三种缺失机制下数据模拟方法及其 SAS 实现*

海军军医大学卫生勤务学系军队卫生统计学教研室(200433)　朱荣慧　秦婴逸　吴　骋△

【提　要】 **目的**　研究真实世界研究中连续型结局变量下多个协变量缺失的数据模拟方法,以期为缺失值填补方法的模拟研究提供参考。**方法**　基于多元线性回归模型和三种缺失机制的定义,模拟了存在混杂因素情况下连续型结局变量的完整数据,设置了完全随机缺失、随机缺失、非随机缺失机制下 3 个协变量均缺失一定比例(如 2%)的情境,并采用 SAS 软件实现缺失机制的模拟过程。**结果**　构建了不同缺失机制下多元协变量缺失的模拟数据集。**结论**　提供的数据模拟方法和 SAS 代码实现了三种不同缺失机制的多元缺失数据模拟,可为相关模拟研究提供参考。

【关键词】 真实世界研究　缺失数据　缺失机制　数据模拟　SAS 代码

【中图分类号】 R821.3　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.05.028

近年来,国内外对真实世界研究(real world study, RWS)的关注度日益增加。真实世界研究是指针对预设的临床问题,在真实世界环境下收集与研究对象健康有关的数据(真实世界数据)或基于这些数据衍生的汇总数据,通过分析获得药物的使用情况及潜在获益-风险的临床证据(真实世界证据)的研究过程[1]。真实世界数据的主要来源是日常所收集的各种与患者健康状况和/或诊疗及保健有关的数据,也称既有数据、日常数据(routine data)。由于既有数据缺乏记录、采集、存储等流程的科研数据质量控制,极有可能存在数据不完整、关键变量缺失、记录不准确等问题,对循证应用造成了困难[2]。其中,完整性是指数据信息的缺失程度,包括变量的缺失和变量值的缺失。对于重要的协变量缺失应当在研究设计阶段尽量避免,若由于实际情况确实无法收集到,可将其视为未观测混杂因素,采用工具变量(instrumental variable, IV)等方法进行处理,而对于变量值的缺失则需谨慎对待。不同研究目的获得的数据,其缺失程度、缺失分布、缺失原因和变量值的缺失机制往往不尽相同。当特定研究的数据缺失比例明显超过同类研究的比例时,会加大研究结论的不确定性,此时需要慎重考虑该数据能否支持产生真实世界证据。对缺失原因的详细分析有助于对数据可靠性的综合判断,如果涉及缺失数据的填补问题,应根据缺失机制的合理假设采用恰当的填补方法[2]。

根据所用理论基础不同,缺失值填补方法可分为基于统计学的填补方法(如均值填补、回归填补、期望最大化填补)和基于机器学习的填补方法(如 K 最近邻填补、基于聚类的填补方法、基于神经网络的填补方法)[3]。实际应用中,需根据具体问题选择合理有效的填补方法进行缺失值处理,进而提高数据质量以及后续分析的准确性。因此,在研究缺失值处理方法时,常常需要模拟不同情境下的数据集,如不同缺失机制、样本量大小、缺失比例等,用于评价不同填补方法的填补性能。但目前关于缺失机制模拟方面的文献大多仅涉及一种数据缺失机制的模拟[4-8],提及三种缺失机制模拟方法的文献大多针对临床纵向数据研究[9-11],其普适性不足。故本文以真实世界研究中常出现的多个协变量缺失为例,基于多元线性回归模型和缺失机制的定义进行模拟,并提供 SAS 代码,以期为相关领域研究者提供参考。

数据的缺失机制

定义完整数据 $M=(m_{ij})$,缺失数据指示变量 $R=(R_{ij})$,M_{obs} 为观测到的数据,M_{mis} 为缺失数据,Φ 为未知参数。Rubin 和 Little[12] 根据给定 M 下 R 的条件分布—$f(R|M,\Phi)$ 的特征将数据的缺失机制分为三类:完全随机缺失(missing completely at random, MCAR),随机缺失(missing at random, MAR)和非随机缺失(missing not at random, MNAR)。完全随机缺失是指缺失数据发生的概率不依赖于 M 的观测值,或缺失值,即:$f(R|M,\Phi)=f(M|\Phi)$。随机缺失是指缺失数据发生的概率只依赖于 M 的观测值,而不依赖于缺失值,即:$f(R|M,\Phi)=f(R|M_{obs},\Phi)$。非随机缺失是指缺失数据发生的概率依赖于 M 的缺失值,即:$f(R|M,\Phi)=f(R|M_{mis},\Phi)$,或缺失数据发生的概率既依赖于 M 的观测值又依赖于 M 的缺失值[13],即:$f(R|M,\Phi)=f(R|M_{obs},M_{mis},\Phi)$。

完整数据模拟方法

模拟研究的设置参照 Morris 等提出的 ADEMP 准则[14],采用蒙特卡洛方法进行模拟,样本量设置为 500,即 $n=500$。每次模拟 1000 次。

* 基金项目:国家自然科学基金(82373687);海军军医大学"深蓝"工程"领航"人才培养对象

△通信作者:吴骋,E-mail: wucheng_wu@126.com

模拟完整数据，设置 1 个连续型结局变量 Y，1 个二分类分组变量 T、7 个协变量 $x_1 \sim x_7$，包括：3 个连续型协变量 $x_1 \sim x_3$、2 个二分类协变量 x_4，x_5、1 个无序多分类协变量 x_6、1 个有序多分类协变量 x_7，构建多元线性回归模型：

$$Y=\beta_0+\beta_T T+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\varepsilon$$

其中，x_1、x_2、x_3 服从多变量正态分布，均数为(1,2,3)，协方差矩阵为$\begin{Bmatrix}3 & 2 & 1\\ 2 & 4 & 1\\ 1 & 1 & 5\end{Bmatrix}$；$x_4$、$x_5$服从二项分布 $x_4 \sim Bernoulli(0.5)$，$x_5 \sim Bernoulli(0.3)$；x_6服从 Tabulated 分布 $x_6 \sim Tabulated(0.5,0.3,0.2)$；$x_7$ 为有序变量，4 个等级各占 25%；ε 服从标准正态分布，$\varepsilon \sim (0,1^2)$。

设置 x_2、x_5为分组和结局之间的混杂因素，通过以下二项分布来生成分组变量 T：

$$T \sim Bernoulli\left(\frac{\exp(\alpha_0+\alpha_1 x_2+\alpha_2 x_5)}{1+\exp(\alpha_0+\alpha_1 x_2+\alpha_2 x_5)}\right)$$

其中 $\alpha_1=2$，$\alpha_2=1$，通过设置 α_0 的参数使 $T=1$ 的概率约等于 30%，根据模拟预实验结果，设置 $\alpha_0=-6.5$。

设置$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7)=(1.5,-0.8,0.2,-0.1,0.3,0.5,-1.2)$，$\beta_T=2$，$\beta_0=2.6$，此时 Y 的均数约为 5.0，标准差约为 1.0。

不同缺失机制的数据模拟方法

设置多变量缺失，分别是 x_1、x_5、x_7，不同变量之间的缺失比例为 1:1:1，总变量值缺失比例设置为 6%，假定缺失值相互独立，则每个变量的缺失比例为 2%。设指示变量 R，当 $R=1$ 时，表示变量缺失，当 $R=0$ 时，表示变量未缺失。

1.完全随机缺失

使用 SAS 生成变量z_1、z_2、z_3 使其分别服从 $Uniform(0,1)$分布，z_1、z_2、z_3 的取值范围为(0,1)，分别对变量z_1、z_2、z_3，计算其缺失比例对应的百分位数，即第 2 百分位数$z_1_missrate$、$z_2\_missrate$、$z_3\_missrate$，再根据 z_1、z_2、z_3 是否落在 0 到其第 2 百分位数之间，判断 R_{x1}，R_{x5}，R_{x7}的值。即，当 $0<z_1 \leqslant z_1\_missrate$，$R_{x1}=1$，当 $0<z_2 \leqslant z_2\_missrate$，$R_{x5}=1$，当 $0<z_3 \leqslant z_3\_missrate$，$R_{x7}=1$。

2.随机缺失

建立回归方程，随机缺失依赖于观测到的变量。设 x_1 的缺失与 x_3 和 Y 相关，x_5的缺失与 x_4、T 和 Y 相关，x_7 的缺失与 x_2 和 Y 相关，即：

$$R_{x1} \sim Bernoulli\left(\frac{\exp(\gamma_{x10}+\gamma_{x11} x_3+\gamma_{x12} Y)}{1+\exp(\gamma_{x10}+\gamma_{x11} x_3+\gamma_{x12} Y)}\right)$$

$$R_{x5} \sim Bernoulli\left(\frac{\exp(\gamma_{x50}+\gamma_{x51} x_4+\gamma_{x52} T+\gamma_{x53} Y)}{1+\exp(\gamma_{x50}+\gamma_{x51} x_4+\gamma_{x52} T+\gamma_{x53} Y)}\right)$$

$$R_{x7} \sim Bernoulli\left(\frac{\exp(\gamma_{x70}+\gamma_{x71} x_2+\gamma_{x72} Y)}{1+\exp(\gamma_{x70}+\gamma_{x71} x_2+\gamma_{x72} Y)}\right)$$

其中$(\gamma_{x11},\gamma_{x12})=(0.30,-0.25)$，$(\gamma_{x51},\gamma_{x52},\gamma_{x53})=(-0.45,0.20,0.35)$，$(\gamma_{x71},\gamma_{x72})=(0.55,-0.15)$，通过设置 γ_{x10}、γ_{x50}、γ_{x70}的参数使 $R_{x1}=1$、$R_{x5}=1$、$R_{x7}=1$ 的概率分别约为 2%。

3.非随机缺失

建立回归方程，非随机缺失依赖于观测到的变量及缺失的变量。与随机缺失设置相同，x_1 的缺失与 x_3 和 Y 相关，x_5的缺失与 x_4、T 和 Y 相关，x_7 的缺失与 x_2 和 Y 相关。同时 x_1、x_5、x_7 的缺失还与其自身相关，即：

$$R_{x1} \sim Bernoulli\left(\frac{\exp(\gamma_{x10}+\gamma_{x11} x_3+\gamma_{x12} Y+\gamma_{x13} x_1)}{1+\exp(\gamma_{x10}+\gamma_{x11} x_3+\gamma_{x12} Y+\gamma_{x13} x_1)}\right)$$

$$R_{x5} \sim Bernoulli\left(\frac{\exp(\gamma_{x50}+\gamma_{x51} x_4+\gamma_{x52} T+\gamma_{x53} Y+\gamma_{x54} x_5)}{1+\exp(\gamma_{x50}+\gamma_{x51} x_4+\gamma_{x52} T+\gamma_{x53} Y+\gamma_{x54} x_5)}\right)$$

$$R_{x7} \sim Bernoulli\left(\frac{\exp(\gamma_{x70}+\gamma_{x71} x_2+\gamma_{x72} Y+\gamma_{x73} x_7)}{1+\exp(\gamma_{x70}+\gamma_{x71} x_2+\gamma_{x72} Y+\gamma_{x73} x_7)}\right)$$

其中$(\gamma_{x11},\gamma_{x12},\gamma_{x13})=(0.30,-0.25,0.20)$，$(\gamma_{x51},\gamma_{x52},\gamma_{x53},\gamma_{x54})=(-0.45,0.20,0.35,-0.82)$，$(\gamma_{x71},\gamma_{x72},\gamma_{x73})=(0.55,-0.15,0.36)$，通过设置 γ_{x10}、γ_{x50}、γ_{x70}的参数使 $R_{x1}=1$、$R_{x5}=1$、$R_{x7}=1$ 的概率分别约为 2%。

模拟缺失机制的 SAS 代码实现

1.完全随机缺失

```
/* simulate missing data */
/* 1. MCAR */
%let missrate=2;/* miss rate */
data mcar_pre;
  set mvn;
  by SampleID;
  call streaminit(1234);
  z1=rand("Uniform");/* z1 */
  call streaminit(123);
  z2=rand("Uniform");/* z2 */
  call streaminit(234);
  z3=rand("Uniform");/* z3 */
run;
/* calculate the specified percentile of z1, z2, z3 */
proc univariate data=mcar_pre noprint;
  by SampleID;
  var z1 z2 z3;
  output out=missvalue pctlpts=&missrate pctlpre=z1 z2 z3;
run;
proc sql;
```

```
    create table mcar_pre1 as
    select a.* ,b.* from mcar_pre a
    left join missvalue b
    on a.SampleID=b.SampleID
    ;
  quit;
  /* assign missing values */
  data mcar;
    set mcar_pre1;
  by SampleID;
    if 0<z1<=z1&missrate. then R_x1=1;else R_x1
=0; /* R_x1 */
    if 0<z2<=z2&missrate. then R_x5=1;else R_x5
=0; /* R_x5 */
    if 0<z3<=z3&missrate. then R_x7=1;else R_x7
=0; /* R_x7 */
    if R_x1=1 then x1=.;
    if R_x5=1 then x5=.;
    if R_x7=1 then x7=.;
  run;
  /* check miss rate */
  proc freq data=mcar;
    by SampleID;
    tables R_x1 R_x5 R_x7/list;
  run;
  2.随机缺失
  /* 2. MAR */
  data mar;
    set mvn;
  by SampleID;
    r_x11=0.30;
    r_x12=-0.25;
    r_x10 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x11*3+r_x12*5);
    r_x51=-0.45;
    r_x52=0.20;
    r_x53=0.35;
    r_x50 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x51*0.5+r_x52*0.3+r_x53*5);
    r_x71=0.55;
    r_x72=-0.15;
    r_x70 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x71*2+r_x72*5);
    call streaminit(1234);
    R_x1 = rand("Bernoulli",(logistic(r_x10+r_
x11*x3+r_x12*Y)));/* R_x1 */
    call streaminit(123);
    R_x5 = rand("Bernoulli",(logistic(r_x50+r_
x51*x4+r_x52*T+r_x53*Y)));/* R_x5 */
    call streaminit(234);
    R_x7 = rand("Bernoulli",(logistic(r_x70+r_
x71*x2+r_x72*Y)));/* R_x7 */
    if R_x1=1 then x1=.;
    if R_x5=1 then x5=.;
    if R_x7=1 then x7=.;
  run;
  /* check miss rate */
  proc freq data=mar;
    by SampleID;
    tables R_x1 R_x5 R_x7/list;
  run;
  3.非随机缺失
  /* 3. MNAR */
  data mnar;
    set mvn;
  by SampleID;
    r_x11=0.30;
    r_x12=-0.25;
    r_x13=0.20;
  r_x10 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x11*3+r_x12*5+r_x13*1);
  r_x51=-0.45;
  r_x52=0.20;
  r_x53=0.35;
  r_x54=-0.82;
  r_x50 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x51*0.5+r_x52*0.3+r_x53*5+r_x54*
0.3);
    r_x71=0.55;
    r_x72=-0.15;
    r_x73=0.36;
  r_x70 = log( &missrate * 0.01/(1-&missrate *
0.01))-(r_x71*2+r_x72*5+r_x73*2.5);
    call streaminit(1234);
    R_x1 = rand("Bernoulli",(logistic(r_x10+r_
x11*x3+r_x12*Y+r_x13*x1)));/* R_x1 */
    call streaminit(123);
  R_x5 = rand("Bernoulli",(logistic(r_x50+r_x51
*x4+r_x52*T+r_x53*Y+r_x54*x5)));/* R_x5
*/
    call streaminit(234);
    R_x7 = rand("Bernoulli",(logistic(r_x70+r_
x71*x2+r_x72*Y+r_x73*x7)));/* R_x7 */
    if R_x1=1 then x1=.;
```

```
    if R_x5 = 1 then x5 = .;
    if R_x7 = 1 then x7 = .;
run;
/* check miss rate */
proc freq data = mnar;
    by SampleID;
    tables R_x1 R_x5 R_x7/list;
run;
```

讨 论

本文以真实世界既有数据中普遍存在的多个协变量缺失为例，利用多元线性回归模型模拟完整数据及完全随机缺失、随机缺失和非随机缺失三种不同缺失机制下的缺失数据。

模拟中设置的结局变量为连续型，采用多元线性回归模型模拟完整数据，对于真实世界研究中常出现的二分类结局变量，可采用 logistic 回归模型进行模拟，但需注意的是，logistic 回归不需要设置误差项 ε，同时模拟时 Y 需进行 inverse logit 转换，并根据二项分布生成，SAS 代码可参照文中分组变量 T 的模拟过程。

不同的数据缺失机制根据其定义进行模拟，在模拟非随机缺失时，本文采用的定义是缺失数据依赖于观测到的变量及缺失变量，模拟时也可以根据定义——缺失数据发生的概率依赖于缺失的变量来进行。模拟时可根据该定义对应进行修改，如以变量 x_7 为例，此时 x_7 的缺失与其本身相关，即 $R_{x7} \sim Bernoulli(logit(\gamma_{x70}+\gamma_{x70}x_7))$。

由于篇幅原因，本文未给出多元线性回归模型完整数据集的 SAS 实现代码，其具体实现方法，读者可参阅相关文献[15]。

结合实际需求，读者可基于本研究进一步拓展，如模拟不同的样本量大小、缺失比例、缺失协变量种类、以及在模型中加入交互项和/或高次项等，以评估不同缺失数据填补方法在不同缺失机制和情境下的性能表现，从而为实际应用中如何选择合适的填补方法提供依据。

参 考 文 献

[1] 国家药品监督管理局.真实世界证据支持药物研发与审评的指导原则(试行)[EB/OL]. https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20200107151901190.html.

[2] 国家药品监督管理局药品审评中心.用于产生真实世界证据的真实世界数据指导原则[EB/OL]. https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539.

[3] 赖晓晨,张立勇,刘辉,等. 基于机器学习的数据缺失值填补:理论与方法[M]. 北京:机械工业出版社,2020.

[4] 肖晓椿. 纵向二分类资料缺失数据处理方法的研究与应用[D].上海:中国人民解放军海军军医大学,2020.

[5] 李业棉,赵芃,杨崙惠,等. 队列研究中纵向缺失数据填补方法的模拟研究[J].中华流行病学杂志,2021,42(10):1889-1894.

[6] 王可,杨弘,田晶,等. 基于 Monte Carlo 模拟的完全随机缺失数据处理方法效果比较[J].中国卫生统计,2020,37(2):298-301.

[7] 杨弘,田晶,王可,等. 混合型缺失数据填补方法比较与应用[J].中国卫生统计,2020,37(3):395-399.

[8] 张彪,韩伟,庞海玉,等. 完全随机缺失条件下连续型随机变量数据缺失插补方法的比较研究[J].中国卫生统计,2015,32(4):605-608+612.

[9] 陈丽嫦,衡明莉,王骏,等. 定量纵向数据缺失值处理方法的模拟比较研究[J].中国卫生统计,2020,37(3):384-388.

[10] 吴秋红,张裕青,李国平,等. 不同模型处理纵向缺失数据的模拟研究及应用[J].中国卫生统计,2013,30(6):855-858+861.

[11] 陈丽嫦,衡明莉,王骏,等. 多种缺失机制共存的定量纵向缺失数据处理方法的模拟比较研究[J].现代预防医学,2020,47(20):3684-3687+3697.

[12] Little R, Rubin D. Statistical analysis with missing data[M]. Hoboken: John Wiley & Sons,2002.

[13] Imbens G, Rubin D. Causal inference in statistics, social, and biomedical sciences[M]. Cambridge: Cambridge University Press, 2015.

[14] Morris T, White I, Crowther M. Using simulation studies to evaluate statistical methods[J]. Statistics in medicine, 2017, 38(11): 2074-2102.

[15] Wicklin R. Simulating data with SAS©[M]. Cary, NC: SAS Institute Inc,2013.

(责任编辑:张 悦)