DOI:10.3724/SP.J.1008.2014.00865

·论　著·

对策理论在线性回归模型自变量重要性估计中的分析及应用

贾孝霞[1,2],伍立志[3],杨　文[2],沈其君[1,2*]
1. 浙江医药高等专科学校基础部,宁波 315000
2. 宁波大学医学院预防医学系,宁波 315000
3. 浙江省疾病预防控制中心环境与职业卫生所,杭州 310051

[摘要] **目的**　基于对策理论中求解Shapley值法构建在线性回归模型中当自变量存在多重共线性时求解自变量相对重要性的方法,同时提出自变量的序列重要性偏R^2值的概念。**方法**　应用Shapley在1953年提出的对策理论中求解Shapley值法,以757例正常人的白细胞(WBC)、红细胞(RBC)、血小板(PLT)及红细胞压积(HCT)4项血常规指标作为自变量,分析这些指标对血红蛋白(HB)的相对重要性大小以及分析序列重要性偏R^2值的实际意义。最后将估计结果和传统的指标及现推荐的指标进行比较。**结果**　最后进入回归模型的自变量有*RBC*、*PLT*和*HCT*,对*HB*的相对重要性估计值分别为0.355 3、0.012 4、0.553 8,用Shapley值法估计的自变量相对重要性值与现最为推荐的优势分析法的估计结果一致。自变量以不同的次序进入回归模型的序列重要性偏R^2值不同。**结论**　*HCT*对*HB*的影响最大,其次是*RBC*,*PLT*影响较小,结果与相关性排序一致,说明用Shapley值法估计自变量的相对重要性具有合理性。

[关键词] 线性模型;相对重要性;序列重要性偏R^2值;对策理论
[中图分类号] R 195.1　[文献标志码] A　[文章编号] 0258-879X(2014)08-0865-05

Analysis and application of game theory in estimating variable importance in linear model

JIA Xiao-xia[1,2], WU Li-zhi[3], YANG Wen[2], SHEN Qi-jun[1,2*]
1. Department of Basic Medical Sciences, Zhejiang Pharmaceutical College, Ningbo 315000, Zhejiang, China
2. Department of Preventive Medicine, Ningbo University School of Medicine, Ningbo 315000, Zhejiang, China
3. Department of Environmental and Occupational Health, Zhejiang Center for Disease Control and Prevention, Hangzhou 310051, Zhejiang, China

[**Abstract**] **Objective**　To apply Shapley value analysis of the game theory for evaluating the relative importance of the predictors in the linear regression when colinearity exists, and to provide a new concept of sequential importance partial R^2. **Methods**　Shapley value analysis of game theory(proposed by Shapley in 1953) was used to evaluate the influencing factors of hemoglobin(HB) in 757 normal adults, by regressing HB on four predictors including the white blood cell(WBC), red blood cell(RBC), blood platelet(PLT) and hematocrit(HCT); meanwhile, the sequential importance partial R^2 was used to analyze its practical significance. Finally the estimated results of Shapley value was compared with others measures including traditional methods and recommended method. **Results**　A succinct set of predictors including *RBC*, *PLT* and *HCT* was identified for establishing a multiply regression, with their relative importance values being 0.355 3, 0.012 4 and 0.553 8, respectively. The results of relative importance were consistent between Shapley value and dominance analysis. Moreover, it was found that the partial R^2 of predictors had different marginal contributions in different orders. **Conclusion**　*HCT* has the largest contribution to *HB*, followed by *RBC*, and *PLT* has the least effect to *HB*. The order of contributions is consistent with the correlation matrix, indicating that the relative importance of the predictors in Shapley value is reasonable.

[**Key words**] linear regression; relative importance; sequential importance partial R^2; game theory
[Acad J Sec Mil Med Univ,2014,35(8):865-869]

[收稿日期] 2013-12-31　[接受日期] 2014-02-21
[基金项目] 国家自然科学基金(81172771). Supported by National Natural Science Foundation of China(81172771).
[作者简介] 贾孝霞,硕士生. E-mail:janemyth@163.com
*通信作者(Corresponding author). Tel: 0574-87600921, E-mail:shenqijun@nbu.edu.cn

线性回归分析中，研究者通常需了解每个自变量对因变量变异的贡献大小即自变量的相对重要性。当自变量间不相关或相关性较弱时，自变量的相对重要性可由一些传统的、简单的指标表示，如标准回归系数的平方、偏相关系数的平方以及半偏相关系数的平方等。但在实际问题中，自变量间不相关的数据并不常见。近年来，国际学者提出几种关于自变量的相对重要性的估计方法，如优势分析法[1-4]、比例边界方差分解法[5]和相对权重[6-8]等。但大多数学者对这些方法的使用仍存在较大的分歧，原因是这些方法构建前提条件不同，有时结果也不尽相同。本研究引用 Shapley 在 1953 年提出的对策理论(game theory)法求解当自变量间存在多重共线性时自变量的相对重要性，并用实际的医学数据探讨对策理论在医学领域中的意义和作用。

1　统计方法原理

1.1　*自变量的序列重要性偏 R^2 值计算*　下面给出本研究提出的自变量的序列重要性偏 R^2 值的概念。假设有 3 个自变量，记为 X_i，X_j 和 X_k，自变量依次以 ijk 次序引入回归方程，分别计算其在不同自变量构成的回归模型中相应的 R^2，分别记为 R_1^2、R_2^2、R_3^2，则 R_1^2 为 X_i 的单独贡献值，R_2^2 为在已经引入 X_i 后再引入 X_j 时 X_i 和 X_j 的联合贡献值，R_3^2 看作是 3 个自变量的总的贡献值。那么 $(R_2^2-R_1^2)$ 表示自变量 X_j 对对应子集的序列重要性偏 R^2 值，同理 $(R_3^2-R_2^2)$ 表示自变量 X_k 对对应子集的序列重要性偏 R^2 值。一般地，假设对因变量有影响的自变量的个数有 p 个，对于某一特定的进入序列，当 $i=1,2,\cdots,p$ 时，$(R_i^2-R_{i-1}^2)$ 为第 i 个自变量在对应序列中的序列重要性偏 R^2 值。当自变量个数为 3 时，所有自变量在不同序列中的序列重要性偏 R^2 值计算如表 1 所示。表 1 中，同一自变量在不同序列中的序列重要性偏 R^2 值不同。

表 1　自变量的序列重要性偏 R^2 值计算

Tab 1　Calculation of sequence importance partial R^2 of each variable

Enter the model order	Sequence importance partial R^2 of each variable		
	X_i	X_j	X_k
ijk	R_i^2	$R_{ij}^2-R_i^2$	$R_{ijk}^2-R_{ij}^2$
ikj	R_i^2	$R_{ik}^2-R_i^2$	$R_{ikj}^2-R_{ik}^2$
jik	$R_{ji}^2-R_j^2$	R_j^2	$R_{jik}^2-R_{ji}^2$
jki	$R_{jki}^2-R_{jk}^2$	R_j^2	$R_{jk}^2-R_j^2$
kij	$R_{ki}^2-R_k^2$	$R_{kij}^2-R_{ki}^2$	R_k^2
kji	$R_{kji}^2-R_{kj}^2$	$R_{kj}^2-R_k^2$	R_k^2

从表 1 可以看出，对于 p 个自变量，有 $p!$ 个不同的进入序列，例如，当自变量个数为 3 时，就有 6 个不同的进入序列。上表列出当自变量个数为 3 时所有自变量以不同的次序进入回归模型时每个自变量的序列重要性偏 R^2 值。

1.2　*自变量重要性的估计*　在实际问题中，应用线性回归模型研究影响因变量的一些因素之间往往存在多重共线性，研究者也关心在自变量的排序未知情况下影响因变量的自变量中每个自变量对因变量变异的贡献大小，也就是将模型的 R^2 如何公平、有效地分配给每个自变量。解决这一问题的方法是对某一自变量在所有引入序列中估计的序列重要性求平均，其原理是以 R^2 作为特征函数利用对策理论求解得出。

对策理论旨在解决的问题是：在一次多人联合参与的工作中，找到一个分配函数将合作产生的总效益公平、有效地分配给联盟中的每位参与者。Shapley 在 1953 年首先提出了具体的解法[9]，因此也被称为 Shapley 值法。后在 1960 年 Roberts 利用若干公理对此方法给出了严格、详细的公式证明与推导，也使得这作为一个公理被研究者广泛使用。在考虑公平、有效地分配总效益时，首先应该注意的问题是在这项工作中的每个参与者的效益之和应该等于所有参与者通过合作产生的总效益。在评估每个参与者的效益时，不能只单独考虑每个参与者单独个人的效益，还应综合考虑与其他参与者的联合贡献。理论上，根据自变量相对重要性概念，在线性回归模型中求解自变量相对重要性可以看作是这一问题的同构问题，同构性解释如下：(1)参与者看作线性回归模型中影响因素或自变量；(2)联盟看作是各个影响因素或自变量的组合；(3)特征函数看作是线性回归模型的 R^2；(4)效益分配看作是线性模型总变异 R^2 的分解；(5)哑参与者看作是与因变量无关的因素或自变量。根据同构性解释，下面首先分析求解自变量重要性的 Shapley 值必须满足的 4 个公理[9]，然后再给出自变量相对重要性的 Shapley 值定理。

假设在回归模型中有 p 个自变量，记为 $X=\{X_1,X_2,\cdots,X_p\}$；设 S 为自变量 X 中任意 $s(s\leqslant p)$ 个自变量组成的一子集；R^2 为用于估计每个子集的

效用的实值特征函数;SV 为贡献分配函数,$SV=(SV_1,SV_2,\cdots,SV_p)$。

公理 1(对称性公理):如果 SV_i 为在 S 中第 i 个因素的贡献值,当这个因素在 S 中记为第 j 个因素时,假定这时子集记为 S',那么 $SV_i[R^2,S]=SV_j[R^2,S']$。解释为,在一个特定的子集中因素的名称对贡献值的确定无影响,贡献值的确定仅对在子集中选择的自变量和特征函数敏感。

公理 2(有效性公理):$\sum_{i=1}^{n} SV_i=R^2$。解释为,各个因素的贡献总和等于模型总变异 R^2,或者各个因素的贡献只能在模型的总变异 R^2 中进行分解。

公理 3(线性公理):如果 X_1 和 X_2 是针对同一自变量集 X 的两个子集,有 $X=X_1\cup X_2$ 且 $X_1\cap X_2=\varnothing$,R_1^2 是对 X_1 的任意特征值函数,R_2^2 是对 X_2 的任意特征函数,且有 $R^2=R_1^2+R_2^2$,则有 $SV[X,R^2]=SV[X_1,R_1^2]+SV[X_2,R_2^2]$。解释为,如果将自变量集 X 分为两个独立且完备的子集,则由该两个子集分别构成的模型的贡献和等于总模型的变异 R^2。

公理 4(哑公理):如果自变量 X_i 为在 S 中一个哑变量,则 $SV_{X_i}[S,R^2]=0$。解释为,在回归模型中,与因变量不相关性的自变量的重要性也为零。

Shapley 值定理:在 p 个自变量 $X=\{X_1,X_2,\cdots,X_p\}$ 和以 R^2 为特征函数的条件下,让 $SV=(SV_1,SV_2,\cdots,SV_p)$ 表示在 R^2 上的一组值,假设每个 SV_i 都满足线性和哑性公理,且 SV 满足对称性和有效性公理,则在 SV 的范围内对任意的 $X_i\in X$ 都有 $SV_i(X)=\sum_{S\subset X/X_i}\gamma_n(S)[R^2(S\cup\{X_i\}-R^2(S))]$,$\gamma_n(S)=\frac{s!\ (p-s-1)!}{p!}$。其中 S 为不包括自变量 X_i 的子集,$S\cup\{X_i\}$ 为包含自变量 X_i 的子集,s 为子集 S 中自变量的个数,p 为所有自变量的个数。经证明 R^2 是唯一一个满足公理 1 至公理 4 的特征函数[9]。

2 实际案例分析

2.1 影响血红蛋白的因素的相关矩阵和回归分析结果 现有 757 例不同年龄正常人的白细胞(WBC)、红细胞(RBC)、血小板(PLT)、红细胞压积(HCT)和血红蛋白(HB)5 项血常规指标,利用 SAS 9.2 统计分析软件,以 HB 为因变量,其他变量为自变量进行回归分析。因变量和自变量的相关系数矩阵见表 2。从表 2 可以看出,影响 HB 的各个影响因素之间存在显著的相关性。首先利用逐步回归法、调整 R^2、C_p 统计量筛选自变量,最后进入回归模型的自变量为 RBC、PLT、HCT。最后估计的回归方程为:$HB=-5.0488+2.4977RBC-0.0065PLT+3.3368HCT$,$R^2=0.9215$,对方程检验 $F=2\,940.61$,$P<0.0001$,说明模型具有统计学意义。

表 2 因变量和自变量的相关系数矩阵

Tab 2 Correlation matrix between dependent variables and independent variables

	WBC	*RBC*	*PLT*	*HCT*	*HB*
WBC	1	-	-	-	-
RBC	0.181 4**	1	-	-	-
PLT	0.254 9**	−0.090 1*	1	-	-
HCT	0.198 9**	0.861 4**	−0.163 0**	1	-
HB	0.176 2**	0.846 1**	−0.176 6**	0.958 8**	1

WBC: White blood cell; RBC: Red blood cell; PLT: Platelet; HCT: Hematocrit; HB: Hemoglobin. *: Correlation is significant at the 0.05 level (2-tailed); **: Correlation is significant at the 0.01 level (2-tailed)

2.2 影响血红蛋白的自变量的序列重要性偏 R^2 值 利用 SAS 9.2 软件,分别计算影响 HB 的 3 个自变量单独的 R^2、不同组合的 R^2 以及 3 个自变量的总 R^2,计算结果如图 1 所示;利用图 1 计算各个自变量的序列重要性偏 R^2 值,结果见表 3。

从表 3 可以看出同一个自变量以不同的次序进入回归模型时的序列重要性偏 R^2 值不同。如本例中由 3 个自变量组成的 6 个不同序列中,同一自变量(如 RBC)的边缘贡献值的变化是从 0.001 5 到 0.715 8。从表中也可以看出对因变量影响最大的自变量是 HCT,最大达到 0.919 3,在测量选择上应首先考虑对 HCT 测量。

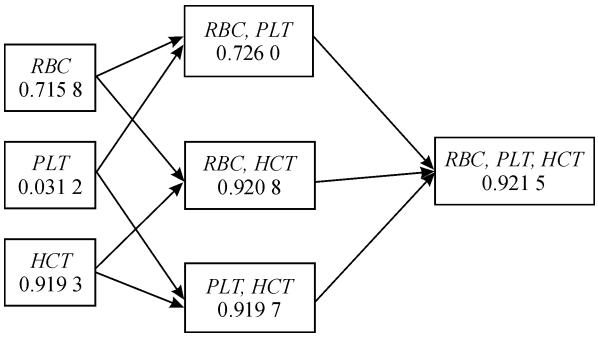

图 1　影响血红蛋白的自变量以不同次序进入模型的 R^2

Fig 1　R^2 of variables entering the model in different orders

2.3　Shapley 值的自变量相对重要性估计　结合 Shapley 值定理给出的公式，利用 SAS 9.2 软件编写相应的程序分别计算 3 个自变量重要性的估计值，事实上，自变量相对重要性估计值是自变量序列重要性偏 R^2 估计值的平均值。同时分别应用标准回归系数的平方（β_{iy}^2）、偏相关系数的平方（partial ρ^2）、乘积尺度（$\beta_i r_{iy}$）、相对权重（ε_i）、优势分析（C_i）和对策理论 Shapley 值（SV_i）6 种方法估计每个自变量的相对重要性。结果见表 4，表中同时给出传统的自变量相对重要性测量方法结果。

表 3　各自变量的序列重要性偏 R^2 值

Tab 3　Sequence importance partial R^2 value of *RBC*, *PLT* and *HCT*

Enter the model order	Sequence importance partial R^2 value of each variable		
	RBC	*PLT*	*HCT*
RBC, *PLT*, *HCT*	0.715 8	0.010 2	0.195 5
RBC, *HCT*, *PLT*	0.715 8	0.000 7	0.205 0
PLT, *RBC*, *HCT*	0.694 8	0.031 2	0.195 5
PLT, *HCT*, *RBC*	0.001 8	0.031 2	0.888 5
HCT, *RBC*, *PLT*	0.001 5	0.000 7	0.919 3
HCT, *PLT*, *RBC*	0.001 8	0.000 4	0.919 3
Average value	0.355 3	0.012 4	0.553 8

RBC：Red blood cell；PLT：Platelet；HCT：Hematocrit

表 4　不同方法计算的自变量相对重要性值（R^2=0.921 5）

Tab 4　Relative importance of variables by different methods（R^2=0.921 5）

Variable	β_{iy}^2	Partial ρ^2	$\beta_i r_{iy}$	ε_i	C_i	SV_i
RBC	0.006 9	0.021 9	0.070 2	0.356 5	0.355 3	0.355 3
PLT	0.000 6	0.007 7	0.004 4	0.013 9	0.012 4	0.012 4
HCT	0.780 1	0.713 3	0.846 8	0.550 7	0.553 8	0.553 8
Total	0.787 6	0.743 0	0.921 5	0.921 5	0.921 5	0.921 5

RBC：Red blood cell；PLT：Platelet；HCT：Hematocrit；β_{iy}^2：Suqare of standard regression coefficient；Partial ρ^2：Suqare of partial correlation coefficient；$\beta_i r_{iy}$：Product measure；ε_i：Relative weight；C_i：Dominance analysis；SV_i：Shapley value

3　讨　论

线性回归模型应用中，自变量之间存在多重共线性时，传统的统计量确定自变量的相对重要性是不完全和失效的[10]，例如，本研究结果显示用标准回归系数平方和偏相关系数的平方计算的 3 个自变量的重要性之和分别等于 0.787 6 和 0.743 0，都与模型总变异相差较大。其他的自变量相对重要性的估计方法所估计的自变量的重要性值总和等于模型 R^2。从分析结果也可以看出，优势分析和对策理论的估计结果几乎一致，原因是优势分析所用的基本思想和对策理论的基本思想一致。自变量间存在多重共线性对回归方程的预测能力并没有太大影响，但却对回归系数的估计和自变量重要性估计影响较大，从而在解释和衡量单个自变量对因变量的作用时产生较大的偏差，特别是自变量之间的共线性水平越高时，造成偏差就越大。本研究在计算序列重要性偏 R^2 值的基础上，借助于对策理论求平均重要性的方法，这样客观、贴切地反映了各个自变量在模型中的作用大小。本研究通过利用影响 *HB* 含量的实例分析发现，在影响 *HB* 的自变量中，*HCT* 对 *HB* 含量的影响最大，重要性估计值为 0.553 8，占模型总变异的 60.10%，其次是 *RBC*，重要性估计值为 0.355 3，占模型总变异的 38.55%，*PLT* 的影响不大，估计值为 0.012 4，占总变异的 1.34%，重要性的排序与相关性排序一致，说明分析结果是合理的。

本研究提出了序列重要性偏 R^2 值的概念，结果显示同一自变量以不同的次序进入模型时计算的序列重要性偏 R^2 值不同，这对一些实际的应用，如疾病的预后影响因素及在疾病预防或控制策略的选择上提供了定量的依据。另外从序列重要性偏 R^2 值

中可以找出影响因变量的自变量中,对因变量的贡献值最大的自变量(即在所有的 $p!$ 个进入序列中 $R^2_{i|1,2,\cdots k-1,k+1,\cdots,p}$ 达到最大),在构建回归模型时应首先考虑将其纳入,依次直到所有的影响因素都纳入为止,可以提高回归模型预测能力[11]。

最后引用 Shapley 值求解自变量重要性的最重要的原因是:它不是一个探索式的理论方法而是基于 4 个公理推导且已经作为 1 个定理使用的方法[9]。另外,Shapley 值法为更加复杂的问题提供了一个比较接近实际的模型,原因是它比较和平均了自变量所有可能的子集构成模型的总变异 R^2[12]。

4 利益冲突

所有作者声明本文不涉及任何利益冲突。

[参 考 文 献]

[1] Budescu D V. Dominance analysis : a new approach to the problem of relative importance of predictors in multiple regression[J]. Psychol Bull, 1993, 114:542-551.

[2] Azen R, Budescu D V, Reiser B. Criticality of predictors in multiple regression[J]. Br J Math Stat Psychol, 2001,54(Pt 2):201-225.

[3] Azen R,Budescu D V. The dominance analysis approach for comparing predictors in multiple regression[J]. Psychol Methods, 2003, 8:129-148.

[4] Budescu D V,Azen R. Beyond global measures of relative importance: some insights from dominance analysis [J]. Organ Res Methods, 2004, 7:341-350.

[5] Grömping U. Estimators of relative importance in linear regression based on variance decomposition[J]. Am Statistician, 2007, 61:139-147.

[6] Johnson J W. A heuristic method for estimating the relative weight of predictor variables in multiple regression[J]. Mult Behav Res, 2000, 35:1-19.

[7] Johnson J W,Lebreton J M. History and use of relative importance indices in organizational research[J]. Organ Res Methods, 2004, 7:238-257.

[8] Tonidandel S, LeBreton J M,Johnson J W. Determining the statistical significance of relative weights[J]. Psychol Methods, 2009, 14:387-399.

[9] Roth A E. The Shapley value:essays in honor of Lloyd S. Shapley[M]. Cambridge: Cambridge University Press, 1988: 330.

[10] Jian B. A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking[J]. J Sens Stud, 2012, 27:87-101.

[11] Beyene J, Atenafu E G, Hamid J S, To T, Sung L. Determining relative importance of variables in developing and validating predictive models[J]. BMC Med Res Methodol, 2009, 9:64-74.

[12] Lipovetsky S, Conklin M. Analysis of regression in game theory approach[J]. Appl Stochastic Models Bus Indus, 2001, 17:319-330.

[本文编辑] 商素芳