数据分析的数学导论期末论文

利用较多缺失数据集的PCI手术成功率预测

1200010714　　马　超

1200010715　　余　冰

1200010698　　储翌尧

北京大学数学科学学院

2014.1.20

摘　要

本文中，我们利用逻辑斯谛回归以及相关的检验技术，选用合理方法选取重要的回归变量，对来自两家医院的2581名接受PCI手术的病人数据进行了分析和比对。从这一份具有较多缺失的数据中，我们最终得到一个预测手术成功与否的公式，其准确率超过80%，同时，通过显著性检验得出了两所医院在手术质量上不具有系统性差距的结论。

关键词：逻辑斯谛回归，数据缺失处理，PCI

目　录

1 问题介绍

1.1 问题背景

PCI手术，又称经皮冠状动脉介入手术，是治疗心肌梗死等严重心血管疾病的有效手段。它具有成功率高、创伤小、恢复快等特征。然而，即便这样，PCI手术也无法做到百分百成功，手术失败的案例时有发生。一般来说，手术成功与否不仅仅与医生在手术过程中的操作有关，还与接受手术的患者本身的各种情况密切相关。简单地说，就是有一些病人由于其自身原因，并不适合接受此手术。那么，如果存在一种技术可以帮助医院判断病人是否适合接受手术，那么无论是医院还是患者都将从中大大获益。本文作者将基于医院提供的大量病人数据，探究PCI 手术成功率与病人各方面身体状况的关系，以期能够给予医院以及患者一些建议和指导。

1.2 数据介绍

本文的研究所用到的数据来自于北京安贞医院和朝阳301医院，共包含2581名接受了PCI手术的病人的健康数据，这些病人中，有1214人来自于安贞医院，其余1367人来自于朝阳301医院。数据共包含73个指标，主要可以分为病人入院检查、手术前和手术后三个部分，数据涵盖面广，分类细致，具有较高的利用价值。在数据的最后，有一个二值变量用来记录手术是否成功。

本数据在样本容量、变量的丰富程度上表现得都相当不错。但是，数据缺失是这组数据存在的一个不可忽略的问题。由于数据中的2581名患者来自于不同的医院，入院接受手术的时间也各不相同，因此，并非每一名患者都有完整的73个指标的记录。大部分指标都存在不同程度的缺失现象，其中有些指标缺失甚为严重。如何对存在大量缺失的数据进行分析，并挖掘出有用的信息，这是本文作者在探究过程中遇到的主要挑战，也是需要重点解决的问题。

1.3 探究目标

基于以上介绍的数据，我们的探究目标主要为以下两点：

1、通过入院检查以及手术前的数据预测手术成功的可能性。

对于医院和医生来说，他们力求能够对每一名病人进行最高效的治疗，因此最为关注的问题是，在手术开始之前判断一名患者是否适合进行PCI手术。如果不适合，医院则需要放弃手术，选择其他的治疗手段。于是，本文需要解决

的第一个问题，就是利用手术开始前的一系列数据记录，预测特定病人的手术成功的可能性有多大。

2、探究两家医院在手术质量上是否存在系统性的差距。

换到病人的角度，作为病人，他们必然希望能够到更好的医院进行治疗。为此，我们将利用来自两所不同医院的数据，探究这两所医院手术质量方面是否存在系统性的差距，即是否其中一家医院的手术质量明显优于另外一家，以期给患者提供一些参考。

2 数据分析与处理

2.1 缺失情况分析

在数据中，“无复流”一项纪录了手术是否成功，是最为重要的指标。这一指标缺失的病人对我们的探究是没有意义的。因此，我们首先去掉了7个没有“无复流”指标的病人。此后，考察每一项指标的缺失程度，发现，几乎完全缺失的指标有6个，分别为：

身高，体重，apoa1，apob，内皮素，术中腺苷

这六个指标的非空记录均不到500个，体现出的患者信息很不完全，如果强行对这些数据进行补全，反而可能会影响原数据的代表性，因此予以除去。另外，由于数据量较大，故对于记录数不到1000，以及所有记录基本都相同的变量，同样予以除去。这一步又除去了9 个变量，剩余58个。后文中的所有分析和处理均是针对这58个变量进行的。

完成上述工作之后，再看每个病人的情况，经统计发现，此时，所有记录都存在的病人共有463个，记录不少于50个的病人有1482个；记录最少的病人只有18个变量非空，记录数少于30 的共有127人。当然如果只考虑手术前的各项数据，拥有全部数据的患者人数会有所上升，达到651名，这一点在夏文宗还会有所阐述。为了使进一步的估计和预测能够顺利进行，接下来要做的就是考虑如何补全表中仍然缺失的数据。

2.2 数据缺失处理

为了操作上的方便，我们选取了一种较为简单的缺失补全方案，即从该变量已有的数据中随机选取一个写到缺失的位置上。这一处理虽然简单，却有其合

理性。以二值变量为例，通常我们使用最大似然法估计其取1的概率。如果设已知的数据有n个，其中取1的有m个，则最大似然估计得到的结果为：

$$p = \frac{m}{n}$$

这个概率与我们在n个已有数据中随机取出一个得到1的概率是相同的。因此，随机取值法事实上并没有改变二值变量的分布，其带来的误差也并不比最大似然法带来的误差大。以上的分析对取少数几个值的属性变量也是适用的。

对于取值相对连续的变量，这样的做法虽然减小了变量的取值空间，但是考虑到已有样本容量较大（超过1000），在没有缺失的位置上，变量的取值已经比较多，因此这样的做法应该不会对变量整体的分布造成较大的影响。另一方面，这种补全数据的方法实现起来极为简单，使用一般的数学软件，如MATLAB、SPSS等，只需短短几行代码就可以完成。

3 利用术前数据的预测结果

完成数据处理之后，我们试图选择合适的模型，寻找出变量和相应变量之间的关系。对于一组自变量和一个响应变量的情形，回归分析通常是一个不错的选择。但是，由于在本数据中响应变量是一个零一变量，因此可以很自然地选择逻辑斯谛回归作为分析的模型。

3.1 逻辑斯谛回归

逻辑斯谛模型通常用于研究二分变量与一系列自变量之间的关系，其定义如下：

定义： 设Y是一个二分变量，$p = P(Y = 1)$，$x_1, x_2, ..., x_k$是自变量，如果p和x_i满足下列关系式：

$$ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

则称二分变量Y与自变量x_i的关系符合逻辑斯谛模型。

为了体现概率p与$x_1, x_2, ..., x_k$的联系，常写$p = p(x) = P(Y = 1|x_1, x_2, ..., x_k)$，并且：

$$p(x) = \frac{exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}$$

接下来的问题就是估计β_i，通常，估计的方法有两种：最大似然估计法和加权最小二乘法。下面仅就k=1的情形作一些解释，k较大的情形实际上只具有形式上的复杂性。

最大似然法

设有下列数据：$x = x_i$时Y的值是y_i，i=1,2,...,n，$y_i = 01$，则有：

$$P(Y = y_i|x_i) = (p(x_i))^{y_i}(1 - p(x_i))^{1-y_i}$$

于是观测值$(x_1, y_1), ..., (x_n, y_n)$对应的似然函数是：

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} (p(x_i))^{y_i}(1 - p(x_i))^{1-y_i}$$

从而，令

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_i} = 0 \qquad (i = 0, 1)$$

就可以得到似然方程组：

$$\sum_{i=1}^{n} (y_i - \frac{exp\{\beta_0 + \beta_1 x_i\}}{1 + exp\{\beta_0 + \beta_1 x_i\}}) = 0$$

$$\sum_{i=1}^{n} (y_i - \frac{exp\{\beta_0 + \beta_1 x_i\}}{1 + exp\{\beta_0 + \beta_1 x_i\}}) x_i = 0$$

在一定条件下，以上似然方程组存在唯一的根$(\hat{\beta}_0, \hat{\beta}_1)$，这就是逻辑斯谛问题的最大似然估计。

加权最小二乘法

此法对数据有一些特殊的要求。设$x = x_i$时对Y作了n_i次观察，且n_i较大，其中{Y=1}发生了r_i次（i=1,2,...,m），这里$x_1, ..., x_m$两两不同。通常用

$$z_i = \ln \frac{r_i + 0.5}{n_i - r_i + 0.5}$$

作为$\ln \frac{p(x_i)}{1-p(x_i)}$的估计值。令

$$v_i = \frac{(n_i+1)(n_i+2)}{n_i(r_i+1)(n_i-r_i+1)} \qquad (i=1,2,...,m)$$

$$\tilde{Q}(\beta_0,\beta_1) = \sum_{i=1}^{m} \frac{1}{v_i}(z_i - \beta_0 - \beta_1 x_i)^2$$

使$\tilde{Q}(\beta_0,\beta_1)$达到最小值的$\tilde{\beta_0}$, $\tilde{\beta_1}$分别称为β_0, β_1的加权最小二乘估计。这里，$\frac{1}{v_i}$就是所谓的权。

可以证明，加权最小二乘估计存在且唯一。

3.2 结果分析

仅考虑入院检查和手术前的数据，除去意义较小的变量后，余下37个。考察2574名患者发现，具有完全记录的患者共有651名。在本问题中，我们首先使用这651名患者的真实数据进行逻辑斯谛回归，然后用其余的患者检验回归得到的模型。我们使用的工具为统计软件SPSS。在回归的过程中我们发现，有部分变量的显著水平很低，我们逐个除去这些显著水平较低的变量，最终得到6 个高度显著的变量，它们分别是：

x2	年龄
x6	糖尿病史
x9	脑梗塞史
x20	心率
x25	血红蛋白
x32	随机血糖

将以上六个变量作为回归变量，再次进行逻辑斯谛回归，得到回归方程，相应的回归系数与显著性见下表：

变量	回归系数β	显著性水平
截距项	5.929	0.000
x2	-0.013	0.109
x20	-0.024	0.000
x25	-0.006	0.049
x32	-0.108	0.000
x6=0	-0.706	0.005
x9=0	0.456	0.198

从表中可以看到，此时除了x2和x9的显著性稍弱以外，其他变量的显著性都很好，截距项与x20的显著性甚至接近0。总的来说，这六个变量对手术结果的影响都是显著的。

代入其余患者的数据进行预测。这里我们规定，对于一名患者，如果预测出的结果p大于0.5，就预言手术不会成功，反之，如果p小于等于0.5，则预言手术会成功。这样预言的成功率为79.36% (1357/1710)。进一步分析具有不同数据完整度的患者的准确率，结果见下表：

缺失数据个数	正确预测率	比例
1	79.74%	1212/1520
2	72.58%	45/62
3	76.47%	13/17
4	78.90%	86/109
5	50%	1/2

从上表中可见，除了缺失5个数据的患者的预测正确率较低外，其余情形下的正确率均相近，而缺失5个数据的患者由于数量太少，故其正确率事实上不具有参考价值。由此可见，数据缺失个数对预测结果的影响不是很大，这说明我们对却是数据的处理有一定成效。

4 不同医院间的系统差别探究

与上一节中的做法略有不同的是，本节中我们加入了一个表示患者所在医院的自变量，若患者来自于安贞医院，则该自变量取1，否则取0。加入这一变量后，对数据重新进行逻辑斯蒂回归，而后对这一标识患者所在医院的自变量检验其显著性。检验的结果为，该变量仅在30%水平下显著，仅凭这一显著水平

我们不能够断言两个医院的系统性差别是显著的。因此，我们得到的结论是：两家医院在手术质量上不具有系统性差距，病人可以选择任何一家医院接受手术。

当然，由于来自朝阳301医院的患者普遍缺少较多数据，因此这一结果的数据基础在朝阳301医院一方稍显薄弱，如果能够根据更加完善的数据进行分析，或许结果会有所不同。

5 结论及展望

在本文中，我们首先对具有较多缺失的数据进行补全处理，使之具备进行回归分析的条件，而后使用逻辑斯谛回归探究自变量与响应变量之间的关系，进行预测和检验。最终，针对文章开头列出的两个问题，我们得到的结论是：

1、使用我们选出的较为显著的六个自变量进行逻辑斯谛回归，利用得到的回归方程对病人的手术结果进行预测，预测的准确率可以达到80%以上，并且这一结果在病人相关检测数据缺失不是很多的情况下可以得到较好的保持。

2、在已有的数据的基础上，不能说明两家医院在手术质量上具有显著的系统性差异。

当然，由于我们水平、时间有限，本文中的分析势必不能够尽善，对于这一数据的处理和挖掘，我们还有一些想法，现列于下，欢迎思考与探究：

1、数据缺失的处理

事实上，我们在本文中使用的数据缺失的处理方法，即随机填补法，是一种最简单的缺失处理方法，其优点在于操作简洁明了。但是简单的方法势必会造成原数据中一些信息的丢失。事实上，处理数据缺失的方法还有很多，如果可以将这些种种方法应用于这组数据上，有理由相信后续的分析会得到更好的结果。

2、术后数据及综合数据的分析与研究

在本文中，我们只谈探究了术前数据对手术结果的影响，诚然，无论是对医院还是病人来说，术前数据都是最为重要的参考数据，因为只有术前数据表现出的异常才有可能让医院和病人放弃手术，选择更加合适的治疗方法。但是这并不意味着术后的数据是没有价值的。手术完成后，病人需要接受一段时间的观察和护理，这时，术后数据就可以给医护人员提供病人情况的参考，并帮助他们制定更好的护理计划，这于病人有助于恢复健康，于医院有助于提高效率、降低成本。因此，对术后数据的分析与研究也是特别值得开展的一项工

作。

参考文献

[1] 李艳芳，苗旺.自变量有缺失的分类问题

[2] 陈家鼎，孙山泽，李东风，刘力平.数理统计学讲义.北京：高等教育出版社，2014

组员介绍及分工

马超

数学科学学院12级本科生，科学与工程计算方向

学号1200010714，邮箱1200010714@pku.edu.cn

承担了数据缺失的补全，本文的撰写工作，并为数据的分析工作提供了一些想法

余冰

数学科学学院12级本科生，科学与工程计算方向

学号1200010715，邮箱1200010715@pku.edu.cn

承担了数据分析工作，完成了大部分的回归分析，及显著性检验工作

储翌尧

数学科学学院12级本科生，科学与工程计算方向

学号1200010698，邮箱1200010698@pku.edu.cn

承担了资料搜寻工作，并为数据分析的工作提供了一些想法

致谢

作为三名计算方向的本科生，本学期的“数据分析的数学导论”课程为我们打开了数据分析的大门，或许我们最终不会登堂入室，但是它使我们能够窥其一斑，领略其中的奥妙，便已令我们受益匪浅。在此过程中，感谢姚老师的谆谆教诲，感谢助教老师认真批改作业所付出的努力，同时，还要感谢课上的同学们在我们疑惑时的耐心帮助。能力一般，水平有限，无所报答，谨祝新年快乐，心想事成。

2015年1月