ISSN 2508-7118

SIES Working Paper Series
No. 368 2017. 6

No. 368

지니계수 보정

홍 민 기

社團法人 서울社會經濟硏究所

지니계수 보정

홍 민 기(한국노동연구원)

표본 조사는 고소득을 잘 포착하지 못하는 한계가 있어서 표본 가구 조사로 계산한 지니계수가 실제의 불평등을 잘 반영하지 못한다는 문제 제기가 있었다. 이러한 문제를 해결하고자 본 연구에서는 가계동향조사의 가중치를 조정하고 고소득 가구 표본을 추가하여 1998년부터 2015년까지 지니계수를 보정하였다. 가중치 보정의 기준이 되는 개인 소득 분포는 국세통계와 가계 조사를 결합하여 구성하였다. 그리고 가구원수와 기타 가구원 소득과 같은 결측 자료(missing data)를 대체(imputation)하여 완전한 자료를 만들고, 지니계수를 반복 추정하는 베이지안 자료 확대 (data augmentation) 방법을 사용하였다.

보정 지니계수는 원래의 지니계수에 비해 10~18% 정도 크다. 표본 가구조사는 고소득을 과소포착하고 있기 때문에 최근 들어 최상위 소득 비중이 커지면서 보정의 효과도 크게 나타난다. 2010년 이후 원래의 지니계수는 약간 하락하는 추세이지만, 보정 지니계수는 증가하는 추세이다.

Ⅰ. 머리말

대표적인 불평등 지표인 지니계수는 표본 가구 조사로 계산하고 있다. 불평등 척도는 고소득의 영향을 많이 받는다 (Cowell and Flachaire, 2007). 표본 가구 조사는 일반적으로 고소득을 잘 포착하지 못하는 한계가 있어서 가구 조사로 계산한 불평등 지표가 실제의 불평등을 잘 반영하지 못할 수 있다. 1980년대 이후 영미권을 중심으로 최상위 소득 비중이 늘어나면서 전체 소득 불평등 추세에 큰 영향을 주고 있다 (Burkhauser, et. al, 2012). 최상위 소득 비중이 늘어나면, 조사 자료로 측정한 소득 불평등 지표가 실제의 불평등도와 차이가 날 가능성이 더 높아진다. 본 연구에서는 고소득을 잘 포착하는 국세통계와 가구 조사 자료를 결합하여 지니계수를 보정하고자 한다.

지니계수를 보정할 때 해결해야 할 문제는 두 가지이다. 첫째, 표본조사가 얼마나 고소득을 과소포착하고 있는지 비교할 수 있는 준거가 있어야 한다. 본 연구에서는 국세통계와 가계동향조사를 결합하여 개인 소득 분포를 만들고 이를 준거로 삼는다. 둘째, 국세통계로 개인 소득을 알 수 있지만, 지니계수를 계산하기 위해서는 가구원수와 가구 소득을 알아야 한다. 개인 소득은 관측 자료이고, 가구원수, 기타가구원 소득은 관측되지 않는 결측 자료(missing data)이다. 본 연구에서는 관측된 자료에 기반하여 결측 자료를 대체(imputation)하여 완전한 자료를 만들어 추정을 하는 베이지안 자료 확대 (Bayesian data augmentation) 방법으로 지니계수를 보정한다.

본 논문의 구성은 다음과 같다. 다음 2장에서는 표본 조사의 소득 과소포착 문제를 해결하는 기존 연구를 개괄한다. 3장에서는 조사 자료와 국세통계의 개인소득 분포를 비교하여 조사 자료의 소득 포착률은 계산한다. 4장에서는 지니계수 보정 방법에 대해 설명하고 5장에서 그 결과를 보여준다. 마지막 6장은 결론이다.

Ⅱ. 기존 연구

표본 조사자료에서의 소득 과소포착(undercoverage) 문제를 해결하는 방식에는 가중치 조정과 대체 방법이 있다 (Hlasny and Verme, 2016)

가중치 조정(reweighting)은 조사 자료의 가중치를 다른 자료와 비교하거나 응답 확률을 고려하여 바꾸는 방법이다. 고소득자들은 조사에 잘 응하지 않는 경향이 있는데 이를 고려하지 않으면 전체 소득 분포가 정확히 조사되지 않을 수 있다. 미국 표본 조사에서는 지역별 비응답 확률을 이용하여 가중치를 바꾸는 방법을 사용한다. 다시 말해, 단위 무응답(unit non-responses) 때문에 발생하는 자료수집 문제를 가중치를 조정하여 해결한다.

대체(replacing)는 주어진 조사 자료로는 원하는 정보를 얻을 수 없을 때 사용된다. 미국 CPS (Current Population Survey) 에서는 일정 금액을 초과하는 소득을 코딩(coding)하여 공개하고 있다. (이른바, top coding). 공개된 CPS 자료로 불평등 지표를 계산하기 위해서는 최상위 소득분포를 복원할 필요가 있는데, Cowell and Flachaire (2007), Jenkins et al. (2011), Lakner and Milanovic(2013)의 연구에서 표본 대체의 방법을 사용하였다. 예를 들어 1억원 이상 소득이 전부 '1억'으로만 표시되어 있을 경우, 최상위 소득에 대한 분포를 가정하여 '1억'으로 표시된 관측치를 다른 값으로 대체한다.

Burkhauser et al. (2012)은 일반화된 베타 2 분포(generalized beta of second kind)를 가정하여 최상위 소득을 대체한 후 이를 비공개 CPS 자료와 비교하였는데, 두 자료로 측정한 불평등 지표의 차이가 매우 적었고, 조세자료로 측정한 불평등 지표와도 차이가 매우 적다고 하였다. 코딩된 소득 자료를 대체하는 방법이 타당하다는 것으로 보여준 것이기도 하고, CPS에서 고소득 과소포착의 문제가 없다는 것을 보여준 것이기도 하다.

미국에서는 조사 자료로 고소득을 파악할 때의 문제가 고소득 코딩이라는 자료 발표과정에서 발생한 것인 반면, 한국에서는 고소득이 표본 조사에 포착되지 않는 문제로부터 발생한다. 예를 들어 2014년 가계동향조사에서는 3억원 이상 소득자가 조사되지 않는다. 가계동향조사로 소득 분포를 파악하려면 표본에 포함되지 않은 고소득자를 표본에 추가하는 방법을 사용할 필요가 있다.

고소득 과소 포착이외에 또 한가지 어려움은 지니계수의 측정 단위와 조세자료에서의 소득 단위가 다르다는 것이다. 지니계수는 가구 단위로 측정되는 반면, 한국에서 조세 납부의 단위는 개인이다. 따라서 국세통계를 기준으로 측정된 개인 소득 분포를 가구 단위의 소득 분포로 전환하여야 한다.

Alvaredo (2011)는 최상위 소득 소득 비중에 대한 정보를 이용하여 지니계수를 보정하는 방법을 제시하였는다. 최상위 집단의 인구비중을 , 소득비중을 S, 최상위 집단의 지니계수 G^*, 최상위를 제외한 하위집단의 지니계수 G^*라고 하자. 최상위 소득이 파레토 분포를 따른다고 가정하면, 지니계수는 다음과 같다.

$$G \quad \frac{\beta-1}{\beta+1} PS + G^*(1-P)(1-S) + S - P$$

여기서 $=\alpha/(\alpha-1)$이고, α는 파레토 계수이다. 최상위 소득 비중, 예를 들어 최상위 소득 1% 집단의 소득비중을 알고 있으면 위 식을 이용하여 조사자료의 지니계수를 보정할 수 있다. 그런데 이 방식은 지니계수를 측정하는 단위와 최상위 소득을 측정하는 단위가 같아야 적용할 수 있다. 최상위 소득 측정의 단위가 개인인 나라에서는 위 방법을 사용할 수 없다.

한국에서는 김낙년·김종일(2013)의 연구에서 통계청의 가구 조사를 국세통계와 비교하여 3개년(1996, 2000, 2010년)의 지니계수를 보정한 바 있다. 이들은 가구조사의 포착률이 1보다 적은 중간 소득 구간에서는 가중치를 조정하고, 가구 조사의 포착률이 0인 고소득 구간에서는 표본 추가하는 방법을 사용하였다. 이 연구의 문제점은 다음과 같다. 첫째, 이들 연구에서는 중간 소득 이하에서는 가중치 조정을 하지 않고 중간 소득 구간에서만 가중치를 조정하는데, 이는 조사자료가 저소득과 고소득을 과대 포착하고, 중간소득을 과소 포착하는 문제는 고려하지 않은 것이다. 둘째, 고소득에서 샘플을 추가할 때 고소득 아래 구간의 가구원 정보의 평균값을 사용하였는데, 소득과 가구원수가 양의 상관관계에 있다는 점을 고려하면 임의적인 방법이다. 셋째, 금융소득이 저축에 비례한다는 가정하에 금융소득을 보정하였는데, 금융소득은 소득세자료의 종합소득에 포함되어 있으므로 불필요한 작업을 한 것이다.

한편, 홍민기(2016)의 연구에서는 국세통계를 기준으로 가계동향조사를 비교하여 가중치를 조정하고, 비반복적인 자료 확대 방법으로 2012년 지니계수를 보정한 바 있다.

본 연구가 기존 연구와 다른 점은 다음과 같다. 첫째, 기존 연구에서는 국세통계를 개인 소득 분포의 기준으로 삼았는데 본 연구에서는 가계동향조사와 국세통계의 개인 소득 분포를 비교하되 표본 조사의 성향도 고려하여 가중치를 조정하는 방식을 택하였다. 둘째, 본 연구에서는 가구원수, 기타 가구원 소득과 같은 결측 자료를 생성하고 지니계수를 보정하는 통계적 방법을 개선하였다. 기존의 연구에서는 결측 자료에 대해 평균값으로 대체하기도 하고, 비반복적인 방법을 사용하였는데, 본 연구에서는 반복적인 베이지안 자료 확대(iterative Bayesian data augmentation)의 방법을 사용하여 기존의 방법을 개선하고자 하였다. 셋째, 기존연구에서는 불연속적으로 한 개 혹은 몇 개 년도에 대해서만 보정값을 계산한 반면, 본 연구에서는 1998년부터 2015년까지 지니계수 보정값을 계산하여 가구 소득 불평등의 추이를 판단할 수 있도록 하였다.

Ⅲ. 가계동향조사와 국세통계의 개인 소득 비교

이 장에서는 '가계동향조사'의 소득과소 포착정도를 파악하기 위해 '가계동향조사 분배지표' 자료와 국세통계의 개인소득 분포를 비교한다. 소득분포의 비교는 근로소득과 종합소득으로 나누어 한다. 국세통계에서 근로소득은 근로소득 연말정산자와 일용근로자의 소득을 합한 것이다. 국세통계에서 종합소득은 이중신고된 근로소득을 제외한 종합소득과 보험방문판매 소득을 합한 것이다. 종합소득에는 배당, 이자, 사업소득, 부동산임대소득, 연금소득, 기타소득이 들어 있다.

'가계동향조사 분배지표' 자료는 '농가경제조사'와 '가계동향조사'를 합한 것이다. 개인소득 분포는 '농가경제조사'의 가구주와 '가계동향조사'의 가구원 소득 정보를 이용하여 계산한다. '가계동향

조사 분배지표'를 이하에서는 간단히 '가계동향조사'라 부른다.

<표 1>에는 2014년 국세통계와 가계동향조사의 소득 분포가 나와 있다. 먼저 근로소득의 분포를 비교하여 보면, 국세통계에 대비하여 가계동향조사에서 파악된 인원의 비율, 즉 포착률은 1천만원이하에서 75.7%, 1-2천만원 구간에서는 91.0%로, 가계동향조사가 저소득 구간에서 과소 포착하고 있다. 반면 포착률은 2-4천만원 구간에서 106.2%, 105.2%로서, 중간 소득 구간에서는 가계동향조사가 과대포착하고 있다. 가계동향조사의 포착률은 6-8천만원 구간에서는 90.4%이고 소득이 증가할수록 포착률이 감소하여 2-3억 구간에서는 13.7%까지 감소한다. 근로소득 3억원 이상 구간에서는 가계동향조사의 포착률이 0이다. 즉, 가계동향조사는 저소득과 고소득 구간을 과소포착하고, 중간소득을 과대포착하며, 초고소득 구간은 전혀 포착하지 못한다. 이는 표본조사에서 전형적으로 나타나는 현상이다.

국세통계와 가계동향조사의 종합소득 분포를 비교한 것이 표의 오른쪽 열에 나와 있다. 종합소득에 대해서는 가계동향조사가 1억원 이하 모든 구간에서 국세통계보다 포착 인원이 많다. 그리고 1억원 이상 구간에서는 가계동향조사의 포착인원이 적다. 고소득 구간에서 가계동향조사가 과소포착하는 것은 표본조사의 특성으로 이해되지만, 저소득 구간에서도 표본조사가 과소포착하는 것으로 보아, 종합소득에 대한 조사는 근로소득에 대한 조사와는 다른 관점에서 판단할 필요가 있다.

<표 1> 국세통계와 가계동향조사의 소득 분포 비교 (2014년)

소득구간	근로소득			종합소득		
	국세통계 (1)	가계 동향조사 (2)	비율 (3)=(2)/(1)	국세통계	가계 동향조사 (4)	조정인원 (4)/(3)
1천만이하	7,679,394	5,812,088	0.757	2,447,667	2,863,177	3,783,058
2천만이하	4,629,046	4,210,934	0.910	901,594	1,221,575	1,342,868
4천만이하	5,229,434	5,553,483	1.062	615,979	1,651,443	1,555,080
6천만이하	2,432,600	2,558,360	1.052	224,528	732,154	696,164
8천만이하	1,334,525	1,206,560	0.904	109,134	196,432	217,265
1억 이하	585,517	363,851	0.621	58,510	69,484	111,816
2억 이하	554,584	124,936	0.225	95,437	39,733	176,374
3억 이하	42,202	5,794	0.137	23,808	3,710	27,026
5억 이하	19,595	0	0	14,707	81	14,707
5억 초과	9,595	0	0	8,639		8,639
합계	22,516,493	19,836,005	0.881	4,500,003	6,777,789	7,932,996

(주) 국세통계에서 근로소득 =근로소득연말정산+일용소득, 종합소득=금융소득(배당, 이자)+사업소득+보험방문판매소득임.

근로소득은 원천 징수되기 때문에 누락되거나 과소 보고될 가능성이 적다. 따라서 국세통계로 구성한 근로소득 분포가 참값이라고 할 수 있다. 참고로 2008년까지는 과세자만 국세통계에 포함되어 있었는데, 2009년부터는 비과세자도 통계에 포함되어 있어서 2009년 이후 국세통계의 정보가 전체

근로소득의 분포를 대표한다고 할 수 있다. 이에 따라 본 연구에서는 근로소득에 대해서는 국세통계와 가계동향조사에서의 분포 차이가 참값과 표본 조사의 차이, 즉 조사 오차(sampling error)라고 간주한다.

국세통계에서 종합소득자는 과세대상자만이 포함된 것이다. 특히 사업소득은 소득을 과소보고할 가능성이 높다. 따라서 국세통계가 실제의 사업소득 분포를 왜곡할 가능성이 있고, 오히려 조사자료에서 사업소득을 더 정확히 파악할 가능성이 있다.

근로소득에 대해서는 국세통계가 정확하고, 사업소득에 대해서는 국세통계가 부정확할 수 있다는 점을 고려하여 본 연구에서는 국세통계와 가계동향조사를 결합하여 사업소득 분포를 계산한다. 종합소득 1억 이상 구간에서는 국세통계에서 파악된 분포를 취한다. 1억 이하 소득구간에 대해서는 가계동향조사의 소득분포가 참값이라고 간주한다. 다만 가계조사의 특성상 저소득을 과소포착하고 중간소득을 과대포착하는 표본추출의 문제는 근로소득의 포착률을 이용하여 해결한다. 가계동향조사에서 표본추출하는 방식이 소득의 종류에 따라 다르다로 볼 이유가 없으므로 근로소득의 비교를 통해 파악한 포착률의 차이가 사업소득에 대해서도 동일하게 적용된다고 볼 수 있다. 이러한 점을 고려하여, 1억원 이하 구간에서는 각 소득구간별로 가계동향조사의 인원을 근로소득 포착률로 나누어 종합소득 조정인원을 계산한다[1].

조정방법은 다음과 같다. 근로소득에 대해 가계동향조사에서 소득구간 에 조사된 인원수를 , 국세통계의 인원수를 T_k^L라고 하면, 근로소득의 소득 구간별 포착율은 $\pi_k^L = S_k^L / T_k^L$ 이 된다. 근로소득의 소득 구간별 포착률은 표본조사와 참값과의 차이를 나타내므로 이를 조사오차 비율이라고 간주한다. 가계동향조사와 국세통계의 종합소득 파악 인원을 각각 S_k^C, T_k^C 라고 하면, 소득구간별 종합소득 인원의 참값 T_k^{C*}은 다음과 같이 조정한다.

$$T_k^{C*} = \max\ S_k^C/\pi_k^L,\ T_k^C]$$

즉, <표 1>에서 나타난 것처럼, 가계동향조사가 전혀 포착하지 못하고 있는 3억원 이상의 구간에서는 국세통계의 인원이 참값이라고 간주하고, 3억원 이하 구간에서는 가계동향조사 인원을 조사오차 비율로 나눈 값이 참값이라고 간주한다.

이렇게 계산된 개인소득분포와 가계동향조사에서의 개인소득분포를 비교한 것이 <표 2>에 나와 있다. 위에서 설명한 바와 같이 국세통계와 가계동향조사를 결합하여 계산한 개인소득 분포가 참값에 해당한다. 2천만원이하 소득에서는 가계동향조사에서 파악된 인원이 국세통계에 비해 적다. 1천만원 이하 소득자는 가계동향조사에서 803만명으로 통합자료 대비 55.8%이다. 1-2천만원 구간에서 가계동향조사의 포착률은 83.5%이다.

---

1) <부록 1>에서는 결합자료를 준거로 하였을 경우와 국세통계만으로 준거 분포를 만들었을 경우 보정의 결과를 비교하였다. 보정 결과의 차이는 3% 미만이다.

<표 2> 국세통계와 가계조사의 개인소득분포 비교 (2014년) (단위 : 명)

소득구간	국세통계+가계동향조사 결합자료 (1)	가계동향조사 (2)	비율 (2)/(1)
0이하	18,139,382	24,399,495	1.345
0초과 1천이하	11,911,203	6,645,312	0.558
1-2천만	6,210,620	5,188,468	0.835
2-4천만	6,583,550	7,229,966	1.098
4-6천만	3,013,013	3,411,417	1.132
6-8천만	1,406,722	1,305,215	0.928
8천-1억	696,730	386,196	0.554
1-2억	671,736	193,181	0.288
2-3억	69,916	4,748	0.068
3-5억	37,186	0	0.000
5억이상	23,940	0	0.000
0이상 합계	30,624,616	24,364,503	0.796

(주) 소득이 0을 초과하는 사람만 계산한 것임. 국세통계에서 0 미만의 소득을 얻고 있는 사람 194,061명이 제외된 것임.

2천만원부터 6천만원까지 구간에서는 가계동향조사에서 파악된 인원이 결합자료보다 많다. 가계동향조사의 포착률은 2-4천만원에서 109.8%, 4-6천만원구간에서 113.2%이다. 가계동향조사의 포착률은 6천-8천만원 구간에서 92.8%이고, 소득이 높아 포착률이 점점 낮아져서 2-3억원 구간에서는 6.8%이다. 그리고 3억원 이상 소득자는 가계동향조사에 파악되지 않는다.

전체적으로 보면, 국세통계와 비교하여 가계동향조사는 중간소득을 과대포착하고 저소득과 고소득을 과소포착하고 있다. 가계동향조사로 측정한 불평등 지수는 결합자료에 비해 낮을 것으로 예상할 수 있다.

Ⅳ. 지니계수의 보정 방법

1. 지니계수의 계산

공식적인 지니계수는 균등화된 가구소득과 가구원수 가중치를 이용하여 계산한다. 가구소득은 근로소득, 사업소득, 재산소득, 사적이전소득을 합한 것으로 가구 시장 소득이라고도 한다. 가구소득을 , 가구원수를 h라고 하면, 균등화된 가구소득은 $y = y/ \quad h$ 이다. 표본조사의 가구 가중치를 w^h라고 할 때, 지니계수에서 사용하는 가구원수 가중치는 $w = w^h h$이다.

지니계수를 계산하는 방법은 다음과 같다 (Shalit, 1985 참조). 먼저 균등화된 가구소득을 작은 것에서 큰 것의 순서대로 정렬한다, 즉 $(j < k \Leftrightarrow y_j^e < y_j^e)$. 소득의 분포함수 (y^e)의 추정량 (estimator)은

(1) $$y^e) = \sum_{j=0}^{i-1} w_j + \frac{w_i}{2} \quad \text{단, } w_0 = 0,\ \sum_{i=1}^{n} w_i = 1$$

와 같다. 지니계수는 y^e와 $F(y^e)$의 가중 공분산(weighted covariance)을 이용하여 다음과 같이 계산된다.

(2) $$G(y,h) = \frac{2}{y^e} \sum_{i=1}^{n} w_i (y_i^e - \overline{y^e})(\hat{F}_i - F),$$

여기서 $\overline{y^e}, \overline{F}$ 는 가중평균값이다. 균등가구소득과 가중치는 가구소득과 가구원수로부터 계산된다. 따라서 지니계수는 가구소득과 가구원수의 함수라는 점을 나타내기 위해 $G(y,h)$라고 표현한다.

2. 지니계수 보정 방법

본 연구에서는 가계동향조사로 계산한 지니계수를 보정하기 위해 가중치 조정과 표본추가의 방법을 사용한다. 가계동향조사가 전혀 포착하지 못하는 고소득구간(예를 들어 2014년 3억원 이상 구간)에서는 표본추가의 방법을 사용하고, 나머지 구간에서는 가중치를 조정한다.

첫 번째로, 소득 포착율이 0인 아닌 구간, 예를 들어 3억원 미만 구간에 대해서는 가계동향조사에서의 개인소득을 국세통계와 비교하여 가구원의 소득이 속한 소득구간에 해당하는 소득포착율의 역수를 곱하여 가계동향조사의 가중치를 조정한다.

구체적인 방법은 다음과 같다. 소득구간 k에 대해, 가계동향조사에서 조사된 인원수를 S_k, 국세통계의 인원수를 T_k라고 하면, 소득 구간별 포착율은 $\pi_k = S_k / T_k$ 이 된다. 그리고 가구 i에 속한 가구원 p의 소득을 x_i^p 라고 하고 가구 가중치를 w_i^h라고 하자. 가구원의 소득 x_i^p가 소득구간 k에 속하는 경우 조정 가구원 가중치는 $w_{ip} = w_i^h / \pi_{ipk}$와 같다. 그러면 조정 가구 가중치 w_{in}^h는 다음과 같이 설정한다.

$$w_{in}^h = \frac{1}{P} \sum_{p=1} w_{ip} = \frac{1}{P} \sum_{p=1}^{P} \frac{w_i^h}{\pi_{ipk}}$$

여기서 P는 가구원 총수이다. 즉, 조정 가구 가중치는 조정 가구원 가중치의 평균값으로 설정한다.

두 번째로, '가계동향조사'가 포착하고 있지 않은 고소득 구간에서는 표본추가의 방법을 사용한다. 그런데, 국세통계에서는 개인 소득을 알 수 있지만, 지니계수를 계산하기 위해 필요한 가구소득과 가구원수는 알 수 없다. 즉, 고소득 구간에서 개인소득 x는 관측되지만, 가구소득 y와 가구원수 h는 관측되지 않는다. 관측되지 않은 자료 $z = (y,h)$는 결측 자료(missing data)이다.

식 (2) 지니계수 계산식을 관측자료와 결측자료의 관점에서 다시 표현하면 다음과 같다. (Tanner and Wong, 1987)

(3) $\theta|x) = \quad G(\theta|z,x)p(z|x)dz$

여기서 $G(\theta|x)$는 관측자료 x가 주어졌을 때 지니계수의 사후 밀도(posterior density)이고, $G(\theta|z,x)$는 완전한 자료가 주어졌을 때 지니계수의 사후 밀도이고, $p(z|x)$는 관측자료 x가 주어졌을 때 결측자료 z에 대한 예측 밀도(predictive density)이다.

식 (3)은 적분형태로 되어 있는데, 해를 구하기 불가능하므로 몬테칼로 방법 (Monte Carlo method)를 사용하여 계산한다. 식 (3)을 계산하기 위한 알고리듬은 다음과 같다.

(a) 예측 밀도 $p(z|x) = p(y,h|x) = p(y|h,x)p(h|x)$ 로부터 $z^{(1)},\cdots,z^{(\ )}$ 즉, $(y,h)^{(1)},\cdots,(y,h)^{(M)}$을 생성한다.

(b) 생성된 자료를 이용하여 지니계수를 계산하고 이를 평균한다.

$$G(\theta|x) \simeq \frac{1}{M}\sum_{m=1}^{M} G(\theta|z^{(m)}, x)$$

$p(h|x)$와 $p(y|h,x)$의 모수값은 처음에는 관측된 자료로 추정을 한다. 그리고 최초에 추정된 모수값에 근거하여 결측자료 z를 생성한다. 다음에는 확대된 자료 (x,z)로 $p(h|x)$와 $p(y|h,x)$의 모수값을 다시 추정한다. 이 과정을 반복하면서 각 단계에서 생성된 확대자료로 지니계수를 계산한다.

자료확대(data augmentation) 방법은 관측할 수 없는 자료나 잠재변수(latent variable)를 모형에 도입하여 반복적으로 최적화나 샘플링을 하는 방법을 가리키며 (van Dyk and Meng, 2001) 다음과 같은 두 단계로 구분된다. (Tanner and Wong (1987) 혹은 McLachlan and Krishnan (2008), 6.4 참조) (a)단계는 x가 주어졌을 때 z의 조건부 밀도를 이용하여 결측자료를 생성하는 것으로, 대치(imputation) 단계이다. (a)단계를 거치면 지니계수를 계산할 수 있는 완전한 자료 (x,z)가 생성되는데 이 완전한 자료가 확대된 자료(augmented data)이다. (b)단계는 확대된 자료로 지니계수의 사후 밀도를 계산하는 '사후' 과정이다.

(a)단계의 $p(y,h|x) = p(y|h,x)p(h|x)$ 를 계산하기 위해서는 개인소득이 주어졌을 때 가구원수의 밀도 $p(h|x)$와 개인소득 및 가구원수가 주어졌을 때 가구소득 y의 밀도 $p(y|h,x)$를 계산해야 한다. 이 예측 밀도들은 가계동향조사를 이용하여 추정한다. $p(h|x)$는 (가구원수-1)을 종속 변수로 하고 가구주 소득의 로그값은 설명 변수로 하여 프와송(poisson) 추정을 한다. $p(y|h,x)$는 각 가구원수 h에 대해 $\ln(y)$를 종속 변수로 하고 $\ln(x)$를 설명 변수로 하여 선형 추정을 한다.

본 연구에서 각 $m\,(=1,...,M)$ 단계마다 추가되는 표본의 개수는 생성된 자료에서의 가중치가 표본조사의 평균 가중치와 근접하도록 결정하였다. 각 구간에서 일양분포(uniform distribution)을 가정하여 생성한 자료는 구간당 100개이며, 새로 생성된 자료의 평균 가중치는 455이다. 그리고 (b)단계 몬테칼로 적분 계산을 위한 시행횟수 M=200 으로 하였다.[2)]

2) 홍민기 (2016)의 연구에서는 가계동향조사 원자료의 일부분을 누락시킨뒤 남은 자료로 보정방법을 적용하고 이를 원자료 지니계수값과 비교하는 방식으로 보정방법을 검증한바 있다. 보정 방법에서는 차이가 있지만 결과의 차이가 크지 않아서 본 논문에서는 검증 과정의 보고를 생략한다.

3. 시기별 자료의 형태에 따른 보정방법의 적용

시기별로 국세통계와 가계동향조사에서 포괄하는 소득, 가구의 범위가 다르다. 크게는 비과세대상자 포함여부, 1인가구 포함여부, 농가가구포함 여부가 시기마다 다르다. <표 3>에서 시기별 자료의 형태를 정리하였다. 일관된 지니계수값을 산출하려면 소득이나 가구의 범위를 같게 만들어야 한다. 본 연구에서는 비과세대상자, 1인가구, 농가가구를 모두 포함한 전체가구를 대상으로 지니계수를 계산하고 보정하고자 하였다.

<표 3> 시기별 자료의 형태

	비과세자 (국세통계)	1인가구 (가계동향조사)	농가가구 (농가경제조사)
2009~2015	포함	포함	포함
2006~2008	×	포함	포함
1998~2005	×	×	포함

2009년부터 2015년까지는 국세통계에 비과세자에 대한 소득통계가 있고, 가계동향조사에 1인가구도 포함되어 있고, 농가경제조사에서 농가가구도 포함되어 있다. 온전한 형태로 자료가 있는 이시기에 대해서는 앞에서 제시한 방법을 그대로 적용하여 지니계수를 보정한다.

2006년부터 2008년까지 근로소득 과세대상자에 대한 정보만 있고, 비과세대상자에 대한 정보는 없다. 이 시기에 1인가구와 농가가구는 포함되어 있다. 과세대상자에 대한 정보만 있으므로 불가피하게 과세대상자만으로 소득분포를 계산하고 이를 가계동향조사 소득분포와 비교하여 지니계수를 보정한다.

가계동향조사에서는 2006년부터 1인가구를 조사에 포함하였고 2005년까지는 2인이상 가구만 조사하였다. 지니계수를 보정할 때에는 1998년부터 2005년까지 1인가구를 가계동향조사 표본에 추가하였다. 1인 가구 추가는 2006년 1인 가구 표본을 각 년도마다 가중치와 소득을 조정하는 방식으로 하였다. 예를 들어, 인구총조사에서 2005년 1인가구는 3,170,675가구로 2006년 33,447,75가구의 94.8%에 해당한다. 따라서 2005년 1인가구의 가중치는 2006년 가구 가중치에 0.948을 곱한다. 즉, 2006년 1인 가구수를 $^{1}_{006}$, $(=1998, \cdots, 2005)$ 년도 1인 가구수를 H_t^1라고 하면, t년도 1인가구의 가중치는 $w_t^1 = H_t^1 / H_{2006}^1$ 와 같이 설정한다. 그리고 2006년과의 물가지수를 고려하여 t년도 1인가구의 소득을 조정한다.

1998년부터 농가경제조사가 있어서 가계동향조사 분배지표 자료를 구성할 수 있다. 그런데 1998년부터 2002년까지는 농가가구의 가구원수에 대한 정보가 없다. 가구원수는 인구총조사의 농가당 가구원수값을 이용하여 대치한다. 그리고 각 농가의 가중치는 인구총조사 농가수를 농가경제조사의 샘플수로 나누어 부여한다.

좀 더 자세히 들어가면, 국세통계연보에 제시된 소득의 범위가 시기마다 다르다. 근로소득 연말정산과 종합소득에 대한 자료는 모든 기간에 걸쳐 있다. 그런데 근로 소득자의 종합소득 신고 현황, 일용 소득자, 보험 방문 판매 사업자에 대한 통계는 2009년부터 있고, 2008년까지는 없다. 2008년까지는 있는 자료로만 개인소득분포를 구성할 수 있다[3] .

<표 4>에서는 비과세자를 포함한 자료로 지니계수를 보정한 결과와 과세자만을 포함한 자료로 지니계수를 보정한 결과를 비교하여 보여주고 있다. 똑같은 보정방법을 사용하였을 때 과세자만을 포함한 자료를 이용하면 지니계수가 약간 높게 계산된다.

자료가 완전한 2009년부터 2014년까지 비과세자를 포함한 자료로 보정한 지니계수값은 0.372, 0.381, 0.401, 0.383, 0.393, 0.394로서 과세자만 포함된 자료로 보정한 지니계수값의 95.3%~97.4%의 비율을 나타낸다. 이 두 값의 관계를 이용하여 과세자에 대한 자료만 있는 2008년 이전에 대해 전체 가구의 지니계수값을 계산한다. 이를 위해서는 보정비율(열(2))을 계산해야 한다. 소득세 납부에서 매년마다 과세자의 비율이 다른데, 과세대상자의 비율에 따라 보정비율도 달라진다고 가정하는 것이 합리적이다. 따라서 본 논문에서는 보정 비율과 과세자 비중의 관계를 다항 스플라인(polynomial spline) 모형으로 추정하고 이를 적용하여 보정 비율을 추정한다. 예를 들어, 2008년에는 과세자 비중이 57.9%이며 보정 비율은 0.961로 추정되었다. 과세자만 포함된 자료로 계산한 지니계수는 0.404이며 여기에 보정 비율을 곱한 값 0.383이 2008년 전체 가구의 지니계수 보정값이다. 이렇게 산출된 보정 비율을 과세자만 있는 자료로 계산한 지니계수(열 (1))에 곱하여 비과세자까지 포함한 전체 가구를 대상으로 한 보정 지니계수값을 계산한다.

<표 4> 지니계수 보정 과정

년도	비과세 포함 지니계수	과세자 지니계수 (1)	과세자 비중	보정 비율 (2)	보정 지니계수 (1)×(2)	표준편차
1998		0.335	0.676	0.970	0.311	0.0001
1999		0.346	0.588	0.960	0.325	0.0002
2000		0.332	0.534	0.956	0.316	0.0003
2001		0.345	0.558	0.960	0.325	0.0003
2002		0.355	0.611	0.954	0.336	0.0003
2003		0.358	0.558	0.960	0.339	0.0003
2004		0.359	0.539	0.957	0.340	0.0004
2005		0.378	0.513	0.952	0.363	0.0004
2006		0.392	0.526	0.955	0.377	0.0004
2007		0.393	0.526	0.955	0.376	0.0003
2008		0.404	0.579	0.961	0.383	0.0005
2009	0.372	0.396	0.597	0.958	0.372	0.0005
2010	0.381	0.404	0.609	0.955	0.381	0.0005
2011	0.401	0.414	0.639	0.955	0.401	0.0006
2012	0.383	0.397	0.673	0.969	0.383	0.0005
2013	0.393	0.407	0.687	0.974	0.393	0.0006
2014	0.394	0.415	0.519	0.953	0.394	0.0006
2015	0.401				0.401	0.0006

3) 1998년부터 2004년까지 국세통계에는 과세표준별 금액만 나와 있다. 2004년과 가장 가까운 2005년 자료에서 소득구간별로 임금/과세표준 비율을 계산한 뒤, 이 비율을 적용하여 과세표준을 소득금액으로 전환한다.

Ⅴ. 지니계수 보정 결과

<표 5>에는 보정한 지니계수, 통계청의 지니계수, 상위 10% 소득비중이 나와 있다. 그리고 <그림 1>에서는 보정 지니계수와 통계청의 지니계수를 비교하여 보여주고 있다. 보정지니계수는 1998년 0.311에서 2006년 0.377까지 급격히 증가하였다. 2007년부터 2012년까지 지니계수는 급격한 상승세를 멈추고 상승과 하락을 반복하고 있었는데, 2013년부터 약간 상승하면서 2015년에는 0.401에 이르고 있다.

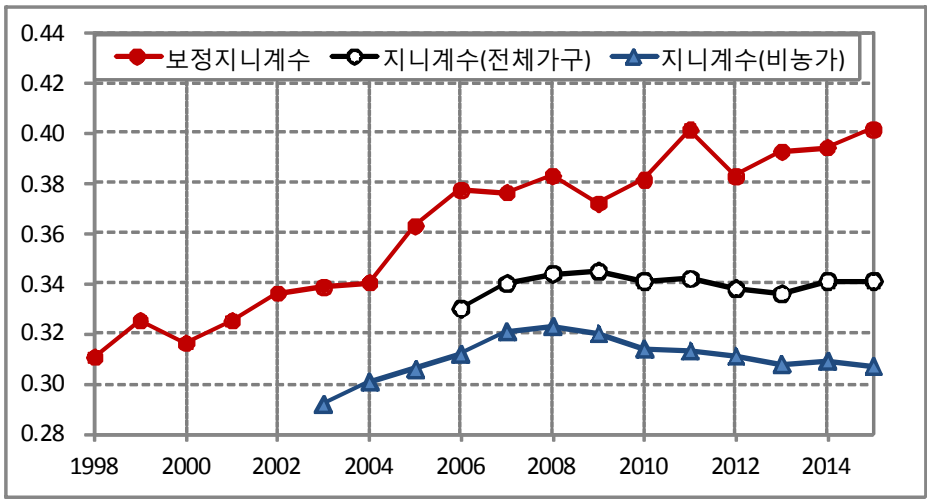
<표 5> 보정지니계수, 통계청 지니계수, 상위 10% 소득비중의 비교

년도	보정 지니계수	통계청 지니계수 전체가구	통계청 지니계수 2인이상비농가	상위10% 소득비중
1998	0.311			0.331
1999	0.325			0.329
2000	0.316			0.364
2001	0.325			0.387
2002	0.336			0.377
2003	0.339		0.292	0.368
2004	0.340		0.301	0.403
2005	0.363		0.306	0.440
2006	0.377	0.330	0.312	0.467
2007	0.376	0.340	0.321	0.463
2008	0.383	0.344	0.323	0.461
2009	0.372	0.345	0.320	0.456
2010	0.381	0.341	0.314	0.464
2011	0.401	0.342	0.313	0.474
2012	0.383	0.338	0.311	0.471
2013	0.393	0.336	0.308	0.473
2014	0.394	0.341	0.309	0.479
2015	0.401	0.341	0.307	0.485

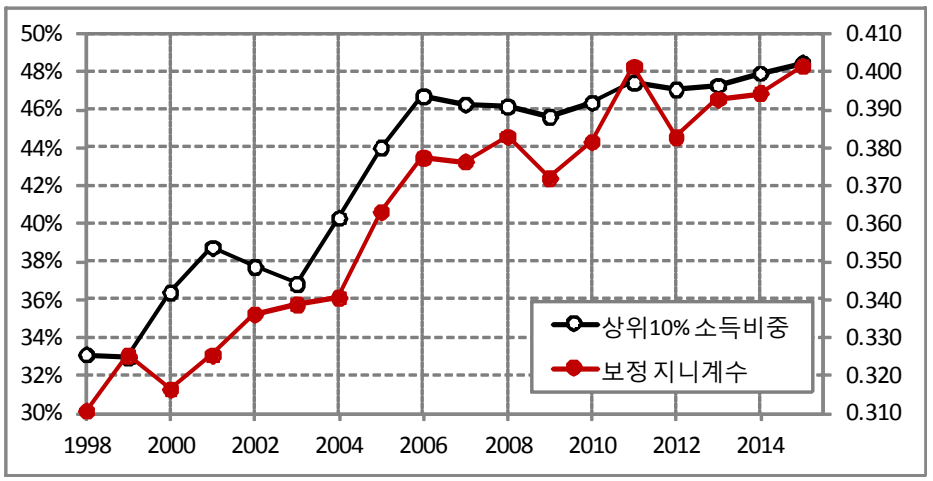
(출처) 상위 10% 소득비중 : 홍민기(2015)

그림에서 나타나듯이, 통계청의 전체가구 지니계수는 2006년 0.330에서 2009년 0.345로 약간 상승하는 추세를 보인다. 이 상승 추세는 같은 기간 보정 지니계수의 추세와 같다. 2010년대에는 추세의 차이가 나타난다. 통계청의 지니계수는 2013년 0.336까지 소폭 하락하는 반면, 본 연구에서 보정한 지니계수는 이 시기에 등락을 거듭하면서도 소폭 상승하고 있다.

[그림 1] 가계동향조사 지니계수와 보정값의 비교



[그림 2] 보정 지니계수(오른쪽 축)와 최상위 소득 비중(왼쪽 축)의 비교



두 지니계수 추세의 차이는 최상위 소득비중 추세와 관련이 있다. 최상위 1% 혹은 5%의 소득은 가계조사에서 과소포착된다. 따라서 최상위 소득 집단의 소득 비중이 커질수록 소득 과소 포착의 정도가 커지고 따라서 보정의 효과도 커진다. 2010년대 부터는 특히 최상위 1% 소득의 비중이 증가하는 추세여서 (홍민기, 2015 참조) 통계청 지니계수와 보정 지니계수의 차이가 커지고 있다.

[그림 2]에서는 보정 지니계수와 최상위 소득 비중을 비교하여 보여주고 있다. 2006년부터 2015년 기간 동안 보정 지니계수와 통계청 지니계수와의 상관계수는 0.08로서 매우 낮은데 비해, 보정지니계수와 최상위 10% 소득비중의 상관계수는 0.885로서 매우 높다. Leigh(2007)는 13개국 자료 연구를 통해 최상위 소득비중과 지니계수가 매우 강한 상관 관계를 갖는다는 것을 보여주었는데, 본 연구에서도 마찬가지 결과를 얻었다.

지니계수를 보정할 때 국세통계를 기준으로 하여 보정을 하였기 때문에 보정의 결과가 국세통계로 계산한 최상위 소득 비중과 상관관계가 높은 당연하게 보일 수도 있지만 그렇지 않다. 최상위 소

득 비중은 개인소득으로 계산한 것이고, 지니계수는 가구소득으로 계산한 것이다. 가구원간의 소득 상관성과 가구원수의 분포에 따라 개인소득과 가구소득 불평등도는 달라진다. 가구원간 소득 상관성이 낮고, 고소득 가구가 가구원이 많은 가구에 집중적으로 분포되어 있다면, 개인소득 불평등도가 높다고 하더라도 가구소득 불평등은 낮을 수 있다. 개인 소득 불평등이 가구 소득 불평등에 영향을 주기는 하지만 가구 분포의 변화에 따라 영향이 다를 수 있다. 일반적으로 1인이나 2인 가구의 증가하면 개인 소득 불평등과 가구소득 불평등의 상관관계가 높아진다.

Ⅵ. 결론

본 연구에서는 가계동향조사의 가중치를 조정하고 표본을 추가하여 1998년부터 2015년까지 지니계수를 보정하였다. 가중치 보정의 기준이 되는 개인소득 분포는 국세통계와 가계조사를 결합하여 구성하였다. 그리고 가구원수와 기타가구원 소득과 같은 결측 자료는 대체를 하여 완전한 자료를 만들고 지니계수를 반복 추정하는 베이지안 자료 확대의 방법을 사용하였다.

원래의 전체가구 지니계수에 비해 보정 지니계수는 10~18% 정도 높은 값을 보였다. 최상위 1%나 5% 소득 비중이 높아질수록 보정의 효과가 컸다. 표본 가계조사는 고소득을 과소포착하고 있기 때문에 최상위 소득비중이 커지면 보정의 효과가 더 커지게 된다.

원래의 지니계수는 2010년 이후 지니계수가 약간 하락하는 추세이지만, 보정 지니계수는 오히려 약간 증가하는 추세이다. 고소득을 과소포착하는 표본조사로 보면 2010년대 이후 소득 불평등이 감소하는 것처럼 보이지만, 보정한 지니계수나 최상위 소득 비중으로 보면 최근에도 소득 불평등이 증가하는 추세를 보여주고 있다.

개인 소득 최상위 10% 비중은 한국이 미국 다음으로 높은데 균등화 가구 시장 소득으로 측정하는 지니계수는 고소득 누락을 보정하더라도 한국이 낮은 편이다. 2013년 한국의 보정 지니계수는 0.393인데, 영국은 0.527, 독일은 0.508, 미국은 0.513, 프랑스는 0.504, 스웨덴은 0.443, 노르웨이는 0.412이다. (OECD statistics). OECD 국가와 비교하여 한국에서 개인소득 불평등이 매우 높지만 가구소득 불평등이 낮은 것은, 외국에 비해 가구원간의 소득 상관성이 낮고, 소득과 가구원수의 상관성이 높기 때문이다. 개인 소득과 가구 소득간의 관계가 어떻게 전개되었는지 추세를 알아보고 소득 불평등에 미치는 함의를 분석하는 연구가 필요하다.

참고문헌

국세통계연보, 각년도

가계동향조사, 각년도

인구총조사, 각년도

농가경제조사, 각년도

김낙년·김종일 (2013), “한국 소득분배 지표의 재검토”, 『한국경제의분석』 19(2), 1-50.

홍민기 (2015), “최상위 소득 비중의 장기 추세 (1958-2013)”, 『경제발전연구』 21(4), 1-34.

홍민기 (2016), 『불평등 지표 개선 연구』, 노동연구원 보고서.

Alvaredo (2011), "A note on the relationship between top income shares and the Gini coefficient", Economic Letters, 110: 274-277.

Burkhauser, R.V., Feng, S., Jenkins, S.P. and Larrimore, J. (2012), “ Recent trends in top income shares in the United States: Reconciling estimates from March CPS and IRS tax return data", Review of Economics and Statistics, 94(2), 371-388.

Cowell, F.A. and Flachaire, E. (2007), “Income distribution and inequality measurement: The problem of extreme values”, Journal of Econometrics, 141(2), 1044-1072.

van Dyk, D. and X. Meng (2001), The Art of Data Augmentation, Journal of Computational and Graphical Statistics, 10(1), 1-50.

Hlasny, V. and P. Verme (2016), Top Incomes and the Measurement of Inequality in Egypt, The World Bank Economic Review, lhw031. doi: 10.1093/wber/lhw031.

Jenkins, S.P., Burkhauser, R.V., Feng, S., and Larrimore, J. (2011) Measuring inequality using censored data: a multple-imputation approach to estimation and inference, Journal of the Royal Statistical Society, 174(1), 63-81.

Lakner, C. and Milanovic, B. (2013) Global income distribution from the fall of the Berlin Wall to the great recession, World Bank Policy Research working paper series #6719.

Leigh, A. (2007), How Closely Do Top Income Shares Track Other Measures of Inequality, The Economic Journal, 117. 589-603.

McLachlan, G. and T Krishnam (2008), The EM Algorithm and Extensions, 2nd Edition, John Wiley & Sons, Inc.

Shalit, H. (1985), Calculating the Gini Index of Inequality for Individual Data, Oxford Bulletin of Economics and Statistics 47(2), 185-189.

Tanner, M. and W. H. Wong (1987), The Calculation of Posterior Distribution by Data Augmentation, Journal of American Statistical Association, 82 (398), 528-540.

<부록 1> 개인소득 준거가 다른 경우

여기서는 개인소득 분포의 준거로 국세통계-가계동향조사를 결합한 자료를 사용하였을 경우와 국세통계로만 사용하였을 경우 보정의 결과가 얼마나 다른지를 설명한다. 보정 결과의 차이를 <부표 1>에서 보여주고 있다.

2015년 보정 지니계수는 결합자료를 준거로 하였을 경우 0.401이고 국세통계를 준거로 하였을 경우 0.398로서 1% 정도 차이가 난다. 2009년부터 비교해보면 준거가 되는 개인소득 분포를 달리해서 발생하는 결과의 차이는 3%미만이다.

결합자료에 비해 국세통계 개인소득 분포는 저소득 인원이 적다. 고소득에서는 두 자료의 차이가 없다. 따라서 국세통계자료로 기준을 삼으면 저소득 가구의 가중치가 과대평가되고, 중간소득 가중치가 과소평가된다. 이 둘의 효과가 상쇄되어 전체 지니계수에 미치는 영향은 매우 적다.

<부표 1> 개인소득 분포의 준거가 다를 경우 보정지니계수 비교

년도	국세통계, 가계조사 결합자료 (1)	국세통계 (2)	비율 (2)/(1)
2009	0.372	0.381	1.02
2010	0.381	0.391	1.03
2011	0.401	0.403	1.00
2012	0.383	0.381	1.00
2013	0.393	0.384	0.98
2014	0.394	0.394	1.00
2015	0.401	0.398	0.99