统计计算

2016 年 4 月 5 日

目录

1. 基于谱分解的回归分析
   - 对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解
   - 对称阵 $\mathbf{S}$ 的谱分解
   - 回归参数的有偏估计
   - 谱分解在岭回归中的应用

1. 优点
   算法比较稳定;
   在解决线性回归中经常出现的近似多重共线性问题 (即 $\mathbf{X}$ 接近非满列秩) 时具有优势.
2. 缺点
   计算量大, 精度较差.

1. 基于谱分解的回归分析
   - 对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解
   - 对称阵 $\mathbf{S}$ 的谱分解
   - 回归参数的有偏估计
   - 谱分解在岭回归中的应用

对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解

假设 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 有谱分解

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \cdots, \lambda_{m+1})\mathbf{U}^{\mathrm{T}},$$

其中 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m+1}$ 是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的 $m+1$ 个特征值 (注意可能存在 i 使得 $\lambda_i = 0$), $\mathbf{U}$ 的列是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的特征向量, 且 $\mathbf{U}$ 为正交矩阵, 于是

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}\mathbf{U}\mathbf{U}^{\mathrm{T}}\beta + \varepsilon = \mathbf{Z}\gamma + \varepsilon$$

其中 $\mathbf{Z} = \mathbf{X}\mathbf{U}, \gamma = \mathbf{U}^{\mathrm{T}}\beta$.

对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解 (续)

显然

$$\|\mathbf{Y}-\mathbf{X}\beta\|^2 = \|\mathbf{Y}-\mathbf{Z}\gamma\|^2,$$

因此 $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ 和 $\mathbf{Y}=\mathbf{Z}\gamma+\varepsilon$ 的最小二乘估计 $\hat{\beta}$ 和 $\hat{\gamma}$ 满足

$$\begin{aligned}
\hat{\beta} &= \mathbf{U}\hat{\gamma}, \\
\hat{\gamma} &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{\dagger}\mathbf{Z}^{\mathrm{T}}\mathbf{Y} = (\mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U})^{\dagger}\mathbf{Z}^{\mathrm{T}}\mathbf{Y} \\
&= \Lambda^{\dagger}\mathbf{Z}^{\mathrm{T}}\mathbf{Y}
\end{aligned}$$

对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解 (续)

$\hat{\beta}$ 的计算方法

1. 求 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \cdots, \lambda_{m+1})\mathbf{U}^{\mathrm{T}},$$

2. 计算

$$\hat{\beta} = \mathbf{U}\,\mathrm{diag}(\lambda_1^{\dagger}, \cdots, \lambda_q^{\dagger})\mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y},$$

Q 的计算方法:

$$Q = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\beta}.$$

目录

1. 基于谱分解的回归分析
   - 对称阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的谱分解
   - 对称阵 $\mathbf{S}$ 的谱分解
   - 回归参数的有偏估计
   - 谱分解在岭回归中的应用

对 S 的谱分解

考虑 $\mathbf{S} = [\mathbf{X}|\mathbf{Y}]^{\mathrm{T}}[\mathbf{X}|\mathbf{Y}]$ 的谱分解

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}},$$

其中，$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_{m_2}), \lambda_1 \geq \cdots \geq \lambda_{m_2} > 0$ 为 $\mathbf{S}$ 的特征值.

对 S 的谱分解 (续)

令 $\mathbf{D} = \mathrm{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_{m_2}})$, 且

$$\mathbf{R} = \mathbf{U}\mathbf{D}^{-1} \triangleq \begin{bmatrix} \mathbf{r}_0^{\mathrm{T}} \\ \vdots \\ \mathbf{r}_{m+1}^{\mathrm{T}} \end{bmatrix},$$

则

$$\hat{\beta}_j = -\mathbf{r}_j^{\mathrm{T}} \mathbf{r}_{m+1} / (\mathbf{r}_{m+1}^{\mathrm{T}} \mathbf{r}_{m+1}), \quad j = 0, 1, \cdots, m,$$

$$Q = 1/(\mathbf{r}_{m+1}^{\mathrm{T}} \mathbf{r}_{m+1}).$$

证明详见教材第 311 页.

目录

线性模型的 LS 估计

线性回归模型作为特殊线性模型，其重要特点是设计矩阵 $\mathbf{X}$ 为满列秩的，于是参数向量 β 可估，一切线性函数 $a'\beta$ 也可估．这就使得线性回归模型的参数估计问题变得相对简单得多．

线性模型参数的 LS 估计具有许多优良性质，其中特别重要的是 Gauss-Markov 定理，它保证了 LS 估计在线性无偏估计类中的方差最小性．如果进一步假设误差服从正态分布，则 LS 估计还具有更多更好的性质．因此，LS 估计成为广泛应用的重要估计．

线性模型 LS 估计的缺点

现代信息处理中，线性模型包含越来越多的自变量，许多实际应用问题表明：在一些情况下 LS 估计并不很理想，在个别情况下可能很不好。

20 世纪 50 年代以来，统计学家们作出了种种努力，试图改进 LS 估计．其中的一个重要方面就是寻求一些新的估计．

线性模型 LS 估计的缺点 (续)

Stein 于 1955 年证明了: 对多元正态总体 $N_r(\theta,\Sigma)$ 的均值 θ 的 LS 估计, 当维数 $r>2$ 时具有不可容许性, 即是说能够找到另外一个估计在均方误差意义下一致优于 LS 估计. 这被称为 Stein 现象.

这个令统计学家为之一惊的发现, 在参数估计理论中开辟了一个崭新的研究领域. 以此为开端, 人们引继提出了许多新的估计, 主要有岭估计、Stein 估计、主成分估计以及特征根估计等. 它们的一个共同点是有偏性, 因此人们把这些估计统称为有偏估计. 从某种意义上讲, 这些估计都改进了 LS 估计.

均方误差

我们先来看一看在什么情况下 LS 估计的性质才明显地变坏.

考虑度量估计优劣的另一个重要标准 ——均方误差 (Mean Square Error, 简记为 MSE).

θ 的估计 $\hat{\theta}$ 的均方误差

$$\mathrm{MSE}(\hat{\theta}) = E\|\hat{\theta} - \theta\|^2 = E(\hat{\theta} - \theta)'(\hat{\theta} - \theta)$$

度量了估计 $\hat{\theta}$ 与未知参数 θ 偏离的大小. 一个好的估计应该有较小的均方误差.

均方误差 (续)

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{tr}\{\mathrm{Cov}(\hat{\theta})\} + \|E\hat{\theta} - \theta\|^2.$$

这是因为

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= E[(\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta)]' \\
&\qquad \cdot [(\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta)] \\
&= E(\hat{\theta} - E\hat{\theta})'(\hat{\theta} - E\hat{\theta}) + \|E\hat{\theta} - \theta\|^2 \\
&= E[\mathrm{tr}\{(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})'\}] + \|E\hat{\theta} - \theta\|^2 \\
&= \mathrm{tr}\{E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})'\} + \|E\hat{\theta} - \theta\|^2 \\
&= \mathrm{tr}\{\mathrm{Cov}(\hat{\theta})\} + \|E\hat{\theta} - \theta\|^2
\end{aligned}$$

因此 LS 估计的 MSE 可以分解为

$$\Delta_1 = \operatorname{tr}\{\operatorname{Var}(\hat{\theta})\} = \sum_{i=1}^{r} \operatorname{Var}(\hat{\theta}_i)$$

和

$$\Delta_2 = \|E\hat{\theta} - \theta\|^2 = \sum_{i=1}^{r} (E\hat{\theta}_i - \theta_i)^2$$

显然, Δ_1 表示了估计 $\hat{\theta}$ 的各个分量方差之和, Δ_2 表示了估计 $\hat{\theta}$ 的各个分量偏差的平方和. 要使 $\operatorname{MSE}(\hat{\theta})$ 尽可能地小, 应该使 Δ_1 和 Δ_2 都比较小.

均方误差 (续)

考虑线性模型 $(\mathbf{Y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, 其中 $\mathbf{X}$ 是满列秩的, LS 估计为 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

因为 $\hat{\beta}$ 为 β 的无偏估计, 在 $\mathrm{MSE}(\hat{\beta})$ 中, $\Delta_2 = 0$, 于是

$$\mathrm{MSE}(\hat{\beta}) = \Delta_1 = \sigma^2 \mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = \sigma^2 \sum_{i=1}^{r} \frac{1}{\lambda_i}$$

其中, $\lambda_1 \geq \cdots \geq \lambda_r > 0$ 为 $\mathbf{X}'\mathbf{X}$ 的特征根.

作业

教材第 343 页: 6.2.

若 $\mathbf{X}'\mathbf{X}$ 至少有一个特征根非常小，非常接近于 0，则设计矩阵 $\mathbf{X}$ 呈病态，即 $\mathbf{X}$ 的列向量之间存在近似的线性关系。这种关系称为设计矩阵 $\mathbf{X}$ 或线性模型的复共线性.

对 LS 估计 $\hat{\beta}$，有 $\Delta_2 = 0$，尽管 Gauss-Markov 定理保证了 $\mathrm{MSE}(\hat{\beta})$ 在线性无偏估计类中最小，但它本身的值却很大，从而 $\mathrm{MSE}(\hat{\beta})$ 将很大.

在这种情况下，LS 估计 $\hat{\beta}$ 不再是 β 的一个良好估计。而导致 LS 估计性质变坏的原因就是复共线性.

复共线性 (续)

产生复共线性的原因是多方面的, 可能是由于自变量之间客观上就有近似的线性关系, 也可能是实际应用中数据收集的局限性所致. 在现代诸如医学信息、生物信息等大型回归问题中, LS估计的性质往往不理想, 甚至可能很差.

岭估计的提出

当矩阵 $\mathbf{X}$ 呈病态情形 (如自变量出现近似线性关系) 时, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 理论上非奇异, 但在计算过程中可能变为奇异的了. 例如

$$\mathbf{X}=\begin{bmatrix}1 & 1 & 1\\ \varepsilon & 0 & 0\\ 0 & \varepsilon & 0\\ 0 & 0 & \varepsilon\end{bmatrix}\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}=\begin{bmatrix}1+\varepsilon^2 & 1 & 1\\ 1 & 1+\varepsilon^2 & 1\\ 1 & 1 & 1+\varepsilon^2\end{bmatrix}.$$

当 $\varepsilon>0$ 时, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 理论上非奇异, 但如果 ε 接近于计算机精度时, $1+\varepsilon^2$ 被舍入为 1, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 就奇异了.

岭估计的提出 (续)

如果用公式 $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ 计算 $\hat{\beta}$, 显然会引起相当大的误差, 从而不稳定.

为了克服上述问题, 可取一个 $k>0$, 用 $\mathbf{X}^{\mathrm{T}}\mathbf{X}+k\mathbf{I}$ 代替 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. 当 k 适当时, 计算结果具有较好的性能.

岭估计

对线性模型 $(\mathbf{Y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$，参数 β 的岭估计 (Ridge Estimation) 为

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

其中，$k > 0$ 称为岭参数.

岭估计是由 Hoerl 和 Kennard 于 1970 年提出的，相关的理论研究和应用得到广泛的重视，成为目前最有影响的一种有偏估计.

岭估计的意义

1. 若 k 取与试验数据 $\mathbf{Y}$ 无关的常数, 则 $\hat{\beta}(k)$ 为线性估计. 取不同的 k 就得到不同的岭估计, 所以上面的定义给出了一个很大的估计类. 特别地, $\hat{\beta}(0)$ 就是 β 的 LS 估计;
2. 岭估计是把 LS 估计中的 $\mathbf{X}'\mathbf{X}$ 换成 $\mathbf{X}'\mathbf{X}+k\mathbf{I}$, 因此当 $\mathbf{X}$ 呈病态时, $\mathbf{X}'\mathbf{X}+k\mathbf{I}$ 的特征根 $\lambda_1+k,\cdots,\lambda_r+k$ 接近于 0 的程度就会降低, 从而打破了原来设计矩阵的复共线性.

岭估计的性质

1. $\hat{\beta}(k) = \mathbf{A}_k\hat{\beta}$, 其中 $\mathbf{A}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}$. 因此, 岭估计是 LS 估计的一个线性变换;
2. $E\hat{\beta}(k) = \mathbf{A}_k\beta$. 因此, 只要 $\mathbf{A}_k \neq \mathbf{I}$, 岭估计就是 β 的有偏估计. 显然 $\mathbf{A}_k = \mathbf{I} \Longleftrightarrow k = 0$, 这表明岭估计类中除了 LS 估计之外, 所有估计均为有偏估计;
3. 对任意 $k > 0, \|\hat{\beta}\| \neq 0$, 均有 $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$. 这表明岭估计是对 LS 估计向原点的压缩;
4. 可以证明: 存在 $k > 0$, 使得 $\mathrm{MSE}(\hat{\beta}(k)) < \mathrm{MSE}(\hat{\beta})$, 这表明存在 $k > 0$, 使得在均方误差意义下, 岭估计优于 LS 估计.

岭参数的选择

引进岭估计的目的就是要减少均方误差，所以我们应该选择使 $\mathrm{MSE}(\hat{\beta}(k))$ 达到最小的 k.

尽管上述性质保证了 k 的存在性，但 k 的最优值不仅依赖于模型的未知参数 β 和 σ^2，而且这种依赖关系没有显示表达式，这就使得最优的 k 的确定十分困难. 而该问题在实际应用中是不可回避的.

统计学家们作了很多工作，提出了十余种选取 k 的方法.

岭迹法

将岭估计 $\hat{\beta}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$ 的分量 $\hat{\beta}_i(k)$ 作为 k 的函数，当 k 在 $[0, +\infty)$ 变化时在平面直角坐标系所描出的图形称为岭迹.

选择 k 的岭迹法为：将 p 个分量 $\hat{\beta}_i(k)$ 的岭迹画在同一个坐标系内，并且兼顾回归系数没有不合理的符号、残差平方和上升不太多等，如果在 k^* 附近 r 条岭迹大体稳定了，就可以考虑取最优的 $k = k^*$.

岭迹法 (续)

在计算岭迹时如果按照岭估计的定义计算 $\hat{\beta}(k)$, 则对每个 k 都要计算 $\mathbf{X}'\mathbf{X}+k\mathbf{I}$ 的逆, 因而计算量很大.

实际应用时, 考虑到

$$\begin{aligned}\hat{\beta}(k) &= (\mathbf{U}\Lambda\mathbf{U}' + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{U}(\Lambda + k\mathbf{I})^{-1}\mathbf{U}'\mathbf{X}'\mathbf{Y} \\ &= \sum_{i=1}^{r}\left(\frac{1}{\lambda_i + k}\right)u_i u_i'\mathbf{X}'\mathbf{Y}\end{aligned}$$

其中, u_i 是 $\mathbf{X}'\mathbf{X}$ 相应于特征根 λ_i 的特征向量. 因此, 计算一次 $\mathbf{X}'\mathbf{X}$ 的特征根和特征向量, 对每个 k 就很容易计算出岭迹了.

岭回归估计的计算

4. 基于谱分解求 $\hat{\beta}(k)$ 的简便算法.

   关键是:

- 新的数据叉积矩阵

$$\mathbf{S}(k) = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I} & \mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \mathbf{Y}^{\mathrm{T}}\mathbf{X} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y} + k \end{bmatrix}$$

  的特征值 $\lambda_i(k)$ 与原 $\mathbf{S}$ 的特征值 λ_i 满足 $\lambda_i(k) = \lambda_i + k$, 而相应的特征向量不变.

- 直接利用教材第 311 页的结果立即可得第 313 页的 (4.7), 即岭回归估计量及残差平方和的计算公式.

问　题?