Statistical and Application 统计学与应用, 2015, 4(3), 133-143
Published Online September 2015 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2015.43015

Hans汉斯

The Application of Variable Selection to Multi-Collinearity Problems

—Based on the Research and Development Input and Output Data

Lei An, Huizhi Jia

School of Mathematics and Statistics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: anlei19890511@126.com

Received: Aug. 6th, 2015; accepted: Aug. 22nd, 2015; published: Aug. 27th, 2015

Abstract

A prerequisite for the promotion of a nation's innovation ability is the input of scientific research, but there are always many multi-collinearity problems among the indexes. In order to know the R&D input-output mode, 31 provinces are divided into two parts to set up ridge regression and PLS regression models separately. The research results show that different areas are influenced by different factors. The Midwest is susceptible to the input of the government and companies, while the technological innovation consciousness of the enterprises in the developed area is stronger.

Keywords

Variable Selection, Multi-Collinearity, Ridge Regression, Partial Least-Squares regression

变量选择方法在多重共线性问题中的应用

—基于全国科技投入产出数据的实例

安　蕾，贾慧芝

云南财经大学统计与数学学院，云南　昆明
Email: anlei19890511@126.com

收稿日期：2015年8月6日；录用日期：2015年8月22日；发布日期：2015年8月27日

文章引用: 安蕾, 贾慧芝. 变量选择方法在多重共线性问题中的应用[J]. 统计学与应用, 2015, 4(3): 133-143.
http://dx.doi.org/10.12677/sa.2015.43015

摘　要

科研投入是提升一国创新能力的前提，但指标之间往往存在较强的多重共线性问题。本文使用岭回归、PLS回归的方法，把我国31个主要的省市自治区分为两类，依次构建R&D投入－产出模型，以期了解我国R&D投入模式。研究结果表明，不同地区受科技投入指标的影响不同，中西部发展地区受政府及企业投入的影响都很显著，而经济较为发达的省市企业的科技创新意识更强。

关键词

变量选择，多重共线性，岭回归，PLS回归

1. 引言

科学研究与试验发展(R&D)能力是衡量一个国家科技创新实力及核心竞争力的关键指标，而科技投入体制对一国的科技发展水平起到决定性作用。针对这方面的研究，国内外的学者都取得了丰富的成果。Griliches (1979, 1986) [1] [2]提出知识生产函数，认为科研产出是研发资本及人力投入的结果。Hitt 等(1996) [3]研究发现企业自主创新能力随着研发经费投入的增加而增加。Inonu (2003) [4]以每百万人口的学术出版物数量及人均 GDP 为标准分类，对经济发展、文化因素与科研产出的关系进行阐述。在国内，余昕等(2007) [5]把 SCI 来源期刊论文量定为科研产出指标，通过对面板数据建立起科研投入产出关系模型，从定量的角度分析发达国家科研产出、科研经费投入、科研人员数及时间等因素的关系。李燕萍等(2009) [6]从环境因素、科研人员、科研经费投入、科研产出四要素的角度建立了影响科研经费有效使用的立体模型。

虽然相关的理论及实证研究较为丰富，但尚存在一些问题。例如科研的投入指标之间并非相互独立，很多情况下存在多重共线性，直接建模可能导致模型的不稳定。另外，现有的研究大多针对单一的产出指标进行影响因素分析，这种不全面的分析可能会导致结果的偏误。在方法的选择上本文尝试使用岭回归及 PLS 回归相结合，一方面可以解决投入指标间存在的多重共线性问题；另一方面，由于本文从多个角度选取投入、产出指标，按经济发展情况分区域构建多个自变量对多个因变量的模型，以期尽可能全面系统的分析科研活动投入体制及各产出指标之间的关系，导致出现分组后样本数少于变量数的情况，而 PLS 回归也能很好的解决这一问题。

2. 数据来源及变量选择

本文数据来自于《中国统计年鉴》及《中国科技统计年鉴》(2013 年)，实际数据为 2012 年全国 31 个省市自治区数据。

根据《中国统计年鉴》科学研究与开发机构部分，研究与试验发展(R&D)投入情况分为人员及经费。结合近年来科研人员对我国科技投入体制的研究[7]，R&D 活动投入指标我们从执行部门、研究方向、及经费来源三个方面进行选取。产出指标从不同的研究机构或执行部门的产出类别进行选取。R&D 投入及产出指标如表 1 所示。

3. 方法论

3.1. 多重共线性判断

考虑到各地区发展情况有很大差异，可能会对模型结果的准确度有影响，我们将样本分为东部经济

Table 1. Research and development input and output index
表 1. 科学研究与试验发展(R&D)投入－产出指标表

投入指标	直接投入	人员		x_1：R&D 人员全时当量(人年)
			按执行部门分	x_2：规上工业企业人员全时当量(人年)
				x_3：研究与开发机构人员全时当量(人年)
				x_4：高等学校人员全时当量(人年)
			按研究方向分	x_5：基础研究人员全时当量(人年)
				x_6：应用研究人员全时当量(人年)
				x_7：试验发展人员全时当量(人年)
		经费		x_8：研究与试验发展(R&D)经费内部支出(万元)
			按研究方向分	x_9：基础研究经费内部支出(万元)
				x_{10}：应用研究经费内部支出(万元)
				x_{11}：试验发展经费内部支出(万元)
			按资金来源分	x_{12}：政府资金经费内部支出(万元)
				x_{13}：企业资金经费内部支出(万元)
				x_{14}：其他资金经费内部支出(万元)
	间接投入	x_{15}：年度科普经费筹集额(万元)		
产出指标				y_1：发表科技论文数(篇)
		按按执行部门分		y_2：规上工业企业新产品开发项目数(项)

注：由于研究与开发机构及高等学校科技产出类别相同，我们定义 y_1 发表科技论文数(篇)为研究与开发机构发表科技论文(篇)及高等学校发表科技论文(篇)之和。

较发达地区(8 个省市：北京、天津、辽宁、上海、江苏、浙江、山东、广东)及中西部发展地区(余下 23 个省市)。选取的指标中，自变量有 15 个，2 个因变量，对东部发达地区建模时，样本个数少于变量个数。另外，考虑到投入指标间往往存在多重共线性，为保证模型的稳定性，我们在建模初要进行共线性判断。

目前有许多常见的多重共线性诊断方法，例如最常见的对自变量的相关系数矩阵进行诊断的方法表明，当自变量间的二元相关系数值很大时，则判定变量间存在多重共线性。然而由于此法中关于相关系数的具体值与共线性的关系无准确的标准，有时即使相关系数值并不太大，但也不能排除准确说不存在多重共线性。另外，容忍度(tolerance)、方差膨胀因子(variance inflation fator, VIF)、条件数(condition number)等都可以作为准则来度量多重共线性。这些判断准则可能不一致，但不失为一个参考。本文采用条件数判断多重共线性，常用 κ 表示，定义为：

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

式子中，λ 为 $X^{\mathrm{T}}X$ 的特征值(X 代表自变量矩阵)，一些研究者认为，当 $\kappa > 15$ 时有共线性问题，$\kappa > 30$ 时，说明共线性问题严重[8]。

如果数据存在多重共线性问题，常用的处理方法有比较经典的主成分分析、逐步回归法及 lasso 回归等，本文选择使用岭回归及 PLS 回归相对比的途径。目前有许多软件都可以进行岭回归及 PLS 回归的运算，但为了更好地普及变量选择方法，本文所有分析都通过可以从网上免费下载的自由软件 R 来实现。

3.2. 岭回归

假定自变量数据矩阵 $X=\{x_{ij}\}$ 为 $n\times p$ 的，通常最小二乘回归寻求那些使得残差平方和最小的系数 β，即

$$\left(\hat{\alpha}^{(ols)},\hat{\beta}^{(ols)}\right)=\arg\min_{(\alpha,\beta)}\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2.$$

岭回归则需要一个惩罚项来约束系数的大小，其惩罚项就是在上面的公式中增加一项 $\lambda\sum_{j=1}^{p}\beta_j^2$，即岭回归的系数既要使得残差平方和小，又不能使得系数太膨胀：

$$\left(\hat{\alpha}^{(ridge)},\hat{\beta}^{(ridge)}\right)=\arg\min_{(\alpha,\beta)}\sum_{i=1}^{n}\left[\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right],$$

这等价于在约束条件 $\sum_{j=1}^{p}\beta_j^2\le s$ 下，满足

$$\left(\hat{\alpha}^{(ridge)},\hat{\beta}^{(ridge)}\right)=\arg\min_{(\alpha,\beta)}\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2.$$

显然这里有确定 λ 或者 s 的问题，一般都用交叉验证或 Mallows Cp 等准则通过计算来确定。其中

$$\hat{\beta}(k)=\left(X'X+kI\right)^{-1}X'y$$

为 β 的岭回归估计，其中 k 成为岭参数，本文采用更方便的可以自动选择岭回归参数的程序包 ridge 中的函数 linearRidge()来实现[8]。

3.3. 偏最小二乘回归

为了研究因变量和自变量之间的统计关系，设有 p 个自变量 $\{x_1,\cdots,x_p\}$ 和 q 因变量 $\{y_1,\cdots,y_q\}$，取 n 个样本观测点，那么自变量与因变量就构成了数据表 $X=\{x_1,\cdots,x_p\}_{n\times p}$ 和 $Y=\{y_1,\cdots,y_q\}_{n\times q}$。为了回归分析的需要，偏最小二乘回归方法先分别在 X 与 Y 中提取出成分 t_1（t_1 是 $x_1,\cdots,x_p$ 的线性组合)和 u_1（u_1 是 $y_1,\cdots,y_q$ 的线性组合)，并要求其需要同时满足两个条件：

1) 根据主成分分析原理，为了能够代表数据表 X 和 Y，首先要求 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：

$$Var(t_1)\to\max$$
$$Var(u_1)\to\max$$

2) 其次要求从自变量中提取的成分 t_1 要在很大程度上能解释对从因变量中提取的成分 u_1，即要求 t_1 和 u_1 的相关性能够达到最大：

$$r(t_1,u_1)\to\max$$

首对成分提取后，偏最小二乘回归分别实施自变量 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、Y 被 t_1 解释后的残余信息进行第二轮的提取，直到能达到一个较为满意的精度。

最后，偏最小二乘回归将通过实施 $y_k\ (k=1,\cdots,q)$ 对从 X 中提取的 m 个成分：$t_1,t_2,\cdots,t_m$ 进行回归，进而表达成 y_k 关于原自变量 $x_1,\cdots,x_p$ 的回归方程[9]。

由于过多的成分可能会出现过拟合现象，因此很多时候，偏最小二乘回归法并不对全部的成分：$t_1,t_2,\cdots,t_A$ 进行回归。因此对于成分数的确定我们就需要有一个标准来进行判断，通常我们使用交叉验证的方法。常见的交叉验证法有“留一验证”，“K 折交叉验证”，“Holdout 验证”等。

交叉验证法将所有样本点随机的分成两部分：第一部分称训练集，用来重新拟合一个偏最小二乘模型；第二部分称测试集，将样本作为测试数据带入已经建好的拟合模型，并求出预测值误差平方和：$\mathrm{PRESS}=\sum\left(y_i-\hat{y}_i\right)^2$，为了将所有的样本都预测一次，我们利用上述方法重复进行 g 次，最后将每个样本的预测误差平方和进行加总构成 PRESS [10]：

$$\mathrm{PRESS}=\sum_{i=1}^{g}\mathrm{PRESS}_j$$

本文选取“留一验证”来计算不同成分数对应的 PRESS 值，选择在成分数尽可能小的情况下，PRESS 最小或几乎不变所对应的成分个数 m，再调整模型重新进行 PLS 回归。

偏最小二乘回归不同于一般的最小二乘法，它的回归系数方差无法得到准确的无偏估计，Miller R.G. (1974) [11]提出了用来估计回归系数的方差的方法：Quenouille-Tukey jackknife。与此方法相对应的，我们在 R 软件的 PLS 包中选取函数 jack.test 检验回归系数的显著性。

4. 实证分析

4.1. 中西部发展地区建模

4.1.1. 共线性判断

中西部发展地区我们抽取 23 个省市进行分析，15 个投入指标，2 个产出指标。读入数据后使用 R 固有的函数 kappa()计算条件数 κ，进行共线性判断。代码如下：

```
w=read.csv("12 发展.csv",header=T)
kappa(w[,1:15])
```

通过 R 软件计算得到：数据 w 的条件数 $\kappa=7225313$，远大于 30，可见 R&D 投入指标间存在严重的多重共线性问题，因此我们就不尝试简单回归，采取岭回归及偏最小二乘回归法对该数据进行回归建模。

4.1.2. 岭回归建模

使用 R 软件自带程序包 ridge 中的 linearRidge()函数进行拟合，代码如下：

```
w1=w[,1:16]#原数据 w 中所有自变量及 y1
w2=w[,c(1:15,17)]#原数据 W 中所有自变量及 y2
library(ridge)
a=linearRidge(y1~.,w1)
b=linearRidge(y2~.,w2)
summary(a)#看显著性、回归系数及岭回归参数
summary(b)
yp=predict(a,w1)
RF=sum((mean(w1$y1)-yp)^2)/sum((w1$y1-mean(w1$y1))^2);RF#求可决系数 R^2
yp=predict(b,w2)
RF=sum((mean(w2$y2)-yp)^2)/sum((w2$y2-mean(w2$y2))^2);RF
```

4.1.3. PLS 回归

使用 R 软件 PLS 程序包中的 plsr()函数进行拟合，代码如下：

```
#数据标准化
x=w[,1:15]
y=w[,16:17]#
X=scale(x)
Y=scale(y)
#初步回归
library(lars)
library(pls)
ap=plsr(Y~X,15,validation="LOO",jackknife=T)#要改，15 个自变量，即 15 个最大因字数,
summary(ap,what="all")#确定因子数，取 1
#修正的偏最小二乘回归
pls2=plsr(Y~X,ncomp=1,validation='LOO',jackknife=T)#要改
coef(pls2)#得到回归系数
jack.test(pls2)#回归参数的显著性检验
predplot(pls2)#画出最终模型的预测效果图
```

4.1.4. 结果分析

将前文回归结果汇总如下表 2，并进行分析：

根据回归系数表可写出 2012 年发展省市针对各因变量 y_i 的回归方程(由于篇幅限制，因变量的回归方程略)。由于数据在偏最小二乘回归前进行过标准化处理，我们可以直接看回归系数来初步判断各自变量对因变量的影响机制，通过对比我们发现：

① 对于中西部发展地区，经费内部支出是影响 R&D 各产出指标最重要的因素(标准化后回归方程的系数最大)。这也与实际情况相符，对于经济欠发达地区，科技投入利用率不高，提高产出主要靠大量增加人力物力投入的粗放型经济发展模式，科技投入的不足严重制约了各省的科技创新能力的提高和科技事业的发展。

② 投入指标按执行部门或研究机构来看，相对于研究机构及高等学校，企业对中西部发展地区科技产出的影响更大，该地区应该重视企业在科技创新中的作用，鼓励企业积极参与科技创新。

为检验回归参数的显著性，我们使用 R 软件 jack.test()函数，并将各回归系数对应的自变量显著情况整理如下表 3：

根据上表偏最小二乘回归结果我们可以看出：对于中西部发展省市，各科技投入指标对产出都起到很明显的促进作用，这与该地区的发展情况相符合，这些地区经济发展相对落后，R&D 人力物力资源都相对匮乏，对科技创新的意识有待加强，因此这些投入指标稍微增加都会对发展中地区的科技产出起到很明显的推动。从岭回归显著性结果我们看出：从资金来源看，影响中西部地区科技产出的最重要因素是企业资金及其他资金，我们应该在确保政府科技投入的前提下，启发企业及社会其他资源的投入。

通常为了判断模型的拟合优度，大家也使用可决系数 R^2，我们认为 R^2 的值越接近 1，说明回归直线对观测值的拟合程度越好。我们也可以使用 R 软件来计算各因变量对应的 R^2，计算出的拟合优度整理如下表 4 所示：

根据上表我们也可看出，使用岭回归及偏最小二乘回归构建的模型对各因变量实际观测值的拟合程度都达到 75%以上，模型拟合效果较好，两种方法都可酌情选用，且分析结果较为可靠。

Table 2. The ridge and PLS regression coefficients of Midwest
表 2. 中西部地区岭回归、偏最小二乘回归系数表

发展数据系数表	岭回归		PLS	
	论文	企业新产品	论文	企业新产品
(Intercept)	1.33E+03	−3.70E+02		
R.D 人员全时当量.人年.	2.04E−02	8.54E−03	8.13E−02	7.65E−02
企业 R.D 人员全时当量	−1.31E−02	1.95E−02	7.35E−02	6.92E−02
机构 R.D 人员全时当量	5.21E−01	−6.41E−02	6.47E−02	6.09E−02
高校 R.D 人员全时当量	7.49E−01	−4.86E−02	6.98E−02	6.57E−02
基础研究人员全时当量	6.32E−02	5.06E−02	6.53E−02	6.14E−02
应用研究人员全时当量	8.13E−02	1.69E−02	7.50E−02	7.06E−02
试验发展人员全时当量	2.50E−02	1.07E−02	7.72E−02	7.26E−02
经费内部支出.万元.	1.04E−03	3.94E−04	8.45E−02	7.95E−02
基础经费	3.25E−02	9.95E−03	7.29E−02	6.86E−02
应用经费	1.06E−02	6.18E−04	7.33E−02	6.89E−02
试验经费	9.74E−04	4.90E−04	8.21E−02	7.72E−02
政府资金	5.18E−04	5.27E−04	6.54E−02	6.15E−02
企业资金	1.63E−03	5.42E−04	7.66E−02	7.21E−02
其他资金	7.20E−02	1.43E−02	7.97E−02	7.50E−02
年度科普经费筹集额.万元.	1.15E−01	6.96E−03	5.72E−02	5.38E−02

Table 3. The effect of the R&D input indexes of Midwest
表 3. 中西部地区 R&D 投入指标显著性表

发展数据显著性表	岭回归		PLS	
	论文	企业新产品	论文	企业新产品
R.D 人员全时当量.人年.		**	***	***
企业 R.D 人员全时当量		***	***	**
机构 R.D 人员全时当量	**	.	***	***
高校 R.D 人员全时当量	**		***	***
基础研究人员全时当量			***	***
应用研究人员全时当量			***	***
试验发展人员全时当量		**	***	***
经费内部支出.万元.	**	***	***	***
基础经费		*	***	***
应用经费			***	***
试验经费	.	***	***	***
政府资金			**	***
企业资金	*	**	***	***
其他资金	.	.	***	***
年度科普经费筹集额.万元.			***	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 4. The R^2 of the regression modle of the Midwest
表 4. 中西部发展地区模型拟合优度表

R 方	岭回归	PLS
y1	0.894201	0.777699
y2	0.807578	0.836394

4.2. 东部发达地区建模

4.2.1. 共线性判断

选取八个经济较为发达的东部沿海省市(北京、天津、辽宁、上海、江苏、浙江、山东、广东)进行建模，15 个自变量，2 因变量，建模过程与中西部发展省市类似，代码略。首先我们对自变量进行共线性判断，计算结果 $\kappa = 1486.796$，远大于 30，数据存在多重共线性问题，另外考虑该地区数据样本量远小于变量个数，选择用岭回归及偏最小二乘回归，运算结果同发展地区，此略。

4.2.2. 结果对比

将岭回归及偏最小二乘回归结果汇总下表 5：

根据上表，我们可以写出相应的各个回归方程(篇幅限制，此处略)。同时我们发现，与中西部发展省市相比，经济较发达省市的回归系数出现负值。例如论文主要是科研单位、高校在基础、应用研究方面的科技产出，该变量受企业、试验发展类科技投入负增长也是合理的。同样的新产品开发项目数主要是规上企业的科技产出，同理可解释该回归方程的负向系数。

使用 R 软件 jack.test()函数检验回归参数的显著性，并将各回归系数对应的自变量显著情况整理如下表 6：

根据上表我们看出经济发达省市模型各变量显著性与发展地区明显不同：

① 对于论文这类科技产出，政府资金对其的影响最大，不受企业资金的影响。研究机构、高校在基础、应用研究领域的科技产出大多为论文、专著形式，投入多、回报期限较长，大多企业不想投资，因此由政府承担起对基础研究的支持作用。

② 从执行部门来看，经济发达省市的产出指标相较与中西部地区，资金投向更加明确，政府资金主要用于支持基础研究，而企业资金主要用于支持企业新产品项目开发，这主要是因为经济较发达省市的 R&D 投入渐渐由大幅度增加科技投入量的粗放型，发展为更加注重经费来源的多元化并提高企业自主开发能力。对于大多数国家而言，由于科技发展的公共品性质导致科技发展初始阶段都依靠政府资金的投入来支持科技发展，但到发展的后期，会逐步转向依靠企业资金的投入，从这个角度来看，我们国家经济较为发达的地区也不例外。

③ 对于规上工业企业的科技产出指标新产品开发项目数我们发现，它受政府资金及企业资金的双重影响都很显著，这主要是由于，这些地区虽然相对于本国其他地区经济发达，但我国科技投入的绝对水平与西方发达国家相比仍然偏低，我们虽然也要像发达国家那样鼓励企业提高科技创新意识，但政府也不能无限制降低科技投入比例，应该继续对企业的科技投入起引导作用。

为分析拟合效果，我们同样可以算出各因变量对应的拟合优度值，汇总如下表 7：

根据上表我们发现，虽然发达省市数据量较少，但经过岭回归及偏最小二乘回归的拟合优度均达到 80%以上，偏最小二乘回归在数据量远少于变量的情况下表现尤为良好。

4.3. 建模方法对比

当数据出现多重共线性问题时，可以使用岭回归或者偏最小二乘回归法，但两种方法各有利弊：

Table 5. The ridge and PLS regression coefficients of East
表 5. 东部地区岭回归、偏最小二乘回归系数表

发达数据	岭回归		PLS	
	论文	企业新产品	论文	企业新产品
(Intercept)	−1.58E+03	7.52E+03		
R.D 人员全时当量.人年.	1.22E−03	1.95E−02	4.14E−02	1.55E−01
企业 R.D 人员全时当量	−3.15E−03	1.91E−02	7.19E−03	1.63E−01
机构 R.D 人员全时当量	1.68E−01	−2.27E−02	9.70E−02	−6.12E−02
高校 R.D 人员全时当量	1.61E+00	−1.02E−02	1.10E−01	−8.21E−03
基础研究人员全时当量	9.85E−02	−2.16E−01	9.88E−02	−4.42E−02
应用研究人员全时当量	−9.65E−02	−2.31E−01	9.55E−02	−1.67E−02
试验发展人员全时当量	2.04E−03	2.44E−02	2.53E−02	1.64E−01
经费内部支出.万元.	1.09E−03	5.21E−04	7.42E−02	1.27E−01
基础经费	7.49E−03	−1.01E−03	1.01E−01	−4.18E−02
应用经费	4.66E−03	−2.96E−03	9.98E−02	−4.66E−02
试验经费	9.91E−04	7.25E−04	4.94E−02	1.51E−01
政府资金	2.57E−03	−3.38E−04	9.93E−02	−5.46E−02
企业资金	3.96E−04	6.06E−04	1.66E−02	1.60E−01
其他资金	2.65E−02	6.89E−03	1.09E−01	4.84E−03
年度科普经费筹集额.万元.	1.12E−01	4.86E−02	1.07E−01	−7.77E−03

Table 6. The effect of the R&D input indexes of East
表 6. 东部地区 R&D 投入指标显著性表

	岭回归		PLS	
发展数据显著性表	论文	企业新产品	论文	企业新产品
R.D 人员全时当量.人年.		***		***
企业 R.D 人员全时当量		***		***
机构 R.D 人员全时当量	**		***	**
高校 R.D 人员全时当量	**			
基础研究人员全时当量		*	***	.
应用研究人员全时当量		**	***	
试验发展人员全时当量		***		**
经费内部支出.万元.	*	*		*
基础经费	**		***	*
应用经费	.	**	***	.
试验经费	.	**		*
政府资金	***	*	**	***
企业资金		**		**
其他资金	*		***	
年度科普经费筹集额.万元.	*	*	***	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 7. The R^2 of the regression modle of the East
表 7. 东部发展地区模型拟合优度表

R 方	岭回归	PLS
y1	0.863094	0.919077
y2	0.805566	0.82586

① 岭回归无法对多个因变量进行组个建模，使用 R 软件进行编程时需要分别对因变量进行回归。但偏最小二乘法可将自变量及因变量分别作为一个整体进行回归。

② 岭回归借助程序包程序包 ridge 中的 linearRidge()函数可以实现自动选择回归参数，而偏最小二乘法需要在初步拟合后，再借助编程人员人工选择成分数进行模型的修订。

③ 当样本数量小于变量数目时，岭回归虽然也可以进行较好的拟合，但偏最小二乘法拟合效果更好。

5. 小结

本文利用岭回归及偏最小二乘回归法对中西部发展省市及东部经济较发达省市的 R&D 投入－产出进行建模，该方法利用其独有信息筛选模式解决了自变量间的多重共线性问题，同时很好的解决了经济发达省市样本量少于变量的问题，两组模型的拟合优度都在 80%以上，拟合效果较好，模型结果具有可参考性。

对于大多数国家而言，由于科技发展的公共品性质导致科技发展的初期阶段，资金来源主要依靠政府投入，而随着科学技术的应用程度的逐渐提高，企业资金投入在经济发达国家的科技投入中起着主要作用[7]。

通过分析我们发现：与国际上发展及发达国家科技投产机制的调整情况类似，对于我国中西部发展省市，R&D 人员全时当量及经费内部支出都对其 R&D 科技产出有明显的促进作用，政府资金、企业资金对 R&D 产出的影响都很显著，应该通过加大投入以获得更多的产出，同时在保证政府科技投入大幅度增加的前提下，引导企业、社会其他资源的投入，以科技创新带动当地经济发展。

对于东部经济较为发达的省市，企业 R&D 人员全时当量及企业资金对 R&D 科技产出指标的影响最显著，其次是其他资金，这主要是由于经济较为发达的省市，其 R&D 投入已渐渐从原来的强调大幅度的科技投入量的粗放型，转变为多目标体系，通过改进投入机制，逐步形成政府、企业和社会共同发展的多渠道的科技投入体系。

基金项目

国家自然科学基金项目“西部民族地区农村劳动力转移培训效应及政策优化研究——以云南民族地区为例”(71263055)。

参考文献 (References)

[1] Griliches, Z. (1979) Issues in assessing the contribution of R&D to productivity growth. *Bell Journal of Economics*, **10**, 92-116. http://dx.doi.org/10.2307/3003321

[2] Griliches, Z. (1981) Market value, R&D, and patents. *Economics Letters*, **7**, 183-187. http://dx.doi.org/10.1016/0165-1765(87)90114-5

[3] Hitt, M.A., Hosdisson, R.E., Johnson, R.A. and Moesel, D.D. (1996) The market for corporate control and firm innovation. *Academy of Management Journal*, **39**, 1084-1119. http://dx.doi.org/10.2307/256993

[4] Inonu, E. (2003) The influence of cultural factors on scientific production. *Scientometrics*, **56**, 137-146. http://dx.doi.org/10.1023/A:1021906925642

[5] 余昕, 王冬, 韩楠, 王欣 (2007) 发达国家科技投入效率初探. *科技进步与对策*, **8**, 129-131.

[6] 李燕萍, 郭玮, 黄霞 (2009) 科研经费的有效使用特征及其影响因素. *科学研究*, **11**, 1685-1691.

[7] 华锦阳, 汤丹 (2010) 科技投入机制的国际比较及对我国科技政策的建议. *科技进步与对策*, **5**, 25-30.

[8] 吴喜之 (2012) 复杂数据统计方法. 中国人民大学出版社, 北京, 25-29.

[9] 王惠文 (1999) 偏最小二乘回归方法及应用. 国防工业出版社, 北京, 151-152.

[10] 齐琛, 方秋莲 (2013) 偏最小二乘建模在 R 软件中的实现及实证分析. *数学理论与应用*, **2**, 104-105.

[11] Miller, R.G. (1974) An unbalanced jackknife. *The Annals of Statistics*, **2**, 880-891. http://dx.doi.org/10.1214/aos/1176342811