第3章 机器学习与因果推断：双重Lasso

2025年7月

本讲关键问题

（1） 线性回归中如何估计单个解释变量前的系数？

（2） 当p/n较大时上述估计量是否依然奏效？

（3） OLS与偏回归估计量为什么也是矩估计量（method of moment）？

（4） 当控制变量维数极高时，偏回归估计量为何依旧稳健？

重点与难点

- 理解双重Lasso方法的动机，适用场景和操作步骤；
- 掌握在实际数据中正确适用Double Lasso, 计算估计量的标准差；
- 从直觉和理论角度理解为何Double Lasso能够适应高维解释变量，得到有限维系数的可靠估计。

一、回顾：线性回归中推断单个解释变量的偏效应

大多数实证研究关心单个或者极少数解释变量对Y产生的平均效应。假设在

$$Y = \alpha D + X'\beta + U$$

中，我们关心D对Y产生的平均效应，这等于要准确估计并推断α。上述问题可以通过偏回归估计量得到：

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \tilde{D}_i \tilde{Y}_i}{\sum_{i=1}^{n} \tilde{D}_i^{\,2}} \tag{3.1}$$

其中

$$\tilde{D}_i = D_i - X_i'\hat{\gamma}_D,$$

$$\tilde{Y}_i = Y_i - X_i\hat{\gamma}_Y.$$

当X维数极高时，甚至超过n时，上述偏回归估计量(3.1)依然能够准确估计D的平均偏效应。**这背后的原理是什么？为什么其他估计量做不到？**

一、回顾：线性回归中推断单个解释变量的偏效应

从矩估计量（method of moment）视角重新理解偏回归：

$$E\left[\left(\tilde{Y}_i - \alpha\tilde{D}_i\right)\tilde{D}_i\right] = 0. \tag{3.2}$$

(3.2)可以视作估计α的矩条件。所谓矩条件，就是以待估参数为未知数的，各个可观测变量须满足的总体期望方程。

$$\tilde{Y} - \alpha\tilde{D} = X'\beta + U - X'\gamma_Y + X'(\alpha\gamma_D) = m(X) + U$$

$$\tilde{D} = D - E(D|X)$$

$$E\left[\left(D - E(D|X)\right)(m(X) + U)\right] = 0$$

$\hat{\alpha}$可以视作矩条件的样本解。 $\hat{\alpha}$的渐近分布：

$$\sqrt{n}(\hat{\alpha} - \alpha) \to^d N\left(0, \left(E\left[\tilde{D}^2\right]\right)^{-1}\sigma^2\right) = N\left(0, \left(E\left[\left(D - E(D|X)\right)^2\right]\right)^{-1}\sigma^2\right).$$

思考题： OLS估计量对应的矩条件是什么？

二、双重Lasso

- **当X维数很高时，上述偏回归估计量是否仍然可靠？**

此时数据生成过程是

$$Y = \alpha D + X'\beta + U, E(U|D, X) = 0. \tag{3.3}$$

其中D是一维随机变量，X是p维随机变量（p可以很大）。

表面上看，post Lasso可以解决以上问题：首先，用Lasso估计Y对(D, X)的回归系数(α, β)，由于Lasso具有变量选择功能，因此β中某些估计值是零。记X前系数估计量不等于零的分量为X_S，然后用Y对(D, X_S)再做OLS回归，这时D前面系数的估计量会被认为是α的一致估计。

- **以上post Lasso是否真的奏效？**

二、双重Lasso

在此，我们考虑四种估计α的方法：

α的估计方法	具体步骤	评论
（1）Lasso $\hat{\alpha}_{lasso}$	用Y对(D,X)进行Lasso，将得到的系数估计量$\hat{\alpha}_{lasso}$作为α的估计量。	由于惩罚项的存在，lasso估计量是系数的有偏估计（压缩估计）。
（2）post Lasso $\hat{\alpha}_{post}$	用Y对(D,X)进行Lasso，记$\hat{\beta}_{lasso}$不等于零的分量是X_{S_1}。用Y对(D,X_{S_1})做OLS，将得到的估计量$\hat{\alpha}_{post}$作为α的估计量。	倘若第1步lasso能够“准确”选择变量，post lasso是一致估计。
（3）double select $\hat{\alpha}_{dlm1}$	用D对X进行Lasso，记系数不等于零的分量对应的X集合是X_{S_2}。用Y对X进行Lasso，记系数不等于零的分量是X_{S_3}。用Y对(D,X_{S_2},X_{S_3})做OLS，将得到的估计量 $\hat{\alpha}_{dlm1}$ 作为α的估计量。	double select比post lasso在最后一步估计中加入了更多的X（那些对D具有预测能力的 X_{S_2}），这缓解了post lasso过度依赖第1步lasso对变量选择的准确性。
（4）double lasso $\hat{\alpha}_{dlm2}$	用D对X进行Lasso，相应的残差是$D_i - X_i'\hat{\gamma}_D = \tilde{D}_i$。用$Y$对$X$进行Lasso，残差是$Y_i - X_i'\hat{\gamma}_Y = \tilde{Y}_i$。用$\tilde{Y}_i$对$\tilde{D}_i$做OLS回归，记$\tilde{D}_i$前面的系数是$\hat{\alpha}_{dlm2}$。	Double lasso是偏回归估计量在X高维情形下的推广，具有良好的小样本性质。

二、双重Lasso

数据生成过程：

$$Y = D + X\beta' + U, p = dimX = 100, n = 100.$$

$$\beta_j = \frac{1}{j^a}, j = 1, \dots, p; X \sim^d N(0, I_p); U \sim^d N(0,1).$$

$$D = X'\gamma + V; \gamma_j = \frac{1}{(p+1-j)^a}, j = 1, \dots, p; V \sim^d N(0,1)/4.$$

（1）用交错验证法选择λ：

	$a=1$			$a=1.5$			$a=2$		
	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std
$\hat{\alpha}_{lasso}$	-0.1017	0.0210	0.1034	-0.2041	0.0606	0.1376	-0.2458	0.0897	0.1712
$\hat{\alpha}_{post}$	-0.0451	0.0177	0.1249	-0.1233	0.0467	0.1774	-0.1548	0.0734	0.2224
$\hat{\alpha}_{dml1}$	0.0935	0.8535	0.9192	-0.0826	0.4146	0.6386	-0.1262	0.3385	0.568
$\hat{\alpha}_{dml2}$	0.8118	1.3487	0.8305	0.274	0.3781	0.5506	0.1684	0.2733	0.495
	$a=2.5$			$a=3$			$a=4$		
	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std
$\hat{\alpha}_{lasso}$	-0.2714	0.1122	0.1963	-0.2864	0.1269	0.2119	-0.2983	0.1393	0.2243
$\hat{\alpha}_{post}$	-0.1756	0.094	0.2514	-0.1884	0.1085	0.2701	-0.1989	0.1211	0.2855
$\hat{\alpha}_{dml1}$	-0.1372	0.3117	0.5412	-0.1484	0.2986	0.5259	-0.1605	0.2798	0.5041
$\hat{\alpha}_{dml2}$	0.127	0.2404	0.4736	0.0995	0.2237	0.4624	0.0632	0.2069	0.4504

二、双重Lasso

（2）同样的数据生成过程用插值法选择λ：

	$a=1$			$a=1.5$			$a=2$		
	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std
$\hat{\alpha}_{lasso}$	-0.7531	0.584	0.1299	-0.8409	0.7251	0.1346	-0.8553	0.75	0.1363
$\hat{\alpha}_{post}$	0.0334	0.0159	0.1217	-0.0518	0.0305	0.1666	-0.0768	0.0479	0.2049
$\hat{\alpha}_{dml1}$	0.0977	0.0351	0.16	0.0070	0.0518	0.2275	-0.0027	0.0844	0.2905
$\hat{\alpha}_{dml2}$	0.0577	0.0127	0.0966	-0.0008	0.0072	0.0846	-0.0042	0.0052	0.0717

	$a=2.5$			$a=3$			$a=4$		
	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std	Bias Ratio	MSE	Std
$\hat{\alpha}_{lasso}$	-0.8608	0.7599	0.1377	-0.8639	0.7655	0.1385	-0.8667	0.7705	0.1388
$\hat{\alpha}_{post}$	-0.0929	0.0642	0.2359	-0.1037	0.0769	0.2573	-0.1146	0.0892	0.2759
$\hat{\alpha}_{dml1}$	-0.0054	0.1164	0.3412	-0.0075	0.141	0.3754	-0.0102	0.1648	0.4059
$\hat{\alpha}_{dml2}$	-0.0043	0.004	0.0634	-0.0044	0.0035	0.0588	-0.0046	0.003	0.055

使用插值法估计λ时，double lasso方法相对于其他方法能得到偏误率更小、小样本性质更好的估计量。

三、双重Lasso为何表现良好？

➢ **为什么上述四个估计量中，只有$\hat{\alpha}_{dml2}$具有良好的小样本性质？背后的理论逻辑是什么？**

在模型

$$Y = \alpha D + X'\beta + U$$

中，我们感兴趣的参数是α。$\hat{\alpha}_{dml2}$的估计步骤对应矩条件

$$E\left[\left(\tilde{Y}(\gamma) - \alpha\tilde{D}(\gamma)\right)\tilde{D}(\gamma)\right] = 0.$$

其中$\tilde{D}(\gamma) = D - X'\gamma_D$，$\tilde{Y}(\gamma) = Y - X'\gamma_Y$，$\gamma = (\gamma_D, \gamma_Y)$，$\alpha$是$\gamma$的函数，即：

$$\alpha(\gamma) = E\left[\left(\tilde{D}(\gamma)\right)^2\right]^{-1} E\left[\tilde{Y}(\gamma)\tilde{D}(\gamma)\right].$$

双重Lasso表现良好的原因在于：可以证明$\alpha(\gamma)$关于γ的导数在真值处等于零，即

$$\frac{\partial\alpha(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_0} = 0. \qquad (3.4)$$

在用双重Lasso估计α时，作为中间步骤的参数$\gamma = (\gamma_D, \gamma_Y)$通过lasso得到。与传统OLS不同，Lasso估计量具有较大的偏误。这是所有机器学习估计量的共同特点。(3.4)保证了$\gamma = (\gamma_D, \gamma_Y)$的估计量的偏误对$\alpha$估计量的影响（偏误）降至最低。

对（3.4）的验证：

$$E\left[\left(\tilde{Y}(\gamma)-\alpha(\gamma)\tilde{D}(\gamma)\right)\tilde{D}(\gamma)\right]=0$$

记$m(\alpha,\gamma)=E[(\tilde{Y}(\gamma)-\alpha\tilde{D}(\gamma))\tilde{D}(\gamma)]$,利用隐函数求导公式：

$$\frac{\partial\alpha(\gamma)}{\partial\gamma}|_{\gamma=\gamma_0}=-\left[\frac{\partial E[(\tilde{Y}(\gamma)-\alpha\tilde{D}(\gamma))\tilde{D}(\gamma)]}{\partial\alpha}\right]|_{\gamma=\gamma_0}^{-1}\times\frac{\partial E[(\tilde{Y}(\gamma)-\alpha\tilde{D}(\gamma))\tilde{D}(\gamma)]}{\partial\gamma}|_{\gamma=\gamma_0}$$

$$\frac{\partial[(\tilde{Y}(\gamma_Y)-\alpha\tilde{D}(\gamma_D))\tilde{D}(\gamma_D)]}{\partial\gamma_D}|_{\gamma=\gamma_0}=E\left[\tilde{Y}(\gamma_Y)X+\alpha 2\tilde{D}(\gamma_D)X\right]|_{\gamma=\gamma_0}=0$$

$$\frac{\partial E[(\tilde{Y}(\gamma_Y)-\alpha\tilde{D}(\gamma_D))\tilde{D}(\gamma_D)]}{\partial\gamma_Y}=E\left[X\tilde{D}(\gamma_D)\right]|_{\gamma=\gamma_0}=0$$

三、双重Lasso为何表现良好？

- **为什么post-lasso估计效果不佳？**

  Post-lasso对应的矩条件是

$$m(\alpha,\gamma)=E[(Y-\alpha D-X'\gamma)D]=0$$

$$\partial_Y m(\alpha,\gamma)\Big|_{\gamma=\gamma_0}=E(DX)\neq 0$$

$\hat{\alpha}_{dml2}$不仅具有良好的小样本性质，它的渐进方差计算也十分简便。可以证明：

$$\sqrt{n}(\hat{\alpha}_{dml2}-\alpha)\rightarrow^{d} N(0,\left(E\left[\tilde{D}^2\right]\right)^{-1}\sigma^2)$$

因此$\hat{\alpha}_{dml2}$的渐进方差估计量是$\frac{(\hat{E}[\tilde{D}^2])^{-1}\hat{\sigma}^2}{n}$,其中$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$，$\hat{e}_i=\hat{\tilde{Y}}_i-\hat{\tilde{D}}_i\hat{\alpha}_{dml2}$，$\hat{\tilde{Y}}_i=Y_i-X_i'\hat{\gamma}_Y$，

$\hat{\tilde{D}}_i=D_i-X_i'\hat{\gamma}_D$，$\hat{E}\left[\tilde{D}^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{D}}_i\right)^2$。

$\hat{\alpha}_{dml2}\hat{X}_i'\hat{\gamma}_Y = \hat{Y}_i$

四、实证案例：穷国增长速率是否更快？

表3.2 用OLS、Post-Lasso和Double Lasso估计经济增长的收敛性

	估计量	标准差	置信区间
OLS	-0.0089	0.0314	[-0.0732,0.0555]
Post-Lasso	-0.0308	0.0148	[-0.0599,-0.0007]
Double Lasso	-0.0293	0.0141	[-0.0573, -0.0013]

案例中Post-Lasso和Double Lasso都使用式子$\lambda = 2c\hat{\sigma}\sqrt{n}z_{1-a/(2p)}$确定$\lambda$，如果用CV确定Post-Lasso过程中的$\lambda$，估计量差不多，标准差会大一点点。

谢 谢!