基于经典统计思想实现多重线性回归分析

谷恒明[1],胡良平[1,2*]

(1.军事医学科学院生物医学统计学咨询中心,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】 本文目的是介绍基于经典统计思想实现多重线性回归分析的方法。首先,概述基于经典统计思想、贝叶斯统计思想和机器学习统计思想建立多重线性回归模型的基本思路;然后以实际问题为例,全面呈现了多重线性回归分析所需要完成的主要任务;最后,总结多重线性回归分析的适用场合及注意事项。结果表明:产生派生变量、进行自变量筛选和共线性诊断、进行异常点诊断等内容是进行多重线性回归分析的主要任务。在多因素试验或观察性研究中,只要结果变量为计量变量,比较常用且有效的做法是进行多重线性回归分析,应尽可能少用单因素差异性分析。

【关键词】 经典统计思想;贝叶斯统计思想;机器学习统计思想;多重线性回归分析;派生变量;自变量筛选;多重共线性诊断;异常点诊断

中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.01.002

Realization of a multiple linear regression analysis based on the classical statistical thought

Gu Hengming[1], *Hu Liangping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, 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 aim of this paper was to realize a multiple linear regression analysis based on the classical statistical method. First of all, given an introduction of the basic idea about building a multiple linear regression model based on the classical statistical thinking, Bayesian statistical thinking and machine learning statistical thinking. Then, adopted a practical problem for instance to comprehensively presented the main tasks performed by a multiple linear regression analysis. Finally, summarized the rational and precautions of the multiple linear regression analysis. The results shown that the main tasks of a multiple linear regression analysis were as follows: to produce the derived variables, to screen and select the variables and col-linearity diagnosis, to diagnose the outliers and so on. The conclusion is that in a multi-factor trial or observational study, as long as the outcome variable is a measurement variable, it is more common and effective to perform a multiple linear regression analysis rather than using the uni-variate difference analyses.

【**Keywords**】 Classical statistical thinking; Bayesian statistical thinking; Machine learning statistical thinking; Multiple linear regression analysis; Derived variables; Independent variable screening; Multiple collinearity diagnosis; Outlier diagnosis

1 基于三种统计思想建立多重线性回归模型的概述[1-3]

1.1 经典多重线性回归分析建模概述

多重线性回归分析是用回归方程定量地刻画一个因变量与多个自变量之间的线性依存关系。其中,因变量是连续型变量,自变量是相互独立的连续型变量(也常包括少量分类变量)。

经典多重线性回归分析的内容包括对自变量的筛选和回归诊断(含多重共线性诊断和异常点诊断)、对回归模型和模型中全部参数的假设检验、对模型拟合效果的评价以及利用求得的回归模型对因变量进行预测。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

对自变量的筛选有八种方法:①前进法;②后退法;③逐步法;④最大 R^2 增量法;⑤最小 R^2 增量法;⑥R^2 选择法;⑦修正 R^2 选择法;⑧Mallow's Cp 统计量选择法。

当自变量之间存在较多的、严重的多重共线性关系时,通常可采用岭回归分析或者主成分回归分析。

设用 Y 代表因变量,X_1、X_2、……、X_m 分别代表 m 个自变量,则多重线性回归模型可表示为:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+\varepsilon \quad (1)$$

式中 β_0 为总体截距,β_1、β_2、…、β_m 分别为各个自变量所对应的总体偏回归系数,ε 为随机误差。偏回归系数 $\beta_i(i=1,2,\cdots,m)$ 表示在其他自变量固定不变的条件下,X_i 每改变一个测量单位时所引起的因变量 Y 的平均改变量。多重线性回归模型的

样本回归方程可以表示为：

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m \quad (2)$$

这里 $\hat{Y}$ 表示 Y 的估计值，b_0、b_1、b_2、……、b_m 为截距和偏回归系数的样本估计值。

1.2 贝叶斯回归分析建模概述

$$Y_i = \mu_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n,$$
$$\mu_i = \beta_0 + \beta_i X_i, \quad (3)$$

这里要为各个参数指定一个先验分布，例如：

$$\pi(\beta_0) = \phi(0, \text{var} = le6)$$
$$\pi(\beta_1) = \phi(0, \text{var} = le6)$$
$$\pi(\sigma^2) = f_{iT}(shape = 3/10, scale = 10/3) \quad (4)$$

经典多重线性回归分析假定自变量前的回归系数是固定的，而贝叶斯回归分析认为参数是随机的。基于贝叶斯统计思想建立回归模型时，要为各个自变量前的参数（即回归系数）和残差指定一个先验分布。可以依靠经验或预试验的结果指定各自合适的先验分布；如果没有办法给定先验分布，可以使用无信息先验（相当于均匀分布）代替。贝叶斯回归分析中没有自变量筛选功能，因此要借助经典多重线性回归分析中筛选方法筛选出来的自变量来建立回归模型。

贝叶斯统计建模可以参考 SAS 软件的 STAT 模块中 MCMC 过程来实现。MCMC 方法即马氏链蒙特卡罗方法，默认算法是使用正态分布随机游动 Metropolis 算法。MCMC 的抽样方法有 Gibbs 抽样、Metropolis 抽样、独立性抽样、随机游动 Metropolis 抽样等[2]。

1.3 机器学习回归分析建模概述

有别于经典统计思想和贝叶斯统计思想，机器学习统计思想则另辟蹊径，它不依赖于概率分布知识、也不依赖于先验分布知识，而是通过基于训练样本的学习获取知识和经验，再用测试样本来验证。属于机器学习的具体方法很多，通常包括决策树法、支持向量机法、神经网络法、随机森林法和集成学习法等[3]。

1.4 三类方法回归建模效果

1.4.1 三类回归建模效果的评价

情形一，样本量较少时：分别使用两种方法建立回归模型，用相对误差绝对值的均值（abserror），残差平方和（SSress）与决定系数（R^2）作为评价指标。

情形二，样本量较多时：①全部数据用来建立模型并比较，评价指标同样本量较少时。②K－Fold 交叉验证，即全部数据拆分为 K 份，其中（K－1）份用作建立模型的训练集，剩下一份当做测试集。训练集拟合效果使用相对误差绝对值的均值（abserror），残差均方（MSE）与决定系数（R^2）作为评价指标；测试集使用相对误差绝对值的均值（abserror），残差均方（MSE）与标准化均方误差（NMSE）作为评价指标。③K－Fold 交叉验证中，K 取值分别为 10、7、4 和 2。当 K 取定一个数值后，分别重复抽取 10 次，即进行 10 次重复建模。

1.4.2 评价指标的具体公式

$$\text{标准化均方误差 NMSE} = \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}, \quad (5)$$

在数值上，NMSE 等于 $1 - R^2$，这里的 R^2 是回归的决定系数，但是，对于测试集来说，其 NMSE 与测试集回归的 R^2 没有什么关系。交叉验证主要关心测试集的 NMSE。

$$\text{残差均方 MSE} = \frac{\sum (y - \hat{y})^2}{N - n - 1}, \quad (6)$$

N 为样本量，n 为自变量个数。从上述的公式中可以看出，残差均方 MSE 是残差平方和与自由度的比值。交叉验证中 K 取值不同，建立模型的训练集和预测使用的测试集样本量是不同的，直接基于残差平方和比较不够合理，因此，需要除以自由度。

说明：BP 神经网络建模效果的评价指标采用公式（5）。

2 实例及基于经典统计思想的回归分析

2.1 问题与数据

【例 1】26 例糖尿病患者的血清总胆固醇（X_1）、甘油三酯（X_2）、空腹胰岛素（X_3）、糖化血红蛋白（X_4）、空腹血糖（Y）的测量值列于表 1，试基于经典统计思想建立血糖与其他几项指标间的多重线性回归方程，并完成其他有关的任务。

2.2 回归分析任务

对例 1 表 1 中的数据，设因变量为 Y，自变量为 X_1、X_2、X_3、X_4，试建立因变量依赖自变量的多重线性回归模型，并做相应的假设检验。

表 1 26 例糖尿病患者血样中有关指标的测定结果

i	X_1	X_2	X_3	X_4	Y
1	5.68	1.9	4.53	8.2	11.2
2	3.97	1.64	7.32	6.9	8.8
…	…	…	…	…	…
25	11.54	10.89	1.2	10.5	20
26	3.84	1.2	6.54	9.6	10.4

注:详细数据见本期第一篇文章《多重线性回归分析的核心内容与关键技术概述》

2.3 采用经典统计思想实现多重线性回归分析的方法

(1)不产生派生变量并采用三种筛选自变量的方法建模

```
data cra1;
input id x1 - x4 y @@;
cards;
……
;
run;
proc reg data = cra1;
  model y = x1 - x4/selection = stepwise sle = 0.5 sls = 0.05;
run;
proc reg data = cra1;
  model y = x1 - x4/selection = forward sle = 0.05;
run;
proc reg data = cra1;
  model y = x1 - x4/selection = backward sls = 0.05;
run;
```

【程序说明】“cards”语句后的省略号代表表 1 中 26 行 6 列数据,其中,第 1 列为“编号”;REG 过程被调用了 3 次,分别采用逐步法、前进法和后退法筛选自变量;“sle = 0.5”代表选变量进入回归模型的显著性水平,其概率值选用 0.5 是非常大的,以便有较多的变量有机会进入回归模型与其他变量进行组合,可以较好地保证单个作用不大但与某些自变量同时存在时作用明显增大的自变量不会被排斥在回归模型之外,这叫做“宽进”;“sls = 0.05”代表已进入回归模型的自变量仍能被保留在回归模型之中的显著性水平,其概率值选用 0.05 是统计学上被公认的显著性水平,这叫做“严出”。

【主要输出结果】本例资料采用上述三种筛选自变量方法所得结果完全相同,其主要结果如下:

方差分析

来源	自由度	平方和	均方	F 值	Pr > F
模型	3	156.15002	52.05001	17.77	<.0001
误差	22	64.44113	2.92914		
校正合计	25	220.59115			

变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
Intercept	4.91480	2.14919	15.31800	5.23	0.0322
x2	0.43796	0.13425	31.17168	10.64	0.0036
x3	-0.29949	0.09699	27.92711	9.53	0.0054
x4	0.81267	0.20624	45.47918	15.53	0.0007

【输出结果解释】第 1 部分表明:总的回归模型具有统计学意义($F = 17.77$、$P < 0.0001$);第 2 部分表明:自变量 X_1 被淘汰掉了,其他 3 个自变量以及截距项均有统计学意义,得到的多重线性回归方程为:

$\hat{Y} = 4.9148 + 0.43796X_2 - 0.29949X_3 + 0.81267X_4$

(2)不产生派生变量并采用逐步法筛选自变量且进行共线性诊断和残差分析等方法建模

上面的 SAS 程序中的数据步程序不变,删除第 2 个和第 3 个过程步程序,第 1 个过程步程序修改如下:

```
proc reg data = cra1;
  model y = x1 - x4/selection = stepwise sle = 0.5
        sls = 0.05 collin collinoint vif tol r stb;
run;
```

【程序说明】“collin”选项是要求系统给出采用“方差比例”算法且未校正回归模型中截距项影响的多重共线性诊断的结果;“collinoint”与前面选项的区别在于“校正了回归模型中截距项影响”;“vif”选项是要求系统给出采用“方差膨胀因子”算法的多重共线性诊断的结果、“tol”等于“1/vif”,即要求系统给出采用“容许度”算法的多重共线性诊断的结果;“r”要求系统给出“残差分析”的计算结果,有助于发现是否存在异常点;“stb”要求系统给出“标准化回归系数”的计算结果。

【主要输出结果及其解释】逐步回归分析的主要结果同上,此处从略。基于 4 种方法进行共线性诊断的结果如下:

参数估计值

变量	自由度	参数估计值	标准误差	t 值	Pr > |t|	标准化估计值	容差	方差膨胀
Intercept	1	4.91480	2.14919	2.29	0.0322	0	.	0
x2	1	0.43796	0.13425	3.26	0.0036	0.38281	0.96430	1.03703
x3	1	-0.29949	0.09699	-3.09	0.0054	-0.37405	0.90487	1.10513
x4	1	0.81267	0.20624	3.94	0.0007	0.48561	0.87426	1.14382

以上为第 1 部分输出结果:倒数第 3 列为“标准化回归系数”,其绝对值越大,表明所对应的自变量对因变量的贡献就越大,由大到小依次为 $X_4 > X_2 > X_3$;倒数第 2 列和第 1 列分别为“容许度”与“方差膨胀因子”方法诊断共线性的结果,只需要看 vif 的数值是否大于 10,大于 10 的那些行上的自变量间存在较严重的共线性。结果表明:3 个自变量间不存在共线性关系。

共线性诊断

个数	特征值	条件指数	偏差比例			
			Intercept	x2	x3	x4
1	3.41419	1.00000	0.00200	0.02574	0.01657	0.00241
2	0.37722	3.00849	0.00185	0.77291	0.16913	0.00055863
3	0.19485	4.18592	0.01758	0.19412	0.59862	0.04365
4	0.01375	15.75998	0.97858	0.00723	0.21568	0.95338

以上为第 2 部分输出结果:这是未对截距项进行校正且依据“方差比例”算法进行共线性诊断的结果。适用于“回归分析模型中截距项无统计学意义”的场合,而本例截距项有统计学意义,故不需要看这部分输出结果。

共线性诊断(截距已调整)

个数	特征值	条件指数	偏差比例		
			x2	x3	x4
1	1.36657	1.00000	0.10439	0.24136	0.31191
2	0.98261	1.17930	0.71961	0.24501	0.00058713
3	0.65082	1.44906	0.17600	0.51362	0.68750

以上为第 3 部分输出结果:这是对截距项进行校正后且依据“方差比例”算法进行共线性诊断的结果。适用于“回归分析模型中截距项有统计学意义”的场合,而本例,截距项有统计学意义,故应该看这部分输出结果。评判是否存在共线性的方法:看 3 个自变量列输出结果的最后一行,当这些数值中有两个或多个数值都很大且接近于 1,那么,它们对应的自变量间存在共线性。本例 3 个自变量间不存在共线性关系。

残差分析的输出结果很多,此处从略。从学生化残差结果来看,没有取值的绝对值大于 2 的观测点;从“Cook´s D”统计量计算结果来看,仅第 25 个观测点的取值为 0.698 大于 0.5,表明此观测点是“可疑的异常点”。

(3)产生派生变量并采用三种筛选自变量的方法建模

所谓产生派生变量,就是除资料中已有的 4 个自变量外,再通过变量变换的方法,引入“新变量”。通常可以引入自变量的二次项,包括各自变量的平方项和任何两个自变量的交叉乘积项。在上面第 1 段 SAS 程序基础上,进行如下修改即可:

```
data cra2;
  set cra1;
z1 = x1 * x1;z2 = x2 * x2;z3 = x3 * x3;z4 = x4 * x4;
z5 = x1 * x2;z6 = x1 * x3;z7 = x1 * x4;z8 = x2 * x3;
z9 = x2 * x4;z10 = x3 * x4;
run;
proc reg data = cra2;
  model y = x1 - x4 z1 - z10/selection = stepwise sle
=0.5sls = 0.05;
run;
proc reg data = cra2;
  model y = x1 - x4 z1 - z10/selection = forward sle =
0.05;
run;
proc reg data = cra2;
  model y = x1 - x4 z1 - z10/selection = backward sls
=0.05;
run;
```

【程序说明】在已经运行原先 SAS 程序的基础

上(即已经创建了 SAS 数据集 cra1),再创建新数据集 cra2,这就是前两个语句的作用。新数据集中增加了 z1 - z10 共 10 个新变量,它们是由原先的 4 个自变量产生的派生变量,代表 10 个二次项;3 个过程步分别采用"逐步法""前进法"和"后退法"筛选自变量,现在的自变量个数为 14 个。

以下是"逐步法"和"前进法"筛选自变量的结果:

方差分析

来源	自由度	平方和	均方	F 值	Pr > F
模型	3	163.18701	54.39567	20.85	<0.0001
误差	22	57.40414	2.60928		
校正合计	25	220.59115			

变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
Intercept	6.63383	1.40149	58.46176	22.41	0.0001
x3	0.72844	0.28459	17.09486	6.55	0.0179
z6	-0.18508	0.05142	33.79889	12.95	0.0016
z7	0.12696	0.02044	100.66247	38.58	<0.0001

以下是"后退法"筛选自变量的结果:

方差分析

来源	自由度	平方和	均方	F 值	Pr > F
模型	3	172.96983	57.65661	26.64	<0.0001
误差	22	47.62132	2.16461		
校正合计	25	220.59115			

变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
Intercept	8.66929	1.02421	155.08554	71.65	<0.0001
z4	0.04547	0.00874	58.55106	27.05	<0.0001
z5	0.05179	0.01170	42.38276	19.58	0.0002
z6	-0.05480	0.01567	26.46656	12.23	0.0020

考察以上两个不同的建模结果可以发现:后退法的建模结果稍好一些,因为其总模型的假设检验结果的 $F=26.64$ 较大(较小者为 $F=20.85$)且模型中所含的项数相同(均为 4 项),总模型和各项假设检验结果均具有统计学意义。

结合前面未引入派生变量时得到的多重线性回归模型,其总模型的 $F=17.77$,小于现在获得的最好的模型对应的 $F=26.64$,故拟合本例资料最好的多重线性回归方程如下:

$$\hat{Y}=8.66929+0.05179X_1\times X_2-0.05480X_1\times X_3+0.45479(X_4)^2$$

(4)产生派生变量并采用后退法筛选自变量且进行共线性诊断和残差分析等方法建模

在运行了前面第 1 个数据步程序(创建数据集 cra1)和第 2 个数据步程序(创建数据集 cra2)的基础上,再运行下面的过程步程序:

```
proc reg data = cra2;
  model y = X1 - X4 Z1 - Z10/selection = backward
        sls = 0.05 collin collinoint vif tol r stb;
run;
```

输出结果较多,此处从略。结果表明:Z_4、Z_5、Z_6 三个项之间不存在共线性关系;残差分析的结果与前面陈述的结果基本相同,但第 25 个观测点对应的 Cook´s D = 0.441 < 0.5,说明它已经不属于"可疑异常点"了。

3 小 结

本文介绍了基于经典统计思想构建多重线性回归模型的主要内容,此法可用于很多多因素临床试验研究和观察性研究中,其主要的应用条件是结果变量为"计量变量",例如主要评价指标为"评分"。值得注意的是:许多实际工作者在对此类资料进行统计分析时,喜欢选用单因素差异性分析,见文献[4-7]。其实,选用多重线性回归分析方法来实现多因素分析,效果更佳。

虽然,在本文中所介绍的实例中,自变量都是计量变量,而在实际使用中,自变量可以是计量的、计数的、定性的。但是,当自变量为多值名义变量时,需要产生哑变量后才能引入多重线性回归模型之中[8],否则,可能得出错误的结论。

参考文献

[1] Kleinbaum DG, Kupper LL, Muller KE, et al. Applied regression analysis and other multivariable methods[M]. 3 版. 北京: 机械工业出版社, 2003: 111-159.

[2] 刘金山, 夏强. 基于 MCMC 算法的贝叶斯统计方法[M]. 北京: 科学出版社, 2017: 118-174.

[3] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 3 版. 北京: 中国人民大学出版社, 2015: 41-56.

[4] 任传波, 姜季妍, 董黎明. 青少年抑郁障碍患者心理社会学特征[J]. 四川精神卫生, 2017, 30(5): 455-457.

[5] 徐华丽, 孙崇勇, 高悦. 大学生人格特质对手机成瘾倾向的影响[J]. 四川精神卫生, 2017, 30(5): 458-462.

[6] 李青青, 张倩. 医学院校大学生情绪状态与学业拖延的关系[J]. 四川精神卫生, 2017, 30(5): 463-465.

[7] 魏国英, 曾丽娟, 周桂成, 等. 精神科护士职业倦怠与工作压力的相关性[J]. 四川精神卫生, 2017, 30(5): 466-469.

[8] 胡良平. 医学统计学——运用三型理论进行现代回归分析[M]. 北京: 人民军医出版社, 2010: 9-33.

(收稿日期:2018-01-29)

(本文编辑:陈 霞)