作业 **5**

1. 若 $\mathbf{x}_{m\times 1}, \mathbf{y}_{n\times 1}$ 为任意两个随机向量，$E(\mathbf{y}|\mathbf{x})$ 称为回归函数，记 $\boldsymbol{\epsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{x})$.

   (a) 证明 $\boldsymbol{\epsilon}$ 与 $\mathbf{x}$ 不相关（因此，$\mathbf{y}$ 可分解为两个不相关的部分：$\mathbf{y} = E(\mathbf{y}|\mathbf{x}) + \boldsymbol{\epsilon}$）.

   (b) 利用 (a) 的结果证明 $\min_{f:R^m\to R^n} E||\mathbf{y} - f(\mathbf{x})||^2 = E||\mathbf{y} - E(\mathbf{y}|\mathbf{x})||^2$ （提示：首先证明 $\hat{f}(\mathbf{x}) = E(\mathbf{y}|\mathbf{x})$ 极小化给定 $\mathbf{x}$ 条件下的条件期望 $E\left(||\mathbf{y} - f(\mathbf{x})||^2|\mathbf{x}\right)$ ）.

   (c) 利用 (a) 的结果证明方差矩阵分解公式：$\mathrm{var}(\mathbf{y}) = \mathrm{var}[E(\mathbf{y}|\mathbf{x})] + E[\mathrm{var}(\mathbf{y}|\mathbf{x})]$ （提示：只需证明 $var(\boldsymbol{\epsilon}) = E[\mathrm{var}(\mathbf{y}|\mathbf{x})]$）.

   (d) 假设 $\mathbf{x}, \mathbf{y}, \mathbf{z}$ 是三个随机向量，证明

$$cov(\mathbf{x}, \mathbf{y}) = cov[E(\mathbf{x}|\mathbf{z}), E(\mathbf{y}|\mathbf{z})] + E[cov(\mathbf{x}, \mathbf{y}|\mathbf{z})].$$

2. 假设 y 是一元响应变量，$\mathbf{x}$ 是 q 元随机向量，$\mathbf{z}$ 是 $(p-q)\times 1$ 随机向量，都是可观测变量。我们感兴趣的问题是研究 y 与 $\mathbf{x}$ 的关系，而 $\mathbf{z}$ 是干扰因素（既与 y 有关也与 $\mathbf{x}$ 有关），需要加以控制。假设如下线性模型

$$y = a + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{z} + \epsilon,\ \epsilon \sim (0, \sigma^2), \epsilon \perp\!\!\!\perp \mathbf{x}, \mathbf{z} \tag{1}$$

   问题是，该模型是否正确地实现了目标 “在控制 $\mathbf{z}$ 的条件下，研究 y 与 $\mathbf{x}$ 的关系”?

   为了考察模型（1）中加项 $\mathbf{c}^\top \mathbf{z}$ 的作用，以及考察系数 $\mathbf{b}$ 的含义，我们令

$$y^\perp = y - \Sigma_{y\mathbf{z}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z}, \quad \mathbf{x}^\perp = \mathbf{x} - \Sigma_{\mathbf{xz}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z},$$

   则可以证明模型（1）蕴含了

$$y^\perp = a + \mathbf{b}^\top \mathbf{x}^\perp + \epsilon,\ \ \epsilon \sim (0, \sigma^2),\ \ \epsilon \perp\!\!\!\perp \mathbf{x}^\perp. \tag{2}$$

   这说明模型 (1) 中线性项 $\mathbf{c}^\top \mathbf{z}$ 确实起到了控制（或消除）$\mathbf{z}$ 影响的作用。进一步，可以证明

$$\mathbf{b} = \Sigma_{\mathbf{xx}\bullet\mathbf{z}}^{-1}\Sigma_{\mathbf{x}y\bullet\mathbf{z}}, \tag{3}$$

   特别地，当 $q = 1$ 时，$b \propto \rho_{xy\bullet\mathbf{z}}$。结果（4）说明了模型（2）中 $\mathbf{x}$ 的系数 $\mathbf{b}$ 确实代表了偏相关系数，干扰因素的影响通过在模型（1）添加 $\mathbf{c}^\top \mathbf{z}$ 得到了消除。

   另一方面，假设模型确实是正确模型，但假如我们没有意识到 $\mathbf{z}$ 是一个潜在的干扰因素，并假设模型

$$y = \alpha + \boldsymbol{\beta}^\top \mathbf{x} + \delta,\ \delta \sim (0, \tau^2), \delta \perp\!\!\!\perp \mathbf{x} \tag{4}$$

   也就是说，我们将模型（1）中的 $\mathbf{c}^\top \mathbf{z}$ 一项归入了误差项：

$$\delta = \epsilon + \mathbf{c}^\top \mathbf{z} - E(\mathbf{c}^\top \mathbf{z}),$$

   其中最后一项 $-E(\mathbf{c}^\top \mathbf{z})$ 是为了保证 $E(\delta) = 0$, 而 $\alpha = a + E(\mathbf{c}^\top \mathbf{z})$, $\boldsymbol{\beta} = \mathbf{b}$。显然 δ 与 $\mathbf{x}$ 不独立，那么基于错误模型 (4), 由课件命题 2

$$\mathbf{b} = \boldsymbol{\beta} = \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}, \tag{5}$$

   与 (3) 相比，这个表达没有控制 $\mathbf{z}$, 因而是错误的表达。

   本题的任务是，假设模型（1）是正确模型，证明（2）和（3）。