응용통계연구 (2011)
24(1), 207–216
DOI: 10.5351/KJAS.2011.24.1.207

일반화추정방정식(GEE)에 대한 부스트랩의 적용

박종선[1] · 전용문[2]

[1]성균관대학교 통계학과, [2]성균관대학교 통계학과

(2010년 10월 접수, 2010년 12월 채택)

요 약

본 논문에서는 일반화추정방정식(GEE)모형에 대한 부스트랩 방법의 적용에 대하여 살펴본다. 다양한 부스트랩 방법들 중 GEE모형에 적용이 가능한 잔차, 쌍 및 점수함수 부스트랩 방법을 가상 및 실제 자료들에 적용한 결과 회귀계수들에 대한 추정치와 표준오차가 점근값들과 차이를 보이는 것으로 나타났다. 따라서 표본수가 크지 않은 경우 부스트랩 방법을 통하여 GEE모형에서의 회귀계수에 대한 추정치화 표준편차를 구하는 것이 효과적임을 알 수 있다.

주요용어: 회귀모형, 일반화추정방정식, 부스트랩.

1. 서론

Liang과 Zeger (1986)는 선형모형을 확장하여 하나의 관측치에 대하여 반복측정되거나 관측치들의 클러스터에 대하여 측정된 자료이면서 반응변수에 대한 정규성 가정이 적절하지 않은 자료에 적용할 수 있는 일반화추정방정식(Generalized Estimating Equation; GEE)을 제안하였다. GEE는 하나의 관측치에서 여러 가지의 다른 실험 조건 혹은 다른 시간에 관측된 종속변수들 간의 결합확률분포에 대한 아무런 가정 없이 단지 각각의 주변확률분포에 대한 가정만을 가지고 모형에 대한 추정이 가능하다는 장점이 있다. 또한 추정에 있어 Wedderburn (1974)의 다변량 유사우도함수를 이용하면 반복측정된 반응변수들 사이의 상관구조가 정확하게 정의되지 않아도 회귀계수의 추정치는 일치성을 만족하며 점근적으로 정규분포를 따르게 된다.

Efron (1979)이 제안한 부스트랩(bootstrap) 방법은 모집단에 대한 속성을 알지 못하여도 주어진 표본에 대한 복원 재표집을 통해 추정치의 분산 등에 관한 추정과 이에 따른 추론을 가능하게 해주는 방법으로, 선형회귀모형에 대한 부스트랩은 Efron (1979)의 잔차 부스트랩에서 시작하여 Freedman (1981), Freedman과 Peters (1984), Wu (1986), Hu와 Zidek (1995) 등 많은 학자들에 의해 다양한 방법들이 제시되었다. 이러한 방법들은 크게 잔차, 점수함수(score function), 그리고 주어진 자료 자체를 재표집하는 쌍 부스트랩 방법 등 세 가지로 나눌 수 있는데 잔차 부스트랩과 자료자체를 재표집하는 쌍 부스트랩의 경우 등분상성질이 만족되지 않는 경우에는 편의가 발생한다. 반면에 Hu와 Zidek (1995)의 점수함수를 이용하는 방법과 Wu (1986)의 방법은 불편성과 일치성을 갖는 것으로 알려져 있다. 특히 Wu의 방법은 잔차에 대하여 다항분포를 가정하고 여기서 추출된 무작위 복원표본이 부스트랩 표본이 되는데 이 방법은 독립이 아닌 점수함수에도 그대로 적용(Lele, 1991)이 가능하다.

Moulton과 Zeger (1991)는 일반화선형모형(Generalized Linear Model)에 대하여 표준화된 피어슨 잔차에 대한 일단계 근사를 이용하는 잔차 부스트랩과 쌍부스트랩을 제안하였으며 Friedl과 Stadlober

[1](110-745) 교신저자: 서울시 종로구 명륜동3가 53, 성균관대학교 통계학과, 교수. E-mail: cspark@skku.edu

(1997), Hu와 Kalbfleisch (2000), Chatterjee와 Bose (2005) 등도 일반화선형모형과 비선형모형에 대한 부스트랩 방법을 연구하였다. 또한 Moulton과 Zeger (1989)는 일반화선형모형에서 반복 측정된 자료에 대한 쌍 부스트랩 방법을 적용할 수 있음을 보였는데 이는 GEE모형의 특별한 형태이므로 이 방법을 GEE모형에 대한 쌍 부스트랩 방법으로 생각할 수 있다.

선형모형에 대한 대표적인 세 가지 부스트랩 방법들은 선형모형의 확장인 일반화선형모형에 적용이 가능하며 비슷한 원리로 일반화선형모형의 확장인 GEE모형에도 적용할 수 있다. 본 논문에서는 잔차 및 쌍 부스트랩 방법과 점수함수 부스트랩 방법을 GEE모형을 따르는 시뮬레이션 자료와 실제 자료에 적용하고 그 결과를 정리 하였다.

논문의 구성은 다음과 같다. 제 2장에서는 일반화선형모형에 대한 잔차, 쌍 및 점수함수 부스트랩 방법을 소개하고 GEE모형에 대한 이들의 적용은 제 3장에서 살펴보기로 한다. 실제 및 모의자료에 대한 세 방법의 결과들과 이들에 대한 비교는 제 4장에서 그리고 마지막으로 결론과 향후 연구방향은 제 5장에 포함하였다.

2. 일반화선형모형에 대한 부스트랩

GEE모형은 일반화선형모형의 확장이고 회귀모형에 대한 부스트랩의 적용방법은 선형회귀모형과 일반화선형모형에 대해 원리적으로 동일한 방법이 사용되므로 본 절에서는 일반화선형모형에 대한 부스트랩 방법들을 살펴보기로 한다.

일반화선형모형은 반응변수가 지수족(exponential family)으로 알려진 분포를 따르는 것으로 가정하며 반응변수가 정규분포를 따른는 경우인 선형회귀모형(linear regression model), 이항분포를 따르는 경우인 로지스틱 회귀모형(logistic regression model), 그리고 포아송분포를 따르는 경우인 포아송 회귀모형(Poisson regression model)을 포함한다. 여기서 대표 관측치를 y로 하는 I개의 관측치 벡터를 $\boldsymbol{y}$라 하고 이에 따르는 k개의 설명변수 벡터를 $\boldsymbol{x}$라 하자. 지수족은 반응변수 y의 밀도함수가

$$f(y;\theta,\phi\,|\,\boldsymbol{x}) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\phi)+c(y,\phi)}\right\} \tag{2.1}$$

의 형태를 가지고 $a(\cdot)$, $b(\cdot)$와 $c(\cdot)$는 알려진 함수들이며 모수 $\theta = \theta(\boldsymbol{x})$는 정준모수라 한다. y의 평균과 분산은 각각 $E(y\,|\,\boldsymbol{x}) = \mu(\boldsymbol{x}) = b^{'}(\theta)$와 $\mathrm{Var}(y\,|\,\boldsymbol{x}) = b^{''}(\theta)a(\phi)$가 되며 설명변수들의 선형결합 $\eta = \sum_1^k x_j\beta_j = \boldsymbol{\beta}^T\boldsymbol{x}$와 평균은 연결함수 g로 다음과 같이 결합되어 있다.

$$\eta(\boldsymbol{x}) = g(\mu(\boldsymbol{x})) = g\left(b^{'}(\theta)\right).$$

$\boldsymbol{\beta}$에 대한 추정치 $\hat{\boldsymbol{\beta}}$는 반복재가중최소제곱법을 이용하여 구할 수 있으며 $t+1$번째의 해는

$$\hat{\boldsymbol{\beta}^{\boldsymbol{t+1}}} = \hat{\boldsymbol{\beta}^{\boldsymbol{t}}} + \left(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}})$$

로 나타낼 수 있다. 여기서, $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{V}}$와 $\hat{\boldsymbol{\mu}}$는 $\hat{\boldsymbol{\beta}^{\boldsymbol{t}}}$의 값으로 계산되었으며 $I\times k$인 행렬 $\boldsymbol{D}$의 원소 D_{ir}는 $D_{ir} = \partial\mu_i/\partial\beta_r$이다. $\boldsymbol{V}$는 대각행렬로 i번째 원소 V_i는 i 번째 관측치의 조건부 분산인 $E(y_i\,|\,\boldsymbol{x}_i)$이다.

우리의 주된 관심은 부스트랩 방법을 이용하여 회귀계수 추정치의 분산 또는 표준오차을 추정하는 것이며 모든 내용은 일차 근사(first order approximation)에 기준을 두고 전개되었다. 따라서 앞에서 언급한 점수함수의 경우 일차 근사식을 사용하였으며 회귀계수의 추정치도 일단계 근사값을 고려하였다.

일반화선형모형에 대한 부스트랩은 Simonoeff와 Tsai (1988), Moulton과 Zeger (1989, 1991) 등에 의하여 연구되었으며 앞에서 언급했던 선형회귀모형에서의 방법과 유사하다. 일반화선형모형에 대한 잔

차 부스트랩은 선형모형의 경우와 달리 각 관측값의 분산이 동일하지 않으므로 이를 표준화 하는 방법에 따라 구분할 수 있다. 본 논문에서는 각 관측치의 표준편차로 표준화한 피어슨의 잔차를 이용하였다. 쌍 부스트랩은 말 그대로 관측치와 설명변수들에 대한 재표본을 이용하는 방법이다. 일반화선형모형의 경우에도 선형모형의 경우와 같이 반응변수와 설명변수들의 벡터에 대한 재표본을 추출하게 된다. 일반화선형모형에 대한 점수함수 부스트랩은 Friedl과 Stadlober (1997)에 의하여 의사가능도함수에 대한 방법으로 여러 방법들에 대한 비교와 함께 소개되었다.

일반화선형모형에 잔차, 쌍 및 점수함수 부스트랩을 적용한 결과는 회귀계수의 부스트랩 추정치에 대한 일차 근사식의 형식으로 다음과 같이 표현할 수 있다.

- 잔차 부스트랩:

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \left(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-\frac{1}{2}}\boldsymbol{T}^*\left(\boldsymbol{y}-\hat{\boldsymbol{\mu}}\right)$$

  여기서 $\boldsymbol{T}^* = \mathrm{diag}(d_i)$이고 $\boldsymbol{d} = (d_1, \ldots, d_I)$는 다항분포$(I; 1/I, \ldots, 1/I)$를 따르는 확률변수이다.

- 쌍 부스트랩:

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \left(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{T}^*\hat{\boldsymbol{V}}^{-\frac{1}{2}}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-\frac{1}{2}}\boldsymbol{T}^*\hat{\boldsymbol{V}}^{-\frac{1}{2}}\left(\boldsymbol{y}-\hat{\boldsymbol{\mu}}\right)$$

  여기서 $\boldsymbol{T}^*$는 잔차 부스트랩의 경우와 같다.

- 점수함수 부스트랩:

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \left(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{T}^{**}\hat{\boldsymbol{V}}^{-\frac{1}{2}}\left(\boldsymbol{y}-\hat{\boldsymbol{\mu}}\right)$$

  여기서 $\boldsymbol{T}^{**} = \mathrm{diag}(d_i)$이고 $\boldsymbol{d} = (d_1, \ldots, d_I)$는 각각이 표준정규분포이고 서로 독립인 $\mathrm{MVN}(\boldsymbol{0}, I_I)$를 따르는 확률변수이다.

모든 방법에서 $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{V}}$와 $\hat{\boldsymbol{\mu}}$는 모든 관측치를 이용한 MLE 추정치 $\hat{\boldsymbol{\beta}}$의 값으로 계산되었다. 종합하면 잔차 부스트랩의 경우에는 피어슨 잔차에 대한 재표본을 사용하고 점수함수의 경우에는 $(\boldsymbol{y}-\hat{\boldsymbol{\mu}})$에 대한 표본을, 그리고 마지막으로 쌍 부스트랩의 경우에는 원자료에 대한 재표본을 이용한다. 여기서 점수함수의 경우에는 일반적인 다항분포가 아닌 Wu (1986)의 방법을 일반화선형모형에 확장한 Lele (1991)가 제안한 방법을 예로 들었다.

3. GEE모형에 대한 부스트랩

동일한 실험단위나 집락에서 하나 이상의 관측치가 추출되거나 동일 개체에 대하여 반복측정된 자료들의 경우에는 그들 사이의 연관성을 고려할 필요가 있다. 반응변수가 정규분포를 따르거나 이에 근사한 경우에는 다양한 종류의 많은 모형들이 존재하였으나 그렇지 않은 경우에 적용할 수 있는 모형으로 1986년에 Liang과 Zeger는 이산형이거나 연속형인 다변량 자료에 대하여 준가능도(quasi-likelihood) 접근법을 이용하는 GEE방법을 제시하였다.

Liang과 Zeger (1986)에 따라 y_{ij}는 i 번째 대상의 j번째 관측치이고 반응변수 벡터 y_i, $i = 1, \ldots, I$의 크기는 $n_i \times 1$라고 가정하자. $\boldsymbol{X}_i$는 $n_i \times k$인 i번째 대상에 대한 설명변수행렬이며 j번째 행은 $\boldsymbol{x}_{ij}^T$으로 두자. 첫번째로 설명변수 $\boldsymbol{X}_i$가 주어졌을 때 반응변수 y_{ij}의 조건부 평균 $\mu_{ij} = E(y_{ij}\,|\,\boldsymbol{x}_{ij})$가 설명변수들의 선형결합 $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$와 일반화선형모형에서와 비슷하게

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} \tag{3.1}$$

와 같이 연결되었다고 가정한다. 여기서, $\boldsymbol{\beta}$는 $k \times 1$는 미지의 회귀모수 벡터이며 g는 미지의 연결함수이다. 두 번째로 2차 적률인 공분산 구조는 i번째 대상에 대한 공분산 행렬 $V_i(\mu)$를

$$\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\mu}) = \frac{\boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\rho}) \boldsymbol{A}_i^{\frac{1}{2}}}{\phi} \tag{3.2}$$

로 두며 $\boldsymbol{A}_i = \boldsymbol{A}(\mu_i)$는 $\mathrm{Var}(y_{ij} \,|\, \boldsymbol{x}_i)$를 j번째 대각원소로 갖는 $n_i \times n_i$인 대각행렬이다. 그리고 $\boldsymbol{R}_i(\rho)$는 $n_i \times n_i$인 상관계수 행렬로 $s \times 1$인 미지의 모수 ρ만의 함수이다. Liang과 Zeger는 정확하게 정의되지 않아도 된다는 의미에서 $\boldsymbol{R}_i(\rho)$를 작업행렬(working correlation) 또는 가상관행렬이라고 불렀다.

이제 대상들 간의 상관이 존재하지 않는다고 가정하면 첫 번째의 두 적률에 대한 정의로부터 미지의 회귀 모수 β와 장애모수 ρ 및 ϕ를 추정하는데 다변량 준가능도함수 (McCullagh, 1983)를 이용할 수 있다. McCullagh와 Nelder (1989)에 의하면 점수벡터 $\boldsymbol{U}_i(\boldsymbol{\beta}) = \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)/\phi$는 선적분으로 정의되는 준가능도함수의 경사도 벡터가 되며 따라서 다변량 준가능도 점수함수는

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{I} \boldsymbol{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{I} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0 \tag{3.3}$$

이 된다. 이 때 $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$이고 $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$이며 $\boldsymbol{V}_i$는 (3.2)에 있다.

일변량의 경우 Wedderburn (1974)은 다중적분 $Q(X\beta, y)$가 β에 대한 일반적인 로그-가능도함수와 비슷한 성질들을 가지고 있음을 증명하였으며 이는 쉽게 다변량의 경우에도 확장될 수 있다. 결과적으로, 점수함수를 통하여 얻어진 β에 대한 추정치는 최대가능도추정치와 비슷한 성질을 갖게 된다.

이제 회귀계수의 추정을 위하여 $\boldsymbol{S}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$로 두면 다음과 같은 일반화추정방정식을 얻게 된다.

$$\sum_{i=1}^{I} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = \boldsymbol{0} \tag{3.4}$$

$\boldsymbol{\beta}$에 대한 추정치 $\hat{\boldsymbol{\beta}}$는 일반화선형모형에서와 비슷하게 반복재가중최소제곱법을 이용하여 구할 수 있으며 $t+1$번째의 해는

$$\hat{\boldsymbol{\beta}^{\boldsymbol{t+1}}} = \hat{\boldsymbol{\beta}^{\boldsymbol{t}}} - \left(\sum_{i=1}^{I} \hat{\boldsymbol{D}}_i^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i \right)^{-1} \sum_{i=1}^{I} \hat{\boldsymbol{D}}_i^T \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i) \tag{3.5}$$

로 나타낼 수 있으며 $\hat{\boldsymbol{D}}_i$, $\hat{\boldsymbol{V}}_i$와 $\hat{\boldsymbol{\mu}}_i$는 $\hat{\boldsymbol{\beta}^{\boldsymbol{t}}}$의 값으로 계산되었다.

위의 추정치는 $\boldsymbol{D} = (\boldsymbol{D}_1^T, \ldots, \boldsymbol{D}_I^T)^T$, $\boldsymbol{S} = (\boldsymbol{S}_1^T, \ldots, \boldsymbol{S}_I^T)^T$라 두고 $\boldsymbol{V}$를 i번째 대각원소가 $\boldsymbol{V}_i$인 크기 $nI \times nI$인 블록대각행렬이라 하면 (3.5)식은 $\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{S}$를 반응변수로 하고 설명변수들을 $\boldsymbol{D}$ 그리고 $\boldsymbol{V}^{-1}$를 가중치로하는 반복재가중최소제곱법과 동일하다.

GEE모형은 반응변수 벡터의 원소들 사이에 가정된 상관행렬구조를 갖는 것을 제외하면 반응변수 벡터의 각 원소들의 주변분포는 일반화선형모형의 가정과 동일하다. 따라서 일반화선형모형모형에 적용이 가능한 부스트랩 방법들을 그대로 GEE모형에 적용할 수 있다. 다만, 일반화선형모형에 대한 부스트랩 방법을 GEE에 적용하려면 반응변수 벡터를 추출단위로하는 복원추출을 통한 재표본을 얻으면 된다. 재표본을 얻는 방법은 앞의 일반화선형모형에 대한 방법들을 모두 적용할 수 있다.

표 **4.1.** 독립인 상관관계를 가정한 로지스틱 자료

추정방법	변수	추정치	표준오차	신뢰구간(95%)		t-값
				하한	상한	
GEE	절편	1.213	0.241	0.741	1.685	5.043
	X_1	−0.192	0.416	−1.007	0.623	−0.462
	X_2	0.328	0.127	0.079	0.577	2.577
	X_3	0.148	0.043	0.064	0.232	3.418
잔차	절편	1.213	0.296	0.633	1.793	4.098
	X_1	−0.192	0.389	−0.954	0.571	−0.494
	X_2	0.329	0.134	0.066	0.592	2.455
	X_3	0.148	0.046	0.058	0.238	3.217
쌍	절편	1.203	0.296	0.623	1.783	4.064
	X_1	−0.179	0.390	−0.943	0.585	−0.459
	X_2	0.328	0.135	0.063	0.593	2.430
	X_3	0.148	0.046	0.058	0.238	3.217
점수함수	절편	1.212	0.295	0.634	1.790	4.109
	X_1	−0.191	0.389	−0.953	0.571	−0.491
	X_2	0.328	0.134	0.065	0.591	2.448
	X_3	0.148	0.046	0.058	0.238	3.217

4. 실증분석

이 장에서는 GEE모형에 부스트랩 방법들을 다양한 자료들에 적용한 결과가 어떻게 나타나는지를 보기 위하여 모의실험자료와 실제 자료에 적용해 보고 그 결과를 비교하였다. 예제 4.1에는 모의실험자료에 대한 결과를 그리고 예제 4.2에는 실제자료에 대한 적용 결과를 포함하였다.

예제 4.1: 모의실험에 사용한 자료는 Hardin과 Hilbe (2002, Section 5.2.4)가 사용한 모의실험자료로 종속변수 Y와 세 개의 독립변수 X_1, X_2, X_3를 포함하고 있다. Y는 0, 1로 된 이항변수이며, X_1은 $U(0,1)$, X_2는 $N(0,1)$을 그리고 X_3는 $U(5,10)$을 따른다. 반복측정으로 인한 내부상관관계가 약 0.4이며 반복측정수는 8이고 관측수가 50이므로 총 표본크기는 400이 되며 교환가능상관을 갖는 이항-로짓(binomial-logit) 모형을 따른다. 각각의 설명변수에 대한 회귀계수는 −0.4, 0.25, 0.15이며 절편은 1.3으로 두었다.

독립(independence), 교환가능(exchangeable), 1차-자기상관(AR-1) 가상관 행렬을 가정하고 GEE모형을 적합한 결과와 부스트랩반복수를 1000으로 하여 쌍 및 잔차 부스트랩을 적용한 결과가 다음의 표 4.1, 4.2, 4.3과 같다. 이 결과를 바탕으로 계수의 추정과 표준오차의 추정에 있어서 부스트랩을 적용한 방법과 GEE방법과의 차이를 계수추정치 및 반복추정치들의 평균과 표준오차 그리고 계수들의 유의성 검정을 위한 t-값을 통하여 살펴보고 추정된 회귀계수에 대한 95% 신뢰구간의 상한/하한을 비교하였다. 부스트랩 방법들에 대한 계수추정치는 반복 재표본에 대한 계수추정치들의 평균값이다.

표 4.1은 반복측정으로 인한 상관관계가 없음을 가정하여 가상관행렬을 단위행렬로 하여 GEE모형을 적합시킨 것이다. 실제 모형에 사용된 상관행렬은 교환가능한 가상관행렬이나 독립을 사용하여 추정한 경우 쌍부스트랩방법에서 얻어진 X_1의 추정치들이 다른 두 방법과 많은 차이를 보였으며 참의 계수값(−0.4)과도 차이가 나는 것을 알 수 있다. 표준오차의 경우에는 모든 방법에 있어서 차이는 크지 않았으나 GEE방법에서 얻어진 점근추정치보다 모든 부스트랩 방법들에서 작게 나타나 점근추정치가 참값을 과대 추정하는 것으로 판단된다.

표 **4.2.** 교환가능인 상관관계를 가정한 로지스틱 자료

추정방법	변수	추정치	표준오차	신뢰구간(95%)		t-값
				하한	상한	
GEE	절편	1.277	0.285	0.718	1.836	4.481
	X_1	−0.332	0.355	−1.018	0.374	−0.907
	X_2	0.252	0.109	−0.466	−0.038	−2.312
	X_3	0.142	0.038	0.068	0.217	3.737
잔차	절편	1.277	0.269	0.750	1.804	4.747
	X_1	−0.332	0.335	−0.989	0.325	−0.991
	X_2	0.252	0.104	0.048	0.456	2.423
	X_3	0.142	0.037	0.070	0.215	3.838
쌍	절편	1.271	0.270	0.742	1.800	4.707
	X_1	−0.324	0.337	−0.985	0.337	−0.961
	X_2	0.252	0.105	0.046	0.458	2.400
	X_3	0.143	0.037	0.071	0.216	3.865
점수함수	절편	1.276	0.268	0.751	1.801	4.761
	X_1	−0.331	0.336	−0.990	0.328	−0.985
	X_2	0.251	0.104	0.047	0.455	2.414
	X_3	0.142	0.037	0.070	0.215	3.838

표 **4.3.** AR-1인 상관관계를 가정한 로지스틱 자료

추정방법	변수	추정치	표준오차	신뢰구간(95%)		t-값
				하한	상한	
GEE	절편	1.333	0.245	0.8530	1.8130	5.4410
	X_1	−0.488	0.372	−1.2170	0.2410	−1.3120
	X_2	0.221	0.113	−0.0010	0.4430	1.9560
	X_3	0.130	0.040	0.0520	0.2080	3.2500
잔차	절편	1.333	0.278	0.6785	0.6742	0.6705
	X_1	−0.487	0.348	−1.1690	0.1950	−1.3990
	X_2	0.221	0.121	−0.0160	0.4580	1.8260
	X_3	0.130	0.044	0.0044	0.2160	2.9550
쌍	절편	1.328	0.279	0.7810	1.8750	4.7600
	X_1	−0.482	0.349	−1.1660	0.2020	−1.3810
	X_2	0.220	0.122	−0.0190	0.4590	1.8030
	X_3	0.130	0.044	0.0440	0.2160	2.9550
점수함수	절편	1.332	0.277	0.7890	1.8750	4.8090
	X_1	−0.487	0.348	−1.1690	0.1950	−1.3990
	X_2	0.220	0.121	−0.0170	0.4570	1.8180
	X_3	0.130	0.043	0.0460	0.2140	3.0230

모형에 사용된 참의 상관행렬과 추정에 사용된 가상관행렬이 같은 경우의 결과가 표 4.2에 포함되어 있다. 이 경우에는 가상관행렬이 참과 다른 두 경우와 다르게 쌍부스트랩을 통하여 얻어진 계수추정치들과 다른 방법들과 거의 차이가 없음을 알 수 있다. 표준오차의 경우 또한 세 방법에 특별한 차이점은 보이지 않으나 점근 추정치과 부스트랩 방법에서 얻어진 추정치에 차이가 거의 없어 이 경우 점근추정치의 과대성이 많이 줄어드는 것으로 보인다.

표 4.3은 반복측정으로 인한 상관관계가 AR-1의 구조를 갖는다는 것을 가정하여 세 방법을 적합시킨 결

표 **4.4.** 천식자료에 대한 적합결과

가상관행렬	변수	GEE		잔차		쌍		점수함수	
		회귀계수	표준오차	회귀계수	표준오차	회귀계수	표준오차	회귀계수	표준오차
독립	절편	1.260	2.695	1.251	3.061	1.298	3.129	1.274	3.072
	Kingston	0.139	0.571	0.143	0.686	0.120	0.733	0.138	0.683
	Age	−0.200	0.259	−0.199	0.282	−0.200	0.287	−0.202	0.283
	Smoke	−0.128	0.424	−0.129	0.393	−0.128	0.437	−0.130	0.393
교환가능	절편	1.275	2.474	1.266	3.052	1.332	3.121	1.290	3.064
	Kingston	0.122	0.696	0.126	0.688	0.102	0.733	0.121	0.686
	Age	−0.204	0.238	−0.203	0.278	−0.205	0.284	−0.205	0.279
	Smoke	−0.094	0.451	−0.095	0.363	−0.089	0.403	−0.096	0.362
AR-1	절편	1.187	2.733	1.178	3.005	1.204	3.058	1.202	3.016
	Kingston	0.261	0.754	0.265	0,652	0.242	0.697	0.261	0.650
	Age	−0.213	0.263	−0.212	0.273	−0.212	0.277	−0.215	0.274
	Smoke	0.077	0.452	0.076	0.374	0.089	0.406	0.076	0.373

과이다. 이 경우에도 X_1의 계수추정치가 다른 두 방법과 GEE 추정치와 차이를 보였다. 표준오차 추정치들은 독립인 경우와 비슷한 양상을 나타내어 GEE 추정치보다 작았다.

실증분석 결과를 정리해 보면 GEE모형의 적합결과는 가상관행렬에 따라 조금씩 차이를 보이는 것으로 나타났다. 이론적으로 가상관행렬을 다르게 하여도 추정량은 일치성을 만족하는 것으로 알려져 있으나 예제에서는 차이를 보여 표본수가 작은 경우 주의할 필요가 있음을 알 수 있다. 또한 참의 상관과 가상관이 같은 경우 세 방법의 결과에는 큰 차이가 없었으나 그렇지 않은 경우 쌍 부스트랩 방법에 의한 추정치가 다른 방법들과 차이를 보였다. 전체적으로 보면 가정된 가상관 행렬에 따라 계수추정치들에는 차이가 있으나 부스트랩 방법들 간에 많은 차이는 나타나지 않았다. 추정오차의 경우 모든 방법에서 GEE 추정치와 차이가 있었으며 이 예제에서는 대부분 GEE 추정치보다 작은 값을 나타내었다.

예제 4.2: 본 예제에서는 실제 자료를 통해 각 방법을 비교하여 보기로 한다. 분석에 사용된 자료는 Ware 등 (1984)이 사용한 자료의 일부분으로 대기오염이 건강에 미치는 영향에 대한 연구이다. 16명의 어린아이를 대상으로 조사되었으며 9세부터 12세까지 4년동안 반복측정되었다. 종속변수는 어린이의 천식여부를 나타내는 이항변수인 wheeze이다. 독립변수로는 Kingston 지방에 거주하는지의 여부를 나타내는 이항변수인 kingston, 어린이의 연령을 나타내는 age, 어린이의 부모가 흡연자인지의 여부를 나타내는 이항변수인 smoke로 구성되어 있다. 이 자료에 대해 앞의 예제에서와 마찬가지로 독립, 교환가능, 1차-자기상관의 세 가지 경우의 가상관행렬을 사용하여 적합하였으며 그 결과가 표 4.4에 있다.

우선 계수 추정치를 살펴보면 잔차 및 점수함수 부스트랩과 GEE 추정치의 경우에는 큰 차이가 없었으나 쌍 부스트랩 방법을 통하여 얻어진 추정치는 나머지 방법들에 보다 대부분의 경우 작은 값을 나타내었다. 표준오차의 경우에는 모든 부스트랩 방법의 결과가 GEE 추정치에 비하여 자료에 따라 일관성있게 크거나 또는 작은 값을 나타내었다. 특히 쌍 부스트랩의 경우에는 나머지 두 부스트랩 방법보다 상대적으로 큰 값을 보였다. 이는 등분산성을 만족하지 않는 경우 쌍 부스트랩 방법에 편의가 있기 때문으로 판단된다.

5. 결론

본 논문에서는 부스트랩 방법들 중 회귀모형에 자주 사용되는 잔차, 쌍 및 점수함수 부스트랩 방법을

GEE모형을 따르는 가상자료와 반응변수가 반복측정된 실제자료에 적용하고 그 결과를 살펴보았다.

GEE모형에서 추정에 사용된 가상관행렬이 참의 상관행렬과 다르더라도 일치성을 만족하는 것으로 알려져 있으나 가상자료에 대한 적용결과 참의 값과 차이가 있는 것으로 나타났다. 다만, 가상관행렬이 참의 상관행렬과 같은 경우에는 계수추정치들 또한 참의 값에 가까운 값을 보임을 알 수 있었다.

부스트랩 방법들을 적용하여 구해진 계수추정치들은 부스트랩 방법에 따르는 차이는 크지 않았으나 가상관행렬이 참과 다른 경우 쌍 부스트랩 방법으로 추정된 값들은 다른 방법들 및 GEE 추정치와 상이한 결과를 나타내어 GEE 모형에서 쌍 부스트랩을 사용할 경우에 주의가 필요한 것으로 판단된다. 표준오차의 경우에는 부스트랩 방법들로 구해진 추정치들 간에는 큰 차이가 없었으나 부스트랩 방법들과 GEE 추정치 간에는 상당한 차이를 나타내고 있어 표준오차의 경우에는 표본이 작은 경우 GEE 추정치가 과대 또는 과소평가되고 있음을 알 수 있다. 표준오차의 경우에도 쌍 부스트랩 방법의 추정치가 다른 두 방법의 추정치보다 크거나 작은 값을 나타내었다.

결론적으로 GEE모형은 일반화선형모형의 확장이므로 선형회귀모형 또는 일반화선형모형에 적용할 수 있는 부스트랩 방법들을 적용하는데 큰 문제는 없었으나 부스트랩 표본에 대한 추정치를 구하는 과정에서 적절히 수렴하지 않는 경우가 종종 나타났다. 하지만 등분산이 아닌 경우 편의가 있는 것으로 알려져 있는 쌍 부스트랩 방법을 제외하고 잔차 및 점수함수 부스트랩의 경우에는 GEE 모형에 적용할 경우 다른 회귀모형에서와 같이 효과적으로 계수의 변동에 대한 추정치를 얻을 수 있을 것으로 보인다. 다만, 본 논문에서 언급한 결과들은 하나의 가상자료와 하나의 실제자료에 대한 적용 결과를 바탕으로 한 것이므로 이를 일반화하기 위해서는 더욱 완벽한 모의실험이 필요하다고 하겠다.

참고문헌

Chatterjee, S. and Bose, A. (2005). Generalized Bootstrapping for estimating equation, *The Annals of Statistics*, **33**, 414–436.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

Freedman, D. (1981). Bootstraping regression model, *The Annals of Statistics*, **9**, 1218–1228.

Freedman, D. and Peters, S. (1984). Bootstraping a regression equation, *Journal of the American Statistical Association*, **79**, 97–106.

Friedl, H. and Stadlober, E. (1997). Resampling methods in generalized linear models useful in environmetrics, *Environmetrics*, **8**, 441–457.

Hardin, J. W. and Hilbe, J. M. (2002). *Generalized Estimating Equations*, Chapman & Hall, New York.

Hu, F. and Kalbfleisch, J. (2000). The estimating function bootstrap (with discussion), *The Canadian Journal of Statistics*, **28**, 449–499.

Hu, F. and Zidek, J. (1995). A bootstrap based on the estimating equations of the linear model, *Biometrika*, **82**, 263–275.

Lele, S. R. (1991). Resampling using estimating equation, In *Estimating Functions* (V.P. Godambe, ed.), 295–304, Oxford University Press.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

McCullagh, P. (1983). Quasi-likelihood function, *The Annals of Statistics*, **11**, 59–67.

McCullagh, P. and Nelder (1989). *Generalized Linear Models* 2^{nd} edition, Chapman & Hall, New York.

Moulton, L. and Zeger, S. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, **45**, 381–394.

Moulton, L. and Zeger, S. (1991). Bootstrapping generalized linear models, *Computational Statistics & Data Analysis*, **11,** 53–63.

Simonoff, J. S. and Tsai, C. L. (1988). Jackknifing & bootstrapping quasi-likelihood estimators, *Journal of Statistical Computation and Simulation*, **30**, 213–232.

Ware, J., Dockery, D., Spiro, A., Speizer, F. and Ferris, B. (1984). Passive smoking, gas cooking and respiratory health of children living in six cities, *Am. Rev. Respir. Dis.*, **129**, 366–374.

Wedderburn, R. (1974). Quasi-likelihood function, generalized linear models, and Gauss-Newton method. *Biometrika*, **61**, 439–447.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *The Annals of Statistics*, **14**, 1261–1295.

Bootstrap Estimation for GEE Models

Chongsun Park[1] · Yong Moon Jeon[2]

[1]Department of Statistics, Sungkyunkwan University
[2]Department of Statistics, Sungkyunkwan University

Abstract

Bootstrap is a resampling technique to find an estimate of parameters or to evaluate the estimate. This technique has been used in estimating parameters in linear model(LM) and generalized linear model(GLM). In this paper, we explore the possibility of applying Bootstrapping Residuals, Pairs, and an Estimating Equation that are most widely used in LM and GLM to the generalized estimating equation(GEE) algorithm for modelling repeatedly measured regression data sets. We compared three bootstrapping methods with coefficient and standard error estimates of GEE models from one simulated and one real data set. Overall, the estimates obtained from bootstrap methods are quite comparable, except that estimates from bootstrapping pairs are somewhat different from others. We conjecture that the strange behavior of estimates from bootstrapping pairs comes from the inconsistency of those estimates. However, we need a more thorough simulation study to generalize it since those results are coming from only two small data sets.

Keywords: Regression model, generalized estimating equation, bootstrap method.

[1]Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myungnyun-Dong, Jongno-Gu, Seoul 110-745, Korea. E-mail: cspark@skku.edu