추정

(1) 가중값 부여 방법

○ 통계학 이론에 의하면, 표본조사로부터 모수를 추정할 때 추정단계에서 가중값을 이용하면 모집단에 대한 특성치인 모수에 대한 불편추정량(unbiased estimator)을 얻는다. 만약 통계분석 과정에서 가중값을 무시하고 분석한 추정치는 심각한 편향(bias)이 발생하게 된다. 표본 크기가 큰 대규모 조사에서 문제가 되는 것은 추정량의 편향이기 때문에 추정과정에서 반드시 가중값을 이용해야 한다.

○ 대부분의 기업체 대상 표본조사에서 k 산업중분류, h 규모의 i번째 기업체에 대한 가중값 w_{khi}를 산출하는 방법으로는 크게 두 가지를 생각할 수 있다. 기업체 대상 조사의 기본 가중값은 첫 번째, 층 내의 모집단 전체 기업수를 그 층의 표본 기업수로 나누어 주는 방법이고, 두 번째, 층 내의 모집단 전체 기업 상용근로자 수를 그 층의 표본기업체의 상용근로자 수로 나누어 주는 방법이다.

- 첫 번째 방법은 일종의 호르비츠-톰슨(Horvitz-Thompson) 추정의 형태이고 두 번째 방법은 일종의 일반화 회귀추정량(Generalized regression estimator; GREG 추정량)의 특수한 경우인 분리 비 추정량(separate ratio estimator)의 형태이다.

○ 첫 번째 방법은 일반적인 표본조사에서 부여하는 방법으로 층별 모집단 특성이 동질적이지만 층별 추출방법이 다르거나 표본단위별로 추출률이 상이한 복합표본조사(complex sample survey)에 널리 이용되는 방법이다. 이 방법의 가중값은 ① 설계가중값, ② 무응답에 대한 조정, ③ 벤치마킹 보정에 대한 조정 등의 세 가지 요인을 결합하여 산정된다. 기업체노동비용조사를 위한 호르비츠-톰슨 추정의 가중값은 다음의 과정으로 부여한다. 가중값은 층화변수인 산업중분류 및 상용근로자수 규모별 기업체 수를 기준으로 부여한다. 이 방법은 기업체 단위를 기준으로 노동비용 통계를 작성할 때 적절한 방안이다.

- 1단계 : 설계 가중값 부여

$$w_{kh1} = \frac{N_{kh}}{n_{kh}}$$

여기서 k : 산업중분류, h : 기업체의 상용근로자수 규모

- 2단계 : 무응답 보정 계수

$$w_{kh2.1} = \frac{n_{kh1}}{r_{kh1}}$$

여기서 r : (k, h) 층의 응답 기업체 수

- 3단계 : 최신 모집단 기업체 수를 벤치마크하기 위한 사후층화 조정 계수

$$w_{p3.12} = \frac{N_{kh,\text{최신기준}}}{\hat{N}_{kh,\text{설계시점}}} = \frac{N_{kh,\text{최신기준}}}{\sum_{(k,h)} (w_{kh1} \times w_{kh2.1})}$$

- 최종 가중값

$$w_{khi,f} = w_{kh1} \times w_{kh2.1} \times w_{p3.12}$$

○ 두 번째 가중값 부여 방법은 기업체 수 기준이 아니라 알려져 있는 모집단 상용근로자 수를 이용하는 방법으로 표본기업체의 상용근로자 수와의 비(ratio)를 이용하여 부여하는 방안이다. 두 번째 방법의 가중값은 추정량의 변동이 작기 때문에 보다 안정적인 통계를 생산할 수가 있다. 두 번째 방법의 가중값은 다음과 같이 부여한다.

$$w_{khi} = \frac{\sum_{i}^{N_{kh}} x_{kh}}{\sum_{i}^{n_{kh}} x_{khi}}$$

여기서 k : 산업중분류, h : 기업체의 상용근로자수 규모, x : 기업체의 상용근로자 수

- 두 번째 가중값을 사용하면 실제로 한 기업체에서 지불한 총 노동 비용은 그 기업체의 상용근로자 수에 비례할 것이므로 상용근로자 수의 비를 사용한 가중값이 더 효율적인 추정을 만들어 낼 것이다. 또한 이러한 가중값을 사용하면 층 내 가중값의 합계가 해당 층의 모집단 총 근로자 수가 나온다는 장점이 있다. 새로운 조사에서 벤치마크 가중값으로 두 번째 가중값을 사용하는 방안의 검토가 필요하다.

○ 기업체노동비용조사의 가중값은 전수층을 포함하여 산업중분류 내의 6개 기업체 규모별로 상용근로자 수와 응답 기업의 조사 근로자 수의 복원배율로 계산한다. 이는 설계가중값에 무응답 조정과 벤치마킹 보정을 하는 가중값을 반영한 층 내의 근로자수 기준으로 비조정(ratio adjustment)의 형태를 사용하는 분리비 추정(separate ratio estimation)의 전형적인 추정 방법이다. 이러한 분리비 추정방법은 층 내에서 각 기업체의 총 노동 비용의 기대값과 분산이 상용근로자 수에 비례하는 초모집단 모형에서 기대분산을 최소화하는 추정법으로 많이 사용된다.

○ 또한, 가중값은 모수 추정의 편향을 제거하지만 추정치의 분산을 크게 하므로 가중값의 변동을 계산하여 작성된 통계에 대한 상대표준오차에 미치는 영향을 감소하기 위해 극단 가중값을 검토하여 조정이 필요하다. 실제 추정 과정에 극단 가중값을 탐색하여 조정하기 위해서는 표본기업체의 가중값을 부여한 후 산업중분류별 가중값의 분포를 분석하여 극단 가중값 경계를 설정하는 것이 바람직하다.

- 산업중분류별 극단 가중값의 경계를 설정하는 방안은 다음과 같다.
  - 통계학의 이론에 근거하여 표본기업체의 개별 가중값이 중앙값 기준 1.5(혹은 3)사분위범위를 초과하거나 혹은 평균 가중값의 2(혹은 3)표준편차를 경계로 설정하는 방안
  - 일반적인 현장조사에서 적용하는 산업중분류별 최소 가중값과 최대 가중값의 비(ratio)를 기준으로 설정하는 방안으로 기준 설정의 검토가 필요하지만 5배(혹은 10배)를 경계로 설정하는 방안
  - 표본기업체의 가중값을 부여한 후 가중값 특성을 검토하여 절대 가중값을 경계로 설정하는 방안 : 현행 조사는 상용근로자 수의 비를 가중값으로 부여한 극단 가중값 기준은 600으로 설정하고 있으며, 2018년 조사 표본기업체의 최대 가중값이 400미만이므로 극단 가중값은 없는 상황임.
- 현행 기업체노동비용조사는 표본층에 대해 산업중분류 및 기업체 규모별로 극단 가중값을 탐색하고 있으며, 표본기업체의 가중값이 조사기획 과정에서 설정한 극단가중값의 경계 내에 모두 존재하므로 극단 가중값이 존재하지 않는다. 하지만 향후에라도 극단 가중값이 존재한다면 기본적으로 설계 시점의 모집단 특성과 벤치마크로 사용하는 최신 모집단 특성의 변화를 고려해 가중값을 조정하는 방안이 필요하다. 극단 가중값을 조정해야 한다면, 층별로 설정한 경계값을 부여하거나 윈저화 등의 방법으로 극단 가중값을 조정하는 방법을 제안한다.

○ 가중값은 표본이 모집단을 얼마나 대표하는 가를 나타내는 확정계수이므로 층별 표본크기가 작으면 모집단을 대표하기 위한 가중값이 지나치게 크게 되어 추정 결과에 심각하게 영향을 주게 된다. 일반적으로 가중값의 변동이 크면 추정치의 분산이 증가해 추정 오차를 크게 하여 조사 결과의 안정성에 영향을 미치게 된다. 이러한 점을 고려하여 새로운 표본설계에서는 산업중분류 및 규모별 최소 표본을 3개로 설정하여 층별 가중값의 영향을 감소하는 방안을 고려해 표본크기를 결정하였다. 실제 조사과정에서 층별 모집단 기업체 수 및 표본기업체 수가 큼에도 조사가 완료된 층별 최종 완료 표본기업체수가 2개 이하라면 인접한 다른 층과 결합하여 가중값을 안정시켜 추정하는 것이 바람직하다. 년도별로 층별 가중값이 안정되면 층별 조사 완료된 표본기업체 수가 추정 결과에 미치는 영향이 감소하여 추정 결과의 안정성이 확보되기 때문이다.

(2) 추정

○ 기업체노동비용조사에서 중요하게 추정하는 모수는 근로자 1인당 월평균 노동비용이다. 근로자 1인당 노동비용의 모수는 다음과 같다.

• 근로자 1인당 월평균 노동비용 : $\theta = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$

- 여기서 x_i는 i번째 기업체 내의 임금 지급 연인원(12개월 합계)이고 y_i는 i번째 기업체 내의 각종 노동비용의 12개월 합계이다.

○ 새로운 표본설계는 전수층과 표본층으로 구분하여 표본설계를 하였으므로 추정은 전수층과 표본층으로 각각 추정하여 결합해 추정하면 된다. 기본적으로 전수층은 전수추출하므로 무응답 및 사후층화 혹은 캘리브레이션의 조정을 통해 집계하고, 표본층은 설계가중값, 무응답 및 사후층화 혹은 캘리브레이션 가중값을 부여하여 추정한다. 하지만 새로운 표본설계에서 제안한 두 번째 가중값은 설계가중값에 무응답에 대한 조정과 벤치마킹 보정 과정이 반영된 분리비 추정을 위한 전형적인 추정 방법의 가중값이므로 전수층도 동일하게 적용되기 때문에 전수층과 표본층을 구분해 추정하지 않고 한 번의 과정으로 추정이 가능하다.

○ 산업중분류별 근로자 1인당 월평균 노동비용의 모수는 비추정량을 이용하여 다음과 같이 추정한다.

• 근로자 1인당 월평균 노동비용의 추정량(전체 근로자 기준) :

$$\hat{\theta} = Y_c + \hat{Y}_s = \frac{\sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}} w_{khi} y_{khi}}{\sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{kh}} w_{khi} x_{khi}}$$

– 여기서 Y_c는 전수층, $\hat{Y}_s$는 표본층의 추정 값, n_{kh}는 k산업중분류, h번째 기업체규모의 표본기업체 수, w_{khi}는 표본기업체에 부여된 가중값, y_{khi}는 표본기업체에서 얻은 관측치(직접노동비용, 간접노동비용, 노동비용 총액 등), x_{khi}는 1년간 임금 지급 연인원임.

• 현행 조사에서 층별 모집단 상용근로자 수와 응답 기업체의 상용근로자 수의 비를 가중값으로 사용하기 때문에 무응답이 있는 경우에도 기업체의 상용근로자 수로 구한 가중값을 그대로 적용할 수 있다. 예를 들어 각 층에서 n_{kh}개의 표본기업체 중에 R_{kh}개의 기업체만이 응답했다고 하더라도 근로자 1인당 월평균 노동비용의 추정량은 다음과 같이 계산이 된다.

– 근로자 1인당 월평균 노동비용의 추정량(무응답 보정) : $\hat{\theta}^{*} = \dfrac{\sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{R_{kh}} w_{khi}^{*} y_{khi}}{\sum_{k=1}^{K}\sum_{h=1}^{H_k}\sum_{i=1}^{R_{kh}} w_{khi}^{*} x_{khi}}$

여기서 w_{khi}^{*}는 응답 기업체에게 최종적으로 부여되는 가중치로 층 내의 모집단 전체 기업 상용근로자 수를 그 층의 응답 기업체의 상용근로자 수로 나누어 준 것으로 사용한다.

○ 산업중분류별, 기업체 규모별의 통계 작성 단위별 노동비용에 대한 각종 평균 통계치는 기본 통계치를 결합해 다음과 같이 추정한다.

• k 산업중분류의 근로자 1인당 월평균 노동비용 추정량 :

$$\hat{\theta}_k = \frac{\sum_{h=1}^{H_k}\sum_{i=1}^{n_{hk}} w_{khi} y_{khi}}{\sum_{h=1}^{H_k}\sum_{i=1}^{n_{hk}} w_{khi} x_{khi}}$$

• h 기업체 규모의 근로자 1인당 월평균 노동비용 추정량 :

$$\hat{\theta}_h = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_{khi}} w_{khi} y_{khi}}{\sum_{k=1}^{K}\sum_{i=1}^{n_{khi}} w_{khi} x_{khi}}$$

○ 산업대분류별 통계 작성 단위별 노동비용에 대한 각종 평균 통계치는 기본 통계치를 결합해 다음과 같이 추정한다.

• G 산업대분류의 근로자 1인당 월평균 노동비용의 추정량 :

$$\hat{\theta}_G = \frac{\sum_{k\in G}^{G_k}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{hk}} w_{khi} y_{khi}}{\sum_{k\in G}^{G_k}\sum_{h=1}^{H_k}\sum_{i=1}^{n_{hk}} w_{khi} x_{khi}}$$

○ 추정량의 분산 추정량과 상대표준오차 : 월 평균 노동비용 총액 등의 비추정량에 대한 분산추정량과 상대표준오차 추정량 및 무응답을 조정한 추정량의 분산추정량은 다음과 같다.

• 분산추정량 $\widehat{var}(\hat{\theta}) = \sum_{k=1}^{K}\sum_{h=1}^{H_k} \dfrac{n_{kh}(1-f_{kh})}{n_h - 1} \sum_{i=1}^{n_{kh}} e_{khi}^{2}$

여기서 $e_{khi} = \dfrac{w_{khi}}{w_{....}} (y_{khi} - \hat{\theta}_{kh}\, x_{khi})$, $\hat{\theta}_{kh} = \dfrac{\sum_{i=1}^{n_{kh}} y_{khi}}{\sum_{i=1}^{n_{kh}} x_{khi}}$,

$$w_{....} = \sum_{k=1}^{K} \sum_{h=1}^{H_k} \sum_{i=1}^{n_{kh}} w_{khi} \ , \quad f_{kh} = \frac{n_{kh}}{N_{kh}}$$

- 표준오차 추정량 : $\widehat{se}\,(\hat{\theta}\,) = \sqrt{\widehat{var}\,(\hat{\theta}\,)}$
- 상대표준오차 추정량 : $\widehat{CV}(\hat{\theta}\,) = \dfrac{\widehat{se}(\hat{\theta}\,)}{\hat{\theta}} \times 100$

- 무응답을 조정한 추정량을 사용한 경우의 분산추정량

$$\widehat{var}\,(\hat{\theta}^{\,*}) = \sum_{k=1}^{K} \sum_{h=1}^{H_k} \frac{R_{kh}\,(1 - f_{kh}\,)}{R_h - 1} \sum_{i=1}^{R_{kh}} e_{khi}^{*2}$$

여기서 $\quad e_{khi}^{*} = \dfrac{w_{khi}^{*}}{w_{....}}\,(y_{khi} - \hat{\theta}_{kh}^{\,*}\, x_{khi}) \quad , \ f_{kh} = \dfrac{R_{kh}}{N_{kh}}, \quad \hat{\theta}^{*}{}_{kh} = \dfrac{\sum_{i=1}^{R_{kh}} y_{khi}}{\sum_{i=1}^{R_{kh}} x_{khi}}$