Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.2 pp. 143~152

Comparative Study on Imputation Procedures in Exponential Regression Model with missing values

Young-sool Park[1] · Soon-kwi Kim[2]

Abstract

A data set having missing observations is often completed by using imputed values. In this paper, performances and accuracy of five imputation procedures are evaluated when missing values exist only on the response variable in the exponential regression model. Our simulation results show that adjusted exponential regression imputation procedure can be well used to compensate for missing data, in particular, compared to other imputation procedures. An illustrative example using real data is provided.

Keywords : MCAR, Imputation procedure, Hot deck, Adjusted exponential regression imputation, Exponential regression model

1. 서론

불완전한 자료(incomplete data)에서 결측값(missing values)의 문제는 많은 데이터 세트에서 빈번히 발생하며, 특히 임상시험(clinical trials) 자료를 다루는 의학, 보건분야 뿐 아니라 사회과학 분야의 실제 상황에서 공통적으로 나타나고 있는 실정이다. 불완전한 자료는 단위누락(missing units)과 항목누락(missing items)이라는 두 가지 형태로 구성된다. 단위누락은 표본단위에서 무반응(무응답)의 결과로 설문 응답을 거절(refusals)하거나 접근하기 어려운(inaccessible) 개체들의 자료이다. 이와 같은 무응답의 형태를 단위 무응답(unit nonresponse)이라고도 한다. 반면에 무응답은 아니지만 몇 가지 질문에 답하지 않은 개체들도 있다. 이런 종류의 무응답의 형태를 항목누락이라고 한다. 결측값이 있는 자료를 분석할 경우, 일반적으로 결측값이 있는 개체들을

1) Associate professor, Dept. of Information Statistics, Kwandong University, Kangnung, Kangwondo 210-701, Korea,
E-mail : yspark@kwandong.ac.kr
2) Professor, Dept. of Information Statistics, Kangnung National University, Kangnung, Kangwondo 210-701, Korea

제거하든지 또는 다른 정보를 이용하여 결측값을 대치하는 방법이 관례이다. 결측값을 대치하기 위해 어떤 정보를 사용하여 결측값을 대치하는 절차를 대치법 절차(imputation procedures)라고 한다.

대치법은 설문의 크기에 관계없이 모수를 추정하는 경우 편향(bias)을 줄여주고 특별한 경우 항목누락인 개체를 제거하는 경우와 비교하여도 정밀도에서의 손실이 적기 때문에 그 동안 널리 사용되었고 또한 폭넓게 개발되어 왔다. 일단 결측된 자료에 값이 대치되면, 완전한 자료인 것 같이 자료를 분석하게 된다.

Efron(1994)은 결측값이 있는 자료에서 추정의 정확성을 평가하기 위해 비모수적인 붓스트랩 방법(bootstrap method)을 사용하였고, 이 방법으로 구한 신뢰구간이 보다 정확한 해답을 제공하는 것으로 판명되었다. Bello(1995)는 몇 가지의 수치적인 대치법(평균대치법(the mean substitution method), EM 알고리즘(EM algorithm), 주성분법(principal component method), 일반적인 반복주성분법(general iterative principal component method), 이상값 분해법(singular value decomposition method))의 성능을 비교 조사하였다. Hegamin-Younger, Forsyth(1998)는 18869명의 자료를 이용하여 네 가지의 대치방법 절차(평균대치법(mean imputation), 조건 평균대치법(conditional mean imputation), 핫덱 대치법(hot deck), 회귀대치법(regression imputation))의 효과를 비교하였다. 이 연구결과에 의하면 반응변수를 예측하려는 경우 평균대치법은 적절하지 않고 대신 회귀대치법이 적절한 것으로 판명되었다.

i번째 개체에 대해 y_i는 변수 Y의 참값이라고 할 때, y_i가 결측값이면 $m_i = 1$, 그렇지 않으면 $m_i = 0$ 이라고 놓자. $x_1, \cdots, x_k$는 관찰된 변수의 집합이라 가정하자. 이때 변수 Y의 값들이 결측될 확률에 따라 결측자료를 세 종류로 구분하였다.

(1) $\mathrm{P}\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = \mathrm{P}\{m_i = 1\}$이면 missing completely at random (MCAR).

(2) $\mathrm{P}\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = \mathrm{P}\{m_i = 1 \mid x_{1i}, \cdots, x_{ki}\}$이면 missing at random (MAR).

(3) $\mathrm{P}\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = \mathrm{P}\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\}$ 또는
$\mathrm{P}\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = \mathrm{P}\{m_i = 1 \mid y_i\}$이면 non-ignorable(NI).
(Little and Rubin, 1987).

본 논문의 주된 목적은 단지 반응변수 Y 만이 결측되었고, 그 결측의 방법이 MCAR 일 경우, 지수회귀계수(exponential regression coefficients)의 추정에 실제로 사용된 5종류의 대치법에 대해 각각의 정확도와 그 효과를 조사하고자 함이다. 5가지의 대치법 절차는

(1) 전체 평균대치법(Grand Mean Imputation Procedure, GM)
(2) 조건 평균대치법(Conditional Mean Imputation Procedure, CM)
(3) 핫덱(Hot-Deck, HD)
(4) 지수회귀대치법(Exponential Regression Imputation Procedure, EI)
(5) 수정된 지수회귀대치법(Adjusted Exponential Regression Imputation

Procedure, AEI)

이고, 이 중에서 EI 방법과 AEI 방법은 이 논문에서 제안된 절차로 3.4절에서 정의하기로 한다. 2절에서는 지수회귀모형을 소개하였고, 3절에서는 5종류의 대치법 절차를 간단히 기술하였다. 4절에서는 모의실험(몬테칼로 실험)과 그 결과를 5, 6절에서는 데이터 세트의 적용과 결론을 각각 서술하였다.

2. 지수회귀모형

생존시간 t_i를 i번째 개체의 사망시간이라 가정하자. 지수회귀모형은 다음과 같이 비례위험모형과 가속화고장모형의 두 방식으로 표현할 수 있다. 지수분포함수에 대한 비례위험모형(proportional hazard model)의 위험함수(hazard rate)는

$$h(t) = \lambda \exp\{\beta_1' x_1 + \cdots + \beta_p' x_p\}, \quad t \geq 0 \text{ and } \lambda > 0$$

로 주어진다. 여기에서 λ는 척도모수(scale parameter)이며, $x_1, \cdots, x_p$는 공변량을 $\beta_1', \cdots, \beta_p'$는 그 회귀계수를 각각 나타낸다. 만약 $Y = \ln t$이라 할 때, Y는 표준극단값분포(standard extreme-value distribution)를 따른다. 동일한 분포함수에 대한 가속화고장모형(accelerated failure-time(AFT) model)은

$$t = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\, \varepsilon$$

와 같이 주어진다. 여기에서 $x_1, \cdots, x_p$는 공변량을 $\beta_0, \beta_1, \cdots, \beta_p$는 절편항과 그 회귀계수를 각각 나타내며, ε는 모수 1를 가지는 지수분포(exponential distribution)를 따른다. 위에서 언급한 비례위험모형과 가속화고장모형 사이에 다음과 같은 관계가 있음을 보일 수 있다.

$$\beta_i = -\beta_i' \quad \text{for} \quad i = 1, 2, \cdots, p,$$
$$\lambda = \exp\{-\beta_0\}.$$

3. 대치법 절차

본 절에서는 반응변수에 결측값이 있는 경우 1절에서 열거한 5종류의 대치법들을 간단히 살펴보고자 한다. 5종류의 대치법 중 처음 3종류는 기존의 널리 사용되는 방법이고, 나머지 2종류는 이 논문에서 제안하는 방법이다.

3.1 기존의 방법

• 전체평균 대치법

전체평균 대치법(GM)은 가장 널리 사용되고, 사용하기 쉬운 대치법이다. 이 절차는 어떤 변수에서 결측된 자료를 제거한 나머지 관측된 자료들의 평균값을 그 변수의 결측값으로 대치하는 방법이다.

• 조건평균 대치법

조건 평균대치법(CM)은 결측값을 대치하기 위해 부가적인 정보를 사용하는 것이다. 부가적인 정보를 이용하여 데이터 세트를 몇 개의 동질적인 집단으로 분할하는 방법이다. 예를 들면, 어느 학급을 순위에 관한 질문의 응답을 근거로 하여 몇 개의 그룹으로 나눌 때 어떤 한 변수의 사분위수를 이용하여 범주화할 수 있을 것이다.

• 핫덱 방법

설문조사에서 반응변수 Y의 값이 결측되지 않은 어느 한 개체의 자료에서 얻은 값을 결측값 Y를 가진 개체의 자료에 대치하는 방법이다. 핫덱(HD) 방법에서 값을 제공하는 개체(donor)를 여러 다양한 방법으로 선택할 수 있다

핫덱 방법에서 각 방(cell)들은 자료의 대치에 중요하다고 간주된 변수에 근거해서 정의된다. 이 변수들은 일반적으로 표본설계에 관련된 변수이거나, 인구통계에 관한 변수들이다. 첫 번째로 데이터 세트는 정의된 방들로 먼저 순서화되어야 하며, 두 번째로 자료의 대치에 밀접하게 관련된 변수들에 의해 순서화되어야 한다. 각 방마다 모든 변수에 결측값이 없는 개체의 값으로 레지스터(register)를 정의한다. 각 개체들에 대해 그 개체의 방들이 결정되며, 만약 변수의 어떤 값이 누락되었다면 그 방의 레지스터의 값으로 대치하게 된다. 반면에 어떤 개체의 자료 값이 완전히 주어졌다면, 그 개체의 자료 값들이 레지스터가 된다. 모든 결측값이 대치될 때까지 이 과정을 반복한다.

3.2 지수회귀대치법과 조정된 지수회귀대치법

반응변수에 결측값이 없는 개체들의 자료를 이용하여 회귀식을 적합시킨다. 이 때 적합된 회귀선은 다음과 같다.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k .$$

위의 식에서 $\hat{y}$는 결측이 있는 개체에 대해 대치하여야 할 종속변수의 추정값이고, $x_1, x_2, \cdots, x_k$는 알려진 공변량들의 값이다. 결측된 종속변수 y 대신 $\hat{y}$를 대치하는 방법을 회귀대치법(regression imputation)이라고 한다.

위에서 설명한 회귀대치법을 지수회귀모형에 적용하여 보자. 가속화고장모형을 결측값이 없는 완전한 자료만을 이용하여 적합시키면, 다음과 같은 적합된 지수회귀모형을 얻을 수 있다.

$$\hat{t} = \exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k), \tag{3.1}$$

여기에서 $\hat{t}$는 적합된 생존시간, $x_1, x_2, \cdots, x_k$는 알려진 공변량들의 값이다. 지수회귀대치법(EI)이란 결측된 생존시간 t_i 대신 적합된 생존시간인 $\hat{t}_i$를 대치하는 방법으로 제안한다. 즉, 본 논문에서 제안하고자 하는 EI 방법(AEI)이란 결측값들을 적합된 생존시간인 $\hat{t}_i$의 값으로 대치하는 방법을 의미한다.

가속화고장모형의 로그 가능도함수 $L(\beta)$는 다음과 같다.

$$L(\beta) = \sum \left(c_i z_i - e^{z_i} \right).$$

여기에서 $z_i = y_i - x_i'\beta$, $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ 그리고 $x_i = (1, x_{1i}, \cdots, x_{ki})'$이다. 절편항에 대한 스코어 방정식은

$$\sum c_i = \sum \left(\frac{t_i}{\hat{t}_i} \right) \tag{3.2}$$

로 주어지며, 생존시간의 추정값 $\hat{t}_i$는 식 (3.1)에 주어졌다. 통상적인 선형회귀모형에서는 관찰값들의 합은 예측값들의 합이 되지만, 식 (3.2)의 항 $\sum c_i$는 절단되지 않은 관찰값의 수를 나타내고 있다. 위의 식을 이용하여 지수회귀모형에서 생존시간의 또 다른 추정값으로 조정된 예측값 $\tilde{t}_i$을

$$\tilde{t}_i = \left(\sum \frac{c_i}{n} \right) \hat{t}_i$$

로 제안할 수 있게 된다. 즉, 조정된 지수회귀대치법(AEI)이란 지수회귀모형에서 결측값들이 존재하는 경우 결측값 대신 이 논문에서 제안한 조정된 예측값 $\tilde{t}_i$를 사용하는 방법을 의미한다. 예측할 수 있듯이 4절에서 두 가지 방법을 굳이 비교하여 본다면 AEI 방법이 EI 방법에 비해 다소나마 나아진 방법임을 알 수 있을 것이다. 여기에서 생존시간 t의 분포는 지수분포를 따른다.

4. 모의실험

본 절에서는 3절에서 소개된 여러 대치법 절차를 시뮬레이션을 통하여 비교 조사하고자 한다. 이 작업의 목적은 5가지의 각기 다른 대치법으로 구한 데이터 세트를 이용하여 지수회귀모형에서의 회귀계수(exponential regression coefficient)를 추정하여 그들의 변화를 탐지하여 서로 다른 여러 대치법들의 성질을 비교, 규명하려는 것이다.

자료는 지수분포 $E(\lambda)$에서 생성하였으며, 여기에서 λ는 모수를 나타낸다. 자료를 생성할 때 사용한 컴퓨터 프로그램은 IMSL의 RNEXP와 SSCAL 서브루틴이다. 따라서 생존시간(survival time) t의 모형은

$$t = E(\lambda = 1) \exp(\beta_0 + c_1\beta_1 + c_2\beta_2)$$

이다. 여기에서 모수 β_0, β_1, β_2의 값들로 1.5, 5.5, -1.5를 각각 사용하였다. 두 공변량의 모수값으로 하나는 양수 또 다른 하나는 음수를 사용하였다. 결측값은 MCAR 방법을 이용하였다. 두 공변량 중 변수 c_1은 자료의 반은 0 그리고 나머지 반은 1을 취하는 범주형변수이고, 변수 c_2는 IMSL의 서브루틴 RNUN을 사용하여 균등분포 $U(0,1)$로부터 생성된 값을 가지는 확률변수이다. 데이터 세트의 10%를 절단된(censored) 자료로 간주하여, 절단상태를 나타내는 변수 status는 0의 값을 가지도록 하였다.

각 표에 주어진 값들은 표본 크기가 n(=40, 60, 100)인 경우 각각 5000번씩 반복 실시한 시뮬레이션 결과이다. 결측된 자료의 비율 (k)은 실제상황에서 일어나는 것을 감안하여 0.05, 0.10, 0.15, 0.20를 고려하였다.

표본의 크기에 따라 결측된 자료가 있는 불완전한 지수회귀자료에서 5가지 대치법 절차를 사용하여, [표 1]에서는 3가지 모수 $(\beta_0, \beta_1, \beta_2)$의 추정값을 각각 제시하였고, [표 2]에서는 [표 1]의 3가지 모수 $(\beta_0, \beta_1, \beta_2)$에 대응하는 평균제곱오차(MSE)들을 각각 나타내었다.

5가지 대치법 절차를 사용한 시뮬레이션 결과에서 일반적으로 여러 가지 유의하여야 할 사항은 다음과 같다.
(1) 결측된 자료의 비율이 증가하면, 대치법 절차의 모수들의 추정값은 참값으로부터 멀어지고 이에 대응하는 평균제곱오차는 증가하는 경향이 있다.
(2) 데이터 세트의 결측비율에 관계없이 표본의 크기가 증가하면 모든 모수의 추정값은 참값에 근접해지며 평균제곱오차는 감소하는 경향이 있다.

5가지 대치법 절차 중에서 차이점을 요약하면 다음과 같다.
(1) 표본의 크기가 $n=40$ 또는 60인 경우, CM과 AEI 방법의 평균제곱오차가 가장 적은 값을 가지게 되고, 표본의 크기가 $n=100$이고 결측비율이 5% 또는 10%일 때 AEI와 CM 방법의 평균제곱오차가, $n=100$이고 결측비율이 15% 또는 20%일 때 AEI와 HD 방법의 평균제곱오차가 가장 적은 값을 가진다.
(2) 모수 $\beta_0, \beta_1, \beta_2$의 추정에 있어서 AEI와 HD 방법이 CM 방법에 비해 적은 편향을 가진다.
(3) GM 방법은 결측비율과 표본의 크기에 관계없이 다른 대치법과 비교하여 상당히 큰 평균제곱오차를 산출하므로 바람직한 방법이라고 볼 수 없겠다.
(4) AEI 방법이 EI 방법보다 다소나마 적은 평균제곱오차를 산출한다.

[표 1] 회귀계수에 근거한 5종류 대치법의 비교

n	k	GM	CM	HD	EI	AEI
40	5%	-3.635123 -4.873974 3.571461	-1.500239 -5.618687 1.499434	-1.531679 -5.611927 1.540333	-1.533301 -5.622315 1.540218	-1.526753 -5.620828 1.532117
	10%	-4.323903 -3.811954 3.089545	-1.475153 -5.641269 1.451774	-1.536520 -5.606614 1.540844	-1.535152 -5.618607 1.536323	-1.520455 -5.616261 1.526687
	15%	-3.864768 -2.526356 0.289728	-1.395989 -5.566090 1.149003	-1.542747 -5.604462 1.554565	-1.529388 -5.620047 1.528153	-1.521009 -5.624607 1.562556
	20%	-4.415373 -2.220041 0.775722	-1.325754 -5.584087 1.098648	-1.545991 -5.598067 1.550889	-1.551888 -5.624381 1.562415	-1.508503 -5.622331 1.561946
60	5%	-4.200583 -4.031991 3.305291	-1.540261 -5.539330 1.447723	-1.609492 -5.498826 1.556051	-1.619278 -5.494202 1.558645	-1.615212 -5.495996 1.557929
	10%	-4.519772 -3.364414 2.817064	-1.500836 -5.564879 1.368722	-1.605988 -5.496493 1.550907	-1.621527 -5.493703 1.557850	-1.613282 -5.494158 1.562955
	15%	-4.473425 -2.373850 1.141999	-1.431176 -5.545960 1.198603	-1.611064 -5.493121 1.556271	-1.633810 -5.485446 1.563964	-1.611423 -5.492766 1.575350
	20%	-4.521713 -2.119308 0.792166	-1.426651 -5.525581 1.139993	-1.610302 -5.487852 1.546610	-1.633040 -5.492458 1.573598	-1.610801 -5.495729 1.615629
100	5%	-3.334405 -3.219060 0.520675	-1.527537 -5.442117 1.314367	-1.558673 -5.454051 1.402601	-1.567348 -5.447823 1.403379	-1.564289 -5.449037 1.403888
	10%	-3.693627 -2.747298 0.353834	-1.508374 -5.433132 1.255393	-1.559703 -5.455675 1.406137	-1.567142 -5.445173 1.393682	-1.557994 -5.445399 1.396877
	15%	-3.966180 -2.332092 0.038747	-1.459855 -5.427676 1.105867	-1.560780 -5.458047 1.410002	-1.560498 -5.438385 1.360402	-1.549130 -5.442465 1.389138
	20%	-4.458676 -2.028631 0.509578	-1.392729 -5.474834 1.029996	-1.565741 -5.457628 1.420577	-1.568136 -5.434446 1.370499	-1.537009 -5.436823 1.396470

여기에서 GM은 전체 평균대치법, CM은 조건 평균대치법, HD는 핫덱, EI는 지수회귀대치법, AEI는 조정된 지수회귀대치법을 각각 의미한다. 각 방의 3×1 벡터는 β_0, β_1, β_2의 추정값을 각각 표시한다. 이 때 모수들의 참값은 $\beta_0 = -1.5$, $\beta_1 = -5.5$, $\beta_2 = 1.5$이다.

[표 2] 평균제곱오차에 근거한 5종류 대치법의 비교

n	k	GM	CM	HD	EI	AEI
40	5%	3.290644	0.204909	0.242663	0.211084	0.211013
	10%	4.657721	0.203242	0.254296	0.216319	0.215266
	15%	5.382910	0.193223	0.266210	0.244044	0.245374
	20%	6.657419	0.222159	0.271079	0.268105	0.266821
60	5%	4.352593	0.119442	0.145434	0.139835	0.139517
	10%	5.244156	0.118754	0.156914	0.146505	0.145638
	15%	6.295015	0.134957	0.167482	0.163793	0.162409
	20%	7.072670	0.145135	0.177154	0.173941	0.174901
100	5%	3.209503	0.084957	0.089958	0.085220	0.085003
	10%	4.616873	0.092629	0.093761	0.090034	0.089019
	15%	6.132301	0.117365	0.098049	0.097506	0.094475
	20%	7.301789	0.140754	0.109728	0.104334	0.101085

5. 실제 사례

5종류의 대치법의 성질을 규명하고 그 차이들을 비교하기 위하여 실제 사례로 Hosmer & Lemeshow(1999, p.4)에 있는 HMO-HIV+study의 데이터 세트를 사용하였다. 데이터 세트는 4개의 변수인 Time(days between entry date and end date), Age, Drug(history of IV drug use), 100명의 환자에 대해 측정된 절단된 변수로 구성되어 있다. 논증의 목적으로 4절의 시뮬레이션에서 사용한 것과 마찬가지로 결측된 자료의 비율 (k)은 5%, 10%, 15%, 20%를 선택하였다. 결측값으로는 반응변수(Time)의 값을 임의로 선택하였다. 결측값이 있는 불완전자료에서 5가지 대치법 절차를 적용한 시뮬레이션 결과는 [표 3]~[표 4]에 주어졌다.

[표 3] 회귀계수에 근거한 5종류 대치법의 비교

k	nonmissing	GM	CM	HD	EI	AEI
5%	-6.151630 0.092092 1.009856	-5.713845 0.077556 1.022298	-5.609066 0.073270 1.130116	-6.079307 0.089992 0.994101	-5.995275 0.086656 1.035377	-6.006202 0.087053 1.034384
10%	-6.151630 0.092092 1.009856	-5.597359 0.073935 0.911987	-5.579041 0.071069 1.111660	-5.946630 0.085387 1.011745	-5.995272 0.086041 1.015121	-6.018138 0.087083 1.016440
15%	-6.151630 0.092092 1.009856	-5.565361 0.072042 0.866754	-5.579041 0.069943 1.143497	-5.940540 0.084477 0.974614	-6.039474 0.085652 1.048344	-6.067515 0.087179 1.048025
20%	-6.151630 0.092092 1.009856	-5.516619 0.069418 0.818621	-5.514032 0.066406 1.177710	-5.753061 0.076774 1.029930	-6.014347 0.083538 1.060857	-6.050556 0.085804 1.066151

여기에서 GM은 전체 평균대치법, CM은 조건 평균대치법, HD는 핫덱, EI는 지수회귀대치법, AEI는 조정된 지수회귀대치법을 각각 의미한다. 각 방의 3×1 벡터는 β_0, β_1, β_2의 추정값을 각각 표시한다. 이 때 모수들의 참값은 $\beta_0 = -6.151630$, $\beta_1 = .092092$, $\beta_2 = 1.009856$ 이다(HMO-HIV+연구).

[표 4] 평균제곱오차에 근거한 5종류 대치법의 비교

k	GM	CM	HD	EI	AEI
5%	0.0640072	0.1030640	0.0018277	0.0083759	0.0072588
10%	0.1057082	0.1205968	0.0140225	0.0081707	0.0059632
15%	0.1215305	0.1154028	0.0152863	0.0046980	0.0028521
20%	0.1467745	0.1451218	0.0531648	0.0071736	0.0044749

AEI 방법은 결측된 자료의 비율이 10%, 15%, 20%일 경우 평균제곱오차를 기준으로 할 때 최적이었고, 결측된 자료의 비율이 5%일 때 HD 방법과 AEI 방법의 평균제곱오차가 제일 적은 것으로 나타났다. 또한 모수 $\beta_0, \beta_1, \beta_2$의 추정에 있어서는 AEI와 HD 방법이 CM 방법에 비해 적은 편향을 보여주고 있다. 실제 데이터 세트의 결과는 표본의 크기가 100일 때, [표 2]에 주어진 시뮬레이션의 결과와 거의 일치함을 볼 수 있다.

6. 결론

본 논문에서는 모의실험을 통하여 5가지 대치법 절차를 비교 분석하였다. 지수회귀모형은 표본의 크기, 모수들의 값과 결측된 자료의 비율에 영향을 받고 있음을 알 수 있다. 시뮬레이션 결과 평균대치법은 결측된 자료를 취급하는 적절한 절차가 아님이 밝혀졌다. 특히, 독립변수들의 지수회귀계수를 추정할 때, AEI 방법은 지수회귀모형에서 매우 유용할 뿐 아니라 실질적인 방법으로 추천할 만 하다. 그러나 AEI 방법을 사용하기로 결정했을 때, 그 방법을 올바르게 적용하기 위해서는 먼저 변수 선택에 관심을 두어야 할 것이다. 유의한 공변량 만을 선택하여 AEI 방법을 적용하는 것도 한 방법이 될 수 있을 것이다. CM 방법을 사용하려면 어느 변수를 기준으로 자료를 그룹화하여야 하는 지 등의 여러 가지 제약이 따르므로 실질적으로 AEI 방법을 추천하고자 한다.

참고논문

1. Bello, A. L. (1995). Imputation techniques in regression analysis: Looking closely at their implementation, *Computational Statistics & Data Analysis*, 20, 45-57.
2. Efron, B. (1994). Missing data, imputation, and bootstrap, *Journal of the American Statistical Association*, 89, 463-474.
3. Hegamin-Younger, C., and Forsyth, R. (1998). A comparison of four imputation procedures in a two-variable prediction system, *Educational and Psychological Measurement*, 58, 197-210.
4. Hosmer, D. W., and Lemeshow, S. (1999). *Applied survival analysis*: John Wiley and Sons, New York.
5. Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*: John Wiley and Sons, New York.

[2002년 12월 접수, 2003년 3월 채택]