·循证医学中的医学统计学问题·

# 回归分析中定性变量的赋值

张晋昕[1]， 李 河[2]

（1. 中山大学公共卫生学院，广州 510080；2. 广东省人民医院，广州 510080）

［摘 要］ 回归分析中对自变量的要求比较宽松，可以是服从正态分布的随机变量，也可以是分类变量及有序变量，参与回归方程的估计时需首先对分类变量和有序变量赋值。实际应用中，分类变量的赋值存在较多的误用，势必导致错误的分析结果。本文给出了最普遍发生的定性变量被错误赋值的情形，剖析了错误的原因，指出对分析结果的严重歪曲。文中阐述了哑变量设置的具体方法和结果的解释，旨在指导读者采用正确的赋值方法，对分类变量采用多个派生的哑变量参与建模计算，从而得到合理的回归分析结果。

［关键词］ 定性变量；回归分析；哑变量

［中图分类号］ R195.1 ［文献标识码］ C ［文章编号］ 1671-5144(2005)03-0169-03

回归分析是探索因变量与自变量间数量上依存关系的方法，根据资料中各变量的不同特性，常用的方法有多重回归分析（multiple regression analysis）、Logistic 回归分析及 Cox 回归分析等。以多重回归分析为例，要求因变量为服从正态分布的定量变量（quantitative variable），而自变量既可以是定量变量，也可以是定性变量（qualitative variable）。这里，着重讨论回归分析中自变量为定性变量时的情形。通常，定量变量的观察结果可以直接用于回归分析，而定性变量需进行合理的赋值方能用于回归分析。本文将结合实例，说明对定性变量如何进行赋值，以期指导读者正确实施回归分析。

例 1 为了研究空气中一氧化氮（NO）的浓度与汽车流量等因素的关系，有人测定了中国北方某城市交通点在单位时间内过往的汽车数（$X_1$）、气温（$X_2$）、空气湿度（$X_3$）、风速（$X_4$）、季节（$X_5$）以及空气中 NO 的浓度（$Y$），数据如表 1 所示。

表 1 中，$Y$ 及 $X_1$~$X_4$ 均为定量变量，只有 $X_5$ 为定性变量，某研究者分别对春、夏、秋、冬赋值以 1、2、3、4，表 1 中最后一列括号内为所赋的值。

以一氧化氮（$Y$）为因变量，其余 5 个可疑影响因素（$X_1$~$X_5$）作为自变量进行多重回归分析，结果如下（由统计软件 SPSS11.5 完成）。

表 1 空气中 NO 浓度与相关因素的监测数据

| 一氧化氮（$Y$） | 车流（$X_1$） | 气温（$X_2$） | 空气湿度（$X_3$） | 风速（$X_4$） | 季节（$X_5$） |
|---|---|---|---|---|---|
| 0.003 | 985 | 20.0 | 80 | 0.45 | 冬（4） |
| 0.076 | 1 444 | 23.0 | 57 | 0.50 | 春（1） |
| 0.001 | 786 | 26.5 | 64 | 1.50 | 冬（4） |
| 0.170 | 1 652 | 23.0 | 84 | 0.40 | 夏（2） |
| 0.156 | 1 756 | 29.5 | 72 | 0.90 | 秋（3） |
| 0.120 | 1 754 | 30.0 | 76 | 0.80 | 夏（2） |
| 0.030 | 1 200 | 22.5 | 69 | 1.80 | 冬（4） |
| 0.120 | 1 500 | 21.8 | 77 | 0.60 | 秋（3） |
| 0.100 | 1 200 | 27.0 | 58 | 1.70 | 春（1） |
| 0.129 | 1 476 | 27.0 | 65 | 0.65 | 秋（3） |
| 0.135 | 1 820 | 22.0 | 83 | 0.40 | 夏（2） |
| 0.098 | 1 436 | 28.0 | 68 | 2.00 | 春（1） |

表 2 多重回归分析结果之一[a]

| Model | Unstandardized Coefficients | | Standardized Coefficients | $t$ | $P$ |
|---|---|---|---|---|---|
| | $b$ | Std.Error | $b'$ | | |
| 1（Constant） | 0.204 | 0.019 | — | 10.931 | 0.000 |
| $X_5$ | −0.043 | 0.007 | −0.896 | −6.394 | 0.000 |

a Dependent Variable：Y

提示只有季节（$X_5$）对空气中一氧化氮浓度（$Y$）有影响（$P<0.001$）。回归系数为负，而季节春夏秋冬被分别事先赋值以 1~4，故提示随着季节由春→夏→秋→冬，空气中一氧化氮浓度（$Y$）逐渐降低。然而此结论经不起推敲！

由常识不难理解，一年中的四个季节，春秋表现出较多的相似。以中国北方为例，冬季多风，气温低、气压偏高，湿度小，人们的户外活动及道路的车流量减少；夏季多雨，气温高、气压偏低，湿度大，人

［作者简介］ 张晋昕（1966-），男，山西榆次人，副教授，医学博士、博士后，主要研究方向为医学时间序列分析、多变量统计分析方法及其医学应用等。E-mail：zjx@gzsums.edu.cn

们的户外活动少而道路的车流量并无明显减少;春秋两季则均在气候状况上呈现出多变的特点,时而风和日丽,时而阴雨霏霏,日温差大,适宜人们的户外活动,道路的车流量增加。由此,无论从何种角度考虑,把春→夏→秋→冬按从小到大的顺序赋值以1~4,即把季节因素作为有序分类变量,都是不恰当的。理论上不难证明,定性变量在回归分析中未能给以正确的赋值,会导致如本例所示的错误结果。

由前述实例不难看出,有必要从回归分析中澄清变量被错误赋值的实质,避免得出被歪曲的结论。以下从各类型变量的定义入手,阐明其在回归分析中的数量特征,以便正确揭示自变量与因变量间数量上的线性依存关系。

(1)**定性变量的种类**:分为分类变量(categorical variable)和有序变量(ordinal variable)两种,前者又称名义变量(nominative variable)或计数资料(numeration data),后者又称等级资料(ranked data)。

(2)**定性变量的原始记录形式**:定性变量的取值称作"水平",例如患者术后出血量分为少、一般、偏多、过多,此为有序变量,习惯上常把各水平取值记作1、2、3、4,较小的取值表示出血量小,而较大的取值表示出血量多。这里,如果把1理解为"较低水平",代表较少出血量,把4理解为"较高水平",代表较多出血量,完全可以和医学实际意义相吻合。再如,ABO分型系统把个体的血型分作A、B、O、AB共4种情形,除直接以字母记录外,也可以依次记作1、2、3、4,这里1却不能理解为"较低水平",4也不能理解为"较高水平",因该变量属于无序分类的名义变量。

(3)**回归分析中有序变量的赋值**:临床体检或实验室检验常用-、±、+和++等来表示测量结果,属于有序变量。如果有理由认为各水平之间是等距离或近似等距离,各测量结果可依次赋值以1、2、3、4等。有序变量的各水平取值间有时并非等距离,如患者的受教育程度,常用的赋值方法是把文盲、小学、中学、大学及以上取作1、2、3、4,事实上,没有理由不可以使用10、12、18、30表示各水平的文化程度,两种赋值方法得到的分析结果并不等价。正确的做法是,按照各水平间合理的(或易解释的)距离,分别赋以一定的数值,这些数值可以距离不相等。

(4)**回归分析中分类变量的赋值**:有些书籍中把变量分为连续型变量(continuous variable)和离散型变量(discrete variable),再把离散型变量分为有序分类和无序分类两种。本文则采用目前卫生部统编供预防医学专业使用的《卫生统计学》(第5版,方积乾主编)中对变量的定义方法,分类变量的取值是无序的。分类变量的取值1、2、3、4……只是为了数据记录的便利而设定的代码,不能由其平均数作为该分类变量的平均水平对资料进行描述,也不能直接参与回归分析等计算。前述例1中,季节就属于分类变量,对若干次观测中对应的季节变量值求均数,没有任何实际意义;用于回归分析时需对变量进行预处理,正确的做法是由季节($X_5$)派生出哑变量(dummy variable)$X_{51}$、$X_{52}$、$X_{53}$(这时,由3个哑变量表示季节,季节在专业上仍为一个因素,却并不以单个变量的具体取值来刻画),春季($X_{51}$=0、$X_{52}$=0、$X_{53}$=0),夏季($X_{51}$=1、$X_{52}$=0、$X_{53}$=0),秋季($X_{51}$=0、$X_{52}$=1、$X_{53}$=0),冬季($X_{51}$=0、$X_{52}$=0、$X_{53}$=1)。值得注意的是,季节有4个水平,却不可以派生出4个哑变量,春季:$X_{51}$=1(其余3个哑变量取0);夏季:$X_{52}$=1(其余3个哑变量取0);秋季:$X_{53}$=1(其余3个哑变量取0);冬季:$X_{54}$=1(其余3个哑变量取0)。否则会导致变量间的不独立,即恒有$X_{51}+X_{52}+X_{53}+X_{54}=1$,不能满足回归分析对自变量间的独立性要求。

(5)**二分类变量的赋值**:二分类变量常用0和1来编码,属于分类变量的特例,也可以称为0-1变量。医学研究中二分类变量很多见,如性别(男和女)、血清学检验(阳性和阴性)、术后5年追踪结局(生存和死亡)。顺便指出,从数学建模的角度而言,二分类变量的赋值可以任取,例如,下述3种关于性别的赋值方案会得到等价的计算分析结果。

$$x_8=\begin{cases}0 & \text{女性}\\ 1 & \text{男性}\end{cases}\qquad x_8=\begin{cases}1 & \text{女性}\\ 2 & \text{男性}\end{cases}\qquad x_8=\begin{cases}5 & \text{女性}\\ 2 & \text{男性}\end{cases}$$

上述赋值方法若作回归分析,则回归系数的绝对值、假设检验的$P$值将完全相同,不影响性别因素及其他任何自变量的分析结果,只是第3种方案的回归系数与第1、2方案得到的回归系数绝对值相反(因第3个方案的赋值方向与前两个方案相反)。

按照第2部分介绍的做法,由季节($X_5$)派生出哑变量$X_{51}$、$X_{52}$、$X_{53}$,多重回归分析结果如下(表3):

逐步回归分析中,首先进入回归方程的变量为过往的汽车数($X_1$),最终有两个变量进入回归方程($X_1$和$X_{51}$),结论为空气中NO的浓度与过往的汽

表 3 多重回归分析结果之二[a]

| Model | Unstandardized Coefficients | | Standardized Coefficients | $t$ | $P$ |
|---|---|---|---|---|---|
| | $b$ | Std.Error | $b'$ | | |
| 1(Constant) | −0.128 | 0.037 | — | −3.497 | 0.006 |
| $X_1$ | 0.000 | 0.000 | 0.892 | 6.233 | 0.000 |
| 2(Constant) | −0.015 | 0.053 | — | −0.280 | 0.786 |
| $X_1$ | 8.824E-05 | 0.000 | 0.501 | 2.601 | 0.029 |
| $X_{51}$ | −0.061 | 0.024 | −0.488 | −2.534 | 0.032 |

a Dependent Variable：Y

车数($X_1$)及“是否为冬季”有关。$b_1$=8.824×10$^{-5}$ 且为正，说明汽车流量越大，空气中 NO 浓度越高；$b_{51}$=−0.61，由于是负值，说明冬季相对于其它季节空气中 NO 浓度较低，而春、夏、秋 3 个季节的更迭不会造成空气中 NO 浓度有统计学意义的变化。

这里给出的例 1 是以多重回归分析为例的，在 Logistic 回归、Cox 回归中对分类变量的赋值方法相同。本文用回归分析的实际资料来说明问题，是因为回归分析的应用非常广泛，而分类变量被错误赋值的情形屡见不鲜。在其它的多因素分析（相关分析、因子分析、聚类分析、判别分析）方法中，也应参照本文介绍的方法对分类变量赋值。

## [参 考 文 献]


[1] 方积乾. 卫生统计学 [M]. 第 5 版. 北京：人民卫生出版社，2003.

[2] 孙尚拱. 医学多变量统计与统计软件[M]. 北京：北京医科大学出版社，2000.

[3] 黄正南. 医用多因素分析 [M]. 第 3 版. 长沙：湖南科学技术出版社，1995.

[4] Jerrold H. Zar. Biomedical Analysis[M]. 4th ed. Upper Saddle River, New Jersey：Prentice Hall International, Inc.，1999.


[收稿日期] 2005−05−13

(上接第 168 页)


results after resection of thoracic esophageal carcinoma [J]. World J Surg, 1978, 2(4)：543−551.

[3] 张锡珍. 食管癌术后放射治疗[J]. 中华放射肿瘤学杂志，1993，2(4)：264−265.

[4] 朱海文，陈国雄，王迎选，等. 食管癌术后放射治疗 [J]. 中华放射肿瘤学杂志，1998，7(1)：46−48.

[5] 张红梅. 胸段食管癌术后放射治疗的临床观察 [J]. 济宁医学院学报，2002，25(3)：50.

[6] Iizuka T, Ide H, Kakegawa T, et al. Preoperative radioactive therapy for esophageal carcinoma. Randomized evaluation trial in eight institutions [J]. Chest, 1988, 93(5)：1054−1058.

[7] Teninere P, Hay JM, Fingerhut A, et al. Postoperation radiation therapy dose not increase survival after curative resection for squamous as shown by a multicenter controlled trial [J]. Surg Gynecol Obstet, 1991, 173(2)：123−129.

[8] Fok M, Sham JST, Chon D, et al. Postoperative radiotherapy for carcinoma of the esophagus：a prospective randomized controlled study [J]. Surgery, 1993, 113(2)：138−147.

[9] Zieren HU, Muller JM, Jacobi CA, et al. Adjuvant postoperative radiation therapy after curative resection of squamous cell carcinoma of the thoracic esophagus：a prospective randomized study [J]. World J Surg, 1995, 19(3)：444−449.

[10] 梅泽如，项其昌，吴维继，等. 食管癌术后预防性放疗前瞻性研究 [J]. 中华放射肿瘤学杂志，1997，6(3)：188−189.

[11] 肖泽芬. 食管癌[M]//殷蔚伯，谷铣之主编. 肿瘤放射治疗学. 第 3 版. 北京：中国协和医科大学出版社，2002：616−617.

[12] 肖泽芬，杨宗贻，梁 军，等. 食管癌根治术后预防性放射治疗的临床价值 [J]. 中华肿瘤杂志，2002，24(6)：608−611.

[13] Xiao ZF, Yang ZY, Liang J, et al. Value of radiotherapy after radical surgery for esophageal carcinoma：a report of 495 patients [J]. Ann Thorac Surg, 2003, 75(2)：325−328.

[14] 陈长生，徐勇勇. 如何进行 Meta 分析 [J]. 中华预防医学杂志，2003，37(2)：138−140.

[15] Pisani P, Parkin DM, Bray F, et al. Estimates of the worldwide mortality from cancers in 1990 [J]. Int J Cancer, 1999, 83(1)：18−29.

[16] Enzinger PC, Mayer RJ. Esophageal Cancer [J]. N Engl J Med, 2003, 349(23)：2241−2252.

[17] 陈文虎. 食管癌的综合治疗 [J]. 肿瘤，2003，23(2)：85−86.

[18] 汤钊猷. 现代肿瘤学 [M]. 上海：上海医科大学出版社，2000：658−691.

[19] 刘 明，李 任，王敬一，等. 食管癌根治术后预防性照射的研究 [J]. 中华放射肿瘤学杂志，1994，3(4)：268.

[20] Beecher HK. The powerful placebo [J]. JAMA, 1955, 159(17)：1602−1606.

[21] 徐勇勇. 医学研究报告中的统计学新概念 [J]. 中华医学杂志，1995，75(3)：178−182.

[22] Friedenreich CM. Methods for pooled analysis of epidemiological studies [J]. Epidemiol, 1993, 4(4)：295−302.


[收稿日期] 2004−07−28