·科研方法专题·

多重线性回归分析的核心内容与关键技术概述

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】 **目的** 本文目的是概述多重线性回归分析的核心内容与关键技术。其核心内容有以下四点:第一,构建多重线性回归模型的方法和求解参数的方法;第二,进行回归诊断的意义和方法;第三,筛选自变量的意义和方法;第四,评价模型拟合效果的方法。其关键技术是如何基于经典统计思想、贝叶斯统计思想和机器学习统计思想实现多重线性回归分析。

【关键词】 多重线性回归模型;回归诊断;共线性;异常点;均方误差;贝叶斯统计;机器学习
中图分类号:R195.1 **文献标识码**:A doi:10.11886/j.issn.1007-3256.2018.01.001

Overview of the core concepts and key techniques in the multiple linear regression analysis

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 paper aims at summarizing the core concepts and key techniques in the multiple linear regression analysis. The core concepts include the follow four points: ①The methods of constructing a multiple linear regression model and finding the solution of the parameters in the model. ②The meanings of implementing regression diagnosis and its methods. ③The meanings and methods of screening the independent variables. ④The methods of appraising the fitting effect of the regression models. The key techniques of the multiple linear regression analysis are involved in three kinds of statistical thoughts which are classical statistical thought, Bayesian statistical thought and the statistical thought of the machine learning.

【**Keywords**】 Multiple linear regression model; Regression diagnosis; Multiple collinearity; Outlier; Mean squared error; Bayesian statistics; Machine learning

1 与多重线性回归分析有关的基本概念

1.1 何为多重线性回归分析

在生物医药学和临床研究中的很多场合下,需要考察因变量如何依赖多因素变化而变化的规律,例如文献[1-2]都涉及到这种需求。此时,所需要的统计分析方法泛称为“多重回归分析”,当因变量为“计量变量”时,常被称为“多重线性回归分析”。

多重线性回归分析就是要求算出相应模型中参数的估计值,对回归方程和参数进行假设检验;在这个过程中,还需要对自变量进行筛选和共线性诊断、对观测点进行异常点诊断,基于构建的多重线性回归模型并在给定的自变量取值条件下对因变量的取值进行预测。多重线性回归模型(对总体而言)与回归方程(对样本而言)的表达式分别见式(1)和式(2)。

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以表示为:

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

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,常假定其服从正态分布。偏回归系数 β_i($i=1,2,\cdots,m$)表示在其他自变量固定不变的情况下,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值,它们是相应总体参数的估计值。

在建立了回归方程以后,就可以对因变量与自变量之间的线性依存关系进行定量分析,进而还可以利用回归方程对因变量进行预测。

1.2 如何在进行多重线性回归分析前估计样本含量

在进行试验研究或调查研究之前,一般都需要估计所需的最低样本含量。在进行多重线性回归分

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

析前,同样也需要估计样本含量。具体的方法可分两步走[3]:

第一步,估计简单直线回归或相关分析时所需要的样本含量。

设因变量为 Y,自变量为 X_1;再设它们之间的简单相关系数 $\rho(\approx r)$,其中,ρ 为两变量之间的总体相关系数,而 r 为两变量之间的样本相关系数。此时,所需要的样本含量 n_1 见式(3):

$$n_1 = 4\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\ln\left(\frac{1+\rho}{1-\rho}\right)}\right] + 3 \qquad (3)$$

在式(3)中,α、β 都表示标准正态分布曲线下两尾端面积之和;$Z_{\alpha/2} = Z_{0.05/2} = 1.96$ 为标准正态分布曲线下两尾端面积之和为 0.05 时对应的横坐标(也被称为临界值),同理,可理解 $Z_{\beta} = Z_{0.1} = 1.28$ 的含义;当总体相关系数 ρ 未知时,一般用样本相关系数 r 取代。

若在简单直线回归模型中新增加($m-1$)个自变量,则为了检验 X_1 的回归系数是否为 0,此时,样本含量 n_m 的计算公式见式(4):

$$n_m = \frac{n_1}{1 - \rho^2_{1|2,3,\cdots,m}} = n_1 VIP_m \qquad (4)$$

在式(4)中,$\rho^2_{1|2,3,\cdots,m}$ 为复相关系数的平方,也称为决定系数,它是以 X_1 为因变量、以其他($m-1$)个自变量为自变量构建($m-1$)重线性回归模型时所算得的决定系数;VIP_m 为方差膨胀因子,其计算公式见式(5):

$$VIP_m = \frac{1}{1 - \rho^2_{1|2,3,\cdots,m}} \qquad (5)$$

值得一提的是:试验或调查之前,如何才能获得公式(3)和公式(4)中的“ρ”和“$\rho^2_{1|2,3,\cdots,m}$”呢?需要依据研究者本人预试验的结果或文献资料中提供同类研究的结果进行估算。

1.3 适合进行多重线性回归分析的数据结构

问题与数据结构见例 1 和表 1。

【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,建立血糖依赖其他几项指标变化的多重线性回归模型(说明:本例的研究者并没有事先估计所需要的样本含量)。

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
3	6.02	3.56	6.95	10.8	12.3
4	4.58	1.07	5.88	8.3	11.6
5	4.6	2.32	4.05	7.5	13.4
6	6.05	0.64	1.42	13.6	18.3
7	4.9	8.5	12.6	8.5	11.1
8	7.08	3	6.75	11.5	12.1
9	3.85	2.11	16.28	7.9	9.6
10	4.65	0.63	6.59	7.1	8.4
11	4.59	1.97	3.61	8.7	9.3
12	4.29	1.97	6.61	7.8	10.6
13	7.97	1.93	7.57	9.9	8.4
14	6.19	1.18	1.42	6.9	9.6
15	6.13	2.06	10.35	10.5	10.9
16	5.71	1.78	8.53	8	10.1
17	6.4	2.4	4.53	10.3	14.8
18	6.06	3.67	12.79	7.1	9.1
19	5.09	1.03	2.53	8.9	10.8
20	6.13	1.71	5.28	9.9	10.2
21	5.78	3.36	2.96	8	13.6
22	5.43	1.13	4.31	11.3	14.9
23	6.5	6.21	3.47	12.3	16
24	7.98	7.92	3.37	9.8	13.2
25	11.54	10.89	1.2	10.5	20
26	3.84	1.2	6.54	9.6	10.4

【对数据结构的解说】在表 1 中,研究者从 26 例糖尿病患者体内抽血,对每份血样进行检查,测定了“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)”5 个定量变量的取值。它们之间的数量关系是客观存在的,而且是并存关系,不存在谁是自变量、谁是因变量。对资料进行统计分析时,可以有多种不同的统计分析目的,仅当希望分析“空腹血糖(Y)”是否随着“X_1、X_2、X_3、X_4”变化而变化时,才人为地将前者视为“因变量”、将后者视为“自变量”。

1.4 多重线性回归分析要求资料应满足的前提条件

结合上面的表 1,呈现多重线性回归分析要求资料应满足的前提条件如下:

第一,受试对象具有同质性。受试对象为 26 例糖尿病患者,而不是患有其他疾病的患者,也不是正常人。这里还有一些隐含的前提条件未明确交代:所患糖尿病的类型、严重程度、病程、是否有糖尿病家族史等基本相同,年龄、性别、职业、身体素质、生活和饮食习惯、锻炼方式和程度等基本相同。

第二,全部变量、特别是因变量是计量变量。经典统计学中,要求进行多重线性回归分析的全部变量都是定量的;在具体使用时,条件有所放松,即自变量中允许有二值定性变量或多值有序变量,若有多值名义变量,需要将其变换为二值的“哑变量(其个数为原多值名义变量的水平数减一)”。

第三,各自变量组合条件下的计量因变量应服从正态分布且方差相等。这一条仅仅局限于“统计理论”上的规定或假定,在实际使用时是无法考证的。因为每个定量自变量的取值都有无穷多个,全部自变量的不同水平组合自然也就有无穷多种了。而在实际研究中,样本含量十分有限,几乎没有两个受试对象在全部自变量上的取值是完全相同的,那么,怎样考察第三个前提条件是否成立呢? 故在实际使用中,只能大大降低要求:只要因变量的全部取值近似服从正态分布即可。

第四,因变量与全部自变量间存在线性关系,而非曲线关系。这一点可以通过绘制因变量与每一个自变量之间的散布图来初步了解。

第五,全部自变量间互相独立,不存在共线性关系。当所拟合的多重线性回归模型中确实存在多重共线性问题时,则此多重线性回归模型的质量就不高。所以,应先进行共线性诊断,当发现存在多重共线性时,要设法消除,然后再构建多重线性回归模型。

1.5 如何进行回归诊断

1.5.1 何为自变量间的多重共线性

所谓多重共线性,就是某几个自变量之间存在线性关系,例如:$X_1 = 0.35X_2 + 1.28X_3 - 21.45X_4$。

1.5.2 如何诊断自变量之间是否存在多重共线性

诊断是否存在多重共线性的方法有多种,如方差分量法、方差膨胀因子法(等价于容许度法)等,计算公式较复杂,此处从略;若借助 SAS 软件中的 REG 过程来拟合多重线性回归模型,只需在model 语句中增加选择项“COLLIN”和“COLLINOINT”或“VIF”和“TOL”即可实现。

1.5.3 何为异常点以及如何诊断

若在回归分析的资料中,所有的观测点(表 1 中就有 26 个观测点)都是“同质的”,则不存在“异常点”;若有少数个体(即观测点)与其他绝大多数个体不同质,则它们就有可能成为异常点。为了便于直观理解异常点,参见图 1。

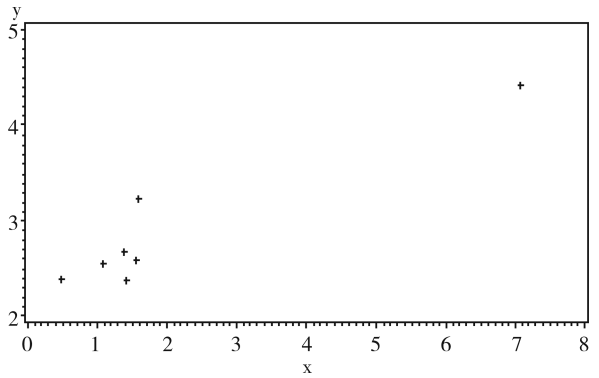

图 1 某实际问题中(x,y)的散布图

图 1 中,最右边的那个点在大多数点的延长线上且偏离得很远,这样的“异常点”可以通过“Cook's D”统计量(其取值远远大于 0.5 就属于异常点)进行诊断;还有一个点在垂直于 x 轴的方向上偏离多数点较远,此类“异常点(在图 1 中偏离较小,暂时可将其视为可疑异常点)”可以通过“学生化残差”统计量(其取值的绝对值大于 2 时就可定为异常点)进行诊断。这两种统计量的计算公式此处从略。

1.6 筛选自变量和剔除异常点的意义和方法

1.6.1 筛选自变量的意义和方法

筛选自变量的意义在于:淘汰掉那些对因变量影响无统计学意义的自变量,使拟合的多重线性回归模型精简且具有更高的价值。在 SAS 软件包中,实现对自变量进行筛选的 model 语句的选项有以下 8 个[4],即:

①selection = forward(采用前进法筛选自变量,变量经假设检验只进不出);

②selection = backward(采用后退法筛选自变量,变量经假设检验只出不进);

③selection = stepwise(采用逐步法筛选自变量,变量经假设检验有进有出);

④selection = maxr(基于最大复相关系数的平方筛选自变量);

⑤selection = minr(基于最小复相关系数的平方筛选自变量);

⑥selection = rsquare(基于复相关系数的平方由大到小且所含自变量个数由少到多给自变量组合排序);

⑦selection = adjrsq(基于校正的复相关系数的平方由大到小且所含自变量个数由少到多给自变量组合排序);

⑧selection = cp(基于 Mallow's C_P 统计量的计算结果筛选自变量,此值越接近当前回归方程中自变量的个数,表明此时的回归方程质量越好)。

1.6.2 剔除异常点的意义和方法

若资料中确实存在严重的异常点,可能会导致所拟合的多重线性回归模型严重歪曲实际情况,很可能会得出误差很大的预测结果;若借助 SAS 软件中的 REG 过程来拟合多重线性回归模型,只需要在 model 语句中增加选择项"r"(即要求进行残差分析)就可实现。

2 构建多重线性回归分析模型方法

2.1 基于经典统计思想构建多重线性回归分析模型

构建多重线性回归分析模型的经典统计思想是:假定模型(1)中的误差项服从正态分布且各自变量组合条件下因变量 Y 的方差相等,在此假定成立的条件下,基于最小二乘原理构造一个偏差函数 Q[5],见式(6):

$$Q = \sum (Y - \hat{Y})^2 = \sum [Y - (b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m)]^2 \quad (6)$$

式(6)中的 Q 就是各观测点上的因变量 Y 与其预测值 $\hat{Y}$ 之间的偏差(也叫做残差)平方和,所谓最小二乘原理实际上就是希望在偏差函数 Q 达到最小值时,求出式(6)中全部回归系数的估计值(包括截距项)。

为了使 Q 达到最小,由高等数学知识可知,将 Q 对 b_0、b_1、b_2、…、b_m 求一阶偏导数并且使之为 0,就可以得到包含 $(m+1)$ 个方程的正规方程组,然后利用求解线性方程组的方法就可由该正规方程组解得各个参数的估计值。为简便起见,下面以导出简单直线回归模型中参数估计值为例,呈现其推导过程:

设简单直线回归方程为:$y_i = a + bx_i$,令:

$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$,求 Q 关于 a、b 的偏导数,并令其为 0,从偏微分方程组中解出 a 和 b,即

$$\frac{\partial Q}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1),$$

$$\frac{\partial Q}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i),\ \text{令}$$

$$\begin{cases} -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_I - an - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_iy_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \Rightarrow$$

$$\begin{cases} a = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right) \\ \sum_{i=1}^{n} xy - a\sum_{i=1}^{n} x_i = b\sum_{i=1}^{n} x_i^2 \end{cases}$$

$$\Rightarrow \begin{cases} a = \frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_iy_i = \left(\frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases}$$

$$\Rightarrow \begin{cases} a = \bar{y} - b\bar{x} \\ \sum xy - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = b\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) \end{cases}$$

$$\Rightarrow \begin{cases} a = \bar{y} - \bar{b} \\ l_{xy} = bl_{xx} \end{cases},\text{得:}$$

$$a = \bar{y} - \bar{b}, b = \frac{l_{xy}}{l_{xx}}$$

按以上程序确定直线回归方程中两个参数估计值,就被称为按最小二乘法或最小平方法进行参数估计。

2.2 基于贝叶斯统计思想构建多重线性回归分析模型

贝叶斯统计思想是充分利用并有效整合"样本信息、总体信息和先验信息",再基于经典统计中的"概率分布知识"和蒙特卡罗统计思想中的"随机抽样"和"统计模拟"技术,构建所需要的多重线性回归模型,进而进行"共轭先验下的贝叶斯推断"和/或"广义先验下的贝叶斯推断"求出多重线性回归模型中的"估计回归系数矩阵"和"估计误差协方差矩阵"。这些内容的数学味过浓,感兴趣的读者参阅文献[6],详细内容此处从略。

2.3 基于机器学习统计思想构建多重线性回归分析模型

机器学习的含义是希望通过对计算机编程,使它能够根据已有的输入数据进行学习[7]。这里所

说的“学习”与人类为了了解未知事物或不会的知识时进行的“学习”是有区别的。在解决不同的问题时,这种学习需要“具体化”。例如,在进行简单直线回归分析时,若基于经典统计思想,前面介绍了依据最小二乘原理可以直接推导出直线回归方程中两个参数估计值的计算公式(7),而利用机器学习方法却无法给出公式(7)。若基于机器学习统计思想,其解决问题的思路如下[8]:

将拟分析的资料随机划分成两部分,分别称为“训练集”与“测试集”,设其样本含量分别为 n 与 k。让计算机在训练集上进行学习,在测试集上设法使均方误差[见式(8)]达到最小值,此时,所得的结果就是“机器学习的结果”(相当于得到了回归模型)。然后,将“机器学习的结果”用于测试集,即把测试集中自变量的数值代入已创建的回归模型求出因变量的估计值,再求出测试集上因变量的残差平方和,进而求出均方误差,见式(9)。

$$MSE_{train}=\frac{1}{n}\sum_{i=1}^{n}(y_i^{train}-\hat{y}_i^{train})^2 \tag{8}$$

$$MSE_{test}=\frac{1}{k}\sum_{i=1}^{k}(y_i^{test}-\hat{y}_i^{test})^2 \tag{9}$$

在上述的计算过程中,用机器学习法实现回归分析时希望达到的最终目的是使式(9) 取得最小值,但在具体实施时,却是基于使式(8) 达到最小值这个目标来不断优化或改进回归模型[见式(10)]中的权重 w 来实现的。

$$\hat{y}=w^Tx \tag{10}$$

在式(10)中,$w\in R^n$ 是参数向量。

值得一提的是:在经典统计思想中,求一个函数的最小值或最大值通常都是基于高等数学中求极值的方法来实现的;而在机器学习统计思想中,是通过事先给定的最小值的“阈值”来实现的。例如,要求式(8)或式(9)小于阈值“10^{-4}”。显然,阈值越小,迭代计算的次数就会越多。有时阈值定得过小,无论迭代计算多少次都难以满足要求,此时可能就会出现所谓不能收敛的情形了。

3 多重线性回归分析模型的假设检验

3.1 对多重线性回归模型进行整体检验

在估计出回归模型的参数以后,需要对回归方程进行显著性检验,检验的原假设为:

$$H_0:\beta_1=\beta_2=\cdots=\beta_m=0 \tag{11}$$

该假设表示所有的偏回归系数都为 0,也就是全部自变量对因变量的作用都没有统计学意义,相应的备择假设为偏回归系数不全为 0。检验的方法是方差分析,其基本思想与简单线性回归相同,将总的离均差平方和分解为回归平方和与残差平方和,然后构造 F 统计量,F 统计量分子与分母的自由度分别为 $\nu_R=m$、$\nu_E=n-m-1$。其计算公式为:

$$F=\frac{SS_R/\nu_R}{SS_E/\nu_E}=\frac{MS_R}{MS_E} \tag{12}$$

式中 MS_R 与 MS_E 分别称为回归均方和残差均方。求出 F 值后查 F 界值表,如果得到的 P 值小于事先确定的显著性水平,就说明回归方程有统计学意义。

3.2 对多重线性回归模型中各参数进行逐一检验

对整个回归方程进行显著性检验之后,还有必要对每一个偏回归系数进行检验,检验的原假设和备择假设分别为 $H_0:\beta_i=0$;$H_1:\beta_i\neq0$。具体检验时,可以根据偏回归平方和构造 F 统计量,也可以采用 t 检验,这两种方法是等价的。

多重线性回归中自变量 X_i 的偏回归平方和用 P_i 表示,它代表从回归方程中剔除 X_i 后回归平方和的减少量,或者在 $m-1$ 个自变量的基础上新增加 X_i 后回归平方和的增加量。偏回归平方和的大小可用来衡量自变量 X_i 在回归中所起作用的大小,它的取值越大,说明 X_i 越重要。对自变量 X_i 进行检验的统计量 F_i 为:

$$F_i=\frac{P_i/1}{SS_E/(n-m-1)} \tag{13}$$

该统计量分子和分母的自由度分别为 1 与 $n-m-1$。

使用 t 检验对偏回归系数进行检验时,检验统计量 t_i 的计算公式为:

$$t_i=\frac{b_i}{S_{b_i}} \tag{14}$$

式中 b_i 是偏回归系数的估计值,S_{b_i}是 b_i 的标准误,t_i 服从自由度为 $\nu=n-m-1$ 的 t 分布。

求出上述检验统计量后,可以查相应的界值表得出 P 值,从而判定 X_i 与 Y 之间是否存在线性关系。

4 多重线性回归分析模型拟合效果的评价

一个多重线性回归分析模型拟合效果如何,可从以下几点来考量:

第一,拟合的多重线性回归方程在整体上有统计学意义;

第二,多重线性回归方程中各回归参数估计值的假设检验结果都有统计学意义;

第三,多重线性回归方程中各回归参数估计值的正负号与专业上的含义相吻合;

第四,根据多重线性回归方程计算出因变量的所有预测值在专业上都有意义;

第五,若有多个较好的多重线性回归方程时,残差平方和较小且多重线性回归方程中所含的自变量的个数又较少者为最佳。

参考文献

[1] 李娇婷, 许昊, 杨黄浩, 等. 利培酮和氯氮平对精神分裂症患者自发性脑活动的不同影响[J]. 四川精神卫生, 2017, 30(1): 27-31.

[2] 赵茜, 史晓宁, 路亚洲, 等. 住院抑郁症患者联合心境稳定剂治疗的影响因素研究[J]. 四川精神卫生, 2017, 30(2): 113-116.

[3] 方积乾. 卫生统计学[M]. 7 版. 北京: 人民卫生出版社, 2017: 238-265.

[4] 胡良平. 科研设计与统计分析[M]. 北京: 军事医学科学出版社, 2012: 513-551.

[5] 胡良平, 毛玮. 外科科研设计与统计分析[M]. 北京: 中国协和医科大学出版社, 2012: 274-300.

[6] 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.

[7] 沙伊· 沙莱夫-施瓦茨,沙伊· 本-戴维. 深入理解机器学习: 从原理到算法[M]. 张文生,译. 北京: 机械工业出版社, 2017: 1-65.

[8] 伊恩· 古德费洛, 约书亚· 本古奥, 亚伦· 库维尔. 深度学习[M]. 赵申剑, 黎彧君, 符天凡, 等译. 北京: 人民邮电出版社, 2017: 63-104.

(收稿日期:2018-01-29)

(本文编辑:陈 霞)

科研方法专题策划人——胡良平教授简介

胡良平, 男, 1955 年 8 月出生, 教授, 博士生导师, 曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授; 现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长,《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 45 部, 参编统计学专著 10 部;发表第一作者学术论文220余篇,发表合作论文 130 余篇, 获军队科技成果和省部级科技成果多项; 参加并完成三项国家标准的撰写工作; 参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中, 为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学, 在全国各地作统计学学术报告 100 余场, 举办数十期全国统计学培训班, 培养多名统计学专业硕士和博士研究生。近几年来, 参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想, 独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。