第14卷 第4期 华侨大学学报自然科学版 Vol. 14 №. 4
JOURNAL OF HUAQIAO UNIVERSITY
1993年10月 (NATURAL SCIENCE) Oct. 1993

秩回归模型用 *BIC* 建模的相容性及应用*

陈 建 伟

(管理信息科学系)

摘要 文〔1〕提出秩和回归模型预测方法,本文研究用 *BIC* 准则选择秩回归变量的相容性问题和建模原则,给出具体的模型设计方法,并在气象预测应用中效果良好.

关键词 秩回归模型,*BIC* 准则,相容性

0 前言

逐步回归法是目前线性回归模型中最常用的建模方法,在国民经济,水文预测等领域中得到广泛的应用. *Hocking*, *Beals* 等学者研究了逐步回归的建模原则,认为逐步回归存在许多缺陷,最主要是可能遗漏最优方程,因它在挑选因子中利用固定的阈值 F。来选择因子,使得最佳因子可能被剔除,不能选入模型中. 其次,利用 F-检验要求预测对象满足正态条件使得应用受到限制. 近年来许多统计学者对如何改进逐步回归做了许多工作,文〔1〕提出秩和逐步回归法,可摆脱正态线性模型的限制,又能尽量利用相关因子的信息. 本文研究用 *AIC*,*BIC* 准则选择秩回归变量的建模原则,在非常弱的条件下证明用 *BIC* 准则具有相容性,克服逐步回归受正态模型和固定阈值 F 的限制. 本设计用于泉州地区的降再量和气温预测有较高的回极率,预测结果与实际吻合.

1 BIC 准则的建模原理和相容性

设秩线性回归模型为

$$R(y_t) = \beta_0 + \beta_1 R(x_{1t}) + \cdots + \beta_p R(x_{pt}) + \varepsilon_t, \tag{1}$$

其中 ε_t 是独立同分布的随机误差,且 $E\varepsilon_t=0$, $D\varepsilon_t=\sigma^2$ 对给定的 N 个观测样本 $(y_t, x_{1t},\cdots,x_{pt})$ $(t=\overline{1,N})$, $(R(y_t),R(x_{1t}),\cdots,R(x_{pt}))$ 是对应的秩 $(t=\overline{1,N})$. 式(1)可写成 $R(Y)=R(x)\beta+\varepsilon$,

本文于 1992-06-26 收到.
福建省自然科学基金和国务院侨办重点学科科研资金的资助项目.

这里

$$\boldsymbol{R}(\boldsymbol{x})=\begin{bmatrix}1 & R(x_{11}) & \cdots & R(x_{p1})\\ 1 & R(x_{12}) & \cdots & R(x_{p2})\\ \vdots & \vdots & & \vdots\\ 1 & R(x_{1N}) & \cdots & R(x_{pN})\end{bmatrix}=(1,x_1,\cdots,x_p)$$

$(Y)=(R(y_1),\cdots R(y_N))^{\tau}$, $\beta=(1,\beta_1,\cdots,\beta_p)^{\tau}$, $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_N)^{\tau}$. 设参数 β_i 的真值为

$$\beta_i=\begin{cases}\beta_{i_s}(\neq 0) & \text{当 } s=1,\cdots,p,\\ 0 & \text{其它,}\end{cases}$$

$I=\{i_1\cdots i_p\}$为非零系数的足标集合. 那么选择模型(1)变量(因子)等价于估计足标 I 的真集. 定义 BIC 准则为

$$BIC(J_s)=\log S(J_s)+\frac{S\log N}{N},\tag{2}$$

其中 $J=\{j_1\cdots j_s\}$是 $J_p=\{1,2,\cdots,p\}$的子集,$S(J)$则基于下列子集 R-回归模型的最小残差平方和. 即对应子集 J_s 的回归模型

$$R(Y)=\boldsymbol{R}[X(J_s)]\beta(J_s)+\varepsilon,\tag{3}$$

这里 $\beta(J_s)=(1,\beta_{j1},\cdots,\beta_{js})$,$\boldsymbol{R}[X(J_s)]=(R(x_{j1}),\cdots,R(x_{js}))$. 由最小二乘法求得对应的参数估计和残差平方和为

$$\hat{\beta}(J_s)=[R^{\tau}(X(J_s)R(X(J_s)]^{-1}R^{\tau}(X(J_s)R(Y)\tag{4}$$

$S(J_s)=R^{\tau}(Y)R(Y)-R(Y)^{\tau}R(X(J_s))[R^{\tau}(X(J_s))R(X(J_s)]^{-1}R^{\tau}(X(J_s))R(Y)$. 令

$$\mathrm{BIC}(J^*)=\min_{1\leqslant s\leqslant p}\mathrm{BIC}(J_s),$$

$\mathrm{BIC}(J_s)$的最小值 $\mathrm{BIC}(J^*)$所对应的足标 $J^*=(i_1,\cdots,i_k)=\hat{I}$ 为 I 的估计值,则称此估计为 BIC(建模原则,其相应于 BIC 准则所确定的最佳 R-回归模型为

$$\hat{R}(Y)=\hat{\beta}_0+\hat{\beta}_{i1}R(x_{i1})+\cdots+\hat{\beta}_{i_k}\cdot R(x_{i_k})\tag{5}$$

下面进一步讨论用 BIC 准则选择秩回归变量的相容性问题.

定理 1 在模型(1)的条件下,则 有(i),(ii),(iii),分别式(6)—(8)

$$\lim_{N\to\infty}\min_{1\leqslant i\leqslant p}d_i^2(N)/\log N=\infty,\tag{6}$$

$$\lim_{N\to\infty}\log\log\max_{1\leqslant i\leqslant p}d_i(N)/\log N=0,\tag{7}$$

$$\max_{1\leqslant i\leqslant N_1}[R(x_{it})]^2=0(d_i^2(N)(\log d_i^2(N)^{-b}),\ (i=1,p,\text{对任 } b>0),\tag{8}$$

其中 $d_i^2(N)=\{R^2(x_{i1})+\cdots+\cdots+R^2(x_{iN})\}$.

证明 (i)由已知条件得$(R(x_{i1}),R(x_{i2}),\cdots,R(x_{iN}))$是对应 N 个样本$(x_{i1},x_{i2},\cdots,x_{iN})$的秩数,则$(R(x_{i1}),R(x_{i2}),\cdots,R(x_{iN}))$是自然数列$(1,2,\cdots,N)$的一个排列,所以对任 $i(1\leqslant i\leqslant p)$有

$$d_i^2(N)=\sum_{j=1}^{N}R^2(x_{ij})=\sum_{k=1}^{N}k^2=\frac{1}{6}N(N+1)(2N+1),$$

$$\lim_{N\to\infty}\min_{1\leqslant i\leqslant p}d_i^2(N)/\log N=\lim_{N\to\infty}\frac{1}{6}N(N+1)(N+2)/\log N=\infty.$$

(ii)由于 $d_i(N)=[\frac{1}{6}N(N+1)(2N+1)]^{\frac{1}{2}}(i=\overline{1,N})$,所以

$$\lim_{N\to\infty}\log\log\max_{1\leqslant i\leqslant p}d_i(N)/\log N=\lim_{N\to\infty}\log\log\left[\frac{1}{6}N(N+1)(2N+1)\right]^{\frac{1}{2}}/\log N$$

$$=\lim_{N\to\infty}\frac{[(N+1)(2N+1)+N(2N+1)+2N(N+1)]N}{\frac{1}{2}(N+1)(2N+1)N\log\frac{1}{6}N(N+1)(2N+1)}$$

$$=\lim_{N\to\infty}\frac{2(1+\frac{N}{N+1}+\frac{N}{2N+1})}{\log\frac{1}{6}N(N+1)(2N+1)}=0.$$

(iii) 由上得 $\max_{1\leqslant t\leqslant N}[R(x_{it})]^2=N^2(i=\overline{1,p})$，对任 $b>0$

$$\lim_{N\to\infty}\frac{\max_{1\leqslant t\leqslant N}[R(x_{it})]^2}{d_i^2(N)(\log d_i^2(N))^{-b}}=\lim_{N\to\infty}\frac{N^2[\log\frac{1}{6}N(N+1)(2N+1)]^b}{\frac{1}{6}N(N+1)(2N+1)}$$

$$=\lim_{N\to\infty}\frac{N^2}{\frac{1}{6}(N+1)(2N+1)}\lim_{N\to\infty}\left[\frac{\log\frac{1}{6}+\log N+\log(N+1)+\log(2N+1)}{N^{\frac{1}{b}}}\right]^b$$

$$=3\lim_{N\to\infty}\left[\frac{\log\frac{1}{6}+\log N+\log(N+1)+\log(2N+1)}{N^{\frac{1}{6}}}\right]^b$$

$$\triangleq 3\lim_{N\to\infty}[g(N)]^b,$$

$$\lim_{N\to\infty}g(N)=\lim_{N\to\infty}\frac{\log\frac{1}{6}+\log N+\log(N+1)+\log(2N+1)}{N^{\frac{1}{b}}}$$

$$=\lim_{N\to\infty}\frac{(\frac{1}{N}+\frac{1}{N+1}+\frac{2}{2N+1})}{\frac{1}{b}N^{\frac{1}{b}-1}}=\lim_{N\to\infty}b(\frac{1}{N}+\frac{1}{N+1}+\frac{2}{2N+1})N^{1-\frac{1}{b}}$$

$$=0\qquad \text{当 } b>0\ (1-\frac{1}{b}<1).$$

故对任 $b>0$，利用初等函性质有

$$\lim_{N\to\infty}[g(N)]^b=\lim_{N\to\infty}e^{b\ln g(N)}=e^{b\log(\lim_{N\to\infty}g(N))}=0,$$

即得对任 $b>0$，必有 $\max_{1\leqslant t\leqslant N}[R(x_{it})]^2=0(d_i^2(N)(\log d_i^2(N)^{-b})(i=\overline{1,p})$. 利用定理 1 及文[4]定理 2 只可得下列结论.

定理 2　设 $E\varepsilon_t^4<\infty$，且 $R[x(J_p)]$ 满足 $\lim_{N\to\infty}\Lambda^{-1}\{R(x(J_p))^tR(x(J_p))\}\Lambda^{-1}=R>0$，那么

$$\lim_{N\to\infty}\hat{I}_N=I\qquad \text{a.s}$$

其中，$\Lambda=\mathrm{diag}(d_1(N),d_2(N),\cdots,d_p(N))$.

由定理 2 可是用 BIC 准则选择秩回归变量是相容性估计.

2 秩和回归模型的 BIC 建模原则

由于 R-变换可能损失因子的信息，为克服此问题，我们引入秩和因子(具体见[1])，从而能集中多个因子相关信息，加强预测对象和因子的相关性．秩和回归模型的 BIC 建模原则为

设预测对象及相关因子分别为 $(y_t, x_{1t}, \cdots, x_{mt})$ $t=\overline{1,N}$，$(x_{10}, \cdots, x_{m0})$ 表示各因子的当前值．

2.1 对历史数据作秩变换

(1)对 Y 作秩变换 $R(y_i)=R_i(i=\overline{1,N})$.

(2)秩相关系数检验，选出与相关的因子作待选因子 $\{x_1, \cdots, x_m\}$．秩相关系数定义为 $R_{X_iY}=1-b(\sum_{k=1}^{N}[R(y_k)-R(x_{ik})]^2/N(N^2-1)$．若 $|R_{X_iY}|>R_e$，则认为 y 与 x_i 有秩相关性，即选入 X_i，否则舍弃．

(3)对 $\{x_1, \cdots, x_m\}$．进行秩变换：$R(x_{it})=R_{it}(i=\overline{1,m})$ 其中 R_{it} 为 $x_{it}\cdots x_{iN}$ 按从小到大顺序排列所对应的秩数；但若 $R_{x_iy}<0$，则 R_{it} 为 $x_{i1}\cdots x_{iN}$ 按从小到大的顺序排列所对应的秩数，当前值 $x_{i0}(i=1, \cdots, m)$ 的秩数与其对应的因子的取 N 个历史数据中最接近的那个数据的秩数．记作 $(R_{i0}(i=\overline{1,m})$.

2.2 取秩数之和构选秩和因子

用 R 型系统分析法或用彼此间相关较密切的 K 个因子归为一组．取它们的秩数之和、再经秩变换后作为一个新的因子 $\{x_1', \cdots, x_p'\}\triangleq J_p$，秩和因子的当前值取为对应秩和因子的几个历史数据中最为接近的那个数据和秩数．

2.3 利用 BIC 准则设计秩和回归模型

(1)对每个固定的 $k(k=1,2,\cdots,p)$，先找出子集 L_s 使得

$$S(L_s)=\min_{J_k}S(J_k) \qquad (S=\overline{1,p}), \tag{8}$$

这里"min"表示在所有 k 个因子的 J_p 的子集中求 $S(J_k)$ 的最小值．$S(J_k)=R^t(Y)R(Y)-R^t(Y)X'(J_k)[X'^t(J_k)X'(J_k)]^{-1}X'^t(J_k)R(Y)$，$R(Y)=(R(y_1), R(y_2), \cdots, R(y_N))^t$.

(2)计算 BIC 准则函数值

$$\mathrm{BIC}(L_s)=\log S(L_s)+\frac{S\log N}{N}, \ (S=\overline{1,p}). \tag{9}$$

求 $\mathrm{BIC}(L_s)$ 的最小值所对应的 $J^*=\{i_1, \cdots, i_{\hat{p}}\}$，即

$$\mathrm{BIC}(J^*)=\min_{1\leqslant s\leqslant p}\mathrm{BIC}(L_s), \tag{10}$$

故所对应于 BIC 准则的最佳秩和回归模型为

$$\hat{R}(y_t)=\hat{\beta}_0+\hat{\beta}_{i_1}x'_{i_1t}+\cdots+\hat{\beta}_{i_{\hat{p}}}x'_{i_{\hat{p}}t}. \tag{11}$$

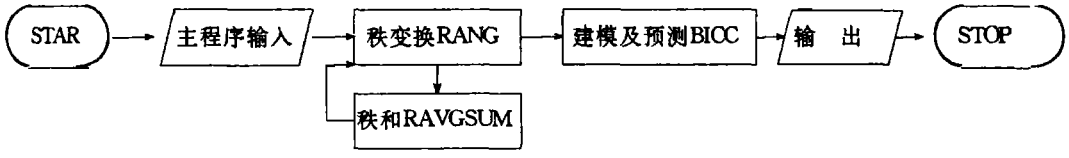
2.4 求 Y 的预极值及回极值

以秩和因子的当前值 $(x'_{i_10}, \cdots, x'_{i_{\hat{p}}0})$ 代入式(11)求得 $\hat{R}(y_0)$，然后将 $\hat{R}(y_0)$ 与 $R(y_t)$ $(t=\overline{1,N})$ 进行比较以最接近的那个历史数值作为 Y_0 的预极值．相似地，以秩和因子 $(x'_{i_1t}, \cdots, x'_{i_{\hat{p}}t})$

$(t=\overline{1,N})$代入式(11)，可求得回极值．

3　计算程序设计

本计算程序设计由主程序模块，秩变换(RANG)模块，秩和(RANGSUM)模块与 BIC 建模及预测(BICC)模块组成．具体流程为



其秩变换模块的功能是对历史数据作秩变换，RNAG 模块的设计摒弃了常用的先排序再由顺序定秩的方法，而是借鉴快速分类方法，边比较大小边求秩，同时也兼顾历史数据相等的情况，方法新颖，计算简便．BICC 模块是用 OLS 方法建立秩和回归模型、并利用 BIC 信息准则选择最佳线性方程，最后作出预报及结果分析．设计中考虑到反复用 OLS 方法，计算回归系数，残差平方和以及 BIC 准则函数值．即

$$\hat{\beta}(J_s)=(X^{r\mathrm{T}}(J_s)X^r(J_s))^{-1}X^{r\mathrm{T}}(J_s)R(Y),$$

$$\hat{S}(J_s)=R^r(Y)[I-X^r(J_s)(X^{r\mathrm{T}}(J_s)X^r(J_s))^{-1}X^{r\mathrm{T}}(J_s)]R(Y)$$

$$BIC(J_s)=\log\hat{S}(J_s)+\frac{S\log N}{N}\quad(S=\overline{1,p}),$$

$$X(J_p)=\begin{bmatrix}1 & x_{11} & \cdots & x_{p1}\\ 1 & x_{12} & \cdots & x_{p2}\\ \vdots & \vdots & & \vdots\\ 1 & x_{1N} & \cdots & x_{pN}\end{bmatrix}$$

其中，$X(J_p)$为 $N\times(1+p)$阶矩阵，对于 $k=\overline{1,p}$，均应对矩阵 $X(J_p)$其它工作矩阵进行数组说明．而针对不同的预极对象，因秩和因子的个数不同，即 p 无法固定下来．若将 p 确定为一个较大的数，势必浪费不少计算内存．因而，设计该模块采用了一维数组，根据 FORTRAN 语言的数组的存储顺序，将二维数组转变成一维数组，从而摆脱了 p 的限制，大大节省内存．我们已用 FORTRAN 语言设计其程序软件，能完成整个计算过程．

4　实例预测及效果分析

根据泉州市气象台 1959—1989 的气象资料，利用本设计方法，对永春 6 月降再量 $Y(100)$进行预极．

利用秩相关系数法，在显著水平 $\alpha=0.05$ 下选出 65 个与 $Y(100)$有统计相关的因子，下面只列出最后参加预测的秩和因子的组成因子．X_1 为崇武 1 月平均气温，X_{57}为崇武 3 月降雨量，X_{108}为崇武 3 月下旬日照，X_{109}为崇武 4 月日照，X_{115}为崇武 5 月日照，X_{294}为永春 1 月上旬平均气温，X_{291}为永春 1 月中旬日照，X_{340}为 2 月最高亚欧$(45'-65'N/0-150'E)$500mb 环

流指数，X_{350}为7月最高亚欧(45′－65′N/0－150′E)500mb环流指数，X_{383}为1月500mb西太平洋副G面积指数，X_{387}为5月500mb西太平洋副G面积指数，X_{391}为9月500mb西太平洋副G面积指数，X_{399}为5月500mb西太平洋副G强度指数，X_{400}为6月500mb西太平洋副G强度指数，X_{401}为7月500mb西太平洋副G强度指数，X_{407}为1月500mb西太平洋平均副G西伸脊点位置，X_{450}为6月155′－180′E 10′N以北副G面积指数，X_{451}为7月155′－180′E 10′N以北副G面积指数，X_{460}为4月110′－150′E 10′N以北副G面积指数，X_{473}为5月65′－180′E 10′N以北副G面积指数，X_{478}为10月65′－180′E 10′N以北副G面积指数，X_{512}为10－11月155′－180′E 10′N以北副G面积指数，X_{530}为上平8月150′－160′E 40′NH，X_{544}为8月H〔60E/60N＋40E/50N〕，X_{553}为9月500mb平均高度均因子．秩和因子及其组成因子的相关系数比较如表1所示．

表1 $x_1'-x_{13}'$的秩相关系数表

x_1'	x_2'	x_3'	x_4'
$R(1)=0.6829$	$R(2)=0.5899$	$R(3)=0.61135$	$R(4)=0.7083$
$R_{xy}(450)=-0.5691$	$R_{xy}(475)=-0.5447$	$R_{xy}(2)=-0.5188$	$R_{xy}(401)=-0.4814$
$R_{xy}(460)=-0.54917$	$R_{xy}(193)=-0.5403$	$R_{xy}(389)=-0.5103$	$R_{xy}(108)=0.4609$
$R_{xy}(530)=-0.5226$	$R_{xy}(472)=-0.51190$	$R_{xy}(105)=0.4734$	$R_{xy}(194)=-0.4481$
$R_{xy}(451)=-0.4974$	$R_{xy}(195)=-0.4897$	$R_{xy}(155)=0.4554$	$R_{xy}(340)=0.4478$
$R_{xy}(1)=-0.4974$	$R_{xy}(457)=-0.4792$	$R_{xy}(388)=-0.4528$	$R_{xy}(512)=-0.4406$
x_5'	x_6'	x_7'	x_8'
$R(5)=0.60278$	$R(6)=0.6829$	$R(7)=0.5967$	$R(8)=0.5844$
$R_{xy}(153)=-0.4682$	$R_{xy}(395)=-0.4740$	$R_{xy}(252)=-0.4345$	$R_{xy}(251)=-0.4367$
$R_{xy}(524)=-0.4406$	$R_{xy}(249)=-0.4509$	$R_{xy}(449)=0.4115$	$R_{xy}(448)=-0.3863$
$R_{xy}(445)=-0.43972$	$R_{xy}(477)=-0.4464$	$R_{xy}(113)=-0.3928$	$R_{xy}(466)=0.3850$
$R_{xy}(107)=0.43815$	$R_{xy}(372)=-0.4149$	$R_{xy}(508)=-0.3887$	$R_{xy}(520)=-0.3833$
$R_{xy}(474)=-0.43726$	$R_{xy}(567)=-0.4122$	$R_{xy}(603)=0.3878$	$R_{xy}(452)=-0.3801$
x_9'	x_{10}'	x_{11}'	x_{12}'
$R(9)=0.5958$	$R(10)=0.7268$	$R(11)=0.5998$	$R(12)=0.5283$
$R_{xy}(383)=-0.4792$	$R_{xy}(407)=0.4249$	$R_{xy}(299)=0.4263$	$R_{xy}(59)=-0.3523$
$R_{xy}(478)=-0.3668$	$R_{xy}(462)=-0.3566$	$R_{xy}(473)=-0.3964$	$R_{xy}(348)=0.3566$
$R_{xy}(291)=0.3659$	$R_{xy}(297)=0.3411$	$R_{xy}(544)=-0.3601$	$R_{xy}(463)=0.3609$
$R_{xy}(399)=-0.3637$	$R_{xy}(3)=0.3382$	$R_{xy}(350)=-0.3548$	$R_{xy}(607)=-0.3503$
$R_{xy}(109)=-0.3633$	$R_{xy}(166)=-0.3378$	$R_{xy}(553)=-0.3538$	$R_{xy}(212)=0.3473$
x_{13}'			
$R(13)=0.4716$			
$R_{xy}(387)=-0.3772$			
$R_{xy}(57)=-0.3646$			
$R_{xy}(400)=-0.3528$			
$R_{xy}(115)=0.3425$			
$R_{xy}(391)=0.3413$			

利用秩和回归模型BIC准则建模原则，得到最优回归模型为

$$\hat{R}(y)=1.4929+0.4766X_1'+0.3882X_4'+0.2283X_{10}'+0.3854X_9'-0.5748X_{13}'$$

由表2可见当允许误差为56.61mm(即极差的15%)时回极率为90%，预测1990年6月份永

春降雨量为 407.00mm. 还对泉州 5 月下旬平均气温等气象指标进行预极，都达到转好的预极效果.

表 2　预测效果分析表

年　份	实 测 值	预 测 值	预测误差	效果评定
1960	401.2	383.8	17.4	✓
1961	133.4	203.6	−70.2	×
1962	393.5	401.2	−8.2	✓
1963	331.4	306.5	24.9	✓
1964	381.2	355.2	26.2	✓
1965	350.5	306.5	44.0	✓
1966	383.8	383.8	0	✓
1967	150.3	150.8	0	✓
1968	407.8	381.4	26.4	✓
1969	277.3	277.3	0	✓
1970	297.8	306.5	−8.7	✓
1971	228.3	255.2	−26.9	✓
1972	423.7	423.7	0	✓
1973	192.3	227.1	−34.8	✓
1974	267.2	297.8	−30.6	✓
1975	267.2	297.8	−30.6	✓
1976	255.2	348.6	−93.4	×
1977	355.2	350.5	4.7	✓
1978	394.9	393.0	1.9	✓
1979	304.1	319.0	−15.9	✓
1980	306.5	303.1	3.4	✓
1981	46.3	46.3	0	✓
1982	269.9	228.3	41.6	✓
1983	221.9	227.1	−5.2	✓
1984	203.6	259.5	−55.9	✓
1985	227.1	221.9	5.2	×
1986	319.0	350.5	−31.5	✓
1987	348.0	267.2	81.4	×
1988	181.3	181.3	0	✓
1989	259.5	212.3	47.2	✓

参　考　文　献

[1] 吴绍敏、陈治典，数理统计及应用概率，(1987)，

[2] Beale, E. M. L., *Technometrics*, 12, 4 (1970), 909−914.

[3] Hocking, R. R., *Biometrics*, 32, 1(1974), 1−49.

[4] An Hong Zhi and Gu Lan, *Acta, math., Appl. sinia (English series)*. 2. 1 (1985), 98−105.

[5] 孙兰芬，应用数学报，13, 1(1990), 74−82.

The Compatibility of Rank and Regression Model formed by BIC Modeling and Its Applications

Chen Jianwei

(*Department of Management Information Science*)

Abstract Starting from the predicted method of rank and regression model proposed by reference (1), the author goes further to the compatibility of rank variables selected by BIC criterion and the principle of modeling. the specific method of model design is given and applied to meterological prediction with good results.

Key words rank and regression model, BIC criterion, compatibility