中山大學
SUN YAT-SEN UNIVERSITY

2012级《多元统计分析与数据挖掘》第5周

2015.3.31

中山大学数学与计算科学学院 黄志洪

因子分析

- 降维的一种方法，是主成分分析的推广和发展
- 是用于分析隐藏在表面现象背后的因子作用的统计模型。试图用最少个数的不可测的公共因子的线性函数与特殊因子之和来描述原来观测的每一分量
- 例子：各科学习成绩（数学能力，语言能力，运动能力等）
- 例子：生活满意度（工作满意度，家庭满意度）
- 例子：薛毅书P522

因子分析的主要用途

- 减少分析变量个数
- 通过对变量间相关关系的探测，将原始变量分组，即将相关性高的变量分为一组，用共性因子来代替该变量
- 使问题背后的业务因素的意义更加清晰呈现

与主成分分析的区别

- 主成分分析侧重“变异量”，通过转换原始变量为新的组合变量使到数据的“变异量”最大，从而能把样本个体之间的差异最大化，但得出来的主成分往往从业务场景的角度难以解释
- 因子分析更重视相关变量的“共变异量”，组合的是相关性较强的原始变量，目的是找到在背后起作用的少量关键因子，因子分析的结果往往更容易用业务知识去加以解释

因子分析使用了复杂的数学手段

- 比主成分分析更加复杂的数学模型
- 求解模型的方法：主成分法，主因子法，极大似然法
- 结果还可以通过因子旋转，使到业务意义更加明显

数学模型：比PCA更复杂的矩阵求解问题

中山大學 SUN YAT-SEN UNIVERSITY

1．数学模型

设 $X=(X_1,X_2,\cdots,X_p)^T$ 是可观测的随机向量，且

$$E(X)=\mu=(\mu_1,\mu_2,\cdots,\mu_p)^T,\quad \mathrm{Var}(X)=\Sigma=(\sigma_{ij})_{p\times p}.$$

因子分析的一般模型为

$$\begin{cases} X_1-\mu_1=a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1\\ X_2-\mu_2=a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2\\ \quad\vdots\\ X_p-\mu_p=a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases}$$

数学模型

$$X = \mu + AF + \varepsilon, \tag{9.22}$$

$$E(F) = 0, \quad \mathrm{Var}(F) = I_m, \tag{9.23}$$

$$E(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2), \tag{9.24}$$

$$\mathrm{Cov}(F, \varepsilon) = 0. \tag{9.25}$$

因子模型性质

中山大學 SUN YAT-SEN UNIVERSITY

2. 因子模型的性质

(1) Σ 的分解

$$\Sigma = AA^T + D. \tag{9.26}$$

(2) 模型不受单位的影响. 若 $X^* = CX$, 则有

$$X^* = \mu^* + A^*F^* + \varepsilon^*,$$

其中 $\mu^* = C\mu$, $A^* = CA$, $F^* = F$, $\varepsilon^* = C\varepsilon$.

(3) 因子载荷不是惟一的. 设 T 是一 m 阶正交矩阵, 令 $A^* = AT$, $F^* = T^TF$, 则模型 (9.22) 可表示为

$$X = \mu + A^*F^* + \varepsilon. \tag{9.27}$$

统计意义

- 因子载荷的意义
- 共同度
- 特殊方差
- 总方差贡献

因子载荷矩阵和特殊方差矩阵的估计

- 主成分法
- 主因子法
- 极大似然法

主成分法

- 通过样本估算期望和协方差阵
- 求协方差阵的特征值和特征向量
- 省去特征值较小的部分，求出A、D
- 程序
- 例子

主因子法

- 首先对变量标准化
- 给出m和特殊方差的估计（初始）值
- 求出简约相关阵R*（p阶方阵）
- 计算R*的特征值和特征向量，取其前m个，略去其它部分
- 求出A*和D*，再迭代计算

极大似然法

- 似然函数
- 极大似然函数
- 算法描述（薛毅书p533）

Jöreskog 和 Lawley 等人 (1967) 提出了一种较为实用的迭代法，使极大似然法逐步被人们采用. 其基本思想是，先取一个初始矩阵

$$D_0 = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2),$$

现计算 A_0, 计算 A_0 的办法是先求 $D_0^{-1/2}\hat{\Sigma}D_0^{-1/2}$ 的特征值 $\theta_1 \geq \theta_2 \geq \theta_p$, 及相应的特征向量 $l_1, l_2, \cdots, l_p$. 令 $\Theta = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_m)$, $L = (l_1, l_2, \cdots, l_m)$ 且令

$$A_0 = D_0^{1/2} L \left(\Theta - I_m\right)^{1/2}. \tag{9.43}$$

再由式 (9.41) 得到 D_1, 然后再按上述方法得到 A_1, 直到满足方程 (9.40) 为止.

方差最大的正交旋转

- 由于因子载荷矩阵不是唯一，有时因子的实际意义会变得难以解释。
- 因子载荷矩阵的正交旋转
- 因子载荷方差
- 载荷值趋于1或趋于0，公共因子具有简单化的结构
- varimax( ) 函数

因子分析函数factanal( )

函数 factanal() 采用极大似然法估计参数，其使用格式为

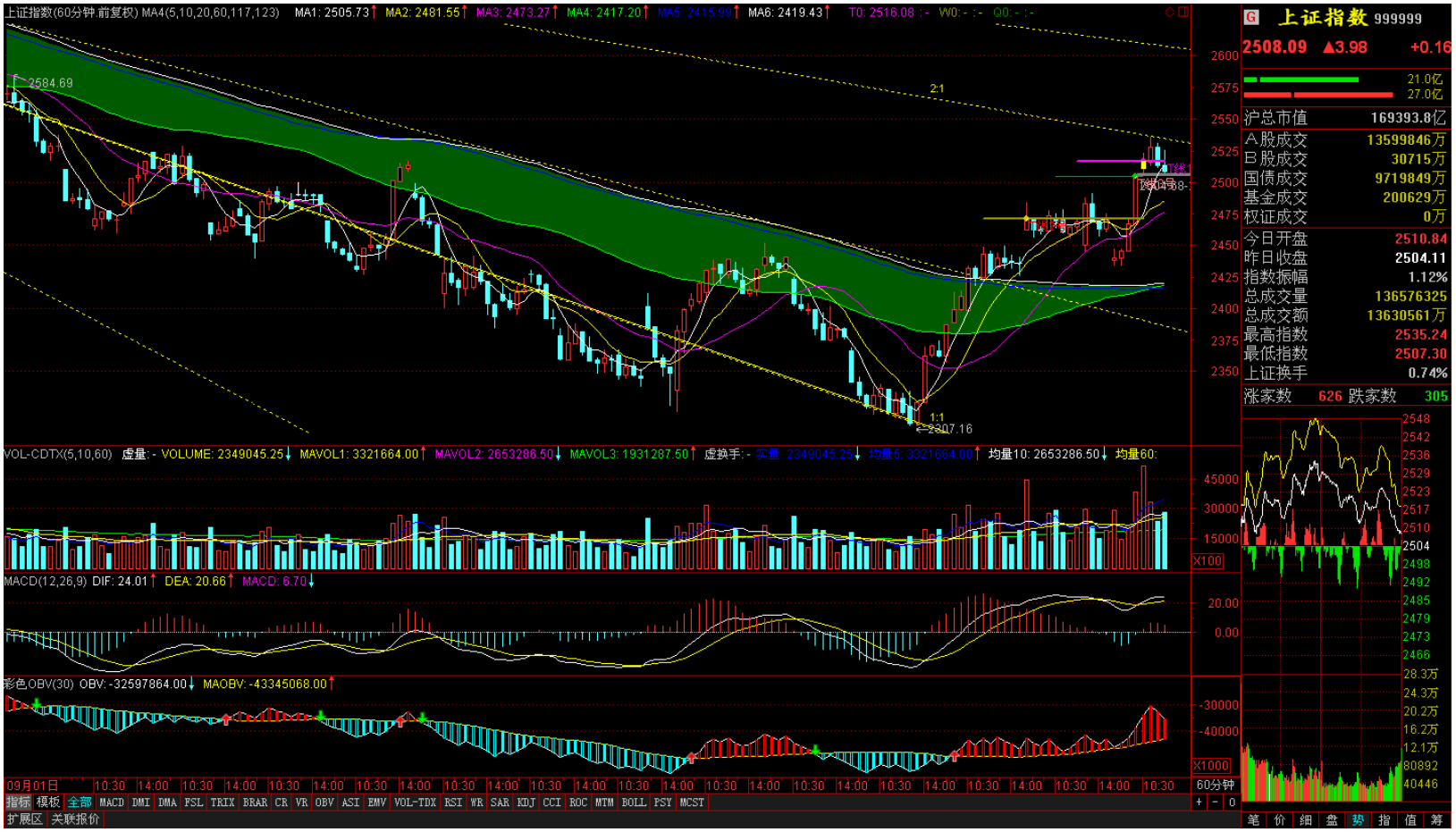
```
factanal(x, factors, data = NULL, covmat = NULL, n.obs = NA,
         subset, na.action, start = NULL,
         scores = c("none", "regression", "Bartlett"),
         rotation = "varimax", control = NULL, ...)
```

其中 x 是数据的公式，或者是由数据 (每个样本按行输入) 构成的矩阵，或者是数据框. factors 是因子的个数. data 是数据框，当 x 由公式形式给出时使用. covmat 是样本的协方差矩阵或样本的相关矩阵，此时不必输入变量 x. scores 表示因子得分的方法，scores="regression"，表示用回归方法计算因子得分，当参数为 scores="Bartlett"，表示用 Bartlett 方法计算因子得分 (具体意义见下小节)，缺省值为 "none""，即不计算因子得分. rotation 表示旋转，缺省值为方差最大旋转，当 rotation="none" 时，不作旋转变换.

因子得分

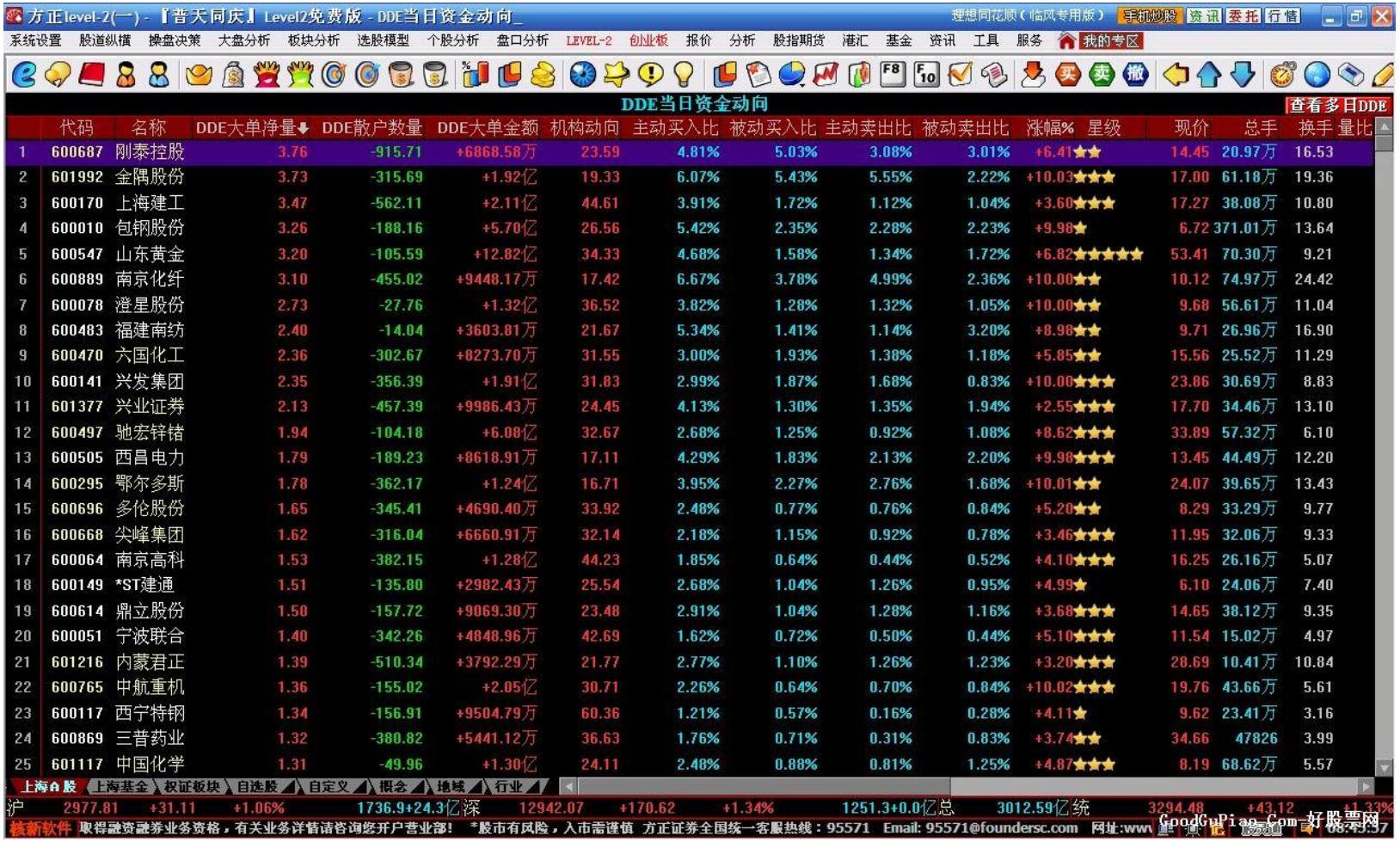
- 薛毅书p543

应用：股票市场指标设计



2015.3.31

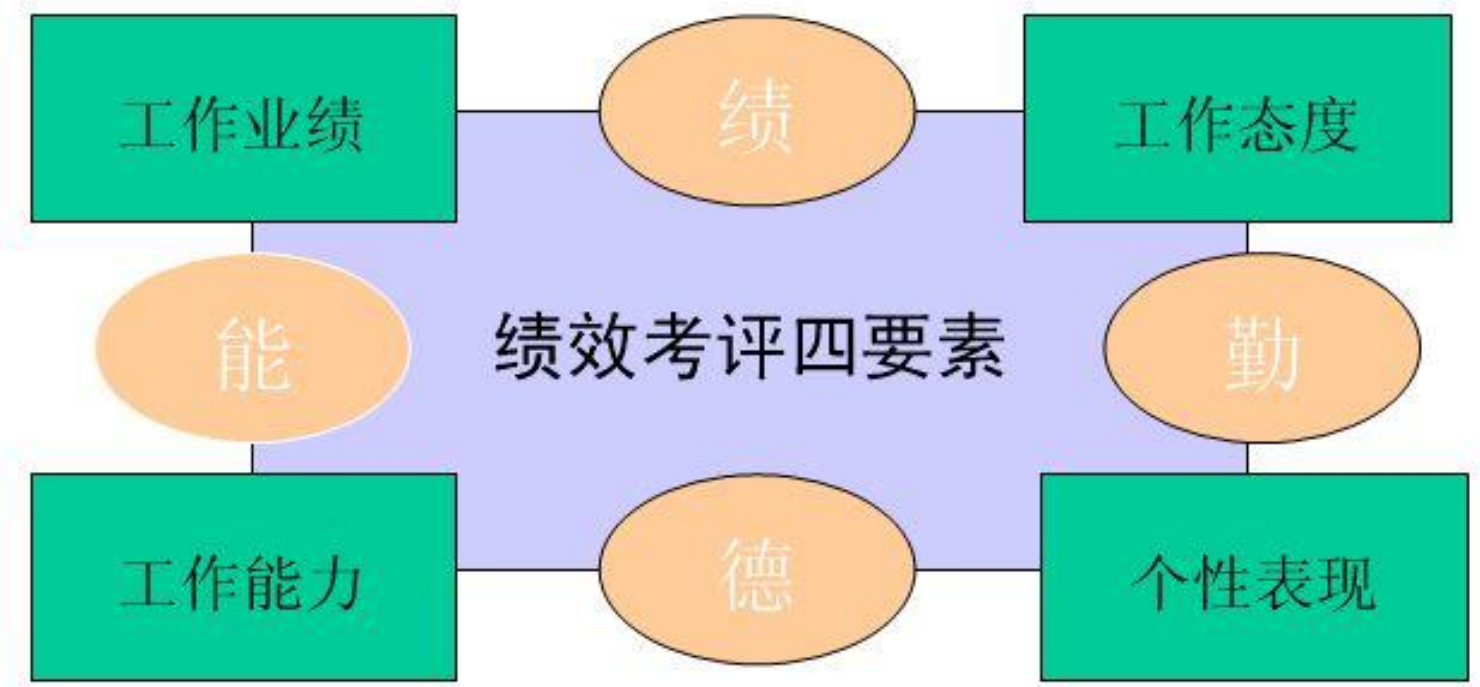
选取股票加入指数



	代码	名称	DDE大单净量	DDE散户数量	DDE大单金额	机构动向	主动买入比	被动买入比	主动卖出比	被动卖出比	涨幅%	星级	现价	总手	换手 量比
1	600687	刚泰控股	3.76	-915.71	+6868.58万	23.59	4.81%	5.03%	3.08%	3.01%	+6.41	★★	14.45	20.97万	16.53
2	601992	金隅股份	3.73	-315.69	+1.92亿	19.33	6.07%	5.43%	5.55%	2.22%	+10.03	★★★	17.00	61.18万	19.36
3	600170	上海建工	3.47	-562.11	+2.11亿	44.61	3.91%	1.72%	1.12%	1.04%	+3.60	★★★	17.27	38.08万	10.80
4	600010	包钢股份	3.26	-188.16	+5.70亿	26.56	5.42%	2.35%	2.28%	2.23%	+9.98	★	6.72	371.01万	13.64
5	600547	山东黄金	3.20	-105.59	+12.82亿	34.33	4.68%	1.58%	1.34%	1.72%	+6.82	★★★★★	53.41	70.30万	9.21
6	600889	南京化纤	3.10	-455.02	+9448.17万	17.42	6.67%	3.78%	4.99%	2.36%	+10.00	★★	10.12	74.97万	24.42
7	600078	澄星股份	2.73	-27.76	+1.32亿	36.52	3.82%	1.28%	1.32%	1.05%	+10.00	★★	9.68	56.61万	11.04
8	600483	福建南纺	2.40	-14.04	+3603.81万	21.67	5.34%	1.41%	1.14%	3.20%	+8.98	★★	9.71	26.96万	16.90
9	600470	六国化工	2.36	-302.67	+8273.70万	31.55	3.00%	1.93%	1.38%	1.18%	+5.85	★★	15.56	25.52万	11.29
10	600141	兴发集团	2.35	-356.39	+1.91亿	31.83	2.99%	1.87%	1.68%	0.83%	+10.00	★★★	23.86	30.69万	8.83
11	601377	兴业证券	2.13	-457.39	+9986.43万	24.45	4.13%	1.30%	1.35%	1.94%	+2.55	★★★	17.70	34.46万	13.10
12	600497	驰宏锌锗	1.94	-104.18	+6.08亿	32.67	2.68%	1.25%	0.92%	1.08%	+8.62	★★★	33.89	57.32万	6.10
13	600505	西昌电力	1.79	-189.23	+8618.91万	17.11	4.29%	1.83%	2.13%	2.20%	+9.98	★★★	13.45	44.49万	12.20
14	600295	鄂尔多斯	1.78	-362.17	+1.24亿	16.71	3.95%	2.27%	2.76%	1.68%	+10.01	★★	24.07	39.65万	13.43
15	600696	多伦股份	1.65	-345.41	+4690.40万	33.92	2.48%	0.77%	0.76%	0.84%	+5.20	★★	8.29	33.29万	9.77
16	600668	尖峰集团	1.62	-316.04	+6660.91万	32.14	2.18%	1.15%	0.92%	0.78%	+3.46	★★★	11.95	32.06万	9.33
17	600064	南京高科	1.53	-382.15	+1.28亿	44.23	1.85%	0.64%	0.44%	0.52%	+4.10	★★★	16.25	26.16万	5.07
18	600149	*ST建通	1.51	-135.80	+2982.43万	25.54	2.68%	1.04%	1.26%	0.95%	+4.99	★	6.10	24.06万	7.40
19	600614	鼎立股份	1.50	-157.72	+9069.30万	23.48	2.91%	1.04%	1.28%	1.16%	+3.68	★★★	14.65	38.12万	9.35
20	600051	宁波联合	1.40	-342.26	+4848.96万	42.69	1.62%	0.72%	0.50%	0.44%	+5.10	★★★	11.54	15.02万	4.97
21	601216	内蒙君正	1.39	-510.34	+3792.29万	21.77	2.77%	1.10%	1.26%	1.23%	+3.20	★★★	28.69	10.41万	10.84
22	600765	中航重机	1.36	-155.02	+2.05亿	30.71	2.26%	0.64%	0.70%	0.84%	+10.02	★★★	19.76	43.66万	5.61
23	600117	西宁特钢	1.34	-156.91	+9504.79万	60.36	1.21%	0.57%	0.16%	0.28%	+4.11	★	9.62	23.41万	3.16
24	600869	三普药业	1.32	-380.82	+5441.12万	36.63	1.76%	0.71%	0.31%	0.83%	+3.74	★★	34.66	47826	3.99
25	601117	中国化学	1.31	-49.96	+1.30亿	24.11	2.48%	0.88%	0.81%	1.25%	+4.87	★★★	8.19	68.62万	5.57

2015.3.31

应用：员工绩效考核指标设计



考核指标设计

评价因素	对评价期间工作成绩的评价要点	评价尺度				
		优	良	中	可	差
勤务态度	A. 严格遵守工作制度，有效利用工作时间。	14	12	10	8	6
	B. 对新工作持积极态度。	14	12	10	8	6
	C. 忠于职守、坚守岗位	14	12	10	8	6
	D. 以协作精神工作，协助上级，配合同事。	14	12	10	8	6
受命准备	A. 正确理解工作内容，制定适当的工作计划。	14	12	10	8	6
	B. 不需要上级详细的指示和指导。	14	12	10	8	6
	C. 及时与同事及协作者取得联系，使工作顺利进行。	14	12	10	8	6
	D. 迅速、适当地处理工作中的失败及临时追加任务。	14	12	10	8	6
业务活动	A. 以主人公精神与同事同心协力努力工作。	14	12	10	8	6
	B. 正确认识工作目的，正确处理业务。	14	12	10	8	6
	C. 积极努力改善工作方法。	14	12	10	8	6
	D. 不打乱工作秩序，不妨碍他人工作。	14	12	10	8	6
工作效率	A. 工作速度快，不误工期。	14	12	10	8	6
	B. 业务处置得当，经常保持良好成绩。	14	12	10	8	6
	C. 工作方法合理，时间和经费的使用十分有效。	14	12	10	8	6
	D. 工作中没有半途而废，不了了之和造成后遗症的现象。	14	12	10	8	6
成果	A. 工作成果达到预期目的或计划要求。	14	12	10	8	6
	B. 及时整理工作成果，为以后的工作创造条件。	14	12	10	8	6
	C. 工作总结和汇报准确真实。	14	12	10	8	6
	D. 工作中熟练程度和技能提高较快。	14	12	10	8	6

多元线性回归的最小二乘解（无偏估计）

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{2n} & \cdots & x_{np} \end{bmatrix},\ \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

则多元线性模型 (6.19) 可表示为

$$Y=X\beta+\varepsilon, \tag{6.20}$$

类似于一元线性回归，求参数 β 的估计值 $\hat{\beta}$, 就是求最小二乘函数

$$Q(\beta)=(y-X\beta)^T(y-X\beta), \tag{6.21}$$

达到最小的 β 值.

可以证明 β 的最小二乘估计

$$\hat{\beta}=\left(X^TX\right)^{-1}X^Ty. \tag{6.22}$$

广义逆的奇异性

$$B = X^{+}Y = (X^{T}X)^{-1}X^{T}Y$$

X^{+}表示X的广义逆（或叫伪逆）。

- 当变量比样本多时，出现奇异性
- 当出现多重共线性时，出现奇异性

例子

- 假设已知x_1，x_2与y的关系服从线性回归型 $y=10+2x_1+3x_2+\varepsilon$

给定 x_1，x_2 的 10 个值，如下表 7.1 的第（2）、（3）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

例子

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：

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

而原模型的参数潼

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

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。

岭回归(Ridge Regression，RR)

- 1962年由Heer首先提出，1970年后他与肯纳德合作进一步发展了该方法
- 先对数据做标准化，为了记号方便，标准化后的学习集仍然用X表示
- 我们称

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

为$\boldsymbol{\beta}$的岭回归估计，其中k称为岭参数。

性质

- 当自变量间存在复共线性时，$|X'X| \approx 0$，我们设想给$X'X$加上一个正常数矩阵kI，（$k > 0$），那么$X'X+kI$接近奇异的程度就会比$X'X$接近奇异的程度小得多。
- 岭回归做为β的估计应比最小二乘估计稳定，当$k=0$时的岭回归估计就是普通的最小二乘估计。

等价模型：惩罚函数

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3.41}$$

剛剛的K
IF有多重貢獻性，BETA會變很大

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t, \tag{3.42}$$

原本的回歸是求抛物體的最小值

IF二維，是個圓
把圓上升OR把抛物體投影

估计族

- 当岭参数为0，得到最小二乘解
- 当岭参数趋向更大时，岭回归系数估计趋向于0

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。

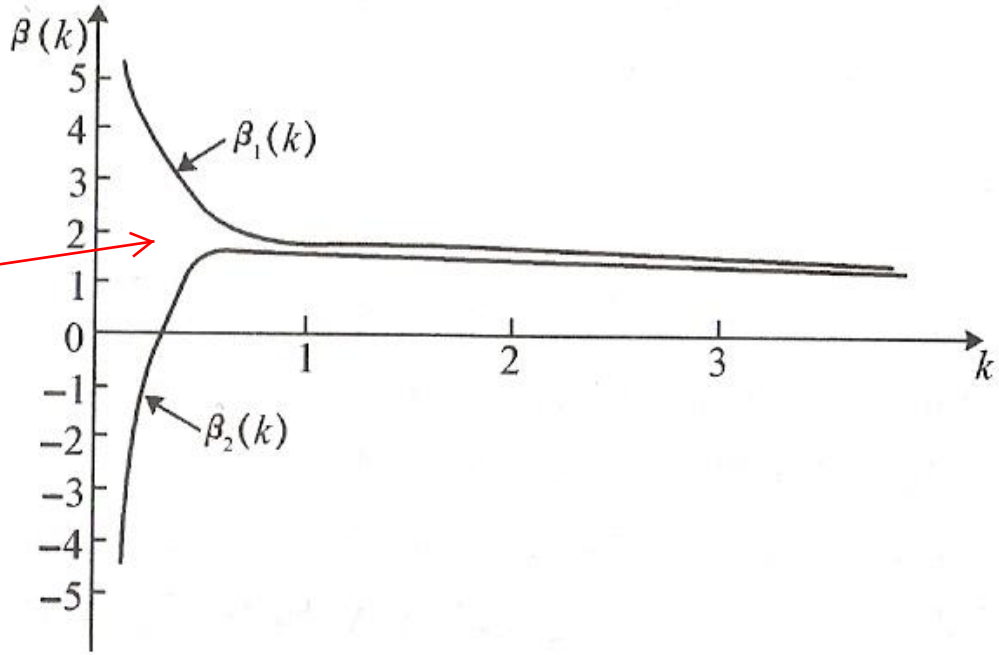
例如对例 7.1 可以算得不同 k 值时的 $\hat{\beta}_1(k)$， $\hat{\beta}_2(k)$，见表 7.2

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

岭迹图

- 当不存在奇异性时，岭迹应是稳定地逐渐趋向于0
- 通过岭迹图观察岭估计的情况，可以判断出应该剔除哪些变量



有喇叭状，多重贡献性

图 7.1

岭回归估计的性质

性质 1 $\hat{\boldsymbol{\beta}}(k)$ 是回归参数 β 的有偏估计。

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

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

要特别强调的是 $\hat{\boldsymbol{\beta}}(k)$ 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。

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。

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。

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

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

岭迹分析

acde有問題

k越大BETA越小,
嶺跡圖趨近於0

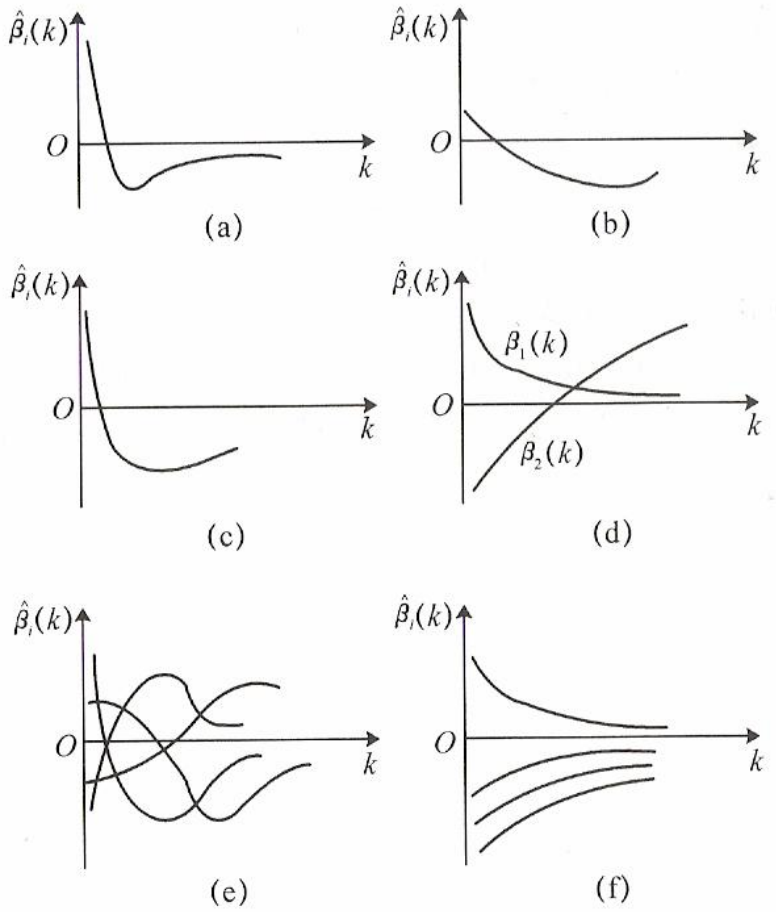


图 7.2

2015.3.31

岭参数的一般选择原则

- 选择k（或 lambda）值，使到

（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实际意义的绝对值；

（4）残差平方和增大不太多。

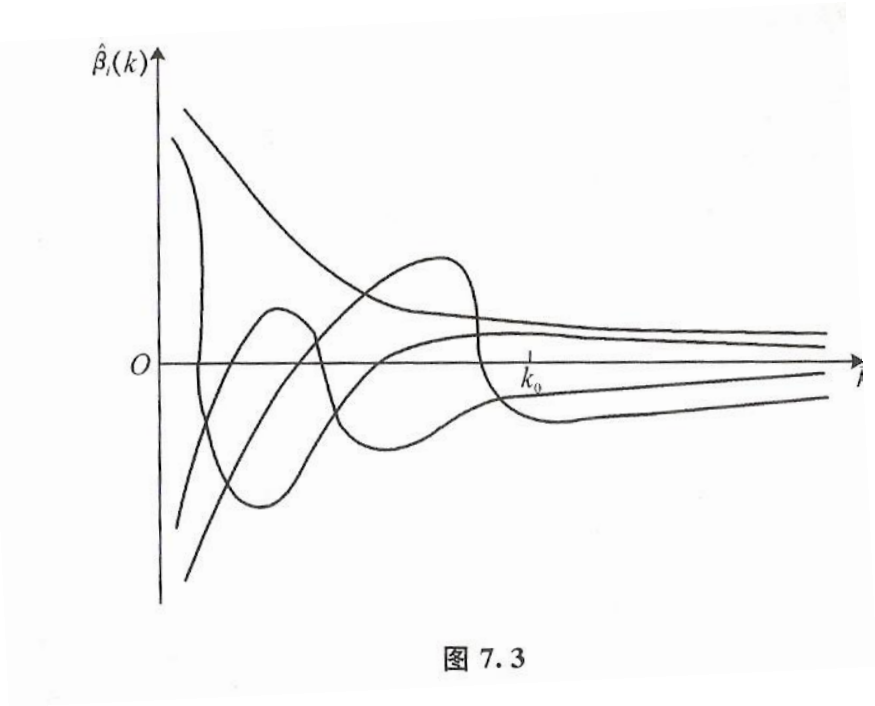


图 7.3

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

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

式中矩阵 $C_{jj}(k)$ 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 $c_{jj}(k) \leqslant 10$。

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，震动趋于零的自变量也可以剔除。

（3）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。

例子

空气污染问题。Mcdonald和Schwing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

x7—Percent of housing units which are sound & with all facilities

住房符合标准的家庭比例数

例子

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

岭迹图

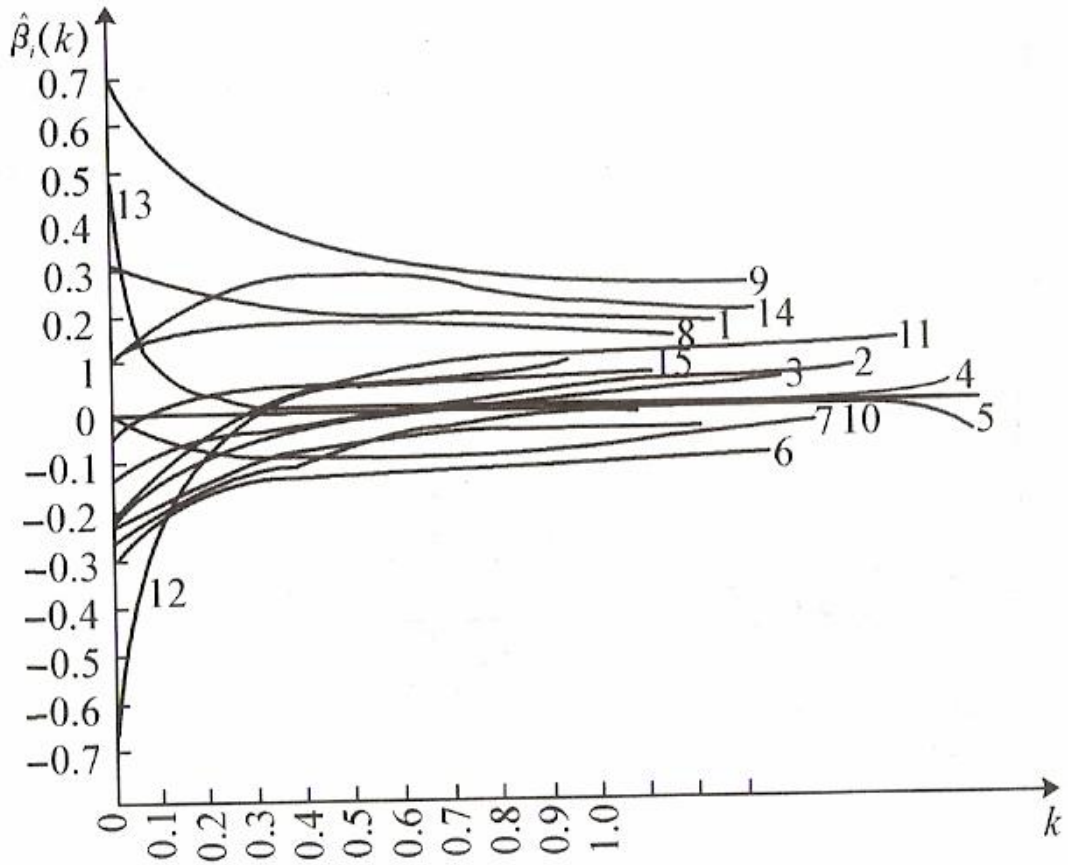


图 7.4

2015.3.31

岭迹分析

- 把15个回归系数的岭迹画到图中，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，因k=0.18时，方差扩大因子接近于1，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k值是不同的。

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x4,x7,x10,x11和x15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x12和x13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x3和x5。
- 这个问题最后剩的变量是x1，x2，x6，x8，x9，x14。

用R语言进行岭回归

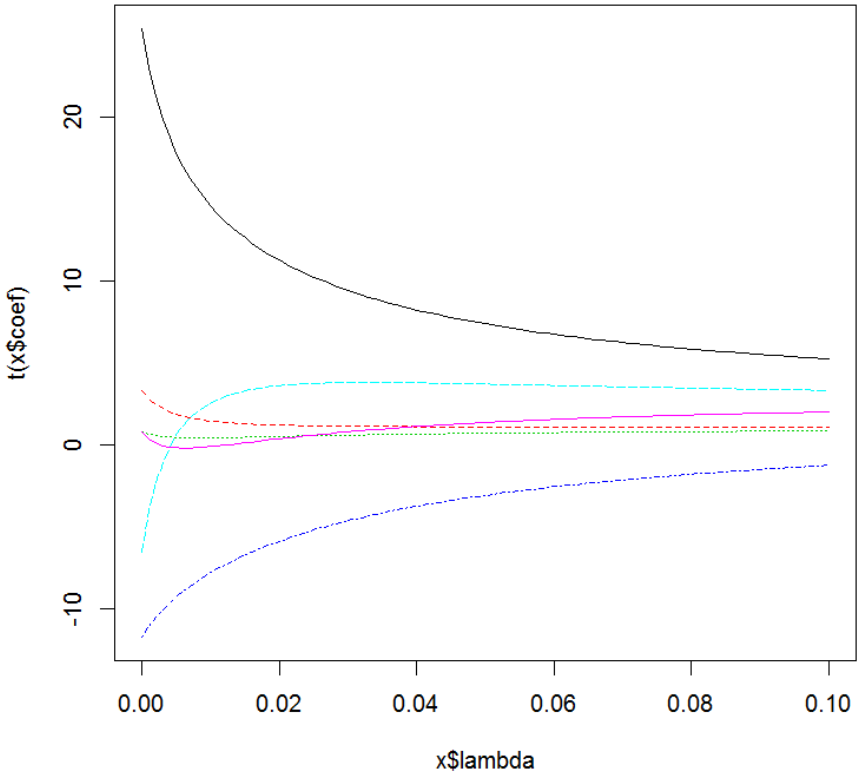
```
> library(MASS)
> longley
          y     GNP Unemployed Armed.Forces Population Year Employed
1947   83.0 234.289      235.6        159.0    107.608 1947   60.323
1948   88.5 259.426      232.5        145.6    108.632 1948   61.122
1949   88.2 258.054      368.2        161.6    109.773 1949   60.171
1950   89.5 284.599      335.1        165.0    110.929 1950   61.187
1951   96.2 328.975      209.9        309.9    112.075 1951   63.221
1952   98.1 346.999      193.2        359.4    113.270 1952   63.639
1953   99.0 365.385      187.0        354.7    115.094 1953   64.989
1954  100.0 363.112      357.8        335.0    116.219 1954   63.761
1955  101.2 397.469      290.4        304.8    117.388 1955   66.019
1956  104.6 419.180      282.2        285.7    118.734 1956   67.857
1957  108.4 442.769      293.6        279.8    120.445 1957   68.169
1958  110.8 444.546      468.1        263.7    121.950 1958   66.513
1959  112.6 482.704      381.3        255.2    123.366 1959   68.655
1960  114.2 502.601      393.1        251.4    125.368 1960   69.564
1961  115.7 518.173      480.6        257.2    127.852 1961   69.331
1962  116.9 554.894      400.7        282.7    130.081 1962   70.551
> 
```

多元线性回归的最小二乘估计

```
> summary(fm1 <- lm(Employed ~ ., data = longley))

Call:
lm(formula = Employed ~ ., data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41011 -0.15767 -0.02816  0.10155  0.45539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.482e+03  8.904e+02  -3.911 0.003560 **
y            1.506e-02  8.492e-02   0.177 0.863141
GNP         -3.582e-02  3.349e-02  -1.070 0.312681
Unemployed  -2.020e-02  4.884e-03  -4.136 0.002535 **
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population  -5.110e-02  2.261e-01  -0.226 0.826212
Year         1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

岭回归

```
names(longley)[1] <- "y  "
 lm.ridge(y ~ ., longley)
plot(lm.ridge(y ~ ., longley,
      lambda = seq(0,0.1,0.001)))
```




```
>
> names(longley)[1] <- "y"
> lm.ridge(y ~ ., longley)
                      GNP     Unemployed  Armed.Forces     Population           Year       Employed
2946.85636017    0.26352725    0.03648291    0.01116105   -1.73702984   -1.41879853    0.23128785
> plot(lm.ridge(y ~ ., longley,
+              lambda = seq(0,0.1,0.001)))
> |
```

岭回归

```
> lm.ridge(y ~ ., longley,lambda = seq(0,0.1,0.001))
                       GNP Unemployed Armed.Forces Population        Year     Employed
0.000  2946.85636 0.26352725 0.03648291  0.011161050 -1.7370298 -1.41879853  0.231287851
0.001  1895.97527 0.23923480 0.03100610  0.009372158 -1.6438029 -0.87657471  0.105607249
0.002  1166.33337 0.22099519 0.02719073  0.008243201 -1.5650260 -0.50108472  0.030290543
0.003   635.78843 0.20661106 0.02440554  0.007514565 -1.4962459 -0.22885815 -0.014755698
0.004   236.65772 0.19485388 0.02230066  0.007043302 -1.4348862 -0.02473192 -0.040566288
0.005   -71.53274 0.18498058 0.02066688  0.006744636 -1.3793225  0.13231532 -0.053663187
0.006  -314.43247 0.17651367 0.01937157  0.006565392 -1.3284596  0.25560068 -0.058119371
0.007  -509.05648 0.16913115 0.01832674  0.006470736 -1.2815187  0.35395451 -0.056588923
0.008  -667.11647 0.16260718 0.01747181  0.006437042 -1.2379217  0.43345188 -0.050860281
0.009  -796.92303 0.15677808 0.01676376  0.006447832 -1.1972245  0.49840118 -0.042171311
0.010  -904.52578 0.15152189 0.01617130  0.006491346 -1.1590755  0.55193667 -0.031397510
0.011  -994.42507 0.14674556 0.01567111  0.006559030 -1.1231903  0.59638825 -0.019168982
0.012 -1070.03184 0.14237663 0.01524553  0.006644564 -1.0893337  0.63352047 -0.005945632
0.013 -1133.97358 0.13835766 0.01488094  0.006743215 -1.0573084  0.66469138  0.007933122
0.014 -1188.30330 0.13464236 0.01456670  0.006851400 -1.0269464  0.69096113  0.022214639
0.015 -1234.64543 0.13119296 0.01429437  0.006966383 -0.9981032  0.71316772  0.036709726
0.016 -1274.29970 0.12797821 0.01405722  0.007086059 -0.9706528  0.73198092  0.051276169
0.017 -1308.31654 0.12497200 0.01384979  0.007208799 -0.9444851  0.74794140  0.065806938
0.018 -1337.55256 0.12215228 0.01366762  0.007333338 -0.9195024  0.76148954  0.080221587
0.019 -1362.71206 0.11950026 0.01350706  0.007458691 -0.8956181  0.77298695  0.094459929
0.020 -1384.37837 0.11699982 0.01336506  0.007584089 -0.8727546  0.78273271  0.108477319
0.021 -1403.03795 0.11463698 0.01323909  0.007708933 -0.8508422  0.79097588  0.122241095
0.022 -1419.09894 0.11239961 0.01312702  0.007832759 -0.8298181  0.79792511  0.135727883
0.023 -1432.90579 0.11027705 0.01302706  0.007955206 -0.8096251  0.80375619  0.148921524
0.024 -1444.75065 0.10825992 0.01293768  0.008075999 -0.7902114  0.80861799  0.161811479
0.025 -1454.88257 0.10633991 0.01285758  0.008194928 -0.7715296  0.81263718  0.174391594
0.026 -1463.51479 0.10450963 0.01278563  0.008311837 -0.7535365  0.81592202  0.186659124
0.027 -1470.83064 0.10276246 0.01272088  0.008426614 -0.7361922  0.81856540  0.198613980
```

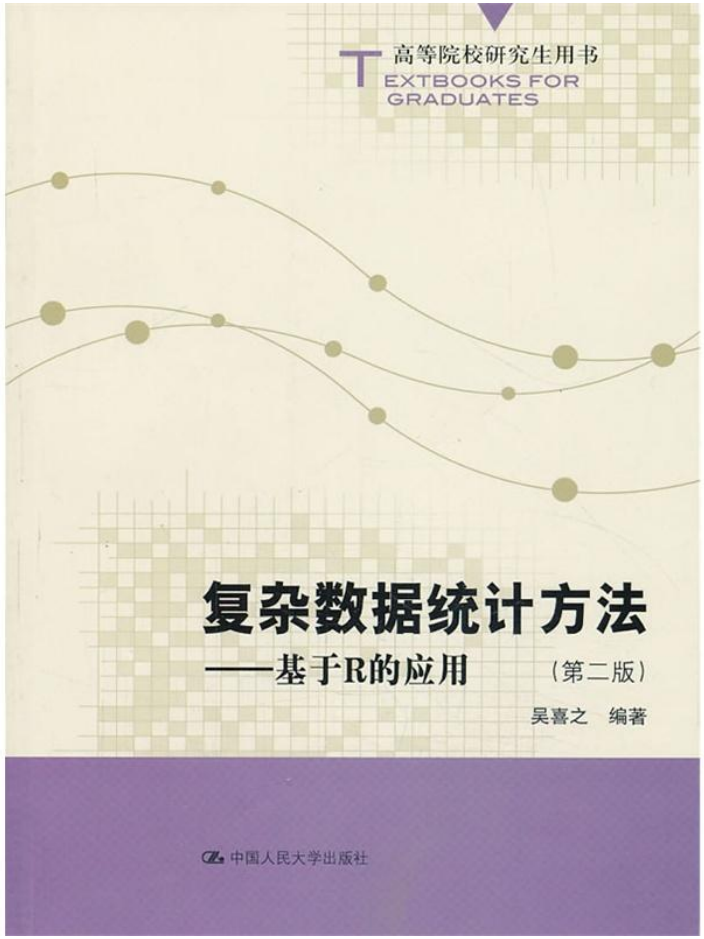
岭回归

```
> select(lm.ridge(y ~ ., longley,lambda = seq(0,0.1,0.001)))
modified HKB estimator is 0.006836982
modified L-W estimator is 0.05267247
smallest value of GCV  at 0.006
```

R的ridge包

```
>
> library(ridge)
> a=linearRidge(GNP.deflator~.,data=longley)
> summary(a)

Call:
linearRidge(formula = GNP.deflator ~ ., data = longley)


Coefficients:
              Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept) -1.247e+03              NA                  NA               NA       NA
GNP          4.338e-02       1.670e+01           3.689e+00            4.526  6.0e-06 ***
Unemployed   1.184e-02       4.286e+00           2.507e+00            1.710   0.0873 .
Armed.Forces 1.381e-02       3.721e+00           1.905e+00            1.953   0.0508 .
Population  -2.831e-02      -7.627e-01           5.285e+00            0.144   0.8853
Year         6.566e-01       1.211e+01           2.691e+00            4.500  6.8e-06 ***
Employed     6.745e-01       9.175e+00           4.996e+00            1.836   0.0663 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.01046912, chosen automatically, computed using 2 PCs

Degrees of freedom: model 3.67 , variance 3.218 , residual 4.123

> |
```

参考书


高等院校研究生用书
EXTBOOKS FOR GRADUATES
复杂数据统计方法
——基于R的应用
（第二版）
吴喜之 编著
中国人民大学出版社

2015.3.31

岭回归的问题

- 岭参数计算方法太多，差异太大
- 根据岭迹图进行变量筛选，随意性太大
- 岭回归返回的模型（如果没有经过变量筛选）包含所有的变量

Thanks

FAQ时间