自变量有缺失的分类问题

李艳芳　　苗旺*†

2013 年 4 月 11 日

摘要

本文研究了自变量有大量缺失时的分类问题和变量选择问题。使用 Logistic 回归模型分类，Lasso 方法选择变量，对缺失数据分别使用完全数据分析、完全随机填补、先学习缺失机制再按缺失机制填补方法进行处理，并做了比较。以安贞医院数据为实际例子，结果表明，按照学习到的缺失机制填补后选择出的模型表现最好，在检验数据集上的误判率 < 0.1。

*北京大学 (Peking University)　　Email: mwfywx@gmail.com

†本文第一章主要由李艳芳完成，第二章第三节及之后的内容由苗旺完成。数据分析共同完成

Contents

1 整理数据

1.1 数据介绍

安贞意义数据中包含 1214 个病人的 73 项指标，其中 37 个术前指标， 36 个术后指标。在这 1214 个观测中只有不到 500 个完全观测。**无复流**是二值变量，目标是根据其他变量建立模型对有无复流进行分类。

如果仅适用完全观测建立模型，将丢失很大一部分信息，使得结果的方差变大，而且可能有很大偏差。所以处理缺失数据成为这个数据集的关键要点。

在 73 个指标中有六个指标（身高、体重、apoa1、apob、内皮素、术中腺苷）完全缺失，所以将其删除。剩下的 67 个指标中，术前指标 32 个，术后指标 35 个。这些指标中有 13 个指标完全观测到，但是在 1214 个观测中只有很少（20 左右）的观测与其他观测所处水平不同，考虑到样本量较大，这样的指标对于最后的分类影响很小，将其删除，最终得到 54 个指标。对这 54 个指标的缺失情况进一步分析发现，其中 26 个指标没有缺失数据，18 个指标缺失数据在整个 1214 个观测中所占比例较小，而另外 13 个指标存在严重的缺失，见表1。

Table 1: 安贞医院数据：变量缺失个数

缺失较少的变量（个数）				缺失很多的变量（个数）			
年龄	2	术前TIMI血流	2	最大扩张压力	169	肌酐	300
罪犯血管血栓数量	2	侧枝循环分级	2	高敏C反应蛋白	300	TNI	300
HDLC	3	入院诊断	16	PCI前CK	300	PCI前CKMB	300
症状到PCI时间	16	收缩压	19	白蛋白	301	血红蛋白	303
舒张压	19	性别	20	LPa	310	球囊扩张次数	315
罪犯血管	27	病变支数	27	BNP	328	支架长度1	341
killip分级	31	心率	48	吸烟史	122		
梗死部位	48	病变位置	48				
罪犯血管狭窄程度	66	中性粒细胞	1				

1.2 数据整理

完全随机填补缺失很少的变量。对于缺失较少的 17 个指标，认为它们是随机缺失对结果影响不大，所以采用完全随机填补的方法将其补充完整，这时候对于这 44 个变量得到了完整的数据集。

对属性变量构造哑变量。对属性变量，取变量等级数 -1 个二值的哑变量代替这个变量。对吸烟史、梗死部位、killip分级、罪犯血管、病变位置、术前TIMI血流、侧枝循环分级、病变支数八个变量构造相应的哑变量。在建立模型时，使用哑变量。

对取值范围很大的变量取对数。有一些变量取值范围很大，小则十以内，大则几千，因此，对其取值取对数($\log(x+0.1)$)，对PCI前CK、PCI前CKMB、肌酐、高敏C反应蛋白、TNI 取对数。

离散化取值奇怪的变量。LPa、BNP 大部分取值为零，其余取值也不规律，因此将其作为二值变量处理。

经过以上的数据整理得到的数据集，就方便数学运算了，在下文所用的数据都是以该数据集为基础。

2 理论

2.1 分类问题 与logistic 回归

Logistic 回归模型是统计学习中经典的分类器，属于对数线性模型。给定一个数据集 $(x_1, y_1)(x_2, y_2)\dots(x_n, y_n)$ ，有监督学习构建模型 $P(Y|X;\theta)$ 。模型主要是参数 θ 的学习过程。当响应变量是二分类问题时，我们选择 logistic 分布来描述 $P(Y|X;\theta)$ 。不妨设 $\pi = P(y = 1|x, \theta)$,则我们设 π 的 logit 变换是自变量的线性函数，即：

$$logit(\pi) = \sum_{i=1}^{n} x_i^T \beta$$

其中 logit 变换是 $logit(\pi) = ln\frac{pi}{1-pi}$ 。这样的回归模型服从下面的条件概率：

$$P(y=1|x) = \frac{e^{\theta^T x}}{1+e^{\theta^T x}}$$
$$P(y=0|x) = \frac{1}{1+e^{\theta^T x}}$$

给定数据后，拟合 logistic 回归模型不采用通常的最小二乘方法，而是采用最大似然法。对于拟合的结果，检验模型或参数的显著性，也不是使用线性模型中的方差分析方法，而是使用与极大似然估计法相联系的卡方统计量进行检验。学习到参数 θ 后，我们利用反 logistic 变换计算得到概率值 $P(y=1|x,\theta)$ ，给定阈值 λ ，当 $P(y=1|x,\theta) > \lambda$ 时，我们将其归为一类，否则归为另一类。这就是 logistic 分类器原理。

2.2 选择变量

当我们使用数据训练分类器的时候，很重要的一点就是要在过度拟合和拟合不足之间达成一个平衡。防止过度拟合的方法就是对模型的复杂度进行约束。模型中解释变量的个数是模型复杂度的一种体现。控制解释变量个数有很多方法，例如变量选择 （feature selection） ，即用 filter 或 wrapper 方法提取解释变量的最佳子集。或变量提取 （feature structure） ，即将原始变量进行某种映射或变换，如主成分方法或因子分析。我们这里使用的是变量选择的方法 lasso[1]（least absolute shrinkage and selection operator）。

lasso 是对回归系数的绝对值之和小于等于一个常数 λ 的约束条件下，使得残差平方和达到最大，来产生某些严格等于零的回归系数，从而达到变量选择的目的。即

$$\hat{\beta} = \arg\max \ell(\beta)$$
$$subject\ \ to \sum_{j=1}^{p} |\beta_j| \leq \lambda$$

其中 $\ell(\beta)$ 表示残差平常和， λ 是调整参数且满足 $\lambda > 0$ ，交叉验证偏似然 （cross-validatin likelihoods） 取最大值时对应的 λ 为最优的 λ 。通过这样的惩罚项，假设 $\hat{\beta}^0$ 是 $\ell(\beta)$ 取最大时得到的系数估计值，如果 $\sum_{j=1}^{p} |\hat{\beta_j}^0| \leq \lambda$ ，那么由 lasso 得到的估计值就是普通的最小二乘估计，若 $\sum_{j=1}^{p} |\hat{\beta_j}^0| >$

λ，则由 lasso 得到的估计值就必须将一部分自变量的系数压缩到零来满足 $\sum_{j=1}^{p} |\hat{\beta_j}^0| \leq \lambda$ 这个约束条件。所以 lasso 的意义就在此，通过把一些无意义或意义极小的自变量的系数压缩到零，筛选出更有意义的自变量使得模型更加精炼，更容易解释。

2.3 缺失数据问题的基本概念

本小节介绍缺失数据问题的形式化，缺失机制，向无环图(DAG)模型，利用 DAG 表示缺失机制。

2.3.1 缺失机制

Rubin(1976)[2] 提出了缺失数据是一种概率现象的观点，并建立了缺失机制的理论。关心的变量为 $Y = (Y_1, Y_2, \cdots Y_p)$，完全数据集记为 $Y = (Y_{i\,obs}, Y_{i\,mis})$，$Y_{obs}$ 是观测到的数据；Y_{mis} 是缺失的数据。定义矩阵 $R = (R_{i\,1}, R_{i\,2} \cdots R_{i\,p}), i = 1, 2 \cdots n$，若 Y_{ik} 观测到，$R_{ik} = 1$，否则 $R_{ik} = 0$；Θ 是与数据有关的参数。矩阵 R 称为*指示矩阵*，反映了个体水平上缺失数据的模式。在研究中经常关心的是总体水平上的数据的缺失模式，缺失机制定义为给定 Y 时 R 的条件分布。

定义 1:（缺失机制） 条件概率 $P[R|Y, \Theta]$ 称为缺失机制。

完全随机缺失(MCAR) $P[R|Y, \Theta] = P[R|\Theta]$,

随机缺失(MAR) $P[R|Y, \Theta] = P[R|Y_{obs}, \Theta]$

非随机缺失(MNAR) $P[R|Y, \Theta] = P[R|Y_{obs}, Y_{mis}, \Theta]$

完全随机缺失和随机缺失表明观测数据能反映缺失机制，非随机缺失表明观测数据不能反映缺失机制，这种情况确定缺失机制需要数据之外的信息。缺失机制还可以用*有向无环图(DAG)描述*。

2.3.2 *DAG*模型描述缺失机制

Fay(1986)[3]提出用*有向无环图(DAG)*描述缺失机制。

定义 2：（有向无环图(DAG)） 首先定义有向图，由结点集合 V 和有向边集合 E 组成的集合称为一个有向图，记为 $G =< V, E >$。其中结点集合 $V = \{V_i\}$ 由给定的结点元素 V_i 组成；一条从 V_i 指向 V_j 的有向边用有序结点对 $< V_i, V_j >$ 表示，有向边集合 $E = \{< V_i, V_j >, V_i, V_j \in V, i \neq j\}$ 。
若有向图 G 没有有向环，称为有向无环图。在有向无环图中，若 $< V_i, V_j > \in E$ ，称 V_i 是 V_j 的父节点， V_j 是 V_i 的子结点，V_i 所有的父结点集合记为 $Pa(V_i)$ 。

把随机变量作为结点，把概率测度引入到*有向无环图*，有下面的定理。

定理 1：有向无环图 $G =< V, E >$ 上 V 的联合概率分布可分解为

$$P(V_1, V_2, \cdots, V_n) = \prod_{i=1}^{n} P[V_i | Pa(V_i)] \tag{2.1}$$

有向边直观地说明变量之间的依赖关系，在 DAG 中还可以反映变量之间的独立性关系。*d-分离*的概念说明边和变量之间的条件独立性的对应关系。

定义 3：（路径 $d-$分离） 在 DAG 中，称路径 p 被结点集 Z *d-*分离，若 p 包含子路径 $i \to m \to j$，或 $i \leftarrow m \to j$，且 $m \in Z$ ；或包含子路径 $i \to m \leftarrow j$，且 m 和其后代 $\in Z$ 。

定义 4：（结点 $d-$分离） 称结点集 X, Y 被结点集 Z $d-$分离，若 X 到 Y 的任一路径被 Z $d-$ 分离。

有了 *d-分离*的概念，可以从利用条件独立性构造出一个满足条件的 DAG，但还不能保证从 DAG 惟一确定独立性条件，下面的*忠实性假定*使条件独立性和 DAG 一一对应起来。

假定 1：（忠实性假定） P 是变量集 Y 上的概率测度，令 $I(P)$ 代表 P 中所包含的所有条件独立关系的集合，一个有向无环图模型 (G, Θ_G) 能生成一个稳定的联合分布当且仅当 $P(G, \theta_G)$ 不包含额外的独立条件，即 $I(P(G, \Theta_G)) \subseteq P(G, \Theta_{G'})$ 对任意参数集合 $\Theta_{G'}$ 成立

*忠实性*假定保证了变量 $d-$分离和条件独立性的对应关系。

定理 2：变量 X, Y 在对应的 DAG 中被变量集 Z d-分离当且仅当对于每一个

与该 DAG 一致的概率分布都有给定 Z 时 X 与 Y 条件独立，即

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_P$$

把关心的变量和相应的指示变量作为 DAG 的节点，引入概率测度，根据分解

$$P(Y_1, Y_2, \cdots, Y_p, R_1, R_2, \cdots, R_p) = P(Y_1, Y_2, \cdots, Y_p)P[R_1, R_2, \cdots, R_p|Y_1, Y_2, \cdots, Y_p]$$

可用 $(Y_1, Y_2, \cdots, Y_p)$ 指向 $(R_1, R_2, \cdots, R_p)$ 的有向边表示缺失机制。

2.3.3 识别性

当有缺失数据时面临的首要问题就是可识别性。可识别之后进行统计推断才有意义。T. J. Rothenberg(1971)[4]对参数模型的可识别性做了详细研究，其定义同样适用于缺失数据问题。设观测到的数据为$\{X, R, RY\}$，潜在的联合分布为 $P(X, Y, R)$，或写成参数形式 $P(X, Y, R; \theta)$。则边缘分布 $P[Y|R = 1]$，$P(R = 0)$ 可由观测数据得到。

定义 5：（联合分布的识别性） 称联合分布 $P(X, Y, R)$ 可识别，若

$$\left.\begin{aligned} P_1[Y|R=1] &= P_2[Y|R=1] \\ P_1(R=0) &= P_2(R=0) \end{aligned}\right\} \Rightarrow P_1(X, Y, R) = P_2(X, Y, R) \qquad a.e.$$

含参数的分布可识别性

$$\left.\begin{aligned} P[Y|R=1;\theta_1] &= P[Y|R=1;\theta_2] \\ P(R=0;\theta_1) &= P(R=0;\theta_2) \end{aligned}\right\} \Rightarrow P(X, Y, R;\theta_1) = P(X, Y, R;\theta_2) \qquad a.e.$$

定义 6：（参数可识别性） 称参数 θ 可识别，若

$$\left.\begin{aligned} P[Y|R=1;\theta_1] &= P[Y|R=1;\theta_2] \\ P(R=0;\theta_1) &= P(R=0;\theta_2) \end{aligned}\right\} \Rightarrow \theta_1 = \theta_2$$

3 处理方法

本章分析安贞医院数据，使用 Logistic 模型分类，Lasso 方法选择变量；对于缺失数据，采用不同方法处理，并说明这些处理方法的使用条件。本章的记号：Y 为响应变量**无复流**；X 为自变量。

3.1 Logistic 回归模型和 Lasso 选择变量

模型 1：（Logistic 回归模型）

$$\log\left(\frac{P[Y=1|X]}{1-P[Y=1|X]}\right)=\beta_0+X\beta_1$$

安贞医院的数据自变量有 55 个，存在多重共线性，使用 Lasso 方法可以方便地选择变量。R 软件包 $glmnet$ 可以实现 Logistic 回归模型的 Lasso 方法选择变量。但是，安贞医院的数据有大量缺失，需要另外的模型处理这些缺失数据。

3.2 处理缺失数据

3.2.1 使用完全数据分析(CompleteAnalysis)

大部分软件处理缺失数据采用完全数据分析方法，即删除有缺失的个体，只用全部变量都观测到的个体分析。安贞医院数据的完全观测的病人有 446 个，约为所有病人的 1/3。

Algorithm 1: CompleteAnalysis

1. 提取完全数据；
2. 以完全数据作为训练样本，剩下的作为检验数据；
3. 用训练样本拟合模型1，用 $glmnet$ 选择变量；
4. 用选择出来的模型判别检验数据，计算误判率。

3.2.2 完全随机填补(MCARImp)

对于有缺失的变量，从其观测到的个体中等概率抽取，填补缺失的观测。

Algorithm 2: MCARImp

1. 分别对每一个变量进行完全随机填补，得到填补后的数据；
2. 从填补后的数据随机抽 900 个病人数据作为训练样本，剩下的 314 个作为检验数据；
3. 用训练样本拟合模型1，用 *glmnet* 选择变量；
4. 用选择出来的模型判别检验数据，计算误判率。

3.2.3 按照缺失机制填补(MechImp)

一些先验信息可以帮助建立缺失机制。安贞医院数据的自变量可以分为几部分：个体属性数据，病史数据；用药史数据；PCI 前测量数据；PCI 中测量数据；PCI 后测量数据。时间在先的变量会影响时间在后的变量和其缺失情况，而且，在观测了这么多变量的情况下，是否缺失对缺失的观测的依赖性变小，有充分的理由认为缺失机制如图1

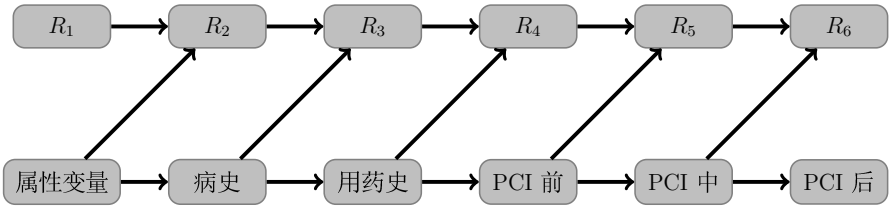

Figure 1: 安贞医院数据：缺失机制 1

图1 中每个方格子中是一个向量，对每个方格子中的缺失机制，认为是，以吸烟史和糖尿病史为例，

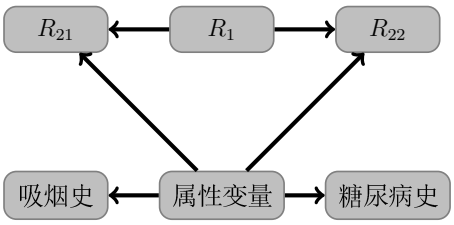

Figure 2: 安贞医院数据：缺失机制 1.1

定理 3：在图1 和 2 的缺失机制下，如果所有变量是二值变量，则其联合分布可识别。

图1 和 2 的缺失机制虽然可以识别，但分析起来仍然复杂，回到第一章介绍安贞医院数据，属性变量，病史，用药史，PCI 中测量数据缺失很少，PCI 前测量数据只有血红蛋白等几个变量缺失较多，PCI 后测量数据只有最大扩张压力等几个变量缺失较多，因此，一个方便的方法就是把这些缺失很少的变量用完全随机缺失机制填补变成完全观测的变量，这时缺失机制变为图3，并且方格子内的局部机制类似图2。

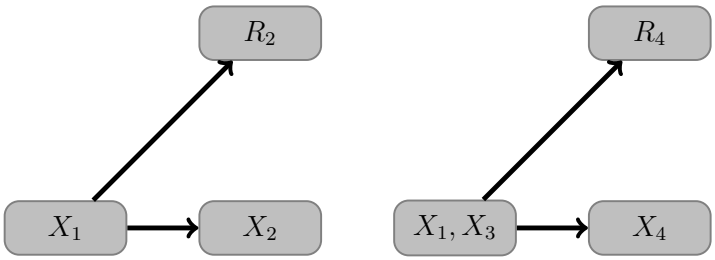

Figure 3: 安贞医院数据：缺失机制 1.2

定理 4：在图3 和 2 的缺失机制下，其联合分布可识别。

在填补缺失的变量时就可以给定完全观测到的变量，从观测到的个体中等概率抽样进行填补。这些完全观测到的变量仍然很多，对缺失指示变量建立 Logistic 回归模型来选择抽样时需要给定的完全观测的变量。

模型 2：

$$\log\left(\frac{P[R=1|X]}{1-P[R=1|X]}\right)=\alpha_0+X\alpha_1$$

仍然用 Lasso 方法选择变量。综上，得到按照学习到的缺失机制填补缺失数据并选择变量的步骤：

Algorithm 3: MechImp

1. 从数据集中分离出变量 X_1, X_2, X_3, X_4，对 X_1, X_3 进行完全随机填补；
2. 对 X_2, X_4 中的变量，拟合模型2，用 Lasso 选择变量；
3. 分别对 X_2, X_4 中的变量，给定第 2 步选择出来的变量进行填补，此时数据集已经没有缺失值；
4. 对第 3 步得到的数据拟合模型1，用 Lasso 选择变量；
5. 用第 4 步得到的模型判别检验数据，计算误判率。

3.2.4 缺失数据的似然方法

数据:

Y 为响应变量；

X 为完全观测的协变量，$\widetilde{X} = \{X_1, \cdots, X_i\}$ 为当前 Logistic 模型选入的完全观测变量集；

M 为有缺失的变量，$\widetilde{M} = \{M_1, \cdots, M_k\}$ 为当前 Logistic 模型选入的完全观测变量集;

R 表示缺失变量是否缺失，$\widetilde{R} = \{R_1, \cdots, R_j\}$ 为当前 Logistic 模型选入的完全观测变量集;取值 1 表示该变量的该观测缺失，0 表示没有没有缺失；

$H = \{H_1, \cdots, H_j\}$ 表示历史测量变量,$\widetilde{H} = \{H_1, \cdots, H_k\}$ 为当前 Logistic 模型选入的完全观测变量集0

模型：

模型 3 :

$$P[Y = 1|X, R] = P[Y|X] = P[Y = 1|\widetilde{X}, \widetilde{M}]$$

$$P[Y = 1|\widetilde{X}, \widetilde{M}] = \frac{1}{1 + \exp\{\alpha_0 + (\widetilde{X}\widetilde{M})\alpha_1\}}$$

模型 4 :

$$P[R_1|X] = P[R_1|H]$$

$$P[R_2|X] = P[R_2|M_1, R_1]$$

$$P[R_3|X] = P[R_3|M_1, R_1]$$

完全似然：

$$\begin{aligned} Likelihood =& P(Y, X, R) \\ =& P[Y|X, R] \cdot P[R|X] \cdot P(X) \\ =& P[Y|\widetilde{X}, \widetilde{M}] \cdot P[\widetilde{R}|X] \cdot P[R/\widetilde{R}|X] \cdot P(X) \end{aligned}$$

由于有缺失数据，完全似然得不到，需要用观测似然：

$$\begin{aligned} Likelihood_{obs} =& P[Y|\widetilde{X}, \widetilde{M_{obs}}] \cdot P[\widetilde{R_{obs}}|X] \cdot P[R_{obs}/\widetilde{R_{obs}}|X] \cdot P(X) \\ & \cdot \int_{\widetilde{X}} P[Y|\widetilde{X}, \widetilde{M_{mis}}] \cdot P[\widetilde{R_{mis}}|X] \cdot P[R_{mis}/\widetilde{R_{mis}}|X] \cdot P(X) d\widetilde{M_{mis}} \end{aligned}$$

对数观测似然：

$$logL_{obs} = \log(Likelihood_{obs})$$

最大似然估计：

$$MLE \qquad maximize \qquad logL_{obs}$$

Logistic 模型逐步选择自变量:

Algorithm 4: 逐步选择变量的似然方法

若 Logistic 模型3 当前包含的自变量为 $\widetilde{X}, \widetilde{M}$，其最大似然为 $\log \widehat{L(\widetilde{X}, \widetilde{M})}$

1. 加入一个新的自变量 X_{add} 加入后做最大似然估计，得到最大似然 $\log L(\widehat{\widetilde{X}, \widetilde{M}, X_{add}})$；
2. 删除一个不再显著的自变量 X_{delete}，通过做似然比检验

$$-2\left(\log L(\widehat{\widetilde{X}, \widetilde{M}, X_{add}}) - \log L(\widehat{\widetilde{X}, \widetilde{M}, -X_{delete}})\right) \sim \chi_1^2$$

3. 所有变量经过加入和删除后得到一个相对满意的模型。

3.3 结果

表2 是完全数据分析，完全随机填补，按照缺失机制填补得到的结果比较，表3 是按照缺失机制填补过程中的结果。

表2 的结果表明，完全数据分析只选出 5 个自变量，用完全观测的病人训练，有缺失的病人检验，误判率为 0.1745(134/768)；完全随机填补，每次等概率抽取 900 个病人数据训练，其余 314 个检验，重复 100 次，得到误判率的 Bootsrap 均值和方差 0.1988(62.42/3140)，0.0200(6.26/314)；按照缺失机制填补，每次等概率抽取 900 个病人数据训练，其余 314 个检验，重复 100 次，得到误判率的 Bootsrap 均值和方差 0.0837(26.28/314)，0.0131(4.12/314)。在 100 次训练中，按照缺失机制填补选择出来的变量很稳定，完全随机填补选择出来的变量变化很大，前者选择出的变量缺失都很少，后者会选出缺失很多的变量如**高敏C反应蛋白，血红蛋白**等。

表3 的结果表明，这些缺失较多的变量不是完全随机缺失，填补缺失值需要给定相应的变量，**吸烟史，PCI前阿司匹林** 是填补缺失值时比较重要的变量。

总之，在选择变量的稳定性和误判率方面，按照缺失机制填补得到的结果比较好，误判率 <0.1 是非常满意的结果。

Table 2: 三种方法的比较。

	MechImp		MCARImp		CompleteAnalysis	
选出的变量	舒张压	0.01350	随机血糖	0.01385	心率	0.03197
	随机血糖	0.11431	IABP	0.45492	IABP	1.23312
	IABP	1.44246	术前TIMI血流.1	-0.74735	PCI术中2b3a	1.52833
	术前TIMI血流.1	1.43612	术前TIMI血流.2	0.45777	罪犯血管血栓数量	0.88437
	术前TIMI血流.2	-0.71785	术前TIMI血流.3	0.17832	PCI术中钙拮抗剂	4.38775
	术前TIMI血流.3	-1.43034	罪犯血管血栓数量	0.05084		
	罪犯血管血栓数量	0.85395	侧枝循环分级.1	-0.50014		
	侧枝循环分级.1	-0.36110	侧枝循环分级.2	0.74425		
	侧枝循环分级.2	-1.29207	侧枝循环分级.3	0.31336		
	侧枝循环分级.3	-1.07242	PCI术中2b3a	-0.04792		
	PCI术中2b3a	2.17533	PCI术中钙拮抗剂	0.15842		
	PCI术中钙拮抗剂	4.99602	PCI术中血栓抽吸	-0.12309		
	PCI术中血栓抽吸	1.42165	罪犯血管.1	14.72711		
	罪犯血管.1	0.08249	罪犯血管.2	0.32532		
	罪犯血管.2	-0.99854	罪犯血管.3	-0.74906		
	罪犯血管.3	0.46717	killip分级.1	-0.18404		
	killip分级.1	1.92809	killip分级.2	-0.02941		
	killip分级.2	-2.56764	killip分级.3	0.89798		
	killip分级.3	1.78512	高敏C反应蛋白	-0.11522		
误判率	26.28(4.12)/314		62.42(6.26)/314		134/768	

Table 3: MechImp 过程

被填补的变量	需要给定的变量	
血红蛋白	吸烟史	PCI前阿司匹林
白蛋白	吸烟史	PCI前阿司匹林
肌酐	吸烟史	PCI前阿司匹林
高敏C反应蛋白	吸烟史	PCI前阿司匹林
TNI	吸烟史	PCI前阿司匹林
PCI前CK	吸烟史	PCI前阿司匹林
PCI前CKMB	吸烟史	PCI前阿司匹林
LPa	吸烟史	PCI前阿司匹林
BNP	吸烟史	PCI前阿司匹林
球囊扩张次数		PCI前阿司匹林
最大扩张压力	预扩张	PCI前阿司匹林
	后扩张	PCI术中硝酸酯
支架长度1	吸烟史	PCI前阿司匹林
		既往 β 受体阻滞剂

4 结论

按照缺失机制填补的结果表明，在判断一个手术后的病人是否会出现无复流时，只需测量表2 中的变量，如果测量的变量没有这么多，也可以通过修正相应变量的系数来判断，不过误判率会大一些。

表2 的结果并不表明，若手术前病人的这些指标使得预测为无复流时就没有做手术的必要，表2 的结果只适用于手术后的病人。

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[2] D. B. Rubin. Inference and missing data (with discussion)[J]. Biometrika. 1976, 63:581 – 592

[3] R. E. Fay. Causal models for patterns of nonresponse[J]. J Amer Statist Assoc. 1986, 81:354 – 365

[4] T. J. Rothenberg. Identification in parametric models[J]. Econometrica: Journal of the Econometric Society, 1971: 577-591.

[5] R. J. A. Little, D. B. Rubin. Statistical Analysis with Missing Data[M]. Wiley, Hoboken, NJ, 2002

[6] Agresti A. Categorical data analysis[M]. Wiley-interscience, 2002.