Journal of the Korean Institute of Industrial Engineers
Vol. 25, No. 4, pp. 510-516, December 1999.

설명변수를 고려한 불완전 사용현장데이터 분석

오영석[1] · 최인수[2] · 배도선[1]

[1]한국과학기술원 산업공학과 / [2]㈜한국패널리서치, 1997년도 한국학술진흥재단 박사 후 연수과정지원으로 수행

Analyses of Incomplete Field Data with Covariates

Young-Seok Oh[1] · In-Su Choi[2] · Do-Sun. Bai[1]

This paper proposes methods of estimating lifetime distribution from incomplete field data under parametric regression models. Failure-record data-failure times and covariates-reported to the manufacturer can be seriously incomplete for satisfactory inference since only reported failures are recorded. This paper assumes that within-warranty data are reported with probability p_1 (≤ 1) and after-warranty data are reported with p_2 ($< p_1$). Methods of obtaining pseudo maximum likelihood estimators(PMLEs) are outlined, their asymptotic properties are studied, and specific formulas for Weibull distribution are obtained. Simulation studies are performed to investigate the effects of follow-up percentage on the PMLEs.

1. 서 론

제품의 수명과 관련된 데이터들은 실험실에서 수행되는 수명시험(life testing)이나 가속수명시험(accelerated life testing)을 통해 얻어지는 것이 보통이다. 그러나 수명시험이나 가속수명시험은 시험시간을 단축하기 위하여 실제 사용조건보다 더 열악한 조건에서 수행되는 것이 보통이므로, 사용현장에서의 제품수명에 대한 정보를 왜곡하여 나타낼 수 있는 위험이 있다. 따라서, 이러한 위험을 줄이고 제품의 수명에 대한 올바른 정보를 얻어내기 위하여는 실제 사용현장에서 얻어진 수명데이터 즉, 사용현장데이터(field data)를 얻고 이를 분석하는 방법들을 연구하는 것이 매우 중요하다.

사용현장데이터는 사용현장에서 제품에 고장이 발생했을 때 서비스센터를 통하여 얻어지는 데이터이다. 즉, 제품이 고장나면 소비자는 보증수리를 받기 위해 서비스센터를 찾게 되고, 이로부터 고장시간, 제조특성, 사용환경 등과 같은 고장에 관련된 정보가 얻어지게 된다. 이러한 사용현장데이터의 중요한 특징중의 하나는 서비스센터에 들어온 제품의 수명만을 알 수 있고, 고장난 제품이라 할지라도 서비스센터에 들어오지 않으면 그 제품의 수명에 대한 정보를 얻을 수 없다는 것이다. 사용현장데이터의 분석에서는 이와 같은 문제점을 해결하기 위해, 서비스센터에 들어온 제품뿐 아니라 들어오지 않은 제품에 대한 정보도 추적조사(follow-up)를 하거나 소비자로부터 설문조사 형식으로 얻어내어 제품의 수명을 추정하는데 사용한다.

사용환경에서 얻어진 고장데이터의 분석에 관한 기존연구로는 Lawless(1983), Suzuki(1985a, 1985b), Kalbfleish와 Lawless (1988), Kalbfleish 등(1991), Lawless(1994), 배도선 등(1995a, 1995b), Lawless 등(1995), Hu와 Lawless(1996), Hu 등(1998) 등이 있다. 이 중에서 특히 Suzuki(1985)는 사용현장데이터와 총 판매제품 중 일정 비율을 샘플하여 그 중 고장이 나지 않은 제품을 추적조사하여 얻은 관측중단시간 데이터를 이용하여 제품의 신뢰도를 추정하는 문제를 다루었다. Kalbfleisch와 Lawless (1988)는 제품의 수명에 영향을 주는 제품의 제조특성, 환경특성 등과 같은 설명변수(covariate)가 존재할 때, 수명분포의 모수와 설명변수가 대수선형 관계임을 가정하고 사용현장데이터와 총 판매제품 중 보증기간 내에 고장이 발생하지 않은 제품의 일정 비율을 추적조사하여 얻은 설명변수에 대한 데이터를 이용하여 수명분포의 모수를 추정하였다. 배도선 등(1995a, 1995b)은 각각 Suzuki(1985), Kalbfleisch와 Lawless(1988)의 연구를 다수고장원인모형과 다수고장원인이 있고 수리 가능한 모형으로 확장하였다.

사용현장데이터의 또 하나의 특징은 보증기간 안에 고장난 제품은 보증제도의 적용을 받으므로 대부분 서비스센터에 들어오지만, 보증기간 이후에 고장난 제품들은 보증제도의 적용

을 받지 못하므로 소비자의 성향에 따라 i)서비스센터를 찾거나 ii)다른 수리업체를 이용하거나 또는 iii)제품을 폐기처분 하므로 고장난 제품중의 일부만이 서비스센터에 들어오게 된다는 것이다.

사용현장데이터의 분석에 관한 기존의 연구들은 사용현장데이터의 이러한 특징을 고려하지 않고 고장난 제품은 모두 서비스센터에 들어온다고 가정하고 있다. 최근 들어 대부분의 회사가 서비스센터의 규모를 확충하고 보증기간 이후에 고장난 제품에 대해서도 저렴한 가격으로 제품이나 부품을 수리 또는 교체하고 있어 보증기간 이후의 사용현장데이터도 비교적 쉽게 얻을 수 있게 되었다. 그러나 실제 사용현장에서는 이들 데이터에 대한 마땅한 분석방법이 없어 제품의 수명에 관련된 품질특성값을 정확히 구하지 못하고 있고, 신제품의 개발이나 제품의 신뢰성 보증 등에 이용되어야 할 귀중한 데이터들이 사장되고 있는 실정이다.

이 논문에서는 이러한 현실적인 필요성에 근거하여 제품의 제조특성 및 환경특성 등을 나타내는 설명변수가 있을 때, 보증기간 이후의 분석시점까지 얻어지는 사용현장데이터와 분석시점에서 추적조사를 통하여 얻게 되는 추적조사 데이터를 이용하여 제품 수명분포의 모수를 추정하는 문제를 다룬다.

이 논문에서 사용되는 기호는 다음과 같다.

N 총 판매 제품수

θ 수명 분포의 모수 벡터

x_i 제품 i 의 설명변수벡터 $(=(1, x_{i1}, \cdots, x_{ik}))$

$f(t_i \mid x_i ; \theta)$ 제품 i 의 설명변수가 x_i일 때 수명 t_i 의 확률밀도함수

$S(t_i \mid x_i ; \theta)$ 제품 i 의 설명변수가 x_i일 때 수명 t_i 의 신뢰도함수

τ_1 제품의 보증시점

τ_2 제품 수명의 분석시점

p_1 보증시점 τ_1 안에 고장난 제품이 서비스센터에 들어올 확률

p_2 τ_1과 τ_2 사이에서 고장난 제품이 서비스센터에 들어올 확률

p_f τ_2까지 서비스센터에 들어오지 않은 제품 중 추적조사되는 제품의 비율

D_1 N 개의 제품 중 τ_1 이전에 서비스센터에 들어온 제품의 집합

D_2 τ_1과 τ_2 사이에서 서비스센터에 들어온 제품의 집합

D_3 τ_2까지 서비스센터에 들어오지 않은 제품의 집합

A_1 추적조사 제품 중 τ_1 이전에 이미 고장난 제품의 집합

A_2 추적조사 제품 중 τ_1과 τ_2 사이에서 이미 고장난 제품의 집합

A_3 추적조사 제품 중 τ_2까지 고장나지 않은 제품의 집합

n_j 집합 D_j에 속한 제품의 수, $j=1,2,3$

a_j 집합 A_j에 속한 제품의 수, $j=1,2,3$

2. 데이터의 형태 및 추적조사 방법

제품이 고장나면 소비자는 보증수리를 받기 위하여 서비스센터를 찾게 되고, 이 때 제품의 고장시간과 제조모형, 제조시간 또는 제조장소 등의 제조특성(manufacturing characteristics) 및 사용자의 특징, 사용조건 등의 환경특성(environmental characteristics)과 같은 설명변수를 얻게 된다. 이와 같이 서비스센터에 들어온 제품으로부터는 고장시간과 설명변수를 얻을 수 있으나, 그렇지 않은 제품의 경우에는 고장여부나 설명변수를 얻을 수 없게 된다. 이때 고장데이터만을 이용하여 수명분포의 모수를 추정할 경우 일치추정량(consistent estimator)을 얻을 수 없으므로(Suzuki, 1985a) 서비스센터에 들어오지 않은 제품에 대한 추가적인 정보를 얻어내는 추적조사를 한다.

서비스센터에 들어오지 않은 제품들은 고장났으나 서비스센터에 들어오지 않은 제품들과 실제로 고장나지 않은 제품들로 이루어져 있으므로, 추적조사를 통하여 얻는 데이터의 형태는 전자의 경우 고장시간과 설명변수이고, 후자의 경우 분석시점까지 고장나지 않았다는 정보와 설명변수이다. 따라서, 제품의 수명분포의 모수를 추정하는 데 사용하는 사용현장데이터는 i) 보증기간 내에 서비스센터에 들어온 제품의 고장시간과 설명변수, ii) 보증시점과 분석시점 사이에 얻어진 고장시간과 설명변수, iii) 추적조사시 보증시점 전에 고장난 제품의 고장시간과 설명변수, iv) 추적조사 시 보증시점과 분석시점 사이에 고장난 제품의 고장시간과 설명변수, v) 추적조사시 고장나지 않은 제품의 설명변수로 구성된다.

추적조사되는 제품을 샘플링하는 방법은 여러 가지가 있으나, 이 논문에서는 서비스센터에 들어오지 않은 제품 중 p_f의 비율로 단순 랜덤 샘플링하는 Kalbfleisch와 Lawless(1988)의 접근 방법을 사용한다. 이러한 추적조사 방법을 사용했을 때의 사용현장데이터의 취득 과정은 <그림 1>과 같다.

데이터의 형태와 추적조사 방법에 따른 가정은 다음과 같다.

가정

① 제품의 보증기간과 고장 시간의 척도는 모두 달력시간이다.

② 고장난 제품이 서비스센터에 들어올 때까지의 시간은 무시할 수 있다.

③ 서비스센터에 들어온 제품의 고장 시간은 정확히 기록된다.

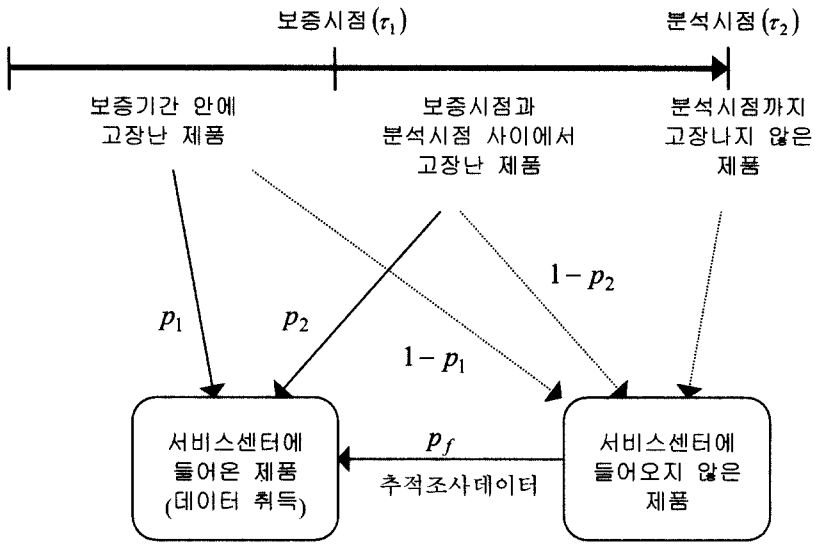

그림 1. 추적조사가 있는 경우 사용현장데이터의 취득 과정.

④ 고장난 제품이 서비스센터에 들어올 보고확률은, 제품의 고장 시간이 보증기간 이내인 경우에는 $p_1(\leq 1)$이고, 보증시점과 분석시점 사이인 경우에는 $p_2(< p_1)$로 시점에 관계없이 일정하다.

⑤ p_1과 p_2는 알려져 있다. 실제로 기업에서는 과거의 경험이나 조사를 통하여 보증기간 전 또는 후에 고장난 제품 중 서비스센터에 들어오는 제품의 비율들을 알 수 있는 경우가 많다.

⑥ 추적조사되는 제품은 랜덤하게 선택되며, 선택된 제품의 설명변수에 대한 정보는 정확히 얻을 수 있다.

⑦ 추적조사 비율 p_f는 0이 아니다.

3. 모수의 추정

이 절에서는 제2절에서 설명한 방법에 의해 수집된 사용현장데이터를 이용하여 제품 수명분포의 모수를 추정하기 위한 의사(pseudo)우도함수를 세우고 의사최우추정량과 이의 점근분산을 구한다.

3.1 의사우도함수

분석시점 τ_2까지 고장난 모든 제품에 대하여 고장시간 t_i를 알고 전체 제품의 설명변수 x_i 들을 안다면, 우도함수 $L_1(\theta)$는

$$L_1(\theta) = \prod_{i:t_i \leq \tau_2} f(t_i \mid x_i;\theta) \prod_{i:t_i > \tau_2} S(\tau_2 \mid x_i;\theta) \qquad (1)$$

가 된다. 또한, 설명변수 x_i 들은 모두 알지만 서비스센터에 들어온 제품에 대한 고장시간만을 안다고 할 때의 우도함수 $L_2(\theta)$는

$$L_2(\theta) = \prod_{i\in D_1} p_1 f(t_i \mid x_i;\theta) \prod_{i\in D_2} p_2 f(t_i \mid x_i;\theta) \prod_{i\in D_3} [(1-p_1)+(p_1-p_2)S(\tau_1 \mid x_i;\theta) + p_2 S(\tau_2 \mid x_i;\theta)] \quad (2)$$

이다. 그러나 모든 제품에 대하여 설명변수를 안다는 것은 현실적으로 불가능하므로, 이 논문에서는 고장나서 서비스센터에 들어왔거나 추적조사되는 제품에 대해서만 고장시간과 설명변수를 알 수 있다고 한다. 따라서, 분석시점까지 서비스센터에 들어오지 않은 제품 중 추적조사되지 않은 제품의 정보는 추적조사로부터 얻은 정보를 이용하여 구하여야 한다.

추적조사를 하면 서비스센터에 들어오지 않은 제품들의 고장데이터와 관측중단데이터중의 일부를 얻을 수 있다. 보증시점 이전과 이후에 얻어진 추적고장데이터의 우도함수에 대한 기여는 각각 $(1-p_1)f(t_i \mid x_i;\theta)$와 $(1-p_2)f(t_i \mid x_i;\theta)$이고, 고장나지 않은 경우에는 $S(\tau_2 \mid x_i;\theta)$가 된다. 그리고 추적조사되지 않은 제품의 경우에는 추적조사를 통하여 얻은 제품의 설명변수 x_i 들을 이용하여 $[(1-p_1)+(p_1-p_2)S(\tau_1 \mid x_i;\theta) + p_2 S(\tau_2 \mid x_i;\theta)]^{\frac{1-p_f}{p_f}}$로 추정할 수 있으므로, D_3에 속하는 제품들의 우도함수에 대한 기여는 추적조사로부터 얻은 정보를 이용하여

$$\prod_{i\in A_1} (1-p_1) f(t_i \mid x_i;\theta) \prod_{i\in A_2} (1-p_2) f(t_i \mid x_i;\theta) \prod_{i\in A_3} S(\tau_2 \mid x_i;\theta)$$

$$\prod_{i\in A_1\cup A_2\cup A_3}[(1-p_1)+(p_1-p_2)S(\tau_1\mid x_i;\theta)+p_2S(\tau_2\mid x_i;\theta)]^{\frac{1-p_f}{p_f}} \tag{3}$$

와 같이 구한다. 이때 $\frac{(1-p_f)}{p_f}$는 $\{n_3-(a_1+a_2+a_3)\}$ / $(a_1+a_2+a_3)$로 서비스센터에 들어오지도 않고 추적조사도 되지 않은 제품의 수와 추적조사된 제품 수의 비를 나타낸다. 이와 같이 추적조사된 부분적인 정보를 이용하여 추적조사되지 않은 나머지 제품들에 대한 정보를 추정하여 구하면 일반적인 우도함수가 아닌 의사우도함수가 된다. 식 (2)에 식 (3)을 대입하여 로그를 취하면 다음과 같은 모수 θ의 의사대수우도함수를 얻을 수 있다.

$$\begin{aligned}\log L_p(\theta) &= n_1\log p_1+n_2\log p_1\\ &+\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}[\log(f(t_i\mid x_i;\theta))]\\ &+a_1\log(1-p_1)+a_2\log(1-p_2)\\ &+\sum_{i\in A_3}\log S(\tau_2\mid x_i;\theta)+\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\log P_i \end{aligned}\tag{4}$$

여기서

$$P_i=(1-p_1)+(p_1-p_2)S(\tau_1\mid x_i;\theta)+p_2S(\tau_2\mid x_i;\theta)$$

이다.

3.2 의사최우추정량 및 추정량의 점근 성질

식(4)를 모수벡터 θ에 대하여 일차편미분한 식 $\partial\log L_p(\theta)/\partial\theta$이 0이 되도록 하는 θ^*를 구하면 이때의 θ^*가 의사최우추정량이 되며, θ^*는 다음과 같은 성질을 갖는다.

[정리 1]

(i) 의사최우추정량 θ^*는 모수 θ의 일치추정량이다.

(ii) N이 커짐에 따라 $\sqrt{N}(\theta^*-\theta)$는 평균이 0이고 분산-공분산 행렬이 $V(\theta)=A(\theta)^{-1}+A(\theta)^{-1}C(\theta)A(\theta)^{-1}$인 다변량 정규분포를 따른다. 여기서 행렬 $A(\theta)$와 $C(\theta)$의 원소 $A(\theta)_{r,s}$와 $C(\theta)_{r,s}$는 각각 다음과 같다.

$$A(\theta)_{r,s}=\lim_{N\to\infty}\frac{1}{N}E\left[-\frac{\partial^2\log L_p(\theta)}{\partial\theta_r\partial\theta_s}\right], \tag{5}$$

$$C(\theta)_{r,s}=\frac{1-p_f}{p_f}\lim_{N\to\infty}\frac{1}{N}E\left[\frac{n_3}{a_1+a_2+a_3-1}\sum_{i\in A_1\cup A_2\cup A_3}(l_{ir}-\overline{l_r})(l_{is}-\overline{l_s})\right], \tag{6}$$

여기서,

$$l_{ir}=\frac{\partial}{\partial\theta_r},\quad \overline{l_r}=\sum_{i\in A_1\cup A_2\cup A_3}\frac{l_{ir}}{a_1+a_2+a_3}$$

이다.

정리1에 대한 증명은 부록에서 다룬다. 식(5)와 (6)의 $A(\theta)_{r,s}$와 $C(\theta)_{r,s}$ 대신 다음에 정의되는 $A_N(\theta^*)_{r,s}$와 $C_N(\theta^*)_{r,s}$를 대입하면 $V(\theta)$의 일치추정량 $V_N(\theta^*)$를 얻을 수 있다.

$$A_N(\theta^*)_{r,s}=-\frac{1}{N}\frac{\partial^2\log L_p(\theta)}{\partial\theta_r\partial\theta_s}, \tag{7}$$

$$C_N(\theta^*)_{r,s}=\frac{n_3(1-p_f)}{Np_f(a_1+a_2+a_3-1)}\sum_{i\in A_1\cup A_2\cup A_3}(l_{ir}-\overline{l_r})(l_{is}-\overline{l_s})'. \tag{8}$$

4. 와이블분포의 경우

제품의 수명이 와이블분포를 따르고 척도모수(scale parameter) α_i가 설명변수 x_i와 대수선형관계 ($\log\alpha_i=x_i'\beta$)를 갖는 경우, 모수가 $\beta=(\beta_1,\cdots\beta_k)'$와 δ인 제품 수명의 확률밀도함수와 신뢰도함수는 각각

$$f(t_i\mid x_i;\beta,\delta)=\delta t_i^{\delta-1}\exp(x_i'\beta)\exp(-t_i^{\delta}e^{x_i'\beta}),\quad t_i>0 \tag{9}$$

$$S(t_i\mid x_i;\beta,\delta)=\exp(-t_i^{\delta}e^{x_i'\beta}),\ t_i>0 \tag{10}$$

이다. 식 (9)와 (10)을 식 (4)에 대입하여 의사대수우도함수를 구하면

$$\begin{aligned}\log L_p(\beta,\delta) &= \sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{x_i'\beta+\log\delta\\ &\quad+(\delta-1)\log t_i-t_i^{\delta}e^{x_i'\beta}\}\\ &+n_1\log p_1+n_2\log p_1+a_1\log(1-p_1)\\ &+a_2\log(1-p_2)-\sum_{i\in A_3}\tau_2^{\delta}e^{x_i'\beta}\\ &+\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\log P_i \end{aligned}\tag{11}$$

이고, 여기서 $P_i=(1-p_1)+(p_1-p_2)\exp(-\tau_1^{\delta}e^{x_i\beta})$ $+p_2\exp(-\tau_2^{\delta}e^{x_i\beta})$이다. 식(11)을 모수 $\beta_r,\ r=1,\cdots,k$과 δ에 대하여 각각 일차 편미분하면

$$\frac{\partial \log L_p}{\partial \beta_r}=\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{x_{ir}(1-W_{i0})\},$$

$$-\sum_{i\in A_3}x_{ir}\tau_2^{\delta}e^{x_i\beta}+\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\left\{\frac{P_{i\beta_r}}{P_i}\right\} \quad (12a)$$

$$\frac{\partial \log L_p}{\partial \delta}=\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{\delta^{-1}+\log t_i-W_{i1}\}$$

$$-\sum_{i\in A_3}\tau_2^{\delta}e^{x_i\beta}\log\tau_2+\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\left\{\frac{P_{i\delta}}{P_i}\right\}, \quad (12b)$$

이고, 여기서

$$P_{i\beta_r}=\frac{\partial P_i}{\partial \beta_r}=-x_{ir}e^{x_i\beta}\{(p_1-p_2)\tau_1^{\delta}\exp(-\tau_1^{\delta}e^{x_i\beta})$$
$$+p_2\tau_2^{\delta}\exp(-\tau_2^{\delta}e^{x_i\beta})\},$$

$$P_{i\delta}=\frac{\partial P_i}{\partial \delta}=-e^{x_i\beta}\{(p_1-p_2)\tau_1^{\delta}\exp(-\tau_1^{\delta}e^{x_i\beta})\log\tau_1$$
$$+p_2\tau_2^{\delta}\exp(-\tau_2^{\delta}e^{x_i\beta})\log\tau_2\},$$
$$W_{ij}=(\log t_i)^j t_i^{\delta}e^{x_i\beta},\ i=1,\cdots,N, j=0,1,2,$$

이다. 식 (12)를 동시에 0으로 하는 $\beta=(\beta_1,\cdots\beta_k)'$와 δ의 의사최우추정치 $\beta^*=(\beta_1^*,\cdots,\beta_k^*)'$와 δ^*는 쉽게 구할 수 없으므로 Newton-Raphson방법과 같은 수치적 방법을 이용하여 구한다.

의사최우추정량 $\beta^*=(\beta_1^*,\cdots,\beta_k^*)'$와 δ^*의 점근분산을 구하기 위해 식 (7)과 (8)로 부터 행렬 $NA_N(\beta^*,\delta^*)$와 $NC_N(\beta^*,\delta^*)$의 원소를 구하면

$$NA_N(\beta^*,\delta^*)_{r,s}=-\frac{\partial^2 \log L_p}{\partial\beta_r\partial\beta_s}$$

$$=\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{x_{ir}x_{is}W_{i0}\}+\sum_{i\in A_3}x_{ir}x_{is}\tau_2^{\delta}e^{x_i\beta},$$

$$-\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\left[\frac{P_{i\beta_r\beta_s}P_i-P_{i\beta_r}P_{i\beta_s}}{P_i^2}\right] \quad (13a)$$

$$NA_N(\beta^*,\delta^*)_{r,k+1}=-\frac{\partial^2 \log L_p}{\partial\beta_r\partial\delta}$$

$$=\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{x_{ir}W_{i1}\}+\sum_{i\in A_3}x_{ir}\tau_2^{\delta}e^{x_i\beta}\log\tau_2,$$

$$-\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\left[\frac{P_{i\beta_r}P_i-P_{i\beta_r}P_{i\delta}}{P_i^2}\right] \quad (13b)$$

$$NA_N(\beta^*,\delta^*)_{k+1,k+1}=-\frac{\partial^2 \log L_p}{\partial\delta^2}$$

$$=\sum_{i\in D_1\cup D_2\cup A_1\cup A_2}\{\delta^{-2}+W_{i2}\}+\sum_{i\in A_3}\tau_2^{\delta}e^{x_i\beta}(\log\tau_2)^2$$

$$-\frac{1-p_f}{p_f}\sum_{i\in A_1\cup A_2\cup A_3}\left[\frac{P_{i\delta\delta}P_i-P_{i\delta}^2}{P_i^2}\right] \quad (13c)$$

$$NC_N(\beta^*,\delta^*)_{r,s}=\frac{n_3(1-p_f)}{p_f(a_1+a_2+a_3-1)}$$
$$\sum_{i\in A_1\cup A_2\cup A_3}(l_{ir}-\overline{l_r})(l_{is}-\overline{l_s}) \quad (14)$$

이다. 여기서,

$$l_{ir}=\frac{\partial}{\partial\beta_r}\log P_i,$$

$$\overline{l_r}=\sum_{i\in A_1\cup A_2\cup A_3}\frac{l_{ir}}{a_1+a_2+a_3},$$

$$P_{i\beta_r\beta_s}=\frac{\partial^2 P_i}{\partial\beta_r\partial\beta_s}$$
$$=x_{ir}x_{is}e^{x_i\beta}\{(p_1-p_2)K_{i1}+p_2K_{i2}\},$$

$$P_{i\beta_r\delta}=\frac{\partial^2 P_i}{\partial\beta_r\partial\delta}$$
$$=x_{ir}e^{x_i\beta}\{(p_1-p_2)K_{i1}\log\tau_1+p_2K_{i2}\log\tau_2\},$$

$$P_{i\delta\delta}=\frac{\partial^2 P_i}{\partial\delta^2}$$

$$=e^{x_i\beta}\{(p_1-p_2)K_{i1}(\log\tau_1)^2+p_2K_{i2}(\log\tau_2)^2\},$$
$$i=1,\cdots,N,\quad r=1,\cdots,k,\quad s=1,\cdots,k,$$

이고, 이때 $K_{ij}=\tau_j^{\delta}\exp(-\tau_j^{\delta}e^{x_i\beta})\{\tau_j^{\delta}\exp^{x_i\beta}-1\}$이다. 식 (13)과 (14)를 이용하면 의사최우추정량의 점근분산을 구할 수 있다.

[예제]

어떤 전자제품이 사용되는 환경이 두 가지로 분류될 수 있고, 이때 사용환경을 나타내는 설명변수 x_i는 0과 1의 값을 가진다. 총 분석제품 수 $N=1,000$으로 이중에 500대는 설명변수가 0의 값을, 나머지 500대는 설명변수가 1의 값을 가지고, x_i가 주어졌을 때 고장시간 t_i는 신뢰도함수가

$$S(t_i \mid x_i;\beta_0,\beta_1)=\exp\{-t_i^{\delta}\exp(\beta_0+\beta_1 x_i)\},\ t_i>0,$$

인 와이블분포를 따른다고 하자. 이때 $\delta=2.0$, β_1은 설명변수의 값이 '1'인 환경이 '0'인 경우에 비하여 고장률이 2배가 되도록 $\beta_1=\log 2=0.6931$로, β_0는 보증기간 12개월 안에 전체 제품 중 고장나는 제품의 비율이 약 5%가 되도록 $\beta_0=-8.3424$로 하였다. $p_1=0.9$, $p_2=0.5$로 가정한 후, IMSL 서브루틴(1997)으로부터 고장데이터를 얻었다. 보증기간 동안 서비스센터에 들어온 전자제품의 수 n_1은 46대로 설명변수가 0인 것이 21대, 설명변수가 1인 것이 25대였다. 분석시점이 24개월일 때, 보증시점과 분석시점 사이에 서비스센터에 들어온 제품의 수 n_2는 64대로 설명변수가 0인 것이 22대, 설명변수가 1인 것이 42대였다. 분석시점 전에 고장났지만 서비스센터에 들어오지 않은 제품과 실제로 분석시점까지 고장나지 않은 제품의 수 n_3는 878대이다. 추적조사비율 $p_f=0.2$로 하고 서비스센터에 들어오지 않은 890대 중에서 178대를 랜덤샘플하여 추적조사 데이터를 얻었다. 추적조사 데이터 중 고장난 것은 15대이고 고장나지 않은 것은 163대였다. 얻어진 고장데이터와 추적조사데이터로 식 (12)의 의사최우추정치를 구해보면

$$\beta_0^*=-8.2952,\quad \beta_1^*=0.5527,\quad \delta^*=2.0188,$$

이고, $V(\theta^*)$는 식 (13)과 (14)로부터

$$V(\theta^*)=\begin{bmatrix} 313.6796 & -27.5242 & -92.2774 \\ -27.5242 & 41.7817 & 0.7347 \\ -92.2774 & 0.7347 & 29.1069 \end{bmatrix}$$

이다. 총 판매제품 수 N이 1,000이므로 의사최우추정치 β_0^*, β_1^*와 δ^*의 점근분산은 각각

$Asvar(\beta_0^*)=0.3137$, $Asvar(\beta_1^*)=0.04178$, $Asvar(\delta^*)=0.02911$

이고, β_0^*, β_1^*와 δ^*는 점근적으로 정규분포를 따르므로 이들에 대한 95% 신뢰구간을 구하면

$$-9.3930\le\beta_0\le-7.1975,$$
$$0.1520\le\beta_1\le0.9533,$$
$$1.6844\le\delta\le2.3532$$

이 된다. 추정된 모수로 제품이 사용되고 있는 환경, 즉 설명변수 x_i의 값에 따라 10%의 제품이 고장나는 시점인 10백분위수의 추정치(단위: 월)를 구해보면 다음과 같다.

표 1. 보고확률과 추적조사비율에 따른 10백분위수의 효율(%)

	$p_1=0.5,\ p_2=0.1$		$p_1=0.5,\ p_2=0.5$	
p_f	$x_i=0$	$x_i=1$	$x_i=0$	$x_i=1$
0.05	13.68	19.80	31.65	22.71
0.1	18.22	27.01	38.09	32.81
0.2	27.07	40.19	42.73	43.58
0.3	37.82	45.36	57.94	51.64
0.5	57.08	58.33	72.14	62.70
0.7	73.58	75.74	82.47	77.11
1.0	100.00	100.00	100.00	100.00
	$p_1=0.9,\ p_2=0.1$		$p_1=0.9,\ p_2=0.5$	
p_f	$x_i=0$	$x_i=1$	$x_i=0$	$x_i=1$
0.05	19.33	26.62	37.53	26.68
0.1	24.71	37.72	46.22	41.48
0.2	34.80	49.11	50.67	54.48
0.3	44.61	56.08	65.89	61.26
0.5	60.64	68.34	77.71	71.52
0.7	76.21	79.60	85.33	81.26
1.0	100.00	100.00	100.00	100.00

설명변수	$t_{0,1}$의 참값	$t_{0,1}^*$
0	21.0314	19.9706
1	14.8718	15.1880

위 표로부터 설명변수 x_i의 값이 0인 환경보다 1인 환경에서 고장이 더 자주 발생함을 알 수 있다. 만약 실제 상황에서 이런 결과가 얻어졌다면 설명변수 1이 의미하는 환경에서 제품이 사용될 경우 제품의 신뢰성을 높이기 위한 방법이 강구되어야 할 것이다.

5. 추적조사비율에 따른 효과

보고확률과 추적조사비율이 의사최우추정량에 미치는 효과를 살펴보기 위하여 예제와 동일한 상황과 모수를 이용하여 p_1이 0.5와 0.9이고 p_2가 0.1과 0.5인 경우, p_f=1.0일 때 얻

어진 10백분위수의 추정값들의 평균제곱오차와 p_f가 0.05, 0.1,0.2,0.3,0.5,0.7,1.0일 때 얻어진 10백분위수의 추정값들의 평균제곱오차의 비로써 추적조사비율에 따른 효율을 구하였다. 모의실험 절차는 배도선 등(1995a)과 유사하게 수행되었다.

<표 1>은 이와 같은 모의실험의 결과로써 수명분포가 와이블분포인 경우 추적조사비율과 보고확률에 따른 10백분위수에 대한 의사최우추정값의 효율을 나타낸 것이다. <표 1>에 의하면 추적조사비율의 증가에 따라 효율이 증가한다는 것과 동일한 추적조사비율에서는 보고확률이 클수록 효율이 높다는 것을 알 수 있다. 또한 보고확률이 큰 경우 추적조사비율이 크지 않더라도 의사최우추정량의 효율이 크게 떨어지지 않음을 알 수 있다.

6. 결 론

이 논문에서는 제품의 환경특성을 나타내는 설명변수가 있는 경우의 사용현장데이터의 분석에 관한 Kalbfleish와 Lawless (1988)의 연구를 데이터가 보증기간 이전 및 이후에 불완전하게 들어오는 상황으로 확장하여 보다 현실성 있는 분석방법을 제시하였다. 제품의 모수가 설명변수와 대수선형관계를 이룬다는 가정하에 의사우도함수를 세운 후 의사최우추정량과 이의 점근성질을 유도하였고, 이를 와이블분포인 경우에 적용하였다. 또한 모의실험을 통하여 보고확률과 추적조사비율이 의사최우추정량에 미치는 효과를 조사한 결과, 보고확률과 추적조사비율이 높을수록 의사최우추정량의 수행도가 좋아짐을 알 수 있었다. 그리고 이 논문에서 아는 값으로 주어진 보고확률을 모르는 경우에 대한 추가적인 연구가 필요할 것이다.

참고문헌

배도선, 최인수, 황용근 (1995), 고장 원인이 여럿인 제품의 사용현장데이터 분석, *응용통계연구*, 8(1), 89-104.

배도선, 윤형제, 최인수 (1995), 수리 가능한 제품의 사용현장데이터 분석, *응용통계연구*, 8(2), 133-145 .

배도선, 최인수, 오영석 (1998), 불완전 보고상황하의 설명변수를 고려한 사용현장데이터 분석, 한국과학기술원 산업공학과 Technical Report #98-20.

Crowder, M. (1986), On consistency and inconsistency of estimating equations, *Econometric*, 2, 305-330.

Hu,X. J. and Lawless,J. F. (1996), Estimation from truncated lifetime data with supplementary information on covariates and censoring times, *Biometrika*, 83, 4, 747-761.

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

Inagaki, N. (1973), Asymptotic relations between the likelihood estimating function and maximum likelihood estimator, *Annals of the Institute of Statistical Mathematics*, 25, 1-26.

Kalbfleisch, J. D. and Lawless,J.F. (1988), Estimation of reliability in field-performance studies, *Technometrics*, 30, 365-388.

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

Lawless, J. F. (1983), Statistical methods in reliability, *Technometrics*, 25, 305-335.

Lawless, J. F. (1994), Adjustments for reporting delays and the prediction of occurred but not reported events, *The Canadian Journal of Statistics*, 22, 15-31.

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

Suzuki, K. (1985a), Estimation of lifetime parameters from incomplete field data, *Technometrics*, 27, 263-272.

Suzuki,K. (1985b), Nonparametric estimation of lifetime distributions from a record of failures and follow - ups, *Journal of the American Statistical Association*, 80, 68-72.

부록: 정리 1의 증명

이 논문에는 정리 1의 간략한 증명 방법만을 언급한다.

(i) $L_p(\theta)$에 포함된 확률밀도함수 $f(t)$가 정칙조건을 따를 때 식 (4)의 일차미분을 $s_p(\theta)=\partial \log L_p(\theta)/\partial\theta$라 하고, $s_p(\theta)$의 기대값을 구하면 0이 되고, 이때 $s_{p(\theta)}=0$를 만족하는 θ^*는 모수 벡터 θ의 일치 추정량이 된다(1986, 1973).

(ii) Inagaki(1986)와 Crowder(1973)는 일치추정량인 경우 $\sqrt{N}(\theta^*-\theta)$는 점근적으로 평균이 0이고 분산-공분산행렬이 $A(\theta)^{-1}B(\theta)A(\theta)^{-1}$인 다변량 정규분포를 따름을 보였다. 이 때의 $A(\theta)$와 $B(\theta)$는 각각 다음과 같다.

$$A(\theta)=\lim_{N\to\infty}\frac{1}{N}E\left(-\frac{\partial s_p}{\partial\theta}\right),$$

$$B(\theta)=\lim_{N\to\infty}\frac{1}{N}E\left(s_p(\theta)s_p(\theta)'\right).$$

식(4)로부터 $s_p(\theta)$를 구한 뒤 정리하면 $A(\theta)^{-1}B(\theta)A(\theta)^{-1}=A(\theta)^{-1}+A(\theta)^{-1}C(\theta)A(\theta)^{-1}$이 됨을 보일 수 있다. 보다 자세한 증명은 배도선 등(1998)에 정리되어 있다.