碩士學位論文

승격시간치유모형(Promotion Time Cure Model)에서의 비모수적 추론

韓國外國語大學校 大學院

統 計 學 科

李 鍾 誠

碩士學位論文

승격시간치유모형(Promotion Time Cure Model)에서의 비모수적 추론

Nonparametric Reasoning in the Promotion Time Cure Model

指導 梁 城 準 敎授

이 論文을 碩士學位 請求論文으로 提出합니다.

2018年 12月

韓國外國語大學校 大學院

統 計 學 科

李 鍾 誠

이 論文을 李鍾誠의 碩士學位 論文으로 認定함.

2018年 12月 7日

審査委員 이석호 (인)

審査委員 양성근 (인)

審査委員 이태욱 (인)

韓國外國語大學校 大學院

요 약

기존 생존분석에서는 치유 된 개체의 존재를 고려하지 않는 경우가 많았지만 실제로는 치유된 개체들이 존재하는 경우가 있다. 이러한 치유된 개체들을 고려한다면 더 정확한 생존함수에 대한 추정이 가능하며, 이를 고려한 것이 Cure Model이다. 본 논문에서는 Cure Model의 갈래중 하나인 Promotion Time Cure Model에서 공변량이 measurement error에 영향을 받는 경우 생존함수의 회귀계수 신뢰구간 추정에 대해 고려한다. Ma & Yin(2008)에서는 Promotion Time Cure Model의 공변량이 measurement error에 영향을 받는 경우에 추정방법이 제안되었으나, measurement error를 정규분포로 가정을 하였다. 따라서 본 논문에서는 measurement error가 정규분포일 때와 아닐 때 생존함수의 회귀계수 신뢰구간을 Ma & Yin(2008)의 추정법과 Bootstrap 방법을 이용하여 비교하였다.

주요용어 : Promotion Time Cure Model, measurement error,
Bootstrap 신뢰구간

목 차

[표 목 차]

1. 서론

기존의 생존분석에서는 모든 개체가 관심 있는 이벤트(failure)를 경험한다고 가정한다. 하지만, 생존분석과 같이 Time-to-event data를 분석할 때 특정 개체의 특정 비율이 관심 있는 이벤트를 결코 경험하지 못하는 경우가 종종 있다. 예를 들어, 특정 질병의 재발까지 시간에 관심이 있는 의학연구에서 일부 환자는 결코 재발을 겪지 않을 것이라고 알려져 있다. 실업기간에 대한 계량경제학연구에서 실업자들은 새로운 직장을 찾지 않을 것이며, 결혼생활에 관한 사회학적 연구에서 어떤 사람들은 평생 동안 결혼하지 않은 채있을 것이다. 각 예시에는 사건을 경험하지 못하는 특정 비율의 개체가 있다. 이 특정 비율의 개체는 치유된다고 한다. 이처럼 실제로는 치유된 개체들이 존재하는 경우가 많으며, 치유된 개체들을 고려하여 생존함수를 추정하면 더 정확한 추정이 가능하다. 이를 고려한 모델이 Cure Model이다.

Time-to-event 변수에 대한 공변량의 영향을 모델링하기 위한 두 가지의 주요 Cure Model이 있는데 첫 번째는 Mixture Cure Model이다. Mixture Cure Model의 생존함수는 $S(t \mid x_1, x_2) = p(x_2) + \{1-p(x_2)\}S_u(t \mid x_1)$이다. 여기서 x_1, x_2는 공변량 벡터이고 $S_u(t \mid x_1)$는 치료되지 않은 개체의 생존함수이며 $p(x_2)$는 x_2의 치유될 확률, 즉 cure fraction이다. 이 모델은 Boag(1949), Berkson & Gage(1952), Farewell(1982), Kuk & Chen(1992), Taylor(1995), Peng & Dear(2000), Sy & Taylor(2000), Peng(2003), Lu(2008)의 연구 논문이 있다. 두 번째는 Cox(1972)모델을 기반으로 cure fraction을 고려한 Promotion Time Cure Model이다.

$$S(t \mid x) = \exp\{-\theta(x)F(t)\} \tag{1.1}$$

여기서 $F(\bullet)$는 적절한 베이스라인 누적분포함수(CDF)이고 $\theta(\bullet)$은 link function으로 일반적으로 $\theta(x) = \exp\{x^{\top}\beta\}$를 사용한다. 이 모델에서 cure fraction은 $\exp\{-\theta(x)\}$이다. Promotion Time Cure Model이 Cox(1972)모델을 기반으로 하는 이유는 $F(\bullet)$를 CDF가 아닌 Cumulative Hazard function으로 정리를 하면 Cox(1972)모델이 되기 때문이다. Promotion Time Cure Model에 대한 논문은 Yakovlev & Tsodikov(1996), Tsodikov(1998a,b, 2001), Chen et al.(1999), Ibrahim et al.(2001), Tsodikov et al.(2003), Zeng et al.(2006), Carvalho Lopes & Bolfarine(2012)등이 있다.

본 논문에서는 Promotion Time Cure Model에서 공변량에 measurement error가 있는 경우를 고려한다. 일부 연속 공변량은 measurement error가 발생할 수 있다. measurement error는 거의 고려되지 않았지만 이를 무시하면 추정의 효율성이 떨어지고 잘못된 결론이 도출 될 수 있다.(Carroll et al., 2006) 이 measurement error를 처리하기 위해서는 형식에 대한 일부 가정이 필요한데, $W = X + U$에 대해 W가 관찰된 공변량의 벡터이고 U가 measurement error의 벡터인 모델을 고려한다. U가 정규분포라고 가정하면 $W = X + U$는 Ma & Yin(2008)에 의해 연구된 measurement error 모델이다. 하지만, Ma & Yin(2008)에 의한 연구는 U가 정규분포라는 가정을 했기 때문에 정규분포가 아닌 다른 분포라면 추정의 효율성이 떨어질 것이다. 따라서 본 논문에서는

Promotion Time Cure Model에서 공변량에 measurement error가 있는 경우 Ma & Yin(2008)의 추정법에 의해 구해진 생존함수의 β계수의 신뢰구간과 Bootstrap 방법에 의해 구해진 β계수의 신뢰구간 비교를 통해 U가 정규분포가 아닐 때 Ma & Yin(2008)의 추정법이 효율성이 떨어짐을 보일 것이다.

논문은 다음과 같이 구성하였다. 2절에서는 Ma & Yin(2008)의 추정법에 대해 간략히 소개하고 3절에서는 Bootstrap 신뢰구간의 3가지 방법에 대해 소개한다. measurement error의 분포에 따른 신뢰구간의 비교는 4절에서 이루어졌으며, 5절에서는 결론을 다룬다.

2. Ma & Yin(2008) 추정법

T_i와 C_i를 각각 i번째 개체의 failure time, censoring time이라고 정의하고 첫 번째 성분이 1인 공변량 벡터 X_i를 정의한다. $Y_i = \min(T_i, C_i)$와 $\Delta_i = I(T_i \le C_i)$가 censoring indicator일 때 (Y_i, Δ_i, X_i)는 독립이고 동일한 분포(iid)라고 가정한다. 이때 생존함수는 다음과 같다.

$$S(t|X) = \exp\left\{-F(t)e^{X^\top\beta}\right\}, \quad W = X + U \tag{2.1}$$

여기에서 measurement error U의 분포는 $N(0, V)$이며, 우선 measurement error를 고려하지 않은 생존함수, $S(t|X)$의 log-likelihood는 다음과 같이 쓸 수 있다.

$$\begin{aligned}\log f(Y, \Delta|X) = {} & \Delta I(Y < \infty)\left\{-F(Y)e^{X^\top\beta} + \log F\{Y\} + X^\top\beta + \log S_C(Y|X)\right\} \\ & + (1-\Delta)I(Y < \infty)\left\{\log f_c(Y|X) - F(Y)e^{X^\top\beta}\right\} \\ & + I(Y = \infty)\left\{\log S_C(\infty|X) - e^{X^\top\beta}\right\}\end{aligned}$$

여기에서 $F\{Y\}$는 Y에서 $F(\bullet)$의 jump size를 나타내며 $F(\bullet)$은 이벤트 시간에만 jump가 있는 right-continuous 함수이다. 설명의 용이성을 위해 $p_i \equiv F\{Y_i\}$, m은 failure time의 순서라고 하면 $\sum_{i=1}^{m} p(i) = 1$하에서 Lagrange multiplier λ에 대

해 $\sum_{i=1}^{n}\log f(Y_i,\Delta_i|X_i)-n\lambda(\sum_{i=1}^{m}p(i)-1)$와 같이 쓸 수 있고, maximizing하여 다음의 식 (2.2)와 (2.3)를 얻을 수 있다.

$$\frac{1}{p(i)}=\sum_{j=1}^{n}I(Y_{(i)}\le Y_j<\infty)e^{X_j^{\top}\beta}+n\lambda,\ i=1,...,m \tag{2.2}$$

$$\sum_{i=1}^{m}p(i)=1 \tag{2.3}$$

(2.2)와 (2.3)의 식을 backfitting방법을 반복하여 풀면 (2.4)의 식을 얻을 수 있다.

$$\sum_{i=1}^{n}\left\{\Delta_i I(Y_i<\infty)-F(Y_i)e^{X_i^{\top}\beta}\right\}X_i=0 \tag{2.4}$$

measurement error를 고려하지 않고 풀었을 때 위의 (2.2), (2.3), (2.4), 3개의 식을 얻을 수 있다. 여기에 measurement error U를 고려하여, X 대신에 W를 넣어 (2.2)와 (2.4)의 식을 풀면 각각 $e^{X_i^{\top}\beta}$는 $e^{W_i^{\top}\beta-\beta^{\top}V\beta/2}$로 X_i와 $e^{X_i^{\top}\beta}X_i$는 W_i와 $e^{W_i^{\top}\beta-\beta^{\top}V\beta/2}(W_i-V\beta)$로 풀 수 있으며, 다음의 식 (2.5)와 (2.6)을 얻을 수 있다.

$$\frac{1}{p(i)} = \sum_{j=1}^{n} \frac{1}{r_j} \sum_{k=1}^{r_j} I(Y_{(i)} \le Y_j < \infty) e^{W_{jk}^{\top}\beta - \beta^{\top}V\beta/2} + n\lambda\,,\ i = 1,...,m \qquad (2.5)$$

$$\sum_{i=1}^{n} \frac{1}{r_i} \sum_{k=1}^{r_i} \left\{ \Delta_i I(Y_i < \infty) W_{ik} - F(Y_i) e^{W_{ik}^{\top}\beta - \beta^{\top}V\beta/2} (W_{ik} - V\beta) \right\} = 0 \qquad (2.6)$$

이번에는 (2.3), (2.5), (2.6) 3개의 식을 위에서 사용했던 backfitting방법을 반복하여 풀면 β에 대한 추정을 할 수 있다.

backfitting방법을 반복하여 추정된 estimator $\hat{\beta}_n$은 다음을 만족한다.

$\sqrt{n}(\hat{\beta}_n - \beta_0) \to N(0, A^{-1}B(A^{-1})^{\top})$에서 $n \to \infty$일 때,

$$A = E(e^{W^{\top}\beta - \beta^{\top}V\beta/2}[F_0(Y)\{V - (W - V\beta)(W - V\beta)^{\top}\} - (W - V\beta)\int_0^Y b_4(y)^{\top} dF_0(y)])^{\top}$$

$$B = \left\{ S_{\beta}(\beta_0, F_0) + S_F(\beta_0, F_0)[\int_0^Y b_4(y)\, dF_0(y)] \right\}^{\otimes 2} \text{ 이다.}$$

여기에서 $S_{\beta}(\beta, F) = -F(Y)e^{W^{\top}\beta - \beta^{\top}V\beta/2}(W - V\beta) + \triangle I(Y < \infty)W$ 이고,

$b_4(y) = \{c_1 - c_2(y)\}^{-1}\{b_1(y) - b_2 - b_3\}$인데 이때 b_1, b_2, b_3, c_1, c_2는 다음과 같다.

① $b_1(y) = E\{I(y \le Y)e^{W^{\top}\beta - \beta^{\top}V\beta/2}(W - V\beta)\}$

② $b_2 = \int_0^{\infty} b_1(y) dF_0(y)$

③ $b_3 = [\int_0^{\infty} \{c_1 - c_2(y)\}^{-1} dF_0(y)]^{-1} \times [\int_0^{\infty} \{b_1(y) - b_2\}/\{c_1 - c_2(y)\} dF_0(y)]$

④ $c_1 = E\left\{e^{W^\top\beta-\beta^\top V\beta/2}F_0(Y)-I(Y<\infty)\triangle\right\}$

⑤ $c_2(y) = E\left\{I(y \leq Y)e^{W^\top\beta-\beta^\top V\beta/2}\right\}$

3. Bootstrap Confidence Interval

특정 모수에 대한 신뢰구간을 구성하려면 추정량의 분산에 대한 정보가 필요하다. 예를 들어 모평균에 대한 신뢰구간은 $\overline{X} \pm z_{\frac{\alpha}{2}} \sigma / \sqrt{n}$과 같은 형태로 주어진다. 즉, 신뢰구간을 구하는 문제는 분산의 추정 문제로 귀결될 수 있다. 하지만, 이것 또한 추정량의 (점근적) 분포가 정규분포일 때에만 적용가능하며 만약 그렇지 않다면 분위수에 대한 추정이 추가로 필요할 것이다.

3절에서는 Bootstrap 신뢰구간의 3가지 종류인 Normal Interval, Pivotal Interval, Percentile Interval을 소개한다.

3.1. Normal Interval

만약 추정량의 분포가 (점근적으로) 정규분포라고 할 수 있다면, 분산을 bootstrap에 근거하여 추정하면 다음과 같은 신뢰구간을 얻을 수 있다.

$$\hat{\beta} \pm Z_{\alpha/2} \sqrt{Var^*(\hat{\beta})} \tag{3.1}$$

여기에서 $Var^*(\hat{\beta})$는 $Var(\hat{\beta})$의 bootstrap 추정량이다.

3.2. Pivotal Interval

만약 추정량의 분포가 정규분포라는 가정을 하기 어렵다면, 추정량의 표본분

포에 대한 정보를 그대로 이용하여 신뢰구간을 구하여야한다. β에 대한 추정량을 $\hat{\beta}$이라 하고 $\hat{\beta}-\beta$의 분포함수를 H하면 여기서 $\hat{\beta}-\beta$를 pivot이라고 한다. H의 하위 α분위수를 u_{α}라고 하면 $P(u_{\alpha/2} \leq \hat{\beta}-\beta \leq u_{1-\alpha/2}) = 1-\alpha$을 쓸 수 있다. 여기서 $u_{\alpha/2}, u_{1-\alpha/2}$를 bootstrap에 근거하여 추정하는 것인데 즉, B회의 샘플링에 의해 얻어진 B개의 추정량의 표본분위수로 $u_{\alpha/2}, u_{1-\alpha/2}$을 추정하고 이를 $u^*_{\alpha/2}, u^*_{1-\alpha/2}$라 하면 β에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$(\hat{\beta}-u^*_{1-\alpha/2}, \hat{\beta}-u^*_{\alpha/2}) \tag{3.2}$$

단, $u^*_{\alpha/2}, u^*_{1-\alpha/2}$는 $\hat{\beta}$가 아니라 $\hat{\beta}-\beta$의 분포의 분위수이다. β는 미지의 모수이므로 실제로는 $u^*_{\alpha/2}, u^*_{1-\alpha/2}$를 바로 얻는 것은 불가능하지만, $\hat{\beta}$의 분포의 분위수를 $v_{\alpha/2}, v_{1-\alpha/2}$라고 하면 $u_{\alpha/2} = v_{\alpha/2} - \beta$이 성립한다. 여기에서 $v_{\alpha/2}$는 bootstrap에 의해 추정가능하고, β는 $\hat{\beta}$로 추정 가능하므로 $u^*_{\alpha/2} = v^*_{\alpha/2} - \hat{\beta}$를 얻을 수 있다. 따라서 bootstrap에 의한 Pivotal Interval은 다음과 같다.

$$(2\hat{\beta}-v^*_{1-\alpha/2}, 2\hat{\beta}-v^*_{\alpha/2}) \tag{3.3}$$

3.3. Percentile Interval

Percentile Interval은 Pivotal Interval보다 더욱 간단한 형태로 주어지는 신

뢰구간이다. $\phi(\hat{\beta})-\phi(\beta)$가 0을 중심으로 대칭인 분포가 되는 적절한 변환함수 ϕ가 존재하는 경우에만 정당화 될 수 있는 것으로 알려져 있다. 즉, 편의추정량에 대해서는 좋은 신뢰구간을 제시할 수 없을 것이며 Percentile Interval은 다음과 같다.

$$(v^*_{\alpha/2}, v^*_{1-\alpha/2}) \tag{3.4}$$

4. 모의실험

모의실험은 오픈 소스인 R을 이용하여 진행되었으며, 모의실험은 Promotion Time Cure Model에서 measurement error의 분포별로 Ma & Yin(2008) 추정법으로 구한 신뢰구간과 bootstrap 방법으로 구한 3가지의 신뢰구간의 Coverage Probability와 Length of Confidence Interval을 비교하는 것이다. 모의실험에 사용할 measurement error의 분포는 4가지를 설정하였으며, 다음과 같다.

$$①U \sim N(0, V^2),\ \ ②U \sim \chi^2(\frac{0.1^2}{2}) - \frac{0.1^2}{2},\ \ ③U \sim Un(-\sqrt{0.03}, \sqrt{0.03}),$$

$$④U \sim \frac{t(3)}{10\sqrt{3}}$$

Ma & Yin(2008) 추정법의 measurement error의 분포에 따른 성능의 차이를 bootstrap 방법과 비교하고자 4가지 경우의 분포를 고려하였으며, 4가지 경우에 각각 bootstrap 횟수는 1000번, 반복수는 400번으로 설정하였다.

생존함수는 $S(t|X_1, X_2) = \exp\{-\exp(-0.3 + X_1 - 0.5X_2)F(t)\}, t > 0$을 고려한다. 여기서 $X_1 \sim Un(0,1)$, $X_2 \sim Ber(0.5)$이고 X_1은 $W = X_1 + U_1$이 관측되도록 measurement error의 영향을 받는다. U_1은 위에서 설정한 4가지의 분포를 사용하고, $E(U_1) = 0$, $Var(U_1) = 0.1^2$로 설정하였으며 카이제곱분포와 유니폼분포, t분포의 경우는 $E(U_1) = 0$, $Var(U_1) = 0.1^2$이 되도록 각각 자유도와 $[a, b]$를 설정하였다.

baseline CDF, $F(t)$에 대해 $t=20$에서 잘린 $\mu=6$의 지수분포를 사용하였고, 최대 이벤트 시간은 20이다. censoring time C는 공변량과 독립이며 $\mu=15$인 지수분포에서 추출했으며 $t=30$에서 잘리며, 평균 치유율은 39%, censoring 비율은 약 15%이다.

샘플의 크기는 $n=150$과 $n=250$ 2가지로 설정하였으며, 모의실험의 결과는 measurement error의 분포에 따라 Coverage Probability와 Length of Confidence Interval을 **[표 4.1]**부터 **[표 4.8]**에 나타내었다.

4.1. $U \sim N(0, V^2)$

<표 4.1> $U \sim N(0, 0.1^2)$**일 경우, Coverage Probability**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		0.920	0.930	0.953	0.958	0.960	0.945
Bootstrap	Normal	0.998	0.990	0.995	1.000	0.993	0.988
	Pivotal	0.998	0.983	1.000	0.985	0.990	0.990
	Percentile	0.940	0.945	0.938	0.960	0.948	0.945

● 소수점 4번째 자리에서 반올림하여 표기

<표 4.2> $U \sim N(0, 0.1^2)$**일 경우, Length of Confidence Interval**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		1.216	1.809	0.969	0.934	1.397	0.747
Bootstrap	Normal	2.115	2.727	1.391	1.559	1.879	0.967
	Pivotal	2.092	2.686	1.382	1.563	1.876	0.965
	Percentile	2.092	2.686	1.382	1.563	1.876	0.965

● 소수점 4번째 자리에서 반올림하여 표기

measurement error가 $U \sim N(0, 0.1^2)$인 경우에 **<표 4.1>** Coverage Probability와 **<표 4.2>** Length of Confidence Interval의 결과를 보면 $n=150$일

때, measurement error의 영향을 받는 X_1의 계수, β_1의 결과가 Ma & Yin(2008)의 추정법으로 구한 신뢰구간의 Coverage Probability가 bootstrap 방법으로 구한 신뢰구간의 Coverage Probability보다 성능이 떨어지는 것을 확인 할 수 있지만 신뢰구간의 길이는 Ma & Yin(2008)의 추정법으로 구한 신뢰구간이 bootstrap 방법으로 구한 신뢰구간보다 좁은 것을 확인할 수 있다. $n=250$일 때, Ma & Yin(2008)의 추정법으로 구한 신뢰구간의 Coverage Probability가 $n=150$일 때보다 정확해져 96%정도의 정확성을 보임을 확인할 수 있으며, 신뢰구간의 길이는 $n=150$일 때보다 차이가 줄기는 했지만 여전히 Ma & Yin(2008)의 추정법으로 구한 신뢰구간이 좁은 범위임을 확인 할 수 있다.

따라서 measurement error가 정규분포일 때, 모의실험의 결과는 Ma & Yin(2008)의 추정법이 bootstrap 방법보다 성능이 더 좋다는 것을 확인할 수 있다.

4.2. $U \sim \chi^2(\frac{0.1^2}{2}) - \frac{0.1^2}{2}$

카이제곱분포의 평균과 분산은 자유도가 k일 때 각각 $k, 2k$이다. 따라서 $Var(U_1) = 0.1^2$을 맞추기 위해서 $k = \frac{0.1^2}{2}$로 설정하였으며, $E(U_1) = 0$을 맞추기 위해서 생성된 난수에서 k만큼 빼서 분석을 하였다.

<표 4.3> $U \sim \chi^2(\frac{0.1^2}{2}) - \frac{0.1^2}{2}$일 경우, Coverage Probability

		$n = 150$			$n = 250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		0.945	0.910	0.938	0.903	0.870	0.933
Bootstrap	Normal	0.995	0.993	0.988	0.995	0.990	0.985
	Pivotal	0.995	0.948	0.990	0.978	0.898	0.985
	Percentile	0.980	0.938	0.940	0.953	0.923	0.948

● 소수점 4번째 자리에서 반올림하여 표기

<표 4.4> $U \sim \chi^2(\frac{0.1^2}{2}) - \frac{0.1^2}{2}$**일 경우, Length of Confidence Interval**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		1.219	1.835	0.976	0.924	1.370	0.746
Bootstrap	Normal	2.031	2.856	1.355	1.619	2.000	0.973
	Pivotal	2.022	2.823	1.351	1.616	1.976	0.968
	Percentile	2.022	2.823	1.351	1.616	1.976	0.968

● 소수점 4번째 자리에서 반올림하여 표기

measurement error, $U \sim \chi^2(\frac{0.1^2}{2}) - \frac{0.1^2}{2}$인 경우에 **<표 4.3>** Coverage Probability와 **<표 4.4>** Length of Confidence Interval의 결과를 보면 $n=150$일 때, 전체적으로 measurement error의 영향을 받는 X_1의 계수, β_1의 Coverage Probability가 β_0나 β_2의 Coverage Probability보다 성능이 떨어지는 것을 확인할 수 있다. Ma & Yin(2008)의 추정법은 좁은 범위의 신뢰구간에서 91%정도의 정확도가 나온 반면, bootstrap 방법은 Ma & Yin(2008)의 추정법보다는 넓은 범위의 신뢰구간에서 약 95%에 근접한 정확도가 나온 것을 확인할 수 있다. bootstrap 방법 중에서는 Percentile Interval과 Pivotal Interval이 같은 범위의 신뢰구간에서 비슷한 정확도가 나온 것을 확인할 수 있다. $n=250$일 때는 Ma & Yin(2008)의 추정법과 bootstrap 방법 중 Pivotal Interval이 신뢰구간의 범위가 좁아짐에 따라 성능이 $n=150$일 때보다 성능이 안 좋아지는 것을 확인할 수 있다. 반면,

bootstrap 방법 중 Percentile Interval의 경우 Pivotal Interval과 같은 범위의 신뢰구간을 구성하면서 92%정도의 정확도가 나온 것을 확인 할 수 있다.

따라서 measurement error가 카이제곱분포일 때, 모의실험 결과는 Ma & Yin(2008)의 추정법보다는 bootstrap 방법이 성능이 더 좋으며, 그 중 Percentile Interval이 가장 좋은 것을 확인할 수 있다.

4.3. $U \sim Un(-\sqrt{0.03}, \sqrt{0.03})$

유니폼분포의 평균과 분산은 각각 $\frac{1}{2}(a+b)$, $\frac{1}{12}(b-a)^2$이다. 따라서 $E(U_1)=0$을 맞추기 위해 $a=-b$로 하여, $\frac{1}{12}(2b)^2=0.1^2$을 만족하는 $b=\pm\sqrt{0.03}$으로 설정하여 분석을 하였다.

<표 4.5> $U \sim Un(-\sqrt{0.03}, \sqrt{0.03})$**일 경우, Coverage Probability**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		0.963	0.948	0.938	0.953	0.963	0.950
Bootstrap	Normal	0.998	0.988	0.995	0.998	0.998	0.993
	Pivotal	0.998	0.990	0.990	0.963	0.998	0.988
	Percentile	0.968	0.935	0.933	0.965	0.940	0.940

● 소수점 4번째 자리에서 반올림하여 표기

<표 4.6> $U \sim Un(-\sqrt{0.03}, \sqrt{0.03})$**일 경우, Length of Confidence Interval**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		1.203	1.808	0.972	0.941	1.398	0.750
Bootstrap	Normal	1.959	2.629	1.328	1.617	1.884	0.986
	Pivotal	1.949	2.610	1.321	1.619	1.884	0.983
	Percentile	1.949	2.610	1.321	1.619	1.884	0.983

● 소수점 4번째 자리에서 반올림하여 표기

measurement error, $U \sim Un(-\sqrt{0.03}, \sqrt{0.03})$인 경우에 **<표 4.5>** Coverage Probability와 **<표 4.6>** Length of Confidence Interval의 결과를 보면 $n=150$일 때, measurement error의 영향을 받는 X_1의 계수, β_1의 결과가 Ma & Yin(2008)의 추정법과 bootstrap 방법 중 Percentile Interval이 비슷하게 나왔지만 신뢰구간의 길이는 Ma & Yin(2008)의 추정법으로 구한 신뢰구간이 Percentile Interval의 신뢰구간보다 좁은 범위인 것을 확인할 수 있다. $n=250$일 때는 Percentile Interval의 신뢰구간이 $n=150$일 때보다 좁아지고 성능도 94%까지 좋아진 것을 확인할 수 있다.

따라서 measurement error가 유니폼분포일 때, 모의실험 결과는 bootstrap 방법 중 Percentile Interval이 Ma & Yin(2008)의 추정법 못지않게 성능이나 신뢰구간의 길이가 좋은 것을 확인할 수 있다.

4.4. $U \sim \frac{t(3)}{10\sqrt{3}}$

t-분포의 경우 자유도 k가 0보다 커야하기 때문에 $Var(U_1)=0.1^2$을 만족하는 자유도 k가 실제로는 존재하지 않는다. 따라서 자유도 k가 3인 t-분포 ($Var(U_1)=3$이 됨)에서 난수를 생성하여, 생성된 난수에서 $10\sqrt{3}$을 나눠 인위적으로 $Var(U_1)=0.1^2$으로 만들어서 분석을 하였다.

<표 4.7> $U \sim \frac{t(3)}{10\sqrt{3}}$일 경우, Coverage Probability

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		0.940	0.935	0.965	0.955	0.945	0.935
Bootstrap	Normal	0.998	0.995	0.993	0.995	0.983	0.990
	Pivotal	0.990	0.993	0.993	0.988	0.973	0.975
	Percentile	0.958	0.963	0.953	0.953	0.943	0.940

● 소수점 4번째 자리에서 반올림하여 표기

<표 4.8> $U \sim \frac{t(3)}{10\sqrt{3}}$**일 경우, Length of Confidence Interval**

		$n=150$			$n=250$		
		β_0	β_1	β_2	β_0	β_1	β_2
Ma & Yin(2008)		1.223	1.821	0.971	0.932	1.393	0.748
Bootstrap	Normal	2.046	2.735	1.362	1.575	1.883	0.970
	Pivotal	2.028	2.689	1.356	1.584	1.880	0.967
	Percentile	2.028	2.689	1.356	1.584	1.880	0.967

● 소수점 4번째 자리에서 반올림하여 표기

measurement error, $U \sim \frac{t(3)}{10\sqrt{3}}$인 경우에 **<표 4.7>** Coverage Probability와 **<표 4.8>** Length of Confidence Interval의 결과를 보면 $n=150$일 때, measurement error의 영향을 받는 X_1의 계수, β_1의 결과가 Ma & Yin(2008)의 추정법보다 bootstrap 방법의 성능이 좋은 것을 확인할 수 있다. 그중 Percentile Interval의 경우는 신뢰구간의 길이가 Ma & Yin(2008)의 추정법보다는 넓은 범위이지만 96%정도의 정확도로 성능이 좋은 것을 확인할 수 있다. $n=250$일 때는 Ma & Yin(2008)의 추정법의 정확도가 약 95%에 근접하게 향상되었으며 신뢰구간의 길이도 더 좁아진 것을 확인할 수 있다. Percentile Interval 또한 여전히 약 95%에 근접하는 정확도를 보이며 신뢰구간의 길이도 좁아져 $n=150$일 때와 마찬가지로 $n=250$일 때도 여전히 성능이 좋은 것을 확인할 수 있다.

따라서 measurement error가 t-분포일 때, 모의실험 결과는 Ma & Yin(2008)의 추정법은 샘플의 크기가 클 경우에 좋았지만, bootstrap 방법 중

Percentile Interval은 샘플의 크기에 상관이 없이 성능이나 신뢰구간의 길이가 좋은 것을 확인할 수 있다.

5. 결론

본 논문에서는 Promotion Time Cure Model에서 공변량이 measurement error의 영향을 받을 때, measurement error의 분포에 따른 Ma & Yin(2008) 추정법과 bootstrap 방법의 비교를 통해 Ma & Yin(2008) 추정법의 성능을 확인하였다. 먼저 Ma & Yin(2008) 추정법에 대해 설명하였으며, 다음으로 Bootstrap Confidence Interval의 3가지 종류에 대해 설명하였다. 추정법들의 성능을 measurement error의 분포에 따라 비교 및 Ma & Yin(2008) 추정법의 성능을 확인하기 위해 분포별 신뢰구간의 Coverage Probability와 Length of Confidence Interval을 계산하여 제시하였다. 모의실험 결과, Ma & Yin(2008) 추정법이 확실히 measurement error의 분포가 정규분포일 때는 성능이 우수하고, 정규분포와 유사한 형태인 0을 기준으로 대칭인 분포일 때도 성능이 나쁘지 않으나 카이제곱 분포일 때는 성능이 현저히 떨어지는 것을 확인할 수 있었다. bootstrap 방법의 경우 전체적으로 좋았으며 특히, 정규분포가 아닐 때는 Percentile Interval이 Ma & Yin(2008)의 추정법보다 우수한 것을 확인할 수 있었다. 또한, 샘플의 개수가 적을 때보다 많을 때 Ma & Yin(2008) 추정법은 성능의 변화 폭이 크며 신뢰구간의 길이는 더 좁게 추정되는 반면 bootstrap 방법 중 Percentile Interval은 성능의 변화 폭이 작으며 신뢰구간의 길이는 마찬가지로 좁게 추정된다.

measurement error가 정규분포일 때는 Ma & Yin(2008) 추정법이 적절하지만 0을 기준으로 대칭인 분포와 카이제곱분포를 포함한 그 외의 분포에서는 Ma & Yin(2008) 추정법을 사용하기에는 문제가 있다고 생각된다. 이러한 분포

에서는 신뢰구간의 길이가 길어 개선의 여지가 있지만 샘플의 크기에 영향을 덜 받으며 일정한 성능을 유지하고 특히, 카이제곱분포의 경우에는 Ma & Yin(2008)의 추정법보다 결과가 우수한 bootstrap 방법 중 Percentile Interval을 사용하는 것이 더 적합하다고 생각된다. 차후에는 measurement error가 0을 기준으로 대칭인 분포와 카이제곱분포를 포함한 그 외의 분포일 때, 새로운 추정법의 대한 연구나 bootstrap 방법 중 Percentile Interval의 신뢰구간 길이에 대한 보완이 필요하다고 생각된다.

참고문헌

[1] Yanyuan Ma and Guocheng Yin (2008). Cure Rate Model with Mismeasured Covariates under Transformation. Journal of the American Statistical Association, Vol. 103, No. 482, pp.743-756.

[2] Biometrika (2017). Inference in a survival cure model with mismeasured covariates using a simulation-extrapolation approach. 104, 1, pp.31-50.

[3] David G. Kleinbaum and Mitchel Klein (2011). Survival Analysis A Self-Learning Text. Third Edition.

[4] AURÉLIE BERTRAND (2017). Survival models with a cure fraction and mismeasured covariates.

[5] Laska, E. M., and Meisner, M. J. (1992). Nonparametric Estimation and Testing in a Cure Rate Model. Biometrics, 48, 1223-1234.

[6] Cook, J. R., and Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statisti cal Association, 89,1314-1328.

[7] Kulich, M., and Lin, D. Y (2000). Additive Hazards Regression With Covariate Measurement Error. Journal of the American Statistical Association, 95, 238-248.

[8] Sy, J. P., and Taylor, J. M. G. (2000). Estimation in a Cox Proportional Hazards Cure Model. Biometrics, 56, 227-236.

[9] Song, X., and Huang, Y (2005). On Corrected Score Approach for Proportional Hazards Model With Covariate Measurement Error. Biometrics,

61, 702-714.

[10] Nakamura, T (1990). Corrected Score Function for Errors-in-Variables Models: Methodology and Application to Generalized Linear Models. Biometrika,11, 127-137.

[11] Nakamura, T (1992). Proportional Hazards Model With Covariates Subject to Mea surement Error. Biometrics, 48, 829-838.