户内人口匹配数据的误用与改进*

——兼与《高等教育扩张与教育机会平等》一文商榷

杨　舸　王广州

提要:户内人口匹配数据被广泛地应用于社会学、人口学及相关领域的研究中,但户内人口匹配数据的选择性偏差常常被研究者忽略。针对这类问题本文利用2000年第五次人口普查和2005年1%人口抽样调查原始抽样数据对户内父子、母子和夫妻关系进行匹配,发现三种匹配数据均存在不同程度的选择性偏差,体现在年龄、性别、流动状况、城乡分布、教育、地区分布等方面。在此基础上,本文对《高等教育扩张与教育机会不平等》一文的匹配数据、分析模型和研究结论进行再检验。发现匹配数据的选择性偏差对分析模型和研究结论的影响是确定的。户内人口匹配数据的偏差不仅影响统计模型因素判断程度的错误,甚至完全改变影响因素的作用方向。为了减小匹配数据偏差的影响,本文提出户内人口匹配数据偏差的调整方法和思路,认为加权和再抽样方法能够在一定程度上弥补"选择性偏差",相比来说,加权模型的调整效果更加理想。

关键词:户内匹配　人口普查数据　加权　再抽样

一、研究问题的提出

以户为单位是人口普查或1%人口抽样调查的主要登记形式。人口普查和1%人口抽样调查不仅需要填报户主信息,而且需要一并申报户内成员及相互关系信息。为了保证户信息和户内人口信息的完整,在人口普查或1%人口抽样调查的原始抽样数据中,均把户作为最小的抽样和可识别的单位,使得原始数据依然保留户和个人两层结构,为研究者的数据挖掘提供了便利。在使用户和户内信息进行数据挖掘时,研究者可以利用"成员与户主的关系"来构建各种户内成员关系,如:夫妻关系、父子关系、母子关系,这些关系被广泛地应用于婚姻、家庭、代际、社会网络、生育、社会流动等问题的研究中。利用以户为单位

* 本项研究为中国社会科学院项目(VZDA2010－15)阶段性研究成果之一。

的原始数据，对户内成员匹配研究的案例很多（李玉柱、姜玉，2009；吴晓刚，2009；李志宏，2004；郭志刚、李睿，2008）。在使用匹配数据进行研究时，首先必须回答户主和户内成员的关系是否确切对应、户内成员登记是否完整，以及户内成员关系是否由于匹配的原因发生扭曲等问题。这些问题不仅是判断匹配样本是否能代表目标总体或是否有选择性的基础，而且也是判断数据是否适合进行相关研究的开始。在实际运用匹配数据的过程中，户内关系的构建都建立在一个假设的基础上，那就是这个关系所涉及的成员都被调查登记在一个“户”中，不能登记在一个“户”内的关系会被排除在研究数据之外。然而，在这种条件下判断出来的“关系”往往容易产生选择性偏差。

二、“户”概念的界定

“户”概念的界定和实际应用与户内成员关系匹配密切相关。但在我国现代统计登记制度中，户的操作定义在发生着变化，特别是户籍管理制度和人口普查（或抽样调查）规则交织在一起，“户口”、“住户”和“家庭”等概念的交叉和重复使得“户”概念的界定和实际操作更加难以把握。

如历次人口普查与人口抽样调查的具体登记情况就不完全一致。1982 年第三次人口普查登记户内成员包括：人住在本户，户口在本地；人住在本户，人来本地一年以上，户口不在本地；人住在本户，人来本地不到一年，但离开户口登记地一年以上；人不住本户，户口在本户，离开本地不足一年。这就是我们常说的“常住人口”口径。1990 年第四次人口普查和 2000 年第五次人口普查均采用了“常住人口”口径，但“五普”的“常住”时间界定由“一年以上”改成了“半年以上”。2005 年全国 1% 人口抽样调查每户的调查对象是：居住在本户的人口和户口登记在本户但人不在本户居住的人，即“现有人口”和“户籍人口”口径。[①] 调查过程中“户内成员关系”信息的采集多数照搬“户口本”，当

① “现有人口”口径调查了所有本户户口本上人口的信息，本可以避免外出流动人口无法与其父亲匹配上的偏差，但由于一些调查和抽样上的误差，使得流出人口的个人信息无法使用。这就是为什么 2005 年 1% 人口抽样调查的原始抽样数据中有很多没有“户主”家庭户的原因。

“人不在户在”这部分人口的信息因严重偏差而被删除后，就出现了许多没有“户主”的家庭户，这就影响到了户内成员关系的匹配。可见，调查登记原则是影响户内关系测量的主要因素，同时，调查实施方法也是影响户内关系测量的重要因素。

三、选择性的产生

家庭生命周期决定了被登记在一起的家庭成员的结构和特点。有家庭关系的人口在以户为单位进行登记时，可能登记在一起也可能无法登记在一起。因此，调查数据的偏差既有主观偏差，也存在客观偏差。然而，从社会学视角研究个体发展的影响因素时，往往需要匹配子女和父母的信息。由于子代的发展需要在其成年之后才能彻底表现出来，这类研究中就需要选取已经成年的子代数据。但成年子女和父辈生活在同一“户”中的可能性却大大降低，许多成年子女由于结婚、外出就业或就学等原因离开父辈居住。此时不论是“常住口径”还是“现有口径”的登记原则，都会将这些离开父辈居住的子女排除在研究数据之外。同样，在研究妇女生育史时，也需要利用户内人口匹配方法，对母亲—子女进行匹配。幼年子女多数与母亲生活在一起，但较年长子女往往已经离家，许多母亲的信息也不能与其全部存活子女匹配上，容易错误估计母亲的生育子女数和初育年龄。与此同时，母亲去世或外出、父母离婚等情况的母亲—子女数据也被排除在外。

四、匹配与未匹配意味着什么？

对原始数据通过户内关系匹配，为研究婚姻、生育、亲子关系、教育等提供了极大便利。完成户内人口匹配后，能够匹配上的成员和未匹配成功的成员往往是两个不同的群体。下文将对我国 2000 年普查的抽样数据和 2005 年 1% 人口抽样调查数据进行父子、母子和夫妻匹配，对比已匹配与未匹配的人口在年龄、性别、城乡和地区分布、教育、婚姻等特征上存在的差别，找出匹配过程可能带来的选择性。

(一)父子匹配

在2000年第五次人口普查的0.95‰的原始数据中,20-30岁人口为209592人,其中能与父亲匹配上的只有70982人,占33.9%;在2005年1%人口抽样调查258万抽样原始数据中,20-30岁人口为372301人,其中能与父亲匹配上的只有115483人,占31.0%。比较发现,无法进行父子匹配的人口比例有增长趋势,这不仅与我国现代化进程中家庭结构的变化和人口流动更加频繁有关,也与调查登记的数据口径变化(由"常住口径"转变为"现有口径")有关。

详细分解匹配过程发现,对于"与户主关系"为户主、配偶、子女、父母、媳婿、兄弟姐妹的个体,我们可以准确判断其父亲是否在该户中,而对于"与户主关系"为孙子女或"其他"的成员,则无法判断其父亲是否在该户中。在2000年和2005年的数据中,绝大部分匹配上的"父子"是准确的,但仍然有0.79%和1.70%的"父子"并不十分准确。这也是利用户内成员匹配数据进行研究分析的风险之一,若将"孙子女"或"其他"的成员都排除掉,又有可能进一步造成新的选择性问题。

从匹配数据"与户主关系"构成可知,父子不能匹配最重要的原因是子女的"自立门户"。在2000年的未匹配人口中,33.49%自己成为了"户主",33.33%是别人的"配偶",14.93%是别人家的"媳婿",还有11.98%的"其他"人多数离家后生活在集体户(学校或工作单位)中。2005年的匹配数据也呈现类似的情况。

对比2000年和2005年匹配上人口与未匹配上人口的特征差别发现:

第一,从年龄和性别结构来看。匹配上人口的年龄结构更年轻,随着年龄的增长,自立门户的可能性上升,父子能匹配的概率则下降。例如2005年20岁能匹配的人口比例为51.63%,25岁下降到32.32%,而到30岁能匹配上的比例仅为18.73%;匹配上人口中男性的比例明显高于未匹配上人口,而女性因为出嫁的原因,能和父亲登记在同一户的比例显著下降。

第二,从户籍特征来看。匹配上人口中农业户口的比例略高于未匹配上人口,差异并不明显;未匹配上人口中流动人口的比例明显高于匹配上人口中流动人口的比例,外出流动是父子不能匹配的重要原因之一。

第三,从社会经济特征来看。匹配上人口的受教育程度向中间集

中,2000 年的数据向初中和高中毕业者集中,2005 年的数据则向高中和大专毕业者集中;就婚姻来看,未匹配上人口中的已婚比例明显高于匹配上人口,女性出嫁和男性结婚后的自立门户是父子不能匹配的重要原因;就职业来看,匹配上人口中的农、林、牧、渔、水利业生产人员的比例明显高于未匹配上人口。

第四,从分布地区来看。未匹配上人口中,北京、上海、福建、广东、新疆等地区所占比重明显高于匹配上人口中这些地区所占比重,这些地区都是流动人口较多的地区。

因此,20 – 30 岁子女的父子匹配数据在总体中占的比例远远小于未匹配上人口,且匹配上人口的分布与总体分布不一致,匹配过程的选择性使得匹配数据的分布在年龄、性别、教育、婚姻、职业、地区等方面均与总体存在一定偏差。

(二) 母子匹配

为了研究育龄妇女的生育史(最常见的是计算妇女的初育年龄或是初婚与初育的间隔时间),需要对数据进行母子匹配。从对 2005 年 1% 人口抽样调查原始数据母亲与其子女进行匹配结果来看,原始数据中共有 51 万育龄妇女有存活子女,只有 26.3 万育龄妇女与其全部存活子女登记在一个家庭户中,占 51.64%,没和任何子女登记在一起的为 15.7 万人,占 30.84%。这意味着只有一半育龄妇女可以准确计算其初育年龄。与父子匹配相比较,母子匹配的比例明显要高。对比 2005 年 1% 人口抽样调查中,母亲和全部存活子女生活在一起的匹配人口与母亲和部分存活子女(或没有和任何一个子女)生活在一起的未匹配人口的特征差别发现:

第一,从年龄结构来看。匹配上母亲的年龄结构更年轻,这说明随着年龄的增长,子女离家的可能性上升,母子能匹配的概率则下降,这与父子匹配的特征完全相同。

第二,从户籍特征来看。与 20 – 30 岁父子匹配无明显选择倾向有所不同,匹配上母亲中非农业户口的比例高于未匹配上母亲的非农比例,非农业人口子女离家的可能性更低;未匹配上母亲中流动人口的比例明显高于匹配上母亲中流动人口的比例,外出流动是母子不能匹配的重要原因之一。

第三,从社会经济特征来看。匹配上母亲的受教育程度高于未匹

配上母亲,未匹配上母亲小学毕业比例高于匹配上母亲,初中和高中毕业比例则低于匹配上母亲;就职业来看,未匹配上母亲的农、林、牧、渔、水利业生产人员的比例高于匹配上母亲,专业技术人员比例则低于匹配上母亲。

由此可见,母子匹配数据在总体中的分布也是有选择性的,匹配过程的选择性使得匹配数据的分布在年龄、流动状况、城乡分布、教育、职业等方面均与总体存在一定偏差。在分析母亲初育年龄的影响因素时,也必须考虑到这些偏差可能产生的影响。

(三)夫妻匹配

2005 年 1% 人口抽样调查数据中的 764195 个在婚妇女中,65.2% 和丈夫生活在一起,另外 34.8% 的在婚妇女并没有和丈夫生活在一起,显然,夫妻匹配的比例明显高于母子匹配,但这并不意味着匹配数据没有选择偏差。

从妻子"与户主关系"构成可以看出夫妻不生活在一起的原因。大多数匹配上妻子是户主的配偶;12.43% 的未匹配上妻子自己是户主,这时丈夫可能外出了,即留守妻子;5.47% 的未匹配上妻子住在娘家,是户主的子女;5.99% 的未匹配上妻子因为住在子女家而未能和丈夫生活在一起;4.31% 的未匹配上妻子是"其他"人,此时她可能在就学或因工作住在集体户中。

比较妻子的特征发现:

第一,从年龄结构来看。未匹配上妻子的年龄结构更年轻,这说明随着年龄的增长,夫妻匹配的概率上升。

第二,从户籍特征来看。匹配上妻子中非农业户口的比例高于未匹配上妻子的非农比例,但差异不大;未匹配上妻子中流动人口的比例明显高于匹配上妻子中的流动人口比例,外出流动是夫妻不能生活在一起的重要原因。

第三,从受教育结构来看。匹配上妻子的受教育程度低于未匹配上妻子,匹配上妻子的小学毕业比例高于未匹配上妻子,初中和高中毕业者比例则低于未匹配上妻子,受教育程度更高的女性,由于工作等原因与丈夫分居的可能性更大。

第四,就分布地区来看,未匹配上妻子在广东、福建、江西、浙江、广西、陕西等地区的比例比匹配上妻子更高,总体构成差异不大。

相对来说，夫妻关系匹配数据是上述这些关系里最“安全”的。但是，匹配数据在总体中的分布还是有一点偏差，体现在年龄、流动状况、城乡分布、教育、地区分布等方面。在研究夫妻教育匹配时，需考虑到挑选带来的教育偏差。

综上所述，各类匹配数据在一些基本特征的选择性倾向上存在一致性。具体来说，在年龄选择性方面，能父子/女匹配上的子女和能母子/女匹配上的母亲均具有年轻化的选择倾向；能和父母匹配上的子女具有男性选择倾向；能匹配上户内成员均具有不流动的选择倾向；能和父母匹配上的子女有未婚选择倾向。另外，匹配成功的关系在受教育、职业、分布上都具有选择倾向(参见表1)。

表1　　三类匹配数据的选择性

特征	父子匹配	母子匹配	夫妻匹配
	对子女来说	对母亲来说	对妻子来说
年龄	有年轻选择倾向	有年轻选择倾向	有年老选择倾向
性别	有男性选择倾向		
户口性质	无明显选择倾向	有非农业户口选择倾向	无明显选择倾向
流动状况	有不流动选择倾向	有不流动选择倾向	有不流动选择倾向
受教育程度	有受教育程度中端的选择倾向	有受教育程度更高的选择倾向	有受教育程度更低的选择倾向
职业	有农、林、牧、渔、水利业的选择倾向	有专业技术人员的选择倾向	无明显职业选择倾向
婚姻	有未婚选择倾向		
分布	有北京、上海、福建、广东等流动人口较多地区的反向选择倾向	无明显地区选择倾向	无明显地区选择倾向

五、对《高等教育扩张与教育机会不平等》一文的再检验

李春玲在《高等教育扩张与教育机会不平等——高校扩招的平等化效应考查》(李春玲，2010。以下简称李文)中，利用2005年1%人口抽样调查的父代—子代匹配数据构建了两组logit regression模型，分析了高等教育机会在哪些方面存在不平等性，并比较“高校扩招前”与

“高校扩招后”的不平等性是否发生了变化和通过作用于哪些因素而发生变化。结果认为:大学扩招(高等教育扩张)并未使高等教育机会的阶层不平等、城乡不平等和民族不平等下降;较高等级的高等教育领域的机会不平等大于较低等级的高等教育领域,尤其表现在阶层不平等和城乡不平等方面。

本文认为,李文所使用的父子匹配数据是匹配率较低、选择性风险较高的户内人口匹配数据,而严重的选择性偏差可能会影响研究结论的可靠性。

(一)重构匹配过程

为了重构研究过程,本文采用与李文完全一致的口径,对2005年1%人口抽样调查数据的一个次级数据集进行了父代—子代匹配,主要围绕“与户主关系”这一变量,本文成功匹配了子女为1975–1985年出生的父子(女)信息118121条(由于一些父亲“上周未从事任何工作”,因此缺少其职业信息,最终进入模型的是95075条),匹配的对应关系如表2。

表2 父代—子代匹配的对应关系和匹配结果

子女	父亲	个案数	百分比
户主	父母	3889	3.29
配偶	岳父母/公婆	252	.21
子女	户主	104463	88.44
子女	配偶	6691	5.66
孙子女①	子女	1720	1.46
孙子女	媳婿	262	.22
兄弟姐妹	父母	844	.71
总计		118121	100.00

而从表3可以看出,本文匹配数据与李文各变量的描述结果有7个方面的不同:首先,李文最终采用的数据为19615条,本文匹配数据

① 当子代与户主关系是孙子女的时候,有时能在同一户中找到两个及以上与户主关系为子女(且性别为男性)的人,这时无法识别哪一个才是该“孙子女”的父亲,因此,将其视作不能匹配成功。

表 3　　李文和本文的分析变量描述性统计　　单位:(%)

	李文			本文		
	总体	1975 – 1979	1980 – 1985	总体	1975 – 1979	1980 – 1985
性别(男)	51. 5	50. 2	51. 8	65. 04	78. 16	60. 03
民族(少数民族)	10. 1	10. 4	10	12. 56	12. 51	12. 58
本人受教育程度						
未上过学	1. 6	2. 5	1. 4	2. 12	2. 58	1. 94
小学	12. 6	17. 2	11. 6	13. 41	16. 75	12. 13
初中	47. 7	46	48. 1	54. 65	57. 09	53. 71
高中	18. 9	17. 6	19. 2	17. 70	14. 74	18. 83
大专	11. 4	10. 1	11. 8	7. 70	5. 81	8. 42
本科	7. 3	5. 9	7. 7	4. 21	2. 74	4. 76
研究生及以上	. 3	. 8	. 3	. 23	. 29	. 20
父亲职业						
管理人员	3. 3	3. 1	3. 3	2. 68	2. 65	2. 69
专业技术人员	6. 7	7	6. 7	6. 06	6. 20	6. 01
办事人员	7	7. 8	6. 8	4. 79	4. 75	4. 81
商业服务业人员	10. 9	9. 8	11. 1	9. 50	7. 90	10. 12
农民	51. 1	55. 3	50. 2	61. 78	67. 12	59. 73
工人	21	17. 1	21. 9	15. 19	11. 38	16. 65
父亲户口(非农业户口)	32. 2	31. 7	32. 4	21. 99	19. 52	22. 93
父亲月收入(≥2000 元)	4	4. 3	3. 9	5. 03	4. 20	5. 35
父亲受教育年限(年)						
均值	7. 29	6. 93	7. 38	7. 92	7. 25	8. 18
标准差	2. 80	3. 00	2. 74	3. 24	3. 33	3. 17
N	19615			95075		

可以进入模型的有 95075 条,样本量相差 7 万条;其次,李文的样本中,子女的男女比例基本一致,本文匹配成功的子女中,男性比例更高,特别是 1975 – 1980 年的匹配子女中 78. 16% 是男性;第三,李文子女的少数民族比例为 10. 1%,本文为 12. 56%;第四,李文匹配子女 19% 受过高等教育,高于本文的 12. 14%;第五,本文匹配父亲职业为农民的

比例高于李文;第六,李文匹配父亲非农业户口比例比本文高出 10 个百分比;第七,本文匹配父亲的受教育年限和收入均比李文更高。

(二)重构模型

同样,本文构建了李文中的两组模型(见表 4)。就第一组模型回归结果[①]来看,基本结论与李文一致,但还是存在一定差异,与李文比较:一,本文显示大学扩招后子女接受高等教育的机会增长程度更大;二,高等教育机会在大学扩招后的城乡差距扩大程度没有那么高;三,男女受教育机会的差异更大;四,父亲因素(不同职业类型、受教育年限、月收入、户口身份等)对子女受教育机会的影响程度没有那么大。

而本文重构的第二组模型回归结果显示:与李文相比,本文认为父亲因素对于子女接受大学本科和大学专科教育机会的影响程度没有那么大;李文认为父亲月收入对子女大学专科教育机会的影响比对大学本科教育机会的影响更大,但本文得出相反的结论,认为父亲收入对子女接受本科教育的影响大于接受专科教育的影响。此外,李文认为大学扩招后,父亲教育背景对于子女接受大学本科和大学专科教育机会的影响不变,而本文分析认为,这种影响是在减少。

由于李文的匹配数据样本量远小于本文的匹配结果,且两种匹配结果计算的模型又存在一定差异,由此我们产生质疑:其匹配过程本身是否存在差错,而李文并没有详细说明其利用 2005 年 1% 人口抽样数据的子样本数据匹配的过程。匹配过程的正确是模型构建的基础,所以最后的研究结论是否可靠也是值得怀疑的。再者,根据上文对父子匹配数据的分析可知,子代的未匹配人口与匹配人口在诸多特征上存在显著差异,且子代数据的匹配率仅为 31%,所以,不加调整地直接使用匹配后的数据进行研究分析可能会出现比较严重的数据选择性问题。

六、户内父子匹配数据的多重选择性与使用前提

直接使用匹配数据看似论据充分,但仍存在最基本的匹配过程所

① 受篇幅所限,本文仅展示第一组模型中模型 3 的回归结果。

表 4　　重构的模型回归结果

自变量	模型3(第一组)			模型4(第二组)			模型5(第二组)		
	B	Exp(B)	S. E.	B	Exp(B)	S. E.	B	Exp(B)	S. E.
父亲职业(参照:农民)									
管理人员	1. 609 **	4. 998	. 122	1. 828 **	6. 220	. 222	1. 523 **	4. 585	. 136
专业人员	1. 138 **	3. 121	. 103	1. 519 **	4. 568	. 209	1. 074 **	2. 928	. 114
办事人员	1. 645 **	5. 183	. 104	1. 919 **	6. 815	. 207	1. 521 **	4. 575	. 115
商业服务业员工	1. 102 **	3. 012	. 096	1. 545 **	4. 688	. 206	1. 011 **	2. 748	. 107
产业工人	. 926 **	2. 524	. 092	1. 402 **	4. 062	. 202	. 847 **	2. 332	. 102
父亲受教育年限	. 245 **	1. 278	. 010	. 246 **	1. 279	. 016	. 223 **	1. 250	. 012
父亲月收入(中高收入)	. 756 **	2. 131	. 083	. 774 **	2. 169	. 099	. 602 **	1. 825	. 097
父亲户口身份(非农)	1. 172 **	3. 229	. 071	1. 382 **	3. 984	. 142	1. 056 **	2. 875	. 080
性别(男性)	−. 469 **	. 625	. 057	−. 476 **	. 621	. 083	−. 401 **	. 670	. 064
民族(少数民族)	−. 212	. 809	. 111	−. 120	. 887	. 177	−. 245 *	. 783	. 124
年龄组(1980 − 1985 年)	. 386 **	1. 471	. 124	. 516 *	1. 675	. 231	. 421 **	1. 524	. 137

续表4

自变量	模型3(第一组)			模型4(第二组)			模型5(第二组)		
	B	Exp(B)	S.E.	B	Exp(B)	S.E.	B	Exp(B)	S.E.
年龄组*管理人员	-.137	.872	.140	-.400	.670	.249	-.099	.906	.157
年龄组*专业人员	-.175	.839	.118	-.253	.776	.233	-.203	.816	.131
年龄组*办事人员	-.202	.817	.119	-.222	.801	.231	-.266*	.767	.133
年龄组*商业服务业人员	-.083	.920	.108	-.367	.693	.228	-.011	.989	.120
年龄组*产业工人	.152	1.165	.102	-.063	.939	.222	.166	1.181	.113
年龄组*父亲受教育年限	-.042**	.959	.012	-.040*	.961	.018	-.043**	.958	.013
年龄组*父亲月收入	-.009	.991	.094	.014	1.014	.111	-.024	.976	.110
年龄组*父亲户口性质	.535**	1.708	.079	.529**	1.696	.156	.467**	1.595	.089
年龄组*性别	.069	1.072	.063	.175	1.191	.092	.012	1.012	.071
年龄组*民族	-.203	.816	.126	-.373	.689	.201	-.132	.876	.141
常数项	-5.429**	.004	.108	-7.490**	.001	.206	-5.434**	.004	.120
-2 log likelihood	45422.12			22484.15			37197.8		
N	95075			95075			91263		

注：* $p<0.05$，** $p<0.01$.

可能导致的选择性。以父子匹配为例，从本文的匹配结果来看，匹配成功的为 118121 人，未匹配成功的人远远大于这个数值，达到 259545 人，在这样的情况下，匹配后的数据能否用于分析父代与子代之间关系问题是值得研究的。虽然匹配成功群体的受教育程度比未匹配成功群体略高，但差异并不很大，匹配成功群体中接受过高等教育的比例为 14.5%，未匹配成功群体中接受过高等教育的比例为 14.13%。表面上来看，“匹配”在受教育方面似乎没有选择性。

进一步分析匹配成功群体的户口状况发现，父代与子代的户口状况相同的更倾向于住在一起。从表 5 可以看出，在匹配成功的群体中，父子(女)都是农业户口占 66.61%，父子(女)都是非农业户口的占 27.57%，只有 5.82% 的父子(女)的户口状况不同。出生在农业户口家庭的人通过自己的努力(如升学)获得了非农业户口到城镇生活，一般是不会再和父母住在一起的，特别是 20－30 岁的年轻人更难以有条件把父亲接来城镇一起居住。在父代—子代被登记在同一个家庭户才能成功匹配的原则下，排除了这些农村家庭背景下获得高等教育的人，可能会夸大城乡受教育机会的差距。

表 5　　匹配成功群体的父代—子代的户口状况　　单位:(%)

子代的户口	父亲的户口		合计
	农业	非农业	
农业	66.61	2.68	69.29
非农业	3.14	27.57	30.71
合计	69.75	30.25	100.00

那么，为什么表面上看，匹配过程对“受教育程度”这一变量的选择性并不明显呢？本文依据“与户主关系”、“现住地”、“户籍所在地”等信息详细分析了父代—子代未能匹配成功的原因，也是父代—子代不住在一起的原因(见表 6)。15.68% 的人因外出流动而没有和父辈住在一起，而这些外出打工的人往往都没有受过高等教育。由此可知，从农村出来受过高等教育者不能进入匹配数据，从农村出来打工没有接受高等教育者也不能进入匹配数据，使得匹配成功人群与未匹配成功人群呈现出差异不大的受教育结构，这不是匹配没有选择性，恰恰是

多重选择性的叠加。

户内匹配数据的处理类似于缺失数据的处理，若未匹配的数据信息符合两个假定：一是未匹配数据完全随机，是否能匹配上与变量的取值无关；二是未匹配数据的分布与匹配数据的分布一致。在符合任意一种假设的基础上，直接使用匹配后数据的多元分析是合理的。但从上文分析可知，本研究的户内匹配数据显然不能符合任一假设，直接使用就有可能产生问题。

表 6 未匹配成功的原因

户类型	未匹配的原因	频数	百分比
家庭户	自立门户	116488	44.88
	市内人户分离	9854	3.80
	外出流动	40703	15.68
	出嫁或入赘	46098	17.76
	其他	8221	3.17
集体户	就业	32554	12.54
	上学	5627	2.17
总计		259545	100.00

七、户内父子匹配数据应用的改进

既然户内成员匹配数据或多或少都会存在匹配选择性问题，如何在研究分析时减少偏差呢？本文试图从数据和方法两个方面着手考虑，找到解决户内人口匹配数据误用问题的方案，利用再抽样和加权技术进行调整计算，以提高相关研究的科学性和可靠性。

匹配选择性产生的结果是：匹配样本的分布偏离于总体的分布。这样的状况就像图 1－a 和图 1－b，解决问题就是要让匹配样本分布尽量接近总体分布。下文将尝试使用两种方法对匹配样本分布进行调整：一是加权，对不同个案给予不同的权重，在权重的倍数效果下，匹配样本分布能基本接近总体分布（如图 1－c）；二是再抽样，在匹配样本中再抽取一个子样本，使得子样本的分布接近总体分布（如图 1－d）。

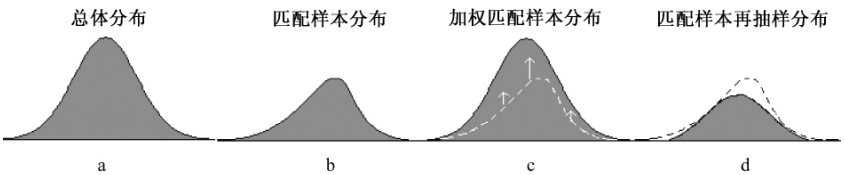

图 1　总体和匹配样本的分布

(一) 加权调整

在实际的过程中,我们无法使得匹配样本加权后的分布与总体分布完全一致,原因有二:一是我们无法对所有影响研究模型的变量进行加权,所找到的影响因素是无法穷尽的;二是许多影响研究模型的变量无法得到其在总体中的分布,自然也无法确定权重。与本文研究模型最相关的变量有:父亲职业、父亲受教育程度、父亲月收入、父亲户口身份、子女受教育程度、子女性别、子女民族、子女年龄。但未匹配的父亲信息无法获得,所以父亲职业、父亲受教育程度、父亲月收入、父亲户口身份均无法获得其在总体中的分布。由此,下文选择子女的性别、户口身份、年龄、受教育程度、是否流动、婚姻 6 个变量作为加权依据,将总体与样本的 6 维交叉表的比值作为权重。加权结果使得匹配样本在这 6 个变量上的分布与总体基本一致。

加权之后依旧重构了两组模型(表 7、表 8)。对这两组模型的分析除了可以发现与原模型不同的结论,还可以发现更多没有发现的问题。

第一,加权后的模型纠正一个错误结论。在李文和本文调整前的模型分析中,有一个不合常理的结论认为,在控制其他变量之后,女性的受教育机会大于男性。加权后的模型则反映女性的受教育机会少于男性。调整前模型结论显然是受"选择性偏差"的影响得出的,农村的女孩受教育机会明显差于男性,但这些没能上大学的女孩往往也早早出嫁,离开父母,她们"选择性"地被更多排除在研究范围之外。

第二,加权后的模型反映一些影响因素的影响程度与李文和本文调整前模型不一致。加权后模型显示,父亲受教育年限对子女接受高等教育的正向影响比原模型更大,子女接受高等教育的城乡差异比原

表 7　　加权后的第一组模型回归结果

自变量	模型 1			模型 2			模型 3		
	B	Exp(B)	S. E.	B	Exp(B)	S. E.	B	Exp(B)	S. E.
父亲职业(参照:农民)									
管理人员	1. 567 **	4. 793	. 035	1. 592 **	4. 914	. 035	1. 804 **	6. 076	. 057
专业人员	1. 026 **	2. 790	. 030	1. 048 **	2. 851	. 030	1. 195 **	3. 303	. 049
办事人员	1. 531 **	4. 624	. 030	1. 542 **	4. 674	. 030	1. 708 **	5. 515	. 050
商业服务业员工	1. 082 **	2. 952	. 026	1. 058 **	2. 879	. 026	1. 151 **	3. 162	. 046
产业工人	1. 061 **	2. 889	. 024	1. 027 **	2. 791	. 024	. 913 **	2. 491	. 044
父亲受教育年限	. 238 **	1. 269	. 003	. 232 **	1. 261	. 003	. 263 **	1. 301	. 005
父亲月收入(中高收入)	. 761 **	2. 140	. 024	. 758 **	2. 133	. 024	. 798 **	2. 221	. 041
父亲户口身份(非农户口)	1. 664 **	5. 280	. 018	1. 689 **	5. 413	. 019	1. 352 **	3. 866	. 034
性别(男性)	. 073 **	1. 075	. 015	. 075 **	1. 078	. 015	. 292 **	1. 339	. 026
民族(少数民族)	−. 392 **	. 675	. 032	−. 401 **	. 670	. 032	−. 250 **	. 779	. 054
年龄组(1980 − 1985 年)				. 381 **	1. 464	. 016	. 807 **	2. 241	. 064

续表 7

自变量	模型 1			模型 2			模型 3		
	B	Exp(B)	S. E.	B	Exp(B)	S. E.	B	Exp(B)	S. E.
年龄组 * 管理人员							-. 327 **	. 721	. 073
年龄组 * 专业人员							-. 226 **	. 798	. 062
年龄组 * 办事人员							-. 244 **	. 784	. 063
年龄组 * 商业服务人员							-. 125 *	. 883	. 056
年龄组 * 产业工人							. 156 **	1. 168	. 053
年龄组 * 父亲受教育年限							-. 047 **	. 954	. 006
年龄组 * 父亲月收入							-. 059	. 943	. 050
年龄组 * 父亲户口性质							. 484 **	1. 622	. 040
年龄组 * 性别							-. 327 **	. 721	. 031
年龄组 * 民族							-. 226 **	. 798	. 066
常数项	-5. 726 **	. 003	. 031	-5. 919	. 003	. 032	-6. 196 **	. 002	. 051
-2 log likelihood	127387. 152			126790. 138			126357. 361		
N	95075			95075			95075		

注：* p < 0. 05，** p < 0. 01，

表 8 加权后的第二组模型回归结果

自变量	模型 4			模型 5		
	B	Exp(B)	S. E.	B	Exp(B)	S. E.
父亲职业(参照:农民)						
管理人员	1.742**	5.711	.101	1.745**	5.724	.064
专业人员	1.406**	4.079	.096	1.148**	3.152	.055
办事人员	1.752**	5.767	.095	1.628**	5.093	.056
商业服务业员工	1.431**	4.184	.095	1.086**	2.963	.052
产业工人	1.220**	3.389	.093	.856**	2.355	.050
父亲受教育年限	.252**	1.286	.007	.247**	1.280	.006
父亲月收入(中高收入)	.762**	2.143	.047	.668**	1.951	.047
父亲户口身份(非农户口)	1.512**	4.535	.067	1.228**	3.415	.038
性别(男性)	.390**	1.476	.039	.198**	1.219	.029
民族(少数民族)	-.286**	.751	.088	-.253**	.776	.060
年龄组(1980 - 1985 年)	.930**	2.533	.113	.830**	2.293	.072
年龄组* 管理人员	-.332**	.717	.122	-.303**	.739	.082
年龄组* 专业人员	-.164	.849	.115	-.268**	.765	.070
年龄组* 办事人员	-.067	.935	.114	-.374**	.688	.071
年龄组* 商业服务业人员	-.266*	.767	.112	-.076	.927	.063
年龄组* 产业工人	.110	1.116	.108	.143*	1.154	.059
年龄组* 父亲受教育年限	-.045**	.956	.009	-.055**	.947	.007
年龄组* 父亲月收入	.014	1.014	.056	-.118*	.888	.058
年龄组* 父亲户口性质	.536**	1.709	.078	.375**	1.454	.046
年龄组* 性别	-.315**	.730	.046	-.291**	.747	.036
年龄组* 民族	-.266*	.767	.106	-.154*	.857	.074
常数项	-8.008**	.000	.093	-6.183**	.002	.057
-2 log likelihood	66048.086			101000.534		
N	95075			91263		

注:* $p<0.05$,**$p<0.01$.

模型更大,扩招对子女受教育机会的正向影响也比原模型大得多。例如:加权后的模型 3 显示,扩招使得子女受教育机会提升了一倍,提升幅度远大于李文模型 3 的结果。

加权后模型显示，父亲职业、父亲受教育年限、父亲月收入、父亲户口身份对子女大学本科和大学专科教育机会都有影响，同时，对大学本科教育机会的影响大于对大学专科的影响。

第三，加权后的模型发现一些过去没有发现的结论。加权后模型中的年龄组与父亲职业、教育、户口性质、子女性别、子女民族的交叉项均呈现显著状态。加权后模型显示：我国高等教育扩招后，子女接受高等教育的机会（不论是大学专科，还是大学本科）受父亲教育、职业的影响减少了，性别、民族的差异也减少了，但城乡差距扩大了。

表 9　　加权前后的主要结论对比

李文结论	加权后结论
受高等教育机会存在不平等性：阶层不平等、城乡不平等、性别不平等、民族不平等	支持原结论，不同的只是“不平等”的程度
大学扩招并未使高等教育机会的阶层不平等、民族不平等下降，而且城乡不平等程度还有明显上升	大学扩招使阶层不平等和民族不平等下降，但城乡不平等反而上升了
高等教育系统内部的等级分层与教育机会不平等有交叉作用，较高等级的高等教育领域（大本）的机会不平等大于较低等级的高等教育领域（大专），尤其表现在阶层不平等和城乡不平等方面	支持原结论
大学扩招不仅没能减少较高等级的高等教育（大学本科）机会的不平等，也没有减少较低等级的高等教育（大学专科）机会的不平等	大学扩招使得两个等级的教育机会阶层不平等和民族不平等均下降了，城乡不平等上升了

（二）再抽样调整

尽管我们期望再抽样得到一个与总体分布完全一致的样本，但实际也无法做到。一是总体的分布是无法完全把握的，一些影响研究模型的变量无法得到其在总体中的分布；二是若考虑到所有影响研究模型的因素，在抽样过程中进行过多的分层反而会使得样本分布偏离总体。本文选择子女的性别、户口身份、年龄、受教育程度、是否流动、婚姻 6 个变量作为再抽样分层的依据，再抽样结果使得匹配样本在这 6 个变量上的分布与总体基本一致（限于篇幅，具体模型分析结果略去）。

再抽样后构建的两组模型，也可以纠正“性别差异”的错误结论，

但其他方面不如加权模型的调整效果理想。原因有以下三个方面:

第一,再抽样的范围受匹配后数据的限制,得到的样本分布与总体分布的接近程度不如加权的方法;第二,所有的抽样都会有抽样误差,再抽样便会再次产生抽样误差,解决的方法是多次再抽样后得到均值,再抽样的次数越多,均值才会越接近总体的均值,但又带来极大的工作量;第三,再抽样使得进入模型的样本量大幅度减少,一些变量的回归系数不显著,使得结论变得不清晰。

八、结论与讨论

人口普查和人口抽样调查不仅是人口学研究的重要数据资源,同时也是经济学、社会学等相关研究的重要信息资源,如何正确使用和正确理解数据背后的客观规律是研究者追求的主要目标。通过本文的研究,我们得出以下基本结论:

第一,父子未匹配人口占目标总体的比例很高。父子未匹配的比例明显大于母子和夫妻关系匹配。我国 2000 年人口普查和 2005 年 1% 人口抽样调查中 20 – 30 岁人口父子不能匹配的占 2/3 左右,能够匹配的仅占 1/3 左右。

第二,匹配与未匹配人口是两个明显不同的群体,因此,在使用人口普查和人口 1% 抽样调查进行户内关系匹配数据时,由于统计口径和登记方式,特别是由于家庭生命周期和人口迁移流动等方面的原因,使匹配人口数据具有明显的选择性。由于匹配人口与未匹配人口的基本特征具有明显差异,而且这种差异可能对研究结论产生很大影响,因此在数据使用过程中首先需要对户内匹配人口与未匹配人口基本特征进行统计检验。

第三,不同的人口群体的数据匹配,由于家庭生命周期特点与年龄、性别等最基本的人口学指标密切相关,因此,通过户内关系进行匹配需要充分考虑年龄、性别、户口类型以及流动迁移特点的不同,总的来说,由于年龄的不同,户内匹配数据选择性偏差不同,年轻子女、中年夫妻匹配的可能性更大一些。

第四,具有选择性偏差的数据在进行统计分析过程中需要考虑偏差的来源和产生的影响,尽可能降低选择性偏差带来的问题,否则有可

能得出错误的结论。

第五，通过对李文的匹配过程再检验，以及利用再抽样和加权方法重构模型，我们发现，尽管李文的研究结论不存在大的差错，但匹配的选择性偏差对研究模型的影响是确定的，不仅会带来影响因素判断程度的错误，甚至完全改变影响因素的作用方向。加权和再抽样方法能够在一定程度上弥补“选择性偏差”，相比来说，加权模型的调整效果更加理想。

总之，本文以2005年1%人口抽样调查原始抽样数据研究户内人口匹配数据，并对李春玲文的研究内容进行了再检验，提出了针对户内人口匹配数据选择性偏差的调整思路，研究重点是讨论户内人口匹配和在使用匹配数据时忽略未匹配人口可能带来的选择性偏差问题。但是，对于2000年人口普查、2005年1%人口抽样调查等原始抽样数据本身存在的问题和由于原始数据偏差所引起的研究偏差并没有涉及。而且，加权和抽样方法所能弥补的“选择性偏差”是有限的，尽管本文前半部分认为“选择性偏差”会夸大城乡教育机会的不平等，但数据调整后依然没能对这一结论进行修正。因此，为了深入研究高等教育扩展与教育机会不平等问题，还需要在基础数据和统计方法方面作进一步的努力。

参考文献：

李春玲，2010，《高等教育扩张与教育机会不平等——高校扩招的平等化效应考查》，《社会学研究》第3期。

李玉柱、姜玉，2009，《80年代以来我国妇女初婚初育间隔变动分析》，《西北人口》第3期。

吴晓刚，2009，《1990－2000年中国的经济转型、学校扩招和教育不平等》，《社会》第5期。

李志宏，2004，《北京市夫妇年龄差分析》，《市场与人口分析》第5期。

郭志刚、李睿，2008，《从人口普查数据看族际通婚夫妇的婚龄、生育数及其子女的民族选择》，《社会学研究》第5期。

作者单位：中国社会科学院人口与劳动经济研究所

责任编辑：谭　深

SOCIOLOGICAL STUDIES 2011

(Bimonthly) 3

Vol. 26 May,2011

PAPER

Abstract: Hume's problem of causal induction has inspired more than two centuries of epistemological discussion. J. S. Mill was the first one to elaborate the methodological principles of causal induction. The Millian principle of "overall similarities" for causal inference laid the foundation for Fisher's randomized experimental design and guided causal analyses in multivariate statistical modeling and qualitative comparative case studies. Causal theories are not, however, induced from empirical correlations. They are produced through leaps of faith and empirically tested through deductive logic. The best research strategy is triangulation that combines theorizing, qualitative analysis, and statistical modeling.

Abstract: The data of household members matched are widely used in sociology, demography and related research, but the selection bias of the matched data is often ignored. This paper uses the raw data of the Fifth Census in 2000 and 1% Population Sample Survey in 2005 to match three kinds of relations, i. e., father and sons, mother and children, husband and wife, and confirms that the selection bias exist in these matched data grouped by age, gender, migration status, rural and urban distribution, education, geographical distribution, and so on. Based on the matched data, this paper re-tests the matched data, analysis models and conclusions of the paper *Expansion of Higher Education and Inequality in Opportunity of Education* and finds out that the selection bias of matched data would influence the accuracy of model analysis and research conclusion. For further reducing the impact of matched data bias, we propose two adjustment methods and find out that re-sampling and weighting methods can reduce the selective bias.