统计计算

2016 年 3 月 31 日

目录

目录

线性模型

线性模型是数理统计学中一种非常重要的模型, 其理论成果十分丰富, 完善, 应用十分广泛. 内容主要包括:

1. 回归分析: 研究变量之间的相互关系, 建立变量之间的经验公式, 以达到预测和控制的目的.
2. 方差分析: 分析哪些因素对指标起显著影响, 并希望知道这些因素在什么时候起最好的影响.

前已述及，回归分析的基本问题是寻找因变量与自变量之间的函数关系 f.

函数 f 的可选范围如果太广，将很难甚至无法得到. 为此，通常将这个可选集合缩小到线性函数类这一可处理的范围.

线性回归分析的基本问题：在某种优化准则下寻求最能逼近观测数据的线性函数.

研究线性模型的重要性

1. 从工具方面来说
   线性问题在数学上有成熟的处理方法，相关成果非常丰富.
2. 从应用方面来说
   (a) 大量线性问题;
   (b) 许多非线性问题可完全转化为线性情形;
   (c) 不能转化的也可采用线性化方法化成线性模型.

研究线性模型的重要性 (续)

1. 部分非线性可以完全转化为线性;
2. 一元多项式回归可化为多元线性模型;
3. 一元非线性函数可用多项式逼近, 因而也可线性化;
4. 将上面的 2, 3 推广到多元情形: 多元多项式; 多元非线性函数用多项式逼近;
5. 复杂的确定性函数的线性组合.

目录

变量之间的关系

1. 确定关系.
2. 相关关系 (这是我们所关心的). 一个实际的例子: 第 320 页例 5.1.

因为有随机项, 不可能获得完全精确的函数关系, 只能在某种优化准则下寻求某个函数最能逼近真实数据.

回归分析是处理因变量 y 与自变量 $x_1, \cdots, x_p$ 之间相关关系的一种有效的统计方法.

回归分析的基本问题是如何找出适当的函数 $f(x_1, \cdots, x_p)$, 使得

$$y = f(x_1, \cdots, x_p) + \varepsilon,$$

其中 ε 为随机误差项 (不可观测), 因此不可能获得完全精确的函数关系, 只能在某种优化准则下寻求某个函数最能逼近真实数据.

变量的类型

研究的手段：从数据出发，通过观测数据寻找自变量与因变量之间的相关关系.

变量的类型：因变量 y 是观测得到的，应是随机变量. 对自变量而言，有两种看法.

回归分析的分类

回归分析的基本问题是建立因变量与自变量之间的某种函数关系 f. 根据函数关系可分为:

1. 线性回归分析
2. 非线性回归分析

回归问题的分类

根据自变量和因变量的个数来分类:

1. 一个自变量对一个因变量: 一对一回归分析
2. 多个自变量对一个因变量: 多对一回归分析
3. 多个自变量对多个因变量: 多对多回归分析

实际中最多的是第 2 种, 这就是常见的多元回归分析.

多元线性回归问题的数学模型

教材第 297 页 §1 开始部分.

1. 线性模型: (1.1) 式;
2. 矩阵形式: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$;
3. 基本假设: $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$;
4. 待估参数: β, σ^2.

多元线性回归问题的数学模型 (续)

回归分析的主要任务:

1. 参数 β 和 σ^2 的估计;
2. 参数 β 的线性函数的统计检验;
3. 预测问题;
4. 回归自变量的筛选.

目录

参数 β 和 σ^2 的估计问题

对 β 的估计方法有很多, 最常用的是最小二乘方法.

称 $\hat{\beta}$ 是参数 β 的最小二乘估计, 如果

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^q} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

称 $Q = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$ 为最小二乘残差平方和.

参数 β 和 σ^2 的估计问题 (续)

最小二乘方法的特点:

1. 不要求更多的统计性质;
2. 具有良好的性能;
3. 计算方便.

参数 β 和 σ^2 的估计问题 (续)

对 σ^2 的估计: $Q/(n-q)$ 是 σ^2 的无偏估计.

回归系数 β 线性假设的统计假设检验

假设

$$H_0: \mathbf{L}\beta = \gamma,$$

其中 $\mathbf{L} \in \mathbb{R}^{s\times q}$, $\operatorname{rank}\mathbf{L} = s$, 且 $\gamma \in \mathbb{R}^s$ 已知.

回归系数 β 线性假设的统计假设检验 (续)

构造统计量

$$F = \frac{(Q_{H_0} - Q)/s}{Q/(n-q)},$$

其中 Q_{H_0} 是 H_0 成立时的最小二乘残差, 即

$$Q_{H_0} = \min_{\mathbf{L}\beta=\gamma} \|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

回归系数 β 线性假设的统计假设检验 (续)

可以证明: 在假设 $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ 下, 当 H_0 成立时, 有

$$F \overset{H_0}{\sim} F(s, n-q).$$

回归方程和回归系数的显著性检验

检验问题: 第 301 页的 H_0 和 $H_0^{(i)}$.

1. 回归方程的显著性检验:

$$\mathbf{L} = \mathbf{I}_m, \gamma = \mathbf{0};$$

2. 回归系数的显著性检验:

$$\mathbf{L} = [0, \cdots, 0, 1, 0, \cdots, 0], \gamma = 0.$$

回归方程和回归系数的显著性检验 (续)

检验统计量:

$$F = \frac{U/m}{Q/(n-q)} \overset{H_0}{\sim} F(m, n-q)$$

$$F_i = \frac{P_i}{Q/(n-q)} \overset{H_0^{(i)}}{\sim} F(1, n-q), \quad i = 1, \cdots, m,$$

其中, $U = Q_{H_0} - Q, P_i = Q_i - Q.$

因此, 关键是计算 Q_{H_0} 和 Q_i.

回归方程和回归系数的显著性检验 (续)

命题

考虑假设

$$H_0 : \mathbf{L}_{s\times q}\beta_{q\times q} = \gamma_{q\times 1},$$

其中 $\operatorname{rank}\mathbf{L} = s$. 在假设 H_0 下必有

$$Q_{H_0} - Q = (\mathbf{L}\hat{\beta} - \gamma)^{\mathrm{T}}\left(\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{L}^{\mathrm{T}}\right)^{(1,2)}(\mathbf{L}\hat{\beta} - \gamma).$$

该命题给出了 $U = Q_{H_0} - Q$ 的直接计算方法. 还需考虑 Q_i 的计算方法.

问　题?

目录

目录

$\hat{\beta}$ 和 Q 的计算

先构造矩阵

$$\mathbf{S} = [\mathbf{X}|\mathbf{Y}]^{\mathrm{T}}[\mathbf{X}|\mathbf{Y}] = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \mathbf{Y}^{\mathrm{T}}\mathbf{X} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \end{bmatrix}$$

其中, $\mathbf{X} \in \mathbb{R}^{n\times q}, q = m+1.$

对 $\mathbf{S}$ 使用广义消去变换 $G_2\mathbf{T}_i$, 可得

$$(G_2\mathbf{T}_q)\cdots(G_2\mathbf{T}_1)\mathbf{S}=$$
$$\begin{bmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)} & (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ -\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \end{bmatrix}.$$

由于
$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$
是 β 的一个最小二乘估计, 且
$$\begin{aligned}
Q &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2 \\
&= (\mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y})^{\mathrm{T}} \\
&\qquad \cdot (\mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y}) \\
&= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{X}^{\mathrm{T}}\mathbf{Y}
\end{aligned}$$

$\hat{\beta}$ 和 Q 的计算 (续)

因此,

$$(G_2\mathbf{T}_q)\cdots(G_2\mathbf{T}_1)\mathbf{S} = \begin{bmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)} & \hat{\beta} \\ -\hat{\beta}^{\mathrm{T}} & Q \end{bmatrix}$$

即由消去变换就可同时求出 $\hat{\beta}$ 和 Q.

特别地, 若 $\mathrm{rank}\mathbf{X} = q$, 即 $\mathbf{X}$ 是满列秩的, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 可逆, 于是广义消去变换成为通常的消去变换, 且

$$\mathbf{T}_q \cdots \mathbf{T}_1 \mathbf{S} = \begin{bmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} & \hat{\beta} \\ -\hat{\beta}^{\mathrm{T}} & Q \end{bmatrix}$$

其中 $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

$\hat{\beta}$ 和 Q 的计算 (续)

综合以上分析可得 $\hat{\beta}$ 和 Q 的算法:

1. 教材第 299 页定理 2.1 (满列秩情形);
2. 教材第 300 页定理 2.2 (一般情形).

检验问题: 第 301 页.

1. 回归方程的显著性检验:

$$\mathbf{L} = \mathbf{I}, \gamma = \mathbf{0};$$

2. 回归系数的显著性检验:

$$\mathbf{L} = [0, \cdots, 0, 1, 0, \cdots, 0], \gamma = 0.$$

检验统计量 (第 301 页):

$$F = \frac{(Q_{H_0} - Q)/m}{Q/(n-m-1)}, F_i = \frac{Q_i - Q}{Q/(n-m-1)},$$

其中

1. Q 表示全部变量均在线性回归模型中时的最小二乘残差平方和;
2. Q_{H_0} 表示全部变量均不在线性回归模型中时的最小二乘残差平方和;
3. Q_i 表示变量 x_i 未在回归模型中时的最小二乘残差平方和.

假设检验 (续)

在假设 H_0 和 $H_0^{(i)}$ 分别成立的条件下, 可以证明检验统计量的分布满足:

$$F \sim F(m, n-m-1),$$
$$F_i \sim F(1, n-m-1).$$

以下考虑 F 和 F_i 的计算, 关键是 $Q_{H_0} - Q$ 和 $Q_i - Q$ 的计算.

Q_{H_0} 的直接计算

命题

考虑假设

$$H_0 : \mathbf{L}_{s\times q}\beta_{q\times 1} = \gamma_{s\times 1},$$

其中 $\operatorname{rank}\mathbf{L} = s$. 在假设 H_0 成立下必有

$$Q_{H_0} - Q = (\mathbf{L}\hat{\beta} - \gamma)^{\mathrm{T}}\left(\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(1,2)}\mathbf{L}^{\mathrm{T}}\right)^{(1,2)}(\mathbf{L}\hat{\beta} - \gamma).$$

上述命题给出了 $Q_{H_0} - Q$ 的一种直接计算方法.

作业

教材第 343 页 6.1.

Q_i 的计算

下面考虑 Q_i 的计算方法.

这里采用消去变换.

回忆: 当全部自变量均在线性模型中时, Q 是通过 $q = m+1$ 个消去变换 $\mathbf{T}_q \cdots \mathbf{T}_1 \mathbf{S}$ 获得的. 即是说: 变量 $x_1, \cdots, x_q$ 在模型中时, 就需要施行消去变换 $\mathbf{T}_1, \cdots, \mathbf{T}_q$.

因此, 当假设 $H_0^{(i)} : \beta_i = 0$ 成立时, 变量 x_i 未出现在模型中, 则不施行消去变换 $\mathbf{T}_i$; 反之, 若 x_i 出现在模型中, 则需施行消去变换 $\mathbf{T}_i$.

Q_i 的计算 (续)

定理

第 301 页定理 2.3.

定理

第 302 页定理 2.4.

目录

设 $\mathbf{S}$ 为对称正定阵, 可对 $\mathbf{S}$ 进行 Cholesky 分解为

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \mathbf{Y}^{\mathrm{T}}\mathbf{X} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{T}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{t}_{\mathbf{y}}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix} \begin{bmatrix} \mathbf{T} & \mathbf{t}_{\mathbf{y}} \\ \mathbf{0}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix},$$

其中 $\mathbf{T}$ 为 $q \times q$ 阶上三角阵. 于是

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{T}^{\mathrm{T}}\mathbf{T},$$
$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{T}^{\mathrm{T}}\mathbf{t}_{\mathbf{y}},$$
$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{t}_{\mathbf{y}}^{\mathrm{T}}\mathbf{t}_{\mathbf{y}} + t_{\mathbf{yy}}^2.$$

从而

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{T}^{-1}\mathbf{t}_{\mathbf{y}},\\ Q &= t_{\mathbf{yy}}^2,\\ Q_{H_0} - Q &= (\mathbf{L}\hat{\beta} - \gamma)^{\mathrm{T}}\left(\mathbf{L}\mathbf{T}^{-1}\mathbf{T}^{-\mathrm{T}}\mathbf{L}^{\mathrm{T}}\right)^{-1}(\mathbf{L}\hat{\beta} - \gamma).\end{aligned}$$

于是可得教材第 303 页算法 2.1.

注: 对于 $Q_{H_0} - Q$, 教材 (2.7) 式需要计算非三角矩阵之逆, 计算复杂度较高, 因此考虑将其分解后再处理.

问　题?

目录

1. 引言
   - 线性模型
   - 回归分析
   - 最小二乘法
2. 基于正规方程的回归分析方法
   - 用消去变换作回归分析
   - 用 Cholesky 分解作回归分析
3. 基于 QR 分解的回归分析方法
   - 基于 Householder 变换的回归分析
   - 基于 Givens 变换的回归分析

见教材第 303–304 页.

优点: 计算精度高, 算法稳定性好.

目录

首先考虑 $\mathbf{X}$ 满列秩的情形.
对矩阵 $[\mathbf{X}|\mathbf{Y}]$, 存在 Householder 矩阵 $\mathbf{H}$, 使得

$$\mathbf{H}[\mathbf{X}|\mathbf{Y}] = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix},$$

其中, $\mathbf{T}$ 为 $m+1$ 阶上三角矩阵, $\mathbf{t_y}$ 为 $m+1$ 维向量, $\mathbf{t_z}$ 为 $n-m-1$ 维向量.

因为

$$[\mathbf{X}|\mathbf{Y}]^{\mathrm{T}}[\mathbf{X}|\mathbf{Y}] = \left(\mathbf{H}[\mathbf{X}|\mathbf{Y}]\right)^{\mathrm{T}}\left(\mathbf{H}[\mathbf{X}|\mathbf{Y}]\right),$$

所以

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \mathbf{Y}^{\mathrm{T}}\mathbf{X} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix},$$

于是 $\hat{\beta} = \mathbf{T}^{-1}\mathbf{t_y}, Q = \mathbf{t_z}^{\mathrm{T}}\mathbf{t_z}$.

上述计算过程见教材第 304 页算法 3.1 (1) — (3).

由于 $\mathbf{X}$ 满列秩,

$$Q_{H_0} - Q = (\mathbf{L}\hat{\beta} - \gamma)^{\mathrm{T}} \left(\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}} \right)^{-1} (\mathbf{L}\hat{\beta} - \gamma).$$

由 $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$ 知

$$\begin{aligned} \mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}} &= \mathbf{L}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{L}^{\mathrm{T}} \\ &= \mathbf{L}\mathbf{T}^{-1}\mathbf{T}^{-\mathrm{T}}\mathbf{L}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{B}, \end{aligned}$$

其中 $\mathbf{B} = \mathbf{T}^{-\mathrm{T}}\mathbf{L}^{\mathrm{T}}$ 是秩为 s 的 $(m+1) \times s$ 阶矩阵.

对 $\mathbf{B}$ 作 QR 分解 (可用 Householder 变换), 使得 $\mathbf{B}=\mathbf{H}_1^{\mathrm{T}}\begin{bmatrix}\mathbf{R}\\\mathbf{O}\end{bmatrix}$, 其中 $\mathbf{R}$ 为 s 阶上三角矩阵. 显然 $\mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{R}^{\mathrm{T}}\mathbf{R}$, 于是

$$\begin{aligned}Q_{H_0}-Q&=(\mathbf{L}\hat{\beta}-\gamma)^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}(\mathbf{L}\hat{\beta}-\gamma)\\&=(\mathbf{L}\hat{\beta}-\gamma)^{\mathrm{T}}(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}(\mathbf{L}\hat{\beta}-\gamma)\\&=\mathbf{Z}^{\mathrm{T}}\mathbf{Z},\end{aligned}$$

其中 $\mathbf{Z}=\mathbf{R}^{-\mathrm{T}}(\mathbf{L}\hat{\beta}-\gamma)$. $Q_{H_0}-Q$ 的计算化为计算 $\mathbf{Z}$, 即三角系数线性方程组 $\mathbf{R}^{\mathrm{T}}\mathbf{Z}=\mathbf{L}\hat{\beta}-\gamma$ 的解.

计算 F (续)

计算的详细过程为教材第 304 页 (4)—(7). 换一个写法为

1. 令 $\mathbf{W} = \mathbf{L}\hat{\beta} - \gamma$;
2. 求方程 $\mathbf{T}^{\mathrm{T}}\mathbf{B} = \mathbf{L}^{\mathrm{T}}$ 的解 $\mathbf{B}$;
3. 对 $\mathbf{B}$ 作 QR 分解, $\mathbf{B} = \mathbf{H}_1^{\mathrm{T}} \begin{bmatrix} \mathbf{R} \\ \mathbf{O} \end{bmatrix}$, 其中 $\mathbf{R}$ 为 s 阶上三角矩阵;
4. 求三角系数阵方程 $\mathbf{R}^{\mathrm{T}}\mathbf{Z} = \mathbf{W}$ 的解 $\mathbf{Z}$;
5. 计算 $Q_{H_0} - Q = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$.

基于 Householder 变换的回归分析

例: 教材第 305 页例 3.1.

$\mathbf{X}$ 非满列秩情形

上面给出的算法都是假定 $\mathbf{X}$ 为满列秩条件下导出的，但实际应用中可能出现 $\mathbf{X}$ 不是满列秩的情况.

此时，$\hat{\beta}$ 的最小二乘估计不唯一，其通解为

$$\hat{\beta} = \mathbf{X}^{\dagger}\mathbf{Y} + (\mathbf{I} - \mathbf{X}^{\dagger}\mathbf{X})\mathbf{z},$$

其中 $\mathbf{z}$ 为任意 $m+1$ 维向量.

$\mathbf{X}$ 非满列秩情形 (续)

究竟取哪一个估计更适当呢? 从自变量对因变量预测的角度来看, 任何一个具体的 $\hat{\beta}$ 都可作为回归方程中的回归系数. 因为

$$\mathbf{X}\hat{\beta} = \mathbf{X}\mathbf{X}^{\dagger}\mathbf{Y} + \mathbf{X}(\mathbf{I} - \mathbf{X}^{\dagger}\mathbf{X})\mathbf{z} = \mathbf{X}\mathbf{X}^{\dagger}\mathbf{Y}$$

与 $\mathbf{z}$ 无关, 而 $\mathbf{X}\hat{\beta}$ 恰好是自变量对因变量的预测.

在实际应用中可取最小范数最小二乘解 $\hat{\beta} = \mathbf{X}^{\dagger}\mathbf{Y}$, 利用 Householder 变换求矩阵广义逆的方法即可求出最小二乘估计.

目录

Givens 变换与 Householder 变换一样可用于计算回归统计量.

更主要地，在增加新的数据或剔除某个异常数据时，Givens 变换更具特殊意义.

基于 Givens 变换的回归分析 (续)

设原始数据阵为 $[\mathbf{X}|\mathbf{Y}]$, 且存在正交矩阵 $\mathbf{H}$ 使得

$$\mathbf{H}[\mathbf{X}|\mathbf{Y}] = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix},$$

则参数的估计可以在此基础上操作, 而不必在增删以后的新数据阵上重复前述算法.

增加一组数据的算法

设增加新数据 $\mathbf{a}^{\mathrm{T}} = [\mathbf{x}_0^{\mathrm{T}}|y_0]$，于是数据阵为 $\begin{bmatrix} \mathbf{X}|\mathbf{Y} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix}$. 令 $\widetilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}$, 则 $\widetilde{\mathbf{H}}$ 为正交矩阵, 且

$$\widetilde{\mathbf{H}} \begin{bmatrix} \mathbf{X}|\mathbf{Y} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{H}[\mathbf{X}|\mathbf{Y}] \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{t}_{\mathbf{y}} \\ \mathbf{0} & \mathbf{t}_{\mathbf{z}} \\ \mathbf{x}_0^{\mathrm{T}} & y_0 \end{bmatrix}$$

$$\triangleq \mathbf{E}_{(n+1)\times(m+2)}.$$

增加一组数据的算法 (续)

对 $\mathbf{E}$ 作 Givens 变换

$$\mathbf{G}_{(m+1)(n+1)} \cdots \mathbf{G}_{2(n+1)} \mathbf{G}_{1(n+1)} \mathbf{E} = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{O} & \mathbf{r_z} \end{bmatrix},$$

其中 $\mathbf{R}$ 为 $m+1$ 阶上三角阵. 即存在正交矩阵 $\mathbf{G} = \mathbf{G}_{(m+1)(n+1)} \cdots \mathbf{G}_{2(n+1)} \mathbf{G}_{1(n+1)} \widetilde{\mathbf{H}}$, 使得

$$\mathbf{G} \begin{bmatrix} \mathbf{X}|\mathbf{Y} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{O} & \mathbf{r_z} \end{bmatrix},$$

从而可得增加新数据后的估计

$$\hat{\beta}^* = \mathbf{R}^{-1} \mathbf{r_y}, \quad Q^* = \mathbf{r_z}^{\mathrm{T}} \mathbf{r_z}.$$

删除一组数据的算法

假设经过 n 次观测，存在正交矩阵 $\mathbf{H}$ 使得

$$\mathbf{H}[\mathbf{X}|\mathbf{Y}] = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix}$$

其中 $\mathbf{T}, \mathbf{t_y}, \mathbf{t_z}$ 均已知.

记删去第 k 组数据 $\mathbf{a}_k^{\mathrm{T}} = [\mathbf{x}_k^{\mathrm{T}}, y_k]$ 后的数据阵为 $[\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}]$.

因为数据的顺序与估计量无关，所以可以反过来看：是数据 $[\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}]$ 增加一组数据 $\mathbf{a}_k^{\mathrm{T}}$ 后成为数据 $[\mathbf{X}|\mathbf{Y}]$.

删除一组数据的算法 (续)

按上面的方法, 设已有正交矩阵 $\widetilde{\mathbf{H}}$ (注意这是待求的), 使得

$$\widetilde{\mathbf{H}}[\widetilde{\mathbf{X}}|\widetilde{\mathbf{Y}}] = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{O} & \mathbf{r_z} \end{bmatrix},$$

于是 $\hat{\beta}_* = \mathbf{R}^{-1}\mathbf{r_y}, Q_* = \mathbf{r_z}^{\mathrm{T}}\mathbf{r_z}$.

令 $\overline{\mathbf{H}} = \begin{bmatrix} \widetilde{\mathbf{H}} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}$, 且

$$\mathbf{E} \triangleq \overline{\mathbf{H}} \begin{bmatrix} \widetilde{\mathbf{X}}|\widetilde{\mathbf{Y}} \\ \mathbf{a}_k^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{H}}[\widetilde{\mathbf{X}}|\widetilde{\mathbf{Y}}] \\ \mathbf{a}_k^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{0} & \mathbf{r_z} \\ \mathbf{x}_k^{\mathrm{T}} & y_k \end{bmatrix}.$$

删除一组数据的算法 (续)

对 $\mathbf{E}$ 作 Givens 变换

$$\mathbf{G}_{(m+1)n}\cdots\mathbf{G}_{2n}\mathbf{G}_{1n}\mathbf{E} = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{r_z} \\ \mathbf{0}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{t_z} \end{bmatrix},$$

注意: 因为 $\mathbf{t_z} = \begin{bmatrix} \mathbf{r_z} \\ t_{\mathbf{yy}} \end{bmatrix}$, 所以 $\mathbf{r_z}$ 和 t_{yy} 是知道的.

令 $\mathbf{G} = (\mathbf{G}_{(m+1)n}\cdots\mathbf{G}_{2n}\mathbf{G}_{1n})^{\mathrm{T}}$ (注意是待求的), 则 $\mathbf{G}$ 为正交矩阵, 且

$$\mathbf{G}\begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{O} & \mathbf{r_z} \\ \mathbf{0}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix} = \mathbf{E} = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{O} & \mathbf{r_z} \\ \mathbf{x}_k^{\mathrm{T}} & y_k \end{bmatrix}.$$

删除一组数据的算法 (续)

因为上式两边矩阵第二行相同, 所以只需计算正交矩阵 $\mathbf{V}$ 使得

$$\mathbf{V}\begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{0}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{x}_k^{\mathrm{T}} & y_k \end{bmatrix}.$$

如果能够求出 $\mathbf{V}$, 则利用已知数据 $\begin{bmatrix} \mathbf{T} & \mathbf{t_y} \\ \mathbf{0}^{\mathrm{T}} & t_{\mathbf{yy}} \end{bmatrix}$, $\begin{bmatrix} \mathbf{R} & \mathbf{r_y} \\ \mathbf{x}_k^{\mathrm{T}} & y_k \end{bmatrix}$ 中的 $\mathbf{R}$ 和 $\mathbf{r_y}$ 就可以求出, 再结合前面的 $\mathbf{r_z}$ 即可求出 $\hat{\beta}_*$ 和 Q_*.

以下的问题是如何计算正交矩阵 $\mathbf{V}$?

删除一组数据的算法 (续)

将 $\mathbf{V}$ 分块为 $\begin{bmatrix} \mathbf{V}_0 & * \\ * & * \end{bmatrix}$, 则 $\mathbf{V}_0\mathbf{T} = \mathbf{R}$, 从而 $\mathbf{V}_0$ 为上三角矩阵.

利用上式最后一行相等可计算 $\mathbf{V}$ 的最后一行.

利用单位正交的条件可逐个求出 $\mathbf{V}$ 的第 j 行, 其中 $j = m+1,\cdots,2,1$. 因为 $\mathbf{V}$ 第 j 行有 $m_2 - j + 1$ 个待定数, 而第 309 页式 (3.7) 中刚好有 $m_2 - j + 1$ 个方程.

删除一组数据的算法 (续)

例: 教材第 309 页例 3.2.

问　题?