复杂抽样调查设计多值有序资料一水平多重Logistic回归分析

王 慧[1],李长平[1,2],胡良平[2,3*]

(1. 天津医科大学公共卫生学院卫生统计学教研室,天津 300070;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029;
3. 军事科学院研究生院,北京 100850
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是比较不同分析策略对复杂抽样调查设计多值有序资料一水平多重logistic回归分析结果的异同。通过实例分析,利用四种不同的分析策略(将复杂抽样视为单纯随机抽样,考虑抽样设计不考虑抽样权重,考虑抽样权重不考虑抽样设计,同时考虑抽样设计和抽样权重)对复杂抽样设计多值有序资料进行建模。在四种不同分析策略的累积logistic回归模型拟合的结果中,自变量的偏回归系数、标准误差及P值均有所不同。在对复杂抽样调查设计的多值有序资料回归建模时,将抽样设计和抽样权重纳入统计分析,会得到更准确、更稳健的分析结果。

【关键词】 复杂抽样设计;多值有序资料;Logistic回归分析;抽样权重
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.05.004

One-level multiple Logistic regression analysis of the multi-value ordered data collected from the complex sampling survey design

Wang Hui[1], *Li Changping*[1,2], *Hu Liangping*[2,3*]

(1. *Department of Health Statistics, School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
**Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【Abstract】 To compare the results of one-level multiple logistic regression analysis of multiple-value ordered data collected from the complex sampling survey design by using different analysis strategies. Four different analysis strategies (treating complex sampling as simple random sampling, considering sampling design without considering sampling weights, considering sampling weights without considering sampling design, and considering both sampling design and sampling weights) were used to model the multi-value ordered data of complex sampling design. In the cumulative logistic regression model fitting results of four different analysis strategies, the partial regression coefficients, standard error and P value of independent variables were all different. In the regression modeling of multi-value ordered data of complex sampling survey design, more accurate and reliable analysis results could be obtained by incorporating sampling design and sampling weights into building regression models.

【Keywords】 Complex sampling survey; Multi-value ordered data; Logistic regression analysis; Sampling weights

调查资料,尤其是临床科研或试验资料,结果变量常为“疗效”(死亡、无效、好转、显效、治愈)或“效果”(优、良、中、差),此类资料被称为多值有序资料[1]。在获取此类资料的调查研究中,为提高样本对总体的代表性和估计的可靠性,研究者常将分层抽样、整群抽样、简单随机抽样组合使用,这种调查被称为复杂抽样调查设计。然而,在对复杂抽样数据进行回归分析时,研究者常常忽略此前采取的抽样设计方法。在不同的抽样阶段下,每个个体所对应的抽样概率不同,抽样权重也就不同,因此,抽样误差估计极为复杂。孙日扬等[2]认为,在复杂抽样调查研究的分析中应考虑抽样权重和观测权重,同时提出了综合权重的概念。在多重线性回归分析中纳入综合权重的分析结果更加准确、稳健。本研究通过不同的分析策略对复杂抽样调查设计多值有序资料进行多重logistic回归分析,并探讨各种分析方法之间的异同。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1 累积多重logistic回归模型的构建与求解

1.1 累积logistic回归模型

结果变量为多值有序变量的logistic回归模型又被称为累积logistic回归模型,它是二值变量logistic

回归模型的扩展[3]，其回归模型见式(1)：

$$y^* = \alpha + \sum_{k=1}^{p} \beta_k x_k + \varepsilon \quad (1)$$

其中 y^* 表示观测现象的内在趋势，不能被直接测量；$xk(k = 1,2,\cdots,p)$ 为 p 个自变量，ε 为误差项。当结果变量有 J 个可能的结局时，相应的取值为 y=1、y=2、… 、y=J，即共有 J-1 个分界点将各相邻类别分开，即：若 $y^* \leqslant \mu_1$，则 y=1；若 $\mu_1 < y^* \leqslant \mu_2$，则 y=2；…；若 $y^* \geqslant \mu_{J-1}$，则 y=J。

给定 x 值的累积概率可以按式(2)表示。其中 $1 - P(y \leqslant j|x)$ 即为 $P(y \geqslant j|x)$，这样就依次将 J 个可能的结局合并成两个，从而进行一般的多重 logistic 回归模型分析。

$$\ln \frac{P(y \leqslant j|x)}{1 - P(y \leqslant j|x)} = \mu_j - (\alpha + \sum_{k=1}^{p} \beta_k x_k + \varepsilon) \quad (2)$$

相应地，累积概率可通过式(3)进行预测：

$$P(y \leqslant j|x) = \frac{e^{\mu_j - \left(\alpha + \sum_{k=1}^{p} \beta_k x_k\right)}}{1 + e^{\mu_j - \left(\alpha + \sum_{k=1}^{p} \beta_k x_k\right)}} \quad (3)$$

SAS 软件在实际运行中，定义 β_{0j} 为各类中截距 α 与分界点 μ_j 的综合，所以上式就转化为式(4)：

$$P(y \leqslant j|x) = \frac{e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}}{1 + e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}} \quad (4)$$

其参数估计采用最大似然法求解，其对数似然方程见式(5)：

$$\sum_{i=1}^{N} x_{ik} \left[y_{ij} - n_{ij} \frac{e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}}{1 + e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}} \right] = 0 \quad (5)$$

对式(5)的求解需要用到非线性迭代算法，一般需要借助统计软件来实现，此处从略。由以上讨论可知，如果结果变量中有 J 个可能的结局，则可获得 J-1 个累积 logit 函数（当进行统计分析时，若有 m 个截距项 β_{0j} 无统计学意义，则只能获得 J-m-1 个累积 logit 函数）。累积 logistic 回归模型应用的假设条件是比例优势假定，其含义是自变量的作用与所有累积 logit 的截断点无关，即对于任意一个自变量 x_k 而言，所有的累积 logit 都有一组相同的参数估计值，只是截距参数有所差别。若不满足比例优势假定条件时，Bender 等[4]建议可以考虑两种方法，一是采用独立的二分类模型，二是采用偏比例优势模型。

1.2 复杂抽样的多值有序 logistic 回归模型

复杂抽样多值有序资料的 logistic 回归模型的构建、求解的思路和方法与单纯随机抽样设计资料的累积 logistic 回归模型基本相同，主要差别在于：复杂抽样的多值有序 logistic 回归模型考虑到了与特定抽样设计条件下对应的“抽样权重”[3]。其参数估计求解于下面的对数似然方程组，见式(6)：

$$\sum_{i=1}^{N} \omega_i x_{ik} \left[y_{ij} - n_{ij} \frac{e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}}{1 + e^{\beta_{0j} + \sum_{k=1}^{p} \beta_k x_k}} \right] = 0 \quad (6)$$

这种结合了抽样权重的似然估计通常被称为加权的最大似然估计或伪似然估计。

2 基于 SAS 的实例分析

2.1 问题与数据

本研究所使用数据为美国卫生与公众服务部开展的医疗支出调查（Medical Expenditure Panel Survey，MEPS），用于对医疗保健的各个方面进行全国性和地区性的评估。MEPS 采用分层整群抽样，抽样权重会根据无响应情况进行调整，并根据当前人口调查的人口控制总量进行调整。在本例中，使用欧洲议会议员提供的 1999 年全年综合数据来研究家庭收入与性别和种族的关系。样本量为 24 618，分层数为 143，群集数为 460。数据存储于 SAS 数据集 MEPS。本例中变量命名及赋值见表 1，分析所用示例数据见表 2。

表 1 MEPS 数据集变量命名及赋值

变 量	变量命名	赋 值
分层变量	stratum	共 143 层
群变量	cluster	共 460 个群
被观测者识别号	ID	-
性别	sex	1=男性，2=女性
人种	race	1=其他人种，2=白种人
家庭收入	income	1=贫穷，2=接近贫穷，3=低收入，4=中等收入，5=高收入
抽样权重	weight	-

表 2 1999 年美国家庭收入情况及影响因素数据（基于 MEPS 数据集）

stratum	cluster	ID	sex	race	income	weight
131	2	1	1	2	4	14137.86
131	2	2	2	2	4	17050.99
131	2	3	1	2	4	35737.55
…	…	…	…	…	…	…
78	10	24614	2	2	5	13224.29
68	1	24615	1	2	4	16793.47
68	1	24616	2	2	4	13627.61
94	1	24618	2	2	1	11660.13

2.2 分析策略

2.2.1 将复杂调查设计资料视为“单纯随机抽样设计资料”

2.2.1.1 SAS 程序

需要调用 LOGISTIC 过程来实现单纯随机抽样资料的累积 logistic 回归。

```
proc logistic data=meps descending;
class sex race;
model income = sex race;
run;
```

【说明】“descending”选项要求对响应变量表中具有较低(1=贫穷)有序值的响应进行建模,class 语句指定分类变量 sex 和 race;model 语句中响应变量为 income,解释变量(即自变量)为 sex 和 race。在此段 SAS 过程步程序之前,应基于表 2 资料创建临时 SAS 数据集 meps,此段 SAS 数据步程序省略了。

2.2.1.2 主要输出结果及解释

最大似然估计分析

参数		自由度	估计	标准误差	Wald 卡方	Pr≥卡方
Intercept	5	1	−0.95	0.0166	3276.486	<0.0001
Intercept	4	1	0.369	0.0157	554.978	<0.0001
Intercept	3	1	1.151	0.0174	4369.013	<0.0001
Intercept	2	1	1.526	0.0191	6406.951	<0.0001
sex	1	1	0.087	0.0115	56.673	<0.0001
race	1	1	−0.29	0.0144	399.159	<0.0001

优比估计

效应	点估计	95% Wald 置信限	
sex 2−1	1.189	1.137	1.244
race 2−1	0.563	0.532	0.596

在形式上,累积 logistic 回归模型分析的结果大致可分为模型基本信息、比例优势假定检验结果、模型拟合信息以及参数估计结果四部分。因篇幅所限,只给出参数估计结果;比例优势假定检验结果显示,χ^2=7.4931,P=0.2766,不拒绝“比例优势假设”的条件,即满足比例优势假定,可采用累积 logistic 回归模型。拟合的累积 logistic 模型给出 4 个截距项以及 sex 和 race 的两个自变量的参数估计值,结果显示,性别和人种对家庭收入的影响均有统计学意义。女性贫穷的风险是男性的 1.189 倍;白种人贫穷的风险比其他人种低 43.7%(=1−0.563)。

2.2.2 考虑抽样设计但不考虑抽样权重

2.2.2.1 SAS 程序

需要调用 surveylogistic 过程来实现复杂随机抽样多值有序资料的累积 logistic 回归模型分析:

```
proc surveylogistic data=meps;
strata stratum;
cluster cluster;
class sex race;
model income(descending) = sex race;
run;
```

【说明】由于研究数据属于分层整群随机抽样调查资料,故在 strata 语句中指定分层变量为 stratum,cluster 语句中指定群集变量为 cluster。

2.2.2.2 主要输出结果及解释

模型信息

数据集	WORK.MEPS	
响应变量	income	income
响应水平数	5	
层变量	stratum	stratum
层数	143	
聚类变量	cluster	cluster
聚类数	460	
模型	累积 Logit	
优化方法	Fisher 评分法	
方差调整	自由度	

响应概略

有序值	income	总频数
1	5	7784
2	4	7722
3	3	3853
4	2	1376
5	1	3883

复杂抽样 logistic 回归主要结果大致可以分为三部分。第一部分是模型的基本信息,可以看到指定的

分层变量和群集，拟合的是累积logistic回归模型；在响应概略表中可以看到因变量income顺序为5、4、3、2、1以及各响应水平的总频数。第二部分模型检验结果均显示整体模型具有统计学意义（P均<0.01）。

最大似然估计分析

参数		估计	标准误差	t值	Pr≥|t|
Intercept	5	−0.9530	0.0409	−23.27	<0.0001
Intercept	4	0.3687	0.0392	9.41	<0.0001
Intercept	3	1.1506	0.0452	25.44	<0.0001
Intercept	2	1.5264	0.0503	30.37	<0.0001
sex	1	0.0865	0.0091	9.53	<0.0001
race	1	−0.2873	0.0352	−8.17	<0.0001

优比估计

效应	点估计	95% 置信限	
sex 2−1	1.189	1.147	1.232
race 2−1	0.563	0.49	0.646

参数估计结果显示性别和人种对家庭收入的影响均具有统计学意义。女性贫穷的风险是男性的1.189倍；白种人贫穷的风险比其他人种低43.7%（=1−0.563）。

2.2.3 考虑抽样权重，不考虑抽样设计

2.2.3.1 SAS程序

需要调用surveylogistic过程来实现复杂随机抽样多值有序资料的累积logistic回归模型分析：

```
proc surveylogistic data=meps;
class sex race;
model income(descending) = sex race;
weight weight;
run;
```

【说明】加入weight语句，指定权重变量weight。

2.2.3.2 主要输出结果及解释

权重变量 weight weight

响应概略

有序值	income	总频数	总权重
1	5	7517	106183852
2	4	7406	86181788
3	3	3657	38927994
4	2	1323	12321996
5	1	3662	32795137

与前文“模型信息”相同的部分此处从略。指定的权重变量在前文2.2.2.2的基础上增加的各响应水平的总权重。模型检验结果均显示整体模型具有统计学意义（P均<0.01）。

最大似然估计分析

参数		估计	标准误差	t值	Pr≥|t|
Intercept	5	−0.7290	0.0202	−36.06	<0.0001
Intercept	4	0.5996	0.0195	30.81	<0.0001
Intercept	3	1.4223	0.0220	64.57	<0.0001
Intercept	2	1.7996	0.0244	73.85	<0.0001
sex	1	0.0963	0.0140	6.86	<0.0001
race	1	−0.3827	0.0178	−21.49	<0.0001

优比估计

效应	点估计	95% 置信限	
sex 2−1	1.212	1.147	1.281
race 2−1	0.465	0.434	0.499

结果显示女性贫穷的风险是男性的1.212倍；白种人贫穷的风险比其他人种低53.5%（=1−0.465）。

2.2.4 同时考虑抽样设计和抽样权重

2.2.4.1 SAS程序

需调用SURVEYLOGISTIC过程来实现复杂随机抽样多值有序资料的累积logistic回归模型分析：

```
proc surveylogistic data=meps;
strata stratum;
    cluster cluster;
```

```
class sex race;
model income(descending) = sex race;
weight weight;
run;
```

【说明】在第“2.2.3.1SAS程序节”的基础上，加入strata语句指定分层变量stratum，加入cluster语句指定群集变量cluster。

2.2.4.2 主要输出结果及解释

与前文“模型信息”相同的部分此处从略。

模型信息

层变量	stratum	stratum
层数	143	
聚类变量	cluster	cluster
聚类数	460	

模型的基本信息在“第2.2.3.2主要输出结果及解释”的基础上增加了关于分层的内容。第二部分模型检验结果均显示模型总体具有统计学意义（P均<0.01）。

最大似然估计分析

参数		估计	标准误差	t 值	Pr≥|t|
Intercept	5	−0.7290	0.0449	−16.23	<0.0001
Intercept	4	0.5996	0.0433	13.85	<0.0001
Intercept	3	1.4223	0.0486	29.28	<0.0001
Intercept	2	1.7996	0.0537	33.52	<0.0001
sex	1	0.0963	0.0119	8.10	<0.0001
race	1	−0.3827	0.0357	−10.73	<0.0001

优比估计

效应	点估计	95% 置信限	
sex 2−1	1.212	1.157	1.27
race 2−1	0.465	0.404	0.535

最后参数估计结果显示，女性贫穷的风险是男性的1.212倍；白种人贫穷的风险比其他人种低53.5%（=1−0.465）。因此，最终建立的四个模型为：

$$P_{\text{高收入}}=\frac{e^{-0.7290+0.0963sex-0.3827race}}{1+e^{-0.7290+0.0963sex-0.3827race}}$$

$$P_{\text{高收入}}+P_{\text{中等收入}}=\frac{e^{0.5996+0.0963sex-0.3827race}}{1+e^{0.5996+0.0963sex-0.3827race}}$$

$$P_{\text{高收入}}+P_{\text{中等收入}}+P_{\text{低收入}}=\frac{e^{1.4223+0.0963sex-0.3827race}}{1+e^{1.4223+0.0963sex-0.3827race}}$$

$$P_{\text{高收入}}+P_{\text{中等收入}}+P_{\text{低收入}}+P_{\text{接近贫穷}}=\frac{e^{1.7996+0.0963sex-0.3827race}}{1+e^{1.7996+0.0963sex-0.3827race}}$$

$$P_{\text{贫穷}}=1-P_{\text{高收入}}-P_{\text{中等收入}}-P_{\text{低收入}}-P_{\text{接近贫穷}}=\frac{1}{1+e^{1.7996+0.0963sex-0.3827race}}$$

2.3 不同分析策略的结果比较

结合上述分析结果可以看出，考虑抽样设计的累积logistic回归模型与普通累积logistic回归模型的结果相比，二者的参数估计值完全相同，但是sex的标准误降低且OR值的置信区间缩窄，说明对分层整群抽样资料进行分析时，若忽视分层信息，则会导致过于保守的检验（P值偏大），同时OR的置信区间也会变宽，容易出现假阳性结果；而race的标准误和OR值的置信区间会增大，本研究认为主要是由于race在群内存在相关性导致的。

只考虑抽样权重的累积logistic回归模型与普通累积logistic回归模型的结果相比，参数估计值和标准误均发生了变化，sex的估计值和标准误变化不大，而在考虑抽样权重后 race 的参数估计值降低，标准误和OR值的置信区间几乎没有变化，所以本研究认为对于存在群内相关性的变量，在加入权重变量后，可在一定程度上校正这种群内相关性导致的预测不稳定。

同时考虑抽样设计和抽样权重的累积logistic回归模型与普通累积Logistic回归模型的结果相比，自变量的参数估计值和标准误均发生了变化，sex的估计值略高，而标准误和置信区间变化不大；race不仅标准误增大了，而且参数估计值也发生了变化，可能是因为race在群变量因素的各个水平中存在相关性，同时在该群变量水平的权重也很小，这也是为什么在考虑了抽样权重后，其标准误仅与考虑群变量的模型相比略有变化，因为它的影响很小。而同时考虑抽样设计和抽样权重的累积logistic回归模型与只考虑抽样权重的累积logistic回归模型相比，sex的参数估计值不变，但其标准误降低、OR值的置信区间变窄，说明在考虑抽样权重的基础上，纳入抽样设计的分析，会使分析结果更加准确和稳健。

3 讨论与小结

3.1 讨论

在社会科学或者卫生领域的研究中，尤其是大规模研究，常涉及多地区或者多中心的抽样，调查对象过于分散，若采用单纯的随机抽样，会出现调查成本高、可行性低的情况[5]，所以研究者经常采用复杂抽样设计，以提高调查的可行性，节约调查的成本支出[6]。但在实际进行复杂抽样调查资料的统计分析时，多数研究者却常常忽略抽样设计，采用单纯随机抽样的普通 logistic 回归模型分析。例如本研究数据是采用动态权重法进行的分层整群随机抽样数据，由于存在群变量，而有可能导致存在群内的相关性，若采用普通的累积 logistic 回归模型分析，会导致较大的假阳性错误；其次，由于普通的累积 logistic 回归模型的应用假设条件是所有样本均来自简单随机抽样，每一个个体被抽中的概率相同[7]，所以不能将抽样权重纳入分析，也会造成信息的损失和结果的偏差。所以在对复杂抽样资料进行统计分析与推断时，将抽样设计和抽样权重正确纳入分析，是分析者应该重点关注的问题。

本文采用 SAS 中的 SURVEYLOGISTIC 过程对复杂随机抽样调查资料进行累积 logistic 回归模型分析，这是一种基于复杂抽样调查设计的分析方法，可以结合抽样设计（分层、整群随机等）和抽样权重进行分析，可以不依赖于模型的假定，充分利用抽样权重、群效应信息等，进一步提高估计结果的准确性和稳定性[8]。考虑到本研究数据是分层整群抽样资料，这类资料也可以通过多水平 logistic 回归模型进行分析，因篇幅所限，此处从略。

3.2 小结

本研究通过分层整群抽样的实例数据进行了不同分析策略的复杂抽样调查多值有序资料的多重 logistic 回归分析，对分析结果给出了解释，并进一步探讨了不同分析策略结果之间的差异，结果表明：在对复杂抽样资料进行统计分析时，将抽样设计和抽样权重纳入分析，会得到更加准确和稳定的分析结果。

参考文献

[1] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京：人民卫生出版社，2012：508-517.

[2] 孙日扬，胡良平. 复杂随机抽样数据的多重线性回归分析方法及其应用[J]. 军事医学，2015，39(5)：380-385.

[3] SAS Institute Inc . SAS/Stat 9.4 user's guide[M]. Cary，NC：SAS Institute Inc，2016：5749-6006，9679-9682.

[4] Bender R，Benner A. Calculating ordinal regression models in SAS and S-Plus[J]. Biom J，2015，42(6)：677-699.

[5] Osborne JW. Best practices in using large，complex samples：the importance of using appropriate weights and design effect compensation [J]. Practical Assessment，Research and Evaluation，2011，16(12)：1-7.

[6] 缪凡，童峰. 复杂抽样数据的 logistic 回归分析方法及其应用[J]. 中国卫生统计杂志，2008，25(6)：577-579.

[7] 王济川，郭志辉. Logistic 回归模型：方法与应用[M]. 北京：高等教育出版社，2001：1-262.

[8] 冯国双，刘德平. 医学研究中的 logistic 回归分析及 SAS 实现[M]. 北京：北京大学医学出版社，2015：97-158.

（收稿日期：2019-09-27）

（本文编辑：吴俊林）