Background and Aim: One of the statistical methods used to analyze the time-to-event medical data is survival analysis. In survival models, the response variable is time to the occurrence of an event. The main characteristic of survival data is the existence of censored data. When we have the distribution of survival time, we can use parametric methods. Among the important and popular distributions that can be used, we can mention the Weibull distribution. If the data derives from a heterogeneous population, simple parametric models (such as Weibull) would not fit the data appropriately. One of the methods which have been introduced to overcome this problem is the use of mixture models.

Methods: To assess the validity of the two-component Weibull mixture model, we use a simulation method on heterogeneous survival data. For this purpose, data with different sample sizes were produced in a batch of 1000. Then, the validity of the model is checked using root mean square error (RMSE) criterion.

Results: It is obtained that increasing the sample size would decrease the RMSE in the parameters. However the maximum observed RMSE in all the parameters was negligible.

Conclusion: The Bayesian Weibull mixture model was a proper fit for the heterogeneous survival data.

Introduction:

Survival studies consist of several statistical methods in which the response variable is time to occurrence of an event and is widely used in medicine, economics, biology, and social sciences(1). The main statistical models used in survival analysis are non-parametric and semi-parametric models such as Kaplan-Meyer and Cox(2, 3). Weibull is one of the most popular models in comparison to the others(4). The simple parametric models are useful for homogeneous data. When we are dealing with a heterogeneous survival data for instance, when different treatments are used for subjects or in the studies about cancer where the recurrence of the disease is probable, applying simple parametric models such as Weibull will not have a good fit(5). The reason is that the Weibull model does not consider the heterogeneous characteristics of
the data. One of the methods which have been proposed to overcome this problem is the use of mixture models (5). Since these models take the multi-modal characteristics of the data into account, they can be a proper substitute for the classic models, and as the classic models are a particular case of mixture models, these can also be used when the data is homogenous (6).

The main idea of the mixture models was proposed by Berkson (1952) for the first time (7). Chen et al. (1985) applied the two-component mixture model to analyze survival data (8). Quian (1994) used a semi-Weibull model which consists of a Weibull part and another survival ratio in order to analyze lung cancer data (9). Marin et al. (2005) implemented the Weibull model on right-censored homogenous survival data by using the Bayesian method with an unknown number of components (10). Erisoghloo in 2010 obtained generalized geometric-exponential mixture model as a new method of survival data analysis (11), and in 2012 in order to analyze lung cancer data, he used Gamma, Weibull, and Log-normal mixture models and evaluated the fit of different models on the data (5). Also, in 2014, employed the method of moment logarithm estimation, which was a new method for mixture models, in order to estimate the model parameters (6). Karakoka and Erisoghloo (2015) compared five different methods of estimation, including Maximum Likelihood, Least-Squares, moments, moment logarithm, and percentage

\[
\begin{align*}
\text{h}(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \\
&= \sum_{k=1}^{g} \pi_k \left( \frac{t}{\beta_k} \right)^{\alpha_k-1} \exp \left( -\left( \frac{t}{\beta_k} \right)^{\alpha_k} \right), t > 0, \alpha, \beta > 0
\end{align*}
\]

In this function, \(\alpha\) is the shape parameter, and \(\beta\) is the scale parameter. Weibull mixture model is defined as below

\[
f(t) = \frac{\alpha}{\beta} \left( \frac{t}{\beta} \right)^{\alpha-1} \exp \left( -\left( \frac{t}{\beta} \right)^{\alpha} \right), t > 0, \alpha, \beta > 0
\]

\[
S(t|g, \pi, \alpha, \beta) = \sum_{k=1}^{g} \pi_k \exp \left( -\left( \frac{t}{\beta_k} \right)^{\alpha_k} \right)
\]

In this function, \(g\) is the number of components of the mixture model and \(\pi_k\) is the mixture component parameter and lies between (0,1) satisfying \(\sum_{k=1}^{g} \pi_k = 1\) condition.

Also, the mixture survival function is

\[
S(t|g, \pi, \alpha, \beta) = \sum_{k=1}^{g} \pi_k \exp \left( -\left( \frac{t}{\beta_k} \right)^{\alpha} \right)
\]
Supposing the number of distribution components is g. Data with the possibility of right-censor is denoted by \( x_i = (t_i, \delta_i) \) which \( \delta_i \) is an indicator function that expresses the censoring state of the ith observation and is 
\[
\delta_i = \begin{cases} 
1 & \text{if it is right-censored} \\
0 & \text{if the event has occurred} 
\end{cases}
\]

Considering the above hypothesis, the likelihood function is then defined as
\[
L = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k \left( \frac{\alpha_k}{\exp(\beta_0 + \beta_1 x_i)} \right) \left( \frac{t_i}{\exp(\beta_0 + \beta_1 x_i)} \right)^{\alpha_k - 1} \delta_i \exp \left( - \frac{t_i}{\beta_0 + \beta_1 x_i} \right)^{\alpha_k}
\]

Bayesian estimation:
In the Bayesian estimation of the parameters, we must choose a prior distribution for the parameters. We use the gamma distribution as a prior for the shape parameter, and we use a normal distribution as a prior distribution for the \( \beta_0 \) and \( \beta_1 \) and Dirichlet distribution as the prior distribution for the \( \pi_k \) Mixture component. The prior distributions described belows:
\[
\pi_k \sim \text{Dirichlet}(\varnothing, ..., \varnothing) \\
\alpha_k \sim \text{Gamma}(\alpha_\alpha, \beta_\alpha) \\
\beta_k \sim \text{Gamma}(\alpha_\beta, \beta_\beta)
\]

Simulation
Since this is the first time that we have covariates in the model, to assess the goodness of the model, we use simulation methods. Here we use mixture Weibull model with the following characteristics:
\[
f(t) = 0.5 \cdot \left( \frac{1.3}{e^{-0.69 - 0.2x}} \right)^{0.3} e^{-\left( \frac{t}{e^{-0.69 - 0.2x}} \right)^{1.3}} + 0.5 \cdot \left( \frac{2.4}{e^{0.69 - 0.2x}} \right)^{1.4} e^{-\left( \frac{t}{e^{0.69 - 0.2x}} \right)^{2.4}}
\]

As a result, the desired mixture survival function becomes:
\[
S(t) = 0.5 \left( \frac{1.3}{e^{-0.69 - 0.2x}} \right)^{1.3} + 0.5 \left( \frac{2.4}{e^{0.69 - 0.2x}} \right)^{2.4}
\]

The main idea behind this is based on:
As we know, survival functions take a value between 0 and 1. So in order to make the generated values random, the stages of the simulation are as follow:

\[
L = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k \left( \frac{\alpha_k}{\exp(\beta_0 + \beta_1 x_i)} \right) \left( \frac{t_i}{\exp(\beta_0 + \beta_1 x_i)} \right)^{\alpha_k - 1} \delta_i \exp \left( - \frac{t_i}{\beta_0 + \beta_1 x_i} \right)^{\alpha_k}
\]

Since the aim of this paper has covariates in the model, the scale parameter should be reparametrized in order to estimate the effectiveness of the covariates. Reparametrisation of the scale parameter according to covariates is
\[
\beta_k = \exp(\beta_0 x_k + \beta_1 x_i)
\]
And the likelihood function becomes
\[
\delta_i \exp \left( - \frac{t_i}{\exp(\beta_0 + \beta_1 x_i)} \right)^{\alpha_k}
\]

x: A uniform random number between 0 and 1.

u: A random number from the uniform 0 and 1 distribution.

A random number is generated from each prior distribution and we name it parameters.

The mixture survival function is equated to u, and as the parameter values are generated, the equation is solved according to t.

If the desired sample size is obtained, the process will end; otherwise, we return to stage 1.

Values greater than 4 are equated to 4 and considered as a censored value.

After the data simulation is completed, model parameters are estimated using the Bayesian method.

Model comparison
In order to assess the effectiveness of the model and evaluate the simulation results, the RMSE index is incorporated, which is:
\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\beta} - \beta)^2}{n}}
\]

Results
In this paper, 1000 sample sets of sample sizes of 10, 20, 50, 100 and 200 have been generated, and the model parameters have been estimated, and then the average and RMSE
indices have been calculated which is presented in table 1.

As long as the value of RMSE decreases as the sample size increases from 10, it can be concluded that, as the sample size increases the model operates well; as a result, mixture models in the presence of covariates would be a proper fit to the estimate the effects of the covariates in heterogeneous populations.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>n=10</th>
<th>n=20</th>
<th>n=50</th>
<th>n=100</th>
<th>n=200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alfa1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>1.229</td>
<td>1.097</td>
<td>1.236</td>
<td>1.265</td>
<td>1.308</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.332</td>
<td>0.321</td>
<td>0.299</td>
<td>0.285</td>
<td>0.281</td>
</tr>
<tr>
<td><strong>Alfa2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>2.369</td>
<td>2.441</td>
<td>2.432</td>
<td>2.352</td>
<td>2.448</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.502</td>
<td>0.460</td>
<td>0.439</td>
<td>0.418</td>
<td>0.380</td>
</tr>
<tr>
<td><strong>Beta01</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>-0.688</td>
<td>-0.690</td>
<td>-0.688</td>
<td>-0.687</td>
<td>-0.687</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.030</td>
<td>0.023</td>
<td>0.020</td>
<td>0.015</td>
<td>0.011</td>
</tr>
<tr>
<td><strong>Beta02</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.691</td>
<td>0.688</td>
<td>0.689</td>
<td>0.689</td>
<td>0.688</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.012</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td><strong>Beta11</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>-0.202</td>
<td>-0.201</td>
<td>-0.198</td>
<td>-0.199</td>
<td>-0.201</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.040</td>
<td>0.031</td>
<td>0.021</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td><strong>Beta12</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>-0.202</td>
<td>-0.202</td>
<td>-0.201</td>
<td>-0.199</td>
<td>-0.201</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.012</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.510</td>
<td>0.492</td>
<td>0.505</td>
<td>0.504</td>
<td>0.502</td>
</tr>
<tr>
<td>Rmse</td>
<td>0.015</td>
<td>0.014</td>
<td>0.014</td>
<td>0.013</td>
<td>0.012</td>
</tr>
</tbody>
</table>

**Conclusion**

In this paper, the Bayesian Weibull mixture model was fitted to the heterogeneous survival data with the help of simulation methods and the proposed model was evaluated by Bayesian estimation methods. In result section, according to the model fitting indices, it was obtained that, this model was a proper fit for the heterogeneous survival data.

**References:**