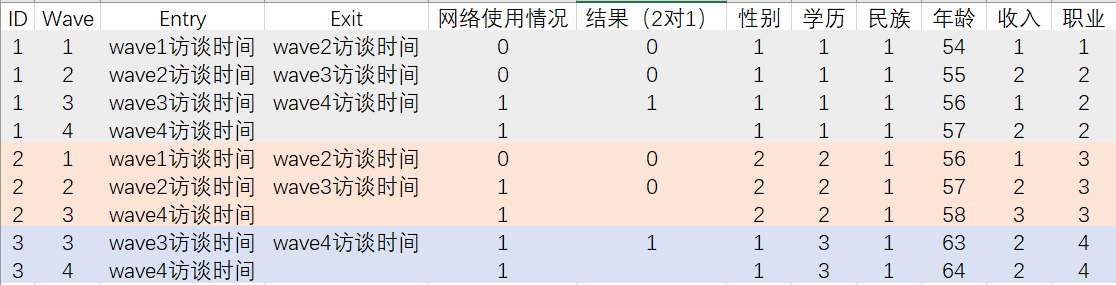
**数据填补策略**

**一、数据填补处理情况**

**数据格式**

****

**（一）协变量**

* **不变的值**

**1、Gender：** Male=1、Female=2

**2、Education level（Edu）：**

* Illiterate=1
* ≤ Elementary school=2
* Middle school=3
* ≥High school=4

**【各组内人数是够的，分组有文献可以参考】**

**3、Nationality（Nat）：**Han=1、Ethnic minorities=2

**【2011年、2015年数据无，录入了2013年、2018年数据】**

**4、Religious belief（Rel）：**Yes=1、No=2

**【2011年、2013年和2015年数据无，录入了2018年数据】**

**5、Work：简化为有无工作（分组有文献参考）：**Yes=1、No=2

已上不变的值均已填补完成，存在以下两个问题：

* 因为是直接录入了各年的数据，而不是只录入第一年然后复制到其余年份，所以同一个参考对象的以上值中存在不完全一致的情况，比方说同一个人前三波教育情况为1，但是最后一波数据原始记录中为2，则该年数据记录为2；经检查这一部分并不是极个别，存在数目较多，所以无法直接删除。

**【对应的修改策略】**以上5个变量需要保证全部一致，所以选择该调查对象**所有数据的众数进行修改**，如果该调查对象只有两年数据且数据不一致，则以**两年中靠前的一年数据为准**。举例说明：2011年、2013年、2015年为1，但2018年为2，这种情况就统一修改为1；2011年为2，2013年为1，这种情况就统一修改为2。

* 完成（1）后，可能依旧有少数空白值由于调查对象所有随访都没有数据而无法填补。**【小龙不用管这一步，后期我们自己在筛选纳入时删除】**
* **变的值**

以下变的值中，“Area”和“Health insurance”两个部分存在一整年都没有数据的情况，其余的空白值都是随机产生的。

**6、Age(years)：** 不分段，直接写入

**【填补策略】**年龄填补比较特别，就是需要符合逻辑，**若丢失为前数据，则有数据年-年份差值；若丢失为后数据，则有数据年+年份差值。**四波数据年份分别为2011、2013、2015、2018，年份差值分别为2、2、3。举例说明：如果该调查对象2011年年龄为45岁，2013年第二波年龄缺失，则在2011年第一波年龄上加上2岁，以47岁作为2013年年龄。

**7、Area：**

* Central of City/Town=1
* Combination zone between urban and rural areas=2
* Village=3

2011年数据、2013年数据两年整年数据都没有，只录入了2015年和2018数据。

**【填补策略】**

* 针对这两年数据，**先通过2015年数据进行填补，若无则用2018年数据填补**。
* Area部分的随机缺失值只有可能出现在2015年或2018年，所以针对随机缺失值可以**通过相邻数据进行填补**，比方说2015年数据缺失，但是2018年数据存在，就以2018年数据填补；2018年缺失则以2015年数据填补。

**8、Marital status（Mari）：**

* Married/Cohabited=1
* Never-married=2
* Divorced/Widowed/separated=3

**【填补策略】**针对随机缺失值可以**通过前一位的相邻数据进行填补**，比方说2013年数据缺失，但是2011年、2015年数据存在，就以2011年数据填补；**如果缺失数据无前一位相邻值，则以相邻后一位数据填补**。

**9、Income（Inc）：**按照个人收入来计算：个人收入=个人工资+退休金

这一部分由许多数据计算得来的，由于有很多人这一部分是缺失的，所以计算最后总收入时就存在许多0值。

**【填补策略】初步先把所有0值转变为NA值**（避免0值导致计算均数错误），然后再通过计算**该调查对象的收入均数进行填补**，比方说2013年数据缺失，2011年数据为30000，2015年数据为40000，就以均数35000填补2013年个人收入部分。

**10、Health insurance（Ins）：**

* 有社保也有商业保险=1
* 有社保没有商业保险=2
* 无保险=3

2015年数据无，录入了2011年、2013年和2018数据。

**【填补策略】**

* 针对这一年数据，**先通过2013年数据进行填补，若无则用2018年数据填补**。
* 针对随机缺失值可以**通过前一位的相邻数据进行填补**，比方说2013年数据缺失，但是2011年数据存在，就以2011年数据填补；**如果缺失数据无前一位相邻值，则以相邻后一位数据填补**。

**11、Living near children（Liv）：**Yes=1、No=2

**【填补策略】**

针对随机缺失值可以**通过前一位的相邻数据进行填补**，比方说2013年数据缺失，但是2011年数据存在，就以2011年数据填补；**如果缺失数据无前一位相邻值，则以相邻后一位数据填补**。

1. **访谈时间：**是以访谈开始的时间为准，Wave 1：2011.05、Wave 2：2013.07、Wave 3：2015.07、Wave 4：2018.07。已填补完成。

**【填补逻辑总结】年龄和个人收入不符合此逻辑（见前细节描述），其余均可以参考**

|  |  |  |  |
| --- | --- | --- | --- |
| 2011 | 2013 | 2015 | 2018 |
| 有 | 有 | 有 | 有 |
| 有 | 有 | 有 | 无 取2015 |
| 有 | 有 | 无 取2013 | 有 |
| 有 | 无 取2011 | 有 | 有 |
| 有 | 无 取2011 | 无 取2011 | 有 |
| 有 | 无 取2011 | 有 | 无 取2015 |
| 有 | 有 | 无 取2013 | 无 取2103 |
| 有 | 无 取2011 | 无 取2011 | 无 取2011 |
| 无 取2013 | 有 | 有 | 有 |
| 无 取2013 | 有 | 有 | 无 取2015 |
| 无 取2013 | 有 | 无 取2013 | 有 |
| 无 取2015 | 无 取2015 | 有 | 有 |
| 无 取2018 | 无 取2018 | 无 取2018 | 有 |
| 无 取2015 | 无 取2015 | 有 | 无 取2015 |
| 无 取2013 | 有 | 无 取2013 | 无 取2013 |
| 无 删除 | 无 删除 | 无 删除 | 无 删除 |

**（二）自变量**

**1、Internet Use（Internet\_Use）：** Yes=1、No=0

**2、Frequency（Freq）：**

* NA=0
* Not Regularly=1
* Almost Every Week=2
* Almost Daily=3

**【填补策略】**

自变量两个变量填补均遵循以下规则：

* 如果该调查对象**所有数据均空白，就暂且不管**，**【后期筛选纳入时我们自己删除】**。
* 如果该调查对象**除了缺失值外的所有数据均为0**，**则缺失值填补为0**。
* 一旦该参考对象**有一年数据出现1/2/3后，如果后面存在缺失值，则延续为1/2/3，如果前面存在缺失值，则填补为0**；如果没有缺失值则为真实随访值。

**（三）因变量**

**1、慢性病发病情况（Chor\_1——14 共14种）**： Yes=1、No=0

* 高血压
* 血脂异常（高血脂或低血脂）
* 糖尿病或血糖升高（包括糖耐量异常和空腹血糖升高）
* 癌症等恶性肿瘤（不包括轻度皮肤癌）
* 慢性肺部疾患如慢性支气管炎或肺气肿、肺心病（不包括肿瘤或癌）
* 肝脏疾病（除外脂肪肝、肿瘤或癌）
* 心脏病（如心肌梗塞、冠心病、心绞痛、充血性心力衰竭和其他心脏疾病）
* 中风（包括脑梗和脑出血）
* 肾脏疾病（不包括肿瘤或癌）
* 胃部疾病或消化系统疾病（不包括肿瘤或癌）
* 情感及精神方面问题
* 与记忆相关的疾病（如老年痴呆症、脑萎缩、帕金森症）
* 关节炎或风湿病
* 哮喘

**2、视力损伤情况（Vis）**（根据6个问题的答案进行判断，如果中老年人报告“失明、常戴眼镜、远视不佳、近视不佳、白内障、青光眼”，则判定为VI视力损伤）：

Yes=1、No=0

**3、抑郁情况（Dep）**（共10道题，每个选项的第1、2、3、4选项分别对应赋值为0、1、2、3分。最后总分为0-30分。≥10分即为抑郁症。）：

Yes=1、No=0

**【填补策略】**

因变量填补均遵循以下规则，**但需要注意每一个调查对象的最后一波数据因为没有之后的随访结果所以都是NA，这里不需要填补**：

* 如果该调查对象**所有数据均空白，就暂且不管**，**【后期筛选纳入时我们自己删除】**。
* 如果该调查对象**除了缺失值外的所有数据均为0，则缺失值填补为0**。
* 一旦该参考对象**有一年数据出现1后，如果后面存在缺失值，则延续为1，如果前面存在缺失值，则填补为0**；如果没有缺失值则为真实随访值。