第7章　岭回归

1. 岭回归估计的定义
2. 岭回归估计的性质
3. 岭迹分析
4. 岭参数 k 的选择
5. 用岭回归选择变量

7.1 岭回归估计的定义

7.1.1 普通最小二乘估计带来的问题

当自变量间存在**复共线性**时，回归系数估计的方差就很大，估计值就很不稳定，下面进一步用一个模拟的例子来说明这一点。

例7-1 假设已知x_1，x_2与y的关系服从线性回归模型

$$y = 10 + 2x_1 + 3x_2 + \varepsilon$$

7.1 岭回归估计的定义

给定 x_1, x_2的10 个值，见表7-1 的第(1)、(2)两行。

表 7-1

序号		1	2	3	4	5	6	7	8	9	10
(1)	x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
(2)	x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
(3)	ε_i	0.8	−0.5	0.4	−0.5	0.2	1.9	1.9	0.6	−1.5	−1.5
(4)	y_i	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

然后用模拟的方法产生10个正态随机数，作为误差项$\varepsilon_1,\varepsilon_2,\ldots\varepsilon_{10}$，见表7-1的第（3）行。然后再由回归模型计算出10个y_i 值，且列在了表7-1的第（4）行。

$$y_i = 10 + 2x_{i1} + 3x_{i2} + \varepsilon_i$$

7.1 岭回归估计的定义

现在我们假设回归系数与误差项是未知的，用普通最小二乘法求回归系数的估计值得：

$$\hat{\beta}_0 = 11.292,\quad \hat{\beta}_1 = 11.307,\quad \hat{\beta}_2 = -6.591$$

而原模型的参数

$$\beta_0 = 10,\ \beta_1 = 2,\ \beta_2 = 3$$

看来相差太大。计算 x_1，x_2 的样本相关系数得 $r_{12} = 0.986$，表明 x_1 与 x_2 之间高度相关。

备注： 注意这里讲的是**估计值和真实值**相差很大，不是偏差很大，偏差是统计学上特定的概念，是指 $E(\hat{\beta}) - \beta$

7.1 岭回归估计的定义

7.1.2 岭回归的定义

岭回归(Ridge Regression，简记为RR)提出的想法是很自然的。

当自变量间存在**复共线性时**，$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$，我们设想给$\boldsymbol{X}'\boldsymbol{X}$加上一个正常数矩阵$k\boldsymbol{I}$ $(k > 0)$，那么$\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}$接近奇异的程度就会比$\boldsymbol{X}'\boldsymbol{X}$接近奇异的程度小得多。

考虑到变量的量纲问题，我们先对数据做标准化，为了计算方便，标准化后的设计阵仍然用$\boldsymbol{X}$表示。

7.1 岭回归估计的定义

我们称

$$\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad (7.2)$$

为 β 的岭回归估计，其中 k 称为岭参数。

（7.2）式中因变量观测向量$\boldsymbol{y}$可以经过标准化也可以未经 标准化。由于假设$\boldsymbol{X}$已经标准化，如果$\boldsymbol{y}$也经过标准化，那 么（7.2）式计算的实际是标准化岭回归估计。

显然，岭回归作为$\boldsymbol{\beta}$的估计应比最小二乘估计稳定，

备注： 当 k=0时的岭回归估计就是普通最小二乘估计。

7.1 岭回归估计的定义

岭回归的另一种定义是**带约束条件**的最小二乘估计，对系数的模平法加上一个限制

$$\hat{\beta} = \arg\min_{\beta}(\mathbf{Y} - \mathbf{X}\beta)^{\top}(\mathbf{Y} - \mathbf{X}\beta), with||\beta||^2 \leq M, M > 0$$

根据**拉格朗日乘数法**，这等价于最小化如下式子

$$\hat{\beta} = \arg\min_{\beta,k}(\mathbf{Y} - \mathbf{X}\beta)^{\top}(\mathbf{Y} - \mathbf{X}\beta) + k(||\beta||^2 - M)$$

求导即可解出式子（7.2）。

思考： M与岭参数K的关系？

7.1 岭回归估计的定义

因为岭参数 k 不是唯一确定的，所以我们得到的岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 实际是回归参数 β 的一个估计族。

例如对例 7.1 可以算得不同 k 值时的 $\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$，见表 7.2

表 7-2

k	0	0.1	0.15	0.2	0.3	0.4	0.5	1.0	1.5	2	3
$\hat{\beta}_1(k)$	11.31	3.48	2.99	2.71	2.39	2.20	2.06	1.66	1.43	1.27	1.03
$\hat{\beta}_2(k)$	−6.59	0.63	1.02	1.21	1.39	1.46	1.49	1.41	1.28	1.17	0.98

7.1 岭回归估计的定义

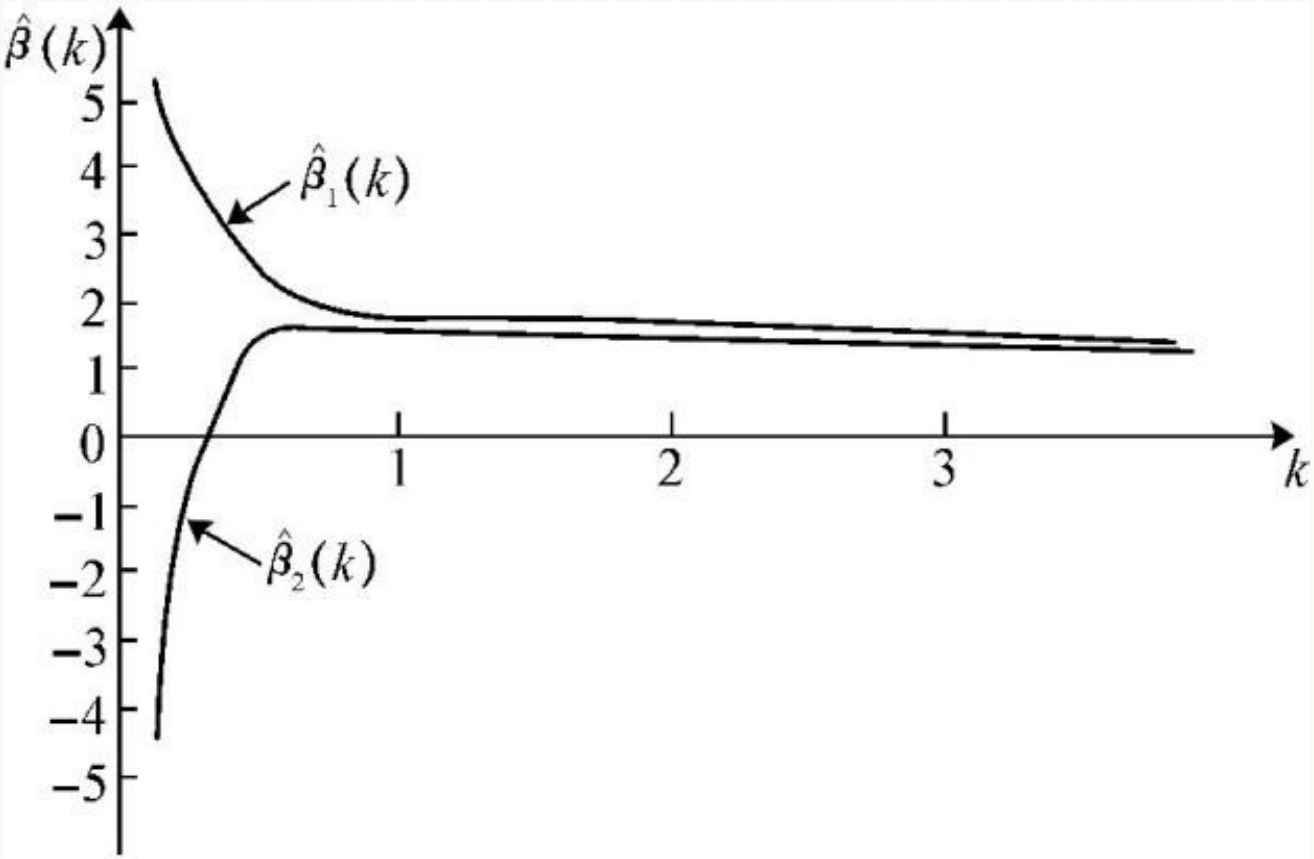


图7-1 岭迹图

7.2 岭回归估计的性质

在本节岭回归估计的性质的讨论中，假定（7.2）式中因变量观测向量y**未经标准化**。

性质 1 $\hat{\beta}(k)$ 是回归参数 β 的有偏估计。

证明：
$$E[\hat{\beta}(k)]=E[(X'X+kI)^{-1}X'y]$$
$$=(X'X+kI)^{-1}X'E(y)$$
$$=(X'X+kI)^{-1}X'X$$

显然只有当 $k=0$ 时，$E[\hat{\beta}(0)]=\beta$；当 $k\neq 0$ 时，$\hat{\beta}(k)$ 是 β 的有偏估计。

要特别强调的是 $\hat{\beta}(k)$ 不再是 β 的无偏估计了，

有偏性是岭回归估计的一个重要特性。

7.2 岭回归估计的性质

性质 2 在认为岭参数 k 是与 y 无关的常数时，$\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，也是 y 的线性函数。

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

$$=(X'X+kI)^{-1}X'X\,\hat{\boldsymbol{\beta}}$$

因此，岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，

根据定义式 $\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 知 $\hat{\boldsymbol{\beta}}(k)$ 也是 y 的线性函数。

这里需要注意的是，在实际应用中，由于岭参数 k 总是要通过数据来确定，因而 k 也依赖于 y，因此从本质上说 $\hat{\boldsymbol{\beta}}(k)$ 并非 $\hat{\boldsymbol{\beta}}$ 的线性变换，也不是 y 的线性函数。

7.2 岭回归估计的性质

性质 3 对任意 $k>0$，$\|\hat{\boldsymbol{\beta}}\| \neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\| < \|\hat{\boldsymbol{\beta}}\|$$

这里 $\|\bullet\|$ 是向量的模，等于向量各分量的平方和。

这个性质表明 $\hat{\boldsymbol{\beta}}(k)$ 可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种向原点的压缩，从 $\hat{\boldsymbol{\beta}}(k)$ 的表达式可以看到，当 $k\to\infty$ 时，$\hat{\boldsymbol{\beta}}(k)\to 0$，即 $\hat{\boldsymbol{\beta}}(k)$ 化为零向量。

7.2 岭回归估计的性质

性质 4 以 MSE 表示估计向量的均方误差，则存在 k>0，使得

$$MSE(\hat{\boldsymbol{\beta}}(k)) < MSE(\hat{\boldsymbol{\beta}})$$

即

$$\sum_{i=1}^{p} E(\hat{\beta}_j(k)-\beta_j)^2 < \sum_{i=1}^{p} D(\hat{\beta}_j)$$

备注：1.存在k>0, 而不是对所有k;
2. MSE分解公式如下

$$MSE(\hat{\theta}) = E(\hat{\theta}-\theta_0)^2 = bias^2 + Var(\hat{\theta})$$

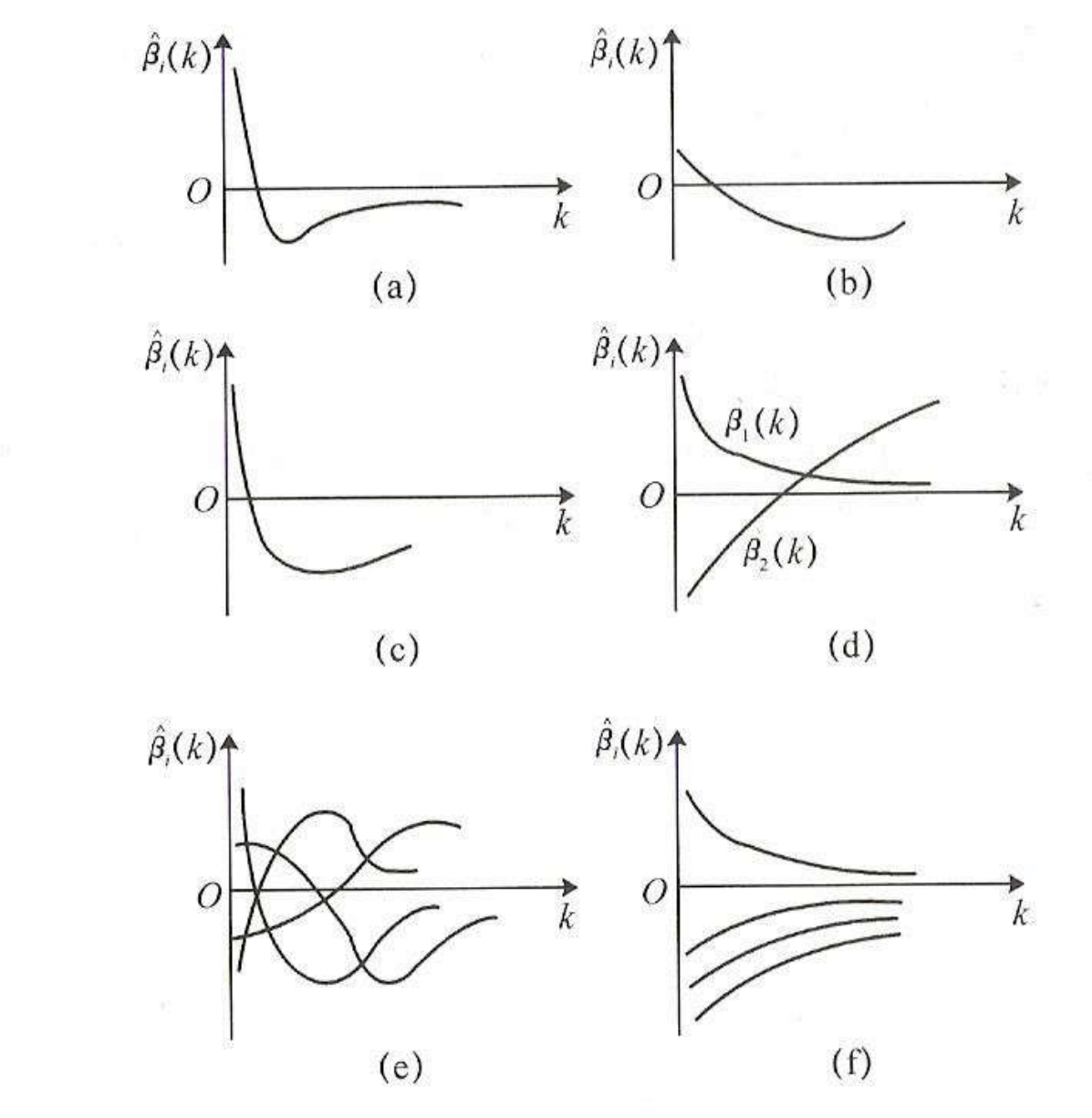
$$bias = E(\hat{\theta}) - \theta_0$$

7.3 岭迹分析

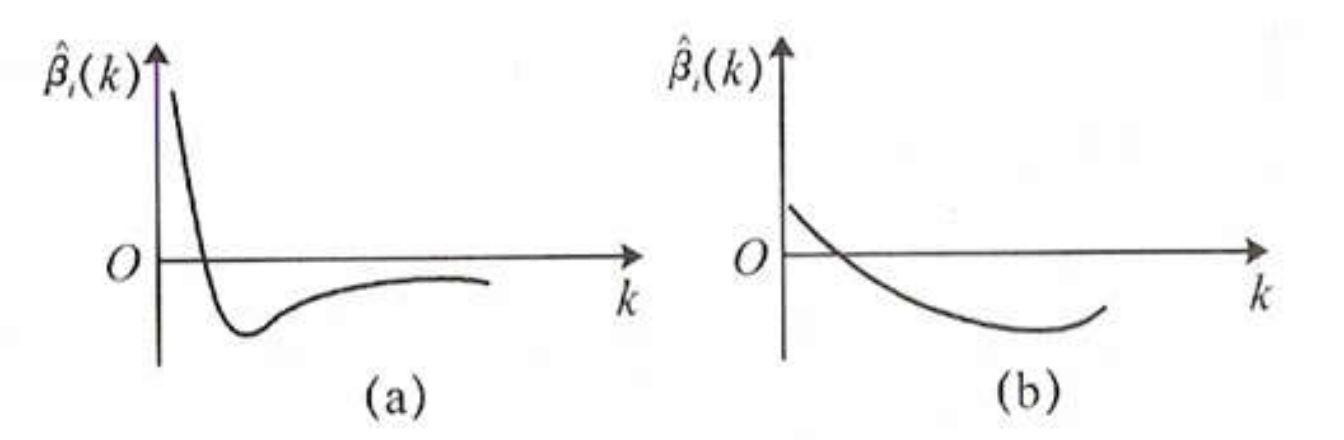
当岭参数k在（0，∞）内变化时，$\hat{\beta}_j(k)$是k的函数，在平面坐标系上把函数$\hat{\beta}_j(k)$描画出来。画出的曲线称为岭迹。在实际应用中，可以根据岭迹曲线的变化形状来确定适当的k值和进行自变量的选择。

在岭回归中，岭迹分析可用来了解各自变量的作用及自变量间的相互关系。下面由图7.2所反映的几种有代表性的情况来说明岭迹分析的作用。

岭迹分析



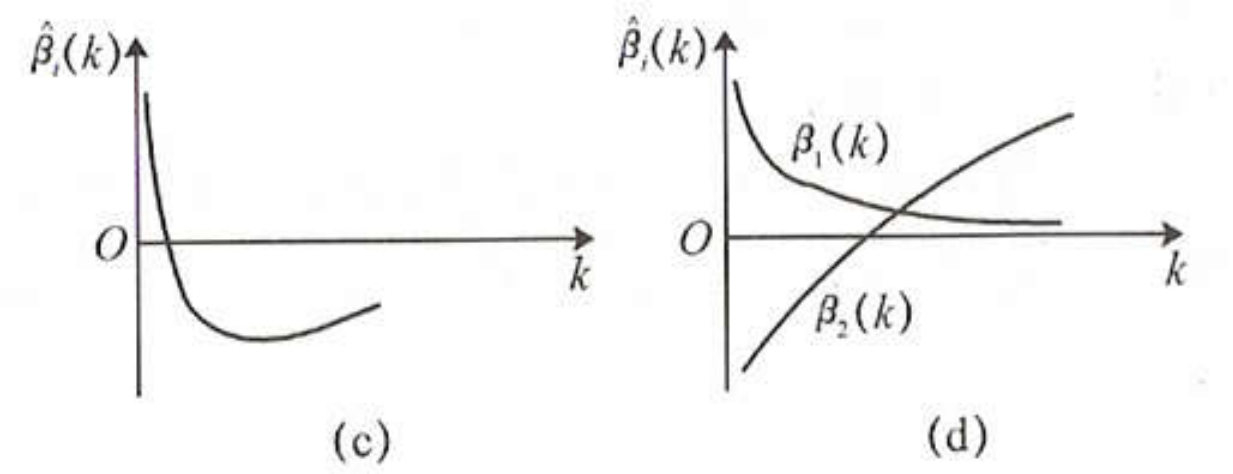
7.3 岭迹分析



（1）在图 7－2（a）中，$\hat{\beta}_j(0)=\hat{\beta}_j>0$，且比较大。从古典回归分析的观点看，应将 x_j 看作对 y 有重要影响的因素。但 $\hat{\beta}_j(k)$ 的图形显示出相当的不稳定性，当 k 从零开始略增加时，$\hat{\beta}_j(k)$ 显著地下降，而且迅速趋于零，因而失去预测能力。从岭回归的观点看，x_j 对 y 不起重要作用，甚至可以剔除这个变量。

（2）图 7－2（b）的情况与图 7－2（a）相反，$\hat{\beta}_j=\hat{\beta}_j(0)>0$，但很接近 0。从古典回归分析的观点看，$x_j$ 对 y 的作用不大。但随着 k 略增加，$\hat{\beta}_j(k)$ 骤然变为负值，从岭回归的观点看，x_j 对 y 有显著影响。

7.3 岭迹分析



(3) 在图 7-2 (c) 中，$\hat{\beta}_j = \hat{\beta}_j(0) > 0$，说明 x_j 比较显著，但当 k 增加时，$\hat{\beta}_j(k)$ 迅速下降，且稳定为负值。从古典回归分析的观点看，x_j 是对 y 有正影响的显著因素。从岭回归的观点看，x_j 是对 y 有负影响的因素。

(4) 在图 7-2 (d) 中，$\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 都很不稳定，但其和却大体上稳定。这种情况往往发生在自变量 x_1 和 x_2 的相关性很强的场合，即在 x_1 和 x_2 之间存在多重共线性。因此，从变量选择的观点看，两者只要保留一个就够了。这可用来解释某些回归系数估计的符号不合理的情形，从实际观点看，β_1 和 β_2 不应有相反的符号。岭回归分析的结果对这一点提供了一种解释。

7.3 岭迹分析

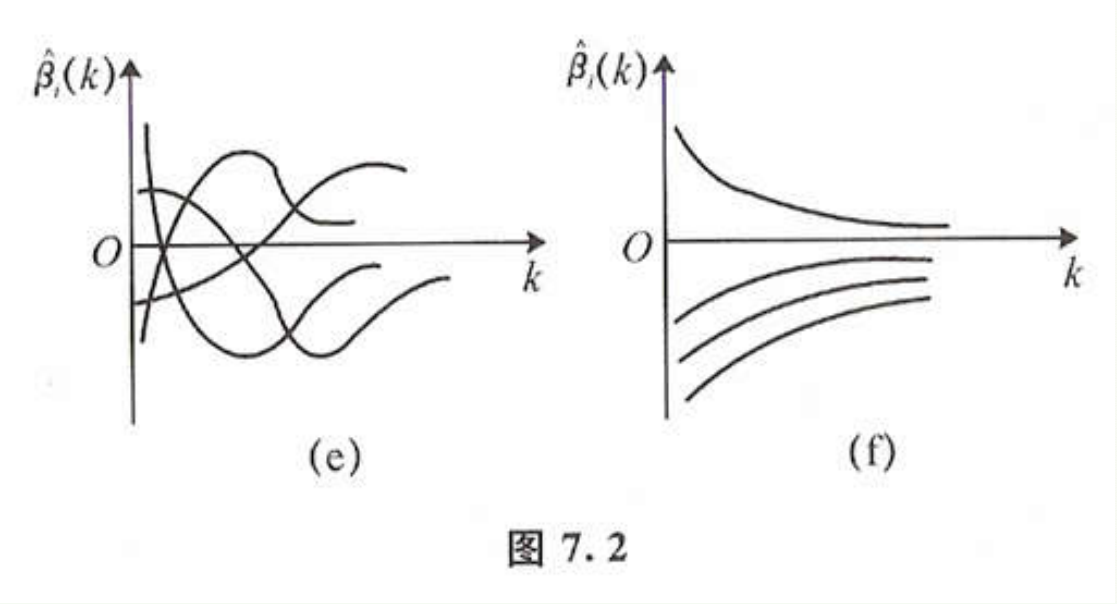


图 7.2

（5）从全局看，岭迹分析可用来估计在某一具体实例中最小二乘估计是否适用，把所有回归系数的岭迹都描在一张 图上，如果这些岭迹线的“**不稳定度**”很大，整个系统呈现 比较“乱”的局面，往往就使人怀疑最小二乘估计是否很好地反映了真实情况，图7-2（e）反映了这种情况。如果情况如图7-2（f）那样，则我们对最小二乘估计可以有更大的信心。当情况介于（e）和（f）之间时，我们必须适当地选择 k 值。

7.4 岭参数 k 的选择

1. 岭迹法

岭迹法选择 k 值的一般原则是：

(1) 各回归系数的岭估计基本稳定；

(2) 用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理；

(3) 回归系数没有不合乎经济意义的绝对值；

(4) 残差平方和增大不太多。

7.4 岭参数 k 的选择

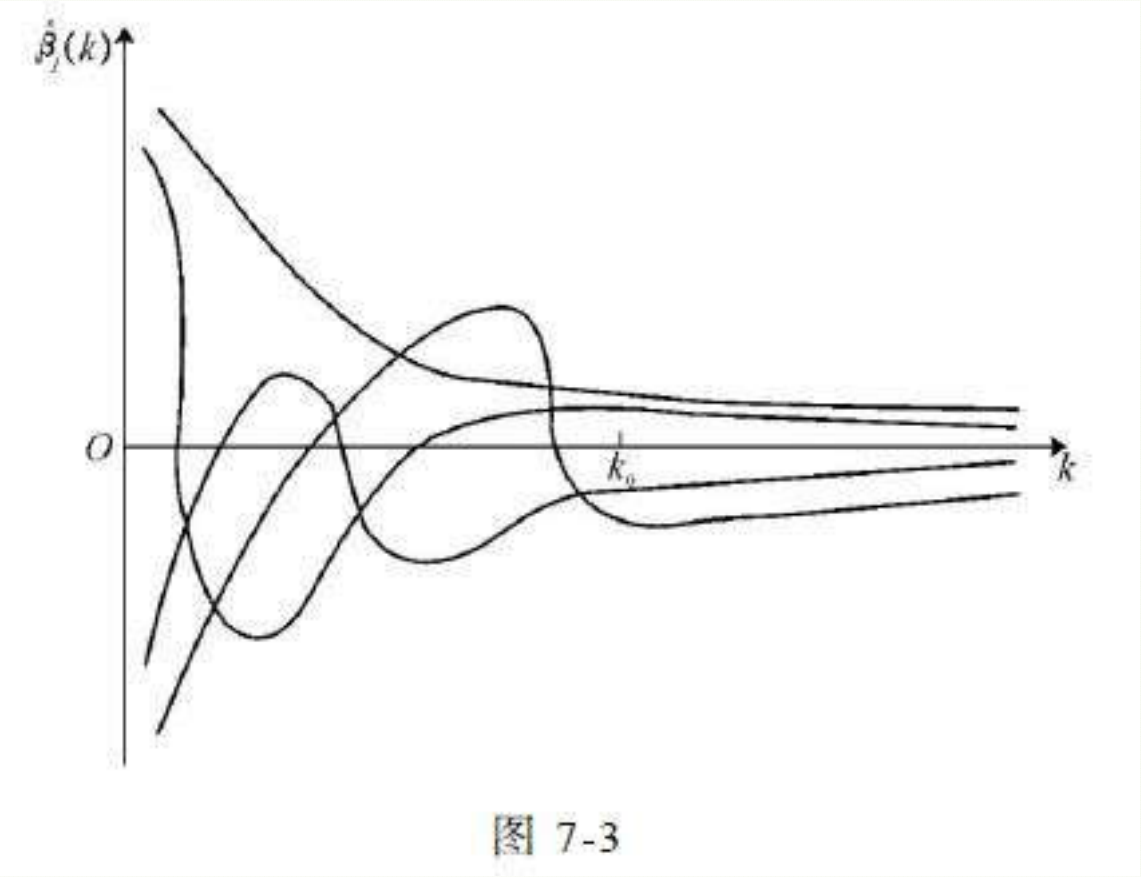


图 7-3

取k_0时，各回归系数的估计值基本上都能相对稳定

7.4 岭参数 k 的选择

例如在图7-3中，当 k 取 k_0 时，各回归系数的估计值基本上都能达到相对稳定。当然，上述种种要求并不总是能达到的。如在例7-1中由图7-1看到，取 k=0.5，岭迹已算平稳。此时 β_1 的估计值比较接近真实值，但是 β_2 还相去甚远。

岭迹法确定 k 值缺少严格的令人信服的理论依据，存在着一定的**主观人为性**，这似乎是岭迹法的一个**明显缺点**。从另一方面说，岭迹法确定 k 值的这种人为性正好是定性分析与定量分析有机结合的地方。

7.4 岭参数 k 的选择

7.4.2 方差扩大因子法

方差扩大因子 c_{jj} 度量了多重共线性的严重程度，计算岭估计 $\hat{\boldsymbol{\beta}}(k)$ 的协方差阵，得

$$
\begin{aligned}
D(\hat{\boldsymbol{\beta}}(k)) &= \operatorname{cov}(\hat{\boldsymbol{\beta}}(k), \hat{\boldsymbol{\beta}}(k)) \\
&= \operatorname{cov}((X'X+kI)^{-1}X'y, (X'X+kI)^{-1}X'y) \\
&= (X'X+kI)^{-1}X' \operatorname{cov}(y, y) X(X'X+kI)^{-1} \\
&= \sigma^2 (X'X+kI)^{-1}X'X(X'X+kI)^{-1} \\
&= \sigma^2 (c_{ij}(k))
\end{aligned}
$$

式中矩阵 $C_{ij}(k)$ 的对角元 $c_{jj}(k)$ 就是岭估计的方差扩大因子。不难看出，$c_{jj}(k)$ 随着 k 的增大而减少。

选择 k 使所有方差扩大因子 $c_{jj}(k) \leqslant 10$。

7.4 岭参数 k 的选择

$c(k)$的对角元素$c_{jj}(k)$为岭估计的方差扩大因子。不难看出，$c_{jj}(k)$随着 k 的增大而减少。用方差扩大因子选择 k 的经验做法是：选择 k 使所有方差扩因子 $c_{jj}(k) \leq 10$ 。当 $c_{jj}(k) \leq 10$时，所对应的 k 值的岭估计$\boldsymbol{\beta}(k)$就会相对稳定。

7.4 岭参数 k 的选择

7.4.3 由**残差平方和**来确定 k 值

岭估计在减小均方误差的同时**增大了残差平方和**，我们希望岭回归的残差平方和 SSE（k）的增加幅度控制在一定的限度以内，可以给定一个大于1的 c 值，要求：

$$\mathrm{SSE}(k) < c\mathrm{SSE} \qquad (7.3)$$

寻找使（7.3）式成立的最大的 k 值。

7.5 用岭回归选择变量

岭回归选择变量的**原则**：

(1) 在岭回归中设计矩阵 $\boldsymbol{X}$ 已经中心化和标准化了，这样可以直接比较标准化岭回归系数的大小。可以剔除掉标准化岭回归系数比较**稳定**且绝对值**很小**的自变量。

(2) 随着 k 的增加，回归系数不稳定，**振动趋于零**的自变量也可以剔除。

(3) 剔除标准化岭回归系数很**不稳定的自变量**。如果依照上述去掉变量的原则，有若干个回归系数不稳定，究竟去掉几个，去掉哪几个，这并无一般原则可循，这需根据去掉某个变量后重新进行岭回归分析的效果来确定。

7.5 用岭回归选择变量

例7.2 空气污染问题。Mcdonald和Schwing在参考文献［18］中曾研究死亡率与空气污染、气候以及社会经济状况等因素的关系。考虑了15个解释变量，收集了**60组样本数据**。

x1—Average annual precipitation in inches 平均年降雨量
x2—Average January temperature in degrees F 1月份平均气温
x3—Same for July 7月份平均气温
x4—Percent of 1960 SMSA population aged 65 or older
年龄65岁以上的人口占总人口的百分比
x5—Average household size 每家人口数
x6—Median school years completed by those over 22
年龄在22岁以上的人受教育年限的中位数

7.5 用岭回归选择变量

x7—Percent of housing units which are sound & with all facilities
住房符合标准的家庭比例数
x8—Population per sq. mile in urbanized areas, 1960 每平方公里人口数
x9—Percent non-white population in urbanized areas,
1960 非白种人占总人口的比例
x10—Percent employed in white collar occupations 白领阶层人口比例
x11—Percent of families with income < $3000
收入在3000美元以下的家庭比例
x12—Relative hydrocarbon pollution potential 碳氢化合物的相对污染势
x13— Same for nitric oxides 氮氧化合物的相对污染势
x14—Same for sulphur dioxide 二氧化硫的相对污染势
x15—Annual average % relative humidity at 1pm 年平均相对湿度
y—Total age-adjusted mortality rate per 100,000
每十万人中的死亡人数

7.5 用岭回归选择变量

计算**X'X**的15个特征为：

4.5272，2.7547，2.0545，1.3487，1.2227

0.9605，0.6124，0.4729，0.3708，0.2163

0.1665，0.1275，0.1142，0.0460，0.0049

注：以上特征根是按照原文献的计算方式，自变量观测阵未包含代表常数项的第一列1

条件数

$$K_{0.5} = \sqrt{\lambda_1 / \lambda_{15}} = \sqrt{4.5275 / 0.0049} = \sqrt{923.918} = 30.396$$

7.5 用岭回归选择变量

进行岭迹分析

把15个回归系数的岭迹画到图7-4中，我们可看到，当k=0.20时岭迹大体上**达到稳定**。按照岭迹法，应取 k=0.2。

若用方差扩大因子法，当 k 在0.02～0.08时，方差扩大因子小于10，故应建议在此范围选取 k 。

由此也看到不同的方法选取的 k 值是不同的。

7.5 用岭回归选择变量

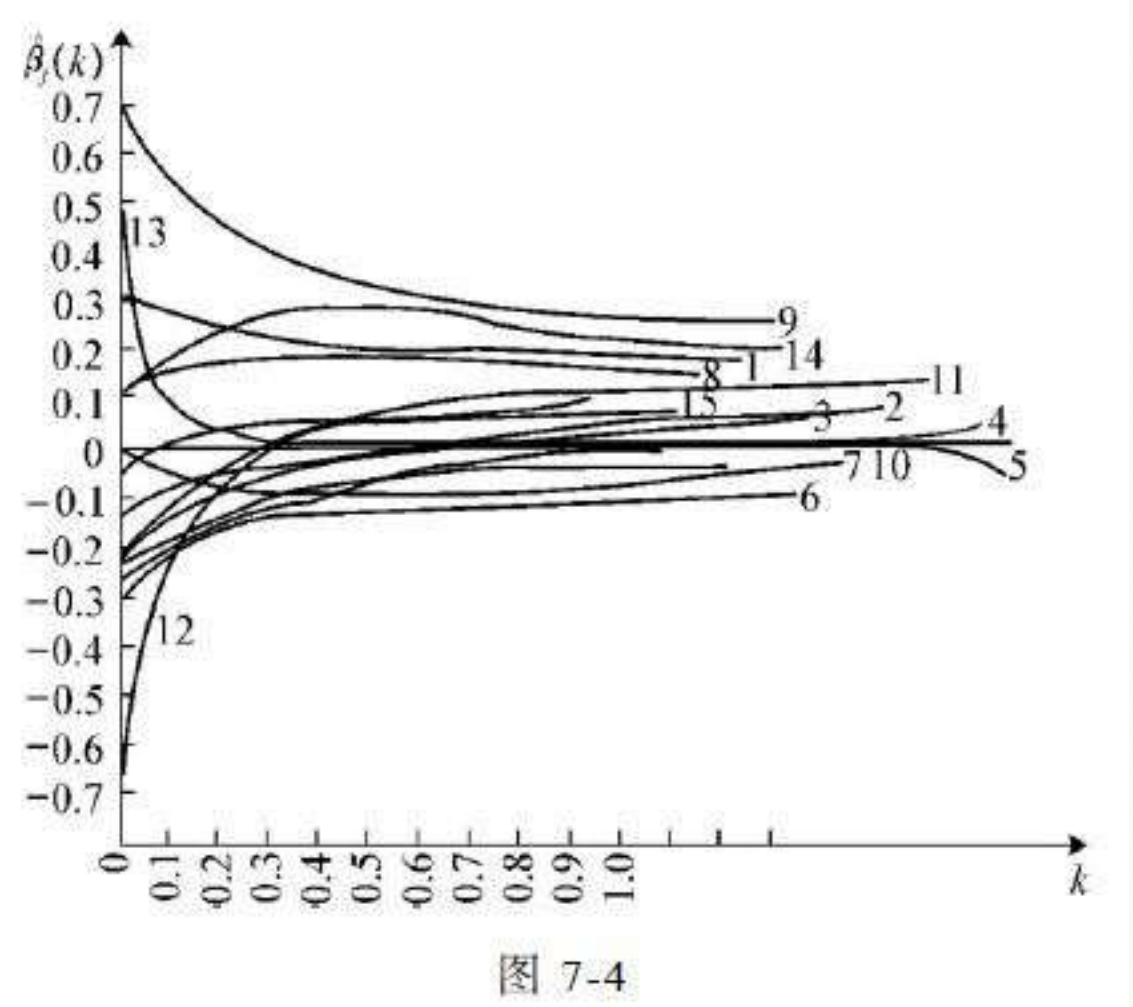


图 7-4

7.5 用岭回归选择变量

在用岭回归进行变量选择时，因为从岭迹看到自变量x_4, x_7, x_{10}, x_{11}和x_{15}有较稳定且绝对值比较小的岭回归系数，根据变量选择的第一条原则，这些自变量可以去掉。

又因为自变量x_{12}和x_{13}的岭回归系数很不稳定，且随着k的增加很快趋于零，根据上面的第二条原则这些自变量也应该去掉。

再根据第三条原则去掉变量x_3和x_5。

这个问题最后剩的变量是x_1, x_2, x_6, x_8, x_9和x_{14}。

7.5 用岭回归选择变量

例7.3 Gorman-Torman例子(见参考文献［2］)。

本例共有10个自变量，**X**已经中心化和标准化了，

X′X的特征根为：

3.692，1.542，1.293，1.046，0.972，

0.659，0.357，0.220，0.152，0.068

最后一个特征根λ_{10}=0.068，较接近于零。

7.5 用岭回归选择变量

$$K_{0.5}=\sqrt{\lambda_1/\lambda_{10}}=\sqrt{3.692/0.068}=\sqrt{54.294}=7.368$$

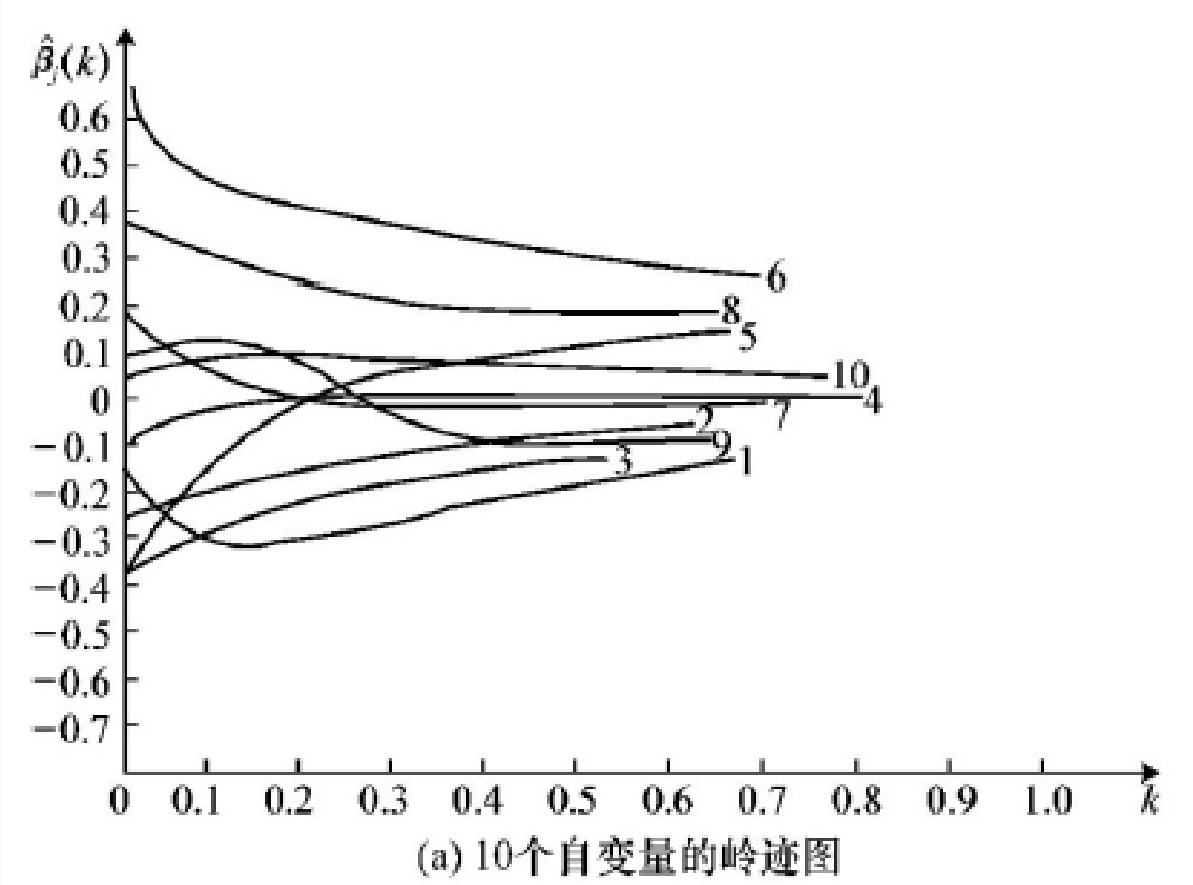
条件数$k\approx 7.4<10$。从条件数的角度看，似乎设计矩阵$\boldsymbol{X}$没有复共线性。但下面的研究表明，做岭回归还是必要的。

关于条件数，这里附带说明它的一个缺陷，就是当$\boldsymbol{X}'\boldsymbol{X}$所有特征根都比较小时，**虽然条件数不大，但多重共线性却存在**。本例就是一个证明。

7.5 用岭回归选择变量

下面做岭回归分析。对 15 个 k 值算出 $\hat{\boldsymbol{\beta}}(k)$，画出岭迹，如图 7-5 (a) 所示。由图 7-5 (a) 可看到，最小二乘估计的稳定性很差。这反映在当 k 与 0 略有偏离时，$\hat{\boldsymbol{\beta}}(k)$ 与 $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(0)$ 就有较大的差距，特别是 $|\hat{\beta}_5|$ 与 $|\hat{\beta}_6|$ 变化最明显。当 k 从 0 上升到 0.1 时，$\|\hat{\boldsymbol{\beta}}(k)\|^2$ 下降到 $\|\hat{\boldsymbol{\beta}}(0)\|^2$ 的 59%，而在正交设计的情形下只下降 17%。这些现象在直观上就使人怀疑最小二乘估计 $\hat{\boldsymbol{\beta}}$ 是否反映了 $\boldsymbol{\beta}$ 的真实情况。

7.5 用岭回归选择变量



(a) 10个自变量的岭迹图

7.5 用岭回归选择变量

另外，因素 x_5 的回归系数的最小二乘估计 $\hat{\beta}_5$ 为负回归系数中绝对值最大的，但当 k 增加时，$\hat{\beta}_5(k)$ 迅速上升且变为正的，与此相反，对因素 x_6，$\hat{\beta}_6$ 为正的，且绝对值最大，但当 k 增加时，$\hat{\beta}_6(k)$ 迅速下降。再考虑到 x_5，x_6 样本相关系数达到 0.84，因此这两个因素可近似地合并为一个因素。

7.5 用岭回归选择变量

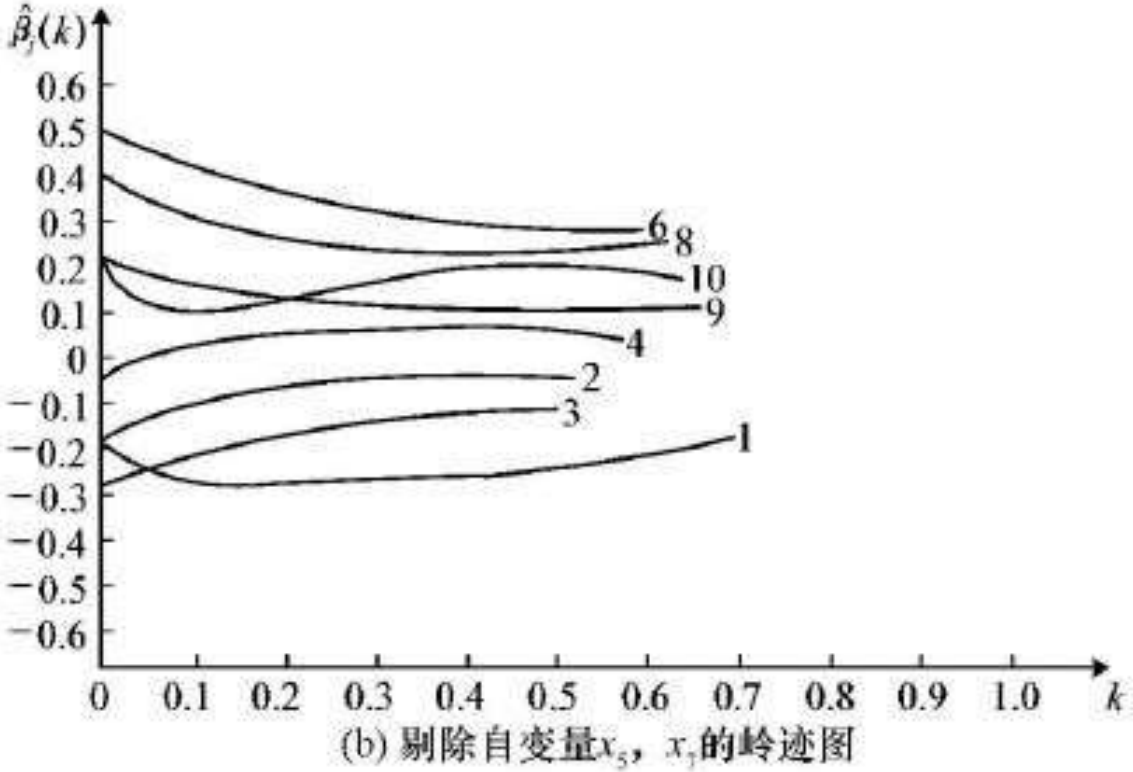
再看 x_7，它的回归系数估计$\hat{\beta}_7$ 绝对值偏高，当 k 增加时，$\hat{\beta}_7(k)$ 很快接近于 0，这意味着 x_7 实际上对 y 无多大影响。至于 x_1，其回归系数的最小二乘估计绝对值看来有点偏低，当 k 增加时，$|\hat{\beta}_1(k)|$ 首先迅速上升，成为对因变量有负影响的最重要的自变量。当 k 较大时，$|\hat{\beta}_1(k)|$ 稳定地缓慢趋于零。这意味着，通常的最小二乘估计对 x_1 的重要性估计过低了。

7.5 用岭回归选择变量

从整体上看，当 k 达到 0.2～0.3 的范围时，各个$\hat{\beta}_j(k)$已大体上趋于稳定，因此，在这区间上取一个 k 值作岭回归可能得到较好的效果。

本例中$\hat{\beta}_5(k)$和$\hat{\beta}_7(k)$当 k 从 0 略增加时，很快趋于 0，于是它们很自然是应该剔除的。去掉它们之后，重作岭回归分析，岭迹基本稳定。因此去掉 x_5 和 x_7 是合理的。

7.5 用岭回归选择变量



(b) 剔除自变量x_5，x_7的岭迹图

图 7-5

7.5 用岭回归选择变量

例7-4 用岭回归方法处理民航客运数据的**多重共线性问题**。

用R软件对例3-3做岭回归分析，其中岭参数 k 及其相应的回归系数的计算结果见表7-3，输出的岭迹图见图7-6(a)，相应的计算代码如下：

7.5 用岭回归选择变量

```
data3.3<-read.csv("D:/data3.3.csv",head=TRUE)
datas<-data.frame(scale(data3.3[,2:7]))
#对样本数据进行标准化处理并转换为数据框的格式存储
library(MASS)                    #加载包 MASS
ridge3.3<-lm.ridge(y~.-1,data=datas,lambda=seq(0,3,0.1))
 #做岭回归，对于标准化后的数据模型不包含截距项，其中 lambda 为岭参数 k 的所有取值
beta<-coef(ridge3.3)    #将所有不同岭参数所对应的回归系数的结果赋给 beta
beta                             #输出 beta
#绘制岭迹图
k<-ridge3.3$lambda       #将所有岭参数赋给 k
plot(k,k,type="n",xlab="岭参数 k",ylab="岭回归系数",ylim=c(-2.5,2.5))
#创建没有任何点和线的图形区域
linetype<-c(1:5)
char<-c(18:22)
for(i in 1:5)
   lines(k,beta[,i],type="o",lty=linetype[i],pch=char[i],cex=0.75)
#画岭迹线
legend(locator(1),inset=0.5,legend=c("x1","x2","x3","x4","x5"),cex=
       0.8,pch=char,lty=linetype)        #添加图例
```

7.5 用岭回归选择变量

表 7-3

k	x_1	x_2	x_3	x_4	x_5
0.0	2.447 39	−2.485 10	−0.083 14	0.530 54	0.563 54
0.1	0.164 17	−0.085 30	−0.110 45	0.587 51	0.387 11
0.2	0.169 52	0.029 65	−0.101 93	0.511 33	0.334 93
0.3	0.184 59	0.084 87	−0.096 48	0.465 10	0.305 35
0.4	0.196 83	0.118 37	−0.092 63	0.434 10	0.286 37
0.5	0.206 18	0.141 00	−0.089 69	0.411 76	0.273 19
0.6	0.213 36	0.157 31	−0.087 31	0.394 81	0.263 53
0.7	0.218 97	0.169 59	−0.085 29	0.381 47	0.256 17
0.8	0.223 42	0.179 16	−0.083 52	0.370 64	0.250 36
0.9	0.227 00	0.186 79	−0.081 94	0.361 63	0.245 68
1.0	0.229 92	0.193 00	−0.080 49	0.354 01	0.241 81
1.1	0.232 33	0.198 13	−0.079 15	0.347 44	0.238 56
1.2	0.234 32	0.202 42	−0.077 89	0.341 71	0.235 79
1.3	0.235 99	0.206 05	−0.076 69	0.336 65	0.233 40
1.4	0.237 38	0.209 15	−0.075 55	0.332 14	0.231 31
1.5	0.238 56	0.211 81	−0.074 46	0.328 09	0.229 47
1.6	0.239 54	0.214 12	−0.073 41	0.324 41	0.227 82
1.7	0.240 37	0.216 12	−0.072 39	0.321 05	0.226 35
1.8	0.241 07	0.217 86	−0.071 40	0.317 96	0.225 01
1.9	0.241 66	0.219 39	−0.070 44	0.315 11	0.223 79

7.5 用岭回归选择变量

续表

k	x_1	x_2	x_3	x_4	x_5
2.0	0.242 14	0.220 74	−0.069 50	0.312 47	0.222 68
2.1	0.242 54	0.221 92	−0.068 59	0.310 00	0.221 65
2.2	0.242 87	0.222 96	−0.067 70	0.307 69	0.220 69
2.3	0.243 13	0.223 88	−0.066 82	0.305 52	0.219 80
2.4	0.243 33	0.224 70	−0.065 97	0.303 48	0.218 97
2.5	0.243 49	0.225 41	−0.065 13	0.301 54	0.218 19
2.6	0.243 60	0.226 04	−0.064 30	0.299 70	0.217 45
2.7	0.243 67	0.226 60	−0.063 49	0.297 96	0.216 75
2.8	0.243 70	0.227 09	−0.062 69	0.296 29	0.216 09
2.9	0.243 70	0.227 52	−0.061 91	0.294 70	0.215 46
3.0	0.243 67	0.227 89	−0.061 14	0.293 17	0.214 86

7.5 用岭回归选择变量

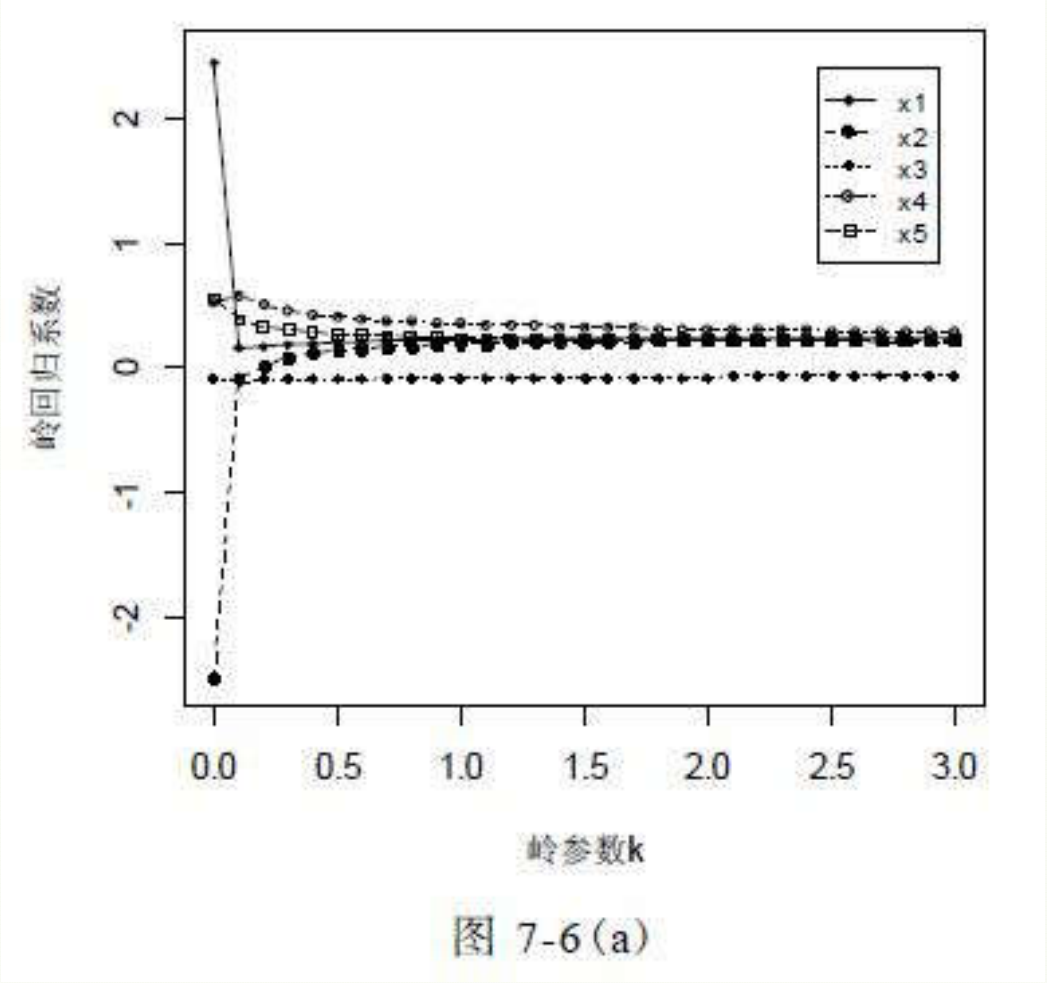


图 7-6(a)

7.5 用岭回归选择变量

从图7-6(a)中可以看到，变量x_2的岭回归系数**从负值迅速变为正值，**β_1和β_2的估计值绝对值都迅速减少，两者之和比较稳定，从岭回归的角度看，x_1和x_2只要保留一个就可以了，x_3, x_4, x_5 的岭回归系数相对稳定。

通过上面的分析，我们决定剔除 x_1，用 y 与其余四个自变量做岭回归。把岭参数的取值范围缩小为 0 到 2 ，步长取0.2，用下面的R程序进行计算：

7.5 用岭回归选择变量

```
ridge13.3<-lm.ridge(y~.-x1-1,data=datas,lambda=seq(0, 2,0.2))
#剔除 x1 后做岭回归
beta1<-coef(ridge13.3)
beta1
k1<-ridge13.3$lambda
#绘制岭迹图
plot(k1,k1,type="n",xlab="岭参数 k",ylab="岭回归系数",ylim=c(-1,1)))
linetype<-c(1:4)
char<-c(18:21)
for(i in 1:4)
   lines(k1,beta1[,i],type="o",lty=linetype[i],pch=char[i],cex=0.75)
legend(locator(1),inset=0.5,legend=c("x2","x3","x4","x5"),cex=
       0.8,pch=char,lty=linetype)
```

7.5 用岭回归选择变量

表 7-4

k	x_2	x_3	x_4	x_5
0.00	−0.23269	−0.13412	0.78770	0.51654
0.20	0.12890	−0.10944	0.56088	0.35889
0.40	0.21694	−0.10248	0.49958	0.32289
0.60	0.25571	−0.09844	0.46902	0.30778
0.80	0.27697	−0.09532	0.44983	0.29967
1.00	0.29002	−0.09262	0.43618	0.29461
1.20	0.29857	−0.09012	0.42571	0.29110
1.40	0.30441	−0.08776	0.41724	0.28847
1.60	0.30847	−0.08549	0.41014	0.28637
1.80	0.31132	−0.08329	0.40400	0.28461
2.00	0.31331	−0.08117	0.39859	0.28307

7.5 用岭回归选择变量

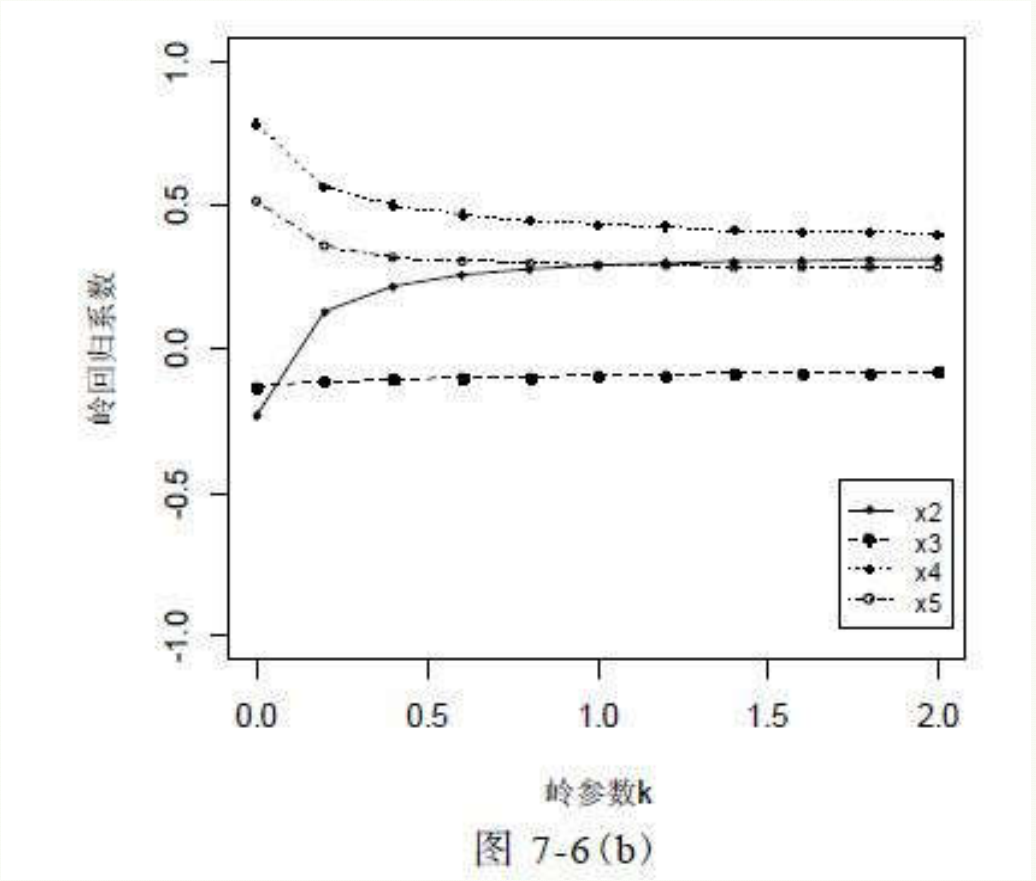


图 7-6(b)

7.5 用岭回归选择变量

由表7-4看到，剔除 x_1 后岭回归系数的变化幅度减小。从岭迹图7-6(b)看出，岭参数 k 大于1.4时，岭参数的**取值基本稳定**，不妨定 k=1.4，此时由表7-4得到样本数据标准化后的岭回归方程为：

$$\hat{y}^* = 0.304x_2^* - 0.0878x_3^* + 0.417x_4^* + 0.288x_5^*$$

此时对应未标准化的岭回归方程为：

$$\hat{y} = 417.394 + 0.069x_2 - 0.007x_3 + 16.970x_4 + 0.223x_5$$

与第6章剔除变量法相比，岭回归方法保留了自变量 x_2，如果希望回归方程中多保留一些自变量，那么岭回归方法是很有用的方法。

7.5 用岭回归选择变量

现在进一步计算出含有全部 5 个自变量的岭回归结果，与普通最小二乘的结果做一个比较。取岭参数 k=2.0，得岭回归方程为

$$\hat{y} = 301.520 + 0.035x_1 + 0.050x_2 - 0.006x_3 + 12.709x_4 + 0.172x_5$$

普通最小二乘回归方程为

$$\hat{y} = 450.91 + 0.354x_1 - 0.561x_2 - 0.007x_3 + 21.578x_4 + 0.435x_5$$

显然岭回归方程比普通最小二乘回归方程的实际意义更为容易解释。