

一、附录

附录 1. 广义随机森林模型

本部分将从计量经济学中条件均值、潜在结果框架下的因果效应等基本概念出发，逐步介绍作为机器学习方法的随机森林模型如何被应用于估计处理效应。

对随机变量 (X_i, Y_i) ，其中 X_i 是 $K \times 1$ 的向量， $i = 1, \dots, n$ ，在最小化均方误差（MSE）准则下，用 X_i 去预测 Y_i 的最优函数是条件期望函数 $\mu(x) = \mathbb{E}(Y_i | X_i = x)$ 。对 $\mu(x)$ 的估计可分为参数方法和非参数方法。参数方法需要对 $\mu(x)$ 的具体函数形式进行假设，例如，目前较为流行的古典线性回归模型假定 $\mu(x)$ 为 X_i 的线性函数。非参数方法不需要假设函数形式，例如 K 近邻法、决策树以及随机森林方法。本文主要关注如何使用随机森林法估计 $\mu(x)$ 。

随机森林方法的基础是决策树（decision tree），决策树应用于回归问题即为回归树（regression tree）。统计学中 CART 算法（Breiman 等, 1984）采用的决策树为二叉树，每次从特征向量 X_i 中选取一个特征将母节点分为两个子节点，直至到达终结点（也称叶节点）。终结点所有训练样本 Y 的平均值即为相应的预测值。在每个节点进行分裂时需要确定分裂变量和该分裂变量的临界值，回归树一般采用最小化残差平方和作为节点的分裂准则。在每次分裂后，要求两个子节点的残差平方和的总和最小，用函数表达如下：

$$\arg \min_{j,s \in P} \text{Error}(j,s) \left[\sum_{d_i \in C_1(j,s)} (Y_i - \bar{Y}_{C_1})^2 + \sum_{d_i \in C_2(j,s)} (Y_i - \bar{Y}_{C_2})^2 \right] \quad (1)$$

其中， j 和 s 分别是分裂变量和临界值， C_1 和 C_2 表示从母节点 P 分裂出的两个子节点， $\bar{Y}_{C_1}$ 和 $\bar{Y}_{C_2}$ 表示落在子节点 C_1 和 C_2 中训练样本的均值。每一步的分裂过程均需要最小化式（1），直到满足设定的终止条件。

为克服决策树模型在预测时可能产生的较大误差，Breiman（1996）提出了装袋法（bagging）。装袋法采用自助法对训练样本进行有放回的抽样，得到 B 个自助样本并构建 B 棵决策树，预测值是 B 棵树的算术均值。由于采用有放回的再抽样，每棵树都有一部分未使用的观测值，称为袋外（Out-of-Bag）观测值，相应地有袋外预测值和袋外误差。在装袋法的基础上，Breiman（2001）进一步提出了随机森林（Random Forest）：在每个分裂节点上仅随机选取部分变量作为分裂变量。使用随机森林预测 $\mu(x)$ 可由下式表达：

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \frac{Y_i I(\{X_i \in L_b(x), i \in S_b\})}{|i: X_i \in L_b(x), i \in S_b|} \quad (2)$$

其中， $L_b(x)$ 表示特征变量为 x 的样本在第 b 棵树所对应的叶节点， S_b 是从全部样本中选取的子样本集合， $S_b \subseteq \{1, \dots, n\}$ 。 $I(\cdot)$ 是示性函数。

在介绍了如何使用随机森林进行估计 $\mu(x)$ 的基础上，下文将介绍如何使用随机森林来进行因果推断和估计处理效应。本文将在潜在结果框架下讨论因果效应。对于 n 个独立同分布的样本 (X_i, Y_i, W_i) ，其中 X_i 是个体特征， Y_i 是结果变量， W_i 是处理状态。 Y_{1i} 和 Y_{0i} 分别代表个体 i 在接受干预和未接受干预状态下的潜在结果。 $\tau_i = Y_{1i} - Y_{0i}$ ，即为个体 i 的因果效应； $\tau = \mathbb{E}[Y_{1i} - Y_{0i}]$ ，即为平均处理效应； $\tau(x) = \mathbb{E}[Y_{1i} - Y_{0i} | X_i = x]$ ，即为条件平均处理效应。当条件平均处理效应为常数，即 $\tau(x) = \tau$ ；并且非混杂性（unconfoundedness）假设成立，即 $(Y_{1i}, Y_{0i}) \perp W_i | X_i$ ，Robinson（1988）指出条件平均处理效应可由下式计算：

$$\circ = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}^{(-i)}(X_i))(W_i - \hat{e}^{(-i)}(X_i))}{\frac{1}{n} \sum_{i=1}^n (W_i - \hat{e}^{(-i)}(X_i))^2} \quad (3)$$

其中, $\hat{m}^{(-i)}(X_i)$ 和 $\hat{e}^{(-i)}(X_i)$ 分别表示在给定协变量 X 的情况下 Y_i 的预测值和个体 i 接受处理的概率, $\hat{m}^{(-i)}(X_i)$ 和 $\hat{e}^{(-i)}(X_i)$ 中的上标 $(-i)$ 表示袋外预测值。

在式 (3) 的基础上, Nie 和 Wager (2021) 进一步提出了一种以 “R-learner” 为基础的估计方法用于估计异质性处理效应:

$$\circ(\cdot) = \arg \min_{\circ} \left\{ \sum_{i=1}^n \left((Y_i - \hat{m}^{(-i)}(X_i)) - \circ(X_i)(W_i - \hat{e}^{(-i)}(X_i)) \right)^2 + \Lambda_n(\circ(\cdot)) \right\} \quad (4)$$

其中, $\Lambda_n(\circ(\cdot))$ 是正则化项, 用于控制 $\circ(\cdot)$ 函数的复杂程度。

从机器学习的角度来看, 随机森林是一种集成学习方法; 从非参数估计角度来看, 随机森林可以被视为一种自适应核函数 (Athey 等, 2019)。我们可以对式 (2) 进行简单的变形:

$$\hat{\mu}(x) = \sum_{i=1}^n \bar{o}_i(x) Y_i, \quad \bar{o}_i(x) = \frac{1}{B} \sum_{b=1}^B \frac{Y_i I(\{X_i \in L_b(x), i \in S_b\})}{|i: X_i \in L_b(x), i \in S_b|} \quad (5)$$

其中, $\bar{o}_i(x)$ 可以理解为自适应核函数, 衡量了第 i 个训练样本落在以 x 为特征向量的叶节点的频率。结合式 (3) 和式 (5), 从核函数视角出发, Athey 等 (2019) 将随机森林应用于估计处理效应的方法为:

$$\circ = \frac{\sum_{i=1}^n \bar{o}_i(x) (Y_i - \hat{m}^{(-i)}(X_i))(W_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^n \bar{o}_i(x) (W_i - \hat{e}^{(-i)}(X_i))^2} \quad (6)$$

其中, $\hat{m}^{(-i)}(X_i)$ 和 $\hat{e}^{(-i)}(X_i)$ 的定义同式 (3) 一致, 在算法实现过程中可以采用回归森林进行估计。此外, 为了更好地捕捉处理效应的异质性, Athey 等 (2019) 提出的广义随机森林不仅关注预测误差, 而且强调分支节点的组间差异。为确保估计结果的可靠性, 在构建森林时, 采用 Wager 和 Athey (2018) 提出的 honest 方法, 一半样本用于构建随机森林, 另一半样本用于估计处理效应。关于估计参数的统计性质, Athey 等 (2019) 对渐进正态性、置信区间等做了详细的讨论。

二、附表

附表 1

描述性统计

变量名	样本量	平均值	标准差	最小值	中位数	最大值
<i>Tgreen</i>	16136	0.231	0.503	0.000	0.000	1.946
<i>Igreen</i>	16136	0.089	0.274	0.000	0.000	1.099
<i>Ugreen</i>	16136	0.175	0.431	0.000	0.000	1.609
<i>Age</i>	16136	2.602	0.438	0.000	2.639	3.912
<i>Soe</i>	16136	0.407	0.491	0.000	0.000	1.000
<i>Size</i>	16136	21.845	1.289	13.08	21.68	28.509
<i>Density</i>	16136	3.371	1.014	0.639	3.307	6.412

<i>Lev</i>	16136	0.415	0.212	0.047	0.405	0.983
<i>Profit</i>	16136	20.984	0.381	14.878	20.887	23.056
<i>Liquid</i>	16136	2.638	3.035	0.221	1.630	19.080
<i>Cflow</i>	16136	0.047	0.070	-0.182	0.046	0.254
<i>Roe</i>	16136	0.066	0.123	-0.715	0.069	0.541
<i>Eps</i>	16136	0.369	0.492	-1.080	0.270	2.301
<i>Inde</i>	16136	0.370	0.053	0.091	0.333	0.800
<i>Share1</i>	16136	0.358	0.150	0.003	0.342	0.900

附表 2 “大气十条”政策对企业绿色创新水平影响的动态效应

变量	<i>Tgreen</i> (1)	<i>Igreen</i> (2)	<i>Ugreen</i> (3)
<i>Before6(2007)</i>	-0.083** (0.032)	-0.036** (0.016)	-0.067** (0.029)
<i>Before5(2008)</i>	-0.038 (0.024)	-0.015 (0.018)	-0.035 (0.022)
<i>Before4(2009)</i>	-0.036 (0.025)	-0.026* (0.015)	-0.023 (0.020)
<i>Before3(2010)</i>	0.000 (0.019)	-0.000 (0.015)	-0.002 (0.017)
<i>Before2(2011)</i>	0.003 (0.023)	0.005 (0.014)	-0.005 (0.021)
<i>Before1(2012)</i>	0.032* (0.017)	0.009 (0.013)	0.022 (0.017)
<i>After1(2014)</i>	0.030 (0.029)	0.010 (0.016)	0.022 (0.025)
<i>After2(2015)</i>	0.065*** (0.025)	0.017 (0.017)	0.057** (0.024)
<i>After3(2016)</i>	0.081*** (0.019)	0.042*** (0.016)	0.056*** (0.017)
<i>After4(2017)</i>	0.064** (0.025)	0.056*** (0.015)	0.026 (0.019)
控制变量	YES	YES	YES
观测值	15783	15783	15783
R^2	0.640	0.500	0.583
企业固定效应	YES	YES	YES
时间固定效应	YES	YES	YES

注：该表是基于式（2）的回归结果，以2013年为参照，检验各年实验组和控制组企业绿色专利水平的差异。

附表3

匹配平衡性检验结果

变量	样本	平均值		标准偏差%	<i>t</i> 检验	
		实验组	控制组		<i>t</i> 值	<i>p</i> 值
<i>Age</i> 上市公司年龄	匹配前	2.602	2.602	0.000	-0.020	0.984
	匹配后	2.601	2.599	0.400	0.250	0.806
<i>Soe</i> 企业性质	匹配前	0.425	0.392	6.800	4.290	0.000
	匹配后	0.422	0.413	1.800	1.070	0.286
<i>Size</i> 公司规模	匹配前	21.916	21.784	10.200	6.470	0.000
	匹配后	21.910	21.894	1.200	0.720	0.474
<i>Density</i> 资本密集度	匹配前	3.508	3.255	24.900	15.900	0.000
	匹配后	3.496	3.474	2.100	1.270	0.204
<i>Lev</i> 杠杆率	匹配前	0.433	0.400	15.600	9.920	0.000
	匹配后	0.431	0.431	0.000	-0.030	0.980
<i>Profit</i> 净利润	匹配前	20.966	20.999	-8.800	-5.600	0.000
	匹配后	20.973	20.966	1.800	1.150	0.251
<i>Liquid</i> 流动比例	匹配前	2.528	2.731	-6.700	-4.220	0.000
	匹配后	2.542	2.585	-1.400	-0.840	0.403
<i>Cflow</i> 现金流水平	匹配前	0.043	0.050	-9.700	-6.150	0.000
	匹配后	0.043	0.042	1.600	1.000	0.317
<i>Roe</i> 资产净收益率	匹配前	0.056	0.075	-15.200	-9.680	0.000
	匹配后	0.058	0.055	2.400	1.400	0.161
<i>Eps</i> 盈利能力	匹配前	0.317	0.413	-19.600	-12.400	0.000
	匹配后	0.322	0.321	0.200	0.130	0.898
<i>Inde</i> 独立董事比例	匹配前	0.371	0.370	2.300	1.470	0.141
	匹配后	0.371	0.372	-2.500	-1.480	0.138
<i>Share1</i> 第一大股东持股比例	匹配前	0.356	0.360	-2.900	-1.820	0.069
	匹配后	0.356	0.351	3.000	1.820	0.069

附表4

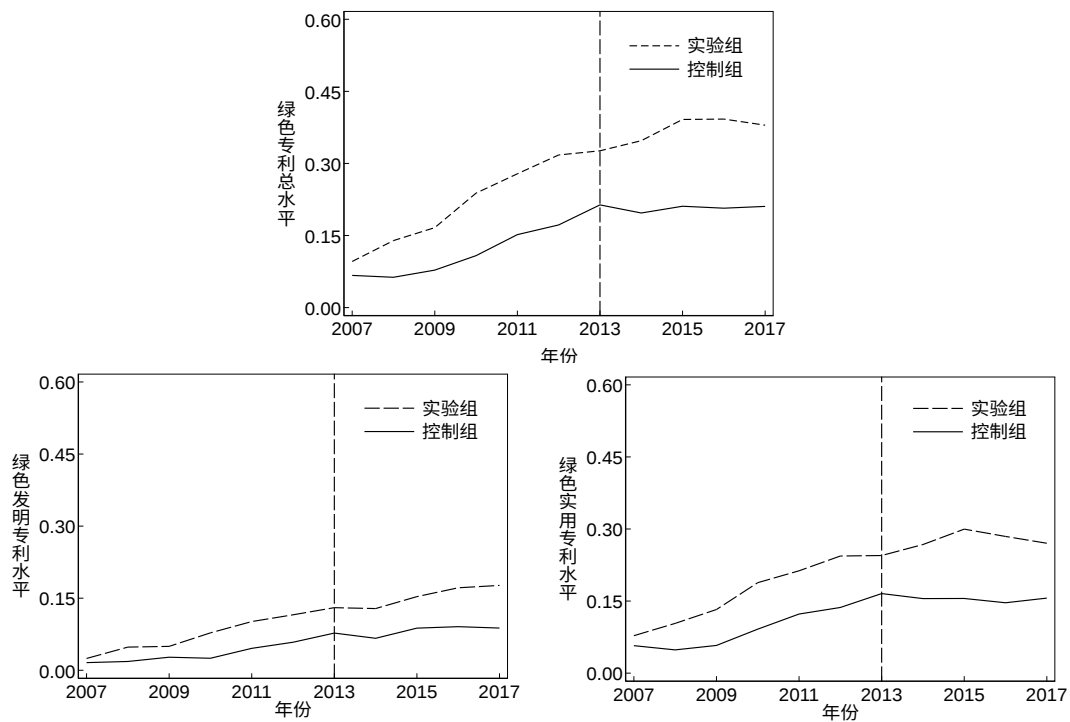
“大气十条”政策与上市公司绿色创新

变量	<i>Tgreen</i>	<i>Igreen</i>	<i>Ugreen</i>
----	---------------	---------------	---------------

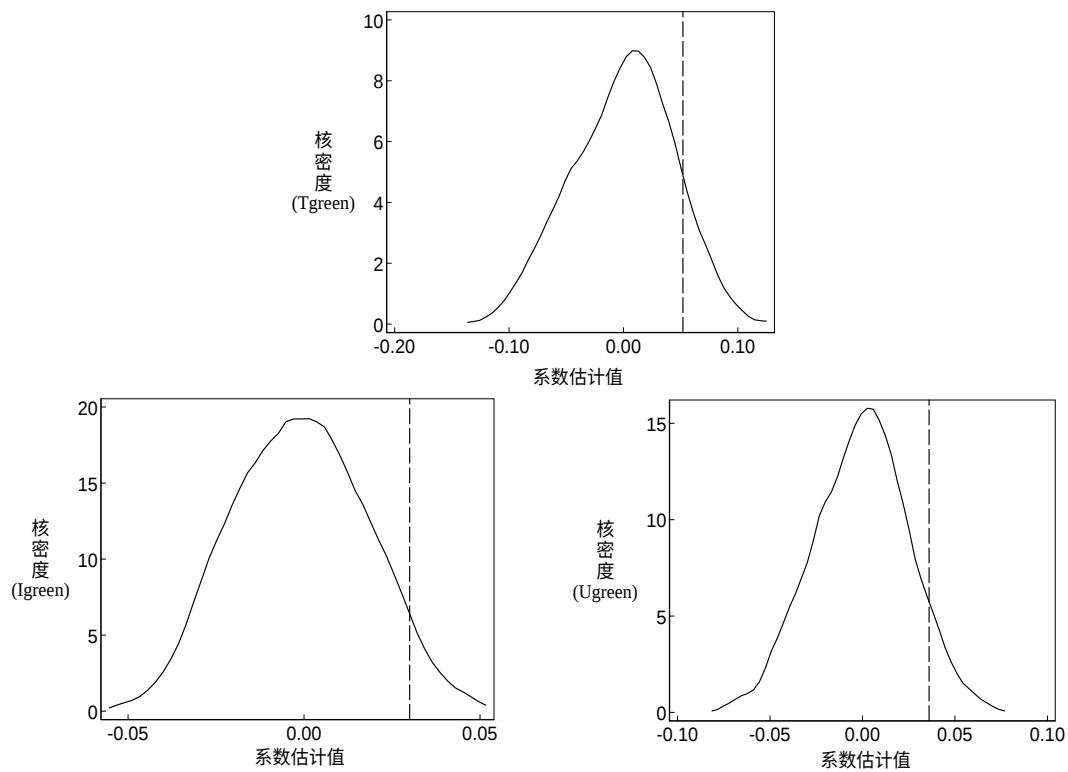
	(1)	(2)	(3)
<i>PolluxPost</i>	0.050*** (0.018)	0.029*** (0.010)	0.034** (0.016)
<i>Age</i>	0.064** (0.031)	0.036* (0.019)	0.039 (0.031)
<i>Soe</i>	0.048* (0.025)	0.028** (0.012)	0.030 (0.021)
<i>Size</i>	-0.944*** (0.191)	-0.536*** (0.117)	-0.751*** (0.162)
<i>Size_Squ</i>	0.023*** (0.004)	0.013*** (0.003)	0.018*** (0.004)
<i>Density</i>	-0.003 (0.009)	0.002 (0.006)	-0.004 (0.009)
<i>Lev</i>	0.118 (0.123)	0.035 (0.081)	0.146 (0.107)
<i>Lev_Squ</i>	-0.103 (0.120)	-0.010 (0.073)	-0.141 (0.108)
<i>Profit</i>	0.015 (0.020)	0.000 (0.014)	0.015 (0.020)
<i>Liquid</i>	0.000 (0.002)	-0.000 (0.001)	0.001 (0.002)
<i>Cflow</i>	0.003 (0.060)	-0.008 (0.029)	0.019 (0.054)
<i>Roe</i>	-0.025 (0.033)	-0.002 (0.021)	-0.022 (0.032)
<i>Eps</i>	-0.029** (0.014)	-0.010 (0.009)	-0.021* (0.012)
<i>Inde</i>	0.050 (0.089)	0.002 (0.061)	0.040 (0.077)
<i>Share1</i>	-0.152** (0.062)	-0.044 (0.036)	-0.138*** (0.049)
<i>Constant</i>	9.525*** (2.175)	5.600*** (1.289)	7.511*** (1.822)
观测值	15783	15783	15783
R ²	0.643	0.503	0.586
企业固定效应	YES	YES	YES
时间固定效应	YES	YES	YES

注：该表与基本回归表2的区别在于引入了企业规模（*Size*）的平方项（*Size_Squ*）。

三、附图



附图 1 绿色专利平行趋势



附图 2 安慰剂检验

注：图中核密度下方的 T_{green} 、 I_{green} 、 U_{green} 分别表示因变量是绿色专利总水平、绿色发明专利水平和绿色实用新型专利水平。

垂直的虚线表示表3中“正确”的估计系数。

附件部分参考文献

- [1] Athey S, Tibshirani J, Wager S. Generalized Random Forests [J]. The Annals of Statistics, 2019, 47(2): 1148-1178.
- [2] Breiman L. Bagging Predictors [J]. Machine Learning, 1996, 24(2): 123-140.
- [3] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32.
- [4] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees. Wadsworth [J]. Biometrics, 1984, 40(3): 358-361.
- [5] Nie X, Wager S. Quasi-oracle Estimation of Heterogeneous Treatment Effects [J]. Biometrika, 2021, 108(2): 299-319.
- [6] Robinson P M. Root- N-Consistent Semiparametric Regression [J]. Econometrica, 1988, 56(4): 931-954.
- [7] Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests [J]. Journal of the American Statistical Association, 2018, 113(523): 1228-1242.