

高净值人群修补技术新进展^{*}

万海远

【摘要】为提高中国住户调查数据中的高净值人群代表性,文章综述了近年来高净值人群修补技术的最新进展,比较各种参数和非参数估计技术的优劣。结果发现,资本化方法、半参数自助法、新拓展幂律分布法和现代数据拼接技术等前沿方法可以有效校正高净值样本,使用不同技术修补后的中国收入差距水平明显上升,甚至变化趋势也可能被逆转,因此有必要修补高净值人群问题。文章建议,在中国的应用中,要尽可能增加样本数量、保留异常值信息、使用半参数自助法及基尼系数指标等,综合考虑高净值数据可得性、研究对象、分析单位和指标协调性等来选择具体的修补技术。加强住户调查数据真实性的核查环节,从源头上提升高净值人群的遵从率,加大税务和行政部门微观数据的开放力度,提供更多、更细的收入分组区间信息,用数字化技术获取高收入人群数据,从而改善住户调查数据对高净值人群的代表性。

【关键词】住户调查 高净值人群 漏损低报 修补技术

【作者】万海远 北京师范大学经济与工商管理学院,教授。

住户调查数据是公共政策制定的基础,但高净值人群缺失和低报问题往往使基于住户调查数据的研究结论出现偏差(Meyer等,2015)。由于顶端人群数量非常有限,简单随机抽样方法很难抽取到足够的样本,同时高净值人群一般不太愿意配合调查而存在较高的瞒报率,导致随机抽样很难捕捉到足够多的高净值样本,从而住户调查数据通常存在右尾端删失,由此会造成基础性统计数字失真(Korinek等,2006)。

近年来,因高净值人群缺失、低报所带来的问题越来越严重(Burkhauser等,2012)。各国基于住户调查数据的研究结果都可能出现偏差,在补充最高1%样本后收入差距的变化趋势甚至可能被改变。中国住户调查数据显示,全国收入差距从2009年开始出现七连降,但考虑高净值样本后的收入差距不仅没有下降,反而出现上升态势(罗楚亮,2019)。在住户调查数据中如何修补高净值样本,成为社会科学领域的重要问题。

^{*} 本文为国家社会科学基金重大项目“中国农村家庭数据库建设及其应用研究”(编号:18ZDA080)的阶段性成果。

近年在如何补充高净值样本方面,出现不少新的估计技术和研究成果,但中国大多数以住户调查数据为基础的实证研究,在应用前都没有认真修补高净值人群缺失问题,由此导致基于住户调查数据的研究结论可能存在偏差,在估算经济社会领域指标平均值或分布状况时还可能出现明显误导。鉴于此,本文试图综述近年来高净值人群修补技术的最新进展,比较不同修补技术的优劣,并结合中国实际情况提出具体的应用建议。

一、行为反应推导技术

(一) 最优避税行为

通常高净值人群除了不愿参加调查外,而且更有可能低报自己的收入,其中既有记忆不清楚等客观原因,也存在为避税而故意瞒报的可能(白重恩等,2015)。在统计稽查体系既定的情况下,收入瞒报与避税动机的关系可能是稳定存在的,收入越高则瞒报程度越高(Hurst等,2014)。一般认为与税收相关的收入种类是税收瞒报的重点对象,Ales等(2016)在关于企业高管的研究中,将个体的基本工薪收入作为控制组,工资性收入中的福利、绩效和奖金作为处理组,由此区分高净值群体收入隐瞒程度的差异。自雇佣者出于避税考虑会有意或无意地瞒报收入,因此是税收规避的重要来源(Lyssiotou,2004)。假定自就业群体的避税动机要高于工资性就业者,通过估计中低收入群体中两个就业类型的隐瞒参数差异并类比到高净值群体,就可以从工资性收入推测出高净值人群被隐瞒的自雇佣收入(Hurst等,2014)。白重恩等(2015)基于家庭成本收益最优化,引入瞒报的成本与收益函数,将个体瞒报带来的少交税作为收益、被税务稽查而被罚款作为成本,由此推导住户收入瞒报的反应函数,并估计高收入组的真实收入。

总的来看,从避税视角研究高净值人群的方法,是假设与收入瞒报有关的交易活动会体现在可观测的指标中,由此根据某些稳健的统计规律予以倒推。这类方法通常会假设自雇性就业的收入瞒报率最高,其他就业群体较少或没有瞒报,但对该假设的合理性并没有进行检验。另外,从避税视角解释收入瞒报仍然值得探讨。中国统计法明确住户调查只是为了总体统计需要,个人收入信息不会泄漏给税务部门。事实上,瞒报更多的还是出于避免露富的考虑,或者是由于记忆误差和保护隐私的需要,因此从避税行为反推高净值人群的方法还需要进一步探讨。

(二) 消费行为模型

通过个体行为数据拟合高净值人群的方法,还包括消费行为模型和慈善捐赠模型等。一般认为消费数据的低估要小得多。王小鲁(2010)提出收入水平越高则消费在其中的占比越低,由此可以估计每组的恩格尔系数或边际消费倾向,并倒推最高群体的收入水平。当然更细的研究会假定工资性收入不存在低估,且家庭食品支出也是准确的,并假定食品支出与瞒报收入存在线性关系,因此,使用恩格尔方程回归得到瞒报程度估计

值,继而推导出自雇佣家庭的瞒报收入(Pissarides 等,1989)。一个改进的方法是需求系统广义矩估计法,它根据户主职业分组并允许不同收入来源的瞒报程度有所不同,认为家庭总收入的瞒报程度为不同收入来源瞒报程度的加权平均,并假设工资性就业为主的家庭不存在瞒报,以此为参照组来估计其他组的瞒报情况(Lyssiotou 等,2004)。进一步的改进研究将工资就业人群的收入和消费参数关系类比到自雇人群的消费函数,进而得到后者的收入瞒报情况(Slemrod,2007)。

总的来看,使用行为反应函数来倒推高净值人群的收入瞒报,其优点是不需要从源头上修正住户调查数据,估计过程简单直接。但这种方法仍需要一些前提假定,如消费倾向与收入水平单调线性递增等。有研究表明,当收入达到一个高门槛点之后,其线性关系并不稳定甚至有逆转的可能(Hurst 等,2014)。另外,这种方法假定个体行为数据是准确的,忽视了其低估的可能,如存在对职业的隐瞒、低估工资性收入中的福利等。而且不同消费支出项实际上存在不同的低估率,这样恩格尔系数就不会出现平衡性变化,因此使用消费平衡性假定来推断高净值群体的真实收入仍存在较多改进空间。

二、遗产税乘数技术

(一) 微观匹配方法

一些国家有完善的资本税和财产存量税政策,或者具有长期的遗产税制度,因此,可以利用遗产税数据来推断高净值样本。遗产税记录提供了居民死亡时点的财产存量水平,而且可以追踪到较长时期的高净值人群(如美国的收入税于1913年建立,遗产税于1916年建立,而家庭消费金融调查SCF最早于1989年开始)。虽然遗产税制的门槛点有较大变化,但基本瞄准了财产最高1%的群体,因此,相对漏损的高净值人群规模是比较稳定的(Kopczuk 等,2004)。

遗产税方法的基本思路是,假定死亡个体是从总人口中抽取的代表性样本,遗产继承人是健在人口的一个子样本,故特定个体的死亡概率 m_i 是抽样概率。如果 m_i 是已知的,就可以通过抽样权重的倒数 $1/m_i$ 来重新加权继承者,从而得到健在人口的分布,由此 $1/m_i$ 也被称为遗产税乘数。这里关键是如何选择特定人群的死亡率,虽然不同年龄和性别人群的死亡率比较容易估计,但特定高净值人群的死亡率显然更低且难以准确估计,而且也要考虑死亡率本身存在的滞后性和自选择性问题。Kopczuk 等(2004)基于政府IRS数据发现,最高10%高净值人群的死亡率非常接近大学以上人群的死亡率水平,因此假定大学以上人群与其他人群的死亡率差异是稳定的,那么基于特定年份二者的差异来调整其他年份各个人群的死亡率,并由此倒推高净值人群的财产,从而就可以反映出各年财产分布的动态变化。

然而,这种方法的问题在于从历史趋势上高净值群体的资本性收入和继承财产比

重存在明显波动,从而遗产税乘数法的估计结果精准度还有待提高。另外,这种方法仅基于个体层面,而家庭人口结构及家庭内部不同成员的财产占比等都有可能影响家庭层面的财产分配,而且这里还存在税收免征和缓征的问题。因此,这种方法不能简单与家庭层面的住户调查数据进行比较(Piketty等,2003)。尤其是越来越多的遗产税规避计划会导致遗产税的实际乘数进一步下降,如把更多资金用于医疗护理、购买长期保险、把资产转移到更多继承人等,从而会低估高净值人群的财产份额(Atkinson等,2011)。而且遗产税方法只能修正最顶端的1%人群,与宏观加总的资金流量表数据比较后,发现遗产税乘数法仍然有较大程度的漏损(Piketty,2014)。

(二) 宏观加总方法

从宏观视角建立国民收入账户的方法,是先在宏观上计算漏损的总财富规模,再通过某些假定分摊到微观家庭上。一些研究尝试利用遗产税总量数据倒推财产分布,并取得很大的成功。如Atkinson等(2011)收集美国的遗产边际税率和实际缴税纪录,从而利用实际税收数据尤其是房产税和遗产税来进行反推,由此获得比较准确的富裕人群数据。Piketty(2014)从宏观视角建立一个国民收入账户以追溯遗漏的财产总额,并比较高净值群体调整前后的财产份额变化,其中估算了美国、英国和日本等20多个具有遗产税制国家的财产差距水平。Piketty等(2019)基于该方法和个人所得税数据还估计了中国的收入差距变化情况。

这个方法的优点是可以直接估计高净值人群的财产份额,尤其是在20世纪70年代住户调查普及之前,只能使用这种方法推算高净值人群问题,尤其是有遗产税制的国家通常每年都会公布这类数据,因此,该方法是连接宏观和微观修补技术的关键纽带(Piketty等,2003)。但这只能在宏观总体上反映漏损的总财产,而不能具体到微观住户,且看不到个体的行为反应,因此,对住户调查数据的改善作用有限(Jones,2015)。资金流量表的收入总量通常远高于住户调查数据,而且在不同年份波动较大,再加上中国当前没有遗产税制,12万元以上个人所得税的分组汇总数据也不再公布,因此当前通过国民收入账户与税收加总数据来倒推的方法在中国并不现实。

三、资本化收入倒推技术

(一) 资本化收入倒推法

高净值群体的重要收入来源是资本性收入,从资本性收入倒推财产成为一个可能的选择,再加上资本性行业的电子化程度高,信息开放程度大,因此成为校正最高1%人群的有效尝试(Kopczuk,2015)。由于资本性收入的获取相对容易,且不同人群具有相对稳定的资本回报率;在规则透明的税法下,这种方法能够捕捉到收入分布的绝大部分人群,由此倒推高净值人群的财产存量成为近年来的重要进展。

假定已知资本回报率 r , 且能观测到实际的资本性收入 k , 通过公式 $W=k/r$ 能获得财产存量水平 W 。由于大多数国家的资本性收入会被征收特定的收入税, 因此从所能观测到的收入税数据出发, 使用政策规定的资本性收入税率反推, 从而得到资本性收入及对应的财产存量 (Saez 等, 2016)。Kopczuk (2015) 研究发现, 资本性收入倒推法估计出的财产差距水平明显高于住户调查修正技术和行政部门数据匹配所得出的财产差距, 这得益于资本性收入倒推法捕捉到更多的高净值样本, 而且指标本身的低报明显更少。所以资本性收入法的优势在于能捕捉到整个分布的绝大多数人群, 而不是遗产税乘法法的很少一部分样本。随着近年来全球税收竞争导致各国遗产税边际税率下降、征收门槛上升、遗产税制覆盖率下降, 资本化方法的优势更加明显 (Kopczuk 等, 2004)。

（二）资本化收入方法的改进

近年来, 资本化方法在国外取得重要进展, 但也存在一些问题 (Saez 等, 2016)。首先, 不是所有财产都会产生资本性收入 (如养老基金、自有住房等), 或产生的资本性收入也不一定能够体现在收入税中。很多财产在持有时无法获得收益, 仅当个体死亡或资产被出售时才能实现资本性收入并进入收入税的范畴, 如农民生产性器械资产、人寿保险资产、养老金等。其次, 不同资产的回报率具有较大差异, 高净值人群的资本回报率明显高于其他人群; 而且富裕人群的高资本回报率还具有选择性 (如只有身体健康才能积极经营或打理资产, 并获得较高资本回报率), 且资本化方法很难纠正税收规避问题 (Piketty, 2014)。Saez 等 (2016) 使用美国私营基金会资本性收入数据发现, 修正上述问题后资本性收入法的表现是稳定的, 在估计财产差距的变化趋势上与遗产税方法的结果非常接近。进一步使用美国消费金融调查 (SCF) 中的收入和财产相互验证后, 也发现资本性收入倒推法具有较好的拟合效果, 尤其是对高净值人群更是如此。

四、多元参数估计技术

（一）分组估计技术

银行证券等金融部门或税务部门会公布高净值人群的收入区间统计, 如不同的收入分组、区间内人数、均值或中位数、不同分组的收入份额等, 在此基础上可以使用参数估计法计算分组的密度函数并拟合高净值人群收入。在把总人口分解为离散组别的差距贡献后, 程永宏 (2006) 建立了分组混合基尼系数的算法, 该算法不依赖于相邻两组分布不重叠的假定。Korinek 等 (2007) 使用不同人群的地理结构差异, 分组拟合各组别的低估函数, 由此用分组加权技术校正了高净值人群问题。Chotikapanich 等 (2007) 基于分组数据放宽组内收入不变的假定, 使用分组数据校正了 8 个东亚国家或地区的收入差距低估情况。Cowell (2011) 证明分组越多或组间的等宽区间越小, 则分组密度函数越倾向于光滑, 也就越接近总体分布。此后 Hajargasht 等 (2012) 基于更细的分组人口份额和

平均收入,推导出通用矩条件和最优权重矩阵,并用于收入分布的广义矩量法估计,从而以一种相对简单的形式表达附加高净值群体的目标函数。

总的来看,分组估计技术简单直接,数据来源广泛,可以相互验证。但部门公布的分组数据可能存在缺失,不同时期的分组区间及提供的有效信息存在较大差异。这种方法还面临从家庭向个体单位的分拆问题,如有的以个体为单位直接扣除,有的以家庭为单位综合计征,故需要基于人口和收入总量再次推测调整(Piketty等,2003)。另外,该方法假定每个分组内部的收入水平是一致的,这一般会显著低估总体差距水平,因此可以视为真实相对差距水平的下限,只有当分组越来越多甚至接近无穷时,估计出的相对差距水平才会更接近真实状况,但这在现实中很难做到(Cowell,2011)。

(二) 多元参数技术

一般认为,不同收入分位上的分布形状有明显的差异,对称钟形正态分布曲线可能使高净值人群的分布过度拖尾(Benhbib等,2018)。Pareto分布对高净值人群的捕捉能力更强,尤其是在特定门槛值之上的高净值人群,但在中低收入端其拟合效果较差;Gamma分布可以很好地模拟收入分配两端,但会夸大中间群体的扭曲程度,因此这些以2个参数为基础的拟合技术还存在改善空间(Arnold,2015)。

基于前面几种分布的优缺点,还可以综合更多参数以拟合现实中的收入分布,如Burkhauser等(2011)观察到高净值人群在某个点上的删失现象,使用了Dagum的3参数技术;Dastrup等(2007)使用4参数的GB2技术,也可以使用5个参数的Generalized Beta技术。Cowell等(2007)认为,这种叠加参数的方法只是技术上更加复杂,而没有从根本上解决住户调查数据内的中高净值样本低估问题。Cowell(2011)认为,使用5个多元参数的方法过于复杂,而解决问题的准确程度并没有显著上升,因此沿着技术难度方向扩展以拟合不同分位点的分布只是一个途径,而另一个更有意义的方向是完善现有的参数估计方法。

五、畸高低值修正技术

(一) 异常值修正技术

抽样分布可能出现两类偏误,一类是高净值住户应该抽但没有抽到,另一类是偶然捕捉到过高的异常高收入样本(Cowell,2011)。在实际的住户调查数据中,这两种情况都有可能遇到。对于第一类误差,上面给出了不同的修正对策。对于第二类误差,理论上不能认为它是对住户数据的不合理反映,因此要尽量保留原始极值可能代表的高净值住户信息。因此,极值修正技术认为,在对数据缩尾后可以适当提高其权重,从而在保留异常值背后信息的同时尽量保持原有数据的分布结构(Li等,2015)。这种方法虽然简单直接,但两种相反效应能够做到恰好抵消的情形并不多见,故该方法仍存在主观性;而且

在缩尾过程中,究竟多大属于异常值、缩尾到何种程度等,并没有客观的判断标准。

已有文献中也存在判别异常值的方法,如使用格拉布斯准则(Grubbs)给定一个置信概率和置信限,凡超过此限的误差就认为不属于随机误差,将其视为异常值剔除。但在剔除异常值后是否能够及如何补充异常值背后所缺失的高净值信息,仅依赖回归后的中位数值去插补是不可能的。后续研究提出了一种新的洛伦茨曲线方法来克服异常值被直接剔除的问题,如 Schluter 等(2002)基于洛伦茨曲线的厚尾行为,提出针对重尾分布要尽量稀释异常值在洛伦茨分布中的影响,这种检验方法既能识别与剔除异常值,还可以尽量保留极值背后所代表的高净值人群信息。

(二) 半参数自助法

一个畸高极端值可能使相对分布差距发生明显变化,纵然使用标准自助法,在厚尾分布下的估计和推断也会受到很大影响,因此是否简单剔除异常值都会带来问题(Burkhauser 等,2011)。对于纯粹在极端值情况下带来的分布困扰,理论和实践证明半参数下的非标准自助法在不同样本中可以实现准确推断。Davidson 等(2007)认为,从总体中随机抽取样本提供了自助法进行精确推断的理想机会,使用蒙特卡洛模拟的结果表明,若真的存在畸高极端值,即使在非常大的样本中使用标准自助法来推断也会带来明显误导。其主要原因是,许多相对差距指数对极端值的分布性质非常敏感,所以他们提供了两个非标准的自助法程序,其在标准自助法失败的情况下仍然有效。

基于上述文献,Cowell 等(2007)进一步提出在估计相对分布差距时,使用半参数法会比直接使用经验分布函数法(对用于生成样本的累积分布函数的估计)更加稳健。在有限样本的情况下,半参数方法下的自助法明显优于其他方法,甚至在有限样本下还能给出准确推断。同时在相对差距指数的选择方面,泰尔零阶指数或对数偏差均值指数对两端的样本比较敏感,而基尼系数是相对去极端化的方法,更不容易受异常值的影响,所以在高净值人群修补技术的应用过程中应作为首选。

六、新拓展幂律拟合技术

(一) 传统幂律技术

收入或财产一般服从典型的幂律分布,即门槛值以上的平均收入与门槛值的比例为常数(Arnold,2015)。考虑到越靠近顶端人群的财产信息越透明准确,传统幂律技术就比较依赖媒体收集的富人榜信息。估计步骤是,先假定高净值人群的财产服从幂律分布,利用富人榜估计幂律分布的关键参数,并根据确定的幂律函数从右向左拟合至主观选定的门槛点,由此可以修补右尾端分布。Cowell 等(2007)认为用幂律技术估计富裕人群分布具有方法学基础,在实践中被学者广泛接受。李实、罗楚亮(2011)提出利用富人榜信息对高净值样本进行修正,并基于拟合的分布函数推算高净值人群规模。王海

港、周开国(2006)对财产分布概率密度函数取对数,对收入对数做最小二乘法回归,认为得到的斜率绝对值就是规模参数,罗楚亮(2019)使用最小二乘法对富豪榜的最低值与住户数据中的最高值进行平滑链接,也直接估计了幂律分布的规模参数。

这种方法的优势在于,富豪榜信息动态更新,能获得年龄、性别、行业分布等人口学特征,这对于了解高净值人群有较大帮助。但这种方法也存在一些问题,如媒体公布的数据质量没有经过严格检验。Piketty(2014)认为,这个榜单中的继承财富占比被大大低估,纵然考虑到税收规避和家庭内部的财产转移问题(遗产税是个体层面,富豪榜通常是家庭或家族层面),这种低估仍然相当严重(Kopczuk,2015)。Li等(2020)也认为胡润和福布斯排行榜数据存在问题,如错误信息统计、个体与家族数据重叠、收入与财产信息混淆等,而且富豪榜没有扣除负债,因此净资产普遍存在高估。在估计方法上,传统幂律技术基本上是先验地假定满足幂律分布,但Cirillo(2013)已经证明大部分数据集实际并不满足幂律分布。Benhabib等(2018)认为,传统方法将频率分布的柱状图绘制在双对数轴上,使用OLS方法估计直线斜率并给出规模参数的做法可能产生系统性误差,如给出更高的标准误、 R^2 值无法指示有效的拟合优度、幂律分布中的门槛约束无法在OLS中体现等。Clauset等(2009)证明这种OLS方法相对主观,规模参数对分布尾部的波动非常敏感,因此会给右尾端的高净值人群推断带来误导。另外,Charpentier等(2019)认为,通过富豪榜数据拟合顶端人群的幂律分布,并从右向左延伸至主观选择的12万这个门槛点,不但门槛值选取过于主观,而且从高估的富豪榜数据延伸至很低的门槛点,使拟合的高净值样本从左右两边都大大增加,并导致总体差距显著上升。

(二) 新拓展幂律技术

近年来,越来越多的文献注意到传统幂律技术中存在的问题,并从不同角度进行拓展。(1)Arnold(2015)改进了传统两参数的幂律技术以适应不同的数据集,特别是改进了多峰核密度下的函数估计,由此修正传统单峰幂律技术所存在的应用偏差。(2)Charpentier等(2019)纠正了幂律技术使用中的几个偏误,如Pareto I模型(Hill估计量)对门槛值的选取非常敏感,估计出的相对差距水平也偏高,而修正后的广义Pareto II技术则不太敏感,估计结果与理想的幂律分布更加接近;Cirillo(2013)提供了幂律技术使用过程中究竟是用Pareto I还是Pareto II的简易检验方法,Jenkis(2017)在比较不同幂律技术所带来的偏差后,认为Pareto II的估计结果更加稳健。(3)Clauset等(2009)提供了检验住户数据是否满足幂律分布的方法,先人工生成幂律数据集并与实际数据进行比较,Cirillo(2013)改进了幂律分布的检验技术。(4)Benhabib等(2018)证明使用最大似然法估计幂律参数可以保证大样本下规模参数的无偏性。虽然小样本下最大似然估计值存在偏差,但在大多数情况下这种偏差可以忽略,至少比最小二乘法的统计误差要小得多。(5)Virkar等(2014)建议结合最大似然法和KS统计量先估算最优门槛值,再据此从相

反方向估计幂律分布的规模参数。

新型幂律拟合技术的优势在于不需要服从幂律分布的前提假定,而是严格检验各种备选分布的拟合优度,使用最大似然比这个统计显著性指标排除不同形式的分布函数,避免对幂律分布的武断假设。使用似然估计法而不是最小二乘法能得到有效的门槛值估计,使用 KS 统计量可以获得准确无偏的规模参数。另外,传统方法主观采信媒体公布的富豪榜最大值,并从右向左拟合到主观给定的 12 万门槛值,而新技术是先估计出幂律分布的最优门槛值,并从左向右拟合出观测数据中未收集到的高净值样本,不需要依赖媒体的数据,研究方法更为客观科学(Arnold, 2015)。但这种方法的问题是,它需要假定已收集到调查数据的尾端值真实可信,一旦存在低估,则由它来推导未收集到的高净值人群数据也会存在一定偏差。

(三) 洛伦茨累积技术

幂律技术隐含的前提是数据在超过某个门槛值之后会遵守幂律分布,现实中可能存在不吻合的情况,故先验地使用给定的分布函数来拟合高净值样本就不合适,所以不提前设定任何函数形式的洛伦茨累积技术就很必要。这种方法的原理是,先用各种不同的分布函数去生成人工数据集并计算对应的累积分布函数,再使用实际调查数据计算累积分布函数,最后检验实际数据与人工数据集的累积分布差异,并从中选定一个最为接近的分布函数(Charpentier 等, 2019)。也可以比较基尼系数的理论分布与洛伦兹技术计算的基尼系数的差异,从任一选择的点开始无限迭代,比较二者的差异,二者越接近说明这种理论分布对实际分布的拟合越好,收入高于该门槛的实际数据就越服从于所给出的分布形式(Cowell, 2011)。通过海量计算并选定分布函数后,就可以插补高净值人群到住户调查数据。

这种方法的最大优点是不提前使用任何一个明确的分布函数假定,而是从各种形式的分布函数中逐一比较,从中择优选取。但这种方法的问题是最开始仍要经验选定调查数据的可能分布形式,且在参数估计过程中要无限试错,这不仅增加计算量,而且试错结果不一定就是最终符合现实的分布(Benhabib 等, 2018)。住户调查数据缺损的高净值样本完全依靠模型来拟合,这会使右尾分布非常敏感,只要住户调查里的中高收入样本发生小的变动,就会导致拟合的右尾分布出现很大变化,因此,建议尽可能增加初始高净值样本数量,以减小数值模拟的区间(Jenkins, 2017)。

七、现代数据拼接技术

高净值样本的修补方法,总体上沿着数据和技术两个方向交替前进。近几年的拟合技术特别注重高质量微观数据与分布技术的融合,在获取高质量微观数据后,与调查数

据拼接就能得到总体分布。

（一）结合微观高净值样本

近年来,得益于社会科学数据的空前增长,各种爬虫技术、互联网技术及政府数据开放,使微观高净值数据显著增加,其中包括3个来源。一是微观的个人所得税数据。随着各国行政部门开放程度的提高,微观所得税数据开始出现,如英国、阿根廷和哥伦比亚等国的税务部门开放了个人所得税样本。二是学者收集的个体高收入数据库。媒体搜集的数据质量难以经受学术标准的审视,有学者建立规范的高净值人群数据库,从行业代表性、领域重叠性和数据独立性角度提出了更高要求。如Li等(2020)围绕高净值人群构建了多个子数据库,包括上市公司高管、私营企业主、金融证券人员和网络红人等。三是部门行政数据。很多国家建立了跨部门、多层次的信息联动平台,广泛引入如银行证券部门的金融资产、交通运输部门的汽车资产、住房建设局的住房信息、医院的医疗保健记录、人力资源部门的社会保险、工商局的企业经营资产等,这都有可能链接到高净值样本。

由于经审计的行政数据具有较高可信性,与住户调查数据进行精确匹配或个体特征倾向得分匹配后发现,美国人口普查数据的社会保险缴存率和缴存金额均存在低估,尤其是高净值人群强制保险外的补充保险和商业保险金额明显低报,因此Kopczuk等(2010)利用这个低估比率反向修正住户调查里高净值人群的实际收入。部门或行政数据的优点是数据准确性高,具有样本量大、测量误差小的特点,而且大多是动态更新数据,这是调查数据所不能比拟的,也比流失率持续上升的追踪调查数据更优。但行政数据只是政府工作记录的成果,数据异质性和指标多样性等还存在较大差距,不同部门的指标定义也不同,难以调整或协调,数据的可复制性较差,普遍缺乏人口学特征。

（二）现代数据拼接技术

由于微观个体层面的所得税数据只包括一定门槛值以上的样本,而门槛值以下的样本无法获得;相反,一般住户调查数据只能获得中低收入样本,高净值样本的代表性不足,因此,数据拼接技术的原理就是要找到一个最优的拼接点,拼接点以下使用住户调查数据,拼接点以上则基于微观高净值数据,由此获得完整的收入或财产分布。现代拼接技术的核心在于这个拼接点的选择,要让拼接后的样本尽可能符合总体分布。

Atkinson(2007)提供了“拼接点”的近似估计方法。在遗漏高净值样本无限少的假定下,如果他们占总体收入的比例为 S ,住户调查数据计算的基尼系数为 G^* ,则总体收入的基尼系数可以被近似表示为 $G=S+(1-S)G^*$ 。在实践应用中,Atkinson等(2011)假定住户调查数据遗漏了0.1%的高净值人群,因此运用上述公式就可以计算拼接后的总体基尼系数。但这种方法的前提是遗漏样本很小,实践中方便起见假定是0.1%,因此该公式仅是对实际情况的近似估计。近年来随着遗漏高净值样本的增加,其结果变得愈发不准确。

Alvaredo(2011)对上述公式进行理论证明,并拓宽了其极端限定,在不再是无限少样本的情况下,发现总体分布的基尼系数公式为 $G=G^{**}PS+G^{*}(1-P)(1-S)+S-P$,其中 G^{*} 是住户调查的基尼系数, G^{**} 是高净值样本的基尼系数, P 和 S 分别是高净值样本占总人口和总收入的比例。

Diaz-Bazan(2015)放宽了对现实情况的假设,认为住户调查不仅会缺失最高值样本,同时次高样本的收入也可能存在低估。例如,在最高 1%群体缺失、次高 1%~5%群体低估、仅其他 95%群体是准确的情况下,之前对总体样本的二段划分是不够的,不仅最高收入人群是 5%、1%还是 0.1%的选取主观,而且住户调查中的次高群体低估明显,因此,上述公式会出现双重偏差,故用两个微观数据来直接拼接会更合适。考虑到税法规定的自我申报门槛值相对较低,门槛值以下的住户调查数据会足够准确,同时该门槛点以上又被税收数据所完全覆盖,所以该文建议使用税法申报门槛点作为拼接点。Jenkins(2017)改进了调查样本与高净值样本的拼接技术,认为选择税法规定的最低申报点为拼接点仍然比较武断,应该结合新型幂律技术拟合高净值样本的幂律函数,并用半参数方法找到最优的幂律门槛值作为两套数据的拼接点。与前面不同的是,该方法更加慎重对待税收数据中的右尾低估问题,并进一步用半参数方法修正税收数据,而不是在门槛值以上用税收数据简单替代调查数据。

现代拼接技术的最大优势是充分结合微观高净值数据和新拓展幂律技术,不需要前置假定任何形式的分布函数,估计过程中也不存在对分布参数的主观假定。只是这个技术的前提仍然是高质量微观数据,很多研究者并不容易获得,同时税收数据本身也存在一定程度的瞒报问题。该方法在收入指标的协调上也面临较大挑战,如调查数据是基于权责发生制定义,而税法是收付实现制来界定;调查数据一般包括政府转移性收入,但税法中没有;税收包括偶然和意外所得,而在调查数据中没有纳入;税收数据一般面向成年就业个体,而不是调查数据的所有家庭成员,二者计算的基尼系数难以直接比较。当然,这些问题基本属于技术上可以克服的,因此该方法成为当前高净值人群的最前沿研究领域。

八、技术比较与简要总结

简单用住户调查数据来研究各国收入分配问题,不但会导致相对差距水平的严重低估,还可能出现方向性误判,所以在使用住户调查数据前需要校正高净值人群缺损问题。在研究实践中,中国大多数基于住户调查数据的实证研究都没有修补高净值人群缺损问题,微观数据中相应人群的代表性不足,可能带来实证研究结论的偏差。

(一) 不同方法比较

从消费和避税行为等间接推导的方法,由于反推过程中依赖很多假设,如假定自雇

性收入存在瞒报而其他收入没有瞒报、消费数据不存在低估等,这显然不太符合现实。在中国,因为信息孤岛原因使不同数据、不同指标间难以实现精确匹配,所以行为反应倒推技术和行政数据匹配方法只能对少数的收入低报问题有帮助,而对高净值样本缺失问题的作用并不大。

由于高净值人群的自营就业收入很多没有进入政府税收范围,行为推导和行政部门匹配等技术难以奏效,但遗产税乘数法能有效弥补该部分高净值人群。然而,遗产税乘数法所能获得的微观样本过少,而且考虑到人均预期寿命提高和平均死亡率下降,用遗产税乘数推导高净值人群的方法也面临较大挑战。中国目前还没有遗产税制,因此遗产税方法也存在较大局限性。对比来看,资本性收入倒推法简单且容易复制,大样本且不依赖样本财产分布区间的假设,但这种方法要注意检验不同人群资本回报率的差异,且该方法不能剔除高净值人群的负债,同时不产生收益的资产也难以计算在内。由于资本性收入的比重越来越高,该技术所估计的高净值人群比重明显高于其他方法。从趋势上看,用遗产税乘数法估计近年来美国财产差距的结果相对稳定,而基于资本化方法发现美国财产差距仍在明显增加,这可能是由于不同方法的分析单位和假定条件不同,其结果不能简单比较。

除了前面使用非参数方法来间接推导高收入样本之外,使用参数估计技术来补充高净值样本是近年来的重要进展。在所有参数估计方法中,分组估计技术的数据来源最为广泛,可以提供不同时期可比较的修补结果,但它假设每组内部的收入一致,由此会低估总体差距水平。幂律技术可以直接修补高净值人群的缺失问题,但要假定数据符合幂律分布,而现实中有很多情况并不满足。叠加更多、更复杂的多元参数技术虽然可以捕捉收入或财产分布中的更多特征,但依然无法解决住户调查数据内中高净值样本的缺失和低估问题。另外,对于极少数的畸高极端值,可以先缩尾并调整权重,再使用非标准的半参数自助法等技术来降低异常值对于总体分布的影响。得益于微观高净值数据的不断丰富,使用现代数据拼接技术来校正高净值样本结合了上述几种修补技术的优势,因此成为近年来的研究前沿。

(二) 总结与启示

在修补高净值人群方面,高质量数据是前提,而更好的估计技术肯定会改善估计质量。综合比较不同的技术方法后,本文发现,行为反应推导法、遗产税乘数法、资本性收入法、多元参数估计技术、幂律拟合技术和现代数据拼接技术等各有利弊,实际估计结果也存在一定差异。这与数据质量和前置假定等有较大关系,使用时要特别注意前提假定和技术要点(Kopczuk, 2015)。

本文建议,在中国住户调查数据的应用研究中,首先要尽量多地保留样本个数,充分利用异常的极端值信息;其次要对极端值进行非标准化的半参数自助法处理;再次要注

意使用基尼系数这种不易受异常值影响的相对差距指数。在此基础上,根据微观数据的实际情况选用修补技术,使用前认真分析各种技术的前置假定和技术本身的优缺点,综合考虑高收入数据的可得性、研究对象、分析单位、指标协调性和研究精度的需要。

由于中国目前没有遗产税制,个人所得税微观数据没有公开,行政部门的高净值人群数据难以获取,再加上中国资本市场并不发达,导致用资本性收入倒推的方法也面临诸多困难,所以在修补中国高净值人群时,可以使用消费等指标来反向推导高净值人群,或使用比较复杂的参数模型或幂律拟合技术修补高净值人群。当然,近年来现代数据拼接技术结合了新拓展幂律技术以拟合微观高净值数据,这种方法不事先假定任何形式的分布函数,也不存在对估计参数的主观假设,因此在中国具有较大的应用前景。

为更好地使用修补技术以纠正中国高净值人群缺失问题,本文建议在住户调查中重视住户调查数据的代表性。一是严格执行《中华人民共和国统计法》关于个体信息保密的规定,加强住户调查数据的真实性核查环节,严防从业人员泄露个体信息,严惩个体隐私数据的买卖行为,在源头上提升高净值人群的配合度和遵从率。二是建议税务部门在早期介入住户调查的抽样设计,结合宏观总体的高净值人群比例数据,优化高净值人群的抽样方案。三是在抽样设计时适当扩大高净值人群抽样比例、严格执行高净值个体的样本轮换规则、减少各种形式的非响应率(如住户联系不上、不配合或拒绝接受访谈等)。四是在现有数据基础上,统计、金融或其他行政部门要增加收入或财产指标的分组区间,提供更多、更细的分组统计描述,如中位数和五等分组数等,从而拓宽高净值人群修补技术的选择空间。五是探索数字化技术获取高净值人群数据,利用大数据增强住户调查的代表性,使现代数据拼接技术在中国具有更大可行性。六是在脱敏后要加大税务等行政部门微观数据的开放力度,这既可以增加个体匹配的多元化信息来源,校正中高净值群体的信息准确度,也能弥补住户调查本身所难以覆盖的高净值样本,从而改善住户调查数据对高净值人群的代表性。

参考文献:

1. 白重恩等(2015):《中国隐性收入规模估计——基于扩展消费支出模型及数据的解读》,《经济研究》,第6期。
2. 程永宏(2006):《二元经济中城乡混合基尼系数的计算与分解》,《经济研究》,第1期。
3. 李实、罗楚亮(2011):《中国收入差距究竟有多大?——对修正样本结构偏差的尝试》,《经济研究》,第4期。
4. 罗楚亮(2019):《高收入人群缺失与收入差距低估》,《经济学动态》,第1期。
5. 王海港、周开国(2006):《中国城乡居民收入分配的不平等程度被低估了吗?——基于帕雷托分布的检验》,《统计研究》,第4期。

6. 王小鲁(2010):《灰色收入与国民收入分配》,《比较》,第3期。
7. Ales L., Sleet C.(2016), Taxing Top CEO Incomes. *The American Economic Review*. 106(11):3331-3366.
8. Alvaredo F.(2011), A Note on the Relationship between Top Income Shares and the Gini Coefficient. *Economics Letters*. 110(3):274-277.
9. Arnold B.C.(2015), *Pareto Distributions*. New York: Taylor & Francis Group.
10. Atkinson A.B., Piketty T., Saez E.(2011), Top Incomes in the Long Run of History. *Journal of Economic Literature*. 49(1):3-71.
11. Atkinson, A.B.(2007), Measuring Top Incomes: Methodological Issues. In: Atkinson A., Piketty T.(Eds.), *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-speaking Countries*. Oxford: Oxford University Press.
12. Benhabib J., Bisin A.(2018), Skewed Wealth Distributions: Theory and Empirics. *Journal of Economic Literature*. 56(4):1261-1291.
13. Burkhauser R.V., Feng S., Jenkins S.P., Larrimore J.(2011), Estimating Trends in US Income Inequality Using the Current Population Survey: The Importance of Controlling for Censoring. *The Journal of Economic Inequality*. 9(3):393-415.
14. Burkhauser R.V., Feng S., Jenkins S.P., Larrimore J.(2012), Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data. *The Review of Economics and Statistics*. 94(2):371-388.
15. Charpentier A., Flachaire E.(2019), Pareto Models for Top Incomes, Université Paris1 Panthéon-Sorbonne (Post-Print and Working Papers). No. Hal-02145024.
16. Chotikapanich D., Griffiths W.E., Rao D.S.P.(2007), Estimating and Combining National Income Distributions Using Limited Data. *Journal of Business & Economic Statistics*. 25(1):97-109.
17. Cirillo P.(2013), Are Your Data Really Pareto Distributed?. *Physica A: Statistical Mechanics and Its Applications*. 392(23):5947-5962.
18. Clauset A., Shalizi C.R., Newman M.E.J.(2009), Power-law Distributions in Empirical Data. *SIAM Review*. 51(4):661-703.
19. Cowell F.A.(2011), *Measuring Inequality*. New York: Oxford University Press.
20. Cowell F.A., Flachaire E.(2007), Income Distribution and Inequality Measurement: The Problem of Extreme Values. *Journal of Econometrics*. 141(2):1044-1072.
21. Dastrup S.R., Hartshorn R., McDonald J.B.(2007), The Impact of Taxes and Transfer Payments on the Distribution of Income: A Parametric Comparison. *The Journal of Economic Inequality*. 5:353-369.
22. Davidson, R., Flachaire, E.(2007), Asymptotic and Bootstrap Inference for Inequality and Poverty Measures. *Journal of Econometrics*. 141(1):141-166.
23. Diaz-Bazan T.V.(2015), Measuring Inequality from Top to Bottom, World Bank Policy Research Working Paper No.7237.
24. Hajargasht G., Griffiths W.E., Brice J., Rao D.S.P., Chotikapanich D.(2012), Inference for Income Distributions Using Grouped Data. *Journal of Business & Economic Statistics*. 30(4):563-575.
25. Hurst E., Li G., Pugsley B.(2014), Are Household Surveys Like Tax Forms? Evidence From Income Underre-

- porting of the Self-employed. *The Review of Economics and Statistics*. 96(1):19–33.
26. Jenkins S.P.(2017), Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica*. 84(334):261–289.
27. Jones C.I.(2015). Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality. *The Journal of Economic Perspectives*. 29(1):29–46.
28. Kopczuk W.(2015), What Do We Know about the Evolution of Top Wealth Shares in the United States. *The Journal of Economic Perspectives*. 29(1):47–66.
29. Kopczuk W., Saez E.(2004), Top Wealth Shares in the United States, 1916–2000: Evidence from Estate Tax Returns. *National Tax Journal*. 57(2):445–487.
30. Kopczuk W., Saez E., Song J.(2010), Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937. *The Quarterly Journal of Economics*. 125(1):91–128.
31. Korinek A., Mistiaen J.A., Ravallion M.(2006), Survey Nonresponse and the Distribution of Income. *The Journal of Economic Inequality*. 4:33–55.
32. Korinek A., Mistiaen J.A., Ravallion M.(2007), An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys. *Journal of Econometrics*. 136(1):213–235.
33. Li Q., Li S., Wan H.(2020), Top Incomes in China: Data Collection and the Impact on Income Inequality. *China Economic Review*. 62(C):36–52.
34. Li S., Wan H.(2015), Evolution of Wealth Inequality in China. *China Economic Journal*. 8(3):264–287.
35. Lyssiotou P., Pashardes P., Stengos T.(2004), Estimates of the Black Economy Based on Consumer Demand Approaches. *The Economic Journal*. 114(497):622–640.
36. Meyer B.D., Mok W.K.C., Sullivan J.X.(2015), Household Surveys in Crisis. *The Journal of Economic Perspectives*. 29(4):199–226.
37. Piketty T.(2014), *Capital in the Twenty-First Century*. Cambridge MA: Harvard University Press.
38. Piketty T., Saez E.(2003), Income Inequality in the United States, 1913–1998. *The Quarterly Journal of Economics*. 118(1):1–39.
39. Piketty T., Yang L., Zucman G.(2019), Capital Accumulation, Private Property, and Rising Inequality in China, 1978–2015. *The American Economic Review*. 109(7):2469–2496.
40. Pissarides C.A., Weber G.(1989), An Expenditure-based Estimate of Britain's Black Economy. *Journal of Public Economics*. 39(1):17–32.
41. Saez E., Zucman G.(2016), Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data. *The Quarterly Journal of Economics*. 131(2):519–578.
42. Schluter C., Trede M.(2002), Tails of Lorenz Curves. *Journal of Econometrics*. 109(1):151–166.
43. Slemrod J.(2007), Cheating Ourselves: The Economics of Tax Evasion. *The Journal of Economic Perspectives*. 21(1):25–48.
44. Virkar Y., Clauset A.(2014), Power-law Distributions in Binned Empirical Data. *The Annals of Applied Statistics*. 8(1):89–119.

(责任编辑:朱 犁)