硕士研究生学位论文

新疆大学

论文题目(中文)：偏最小二乘回归算法改进及应用

论文题目(外文)：A Modified Algorithm of Partial Least Squares Regression and Its Application

研究生姓名：丁磊

学科、专业：应用数学

学位类别：理学硕士

研究方向：概率统计及其应用

导师姓名 职称：胡锡健 副教授

论文答辩日期 2007年 6月 3日

学位授予日期 2007年 7月 日

摘　　要

偏最小二乘回归(Partial Least Squares Regression,PLSR) 是一种先进的多元统计分析方法, 于1964 由瑞典计量经济学家 Herman Wold 等人首次提出, 主要用来解决多元回归分析中的变量多重相关性或变量多于样本点等实际情况的问题. 由于它集多元线性回归分析, 主成份分析和典型相关分析的基本功能为一体, 因此在国外被誉为第二代多元统计分析方法, 该方法目前已广泛应用于化学计量, 工业设计, 计量经济学等各个领域.

本文的主要内容可以概述如下:

第一部分主要阐述了偏最小二乘方法的历史和现状, 并对偏最小二乘回归近期的热点问题进行了总结.

第二部分详细介绍了偏最小二乘回归的基本思想, 数学原理和单因变量偏最小二乘的算法推导, 并利用该方法对防治沙尘暴研究进行了偏最小二乘回归建模分析, 从中发现抑制沙尘暴的根本办法不是治理沙漠, 而是要控制土地沙漠化和抑制裸露农田起尘.

第三部分在回归分析中经常存在自变量过多并且之间存在多重相关性现象, 为了寻找对因变量有重要影响的自变量, 本文提出了偏最小二乘向前逐步回归法, 并对该方法进行了详细的理论推导. 同时, 运用SAS软件, 利用该方法对化工领域的典型数据进行建模分析, 结果发现, 该方法易于操作, 具有一定的实用性. 另外, 在多指标体系中建立综合评价指数时, 往往会遇到指标变量集合间存在多重相关性问题, 而传统的主成份分析并不能解决该问题, 针对这种情况, 本文采用PLS路径分析的思想, 构建综合评价指标, 对中国西部城市综合评价进行实证分析.

第四部分针对偏最小二乘回归无法对未来值进行预测的问题, 采用了偏最小二乘时间序列预测模型. 一方面, 针对因子间的多重相关性现象, 采用偏最小二乘回归建模, 从而明确各因子对因变量的影响程度; 另一方面, 根据构成因子数据的特点, 利用 AR(p) 模型对各因子未来值进行预测, 然后将其代入已建成的偏最小二乘回归方程, 从而实现对因变量未来值进行预测. 本文利用该方法对烟台市年生活用水量进行了实证分析.

关键词： 偏最小二乘回归; 多重共线性; 变量选择; 算法改进; 潜变量; 主成分.

Abstract

Partial least squares regression (PLSR) is a sophisticated multivariate analysis method ,which was first raised by economist Herman Wold and others in Sweden in 1964.It is Mainly used to solve multivariate regression analysis of multiple variables relevance or variable sample points more than the actual cases. As It combines multiple linear regression analysis, principal components analysis and canonical correlation analysis of the basic functions of integration. Therefore it is known as the second generation of multivariate statistical analysis. The method has been widely used in chemical measurement, industrial design, econometrics, and other areas.

The main content of this paper can be summarized as follows :

The first part deals mainly with the partial least squares method of history and current situation as well as the recent hot issues summarized.

The second part details the partial least squares regression to the basic idea, mathematical principles and single dependent variable PLS algorithm is derived. And the method is used to control for the study of sandstorms, finding inhibition from the fundamental approach is not treating the desert, but to control and curb the desertification of land bare farmland dust.

The third part of the regression analysis, there is often too many variables exist between multiple and related phenomena, In order to find the variables are important among all variables, this paper presents a forward-stepwise-partial least squares, and the method carried out a detailed theoretical derivation. Meanwhile,by the use of SAS software, the results showed that this method is easy to operate with a certain practicality. In addition, multiple indicator system to establish a comprehensive evaluation index, often encountered indicator variables exist between multiple sets related issues, and the traditional principal component analysis does not solve the problem, according to this situation, this

paper use PLS Path Analysis and Construct comprehensive evaluation indicators for China's western city's comprehensive evaluation of empirical analysis. The forth part , Partial least squares regression is unable to predict the future value of the issue, therefore we presented the PLS and time series forecasting model. On one hand, address the multiple factors related phenomenon, using partial least squares regression modeling, so clearly the result of variable factors on the extent of the impact. On the other hand, according to the data form factor characteristics, using AR (p) model for the future-value forecast, then their generation has been built into the partial least squares regression equation which forecasts the future value of the dependent variable. We used the method to analyze the water consumption of Yantai City.

Keywords: Partial Least Square Regression, multivariate collinearity, variable selection, latent variable, principle component.

学位论文独创性声明

本人声明，所呈交的学位论文系本人在导师指导下独立完成的研究成果. 文中依法引用他人的成果，均已做出明确标注或得到许可. 论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果. 与我一同工作的同志对本研究所做的任何贡献均已在论文中做出了明确的说明并表示谢意.

本人如违反上述声明，愿意承担由此引发的一切责任和后果.

论文作者签名：丁磊　　日期：2007 年 5 月 28 日

学位论文知识产权权属声明

本人的学位论文是在学期间在导师指导下完成的，知识产权归属学校. 学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利. 本人离校后发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为新疆大学.

本学位论文属于：

保密 □，在　　年解密后适用于本声明.

不保密 □.

(请在以上方框内打"√")

论文作者签字：丁磊　　日期：2007 年 5 月 28 日

导师签字：胡锡虎　　日期：2007 年　月　日

第一章 综 述

§1.1 引言

在化学研究中, 经常需要利用一些可以控制(或容易测量)的变量(解释变量)去解释, 控制或预测另外一些变量(反应变量), 一个常用的统计建模方法是一般最小二乘(OLS)的多元线性回归. 然而, 只有当解释变量(1)数目较少, (2)无多重共线性, (3)各解释变量与反应变量之间的关系易于解释时, 多元回归的一般最小二乘估计才具有优良的理论特性: 比如最佳线性无偏估计（BLUE）, 才能较好地拟合数据, 并能对结果给予比较合理的解释. 若数据不能完全满足以上三个条件, 则多元线性回归的一般最小二乘估计就会失效.

为了处理这种违背假设的数据, 统计学家探索性地对一般最小二乘估计进行了多种改进. 为了克服多重共线性影响, 发展了一系列有偏估计方法: 岭估计, 压缩估计, 主成份估计及特征根估计等, 或是运用逐步回归等方法去掉一部分解释变量. 但是对于观察个体数较少, 甚至少于解释变量的情况, 以上方法均不适用, 并且这些方法仍然存在着各种各样的问题或不足: 要么解释性不够好, 要么模型拟合精度不够高, 要么预测精度不够理想等[1,2,3,].

对此, 欧洲经济计量学家, 如 Herman Wold 等, 提出并发展起来的一种新的统计方法偏最小二乘 (Partial Least Squares,简写PLS)[4,5]. 它是一般最小二乘的一种拓展, 能够克服第一代多元统计分析方法的一些不足, 但当时偏最小二乘回归在统计理论上还有很多的问题没有完全解决, 在应用领域也没有取得大的进展. 因此没有引起统计学界和应用领域研究人员的足够重视. 直至上世纪80年代, 计量化学研究者首先将偏最小二乘回归成功地运用于计量化学, 而后工业设计工作者应用该方法同样获得巨大成功, 才真正引起了各方面极大的关注[6,7]. 由此偏最小二乘回归的统计理论和算法研究取得了极大的发展, 其应用也迅速扩展到了其他领域, 如管理科学, 教育评测学, 药理学[8]. 到了上世纪 80 年代末至90 年代初, 非线性迭代偏最小二乘（NIPLS）形成多种算法变种, 最早由Heman Wold提出的NIPLS算法发展出迭代法, 特征根法, 奇异值分解法等诸多的算法, 它们极大的丰富了偏最小二乘算法[8]. 随着对偏最小二乘回归理

论, 算法和性质的进一步深入研究, De Jong于 1993 年提出了一种与NIPLS不完全相同的算法, 即简单偏最小二乘 (Sample Partial Least Squares Regression)[9], 同样实现了偏最小二乘回归的基本思想.

1996 年 10 月, 在法国高等商业教育组织机构HFCCISIA-CERESTA的组织和资助下, 有关偏最小二乘回归方法, 理论和应用的第一次国际学术专题研讨会在法国巴黎召开, 来自世界各地的著名偏最小二乘专家分析和介绍了他们各自关于 PLS 方法的最新进展及研究成果, 以及在计量化学, 工业设计, 市场分析和金融分析等领域的应用. 这次会议极大的激起了统计学家及相关应用领域专家对偏最小二乘回归的研究热情, 促进了偏最小二乘回归理论和算法的进一步深入发展. 现在, PLS 方法的国际研讨会每两年举办一次, 2005年 9 月在西班牙巴赛罗纳举办了 PLS 的第五次会议. 在国外, 有关 PLS 方法的理论, 性质, 算法及典型应用等方面的前沿研究成果一般发表在Journal of chemometries 和 chemamatrics and Zntelligent Laboratory Systems 等专业期刊上.

在我国, 大多数研究者仅是借用外国现成的软件进行PLS的应用研究, 只有部分研究者对 PLS 方法进行了较为深入地研究, 如王惠文编写了 ≪偏最小二乘回归方法及其应用≫ 对PLS进行了详细而深入地阐述[10,11]; 许青松和梁益曾等人提出了广义 PLS 算法, 并研究了 Monte Carlo 交叉验证法用于 OLS 成分数目的确定: 吴喜之提出了一种改进的偏最小二乘回归法 [12,13], 进一步推动了偏最小二乘的理论研究与发展.

目前, 在对偏最小二乘回归理论和应用研究中, 偏最小二乘主要有两个热点领域: 一个是非线性偏最小二乘回归模型的研究 [14,15,16], 另一个是关于PLS路径分析的研究 [17,18,19].

§1.2 本文的主要工作

文章分为三个部分:

第一部分在文献阅读的基础上, 叙述了偏最小二乘回归的理论的历史现状, 并对PLS近期的热点问题进行了阐述.

第二部分详细阐述偏最小二乘回归的基本思想, 数学原理和基本算法, 并将偏最小二乘建模方法运用到沙尘暴防治研究中.

第三部分针对回归分析中自变量选择问题, 提出了偏最小二乘向前逐步回

归选择变量法, 本文将该方法与逐步回归方法进行了对比分析. 同时, 本文对偏最小二乘的热点领域PLS路径模型进行了研究, 本文利用PLS路径模型, 通过构建多指标系数评估指数, 对中国西部省区进行了综合分析.

第四部分针对偏最小二乘回归无法对未来值进行预测的问题, 提出了偏最小二乘时间序列预测模型. 本文利用该方法对烟台市年生活用水量进行了实证分析.

§1.3 本文的创新之处

1. 在文中的第三部分, 提出了偏最小二乘向前逐步回归选择变量法, 并对该方法进行了详细的理论推导. 该方法主要运用一元或多元线性回归, 借助对回归系数的显著性检验来寻找对因变量有显著作用的变量, 然后, 利用选择的变量构建成分进行回归建模. 成分提取停止的原则除了运用交叉有效性外, 也可选择使用本文的方法: 直到所有变量对因变量都不显著时停止构建新成分. . . .

2. 在文中的第四部分将偏最小二乘回归模型和时间序列预测模型相结合. 该模型的优势在于, 一方面, 避免了传统时间序列建模方法只能对单一序列进行样本预测, 但对于由众多因素影响的序列无法确定各影响因素的影响程度. 针对针对这种现象, 采用偏最小二乘回归建模, 从而明确各因子对因变量的影响程度. 另一方面, 由于各因子均为带有明显趋势项的非平稳时间序列, 按照传统的时间序列建模方法, 经过多次尝试均无法建立满意的AR(p)模型, 因此对这些序列采取下列步骤: 1 将序列进行平稳化处理, 剔除时间趋势项. 2 对提取的残差序列通过子观察自相关和偏自相关图确定模型的阶数及形式. 3 对原序列进行方程检验, 形成预测模型. 这样可以得到更为合理的预测结果. 最后, 将各因子的预测值代入偏最小二乘回归方程, 从而实现了偏最小二乘回归对未来值的预测.

第二章　偏最小二乘的理论与方法

偏最小二乘回归是一种先进的多元统计分析方法, 能够有效的解决变量间的多重相关性问题, 本部分主要介绍了偏最小二乘的基本思想, 基本原理和单因变量偏最小二乘的算法推导. 并在此基础上对防治沙尘暴进行了偏最小二乘回归建模分析.

2.1 基本思想

偏最小二乘回归的目的是在解释变量空间里寻找某些线性组合, 以能更好地解释反应变量的变异信息. 假定所有解释变量与反应变量均有关联. 偏最小二乘的基本思想如图 2.1 所示:

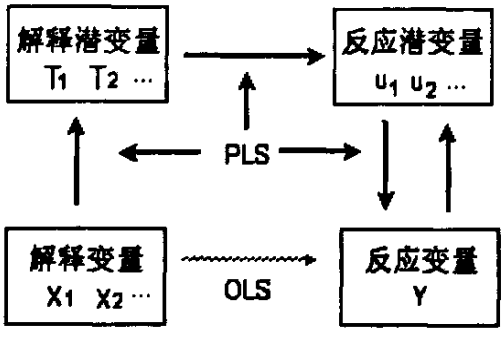

图 2.1 PLS 建模示意图

一般的最小二乘多重回归直接建立反应变量关于解释变量的线性回归模型, 反映二者之间的线性关系(图中虚线箭头表示); 而偏最小二乘则是建立解释潜变量关于反应潜变量的线性回归模型, 间接反映解释变量与反应变量之间的关系(图中实线箭头表示).

该方法同时从解释变量与反应变量中提取两组潜变量(也称主成分), 它们分别是解释变量与反应变量的线性组合, 满足以下两个条件: (1) 两组潜变量分别最大限度地承载解释变量或反应变量的变异信息; (2)相互对应的解释潜变量与反映潜变量之间相关性最大化[20].

2.2. 数学原理

设有 q 个因变量 $\{y_1,\cdots,y_q\}$ 和 p 个自变量 $\{x_1,\cdots,x_p\}$, n 个样本点,由此构成的自变量数据表 $X=(x_1,\cdots,x_p)_{n\times p}$, 因变量数据表 $Y=(y_1,\cdots,y_q)_{n\times q}$.

偏最小二乘回归分别在 X 与 Y 中提取出成分 t_1 和 u_1, 在提取这两个成分时, 为了回归分析的需要, 有下列两个要求:

(1) t_1 和 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 和 u_1 的相关程度能够达到最大.

这两个要求表明, t_1 和 u_1 应尽可能好地代表数据表 X 与 Y, 同时自变量的成分 t_1 对因变量 的成分 u_1 又有最强的解释能力. 在第一个成分 t_1 和 u_1 被提取后, 偏最小二乘回归分别实现 X 对 t_1 的回归以及 Y 对 t_1 的回归. 如果回归方程已经达到满意的精度, 则算法终止; 否则, 将利用 X 被 t_1 解释后的残余信息及 Y 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, 直到能达到一个满意的精度为止. 若最终对 X 共提取了 m 个成份 $t_1,\cdots,t_m$, 偏最小二乘回归将通过施行 $y_k, k=1,2,\cdots,q$ 对 $t_1,\cdots,t_m$ 的回归, 然后再表达成 y_k 关于原变量 $x_1,\cdots,x_p$ 的回归方程.

2.3. 单因变量PLS算法的推导

记 $F_0(F_0 \in R^n)$ 是因变量 y 的标准化变量, E_0 是自变量集合X 的标准化矩阵.

按照偏最小二乘回归的第 1 步, 首先从F_0中抽取 1 个成分 $u_1, u_1 = F_0c_1, \|c_1\| = 1$; 从 E_0 中抽取一个成分 $t_1, t_1 = E_0w_1, \|w_1\| = 1$.

由于$F_0(F_0 \in R^n)$只是1个变量, 所以, 是一个标量. 而由于 $\|c_1\| = 1$ 所以, $c_1 = 1$, 即有

$$u_1 = F_0$$

另外, 根据式 $F_0^T E_0 E_0^T F_0 C_1 = \theta_1^2 C_1$, 其中, θ_1 是矩阵 $E_0^T F_0 F_0^T E_0$ 的最大特征值[10]. 所以

$$\theta_1^2 = \|E_0^T F_0\|^2$$

另一方面, 利用 $w_h = \frac{1}{\theta_h} E_{h-1}^T u_h$ [10], 可以求出

$$w_1 = \frac{1}{\theta_1} E_0^T u_1 = \frac{E_0^T F_0}{\|E_0^T F_0\|}$$

因为 E_0 是标准化矩阵, F_0 是标准化变量, 所以, 有

$$E_0^T F_0 = (E_{01}, E_{02}, \cdots, E_{0p})^T F_0 = (r(x_1, y), r(x_2, y), \cdots, r(x_p, y))^T$$

则

$$w_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} r^2(x_j, y)}} \begin{bmatrix} r(x_1, y) \\ r(x_2, y) \\ \vdots \\ r(x_p, y) \end{bmatrix}$$

$$t_1 = E_0 w_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} r^2(x_j, y)}} [r(x_1, y)E_{01} + r(x_2, y)E_{02} + \cdots + r(x_p, y)E_{0P}] \tag{2.3.1}$$

下面, 实施 E_0 在 t_1 上的回归及 F_0 在 t_1 上的回归,即

$$E_0 = t_1 p_1^T + E_1$$

$$F_0 = t_1 r_1 + F_1$$

式中: p_1, r_1 是回归系数(r_1 是标量).

即

$$p_1 = \frac{E_0^T t_1}{\|t_1\|^2}, \quad r_1 = \frac{F_0^T t_1}{\|t_1\|^2}.$$

记残差矩阵

$$E_1 = E_0 - t_1 p_1^T = (E_{11}, E_{12}, \cdots, E_{1p})$$

$$F_1 = F_0 - t_1 r_1$$

然后, 进行偏最小二乘回归的第 2 步, 以 E_1 取代 E_0, 以 F_1 取代 F_0, 以同样的方法重复第 1 步的工作得到

$$w_2 = \frac{E^T 1 F_1}{\|E^T 1 F_1\|} = \frac{1}{\sum_{j=1}^{p} Cov^2(E_{1j}, F_1)} \begin{bmatrix} Cov^2(E_{11}, F_1) \\ Cov^2(E_{12}, F_1) \\ \vdots \\ Cov^2(E_{1P}, F_1) \end{bmatrix}$$

$$t_2 = E_1 w_2$$

施行 E_1, F_1 对 t_2 的回归有

$$E_1 = t_2 p_2^T + E_2$$

$$F_1 = t_2 r_2^T + F_2$$

其中

$$p_2 = \frac{E_1^T t_2}{\|t_2\|^2}, \quad r_2 = \frac{F_1^T t_2}{\|t_2\|^2}$$

依此类推偏最小二乘回归的第 3 步, 第 4 步 $\cdots$ 最后, 可用交叉有效性确定偏最小二乘回归中成分 t_h 的提取个数, 停止迭代.

在得到的成分 $t_1, t_2, \cdots, t_m (m < A, A = rank(X))$ 后, 有 F_0 关于 t_h 的回归模型为

$$F_0 = r_1 t_1 + r_2 t_2 + \cdots + r_m t_m + F_m$$

2.4 偏最小二乘成分的确定—交叉有效性

在单因变量的偏最小二乘回归中, 交叉有效性的定义如下. 记 y_i 为原始数据, $t_1, t_2, \cdots, t_m$ 是在偏最小二乘回归过程中提取的成分. $\hat{y}_{hi}$ 是使用全部样本点并提取 h 个成分回归建模后,第 i 各样本点拟合值. $\hat{y}_{h(-i)}$ 是在建模时删去样本点 i, 取 h 个成分回归建模后,在用此模型计算 y_i 的 拟合值.记

$$S_{SS,h} = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2$$

$$S_{PRESS,h} = \sum_{i=1}^{n} (y_i - \hat{y}_{h(-i)})^2$$

$$Q_h^2 = 1 - \frac{S_{PRESS,h}}{S_{SS,h-1}}$$

当 $Q_h^2 \geq 0.0975$ (即 $\sqrt{S_{PRESS,h}} \leq 0.95\sqrt{S_{SS,h-1}}$)时, 引进新的主成分 t_h 会对模型的预测能力有明显的改善作用.

2.5 实证分析

变量与数据选取

本部分利用PLS回归方法, 建立偏最小二乘模型, 为防止农田沙化, 降低风蚀提供实证参考意见. 本节的案例数据有中国农业大学路明教授领导沙尘暴防治课题组提供. 我们主要选取了 7 个自变量:

x_1— 土壤含水量

x_2— 土壤颗粒直径

x_3— 地表覆盖率

x_4— 沙地

x_5— 传统耕作农田

x_6— 退化草地

x_7— 免耕法农田

我们将这 7 个变量来预测各样农田土壤风蚀量 y, 寻找防止农田沙化, 降低风蚀的关键因素.

表 1：各样农田土壤风蚀量与影响因素

序号	风蚀量 $/(g \cdot cm^{-2})$	土壤含水量/%	土壤颗粒直径/mm	地表覆盖率/%	沙地	传统耕作农田	退化草地	免耕法农田
1	11.6738	3.6227	0.6506	12.4	1	0	0	0
2	13.8116	3.6227	0.6506	12.4	1	0	0	0
3	15.26	3.6227	0.6506	12.4	1	0	0	0
4	12.1596	3.6227	0.6506	12.4	1	0	0	0
5	6.021	6.2909	0.266	13.8	0	1	0	0
6	8.598	6.2909	0.266	13.8	0	1	0	0
7	10.3952	6.2909	0.266	13.8	0	1	0	0
8	7.3308	6.2909	0.266	13.8	0	1	0	0
9	3.689	10.021	0.3366	45.4	0	0	1	0
10	5.3386	10.021	0.3366	45.4	0	0	1	0
11	5.9706	10.021	0.3366	45.4	0	0	1	0
12	4.8934	10.021	0.3366	45.4	0	0	1	0
13	2.768	8.8827	0.3386	58.5	0	0	0	1
14	4.1674	8.8827	0.3386	58.5	0	0	0	1
15	4.3572	8.8827	0.3386	58.5	0	0	0	1
16	4.111	8.8827	0.3386	58.5	0	0	0	1

注: 沙地, 传统耕作农田, 退化草地, 免耕法农田为虚拟变量, 表示不同的耕地类型,1– 是,2– 否.

实证结果

通过计算可知, 各自变量中各个变量高度相关, 其中土壤含水量 x_1 与沙地

x_4 的相关系数高达 98.05%, 土壤含水量 x_1 与土壤覆盖率 x_3 的相关系数高达 84.99%, 同时, 我们分别计算出各自变量的条件指数见表 2, 进一步判断各自变量的共线性程度.

表 2：多重共线性诊断

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7
条件指数	1	1.38	1.61	1914878	1914878	1914878	1914878

由于 x_4, x_5, x_6, x_7 的条件指数远远大于 30, 所以 x_4, x_5, x_6, x_7 间存在严重的共线性问题.

事实上, 通过数据, 我们也可以发现 $x_4 + x_5 + x_6 + x_7 = 1$. 下面我们运用PLS回归方法来克服这种缺陷.

应用交叉检验方法, 计算出 Y 的交叉有效性.

表 3：交叉有效性

成分个数	Q_h^2	临界值
1	0.86142	0.0975
2	-0.3977	0.0975

由于 $Q_2^2 < 0.00975$, 所以选择 $h = 1$, 即采用成分 t_1 做偏最小二乘回归模型, 预测效果最好.

PLS模型的精度分析

下面我们对预测模型做详细的精度分析, 具体结果见表 4.

表 4：精度分析

成分	t_1
RdX	52.3253%
RdY	89.8514%

表中符号 RdX 表示成分 t_h 对 X 的解释能力 $Rd(x_j, t_h)$, 其中

$$Rd(x_j, t_h) = \frac{1}{p}\sum_{j=1}^{p} r^2(x_j, t_h) \quad h = 1$$

$r^2(x_j,t_h)$ 表示 x_j 与 t_h 的相关系数, 符号 RdY 表示成分 t_h 对 Y 的解释能力 $Rd(Y,t_h)$, 其中

$$Rd(Y,t_h)=r^2(Y,t_h)\quad h=1$$

从表中的数据可知, 第一主成分 t_1 解释了原变量系统中 (52.33%) 的变异信息, 同时解释了因变量系统中 89.85% 的变异信息. 由此可见, 我们所提出的主成分对自变量, 因变量具有较好的解释性, 模型具有令人满意的精度要求.

PLS变量投影重要性指标VIP分析

在PLS分析的辅助分析技术中, 变量投影重要性指标VIP可以指出每一个自变量在解释因变量集合时的作用的重要性, 计算公式为:

$$VIP_j=\sqrt{\frac{p}{Rd(Y;t_1,t_2)}\sum_{h=1}^{2}Rd(Y;t_h)w_{hj}^2}$$

其中 VIP_j 表示第 j 个自变量 x_j 的投影重要性指标, p 表示自变量的个数, w_{hj} 是轴 w_h 的第 j 个分量, 它被用于测量 x_j 对构造 t_h 成分的边际贡献. 且对任意的 $h=1,2$, 总有 $\sum_{j=1}^{p}w_{jh}^2$. $Rd(Y;t_1,t_2)$ 代表主要成分对 Y 的累计解释能力. VIP_j 的取值见表 5:

表 5: 各变量的 VIP 值

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7
VIP	1.313	1.116	1.211	1.254	0.122	0.564	0.811

有表5 VIP 值知, 风蚀量与土壤颗粒直径土壤含水量地表覆盖率高度相关; 从不同类型的农田来看, 风蚀量与免耕法农田的相关程度最大, 说明了施行免耕法农田相对与其他农田对于保护土壤风蚀 有着最好的效果.

预测模型的建立

通过提取两个成分建立最终的回归模型

$$y=-0.236x_1+0.215x_2-0.231x_3+0.239x_4+0.011x_5-0.067x_6-0.183x_7$$

为了进一步考察模型的拟合程度, 我们绘制实际值与预测值图形, 进行对比分析.

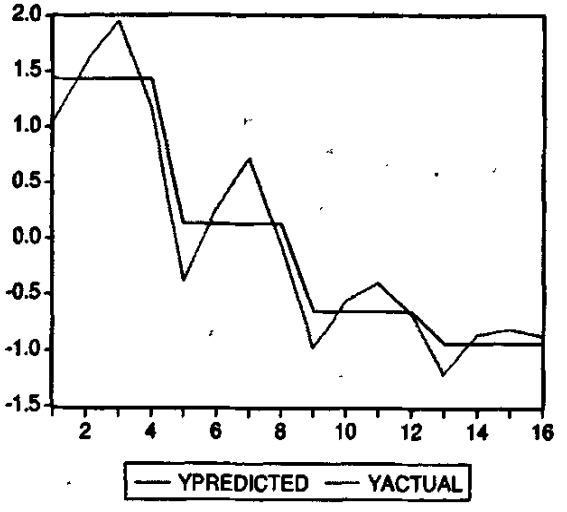

图 2.2 因变量观测值与预测值图

从图中可以看出, 模型的拟合效果还是比较好的(相对误差3.27%).

小结

通过上面的分析, 我们发现, 土壤含水量和地表覆盖率与风蚀量起负向作用, 增大土壤的含水量和地表覆盖率对降低土壤的风蚀有良好的作用, 土壤颗粒直径则起正向作用, 即土壤颗粒越大, 则越容易引起风蚀. 在不同类型的农田中, 沙化农田的风蚀程度是最高的, 采用免耕技术农田的风蚀量是最低的.

第三章 偏最小二乘算法改进和基于PLS路径模型的实证分析

本章对回归分析中自变量选择问题, 提出了偏最小二乘向前逐步回归选择变量法,利用PLS路径模型, 通过构建多指标系数评估指数, 对中国西部省区进行了综合分析.

3.1 引言

采用回归分析方法处理的变量问题时, 首先要选定回归变量, 在应用 PLS 方法时, 一方面, 可利用 PLS 方法本身提取的成分或者辅助分析技术进行多变量的选择, 如基于 PLS 成分的变量筛选法[21,22,23], 变量投影重要性指标VIP(variable importance in projection)[24]等; 另一方面, 可借用其他方法作为PLS的前置预处理以实现变量选择, 如交互变量选择法IVS(Iterative variable selection)[25,26,27]等. 本文在上述文献研究的基础上提出了偏最小二乘向前逐步回归法, 主要通过修正PLS权重或系数以消除模型中无用变量. 该方法可借助一元或多元线性回归, 通过对回归系数的统计检验来实现.

3.2 算法改进

下面通过把提取成分的每一步都与一个一元或者多元 OLS 回归联系起来, 给出这个算法的一种新的阐述.

• 计算PLS的第一个成分 t_1

第一个成分 t_1 定义如下:

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cov}^2(y, x_j)}} \sum_{j=1}^{p} \mathrm{Cov}(y, x_j) x_j \tag{3.2.1}$$

变量 x_j 的权重 $\mathrm{Cov}(y, x_j)$ 也可以写成 $Cor(y, x_j)^* s(y) s(x_j)$, 其中 $s(y)$ 和 $s(x_j)$ 分别是 y 和 x_j 的方差. 接下来, 为了使变量 x_j 在构造 t_1 时贡献很大,需要 x_j 与 y 有很强的相关性.

而在 OLS 单变量回归中, $\mathrm{Cov}(y, x_j)$ 是 y 和修正的解释变量 $x_j/Var(x_j)$ 的回归系数:

$$y = a_{1j}\left(\frac{x_j}{Var(x_j)}\right) + \varepsilon.$$

事实上,

$$a_{1j} = \frac{\mathrm{Cov}\left(y, x_j/Var(x_j)\right)}{Var\left(x_j/Var(x_j)\right)} = \mathrm{Cov}(y, x_j).$$

这样, 通过对回归系数 a_{1j} 的检验可以用来估计变量 x_j 在构造 t_1 时的重要性, 在这个基础上, 我们可以直接研究 y 关于 x_j 的回归:

$$y = a'_{1j}x_j + \varepsilon.$$

事实上, 检验系数 a_{1j} 或 a'_{1j} 是否显著不为零是等价的. 在 (3.2.1) 中, 用 0 替代每一个非显著的协方差, 从而去掉相关的解释变量. 因此, t_1可写成

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{1j}^2}} \sum_{j=1}^{p} a_{1j}x_j. \tag{3.2.2}$$

• 计算PLS的第二个成分 t_2

首先, 作 y 关于 t_1 和 $x_j, (j = 1, \cdots, p)$ 的多元回归, 寻找除 t_1 外, 对 y 有显著贡献的变量 x_j:

$$y = c_1 t_1 + a_{2j}x_j + \varepsilon;$$

同时作回归:

$$x_j = p_{1j}t_1 + x_{1j}, \quad j = 1, \cdots, p.$$

其中 x_{1j} 是回归方程的残差.

第二个成分 t_2 定义为:

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} Cov^2(y, x_{1j})}} \sum_{j=1}^{p} Cov(y, x_{1j})x_{1j}.$$

而在 OLS 多变量回归中, $Cov(y, x_{1j})$ 也是 y 关于 t_1 和$x_{1j}/Var(x_{1j})$ 的回归系数 a_{2j},

$$y = c_{1j}t_1 + a_{2j}\Big(\frac{x_{1j}}{Var(x_{1j})}\Big) + \varepsilon.$$

这样, 关于回归系数 a_{2j} 在构造 t_2 时的重要性的检验, 也可以变成直接研究 y 关于 t_1 和 x_j 的回归:

$$y = c_{1j}^{'}t_1 + a_{2j}^{'}x_j + \varepsilon.$$

事实上,

$$\begin{aligned} y &= c_1t_1 + a_{2j}x_j = c_1t_1 + a_{2j}(p_{1j}t_1 + x_{1j}) + \varepsilon \\ &= (c_1 + a_{2j}p_{1j})t_1 + a_{2j}x_{1j} + \varepsilon. \end{aligned}$$

按照第一步的做法, 可以去掉成分 t_2 中不重要的解释变量. 所以, t_2 可写成

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{2j}^2}} \sum_{j=1}^{p} a_{2j}x_{1j}. \tag{3.2.3}$$

• 计算PLS的第 h 个成分 t_h

第 h 个成分 t_h 定义为:

$$t_h = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{hj}^2}} \sum_{j=1}^{p} a_{hj}x_{(h-1)j}, \tag{3.2.4}$$

其中, a_{hj} 也是 y 关于$t_1, \cdots, t_{h-1}$ 和 $x_{(h-1)j}$ 的回归系数:

$$\begin{aligned} y &= c_1t_1 + c_2t_2 + \cdots + c_{h-1}t_{h-1} + a_{hj}(p_{1j}t_1 + p_{2j}t_2 \\ &\quad + \cdots + p_{(h-1)j}t_{(h-1)j} + x_{(h-1)j}) + \varepsilon \\ &= (c_1 + a_{hj}p_{1j})t_1 + (c_2 + a_{hj}p_{2j})t_2 + \cdots + \\ &\quad (c_{h-1} + a_{hj}p_{(h-1)j})t_{h-1} + a_{hj}x_{(h-1)j} + \varepsilon. \end{aligned}$$

停止原则: 当所有的回归系数显著为零时, 停止构建新成分.

3.3 实例分析

下面将通过一个具体的案例分析, 进一步理解该方法的工作过程和特点. 这是Cornell在 1990 年采用的一个化工方面的例子. 此后, 又被Wold, Tenenhaus 等人多次引用, 成为单因变量偏最小二乘回归的一个经典案例. 该例中, 有 7 个自变量 x_1—x_7, 因变量记为 y, 即:

x_1—直接蒸馏成分;

x_2—重整汽油;

x_3—原油热裂变;

x_4—原油催化裂化油;

x_5—聚合物;

x_6—烷基化物;

x_7—天然香精;

y—原辛烷值,

变量间的相关系数矩阵如下图所示:

表 6：相关系数矩阵

r$(\cdot,\cdot)$	x_2	x_3	x_4	x_5	x_6	x_7	y
x_1	0.10	0.999	0.37	−0.55	−0.80	0.60	−0.84
x_2		0.10	−0.54	−0.29	−0.19	−0.59	−0.07
x_3			0.37	−0.55	−0.80	0.61	−0.84
x_4				−0.21	−0.64	0.92	−0.71
x_5					0.46	−0.27	0.49
x_6						−0.66	0.98
x_7							−0.74

由表 6 知 $r(x_1,x_3)=0.999$, x_1 和 x_3 高度相关; $r(x_4,x_7)=0.92$, x_4 和 x_7 高度相关, 下面用修正的偏最小二乘进行回归分析.

• 计算PLS的第一个成分 t_1:

首先, 分别做 y 关于$x_1,x_2,\cdots x_7$ 的一元线性回归来寻找与 y 显著相关的变量 x_j, 回归系数显著性检验如下表:

表 7：回归系数显著性检验结果

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7
t值	−4.19	−0.02	−4.86	−3.16	1.80	18.09	−3.49
p值	0.0019	0.8269	0.007	0.0102	0.1028	< 0.0001	0.0058

由上表知, 在 $\alpha = 0.05$ 的水平下, 筛选掉变量 x_2, x_5 仅选择与 y 显著相关的变量 x_1, x_3, x_4, x_6, x_7 . 所以,

$$\begin{aligned} t_1 &= \frac{-0.837x_1 - 0.838x_3 - 0.707x_4 + 0.985x_6 - 0.741x_7}{\sqrt{0.837^2 + 0.838^2 + 0.707^2 + 0.985^2 + 0.741^2}} \\ &= 0.4526x_1 - 0.453x_3 - 0.382x_4 + 0.553x_6 - 0.401x_7. \end{aligned}$$

• 计算PLS的第二个成分 t_2:

首先寻找对构建 t_2 由显著贡献的变量 x_j, 做 y 关于 t_1 和 x_j 的多元回归, 在显著性水平 $\alpha = 0.05$, 变量的 p 值如下：

表 8：回归系数显著性检验结果

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7
t值	0.51	15.27	0.4	1.17	0.33	7.59	2.04
p值	0.62	< 0.0013	0.702	0.8723	0.7462	< 0.0001	0.0722

由上表知, 在 $\alpha = 0.05$ 的水平下, 只有变量 x_2, x_6 与 t_2 显著相关.

$$\begin{aligned} t_2 &= \frac{-0.196x_{12} + 0.652x_{16}}{\sqrt{0.196^2 + 0.652^2}} \\ &= 0.197x_1 - 0.294x_2 + 0.207x_3 + 0.174x_4 + 0.683x_6 + 0.183x_7. \end{aligned}$$

• 计算PLS的第三个成分 t_3:

首先寻找对构建 t_3 由显著贡献的变量 x_j, 做 y 关于 t_1, t_2 和 x_j 的多元回归, 在显著性水平 $\alpha = 0.05$, 变量的 p 值如下：

表 9：回归系数显著性检验结果

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7
t值	2.33	2.53	2.3	−2.75	−0.52	2.53	−1.82
p值	0.048	0.0353	0.049	0.0249	0.6159	0.0354	0.1057

由上表知, 在 $\alpha = 0.05$ 的水平下, 只有变量 x_1, x_2, x_3, x_4, x_6 与 t_3 显著相关. 所以

$$\begin{aligned} t_3 &= \frac{0.15909x_{21} + 0.23995x_{22} + 0.17944x_{23} - 0.14673x_{24} + 0.75893x_{26}}{\sqrt{0.15909^2 + 0.23995^2 + 0.17944^2 + 0.14673^2 + 0.75893^2}} \\ &= 0.329x_1 + 0.282x_2 + 0.36x_3 - 0.05x_4 + 0.73x_6 + 0.13x_7. \end{aligned}$$

• 计算PLS的第四个成分 t_4:

首先寻找对构建 t_4 由显著贡献的变量 x_j, 做 y 关于 t_1, t_2, t_3 和 x_j 的多元回归, 在显著性水平 $\alpha = 0.05$, 变量的 p 值如下:

表 10 : 回归系数显著性检验结果

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7
t值	−0.13	−0.05	−0.11	−0.42	0.35	−0.05	0.83
p值	0.9012	0.9634	0.9193	0.6863	0.7362	0.9620	0.4346

由上表知, 在 $\alpha = 0.05$ 的水平下, 没有显著性解释变量. 故保留三个PLS成分, 运算停止.

三个成分的PLS回归标准方程

$$\hat{y} = 0.511t_1 + 0.389t_2 + 0.185t_3.$$

化为原始变量的回归方程:

$$\begin{aligned} y &= 87.75902 - 6.338936x_1 - 2.07433x_2 - 10.46856x_3 \\ &\quad -3.987578x_4 + 15.04312x_6 - 26.78028x_7. \end{aligned}$$

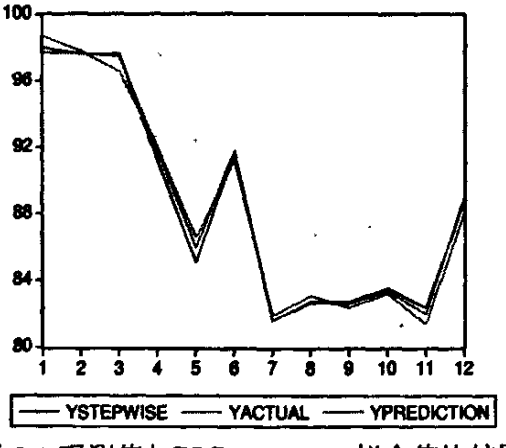

图 3.1 观测值与PLS，stepwise拟合值比较图

通过对比, 发现该方法比我们常用的逐步回归选择变量法拟合效果要好, 并且能够最大限度的保留被解释变量显著相关的解释变量. 因此具有一定的可行性.

3.4 基于PLS模型的实证分析

3.4.1 PLS路径模型

Wold(1985)提出PLS路径模型[28], 该模型主要有两部分组成. 第一是测度模型(又称“外部模型”): 用于描述显变量与隐变量之间的关系; 第二是结构模型(又称“内部模型”): 用于描述隐变量之间的关系. 参见图 3.2

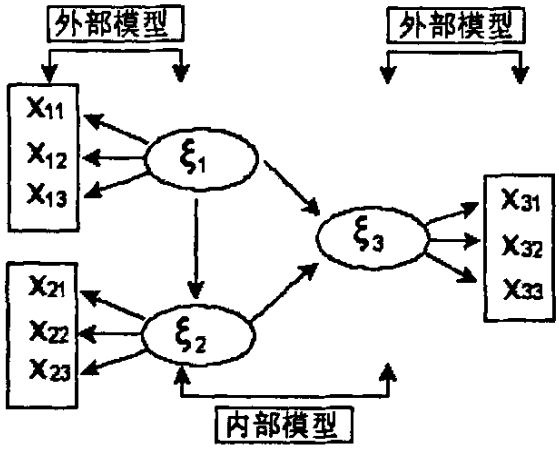

图 3.2 PLS模型路径示例

PLS路径模型的结构与假设条件

假设对于 n 个观测样本点有 J 组变量 $x_j = (x_{j1}, x_{j2}, \cdots, x_{jk})$, 其中变量 x_{jh} 称为“显变量”, 设它们都是中心化的变量(即变量的均值为零). 另外, 还假设每一组显变量都大致是“一维”的,即该组中每一个显变量都受到同一个标准化的隐变量 ξ_j 的影响, 第 j 组显变量 x_{jh} 与其隐变量 ξ_j 的关系, 通过外部模型(一元回归)表达为

$$X_{jh} = \pi_{jh}\xi_{jh} + \varepsilon_{jh},$$

其中, ξ_j 的均值为 0, 标准差为 1; 误差项 ε_{jh} 均值为 0, 且与隐变量 ξ_j 不相关; π_{jh} 为回归系数.

为了检验一组显变量是否符合 Unidimensionality 条件, 常用的检验方法是显变量组的主成分分析:

一般规定, 如果一组显变量相关系数矩阵第一个特征值大于 1, 其他特征值均小于 1, 那么可以认为这组显变量是唯一度的.

另一方面还可以通过内部模型, 描述 J 个隐变量 ξ_j 之间的关系, 其形式为

$$\xi_j = \sum_{i \neq j} \beta_{ij} \xi_i + \zeta_j \tag{3.3.2}$$

其中误差项 ζ_j 应满足均值为零且与 $\xi_i (i \neq j)$ 不相关的假设, β_{ji} 为回归系数.

PLS 路径模型中的参数估计方法

要对显变量组 $X_j (j = 1, 2, \cdot, J)$ 所对应的隐变量 ξ_j 进行估计, 可以有 2 种方法: 一种是根据显变量与隐变量之间的关系, 对隐变量进行计算, 这称为外部估计; 另一种方法是通过对隐变量之间的关联关系进行计算, 这称为内部估计.

1 外部估计

隐变量 ξ_j 可以有显变量 $x_{jh} (j = 1, 2, \cdots, J; h = 1, 2, \cdots, p_j)$ 的线性组合来估计, 记该估计量为 Y_j. 由于在模型假定中假设隐变量 ξ_j 是标准化的, 因此, 有

$$Y_j = (\sum_{h=1}^{p_j} w_{jh} x_{jh})^* = (X_j w_j)^* \tag{3.3.3}$$

式中: w_j 为权重向量, 星号 $*$ 表示对估计量进行标准化处理.

2 内部估计

根据结构模型也可以认为, 隐变量 ξ_j 还可以通过与之关联的其他隐变量进行估计. 这样得到的估计量被称为内部估计量, 记为 Z_j, 有

$$Z_j = (\sum_{i:\beta_{ji} \neq 0} e_{ji} Y_i)^* \tag{3.3.4}$$

式中: β_{ji} 为式 (3.3.2) 的方程系数; e_{ji} 称为内部权数. e_{ji} 的计算方法为:

$$e_{ji} = sign(r(Y_j, Y_i)) = \begin{cases} 1 & r(Y_j, Y_i) > 0 \\ -1 & r(Y_j, Y_i) < 0 \\ 0 & r(Y_j, Y_i) = 0 \end{cases}$$

式中: $sign$ 是符号函数; $r(Y_j, Y_i)$ 是外部估计量 Y_j 与 Y_i 的相关系数.

对于权重向量 w_j 的估计方法; 采用下列模式:

$$w_j = \frac{1}{n} X_j^T Z_j \tag{3.3.5}$$

其中, 权重向量 w_j 是变量 X_j 与 Z_j 的相关系数. 对于标准化的变量, 实际上 w_j 是 Z_j 对 X_j 作偏最小二乘回归的第一个成分的权数, 及偏最小二乘回归的第一个轴向量.

综上所述, PLS 路径分析采用迭代算法来计算隐变量, 最后根据隐变量的估计值, 计算测量模型与结构模型, 具体步骤如下.

第一步 取向量 Y_j 的初始值等于 x_{j1}.

第二步 通过式 (3.3.4), 计算 Z_j 的估计值, 即

$$Z_j = (\sum_{i:\beta_{ji}\neq 0} e_{ji} Y_i)^*$$

式中: $e_{ji} = sign(r(Y_j, Y_i))$.

第三步 根据 Z_j 的估计值, 通过式 (3.3.5) 计算权重向量 w_j 为

$$w_j = \frac{1}{n} X_j^T Z_j$$

第四步 利用计算得到的 w_j, 通过式 (3.3.3), 计算新的 Y_j 为

$$Y_j = (\sum_{h=1}^{p_j} w_{jh} x_{jh})^* = (X_j w_j)^*$$

再回到第二步, 直到计算收敛为止, 以最终的到的 Y_j 作为对隐变量 ξ_j 的估计值 $\widehat{\xi_j}$.

第五步 最后, 在用估计量 $\widehat{\xi_j}$ 代替隐变量 ξ_j 后, 运用普通最小二乘的多元回归方法, 来估计模型 (3.3.2) 中的各个系数.

3.4.2 实证分析

下面以中国西部 12 城市的发展水平为例, 说明 PLS 路径模型在构建综合评估指数进行综合评价的应用.本文采用LV-PLS1.8软件进行测算.

为了采用PLS路径模型对城市综合实力进行评估, 选取 2005 年中国西部 11 省/区/直辖市 作为分析对象, 采用下列三组指标进行分析: 见表 11.

表 11：西部省/区/直辖市/综合评价隐变量及显变量

隐变量		模型中变量名	显变量	
综合评价	经济发展 ξ_1	Economy	人均 GDP(元)	x_{11}
			人均固定资产投资(元)	x_{12}
			人均财政收入(元)	x_{13}
	生活水平 ξ_2	Living	职工人均工资(元)	x_{21}
			人均生活消费支出(元)	x_{22}
			人均可支配收入(元)	x_{23}
	教育水平 ξ_3	Education	平均每万人拥有高校教师数(人)	x_{31}
			平均教育经费(元)	x_{32}
			每万人公共图书馆藏书(册)	x_{33}

对三组变量分别做主成分分析，用于进行 Unidimensionality 检验，结果见表:

表 12：Unidimensionality 检验

变量组	第一主成分的特征值	第二主成分的特征值
经济水平	3.452	0.216
生活水平	2.963	0.175
教育水平	2.394	0.6172

从表 7 中可以看到，三个变量组的 Unidimensionality 检验都显示通过，下面建立城市综合评价的PLS路径模型如图 3.3：

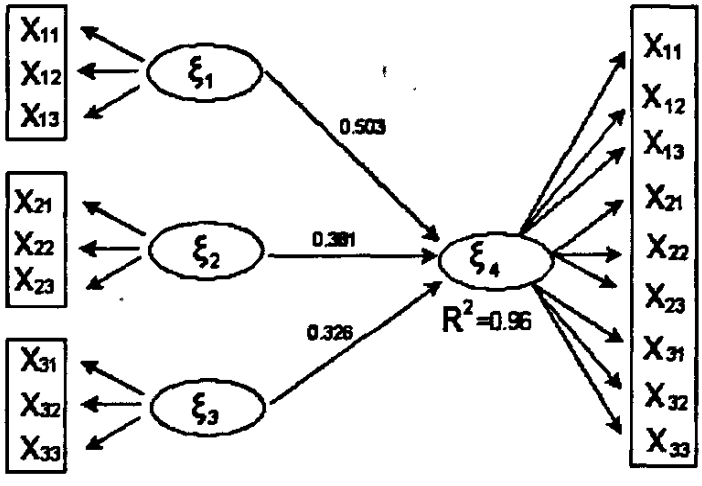

图 3.3 城市综合评价路径设计

对所有的显变量进行标准化处理, 然后利用 PLS 路径模型进行计算. 计算结果见表 13.

表 13 说明, 各组显变量与其相应的隐变量的相关程度均较高, 说明隐变量很好的概括了显变量组所包含的信息, 同时, 隐变量 ξ_4 对 ξ_1, ξ_2, ξ_3 的多元回归方程的 $R^2 = 0.96$, 说明, ξ_4 对 ξ_1, ξ_2, ξ_3 的概括程度相当高, 可以最大程度的表原是变量的信息.

表 13 变量的外部权和隐变量的相关系数

隐变量		经济发展			生活水平			教育水平		
内部权重		0.503			0.381			0.326		
显变量		人均GDP	人均固定资产投资	人均财政收入	职工人均工资	人均生活消费支出	人均可支配收入	平均每万人拥有高校教师数	平均教育经费	每万人公共图书馆藏书
外部权重	经济发展	0.438	0.428	0.269						
	生活水平				0.371	0.406	0.355			
	教育水平							0.384	0.796	0.062
相关系数	经济发展	0.961	0.837	0.785						
	生活水平				0.885	0.97	0.933			
	教育水平							0.576	0.783	

通过外部模型,可以利用外部权数直接计算出"综合评价"隐变量从而实现综合考察各个省/直辖市/区的总体发展情况见表 14.

表 14 西部 11 城市综合排名

省/区/直辖市	得分	综合排名	省/区/直辖市	得分	综合排名
新疆	1.92	1	四川	-0.71	7
重庆	0.62	2	云南	-0.82	8
内蒙古	0.48	3	广西	-0.87	9
宁夏	0.37	4	甘肃	-1.31	10
陕西	0.24	5	贵州	-1.46	11
青海	-0.06	6			

由上表, 可以看出西部各个地区发展的综合情况, 由于该模型所用的显变量均为人均水平, 故由于各地区人口因素影响可能会对模型评价有所影响.

小结: 本文利用PLS路径模型对我国西部地区各省/区/直辖市进行了综合评价, 该方法通过了一个路径模型反映了各个隐变量之间的关系及其与综合评价指数之间的关系, 充分提出原始变量的信息构造出综合评价指数, 在很大程度上克服多重共线性的影响, 从而使评价结果更趋于合理.

第四章 AR(p)模型和偏最小二乘回归藕合模型

§4.1 时间序列模型[29]

时间序列分析的基本思想认为, 某一变量要素在随时间变化的过程中任一时刻的变化和前期要素变化有关. 利用这种关系建立适当的模型来描述这一要素变量的变化规律, 然后利用所建立的模型做出变量要素未来时刻的预测值估计.

对于某一非平稳时间序列x_t, 首先需要判断时间序列特性, 如趋势性, 突变, 周期性等, 若存在, 则要剔除趋势性, 突变, 周期性等, 使非平稳序列平稳化. 对经过平稳化处理后得到的新序列可以按照平稳序列AR(p)模型进行计算, 模型的阶数可以根据AIC准则来确定.

下面介绍传统的Box—Jenkins时间序列建模方法[30]:

一. 将时间序列转换成平稳的;

二. 选取最适合的模型和阶数;

三. 寻找模型的参数;

步骤一包括运用差分的方法将时间序列转换成平稳的; 步骤二包括在判断最佳模型时所用的尝试方法, 此工作非常的繁琐, 通过对相关图及偏相关图的分析, 确定模型的形式; 最后一步相对较容易, 通常采用最小二乘拟合估计或极大似然估计方法. 建模的过程很复杂, 在一定程度上它要求猜测, 推导及经验等才能做出一个比较好的模型来.

由于本文中的序列都具有趋势项, 用传统的“Box—Jenkins”方法建模, 经过多次反复尝试检验, 都未得到好的模型. 因此对这些序列采取下列步骤:

1 将序列进行平稳化处理, 剔除时间趋势项.

2 对提取的残差序列通过自相关和偏自相关图确定模型的阶数及形式.

3 对原序列进行方程检验, 形成预测模型.

AR(p) 模型[31]

p 阶自回归模型可以表示为 $x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \alpha_t$, φ_k 为模型系数, α_t 为白噪声. 由如下公式可得 φ_k $(k = 1, 2, \cdots, p)$的估计值$\hat{\varphi}_k$:

$$\begin{pmatrix} \hat{\varphi}_1 \\ \hat{\varphi}_2 \\ \vdots \\ \hat{\varphi}_p \end{pmatrix} = \begin{pmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{pmatrix}$$

式中 $r_\tau(\tau = 1, 2, \cdots, p)$ 为样本序列 x_t 的自相关系数.

§4.2 建立偏最小二乘回归的城市用水量模型

1. 样本数据的选取

数据来自烟台市统计年鉴(1980-2000)[32], 从中选取7个因子: x_1 为总人口数(万人), x_2 为固定资产值(万元), x_3 为工业单位个数(个), x_4 为国内生产总值GDP(万元), x_5 为人均国民生产总值(元), x_6 为人均日常用水量(L), x_7 为供水能力($\times 10^4 m^3$/d), y 为水资源量($\times 10^4 m^3$).

2. 多重相关性诊断

利用方差膨胀因子对各自变量进行诊断, 检查其间是否存在多重相关性. 自变量 x_j 的方差膨胀因子记为(VIF), 其表达式为 $(VIF)_j = (1 - R_j^2)^{-1}$. 式中, $r_j{}^2$是以 x_j 为因变量时对其它自变量回归的复测定系数. 所有x_j变量中最大的VIF_j, 通常被用来作为变量多重相关性的指标. 如果最大的 VIF_j 超过10,即$r_j{}^2 < 0.9$, 表示多重相关性将严重影响最小二乘的估计值, 则自变量之间存在高度相关现象. 诊断结果见表1, 由表1可知, $r(x_4, x_5) = 0.998$, 则$r(x_4, x_5)^2$=0.996>0.9, $(VIF)_{max}$=250>10, 因此自变量之间存在多重相关性.

表4.1 变量相关系数

r	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
x_1	1							
x_2	0.848	1						
x_3	0.826	0.645	1					
x_4	0.833	0.974	0.562	1				
x_5	0.833	0.974	0.565	0.998	1			
x_6	0.507	0.577	0.68	0.484	0.489	1		
x_7	0.75	0.913	0.449	0.972	0.972	0.465	1	
y	0.871	0.909	0.535	0.95	0.949	0.327	0.926	1

3. 建立传统最小二乘多元回归模型

采用传统最小二乘法建立的多元回归模型为:

$$\begin{aligned} y &= 36.129x_1 + 0.00263x_2 - 0.23575x_3 + 0.00373x_4 - 2.73786x_5 \\ &\quad +6.85515x_6 + 41.6762x_7 - 13649.6 \end{aligned}$$

由上式可以看出, y (水资源量)与 x_5 (人均国民生产总值) 以及 x_3 (工业单位个数)的系数分别为 -2.73786 和-0.23575, 表明人均国民生产总值与工业单位个数增加, 水资源量反而减少, 显然不符合实际. 这说明各自变量间存在多重相关性, 不能用传统的最小二乘法建立回归模型, 否则, 会导致自变量对因变量的贡献程度无法解释, 与实际情况相违背.

4. 建立偏最小二乘回归模型

现将因变量序列 $y_i(i = 1,\cdots,21)$ 自变量序列 $x_{ij}(i = 1,\cdots,21; j = 1,\cdots,7)$ 进行标准化处理，得到自变量和因变量的标准化序列 F_0 与 E_{0j}. 采用Matlab7.0编程计算:

通过提取两个成分t_1, t_2，得到标准化回归预测方程:

$$\begin{aligned} F_0 &= 0.3156E_{01} + 0.1519E_{02} + 0.1586E_{03} + 0.2136E_{04} \\ &\quad +0.2121E_{05} - 0.26E_{06} + 0.237E_{07} \end{aligned}$$

$Q^3 = -0.07 < 0.095$，计算完毕。 将数据还原, 最后有原变量的偏最小二乘回归方程:

$$\begin{aligned} \hat{y} &= 20.33x_1 + 0.0007x_2 + 0.64x_3 + 0.000118x_4 \\ &\quad +0.074x_5 - 7.44x_6 + 10.34x_7 - 5807 \end{aligned}$$

从上式可以看出, 除了x_6的回归系数是负值, 其余都是正值, 与实际情况相符合. 由方程的系数可以看出, 总人口数(x_1)和供水能力(x_7)对烟台市的年生活用水量有较大影响. 其方程系数分别为20.33, 10.34, 并且呈正相关. 说明人口越多供水能力越强, 烟台市年生活用水越多.

§4.3 AR模型预测各项因子

由原始数据可以看出各项因子都有明显的逐年增大趋势, 因此推断预测模型中含有趋势项, 通过最小二乘估计出趋势项的直线方程, 再剔除趋势项后发现

回归残差为平稳序列. 这时对残差进行AR模型预测. AR模型的阶数p由AIC准则确定, 最后将预测的平稳序列再加上趋势项得到各因子未来年份的预测值, 各变量的自回归预测模型如下:

$$X_{1t} = 586.016 + 3.24t + 1.35AR(1) - 0.485AR(2)$$

$$R^2 = 0.99 \qquad DW = 1.96$$

$$X_{2t} = -948772.8 + 105196.7t + 1.6043AR(1) - 1.44AR(2) + 0.669AR(3)$$

$$R^2 = 0.96 \qquad DW = 2.2$$

$$X_{3t} = 1659.28 + 51.84t + 1.026AR(1) - 0.0377AR(2)$$

$$R^2 = 0.85 \qquad DW = 2.1$$

$$X_{4t} = -12553204 + 1008482t + 1.91AR(1) - 1.354AR(2) + 0.399AR(3)$$

$$R^2 = 0.997 \qquad DW = 2.25$$

$$X_{5t} = -26372.8 + 1782.97t + 1.983AR(1) - 1.526AR(2) + 0.503AR(3)$$

$$R^2 = 0.997 \qquad DW = 2.23$$

$$X_{6t} = 51.97 + 7.509t + 0.689AR(1) - 0.569AR(2)$$

$$R^2 = 0.63 \qquad DW = 1.99$$

$$X_{7t} = -33.389 + 7.24t + 1.336AR(1) - 0.61AR(2) + 0.07AR(3)$$

$$R^2 = 0.95 \qquad DW = 2.004$$

对各自变量预测模型进行误差验算, 平均相对误差为2.35%(计算过程略), 将模型的预测值代入偏最小二乘回归模型, 并与实际情况作比较,结果见表4.2

表4.2 组合预测方法分析验算

年份	实际数据	预测数据	相对误差
1997	11626	10847.77	0.0717
1998	11536	11209.49	0.0291
1999	11276	11360.96	-0.007
2000	11309	11852.34	-0.045

由表4.2知, 预测结果平均相对误差为1.22%, 可以满足城市需水量预测要求. 根据各自变量的预测模型, 计算出各自变量未来年份的预测值, 计算结果见表4.3.

表4.3 2005-2010自变量预测

年份	x_1	x_2	x_3	x_4	x_5	x_6	x_7
2005	655.685	1207024	2455.981	12120266	18830.83	4776.31	138.722
2006	658.049	1270643	2483.229	13012643	20235.83	5506.658	146.719
2007	660.417	1333872	2518.483	13899247	21621.7	6217.884	184.398
2008	662.783	1396533	2544.893	14783473	23002.26	6923.797	161.841
2009	665.146	1458786	2569.301	15668983	24390.37	7637.256	169.142
2010	667.505	1520829	2592.305	16557896	25792.39	8364.621	176.375

将所预测数据代入偏最小二乘回归模型, 即得到未来年份城市需水量的预测值, 计算结果见表4.3.

表4.3 2005-2010烟台市年生活用水量的预测

2005	2006	2007	2008	2009	2010
12800.97	13185.34	13540.03	13883.75	14230.85	14584.58

§4.4结果分析

由以上各步的计算结果可知: x_1 总人口数(万人), x_2 固定资产值(万元), x_3 工业单位个数(个), x_4 国内生产总值GDP(万元), x_5 人均国民生产总值(元), x_6 人均日常用水量(L), x_7 供水能力($\times 10^4 m^3/d$), y 水资源量($\times 10^4 m^3$), 在2005-2010年具有明显的增长趋势; 并且, 预测得到的2005-2010年的烟台市年生活用水量也相应的呈现增长趋势. 这说明随着社会的发展, 城市逐渐扩张, 城市人口不断地增加, 人民的生活水平也相应的提高了, 社会对水量的需求也会越来越高.

小结

(1) 偏最小二乘回归方法结合自变量集合 X 与因变量 y ,提取了互不相关的主成分, 再对这些主成份进行普通的多元回归, 解决了原自变量的多重相关性问题.

(2) 由于各项因子存在明显增大的趋势, 所以, 在用AR(p)模型对各项因子预测前先剔除其趋势项, 这样可以得到更为合理的预测结果.

(3) 将(2)中各因子的预测值代入(1)中的偏最小二乘回归方程, 从而实现了偏最小二乘回归对未来年份的预测, 可为烟台市的供水计划等提供参考依据.

总结

本文根据单因变量偏最小二乘算法, 提出了偏最小二乘向前选择变量法. 针对偏最小二乘无法对未来值预测的缺点, 采用了偏最小二乘与时间序列模型相结合的做法. 另外, 运用PLS路径模型分析对中国西部城市的综合评价进行了实证分析. 取得的结果如下:

(1) 采用偏最小二乘法能够最大限度的保留对因变量影响的重要变量, 对于我们寻找重要变量提供了依据.

(2) 偏最小二乘时间序列模型由于充分利用各自的优点, 因此, 对未来值预测时具有更好的解释性.

(3) PLS路径模型考虑了各指标之间的关系, 因此在构建综合评价指标时更具合理性.

进一步研究的思路:

(1)进一步研究PLS路径模型与结构方程的思想, 弄清它们之间的区别与联系找出二者的结合点.

(2) 偏最小二乘回归算法改进的理论研究.

(3) 偏最小二乘回归与ARIMA模型及非线性时间序列相结合的理论及应用.

由于本人才疏学浅, 文中出现疏漏错误在所难免, 恳请读者批评指正.

参考文献

[1] C.R.Rao, H.Toutenburg, Linear Models: Least Squares and Alternatives. [M], Springer-Verlag New York Inc. 1995.

[2] 童恒庆. 经济回归模型及计算. [M], 武汉: 湖北科学技术出版社, 1999.

[3] 任若恩,王惠文.多元统计数据分析—理论.方法.实例 [M], 北京: 国防工业出版社. 1998.

[4] P.Geladi.Wold, Herman, The Father of PLS. [J], Chemometrics and Intelligent Laboratory Syetem,1992,15(1)7-8.

[5] Wold, Herman. Path models with latent variables: The NIPALS approach. In Quantitative Sociology: International perspectives on mathematical and statistical model building, [M], CEd.S.H.M. Blalock et al, Academic Press, NY, 1975, 307-357

[6] Wold.H, Partial Least Squares, [M], [A], In Samuel Kotz and NOrman L.Johnson, eds, Encyclopedia of Statistical Science, Vol, 6, New York: Wiley, 1985, 581-591.

[7] Svante Wold,Michael Sjolstrom, LennartErikson. PLS-regression:a basic tool of chemometrics. [J], Chemometrics and Intelligent Laboratory Systems, 2001,58: 109-130.

[8] Richard Noonan, Herman Wold, Evaluating school systems using Partial Least Squares, Evaluation in education 1983, 7(3).

[9] De Jong S. SIMPLS: AN alternative approach to partial least squares regression. [J], Chemometrics and Intelligent Laboratory Systems, 1993, 18: 251-263.

[10] 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.

[11] 王惠文, 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.

[12] Wu xizhi, Partial least squares regression and its problems, a lecture given to university of North Carolina-Chapel Hill Columbia university, University of Pennsylvania, Auburn University and Cmory University, 2001.

[13] Cheng Bo, Wu xizhi, A modification of the PLS method,[J], Advances in Mathematics,1999,28(4): 375.

[14] Wold.S, Kettaneh N, Skagerberg B., Nonlinear PLS modeling, [J], Chemometrics and intelligent laboratory systems, 1989, 7: 53-65.

[15] Frank I E,Modern nonlinear regression methods, [J], Chemometrics and intelligent laboratory systems, 1995, 27: 1-9.

[16] Durand J F, Local polynomial additive regression though PLS and Splines: PLSS, [J], Chemometrics and intelligent laboratory systems, 2001, 58: 235-246.

[17] Guinot C, Latreille J.Tenechaus M.PLS path modeling and multiple table analysis.Applationto to the Cosmetic habits of women in Ile-de-Frace[J], Chemometrics and intelligent laboratory systems, 2001, 58: 247-259.

[18] 王惠文,付凌晖, PLS路径模型在建立综合评价指数中的应用, [J], 系统工程理论与实践, 10:80-85, 2004.

[19] Bayol MP,Delafoye A,Tellier C.Tenenhaus M., Use of PLS path modeling to estimate the European Consumer Satisfaction. Index(ESCI) Model,[J], Statistica Applacata-In Talian Journal of Applied statistics, 2000, 12(3): 361-375.

[20] GarthWaite PH. An interpretation of partial least squares, [J], JASA, 1994(1): 122.

[21] J.P.Gauchi, P.Chagnon,Comparison of selection methods of explanatory variable in PLS regression with application to manufacturing process data, [J], Chemometrics and intelligent laboratory systems, 2001, 58: 171-193.

[22] A.Lazraq, R.Cleroux, The PLS multivariate regression model: testing the significate of successive PLS components, [J], Journal of chenometrics, 15(2001), 523-536.

[24] A.Lazraq, R.Cleroux, Selecting both latent and explanatory variables in the PLS1 regression model, [J], Chemometrics and intelligent laboratory systems, 2003, 66: 117-126.

[25] Lingren, F.Geladi, P.Rannar, S.Wold, S., Interactive variable selection(IVS) for PLS Part I: theory and algorithms,[J], J.Chemometrics, 1994, 8, 349-362.

[26] Marx, B.D, Interatively reweighted partial least squares estimation for generalized linear regression,[J], Technornetrics , 1996, 38(4): 374-381.

[27] Forina, M, Casolino, C., Pizarro Millan, C. Iterative predictor weighting PLS (IPW): a technique for the elimination of useless predictors in regression problems, [J], J. Chemometrics, 1999, 13, 165-184

[28] Wold.H, Partial Least Squares, [A], Kotz S Johnson N L. Encyclopedia of Statistical Science, [M], New York: John Wiley & Sons, 1985.

[29] 何书元, 应用时间序列分析[M], 北京: 北京大学出版社, 2003

[30] 吴怀宇, 时间序列分析与综合. [M], 北京:高等教育出版社, 2000

[31] [美]Damodar N. Gujarati达摩达尔. N. 古扎拉剃, 计量经济学基础(上,下)中国人民大学出版社[M]

[32] 孟凡德, 刘贤赵, 烟台市水资源承载力变化的驱动力研究烟台师范学院学报(自然科学版), [J], 2003, 19(1): 46-50

研究成果

1. 新疆农村居民消费水平及结构的统计分析, 昌吉学院学报, 2006(12).
2. 基于时间序列偏最小二乘回归藕合模型的城市用水量的预测(待发表)

致谢

本文是在我的导师胡锡健副教授的悉心指导和不断鼓励下完成的. 本文的选题, 构思, 直至最后的完成, 胡老师都投入了许多的关注, 倾注了大量的心血. 我深深感到, 在三年的硕士研究生学习期间, 我的每一点进步都凝聚着胡老师大量的精力和汗水. 胡老师严谨的治学作风, 渊博的学识, 缜密的逻辑思维和富于创造, 开拓的思想; 平易近人, 诲人不倦的崇高品质都深深地感染着我. 特别是他在这几年中对我的严格要求和淳淳教导, 都将使我受益终生. 在此, 我向胡老师表示最诚挚的敬意和最衷心的感谢! 感谢胡老师三年来的谆谆教诲和培养之恩!

另外, 金融与统计教研室的吴黎军, 师恪老师给了我许多帮助, 我的成长和他们是分不开的! 感谢新疆大学数学与系统科学学院的各位领导和老师的关心和帮助!

同时, 感谢李涛, 李青, 覃龙, 于兰, 张自武和师姐郑彦玲在我做论文期间给予的帮助. 特别要感谢我的父母和妻子在我上研究生的三年期间给予的关心和支持.

本文的写作参考了国内外一些专家的论文成果, 在此一并向作者们表示衷心的感谢!