China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-35

系列编辑：谢宇　责任编辑：赵启琛

中国家庭追踪调查
2016 年数据库介绍及数据清理报告

吴琼 戴利红 甄祺 张婧申 谷丽萍 张聪 赵方圆

2018.10.23

一． 背景介绍

CFPS2016 为中国家庭追踪调查的第四轮全国调查，集中的面访时间为 2016 年 7 月至 11 月，加上后期的外出家庭追访以及电话调查，调查执行期持续到 2017 年 5 月。2017 年 9 月 CFPS 项目组发布了成人、少儿、经济库的测试版本，2018 年 4 月发布了家庭关系库，2018 年 7 月发布了成人、少儿、经济库的正式版本以及个人跨年库。这份技术报告针对 2018 年 4 月及之后发布的各库清理过程进行梳理，并向用户介绍使用这些数据库的注意事项。

CFPS2016 最终完成家庭层面有效样本 14763 户，个人有效样本 45319 份。以 2014 年调查完访样本为基础，CFPS2016 在家庭层面的追踪率是 89%，个人层面追踪率是 82%。如果以 2010 年基线调查在家庭关系库中界定的 57155 名家庭成员为基础，经过六年之后，CFPS2016 成功追踪到该基线样本的 69%。在所有 CFPS2016 的样本中，约两成左右由电话访问完成。

从问卷内容上看，此轮问卷基本保持了与 CFPS2014 相同的模块，但将以前个人问卷中分属成人和少儿问卷但内容相同的一些模块改成了共用模块。这步变更对数据的使用将产生两方面的影响：在 CFPS2016 内部，成人和少儿模块中相同的问题会有同样的变量名，为同一轮数据的跨库使用增加了便捷性；但在跨年间，CFPS2016 的这个变动有可能造成部分变量在不同轮次间变量名发生改动，需要用户比对问卷进行确认。CFPS2016 还新增了部分问卷采集内容，也相应地删除了部分题目，具体的变化内容可以从 CFPS 项目网站上的《历年问卷内容变动表》中获取。CFPS2016 在问卷结构上的另一个重大调整是大幅度扩充了电访调查的内容，除了认知测试只在面访中提问之外，其他所有问题在面访和电访问卷中都是一致的。

二． 2016 年问卷数据清理步骤

1. 中断样本的确认

在访问已经开始后，由于各种原因（如受访者中途退出，访问系统问题以及其它原因）而需要中止访问的样本都属于调查中的中断样本。中断样本中的大部分在后期成功补充完成并收录在数据库中，但有少量处于未完成状态。我们针对这种未完成状态的中断样本，检查其问卷进展的完整度，如果超过 80%的完成率，则将其纳入发布数据集中。根据这个筛选标准，CFPS2016 共纳入 78 个中断样本观测，其中家庭成员库 35 个，家庭经济库 8 个，成人库 28 个，少儿库 7 个。中断样本以 interrupt 变量来指征，中断样本的 interrupt 变量值取 1。

2. 各库样本编码清理

CFPS2016 的各库样本编码清理工作包含以下环节。1）清理各库内部样本的重复情况。各库内部 id 的重复主要来自两个方面的原因：一是由于在执行过程中启用了备用问卷，这些重复样本的确定绝大部分可以通过结果代码是否指征该条样本正常完成来进行判定；二是同一个体被多个关联家庭认为是自家成员，这些重复样本的确定需要结合个人的在家状态和问卷完访时间来综合判断。2）清理跨库间样本编码的逻辑关系，确保所有数据集的观测以家庭成员库为出发点，避免个人库的样本无特定家庭归属的情况。3）以家庭关系库为基础，确认和清理所有关联家户中相同名字但不同 pid 的个体。关联家庭中同名不同 pid 出现的主要原因是往期或是当期调查中进入 CFPS 的个体在当期关联家庭中以新人的身份再次进入 CFPS，我们对此类样本会综合考虑完访形式、完访时间、在家状态、数据完整性等因素，保留合适的问卷数据，并将其 pid 在各库中统一为个人初次进入 CFPS 调查时的 pid。CFPS2016 中共清理此类样本约 740 条。4）以家庭成员库为基础，清理各问卷中涉及到的家庭成员列表中的样本编码。5）两条在 2018 年执行时发现年龄矛盾的样本从 2016 年少儿库中删除。

3. 数值变量的录音核查

CFPS 清理过程中对数值的分布进行核查时，会发现一些过大或过小的奇异值。针对这些存疑观测，我们在清理中将部分变量的录音数据进行提取，如果发现存在访员记录错误的情况，则对该条数据进行相应更新。由于受访者回答的录音质量效果不如访员提问的录音质量，并非所有的存疑数据我们都能听到有效的录音信息。表 1 列出了通过录音核查改动超过 10 条的变量及各变量通过录音信息所更新的观测条数。

表 1　CFPS2016 存疑观测的录音核查涉及变量

变量名	变量标签	修改观测数
家庭经济库		
FQ5	房屋购建成本（万元）	142
FM401	全部经营总资产（万元）	136
FS6V	耐用消费品总值	40
FT1	您家现金及存款总额（元）	38
FR2	其他房产市价（万元）	34
FQ5	房屋构建成本（万元）	33
FM401	全部经营总资产(万元)	30
FT302	所有住房房贷总支出（万元）	25
FQ6	房子当前市价（万元）	24
FT601	亲友借款待偿额(元)	23

FT101	您家定期存款总额（元）	21
FT501	您家待偿贷款额（元）	21
FEXP	过去 12 个月总支出（元）	17
FS7V	农用机械总值	17
FINC	过去 12 个月总收入（元）	12
FO7	工资收入总额（元）	10
FP506	住房维修费（元/年）	10
FT401	购房建房装修借款额度（元）	10
成人库		
QP702	一周锻炼时长（小时）	23
QG801_A_1	每月实物福利（元）-免费早/中/晚餐	17
QQ402	午休时长（分）	17
QF302_A_1	子女 1 向你提供经济帮助的额度（元）	16
KS1011	非周末学习时间（小时）	15
QQ403A	晚上睡觉时间（点）	12
KU102A	每月公付手机费（元）	10
QG3011	上下班单程时间（分钟）	10

4. 数据整合

根据访问模式的不同和回答人的不同，CFPS 针对同样的受访者可能会有不同的问卷。访问模式会决定受访者使用的是面访问卷还是电访问卷，这两类问卷内容在 CFPS2016 之前的差别很大，但由于 CFPS2016 中电访与面访已经实施高度整合，二者内容的差别只在认知测试部分。CFPS2016 在执行时继续沿用 CFPS2014 时面访为主，电访为辅的原则。电访比例相比前轮继续上升。CFPS2016 中家庭成员、家庭经济、成人、少儿问卷中电访所占比例分别为 18.89%、16.29%、18.07%和 19.92%。

根据回答人的不同，问卷可能是受访者自己回答的自答问卷或知情人回答的代答问卷。自答问卷和代答问卷的差别较大，代答问卷是自答问卷的精简版本。在整合自答和代答问卷数据库时，我们对两种问卷进行了逐题比对，将在两套问卷中完全相同的问题取相同的变量名，而有所不同的题采用不同的变量名。当某个样本既存在自答问卷又存在代答问卷时，我们在最终发布版数据集中只保留了其自答问卷的内容。发布数据集中用两个变量 selfrpt 和 proxyrpt 来标明受访者数据来自受访者的自答或代答问卷。

表 2 中列出了自答与代答问卷中不完全匹配的变量，我们分别给出其相应的变量名。研究者可根据具体的研究需要对其进行再处理。

表 2　2016 年成人自答和代答问卷不完全匹配问题信息列表

题号	自答问卷变量	代答问卷变量	不匹配情况
W1R	CFPS2016EDU	PW1R	PW1R 是代答人回答的受访者的最高学历；CFPS2016EDU 是综合各种信息之后，清理得到的受访者最高学历。推荐使用 CFPS2016EDU
GC103 GC103A	EGC103	PGC103A	自答问卷提问的是修正后的 2014 年调查时的主要工作是否为 2016 年调查时的主要工作；代答问卷提问的是未修正的 2014 年调查时的主要工作是否为 2016 年调查时的主要工作
G501	QG501	PG501	自答问卷提问的是与工作单位或雇主在合同中是否约定了合同期限；代答问卷提问的是过去 12 个月，受访者有几个月从事了这份工作
Z101	QZ101_S_n	PZ100_S_n	自答问卷提问的是其他参与回答问卷的家庭成员 n；代答问卷提问的是参与代答问卷的家庭成员 n
Z201	QZ201	PZ201	自答问卷提问的是受访者的理解能力；代答问卷提问的是代答者的理解能力
Z207	QZ207	PZ207	自答问卷提问的是受访者的智力水平；代答问卷提问的是代答者的智力水平
Z209	QZ209	PZ209	自答问卷提问的是受访者对调查的兴趣；代答问卷提问的是代答者对调查的兴趣
Z211	QZ211	PZ211	自答问卷提问的是受访者的可信度；代答问卷提问的是代答者的回答的可信度
Z5	QZ5	PZ5	自答问卷提问的是受访者急于结束调查的程度；代答问卷提问的是代答者急于结束调查的程度

5. 综合变量的添加

除了从问卷中直接生成的变量之外，CFPS 发布数据中还包括了部分项目团队人员基于问卷变量后期生成的综合变量。这些综合变量的基本情况如下：

1） 家庭收入系列变量（家庭经济库）

家庭收入包括总的家庭收入、人均家庭收入和具体分项收入（家庭工资性收入、经营性收入、转移性收入、财产性收入和其他收入）。其中工资性收入是指家庭成员从事农业受雇或非农受雇工作争取的税后工资、奖金和实物形式的福利。经营性收入是指家庭从事农林牧副渔业生产经营扣除成本后的净收入（包括自产自销部分），以及从事个体经营和开办私营企业获得的净利润。转移性收入是指家庭通过政府的转移支付（如养老金、补助、救济）和社会捐助获取的收入。财产性收入是指家庭通过投资或出租土地、房屋、生产资料等获得的收入。其他收入是指通过亲友的经济支持和赠予获取的收入。CFPS2016 在经营性收入、转移性收入、财产性收入和其他收入方面的设计与 CFPS2014 相同。为了方便用户与 CFPS2010 基线数据进行比较，我们还同时生成一版与基线可比的系列变量。相关变量列表见表 3。

2） 家庭支出系列变量（家庭经济库）

家庭支出包括家庭总支出以及分类别的四大类支出：居民消费性支出（包含食品、衣着、居住、家庭设备及日用品、交通通讯、文教娱乐、医疗保健、其他消费性支出），转移性支出（包括家庭对非同住亲友的经济支持、社会捐助以及重大事件中人情礼），保障性支出（包括家庭购买各类商业保险）和建房购房贷款支出。CFPS2016 家庭支出方面的设计与 CFPS2014 相同。相关变量列表见表 3。

以上两类变量在家庭经济库中的具体变量列表如下。

表 3　CFPS2016 家庭经济库综合变量

变量名	变量标签
家庭收入系列变量	
fWAGE_1	工资性收入（调整后）
fWAGE_2	工资性收入（与 2010 年可比）
foperate_1	经营性收入
foperate_2	经营性收入（与 2010 年可比）
ftransfer_1	转移性收入
ftransfer_2	转移性收入（与 2010 年可比）
fproperty_1	财产性收入
fproperty_2	财产性收入（与 2010 年可比）
FELSE_1	其他收入
FELSE_2	其他收入（与 2010 年可比）
FINCOME1	全部家庭纯收入
FINCOME2	全部家庭纯收入(与 2010 可比)
fincome1_per	人均家庭纯收入

fincome2_per	人均家庭纯收入(与 2010 可比)
fincome1_per_p	人均家庭纯收入分位数
fincome2_per_p	人均家庭纯收入分位数（与 2010 可比）
家庭支出系列变量	
PCE	居民消费性支出-加总
FOOD	食品支出-调整
DRESS	衣着鞋帽支出
HOUSE	居住支出-调整
DAILY	家庭设备及日用品支出-调整
MED	医疗保健支出
TRCO	交通通讯支出-调整
EEC	文教娱乐支出
OTHER	其他消费性支出
EPTRAN	转移性支出
EPWELF	福利性支出
MORTAGE	房贷支出
EXPENSE	家庭总支出
其它	
urban16	基于国家统计局资料的城乡分类

3） 个人主要工作工资收入变量插补（成人库）

在 CFPS2016 调查中，由于操作不当，对于 CFPS2014 到 CFPS2016 两次调查间主要工作没有发生变动的人群（n=4901），CFPS 未采集到其主要工作相关信息，造成了这一部分人群主要工作的工资收入缺失。在原始数据库中，这部分人群的 2016 年主要工作总收入(incomeb)为缺失（-8），他们的工作总收入（即主要工作收入和一般工作收入之和）也为缺失（-8）。我们根据缺失样本在 CFPS2016 中所采集到的其他个体特征（性别、年龄、教育水平等）以及他们在 CFPS2014 中采集到的主要工作收入，利用模型生成了 2016 年主要工作收入的插补值（incomeb_imp）。用户可以根据该主要工作收入插补值生成自己个人工作总收入（主要工作收入+一般工作收入）。这些样本的个人收入虽然缺失，但由于家庭收入是在家庭层面采集的，所以收入基本不受影响。

4） 认知水平（成人和少儿库）

CFPS2016 的认知测量与 CFPS2012 的设计基本相同，沿用了美国健康与退休调查（Health and Retirement Study）中的记忆题与数列题。其中记忆题包含 4 套难度类似的平行测试，受访者随机接受四套测试中的任意一套进行记忆评估。每套测试中有 10 个常见名词，访员将这 10 个名词清晰匀速地读出给受访者后，要求受访者即时按任意顺序回忆并说出他们能记得的名词，我们将这次记忆测试回答出的正确词汇个数称作受访者的即刻记忆得分

（immediate word recall）,记录在变量 IWR1（字词回忆：第一轮）中。如果受访者第一次没有正确回忆起任何词语，访员会再次读出 10 个名词，受访者有第二次机会再回忆一次，这次的分数记录在 IWR2 中（字词回忆：第二轮）。IWR（字词回忆：第一轮&第一轮）是综合了第一轮和第二轮测试的得分。当即时记忆测试结束之后，访员继续调查问卷中的其他内容。约在五分钟后，会让受访者尝试回忆之前听到的 10 个词汇，受访者在这次的记忆测试得分被称为其延时记忆得分（delayed word recall）,记录在变量 DWR（字词延时回忆）中。

数列题采用的是二阶段适应性测试方法，它的基本思想是第一阶段给所有人相同难度的试题，根据第一阶段测试的得分情况决定第二阶段测试的难度，得分较低的第二阶段将得到较难的测试，而得分较高的受访者在第二阶段将得到较容易的测试。具体的测试介绍可参考技术报告 CFPS-31：《中国家庭追踪调查 2012 年数列测试题》。2012 年在设计上存在一个缺陷，数列题在两道例题完成之后询问受访者是否清楚测试流程，只要受访者表达存在困惑或者不想回答，则跳过该测试。这样一道明确的筛选题导致数列题出现了大量的缺失，在中老年群体中缺失比例超过一半。因此在 CFPS2016 中，我们将这一套筛选题删除，无应答率显著降低。对于适应性的测试，传统的计分方法（按答对题数计算总分）并不适用，因为不同人群拿到的试题难度有系统性的差别。适应性测试的计分一般基于现代测量学理论的“项目反应理论”模型计算。由于这种计分方法较为专业，为了省去用户自行处理的麻烦，我们在发布数据库中直接提供了由 Rasch 模型计算出来的得分，该分数在 CFPS 2016 中的变量名为 NS_W。同时，我们还生成了变量 NS_WSE，用来表示与 NS_W 相应的标准误。

5） CESD 抑郁变量

CFPS2016 中采用的是 Center for Epidemiologic Studies Depression Scale (CES-D)这套量表来测试个人的抑郁水平。这套量表有多种形式，在 CFPS2012 中，我们使用的是包含 20 道题的 CESD20。但实地调查的反馈显示该套量表用在 CFPS 个人问卷中显得题量过多，受访者接受程度不高。于是在 CFPS2016 中我们调整设计，改用了该套量表的精简模式，将题量从 20 道题减少到 8 道题。同时为了能有效比对不同轮次间的抑郁分数，我们选择了面访人群中随机 1/5 的样本依然延用 CESD20，剩下 4/5 的样本使用 CESD8。基于这个设计，我们数据处理人员在后期将两套题目的分数进行了对等的操作，使用的方法是百分位数等化方法(equipercentile equating)，生成了可比的分数 CESD20sc（构建的 CESD20 总分）。这个分数保持了 CESD20 的打分区间，与 CFPS2012 中的 CESD20 量表得分也是可比的。除了综合变量分数 CESD20sc，我们也同时保留了原始的单题分数，用户也可以自行生成自己认为更

合适的对等分数。

6. 各类编码工作

1） 职业编码和行业编码

CFPS2016 采集了受访者的详细工作信息，涵盖了自家农业生产活动、农业打工、受雇、非农自雇以及家庭帮工。工作信息相关变量很多是文字信息，出于以下两点考虑，这些原始变量不在数据库中发布：1）涉及到与隐私相关的具体工作单位信息；2）文字所含信息对于多数分析者来说难以直接运用。因此我们组织工作人员对这些原始的信息进行职业和行业的编码，生成不包含隐私信息且较方便分析的数据。为了方便用户使用，我们在成人数据库当中生成变量行业编码（QGA4CODE），以及变量职业编码（QGA401CODE）。

2） 疾病和死亡编码

在 CFPS2016 调查中，发现在与上次调查间，共有 694 人由家庭成员汇报为去世状态，对于这些个人，我们询问了其死亡原因，然后由访员在现场对死亡原因进行编码。我们生成个体去世原因变量（TA401_A16_P）供用户使用。

CFPS2016 年成人和少儿问卷中，分别在健康模块询问了关于疾病的信息。在成人问卷当中，我们询问了关于慢性疾病的信息，并在后期处理过程中生成变量慢性疾病编码（QP402ACODE 和 QP402BCODE）。在儿童问卷中，我们询问了儿童的疾病情况，并生成变量过去 12 个月最严重疾病编码（PC5_CODE）和出生后最严重疾病编码（PC5_2010CODE）。

3） 地址编码和城乡状态

与往期数据库相同，CFPS2016 的地址信息给出了三级编码：省码、区县码和村居码，其中省码（provcd16）是国标码，用户可以知道是哪个具体省，但区县码(countyid16)和村居码(cid16)均为数据管理员后期分配的伪码。我们同时提供了按 2016 年国家统计局网站上定义的各样本所在村居的城乡性质（urban16）。

4） 其他编码

除了上述的行业编码、职业编码和地址编码之外，CFPS2016 还对职业期望（成人库中变量为 KS801CODE，少儿库中变量为 WD101CODE）、行政管理职务（成人库中变量为 QG1401CODE，少儿库中变量为 QG14）、高等教育院校（成人库中变量为

PS1CODE_COLLEGE）等信息进行了编码，以供研究使用。

三． 数据库简介

CFPS2016 全国追踪调查以 2010、2012 和 2014 年全国调查所界定出来的家庭为基础，发放的样本包括 2014 年完访的所有家庭以及 2010 年或 2012 年完访但 2014 年并未成功追踪的家庭[1]。CFPS2016 访问问卷包括家庭成员问卷、家庭经济问卷、成人问卷、以及少儿问卷，这四套问卷产生相应的四个问卷数据库，同时我们生成了个人层面的跨年核心变量库。CFPS2016 五个基础数据库的情况如表 4 所列。

表 4　CFPS2016 年各库基本状况

数据库	样本量	变量数
成人数据库	36892	1096
少儿数据库	8427	635
家庭关系数据库	58179	286
家庭经济数据库	14019	329
跨年核心变量库	70282	109

1. 跨年核心变量库：

跨年核心变量库是个人层面的数据库，它包含了自 2010 年 CFPS 基线调查以来所有进入 CFPS 样本的个人基本信息。样本中包括 57,155 名基线基因成员，追踪调查时新加入的 4875 名基因成员以及追踪调查时新加入的其他成员 8252 名，共 70,282 名个人。核心变量库中收录的变量可以分为三类：第一类是基线变量（time constant variables），如出生年、性别、民族，这些变量对于每个个体样本来说只有一个值。第二类是跨年变量（time varying variables），包括婚姻、收入（个人、家庭人均）、工作状态（在业、失业、退出劳动力市场）、户口状况（城乡）、居住地城乡状态、是否经济上是一家、是否物理地址上居住在一起、教育(是否在学、最高学历、上学/离校阶段)、迁移。这些变量每轮调查都有一个相应的变量，

[1] 去除那些在往期调查中已经确认的所有家庭成员已经死亡的家庭。

各轮之间数值可能不等。第三类变量是与访问过程相关的其他变量，包括访问状态、基因成员类型、死亡状态（死亡时间、死因编码）以及各轮权数。

由于追踪数据各年数值可能存在一定的不一致性，我们在确定跨年库数值时对某一轮的数据进行了年内数据的筛选(即当年的数据库中如果有后期创建的最佳变量时我们取该值)，但并没有进行跨年数据的一致性处理，因为跨年数据的不一致性问题比较复杂，除非在确定某一轮数据有误的情况下我们会将当年数据进行更新，其他无法确认的情况我们保留了各年的数值，用户使用时需要根据自己的研究需要进一步处理。

2. 成人库和少儿库：

与往年相比，CFPS2016 中成人和少儿库中共用模块变量的比例加大，两个数据库的相似程度比往年有了进一步提高。成人库包括往期调查界定出来的基因成员中 CFPS2016 调查时年龄处在 16 岁及以上的个人，以及 2016 年新增家庭成员中年龄处在 16 岁及以上的个人。少儿库包括往期调查界定出来的基因成员中 CFPS2016 调查时年龄处在 16 岁以下的个人，以及 2016 年新增家庭成员中年龄处在 16 岁以下的个人，其中 10 岁及以下的家庭成员只有家长的代答问卷，10 岁到 15 岁的家庭成员既有家长的代答问卷，也有个人的自答问卷。无论是成人库还是少儿库，问卷的实际回答人如果是受访者本人，则 selfrpt=1,表明使用了自答问卷；如果问卷的实际回答人是了解受访者情况的其他家人，则 proxyrpt=1，表明使用了代答问卷。在成人库中，代答问卷在以下两种情况下会启动，一是当家庭中有成员外出时，会先由原家庭成员完成一份代答问卷以捕捉外出个人的基础信息；二是当受访者本人由于身体的原因不适合自己回答问卷时(如老年失智症患者)，也将由其家人替他完成一份代答问卷。对于我们成功追访到的外出人员，他的原始数据中既有自答数据，也有代答数据，但在最终数据集中，我们只保留了这一部分样本的自答数据。

3. 家庭成员关系库：

家庭成员关系库以家庭成员为单位，包括 2010 年基因成员及之后调查年新增的家庭成员的配偶、父母及子女的基本信息。2016 年家庭成员关系库中包括来自 14763 个家庭的 58179 条个人样本。熟悉往年家庭关系库的 CFPS 用户需要注意， CFPS2016 的家庭库有一个重大调整，那就是我们将所有个人都只保留了一条观测，其中优先保留其在当前家庭的观测，如果当前家庭在家庭层面未完访，则保留个体在上一级家庭的观测。在以往的关系库中，同样的个人有可能出现在多条观测中，这是因为我们将另组家庭成员分别放在原家庭列表和另组

家庭列表中，并用是否在家（co_aXX_p=1 表示在家，0 表示离开原家庭，其中 XX 表示年份）来表明该个体在 2012 年经济上属于哪个家户。此种安排是为了更好的显示个体在不同轮次间的动态过程。详细情况可参见《2012 年家庭成员库的分解与家庭关系库的构建》。

4. 家庭经济库：

家庭经济库以家庭为单位，包括往期调查所界定出来的原生家庭以及在 2016 年调查时发现由家庭因婚姻变化、子女经济独立等原因所派生出来的新组家庭。在 2016 年家庭经济库的 14019 户中，有 1164 户为当年调查时所界定的另组家庭。访问方式为面访或电访。

在 CFPS2016 调查中，对于外出单元与原家庭的经济联系，存在原家庭和外出单元自己的双重界定。当二者的界定不一致时，对家庭经济问卷数据有一定的影响。在家庭经济问卷的开始，访员会根据之前家庭成员问卷的回答情况给受访者列出目前与该家庭有经济联系的每个个人，然后提醒受访者经济问卷的问题应考虑到该列表中的所有家庭成员。如前所述，当双重界定之间不一致时，有可能产生家庭成员在多个关联家庭中重复出现的状况。为了便于用户在后期对这些家庭的数据进行调整，我们在经济问卷中生成了一系列变量来指征与一个家庭有关联的家户号，以及与每一个关联家户之间是否有经济问卷指待家庭成员的重叠，该系列变量如下表 5。

表 5　家庭经济库中经济问卷关联家庭相关变量信息

变量名	变量标签	值标签
overlapfid1	经济问卷关联家户 1	
overlapfid2	经济问卷关联家户 2	
overlapfid3	经济问卷关联家户 3	
overlapfid1type	与关联家户 1 相关类型	1=fid16 与关联家户完全相等；2=fid16 与关联家户交叉，即 fid16 与关联家户有部分共同成员，但是不等；3=fid16 完全包含于关联家户；4=fid16 完全包含关联家户
overlapfid2type	与关联家户 2 相关类型	
overlapfid3type	与关联家户 3 相关类型	