The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.5.753
32(5), 753–761

A concordance test for bivariate interval censored data using a leverage bootstrap

Yang-Jin Kim[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received August 27, 2019; Revised September 22, 2019; Accepted October 4, 2019)

Abstract

A test procedure based on a Kendall's τ statistic is proposed for the association of bivariate interval censored data. In particular, a leverage bootstrap technique is applied to replace unknown failure times and a classical adjustment method is applied for treating tied observations. The suggested method shows desirable results in simulation studies. An AIDS dataset is analyzed with the suggested method.

Keywords: AIDS study, Association, Bivariate interval censored data, Kendall's τ, Leverage bootstrap

1. 서론

두 변수간의 연관성에 대한 다양한 검정방법들이 연속형과 범주형 자료에서 개발되었다. 다양한 종류의 중도 절단자료(censoring)를 포함한 생존 자료 분석에서도 두 변수들간의 연관관계에 대한 많은 방법들이 개발되어왔다 (Kalbfleish와 Prentice, 2002). 이러한 연관성 연구를 통해 질병의 발병원인을 추적하거나 향후 발병에 대한 주요한 지시요인을 규명할 수 있다. 예를 들어, 일란성 쌍둥이를 대상으로 한 유방암 발생 여부에 대한 연구에선 쌍둥이 자매들간의 유방암 발병 시점 연관성은 환경 인자의 영향력을 측정하는 데 도움이 될 수 있다. 또 다른 예로 후천성 면역 결핍 증후군(acquired immune deficiency syndrome; AIDS)을 가진 환자의 인간 면역 결핍 바이러스(human immunodeficiency virus; HIV) 감염 시점과 바이러스 잠복 시간과의 연관성 여부를 통해 신약의 효과성을 검증할 수 있다. 생존 자료 분석의 주요 목적은 중도 절단을 통한 불완전한 자료 형태로 인한 정보의 손실을 최소화하고자 한다. 이를 위한 한 가지 방안으로 적절한 보정 또는 재표본 방법을 이용하여 기존의 통계량의 확장을 고려해왔다. 그 예로, 본 연구의 목적인 두 변수간의 연관성 연구를 위해 Kendall's τ 통계량의 적용이 고려되어진다. Kendall's τ는 중도 절단된 이변량 생존 자료에서 가장 널리 사용되는 통계량으로 계산의 용이성과 검정력이 여러 논문들을 통해 논의되었다 (Brown 등, 1974; Oakes, 1982, 2008; Wang과 Wells, 2000; Lakhal-Chaieb 등, 2009). Kendall's τ 통계량은 분포의 가정없이 두 변수쌍의 순위의 일치성을 통해 계산된다. 좀 더 자세하게 설명하면, τ는 두 개의 확률 차, 일치쌍(concordant pair)의 확률과 비일치쌍(discordant pairs)의 확률 차이로 이변량 자료 (T_1, T_2)에 대해

$$\tau = \Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\} - \Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\}.$$

This work was supported by the Korea National Research fund (NRF-2017R1D1A1B03030578).
[1]Department of Statistics, Sookmyung Women's University, 100-gil Chengpa-Dong, Yongsan-Gu, Seoul 04310, Korea. E-mail: yjin@sookmyung.ac.kr

로 표현되는데 Kendall's τ의 범위는 $-1 \le \tau \le 1$이며 $\tau = 0$은 두 변수의 독립성을 의미한다. 완전한 자료 하에서 τ는 다음과 같이 추정될 수 있다.

$$\hat{\tau} = \frac{\sum_{1 \le i < j \le n} a_{ij} b_{ij}}{\binom{n}{2}},$$

여기서 $T_{1i} > T_{1j}$이면 $a_{ij} = 1$이고 $T_{1i} < T_{1j}$이면 $a_{ij} = -1$으로 정의된다. 비슷하게 $T_{2i} > T_{2j}$이면 $b_{ij} = 1$이고 $T_{2i} < T_{2j}$이면 $b_{ij} = -1$이 된다. 따라서 $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0$에 대해서는 $a_{ij}b_{ij} = 1$이며 $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0$에 대해서는 $a_{ij}b_{ij} = -1$이 된다. 또 다른 추정량으로 다음의 결합 분포와의 관계를 이용할 수 있다.

$$\begin{aligned} \tau = T(F) &= 4 \Pr(T_{1i} > T_{1j}, T_{2i} > T_{2j}) - 1 \\ &= 4 \int_0^\infty \int_0^\infty S(x,y) dS(x,y) - 1, \end{aligned} \tag{1.1}$$

여기서 $S(x,y) = \Pr(T_1 \le x, T_2 \le y)$는 이변량 생존 함수(bivariate survival function)이며 $S_F = \{(x,y) : S(x,y) > 0\}$의 범위안에서 정의된다. 따라서 S의 추정량 $\hat{S}$을 이용하여 $\hat{\tau} = T(\hat{S}) = \hat{\tau}_0$를 추정할 수 있다.

$\hat{\tau}_0$의 대표본 성질은 이른바 functional δ 방법에 의해 $\sqrt{n}\{\hat{S}(x,y) - S(x,y)\}$의 근사적 분포가 평균이 0인 Gaussian process을 만족할 때, $\sqrt{n}(\hat{\tau}_0 - \tau)$이 근사적으로 정규분포로 수렴함이 증명되었다 (Gill, 1989). 하지만 Wang과 Wells (2000)은 중도절단에 의한 불완전한 영역(incomplete supports)으로 인해 중도 절단 자료에 대한 추정량 (1.1)의 적용은 과소추정됨을 보였다. 본 연구에서는 이변량 구간 중도 절단 자료의 연관성을 검정하기 위해 Kendall's τ 통계량을 이용한 검정방법을 제안한다. 구간 중도 절단 자료(interval censored data)는 관심있는 사건의 정확한 발생 시점 대신에 그 시점을 포함한 구간을 포함한 자료이다. 이러한 자료형태는 관측연구가 있는 종단 연구(longitudinal study)에서 발생되거나 물리적으로 정확한 발병 시점 관측이 불가능한 임상연구(예를 들어 종양 발생 시점)에서 흔히 관측될 수 있다. 이러한 구간 중도 절단 자료는 일반적인 우중도 절단자료보다 자료의 불완전성이 더 심각할 수 있으며 이는 곧 추정량의 근사적 성질에도 영향을 미치게 된다. 예를 들어, 우중도 절단 자료하에서 생존함수의 추정치, Kaplan-Meier 추정량의 수렴 속도가 $\sqrt{n}$인데 반해, 구간 중도 절단 자료하에서의 분포 추정량 $\hat{S}$는 이보다 더 느린 수렴속도를 가지게 된다 (Sun, 2006). 그럼에도 불구하고 여러 방법을 통해 구간 중도 절단 자료하에서 다양한 회귀분석방법들이 개발되어 왔으며 이들 회귀계수는 일반적인 수렴속도와 함께 정규분포를 가짐이 증명되었다. 이변량 구간 중도 절단 자료에 대한 연구에 대해서도 여러 학자들에 의해 논의되었다. Betensky와 Finkelstein (1999)은 다중 대체 방법(multiple imputation)을 이용하여 구간중도 절단에 대한 τ를 추정하고자 하였다. 하지만 이 방법은 $\tau \neq 0$일 때 편의 추정량을 가져올 수 있다. Bogaerts와 Lesaffre (2008)는 이변량 구간 중도 절단 자료에 모수 분포를 가정하였다. 즉, 이변량 로그 정규 분포를 가정한 후 적절한 격자점과 해당되는 결합 확률 질량 함수를 추정한 후 이들값을 이용하여 τ를 추정하였다. copula 함수의 이변량 구간 중도절단자료의 연관성 연구에 대한 적용도 여러 학자들에 의해 고려되었다. Ding과 Wang (2004)은 이변량 current status data의 연관성을 추정하기 위해 Clayton copula 함수를 이용한 이단계 추정을 적용하였다. 일 단계에서는 비모수 방법을 이용하여 결합 분포를 추정하였으며 이 단계에서는 copula 계수를 각각 추정하였다. 이들 방법은 Sun 등 (2006)에 의해 이변량 구간 중도 절단 자료로 확장되었다. 본 논문에서 이변량 구간 중도 절단자료의 연관성을 추정하고 검정하기 위해 재표본 방법인 붓스트랩 방법의 일종인 지렛대 붓스트랩(leverage bootstrap)를 적용하고 한다.

먼저 2장에서는 지렛대 붓스트랩에 대한 소개를 하며 일변량 구간 중도 절단자료의 적용예를 정리한다.

3장에서는 Kendall's τ 추정을 위한 이 방법을 확장한다. 4장에서는 제안된 방법의 적절성을 검정하기 위해 몇가지 모의실험을 실시하고 실제 자료에 적용하고자 한다. 관련된 향후 연구를 5장에서 제안한다

2. 지렛대 붓스트랩(leverage bootstrap)

본 장에서 일변량 구간 중도절단 자료에 대한 지렛대 붓스트랩방법의 적용을 정리한다 (Ren, 2003).

$F(x) = \Pr(X \le x)$를 $X_1, \ldots, X_n$의 분포함수라 하고 F_0를 알려진 연속형 분포함수라 할 때, 다음의 가설을 검정하고자 한다.

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0. \tag{2.1}$$

위 귀무가설을 검정하기 위해 다음의 Cramer-Von Mises 검정 통계량이 적용될 수 있다.

$$\hat{T}_n = n \int_0^\infty \left(\hat{F}_n(x) - F_0(x) \right)^2 dF_0(x),$$

여기서 $\hat{F}_n$은 귀무가설하에서 완전 자료$(X_1, \ldots, X_n)$의 경험적 분포 함수(empirical distribution)라 할 때, $n \to \infty$일 때, $\hat{T}_n \to W$를 만족하게 되는데 여기서 W는 가우시안 확률과정이다 (Shorack과 Wellner, 1986). 이러한 성질은 우중도 절단자료하에서도 적용이 되며 $\hat{F}_n$을 이용하여 일그룹 적합도 검정에 적용될 수 있다 (Andersen 등, 1993). 그러나 이러한 함수 대입 추정량(functional plug-in method)은 구간 중도 절단 자료에 대해 적합하지 않다. 그 이유는 $\hat{F}_n$의 수렴 속도와 관련되어 있는데 구간 중도 절단의 관측 영역에 대한 성김성(coarseness) 때문에 어느 시점에서는 일반적인 수렴속도 $\sqrt{n}$보다 느린 $n^{1/3}$을 가지게 된다. 이러한 불안정한 성질 때문에 $\hat{T}_n$의 구간 중도 절단 자료에 대한 적용은 적합하지 못하다. 이를 극복하기 위해 Geskus와 Groeneboom (1999)는 평활 함수 $K(\cdot)$를 적용하는 것을 고려해 보았다. 그들의 연구에 의하면 $\sqrt{n}(K(\hat{F}_n) - K(F))$은 근사적으로 정규분포를 따르지만 귀무 가설 (2.1)하에서는 $\hat{T}_n$에 대한 표준화 과정을 유도할 수 없어 이러한 근사적 성질이 적용될 수 없었다. 이에 Ren (2003)은 붓스트랩 방법의 변형인 지렛대 붓스트랩 기법을 구간 중도 절단자료에 적용할 것을 제안하였다. 일반적인 비모수 붓스트랩 방법은 원자료로부터 추출된 붓스트랩 표본을 이용하여 관심있는 추정량 계산과 검정 통계량을 유도한다. 즉, $\hat{F}_n$으로부터 추출된 붓스트랩 표본 $(\hat{X}_{n1}, \ldots, \hat{X}_{nm})$을 이용하여 구한 $\hat{H}_n(\hat{X}_{n1}, \ldots, \hat{X}_{nm})$은 근사적으로 $H_n(X_1, \ldots, X_n)$에 수렴함이 알려져 있다 (Efron, 1967). 여기서 $H_n(X_1, \ldots, X_n)$은 관측된 자료로부터 계산되는 추정량 또는 검정 통계량이다. 하지만 일반적인 붓스트랩 표본의 구간 중도 절단자료에 대한 적용은 바람직하지 못한 결과를 가져왔다. 즉, 구간 중도 절단 자료에 대한 일반적인 붓스트랩 방법의 적용은 여전히 불완전한 자료를 가져왔으며 $\tilde{H}_n$의 점근적 성질 (수렴성과 수렴속도)의 향상에 기여하지 못하였다.

이러한 자료의 불완전성과 관련된 문제점을 극복하기 위해, 지렛대 붓스트랩의 적용이 고려되었다. 즉, $\hat{F}$으로부터 완전 자료를 샘플링함으로써, 더 이상 구간 중도 절단된 자료를 사용할 필요가 없다. 더 자세하게 재표본 과정을 설명하면 첫 번째 단계에서는 구간 중도 절단 자료를 이용하여 $\hat{F}_n$를 구하는 것이다. 두 번째 단계에서는 $\hat{F}_n$에서 모의 완전 자료(pseudo complete data)인 $(T^*_{n1}, \ldots, T^*_{nm})$을 랜덤추출하게 된다. 이를 지랫대 붓스트랩 표본(leverage bootstrap data)이라 한다. 여기서 m은 재표본 수를 의미하며 $n \to \infty$일 때, $m \to \infty$와 $m/n \to 0$을 만족시킨다. 세 번째 단계에서는 이러한 붓스트랩 표본을 이용하여 통계적 추론 방법을 적용한다. Figure 2.1은 일반적인 붓스트랩 방법과 지렛대 붓스트랩 방법을 보여준다 (Ren, 2003). Ren (2003)은 귀무가설 (2.1)을 검정하기 위해 그리고 Yuen 등 (2006)은 K 표본 동일성 검정 $(H_0 : F_1 = \cdots = F_K)$을 위해 지렛대 붓스트랩 방법을 각각 적용하였으며 점근적 성질을 증명하였다. 이 두 연구에서 m에 대한 선택은 표본 n에 근거하여 유도되었다. 본 연구에서는 그들이 유도한 공식을 이용하여 계산된 m을 적용할 것이다.

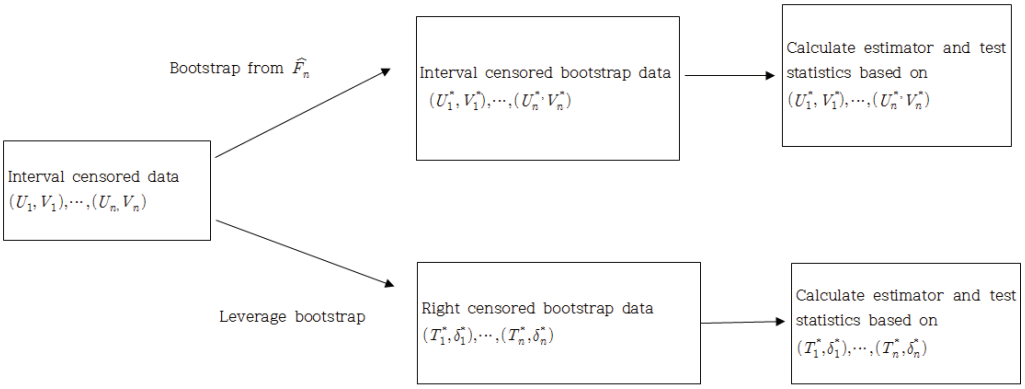

Figure 2.1. Application of leverage bootstrap and ordinary bootstrap to interval censored data.

3. 이변량 구간 중도절단 자료의 연관성 검정

(T_1, T_2)은 이변량 생존 함수 $S(t_1, t_2) = \Pr(T_1 \geq t_1, T_2 \geq t_2)$를 가지는 이변량 확률 변수이다. 이변량 구간 중도 절단 자료는 (U_1, V_1)와 (U_2, V_2)로 표현되면 이때 이들 쌍은 $U_1 < T_1 \leq V_1$와 $U_2 < T_2 \leq V_2$을 만족하게 된다. 일반적으로 이러한 구간 중도 절단자료는 관측 연구 또는 혈액 관련 연구등에서 발생된다. 예를 들어, AIDS 환자의 정확한 HIV 감염시점과 AIDS 발현시점은 알려져 있지 않다. 이는 체내 감염 메카니즘과 혈청 관련 시스템으로 결정되는 것으로 정확한 시점을 관측하는 것은 불가능하다. 대신 병원 또는 보건소와 같은 의료기관을 방문하여 혈액 채취 후 여러가지 검사를 통해서만 감염여부와 발현여부를 확진받게 된다. 여기서 이러한 방문시점(또는 관측 시점)은 관심있는 사건의 발생시간과 일반적으로 독립적이라고 간주한다. 이를 비정보적 중도절단(noninformative censoring) 또는 비정보적 관측(noninformative observational time)이라고 한다. 만약 이 가정이 위배되는 경우 이를 고려한 방법이 적용되어야 한다 (Kim, 2006). 이제 우리의 관심인 지렛대 붓스트랩방법을 이용한 이변량 연관성 검정을 위해 다음의 단계로 적용해 보자.

1단계. 이변량 구간 중도 절단 자료에 대한 이변량 분포 함수 $F(t_1, t_2) = \Pr(T_1 \leq t_1, T_2 \leq t_2)$를 추정한다. 이를 위해 다양한 방법이 적용될 수 있으며 본 연구에서는 비모수 방법을 이용한다. 특히 이변량 구간 중도 절단의 비모수 최대 우도 추정량은 우도 함수 $L(S) = \prod_{i=1}^{n} F(D_i)$,

$$F(D_i) = F(V_{1i}, V_{2i}) - F(V_{i1}, U_{2i}) - F(U_{i1}, V_{2i}) + F(U_{i1}, U_{i2})$$

를 최대화를 만족시키는 것으로 이를 위해 우도 함수는 $L(p) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} p_j$로 표현된다. 여기서 $p = (p_1, \ldots, p_m)$, $p_j = F(H_j) = F(s_{1j}, s_{2j}) - F(r_{1j}, s_{2j}) - F(s_{1j}, r_{2j}) + F(r_{1j}, r_{2j})$이다. 지시함수 $\alpha_{ij} = I(H_j \in (U_{i1}, V_{i1}] \times (U_{i2}, V_{i2}])$로 정의되면 $H_j = \{(r_{1j}, s_{1j}] \times (r_{2j}, s_{2j}]\}$로 분리된 사각형(disjoint rectangle)을 표현한다. 이러한 분리된 사각형을 구하기 위해 다양한 알고리즘들이 여러 학자들에 의해 제안되었다 (Betensky와 Finkelstein, 1999; Gentleman과 Vandal, 2002; Bogaerts와 Lesaffre, 2004). 특히 Maathuis (2005)가 개발된 R 패키지 MLEcens가 제공하는 알고리즘은 가장 빠른 속도를 제공하는 것으로 알려져 있다.

2단계. 적절하게 선택된 m을 이용하여 모의(pseudo) 완전 이변량 생존 시간들로 구성된 지렛대 표본 $\{t_{1l}^{*b}, t_{2l}^{*b},\ l = 1, \ldots, m;\ b = 1, \ldots, B\}$을 구한다. 이 표본을 이용하여 $\hat{\tau}^{*b}$와 분산을 구한다.

3단계. 2단계에서 $(\hat{\tau}^{*1}, \ldots, \hat{\tau}^{*B})$와 분산을 이용하여 검정 통계량 $(z^{*1}, \ldots, z^{*B})$을 구한다. 새로운 추정값은 B개의 $\hat{\tau}^*$ 값의 평균이 된다. 유의 수준 α 값에 해당되는 임계치와 절대치를 비교하여 더 큰값의 비율로 검정력을 계산하였습니다. 이를 이용하여 귀무가설 $(H_0 : \tau = 0)$의 기각 여부를 결정한다.

2단계에서 추출된 모의 완전 이변량 생존 시간은 동점 자료를 포함할 수 있으며 따라서 이러한 동점 자료의 효과를 보정하기 위해 다음의 통계량이 적용된다 (Kendall과 Gibbons, 1990)

$$\hat{\tau}_b = \frac{\tilde{S}}{\sqrt{\frac{1}{2}n(n-1) - W}\sqrt{\frac{1}{2}n(n-1) - Q}}, \tag{3.1}$$

여기서 $\tilde{S}$는 일치 쌍수과 비일치 쌍수의 차이이며 W와 Q는 다음과 같이 정의된다.

$$W = \frac{1}{2}\sum_j w_j(w_j - 1), \qquad Q = \frac{1}{2}\sum_l q_l(q_l - 1),$$

여기서 w_j와 q_l는 T_1과 T_2에서 동점수이다. 귀무가설 $H_0 : \tau = 0$을 검정하기 위해, 식 (3.1)의 분산은 다음과 같이 정의된다.

$$\begin{aligned}\mathrm{Var} = \frac{1}{18}&\left\{m(m-1)(2m+5) - \sum u(u-1)(2u+5) - \sum v(v-1)(2v+5)\right\} \\ &+ \frac{1}{9m(m-1)(m-2)}\left[\sum u(u-1)(u-2)\right]\left[\sum v(v-1)(v-2)\right] \\ &+ \frac{1}{2m(m-1)}\left[\sum u(u-1)\right]\left[\sum v(v-1)\right],\end{aligned}$$

여기서 검정 통계량 $Z^b = \hat{\tau}_b/\sqrt{\mathrm{Var}} \sim N(0,1)$을 따르게 된다.

4. 모의 실험과 자료 분석

4.1. 모의실험

제안된 방법의 적합성을 확인하기 위해 다양한 τ 값들하에서 추정량의 편이여부와 검정력을 조사한다. 각 모의 실험은 500번의 반복이 시행되며 세 가지 표본 크기 $(n = 100, 200, 400)$이 적용된다. 여기서 $H_0 : \tau = 0$은 두 변수의 독립성을 의미한다. 그 밖에 여러 가지 τ 값들을 적용함으로써 검정력을 검토하고자 한다. 이변량 생존 시간을 생성하기 위해 Clayton 모형이 적용된다. 즉, α 값이 주어질 때, 다음의 결합 생존 함수를 구할 수 있다. 여기서 $\alpha = (0.5, 1, 2, 8)$는 $\tau = (0.2, 0.3, 0.5, 0.8)$를 의미한다.

$$S(t_1, t_2) = \left[S_1^{-\alpha}(t_1) + S_2^{-\alpha}(t_2) - 1\right]^{-\frac{1}{\alpha}},$$

여기서 주변 생존 함수는 $S_k(t) = \exp(-0.5t)$, $k = 1, 2$이다. 구간 중도 절단 시간을 생성하기 위해 20개의 방문 소요 시간을 균일분포 $s \sim U(0, 0.5)$로부터 생성한다. 이때, 구간 중도 절단 시간 (U_{ki}, V_{ki})은

$$U_{ki} = \sum_{j=1}^{l-1} s_j < T_{ki} < \sum_{j=1}^{l} s_j = V_{ki}, \quad k = 1, 2$$

으로 정한다. Table 4.1과 Table 4.2는 서로 다른 m 값들을 이용하 모의실험을 시행한 결과를 보여준다. 즉, Table 4.1은 $m = n^{1/3}$, Table 4.2에서는 $m = n^{4/9}$를 적용한 결과를 보여준다. 여기서 200개의

Table 4.1. Test size and empirical power with $m = n^{1/3}$ at $n = 100$ and 200

	τ	$n = 100$		$n = 200$		$n = 400$	
		Estimate	Emp. power	Estimate	Emp. power	Estimate	Emp. power
Independence	0.00	0.002	0.043	−0.003	0.050	0.002	0.054
Dependence	0.20	0.208	0.053	0.200	0.086	0.202	0.118
	0.33	0.330	0.086	0.338	0.154	0.331	0.217
	0.50	0.502	0.162	0.503	0.308	0.502	0.325
	0.80	0.805	0.542	0.807	0.791	0.800	0.876

Table 4.2. Test size and empirical power with $m = n^{4/9}$ at $n = 100$ and 200

	τ	$n = 100$		$n = 200$		$n = 400$	
		Estimate	Emp. power	Estimate	Emp. power	Estimate	Emp. power
Independence	0.00	0.004	0.067	−0.002	0.0601	0.001	0.056
Dependence	0.20	0.202	0.126	0.207	0.155	0.201	0.183
	0.33	0.338	0.237	0.331	0.319	0.335	0.412
	0.50	0.501	0.478	0.506	0.644	0.504	0.769
	0.80	0.802	0.942	0.809	0.991	0.810	0.998

붓스트랩 표본($B = 200$)이 사용되었다. 표에서는 추정된 $\hat{\tau}$의 평균과 검정력(500번의 반복 시행중에 0.05의 명목값과 비교하여 귀무가설을 기각할 비율)을 보여준다. 모든 경우에 대해서 추정량은 불편 추정량을 만족하며 표본의 크기가 커질수록 검정력(empirical power)이 커짐을 확인할 수 있다. 특히 두 표의 결과를 비교해볼 때, $m = n^{1/3}$이 귀무가설하에서의 type 1 error인 0.05에 더 가까운 값을 보여주었다.

4.2. AIDS 자료에 대한 적용

앞 절에서 제안된 방법을 이용해서 AIDS 환자의 HIV 감염시간과 바이러스 잠복시간과의 연관관계 여부를 검정하고자 한다. 분석할 자료는 수혈로 인해 바이러스에 감염된 188명 혈우병 환자에 대한 기록이다. 이들 환자 중 집중 치료를 받은 97명의 환자를 대상으로 두 시점간의 연관관계를 조사하고자 한다. 본 자료는 De Gruttola와 Lagakos (1989)와 Kim (2006)에 의해 분석되었다. 바이러스 감염시점과 AIDS 발현 시점을 각각 T와 Y라고 할 때, 이들 사건의 정확한 발생시점은 물리적으로 관측할 수 없다. 왜냐하면 이러한 진단은 혈액 체취 후 정밀 검사 결과가 필요하기 때문이다. 따라서 분석 자료는 환자가 마지막으로 음성 반응을 가진 병원 방문 시점(TL)과 양성 반응을 가진 첫 번째 병원 방문시점(TR)으로 구성된 구간 중도 절단 자료가 된다. 즉, 바이러스 감염 시점 T와 두 병원 방문 시점 (TL, TR)은 $(TL < T < TR)$의 관계를 가진다. 비슷하게 AIDS 발현 시점도 혈액 체취를 위한 병원 방문 시점에 의해 구간 중도 절단되어진다 $(YL < Y < YR)$. 본 연구의 주요 관심은 바이러스 감염시점과 바이러스 잠복 시간과의 연관관계이므로 잠복 시간 $(Z = Y - T)$을 정의할 필요가 있다. 여기서 잠복 시간은 HIV 감염시점에서부터 AIDS 발현시간까지로 정의되므로 잠복 시간 또한 구간 중도 절단됨을 알 수 있으며 이는 $(ZL = YL - TR < Z < YR - TL)$으로 표현된다. 귀무가설 $H_0 : \tau = 0$을 검정하기 위해 200 지렛대 붓스트랩 표본이 사용되었으며 각 표본에서는 두 가지 m값이 각각 적용되었다. $m = n^{1/3} \approx 5$가 적용되었을 때, $\hat{\tau} = -0.119$와 p-value $= 0.222$이었으며 $m = n^{4/9} \approx 8$가 적용될 경우, $\hat{\tau} = -0.110$로 p-value $= 0.220$로 두 경우에서 모두 귀무가설이 기각되지 못했다. 즉, 바이러스 감염과 잠복 시간간에는 음의 연관관계가 있었지만 통계적으로 유의미하지는 않았다.

5. Discussion

본 논문에서는 이변량 구간 중도 절단 자료의 연관성 여부를 검정하기 위해 붓스트랩 방법을 적용하였다. 일반적인 붓스트랩 방법을 적용할 경우 τ에 대한 편의 추정량을 보여주었다. 이에 좀 더 완전한 자료 형태를 구하기 위한 지렛대 붓스트랩 방법이 적용되었다. Ren (2003)과 Yuen 등 (2006)은 이 방법을 일변량 구간 중도 절단 자료에 적용하여 일표본 적합도 검정 ($H_0 : F = F_0$)과 k 표본 동질성 ($H_0 : F_1 = \cdots = F_k$) 검정을 실시하였다. 그들의 연구를 이변량 구간 중도 절단 자료에 적용함으로 τ 통계량에 근거한 검정방법을 제안하였다. 특히 본 논문에서 사용한 지렛대 붓스트랩 방법은 Bickel과 Ren (2001)이 제안한 m out of n 붓스트랩 방법과 매우 유사하며 여기서 재표본 크기 m은 표본 크기의 함수로 유도되었다. 본 논문에서 Yuen 등 (2006)에서 제시된 두 가지 m을 적용하였다. 하지만 이 공식은 일변량 구간 중도 절단 자료에 근거한 것으로 본 논문에 적용하는 것은 부적절할 수 도 있다. 그럼에도 불구하고 모의실험의 결과에 의하면 추정값은 참값에 가까웠으며 검정력도 만족스러웠다. 본 논문의 이러한 한계를 인지하고 이변량 검정에 적합한 m의 공식을 유도를 향후 연구 주제로 삼고자 한다.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Betensky, R. and Finkelstein, D. F. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine*, **18**, 3101–3109.

Bickel, P. J. and Ren, J. J. (2001). The bootstrap in hypothesis testing, *Lecture Notes-Monograph Series State of the Art in Probability and Statistics*, **36**, 91–112.

Bogaerts, K. and Lesaffre, E. (2004). A new fast algorithm to find the regions of possible support for bivariate interval censored data, *Journal of Computational and Graphical Statistics*, **13**, 330–340.

Bogaerts, K. and Lesaffre, E. (2008). Estimating local and global measures of association for bivariate interval censored data with a smooth estimate of the density, *Statistics in Medicine*, **28**, 5941–5955.

Brown, B. W., Hollander, M., and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies, *Reliability and Biometry*, 327–354.

De Gruttola, V. and Lagakos, S. E. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**, 1–12.

Ding, A. A. and Wang, W. (2004). Testing independence for bivariate current status data, *Journal of the American Statistical Association*, **99**, 145–155.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 831–853.

Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data, *Canadian Journal of Statistics*, **30**, 557–571.

Geskus, R. and Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval-censoring, case2, *Annals of Statistics*, **27**, 627–674.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I), *Scandinavian Journal of Statistics*, **16**, 97–128.

Kalbfleish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed), John Wiley, New York.

Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*, Oxford University Press, Oxford.

Kim, Y. J. (2006). Regression analysis of doubly censored data with frailty, *Biometrics*, **62**, 458–464.

Lakhal-Chaieb, L., Rivest, L. P., and Beaudoin, D. (2009). IPCW estimator for Kendall's tau under bivariate censoring, *The International Journal of Biostatistics*, **5**, 8.

Maathuis, M. H. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval censored data, *Journal of Computational and Graphical Statistics*, **14**, 352–362.

Oakes, D. A. (1982). Concordance test for independence in the presence of censoring, *Biometrics*, **38**, 451–455.

Oakes, D. A. (2008). On consistency of Kendall's tau under censoring, *Biometrika*, **95**, 997–1001.

Ren, J. J. (2003). Goodness of fit tests with interval censored data, *Scandinavian Journal of Statistics*, **30**, 211–226.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, Wiley, New York.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Sun, L., Wang, L., and Sun, J. (2006). Estimation of the association for bivariate interval-censored failure time data, *Scandinavian journal of Statistics*, **33**, 637–649.

Yuen, K. C., Shi, J., and Zhu, L. (2006). A k-sample test with interval censored data, *Biometrika*, **93**, 315–328.

Wang, W. and Wells, M. (2000). Estimation of Kendall's tau under censoring, *Statistica Sinica*, **10**, 1199–1215.

지렛대 붓스트랩을 이용한 이변량 구간 중도 절단 자료의 일치성 검정

김양진[a,1]

[a]숙명여자대학교 통계학과

(2019년 8월 27일 접수, 2019년 9월 22일 수정, 2019년 10월 4일 채택)

요 약

본 논문에서는 이변량 구간 중도 절단 자료의 연관성 검정을 연구하고자 한다. Kendall's τ 통계량은 분포의 가정을 필요로 하지 않는 비모수방법으로 연관성 검정을 위해 빈번히 적용되고 있다. 본 논문에서도 이러한 τ 통계량을 이용한 검정을 하기 위해 붓스트랩 방법을 적용시킨다. 일반적인 비모수 붓스트랩 방법의 구간 중도 절단에 적용은 편의된 결과를 보여주었다. 이는 구간 중도 절단자료의 불완전성(incompleteness)과 관련된 것으로 이를 극복하기 위해 지렛대 붓스트랩 방법을 적용하였다. 추정된 분포에 근거하여 구간 중도 절단 대신 모의 완전한 표본(pseudo complete data)을 추출하는 것이다. 본 논문에서는 재표본의 크기 m을 결정하기 위해 기존 연구자의 공식을 이용하였다. 시행된 모의 실험의 결과는 바람직한 제 1종 오류값과 좋은 검정력을 보였주었으며 실제 적용 예로 AIDS 자료에서 HIV 감염시점과 바이러스 잠복 시간과의 연관성 여부를 검정해보았다.

주요용어: AIDS 연구, 이변량 구간 중도 절단, 연관성, Kendall's τ, 지렛대 붓스트랩

본 연구는 한국연구재단의 연구비 지원에 의해 시행되었습니다 (NRF-2017R1D1A1B03030578).

[1](04310) 서울시 용산구 청파로 길 100, 숙명여자대학교 통계학과. E-mail: yjin@sookmyung.ac.kr