第 7 章　岭回归

李　杰

数据科学学院，浙江财经大学

2019 年 12 月 26 日

内容提要

7.1 岭回归估计的定义

☞ 普通最小二乘估计的问题

- 当资料矩阵 $\mathbf{X}$ 呈病态时, $\mathbf{X}$ 的各列间有较强的近似线性关系. 而系数估计量因方差 $\mathrm{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}$ 很大变得极不稳定, 甚至出现估计值的符号与实际经济意义不符合的情形.
- Monte Carlo 实验

```
x1 <- c(1.1,1.4,1.7,1.7,1.8,1.8,1.9,2.0,2.3,2.4)
x2 <- c(1.1,1.5,1.8,1.7,1.9,1.8,1.8,2.1,2.4,2.5)
epsilon <- rnorm(10)
y <- 10+2*x1+3*x2+epsilon

test_lm <- lm(y~x1+x2)
summary(test_lm)

cor.test(x1,x2)
```

岭回归的定义

☞ 普通最小二乘估计的问题

- 当资料矩阵 $\mathbf{X}$ 呈病态时, $\mathbf{X}$ 的各列间有较强的近似线性关系. 而系数估计量因方差 $\mathrm{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}$ 很大变得极不稳定, 甚至出现估计值的符号与实际经济意义不符合的情形.

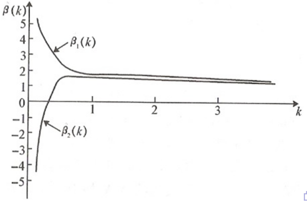
☞ 岭回归 (ridge regression, RR)

- 当**标准化**后的资料矩阵 $\mathbf{X}$ 中的变量存在严重多重共线性时, $|\mathbf{X}'\mathbf{X}| \approx 0$. 但 $|\mathbf{X}'\mathbf{X} + k\mathbf{I}|$ $(k>0)$ 接近奇异的程度比 $\mathbf{X}'\mathbf{X}$ 低很多. 称

$$\hat{\beta}(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}$$

为参数向量 β 的**岭回归估计**, k 为**岭参数**. $\mathbf{y}$ 可以标准化, 也可以没标准化.

- 岭参数 k 不唯一.



7.2 岭回归估计的性质

☞ **性质 1:** $\hat{\beta}(k)$ 是参数 β 的有偏估计.

【证】$\mathrm{E}[\hat{\beta}(k)] = \mathrm{E}[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{y}] = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}'\beta$

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

【证】$\hat{\beta}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\underline{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$

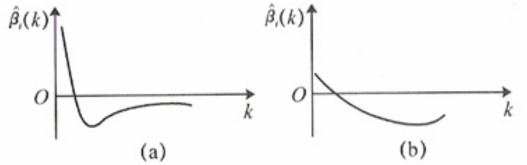
☞ **性质 3:** 对任意的 $k > 0$, $\|\hat{\beta}\| \neq 0$, 总有 $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$.

【注】该性质表明 $\hat{\beta}(k)$ 是对 $\hat{\beta}$ 向零的压缩. 当 $k \to \infty$ 时, $\hat{\beta}(k) \to \mathbf{0}$.

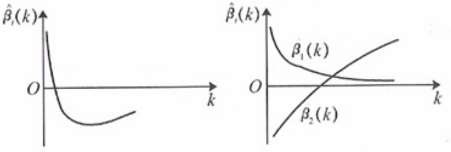
☞ **性质 4:** 总存在岭参数 k, 使得 $\mathrm{MSE}[\hat{\beta}(k)] < \mathrm{MSE}[\hat{\beta}]$

☞ 岭迹分析

- **岭迹:** 当参数 k 在 $(0,+\infty)$ 内变化时, $\hat{\beta}_j(k)$ $(j=0,2,\cdots,p)$ 是 k 的函数. 在平面坐标系上将 $\hat{\beta}_j(k)$ $(j=0,2,\cdots,p)$ 描绘出来, 画出的曲线叫**岭迹**.
- 岭迹分析的作用: 可用来确定岭参数 k 的值, 以及选择变量.
- **岭迹分析:**
  - ✧ **(a) 图:** $\hat{\beta}_j(0)=\hat{\beta}_j>0$, 且 $|\hat{\beta}_j(0)|$ 比较大. 从古典回归分析角度看来, x_j 对 y 有重要影响. 但 $\hat{\beta}_j(k)$ 的岭迹显示出很不稳定, 当 k 从零开始增大时, $\hat{\beta}_j(k)$ 显著地下降, 而且迅速趋于零, 因而失去预测能力. 故从岭回归的角度看来, x_j 对 y 影响不大, 应该剔除该变量.
  - ✧ **(b) 图:** 与 (a) 图相反, $\hat{\beta}_j(0)=\hat{\beta}_j>0$, 但 $|\hat{\beta}_j(0)|$ 接近零. 从古典回归分析角度看来, x_j 对 y 影响不大. 但 $\hat{\beta}_j(k)$ 的岭迹显示当 k 略增时, $\hat{\beta}_j(k)$ 快速为负值, 且 k 增大时 $\hat{\beta}_j(k)$ 没有趋于零. 故从岭回归的角度, x_j 对 y 有显著影响.

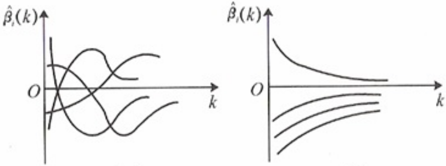


- **(c) 图:** $\hat{\beta}_j(0) = \hat{\beta}_j > 0$, 说明 x_j 对 y 有重要影响. 但 k 略增时, $\hat{\beta}_j(k)$ 快速减小为负值, 从古典回归分析看来, x_j 对 y 有正面影响. 但从岭回归的角度, x_j 对 y 有负面影响.
- **(d) 图:** $\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 都不稳定, 但二者得"和"却大致稳定, 这表明 x_1 和 x_2 之间有较强的相关系, 从变量选择的角度来看, 二者选其一即可



(c) (d)

- **(e) 图和 (f) 图:** 从全局看, 岭迹分析可用来估计某具体实例中最小二乘估计是否适用. 如果所有的岭迹比较"乱", 如图 (e), 稳定性差, 则可怀疑 OLS 可能不适用. 如图 (f) 则可放心使用 OLS.

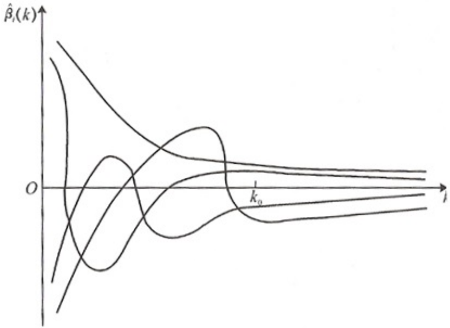


(e) (f)

7.4 岭参数 k 的选择

☞ **岭迹法**

- 各回归系数的岭估计基本稳定;
- 用最小二乘估计的符号不合理的回归系数, 其岭估计的符号变得合理;
- 回归系数没有不符合经济意义的绝对值;
- 残差平方和增加不多.



☞ 方差扩大因子法

- **方差扩大因子:** 因为

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}(k)\right) &= \mathrm{Cov}\left(\hat{\beta}(k), \hat{\beta}(k)\right) = \mathrm{Cov}\left(\left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}\right)\\
&= \left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\mathbf{X}'\mathrm{Cov}(\mathbf{y},\mathbf{y})\mathbf{X}\left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+k\mathbf{I}\right)^{-1}\\
&= \sigma^2\mathbf{c}(k)
\end{aligned}$$

  $\mathbf{c}(k)$ 的对角线元素 $c_{jj}(k)$ 即为岭估计的方差扩大因子.
- **应用:** 选择 k, 使得所有的 $c_{jj}(k) \leqslant 10$.

☞ 由残差平方和确定 k 值(略)

☞ **案例：民航客运数据分析**

```
library("car")                          # 提供 vif 函数
library("MASS")                         # 提供 lm.ridge 函数
                                        # 获取数据
civil <- read.csv("E:\\Documents\\ZUFE\\数科学院\\教学课件
                        \\应用回归分析\\数据\\civil.csv")

civil <- civil[,-1]                     # 数据预处理
civil_lm <- lm(guests~income+consume+realguests
            +civilmiles+travors,data=civil) # 做普通最小二乘回归
vif(civil_lm)                           # 方差扩大因子检验
```

vif 检验结果表明数据中的多重共线性程度比较严重，下面用“岭回归”方法解决该问题。

```
civil_scale <- scale(civil)                 # 数据标准化
civil_s <- data.frame(civil_scale)          # 生成数据框
                                            # 岭回归
civil_ridge <- lm.ridge(guests~-1+income+consume+realguests
     +civilmiles+travors,data=civil_s,lambda=seq(0,3,length=30))

civil_beta <- coef(civil_ridge)    # 提取岭回归系数

k <- civil_ridge$lambda                     # 提取岭回归系数 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)
lcol <- c("red","green","brown","blue","purple")
for(i in 1:5)
lines(k,civil_beta[,i],type="o",lty=linetype[i],pch=char[i],
                              col=lcol[i],cex=0.3)  # 绘制岭迹图
                                                    # 添加图例
legend(locator(1),inset=0.5,legend=c("income","consume","realguests",
      "civilmiles","travors"),cex=0.8,pch=char,lty=linetype,col=lcol)
```

```
# 删除变量 income 后用剩余变量再做岭回归
                                                    # 岭回归
civil_ridge <- lm.ridge(guests~-1+consume+realguests+civilmiles
                  +travors,data=civil_s,lambda=seq(0,2,length=20))

civil_beta <- coef(civil_ridge)                     # 提取岭回归系数

k <- civil_ridge$lambda                             # 提取岭回归系数 k
plot(k,k,type="n",xlab="岭参数k",ylab="岭回归系数",ylim=c(-1,1))
                                       # 创建没有任何点和线的图形区域
linetype <- c(1:4)
char <- c(18:21)
lcol <- c("red","green","brown","blue")
for(i in 1:4)                                       # 绘制岭迹图
lines(k,civil_beta[,i],type="o",lty=linetype[i],pch=char[i],
                            col=lcol[i],cex=0.3)    # 添加图例
legend(locator(1),inset=0.5,legend=c("consume","realguests","civilmiles",
                         "travors"),cex=0.8,pch=char,lty=linetype,col=lcol)
```

从岭迹可以看出，当 $\lambda = 1.4$ 是岭迹趋于平稳，且有相对合理的解释.

```
civil_ridge <- lm.ridge(guests~-1+consume+realguests+civilmiles+travors,
                                                    data=civil_s,lambda=1.4)
civil_ridge_coef <- coef(civil_ridge)
```

$$\widehat{guests}^* = 0.304consume^* - 0.088realguests^* + 0.417civilmiles^* + 0.288travors^*$$

```
civil_mean <- attr(civil_scale,"scaled:center")    # 提取中心化数据的均值
civil_sd <- attr(civil_scale,"scaled:scale")       # 提取中心化数据的标准差
                                         # 还原成原始数据后的模型系数估计值
civil_ori_coef <- civil_ridge_coef * (civil_sd[1]/civil_sd[3:6])
civil_ori_inte <- civil_mean[1] - civil_mean[3:6] %*% civil_ori_coef
                                         # 还原成原始数据后的模型截距估计值
```

对应的未标准化的岭回归方程为

$$\widehat{guests} = 417.394 + 0.069consume - 0.0070realguests + 16.790civilmiles + 0.223travors$$

☞ 案例: 法国经济数据分析

```
library(MASS)                                           # 提供 lm.ridge 函数
library(car)                                            # 提供 vif 函数

france <- read.csv("D:\\documents\\MyDoc\\JobInZufe\\Courseware
                                \\应用回归分析\\数据\\france.csv")

france_lm <- lm(import~GNP+saving+consume,data=france)  # OLS
vif(france_lm)                                          # vif 检验

france_scale <- scale(france)                           # 数据标准化
france_s <- data.frame(france_scale)                    # 转化成数据框
                                                        # 基于标准化数据做岭回归
france_ridge <- lm.ridge(import~-1+GNP+saving+consume,data=france_s,
                          lambda=seq(0,0.4,length=200))
france_beta <- coef(france_ridge)                       # 提取岭回归系数
k <- france_ridge$lambda                                # 提取岭参数
plot(k,k,type="n",xlab="岭参数 k",ylab="岭回归系数",ylim=c(-0.5,1.5))
                                                        # 绘制空图
```

```
linetype <- c(1:3)
char <- c(18:20)
lcol <- c("red","green","blue")
for(i in 1:3)                                            # 绘制岭迹
   lines(k,france_beta[,i],type="o",lty=linetype[i],pch=char[i],
                                  col=lcol[i],cex=0.3)

legend(locator(1),inset=0.8,legend=c("GNP","saving","consume"),cex=0.8,
pch=char,lty=linetype,col=lcol)
```

从岭迹可以看出，当 $\lambda = 0.15$ 是岭迹趋于平稳，且有相对合理的解释.

```
france_ridge <- lm.ridge(import~-1+GNP+saving+consume, # 选择合适的岭参数后
做岭回归
                              data=france,lambda=0.15)
france_ridge_coef <- coef(france_ridge)
france_ridge_coef
```

故岭估计模型为

$$\widehat{import}^* = 0.053GNP^* + 0.538saving^* + 0.071consume^*$$

```
france_mean <- attr(france_scale,"scaled:center") # 取原始数据的均值
france_sd <- attr(france_scale,"scaled:scale")    # 取原始数据的标准差
                                       # 计算原始数据对应的回归系数与截距
france_ori_coef <- france_ridge_coef*(france_sd[1]/france_sd[2:4])
france_ori_inte <- france_mean[1]-france_mean[2:4]%*%france_ori_coef
france_ori_coef
france_ori_inte
```

对应的未标准化的岭回归方程为

$$\widehat{import} = 13.26 + 0.008GNP + 1.481saving + 0.016consume$$

☞ 案例：法国经济数据分析(基于原始数据)

```
library(MASS)

france <- read.csv("D:\\documents\\MyDoc\\JobInZufe\\Courseware
                                \\应用回归分析\\数据\\france.csv")

france_ridge <- lm.ridge(import~1+GNP+saving+consume,data=france,
                          lambda=seq(0,0.4,length=200))
par(mai=c(0.9,0.9,0.2,0.2))
plot(france_ridge)         # 绘制岭迹图
select(france_ridge)       # 选择岭参数
```

从岭迹可以看出，当 $\lambda = 0.04$ 是岭迹趋于平稳，且有相对合理的解释.

```
lm.ridge(import~1+GNP+saving+consume,data=france,lambda=0.04)
```

故岭估计模型为

$$\widehat{import} = -9.496 + 0.0198GNP + 0.598saving + 0.183consume$$

7.5 用岭回归选择变量

☞ 用岭回归选择变量的原则

- 假定资料矩阵 $\mathbf{X}$ 各列已标准化. 则标准化岭回归系数比较稳定且绝对值很小的自变量.
- 当 k 值很小时, 标准化岭回归系数的绝对值并不小, 但是不稳定, 随着 k 值增大迅速趋于零, 类似这种岭回归系数不稳定且区域零的自变量, 也可删除.
- 提出标准化回归系数很不稳定的自变量.

☞ 应用