复杂调查资料的特点与统计分析方法概述

崔 壮[1],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.军事医学科学院生物医学统计学咨询中心,北京 100850;
3.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
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 复杂抽样是在抽样过程中采用除一阶段单纯随机抽样外,其他抽样方法或其组合的抽样方案。本文对复杂抽样资料的特点、基于复杂调查资料进行差异性分析、多重回归分析以及进行生存资料多重回归分析的要点进行宏观概述。为科研工作者进行复杂抽样资料的分析提供参考和借鉴。

【关键词】 复杂调查;特点;抽样权重;统计分析技术;多重回归分析
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2017.05.004

Overview for the features of complex survey data and its analytical techniques

Cui Zhuang[1], *Hu Liangping*[2,3*]

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

【**Abstract**】 Complex sampling is the sampling plan of other sampling methods or their combination, except a simple random sampling of one stage in the process of sampling. This paper presented a macro overview of the characteristics of complex sampling data, the main points of the difference analysis and multiple regression analysis based on the complex survey data, and the key points of multiple regression analysis of survey survival data. The paper could provide references for the researchers to better understand and implement the analysis of complex sampling data.

【**Keywords**】 Complex survey; Feature; Sampling weight; Analytical techniques of statistics; Multiple regression analysis

1 复杂抽样资料的特点

1.1 何为复杂抽样

复杂抽样是指在抽样过程中采用除一阶段单纯随机抽样外,其他抽样方法或其组合的抽样方案,通过复杂抽样完成的调查称为复杂调查[1]。复杂抽样通常具有分层、整群、不等概率或多阶段设计等方法,其产生的样本称为复杂样本。复杂抽样有以下优点:节省人力物力,使大规模调查更具可行性;可灵活调整样本量在各级抽样单位中的分配;可通过改变抽样比来提高子总体的代表性和估计的可靠性。因此,目前在社会科学领域以及卫生领域调查研究中[2],尤其是大规模调查,一般涉及多地区或多中心的抽样问题,由于单纯随机抽样因调查对象过于分散、成本高且可行性较低[3],故选择复杂抽样设计。

1.2 分析复杂抽样资料的困难

复杂随机抽样中每个阶段的抽样方法不一定相同,其抽样误差的计算随着抽样阶段及抽样方法的增多变得极为复杂。然而,研究者在统计分析时,常忽略之前采取的抽样设计方法,将资料均视为来自单纯随机抽样设计下获得的资料来处理。实际上,在不同抽样率下得到的等量样本量的样本数据所包含的信息是不同的,即“抽样权重”不同[4]。有研究[5]显示,对分层抽样获得的复杂调查数据进行列联表的卡方检验、构建 *OR* 的 95% 置信区间时,若忽视分层,会导致过于保守的检验(*P* 值偏大),*OR* 的置信区间通常也会变宽;而对于整群抽样,通常会产生相反的影响,若忽视整群效应,会获得偏小的 *P* 值和更窄的置信区间,而事实上的置信区间并非如此精确。

1.3 需要引入权重

文献[4]认为,在抽样调查研究中将观测对结果的贡献程度考虑在内,在分析中应考虑抽样权重和观测权重,同时也提出了综合权重的概念。研究显示纳入综合权重的结果更加灵敏且准确、稳健。

观测权重是基于综合评价中权重系数的思想,

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

在回归分析中引入反映每个个体或观测对总体的重要性的度量,表示在其他观测不变的情况下,该观测的变化对结果的影响程度。常用的有经验权重法、试验次数权重法和贡献权重法等[4]。

抽样权重是在抽样研究中,为反映所抽取的样本中各个观测在总体中的重要程度,或样本中各个观测代表总体中个体的数目。抽样权重的大小与抽样方法有关,分为基础抽样权重、调整抽样权重与总抽样权重[4]。

综合权重是在对随机抽样所得的数据进行统计分析时,不仅考虑抽样权重,还将观测权重考虑在内,计算各个观测对结果总的重要程度。其计算方法是:综合权重 = 观测权重 × 抽样权重[4]。

但是,随着抽样率的变化和样本的不同,同一个观测对模型拟合的贡献是不同的。而对于不同的抽样率和样本中同一个观测的观测权重应当是不同的。同时,基于观测权重得到的综合权重也应当随样本的变化而变化。因此,如何动态地计算观测权重与综合权重仍需进一步研究[4]。

1.4 如何准确估计抽样误差

实际研究中,大多数大规模的样本并非通过简单随机抽样获得的,或通过分层减小方差,对感兴趣的领域进行估计,或通过分群来降低成本。在复杂调查中,采用复杂的抽样方式获得的数据通常不是独立的,并且每个样本被抽到的概率是不相等的。但标准统计软件一般是按假定观测单位是独立等分布的条件下编写的分析程序,可以给出均值等统计量的正确估计,但这时标准误、置信区间和假设检验往往是不正确的,在计算时并未考虑抽样设计[6],如直接采用 SAS 中的 SUMMARY、FREQ、MEANS、REG 等标准统计分析过程来分析复杂抽样数据会导致统计推断错误。目前,SAS 9.0 或更高的版本可以通过 SURVEYMEANS、SURVEYFREQ、SURVEYREG、SURVEYLOGISTIC 和 SURVEYPHREG 等过程进行复杂调查资料的分析[7]。

1.5 复杂调查中方差的估计方法

在复杂抽样中,抽样权重包含了构造点估计所需的全部信息,但它不包含标准误估计的任何信息,因此仅仅知道抽样权重并不能进行统计推断。统计量的方差取决于任何一个单元的入选概率,因此需要除抽样权重以外更多关于抽样设计的信息。对于复杂调查中方差的估计方法,主要包括线性化、随机组、重抽样以及广义方差函数等[8]。

Taylor 级数线性近似法(Taylor Series Linearization, TSL):复杂调查方差估计中的理论特性是被研究得最透彻、最常采用的方法,其基本思想是利用 Taylor 级数方法将非线性统计量线性化,然后计算方差的估计值[9-10]。但计算过于繁琐,在包含权数的复杂函数中难以应用,对估计的每个非线性统计量都需一个单独的方差计算公式,还需要进行专门的设计,每个统计量的计算方法都不同。准确度取决于样本量,样本量不够大,方差的估计通常偏低。

重抽样法:分层多阶段抽样中采用重抽样方法,通过从完整样本中抽取子样本计算估计值,避免了求偏导数的过程。主要包括平衡重复复制法(Balanced Repeated Replication, BRR)、刀切法(Jackknife Repeated Replication, Jackknife)和 Bootstrap 法。

Jackknife 法:基本思想是将总体分成 k 组,每次抽取时从中去掉一组,得到的多个二次抽样样本,每个二次样本可得到一个均数或者率的估计值,根据估计值的差异估计方差[11],属于较为全能的方法。每层多于两个群组的分层多阶段抽样中,BRR 法不再适用,Jackknife 法则有较好的表现。对于某些统计量估计方差结果不佳,如简单随机抽样中分位数的方差估计效果较差。

BRR 法:基本思想是假设总体分成 L 层,从每层随机抽取两个样本单位,共抽取 2L 次,产生 2L 个样本,得到多个均数或率的估计值,利用多个估计值的差异估计方差[8,12]。BRR 几乎可应用于所有统计量,但通常只能用于每层只有两个群组(PSU)或能转化为每层有两个 PSU 的设计。与 Jackknife 法和 Bootstrap 法相比,BRR 法计算量相对较小。抽样设计在每层中有两个群组,估计的是有放回抽样的方差,可能会高估方差。

Bootstrap 法:适用于通常抽样设计中的非光滑函数(如分位数),但计算量大于 Jackknife 和 BRR。

2 基于复杂调查资料进行差异性分析的要点

2.1 SURVEYFREQ 过程简介

PROC SURVEYFREQ 根据获得的调查数据的抽样设计计算误差估计值,调查设计可以是一个复杂的抽样调查,如分层抽样、整群抽样以及不平衡加权,PROC SURVEYFREQ 提供了很多误差估计的方法,包括 TSL、BRR 法和 Jackknife 法。

SURVEYFREQ 过程利用样本调查数据生成单向

到多向频率表和交叉表。这些表包括人口总数、人口比例(总体比例,行和列比例)以及相应的标准误差的估计、置信限度、变异系数和模型的效果评价。

对于单向频率表,PROC SURVEYFREQ 提供了针对抽样设计的调整 Rao – Scott 卡方拟合优度检验;对于双向频率表,PROC SURVEYFREQ 提供了基于行和列之间无关联的检验。这些检验包括 Rao – Scott 卡方检验、Rao – Scott 似然比检验、Wald 卡方和 Wald 对数线性卡方检验。

以下语句说明了 PROC SURVEYFREQ 的用法:

```
PROC SURVEYFREQ < options > ;
BY variables ;
CLUSTER variables ;
REPWEIGHTS variables < / options > ;
STRATA variables < / option > ;
TABLES requests < / options > ;
WEIGHT variable ;
```

PROC SURVEYFREQ 语句调用该过程,识别要分析的数据集,并指定方差估计方法。PROC SURVEYFREQ 语句是必需的。TABLES 语句指定频率或交叉表,以及这些表的统计量和检验结果。STRATA 语句列出了在分层设计中的分层变量。CLUSTER 语句指定在整群设计中的群组变量。WEIGHT 语句指定抽样权重变量。REPWEIGHTS 语句指定经过 BRR 法或者 Jackknife 法估计误差后的重新加权变量,BY 语句对以 BY 变量分组的各个亚族分别进行完全独立的分析。

SURVEYFREQ 与 FREQ 过程的不同点主要体现在 PROC SURVEYFREQ 后可以根据需要选择不同的误差估计方法,比如 VARMETHOD = TAYLOR, VARMETHOD = BRR, VARMETHOD = BRR (fay = c)(c 是一个相关系数), VARMETHOD = JACKKNIFE, 并且可以使用 CLUSTER 语句、REPWEIGHTS 语句、STRATA 语句。

2.2 SURVEYMEANS 过程简介

SURVEYMEANS 过程通过计算调查资料的统计量来估计调查人群的特征。通过该过程可以估计均数、合计、百分位数、四分位数间距。PROC SURVEYMEANS 也可以进行域分析,即对一个亚人群或者区域进行估计。该过程也可以估计误差、置信区间以及进行 t 检验。PROC SURVEYMEANS 运用基于复杂抽样设计的 TSL 或者运用 BRR 来估计抽样误差,该过程适用于复杂抽样过程如分层抽样、整群抽样和不平衡加权抽样设计。

以下语句说明了 PROC SURVEYMEANS 的用法:

```
PROC SURVEYMEANS < options > < statistic – keywords > ;
BY variables ;
CLASS variables ;
CLUSTER variables ;
DOMAIN variables < variable_variable variable_variable_variable ...
> < / option > ;
RATIO < 'label' > variables / variables ;
REPWEIGHTS variables < / options > ;
STRATA variables < / option > ;
VAR variables ;
WEIGHT variable ;
```

PROC SURVEYMEANS 选择输入要分析的数据集,指定要计算的统计量以及误差估计方法。VAR 语句指定要分析的变量。CLASS 语句指定要被分析数值变量转换为分类变量。STRATA 语句列出在分类设计中进行分类的变量。CLUSTER 语句指定在整群设计中群组变量。DOMAIN 语句列出域分析或者亚人群分析的变量,RATIO 语句指定要进行率分析的均数或者百分位数,WEIGHT 语句指定抽样权重变量,REPWEIGHTS 语句指定经过 BRR 或者 Jackknife 法估计误差后的重新加权变量,BY 语句对以 BY 变量分组的各个亚族分别进行完全独立的分析。

SURVEYMEANS 与 MEANS 过程的不同点主要体现在 PROC SURVEYMEANS 后可以根据需要选择不同的误差估计方法,比如 VARMETHOD = TAYLOR, VARMETHOD = BRR, VARMETHOD = BRR (fay = c)(c 是一个相关系数),并且可以使用 CLUSTER 语句、DOMAIN 语、REPWEIGHTS 语句和 STRATA 语句。

3 基于复杂调查资料进行多重回归分析的要点

3.1 SURVEYREG 过程简介

PROC SURVEYREG 过程可以对调查资料的数据进行回归分析。该过程可以处理复杂的抽样设计资料包括分层设计、整群设计和不平衡加权数据。该过程适用于符合线性模型的测量数据,并计算回归系数以及变量 – 协变量矩阵。该过程还为模型效应和模型参数的任何指定的可估线性函数提供了假设检验。利用回归过程可以计算样本调查数据的预测值。PROC SURVEYREG 基于广义最小二乘估计法采用逐步法估计回归系数,该过程假定回归系数在不同层和基本抽样单元上是不变的。为了估计回归系数的方差 – 协方差矩阵,PROC SURVEYREG

过程运用基于复杂抽样设计的 TSL 或者运用 BRR 估计抽样误差。

以下语句说明了 PROC SURVEYREG 的用法：

```
PROC SURVEYREG < options > ;
BY variables ;
CLASS variables ;
CLUSTER variables ;
CONTRAST 'label' effect values < ... effect values > < / options > ;
DOMAIN variables < variable_variable variable_variable_variable ... > ;
EFFECT name = effect - type ( variables < / options > ) ;
ESTIMATE < 'label' > estimate - specification < / options > ;
LSMEANS < model - effects > < / options > ;
LSMESTIMATE model - effect lsmestimate - specification < / options > ;
MODEL dependent = < effects > < / options > ;
OUTPUT < keyword < = variable - name > ... keyword < = variable - name > > < / option > ;
REPWEIGHTS variables < / options > ;
SLICE model - effect < / options > ;
STORE < OUT = > item - store - name < / LABEL = 'label' > ;
STRATA variables < / options > ;
TEST < model - effects > < / options > ;
WEIGHT variable ;
```

语句 PROC SURVEYREG 和语句 MODEL 是必需的，如果模型包含分类效应，则必须采用 CLASS 语句来对变量进行分类，并且 CLASS 语句一定要位于 MODEL 语句之前，如果还要使用 CONTRAST 语句或者 ESTIMATE 语句，则 MODEL 语句一定要在 CONTRAST 语句或者 ESTIMATE 语句之前。语句 CLASS、CLUSTER、CONTRAST、EFFECT、ESTIMATE、LSMEANS、LSMESTIMATE、REPWEIGHTS、SLICE、STRATA、TEST 可以多次使用，而语句 MODEL、WEIGHT、STORE、OUTPUT 只能使用一次。CLASS 语句指定分层变量，CLUSTER 语句指定整群设计中群组变量，DOMAIN 语句指定域分析的变量，MODEL 语句指定响应变量和协变量，REPWEIGHTS 语句指定经过 BRR 法或者 Jackknife 法估计误差后的重新加权变量。

SURVEYREG 与 REG 过程的不同点主要体现在 PROC SURVEYREG 过程后可以根据需要选择不同的误差估计方法，比如 VARMETHOD = TAYLOR，VARMETHOD = BRR，VARMETHOD = BRR (fay = c)（c 是一个相关系数），并且可以使用 CLUSTER 语句、DOMAIN 语句、STRATA 语句。

3.2 SURVEYLOGISTIC 过程简介

SURVEYLOGISTIC 过程基于最大似然法对离散响应测量数据的线性逻辑回归模型进行拟合。对于统计推断，SURVEYLOGISTIC 适用于分层抽样、整群抽样和不平衡加权抽样得到的数据进行统计分析。用 Fisher 评分算法或者 Newton - Raphson 算法来进行最大似然估计，并且可以为参数估计指定初始值，在 ordinallogistic 回归中可以用 probit 函数或 log - log 函数来替换 logit 函数，作为连接函数。优势比的估计值可以和参数估计一起显示，并且可以根据需要自行指定所需的解释变量。回归参数的误差和优势比的计算一般采用基于复杂抽样设计的 TSL 或 BRR 进行估计。

以下语句说明了 PROC SURVEYLOGISTIC 的用法：

```
PROC SURVEYLOGISTIC < options > ;
BY variables ;
CLASS variable < (v - options) > < variable < (v - options) > ... > < / v - options > ;
CLUSTER variables ;
CONTRAST 'label' effect values < ,. . . effect values > < / options > ;
DOMAIN variables < variable_variable variable_variable_variable ... > ;
EFFECT name = effect - type ( variables < / options > ) ;
ESTIMATE < 'label' > estimate - specification < / options > ;
FREQ variable ;
LSMEANS < model - effects > < / options > ;
LSMESTIMATE model - effect lsmestimate - specification < / options > ;
MODEL events/trials = < effects < / options > > ;
MODEL variable < (v - options) > = < effects > < / options > ;
OUTPUT < OUT = SAS - data - set > < options > < / option > ;
REPWEIGHTS variables < / options > ;
SLICE model - effect < / options > ;
STORE < OUT = > item - store - name < / LABEL = 'label' > ;
STRATA variables < / option > ;
< label: > TEST equation1 < , ... , equationk > < / options > ;
UNITS independent1 = list1 < ... independentk = listk > < / option > ;
WEIGHT variable ;
```

语句 CLASS、CLUSTER、CONTRAST、EFFECT、ESTIMATE、LSMEANS、LSMESTIMATE，REPWEIGHTS、SLICE、STRATE、TEST 可以在程序中出现多次，而语句 MODEL、WEIGHT、STORE、OUTPUT、UNITS 只能用一次，并且 CLASS 语句必须在 MODEL 语句之前出现使用，CONTRAST 语句必须位于 MODEL 语句之后。BY 语句指定分组变量，CLASS 语句指定分层变量，CLUSTER 语句指定整群设计中群组变量，DOMAIN 语句指定域分析的变量，MODEL 语句指定

响应变量和协变量,REPWEIGHTS 语句指定经过 BRR 法或 Jackknife 法估计误差后的重新加权变量。

SURVEYLOGISTIC 和 LOGISTIC 过程的不同点主要体现在 SURVEYLOGISTIC 后可以根据需要选择不同的误差估计方法,比如 VARMETHOD = TAYLOR, VARMETHOD = BRR, VARMETHOD = BRR (fay = c)(c 是一个相关系数),并且可以使用 DOMAIN 语句、REPWEIGHTS 语句。

4 基于复杂调查资料进行生存资料多重回归分析的要点

SURVEYPHREG 过程执行基于 Cox 比例风险模型的抽样调查数据的回归分析。当有合适的解释变量可用时,Cox 的半参数比例风险回归模型被广泛应用于分析生存数据,并估计危险率,该过程提供基于复杂抽样设计资料的误差估计以及置信区间、有关参数和模型效应的假设检验。SURVEYPHREG 提供了几种优化的技术以最大限度地提高对数似然值,风险比可以和参数估计一同计算得到,回归参数的抽样误差和风险比可以通过基于复杂抽样设计的 TSL 或者运用 BRR 估计得到。

以下语句说明了 PROC SURVEYPHREG 的用法:

```
PROC SURVEYPHREG < options > ;
BY variables ;
CLASS variable < (options) > < . . . variable < (options) > > < /options > ;
CLUSTER variables ;
DOMAIN variables < variable_variable variable_variable_variable . . . > ;
ESTIMATE < 'label' > estimate - specification < / options > ;
FREQ variable ;
LSMEANS < model - effects > < / options > ;
LSMESTIMATE model - effect lsmestimate - specification < / options > ;
MODEL response < * censor(list) > = effects < /options > ;
NLOPTIONS < options > ;
OUTPUT < OUT = SAS - data - set > < keyword = name . . . keyword = name > < /options > ;
REPWEIGHTS variables < / options > ;
SLICE model - effect < / options > ;
STRATA variables < /option > ;
STORE < OUT = > item - store - name < / LABEL = 'label' > ;
TEST < model - effects > < / options > ;
WEIGHT variable ;
```

语句 PROC SURVEYPHREG 和 MODEL 是必需的,并且 CLASS 语句必须在 MODEL 语句之前出现,MODEL 语句指定要分析模型,CLASS 语句指定进行分类的变量,STRATA 语句指定分层变量,CLUSTER 语句指定在整群设计中群组变量,WEIGHT 语句指定抽样权重变量,NLOPTIONS 语句指定优化算法,REPWEIGHTS 语句指定经过 BRR 法或者 Jackknife 法估计误差后的重新加权变量,DOMAIN 语句罗列出进行亚人群或者域分析的变量,BY 语句指定变量分组后分别进行分析。

SURVEYPHREG 和 PHREG 过程的不同点主要体现在 PROC SURVEYPHREG 后可以根据需要选择不同的误差估计方法,比如 VARMETHOD = TAYLOR, VARMETHOD = BRR, VARMETHOD = BRR (fay = c)(c 是一个相关系数),并且可以使用 DOMAIN语句、REWEIGHTS 语句、NLOPTIONS 语句。

参考文献

[1] 姜博,王丽敏,刘艳,等. 复杂抽样数据统计分析方法回顾[J]. 中国卫生统计,2015,32(4):721 - 723, 726.

[2] 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.

[3] Anderson KM, Wilson PW, Odell PM, et al. An updated coronary risk profile. A statement for health professionals[J]. Circulation, 1991, 83(1): 356 - 418.

[4] 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.

[5] Sharon L. Sampling: Design and Analysis[M]. Boston: Thomson Brooks Cole, 2009: 291 - 355.

[6] SAS Institute Inc. SAS /STAT 9.3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 7207 - 7547.

[7] 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, 2008, 25(6): 577 - 579.

[8] 王晓荣, 赵俊康, 王彤. 复杂抽样下的截取回归模型在医学研究中的应用[J]. 中国卫生统计, 2012, 29(5): 691 - 697.

[9] 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.

[10] West BT. Statistical and methodological issues in the analysis of complex sample survey data: practical guidance for trauma researchers[J]. J Trauma Stress, 2008, 21(5): 440 - 447.

[11] KrewskiD, Rao JNK. Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods[J]. Ann Stat, 1981, 9(5): 1010 - 1019.

[12] 吕萍. 重权数在复杂调查的方差估计中的应用[J]. 统计研究, 2011, 28(2): 93 - 99.

(收稿日期:2017 - 08 - 17)

(本文编辑:陈 霞)