

住户追踪调查中的样本维护策略

吕萍^a,严洁^b

(北京大学 a.中国社会科学调查中心;b.政府管理学院,北京 100871)

摘要:对一个大型的追踪调查来说,长期的追踪调查增加调查对象的回答负担,样本老化、样本流失等导致追踪样本不能很好地反映调查总体的变化特征,有效的样本维护策略是十分重要的。文章在介绍了追踪调查的优缺点基础上,主要对住户追踪调查的追踪规则,住户追踪调查中的一些维护策略,包含影响样本维护的因素、常规和非常规的样本维护措施进行探讨。

关键词:追踪调查;住户追踪调查;样本维护;样本流失

中图分类号:C811 **文献标识码:**A **文章编号:**1002-6487(2015)07-0004-04

0 引言

追踪调查逐渐成为抽样调查的主要方法之一。追踪调查主要适用于收集样本单元在不同时间上的连续资料,通过不同时间的调查资料的统计分析探索社会现象随时间变化而发生的变化特征及不同现象之间的因果关系。追踪调查又分为初访调查和追访调查,各期之间的间隔一般都很长。跟踪调查数据的出现,有效地推动了社会科学和政府科学决策的发展。以PSID为例,美国社会科学的定量研究中大量的期刊文章运用了PSID的数据,其中包含4位诺贝尔奖得主。与此同时,美国联邦政府的房屋与城市发展部、健康与人类服务部、财政部、农业部、劳工部、社会保障部、白宫、国会等政府部门,大量使用了PSID的数据作为决策依据。由于跟踪数据对学术发展的重要贡献,在美国国家科学基金会评选的50年资助的最有价值的50个项目中,PSID是唯一的社会科学项目。由此可见追踪调查是十分重要的。

中国的追踪调查起步较晚,近几年才逐渐出现了一些大型的追踪调查,一方面是由于追踪调查需要投入大量的人力、物力和财力,另一方面,我国追踪调查的理论方法研究相对落后,限制了连续性调查的发展。但是,随着一些高校等相关科研机构、政府部门陆续建立调查研究中心,例如北京大学中国社会科学调查中心、中国人民大学中国调查与数据中心等的成立,中国家庭动态跟踪调查、中国健康与养老追踪调查等大型追踪调查的展开,大大地推动中国追踪调查的发展。

1 追踪调查存在的问题

追踪调查存在一些缺点,长期的追踪调查增加了调查

对象的回答负担,样本老化、样本流失等导致追踪样本不能很好地反映调查总体的变化特征。追踪调查的抽样设计、抽样误差和非抽样误差的研究比截面调查更加复杂,在追踪调查中主要存在以下问题:

(1)在追踪调查的初访设计中需要考虑更多的因素,不仅要考虑概率抽样,也要考虑调查总体、样本总体的变化、实际调查中样本追踪的困难程度、样本流失和样本无应答等的影响。因此,在首期调查的抽样设计时,不仅要考虑样本的代表性,也需要考虑以后各期追踪样本的有效性和样本量的变化,及各期样本的无应答和样本流失的影响。样本追踪的成功率是追踪调查关注的焦点问题之一,尽管采取很多的方法提高样本的应答率,但是调查的流失率仍在逐年增加。有些追踪调查为了能够有效地追踪样本,甚至舍弃随机抽样,将调查对象是否愿意长期参与、以后各期是否更容易追踪到作为选择研究对象的标准,这种方式虽然提高了追踪率,但却损失了调查的效率。因此,在追踪调查的初访设计中,一定要结合调查目的和目标群体的特征确定样本量和抽样设计。在实际调查中也采取有效的方式提高样本的追踪率,根据调查对象的特点、调查方式和间隔时间等因素,制定一组明确的、切实可行的追踪规则和样本维护策略。

(2)样本老化问题是追踪调查中重点关注的问题之一。追踪调查需要对同一样本单元反复调查,而前期调查很容易改变调查者的回答习惯,对后期调查结果产生影响,造成回答偏差。另外,对同一样本单元进行反复调查,加重被调查者的负担,容易导致样本老化,使被调查者产生一定的厌倦情绪,不如实回答调查信息、拒绝接受调查等,导致抽样框误差、无应答误差和计量误差,影响调查数据的质量。此时,一方面需要对样本单元定期维护,让受访者了解调查的意义,提高数据的质量。另一方面通过有效的统计方法降低各种应答率偏差,提高数据质量,追踪

基金项目:国家自然科学基金资助项目(71171004);国家社会科学基金资助项目(11CTJ005)

作者简介:吕萍(1981-),女,山东泰安人,副研究员,研究方向:统计调查方法。

调查获得更加丰富的数据,也为评估、控制、深入研究非抽样误差提供更多机会。

(3)样本无应答。无应答也是追踪调查的难点问题之一。在截面调查中,无应答包含单元无应答和项目无应答两种。单元无应答是指某一样本单元没有接受调查,项目无应答是指某一单元接受了调查,但由于调查员遗漏、被调查拒访等原因导致调查问卷中的某个问题没有应答。追踪调查除了单元无应答和项目无应答外,还有期单元无应答和期项目无应答,因此追踪调查中的无应答比截面调查更加复杂,导致更多的无应答误差。此时,一方面通过加强访员培训提高访员的访问技术,采用多次访问、激励措施等方式降低无应答,另一方面通过加权、插补的方法对数据进行事后调整。因此追踪调查获得更多的信息,也为各种缺失数据的方法研究提供了更加丰富的数据资料 and 实践经验。

(4)样本流失是追踪调查无应答的一种特殊形式,是追踪调查的一个重要特征,也是追踪调查的一个挑战。样本流失是指在首次调查之后,样本单元由于去世、搬迁、出国和拒绝访问等原因无法追踪到,其中样本由于搬迁而无法找到是样本流失最主要的原因,虽然制订有效的样本维护策略可以防止样本流失,但由于实际调查的复杂性,样本流失仍是无法避免的,尤其是当样本流失的比例过大,流失样本具有特殊性质时,最终得到的估计量将是严重有偏的,无法反映总体的变化。随着社会经济的发展,样本流失的比例也在逐年增加,此时,一方面采取有效的措施进行样本维护,通过统计方法对样本进行有效分类采取不同的维护策略。其次,定期补充样本从而保证样本的代表性。最后,通过事后分层调整等加权方法提高样本的代表性,但当样本流失的群体与接受调查的群体差距较大时,无论是何种方法都会无法弥补精度的损失。

2 住户追踪调查

住户追踪调查是一种特殊的追踪调查,是以住户为群进行的追踪调查,这种追踪调查不仅可以获得个体层面的信息,而且能够获得个体所在住户的信息。为了解本国或某地区住户的生活、环境、教育、健康等问题,许多国家都开展住户跟踪调查,为社会政策分析、发展规划、计划管理等提供数据支持。住户追踪调查已成为人们组织国家统计体系、制定政策的核心战略内容,美国的PSID项目即为典型的住户追踪调查项目,此外,英国的BHPS、德国的GSOEP、台湾的PSFD,欧盟ECHP,澳大利亚的HILDA,瑞士的SHP,英国的UKHLS等都是典型的住户调查。

中国的追踪调查起步较晚,其中最具代表性的是北京大学中国社会科学调查中心主持的中国健康与养老追踪调查(CHARLS)和中国家庭动态跟踪调查(CFPS)。大型追踪调查的实施一方面大大推动中国追踪调查的理论和应用的发展,另一方面对追踪调查在中国的应用也提出了更高的要求。保证追踪调查的效率,提高样本的追踪率,

减少追踪调查的样本流失是一个迫切需要解决的问题。本文主要针对住户追踪调查中样本流失问题进行初步探讨,并结合笔者的实际工作经验初步探索在中国住户追踪调查中如何制定有效地样本维护策略。

在实际调查中,住户追踪调查有不同的分类方法。按照调查对象划分,追踪调查一方面可以在每个样本住户中只抽选一人(随机抽选或直接选取户主)作为长期跟踪调查的对象,另一方面也可以将样本住户中的所有人(或满足调查条件的多个个体)作为长期跟踪调查的对象。按照调查目的划分,若追踪调查目的是对各个时点样本特征的变化进行分析,则只需要对初始的样本总体进行跟踪;若调查目的是为了需要满足对调查总体的特征的估计,则需要考虑调查总体和样本总体的变化,在实际调查中,追踪调查的目的多为后者。按照追踪对象划分,有的调查只针对初期的所有样本住户进行追踪,有的调查只针对初期完访的住户进行追踪调查,有研究表明,初期调查的拒绝回答者,后期接受调查的概率很小,因此大多追踪调查为了节省经费只对初访调查中完成样本住户或样本个人进行调查。

对于一个大型的追踪调查,随着时间的变化,调查总体和样本总体都会发生变化。在调查总体中,总体单元以出生、迁入等方式进入调查总体,总体单元以死亡、迁出等方式离开调查总体,导致调查总体的变化。而样本总体的变化包括样本住户和样本个体的变化,其中样本住户的变化一方面是由于样本住户的无回答、流失等导致样本住户减少,另一方面由于样本住户内个体的出生、死亡、迁入、迁出、婚姻等变化导致的样本住户的变化。在追踪调查中,为了满足对调查总体特征进行估计的目的,在住户追踪调查过程中,需要在分析明确调查总体和样本总体及其变化的基础上,确定样本的追踪规则和样本维护策略。

3 追踪规则

对于住户追踪调查来说,追访追踪规则是十分重要的,即在调查中需要确定哪些个体是需要每年追访的,哪些个体是不需要每年追访的。在追踪调查中,理论上应该定义初访的所有样本住户及其满足条件的家庭成员为初始追踪对象。但在实际调查中,为了节省经费,常常定义追踪住户是初访调查中家中至少完成了一份调查问卷的住户为初始样本住户,样本住户中的所有满足条件的样本个体定义为初始样本单元(有些调查中也成为基因成员),并对初始样本单元固定时间进行追踪。由上文,随着时间的变化,调查总体和样本总体都会发生变化。不同的追踪规则下,定义的调查总体和样本总体是不同的。追踪调查中的调查总体主要分为以下几类:

(1)初访调查总体。即初访调查中的调查总体,是一个静态的总体。

(2)初访调查总体和初访样本的后代。

(3)初访调查总体和追踪期与初访样本同住的个体。

(4)初访调查总体和曾经在各期调查中接受调查的个体。

(5)初访调查总体和所有层在调查期接受调查的个体以及与之同住的家庭成员。

由上述调查总体产生的追踪规则是:

(1)只追踪初访样本(纵向样本),是最为简单的追踪调查。通过对初访样本进行长期追踪获得初访样本的变化特征,其中样本流失、无回答是一个需要注意的问题,样本维护对此类调查至关重要。

(2)追踪初访样本和调查期与之同住的家庭成员,当同住人员离开该样本住户,并且进入的住户不再包含纵向样本时,则不再进行追踪。

(3)追踪初访样本及其后代,在这种追踪规则下,即使后代离开样本住户也需要进行追访。

(4)追踪初访样本和在追踪调查中曾经接受过调查的同住个体,当这些同住个体离开样本住户后不再进行追踪。

(5)追踪在调查中曾经接受过调查的所有个体,以及这些样本在调查期的同住个体。

由上,不同的追踪规则不仅需要确定调查初访样本,而且需要明确由追踪规则组成的间接样本。几乎所有的追踪调查的第一个标准是共有—个家(住宅单位),另一个标准是存在—定的“共同生活安排”,这个标准在每个国家中略有不同,例如LISS提出的每周至少4天在家吃住的标准,该调查定义初访调查中至少完成家庭成员问卷的住户为初访住户,初访住户中的所有家庭成员为初访样本成员。调查中的追踪对象是户主、户主的伴侣、户主的子女(子女通常指在该家吃住每周至少四天以上的孩子)、其他家庭成员(通常指在该家吃住每周至少四天以上的成员,例如户主的父母等)。由于实际情况的复杂性,在LISS调查中还包含在初访调查中不在家中,但有意向—年内回家并把该住户作为主要居住地的人同样被视为住户的初访样本,其中,暂住人员,例如实习、旅游、度假、短时间住院,这些人视为即将放回的样本住户的初始家庭成员。

BHPS用同样的区分方式,但它增加了同居者,严格来说这些人不算是纵向受访者或初始样本成员,但他们与某个初始样本单元有较强的联系,如(祖)父母或回家居住的孩子,但是当该同居者离开样本住户后则不再进行访问。而在GSOEP中,所有成员曾经参加过调查的样本成员都被跟踪,只要曾经接受过调查都将被长期跟踪。ECHP和PSID区分纵向基因成员和同居者,只追踪基因成员的分裂住户,PSID只对样本住户的—个人进行调查。

不同的调查有不同的追踪规则,为了减少追踪调查的样本流失,就需要针对不同的调查总体、样本总体以及追踪样本追踪规则制定不同的样本维护策略。

4 样本维护的基本措施

追踪调查中的样本流失与管理学中的客户流失有相

似之处,受到各个方面的多个因素的共同影响。样本流失有多种分类方式,根据流失原因的不同,可将流失分为主动流失(样本个体不满意该调查拒绝接受调查)和被动流失(由于出国,进入调查机构不再满足调查要求,或由于搬迁、生病、死亡等无法接受调查)。按流失形式又可分为易流失样本(样本住户不情愿接受调查,经过多次说服接受了调查但是调查问卷存在大量的项目无回答)与主动流失样本(由于某种原因,样本几乎终止了调查)等等。在追踪调查中,一般将样本住户分为稳定住户或个人(例如每轮都接受调查、对调查十分感兴趣,调查问卷的回答质量较高等),易流失住户或个人(调查者不情愿接受调查、经常不参加调查、问卷质量较差等、曾多次拒绝访问和住户内的完访率较低等)以及流失住户或个人(样本明确表示不再参加调查,并且经过调查组织者多次说服失败不再参加调查)。由于追踪调查中有大量前期调查的问卷数据和并行数据,利用数据挖掘等统计方法确定流失住户的特征和流失趋势,对不同的住户制定不同的样本维护策略,提高样本追踪率,减少样本流失。导致追踪调查中样本流失有很多方面,需要利用前期的调查数据、并行数据对样本进行统计分析,针对样本的不同特点制定不同的维护策略,尽量避免样本流失。不同的调查项目有不同的维护方式,通常分为常规维护和特殊维护。

4.1 常规的样本维护策略

常规的样本维护策略,主要体现在在调查季前、调查季后、各个重要的节日的维护。

(1)调查季前的样本维护

在调查正式开始前,有必要以调查机构的名义给受访户邮寄信件和相关资料,—般以“致受访户的一封信”形式,主要是告之调查何时何地开始,同时将当前调查项目的进展情况、本次调查的执行时间、执行周期以及目前调查取得的成果、项目的进展告之受访户,使其具有主人翁的意识,有—份参与实际调查的自豪感和使命感。同时发放—份“联系信息变更表”,若受访者有信息更改可以通过信件、邮件、短信的方式给调查机构,调查机构将根据受访者提供的信息进一步完善样本维护地址和联系信息库,对于给出信息变更表的人员给予适当的奖励,并鼓励受访者继续参与调查并提前告知方便的访问时间。

(2)调查结束后维护

在调查结束后,调查者可以以机构或调查项目组的名义给受访者邮寄—份感谢信,主要是对受访户表示感谢,并说明此次调查的进展和完成情况,并简单介绍本年度会有什么样的维护措施和奖励措施,同时发送—份“联系信息变更表”,鼓励受访者在个人或家庭信息发生变化及时通知调查机构。

调查季前、后的样本维护在各类调查中都存在,但追踪调查有更加丰富的信息,因此在具体执行中,需要结合受访者的特征给出不同的信件,进行不同的维护,例如对拒访户、应答户以及无联系的住户给出不同的信件。随着追踪调查的开展,充分挖掘调查获得各类信息,将受访住

户分类,采用不同的样本维护方案是十分重要的,例如利用分类树、聚类、神经网络、logistic 回归模型、支持向量机等数据挖掘方法将受访住户分为易流失人群和稳定人群,进而采用不同的维护策略。

另一方面,计算机辅助调查广泛的应用于大型的追踪调查中,同样的,在样本维护中利用计算机辅助面访调查、计算机辅助电访调查、计算机辅助网络调查等,借助于计算机和网络进行样本维护是十分有效的,同时可以大大减少维护费用。首先,分析前期的数据资料,包含计算机调查产生的联系结果代码、访员观察、访问时间、访问时长等利用统计方法挖掘信息,将用户分为多种类型。例如,通过受访户提供的联系信息进行维护,对于稳定住户,首先进行网络维护,通过受访户提供的电子邮件、QQ、MSN 等发送电子邮件进行维护,对于电子邮件是退信、无网络邮件等,并且在前期已经证明邮寄地址有效的情况下,通过邮寄信件的方式完善,而对于易流失的住户则最好采用电话或实地维护。

(3) 节日维护

在中国,节日是非常重要的,尤其是春节。在追踪调查中,若频繁的与受访户联系,会打扰受访户,容易导致受访户的反感,但若长期不联系则容易导致与受访者的疏离。春节作为中国人的一个重大节日,在这个节日前与受访户联系,提前预祝新春快乐是进行样本维护的一个绝佳的机会,在问候的同时询问一下受访者家中人员、联系信息的变化,并给受访者邮寄春节贺卡,一方面表示对受访者的春节的问候,另一方面获得了受访者的最新联系信息。因此在春节维护时多通过电话维护方式并辅以少量的实地维护,设计一份有效的春节维护问卷是十分重要的,例如,对本项目的满意程度啊、是否收到之前的维护信件、是否愿意参加下一次的调查等。本次获得维护问卷数据以及并行数据,并结合之前的调查数据信息对受访户进行分析,以便及时预测受访者倾向流失的情况。当然,在春节维护中,若可以利用前期的信息对样本住户进行有效分类,将其中的易流失人群通过有经验的样本维护人员亲自上门派送新年贺卡,并表示春节问候,也将会是收到意想不到的维护效果。

一个长期的大型追踪调查,除了借助网络、电话、邮寄等方式的样本维护外,还需要进行一定的实地维护,因此需要大量的样本维护人员。在实际中为了节省经费,经常在样本点招聘地方样本人员,他们作为调查机构与受访者的一个纽带,一方面这些人了解当地的风俗,与受访户更加贴近,可以方便及时地向受访户传达调查机构的信息并及时的反馈受访户的信息,并将当地大范围的拆搬迁、及时反馈给调查机构。

4.2 非常规的样本维护措施

在常规维护的基础上,还需要进行一些非常规的样本维护措施,非常规的维护是常规维护外的特殊维护。主要有以下几种方式:

(1) 不定期或定期举办抽奖活动。例如在适当时期对一些连续参加调查并且调查数据质量较好的住户,根据调查项目的经费情况抽取部分住户给予一定的奖励,可以奖励现金、礼物或组织旅游、参观等活动,并在调查机构网站上进行宣传,激励受访者继续参加调查。

(2) 搭车调查的维护。有时一些调查为了节省经费,会在大型追踪调查的基础上进行小型的子样本调查或搭车调查,该搭车调查获得的数据(包含部分问卷数据和并行数据)可以视为一次非常规的样本维护,获得数据与其他数据一起对样本户进行分析,修正样本维护策略。

(3) 根据样本住户特征制定的特殊的维护策略。例如天灾、丧病等,调查机构给予特殊的帮助。

4.3 样本维护系统

除了对样本进行上述常规和特殊维护外,一个长期的大型调查,同样需要建立一个样本维护系统,记录样本的变化信息,并利用样本维护系统定期去样本进行维护,例如在各类节日调查季前后等定期发送短信、邮件等,专人处理受访户回馈的电话、邮件等信息,针对不同的住户采用不同的样本维护策略。例如,首先将住户按照调查数据分析分为稳定住户和易流失住户,对稳定住户并且有电子邮件、邮寄地址、手机、QQ、MSN 等信息,通过网络或电子邮件进行维护,节省大量的经费。而对于易流失的住户根据样本住户的特征进行电话或实地维护,提高应答率,同时更好地了解受访者的特征,进而采取有效的措施将其转化为稳定住户。

5 总结

随着中国社会、经济的发展,追踪调查成为为了解社会发展变迁的主要调查方法之一。但是对于一个大型的追踪调查来说,长期的追踪调查容易到样本老化、样本流失等多种问题,影响样本的代表性。因此,有效的样本维护策略是十分重要的,本文简要介绍了追踪调查的优缺点,住户追踪调查中的一些维护策略,包含影响样本维护的因素、常规和非常规的样本维护措施。为进一步解决中国追踪调查中的样本维护问题提供参考。

参考文献:

- [1]Kasprzyk D, Duncan G, Kalton G. Panel Survey[M].New York:John Wiley&Sons,1989.
- [2]Lynn P. Methodology of Longitudinal Surveys[M]. New York: Wiley, 2009.
- [3]Herrio R A, Kasprzyk D. The Survey of Income and Program Participation[J]. Proceedings Social Statistica l Section,American Statistica l Association, 1984, 52(1).
- [4]Kish L, Survey Sampling[M].New York: Wiley,1965.
- [5]Groves I R M, Couper M P. Nonresponse in Household Interview Survey[M].New York: Wiley,1998.

(责任编辑/亦民)