A Bayesian Analysis of Unobserved Heterogeneity for Unemployment Duration Data in the Presence of Interval Censoring

M. Ganjali® **T. Baghfalaki** **D. Berridge**

Shahid Beheshti University Shahid Beheshti University Lancaster University

ABSTRACT

In this paper, we discuss Bayesian inference of unobserved heterogeneity for unemployment duration data in the presence of right and interval-censoring, and non-proportionality. We employ accelerated failure time models with three different distributional assumptions: log-logistic, log-normal, and Weibull models, and use members of an exponential family of distributions for considering unobserved heterogeneity. We adopt a Bayesian approach, using Markov Chain Monte Carlo via WinBUGS software, to analyze the data. The proposed approach is illustrated using the unemployment duration data set of Iran in 2009. A sensitivity analysis using different latent variable models of the exponential family is also considered. After checking convergence, using the Gelman-Rubin diagnostic test, we compared different distributional assumptions using the DIC_3 criterion. Our findings reveal significant discrepancies in unemployment duration based on different covariates for the sample population of Iran in 2009.

Key words: *Accelerated Failure Time Model, Bayesian Analysis, Interval Censoring, Kaplan-Meier Method, MCMC*
JEL Classifications: C11, C41

1. INTRODUCTION

Ordinary regression models of unemployment duration data implicitly assume that, given the measured covariates, the sample population is homogenous: that is, all individuals have the same risk for the event of interest. This assumption is not realistic since demographic differences about which we have no information exist. Sometimes, the undue financial burden of collecting all relevant explanatory information leads to the neglect of some covariates, resulting in unobserved heterogeneity.

Heterogeneity can also be a consequence of disregarding group-specific or individual-specific variation. For example, different people can have distinct genetic characteristics or different employment habits, which may be incorrectly ignored in the study. In multi-city studies there often exist sources of heterogeneity between cities, which may include geographical differences, e.g. different work habits of staff in different cities (Komarek et al., 2007).This is a special case of the omitted variables problem, with its resulting biases.

® M. Ganjali, Professor of Statistics, Department of Statistics, Shahid Beheshti University, Tehran, Iran, (email: m-ganjali@sbu.ac.ir).
T. Baghfalaki, Department of Statistics, Shahid Beheshti University, Tehran, Iran, (email: t_baghfalaki@sbu.ac.ir)
D. Berridge, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, UK, (email: d.berridge@lancaster.ac.uk).

As discussed by Omori and Johnson (1993) and Greene (2003), if the model specification is incomplete due to unobserved but relevant and systematic individual differences in the distribution of the duration data, then inferences based on the improperly specified model are likely to be biased. Unobserved heterogeneity and frailty models try to reduce heterogeneity in the sample. Vaupel et al. (1979) were the first to recognize that unobserved heterogeneity could lead to bias in parameter estimates and to incorrect inferences about the lifetime distribution for univariate and independent survival times. By means of simulation, Tempelman and Gianola (1996) demonstrated the necessity of accounting for unobserved heterogeneity to make correct inferences about model parameters.

In unobserved heterogeneity of time-to-event models, Duchateau et al. (2002) used the frailty proportional hazard model with a center-specific random effect to investigate heterogeneity between centers in multi-center trials. They proposed log-normal and gamma densities for the center-specific random effect and used the expectation-maximization (EM) algorithm in their analysis. Also, Legrand et al. (2005) proposed a Bayesian approach to look for heterogeneity between centers in a proportional hazard model. They modeled heterogeneity by including a center-specific random effect and a random treatment by center interaction. Yamaguchi and Ohashi (1999) also discussed the use of frailty modeling to investigate heterogeneity between centers in time-to-event data. Campolieti (2001) conducted a Bayesian analysis of unobserved heterogeneity of duration data. The model employed by Campolieti (2001) uses Dirichlet mixtures of normals, which includes Heckman and Singer's (1982, 1984) approach as a special case. Some tests of heterogeneity have been proposed in recent years, where a homogeneity test is the testing of the degeneration of the random effect distribution at point zero. Vindenes et al. (2008) have mentioned that usual heterogeneity can be important, particularly in understanding the behavior of small populations, due to its impact on demographic variance. For example, Knape et al. (2011) handled heterogeneity in the island population of Silvereyes by using a random effects model.

In this paper, we use a latent variable model to handle unobserved group-specific heterogeneity in the analysis of unemployment duration data. Introducing this unobserved random factor modifies the accelerated failure time (AFT) model. We discuss a Bayesian latent variable model which is a member of the exponential family of distributions for heterogeneity. To the best of our knowledge, this model has not yet been discussed in the unobserved heterogeneity literature.

We employ the proposed methodology to analyze unemployment duration data for Iran in 2009. In this study, we group the data according to the province in which an individual lives. It is essential to consider a heterogeneity factor for modeling because of non-homogeneity among different provinces. As will be discussed in Section 4, the data set does not satisfy the proportionality assumption; hence an AFT model is used for modeling this data set.

We use three distributions: log-logistic, log-normal, and Weibull for the duration of unemployment. The validity of these three distributional assumptions is supported by goodness-of-fit tests based on some probability plots. We compare the models with different distributional assumptions using a Bayesian criterion and obtain the results of the Bayesian implementation using the available software WinBUGS (Spiegelhalter et al., 2003). Next, we check the convergence of parameters using the Gelman-Rubin criterion in the BOA (Bayesian Output Analysis) package. We then present a sensitivity analysis with respect to different random effects distributional assumptions.

The remainder of the paper is organized as follows: Section 2 discusses a brief review of unobserved heterogeneity modeling. Section 3 presents the Bayesian model and the computational approach to handling unobserved heterogeneity in the modeling of our data set. Section 4, after describing the response and explanatory variables of interest, implements a goodness-of-fit test, checks the convergence of the MCMC model, and discusses the results and a comparison of the performance of different distributions. Section 5 contains some concluding remarks.

2. UNOBSERVED HETEROGENEITY MODELING

Population homogeneity, a common assumption in much data analysis, assumes that all differences in the population have been captured by the measured variables in the study. However in reality, unobserved heterogeneity, caused by omitting relevant variables, exists in many settings. There are several approaches that may be used to investigate the consequences of unobserved heterogeneity in data analysis. The basic idea behind these methods recognizes that there is an unobserved or latent variable, which may have a discrete or a continuous distribution.

One such method is the use of a frailty model. The frailty approach to modeling unobserved covariates is based on the choice of a frailty distribution. Different distributions are commonly assumed for frailty, for example gamma, log-normal, inverse Gaussian, and log-t. (Wienke, 2011). In time-to-event data, the hazard function of an individual, $\mu(t|X,Z)$, depends on an unobservable time-independent random variable Z. In the multiplicative hazard framework, Z acts multiplicatively on the baseline hazard function $\mu_0(t)$ as follows:

$$\mu(t \mid X, Z) = Z\mu_0(t)e^{X'\beta}$$

where $X=(X_1,...,X_k)$ and $\beta=(\beta_1,...,\beta_k)$ are vectors of explanatory variables and regression coefficients, respectively. Z is a nonnegative latent variable, where $E[Z]=1$ and its variance is interpreted as a measure of heterogeneity.

Another way of considering unobserved heterogeneity is to use a latent class approach. Heckman and Singer (1982, 1984) proposed a nonparametric random effects approach based on the latent class. Their formulation is similar to an ordinary frailty model with a discrete distributional assumption for the frailty term as the latent class. In their work, the likelihood function for the i^{th} subject is given by:

$$\ell_i = \sum_{w=1}^{K} \pi_w \mu(t_i \mid x_i, \theta_w)^{\delta_i} S(t_i \mid x_i, \theta_w)$$

where π_w is the proportion of the population belonging to the latent class w with latent variable θ_w, such that the number of latent classes is K, and δ_i is the right-censoring indicator. Hagenaars and McCutcheon (2002) proposed a latent class approach to handling heterogeneity in the hazard modelling of survival data. There are some parametric and non-parametric methods for the estimation of latent class parameters (Hagenaars and McCutcheon, 2002).

The most widely used method in duration modelling with explanatory variables is the broadly applicable Cox proportional hazards model (Cox, 1972). The Cox model, which is semi-parametric, has enjoyed tremendous success in applied work with the availability of software enabling model estimation and inference. Nonetheless one can not always use this method since the validity of the proportionality assumption has to be confirmed before using this

model. When the proportionality assumption of the Cox regression model is not valid, AFT models with different distributional assumptions may be used as an alternative approach.

Komarek et al. (2005) considered the following shifted and scaled penalized Gaussian mixture model:

$$\log(T_i) = \alpha + \beta' x + \sigma\varepsilon$$

where x is a vector of explanatory variables, α is an intercept parameter, β is a vector of regression coefficients, and σ is a scale parameter. The distribution of the error term, ε, in their proposed model is:

$$f(\varepsilon) = \sum_{j=1}^{g} w_j(a)\phi_{\mu_j,\sigma_0^2}(\varepsilon)$$

where $\phi_{\mu_j,\sigma_0^2}(\varepsilon)$ is the Gaussian density with mean μ_j and variance σ_0^2, and $w_j(a)$, $j = 1,..., g$ are the mixture coefficients, which are specified as:

$$w_j(a) = \frac{e^{a_j}}{\sum_{i=1}^{g} e^{a_i}}$$

such that $0 < w_j(a) < 1$ and $\Sigma_{i=1}^{g} w_j(a) = 1$.

A random effect AFT model is discussed by Pan and Louis (2000), who consider a non-parametric Kaplan-Meier estimate in their estimation procedure. Laird and Ware's (1982) random effect AFT model (Komarek et al., 2007) is an AFT model with a random effect term to account for heterogeneity. Komarek et al. (2007) generalized the random effect AFT model by using a penalized Gaussian mixture as the error distribution. Their computation is based on Markov chain Monte Carlo (MCMC) techniques.

In these models, usually a gamma distributional assumption is used for the random effects. In this paper, we use a random effect AFT model as $\log(T) = x'\beta + b + \varepsilon$ assuming the random effect (b) follows a distribution which is a member of the exponential family distribution. We then perform some sensitivity analyses with respect to the change in the distributional assumption for the random effect.

A goodness-of-fit probability plot (see Section 4), allows us to use three different distributional assumptions: log-logistic, log-normal and Weibull when analyzing our duration data (Green, 2003). In the following section, we adopt a Bayesian approach to handling group-specific heterogeneity in a random effects AFT model.

3. BAYESIAN RANDOM EFFECTS AFT MODELS FOR GROUP-SPECIFIC HETEROGENEITY

Latent variable models may be used for analyzing event time data, when there is no reason to reject an assumption of unobserved heterogeneity. In our proposed model, we use an AFT model for analyzing unemployment duration data of Iran. We consider the effect of some explanatory variables, but we predict that some variables may exist that are not recorded in this study such as, for example, geographical differences between provinces or different economic conditions. Therefore, we define a latent variable that varies among provinces to handle this unobserved heterogeneity in our proposed model. Let T_{ij} denote unemployment duration for the i^{th} subject in the j^{th} province ($i = 1,..., n_j$; $j = 1,..., m$). Under the exponential

family distributional assumption, we define a random effect b_j the estimated value of which distinguishes between homogenous and non-homogenous provinces.

The random effect AFT model for group-specific heterogeneity is given by:

$$\log(T_{ij}) = x_{ij}^{'}\beta + b_j + \varepsilon_{ij}, i = 1,\ldots,n_j; j = 1,\ldots,m \qquad (3.1)$$

where $n=\Sigma_{j=1}^{m} n_j$ is the total sample size, ε_{ij} s are independent and identical error terms, which are distributed as normal, logistic, or extreme value, $x_{ij}=(x_{ij1},..., x_{ijp})'$ are p-dimensional vectors of explanatory variables for the i^{th} individual in the j^{th} province, and $\beta_{ij}=(\beta_1,..., \beta_p)'$ is the vector of coefficients.

Suppose that n independent and identical vectors of $(t_{ij,l},t_{ij,u},x_{ij}^{'})^{'}$, where t_{ij} is the time-to-event for the i^{th} individual in the j^{th} group, are observed, such that $t_{ij}\in(t_{ij,l},t_{ij,u}]$. For a right-censored observation $t_{ij,u}=\infty$, and for an exactly observed event time $t_{ij,l}=t_{ij,u}=t_{ij}$. The x_{ij} is a $p\times1$ vector of explanatory variables.

A random variable Z follows an s–dimensional exponential family distribution, if its density is of the form:

$$h(z;\nu) = e^{\sum_{i=1}^{s}\nu_i T_i(z)-A(\nu)} \times h_0(z). \qquad (3.2)$$

This form of exponential family is said to have canonical form, and v_i, $i=1,...,s$, are canonical parameters.

The advantage of our proposed method, in contrast to former random effect AFT models, is that it can consider various significant distributional assumptions for duration time (survival time) where selection among these distributions can easily be justified using a graphical method. In our proposed method, we use different members of the exponential distribution family for the random effect and perform a sensitivity analysis to examine the effect of a change in distribution on the results. Also, we conduct all the computations for parameter estimation using available software WinBUGS, where we implement the Bayesian criterion DIC_3 for model comparison. Also, we demonstrate the random effect's influence distinguished between the homogenous and non-homogenous area of the population (illustrated in Figure 4.5).

In the following subsections we explain the Bayesian implementation of our approach.

3.1. Log-normal and Log-logistic Random Effect AFT Models in Bayesian Perspective

The structure of the log-normal model is given by:

$$T_{ij} \mid b_j \sim LN(x_{ij}^{'}\beta + b_j, \sigma^2),\ b_j \sim h(b_j;\nu) \qquad (3.3)$$

where $h(.,\nu)$ is given by (3.2). Let $\theta=(\beta,\sigma,\nu)$. The likelihood function for this model is given by:

$$L(\theta;t) = \prod_{j=1}^{m}\left[\int_0^{\infty}\prod_{i=1}^{n_j}\left[\left\{\int_{t_{ij,l}}^{t_{ij,u}}\frac{1}{\sigma u}\phi(\log(u);x_{ij}^{'}\beta+b_j,\sigma^2)du\right\}^{\delta_{ij}} \times\left\{\frac{1}{\sigma t_{ij}}\phi(\log(t_{ij});x_{ij}^{'}\beta+b_j,\sigma^2)\right\}^{1-\delta_{ij}}\right]h(b_j;\nu)db_j\right]$$

where δ_{ij} is an indicator function, which takes value one for complete data and value zero for interval-censored observations. For right-censored observations $t_{ij,u}=\infty$. Also, $\phi(.;\mu,\sigma^2)$ denotes the density function of the normal distribution with mean μ and variance σ^2, and is a member of the exponential family distribution.

For a Bayesian structure, as mentioned by Ibrahim et al. (2002), the following distributions are appropriate proposals for priors:

$$\beta|\sigma^2 \sim N_p(\mu_0,\sigma^2 V_0), \quad \sigma^2 \sim I\Gamma(a,b), \quad \nu \sim \pi(\nu;\gamma). \tag{3.4}$$

The hyper-parameters of these priors may be selected such that one can have low-informative prior distributions.

Via the hierarchical structure of (3.3) and prior distributions of (3.4), the joint posterior distribution for all parameters is given by:

$$\pi(\beta,\sigma^2,\nu,b|t,x) \propto \prod_{j=1}^{m}\prod_{i=1}^{n_j}\int_{t_{ij,l}}^{t_{ij,u}} \frac{1}{\sigma u}\phi(\log(u);x_{ij}'\beta+b_j,\sigma^2)du \times h(b_j;\nu)\times\pi(\nu;\gamma)\times(\sigma^2)^{-p/2}e^{-\frac{1}{2\sigma^2}(\beta-\mu_0)'V_0^{-1}(\beta-\mu_0)}\times(\sigma^2)^{-a-1}e^{-\frac{b}{\sigma^2}} \tag{3.5}$$

The posterior distribution for the above model specification does not have closed form solutions for the parameters. To perform the Bayesian analysis, MCMC techniques can be used to sample the joint posterior distribution of these models. One special MCMC type approach, which requires only the specification of the conditional posterior distribution for each parameter, is the Gibbs sampler (Casella and George, 1992). For implementation of the Gibbs sampler, we need the full conditional distributions. These are given as follows:

$$\begin{aligned}
&f(t_{ij}|b_j,\beta,\sigma,\nu,x_{ij}) \propto f_{LN}^{\delta_{ij}}(t_{ij};x_{ij}'\beta+b_j,\sigma^2)\times[F_{LN}(t_{ij,u};x_{ij}'\beta+b_j,\sigma^2)-F_{LN}(t_{ij,l};x_{ij}'\beta+b_j,\sigma^2)]^{1-\delta_{ij}}\\
&\pi(\sigma|b_j,t_{ij},\beta,\nu,x_{ij}) \propto f_{LN}^{\delta_{ij}}(t_{ij};x_{ij}'\beta+b_j,\sigma^2)\times[F_{LN}(t_{ij,u};x_{ij}'\beta+b_j,\sigma^2)-F_{LN}(t_{ij,l};x_{ij}'\beta+b_j,\sigma^2)]^{1-\delta_{ij}}\times f_N(\beta;\mu_0,\sigma^2V_0)\times I\Gamma(\sigma;a_0,b_0)\\
&\pi(b_j|t_{ij},\beta,\sigma,\nu,x_{ij}) \propto f_{LN}^{\delta_{ij}}(t_{ij};x_{ij}'\beta+b_j,\sigma^2)\times h(b_j;\nu)\times[F_{LN}(t_{ij,u};x_{ij}'\beta+b_j,\sigma^2)-F_{LN}(t_{ij,l};x_{ij}'\beta+b_j,\sigma^2)]^{1-\delta_{ij}}\\
&\pi(\nu|\beta,t_{ij},b_j,\sigma,x_{ij}) \propto h(b_j;\nu)\times\pi(\nu;\gamma),\pi(\beta|t_{ij},b_j,\sigma,\nu,x_{ij}) \propto f_{LN}^{\delta_{ij}}(t_{ij};x_{ij}'\beta+b_j,\sigma^2)\times f_N(\beta;\mu_\beta,\Sigma_\beta)\times[F_{LN}(t_{ij,u};x_{ij}'\beta+b_j,\sigma^2)-F_{LN}(t_{ij,l};x_{ij}'\beta+b_j,\sigma^2)]^{1-\delta_{ij}}
\end{aligned} \tag{3.6}$$

where

$$\mu_\beta = \left(x_{ij}(\log t_{ij}-b_j)+(\sigma^2V_0)^{-1}\mu_0\right)\times\left(x_{ij}x_{ij}'+(\sigma^2V_0)^{-1}\right)$$

and

$$\Sigma_\beta = \sigma^2\left(x_{ij}x_{ij}'+(\sigma^2V_0)^{-1}\right)^{-1}$$

Gibbs sampling is needed for the implementation of equation (3.6),which can be done using the WinBUGS software (Spiegelhalter et al., 2003).

The log-logistic AFT model and its Bayesian implementation have structures similar to those of the log-normal AFT model, wherein the log-normal distribution should be replaced by the log-logistic distribution.

3.2. Weibull Random Effect AFT Model in Bayesian Perspective

Similar to Section 3.1, the structure of the Weibull random effect AFT model is given by:

$$T_{ij}|b_j : Weib(\lambda_{ij},\eta),\ log(\lambda_{ij}) = -\{x_{ij}'\beta+b_j\},\ b_j : h(b_j;\nu) \tag{3.7}$$

For $\theta=(\beta,\sigma,\nu)$, the likelihood function of this model is given by:

$$L(\theta;t)=\prod_{j=1}^{m}\left\{\left[\int_0^{\infty}\prod_{i=1}^{n_j}\left\{\int_{t_{ij,l}}^{t_{ij,u}} f_{Wei}(u;\lambda_{ij},\eta)du\right\}^{\delta_{ij}}\times f_{Wei}^{1-\delta_{ij}}(t_{ij};\lambda_{ij},\eta)\right]h(b_j;\nu)db_j\right\}$$

where $f_{Wei}(.;\lambda_{ij},\eta)$ denotes the Weibull density function with parameters λ_{ij} and η. The independent prior distributions for the Bayesian structure are given by:

$$\beta\sim N_p(\mu_0,V_0),\quad \eta^2\sim I\Gamma(a,b),\quad \nu\sim\pi(\nu;\gamma). \tag{3.8}$$

The hyper-parameters of these priors are again selected such that one can have the low-informative priors. Combining the complete likelihood function and the prior distributions (3.8), the joint posterior distribution of the parameters is given by:

$$\pi(\beta,\eta,\nu,b|\,t_{ij},x_{ij})\propto\prod_{j=1}^{m}\prod_{i=1}^{n_j}\int_{t_{ij,l}}^{t_{ij,u}} f_{Wei}(u;\lambda_{ij},\eta)du\times h(b_j;\nu)\times\pi(\nu;\gamma)\times e^{-\frac{1}{2}(\beta-\mu_0)'V_0^{-1}(\beta-\mu_0)}\times(\sigma^2)^{-a-1}e^{-\frac{b_j}{\sigma^2}} \tag{3.9}$$

Also, hierarchical structures for the Gibbs sampler implementation are given by:

$$\begin{aligned}
&f(t_{ij}\,|\,b_j,\beta,\eta,\nu,x_{ij})\propto f_{Wei}^{\delta_{ij}}(t_{ij};\exp(-x_{ij}'\beta+b_j),\eta)\times[F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)-F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)]^{1-\delta_{ij}},\\
&\pi(\beta|\,t_{ij},b_j,\eta,\nu,x_{ij})\propto f_{Wei}^{\delta_{ij}}(t_{ij};\exp(-x_{ij}'\beta+b_j),\eta)\times f_N(\beta;\mu_0,V_0)\times[F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)-F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)]^{1-\delta_{ij}},\\
&\nu|\beta,t_{ij},b_j,\eta,x_{ij}\propto h(b_j;\nu)\times\pi(\nu;\gamma),\\
&\pi(\eta\,|\,b_j,t_{ij},\beta,\nu,x_{ij})\propto f_{Wei}^{\delta_{ij}}(t_{ij};\exp(-x_{ij}'\beta+b_j),\eta)\times f_{I\Gamma}(\eta;a_0,b_0)\times[F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)-F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)]^{1-\delta_{ij}},\\
&\pi(b_j\,|\,t_{ij},\beta,\eta,\nu,x_{ij})\propto f_{Wei}^{\delta_{ij}}(t_{ij};\exp(-x_{ij}'\beta+b_j),\eta)\times h(b_j;\nu)\times[F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)-F_{Wei}(t_{ij,u};\exp(-x_{ij}'\beta+b_j),\eta)]^{1-\delta_{ij}},
\end{aligned} \tag{3.10}$$

where δ_{ij} is an indicator function, which takes value one for complete data and value zero for interval censored observations. For right-censored observations $t_{ij,u}=\infty$. The equations in (3.10) provide the opportunity to program the Bayesian implementation in the WinBUGS software.

3.3. Measuring unobserved heterogeneity

The amount of unobserved heterogeneity is determined by the size of the standard deviation of the latent variable distribution, such that, the larger the standard deviation of the latent variable, the stronger the unobserved heterogeneity. In the exponential family (2), the unobserved heterogeneity factor, which is the covariance of the random variable Z, can be computed by:

$$\operatorname{cov}(T_j(z),T_k(z))=\frac{\partial^2}{\partial\nu_j\partial\nu_k}A(\nu),$$

$$\operatorname{var}(b_j)=\frac{\partial^2}{\partial b_j^2}A(b_j),$$

where the form of $A(b_j)$ depends on the selected member of the exponential family of distributions (Lehmann and Casella, 1998, p. 23-32).

4. APPLICATION

4.1. The Data Set

The data set used in this paper is extracted from a follow-up study conducted by the Statistical Center of Iran. In these data, the labor force status of people is recorded in two seasons (spring and summer) in 2009. We have selected the individuals who are observed on both seasons and are unemployed in spring (unemployed individuals answer a question about their duration of unemployment in spring). The data contains detailed individual information for a

random sample of the population aged 14 and older. The vector of explanatory variables includes personal characteristics such as gender, age, the place of residence, current marital status, education status, and the number of household members. Details of the categories of explanatory variables and their percentages are listed in Table 4.1. In this study, we are concerned with the existence of a set of covariates, which may have been omitted or may not have been considered in each province of the study.

Explanatory variable	Categories	Percentage
Current marital status	married	2.94 %
	widow(er) or divorced	7.01 %
	single	6.99 %
Gender	female	2.25 %
	Male	7.75 %
Age	< 20	1.26 %
	21–25	3.78 %
	26–30	2.49 %
	30 >	2.47 %
Education status	under diploma	4.64 %
	diploma	3.07 %
	associate of arts (or science)	8.10 %
	MA and upper	1.48 %
Number of household members	one or two	8.21 %
	three	8.83 %
	four and more	3.39 %
Residence	rural	2.77 %
	urban	7.23 %

Table 4.1 Different categories of chosen explanatory variables along with their percentages.

	Frequency	Percent
Right-censored data[1]	743	55.6
Completed observation[2]	121	9.0
Interval–censored data[3]	473	35.4

Table 4.2 Employment status in the summer of 2009 of unemployed individuals in the spring of 2009.
Notes: [1]: Still unemployed in summer, [2]: Duration is recorded, [3]: Duration is recorded in an interval.

Table 4.2 gives the frequencies and percentages for different categories of unemployment status in the summer of 2009, for unemployed individuals in the spring of 2009. This table shows that of the 1337 individuals in the study, 743 individuals remained unemployed in the summer, and 473+121 individuals became employed. Unfortunately, the exact duration of unemployment has only been recorded for 121 individuals. For the other 473 individuals we only know that their shift to employment happened during a 3-month period. The employment duration of these individuals can be considered as interval-censoring. Figure 4.1 illustrates the survival curve for unemployment duration. Points on this curve estimate the proportion of individuals who remained unemployed over time.

For a preliminary description of the explanatory variables in the data set, Figure 4.2 shows the Kaplan-Meier estimate of the survival curves of unemployment duration for different categories of explanatory variables. For example, according to Figure 4.2(a) females remained unemployed for a longer period than males.

Figure 4.1 Survival curve of unemployment duration along with its confidence bands.

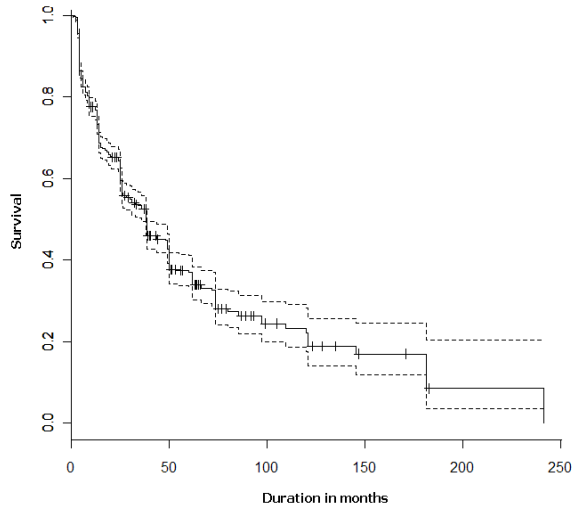
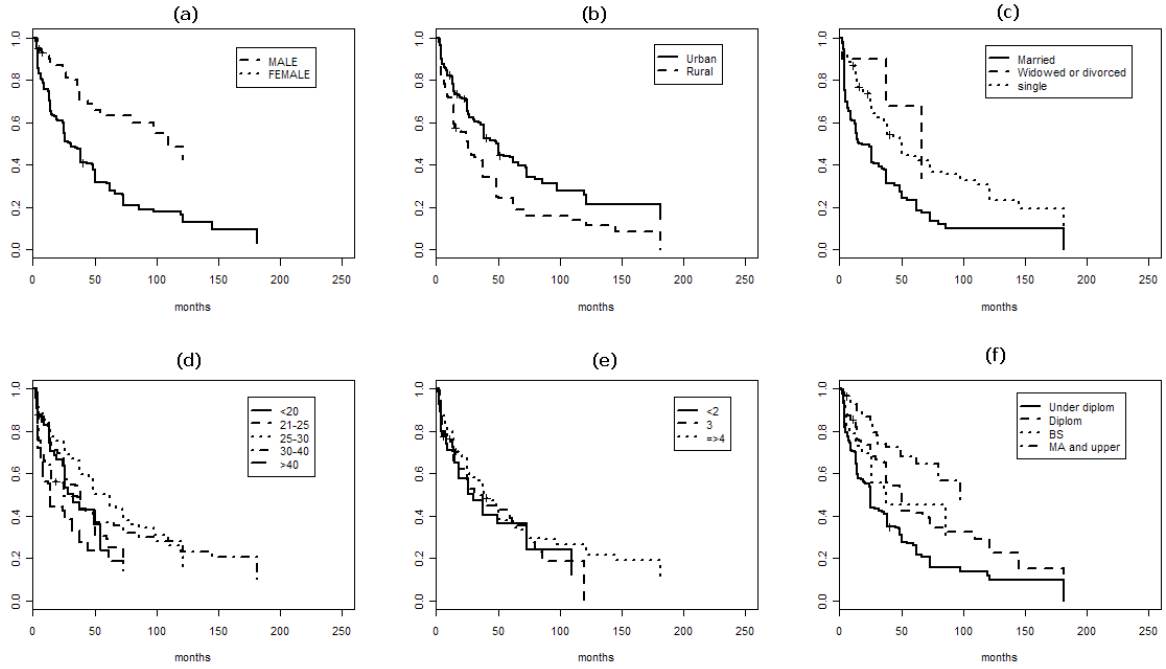


Figure 4.2 Kaplan-Meier estimates of the survival curves of unemployment duration by (a): Gender, (b): Place of residence, (c): Current marital status, (d): Age group, (e): Number of household members, (f): Educational level.



4.2. Selection of Theoretical Distribution Based on the Probability Plot

The probability plot is one of the most frequently applied methods for checking the distributional assumption. In this paper, we use the method of Lee and Wang (2003) with some adjustments.

As mentioned by Lee and Wang (2003), if the theoretical distribution is adequate for the data, a graph of log(t) versus a function of the sample cumulative distribution function will be close to a straight line. In other words, a fitted linear regression for log(t) and that function of the cumulative distribution function is a good index for the selection of a theoretical distribution.

The regression lines for the log-normal, log-logistic, and Weibull distributions are given by:

$$\log t_i = \frac{1}{\gamma}\log\frac{1}{\lambda} + \frac{1}{\gamma}\log(\log(\frac{1}{1-F(t_i)})) + e_i$$
$$\log t_i = \mu + \sigma\Phi^{-1}(F(t_i)) + e_i, \qquad (4.11)$$
$$\log t_i = \mu + \sigma\log(\frac{1}{1-F(t_i)} - 1) + e_i,$$

where e_is are the error terms of the regression models. Thus, a quick goodness-of-fit test is a regression line of $\log(t_i)$ versus a function of $\widehat{F}(t_i)$, where $\widehat{F}(t_i)$ is an estimate of $F(t_i)$. This method can be summarized in the following steps:

- Select a theoretical distribution for the survival time T.
- Estimate the cumulative distribution function. There are several approaches: the most famous is the Kaplan-Meier estimate, another method used by Lee and Wang (2003) is the use of $(i-0.5)/n$ for the i^{th} ordered time values, $i=1,...,n$. In this method, right-censored observations are considered only in sorting the index i. For interval-censored data, midpoint imputation may be used.
- Fit a linear regression for $\log(T)$ and the function of the cumulative distribution function.

The R-squared values of the fitted lines for the log-normal, log-logistic, and Weibull models are summarized in Table 4.3. This table shows that these three distributions are potential candidates when analyzing these data.

	Multiple R-squared	
	$(i-0.5)/n$	Kaplan-Meier
log-normal model	0.977	0.914
log-logistic model	0.958	0.965
Weibull model	0.915	0.935

Table 4.3 values of R-squared for fitted regression models in probability plot approach.

4.3. Modeling unemployment duration in Iran

In this section, we analyze the data set based on the proposed methods given in Section 3. The explanatory variables in this data set were listed in Table 4.1. In the following, we consider age as a continuous covariate. If a categorical variable has g categories then $g-1$ dummy variables need to be created, and consequently, $g-1$ regression coefficients have to be estimated. For example, for current marital status, the dummy variables are defined as follows:

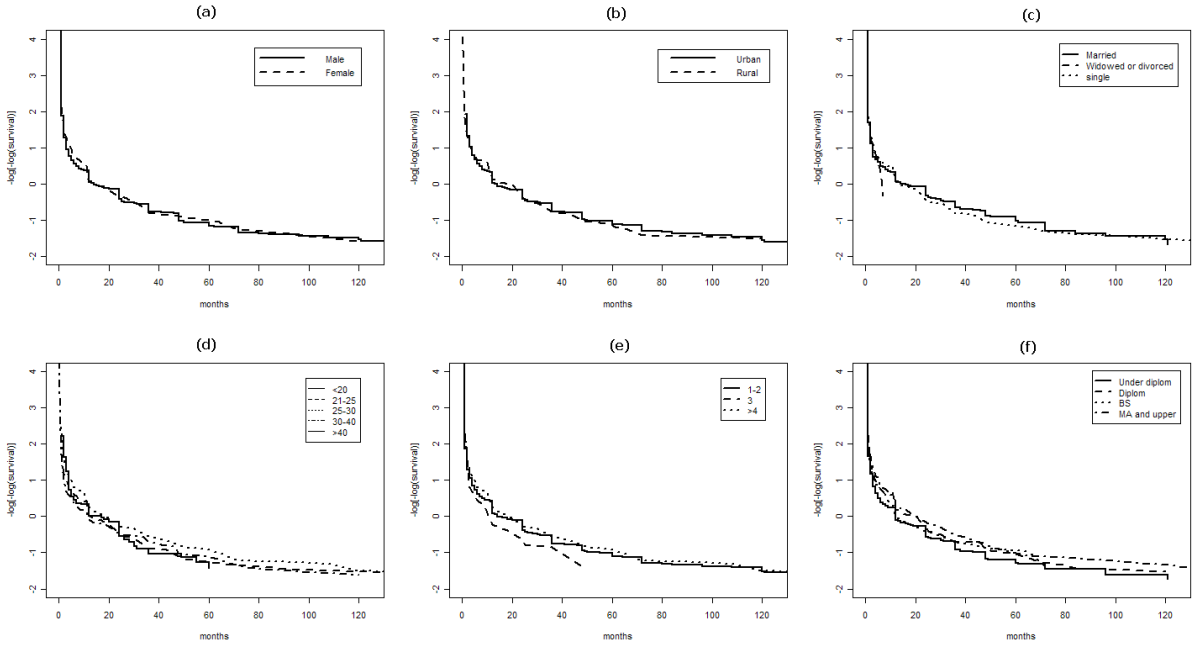
$$\mathbf{mar1} = \begin{cases} 1 & \textbf{if married} \\ 0 & \textbf{o.w.} \end{cases}, \quad \mathbf{mar2} = \begin{cases} 1 & \textbf{if widow(er) or divorced} \\ 0 & \textbf{o.w.} \end{cases},$$

and mar1 = mar2 = 0 defines single status.

The Cox proportional hazard model is a common model to be applied, because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. However, one may use a graphical test for checking the validity of the proportionality assumption (Deshpande and Purohit, 2005; page 189). We have checked the proportionality assumption using log−log plots of unemployment duration for different levels of explanatory

variables. Figure 4.3 shows these plots. In some panels of this figure (for example panel c) the log−log survival functions are not parallel for different categories of explanatory variables, therefore the proportional hazards assumption is not valid. Consequently, the proposed AFT models of Section 3 should be applied.

Figure 4.3 Graphical test for proportional hazards: Kaplan-Meier log-log plots of unemployment duration by (a): Gender, (b): Place of residence, (c): Current marital status, (d): Age group, (e): Number of household members, (f): Educational level.



The AFT random effect model for group-specific heterogeneity is given by:

$$\log(T_{ij}) = \beta_0 + \beta_1 mar1_{ij} + \beta_2 mar2_{ij} + \beta_3 sex_{ij} + \beta_4 age_{ij} + \beta_5 edu1_{ij} + \beta_6 edu2_{ij} + \beta_7 edu3_{ij}$$
$$+ \beta_8 num1_{ij} + \beta_9 num2_{ij} + \beta_{10} res_{ij} + b_j + \varepsilon_{ij}, \quad i = 1,...,n_j;\ j = 1,2,...,30. \tag{4.12}$$

In this model, *mar*1 and *mar*2 denote current marital status, and *sex* and age denote the gender and age of individuals, respectively. Dummy variables *edu*1, *edu*2, and *edu*3 are used for education levels, *num*1 and *num*2 denote categories of household sizes, and *res* is used to represent living area. Also, we have considered the above-mentioned three distributional assumptions for the duration data.

For the random effect b_j, we have considered different members of the exponential family of distributions. Vaupel et al. (1979) proposed the use of a gamma distribution for b_j, with a mean of one and a variance of $1/\gamma$, where γ is the unknown parameter to be estimated. Several authors have proposed incorporating a gamma-distributed random term (Tuma and Hannan, 1984; Lancaster, 1990). The analytically tractable and readily computational properties of the gamma distribution are important reasons, which are mentioned by Vaupel et al. (1979), for selecting the gamma as an appropriate mixing distribution. Moreover, as Hagenaars and McCutcheon (2002) mentioned, it is a flexible distribution that takes on a variety of shapes as the dispersion parameter γ varies. As mentioned in Section 3.3, the amount of unobserved heterogeneity in this model is determined by the standard deviation of b_j, which is $1/\sqrt{\gamma}$. Also, the normal mixing distribution is the most important base for Laird and Ware's (1982) random effect model.

In our analysis, we ran two MCMC chains with 30000 iterations for each. Then, we discarded the first 10000 iterations as burn-in and retained 20000 for the posterior analysis. We checked the convergence of parameter estimates using the Gelman and Rubin diagnostic test (Gelman and Rubin, 1992) for all models. These approaches were implemented using the BOA package.

As the DIC criterion (Spiegelhalter et al., 2002), which is automatically calculated by WinBUGS, is not adequate for model comparison (Celeux et al., 2006; DeIorio and Robert, 2002), in this study we used the DIC_3 criterion (Celeux et al., 2006).

Let Θ and $Z=(z_1,...,z_N)'$ be the entire model parameters and data, respectively. DIC_3 is given by:

$$DIC_3 = -4E_\theta\left[\log \mathbf{f(z}|\theta\mathbf{)}\,|\mathbf{z}\right] + 2\log \hat{f}(z),$$

where $\hat{f}(z) = \Pi_{i=1}^{N} \hat{f}(z_i) = E_\theta\left[\mathbf{f(z}|\theta\mathbf{)}\,|\mathbf{z}\right]$. The smaller the value of DIC_3 is, the better the fit of the model.

	Log-Logistic	Log-Normal	Weibull
Parameters	Est. (S.E.)	Est. (S.E.)	Est. (S.E.)
Intercept	3.192 (0.275)	3.082 (0.252)	3.614 (0.300)
Gender			
Female	0.826 (0.135)	0.794 (0.129)	0.881 (0.146)
Baseline (male)	-	-	-
Current marital status			
Married	-0.951 (0.117)	-0.909 (0.117)	-0.815 (0.103)
Widow(er) or divorced	-0.178 (0.571)	-0.102 (0.544)	-0.086 (0.644)
Baseline (single)	-	-	-
Education level			
Under diploma	-0.848 (0.174)	-0.813 (0.165)	-0.801 (0.189)
Diploma	-0.462 (0.177)	-0.418 (0.167)	-0.417 (0.193)
BS	-0.873 (0.218)	-0.816 (0.208)	-0.809 (0.232)
Baseline(MA and higher)	-	-	-
Age	0.011 (0.004)	0.010 (0.006)	0.013 (0.002)
Number of household members			
One or two	-0.309 (0.197)	-0.278 (0.191)	-0.260 (0.183)
Three	0.024 (0.122)	0.052 (0.123)	-0.095 (0.118)
Baseline (four and more)	-	-	-
Living area			
Rural	-0.532 (0.099)	-0.479 (0.098)	-0.536 (0.093)
Baseline (urban)	-	-	-
Scale	0.742 (0.025)	1.651 (0.103)	0.980 (0.032)
HF	0.596 (0.149)	0.428 (0.093)	0.607 (0.136)
DIC_3	7814.435	7776.615	9011.668

Table 4.4 Bayesian parameter estimates and standard errors for AFT models under gamma latent model.

The results of using gamma and normal models are shown in Tables 4.4 and 4.5, respectively. Also, Table 4.6 shows the results of using a squared normal latent model. In these tables *HF* denotes Heterogeneity Factor, which is the standard error of the random effect b_j. The best model according to the DIC_3 criterion is the log-normal AFT model, with the squared normal latent model. The estimated *HF* in this model is highly significant.

Also, Table 4.6 shows that married persons have shorter unemployment duration than singles and widow(er)s or divorced people, fixing values of the other explanatory variables. The effects of other explanatory variables can be interpreted in a similar manner (see, Section 5).

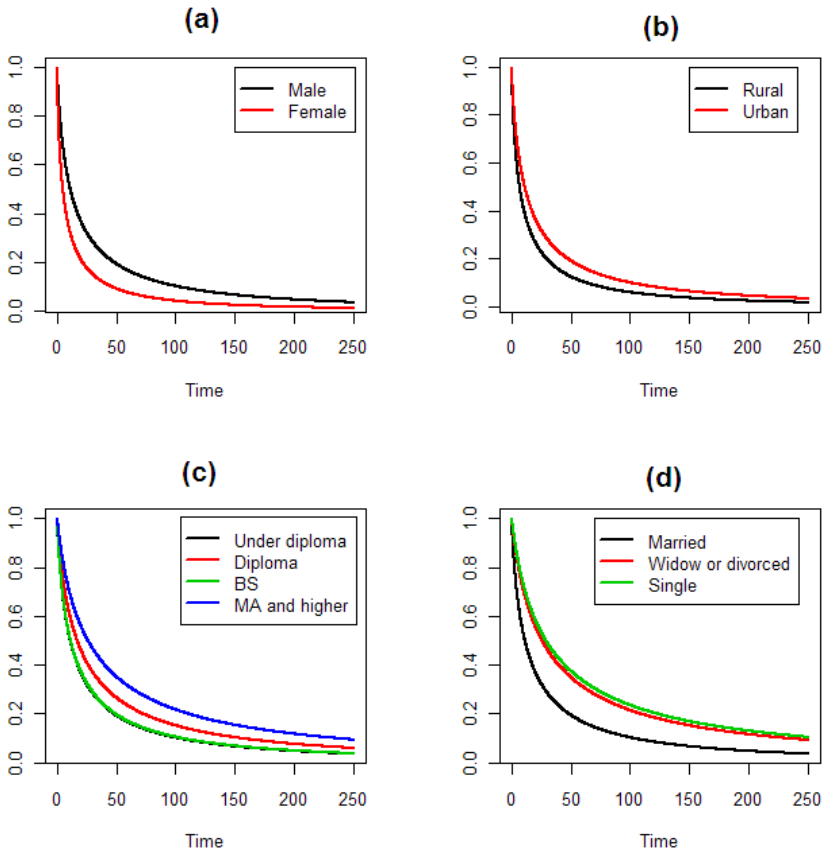
	Log-Logistic	Log-Normal	Weibull
Parameters	Est. (S.E.)	Est. (S.E.)	Est. (S.E.)
Intercept	3.097 (0.265)	3.081 (0.243)	1.835 (0.312)
Gender			
Female	0.824 (0.134)	0.790 (0.128)	0.879 (0.147)
Baseline (male)	-	-	-
Current marital status			
Married	-0.950 (0.118)	-0.904 (0.115)	-0.805 (0.101)
Widow(er) or divorced	-0.088 (0.574)	-0.103 (0.537)	-0.075 (0.632)
Baseline (single)	-	-	-
Education level			
Under diploma	-0.854 (0.175)	-0.807 (0.158)	-0.804 (0.183)
Diploma	-0.472 (0.177)	-0.419 (0.162)	-0.424 (0.189)
BS	-0.881(0.222)	-0.820 (0.202)	-0.813 (0.229)
Baseline (MA and higher)	-	-	-
Age	0.012 (0.007)	0.009 (0.006)	0.012 (0.006)
Number of household members			
One or two	-0.299 (0.195)	-0.267 (0.194)	-0.261 (0.182)
Three	0.025 (0.120)	0.063 (0.123)	-0.099 (0.117)
Baseline (four and more)	-	-	-
Living area			
Rural	-0.519 (0.099)	-0.486 (0.097)	-0.533 (0.095)
Baseline (urban)	-	-	-
Scale	0.742 (0.025)	1.652 (0.101)	0.977 (0.030)
HF	2.516 (0.467)	2.001 (0.302)	5.425 (0.878)
DIC_3	7809.755	7773.382	9015.654

Table 4.5 Bayesian parameter estimates and standard errors for AFT models under normal latent model.

	Log-Logistic	Log-Normal	Weibull
Parameters	Est. (S.E.)	Est. (S.E.)	Est. (S.E.)
Intercept	3.049 (0.268)	3.096 (0.241)	3.440 (0.273)
Gender			
Female	0.830 (0.138)	0.791 (0.129)	0.885 (0.145)
Baseline (male)	-	-	-
Current marital status			
Married	-0.955 (0.119)	-0.899 (0.116)	-0.818 (0.107)
Widow(er) or divorced	-0.163 (0.568)	-0.086 (0.537)	-0.048 (0.652)
Baseline (single)	-	-	-
Education level			
Under diploma	-0.832 (0.180)	-0.811 (0.159)	-0.798 (0.181)
Diploma	-0.446 (0.185)	-0.424 (0.163)	-0.411 (0.185)
BS	-0.856(0.229)	-0.825 (0.204)	-0.797 (0.229)
Baseline (MA and higher)	-	-	-
Age	0.012 (0.006)	0.009 (0.004)	0.013 (0.006)
Number of household members			
One or two	-0.302 (0.200)	-0.271 (0.193)	-0.254 (0.182)
Three	0.020 (0.122)	0.059 (0.124)	-0.089 (0.120)
Baseline (four and more)	-	-	-
Living area			
Rural	-0.526 (0.095)	-0.489 (0.098)	-0.519 (0.092)
Baseline (urban)	-	-	-
Scale	0.742 (0.003)	1.654 (0.103)	1.016 (0.033)
HF	4.425 (0.745)	2.001 (0.302)	4.805 (0.935)
DIC_3	7813.242	7772.309	9008.878

Table 4.6 Bayesian parameter estimates and standard errors for AFT models under squared normal latent model.

Figure 4.4 S($t|\theta,x$) by significant explanatory variables in model (4.13) under the log-normal AFT model with squared normal random effect distribution



One of the most important graphical representations in survival analysis is the marginal survival distribution plot. The marginal survival distribution function is given by:

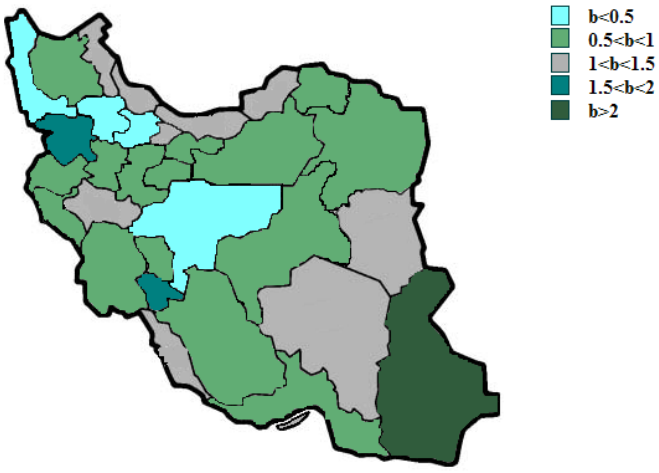
$$S(t;\theta,x) = \int S(t;\theta,b,x)h(b;\nu)db.$$

Figure 4.4 illustrates predicted plots of the marginal survival function for different categories of the explanatory variables in model (4.13) under the best fitting distributional assumption. Panel (a) demonstrates the effect of gender on the marginal survival distribution while holding fixed values of the other explanatory variables: namely, married, living in urban, twenty-nine year old educated to under diploma level. This panel shows that the duration of unemployment with these properties is longer for a female than that for a male. Panel (b) shows the effect of living area on the marginal survival distribution while holding fixed other explanatory variables: a married man, twenty-nine year old educated to under diploma level. Panel (c) shows the effect of education levels on the marginal survival distribution while holding fixed the explanatory variables: a married man, twenty-nine year old educated to under diploma level and a household size greater than three. Panel (d) shows the effect of current marital status on the marginal survival distribution while holding fixed the explanatory variables: a married man, living in urban, and twenty-nine year old.

A comparison of the results for the log-normal AFT model, the best fitting model, in Tables 4.4 to 4.6 can be regarded as a sensitivity analysis of the effect on parameter estimates and model comparison criteria of changing the latent variable distribution. This shows that, for these data points, the results are not sensitive to the choice of latent variable distribution.

As a final component of our discussion on unobserved heterogeneity, Figure 4.5 shows the unobserved heterogeneity of different provinces. On this map, we categorize the posterior mean of the latent variable of the best fitting model for different provinces. Through a close look at this figure, one can imply that the provinces with the same color have similar levels of unobserved heterogeneity. Also, the larger is the unobserved heterogeneity factor, the darker is the color in this figure and the larger is the duration of unemployment. The larger values of b are for the rural areas; for example, Sistan and Baluchestan province (which includes Afghan immigrants) has the largest value of b. This figure shows that most of the provinces have similar values of b [unobserved heterogeneity factor in the interval (0.5,1)].

Figure 4.5 Posterior means of latent random effects obtained by the best fitting model.



5. CONCLUSION

In this paper, we analyzed unemployment duration data of Iran in 2009, containing right and interval-censored observations. Using the WinBUGS software, we adopted a Bayesian approach to handle unobserved heterogeneity.

For unemployment duration, we considered accelerated failure time models with various distributional assumptions selected using proposed R^2 regression indices. A latent exponential family distribution was considered for unobserved heterogeneity. We checked the convergence of the MCMC approach via a Gelman-Rubin diagnostic test. Finally, we compared different distributional assumptions using the DIC_3 criterion. A log-normal AFT model with squared normal latent variable was selected as the best fitting model. Our study demonstrated that it is essential to consider the heterogeneity factor in the modeling of unemployment duration of Iran in 2009.

The results of our proposed method revealed significant difference in unemployment duration based on different explanatory variables: for example, married persons have shorter unemployment durations than singles persons and widow(er)s or divorced people (fixing values of the other explanatory variables). Females have longer unemployment durations than males. Also, in Iran, people with the under diploma have the shortest unemployment durations. Unemployment duration for families with one or two members is shorter than that for families with larger numbers of family members, and people living in urban areas have longer unemployment duration than people living in rural areas.

Using the proposed method provinces with similar levels of unobserved heterogeneity can be identified by inspecting the estimated values of the latent variable.

The proposed method presents the possibility of considering different members of the family of random effects distributions and the use of the available software WinBUGS for model implementation. The proposed approach can be extended for analyzing unemployment data with competing risks, where the individuals may remain unemployed, become employed, or become economically inactive.

REFERENCES

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician*, 46 (3), 167–174.

Campolieti, M. (2001). Bayesian Semiparametric Estimation of Discrete Duration Models: An Application of the Dirichlet Process Prior. *Journal of Applied Economics*, 16, 1–22.

Celeux, G., F. Forbes, C. P. Robert and D. M. Titterington (2006). Deviance information criteria for missing data models. *Bayesian Anaysis*, 1(4), 651–674.

Cox, D. R. (1972). Regression models and life tables. *Journal of Royal Statistical Society B*, 34, 187–220.

DeIorio, M. and C. P. Robert (2002). Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, 64, 629–630.

Deshpande, J. V. and S. G. Purohit (2005). *Life Time Data: Statistical Models and Methods*. World Scientific, Singapore.

Duchateau, L., P. Janssen, P. Lindsey, C. Legrand, R. Nguti and R. Sylvester (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis*. 40 (3), 603–620.

Gelman, A. and D. B. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7, 457–511.

Greene, W. H. (2003). *Econometric Analysis.* 5th edition, Prentice Hall.

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Heckman, J. J. and B. Singer, (1982). Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography, e*d. K. Land and A. Rogers. New York: Academic.

Heckman, J. J. and B. Singer (1984). The identifiability of the proportional hazard model. *Review of Economic Studies*, 51, 231–241.

Ibrahim, J. G., M. H. Chen and D. Sinha (2002). *Bayesian Survival Analysis*. Springer-Verlag, New York.

Knape, J., N. Jonzen, M. Skold, J. Kikkawa and H. McCallum (2011). Individual heterogeneity and senescence in Silvereyes on Heron Island. *Ecology*, 92, 813–820

Komarek, A., E. Lesaffre and C. Legrand (2007). Baseline and treatment effect heterogeneity for survival times between centers using a random effects accelerated failure time model with flexible error distribution. *Statistics in Medicine*, 26, 5457–5472.

Komarek, A., A. Lesaffre and J. F. Hilton (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14, 726–745.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Lancaster, T. (1990). *The Economic Analysis of Transition Data*. Cambridge: Cambridge University Press.

Lee, E. T. and G. W. Wang (2003). *Statistical Methods for Survival Data Analysis. Lifetime Learning Publications*. Wiley, New York.

Legrand, C., V. Ducrocq, P. Janssen, R. Sylvester and L. Duchateau (2005). A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. *Statistics in Medicine*, 24, 3789–3804.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. 2nd edition, New York: Springer.

Omori, Y. and R. A. Johnson (1993). The Influence of Random Effects on the Unconditional Hazard Rate and Survival Function. *Biometrika*, 80, 910–924.

Pan, W. and T. A. Louis (2000). A linear mixed-effects model for multivariate censored data. *Biometrics*, 56, 160–166.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin and A. Lindevan der (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society-Series B*, 64, 583–616.

Spiegelhalter. D. J., A. Thomas, N. Best and D. Lunn (2003). *WinBUGS Examples, MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health*. Imperial College School of Medicine, UK.

Tempelman, R. J. and D. Gianola (1996). A mixed effects model for overdispersed count data in animal breeding. *Biometrics*, 52, 265–279.

Tuma, N. B. and M. T. Hannan (1984). *Social Dynamics: Models and Methods*. New York: Academic.

Vaupel, J. W., K. G. Manton and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–54.

Vermunt, J. K. (2002). A General Latent Class Approach to Unobserved Heterogeneity in the Analysis of Event History Data. In *Applied Latent Class Analysis*. 1st ed. Cambridge:

Cambridge University Press, 383-407. http://dx.doi.org/10.1017/CBO9780511499531.015

Vindenes, Y., S. Engen and B. E. Saether (2008). Individual heterogeneity in vital parameters and demographic stochasticity. *American Naturalist*, 171, 455–467.

Wienke, A. (2011). *Frailty models in survival analysis*. Chapman Hall/CRC.

Yamaguchi, T. and Y. Ohashi (1999). Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. *Statistics in Medicine*, 18, 1961–1971.