연구논문

# 패널조사 웨이브 무응답의 대체방법 비교*

Comparisons of Imputation Methods for Wave Nonresponse in Panel Surveys

김규성** · 박인호***
Kyu−Seong Kim · Inho Park

본 논문에서는 패널조사에서 발생하는 웨이브 무응답을 대체하는 방법을 고찰하였다. 패널조사에서는 이전 조사 데이터를 무응답 대체에 활용할 수 있기 때문에 이러한 성질을 이용하면 횡단면 무응답 대체보다 더 효과적인 웨이브 무응답 대체법을 찾을 수 있다. 먼저 웨이브 무응답 대체를 사용하는 해외의 주요 패널조사를 살펴보고, 웨이브 무응답 대체방법 중 종단면 회귀대체법, 이월대체법, 최근방 회귀대체법, 그리고 행렬대체법을 고찰하였다. 그리고 웨이브 무응답 대체법의 성능을 비교하기 위하여 한국복지패널 데이터를 대상으로 모의실험을 실시하였다. 성능을 비교하기 위하여 평균대체, 회귀대체, 비대체, 최근방 대체, 핫덱 대체를 고려하였고 성능평가 지표로는 예측 정확성 지표와 추정 정확성 지표를 이용하였다.

모의실험 결과 비대체, 행렬대체는 두 지표 모두 우수했고, 회귀대체, 종단면 회귀대체, 이월대체는 예측 정확성은 우수한 반면 추정 정확성은 다소 떨어졌으며, 반대로 최근방 회귀대체, 최근방 대체, 핫덱 대체는 예측 정확성은 떨어지나 추정 정확성은 높은 것으로 나타났다. 마지막으로 평균 대체는 두 지표 모두 좋지 않았다.

**주제어**: 이월대체, 종단면 회귀대체, 최근방 회귀대체, 행렬대체

We compare various imputation methods for compensating wave nonresponse that are commonly adopted in many panel surveys. Unlike the cross−sectional survey, the panel survey is involved a time−effect in nonresponse in a sense that nonresponse may happen for some but not all waves. Thus, responses in neighboring waves can be used as powerful predictors for imputing wave nonresponse such as in longitudinal regression imputation, carry−over imputation, nearest neighborhood regression imputation and

* 이 논문은 2009년도 한국은행 경제통계국 연구용역비 지원에 의한 것임.

** 교신저자(corresponding author): 서울시립대학교 통계학과 교수 김규성.
E−mail: kskim@uos.ac.kr

*** 한국은행 경제통계국 전문연구원

row−column imputation method. For comparison, we carry out a simulation study on a few income data from the Korean Welfare Panel Study based on two performance criteria: predictive accuracy and estimation accuracy. Our simulation shows that the ratio and row−column imputation methods are much more effective in terms of both criteria. Regression, longitudinal regression and carry−over imputation methods performed better in predictive accuracy, but less in estimation accuracy. On the other hand, nearest neighborhood, nearest neighbor regression and hot−deck imputation show higher performance in estimation accuracy but lower predictive accuracy. Finally, the mean imputation shows much lower performance in both criteria.

**Key words** : carry−over imputation, longitudinal regression imputation, nearest neighbor regression imputation, row−column imputation

## Ⅰ. 서론

패널조사에서는 초기표본을 반복조사하므로 횡단면 조사에서는 발생하지 않는 시간에 대한 효과가 나타난다. 무응답과 관련하여 횡단면 조사에서는 단위무응답(unit nonrsponse)과 항목무응답(item nonrsponse)으로 구분하는 것으로 충분하지만, 패널조사에서는 웨이브 무응답(wave nonresponse)이 추가된다. 동일 조사단위에 대하여 여러 시점에 걸쳐 정보를 수집할 때 하나 이상의 시점에서 응답을 얻지 못하는 경우 이러한 무응답을 웨이브 무응답이라고 한다. 예컨대 어느 표본가구가 1차 조사에서는 응답을 하고 2차 조사에서는 응답을 하지 않다가 3차 조사에서 다시 응답을 하였다면, 이러한 2차년도 무응답이 웨이브 무응답이다. 웨이브 무응답은 횡단면(cross sectional) 관점에서는 단위무응답으로 간주되어 가중치조정법으로 처리될 수 있고, 종단면(longitudinal) 관점에서는 항목무응답으로서 대체방법으로 처리될 수 있다.

〈표 1〉은 한국보건사회연구원이 실시하고 있는 한국복지패널의 응답 현황을 보여 주고 있다(한국보건사회연구원 2008). 1차 웨이브(2006년 조사)부터 3차 웨이브(2008년 조사)까지 모두 응답한 가구는 6,101가구로 전체 초기표본가구의 83.95%이다. 나머지 1,166가구(16.05%)는 3차례 웨이브 중 한 웨이브 이상 응답을 하지 않았다. 예컨대 웨이브3의 무응답 가구는 953가구로서 3차 웨이브 자료만 가지고 분석을 한다면 953가구는 가중치조정을 통

**〈표 1〉 한국복지패널의 웨이브 응답 현황**

(단위: 명, %)

| 번호 | 구분 | | 웨이브1 (2006년) | 웨이브2 (2007년) | 웨이브3 (2008년) | 응답 및 응답률 | |
|---|---|---|---|---|---|---|---|
| 1 | 완전응답 | | √ | √ | √ | 6,101 | 83.95 |
| 2 | 소멸패턴 | | √ | √ | | 410 | 5.64 |
| 3 | | | √ | | | 534 | 7.35 |
| 4 | 비소멸 패턴 | 재진입 | √ | | √ | 27 | 0.37 |
| 5 | | 나중 진입 | | | √ | 126 | 1.73 |
| 6 | | | | √ | √ | 60 | 0.83 |
| 7 | | 늦은 진입 후 소멸 | | √ | | 9 | 0.12 |
| 8 | 완전무응답 | | | | | - | - |
| | | | 7,072 | 6,580 | 6,314 | 7,267 | 100 |

하여 처리가 될 것이다. 그러나 이 중에서 완전무응답인 경우를 제외하면 동일한 가구에서 동일한 조사항목에 대하여 웨이브1이나 웨이브2의 응답이 있으므로 분석에 이를 이용할 수 있다. 즉, 동일한 항목을 시차를 두고 여러 번 조사하고 이 중 일부 시점에서 무응답이 발생한 경우이므로 다른 시점의 응답을 이용하여 웨이브 무응답을 항목무응답으로 간주하고 대체 처리할 수 있는 것이다.

패널조사의 웨이브 무응답 대체에는 횡단면 대체법과 웨이브 무응답 대체법을 동시에 고려할 수 있다. 1차 웨이브에서는 항목무응답에 대한 처리는 횡단면 무응답 대체가 불가피하다. 그러나 2차 웨이브와 3차 웨이브에서는 횡단면 무응답 대체법뿐만 아니라 웨이브 무응답 대체법도 가능하다. 본질적으로 횡단면 무응답 대체는 동 시점의 보조정보를 대체에 이용하는 것이고 웨이브 무응답 대체는 이전 시점의 동일 항목 응답을 이용하는 것이므로, 만일 관심변수의 응답이 시간 변화에 따라 크게 다르지 않고 안정된 패턴을 보인다면 횡단면 대체보다는 이전 시점의 응답을 이용하는 웨이브 무응답 대체가 더 효과적이다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 외국의 주요 패널조사에서 사용하고 있는 웨이브 무응답 대체법을 간단하게 살펴본다. 본 논문에서 검토하는 5개 주요 패널조사는 다른 연구에서도 많이 인용하는 패널조사인데, 소득과 같은 주요 변수의 무응답은 웨이브 무응답 대체를 하고 있는 것으로 나타났다. 이들 조사에서 사용하는 무응답 대체 방법은 웨이브 무응답 처리를 도입하고자 하는 패널조사에서 참고가 될 것이다. 3절에서는 무응답 대체법을 살펴본다. 횡단면 조사의 무응답 대체 기법은 우리나라에서도 이미 많이 소

개되었고 적용사례도 다수 알려져 있다(예를 들면 김영원 · 조선경 1996; 김규성 2000; 김규성 외 2005; 조영숙 외 2008). 그리고 횡단면 무응답 대체법은 패널조사의 웨이브 무응답 처리에서도 여전히 유용한 기법이다. 웨이브 무응답 대체 방법으로 종단면 회귀대체법, 이월대체법, 최근방 회귀대체법, 행렬 대체법을 고찰하기로 한다. 이어서 무응답 대체 방법을 선택하는 기준에 대하여 설명한다. 4절에서는 3절에서 고찰한 웨이브 무응답 대체법과 횡단면 무응답 대체법을 비교하기 위하여 실시한 모의실험을 논의한다. 모의실험에 사용한 데이터는 한국보건사회연구원의 한국복지패널 데이터로 이 중 소득변수를 대상으로 하였다. 마지막으로 5절에서는 연구내용을 요약한다.

## II. 웨이브 무응답 대체 사례

### 1. 영국의 가구패널조사

영국의 가구패널조사(British Household Panel Survey; BHPS)는 영국의 개인 및 가구 단위에서 일어나는 사회적, 경제적 변화에 대한 이해를 증진시키기 위해 가구의 주거상태, 소비지출, 고용상태, 재무관련, 그리고 건강 등을 조사하고 사회경제구조의 변화추세를 조사한다. 조사대상은 16세 이상의 성인이고 1991년 1차년도가 시작된 이래 현재 17차(2007년) 자료까지 공개되고 있다. 1차 조사에서는 약 5,500가구(약 10,000가구원)를 표본으로 구성하여 조사하였고, 9차년도 조사부터는 잉글랜드와의 정부정책 효과성분석을 위해 스코틀랜드와 웨일즈 지역의 표본 수를 약 1,500가구 확대하였으며, 11차년도에는 북 아일랜드 표본 약 2000가구를 추가하면서 영국의 전 지역을 포괄하는 표본을 갖게 되었다. 현재 BHPS의 전체 표본 수는 약 10,000가구이다(Tayler et al. 2007).

BHPS에서는 모든 웨이브의 소득변수와 가구비용 변수의 무응답이 대체된다. 범주형 소득관련 변수는 핫덱 대체법으로 무응답을 처리하고 있고 연속형 소득관련 변수는 최근방 회귀대법으로 무응답을 처리하고 있다.

### 2. 호주의 가구, 소득 및 노동력변화 조사

호주의 가구, 소득 및 노동력변화 조사(Household, Income and Labour Dynamics in Australia; HILDA)는 2001년 이래 매년 시행되는 광범위한 사회경제조사로서, 특히 가족과 가구의 형태 그리고 소득과 직업에 초점을 둔 대규모 조사이다(Watson 2004; Watson

et al. 2008). 2001년도 1차 조사에서는 7,682가구와 가구의 성인을 대상으로 19,914명의 가구원을 대상으로 조사했다. 연간 총소득은 지난해를 기준으로 하여 모든 직종에서 받은 총임금과 자영업자나 농어업 경영자의 소득 그리고 배당, 수익, 손실 등을 합하여 얻는다.

무응답 가구는 응답 가구를 이용하여 가중치 조정법으로 처리한다. 그리고 응답가구나 무응답 가구의 무응답 가구원은 응답 가구원으로 가중치를 조정한다. 응답자의 소득 항목이 무응답일 경우는 대체 처리한다. 따라서 가구 수준에서 소득변수는 대체 처리한 데이터가 제공된다.

소득관련 변수의 무응답은 일차적으로 대체군 안에서 행렬대체법을 이용하여 처리한다. 대체군은 나이 그룹으로 묶어 15~19세, 20~24세, 25~34세, 35~44세, 45~54세, 55~64세, 그리고 65+로 구분한다. 일부의 경우, 예를 들어 최근 조사에 새로 진입하는 가구는 직전 응답이 없으므로 영국의 BHPS에서와 같은 최근방 회귀대체법을 사용하여 무응답을 대체한다. HILDA 2.0에서는 무응답 소득변수를 최근방 회귀대체를 하였으며, HILDA 3.0에서는 행렬대체법을 채택하여 소득변수를 대체하고 있다.

### 3. 독일의 사회 · 경제패널조사

독일의 사회 · 경제 패널조사(German Socio-Economic Panel; GSOEP)는 유럽의 가구 패널조사 중 가장 오래된 조사로서 1984년에 시작되었다. 1984년 1차 조사에서는 서독지역만을 조사대상 지역으로 하였으나 1990년과 1995년에 동독지역의 가구가 표본에 포함되었다. 연간 총근로소득은 10개의 질문에서 얻는다. 전년도의 근로소득과 자영업자 월소득을 소득의 종류에 따라 누적하여 연소득을 구한다(Frick & Grabka 2007).

독일의 GSOEP에서는 소득변수에서 발생하는 무응답의 대체를 위하여 과거의 동일한 소득변수가 사용 가능한 경우에 행렬대체법을 사용하여 대체한다. 일부 경우에는 다른 방법들을 병용하기도 한다. 무응답률이 낮을 때에는 평균대체를 사용하고, 소득구조가 복잡한 경우에는 회귀대체를 사용한다.

### 4. 미국의 소득 및 프로그램 참여 조사

미국의 소득 및 프로그램 참여 조사(Survey of Income and Program Participation; SIPP)는 소득의 원천과 금액, 노동관련 정보, 사회복지 프로그램 참여 정도와 수혜자 여부 및 기타 인구사회학적 특성들을 조사하여 현존하는 정부의 복지프로그램의 효과를 검증하

고 미래의 대상자 규모와 소요비용을 추계하기 위한 조사이다. 특히 SIPP는 약 70개 항목의 현금소득 및 현물소득에 대한 다양한 정보를 제공하며 조세, 자산, 부채 및 정부의 소득이전 프로그램의 수혜 여부와 급여수준 등에 대해서 조사하고 있다(U.S. Bureau of the Census 2001).

미국의 SIPP에서는 두 가지 대체방법을 사용한다. 항목무응답은 축차 핫덱(sequential hot deck)을 사용하여 대체하고, 웨이브 무응답은 랜덤 이월 대체방법(random carryover method)을 이용하여 대체한다.

### 5. 캐나다의 노동 및 소득 변화 조사

캐나다의 노동 및 소득 변화 조사(Survey of Labour and Income Dynamics; SLID)에서는 개인과 가족들의 경제적인 풍요와 이러한 풍요에 영향을 미치는 요인들에 대한 연구를 기본 목적으로 하고 있으며, 소득 및 노동시장 그리고 가족 등 다방면에 걸친 동태적인 변화를 조사하고 있다.

SLID에서는 웨이브 무응답 대체방법으로 전년도 응답값이 있을 때에는 직전값 전진 이월대체를 실시하며 전년 데이터가 없는 경우에는 최근방 대체를 실시하고 있다.

## III. 웨이브 무응답 대체방법

패널조사의 웨이브 무응답은 횡단면 무응답 대체법으로도 처리될 수 있다. 횡단면 조사의 무응답 대체법은 이미 많이 알려져 있으므로 과거 시점의 조사 데이터를 웨이브 무응답 대체에 사용하지 않고 횡단면 무응답 대체법을 패널조사에 그대로 적용하면 된다. 횡단면 무응답 대체법에 대한 구체적인 설명은 다른 문헌을 참고하기 바란다(예를 들면 Little & Rubin 2002; 김규성 2000 등). 본 논문에서는 웨이브 무응답 대체법과의 비교를 위하여 횡단면 대체법 중 평균대체법, 회귀대체법, 비대체법, 최근방 대체법, 핫덱 대체법을 다음절의 모의실험에서 사용하였다.

### 1. 웨이브 무응답 대체법

웨이브 무응답 대체방법을 설명하기에 앞서 다음과 같은 기호를 사용한다. $y$를 대체를 실시하려고 하는 항목 조사변수라고 하고 $x$를 보조변수라고 하자. 조사 시점(wave)은 첨

자 $t(t=1,...,T)$로 나타내고, 조사 단위는 $j$로 표시하며, 대체군은 $g(g=1,...,G)$로 나타내자. 그러면 $t$시점의 대체군 $g$의 $j$번째 단위의 관심 항목 변수는 $y_{gjt}$가 되고 대응하는 보조변수는 $x_{gjt}$이다. 또한 무응답을 처리한 값을 대체값 $y^*_{gjt}$로 나타내자. 응답 단위에서는 $y^*_{gjt}=y_{gjt}$이고 무응답 단위에서는 무응답이 $y^*_{gjt}$로 대체된다.

대체군 $g$의 표본 수를 $n_{gt}$라 하고, 이 중 응답자 수를 $m_{gt}$, 그리고 무응답자 수를 $r_{gt}$라고 하면 $n_{gt}=m_{gt}+r_{gt}$가 된다. $t$시점의 전체 표본 수는 $n_t=\sum_{g=1}^{G} n_{gt}$이다.

1) 종단면 회귀 대체

일반적으로 패널조사에서는 동일 항목에 대하여 두 시점 간의 응답 사이에 강한 상관관계가 있을 가능성이 크다. 따라서 보조변수로 직전 조사의 응답을 사용한 회귀모형은 유의한 모형이 된다.

$$y_{gjt}=\beta_{gto}+\beta_{gt1}\,y_{gj(t-1)}+\epsilon_{gjt}$$

이 모형 중 $\beta_{gt0}=0$인 경우가 비례 변화 모형이고 $\beta_{gt1}=0$인 경우는 가법 변화 모형이다.

종단면 웨이브 무응답 대체값으로는 추정된 회귀모형으로부터 적합치을 구해 사용한다.

$$\begin{aligned} y^*_{gjt} &= \widehat{\beta_{gt0}}+\widehat{\beta_{gt1}}y_{gj(t-1)} \\ &= \bar{y}^{(r)}_{gt}+\widehat{\beta_{gt1}}(y_{gj(t-1)}-\bar{y}^{(r)}_{g(t-1)}) \end{aligned}$$

여기에서 $\bar{y}^{(r)}_{gt}$은 $t$ 시점의 $g$ 대체군의 응답자 평균이다.

2) 이월대체

이월 대체는 종단면 회귀대체의 극단적인 경우로 볼 수 있다. 항목 응답이 시간변화에 안정적인 경우, 위 회귀모형에서 $\beta_{gt0}=0$, $\beta_{gt1}=1$, $\epsilon_{gjt}=0$을 가정할 수 있다. 즉, $y_{gjt}=y_{gj(t-1)}$. 그러면 웨이브 무응답은 직전 조사값으로 대체될 것이다. 즉,

$$y^*_{gjt}=y_{gj(t-1)}$$

이와 같이 직전조사의 응답값으로 대체하는 경우를 직전값 전진 이월대체(last value carried forward)라고 한다.

웨이브 무응답 처리에서는 이후 조사의 응답값도 이용 가능하므로 이후 조사의 응답값을 대체값으로 사용하면 직후값 후진 이월대체(next value carried backward)가 된다.

$$y^{*}_{gjt} = y_{gj(t+1)}$$

혹은 직전과 직후에 응답이 있는 경우 둘 중의 하나를 랜덤하게 이월대체할 수 있는데 이를 랜덤 이월대체(random carryover)라고 한다. 즉,

$$y^{*}_{gjt} = y_{gj(t-1)} \text{ 혹은 } y_{gj(t+1)}, \text{ 각 확률은 } 0.5$$

이월대체 방법은 항목값이 웨이브 변화에 매우 안정적이라는 가정을 하고 있으므로 이 방법은 웨이브 간 차이의 변화를 과소평가하는 단점이 있다.

3) 최근방 회귀대체 혹은 예측평균 매칭

횡단면 무응답의 최근방 대체법과 마찬가지로 웨이브 무응답에서도 최근방 대체법을 사용할 수 있다. 횡단면 최근방 대체법은 최근방 단위를 보조변수의 거리로 찾는 반면, 웨이브 최근방 대체법에서는 최근방 단위를 동일 항목의 전시점 응답으로 찾는다는 점이 다르다. 즉,

$$y^{*}_{gjt} = y_{gkt} : \ |y_{gj(t-1)} - y_{gk(t-1)}| \leq |y_{gj(t-1)} - y_{gl(t-1)}|, \quad (l \neq k)$$

최근방 회귀대체(nearest neighbor regression imputation) 혹은 예측평균 매칭(predictive mean matching) 방법은 최근방 단위를 회귀모형 적합값으로 찾고 대응하는 값을 대체하는 방법이다(Little & Rubin 2002). 즉,

$$y^{*}_{gjt} = y_{gkt} : \ |\hat{y}_{gjt} - \hat{y}_{gkt}| \leq |\hat{y}_{gjt} - \hat{y}_{glt}|, \quad (l \neq k)$$

여기에서 적합값 $\hat{y}_{gjt}$는 직전 조사의 응답을 보조변수로 사용하여 구한다. 즉,

$$\hat{y}^{*}_{gjt} = \widehat{\beta_{gt0}} + \widehat{\beta_{gt1}} y_{gj(t-1)}$$

4) 행렬대체

행열 대체법(row−and−column imputation method)은 이웃 단위 정보(행)와 과거 시점 정보(열)를 동시에 이용하는 대체방법이다(Little and Su 1989). 이 방법은 행(단위)과 열(시점) 정보의 조합에 의하여 시점이 지나도 매우 유사한 응답 정보를 이용할 수 있는 이점이 있다. 이 방법에서는 행 효과(개인 효과), 열 효과(웨이브 효과), 그리고 잔차 효과를 구한 후 세 효과를 곱하여 대체값을 얻는다.

먼저 열(wave) 효과는 다음과 같이 구한다.

$$c_{gt} = \frac{\bar{y}_{gt}}{\bar{y}_g}, \quad t = 1,2,\cdots,T, \quad g = 1,2,\cdots,G$$

여기에서 $\bar{y}_{gt}$는 $t$시점의 대체군 $g$에서 계산한 표본평균으로 $T$시점에 걸쳐 완전 응답한 단위에서 계산한다. 또한 $\bar{y}_g = \sum_{t=1}^{T} \bar{y}_{gt} / T$ 이다.

다음으로 행(person) 효과는 다음과 같이 구한다.

$$\bar{y}_g^{(j)} = \frac{1}{T_j} \sum_{t=1}^{T_j} \frac{y_{gjt}}{c_{gt}}$$

여기에서 $\sum_{t=1}^{T_j}$는 $j$번째 개인의 응답의 합을 의미하고, $T_j$는 응답한 웨이브의 수이다. 이어서 각각의 단위를 $\bar{y}_g^{(j)}$에 의해 정렬하고 무응답 단위 $(gj)$를 대체군 $g$에서 가장 가까운 응답 단위와 연계(matching)한다. 이를 $(gl)$이라고 하자. 그러면 웨이브 무응답 데이터 $y_{gjt}$는 다음과 같이 행렬대체된다.

$$y_{gjt}^{*} = [\bar{y}_g^{(j)}][c_{gt}][\frac{y_{glt}}{\bar{y}_g^{(l)} c_{gt}}] = y_{glt} \frac{\bar{y}_g^{(j)}}{\bar{y}_g^{(l)}}$$

괄호 안의 세 항은 각각 행 효과, 열 효과 그리고 잔차 효과를 나타낸 것이다. 처음 두 개의 항은 예측평균을 추정한 것이고, 마지막 항은 연결된 자료를 이용한 대체방법의 확률적 요소이다(Starick 2005; Starick & Watson 2007).

## 2. 대체방법 평가기준

본질적으로 무응답 대체는 예측의 한 형태이다. 현재 가지고 있는 정보를 최대로 이용하여 무응답값을 예측한 후 예측값으로 실제값을 대신하는 것이 대체이다. 당연히 대체방법에 따라 예측값은 달라지므로 대체 방법을 평가하여 문제에 맞는 대체방법을 찾는 것이 중요하다.

대체방법을 선택하는 기준으로 정확성(accuracy)과 적절성(plausibility)이 있다(Little 1988; Chambers 2000). 정확성은 다시 세부적으로 예측 정확성(predictive accuracy), 순위 정확성(ranking accuracy), 분포 정확성(distributional accuracy), 그리고 추정 정확성(estimation accuracy)으로 나누어 볼 수 있다. 예측 정확성은 대체 절차가 실제값의 데이터 구조를 유지하는지를 나타내는 것이다. 즉, 실제값에 가까운 대체값이 예측 정확성이 높다고 할 수 있다. 순위 정확성은 대체 절차가 대체값의 순위를 유지하는지를 나타낸다. 즉, 실제값의 순위와 대체값의 순위를 유지하는 대체법이 순위 정확성이 높은 대체법이다. 분포 정확성은 실제값의 분포와 대체값이 분포가 얼마나 일치하는가를 보는 것이다. 대체후 대체값의 주변분포와 고차항 분포가 실제값의 분포와 비슷할수록 분포 정확성이 높은 대체방법이다. 추정 정확성은 실제값으로 분석한 추정결과와 대체값으로 분석한 추정결과 간의 관계를 보는 것이다. 추정결과가 비슷할수록 추정 정확성이 높다고 할 수 있다.

적절성은 대체 후의 값이 그럴듯하여야 함을 의미한다. 특히 통계조사에서는 에디팅 과정에서 데이터들의 일치성을 점검하는데, 잘못하면 대체로 인하여 데이터의 일치성이 상실될 수도 있다. 따라서 에디팅 편집규칙에 어긋나지 않는 값으로 대체되어야 한다. 그리고 이러한 바람직한 대체법 평가기준은 대체과정에 반영되어야 한다. 본 논문에서는 이들 중 예측 정확성과 추정 정확성 측도만을 살펴보았다.

### 1) 예측 정확성 측도

예측 정확성은 응답값과 대체값의 차이를 비교하여 평가할 수 있다. 대체값이 응답값과 유사할수록 예측이 정확하다고 평가하는 것이다. 유사성은 대체값과 응답값의 거리로써 측정한다. 절대거리($|y^*_{gjt} - y_{gjt}|$)를 사용한 지표가 평균절대편차(Mean Absolute Deviation: MAD)이고 제곱거리($(y^*_{gjt} - y_{gjt})^2$)를 사용한 지표가 제곱근평균제곱편차(Root Mean Square Deviation: RMSD)이다.

- 평균절대편차 : $MAD_t = \frac{1}{n_t} \sum_{g=1}^{G} \sum_{j=1}^{n_{gt}} |y^*_{gjt} - y_{gjt}|$

- 제곱근평균제곱편차 : $RMSD_t = \sqrt{\frac{1}{n_t}\sum_{g=1}^{G}\sum_{j=1}^{n_{gt}}(y_{gjt}^{*} - y_{gjt})^2}$

여기에서 $n_t = \sum_{g=1}^{G} n_{gt}$는 $t$ 시점의 표본 수이다.

조사변수의 척도의 영향을 배제하기 위해서는 위 지표를 각 변수의 평균으로 나눈 상대지표를 사용하는 것이 효과적이다. 상대평균절대편차(Relative Mean Absolute Deviation; RMAD)와 상대제곱근평균제곱편차(Relative Root Mean Square Deviation; RRMSD)는 다음과 같다.

- 상대평균절대편차 : $RMAD_t = \frac{1}{\overline{Y}_t}\frac{1}{n_t}\sum_{g=1}^{G}\sum_{j=1}^{n_{gt}}|y_{gjt}^{*} - y_{gjt}|$
- 상대제곱근평균제곱편차 : $RRMSD_t = \frac{1}{\overline{Y}_t}\sqrt{\frac{1}{n_t}\sum_{g=1}^{G}\sum_{j=1}^{n_{gt}}(y_{gjt}^{*} - y_{gjt})^2}$

여기에서 $\overline{Y}_t$는 $t$시점의 모평균이다. 위의 네 가지 지표는 모두 작은 값을 가질수록 더 좋은 대체방법이다.

2) 추정 정확성 측도

추정 정확성은 응답값과 대체값으로 평균(1차 적률), 분산(2차 적률), 왜도(3차 적률) 그리고 첨도(4차 적률)을 구한 후 두 값의 차이로 평가한다. 그리고 예측 정확성에서와 마찬가지로 조사변수의 척도에 대한 영향을 배제하기 위하여 모평균으로 나누어 준 상대편차를 지표값으로 사용한다.

$$RD_t^k = \frac{1}{\overline{Y}_t}\frac{1}{n_t}\sum_{g=1}^{G}\sum_{j=1}^{n_{gt}}(y_{gjt}^{*k} - y_{gjt}^{k}), \quad k = 1,2,3,4$$

여기서 $k=1$인 경우가 상대편향(Relative Bias, RB)이고, $k=2$인 경우가 분산의 상대편차(Relative Deviation of Variance; RDV)이며, $k=3$인 경우가 왜도의 상대편차(Relative Deviation of Skewness; RDS), 그리고 $k=4$인 경우가 첨도의 상대편차(Relative Deviation of Kurtosis; RDK)이다. 상대편차가 0에 가까울수록 추정의 정확성은 높다고 할 수 있다.

# IV. 모의실험

## 1. 모집단 구성

본 절에서는 앞에서 설명한 항목무응답 대체방법을 한국복지패널조사에 적용하여 모의실험을 실시하였다. 서론에서 언급한 바와 같이 한국복지패널의 총 패널가구는 7,267가구이다. 그리고 조사의 목적상 저소득 가구의 추출률이 일반가구의 추출률보다 더 크게 적용되었다. 본 논문에서는 복지패널 표본을 모집단으로 간주하고 부차 표본을 뽑아 무응답 대체방법을 비교하고자 한다. 이런 목적으로 완전응답을 한 패널가구 중 3,370 가구를 뽑아 모의실험 모집단으로 하였다. 그리고 패널가구의 3개년 가처분 소득을 관심변수로 하였고 가구 총비용을 보조변수로 하였다. 일반가구와 저소득 가구에서 구한 기초통계량이 〈표 2〉에 주어져 있다.

## 2. 모의실험 방법

### 1) 표본선정

모집단을 일반가구 층과 저소득가구 층으로 구분한 후 각 층에서 표본을 선정하는데, 표본선정 방법은 단순확률추출(SRS)과 확률비례추출(PPS)를 고려한다. 한국복지패널은 횡단면 가중치 및 종단면 가중치를 이용자에게 제공하고 있다. 본 모의실험에서는 1차년도 횡단면 가중치를 확률비례표본 추출에 이용하기로 한다.

**〈표 2〉 모의실험 모집단의 기초통계량**

(단위: 가구 수, 만원)

| 구 분 | 빈 도 | 항 목 | 평 균 | 표준편차 | 최소값 | 최대값 |
|---|---|---|---|---|---|---|
| 일반가구 | 2,364 | 총비용 | 218.85 | 96.87 | 40.0 | 595 |
| | | 2005년 소득 | 2,996.07 | 1,354.15 | 880.8 | 8,316 |
| | | 2006년 소득 | 3,256.16 | 1,629.15 | -156.0 | 8,914 |
| | | 2007년 소득 | 3,493.54 | 1,771.97 | -466.0 | 9,673 |
| 저소득가구 | 1,006 | 총비용 | 71.45 | 42.43 | 10.3 | 250 |
| | | 2005년 소득 | 726.07 | 405.42 | -156.0 | 2,098 |
| | | 2006년 소득 | 1,014.61 | 646.01 | -84.0 | 2,988 |
| | | 2007년 소득 | 1,156.89 | 795.51 | -48.8 | 4,310 |

표본수는 100개, 200개, 300개를 고려한다. 각 층별로 50개, 100개, 150개의 표본을 선정한다.

### 2) 무응답 발생

무응답을 발생시키기 위하여 두 가지 응답 메커니즘을 고려한다. 첫 번째는 완전확률응답(Missing Completely At Random: MCAR) 메커니즘으로 무응답이 완전히 랜덤하게 발생하는 경우이다. 응답을 나타내는 변수를 $R$로 표현하고, 응답은 $R=1$, 무응답은 $R=0$로 표현한다고 하자. 그러면 첫 번째 완전확률응답 메카니즘은 다음과 같이 된다.

$$\Pr(R_g = 1) = c_g, \quad g = 1,2$$

본 모의실험에서는 일반가구 층과 저소득 가구 층을 대체군으로 하였고, 각 대체군에서 응답확률은 동일하게 한다.

두 번째는 응답확률이 조사변수에 연관된 경우로 일반가구에서는 고소득 가구의 무응답률을 높게 하고 저소득가구에서는 저소득 가구의 무응답률을 높게 한다. 두 번째 응답 메커니즘인 비확률 무응답(Not-Missing at Random: NMAR)을 표현하면 다음과 같다.

$$\Pr(R_g = 1|y,x) \propto \frac{1}{1+\exp(-\alpha_g - \beta_g x - \gamma_g y)}, \quad g = 1,2$$

즉, 응답이 조사변수 $y$와 보조변수 $x$에 영향을 받는 경우이다.

응답률은 패널조사임을 고려하여 1차년도는 94%, 2차년도는 88%, 3차년도는 82%의 응답을 한 경우를 고려한다.

### 3) 무응답 대체 및 반복수

무응답을 대체하기 위한 횡단면 대체법 중에서는 (i) 평균대체 (ii) 핫덱 대체 (iii) 회귀대체 (iv) 비대체 (v) 최근방 대체를 고려하고, 웨이브 무응답 대체법 중에서는 (vi) 종단면 회귀대체 (vii) 이월대체 (viii) 최근방 회귀대체 그리고 (ix) 행렬대체법을 고려한다.

### 4) 대체방법 평가

대체방법의 평가를 위하여 예측 정확성 지표와 추정 정확성 지표를 사용한다. 예측 정확성 지표로는 상대평균절대편차(RMAD), 상대제곱근평균제곱편차(RRMSD)을 사용하고, 추정 정확성 지표로는 상대편향(RB), 분산의 상대편차 (RDV), 왜도의 상대편차 (RDS), 첨도의 상대편차 (RDK)를 사용한다.

**〈표 3〉 대체방법의 정확성 지표값**

(표본 수 200, 3차년도 변수, 무응답률 18%)

| 표집방법 | 응답패턴 | 대체방법 | 예측 정확성 | | 추정 정확성 | | | |
|---|---|---|---|---|---|---|---|---|
| | | | RMAD ($\times 10^{-2}$) | RRMSD ($\times 10^{-1}$) | RB ($\times 10^{-3}$) | RDV ($\times 10$) | RDS ($\times 10^{5}$) | RDK ($\times 10^{9}$) |
| SRS | MCAR | 평균대체 | 6.463 | 2.058 | 0.464 | -11.625 | -12.290 | 10.114 |
| | | 회귀대체 | 4.573 | 1.551 | -0.074 | -6.488 | -6.888 | -5.900 |
| | | 비대체 | 4.617 | 1.596 | -0.409 | -3.013 | -2.926 | -2.425 |
| | | 핫덱대체 | 8.980 | 2.930 | 0.009 | -0.064 | -0.046 | -0.023 |
| | | 최근방대체 | 6.404 | 2.156 | -1.832 | -1.321 | -0.984 | -0.776 |
| | | 종단면회귀 | 4.587 | 1.533 | 0.087 | -7.461 | -7.792 | -6.589 |
| | | 이월대체 | 5.080 | 1.795 | -23.894 | -17.329 | -11.853 | -8.255 |
| | | 최근방회귀 | 6.391 | 2.152 | -1.553 | -1.034 | -0.740 | -0.573 |
| | | 행렬대체 | 5.135 | 1.765 | 0.067 | -0.853 | -1.021 | -0.261 |
| SRS | NMAR | 평균대체 | 6.409 | 2.056 | -1.632 | -13.849 | -13.950 | -11.281 |
| | | 회귀대체 | 4.601 | 1.567 | -0.275 | -6.921 | -7.348 | -6.320 |
| | | 비대체 | 4.677 | 1.623 | 0.245 | -2.666 | -2.755 | -2.329 |
| | | 핫덱대체 | 8.964 | 2.928 | -4.114 | -3.566 | -2.538 | -1.772 |
| | | 최근방대체 | 6.406 | 2.159 | -1.202 | -1.275 | -1.083 | -0.886 |
| | | 종단면회귀 | 4.609 | 1.540 | -0.116 | -8.065 | -8.437 | -7.165 |
| | | 이월대체 | 5.101 | 1.802 | -24.216 | -18.064 | -12.570 | -8.844 |
| | | 최근방회귀 | 6.395 | 2.156 | -0.999 | -1.158 | -0.991 | -0.811 |
| | | 행렬대체 | 5.207 | 1.817 | -0.204 | -0.784 | -0.278 | 2.935 |
| PPS | MCAR | 평균대체 | 7.045 | 2.195 | -0.144 | -13.418 | -14.705 | -12.627 |
| | | 회귀대체 | 4.921 | 1.659 | -0.599 | -7.724 | -8.502 | -7.582 |
| | | 비대체 | 4.965 | 1.711 | -0.781 | -3.741 | -3.750 | -3.205 |
| | | 핫덱대체 | 9.727 | 3.097 | 0.836 | 0.893 | 0.864 | 0.793 |
| | | 최근방대체 | 6.746 | 2.259 | -1.065 | -1.126 | -1.079 | -0.974 |
| | | 종단면회귀 | 4.929 | 1.630 | -0.291 | -8.582 | -9.325 | -8.238 |
| | | 이월대체 | 5.541 | 1.929 | -27.146 | -20.493 | -14.677 | -10.708 |
| | | 최근방회귀 | 6.804 | 2.284 | -0.923 | -0.907 | -0.934 | -0.911 |
| | | 행렬대체 | 5.495 | 1.867 | -0.425 | -1.452 | -1.390 | -1.113 |
| PPS | NMAR | 평균대체 | 6.716 | 2.111 | -2.730 | -16.377 | -16.767 | -13.963 |
| | | 회귀대체 | 4.775 | 1.602 | -1.070 | -7.830 | -8.345 | -7.363 |
| | | 비대체 | 4.871 | 1.665 | -0.068 | -2.485 | -2.239 | -1.689 |
| | | 핫덱대체 | 9.697 | 3.105 | -3.138 | -2.238 | -1.169 | -0.405 |
| | | 최근방대체 | 6.746 | 2.263 | -3.885 | -2.242 | -0.954 | -0.322 |
| | | 종단면회귀 | 4.950 | 1.603 | -1.838 | -10.336 | -10.728 | -9.301 |
| | | 이월대체 | 5.539 | 1.929 | -28.729 | -22.365 | -16.174 | -11.784 |
| | | 최근방회귀 | 6.753 | 2.270 | -3.515 | -2.337 | -1.233 | -0.662 |
| | | 행렬대체 | 5.605 | 1.861 | -1.892 | -2.262 | -1.799 | -1.295 |

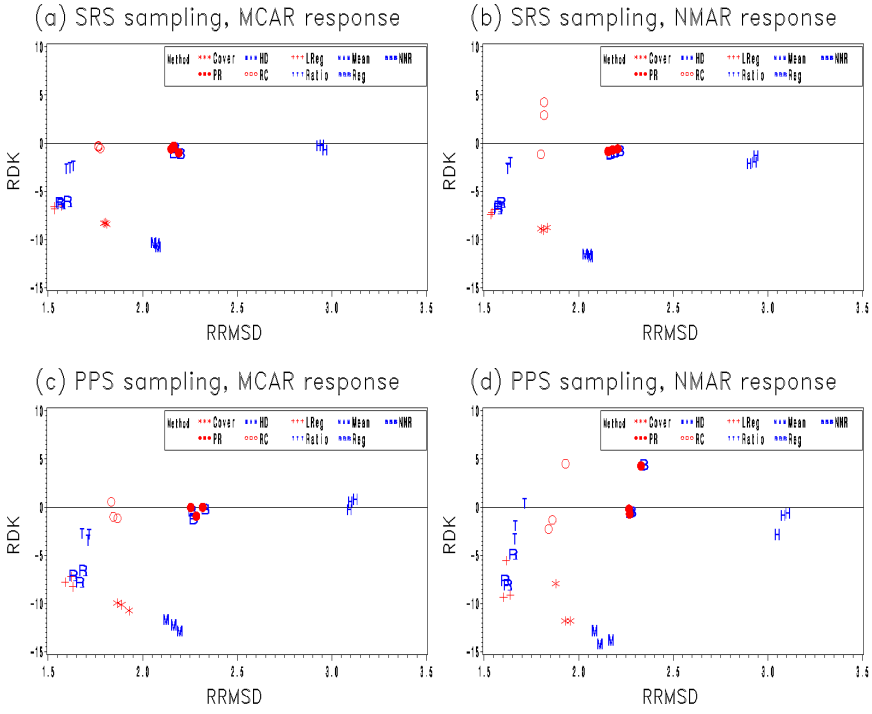
주1) SRS : 단순확률표집, PPS : 확률비례표집

주2) MCAR : 완전확률무응답 패턴, NMAR : 비확률무응답 패턴

모의실험은 세 종류의 표본수($n = 100, 200, 300$), 두 종류의 표본추출 방법(SRS, PPS), 2종류의 응답 메카니즘(MCAR, NMAR), 세 종류의 무응답확률(1차년도는 6%, 2차년도는 12%, 3차년도는 18%)별로 9개 대체방법을 적용한 후 예측 정확성 지표 2개와 추정 정확성 지표 4개를 계산한다. 그리고 이를 1,000회 반복하여 1,000회 평균을 구한다.

## 3. 모의실험 결과

모의실험 결과 중 표본 수가 200인 경우의 3차년도 변수에 대한 6개 지표값이 〈표3〉에 주어져 있다. 표본수가 100인 경우와 300인 경우는 표본 수가 200인 경우와 유사한 경향을 보이기 때문에 결과 제시는 생략하기로 한다. 또한 1차년도 변수와 2차년도 변수에 대해서도 무응답률이 다르므로(1차년도 6%, 2차년도 12%, 3차년도 18%) 지표값의 단위는 다르지만 경향이 비슷하게 나타났기 때문에 결과 제시는 생략한다.




〈그림1〉 평가지표 산점도

(대체방법: *이월대체, H 핫덱대체, +종단면 회귀대체, M 평균대체, B 최근방대체, ·최근방 회귀대체, ◦ 행렬대체, T 비대체, R 회귀대체)

모의실험 결과를 요약하면 다음과 같다.

예측 정확성 측면에서는 비대체법, 회귀대체법, 종단면 회귀대체법이 가장 우수한 것으로 나타났고, 이월 대체법, 행렬 대체법이 그 다음으로 좋게 나타났다. 그 다음은 평균대체법, 최근방 회귀대체법, 최근방 대체법이 좋게 나타났고, 핫덱 대체법이 가장 좋지 않은 것으로 나타났다 (<표 3>, <그림1> 참조). 예측 정확성이 가장 좋게 나타난 세 방법은 모두 전년도 자료와 현재 연도 자료의 선형성에 근거하는 대체법으로, 패널조사의 속성에 부합하여 좋은 성능을 보이는 것으로 해석할 수 있다. 핫덱 대체, 최근방 대체, 평균대체법은 전년도 자료를 이용하지 않는 횡단면 방법으로 전년도 자료를 이용하는 방법에 비하여 성능이 좋지 않았다.

추정 정확성 측면에서는 비대체, 최근방 대체, 최근방 회귀대체, 행렬대체, 핫덱 대체가 우수한 것으로 나타났고, 그 다음으로 회귀대체와 종단면 회귀대체가 성능이 좋은 것으로 나타났으며, 이월대체와 평균대체는 성능이 가장 좋지 않은 것으로 나타났다. 전년도 자료를 이용하는 최근방 회귀대체법, 행렬대체법은 추정의 정확성이 높은 반면 이월대체법은 상대적으로 정확성이 높지 않게 나타났다. 이는 비록 패널조사에서 조사자료의 연도별 상관성이 높기는 하지만 이월대체처럼 전년도 데이터를 그대로 사용하는 것은 효과적인 방법이 아님을 말해 준다. 핫덱 대체는 전년도 자료를 전혀 이용하지 않음에도 불구하고 추정의 정확성이 매우 우수한 것으로 나타났다. 그 이유는 핫덱 표본이 응답자 중에서 다시 뽑은 재표본(replicate)으로 응답자를 대표하므로 추정의 정확성이 높게 나타난 것이다(<그림1>의 (a)와 (c)). 그런데 <그림1>의 (b)와 (d)에서 볼 수 있듯이 응답 메카니즘이 NMAR이면 핫덱 방법의 RDK 지표값은 다소 커지는 경향이 있다. 즉, 조사변수와 연관된 무응답이 발생하면 핫덱 표본은 치우친 응답표본만을 대표하므로 추정의 정확성은 저하되게 된다.

본 모의실험은 표본추출방법과 응답 메커니즘에 따른 대체방법의 차이를 확인하기 위하여 표본추출 방법 2가지와 응답 메커니즘 두 가지를 조합하여 총 4종류의 서로 다른 응답을 얻은 후 대체방법을 적용하였다. 네 가지 조합에 따른 비교를 <그림1>에서 쉽게 할 수 있다. <표 1>에서 보면 9개 대체 방법은 표본추출방법이나 무응답 발생 메커니즘에 영향을 받기는 하지만 상호 비교에 있어서 그 패턴은 크게 변하지 않았다. 즉, 6개 정확성 지표에 대하여 다소의 크기 차이는 있지만 그 우열은 크게 변하지 않는다. 조금 구체적으로 살펴보면, 행렬대체법은 표본추출방법에 다소 영향을 받기는 하지만 응답 패턴에 더 큰 영향을 받고, 최근방 대체와 최근방 회귀대체, 핫덱 대체, 이월대체는 표본추출방법과 응답 패턴에 동시에 영향을 받는다. 비대체. 회귀대체, 종단면 회귀대체, 평균대체 등은 표본추출

방법 및 응답 패턴에 영향을 받기는 하지만 그 정도는 크지 않다.

대체방법별로 보면, 비대체와 행렬대체는 예측 정확성과 추정 정확성이 모두 우수하게 나타났고, 회귀대체, 종단면 회귀대체, 이월대체는 예측 정확성은 우수한 반면 추정 정확성은 다소 떨어지는 것으로 나타났다. 반면 최근방 대체, 최근방 회귀대체, 핫덱 대체는 예측 정확성은  떨어지나 추정 정확성은 우수한 것으로 나타났다. 마지막으로 평균대체는 두 정확성 모두 좋지 않은 것으로 나타났다.

## V. 결론

본 논문에서는 패널조사에서 활용 가능한 웨이브 무응답 대체법을 고찰하고 모의실험을 통하여 각 방법의 성능을 비교하였다. 대체방법의 성능은 예측 정확성 지표와 추정 정확성 지표를 사용하여 수치화 하였다. 본 연구에서 고찰한 웨이브 무응답 대체방법은 종단면 회귀대체, 이월대체, 최근방 회귀대체, 행렬대체이고 비교대상으로 삼은 횡단면 무응답 대체법은 평균대체, 회귀대체, 비대체, 최근방 대체, 핫덱 대체를 고려하였다. 모의실험 결과, 비대체와 행렬대체는 두 정확성이 모두 우수하게 나타났다. 회귀대체, 종단면 회귀대체, 이월 대체는 예측 정확성만 우수했으며 최근방 대체, 최근방 회귀대체, 핫덱 대체는 추정 정확성만 우수했다. 마지막으로 평균대체는 두 정확성 모두 좋지 않은 것으로 나타났다.

본 논문에서 얻은 모의실험 결과는 한국복지패널 데이터로부터 얻은 경험적인 결과이기 때문에 이를 일반화하기는 어렵다. 그러나 많은 패널조사의 속성상 동일 조사항목에 대한 조사변수는 시점 간에 매우 밀접한 상관성이 있기 때문에, 비록 조사는 다를지라도 본 논문의 결과는 다른 조사에서 웨이브 무응답 대체방법을 평가하고 선정하는데 좋은 참고가 되리라고 생각된다. 대체방법에 대한 이론적인 검토와 더 일반화된 대체방법의 비교는 향후 연구로 남긴다.

### 참고문헌

김규성. 2000. “무응답 대체 방법과 대체 효과.”《조사연구》1(2): 1-14.

김규성 · 이기재 · 김진. 2005. “농어가경제조사에서 가중핫덱 무응답 대체법의 활용.”《응용통계연구》18(2): 311-328.

김영원 · 조선경. 1996. “표본조사에서 항목 무응답 대체 방법.”《한국통계학회논문집》3(3): 145-159.

조영숙 · 천영민 · 황대용. 2008. “농촌생활지표조사에서 무응답 대체: 사례.”《응용통계연구》21(1): 95-107.

한국보건사회연구원. 2008.《한국복지패널 3차년도 조사자료: User's Guide》.

Chambers, R. 2000. “Evaluation Criteria for Statistical Editing and Imputation.” *Working Paper for the Euroedit Project on the Development and Evaluation of New Methods for Editing and Imputation*. University of Southampton: UK.

Frick, J. R. and Grabka, M. M. 2007. “Item Non-response and Imputation of Annual Labor Income in Panel Surveys from a Cross-national Perspective.” *SOEP papers on multidisplinary panel data research*. No. 49. Berlin: DIW.

Little, R. A. 1988. “Missing-data Adjustemnts in Large Surveys.” *Journal of Business and Economic Statistics* 6: 287-296.

Little, R. J .A. and Rubin, D. B. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley.

Little, R. J. A. and Su, H. L. 1989. “Item Non-response in Panel Surveys.” In D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh.(Eds.), *Panel Surveys*. New York: Wiley.

Starick, R. 2005. “Imputation in Longitudinal Surveys: The Case of HILDA.” *Research Paper*. Australian Bureau of Statistics.

Starick, R. and Watson, N. 2007. “Evaluation of Alternative Income Imputation Methods for the HILDA Survey.” *HILDA Project Discussion Paper Series* No. 1/07. June 2007.

Taylor, M.F. (Ed.). with Brice, J., Buck, N. and Prentice-Lane, E. 2007. *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex.

U.S. Bureau of the Census. 2001. *Survey of Income and Program Participation User's Guide*.

Watson, N. 2004. “Income and Wealth Imputation for Wave 1 and 2.” *HILDA Project Technical Paper Series* No. 3/04 July 2004. University of Melbourne.

Watson, N. (Ed.). 2008. “*HILDA User Manual-Release* 6.” Melbourne Institute of Applied Economic and Social Research. University of Melbourne.

[접수 2010/1/30, 수정 2010/2/20, 게재확정 2010/2/26]