Statistics and Application 统计学与应用, 2023, 12(5), 1376-1384
Published Online October 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.125142

Hans汉斯

二分类变量缺失数据处理方法的比较研究

余雪勤

重庆理工大学理学院，重庆

收稿日期：2023年9月23日；录用日期：2023年10月19日；发布日期：2023年10月26日

摘　要

本文介绍了随机缺失模式下一些常用的插补方法，着重介绍了多重插补法和回归插补法两种方法，并且通过模拟实际案例中的响应变量不同的缺失率进一步探讨了这几种方法的插补效果。结果表明，在缺失率较低的情况下，基于逻辑回归的多重插补与回归插补效果差别不大，但基于逻辑回归的多重插补下，插补1次和插补5次后的模型个别参数系数及标准误与完整数据系数差别较大；然而在缺失率较大的情况下，基于逻辑回归的多重插补的效率明显低于回归插补，插补1次的效果与插补5次的效果差别不大，插补后参数系数及标准误与完整数据系数差别大。

关键词

二分类变量，随机缺失，回归插补，多重插补

Comparative Study on Methods for Handling Missing Data in Binary Variables

Xueqin Yu

School of Science, Chongqing University of Technology, Chongqing

Received: Sep. 23rd, 2023; accepted: Oct. 19th, 2023; published: Oct. 26th, 2023

Abstract

This article introduces some commonly used imputation methods for random missing patterns, with a focus on two methods: multiple imputation and regression imputation. It further explores the imputation effectiveness of these methods by simulating different missing rates for the response variable in real-life cases. The results show that, at lower missing rates, there is not much difference in the effectiveness between multiple imputation based on logistic regression and regression imputation. However, under multiple imputation based on logistic regression, the estimated coefficients and standard errors of the model after 1 or 5 imputations differ significantly

文章引用: 余雪勤. 二分类变量缺失数据处理方法的比较研究[J]. 统计学与应用, 2023, 12(5): 1376-1384.
DOI: 10.12677/sa.2023.125142

from those of the complete data set. On the other hand, at higher missing rates, multiple imputation based on logistic regression is noticeably less efficient than regression imputation. The effectiveness does not differ much between 1 and 5 imputations, but the estimated coefficients and standard errors after imputation differ greatly from those of the complete data set.

Keywords

Binary Variables, Missing at Random, Regression Imputation, Multiple Imputation

1. 引言

1.1. 研究背景

数据作为现代科技的核心，已经渗透到社会各个领域。但是，由于数据质量的问题，缺失数据问题成为数据处理过程中极为普遍的问题之一。缺失数据往往会引起统计推断和机器学习算法的偏差，影响数据分析和建模的准确性和可靠性。因此，缺失数据的插补技术是数据处理中非常重要的方法之一。已有的插补方法大多是基于平均值、回归、KNN 等方法[1]，但这些方法在实际应用中存在一定的限制和缺陷。在现代人工智能技术的推动下，基于机器学习的缺失数据插补方法逐渐成为研究热点。在这些方法中，基于二分类数据的插补方法已经成为目前研究的重点。因此，本文将基于二分类响应变量的缺失数据作为研究对象，旨在提出一种针对二分类响应变量缺失数据的优秀可行的插补方法，改进数据处理技术，提高模型的性能，推进数据科学的发展。

1.2. 研究现状

缺失数据是近年来统计学和许多相关学科研究中的热门课题之一，对于这个问题研究，已经取得了大量的成果。对缺失数据的统计分析一直是统计学界相当关注的一个领域。统计学界相当感兴趣的领域，许多处理缺失数据问题的工具都已经被开发出来了。例如有测量误差的数据可以被看作是缺失数据的一个特例，二阶段抽样也可以被看作是一个有计划的缺失数据问题，其中关键项目只有在第二阶段的样本中被观察到。许多采用潜在变量的统计问题也可以被看作是缺失数据问题。庞新生[2]介绍了多重插补程序的三种数据插补方法：回归预测法、倾向得分法和蒙特卡罗的马氏链方法，并且对多重插补的插补效果进行推断，指出多重插补存在的问题。肖亚明[3]等人对分类变量进行删除法、基于对数线性模型的多重填补法和基于潜在类型的多重填补法处理分类数据的效果比较，结果表明基于潜在类别的多重填补法处理分类数据较好。

1.3. 研究意义

缺失数据问题一直是数据处理中的难点问题之一，影响了数据的准确性和可靠性。此外，传统的缺失数据插补方法存在一定的局限性。通过本文对二分类响应变量随机缺失处理方法的比较研究，为数据处理提供更加优秀和可行的方法，有助于提高模型的性能、促进数据分析和挖掘的准确性。通过对二分类缺失数据插补方法的比较，旨在提出一种适应于二分类插补的处理方法。

2. 相关理论和数据

2.1. 缺失机制的分类

假设在数据集(X,Y,T)中，(X,T)完全观测，响应变量Y可能存在缺失。可以将数据集Y分成两部分，记为$Y=(Y_o,Y_m)$，其中Y_o为观测数据集，Y_m为缺失数据集。用随机向量R表示Y是否缺失，$R=1$表示被观测到，$R=0$表示Y缺失。

缺失数据的研究与传统统计方法的研究区别在于，要从完全观测数据中到有效的估计，关键在于缺失机制的确定。缺失机制就是在给定观测数据和缺失数据下，指示变量R的概率分布，即$P(R|Y_o,Y_m)$。目前关于缺失数据主要集中于三种缺失机制[4]的研究上：

第一种是完全随机缺失，记为 MCAR，它指的是数据缺失与否与缺失数据和观测数据均无关，即$P(R|Y_o,Y_m)=P(R)$。在 MCAR 下，由于数据和未知参数在观测数据下仍满足零均值条件，那么通过对观测数据的统计推断，就能得到统计模型的一个有效的估计。可以采用的方法有广义估计方程等矩估计方法。

第二种是随机缺失，记为 MAR，它指的是数据的缺失仅与完全观测数据有关而与缺失数据无关，$P(R|Y_o,Y_m)=P(R|Y_o)$。在 MAR 下，只有基于似然函数的方法才能到处有效估计，而基于观测数据的矩估计和估计方程方法将失效。

第三种是非随机缺失，记为 NMAR，它指的是数据的缺失与缺失数据也有关。对应于可忽略的缺失机制，这种缺失机制也被成为是不可忽略的。当数据非随机缺失时，需要Y和R的联合分布函数，才能导出一个有效的估计。

2.2. 缺失模式分类

数据缺失的模式[5]常见有单调缺失模式和任意缺失模式。单调缺失模式是指一个变量中的受访者集合总是另一个变量的受访者集合的子集，这可能会包含进一步的子集。数据的缺失状态呈现出阶梯状的样式，即对第i个观测在第j个指标上的取值而言，若Y_{ij}缺失，则该观测在其后的任意一个指标上的取值也是缺失的。任意缺失模式中数据缺失具有随意性，即使通过行列变换也无法看出任何规律。在实际研究中也更为常见。单变量缺失可视为单调缺失模式的一个特例。

2.3. 数据来源

根据美国疾病控制预防中心的数据，现在美国 1/7 的成年人患有糖尿病。但是到 2050 年，这个比例将会快速增长至高达 1/3。以 UCL 机器学习库里的一个糖尿病数据集为样本。该数据集是一个多变量数据集，包含 768 个女性样本的生理特征，各有 9 个特征(怀孕次数，血糖，血压，皮脂厚度，胰岛素，BMI 身体质量指数，糖尿病遗传函数，年龄，结果)，因变量为结果，0 代表未患糖尿病，1 代表患有糖尿病，在 768 个数据点中，500 个被标记为 0，268 个标记为 1。

3. 缺失值处理

3.1. 数据缺失模式

本文旨在研究二分类数据单变量缺失模型下，多种缺失数据处理方法的比较研究。现模拟数据的缺失情况，在糖尿病数据集中，协变量为糖尿病数据集 768 名女性样本的 8 个生理特征，协变量不存在缺失，缺失变量为响应变量(是否患有糖尿病)，是二分类变量，当$y=0$时，表示未患有糖尿病；否则，代表患有糖尿病。现模拟响应变量在 6.25%、19.1%、36.9%三种不同缺失比例下，不同的缺失比例通过调

整缺失机制的参数来实现。采用多重插补、回归插补对缺失的数据集进行填补，然后通过插补后的完整数据集建立逻辑回归模型，比较各参数系数估计值及其标准误。以及模型判别准确率来比较插补方法的效果。本文通过模拟研究逻辑回归参数估计量的性能及模型判别准确率来分析插补方法的效果。假设 $X_1, X_2, \cdots, X_8$ 无缺失值，总是能被观测到，然而数据中“结果” Y 是随机缺失的。由于 X_1, X_2, X_6 这三个完全观测变量与 Y 相关程度较强，故选这三个变量对 Y 设置随机缺失。设 R_i 为 0~1 变量，用来表示第 i 个个体的响应变量信息是否是可以被观测的，当响应变量是完全观测的 R_i 指示为 1，否则 R_i 指示为 0。通过下面的逻辑回归模型来生成 R_i：

$$\operatorname{logit}\left(p\left(R_i = 0 \mid X_i\right)\right) = -\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i6}, \tag{1}$$

其中，$\operatorname{logit}(p) = \log(p/1-p)$，通过设置不同的 $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ 参数来实现不同的缺失率[6]。

3.2. 研究思路

现模拟该数据集在 6.25%、19.1%、36.9%三种不同缺失比例下，然后采用完整个案分析法、多重插补法、回归插补法对缺失后的数据集进行处理，其中多重插补方法同时考察插补 1 次、5 次对处理效果的影响。然后不同缺失比例下不同方法的模型的参数估计，以及模型判别的正确率，并对这些估计值进行汇总分析，比较这些方法的参数估计。

3.3. 处理方法

3.3.1. 完整个案分析法(CC)

完整个案分析法又称删除的方法[7]，完整个案分析法是指所有个体的多次观测记录均被检测到且用于分析变量值，含有数据缺失的均略去，仅利用可以完全观测的样本进行统计推断。完整个案分析法虽然简单，数据结构平衡，可获得完整的数据矩阵；但丢掉了大量不完全严格不能中的信息，因而会损失估计的效率，出现较大的偏倚，结果过于保守或夸大其效果，数据缺失资料分析不推荐使用该法。

3.3.2. 多重插补法-logistic 回归[8]

二分类变量常用 logistic 回归模型进行插补。常利用拟合 logistic 回归模型的基础上，根据拟合回归参数的后验分布值，可得到一个新的 logistic 回归模型，完成对缺失数据的插补。其具体步骤为：

第一步，假设有 n 个观测，其中有 p 个自变量和一个二元分类变量 Y，响应变量 Y 取值为 0 或 1，协变量为 $X_1, X_2, \cdots, X_p$，利用在 $X_1, X_2, \cdots, X_p$ 变量上的完整观测数据，建立 Y 与协变量之间的线性模型，即公式为：

$$\operatorname{Logistic}\left(p_1\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \tag{2}$$

其中，回归参数的估计值记为 $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)$，用来衡量各个变量对结果的影响，与之相关的协方差阵记为V

$$p_1 = p\left(y = 1 \mid x_1, x_2, \cdots, x_p\right), \tag{3}$$

$$\operatorname{logit}\left(p_1\right) = \frac{p_1}{1 - p_1}, \tag{4}$$

第二步，根据参数的后验分布模拟得到新的参数，具体来说，从参数估计值 $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)$ 的后验分布预测分布中，获得新的参数估计值 β_*。其公式为：

$$\beta_* = \hat{\beta} + V_i' Z, \tag{5}$$

其中，V_i' 是对V进行Cholesky分解后得到的上三角矩阵[8]，Z是一个向量，由$p+1$个独立的随机正态变量组成。如果Y_i缺失，可通过下式计算得到$Y=1$的概率：

$$p_1=\frac{\exp\left(\beta_{*0}+\beta_{*1}x_1+\cdots+\beta_{*p}x_p\right)}{1+\exp\left(\beta_{*0}+\beta_{*1}x_1+\cdots+\beta_{*p}x_p\right)}, \tag{6}$$

然后产生一个随机均匀变量u，其取值介于0和1之间。如果u的值小于p_1，则缺失数据填补为0，否则，填补值为1。

第三步，随机生成r次u，得到r组可能插补值，用y_{ir}表示第r组的第i个观测值的响应变量插补值，其次使用每组的y值和$x_1,x_2,\cdots,x_p$进行逻辑回归建模，可以得到r组逻辑回归模型。假设第r组逻辑回归模型的参数为$\beta_{0r},\beta_{1r},\cdots,\beta_{pr}$。

第四步，使用均值回归的方法结合r组模型参数。具体来说，对于第j个参数，我们可以计算出r组参数的平均值$\overline{\beta}_j$和标准误se_j。然后可以使用以下公式来计算回归系数的t值和P值：

$$t_j=\frac{\overline{\beta}_j}{se_j}, p_j=2P\left(|T|>|t_j|\right), T\sim t_{n-r}, \tag{7}$$

其中n是观测值的总数，并且假设每组插补得到的数据是完全独立的。

3.3.3. 回归插补法

回归插补[9]的原理就是利用已知协变量的观测信息与样本中的有效回答记录建立回归模型，模型表明了目标变量与辅助变量之间的关系，然后根据辅助信息，利用建立的回归模型对缺失值进行估计，于是第i个缺失值的估计值可以表示为：

$$Y_i^*=f\left(X\right)+e_i, \tag{8}$$

首先利用已知的观测信息建立逻辑回归模型：

$$f\left(x_i\right)=\frac{\exp\left(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}\right)}{1+\exp\left(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}\right)}, \tag{9}$$

应用回归填补法时，需要注意e_i的处理，需要构造随机残差e的数据集，根据辅助变量X将样本单位分层在各层中将观测数值与其均值的离差视为残差。再用回归法得到$\hat{Y}_i$后，在该层的残差集中随机抽取$\hat{Y}_i$的残差项，并将其和作为缺失数据插补为1的概率值，将有缺失的y_i根据概率进行填补，如果概率值大于0.5，插补值就为1，概率值小于0.5，插补值就为0。因此只有当辅助变量和目标变量之间存在高度的相关关系时，采用回归方法就是比较的。

4. 处理结果比较

本文所选取的案例数据包含768个样本点，9个特征，表1是对数据变量集的描述。

Table 1. Variable description
表1. 变量描述

变量名称	描述	取值
Outcom(Y)	目标结果，是否患有糖尿病(1患病，0未患病)	二元变量
Pregnancies(X_1)	怀孕次数	连续型
Glucose(X_2)	口服葡萄糖耐量试验中两小时后的血浆葡萄糖浓度	连续型

Continued

BloodPressure(X_3)	血压(以毫米汞柱为单位)	连续型
SkinThickness(X_4)	皮脂厚度	连续型
Insulin(X_5)	两小时血清胰岛素浓度(以 μU/ml 为单位)	连续型
BMI(X_6)	身体质量指数	连续型
DiabetesPedigreeFunction(X_7)	糖尿病遗传函数	比例值
Age(X_8)	年龄	离散型

$$\operatorname{logit}\left(p\left(y=1\right)\right)=-7.955+0.153x_1+0.034x_2-0.012x_3+0.084x_6+0.911x_7. \tag{10}$$

由于在调查中，数据往往会出现缺失的情况，从而得到不完整的数据集。对于数据缺失的情况，本文对得到的 768 个糖尿病完整数据集的响应变量“结果”进行随机缺失，产生不同缺失率下的缺失数据，然后再对缺失数据集进行插补，由插补后的数据集进行参数估计和模型建立，并且进行模型判别正确率的检验，运用 R 软件进行分析病得到分析结果。

4.1. 缺失比例为 6.25%

本文采用 UCL 机器学习库里的一个糖尿病数据集为样本。选取怀孕次数、血糖、血压、BMI 身体质量指数、糖尿病遗传函数作为完全观测变量，但响应变量“结果”是随机缺失的个体，该数据集是包含 $n=768$ 个女性样本的生理特征。R_i 为 0~1 变量，用来表示第 i 个个体的“结果”信息是否可以被观测到，通过对“结果”缺失的概率建立 logistic 回归模型，来验证上述所提方法在实际生活中的可实施性。当“结果”信息被完全观测时，R_i 指示为 1；当“结果”信息部分丢失或者完全丢失时，R_i 指示为 0。假设协变量是完全变量，响应变量缺失机制为 MAR，y 的缺失与 x_1,x_2,x_6 有关[10]，参数设置见公式(11)如下：

$$\operatorname{logit}\left(p\left(R_i=0\mid X_i\right)\right)=-5-2X_{i1}+X_{i2}-2X_{i6}, i=\left(1,2,\cdots,768\right), \tag{11}$$

从而得到了响应变量随机缺失 6.25%的数据集，然后分别采用完整个案分析法、多重插补逻辑回归法和回归插补的方法分别得到不同的填补数据集下模型的参数估计及标准误差，并且得到模型判别正确率，见表 2：

Table 2. Parameter estimates and standard errors at 6.25% missing rate
表 2. 缺失率 6.25%下参数估计及标准误差

处理方法	名称	β_0	β_1	β_2	β_3	β_6	β_7	模型判别准确率
CC	系数 标准误	−8.047 0.702	0.145 0.030	0.037 0.004	−0.012 0.005	0.082 0.015	0.885 0.303	77.3%
MI/logit 填补 1 次	系数 标准误	−6.079 0.587	0.149 0.026	0.020 0.003	−0.009 0.005	0.089 0.013	0.681 0.267	78%
MI/logit 填补 5 次	系数 标准误	−5.931 0.580	0.157 0.026	0.019 0.003	−0.009 0.005	0.083 0.013	0.711 0.266	78.4%
回归插补	系数 标准误	−8.40 0.699	0.143 0.028	0.039 0.004	−0.012 0.005	0.081 0.014	0.883 0.301	77.2%

当数据缺失比例为 6.25%时，三种不同方法处理后，最终模型的系数、标准误和模型的判别正确率几乎都非常接近。回归插补法和完整个案分析法的参数及其标准误差的估计与完整数据分析结果差别较

小。在缺失率较低的情况下，多重插补的效果略较于其他两种方法好，但个别参数及标准误差的估计值与完整数据分析结果差别较大[11]。

4.2. 缺失比例为 19.1%

通过对“结果”缺失的概率建立 logistic 回归模型，来验证上述所提方法在实际生活中的可实施性。R_i 为 0~1 变量，用来表示第 i 个个体的“结果”信息是否可以被观测的，当“结果”信息被完全观测时，R_i 指示为 1；当“结果”信息部分丢失或者完全丢失时，R_i 指示为 0。通过改变$(\alpha_0,\alpha_1,\alpha_6)$的值，使得缺失率达到了 19.1%。假设协变量是完全变量，响应变量缺失机制为 MAR，y 的缺失与 x_1,x_2,x_6 有关，参数设置见公式(12)如下：

$$\text{logit}\left(p\left(R_i=0\mid X_i\right)\right)=-3-2.5X_{i1}+X_{i2}-2.5X_{i6},i=(1,2,\cdots,768), \tag{12}$$

得到了响应变量随机缺失 19.1%的数据集，分别采用完整个案分析法、多重插补逻辑回归法和回归插值的方法分别得到不同的填补数据集下模型的参数估计及标准误差，并且得到模型判别正确率，见表 3：

Table 3. Parameter estimates and standard errors at 19.1% missing rate
表 3. 缺失率 19.1%下参数估计及标准误

处理方法	名称	β_0	β_1	β_2	β_3	β_6	β_7	模型判别准确率
CC	系数 标准误	−7.881 0.729	0.148 0.034	0.037 0.004	−0.015 0.006	0.083 0.017	0.709 0.317	77.2%
MI/logit 填补 1 次	系数 标准误	−4.171 0.515	0.137 0.025	0.006 0.003	−0.004 0.004	0.086 0.012	0.518 0.248	71.4%
MI/logit 填补 5 次	系数 标准误	−4.378 0.523	0.111 0.024	0.006 0.003	−0.003 0.004	0.093 0.013	0.479 0.248	70.3%
回归插补	系数 标准误	−8.913 0.730	0.150 0.029	0.046 0.004	−0.016 0.005	0.075 0.015	0.688 0.311	77.1%

当数据缺失比例为 19.1%时，三种不同方法处理后，最终模型的系数、标准误差别较大，模型的判别正确率都大于 70%，但完整个案分析法的模型判别的准确率较高于多重插补方法。多重插补在缺失率较高的情况下不具有优势，然而完整个案分析法和回归插补法都比多重填补方法较好。但多重插补仍保持其大于 70%的模型判别准确率。除了完整个案分析法和回归插补的参数及其标准误差的估计与完整数据分析结果差别较小，多重填补方法的个别参数及其标准误差的估计与完整数据分析结果差别较大。由于数据的随机性较大，插补效率并不是很高，特别是插补次数越多的情况下，系数并没有体现其显著性。

4.3. 缺失比例为 36.9%

通过对“结果”缺失的概率建立 logistic 回归模型，来验证上述所提方法在实际生活中的可实施性。设 R_i 为 0~1 变量，用来表示第 i 个个体的响应变量信息是否是可以被观测的，当“结果”信息被完全观测时，R_i 指示为 1；当“结果”信息部分丢失或者完全丢失时，R_i 指示为 0。通过改变$(\alpha_0,\alpha_1,\alpha_6)$的值，使得缺失率达到了 36.9%。假设协变量是完全变量，响应变量缺失机制为 MAR，y 的缺失与 x_1,x_2,x_6 有关，参数设置见公式(13)如下：

$$\text{logit}\left(p\left(R_i=0\mid X_i\right)\right)=-4-2.5X_{i1}+X_{i2}-3X_{i6},i=(1,2,\cdots,768), \tag{13}$$

得到了响应变量随机缺失 36.9%的数据集，然后分别采用完整个案分析法、多重插补逻辑回归法和回归插补的方法分别得到不同的填补数据集下模型的参数估计及标准误差，并且得到模型判别正确率，见表 4：

Table 4. Parameter estimates and standard errors at 39.1% Missing rate
表 4. 缺失率 39.1%参数估计及标准误

处理方法	名称	β_0	β_1	β_2	β_3	β_6	β_7	模型判别准确率
CC	系数	−8.127	0.158	0.036	−0.013	0.092	0.520	77.3%
	标准误	0.840	0.040	0.005	0.006	0.023	0.353	
MI/logit 填补 1 次	系数	−1.842	0.074	−0.004	−0.005	0.078	0.066	58.9%
	标准误	0.454	0.024	0.002	0.004	0.012	0.234	
MI/logit 填补 5 次	系数	−1.984	0.570	−0.004	−0.004	0.083	−0.033	58.9%
	标准误	0.460	0.023	0.002	0.004	0.012	0.234	
回归插补	系数	−10.829	0.194	0.054	−0.018	0.100	0.587	77.0%
	标准误	0.852	0.032	0.004	0.006	0.016	0.333	

当数据缺失比例为 39.1%时，三种不同方法处理后，最终模型的系数、标准误和模型的判别正确率差别较大。除了完整个案分析法的参数及其标准误差的估计与完整数据分析结果差别较小，多重填补方法与回归插补法的个别参数及其标准误差的估计与完整数据分析结果差别较大，尤其是多重插补得到的参数估计值与完整数据分析得到的参数相差甚大。在缺失率较大的情况下，多重插补就失去其优势，不仅模型个别参数不显著，而且最终模型的判别准确率远低于完整个案分析法和回归插补法。基于逻辑回归模型的多重插补方法具有随机性，数据缺失越多，插补次数越多的情况下，产生的误差往往比较大，最后得到的模型判别准确率也比较大且模型系数不体现其显著性。完整个案分析法的模型判别的准确率与回归插补法得到的模型判别准确率相差不大，都高于多重填补的方法，缺失率越高，多重填补方法就失去了填补的优势，特别是在二分类变量的插补中。随着缺失率的增大，多重插补法处理后的参数准确度下降，处理效果不佳。

5. 结论与讨论

在完整数据集基础上模拟不同缺失率的随机缺失数据集，采用三种方法对缺失数据集进行参数估计，并且比较模型的判别正确率。当缺失率较低时，用回归插补法与多重插补法以及完整个案分析法进行处理的结果差别不太，与之相比，处理后的结果与完整数据集的分析结果较为接近。多重插补就体现其优势，模型判别准确率较其他两种方法更优。当数据缺失率一般时，回归插补处理后的结果与完整个案分析法较多重填补法更接近完整数据的分析结果，且填补 5 次的结果与填补 1 次比较效果差异不显著，插补次数越多，由于数据插补的随机性，得到的数据插补准确率越低，模型判别准确率下降；且填补方法的效果并不比完整个案分析法好。然而缺失率较高时，填补 5 次的结果与填补 1 次的效果均没有体现其优越性，反而比完整个案分析法更劣，且其中个别的参数及其标准误差的估计与完整数据分析结果差别较大，有的系数甚至不呈现其显著性。

在二分类响应变量随机缺失、其他协变量不存在缺失的情况，应用多重插补、回归插补对缺失的二分类数据进行填补，结果表明在不同缺失率下选择的插补方法会有所不同，因此在选择分析方法时我们需要注意原始数据的特点具体问题具体分析也可通过比较选择出最适合的统计方法[12]。缺失数据插补效果取决于辅助变量的选取，辅助变量与目标变量的关系越显著，在 MAR 机制下，插补效果越好，当辅助变量较多时，需要对辅助变量进行筛选，尽可能选取与目标变量关系显著的变量。目前各种关于二分类数据缺失的插补方法也在不断探讨与完善，我们需要进一步深入学习。

参考文献

[1] 岳勇, 田考聪. 数据缺失及其填补方法综述[J]. 预防医学情报杂志, 2005(6): 683-685.

[2] 庞新生. 缺失数据插补处理方法的比较研究[J]. 统计与决策, 2012(24): 18-22. https://doi.org/10.13546/j.cnki.tjyjc.2012.24.003

[3] 肖亚明, 陈永杰, 王玉鹏, 刘美娜. 分类变量缺失数据处理方法有效性的比较研究[J]. 中国卫生统计, 2016, 33(2): 186-189.

[4] 袁中萸. 多元线性回归模型中缺失数据填补方法的效果比较[D]: [硕士学位论文]. 长沙: 中南大学, 2008.

[5] 周敏. 多分类等级量表数据缺失填补方法的比较研究[D]: [硕士学位论文]. 沈阳: 中国医科大学, 2022. https://doi.org/10.27652/d.cnki.gzyku.2022.000407

[6] 于力超. 协变量数据缺失情形下的参数估计方法[J]. 统计与决策, 2018, 34(17): 9-13. https://doi.org/10.13546/j.cnki.tjyjc.2018.17.002

[7] 王曼, 施念, 花琳琳, 等. 成组删除法和多重填补法对随机缺失的二分类变量资料处理效果的比较[J]. 郑州大学学报(医学版), 2012, 47(5): 642-645.

[8] 解东方. 心血管病流行病调查中缺失数据填补方法的比较及模拟研究[D]: [博士学位论文]. 北京: 北京协和医学院, 2014.

[9] 戴明锋, 金勇进, 查奇芬, 等. 二分类 Logistic 回归插补法及其应用[J]. 数学的实践与认识, 2013, 43(21): 162-167.

[10] 肖亚明, 陈永杰, 王玉鹏, 等. 分类变量缺失数据处理方法有效性的比较研究[J]. 中国卫生统计, 2016, 33(2): 4.

[11] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述[J]. 计算机工程与应用, 2021, 57(14): 27-38.

[12] 鲍晓蕾, 高辉, 胡良平. 多种填补方法在纵向缺失数据中的比较研究[J]. 中国卫生统计, 2016, 33(1): 45-48.