3.1 数据的回归分析

王正明　易泰河

系统工程学院　军事建模与仿真系

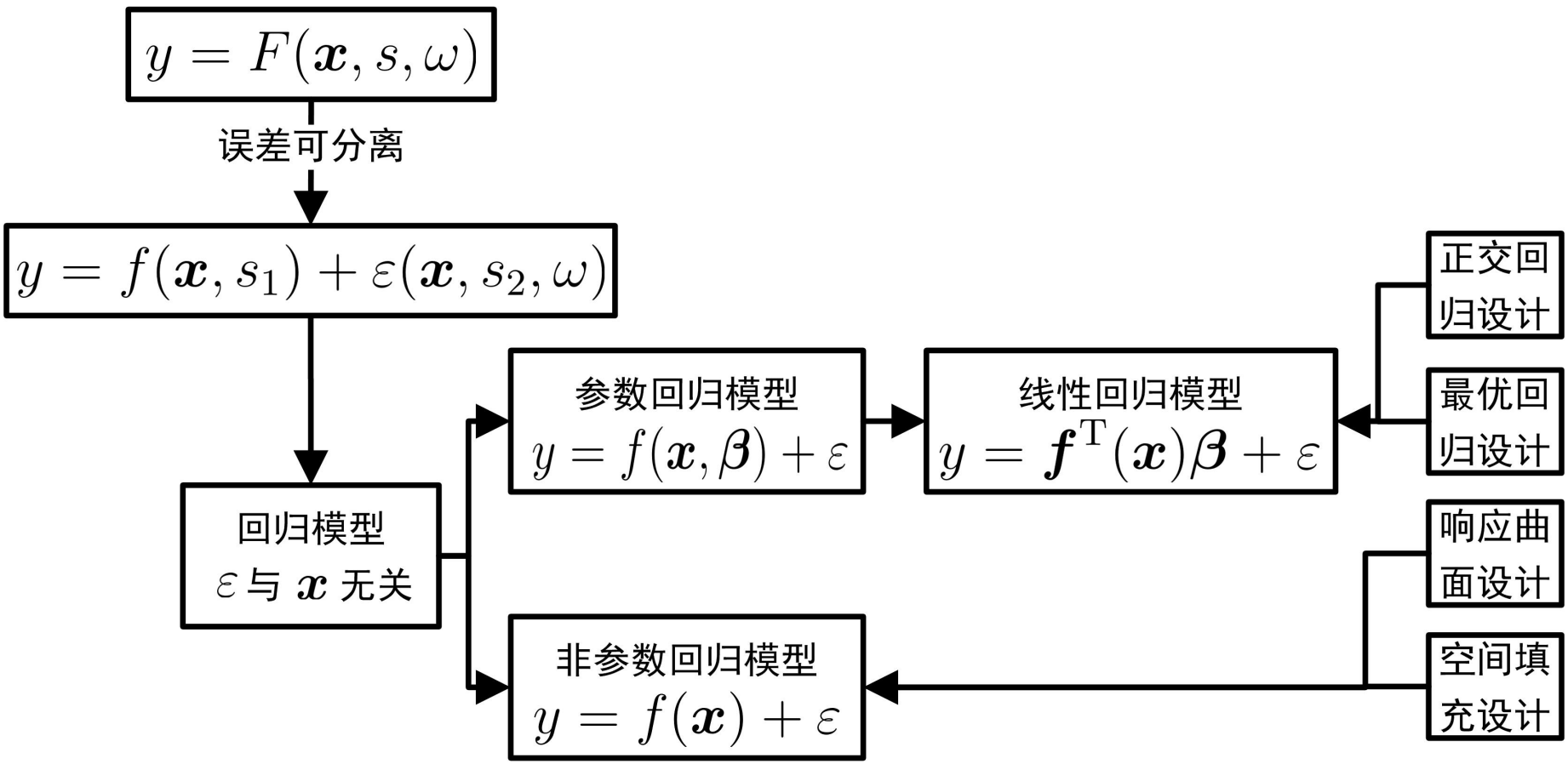
2019 年 12 月 23 日

引言

- 固定效应模型: 从处理到效应的映射.
- 回归试验: 寻找响应变量与诸定量因子之间的数量规律, 解决预测、控制和优化的问题.
- 基本思想是把试验安排、数据处理和参数估计的精度统一考虑.

响应模型于试验设计之间的关系

- 根据响应模型 $y = F(\boldsymbol{x}, s, \omega)$ 的形式分类.



- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 x_2 + \varepsilon$ 是线性回归模型吗?

设 $\boldsymbol{x} \in \mathbb{R}^p$, 为什么线性回归模型要写作

$$y = \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\beta} + \varepsilon = [f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x})] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} + \varepsilon$$

而不写作 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$?

对回归模型的两种理解

例 (机理模型)

根据变量之间的物理关系建立的模型. 如万有引力定律:

$$F = G\frac{m_1 m_2}{r^2}.$$

例 (经验模型)

从数据中归纳得到的模型, 如《遗传的身高向平均数方向的回归》

$$y = 33.73 + 0.516x.$$

问: 胡克定律 $F = -kx$ 是经验模型还是机理模型?

问: 回归模型反映了变量之间的因果关系还是相关关系?

日期	最快配速	最慢配速	平均配速	跑步时长	跑步距离 (km)	消耗能量 (kcal)
$12-10$	$5'49''$	$6'14''$	$5'56''$	$00:30:08$	6.06	351
$12-09$	$5'08''$	$6'09''$	$5'18''$	$01:17:38$	14.17	983
$12-04$	$4'50''$	$5'34''$	$5'28''$	$00:28:04$	5.29	367
$12-02$	$5'23''$	$5'59''$	$5'38''$	$00:45:42$	8.10	562
$12-01$	$4'34''$	$5'41''$	$5'02''$	$00:25:25$	5.04	350
$11-29$	$4'55''$	$5'42''$	$5'19''$	$00:37:18$	7.00	485
$11-27$	$4'50''$	$5'30''$	$5'15''$	$00:28:14$	5.36	372
$11-26$	$5'14''$	$5'49''$	$5'24''$	$01:15:50$	14.03	974
$11-22$	$5'10''$	$6'05''$	$5'27''$	$00:38:30$	7.04	488
$11-21$	-	-	$7'10''$	$00:01:55$	0.26	18
$11-20$	$4'33''$	$6'00''$	$5'10''$	$00:26:08$	5.04	350
$11-18$	$5'16''$	$5'57''$	$5'40''$	$00:59:33$	10.50	728
$11-15$	$5'23''$	$6'15''$	$5'45''$	$00:40:39$	7.05	489
$11-14$	$5'13''$	$7'00''$	$5'41''$	$00:40:30$	7.11	494
$11-13$	$4'44''$	$5'26''$	$5'01''$	$00:25:07$	5.00	347
$11-12$	$3'55''$	$5'53''$	$4'54''$	$00:34:53$	7.11	493
$11-11$	$4'57''$	$6'32''$	$5'34''$	$01:01:22$	11.01	764
$11-06$	$4'49''$	$5'45''$	$5'12''$	$00:43:09$	8.28	574
$11-04$	$5'00''$	$5'52''$	$5'19''$	$00:45:35$	8.54	593

第三章　回归试验设计

本节教学目的

1. 掌握线性模型的参数估计;
2. 掌握线性模型的假设检验;
3. 理解病态性的概念, 以及它与效应混杂的关系;
4. 从回归分析的角度理解方差分析.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

(1) 线性模型

(2) 回归系数的估计

(3) 估计量的性质

(4) 方差参数的估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

- 设有 n 次观测: $y_i = \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \cdots, n.$
- 记 $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^{\mathrm{T}}$,

$$\boldsymbol{X} = \begin{bmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots & f_m(\boldsymbol{x}_1) \\ f_1(\boldsymbol{x}_2) & f_2(\boldsymbol{x}_2) & \cdots & f_m(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(\boldsymbol{x}_n) & f_2(\boldsymbol{x}_n) & \cdots & f_m(\boldsymbol{x}_n) \end{bmatrix}.$$

则线性模型可写成矩阵形式: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$

- 称 $\boldsymbol{X}$ 为广义设计矩阵, 它由设计矩阵决定.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

(1) 线性模型

(2) 回归系数的估计

(3) 估计量的性质

(4) 方差参数的估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

- 假定 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{C})$, 则 $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{C})$ 的密度函数为:

$$\frac{1}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{C})^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right\}.$$

- $\boldsymbol{\beta}$ 的极大似然估计为

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{\beta}} \left\{(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right\}.$$

- $Q(\boldsymbol{\beta}) := (\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})$ 表示加权残差平方和, 因此 $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ 也是参数 $\boldsymbol{\beta}$ 的加权最小二乘估计.

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = \hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{y}.$$

注 在 $\boldsymbol{C}$ 已知的情况下, 可以将 $\boldsymbol{C}$ 变换成单位矩阵.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

(1) 线性模型

(2) 回归系数的估计

(3) 估计量的性质

(4) 方差参数的估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

定义

β 的估计 $\hat{\beta}$ 的均方误差 (mean square error) 定义为

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) := \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right].$$

- 对于一维参数而言,

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{Var}(\hat{\beta}) + \mathrm{bias}^2(\hat{\beta}),$$

即均方误差等于偏差的平方与方差的和.

定理 (最小二乘估计的性质)

线性模型中参数 β 的加权最小二乘估计 $\hat{\beta}$ 具有如下性质:

(1) $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X})^{-1}\right)$;

(2) 设 $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ 为矩阵 $\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}$ 的特征值, 则 $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \sum_{k=1}^{m} \lambda_k^{-1}$.

推论:

(1) 对任意矩阵 $\boldsymbol{A}_{k\times m}$, $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{A}^{\mathrm{T}}\right)$;

(2) $\forall \boldsymbol{c} \in \mathbb{R}^m$, $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{c}\right)$;

(3) 任意点 $\boldsymbol{x} \in \mathcal{X}$ 处响应值的预测:

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x})\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\beta}, \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x})(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{f}(\boldsymbol{x})\right).$$

- 响应变量的回归值为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

称 $\boldsymbol{H}$ 为帽子矩阵, 它是幂等矩阵, $\boldsymbol{H}^2 = \boldsymbol{H}$, $\mathrm{rank}(\boldsymbol{H}) = m$;

- 称 $e := y - \hat{y} = (I - H)y$ 为残差向量,

$$\hat{y}^{\mathrm{T}} e = 0, \quad e^{\mathrm{T}} e + \hat{y}^{\mathrm{T}}\hat{y} = y^{\mathrm{T}} y$$

可将 e、$\hat{y}$ 和 y 理解为直角三角形的三条边;

- 由于 $X^{\mathrm{T}} C^{-1} e = 0$, 当 $C = \sigma^2 I$, 且 $f(x)$ 中有常数 1 (对应回归方程中有截距项) 时, $\sum_{i=1}^{n} e_i = 0$.

- 响应变量的回归值为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

称 $\boldsymbol{H}$ 为帽子矩阵, 它是幂等矩阵, $\boldsymbol{H}^2 = \boldsymbol{H}$, $\mathrm{rank}(\boldsymbol{H}) = m$;

- 称 $\boldsymbol{e} := \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ 为残差向量,

$$\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{e} = 0, \quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$

可将 $\boldsymbol{e}$、$\hat{\boldsymbol{y}}$ 和 $\boldsymbol{y}$ 理解为直角三角形的三条边;

- 由于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{e} = 0$, 当 $\boldsymbol{C} = \sigma^2\boldsymbol{I}$, 且 $f(x)$ 中有常数 1 (对应回归方程中有截距项) 时, $\sum_{i=1}^{n} e_i = 0$.

- 响应变量的回归值为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

称 $\boldsymbol{H}$ 为帽子矩阵, 它是幂等矩阵, $\boldsymbol{H}^2 = \boldsymbol{H}$, $\mathrm{rank}(\boldsymbol{H}) = m$;

- 称 $\boldsymbol{e} := \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ 为残差向量,

$$\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{e} = 0, \quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$

可将 $\boldsymbol{e}$、$\hat{\boldsymbol{y}}$ 和 $\boldsymbol{y}$ 理解为直角三角形的三条边;

- 由于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{e} = \boldsymbol{0}$, 当 $\boldsymbol{C} = \sigma^2\boldsymbol{I}$, 且 $\boldsymbol{f}(\boldsymbol{x})$ 中有常数 1 (对应回归方程中有截距项) 时, $\sum_{i=1}^{n} e_i = 0$.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

(1) 线性模型

(2) 回归系数的估计

(3) 估计量的性质

(4) 方差参数的估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

- 假定 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C})$, 相关矩阵 $\boldsymbol{C}$ 已知, 方差 σ^2 未知, 则随机向量 $\boldsymbol{y}$ 的密度函数为

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \det(\boldsymbol{C})^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}.$$

- 参数 $(\boldsymbol{\beta}, \sigma^2)$ 的对数似然函数为:

$$\ell(\boldsymbol{\beta}, \sigma^2) \propto -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).$$

- 假定 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C})$, 相关矩阵 $\boldsymbol{C}$ 已知, 方差 σ^2 未知, 则随机向量 $\boldsymbol{y}$ 的密度函数为

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \det(\boldsymbol{C})^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}.$$

- 参数 $(\boldsymbol{\beta}, \sigma^2)$ 的对数似然函数为:

$$\ell(\boldsymbol{\beta}, \sigma^2) \propto -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).$$

- 求导, 并令导数等于 0

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \\ \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\dfrac{1}{2\sigma^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}, \end{cases}$$

- 求解得到参数的极大似然估计为

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{y}, \\ \widehat{\sigma}^2_{\mathrm{ML}} = \dfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{cases}$$

- $\hat{\sigma}^2_{\mathrm{ML}}$ 是有偏的, 即 $\mathbb{E}\left(\hat{\sigma}^2_{\mathrm{ML}}\right) \neq \sigma^2$!

- 求导, 并令导数等于 0

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \\ \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\dfrac{1}{2\sigma^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}, \end{cases}$$

- 求解得到参数的极大似然估计为

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{y}, \\ \widehat{\sigma}^2_{\mathrm{ML}} = \dfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{cases}$$

- $\hat{\sigma}^2_{\mathrm{ML}}$ 是有偏的, 即 $\mathbb{E}\left(\hat{\sigma}^2_{\mathrm{ML}}\right) \neq \sigma^2$!

- 求导, 并令导数等于 0

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \\ \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\dfrac{1}{2\sigma^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}, \end{cases}$$

- 求解得到参数的极大似然估计为

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{y}, \\ \widehat{\sigma}^2_{\mathrm{ML}} = \dfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{cases}$$

- $\hat{\sigma}^2_{\mathrm{ML}}$ 是有偏的, 即 $\mathbb{E}\left(\hat{\sigma}^2_{\mathrm{ML}}\right) \neq \sigma^2$!

定理 (方差的无偏估计)

定义加权残差平方和

$$\mathrm{RSS} := (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

则

- $\mathrm{RSS}/\sigma^2 \sim \chi^2(n-m)$ 且与 $\hat{\boldsymbol{\beta}}$ 独立;
- $\widehat{\sigma^2} = \mathrm{RSS}/(n-m)$ 是 σ^2 的无偏估计.

定理 (引理 1.5)

设 $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{A}$ 为 n 阶对称矩阵. 则

(1) $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\xi} \sim \chi^2 (\mathrm{rank}(\boldsymbol{A}))$ 当且仅当 $\boldsymbol{A}$ 为幂等矩阵, 即 $\boldsymbol{A}^2 = \boldsymbol{A}$;

(2) 设 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$ 均为对称非负定的幂等矩阵, 则 $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A}_1 \boldsymbol{\xi}$ 和 $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A}_2 \boldsymbol{\xi}$ 互相独立当且仅当 $\boldsymbol{A}_1 \boldsymbol{A}_2 = \mathbf{0}$;

(3) 如果 $\boldsymbol{A}$ 可分解为 k 个对称矩阵 $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_k$ 的和, 则下列命题等价:

  (a) $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A}_1 \boldsymbol{\xi}, \cdots, \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{\xi}$ 互相独立, 且 $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A}_i \boldsymbol{\xi} \sim \chi^2 (\mathrm{rank}(\boldsymbol{A}_i))$ 对 $i = 1, \cdots, k$ 都成立;

  (b) $\boldsymbol{A}$ 为幂等矩阵, 且 $\mathrm{rank}(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A}_1) + \cdots + \mathrm{rank}(\boldsymbol{A}_k)$.

(4) 设 $\boldsymbol{B}$ 为 $m \times n$ 阶矩阵, $\boldsymbol{A}$ 为幂等矩阵, $\boldsymbol{B}\boldsymbol{A} = \mathbf{0}$, 则线性型 $\boldsymbol{B}\boldsymbol{\xi}$ 与二次型 $\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\xi}$ 互相独立.

证明.

注意到

$$\mathrm{RSS} = \boldsymbol{\varepsilon}^{\mathrm{T}} \left[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \right] \boldsymbol{\varepsilon},$$

令 $\tilde{\boldsymbol{\varepsilon}} = \sigma^{-1} \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{\varepsilon}$, 则 $\tilde{\boldsymbol{\varepsilon}}$ 为 n 维标准 Gauss 随机向量, 且

$$\mathrm{RSS} = \sigma^2 \tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}} \left[\boldsymbol{I} - \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-\frac{1}{2}} \right] \tilde{\boldsymbol{\varepsilon}},$$

故 $\mathrm{RSS}/\sigma^2 \sim \chi^2 \left(\mathrm{tr}(\boldsymbol{I} - \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-\frac{1}{2}}) \right)$. 独立性由下述等式保证:

$$\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-\frac{1}{2}} \left[\boldsymbol{I} - \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-\frac{1}{2}} \right] = \boldsymbol{0}.$$

例 (单因子固定效应模型的参数估计)

回忆单因子试验固定效应模型,

(1) 利用线性模型参数估计的理论给出诸效应的估计;

(2) 回忆正交对照的概念, 并利用线性模型参数估计的理论给出对照的估计.

小结

- 线性模型 $\boldsymbol{y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C}\right)$ 的参数估计为

$$\begin{cases} \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{y}, \\ \widehat{\sigma^2} = \dfrac{1}{n-m} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{cases}$$

- $\hat{\boldsymbol{\beta}}$ 服从正态分布, 且与 $\hat{\sigma^2}$ 独立;
- $\frac{1}{\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \sim \chi^2(n-m)$;
- 矩阵 $\boldsymbol{X}^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{X}$ 非常重要!

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

(1) 回归模型的显著性检验

(2) 回归系数的逐个检验

(3) 回归系数的分组检验

3.1.3 病态性及其处理方法

模型假定

(1) 向量函数 $\boldsymbol{f}(\boldsymbol{x})$ 含有常数分量 1, 即 $f_1(\boldsymbol{x}) \equiv 1$, 可理解为线性模型中有截距项;

(2) 假定 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, 对于 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{C})$ 的情形, 令

$$\boldsymbol{y}' = \boldsymbol{C}^{-1/2}\boldsymbol{y}, \quad \boldsymbol{X}' = \boldsymbol{C}^{-1/2}\boldsymbol{X},$$

则 $\boldsymbol{y}' \sim N(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

模型假定

(1) 向量函数 $\boldsymbol{f}(\boldsymbol{x})$ 含有常数分量 1, 即 $f_1(\boldsymbol{x}) \equiv 1$, 可理解为线性模型中有截距项;

(2) 假定 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, 对于 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C})$ 的情形, 令

$$\boldsymbol{y}' = \boldsymbol{C}^{-1/2}\boldsymbol{y}, \quad \boldsymbol{X}' = \boldsymbol{C}^{-1/2}\boldsymbol{X},$$

则 $\boldsymbol{y}' \sim N(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$.

- 从整体上检验自变量对因变量是否有显著影响:

$$H_0 : \boldsymbol{\beta}_{-1} = \mathbf{0}, \quad \text{v.s.} \quad H_1 : \boldsymbol{\beta}_{-1} \neq \mathbf{0}.$$

$\boldsymbol{\beta}_{-1}$ 表示去除了截距项后的回归系数向量.

- 检验方法是作方差分析

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

- $SS_E := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 为恰为残差平方和;
- 称 $SS_R := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 为回归平方和.

- 从整体上检验自变量对因变量是否有显著影响:

$$H_0 : \boldsymbol{\beta}_{-1} = \mathbf{0}, \quad \text{v.s.} \quad H_1 : \boldsymbol{\beta}_{-1} \neq \mathbf{0}.$$

$\boldsymbol{\beta}_{-1}$ 表示去除了截距项后的回归系数向量.

- 检验方法是作方差分析

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

- $SS_E := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 为恰为残差平方和;
- 称 $SS_R := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 为回归平方和.

- 从整体上检验自变量对因变量是否有显著影响:

$$H_0 : \boldsymbol{\beta}_{-1} = \mathbf{0}, \quad \text{v.s.} \quad H_1 : \boldsymbol{\beta}_{-1} \neq \mathbf{0}.$$

$\boldsymbol{\beta}_{-1}$ 表示去除了截距项后的回归系数向量.

- 检验方法是作方差分析

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

- $SS_E := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 为恰为残差平方和;
- 称 $SS_R := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 为回归平方和.

定理 (回归模型的显著性检验)

如果回归函数中 $f_1(\boldsymbol{x}) \equiv 1$, 则当 H_0 为真时,

$$F := \frac{SS_R/(m-1)}{SS_E/(n-m)} \sim F(m-1, n-m).$$

特别地, 对于给定的显著性水平 α, 检验问题

$$H_0 : \boldsymbol{\beta}_{-1} = \mathbf{0}, \qquad H_1 : \boldsymbol{\beta}_{-1} \neq \mathbf{0}$$

的拒绝域为 $F > F_{1-\alpha}(m-1, n-m)$.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

(1) 回归模型的显著性检验

(2) 回归系数的逐个检验

(3) 回归系数的分组检验

3.1.3 病态性及其处理方法

- 意义: 剔除不重要的分量, 简化模型

- 检验 $f(\boldsymbol{x})$ 的第 $j>1$ 个分量 $f_j(\boldsymbol{x})$ 是否显著:

$$H_{0j}: \beta_j = 0, \quad \text{v.s.} \quad H_{1j}: \beta_j \neq 0.$$

- 因 $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{M_{jj}}} \sim N(0,1)$$

M_{jj} 表示矩阵 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 的第 j 个主对角元.

- 意义: 剔除不重要的分量, 简化模型
- 检验 $\boldsymbol{f}(\boldsymbol{x})$ 的第 $j > 1$ 个分量 $f_j(\boldsymbol{x})$ 是否显著:

$$H_{0j} : \beta_j = 0, \quad \text{v.s.} \quad H_{1j} : \beta_j \neq 0.$$

- 因 $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{M_{jj}}} \sim N(0, 1)$$

M_{jj} 表示矩阵 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 的第 j 个主对角元.

- 意义: 剔除不重要的分量, 简化模型
- 检验 $\boldsymbol{f}(\boldsymbol{x})$ 的第 $j>1$ 个分量 $f_j(\boldsymbol{x})$ 是否显著:

$$H_{0j}:\beta_j=0, \quad \text{v.s.} \quad H_{1j}:\beta_j\neq 0.$$

- 因 $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$,

$$\frac{\hat{\beta}_j-\beta_j}{\sigma\sqrt{M_{jj}}}\sim N(0,1)$$

M_{jj} 表示矩阵 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 的第 j 个主对角元.

- 根据 t 分布的定义, 给定检验水平 α, 当

$$|t_j| = \left| \frac{\hat{\beta}_j}{\sqrt{M_{jj}\mathrm{RSS}/(n-m)}} \right| \geq t_{\alpha/2}(n-m)$$

时, 拒绝原假设 H_0, 认为 β_j 显著异于 0; 否则接受 H_0, 认为 $f_j(\boldsymbol{x})$ 对 y 无显著影响.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

(1) 回归模型的显著性检验

(2) 回归系数的逐个检验

(3) 回归系数的分组检验

3.1.3 病态性及其处理方法

回归系数的分组检验问题

设 $1 \leq r \leq m-1,\ k+r=m,$

$$\boldsymbol{X}=[\boldsymbol{X}_k, \boldsymbol{X}_r], \quad \boldsymbol{\beta}=[\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{\beta}_r^{\mathrm{T}}]^{\mathrm{T}}.$$

考虑检验问题

$$H_0: \boldsymbol{\beta}_r=\mathbf{0}, \quad \text{v.s.} \quad H_1: \boldsymbol{\beta}_r \neq \mathbf{0}.$$

- 称 $\boldsymbol{y}=\boldsymbol{X}_k\boldsymbol{\beta}_k+\boldsymbol{X}_r\boldsymbol{\beta}_r+\boldsymbol{\varepsilon}$ 为全模型, 它的参数估计

$$\begin{cases}\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y},\\ \mathrm{RSS}=\boldsymbol{y}^{\mathrm{T}}\left[\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{y}.\end{cases}$$

- 称 $\boldsymbol{y}=\boldsymbol{X}_k\boldsymbol{\beta}_k+\boldsymbol{\varepsilon}$ 为选模型, 它的参数估计为

$$\begin{cases}\widetilde{\boldsymbol{\beta}}_k=(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{y},\\ \mathrm{RSS}_{H_0}=\boldsymbol{y}^{\mathrm{T}}\left[\boldsymbol{I}-\boldsymbol{X}_k(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\right]\boldsymbol{y}.\end{cases}$$

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\sigma^2} \sim \chi^2(r), \quad \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/r}{\mathrm{RSS}/(n-m)} \sim F(r, n-m).$$

- 取 $r = 1$, 可以依次检验各变量的显著性;
- 取 $r = m - 1$, 与前面回归模型显著性检验定理一致;
- 可推广到检验线性假设

$$H_0 : \boldsymbol{G\beta} = \mathbf{0} \quad \text{v.s.} \quad H_1 : \boldsymbol{G\beta} \neq \mathbf{0},$$

其中 $\boldsymbol{G}$ 为 $r \times m$ 行满秩矩阵, $r \leq m$.

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\sigma^2} \sim \chi^2(r), \quad \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/r}{\mathrm{RSS}/(n-m)} \sim F(r, n-m).$$

- 取 $r = 1$, 可以依次检验各变量的显著性;
- 取 $r = m - 1$, 与前面回归模型显著性检验定理一致;
- 可推广到检验线性假设

$$H_0 : \boldsymbol{G\beta} = \mathbf{0} \quad \text{v.s.} \quad H_1 : \boldsymbol{G\beta} \neq \mathbf{0},$$

其中 $\boldsymbol{G}$ 为 $r \times m$ 行满秩矩阵, $r \leq m$.

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\sigma^2} \sim \chi^2(r), \quad \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/r}{\mathrm{RSS}/(n-m)} \sim F(r, n-m).$$

- 取 $r = 1$, 可以依次检验各变量的显著性;
- 取 $r = m - 1$, 与前面回归模型显著性检验定理一致;
- 可推广到检验线性假设

$$H_0 : \boldsymbol{G\beta} = \mathbf{0} \quad \text{v.s.} \quad H_1 : \boldsymbol{G\beta} \neq \mathbf{0},$$

其中 $\boldsymbol{G}$ 为 $r \times m$ 行满秩矩阵, $r \leq m$.

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\sigma^2} \sim \chi^2(r), \quad \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/r}{\mathrm{RSS}/(n-m)} \sim F(r, n-m).$$

- 取 $r = 1$, 可以依次检验各变量的显著性;
- 取 $r = m - 1$, 与前面回归模型显著性检验定理一致;
- 可推广到检验线性假设

$$H_0 : \boldsymbol{G\beta} = \mathbf{0} \quad \text{v.s.} \quad H_1 : \boldsymbol{G\beta} \neq \mathbf{0},$$

其中 $\boldsymbol{G}$ 为 $r \times m$ 行满秩矩阵, $r \leq m$.

例 (单因子固定效应模型的参数估计)

回忆单因子试验固定效应模型,

(1) 利用线性模型假设检验的理论给出诸效应的显著性检验;

(2) 回忆正交对照的概念, 并利用线性模型假设检验理论给出对照的显著性检验.

例 (国家财政收入)

影响一个国家或地区财政收入的因素包括国内生产总值、财政支出、商品零售价指数等. 选择包括中央和地方税收的“国家财政收入”中的“各项税收”作为响应变量, 下表是来源于《中国统计年鉴》1978-2011 年有关财政收入 y、国内生产总值 x_1、财政支出 x_2、商品零售物价指数 x_3 的数据. 其中变量 y, x_1, x_2 的单位都是亿元人民币.

年份	y	x_1	x_2	x_3	年份	y	x_1	x_2
1978	519.28	3645.2	1122.09	100.7	1995	6038.04	60793.7	6823.72
1979	537.82	4062.6	1281.79	102.0	1996	6909.82	71176.6	7937.55
1980	571.70	4545.6	1228.83	106.0	1997	8234.04	78973.0	9233.56
1981	629.89	4891.6	1138.41	102.4	1998	9262.80	84402.3	10798.18
1982	700.02	5323.4	1229.98	101.9	1999	10682.58	89677.1	13187.67
1983	775.59	5962.7	1409.52	101.5	2000	12581.51	99214.6	15886.50
1984	947.35	7208.1	1701.02	102.8	2001	15301.38	109655.2	18902.58
1985	2040.79	9016.0	2004.25	108.8	2002	17636.45	120332.7	22053.15
1986	2090.73	10275.2	2204.91	106.0	2003	20017.31	135822.8	24649.95
1987	2140.36	12058.6	2262.18	107.3	2004	24165.68	159878.3	28486.89
1988	2390.47	15042.8	2491.21	118.5	2005	28778.54	184937.4	33930.28
1989	2727.40	16992.3	2823.78	117.8	2006	34804.35	216314.4	40422.73
1990	2821.86	18667.8	3083.59	102.1	2007	45621.97	265810.3	49781.35
1991	2990.17	21781.5	3386.62	102.9	2008	54223.79	314045.4	62592.66
1992	3296.91	26923.5	3742.20	105.4	2009	59521.59	340902.8	76299.93
1993	4255.30	35333.9	4642.30	113.2	2010	73210.79	401512.8	89874.16
1994	5126.88	48197.9	5792.62	121.7	2011	89738.39	472881.6	109247.79

例 (国家财政收入)

用 R 建立响应变量 y 关于 $\boldsymbol{f}(x_1, x_2, x_3) = [1, x_1, x_2, x_3]^{\mathrm{T}}$ 的线性回归模型, 操作步骤如下.

(1) 读入数据. 将表中的 Excel 类型的数据保存为文本文档 data1.txt, 然后使用如下代码读入数据.

```
yx <- read.table("data1.txt", header = T);
y  <- yx[,2];
x1 <- yx[,3];
x2 <- yx[,4];
x3 <- yx[,5];
```

(2) 拟合模型并展示结果. 使用代码

```
fm <- lm(y ~ x1 + x2 + x3, data = yx);
summary(fm)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = yx)

Residuals:
    Min      1Q  Median      3Q     Max
-2928.0 -637.3   87.6   422.4 3082.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.936e+03 3.163e+03 -1.560 0.129187
x1           4.298e-02 1.092e-02 3.934 0.000457 ***
x2           6.336e-01 4.915e-02 12.891 9.12e-14 ***
x3           4.257e+01 2.962e+01 1.437 0.160996
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
    0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1004 on 30 degrees of
    freedom
Multiple R-squared: 0.9982,^^IAdjusted R-squared:
    0.9981
F-statistic: 5690 on 3 and 30 DF, p-value: < 2.2e-16
```

例 (国家财政收入)

拟合的线性回归方程为:

$$\hat{y} = -4936 + 0.04298x_1 + 0.6336x_2 + 42.57x_3.$$

- 从 t 检验来看, x_1 和 x_2 对响应变量的影响是显著的, 而 x_3 对响应变量的影响是不显著的.
- 从 F 检验来看, 可以认为所建立的回归方程显著有效.
- Multiple R-Squared 称为复相关系数, 其定义为

$$R^2 := \frac{SS_R}{SS_T} = 1 - \frac{\mathrm{RSS}}{SS_T},$$

其值越接近于 1, 自变量的解释程度越高.

例 (国家财政收入)

自变量个数越多, 残差就越小. 当自变量个数与观测数据个数相同时, 残差平方和可以为 0, $R^2 = 1$. 因此, 复相关系数仅表达了模型的拟合程度, 这未必合适. 为消除自变量个数的影响, 统计学家提出调整的相关系数(Adjusted R-squared) 的概念, 其定义为

$$R_A^2 := 1 - \frac{\mathrm{RSS}/(n-m)}{SS_T/(n-1)},$$

注意, 计算自变量个数 m 时, 应当把截距项也算进来, 即本例中 $m = 3 + 1 = 4$. 本例中, $R_A^2 = 0.9981$, 表明回归方程拟合较好.

课堂小结

- 回归模型的显著性检验: 方差分析;
- 回归系数的逐个检验: t 检验;
- 回归系数的分组检验: F 检验;
- 前两个检验都可以作为后一个检验的特例.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

(1) 病态性及其后果

(2) 病态性的诊断

(3) 病态性的处理方法

- 当矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 存在小的特征值, 即病态时, 最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的均方误差可能会变得很大.
- 造成矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 病态的原因有两个:
  (1) $n < m$, 这类问题在高维数据中常常出现, 近 20 年来已成为统计学家研究的热点问题之一;
  (2) 矩阵 $\boldsymbol{X}$ 的复共线性, 即存在不全为 0 的常数 $c_1, \cdots, c_m$, 使得

$$\|c_1\boldsymbol{X}_1 + \cdots + c_m\boldsymbol{X}_m\|_2 \approx 0,$$

  其中 $\boldsymbol{X}_j$ 表示广义设计矩阵 $\boldsymbol{X}$ 的第 j 列.
- $\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \approx 0$ 可以作为复共线性的诊断判据.

- 当矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 存在小的特征值, 即病态时, 最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的均方误差可能会变得很大.
- 造成矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 病态的原因有两个:
  (1) $n < m$, 这类问题在高维数据中常常出现, 近 20 年来已成为统计学家研究的热点问题之一;
  (2) 矩阵 $\boldsymbol{X}$ 的复共线性, 即存在不全为 0 的常数 $c_1, \cdots, c_m$, 使得

$$\|c_1\boldsymbol{X}_1 + \cdots + c_m\boldsymbol{X}_m\|_2 \approx 0,$$

  其中 $\boldsymbol{X}_j$ 表示广义设计矩阵 $\boldsymbol{X}$ 的第 j 列.
- $\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \approx 0$ 可以作为复共线性的诊断判据.

- 当矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 存在小的特征值, 即病态时, 最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的均方误差可能会变得很大.
- 造成矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 病态的原因有两个:
  (1) $n < m$, 这类问题在高维数据中常常出现, 近 20 年来已成为统计学家研究的热点问题之一;
  (2) 矩阵 $\boldsymbol{X}$ 的复共线性, 即存在不全为 0 的常数 $c_1, \cdots, c_m$, 使得

$$\|c_1\boldsymbol{X}_1 + \cdots + c_m\boldsymbol{X}_m\|_2 \approx 0,$$

  其中 $\boldsymbol{X}_j$ 表示广义设计矩阵 $\boldsymbol{X}$ 的第 j 列.
- $\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \approx 0$ 可以作为复共线性的诊断判据.

- 当矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 存在小的特征值, 即病态时, 最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的均方误差可能会变得很大.
- 造成矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 病态的原因有两个:
  (1) $n < m$, 这类问题在高维数据中常常出现, 近 20 年来已成为统计学家研究的热点问题之一;
  (2) 矩阵 $\boldsymbol{X}$ 的复共线性, 即存在不全为 0 的常数 $c_1, \cdots, c_m$, 使得

$$\|c_1\boldsymbol{X}_1 + \cdots + c_m\boldsymbol{X}_m\|_2 \approx 0,$$

  其中 $\boldsymbol{X}_j$ 表示广义设计矩阵 $\boldsymbol{X}$ 的第 j 列.
- $\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \approx 0$ 可以作为复共线性的诊断判据.

例

假设 x_1, x_2 与 y 之间存在线性关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

其中 $\beta_0 = 10$, $\beta_1 = 2$, $\beta_2 = 3$. 考虑如下设计矩阵

$$\boldsymbol{X} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.1 & 1.4 & 1.7 & 1.7 & 1.8 & 1.8 & 1.9 & 2.0 & 2.3 & 2.4 \\ 1.1 & 1.5 & 1.8 & 1.7 & 1.9 & 1.8 & 1.8 & 2.1 & 2.4 & 2.5 \end{bmatrix}^{\mathrm{T}}$$

- 用随机模拟产生正态分布的随机误差的 10 次观测值

$$\boldsymbol{\varepsilon} = [0.8, -0.5, 0.4, -.05, 0.2, 1.9, 1.9, 0.6, -1.5, -0.5]^{\mathrm{T}}.$$

由模型 $y = 10 + 2x_1 + 3x_2 + \varepsilon$ 得到 y 的 10 次观测值:

$$\boldsymbol{y} = [16.3, 16.8, 19.2, 18.0, 19.5, 20.9, 21.1, 20.9, 20.3, 22.0]^{\mathrm{T}}.$$

- 利用最小二乘法得到 β 的估计为

$$\hat{\boldsymbol{\beta}} = [11.292, 11.307, -6.591]^{\mathrm{T}}.$$

- 由于因果关系的复杂性, 有时这种复共线性是无法避免的.

- 用随机模拟产生正态分布的随机误差的 10 次观测值

$$\boldsymbol{\varepsilon} = [0.8, -0.5, 0.4, -.05, 0.2, 1.9, 1.9, 0.6, -1.5, -0.5]^{\mathrm{T}}.$$

由模型 $y = 10 + 2x_1 + 3x_2 + \varepsilon$ 得到 y 的 10 次观测值:

$$\boldsymbol{y} = [16.3, 16.8, 19.2, 18.0, 19.5, 20.9, 21.1, 20.9, 20.3, 22.0]^{\mathrm{T}}.$$

- 利用最小二乘法得到 $\boldsymbol{\beta}$ 的估计为

$$\hat{\boldsymbol{\beta}} = [11.292, 11.307, -6.591]^{\mathrm{T}}.$$

- 由于因果关系的复杂性, 有时这种复共线性是无法避免的.

- 用随机模拟产生正态分布的随机误差的 10 次观测值

$$\boldsymbol{\varepsilon} = [0.8, -0.5, 0.4, -.05, 0.2, 1.9, 1.9, 0.6, -1.5, -0.5]^{\mathrm{T}}.$$

由模型 $y = 10 + 2x_1 + 3x_2 + \varepsilon$ 得到 y 的 10 次观测值:

$$\boldsymbol{y} = [16.3, 16.8, 19.2, 18.0, 19.5, 20.9, 21.1, 20.9, 20.3, 22.0]^{\mathrm{T}}.$$

- 利用最小二乘法得到 $\boldsymbol{\beta}$ 的估计为

$$\hat{\boldsymbol{\beta}} = [11.292, 11.307, -6.591]^{\mathrm{T}}.$$

- 由于因果关系的复杂性, 有时这种复共线性是无法避免的.

当线性回归模型存在复共线性时，若仍然采用普通的最小二乘法估计模型参数，会产生如下不良后果:

(1) 完全复共线性下模型参数的最小二乘估计不存在.

(2) $\hat{\boldsymbol{\beta}}$ 的方差矩阵 $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ 的对角元会变得很大.

(3) 复共线性下自变量的显著性检验没有意义.

(4) 模型的预测变得不可靠.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

(1) 病态性及其后果

(2) 病态性的诊断

(3) 病态性的处理方法

例 (猪肉价格)

下表给出了我国 1991-2006 年猪肉价格及其影响因素数据:

- y 表示猪肉价格, 单位为元/千克;
- x_1 表示消费价格指数, 即 CPI;
- x_2 表示人口数量, 单位为亿;
- x_3 表示年末存栏量, 单位为万头;
- x_4 表示城镇居民可支配收入, 单位为元;
- x_5 表示玉米价格, 单位为元/吨;
- x_6 表示猪肉生产量, 单位为万吨.

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1990	9.84	103.1	14.39	36241	1510.2	686.7	2281
1991	10.32	103.4	12.98	36965	1700.6	590	2452
1992	10.65	106.4	11.60	38421	2026.6	625	2635
1993	10.49	114.7	11.45	39300	2577.4	726.7	2854
1994	9.16	124.1	11.21	41462	3496.2	1004.2	3205
1995	10.18	117.1	10.55	44169	4283	1576.7	3648
1996	14.96	107.9	10.42	36284	4838.9	1481.7	3158
1997	11.81	102.8	10.06	40035	5160.3	1150.8	3596
1998	10.77	99.2	9.14	42256	5425.1	1269.2	3884
1999	8.38	98.6	8.18	43020	5854	1092.5	3891
2000	8.74	100.4	7.58	44682	6280	887.5	4031
2001	10.18	100.7	6.95	45743	6859.6	1060	4184
2002	9.85	99.2	6.45	46292	7702.8	1033.3	4327
2003	10.7	101.2	6.01	46602	8472.2	1087.5	4519
2004	13.97	103.9	5.87	48189	9421.6	1288.3	4702
2005	13.39	101.8	5.89	50335	10493	1229.2	5011
2006	14.03	101.5	5.28	49441	13172	1280	5197

例 (猪肉价格)

记 $y^* = \log y,\ z_1 = \log x_1, \cdots, z_6 = \log x_6$. 建立线性回归模型:

$$y^* = \beta_0 + \sum_{j=1}^{6} \beta_j z_j + \varepsilon.$$

将表中数据存储为 txt 格式, 命名为 “porkpricedata.txt” . 在 R 中利用代码块

```
porkpricedata <- read.table("porkpricedata.txt");
y  <- log(porkpricedata[,2]);
z1 <- log(porkpricedata[,3]);
z2 <- log(porkpricedata[,4]);
z3 <- log(porkpricedata[,5]);
z4 <- log(porkpricedata[,6]);
z5 <- log(porkpricedata[,7]);
z6 <- log(porkpricedata[,8]);
ppd_result <- lm(y ~ z1 + z2 + z3 + z4 + z5 +z6);
summary(ppd_result)
```

```
Call:
lm(formula = y ~ z1 + z2 + z3 + z4 + z5 + z6)

Residuals:
      Min        1Q    Median        3Q       Max
-0.240615 -0.094805  0.003942  0.100881  0.201397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.03498 17.44684  0.862   0.409
z1           0.26583   1.02862  0.258   0.801
z2          -0.89505   1.00900 -0.887   0.396
z3           0.08584   3.12736  0.027   0.979
z4           0.46123   1.08240  0.426   0.679
z5           0.42119   0.36391  1.157   0.274
z6          -2.40388   3.08755 -0.779   0.454

Residual standard error: 0.156 on 10 degrees of
   freedom
Multiple R-squared: 0.4684,^^IAdjusted R-squared:
   0.1494
F-statistic: 1.468 on 6 and 10 DF, p-value: 0.2817
```

- 部分参数的估计标准误差极大, 如 $\hat{\beta}_3 = 0.08584$, 但其标准误差为 3.12736.
- 这可能是由于变量之间存在复共线性造成的, 需要对是否存在复共线性进行诊断.
- 方差膨胀因子法是用于诊断回归方程是否存在复共线性的常用方法.

- 部分参数的估计标准误差极大, 如 $\hat{\beta}_3 = 0.08584$, 但其标准误差为 3.12736.
- 这可能是由于变量之间存在复共线性造成的, 需要对是否存在复共线性进行诊断.
- 方差膨胀因子法是用于诊断回归方程是否存在复共线性的常用方法.

- 部分参数的估计标准误差极大, 如 $\hat{\beta}_3 = 0.08584$, 但其标准误差为 3.12736.
- 这可能是由于变量之间存在复共线性造成的, 需要对是否存在复共线性进行诊断.
- 方差膨胀因子法是用于诊断回归方程是否存在复共线性的常用方法.

- 首先对广义设计矩阵 $\boldsymbol{X}$ 作中心标准化处理, 即

$$X_{ij}^{*}=\frac{X_{ij}-\frac{X_{\cdot j}}{n}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij}-\frac{X_{\cdot j}}{n}\right)^{2}}},$$

  得到中心标准化后的设计矩阵 $\boldsymbol{X}^{*}$.
- 中心标准化可借助添加包 `MASS` 中的函数 `scale()` 来实现.
- $(\boldsymbol{X}^{*})^{\mathrm{T}}\boldsymbol{X}^{*}$ 为诸各列 $f_j(\boldsymbol{x}_i)$ 之间的相关阵. 记 $\boldsymbol{C}=(c_{ij})=\left[(\boldsymbol{X}^{*})^{\mathrm{T}}\boldsymbol{X}^{*}\right]^{-1}$.
- 自变量 $f_j(\boldsymbol{x})$ 的方差膨胀因子(variance inflation factor) 定义为 $\mathrm{VIF}_j:=c_{jj}$.

- 由于最小二乘估计的方差阵为 $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1}$, 可以验证 $\mathrm{Var}(\hat{\beta}_j) = c_{jj}\sigma^2 / L_{jj}$, 其中

$$L_{jj} = \sum_{i=1}^{n} \left(X_{ij} - \frac{X_{\cdot j}}{n} \right)^2$$

完全由 $\boldsymbol{X}$ 的第 j 列决定. 因而用 c_{jj} 作为度量 $f_j(\boldsymbol{x})$ 的方差膨胀程度的因子是非常合适的.

- 记 R_j^2 为利用其余自变量来拟合第 j 个自变量的复相关系数, 可以证明

$$c_{jj} = \frac{1}{1 - R_j^2},$$

故 VIF_j 越大, 表明第 j 个自变量越容易由其余自变量线性表出, 如果 $\mathrm{VIF}_j \geq 10$, 则可以认为存在复共线性.

- 由于最小二乘估计的方差阵为 $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1}$, 可以验证 $\mathrm{Var}(\hat{\beta}_j) = c_{jj}\sigma^2 / L_{jj}$, 其中

$$L_{jj} = \sum_{i=1}^{n} \left(X_{ij} - \frac{X_{\cdot j}}{n} \right)^2$$

完全由 $\boldsymbol{X}$ 的第 j 列决定. 因而用 c_{jj} 作为度量 $f_j(\boldsymbol{x})$ 的方差膨胀程度的因子是非常合适的.

- 记 R_j^2 为利用其余自变量来拟合第 j 个自变量的复相关系数, 可以证明

$$c_{jj} = \frac{1}{1 - R_j^2},$$

故 VIF_j 越大, 表明第 j 个自变量越容易由其余自变量线性表出, 如果 $\mathrm{VIF}_j \geq 10$, 则可以认为存在复共线性.

例 (猪肉价格)

添加包 DAAG 提供了计算 VIF 的函数, 其代码为:

```
install.packges("DAAG")
library(DAAG)
vif(ppd_result, digits = 3)
```

结果如下:

```
   z1     z2     z3     z4     z5     z6
 2.92  66.30  73.40 315.00   7.47 385.00
```

z_6 的方差膨胀因子高达 385.00, 表明本例自变量之间的复共线性非常强.

3.1 数据的回归分析

3.1.1 线性模型的参数估计

3.1.2 线性模型的假设检验

3.1.3 病态性及其处理方法

(1) 病态性及其后果

(2) 病态性的诊断

(3) 病态性的处理方法

根据病态性产生的原因，从以下几个角度着手处理:

- 从数据着手，增加样本量;
- 从模型着手，剔除一些不重要的自变量，或引入一些先验信息，或干脆改变模型的形式;
- 从参数估计方法着手，如引入有偏估计，通过降低方差来减小均方误差.

尽管这些方法的角度不同，但其数学形式可能一致.

根据病态性产生的原因，从以下几个角度着手处理:

- 从数据着手，增加样本量;
- 从模型着手，剔除一些不重要的自变量，或引入一些先验信息，或干脆改变模型的形式;
- 从参数估计方法着手，如引入有偏估计，通过降低方差来减小均方误差.

尽管这些方法的角度不同，但其数学形式可能一致.

总结

1. 线性模型的参数估计
   - 回归系数的加权最小二乘估计
   - 方差的无偏估计
   - 响应值的预测
2. 线性模型的假设检验
   - 回归模型的显著性检验
   - 回归系数的逐个检验
   - 回归系数的分组检验
3. 病态性: 矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}$ 近似不可逆.