解决多元线性回归中多重共线性问题的方法分析

谢小韦，印凡成
河海大学理学院，南京 (210098)
E-mail：xiexiaowei@hhu.edu.cn

摘　要：为了解决多元线性回归中自变量之间的多重共线性问题，常用的有三种方法：岭回归、主成分回归和偏最小二乘回归。本文以考察职工平均货币工资为例，利用三种方法的SAS程序进行了回归分析，根据分析结果总结出三种方法的优缺点，结果表明如果能够使用定性分析和定量分析结合的方法确定一个合适的 k 值，则岭回归可以很好地消除共线性影响；主成分回归和偏最小二乘回归采用成份提取的方法进行回归建模，由于偏最小二乘回归考虑到与因变量的关系，因而比主成分回归更具优越性。

关键词：多重共线性；岭回归；主成分回归；偏最小二乘回归

1. 引言

现代化的工农业生产、社会经济生活、科学研究等各个领域中，经常要对数据进行分析、拟合及预测，多元线性回归是常用的方法之一。多元线性回归是研究多个自变量与一个因变量间是否存在线性关系,并用多元线性回归方程来表达这种关系，或者定量地刻画一个因变量与多个自变量间的线性依存关系。

在对实际问题的回归分析中，分析人员为避免遗漏重要的系统特征往往倾向于较周到地选取有关指标，但这些指标之间常有高度相关的现象，这便是多变量系统中的多重共线性现象。在多元线性回归分析中，这种变量的多重相关性常会严重影响参数估计，扩大模型误差，破坏模型的稳健性，从而导致整体的拟合度很大，但个体参数估计值的 t 统计量却很小，并且无法通过检验。由于它的危害十分严重，存在却又十分的普遍，因此就要设法消除多重线性的不良影响。

常用的解决多元线性回归中多重共线性问题的模型主要有主成分回归、岭回归以及偏最小二乘回归。三种方法采用不同的方法进行回归建模，决定了它们会产生不同的效果。本文以统计职工平均货币工资为例，考察一组存在共线性的数据，运用 SAS 程序对三种回归进行建模分析，并对结果进行比较，总结出它们的优势与局限，从而更好地指导我们解决实际问题。

2. 共线性诊断

拟合多元线性回归时，自变量之间因存在线性关系或近似线性关系，隐蔽变量的显著性，增加参数估计的方差，导致产生一个不稳定的模型，因此共线性诊断的方法是基于自变量的观测数据构成的矩阵 $X^T X$ 进行分析，使用各种反映自变量间相关性的指标。共线性诊断常用统计量有方差膨胀因子 VIF (或容限 TOL)、条件指数和方差比例等。

一般认为：若 VIF>10，说明模型中有很强的共线性关系；若条件指数值在 10 与 30 间为弱相关，在 30 与 100 间为中等相关，大于 100 为强相关；在大的条件指数中由方差比例超过 0.5 的自变量构成的变量子集就认为是相关变量集[1]。

3. 三种解决方法

岭回归基本思想: 当出现多重共线性时，有 $|X^T X| \approx 0$，从而使参数的 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 很不稳定，出现不符合含义的估计值，给 $X^T X$ 加上一个正常数矩阵 $KI(K > 0)$，则 $|X^T X + KI|$ 等

于 0 的可能性就比$|X^TX|$的可能性要小得多，再用$\hat{\beta}=(X^TX+KI)^{-1}X^TY$来估计,$\hat{\beta}$比用普通最小二乘估计的$\hat{\beta}$要稳定得多。

主成分回归基本思想：观察 n 个样本点，得到因变量 y 和 p 个自变量$x_1,x_2,\cdots,x_p$关系，设自变量$x_0=(x_1,x_2,\cdots,x_p)$间的相关数矩阵记为$\boldsymbol{R}$。

主成分回归方法完全撇开因变量 y ，单独考虑对自变量集合做主成分提取。其过程是:

1) 求$\boldsymbol{R}$的前m个非零特征值$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m>0$，以及相应的特征向量$u_1,u_2,\cdots,u_m$；

2) 求m个主成分:$F_h=X_0u_h \qquad h=1,2,\cdots,m$

偏最小二乘回归的基本思想： 首先在自变量集中提取第一潜因子t_1（t_1是$x_1,x_2,\cdots,x_m$的线性组合, 且尽可能多地提取原自变量集中的变异信息, 比如第一主成分)；同时在因变量集中也提取第一潜因子u_1,并要求t_1与u_1相关程度达最大。然后建立因变量 Y 与t_1的回归，如果回归方程已达到满意的精度, 则算法终止。否则继续第二轮潜在因子的提取, 直到能达到满的精度为止。若最终对自变量集提取l个潜因子$t_1,t_2,\cdots,t_l$，偏最小二乘回归将通过建立 Y 与$t_1,t_2,\cdots,t_l$的回归式, 然后表示为 Y 与原自变量的回归方程式[2]。

4. 实例分析

全国单位大体分成三大类：国有单位，城镇集体单位和其他单位，考虑到职工的平均工资主要和这三类单位的工资有关，为了研究和分析我国职工的平均工资，需建立一个以职工平均工资为因变量，三类单位的工资为自变量的回归方程。

考察职工平均货币工资指数 Y 与国有单位货币工资指数 x1，城镇集体单位货币工资指数 x2， 其他单位货币工资指数 x3 等三个自变量有关。现收集 1991 年至 2005 年共 15 年的数据，如表 1 所示。

表 1　职工货币工资指数

Table 1 The index of staff’s monetary wage

货 币 工 资　　指　数 (上年=100)				
年数	国有单位x1	城镇集体单位 x2	其他单位x3	平均工资y
1991	108.5	111.0	116.1	109.3
1992	116.2	113.0	114.4	115.9
1993	122.7	122.9	125.2	124.3
1994	135.8	125.2	126.9	134.6
1995	117.3	121.1	118.4	121.2
1996	111.6	109.4	110.7	112.9
1997	107.4	104.9	106.4	104.2
1998	106.1	102.5	97.7	106.6
1999	111.4	108.3	109.6	111.6

2000	111.8	108.5	111.8	112.3
2001	117.0	109.7	110.5	116.0
2002	115.1	111.6	108.8	114.3
2003	113.3	113.2	110.3	113.0
2004	114.8	113.1	111.6	114.1
2005	115.4	115.0	112.2	114.6

运用 SAS 程序对这组数据进行共线性诊断，输出结果见图 1。

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-8.38080	6.49068	-1.29	0.2231	0
x1	1	0.74938	0.10317	7.26	<.0001	4.15932
x2	1	0.34460	0.17075	2.02	0.0686	8.90299
x3	1	-0.01407	0.13048	-0.11	0.9161	6.57394

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	x1	x2	x3
1	3.99673	1.00000	0.00018214	0.00005441	0.00002072	0.00003539
2	0.00236	41.16127	0.91903	0.03317	0.00659	0.02860
3	0.00068155	76.57817	0.00053669	0.77970	0.01414	0.29860
4	0.00022778	132.46166	0.08025	0.18708	0.97925	0.67277

图 1 数据共线性诊断的部分结果

Fig .1 Collinearity Diagnostics of the data (part)

由图 1 的共线性诊断结果可以知最大条件指数 132.46>100，说明 4 个自变量间有强相关性，与最大条件指数在一行的 3 个变量中有 2 个变量的方差比例都大于 0.5，可见这 4 个变量是一个具有强相关的变量集。

由此得到回归方程为：

$$y = -8.380 + 0.749x1 + 0.345x2 - 0.014x3$$

可以看到变量 x3 的系数为负，这与实际情况不符。出现此现象的原因是变量 x1 与 x2，x3，x4 线性相关($\rho(x_1, x_2) = 0.9756, \rho(x_1, x_3) = 0.9207, \rho(x_1, x_4) = 0.9268$)，此处也可看出这 4 个变量是多重相关的变量集。

4.1 运用岭回归 SAS 程序进行回归分析

为了消除变量之间的多重共线性关系，用岭回归方法来建立回归方程产，并用 SAS 程序进行岭回归分析，部分结果见图 2、3，从岭迹图中可以看出，当 $k \geq 0.02$ 后，岭迹图趋于稳定。

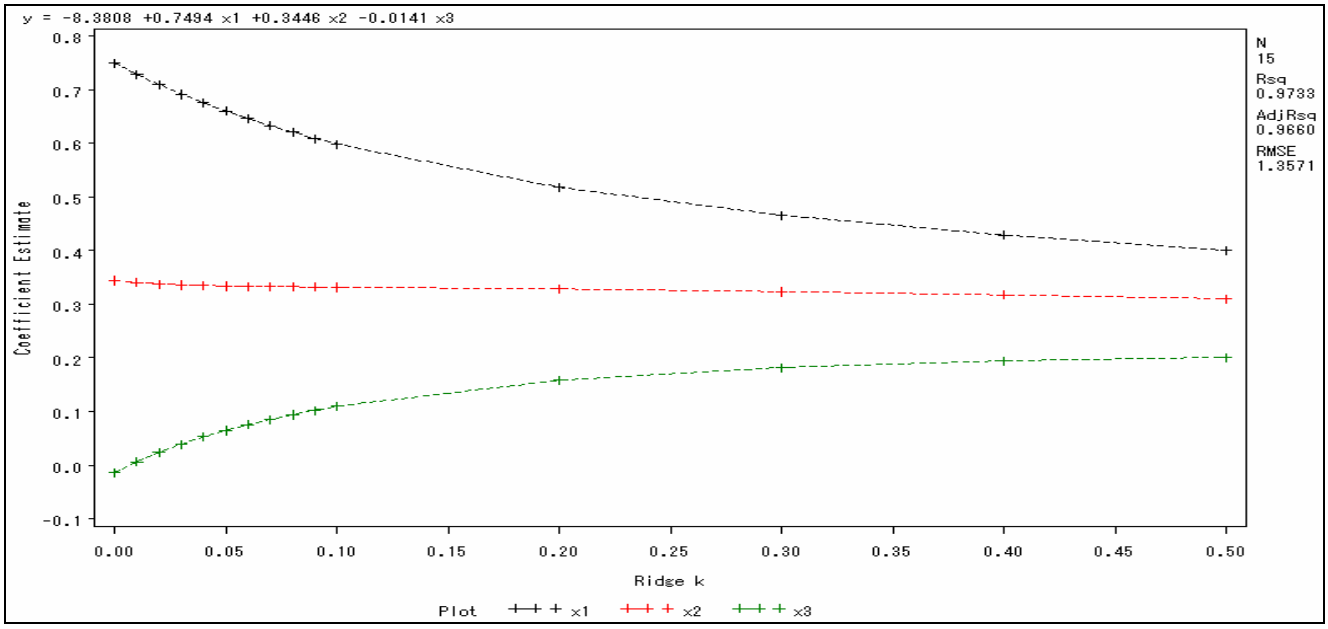

图 2　职工平均工资的岭迹图

Fig. 2　Ridge mark chart of staff's average wage

取 $k=0.02$ 的岭回归估计来建立岭回归方程，由图 3 可以写出岭回归方程式为:

$$y=-7.312+0.709x1+0.338x2+0.024x3$$

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	x1	x2	x3	y
1	MODEL1	PARMS	y	.	.	1.35714	-8.3808	0.74938	0.34460	-0.01407	-1
2	MODEL1	RIDGEVIF	y	0.00	.	.	.	4.15932	8.90299	6.57394	-1
3	MODEL1	RIDGE	y	0.00	.	1.35714	-8.3808	0.74938	0.34460	-0.01407	-1
4	MODEL1	RIDGEVIF	y	0.01	.	.	.	3.66978	6.91260	5.32043	-1
5	MODEL1	RIDGE	y	0.01	.	1.36101	-7.8348	0.72860	0.34066	0.00622	-1
6	MODEL1	RIDGEVIF	y	0.02	.	.	.	3.27570	5.53341	4.42491	-1
7	MODEL1	RIDGE	y	0.02	.	1.37114	-7.3119	0.70945	0.33802	0.02375	-1
8	MODEL1	RIDGEVIF	y	0.03	.	.	.	2.95047	4.53787	3.75848	-1
9	MODEL1	RIDGE	y	0.03	.	1.38577	-6.8064	0.69177	0.33623	0.03910	-1
10	MODEL1	RIDGEVIF	y	0.04	.	.	.	2.67706	3.79536	3.24629	-1

图 3　职工平均工资数据输出数据集(部分)

Fig. 3　Output of staff's average wage data (part)

可以看出各个回归系数的方差膨胀因子均小于 6，岭回归方差的均方根误差为 1.37114，虽比普通最小二乘回归方程的均方根误差(1.35714)有所增大，但增加不多。

4.2　运用主成分回归 SAS 程序进行回归分析

运用 SAS 程序可以得出删去第三个主成分后的主成分回归方程，结果见图 4。主成分回归方程为：

$$y=-7.701+0.767x1+0.274x2+0.033x3$$

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	x1	x2	x3	y
1	MODEL1	PARMS	y	.	.	1.35714	-8.38080	0.74938	0.34460	-0.01407	-1
2	MODEL1	IPCVIF	y	.	1	.	.	3.46990	0.33931	1.87637	-1
3	MODEL1	IPC	y	.	1	1.30996	-7.70125	0.76721	0.27352	0.03275	-1
4	MODEL1	IPCVIF	y	.	2	.	.	0.11692	0.12630	0.12169	-1
5	MODEL1	IPC	y	.	2	2.15413	-7.34473	0.33697	0.39619	0.34584	-1

图 4　职工平均工资数据主成分回归的结果

Fig. 4　Principal Component Regression's result of staff's average wage data

可以看出各个回归系数的方差膨胀因子均小于 3.5；主成分回归方程的均方根误差为 1.30996，比普通最小二乘回归方程的均方根误差(1.35714)有所减小。

4.3 运用偏最小二乘回归 SAS 程序进行回归分析

最后，使用 SAS 软件中得 PLS 过程完成偏最小二乘回归分析，输出结果见图 5。

The PLS Procedure

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS
0	1.071429
1	0.33616
2	0.21046
3	0.228461

Minimum root mean PRESS 0.2105
Minimizing number of factors 2

Obs	_LV_	_TYPE_	x1	x2	x3	_X_	y	_Y_
1	2	SIM	.	.	.	.	.	.
2	.	CENTER	114.960	112.627	112.707	.	114.993	.
3	.	SCALE	7.170	6.338	7.127	.	7.365	.
4	1	R	0.615	0.578	0.539	.	.	.
5	2	R	0.779	-0.156	-0.611	.	.	.
6	1	B	0.358	0.337	0.314	1.00000	.	.
7	2	B	0.741	0.260	0.013	1.00000	.	.
8	1	PQ	0.569	0.588	0.575	.	1.000	.
9	2	PQ	0.752	-0.219	-0.622	.	1.000	.
10	0	V	1.000	1.000	1.000	1.00000	1.000	1.00000
11	1	V	0.888	0.946	0.906	0.91323	0.928	0.92752
12	2	V	0.995	0.955	0.979	0.97605	0.973	0.97316

图 5 职工平均工资数据偏最小二乘回归的结果

Fig.5 Partial Least Square Regression's result of staff's average wage data

由估计值可以写出标准化回归方程为 $\tilde{y}=0.741\tilde{x}1+0.260\tilde{x}2+0.013\tilde{x}3$，用原始变量可表示为 $y=-7.973+0.761x1+0.302x2+0.013x3$ [3]。偏最小二乘回归方程中回归系数的符号都是有意义的。可知偏最小二乘回归方程的均方根误差为 1.18075，比普通最小二乘回归方程的均方根误差(1.35714)有所减小，且比主成分回归方程的均方根误差为 1.30996 也有所减小。

由实例看出，对于这组数据的处理，三种方法中岭回归的效果相对较差，主成分回归次之，偏最小二乘回归的计算结果更为可靠。

5. 结论

岭回归估计量的质量取决于 k 值的选取，一般认为：在通过岭迹图和方差膨胀因子来选择 k 值时，其判断方法是选择一个尽可能小的 k 值，在这个较小的 k 值上，岭迹图中回归系数已变得比较稳定，并且方差膨胀因子也变得足够小。从上面的实例中可以看出岭回归的效果相对较差，这是由于 k 值的确定存在一定的人为因素，所以在确定 k 值的时候要把定性分析和定量分析有机的结合起来。这样才能充分发挥岭回归的优点。

利用主成分回归的方法使主成分之间不再存在自相关现象，这就对解决多重相关性下的回归建模问题给出了某种希望。这种成分提取的思路是十分可取的，但利用主成分进行的回归很多时候结果往往不够理想，原因在于，在上述成分提取过程中，完全没有考虑与因变量 y 的联系。这样所得到的第 1 (或前几个) 主成分可能会对自变量系统有很强的概括能力，而对 y 的解释能力却变得十分微弱。

偏最小二乘回归也采用成份提取的方式进行回归建模，但其思路与主成分回归却有很大的不同。它在对自变量进行信息综合时，不但考虑要最好的概括自变量系统中的信息，而且

要求所提取的成分必须对因变量有一定的解释性。分析结果表明，与主成分回归相比，偏最小二乘回归更具有先进性，其计算结果更为可靠。偏最小二乘回归法尤其适用于变量数目巨大的情况下，如果数据中变量的个数多，偏最小二乘回归的优点更能充分的显示出来。

在解决多重共线性问题的三种方法中，岭回归的关键在于 k 值的选择，而 k 的选择存在一定的人为因素，如果能够使用定性分析和定量分析结合的方法确定 k 值，则岭回归可以很好地消除共线性影响。主成分回归只注重尽可能多的概括自变量系统中的信息，对因变量的解释性毫不考虑，相比之下，偏最小二乘回归不单概括了自变量的信息，还考虑到了提取的成分对因变量有最好的解释，其计算结果更可靠，因此它比主成分回归更具优越性，这种优势在自变量数目巨大的情况下表现地尤为突出。

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，1999:67-84.
[2] 高惠璇. 处理多元线性回归中自变量共线性的几种方法[J]，数理统计与管理，2000，9（5）:
[3] 高惠璇. 两个多重相关变量组的统计分析[J]，数理统计与管理，2002，3（2）:

Analysis of methods to solve the problem of multi-correlation between variables in multi-linear regression

Xie Xiaowei， Yin Fancheng
hohai university，Nanjing（210098）

Abstract

In order to solve the problem of multi-correlation between variables in multi-linear regression, three methods are commonly used: Ridge Regression, Principal Component Regression and Partial Least Square Regression. This paper takes staff's average wage statistics as an example, using the three methods' SAS procedure to make regression analysis. Based on the result, their advantages and disadvantages are summed up. The results also indicate that, if the union of qualitative analysis and quantitative analysis can determine an appropriate k value, Ridge Regression will be good to eliminate the influence of multi-correlation. Both Principal Component Regression and Partial Least Square Regression create regression modeling through extracting ingredients, as Partial Least Square Regression takes account the relation between dependent variable, it's superior to Principal Component Regression.

Keywords: multi-correlation，Ridge Regression，Principal Component Regression，Partial Least Square Regression