

违背 Gauss-Markov 条件

① 异方差性: Heteroscedasticity

② 自相关性: Autocorrelation

1. 广义最小二乘估计 (Generalized Least Squares Estimation, GLSE)

考虑去除 Gauss-Markov 条件的一般线性模型:

$$y = X\beta + \varepsilon$$

其中 $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ 是满足 $\text{rank}(X) = p < n$ 的固定设计矩阵, $\beta \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$.

只要求随机误差 ε 满足

$$\left. \begin{aligned} E(\varepsilon) &= 0 \\ \text{Cov}(\varepsilon) &= \Sigma \in \mathbb{R}^{n \times n} \text{ 为正定阵} \end{aligned} \right\}$$

GLSE:

若 Σ 已知, 做变换

$$y^* = \Sigma^{-\frac{1}{2}} \cdot y$$

$$X^* = \Sigma^{-\frac{1}{2}} \cdot X$$

$$\varepsilon^* = \Sigma^{-\frac{1}{2}} \cdot \varepsilon$$

则一般线性模型变为:

$$y^* = X^* \beta + \varepsilon^*$$

且满足 Gauss-Markov 条件:

$$E(\varepsilon^*) = 0$$

$$\text{cov}(\varepsilon^*) = I$$

Proof: $\text{cov}(\varepsilon^*) = \text{cov}(\Sigma^{-\frac{1}{2}} \varepsilon) = \Sigma^{-\frac{1}{2}} \cdot \text{cov}(\varepsilon) \cdot (\Sigma^{-\frac{1}{2}})^T = \Sigma^{-\frac{1}{2}} \cdot \Sigma \cdot \Sigma^{-\frac{1}{2}} = I$

于是, 由 OLS 可得: $\hat{\beta}^* = \underset{b \in \mathbb{R}^p}{\text{argmin}} (y - Xb)^T \Sigma^{-1} (y - Xb) = \underset{b \in \mathbb{R}^p}{\text{argmin}} \|y^* - X^*b\|^2$

$$\hat{\beta}^* = (X^{*T} X^*)^{-1} \cdot X^{*T} \cdot y^*$$

$$= \underline{(X^T \Sigma^{-1} X)^{-1} \cdot X^T \Sigma^{-1} \cdot y} \quad \leftarrow \text{GLSE}$$

对于原始一般线性模型 $y = X\beta + \varepsilon$, 我们称 $\hat{\beta}^*$ 为 β 的 GLSE, 显然 $\hat{\beta}^*$ 满足:

$$\mathbb{E}(\hat{\beta}^*) = \beta$$

$$\text{cov}(\hat{\beta}^*) = (X^T \Sigma X)^{-1} \quad \because \text{cov}(y) = \Sigma$$

即如果直接求 OLS $\underline{\hat{\beta} = (X^T X)^{-1} X^T y}$, 则有 $\leftarrow \text{OLSE}$

$$\mathbb{E}(\hat{\beta}) = \beta$$

$$\text{cov}(\hat{\beta}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} \quad \because \text{cov}(y) = \Sigma$$

比较 GLSE 和 直接 OLSE

由 Gauss-Markov 定理, GLSE 的 $\hat{\beta}^*$ 在所有对 β 的线性无偏估计中, 具有最小的协方差 (BLUE), 即有对 β 的任意线性无偏估计 $\tilde{\beta}$, 有

$$[\text{cov}(\tilde{\beta}) - \text{cov}(\hat{\beta}^*)] \text{ 是半正定的}$$

同样, OLSE $\hat{\beta}$ 不如 GLSE $\hat{\beta}^*$

$$[\text{cov}(\hat{\beta}) - \text{cov}(\hat{\beta}^*)] \text{ 是半正定的}$$

若 Σ 未知: ① 异方差性 ② 自相关性

2. 异方差性 Heteroscedasticity

假设随机误差具有不同的方差：

$$\begin{aligned}\text{cov}(\varepsilon) &= \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2) \\ &= \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}\end{aligned}$$

这种数据特征称为 异方差性 Heteroscedasticity

2.1 判断异方差性 Diagnosis of Heteroscedasticity

首先，按照 OLSE 估计参数 β 得到 $\hat{\beta}$ ，其仍然是无偏的，

$$E(\hat{\beta}) = \beta$$

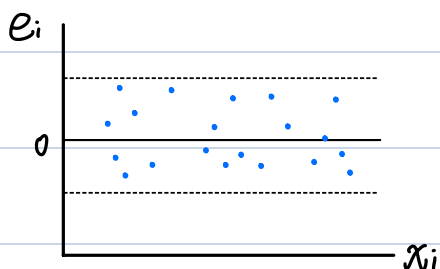
而后计算残差，

$$e_i = y_i - \hat{y}_i = y_i - x_i^T \hat{\beta}$$

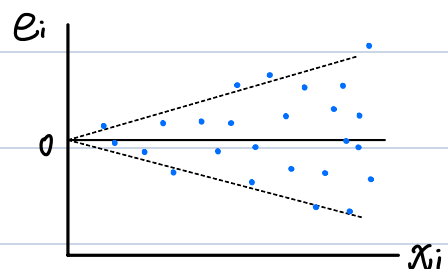
考虑用 e_i 代替 ε_i ，比较 $\{x_i \mid i=1, 2, \dots, n\}$ 与 $\{e_i \mid i=1, 2, \dots, n\}$ 的关联。Tip: 先考虑一元情形， $x_i \in \mathbb{R}$

① Method 1: 残差图

绘制 (x_i, e_i) 图像



e_i 不随 x_i 变化，无异方差性



有异方差性

② Method 2: Spearman 相关系数

(1) $\{x_i | i=1, 2, \dots, n\}$ 与 $\{|e_i| | i=1, 2, \dots, n\}$ 的 Pearson 相关系数

$$r_p = \frac{\sum (x_i - \bar{x})(|e_i| - \overline{|e|})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (|e_i| - \overline{|e|})^2}}$$

(2) $\{x_i | i=1, 2, \dots, n\}$ 与 $\{|e_i| | i=1, 2, \dots, n\}$ 的 Spearman 相关系数

= Pearson Correlation of Ranking

记 a_i 为 x_i 在 $\{x_i\}$ 中的排名, b_i 为 $|e_i|$ 在 $\{|e_i|\}$ 中的排名, 则有它们的 Spearman 相关系数为:

$$r_s = \frac{\sum (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum (a_i - \bar{a})^2 \sum (b_i - \bar{b})^2}}$$

特别地, 当 $\{x_i\}$ 中两两不同, $\{|e_i|\}$ 中两两不同时, 可简化为

$$r_s = 1 - 6 \sum \frac{(a_i - b_i)^2}{n(n^2 - 1)}$$

Tips: 排名, 例如 $\{1, 5, 3, 2\}$ 的降序为 $\{5, 3, 2, 1\}$, 于是原集合的 ranking 为 $\{4, 1, 2, 3\}$

若有相同, 取平均: 例如 $\{5, 2, 5, 1\}$ 的降序为 $\{5, 5, 2, 1\}$, 则 ranking 为 $\{1.5, 3, 1.5, 4\}$

P-Value 检验

基于 r_s 计算 P 值: 若 P 值小于显著性水平, 则认为 $\{x_i\}$ 与 $\{|e_i|\}$ 有关联
即存在异方差性 (原假设 $r_s = 0$)

① Method 1: 近似 t 分布

当 $n > 8$ 时, 近似分布 (构造枢轴量)

$$t = \frac{\sqrt{n-2} \cdot r_s}{\sqrt{1-r_s^2}} \sim t(n-2)$$

② Method 2: 排列检验 Permutation Test

Step 1:

$\{e_i\}$ 不变

Step 2:

取较大的 $K \in \mathbb{N}_+$, 对 $\{x_i\}$ 重新随机排列 K 次得到 K 个不同的序列

分别记为 $x_i^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}) \quad k = 1, 2, \dots, K$

Step 3:

分别计算 $\{x_i^{(k)}\}$ 与 $\{e_i\}$ 的 Spearman 相关系数 $r_s^{(k)}$

Step 4:

比较原序列的 $|r_s|$ 不大于这 K 个新 $|r_s^{(k)}|$ 的频率, 作为 p 的近似,

即有:

$$p \approx \frac{1}{K} \cdot \sum_{k=1}^K I(|r_s| \leq |r_s^{(k)}|)$$

其中 $I(\cdot)$ 为示性函数

Proof

$H_0: \{x_i\}$ are not associated with $\{e_i\}$ or $|r_s| = 0$

↓

在 H_0 成立时, $\{x_i^{(k)}\}$ 也与 $\{e_i\}$ 无关

↓

在 H_0 成立时, $r_s, r_s^{(1)}, \dots, r_s^{(K)}$ 同分布

↓

$r_s, r_s^{(1)}, \dots, r_s^{(K)}$ 为 $\{r_s\}$ 总体的样本, 经验分布当 $K \rightarrow \infty$ 时
近似总体分布

↓

H_0 成立 ($|r_s| = 0$) 的极端情况为 $|r_s^{(k)}| > 0$ 即

$$\frac{1}{K} \sum \mathbb{I}(|r_s^{(k)}| \geq |r_s|)$$

代表 H_0 成立的极端可能

REMINDER

p 越小, 代表原 H_0 非常奇怪, 重新抽样也无法比 H_0
更奇怪, 故拒绝 H_0 .

Tips: 多元情形

残差定义不变: $e_i = y_i - x_i^T \hat{\beta}$

但 $\{x_i\}$ 变为 $x_{i1}, x_{i2}, \dots, x_{ip}$ ($i=1, 2, \dots, p$) 分别比较

2.2 应对异方差性 Solutions to Heteroscedasticity

目标: 求出 Σ , 而我们已经假设 $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$, 故只需求 σ_i^2 . 而我们认为 σ_i^2 受 x_i 影响, 于是可以构建

$$E(\varepsilon_i^2) = \sigma_i^2 = g(x_i)$$

利用 $\{\varepsilon_i^2\}$ 与 $\{x_i\}$ 回归求解 $g(\cdot)$

① Method 1: 加权最小二乘估计 (Weighted Least Squares Estimation, WLSE)

假设 $\sigma_i^2 = x_i^m$

为简化模型, 假设 $\exists m \in \mathbb{R}$ 使 $\sigma_i^2 = x_i^m$, 关于 m 求解可以使用对数似然函数或交叉验证。得到 m 后:

$$\hat{\Sigma} = \text{diag}(x_1^m, x_2^m, \dots, x_n^m)$$

从而求出 $\hat{\beta}^*$ 为

$$\hat{\beta}_w = (X^T \hat{\Sigma}^{-1} X)^{-1} \cdot X^T \hat{\Sigma}^{-1} \cdot y \quad \leftarrow \text{WLSE}$$

不同于 OLSE: $\hat{\beta} = \arg\min_{b \in \mathbb{R}^p} \|y - Xb\|^2$

利用 GLSE: $\hat{\beta}^* = \arg\min_{b \in \mathbb{R}^p} (y - Xb)^T \Sigma^{-1} (y - Xb)$

有 WLSE: $\hat{\beta}_w = \arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^n \underbrace{\sigma_i^{-2}}_{\text{加权}} (y_i - x_i^T b)^2 \quad (\sigma_i^2 = x_i^m)$

② Method 2: 方差稳定性变换 (Variance-Stabilizing Transformation)

假设 $\sigma_i^2 = [\phi(\mu_i)]^2$ 其中 $\mu_i = E(y_i) = x_i^T \beta$

假设 $\tilde{y}_i = h(y_i)$ 且 $\tilde{y}_i$ 满足 $\text{var}(\tilde{y}_i) = \sigma^2 \in \text{Const}$

根据以上假设, 使 $\{\tilde{y}_i\}$ 对 $\{x_i\}$ 回归即可, 下面推导 $h(\cdot)$

对 $\tilde{y}_i = h(y_i)$ 在 $y_i = \mu_i (= E(y_i) = x_i^T \beta)$ 作一阶泰勒展开:

$$\tilde{y}_i \approx h(\mu_i) + h'(\mu_i) \cdot (y_i - \mu_i)$$

故有 $\tilde{y}_i$ 的方差为:

$$\sigma^2 = \text{var}(\tilde{y}_i) \approx [h'(\mu_i)]^2 \cdot \text{var}(y_i)$$

即由假设 $\text{var}(y_i) = \sigma_i^2 = [\phi(\mu_i)]^2$ 有

$$h'(\mu_i) \approx \frac{\sigma}{\phi(\mu_i)}$$

积分后可得:

$$\begin{aligned} \tilde{y}_i = h(y_i) &\approx \sigma \int [\phi(y_i)]^{-1} dy_i \\ &\propto \int \{\phi(y_i)\}^{-1} dy_i \end{aligned}$$

其中 " $a \propto b$ " 表示 $\exists c \in \text{Const s.t. } a = cb$

Example 1: 若 $\sigma_i^2 \propto \mu_i$, 则 $\phi(\mu_i) \propto \mu_i^{\frac{1}{2}}$, 故取 $\tilde{y}_i = h(y_i) = y_i^{\frac{1}{2}}$

(注: 当 $y_i \sim \text{Poi}(\mu_i)$ 时, 满足 $\sigma_i^2 = \mu_i = \lambda_i$)

Example 2: 若 $\sigma_i^2 \propto \mu_i^2$, 则 $\phi(\mu_i) \propto \mu_i$, 故取 $\tilde{y}_i = h(y_i) = \log y_i$

(注: 当 $y_i \sim T(\alpha, \lambda_i)$ 时, 满足 $\sigma_i^2 = \frac{\alpha}{\lambda_i^2}$ $\mu_i^2 = (\frac{\alpha}{\lambda_i})^2$ 即 $\sigma_i^2 \propto \mu_i^2$)

3. 自相关性 Autocorrelation

Σ 不是对角阵, 这种数据特征称为 自相关性

引入

当自相关性存在时, OLS 不再是 β 的最小方差线性无偏估计。即使 $\text{var}(\varepsilon_i) = \sigma^2$, 估计量 $\frac{\|e\|^2}{n-p}$ 对 σ^2 也可能是有偏的:

$$e = (I - H)y, \quad H = X(X^T X)^{-1} X^T$$

于是有:

$$\because E(y) = X\beta \Rightarrow E(e) = 0$$

$$\text{cov}(e) = E(e e^T) = \text{cov}((I - H)y) = (I - H)\Sigma(I - H)$$

所以有:

$$E\{(e - E(e))(e - E(e))^T\} = E(e e^T)$$

$$E(\|e\|^2) = E(e^T e) = E(\text{tr}(e^T e)) \quad (\because e^T e \in \mathbb{R})$$

$$= E(\text{tr}(e e^T)) \quad (\because \text{tr}(AB) = \text{tr}(BA))$$

$$= \text{tr}(E(e e^T)) = \text{tr}[(I - H)\Sigma(I - H)]$$

$$= \text{tr}((I - H)\Sigma)$$

$$= \text{tr}(\Sigma) - \text{tr}(H\Sigma)$$

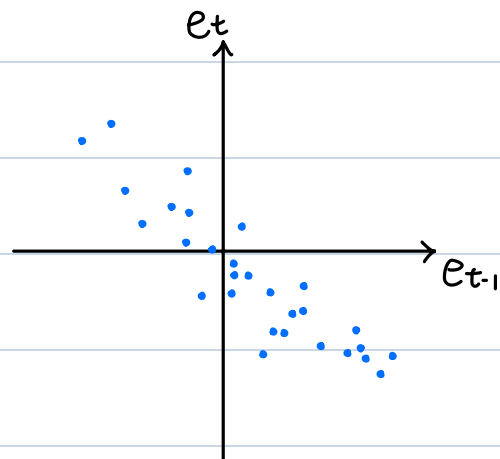
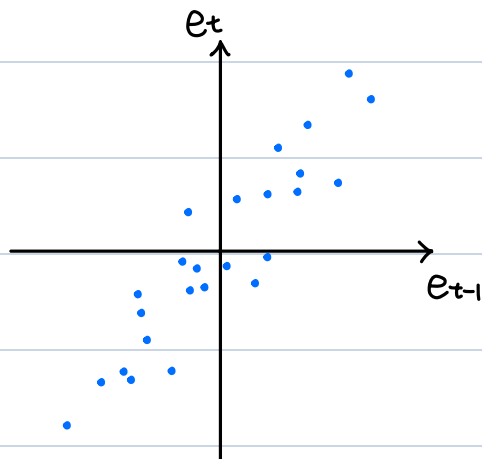
无法化简, 故 $\frac{\|e\|^2}{n-p}$ 对 σ^2 也可能是有偏的

3.1 判断自相关性 Diagnosis of Autocorrelation

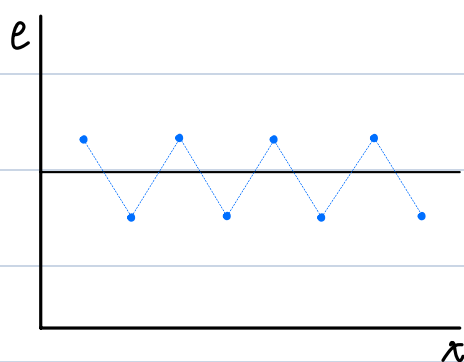
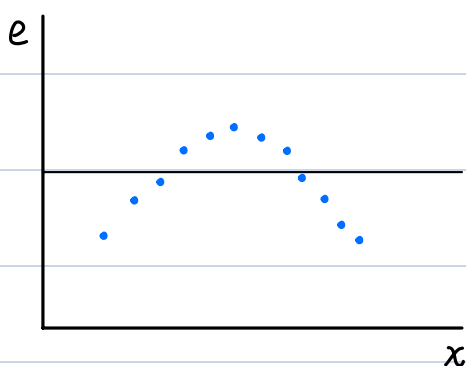
类似地, 利用 OLS 计算 $e_t = y_t - x_t^T \hat{\beta}$

① Method 1:

绘制 (e_{t-1}, e_t) 图像, 判断一阶自相关性



对于单变量模型, 绘制 (x_t, e_t) 图像



② Method 2: 一阶自相关系数

(1) 一阶自相关系数 ρ :

$$\rho = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sqrt{\sum_{t=2}^n e_t^2 \cdot \sum_{t=2}^n e_{t-1}^2}}$$

(2) 一阶自相关系数的估计 $\hat{\rho}$

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sqrt{\sum_{t=2}^n e_t^2 \cdot \sum_{t=2}^n e_{t-1}^2}}$$

③ Method 3: Durbin-Watson (DW) 检验

一种适用于小样本检验的方法, 只针对 一阶 自回归的序列, 具体地为

假设:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (t=2, 3, \dots, n)$$

其中 u_t 是不相关序列，我们要检验：

$$H_0: \rho = 0 \quad \text{v.s.} \quad H_1: \rho \neq 0$$

DW 检验统计量为

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

$$\approx 2(1 - \hat{\rho}) \in [0, 4]$$

Proof: $DW = (\sum_{t=2}^n e_t^2)^{-1} [\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2 \sum_{t=2}^n e_t e_{t-1}] \approx (\sum_{t=2}^n e_t^2)^{-1} [2 \sum_{t=2}^n e_t^2 - 2 \sum_{t=2}^n e_t e_{t-1}]$

$$= 2(1 - (\sum_{t=2}^n e_t^2)^{-1} \sum_{t=2}^n e_t e_{t-1})$$

$$\text{而 } \hat{\rho} = (\sum_{t=2}^n e_t^2 \cdot \sum_{t=2}^n e_{t-1}^2)^{-\frac{1}{2}} \sum_{t=2}^n e_t e_{t-1} \approx (\sum_{t=2}^n e_t^2)^{-1} \sum_{t=2}^n e_t e_{t-1}$$

$$\therefore DW \approx 2(1 - \hat{\rho})$$

DW 检验查表：

根据样本数 n ，自变量个数 k (包括常数项)，查表：

$0 \leq DW \leq d_L$	随机误差 $\{\varepsilon_t: t=1, \dots, n\}$ 间存在正相关
$d_L < DW \leq d_U$	不能判定是否有自相关
$d_U < DW < 4 - d_U$	随机误差 $\{\varepsilon_t: t=1, \dots, n\}$ 间无自相
$4 - d_U \leq DW < 4 - d_L$	不能判定是否有自相关
$4 - d_L \leq DW \leq 4$	随机误差 $\{\varepsilon_t: t=1, \dots, n\}$ 间存在负相关

n 是观察值的数目， k 是解释变量的数目，包括常数项

5% 的上下界

n	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83

3.2 应对自相关性 Solutions to Autocorrelation

① Method 1: 迭代法

假设随机误差具有 一阶 自相关性, 即

$$y_1 = x_1^T \beta + \varepsilon_1$$

$$y_t = x_t^T \beta + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

其中 ① $E(\varepsilon) = 0$ $\text{var}(\varepsilon) = \sigma^2$, 满足 同方差假设

② $E(\varepsilon_{t-1} u_t) = 0$ $E(u_t) = 0$ $\text{cov}(u_t, u_s) = \sigma^2 \cdot I(s=t)$, 满足 u_t 是
不相关序列, u_t 满足 Gauss - Markov 条件

③ ρ 显然为 ε_t 与 ε_{t-1} 的相关系数

于是, 我们得到 一阶线性关联 的线性模型, 采用 一阶差分:

$$y'_t = y_t - \rho y_{t-1}$$

$$x'_t = x_t - \rho x_{t-1}$$

而后使用 $\{y'_t\}$ 和 $\{x'_t\}$ 计算 OLSE

Tip: 迭代发生在 u_t 不满足 Gauss - Markov 条件时

② Method 2: Cochrane-Orcutt (CO) 迭代

过程类似, 结束条件为 $|\hat{\rho}_{t+1} - \hat{\rho}_t| < \delta$

4 Box-Cox 变换

Box-Cox : 综合处理

对于 $\{y_i | i=1, 2, \dots, n\}$ 全正, Box-Cox 变换为:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases} \quad i=1, 2, \dots, n$$

其中 λ 为待定参数 (注: y_i 非正, 先变为 $y_i + a$)

因为 λ 待定, 于是上式为一个变换族。我们~~希望~~变换后满足:

$$y^{(\lambda)} = (y_1^{(\lambda)}, y_2^{(\lambda)}, \dots, y_n^{(\lambda)})^T \sim N(X\beta, \sigma^2 I)$$

下面, 求解 λ

求解:

在假设 $y^{(\lambda)} \sim N(X\beta, \sigma^2 I)$ 条件下, 固定 λ 有 β 与 σ^2 的似然函数:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-(2\sigma^2)^{-1} \cdot \|y^{(\lambda)} - X\beta\|^2\right\}$$

$$\Rightarrow L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-(2\sigma^2)^{-1} \cdot \|y^{(\lambda)} - X\beta\|^2\right\} \cdot |J|$$

$$\text{其中 } |J| = \left| \frac{dy^{(\lambda)}}{dy} \right| = \prod_{i=1}^n y_i^{\lambda-1}$$

↑
代入 λ 后

求出 β 与 σ^2 的 MLE 为:

$$\hat{\beta}_{MLE}(\lambda) = (X^T X)^{-1} X^T \cdot y^{(\lambda)}$$

$$\hat{\sigma}_{MLE}^2(\lambda) = \frac{1}{n} SSE(\lambda, y^{(\lambda)})$$

$$\text{其中 } SSE(\lambda, y^{(\lambda)}) = y^{(\lambda)T} \cdot (I - H) \cdot y^{(\lambda)} = y^{(\lambda)T} \cdot \{I - X(X^T X)^{-1} X^T\} \cdot y^{(\lambda)}$$

$$(y - \hat{y})^T (y - \hat{y}) = [(I - H)y]^T [(I - H)y] = y^T (I - H)y$$

将求得的 $\hat{\beta}_{MLE}(\lambda)$ 和 $\hat{\sigma}_{MLE}^2(\lambda)$ 代回, 取对数, 略去无 λ 项:

$$\log L_{\max}(\lambda) = -n \cdot \log \frac{SSE(\lambda, y^{(1)})}{2} + \log |J|$$

$$= -n \cdot \log \frac{SSE(\lambda, z^{(1)})}{2}$$

其中 $z^{(1)} = (z_1^{(1)}, z_2^{(1)}, \dots, z_n^{(1)})^T = |J|^{-\frac{1}{n}} \cdot y^{(1)}$ ↙ 乘 λ

而 $SSE(\lambda, z^{(1)}) = z^{(1)T} \cdot (I - H) \cdot z^{(1)}$

所以，目标为 $\min SSE(\lambda, z^{(1)})$ 的 λ^*

5. 强影响点, Influential Point

① 定义：远离自变量平均的，会将直线向自身拟合的点，

② 模型：

原始模型： $E(y) = X\beta \quad \text{cov}(y) = \sigma^2 I$

移除1点： $y_{-i} = X_{-i}\beta + \varepsilon_{-i}$ 移除样本 i 后拟合

$E(\varepsilon_{-i}) = 0 \quad \text{cov}(\varepsilon_{-i}) = \sigma^2 I$

Tips: $y, \varepsilon \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p} \quad \beta \in \mathbb{R}^p$

$y_{-i}, \varepsilon_{-i} \in \mathbb{R}^{n-1} \quad X_{-i} \in \mathbb{R}^{(n-1) \times p} \quad \beta \in \mathbb{R}^p$

③ OLS 估计

$$\hat{\beta}_{-i} = \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i e_i$$

原始 OLS

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

移除后 OLS

$$\hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i}$$

$$\begin{aligned}\hat{\beta}_{-i} &= \hat{\beta} - (1 - h_{ii})^{-1}(X^T X)^{-1} x_i y_i + (1 - h_{ii})^{-1}(X^T X)^{-1} x_i x_i^T \hat{\beta} \\ &= \hat{\beta} - (1 - h_{ii})^{-1}(X^T X)^{-1} x_i e_i.\end{aligned}$$

④ Cook 距离

样本 i 是否为强影响点, 取决于 $\hat{\beta}$ 与 $\hat{\beta}_{-i}$ 之间的差距

定义 Cook 距离:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T X^T X (\hat{\beta} - \hat{\beta}_{-i})}{p \hat{\sigma}^2}$$



其中 $\hat{\sigma}^2 = \frac{1}{n-p} \|y - X\hat{\beta}\|^2$



Theorem 1 $\hat{\beta}_{-i}$ 与 D_i 的简化表达:

$$h_{ii} = x_i^T (X^T X)^{-1} x_i$$



$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}$$

于是有:

$$\hat{\beta}_{-i} = \hat{\beta} - \frac{(X^T X)^{-1} x_i}{1 - h_{ii}} e_i$$



$$D_i = \frac{h_{ii}}{p(1 - h_{ii})} r_i^2$$

其中 $\hat{\beta} = (X^T X)^{-1} X^T y$



Tips: ① e 为残差: 由原始 OLS: $\hat{\beta} = (X^T X)^{-1} X^T y$ 得

$$E(e) = 0$$

$$\begin{aligned}\text{cov}(e) &= \text{cov}(y - \hat{y}) = \text{cov}(y - Hy) \quad \text{其中 } H = X(X^T X)^{-1} X^T \\ &= \text{cov}(y) \cdot (I - H)\end{aligned}$$

② r_i 为 e_i 标准化

③ 证明见习作

⑤ 判断强影响点, 并处理

经验判断: Cook 距离 $D_i > 1$

处理: 删去该样本点

6. 异常点, Outlier

① 定义: 少量来自不同模型的观测值点

② 模型:

原始模型: $y \sim N(X\beta, \sigma^2 I)$

增加异常点: $y \sim N(X\beta + d_i\eta, \sigma^2 I)$

其中 $d_i = (0, 0, \dots, 0, 1, 0, \dots, 0)^T \in \mathbb{R}^n$

$\eta \in \mathbb{R}$

↓
被干扰的模型

$$E(y_i) = x_i^T \beta \quad (i \neq j)$$

$$E(y_j) = x_j^T \beta + \eta$$

③ 检验异常点

$$H_0: \eta = 0 \quad \text{v.s.} \quad H_1: \eta \neq 0$$

Tips: 可以等价于 $p+1$ 个自变量, 检验某个 $\beta_j = 0$ 与否

$$X' = (X, d_j) \in \mathbb{R}^{n \times (p+1)} \quad d_j \in \mathbb{R}^n$$

$$\beta' = \begin{pmatrix} \beta \\ \eta \end{pmatrix} \in \mathbb{R}^{p+1} \quad \beta \in \mathbb{R}^p \quad \eta \in \mathbb{R}$$

Theorem 2 对 β' 的 β, η 估计 OLSE 得:

$$\tilde{\beta} = \hat{\beta}_{-j} \in \mathbb{R}^p$$

$$\tilde{\eta} = \frac{e_j}{1 - h_{jj}} \in \mathbb{R}$$

$$\text{其中 } \hat{\beta}_{-i} = \hat{\beta} - \frac{(X^T X)^{-1} x_i}{1 - h_{ii}} e_i$$

$$h_{ii} = x_i^T (X^T X)^{-1} x_i$$

$$\hat{\beta} \text{ 为原始 OLSE: } \hat{\beta} = (X^T X)^{-1} X^T y \leftarrow \text{排除 } d_j \text{ 与 } \eta \text{ 的 } \hat{\beta} \in \mathbb{R}^p$$

Proof: 显然, 有:

$$d_j^T y = y_j$$

$$d_j^T d_j = 1$$

$$X^T d_j = x_j$$

将 $\beta' = (\beta^T, \eta)^T \in \mathbb{R}^{p+1}$ 作为参数做 OLSE

其中 $X' = (X, d_j) \in \mathbb{R}^{n \times (p+1)}$ 作为 Designed Matrix

$$\hat{\beta}' = (X'^T X')^{-1} X'^T \cdot y$$

$$\therefore \begin{bmatrix} \tilde{\beta} \\ \tilde{\eta} \end{bmatrix} = \left\{ \begin{bmatrix} X^T \\ d_j^T \end{bmatrix} \cdot [X, d_j] \right\}^{-1} \cdot \begin{bmatrix} X^T \\ d_j^T \end{bmatrix} \cdot y = \begin{bmatrix} X^T X & x_j \\ x_j^T & 1 \end{bmatrix}^{-1} \begin{bmatrix} X^T y \\ y_j \end{bmatrix}$$

根据分块矩阵求逆公式, 并考虑到 $h_{ii} = x_i^T (X^T X)^{-1} x_i$:

$$\begin{aligned}
 \begin{pmatrix} \tilde{\beta} \\ \tilde{\eta} \end{pmatrix} &= \begin{pmatrix} (X^T X)^{-1} + (X^T X)^{-1} x_j x_j^T (X^T X)^{-1} / (1 - h_{jj}) & -(X^T X)^{-1} x_j / (1 - h_{jj}) \\ -x_j^T (X^T X)^{-1} / (1 - h_{jj}) & (1 - h_{jj})^{-1} \end{pmatrix} \begin{pmatrix} X^T y \\ y_j \end{pmatrix} \\
 &= \begin{pmatrix} \hat{\beta} + (X^T X)^{-1} x_j x_j^T \hat{\beta} / (1 - h_{jj}) - (X^T X)^{-1} x_j y_j / (1 - h_{jj}) \\ y_j / (1 - h_{jj}) - x_j^T \hat{\beta} / (1 - h_{jj}) \end{pmatrix} \\
 &= \begin{pmatrix} \hat{\beta} - (X^T X)^{-1} x_j e_j / (1 - h_{jj}) \\ e_j / (1 - h_{jj}) \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{-j} \\ e_j / (1 - h_{jj}) \end{pmatrix},
 \end{aligned}$$

④ 检验统计量

Theorem 3 若 $H_0: \eta = 0$ 成立, 则有

$$F_j = \frac{(n-p-1) r_j^2}{n-p-r_j^2} \sim F(1, n-p-1)$$

Tip: ① 证明见习作

② “可以等价为 $p+1$ 个自变量, 检验某个 $\beta_j = 0$ 与否”

Chapter 5 ☆

假设检验: 求出 F_j 和 p -Value

① $F_j > \alpha$ 或 $p < \alpha_{1-\alpha}$ 时, 拒绝 H_0 .

② 拒绝 $H_0: \eta = 0$, 说明自变量 j 特征是异常值

区别:

强影响点: 样本点, 离群 $\Rightarrow$ n 变化

异常点: 自变量(特征)错误 $\Rightarrow$ p 变化

