DOI: 10.13656/j.cnki.gxkx.2009.04.003
广西科学 Guangxi Sciences 2009, 16(4): 400~402, 413

缺失数据下线性模型中缺失值处理方法的比较*

Comparison of Methods to Handle Missing Values in Linear Models with Missing Data

李英华[1],刘 妍[2],秦永松[1]
LI Ying-hua[1], LIU Yan[2], QIN Yong-song[1]

(1.广西师范大学数学科学学院,广西桂林 541004; 2.广西师范大学附属外国语学校,广西桂林 541004)

(1. School of Mathematical Sciences, Guangxi Normal University, Guilin, Guangxi, 541004, China; 2. Foreign Language School Attached to Guangxi Normal University, Guilin, Guangxi, 541004, China)

摘要:在响应变量随机缺失的线性模型中,利用 R统计软件模拟比较完全样本法、固定填补法和分数线性回归填补法得到的回归系数、响应变量均值、响应变量的分布函数、响应变量的分位数估计,并用标准误差 (*SE*)评判其优劣.结果表明,除固定填补法外,无论采用其余哪种方法,随着样本容量的增大,评判值 *SE* 减小,样本容量越大,估计也就越精确;缺失概率的大小也影响估计的精度,缺失概率越大,相应的评判值 *SE* 越大,估计的精度也就越差;另外,在分数线性回归填补法中, $J=5$ 的结果总是比 $J=1$ 的结果好,这说明随着 J 的增大,其估计精度也随着提高.

关键词:线性模型 缺失数据 固定填补 分数填补
中图法分类号: O212 文献标识码: A 文章编号: 1005-9164(2009)04-0400-03

Abstract When the response variable is missing at random in a linear model, three means are considered to handle missing values. They are deleting cases with missing values, deterministic and fractional linear regression imputations. Based on these methods, three estimators are studied for the regression parameters such as the mean, the distribution functions and the quantiles of the response variable. Simulations using statistical R software are conducted to compare the performances of three estimators. The results show that if we use the methods except for the deterministic imputation, the values of *SE* decrease and the estimations are more accurate as the sample sizes increase. We can also see that the values of *SEs* increase and the estimatiors are less accurate as the response probabilities decrease. The estimatiors are more accurate at $J=5$ than that at $J=1$, which shows that the accuracy of the estimators increases as J increases based on the fractional regression imputation.

Key words linear model, missing data, deterministic imputation, fractional imputation

线性模型有很强的实际应用背景,在生物、医学、经济、金融、环境科学及工程技术等领域的数据分析中应用广泛.然而,在许多实际问题中,由于各种人为或其它不可知因素,都容易导致大量缺失数据的产生.从而线性模型在缺失数据下的研究成为焦点.

缺失数据的处理往往采用简单的成对删除,这很容易实现,在少量缺失数据时也是可行的,但是可能导致严重的估计偏差.通常对缺失数据进行填补是比较有效的.文献[1,2]中列举了填补法的主要优点.文献[3,4]用(固定)回归填补方法填补有缺失的线性模型中的数据,但没有考虑其它缺失数据的处理方法.对各种处理方法优劣的比较是一个没有涉及的课题.本文同时考虑缺失值的 3种处理方法:完全样本法、固定填补法和分数线性回归填补法,基于这 3种处理方法得到回归系数、响应变量均值、分布函数、分位数

收稿日期: 2008-11-17
作者简介:李英华(1984-),女,硕士研究生,主要从事数理统计研究.
* 国家自然科学基金项目(10661003),广西科学基金项目(0728092),教育部留学回国人员科研启动基金项目([2004]527)资助.

的 3种估计形式,并用 R统计软件进行数值模拟,分别比较回归系数、均值、分布函数、分位数基于这 3种处理方法得到的估计的优劣.所得结果为实际应用中估计方法的选取提供依据.

1 基于 3种不同处理方法得到的回归系数、均值、分布函数和分位数估计

考虑线性回归模型: $y_i = x_i'U + X, i = 1,\cdots,n$, 其中 x_i 为 p维随机设计向量, y_i 为响应变量, U为 p维未知参数,误差序列 $\{X\}$ 独立同分布,且 $EX = 0, 0 < e^2 = VarX < \infty$.

设有不完全 iid.样本 $\{(x_i, y_i, W), 1 \leqslant i \leqslant n\}$,其中 $\{x_i, 1 \leqslant i \leqslant n\}$ 可全部观测到, $\{y_i, 1 \leqslant i \leqslant n\}$ 有缺失, W为指示 y_i 缺失的变量,即

$$W = \begin{cases} 0, \text{如果 } y_i \text{ 缺失}, \\ 1, \text{如果 } y_i \text{ 不缺失}. \end{cases}$$

用 (x, y, W) 记 $\{(x_i, y_i, W), 1 \leqslant i \leqslant n\}$ 所对应的总体.我们假定 $\{y_i\}$ 满足 MAR缺失机制,即 $P(W = 1 | x, y) = P(W = 1 | x) = P(x)$,也即给定 x之下, y与 W条件独立.

为方便,引入记号:

$$r = \sum_{i=1}^{n} W, m = n - r, s_r = \{i: W = 1, i = 1,\cdots,n\}, s_m = \{i: W = 0, i = 1,\cdots,n\},$$

分别表示没有缺失数据的单元数、有缺失数据的单元数、没有缺失数据的单元集和有缺失数据的单元集.

1.1 基于完全样本下的估计

利用观察到的样本数据对 $\{(x_i, y_i), i \in s_r\}$ 给出回归系数、均值、分布函数、分位数的估计,即

$$\hat{U}_{n1} = \left(\sum_{i=1}^{n} W x_i x_i'\right)^{-1} \sum_{i=1}^{n} W x_i y_i), \quad (1.1)$$

$$\bar{Y}_{n1} = \frac{1}{r}\sum_{i=1}^{n} W y_i, \quad (1.2)$$

$$\hat{F}_{n1}(y) = \frac{1}{r}\sum_{i=1}^{n} \{W I(y_i \leqslant y)\}, \quad (1.3)$$

$$\hat{\theta}_{q,1} = \inf_u \{\hat{F}_{n1}(u) \geqslant q\} = \hat{F}_{n1}^{-1}(q). \quad (1.4)$$

1.2 基于固定补足下的估计

用固定补足方法来填补缺失的响应变量,当 y_i 缺失时,用其预测值来补足,即 $y_{i1}^* = x_i'\hat{U}_{n1}, i \in s_m$,作为 $y_i, i \in s_m$ 的填补值.补足后,记 $\tilde{y}_{i1} = W y_i + (1 - W) y_{i1}^*, i = 1,\cdots,n$,从而基于固定补足后的"完全样本"得到回归系数、均值、分布函数、分位数的估计为

$$\hat{U}_{n2} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i \tilde{y}_{i1}), \quad (1.5)$$

$$\bar{Y}_{n2} = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_{i1}, \quad (1.6)$$

$$\hat{F}_{n2}(y) = \frac{1}{n}\sum_{i=1}^{n} \{W I(y_i \leqslant y) + (1 - W) I(y_{i1}^* \leqslant y)\}, \quad (1.7)$$

$$\hat{\theta}_{q,2} = \inf_u \{\hat{F}_{n2}(u) \geqslant q\} = \hat{F}_{n2}^{-1}(q). \quad (1.8)$$

1.3 基于分数线性回归填补下的估计

利用随机方法,并借鉴于文献 [2]中分数线性回归填补法,即用 $y_{i2}^* = x_i'\hat{U}_{n1} + \bar{X}_1^*, i \in s_m$,作为 $y_i, i \in s_m$ 的填补值,其中 $\bar{X}_1^* = J^{-1}\sum_{l=1}^{J} X_l^*, J \geqslant 1, \{X_l^*, l = 1,\cdots,J\}$ 为从 $\{y_j - x_j'\hat{U}_{n1}, j \in s_r\}$ 中独立重复地抽取的 J个样本.补足后,记 $\tilde{y}_{i2} = W y_i + (1 - W) y_{i2}^*, i = 1,\cdots,n$,从而基于随机补足后的"完全样本"得到回归系数、均值、分布函数、分位数的估计为

$$\hat{U}_{n3} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i \tilde{y}_{i2}\right), \quad (1.9)$$

$$\bar{Y}_{n3} = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_{i2}, \quad (1.10)$$

$$\hat{F}_{n3}(y) = \frac{1}{n}\sum_{i=1}^{n} \{W I(y_i \leqslant y) + (1 - W) J^{-1}\sum_{l=1}^{J} I(x_i'\hat{U}_{n1} + X_l^* \leqslant y)\}, \quad (1.11)$$

$$\hat{\theta}_{q,3} = \inf_u \{\hat{F}_{n3}(u) \geqslant q\} = \hat{F}_{n3}^{-1}(q). \quad (1.12)$$

2 数值模拟

用 R统计软件进行模拟,分别比较 3种方法所得回归系数、均值、分布函数、分位数的优劣.

2.1 模拟数据的选取

(1) 模型: 取 $y_i = x_i U + X$,其中 $U = 1, x_i \sim N(1, 1), X \sim N(0, 1), 1 \leqslant i \leqslant n$.

(2) 响应变量不缺失的概率: 在 MAR和 MCAR缺失机制下考虑两种缺失情形.

情形 1(MAR):

$$P_1(u) = P(W = 1 | x = u) = \begin{cases} 0.8 + 0.2 | u - 1 |, \text{当 } | u - 1 | \leqslant 1 \text{时}, \\ \max\{1 - 0.05 | u - 1 |, 0\}, \text{其它}. \end{cases}$$

此种情况下,响应变量不缺失的期望 $EW = EP_1(x) \approx 0.9$.

情形 2(MCAR):

$P_2(u) = P(W = 1 | x = u) = 0.6$,对任意的 x.此种情况下,响应变量不缺失的概率期望 $EW = EP_2(x) = 0.6$.

(3) 样本容量: $n = 60, 120, 240, 300, 500, 600, 1000$.

(4) 重复次数: $N = 1000$.

2.2 模拟结果的评判准则

用标准误差 (SE) 评判估计的优劣, SE越小说

明此种方法越好,这里 $SE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta)^2}$,其中,$\hat{\theta}_i$表示第 i次估计值,$i=1,\cdots,N$,θ表示真值.

2.3 模拟结果及分析

在模拟中,$SE1$ $SE2$ $SE3$和 $SE4$分别表示基于完全样本的估计、基于固定填补的估计、基于分数线性回归填补($J=1$的单一填补)的估计和基于分数线性回归填补($J=5$的多重填补)的估计的 SE值.

从图 1结果可以看出,无论响应变量的缺失概率如何,随着样本容量的增大,$SE1$和 $SE2$都是最小的,两者相差不大,在样本容量为 60时,$SE2$是最小.这说明在样本容量很小时,基于固定填补得到的回归系数估计是一个比较好的估计,随着样本容量的增大,基于完全样本和基于固定填补都是比较好的估计回归系数的方法.

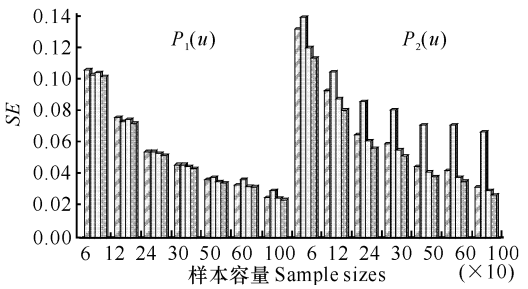

图 1 回归系数的估计

Fig. 1 SE for regression parameters

$SE1$ $SE2$ $SE3$ $SE4$

从图 2结果可以看出,无论样本容量与响应变量的缺失概率如何,$SE2$都是最小的,这说明在此情况下,基于固定填补下的估计是一个较好的估计方法.

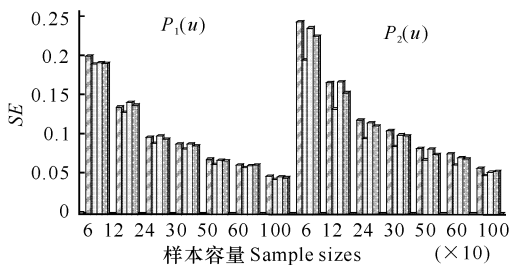

图 2 均值的估计

Fig. 2 SE for mean

$SE1$ $SE2$ $SE3$ $SE4$

从图 3结果可以看出,在 MAR下,样本容量 < 500时,基于不同的处理方法得到的估计相差不大,样本容量> 300时,第二种估计的 SE值越来越大,即此时基于固定填补下的估计是不可取的估计方法;在 MCAR下,基于多重分数线性回归填补下的估计是一个比较好的估计,随着样本容量的增大,采用基于固定填补下的估计的效果显得越来差,甚至比不上基于完全样本下的估计.无论样本容量与缺失概率如何,$SE4$最小,所以在缺失数据下,对分布函数进行估计时,基于多重分数线性回归填补是首选的方法.

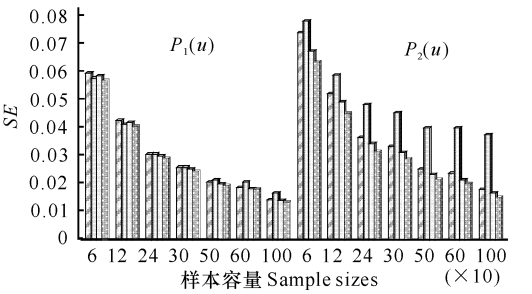

图 3 分布函数在 $y=2$时的估计

Fig. 3 SE for distribution function $y=2$

$SE1$ $SE2$ $SE3$ $SE4$

从图 4结果可以看出,无论响应变量的缺失概率如何,当样本容量> 120时,基于固定填补下的估计的效果越来越差,甚至比不上基于完全样本下的估计.除了在 MCAR下,样本容量为 60的情况外,$SE4$总是保持最小,这说明采用基于多重分数线性回归填补方法得到的分位数估计最接近于真实值,即这种补足法最好.

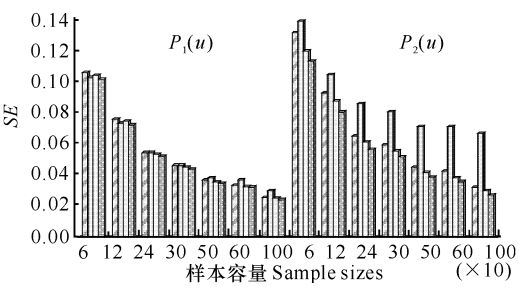

图 4 分位数在 $q=0.25$时的估计

Fig. 4 SE for quantile $q=0.25$

$SE1$ $SE2$ $SE3$ $SE4$

3 结束语

一般地,在数据出现缺失的线性模型中,估计其回归系数时,基于完全样本和基于固定填补方法都是可取的,而基于完全样本的方法简单,易于操作,使用成本低,建议实际应用中在估计回归系数时直接采用基于完全样本的估计方法;估计其均值时,基于固定填补下的估计更接近于真实值;在对分布函数和分位数进行估计时,随机填补方法将会显示出其优越性,基于多重分数线性回归填补方法优于其它方法.

(下转第 413页 Continue on page 413)

3 结束语

从以上的计算机模拟中,我们发现加权网络上舆论传播的一些新特点: 连接条数越大,系统就越容易形成一致意见;权重的引入不利于一致意见的形成;异步更新方式较同步更新方式不利于系统一致意见终态的形成.

参考文献:

[1] Strogatz S H, Watts D J. Collective dynamics of scaling in random networks[J]. Nature, 1998, 393: 440–442.

[2] Sousa A O. Consensus formation on a triad scale-free network[J]. International Journal of Modern Physics C, 2004, 12(10): 1537–1544.

[3] Bernardes A T, Stauffer D, Kertész J. Election results and the sznajd model on barabási network[J]. Eur Phys J B, 2002, 25: 123.

[4] González M C, sousa A O, Herrmann H J. Opinion formation on a deterministic pseudo-fractal network[J]. International Journal of Modern Physics C, 2004, 15(1): 45–57.

[5] Sznajd Weron, Weron R. A simple model of price formation[J]. Int J Mod Phys C, 2002, 13(1): 115–123.

[6] Bernardes A T, Costa U M S, Araujo A D, et al. Damage spreading, coarsening dynamics and distribution of political votes in sznajd model on square lattice[J]. International Journal of Modern Physics C, 2001, 12(2): 159–167.

[7] Reka Albert, Albert-László Barabási. Statistical mechanics of complex networks[J]. Review of Modern Physics, 2002, 74: 47-86.

[8] 方锦清.迅速发展的复杂网络研究与面临的挑战[J].自然杂志, 2005, 27(5): 269–273.

[9] 刘涛,陈忠,陈晓荣.复杂网络理论及其应用研究概述[J].系统工程, 2005, 23(6): 1–7.

[10] Wang Wenxu, Wang Binghong, Bo Hu, et al. General dynamics of topology and traffic on weighted technological networks[C]. Physical Review Letters 94, 2005: 188702.

(责任编辑: 邓大玉)

(上接第 402页 Continue from page 402)

从模拟结果可以看出,样本容量影响估计,除固定填补法外,无论采用哪种方法,随着样本容量的增大,评判值 *SE* 减小,样本容量越大,估计也就越精确;缺失概率的大小也影响估计的精度,缺失概率越大,相应的评判值 *SE* 越大,估计的精度也就越差;另外,在分数线性回归填补法中,*J* = 5的结果总是比 *J* = 1的结果好,这说明随着 *J* 的增大,其估计精度也随着提高.无论采用哪种处理方法,都无法避免主观因素对原系统的影响,并且在缺失值过多的情形下将整个数据集完整化是不可行的.故针对各种实际问题,要注意分清问题的实质,合理地运用各种缺失数据的处理方法.

参考文献:

[1] Brick J M, Kalton G. Handling missing data in survey research[J]. Statist Methods Med Res, 1996, 5: 215–238.

[2] Qin Y, Rao J N K, Ren Q. Confidence interval for marginal parameters under fractional linear regression imputation for missing data[J]. J Multivariate Anal, 2008, 99: 1232–1259.

[3] Wang Q, Rao J N K. Empirical likelihood-based inference in linear models with missing data[J]. Scand J Statist, 2002, 29: 563–576.

[4] Wang Q, Rao J N K. Empirical likelihood for linear regression models under imputation for missing responses[J]. Canad J Statist, 2001, 29: 597–608.

[5] Rubin D B. Multiple imputation after 18+ Years[J]. J of the Amer Statist Assoc, 1996, 91: 473–489.

(责任编辑: 尹 闯)