高斯马尔科夫定理

在 OLS 那篇文章里，我提到高斯马尔科夫定理证明了 OLS 有一些特别好的性质，具体来说，当误差项均值为 0 时，OLS 得到的 w 无偏（unbiased），如果各误差项方差相同，OLS 得到的 w 是最佳无偏线性估计（BLUE, best linear unbiased estimator）。这篇文章里我会解释什么是无偏、什么是最佳无偏线性估计、如何证明 OLS 具有这些性质，并由此展开讨论 OLS 的局限。

如何评价 OLS?

在评价 OLS 之前，我们先定义评价的标准，什么是好的估计？这里我们采用频率学派的标准，即偏差（Bias）和方差（Variance）。

真实的数值表示为 w，我们基于样本估计出的数值表示为 $\hat{w}$，由于存在误差 ϵ，ϵ 是随机变量，影响了 y，因此 y 是随机变量，并影响了通过数据估计得到的$\hat{w}$，在 OLS 中 $\hat{w} = (X^TX)^{-1}X^Ty$，因此 $\hat{w}$ 是随机变量，并有对应的分布。

如上图所示，我们希望 $\hat{w}$的均值接近 w，也就是偏差 $\mathrm{E}(\hat{w}) - w$ 尽量小，当 $\mathrm{E}(\hat{w}) = w$ 时，$\hat{w}$ 就是无偏（Unbiased）估计。

我们希望 $\hat{w}$ 给出的结果 波动小，也就是方差$\mathrm{Var}(\hat{w})$ 尽量小，如果 $\mathrm{Var}(\hat{w})$ 是所有估计里最小的，$\hat{w}$ 就是最佳估计。如果 $\mathrm{Var}(\hat{w})$ 是所有无偏线性估计里最小的，$\hat{w}$ 就是最佳无偏线性估计（BLUE, best linear unbiased estimator）。

期望、协方差

沿用上篇文章的符号，列向量 w 的期望定义为

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}, \quad \mathrm{E}(w) = \begin{bmatrix} \mathrm{E}(w_0) \\ \mathrm{E}(w_1) \\ \mathrm{E}(w_2) \end{bmatrix} = \begin{bmatrix} \bar{w_0} \\ \bar{w_1} \\ \bar{w_2} \end{bmatrix}$$

协方差矩阵是方差以及协方差的矩阵形式，w 的协方差矩阵定义为

$$\begin{aligned}
\mathrm{Var}(w) &= \mathrm{E}\left[[w - \mathrm{E}(w)][w - \mathrm{E}(w)]^T\right] \\
&= \mathrm{E}\left[\begin{pmatrix} w_0 - \bar{w_0} \\ w_1 - \bar{w_1} \\ w_2 - \bar{w_2} \end{pmatrix} \begin{pmatrix} w_0 - \bar{w_0} & w_1 - \bar{w_1} & w_2 - \bar{w_2} \end{pmatrix}\right] \\
&= \begin{bmatrix} \mathrm{E}[(w_0 - \bar{w_0})(w_0 - \bar{w_0})] & \mathrm{E}[(w_0 - \bar{w_0})(w_1 - \bar{w_1})] & \mathrm{E}[(w_0 - \bar{w_0})(w_2 - \bar{w_2})] \\ \mathrm{E}[(w_1 - \bar{w_1})(w_0 - \bar{w_0})] & \mathrm{E}[(w_1 - \bar{w_1})(w_1 - \bar{w_1})] & \mathrm{E}[(w_1 - \bar{w_1})(w_2 - \bar{w_2})] \\ \mathrm{E}[(w_2 - \bar{w_2})(w_0 - \bar{w_0})] & \mathrm{E}[(w_2 - \bar{w_2})(w_1 - \bar{w_1})] & \mathrm{E}[(w_2 - \bar{w_2})(w_2 - \bar{w_2})] \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{Var}(w_0) & \mathrm{Cov}(w_0, w_1) & \mathrm{Cov}(w_0, w_2) \\ \mathrm{Cov}(w_0, w_1) & \mathrm{Var}(w_1) & \mathrm{Cov}(w_1, w_2) \\ \mathrm{Cov}(w_0, w_2) & \mathrm{Cov}(w_1, w_2) & \mathrm{Var}(w_2) \end{bmatrix}
\end{aligned} \tag{1}$$

我们知道期望和协方差具有以下性质

$$
\begin{aligned}
\mathrm{E}(A+B) &= \mathrm{E}(A)+\mathrm{E}(B) \\
\mathrm{E}(A^2) &= \mathrm{E}(A)^2+\mathrm{Var}(A)=\mu^2+\sigma^2 \\
\mathrm{Cov}(A,B) &= \mathrm{E}(AB)-\mathrm{E}(A)\mathrm{E}(B) \\
\mathrm{Cov}(A,B) &= 0 \quad \text{当且仅当 } A,B \text{ 相互独立}
\end{aligned}
\tag{2}
$$

$\hat{w}$ 的期望

在上篇文章中，假设真实存在 $y=Xw+\epsilon$, 这里的 w 是个确定值，因此 $\mathrm{E}[w]=w$，通过最小二乘法估计得出 $\hat{w}=(X^TX)^{-1}X^Ty$，这里的 $\hat{w}$ 是随机变量，因此

$$
\begin{aligned}
\mathrm{E}(\hat{w}) &= \mathrm{E}[(X^TX)^{-1}X^Ty] \\
&= \mathrm{E}[(X^TX)^{-1}X^T(Xw+\epsilon)] \\
&= \mathrm{E}[(X^TX)^{-1}X^TXw]+\mathrm{E}[(X^TX)^{-1}X^T\epsilon] \Longleftarrow (X^TX)^{-1}X^TX=I \\
&= \mathrm{E}[w]+\mathrm{E}[(X^TX)^{-1}X^T\epsilon]
\end{aligned}
\tag{3}
$$

此时我们引入假设 1：$\mathrm{E}(\epsilon)=0$，**假设 2**：X 为确定值（文末会讨论这个假设），$\mathrm{E}(X\epsilon)=X\mathrm{E}(\epsilon)=0$，可得

$$
\begin{aligned}
E(\hat{w}) &= \mathrm{E}[w]+\mathrm{E}[(X^TX)^{-1}X^T\epsilon] \\
&= \mathrm{E}[w]+0 \\
&= w
\end{aligned}
\tag{4}
$$

由上可得，当 $\mathrm{E}(\epsilon)=0$ 时，$\mathrm{E}[\hat{w}]=w$，最小二乘法得到的 $\hat{w}$ 无偏（unbiased）。

$\hat{w}$ 的协方差矩阵

将 $y=Xw+\epsilon$ 代入 $\hat{w}$，得

$$
\begin{aligned}
\hat{w} &= (X^TX)^{-1}X^Ty \\
&= (X^TX)^{-1}X^T(Xw+\epsilon) \\
&= (X^TX)^{-1}X^TXw+(X^TX)^{-1}X^T\epsilon \\
&= w+(X^TX)^{-1}X^T\epsilon
\end{aligned}
\tag{5}
$$

由于 $E(\hat{w})=w$，可得协方差矩阵

$$
\begin{aligned}
\mathrm{Var}(\hat{w}) &= \mathrm{E}\big[[\hat{w}-\mathrm{E}(\hat{w})][\hat{w}-\mathrm{E}(\hat{w})]^T\big] \\
&= \mathrm{E}\big[(\hat{w}-w)(\hat{w}-w)^T\big]
\end{aligned}
\tag{6}
$$

代入 $\hat{w}=w+(X^TX)^{-1}X^T\epsilon$，得

$$
\begin{aligned}
\mathrm{Var}(\hat{w}) &= \mathrm{E}\big[(\hat{w}-w)(\hat{w}-w)^T\big] \\
&= \mathrm{E}\big[[w+(X^TX)^{-1}X^T\epsilon-w][w+(X^TX)^{-1}X^T\epsilon-w]^T\big] \\
&= \mathrm{E}\big[[(X^TX)^{-1}X^T\epsilon][(X^TX)^{-1}X^T\epsilon]^T\big] \\
&= \mathrm{E}\big[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\big]
\end{aligned}
\tag{7}
$$

因为 $\mathrm{E}(\epsilon)=0$，可得

$$
\begin{aligned}
\mathrm{Var}(\epsilon) &= \mathrm{E}\big[[\epsilon-\mathrm{E}(\epsilon)][\epsilon-\mathrm{E}(\epsilon)]^T\big] \\
&= \mathrm{E}(\epsilon\epsilon^T) \\
&= \begin{bmatrix} \mathrm{E}(\epsilon_1^2) & \mathrm{E}(\epsilon_1\epsilon_2) & \dots & \mathrm{E}(\epsilon_1\epsilon_i) \\ \mathrm{E}(\epsilon_2\epsilon_1) & \mathrm{E}(\epsilon_2^2) & \dots & \mathrm{E}(\epsilon_2\epsilon_i) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\epsilon_i\epsilon_1) & \mathrm{E}(\epsilon_i\epsilon_2) & \dots & \mathrm{E}(\epsilon_i^2) \end{bmatrix}
\end{aligned} \tag{8}
$$

现在我们需要引入**假设 3.1**：任意项 ϵ_i 与 ϵ_j 独立，因此 $\forall i \neq j$，$\mathrm{Cov}(\epsilon_i,\epsilon_j)=\mathrm{E}(\epsilon_i\epsilon_j)=\mathrm{E}(\epsilon_i)\mathrm{E}(\epsilon_j)=0$。

对角线上的 $\mathrm{E}(\epsilon_j^2)=\mathrm{Var}(\epsilon_j)$，引入**假设 3.2**：任意项 $\mathrm{Var}(\epsilon_j)$ 为定值 σ^2，也就是 $\mathrm{Var}(\epsilon_1)=\mathrm{Var}(\epsilon_2)...=\mathrm{Var}(\epsilon_i)=\sigma^2$。

把假设 3.1 和假设 3.2 代入 $\mathrm{Var}(\epsilon)$，可得

$$
\begin{aligned}
\mathrm{Var}(\epsilon) &= \mathrm{E}(\epsilon\epsilon^T) \\
&= \begin{bmatrix} \mathrm{E}(\epsilon_1^2) & \mathrm{E}(\epsilon_1\epsilon_2) & \dots & \mathrm{E}(\epsilon_1\epsilon_i) \\ \mathrm{E}(\epsilon_2\epsilon_1) & \mathrm{E}(\epsilon_2^2) & \dots & \mathrm{E}(\epsilon_2\epsilon_i) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\epsilon_i\epsilon_1) & \mathrm{E}(\epsilon_i\epsilon_2) & \dots & \mathrm{E}(\epsilon_i^2) \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \mathrm{E}(\epsilon_i\epsilon_2) & \dots & \sigma^2 \end{bmatrix} \\
&= \sigma^2 I
\end{aligned} \tag{9}
$$

回到 $\mathrm{Var}(\hat{w})$，这里的 X 视为确定值（X 是确定值还是随机变量的讨论后文会提及），所以可以提出来，得

$$
\begin{aligned}
\mathrm{Var}(\hat{w}) &= \mathrm{E}\big[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\big] \\
&= (X^TX)^{-1}X^T\mathrm{E}(\epsilon\epsilon^T)X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TIX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned} \tag{10}
$$

高斯马尔科夫定理

用反证法，假设存在比 OLS 更好的无偏线性估计 $\tilde{w}$，M 为任意矩阵，设

$$
\tilde{w} = My
$$

之所以称之为线性估计，是因为 $\tilde{w}$ 是 y 的线性函数，即 $\tilde{w}=f(y)=My$，OLS 也是 y 的线性函数，即 $\hat{w}=g(y)=(X^TX)^{-1}X^Ty$。

可得 $\tilde{w}$ 的期望是

$$\begin{aligned} E(\tilde{w}) &= E(My) \\ &= E[M(Xw+\epsilon)] \\ &= E(MXw+M\epsilon) \\ &= E(MXw) \end{aligned} \tag{11}$$

为了使 $\tilde{w}$ 无偏，即 $E(\tilde{w})=E(MXw)=E(w)=w$，$MX=I$ 必须恒成立。

由于 M 为任意矩阵，我可以将 M 改写为 $(X^TX)^{-1}X^T+C$，C 是任意矩阵。只要 M 存在，我肯定能找到满足 $(X^TX)^{-1}X^T+C=M$ 的 C，这里没有任何技术含量，不要想太多。

由于 $MX=I$ 必须恒成立，因此 $CX=0$，证明如下

$$\begin{aligned} (X^TX)^{-1}X^T+C &= M \\ [(X^TX)^{-1}X^T+C]X &= MX \\ [(X^TX)^{-1}X^T+C]X &= I \\ (X^TX)^{-1}X^TX+CX &= I \\ CX &= 0 \end{aligned} \tag{12}$$

由于 $\tilde{w}$ 无偏，$MX=I$，$\mathrm{E}(\epsilon\epsilon^T)=\sigma^2 I$，可得

$$\begin{aligned} \mathrm{Var}(\tilde{w}) &= \mathrm{E}\left[[\tilde{w}-\mathrm{E}(\tilde{w})][\tilde{w}-\mathrm{E}(\tilde{w})]^T\right] \\ &= \mathrm{E}\left[(\tilde{w}-w)(\tilde{w}-w)^T\right] \\ &= \mathrm{E}\left[[M(Xw+\epsilon)-w][M(Xw+\epsilon)-w]^T\right] \\ &= \mathrm{E}\left[(M\epsilon)(M\epsilon)^T\right] \\ &= \mathrm{E}(M\epsilon\epsilon^TM^T) \\ &= M\mathrm{E}(\epsilon\epsilon^T)M^T \\ &= \sigma^2 MM^T \end{aligned} \tag{13}$$

由于 $(X^TX)^{-1}X^T+C=M$，$CX=0$ 以及 $X^TC^T=0$，因此

$$\begin{aligned} \mathrm{Var}(\tilde{w}) &= \sigma^2 MM^T \\ &= \sigma^2[(X^TX)^{-1}X^T+C][(X^TX)^{-1}X^T+C]^T \\ &= \sigma^2\left[(X^TX)^{-1}X^T[(X^TX)^{-1}X^T]^T+(X^TX)^{-1}X^TC^T+C[(X^TX)^{-1}X^T]^T+CC^T\right] \\ &= \sigma^2\left[(X^TX)^{-1}+CC^T\right] \end{aligned} \tag{14}$$

因为对于任意矩阵 A，$AA^T\geq 0$ 恒成立，所以

$$\mathrm{Var}(\tilde{w})-\mathrm{Var}(\hat{w})=\sigma^2\left[(X^TX)^{-1}+CC^T\right]-\sigma^2(X^TX)^{-1}=\sigma^2(CC^T)\geq 0$$

也就是说，$\mathrm{Var}(\tilde{w})\geq\mathrm{Var}(\hat{w})$，比 $\hat{w}$ 更好的无偏线性估计不存在，因此 OLS 估计是最佳无偏线性估计。

假设

回顾上面的证明，为了证明 OLS 估计无偏，我们需要**假设 1**：$\mathrm{E}(\epsilon)=0$；**假设 2**：X 为确定值。

为了得到 OLS 估计的协方差矩阵和证明 OLS 估计是最佳无偏线性估计，我们需要**假设 3.1**：$\forall i \neq j$，ϵ_i 与 ϵ_j 独立，$\mathrm{Cov}(\epsilon_i, \epsilon_j=0)$；**假设 3.2**：$\mathrm{Var}(\epsilon_1)=\mathrm{Var}(\epsilon_2) \ldots=\mathrm{Var}(\epsilon_i)=\sigma^2$。这两个假设可以合并为**假设 3**：$\mathrm{Var}(\epsilon)=\sigma^2 I$。

注意，证明 OLS 估计是最佳无偏线性估计不需要假设 ϵ 呈正态分布。

X 是确定值还是随机变量?

在上文的证明里，我们将 X 视作确定值，如果数据来源是可控的实验，X 是实验设计者定义的数值，我给小白鼠甲 1 粒药丸，小白鼠乙 2 粒药丸，小白鼠丙 3 粒药丸... 那么将 X 视作确定值是说得通的，$y=Xw+\epsilon$ 中只有 ϵ 和 y 是随机变量，其中 y 的随机性只来自于 ϵ。但在大部分情况下，X 是抽样得到的，因此 X 应该视作随机变量，$y=Xw+\epsilon$ 中 X、ϵ 和 y 都是随机变量，y 的随机性来自于 X 和 ϵ，因此假设需要调整，例如假设 1 和假设 2 合并为 $\mathrm{E}(\epsilon|X)=0$，即样本 X 与误差 ϵ 不相关（均值独立）的条件下误差均值为零， OLS 估计的期望、协方差矩阵和证明也需要调整，但 OLS 估计是最佳无偏线性估计依然成立，可以参考 Linear regression with random regressors、Regression inference assuming predictors are fixed、Independent variable = Random variable? 以及 Discussion of the Gauss-Markov Theorem。

最小二乘法的局限

虽然高斯马尔科夫定理证明了 OLS 估计是最佳无偏线性估计，但是 OLS 并不万能，依然有局限性。

首先，最大的局限性是其过于看重无偏性。传统统计学理论认为，我们应该先找到无偏估计，再从这些估计里挑选出方差最小的，即便有偏估计的方差比无偏估计的方差更小，因为偏离了真实值，所以没有意义。

虽然传统统计学的思路听起来很有道理，但机器学习领域（尤其是神经网络领域）并不认同这个思路，German et al. (1992) 认为我们应该把偏差和方差综合考虑，即考虑估计的「泛化」能力，这个泛化能力被定义为均方误差（MSE），估计值 $\hat{w}$ 与真实值 w 的欧式距离，由于 $\hat{w}$ 是随机变量，所以将距离取均值，即

$$
\begin{aligned}
MSE &= \mathrm{E}[(w-\hat{w})^2] \\
&= \mathrm{E}\left[[w-\mathrm{E}(\hat{w})+\mathrm{E}(\hat{w})-\hat{w}]^2\right] \\
&= \mathrm{E}\left[[w-\mathrm{E}(\hat{w})]^2+[\hat{w}-\mathrm{E}(\hat{w})]^2+2[w-\mathrm{E}(\hat{w})][\mathrm{E}(\hat{w})-\hat{w}]\right] \\
&= \mathrm{E}\left[[w-\mathrm{E}(\hat{w})]^2\right]+\mathrm{E}\left[[\hat{w}-\mathrm{E}(\hat{w})]^2\right]+\mathrm{E}\left[2[w-\mathrm{E}(\hat{w})][\mathrm{E}(\hat{w})-\hat{w}]\right]
\end{aligned}
\tag{15}
$$

由于 $\mathrm{E}\left[[\mathrm{E}(\hat{w})-\hat{w}]\right]=0$，因此

$$
\begin{aligned}
MSE &= \mathrm{E}[(w-\hat{w})^2] \\
&= \mathrm{E}\left[[w-\mathrm{E}(\hat{w})]^2\right]+\mathrm{E}\left[[\hat{w}-\mathrm{E}(\hat{w})]^2\right]+\mathrm{E}\left[2[w-\mathrm{E}(\hat{w})][\mathrm{E}(\hat{w})-\hat{w}]\right] \\
&= \mathrm{Bias}^2(\hat{w})+\mathrm{Var}(\hat{w})
\end{aligned}
\tag{16}
$$

由上可得，估计的均方误差（MSE）可分解为估计的偏差和方差，如果我们让偏差高一点，使方差降低，使模型更「平滑」，效果也许可以比无偏估计更好，像是岭回归（Ridge regression）和 LASSO 等就是通过增加偏差，使模型更「平滑」，取得了比 OLS 更好的泛化能力。这个 rule of thumb 被称为偏差方差取舍（Bias-Variance Tradeoff），但并不意味着提高偏差就一定能降低方差，我们也很难找到 MSE 最低点，它只是方便我们直觉上理解和记忆。

其次，在贝叶斯方法中，最小二乘法只是一种特殊情况，贝叶斯学派预先假设 w 的先验分布来得出$P(w|X)$ 的后验分布，通过后验分布估计参数得到 $\hat{w}$，这是和频率学派完全不同的思路。

第三，高斯马尔科夫定理的假设可能不满足。对于假设 $\mathrm{E}(\epsilon|X)=0$，如果 ϵ 中包括了我们未考虑的变量影响了数据 X，或者 X 与 y 相互影响，那么 X 和 ϵ 不独立，OLS 估计是有偏的，即计量经济学领域研究的内生性问题，需要引入工具变量和 2SLS 来解决。对于假设 $\mathrm{Var}(\epsilon)=\sigma^2I$，如果数据是时间序列，$\epsilon_t1$ 可能影响了 ϵ_t2，即自相关，$\mathrm{Var}(\epsilon)\neq\sigma^2I$我们需要使用 GLS 等方法来解决。

最后，高斯马尔科夫定理针对的是线性估计，如果改用非线性估计也许可以取得更好的效果，例如决策树、随机森林、神经网络、Kernel 等等，线性估计的优势在于计算简单、可以检验显著性（p 值），但在计算力和工具高度发达的今天，`import scikit-liearn`、 `import keras` 再写两行代码就能进行非线性估计，用交叉验证和 Bootstrap 就可以检验模型的泛化能力，还要什么 p 值?