

오염된 임의절단 자료의 선형회귀분석시 회귀계수 추정에 관한 연구

박춘일* 오대호** 송민구**

Regression Coefficient Estimate for Linear Regression
Analysis case of a Polluted by Arbitrary Truncated Data

Park Choon-Il · Oh Dae-Ho · Song Min-Gu

Abstract

Several methods for linear regression with censored data are considered, that is, iterated least square estimation, maximum likelihood estimation by EM algorithm, Kaplan-Meier estimator, Kour, Susarla & Van Ryzin estimator. Maximum likelihood estimates of parameters in linear models with censored normal responses may be simply obtained using the EM algorithm. The iterative computations required for the regression coefficients are identical to those described by Schmee and Hahn for least squares estimates, but those for the variance estimates are different. KSVR estimator is an estimator of β that does not require iteration. When we estimate regression coefficient estimation of contaminated and censored data, KSVR estimator make an error of regression coefficient estimation because of sensitive to contaminated data such as, least square estimator. In this study, modified estimator based on the least trimmed mean square in regression would be suggested in order to overcome these phenomena.

Key words : Contaminated and Censored data, Trimmed mean, KSVR estimator, EM algorithm, Kaplan-Meier estimator

* 해양대학교 응용수학과 부교수

** 동국대학교 통계학과

I. 서론

변수들 사이의 통계적 관계는 임의 절단된 관찰치를 포함하는 자료를 접합시키지 않으면 안되는 경우가 흔히 발생한다. 예를 들면 수명 검정에서 가속되는 스트레스를 처리하는 경우와 그리고 산출된 자료가 어떤 모형에서 잘못 가정된 평균 시간과 스트레스와의 관련된 경우를 예를 들 수 있다. 하지만, 실제적인 면에서 임의 절단을 포함하는 검정 프로그램에서의 자료가 종종 분석되어진다. 우리는 임의 절단된 자료에서 임의 절단을 제거하기 위해 두 가지 편법들 중에서 하나를 사용하여 완전한 자료를 위한 표준적인 최소자승 기법에 의해 틀리게 분석하는 경우가 종종 발생한다는 것을 알고 있다. 첫째로는 측정하지 못하는 시간을 임의 절단된 시간으로 가정하는 수명 시간을 분석하는 경우에, 마치 그것을 임의 절단되지 않은 관찰치로 취급한다. 둘째는 임의 절단된 관찰치를 무시하는 경향이 있다. 이러한 절차의 모두는 회귀선의 편의 추정을 야기한다. 임의 절단이 존재하는 자료의 선형회귀분석시 회귀 계수를 추정하기 위한 여러 가지 방법들에 관한 활발한 연구가 계속 되고 있는데, β 를 추정하기 위해서 부분적 우도 접근법인 COX(1972) 추정량, 가중합(Weighted sum of square)을 최소화시키는 Miller(1976)의 추정량, Buckley와 James(1979)는 절단된 자료가 포함된 경우의 의사 확률 변수(pseudo random variable)를 정의하고, 이에 대한 잔차자승합을 최소화시키는 추정량을 제안하였으며, Schmeek과 Hahn(1972)는 Gaussian 추정의 임의 절단된 값을 대체하는 반복적인 최소자승 추정을 제안하였고, Aitichim(1981)은 EM 알고리즘(Dempster, Laird & Rubin)을 사용하여 임의 절단이 존재하는 경우 최우추정을 하였다. 또한, Rubin과 Merg(1990)은 EM 알고리즘에서 E-단계의 몬테카를로 기법을 적용시킨 Monte Carlo EM을 사용하였고, Louis(1982)는 EM 알고리즘의 수렴 속도를 향상시키는 Louis Turbo EM을 사용하였다. Tanner와 Wang(1987)의 Data Augmentation 알고리즘 Wei와 Tanner(1990)의 poor Man's Data Augmentation 알고리즘 등이 있다. 그러나 이러한 추정량들은 반복이 필요하다는 단점이 있다. 따라서 이러한 문제점을 극복하기 위해서 Koul, Susarla와 Van Ryzin(1981) 반복이 요구되지 않는 β 의 추정량을 제안하였고, Lemgans(1987)은 비반복 추정량의 대안을 제시하였는데 그것은 Koul 등의 접근 방법보다 향상된 점이 있으며 Fygenon과 Zhou(1992)는 Koul 등이 제안한 추정량을 수정하여 새로운 추정량을 제시하였다. 그러나 Koul 등과 Lemgans, Fygenon과 Zhou가 제안한 추정량은 오염된(contamination) 임의 절단 자료일 때, 회귀선의 기울기를 잘못 추정하는 심각한 문제가 발생한다. 본 연구에서는 임의 절단된 자료의 회귀 계수 추정법인 EM 알고리즘을 이용한 최우추정법, 반복적인 최소자승법, Kaplan-Meier 추정량, KSV 추정량 등에 관하여 고찰하였으며, 앞에서 언급한 문제점의 해결 방안으로서 오염된 자료에 민감하지 않고 주어진 자료의 활용면에서도 의미가 있는

최소절단 평균회귀 (least trimmed mean of square regression) 에 근거하여 오염된 임의 절단 자료의 선형회귀 모형에서 회귀 계수 추정을 위한 추정량 $\hat{\beta}_{MTMKSV}$ 를 제안하고, 이것을 모의실험을 통하여 다른 추정량과 비교하고자 한다 .

II. 본 론

1. EM 알고리즘의 개요

EM 알고리즘은 불완전 자료하에서 MLE를 추정하는 가장 일반적인 반복 알고리즘이다. EM 알고리즘의 과정은 결측자료 Y_{mis} 에 대해서 ① Y_{mis} 를 추정치로 대치하고 ② 모수를 추정하여 ③ 다시 모수 추정치하에서 Y_{mis} 를 보정하고 ④ 다시 모수를 재추정하는 반복 과정을 거치게 된다.

EM 알고리즘은 E-단계 (Expectation Step)와 M-단계 (Maximization Step)로 이루어져 있다.

E-단계에서는 관찰치와 현재 추정된 모수에 대한 결측치의 조건부 기대값을 구한후 이 값을 결측치에 대치한다. 그런데 EM 알고리즘에서는 결측치가 Y_{mis} 자체를 의미하는 것이 아니라, $l(\theta|Y_{obs}, Y_{mis})$ 로 표현되는 Y_{mis} 의 함수로 취급한다는 것이다. 특히 $\theta^{(t)}$ 가 θ 의 현재 반복 추정치라 할 때 E-단계는 $l(\theta|Y)$ 의 기대값

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y)f(Y_{mis}|Y_{obs}, \theta^{(t)})dY_{mis} \quad (1-1)$$

를 찾는다. 이때 M-단계는 모든 θ 에 대해 기대 우도를 최대화하는 즉,

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) \quad \forall \theta \quad (1-2)$$

을 만족하는 $\theta^{(t+1)}$ 를 결정하는 것이다.

2. EM 알고리즘을 이용한 회귀 계수 추정

다음과 같은 선형 회귀 모형을 고려하자.

$$t_i = \beta_0 + \beta_1 X_i + \sigma \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2-1)$$

여기서, $\varepsilon \sim N(0, 1)$, t_i 는 i 번째 측정하지 못한 시간. m 개의 관찰치는 임의 절단되지 않은 자료이고 나머지 $n-m$ 개의 관찰치는 임의 절단된 자료이다. 여기서, 대수 추가 사후 확률은

$$-n \log \sigma - \sum_{i=1}^m (t_i - \beta_0 - \beta_1 X_i)^2 / 2\sigma^2 - \sum_{i=m+1}^n (Z_i - \beta_0 - \beta_1 X_i)^2 / 2\sigma^2 \quad (2-2)$$

여기서, Z_j 는 경우 j 에서 관측하지 못한 시간이다. 조건부 사전 분포 $P(Z_i | \beta_0, \beta_1, \sigma, c_i)$ 는 조건부 정규 분포를 한다. 관측하지 못한 시간 Z_i 는 C_i 보다 크다. E-단계의 계산을 위해 다음 식이 유도된다.

$$-n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^m (t_i - \beta_0 - \beta_1 X_i)^2 - \frac{1}{2\sigma^2} \sum_{i=m+1}^n \{E(Z_i^2 | \beta_0, \beta_1, \sigma, Z_i > C_i) - 2(\beta_0 + \beta_1 X_i)E(Z_i | \beta_0, \beta_1, \sigma, Z_i > C_i) + (\beta_0 + \beta_1 X_i)^2\} \quad (2-3)$$

$$E(Z_i^2 | \beta_0, \beta_1, \sigma, Z_i > C_i) = \mu_i^2 + \sigma^2 + \sigma(C_i + \mu_i)H\left(\frac{C_i - \mu_i}{\sigma}\right) \quad (2-4)$$

$$E(Z_i | \beta_0, \beta_1, \sigma, Z_i > C_i) = \mu_i + \sigma H\left(\frac{C_i - \mu_i}{\sigma}\right) \quad (2-5)$$

여기서, $\mu_i = \beta_0 + \beta_1 X_i$, $H(x) = \phi(x) / \{1 - \Phi(x)\}$, 그리고 $\phi(x)$ 와 $\Phi(x)$ 는 표준 정규 분포의 밀도와 분포 함수이다.

식 (2-5)에서 다음 식을 유도할 수 있다.

$$\begin{aligned}
 E(Z_i | \beta_0, \beta_1, \sigma, Z_i > C_i) &= \beta_0 + \beta_1 X_i + \sigma Z \left(\varepsilon_i | \varepsilon_i > \frac{C_i - \beta_0 - \beta_1 X_i}{\sigma} \right) \\
 &= \beta_0 + \beta_1 X_i + \sigma \left\{ \int_{\frac{C_i - \beta_0 - \beta_1 X_i}{\sigma}}^{\infty} \omega \phi(\omega) d\omega \right\} / \left\{ 1 - \Phi \left(\frac{C_i - \beta_0 - \beta_1 X_i}{\sigma} \right) \right\} \\
 &= \beta_0 + \beta_1 X_i + \sigma \phi \left(\frac{C_i - \beta_0 - \beta_1 X_i}{\sigma} \right) / \left\{ 1 - \Phi \left(\frac{C_i - \beta_0 - \beta_1 X_i}{\sigma} \right) \right\} \\
 &= \beta_0 + \beta_1 X_i + \phi \left(\frac{C_i - \beta_0 - \beta_1 X_i}{\sigma} \right)
 \end{aligned} \tag{2-6}$$

M-단계에서는 다음과 같이 전개된다.

$$\frac{\partial Q}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^m (t_i - \beta_0 - \beta_1 X_i) + \sum_{i=m+1}^n \{ E(Z_i) - \beta_0 - \beta_1 X_i \} = 0 \tag{2-7}$$

$$\frac{\partial Q}{\partial \beta_1} = 0 \rightarrow \sum_{i=1}^m X_i (t_i - \beta_0 - \beta_1 X_i) + \sum_{i=m+1}^n X_i \{ E(Z_i) - \beta_0 - \beta_1 X_i \} = 0 \tag{2-8}$$

$$\frac{\partial Q}{\partial \beta_0} = 0 \rightarrow \frac{\sum_{i=1}^m (t_i - \beta_0 - \beta_1 X_i)^2}{\sigma^4} + \frac{\sum_{i=m+1}^n E(Z_i) - 2(\beta_0 + \beta_1 X_i)E(Z_i) + (\beta_0 - \beta_1 X_i)^2}{\sigma^4} - \frac{n}{\sigma^2} = 0 \tag{2-9}$$

β^{i+1} 을 얻기 위해서 $E(Z_i | \beta_0^i, \beta_1^i, \sigma^i, Z_i > C_i)$ 에 의해서 C_i 를 대체시킨다. 그리고 최소자승 추정법을 적용한다.

σ_{i+1}^2 을 얻기 위해서는 다음 식이 유도된다.

$$\sigma_{i+1}^2 = \frac{\sum_{j=1}^m (t_j - \mu_j^i)^2}{n} + \frac{\sigma_i^2 \sum_{j=m+1}^n \left[1 + \left(\frac{C_j - \mu_j^i}{\sigma_i} \right) H \left(\frac{C_j - \mu_j^i}{\sigma_i} \right) \right]}{n} \tag{2-10}$$

여기서, $\mu_j^i = \beta_0^i + \beta_1^i X_j$ 이다.

3. 반복적인 최소자승추정법

스트레스 x 와 측정하지 못한 평균 시간 μ_x 사이의 관심있는 부분의 범위에서 표준적인 단순 회귀 모형을 가정한다.

$$\mu_x = \beta_0 + \beta_1 x \quad (3-1)$$

여기서, 어떤 스트레스에서 측정하지 못한 시간은 표준편차 σ 를 가지는 정규 분포를 하고, β_0, β_1 과 σ 는 미지의 모수이다. 하나 또는 그 이상의 단위는 몇몇의 스트레스의 각각으로 검정되어진다. 절단된 정규 분포의 잘 알려진 성질을 이용하면, 이러한 단위를 위한 측정하지 못한 시간의 기대값 μ_x^* 은 다음과 같다.

$$\mu_x^* = \mu_x + \sigma f(z) / [1 - F(z)] \quad (3-2)$$

여기서,

$$z = (c_x - \mu_x) / \sigma$$

단 c_x 는 스트레스 x 에서 특별히 소비하는 임의 절단 시간을 나타내고, $f(\cdot)$ 과 $F(\cdot)$ 는 정규 분포 밀도 함수 및 분포 함수를 나타낸다. 이러한 상황에서 반복적인 최소자승 절차는 다음과 같다.

【반복 0】

단계1. 마치 중도 절단된 시간을 측정하지 못한 것처럼 바깥쪽으로 벗어난 것으로 취급하는 표준적인 최소 자승 회귀 분석을 사용하여 선형 관계를 접합시킨다.

$$\hat{\beta}_0^{(0)}, \hat{\beta}_1^{(0)} \text{ 그리고 } \hat{\sigma}^{(0)} \text{ 는 } \beta_0, \beta_1 \text{ 과 } \sigma \text{의 초기 추정치를 나타낸다.}$$

단계2. 각각의 바깥쪽으로 벗어난 즉, 측정하지 못한 절대 평균 시간을 추정하기 위해서 단계1로부터 접합된 초기 회귀선을 사용한다.

$$\hat{\mu}_x^{(0)} \text{ 는 스트레스 } x \text{에서 추정치를 나타낸다.}$$

방정식 (3-1)와 (3-2)를 사용하여 각각의 바깥쪽으로 벗어난 x 를 위한 측정하지 못한 시간 평균 $\hat{\mu}_x^{*(0)}$ 을 추정하기 위해서 σ 와 μ_x 의 미지값을 대신해서 $\hat{\sigma}^{(0)}$ 와 $\hat{\mu}_x^{(0)}$ 의 초기 추정치를 사용한다.

【반복 1】

단계1. 바깥쪽으로 벗어난 것을 위해서 이전 반복 단계2에서 얻어진 측정하지 못한 시간의 평균 $\hat{\mu}_x^{*(0)}$ 를 추정하는데 사용하여, 수정된 최소 자승 회귀를 얻는다.

$\hat{\beta}_0^{(0)}$, $\hat{\beta}_1^{(0)}$ 그리고 $\hat{\sigma}^{(0)}$ 는 $\hat{\beta}_0$, $\hat{\beta}_1$ 과 σ 의 새로운 추정치를 나타낸다.

단계2. 측정하지 못한 평균 시간을 재추정하기 위해서 현 재반복의 단계1에서 얻어진 최소 자승 추정치를 사용하여 이전의 반복 단계 x 를 반복한다.

【반복 2】

이상의 절차를 수렴할 때까지 계속한다.

요약하면, Schmee & Hahn (1979)가 제시한 반복적인 최소자승 방법은 Gaussian 추정의 임의절단 값을 대체하는 것이다. 또한, 앞에서 제시한 EM 알고리즘을 이용한 최우 추정법과 비교할때 계산된 σ 의 값에서 약간의 차이가 나타났다.

4. Kaplan-Meier 추정량

다음과 같은 선형 회귀 모형을 가정한다.

$$t_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (4-1)$$

여기서, t_i : 실제수명시간 (life time)을 나타내고, X_i 는 관측치이며, 오차 $\varepsilon_i \sim N(0, \sigma^2)$ 이다.

Kaplan-Meier은 자승합을 최소화시키는 b_0 와 b_1 의 추정치 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 의 최소 자승 추

정치를 다음과 같이 제시하였다.

$$\sum (t_i - \beta_0 - \beta_1 X_i)^2 \quad (4-2)$$

(4-2)식은 다음과 같이 표현할 수 있다.

$$\frac{1}{n} \sum_{i=1}^n (t_i - \beta_0 - \beta_1 X_i)^2 = \int z^2 d\hat{F}_{\beta_0, \beta_1}(z) \quad (4-3)$$

여기서, $\hat{F}_{\beta_0, \beta_1}$ 는 $z = t_i - \beta_0 - \beta_1 X_i$ ($i = 1, 2, \dots, n$)을 바탕으로 하는 표본분포의 추정치이다. 임의 절단이 있는 경우, β_0 와 β_1 을 최소화하기 위해서 $\int z^2 d\hat{F}_{\beta_0, \beta_1}(z)$, 여기서, $\hat{F}_{\beta_0, \beta_1}$ 을 임의 절단 표본에 바탕을 둔 F 의 Kaplan-Meier 추정이라 한다.

$$z = t_i - \beta_0 - \beta_1 X_i \quad (i = 1, 2, \dots, n) \quad (4-4)$$

적분에서 고정된 β_0 와 β_1 을 위해서 가중 자승함은 다음과 같이 표현된다.

$$\sum_{uc} \omega_i(0, B_i)(t_i - \beta_0 - \beta_1 X_i)^2 \quad (4-5)$$

또한, 최소값을 위한 수치적 탐색은 탐색이 1차원일 때 의미가 있다. 최소화시키는 β_0 의 값은

$$\hat{\beta}_0^{KM} = \sum_{uc} \omega_i(0, B_i)(t_i - \beta_1 X_i) \quad (4-6)$$

β_0 에 관해서 (4-6)식을 최소화시키는 β_1 은 미분에 의해서 얻어질 수 있다. 결국 최소치를 얻기 위한 탐색은 다음과 같은 식으로 유추되어진다.

$$f(\beta_1) = \sum_{uc} \omega_i(B_i)(t_i - \hat{\beta}_0^{KM} - \beta_1 X_i)^2 \quad (4-7)$$

여기서, $\omega_i(\beta_1) = \omega_i(0, B_i)$, 결국 $\hat{\beta}_0^{KM}$ 과 $\hat{\beta}_1^{KM}$ 을 최소화하기 위한 가중 자승함의 수치적 값은 다음과 같다.

$$\sum_{uc} \omega_i (\hat{\beta}_1^{KM})(t_i - \hat{\beta}_0^{KM} - \hat{\beta}_1^{KM} X_i)^2 \quad (4-8)$$

식 (4-7)을 최소화시키는 $\hat{\beta}$ 를 추정하면 된다. 그러나 식 (4-7)에서 $\hat{\beta}$ 을 구하는 것이 어렵기 때문에 Kaplan-Meier은 다음과 같은 반복 추정량을 제안하였다.

단계1) 초기 추정치를 다음과 같이 구한다.

$$\hat{\beta}_0 = \frac{\sum_{uc} t_i (X_i - \bar{X}^{uc})}{\sum_{uc} (X_i - \bar{X}^{uc})^2} \quad (4-9)$$

여기서, \bar{X}^{uc} 는 임의 절단되지 않는 t_i 와 연관된 X_i 의 평균이다.

$$\text{단계2) } \sum_{uc} \omega_i (\hat{\beta}_i)(t_i - \beta_0 - \beta_1 X_i)^2 \text{ 계산한다.} \quad (4-10)$$

$$\text{단계3) } \omega_i^*(\hat{\beta}_0) = \omega_i(\hat{\beta}_0) / \sum_{uc} \omega_j(\hat{\beta}_0) \quad (4-11)$$

로 정의 한다. 그러면, 다음과 같은 새로운 추정량이 구해진다.

$$\hat{\beta}_1^{KM} = \frac{\sum_{uc} \omega_i^*(\hat{\beta}_0) t_i (X_i - \bar{X}_i^*)}{\sum_{uc} \omega_i^*(\hat{\beta}_0) (X_i - \bar{X}_i^*)^2} \quad (4-12)$$

여기서, $\bar{X}^* = \sum \omega_i^*(\hat{\beta}_0) X_i$ 이다.

단계4) 위의 과정을 추정치가 수렴할 때까지 반복한다.

5. KSV(Koul, Susarla와 Van Ryzin) 추정량

Koul, Susarla와 Van Ryzin은 반복이 필요없는 추정량을 제안하였다. 다음과 같은 다중 회귀 모형을 가정한다.

$$T_i = C_i \beta + \varepsilon_i, \quad 1 \leq i \leq n$$

여기서, $C_i = (1, x_{i1}, \dots, x_{ik})$ 는 계획 행렬 C_n 의 i 번째 행이고, $\beta = (\beta_0, \beta_1, \dots)$ 는 모수벡터이며 $\{\varepsilon_i\}$ 는 평균이 0이고, 분산이 σ^2 인 독립인 확률 변수이고, $\{Y_i\}$ 는 $\{\varepsilon_i\}$ 와 독립인 임의의 절단된 확률 변수이다. 우리는 n 이 충분히 크면 $(C_n^T C_n)^{-1}$ 가 존재한다고 가정하자. 따라서 $\hat{\beta}$ 는 다음과 같다.

$$\hat{\beta} = (C_n^T C_n)^{-1} C_n \hat{W} \tag{5-2}$$

$$\hat{W} = (\hat{W}_1, \dots, \hat{W}_n), \quad \text{여기서 } \{\hat{W}_i, 1 \leq i \leq n\}$$

$$\hat{W}_i = \delta_i Z_i [Z_i \leq M_n], \quad d_i = x_i - \bar{x} \quad \text{그리고} \quad \hat{W}_i = W_i [\hat{G}(Z_i)]^{-1} \quad 1 \leq i \leq n$$

$$\delta_i = [T_i < Y_i], \quad Z_i = \min(T_i, Y_i) \quad \text{이다.}$$

그리고 j 번째의 층으로부터 Kaplan-Meier 추정량을 추출한다.

(즉, $\hat{G}_j(z)$ 는 $\{Y_{ij}, 1 - \delta_{ij} : i = 1, \dots, n_j : j = 1, \dots, k\}$ 로부터 추정된다. n_j 는 j 번째 층에서의 관찰치의 수이다.)

$$T_{ij}^{(a)} = \delta_{ij} T_{ij} / (1 - \hat{G}_j(Y_{ij})) \tag{5-3}$$

여기서,

$$\delta_{ij} = \begin{cases} 1, & T_{ij} \leq C_{ij} \\ 0, & \text{o.w} \end{cases}$$

$(\hat{T}_{ij}^{(a)}, C_{ij}^T)$ 를 위한 최소 자승 절차를 적용하면 수정된 추정량은 다음과 같다.

$$\hat{\beta}_{MKS\hat{V}} = (C^T C)^{-1} C^T \hat{T}^{(a)} \tag{5-4}$$

6. 제안된 $\hat{\beta}_{TMKSV}$ 추정량

(5-1)의 회귀 모형을 고려한다. 이때 실제로 관찰한 자료는 T_i 가 아니고 우절단된 Y_i 이므로

$$E(Y_i) \neq c_i\beta, \quad 1 \leq i \leq n \quad (6-1)$$

따라서, 우절단된 자료를 고려하여,

$$Y_{ij}^{(a)} = \frac{\delta_{ij} Y_{ij}}{1 - G(Y_{ij})}, \quad 1 \leq i \leq n_j, \quad j = 1, \dots, k, \quad (6-2)$$

여기서,,

$$\delta_{ij} = \begin{cases} 1, & T_{ij} \leq C_{ij} \\ 0, & o.w \end{cases}$$

그리고 $E(Y_{ij}^{(a)})$ 는 다음과 같이 구할 수 있다.

$$\begin{aligned} E(Y_{ij}^{(a)}) &= \int_0^{\infty} \frac{\nu}{1 - G(\nu)} (1 - G(\nu)) dF(\nu) \\ &= \int_0^{\infty} \nu dF(\nu) \\ &= c_i\beta \end{aligned} \quad (6-3)$$

사실 $Y_1^{(a)}, \dots, Y_n^{(a)}$ 를 모두 관찰할 수 없으므로, 우리는 아래의 추정량을 사용하기로 한다.

$$\widehat{Y}_{ij}^{(a)} = \frac{\delta_{ij} Y_{ij}}{1 - \widehat{G}(Y_{ij})}, \quad i = 1, \dots, n, \quad j = 1, \dots, k \quad (6-4)$$

여기서,

$$\delta_{ij} = \begin{cases} 1, & T_{ij} \leq C_{ij} \\ 0, & o.w \end{cases} \quad \text{이고 } \widehat{G} \text{는 } G \text{에 대한 추정량이다.}$$

따라서, 최소제곱 절단평균 회귀 (least trimmed mean of square regression) 추정법을 이용한 추정량 $\hat{\beta}_{TMKSV}$ 을 다음과 같이 제안한다.

$$\hat{\beta}_{TMKSV} = \underset{\hat{\beta}}{\text{minimize}} \text{ trimed mean } \gamma_i^{(a)} \quad (6-5)$$

여기서, $\gamma_{ij} = \hat{y}_{ij}^{(a)} - c_i \hat{\beta}_{TMKSV}$ 이고, $100\alpha\%$ 절단 평균은 다음과 같이 정의한다.

$$T(\alpha) = \frac{1}{n(1-2\alpha)} \left[k(X_{(r+1)} + X_{(n-r)} + \sum_{i=r+2}^{n-r-1} X_{(i)}) \right], \quad (6-6)$$

여기서, $r = [an]$, $k = 1 - (an - r)$, 그리고 α 는 절단을 (trimed rate) 이다. 그런데 일반적으로 (6-4) 식으로 추정된 $\hat{y}_{ij}^{(a)}$ 는 실제의 자료보다 큰 값을 가지는 경향이 있기 때문에 제안된 추정량 $\hat{\beta}_{TMKSV}$ 은 과대 추정의 문제점이 발생할 수 있다. 이러한 문제점을 극복하기 위해서 $\hat{y}_{ij}^{(a)}$ 의 변환 (transformation)을 취하는 방법을 생각한다. 즉,

$$\hat{y}_{ij}^{(a)} = \ln \hat{y}_{ij}^{(a)}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, k \quad (6-7)$$

따라서 (5-5) 식을 이용하여 수정 제안된 추정량을 $\hat{\beta}_{MTMKSV}$ 은 다음과 같다.

$$\hat{\beta}_{MTMKSV} = \underset{\hat{\beta}}{\text{minimize}} \text{ trimed mean } \gamma_i^{*2} \quad (6-8)$$

여기서, $\gamma_{ij}^* = \hat{y}_{ij}^{(a)} - c_i \hat{\beta}_{MTMKSV}$ 이다.

예제 1.

<표1>은 Crawford에 의해서 처음으로 보고된바 있는 40개의 기관에서 전기의 절연의 수명 검정에서 온도에 의해서 가속화되는 결과를 나타낸다. 10개의 기관은 각각 4종류의 온도에서 실험되었다. 실험은 각각의 온도에서 다른 시간에 종료하였다. 모형은 아래의 조건을 가정한 자료에 의해 분석한다.

- (1) 임의의 온도에서 측정하지 못한 시간의 분포는 대수 정규 분포이다.
- (2) 측정하지 못한 분포에서 대수 정규 시간의 표준편차 σ 는 상수이다.
- (3) 측정하지 못한 시간의 대수 평균 μ_x 는 절대온도 T 를 $x = 1000/(T+273.2)$ 로 변환

하는 상호관계를 나타내는 선형 함수이다.

〈표1〉 다양한 검정 온도에서 절연 수명 시간

실 험 온 도			
150℃	170℃	190℃	220℃
8064*	1764	408	408
8064*	2772	408	408
8064*	3444	1344	504
8064*	3542	1344	504
8064*	3780	1440	504
8064*	4860	1680*	528*
8064*	5196	1680*	528*
8064*	5488*	1680*	528*
8064*	5488*	1680*	528*
8064*	5488*	1680*	528*

〈표2〉 절연 수명 자료의 EM 알고리즘 및 반복적인 최소제곱을 이용한 회귀계수 추정치의 비교

	추 정 형 태	
	반복적 최소제곱법 (17번 반복)	EM 알고리즘 이용한 최우추정법 (16번 반복)
β_0	-5.818	-6.019
β_1	4.204	4.311
대수표준편차(σ)	0.2041	0.2592
각각의 온도에서 수명 시간의 중위수		
220℃	208	530
190℃	1812	1940
170℃	4654	5080
150℃	13060	14680
130℃	40638	47000

예제 2.

〈표3〉에서처럼 우절단 모의자료가 오염된 자료(contaminated data)일 때 $\hat{\beta}_{KSV}$ 추정량과 본 연구에서 제안한 $\hat{\beta}_{MTMKSV}$ 추정량을 비교 하는데, $\hat{\beta}_{MTMKSV}$ 추정법에서 절단률(trimd rate)는 10%, 30%, 50%일 경우로 하겠다.

〈표3〉 모의 자료 발생 방법

$T_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 단 $\beta_0 = 1$ $\beta_1 = 1.5$	
$\varepsilon_i \sim N(0, 0.2^2)$	$n = 50$
$X_i \sim U(3, 6)$	$n = 50$
a) $C_i \sim U(8, 8.5)$	$n = 50$
b) $C_i \sim U(7.8, 8.3)$	
c) $C_i \sim U(7.4, 7.9)$	
$CY_i \sim N(3, 0.2^2)$	$n = 7$
$CY_i \sim N(7.0, 0.2^2)$	$n = 7$

〈표4〉 모의 자료에서 $\hat{\beta}_{KSV}$ 와 $\hat{\beta}_{MTMKSV}$ 의 비교

절단백분율	추정방법	잔차자승의 절단율	추 정 량		모 수	
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
16.7%	$\hat{\beta}_{KSV}$		14.8743	-0.9243	1	1.5
	$\hat{\beta}_{MTMKSV}$	10%	-0.7328	0.7456		
		30%	-1.2457	0.9324		
26.7%	$\hat{\beta}_{KSV}$		17.2675	-2.2574	1	1.5
	$\hat{\beta}_{MTMKSV}$	10%	-1.5428	0.7248		
		30%	-1.4836	0.8324		
36.7%	$\hat{\beta}_{KSV}$		16.2481	-2.8043	1	1.5
	$\hat{\beta}_{MTMKSV}$	10%	-1.3125	0.8742		
		30%	-1.2944	1.1394		
		50%	-1.2539	1.2411		

III. 결 론

본 연구에서는 임의 절단 자료의 선형회귀분석시 회귀계수 추정에 관하여 살펴보았는데 그것은, EM 알고리즘을 이용한 최우추정법, 반복적인 최소자승법, Kaplan-Meier 추정량, Koul, Susarla와 Van Ryzin이 제안한 추정량, 본 연구에서 제안한 $\hat{\beta}_{MTMKSV}$ 추정량,

$$\hat{\beta}_{MTMKSV} = \underset{\hat{\beta}}{\text{minimize}} \text{ trimed mean } \gamma_i^{*2}$$

이다. 그런데 EM 알고리즘을 이용한 최우추정법, 반복적인 최소자승법, Kaplan-Meier 추정량 등은 반복(iteration)이 필요하다는 결점이 있다. 따라서 Koul, Leurgans, Fygenon와 Zhou등은 이러한 단점을 보완하는 비반복 추정량을 제안하였다. 그러나 임의 절단된 자료에 오염된 자료가 존재할 경우에는 최소자승추정법에서 나타나는 것과 같이 회귀계수의 추정치가 의미가 없게 된다. 따라서 오염된 자료가 포함된 임의절단 모의실험 자료일때, $\hat{\beta}_{KSV}$ 추정량과 본 논문에서 제안된 $\hat{\beta}_{MTMKSV}$ 추정량을 비교해 본 결과, $\hat{\beta}_{KSV}$ 추정량은 양의 기울기를 음의 기울기로 잘못 판정하는 경우가 발생 하였으나, 본 논문에서 제안한 $\hat{\beta}_{MTMKSV}$ 는 이러한 경우에는 비교적 정확한 추정량을 제공하였다. 또한 $\hat{\beta}_{MTMKSV}$ 에서 절단율(trimed rate)이 높을수록 추정효율이 향상됨을 볼 수 있었다.

참 고 문 헌

1. Buckley. J. and I. James (1979). Linear regression with censored data, *Biometrika* 66, 429-436.
2. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood From Incomplete Data via the EM Algorithm"(with discussion), *J. Roy. Statist. Soc., Ser. B*, 39, 1-38.
3. Heller. G. and J. Simonoff (1990). A comparison of estimators for regression with a censored response variable, *Biometrika* 77, 515-520.
4. James, I. R. & Smith, P. J. (1984). Consistency results for linear regression with censored data. *Ann. Statist.* 12, 590-600.
5. Koul. H., V. Susarla and J. Van Ryzn (1981). Regression analysis with randomly right-censored data, *Ann. Statist.* 9, 1276-1288.

6. Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* 74, 301-9.
7. Mendel Fyngenson & Mai Zhou. (1992). Modifying the Koul, Susarla and Van Ryzin estimator for linear regression models with right censoring, 295-299.
8. Miller, R. G. and J. Halpern (1982). Regression with censored data, *Biometrika* 69, 521-531.
9. Schmee, J., and Hahn, G. J. (1979). "A Simple Method for Regression Analysis With Censored Data," *Technometrics*, 21, 471-432.
10. Schneider, H. & Weissfeld, L. (1986). Estimation in linear models with censored data. *Biometrika* 73, 741-5.
11. Simon, G. A. & Simonoff, J. S. (1986). Diagnostic plots for missing data in least square regression. *J. Am. Statist. Assoc.* 81, 501-9.
12. Smith, P. J. (1988). Asymptotic properties of linear regression estimations under a fixed censorship model. *Aust. J. Statist.* 30, 52-66.
13. Stephen M. Stigler (1973). "The Asymptondistribution of the trimmed mean", *The Annals of Statistics*, Vol. 1, 472-477.
14. Wei, G.C.G. and Tanner, M.A(1990). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm", *Journal of the American Statistical Association*, 85, 699-704.
15. Wei, G.C.G. and Tanner, M.A(1990). "Poster Computation for Censored Regression Data", *Journal of the American Statistical Association*, 85, 829-839.
16. Weissfeld, L. A. & Schneider, H. (1987). Inferences based on the Buckley-James procedure. *Comm. Statist. A* 16, 177-87.