复杂抽样调查设计多值名义资料一水平多重Logistic回归分析

刘媛媛[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：李长平，E-mail：1067181059@qq.com）

【摘要】 本文目的是介绍复杂抽样调查设计多值名义资料一水平多重logistic回归模型构建，并探讨不同策略之间的差异。采用SAS中的LOGISTIC过程和SURVEYLOGISTIC过程，分别按照是否考虑抽样设计与是否考虑抽样权重共4种分析策略对数据构建广义logistic回归模型，并比较结果。不同分析策略所得结果显示，不仅参数估计值、回归系数标准误、*OR*值及其置信区间的估计值有所差别，而且对纳入模型的解释变量也有影响。因此，在对复杂抽样调查设计多值名义资料构建广义logistics回归模型时，既要考虑抽样设计，又要兼顾抽样权重，否则即使样本量足够大，也会导致错误的推断结论。

【关键词】 复杂抽样；多值名义资料；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/scjsws20191119005

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

Liu Yuanyuan[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: Li Changping, E-mail: 1067181059@qq.com*）

【Abstract】 The purpose of this article was to introduce the construction of multiple logistic regression models with multi-value nominal data collected from the complex sampling survey design, and to explore the differences between different strategies. Using the LOGISTIC procedure and the SURVEYLOGISTIC procedure in SAS software, generalized logistics regression models were constructed based on whether the sampling design or the sampling weights were considered, and the results were compared. The results obtained by different analysis strategies showed that not only the values of parameter estimation, the standard error of the regression coefficients, the *OR* value and its confidence intervals were different, but also the explanatory variables in the established models were also different. When constructing a generalized logistics regression model for multi-value nominal data of complex sampling design, both the sampling design and the sampling weights should be considered. Otherwise, even if the sample size was large enough, it would lead to the erroneous inference conclusions.

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

在调查研究中，常见的结果变量及其取值除了二值资料、多值有序资料之外，还包括如血型“A型、B型、O型、AB型”或疾病分型“A型、B型、C型”这样的资料，称为多值名义资料。此类资料特指因变量或结果变量为多值名义变量，而自变量可以是定性的、定量的或混合型的资料[1]。现在，复杂抽样调查设计在实际调查研究中使用越来越多，对由此获得的复杂抽样数据进行统计分析时，需充分考虑由不同的抽样方法而产生的不同“抽样权重”。本文通过不同分析策略对复杂抽样调查设计多值名义资料进行多重logistic回归分析，并探讨不同策略之间的差异。

1 多值名义资料多重logistic回归模型简介

1.1 简单随机抽样下多值名义资料多重logistic回归模型的构建

对于结果变量为多值名义变量的logistic回归

项目基金：国家高技术研究发展计划课题资助（2015AA020102）；国家自然科学基金项目（81803333）

模型，其结果变量的多个取值之间是“无序的”，假设结果变量 Y 的取值的类别个数为（D+1）个，这时，总是以其中一个取值类别作为对照，将其他类别与对照类别进行比较，共生成 D 个 logistic 回归模型，所构建的 logistic 回归模型也被称为扩展的 logistic 回归模型或广义 logit 模型[2]。见式（1）。

$$\log\left[\frac{Pr(Y=i|x)}{Pr(Y=D+1|x)}\right]=\alpha_i+x\beta_i, i=1,\cdots,D \quad (1)$$

其中，$\alpha_1,\cdots,\alpha_D$ 是 D 个截距参数，$\beta_1,\cdots,\beta_D$ 是 D 个参数组成的向量，β_i 代表第 i 类相对于第（D+1）类的回归系数向量，x 代表协变量向量。此模型最早由 McFadden[3] 介绍，并被作为多项 logit 模型而熟知。

对上式进行转换可得式（2）：

$$P_i=\frac{e^{\beta_i x}}{1+\sum_{i=1}^{D}e^{\beta_i x}}, i=1,\cdots,D \quad (2)$$

因为所有（D+1）类的概率之和必须为 1，所以第（D+1）类的概率为式（3）[4]：

$$P_{D+1}=\frac{1}{1+\sum_{i=1}^{D}e^{\beta_i x}}, \quad i=1,\cdots,D \quad (3)$$

1.2 复杂抽样下多值名义资料多重 logistic 回归模型的构建

对于复杂抽样下多值名义资料多重 logistic 回归模型来说，建模时通过使用伪对数似然函数来估计模型参数。当结果变量为多值名义资料时，构建广义 logit 模型将使用 logit 连接函数拟合每个响应类别的预期比例与参考类别的预期比例的比值[2]。此时，广义 logit 模型即为式（4）：

$$\log\left[\frac{\pi_{hijd}}{\pi_{hij(D+1)}}\right]=x_{hij}\beta_d \quad (4)$$

其中，d=1, 2, …, D。模型参数向量为 $\beta_d=(\beta_{d1},\beta_{d2},\cdots,\beta_{dk})'$。$\pi_{hij}$ 为结果变量的期望向量。x_{hij} 为第 h 层第 i 个群集第 j 个单位解释变量的 k 维行向量。

利用伪对数似然函数对模型参数进行估计，求解最大似然估计值。见式（5）：

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\omega_{hij}D_{hij}\left[\mathrm{diag}(\pi_{hij})-\pi_{hij}\pi'_{hij}\right]^{-1}(y_{hij}-\pi_{hij})=0 \quad (5)$$

在式（5）中，D_{hij} 为连接函数关于 θ 的偏导数矩阵，θ 为回归系数的列向量，$\theta=(\beta'_1,\beta'_2,\cdots,\beta'_D)'$，$\omega_{hij}$ 为抽样权重，y_{hij} 为变量 Y 的前 D 个类别的指示变量组成的一个 D 维的列向量[5]。

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面板调查（Medical Expenditure Panel Survey，MEPS）的数据，对医疗保健的各个方面进行评估[2]。该研究采用分层整群抽样，抽样权重根据无响应情况和当前人口调查的人口控制总量进行调整。在本例中，利用 1999 年全年数据来研究医保覆盖情况与人口学变量之间的关系。数据存储于 SAS 数据集 MEPS，样本量为 24 618，变量为 8 个，具体变量名及赋值见表 1。

表 1 数据集中变量名及赋值或单位

变　量	变量名	赋值或单位
分层变量	stratum	个
群变量	cluster	个
性别	sex	1=男性，2=女性
人种	race	1=美国印第安人，2=爱斯基摩人，3=亚洲或太平洋岛民，4=黑人，5=白人
家庭收入水平	income	1=贫穷，2=接近贫穷，3=低收入，4=中等收入，5=高收入
健康保险覆盖情况	insurance	1=全年有私人保险，2=全年只有公共保险，3=全年没有保险
全年卫生保健总支出	expenditure	美元
抽样权重	weight	具体数值

2.2 分析策略

2.2.1 按单纯随机抽样进行分析

既不考虑抽样设计，也不考虑抽样权重：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

基于表 1 及其具体数据创建临时 SAS 数据集 MEPS 所对应的 SAS 数据步程序从略。调用 LOGISTIC 过程来实现单纯随机抽样设计资料的广义 logit 模型。

```
proc logistic data=meps;
```

```
class sex race income;
model insurance=sex race income expenditure/ link=glogit;
run;
```

【说明】class语句指定分类变量sex、race、income；model语句中响应变量为*Y*= insurance，以insurance=3为参考类别，解释变量（即自变量）为sex、race、income和expenditure。在MODEL语句中指定了LINK=GLOGIT选项，即指定拟合广义logit回归模型，即扩展的多重logistic回归模型。

2.2.1.2 主要输出结果及解释

最大似然估计分析

参数		insurance	自由度	估计	标准误差	Wald卡方	Pr>卡方
Intercept		ANY PRIVATE	1	2.574	15.748	0.027	0.870
Intercept		PUBLIC ONLY	1	1.806	15.748	0.013	0.909
SEX	FEMALE	ANY PRIVATE	1	0.063	0.020	10.016	0.002
SEX	FEMALE	PUBLIC ONLY	1	0.141	0.024	35.971	<0.001
race	ALEUT, ESKIMO	ANY PRIVATE	1	6.914	62.992	0.012	0.913
race	ALEUT, ESKIMO	PUBLIC ONLY	1	7.648	62.992	0.015	0.903
race	AMERICAN INDIAN	ANY PRIVATE	1	−1.812	15.749	0.013	0.908
race	AMERICAN INDIAN	PUBLIC ONLY	1	−1.888	15.749	0.014	0.905
race	ASIAN OR PACIFIC ISLANDER	ANY PRIVATE	1	−1.918	15.748	0.015	0.903
race	ASIAN OR PACIFIC ISLANDER	PUBLIC ONLY	1	−2.097	15.748	0.018	0.894
race	BLACK	ANY PRIVATE	1	−1.575	15.748	0.010	0.920
race	BLACK	PUBLIC ONLY	1	−1.587	15.748	0.010	0.920
income	HIGH INCOME	ANY PRIVATE	1	1.484	0.044	1144.682	<0.001
income	HIGH INCOME	PUBLIC ONLY	1	−0.442	0.062	51.098	<0.001
income	LOW INCOME	ANY PRIVATE	1	−0.257	0.039	43.628	<0.001
income	LOW INCOME	PUBLIC ONLY	1	−0.081	0.046	3.059	0.080
income	MIDDLE INCOME	ANY PRIVATE	1	0.629	0.035	322.482	<0.001
income	MIDDLE INCOME	PUBLIC ONLY	1	−0.436	0.047	87.628	<0.001
income	NEAR POOR	ANY PRIVATE	1	−0.813	0.058	193.372	<0.001
income	NEAR POOR	PUBLIC ONLY	1	0.231	0.061	14.425	0.000
expenditure		ANY PRIVATE	1	0.000	0.000	261.668	<0.001
expenditure		PUBLIC ONLY	1	0.000	0.000	309.626	<0.001

优比估计

效应	insurance	点估计	95% Wald置信限	
SEX MALE−FEMALE	ANY PRIVATE	1.134	1.049	1.227
SEX MALE−FEMALE	PUBLIC ONLY	1.327	1.210	1.455
race WHITE−ALEUT, ESKIMO	ANY PRIVATE	>999.999	<0.001	>999.999
race WHITE−ALEUT, ESKIMO	PUBLIC ONLY	>999.999	<0.001	>999.999
race WHITE−AMERICAN INDIAN	ANY PRIVATE	0.816	0.567	1.174
race WHITE−AMERICAN INDIAN	PUBLIC ONLY	1.206	0.811	1.792
race WHITE−ASIAN OR PACIFIC ISLANDER	ANY PRIVATE	0.734	0.613	0.879
race WHITE−ASIAN OR PACIFIC ISLANDER	PUBLIC ONLY	0.978	0.787	1.215
race WHITE−BLACK	ANY PRIVATE	1.035	0.927	1.156
race WHITE−BLACK	PUBLIC ONLY	1.629	1.444	1.838
income NEGATIVE OR POOR−HIGH INCOME	ANY PRIVATE	12.506	10.948	14.286

		优比估计		
income NEGATIVE OR POOR-HIGH INCOME	PUBLIC ONLY	0.310	0.263	0.366
income NEGATIVE OR POOR-LOW INCOME	ANY PRIVATE	2.192	1.939	2.478
income NEGATIVE OR POOR-LOW INCOME	PUBLIC ONLY	0.445	0.393	0.505
income NEGATIVE OR POOR-MIDDLE INCOME	ANY PRIVATE	5.320	4.742	5.968
income NEGATIVE OR POOR-MIDDLE INCOME	PUBLIC ONLY	0.312	0.275	0.355
income NEGATIVE OR POOR-NEAR POOR	ANY PRIVATE	1.258	1.067	1.484
income NEGATIVE OR POOR-NEAR POOR	PUBLIC ONLY	0.608	0.518	0.714
expenditure	ANY PRIVATE	1.000	1.000	1.000
expenditure	PUBLIC ONLY	1.000	1.000	1.000

这里仅列出部分广义 logit 回归模型分析结果。其中模型参数的假设检验分别使用似然比检验、评分检验和 Wald 检验三种方法，结果显示回归模型有统计学意义。最大似然估计结果显示，性别、家庭收入水平和全年卫生保健总支出对健康保险覆盖情况的影响均有统计学意义；优势比估计结果显示，相对于全年没有保险者而言，女性、家庭收入水平非贫穷者、全年卫生保健总支出高者倾向于全年有私人保险；男性、家庭收入水平非贫穷者、全年卫生保健总支出高者倾向于全年只有公共保险。

2.2.2 考虑抽样设计，但不考虑抽样权重

2.2.2.1 SAS 程序

调用 SURVEYLOGISTIC 过程来实现复杂抽样调查设计多值名义资料的广义 logit 回归模型。

```
proc surveylogistic data=meps;
strata stratum;
cluster cluster;
class sex race income;
model insurance = sex race income expenditure/ link=glogit;
run;
```

【说明】STRATA 语句用于指定在分层抽样设计中的分层变量，CLUSTER 语句指定整群抽样设计中的群变量。其他解释同上。

2.2.2.2 主要输出结果及解释

SAS 输出结果很多，由于篇幅限制，此部分结果从略。由输出结果得知：性别、人种、家庭收入水平和全年卫生保健总支出对健康保险覆盖情况的影响均有统计学意义。

2.2.3 不考虑抽样设计，但考虑抽样权重

2.2.3.1 SAS 程序

调用 SURVEYLOGISTIC 过程来实现复杂抽样调查设计多值名义资料的广义 logit 回归模型。

```
proc surveylogistic data= meps;
class sex race income;
model insurance = sex race income expenditure/ link=glogit;
weight weight;
run;
```

【说明】WEIGHT 语句指定权重变量，其他解释同上。

2.2.3.2 主要输出结果及解释

由于篇幅限制，SAS 输出结果从略。由输出结果得知：性别、人种、家庭收入水平和全年卫生保健总支出对健康保险覆盖情况的影响均有统计学意义。

2.2.4 同时考虑抽样设计和抽样权重

2.2.4.1 SAS 程序

调用 SURVEYLOGISTIC 过程来实现复杂抽样调查设计多值名义资料的广义 logit 模型。

```
proc surveylogistic data= meps;
strata stratum;
cluster cluster;
weight weight;
class sex race incom;
model insurance = sex race income expenditure/ link=glogit;
run;
```

【说明】分别用 STRATA 语句、CLUSTER 语句、WEIGHT 语句指定复杂抽样中的分层变量、群变量、权重变量，CLASS 语句指定分类变量；MODEL 语句中

结果变量为insurance，以insurance=3为参考类别，解释变量为sex、race、income和expenditure。在MODEL语句中指定LINK=GLOGIT选项，即指定拟合广义logit回归模型。

2.2.4.2 主要输出结果及解释

由于篇幅限制，SAS输出结果从略。由输出结果得知：性别、人种、家庭收入水平和全年卫生保健总支出对健康保险覆盖情况的影响均有统计学意义。相对于全年没有保险者而言，女性、爱斯基摩人（相对于白人）、家庭收入水平非贫穷者、全年卫生保健总支出高者倾向于全年有私人保险，而男性、人种为美国印第安人或亚洲或太平洋岛民或黑人（相对于白人）者、全年卫生保健总支出低者倾向于无保险；女性、人种非白人、家庭收入水平贫穷者、全年卫生保健总支出高者倾向于全年只有公共保险。

2.3 不同分析策略的结果比较

不考虑复杂抽样的普通广义logit回归模型与仅考虑抽样设计的广义logit回归模型所得回归系数及*OR*值的参数估计值相同，仅回归系数的标准误及*OR*值的95% *CI*不同，而其变化有的增大有的减小。说明是否考虑抽样方法对广义logit回归模型参数估计存在影响。

考虑抽样权重与同时考虑抽样设计和抽样权重之后构建的广义logit回归模型所得回归系数及*OR*值的参数估计值相同，却与前两种分析策略结果不同。而且这两种分析策略得到的回归系数标准误及*OR*值的95% *CI*也有增大或减小的区别。race变量在不考虑抽样权重时，对健康保险覆盖情况无影响；但在考虑抽样权重后，race变量的不同情况对健康保险覆盖情况的影响有统计学意义。说明在对复杂抽样调查设计多值名义资料构建广义logit回归模型时，首先应考虑研究采用的抽样方法，由此计算相应的抽样权重，否则可能产生较大偏差[5]。

3 讨论与小结

抽样调查是调查研究中相对简单易行且代表性较好的方法之一，但单一的抽样方法在实际应用中存在一些缺点，所以复杂抽样的思想和方法应运而生，由复杂抽样方法获得的样本称为复杂样本[6]。由于复杂随机抽样每个阶段的抽样方法可能不同，所以其抽样误差的计算相当复杂。因此，在对复杂样本进行统计分析时，既要充分考虑多种抽样方法联合使用对抽样误差的影响，又要注意不同抽样率下抽样权重的不同，否则会使参数及其置信区间等的估计产生偏差。

为了探讨在复杂抽样或单纯随机抽样基础上进行统计分析的差异，本研究分别采用SAS软件中的LOGISTIC过程和SURVEYLOGISTIC过程，按照是否考虑抽样设计与是否考虑抽样权重共4种分析策略对数据进行统计分析。由于LOGISTIC过程可采用逐步回归法对自变量进行筛选，而SURVEYLOGISTIC过程不支持，所以本研究并未使用该选项。结果显示，如果在统计分析中忽视“复杂抽样”或“抽样权重”，不仅会对参数估计值、回归系数标准误、*OR*值及其置信区间的估计产生影响[6]，而且对纳入广义logit回归模型的解释变量也有影响。由于复杂抽样中的抽样权重包含进行参数点估计时所需的信息，但不包含标准误估计的信息，因此，在SURVEYLOGISTIC过程中需对方差进行估计。正确的方差估计包括每一个抽样阶段的方差估计和联合抽样概率[7]。SAS中可采用Taylor级数线性近似法（线性化）、重抽样等方法，如不进行设置，则默认前者方法，这也是该过程与LOGISTIC过程的主要区别。因此，在实际研究中，利用样本数据对总体进行统计推断时，必须对样本的设计类型加以考虑，不然即使样本量足够大，也会导致错误的推断结论[7]。

本文通过实例研究，按照不同的分析策略分别对结果变量为多值名义变量的分层整群抽样数据构建广义logit回归模型，通过对结果的解释和比较，发现在对复杂抽样调查设计多值名义资料进行多重logistic回归分析时，既要考虑抽样设计，又要兼顾抽样权重，以得到更准确的分析结果。

参考文献

[1] 胡良平. 面向问题的统计学——(2)多因素设计与统计分析[M]. 北京：人民卫生出版社，2012：505-507.

[2] SAS Institute Inc . SAS/STAT® 15.1 User's Guide[M]. Cary, NC：SAS Institute Inc，2018：9681-9739.

[3] McFadden D. Conditional logitanalysis of qualitative choice behavior[M]. In Frontiers in Econometrics. New York：Academic Press，1974：105-142.

[4] Agresti A. Categorical data analysis[M]. 2nd edition. New York：John Wiley & Sons，2002：165-210.

[5] 孙日扬. 复杂随机抽样数据的多重线性与多重logistic回归分析方法及其应用[D]. 北京：中国人民解放军军事医学科学院，2015.

[6] 崔壮，胡良平. 复杂调查资料的特点与统计分析方法概述[J]. 四川精神卫生，2017，30(5)：410-414.

[7] 刘建华，金水高. 复杂抽样调查总体特征量及其方差的估计[J]. 中国卫生统计，2008，25(4)：377-379.

（收稿日期：2019-11-19）

（本文编辑：吴俊林）