김일성종합대학학보 주체109(2020)년
수 학 제66권 제2호

# 구간형불완전자료인 경우 한가지 선형회귀모형의 파라메터추정문제

리 광 선

지난 시기에는 관측자료가 구간형불완전자료로 주어지는 경우 그에 대한 통계적추론 문제가 많이 연구되지 못하였다.[1, 3－5]

선행연구[2]에서는 간단한 선형회귀모형인 경우 구간형불완전자료에 기초한 반응변량의 추정값을 그 구간의 중간점으로 취급한 단순한 방법이 연구되였으나 이 방법은 구간의 길이에 따라 오차가 상대적으로 크게 변하는 부족점을 가지고있다.

또한 조건부수학적기대값을 리용하여 반응변량의 추정량을 구한 선행연구[1]의 방법 역시 $Y$의 분포함수 $F(x)$가 미지인 경우 추정량을 구할수 없는 부족점을 가지고있다.

론문에서는 구간형불완전자료인 경우 선형회귀모형의 반응변량의 한가지 추정량을 구성하는 방법과 그 특성을 연구하고 그에 기초하여 회귀모형의 파라메터추정문제에 대한 연구를 진행하였다.

## 1. 문 제 설 정

$Y$는 어떤 수명시간을 표시하는 우연량이라고 하자.

이때 $Y$의 정확한 수명이 관측되지 못하고 어떤 구간 $(u,\ v)$ $(u<v)$ 사이에 있다는것만을 안다고 하자. 즉 관측결과는 $(\delta_1,\ \delta_2,\ u,\ v)$로 표시할수 있다.

여기서 $\delta_1=I(Y\le u),\ \ \delta_2=I(u<Y\le v)$이다.

그리고 $I(x)$는 정의함수를 의미한다.

이제 다음과 같은 선형회귀모형을 고찰하기로 한다.

$$y_i=\alpha+\beta x_i+\varepsilon_i\ \ (i=\overline{1,\ n}) \tag{1}$$

여기서 $\alpha,\ \beta$는 미지인 회귀결수들, $\varepsilon_i(i=\overline{1,\ n})$는 독립이고 동일분포인 우연량들, 그 분포함수는 $F_\varepsilon(\cdot),\ E\varepsilon_i=0,\ E\varepsilon_i^2=\sigma^2$이라고 하자.

이때 $y_i\ (i=\overline{1,\ n})$가 관측되지 못하고 다만 $(u_i,\ v_i)\,(i=\overline{1,\ n})$가 관측되는 경우에 $y_i$를 추정하는 문제와 회귀결수 $\alpha,\ \beta$의 추정량을 구하는 문제를 연구하려고 한다.

## 2. $Y$의 추정량과 그 성질

이제 모형 (1)의 반응변량 $Y$의 추정량을 구성하는 문제를 고찰하자.

만일 $y_i\le u_i$인 경우에는 $y_i$가 정확히 관측되고 $u_i<y_i\le v_i$인 경우에는 $y_i$가 관측되지 않고 다만 $(u_i,\ v_i)$가 관측되게 된다.

선행연구[1]에서는 조건부수학적기대값을 리용하여 $y_i$를 다음과 같이 추정하였다.

$$y_i^* = y_i\delta_{1i} + E(y \mid u < y \le v)\delta_{2i} + E(y \mid y > v)(1-\delta_{1i}-\delta_{2i}) \tag{2}$$

여기서 $\delta_{1i} = I(y_i \le u_i),\ \delta_{2i} = I(u_i < y_i \le v_i)$ 이다.

이제 $y_i$의 추정량 $y_i^*$의 성질을 보면 다음과 같다.

$(u_i,\ v_i)\ (i=\overline{1,\ n})$는 서로 독립이고 동일분포 $H(x,\ y)$에 따른다고 하자. 그러면

$$Ey_i^* = Ey_i\ (i=\overline{1,\ n}),\ Var(y^*) \le Var(y)$$

가 성립한다.

따라서 $y_i$의 추정량 $y_i^*$이 비교적 좋은 추정량이라는것을 알수 있다.

그러나 $Y$의 분포함수 $F(x)$가 미지인 경우에는 추정량 $y_i^*$을 구할수 없게 된다.

이러한 부족점을 극복하기 위하여 다음과 같이 추정량을 구성하기로 하자.

$$y_i^* = f_1(y_i)\delta_{1i} + f_2(u_i,\ v_i)\delta_{2i} + f_3(v_i)(1-\delta_{1i}-\delta_{2i}) \tag{3}$$

여기서 $f_1$, $f_2$, $f_3$은 다음과 같은 조건을 만족시키는 함수들이라고 하자.

① $f_1(y_i)(1-H_u(y_i)) + \iint\limits_{u_i<y_i\le v_i} f_2(u,\ v)h(u,\ v)dudv + \int\limits_0^{y_i} h_v(v)f_3(v)dv = y_i$

② $f_1$, $f_2$, $f_3$은 $H(u,\ v)$, $h_u$, $h_v$와는 관계가 있고 $F(x)$와는 무관계한 함수들이다.

그러면 다음과 같은 사실이 성립한다.

정리 1 $y_i$의 추정량

$$y_i^* = f_1(y_i)\delta_{1i} + f_2(u_i,\ v_i)\delta_{2i} + f_3(v_i)(1-\delta_{1i}-\delta_{2i})$$

에 대하여

$$Ey_i^* = Ey_i\ (i=\overline{1,\ n})$$

이다. 여기서 $f_1$, $f_2$, $f_3$은 우의 조건 ①, ②를 만족시키는 함수들이다.

증명 $Ey_i^* = E[f_1(y_i)\delta_{1i} + f_2(u_i,\ v_i)\delta_{2i} + f_3(v_i)(1-\delta_{1i}-\delta_{2i})] =$

$$= \iint\limits_{y_i\le u_i} f_1(y_i)dF_i(y_i)h_u(u)du + \iiint\limits_{u_i<y_i\le v_i} f_2(u,\ v)h(u,\ v)dudvdF_i(y_i) +$$

$$+ \iint\limits_{y_i>v_i} f_3(v)dF_i(y_i)h_v(v)du = \int\limits_{-\infty}^{+\infty} f_1(y_i)(1-H_u(y_i))dF_i(y_i) +$$

$$+ \int\limits_{-\infty}^{+\infty} \iint\limits_{u_i<y_i\le v_i} f_2(u,\ v)h(u,\ v)dudvdF_i(y_i) + \int\limits_{-\infty}^{+\infty}\int\limits_0^{y_i} f_3(v)h_v(v)dudF_i(y_i) =$$

$$= \int\limits_{-\infty}^{+\infty}\left[ f_1(y_i)(1-H_u(y_i)) + \iint\limits_{u_i<y_i\le v_i} f_2(u,v)h(u,v)dudv + \int\limits_0^{y_i} f_3(v)h_v(v)du \right] dF_i(y_i) =$$

$$= \int\limits_{-\infty}^{+\infty} y_i dF_i(y_i) = Ey_i$$

여기서 마지막식은 조건 ①로부터 나온다. 즉 $y_i$의 추정량 $y_i^*$은 $Ey_i$의 불편추정량이 된다.(증명끝)

정리 2 $f_1$, $f_2$, $f_3$은 조건 ①, ②와 $f_1(z)=f_2(z)=f_3(z)$를 만족시키는 함수들이라고 하자. 여기서

$$z=\begin{cases} y, & y\le u \\ f(u,\ v), & u<y\le v \\ v, & y>v \end{cases}$$

이다. 이때 $y_i$의 추정값으로 $y_i^*$을 취하면 다음과 같은 사실이 성립한다.

$$Var(y_i^*)=\inf_{f_1,\ f_2,\ f_3} Var(y)$$

증명 정리 1로부터 $y_i^*$과 $y_i$의 수학적기대값은 같다.

조건 ①로부터

$$f_1=(1-H_u)^{-1}\left(y-\iint\limits_{u<y\le v} f_2(u,\ v)h(u,\ v)dudv-\int_0^y h_v\, f_3(v)dv\right)$$

이고 $y_i^*$의 2차모멘트를 계산하면

$$a=E(y_i^*)^2=E\left[\delta_1 f_1+\delta_2 f_2+(1-\delta_1-\delta_2)f_3\right]^2=$$

$$=E(f_1^2+f_2^2+f_3^2)=\iint\limits_{y\le u} f_1^2 h_u dudF+\iiint\limits_{u<y\le v} f_2^2 h(u,\ v)dudvdF+\iint\limits_{y>v} f_3^2 h_v dvdF=$$

$$=\int_{-\infty}^{+\infty} f_1^2(1-H_u(y))dF+\int_{-\infty}^{+\infty}\iint\limits_{u\le y\le v} f_2^2 h(u,\ v)dudvdF+\int_{-\infty}^{+\infty}\int_0^y f_3^2 h_v dvdF$$

이므로

$$a=\int_{-\infty}^{+\infty}(1-H_u)^{-1}\left(y-\iint f_2(u,\ v)h(u,\ v)dudv-\int_0^y h_v\, f_3(v)dv\right)^2 dF+$$

$$+\int_{-\infty}^{+\infty}\iint\limits_{u<y\le v} f_2^2 h(u,\ v)dudvdF+\int_{-\infty}^{+\infty}\int_0^y f_3^2 h_v dvdF=\int_{-\infty}^{+\infty}(1-H_u)^{-1}bdF$$

이다. 여기서 $b$는

$$b=b(f_2,\ f_3)=\left[y-\iint\limits_{u<y\le v} f_2 h(u,\ v)dudv-\int_0^y h_v\, f_3(y)dv\right]^2+$$

$$+(1-H_u)\iint\limits_{u<y\le v} f_2{}^2 h(u,\ v)dudv+(1-H_u)\int_0^y f_3{}^2 h_v dv$$

이다. $F$의 임의성으로부터 결국 $a$를 최소화하는 문제는 $b$를 최소화하는것과 동등하다.

이제 $b$를 최소화하는 문제를 고찰하자.

$\Delta=\Delta(u,\ v)$, $\rho_1$, $\rho_2$를 파라메터라고 할 때 $b(f_2+\rho_1\Delta,\ f_3+\rho_2\Delta)$를 구하면

$$\left.\frac{\partial b(f_2+\rho_1\Delta,\ f_3+\rho_2\Delta)}{\partial\rho_1}\right|_{\rho_1=0,\ \rho_2=0}=-y\iint\Delta h(u,\ v)dudv+\iint f_2 h(u,\ v)dudv\iint\Delta h(u,\ v)dudv+$$

$$+\int_0^y h_v f_3 dv \iint \Delta h(u,\ v)dudv + (1-H_u)\iint f_2 \Delta h(u,\ v)dudv = 0$$

이고 마찬가지로

$$\left.\frac{\partial b(f_2+\rho_1\Delta,\ f_3+\rho_2\Delta)}{\partial \rho_2}\right|_{\rho_1=0,\ \rho_2=0} = 0$$

또는

$$\iint \Delta h(u,\ 1-H_u)\left[-y+\iint f_2 h(u,\ 1-H_u)dud(1-H_u)+\int_0^y h_{1-H_u} f_3 d(1-H_u)+(1-H_u)f_2\right]\cdot$$

$$\cdot dud(1-H_u)=0$$

$$\iint \Delta h(u,\ v)\left[-y+\iint f_2 h(u,\ v)dudv+\int_0^y h_v f_3 dv+(1-H_u)f_3\right]dudv=0$$

이 성립한다. $\Delta$의 임의성으로부터 간단히 하면

$$-y+\iint f_2 h(u,\ v)dudv+\int_0^y h_v f_3 dv+(1-H_u)f_2=0 \tag{4}$$

$$-y+\iint f_2 h(u,\ v)dudv+\int_0^y h_v f_3 dv+(1-H_u)f_3=0 \tag{5}$$

이다. 또한 조건 ①로부터

$$-y+\iint f_2 h(u,\ v)dudv+\int_0^y h_v f_3 dv+(1-H_u)f_1=0 \tag{6}$$

이므로 식 (4)－(6)을 종합하면 $f_1(z)=f_2(z)=f_3(z)$를 얻는다.

여기서

$$z=\begin{cases} y, & y\le u \\ f(u,\ v), & u<y\le v \\ v, & y>v \end{cases}$$

이고 $z=f(u,\ v)$는 $z$가 $u,\ v$의 함수로 확정된다는것을 표시한다.(증명끝)

정리 2는 분산이 최소로 되는 추정량이 존재한다는것을 보여준다.

## 3. 선형회귀모형파라메터의 추정

선형회귀모형 (1)의 미지인 파라메터 $\alpha,\ \beta$의 추정량을 구하기로 한다.

$F_i(.)$는 $y_i$의 분포함수라고 하자.

이제 $\{y_i\}_{i=1}^n$이 구간형불완전자료로 주어졌을 때 최소두제곱법을 리용하여 회귀결수 $\alpha,\ \beta$의 추정량을 구하기로 한다.

앞에서 구한 반응변량 $y_i$의 추정량 $y_i^*$을 리용하면 $\alpha,\ \beta$의 최소두제곱추정량은 각각 다음과 같다.

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i^*}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\alpha}_n = \bar{y}^* - \hat{\beta}_n \bar{x} \tag{7}$$

여기서

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^*$$

이다.

이때 다음의 정리가 성립한다.

정리 3 $\alpha$, $\beta$의 최소두제곱추정량 $\hat{\alpha}_n$, $\hat{\beta}_n$은 $\alpha$, $\beta$의 불편추정량으로 된다. 즉

$$E\hat{\beta}_n = \beta, \quad E\hat{\alpha}_n = \alpha$$

다음으로 회귀모형의 오차분산 $\sigma^2$의 추정량을 구하기로 한다.

$\sigma^2$의 추정량으로서 $\hat{\sigma}_n{}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^* - (\hat{\alpha}_n^2 + 2\hat{\alpha}_n\hat{\beta}_n\bar{x} + \hat{\beta}_n^2 x^2)$이라고 하면 다음과 같은 사실이 나온다.

정리 4 $\sup_{1\le i\le n}(Var(y_i^*) - \sigma^2) = O(n)$이라고 하면

$$\lim_{n\to\infty} E\hat{\sigma}_n^2 = \sigma^2$$

이 성립한다.

## 참 고 문 헌

[1] G. Gomez et a1.; Statistics in Medicine, 22, 409, 2003.

[2] T. Rebekka, G. Guadaluoe; Statistics in Medicine, 23, 3377, 2004.

[3] M. Tan, G. L. Tian; Estimating Restricted Normal Means Using the EM Type Algorithms and IBF Sampling, World Scientific Publishing, 101~105, 2003.

[4] 郑祖康; 应用概率统计, 20, 119, 2004.

[5] 王蓉华; 强度与环境, l33, 61, 2006.

주체108(2019)년 12월 15일 원고접수

## Study on the Parameter Estimation for a Linear Regression Model with Interval Censored Data

*Ri Kwang Son*

In this paper, we study a method to estimating parameters for a linear regression model with interval censored data, and then consider properties of the estimated volumes.

Keywords: parameter estimate, regression model, interval censored data