Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 843–858

BLS 보정 방법의 민감도에 관한 연구†

이석진[1], 신기일[2]

요 약

사업체 조사에서 사용되는 BLS 보정방법은 표본으로 선택된 사업체의 설계 가중치를 사후에 보정해 줌으로써 모집단의 대표성을 유지해 줄 뿐만 아니라 더욱 정확한 총계 추정을 가능하게 해준다. 일반적으로 BLS 보정은 설계가중치와 표본 틀(Sample frame)의 보조변수를 이용하여 정해지기 때문에 사용된 보조변수에 따라 보정인자의 결과 값이 다르게 된다. 본 논문에서는 보조변수가 있는 경우, 보조 변수가 없는 경우 그리고 다른 보조변수를 사용할 경우에 최종가중치가 어떻게 변하는지를 파악하기 위해 보조변수의 민감도를 살펴보았다. 본 논문에서 사용된 자료는 노동부의 2007년 매월노동통계 자료이다.

주요용어: 표본설계; 가중치; 무응답 보정; 이상점 보정.

1. 서론

표본설계에서 가장 중요한 부분은 표본 추출 방법과 가중치 계산방법이다. 최근 잦은 휴폐업과 무응답 그리고 이상치 등 조사과정에서 발생하는 비표본 오차는 점차 커지고 있는 실정이다 (Kish, 1965). 따라서 비표본 오차의 중요성은 점차 커지고 있으며 비표본 오차를 줄이는 총계추정방법은 매우 중요하다고 하겠다. 비표본 오차를 줄이는 하나의 방법인 결측치(missing value) 대체법에 관한 연구가 활발히 진행되고 있으며 (Rubin, 1987; Little과 Rubin, 2002) 국내에서도 결측치 대체법에 관한 연구는 활발히 진행되고 있다 (김규성, 2000; 이진희와 신기일, 2007). 또한 대체법과는 다른 방법인 갈퀴법을 사용함으로써도 비표본 오차를 줄일 수 있다. 갈퀴법에 관한 내용은 Oh와 Scheuren (1987) 또는 신민웅과 이상은 (2001)을 살펴보기 바란다. 그러나 비표본 오차를 줄이는 총계추정방법은 가중치 계산법을 통하여 이루어질 수도 있기 때문에 가중치에 관한 연구는 매우 의미가 있다고 하겠다. 일반적인 가중치 계산 방법은 기본적으로 표본 추출 방법에 의해 크게 좌우된다. 이는 표본 추출방법에 의해 기본 가중치 또는 설계 가중치가 결정되기 때문이다. 만약 모든 일이 잘 진행된다면 총계 추정치(Estimate of total) 계산에는 설계 가중치가 사용될 것이다. 그러나 실제 조사를 진행하다보면 많은 일이 발생한다. 예를 들어 무응답이 발생할 것이고, 이상점이 발생할 것이다. 또한 사업체가 폐업할 것이고, 새로운 신규진입

† 이 논문은 2008년 한국외국어대학교 교내연구비의 지원을 받았음.
1) (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 석사과정.
2) (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수.
교신저자: keyshin@hufs.ac.kr

사업체가 있을 것이다. 이러한 모든 상황은 조사가 이루어지기 전에는 알 수 없는 상황이다.

일반적으로 가중치 계산은 크게 두 가지로 나누어질 수 있다. 즉 조사전의 설계 당시에 구해지는 설계 가중치 그리고 조사 후에 계산되어지는 사후 조정이다. 결국 최종 가중치는 설계 가중치에 사후조정을 곱하는 형식으로 이루어진다. 본 논문에서 연구될 보정 방법은 BLS 방법으로 미국 노동통계청(Bureau of labor statistics)이 사업체 조사를 위해 만든 것이다. 이 방법을 사용하기 위해서는 기본적으로 표본에 주어진 설계가중치와 표본 틀에 보조변수가 있어야 한다. 일반적으로 보조변수로는 종사자수가 사용된다. 이 두 수치는 큰 어려움 없이 얻을 수 있다. 국내에서도 사업체 조사를 위해 사업체기초통계조사가 표본 틀로 사용되고 있다. 사업체기초통계조사를 이용하게 되면 설계가중치 뿐 아니라 보조변수인 종사자 수 또한 얻을 수 있다. 노동부는 사업체기초통계조사 자료를 근본으로, 노동부에서 필요한 자료로 가공한 사업체실태조사 자료를 만들고 있다. 이 사업체실태조사가 노동부에서 실시하는 각종 표본 설계의 표본 틀로 이용되고 있다. 사업체실태조사 자료에는 상용근로자 수뿐만 아니라 임시 및 일용 근로자 수도 함께 수록되어 있다. 또한 본 논문에서 사용 된 자료인 매월노동통계 조사자료의 경우에도 같은 자료가 수록되어 있다. 따라서 미국의 BLS 방법을 이용하여 보정을 할 경우 상용근로자를 기준으로 할 것인지 또는 임시 및 일용 근로자를 포함한 총 근로자수를 사용할 것인지가 고민이 될 수 있다. 이에 본 논문에서는 총계추정에 사용되는 보조변수가 얼마나 영향을 미치는지 살펴보았다. 2장에는 미국의 BLS 방법을 설명하였고, 3장에는 보조변수가 총계추정에 미치는 영향력인 민감도에 관하여 이론적인 전개를 실시하였으며 4장에는 2007년 매월노동통계 자료를 이용하여 모의실험을 실시하였다. 5장에 결론이 있다.

2. BLS(Bureau of Labor Statistics) 보정 방법

BLS 방법은 미국 노동통계청(Bureau of Labor Statistics)에서 사업체조사에 사용하는 방법이다. BLS 가중치 방법의 가장 대표적인 보정 단계는 다음의 4단계이다. 즉 집계 후 보정(Re-aggregation Adjustment), 무응답 보정(Non-response Adjustment), 이상점 보정(Outlier Adjustment) 그리고 벤치마크 보정(Benchmark Adjustment)이다. 이외에 다른 보정 방법이 발표되었으나 본 연구에서는 대부분의 표본조사에서 가장 많이 활용되는 위의 4단계만을 간단히 설명하였다. 이 방법에 관한 자세한 내용과 다음의 식 (2.2)와 (2.3)의 가중치 합이 모집단 수가 되는 증명은 이상은 (2008)을 참조하기 바란다. 또한 BLS 방법에 관한 자세한 내용은 Burdete (2003)을 참조하기 바란다.

2.1. 집계 후 보정(Re-aggregation Adjustment: REAG)

REAG는 표본 조사 후 조사 결과의 차이가 큰 경우(일반적으로 2배 이상인 경우)에 적용하는 보정 방법이다. 예를 들면 사업체의 종사자수가 표본설계 당시 10명에서 25명으로 2배 이상이 된 경우의 보정방법이다. 이때 보정을 하지 않게 되면 그 표본이 대표하는 모든 사업체의 종사자 수가 2배 이상으로 추정되어 과대추정의 원인이 될 수 있다. 반대의

경우, 즉 과소추정의 원인이 되는 경우에도 이를 보정할 수 있다. 다음의 보정 공식을 이용하여 이를 보정하게 된다.

$$\text{Reag Factor} = f_i^{REAG} = \frac{\sum y_i^D}{\sum y_i^O}, \tag{2.1}$$

여기서 y_i^D는 표본설계 시 변수의 값(조사 전, Design)이며 y_i^O는 조사 후 얻어진(Observed) 변수의 값이다. 이 보정 방법은 가중치를 보정하는 것이 아니라 자료값을 보정한다.

2.2. 무응답 보정(Nonresponse Adjustment: NRA)

NRA는 가장 많은 관심을 갖는 방법이며 또한 다양한 방법이 제시되고 있다. BLS의 무응답 보정방법은 다음과 같다.

$$\text{NRA Factor} = f_i^{NRA} = \frac{\sum_{j\in h} w_j y_j^D}{\sum_{j\in h_0} w_j y_j^D}, \tag{2.2}$$

여기서 w_j는 표본설계 시 주어진 기본 가중치이고, h는 조사시점에서 존재하지 않는 조사대상자(폐업)를 제외한 모든 조사대상자 집단을 나타내며, h_0는 조사시점에서 존재하지 않는 조사대상자(폐업)와 무응답을 한 조사 대상자를 제외한 모든 조사대상자 집단을 나타낸다.

2.3. 이상점 보정(Outlier Adjustment Factor: OAF)

OAF는 조사결과가 이상점으로 판명이 났을 때 이 결과를 보정하기 위한 방법으로 이상점에 부여된 가중치를 “1” 값으로 환원하게 만든다. 이를 위한 공식은 다음과 같다.

$$\text{Outlier Factor} = f_i^{OAF} = \frac{1}{w_i \times f_i^{REAG} \times f_i^{NRA}}. \tag{2.3}$$

이상점으로 판명된 값을 환원한 후에는 이상점이 아닌 자료의 가중치를 보정하여 전체 가중치의 합이 유지되도록 하여야 한다. 즉 모집단 수는 정해진/알려진 것으로 전체 가중치의 합은 모집단 수와 같아야 하므로 이상점에서 환원된 가중치의 나머지 부분을 이상점이 아닌 자료에 추가로 배분해 주어야 한다. 이를 위하여 먼저 식 (2.2)의 분자와 분모를 각각 $\text{NOF}_1 = \sum_{j\in h} w_j y_j^D$, $\text{NOF}_2 = \sum_{j\in h_0} w_j y_j^D$라 하자. 그리고 $\text{ANOF}_1 = \text{NOF}_1 - \sum_j y_j^{D,out}$, $\text{ANOF}_2 = \text{NOF}_2 - \sum_j w_j y_j^{D,out}$라 하자. 여기서 $y_j^{D,out}$은 이상점 자료를 의미하며 ANOF_1는 NOF_1에서 이상점으로 판명된 자료의 합을 뺀 것이고, ANOF_2는 NOF_2에서 이상점으로 판명된 자료의 가중치를 곱한 자료의 합을 뺀 것이다. 그러면 무응답 보정 값은 다음과 같이 구해진다.

$$\text{Non-Outlier Factor} = f_i^{OAF(no)} = \frac{\text{ANOF}_1/\text{ANOF}_2}{f_i^{NRA}}. \tag{2.4}$$

2.4. 벤치마크 보정(Benchmark Adjustment Factor: BMF)

BMF는 가중치 보정 단계 중에서 마지막 단계로 조사시점의 알려진 최신 정보를 기준으로 가중치를 보정한다. 사용되는 Benchmark에 따라 가중치의 값이 크게 변하기 때문에 벤치마크의 설정은 신중해야 하며 많은 경우 센서스나 행정자료를 이용한다. 벤치마크의 보정방법은 다음과 같다.

$$\text{Benchmark Factor} = f_i^{BMF} = \frac{\text{Target value}_i(\text{Benchmark value}_i)}{\text{Reported value}_i}, \tag{2.5}$$

즉 Target value 혹은 Benchmark value가 주어지면 조사시점에서 얻은 추정치와의 비율을 이용하여 보정한다.

2.5. 최종 가중치(Final Weight)

전술한 4가지 단계를 마친 후 얻어진 최종 가중치는 다음과 같다.

$$w_i^{Final} = w_i \times f_i^{REAG} \times f_i^{NRA} \times f_i^{OAF} \times f_i^{BMF}. \tag{2.6}$$

식 (2.6)에 의해 최종가중치가 정해지면 BLS에 의한 추정량, $\hat{t_B}$은 다음과 같다.

$$\hat{t_B} = \sum_{i=1}^{n} w_i^{Final} y_i. \tag{2.7}$$

3. 민감도 분석

본 논문에서는 BLS에서 사용하는 4단계 중에서 큰 관심을 끌고 있는 무응답 보정과 이상점 보정을 살펴보았다. 먼저 무응답 보정에서 사용되고 있는 요인을 살펴보자. 일반적으로 관심변수와 상관이 높다고 생각되는 자료가 있을 경우 식 (2.2)을 사용하는 것은 당연하다. 그러나 여기서 얻어진 가중치는 관심 있는 변수 하나에 적용되는 것이 아니라 여러 관심 있는 변수에 사용된다. 예를 들어 노동부에서 생산하는 매월노동통계는 상용근로자수를 추정할 뿐 아니라 일인당 평균 근로시간과 일인당 평균 총임금을 추정한다. 일반적으로 상용근로자수와 평균 근로시간 또는 상용 근로자수와 일인당 평균 총임금 사이에는 큰 상관관계가 존재하지 않는다. 따라서 보정에 사용되는 보조변수, 즉 y_i^D의 영향력을 알아보는 것은 매우 중요하다. 이 같은 내용은 이상점 보정에도 똑같이 적용된다.

이제 주어진 보조변수 y_i^D의 영향력을 살펴보기 위하여 y_i^D에 일종의 교란요인(disturbance factor)인 x_i를 더한 새로운 변수 z_i, 즉 $z_i = x_i + y_i^D$를 생성하였다. 그리고 z_i를 식 (2.2)와 (2.4)의 y_i^D 대신에 대입하였다. 그러면 x_i를 변화시키면서 보조변수의 영향력을 살펴볼 수가 있게된다. 즉 x_i가 커지면 y_i^D의 영향력은 줄어들게 되고 x_i가 "0"에 가까우면 y_i^D의 영향력에는 변화가 없게된다. 본 논문에서는 이를 민감도라고 부르겠다. 본 논문에서는 다음의 네 경우를 살펴보았다.

Case 1: $x_i = 0$인 경우, 즉 $z_i = y_i^D$인 경우. 미국 BLS에서 사용.

Case 2: $X_i \sim \text{Poi}(\lambda)$, $\lambda > 0$이고, $z_i = x_i + y_i^D$인 경우. x_i와 y_i^D는 독립.

Case 3: $X_i \sim \text{Uni}(0, a)$, $a > 0$이고, $z_i = x_i + y_i^D$인 경우. x_i와 y_i^D는 독립.

Case 4: $z_i = 1$인 경우. 즉, $x_i = 0$이고 $y_i^D = 1$인 경우. 호주에서 사용.

사용된 보조변수 y_i^D가 종사자수이므로 x_i는 포아송분포를 선택하였다. 또한 포아송분포는 $E(X_i) = \lambda$, $\text{Var}(X_i) = \lambda$이므로 분산의 영향력을 살펴보기 위해 평균보다 분산이 매우 빨리 커지는 분포인 균일 분포를 선택하여 이의 영향력도 살펴보았다.

3.1. 무응답 보정에서의 민감도 분석

무응답 보정을 위한 보정인자는 다음과 같다.

$$f_i^{NRA} = \frac{\sum_{j\in h} w_j y_j^D}{\sum_{j\in h_0} w_j y_j^D}.$$

먼저 다른 모든 상황이 정상적이고 보정은 무응답 보정만이 필요하다고 가정하자. 그리고 분석을 단순화 하기 위하여 설계 가중치는 같은 셀 또는 층 안에서는 모두 같은 값인 $w_i = w$라 가정하자. 그러면

$$f_i^{NRA} = \frac{\sum_{j\in h} y_j^D}{\sum_{j\in h_0} y_j^D} \tag{3.1}$$

이 된다. 이제 민감도를 살펴보기 위하여 다음을 정의하자.

$$f_z^{NRA} = \frac{\sum_{j\in h} z_j}{\sum_{j\in h_0} z_j}. \tag{3.2}$$

Case 1: $x_i = 0$이므로 $f_z^{NRA} = f^{NRA}$인 미국 BLS와 같게 된다.

Case 2: $X_i \sim \text{Poi}(\lambda)$이므로 $E(X_i) = \lambda$, $\text{Var}(X_i) = \lambda$이 된다.

I) λ가 작은 경우

$$f_z^{NRA} = \frac{\sum_{j\in h} z_j}{\sum_{j\in h_0} z_j} = \frac{\sum_{j\in h} \left(x_j + y_j^D\right)}{\sum_{j\in h_0} \left(x_j + y_j^D\right)} = \frac{\sum_{j\in h} x_j + \sum_{j\in h} y_j^D}{\sum_{j\in h_0} x_j + \sum_{j\in h_0} y_j^D} \approx \frac{\sum_{j\in h} y_j^D}{\sum_{j\in h_0} y_j^D}$$

이 되어 $f_z^{NRA} = f^{NRA}$인 Case 1의 경우와 같은 결과를 준다.

II) λ가 큰 경우

$$f_z^{NRA} = \frac{\displaystyle\sum_{j\in h} z_j}{\displaystyle\sum_{j\in h_0} z_j} = \frac{\displaystyle\sum_{j\in h}\left(x_j + y_j^D\right)}{\displaystyle\sum_{j\in h_0}\left(x_j + y_j^D\right)} = \frac{\displaystyle\sum_{j\in h} x_j + \sum_{j\in h} y_j^D}{\displaystyle\sum_{j\in h_0} x_j + \sum_{j\in h_0} y_j^D}$$

이 된다. 위와는 반대로 λ가 큰 경우에는 $\sum_{j\in h} x_j$, $\sum_{j\in h_0} x_j$가 $\sum_{j\in h} y_j^D$, $\sum_{j\in h_0} y_j^D$보다 각각 매우 커지게 된다. 따라서 위 식은

$$\approx \frac{\displaystyle\sum_{j\in h} x_j}{\displaystyle\sum_{j\in h_0} x_j} \approx \frac{n_h}{n_{h_0}} \tag{3.3}$$

이 되어 Case 4의 결과와 같게 된다. 여기서 n_{h_0}는 h_0에 속한 사업체 수이고 n_h는 h에 속한 사업체 수이다.

Case 3: $X_i \sim \mathrm{Uni}(0, a)$이므로 $E(X_i) = a/2$이고 $\mathrm{Var}(X_i) = a^2/12$이다. 따라서 작은 a의 경우 Case 1, 큰 a인 경우 식 (3.3)과 같은 결과를 갖게된다. Case 2의 경우는 분산과 평균이 같은 반면 균일분포의 경우 a가 증가할수록 빠르게 분산이 증가하기 때문에 이러한 조건에서 얻어진 결과를 살펴보기 위한 모의실험이 Case 3이다.

Case 4: $f_z^{NRA} = n/n_0$인 호주 경우이다.

3.2. 이상점 보정에서의 민감도 분석

이상점 보정인자는 이상점 자료를 위한 이상점 보정인자와 이상점이 아닌 자료를 위한 이상점 보정인자가 있다. 이상점 자료를 위한 이상점 보정인자는 4가지 Cases에 같은 값을 주기 때문에 특별한 내용이 없다. 따라서 이 절에서는 이상점이 아닌 자료를 위한 이상점 보정인자를 살펴보았다. 본 논문에서는 이상점 보정인자만의 민감도를 살펴보기 위해 무응답 보정인자는 "1"로 가정하였다. 따라서 식 (2.4)은 다음과 같이 된다.

$$f_i^{OAF(no)} = \frac{\mathrm{ANOF}_1}{\mathrm{ANOF}_2} = \frac{\displaystyle\sum_h w_j y_j^D - \sum_j y_j^{D,out}}{\displaystyle\sum_{h_0} w_j y_j^D - \sum_j w_j y_j^{D,out}}. \tag{3.4}$$

이제 민감도를 살펴보기 위해 다음을 고려하자.

$$f_z^{OAF(no)} = \frac{\displaystyle\sum_h w_j z_j - \sum_j z_j^{out}}{\displaystyle\sum_{h_0} w_j z_j - \sum_j w_j z_j^{out}}$$

$$
= \frac{\left(\sum_h w_j x_j - \sum_j x_j^{out}\right) + \left(\sum_h w_j y_j^D - \sum_j y_j^{D,out}\right)}{\left(\sum_{h_0} w_j x_j - \sum_j w_j x_j^{out}\right) + \left(\sum_{h_0} w_j y_j^D - \sum_j w_j y_j^{D,out}\right)}. \tag{3.5}
$$

Case 1: $x_i = 0$이므로 $f_z^{NRA} = f^{NRA}$인 미국 BLS와 같게 된다.

Case 2: $X_i \sim \text{Poi}(\lambda)$이므로 3.1절과 같이

I) λ가 작은 경우

$$
f_z^{OAF(no)} \approx \frac{\sum_h w_j y_j^D - \sum_j y_j^{D,out}}{\sum_{h_0} w_j y_j^D - \sum_j w_j y_j^{D,out}} \tag{3.6}
$$

가 되어 Case 1의 결과가 된다. 또한

II) λ가 큰 경우

$$
f_z^{OAF(no)} \approx \frac{\sum_h w_j x_j - \sum_j x_j^{out}}{\sum_{h_0} w_j x_j - \sum_j w_j x_j^{out}} \tag{3.7}
$$

이 된다. 이제 $x_j = \lambda + e_j$라 하자. 그러면 식 (3.8)은

$$
f_z^{OAF(no)} \approx \frac{\left(\sum_h w_j \lambda - \sum_{j,out} \lambda\right) + \left(\sum_h w_j e_j - \sum_j e_j^{out}\right)}{\left(\sum_{h_0} w_j \lambda - \sum_{j,out} w_j \lambda\right) + \left(\sum_{h_0} w_j e_j - \sum_j w_j e_j^{out}\right)} \tag{3.8}
$$

이 된다. 따라서 큰 λ에 대해서

$$
f_z^{OAF(no)} \approx \frac{\sum_h w_j \lambda - \sum_{j,out} \lambda}{\sum_{h_0} w_j \lambda - \sum_{j,out} w_j \lambda} = \frac{\sum_h w_j - \sum_{j,out} 1}{\sum_{h_0} w_j - \sum_{j,out} w_j} \tag{3.9}
$$

이 되어 Case 4와 같은 결과를 갖게 된다.

Case 3: Case 2와 같이 작은 a인 경우 Case 1, 큰 a인 경우는 Case 4와 같은 결과를 갖게 된다.

Case 4: 식 (3.4)에서 y_i^D에 "1"을 대입한 식 (3.9)가 된다. 따라서 호주 결과와 같게된다.

4. 모의실험

4.1. 자료 설명

3절의 이론적 민감도 분석 결과를 위한 모의실험이 이루어졌다. 모의실험에 사용된 자료는 2007년 매월노동통계 자료이다. 이 자료는 약 7,000개로 이루어졌으며 매월 월별 상용근로자수, 월별 1인당 평균 근로시간, 월별 1인당 평균 총임금 등을 조사한다. 본 모의실험에서는 이 자료를 모집단으로 가정하고 분석하였다. 민감도 분석을 위한 네가지 Cases를 살펴보기 위해 생성된 자료는 다음과 같다. 먼저 Case 1은 표본 설계시에 존재하는 상용근로자수 Y_i^D를 그대로 사용한다. Case 2를 위한 설계는 다음과 같다. 먼저 약 7,000개의 자료 X_{POI}를 평균이 λ인 포아송분포, Poi(λ)를 이용하여 생성한다. 여기서 λ = 30, 50, 100, 300 그리고 1,000을 사용하였다. 다음으로 생성된 자료 X_{POI}를 원자료 Y^D에 더한 Z_{POI}를 만든다. 다음으로 Case 3을 위하여 약 7,000개 자료 X_{UNI}를 균일분포, Uni(0, a)를 이용하여 생성한다. 여기서 $a = 100$, 1,000, 5,000, 10,000 그리고 20,000을 사용하였다. 생성된 자료 X_{UNI}를 원자료 Y^D에 더한 Z_{UNI}를 만든다. Cases 4를 위하여 $Z_{AU} = 1$도 생성한다.

이론적으로 $\mathrm{Cov}(y_i^D, z) = \mathrm{Cov}(y_i^D, y_i^D + x)$에서 x와 y_i^D가 독립이므로 이 값은 $\mathrm{Var}(y_i^D)$가 된다. 또한 $\mathrm{Var}(z) = \mathrm{Var}(y_i^D) + \mathrm{Var}(x)$이다. 따라서

$$\begin{aligned}\rho_{y_i^D, z} &= \frac{\mathrm{Cov}(y_i^D, z)}{\sqrt{\mathrm{Var}(y_i^D)}\sqrt{\mathrm{Var}(z)}} \\ &= \frac{\mathrm{Var}(y_i^D)}{\sqrt{\mathrm{Var}(y_i^D)}\sqrt{\mathrm{Var}(y_i^D) + \mathrm{Var}(x)}} \\ &= \sqrt{\frac{\mathrm{Var}(y_i^D)}{\mathrm{Var}(y_i^D) + \mathrm{Var}(x)}}\end{aligned}$$

가 된다. 자료에서 $\hat{\mathrm{Var}}(y_i^D) = 432,586$으로 추정되었고, $x \sim$ Poi(λ), $x \sim$ Uni(0, a)에서 분산이 쉽게 계산되므로 $x \sim$ Uni(0, 10,000)과 $x \sim$ Uni(0, 20,000)일 때 상관계수가 각각 0.22와 0.11이 되는 것을 쉽게 알 수 있다. 따라서 본 논문에서는 상관계수가 약 0.1에서 0.999까지인 경우를 살펴보았다.

이렇게 생성된 자료를 합쳐서 한 파일에 저장하였다. 이는 표본 추출시 같은 사업체 자료가 추출되어 비교가 의미를 가질 수 있도록 하기 위함이다.

무응답 보정 방법과 이상점 보정 방법의 민감도 분석을 위하여 무응답 자료와 이상점 자료를 모집단 자료에 생성하였다. 먼저 무응답 보정 방법을 위하여 모집단 자료에서 20%를 선택하였다. 이 중에서 랜덤으로 10%는 무응답으로 나머지는 90%는 폐업으로 정한 후 이를 포함하는 새로운 모집단을 만들었다. 따라서 새로운 모집단에는 무응답과 폐업이 정해진 비율만큼 들어있다. 이렇게 만든 새로운 모집단 자료에서 2,000개의 자료를 랜덤으로 추출하였다. 그러면 약 $2{,}000 \times 0.2 \times 0.1$인 40개의 무응답 사업체가, 약 $2{,}000 \times 0.2 \times 0.9$인 360개의 폐업인 사업체가 그리고 나머지 약 1,600개인 정상적인 자료가 추출된다. 같은 방법으로 이 모집단에서 2,000개의 표본을 3,000번 반복하여 추출한 후

표 4.1: 가중치(Poi)

무응답비율	Case 1	Poi(30)	Poi(50)	Poi(100)	Poi(300)	Poi(1000)	Case 4
				평균 가중치(Poi)			
2%	3.6192	3.6187	3.6185	3.6182	3.6177	3.6174	3.6173
6%	3.7996	3.7981	3.7974	3.7963	3.7948	3.7940	3.7936
10%	3.9782	3.9761	3.9752	3.9738	3.9718	3.9709	3.9705
14%	4.1532	4.1509	4.1500	4.1486	4.1470	4.1465	4.1465
18%	4.3304	4.3277	4.3267	4.3251	4.3233	4.3229	4.3231
				가중치 분산(Poi)			
2%	0.0045	0.0031	0.0025	0.0016	0.0006	0.0002	0.0001
6%	0.0143	0.0098	0.0080	0.0052	0.0019	0.0007	0.0004
10%	0.0251	0.0174	0.0141	0.0093	0.0034	0.0012	0.0008
14%	0.0335	0.0231	0.0187	0.0122	0.0044	0.0016	0.0011
18%	0.0460	0.0318	0.0258	0.0169	0.0062	0.0022	0.0015

결과를 비교하였다. 새롭게 만들어진 모집단에서 무응답의 비율과 폐업의 비율을 (30%, 70%), (50%, 50%), (70%, 30%), (90%, 10%)와 같이 조정하여 무응답과 폐업의 비율에 의한 결과도 함께 살펴보았다. 따라서 전체 자료에 대한 무응답 비율은 2%, 6%, 10%, 14% 그리고 18%이다. 다음으로 이상점 보정 방법을 위해서는 먼저 7,000개 자료에서 2,000개 자료를 추출한다. 이 2,000개 자료 중에서 3%의 자료를 추출한다. 추출한 자료의 자료 값에 4를 곱하여 이상점을 만든다. 이 같은 실험을 3,000번 반복하여 실시하였다. 또한 이상점의 비율을 3% 이외에 5% 그리고 10%도 함께 살펴보았다.

4.2. 비교 통계량

각 Case의 결과를 비교하기 위해 사용된 비교 통계량은 Rao (2003)에서 비교를 위해 제안된 통계량을 사용하였으며 정의는 다음과 같다.

$$\mathrm{MSE} = \frac{1}{3000}\sum_{k}^{3000}\left(t_{B,k} - \hat{t}_{B,k}\right)^2, \qquad \mathrm{RB} = \frac{1}{3000}\sum_{k}^{3000}\frac{t_{B,k} - \hat{t}_{B,k}}{t_{B,k}},$$

$$\mathrm{RMSE} = \frac{1}{3000}\sum_{k}^{3000}\left(\frac{t_{B,k} - \hat{t}_{B,k}}{t_{B,k}}\right)^2, \qquad \mathrm{ARE} = \frac{1}{3000}\sum_{k}^{3000}\left|\frac{t_{B,k} - \hat{t}_{B,k}}{t_{B,k}}\right|,$$

$$\mathrm{RRMSE} = \left\{\frac{1}{3000}\sum_{k}^{3000}\left(\frac{t_{B,k} - \hat{t}_{B,k}}{t_{B,k}}\right)^2\right\}^{\frac{1}{2}}, \qquad \mathrm{Ratio} - \mathrm{MSE} = \frac{\mathrm{MSE}_*}{\mathrm{MSE}_{CASE1}},$$

여기서 $t_{B,k}$는 모집단 총계이고, $\hat{t}_{B,k}$는 k번째 반복에서 구해진 총계 추정치이다.

4.3. 무응답 보정

식 (3.2)를 이용하여 얻어진 무응답 보정인자를 식 (2.6)에 대입하고, f_i^{REAG}, f_i^{OAF}, f_i^{BMF}를 모두 "1"로 하여 구한 최종 가중치의 평균과 분산을 표 4.1에 정리하였다. 모든

표 4.2: 평균 NAR-Factor(Poi)

무응답비율	Case 1	Poi(30)	Poi(50)	Poi(100)	Poi(300)	Poi(1000)	Case 4
2%	1.0256	1.0254	1.0254	1.0253	1.0251	1.0251	1.0250
6%	1.0767	1.0762	1.0761	1.0758	1.0753	1.0751	1.0750
10%	1.1273	1.1267	1.1264	1.1260	1.1255	1.1252	1.1251
14%	1.1769	1.1762	1.1760	1.1756	1.1751	1.1750	1.1750
18%	1.2271	1.2263	1.2260	1.2256	1.2251	1.2250	1.2250

표 4.3: 각 방법의 비교 결과(Poi)

무응답비율	Case 1	Poi(30)	Poi(50)	Poi(100)	Poi(300)	Poi(1000)	Case 4
				MSE(Poi)			
2%	6,183E6	6,175E6	6,177E6	6,187E6	6,233E6	6,287E6	6,333E6
6%	6,755E6	6,773E6	6,797E6	6,870E6	7,073E6	7,296E6	7,470E6
10%	7,113E6	7,114E6	7,144E6	7,248E6	7,577E6	7,954E6	8,253E6
14%	7,154E6	7,194E6	7,254E6	7,421E6	7,909E6	8,445E6	8,865E6
18%	7,244E6	7,297E6	7,377E6	7,600E6	8,253E6	8,970E6	9,529E6
				RB(Poi)			
2%	0.0025	0.0026	0.0026	0.0027	0.0027	0.0028	0.0028
6%	−0.0005	−0.0002	−0.0002	0.0001	0.0001	0.0002	0.0002
10%	−0.0021	−0.0018	−0.0017	−0.0016	−0.0014	−0.0015	−0.0015
14%	0.0016	0.0018	0.0019	0.0019	0.0018	0.0016	0.0014
18%	0.0009	0.0011	0.0012	0.0012	0.0010	0.0007	0.0003
				RMSE(Poi)			
2%	0.0085	0.0085	0.0085	0.0086	0.0086	0.0087	0.0088
6%	0.0085	0.0085	0.0086	0.0086	0.0089	0.0092	0.0094
10%	0.0082	0.0082	0.0082	0.0083	0.0087	0.0091	0.0095
14%	0.0075	0.0076	0.0076	0.0078	0.0083	0.0089	0.0093
18%	0.0070	0.0071	0.0072	0.0074	0.0080	0.0087	0.0093
				ARE(Poi)			
2%	0.0743	0.0743	0.0743	0.0744	0.0746	0.0749	0.0752
6%	0.0742	0.0743	0.0744	0.0748	0.0758	0.0769	0.0777
10%	0.0727	0.0727	0.0727	0.0731	0.0745	0.0763	0.0777
14%	0.0696	0.0698	0.0702	0.0710	0.0733	0.0757	0.0775
18%	0.0675	0.0677	0.0681	0.0691	0.0720	0.0751	0.0775
				RRMSE(Poi)			
2%	0.0926	0.0925	0.0925	0.0926	0.0930	0.0934	0.0937
6%	0.0923	0.0924	0.0926	0.0931	0.0944	0.0959	0.0971
10%	0.0905	0.0905	0.0907	0.0914	0.0934	0.0957	0.0975
14%	0.0869	0.0871	0.0875	0.0885	0.0914	0.0944	0.0967
18%	0.0839	0.0842	0.0847	0.0859	0.0896	0.0934	0.0962
				Ratio-MSE(Poi)			
2%	1.0000	0.9989	0.9990	1.0008	1.0081	1.0170	1.0244
6%	1.0000	1.0027	1.0062	1.0170	1.0470	1.0800	1.1059
10%	1.0000	1.0002	1.0045	1.0190	1.0653	1.1183	1.1603
14%	1.0000	1.0057	1.0140	1.0373	1.1056	1.1804	1.2391
18%	1.0000	1.0073	1.0184	1.0492	1.1393	1.2383	1.3155

표 4.4: 가중치(Uni)

무응답비율	Case 1	Uni(100)	Uni(1000)	Uni(5000)	Uni(10000)	Uni(20000)	Case 4
			평균 가중치(Uni)				
2%	3.6193	3.6186	3.6174	3.6173	3.6173	3.6171	3.6172
6%	3.7979	3.7963	3.7942	3.7944	3.7938	3.7938	3.7939
10%	3.9745	3.9723	3.9708	3.9703	3.9704	3.9700	3.9707
14%	4.1573	4.1532	4.1482	4.1468	4.1482	4.1474	4.1473
18%	4.3364	4.3312	4.3247	4.3229	4.3238	4.3225	4.3228
			가중치 분산(Uni)				
2%	0.0045	0.0026	0.0004	0.0002	0.0002	0.0002	0.0001
6%	0.0140	0.0079	0.0012	0.0007	0.0007	0.0007	0.0005
10%	0.0230	0.0130	0.0021	0.0012	0.0012	0.0012	0.0008
14%	0.0362	0.0201	0.0031	0.0017	0.0016	0.0016	0.0011
18%	0.0512	0.0289	0.0046	0.0024	0.0022	0.0023	0.0016

표 4.5: 평균 NAR-Factor(Uni)

무응답비율	Case 1	Uni(100)	Uni(1000)	Uni(5000)	Uni(10000)	Uni(20000)	Case 4
2%	1.0256	1.0254	1.0251	1.0250	1.0250	1.0250	1.0250
6%	1.0762	1.0757	1.0752	1.0752	1.0750	1.0751	1.0751
10%	1.1262	1.1256	1.1252	1.1250	1.1251	1.1250	1.1252
14%	1.1781	1.1769	1.1755	1.1751	1.1755	1.1752	1.1752
18%	1.2288	1.2273	1.2255	1.2250	1.2252	1.2249	1.2250

경우에 있어 평균가중치는 일정한 것으로 판단되나, 가중치의 분산은 λ가 커질 수록 작아지는 것을 알 수 있다. Case 4의 경우는 추출된 자료의 수가 랜덤으로 추출되어 그 값이 일정하지 않지만 크게 변하지는 않기 때문에 분산이 작게 나타났다. 또한 표 4.2는 무응답 보정 인자의 평균 값이며 표 4.1의 평균 가중치와 같이 거의 일정하다.

표 4.3은 제안된 여러 비교 통계량을 이용하여 각각의 Case별 우수성을 비교하였다. 총계 추정에 가장 많은 정보를 사용한 경우인 Case 1의 MSE가 가장 좋은 것으로 나타났다. λ가 30에서 1,000으로 커질수록 MSE도 커지고 있으며 아무런 보조정보가 없는 경우인 Case 4가 가장 큰 MSE를 보여 주고 있다. MSE 뿐만 아니라 다른 비교 통계량도 같은 결론을 나타내고 있다. 무응답 비율이 증가할수록 Case 1의 효율성이 높으며 Ratio-MSE를 보면 무응답 비율이 18%인 경우 그 비율은 1.3155에 이르는 것을 확인할 수 있다. 그러나 모집단 총계와 총계 추정치와의 차이인 편향(Bias)은 상대편향인 RB를 이용하여 확인할 수 있으며 매우 숫자가 작아 편향은 존재하지 않은 것으로 판단된다.

표 4.4와 4.5는 Case 1과 Case 3 그리고 Case 4를 비교한 결과로 가중치의 평균과 무응답 보정인자의 평균은 모든 경우에서 일정한 것으로 나타났다. 가중치의 분산은 Case 1과 Case 3 그리고 Case 4의 순으로 줄어들고 있으며 Case 2에 비해 Case 3이 Case 4의 결과에 더욱 근접한 것을 알 수 있다. 이는 Case 2의 포아송 분포에 사용된 분산보다 Case 3의 균일 분포에 사용된 분산이 크기 때문인 것으로 판단된다.

표 4.6: 각 방법의 비교 결과(Uni)

무응답비율	Case 1	Uni(100)	Uni(1000)	Uni(5000)	Uni(10000)	Uni(20000)	Case 4
				MSE(Uni)			
2%	6,385E6	6,390E6	6,502E6	6,571E6	6,586E6	6,590E6	6,592E6
6%	6,731E6	6,737E6	7,062E6	7,263E6	7,311E6	7,320E6	7,331E6
10%	6,814E6	6,932E6	7,666E6	8,048E6	8,078E6	8,160E6	8,160E6
14%	7,298E6	7,423E6	8,420E6	9,011E6	9,094E6	9,099E6	9,150E6
18%	7,755E6	7,854E6	9,099E6	9,922E6	10,060E6	10,165E6	10,166E6
				RB(Uni)			
2%	−0.0020	−0.0019	−0.0018	−0.0018	−0.0018	0.0017	−0.0018
6%	0.0011	0.0014	0.0015	0.0013	0.0014	0.0014	0.0014
10%	0.0018	0.0020	0.0016	0.0015	0.0015	0.0015	0.0013
14%	−0.0008	−0.0003	−0.0001	−0.0001	−0.0005	−0.0003	−0.0003
18%	−0.0042	−0.0036	−0.0034	−0.0034	−0.0036	−0.0034	−0.0035
				RMSE(Uni)			
2%	0.0088	0.0088	0.0090	0.0091	0.0091	0.0091	0.0091
6%	0.0085	0.0085	0.0089	0.0091	0.0092	0.0092	0.0092
10%	0.0078	0.0080	0.0088	0.0093	0.0093	0.0094	0.0094
14%	0.0077	0.0078	0.0089	0.0095	0.0096	0.0096	0.0097
18%	0.0075	0.0076	0.0088	0.0096	0.0098	0.0099	0.0099
				ARE(Uni)			
2%	0.0755	0.0754	0.0760	0.0763	0.0764	0.0764	0.0764
6%	0.0744	0.0743	0.0759	0.0769	0.0772	0.0772	0.0773
10%	0.0714	0.0721	0.0759	0.0776	0.0779	0.0781	0.0781
14%	0.0697	0.0706	0.0754	0.0780	0.0784	0.0784	0.0787
18%	0.0689	0.0694	0.0752	0.0788	0.0793	0.0797	0.0798
				RRMSE(Uni)			
2%	0.0942	0.0942	0.0950	0.0955	0.0956	0.0957	0.0957
6%	0.0922	0.0923	0.0944	0.0958	0.0961	0.0962	0.0962
10%	0.0885	0.0893	0.0939	0.0962	0.0964	0.0969	0.0969
14%	0.0878	0.0886	0.0943	0.0976	0.0980	0.0981	0.0983
18%	0.0868	0.0874	0.0940	0.0982	0.0989	0.0994	0.0994
				Ratio-MSE(Uni)			
2%	1.0000	1.0008	1.0183	1.0292	1.0315	1.0321	1.0325
6%	1.0000	1.0010	1.0492	1.0790	1.0862	1.0875	1.0891
10%	1.0000	1.0174	1.1251	1.1812	1.1855	1.1975	1.1975
14%	1.0000	1.0172	1.1537	1.2347	1.2461	1.2468	1.2537
18%	1.0000	1.0128	1.1733	1.2795	1.2972	1.3108	1.3109

표 4.6은 표 4.3과 같은 결과를 주고 있다. 균일 분포의 $a = 10{,}000,\ 20{,}000$인 경우 Case 4의 결과와 거의 일치하고 있다. 실제로 Uni(10,000)과 Uni(20,000)을 사용할 때 모의실험에서 얻은 y_i^D와 z와의 상관계수 추청량은 0.22와 0.11로 나타났다. 따라서 상관관계가 낮으면 Case 4와 같은 결과를 주는 것을 확인할 수 있다.

표 4.7: 평균 Non-Outlier-Factor(Poi)

이상점비율	Case 1	Poi(30)	Poi(50)	Poi(100)	Poi(300)	Poi(1000)	Case 4
3%	1.0221	1.0221	1.0221	1.0221	1.0221	1.0221	1.0222
5%	1.0386	1.0384	1.0383	1.0381	1.0379	1.0378	1.0377
10%	1.0794	1.0793	1.0793	1.0793	1.0794	1.0795	1.0796

표 4.8: 각 방법의 비교 결과(Poi)

이상점비율	Case 1	Poi(30)	Poi(50)	Poi(100)	Poi(300)	Poi(1000)	Case 4
				MSE(Poi)			
3%	7,222E6	7,223E6	7,228E6	7,243E6	7,290E6	7,344E6	7,387E6
5%	7,414E6	7,427E6	7,441E6	7,475E6	7,571E6	7,673E6	7,752E6
10%	8,060E6	8,093E6	8,121E6	8,189E6	8,376E6	8,573E6	8,724E6
				RB(Poi)			
3%	0.0030	0.0029	0.0029	0.0029	0.0028	0.0027	0.0027
5%	−0.0021	−0.0020	−0.0019	−0.0019	−0.0018	−0.0017	−0.0017
10%	−0.0015	−0.0016	−0.0016	−0.0017	−0.0020	−0.0022	−0.0024
				RMSE(Poi)			
3%	0.0064	0.0064	0.0064	0.0064	0.0065	0.0065	0.0066
5%	0.0063	0.0063	0.0064	0.0064	0.0065	0.0066	0.0067
10%	0.0064	0.0064	0.0065	0.0065	0.0067	0.0068	0.0069
				ARE(Poi)			
3%	0.0643	0.0643	0.0644	0.0645	0.0647	0.0649	0.0651
5%	0.0636	0.0637	0.0638	0.0640	0.0645	0.0650	0.0654
10%	0.0641	0.0642	0.0643	0.0646	0.0654	0.0661	0.0666
				RRMSE(Poi)			
3%	0.0801	0.0801	0.0801	0.0802	0.0805	0.0808	0.0810
5%	0.0797	0.0798	0.0799	0.0801	0.0806	0.0811	0.0815
10%	0.0800	0.0802	0.0803	0.0806	0.0815	0.0825	0.0832
				Ratio-MSE(Poi)			
3%	1.0000	1.0001	1.0008	1.0029	1.0095	1.0170	1.0228
5%	1.0000	1.0018	1.0036	1.0083	1.0213	1.0351	1.0457
10%	1.0000	1.0041	1.0076	1.0160	1.0392	1.0636	1.0824

4.4. 이상점 보정

Case 1과 포아송 분포를 이용한 Case 2 그리고 Case 4를 비교한 표 4.7를 살펴보면 이상점 보정값이 Case에 상관없이 일정한 것을 확인할 수있다. 또한 표 4.8의 비교통계 결과를 살펴보면 무응답 보정 결과와 같은 추세를 유지하고 있다. 즉 λ가 커질수록 MSE 등 비교 통계량 값이 커지고 있으며, 상대 편향은 거의 "0"에 가까운 결과를 주고 있다.

Case 1과 균일 분포를 이용한 Case 3 그리고 Case 4를 비교한 결과를 표 4.9와 4.10에 정리하였다. 그 내용은 포아송 분포를 이용한 결과와 유사한 것을 확인할 수 있다.

표 4.9: 평균 Non-Outlier-Factor(Uni)

이상점비율	Case 1	Uni(100)	Uni(1000)	Uni(5000)	Uni(10000)	Uni(20000)	Case 4
3%	1.0221	1.0221	1.0221	1.0222	1.0220	1.0221	1.0222
5%	1.0386	1.0383	1.0379	1.0378	1.0379	1.0378	1.0377
10%	1.0794	1.0793	1.0795	1.0795	1.0797	1.0797	1.0796

표 4.10: 각 방법의 비교 결과(Uni)

이상점비율	Case 1	Uni(100)	Uni(1000)	Uni(5000)	Uni(10000)	Uni(20000)	Case 4
				MSE(Uni)			
3%	7,222E6	7,232E6	7,309E6	7,353E6	7,376E6	7,399E6	7,387E6
5%	7,414E6	7,441E6	7,637E6	7,705E6	7,725E6	7,747E6	7,752E6
10%	8,060E6	8,128E6	8,492E6	8,686E6	8,708E6	8,692E6	8,724E6
				RB(Uni)			
3%	0.0030	0.0029	0.0028	0.0027	0.0028	0.0027	0.0027
5%	−0.0021	−0.0019	−0.0018	−0.0018	−0.0018	−0.0018	−0.0017
10%	−0.0015	−0.0016	−0.0022	−0.0023	−0.0024	−0.0025	−0.0024
				RMSE(Uni)			
3%	0.0064	0.0064	0.0065	0.0065	0.0066	0.0066	0.0066
5%	0.0064	0.0064	0.0065	0.0066	0.0066	0.0066	0.0066
10%	0.0064	0.0066	0.0067	0.0069	0.0069	0.0069	0.0069
				ARE(Uni)			
3%	0.0643	0.0644	0.0648	0.0650	0.0651	0.0652	0.0651
5%	0.0636	0.0638	0.0649	0.0652	0.0653	0.0654	0.0654
10%	0.0641	0.0644	0.0658	0.0664	0.0666	0.0665	0.0666
				RRMSE(Uni)			
3%	0.0801	0.0801	0.0806	0.0808	0.0809	0.0810	0.0810
5%	0.0797	0.0799	0.0809	0.0813	0.0814	0.0815	0.0815
10%	0.0800	0.0803	0.0821	0.0830	0.0831	0.0831	0.0832
				Ratio-MSE(Uni)			
3%	1.0000	1.0014	1.0120	1.0181	1.0213	1.0245	1.0228
5%	1.0000	1.0037	1.0302	1.0393	1.0420	1.0449	1.0457
10%	1.0000	1.0085	1.0536	1.0776	1.0803	1.0784	1.0824

5. 결론

본 논문에서는 표본 설계시에 갖고 있는 보조변수가 총계 추정에 어떤 영향을 미치는지 살펴보았다. 모의실험은 표본설계 당시의 사업체 상용 근로자수가 월별 총계 추정에 미치는 영향 분석에 초점을 맞추었다. 즉 노동부가 발표하는 매월노동통계의 경우, 월별 상용근로자수, 월별 1인당 평균 총임금, 월별 1인당 평균 근로시간 등을 조사하여 발표한다. 이 경우, 표본 설계의 표본 틀에 있는 종사자수는 월별 상용근로자수 추정에 영향을 줄 수 있다. 이는 이들 두 자료 사이에 높은 상관관계가 있기 때문이다. 반면에 모집단 종사자수와 월별 평균 임금총액 또는 모집단 종사자수와 1인당 평균 근로시간 사이에는 높은 상관

관계가 있지 않다. 그러나 한번 결정된 가중치는 월별 상용근로자수, 월별 1인당 평균 총임금, 월별 1인당 평균 근로시간 등에 똑같이 적용된다. 따라서 2장에서 사용한 보정 방법이 각 변수에 어느 정도 영향을 주는지, 특히 보정 공식에 사용한 y_i^D의 영향력을 분석하는 것은 매우 중요하다고 하겠다. 3장의 이론적인 민감도 분석과 4장의 모의실험 결과를 종합해 보면 관심 변수의 총계 추정과 관계가 높은 보조변수를 사용하여 보정을 할 경우 좋은 결과를 얻을 수 있음을 확인할 수 있었다. 즉 표 4.3과 4.6의 Ratio-MSE(Poi)와 Ratio-MSE(Uni)를 살펴보면 보조변수를 사용하지 않은 경우인 Case 4의 무응답 비율이 18%일 때 Case 1에 비해 약 30%의 MSE가 증가하는 것을 확인할 수 있다. 또한 표 4.8과 4.10의 이상점에 관한 모의실험 결과 중 Ratio-MSE(Poi)와 Ratio-MSE(Uni)를 살펴보면 보조변수를 사용하지 않은 경우인 Case 4의 이상점 비율이 10%일 때 Case 1에 비해 약 8%의 MSE가 증가하는 것을 알 수 있다. 그러나 4.1절의 상관계수 공식을 이용하면 $U(1{,}000)$인 경우 상관계수는 0.915, $U(5{,}000)$인 경우의 상관계수는 0.415이다. 표 4.6을 살펴보면 상관계수가 0.915일 때 Ratio-MSE는 1.1733 그리고 상관계수가 0.415일 때 1.2795이다. 또한 독립일 때 1.3109이다. 즉 상관계수가 0.415일 때와 독립인 경우의 Ratio-MSE가 비슷하게 나타나고 있다. 따라서 보조변수가 있는 경우 보조변수를 사용하는 것이 바람직하지만 높은 상관관계가 없다면 큰 효과를 기대하기는 어려울 것으로 판단된다.

참고문헌

김규성 (2000). 무응답 대체 방법과 대체 효과, <한국조사연구학회>, **1**, 1–14.

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

이상은 (2008). 표본조사에 따른 추정방법 비교: 가중치 조정기법을 중심으로, <응용통계연구>, **21**, 413–427.

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

Burdete, T. (2003). Survey of Occupational Injuries and Illnesses, Sample Design, *Bureau of Labor Statistics.*

Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.

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

Oh, H. L. and Scheuren, F. (1987). Modified raking ratio estimation, *Survey Methodology*, **13**, 209–219.

Rao, J. N. K. (2003). *Small Area Estimation*, John Wiley & Sons, New York.

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

[2008년 7월 접수, 2008년 8월 채택]

A Study on the Sensitivity of the BLS Methods†

Seok-Jin Lee[1)], Key-Il Shin[2)]

Abstract

BLS adjustment methods have been able to provide more accurate estimates of total and make samples represent population characteristics by post-adjustment of design weights of samples. However, BLS methods use additional data, for instance number of employee, without this information or using other information, give different weight adjustment factors. In this paper we studied the sensitivity of the variables used in BLS adjustment. The 2007 monthly labor survey data is used in analysis.

Keywords: Sampling survey; weight; non-response adjustment; outlier adjustment.

† This paper was supported by the Hankuk University of Foreign Studies research fund 2008.

1) Graduate student, Department of statistics, Hankuk University of Foreign Studies, Yonginsi, Kyonggy 449-791, Korea.

2) Professor, Department of statistics, Hankuk University of Foreign Studies, Yonginsi, Kyonggy 449-791, Korea. Correspondence: keyshin@hufs.ac.kr