International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 20(2), 2020, pp 225-242

Parametric Regression Models for Analyzing Lifetime Data with Incomplete Covariates Using the EM Algorithm

Md. Rezaul Karim[1*] and M. Ataharul Islam[2]

[1]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh.
Email: mrezakarim@yahoo.com

[2]Institute of Statistical Research and Training, University of Dhaka, Bangladesh. Email: mataharul@yahoo.com

[*]Correspondence should be addressed to Md. Rezaul Karim
(Email: mrezakarim@yahoo.com)

[Received July 28, 2020; Accepted September 16, 2020]

Abstract

The parametric regression models are employed extensively in both reliability and survival analyses for identifying factors or covariates associated with lifetimes of an object. There are situations where the information on some covariates associated with some lifetimes are not available. For example, in the case of warranty claims database, the information on covariates (e.g., failure modes, usage conditions, operating environments, etc.) are known for the objects that fail within the warranty period and are unknown for the censored objects. This article applies the Weibull and lognormal parametric regression models for modeling the lifetimes as a function of covariates. The expectation maximization (EM) algorithm is used to obtain the maximum likelihood estimates of the parameters of the model because of incomplete information on some covariates. An example based on real field data of an automobile component is given to estimate the distribution quantiles and reliability functions at different conditions of covariates for assessing the performance of the component with respect to those covariates. It also traces the best and worst conditions of covariates and recommends that efforts should be concentrated at determining and reducing the risks associated with the root causes of failures in the worst conditions of covariates for improving the reliability of the component.

Keywords: Parametric regression model, Weibull regression model, Lognormal regression model, Warranty claims data, Reliability, Covariates, EM algorithm.

AMS Classification: 62N02, 62N05.

1. Introduction

Reliability and survival analyses are the specialized fields of mathematical statistics and are developed to deal with special type of time-to-event random variables (lifetime, failure time, survival time, etc.). In the case of reliability analysis, our concern is to address the characteristics of survival times of products (item, equipment, component, subsystem, system, etc.), whereas in the case of survival analysis, we address the characteristics of lifetimes arising from problems associated with living organisms (plant, animal, individual, person, patient, etc.) (Karim & Islam, 2019). In this paper by an object we mean item, equipment, component, subsystem, system, etc. among products; and plant, animal, individual, person, patient, etc. among living organisms in an experiment/study.

Reliability of a product conveys the concept of dependability and successful operation or performance. It is a desirable property of great interest to both manufacturers and consumers. The time-to-failure or lifetime of an object is intimately linked to its reliability and this is a characteristic that will vary from system to system even if they are identical in design and structure. For example, if we use the same automobile component in different automobiles and observe their individual failure times, we would not expect them all to have the same failure times. The times to failure for the components used in different automobiles would be different and be defined by a random variable. The behavior of the random variable can be modeled by a probability distribution which is a mathematical description of a random phenomenon consisting of a sample space and a way of assigning probabilities to events. The basis of reliability analysis is to model the lifetime by a suitable probability distribution and to characterize the life behavior through the selected distribution (Karim & Islam, 2019).

A salient feature of modern industrial societies is that new products are appearing on the market at an ever increasing pace. This is due to (i) rapid advances in technology and (ii) increasing demands of customers, with each a driver of the other (Blischke, Karim, & Murthy, 2011). Customers need assurance that a product will perform satisfactorily over its designed useful life. This depends on the reliability of the product, which, in turn, depends on decisions made during the design, development and production of the product. One way that manufacturers can assure customers of satisfactory product performance is through reliability. Automotive manufacturing companies analyze field reliability data to enhance the quality and reliability of their products and to improve customer satisfaction. In recent years, many manufacturers have utilized the warranty database as a prime source of field reliability data, which can be collected economically and efficiently through repair service networks. Warranty claim data are superior to

laboratory test data in the sense that they contain information on the actual environment in which the product is used. Therefore, a number of procedures have been developed for collecting and analyzing warranty claim data, e.g., (Blischke, Karim, & Murthy, 2011) and the references given therein, (Yang, He, & He, 2016), (Khoshkangini, Pashami, & Nowaczyk, 2019), (Khoshkangini, et al., 2020).

There are situations where the lifetime of an object depend on some explanatory variables or covariates, e.g., the lifetime of an automobile depends on the operating environment and the survival time of a patient depends on the treatment condition. If important explanatory variables are ignored in an analysis, it is possible that resulting estimates of quantities of interest (e.g., distribution quantiles or failure probabilities) could be biased seriously (Meeker & Escobar, 1998). Regression analysis is useful for modeling the lifetime as a function of the covariates or predictors. The use of linear regression models assuming normality assumption is very limited in reliability and survival analyses due to the fact that: (i) the lifetime variables are non-negative and skewed and (ii) the relationship between lifetimes and explanatory variables are not directly linear (Karim & Islam, 2019). Due to the nature of the data in reliability and survival analyses, it is not a practical option to use a linear regression model, and the parametric regression models are employed extensively for identifying factors or covariates associated with the lifetime of an object. If we want to know the probability that an object will last longer than a certain lifetime with respect to particular circumstances or characteristics or covariates, the parametric regression models can be applied. However, there are situations where the information on some covariates associated with some lifetimes are not available. For example, in the case of warranty claims database, the information on covariates (e.g., failure modes, usage conditions, operating environments, etc.) are known for the objects that fail within the warranty period and are unknown for the censored objects.

In this paper, an approach is discussed for modelling the reliability of a specific system (unit) of automotive components based on field failure warranty data. The unit's lifetime depends on some explanatory variables or covariates such as the automobile operating environment or used region, types of automobile that use the unit, and the types of failure mode. If a unit fails within the warranty period, the information on covariates can be known from the warranty database; however, such information is unknown for the censored units. The principal aim of the paper is to fit the Weibull and lognormal regression models for the lifetime of the unit which depends upon a vector of categorical covariates and to assess the reliability of the unit as a function of those covariates. The expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) by the method of

weights proposed in (Ibrahim, 1990) is used to estimate the parameters of the models.

The outline of the paper is as follows. Section 2 discusses the parametric regression model for modelling the lifetime variable as a function of covariates. Section 3 presents the parameter estimation procedure using the EM algorithm by the method of weights for incomplete covariates. An example based on real field data of automobile components is given in section 4. Section 5 concludes the paper with a discussion and possible implementation issues for future research.

2. Parametric Regression Model

Let the lifetime random variable denoted by T [and $Y = \log(T)$] depends on a vector of explanatory variables or covariates, X = (X_1, …,X_p). Regression analysis of lifetimes involves specifications for the distribution T, $f(t \mid X, \beta, \sigma)$, for given X, upon which lifetime may depend. The general form of the parametric regression model is

$$Y_{ij} = \mu(X_j) + \sigma\varepsilon_{ij} \tag{1}$$

where Y_{ij} is the i^{th} log lifetime $[Y_{ij} = \log(T_{ij})]$ for a given explanatory variable X_j, σ > 0, and ε_{ij} is a random variable that does not depend on X_j $(j = 1,2,\ldots,p; i = 1,2,\ldots,n_j)$. Expression (1) can be rewritten in terms of matrices so that the model is given by

$$Y = \mu(\mathbf{X}) + \sigma\varepsilon \tag{2}$$

where $\mu(\mathbf{X})$ is a location parameter and σ is a scale parameter. The distribution of T (where $Y = \log T$) depends on the assumed distribution of ε, see, (Kalbfleisch & Prentice, 2002), (Nelson, 1990), and (Lawless, 2003). A variety of functional forms for $\mu(\mathbf{X})$ or ($\alpha(\mathbf{X}) = \exp(\mu(\mathbf{X}))$ in (2) have been proposed, but the most useful form is perhaps the log-linearmodel (Lawless, 1982), for which

$$\mu(\mathbf{X}) = \mathbf{X}\beta \tag{3}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is the $1 \times p$ vector of independent (or *regressor*) variables and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients. This form is comparatively simple to apply and available in many statistical software packages.

Let $\theta = (\beta', \sigma, \gamma')'$ be a $(p+r+1)\times 1$ vector of all parameters in the model, where $\beta = (\beta_1, \ldots, \beta_p)'$ represents a vector of regression coefficients, σ is a scale parameter,

and $\gamma = (\gamma_1, ..., \gamma_r)'$ is the parameter vector associated with the distribution of covariates X, $g(X \mid \gamma)$. The complete-data log-likelihood function based on n independent observations can be written as

$$l_{X,t}(\theta \mid X, t) = \sum_{i=1}^{n} l_{X,t}(\theta \mid X_i, t_i) = \sum_{i=1}^{n} l_{t|X}(\beta, \sigma \mid X_i, t_i) + \sum_{i=1}^{n} l_X(\gamma \mid X_i), \quad (4)$$

where $l_{X,t}(\theta \mid X_i, t_i)$ is the complete-data log-likelihood of θ for the ith observation based on the joint distribution of (X, t); $l_{t|X}(\beta, \sigma \mid X_i, t_i)$ is the log-likelihood based on the conditional distribution of $t|X$; and $l_X(\gamma \mid X_i)$ is the contribution from the marginal distribution of X. Many specifications are possible for $f(t \mid X, \beta, \sigma)$ and $g(X \mid \gamma)$ depending on the problem and the nature of the available data. In this paper we consider the two parameters Weibull and lognormal distributions for $f(t \mid X, \beta, \sigma)$ and the multinomial distribution for $g(X \mid \gamma)$.

2.1. Log-Location-Scale Regression Model

A random variable Y belongs to the location-scale family of distributions if its cumulative density function (cdf) can be expressed as

$$\Pr(Y \le y) = F\left(y \mid \mu(X), \sigma\right) = \Phi\left(\frac{y - \mu(X)}{\sigma}\right), \; y > 0,$$

where $\Phi(z)$ is the cdf of z and it does not depend on any unknown parameters. A random variable T belongs to the log-location-scale family distribution if $Y = \log(T)$ is a member of the location-scale family. The log-location-scale distribution regression model can be expressed as

$$\Pr(T \le t) = F\left(t \mid \mu(X), \sigma\right) = F(t \mid X, \beta, \sigma) = \Phi\left[\frac{\log(t) - \mu(X)}{\sigma}\right], \; t > 0. \quad (5)$$

with location parameter dependent on X, $\mu(X) = \beta' X$, and scale parameter σ does not depend on X. The Weibull, lognormal and loglogistic distributions are the special cases of this model. The quantile function of the model (5)

$$\log\left[t_p(X)\right] = y_p(X) = \mu(X) + \Phi^{-1}(p)\sigma \quad (6)$$

is linear in X. Such a relationship between $t_p(X)$ and X sometimes known as "loglinear relationship". Choosing Φ determines the shape of the distribution for a particular value of X. As mentioned in (Meeker & Escobar, 1998), $\Phi = \Phi_{\text{sev}}$ (smallest extreme value), $\Phi = \Phi_{\text{nor}}$ (normal) and $\Phi = \Phi_{\text{logis}}$ (logistic) are used for Weibull, lognormal and loglogistic distributions, respectively.

The likelihood function for a combination of *n* independent exact-failure and right-censored observations can be written as

$$L(\beta,\sigma)=\prod_{i=1}^{n}\left[\frac{1}{\sigma t_i}\phi\left(\frac{\log(t_i)-\mu_i(X)}{\sigma}\right)\right]^{\delta_i}\left[1-\Phi\left(\frac{\log(t_i)-\mu_i(X)}{\sigma}\right)\right]^{1-\delta_i} \tag{7}$$

where $\phi(z)$ is the probability density function (pdf) of *z*, $\mu_i(X)=\beta' X_i$, δ_i =1 for an exact-failure time and δ_i =0 for a right-censored observation.

2.2. Weibull Regression Model

It is convenient to use a simple alternative parameterization for the Weibull distribution based on the relationship between the Weibull distribution and the smallest extreme value (SEV) distribution. In (5), if we assume Φ as the cdf of the smallest extreme value distribution, i.e., $\Phi=\Phi_{\text{sev}}$, we get the Weibull regression model with location parameter dependent on *X*, $\mu(X)=\beta' X$, and scale parameter σ. Under this model, the density function of *T* given *X* can be written as

$$f(t\mid X,\beta,\sigma)=\frac{1}{\sigma t}\exp\left[\left(\frac{\log(t)-\mu(X)}{\sigma}\right)-\exp\left(\frac{\log(t)-\mu(X)}{\sigma}\right)\right],\ t>0. \tag{8}$$

The survivor function for this model

$$S(t\mid X,\beta,\sigma)=\exp\left[-\exp\left(\frac{\log(t)-\mu(X)}{\sigma}\right)\right],\ t>0\ . \tag{9}$$

Conditional on covariates, for unit *i*, the log-likelihood function for β and σ, $l_{t|X}(\beta,\sigma\mid\delta_i,X_i,t_i)$, can be obtained using (7), (8) and (9). For more detailed explanations of Weibull regression model, see (Meeker & Escobar, 1998), (Lawless, 2003), (Karim & Suzuki, 2007), (Blischke, Karim, & Murthy, 2011), and (Karim & Islam, 2019).

2.3. Lognormal Regression Model

In (5), if we assume Φ as the cdf of the normal distribution, i.e., $\Phi=\Phi_{\text{nor}}$, we get the lognormal regression model with scale parameter $\alpha(\mathbf{X})=\exp(\mu(\mathbf{X}))=\exp(\beta' X)$ dependent on *X*, and shape parameter σ. Under this model, the density function of *T* given *X* can be written as

$$f(t\mid X,\beta,\sigma)=\frac{1}{\sigma t}\phi_{\text{nor}}\left(\frac{\log(t)-\mu(X)}{\sigma}\right)=\frac{1}{\sigma t\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu(X)}{\sigma}\right)^2\right],\ t>0, \tag{10}$$

and the survivor function becomes

$$S(t \mid X, \beta, \sigma) = 1 - \Phi_{\text{nor}}\left(\frac{\log(t) - \mu(X)}{\sigma}\right) = \int_t^{\infty} \frac{1}{\sigma t' \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log(t') - \mu(X)}{\sigma}\right)^2\right] dt', t > 0, \quad (11)$$

where ϕ_{nor} and Φ_{nor} are the pdf and cdf of the standard normal distribution. Conditional on covariates, for unit *i*, the log-likelihood function for β and σ, $l_{t|X}(\beta, \sigma \mid \delta_i, X_i, t_i)$, under this model can be obtained by using (10) and (11) in (7). More details on lognormal regression model can be found, for example, in (Meeker & Escobar, 1998), (Kalbfleisch & Prentice, 2002), (Lawless, 2003), and (Karim & Islam, 2019).

2.4. Distribution of Covariates

Suppose *A* denotes a set of all possible combinations of levels of covariate vectors for any individual, *r* denotes the number of elements in *A*, $\boldsymbol{X}^{(k)}$ denote the *k*th covariate vector in *A* and m_k be the expected count in covariate class *k*, $k = 1, \ldots, r$. As mentioned in (Karim & Suzuki, 2007), the counts of individuals having each of the possible covariate assignments, m_k, are distributed as multinomials (n, $\boldsymbol{\gamma}$), where the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)'$ and γ_k denote the probability that an individual is of covariate type *k*, $k = 1, \ldots, r$. Thus, the log-likelihood for $\boldsymbol{\gamma}$, $l_{\boldsymbol{X}}(\boldsymbol{\gamma}|\mathbf{X})$, can be expressed as

$$l_x(\boldsymbol{\gamma}|\mathbf{X}) = \sum_{i=1}^{n} l_X(\gamma \mid X_i) \propto \sum_{k=1}^{r-1} m_k \log(\gamma_k) + \left(n - \sum_{k=1}^{r-1} m_k\right) \log(\gamma_r) \quad (12)$$

where $n - \sum_{k=1}^{r-1} m_k$ is the expected number of individuals belonging to the last covariate class *r* and $\gamma_r = 1 - \sum_{k=1}^{r-1} \gamma_k$.

3. Parameter Estimation via the EM algorithm

If all the values of response variable t_i and the corresponding covariates X_i are observed for $i = 1, \ldots, n$, then the log-likelihood function for the Weibull regression model can be derived by using Eqs. (7), (8), (9) and (12). Similarly, the log-likelihood function for the lognormal regression model can be derived by using Eqs. (7), (10), (11) and (12). These log-likelihood functions are then maximized separately to obtain the MLEs of $\boldsymbol{\theta}$ for the Weibull and lognormal models. However, we deal with the problem in which for censoring time τ_i, if $t_i \leq \tau_i$, the values of t_i, τ_i, and $\boldsymbol{X}_i$ are observed, whereas if $t_i > \tau_i$, the number of censored units and the value of τ_i are known but $\boldsymbol{X}_i$ are unknown. Section 4 describes in detail the nature of available data. Since it is a problem of incomplete categorical covariates

(Lipsitz & Ibrahim, 1996a), we apply the EM algorithm (Dempster, Laird, & Rubin, 1977) to estimate the parameters of the models. The EM algorithm consists of two steps to iterate: the E-step determines the conditional expectation of complete-data log-likelihood given observed data and the M-step maximizes that conditional expected log-likelihood. The detailed explanation of the EM algorithm can be found in (Dempster, Laird, & Rubin, 1977) and (McLachlan & Krishnan, 1997).

Let $\boldsymbol{X}_i = (\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},\,i})$, where $\boldsymbol{X}_{\text{obs},i}$ and $\boldsymbol{X}_{\text{mis},\,i}$ denote the observed and missing components of $\boldsymbol{X}_i$, respectively. Following references (Ibrahim, 1990) and (Lipsitz & Ibrahim, 1996b), the E-step of the EM algorithm at the (s+1)st iteration can be written as

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)}) &= \sum_{i=1}^{n} Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)} = E\left[l(\boldsymbol{\theta}; t_i, \delta_i, X_i \mid X_{\text{obs},i}, t_i, \delta_i, \boldsymbol{\theta}^{(s)} \right] \\ &= \sum_{i=1}^{n} \sum_{X_{\text{mis},i(j)}} w_{ij}\left(\boldsymbol{\theta}^{(s)}\right) l_{t|X}\left(\beta, \sigma \mid \delta_i, X_i, t_i\right) + \sum_{i=1}^{n} \sum_{X_{\text{mis},i(j)}} w_{ij}\left(\boldsymbol{\theta}^{(s)}\right) l_X\left(\boldsymbol{\gamma} \mid X_i\right) \end{aligned} \tag{13}$$

where$j$ indexes all possible combinations of levels of missing covariates for i, and

$$\begin{aligned} w_{ij}\left(\boldsymbol{\theta}^{(s)}\right) &= \Pr\left[X_{\text{mis},i} \mid X_{\text{obs},i}, t_i, \delta_i, \boldsymbol{\theta}^{(s)} \right] \\ &= \frac{f\left(t_i, \delta_i \mid X_{\text{mis},i}(j), X_{\text{obs},i}, \beta^{(s)}, \sigma^{(s)}\right) g\left(X_{\text{mis},i}(j), X_{\text{obs},i} \mid \boldsymbol{\gamma}^{(s)}\right)}{\sum_{X_{\text{mis},i(j)}} f\left(t_i, \delta_i \mid X_{\text{mis},i}(j), X_{\text{obs},i}, \beta^{(s)}, \sigma^{(s)}\right) g\left(X_{\text{mis},i}(j), X_{\text{obs},i} \mid \boldsymbol{\gamma}^{(s)}\right)} \end{aligned} \tag{14}$$

represent the weights, which can be interpreted as the posterior probabilities of the missing values (Lipsitz & Ibrahim, 1996b). Assume that here n_i new observations have been created for each of the possible missing covariates for observation i, given the response t_i and the observed covariate $X_{\text{obs},i}$. If $N = \sum_{i=1}^{n} n_i$ denotes the total number of new observations, then the double subscripted weights w_{ij} can be replaced with a single subscripted weights, say, v_i, for i =1, …, N (Karim & Suzuki, 2007).

In the M-step, the first term of the expression of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)})$ given in Eq. (13) can be maximized by using any program/package (e.g. S-Plus, R-language) used in failure time regression modeling that allows weights for observations or by applying the Newton-Raphson iterative method to obtain $\hat{\boldsymbol{\beta}}^{(s+1)}$ and $\hat{\sigma}^{(s+1)}$. The maximization of the second term of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)})$ is similar to the maximization of a multinomial likelihood and the parameters are estimated by summing up the expected numbers belonging in each of the r covariate classes dividing by the total

number in all classes; i.e. the parameter $\boldsymbol{\gamma}$ is updated at the (s+1)st iteration in the M-step by

$$\hat{\gamma}_k^{(s+1)} = \frac{m_k^{(s)}}{N}, k = 1, \cdots, r \ . \tag{15}$$

Iterating between the E- and M-steps until they meet a convergence criterion, the EM algorithm finds the MLE of $\boldsymbol{\theta}$. The overall operation of the EM algorithm for estimating the parameters of the models is shown by the block diagram given in Figure 1.

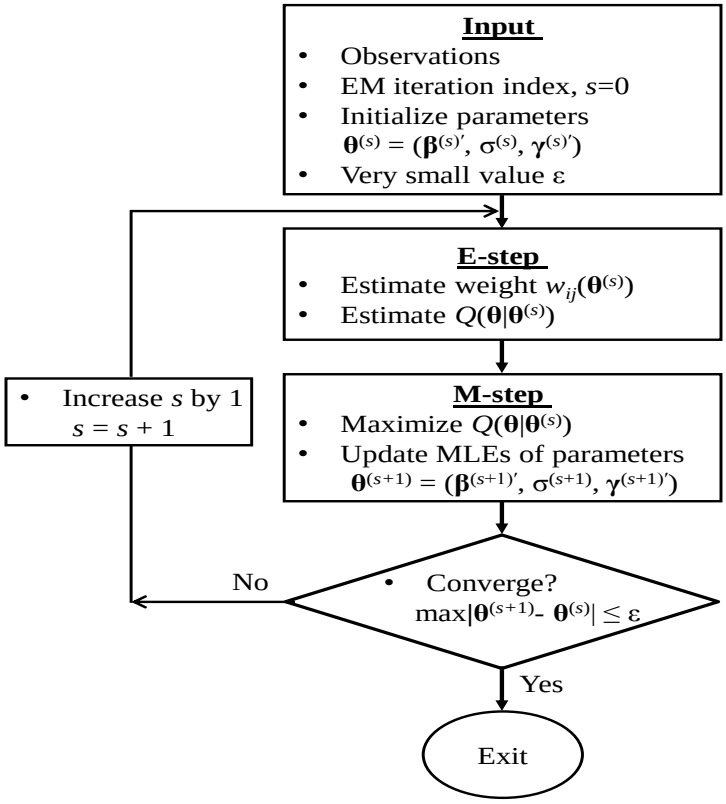

Figure 1: Block diagram of the EM algorithm for estimating the parameters of the models

(Louis, 1982) proposed a method for obtaining the observed information matrix of $\hat{\boldsymbol{\theta}}$ when using the EM algorithm. This method is formulated in (Lipsitz & Ibrahim, 1996a) to apply in the case of missing covariates. The same method applies here to estimate the asymptotic variances of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$.

4. Example

This example illustrates a set of warranty claims data for a specific system (unit/component) of automobile. The units were produced during one and half year, sold during three years, warranty claims were recorded during four years observational period under the one-dimensional warranty of 18 months. A portion of the data frame is given in Table 1, where Age (T) = age of the component, Frequency (d) = frequency of failure or censored items, Indicator (δ) = failure/censored indictor (1 for an exact-failure time and 0 for a right-censored observation), Region (X_1) = component used region [Region1 (R1), Region2 (R2), Region3 (R3), Region4 (R4)], Auto type (X_2) = types of automobile in which the unit is used [Auto1 (A1), Auto2 (A2)], and Failure mode (X_3) = failure modes of the failed units [Mode1 (M1), Mode2 (M2), Mode3 (M3)].

Table 1: A portion of automobile component data frame that used in the Example

No. (i)	Age (T)	Frequency (d)	Indicator (δ)	Region (X_1)	Auto type (X_2)	Failure mode (X_3)
1	t_1	d_1	δ_1=1	R1	A1	M1
2	t_2	d_2	δ_2=0			
3	t_3	d_3	δ_3=1	R2	A2	M3
4	t_4	d_4	δ_4=0			
5	t_5	d_5	δ_5=1	R4	A2	M2
6	t_6	d_6	δ_6=0			

If a unit fails within its warranty period, information on a number of variables including the age of the unit and the corresponding covariates $\boldsymbol{X} = (X_1, X_2, X_3)'$ are recorded in the warranty database. The censoring ages and the age-based number of censored units are calculated based on data provided by the production and sales departments, but the covariate values for censored units are unknown. Examination of the original recorded data revealed a few types of typographical error. After editing the relevant errors, among 126000 sold products 7366 claims were considered for analysis. The information regarding the names of the unit, failure modes, and used regions are not disclosed here to protect the proprietary nature of the information.

Our interest in this example is to investigate how the age-based lifetime (age, T) of the unit differs with respect to three categorical covariates: Region (X_1), Auto (X_2) and Mode (X_3). The number of observed failures belonging in R1, R2, R3,

R4, A1, A2, M1, M2, and M3 are respectively 4202, 668, 387, 2109, 2240, 5126, 2577, 2433, and 2356. Without loss of generality, {Region1 (R1), Auto1 (A1), Mode1 (M1)} is assumed as the reference or baseline level, the level against which other levels are compared. Then the covariate vector $\boldsymbol{X} = (1, X_1, X_2, X_3)'$ can be rewritten as $\boldsymbol{X} = (1, X_{R2}, X_{R3}, X_{R4}, X_{A2}, X_{M2}, X_{M3})'$ under the assumption that all of the six dichotomous covariates $X_{(\,)}$'s take values 1 or 0 to indicate the presence or absence of it. The Weibull and lognormal regression models, discussed in Section 2, are assumed for age T, $f(t, \delta| \boldsymbol{x}, \boldsymbol{\beta}, \sigma)$, with $\mu(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X} = \beta_0 + \beta_{R2}X_{R2} + \beta_{R3}X_{R3} + \beta_{R4}X_{R4} + \beta_{A2}X_{A2} + \beta_{M2}X_{M2} + \beta_{M3}X_{M3}$. The EM algorithm, discussed in the previous section, is applied to estimate the parameters and their asymptotic variances. There are $r = 24$ possible covariate classes and $g(\boldsymbol{X}|\boldsymbol{\gamma})$ is assumed to be a multinomial with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{24})'$. Table 2 and Table 3 summarize the numerical results obtained via the EM algorithm for the Weibull and lognormal regression models, respectively. Programing codes written in R (Web site http://cran.r-project.org/) using the survival library for estimating the parameters of the models.

Table 2: MLEs of parameters for the Weibull regression model

Parameters	MLEs	Std. Error	*Z* value	*p* value	95% Confidence limits	
					Lower limit	Upper limit
β_0	4.3477	0.0223	194.5382	0.00E+00	4.3039	4.3915
β_{R2}	0.4474	0.0416	10.7664	2.69E-26	0.3659	0.5288
β_{R3}	0.4052	0.0574	7.0567	6.13E-12	0.2927	0.5177
β_{R4}	0.0947	0.0201	4.7044	6.24E-06	0.0552	0.1341
β_{A2}	-0.1369	0.0218	-6.2782	1.10E-09	-0.1797	-0.0942
β_{M2}	-0.0117	0.0204	-0.5729	3.39E-01	-0.0518	0.0284
β_{M3}	0.0258	0.0208	1.2439	1.84E-01	-0.0149	0.0666
σ	0.5811	0.0024	-174.45*	0.00E+00	0.5764	0.5858

* Test statistic for H_0: $\sigma = 1$.

Table 3: MLEs of parameters for the lognormal regression model

Parameters	MLEs	Std. Error	*Z* value	*p* value	95% Confidence limits	
					Lower limit	Upper limit
β_0	4.9915	0.0255	195.9747	0.00E+00	4.9416	5.0414
β_{R2}	0.4278	0.0471	9.0824	4.88E-19	0.3355	0.5202
β_{R3}	0.3579	0.0628	5.6944	3.63E-08	0.2347	0.4811
β_{R4}	0.1096	0.0246	4.4649	1.87E-05	0.0615	0.1578
β_{A2}	-0.2643	0.0262	-10.0759	3.59E-23	-0.3157	-0.2129
β_{M2}	-0.0419	0.0248	-1.6928	9.52E-02	-0.0904	0.0066
β_{M3}	0.0244	0.0252	0.9653	2.50E-01	-0.0251	0.0738
σ	1.4000	0.0036	111.4849*	0.00E+00	1.3930	1.4071

* Test statistic for H_0: $\sigma = 1$.

In Table 2 and Table 3, very small p values for all of the regression coefficients, except for β_{M2} and β_{M3} (where $p \geq 0.05$), indicate a strong evidence in favour of the dependency of lifetime on those covariates. Also, the both models reject the hypothesis, H_0: $\sigma = 1$.

A graphical method is applied based on the examination of residuals to assess the adequacy of the distributional assumptions. For the assumed models, the standardized residuals or censored Cox–Snell residuals (Cox & Snell, 1968) is defined in (Meeker & Escobar, 1998, p. 443) as

$$\hat{\varepsilon}_i = \exp\left[\frac{\log(t_i) - \hat{\mu}(X)}{\hat{\sigma}}\right], i = 1, \cdots, n \,. \tag{15}$$

When t_i is a censored observation, the corresponding residual is also censored and can be estimated using the complete data residual summed over the missing data at the last iteration of the EM algorithm, like the estimation of the denominator of weights, $w_{ij}(\boldsymbol{\theta})$ (Karim & Suzuki, 2007). If the model fits the data well, the Cox-Snell residuals should approximately follow a unit exponential distribution, hence the cumulative hazard function of the residuals should be $H_\varepsilon(\hat{\varepsilon}) = \hat{\varepsilon}$ or $-\log[S_\varepsilon(\hat{\varepsilon})] = \hat{\varepsilon}$ (Collett, 2015). Figure 2 shows the plots of the estimated Cox-Snell residuals, $\hat{\varepsilon}$, versus $-\log[S_\varepsilon(\hat{\varepsilon})]$, respectively for the Weibull (left side) and lognormal (right side) regression models.

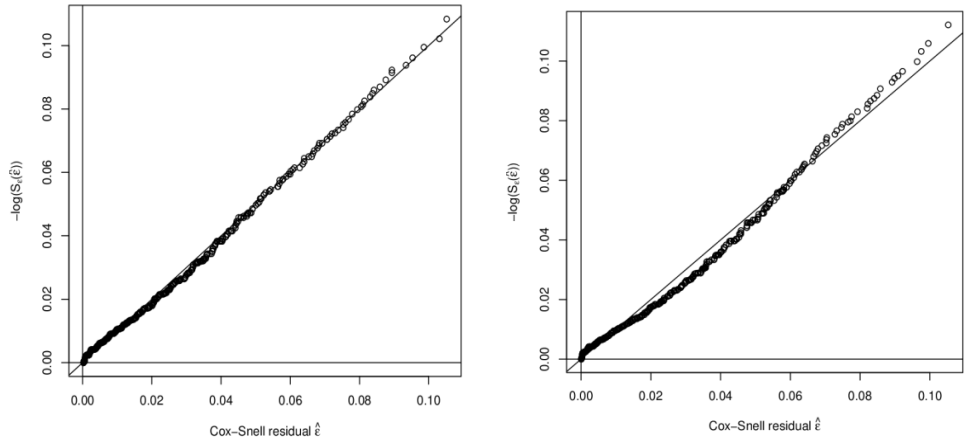

Figure 2: Cox-Snell residual plot (left side for Weibull model, right side for lognormal model)

We see from the Cox-Snell residual plots, Figure 2, that the residuals fall on a straight line through the origin with a slope approximately one for the Weibull model comparatively better than the lognormal model. This indicates that the Weibull model gives a reasonable and better fit to the data than the lognormal model. Also the performance between two models is compared using the Akaike Information Criterion (AIC). The AIC is a measure of the goodness of fit of statistical models that is based on the concept of entropy. The formula for AIC is

$$\text{AIC} = -2(\text{loglikelihood}) + 2(k + c) \,. \tag{16}$$

where k denotes the number of covariates in the model not including the constant terms and c is the number of model-specific distributional parameters. Lower AIC values indicate a better model fit. The AIC values for the Weibull and lognormal models are respectively 132213.6 and 132568.5, indicate that the Weibull model provides better fit than the lognormal model for the observed data. Therefore, the Weibull regression model will be considered in the next analysis. (Karim & Suzuki, 2007) also applied the Weibull regression model for analyzing similar type of warranty claims data and from simulation studies they showed that the EM algorithm is applicable in the case of missing covariates for censored units.

In many applications of Weibull distribution, interest centers on quantiles/ percentiles (or B_p life) rather than the distributional parameters. For example, in the automobile industry of Japan, B10 lifetime (means the 10^{th} percentile) is the most popular reliability index (Suzuki, 1985a). Table 4 shows the maximum likelihood estimates for the pth quantile,

$$\hat{t}_p(X) = \exp\left[\hat{\mu}(X) + \Phi_{\text{sev}}^{-1}(p)\hat{\sigma}\right] \tag{17}$$

for given $\boldsymbol{X}$ (specified levels of covariates) and p (=0.10, 0.25, 0.50, 0.75), where $\Phi_{\text{sev}}^{-1}(p)$ means the pth quantile of the standardized SEV distribution. The $\hat{t}_p(X)$ estimates the lifetime at which 100p% of the sample lies at or below that lifetime for given covariate X. For example, $\hat{t}_{0.5}(X)$ provides the estimated median lifetime of the component for given X. In Table 4, *X_a.b.c* means the covariates {Region = *a*, Auto = *b*, Mode = *c*}.

Table 4 shows that when covariates of used region, auto type, and failure mode are fixed respectively as Region1, Auto1, and Mode1, the ML estimate of $t_{0.10}(X)$ is 20.91. This estimate implies that 10 percent of the units are expected to fail at age 20.91 (the measurement unit is omitted). The estimates of quantiles for other values of p and for different conditions of covariates can be interpreted similarly.

Table 4: MLEs of *p*th quantiles at specific conditions of covariates (*X_a.b.c* means the covariates {Region = *a*, Auto = *b*, Mode = *c*})

No.	Covariates, $\boldsymbol{X}$	The *p*th quantile, $\hat{t}_p(\mathbf{X})$, for given $\boldsymbol{X}$ and *p*			
		$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$
1	X_1.1.1	20.91	37.48	62.47	93.46
2	X_1.1.2	20.66	37.04	61.74	92.37
3	X_1.1.3	21.45	38.46	64.11	95.90
4	X_1.2.1	18.23	32.68	54.48	81.50
5	X_1.2.2	18.02	32.30	53.84	80.55
6	X_1.2.3	18.71	33.54	55.90	83.63
7	X_2.1.1	32.70	58.62	97.72	146.19
8	X_2.1.2	32.32	57.94	96.58	144.48
9	X_2.1.3	33.56	60.16	100.28	150.01
10	X_2.2.1	28.52	51.12	85.22	127.48
11	X_2.2.2	28.19	50.53	84.22	126.00
12	X_2.2.3	29.26	52.46	87.45	130.82
13	X_3.1.1	31.35	56.20	93.68	140.15
14	X_3.1.2	30.99	55.55	92.59	138.51
15	X_3.1.3	32.17	57.67	96.14	143.82
16	X_3.2.1	27.34	49.01	81.70	122.21
17	X_3.2.2	27.02	48.44	80.74	120.79
18	X_3.2.3	28.06	50.29	83.83	125.41
19	X_4.1.1	22.98	41.20	68.68	102.74
20	X_4.1.2	22.72	40.72	67.88	101.54
21	X_4.1.3	23.58	42.28	70.47	105.43
22	X_4.2.1	20.04	35.93	59.89	89.59
23	X_4.2.2	19.81	35.51	59.19	88.55
24	X_4.2.3	20.57	36.87	61.46	91.94

The condition of covariates {Region2, Auto1, and Mode3} is the best as the lifetime is maximum and the condition {Region1, Auto2, and Mode2} is the worst as the lifetime is minimum for the unit among the 24 conditions of covariates. Industrial personnel who are responsible for reliability, safety, and design decisions for the unit are interested to know whether a redesign would be needed to meet the design life specification for specified levels of the covariates. The results in Table 4 would be useful for this requirement.

Figure 3 shows the estimates of reliability function $\hat{R}(t \mid \mathbf{X})$ (or survival function $\hat{S}(t \mid \mathbf{X})$) of the unit for all 24 possible combinations of covariates (left side) and for the best-three and worst-three conditions of covariates (right side).

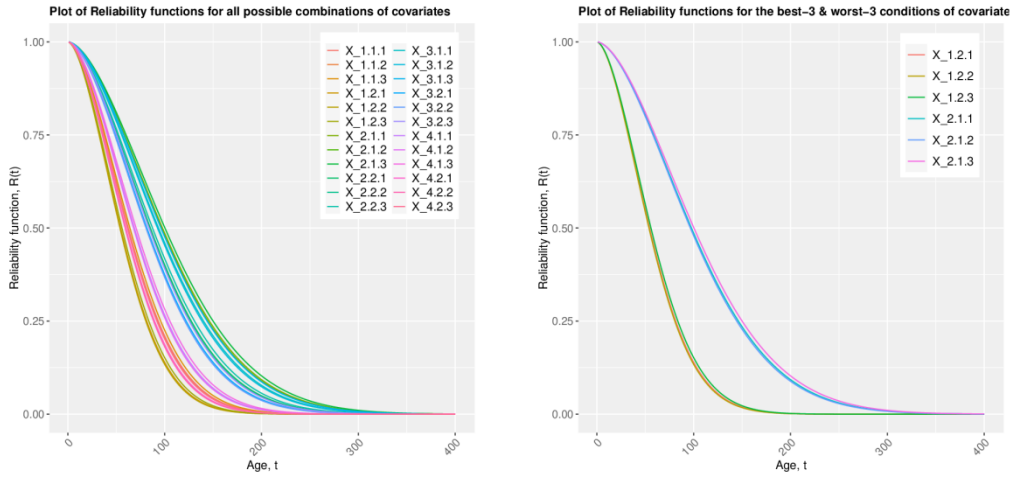

Figure 3: Reliability functions (left side for all 24 possible combinations of covariates, right side for the best-three and worst-three conditions of covariates)

It is a bit complicate to separate and explain the individual reliability function from the left side of Figure 3, therefore, we explain the right side of the figure. The best three conditions of covariates based on the estimated reliability are (i) X_2.1.3 {Region2, Auto1, and Mode3}, (ii) X_2.1.1 {Region2, Auto1, and Mode1} and (iii) X_2.1.2 {Region2, Auto1, and Mode2}. The right side of the figure gives $R_{X\_2.1.3}(t{=}100.28) = 0.50$, indicates the probability that an unit will last longer than age 100.28 under the covariates {Region2, Auto1, and Mode3} is 0.50. Similarly, $R_{X\_2.1.1}(t{=}97.72) = 0.50$ and $R_{X\_2.1.2}(t{=}96.58) = 0.50$. The worst three conditions of covariates based on the estimated reliability are (i) X_1.2.2 {Region1, Auto2, and Mode2}, (ii) X_1.2.1 {Region1, Auto2, and Mode1} and (iii) X_1.2.3 {Region1, Auto2, and Mode3}. It can be seen that $R_{X\_1.2.2}(t{=}53.84) = 0.50$, shows the probability that the unit will survive more than age 53.84 is 0.50. Similarly, we get $R_{X\_1.2.1}(t{=}54.48) = 0.50$ and $R_{X\_1.2.3}(t{=}55.90) = 0.50$.

The median lifetimes at the worst three conditions of covariates are about half compared to the best three conditions of covariates. The Figure 3 suggest that Region2 and Auto1 are the favorable covariates and Region1 and Auto2 are the harshest covariates for the unit. Therefore, to improve the reliability and to reduce

the warranty costs, efforts should be concentrated to improve the unit for surviving in the harshest conditions of covariates {Region1 and Auto2}. That is, the harshest conditions covariates should be the targets of investigation to the manufacturer aimed at determining the root causes of failures and to eliminate or to reduce the risks associated those root causes. Design changes may be needed to protect the component from the harshest environmental effects encountered in the Region1 and Auto2.

The amounts of production and sales for different time intervals are collected from sources other than the warranty claims database of the automobile component considered in the example. The age-based number of censored units are calculated using the amounts of sales and failures. One of the limitations of the data is that there is a possibility to be minor differences between the calculated number of censored units and the exact censored units for some time intervals.

5. Concluding Remarks

The Weibull and lognormal regression models have been considered for modeling the lifetime of an object as a function of covariates. Because of missing covariates for censored items, the EM algorithm is applied to estimate the model parameters and their confidence intervals. A set of warranty claims data of an automobile component is considered as an example. For the example component, the model selection criterion, AIC, and the Cox-Snell residual plot suggested that the Weibull regression model provides a reasonable and better fit to the data than the lognormal regression model.

Estimates of the quantiles and reliability functions of the Weibull model at all possible conditions of covariates were presented for assessing the performance of the component with respect to these covariates. It is observed that the condition of covariates {Region2, Auto1, and Mode3} is the best as the lifetime is maximum and the condition {Region1, Auto2, and Mode2} is the worst as the lifetime is minimum for the component among the 24 conditions of covariates. It is also observed that the Region2 and Auto1 are the favorable and Region1 and Auto2 are the harshest conditions of covariates for the component. To improve the reliability and to reduce the warranty costs, efforts should be concentrated to improve the component for surviving in the harshest conditions of covariates {Region1 and Auto2}. That is, the harshest conditions covariates should be the targets of investigation to the manufacturer aimed at determining the root causes of failures and to eliminate or to reduce the risks associated those root causes.

Design changes may be needed to protect the component from the harshest environmental effects encountered in the Region1 and Auto2.

An important extension of future study is to use modelling of lifetime based on mileage (actual usage measured in km/mile) or to consider the mileage accumulation rate or usage rate (usage per unit of age) as another variable in the model, which would be useful in many applications. However, as suggested in (Suzuki, 1985a), (Suzuki, 1985b), (Lawless, Hu, & Cao, 1995), (Hu, Lawless, & Suzuki, 1998), (Attardi, Guida, & Pulcini, 2005), and (Rai & Singh, 2005) that the usage-based effective estimation requires additional supplementary information about mileage accumulation from sources other than the warranty claims database.

References

[1] Attardi, L., Guida, M.and Pulcini, G. (2005). A mixed-Weibull regression model for the analysis of automotive warranty data. Reliability Engng and System Safety, 87, 265–273.

[2] Blischke, W. R., Karim, M. R.and Murthy, D. N. (2011). Warranty Data Collection and Analysis. Springer-Verlag London Ltd.

[3] Collett, D. (2015). Modelling Survival Data in Medical Research (3rd ed.). Chapman & Hall, London.

[4] Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). Journal of the Royal Statistical Society, A, 30, 248–275.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statist. Soc. B, 39, 1–38.

[6] Hu, X. J., Lawless, J. F. and Suzuki, K. (1998). Nonparametric estimation of a lifetime distribution when censoring times are missing. Technometrics, 40, 3–13.

[7] Ibrahim, J. G. (1990). Incomplete data in generalized linear models. J. Am. Statist. Ass., 85, 765–769.

[8] Kalbfleisch, J. D. and Prentice, R. L. (2002). The statistical Analysis of Failure Time Data (2nd ed.). John Wiley & Sons, Inc., New Jersey.

[9] Karim, M. R. and Islam, M. A. (2019). Reliability and Survival Analysis. Springer Nature Singapore Pte Ltd.

[10] Karim, M. R. and Suzuki, K. (2007). Analysis of Warranty Data with Covariates. Proceedings of the Institute of Mechanical Engineering, Part O, Journal of Risk and Reliability, 221(4), 249-255.

[11] Khoshkangini, R., P., S. M., Berck, P., Gholami S., S., Pashami, S., Nowaczyk, S. and Niklasson, T. (2020). Early Prediction of Quality Issues in Automotive Modern Industry. Information, 11, 354.

[12] Khoshkangini, R., Pashami, S. and Nowaczyk, S. (2019). Warranty Claim Rate Prediction Using Logged Vehicle Data. In Progress in Artificial Intelligence (pp. 663 – 674). Springer.

[13] Lawless, J. F. (1982). Statistical models and methods for lifetime data (1st ed.). Wiley, New York.

[14] Lawless, J. F. (2003). Statistical models and methods for lifetime data (2nd ed.). John Wiley & Sons, Inc., New Jersey.

[15] Lawless, J. F., Hu, X. J. and Cao, J. (1995). Methods for the estimation of failure distributions and rates from automobile warranty data. Lifetime Data Analysis, 1, 227–240.

[16] Lipsitz, S. R. and Ibrahim, J. G. (1996a). Using the EM algorithm for survival data with incomplete categorical covariates. Lifetime Data Analysis, 2, 5–14.

[17] Lipsitz, S. R. and Ibrahim, J. G. (1996b). A conditional model for incomplete covariates in parametric regression models. Biometrika, 83, 916–922.

[18] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. J. R. Statist. Soc. B, 226–233.

[19] McLachlan, G. J. and Krishnan, T. (1997). The EM algorithm and extensions. John Wiley & Sons, Inc., New York.

[20] Meeker, W. Q. and Escobar, L. A. (1998). Statistical Methods for Reliability Data. Wiley Interscience, New York.

[21] Nelson, W. (1990). Accelerated testing: statistical models, test plans, and data analysis. Wiley, New York.

[22] Rai, B. and Singh, N. (2005). A modeling framework for assessing the impact of new time/mileage warranty limits on the number and cost of automobile warranty claims. Reliability Engng and System Safety, 88, 157–169.

[23] Suzuki, K. (1985a). Estimation of lifetime parameters from incomplete field data. Technometrics, 27, 263–271.

[24] Suzuki, K. (1985b). Nonparametric estimation of lifetime distribution from a record of failures and follow-ups. J. Am. Statist. Ass., 80, 68–72.

[25] Yang, D., He, Z. and He, S. (2016). Warranty claims forecasting based on a general imperfect repair model considering usage rate. Reliab Eng Syst Saf, 145, 147-154.