), *Regression models and life-tables*, "Journal of the Royal Statistical Society", Series B, vol. 34, no. 2, pp. 187–220.

Drobnič S., Frątczak E. (2001), *Employment Patterns of Parried Women in Poland, Careers of Couples in Contemporary Society*, Oxford University Press, New York.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2000), *Bayesian Data* Analysis, Chapman & Hall/CRC, London.

Generations and Gender Programme, http://www.ggp-i.org/ [accessed: 1.09.2016].

Gilks W., Wild P. (1992), *Adaptive rejection sampling for Gibbs sampling*, "Applied Statistics", no. 41, pp. 337–348.

Gill J. (2008), *Bayesian Method: a Social and Behavioral Sciences Approach*, Chapman & Hall/CRC, London.

Grzenda W. (2013), *The significance of prior information in Bayesian parametric survival models*, “Acta Universitatis Lodziensis. Folia Oeconomica”, no. 285, pp. 31–39.
Ibrahim J.G., Chen M-H., Sinha D. (2001), *Bayesian Survival Analysis*, Springer-Verlag, New York.
Jeffreys H. (1961), *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.
Kalbfleisch J.D., Prentice R.L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley Series in Probability and Statistics, Hoboken, New Jersey.
Kass R.E., Raftery A.E. (1995), *Bayes factors*, “Journal of the American Statistical Association”, no. 90, pp. 773–795.
Kim S.W., Ibrahim J.G. (2000), *On Bayesian inference for parametric proportional hazards models using noninformative priors*, “Lifetime Data Analysis”, no. 6, pp. 331–341.
Lancaster T. (1979), *Econometric methods for the duration of unemployment*, “Econometrica”, vol. 47, no. 4, pp. 939–956.
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.
Lawless J.L. (2003), *Statistical Models and Methods for Lifetime Data*, Wiley-Interscience, Hoboken, New Jersey.
Lee E.T., Wang J.W. (2003), *Statistical Methods for Survival Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.
Marzec J. (2008), *Bayesian Models of Variable Quality and Limited Research Loans in Default*, Cracow University of Economics, Cracow.
Newton M.A., Raftery A.E. (1994), *Approximate Bayesian inference by the weighted likelihood bootstrap*, “Journal of the Royal Statistical Society”, Series B (Methodological), vol. 56, no. 1, pp. 3–48.
Osiewalski J. (2001), *Bayesian Econometrics Applications*, Cracow University of Economics, Cracow.
Raftery A. (1996), *Bayesian model selection in social research*, “Sociological Methodology”, no. 25, pp. 111–163.
Spiegelhalter D., Best N., Carlin B., Linde A. van der (2002), *Bayesian measures of model complexity and fit*, “Journal of the Royal Statistical Society”, Series B, no. 64, pp. 583–639.
Walker S., Mallick B.K. (1999), *A Bayesian semiparametric accelerated failure time model*, “Biometrics”, no. 55, pp. 477–483.
Wei L.J. (1992), *The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis*, “Statistics in Medicine”, no. 11, pp. 1871–1879.
Wilks S.S. (1935), *The likelihood test of independence in contingency tables*, “The Annals of Mathematical Statistics”, no. 6, pp. 190–196.
Wilks S.S. (1938), *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, “The Annals of Mathematical Statistics”, no. 9, p. 60–62.

## Modelowanie czasu trwania pierwszej pracy z wykorzystaniem Bayesowskich modeli przyspieszonej porażki AFT

**Streszczenie:** W niniejszym artykule poddano analizie czas trwania pierwszej pracy osób w wieku 18–30 lat. Celem badania jest znalezienie rozkładu, który najlepiej opisuje badane zjawisko. W modelowaniu wykorzystano modele przyspieszonej porażki AFT w ujęciu Bayesowskim. Wykorzystanie podejścia Bayesowskiego rozszerzyło dotychczasowe badania przez możliwość uwzględnienia w badaniu informacji *a priori* oraz umożliwiło porównywanie rozkładów parametrów modeli. Ponadto dało możliwość porównania mocy wyjaśniającej konkurencyjnych modeli na gruncie teorii Bayesowskiej. Z wykorzystaniem zaproponowanych metod porównano czas trwania pierwszej pracy dla kobiet i mężczyzn.

**Słowa kluczowe:** parametryczne modele przeżycia, modele AFT, podejście Bayesowskie, MCMC, zatrudnienie

**JEL:** J630



Received: 2017-01-07; verified: 2017-03-23. Accepted: 2017-07-31