正态线形模型下缺失值的贝叶斯多重插补

——基于柑橘数据的分析

潘传快[1,2],熊　巍[2],祁春节[2]

(1.武汉纺织大学 经济学院,湖北 武汉 430073;
2.华中农业大学 经济管理学院,湖北 武汉 430070)

摘　要　缺失值是调查中普遍存在的问题,利用变量之间的相关关系,可以通过正态线形模型利用不存在缺失值的变量对存在缺失值的变量进行插补。较之单一插补,多重插补更能有效地估计总体方差,因此更多地被使用;特别是采用贝叶斯多重插补,其模型的差数和残差估计均来自相应后验分布的随机抽取,这样对总体方差的估计更为精确。通过大量模拟试验,发现贝叶斯多重插补较之单一插补和一般多重插补能构建更宽的置信区间从而有更准确的总体参数覆盖率,这点在数据缺失比重很大时优势更明显。

关键词　缺失值;贝叶斯;多重插补;模拟;正态线性模型

中图分类号:F 222　　**文献标识码**:A　　**文章编号**:1008-3456(2017)01-0072-06

DOI编码:10.13300/j.cnki.hnwkxb.2017.01.009

调查数据中存在缺失值是极为常见的问题,受访者对调查问题无法做出回答或者不愿做出回答就产生了缺失值,甚者缺失值本身就是变量样本空间的一个点[1]。比如在对柑橘种植户的调查中,种植户并不一定统计过自己家庭一年的收入或者不愿意回答,那么在家庭收入这个变量上就会有缺失值。或者农户由于文化程度偏低,对问卷错误理解,从而使问卷失效[2]。处理缺失值首先能想到的方法是删除含缺失值的所有单元数据(完全个案删除),大部分统计软件(诸如SPSS、STATA、R)也默认这种处理方法。如果缺失个别数据,这种方法是可取的,在有些场合下甚至是更优的选择[3]。但是如果大量数据缺失,这种删除会抛弃很多数据;有些时候数据看起来缺失的不多,但由于存在大量的变量,对个别变量极小比例的缺失如果采用完全个案删除综合起来都会引致大量数据的删除[4-5]。对缺失数据进行插补被认为是一个更可取的方法,因为插补不会损失原有的数据信息,而且如果插补方法得当,还可以对数据信息进行有效的补充[1,5]。有效利用变量之间的相关关系,构建线形模型,利用不存在缺失值的变量对存在缺失值的变量进行插补,而这种插补带有预测性,这样就能对数据信息进行一定的补充。比如柑橘种植户虽然无法或不愿回答家庭收入,但对自己的柑橘产量是记得很清楚而且原意回答的,这样柑橘产量这个变量往往不会有缺失,而我们相信种植户的家庭收入跟柑橘产量相关性是很高的,我们就可以借助线形模型以柑橘产量为辅助变量来插补家庭收入。简单的插补是单一插补,即利用数据的未缺失部分估计线形模型的参数,然后借助该模型对缺失值进行点估计;但单一插补往往低估了变量的方差,容易产生一个过窄的置信区间或者太显著的检验统计量(容易拒绝零假设)[6]。多重插补能有效解决这个问题,多重插补不产生单一的插补值,而是随机产生多个插补值,形成多个"完整"的数据,然后利用这多个数据汇总成一个估计量,以不同插补值之间的差异来弥补单一插补低估总体方差的不足[6-7]。普通的多重插补为在单一插补的基础上加上一个随机干扰项,但这种插补仍然低估了总体方差,因为缺失值的存在使线形模型的参数也不是确定的[3]。因此使

收稿日期:2016-07-26
基金项目:国家社会科学基金项目"改革农产品价格形成机制研究"(16BJY136);国家现代农业(柑橘)产业技术体系(MATS)专项经费资助项目(CARS-27-08B);2015年华中农业大学研究生课程建设项目"中级计量经济学"(2015KJ15)。
作者简介:潘传快(1979-),男,讲师,博士研究生;研究方向:统计分析、农业经济。

用随机的线形模型参数进行多重插补是更好的选择,参数的随机抽取有两种主要的方法:贝叶斯法和重抽样方法(Jacknife 法和 Bootstrap 法)[8-9]。贝叶斯方法是指模型的参数以及残差的产生来自它们相应后验分布的随机抽取[10-11]。图 1 展现了对同一缺失值不同方法的插补结果,作为比较加入了随机插补(a):插补值来自未缺失数据的一个随机抽取。观察发现多重插补对同一缺失值得到不同的插补值,利用不同插补值之间的差异来弥补对总体差异的低估;而贝叶斯多重插补由于模型参数也是随机产生的,因此其插补值之间的差异更大。

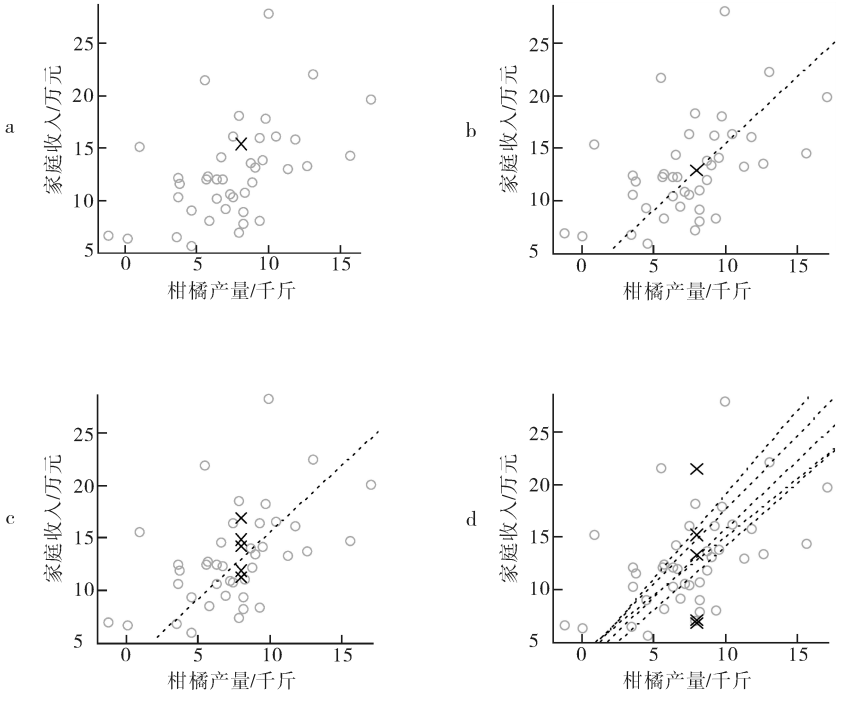

图 1　四种插补方法的比较:(a)随机单一插补; (b)回归单一插补; (c)普通多重插补; (d)贝叶斯多重插补

目前有很多研究已经将贝叶斯多重插补用于实际调查数据[12-13],但调查数据是唯一的不可重复的,无法事先知道缺失值的真值和总体参数,因此根据对调查数据的插补无法获知贝叶斯多重插补对比单一插补和普通多重插补到底有多大的优势。为此,本文拟采取模拟分析的方法,模拟产生缺失数据再进行大量的插补试验以排除偶然性,以期很客观地分析贝叶斯多重插补较之其他插补方法的优势。

一、模型和方法

1.假　设

贝叶斯多重插补模型的构建基于以下假设:

①数据分为 Y 和 X 两部分,其中 Y 来自一元正态总体,存在缺失值;而 X 是来自 q 元正态总体的完整数据不存在缺失值。

②Y 缺失机制为随机缺失(MAR),即 Y 的缺失只跟 X 有关,跟本身无关[5,12]。

③Y 跟 X 之间是线性关系,即:

$$Y\sim N(X\beta,\sigma^2) \tag{1}$$

式(1)中 β 是 q 维向量,而 σ^2 是标量。这样我们就可以构建正态线性模型,利用 X 插补 Y 的缺失部分。

2.贝叶斯多重插补模型

设数据有 n 个观测值,含有缺失值的变量 Y 有 n_1 个观测未缺失,记为 Y_1,有 n_0 个观测缺失,记为 Y_0。不含缺失值的 X(X 为向量或矩阵)也分为两部分:X_1 和 X_0,X_1 为 Y_1 对应的部分,X_0 为 Y_0

对应的部分。根据未缺失数据 Y_1 和 X_1,可以获得式(1)中模型的最小平方估计 $\hat{\beta}_1$ 和 $\hat{\sigma}_1^2$:

$$\hat{\beta}_1=(X_1^TX_1)-1(X_1^TY_1) \tag{2}$$

$$\hat{\sigma}_1^2=(Y_1-X_1\hat{\beta}_1)^T(Y_1-X_1\hat{\beta}_1)/(n_1-q) \tag{3}$$

贝叶斯多重插补的重点是利用 $\hat{\beta}_1$ 和 $\hat{\sigma}_1^2$ 为 β 和 σ^2 构建后验分布,其中 σ^2 可构建 χ^2 分布如下:

$$\frac{\hat{\sigma}_1^2}{\sigma^2}(n-q)\sim\chi^2(n-q) \tag{4}$$

在给定 σ^2 的情况下,β 的后验分布为:

$$\beta|\sigma^2\sim N(\hat{\beta}_1,\sigma^2\ (X_1^tX_1)^{-1}) \tag{5}$$

根据 β 和 σ^2 的后验分布,对缺失值 Y_0 的插补使用模型:

$$Y_*=X_0\beta_*+Z_1\sigma_* \tag{6}$$

式(6)中 Z_1 是来自标准正态分布 $N(0,1)$ 的 n_0 个随机抽取,σ_* 和 β_* 分别来自式(4)和式(5)的一个随机抽取。其中:

$$\sigma_*^2=\hat{\sigma}_1^2(n_1-q)/g \tag{7}$$

式(7)中 g 是来自 $\chi^2(n_1-q)$ 的一个随机抽取。则:

$$\beta_*=\hat{\beta}_1+\sigma_*\ ((X_1^TX_1)^{-1})^{1/2}Z_2 \tag{8}$$

式(8)中,$((X_1^TX_1)^{-1})^{1/2}$ 指 $(X_1^TX_1)^{-1}$ 的方根(Cholesky 分解),Z_2 是来自标准正态分布 $N(0,1)$ 的 q 个随机抽取。

为了消除奇异矩阵问题,式(8)中的 $(X_1^TX_1)^{-1}$ 可以用下式来代替:

$$V=((X_1^TX_1)^{-1}+diag((X_1^TX_1)^{-1})\kappa)^{-1} \tag{9}$$

式(9)中 κ 是一个岭参数(ridge parameter),一般为一个接近于零的正数(比如:0.000 01)。

3.**参数估计**

贝叶斯插补后会得到 m 个完整的数据,需要用这 m 个数据完成估计或检验。其中均值和(内部)方差点估计分别为:

$$\bar{Y}=\frac{\sum_{l=1}^{m}\bar{Y}_l}{m}\text{和}\bar{S}^2=\frac{\sum_{l=1}^{m}S_l^2}{m} \tag{10}$$

而插补后的总方差 T 被分解成两部分:数据内部的方差和数据之间的方差:

$$T=\bar{S}^2+(1+\frac{1}{m})B \tag{11}$$

式(11)中,B 表示 m 个完整数据之间的方差,计算公式为:

$$B=\frac{\sum_{l=1}^{m}(\bar{Y}_l-\bar{Y})^2}{m-1} \tag{12}$$

如此,μ 的一个 $100(1-\alpha)\%$ 置信区间为:

$$\bar{Y}\pm t_v(\alpha/2)\sqrt{T} \tag{13}$$

式(13)中 v 为学生-t 分布的自由度,计算公式[5]为:

$$v=(m-1)(1+\frac{1}{r})^2 \tag{14}$$

式(14)中 r 表示因数据缺失相对增加的方差,计算公式为:

$$r=\frac{(1+\frac{1}{m})B}{\bar{S}^2} \tag{15}$$

4.**缺失信息的判断**

为了判定插补值对总体方差的影响,引入两个指标:λ 值和 γ 值[9]。λ 是指缺失值产生的方差占总方差的比重,计算公式为:

$$\lambda=\frac{B+B/m}{T} \tag{16}$$

而 γ 是指缺失信息比,计算公式为:

$$\gamma=\frac{r+2/(v+3)}{r+1} \tag{17}$$

二、模拟分析

利用计算机进行模拟分析是现代统计的重要分析方法,采取模拟分析有很多优点:因为是模拟,所以可以使整个插补模型的假设条件完全得到满足;模拟分析可以事先知道总体参数、缺失值的真值,这样方便计算插补值的误差;鉴于计算机的强大计算能力,模拟分析可以获取大量样本或者进行大量模拟试验,这样以避免单个样本或者单次模拟试验结果的偶然性,最终得到稳定的结论。本文所有的数据模拟以及分析都基于 R 语言。

1.**数据模拟与插补**

事先设定 X 为来自正态总体 $N(6.8,3.5)$的一个容量为 100 的样本,则:

$$Y\sim N(2.39+1.27X,0.47^2) \tag{18}$$

对式(18)中的 Y 随机缺失 50 个数(缺失比例为 50%),大的缺失比例更容易对比不同的插补方法的优劣。模拟的数据及缺失情况显示在图 2 中,从该图中可看出家庭收入的缺失跟本身无关,只跟柑橘产量有关,符合随机缺失(MAR)的假定。

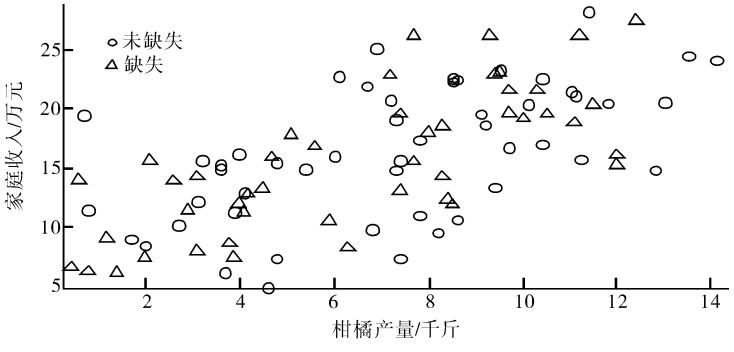

图 2　模拟数据的缺失情况

对缺失数据分别采用单一插补、普通多重插补和贝叶斯多重插补进行 100 次的插补;再将 100 次插补后的参数估计结果求平均数,这样做的目的是避免单次插补结果的偶然性,结果列在表 1 中。

表 1　三种插补方法的估计

插补方法	点估计	标准误	95%置信下限	95%置信上限
单一插补	1.269 246	0.009 6	1.252 3	1.288 2
普通多重插补	1.269 315	0.017 2	1.233 2	1.305 5
贝叶斯多重插补	1.269 296	0.021 0	1.221 6	1.317 0

2.**比　较**

为了比较三种插补方法的效果,再构建两个指标:偏差和覆盖率。偏差是指每一次插补后的点估计的均值与真实值之差的平均数;覆盖率是指根据每一次插补后的数据构建的置信区间中能包括真实值的置信区间所占的比重。表 2 记录了三种插补方法的偏差和覆盖率,发现三种方法产生点估计的偏差几乎没有区别,都略微较真实值偏小,幅度为 −0.007 左右。但是在覆盖率上,单一插补构建的置信区间只能覆盖 69%的真实值,而普通多重插补明显改善,覆盖了 93%的真实值,但是还是低于

事先设定的95%。而采用贝叶斯多重插补,覆盖率能达到97%,高于事先设定的95%置信水平。为什么会这样呢,表2还计算了每一次插补的置信区间(CI)的平均宽度,发现单一插补的CI宽度是明显偏窄的,这是因为单一插补的结果明显低估了总体方差从而使置信区间宽度变窄;普通多重插补有所改进,但CI宽度仍然偏窄,而贝叶斯多重插补的CI最宽。

表2进一步总结了三种插补方法的缺失信息。单一插补的λ值为0,表示缺失信息比为0,插补值未给总体贡献差异,这显然不符合实际;而普通多重插补的γ值和λ值加大,说明插补的缺失值为总体数据贡献了差异,但仍然偏小;只有贝叶斯多重插补的γ值和λ值接近50%,这与事先设定的50%的缺失比重接近。

表2 三种插补方法的效果比较

插补方法	偏差	覆盖率	CI宽度	γ	λ
单一插补	−0.0007 5	0.69	0.038 0	0.020 2	0.000 0
普通多重插补	−0.0006 8	0.93	0.072 3	0.390 1	0.340 4
贝叶斯多重插补	−0.0007 0	0.97	0.095 5	0.582 2	0.522 1

3.**不同缺失比重对比**

为了验证在不同缺失比重下三种插补方法的效果,把缺失比重设为20%、50%和80%,然后比较三种插补方法构建的100次插补的置信区间的覆盖率。观察发现,多重插补特别是贝叶斯多重插补在缺失比重很高的情况下效果更明显(见表3)。当缺失比重只有20%的时候,即使是单一插补也有85%的覆盖率,而普通的多重插补有93%的覆盖率,跟贝叶斯多重插补接近;当缺失比重上升到50%时,单一插补覆盖率下降为69%,但普通多重插补尚有93%的覆盖率;当缺失比重上升到80%时,单一插补的覆盖率就只有28%,而普通多重插补的覆盖率下降为75%。由此可见,不管缺失比重有多高,贝叶斯多重插补的覆盖率始终保持在97%。

表3 不同缺失比重下三种插补方法的覆盖率

缺失比重/%	单一插补	普通多重插补	贝叶斯多重插补
20	0.85	0.93	0.97
50	0.69	0.93	0.97
80	0.28	0.75	0.97

4.**试验次数对贝叶斯多重插补的影响**

在100次试验中发现贝叶斯插补后真实值的覆盖率大于事先设定的95%,这是否意味着贝叶斯多重插补会通过扩大估计标准误来提高覆盖率?若果真如此,则虽然提高了覆盖率但是估计精确度下降了。为此,将试验次数分别提高到1 000次和10 000次,结果列在表4中。发现当试验次数提高到1 000次覆盖率下降为95%,提高到10 000次,发现进一步下降。说明贝叶斯多重插补进行估计的覆盖率接近95%水平,贝叶斯插补的估计是准确的,并没有扩大置信区间的宽度。同时还发现随着试验次数的增加,估计偏差进一步下降直至接近0,说明贝叶斯多重插补得到了一个无偏有效的点估计。此外γ值和λ值随着试验次数的增加也更接近事先设定的缺失比重50%,说明贝叶斯插补并没有扩大或缩小缺失信息。

表4 不同试验次数下贝叶斯多重插补的效果比较

次数	偏差	覆盖率	CI宽度	γ	λ
100	−0.000 704	0.970	0.095 5	0.582 2	0.522 1
1 000	−0.000 261	0.950	0.092 2	0.548 5	0.490 7
10 000	−0.000 055	0.944	0.092 1	0.549 2	0.491 5

三、结 论

对于缺失值的处理,插补是比删除更好的解决方法,因为插补不会损失任何原始信息。通过对模拟产生的缺失数据进行大量重复地插补,比较后可发现单一插补虽然不会产生偏差,但是极大地低估了总体方差从而对总体参数的区间估计覆盖率偏低,普通的多重插补虽然能改善但是仍然偏低,只有贝叶斯多重插补能有效地解决这个问题,得到准确的覆盖率。进一步试验发现,在缺失比重很大的时候贝叶斯多重插补的效果更为明显。通过增加试验次数,还发现贝叶斯多重插补能得到一个无偏有

效的点估计,以及一个接近事先设定置信水平的区间估计。

但是,正如 Little 等所说的,插补是有风险的,它产生了一个幻觉,让研究者以为得到了一个完整数据,事实上这个数据是虚构的;特别是当需插补的数据背离模型假设时这个风险尤大[1]。此外,多重插补得到的不是一个单一的数据而是多个数据,而且由于插补是随机的,因此得到的插补数据也不是唯一的,每次插补都会得到不同的结果(尽管在 R 语言中可以设定随机数种子让结果固定),这给研究者展示数据带来了困难。因此我们在使用贝叶斯多重插补时,必先检验数据是否符合模型假设,然后明确插补的目的是为了对参数进行估计和检验,否则单一插补乃至删除或许是更好的选择。

参 考 文 献

[1] LITTLE R J,RUBIN D B.Statistical analysis with missing data[M].Hoboken:Wiley John & Sons,2002.

[2] 熊肖雷,李冬梅.农户参与农业标准化生产意愿的影响因素——基于四川种植业农户的调查与实证[J].华中农业大学学报(社会科学版),2014(6):51-57.

[3] ALLISON P D.Missing data[J].Thousand Oaks ca sage quantitative applications in the social sciences,2002,17(9):285-314.

[4] ALLISON P D.Estimation of linear models with incomplete data[J].Sociological methodology,1987,17(1):71-103.

[5] RUBIN D B.Inference and missing data[J].Biometrika,1976,63(3):581-592.

[6] RUBIN D B.Multiple imputation for nonresponse in surveys[M].New York:John Wiley & Sons,2004.

[7] YING G,LITTLE R J.Bayesian multiple imputation for assay data subject to measurement error[J].Journal of statistical theory & practice,2013,7(2):219-232.

[8] RAO J N,SHAO J.Jackknife variance estimation with survey data under hot deck imputation[J].Biometrika,1992,79(4):811-822.

[9] SHAO J,CHEN Y.Balanced repeated replication for stratified multistage survey data under imputation[J].Journal of the American statistical association,1998,93(442):819-831.

[10] SI Y,REITER J P.Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys[J].Journal of educational & behavioral statistics,2013,38(5):499-521.

[11] SUN M,BUTAR F B.Bayesian multiple imputation and maximum likelihood methods for missing data[J].Section on research methods,2014(28):3175-3180.

[12] VAN B S.Flexible imputation of missing data[M].Florida:CRC Press,2012.

[13] MENG X L.Multiple-imputation inferences with uncongenial sources of input[J].Statistical science,1994,9(4): 538-558.

(责任编辑:刘少雷)