연구논문

# 순환표본의 결합을 위한 가중치 산출에 대한 연구*

A Study on the Construction of Weights for Combined Rolling Samples

송종호** · 박진우*** · 변종석**** · 박민규*****
Jongho Song · Jinwoo Park · Jongseok Byun · Mingue Park

순환표본조사를 시행할 경우 매 순환주기별로 적절한 통계적 신뢰도를 가진 전체 모집단 특성이 추정될 수 있는 반면에, 작은 표본크기로 인하여 통계적 신뢰도가 높은 소지역 추정량의 산출은 어렵다. 따라서 소지역 추정량은 일반적으로 일정 주기 후 혹은 전체조사가 마무리된 후 독립적인 순환표본들을 결합하여 얻어진 최종표본을 통해 산출된다. 본 연구에서는 순환표본을 결합하여 추정량을 만들 때 필요한 가중치 산출의 문제를 고려하였다. 기존의 연구들이 각 조사에 따른 경험을 바탕으로 조사별로 가능한 순환표본 결합 가중치를 정의하였으나, 본 연구에서는 모든 가능한 관심변수에 적용 가능하도록 표본설계변수에만 의존하는 모형을 설정하고 주어진 모형하에서의 최량선형불편예측치(Best Linear Unbiased Predictor: BLUP)를 고려하였다. 모의실험을 통하여 각 모형 하에서 정의되는 여러 BLUP을 비교하여 모형변화에 강건한 추정량을 제안하고 그 결과를 제4기 국민건강영양조사에 적용하였다.

**주제어**: 선형모형, 가중치, BLUP, 국민건강영양조사

Although it is possible to provide statistically reliable estimators of the entire population parameters based on each independent rolling sample, estimators of the small areas may not have the required statistical efficiency. Thus, in general, small area estimators are calculated based on the combined rolling sample after entire rolling sample survey is finished. In this study, we considered the construction of weights that

* 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2009−0074328)

** 고려대학교 통계학과 박사과정

*** 수원대학교 통계정보학과 교수

**** 한신대학교 정보통계학과 교수

***** 교신저자(corresponding author) : 고려대학교 통계학과 부교수 박민규.
E−mail : mpark2@korea.ac.kr

is necessary in the analysis of the combined rolling sample. Unlike the past studies that provided the empirical results for the corresponding specific rolling sample survey, we considered linear models that depends only on design variables and rolling period and provided the corresponding Best Linear Unbiased Predictor(BLUP). Through a simulation study, we proposed the estimators for the population parameters that are robust to model failure and the BLUP under the assumed model. The results are applied to the 4th Korea National Health and Nutrition Examination Survey.

**Key words** : linear model, weight, BLUP, National Health and Nutrition Examination Survey

## Ⅰ. 서론

표본조사(sample survey)의 중요한 목적 중의 하나는 표본에서 조사된 정보를 이용하여 관심대상인 모집단의 특성을 추론하는 것이다. 센서스나 표본크기가 큰 대규모 조사에서는 표집오차(sampling error)가 없거나 크지 않은 반면, 비표집오차(non−sampling error)가 커질 가능성이 크기 때문에 큰 표본크기만을 고집하는 것은 바람직하지 못하다. 또한 일반적으로 표본크기가 커질수록 조사기간이 길어지며 이로 인하여 조사시점과 자료제공시점 사이에 차이가 발생하는 문제가 야기될 수 있다. 따라서 성공적인 표본조사를 위해서는 모집단의 분포 및 특성을 대표할 수 있는 표본추출뿐 아니라 모집단의 정보를 제공하는 시점도 함께 고려되어야 할 것이다. 순환표본조사(rolling sample survey)는 대규모의 표본조사에서 발생하는 시간 및 비용 상의 현실적인 제약과 자료제공시점과 조사시점과의 불일치 문제를 해결하기 위해 흔히 사용된다.

순환표본조사는 크기가 전체인구의 1/F인 k개의 독립적인 확률표본을 구성하여 정해진 각 시점별로 독립적인 k 표본의 조사를 진행하는 방법이다. 각 확률표본은 어떤 큰 지역에 대한 추정량을 제공할 수 있어야 하며, 큰 지역에 대한 더욱 정확한 추정량이나 소지역의 추정량을 얻기 위해서는 연속적으로 조사되는 확률표본을 누적하여 사용할 수 있다. 특별히 k=F인 경우를 “순환센서스(rolling census)”라 부른다. 여기서 지역은 지리적 지역과 인구학적인 그룹을 모두 포함하는 개념이다. 각 조사시점별로 얻어진 자료의 누적을 통하

여 새로운 표본을 구성하고 이를 이용하여 모집단에 대한 분석을 실시하는 순환표본조사 또는 순환센서스에 대한 자세한 내용은 Kish(1979; 1990; 1998; 1999)를 참조하면 된다.

본 연구에서는 외국의 두 가지 순환표본조사 사례를 중심으로 선행연구에서 적용된 순환표본의 결합 및 누적에 관한 여러 가지 접근방법을 소개할 것이다. 그리고 조사된 대부분의 변수에 적용 가능한 보편적인 결합가중치를 유도하기 위하여 순환표본 조사시점과 표본설계변수만을 고려한 선형모형들을 가정하고 각 모형 하에서 최량선형불편예측치(Best Linear Unbiased Predictor : BLUP)를 정의한다. 모의실험을 통하여 각 모형 하에서 정의된 결합가중치를 비교하고 이를 질병관리본부에서 실시하고 있는 순환조사인 제4기 국민건강영양조사의 2007년과 2008년 순환표본의 결합에 적용하여 분석하고자 한다.

## Ⅱ. 외국의 순환표본조사 사례

### 1. 미국 지역사회조사(American Community Surveys: ACS)

미국 지역사회조사는 경제적, 사회적 그리고 미국 전역 공동체의 가구 특성에 대한 기본적인 정보를 제공하기 위하여 미국 인구조사국(U.S. Bureau of the Census)에서 수행하는 순환표본조사이다. 미국 지역사회조사의 주된 목적은 다음의 두 가지이다. 첫 번째 목적은 기존의 센서스 long form 조사로부터 얻어진 추정치와 유사한 평균제곱오차(Mean Squared Error : MSE)를 가지면서 모든 지역에 대한 정보를 10년 주기로 제공하는 것이다. 두 번째 목적은 지역에 대한 변화하는 정보를 얻기 위하여 주어진 관심 지역의 추정치를 매년 제공하는 것이다(US Census Bureau 2006).

미국 지역사회조사의 방법론적 기초를 이루는 순환표본의 아이디어는 Leslie Kish가 1970년대 말에 소개한 연속측정(continuous measurement)의 개념에서 출발한다(Alexander 1998; 2002). 인구조사국은 순환표본조사를 센서스 long form의 대체후보로 간주하고 연구를 진행하였으며 이 과정을 통하여 Leslies Kish가 제안한 “연속측정”의 개념이 미국 지역사회조사라 불리는 대규모의 표본조사로 발전하였다. 미국 지역사회조사와 관련된 국내의 연구결과는 전광희(2008)와 김규성(2009)을 참조하면 된다.

미국 지역사회조사 표본은 인구조사국에서 지원하는 Master Address File(MAF)를 추출틀로 사용하여 2단계 집락추출법(two−stage cluster sampling)을 통해 추출된다. 1단계

추출에서는 MAF로부터 얻어진 전체 주소에서 20%를 표본으로 추출하고, 남은 80%는 4개의 동일한 그룹에 할당한다. 5개로 분할된 그룹을 순서화하고 각각을 매년 순환한다. 그리고 2단계에서 해당년도의 그룹 안에서 소규모 지역에 대한 추론이 가능하도록 층별 추출확률을 다르게 한 층화추출 방법을 적용한다(Hefter & Williams 2008). 데이터의 수집은 메일, 전화, 방문의 연속적인 3가지 방법을 이용하여 3개월에 걸쳐 수행한다.

Kish(1998; 1999)는 미국 지역사회조사에서 연 단위의 순환표본을 결합할 때 사용되는 가중치의 산출을 위해 여러 가지 방법을 제안하였다. Kish는 계절성이나 우연한 변동이 아닌 장기적인 변동을 고려하기 위해 연 단위 순환표본의 가중평균을 통하여 10년의 순환표본을 결합하는 것을 고려하였다. 연 단위 순환표본의 평균을 $\bar{y}_i$, 해당가중치를 $r_i$라고 하면 10년 간의 순환표본의 결합을 통한 평균은 $\bar{y}_t = \sum_{i=1}^{10} r_i \bar{y}_i$ 으로 표현되며, $\sum_{i=1}^{10} r_i = 1$이다. 미국 지역사회조사의 연 단위의 순환표본의 결합을 위해 Kish(1998; 1999)와 Alexander(2002)는 첫째로, 마지막 년도에 모든 가중치를 부여하는 방법($r_{10} = 1$) 둘째로, 매년 같은 가중치(equal weight)를 부여하는 방법($r_i = 0.1$) 셋째로, 최신의 순환표본에 더 많은 가중치를 부여하는 방법($r_1 \leq r_2 \leq \cdots \leq r_{10}$) 등을 제시하였다.

## 2. Health Care Survey of Department of Defense Beneficiaries(HCSDB)

HCSDB는 미국 국방부에서 현역 또는 은퇴군인과 그 가족에게 제공되는 의료서비스인 Military Health System(MHS)을 관리하기 위해 실시되는 연 단위의 조사이다. 1995년 처음 실시되어 2001년 이후의 조사에서는 각 분기별로 독립적인 조사가 진행되는 순환표본조사로 하고 있다. HCSDB의 분기별 조사는 MHS 수혜자들에 대한 최신의 정보를 보다 자주 제공해 주는 최초의 도구가 되었다. 또한 MHS에 대한 정보를 수집하는 많은 조사들 중에서 HCSDB만이 유일하게 과거의 12개월 동안 전 세계에 있는 MHS 수혜자의 건강관리경험을 측정하고 있다. 각 분기별 자료를 결합하여 연 단위의 자료를 생산함과 동시에 각 분기별 조사에 추가 질의응답을 포함하여 분기별 자료를 생산한다. HCSDB의 가장 유용한 특징 중 하나는 매년 조금씩 변하는 핵심 질문과 각 분기별로 변화하는 보조 질문을 결합하여 사용한다는 것이다. 핵심 질문을 통해서는 시간에 따라 변화하는 적용범위, 적용대상, 만족도 등의 경향성을 파악할 수 있으며 보조 질문들을 통해서는 조사 이용자의 변화 패턴을 파악할 수 있다. 각 연도별 HCSDB의 조사방법과 결과는 MHS 홈페이지인 'http://www.tricare.mil/survey/hcsurvey/'를 통해 제공되고 있다.

2001 HCSDB 분기별 조사에서의 표본은 층화추출방법을 사용하여 추출하였다. 각 층은 126개의 군수용 지역과 6개의 등록수혜자 그룹으로 구성되었다. 조사는 분기당 독립적으로 선택된 45,000명의 MHS 수혜자를 대상으로 메일 조사를 통하여 실시된다. 각 분기별 자료에 가중치가 부여되며 4분기의 조사가 모두 종료된 후 각 분기의 자료를 누적한다. 그것으로 하나의 결합자료를 생성하여 소규모 지역에 대한 추정에 있어 더욱 신뢰할 수 있는 정보를 제공하고 있다.

Friedman(2003)은 HCSDB에서 얻어지는 분기별 순환표본의 여러 가지 가능한 결합방법을 비교하였다. 순환표본의 결합을 위해 고려된 가중치 부여방법은 분기마다 동일한 가중치를 부여하는 방법과 가장 최근의 순환표본에 더 큰 가중치를 부여하는 방법, 그리고 특정 지역을 추정하기 위해 각 분기별 지역의 크기의 비율을 고려한 방법이다. 고려된 순환표본의 결합방법들은 추정량의 상대오차를 통하여 추정량을 비교하였다. 상대오차로는 $RE\_P = (Y_t - Y_{equal})/Y_{equal}$ 가 사용되었다. 분기별 평균의 변화 양상이 뚜렷한 경우에는 최근의 순환표본에 더 큰 가중치를 부여하는 방법이 더 좋은 성능을 보여 주었으나, 분기별 변화 양상이 없는 경우에는 고려한 결합방법들이 비슷한 성능을 보였다.

## Ⅲ. 순환표본의 결합을 위한 가중치 유도

선행연구에서 고려된 순환표본조사 자료의 결합에서는 각 순환표본들로부터 얻은 각각의 통계량에 가중치를 부여하여 그 가중합을 유도하는 방법들을 제시하고 있다. 이는 일반적으로 각 시점별 통계량들이 선형추정량임을 감안할 때 기존 연구들은 각 시점별로 계산된 가중치에 일정한 상수를 곱하여 주는 가장 단순한 형태의 가중치 변환이라고 간주할 수 있다. 본 연구에서는 가능한 모든 변수에 적용 가능한 보편적인 결합가중치를 유도하기 위하여 순환표본 조사시점과 표본설계변수만을 이용한 선형모형들을 고려하여 각 모형 하에서 BLUP를 정의하고 이를 비교하고자 한다.

순환표본들의 결합을 위한 가중치의 유도에 앞서 각 순환표본 조사시점별 모평균과 모집단 총합의 추정량을 정의하자. 순환표본 조사시점 $t$에서 $h$번째 층에 속한 $i$번째 관측치의 관심변수 값을 $y_{thi}$, 모집단 총합 추정을 위해 관측치에 부여되는 가중치를 $w_{thi}$로 정의하자. 각 순환표본 조사시점별 관심변수 $Y$의 모집단 총합과 모평균의 선형추정량을 각각

$$\hat{t}_t = \sum_h \sum_i w_{thi} y_{thi} \tag{1}$$

$$\hat{\bar{y}}_t = \frac{\sum_h \sum_i w_{thi} y_{thi}}{\sum_{h'} \sum_{i'} w_{th'i'}} = : \sum_h \sum_i \alpha_{thi} y_{thi} \tag{2}$$

로 정의하자. 여기서 $\alpha_{thi} = \dfrac{w_{thi}}{\sum_{h'} \sum_{i'} w_{th'i'}}$ 이다.

각 시점의 순환표본을 결합하여 얻은 전체 표본을 이용하여 정의할 수 있는 가장 일반적인 추정량은 각 관측치에 부여된 가중치인 (1)의 $w_{thi}$ 또는 (2)의 $\alpha_{thi}$를 개별적으로 보정하여 얻어지는 추정량으로 아래와 같이 표현할 수 있다.

$$\hat{\bar{y}}^{\,\mathrm{I}} = \sum_t \sum_h \sum_i a_{thi} \alpha_{thi} y_{thi} \tag{3}$$

여기서 $\alpha_{thi}$는 (2)의 가중치를 나타내며 (3)을 정의하기 위하여 사용되는 가중치 $a_{thi}$는 변수 $Y$에 대한 적절한 분포의 가정, 표본추출법 그리고 사용자의 편의성 등의 문제에 의존하여 결정된다. 각 관측치에 각각 다른 보정치 $a_{thi}$를 적용하는 경우, 각 변수에 따라 적절한 모형의 정의가 필요하게 되므로 극단의 경우 변수의 수만큼 가중치가 정의되는, 현실적으로 가능하지 않은 대안이 된다.

고려되는 모든 관심변수 $Y$에 대한 보편적인 모형을 설정하기 위하여 대부분의 대규모 조사에서 사용되는 층화추출법을 고려하자. 시점 $t$에 조사된 순환표본의 층 $h$의 모평균 추정량 $\hat{\bar{y}}_{th}$을 아래와 같이 정의하자.

$$\hat{\bar{y}}_{th} = \frac{\sum_{(t'h'i') \in s} w_{t'h'i'}\, y_{t'h'i'}\, d_{t'h'i'}}{\sum_{(t'h'i') \in s} w_{t'h'i'}\, d_{t'h'i'}} \tag{4}$$

여기서 $d_{t'h'i'} = \begin{cases} 1, & \text{if } (t'h'i') \in s_{th} \\ 0, & \text{if } (t'h'i') \notin s_{th} \end{cases}$, $\quad t = 1,2,\cdots,T, \;\; h = 1,\cdots,H$ 이다. 각 조사시점별 각 층으로부터 단순임의표본이 추출되었다면 (4)는 다음과 같이 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{th} = \frac{1}{n_{th}}\sum y_{t'h'i}$$

본 연구에서는 고려 변수에 따라 그 값이 다르게 결정될 수 있는 (3)의 $a_{thi}$대신 조사시점과 표본설계변수인 층에만 의존하는 추정량

$$\hat{\bar{y}}^{\text{II}} = \sum_t \sum_h a_{th}\hat{\bar{y}}_{th} \tag{5}$$

를 고려한다. 또한 정의된 $\hat{\bar{y}}_{th}$에 대한 적절한 모형을 설정하기 위하여 먼저 각 시점의 순환표본 $s_t$가 층화추출표본이고 각 시점에 사용되는 층화변수가 동일함을 가정하자. 일반적으로 순환표본추출은 각 시점별로 추출되기보다는 1차 조사시점에서 모두 추출됨으로 주어진 가정은 현실적으로 합당하다. 이러한 가정 하에서 $\hat{\bar{y}}_{th}$에 대한 가장 보편적인 모형은 다음과 같다.

$$\hat{\bar{y}}_{th} \sim \left(\mu_{th}\ ,\ \frac{\sigma_h^2}{n_{th}}\right) \tag{6}$$

여기서 $\sigma_h^2$은 순환표본의 조사 시기에 의존하지 않는 모수로서 각 층 내에서 시간이 경과함에 따라 분산구조는 변화하지 않는 경우를 가정하고 있다. 따라서 $\hat{\bar{y}}_{th}$의 분산은 각 시점별, 층별 표본크기 $n_{th}$에 따라 결정된다. 모든 관심 변수 $Y$에 적용할 수 있는 보편적인 모형을 가정하기 위하여 $\mu_{th}$가 순환표본 조사시점과 설계변수인 층에만 의존하는 8가지 모형들을 고려한다.

본 연구에서는 순환표본의 결합 방법들을 비교하기 위하여 전체 조사 기간인 $T$시점 동안의 모집단 평균을 관심 모수로 고려하였다. 고려되는 8가지 가능한 모형과 각 모형 하에서 전체 $T$시점 동안의 모집단 평균의 BLUP는 다음과 같다. 또한 정의된 각 추정량들은 선형추정량으로서 본 연구에서 고려하고 있는 추정량 (5)의 형태를 가지고 있다. 따라서 순환표본의 결합을 위한 가중치는 각 모형 하에서 유도된 $a_{th}$에 의하여 정의된다.

**[모형1]** $\mu_{th} = \bar{y}_N, \quad \sigma_h^2 = \sigma^2$

$\hat{\bar{y}}_{th}$의 평균이 시점과 층에 의존하지 않으며 분산이 동일한 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_N + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma^2}{n_{th}}\right) \tag{7}$$

주어진 모형 (7)에서 정의되는 $\bar{y}_N$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}^{\,\mathrm{I}}_{N,BLUP} = \frac{\sum_t \sum_h n_{th} \hat{\bar{y}}_{th}}{\sum_t \sum_h n_{th}} \tag{8}$$

따라서 [모형1]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$a_{th} = \frac{n_{th}}{\sum_{t'} \sum_{h'} n_{t'h'}}$$

**[모형2]** $\mu_{th} = \bar{y}_N, \quad \sigma_h^2$

$\hat{\bar{y}}_{th}$의 평균이 시점과 층에 의존하지 않으며 층별 분산이 서로 다른 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_N + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma_h^2}{n_{th}}\right) \tag{9}$$

주어진 모형 (9)에서 정의되는 $\bar{y}_N$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}^{\,\mathrm{II}}_{N,BLUP} = \frac{\sum_t \sum_h \left(n_{th}/\sigma_h^2\right) \hat{\bar{y}}_{th}}{\sum_t \sum_h \left(n_{th}/\sigma_h^2\right)} \tag{10}$$

따라서 [모형2]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$a_{th} = \frac{n_{th}/\sigma_h^2}{\sum_{t'} \sum_{h'} \left(n_{t'h'}/\sigma_{h'}^2\right)}$$

**[모형3]** $\mu_{th} = \bar{y}_{N,h}$, $\sigma_h^2 = \sigma^2$

$\hat{\bar{y}}_{th}$의 평균이 시점에 의존하지 않고 층에 대한 효과를 가지며 분산이 동일한 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,h} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma^2}{n_{th}}\right) \tag{11}$$

주어진 모형 (11)에서 정의되는 $\bar{y}_{N,h}$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}_{N,h} = \frac{\sum_t n_{th} \hat{\bar{y}}_{th}}{\sum_t n_{th}} \tag{12}$$

$\bar{y}_{N,h}$의 BLUP (12)를 통하여 $\bar{y}_N$의 BLUP를 구하면 다음과 같다.

$$\hat{\bar{y}}^{\text{III}}_{N,BLUP} = \sum_h \frac{N_h}{N} \hat{\bar{y}}_{N,h} = \sum_h \frac{N_h}{N} \sum_t \frac{n_{th}}{n_h} \hat{\bar{y}}_{th} \tag{13}$$

따라서 [모형3]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$a_{th} = \frac{\sum_{t'} N_{t'h}}{\left(\sum_{t'} \sum_{h'} N_{t'h'}\right)\left(\sum_{t'} n_{t'h}\right)} n_{th}$$

**[모형4]** $\mu_{th} = \bar{y}_{N,h}$, $\sigma_h^2$

$\hat{\bar{y}}_{th}$의 평균이 시점에 의존하지 않고 층에 대한 효과를 가지며 층별 분산이 서로 다른 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,h} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma_h^2}{n_{th}}\right) \tag{14}$$

주어진 모형 (14)에서 정의되는 $\bar{y}_{N,h}$와 $\bar{y}_N$의 BLUP는 각각 [모형3]의 (12), (13)이며 [모형4]에서 정의되는 효율적인 가중치 $a_{th}$는 [모형3]과 일치한다.

**[모형5]** $\mu_{th} = \bar{y}_{N,t}, \quad \sigma_h^2 = \sigma^2$

$\hat{\bar{y}}_{th}$의 평균이 층에 의존하지 않고 시점에 대한 효과를 가지며 분산이 동일한 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,t} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma^2}{n_{th}}\right) \tag{15}$$

주어진 모형 (15)에서 정의되는 $\bar{y}_{N,t}$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}_{N,t} = \frac{\sum_h n_{th} \hat{\bar{y}}_{th}}{n_t} \tag{16}$$

$\bar{y}_{N,t}$의 BLUP (16)을 통하여 $\bar{y}_N$의 BLUP를 구하면 다음과 같다.

$$\hat{\bar{y}}_{N,BLUP}^{\mathrm{IV}} = \sum_t \frac{N_t}{N} \hat{\bar{y}}_{N,t} = \sum_t \frac{N_t}{N} \sum_h \frac{n_{th}}{n_t} \hat{\bar{y}}_{th} \tag{17}$$

따라서 [모형5]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$a_{th} = \frac{\sum_{h'} N_{th'}}{\left(\sum_{t'} \sum_{h'} N_{t'h'}\right)\left(\sum_{h'} n_{th'}\right)} n_{th}$$

**[모형6]** $\mu_{th} = \bar{y}_{N,t}, \quad \sigma_h^2$

$\hat{\bar{y}}_{th}$의 평균이 층에 의존하지 않고 시점에 대한 효과를 가지며 분산이 서로 다른 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,t} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma_h^2}{n_{th}}\right) \tag{18}$$

주어진 모형 (18)에서 정의되는 $\bar{y}_{N,t}$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}_{N,t} = \frac{\sum_h \left(n_{th}/\sigma_h^2\right) \hat{\bar{y}}_{th}}{\sum_h \left(n_{th}/\sigma_h{}^2\right)} \tag{19}$$

$\bar{y}_{N,t}$의 BLUP (19)를 통하여 $\bar{y}_N$의 BLUP를 구하면 다음과 같다.

$$\hat{\bar{y}}^{\,\mathrm{V}}_{N,BLUP} = \sum_t \frac{N_t}{N}\hat{\bar{y}}_{N,t} = \sum_t \frac{N_t}{N}\sum_h \frac{n_{th}/\sigma_h^2}{\left(\sum_h n_{th}/{\sigma_h}^2\right)}\hat{\bar{y}}_{th} \tag{20}$$

따라서 [모형6]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$a_{th} = \frac{\sum_{h'} N_{th'}}{\left(\sum_{t'}\sum_{h'} N_{t'h'}\right)\left(\sum_{h'} n_{th'}/\sigma_{h'}^2\right)}\frac{n_{th}}{\sigma_h^2}$$

**[모형7]** $\mu_{th} = \bar{y}_{N,th}$, $\sigma_h^2 = \sigma^2$

$\hat{\bar{y}}_{th}$의 평균이 층과 시점의 교호효과를 가지며 분산이 동일한 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,th} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma^2}{n_{th}}\right) \tag{21}$$

주어진 모형 (21)에서 정의되는 $\bar{y}_{N,th}$의 BLUP는 다음과 같다.

$$\hat{\bar{y}}_{N,th} = \hat{\bar{y}}_{th} \tag{22}$$

$\bar{y}_{N,th}$의 BLUP (22)를 통하여 $\bar{y}_N$의 BLUP를 구하면 다음과 같다.

$$\hat{\bar{y}}^{\,\mathrm{VI}}_{N,BLUP} = \sum_t\sum_h \frac{N_{th}}{N}\hat{\bar{y}}_{N,th} = \sum_t\sum_h \frac{N_{th}}{N}\hat{\bar{y}}_{th} \tag{23}$$

따라서 [모형7]에서 정의되는 효율적인 가중치 $a_{th}$는 다음과 같다.

$$\alpha_{th} = \frac{N_{th}}{\sum_{t'}\sum_{h'} N_{t'h'}}$$

**[모형8]** $\mu_{th} = \bar{y}_{N,th}$, $\sigma_h^2$

$\hat{\bar{y}}_{th}$의 평균이 층과 시점의 교호효과를 가지며 분산이 서로 다른 경우로 다음의 선형모형으로 표현된다.

$$\hat{\bar{y}}_{th} = \bar{y}_{N,th} + \epsilon_{th}, \quad \epsilon_{th} \sim \left(0, \frac{\sigma_h^2}{n_{th}}\right) \tag{24}$$

주어진 모형 (24)에서 정의되는 $\bar{y}_{N,th}$와 $\bar{y}_N$의 BLUP는 각각 [모형7]의 (22), (23)이며 [모형8]에서 정의되는 효율적인 가중치 $a_{th}$는 [모형7]과 일치한다.

## Ⅳ. 모의실험

이 장에서는 Ⅲ장에서 살펴보았던 조사시점별, 층별 추정량인 $\hat{\bar{y}}_{th}$의 가능한 8가지 모형으로부터 제시된 추정량들을 모의실험을 통해 비교한다. 모의실험을 위한 모집단은 총 4개의 층으로 구성되어 있으며 2개 시점의 순환표본을 고려하였다. 이는 본 연구 결과가 실제 적용될 질병관리본부의 국민건강영양조사가 현재시점 기준 2개 연도 자료가 축적되었기 때문이다. 층별, 조사시점별 표본배분방법으로는 시점별 배분된 표본의 크기에 따라서 두 경우를 고려하였다. 또한 각 시점별, 층별 표본추출방법으로 단순임의추출법을 고려하였고 따라서 $\hat{\bar{y}}_{th}$은 단순평균 $\bar{y}_{th}$로 표현된다. Ⅲ장에서 제시된 8가지 모형에 대한 6가지 추정량을 2개 시점의 순환표본의 경우에 적용하면 다음과 같이 표현된다.

추정량Ⅰ : $\hat{\bar{y}}_N^{\mathrm{I}} = \sum_{t=1}^{2}\sum_{h=1}^{4}\frac{n_{th}}{n}\bar{y}_{th}$ (모형1 하에서의 BLUP)

추정량Ⅱ : $\hat{\bar{y}}_N^{\mathrm{II}} = \sum_{t=1}^{2}\sum_{h=1}^{4}\frac{n_{th}^*}{n^*}\bar{y}_{th}$ (모형2 하에서의 BLUP)

추정량Ⅲ : $\hat{\bar{y}}_N^{\mathrm{III}} = \sum_{h=1}^{4}\frac{N_h}{N}\sum_{t=1}^{2}\frac{n_{th}}{n_h}\bar{y}_{th}$ (모형3, 모형4 하에서의 BLUP)

추정량Ⅳ : $\hat{\bar{y}}_N^{\mathrm{IV}} = \sum_{t=1}^{2}\frac{N_t}{N}\sum_{h=1}^{4}\frac{n_{th}}{n_t}\bar{y}_{th}$ (모형5 하에서의 BLUP)

추정량Ⅴ : $\hat{\bar{y}}_N^{\mathrm{V}} = \sum_{t=1}^{2}\frac{N_t}{N}\sum_{h=1}^{4}\frac{n_{th}^*}{n_t^*}\bar{y}_{th}$ (모형6 하에서의 BLUP)

추정량Ⅵ : $\hat{\bar{y}}_N^{\mathrm{VI}} = \sum_{t=1}^{2}\sum_{h=1}^{4}\frac{N_{th}}{N}\bar{y}_{th}$ (모형7, 모형8 하에서의 BLUP)

여기서 $N=\sum_t\sum_h N_{th}$, $N_h=\sum_t N_{th}$, $N_t=\sum_h N_{th}$, $n=\sum_t\sum_h n_{th}$ $n_{h=}\sum_t n_{th}$, $n_{t=}\sum_h n_{th}$, $n_{th}^{*}=n_{th}/\sigma_h^2$, $n^{*}=\sum_t\sum_h n_{th}^{*}$, $n_t^{*}=\sum_t n_{th}^{*}$ 이다. 추정량Ⅱ와 추정량 V는 모집단의 층별 분산이 고려된 추정량으로, 실제 모집단의 분산은 일반적으로 주어지지 않으므로 표본으로부터 추정하여 사용하여야 한다. 모의실험을 위해서는 모집단의 분산이 알려져 있다고 가정하고 추정량을 정의하였다.

### 1. 모집단 생성 및 순환표본의 구성

2개 시점에 대한 순환표본을 구성하기 위하여 모집단을 4개의 층으로 구성하였으며 현실 상황에 더 적합하도록 각 층의 크기를 다르게 하였다. Ⅲ장에서 살펴본 8가지 모형을 모집단으로 구성하였으며 각 모형의 모수는 다음과 같이 정의하였다.

모형1: $\bar{y}_{th}=\bar{y}_N+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma^2/n_{th}\right),\ \bar{y}_N=0,\ \sigma^2=1.5^2$

모형2: $\bar{y}_{th}=\bar{y}_N+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma_h^2/n_{th}\right),\ \bar{y}_N=0,\ \sigma_{h=i}^2=1.5^i, i=1,..,4$

모형3: $\bar{y}_{th}=\bar{y}_{Nh}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma^2/n_{th}\right)$
$\bar{y}_{N,h=1}=5,\ \bar{y}_{N,h=2}=3,\ \bar{y}_{N,h=3}=-3,\ \bar{y}_{N,h=4}=-5,\ \sigma^2=1.5^2$

모형4: $\bar{y}_{th}=\bar{y}_{Nh}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma_h^2/n_{th}\right)$
$\bar{y}_{N,h=1}=5,\ \bar{y}_{N,h=2}=3,\ \bar{y}_{N,h=3}=-3,\ \bar{y}_{N,h=4}=-5$
$\sigma_{h=i}^2=\left|\bar{y}_{N,h=i}\right|, i=1,..,4$

모형5: $\bar{y}_{th}=\bar{y}_{Nt}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma^2/n_{th}\right),\ \bar{y}_{N,t=1}=5,\ \bar{y}_{N,t=2}=-5,\ \sigma^2=1.5^2$

모형6: $\bar{y}_{th}=\bar{y}_{Nt}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma^2/n_{th}\right)$
$\bar{y}_{N,t=1}=5,\ \bar{y}_{N,t=2}=-5,\ \sigma_{h=i}^2=1.5^i, i=1,..,4$

모형7: $\bar{y}_{th}=\bar{y}_{Nth}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma^2/n_{th}\right),\ \sigma^2=1.5^2$
$\bar{y}_{N,t=1,h=1}=5,\ \bar{y}_{N,t=2,h=1}=-5,\ \bar{y}_{N,t=1,h=2}=3,\ \bar{y}_{N,t=2,h=2}=-3$
$\bar{y}_{N,t=1,h=3}=-3,\ \bar{y}_{N,t=2,h=3}=3,\ \bar{y}_{N,t=1,h=4}=-5,\ \bar{y}_{N,t=2,h=4}=5$

모형8: $\bar{y}_{th}=\bar{y}_{Nth}+\epsilon_{th},\ \epsilon_{th}\sim\left(0,\sigma_h^2/n_{th}\right),\ \sigma_{h=i}^2=\left|\bar{y}_{N,h=i}\right|, i=1,..,4$
$\bar{y}_{N,t=1,h=1}=5,\ \bar{y}_{N,t=2,h=1}=-5,\ \bar{y}_{N,t=1,h=2}=3,\ \bar{y}_{N,t=2,h=2}=-3$
$\bar{y}_{N,t=1,h=3}=-3,\ \bar{y}_{N,t=2,h=3}=3,\ \bar{y}_{N,t=1,h=4}=-5,\ \bar{y}_{N,t=2,h=4}=5$

〈표 1〉 표본 배분

| | 모집단 | | 순환표본(1) $n_{t=1,h} = n_{t=2,h}$ | | 순환표본(2) $n_{t=1,h} = 2n_{t=2,h}$ | |
|---|---|---|---|---|---|---|
| | $t$=1 | $t$=2 | $t$=1 | $t$=2 | $t$=1 | $t$=2 |
| $h$=1 | 300 | 300 | 30 | 30 | 40 | 20 |
| $h$=2 | 240 | 240 | 24 | 24 | 32 | 16 |
| $h$=3 | 180 | 180 | 18 | 18 | 24 | 12 |
| $h$=4 | 120 | 120 | 12 | 12 | 16 | 8 |

위와 같이 구성된 8가지 모형을 바탕으로 2개 시점의 순환표본을 추출하였다. 표본추출 방법은 층화추출법을 고려하였으며 표본 배분은 각 층 내의 조사단위들의 크기에 비례하여 각 층의 표본크기를 배분하는 비례배분법(proportional allocation)을 실시하였다. 층별, 시점별 모집단과 고려하고 있는 두 순환표본의 크기는 〈표 1〉과 같다. 고려된 순환표본은 두 시점에 따라 층별 표본크기가 변하지 않는 순환표본(1)과 한 시점의 층별 표본크기가 다른 시점의 층별 표본크기의 2배가 되는 순환표본(2) 두 가지이다. 이는 본 연구결과가 실제 적용될 질병관리본부의 국민건강영양조사의 2개 연도 순환표본의 구성이 실제 순환표본(1)과 같이 설계되었으나 조사상에서 순환표본(2)와 같이 진행된 것을 반영하기 위한 것이다.

## 2. 추정량 비교

1절에서 정의된 8가지 모형을 바탕으로 구성된 두 가지 순환표본에 대하여 추정량들의 성능을 효율성(efficiency)과 강건성(robustness) 측면에서 비교한다. 각 모형에 대하여 모집단을 생성하고 모집단으로부터 순환표본 구성에 맞게 표본을 추출한 후 6가지 추정량을 계산하는 과정을 10,000번 반복 시행하여 각 추정량에 대한 Monte Carlo 평균과 분산, 편의(Bias) 그리고 평균제곱오차(MSE)를 계산하였다. 추정량들의 효율성 비교를 위해서는 상대평균제곱오차(Relative Mean Squared Error; RMSE)를 사용하였으며, 상대평균제곱오차는 각 모형에서의 BLUP 추정량을 기준으로 다음과 같이 정의하였다.

$$RMSE = \frac{MSE\left(\hat{\bar{y}}_N\right)}{MSE\left(\hat{\bar{y}}_{N,BLUP}\right)}$$

1) 순환표본(1) : $n_{1,h} = n_{2,h}$

두 시점의 각 층별 표본크기가 동일하며 각 층의 모분산이 같은 경우, 즉 $n_{t=1,h} = n_{t=2,h}$, $\sigma_h^2 = \sigma^2$인 경우에는 고려한 6개의 추정량이 모두 아래와 같이 표현된다.

$$\text{추정량[1]} : \quad \hat{\bar{y}}_N^{(1)} = \sum_{t=1}^{2}\sum_{h=1}^{4} \frac{n_{th}}{n}\bar{y}_{th}$$

그러나 각 층의 모분산이 다른 경우에는 추정량을 구하는 과정에서 모분산 정보의 사용 여부에 따라서 추정량 Ⅰ, Ⅲ, Ⅳ, Ⅵ는 추정량[1], 추정량 Ⅱ, Ⅴ는 아래와 같이 표현된다.

$$\text{추정량[2]} : \quad \hat{\bar{y}}_N^{(2)} = \sum_{t=1}^{2}\sum_{h=1}^{4} \frac{n_{th}/\sigma_h^2}{\left(\sum_{t'}\sum_{h'}\left(n_{t'h'}/\sigma_{h'}^2\right)\right)}\bar{y}_{th}$$

정의된 8가지 모집단에 대하여 2개 연도의 순환표본을 추출하여 6가지 추정량의 Monte Carlo 성질을 비교한 결과는 〈표 2〉와 같다.

**〈표 2〉 순환표본(1)에서의 8가지 모형 비교**

| | | Mean | Var | Bias | MSE |
|---|---|---|---|---|---|
| 모형1 | 추정량[1] | -0.0007 | 0.0134 | 0.0007 | 0.0121 |
| 모형3 | 추정량[1] | 1.2860 | 0.0134 | 0.0001 | 0.0121 |
| 모형5 | 추정량[1] | 0.0002 | 0.0134 | 0.0003 | 0.0121 |
| 모형7 | 추정량[1] | -0.0002 | 0.0134 | 0.0002 | 0.0120 |
| 모형2 | 추정량[1] | 0.0003 | 0.0157 | 0.0003 | 0.0142 |
| | 추정량[2] | 0.0001 | 0.0131 | 0.0001 | 0.0120 |
| 모형4 | 추정량[1] | 1.2860 | 0.0237 | 0.0002 | 0.0213 |
| | 추정량[2] | 1.0717 | 0.0223 | 0.2141 | 0.0661 |
| 모형6 | 추정량[1] | 0.0000 | 0.0240 | 0.0000 | 0.0216 |
| | 추정량[2] | 0.0002 | 0.0225 | 0.0002 | 0.0204 |
| 모형8 | 추정량[1] | -0.0009 | 0.0156 | 0.0008 | 0.0141 |
| | 추정량[2] | -0.0006 | 0.0130 | 0.0005 | 0.0120 |

〈표 3〉 모형과 추정량에 따른 RMSE

| | | 추정량[1] | 추정량[2] |
|---|---|---|---|
| 시간의 효과가 없는 모형 | 모형2 | 1.183 | 1.000 |
| | 모형4 | 1.000 | 3.103 |
| 시간의 효과가 있는 모형 | 모형6 | 1.059 | 1.000 |
| | 모형8 | 1.000 | 0.851 |

층별 분산이 모두 동일한 모형 1, 3, 5, 7 하에서 추정량[1]이 불편추정량임을 확인할 수 있다. 층별 분산이 다른 모형 2, 4, 6, 8의 경우, 모형8의 경우를 제외하고 해당 모형 하에서 BLUP가 다른 추정량보다 작은 MSE를 갖는다. 모형8의 경우 BLUP로 정의되는 추정량[1]이 상대적으로 큰 MSE를 갖는데 이는 모형8 하에서 정의되는 각 순환표본시점별, 층별 추정량인 $\bar{y}_{th}$가 $\bar{y}_{th,N}$의 BLUP지만 이들의 가중평균을 구하는 과정에서 모집단의 분산 정보가 적절하게 사용되지 않았기 때문인 것으로 판단된다. 모형4 하에서 추정량[2]의 비효율성은 추정량의 편의로 인한 것임을 확인할 수 있다.

순환표본(1)을 사용할 경우 모형 2, 4, 6, 8 하에서 정의된 두 가지 추정량의 RMSE는 〈표 3〉과 같다. 추정량[2]는 모형8 하에서 RMSE가 0.851로 추정량[1]보다 높은 효율성을 나타내지만 모형4 하에서의 RMSE는 3.103으로 매우 크게 나타나고 있다. 반면에 추정량[1]의 RMSE는 모형2 하에서 1.183, 모형6 하에서 1.059로 나타나고 있다. 즉 추정량[1]의 경우 시점효과 유무에 따른 모형 변화에 대하여 추정량[2]보다 강건한 모습을 보여 주고 있다.

2) 순환표본(2) : $n_{1,h} = 2n_{2,h}$

한 시점의 층 표본크기가 다른 시점의 2배이며 각 층의 모분산이 같은 경우, 즉 $n_{1,h} = 2n_{2,h}$, $\sigma_h^2 = \sigma^2$인 경우 평균이 조사시점에 따라 변하지 않는 모형 하에서 얻어진 BLUP Ⅰ, Ⅱ, Ⅲ은 추정량[1]로 표현되며, 평균이 조사시점에 영향을 받는 모형 하에서 얻어진 BLUP Ⅳ, Ⅴ, Ⅵ 은

$$\text{추정량[3]} : \quad \hat{\bar{y}}_N^{(3)} = \sum_{t=1}^{2}\sum_{h=1}^{4}\frac{N_{th}}{N}\bar{y}_{th}$$

로 표현된다.

각 층의 모분산이 다른 경우, 평균이 조사시점에 영향을 받지 않는 모형 하에서 모분산 정보를 사용하지 않은 BLUP Ⅰ, Ⅲ은 추정량[1], 모분산 정보를 사용한 BLUP Ⅱ는 추정량[2]로 표현된다. 평균이 조사시점에 영향을 받는 모형 하에서 모분산 정보를 사용하지 않은 BLUP Ⅳ, Ⅵ는 추정량[3], 모분산 정보를 사용한 BLUP Ⅴ는

$$\text{추정량[4]} : \quad \hat{\bar{y}}_N^{(4)} = \sum_{t=1}^{2} \frac{N_t}{N} \sum_{h=1}^{4} \frac{n_{th}^*}{n_t^*} \bar{y}_{th}$$

로 각각 표현된다.

**〈표 4〉 순환표본(2)에서의 8가지 모형 비교**

| | | Mean | Var | Bias | MSE |
|---|---|---|---|---|---|
| 모형1 | **추정량[1]** | -0.0007 | 0.0134 | 0.0006 | 0.0121 |
| | 추정량[3] | -0.0005 | 0.0151 | 0.0005 | 0.0138 |
| 모형3 | **추정량[1]** | 1.2854 | 0.0134 | 0.0004 | 0.0121 |
| | 추정량[3] | 1.2852 | 0.0151 | 0.0006 | 0.0137 |
| 모형5 | 추정량[1] | 0.4278 | 0.0132 | 0.4278 | 0.1950 |
| | **추정량[3]** | -0.0008 | 0.0149 | 0.0007 | 0.0136 |
| 모형7 | 추정량[1] | 1.6667 | 0.0135 | 1.6669 | 2.7906 |
| | **추정량[3]** | 0.0001 | 0.0152 | 0.0003 | 0.0138 |
| 모형2 | 추정량[1] | 0.0004 | 0.0156 | 0.0003 | 0.0140 |
| | **추정량[2]** | 0.0004 | 0.0130 | 0.0003 | 0.0120 |
| | 추정량[3] | 0.0005 | 0.0176 | 0.0005 | 0.0161 |
| | 추정량[4] | 0.0005 | 0.0147 | 0.0004 | 0.0137 |
| 모형4 | **추정량[1]** | 1.2861 | 0.0156 | 0.0002 | 0.0140 |
| | 추정량[2] | 2.7144 | 0.0130 | 1.4286 | 2.0527 |
| | 추정량[3] | 1.2858 | 0.0176 | 0.0000 | 0.0161 |
| | 추정량[4] | 2.7141 | 0.0147 | 1.4283 | 2.0536 |
| 모형6 | 추정량[1] | 0.4292 | 0.0156 | 0.4292 | 0.1982 |
| | 추정량[2] | 0.9052 | 0.0130 | 0.9051 | 0.8312 |
| | 추정량[3] | 0.0007 | 0.0175 | 0.0007 | 0.0160 |
| | **추정량[4]** | 0.0004 | 0.0146 | 0.0004 | 0.0135 |
| 모형8 | 추정량[1] | 1.6662 | 0.0156 | 1.6662 | 2.7902 |
| | 추정량[2] | 1.6662 | 0.0130 | 1.6661 | 2.7880 |
| | **추정량[3]** | -0.0006 | 0.0175 | 0.0006 | 0.0160 |
| | 추정량[4] | -0.0006 | 0.0146 | 0.0007 | 0.0136 |

정의된 8가지 모집단에 대하여 2개 연도의 순환표본(2)를 추출하여 6가지 추정량의 Monte Carlo 성질을 비교한 결과는 〈표 4〉와 같다. 층별 분산이 같은 모형 1, 3, 5, 7의 경우 해당 모형 하에서 BLUP가 다른 추정량보다 작은 MSE를 갖는다. 모형5 하에서 추정량[3]의 비효율성은 추정량의 편의로 인한 것임을 확인할 수 있다. 또한 층별 분산이 다른 모형 2, 4, 6, 8의 경우, 순환표본(1)과 같이 모형8의 경우를 제외하고 해당 모형하에서 BLUP가 다른 추정량들보다 작은 MSE를 갖는다. 모형4에서는 추정량[2]가, 모형6과 모형8에서는 추정량[1]과 추정량[2]가 비효율성을 보이며 이것은 추정량의 편의로 인한 것임을 확인할 수 있다.

순환표본(2)를 적용한 경우 8가지 모형 하에서 정의된 6가지 추정량의 RMSE는 〈표 5〉와 같다. 시간의 효과가 없는 모형에서는 추정량[1]이, 그리고 시간의 효과가 있는 모형에서는 추정량[3]이 전반적으로 우수하게 나타나고 있다. 추정량[2]와 추정량[4]의 경우 주어진 모형의 가정이 만족되지 않으면 MSE가 크게 증가하므로 모형 가정에 매우 민감한 것을 알 수 있다. 모든 가능한 모형을 고려한 결과 추정량[3], 즉 층별, 시점별 모집단 크기를 이용한 추정량이 MSE의 변화가 가장 작게 나타나 강건성이 좋은 추정량으로 판단된다.

결론적으로 순환표본(1)의 경우, 층별 그리고 시점별 표본크기를 이용한 추정량[1]의 성능이 안정적으로 나타나며 순환표본(2)의 경우, 층별 그리고 시점별 모집단 크기를 이용하여 정의된 추정량[3]이 안정적인 결과를 제공하는 것으로 모의실험 결과가 나타났다.

**〈표 5〉 모형과 추정량에 따른 RMSE**

| | | 추정량[1] | 추정량[3] | 추정량[2] | 추정량[4] |
|---|---|---|---|---|---|
| 시간의 효과가 없는 모형 | 모형1 | 1.000 | 1.140 | | |
| | 모형2 | 1.167 | 1.342 | 1.000 | 1.142 |
| | 모형3 | 1.000 | 1.132 | | |
| | 모형4 | 1.000 | 1.150 | 146.621 | 146.686 |
| 시간의 효과가 있는 모형 | 모형5 | 14.338 | 1.000 | | |
| | 모형6 | 14.681 | 1.185 | 61.570 | 1.000 |
| | 모형7 | 202.217 | 1.000 | | |
| | 모형8 | 174.388 | 1.000 | 174.250 | 0.850 |
| 합 계 | | 409.791 | 8.950 | 383.442 | 149.677 |

## Ⅴ. 사례연구 : 제4기 국민건강영양조사

제4기 국민건강영양조사는 기존의 3년 주기의 조사와 달리 3년 동안 상시적으로 조사를 수행할 수 있는 순환조사 방식을 통하여 이루어졌다. 이는 국민건강영양조사의 일부인 검진조사를 위한 전문인력이 제한되어 있음에 일부 기인한다. 전문인력을 통한 자료수집은 측정자 간의 변동을 최소화시킬 수 있으며, 이러한 순환조사는 미국의 Nation Centers for Health Statistics(NCHS)에서 실시하는 National Health and Nutrition Examination Survey(NHANES)에서 적용되고 있다.

제4기 국민건강영양조사 표본은 3개 연도에 걸쳐 독립적인 3개의 순환표본으로 구분된다. 각각의 독립적인 순환표본은 전국 모집단 가구를 대표할 수 있도록 구성되어 있다. 제4기 국민건강영양조사를 위한 표본설계에 대하여서는 이계오(2007)을 참조하면 된다. 추출된 각각의 순환표본을 통해 매년 전국 단위의 통계를 생산함과 동시에 3개 연도 순환표본의 결합을 통하여 시·도 단위의 통계를 생산한다.

본 장에서는 3개 연도의 순환표본의 결합에 앞서 조사가 완료된 2007년, 2008년의 순환표본을 바탕으로 앞서 제시한 추정량을 적용하여 2개 연도의 순환표본의 가능한 결합방법들을 살펴보고 또한 추후 3개년 순환표본 결합에 대한 적절한 방법을 살펴보고자 한다.

제4기 국민건강영양조사의 2개 연도 순환표본의 크기는 최초 설계시 2007년과 2008년 모두 200개의 조사구를 대상으로 조사할 예정이었으나 실제 조사과정에서 2007년 100개 조사구, 2008년 200개의 조사구가 조사되었으며, 표본설계 시 사용된 층화방법을 고려해 볼 때 층별 분산의 동질성에 대한 가정을 만족한다고 판단된다. 모의실험에서 고려한 순환표본(2)의 경우에서 살펴본 바와 같이 추정량[2]와 추정량[4]의 경우 그 효율성이 모수에 크게 의존하는 문제점이 있어 제4기 국민건강영양조사 자료의 분석을 위해서 추정량[1]과 추정량[3]만을 고려한다. 2개 연도 자료를 이용하여 정의되는 추정량은 다음과 같이 표현된다.

$$\hat{\bar{y}}_N^{(1)} = \sum_{t=1}^{2}\sum_{h}\frac{n_{th}}{n}\hat{\bar{y}}_{th} = \frac{1}{3}\hat{\bar{y}}_{(1)} + \frac{2}{3}\hat{\bar{y}}_{(2)}$$

$$\hat{\bar{y}}_N^{(3)} = \sum_{t=1}^{2}\sum_{h}\frac{N_{th}}{N}\hat{\bar{y}}_{th} = \frac{1}{2}\hat{\bar{y}}_{(1)} + \frac{1}{2}\hat{\bar{y}}_{(2)}$$

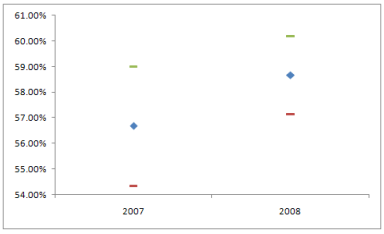
여기서 $\hat{\bar{y}}_{(i)} = \sum_h \left(n_{ih} / \sum_{h'} n_{ih'}\right) \hat{\bar{y}}_{ih}$ , $i = 1,2$ 은 각 순환표본으로부터 구하여진 모평균의 추정량을 나타낸다. 즉 각각의 순환표본으로부터 구해진 추정량 $\hat{\bar{y}}_{(i)}$에 가중치를 부여한 것으로, 층과 시점에 대하여 추정량[1]은 표본크기에 대한 정보를, 추정량[3]은 모집단 크기에 대한 정보를 사용한 것이다.

본 연구에서 고려한 주요 관심변수는 건강설문조사의 월간음주여부와 현재흡연여부, 검진조사의 비만유병여부와 당뇨유병여부, 영양조사의 에너지(식품섭취)와 식품섭취량이다. 총 6개의 변수에 대하여 두 가지 결합 방법을 적용한 결과는 〈표 6〉~〈표 11〉과 같다. 모든 변수에서 두 결합 방법을 적용한 결과 모든 변수에서 추정량[1]의 표준오차가 추정량[3]보다 작게 나타남을 확인할 수 있다. 이는 모의실험 결과에서 살펴 본 바와 같이 시점별 평균의 차이가 없는 경우 추정량[1]이 더 효율적이기 때문이다. 한편 추정량[3]의 계산되어진 추정량의 오차가 (1)보다 크게 나타나지만 그 차이가 작고 추정치 역시 (1)과 큰 차이가 없음을 알 수 있다.

**〈표 6〉 월간음주여부**

(단위: 명)

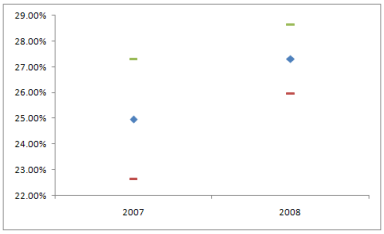
| 추정량 | 표본수 | 비율(%) | S.E(%) |
|---|---|---|---|
| 2007년 | 2979 | 56.68 | 1.168 |
| 2008년 | 6796 | 58.67 | 0.771 |
| (1) | 9775 | 58.01 | 0.647 |
| (3) | 9775 | 57.68 | 0.700 |




**〈표 7〉 현재흡연여부**

(단위: 명)

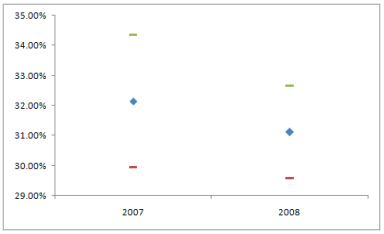
| 추정량 | 표본수 | 비율(%) | S.E(%) |
|---|---|---|---|
| 2007년 | 2980 | 24.96 | 1.169 |
| 2008년 | 6797 | 27.30 | 0.681 |
| (1) | 9777 | 26.53 | 0.584 |
| (3) | 9777 | 26.14 | 0.657 |



〈표 8〉 비만유병여부(19세 이상)

(단위: 명)

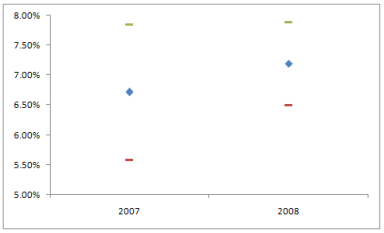
| 추정량 | 표본수 | 비율(%) | S.E(%) |
|---|---|---|---|
| 2007년 | 2997 | 32.15 | 1.105 |
| 2008년 | 6727 | 31.11 | 0.779 |
| (1) | 9724 | 31.46 | 0.630 |
| (3) | 9724 | 31.63 | 0.676 |




〈표 9〉 당뇨유병여부(10세 이상)

(단위: 명)

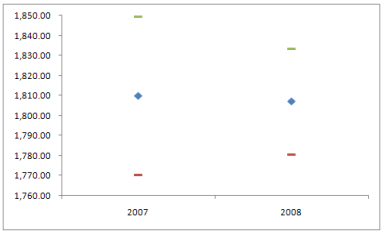
| 추정량 | 표본수 | 비율(%) | S.E(%) |
|---|---|---|---|
| 2007년 | 3277 | 6.72 | 0.568 |
| 2008년 | 7345 | 7.19 | 0.354 |
| (1) | 10622 | 7.03 | 0.302 |
| (3) | 10622 | 6.95 | 0.352 |




〈표 10〉 에너지(식품섭취)

(단위: 명)

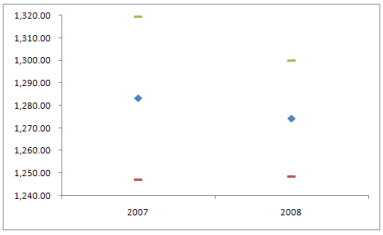
| 추정량 | 표본수 | Mean | S.E |
|---|---|---|---|
| 2007년 | 4091 | 1809.86 | 19.791 |
| 2008년 | 8628 | 1806.99 | 13.391 |
| (1) | 12719 | 1808.05 | 11.137 |
| (3) | 12719 | 1808.54 | 12.223 |




〈표 11〉 식품섭취량

(단위: 명)

| 추정량 | 표본수 | Mean | S.E. |
|---|---|---|---|
| 2007년 | 4091 | 1283.22 | 18.080 |
| 2008년 | 8628 | 1274.19 | 13.079 |
| (1) | 12719 | 1279.05 | 10.536 |
| (3) | 12719 | 1277.52 | 11.339 |



제4기 국민건강영양조사는 3개의 순환표본을 고려한 조사로서 앞으로 2009년도 순환표본에 대한 자료를 추가하여 3개 연도의 순환표본에 대한 결합을 고려하여야 한다. 본 연구의 결과를 살펴볼 때 각 연도별 평균의 통계적 차이가 예측되는 경우, 모집단의 정보를 이용한 일종의 조사시점을 사후층으로 이용한 사후층화 추정량이 적절하다고 판단된다. 또한 연도별 모평균의 변화가 크지 않은 경우, 모집단 정보대신 표본 정보를 이용한 추정량의 통계적 효율성이 높게 나타난다. 국민건강영양조사의 주요 조사 내용이 단기간에 급격한 변화를 나타내지 않는 것과 순환조사의 총 시기가 3년으로 길지 않음을 고려할 때 표본 정보를 이용한 추정량의 가중치가 순환표본 결합을 위해 사용할 수 있을 것으로 판단된다. 그러나 실제 가중치 작성을 위해서는 조사에 포함된 모든 변수들의 연도별 변화에 대한 검토가 선행되어야 할 것이다.

## Ⅵ. 결 론

본 연구에서는 순환표본의 결합을 위한 가중치의 산출을 위하여 많은 관심변수에 적용 가능한 선형모형들을 고려하고 각 모형 하에서 정의되는 BLUP의 성능을 비교하였다. 모의실험을 통하여 살펴본 결과, 두 시점의 각 층별 표본크기가 동일한 순환표본(1)과 한 시점의 층 표본크기가 다른 시점의 2배인 순환표본(2)에서 모형8의 경우를 제외하고 해당 모형 하에서 BLUP가 다른 추정량들보다 작은 MSE를 갖는다. 모형 4, 5, 6, 7에서 나타나는 추정량의 비효율성은 추정량의 편의로 인한 것임을 확인하였다. 모형8의 경우 BLUP로 정의되는 추정량[1]이 상대적으로 큰 MSE를 갖는데 이것은 모집단의 분산 정보가 적절하게 사용되지 않았기 때문으로 여겨진다. 순환표본(1)에서는 추정량[1]이 시간의 효과 여부에 따른 모형 변화에 대하여 추정량[2] 보다 강건함을 보여주고 있다. 순환표본(2)에서는 시간의 효과가 없는 모형에서는 추정량[1]이, 시간의 효과가 있는 모형에서는 추정량[3]이 MSE 측면에서 우수하게 나타나고 있으며, 모든 모형을 고려하면 추정량[3]이 MSE의 변화가 가장 작게 나타나 강건성이 좋은 추정량으로 판단된다.

순환표본의 결합에 대한 본 연구에서는 기존의 순환조사에서 적용된 방법론적인 접근방식에서 벗어나 실질적으로 적용 가능한 선형모형을 가정하고, 그 모형하에서 BLUP를 제시함으로써 주어진 모형 하에서 효율적인 결합방법을 제고할 수 있는 방안들을 소개하였다. 순환표본의 결합에 앞서 순환표본조사와 표본구조에 대한 정확한 이해를 통하여 제안

된 여러 가지 모형을 고려하고 가능한 모형 하에서 모형 오류에 강건한 추정량을 유도하는 방안이 본 연구를 통해 고려될 수 있을 것이다.

## 참고문헌

김규성. 2009. “인구주택총조사 대안 방법으로의 순환총조사.”《조사연구》10(1): 97−114.

이계오. 2007. “제4기 국민건강·영양조사 표본설계.” 최종보고서.

전광희. 2008. “미국 센서스의 변화와 향후 전망: 2000년의 경험과 2010의 계획을 중심으로.” 《한국인구학》31(2): 101−132.

Alexander, C. H. 1998. “Recent Developments in the American Community Survey.” *Proceedings of the American Statistical Association, Survey Research Methods Section* 91−100.

Alexander, C. H. 2002. “Still Rolling; Leslie Kish's ‘Rolling Samples’ and The American Community Survey.” *Survey Methodology* 28(1): 35−41.

Friedman, E. M. 2003. “Combined Estimates from Four Quarterly Survey Data Sets.” *Proceedings of the American Statistical Association, Survey Research Methods Section* 1064−1069.

Kish, L. 1979. “Samples and Censuses.” *International Statistical Review* 47, 99−109.

Kish, L. 1990. “Rolling Samples and Censuses.” *Survey Methodology* 16(1): 63−71.

Kish, L. 1998. “Space/Time Variations and Rolling Samples.” *Journal of Official Statistics*, 14(1): 31−46.

Kish, L. 1999. “Cumulating/Combining Population Surveys.” *Survey Methodology* 25(2): 129−138.

Hefter, S. P. and Williams, A. L. 2008. “America Community Survey: Sample Design Issues and Challenges.” *Proceedings of the American Statistical Association, Survey Research Methods Section* 3452−3459.

US Census Bureau. 2006. "Design and Methodology, American Community Survey.” *Technical Paper* 67.

[접수 2010/2/18, 수정 2010/3/12, 게재확정 2010/3/15]