基于 Gram-Schmidt 过程的多项式回归建模方法

王惠文 郭丽娟

(北京航空航天大学 经济管理学院，北京 100191)

摘 要：多项式回归模型是一种常用的非线性回归方法.由于在多项式回归模型中，自变量之间往往存在较强的相关关系，采用普通最小二乘回归方法来估计回归系数会存在较大的计算误差.为了提高多项式回归模型的预测准确性和可靠性，提出一种基于 Gram-Schmidt 过程进行多项式回归的建模方法，可以实现自变量集合的正交化，克服自变量集合多重共线对回归建模的不良影响，从而有效地运用最小二乘建立回归模型.同时可以进行信息筛选有效选取对因变量有显著解释作用的自变量，排除自变量中的冗余信息.采用仿真数据分析，检验了该方法的有效性.

关 键 词：Gram-Schmidt 过程；多项式回归；多重相关性

中图分类号：O 212

文献标识码：A **文 章 编 号**：1001-5965(2008)11-1349-04

Polynomial regression modeling based on Gram-Schmidt process

Wang Huiwen Guo Lijuan

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: The polynomial regression model is a widely applied nonlinear regression method. Since the high correlation exists among independent variables in the polynomial regression model, it will induce excessive computational error to estimate coefficients with the ordinary least square regression. A method of polynomial regression modeling based on Gram-Schmidt process which can achieve the orthogonalization of the independent variables and overcome the adverse effects of multicollinearity to regression modeling was proposed, so as to apply ordinary least square to regression modeling effectively. The independent variables including notable explaining information can be selected effectively, at the same time redundant information is deleted. Simulation data analysis was adopted to test the effectiveness of the method.

Key words: Gram-Schmidt process; polynomial regression; multiple correlation

在实际工作中，人们在采用回归模型解释因果变量间的相关关系时，经常会遇到自变量之间存在幂乘关系的情况.在这种情况下，多项式回归模型成为一个合理的选择.除此之外，由于许多复杂的非线性模型都可以用适当项数的多项式去逼近，因此在函数关系比较复杂的问题中，经常可以忽略因变量与自变量的确切函数关系，用多项式模型进行回归分析.

对于多项式回归模型，由于自变量之间存在较强的相关关系，因而如果采用普通最小二乘回归方法(OLS，Ordinary Least Square)来估计变量的回归系数，则会存在较大的计算误差.在此情况下，虽然 OLS 得到的估计量依然是总体参数的最小方差无偏估计量，但其估计方差却会非常大.如果仍然采用 OLS 拟合回归模型，则模型的精度和稳定性都不能得以保证，并且回归建模的结果还会出现许多反常现象.因此，为了解决这个问题，人们采用了多种方法，其中主成分分析和偏最

收稿日期：2007-11-13

基金项目：国家自然科学基金资助项目(70371007，70521001，70531010)；北京市自然科学基金资助项目(9052006)

作者简介：王惠文(1957－)，女，辽宁大连人，教授，wanghw@vip.sina.com.

小二乘回归方法[1-3]被认为能够取得较好的效果.

从两种方法的计算过程来看,都是首先得到代表原始自变量集合大部分信息量的新综合变量,然后使用这些综合变量对因变量进行回归分析.由于得到的新综合变量之间是相互正交,所以避免了直接利用最小二乘方法所遇到的问题,并且提高了预测的准确性和可靠性[4].

在本文中,将提出一种基于 Gram-Schmidt 过程[5]的回归方法,提取对因变量有最佳解释能力的自变量,并将这些自变量变换成直交变量,从而有效地运用最小二乘建立回归模型.

1 方法原理

本节首先给出论文研究的理论依据,然后提出基于 Gram-Schmidt 过程的多项式回归建模的方法.多项式回归大致可以分为以下 2 个步骤:

1)若自变量间存在幂乘关系,或者存在更复杂的非线性关系,但具体模型形式又难以准确判断时,可以用包含 $p+1$ 个待估参数的 p 次多项式来近似表达这个未知的函数关系.p 次多项式的形式为

$$y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_p x_i^p \quad i=1,2,\cdots,n \tag{1}$$

2)用变量变换法将多项式模型转化为多元线性模型,即令 $x_{1i}=x_i,x_{2i}=x_i^2,\cdots,x_{pi}=x_i^p$,即得线性模型

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi} \tag{2}$$

这样就可以把任意多项式回归转化为多元线性回归模型来求解[6].

多项式回归模型的形式可以表示为

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}+\varepsilon_i \tag{3}$$

其中,ε_i 为随机误差.

记 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}$,$\boldsymbol{x}_j=[x_{j1},x_{j2},\cdots,x_{jn}]^{\mathrm{T}}$,$j=1,\cdots,p$,$\boldsymbol{\varepsilon}=[\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n]^{\mathrm{T}}$.

下面介绍 Gram-Schmidt 过程及其反变换的方法.

任意一组线性无关的变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$,总可以经过如式(4)所示的 Gram-Schmidt 过程[5]变换成一组正交变量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p$.

$$\left.\begin{aligned}
\boldsymbol{z}_1&=\boldsymbol{x}_1\\
\boldsymbol{z}_2&=\boldsymbol{x}_2-\frac{\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1\\
\boldsymbol{z}_3&=\boldsymbol{x}_3-\frac{\boldsymbol{x}_3^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1-\frac{\boldsymbol{x}_3^{\mathrm{T}}\boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2}\boldsymbol{z}_2\\
&\vdots\\
\boldsymbol{z}_p&=\boldsymbol{x}_p-\sum_{k=1}^{p-1}\frac{\boldsymbol{x}_p^{\mathrm{T}}\boldsymbol{z}_k}{\boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_k}\boldsymbol{z}_k
\end{aligned}\right\} \tag{4}$$

式(4)中的 $\boldsymbol{z}_j$ 称为 Gram-Schmidt 变量,等式左侧 $\boldsymbol{z}_j$ 所对应的 $\boldsymbol{x}_j$ 称为与 $\boldsymbol{z}_j$ 关联的变量.

记 $\boldsymbol{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p)$,$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p)$.依据式(4),可以得到[7]

$$\boldsymbol{X}=\boldsymbol{Z}\tilde{\boldsymbol{R}} \tag{5}$$

其中

$$\tilde{\boldsymbol{R}}=\begin{bmatrix}
1 & r_{21} & r_{31} & \cdots & r_{p_1}\\
0 & 1 & r_{32} & \cdots & r_{p_2}\\
0 & 0 & 1 & \cdots & r_{p_3}\\
\vdots & \vdots & \vdots & & \vdots\\
0 & 0 & 0 & \cdots & 1
\end{bmatrix} \tag{6}$$

$$r_{jk}=\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{z}_k}{\boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_k}\quad j=2,\cdots,p;k=1,\cdots,p-1 \tag{7}$$

显然可以得到以下 2 个结论[7]:

1)当变量集合 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的秩为 $s(s<p)$ 时,对它们做 Gram-Schmidt 正交变换后,得到 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p$.这其中,必有 s 个变量是相互直交的,而剩下的 $p-s$ 个为 0 向量.

由此可以看出,Gram-Schmidt 过程除了可以将变量集合中的信息进行正交分解,还具有排除冗余变量(即被变换成 0 向量的那些变量)的功能.

2)当变量集合 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的秩为 p 时,经过 Gram-Schmidt 过程的处理后,得到 p 个非 0 的相互正交的向量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p$.

此时,由式(5)容易得到,Gram-Schmidt 过程的反变换为

$$\boldsymbol{Z}=\boldsymbol{X}\tilde{\boldsymbol{R}}^{-1} \tag{8}$$

根据 Gram-Schmidt 过程的反变换,如果 $\boldsymbol{Y}$ 对 Gram-Schmidt 变量 $\boldsymbol{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p)$ 的回归方程为 $\hat{\boldsymbol{Y}}=\boldsymbol{Z}\hat{\boldsymbol{\beta}}$,则还原成 $\boldsymbol{Y}$ 对 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p)$ 的回归方程,得到

$$\hat{\boldsymbol{Y}}=\boldsymbol{X}\tilde{\boldsymbol{R}}^{-1}\hat{\boldsymbol{\beta}} \tag{9}$$

容易验证,利用 Gram-Schmidt 正交过程对自变量进行处理将完全不破坏 OLS 中的 Gauss-Markov 条件,因此可以使用 OLS 的参数检验技术[6].

基于 Gram-Schmidt 过程的多项式回归分析是通过 Gram-Schmidt 变换过程,在自变量集合中提取对因变量有最佳解释能力的信息,剔除对因变量无显著解释作用的信息以及重复解释的信息,并将提取出来的自变量变换成若干直交变量.在这个新的自变量空间中,由于变量之间是相互正交,所以使用这些对因变量有最佳解释能力的新变量进行回归分析时,将不再会遇到使用普通最小二乘回归分析时出现的问题.基于 Gram-Schmidt 过程的多项式回归分析的计算步骤为:

1) 对于 p 次多项式(1),令 $x_{1i}=x_i, x_{2i}=x_i^2, \cdots, x_{pi}=x_i^p$.

2) 对变量集合 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$ 做中心化处理,使每一个变量的均值为 0.

3) 令 $\boldsymbol{z}_j^1=\boldsymbol{x}_j, j=1,2,\cdots,p$.分别建立 $\boldsymbol{Y}$ 关于 $\boldsymbol{z}_j^1(j=1,2,\cdots,p)$的一元回归模型.在通过 t-检验的自变量中,选 t-value 绝对值最大的 Gram-Schmidt 变量进入模型,不妨设第一个被选中的关联变量为 $\boldsymbol{x}_1$,即 $\boldsymbol{z}_1=\boldsymbol{x}_1$.

4) 令

$$\boldsymbol{z}_j^2=\boldsymbol{x}_j-\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1 \qquad j=2,3,\cdots,p \tag{10}$$

分别求 $\boldsymbol{Y}$ 关于 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_j^2(j=2,\cdots,p)$的二元线性回归.在通过 t-检验的自变量中,选 t-value 绝对值最大的 Gram-Schmidt 变量进入模型.这里不妨设第二个被选中的关联变量为 $\boldsymbol{x}_2$,即 $\boldsymbol{z}_2=\boldsymbol{x}_2-r_{21}\boldsymbol{z}_1$.

5) 令

$$\boldsymbol{z}_j^3=\boldsymbol{x}_j-\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1-\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2}\boldsymbol{z}_2 \qquad j=3,4,\cdots,p \tag{11}$$

分别求 $\boldsymbol{Y}$ 关于 $\boldsymbol{z}_1, \boldsymbol{z}_2$ 与 $\boldsymbol{z}_j^3(j=3,4,\cdots,p)$的三元线性回归.在通过 t-检验的变量中,选 t-value 绝对值最大的 Gram-Schmidt 变量进入模型.不妨设第 3 个被选中的关联变量为 $\boldsymbol{x}_3$,即 $\boldsymbol{z}_3=\boldsymbol{x}_3-r_{31}\boldsymbol{z}_1-r_{32}\boldsymbol{z}_2$.

重复上面的步骤,直到模型外面的所有变量经 Gram-Schmidt 处理后都不能通过 t-检验为止.最终得到回归模型:

$$\hat{y}_i=\alpha_0+\alpha_1 z_{1i}+\cdots+\alpha_m z_{mi} \tag{12}$$

由于 $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_m$ 是原自变量的线性组合,所以最后可以由 Gram-Schmidt 过程的反变换得到 y_i 关于 $x_{1i}=x_i, x_{2i}=x_i^2, \cdots, x_{pi}=x_i^p$ 的回归模型:

$$\hat{y}_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}^2+\cdots+\beta_m x_{pi}^m \tag{13}$$

2　仿真实验

下面采用仿真数据来检验以上方法的有效性.

仿真数据的生成:令变量 $\boldsymbol{x}_1$ 为服从均值为 0,方差为 25 的正态分布的 50 个随机数.然后分别令 $\boldsymbol{x}_2=\boldsymbol{x}_1^2$;$\boldsymbol{x}_3=\boldsymbol{x}_1^3$;$\boldsymbol{Y}=\boldsymbol{x}_1+\boldsymbol{x}_2+\boldsymbol{x}_3+\boldsymbol{\varepsilon}$,其中 $\boldsymbol{\varepsilon}\sim N(0,1)$.

这样,得到 50 个随机向量$(x_{1i}, x_{2i}, x_{3i}, y_i)$,$i=1,2,\cdots,50$.

应用仿真数据检验基于 Gram-Schmidt 过程的多项式回归建模方法的过程如下:

1) 对数据向量$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{Y})$进行标准化处理.为简便记,仍记为$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{Y})$.计算$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$之间的相关系数矩阵 $\boldsymbol{R}$:

$$\boldsymbol{R}=\begin{bmatrix} 1.000 & -0.161 & 0.840 \\ -0.161 & 1.000 & -0.188 \\ 0.840 & -0.188 & 1.000 \end{bmatrix}$$

很显然,自变量 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_3$ 之间存在较高的相关性.

2) 提取综合变量

首先,令 $\boldsymbol{z}_j^1=\boldsymbol{x}_j, j=1,2,3$,分别建立 $\boldsymbol{Y}$ 关于 $\boldsymbol{z}_j^1(j=1,2,3)$的一元回归模型,得到相应的 t-value 如表 1.

表 1　一元线性回归模型中自变量 z_j^1 的 t-value

变　量	z_1^1	z_2^1	z_3^1
t-value	10.744	−0.752	84.863

由于样本容量较大,可以取 t-检验的阈值为 $t_{0.025}(50-1-1)\approx 1.96$.在表 1 通过 t-检验的自变量中,选择 t-value 绝对值最大的自变量为 $\boldsymbol{z}_1$,即 $\boldsymbol{z}_1=\boldsymbol{x}_3$.分别将 $\boldsymbol{x}_1, \boldsymbol{x}_2$ 与 $\boldsymbol{z}_1$ 做 Gram-Schmidt 正交变换,将得到的变量分别与 $\boldsymbol{z}_1$ 一道做关于 $\boldsymbol{Y}$ 的二元线性回归.得到这些变量的 t-value 如表 2.

表 2　二元线性回归模型中自变量 z_j^2 的 t-value

变　量	z_1^2	z_2^2
t-value	0.533	78.127

在二元回归中,仍然取 t-检验的阈值为 $t_{0.025}(50-2-1)\approx 1.96$.在表 2 中选择 t-value 绝对值最大的 $\boldsymbol{z}_2^2$ 作为 $\boldsymbol{z}_2$,即 $\boldsymbol{z}_2=\boldsymbol{x}_2-\frac{\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1$.然后,将 $\boldsymbol{x}_1$ 与 $\boldsymbol{z}_1, \boldsymbol{z}_2$ 做 Gram-Schmidt 正交变换,将得到的变量与 $\boldsymbol{z}_1, \boldsymbol{z}_2$ 一道做关于 $\boldsymbol{Y}$ 的三元线性回归.得到的变量的 t-value 为 20.752.

在三元线性回归中,t-检验的阈值为 $t_{0.025}(50-3-1)\approx 1.96$.$\boldsymbol{z}_3^3$ 通过检验,被选入模

型，即 $\boldsymbol{z}_3=\boldsymbol{x}_1-\frac{\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1-\frac{\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2}\boldsymbol{z}_2$.

根据上面的计算过程，还可以得到 Gram-Schmidt 正交变换后的综合变量与原变量之间的关系为

$$\begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{z}_2\\ \boldsymbol{z}_3\end{bmatrix}=\begin{bmatrix}0 & 0 & 1\\ 0 & 1 & 0.1882\\ 1 & 0.0031 & -0.8392\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \boldsymbol{x}_3\end{bmatrix}$$

3）回归分析

使用 Gram-Schmidt 变换后的综合变量（$\boldsymbol{z}_1$，$\boldsymbol{z}_2$，$\boldsymbol{z}_3$）对因变量 $\boldsymbol{Y}$ 进行回归，得到的回归方程为

$$\hat{\boldsymbol{Y}}=0.997\boldsymbol{z}_1+0.083\boldsymbol{z}_2+0.012\boldsymbol{z}_3$$

该方程的检验结果见表 3.

表 3 基于 Gram-Schmidt 过程的回归分析结果

参数	R^2	F 检验值	α_1 的 t 检验值	α_2 的 t 检验值	α_3 的 t 检验值
分析结果	0.995	148306.4	3059.312	248.806	20.752

依据 Gram-Schmidt 过程的反变换，$\boldsymbol{Y}$ 与原始自变量集合之间的回归方程可以表示为

$$\hat{y}_i=-0.337+0.983x_{1i}+1.015x_{1i}^2+1.001x_{1i}^3$$

回归方程系数的 95% 的置信区间见表 4.

表 4 回归方程系数的 95% 的置信区间

回归系数	β_1	β_2	β_3
置信区间	[0.901,1.147]	[1.003,1.015]	[0.999,1.001]

从仿真结果可以看到，由于原始自变量集合（$\boldsymbol{x}_1$，$\boldsymbol{x}_2$，$\boldsymbol{x}_3$）与因变量的相关程度都较高，不存在冗余变量信息，所以所有解释变量都被选入多项式模型.

回归方程的因变量的拟和值与原始因变量值的关系由图 1 表示.

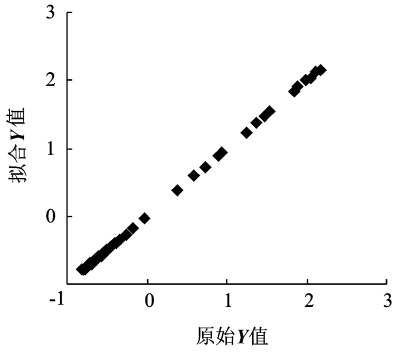

图 1 因变量拟合值和原始值的关系

从图 1 可以看出，在仿真实验中，采用基于 Gram-Schmidt 过程的多项式回归建模方法得到的模型很好地拟和了原仿真模型.

3 结束语

本文提出一种基于 Gram-Schmidt 过程的多项式回归分析方法. 从该方法的计算过程来看，由于通过 Gram-Schmidt 过程得到的新综合变量之间是相互直交的，因此有效克服了自变量集合多重共线对多项式回归建模的不良影响，从而提高了建模的准确性和可靠性.

这种方法得到的综合变量不但较好的代表了原始自变量集合，而且保证了被选入模型的综合变量与因变量相关程度也最大. 在自变量集合中，如果存在与因变量相关程度较低或对因变量无解释作用的变量的情况下，该方法也可以最大限度地提取与因变量相关的自变量信息，同时排除自变量集合中的冗余信息. 此外，该模型除了利用 Gram-Schmidt 过程进行信息提取以外，其他建模步骤都和普通最小二乘方法一样，因此更容易被熟悉 OLS 的应用人员接受，便于推广应用.

参考文献（References）

[1] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method[M]. Heidelberg: Springer-Verlag, 1983

[2] Tenenhaus M. L'approche PLS[M]. Revue de Statistique Appliquée, 1998

[3] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京：国防工业出版社，1999：93－107
Wang Huiwen. Partial least-squaresregression-method and applications[M]. Beijing: National Defense Industry Press, 1997, 93－107(in Chinese)

[4] 付凌晖，王惠文. 多项式回归的建模方法比较研究[J]. 数理统计与管理，2004，23(1)：48－52
Fu Linghui, Wang Huiwen. A comparative reseach of polynomial regression modelling methods[J]. Application of Statistics and Management, 2004, 23(1): 48－52(in Chinese)

[5] Jain S K, Gunawarddena A D. Linear algebra: an interactive approach[M]. Bejing: China Machine Press, 2003

[6] 内特，沃塞曼，库特纳. 应用线性回归模型[M]. 张勇，王国明，赵秀珍译. 北京：中国统计出版社，1990
Neter, Wasserman, Kutner. Applied linear regression model [M]. Translated by Zhang Yong, Wang Guoming, Zhao Xiuzhen. Beijing: China Statistics Press, 1990(in Chinese)

[7] 王惠文，陈梅玲，Gilbert Saporta. Gram-Schmidt 回归及在刀具磨损预报中的应用[J]. 北京航空航天大学学报，2008，34(6)：729－733
Wang Huiwen, Chen Meiling, Gilbert Saporta. Gram-Schmidt regression and application in cutting tool abrasion prediction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(6): 729－733(in Chinese)

（责任编辑：吴 琼）