中山大学学报 (自然科学版)
第 37卷 第 5期 ACTA SCIENTIARUM NATURALIUM Vol. 37 No. 5
1998年 9月 UNIVERSITATIS SUNYATSENI Sep. 1998

逐步回归分析的拓展*

张华嘉 舒 元
(中山大学岭南学院, 广州 510275)

摘 要 给出一种方法, 在逐步回归后, 使每一前进 (后退) 步为一单元步, 用 F 检验来判别这单元步的去留, 使回归式子比逐步回归纳入更多的解析变量, 同时能减低向后消元法产生多重共线性的可能性.

关键词 逐步回归, 单元步, 多重共线性, F检验
分类号 O 212.1

1 提出问题

在回归分析中, 对解析变量的选择很重要. 逐步回归法能使回归式子保留几个最为显著的解析变量, 但由于其每一前进步只选择 1个显著变量, 很可能排斥了一些重要变量.

在表 1中, 1代表农业, 21代表制造业, 22代表建筑业, 31代表邮电运输业, 32代表商业, 33代表非物质服务业, Xij 代表 ij 行业滞后一年的国内生产总值. X 代表实际国内生产总值的增长量.

下面的所有回归都去掉了 1991年 (由于回归滞后一阶, 实际是去掉 1990年).

用向后消元法得

$$X = -2\,866.63 + 3.26X_1 - 0.97X_{21} + 12.66X_{22} + 6.93X_{31} - 6.50X_{32} \quad (1)$$
$$(-5.20)\quad (4.50)\quad (-3.02)\quad (3.19)\quad (2.12)\quad (-5.35)$$
$$R^2 = 0.94 \qquad DW = 2.00 \qquad ESS = 314\,557.0$$

用逐步回归法得

$$X = -2\,845.68 + 3.69X_1 - 3.04X_{32} \quad (2)$$
$$(-5.73)\quad (5.88)\quad (-4.10)$$
$$R^2 = 0.87 \qquad DW = 1.45 \qquad ESS = 680\,598.9$$

与 (1) 式相比, (2) 式中保留了最显著变量 X_1, X_{32}而去掉了变量 X_{21}, X_{22}, X_{31}.

人们面临的问题是: 一方面逐步回归能避免共线性, 但入选的变量很少. 在某些情况下与分析目标不符; 另一方面, 向后消元法虽然纳入较多变量, 但容易导致共线性, 从而损害了回归结论的精确性. 为避免多重共线性又能纳入较多变量, 目前有 2类方法, 第一类方法如岭回归, 主成分回归, 偏最小二乘法.[1]但这些方法是以有偏估计为代价的. 另一

* 收稿日期: 1998-03-02 张华嘉, 男, 34岁, 讲师

表 1 1978~ 1995年各行业实际产值和实际国内生产总值增长量

Tab. 1 The real output values in various industries and the increase of real GDP from 1978 to 1995 10^8元

年份	X	X_1	X_{21}	X_{22}	X_{31}	X_{32}	X_{33}
1978		1 018. 40	1 607. 00	138. 20	172. 80	265. 50	422. 20
1979	275. 43	1 080. 52	1 746. 81	140. 96	186. 11	288. 86	452. 65
1980	304. 42	1 065. 25	1 966. 97	178. 55	196. 65	285. 15	500. 90
1981	221. 07	1 139. 59	2 000. 72	184. 22	200. 45	370. 64	514. 87
1982	36. 24	1 270. 96	2 116. 42	190. 58	223. 78	385. 24	618. 06
1983	887. 90	1 375. 86	2 322. 12	223. 05	246. 24	469. 40	698. 16
1984	81. 80	1 554. 08	2 667. 62	247. 38	283. 05	570. 29	834. 10
1985	829. 92	1 582. 59	3 152. 93	302. 24	321. 24	734. 90	939. 36
1986	619. 72	1 634. 53	3 458. 26	350. 20	362. 36	812. 70	1 062. 24
1987	880. 66	1 711. 93	3 914. 65	412. 80	398. 65	922. 08	1 238. 40
1988	956. 76	1 754. 70	4 512. 46	408. 10	473. 13	965. 89	1 614. 04
1989	384. 15	1 808. 68	4 740. 65	408. 10	473. 13	965. 89	1 614. 04
1990	376. 91	1 942. 09	4 899. 74	412. 94	513. 56	919. 96	1 690. 10
1991	938. 64	1 987. 92	5 605. 22	452. 47	571. 10	961. 38	1 866. 50
1992	1 587. 36	2 081. 61	6 791. 18	547. 55	631. 07	1 086. 96	2 102. 60
1993	1 717. 82	2 179. 38	8 155. 53	646. 09	709. 34	1 158. 64	2 359. 65
1994	1 830. 17	2 266. 96	9 698. 25	734. 53	776. 74	1 255. 28	2 600. 05
1995	1 656. 21	2 368. 97	11 059. 37	906. 54	850. 52	1 364. 67	2 787. 45

资料来源:《中国统计年鉴》① 各年算出

类方法是先用向后消元法得

$$Y = U_0 + U_1 X_1 + \cdots + U_n X_n \tag{3}$$

再用一些方法检验 $X_1, \cdots, X_n$ 是否有多重共线性, 然后去掉引起共线性的变量, 如 Farrar-Glauber检验和 Klein方法[2]. Farrar-Glauber检验, 对任一 $i = 1, 2, \cdots, n$, 对 $X_i = U_0 + U_1 X_1 + \cdots + U_{i-1} X_{i-1} + U_{i+1} X_{i+1} + \cdots + U_n X_n$ 回归, 得判别系数 R_i^2, 令

$$F_i = \frac{R_i^2 / (n-1)}{(1-R_i^2) / (N-n)} \sim F(n-1, N-n) \tag{4}$$

如 $F_i > F_{0.05}$, 则认为 X_i 是产生共线性的因素, 从而在回归中去掉 X_i;

如 $F_i < F_{0.05}$, 则认为 X_i 不是产生共线性的因素, 从而在回归中保留 X_i.

这方法是利用公式[3] $D(\hat{U}_i) = e^2 / (1-R_i^2) \sum_{j=1}^{N} (X_{ij} - \bar{X}_i)^2$ 当 R_i^2 越接近 1时, $D(\hat{U}_i)$ 越大, 从而估计精确性越差. $F_i > F_{0.05}$ 只是利用 F 分布给出了 1个临界值, 说明 R_i^2 已足够接近 1, 从而 $D(\hat{U}_i)$ 足够大, 应该去掉 X_i.

Klein判别公式: 设 $V_{X_i X_j}$ 为 X_i, X_j 的样本相关系数. R^2 为 (3) 式的判别系数. 如 $V_{X_i X_j} \geqslant R^2$ 时, $D(\hat{U}_i)$ 足够大, 应该去掉 X_i.

在用上述 2个方法判别 X_i 的去留时, 会产生逐步回归的效果, 即在消除多重共线性同时, 可能删掉一些对 Y 有解析力的变量, 如用 Farrar-Glauber检验和 Klein判别公式对表 1的数据进行回归, 都得到 (2)式, 与 (1)式相比, 可知 (2)式漏掉几个有解析力的变量.

在回归方程显著的解析变量超出了逐步回归入选范围情况下, 本文提出 1个根据其对

① 中国统计年鉴. 北京: 中国统计出版社, 1994~ 1997

Y的解析力来判别解析变量去留的方法.

2 本文方法

设 Y为被解析变量, $X_1,\cdots,X_n$; $Z_1,\cdots,Z_k$ 为解析变量, 得回归式子 $Y=U_1X_1+U_2X_2+\cdots+U_nX_n+V_1Z_1+\cdots+V_kZ_k$, 用逐步回归法得

$$Y=U_1X_1+U_2X_2+\cdots+U_nX_n \tag{5}$$

在保留 $X_1,\cdots,X_k$ 的前提下用向后消元法回归, 得

$$Y=U_1X_1+U_2X_2+\cdots+U_nX_n+V_1Z_1+\cdots+V_mZ_m \tag{6}$$

本文方法要点: ① 单元步的定义,② 在只含 1个单元步情形时判别这单元步去留的方法, ③ 在有多个单元步时判别这些单元步去留的方法.

下面给出一种方法判别 $Z_1,\cdots,Z_m$ 中哪些变量可以保留, 哪些变量应该删去.

2.1单元步的定义

在 (6) 式中, 称 $Z_{i_1},Z_{i_2},\cdots,Z_{i_l}$为一单元步. 如去掉 $Z_{i_1},\cdots,Z_{i_l}$中任一部分变量, 则 $Z_{i_1},\cdots,Z_{i_l}$中其余变量总存在不显著变量. 如去掉所有 $Z_{i_1},\cdots,Z_{i_l}$变量时, (6)式中其余解析变量都显著.

找单元步变量的方法: 在 (6) 式中, 不妨设 $Z_1,\cdots,Z_m$ 的显著程度不断下降, 在 (6) 式中去掉 Z_m 如剩下的解析变量显著, 则 Z_m为一单元步. 如剩下的变量有部分不 显著, 不妨设 Z_{m-1}最不显著, 去掉 Z_{m-1}. 重复上述步骤, 只要剩下变量出现不显著, 就去掉最不显著变量, 直到剩下所有变量都显著为止. 不妨设去掉 $Z_{m-k},\cdots,Z_m$后, 剩下所有变量都显著, 这时可通过定义 1验证 $Z_{m-k},\cdots,Z_m$是否为一单元步.

由于在 (1) 式中去掉 X_{21}, X_{22}, X_{31}中任何 1个或 2个变量, 这 3个中剩下的变量都不显著, 如同时去掉这 3个变量, (1) 式变为 (2) 式, 各变量显著, 所以上述 3个变量组成 1个单元步.

2.2只含 1个单元步情形的判别方法

引用 (5), (6) 式, 设 (5) 式为逐步回归的结果, (6) 式为向后消元法的结果, 且 $Z_1,\cdots,Z_m$ 组成一单元步. 设 (5), (6) 式的残差平方和分别为 ESS_1 和 ESS_2.

令

$$F=\frac{(ESS_1-ESS_2)/n}{ESS_2/(N-m-n)}\sim F(n,N-m-n) \tag{7}$$

如 $F>F_{0.05}$, 单元步 $Z_1,\cdots,Z_m$ 对 Y有解析力, 则单元步所有变量应该保留;

如 $F<F_{0.05}$, 单元步 $Z_1,\cdots,Z_m$ 对 Y无解析力, 则单元步应该删去.

在这里保留或删除整个单元步,而不是其中一部分. 这是因为仅删除任一部分变量,该单元步中余下的某些变量就不显著. 由于含有不显著变量, 应以删除, 直到删除整个单元步, 这就是单元步的含义.

2.3含多个单元步情形的判别方法

引用 (5), (6) 式, 设 (6) 式中, $Z_1,\cdots,Z_t$ 为一单元步, $Z_{t+1},\cdots,Z_m$ 为另一单元步, 用只含 1个单元步情形的判别方法分别判别这 2单元步的去留: $Y=U_1X_1+\cdots+U_nX_n+V_1Z_1+\cdots+V_tZ_t$, 得 F_1为 (7) 式的 F值; $Y=U_1X_1+\cdots+U_nX_n+V_{t+1}Z_{t+1}+\cdots+V_mZ_m$, 得 F_2为 (7) 式的 F值.

不妨设 $F_1>F_2$, $F_1>F_{0.05}$, 则 $Z_1,\cdots,Z_t$ 可保留, 在此基础上, 用只含 1个单元步情形

的判别方法判别 $Y = U_1X_1 + \cdots + U_nX_n + V_1Z_1 + \cdots + V_tZ + V_{t+1}Z_{t+1} + \cdots + V_mZ_m$ 来决定 Z_{t+1}, $\cdots$, Z_m 可否保留.

这个方法的本质是找出各单元步后, 由于同一单元步的变量是同时去留, 因此可当每一单元步为一变量进行回归, 特别地可以进行逐步回归. 这就使得原逐步回归每一前进(后退) 步为一变量, 而现在推广为 1个单元步.

最后, 用本文方法、Farrar-Glauber检验和 Klein方法来判别表 1中变量的去留, 并比较 3个结论.

用本文方法: 由于 (1)式中 $ESS_1 = 314\ 557.0$, (2)式中 $ESS_2 = 680\ 598.9$, 由 (7)式得, $F = 3.88 > 3.71 = F_{0.05}(3, 10)$, 所以单元步 X_{21}, X_{22}和 X_{31}对 X 有解析力, 应以保留, 即 (1)式可以接受.

Farrar-Glauber检验: 对解析变量之间进行回归分析, 得 X_{31}对其他解析变量回归所得的判别系数 $R^2 = 0.998\ 786$最大. 由 (4)式得 $F = 1\ 234.09 > 3.37 = F_{0.01}(6, 9)$, 所以, X_{31}是产生多重共线性的原因, (1)式应去掉 X_{31}后再进行回归, 回归结果与 (2)式相同.

Klein方法: X_{31}与 X_{21}的样本相关系数为 0.989, 其值大于 (1)式的判别系数 $R^2 = 0.94$. 由 Klein判别得 X_{31}与 X_{21}是产生共线性的原因, 所以 (1)式应先去掉 X_{31}或 X_{21}后再回归. 其结果都是变为 (2)式.

总结来说, 用本文方法判别解析变量的去留, 得 (1)式. 用逐步回归、Farrar-Glauler检验和 Klein方法判别则得 (2)式. 比较 (1), (2)式可认为 (1)式更可取. 理由是: (1)式比 (2)式有更多的解析变量, 对行业结构分析有利. (1)中所反映关于 X_{21}, X_{22}, X_{31}的情况与经济直观相符.

参 考 文 献

1 何晓群. 回归分析与经济数据建模. 北京: 中国人民大学出版社, 1997. 297, 321, 356
2 吴承业. 统计计量学概论. 北京: 中国铁道出版社, 1987. 125~130
3 刘振亚. 计量经济学教程. 北京: 中国人民大学出版社, 1997. 121

Extension of Stepwise Regression

Zhang Huajia[*] *Shu Yuan*

Abstract A method is proposed to incorporate more independent variables into the regression than the stepwise regression does, and to reduce the possibility of multi-collinearity caused by backward regression. Each forward (backward) step following the stepwise regression is a unit step and F test is used to judge whether this unit step is remained.

Keywords stepwise regression, unit step, multi-collinearity, F test

[*] Lingnan College, Zhongshan University, Guangzhou 510275, China