한국통계학회논문집
2010, 17권, 6호, 899–907

DOI: 10.5351/CKSS.2010.17.6.899

패널자료의 무응답 대체법

박기덕[a], 신기일[1,a]

[a]한국외국어대학교 통계학과

요 약

무응답 대체(non-response imputation) 방법에 관한 많은 이론과 방법이 제안되었으며 실제 자료 분석에 이용되고 있다. 흔히 횡단면 무응답 대체를 위하여 다중 대체법(multiple imputation)이 사용되고 있으며 2차년도 이상의 패널자료에는 종시점회귀대체법(cross-wave regression imputation)이 사용되고 있다. 본 연구에서는 패널자료 분석을 위하여 종시점회귀대체법의 일반 형태인 시계열 대체법과 횡단면 무응답 대체법을 결합한 시계열-횡단면 다중 대체법을 제안하였다. 노동부의 매월노동통계 자료를 이용하여 제안한 방법과 기존의 종시점회귀대체법을 비교하여 우수함을 보였다.

주요용어: 무응답 대체, 종시점회귀대체법, 이월대체법, 다중대체.

1. 서론

표본조사에서 고려해야 할 중요한 내용 중의 하나는 비표본오차 관리이다. 특히 최근의 표본조사에서는 무응답에 의한 결측 비율이 높아지는 추세를 보이고 있다. 따라서 무응답이 발생했을 경우 이를 해결하기 위한 여러 방법이 제안되었다. 그 중에서 대표적으로 사용되고 있는 방법이 무응답 대체이다. 무응답 대체에 관한 중요한 결과는 Little과 Rubin (2002)과 Rubin (1987)에 나와 있다. 국내에서도 무응답 대체에 관한 연구가 진행되고 있다. 예를 들어 김규성 (2000)은 무응답 대체와 그 효과를 분석하였으며 또한 신민웅과 이상은 (2001)은 무응답에 관한 일반적인 내용을 정리하였다. 그러나 이러한 내용들은 과거의 자료를 사용하지 않고 횡단면 자료가 얻어진 경우를 다루고 있다. 최근 이진희와 신기일 (2007)은 횡단면 자료가 아닌 종단면 자료가 주어진 경우에서의 무응답 대체법을 연구하였다. 이 연구는 시계열의 자기상관관계와 자료의 공간상관관계를 이용한 무응답 대체로 종단면 정보와 횡단면 정보를 모두 이용한 대체 방법의 하나로 볼 수 있다.

최근 국내외적으로 많은 패널 자료가 얻어지고 있다. 패널자료는 횡단면 정보 뿐 아니라 종단면 정보를 함께 얻을 수 있기 때문에 정확한 정책입안과 깊이 있는 연구를 위해 필요한 정보를 주고 있다. 패널조사에서 무응답이 발생하게 되면 종단면 정보를 얻을 수 없게 되기 때문에 무응답을 대체하는 것은 매우 중요하다. 특히 김호진 등 (2008)에서와 같이 패널조사에서 얻어진 자료를 이용자들에게 배포할 경우에는 무응답과 같이 결측된 자료를 배포하지 않고 완전한 자료를 배포하여야 하기 때문에 결측치 대체는 반드시 필요하다.

1차년도 뿐만 아니라 2차년도 이후에 발생한 무응답 대체의 경우 시계열적인 연관성을 유지해야 하고 또한 이를 고려한 대체법이 사용되어야 한다. 패널자료에서의 무응답 대체에 관한 국내 연구는 여러 대체방법을 사용하여 GEE 추정량의 효과를 비교분석한 김동욱과 노영화 (2003)가 있다. 이

이 연구는 2010년도 한국외국어대학교 교내연구비에 의해 수행되었음.

[1] 교신저자: (449-791)경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 자연과학대학 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

논문에서는 범주형 반복 측정 자료 분석에서 설명변수에 부분적으로 결측이 발생할 경우 여러 대체법을 사용하여 대체한 후 그 효율을 살펴보았다. 특히 여러 대체법 중에서 종시점회귀대체법(cross-wave regression imputation)과 이의 간편한 방법인 이월대체법(carry-over imputation)과 같은 패널자료에 사용할 수 있는 방법을 소개하였다. 종시점회귀대체와 이월대체에 관한 자세한 내용은 Lepkowski (1989)를 참조하기 바란다.

패널자료처럼 2차년도 이상의 자료에서 무응답이 발생한 경우에는 횡단면 무응답 대체법을 사용하는 것보다 패널자료에 맞는 대체법을 사용하는 것이 더 우수한 결과를 줄 수 있을 것이다. 일반적으로 간단하면서 효과적인 종단면 대체법은 종시점회귀대체법이다. 그러나 단순히 종시점회귀대체법을 사용하는 것보다 종시점회귀대체법과 횡단면 대체법을 결합한 방법을 사용하는 것이 횡단면 정보와 종단면 정보를 모두 사용하기 때문에 우수한 결과를 줄 수 있다. 본 논문에서는 종시점회귀대체법의 일반형태인 시계열-횡단면 대체법을 제안하였으며 종시점회귀대체법과 제안된 대체법을 비교하여 그 우수성을 보였다. 본 논문에서 결측 메카니즘은 MAR(missing at random)을 가정하였으며 분석에 사용된 자료는 노동부의 매월노동통계자료이다.

본 논문의 구성은 다음과 같다. 먼저 2절에서 종시점회귀대체법과 이 방법의 특별한 경우인 이월대체법 그리고 간단한 시계열 분석 방법을 소개하였다. 3절에서는 다중대체법에 관한 내용이 설명되었으며 4절에서는 이를 결합한 새로운 대체법을 설명하였다. 5절에 이 방법들을 비교하기 위한 모의실험 결과가 나와 있으며 6절에 종합적인 결론이 있다.

2. 종단면 대체법

본 논문에서 연구된 종단면 대체법은 종시점회귀대체법과 이의 간편 방법인 이월대체법이다. 이 방법은 Lepkowski (1989)에 소개되었으며 국내에서는 김동욱과 노영화 (2003)에서 다른 대체법과 그 효율성이 비교된 방법이다. 참고로 김동욱과 노영화 (2003)에서는 cross-wave regression imputation을 횡시점회귀대체법이라는 용어로 사용하였다. 또한 미국 Census Bureau의 패널자료인 SIPP(The Survey of Income and Program Participation)에서도 carry-over with R과 같은 이월대체의 응용방법이 사용되고 있다. 이 절에서는 김동욱과 노영화 (2003)의 내용을 간단히 살펴보았다.

2.1. 종시점회귀대체

종시점회귀대체는 조사 시점의 순서에 따라 개별 표본을 정렬한 후 무응답이 있는 t 시점의 자료 y_t를 종속변수로 하고 가장 가까운 과거 시점인 $t-1$ 시점의 응답값 y_{t-1}을 독립변수로 하는 단순회귀모형을 기본으로 한다. 이 경우 먼저 무응답이 있는 조사 단위를 제외한 완전한 자료를 만든 후 회귀모형을 적합한다. 추정된 회귀계수 $\hat{\beta}_0, \hat{\beta}_1$을 이용하여 적합값 $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1}$을 무응답 y_t의 대체값으로 사용한다. 연속적으로 y_t와 y_{t-1}에 무응답이 있는 경우는 $t-1$ 시점의 값 y_{t-1}을 종속변수로 하고 y_{t-2}를 독립변수로 하는 단순회귀모형을 사용하여 $\hat{y}_{t-1}$을 구한다. 구해진 $\hat{y}_{t-1}$으로 무응답을 대체하였으므로 이 대체값을 이용하면 t 시점의 무응답 값을 대체할 수 있다.

2.2. 이월대체

이월대체는 종시점회귀대체법에서 구한 모형 $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1}$에서 오차항 없이 $\hat{\beta}_0 = 0$과 $\hat{\beta}_1 = 1$을 대입하여 무응답 값을 대체하는 방법이다. 즉 $\hat{y}_t = y_{t-1}$로 대체하는 방법으로 전 차수의 응답값을 무응답 대체값으로 사용하는 방법이다. 이 방법은 종시점회귀대체의 간편한 방법으로 생각될 수 있다. 물론 연속 시점에서 무응답이 있는 경우는 응답이 있는 가장 가까운 과거 시점의 응답값으로 대체하게 된다. 이 방법은 미국 Census Bureau에서의 패널자료인 SIPP의 결측 대체를 위하여 사용되었다. 이

때 사용한 방법은 이월대체를 응용한 방법으로 carry over, with random R, method와 carry over, with population R, method이다 (Tremblay, 1994).

2.3. 시계열 모형 대체법

2.3.1. 시계열 분석 이론

시간의 흐름에 따라 얻어진 시계열 자료는 일반적으로 ARIMA 모형을 이용하여 분석하며, ARIMA(p,d,q) 모형 중에서 패널자료에 적합한 모형인 AR(p) 모형은 다음과 같다.

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + a_t,$$

여기서 a_t는 백색잡음과정(white noise process)으로 회귀모형의 오차에 해당되며 $a_t \sim iid(0, \sigma_a^2)$이라 가정한다. 또한 정상성(stationary)을 만족한다고 가정한다.

시계열 모형을 이용하여 모수를 추정하고 예측을 하기 위해서는 충분한 길이의 자료가 확보되어야 하며 이때 사용하는 간단한 모수추정법은 최소제곱추정법(least squares estimation)이다. 모형이 식별되고 모수가 추정되어 모형 검진이 이루어지면 이를 기반으로 예측(forecast)을 하게 된다. 예측 값은 사용된 모형이 AR(p)이므로 회귀분석에서 사용하는 예측 기법이 그대로 적용될 수 있다.

2.3.2. 패널자료 적용

패널자료의 경우 충분히 긴 기간 동안 자료가 얻어질 수 없기 때문에 일반적인 시계열 분석 기법을 사용할 수 없고 패널자료분석 또는 종단면 자료 분석에서 사용되는 방법이 고려될 수 있다. 즉 충분히 긴 기간이 조사되지 않기 때문에 패널자료의 경우에는 AR(p) 모형의 모형 식별에서 큰 차수를 정하는 것은 타당하지 않다. 결국 AR(1) 또는 AR(2) 모형 등 작은 차수의 모형을 선택하는 것이 타당하다. 모수 추정의 경우도 자료의 길이가 짧기 때문에 일반적인 시계열 모수 추정과 달리 패널자료를 이용한 모수 추정 기법을 사용하여야 한다. 이 경우는 먼저 자료를 적당한 셀로 분류한 후 그 셀에 포함된 자료를 이용하여 회귀분석과 같은 방법을 이용하여 추정한다.

만약 자료가 1차년도와 2차년도, 즉 2년간 자료가 얻어졌다면 선택할 수 있는 모형은 AR(1) 모형이다. 이 경우의 모형은 다음과 같다.

$$y_t - \mu = \phi(y_{t-1} - \mu) + a_t, \tag{2.1}$$

여기서 a_t는 오차에 해당되는 백색잡음과정이고 식 (2.1)은 다음과 같이 표현될 수 있다.

$$y_t = (1 - \phi)\mu + \phi y_{t-1} + a_t. \tag{2.2}$$

이제 $(1 - \phi)\mu = \beta_0, \phi = \beta_1, a_t = \varepsilon_t$이라 표시하면 AR(1) 모형은 다음의 단순회귀모형과 같아진다.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t.$$

따라서 종시점회귀대체와 완전히 같은 모형이 된다. 그러나 여기서 주의해야 할 것은 AR(1) 모형이 정상 시계열을 만족하기 위해서 기울기 β_1의 추정량의 절대값은 반드시 '1'보다 작아야 한다는 것이다. 즉 $|\beta_1| < 1$을 만족해야 한다. 이 조건은 시계열 분석의 매우 기본적인 가정인 정상성 조건에 해당된다. 그러나 짧은 구간, 즉 패널자료와 같은 경우에는 이 조건을 만족해야 하는 지에 관하여 고려할 필요가 있다. 무응답 대체를 위한 예측 값은 추정된 절편 $\hat{\beta_0}$와 $\hat{\beta_1}$ 그리고 전년도 자료인 y_{t-1}을 다음의 식에 대입하여 구한다.

$$\hat{y_t} = \hat{\beta_0} + \hat{\beta_1} y_{t-1}. \tag{2.3}$$

다음으로 $\beta_0 = 0$인 경우를 고려할 수 있다. 그러면 모형은 다음과 같이 표현 될 수 있다.

$$y_t = y_{t-1} + a_t. \tag{2.4}$$

이 모형은 시계열 모형 중에서 대표적인 비정상 시계열 모형인 확률보행과정(random walk process)이다. 모형 (2.4)의 특징은 모수를 추정할 필요가 없다는 것이며 또한 절편이 없다는 것이다. 물론 절편이 있는 경우의 확률보행과정(random walk process with drift)도 있으나 이 모형은 직선의 추세가 존재한다는 가정 하에서 사용하는 모형이다. 확률보행과정의 예측 값은 전 시점 자료 값으로 결정되기 때문에 따라서 이월대체 결과와 완전히 일치하게 된다. 3차년도 이상의 자료가 얻어진 경우에는 AR(2) 이상의 모형을 적합할 수 있다. 이 경우에도 정상성을 만족하는 시계열 모형을 적합하는 것을 고려할 필요가 있다. 물론 자료에 따라서는 확률보행과정을 이용할 수도 있다. 이러한 모든 내용은 ARIMA 모형 분석방법을 이용할 수 있으며 자세한 내용은 Wei (1990)를 살펴보기 바란다

3. 다중 대체법

횡단면 대체법에 관한 많은 내용이 이미 알려져 있다. 하나의 대체값을 구하여 대체하는 방법인 단순대체법으로는 평균값대체법, 핫덱대체법 등이 있다. 또한 부가적 정보나 다른 자료로 부터 타당한 응답을 얻을 수 있을 때 두 개 또는 그 이상의 변수들 간에 존재하는 관계를 이용하여 대체하는 비추정대체법과 회귀대체법 등이 있다. 본 논문에서는 최근 그 효용성이 알려진 다중 대체법(multiple imputation)에 관하여 살펴보았다. 특히 이 절에서는 윤성철 (2004)의 다중 대체법에 관한 연구논문의 일부를 간단히 설명하였으며 다중 대체법에 관한 자세한 내용은 Rubin (1987)을 참고하기 바란다. 일반적으로 단순 대체법(single imputation)은 결측치를 가진 자료 분석에 사용하기가 용이하고, 통계적 추론에 사용된 통계량의 효율성 및 일치성 등의 문제를 부분적으로 보완 해준다. 그러나 추정량의 표준오차가 과소추정되는 문제점을 가지고 있다. 이에 단순 대체법의 문제를 보완할 수 있는 다중 대체법이 제안되었는데 다중 대체법은 대체를 한번 하지 않고 m번의 대체를 통한 m개의 가상적 완전한 자료를 만들어 분석하는 방법으로, 분석은 3단계, 대체(imputations step), 분석(analysis step), 결합(combination step)으로 구성되어 있다. 각 단계를 간단히 설명하면 다음과 같다.

3.1. 대체단계(Imputation Step)

결측/무응답을 가진 자료 분석을 할 때 일반적인 결측 메카니즘에 대해서 MAR을 가정을 하게 되며 이 가정 하에서 보편적으로 사용되는 대체법들이 의미를 갖게 된다. 대체 단계에서 사용되는 모형은 매우 다양하며 결측 자료의 형태가 단조일 때 모수적 모형으로는 회귀방법(regression method), 비모수적 모형으로는 성향 방법 등이 있고, 단조가 아닌 경우에는 보통 EM, MCMC(Markov Chain Monte Carlo) 방법이 주로 사용되고 있다. 이러한 대체 방법을 이용해서 가상의 완전한 자료 m개를 생성하게 된다. 가상 데이터 세트 수 m을 많이 생성하면 할수록 보다 바람직한 결과를 도출할 수 있지만, 분석에 있어서 너무 많은 시간이 소요되므로 일반적으로 m은 3에서 5 정도면 충분하다고 알려져 있다.

3.2. 분석단계(Analysis Step)

이 단계에서는 대체단계에서 만든 m개의 완전한 가상 자료 각각을 표준적 통계 분석을 통하여 관심이 있는 추정량 θ_i와 분산 $V_i\,(i = 1, \ldots, m)$을 계산한다. 즉 원자료가 결측/무응답이 없는 자료일 때 분석하는 표준적 통계분석법과 같은 분석기법을 사용하게 된다.

3.3. 결합단계(Combination Step)

이 단계에서는 분석단계에서 생성된 m개의 추정량과 분산의 결합을 통한 통계적 추론을 한다. 추

론 방법 및 결합은 다음과 같다.

1) 다중 대체 추정량: $\theta_{MI} = 1/m \sum_{i=1}^{m} \theta_i$

2) 다중 대체 추정량의 분산: $V_{MI} = V + (1 + m^{-1})B$

대체내 분산(Within imputation variance): $V = 1/m \sum V_i$

대체간 분산(Between imputation variance): $B = 1/(m-1) \sum_{i=1}^{m} (\theta_i - \theta_{MI})^2$

3) 통계적 추론: MI의 $(1-\alpha) \cdot 100\%$ 신뢰구간

$\theta_{MI} \sim T$분포: $\theta_{MI} \pm t_\alpha(df) \sqrt{V_{MI}}$, $df = (m-1)[1 + V/\{(1+m^{-1})B\}]^2$

본 논문에서는 문제를 간단히 해결하기 위해 다중 대체법으로 5개의 자료세트를 생성하였다. 특히 다중 대체법을 사용하는 가장 중요한 이유는 분산의 과소 추정을 방지하는 것이다. 그러나 본 논문의 목적이 분산 추정에 있지 않기 때문에 5개 자료세트의 평균을 구하여 하나의 자료세트를 만들었다. 이는 $\theta_{MI} = 1/m \sum_{i=1}^{m} \theta_i$이기 때문에 평균 자료를 사용하게 되면 추정에 관해서는 다중 대체법의 효과를 그대로 유지할 수 있기 때문이다.

4. 제안된 모형

시계열 모형을 이용한 대체법은 종시점회귀대체와 이월대체를 포함하는 모형이다. 그러나 패널자료만이 갖고 있는 특징이 있을 수 있으므로 시계열 모형을 응용하여 적용할 필요가 있다. 즉 정상성 조건을 완화하는 것도 고려할 필요가 있다. 한편으로는 간단히 횡시점 정보만을 이용하여 대체할 수도 있다. 그러나 횡시점 대체의 대체효과를 높이기 위해서는 대체에 필요한 추가적인 독립변수를 확보할 필요가 있다. 의미 있는 독립변수가 많을 경우 대체의 정확도는 높아질 수 있기 때문이다. 결국 시계열 모형에서 사용된 자료, 즉 1차년도 자료와 횡시점에서 얻어진 독립변수를 동시에 고려한 모형을 설정할 수 있으며 그 결과로 더욱 정확한 대체가 이루어질 수 있다.

다음은 2차년도 독립변수가 하나인 종시점회귀대체법 모형이다.

$$y_t = \beta_0 + \beta_1 x_t + \phi y_{t-1} + a_t, \tag{4.1}$$

여기서 y_t는 2차년도 패널자료, y_{t-1}는 1차년도 패널자료, x_t는 2차년도 패널자료에서 얻어진 독립변수 그리고 a_t는 오차에 해당되는 백색잡음과정이다.

식 (4.1)은 단변량 시계열 모형에 추가로 독립변수가 있을 때 사용할 수 있는 모형인 ARIMAX 모형 또는 전이함수모형(transfer function model)의 하나이다. 일반적인 전이함수 모형은 $y_t = v_i(B)x_{i,t} + n_t$으로 표현될 수 있으며 여기서 $v_i(B) = v_{i0} + v_{i1}B + v_{i2}B^2 + \cdots$이고 B는 $By_t = y_{t-1}$인 후진연산자(backward shift operator)이다. 그리고 y_t는 출력시계열, $x_{i,t}$는 입력시계열 그리고 n_t는 잡음과정이다. 이에 관한 내용은 Wei (1990)을 살펴보기 바란다.

식 (4.1)을 살펴보면 1차 년도에 해당되는 자료(종단면 자료)와 2차년도 패널자료의 독립변수(횡단면 자료)를 동시에 사용한 모형이며 대체 결과는 종단면 정보만을 사용하는 종시점회귀대체 또는 이월대체에 비해 우수할 것으로 판단된다. 또한 2차년 도에서 얻어진 횡단면 자료만을 사용하는 것보다도 우수할 수 있다. 이러한 모형은 독립변수가 2개인 다중 대체법을 이용하여 쉽게 대체 할 수 있게 된다.

5. 모의실험과 자료 분석

본 논문에서 사용된 자료는 2007년 매월노동통계의 임금 자료이다. 매월노동통계자료는 종사자수로 규모를 나누어 층을 구성하는데 본 논문에서는 종사자수 규모가 10–29인: 1층, 100–299인: 2층인

표 1: 사업체당 임금의 기초 통계량

규모	평균		분산		상관계수	기울기
	3월	5월	3월	5월		
10–29인	8,693.95	8,497.49	74,465,136	69,883,357	0.905	0.82
100–299인	86,730.76	84,325.62	6,619,504,471	6,244,846,620	0.943	0.92

표 2: 규모 10–29인의 모의실험 결과

구간	결측치 비율	Method	MSE	BIAS	MAE
6구간	3%	t_R	628,447,072	−1,298	19,252
		t_C	789,025,691	−11,881	21,630
		t_{SM}	294,730,348	−1,017	13,642
	5%	t_R	1,173,401,379	−1,989	26,630
		t_C	1,525,059,271	−19,273	29,983
		t_{SM}	567,312,882	−872	18,780
	10%	t_R	2,535,434,722	292	39,958
		t_C	3,377,642,181	−35,032	46,408
		t_{SM}	1,141,690,203	1,787	27,059
	15%	t_R	4,192,659,717	−2,745	51,377
		t_C	6,092,860,598	−54,487	64,108
		t_{SM}	1,841,207,585	−1,878	34,521
	20%	t_R	5,335,734,870	1,915	57,543
		t_C	8,176,815,584	−67,283	75,569
		t_{SM}	2,462,937,212	−504	40,019
8구간	3%	t_R	718,529,075	−174	20,374
		t_C	849,872,918	−10,609	22,029
		t_{SM}	257,481,272	−496	12,694
	5%	t_R	1,135,650,477	−628	26,238
		t_C	1,368,379,633	−17,347	28,529
		t_{SM}	410,189,790	614	15,930
	10%	t_R	2,705,354,606	−2,881	40,866
		t_C	3,765,173,580	−37,280	48,929
		t_{SM}	934,957,487	−1,819	24,384
	15%	t_R	3,977,806,747	−2,606	50,138
		t_C	6,045,075,263	−54,012	63,192
		t_{SM}	1,460,127,802	−1,088	30,500
	20%	t_R	5,121,103,443	−2,784	56,530
		t_C	8,449,173,640	−70,045	76,681
		t_{SM}	1,863,340,311	−2,078	34,772

두 층을 살펴보았다. 또한 패널자료의 특성에 맞게 3월과 5월 자료를 사용하였다. 1층의 경우 자료수는 1,763개이고 2층인 경우는 920개이다. 각 층의 평균, 분산, 상관계수 그리고 기울기를 살펴본 결과는 표 5.1에 나와 있다. 기울기가 약 0.82와 0.92이므로 종시점회귀대체가 이월대체보다 우수한 결과를 줄 것으로 예상된다.

모의실험에 사용된 종시점회귀대체법은 식 (2.3)을 사용하였다. 즉 단순회귀모형에서 3월 자료를 y_{t-1}, 5월 자료를 y_t로 하여 회귀계수 $\hat{\beta}_0, \hat{\beta}_1$을 추정한 후 결측치에 예측값을 대입하여 총계를 추정하였다. 이월대체법은 식 (2.4)를 사용하였다. 즉 $\hat{y}_t = y_{t-1}$으로 5월 자료의 결측치에 3월 자료를 대체하는 방법이며 이 결과를 이용하여 총계를 추정하였다. 다음으로 본 논문에서 제안한 방법은 횡시점 정보를 종시점 정보에 추가하는 것이다. 일반적으로 횡시점 정보를 얻는 것은 쉬운 일이 아니다. 표 1을 보면 규모 10–29인 경우에 비해 규모 100–299인 규모에서는 사업체당 지불하는 임금이 약 10배인 것을 알

표 3: 규모 100–299인의 모의실험 결과

구간	결측치 비율	Method	MSE	BIAS	MAE
6구간	3%	t_R	21,645,799,206	−1,449	112,699
		t_C	26,159,671,353	−70,253	118,229
		t_{SM}	10,853,292,324	−6,909	81,422
	5%	t_R	35,503,751,782	−3,064	145,434
		t_C	47,399,423,445	−113,056	463,048
		t_{SM}	19,893,281,745	−10,101	111,770
	10%	t_R	79,034,383,995	166	221,031
		t_C	113,935,192,124	−219,278	263,309
		t_{SM}	41,563,549,562	−9,128	159,894
	15%	t_R	114,075,121,932	8,911	268,353
		t_C	190,555,463,807	−313,875	356,970
		t_{SM}	58,731,626,094	−3,725	192,272
	20%	t_R	172,279,683,884	−4,945	333,149
		t_C	313,727,616,780	−445,064	467,450
		t_{SM}	85,934,098,224	−12,089	232,206
8구간	3%	t_R	20,072,087,559	8,041	108,003
		t_C	23,110,051,394	−58,526	111,327
		t_{SM}	6,467,938,739	−985	63,207
	5%	t_R	34,436,399,113	−1,338	144,894
		t_C	45,328,571,838	−112,189	161,356
		t_{SM}	10,218,684,058	−7,695	80,964
	10%	t_R	68,964,008,770	−238	207,021
		t_C	106,776,413,556	−220,113	259,791
		t_{SM}	21,875,469,259	−1,613	115,795
	15%	t_R	123,604,118,641	−13,928	282,806
		t_C	212,146,859,268	−347,857	377,602
		t_{SM}	32,675,971,607	−8,426	144,281
	20%	t_R	159,753,693,210	−14,998	318,629
		t_C	312,785,609,238	−453,094	470,903
		t_{SM}	47,770,807,171	−2,957	175,397

수 있다. 물론 조사표에는 사업체마다 지불한 총임금을 기재하여야 하기 때문에 이 값이 얻어진다. 본 논문의 자료 분석에서도 임금과 관련된 횡시점 자료를 얻는 것은 쉬운 일이 아니다.

최근 장애인고용패널조사에서는 조사표에 연속형 자료외에 추가로 범주형 자료의 기입항목을 마련하였다. 즉 정확한 연속형 자료의 기입을 거부할 경우 이 연속형 자료에 해당되는 범주형 자료를 반드시 기입하도록 하였다. 이 범주형 자료는 대체에 필요한 많은 정보를 보유하고 있다. 본 모의실험에서는 임금과 관련되어 사용할 수 있는 횡단면 독립변수가 존재하지 않아, 2차년도의 범주형 임금자료를 횡시점 대체의 독립변수로 사용하였다. 따라서 새로운 대체법은 식 (4.1)에서 x_t는 5월의 범주형 임금자료, y_{t-1}은 3월의 임금자료를 사용하여 다중 대체법을 적용하는 것이다. SAS의 Proc MI를 사용하여 5개의 자료 세트를 얻은 후 이를 평균하여 총계를 추정하였다.

사용한 비교통계량은 MSE, BIAS 그리고 MAE로 정의는 다음과 같다.

$$\text{MSE} = \frac{1}{R}\sum_{i=1}^{R}(t_i - \hat{t}_i)^2, \qquad \text{BIAS} = \frac{1}{R}\sum_{i=1}^{R}(t_i - \hat{t}_i), \qquad \text{MAE} = \frac{1}{R}\sum_{i=1}^{R}\left|t_i - \hat{t}_i\right|,$$

여기서 t_i는 총계의 참값이고 $\hat{t}_i$는 총계 추정값이며 $R = 1{,}000$의 반복을 실시하였다. MAR 결측 메카니즘을 이용하여 결측하였으며, 연속형 자료를 범주형 자료로 변환하여 x_t 자료를 생성하였다. 결측 비

율은 각각 3, 5, 10, 15 그리고 20%이다. 또한 범주형 자료는 4, 6, 8, 10 구간으로 나누어 구간의 수에 따라 얼마나 총계 추정에 영향을 주는 지 살펴보았다. 구간 수별 결과가 유사하여 본 논문에서는 6 구간과 8 구간의 결과를 표로 나타내었으며 표 2와 3에서 t_R은 종시점회귀대체, t_C는 이월대체 그리고 t_{NM}은 제안된 방법을 의미한다.

표 2의 모의실험 결과를 살펴보면 먼저 이월대체인 t_C에 비해 종시점회귀대체인 t_R이 MSE, BIAS 그리고 MAE 등 모든 비교통계량에서 우수한 결과를 나타내고 있다. 특히 결측비율이 증가할 수록 그 차이는 커지고 있다. 제안된 방법으로 구해진 t_{SM}은 횡시점 정보와 종시점 정보를 모두 사용하기 때문에 다른 두 방법에 비해 매우 우수한 결과를 주고 있다. 특히 구간의 수가 증가할수록 추가적인 정보가 많아져 더욱 우수한 결과를 주고 있다. 그러나 실제 조사에서 구간 수가 많아지면 응답자의 응답부담은 커지게 되므로 적절한 구간 수를 정해야 한다. 이러한 결과는 2층의 결과인 표 3에서도 공통적으로 얻어지고 있다.

6. 결론

이월대체법은 두 시점간의 원자료를 분석하지 않고 전 시점의 자료로 결측치를 대체하는 간편하면서도 효과적인 대체법이다. 그러나 기울기가 '1'이 아닌 경우에는 그 효과가 떨어지는 단점이 있다. 반면에 원자료 분석이 가능하다면 종시점회귀대체법을 사용할 수 있다. 대부분의 경우, 특히 기울기가 '1'이 아닌 경우 종시점회귀대체법이 우수한 결과를 주고 있다. 본 논문에서는 종시점회귀대체법을 확장한 방법을 제안하였다. 즉 횡시점 정보와 종시점 정보를 모두 이용할 수 있는 방법을 제안하였으며 종시점 정보만을 이용하는 방법에 비해 우수한 결과를 주고 있다. 많은 경우 주요 변수와 상관관계가 높은 자료를 찾는 것은 쉬운 일이 아니다. 그러나 연속형 주요 변수에 추가하여 범주형 변수를 얻어 이 정보를 대체법에 추가한다면 본 논문에서 얻어진 결과처럼 정도 높은 대체를 할 수 있을 것이다.

참고 문헌

김규성 (2000). 무응답대체방법과 대체효과, <조사연구>, **1**, 1–14.

김동욱, 노영화 (2003). 대체방법별 GEE추정량 비교, <응용통계연구>, **16**, 407–426.

김호진, 류정진, 장영석, 류기섭 (2008). <제1차 장애인고용패널조사>, 한국장애인고용촉진 공단 고용개발원.

신민웅, 이상은 (2001). <표본조사를 위한 표본설계>, 교우사.

윤성철 (2004). 결측값의 대치법, <예방의학회지>, **37**, 211–219.

이진희, 신기일 (2007). 공간-시계열 모형을 이용한 결측 대체 방법에 관한 연구, <응용통계연구>, **20**, 499–514.

Lepkowski, J. M. (1989). *Treatment of Wave Nonresponse in Panel Survey*, In Panel Survey, Hohn Wiley & Sons, 348–374.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Survey*, John Wiley & Sons, New York.

Tremblay, A. (1994). *Longitudinal Imputation of SIPP Food Stamp Benefits, U.S Department of Commerce*, Bereau of the Census, **208**, http://www.census.gov/sipp/workpapr/wp208.pdf.

Wei, W.W.W (1990). *Time Series Analysis, Univariate and Multivariate Methods*, Addison Wesley, New York.

2010년 8월 접수; 2010년 10월 채택

Non-Response Imputation for Panel Data

Gi-Deok Pak[a], Key-Il Shin[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies

Abstract

Several non-response imputation methods are suggested, however, mainly cross-sectional imputations are studied and applied to this analysis. A simple and common imputation method for panel data is the cross-wave regression imputation or carry-over imputation as a special case of cross-wave regression imputation. This study suggests a multiple imputation method combined time series analysis and cross-sectional multiple imputation method. We compare this method and the cross-wave regression imputation method using MSE, MAE, and Bias. The 2008 monthly labor survey data is used for this study.

Keywords: Non-response imputation, cross-wave regression imputation, carry-over imputation, multiple imputation.

This research was supported by the research fund of Hankuk University of Foreign Studies, 2010.

[1] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, San 89, Wangsan, Mohyun, Yongin, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr