한국경제학보 제22권 제1호
The Korean Journal of Economics Vol. 22, No. 1 (Special Issue 2015)

통계청 가계조사자료에 기초하여 계산된 상위소득점유율*

신 주 형 · **김 윤 미** · **김 태 환**

요 약

Piketty(2014)의 방식을 따라서 국세청의 조세자료를 사용하여 계산된 상위소득점유율의 추정결과가 김낙년·김종일(Kim and Kim, 2015)의 연구에서 보고되었다. 조세자료의 이용을 비판하는 사람들은 납세자들이 세액이 적게 산출될 수 있는 방법으로 소득을 신고할 유인이 있다는 점을 지적한다. 또한 조세자료에서는 납세자들의 사회·경제적 특성들을 함께 파악하기가 어렵다. 이러한 점들 때문에 기존의 많은 불평등 연구는 가계조사자료를 이용하였다. 이 두 연구 분야를 연결하기 위해서 Burkhauser et al.(2012)은 미국에서의 가계조사에 해당하는 CPS를 이용하여 상위소득점유율을 계산하여 이를 조세자료를 이용하여 계산된 상위소득점유율과 비교·분석하였다. 본 논문에서도 이러한 문제의식에 입각해 통계청에서 매년 실시하는 가계조사를 통해 생성된 가계조사자료를 사용하여 상위소득점유율을 계산하였고, 이러한 결과를 조세자료의 결과와 비교·분석하였다.

주제분류 : B030104, B030400
핵심 주제어 : 상위소득점유율, 가계조사, GB2분포

Ⅰ. 서 론

노동과 자본의 분배 몫인 임금과 이윤은 각 요소의 한계생산성에 의해서

* 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013S1A3A2053799).
** 연세대학교 경제학과 대학원생, e-mail: joohyungshin@gmail.com
*** 서울시립대학교 경제학과 부교수, e-mail: kimy@uos.ac.kr
**** 교신저자, 연세대학교 경제학과 교수, e-mail: tae-hwan.kim@yonsei.ac.kr

시장에서 자동적으로 그리고 공정하게 결정된다는 한계생산성이론이 근대 경제학의 핵심토대를 이루고 있기 때문에, 분배이론은 흥미 있는 연구주제가 되지 못하였다. 그럼에도 프랑스(Piketty, 2003), 미국(Piketty and Saez, 2003), 영국(Atkinson, 2005) 등에서 일군의 경제학자들은 분배이론 특히 소득불평등을 지속적으로 연구하였고 그 결과가 Piketty(2014)의 〈21세기 자본〉으로 집대성되어 경제학계에 큰 반향을 초래하였다. 그 이유는 미국을 포함한 많은 선진 국가들에서 소득불평등이 심각한 수준이며 진행속도가 매우 빠르다는 것이다. 예를 들어, 미국의 경우 1950-1960년대에 상위10%가 약 30%의 소득을 점유하였는데, 1970년대부터 이 수치가 급속히 상승하였고, 2013년 현재 약 47%의 소득을 점유하고 있다.

한국의 경우에는 1970년대 이후 소득불평등에 대한 꾸준한 연구가 진행되어 왔지만 주로 가계조사 등과 같은 표본에 근거하여 계산된 지니계수 등의 지표를 사용하였다(주학중, 1979; 김대모·안국신, 1987; 윤기중, 1997; 정수빈·민은기·김태환, 2014). 반면에, Piketty(2014)의 방식을 따라서 상위소득점유율에 의거하여 한국의 소득불평등을 연구한 결과는 김낙년·김종일(Kim and Kim, 2015)등에서 찾아볼 수 있다. 김낙년·김종일(2015)은 조세자료를 사용하여 1933년부터 2012년까지 상위 소득계층의 소득점유율을 구하였고, 이 결과가 파리경제대학에서 운영하는 〈The World Top Incomes Database〉에 등재되었다.

조세자료는 거의 모집단에 근접한 자료이기에 샘플링에러를 제거할 수 있다는 장점이 있다. 하지만, 납세자들이 세액이 적게 산출될 수 있는 방법으로 소득을 신고할 유인이 있고, 또한 이러한 유인은 세제 및 세율에 따라서 달라진다는 점이 조세자료의 한계로 지적된다. 이러한 조세자료의 문제점은 Piketty and Saez(2003)에서도 자세히 논의되고 있다. 반면에 가계조사가 이상적으로 설계되고 참여자들의 익명성이 완벽히 보장된다면 이러한 문제점을 해결할 수 있다. 하지만 현실적으로는 가계조사에서도 낮은 응답률, 소득의 과소보고, 측정오차 등의 문제가 발생된다. 그럼에도 불구하고 소득불평등을 연구한 기존의 대부분의 논문은 가계조사자료를 사용하였다. 이러한 두 연구 분야를 연결하려는 시도로 Burkhauser et al.(2012)은 미국에서의 가계조사에 해당하는 CPS(Current Population Survey)를 이용하여 상위소득점유율을 계산하여 이를 조세자료를 이용하여 계산된 결과와

비교·분석하였다. 본 논문에서도 이러한 문제의식에 입각해 통계청에서 발표하는 가계조사자료를 사용하여 상위소득점유율을 계산하였고 이러한 결과를 조세자료를 이용하여 계산된 김낙년·김종일(2015)의 상위소득점유율과 비교·분석하였다.

Ⅱ. 자료에 대한 설명

우리나라 가계조사의 경우 1990년부터 2002년까지는 2인 이상 도시가구가, 2003년부터 2005년까지는 2인 이상 전국가구가 조사 대상이었고, 2006년부터 현재까지는 1인 이상 전국가구가 조사대상이다(각각 '도시가계조사,' '가계조사,' '가계동향조사'로 불림). 본 연구는 이 세 가지 조사자료 모두를 분석대상으로 하고 있고, 기간은 1990년부터 2013년까지이다. 매년 조사대상이 되는 가구 수는 약간씩 변동되는데 마지막해인 2013년도의 조사대상 가구 수는 10,046가구였다. 아래에서는 이 세 가지 자료를 모두 통틀어서 '가계조사'라고 부르기로 한다. 전 기간을 걸쳐서 비겸업, 비외국인 혈연 가구만이 조사대상이고 농가, 임가, 어가는 조사대상서 제외된다. 김낙년·김종일(2015)과 가능한 한 맞춰서 과세대상소득인 급여소득, 상여금, 사업소득, 임대소득, 재산소득, 이자소득, 배당소득, 퇴직금, 연금 등을 포함하였고, 과세대상이 아닌 이전소득은 제외하였다.

Ⅲ. 상위소득점유율 추정방법

본 연구의 목적은 가계조사 자료를 이용하여 가구 소득분포에서 도출한 소득점유율을 구하는 것이다. 즉, 본 연구에서 상위10%는 상위10%의 개인들이 아니라 상위10%의 가구들을 의미한다. 이는 가계조사의 단위가 개인이 아니라 가구이기 때문이다.[1] 소득분포를 추정하는데 있어서 기존에는

1) 가계조사를 이용하여 'Equivalence Scales'라는 방법을 사용하면 개인의 소득분포를 추정할 수도 있다. 하지만, 이를 위해서는 모든 가구 구성원의 나이 정보가 필요한데, 이러한 정보가 가계조사자료에 제공되어 있지 않다.

파레토분포를 많이 사용하였지만, 최근에는 더욱 유연한 Generalized Beta of the Second Kind(GB2)가 주로 사용된다(Jenkins, 2009). GB2분포는 다음과 같다; $f(x)=\frac{ax^{ap-1}}{b^{ap}B(p,q)\left[1+(x/b)^a\right]^{p+q}}$. GB2분포는 양수인 x에 대해서만 정의되고, 분포의 모습을 결정하는 4개의 모수인 a,b,p,q 는 모두 양수이다. 또한 $\Gamma(z)$를 감마함수라고 할 때, $B(p,q)$는 $B(p,q)=\Gamma(p)\Gamma(q)/\Gamma(p+q)$로 정의되는 베타함수이다.

1. 분포의 모수를 알고 있을 경우 상위소득점유율 추정방법

소득분포가 GB2분포를 따르고 GB2분포의 4개의 모수를 모두 안다면, 아래와 같이 상위소득점유율을 구할 수 있다. 먼저 $H(z)$라는 함수를 정의한다; $H(z)=\int_z^\infty f(x)dx$. 즉, $H(z)$는 소득수준이 z보다 높은 가구의 비율이다. 여기에 국민경제 전체 가구 수인 n을 곱한 $n\times H(z)$는 소득이 z보다 높은 가구의 수이다. 또한 $G(z)$라는 함수를 정의한다; $G(z)=\int_z^\infty xf(x)dx$. 즉, $G(z)$는 소득수준이 z보다 높은 가구들의 평균소득이다. 따라서 $n\times G(z)$는 소득이 z보다 높은 가구들의 총소득이다. 따라서 다음과 같이 정의된 $A(z)=\frac{G(z)}{H(z)}$는 z보다 높은 소득구간에 속하는 가구들의 평균소득이 된다.

만약, n_z이 소득이 z보다 높은 전체가구의 수이고 I_0가 국민경제의 전체 소득이라면, '소득이 z보다 높은 가구들의 소득 총합이 국민경제의 전체 소득에서 차지하는 비중'은 다음과 같이 계산 된다; $\frac{n_z\times A(z)}{I_0}$. 예를 들어, 상위 10%의 소득점유율을 구하려면 $\int_z^\infty f(x)dx$가 10%가 되는 $z=z_{0.1}$의 값을 구한 다음, 위의 과정을 거쳐서 $A(z_{0.1})$를 구한다. 그 다음 소득이 $z_{0.1}$보다 많은 가구의 수 $n_{z_{0.1}}$와 국민경제 전체소득 I_0를 구해 위의 식을 사용하여 상위10%가구의 소득점유율을 계산할 수 있다.[2)]

2) 본 논문에서 사용한 소득점유율 계산방식은 김낙년·김종일(2015)에서 사용한 계산방식과 기본적으로 동일하다. 차이점은 김낙년·김종일(2015)에서는 파레토분포를 사용한 것이고, 본 논문에서는 보다 일반적인 GB2분포를 사용한다는 점이다. 파레토분포를 사용하면 z보다 높은 소득구간에 속하는 가구들의 평균소득에 해당되는

위의 계산을 위해서는 기준모집단(reference population)과 기준소득(reference income)에 해당되는 '국민경제 전체 가구 수'와 '국민경제 전체 소득'을 알아야 한다. 이 두 정보를 어떻게 추정하는가에 따라서 계산된 소득점유율에 상당한 차이가 존재할 수 있다. 개인소득분포를 논의하는 김낙년·김종일(2015)의 연구에서는 reference population을 통계청에서 보고하는 '20세 이상 총인구수'로 설정하고 있다. 반면에 가구소득분포를 분석하는 본 연구에서는 전체가구 수를 추정하여야 한다. 국민경제 전체 가구 수인 n은 두 가지 방법으로 추정할 수 있는데, 첫째는 가계조사자료에서 보고되는 가중치의 합으로 구할 수 있다. 통계청에서 계산하는 이 가중치는 i번째 가구가 모집단에서 추출될 확률의 역수로써 각 가구가 대표하는 가구의 수를 나타낸다. 둘째는 통계청의 '장래가구추계' 자료를 이용한다. 소득이 $z_{0.1}$보다 많은 가구의 수 $n_{z_{0.1}}$는 $n\times 0.1$로 구한다. 전체소득 I_0를 구하는 방법도 앞에서 전체 가구 수를 구하는 것처럼 두 가지 방법으로 추정할 수 있다. 첫째는 가계조사자료상 소득×가중치의 합을 이용할 수 있고, 둘째는 김낙년·김종일(2015)이 구한 방법대로 한국은행 경제통계시스템 국민계정 중 소득계정에서 통상적인 임금이 아니거나 비과세소득인 요소들은 제거한 전체 소득을 이용한다.

2. 분포의 모수를 알고 있지 못할 경우 상위소득점유율 추정방법

위의 모든 논의의 전제조건은 GB2분포의 모수인 a,b,p,q를 연구자가 알고 있다는 것이다. 하지만, 현실적으로 연구자가 이러한 모수들을 알 수 없으며 소득분포자료로부터 추정되어져야 한다. 본 연구에서는 MLE(maximum likelihood estimation)추정을 하였다. 즉, 각 년도마다 다음의 우도함수(log-likelihood function)를 극대화하는 계수 추정치 $(\hat{a},\hat{b},\hat{p},\hat{q})$ 를 구한다; $\ln L=\sum_{i=1}^{N} w_i \ln f(x_i)$. 여기서 N은 표본의 관찰치 수를 의미하고 w_i는 가구 i에게 부여되는 가중치이다. 이 가중치가 1이면 일반적인 MLE가 된

$A(z)$가 z와 파레토분포의 모수들의 간단한 함수로 표현되어 계산이 간단해진다. 반면에 보다 일반적인 GB2분포를 사용하면 추정(fitting)을 향상시킬 수 있다는 장점이 있지만, 본문에서 설명한 바와 같이 수치적분이 필요하다.

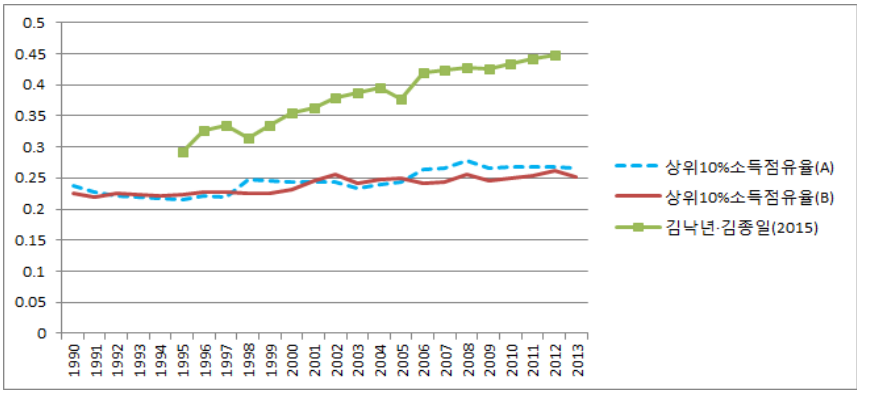
다. 본 연구의 추정결과에 의하면 가중치를 사용한 MLE와 사용하지 않은 MLE결과에 큰 차이가 존재하지 않았지만, Burkhauser et al.(2012)의 방법을 따라서 가중치를 사용한 MLE를 적용한다. 위의 우도함수를 추정할 때 모든 모수들이 양수라는 제약을 주어야하는데, 이러한 제약으로 인해 초래되는 극대화의 기술적인 문제점(예를 들면, local maximum으로의 수렴)들을 해결하기 위하여, Chib and Kang(2013)에서 사용하듯이 stochastic optimization과 standard Newton-Raphson deterministic optimizer를 결합한 형태의 optimization를 사용한다. MLE에 의해서 모수들이 추정되면, 이러한 추정치를 바탕으로 추정된 GB2분포를 구할 수 있다. 추정된 GB2분포를 $\hat{f}(x)$라고 한다면, 앞에서 설명한 모든 계산방식을 $f(x)$ 대신 $\hat{f}(x)$를 사용하여 진행하면 된다. 이때, 모든 적분은 Matlab에서 수치적분(numerical integration)으로 계산하였다.

Ⅳ. 추정결과 및 향후연구과제

위의 방법에 기초하여 계산된 '상위10%소득점유율'은 아래의 〈그림 1〉에 나타나 있다. 〈그림 1〉에는 세 가지 종류의 상위10%소득점유율 시계열이 있는데, 첫째는 3장의 방법을 사용하되 전체 가구 수와 전체 소득을 가계조사자료에서 보고되는 가중치를 사용한 경우〔상위10%소득점유율(A)〕에 해당되고, 둘째는 역시 3장의 방법을 사용하지만, 전체 가구 수를 장래가구추계 자료에서 구하고 전체 소득을 국민계정에서 구해서 계산된 경우〔상위10%소득점유율(B)〕이며, 세 번째는 김낙년·김종일(2015)에서 보고된 상위10%소득점유율이다.

먼저, 가계조사자료에 의거하여 계산된 두 가지 종류의 소득점유율(A, B)은 큰 차이를 보이지 않으며 많은 경우 움직이는 방향도 일치하고 있다. 1990년에 상위10%의 가구는 전체 소득의 약 23%를 점유한다. 이후 90년대에는 다소 완화되는 경향이 있으나, A시계열에 의하면 외환위기 직후인 1998년부터 25%정도로 급등하였고, B시계열의 의하면 외환위기 직후부터 완만한 상승세를 보이고 있다. 이후의 추세는 A시계열의 경우에는 2000년대 초반부터 지속적인 완만한 상승세를 보이고 금융위기가 있었던

【그림 1】 상위10%소득점유율



주: 3장에서 설명한 GB2분포를 사용하여 추정한 상위10%소득점유율; (A)는 가구수와 총소득을 가계조사자료내에서 구한 것이고, (B)는 가구수와 총소득을 통계청 및 국민소득계정에서 구한 것임. 김낙년·김종일(2015) 시계열이 1995년부터 시작되는 이유는 종합소득을 기준으로 한 이 시계열이 1933-1940년, 1979-1985년, 1995-2012년에만 계산되어 있기 때문임.

2008년에 28%로 정점을 찍은 후에 다소 감소한 후에 큰 변동이 없다. 반면에 B시계열의 경우에는 2002년에 약 25%의 점유율로 상승한 후에 2013년까지 큰 변동 없이 유지되고 있다. 따라서 통계청의 가계조사자료에 의하면 한국사회에서 상위10%가구들의 소득점유율은 해당 24년의 기간 동안 22%에서 28%의 범위 내에서 급격한 변화 없이 안정적으로 유지되어왔다고 할 수 있다. 그러나 이러한 다소 낙관적인 묘사는 김낙년·김종일(2015)에서 보고된 상위10%개인들의 소득점유율과도 매우 다른 모습을 보여주고 있다. 조세자료에 기초해서 계산된 상위10%개인들의 소득점유율은 1995년부터 2012년까지 두 차례의 일시적 감소(1998년과 2005년)만 있었을 뿐 전 기간에 있어서 가파른 상승세를 보이고 있다. 1995년에는 상위10%가 전체소득 중에서 29%를 가져갔지만, 2012년도에 이르면 전체소득 중에서 약 49%를 가져가고 있다.

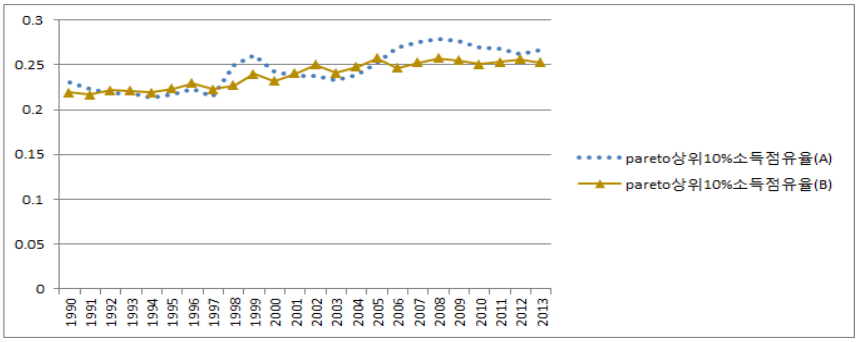
가계조사자료와 조세자료가 왜 이처럼 커다란 차이를 보이는가에 대해 심도 깊게 규명하는 것은 추후의 중요한 연구과제일 것으로 판단된다. Burkhauser et al.(2012)의 연구에 의하면 미국의 경우에도 가계조사자료에 기초한 소득점유율이 조세자료에 기초한 소득점유율보다 낮은 경향을 보이고는 있지만, 〈그림1〉에서와 같은 커다란 차이는 보고되지 않는다. 첫 번째 가능한 설명은 가계조사가 진행될 때 고소득층 가구들의 응답률이 낮

을 수 있고 그러한 경향이 한국사회의 어떤 특수성으로 인해 증폭되었다는 것이다. 만약 상위소득층의 비응답률(non-response rates)이 심하다면 가계조사자료에 기초한 상위소득점유율 추정에 심각한 '비응답편의(non-response bias)'가 존재할 수 있다. 일반적으로 가계조사자료에서 이러한 편의로부터 자유롭기 위해서는 최소한 85%의 응답률이 필요한 것으로 알려져 있다(Groves, 2006). 또한 현대자본주의경제의 최상층부에 위치한 최고소득층은 일반적으로 베블린이 언급했던 전근대적 유한계층이라기보다는 CEO등을 포함한 '일하는 부유층(Working Rich)'에 해당된다. 가계조사가 진행되는 방식은 참여가구에게 발생한 소득은 물론 식료품 및 의류 등 실생활에서의 소비지출항목에 대한 가계부를 1년 동안 작성하도록 요구하고 있기 때문에 Working Rich들이, 설사 그들이 조사의 익명성을 신뢰하고 소득을 공개하는 것에 대한 거부감이 전혀 없더라고 하더라도, 이러한 가계조사에 참여할 가능성은 낮아 보인다. 두 번째 가설은 고소득자들 중에 일부가 조사에 참여하였다고 하더라도 그들 중 일부는 소득을 과소보고한다는 것이다. 소득의 과소보고는 조사 진행 과정이 1년 동안 진행되면서 익명성이 유지되기 힘들다는 측면에 의해 유발될 수도 있고, 가계조사에서 측정하고자하는 소득의 개념에 대한 불확실성에 의해 유발될 수도 있다. 예를 들어, 고소득층의 주요 소득인 자본이익의 경우 어떤 항목에 기입해야할지가 불분명하며 또 다른 소득인 주식인수권(stock options)의 경우에는 해당 항목이 아예 포함되어 있지 않다. 세 번째 가능한 설명은 측정오차(measurement error)이다. 가계조사의 경우 응답자들의 신원보호를 위해, 또는 이상치(outliers)나 기록실수(recording error)의 방지를 위해서 응답자가 기입한 소득 자료의 조정이 이루어지는데, 이러한 전 과정에서 측정오차가 발생할 수 있다. 마지막으로 조세자료에 존재하는 문제점, 즉 납세자들이 세액이 적게 산출될 수 있는 방법으로 소득을 신고할 유인이 있다는 문제점이 조세자료와 가계조사자료의 결과에 어떠한 영향을 미치는지도 추후에 함께 연구해야할 것으로 판단된다.

가계조사자료에 대한 추가적인 분석결과가 〈그림 2〉에서 보고되고 있다. 〈그림 1〉에서는 GB2분포를 사용하여 소득점유율을 추정하였는데, 〈그림 2〉에서는 각 해당연도의 소득자료에서 상위소득10%를 추정할 수 있도록 모수의 일부에 제약을 부여한 파레토분포를 사용하여 소득점유율을 추정하

였다. 〈그림 2〉에서 보고된 소득점유율의 시계열을 살펴보면, 〈그림 1〉에서의 결과와 큰 차이가 없다는 것을 알 수 있다.

【그림 2】 파레토분포를 사용한 상위10%소득점유율



주: 3장에서 설명한 계산방식을 사용하되, 파레토분포를 사용하여 추정한 상위10%소득점유율; (A)는 가구수와 총소득을 가계조사자료내에서 구한 것이고, (B)는 가구수와 총소득을 통계청 및 국민소득계정에서 구한 것임.

투고 일자: 2015. 3. 27. 심사 및 수정 일자: 2015. 4. 18. 게재 확정 일자: 2015. 4. 25.

◈ 참 고 문 헌 ◈

김대모 · 안국신 (1987), 『한국의 소득분배 및 그 결정요인과 분배문제에 대한 국민의 의식구조』, 문교부.

윤기중 (1997), 『한국경제의 불평등 분석』, 박영사.

정수빈 · 민은기 · 김태환 (2014), "지니계수의 확장 및 이를 이용한 한국사회의 소득불평등 요인 분석," 『사회경제평론』, 제45호, pp.185-230.

주학중 (1979), 『한국의 소득분배와 결정요인』, 한국개발연구원.

Atkinson, A. B. (2005), "Top Incomes in the UK over the Twentieth Century," *Journal of the Royal Statistical Society, Series A*, Vol. 168, No. 2, pp.325-343.

Burkhauser, R. V., S. Feng, S. P. Jenkins and J. Larrimore (2012), "Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data," *The Review of Economics and Statistics*, Vol. 94, No. 2, pp.371-388.

Chib, S. and K. H. Kang (2013), "Change-Points in Affine

Arbitrage-Free Term Structure Models," *Journal of Financial Econometrics*, Vol. 11, No. 2, pp.302-334.

Groves, R. M. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, Vol. 70, No. 5, pp.646-675.

Jenkins, S. P. (2009), "Distributionally-Sensitive Inequality Indices and the GB2 Income Distribution," *The Review of Income and Wealth*, Series 55, No. 2, pp.392-398.

Kim, N. N. and J. Kim (2015), "Top Incomes in Korea, 1933-2010: Evidence from Income Tax Statistics," Hitotsubashi Journal of Economics, forthcoming.

Piketty, T. (2014), *Capital in the Twenty-First Century*, The Belknap Press of Harvard University Press.

_____ (2003), "Income Inequality in France, 1901-1998," *Journal of Political Economy*, Vol. 111, No. 5, pp.1004-1042.

Piketty, T. and E. Saez (2003), "Income Inequality in the United States, 1913-1998," *Quarterly Journal of Economics*, Vol. 118, No. 1, pp.1-39.

Top Income Shares Based on Household Income and Expenditure Survey Data

Joo-Hyung Shin* · Yunmi Kim** · Tae-Hwan Kim***

Abstract

Following the methodology proposed by Piketty (2014), Kim and Kim (2015) have estimated top income shares in Korea based on income tax data published by the National Tax Service. However, a potential problem with tax data is that tax filers can have financial incentives to legally manipulate the way to report their income to minimize their tax liabilities. To reconcile the problem, Burkhauser et al. (2012) have used CPS data in the USA to compute top income shares and compared their results with the outcomes based on tax data. We have used the Household Income and Expenditure Survey data published by Statistics Korea to compute top income shares in Korea and compared our results with Kim and Kim (2015)'s results.

KRF Classification : B030104, B030400
Key Words : top income shares, household income and expenditure survey, GB2 distribution

* Graduate Student, School of Economics, Yonsei University, e-mail: joohyungshin@ gmail.com
** Professor, Department of Economics, University of Seoul, e-mail: kimy@ uos.ac.kr
*** Corresponding Author, Professor, School of Economics, Yonsei University, e-mail: tae-hwan.kim@yonsei.ac.kr