

回归分析样本试卷

1. (20分) 某同学分析数据时假设了如下线性模型

$$y_i = a + bx_i + \gamma'z_i + \epsilon_i, i = 1, \dots, n, n = 57$$

其中 x_i 是感兴趣的自变量, z_i 是 5×1 的控制变量 (向量), γ 是 5×1 的参数向量。她认为 $b \neq 0$, 其原假设为 $H_0: b = 0$ 。假设 $\epsilon_i, i = 1, \dots, n$ iid $\sim N(0, \sigma^2)$, 且与 x_i 和 z_i 独立。应用最小二乘法拟合数据得到 $\hat{b} = 3.76$, 其标准差为 1.88, 复相关系数平方 $R^2 = 0.81$ 。假设自变量 x 与 z 的相关性度量 $R_x^2 = \frac{\|\hat{x} - \mathbf{1}\bar{x}\|^2}{\|\mathbf{x} - \mathbf{1}\bar{x}\|^2} = 0.5$ (这里 \hat{x} 是 $\mathbf{x} = (x_1, \dots, x_n)'$ 在其它自变量张成空间上的投影)。

- 所有拟合值的算术平均大于、等于还是小于 \bar{y} ?
 - 计算 H_0 的 t -检验统计量的值, 它在原假设下服从什么分布?
 - 计算回归方程的显著性的 F 检验统计量的值, 它在原假设下服从什么分布?
 - 计算响应变量与拟合值的样本相关系数。
 - 求在模型中删除变量 z_i 之后 x_i 的回归系数估计的方差。
2. (10分) 数据 $(y_i, x_i, z_i), i = 1, \dots, n$ 中 y_i 是体重, x_i 是身高, $z_i = 0, i = 1, \dots, n_0$ 代表女性; $z_i = 1, i = n_0 + 1, \dots, n$ 代表男性. 对女性和男性分别假设模型:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = 1, \dots, n_0, \\ y_i &= \alpha_0 + \alpha_1 x_i + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = n_0 + 1, \dots, n \end{aligned}$$

- (a) 试说明上述两个模型可以合并成一个模型:

$$y_i = \beta_0 + \beta_1 x_i + \gamma_0 z_i + \gamma_1 z_i x_i + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = 1, 2, \dots, n,$$

- (b) 在合并模型中如何检验 $H_0: \beta_1 = \alpha_1$? 检验统计量在原假设下服从什么分布?

3. (10分) 假设线性模型 $Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}, \epsilon \sim (0, \sigma^2)$, 假设 X 列满秩。记 β 的 LS 估计为 $\hat{\beta}, \hat{Y} = X\hat{\beta}$ 。设 A 为任意一个 $p \times n$ 矩阵, 使得 AX 可逆。令 $\tilde{\beta} = (AX)^{-1}AY$ 以及 $\tilde{Y} = X\tilde{\beta}$ 。

- (a) 证明 $\tilde{\beta}$ 是 β 无偏估计。

- (b) 对任何随机向量 $Y_{n \times 1}^*$, 其均方误差定义为 $m(Y^*) = E\|Y^* - X\beta\|^2$ 。证明 $m(\tilde{Y}) \geq m(\hat{Y})$ 。

4. (20分) 假设线性模型 $Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}, \epsilon \sim (0, \sigma^2)$, 假设 X 列满秩。记 β 的 LS 估计为 $\hat{\beta}, \hat{Y} = X\hat{\beta}$ 。划分 $X = (X_1, X_2)$, 其中 X_1 为 $n \times k$ 矩阵, X_2 为 $n \times (p - k)$ 矩阵 ($1 \leq k < p$), 划分 $\beta = (\beta_1', \beta_2')'$, β_1 和 β_2 分别为 $k \times 1$ 和 $(p - k) \times 1$ 向量。记 $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2)'$, 其中 $\tilde{\beta}_1 = (X_1' X_1)^{-1} X_1' Y, \tilde{\beta}_2 = 0$ 。假设 $X_1' X_2 = 0$ 。

(a) 记 $\tilde{\beta}$ 的均方误差矩阵为 $M(\tilde{\beta}) = E(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'$, 证明

$$M(\tilde{\beta}) = \begin{pmatrix} \sigma^2(X_1'X_1)^{-1} & 0 \\ 0 & \beta_2\beta_2' \end{pmatrix}$$

(b) 证明当 $\|X_2\beta_2\|^2 \leq \sigma^2$ 时, $M(\tilde{\beta}) \leq M(\hat{\beta})$.

5. (20分) 为了使用天平测量两个物体的重量 α 和 β , 现测量 $\alpha + \beta$ 两次, 得到测量值 y_1, y_2 ; 测量 $\alpha - \beta$ 两次, 得到测量值 y_3, y_4 。假设各次测量误差独立, 服从 $N(0, \sigma^2)$ 分布且与真实重量是可加的。

(a) 写出上述测量方案的线性回归模型的矩阵形式, 特别地, 设计阵 X 的具体形式。并求出 α, β 的 LS 估计及其方差。

(b) 求出原假设 $H_0: \alpha = \beta$ 的 F 检验统计量 (以 y_1, \dots, y_4 表示), 该统计量在原假设下的分布是什么?

(c) 另外一种方案是单独测量 α, β 各 m 和 n 次, 为了得到不大于第一种方案的总估计方差(α, β 的LS估计的方差之和), 那么总测量次数 $m + n$ 最少应该是多少?

6. (20分) 假设数据 $(x_i, y_i, z_i), i = 1, 2, \dots, n$ 满足线性模型

$$y_i = a + bx_i + cz_i + \epsilon_i, \quad \epsilon_1, \dots, \epsilon_n \text{ iid } \sim (0, \sigma^2),$$

其中 a, b, c, σ^2 为未知参数, 误差项 ϵ_i 与 (x_i, z_i) 独立。模型也可等价地写成

$$\mathbf{y} = \mathbf{1}a + \mathbf{x}b + \mathbf{z}c + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (0, \sigma^2 I_n),$$

其中 $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{1} = (1, 1, \dots, 1)'$, $\mathbf{x} = (x_1, \dots, x_n)'$, $\mathbf{z} = (z_1, \dots, z_n)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$. 已知 b 的最小二乘估计

$$\hat{b} = \mathbf{x}^\perp' \mathbf{y} / \|\mathbf{x}^\perp\|^2, \quad (1)$$

其中 $\mathbf{x}^\perp = \mathbf{x} - P_{\mathbf{1}, \mathbf{z}} \mathbf{x}$, 其中 $P_{\mathbf{1}, \mathbf{z}}$ 是 $\mathbf{1}, \mathbf{z}$ 张成的空间对应的投影阵。

(a) 由上述 \hat{b} 的公式(1), 证明 b 的最小二乘估计为

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x} - (z_i - \bar{z})\hat{\gamma}) y_i}{\sum_{i=1}^n (x_i - \bar{x} - (z_i - \bar{z})\hat{\gamma})^2},$$

其中 $\hat{\gamma} = s_{xz}/s_{zz} = \sum_{i=1}^n (z_i - \bar{z})x_i / \sum_{i=1}^n (z_i - \bar{z})^2$ 。

(b) 证明 b 的最小二乘估计的方差

$$\text{var}(\hat{b}|x, z) = \frac{\sigma^2}{(1 - r_{xz}^2)s_{xx}},$$

其中 $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, r_{xz} 为 $(x_i, z_i), i = 1, \dots, n$ 的样本相关系数。

(c) 相比于 x_i 与 z_i 不相关的情形, b 的LS估计的方差增加还是减少了多少倍?

(d) 假设已知回归方程的显著性检验是高度显著的, 即 $H_0: b = c = 0$ 的 F -检验高度显著。利用(b)和(c)的结果, 说明什么情况下有可能原假设 $H_{01}: b = 0$ 和 $H_{02}: c = 0$ 的 t -检验都不显著。