China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-45

系列编辑：谢宇　责任编辑：赵逸文

中国家庭追踪调查
2022 年数据库介绍及数据清理报告

吴琼 孙妍 戴利红 甄祺 谷丽萍 张凯雯 吕萍

2024.10

中国家庭追踪调查 2022 年数据库介绍及数据清理报告

一．CFPS2022 总体介绍

1.1 CFPS2022 的访问对象

2022 年，北京大学中国社会科学调查中心实施“中国家庭追踪调查”（CFPS）项目的第七次全国调查。此轮追踪调查涵盖自基线以来所有基因成员所在的家庭及个人，即便他们上一轮次可能没有完成访问。

在家庭层面，我们共发放超过 1.95 万户家庭样本，加上在本轮调查中新生成的家庭样本，最终共涵盖超过 2.3 万户需要访问的家庭单元。家庭层面的访问包括家庭成员问卷（后续产生家庭成员关系库）及家庭经济问卷（后续产生家庭经济库）。其中家庭成员问卷界定了家庭中的各类成员状态、特征及相关之间的关系，是 CFPS 中每个家户调查的起始点。家庭成员问卷完访后，才会产生包括家庭经济问卷以及针对每一位符合访问资格的家庭成员的个人问卷。

1.2 执行总体情况

2022 年共有 698 名访员投入到 CFPS 的访问工作中，执行期于 2022 年 5 月开始，至 2022 年年底结束，约七成访问集中在 7 月份到 9 月份完成。每份问卷的访问年月信息可在发布数据集中查找[1]。由于疫情的影响，2022 年的访问形式沿用 2020 年的模式，以电访为优先。如果受访对象主动要求面访且符合面访条件的，访员再采用面访。在电访为主的背景下，为了提高访问成功率，我们制定了一系列的访问策略：在遵守当地公共卫生政策要求的前提下，制定区县内开展面访的访问策略，充分利用面访访员在实地的优势，提升访问成功率；访问系统全面升级，电话访问和面对面访问充分结合，访员可根据实际情况及受访者意愿随时变换访问形式，提升访问效率和完成率；在电访开始前，事先通过短信平台向受访者发放调查告知短信，以协助电话访员更好地联系受访者。

在大规模电访的背景下，CFPS2022 的家庭层面完访率与 CFPS2020 相比有所降低，截面应答率为 58%；但家庭层面跨轮应答率为 78%，较 CFPS2020 有小幅上升。家庭成员问卷完成

[1] 各数据集中访问年和月分别是 cyear 和 cmonth 变量。

后，共生成近 5.8 万份户内下级问卷（含家庭经济及各类个人问卷），下级问卷完访率为73.3%，其中总体访问负担最重、耗时最长的个人自答问卷的完访率为 65.5%。电访模式下访问单元由家庭户转为独立个体后，户内个人完访难度加大，具体表现为个人层面样本拒访比例及转代答的比例明显升高、个人自答问卷无电话的比例较高，其失联率在各问卷类型中居高。

1.2 问卷设计调整

CFPS 作为一项追踪调查，核心模块及其问卷问题在可能的条件下都尽量维持之前的设计。因此，2022 年问卷的总体设计延续了 2020 年的基本结构。但项目组结合搭载项目需求以及问卷设计的总体规划，对问卷内容进行了增删，其中问卷新增的主要内容如下：

1） 行为和精神状态模块：个人问卷中添加夫妻家务分工信息。

2） 主观态度模块：个人问卷中添加成长型/固定型思维。

3） 网络模块：个人问卷中添加数字金融服务的使用情况（机房限制类数据）。

4） 亲子互动模块：少儿家长代答问卷添加家长评测的 3-6 岁儿童的能力发展指数。

5） 学校基本情况模块：少儿家长代答问卷添加家校互动问题。

6） 住房模块：家庭经济问卷添加房屋出售情况、当前住房状况。

7） 中等收入群体：家庭经济问卷添加中等收入群体模块，测量其主观认知（机房限制类数据）。

除此之外，各问卷还存在一些细节上的调整，具体问卷内容可以通过项目网站的“文档中心”栏目下的“调查问卷”进行查询。在调查问卷页面，我们除了各年份的问卷之外，还提供了将历年问卷汇总在一起的文档《CFPS2010-2022 历年问卷内容汇总表》，方便用户查询问卷在不同轮次中的变化。

1.3 数据分级管理

为了最大程度保护受访者隐私信息，CFPS 采用变量分级共享机制。所有可能识别到个体的隐私变量被直接从共享数据集中删除，数据用户无法在共享数据集中识别到个人。在用

户可用的共享数据集中，项目组综合考虑数据安全、合作方要求、国内外调查界惯例等，将变量分成三个等级。绝大部分问卷变量被划归公开数据集，注册用户可在官方数据平台下载使用。除公开数据集之外，还有两级限制类数据。其中，一般限制性数据的安全等级稍高于公开数据集，用户提交额外申请说明研究目的，申请通过审核后方可在平台上下载使用。机房使用类限制数据是目前可供用户使用的最高安全级别的数据，该类数据只能在北京大学中国社会科学调查中心的机房中使用。详细信息可以在项目网站“数据中心”页面下的“限制数据”栏目查询。

北京大学中国社会科学调查中心对所有限制类数据的申请和使用都有详细记录。数据用户通过调查中心的官方渠道申请和使用数据是维护数据安全和保护用户自身科研成果的最佳方法。违规使用数据可能导致论文被撤稿、机房权限被永久取消。

二．CFPS2022 数据清理

CFPS2022 的数据清理涉及到以下主要环节。在清理过程中，我们并非按次序进行各环节的工作，而是根据需要并行开展多个步骤。由于不同的环节之间互相关联，每项清理步骤都包含多轮的迭代过程。

2.1 结果代码清理

样本层面清理主要依赖于问卷访问系统记录的结果代码和问卷数据本身。结果代码是每条样本在访问过程中及访问结束后呈现出的访问状态代码。根据 CFPS 访问系统的设计，所有样本的初始结果代码均为 0000，当样本达成完访状态时其结果代码应为 1001，而各种中间状态（如样本中断）以及其他最终访问状态（如联系不上）各有对应的代码呈现。

在访问过程中，结果代码的清理目标是每条完访样本均有唯一对应的初始样本或上级问卷，且同一家庭或个人样本的同类问卷样本最多有一条为有效完访问卷。在访问结束后，结果代码的清理目标是所有开启的问卷均有合理的最终结果代码。清理时，我们运用的主要原则包括以下几条：1）若同一问卷的原样本和备用样本均出现了 1001，需与执行团队沟通，确认唯一的保留条目；2）若个人或家庭经济问卷出现了 1001 而上级家庭成员问卷未出现 1001，需与执行团队确认原因并确保存在完访的上级问卷；3）若个人代答问卷出现了 1001

（即样本开启了个人代答且完访），则对应的个人自答问卷的样本结果代码需明确显示为“生成代答”；4）被判定为非有效问卷的样本视为“无效问卷”，删除其对应的问卷数据。

最终清理完成的结果代码库中，总观测条数为 92925 条，其中被判定为结果代码层面完访的观测 63460 条，包括家庭成员问卷库中的 12973 条家庭样本，家庭经济问卷 10740 条，个人自答问卷 23168 条，少儿家长代答问卷 6060 条，个人代答问卷 2287 条，第三方搭载问卷 8232 条。家庭代答问卷此次被包含于家庭成员问卷之内，不单独生成结果代码。非完访样本中有约 38%为受访者拒访导致访问无法完成，约 23%为受访者失联导致访问无法完成，约 12%为备份样本生成或访问流程中的操作造成的重复，此外还包含各种致使访问无法完成的情况，如受访者由于身体原因无法交流而生成代答、受访者去世等。

经过如上步骤后，访问状态被确认为完访的问卷还需经历更进一步的梳理才能纳入发布库，如在家庭成员库中，我们会删除不包含基因成员的家庭，家庭的观测数目可能会有所减少；在各库中我们会按一定的标准纳入一部分中断样本（参照 2.3），并将自答和代答样本进行合并。这些都使得结果代码库中完访样本数目与表 2 中发布库的数目存在一定程度上的差异。

2.2 调查过程中的问卷数据实时清理

项目组从 2014 年开始尝试在调查过程中对部分变量进行重点监控，一方面旨在对问卷数据中存疑的信息进行及时的确认或更新；另一方面它可以实时地反映访员在访问过程中的不规范行为，对访员进行提醒和干预。在 2022 年调查的数据实时清理中，我们使用微信小程序与访员互动，访员可以便捷登录，在安全的环境中及时查看项目组提供的数据清理反馈。

CFPS2022 实时数据清理的重点包括两大部分内容：一是经济库中涉及到金额的变量，尤其是在“万元”和“元”之间容易引起单位混淆的变量；二是自由文本输入时不完整的信息筛选，比如对职业、疾病等信息的记录。2022 年，我们在反馈系统中提交了 3596 条信息，其中提醒类 2500 条，主要是职业或疾病类文本变量，需要访员反馈的信息有 1096 条。在需要访员反馈的信息中，访员反馈率为 81.5%，确认需要修改的有 710 条。改动较多的变量有以下几类：1）经济库的数值改动：包括单位混淆（如万元与元混淆），快捷键误用（本应为 0 的数值误用选择题中表示“否”的 5），以及参考时间段混淆（如在月支出的题目中回答了

年支出）等。数值改动较多的变量包括 FM401、FQ5、FQ6、FR2、FT301、FT302、FT401N 等，我们根据访员反馈更新条目 150 条；2）个人库和少儿库中的职业和疾病类文本变量，除去提醒类反馈，还有需访员重新确认的内容，主要是文本框中只包括数字或者核心变量不规范等问题，最后根据访员反馈更新 556 条相关文本内容；3）个人库中的数值改动，包含金额以及时间相关的数值型变量，涉及到 QG12、QQ402 等变量，最后根据反馈确定更新条目 4 条。

实时清理还包括一些重点变量的监测、重点模块内部和大模块之间的逻辑跳转检查等。这些是为了在调查过程中及时发现问题后与设计团队沟通，并根据需要调整。

2.3 调查后整体清理

CFPS2022 调查的执行工作在 2022 年底整体结束之后，数据管理团队开始进行调查后期的数据清理工作。在样本层面，我们在结果代码清理的基础上，结合问卷库实施了进一步的样本编码确认工作，具体包含以下主要环节。1）确认各库的中断样本是否加入发布库。各库根据自身特点采用不同的纳入标准(如成员库为至少完成家庭成员问卷 A、B、C 三个模块；经济库为至少完成“家庭收入”相关模块；个人问卷和家长代答问卷为至少完成问卷的 50%)。成员库最终纳入 132 条中断样本家户（其中包含 620 个人），经济库纳入 92 条观测，个人库纳入个人自答问卷 558 条，少儿家长代答库纳入 20 条。中断样本在各问卷中均以 interrupt 变量来指征，中断样本的 interrupt 变量值取 1。由于中断样本只回答了问卷的一部分内容，这些样本数据的不少变量会呈现出“-8”（不适用），因为受访者可能在回答那一部分问题时已经终止该问卷的访问；2) 以家庭关系库为出发点，确定有效家庭样本和个人样本，删除中间过程中产生的无效样本或重复样本。延续历年追踪调查家户清理的原则，每位完成访问的基因成员在当年一定存在且只存在唯一的家庭归属；每个有效的家户样本一定包含至少一位基因成员（有关基因成员的解释，请参考项目网站上的“用户手册”）；3）根据家庭关系库界定出的有效家庭和个人样本，整理问卷内部和跨问卷的重复样本编码，决定家庭经济库、个人库、少儿家长代答库的发布样本，并根据需要调整样本编码。

在变量层面，我们在调查实时清理反馈的基础上，进一步对各模块内部的逻辑跳转、重点变量分布进行核查。对于部分指标，我们将其与其他来源的宏观指标进行对比分析。对于个人库来说，由于问卷存在自答和代答两种模式，我们在发布数据集中将二者整合，方便用户使用。自答是个人问卷数据的优选模式，如果同时存在个人自答和个人代答问卷，发布的

个人数据集会优先保留个人自答数据。两种情况下会产生代答问卷：一是在个人问卷访问阶段，访问对象由于身体原因无法完成问卷，在访员申请且访问督导审核同意的情况下，由熟悉访问对象的家人完成代答问卷。这种情况的代答申请和审核需要遵循严格的标准，确保访问对象确系由于身体原因无法完成时才启用代答。对于 15 岁及以下的受访者，可能同时存在少儿家长代答问卷和个人代答问卷的样本，我们优先保留内容更丰富的少儿家长代答数据。二是在家庭成员问卷的访问过程中，问卷回答人对于所有物理外出和经济独立的基因、核心人员（有关“物理外出”和“经济独立”以及人员类型的定义，请参考项目网站上的“用户手册”）完成代答问卷，此类设计是为了最大限度地收集外出人员的基本信息。家庭代答问卷并不影响外出人员个人问卷的生成，访员依然需要对外出人员进行个人问卷的采集，并尽可能完成他们的个人自答问卷。

整体清理环节的另外一部分重要工作跟元数据的整理有关，元数据包括变量名、变量标签和值标签。元数据整理的目标是让用户使用时更容易正确地理解数据的涵义。除了常规的标准化操作之外，对于追踪调查来说，元数据的整理还包括跨年、跨库的元数据尽量保持一致。

2.4 综合变量构建

除了访问时通过调查问卷直接采集的变量之外，CFPS 发布数据中还包括了项目组基于问卷原始变量后期生成的综合变量, 比如基于各种收入相关问题生成的家庭总收入，基于针对各类人群进行教育信息采集的变量生成的最高学历变量等等。CFPS 采用计算机化的访问系统，在问卷设计中不仅要结合本轮采集的其他变量，还需要参考往期数据的加载信息，问卷逻辑非常复杂。这样的设计方式使得访问的过程更加流畅，受访者也无需重复回答已经采集过的信息或与其他的特定情况无关的信息，但是对于用户来说，复杂的跳转逻辑意味着数据结构更加抽象。

故综合变量的存在旨在降低用户们使用相关信息的难度。综合变量的基本算法可以通过项目网站上“数据文档”页面的“综合变量查询表”进行查询。生成综合变量的基础变量同样也在发布库中，如果用户觉得项目组提供的算法不能满足他们具体的研究需求，也可以直接生成基于其他算法的变量。我们将在数据库介绍部分对每个数据集出现的综合变量进行进一步介绍。

2.5 文本编码

CFPS 问卷中不少问题的应答是自由文本形式的，并非数值或固定选项答案。典型的自由文本式回答的问题包括工作内容（职业）、工作单位（行业）、疾病等。原始的自由文本信息无法直接放在发布库中，一是出于数据安全的考虑，因为其中可能包含个人信息；二是对于绝大部分用户来说，原始的自由文本难以直接分析。CFPS 数据管理团队对于每类自由文本使用一定的编码规则，将其转化成可以放在发布数据集中的编码。具体编码信息可以通过技术报告《CFPS-41：中国家庭追踪调查文本编码技术报告》进行查询。编码之后各个数值代表的具体含义可以从项目网站上“数据文档”页面的 codebook 中查找。

职业和行业编码。 CFPS2022 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。我们对这些原始的信息进行编码后，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，我们在个人数据库中生成行业编码系列变量，包括实习工作行业（QGA4CODE）、第一份工作行业（KGD3CODE）、主要工作行业（QG302CODE）；职业编码系列变量，包括实习工作职业（QGA401CODE）、第一份工作职业（KGD4CODE）、主要工作职业（QG303CODE）、配偶/同伴职业系列变量（QEA203CODE、EEB4022_A_1CODE）、受访者 14 岁时父母职业（QV103CODE、QV203CODE）。

我们根据生成的职业编码，创建了职业威望系列变量：与实习工作职业编码 QGA401CODE 相关的三个职业威望变量（qga401code_isco、qga401code_isei、qga401code_siops）、与第一份工作职业编码 KGD4code 相关的三个变量（kgd4code_isco、kgd4code_isei、kgd4code_siops）、与主要工作职业编码 QG303code 相关的四个变量（qg303code_isco、qg303code_isei、qg303code_siops、qg303code_egp）、配偶/同伴职业和受访者 14 岁时父母职业对应的职业威望。有关这些职业威望的具体计算方法，用户可以参考技术报告《CFPS-10：中国家庭追踪调查 2010 年职业社会经济地位测量指标构建》。相关转换文件也可以在“数据文档”下的“文本编码”页面查询。

疾病和死亡编码。CFPS2022 调查在问卷的两个位置采集死亡原因，一是在家庭成员问卷中由成员回答人提供家庭成员中是否有人去世以及去世原因；二是在个人问卷中，对于初次确认死亡状态的配偶进行死亡原因的采集。访员在现场对死亡原因进行编码，死亡相关的信息除了体现在如上提及的家庭关系库以及个人库，还包含在跨年个人核心变量库中。

CFPS2022 在个人问卷中询问了关于慢性疾病的信息，并在后期处理过程中生成慢性疾

病编码变量（QP402Acode 和 QP402Bcode)，还询问了过去 12 个月最严重的疾病，生成编码（QP5CODE）。在少儿家长代答问卷中，我们询问了“儿童过去 12 个月最严重疾病”并在后期生成编码变量（WC5Ncode）；对于之前年份未进行相关数据采集的受访者，我们还会询问其出生后最严重疾病，并形成后期编码（WC5_2010code）。

CFPS2016 对“过去两周哪些身体不适”这一问题（QP302）进行了设计上的调整，由之前的选择题变成了文本开放题，一直延续至 2022 年。对于这个问题，我们于 2018 年增加了基于受访者所回答文本的后期编码，即变量 QP302CODE。2022 年我们依旧对此变量文本进行了处理，生成 QP302CODE 编码。

其他编码。除了上述的编码之外，CFPS2022 还对职业期望（个人库中变量为 QS801_B_2CODE，少儿家长代答库中变量为 WD101CODE）、行政管理职务（个人库中变量为 QG1401CODE）以及学校基本情况进行了编码，其中学校相关的编码涉及多个变量。我们采集了正在上学的受访者的就读学校，涵盖了小学、初中一直到博士阶段的院校名称。对于中专、大专及其以上的学校，我们采用和之前年份相同的编码方案形成了中等职业和高等教育的学校类型编码（QS1_B_1CODE）。除此之外，我们依据教育部发布的《中等职业学校专业目录》（2010 年修订版）生成了专业编码。在个人自答库有针对在读的职业初中学生的所学专业（QS401CODE）、在读的职业高中学生的专业（QS501_B_1CODE）、初中阶段所学专业（KW1002_B_1CODE）、高中阶段所学专业（KW1002_B_2CODE）进行的编码；在少儿家长代答库中，专业变量有初中在读学生所学专业（WS401CODE）与高中在读学生的专业（WS501CODE）。我们还依据教育部发布的《学位授予和人才培养学科目录(2011 年)》生成了学科编码，包括学校基本情况模块的大专在读学生所学专业的学科（QS701NCODE）和本、硕、博在读学生所学专业的学科（QS9NCODE），教育史模块的大专阶段主修专业的学科（KW1003_B_1CODE）、大学阶段主修专业的学科（KW1003_B_2CODE）、硕士阶段主修专业的学科（KW1003_B_3CODE）和博士阶段主修专业的学科（KW1003_B_4CODE）。

2.6 地址编码

与往年相同，CFPS2022 的地址信息给出了三级编码：省级（provcd）、区县级(countyid)和村居级(cid)，分别代表家庭或者个人在接受当期 CFPS 调查时的居住地址。CFPS 家庭层面的地址信息来自家庭成员问卷回答人，如果该信息存在缺失，我们也会尝试利用在家个人的个人问卷信息、往年的地址信息进行适量补充。CFPS 个人层面的地址信息主要来自个人

地址模块、EHC 地址模块，当该信息缺失时，我们也会根据家庭层面的地址（针对在家个人）以及离家单元地址（针对外出个人）进行适量补充。如果样本是在家个人，他们的地址取自家庭地址；如果样本是外出个人，他们的地址则来自于家庭成员问卷中的外出单元模块。由于外出单元模块中的地址为家人代答，信息不完整的现象较为常见；同时由于外出单元地址中村居层面不是结构化信息，这些外出样本的地址在村居编码(cid22)以及城乡属性(urban22)上存在系统性的缺失。代答样本中由于离家比例较高，再加上没有自答问卷所提供的 EHC 地址，在地址上的缺失情况相对于自答样本更为突出。城乡性质(urban22)采用 2022 年国家统计局网站上对村居的定义。对于部分村居样本无法准确进行编码的情况，我们通过地图、以往各期城乡属性对其村居的城乡性质进行了一定的补充。需要注意的是对于家庭代答的个体样本（proxytype=2），他们的个人地址信息来自于原家庭的家庭层面住址或外出地址模块，具体取决于家庭代答个体是否离家。

发布数据集中的省码是国标码，可以对应到具体省份，但区县和村居码均为 CFPS 项目编制的虚拟编码，不是标准代码，数值本身没有实质性含义，也无法与外部数据链接。需要使用更具体地址信息的用户请参考 CFPS 项目网站上“数据中心”下面的“限制数据”栏目。CFPS 的区县和村居的虚拟编码在编制过程中基本实施与国标码一一对应的关系，跨年间可以比较。当样本家庭所居住区域的国标码发生变动时，CFPS 数据集中对应的地址编码也会发生变动。需要用户们特别注意的是 CFPS 的具体区县信息属于限制类数据，请大家按照官方网站上“数据中心”栏目的“限制数据”页面相关信息进行规范操作，保护自己的科研成果。过去数年间，已经出现多次因为用户不了解数据使用规范而给自己带来负面影响的实例：包括文章被接受后，无法按照期刊的要求获得项目组的官方授权，影响发表进度；用户发表时不清楚数据共享规范，导致文章发表后需要添加勘误说明并存在撤稿风险等。知晓 CFPS 数据管理规定的用户如果违规，可能导致论文被撤稿以及调查中心机房使用权限的永久取消。

在地址相关部分，还涉及如下几个用户常见的疑问。一是虽然 CFPS 的 2010 年基线调查只涵盖 25 个省市，但在追踪调查中，人口的地区间流动会扩大地址所在的范围。从 2012 年追踪调查年开始，样本就不再局限于基线抽样时的 25 个省市和 160 多个区县。二是同一个家庭内部的个人地址可能不同，这是因为 CFPS 界定家庭成员时是以经济联系为基础，并不要求家庭中的个人都居住在同一个物理地址上，相关信息可以参考用户手册中“经济独立”相关内容。

2.7 加权

CFPS2022 权数包括家庭和个人层面权数。家庭层面权数针对所有包含基因成员且完成了经济问卷的家庭，只包含截面权数，不包含追踪权数，这是由于家庭内部基因成员的死亡、婚姻和迁移情况在不断变化，跨年数据之间缺乏类似于个人的可比性。个人层面权数针对 CFPS 定义的基因成员（包括基线界定的基因成员和后期追踪调查时新产生的基因成员），存在截面权数和追踪权数。

个人层面的截面权数针对 2022 年完成个人访问的基因成员，它的基本算法包括如下步骤：一、计算 2022 年个人基础权数，其中基线基因成员的基础权数是 2010 年所有基因成员的 2010 年事后分层权数），追踪调查时新产生基因成员的权数是其父亲或母亲的基线个人基础权数或父母二人个人基础权数的均值；二、计算 2022 年个体的流失权数，个人截面无回答权数是基础权数和流失权数的乘积；三、为避免权数波动过大，我们对权数进行了极值调整，此处我们使用基于权数分布的分位数权数截取方法（取分位数是 1%和 99%）对上海市的权数进行权数极值调整。其他地区权数波动较小，未做极值调整。四、对权数进行标准化操作，即对每个子总体（subpopulation），由总量权数除以该子总体内总量权数的均值得出标准化权数，进而得到最终的个人截面权数。我们在个人库和少儿家长代答库中发布了个人横截面调整权数（rswt_natcs22n）。

个人层面的面板权数针对 2010 年个人问卷（少儿问卷或者成人问卷）完访并且在 2022 年也完成了个人问卷（少儿问卷或个人问卷）的基因成员，它的计算方法同样是在 2022 年基础权数基础上再考虑 2022 年的流失情况，然后经过权数极值调整和权数的标准化，最终形成 2022 年的面板无回答调整权数，即最终的个人面板权数。我们在个人库和少儿家长代答库中发布了个人面板无回答调整权数（rswt_natpn1022n）。

CFPS2022 的家庭横截面权数是在 CFPS2022 个人横截面的基础设计权数的基础上得到的，即家庭横截面权数是家中所有基因成员的个人基础设计权数的均值，在此基础上我们考虑了家庭层面的样本流失，形成家庭层面的流失权数。然后经过权数极值调整和权数的标准化得到最终的家庭横截面权数。我们在家庭经济库中发布了家庭横截面无回答调整权数（fswt_natcs22n）。

跟往年一样，如果用户进行截面数据分析，可以使用家庭或个人层面的截面权数；如果用户分析个人在 2010-2022 年的变化，则可以使用我们提供的个人面板权数。如用户希望了解权数构建的更详细信息，可以阅读技术报告《CFPS-17 中国家庭追踪调查 2010 年基线调查权数计算》、《CFPS-44：中国家庭追踪调查 2020 年权数调整报告》

三．数据库简介

CFPS2022 主体问卷包括家庭成员问卷、家庭经济问卷、个人自答问卷、个人代答问卷以及少儿家长代答问卷。在家庭成员问卷访问过程中，我们会让家庭问卷回答人对于外出的个人（包括经济离家和物理离家）提供一份家庭代答问卷。调查问卷发布在项目网站“文档中心”下面的“调查问卷”栏目。

在数据发布时，我们将个人自答问卷数据和个人层面的代答问卷数据进行了整合，形成了针对 10 岁及以上个体的个人库。其余各个问卷（成员问卷、家庭经济问卷、少儿家长代答问卷）分别对应一个单独的数据集。每个数据集相应的 codebook 发布在项目网站“文档中心”下面的“数据文档”栏目，用户可以通过 codebook 查询变量中数值代表的含义。CFPS2022 数据库基本情况如表 2 所列。跨年个人样本综合变量库（crossyearid）将后续单独发布。

表 2 CFPS2022 年各库基本状况[2]

数据库	样本量	变量数
家庭关系库	47328[3]	298
家庭经济库	10726	331
个人库	27001	1283
少儿家长代答库	6228	311

3.1 家庭成员关系库(famconf)

2022 年家庭成员关系库中包括来自 12251 个家庭的 47328 条个人样本，其中基因成员占比 83.0%，核心成员 12.2%，非核心成员 4.8%。家庭关系库以 CFPS 界定的每个家庭成员

[2] 表 2 统计值基于的数据集版本分别如下：家庭关系库 1.0 版，个人库 1.0 版，少儿家长代答库 2.0 版，家庭经济库 1.0 版。
[3] 共涉及 12251 个家户。

为一行，家庭成员以 pid 标识，包括 2010 年基因成员及之后调查年新增的家庭成员，与每个家庭成员的关系人信息。关系人包括配偶(_s 系列变量)、父亲（_f 系列变量）、母亲（_m 系列变量）及子女（_c1-_c10 系列变量）的基本信息。关系人的 pid 变量为 pid_f、pid_m、pid _s、pid _c1-pid _c10。除了常规的 pid 之外，关系人的 pid 变量中包含多类特殊值，它们的涵义如下：1）-8 表示没有采集到对应关系人，可能造成此项的原因包括受访者不清楚关系人信息、不存在此关系人、或者当期没有继续提问而继承了往期的数据；2）77 表示存在关系人，但不在 CFPS 个人样本库中，不会采集其他信息。

根据 CFPS 个人样本追踪的规则，在家庭关系库中对每个成员定义了样本类型：基因成员指示变量是 gene，其中 1 表示是基因成员，0 表示否。基因成员在 CFPS 调查中永久追踪，产生个人问卷；核心成员指示变量为 coremember22，其中 1 表示是核心成员，在当前轮次实施追踪，产生个人问卷；如果下一轮次不符合核心成员的身份则不继续追踪。需要注意的是产生个人问卷并不意味着个人问卷的完访，因此并非所有的基因成员和核心成员都存在有效的完访个人问卷。没有成功完成个人访问的个体不会出现在个人库中，因此关系库的成员数目要大于完成个人层面问卷的成员数目。如上两个变量是综合变量，取值基于相应成员与现有的基因成员的关系，如果在清理中发现之前的关系界定错误而调整，这两个变量也会随之更新。在跨年使用数据时，应以当前年份最新的数据为准。

CFPS2022 调查时同处一个家庭的成员拥有同样的家户号 fid22，其中 co_a22_p 数值为 1 的代表本轮调查时界定出的该家庭的家庭成员，co_a22_p 数值为 0 的表示相应的个人曾经属于 fid22 这个家庭，但是当前轮次由于各种原因（经济独立、存疑、去世等）已不属于当前的 fid22 家庭，且需要追踪的成员在 2022 年调查中其所在的家户单元未成功完访。如果其离家单元成功完访，这些成员将会被赋予新的 fid22，用户可以通过往期的家户号（fid20，fid18 等）追溯他们的上级源头家庭。因此对于同一个观测，如果不同年份的家户号数值不同，说明该家户曾经发生过由于部分成员经济独立所导致的家户分裂。co_a22_p 表示成员与 fid22 之间的经济联系，也是 CFPS 界定家庭成员的基础。除此之外，tb6_a22_p 变量用来表示成员是否物理上居住在 fid22 所在的家庭住址。在 CFPS 的设计中，经济上同属一个家庭的成员可以居住于不同的地址。

CFPS2022 家庭成员问卷中关于离家原因（家庭成员问卷中的 A3 题目）的选项设计较往期有不同。和 CFPS2020 对比，CFPS2022 的 A3 题新增了 5 个选项，分别是选项是 7（“外出读书”）、8（“外出打工/工作”）、9（“分家”）、10（“婚嫁”）、11（“离婚”），事实上它们是

从以往的 77（“备注”）选项中进一步分离出来的占比较多的类型。这几个选项含义存在于清理后的 CFPS2020 发布数据中，只是对应数值有所不同。为了保持跨年可比，将新增选项值采用与 2020 年家庭关系库中系列变量 TB601_A20_*（离家原因）一样的数值对应关系。另外，对于“77”类别（77. 其他原因【请记录受访者原话】）所采集的开放性文本信息进行了归类编码，在原有分类基础上增加以下类别：出境、探亲、迁移、外出就医。

综合变量 CFPS2022_INTERV_*表示 2022 年个人问卷是否完访，它比往期变量值的分类更加细致：1 代表个人层面的问卷完访；0 代表产生了个人层面的问卷但未完访；-8 不适用，表示在该家庭中不需要产生个人层面的问卷，包含两类个人：1、不是该家庭成员，2 是家庭成员且为非核心成员。除此之外，CFPS2022 关系库与 CFPS2020 保持了一致的变量结构。关系库是成员库经由复杂的清理过程生成的，变量与问卷中的相关问题的对应关系不如其他问卷那样明显。为了便于用户了解变量来源，表 3 列出了这些变量的基本含义以及它们在问卷中对应的问题。

表 3　CFPS2022 家庭关系库变量说明

变量名	变量标签	家庭成员问卷问题编号	算法简要描述
FID_PROVCD22	2022 年家庭省级国标码		根据地址模块清理省级国标码
FID_COUNTYID22	2022 年家庭区县顺序码		根据地址模块清理后区县国标码再重新编码
FID_CID22	2022 年家庭村居顺序码		根据地址模块清理后村居国标码再重新编码
FID_URBAN22	2022 年家庭城乡分类（基于国统局）		家庭所在村居根据国统局的城乡类型进行村居属性的分类
SUBSAMPLE	是否在全国再抽样样本中		基于基线原家庭（fid_base）对应的抽样信息
SUBPOPULATION	抽样子总体		基于基线原家庭（fid_base）对应的抽样信息
GENETYPE22	2022 年基因类型		根据 2022 年家庭成员类型表、是否是基因成员重新编码
FidXX（如 fid22）	2010-2022 对应的家庭编码		历年的家庭归属
FAMILYSIZE22	家庭成员人数		汇总同一个 fid22 内部 co_a22_p=1 的人员总数

TB2_A_P	个人性别	BC2、E1、D105	成员问卷新采集的、个人问卷中采集的、及往年已有的性别信息的综合
TB1Y_A_P	个人出生年	BC3、E2、D104	成员问卷新采集的、个人问卷中采集的、及往年已有的出生年信息的综合
TB1M_A_P	个人出生月	BC3、E2、D104	成员问卷新采集的、个人问卷中采集的、及往年已有的出生月信息的综合
TB3_A22_P	个人婚姻	BC4、E3	成员问卷新采集的、个人问卷中采集的、及往年已有的婚姻信息的综合
TB4_A22_P	个人最高学历	BC5、E4	成员问卷新采集的、个人问卷中采集的、及往年已有的最高学历信息的综合
HUKOU_A22_P	个人户口	BC6、E5、D106	成员问卷新采集的、个人问卷中采集的、及往年已有的户口信息的综合
TB6_A22_P	个人是否居住在家	A2、A201	
CO_A22_P	个人是否与该家庭经济上是一家人	F102、B1	优先以离家人自己主观判断是否经济独立为准，其次以原生家庭的主观判断来界定
OUTPERS_R_WHERE22_P	离家人（个人）的居住区域	G1、H1	整合成员问卷中单人离家和多人离家的信息
TB602ACODE_A22_P	离家（个人）省国标码	G101 、H101	整合成员问卷中单人离家和多人离家省份信息
TB601_A22_P	离家（个人）原因	A105、A3	值（1-6）保留问卷原选项值，将问卷选项 7、8、9 10、11 调整成与往期一致，将 77 采集的文本信息进行归类，并重新编码。
TB602CCODE_A22_P	离家（个人）区县顺序码	G102 、H102	整合成员问卷中单人离家和多人离家区县信息，并重新编码。
OUTUNIT22	外出单元序号	F1	将原生家庭把离家人员划分在不同的单元，依次顺序编码
gene	是否是基因成员	“CFPS 家庭人员类型”表	个人是否是基因成员
COREMEMBER22	是否是核心成员	“CFPS 家庭人员类型”表	个人是否是调查当年的核心成员
CFPS2022_INTERV_P	个人本轮个人问卷是否完访		汇总各类个人问卷的完访情况

ALIVE_A22_P	个人是否健在	A3	成员问卷新采集的、及往年已有的死亡信息的综合。
TA4Y_A22_P	个人去世年份	A4	成员问卷新采集成员去世年份信息
TA4M_A22_P	个人去世月份	A4	成员问卷新采集成员去世月份信息
TA401_A22_P	个人去世原因	A401	成员问卷新采集成员去世原因信息
pid_*	父亲、母亲、配偶、10 个孩子的样本编码	C2、C3、C4、C5	2022 年新采集信息与历年家庭关系信息的整合
C105_A22_P	个人进入该家庭的原因	C105	
RTYPE_END22	家庭成员问卷中 rtype 的含义	见家庭成员问卷“CFPS 家庭人员类型”中的说明	
FID_BASE	基线家庭样本编码		样本所在 fid22 回溯到 2010 年基线调查时的源头家庭
PSU	基线家庭初级抽样单位		Fid_base 在 2010 年基线采样时所对应的 PSU
ADS1_22	是否搬家	ADS1	
KZ103_22	访问使用的主要语言	Z103	
INTERVIEWERID22	访员编码		执行层面信息
Interrupt	是否中断样本		问卷是否完成

3.2 家庭经济库(famecon)

家庭经济库以家庭为单位，fid22 为每个家庭的唯一标识符。如前所述，该家户在之前年份的家户号用 fid10-fid20 标识。对于跨年间家户未发生变动的家庭来说，其家户号维持不变；但如果跨年间家户发生分裂，被认定为依然在“原家庭”的那些家庭成员所在的家

户号依然不变，只有被认定为是“新家庭”的家庭成员所在的家户号才会发生改变。因此我们不能完全依据家户号来判断一个家庭是否发生结构上的变化。准确判断一个家户是否与上一轮调查完全一致要依赖于家庭关系库中的家庭成员构成。如果需要了解关于家户分裂的相关信息，用户可以参考“用户手册”中“家庭变迁”相关内容。

经济库中的样本包括往期调查所界定出来的原家庭以及在 2022 年调查时因婚姻变化、子女经济独立等原因所派生出来的新组家庭。访问方式为面访或电访（以 iwmode 标识），面访和电访的问卷内容基本相同。由于疫情原因，CFPS2022 家庭经济库电访比例为 96.4%。家庭经济库中的所有观测都来自于家庭关系库，但并非所有存在于家庭关系库的家户都一定会在经济库中有观测。少量家庭在完成关系库之后并未完成家庭经济问卷的回答。

在 CFPS2022 家庭经济库中，有两个变量与家庭规模相关，它们分别是 familysize22 和 fml_count。其中 familysize22 是调查结束后项目组生成的，它基于家庭关系库的信息，由关系库中属于同一个家庭的成员数确认（fid22 相同，且 co_a22_p=1）。Fml_count 是调查实时的加载变量，由调查在进行过程中由家庭关系问卷中直接加载过来，代表了访问时调查系统界定的家庭成员数目，这个数目是受访者在回答家庭经济库时的参考。二者在 95.9%的家户中都相同，剩余不一致的主要原因在于两个方面：一方面，如果原家庭中存在离家单元，当离家单元对自己的经济关系与原家庭认定不一致时，我们将以离家单元自身的界定为准。譬如原家庭回答人可能认为离家单元与自己有经济联系依然是一家人，但离家单元自身界定自己已经经济独立，变成了独立的家户，这时 familysize22 和 fml_count 可能会出现差异。另一方面的原因是当同一个人有可能被多个关联家庭认定为自家成员时，根据唯一归属和清理原则，只保留该个体在其中一个家庭中，这时另一个家庭的实际家庭规模就低于原始值。如下是项目组根据经济库中的基础变量生成的有关家庭收入、支出、资产相关的综合变量。

3.2.1 家庭收入

家庭收入综合变量包括总的家庭收入（fincome1）、人均家庭收入（fincome1_per）和具体分项收入（家庭工资性收入 fwage_1、经营性收入 foperate_1、转移性收入 ftransfer_1、财产性收入 fproperty_1 和其他收入 felse_1）[4]。其中工资性收入（fwage_1）是指家庭成员

[4] 本节中介绍的家庭收入相关的各类综合变量的基本算法可以从项目网站上 “数据文档”页面的“综合

从事农业受雇或非农受雇工作的税后工资、奖金和实物形式的福利。关于工资性收入，除了家庭经济库的数据之外，个人问卷中也采集了个体的工资信息，我们在综合变量生成的过程中引入了个人自答中的工资信息。我们将经济问卷所采集的工资性总收入与所有完成个人问卷的家庭成员的工资性收入的总和进行比较，最终的家庭总工资收入的取值为二者中的高值。同 CFPS2020 一样，我们在 CFPS 2022 工资性收入的计算中未引入个人代答数据，用户如果需要可以自行计算。

经营性收入（foperate_1）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。具体计算方法为农业净收入和私营企业、个体经营收入之和减去农业经营成本得到的结果，当收入低于成本时则此值置为 0，也即经营性收入不为负值。如果用户需要了解经营亏损的情况，可以依据问卷数据中的相关变量自行计算。

转移性收入（ftransfer_1）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。用户需要留意的是 CFPS 的转移性收入综合变量与国统局基于住户收支调查计算的转移性收入在口径上存在差异：国统局的口径中包括养老和退休金、社会救济和补助、政策性生产生活补贴、报销医疗费、住户之间赡养收入，而不包括住户之间实物馈赠和一次性的拆迁、土地征用补偿等收入；而 CFPS 中没有直接对医疗报销进行采集，且来自不同住家人的经济帮助也没有区分现金和实物。

财产性收入（fproperty_1）是指家庭通过投资、出租土地、房屋、生产资料等获得的收入。同样的，CFPS 计算的财产性收入综合变量与国统局也存在重要差异：国统局的财产性收入中包含了城市住宅租金折算值，而 CFPS 变量中不包含。CFPS2020 和 CFPS2022 的财产性收入中加入了金融投资收入。

其他收入（felse_1）是指通过亲友的经济支持和赠予获取的收入。这部分在国统局的计算中是转移性收入的一部分。

在生成家庭总收入 fincome1 变量时，我们采用了如下步骤：1）原始数据进行清理后，分别生成工资性收入、经营性收入、财产性收入、转移性收入、其他收入这五项收入。2）

变量查询表”进行查询。

将前一步生成的五个分项收入进行加总，将加总的数值与经济问卷受访者所回答的家庭年总收入（finc）进行比较，最终的家庭总收入综合变量取值为二者中的高值。

人均收入为总收入 fincome1 除以家庭人口数，注意此处的家庭人口数为经济问卷访问过程中由成员问卷加载过来的 fml_count，而非 familysize22。此处选择使用 fml_count 的逻辑是在家庭经济问卷数据采集时，受访者的回答是基于 fml_count 的对应家庭。这个变量定义了经济问卷访问过程中的家庭规模，与后期清理过后生成的家庭规模变量 familysize22 稍有不同。

用户在使用时需要注意作为综合变量的家庭收入（fincome1）与问卷中原有自报家庭总收入变量（finc）的差异。综合变量并非受访者直接提供，而是数据管理人员后期生成的。为避免用户误用，CFPS2022 的数据中不再单独生成 fincome2（与 2010 年可比）及其系列变量，用户如有需要，可根据“综合变量查询表”中的算法自行生成。

3.2.2 家庭支出

家庭支出综合变量包括家庭总支出（EXPENSE）以及分类别的四大类支出，它们分别是居民消费性支出 PCE（包含食品 FOOD、衣着 DRESS、居住 HOUSE、家庭设备及日用品 DAILY、交通通讯 TRCO、文教娱乐 EEC、医疗保健 MED、其他消费性支出 OTHER），转移性支出 EPTRAN（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 EPWELF（包括家庭购买各类商业保险），建房购房贷款支出 MORTAGE。CFPS2022 家庭支出方面的设计与 CFPS2020 相同。各支出变量的算法均可参照项目网站上的“综合变量查询表”。

与国家统计局住户收支调查相关数据口径相比，CFPS 在各类支出上也与该口径存在着差别。比如国统局的算法中自有住房折算租金会被算入居住支出内，交通通讯设备的购买和维护费会被算入交通通讯支出；而 CFPS 的支出算法未纳入折算租金，交通通讯设备相关支出被计算在生活用品和服务里支出当中，交通通讯支出仅包括每月本地交通费、邮电通讯费换算出的年支出费用。

家庭总支出的算法类似于家庭总收入。一方面我们可以通过分项加总的方式得出家庭总支出（消费性支出 PCE、转移性支出 EPTRAN 、保障性支出 EPWELF 、建房购房贷款支出 MORTAGE），另一方面受访者在经济问卷的最后需要给出家庭总支出（fexp）。最终的家庭

总支出综合变量以加总所得的支出为主，当分项信息中受访者无法给出具体数值或者分项加总的数值小于 100 且自报总支出大于 100 时，家庭总支出综合变量的取值来自于自报总支出。

3.2.3 家庭资产

家庭资产包括家庭总体净资产（total_asset）和分类别的各项资产，其中包括住房净资产（houseasset_net）、土地资产（land_asset）、生产性固定资产（fixed_asset）、金融资产（finance_asset）、耐用消费品（durables_asset）和房贷外金融负债（nonhousing_debts）。其中住房净资产只针对家庭成员完全或部分拥有产权时才进行计算。具体而言，住房净资产为现住房价值（resivalue）和其他房产的总价值（otherhousevalue）减去房贷（house_debts）所得，土地资产为农业经营收入的估算产物（具体估算方式与往年相同，可参考项目网站上“数据文档”页面的“综合变量查询表”），生产性固定资产为经营资产（company）和农用器械价值（agrimachine）之和，金融资产为存款（savings）、金融产品（financial_product）和他人欠自家款项（debit_other）之和。房贷外金融负债则包含了来自银行、亲友及其他机构的借贷总额。最终家庭的净资产（total_asset）等于各项资产加总减去各项负债加总。

3.2 个人库(person)

CFPS2022 与 CFPS2020 的设计一致，个人库包括所有 10 岁及以上个人的问卷数据，个人样本由 pid 唯一标识。Pid 在跨年跨问卷的数据集中保持不变，个人层面的不同数据集都可以通过 pid 来进行链接。个人库样本数目小于关系库中 10 岁及以上家庭成员的数目，因为并非所有家庭成员都完成了个人问卷, 部分家庭成员由于联系不上、拒访等原因没有完成个人问卷。除了个人标识符 pid 之外，个人库还给出了个人所在的家户号 fid22，如果需要将个人问卷与家庭问卷进行链接，则可以通过 fid22 作为链接变量进行跨库链接。

个人问卷的跳转逻辑较为复杂，而对跳转逻辑的准确理解是我们使用 CFPS 数据的基础，尤其是正确处理缺失数据（如“-8”代表的“不适用”），更需要依赖于对跳转逻辑的深刻理解。用户可以通过观看官方数据平台上“培训视频”栏目下的“读懂 CFPS 问卷”来进一步了解，还可以借助项目网站上的往年 CFPS 问卷的“逻辑流程图”（“文档中心” →“数据文档”）来整理思路。

CFPS2022 个人库包含个人自答、个人代答（针对因为身体原因无法完成自答）和家庭代答（针对离家单元中的个人，一般由成员问卷的回答人统一代答）的样本，访问模式为面

访和电访（用 self_iwmode 和 proxy_iwmode 进行标识）。由于疫情原因，CFPS2022 个人库电访比例为 96.1%。个人自答的面访和电访的区别主要在于认知模块，面访中包括识字和数学这两个认知测试，但电访中没有；访员观察模块也是只适用于面访受访者。除此之外，其他问题在面访和电访中已经统一，个人自答的样本使用 selfrpt 变量标识，当 selfrpt＝ 1 时，该样本为个人自答样本。

在个人库中，我们用变量 proxytype 标识出该样本是来自于家庭代答还是个人代答，注意家庭代答的样本可能来自上一级家户的家庭成员，而个人代答一般由来自当前家户的家庭成员代答，可通过变量 respc1pid 来查看代答人的样本编码，respc1pid 取值为 77 是指代答人不在从家庭成员加载的列表中，取值为-8 表示按照代答人选择的序号并未在列表中找到对应的 pid。从问卷内容上，家庭代答和个人代答的问卷内容完全相同；但代答问卷和自答问卷的差别较大，代答问卷从内容上到具体问题上均为自答问卷的简版。在整合自答和代答数据的过程中，我们对自答和代答问卷的问题进行了比对，如果提问方式上没有任何差别，我们则统一了变量名（以自答变量名为准），作为同一个变量进行处理；但如果提问方式上有差异，我们则保持了自答和代答各自的变量。

需要特殊说明的是，个人库EHC-RESI模块中，由于系统访问时的操作原因，可能会存在变量缺失或者冗余的情况，用户可在使用时按需清理。为了便于用户使用，我们对原始数据集中的性别和年龄变量进行了整合。原始数据集中包含GENDER_PRE和GENDER_UPDATE变量，分别代表基于往期数据或家庭关系库的加载变量和通过访问更新后的变量。在此基础上，我们生成了信息最为完整的GENDER变量，既包括清理的内容，也包括通过跨年数据，以及关系库补充的内容。年龄相关变量包括AGE、IBIRTHY、IBIRTHY_UPDATE三个变量，分别为通过出生年份计算的年龄、加载出生年份信息，以及更新后的出生年份信息。对于年龄的缺失，同样会根据跨年库和关系库进行填补，均补充到AGE变量中，AGE的信息最为完整。

为了便于用户使用，个人库添加了如下综合变量。

3.2.1 受雇工作的工资性收入 emp_income

个人库中的 emp_income 变量的基础是系统自动生成的，基本算法是 incomeA（一般工作的总工资性收入）+incomeB（主要工作的工资性收入），也即 emp_income 变量主要反映个体过去 12 个月（持续到访问最近一年的工作）的工资性收入。需要注意的是，

数据集中的 emp_income 变量存在大量的“-8”（不适用）[5]，这是因为问卷只针对受雇类型的工作进行“工资性收入”的相关提问，如果一般工作或主要工作并非受雇（譬如从事个体/私营经济/其它自雇），则不对工资性收入进行提问。用户可以通过 lastyjob（工作是否延续到最近一年指示变量）以及 jobclass（工作类型）等系列变量对相关问题的跳转进行确认。注意此处的 emp_income 与 2018 年数据集中 income 变量口径一致，因为很多用户对 income 变量存在理解上的误区，我们从 CFPS2020 开始将其变量名进行了更新，后续也会对之前年份的数据集进行相应更新。

我们针对系统自动生成的 emp_income 变量进行了如下修正：当原始变量为拒绝回答或不知道（-1 或-2）时，问卷设计进行了展开式提问，我们生成了相应的收入区间变量，一般工作为 GC2051_MIN_A_N,GC2051_MAX_A_N 系列，主要工作为 QG1201_MIN 和 QG1201_MAX， 并利用区间的两个端点（最大值、最小值）取均值的方法来获得估计值。如果最终估计的区间在两端，也即比最小值还小，或是比最大值还大，我们则采用最小值的二分之一或最大值的两倍取估计值。

关于问卷中工作总收入的校验（QG1202），IncomeA、IncomeB 二者如果只有其中一项有值，可以不用进行题目 QG1202 工作总收入的校验。因此，我们对原始的 QG1202 数值进行了修正，如果只有一项有值并且还提问了 QG1202，那么 QG1202 更新为-8。同时我们也对问卷进行了更新，将“QG1202 工作总收入校验”这道题目前的跳转条件“若 Income>0”更新为了“若 IncomeA 和 IncomeB 同时有值”的情况下，再继续提问 QG1202。

3.2.2 认知水平

由于认知测试只针对面访样本，接受电话访问的个人没有相关数据，因此 CFPS2022 中完成了认知测试的样本量和 CFPS2020 一样很小。CFPS2022 认知测试从设计上沿袭了 CFPS2014 的问卷，包括识字测试和数学测试两部分。识字测试和数学测试的原始设计是按照受访者的教育水平采用不同的起点来进行测试，以提高测试效率，但在基线调查时，我们发现不少受访者在起点问题上就选择“不知道”或者打错。为了更准确地估计受访者的认知水平，我们从 CFPS2014 开始，允许受访者在较高起点的首道题答错后，降到低一级起点再尝试。我们按照每答对一道题记一分的方式计算受访者在识字和数学方面的总分，由变量 WORDTEST22_SC2 和 MATHTEST22_SC2 表示。同时，我们另外生成了一套假设起点固定

[5] 有关“-8”（不适用）的解释，请查看官方数据平台“文档下载”页面下的培训视频：“如何读懂 CFPS 问卷”。

时受访者有可能得到的分数，以确保其与 CFPS2010 认知分数的可比性，由变量 WORDTEST22 和 MATHTEST22 表示。两种算法的得分相关度非常高，其中识字测试的两套得分相关系数接近 1，数学测试的相关系数也超过 0.98。如果用户使用 2014 年、2018 年或 2022 年认知测试数据时，我们推荐使用_SC2 系列变量，但如果分析中包含有 2010 年数据时，我们推荐使用与 2010 年可比变量，也即 WORDTEST22 和 MATHTEST22。

3.2.3 CESD 抑郁变量

CFPS2022 中采用 Center for Epidemiologic Studies Depression Scale (CES-D)量表来测试个人的抑郁水平。与 CFPS2020 的设计基本相同，生成了精简的 8 道题版本的 CES-D8。CES-D8 的得分是将两道反向提问的问题变量进行正向调整后，和其他六道题一起共同加总所得。在缺失值的处理上，由于总体的缺失比例较低（0.36%），我们采用了成列删除的方式，只针对那些八道题均不缺失的样本进行了综合变量的计算。

CESD8 代表的是基于 8 道题版本的分数，用户可以自行根据研究需要对缺失部分变量的 CESD 数据进行相应处理。2022 数据集中保留了原始的单题分数，用户也可以自行生成更合适的分数。还需要提醒用户的是 CES-D 原始量表中单题的分数取值是在 0-3 之间，而 CFPS 数据集中的单题取值是 1-4 之间，这使得 CFPS 中 CES-D 的总分区间跟一些文献中不一致，大家注意进行简单转化即可[6]。使用 CFPS2016 及 CFPS2012 数据的用户可能会发现，我们曾经使用过 20 道题版本的 CES-D20 量表。为此，我们的发布数据集同时提供了 CESD20sc 变量，代表基于这八道题构建出来的与 20 道题版本对等的分数。用户可以参考技术报告《CFPS-35 中国家庭追踪调查 2016 年数据库介绍及数据清理报告》了解 CESD20sc 的构建方法。

3.2.4 教育

教育模块的跳转非常复杂，我们综合加载数据以及不同模块的信息，生成了四个教育系列综合变量：受访者已完成的最高学历 CFPS2022EDU、受访者离校/上学阶段 CFPS2022SCH、受访者已完成的受教育年限 CFPS2022EDUY 和经插补之后的受访者已完成的受教育年限 CFPS2022EDUY_IM。

在生成教育综合变量时，我们会先根据问卷的跳转情况，将跳转到不同模块的人群进行划分，以区分每一种在读/离校的人群。针对个人自答，我们会结合问卷中 C 部分教育的 QC3、

[6] 如针对 CES-D20，原始量表的分数区间为 0-60，CFPS 数据集中为 20-80，我们只需将后者减去 20 可得与前者直接可比的区分。

QC5 和 W 部分教育史的 KW01、WEDU 以及 EDU_LAST、R1_LAST 计算受访者已完成的最高学历和离校/上学阶段；针对少儿家长代答、个人代答和家庭代答，这几个没有教育史内容的版块，我们会根据 EDU_LAST、R1_LAST 和 QC3、QC5 计算受访者已完成的最高学历和离校/上学阶段。初步整理完最高学历和离校/上学阶段之后，我们会对这两个变量之间的相互关系进行一轮逻辑校验和修正，原则上来说，离校/上学阶段应该等于或高于最高学历。此外，我们还会结合受访者历年的教育综合变量对 2022 年的教育数据进行填补和修正，尽量在跨年之间达到逻辑上合理。

在以上两个变量清理的基础上，我们再计算受访者已完成的受教育年限，具体步骤如下。（1）当离校/上学阶段等于最高学历时，我们将最高学历转换为其相对应的受教育年限。（2）当离校/上学阶段不等于最高学历时，受教育年限变量则会依据最高学历、离校/上学阶段这两个变量为基础进行两种算法的估计，然后在两种算法中取高值，生成受教育年限综合变量。当以最高学历为基础时，受教育年限等于最高学历转化成的受教育年限；当以离校/上学阶段为依据时，受教育年限等于离校/上学阶段的前一个阶段所对应的受教育年限加上后续未完结教育阶段的已完成年限。

若受访者未完结阶段的已读年数缺失，但最高学历和离校上学阶段不缺失，按照上述方式生成的受教育年限可能缺失。我们会使用 hot deck 方法对该变量的取值进行插补，生成插补版的教育年限（CFPS2022EDUY_IM），具体是按照排序后的个人 ID（pid）找到上一位与该受访者最高学历阶段相同且同样没有完成该阶段、但是该阶段已读年数不缺失的受访者，取其已读年数补充。CFPS2022 教育变量主要有 4 个数据源，分别是个人自答、父母代答、个人代答和家庭代答，同一个受访者根据不同的来源可能会算出不同的教育变量值，所以我们也对这四个来源里出现了两次及以上的受访者的教育变量选择了更合理的教育值。除此之外，家庭成员关系库中还有关于最高学历的相关内容，用户也可以考虑根据关系库的学历内容进行适当填充。有关教育变量清理的更详细内容请参考技术报告《CFPS-36：中国家庭追踪调查 2010 年教育程度相关变量清理与评估》。

此外，用户如果希望自己通过原始变量来进行整理或者核对，需要注意问卷中涉及教育阶段的原始值和项目组最终发布的教育综合变量取值有所不同。为了与历年的教育综合变量可比，我们在清理过程中进行了数值转换，譬如在问卷中“小学”教育程度是“3”，我们在算综合变量时会把其转换为“2”。数据集中各个数值的具体含义可以通过 2022 年的 codebook

来查询[7]。

3.2.5 当前工作状态 employ

个人自答的 employ 变量是系统根据受访者在【GB 当前工作状态确认】模块回答的信息自动生成的，基本算法如下：（1）如果受访者①过去一周至少工作了 1 个小时、或②能够在确定的时间或者 6 个月以内回到原来的工作岗位、或③从事个体经营活动，但是目前处于生意淡季，等过一段时间还会继续经营、或④从事农业方面的工作但是目前处于农闲季节，则判定受访者有工作，employ=1；（2）如果受访者过去一个月找过工作，且如有工作机会能在两周内开始工作，则判定受访者失业，employ=0；(3)如果受访者过去一个月没有找过工作，或如有工作机会不能在两周内开始工作，则判定退出劳动力市场，employ=3；(4)其他情况，employ=-8。

个人代答和家庭代答问卷并没有采用整套【GB 当前工作状态确认】模块，但是我们根据GB1的信息补充填补了部分代答受访者的当前工作状态。如果受访者过去一周至少工作了1个小时，则判定受访者有工作，employ=1；其他情况，由于信息不足，无法判断受访者是在业、失业还是退出劳动力市场，故employ赋值为“-10 无法判断”,假如该问题的原值为“-8”，则employ=“-8”。

有关受访者就业或失业状态的设计，用户可以参考项目网站上“用户手册”中“就业状态界定”章节相关内容。从统计口径上来说，CFPS定义的失业与国家统计局调查失业率之间存在着系统性差异，感兴趣的用户可以参考CFPS公众号（ISSS_CFPS）推送文章：《CFPS小课堂|CFPS失业率的计算方法》。

3.3 少儿家长代答库（childproxy）

少儿家长代答库包含所有 0-15 岁少儿的家长代答数据。少儿家长代答库的观测单位为少儿，每个被访问到的孩子为一行，以 pid 标识。代答人一般为最熟悉孩子的家庭成员，可能是父母，还可能是其他家庭成员。代答人的样本编号以 respc1pid 来标识，和个人库中的代答样本一样，respc1pid 取值为 77 是指代答人不在从家庭成员加载的列表中。与 CFPS2020 的设置相同，CFPS2022 的少儿家长代答库只有家长代答数据，没有 10-15 岁少儿的自答部分，自答数据在个人库中。

少儿家长代答库包括少儿家长代答样本，以及少量来自家庭代答的 0-15 岁少儿样本。

[7] Codebook 的位置在项目网站上“文档中心”→ “数据文档”→“Codebook 及 SAS 值标签”。

数据库中同样用标识符变量 proxytype 对少儿家长代答（proxytype=3）和家庭代答（proxytype=2）进行区分，家庭代答和个人代答问卷设计类似，因此家长代答问卷的内容比家庭代答丰富。访问模式同样也有电面访两种，可使用变量 IWmode 进行面访(IWmode=1)和电访(IWmode=2)的识别。

相关链接

- 北京大学中国社会科学调查中心网站

http://www.isss.pku.edu.cn/

- CFPS 官方网站

http://www.isss.pku.edu.cn/cfps/

- CFPS 数据平台

http://www.isss.pku.edu.cn/cfps/download/login

- CFPS 限制类数据管理规定

https://www.isss.pku.edu.cn/cfps/sjzx/xzsj/index.htm

- CFPS 论文发表注意事项

https://www.isss.pku.edu.cn/cfps/cjwt/fbxg/index.htm

CFPS 项目公众号（ISSS_CFPS）