回归建模的基础与要领(Ⅲ)
——变量状态与相互间关系

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍回归建模的基础与要领之三,即"变量状态与相互间关系"。首先,介绍"因变量状态"与"自变量状态";其次,介绍"自变量间相互关系",即"自变量间相互独立""自变量间有线性关系"和"自变量间有非线性关系";最后,介绍"自变量与因变量间关系",包括"自变量与因变量间无任何数量关系""自变量与因变量间有间接数量关系"和"自变量与因变量间有直接数量关系"。很明显,清楚"变量状态和变量间关系"是构建合理回归模型的重要基础与要领之一。

【关键词】 自变量;因变量;相互独立;多重共线性;非线性关系
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.06.003

The basis and essential of the regression modeling(Ⅲ)
——the variables´status and their relationships

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The purpose of this paper was to introduce the third aspect of the basis and essential of the regression modeling: the variables´status and their relationships. Firstly, "the status of the dependent variable" and "the status of the independent variables" were introduced. Secondly, "the relationships among the independent variables" were listed, such as "the independence among the independent variables", "the linear relationship among the independent variables" and "the nonlinear relationship among the independent variables". Thirdly, "the relationships between the dependent variable and the independent variables" were given as follows: ①There were no any relationship between the dependent variable and the independent variables. ②There were indirect quantity relationship between the dependent variable and the independent variables. ③There were direct quantity relationship between the dependent variable and the independent variables. It was obviously that the variables´status and their relationships understood fully were one of the basis and essential of building a rational regression model.

【**Keywords**】 Independent variable; Dependent variable; Independence each other; Multiple collinearity; Nonlinear relationship

1 概 述

回归分析是研究因变量如何依赖自变量变化而变化的规律的重要统计分析方法之一,然而,回归分析的基本要素涉及两个方面,其一,变量状态及相互间关系;其二,样品(测定变量取值的对象)状态及相互间关系。因篇幅所限,本文仅讨论前述的"第一个要素"。

2 变量状态

2.1 因变量状态

一般来说,可将因变量分为四种状态,即计量的、计数的、有序的(也被称为等级的)和定性的;事实上,在实际应用中,还有一种状态,即"相异性"或"相似性"大小的度量,被称为"非度量型数据"[1]。例如,度量 100 种汽车彼此两两之间的相似程度,可以定义一些"数字"来表示任何两辆汽车之间的相似程度,但它们可能仅代表一种"相似程度"上的"顺序关系",并不代表"数量大小"上的"顺序关系";再比如:现有 50 种不同风味的菜肴,让 10 位鉴赏家品尝,每位鉴赏家给每种菜肴评一个分,这个"分"就被称为"偏好得分"。各鉴赏家所评出的"偏好得分"之间是不可比的。显然,"非度量型变量"不适合用作回归分析中的"因变量",但可用于"非度量型多维尺度分析"[1]或"结合分析"[2]之中。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2 自变量状态

自变量状态也有“计量的、计数的、有序的和定性的”四种，但从回归模型构建与求解的“最初理论和方法”中可隐约体察到：统计学的先驱者们默认自变量都是“计量的”。不知从何时开始，统计学上接受了“定性的自变量”，并将“二值定性自变量”赋予两个不等的数值（通常分别取 0 与 1），而将具有 k 水平的多值名义变量改造成彼此有一定联系的（$k-1$）个“哑变量”（因为它们都以同一个“水平”为基准）。严格地说，这（$k-1$）个哑变量应当同时进入或剔除回归模型，因为每一个哑变量都只利用了全部数据集中一部分“样品或观测”。具体来说，就是基准水平组的样品和其对比组的样品。

3 自变量间相互关系

3.1 自变量间相互独立

经典统计学的回归分析要求：自变量间应相互独立。然而，在解决实际问题时，存在两方面的困难：第一，如何方便快捷地证明给定资料中的自变量间是相互独立的；第二，若基于专业知识和/或统计学知识，得知某些自变量间并非相互独立，如何合理处置？

3.2 自变量间有线性关系

3.2.1 自变量间有线性关系及共线性诊断

如何发现自变量间存在线性关系呢？这在统计学上被称为“共线性诊断”。很多通用统计软件都有这方面的功能，例如：SAS 软件的 REG 过程中，可用“条件数和方差分量”和/或“方差膨胀因子或容许度”[2]来实现共线性诊断。

3.2.2 如何消除共线性的影响

一般来说，当自变量间存在多重共线性时，先通过自变量筛选，可以淘汰出一些自变量，再对保留在回归模型中的全部自变量进行共线性诊断。若此时自变量间仍存在共线性，可采取以下两种方法消除共线性的影响：其一，采用主成分回归分析法，即先对全部自变量进行主成分分析，再以全部主成分变量（它们之间互相独立）为“新自变量”，创建因变量 Y 依赖新自变量的回归模型；其二，直接采用岭回归分析法构建多重线性回归模型。采用前述两种方法对同一个资料构建多重线性回归模型，发现岭回归分析优于主成分回归分析。因为主成分回归分析不能克服某些回归系数的正负号违反专业知识的弊端，而岭回归分析很好地解决了这个问题[3-4]。

3.2.3 自变量间有非线性关系

到目前为止，在进行多重回归分析时，建模者很少考虑“自变量间有非线性关系”的问题。由基本常识可知，既然自变量间有“共线性关系”，那就可能存在“共非线性关系”。只是从统计学角度来看这种情况非常难以驾驭，故迄今为止，似乎尚无现成的统计模型能处理此问题。这也足以说明统计学远未达到尽善尽美的程度。

4 自变量与因变量间的关系

4.1 自变量与因变量间无任何数量关系

在对资料进行回归建模之前，人们赋予资料一个“隐含假定”：自变量与因变量间存在数量联系。至于这种联系的密切程度是很弱、少许、中等、较强还是很强，取决于不同的自变量及因变量在全部观测对象上的取值或表现，需要借助统计学上的假设检验来作出推断。然而，在实际问题中，确有一些自变量与因变量间没有任何关系，此时，经过假设检验或许还能得出：这些自变量对预测因变量的值具有统计学意义！如何才能发现这种“无中生有”的错误结论？

在 SAS/STAT 9.3 中有一个“试验性过程”叫做“ADAPTIVEREG”，它的含义是“适应性回归分析过程”。该过程的“初衷”是能根据自变量与因变量的“数量表现”，灵活且有针对性地度量出各自变量对因变量影响的“重要性”，从而发现那些与因变量无关的“自变量”。然而，令人失望的是：人为设定一些与因变量无关的自变量，采用前述提及的“ADAPTIVEREG”过程建模，仍然找出了几个“重要的自变量”。SAS 程序和计算结果如下：

```
data artificial;
  drop i;
  array X{10};
  do i=1 to 400;
    do j=1 to 10;
      X{j}=ranuni(1);
    end;
    Y=40* exp(8* ((x1-0.5) * *2+(x2-0.5)
* *2)) /
      (exp(8* ((x1-0.2) * *2+(x2-0.7) *
*2)) +
```


?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
        exp(8*((x1-0.7)**2+(x2-0.2)**2)))+rannor(1);
      output;
   end;
run;
proc corr data=artificial;
      var y;
      with x3-x10;
run;
```

```
proc adaptivereg data=artificial;
      model y=x3-x10;
run;
```

【SAS 程序说明】

在 SAS 数据步中，创建了 10 个自变量 $x_1 \sim x_{10}$，将它们放入一个数组“$X\{\ \}$”中，它们的取值为服从均匀分布的“随机数”；创建了一个因变量 y，它是“x_1”与“x_2”的曲线函数，其函数的表达式见下面的式(1)：

$$y=\frac{40\exp(8((x_1-0.5)^2+(x_2-0.5)^2))}{\exp(8((x_1-0.2)^2+(x_2-0.7)^2))+\exp(8((x_1-0.7)^2+(x_2-0.2)^2))} \quad (1)$$

共有 400 个观测值，即样本含量为 400。也就是说，y 仅与“x_1”和“x_2”有曲线关系，而与“$x_3 \sim x_{10}$”无关。

在第 1 个 SAS 过程步中，进行 y 与“$x_3 \sim x_{10}$”之间的 Pearson 相关分析；在第 2 个 SAS 过程步中，由“model 语句”可知，试图创建 y 依赖“$x_3 \sim x_{10}$”的多重线性回归模型。

【SAS 主要输出结果】

Pearson 相关系数，N = 400

Prob > |r| under H0: Rho = 0

x	y
x_3	0.00403
	0.9360
x_4	0.07957
	0.1121
x_5	0.02107
	0.6744
x_6	−0.00101
	0.9839
x_7	−0.01501
	0.7648
x_8	0.06333
	0.2063
x_9	0.02017
	0.6876
x_{10}	−0.03156
	0.5291

“$x_3 \sim x_{10}$”后面均有两行计算结果，上行代表“Pearson 相关系数”、下行代表“对应的 P 值”。以上结果表明，y 与“$x_3 \sim x_{10}$”中的任何一个之间的 Pearson 相关系数都很小，假设检验的结果均无统计学意义，也就是说，y 与“$x_3 \sim x_{10}$”之间的任何一个都是互相独立的。

变量重要性

变量	基数	重要性(%)
x_3	6	100.00
x_4	2	60.87
x_7	2	42.66
x_8	1	16.58

此结果表明：在 8 个与因变量无关的自变量中，找出了 4 个比较重要的自变量，其中，x_3 与 x_4 对因变量 y 影响的重要性分别为 100.00% 与 60.87%。显然，这个结论是错误的！若采用 SAS/STAT 中的“REG 过程”并分别借助逐步法、后退法和前进法“筛选自变量”，其 SAS 过程步程序如下：

```
proc reg data=artificial;
      model y=x3-x10/selection=stepwise sle=0.9
sls=0.05;
run;
proc reg data=artificial;
      model y=x3-x10/selection=backward sls=0.05;
run;
proc reg data=artificial;
      model y=x3-x10/selection=forward sle=0.05;
run;
```

【SAS 输出结果】

上面三个过程步运行的结果相同，均没有一个自变量被保留在回归模型中。这个结果反映了真实的情况。

然而，当人为假定模型中不包含截距项(在前面三个过程步的“model 语句”的“/”之后加上一个选项“NOINT”)时，三个过程步运行的结果相同，其最终结果如下：

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	4	5985.05000	1496.26250	127.41	<0.0001
误差	396	4650.41446	11.74347		
未校正合计	400	10635			
变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
x_4	2.40245	0.51390	256.65707	21.86	<0.0001
x_5	1.77945	0.50770	144.26047	12.28	0.0005
x_8	1.84510	0.54490	134.65135	11.47	0.0008
x_9	1.33568	0.52141	77.06289	6.56	0.0108

据此，可写出 4 重线性回归模型如下：

$\hat{y} = 2.40245x_4 + 1.77945x_5 + 1.84510x_8 + 1.33568x_9$

该 4 重线性回归模型的“$R^2 = 0.5627$”，模型的假设检验结果为：$F = 127.41$、$P < 0.0001$，说明此模型具有统计学意义。

显然，这个结果在统计学上是“相当好的”；然而，它确实严重违背了真实情况！

由此可知：当研究者对所研究变量之间的“真实情况”一无所知时，必须依据“基本常识”和“专业知识”作出有一定依据的“假定”，运用统计学的各种技术方法构建多重回归模型，再回到实践中去检验回归模型的实用价值。

4.2 自变量与因变量间有间接数量关系

在实际问题中，自变量与因变量间有间接数量关系的情形是最常见的。例如：若以正常成年人“心像面积”为因变量，以其“身高、体重、体重指数、胸围”为自变量，则后者对前者的影响是“间接的”，而且具有一定的“数量关系”。再例如：若以正常成年人“身体健康指数（假定其存在）”为因变量，以其“血糖生化指标（如空腹血糖、餐后 2 小时血糖、空腹胰岛素、餐后 2 小时胰岛素、糖化血红蛋白、胰岛素抵抗指数、胰岛素敏感指数等）”“血脂生化指标（甘油三脂、总胆固醇、低密度脂蛋白胆固醇、高密度脂蛋白胆固醇、载脂蛋白 α、载脂蛋白 β 等）”“肝功能指标（门冬氨酸氨基转移酶、谷丙冬氨酸氨基转移酶、谷草/谷丙、γ - 谷氨酰转肽酶、血清总蛋白、白蛋白、球蛋白、白球比、总胆红素、直接胆红素、间接胆红素等）”“肾功能指标（肌酐、尿素氮、尿酸等）”“炎症因子指标（TNF - α、IL - 6、C 反应蛋白、MCP - 1 等）”“脂肪因子指标（瘦素、脂联素、游离脂肪酸等）”“内毒素”“肠泌肽指标（胰高血糖素样肽 - 1 和葡萄糖依赖性促胰岛素多肽）”“代谢组学检测指标（胰高血糖素样肽 - 1、YY 肽等）”“DNA 甲基化检测指标”和“各种基因检测指标”为自变量，则后者对前者的影响是“间接的”，而且具有一定的“数量关系”。

类似上面的例子，在人体身心、自然界、人与自然之间，只要找出“因变量”，就有大量的“自变量”与其有间接的数量关系。

4.3 自变量与因变量间有直接数量关系

在现实问题中，自变量与因变量间有直接数量关系的情况相对较少。一个最常见的例子如下：若以“药物种类”“剂量大小”“作用时间”和“给药途径”等作为自变量，而以“生物体作出的反应”为“因变量”，则自变量与因变量间存在直接数量关系；再比如，在农业试验研究中，若以“作物品种”“耕种方式”“土壤成分”“灌溉方式”“降雨量多少”等作为“自变量”，以“作物产量或品质”作为因变量，则自变量与因变量间也有直接数量关系。

5 讨论与小结

在研究因变量是否依赖多个自变量变化而变化的规律时，统计学教科书上通常都“理直气壮”地引导使用者直接构建“多重线性回归模型”。由基本常识和专业知识可知，在实际问题中，可能某些自变量完全独立于因变量，也可能某些自变量与因变量间存在着某种复杂的“曲线关系”，更多情况下，人们遗漏了很多“间接或直接”影响因变量的自变量（这正是很多试验设计质量不高的科研项目存在的“严重瑕疵”）。所以，人们最习惯使用的“多重线性回归分析方法”，只是对变量间关系的一种“理想化、简单化”处理方法，其结果“仅供参考”。

比较稳妥的做法是：第一，要力争科研设计无懈

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

可击(至少要做到:对因变量可能有影响的自变量不会被遗漏);第二,有标准操作规程并按其实施科学研究;第三,有实时精准的质量控制策略并得到严格落实;第四,有经得起推敲且系统全面的“统计分析计划”,单从“统计建模”方面来说,应先对资料进行“探索性分析”,以便对某些变量采取合适的变量变换、引入必要的“派生变量”[3-4]、采取多种可能的“统计模型”拟合资料,从构建的多个高质量回归模型中,优中选优;然后,将足够大样本量的“测试数据集(未参与回归建模计算)”带入求得的“最优”回归模型,考察其“精准程度”。仅当“精准程度”达到专业要求时,才可以使用已构建的回归模型去解决所研究的实际问题。

参考文献

[1] 胡良平. 面向问题的统计学——(3)试验设计与多元统计分析[M]. 北京:人民卫生出版社,2012:318-332.

[2] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京:人民卫生出版社,2012:215-228,527-540.

[3] 胡良平. 主成分分析应用(Ⅰ)——主成分回归分析[J]. 四川精神卫生,2018,31(2):128-132.

[4] 胡良平. 岭回归分析[J]. 四川精神卫生,2018,31(3):193-196.

(收稿日期:2018-12-12)

(本文编辑:唐雪莉)

(上接第 497 页)

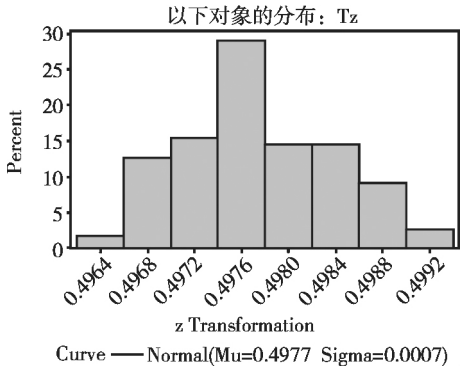

图 6 经 Box-Cox 变换后的某地 110 名健康男性体重(kg)的频数分布直方图

参考文献

[1] 胡良平. 面向问题的统计学——(1)科研设计与统计基础[M]. 北京:人民卫生出版社,2012:258-311.

[2] 茆诗松. 统计手册[M]. 北京:科学出版社,2006:59,111.

[3] 胡良平. 计数资料回归分析基础知识[J]. 四川精神卫生,2018,31(5):385-393.

[4] SAS Institute Inc. STAT SAS 9.3 User s Guide[M]. Cary,NC:SAS Institute Inc,2011:7761-8002.

(收稿日期:2018-12-12)

(本文编辑:唐雪莉)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net