定量社会科学的因果推断

Causal Inference in Quantitative Social Sciences

江　艇
中国人民大学经济学院

Last updated: April 18, 2020

Lecture 2　线性回归理论回顾（上）

样本线性回归 在样本容量为 n 的样本中,寻找能近似 y 的 $x_1, x_2, \ldots, x_K$（其中 $x_1 \equiv 1$）的最佳线性组合

$$b_1 + b_2 x_2 + \ldots + b_K x_K$$

$$\min_{\{b_1,\ldots,b_K\}} S = \sum_{i=1}^{n} \left(y_i - b_1 - b_2 x_{i2} - \ldots - b_K x_{iK}\right)^2$$

$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} \left(y_i - b_1 - b_2 x_{i2} - \ldots - b_K x_{iK}\right) = 0$$

$$\frac{\partial S}{\partial b_2} = -2 \sum_{i=1}^{n} x_{i2} \left(y_i - b_1 - b_2 x_{i2} - \ldots - b_K x_{iK}\right) = 0$$

$$\vdots$$

$$\frac{\partial S}{\partial b_K} = -2 \sum_{i=1}^{n} x_{iK} \left(y_i - b_1 - b_2 x_{i2} - \ldots - b_K x_{iK}\right) = 0$$

向量表示

$$\min_{\mathbf{b}} S(\mathbf{b}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\mathbf{b}\right)^2$$

$$\mathbf{x}_i = (1 \; x_{i2} \; \ldots \; x_{iK})', \; \mathbf{b} = (b_1 \; \ldots \; b_K)'$$

$$\sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i'\mathbf{b}\right) = \mathbf{0}$$

$$\mathbf{b} = \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

几种特殊情形

- 一元回归

$$y_i = b_0 + b_1 x_i + e_i$$

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}}(x, y)}{\widehat{\mathrm{Var}}(x)}$$

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \bar{y} - b_1 \bar{x}$$

- 仅含截距项回归

$$y_i = b_0 + e_i \Rightarrow b_0 = \bar{y}$$

- 不含截距项的一元回归

$$y_i = b_1 x_i + e_i \Rightarrow b_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

- 去均值后的一元回归

$$y_i - \bar{y} = b_1(x_i - \bar{x}) + e_i$$

- 含截距项的虚拟变量一元回归

$$y_i = b_0 + b_1 D_i + e_i \Rightarrow b_0 = \bar{y}_{D=0},\ b_1 = \bar{y}_{D=1} - \bar{y}_{D=0}$$

- 不含截距项的虚拟变量一元回归

$$y_i = b_0 D_{0i} + b_1 D_{1i} + e_i \Rightarrow b_0 = \bar{y}_{D=0},\ b_1 = \bar{y}_{D=1}$$

- (共线性与虚拟变量陷阱)

条件期望函数

- 条件期望函数是基于 $\mathbf{x}$ 的信息对 y 的最佳预测（在最小化均方误差意义上）。

$$\mathbb{E}(y|\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}\left(y - f(\mathbf{x})\right)^2$$

- 定义期望残差 $\tilde{\varepsilon} \triangleq y - \mathbb{E}(y|\mathbf{x})$，则以下结论自动满足：
  - $\tilde{\varepsilon}$ 和 $\mathbf{x}$ 均值独立：$\mathbb{E}(\tilde{\varepsilon}|\mathbf{x}) = 0.$
  - $\tilde{\varepsilon}$ 期望为零：$\mathbb{E}(\tilde{\varepsilon}) = 0.$
  - $\tilde{\varepsilon}$ 和 $\mathbf{x}$ 不相关：$\mathbb{E}\left(\mathbf{x}\tilde{\varepsilon}\right) = 0.$
  - $\tilde{\varepsilon}$ 和 $\mathbf{x}$ 的任何函数均值独立：$\mathbb{E}\left(\tilde{\varepsilon}|f(\mathbf{x})\right) = 0.$
  - $\tilde{\varepsilon}$ 和 $\mathbf{x}$ 的任何函数不相关：$\mathbb{E}\left(f(\mathbf{x})\tilde{\varepsilon}\right) = 0.$

总体线性回归

- 定义总体最小二乘问题

$$\min_{\boldsymbol{\beta}} \mathbb{E}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2$$

$$\mathbb{E}\left(\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right) = \mathbf{0}$$

  称这个问题的解 $\mathbf{x}_i'\boldsymbol{\beta}$ 为总体回归函数。

- 若条件期望函数为线性，则总体回归函数等于条件期望函数。

$$\mathbb{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

- 若条件期望函数为非线性，则总体回归函数是对条件期望函数的最佳线性预测（在最小化均方误差意义上）。[1]

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}^*} \mathbb{E}\left(\mathbb{E}(y_i|\mathbf{x}_i) - \mathbf{x}_i'\boldsymbol{\beta}^*\right)^2$$

[1] 因此，也有人把条件期望函数 $\mathbb{E}(y_i|\mathbf{x}_i)$ 本身称为总体回归函数，而把 $\mathbf{x}_i'\boldsymbol{\beta}$ 称为线性回归的总体回归函数。

- 定义总体回归残差 $\varepsilon_i \triangleq y_i - \mathbf{x}_i'\boldsymbol{\beta}$，则以下结论自动满足（特别注意 ε 和 $\tilde{\varepsilon}$ 的区别）：
  - 无论条件期望函数是否为线性，ε_i 和 $\mathbf{x}_i$ 不相关：$\mathbb{E}(\mathbf{x}_i\varepsilon_i) = 0$.
  - 若条件期望函数为线性，则 ε_i 和 $\mathbf{x}_i$ 均值独立：$\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$.

样本线性回归是对总体线性回归的估计

- 可以把样本最小二乘问题的解视作总体最小二乘问题的解的矩估计量。

$$\mathbb{E}\left(\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right) = 0$$

$$\boldsymbol{\beta} = \left(\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')\right)^{-1}\mathbb{E}(\mathbf{x}_i y_i)$$

$$\mathbf{b} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i\right)$$

因为总体线性回归是对条件期望函数的预测，因此样本线性回归也可以视作是对条件期望函数的估计。

- 当条件期望函数为线性，OLS 估计量 $\mathbf{b}$ 是对总体回归函数——也即条件期望函数——的无偏估计。

$$\mathbb{E}\left(\mathbf{b}|\mathbf{X}\right) = \boldsymbol{\beta}$$

- 当条件期望函数为非线性，OLS 估计量 $\mathbf{b}$ 是对总体回归函数——也即条件期望函数的最佳线性预测——的一致估计。

$$\mathbf{b} \rightarrow_p \boldsymbol{\beta}$$

结构模型

- 第 1 讲介绍的因果模型可以被一般化地写作如下形式（暂时不区分核心解释变量 D 和控制变量 X）：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

这一模型也被称为结构模型，其形式似乎与回归模型完全相同，但含义截然不同。

  - 因果模型的结构反映了我们关于 y 和 $\mathbf{x}$ 之间关系的先验知识。特别而言，模型的线性性质，以及 $\mathbf{x}$ 和 ε 之间的关系，是模型的假设，没有这些假设，我们无从识别 $\boldsymbol{\beta}$。因此称 $\boldsymbol{\beta}$ 为结构参数，称 ε 为结构误差。
  - 线性回归模型是总体回归问题的结果，因此在线性回归模型中，$\mathbf{x}$ 和 ε 之间的关系不是假设，而是结论。

- 若结构误差 ε 和 $\mathbf{x}$ 均值独立，即 $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$，则意味着条件期望函数为线性，$\mathbb{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$。我们已经知道当条件期望函数为线性时，OLS 估计量是条件期望函数的无偏估计，那么它也就是因果参数（结构参数）的无偏估计。
- 类似地，若结构误差 ε 和 $\mathbf{x}$ 不相关，即 $\mathbb{E}(\mathbf{x}_i\varepsilon_i) = 0$，因为总体回归残差总是和 $\mathbf{x}$ 不相关，可知此时结构模型和回归模型等价，OLS 估计量是总体回归函数的一致估计，那么它也就是因果参数（结构参数）的一致估计。

- 例 1：假定数据生成过程为

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \tilde{\varepsilon}_i,\ \mathbb{E}(\tilde{\varepsilon}_i | x_i, w_i) = 0$$

我们写下的结构模型为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

则

$$\varepsilon_i \triangleq \beta_2 w_i + \tilde{\varepsilon}_i$$

此时，如果保持 w_i 和 $\tilde{\varepsilon}_i$ 不变（也就是保持 ε_i 不变），x_i 增加一个单位，对 y_i 的因果效应应该是 β_1.
但如果我们考察回归模型

$$y_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i$$

除非 w_i 与 x_i 均值独立，否则线性回归无法得到对因果效应 β_1 的无偏估计。

例如，假定 w_i 和 x_i 服从联合正态分布

$$\begin{pmatrix} x_i \\ w_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_w \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xw}\sigma_x\sigma_w \\ \cdot & \sigma_w^2 \end{pmatrix} \right)$$

此时 $\mathbb{E}(y_i|x_i)$ 是 x_i 的线性函数，线性回归可以得到对它的无偏估计。

$$\begin{aligned} \mathbb{E}(y_i|x_i) &= \beta_0 + \beta_1 x_i + \mathbb{E}(\varepsilon_i|x_i) \\ &= \beta_0 + \beta_1 x_i + \beta_2 \mathbb{E}(w_i|x_i) \\ &= \beta_0 + \beta_1 x_i + \beta_2 \left[\mu_w + \rho_{xw}\frac{\sigma_w}{\sigma_x}(x_i - \mu_x) \right] \\ &= \underbrace{\left[\beta_0 + \beta_2 \left(\mu_w - \rho_{xw}\frac{\sigma_w}{\sigma_x}\mu_x \right) \right]}_{\triangleq \gamma_0} + \underbrace{\left(\beta_1 + \beta_2 \rho_{xw}\frac{\sigma_w}{\sigma_x} \right)}_{\triangleq \gamma_1} x_i \end{aligned}$$

- 例 2：假定数据生成过程为

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \tilde{\varepsilon}_i,\ \mathbb{E}(\tilde{\varepsilon}_i | x_i) = 0$$

我们写下的结构模型为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

则

$$\varepsilon_i \triangleq \beta_2 x_i^2 + \tilde{\varepsilon}_i,\ \mathbb{E}(\varepsilon_i | x_i) \neq 0$$

但如果我们考察回归模型

$$y_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i$$

容易证明

$$\gamma_0 = \beta_0 + \beta_2 \left[\frac{\left(\mathbb{E}x_i^2\right)^2 - \left(\mathbb{E}x_i\right)\left(\mathbb{E}x_i^3\right)}{\left(\mathbb{E}x_i^2\right) - \left(\mathbb{E}x_i\right)^2} \right]$$

$$\gamma_1 = \beta_1 + \beta_2 \left[\frac{\left(\mathbb{E}x_i^3\right) - \left(\mathbb{E}x_i\right)\left(\mathbb{E}x_i^2\right)}{\left(\mathbb{E}x_i^2\right) - \left(\mathbb{E}x_i\right)^2} \right]$$

如果保持 $\tilde{\varepsilon}_i$ 不变, x_i 增加一个单位, 对 y_i 的因果效应应该是 $\beta_1+2\beta_2 x_i$, 平均因果效应应该是 $\beta_1 + 2\beta_2 (\mathbb{E}x_i)$, 课件线性回归无法得到对这两者的一致估计。

- 例 3：D 表示是否上过大学，y 表示工资水平，由于 D 为虚拟变量，因此 $\mathbb{E}(y_i|D_i)$ 必然可以表示为 $\mathbb{E}(y_i|D_i) = \gamma_0 + \gamma_1 D_i$，事实上

$$\gamma_0 = \mathbb{E}(y_i|D_i = 0)$$
$$\gamma_1 = \mathbb{E}(y_i|D_i = 1) - \mathbb{E}(y_i|D_i = 0)$$

  线性回归模型

$$y_i = \gamma_0 + \gamma_1 D_i + \varepsilon_i \tag{2.1}$$

  总是可以通过 OLS 得到无偏估计。但 γ_1 并不具有因果含义，它只表示总体中上过大学人群和没上过大学人群的平均工资差异。

  模型 (2.1) 试图回答的是如下的**预测性问题 (predictive question)**：“**如果我们观测到 D 的取值为 D_0**，我们预期 y 的取值为何？”

相反，当我们写下结构模型

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{2.2}$$

我们是把 ε 解释为“影响工资水平的不可观测的能力或积极性”，那么 ε 很有可能和 D 相关（能力越强的人越倾向于上过大学）。此时 y 对 D 的 OLS 回归就无法得到因果效应 β_1 的无偏估计。

模型 (2.2) 试图回答的是如下的**因果性问题 (causal question)**：“**如果我们干预人们的上大学行为，将 D 的取值设定为 D_0**，我们预期 y 的取值为何？”