International Journal of Reliability and Applications
Vol. 4, No. 2, pp. 79-95, 2003

Analysis of Marginal Count Failure Data by using Covariates

Md. Rezaul **Karim** *
Department of Statistics, University of Rajshahi
Rajshahi - 6205, BANGLADESH

Kazuyuki **Suzuki**
Department of Systems Engineering
The University of Electro-Communications, Tokyo 182-8585, JAPAN

Abstract. Manufacturers collect and analyze field reliability data to enhance the quality and reliability of their products and to improve customer satisfaction. To reduce the data collecting and maintenance costs, the amount of data maintained for evaluating product quality and reliability should be minimized. With this in mind, some industrial companies assemble warranty databases by gathering data from different sources for a particular time period. This "marginal count failure data" does not provide (i) the number of failures by when the product entered service, (ii) the number of failures by product age, or (iii) information about the effects of the operating season or environment.

This article describes a method for estimating age-based claim rates from marginal count failure data. It uses covariates to identify variations in claims relative to variables such as manufacturing characteristics, time of manufacture, operating season or environment. A Poisson model is presented, and the method is illustrated using warranty claims data for two electrical products.

Key Words : *field reliability data, seasonal effect, MLE, EM algorithm, AIC.*

1. INTRODUCTION

*Corresponding author.
E-mail address: mrkarim@librabd.net

Manufacturers analyze field reliability data to enhance the quality and reliability of their products and to improve customer satisfaction. A prime source of field reliability data is warranty claim data, which is collected economically and efficiently through service networks. For many products, it can be assumed that the warranty is sufficient inducement for customers to report all failures within the warranty period. By collecting and analyzing field reliability data from warranty claims, manufacturers can predict future claims; determine whether a recall, halt in production, or modification is necessary; ascertain whether product reliability is affected by the manufacturing process or usage environment; and compare failure rates among similar or competing products. However, due to the diffuse organizations of service departments or repair service networks and to reduce the data collecting and maintenance costs, many industrial companies construct warranty databases by gathering data from several different departments for particular time periods. For example, they use the monthly sales amounts, N_y, $y = 1, 2, \cdots, Y$, provided by the sales department, as in Table 1, and the number of claims registered for a given month, r_j, $j = 1, 2, \cdots, T$, provided by the service department, as in Table 2.

Table 1. Information from sales department
— monthly sales amounts —

Sales month, y	1	2	$\cdots$	y	$\cdots$	Y
Sales amount, N_y	N_1	N_2	$\cdots$	N_y	$\cdots$	N_Y

Table 2. Information from maintenance department
— monthly number of claims registered —

Recorded month, j	1	2	$\cdots$	j	$\cdots$	T
Number of claims, r_j	r_1	r_2	$\cdots$	r_j	$\cdots$	r_T

Table 3 illustrates the general structure of the data, obtained by combining data from Tables 1 and 2. In Table 3, $\{r_{yt}\}$ be the number of products sold in the yth month which failed after t months (at age t) for $t = 0, 1, \cdots, \min(W-1, T-y)$, where T ($T \geq Y$) is the number of observed months, W is the length of the warranty period, and

$$r_j = \sum_{y=\max(1,j-W+1)}^{\min(j,Y)} r_{y,j-y}, \quad j = 1, 2, \cdots, T,$$

be the count of failures occurring in the jth month. $\{r_j\}$ is called the *marginal count failure data*, which can be observed, and $\{r_{yt}\}$ is the complete data, which can not be observed (see, e.g., Karim, Yamamoto and Suzuki, 2001a & 2001b; and Karim and Suzuki, 2002).

Table 3. General data structure of monthly counted warranty claims

(Left portion of the Table)

		Warranty claims $\{r_{yt}\}$ in a calendar time (month, j)					
y	N_y	1	2	$\cdots$	W	$W+1$	$\cdots$
1	N_1	r_{10}	r_{11}	$\cdots$	$r_{1,W-1}$		
2	N_2		r_{20}	$\cdots$	$r_{2,W-2}$	$r_{2,W-1}$	
$\vdots$	$\vdots$			$\cdots$	$\cdots$	$\cdots$	$\cdots$
y	N_y						
$y+1$	N_{y+1}						
$\vdots$	$\vdots$						
Y	N_Y						
	$\{r_j\}$	r_1	r_2	$\cdots$	r_W	r_{W+1}	$\cdots$

(Right portion of the Table)

		Warranty claims $\{r_{yt}\}$ in a calendar time (month, j)					
y	N_y	y	$y+1$	$\cdots$	Y	$\cdots$	T
1	N_1						
2	N_2						
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y	N_y	r_{y0}	r_{y1}	$\cdots$	$r_{y,Y-y}$	$\cdots$	$r_{y,T-y}$
$y+1$	N_{y+1}		$r_{y+1,0}$	$\cdots$	$r_{y+1,Y-y-1}$	$\cdots$	$r_{y+1,T-y-1}$
$\vdots$	$\vdots$			$\cdots$	$\cdots$	$\cdots$	$\cdots$
Y	N_Y				r_{Y0}	$\cdots$	$r_{Y,T-Y}$
	$\{r_j\}$	r_y	r_{y+1}	$\cdots$	r_Y	$\cdots$	r_T

Note: Y is the total number of months of sale; T is the number of observed months ($Y \leq T$); W is the length of the warranty period; in this Table $W < Y < T$. N_y and $\{r_j\}$ (marginal count) are observed data; $\{r_{yt}\}$ (complete data) are not observed.

Sometimes manufacturers form populations by month of production instead of month of sale. In this case, in Table 1 and the first two columns of Table 3 will replace with the month of production (x) and the number of products produced in month x, M_x, $x = 1, 2, \cdots, X \leq Y$; and $\{r_{xt}\}$ will be the number of products produced in the xth month which failed at age t for $t = 0, 1, \cdots, \min(W-1, T-x)$. This situation is popular when manufacturers want to investigate the effect of some engineering changes such as product design, manufacturing and assembly changes, the performance of the products before and after the engineering changes, etc.

Marginal count aggregation is popular in many industries and arises when companies report only the failure counts for a product rather than unit-wise outcomes, due to limited resources for the collection of such data from worldwide maintenance departments. For example, some companies manufacture electrical consumer products (e.g., personal computers, television, facsimiles, telephones, camcoder, etc.) and collect data on warranty claims for their products through worldwide networks

and report only the regular counts of failures for their products. Masuda, Usui and Suzuki (1999), Karim et al. (2001a & 2001b), Karim and Suzuki (2002) and Suzuki, Karim, and Wang (2001) all discussed the same type of marginal count data for a warranted product.

Manufacturers can collect information on the performance of their products from warranty claim data easily and cheaply. But the information from warranty claim data is sometimes incomplete. A class of problems involving incomplete information can be identified in a warranty database, especially when it is constructed based on marginal count failure data (Karim, 2001). Many factors contribute to product failures that result in warranty claims. The most important factors are the age (the time in service) of the product and the effects of the manufacturing characteristics, time of manufacture and the operating seasons or environments. The age-based (or age-specific) analysis of product failure data has engendered considerable interest in the literature (Kalbfleisch, Lawless, and Robinson, 1991; Kalbfleisch and Lawless, 1996; Lawless, 1998; Karim et al., 2001a & 2001b; and Karim and Suzuki, 2002). The marginal count data contains the aggregated number of failures of different ages, which were produced in different periods and operated in different environments, does not provide directly the age-based number of failures nor the information on the effects of the manufacturing characteristics and operating seasons or environments.

Karim et al. (2001a) developed methods for age-based failure analysis based on marginal count data, without considering the effects of the manufacturing characteristics or operating seasons. That is, they assumed that the expected number of failures per product depends only on the age of the product and is independent of other factors. However, in manufacturing, products may be produced over a period of time, using two or more different raw materials, machines, or environments and operate in different usage environments. Therefore, sometimes the failure distributions differ from period to period, as a result of the effects of manufacturing characteristics or of different usage environments or seasonal effects.

Lindley and Singpurwalla (1986) proposed a multivariate distribution for the life-lengths of the components of a system which operates in random environments. They assumed that the component life-lengths are independent exponential variables and the effect of the operating environments change the original failure rates λ_1 and λ_2 to $\eta\lambda_1$ and $\eta\lambda_2$, where the uncertainty in η is described by a gamma distribution. Li (2000) proposed a method for obtaining a self-consistent estimator of the lifetime distribution using observations from two different environments, where the age-based observations of the unit in both environments are available. Karim et al. (2001b) assumed that due to extraneous causes, a change in the production process may take place at some point in time after which the lifetime distribution of the products is different from the distribution before the change. They investigated a method to detect a single change-point when only the marginal failure counts for the products are available.

This article develops a method for estimating age-based expected number of failures from marginal count failure data. It uses covariates to identify variations in

claims relative to variables such as manufacturing characteristics, time of manufacture, operating season or environment. Poisson log-linear models are employed to analysis warranty claims data for two electrical products. It discusses the maximum likelihood estimation of the mean number of failures per product at age t, and the effect of the manufacturing characteristics and the operating environment on the failure.

The structure of this paper is as follows. The proposed models are discussed in Section 2. Section 3 deals with estimation. Section 4 presents a criterion to select the best approximation model among a set of candidate models. Section 5 analyses two sets of data for marginal count of warranty claims. Finally, Section 6 summarizes the paper and presents concluding remarks.

2. MODELING

Kalbfleisch et al. (1991) and Lawless (1998) modeled the warranty claims using a Poisson model and discussed the age-specific estimation of expected number of claims based on complete data. Karim et al. (2001a, 2001b) used a nonhomogeneous Poisson process to model the failure counts for the repairable products and assumed that for each sales month y, $y = 1, 2, \cdots, Y$, the r_{yt}, $t = 0, 1, \cdots, \min(T - y, W - 1)$ (in Table 3) are independently distributed as Poisson distributions with mean $N_y\lambda_t$, that is,

$$r_{yt} \sim \text{Poisson}(N_y\lambda_t) \tag{2.1}$$

where λ_t is the mean number of failures at age t per product.

Model (2.1) is appropriate when the expected number of age t claims per product does not vary substantially with respect to calendar time or other factors. However, sometimes expected claims may depend on factors such as manufacturing conditions or the environment in which the product is used. In this article we consider different manufacturing characteristics depending on different production periods and assume that the lifetime of the product depends on that characteristics and operating seasons. In this case, covariates are associated with production periods or operating seasons, and a model for the way that the covariates affect the expected claims is postulated. The Poisson model (2.1) could be extended in the usual way to allow the covariate analysis. Poisson regression models for count data such as claims have been widely studied (e.g., Lawless, 1987; Kalbfleisch et al., 1991; and Kalbfleisch and Lawless, 1996).

We suppose that there is a vector of covariates z associated with different groups of products that were produced in different time periods or operated in different seasons. The expected number of claims at age t entering service on month y, $\{r_{yt}\}$, can be conveniently modeled in the log linear form

$$E\{r_{yt}\} = N_y\lambda_t e^{z'\beta} \tag{2.2}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters. Kalbfleisch et al. (1991) and Kalbfleisch and Lawless (1996) mentioned that the specification similar to (2.2) is a very useful way to model covariate effects with current event or count data and the covariate $\boldsymbol{z}$ can specify, for example, manufacturing characteristics or time of manufacture, model lines, and, if they are available, even time-varying factors such as environmental variables.

From model (2.2), we consider two models: Model $\mathcal{M}1$, given is Section 2.1, and Model $\mathcal{M}2$, given is Section 2.2, where $\boldsymbol{z}$ specifies respectively different operating seasons and different production periods.

2.1 Model for the effects of operating seasons

Usually the failure distributions of the products operating in different seasons or environments are distinct. An example of it is that the expected number of failures of many electric products are larger while operating in summer season than the number of failures while operating in winter season. And the difference between the number of failures in two different seasons is large, especially for the products which are used outside of houses, buildings, etc. Here we assume that there are S different seasons or environments denoted by $\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_S$ in a year (in 12 months) and the season $\mathcal{E}_s$ consists of the months from $d_{s-1}+1$ to d_s, where d_s means the sth calendar-month of a year, $d_0 = 0$, $s = 1, 2, \cdots, S \leq 12$ (refer Example 1 in more detail). Throughout this article, the time unit is assumed as one month for both counting and age. Of course the method can be applied to other unit of count intervals, for example, weekly counts or quarterly counts.

We consider the effects of different operating seasons via the regression analysis. Let us define a covariate vector $\boldsymbol{z} = (z_1, z_2, \cdots, z_S)'$ where $z_s = 1$ if the product was operated in season s and 0 if not $(s = 1, 2, \cdots, S)$, and $\boldsymbol{\beta}$ is a $S \times 1$ vector of regression parameters. Let $K_{[j]}$ takes only integer-values, such that $K_{[j]} = \text{trunc}(\frac{j-1}{12}) + 1$ for $j = 1, 2, \cdots, T$, where "trunc(x)" is used to truncate a number x to the next nearest integer towards 0. For $y = 1, 2, \cdots, Y$; $t = 0, 1, \cdots, \min(W-1, T-y)$ and $s = 1, 2, \cdots, S$, if we define

$$\boldsymbol{z}'\boldsymbol{\beta} = \boldsymbol{z}'\boldsymbol{\beta}(y,t) = \begin{cases} \beta_s, & \text{if } 12(K_{[y+t]}-1) + d_{s-1} + 1 \leq (y+t) \leq 12(K_{[y+t]}-1) + d_s \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

then $E(r_{yt})$ in (2.2) can be expressed in the form

$$E(r_{yt}) = m_{yt} = N_y \lambda_t \exp(\boldsymbol{z}'\boldsymbol{\beta}(y,t)), \tag{2.4}$$

and under the Poisson assumption the distribution of r_{yt} is

$$r_{yt} \sim \text{Poisson}(m_{yt}). \tag{2.5}$$

We call this model Model $\mathcal{M}1$. Given the set of independent complete observations $\{r_{yt}\}$ from different operating seasons, the complete data log likelihood can be

written as

$$\log L_c(\lambda_t, \beta; r_{yt}) = \sum_{y=1}^{Y} \sum_{t=0}^{\min(W-1,T-y)} \{-m_{yt} + r_{yt}\log(m_{yt}) - \log(r_{yt}!)\}. \quad (2.6)$$

The problem we address here is to estimate λ_t, $t = 0, 1, \cdots, \min(W-1, T-1)$, and β when the only data available are the marginal counts, $\{r_j\}$, $j = 1, 2, \cdots, T$, rather than $\{r_{yt}\}$ (see Table 3). The sum of independent Poisson random variables is also a Poisson variable, hence under model (2.5), $\{r_j\}$, $j = 1, 2, \cdots, T$, are independently distributed according to Poisson with mean $m_j = \sum_{y=\max(1,j-W+1)}^{\min(Y,j)} m_{y,j-y}$, where $m_{y,j-y}$ is given in (2.4). Therefore, the observed data log likelihood is

$$\log L(\lambda_t, \beta; r_j) = \sum_{j=1}^{T} \{-m_j + r_j \log(m_j) - \log(r_j!)\}. \quad (2.7)$$

In this model $\beta_u > \beta_s$, $\beta_u < \beta_s$, and $\beta_u = \beta_s$ respectively suggest that the operating environment $\mathcal{E}_u$ is a more harsh, a more mild, and the same as that of the another environment $\mathcal{E}_s$, $u = 1, 2, \cdots, S$; $s = 1, 2, \cdots, S$; $u \neq s$. Model $\mathcal{M}1$ included the model presented in Karim et al. (2001a) as a special case when $\beta_s = 0$, $\forall s$. Also if we put $S = 2$, $\log(\beta_1) = \eta$ and $\beta_2 = 0$, Model $\mathcal{M}1$ becomes the model discussed in Karim and Suzuki (2002) where they assumed only two different seasons in a year and the effect of the environment $\mathcal{E}_1$ is to either increase or decrease the parameter λ_t by a common positive factor η. Model $\mathcal{M}1$ is applied in Section 5.1 to analyze a data set of field failure warranty data.

2.2 Model for the effects of manufacturing characteristics

Kalbfleisch et al. (1991) and Kalbfleisch and Lawless (1996) discussed a method with an example where they examined the claims experience of cars manufactured in six distinct production periods. We separate the production period X in V, $(V \leq X)$, different groups, $\mathcal{G}_v$, such that $\{r_{xt}\}$ were produced in group $\mathcal{G}_v$ if $h_{v-1} + 1 \leq x \leq h_v$, with $h_0 = 0$, $v = 1, 2, \cdots, V$; $x = 1, 2, \cdots, X$. Again we define a covariate vector $z = (z_1, z_2, \cdots, z_V)'$ where $z_v = 1$ if the product was produced in group v and 0 if not $(v = 1, 2, \cdots, V)$, and $\boldsymbol{\beta}$ be a $V \times 1$ vector of regression parameters. In this case, if we define

$$z'\boldsymbol{\beta} = z'\boldsymbol{\beta}(x) = \begin{cases} \beta_v, & \text{if } h_{v-1} + 1 \leq x \leq h_v \\ 0, & \text{otherwise} \end{cases} \quad (2.8)$$

for $x = 1, 2, \cdots, X$; $v = 1, 2, \cdots, V$. Then $E(r_{xt})$ can be expressed as

$$E(r_{xt}) = \mu_{xt} = M_x \lambda_t \exp(z'\boldsymbol{\beta}(x)), \quad (2.9)$$

and

$$r_{xt} \sim \text{Poisson}(\mu_{xt}). \tag{2.10}$$

We call this model Model $\mathcal{M}2$. Under this model, the complete data likelihood becomes

$$\log L_c(\lambda_t, \boldsymbol{\beta}; r_{xt}) = \sum_{x=1}^{X} \sum_{t=0}^{\min(W-1,T-x)} \{-\mu_{xt} + r_{xt}\log(\mu_{xt}) - \log(r_{xt}!)\} \tag{2.11}$$

and $\{r_j\}$, $j = 1, 2, \cdots, T$, are independently distributed as Poisson with mean $E(r_j) = \mu_j = \sum_{x=\max(1,j-W+1)}^{\min(X,j)} \mu_{x,j-x}$, and the observed data log likelihood becomes

$$\log L(\lambda_t, \boldsymbol{\beta}; r_j) = \sum_{j=1}^{T} \{-\mu_j + r_j \log(\mu_j) - \log(r_j!)\}. \tag{2.12}$$

This model is illustrated using a real data set in Section 5.2.

We estimate the parameters $\{\lambda_t\}$ and $\boldsymbol{\beta}$ of the Models $\mathcal{M}1$ and $\mathcal{M}2$ by using the maximum likelihood estimation method. Although the observed data likelihoods can be maximized directly by using numerical maximization techniques to obtain the maximum likelihood estimators, but it sometimes give negative estimates for some λ_t. This is because of insufficient data, that is, the sample is not large enough for the consistency result. In contrast, if the EM algorithm is used, the estimates become non-negative (Karim et al., 2001a). The EM algorithm is most effective when maximization of the conditionally expected complete data log likelihood given the observed data is easier than maximization of the observed data log likelihood. But one of the main criticisms of the algorithm is that it is slow to converge. We use the EM algorithm by accounting for non-negativity constraints and by taking advantage of simple maximization procedures.

3. ESTIMATION OF PARAMETERS VIA THE EM ALGORITHM

The EM algorithm is an algorithm to find an MLE for incomplete data problems (Dempster, Laird and Rubin, 1977; McLachlan and Krishnan, 1997). It consists of an E-step, computation of conditional expectation of the log-likelihood function based on the complete data, given the observed data, and an M-step, maximization of that expectation. For the Model $\mathcal{M}1$, the E-step and the M-step are defined as follows:

E-step

Let $\lambda_t^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ be the current estimate of the parameter λ_t and $\boldsymbol{\beta}$ at the kth iteration. At the $(k+1)$th iteration, the E-step computes the conditional expected

log likelihood of the complete-data, given the observed data and current fit $\lambda_t^{(k)}$ and $\beta^{(k)}$,

$$\begin{aligned} Q\left(\lambda_t, \beta|\lambda_t{}^{(k)}, \beta^{(k)}\right) &= \mathrm{E}_{\lambda_t^{(k)},\beta^{(k)}}[\log L_c\left(\lambda_t, \beta; r_{yt}\right)|r_1, r_2, \cdots, r_T] \\ &\propto \sum_{y=1}^{Y} \sum_{t=0}^{\min(W-1,T-y)} \{-N_y \lambda_t \exp(z'\beta(y,t))\} \\ &+ \sum_{y=1}^{Y} \sum_{t=0}^{\min(W-1,T-y)} \left\{\mathrm{E}_{\lambda_t^{(k)},\beta^{(k)}}[r_{yt}|r_{y+t}] \log[N_y \lambda_t \exp(z'\beta(y,t))]\right\} \end{aligned} \tag{3.1}$$

where

$$\mathrm{E}_{\lambda_t^{(k)},\beta^{(k)}}[r_{y,j-y}|r_j] = \frac{r_j N_y \lambda_{j-y}^{(k)} \exp(z'\beta^{(k)}(y, j-y))}{\sum_{y=\max(1,j-W+1)}^{\min(Y,j)} N_y \lambda_{j-y}^{(k)} \exp(z'\beta^{(k)}(y, j-y))}, \; j = 1, 2, \cdots, T \tag{3.2}$$

because the conditional distribution of $\{r_{y,j-y}\}$ given r_j is a multinomial with sample size r_j and probabilities

$$\Pr\{r_{y,j-y}|r_j\} = \frac{N_y \lambda_{j-y}^{(k)} \exp(z'\beta^{(k)}(y, j-y))}{\sum_{y=\max(1,j-W+1)}^{\min(Y,j)} N_y \lambda_{j-y}^{(k)} \exp(z'\beta^{(k)}(y, j-y))}, \tag{3.3}$$

and $z'\beta^{(k)}(y, j-y)$ is obtained from (2.3) by inserting $\beta^{(k)}$ instead of β.

M-step

The M-step maximized the $Q\left(\lambda_t, \beta|\lambda_t{}^{(k)}, \beta^{(k)}\right)$ function. Differentiating $Q\left(\lambda_t, \beta|\lambda_t{}^{(k)}, \beta^{(k)}\right)$ (let us denote it by Q) with respect to λ_t, $t = 0, 1, \cdots, \min(W-1, T-1)$, we find that

$$\frac{\partial Q}{\partial \lambda_t} = \sum_{y=1}^{\min(T-t,Y)} \left\{\frac{\mathrm{E}_{\lambda_t^{(k)},\beta^{(k)}}[r_{yt}|r_{y+t}]}{\hat{\lambda}_t} - N_y \exp(z'\beta(y,t))\right\} = 0. \tag{3.4}$$

Let f_s denote the frequency of a season s occurs within the total observed period T, then

$$f_s = \begin{cases} 1 + \mathrm{trunc}\left(\frac{T}{12+d_{s-1}+1}\right), & \text{if } d_{s-1}+1 \le \min(12, T) \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

where $s = 1, 2, \cdots, S$. The differentiation of Q with respect to β, can be written as

$$\frac{\partial Q}{\partial \beta_s} = \sum_{i=1}^{f_s} \sum_{y=\max(l_{is}-W+1,1)}^{\min(Y,u_{is})} \sum_{t=\max(l_{is}-y,0)}^{\min(W-1,T-y,u_{is}-y)} \{\mathrm{E}_{\lambda_t^{(k)},\beta^{(k)}}[r_{yt}|r_{y+t}]$$

$$-N_y\lambda_t \exp(\boldsymbol{z}'\hat{\boldsymbol{\beta}}(y,t))\} = 0, \tag{3.6}$$

where $l_{is} = 12(i-1) + d_{s-1} + 1$, $u_{is} = 12(i-1) + d_s$ and $s = 1, 2, \cdots, S$.

Within every M-step of the two-step EM algorithm, we seek to maximize (3.1) by solving the estimating equations (3.4) and (3.6). The estimating equations (3.4) and (3.6) indicate that within the iterative procedure of the EM method, the function Q must be maximized using a numerical method. The equations (3.4) and (3.6) are equivalent to (3.2) and (3.3) of Lawless and Nadeau (1995), so the steps leading to solution of the likelihood equations in that paper can be applied with a simple modification to solve (3.4) and (3.6). Below we describe a procedure to solve for the MLE using a combination of two distinct iterative methods, one embedded within the other. Solving equations (3.4) and (3.6), the M-step on the $(k+1)$th iteration leads to

$$\hat{\lambda}_t^{(k+1)} = \frac{\sum_{y=1}^{\min(T-t,Y)} \mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{yt}|r_{y+t}]}{\sum_{y=1}^{\min(T-t,Y)} N_y \exp(\boldsymbol{z}'\boldsymbol{\beta}(y,t))}, \quad t = 0, 1, \cdots, \min(W-1, T-1), \tag{3.7}$$

and

$$\hat{\beta}_s^{(k+1)} = \log\left\{ \frac{\sum_{i=1}^{f_s} \sum_{y=\max(l_{is}-W+1,1)}^{\min(Y,u_{is})} \sum_{t=\max(l_{is}-y,0)}^{\min(W-1,T-y,u_{is}-y)} \mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{yt}|r_{y+t}]}{\sum_{i=1}^{f_s} \sum_{y=\max(l_{is}-W+1,1)}^{\min(Y,u_{is})} \sum_{t=\max(l_{is}-y,0)}^{\min(W-1,T-y,u_{is}-y)} N_y\lambda_t} \right\}, \tag{3.8}$$

$s = 1, 2, \cdots, S$.

One way is to find $\hat{\lambda}_t^{(k+1)}$ and $\hat{\beta}_s^{(k+1)}$ by alternating between (3.7) and (3.8). First estimate the $\hat{\lambda}_t^{(k+1)}$'s from the current $\beta_s^{(k)}$ estimate using (3.7) (step 1) and in (3.8) treat the λ_t's as fixed and solve for $\hat{\beta}_s^{(k+1)}$'s (step 2). Repeated iterations of steps 1 and 2 generally converge to the estimates $\hat{\lambda}_t$ $(t = 0, 1, \cdots, \min(W-1, T-1))$ and $\hat{\beta}_s$ $(s = 1, 2, \cdots, S)$ that satisfy (3.4) and (3.6). Iterating between the E-step and the M-step until it meets a convergence criterion, the EM algorithm finds the MLE of λ_t, $t = 0, 1, \cdots, \min(W-1, T-1)$ and $\boldsymbol{\beta}$.

Similarly, for Model $\mathcal{M}2$ the E-step and the M-step are defined as follows:

E-step

For Model $\mathcal{M}2$ at the $(k+1)$th iteration in the E-step we have to compute

$$\begin{aligned} Q\left(\lambda_t, \boldsymbol{\beta} | \lambda_t{}^{(k)}, \boldsymbol{\beta}^{(k)}\right) \propto & \sum_{x=1}^{X} \sum_{t=0}^{\min(W-1,T-x)} \{-M_x\lambda_t \exp(\boldsymbol{z}'\boldsymbol{\beta}(x))\} \\ & + \sum_{x=1}^{X} \sum_{t=0}^{\min(W-1,T-x)} \left\{ \mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{xt}|r_{x+t}] \log[M_x\lambda_t \exp(\boldsymbol{z}'\boldsymbol{\beta}(x))] \right\} \end{aligned} \tag{3.9}$$

where

$$\mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{x,j-x}|r_j] = \frac{r_j M_x \lambda_{j-x}^{(k)} \exp(\boldsymbol{z}'\boldsymbol{\beta}^{(k)}(x))}{\sum_{x=\max(1,j-W+1)}^{\min(X,j)} M_x \lambda_{j-x}^{(k)} \exp(\boldsymbol{z}'\boldsymbol{\beta}^{(k)}(x))}, \quad j = 1, 2, \cdots, T, \tag{3.10}$$

and $\boldsymbol{z}'\boldsymbol{\beta}^{(k)}(x)$ is obtained from (2.8) by inserting $\boldsymbol{\beta}^{(k)}$ instead of $\boldsymbol{\beta}$.

M-step

At the $(k+1)$th iteration the M-step finds

$$\hat{\lambda}_t^{(k+1)} = \frac{\sum_{x=1}^{\min(T-t,X)} \mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{xt}|r_{x+t}]}{\sum_{x=1}^{\min(T-t,X)} M_x \exp(\boldsymbol{z}'\boldsymbol{\beta}(x))}, \quad t = 0, 1, \cdots, \min(W-1, T-1), \tag{3.11}$$

and

$$\hat{\beta}_v^{(k+1)} = \log\left\{ \frac{\sum_{x=h_{v-1}+1}^{\min(X,h_v)} \sum_{t=0}^{\min(W-1,T-x)} \mathrm{E}_{\lambda_t^{(k)},\boldsymbol{\beta}^{(k)}}[r_{xt}|r_{x+t}]}{\sum_{x=h_{v-1}+1}^{\min(X,h_v)} \sum_{t=0}^{\min(W-1,T-x)} M_x \lambda_t} \right\}, \quad v = 1, 2, \cdots, V. \tag{3.12}$$

The same procedure as used for Model $\mathcal{M}1$ can be applied here to find the MLE of λ_t and $\boldsymbol{\beta}$.

The likelihood-ratio test of hypothesis can be done to test whether there are significant effects of manufacturing characteristics or operating seasons on the products reliability, by setting the hypothesis: $\mathrm{H}_0 : \boldsymbol{\beta} = 0$ against the alternative $\mathrm{H}_1 : \boldsymbol{\beta} \neq 0$. This test can be performed based on the statistic

$$-2 \log\left[L(\hat{\boldsymbol{\theta}}^*; r_1, r_2, \cdots, r_T)/L(\hat{\boldsymbol{\theta}}; r_1, r_2, \cdots, r_T)\right] \tag{3.13}$$

which is asymptotically distributed as chi-square with degree of freedom equals to the number of elements of the vector $\boldsymbol{\beta}$. Here $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$ stand for the unrestricted maximum likelihood estimate of the parameters and that under null hypothesis, respectively.

4. LOCATIONS OF DIFFERENT OPERATING SEASONS AND CHANGE-POINTS

In practice, it is often the case that for some products the locations (beginning month – end month) of different operating seasons, $d_1, d_2, \cdots, d_S$, are unknown. In this case, it is important to determine $d_1, d_2, \cdots, d_S$ based on the available data. One possible way of this determination is to fit the Model $\mathcal{M}1$ by considering all possible locations of d_s, $s = 1, 2, \cdots, S$, and to find the best model among a set

of models. Thus, this method can be formulated within the framework of model selection.

Akaike (1973, 1974) gave an information theoretic interpretation of the likelihood function and proposed it as a model selection criterion, which is known as *Akaike Information Criterion* (AIC), and is defined as

$$\text{AIC} = -2 \times (\text{maximum log-likelihood}) + 2 \times (\text{the number of parameters}). \quad (4.1)$$

AIC has been proved to work well in selecting the best model from a set of models (see, e.g., Sakamoto, Ishiguro, and Kitagawa, 1986; and Burnham and Anderson, 1998). In this problem, the AIC can be applied as follows. Let $\hat{\boldsymbol{\theta}}_{(d_1,d_2,\cdots,d_S)} = (\hat{\lambda}_0, \cdots, \hat{\lambda}_{\min(W-1,T-1)}, \hat{\boldsymbol{\beta}}')'$ be the MLE of $\boldsymbol{\theta}_{(d_1,d_2,\cdots,d_S)} = (\lambda_0, \cdots, \lambda_{\min(W-1,T-1)}, \boldsymbol{\beta}')'$ for the locations $d_1, d_2, \cdots, d_S$, and let $\log L(\hat{\boldsymbol{\theta}}_{(d_1,d_2,\cdots,d_S)})$ be the corresponding maximized value of the log likelihood obtained from (2.7). Model $\mathcal{M}1$ has $\{\min(W,T) + S\}$ parameters, and the AIC becomes

$$\text{AIC}(d_1, d_2, \cdots, d_S) = -2 \log L(\hat{\theta}_{(d_1,d_2,\cdots,d_S)}) + 2\{\min(W,T) + S\}. \quad (4.2)$$

$\text{AIC}(d_1, d_2, \cdots, d_S)$ is minimized over all possible values of $d_1, d_2, \cdots, d_S$ in order to select the best model for the data, and to detect the suitable locations for $d_1, d_2, \cdots, d_S$.

Furthermore, in manufacturing, products may be produced over a period of time, using two or more different raw materials, machines or environments. Due to some changes in the raw materials and/or in the production process, the quality of the products sometimes undergo abrupt changes at some points in time. If the times at which the changes occurred are unknown the situation is known as change-point problem. Model $\mathcal{M}2$ can be employed to explain this problem. We assume that the changes occur at different months of production $h_1, h_2, \cdots, h_V$ for which the quality of the products produced in different production periods, $\mathcal{G}_v$, become different. Model $\mathcal{M}2$ has $\{\min(W,T) + V\}$ parameters, and using the likelihood (2.12) for this model the AIC can be expressed as

$$\text{AIC}(h_1, h_2, \cdots, h_V) = -2 \log L(\hat{\theta}_{(h_1,h_2,\cdots,h_V)}) + 2\{\min(W,T) + V\}. \quad (4.3)$$

The change-points that minimize the $\text{AIC}(h_1, h_2, \cdots, h_V)$ are considered to be the most likely locations for the change-points.

5. APPLICATION OF THE MODELS

5.1 Example 1: Determination of Harsh Environment

In this example, we analyze a data set for the warranty claims of an electrical product given in Karim and Suzuki (2002). The name of the product is not given

to protect proprietary information. The products were sold during 15 month period from September 1996 to November 1997 with a warranty of one year, and the reported number of claims were collected over 19 month period from September 1996 to March 1998. Therefore, $Y = 15$, $T = 19$, $W = 12$. The monthly unit sales numbers (N_y) and monthly claim numbers (r_j) were collected respectively from the sales departments and the maintenance departments, where $(N_1, N_2, \cdots, N_{15}) = (31373, 20326, 16915, 21686, 8172, 17522, 22454, 14696, 5594, 4325, 3405, 3022, 4025, 1359, 1709)$ and $(r_1, r_2, \cdots, r_{19}) = (4, 11, 28, 27, 47, 47, 127, 152, 181, 178, 187, 196, 235, 224, 185, 159, 190, 171, 184)$.

The manufacturer of the products assumed that the expected failures may be higher in summer season than other seasons. The purposes of the analysis of this data set are to investigate whether failures increase in the summer season, that is, to determine the harsh operating season and to estimate the age-based expected number of failures per product. To do this, we fit the Model $\mathcal{M}1$ with two seasons ($S = 2$): Season 1 ($\mathcal{E}_1$) consists of the months from 1 to d_1 and Season 2 ($\mathcal{E}_2$) consists of the months from $d_1 + 1$ to d_2. The model will not be identifiable unless the effect of one season is fixed and known. Without loss of generality, we will assume $\beta_1 \equiv 0$, and will refer to season 1 ($\mathcal{E}_1$) as the "baseline season". Then we fit Model $\mathcal{M}1$ with parameters $\lambda_0, \lambda_1, \cdots, \lambda_{11}$ and $\beta = (0, \beta_2)'$, and calculate the AIC(d_1, d_2) from (4.2) for all possible locations of d_1 and d_2. The calculated values of the first 5 minimum AIC out of the 66 values of AIC are shown in Table 4.. This table shows that the first minimum AIC is attained at $d_1 = 6$ and $d_2 = 10$ and the second minimum AIC attained at $d_1 = 6$ and $d_2 = 9$.

Table 4. $\hat{\beta}_2$ and AIC(d_1, d_2) for different locations of d_1 and d_2

d_1	d_2	$\hat{\beta}_2$	AIC(d_1, d_2)
6	10	0.680	182.435
6	9	0.711	183.648
3	7	0.550	184.075
4	7	0.627	187.581
6	7	0.747	193.970

Note: Only the 5 minimum AIC are shown in this table out of the 66 possible values of AIC.

To test whether there are significant seasonal effects on the products reliability, we set the hypothesis: $H_0 : \beta_2 = 0$ "there is no seasonal effect", against the alternative $H_1 : \beta_2 \neq 0$ "there exists seasonal effect". This test can be performed using the likelihood ratio test based on the statistic (3.13), which is asymptotically distributed as chi-square with one degree of freedom. At the location of the first minimum AIC the calculation of the test statistic in (3.13) is $-2\log\{(-117.610)/(-78.217)\} = 78.79$. Since the critical value for the test with $\alpha = 0.05$ is 3.84, we reject H_0 for the

data set and conclude that the model with seasonal effect is more appropriate. The second minimum AIC attained at $d_1 = 6$ and $d_2 = 9$, and gives the value very close to the first minimum AIC and also reject $H_0 : \beta_2 = 0$.

Therefore, this data set indicates that the harsh environment $\mathcal{E}_2$ was from $d_1+1 = 6+1$ (March) to $d_2 = 10$ (June) or $d_1 + 1 = 6 + 1$ (March) to $d_2 = 9$ (May). The reason against the determination of the months March to May/June as the harsh environment is that in Japan these products are used in the months March to May more times compared with any other months. That is, the usage (the actual operating time) of the product is maximum in the months March to May, and because of the effects of high usage rate the expected numbers of failures in these months are high.

The age-based expected number of failures per product in the operating period March to May is $\exp(\hat{\beta}_2) = 2.036$ times higher than for the another period June to February as: $(\hat{\lambda}_0, \hat{\lambda}_1, \cdots, \hat{\lambda}_{11}) = (0.00013,\ 0.00025,\ 0.00068,\ 0.00034,\ 0.00063,\ 0.00013,\ 0.00089,\ 0.00074,\ 0.00074,\ 0.00274,\ 0.00269,\ 0.00359)$. This result indicates that the expected number of claims are increasing with respect to age.

5.2 Example 2: Determination of Change-points

In this example, we consider data on warranty claims of a home electrical product, discussed by Karim et al. (2001b). The marginal counts of warranty claims were collected over a 30-month period for the product with a warranty of one year. That is, $Y = 30$, $T = 30$, and $W = 12$. The monthly unit sales numbers and monthly claim numbers are given. Karim et al. (2001b) analyzed this data to investigate whether a change unknown to the manufacturer occurred in the raw materials and/or in the manufacturing process and to detect the location of the change, if it has occurred. To investigate the effect of manufacturing characteristics, here we consider the numbers of products produced in month x, (M_x), equal to the sales amounts of that month, (N_x); which in general is not true. The data given in Karim et al. (2001b) are $(M_1, M_2, \cdots, M_{30}) = (1750,\ 490,\ 2098,\ 3295,\ 652,\ 948,\ 1864,\ 1390,\ 1851,\ 1228,\ 1819,\ 2262,\ 2226,\ 2224,\ 2846,\ 6394,\ 1676,\ 2651,\ 3799,\ 2533,\ 2419,\ 2327,\ 4199,\ 2612,\ 3944,\ 3008,\ 3442,\ 5931,\ 2193,\ 3272)$ and $(r_1, r_2, \cdots, r_{30}) = (3,\ 7,\ 10,\ 27,\ 32,\ 25,\ 39,\ 57,\ 69,\ 65,\ 80,\ 117,\ 162,\ 197,\ 268,\ 324,\ 314,\ 276,\ 242,\ 302,\ 293,\ 308,\ 408,\ 381,\ 456,\ 481,\ 463,\ 516,\ 451,\ 331)$.

If the manufacturer can detect the location of the change, they can clarify the cause of the change as well as assess the actual quality and reliability of their products. Karim et al. (2001b) detected a single change-point in this data set. Their method requires a large number of parameters equals $\min(W, T) + \min(W, T - \tau)$, for a change-point at τ, $1 \leq \tau \leq Y$. This number increases if double or multiple change-points are considered and if the dimension of the parameter space is larger than the number of observed variables $\{r_j : j = 1, 2, \cdots, T\}$, the parameters can not be estimated uniquely. The manufacturer assumed that there may exists more than one change-point in the 30-month production period.

Our objective here will be to investigate for more than one change-points, that is, to compare the products produced in different production periods and to find the best locations for the changes. We divide the production period $X = 30$ in $V = 3$ different groups: $\mathcal{G}_1$ from the month 1 to h_1, $\mathcal{G}_2$ from h_1+1 to h_2, and $\mathcal{G}_3$ from h_2+1 to $h_3 \equiv X$. Therefore, $z = (z_1, z_2, z_3)'$ and $\beta = (\beta_1, \beta_2, \beta_3)'$. Again we consider $\beta_1 \equiv 0$, and assume production period for group $\mathcal{G}_1$ as the "baseline production period". Using Model $\mathcal{M}2$ and equation (4.3), we compute AIC(h_1, h_2, h_3) for all possible locations of $\{h_1, h_2, h_3\}$. The first 5 minimum values of AIC out of 406 are given in Table 5.

Table 5. $\hat{\beta}_2$, $\hat{\beta}_3$ and AIC(h_1, h_2, h_3) for different locations of $\{h_1, h_2, h_3\}$

h_1	h_2	$\hat{\beta}_2$	$\hat{\beta}_3$	AIC(h_1, h_2, h_3)
5	11	1.215	1.883	288.622
4	8	1.473	1.076	288.752
4	7	1.550	1.067	289.408
4	6	1.878	1.172	293.137
6	11	1.147	1.796	294.717

Note: Only the 5 minimum AIC are shown in this table out of the 406 possible values of AIC; $d_0 = 0$ & $d_3 = 30$.

This table shows that the first minimum AIC is attained at $\{h_1 = 5, h_2 = 11, h_3 = 30\}$. This indicates that three groups $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ consist of the production months 1 to 5, 6 to 11 and 12 to 30 respectively. The age-based expected number of failures per product for the baseline production period (group $\mathcal{G}_1$) is as follows: $(\hat{\lambda}_0, \hat{\lambda}_1, \cdots, \hat{\lambda}_{11})$ = (0.004534, 0.004610, 0.002032, 0.000003, 0.000278, 0.000062, 0.000195, 0.000606, 0.001132, 0.002301, 0.002698, 0.001608). The estimates $\hat{\beta}_2 = 1.215$ and $\hat{\beta}_3 = 1.883$ suggest that the expected failures for group $\mathcal{G}_2$ (months 6 to 11) and for group $\mathcal{G}_3$ (months 12 to 30) are substantially higher than that of group $\mathcal{G}_1$.

6. CONCLUDING REMARKS

This paper proposed a method for analyzing marginal count failure data by using covariates associated with different groups based on production periods and operating seasons. Poisson log-linear models are applied to model the failure counts and the EM algorithm is used to obtain the maximum likelihood estimates of the mean number of failures per product at age t, and the effects of the operating seasons or the manufacturing characteristics. A method is proposed for detecting the locations of different operating environments and change-points occurred in the production process. The detection method is formulated within the framework of model selection, thus the Akaike Information Criterion (AIC) is utilized for this

purpose. An advantage of the proposed method is that it needs only the maximized log-likelihood function and does not require any complicated distribution theory. The method was applied to two data sets from two warranty databases of electrical products.

An extension of the method considering the sales lag and reporting delay would be useful. Also an area for further investigation is to extend the theoretical properties of the method. A more detailed investigation on the properties and applications of the method is expected to be performed in future research.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. in *Proc. 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds.), Akademiai Kiado, Budapest, pp. 267-281.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automatic Control* **AC-19**, 716-723.

Burnham, K. P. and Anderson, D. R. (1998). *Model selection and inference : a practical information-theoretic approach*, New York : Springer.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, **39** pp. 1-38.

Kalbfleisch, J. D. and Lawless, J. F. (1996). Statistical analysis of warranty claims data. Chapter 9. In *Product Warranty Handbook*, Eds. W.R. Blischke and D.N.P. Murthy. New York: Marcel Dekker.

Kalbfleisch, J. D., Lawless, J. F and Robinson, J. A. (1991). Methods for the analysis and prediction of warranty claims, *Technometrics*, **33**, 273-285.

Karim, M. R. (2001). Statistical analysis of failure data based on marginal counts, *Ph.D. Dissertation*, The University of Electro-Communications, Tokyo, Japan.

Karim, M. R. and Suzuki, K. (2002). Analysis of marginal count failure data under different environmental conditions, *Journal of the Indian Statistical Association*, (in press).

Karim, M. R., Yamamoto, W. and Suzuki, K. (2001a). Statistical analysis of marginal count failure data, *Lifetime Data Analysis*, **7**, 2, 173-186.

Karim, M. R., Yamamoto, W. and Suzuki, K. (2001b). Change-point detection from marginal count failure data, *Journal of the Japanese Society for Quality Control*, **31**, 318-338.

Lawless, J. F. (1987). Regression methods for Poisson process data, *Journal of American Statistical Association*, 82, 808–815.

Lawless, J. F. (1998). Statistical analysis of product warranty data, *International Statistical Review*, 66, 1, 41-60.

Lawless, J.F. and Nadeau, J.C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events, *Technometrics*, **37**, 158-168.

Li, L. (2000). Estimation of distribution functions using data from different environments. *Lifetime Data Analysis*, 3, 171-279.

Lindley. D. V. and Singpurwalla, N. D. (1986). Multivariate distributions for the life lengths of components of a system sharing a common environment, *J. Appl. Prob.*, 23, 418–431.

Masuda, A, Usui, M. and Suzuki, K. (1999). Analysis of product life based on limited product history, *Journal of Reliability Engineering Association of Japan*, 21, 3, 122-130 (in Japanese).

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

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike information criterion statistics*, D. Reidel Publishing Co., KTK Scientific Publishers, Tokyo.

Suzuki, K., Karim, M. R. and Wang, L. (2001). Statistical Analysis of Reliability Warranty Data, *Handbook of Statistics: Advances in Reliability*, **20**, eds. N. Balakrishnan and C. R. Rao, Elsevier Science, 585-609.