한국통계학회논문집
2011, 18권, 3호, 367–376

DOI: 10.5351/CKSS.2011.18.3.367

수정된 BLS 가중치보정법

박정준[a], 조기종[a], 이상은[b], 신기일[1,c]

[a]고려대학교 환경생태공학부, [b]경기대학교 응용정보통계학과, [c]한국외국어대학교 통계학과

요 약

BLS 가중치보정법은 사업체 조사 시 발생한 무응답 및 이상점을 처리하기 위해 사용하는 가중치 보정방법중의 하나이다. 최근의 연구에 의하면 총계 추정에 있어 BLS 무응답 가중치보정법의 결과가 비추정법을 사용한 대체 결과와 일치하는 것으로 알려졌다. 본 논문에서는 이상점과 무응답이 동시에 있는 경우, BLS 무응답 가중치보정법을 비추정 대체법으로 바꾸어 총계를 추정하는 새로운 방법을 제안하였다. 매월 노동통계 자료를 이용한 모의실험을 통하여 제안된 방법의 우수성을 확인하였다.

주요용어: 회귀대체법, 무응답 보정, 무응답 대체, 이상점 가중치 보정.

1. 서론

표본설계 시 고려해야할 가장 중요한 내용은 모집단을 잘 대표하는 표본을 추출하는 것과 이에 맞는 추정량을 결정하는 것이다. 그러나 대표성을 갖는 표본이라 하여도 실사 중에 무응답이 발생하거나 이상점이 발생한 경우에는 이의 영향력을 최소로 할 수 있는 타당한 추정 방법을 사용하여야 한다. 사업체 조사에서 무응답이 발생하거나 이상점이 발생한 경우에 사용하는 방법 중의 하나가 BLS 가중치보정법이다 (Burdete, 2003). 이상은 (2008), 이석진과 신기일 (2008), 김석과 신기일 (2009)은 BLS 가중치보정법의 효율성을 비교하여 우수성을 확인하였다. BLS 가중치보정법의 핵심 내용은 무응답 보정과 이상점 보정이다. 최근 이상은과 신기일 (2010)은 총계 추정 시 BLS 무응답 가중치보정법이 비추정(ratio estimation) 대체와 일치하는 것을 밝혔다. 또한 무응답 효과를 줄이기 위해 결측치를 직접 대체하는 무응답 대체법에 관한 많은 이론이 개발되었으며 Little (1987)과 Little과 Rubin (2002)은 이에 관한 내용을 정리하였다.

본 논문에서는 BLS 가중치보정법의 무응답 보정과 이상점 보정을 사용하여 얻은 총계 추정량과 비추정 대체법과 회귀대체법을 사용하여 무응답을 대체한 후 이상점 보정을 실시하여 얻은 총계 추정량의 우수성을 비교하였다. 추가로 회귀대체법을 고려한 이유는 비추정 대체에 비해 회귀대체(regression imputation)가 특정한 조건하에서 우수한 것이 이미 잘 알려져 있기 때문이다 (신민웅과 이상은, 2001). 본 연구에서는 결측 메카니즘과 이상점 발생 메카니즘은 MAR(missing at random)을 따른다고 가정하였다.

본 논문의 구성은 먼저 2절에서 BLS 무응답 가중치보정법과 이상점 보정법을 이용한 총계 추정법을 간단히 설명하였으며 비추정 대체법, 회귀대체법과 이상점 보정법을 접목한 새로운 총계 추정법을 제안하였다. 3절에서는 모의실험을 통하여 제안된 방법의 우수성을 살펴보았으며 4절에 결론이 있다.

이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2008-313-C00141).

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

2. 총계추정량

2.1. BLS 보정법을 이용한 총계 추정량

BLS 가중치보정법은 미국 노동통계국이 사업체 조사 시 표본 틀의 자료와 실사에서 얻어진 자료의 상관관계가 높은 경우에 유용하게 사용하는 방법으로 무응답과 이상점이 랜덤하게 발생(missing at random; MAR)한 경우에 사용하는 방법이다. BLS 가중치보정법은 집계 후 보정, 무응답 보정, 이상점 보정, 밴치 마크 보정 등 여러 보정방법을 결합하여 사용하는 방법이다. Burdete (2003)은 미국 노동통계국에서 조사하고 있는 Survey of Occupational Injuries and Illnesses 자료를 이용하여 각각의 보정법을 자세히 설명하였다. 본 연구에서는 이 중에서 중요하게 사용되는 두 보정방법인 무응답 보정과 이상점 보정을 이용하여 최종 가중치를 구하는 방법을 살펴보았다. 이때 각 층별로 보정법이 적용되기 때문에 본 논문에서는 하나의 층이 있는 경우만을 고려하였다.

2.1.1. 무응답 보정(Non-response adjustment; NRA)

실사에서는 무응답이 흔히 발생하고 그 영향력이 매우 크기 때문에 NRA는 가장 많은 관심을 받고 있다. 이를 위한 BLS 무응답 가중치보정 공식은 다음과 같다.

$$\text{NRA factor} = f_i^{NRA} = \frac{\sum_{j=1}^{n} w_j y_j^D}{\sum_{i=1}^{r} w_i y_i^D},$$

여기서 n은 표본 수이며 실제 조사된 유효 표본 수는 r이다. 또한 y_i^D는 표본설계 당시의 표본 틀 자료값이며 w_i는 표본설계 당시 구한 설계가중치(design weight)이다. 일반적으로 같은 층에서는 같은 설계가중치를 갖고 있으므로 $w_i = w$, $i = 1, \ldots, n$이 되어

$$\text{NRA factor} = f_i^{NRA} = \frac{\sum_{j=1}^{n} y_j^D}{\sum_{i=1}^{r} y_i^D} = f^{NRA}$$

가 된다.

2.1.2. 이상점 보정(Outlier adjustment)

이상점 보정은 특정 자료들이 이상점으로 판정되었을 때 그 영향력을 줄이기 위해 사용하는 가중치 보정방법이다. 이상점으로 판정된 자료의 가중치는 모두 "1"이 되도록 하고 이상점이 아닌 자료의 가중치는 다음의 공식을 이용하여 보정하게 된다. 따라서 보정 인자는 이상점인 자료와 이상점이 아닌 자료에 따라 다르게 적용된다.

이상점인 경우의 이상점 보정인자:

$$f_1^{OAF} = \frac{1}{w \times f^{NRA}}.$$

이상점이 아닌 경우의 이상점 보정인자:

$$f_2^{OAF} = \frac{\text{ANOF}_1 / \text{ANOF}_2}{f^{NRA}},$$

여기서 $\text{ANOF}_1 = \sum_{i=1}^{n} w y_i^D - \sum_{j=1}^{c} y_j^{D,out}$이고 $\text{ANOF}_2 = \sum_{i=1}^{r} w y_i^D - \sum_{j=1}^{c} w y_j^{D,out}$이며 c는 이상점 개수이다. 또한 $y_i^{D,out}$은 이상점에 해당되는 표본설계 당시의 표본 틀 자료값이다.

2.1.3. 최종 가중치(Final weight)

이상의 두 가중치 보정인자가 얻어지게 되면 최종 가중치는 다음과 같다.

$$w_i^{Fin} = w \times f^{NRA} \times f_i^{OAF}.$$

따라서 이상점인 경우의 최종 가중치는 $w_1^{Fin} = 1$이 되고 이상점이 아닌 자료의 경우 최종 가중치는

$$w_2^{Fin} = w \times \left(\frac{\text{ANOF}_1}{\text{ANOF}_2} \right)$$

이 된다.

2.2. 제안된 총계 추정량

2.2.1. 비추정 대체법을 이용한 BLS 가중치보정법의 총계 추정량

이제 가중치를 이용한 총계 추정량 대신 비추정 대체를 사용한 총계 추정량을 살펴보자. 먼저 비추정으로 대체한 무응답 대체값을 $\hat{y}_j^{m(R)}$이라 하자. 그러면 $\hat{y}_j^{m(R)} = \hat{R} y_j^D$이다. 여기서 $\hat{R} = \sum_{i=1}^r y_i^O / \sum_{i=1}^r y_i^D$이고 y_i^O는 조사된 실사 자료 값, y_i^D는 표본설계 당시의 표본 틀 자료값이다. 최근 이상은과 신기일 (2010)은 무응답만이 존재하는 경우 BLS 무응답 가중치보정법이 비추정 대체법과 같다는 것을 증명하였다. 따라서 비추정으로 무응답을 추정하여 대체한다면 무응답에 의한 결측값이 없으므로 무응답 보정인자는 $f^{NRA} = 1$이 된다.

2.2.2. 무응답 회귀 대체(Regression imputation)

본 연구에서는 표본으로 n개가 배정되었으나 유효 표본 r개가 조사된 것을 가정하였다. 따라서 $n - r$개의 무응답 자료가 존재한다. 이 경우 r개의 조사자료, y_i^O와 이에 해당되는 r개의 표본 틀 자료, y_i^D를 이용하여 회귀대체를 할 수 있다. 즉 다음의 단순회귀모형을 이용하여 회귀대체를 실시한다.

$$y_i^O = \beta_0 + \beta_1 y_i^D + \epsilon_i, \quad i = 1, \ldots, r.$$

결국 추정된 회귀계수, $\hat{\beta}_0, \hat{\beta}_1$와 다음 식을 이용하여 결측치를 계산한다.

$$\hat{y}_j^{m(LR)} = \hat{\beta}_0 + \hat{\beta}_1 y_j^D, \quad j = 1, \ldots, n - r,$$

여기서 $\hat{y}_j^{m(LR)}$은 회귀대체법에 의한 결측치 대체값을 의미한다.

2.2.3. 비추정 대체와 회귀대체 비교

비추정 대체를 위한 일반 모형은 다음과 같다.

$$y_i = R x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

여기서 $\epsilon_i \sim iid(0, x_i \sigma^2)$이다. 또한 회귀대체를 위한 일반 모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

여기서 $\epsilon_i \sim iid(0, \sigma^2)$이다. 따라서 두 모형은 기본적으로 다른 형태를 취하고 있다. 즉 회귀대체 모형과 비교하면 비추정 대체 모형의 경우 절편이 없으며 분산이 일정하지 않은 경우에 타당하도록 만

들어졌다. 따라서 어떤 모형을 사용하는 것이 타당한지는 위의 두 모형에 주어진 가정이 자료에 얼마나 타당한가에 달려있다. 비추정 대체의 경우 $\hat{R} = \bar{y}/\bar{x} = \sum_{i=1}^{r} y_i / \sum_{i=1}^{r} x_i$이며 따라서 대체는 $\hat{y}^{m(R)} = \hat{R}x_i$을 이용한다. 또한 $\hat{\beta}_0, \hat{\beta}_1$을 단순회귀추정량이라 할 때 회귀대체는 $\hat{y}^{m(LR)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$을 이용한다. 비추정을 이용한 총계 추정과 회귀추정을 이용한 총계 추정의 경우 다음의 조건하에서는 회귀추정 결과가 우수하다는 것이 알려져 있다 (이상은과 신민웅, 2001). 즉 $-\rho^2 S_y^2 < R^2 S_x^2 - 2R\rho S_y S_x$이면 회귀 추정이 우수하다. 여기서 $R = Y/X$, $\rho = \mathrm{corr}(X_i, Y_i)$, $S_x^2 = 1/(N-1)\sum_{i=1}^{N}(x_i - \bar{X})^2$, $S_y^2 = 1/(N-1)\sum_{i=1}^{N}(y_i - \bar{Y})^2$이고, X, Y는 각 자료의 총계이다.

이에 본 논문에서는 수정된 BLS 가중치보정법을 위하여 비추정 대체에 추가하여 회귀대체도 고려하였다.

2.2.4. 이상점 보정(Outlier adjustment)

2.2.1절과 2.2.2절에서 모든 무응답이 대체되었으므로 무응답 가중치보정은 필요없게 된다. 즉 무응답 보정인자, $f^{NRA} = 1$이 된다. 따라서 이상점 보정인자는 다음의 식이 된다.

- 이상점인 경우의 이상점 보정인자:

$$f^{OAF} = \frac{1}{w}$$

- 이상점이 아닌 경우의 이상점 보정인자:

$$f^{OAF(N)} = \frac{\mathrm{ANOF}_1^{(N)}}{\mathrm{ANOF}_2^{(N)}},$$

여기서 $\mathrm{ANOF}_1^{(N)} = \sum_{i=1}^{n} wy_i^D - \sum_{j=1}^{c} y_j^{D,out}$이고 $\mathrm{ANOF}_2^{(N)} = \sum_{i=1}^{n} wy_i^D - \sum_{j=1}^{c} wy_j^{D,out}$이며 c는 이상점 개수이다. 즉 무응답이 없는 상태이므로 $\mathrm{ANOF}_2^{(N)}$은 ANOF_2와 다른 이상점 보정인자가 구해진다.

2.2.5. 최종 가중치(Final weight)

2.2.1절에서 2.2.4절의 결과를 이용하면 이상점인 경우의 최종 가중치는 "1"이 되고 이상점이 아닌 자료의 경우 최종 가중치는

$$w_2^{Fin} = w \times \frac{\mathrm{ANOF}_1^{(N)}}{\mathrm{ANOF}_2^{(N)}}$$

이 되어 비추정 대체법과 회귀대체법 모두 같은 형태의 최종 가중치를 얻게 된다.

3. 모의실험

BLS 가중치보정법을 이용한 총계 추정량, 비추정 대체를 이용한 총계 추정량, 회귀대체를 이용한 총계 추정량의 우수성을 살펴보기 위하여 모의실험이 실시되었다. 모의실험에는 김석과 신기일 (2009)과 이상은과 신기일 (2010)에서 사용한 자료를 사용하였다. 이 자료는 2007년 3월과 5월의 매월노동통계 자료 중에서 매출액 자료이다. 이상은과 신기일 (2010)에서와 같이 3개 층의 자료가 사용되었으나, 본 논문에서는 무응답과 이상점이 있는 경우에 대체법의 효과를 살펴보는 것이 주된 목적이므로 이 목적에 맞지 않는 일부의 자료를 제거한 후, 즉 5월 자료가 "0"인 자료를 제거한 후 모의실

표 1: 층별 모집단 특징

층	표본 수	상관계수	회귀대체		비추정대체
			β_0	β_1	R
C1	1117	0.884	575.51	0.81	0.951
C2	963	0.943	4788.62	0.92	0.977
C3	397	0.945	64017	0.82	0.890

험이 실시되었다. 각 층은 사업체 규모에 따라 나누어졌으며 C1층: 5–9인, C2층: 100–299인, C3층: 500인 이상을 의미한다. 표 1은 각 층별 모집단 특징을 나타내고 있다.

이상은과 신기일 (2010) 자료와 비교하면 C1층의 경우 5월 자료에 "0"인 자료가 많아 상대적으로 많은 자료가 제거되었으며, C2층의 경우에는 20개의 자료가 제거되었다. 반면 C3층은 대규모 사업체이므로 매출액이 "0"인 자료가 없어 제거된 자료가 없다. 자료의 제거에도 불구하고 상관계수와 회귀계수의 기울기 추정값은 거의 변화가 없다. 따라서 이상은과 신기일 (2010)의 모의실험 결과와 본 모의실험 결과를 연결하여 비교하는 것도 의미가 있을 수 있다. 이상은과 신기일 (2010)의 모의실험 결과는 이상점이 없는 경우의 결과로 C1층의 경우 회귀대체가 우수한 결과를 주고 있으며 C2층은 거의 같은 결과를 C3층은 비추정 대체가 우수한 결과를 주고 있다.

본 논문에서 사용한 비교 통계량은 MSE, Bias, RMSE, ARE로 각 통계량의 정의는 다음과 같다.

$$\text{MSE} = \frac{1}{R}\sum_{k=1}^{R}(t_k - \hat{t}_k)^2,$$

$$\text{Bias} = \frac{1}{R}\sum_{k=1}^{R}(t_k - \hat{t}_k),$$

$$\text{RMSE} = \frac{1}{R}\sum_{k=1}^{R}\left(\frac{t_k - \hat{t}_k}{t_k}\right)^2,$$

$$\text{ARE} = \frac{1}{R}\sum_{k=1}^{R}\left|\frac{t_k - \hat{t}_k}{t_k}\right|,$$

여기서 t_k는 모집단 총계이고, $\hat{t}_k$는 각 방법에서 얻어진 총계 추정량이다. 반복 수 R은 1,000을 이용하였으며, 각 모집단으로부터 각각, 5%, 10% 그리고 20%의 결측비율을 사용하여 자료를 구성하였다. 또한 이상점의 수는 각각 3, 7 그리고 10개를 사용하였다. 이상점은 5월 자료값에 6을 곱하여 만들었으며 가중치 $w = 10$을 사용하였다. 결측 메카니즘과 이상점 발생 메카니즘은 MAR를 따른다고 가정하였다. 결과는 표 2~4에 나와 있다.

먼저 표 2의 MSE 결과를 살펴보면 t_{BLS}와 t_{NBLS}는 이상점의 수가 3, 7에서 큰 차이를 보이고 있지 않다. 그러나 이상점의 수가 10개인 경우에는 제안된 방법이 우수한 것을 확인할 수 있다. 이는 자료의 수가 1100여개가 되기 때문에 이상점의 수가 큰 경우에만 차이를 보이는 것으로 판단된다. t_{REG}의 결과를 살펴보면 t_{NBLS}에 비해 우수한 결과를 보이고 있다. 이는 새롭게 제안한 방법에 의한 결과라기보다는 비추정대체보다 회귀대체가 우수하기 때문인 것으로 판단된다. 결론적으로 대체법을 사용하여 만든 제안된 총계 추정량이 우수한 결과를 주고 있다. Bias를 기준으로 살펴보면 모든 결과에서 음의 값을 갖고 있다. 또한 작은 수의 이상점이 있는 경우에는 t_{BLS}의 결과가 우수한 것을 확인할 수 있으며, 이상점 수가 많아지게 되면 모든 방법에서 유사한 결과를 주고 있다. RMSE와 ARE를 기준으로 살펴보면 이상점의 수가 작은 경우에는 유사한 결과를 주고 있으며 이상점 수가 많아지면 제안된 방법이 우수한 결과를 주고 있음을 확인할 수 있다.

표 2: 제안된 방법의 우수성 비교 결과(C1층)

비교통계량	추정량	비교통계량	결측 비율		
			5%	10%	20%
MSE	3	t_{BLS}	19153598820	39144383982	111263670314
		t_{NBLS}	19160659329	39205700832	111870375721
		t_{REG}	16958343507	35643150123	102424305658
	7	t_{BLS}	24695582750	64665726318	239397535546
		t_{NBLS}	24688671535	64596094042	237761925922
		t_{REG}	22250190149	60215688745	227093567933
	10	t_{BLS}	38720799129	100565531607	398360924541
		t_{NBLS}	38192720106	98993429224	383442194966
		t_{REG}	35728159174	94016617023	373235659877
Bias	3	t_{BLS}	−8348.74	−30355.04	−110694.75
		t_{NBLS}	−9627.00	−32864.69	−115766.46
		t_{REG}	−8564.19	−35593.36	−116485.40
	7	t_{BLS}	−56723.80	−126339.79	−341817.01
		t_{NBLS}	−57961.10	−128589.57	−344641.62
		t_{REG}	−59650.44	−128953.00	−344710.86
	10	t_{BLS}	−111334.38	−222790.67	−505271.13
		t_{NBLS}	−111179.16	−222250.94	−498844.58
		t_{REG}	−109577.98	−222591.96	−502885.69
RMSE	3	t_{BLS}	0.000009912	0.000020247	0.000057484
		t_{NBLS}	0.000009915	0.000020279	0.000057795
		t_{REG}	0.000008775	0.000018428	0.000052899
	7	t_{BLS}	0.000012271	0.000032065	0.000118519
		t_{NBLS}	0.000012269	0.000032038	0.000117767
		t_{REG}	0.000011059	0.000029861	0.000112443
	10	t_{BLS}	0.000018556	0.000048571	0.000191280
		t_{NBLS}	0.000018313	0.000047843	0.000184275
		t_{REG}	0.000017116	0.000045379	0.000179259
ARE	3	t_{BLS}	0.002435498	0.003507251	0.005954216
		t_{NBLS}	0.002434256	0.003506270	0.005966319
		t_{REG}	0.002311227	0.003357674	0.005811106
	7	t_{BLS}	0.002736818	0.004406277	0.008746155
		t_{NBLS}	0.002737089	0.004408974	0.008748194
		t_{REG}	0.002612392	0.004328456	0.008701432
	10	t_{BLS}	0.003263519	0.005636026	0.011649000
		t_{NBLS}	0.003247364	0.005603722	0.011481000
		t_{REG}	0.003225631	0.005540522	0.011474000

표 3의 결과를 살펴보면 C1층에 비해 자료값이 크고 자료의 수도 상대적으로 작기 때문에 이상점의 영향이 크게 나타나고 있다. 물론 표 3에서도 이상점의 수가 3인 경우에는 유사한 결과를 주고 있다. 또한 Bias의 경우에는 t_{BLS}가 우수한 결과를 주는 경우도 있다. 그러나 이상점의 수가 7, 10개인 경우 모든 비교 통계량을 기준으로 살펴보았을 때 t_{BLS}에 비해 t_{NBLS}와 t_{REG}가 우수한 결과를 주고 있음을 확인 할 수 있다. 또한 이상점의 수가 10인 경우에는 t_{NBLS}와 t_{REG}의 결과가 유사하여 제안된 방법이 우수한 결과를 주고 있다.

표 4의 결과는 이상점의 영향이 더욱 확대되고 있으며 이상점의 수가 작은 경우에는 유사한 결과를 주는 반면 이상점 수가 많은 경우에는 제안된 방법이 매우 우수한 것을 확인 할 수 있다. 특히 비추정 대체, 회귀대체 등 자료에 맞는 타당한 대체법을 사용한다면 수정된 BLS 가중치보정법의 총계 추정량은 매우 우수한 결과를 줄 것으로 판단된다.

표 3: 제안된 방법의 우수성 비교 결과(C2층)

비교통계량	추정량	비교통계량	결측 비율		
			5%	10%	20%
MSE	3	t_{BLS}	4.2787E12	1.0591E13	3.3726E13
		t_{NBLS}	4.2679E12	1.0542E13	3.3235E13
		t_{REG}	4.1922E12	1.0406E13	3.2669E13
	7	t_{BLS}	1.0174E13	2.9789E13	1.1344E14
		t_{NBLS}	9.8661E12	2.8512E13	1.0563E14
		t_{REG}	9.6786E12	2.8347E13	1.0545E14
	10	t_{BLS}	2.4232E13	6.6738E13	2.3333E14
		t_{NBLS}	2.3005E13	6.1795E13	2.0909E14
		t_{REG}	2.2802E13	6.1453E13	2.0969E14
Bias	3	t_{BLS}	−642730.57	−1455687.35	−3094232.80
		t_{NBLS}	−644751.80	−1457655.15	−3085476.68
		t_{REG}	−621760.37	−1461371.94	−3080009.67
	7	t_{BLS}	−2011324.33	−4120180.23	−8835588.02
		t_{NBLS}	−1972489.14	−4025105.82	−8546378.72
		t_{REG}	−1967373.21	−4043153.50	−8556862.30
	10	t_{BLS}	−3948552.73	−6844880.71	−13539915.24
		t_{NBLS}	−3844800.65	−6588985.31	−12851090.36
		t_{REG}	−3847316.57	−6572454.69	−12881830.42
RMSE	3	t_{BLS}	0.000006164	0.000015253	0.000048247
		t_{NBLS}	0.000006149	0.000015184	0.000047560
		t_{REG}	0.000006041	0.000014985	0.000046735
	7	t_{BLS}	0.000013837	0.000040604	0.000154416
		t_{NBLS}	0.000013426	0.000038891	0.000143944
		t_{REG}	0.000013163	0.000038662	0.000143646
	10	t_{BLS}	0.000031766	0.000087783	0.000308254
		t_{NBLS}	0.000030176	0.000081355	0.000276547
		t_{REG}	0.000029897	0.000080890	0.000277229
ARE	3	t_{BLS}	0.001887692	0.003020018	0.005458025
		t_{NBLS}	0.001885148	0.003013756	0.005426674
		t_{REG}	0.001887606	0.003006936	0.005377157
	7	t_{BLS}	0.002878796	0.005208988	0.010620000
		t_{NBLS}	0.002837904	0.005103806	0.010283000
		t_{REG}	0.002827965	0.005131060	0.010283000
	10	t_{BLS}	0.004667953	0.007963047	0.015687000
		t_{NBLS}	0.004552411	0.007673474	0.014899000
		t_{REG}	0.004544770	0.007672740	0.014932000

4. 결론

본 논문에서는 BLS 가중치보정법을 수정하여 총계를 추정하는 새로운 방법을 제안하였다. BLS 가중치보정법은 결측치와 이상점이 있는 경우에 가중치를 보정하여 총계를 추정한다. 본 논문에서는 이상점과 무응답 결측이 동시에 있을 경우, 무응답 결측의 영향력을 최소로 하기 위해 설계가중치를 보정하는 방법 대신 결측을 대체하는 대체법의 사용을 제안하였다. 이 경우 이상점이 있기 때문에 제안된 방법을 사용하기 위해 이상점 가중치 보정공식이 수정되었다. 달라진 보정방법이 총계 추정에 얼마나 영향을 주는지 살펴보았으며 모의실험 결과 BLS 무응답 가중치보정법을 사용하는 것보다 대체법을 사용하는 제안된 새로운 방법이 우수한 총계 추정값을 주는 것을 확인하였다.

표 4: 제안된 방법의 우수성 비교 결과(C3층)

비교통계량	추정량	비교통계량	결측 비율		
			5%	10%	20%
MSE	3	t_{BLS}	1.0920E+15	2.1876E+15	1.0251E+16
		t_{NBLS}	1.0477E+15	2.0739E+15	8.4186E+15
		t_{REG}	1.0777E+15	2.2395E+15	9.2273E+15
	7	t_{BLS}	6.5812E+15	9.7367E+15	3.4461E+16
		t_{NBLS}	5.5745E+15	8.1564E+15	2.5407E+16
		t_{REG}	6.3360E+15	8.5264E+15	2.6302E+16
	10	t_{BLS}	2.9775E+16	2.6558E+16	9.0338E+16
		t_{NBLS}	2.3920E+16	2.0975E+16	5.7952E+16
		t_{REG}	2.5773E+16	2.1571E+16	5.9001E+16
Bias	3	t_{BLS}	−2732196.35	−7765394.61	−32838267.37
		t_{NBLS}	−2981884.61	−8283071.83	−32329441.18
		t_{REG}	−3100155.10	−8498029.31	−33180147.22
	7	t_{BLS}	−27479354.31	−38863287.90	−98327879.92
		t_{NBLS}	−26482940.99	−37307343.97	−91102015.52
		t_{REG}	−26954832.21	−38453737.81	−91692440.02
	10	t_{BLS}	−60168670.43	−85609234.53	−173614399.80
		t_{NBLS}	−56525980.53	−79355923.96	−152271903.00
		t_{REG}	−57181442.44	−79335782.28	−151912191.90
RMSE	3	t_{BLS}	0.000082665	0.000169286	0.000703262
		t_{NBLS}	0.000079927	0.000162051	0.000594574
		t_{REG}	0.000081458	0.000171015	0.000643981
	7	t_{BLS}	0.00037871	0.00059521	0.002115348
		t_{NBLS}	0.000327097	0.000506836	0.001609831
		t_{REG}	0.000363544	0.000529748	0.001667253
	10	t_{BLS}	0.001423719	0.001497786	0.005032896
		t_{NBLS}	0.001161897	0.001198174	0.003347427
		t_{REG}	0.001240221	0.001227779	0.003392787
ARE	3	t_{BLS}	0.006433023	0.009481349	0.017064000
		t_{NBLS}	0.006370682	0.009367678	0.016517000
		t_{REG}	0.006037353	0.009121514	0.016730000
	7	t_{BLS}	0.011006000	0.014848000	0.029896000
		t_{NBLS}	0.010629000	0.014221000	0.027780000
		t_{REG}	0.010371000	0.014272000	0.027707000
	10	t_{BLS}	0.017478000	0.023853000	0.045525000
		t_{NBLS}	0.016552000	0.022207000	0.040269000
		t_{REG}	0.016476000	0.022375000	0.040094000

비추정 대체와 회귀대체 중 하나를 선택하는 것은 자료의 현상을 파악하여 결정하게 된다. 어떤 방법이 선택되는지에 상관없이 본 논문에서 제안한 방법을 사용하게 되면 기존의 BLS 가중치보정법의 총계 추정량에 비해 우수한 결과를 줄 것으로 판단되며 자료에 맞는 적절한 대체법이 사용된다면 그 효과는 더욱 클 것으로 판단된다.

참고 문헌

김석, 신기일 (2009). 상관관계와 표본 크기에 따른 BLS 무응답 보정의 효율성 비교, <응용통계연구>, **22**, 1301–1313.

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

이상은 (2008). 표본조사에 따른 추정방법비교: 가중치조정기법을 중심으로, <응용통계연구>, **21**, 413–427.

이상은, 신기일 (2010). BLS 무응답 보정법을 이용한 대체법과 이월대체법에 관한 연구, <응용통계연구>, **23**, 909–921.

이석진, 신기일 (2008). BLS 보정 방법의 민감도에 관한 연구, <통계학회논문집>, **15**, 843–858.

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

Little, R. J. A. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

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

2011년 1월 접수; 2011년 3월 채택

Modified BLS Weight Adjustment

Jung-Joon Park[a], Kijong Cho[a], Sang Eun Lee[b], Key-Il Shin[1,c]

[a]Department of Environmental Science and Ecological Engineering, Korea University
[b]Department of Applied Statistics, Kyonggi University
[c]Department of Statistics, Hankuk University of Foreign Studies

Abstract

BLS weight adjustment is a widely used method for business surveys with non-responses and outliers. Recent surveys show that the non-response weight adjustment of the BLS method is the same as the ratio imputation method. In this paper, we suggested a modified BLS weight adjustment method by imputing missing values instead of using weight adjustment for non-response. Monthly labor survey data is used for a small Monte-Carlo simulation and we conclude that the suggested method is superior to the original BLS weight adjustment method.

Keywords: Regression imputation, ratio imputation, non-response adjustment, outlier adjustment.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2008-313-C00141).

[1] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yonginsi, Kyunggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr