数据清洗过程详述

目录

1 PFP

要研究 CHARLS 数据库中个体的衰弱情况，需要选取一个合适的量表，再从 CHARLS 中抽取量表中对应的指标，本文选取 PFP 量表。

1.1 PFP 指标

PFP 对应抽取的指标有 13 个，分别是：

（1）体重测量（Weight Measurement）；

（2）身高测量（Height Measurement）；

（3）左手握力第一次测量（Left Hand-1, kg）；

（4）右手握力第一次测量（Right Hand-1, kg）；

（5）左手握力第二次测量（Left Hand-2, kg）；

（6）右手握力第二次测量（Right Hand-2, kg）；

（7）步速第一次测量（Walking Speed Time-1）；

（8）步速第二次测量（Repeat the Measurement）；

（9）持续高消耗性体力活动至少十分钟（Do Vigorous Activities At Least 10 Minutes Continuously）；

（10）持续中消耗性体力活动至少十分钟（Do Moderate Activities At Least 10 Minutes Continuously）；

（11）持续低消耗性体力活动至少十分钟（Walking At Least 10 Minutes Continuously）；

（12）感到生活无法继续（Could Not Get Going）；

（13）感觉自己做的一切都是徒劳的（Felt Everything I Did Was An Effort）。

1.2 界值的选取

界值的选取主要参考了以往使用使用 CHARLS 数据进行衰弱研究的文献[1]。

表 1.2-1 握力减退界值表

性别	$BMI(kg/m^2)$	握力减退临界值kg
女性	≤ 20.0	15.0
	$22.0-22.1$	17.5
	$22.1-24.8$	17.5
	≥ 24.8	20.0
男性	≤ 20.6	25.2
	$20.6-23.2$	28.5
	$23.2-25.9$	30.0
	≤ 25.9	30.0

表 1.2-2 步速减慢界值表

性别	身高（cm）	步速减慢临界值
女性	≤ 151	0.36
	> 151	0.43
男性	≤ 163	0.45
	> 163	0.48

1.3 各项指标的判定

（1）**体重减轻(Weight Loss)**：相较上次调查体重减轻≥5 公斤或身体质量指数$BMI \leq 18.5\ kg/m^2$，则体重减轻条目取值为 1，否则为 0；

（2）**握力减退(Weakness)**：惯用手的最大握力（无惯用手则用两只手的均值），以站姿测试两次，取每次两次均值。因惯用手做过手术或者发生任何肿胀、发炎、严重疼痛受伤等原因无法进行测试者则握力值缺失。当握力低于正常值（低于加权人群分布 20%，经性别和 BMI 的四分位数调整）。当握力取值低于临界值则握力减退条目取值为 1，否则为 0；

（3）**步速减慢(Slowness)**：测试短距离(2.5 米)步行速度，测两次取平均值。如果因为手术、外伤等原因无法测试则该条目缺失。当步速低于正常值（低于加权人群分布 20%，调整性别和身高)。具体划分界值如表 1.2-2 所示，当步速取值低于临界值则步速减慢条目取值为 1，否则为 0：

（4）**体力活动水平减弱(Low Energy Expenditure)**：如果受访者自报通常一周内连续行走的时间低于 10 分钟，则体力活动水平减弱条目取值为 1，否则为 0：

（5）**自报疲劳(Exhaustion)**：使用抑郁量表(CES-D-10)中的“我觉得做任何事都很费劲”以及“我觉得我无法继续我的生活”两个条目进行测量，研究对象回答很少或者根本没有(<1 天)”、“不太多（1-2 天)”、“有时或者说有一半的时间(3-4 天)”、“大多数的时间(5-7 天)”时依次取值为 0-3，得分范围为 0-6，当得分≥3 则自报疲劳条目取值为 1，否则为 0。

2 MMSE

要研究 CHARLS 数据库中个体的认知障碍情况，需要选取一个合适的量表，再从 CHARLS 中抽取量表中对应的指标，本文选取 MMSE 量表。

2.1 MMSE 指标

MMSE 对应抽取的指标有 30 个，分别是：

（1）现在是哪年（Checking Year）；

（2）现在是哪月（Checking Month）；

（3）现在是哪日（Checking Date）；

（4）今天是星期几（Checking Day of Week）；

（5）现在什么季节（Checking Season）；

（6）这里是哪个省份（Checking State）；
（7）这里是哪个城市（Checking County）；
（8）这里是哪个区（Checking City）；
（9）这里是哪一层楼（Checking Floor）；
（10）这个地方叫什么（Checking Address）；
（11）重复一次：皮球（Repeated Time 1: Ball）；
（12）重复一次：国旗（Repeated Time 1: Flag）；
（13）重复一次：树木（Repeated Time 1: Tree）；
（14）100-7*1（Specific Result from 100-7）；
（15）100-7*2（Specific Result from dc014_w4_2-7）；
（16）100-7*3（Specific Result from dc014_w4_3-7）；
（17）100-7*4（Specific Result from dc014_w4_4-7）；
（18）100-7*5（Specific Result from dc014_w4_5-7）；
（19）延时回忆：皮球 国旗 树木（Delayed Recall: Ball Flag Tree）；
（20）延时回忆：皮球 国旗 树木（测量 1）（Delayed Recall: Ball Flag Tree-1）；
（21）延时回忆：皮球 国旗 树木（测量 2）（Delayed Recall: Ball Flag Tree-2）；
（22）请问这是什么？（手表）（Watch Correct）；
（23）请问这是什么？（铅笔）（Pencil Correct）；
（24）请重复“四十四只石狮子”（Repeat Correct）；
（25）请受访者用右手拿纸（Hand Correct）；
（26）请受访者折纸（Folds Correct）；
（27）将折好的纸放在腿上（Leg Correct）；
（28）按照这张纸上内容做（闭眼）（Close Your Eye）；
（29）请受访者写正确有意义的句子（Sentence Correct）；
（30）模仿画图（Draw Correct）。

2.2 各项指标的判定

（1）**方向感（Orientation）**：现在是哪年、现在是哪月、现在是哪日、今天是星期几、现在什么季节、这里是哪个省份、这里是哪个城市、这里是哪个区、这里是哪一层楼、这个地方叫什么，共十项指标，每回答正确一个得一分；

（2）**不知道叫什么中文名了（Registration）**：重复一次：皮球、重复一次：国旗、重复一次：树木，共三项指标，每重复正确一个得一分；

（3）**注意力和计算能力（Attention and Calculation）**：计算 100-7*1、100-7*2、100-7*3、100-7*4、100-7*5，共五项指标，每回答正确一个得一分；

（4）**记忆能力（Recall）**：延时回忆：皮球 国旗 树木、延时回忆：皮球 国旗 树木（测量 1）、延时回忆：皮球 国旗 树木（测量 2），共三项指标，每回答正确一个得一分；

（5）语言与实践（Language and Praxis）：请问这是什么？（手表）、请问这是什么？（铅笔）、请重复“四十四只石狮子”、请受访者用右手拿纸、请受访者折纸、将折好的纸放在腿上、按照这张纸上内容做（闭眼）、请受访者写正确有意义的句子、模仿画图，共九项指标，每正确一个得一分。

3 数据清洗流程

在充分了解了 PFP 量表和 MMSE 量表后，即从原始数据中选取指标，开始数据的清洗。但首先我们**剔除了年龄在 45 岁以下的人**，以便于后期的分析。

3.1 PFP 指标的抽取与计算

PFP 指标一部分（前 8 项）储存于 CHARLS 2015 年数据的体检信息表（Biomarkers）中，另一部分（后 5 项）存储与 CHARLS 2015 年数据的健康状况和功能表（Health Status and Functioning）中。抽取出对应指标即可计算得分。

PFP 量表得分范围为 0-5 分，当得分≥3 则定义为衰弱，取值为 1，否则为 0。对于样本中缺失的数据，我们先剔除了缺失指标个数大于 1 的记录，再使用 MICE[2]模型进行插补。

3.2 MMSE 指标的抽取与计算

MMSE 指标一部分使用了 CHARLS 2015 年数据的健康状况和功能表（Health Status and Functioning）第 1~5 项、14~18 项以及第 30 项，剩余项使用 CHARLS 2018 年数据的认知表（Cognition）进行填补。

在借助 MMSE 量表计算出得分后，我们参照年龄 MMSE 量表给出的教育水平划分（即表 3.2-1）来进一步的进行认知障碍的判断，小于对应的界值即判断为认知障碍。对于样本中缺失的数据，我们先剔除了缺失指标个数大于 3 的记录，再使用 MICE 模型进行插补。

表 3.2-1 受教育程度界值表

受教育程度	认知障碍界值
初中	21
高中	23
大学	24

3.3 得到最终结果

分别得到衰弱、认知障碍的判断结果后。将对应的两表做**自然连接**，即筛选出二者共有的人的记录。

参考文献

[1] WU C,SMIT E,XUE Q L,et al.Prevalence and Correlates of Frailty among Community-Dwelling Chinese Older Adults:The China Health and Retirement Longitudinal Study[J].J GerontolA Biol Sci Med Sci,2017.DOI:10.1093/gerona/glx098.

[2] 这模型我也找不到参考文献，先放个网址吧：mice · PyPI