经济与商务实证研究方法
第二讲

陈志远

人大商学院硕博课程

版本：2025 年春期学期

目录

1 线性回归

1.1 线性回归与 CEF

- 随机变量 Y_i 关于 X_i 的线性回归方程为：

$$Y_i = \mathbf{X}_i' b + e_i \tag{1}$$

- 线性回归的估计系数来自于求解如下最小化问题

$$\beta = \underset{b}{argmin} E[(Y_i - \mathbf{X}_i' b)^2] \tag{2}$$

$\beta = (\beta_1, \cdots, \beta_k)'$,$b = (b_1, \cdots, b_k)'$, $b \in \mathbb{R}^k$, $\mathbf{X}_i' = (X_{i1}, X_{i2}, ..., X_{ik})$, 即 $\mathbf{X}_i' b = \sum_{j=1}^k X_{ij} \cdot b_j$.

- 求解 $\underset{b}{min} E[(Y_i - \mathbf{X}_i'b)^2]$ 的一阶条件：

$$\begin{aligned} 2E[\mathbf{X}_i(Y_i - \mathbf{X}_i'\beta)] = 0 &\Leftrightarrow E[\mathbf{X}_i(Y_i - \mathbf{X}_i'\beta)] = \mathbf{0} \\ &\Rightarrow \quad E(\mathbf{X}_i Y_i) = E(\mathbf{X}_i \mathbf{X}_i') \cdot \beta \\ &\Rightarrow \quad \beta = \underbrace{E(\mathbf{X}_i \mathbf{X}_i')^{-1}}_{\text{k-by-k matrix}} \underbrace{E(\mathbf{X}_i Y_i)}_{\text{k-by-1 vector}} \end{aligned} \tag{3}$$

  - 其中一阶条件 $E[\mathbf{X}_i(Y_i - \mathbf{X}_i'\beta)] = E\left[\mathbf{X}_i e_i\right] = \mathbf{0}$, 即有 $E\left[X_{ij} e_i\right] = 0$.
  - 此正交条件即为矩条件 (moment condition)，也可以认为是参数识别 (parameter identification) 条件。

- 考虑仅有常数项与单个解释变量的情形：$\mathbf{X}_i' = (1, X_i)$, 此时

$$Y_i = \alpha + \beta_1 X_i + e_i \tag{4}$$

$$\text{两边取期望: } E\left(Y_i\right) = \alpha + \beta_1 E(X_i) + E(e_i) \tag{5}$$

由一阶条件 $E\left(\mathbf{X}_i e_i\right) = 0$ 可知 $E(X_i e_i) = 0$ 且 $E(e_i) = 0$, 从而有

$$E(X_i Y_i) = \alpha E(X_i) + \beta_1 E(X_i^2) \tag{6}$$

联立 (5) 与 (6)，可得：

$$\begin{cases} \beta_1 = \frac{E(X_i Y_i) - E(X_i) E(Y_i)}{E\left(X_i^2\right) - E(X_i)^2} = \frac{Cov(Y_i, X_i)}{V(X_i)} \\ \alpha = E(Y_i) - \beta_1 E(X_i) \end{cases}$$

命题 1. 回归系数分解公式*:*

$$\beta_k = \frac{Cov(Y_i, \tilde{X}_{ik})}{V(\tilde{X}_{ik})}, \quad X_{ik} = \mathbf{X}_{i,-k}'\gamma + \tilde{X}_{ik}$$

其中 $\widetilde{X}_{ki}$ 是将第 k 个解释变量 X_{ik} 回归到其他所有解释变量 $\mathbf{X}_{i,-k}$ 得到的残差项。

证明. 根据回归方程 $Y_i = \alpha + \beta_1 X_{i1} + ... + \beta_k X_{ik} + ... + \beta_K X_{iK} + e_i$, 将 $\tilde{X}_{ik} = X_{ik} - \mathbf{X}_{i,-k}'\gamma$ 代入该回归方程可得：

$$Y_i = \beta_k \left(\mathbf{X}_{i,-k}'\gamma + \tilde{X}_{ik}\right) + \mathbf{X}_{i,-k}'\beta_{-k} + e_i$$

根据矩条件 $E(\mathbf{X}_{i,-k}\tilde{X}_{ik})=0$ 和 $E(\tilde{X}_{ik})=0$，可知 $Cov(\mathbf{X}_{i,-k},\tilde{X}_{ik})=0$, 从而有

$$Cov(Y_i,\,\tilde{X}_{ik})=\beta_k V(\tilde{X}_{ik})+Cov(e_i,\tilde{X}_{ik})$$

□

注 1. 假设变量 X_{ik} 为研究的核心解释变量，当调整控制变量时，回归系数符号可能会改变：

$$\begin{aligned}\beta_k&=\frac{Cov(Y_i,\tilde{X}_{ik})}{V(\tilde{X}_{ik})}\\&=\frac{1}{V\left(\tilde{X}_{ik}\right)}\left[Cov\left(Y_i,\,X_{ik}\right)-Cov(Y_i,\mathbf{X}'_{i,-k})\gamma\right]\end{aligned}$$

在实证研究中，通常通过逐步添加重要控制变量的方式来佐证结果的稳健性。

1.2 线性回归模型的基本性质

- 线性回归是所有计量模型的基础形式，以下定理展示了线性回归模型的优良性质。

定理 1. （*Regression Justification 1*）线性条件期望函数定理。假设条件期望函数是线性函数，则线性回归函数为总体样本的条件期望函数，即 $E\left(Y_i|\mathbf{X}_i\right)=\mathbf{X}_i'\beta$。

证明. 假设 $E(Y_i\mid X_i)=X_i'\beta^*$，根据 CEF 正交分解定理，有 $E[\mathbf{X}_i(Y_i-E(Y_i\mid\mathbf{X}_i))]=0$，从而 $E[\mathbf{X}_i(Y_i-\mathbf{X}_i'\beta^*)]=0$，$\beta^*$ 取值满足 $\beta^*=E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_iY_i)=\beta$。 □

注 2. 该定理说明当总体样本的条件期望函数为线性函数时，线性回归的系数与总体样本系数一致。

定理 2. *(Regression Justification 2)* 最优线性预测定理。对于所有关于 $\mathbf{X}_i$ 的线性方程：
(i) 线性回归函数 $\mathbf{X}_i'\beta$ 是 Y_i 的最佳线性预测量*;*
(ii) $\mathbf{X}_i'\beta$ 是具有最小均方误差 *(MSE)* 的条件期望函数 $E(Y_i|X_i)$ 的线性逼近函数。

证明. (i) $\beta=E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_iY_i)$ 是最小化问题 $E\left[(Y_i-\mathbf{X}_i'\beta)^2\right]$ 的解；
(ii) 需要证明 $\beta=\underset{b}{argmin}E[(E(Y_i\mid\mathbf{X}_i)-\mathbf{X}_i'b)^2]$，注意到

$$\begin{aligned}(Y_i-\mathbf{X}_i'b)^2&=[(Y_i-E(Y_i\mid\mathbf{X}_i))+(E(Y_i\mid\mathbf{X}_i)-X_i'b)]^2\\&=[Y_i-E(Y_i\mid X_i)]^2+[E(Y_i\mid X_i)-X_i'b]^2+2[Y_i-E(Y_i\mid X_i)][E(Y_i\mid X_i)-X_i'b]\end{aligned}$$

对上式取期望，有

$$E\left[(Y_i-\mathbf{X}_i'b)^2\right]=E[Y_i-E(Y_i\mid\mathbf{X}_i)]^2+E[(E(Y_i\mid\mathbf{X}_i)-\mathbf{X}_i'b)^2]$$
$$+2E\left[(E(Y_i\mid\mathbf{X}_i)-\mathbf{X}_i'b)\,\varepsilon_i\right]$$

根据 CEF 正交分解定理，有 $E\left[(E(Y_i\mid X_i)-X_i'b)\,\varepsilon_i\right]=0$，即有

$$\beta=argmin_bE\left[(Y_i-\mathbf{X}_i'b)^2\right]=argmin_bE[(E(Y_i\mid\mathbf{X}_i)-\mathbf{X}_i'b)^2]$$

□

注 3. 即使总体的条件期望函数是非线性的，回归函数提供了 CEF 的最佳线性逼近！

- **分组回归**：线性回归中的系数可以理解为 $E\left(Y_i|\mathbf{X}_i\right)$ 关于 $\mathbf{X}_i$ 的回归

$$\begin{aligned}\beta&=E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_iY_i)\\&=E(\mathbf{X}_i\mathbf{X}_i')^{-1}E\left[\mathbf{X}_iE(Y_i\mid\mathbf{X}_i)\right]\end{aligned}$$

  - 比如 $\mathbf{X}_i$ 是 i 的性别、出生地，没有 Y_i 的直接信息，但可以计算 $E(Y_i\mid\mathbf{X}_i)$ 分组的平均值
  - 当 $\mathbf{X}_i$ 是离散的时候，

$$E[(E(Y_i\mid X_i)-X_i'b)^2]=\sum[E(Y_i\mid\mathbf{X}_i=\mathbf{u})-\mathbf{u}'b]^2g_x(\mathbf{u})$$

    其中 $g_x(\mathbf{u})$ 为概率密度函数

1.3 线性回归的统计推断

- 在现实情况，CEF 和 β 均不知道，但频数学派 (Frequentist) 认为 β 真实值是确定的，但基于随机变量的估计量本身具有随机性。
- 统计推断（statistical inference）用来刻画估计量的概率分布性质

$$\beta=E(X_iX_i')^{-1}E(X_iY_i)\overset{\text{Sample Analog}}{\longrightarrow}\hat{\beta}=(\sum_iX_iX_i')^{-1}\sum_iX_iY_i$$

  - 样本假设：$W_i=(\mathbf{X}_i,\,Y_i)$ i.i.d
  - 样本类比：$\begin{cases}E[W_i]=\frac{1}{N}\sum_{i=1}^NW_i\\E[W_iW_i']=\frac{1}{N}\sum_{i=1}^NW_iW_i'\end{cases}$
  - 该独立同分布假定保证了大数定律和中心极限定理的适用性

定理 3. 大数定律（*Law of Large Numbers*）：若随机变量 $\{X_i\}_{i=1}^{N}$ *i.i.d*, 则有 $\bar{X}=\frac{1}{N}\sum_i X_i$ 依概率收敛于 $E\left(X_i\right)$, 即 $plim\bar{X}=E(X_i)$.

定理 4. 中心极限定理（*Central Limit Theorem*）：若随机变量 $\{X_i\}_{i=1}^{N}$ 分布满足 *i.i.d.* 且均值和方差分别为 μ 和 σ^2, 则有

$$\sqrt{N}\left(\bar{X}-\mu\right)\xrightarrow{d}\mathcal{N}\left(\mu,\,\sigma^2\right)$$

定理 5. *Slutsky* 定理：$a_n\xrightarrow{d}F(\cdot)$，$plimb_n=b$ 则 $a_n+b_n\xrightarrow{d}a_n+b$, $a_nb_n\xrightarrow{d}a_nb$。

- OLS 系数估计量 $\hat{\beta}$ 的渐进分布:

$$\begin{aligned}\hat{\beta}&=(\sum_i\mathbf{X}_i\mathbf{X}_i')^{-1}\sum_i\mathbf{X}_iY_i\\&=(\sum_i\mathbf{X}_i\mathbf{X}_i')^{-1}\sum_i\mathbf{X}_i\left(\mathbf{X}_i'\beta+e_i\right)\\&=\beta+(\sum_i\mathbf{X}_i\mathbf{X}_i')^{-1}\sum_i\mathbf{X}_ie_i\\\Rightarrow\quad\sqrt{N}(\hat{\beta}-\beta)&=(\frac{1}{N}\sum_i X_iX_i')^{-1}\frac{1}{\sqrt{N}}\sum_i X_ie_i\end{aligned}$$

  - 经典误差独立假设要求 $E(e_i\mid\mathbf{X}_i)=0$ or $E(e_i)=0\Rightarrow$ 无偏估计量 $E(\hat{\beta})=\beta$
  - By LLN, $\text{plim}\frac{1}{N}\sum_i\mathbf{X}_i\mathbf{X}_i'=E(\mathbf{X}_i\mathbf{X}_i')$;
  - By CLT:$\frac{1}{\sqrt{N}}\sum_i\mathbf{X}_ie_i\xrightarrow{d}N(0,E(\mathbf{X}_i\mathbf{X}_i'e_i^2))$
  - 运用 Slutsky 定理有

$$\sqrt{N}(\hat{\beta}-\beta)\xrightarrow{d}\mathcal{N}\left(0,\,E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_i\mathbf{X}_i'e_i^2)E(\mathbf{X}_i\mathbf{X}_i')^{-1}\right)\tag{7}$$

注 4. 对于随机变量 X 和 $a\in\mathbb{R}$, 有 $V(aX)=a^2V(X).V(Ax)=AE(XX')A$ 对于矩阵 $\mathbf{A}\in\mathbb{R}^{K\times K}$ 和随机向量 $\mathbf{X}$, 有

$$\begin{aligned}V(\mathbf{AX})&=E\left[\left(\mathbf{AX}-E(\mathbf{AX})\right)\left(\mathbf{AX}-E(\mathbf{AX})\right)'\right]\\&=E\left[\mathbf{A}\left(\mathbf{X}-E(\mathbf{X})\right)\left(\mathbf{X}'-E(\mathbf{X}')\right)\mathbf{A}'\right]\\&=\mathbf{A}E\left[\left(\mathbf{X}-E(\mathbf{X})\right)\left(\mathbf{X}'-E(\mathbf{X}')\right)\right]\mathbf{A}'\\&=\mathbf{A}V\left(\mathbf{X}\right)\mathbf{A}'\end{aligned}$$

  - 可以使用样本类比得到估计此渐进分布的方差估计量为:

$$\left(\frac{1}{N}\sum_i\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{N}\sum_i\mathbf{X}_i\mathbf{X}_i'\hat{e_i}^2\right)\left(\frac{1}{N}\sum_i\mathbf{X}_i\mathbf{X}_i'\right)^{-1}$$

其中残差项 $\hat{e}_i=Y_i-\mathbf{X}_i'\hat{\beta}$。

- **异方差稳健的标准误（White，1980）**
  - 在同方差假设条件下：$E(e_i^2 \mid \mathbf{X}_i) = \sigma^2$，则有 $E(\mathbf{X}_i\mathbf{X}_i'e_i^2) = E(\mathbf{X}_i\mathbf{X}_i'E(e_i^2 \mid \mathbf{X}_i)) = \sigma^2 E(\mathbf{X}_i\mathbf{X}_i')$，此时系数估计量渐进分布的方差为：

$$
\begin{aligned}
V\left[\sqrt{N}(\hat{\beta}-\beta)\right] &= E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_i\mathbf{X}_i'e_i^2)E(\mathbf{X}_i\mathbf{X}_i')^{-1} \\
&= E(\mathbf{X}_i\mathbf{X}_i')^{-1}\sigma^2
\end{aligned}
$$

  - 当用线性模型来近似 CEF 时：

$$
\begin{aligned}
E[(Y_i - \mathbf{X}_i'\beta)^2 \mid \mathbf{X}_i] &= E\left\{[(Y_i - E(Y_i \mid \mathbf{X}_i)) + (E(Y_i \mid \mathbf{X}_i) - \mathbf{X}_i'\beta)]^2 \mid \mathbf{X}_i\right\} \\
&= V(Y_i \mid \mathbf{X}_i) + [E(Y_i \mid \mathbf{X}_i) - \mathbf{X}_i'\beta]^2
\end{aligned}
$$

    即使 $V(Y_i|\mathbf{X}_i)$ 是常数，但是均方误差会随着回归线与 CEF 的偏离加大而增加。
  - 尽管在理论上容易说明，但异方差问题是一个实证问题：异方差对标准误的影响没有确定的规律。在 Mostly Harmless Econometrics 中，作者建议如果异方差标准误与同方差标准误相差超过 30% 以上，则表示可能是有限样本导致的标准误计算误差。

1.4 线性回归的因果含义

- **实证问题**：受教育年限与收入水平的因果关系
  - 如果没有读高中/本科/硕士/博士，收入水平多高？
  - 辍学决策："The Road Not Taken"by Robert Frost
- 二元教育决策变量下的因果分析：
  - $C_i \in \{0, 1\}$
  - 潜在结果：

$$
\text{潜在收入结果} = \begin{cases} Y_i^1 & if \quad C_i = 1 \\ Y_i^0 & if \quad C_i = 0 \end{cases}
$$

  - 根据之前的分析有：

$$
\begin{aligned}
\underbrace{E[Y_i|C_i=1] - E[Y_i|C_i=0]}_{\text{Observed difference in earnings}} = &\underbrace{E[Y_i^1 - Y_i^0|C_i=1]}_{\text{Average treatment effect on the treated}} \\
&+ \underbrace{E[Y_i^1|C_i=1] - E[Y_i^0|C_i=0]}_{\text{Selection bias}} \qquad (8)
\end{aligned}
$$

注 5. 如果存在选择性偏误，$E[Y_i^1|C_i=1]-E[Y_i^0|C_i=0]>0$, 教育对收入的影响会被高估。

- **条件独立假定（Conditional Independence Assumption, CIA）**: 给定可以观测的特征 $\mathbf{X}_i$, 没有选择性偏误:

$$\left\{Y_i^1, Y_i^0\right\} \perp C_i|\mathbf{X}_i$$

换言之，$E[Y_i|\mathbf{X}_i, C_i=1]-E[Y_i|\mathbf{X}_i, C_i=0]=E[Y_i^1-Y_i^0|\mathbf{X}_i]$.

注 6. 在 CIA 下，可以得到平均处理效应 (ATE)。若只需要得到 ATT, 只需要假定 $Y_i^1 \perp C_i|\mathbf{X}_i$ 或者 $Y_i^0 \perp C_i|\mathbf{X}_i$.

- 多元取值教育变量下的因果分析：
  - 假设个体收入满足方程: $Y_i^s=f_i(s)$, 该方程刻画了不同 s 取值情况下的收入水平;
  - 实际观测的个体收入水平为:

$$Y_i=\sum_{s\in\mathcal{S}}\mathbb{I}(s_i=s)Y_i^s$$

  - **CIA**：$Y_i^s \perp s_i|\mathbf{X}_i$**,** $\forall s$，即给定 $\mathbf{X}_i$, s_i 是随机分配的, 可以认为 s_i 的决策方程如下：

$$s_i=g\left(\mathbf{X}_i\right)+\epsilon_i$$

    其中 $Y_i^s \perp \epsilon_i|\mathbf{X}_i$.
  - 特征 $\mathbf{X}_i$ 处的平均因果效应：
    * k 年教育经历对收入的因果影响为: $E[f_i(s)-f_i(s-4)|\mathbf{X}_i]$
    * 但是对于任意个体 i, 只能观测到 $f_i(s=s_i)$，即其实际接受教育年限和收入水平
    * 在 CIA 假定下有：

$$\begin{aligned}E[Y_i|\mathbf{X}_i, s_i=s]-E[Y_i|\mathbf{X}_i, s_i=s-1]\\=E[f_i(s)-f_i(s-1)|\mathbf{X}_i]\end{aligned}$$

注 7. 比如要研究高中毕业的收入效应，根据 CIA 可知:

$$\begin{aligned}&E[f_i(12)|\mathbf{X}_i, s_i=12]-E[f_i(11)|\mathbf{X}_i, s_i=11]\\&=E[f_i(12)-f_i(11)|\mathbf{X}_i, s_i=12]\\&=E[f_i(12)-f_i(11)|\mathbf{X}_i]\end{aligned}$$

此时, 根据 CIA, 给定可以观测的特诊, 高中辍学决策跟潜在收入无关, 并且在 $\mathbf{X}_i$ 处有 ATT 与 ATE 相等。

- 总体平均因果效应:

$$
\begin{aligned}
&E\left\{E[Y_i|\mathbf{X}_i, s_i = s] - E[Y_i|\mathbf{X}_i, s_i = s-1]\right\} \\
&= E\{E[f_i(s) - f_i(s-1)|\mathbf{X}_i]\} \\
&= E[f_i(s) - f_i(s-1)]
\end{aligned}
$$

- 分样本平均因果效应:

$$E[f_i(s) - f_i(s-1)|s_i = s]$$

注 8. 如果要研究本科教育的平均因果效应，$E\left[f_i(16) - f(12)\,|s_i = 16\right]$ 表示接受**本科教育**对于**本科生**收入的平均效应，而 $E\left[f_i(16) - f_i(12)\right]$ 表总体平均效应。

注 9. 总体平均因果效应可以通过如下计算式得到

$$E[f_i(s) - f_i(s-1)] = \int E[f_i(s) - f_i(s-1)|\mathbf{X}_i = x]dF_{\mathbf{X}}(x)$$

即为关于 $\mathbf{X}_i$ 概率密度函数的加权平均。

注 10. 因果效应的计算过程实质上是一个基于个体特征变量 $\mathbf{X}_i$ **的匹配估计量 (Matching Estimator):**

（1）对于任意 $\mathbf{X}_i$ 取值，比较具有“相同”特征的处理组 (如 $s_i = s$) 和对照组 (如 $s_i = s-1$)

（2）根据 $\mathbf{X}_i$ 概率密度函数进行加权平均

- 因果效应的线性模型
  - 假设个体收入方程具有如下形式:

$$f_i(s) = \alpha + \rho s + \eta_i, \quad s \in \mathcal{S} \tag{9}$$

    * 该模型刻画了不同教育年限 s 下的**潜在收入水平。**

  - 将数据 (s_i, Y_i) 代入上述模型可得:

$$Y_i = \alpha + \rho s_i + \underbrace{\mathbf{X}_i'\boldsymbol{\gamma} + \nu_i}_{\equiv \eta_i} \tag{10}$$

  - s_i 可能通过个体特征 $\mathbf{X}_i$ 与 η_i 相关
    * **基于可观测变量的选择假设(Selection-on-observeables Assumption)**: $E[\eta_i|\mathbf{X}_i] = \mathbf{X}_i'\boldsymbol{\gamma} \Leftrightarrow E[v_i|\mathbf{X}_i] = 0$ 利用 CIA 可得:

$$
\begin{aligned}
E[f_i(s)|\mathbf{X}_i, s_i] &= E[f_i(s)|\mathbf{X}_i] \\
&= \alpha + \rho s + \mathbf{X}_i'\boldsymbol{\gamma}
\end{aligned}
$$

这说明模型 (10) 满足如下线性无关假定：

$$E[Y_i|\mathbf{X}_i, s_i] = \alpha + \rho s_i + \mathbf{X}_i'\boldsymbol{\gamma} \tag{11}$$

注 11. **教育投资结构模型**：考虑个体 i 的教育决策问题:

$$\max_s \{f_i(s) - C_i(s; \mathbf{X}_i)\}$$

其中 $C_i(s; \mathbf{X}_i)$ 为接受教育的成本。从而最优的教育决策为

$$f_i'(s^*) = C_i'(s^*; \mathbf{X}_i) \Rightarrow s^* = g(\mathbf{X}_i)$$

添加个体扰动之后，个体最优教育投资决策方程为：

$$s_i = s^* + u_i = g(\mathbf{X}_i) + u_i$$

CIA 假定要求:

$$Y_i^s \perp u_i|\mathbf{X}_i$$