偏最小二乘建模在 R 软件中的实现及实证分析

齐 琛 方秋莲
(中南大学数学与统计学院,长沙,410075)

摘 要 通过介绍偏最小二乘(PLS)的建模和显著性检验原理,解决了小样本多变量且变量间存在多重共线性的回归问题,建立了多变量对多变量的回归模型,并使用 R 软件(版本为 R i386 2.15.1)实现了 PLS 建模;最后基于葡萄和葡萄酒理化指标数据进行了实证分析.

关键词 偏最小二乘 R 语言 jackknife 方差 显著性检验

Partial Least Squares Modelling with R Software and Empirical Analysis

Qi Chen Fang Qiulian
(School of Mathematics and Statistics, Central South University, Changsha 410075, China)

Abstract This paper introduces the Partial Least Squares(PLS) method and its significance test principle for modelling regression problems in which sample size is small and there is multicollinearity among observable variables, and furthermore, illustrates how to set up PLS models with the R software. An example to model the relation of the physicochemical indexes between grapes and wine is given to demonstrate the modelling process.

Key words PLS R Language Jackknife Variance Significance Test

1 引言

人们总能通过普通最小二乘回归进行数据的分析和预测;然而,当选取的解释变量过多而样本很少时,无法使用普通最小二乘法进行建模. Herman Wold 在 20 世纪 70 年代的经济学研究中提出偏最小二乘法(PLS),它能够在小样本的情况下实行多变量对多变量的回归建模. 1998 年王惠文[1]对偏最小二乘回归方法及其应用进行了详尽的解说,在 2005 年与吴载斌、孟洁[2]一起,对最小二乘法再进行了进一步的扩展,并提出回归系数的显著性检验方法;高惠璇[3]用具体例子对最小二乘回归、主成分回归和偏最小二乘回归进行比较分析,并使用 SAS 软件实现了 PLS 建模. R 软件是一款具有强大统计分析功能的开源软件,利用 R 软件进行偏

湖南省自然科学基金资助项目(编号:12JJ5002)
收稿日期:2013 年 4 月 28 日

最小二乘回归建模,可以得到理想的模型,并能够对回归系数进行显著性检验,帮助人们发现变量的主要影响因素,进行下一步数据探索分析. 本文介绍偏最小二乘的回归原理,以及如何使用 R 软件中的 pls 包[4] 实现 PLS 建模,并尝试用 PLS 对葡萄和葡萄酒理化指标数据进行实证分析.

2 偏最小二乘回归及显著性检验原理

2.1 偏最小二乘回归原理

设有 q 个因变量$\{y_1,y_2,\cdots,y_q\}$ 和 p 个自变量 $\{x_1,x_2,\cdots,x_p\}$,观测了 n 个样本点,由此构成了自变量与因变量的数据表 $X_{n\times p}$ 和 $Y_{n\times q}$. 偏最小二乘回归分别在 X 与 Y 中提取出成分 t_1 与 u_1(即: t_1、u_1 分别是 $x_1,x_2,\cdots,x_p$、$y_1,y_2,\cdots,y_q$ 的线性组合). 在提取这两个成分时,需满足以下两个条件:

(1) t_1 与 u_1 应尽可能大地携带它们各自数据表中的变异信息;

(2) t_1 与 u_1 的相关程度达到最大.

这两个要求表明,t_1 与 u_1 应尽可能好地代表数据 X 与 Y,同时自变量的成分 t_1 对因变量的成分 u_1 又有较强的解释能力.

在第一个成分 t_1 与 u_1 被提取后,偏最小二乘回归分别实施 X 对 t_1 的回归及 Y 对 t_1 的回归. 如果回归方程已经达到满意的精确度,则算法终止;否则将利用 X 与 Y 分别被 t_1 解释后的残余信息进行第二轮的成分提取. 如此往复,直到能达到一个较满意的精度为止. 若最终对 X 提取了 m 个成分 $t_1,\cdots t_m$,偏最小二乘回归将实行 y_k 对 $t_1,\cdots,t_m$ 的回归,然后再表达成 y_k 关于原变量 $x_1,x_2,\cdots,x_p$ 的回归方程($k=1,2,\cdots,q$).

2.2 偏最小二乘回归步骤

Step 1: 先将数据进行标准化处理,得到标准化后的 X 与 Y 矩阵,记第 i 对成分为 t_i 与 u_i,并且 $t_i=Xw_i$, $u_i=Yc_i$. 于是对第一对成分的提取,即求解以下优化问题:

$$\max \ <Xw_1,Yc_1>$$

$$s.t.\begin{cases} w_1'w_1=1 \\ c_1'c_1=1 \end{cases}$$

只需求出矩阵 $M=X'YY'X$ 的特征值与特征向量,其最大特征值 λ_1 对应的特征向量即为所求的 w_1,目标函数值等于 $\sqrt{\lambda_1}$.

Step 2: 分别做 $y_1,y_2,\cdots,y_q$ 和 $x_1,x_2,\cdots,x_p$ 对 t_1 的回归

$$\begin{cases} X=t_1p_1'+E_1 \\ Y=t_1q_1'+F_1 \end{cases} \tag{2.1}$$

其中，回归系数向量 $p_1' = (p_{11},\cdots,p_{1p})$，$q_1' = (q_{11},\cdots,q_{1p})$；$E_1$ 与 F_1 是回归方程的残差阵，p_1' 和 q_1' 可由简单最小二乘法的原则求得.

Step 3：用 E_1 与 F_1 代替 X 与 Y 进行前两个步骤求得第二对成分，依次循环. 设 X 的秩为 $r(r \leqslant p)$，则存在 r 个主成分，使得

$$\begin{cases} X = t_1 p_1' + \cdots t_r p_r' + E_r \\ Y = t_1 q_1' + \cdots t_r q_r' + F_r \end{cases} \tag{2.2}$$

再把 $t_i = Xw_i$ 带入(2.2)即可得到 $y_1, y_2, \cdots, y_q$ 分别对 $x_1, x_2, \cdots, x_p$ 的回归方程

$$y_i = \beta_{j1} x_1 + \cdots \beta_{jp} x_p \quad (i = 1,2,\cdots,q) \tag{2.3}$$

2.3 根据交叉验证结果选择模型的成分个数

若选取成分的个数过多，会很容易出现过度拟合的问题，因此我们需要一个有效的原则来确定成分的个数. 采用类似抽样测试的工作方式，把所有样本点分成两部分：第一部分用来重新拟合一个偏最小二乘模型，第二部分的样本点作为测试数据；带入拟合模型中求得预测值误差平方和 $PRESS = \sum (y_i - \hat{y}_i)^2$. 再以这种方式重复 g 次，直到所有的样本都被预测了一次，最后把每个样本的预测误差平方和加总，称为 PRESS.

$$PRESS = \sum_{i=1}^{g} PRESS_j \tag{2.4}$$

常见的交叉验证方法有“留一验证”，“K 折交叉验证”，“Holdout 验证”等方法，选取一种方法分别求出第 1 ~ r 个成分对应的 PRESS 值，取 PRESS 最小的或者 PRESS 几乎不再变化的成分个数作为最终模型选取的成分个数 m.

2.4 回归系数的显著性检验

在 R 软件的 pls 包(package pls)中提供了函数 jack.test，用来检验回归系数的显著性. 由于偏最小二乘法不同于一般最小二乘法，它的回归系数方差无法得到准确的无偏估计. Miller 提出了 Quenouille – Tukey jackknife 方法来估计回归系数的方差：先抽出 g 个样本子集，然后用只去除一个子集的样本做一次偏最小二乘的回归系数估计，记去除第 i 个样本子集对应的回归系数为 β_{-i}，则 jackknife 方差估计为

$$\mathrm{var}(\beta_i) = \frac{g-1}{g} \sum_{t=1}^{g} (\beta_{-i} - \bar{\beta})^2 \tag{2.5}$$

其中 $\bar{\beta}$ 是 β_{-i} 的均值，最常见的重抽样法是留一 jackknife 法，即每次选一个样本点，于是共有 n 个样本子集(即 $g = n$). 在估计出方差后，类似于普通最小二乘法，求出 β_i 对应的 t 统计量，再进行均值是否为零的假设检验. 由于偏最小二乘回归系数确定的复杂公式，我们至今无法确定准确的 t 分布，在 R 软件中默认为服从自由度为 m 的 t 分布(m 为建模使用成分的个数).

3 在 R 软件中的实现

首先需要在加载 R 的程序包 pls; pls 包是由 Bjørn - Helge Mevik,Ron Wehrens 和 Kristian Hovde Liland 创建,专门用来做偏最小二乘回归的程序包. 代码如下:

```
>library("pls",lib.loc="C:/Program Files/R/R-2.15.1/library")
```

再导入自变量和因变量的样本数据,并且使用 scale() 函数将数据进行标准化消除量纲的影响. 记标准化之后的自变量为 X,因变量为 Y,进行 PLS 回归的代码如下:

```
>pls1<-plsr(Y~X,validation="LOO",jackknife=TRUE)
#进行偏最小二乘回归,模型存为对象 pls1
> summary(pls1,what="all") #显示回归结果(包括 PRESS 与变异解释度)
```

其中,validation="LOO" 表示使用留一交叉验证计算 PRESS,jackknife=TRUE 表示使用 jackknife 法估计回归系数方差(为后面的显著性检验做准备). 在没给定成分个数的情况下,会默认使用所有的主成分进行回归,因此我们需要在选择的成分个数尽可能小的前提下,选择使 PRESS 最小或几乎不变的成分个数. 假设选定了成分个数为 m,重新进行回归,并对回归系数假设检验,代码如下:

```
>pls2<-plsr(Y~X,ncomp=m,validation="LOO",jackknife=TRUE)
#"ncomp=m"表示模型成分个数为 m
> jack.test(pls2)
```

另外还可以使用 coef() 函数得到回归系数,scores() 得到得分矩阵,loadings() 得到载荷矩阵,predict() 得到对应样本的预测值,以及 plot() 函数将结果以图的形式展现.

4 基于葡萄和葡萄酒理化指标的 PLS 实证分析

葡萄酒是由葡萄精细酿造而成,因此二者的理化指标之间必然存在一定的联系. 本文采用中国 2012 年数学建模大赛 A 题中提供的数据,对红葡萄酒的理化指标与酿酒红葡萄的理化指标进行最小二乘法建模分析(以下的葡萄酒与酿酒葡萄均指红葡萄酒与酿酒红葡萄).

4.1 建模过程

Step 1: 导入数据,并进行数据的标准化:

```
> G1<-read.csv("K:\\WORK\\论文\\R\\grape.csv")
> W1<-read.csv("K:\\WORK\\论文\\R\\wine.csv")
> X<-scale(G1)
> Y<-scale(W1)
```

得到的自变量 X 是 27×59 的矩阵，$X_1 \sim X_{59}$ 依次代表酿酒葡萄理化指标如下(在此只列出部分名称，具体参见“2012 年数学建模大赛 A 题附件 2”)：

氨基酸总量、天门冬氨酸、苏氨酸、丝氨酸、谷氨酸…果皮颜色 H、果皮颜色 C.

得到的因变量 Y 是 27×15 的矩阵，$X_1 \sim Y_{15}$ 依次代表葡萄酒理化指标如下:

花色苷、单宁、总酚、酒总黄酮…色泽 b* 、色泽 H、色泽 C.

Step 2: 进行初步偏最小二乘回归:

```
>pls1 <- plsr(Y~X, ncomp=10, validation = "LOO", jackknife = TRUE)
>summary(pls1,what="all")
```

#注: R 中默认最多只能显示 25 个主成分对应的各项结果，此处已达到最大个数 25)

选取部分结果如表 1 所示:

表 1 初步模型拟合结果(部分)

VALIDATION: RMSEP					
Cross-validated using 27 leave-one-out segments.					
	Response: Y1				
	(Intercept)	1 comps	2 comps	3 comps	4 comps...
CV	1.019	0.7519	0.7602	0.7106	0.6757
adjCV	1.019	0.7490	0.7540	0.7058	0.6759...
TRAINING: % variance explained					
	1 comps	2 comps	3 comps	4 comps	
X	16.17299	24.845	35.872	46.10	
Y1	65.38198	74.147	81.728	82.46	
Y2	83.15187	83.729	84.343	84.38...	

其中 CV 即为不同主成分个数对应的 PRESS，adjcv 为调整后的 PRESS，“TRAINING: % variance explained”一栏为主成分对各变量的累积贡献率.

由结果可知，主成分个数为 3 个时，模型在经过留一交叉验证法后求得的 PRESS 总和最小，随着成分个数的增加，PRESS 值也没有太大改变，并且 3 个成分对各个因变量的累积贡献率也基本达到了 85%，因此定下回归的成分个数 $m=3$.

Step 3: 根据成分数 $m=3$，建立最终模型:

```
>pls2 <- plsr(Y~X,ncomp=3,validation = "LOO",jackknife = TRUE)
>coef(pls2) #得到回归系数
```

得到回归系数后，便能写出各因变量对所有解释变量的回归方程，下面写出 Y_1 对各解释变量的回归方程(由于变量太多，因此中间有省略)：

$$Y_1 = -0.0066X_1 + 0.046X_2 + 0.0087X_3 - 0.041X_4 + 0.0057X_5 - 0.012X_6$$
$$+ \cdots - 0.036X_{55} - 0.018X_{56} - 0.00049X_{57} + 0.012X_{58} - 0.017X_{59}$$

4.2 模型拟合效果分析

使用 validationplot( ) 函数可以画出 PLS 模型在不同主成分数下对应的 RMSEP(由留一交叉验证法算得的均方预测误差根) ,对初始模型的结果进行画图,如图 1 所示:

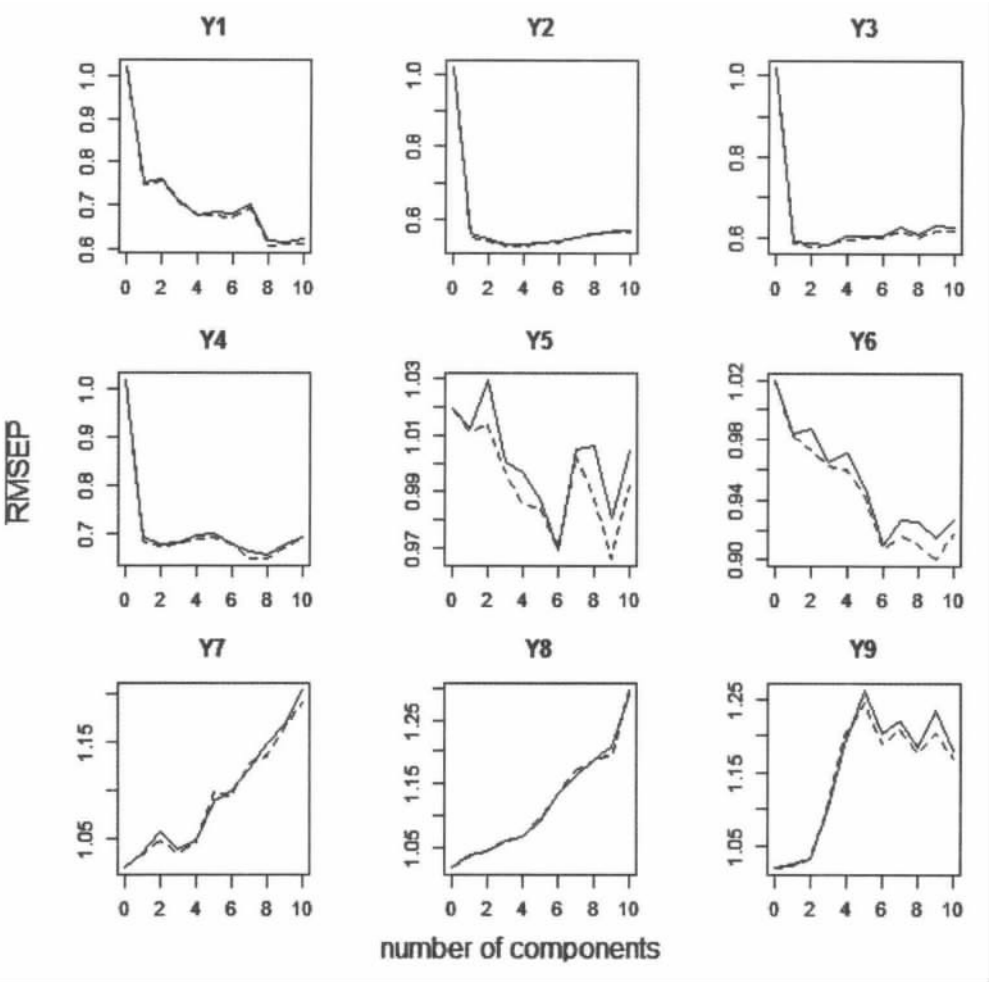

图 1 不同成分数对应的均方误差图(此处仅截取前 9 个因变量)

图 1 中纵坐标“RMESP”表示均方预测误差根,横坐标为不同模型的成分个数; 由图 1 可知大部分因变量在成分数为 3 时对应的均方误差根最小,证明选择 3 个成分参与建模是正确的.

使用 predplot(pls2) 函数画出最终模型的预测效果图,如图 2 所示.

图 2 中纵坐标为各因变量的预测值,横坐标为各因变量的实际测量值; 散点集中分布在主

对角线上则说明预测效果很好. 图 2 中 15 张预测图(对应 15 个因变量) 的散点大致都分布在对角线上,说明最终模型的拟合效果较好. 然而,Y8 对应的散点图几乎是一条垂直的线,预测很糟糕; 查阅原始数据得知第 26 个样本点的 Y8 原始值是 1.6239,远大于其他只有 0.02 左右的数据,因此 26 号样本为模型的强影响点,它使得对 Y8 的拟合效果很差; 可以考虑将其剔除重新进行预测. 对于其他的变量,可以直观的看出效果是不错的.

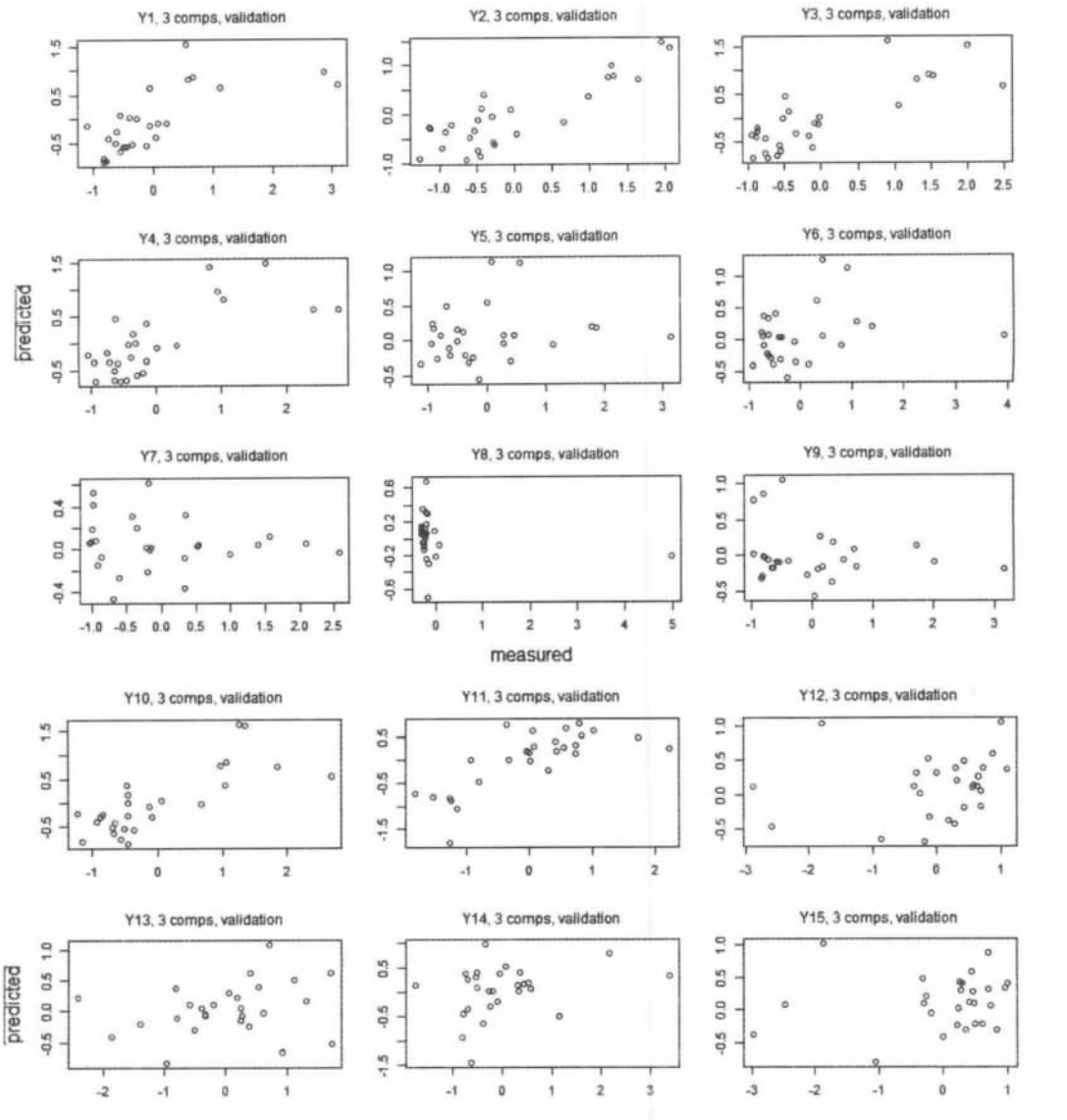

图 2 预测效果图

4.3 回归参数的显著性检验

使用 jack. test( ) 函数进行检验,挑出与通过检验的回归系数对应的自变量,整理结果如表

2 所示(由于篇幅限制,只列出前 2 个因变量对应的结果) :

表 2 回归系数显著性统计表

葡萄酒理化指标 Y	酿酒葡萄理化指标 X	系数符号	显著程度	葡萄酒理化指标 Y	酿酒葡萄理化指标 X	系数符号	显著程度
	X19	+	.		X28	+	* * *
	X21	+	.		X29	+	* * *
	X27	+	* *		X30	+	*
Y1	X28	+	*		X34	+	* *
	X29	+	* *		X36	+	*
	X37	+	*		X38	+	*
	X40	+	*	Y2	X40	+	* *
	X1	+	*		X41	+	*
	X6	+	.		X45	+	*
	X11	+	.		X49	+	* *
Y2	X18	−	.		X51	−	.
	X21	+	* * *		X52	+	*
	X26	+	*		X53	+	*
	X27	+	* * *		X55	+	*

注: 显著性符号表示 ‘* * *’极其显著,‘* *’非常显著,‘* ’很显著,‘.’较显著.

通过显著性检验可以知道各因变量(葡萄酒理化指标) 受哪些自变量(酿酒葡萄理化指标) 的影响较大,及其受影响程度. 由表 2 可知,对因变量 Y1 花色苷而言,对其有显著影响的自变量有 X19 蛋白质、X21 花色苷、X27 DPPH 自由基、X28 总酚、X29 单宁、X37 杨梅黄酮、X40 异鼠李素,并且均是对其有正向的影响. 酿酒葡萄的花色苷量对葡萄酒的花色苷值有正向影响是显然成立的,此处也通过了显著性检验,因此可初步判断此模型与实际相符,对于其他影响显著的变量便是给了我们一个探索点,可以通过别的方法深入探讨. 同理于其他 14 个因变量,显著性检验可以让我们初步了解到各因变量的受影响因素.

5 总结

偏最小二乘回归能够解决许多以往用普通多元线性回归不能解决的问题,在解释变量个数大于样本个数的情况下也能建立出很有效的模型. 本文主要介绍的是使用 R 软件的 pls 包进行偏最小二乘建模,成功地对葡萄酒和酿酒葡萄的理化指标进行了偏最小二乘的建模. 然而由于 PLS 公式的复杂性,对于回归系数的方差估计至今没有特别完善的方法,因此我们需要辩证地看待 jack. test 显著性检验的结果,以它作为一个研究的参考,再进一步进行深入分析.

参考文献

[1] 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.

[2] 王惠文. 吴载斌. 孟洁. 偏最小二乘回归的线性与非线性方法 [M]. 北京: 国防工业出版社, 2006.

[3] 高惠璇. 两个多重相关变量组的统计分析(3)(偏最小二乘与 PLS 过程) [J]. 数理统计与管理, 2001,21(2):58 - 64.

[4] Bjørn - Helge Mevik. Ron Wehrens. Kristian Hovde Liland. Partial Least Squares and Principal Component regression [M/OL]. http://mevik.net/work/software/pls.html,2011 - 11 - 27.

[5] Randall D. Tobias. An Introduction to Partial Least Squares Regression [J/OL]. http://support.sas.com/techsup/tech note/ts509.pdf.

[6] B. Eforn. C. Stein. Tthe Tackknife Estimate Of Variance [J]. The Annals of Statistics, 1981.

[7] 蒋红卫. 夏结来. 偏最小二乘回归及其应用 [J]. 第四军医大学学报,2003,24(3).

[8] 王怀亮. 交叉验证在数据建模模型选择中的应用 [J]. 商业经济,2011,11.

[9] Sophia Yuditskaya. An Overview of Methods in Linear Least - Squares Regression [J]. Pattern Recognition and Analysis,2010,11.

[10] Bootstrap and Jackknife Estimation of Sampling Distributions [M/OL]. http://www.stat.washington.edu/jaw/COURSES/580s/581/LECTNOTES/ch8.pdf.