Statistics and Application 统计学与应用, 2025, 14(5), 97-109
Published Online May 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.145129

Hans汉斯

随机缺失协变量下基于多重插补的Logistic回归分析

杨　越，余雪勤，杨　逊

重庆理工大学理学院，重庆

收稿日期：2025年4月21日；录用日期：2025年5月13日；发布日期：2025年5月23日

摘　要

在抽样调查研究中，针对含缺失数据集的总体参数估计问题，当前学术界的研究焦点呈现明显的不均衡性。现有方法体系主要集中于因变量缺失场景下的参数估计(如响应变量缺失下的回归分析)，而对协变量缺失情形下的统计推断方法探索研究较少。本文在协变量数据缺失机制为随机缺失的情形下，考虑用多重插补法来进行插补得到完整数据并建立Logistic回归模型，从多重插补估计量与设定模型参数的差异以及多重插补得到的方差估计来评估多重插补法的精度，并通过模拟实验对比了不同缺失率下多重插补估计量表现。最后对协变量存在随机缺失的德国信贷数据进行了多重插补后逻辑回归建模分析，比较了两种多重插补方法的效果。

关键词

随机缺失，多重插补，Logistic模型

Logistic Regression Analysis Based on Multiple Imputation under Random Missing Covariates

Yue Yang, Xueqin Yu, Xun Yang

School of Science, Chongqing University of Technology, Chongqing

Received: Apr. 21st, 2025; accepted: May 13th, 2025; published: May 23rd, 2025

Abstract

In sampling survey research, there is a significant imbalance in the current academic focus on the estimation of population parameters with missing datasets. Existing methods mainly focus on situations

文章引用: 杨越, 余雪勤, 杨逊. 随机缺失协变量下基于多重插补的 Logistic 回归分析[J]. 统计学与应用, 2025, 14(5): 97-109. DOI: 10.12677/sa.2025.145129

where the dependent variable data is missing (such as regression analysis in response variable missing scenarios), while relatively little research has been conducted on scenarios involving missing covariates. Under the condition of missing covariate missing data mechanism, this paper considered the multiple interpolation method to get the complete data and establish the Logistic regression model, evaluated the accuracy of the variance estimate obtained from the difference between the parameters and the multiple interpolation, and compared the performance of the multiple interpolation estimator under different missing rates through simulation experiments. Finally, multiple imputation logistic regression modeling analysis was conducted on German credit data with random missing covariates, and the effects of the two multiple imputation methods were compared.

Keywords

Random Missing, Multiple Imputation, Logistic Model

1. 引言

在实际数据分析场景中，受人为操作失误、设备采集误差及数据获取成本制约等多重因素影响，缺失数据集广泛存在于各类研究实践中。统计推断的可靠性本质上取决于数据的信息完整度，当关键变量存在系统性缺失时，传统基于完全观测数据的分析方法(如线性回归模型)将产生显著偏差，尤其在协变量维度较高或缺失比例较大的复杂场景下。这种偏差在缺失机制与变量取值存在内在关联时尤为突出——例如医疗健康调查中，重症患者可能因心理抵触隐瞒病情数据，若简单剔除此类样本，将导致疾病严重程度评估产生向下偏误。因此，构建包含丰富辅助变量的多维数据体系显得尤为重要。针对缺失值处理，现有研究主要采用数据插补技术(如基于模型的回归插补、基于相似度的近邻插补)和似然函数修正方法(联合建模变量分布与缺失机制) [1]。然而值得注意的是，这些方法论多建立在协变量完全观测的理想假设之上，而现实调研中协变量缺失现象同样普遍，例如涉及家庭收入、教育程度等敏感信息时，受访者可能选择性隐瞒部分特征数据，或受认知水平限制无法准确回答特定问题。本研究聚焦于因变量完整但协变量存在随机缺失的情形下，通过构建协变量缺失下研究更稳健、有效的统计推断方法。

2. 相关理论

2.1. 符号说明

假设有 N 个观测数据，有二元响应变量 Y 和 p 个协变量 X_i，$X_i=\left(x_{i1},\cdots,x_{ip}\right)$，其中响应变量 Y 没有缺失值，得到完整观测 y_i，$i=1,\cdots,N$。协变量含有缺失值的情形下，记第 i 条数据协变量的所有观测值为 $X_{obs,i}$，缺失值为 $X_{mis,i}$，$X_i=\left(X_{obs,i},X_{mis,i}\right)$，记 $N\times p$ 维协变量缺失指示变量矩阵 $r=\left(r_1,\cdots,r_N\right)^{\mathrm{T}}$，若第 i 条数据第 k 个协变量数据未缺失 $r_{ik}=0$，否则 $r_{ik}=1$。

2.2. 随机缺失

缺失机制描述的是包含缺失的变量和其它变量的联系，在不同的缺失机制下，缺失的概率与各变量的相关性或非相关性。针对不同类型的缺失数据，研究其机制具有重要的理论意义和应用价值。Rubin, D. B. (1976) [2]对丢失数据的机制进行了研究，将其划分为随机缺失(MAR)、完全随机缺失(MCAR)和非随机缺失(NMAR)。本文研究的是在协变量随机缺失条件下，利用多重插值方法进行 Logistic 回归分析。记

P 是模型发生缺失的概率，c 是模型的常数[3]。

随机缺失是指缺失数据项与带缺失数据部分的变量没有关系，即是独立的，而只与完全观测数据存在关系。其公式可以表示为：

$$p\left(r_{ik}=1\left|y_i,X_{obs,i},X_{mis,i},c\right.\right)=p\left(r_{ik}=1\left|y_i,X_{obs,i},c\right.\right) \tag{1}$$

随机缺失在现实中比较常见，只要满足缺失项与带缺失项的变量无关即可。比如某些调研的一些私密问题，对于被调查者是否回答该问题，与性别、年龄等相关，造成该项回答的缺失就是随机缺失。检测化验单中其中两项对诊断疾病结果的功能相似，这两项中有一项可检查可不检查等，造成该项回答的缺失也是随机缺失。

2.3. 多重插补

多重插补是从单一插补方法衍变过来的一种插值方法，反映了某些数据集的不确定性并较好地保持了变量之间的良好关系。Rubin 在 1977 年首次提出多重插补法，后来被人们逐步地加以改进与合成，目前已成为一种较为完整的方法[4]。其主旨如下：

(i) 对于每一个遗漏的数值，我们将用 M 个可能的估算来进行插值，从而得到 M 个完全的数据集。

(ii) 对于每一组完备的数据，我们都可以采用同样的方法，对其进行分析，从而获得 M 组的解析结果。

(iii) 对 M 个插补数据集进行综合，得出对目标变量的统计推断。

该模型的本质是一种仿真方法，它在特定的情况下对模型参数进行了建模，从而可以用来对模型中的数据进行估计。所谓多重插补，就是由插值，解析，合并三个步骤组成的。

多重插补是基于贝叶斯理论，将贝叶斯方法中的参数看作是随机性的。贝叶斯理论的主要结论是参数的后验分布可由该参数的似然函数与先验分布表示：

$$P\left(\theta\left|data\right.\right)=\frac{P\left(y_i\left|\theta\right.\right)P\left(X_i\left|\theta\right.\right)}{\int P\left(y_i\left|\theta\right.\right)P\left(X_i\left|\theta\right.\right)\mathrm{d}\theta}\propto P\left(y_i\left|\theta\right.\right)P\left(X_i\left|\theta\right.\right) \tag{2}$$

插补问题本质上是一个预测问题。多重插补法认为缺失值是随机的，它的值可以通过已观测到的值进行预测和插值。因此，Kim, J. K.和 Shao J. [5]将贝叶斯方法应用于预测来处理填充问题。在 $X_i=\left(X_{obs,i},X_{mis,i}\right)$ 的缺失数据设定下，给定当前观测值 $X_{obs,i}$， $X_{mis,i}$ 的后验预测分布可表示为：

$$P\left(X_{mis,i}\left|X_{obs,i},y_i,\eta\right.\right)\propto P\left(y_i\left|X_i,\beta\right.\right)P\left(X_i\left|\alpha\right.\right) \tag{3}$$

其中 $\eta=\left(\alpha,\beta\right)$。特别是，当协变量 X 均是离散类型时，

$$P\left(X_{mis,i}\left|X_{obs,i},y_i,\eta\right.\right)=\frac{P\left(y_i\left|X_{obs,i},X_{mis,i},\beta\right.\right)P\left(X_{mis,i}\left|X_{obs,i},\alpha\right.\right)}{\sum_{X_{mis,i}}P\left(y_i\left|X_{obs,i},X_{mis,i},\beta\right.\right)P\left(X_{mis,i}\left|X_{obs,i},\alpha\right.\right)} \tag{4}$$

分母中包含了对于第 i 个数据项中缺少的所有协变数的全部可能的总和。如果所有的协变量都是连续的，那么在分母中的求和符号就变成了积分数值。通常，协变量可以分为离散和连续两种类型，记第 i 个数据中的离散性和连续性的所有可能的取值范围为 Ω：包含丢失值的离散类型的协变量的集合可以用 $X_{mis,i}^{(1)}$ 表示；包含丢失值的连续类型的协变量的集合可以用 $X_{mis,i}^{(2)}$ 来表示，那么：

$$P\left(X_{mis,i}\left|X_{obs,i},y_i,\eta\right.\right)=\frac{P\left(y_i\left|X_{obs,i},X_{mis,i},\beta\right.\right)P\left(X_{mis,i}\left|X_{obs,i},\alpha\right.\right)}{\int_{\Omega}\sum_{X_{mis,i}^{(1)}}P\left(y_i\left|X_{obs,i},X_{mis,i},\beta\right.\right)P\left(X_{mis,i}\left|X_{obs,i},\alpha\right.\right)\mathrm{d}X_{mis,i}^{(2)}} \tag{5}$$

为获得插补值 $X_{mis,i}$，需要获得 $X_{mis,i}$ 后验分布 $P\left(X_{mis,i}\middle|X_{obs,i},y_i,\eta\right)$ 的参数 η，然后从该后验分布中抽取 $X_{mis,i}$，获得参数 η 的方法有两种：在 SAS ProcMI 软件中，假定数据服从多变量正态分布，并假定协变量的丢失机理是 MAR，从而得到插值结果。但是，这个方法要求我们假设所有的人都服从多个正态分布。二是利用 Gibbs 采样法进行插值，而无需对总体分布作任何假定[6]。本项目拟采用数据增广(Data augmentation) [7]实现对 (α,β) 的 Gibbs，以充分利用观测数据中的信息。在以下两个公式中，体现了数据增强算法的基本概念：

$$f\left(\eta\middle|X_{obs,i},y_i\right)=\int_{S\left(X_{mis,i}\middle|X_{obs,i},Y_i\right)}f\left(\eta\middle|X_{obs,i},y_i,X_{mis,i}\right)f\left(X_{mis,i}\middle|X_{obs,i},y_i\right)\mathrm{d}X_{mis,i} \tag{6}$$

$$f\left(X_{mis,i}\middle|X_{obs,i},y_i\right)=\int_{S\left(\eta\middle|X_{obs,i},Y_i\right)}f\left(X_{mis,i}\middle|X_{obs,i},y_i,\eta\right)f\left(\eta\middle|X_{obs,i},y_i\right)\mathrm{d}\eta \tag{7}$$

这里，$S\left(X_{mis,i}\middle|X_{obs,i},Y_i\right)$ 和 $S\left(\eta\middle|X_{obs,i},Y_i\right)$ 表示在给定的观察情况下，缺失的协变量数值以及参数的分布。设第 k 步迭代所得参数的后验分布 $f_k\left(\eta_k\middle|X_{obs,i},y_i\right)$，则第 $k+1$ 步参数 η 的后验分布为：

$$f_{k+1}\left(\eta_{k+1}\middle|X_{obs,i},y_i\right)=\int g\left(\eta_k,\eta_{k+1}\right)f_k\left(\eta_k\middle|X_{obs,i},y_i\right)\mathrm{d}\eta_k \tag{8}$$

其中：

$$g\left(\eta_k,\eta_{k+1}\right)=\int f\left(\eta_{k+1}\middle|X_{obs,i},X_{mis,i},y_i\right)f\left(X_{mis,i}\middle|X_{obs,i},y_i,\eta_k\right)\mathrm{d}X_{mis,i} \tag{9}$$

Tanner and Wong (1987)已经证明了在某些正则项下，f_{k+1} 服从分布渐近收敛到参数的后验分布 $f\left(\eta\middle|X_{obs,i},y_i\right)$。基于上述两个公式，本文提出了一种迭代法：

首先，对参数进行初值设置 η_0，并从第 k 个步骤的后验分布 $f_k\left(\eta_k\middle|X_{obs,i},y_i\right)$ 中选择抽取 M 个参数值 $\left\{\eta^{(j)}\right\}_{j=1}^{M}$，对每一个 $\eta^{(j)}$，代入 $f\left(X_{mis,i}\middle|X_{obs,i},y_i,\eta^{(j)}\right)$(计算公式见式(6)和式(7))，对第 k 个步进行插值，然后在步骤 k 处获得 M 个插补值 $\left\{X_{mis,i}^{(j)}\right\}_{j=1}^{M}$。

接着，我们将第 M 步骤的插值结果替换成 M 个完备的数据集合，利用该集合中的参数的后验分布，计算出第 $k+1$ 个步骤中的参数的后验分布：

$$f_{k+1}\left(\eta_{k+1}\middle|X_{obs,i},y_i\right)=\frac{1}{M}\sum_{j=1}^{M}f\left(\eta\middle|X_{obs,i},X_{mis,i}^{(j)},y_i\right) \tag{10}$$

通过重复多次，获得参数的后验分布，将其替换为缺失的协变量的后验分布，从而获得缺失的协变量的分布，然后从这个后验分布 $X_{mis,i}^{(j)}$ 中提取，如此反复进行可以得到 $X_{mis,i}^{*(1)},\cdots,X_{mis,i}^{*(M)}$。

利用此插值方法求出多项式插值 η 的估计：

$$\hat{\eta}=\frac{1}{M}\sum_{j=1}^{M}\hat{\eta}^{(j)}$$

由第 i 个插补后的完整数据集得到参数的方差估计 $V^{(j)}$，$W=\frac{1}{M}\sum_{m=1}^{M}\hat{V}^{(j)}$ 为 M 个方差估计的均值，称为组内方差均值，定义组间方差均值为 $B=\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\eta}^{(j)}-\hat{\eta}\right)\left(\hat{\eta}^{(j)}-\hat{\eta}\right)^{\mathrm{T}}$，多重插补得到的方差估计为：

$$\hat{V}=W+\left(1+\frac{1}{M}\right)B \tag{11}$$

2.4. Logistic 回归模型

Logistic 回归模型，虽然它的名字是“回归”，但实际是一种分类学习方法。设 Y 是布尔型因变量，X 是 p 维相互独立的协变量，$p_i = P(Y=1|X)$ 表示在 X 的条件下 Y 取值为 1 的概率，有：

$$P(Y=1|X)=\frac{1}{1+e^{-z}}$$

其中 $z=\beta_0+\sum_{i=1}^{p}\beta_i * X_i$。

$1-p_i = P(Y=0|X)$ 表示在 X 的条件下 Y 取值为 0 的概率，则

$$1-p_i = P(Y=0|X)=1-\frac{1}{1+e^{-z}}$$

两者之比取对数可得到“对数几率”

$$\ln\frac{p_i}{1-p_i}=\beta_0+\sum_{i=1}^{p}\beta_i * X_i$$

回归模型中最为熟悉的估计方法是最小二乘估计，但最小二乘估计适用于线性模型．由于极大似然估计与最小二乘估计同等的估计性能，且极大似然估计可用于非线性模型，因而 Logistic 模型一般用极大似然估计来估计模型的参数。给定数据集 $\{(x_i,y_i)\}_{i=1}^{n}$，模型最大化“对数似然”

$$l(\beta_0,\beta_i)=\ln\left(\prod_{i=1}^{n}p_i^{Y_i}*(1-p_i)^{1-Y_i}\right)=\sum_{i=1}^{n}Y_i\ln\left(\frac{p_i}{1-p_i}\right)+\ln(1-p_i)$$

进一步，

$$l(\beta_0,\beta_i)=\sum_{i=1}^{n}\left[Y_i*\left(\beta_0+\sum_{i=1}^{p}\beta_i*X_i\right)-\ln\left(1+e^{\beta_0+\sum_{i=1}^{p}\beta_i*X_i}\right)\right]$$

对参数向量 β 求偏导并令其为 0，求得的参数即为模型参数的估计值。

3. 模拟研究

本节应用模拟研究对提出的插补方法进行验证。假设样本量 $N=500$，第 i 条数据因变量 y_i 为 0~1 变量，协变量 $X_i=(X_{i1},X_{i2},X_{i3})$，$y_i$ 通过如下 logistic 模型独立生成：

$$\text{logit}(p(y_i=1))=-1-X_{i1}+X_{i2}-X_{i3},\ i=1,\cdots,500 \tag{12}$$

故模型参数真实值为 $(\beta_0,\beta_1,\beta_2,\beta_3)=(-1,-1,1,-1)$。

采用如下方式生成协变量值：

$$X_{i1}\sim N(0,0.5^2),\quad X_{i2}\sim N(3,1.5^2),\quad X_{i3}\sim N(2,1)$$

假设 X_{i1} 和 X_{i2} 没有缺失，X_{i3} 的随机缺失机制如下(缺失率分别为 0.044、0.124、0.42)：

$$\begin{aligned}
&\text{logit}(p(r_{i3}=1|X_{i1},X_{i2},y_i))=1-2X_{i1}-4X_{i2}+3y_i,\ i=1,\cdots,500\\
&\text{logit}(p(r_{i3}=1|X_{i1},X_{i2},y_i))=1-2X_{i1}-2X_{i2}+3y_i,\ i=1,\cdots,500\\
&\text{logit}(p(r_{i3}=1|X_{i1},X_{i2},y_i))=1-2X_{i1}-X_{i2}+3y_i,\ i=1,\cdots,500
\end{aligned} \tag{13}$$

可见协变量 X_{i3} 的缺失机制为 MAR。采用上面的数据生成方法，得到样本量为 500 的模拟数据。再

采用上述协变量数据缺失机制，得到只有 X_{i3} 含缺失的模拟数据，对该数据进行 5 次多重插补得到 X_{i3} 不同缺失率下模型(14)的参数估计，从估计值与真实值的偏差、基于模拟数据参数估计的标准差和均方误差看估计结果与本文设定的模型(12)之间的符合程度。其中标准差是由式(11)多重插补得到的方差估计开根号而来，均方误差是五次插补所得参数估计值与真实值的差值平方的均值。

$$\operatorname{logit}\left(p\left(y_i=1\right)\right)=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3},\ i=1,\cdots,500 \tag{14}$$

所得结果如下(保留小数点后 4 位)：

Table 1. Results of multiple imputation with X_{i3}-missing rate of 0.044
表 1. X_{i3} 缺失率为 0.044 下多重插补结果

参数	真实值	估计值	偏差	标准差	均方误差
β_0	−1	−0.7712	0.2288	0.3874	0.0542
β_1	−1	−0.8330	0.1670	0.2331	0.0279
β_2	1	1.0705	0.0705	0.1122	0.0052
β_3	−1	−1.2140	0.2140	0.1498	0.0459

Table 2. Results of multiple imputation with X_{i3}-missing rate of 0.124
表 2. X_{i3} 缺失率为 0.124 下多重插补结果

参数	真实值	估计值	偏差	标准差	均方误差
β_0	−1	−0.6387	0.3613	0.4202	0.1494
β_1	−1	−0.8187	0.1813	0.2520	0.0389
β_2	1	1.0878	0.0878	0.1173	0.0085
β_3	−1	−1.3097	0.3097	0.1652	0.0979

Table 3. Results of multiple imputation with X_{i3}-missing rate of 0.42
表 3. X_{i3} 缺失率为 0.42 下多重插补结果

参数	真实值	估计值	偏差	标准差	均方误差
β_0	−1	−1.2652	0.2652	0.5455	0.1674
β_1	−1	−0.8474	0.1526	0.2317	0.0278
β_2	1	0.9310	0.0690	0.1013	0.0050
β_3	−1	−0.7056	0.2944	0.2302	0.1125

由以上表 1~3 可知，协变量 X_{i3} 不同缺失率下参数的偏差都不大，且各参数的标准差和均方误差也很可观，说明对于随机缺失情况，多重插补效果不错。

4. 实例论证

实例数据是德国的信用评估数据，来源于 UCI 数据集，总样本量为 1000，协变量的维度为 11。实例

数据是不存在缺失的数据集。本文模拟协变量支票账户状态和财产存在随机缺失的情况下，将多重插补后 Logistic 回归得到的估计参数与完整无缺失数据 Logistic 回归得到的参数进行对比分析。

4.1. 数据预处理

4.1.1. 变量说明

本文选取德国信用评估数据中 12 个变量，分别为现有支票账户状态、月持续时间、信用记录、个人身份和性别、财产、年龄、其他分期付款计划、电话、是否外籍和目前就业情况，因变量为是否违约。各变量说明见表 4。

Table 4. Variable description
表 4. 变量描述

变量	变量描述	变量	变量描述
支票账户状态	1：... < 0 德国马克；2：0 ≤ ... < 200 德国马克；3：≤200 德国马克；4：没有支票账户	信用记录	0：没有贷记/所有贷记都已按时归还；1：这家银行的所有贷款都已按时偿还；2：到目前为止已按时归还的现有信贷；3：过去延迟付款；4：其他现有信贷(不在本行)
年龄	取值在[19, 75]之间	其他分期付款计划	1：银行；2：商店；3：没有
储蓄账户/债券	1：... < 100 德国马克；2：100 ≤ ... < 500 德国马克；3：500 ≤ ... < 1000 德国马克；4：... ≥ 1000 德国马克；5：未知/无	目前就业情况	1：失业；2：...不到 1 年；3：1 ≤ ... 不到 4 年；4：4 ≤ ...不到 7 年；5：... ≥ 7 年
电话	1：无；2：有，登记在顾客名下	是否外籍	1：是的；2：不是

支票账户状态取值为 1~4。账户资金少于 0 德国马克取值为 1；在[0, 200)之间取值为 2；大于等于 200 德国马克取值为 3；没有支票账户就取值 4。

信用记录取值为 0~4。0 意味着不存在信用记录或所有贷记都已按时归还；1 意味着该银行的全部借款均已按期偿付；1 表示这家银行的所有贷款都已按时偿还；2 表示到目前为止已按时归还的现有信贷；3 表示过去延迟付款；4 表示其他不在本行的现有信贷。

储蓄账户或债券取值为 1~5。1 表示账户资金小于 100 德国马克；2 表示账户资金 100~500 德国马克；3 表示账户资金为 500~1000 德国马克；4 表示账户资金大于等于 1000 德国马克；5 表示未知或无储蓄账户。

目前就业情况取值为 1~5。1 表示失业；2 表示就业不到 1 年；3 表示就业 1~4 年；4 表示就业 4~7 年；5 表示就业大于等于 7 年。

当性别为男且处于离婚或分居时，身份和性别取值为 1；女性处于离婚或分居或已婚时，身份和性别取值为 2；是单身男性时身份和性别取值为 3；已婚或丧偶男性身份和性别取值为 4；单身女性身份和性别取值为 5。

财产取值为 1~4。1 指不动产房产；2 指非不动产，但可供住房互助社之存款协定或人寿保险；3 表示不是前两种情况而是汽车或其它，不在储蓄账户或债券中；4 表示未知或无属性。

数据年龄和月持续时间存在异常值，两者箱线图如下图 1：

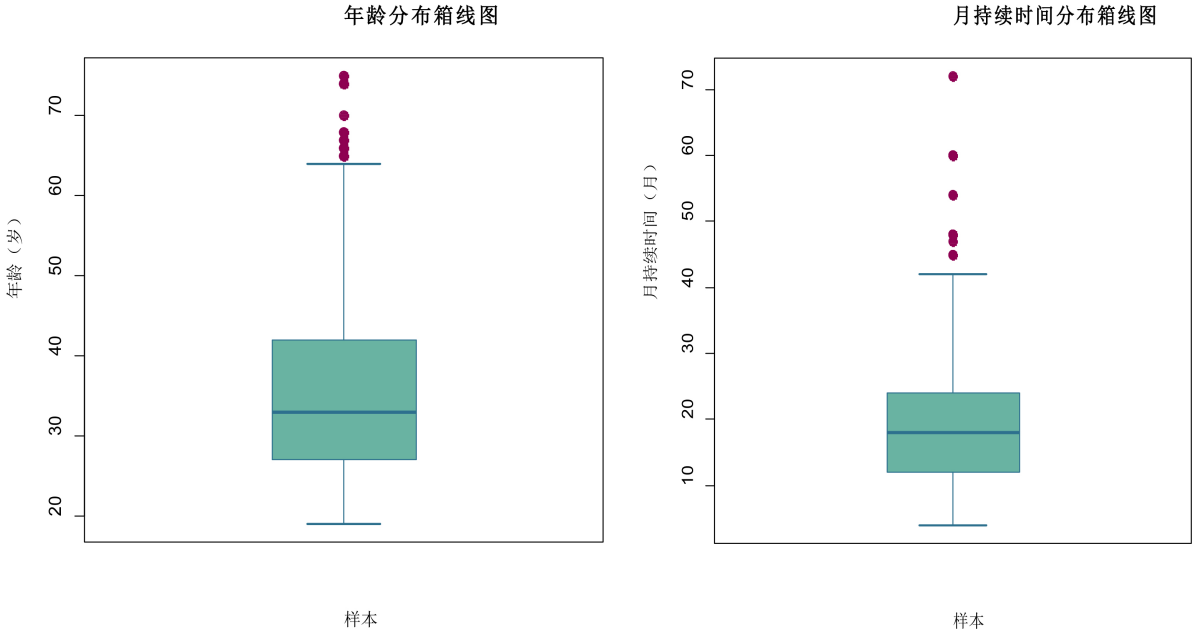

Figure 1. Box-and-Line Diagram
图 1. 箱线图

剔除离群点及数据中的重复数值，获得(903, 12)维的数据。然后，为了去除量纲，对数据进行标准化。

4.1.2. 变量筛选

对于变量筛选，首先是进行协变量与因变量以及因变量之间的 pearson 相关性检验，结果分别见表 5 和图 2。

Table 5. Pearson correlation test between covariates and dependent variables
表 5. 协变量与因变量的 pearson 相关性检验

协变量	支票账户状态	月持续时间	信用记录	储蓄账户/债券	目前就业情况	身份和性别
p 值	2.44×10^{-30}	6.49×10^{-12}	2.42×10^{-13}	1.21×10^{-8}	0.00023679	0.0052613
协变量	财产	年龄	其他分期付款计划	电话	是否外籍	
p 值	5.97×10^{-6}	0.0039253	0.000501853	0.2492787	0.00941192	

由表可知，除电话与因变量的相关性检验 p 值外，其余协变量与因变量的相关性检验 p 值都小于 0.05，拒绝原假设，即认为协变量与因变量之间不是独立的。考虑删除电话变量。

从图 2 中我们可以看到，各个协变量并不是完全独立的，它们之间的相关系数的绝对值大小在[0.001907967，0.350847483]之间(除开变量自身相关性)。

由于本文建立的是协变量与因变量的 Logistic 回归模型，故在筛选变量环节分别进行协变量与因变量的单因素 Logistic 模型显著性检验和协变量与因变量的 Logistic 模型显著性检验。检验结果如下表 6、表 7。

协变量与因变量的单因素 Logistic 模型系数显著性检验中，电话的 p 值为 0.249，大于 0.05，之前的相关性检验也没有通过，故删除电话这一列变量。其余协变量与因变量的 p 值均小于 0.05，通过检验。

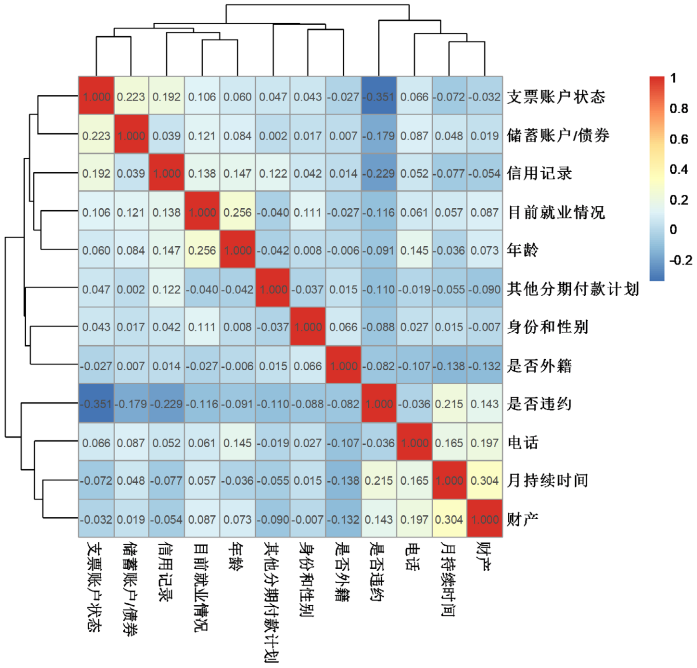

Figure 2. Heatmap of correlation of variables
图 2. 各变量之间相关性热图

Table 6. Significance test of univariate logistic model for covariates and dependent variables
表 6. 协变量与因变量的单因素 Logistic 模型显著性检验

变量	参数估计	标准误	z 值	Pr(>|z|)
截距项	−0.84859	0.06909	−12.283	<2e−16***
电话	−0.08019	0.06957	−1.153	0.249

注：***表示 $p < 0.001$，表明结果在 99.9%的置信水平下具有极强的统计显著性。

Table 7. Logistic model significance test of covariates and dependent variables
表 7. 协变量与因变量的 Logistic 模型显著性检验

变量	参数估计	标准误	z 值	Pr(>|z|)
截距项	−1.13034	0.08777	−12.878	<2e−16***
月持续时间	0.39605	0.08061	4.913	8.97e−07***
信用记录	−0.34714	0.08346	−4.159	3.19e−05***
储蓄账户/债券	−0.34144	0.08952	−3.814	0.000137***
目前就业情况	−0.16461	0.08215	−2.004	0.045094*
身份和性别	−0.17073	0.07810	−2.186	0.028818*
财产	0.19460	0.08476	2.296	0.021678*
年龄	−0.12717	0.08449	−1.505	0.132273
其他分期付款计划	−0.20514	0.07514	−2.730	0.006333**
是否外籍	−0.23383	0.11427	−2.046	0.040728*

注：***表示 $p < 0.001$，表明结果在 99.9%的置信水平下具有极强的统计显著性；**表示 $p < 0.01$，表明结果在 99%的置信水平下具有高度统计显著性；*表示 $p < 0.05$，表明结果在 95%的置信水平下具有统计显著性。

由表可知，只有年龄的 p 值大于 0.05，考虑删除。

4.2. 建模分析

4.2.1. 完整数据集的 Logistic 回归建模

考虑删除变量年龄，建立完整数据集的 Logistic 回归建模如下表 8：

Table 8. Logistic regression modeling results of the complete dataset
表 8. 完整数据集的 Logistic 回归建模结果

变量	参数估计	标准误	z 值	Pr(>|z|)
截距	−1.21997	0.09314	−13.098	<2e−16***
支票账户状态	−0.74681	0.09127	−8.183	2.78e−16***
月持续时间	0.30267	0.08655	3.497	0.000470***
信用记录	−0.34274	0.08841	−3.877	0.000106***
储蓄账户/债券	−0.23958	0.09276	−2.583	0.009799**
目前就业情况	−0.27569	0.08588	−3.210	0.001326**
身份和性别	−0.16453	0.08279	−1.987	0.046895*
财产	0.18505	0.08851	2.091	0.036553*
其他分期付款计划	−0.18261	0.07927	−2.304	0.021249*
是否外籍	−0.23383	0.11427	−2.046	0.040728*

注：***表示 $p < 0.001$，表明结果在 99.9%的置信水平下具有极强的统计显著性；**表示 $p < 0.01$，表明结果在 99%的置信水平下具有高度统计显著性；*表示 $p < 0.05$，表明结果在 95%的置信水平下具有统计显著性。

从表可以看出，筛选后的每个变量与因变量建立的 Logistic 回归模型中系数都是显著的，它们的 p 值都明显小于 0.05。

将以上变量分别依续赋为 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8 和 X_9，令是否违约为 Y。完整数据集的 Logistic 回归模型如下(保留小数点后两位)：

$$\begin{aligned}\operatorname{logit}\left(P\left(Y=1\right)\right) &= -1.22-0.75X_1+0.3X_2-0.34X_3-0.24X_4\\ &\quad -0.28X_5-0.16X_6+0.19X_7-0.18X_8-0.27X_9\end{aligned}$$

从模型表达式可以看出，月持续时间和财产与违约概率是正相关的，月份持续时间越久，财产越少，违约的概率就越大，符合实际情况；除截距项外，支票账户状态系数绝对值最大，该变量最为重要，每增加一单位该变量，违约概率的对数几率就下降 0.75，支票越多违约概率越低，也符合实际情况。

4.2.2. 多重插补后 Logistic 回归建模

根据金融风控领域的实际经验，储蓄账户/债券、财产和其他分期付款计划这三个变量是高缺失风险变量，三者缺失概率通常都大于 20%。故在本文假设储蓄账户/债券、财产和其他分期付款计划这三个变量存在缺失，即 X_4、X_7 和 X_8 存在缺失(发生缺失后分别用 mX_4、mX_7 和 mX_8 表示)。

储蓄账户/债券随机缺失机制如下：

$$\operatorname{logit}\left(p\left(r_{i4}=0\middle|y_i,X_{obs,i},c\right)\right)=1-2X_1-2X_2-0.3X_3-0.4X_5-X_6-X_9-0.3Y$$

财产随机缺失机制如下：

$$\operatorname{logit}\left(p\left(r_{i7}=0\middle|y_i,X_{obs,i},c\right)\right)=1-2X_1-2X_2-0.5X_3-0.2X_5-X_6+X_9-0.1Y$$

其他分期付款计划随机缺失机制如下：

$$\text{logit}\left(p\left(r_{i8}=0\middle|y_i,X_{obs,i},c\right)\right)=1-2X_1-2X_2-0.3X_3-0.4X_5-X_6-X_9-0.2Y$$

总共 903 个样本中，有 315 个样本不存在任何缺失，2 个样本缺失 X_8 这一个变量，30 个样本缺失 X_7 这一个变量，一个样本同时缺失 X_7 和 X_8 这两个变量，30 个样本同时缺失 X_4 和 X_8 这两个变量，525 个样本同时缺失 X_4、X_7 和 X_8 这三个变量。变量 X_4 即储蓄账户/债券缺失率为 555/903 = 0.615，变量 X_7 即财产缺失率为 556/903 = 0.616，变量 X_8 即其他分期付款计划缺失率为 558/903 = 0.618。

运用 mice 包中 mice 函数进行 5 次插补，对于连续型变量分别进行基于正态假设的插补(norm)和预测均值匹配(pmm)插补，插补结果如下图 3、图 4：

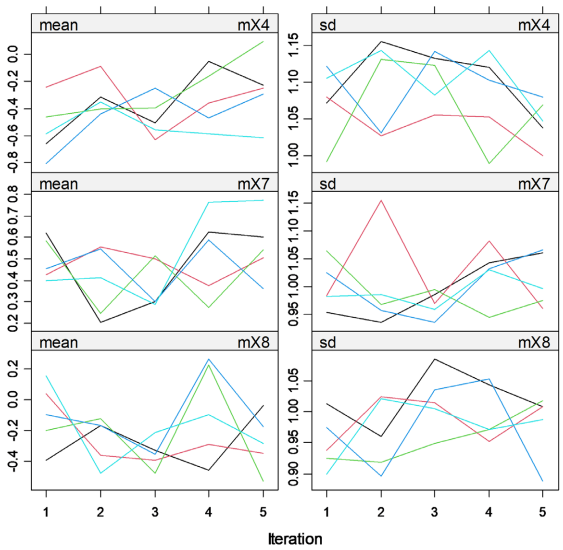

Figure 3. Missing variable interpolation trace map of norm
图 3. norm 方法的缺失变量插补线迹图

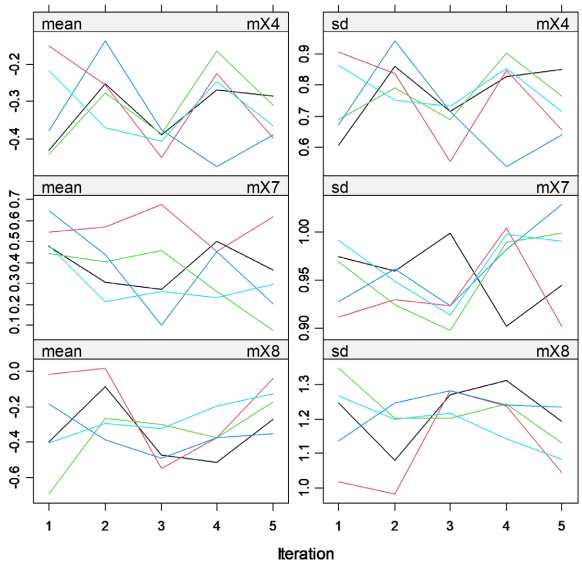

Figure 4. Missing variable interpolation trace map of pmm
图 4. pmm 方法的缺失变量插补线迹图

由图可以看出，储蓄账户/债券、财产和其他分期付款计划的均值以及标准差的插补线迹充分混合且没有趋势，表明算法收敛良好。

运用 With 函数，将插补得到的 5 个“完整数据”做相同的回归建模，再用 Pool 函数，得到 Logistic 回归建模参数的综合结果。见下表 9、表 10：

Table 9. Multiple interpolation modeling results for the norm method
表 9. norm 方法的多重插补建模结果

参数	完整数据	多重插补法	偏差	标准差	2.5%置信下限	97.5%置信上限
β_0	−1.220	−1.484	0.264	0.18	−1.89	−1.0735
β_1	−0.747	−0.715	0.032	0.11	−0.95	−0.4839
β_2	0.303	0.405	0.102	0.09	0.23	0.5818
β_3	−0.343	−0.245	0.098	0.12	−0.48	−0.0065
β_4	−0.240	−0.078	0.162	0.2	−0.58	0.4239
β_5	−0.276	−0.329	0.053	0.1	−0.54	−0.1199
β_6	−0.165	−0.022	0.143	0.12	−0.29	0.2437
β_7	0.185	0.265	0.08	0.16	−0.1	0.6302
β_8	−0.183	−0.565	0.382	0.23	−1.13	−0.0012
β_9	−0.265	−0.293	0.028	0.13	−0.55	−0.0344

Table 10. Multiple interpolation modeling results for the pmm approach
表 10. pmm 方法的多重插补建模结果

参数	完整数据	多重插补法	偏差	标准差	2.5%置信下限	97.5%置信上限
β_0	−1.220	−1.381	0.161	0.114	−1.609	−1.153
β_1	−0.747	−0.728	0.018	0.112	−0.958	−0.499
β_2	0.303	0.364	0.061	0.092	0.181	0.547
β_3	−0.343	−0.292	0.051	0.093	−0.475	−0.108
β_4	−0.240	−0.088	0.151	0.161	−0.442	0.266
β_5	−0.276	−0.312	0.036	0.094	−0.498	−0.126
β_6	−0.165	−0.086	0.078	0.095	−0.277	0.104
β_7	0.185	0.282	0.097	0.135	−0.021	0.584
β_8	−0.183	−0.414	0.232	0.107	−0.653	−0.176
β_9	−0.265	−0.283	0.018	0.134	−0.548	−0.019

由表可知，变量的多重插补建模拟合效果两种方法还是很不错的，运用两种多重插补法得到的参数同完整数据得到的参数正负相同，偏差以及标准差都不大。方法 pmm 整体上优于 norm 方法。

基金项目

重庆理工大学研究生创新项目资助(gzlcx20233310)。

参考文献

[1] 金勇进, 邵军. 缺失数据的统计处理[M]. 北京: 中国统计出版社, 2009.

[2] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.1093/biomet/63.3.581

[3] 樊思敏. Logistic 模型缺失协变量的半参数方法研究及其应用[D]: [硕士学位论文]. 长春: 长春理工大学, 2019.

[4] 庞新生. 缺失数据多重插补处理方法的算法实现[J]. 统计与决策, 2012(11): 88-90.

[5] Kim, J.K. and Shao, J. (2013) Statistical Methods for Handling Incomplete Data. Chapman & Hall.

[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.

[7] Tan, M.T., Tian, G.L. and Ng, K.W. (2010) Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation. Chapman & Hall.