第六章 地理系统要素的逐步回归分析

一 最优回归方程的选择

- **为什么要进行地理要素逐步回归**
  - 在水文、气象、地震等预报工作中，以及进行人文地理、经济地理要素分析时，需要选出对**研究变量有影响的因子**
  - **不同的影响因子对于因变量的影响程度不同**，如何选出影响最大的因子?
  - 首先选取大量的因子进行考虑，但**如何对这些因子进行筛选**，选出对因变量影响最大的因子，是本章要掌握的关键

一 最优回归方程的选择

原始数据

编号	x1	x2	x3	x4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

一 最优回归方程的选择

- 最优回归方程
  - 回归方程中所包含的自变量越多，回归平方和就愈大，而剩余平方和就愈小，因此剩余方差（相应的剩余标准差）一般也越小
  - 一个合理的回归方程应该只包含显著的因子，而不应该包含不显著的因子
  - 一个合理的回归方程应该包括一些在现有条件下容易测定的因子
  - 所谓的回归方程就是包含对y显著的变量，而不包括对y不显著变量的回归方程

1 最优回归方程的选择

- 最优回归方程的选择：设有y与x_1，x_2，x_3，x_4的一组观测数据，如表6-1
  - 从所有可能的变量组合中挑选最优者：即把所有包含1个，2个，…直至所有变量的线性回归方程全部计算出来，并对每个方程及自变量做显著性检验，然后从中选出一个方程，要求该方程中所有的变量全部显著，且剩余均方和较小
  - 统计自变量和因变量之间的所有方程
  - 计算全部方程的显著性及其剩余均方和
  - 选出最优方程

一 最优回归方程的选择

全部可能的线性回归结果的比较

	b0	b1	b2	b3	b4	$S_{剩}$	$f_{剩}$	a^2
						2175.76	12	
(1)	81.4793	1.8687**				1265.69	11	115.06
(2)	57.4236		0.7891 **			906.34	11	82.39
(3)	110.2026			-1.2558(*)		1939.40	11	176.31
(4)	117.5679				-0.7382 **	883.87	11	80.35
(5)	52.5773	1.4683 **	0.6623 **			57.90	10	5.79
(6)	72.3490	2.3125 *		0.4945		1227.07	10	122.71
(7)	103.0973	1.4400 **			-0.6140 **	74.76	10	7.48
(8)	72.0747		0.7313 **	-1.0084 **		415.44	10	41.54
(9)	94.1600		0.3109		-0.4569	868.88	10	86.89
(10)	131.2824			-1.1999 **	-0.7246 **	175.74	10	17.57
(11)	48.1936	1.6959 **	0.6569 **	0.2500		48.11	9	5.35
(12)	71.6842	1.4519 **	0.4161(*)		-0.2365	47.97	9	5.33
(13)	111.6844	1.0519 **		-0.4100(*)	-0.6428 **	50.84	9	5.65
(14)	203.6418		-0.9234 **	-1.4480 **	-1.5570 **	73.82	9	8.20
(15)	62.4052	1.5511(*)	0.5101	0.1019	-0.1441	47.86	8	5.98

1 最优回归方程的选择

- **方法优点**
  - 总是可以找到一个最优方程
- **方法缺点**
  - 由于地理影响因素的复杂性，实际工作中单个变量的影响因素往往很多
  - 如果存在n个影响因子，就会有2^n-1个方程存在，降低计算效率

一 最优回归方程的选择

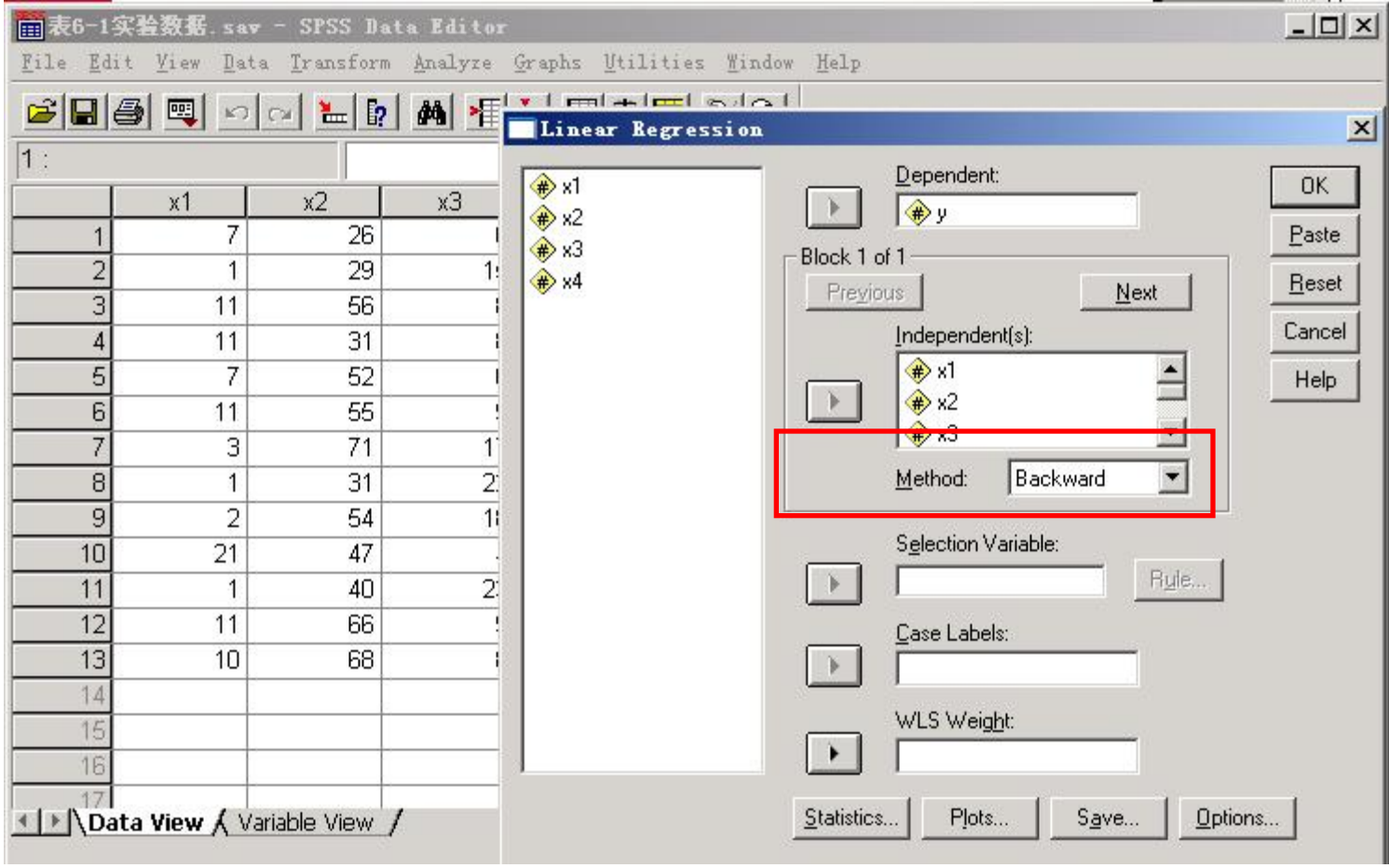
- 不显著因子的逐次剔除
  - 首先建立包含全部自变量的回归方程（15）
  - 对每一个因子做显著性检验，剔除不显著因子中偏回归平方和最小（5.33）的一个因子x_3，重新建立方程（12）（偏回归平方和越小，则剩余平方和越大，得到的方程越不可靠）
  - 对方程（12）的每一个因子做显著性检验，剔除不显著因子x_4，再重新建立方程（5）
  - 在不显著因子较少时可以采用，反之则会造成很大的工作量
  - SPSS-analyze-regression-linear-backward

一 最优回归方程的选择

全部可能的线性回归结果的比较

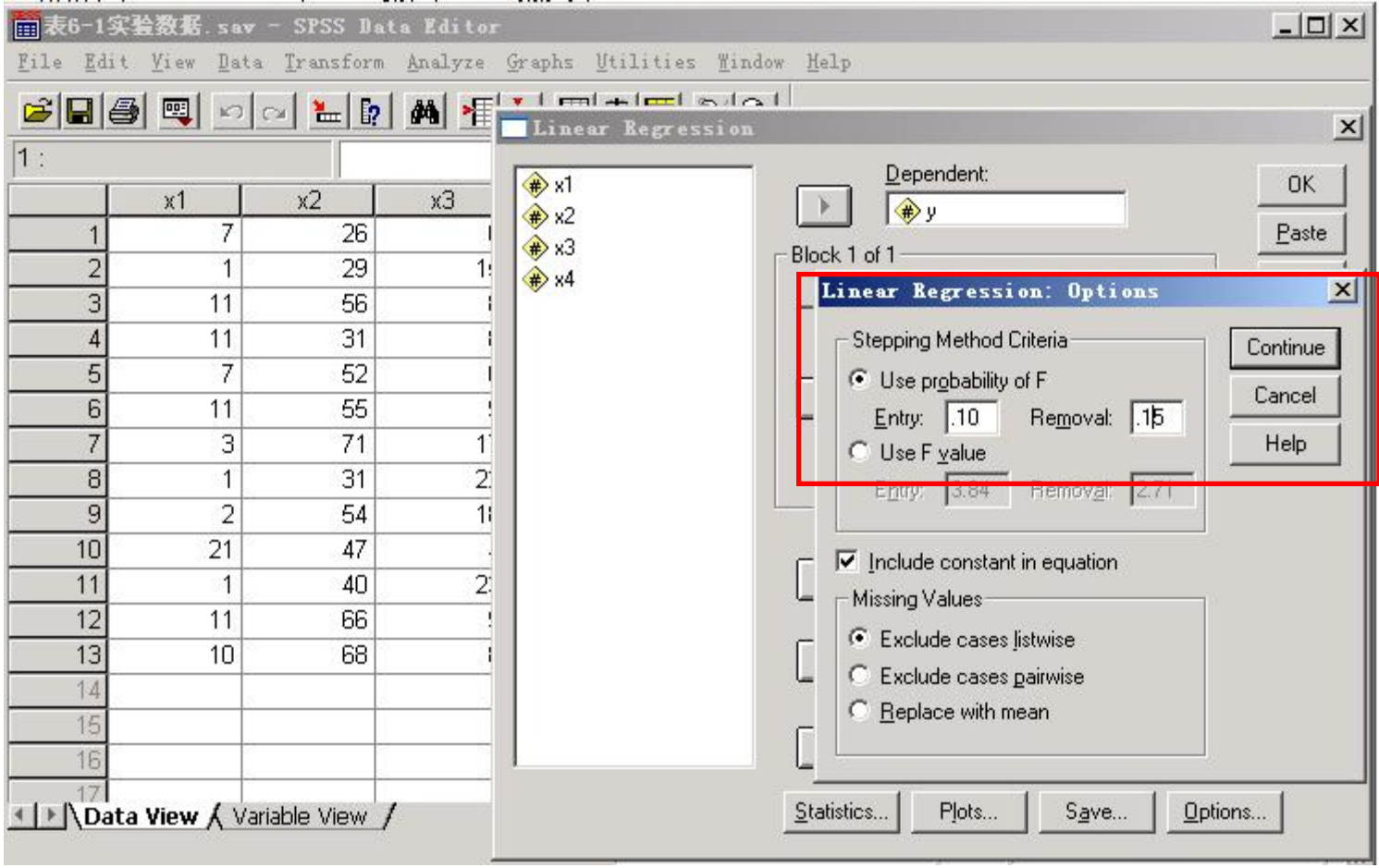
	b0	b1	b2	b3	b4	$S_{剩}$	$f_{剩}$	a^2
						2175.76	12	
(1)	81.4793	1.8687**				1265.69	11	115.06
(2)	57.4236		0.7891 **			906.34	11	82.39
(3)	110.2026			-1.2558(*)		1939.40	11	176.31
(4)	117.5679				-0.7382 **	883.87	11	80.35
(5)	52.5773	1.4683 **	0.6623 **			57.90	10	5.79
(6)	72.3490	2.3125 *		0.4945		1227.07	10	122.71
(7)	103.0973	1.4400 **			-0.6140 **	74.76	10	7.48
(8)	72.0747		0.7313 **	-1.0084 **		415.44	10	41.54
(9)	94.1600		0.3109		-0.4569	868.88	10	86.89
(10)	131.2824			-1.1999 **	-0.7246 **	175.74	10	17.57
(11)	48.1936	1.6959 **	0.6569 **	0.2500		48.11	9	5.35
(12)	71.6842	1.4519 **	0.4161(*)		-0.2365	47.97	9	5.33
(13)	111.6844	1.0519 **		-0.4100(*)	-0.6428 **	50.84	9	5.65
(14)	203.6418		-0.9234 **	-1.4480 **	-1.5570 **	73.82	9	8.20
(15)	62.4052	1.5511(*)	0.5101	0.1019	-0.1441	47.86	8	5.98

1 最优回归方程的选择



- 自变量选择x_1，x_2，x_3和x_4，因变量选择y，方法选择backward
- 点击options设置置信度

最优回归方程的选择



	x1	x2	x3
1	7	26	
2	1	29	
3	11	56	
4	11	31	
5	7	52	
6	11	55	
7	3	71	
8	1	31	
9	2	54	
10	21	47	
11	1	40	
12	11	66	
13	10	68	
14			
15			
16			
17			

- 逐步回归的原则中选择引入的置信度为0.10，移除的置信度为0.15
- 点击continue按钮

1 最优回归方程的选择

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	62.405	70.071		.891	.399
	x1	1.551	.745	.607	2.083	.071
	x2	.510	.724	.528	.705	.501
	x3	.102	.755	.043	.135	.896
	x4	-.144	.709	-.160	-.203	.844
2	(Constant)	71.648	14.142		5.066	.001
	x1	1.452	.117	.568	12.410	.000
	x2	.416	.186	.430	2.242	.052
	x4	-.237	.173	-.263	-1.365	.205
3	(Constant)	52.577	2.286		22.998	.000
	x1	1.468	.121	.574	12.105	.000
	x2	.662	.046	.685	14.442	.000

a. Dependent Variable: y

- 从剔除的过程来看，刚开始包含4个因变量，之后随着显著性检验的不断进行，先剔除x_3，之后再剔除x_4，得到最优回归模型
- 模型只包含常数项和自变量x_1和x_2

一 最优回归方程的选择

全部可能的线性回归结果的比较

	b0	b1	b2	b3	b4	$S_{剩}$	$f_{剩}$	a^2
						2175.76	12	
(1)	81.4793	1.8687**				1265.69	11	115.06
(2)	57.4236		0.7891 **			906.34	11	82.39
(3)	110.2026			-1.2558(*)		1939.40	11	176.31
(4)	117.5679				-0.7382 **	883.87	11	80.35
(5)	52.5773	1.4683 **	0.6623 **			57.90	10	5.79
(6)	72.3490	2.3125 *		0.4945		1227.07	10	122.71
(7)	103.0973	1.4400 **			-0.6140 **	74.76	10	7.48
(8)	72.0747		0.7313 **	-1.0084 **		415.44	10	41.54
(9)	94.1600		0.3109		-0.4569	868.88	10	86.89
(10)	131.2824			-1.1999 **	-0.7246 **	175.74	10	17.57
(11)	48.1936	1.6959 **	0.6569 **	0.2500		48.11	9	5.35
(12)	71.6842	1.4519 **	0.4161(*)		-0.2365	47.97	9	5.33
(13)	111.6844	1.0519 **		-0.4100(*)	-0.6428 **	50.84	9	5.65
(14)	203.6418		-0.9234 **	-1.4480 **	-1.5570 **	73.82	9	8.20
(15)	62.4052	1.5511(*)	0.5101	0.1019	-0.1441	47.86	8	5.98

一 最优回归方程的选择

- 变量逐个引入法：
  - 先计算各因子与y的相关系数，将其绝对值最大的一个因子x_4引入方程得到方程（4）（剩余均方和80.35在引入一个参数时最小）
  - 对回归平方和进行检验，结果是显著的
  - 然后找出余下的因子中与y的偏相关系数最大的那个因子x_1，（首先排除方程5-10中无x_4的方程，然后再比较剩余均方和）经检验结果显著，将其引入方程，得到方程（7）
  - 再找到余下的因子中与y偏相关系数最大的x_2，（5.33＜5.65）经检验，该因子也要引入，得到方程（12）
  - X_3经检验不显著，不再引入

一 最优回归方程的选择

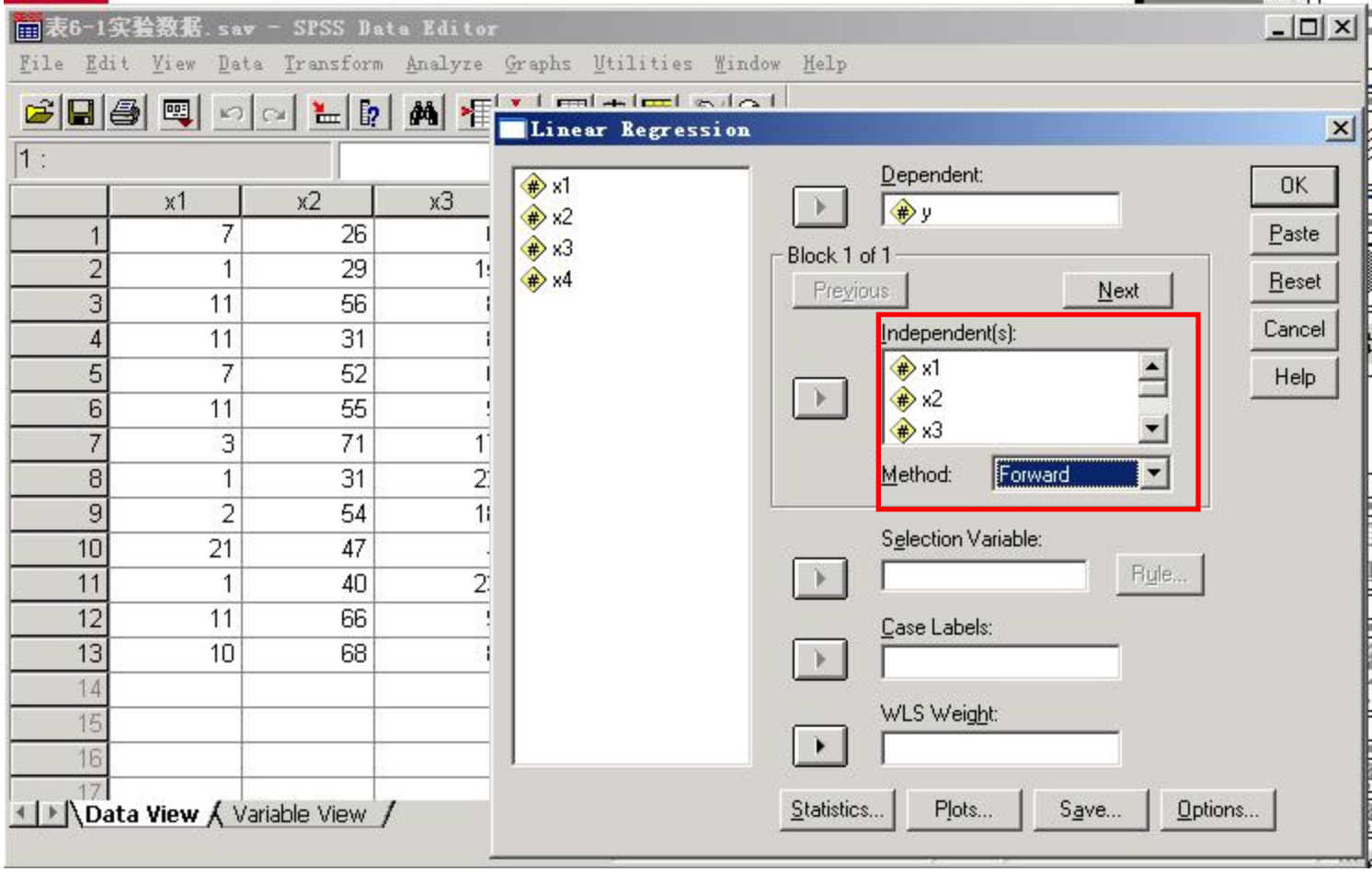
全部可能的线性回归结果的比较

	b0	b1	b2	b3	b4	$S_{剩}$	$f_{剩}$	a^2
						2175.76	12	
(1)	81.4793	1.8687**				1265.69	11	115.06
(2)	57.4236		0.7891 **			906.34	11	82.39
(3)	110.2026			-1.2558(*)		1939.40	11	176.31
(4)	117.5679				-0.7382 **	883.87	11	80.35
(5)	52.5773	1.4683 **	0.6623 **			57.90	10	5.79
(6)	72.3490	2.3125 *		0.4945		1227.07	10	122.71
(7)	103.0973	1.4400 **			-0.6140 **	74.76	10	7.48
(8)	72.0747		0.7313 **	-1.0084 **		415.44	10	41.54
(9)	94.1600		0.3109		-0.4569	868.88	10	86.89
(10)	131.2824			-1.1999 **	-0.7246 **	175.74	10	17.57
(11)	48.1936	1.6959 **	0.6569 **	0.2500		48.11	9	5.35
(12)	71.6842	1.4519 **	0.4161(*)		-0.2365	47.97	9	5.33
(13)	111.6844	1.0519 **		-0.4100(*)	-0.6428 **	50.84	9	5.65
(14)	203.6418		-0.9234 **	-1.4480 **	-1.5570 **	73.82	9	8.20
(15)	62.4052	1.5511(*)	0.5101	0.1019	-0.1441	47.86	8	5.98

一 最优回归方程的选择

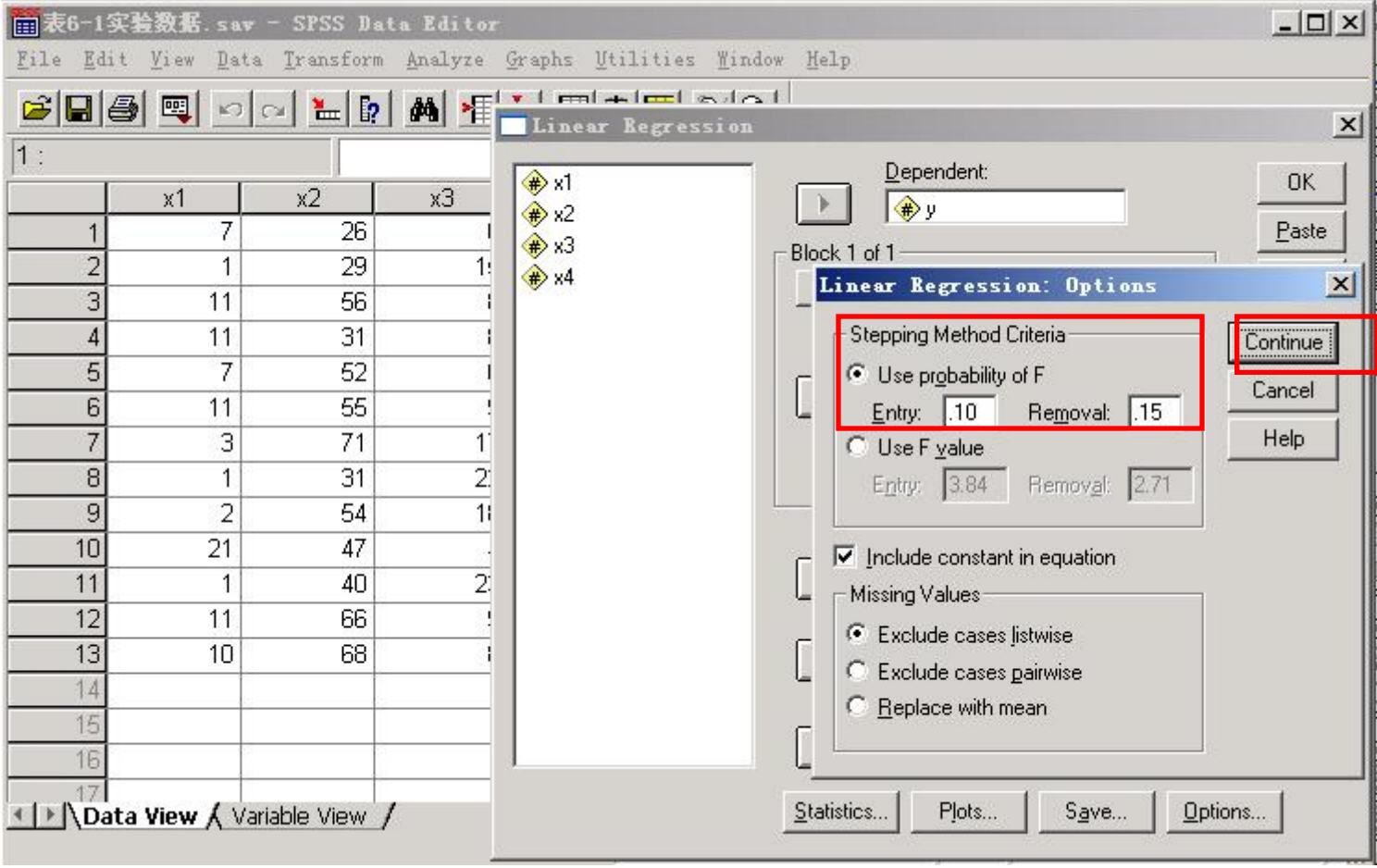
- 缺点：
  - 最后得到的方程并不是最优方程，某一个因素的引入可能会降低其他因素的显著度
  - 显著性检验可能会出现低效率的重复，需要不断的检验，剔除不显著因子
  - 引入的变量和原来存在的变量之间可能存在着一定的相关关系，当某一变量引入之后，会导致原来引入的变量显得不再重要，降低其显著程度
  - SPSS-analyze-regression-linear-forward

一 最优回归方程的选择



- 自变量选择x_1，x_2，x_3和x_4，因变量选择y，方法选择forward
- 点击options按钮设置置信度

1 最优回归方程的选择



	x1	x2	x3
1	7	26	
2	1	29	
3	11	56	
4	11	31	
5	7	52	
6	11	55	
7	3	71	
8	1	31	
9	2	54	
10	21	47	
11	1	40	
12	11	66	
13	10	68	
14			
15			
16			
17			

- 设置引入自变量的置信度为0.10，移除自变量的置信度为0.15
- 点击continue按钮

1 最优回归方程的选择

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	117.568	5.262		22.342	.000
	x4	-.738	.155	-.821	-4.775	.001
2	(Constant)	103.097	2.124		48.540	.000
	x4	-.614	.049	-.683	-12.621	.000
	x1	1.440	.138	.563	10.403	.000
3	(Constant)	71.648	14.142		5.066	.001
	x4	-.237	.173	-.263	-1.365	.205
	x1	1.452	.117	.568	12.410	.000
	x2	.416	.186	.430	2.242	.052

a. Dependent Variable: y

- 从引入的过程来看，首先引入x4，其系数均能通过显著性检验（$p \leq 0.1$），接着引入x1，也能通过显著性检验，最后引入x2，也能通过显著性检验
- 自变量x2 的引入导致x4不显著，因此得到的回归方程并不是最优的回归方程，说明x2和x4之间存在着很强的相关性

1 最优回归方程的选择

Correlations

		y	x1	x2	x3	x4
Pearson Correlatio	y	1.000	.731	.816	-.535	-.821
	x1	.731	1.000	.229	-.824	-.245
	x2	.816	.229	1.000	-.139	-.973
	x3	-.535	-.824	-.139	1.000	.030
	x4	-.821	-.245	-.973	.030	1.000
Sig. (1-tailed)	y	.	.002	.000	.030	.000
	x1	.002	.	.226	.000	.209
	x2	.000	.226	.	.325	.000
	x3	.030	.000	.325	.	.462
	x4	.000	.209	.000	.462	.
N	y	13	13	13	13	13
	x1	13	13	13	13	13
	x2	13	13	13	13	13
	x3	13	13	13	13	13
	x4	13	13	13	13	13

- 从相关系数分析表也可以看出，x_2除了和y之间的相关性比较高之外，还和x_4之间的相关性比较高，因此x_2的引入导致x_4的作用下降
- 回归的模型见方程12

一 最优回归方程的选择

	b0	b1	b2	b3	b4	$S_{剩}$	$f_{剩}$	a^2
						2175.76	12	
(1)	81.4793	1.8687**				1265.69	11	115.06
(2)	57.4236		0.7891 **			906.34	11	82.39
(3)	110.2026			-1.2558(*)		1939.40	11	176.31
(4)	117.5679				-0.7382 **	883.87	11	80.35
(5)	52.5773	1.4683 **	0.6623 **			57.90	10	5.79
(6)	72.3490	2.3125 *		0.4945		1227.07	10	122.71
(7)	103.0973	1.4400 **			-0.6140 **	74.76	10	7.48
(8)	72.0747		0.7313 **	-1.0084 **		415.44	10	41.54
(9)	94.1600		0.3109		-0.4569	868.88	10	86.89
(10)	131.2824			-1.1999 **	-0.7246 **	175.74	10	17.57
(11)	48.1936	1.6959 **	0.6569 **	0.2500		48.11	9	5.35
(12)	71.6842	1.4519 **	0.4161(*)		-0.2365	47.97	9	5.33
(13)	111.6844	1.0519 **		-0.4100(*)	-0.6428 **	50.84	9	5.65
(14)	203.6418		-0.9234 **	-1.4480 **	-1.5570 **	73.82	9	8.20
(15)	62.4052	1.5511(*)	0.5101	0.1019	-0.1441	47.86	8	5.98

一 最优回归方程的选择

- 逐步回归分析
  - 将因子一个个引入，引入因子的条件是该因子的偏回归平方和在没有进入方程的其余因子当中为最大，而且经过检验是显著的
  - 每引入一个新因子之后，在新的方程的基础上，再在已进入方程的因子中找出偏回归平方和最小的一个并作检验，如不显著则将剔除
  - 每引入一个变量或剔除一个变量前后，就要做F检验，直到最后没有显著的变量可以引入，也没有不显著的变量需要剔除为止

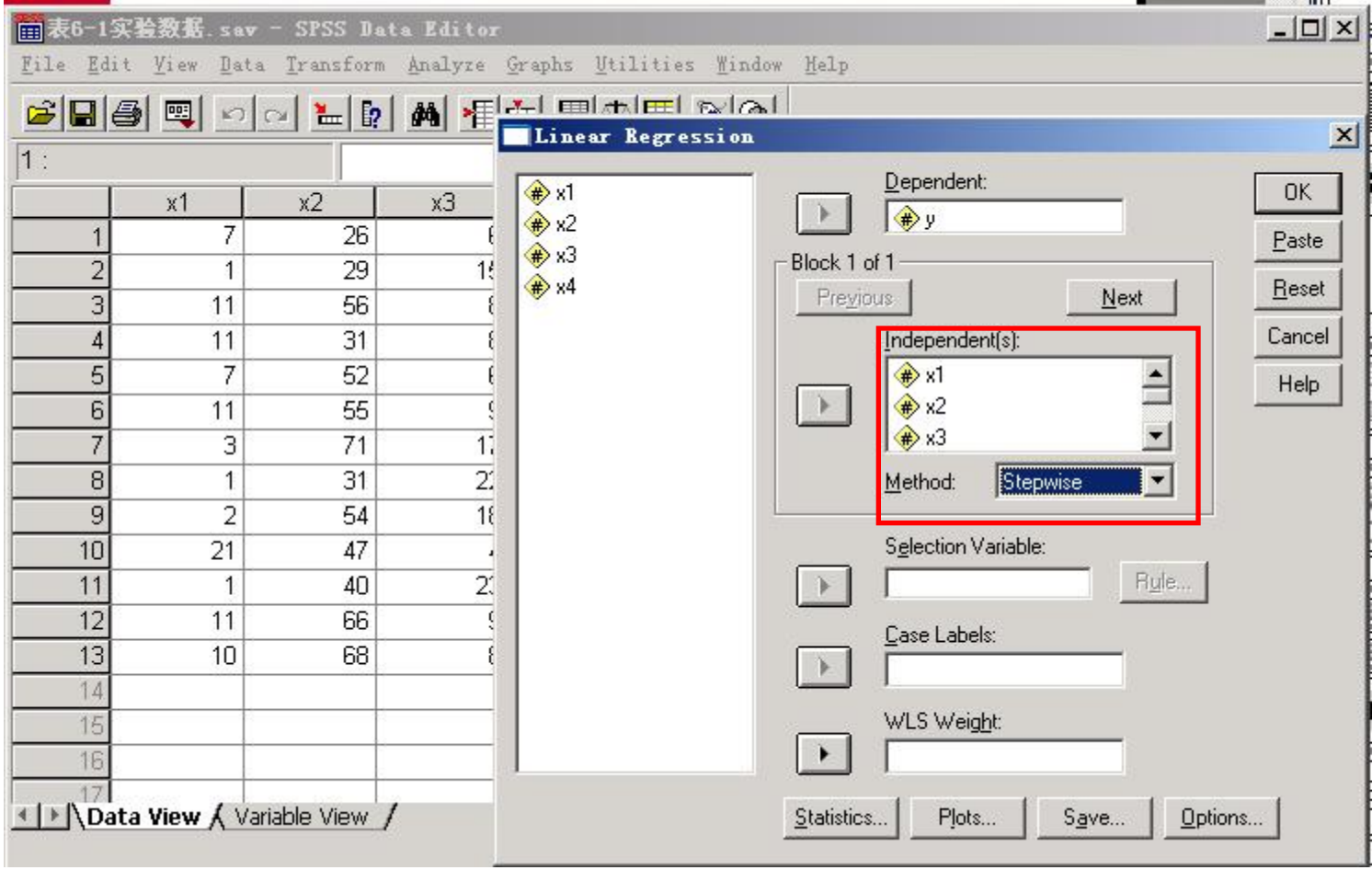
一 最优回归方程的选择

- **逐步回归基本原理**
  - 边引入，边剔除
  - 引入前显著性检验，剔除前显著性检验
  - 直到没有显著的变量可以引入，没有不显著的变量可以剔除

1 最优回归方程的选择

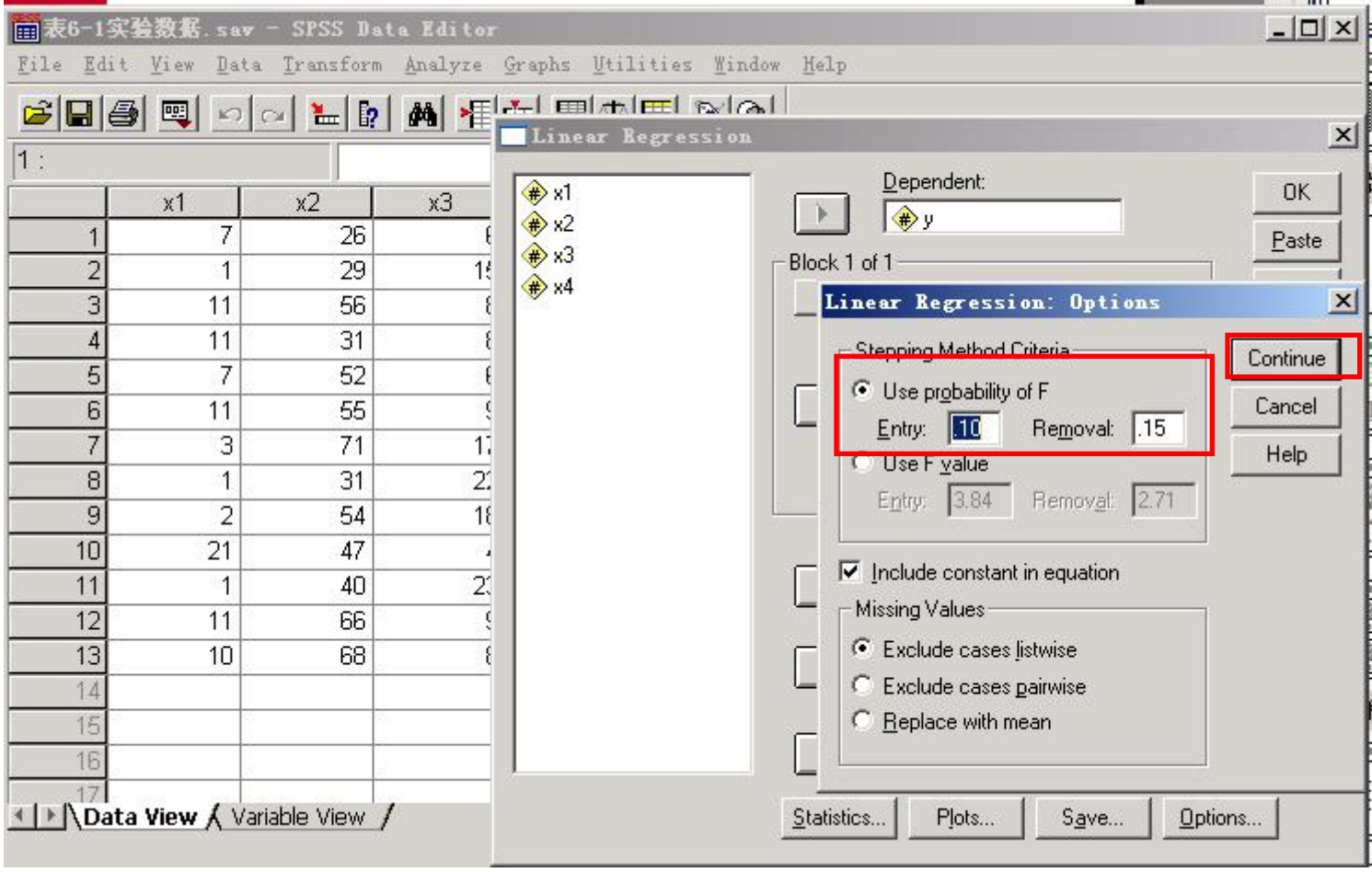
- 逐步回归在现实生活中的应用
  - 岗位的优胜劣汰：通过不断的引人，是集团内部的效益达到最大化
  - 达尔文的生物进化论：物种的侵入可能会取代原有物种在栖息地上的地位，逐步淘汰不适应的物种，典型如物种对全球变化的响应

1 最优回归方程的选择



- 选择x_1，x_2，x_3和x_4作为自变量引入，y作为因变量引入，方法选择stepwise
- 点击options选择置信度

最优回归方程的选择



- 选择引入的置信度为0.10，移除的置信度为0.15
- 点击continue按钮

1 最优回归方程的选择

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	117.568	5.262		22.342	.000
	x4	-.738	.155	-.821	-4.775	.001
2	(Constant)	103.097	2.124		48.540	.000
	x4	-.614	.049	-.683	-12.621	.000
	x1	1.440	.138	.563	10.403	.000
3	(Constant)	71.648	14.142		5.066	.001
	x4	-.237	.173	-.263	-1.365	.205
	x1	1.452	.117	.568	12.410	.000
	x2	.416	.186	.430	2.242	.052
4	(Constant)	52.577	2.286		22.998	.000
	x1	1.468	.121	.574	12.105	.000
	x2	.662	.046	.685	14.442	.000

a. Dependent Variable: y

- 首先引入x_4，然后引入x_1，最后引入x_2，发现x_2的引入导致x_4不显著，即通不过t检验，而且达到了移除的标准（$p=0.205\geq0.15$）
- 将x_4剔除，得到最优回归模型为4，即只包含常数项，x_1和x_2，且所有自变量系数均能够通过显著性检验

2 逐步回归的数学模型-自学

对变量重新编号

在逐步回归分析中，常令自变量个数K=n-1并记$y_\alpha=x_{\alpha n}$，这时的数学模型为

$$x_{\alpha n}=\beta_0+\beta_1 x_{\alpha 1}+\beta_1 x_{\alpha 1}+\ldots+\beta_1 x_{\alpha 1}+\varepsilon_\alpha,\alpha=1,2,\ldots,N$$

在这种编号下，各类偏差平方和分别为

$$\begin{cases} S_{总}=\sum_{\alpha}(x_{\alpha n}-\bar{x}_n)^2 \\ S回=\sum_{\alpha}(\hat{x}_{\alpha n}-\bar{x}_n)^2 \\ S剩=\sum_{\alpha}(x_{\alpha n}-\hat{x}_{\alpha n}) \\ x_j\text{的偏回归平方和：}Q_j=S_{回}-S_{回}'=b_j^2/C_{jj} \end{cases}$$

其中x_n为地理要素观测值x_{1n}，$x_{2n},\ldots x_{Nn}$的算术平均数，$\hat{x}_{\alpha n}$为所求得的回归方程$x_n=b_0+b_1x_1+b_2x_2+\ldots+b_{n-1}x_{n-1}$的第$\alpha$个回归值，$C_{jj}$为相关矩阵$L^{-1}$矩阵对角线上的元素

2 逐步回归的数学模型-自学

- **对数据进行标准化变换**

由于不同的$x_j(j=1,2,...n)$，它们的取值范围和单位都不相同，所以为了再无量纲影响下进行计算，我们在逐步回归中常将所有数据先进行“标准化”变换，即令

$$z_{aj}=\frac{x_{aj}-\bar{x}_j}{\sigma_j},\ j=1,2,...,n$$

其中$x_j=\frac{1}{N}\sum_{\alpha}x_{\alpha j},\sigma_j^{\ 2}=l_{jj}=\sum_{\alpha}(x_{aj}-\bar{x}_j)^2$。在变换（6-4）下，其数学模型为

$$z_{\alpha n}=\beta_0'+\beta_1'z_{\alpha 1}+\beta_2'z_{\alpha 2}+...+\beta_{n-1}'z_{\alpha,n-1}+\varepsilon$$

2 逐步回归的数学模型-自学

对数据进行标准化变换

这时，模型6-5的结构矩阵X与观察值的矩阵Y分别为

$$\begin{bmatrix} 1 & \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{1,n-1}-\bar{x}_{n-1}}{\sigma_{n-1}} \\ 1 & \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{2,n-1}-\bar{x}_{n-1}}{\sigma_{n-1}} \\ \cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \\ 1 & \frac{x_{N1}-\bar{x}_1}{\sigma_1} & \frac{x_{N2}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{N,n-1}-\bar{x}_{n-1}}{\sigma_{n-1}} \end{bmatrix}$$

$$y=\begin{bmatrix} \frac{x_{1n}-\bar{x}_n}{\sigma_n} \\ \frac{x_{2n}-\bar{x}_n}{\sigma_n} \\ \cdots\cdots \\ \frac{x_{Nn}-\bar{x}_n}{\sigma_n} \end{bmatrix}$$

2 逐步回归的数学模型-自学

而系数矩阵A与常数项矩阵B分别为

$$A=X'X=\begin{bmatrix} N & 0 & 0 & \cdots & 0 \\ 0 & r_{11} & r_{12} & \cdots & r_{1,n-1} \\ 0 & r_{21} & r_{22} & \cdots & r_{2,n-1} \\ \cdots & \cdots & \cdots & & \cdots \\ 0 & r_{n-1,1} & r_{n-1,2} & \cdots & r_{n-1,n-1} \end{bmatrix} = \begin{bmatrix} N & 0 \\ 0 & R \end{bmatrix}$$

$$B=X'Y=\begin{bmatrix} 0 \\ r_{1n} \\ r_{2n} \\ \cdots \\ r_{n-1,n} \end{bmatrix}$$

其中$r_{ij}(i, j=1, 2, \ldots, n)$为变量$x_i$和$x_j$的相关系数

$$r_{ij}=\frac{\sum_{\alpha}(x_{\alpha i}-x_i)(x_{\alpha j}-x_j)}{\sigma_i \sigma_j}$$

而R是系数矩阵A右下角的$n-1$阶对称方阵，它的元素都是变量x_1，x_2，$\ldots x_{n-1}$间的相关系数，故亦称R为相关系数矩阵

2 逐步回归的数学模型-自学

在变换$6-4$下模型$6-5$中的常数项β_0'的估计值$\hat{\beta}_0'=0$，这是因为

$$\hat{\beta}_0' = Z_n - \sum_{j=1}^{n-1} \hat{\beta}_0' Z_j$$

$$= 0 - \sum_{j=1}^{n-1} \hat{\beta}_j' 0$$

$$= 0$$

由此，模型$6-5$可改写为$z_{\alpha n} = \beta_1' z_{\alpha 1} + \beta_2' z_{\alpha 2} + \ldots + \beta_{n-1}' z_{\alpha,n-1} + \varepsilon_\alpha$

它的系数矩阵就是相关系数矩阵R，它的常数项矩阵就是从B中把第一个元素0剔除后的矩阵，即

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1,n-1} \\ r_{21} & r_{22} & \cdots & r_{2,n-1} \\ \cdots & \cdots & & \cdots \\ r_{n-1,1} & r_{n-1,2} & \cdots & r_{n-1,n-1} \end{bmatrix}$$

$$B=\begin{bmatrix} r_{1n} \\ r_{2n} \\ \cdots \\ r_{n-1,n} \end{bmatrix}$$

2 逐步回归的数学模型-自学

模型6－6与模型6－1回归系数之间的关系　设由模型6-6求得的回归方程为

$z_n = d_1 z_1 + d_2 z_2 + \ldots + d_{n-1} z_{n-1}$

将变换6－4代入，即得

$$\frac{x_n - \bar{x}_n}{\sigma_n} = d_1 \frac{x_1 - \bar{x}_1}{\sigma_1} + d_2 \frac{x_2 - \bar{x}_2}{\sigma_2} + \ldots + d_{n-1} \frac{\hat{x}_{n-1} - \bar{x}_{n-1}}{\sigma_{n-1}}$$

或$$\hat{x}_n = \left(\bar{x}_n - \frac{\sigma_n}{\sigma_1} d_1 \bar{x}_1 - \frac{\sigma_n}{\sigma_2} d_2 \bar{x}_2 - \ldots - \frac{\sigma_n}{\sigma_{n-1}} d_{n-1} \bar{x}_{n-1} \right) + \frac{\sigma_n}{\sigma_1} d_1 \bar{x}_1 + \frac{\sigma_n}{\sigma_2} d_2 \bar{x}_2 + \ldots + \frac{\sigma_n}{\sigma_{n-1}} d_{n-1} \bar{x}_{n-1}$$

比较6－3和6－8可得

$$\begin{cases} b_j = \dfrac{\sigma_n}{\sigma_j} d_j, \; j = 1, 2, \ldots, n-1 \\ b_0 = x_n - \sum_{j=1}^{n-1} b_j \bar{x}_j \end{cases}$$

因而只要能求得6－7式中的d_j，也就能求得6－3中的b_j，今后讨论将只在模型6-6下进行

2 逐步回归的数学模型-自学

在模型6－6下的各种平方和的计算

在模型6－6下总的偏差平方和、回归平方和、剩余平方和、z_j的偏回归平方和分别记为$\%S_{总}$、$\%S_{回}$、$\%S_{剩}$、V_j，注意到$z_j=\frac{1}{N}\sum_{\alpha} z_{\alpha j}=0, j=1,2,...,n$，立即可得

$$\begin{cases} \%S_{总}=\dfrac{1}{\sigma_n^{\,2}}S_{总}=1 \\ \%S_{回}=\dfrac{1}{\sigma_n^{\,2}}S_{回} \\ \%S_{剩}=\dfrac{1}{\sigma_n^{\,2}}S_{剩} \\ V_j=\dfrac{1}{\sigma_n^{\,2}}Q \end{cases}$$

譬如$\%S_{总}=\sum_{\alpha}(z_{\alpha n}-\bar{z}_n)=\sum_{\alpha} z_{an}^{\,2}=\sum_{\alpha}\left(\dfrac{x_{an}-\bar{x}_n}{\sigma_n}\right)^2=1$

3 逐步回归分析的步骤-自学

为了求解求逆计算回归系数，我们将系数矩阵R与常数项矩阵B放在一起组成组成增广矩阵，同时为了检验的方便，又在此矩阵中添上一行（$r_{n1}, \ldots, r_{nn}$），从而组成一个方阵，基座$R^{(0)}$：

$$R^{(0)} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1,n-1} & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2,n-1} & r_{2n} \\ \cdots & \cdots & & \cdots & \cdots \\ r_{n-1,1} & r_{n-1,2} & \cdots & r_{n-1,n-1} & r_{n-1,n} \\ r_{n1} & r_{n2} & \cdots & r_{n,n-1} & r_{nn} \end{bmatrix} = \begin{bmatrix} R & B \\ B' & r_{nn} \end{bmatrix}$$

$R^{(0)}$中所有元素都是相关系数，也称相关系数矩阵。下面我们就利用$R^{(0)}$来进行逐步回归分析

3 逐步回归分析的步骤-自学

首先从n-1个因子$z_1,\ldots,z_{n-1}$中引入一个因子z_{k_1}，建立一元线性回归方程。

显然，这z_{k_1}是所有因子中偏回归平方和（或称方差贡献）为最大的因子，因为这能使剩余平方和减少的最多

为了计算各个一元线性回归方程中因子的偏回归平方和，就要计算各回归方程$Z_n=d_j^{(1)}Z_j,\ j=1,2,\ldots,n-1$

的系数$d_j^{(1)}$（上标（1）表示引入的第一个变量）及相应的正规方程系数矩阵R的逆矩阵C，由求解求逆变换的性质可知，这只要对$R^{(0)}$作消去Z_j的变换即可。

3 逐步回归分析的步骤-自学

记$R^{(0)}$经过求解求逆变换L_j后所得的矩阵中的元素为$r_{uv}^{(1j)}$（上式1j中的1表示对矩阵$R^{(0)}$的第一次变换，j表示消去Z_j），则

$$d_j^{(1)} = r_{jn}^{(1j)} = \frac{r_{jn}}{r_{jj}}$$

$$c_{jj}^{(1)} = r_{jj}^{(1j)} = \frac{1}{r_{jj}}$$

在z_j的一元线性回归方程中，z_j的偏回归平方和

$$V_j^{(1)} = \frac{[d_j^{(1)}]^2}{c_{jj}^{(1)}} = \frac{r_{jn}^{\ 2}}{r_{jj}}$$

上式表明$V_j^{(1)}$完全可用$R^{(0)}$中的元素表示。

3 逐步回归分析的步骤-自学

从$V_j^{(1)}(j=1,2,\ldots n-1)$中找出

最大的一个，记为$V_{k_1}^{(1)}$，即$V_{k_1}^{(1)}=\max\limits_{1\leqslant j\leqslant n-1} V_j^{(1)}$

并对其做显著性检验，由于回归方程$Z_n=d_{k_1}^{(1)}Z_{k_1}$的剩余平方和

$S_{剩}^{(1)}=S_{总}-V_{k_1}^{(1)}=r_{nn}-V_{k_1}^{(1)}=1-V_{k_1}^{(1)}$

其自由度$f_{剩}^{(1)}$=N-1-因子数=N-2(N为样本数)，于是F统计量为

$$F_1=\frac{V_{k_1}^{(1)}}{S_{剩}^{(1)}/f_{剩}^{(1)}}=\frac{V_{k_1}^{(1)}}{r_{nn}-V_{k_1}^{(1)}}(N-2)$$

为了区别是引入因子时作的检验，还是剔除因子时做的检验，将引入时做检验的F统计量记为F_1，剔除时记为F_2。当$F_1>F_a$（1，N-2）时，可将变量Z_{k_1}引入，并对$R^{(0)}$做求解求逆变换L_{k_1}得到$R^{(1)}=L_{k_1}R^{(0)}$

3 逐步回归分析的步骤-自学

求解求逆紧凑变换的公式为：

$$L_{k_l}:\begin{cases} r_{ij}^{(l)} = r_{ij}^{(l-1)} - r_{ik_l}^{(l-1)} r_{k_l j}^{(l-1)} / r_{k_l k_l}^{(l-1)} i, j \neq k_l \\ r_{k_l j}^{(l)} = r_{k_l j}^{(l-1)} / r_{k_l k_l}^{(l-1)}, j \neq k_l \\ r_{ik_l}^{(l)} = -r_{ik_l}^{(l-1)} / r_{k_l k_l}^{(l-1)}, i \neq k_l \\ r_{k_l k_l}^{(l)} = 1 / r_{klkl}^{(l-1)} \end{cases}$$

上式当$l=1$时即为变换l_{k_1}

由于刚刚引入第一个因子，而且回归方程中没有别的变量，因此不需要再做是否要剔除的检验。如$F_1 < F_a$，表示在这一组变量中，没有一个与Z_n有显著的线性关系，建立回归方程工作就此结束

3 逐步回归分析的步骤-自学

继续挑选第二个因子，同样要求该因子是二元线性回归方程中的偏回归平方和，是所有因子（除Z_{k_1}外）中最大

在第一步中，我们已得到$R^{(1)}=L_{k_1}R^{(0)}$，和前面的道理一样，我们同样希望用$R^{(1)}$中的元素来求得二元回归中的偏回归平方和

记$R^{(1)}$经过L_j变换$(j\neq k_1)$的偏回归平方和$V_j^{(2)}$为

$$V_j^{(2)}=\frac{[d_j^{(2)}]^2}{c_{jj}^{(2)}}=\frac{[r_{jn}^{(1)}]^2}{r_{jj}^{(1)}}$$

记$V_{k_2}^{(2)}=\max_{j\neq k_1}V_j^{(2)}$

因子Z_{k_2}要不要引入回归方程，要由对$V_{k_2}^{(2)}$做F检验后决定。对于回归方程

$Z_n=d_{k_1}^{(2)}Z_{k_1}+d_{k_2}^{(2)}Z_{k_2}$

来讲，其剩余平方和$S_{剩}^{(2)}=S_{剩}^{(1)}-V_{k_2}^{(2)}$

3 逐步回归分析的步骤-自学

由于$S_{剩}^{(1)}=r_{nn}-V_{k_1}^{(1)}=r_{nn}-\frac{(r_{k_1n})2}{r_{k_1k_1}}=r_{nn}-\frac{r_{k_1n}r_{nk_1}}{r_{k_1k_1}}=r_{nn}^{(1)}$

其中倒数第二个等号，是由于$r_{k_1n}=r_{nk_1}$，最后一个等号是由上述L_{k_1}求解求逆变换的公式所决定的。于是$S_{剩}^{(2)}=r_{nn}^{(1)}-V_{k_2}^{(2)}$

由于$f_{剩}^{(2)}=N-3$

故这是F统计量$F_1=\frac{V_{k_2}^{(2)}}{S_{剩}^{(2)}/(N-3)}=\frac{V_{k_2}^{(2)}}{r_{nn}^{(1)}-V_{k_2}^{(2)}}(N-3)$

当$F_1>F_\alpha(1,N-3)$时，因子Z_{k_2}应该引入，否则就不引入，逐步回归分析结束。

如$F_1>F_\alpha$，则对$R^{(1)}$施行变换L_{k_2}，得到$R^{(2)}=L_{k_2}R^{(1)}$。

3 逐步回归分析的步骤-自学

当引子Z_{k_2}引入后，应该检验Z_{k_1}要不要剔除。这要对Z_{k_1}在二元回归方程

$Z_n=d_{k_1}^{(2)}Z_{k_1}+d_{k_2}^{(2)}Z_{k_2}$

中的偏回归平方和做F检验。

记$R^{(2)}=L_{k_2}R^{(1)}$，由变换的性质知道

$d_{k_1}^{(2)}=r_{k_1 n}^{(2)}$　　$C_{k_1k_1}^{(2)}=r_{k_1k_1}^{(2)}$

故6-11中Z_{k_1}的偏回归平方和$V_{k_1}^{(2)}=\dfrac{[d_{k_1}^{(2)}]^2}{C_{k_1k_1}^{(2)}}=\dfrac{[r_{k_1n}^{(2)}]^2}{r_{k_1k_1}^{(2)}}$

3 逐步回归分析的步骤-自学

6－11式的剩余平方和

$$S_{剩}^{(2)}=r_{nn}^{(1)}-V_{k_2}^{(2)}=r_{nn}^{(1)}-\frac{[r_{k_2n}^{(1)}]^2}{r_{k_2k_2}^{(1)}}=r_{nn}^{(1)}-\frac{r_{nk_2}^{(1)}r_{k2n}^{(1)}}{r_{k_2k_2}^{(1)}}=r_{nn}^{(2)}$$

$$f_{剩}^{(2)}=N-1-2=N-3$$

于是检验Z_{k_1}是否显著的统计量为：$F_2=\dfrac{V_{k_1}^{(2)}}{S_{剩}^{(2)}/f_{剩}^{(2)}}=\dfrac{[r_{k_1n}^{(2)}]^2}{r_{nn}^{(2)}}(N-3)$

当$F_2>F_\alpha(1,N-3)$时，因子Z_{k_1}不要剔除，接下去再考虑选入第三个因子Z_{k_2}；

当$F\leqslant F_\alpha$（1，N-3）时，因子Z_{k_1}应剔除，这时需对$R^{(2)}$作变换L_{k_1}，得$R^{(3)}$，从而建立只包括Z_{k_2}的回归方程，然后再考虑引入新的因子。

3 逐步回归分析的步骤-自学

重复步骤2、3，一般地讲，在建立了l个因子Z_{k_1}，Z_{k_2}，...，Z_{k_l}的回归方程

$$\hat{Z}_n = d_{k_1} Z_{k_1} + d_{k_2} Z_{k_2} + ... + d_{k_l} Z_{k_l}$$

之后，假定这时$R^{(0)}$经过l次变换成了$R^{(l)}$，那么由于新因子Z_{k_l}的引入，首先要对原有的因子Z_{k_1}，Z_{k_2}，...，$Z_{k_{l-1}}$做剔除的F检验，即对其中偏回归平方和最小的变量，设为Z_{k_j} $(1 \leqslant j \leqslant l-1)$做检验，如$F_2 > F_\alpha$，则无需剔除，可进一步引入新的变量；如果$F_2 < F_\alpha$，则$Z_{k_j}$应予以剔除，并对$R^{(l)}$进行$L_{k_j}$变换，得到$R^{(l+1)}$。

必须注意，这时还不能引入变量，而应该重复做剔除检验，直到没有要剔除的变量为止，才可以引入变量。理论和实际计算都证明可能相继地剔除几个自变量。具体步骤如下：

3 逐步回归分析的步骤-自学

(1)对$t=1,2,...,l-1,$计算因子Z_{k_t}的偏回归平方和

$$V_{k_t}^{(l)}=\frac{[r_{k_t n}^{(l)}]^2}{r_{k_t k_t}^{(l)}}$$

从中找出$V_k^{(l)}=\min\limits_{1\leqslant i\leqslant l-1} V_{k_t}^{(l)}$

对因子Z_k做F检验，由于$S_{剩}^{(l)}=r_{nn}^{(1)}$

$$f_{剩}^{(l)}=N-l-1$$

所以$F_2=\frac{V_k^{(l)}}{r_{nn}^{(l)}}(N-l-1)$

当$F_2<F_\alpha(1,N-l-1)$时，对$R^{(l)}$作变换L_k，得到$R^{(l+1)}=L_k R^{(l)}$。然后重复上述步骤，对因子$Z_{k_t}(k_t\neq k)$重做检验，直到没有因子要剔除时为止；

当$F_2>F_C(1,N-l-1)$时，下一步考虑引入新变量。

3 逐步回归分析的步骤-自学

（2）设经过以上剔除检验后，回归方程中包含$x_{k_1}, x_{k_2}, ..., x_{k_l}$共$l$个因子，系数矩阵为$R^{(l)}$。对$j \neq k_1, k_2, ...k_l$的因子$Z_j$，计算偏回归平方和

$$V_j^{(l+1)} = \frac{(r_{jn}^{(l)})^2}{r_{jj}^{(l)}}$$

从中找出最大值$V_{k_{l+1}}^{(l+1)} = \max_{j \neq k_1, ...k_l} V_j^{(l+1)}$

此时$S_{剩}^{(l+1)} = S_{剩}^{(l)} - V_{k_{l+1}}^{(l+1)} = r_{nn}^{(1)} - V_{k_{l+1}}^{(l+1)}$

对因子$Z_{k_{l+1}}$做F检验$F_1 = \frac{V_{k_{l+1}}^{(l+1)}}{r_{nn}^{(l)} - V_{k_{l+1}}^{(l+1)}}(N-l-2)$

3 逐步回归分析的步骤-自学

当$F_1 > F_\alpha(1, N-l-2)$时，引入因子$Z_{k_{l+1}}$，对$R^{(1)}$做变换$L_{k_{l+1}}$,得到$R^{(l+1)} = L_{k_{l+1}} R^{(l)}$。然后进行剔除检验，即转到第一步。

当$F_1 < F_\alpha(1, N-l-2)$时，挑选因子结束。这时就建立了l元线性回归方程6-12，其回归系数为$d_{k_t}^{(l)} = r_{k_t n}^{(l)}, t = 1, 2, ..., l$

回归方程的均方差估计值 $\hat{\sigma} = \sqrt{\frac{S_{剩}^{(l)}}{N-l-1}} = \sqrt{\frac{r_{nn}^{(l)}}{N-l-1}}$

回归方程6-12的复相关系数$R = \sqrt{1 - r_{nn}^{(l)}}$

然后使用$6-9$式求出模型$6-1$下的回归系数b_0, b_j建立回归方程

$x_n = b_0 + b_1 x_1 + ... + b_{n-1} x_{n-1}$

4 逐步回归在地理系统分析中的应用实例

- 我们以台风暴雨的逐步回归为例，说明地理要素进行逐步回归分析的过程
- 设研究变量y为我国东南沿海地区一次登陆台风所造成的24h暴雨量；经分析，找出影响暴雨量的主要因素有：x_1，x_2，x_3，x_4，x_5，x_6，x_7，具体资料见表6-3

逐步回归在地理系统分析中的应用实例

计算步骤：

（1）首先规定一个F^*值，作为F检验用的临界值。按理每一步F_α值均不同

F_1：$F(1, N-1-1)$

F_2：$F(l, N-l-1)$

纠错

其中l为选上的因子的个数，但由于在一般情况下N？l，所以对于给定的显著性水平α来讲，所有的$F_\alpha(1, N-1-1)$和$F_\alpha(l, N-l-1)$都近似相等。故为方便期间，我们可取一个定数F^*作为F检验的标准。为了使最终的回归方程中包含较多的变量，F水平不宜取得过高。根据原始数 据的个数N及估计可能选入方程的变量个数l，按$N-l-1$计算自由度。由显著性水平α及自由度f值，查表决定F^*值。本例取$F^*=3.5$

逐步回归在地理系统分析中的应用实例

- 实际应用过程中一般采用置信度，比如引入的置信度设置为0.05，移除的置信度设置为0.10，或者引入的置信度设置为0.10，移除的置信度设置为0.15，注意移除的变量永远>引入的变量所设置的置信度，否则模型将会陷入死循环
- F水平在研究精度要求不高时，可以适当放宽，一般引入变量的最高设置为0.10-0.15之间，不易过大，否则容易引入过多的无谓变量，但不能小于0.05，有可能导致一个变量都引入不进，地理模型不同于数学模型

4 逐步回归在地理系统分析中的应用实例

（2）计算$R^{(0)}$:

$$r_{ij}=\frac{l_{ij}}{\sigma_i\sigma_j}=\frac{\sum_{\alpha}(x_{\alpha i}-\bar{x}_i)(x_{\alpha i}-\bar{x}_j)}{\sigma_i\sigma_j},\text{其中}\sigma_j=\sqrt{l_{ji}}$$

先计算x_i与σ_j，然后算得$R^{(0)}$。

$$r_{21}=\frac{l_{21}}{\sigma_2\sigma_1}=\frac{\sum_{\alpha}(x_{\alpha 2}-\bar{x}_2)(x_{\alpha 1}-\bar{x}_1)}{\sigma_2\sigma_1}(i=2,j=1)$$

$$r_{12}=\frac{l_{12}}{\sigma_1\sigma_2}=\frac{\sum_{\alpha}(x_{\alpha 1}-\bar{x}_1)(x_{\alpha 2}-\bar{x}_2)}{\sigma_1\sigma_2}(i=1,j=2)$$

- **具体参见表6-4（p126）**
- **数据来源来自于表6-3（p125：当a=1时，为第一行第二列数据与第二列平均值之差乘以第一行第一列数据与第一列平均值之差加上a=2时，第二行第二列数据与第二列平均值之差乘以第二行第一列数据与第一列平均值之差加上a=n。。。）**

4 逐步回归在地理系统分析中的应用实例

（3）选因子

选第一个因子　对j=1, . . . , 7, 计算偏回归平方和 $V_j^{(1)}=\dfrac{(r_{jn}^{(0)})^2}{r_{jj}^{(0)}}$,由 $R^{(0)}$得到

$$V_1^{(1)}=\frac{(r_{1n}^{(0)})^2}{r_{11}^{(0)}}=\frac{(0.4208)^2}{1}=0.1771;V_2^{(1)}=\frac{(r_{2n}^{(0)})^2}{r_{22}^{(0)}}=\frac{(-0.1003)^2}{1}=0.0100$$

$$V_3^{(1)}=\frac{(r_{3n}^{(0)})^2}{r_{33}^{(0)}}=\frac{(-0.2733)^2}{1}=0.0747;V_4^{(1)}=\frac{(r_{4n}^{(0)})^2}{r_{44}^{(0)}}=\frac{(0.0015)^2}{1}=0.0000$$

$$V_5^{(1)}=\frac{(r_{5n}^{(0)})^2}{r_{55}^{(0)}}=\frac{(0.1528)^2}{1}=0.0233;V_6^{(1)}=\frac{(r_{6n}^{(0)})^2}{r_{66}^{(0)}}=\frac{(-0.3534)^2}{1}=0.1249$$

$$V_7^{(1)}=\frac{(r_{7n}^{(0)})^2}{r_{77}^{(0)}}=\frac{(0.1670)^2}{1}=0.0279$$

x_1的$V_1^{(1)}$最大，即 $\max V_j^{(1)}=V_1^{(1)}=0.1771$

作F检验，$F_1=\dfrac{V_1^{(1)}(N-2)}{r_{nn}-V_1^{(1)}}=\dfrac{0.1771}{1-0.1771}\times 27=5.81>3.5$

所以引进x_1，其相关系数为0.4208。然后对$R^{(0)}$以$r_{11}^{(0)}$为主元进行L_1变换，得到$R^{(1)}$

4 逐步回归在地理系统分析中的应用实例

$R^{(1)}$是基于求逆紧凑变换公式及$R^{(0)}$得到，其中求逆紧凑变换k_1为对应的已经引入自变量的下标，比如第一次引进x_1，则$k_1=1$然后根据i, j, k之间的关系按照相应的公式进行求解，比如r_{23},则$i, j \neq k_1$,即选用求逆紧凑变换公式1求解，而r_{11}中，$i, j = k_1$，即选用求逆紧凑变换公式4求解。

选第二个因子，对$j \neq 1$,计算偏回归平方和$V_j^{(2)} = \dfrac{(r_{jn}^{(1)})^2}{r_{jj}^{(1)}}$,选出$x_6$的

偏回归平方和最大，即$\max\limits_{j\neq 1} V_j^{(2)} = V_6^{(2)} = \dfrac{(-0.3204)^2}{0.9938} = 0.1033$

- **求逆紧凑变换公式具体见P122&P121**

4 逐步回归在地理系统分析中的应用实例

即还是选出去掉x_1代表的第一行最后一列系数后，比较除最后一行之外的其它行最后一列系数的平方与对应的对角线值的商的大小，选择最大的偏回归平方和

做F检验$F_1 = \dfrac{V_6^{(2)}(N-3)}{r_{nn}^{(1)} - V_6^{(2)}} = \dfrac{0.1033}{0.8228 - 0.1033} \times 26 = 3.73 > 3.5$

所以可引入x_6，此时回归方程的复相关系数为$R = 0.530$.然后对$R^{(1)}$以$r_{66}^{(1)}$进行L_6变换，得到$R^{(2)}$.

对应的自由度为$N-1$-引入因子数，因为引入了2个因子x_1和x_6，所以自由度为N-3，

4 逐步回归在地理系统分析中的应用实例

剔除检验 x_6刚引进，无需检验，只需对x_1的偏回归平方和做检验：

$$V_1^{(2)}=\frac{(r_{1n}^{(2)})^2}{r_{11}^{(2)}}=\frac{0.3955^2}{1.0061}=0.1564/1.0061=0.1554$$

$$F_2=\frac{V_1^{(2)}}{r_{nn}^{(2)}}(N-3)=\frac{0.1554}{0.719}\times 26=5.622>3.5$$

故x_1不需剔除。实际上，在引入第二个因子后，无须做剔除的检验

选第三个因子　对$j\neq 1,6$，计算偏回归平方和$V_j^{(3)}=\frac{(r_{jn}^{(2)})^2}{r_{jj}^{(2)}}$得到$x_5$的方差

贡献最大，即$\max\limits_{j\neq 1,6} V_j^{(3)}=V_5^{(3)}=\frac{0.2965^2}{0.8938}=0.09836$

4 逐步回归在地理系统分析中的应用实例

作F检验$F_1=\dfrac{V_5^{(3)}}{r_{nn}^{(2)}-V_5^{(3)}}\times(N-4)=\dfrac{0.09836}{0.7195-0.09836}\times 25=3.9588>3.5$

所以可引进x_5,此时回归方程的复相关关系系数为$R=0.616$。然后对$R^{(2)}$以$r_{55}^{(2)}$作变换L_5得到$R^{(3)}$。

剔除检验　由于因子x_5的引入，又需对原有因子x_1, x_6重做检验

计算$V_1^{(3)}$与$V_6^{(3)}$，得$\min\limits_{j=1,6} V_j^{(3)}=V_6^{(3)}=0.16224$

$$F_2=\frac{V_6^{(3)}(N-4)}{r_{nn}(3)}=6.5303>3.5$$

故x_1与x_6都无须剔除

4 逐步回归在地理系统分析中的应用实例

选第四个变量对$j \neq 1,5,6$，计算$V_j^{(4)}=(r_{jn}^{(3)})^2/r_{jj}^{(3)}$，由$R^{(3)}$可得

$$\max_{j\neq 1,5,6} V_j^{(4)} = V_3^{(4)} = \frac{0.2032^2}{0.6877} = 0.0599$$

$$F_1 = \frac{V_3^{(4)}(N-5)}{r_{nn}^{(3)} - V_3^{(4)}} = \frac{0.0599}{0.621-0.0599} \times 24 = 2.563 < 3.5$$

设x_3不能引入。

至此为止，共选进3个因子：x_1, x_5, x_6，挑选因子工作结束。

逐步回归在地理系统分析中的应用实例

（4）计算回归系数：从$R^{(3)}$可得模型6-6的参数估计

d_1=0. 4227

d_5=0. 3317

d_6=-0. 4249

从而模型6-1的参数估计为

$$b_1=\frac{\sigma_n}{\sigma_1}d_1=\frac{1039.3}{15.589}\times 0.4227=28.1809$$

$$b_5=\frac{\sigma_n}{\sigma_5}d_5=\frac{1039.3}{6.768}\times 0.3317=50.9361$$

$$b_6=\frac{\sigma_n}{\sigma_6}d_6=\frac{1039.6}{27.870}\times(-0.4247)=-15.8449$$

$$b_0=x_n-b_1x_1-b_5x_5-b_6x_6$$

$$=374.9-28.1809\times 4.092-50.9361\times 0.083-(-15.8449\times 7.7)=377.3618$$

所以预报方程是$y=377.4+28.19x_1+50.93x_5-15.85x_6$

- **计算回归系数参见P126最后一次变换最后一列数据**
- **参数估计公式见P120**

4 逐步回归在地理系统分析中的应用实例

（5）计算$S_{总}$，$S_{回}$，$S_{剩}$对方程做显著性检验，及求$\hat{\sigma}$:

$S_{总}=\sigma_n^2=1080144.49$

$S_{剩}=\sigma_n^2 r_{nn}^{(3)}=670877.69$

$S_{回}=\sigma_n^2(1-r_{nn}^{(3)})=1039.32\times(1-0.621)=409266.7473$

$$F=\frac{S_{回}/3}{S_{剩}/25}=13.66>F_{0.05}(3.25)=2.99$$

此方程是显著的。

逐步回归的SPSS实现过程

逐步回归在地理系统分析中的应用实例

台风暴雨资料.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

12 :

	x1	x2	x3	x4	x5	x6	x7	y	var
1	14.50	27.0	8.8	2.0	-.5	8.0	248.0	900	
2	7.57	27.7	10.8	7.0	.8	5.0	81.0	354	
3	1.94	28.3	13.6	13.0	-.2	1.7	124.8	566	
4	3.04	27.3	12.1	13.0	.2	1.5	314.6	521	
5	8.07	28.5	5.7	-2.0	-.6	2.7	110.4	333	
6	4.64	28.5	15.8	14.0	1.4	2.0	109.6	259	
7	3.02	27.4	5.4	.0	.6	4.6	110.0	589	
8	6.20	28.2	12.0	12.0	.0	2.5	373.0	416	
9	2.69	29.0	12.7	6.0	1.3	15.7	87.8	289	
10	2.85	27.5	5.0	12.0	.0	6.8	152.2	254	
11	1.02	27.0	20.7	1.0	1.0	10.0	148.5	209	
12	1.62	27.5	7.0	4.0	1.5	6.0	48.0	428	
13	7.02	27.3	5.8	-17.0	1.8	10.0	230.0	673	
14	2.09	27.3	14.5	-11.0	.0	8.5	110.5	395	
15	.83	28.7	11.8	-13.0	2.3	4.0	125.0	327	
16	4.56	27.0	7.0	-4.0	-.3	4.0	240.0	829	
17	5.43	29.0	7.2	-4.0	-1.5	4.0	157.2	266	
18	4.05	26.9	4.2	-1.0	-.3	2.8	80.0	653	

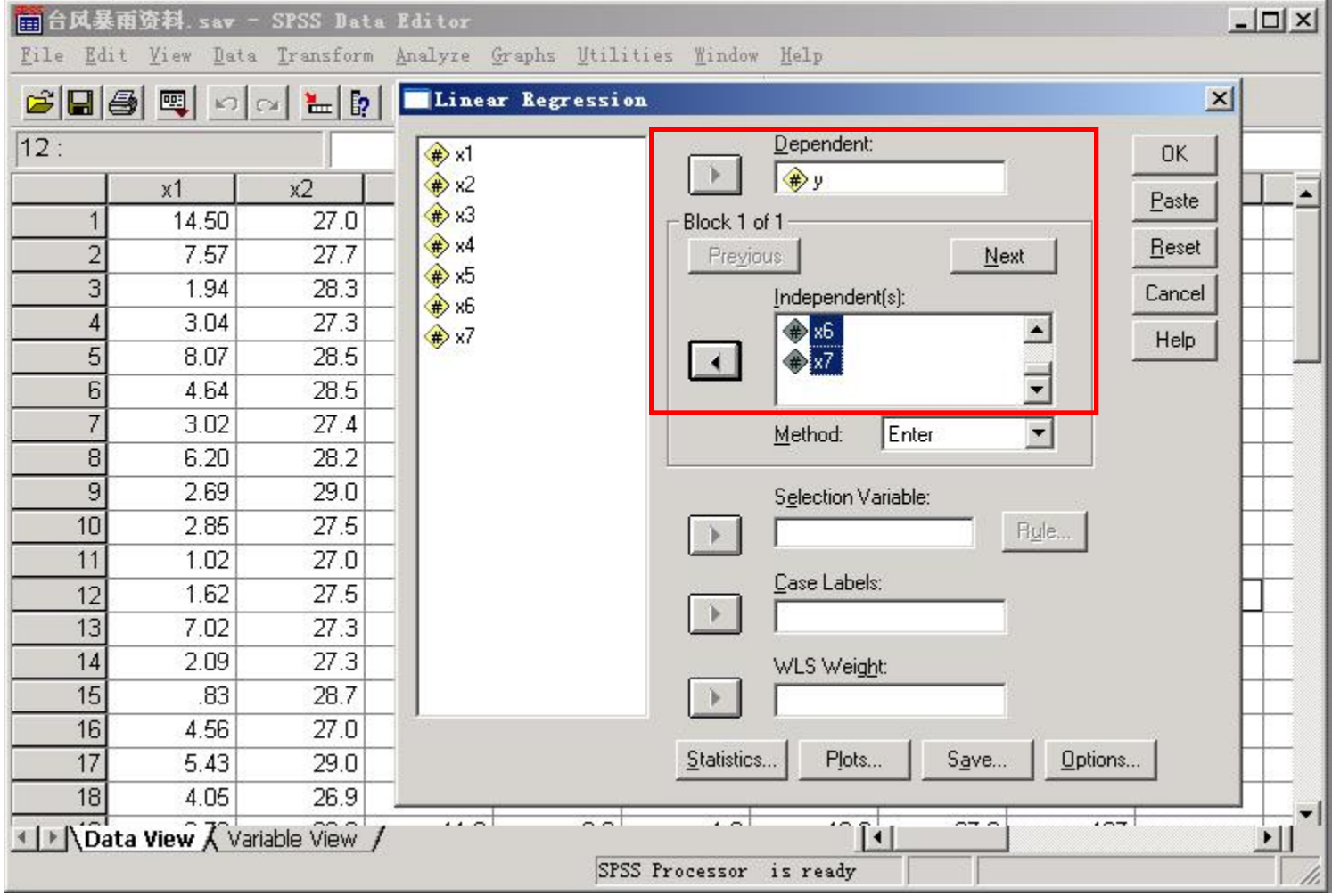
Data View | Variable View

SPSS Processor is ready

（1）将所有的变量依次输入到数据栏中

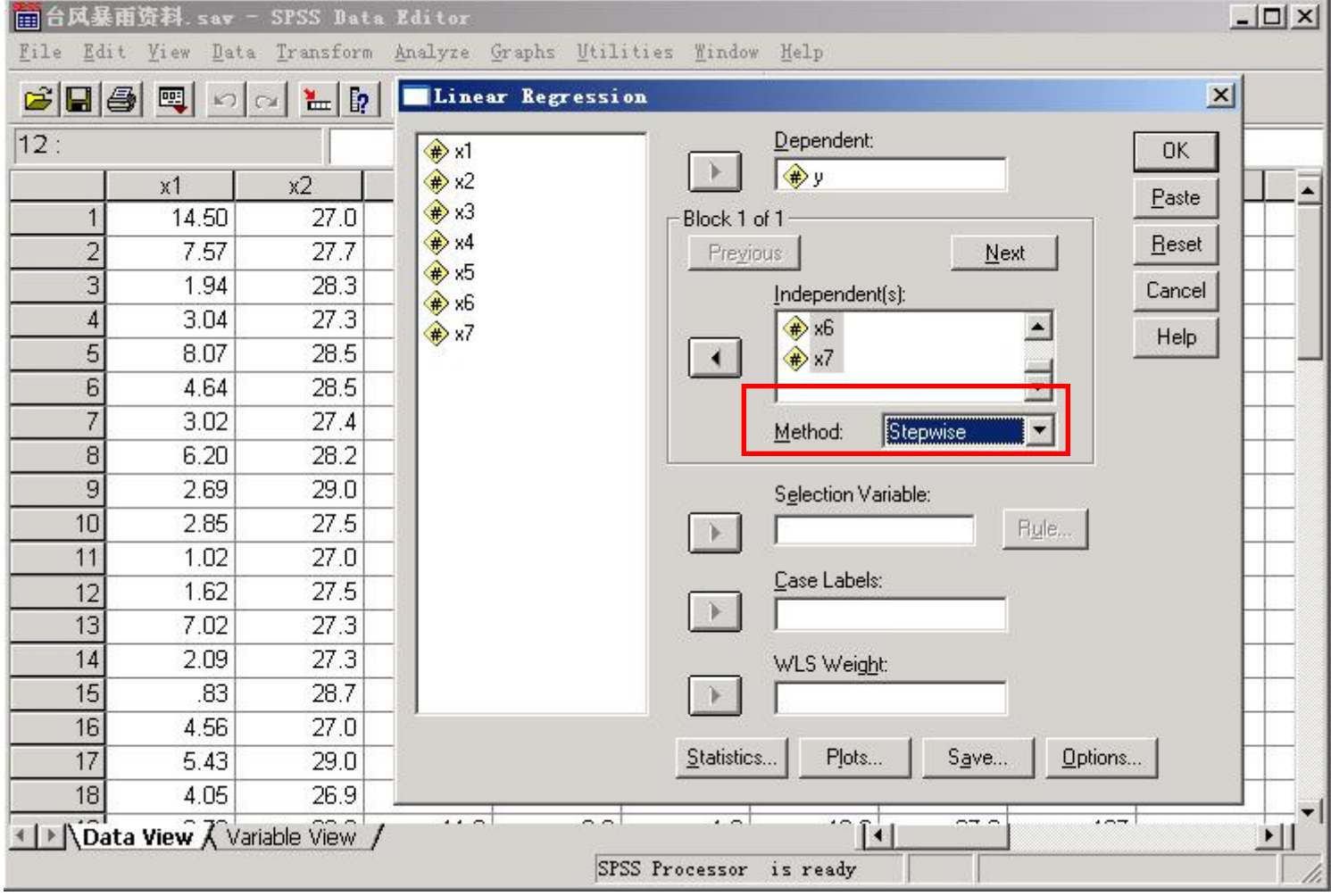
（2）注意设置变量的类型及小数点位数

4 逐步回归在地理系统分析中的应用实例



（1）将y选入因变量对话框，将$x_1,x_2,x_3,x_4,x_5,x_6,x_7$移入自变量对话框

4 逐步回归在地理系统分析中的应用实例



（1）方法里面选择stepwise

（2）其它采用默认选项，点击ok

4 逐步回归在地理系统分析中的应用实例

Descriptive Statistics

	Mean	Std. Deviation	N
y	371.48	197.577	29
x1	4.0921	2.94586	29
x2	27.572	.9137	29
x3	10.903	6.1897	29
x4	.000	9.6103	29
x5	.083	1.2762	29
x6	7.700	5.2648	29
x7	168.872	82.5405	29

（1）所有因变量和自变量的统计分析表，包括平均值，标准偏差和样本数

逐步回归在地理系统分析中的应用实例

Correlations

		y	x1	x2	x3	x4	x5	x6	x7
Pearson Correlation	y	1.000	.415	-.118	-.286	-.025	.133	-.332	.180
	x1	.415	1.000	-.182	-.069	.002	-.106	-.078	.173
	x2	-.118	-.182	1.000	-.091	.196	.150	-.182	-.290
	x3	-.286	-.069	-.091	1.000	-.358	.455	.451	-.117
	x4	-.025	.002	.196	-.358	1.000	-.269	-.404	.056
	x5	.133	-.106	.150	.455	-.269	1.000	.315	-.252
	x6	-.332	-.078	-.182	.451	-.404	.315	1.000	-.004
	x7	.180	.173	-.290	-.117	.056	-.252	-.004	1.000
Sig. (1-tailed)	y	.	.013	.271	.066	.449	.245	.039	.176
	x1	.013	.	.172	.361	.496	.292	.343	.185
	x2	.271	.172	.	.319	.154	.219	.172	.063
	x3	.066	.361	.319	.	.028	.007	.007	.273
	x4	.449	.496	.154	.028	.	.079	.015	.386
	x5	.245	.292	.219	.007	.079	.	.048	.093
	x6	.039	.343	.172	.007	.015	.048	.	.492
	x7	.176	.185	.063	.273	.386	.093	.492	.
N	y	29	29	29	29	29	29	29	29
	x1	29	29	29	29	29	29	29	29
	x2	29	29	29	29	29	29	29	29
	x3	29	29	29	29	29	29	29	29
	x4	29	29	29	29	29	29	29	29
	x5	29	29	29	29	29	29	29	29
	x6	29	29	29	29	29	29	29	29
	x7	29	29	29	29	29	29	29	29

（1）不同参数之间的相关系数；单尾检验及其样本数

4 逐步回归在地理系统分析中的应用实例

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	257.578	58.862		4.376	.000
	x1	27.836	11.743	.415	2.370	.025

a. Dependent Variable: y

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	x2	-.044[a]	-.243	.810	-.048	.967
	x3	-.259[a]	-1.508	.143	-.284	.995
	x4	-.026[a]	-.144	.887	-.028	1.000
	x5	.179[a]	1.020	.317	.196	.989
	x6	-.301[a]	-1.782	.086	-.330	.994
	x7	.111[a]	.618	.542	.120	.970

a. Predictors in the Model: (Constant), x1

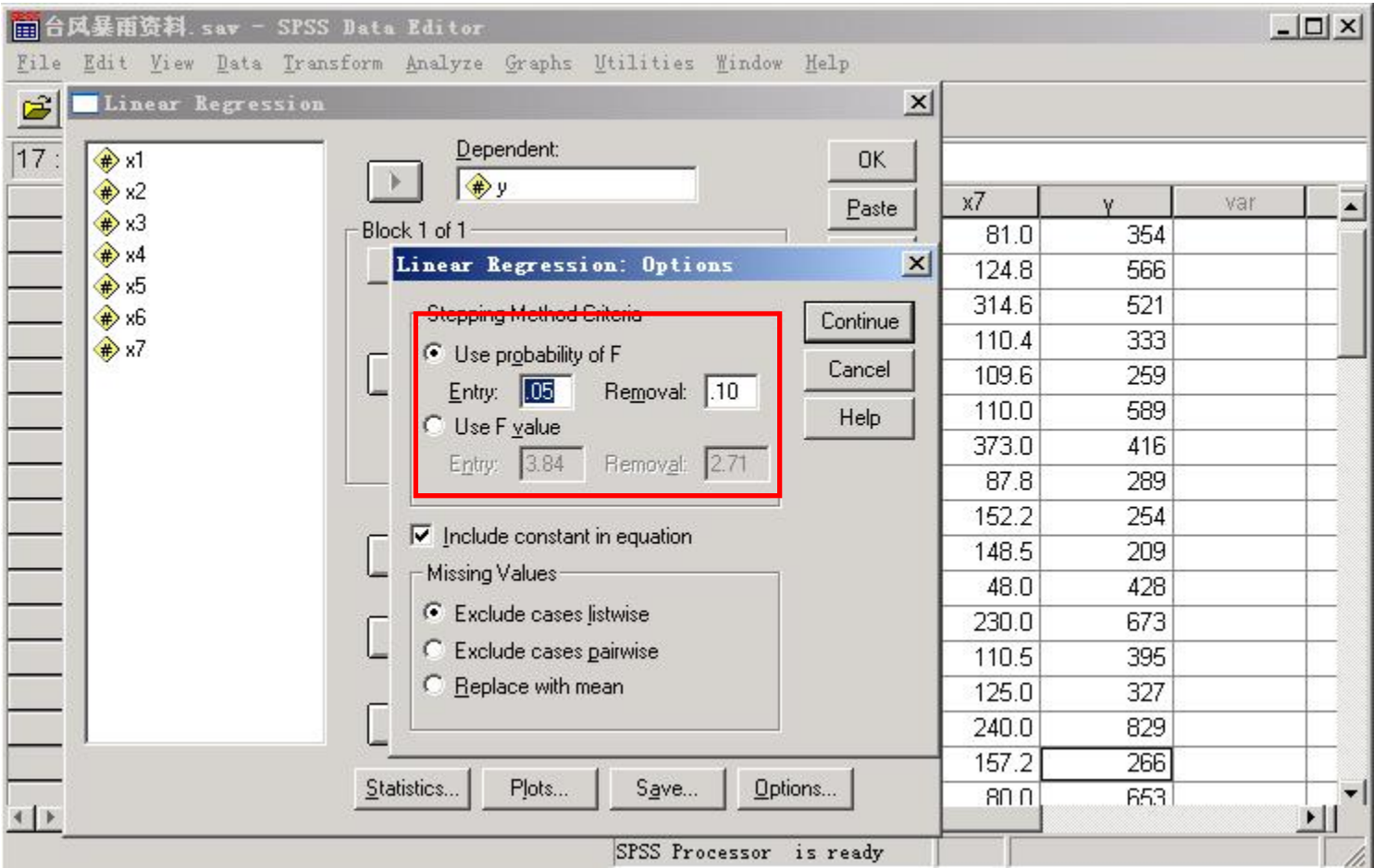
b. Dependent Variable: y

（1）结果发现：模型只引入了一个参数x_1，而其他参数均被排除

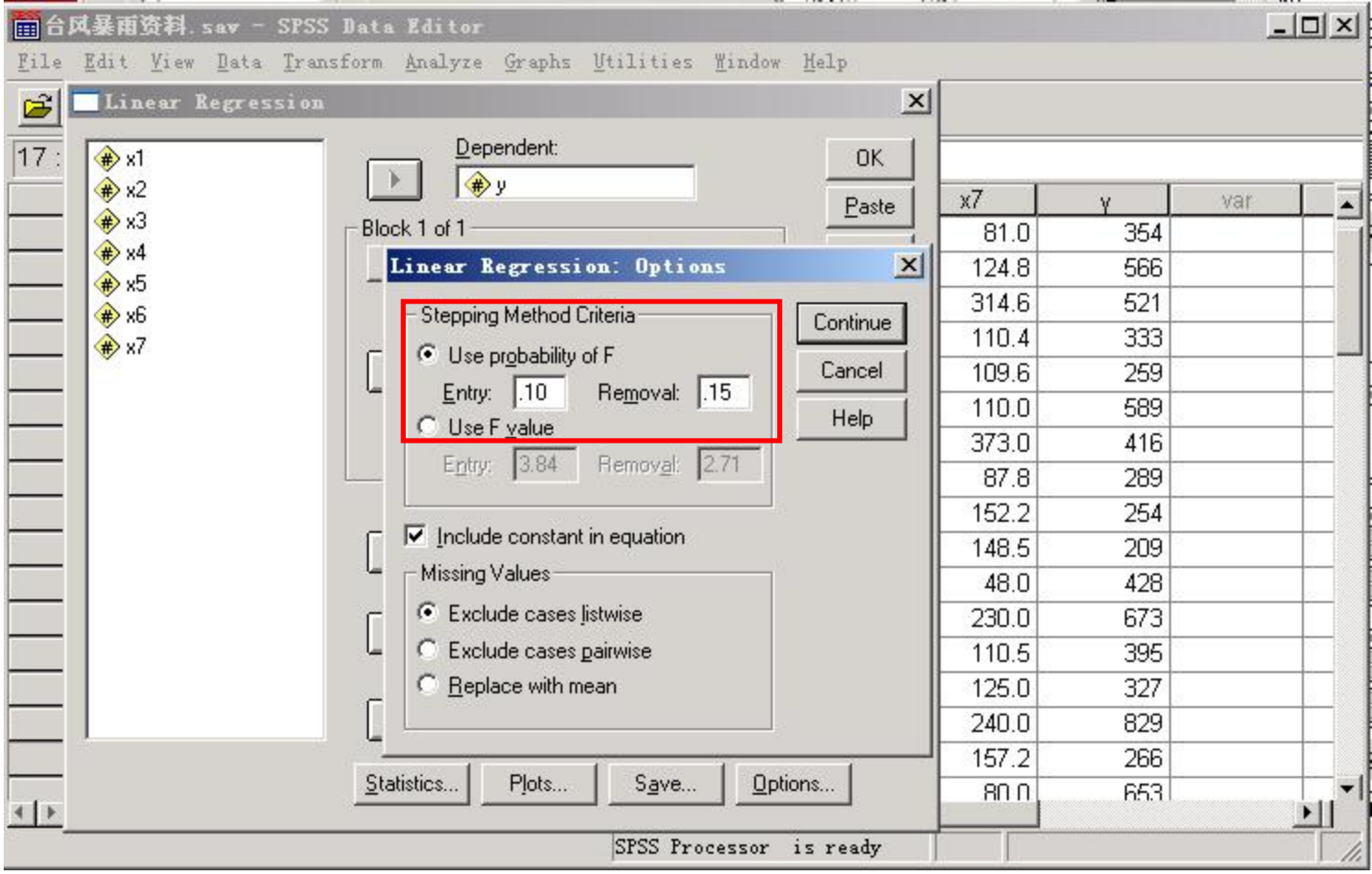
（2）为什么和书上所做的回归分析有所不同，原因何在？？

逐步回归在地理系统分析中的应用实例

- 不同置信度下引入的函数和排除的函数存在差异，模型设置的置信度越高，引入的越少，排除的越多，反之，引入的增加，排除的减少
- 逐步回归原则：在置信度能够满足基本要求的情况下，尽可能的增加自变量的数目
- 默认置信度为进入0.05，移除0.10



逐步回归在地理系统分析中的应用实例



- 将逐步回归的置信度改为进入0.10，移除0.15
- 点击continue，其他同上

4 逐步回归在地理系统分析中的应用实例

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	x1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .100, Probability-of-F-to-remove >= .150).
2	x6	.	Stepwise (Criteria: Probability-of-F-to-enter <= .100, Probability-of-F-to-remove >= .150).
3	x5	.	Stepwise (Criteria: Probability-of-F-to-enter <= .100, Probability-of-F-to-remove >= .150).

a. Dependent Variable: y

- 模型引入的变量包括x1，x6和x5
- 采用的方法为stepwise，引入的准则为置信度≤0.1，移出的准则为置信度≥0.15
- 引入模型的先后顺序为x1，x6和x5

4 逐步回归在地理系统分析中的应用实例

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.415[a]	.172	.142	183.056	
2	.512[b]	.262	.206	176.096	
3	.586[c]	.344	.265	169.377	2.610

a. Predictors: (Constant), x1

b. Predictors: (Constant), x1, x6

c. Predictors: (Constant), x1, x6, x5

d. Dependent Variable: y

- 依照新的引入和移出原则，一共产生了3个不同的模型，其中第一个模型只引入了x_1，第二个模型引入了x_1和x_6，第三个模型引入了x_1，x_6和x_5
- 从逐步回归的角度来说，在给定的置信度下，引入的变量尽可能多
- 模型中引入的变量越多，其决定系数R^2越大，即自变量能够解释的因变量变化的百分比越多，残差平方和越小

4 逐步回归在地理系统分析中的应用实例

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	188269.6	1	188269.636	5.618	.025[a]
	Residual	904759.6	27	33509.615		
	Total	1093029	28			
2	Regression	286777.9	2	143388.960	4.624	.019[b]
	Residual	806251.3	26	31009.666		
	Total	1093029	28			
3	Regression	375816.3	3	125272.098	4.367	.013[c]
	Residual	717212.9	25	28688.518		
	Total	1093029	28			

a. Predictors: (Constant), x1

b. Predictors: (Constant), x1, x6

c. Predictors: (Constant), x1, x6, x5

d. Dependent Variable: y

- 模型参数方差分析表
- 三个模型在给定的置信度下均可以通过检验，即p≤0.1，因此所有模型均适用
- 但从逐步回归的角度，认为模型3能够更好的解释地理现象
- 随着模型参数引入的增多，在总平方和不变的情况下，回归平方和增加，而剩余平方和减小（第二列）

4 逐步回归在地理系统分析中的应用实例

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations		
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	257.578	58.862		4.376	.000			
	x1	27.836	11.743	.415	2.370	.025	.415	.415	.415
2	(Constant)	351.075	77.189		4.548	.000			
	x1	26.252	11.332	.391	2.317	.029	.415	.414	.390
	x6	-11.301	6.341	-.301	-1.782	.086	-.332	-.330	-.300
3	(Constant)	367.373	74.818		4.910	.000			
	x1	27.910	10.940	.416	2.551	.017	.415	.454	.413
	x6	-14.801	6.414	-.394	-2.308	.030	-.332	-.419	-.374
	x5	46.737	26.529	.302	1.762	.090	.133	.332	.285

a. Dependent Variable: y

- 模型系数表
- 三个模型的系数（见B列），其中constant项对应的为常数项，其它依次为各个自变量前面的系数
- 所有模型的系数均能够通过t检验，在p=0.1置信度下均是可靠的
- 依据逐步回归原理，选择模型三前面的系数作为最优模型
- 模型引入的先后顺序也可以从偏回归系数进行判断

4 逐步回归在地理系统分析中的应用实例

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	x2	-.044[a]	-.243	.810	-.048	.967
	x3	-.259[a]	-1.508	.143	-.284	.995
	x4	-.026[a]	-.144	.887	-.028	1.000
	x5	.179[a]	1.020	.317	.196	.989
	x6	-.301[a]	-1.782	.086	-.330	.994
	x7	.111[a]	.618	.542	.120	.970
2	x2	-.110[b]	-.620	.541	-.123	.928
	x3	-.155[b]	-.815	.423	-.161	.795
	x4	-.176[b]	-.956	.348	-.188	.836
	x5	.302[b]	1.762	.090	.332	.894
	x7	.114[b]	.661	.515	.131	.970
3	x2	-.180[c]	-1.048	.305	-.209	.887
	x3	-.315[c]	-1.669	.108	-.322	.688
	x4	-.128[c]	-.705	.488	-.142	.813
	x7	.201[c]	1.192	.245	.236	.907

a. Predictors in the Model: (Constant), x1

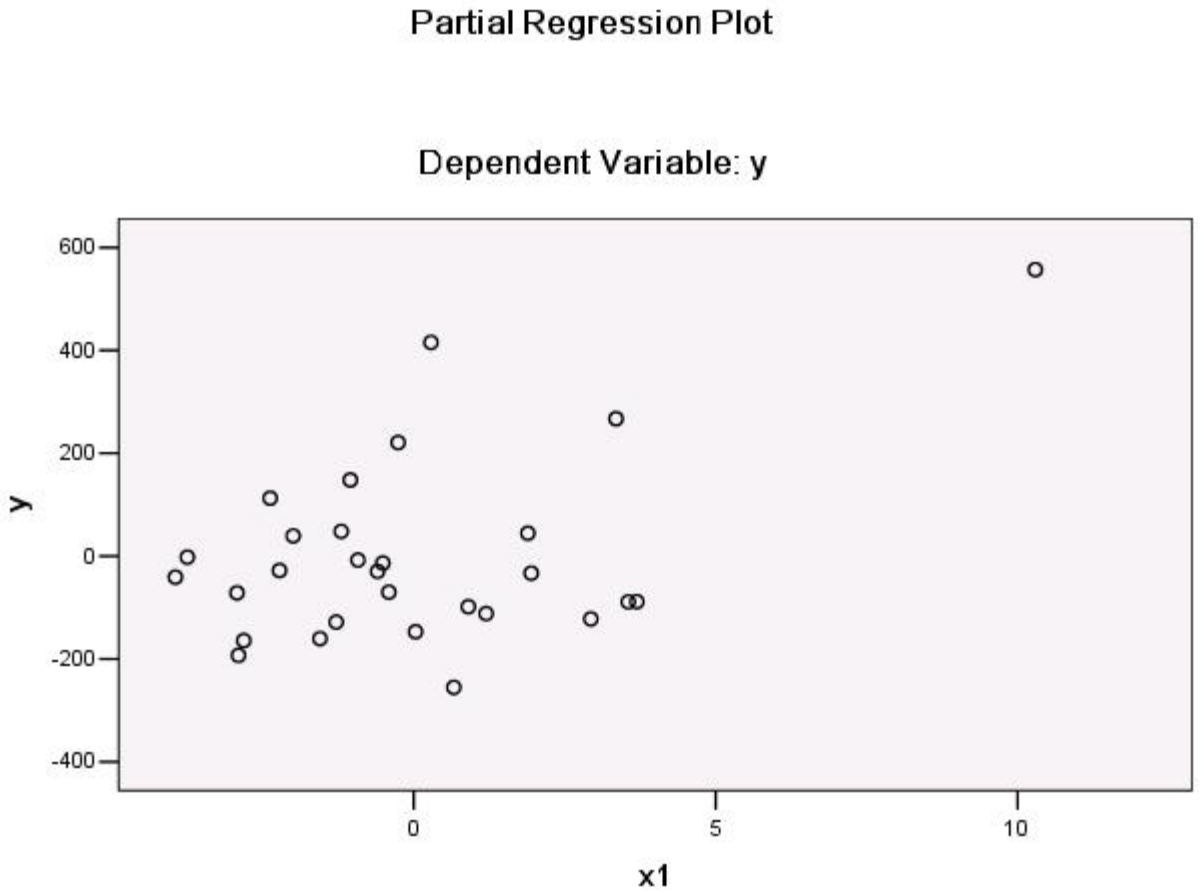
b. Predictors in the Model: (Constant), x1, x6

c. Predictors in the Model: (Constant), x1, x6, x5

d. Dependent Variable: y

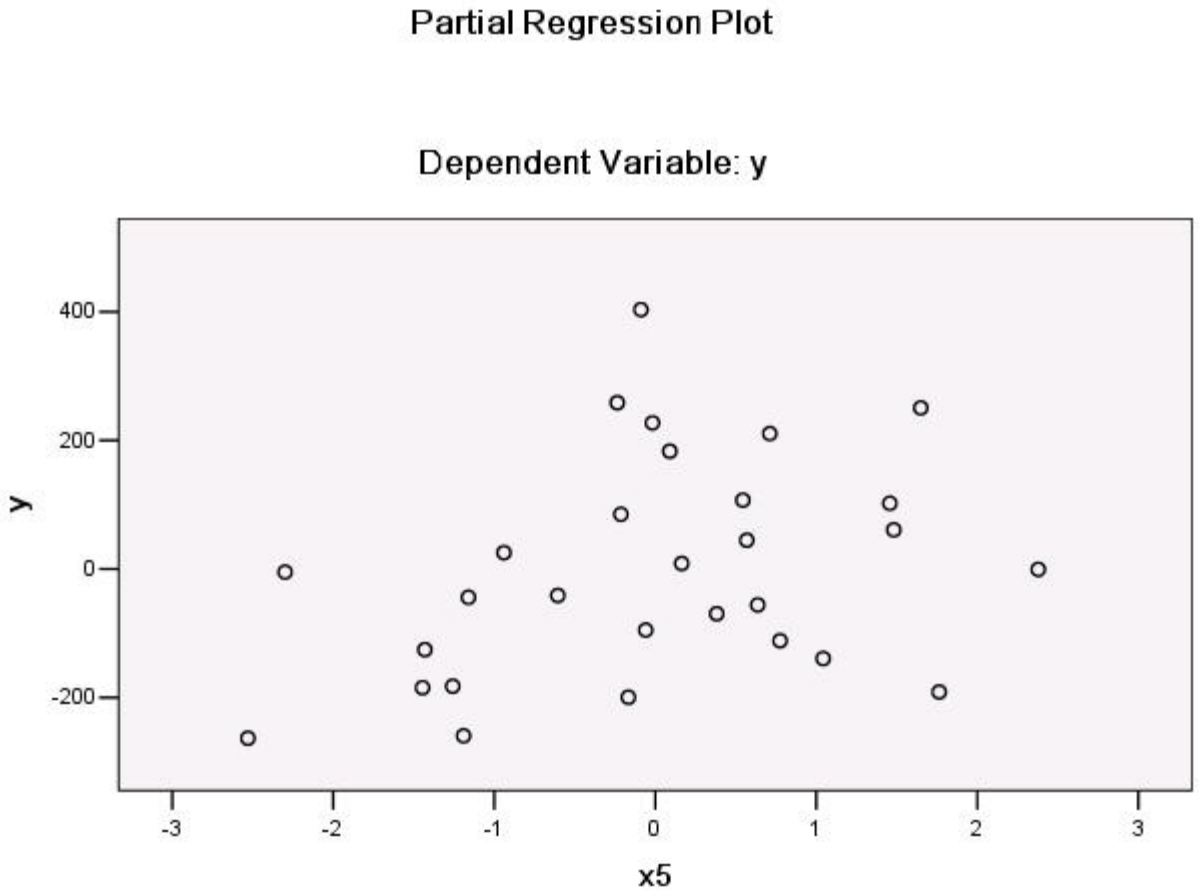
- 从模型3可以看出，x_2,x_3,x_4,x_7均从模型中被排除，其系数均没有通过t检验，即$P \geq 0.1$

4 逐步回归在地理系统分析中的应用实例



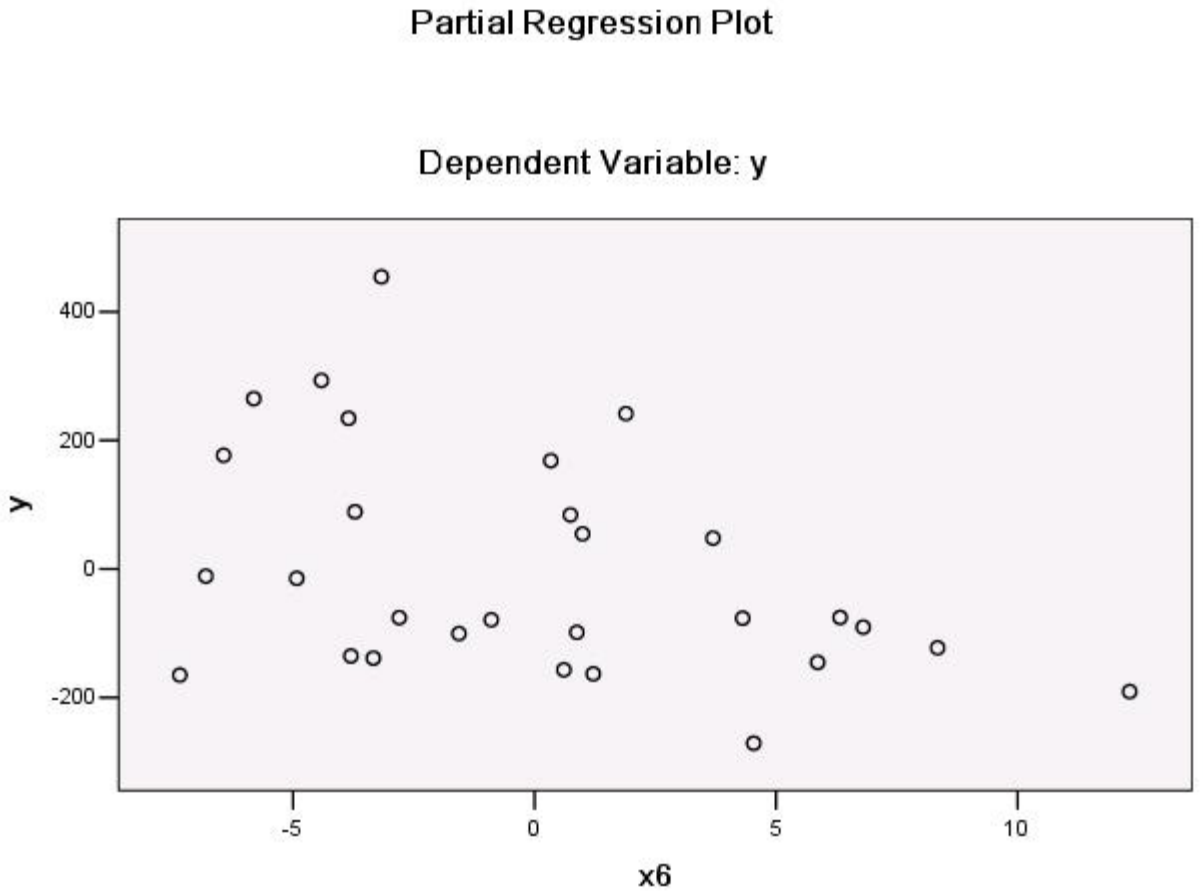
- 从模型中可以看出，在x_5和x_6不变的情况下，散点图表现出x_1和y呈现出显著的正偏相关

4 逐步回归在地理系统分析中的应用实例



- 从模型中可以看出，在x_1 和x_6不变的情况下，散点图表现出x_5和y呈现出显著的正偏相关

4 逐步回归在地理系统分析中的应用实例



- 从模型中可以看出，在x_1 和x_5 不变的情况下，散点图表现出x_6和y呈现出显著的负偏相关

4 逐步回归在地理系统分析中的应用实例

- 逐步回归的MATLAB实现
- %输出变量矩阵
- x1=[14.5 7.57 1.94 3.04 8.07 4.64 3.02 6.20 2.69 2.85 1.02 1.62 7.02 2.09 0.83 4.56 5.43 4.05 3.78 1.11 7.17 5.00 3.88 0.74 3.05 0.30 3.44 5.94 3.12];
- x2=[27.0 27.7 28.3 27.3 28.5 28.5 27.4 28.2 29.0 27.5 27.0 27.5 27.3 27.3 28.7 27.0 29.0 26.9 28.0 29.0 27.0 26.0 27.0 26.5 27.8 28.0 28.0 25.0 27.2];
- x3=[8.8 10.8 13.6 12.1 5.7 15.8 5.4 12.0 12.7 5.0 20.7 7.0 5.8 14.5 11.8 7.0 7.2 4.2 11.6 13.6 11.0 33.6 16.0 -1.2 13.4 11.0 8.0 10.0 9.1];

逐步回归在地理系统分析中的应用实例

- x4=[2.0 7.0 13.0 13.0 -2.0 14.0 0.0 12.0 6.0 12.0 1.0 4.0 -17.0 -11.0 -13.0 -4.0 -4.0 -1.0 8.0 -3.0 2.0 -27.0 -7.0 6.0 -7.0 -7.0 -4.0 1.0 6.0];
- x5=[-0.5 0.8 -0.2 0.2 -0.6 1.4 0.6 0.0 1.3 0.0 1.0 1.5 1.8 0.0 2.3 -0.3 -1.5 -0.3 -1.0 -0.5 -1.0 2.7 1.0 -2.0 -1.7 -0.7 -0.2 -2.7 1.0];
- x6=[8.0 5.0 1.7 1.5 2.7 2.0 4.6 2.5 15.7 6.8 10.0 6.0 10.0 8.5 4.0 4.0 4.0 2.8 12.2 14.0 10.6 23.3 9.5 9.0 2.7 8.0 11.7 5.2 11.3];
- x7=[248.0 81.0 124.8 314.6 110.4 109.6 110.0 373.0 87.8 152.2 148.5 48.0 230.0 110.5 125.0 240.0 157.2 80.0 97.0 144.0 157.3 206.4 134.0 368.0 165.2 144.2 256.0 201.6 173.0];

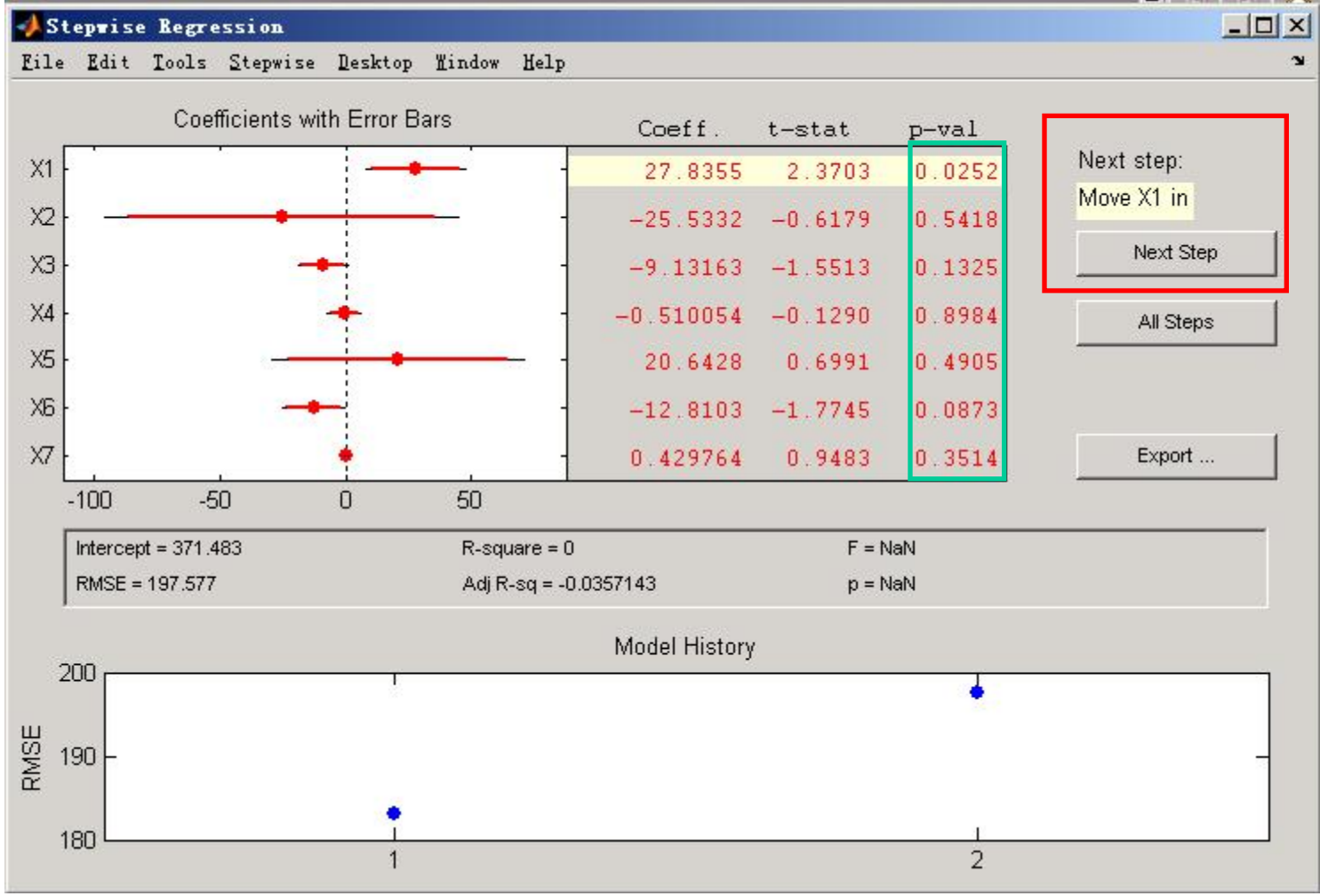
4 逐步回归在地理系统分析中的应用实例

- 【函数名称】：stepwise
- 【函数功能】：创建多元线性回归法建模的交互式图形环境
- 【调用格式】：
- stepwise（x，y）
- stepwise（x，y，inmodel，penter，premove）
- 【参数说明】：
- X为p元线性模型解释变量的n个观测值的n*p矩阵
- Y为p元线性模型因变量的n个观测值n*1向量
- Inmodel为各个解释变量在最终回归方程中地位的说明（1表示在方程中，0表示不再方程中）
- Penter为模型检验的显著性水平上限值（缺省设置为0.05）
- Premove为模型检验的显著性水平下限值（缺省设置为0.10）

逐步回归在地理系统分析中的应用实例

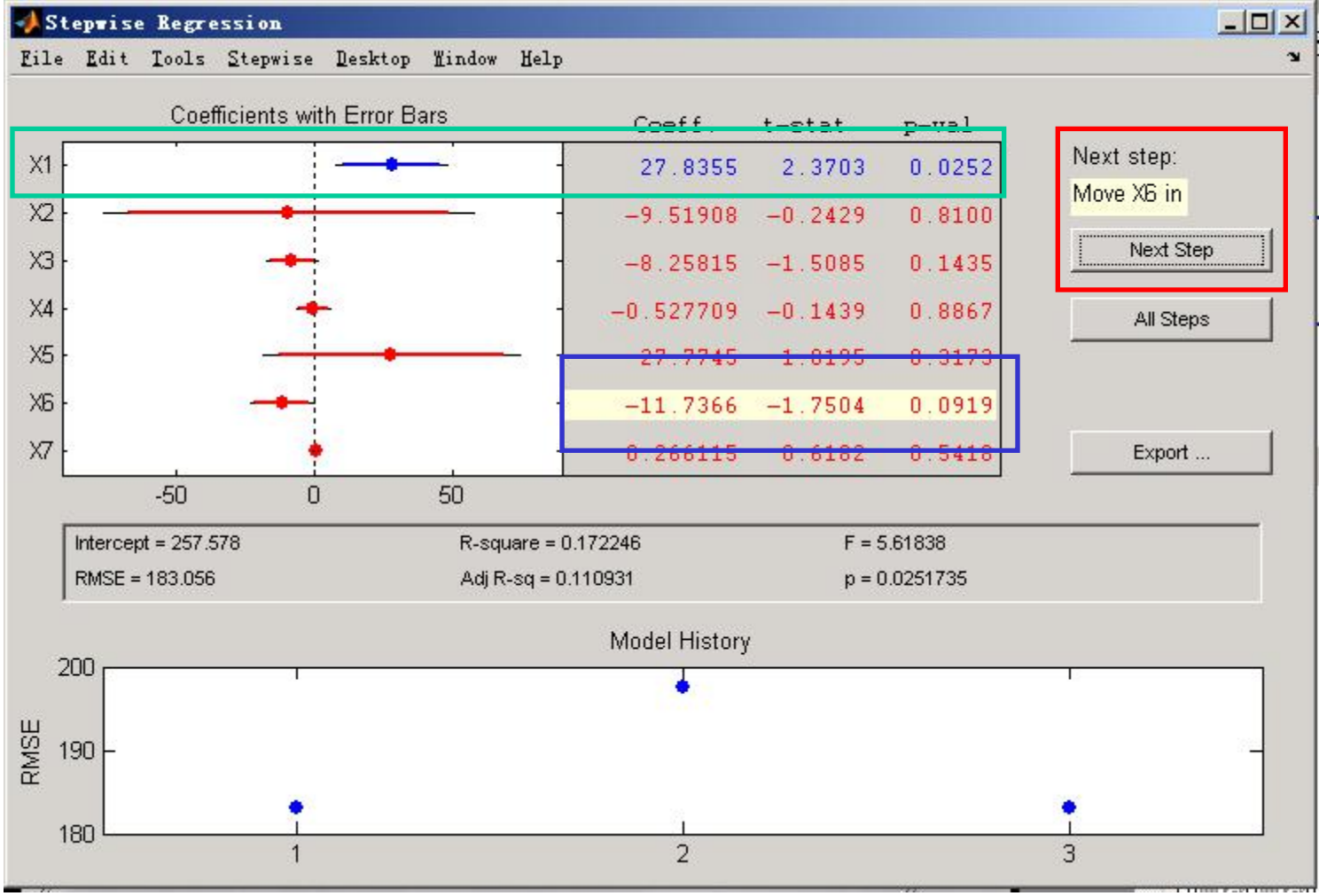
- %将所有矩阵进行转置，构造cat函数所需的变量
- x11=x1';x22=x2';x33=x3';x44=x4; x55=x5'; x66=x6';x77=x7';
- y=y';
- %构造大矩阵，函数采用cat函数，2表示沿列进行拼接
- x=cat(2,x11,x22,x33,x44,x55,x66,x77);
- %采用stepwise函数进行拟合分析
- stepwise(x,y,1,0.10,0.15)

4 逐步回归在地理系统分析中的应用实例



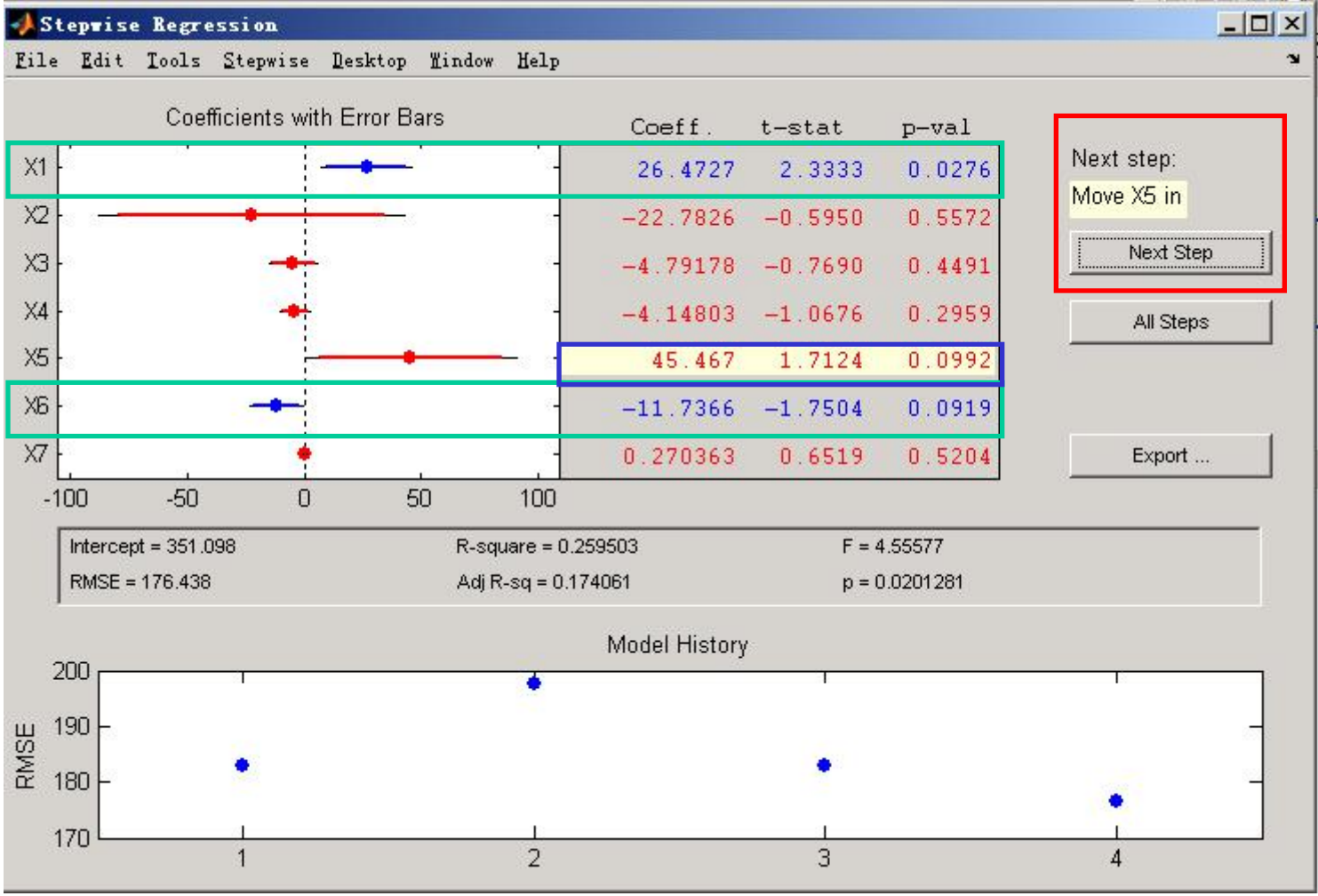
- 第一步，通过对所有自变量系数进行t检验，发现x_1对应的p最小，因此首先将x_1移入
- 点击next step

4 逐步回归在地理系统分析中的应用实例



- 选择next step后出现move x_6 in，即输入自变量x_6，图中蓝色的部分表示已经引入的变量，红色部分表示正在引入或者还没有引入的变量，黄色条状重点标注的为正在引入的变量即x_6
- Intercept为截距，即constant值，R-square为决定系数，F为F检验值，RMSE为残差平方和，AdjR-sq为调整后的决定系数，P为显著性检验值，即引入的变量x_6能够通过置信度为0.1的检验

逐步回归在地理系统分析中的应用实例



- 点击next step显示下一步move x_5 in，同时蓝色部分显示的x_1和x_6已经成功引入，黄色条状显示正在引入的变量，红色部分显示尚未引入的变量
- 其余参数intercept等同上，结果显示随着自变量引入的增加，残差平方和减小，显著性检验表明，其能够通过p=0.1检验

4 逐步回归在地理系统分析中的应用实例

- 点击next step显示move no terms，说明在当前的置信度下已经没有变量可以引入，整个模型的引入过程结束，next step按钮变暗
- 左上框蓝色部分显示的为最终引入的变量，x_1，x_5和x_6，显示方式为系数加标准误差，右边coeff.为不同自变量对应的系数，t-stat为不同系数的t检验值，p-val为显著性检验的结果，显示均能通过引入的显著性检验，即$p \leq 0.1$
- 红色部分为没有引入的变量，原因在于$p\text{-value} \geq 0.1$，即不满足引入变量的条件，方程的系数检验在特定的置信度下可以通过，说明拟合的方程比较可信

4 逐步回归在地理系统分析中的应用实例

- 得到的回归模型为
- Y=366.268+28.1624x1+45.467x5-15.1861x6
- 随着引入变量的增多，不同软件由于对小数位数处理不同，得到的结果会有所出入，引入的变量越少，不同软件计算的系数估计值差异越小

Coefficients[a]

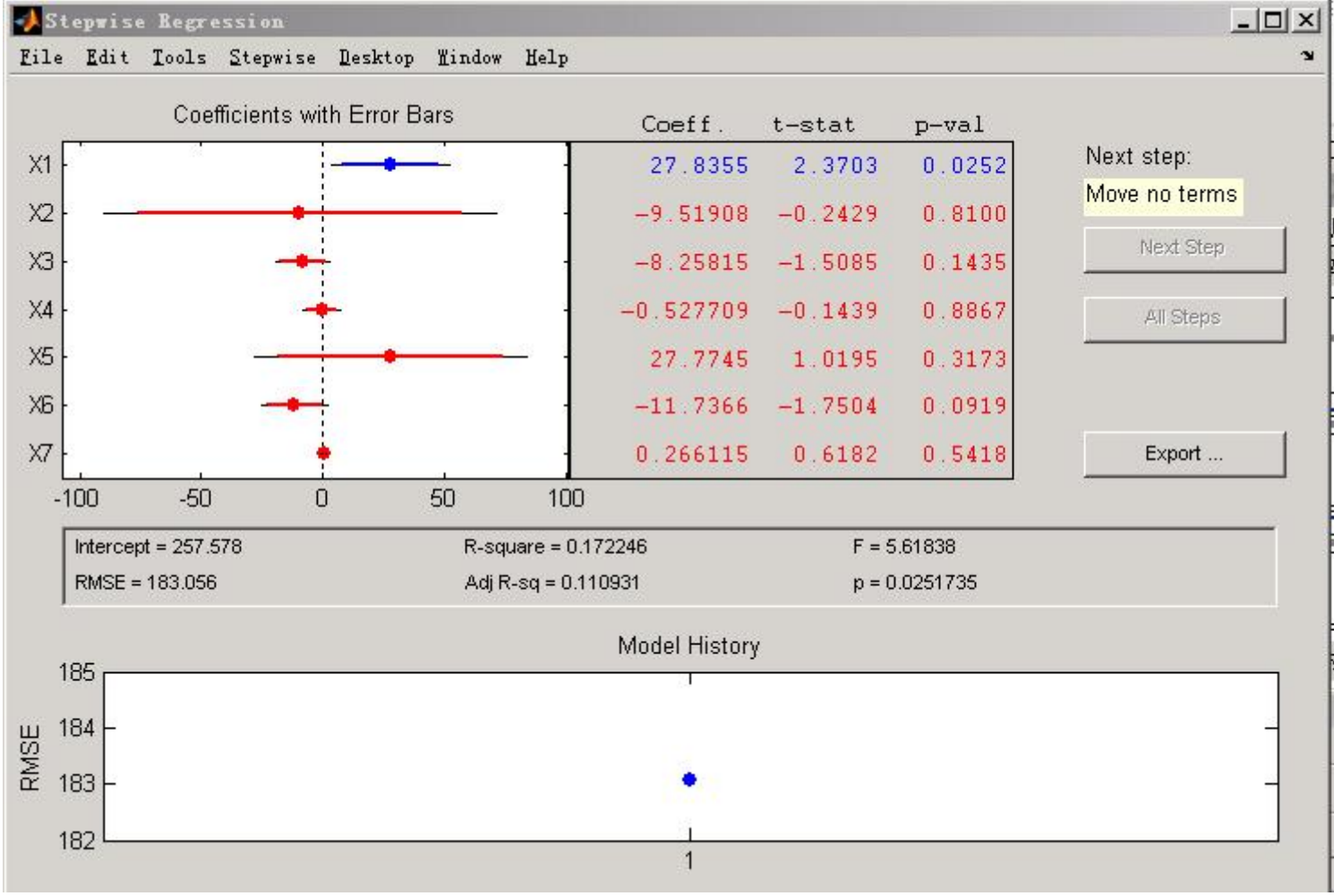
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	257.578	58.862		4.376	.000			
	x1	27.836	11.743	.415	2.370	.025	.415	.415	.415
2	(Constant)	351.075	77.189		4.548	.000			
	x1	26.252	11.332	.391	2.317	.029	.415	.414	.390
	x6	-11.301	6.341	-.301	-1.782	.086	-.332	-.330	-.300
3	(Constant)	367.373	74.818		4.910	.000			
	x1	27.910	10.940	.416	2.551	.017	.415	.454	.413
	x6	-14.801	6.414	-.394	-2.308	.030	-.332	-.419	-.374
	x5	46.737	26.529	.302	1.762	.090	.133	.332	.285

a. Dependent Variable: y

逐步回归在地理系统分析中的应用实例

- %将所有矩阵进行转置，构造cat函数所需的变量
- x11=x1';x22=x2';x33=x3';x44=x4; x55=x5'; x66=x6';x77=x7';
- y=y';
- %构造大矩阵，函数采用cat函数，2表示沿列进行拼接
- x=cat(2,x11,x22,x33,x44,x55,x66,x77);
- %采用stepwise函数进行拟合分析
- stepwise(x,y,1,0.05,0.10)

4 逐步回归在地理系统分析中的应用实例



- 运行stepwise（x，y，1，0.05，0.10），便会得到如上所示表格。置信度降低后，只有左上框所示的蓝色部分即变量x1被引入，而其它变量均被移出
- Next step显示 move no terms，且next step按钮变成灰色，说明已经没有变量可以引入
- 从模型的参数来看，其能够通过置信度为引入时$p \leq 0.05$的要求，即所拟合的模型能够通过显著性检验，其结果是可信的

逐步回归在地理系统分析中的应用实例

- 最终得到的模型表达方程为
- Y=257.578+27.8355x1
- 不同软件计算出的结果可能会有小的出入，属于正常结果
- 模型的拟合属于一个不断逼近的过程，不存在完全精确的参数，不同软件的处理方式不同，导致结果有所差异

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	257.578	58.862		4.376	.000
	x1	27.836	11.743	.415	2.370	.025

a. Dependent Variable: y