合成控制法中的安慰剂检验改进研究*

——基于准标准化转换的统计推断

连玉君　李　鑫

内容提要：合成控制法的统计推断主要基于以置换检验（permutation test）为基本思想的安慰剂检验。为了应对干预前拟合欠佳对统计推断的影响，前期文献往往需要人为删除部分拟合欠佳的样本，存在较强的主观性。本文通过准标准化转换来惩罚安慰剂检验过程中的噪音成分，提升干预后时段政策效果的可比性，以在不删除观察值的情况下实施安慰剂检验。研究表明，准标准化转换具有如下优势：首先，能有效提升干预前控制组与实验组的可比性，减少噪音成分对检验结果的影响，蒙特卡洛模拟分析证实了该结论的稳健性；其次，基于转换样本的安慰剂检验无需人为删除拟合欠佳的控制组样本，使检验结果更为客观；最后，以转换样本为基础，可以使用Bootstrap法构造各个时点上政策效果的置信区间，从而使合成控制法的统计推断框架与传统统计推断保持一致。

关键词：合成控制法；安慰剂检验；自抽样；置信区间

DOI: 10.19343/j.cnki.11–1302/c.2022.08.008

中图分类号：C81　　**文献标识码：**A　　**文章编号：**1002–4565(2022)08–0115–14

Improved Placebo Test for Synthetic Control Method: Inference with Quasi-standardized Transformation

Lian Yujun & Li Xin

Abstract: The inference of the synthetic control method mainly relies on the “placebo test”, which is based on the permutation test. However, because of the poor fitting in the pre-intervention period which results in ineffective inference, previous studies choose to subjectively delete the control units with poor pre-intervention fitting. This article introduces a process of quasi-standardized transformation to eliminate the noise component in the placebo test in order to improve the comparability of treatment effects in the post-intervention period and to ensure that we can implement placebo tests without subjectively deleting any units. We find that our method has several advantages: Firstly, it can effectively improve the comparability of control units and treated units during the pre-intervention period, reducing the influence of noise components on the test result. Our Monte Carlo simulation results further proves the robustness of this conclusion. Secondly, itavoids the manual deletion of the poor fitting control unitsso that the inference can be more objective. Finally, we can use the data, transformed by the quasi-standardized processing, and the Bootstrap method to construct the confidence interval of the treatment effects, thereby ensuring that our statistical inference framework for the synthetic control

*基金项目：国家自然科学基金面上项目“中国上市公司财务决策中的同群效应研究”（71672206）；广东省自然科学基金面上项目“股权质押的风险传染与防范措施研究”（2020A1515011291）。

method can be consistent with traditional inference methods.

Key words: Synthetic Control Method; Placebo Test; Bootstrap; Confidence Interval

一、引言

合成控制法（synthetic control method，SCM）由Abadie等（2010）提出，被视为过去十五年内政策评估文献中最具创新的估计方法（Athey和Imbens，2017）。其基本思想是，用控制组样本的加权平均（合成实验组）估计实验组的“反事实”结果，进而用政策干预后实验组与合成实验组的差异来定量评估政策效果。

尽管合成控制法的应用日益广泛，但其小样本特征使得传统统计推断方法存在严重的偏误。因此，自Abadie等（2010）提出该方法以来，文献中主要采用以置换检验（permutation test）为基本思想的安慰剂检验来进行统计推断。假设控制组中的每个个体均经历了“政策冲击”（事实上没有），此时针对其估计出的“政策效果”只包括“安慰剂效应”（随机噪音），而针对实验组估计的“政策效果”则同时包括“安慰剂效应”和“实际政策效果”。如果“实际政策效果”不为零，则针对实验组估计出的“政策效果”应该与基于控制组个体估计出的“政策效果”存在显著差异。

但后续研究发现，上述检验方法存在两个问题。一是“政策干预的不可比性”。部分控制组个体在干预前（政策发生前）拟合欠佳，导致干预后（政策发生后）的安慰剂检验包含了过多的噪音成分，从而得到过多“虚假”的统计检验结果（Ferman和Pinto，2017）。二是“临界值缺失问题”。为了解决可比性问题，文献中多采用Abadie等（2010）的做法，删除那些在干预前拟合较差的控制组个体，但由于删除标准缺少理论基础，因此具有很强的主观性[①]。

本文提出的准标准化转换方法可以避免上述局限，在不删除样本的情况下进行统计推断。本文借鉴了加权最小二乘法的思路：对干预前拟合欠佳的控制组个体（噪音成分较高）赋予较小的权重，以提高实验组与控制组的可比性。具体而言，本文的转换方式为 $Gap^{s}_{jt}=\varphi_j\times Gap_{jt}$，其中 $\varphi_j=1\big/RMSPE^{Pre}_{j}$，$Gap_{jt}$ 为个体 j 在第 t 时点的政策效果，$RMSPE^{Pre}_{j}$ 为干预前的预测均方根误差[②]。由于上述变换并不是严格的标准化系数的处理过程，本文将 Gap^{s}_{jt} 称为准标准化转换效应（quasi-standardized treatment effects），简称为qsd-TEs。

本文研究结果表明，准标准化转换能够有效提升控制组在干预前的拟合程度，尤其是对于干预前噪音较大的控制组个体，提升效果十分显著。同时，蒙特卡洛模拟分析从噪音成分大小和权重分布角度检验了准标准化方法的有效性。

本文的贡献主要体现在以下几个方面。其一，本文提出的准标准化方法能够有效降低安慰剂检验过程中由于控制组拟合欠佳导致的噪音，增强了实验组与控制组之间的可比性，避免人为删除干预前拟合欠佳的样本，以克服前期文献（如Abadie等，2010）中普遍存在的“主观挑选（cherry-picking）”问题。其二，本文可以基于准标准化转换后的修正数据，使用Bootstrap来构造政策效果的置信区间，实现大样本条件下的统计推断。本文的改进方法与Abadie等（2010）和Abadie等（2015）文中使用的统计量——“预测均方误差比”本质上具有相似之处[③]，但是后者只能从整体上评估政策效应的统

---

①附录 Part I中使用 Abadie 等（2010）禁烟法案为例，进一步阐述了政策效果不可比性和临界值缺失问题。因篇幅所限，见《统计研究》网站所列附件。下同。

②Abadie 等（2010）中的 Gap_{jt} 是指实验组实际人均香烟消费量与控制组加权人均香烟消费量之差，用于衡量干预后时期的政策效果以及干预前时期的拟合效果。关于 Gap_{jt} 与 Gap^{s}_{jt} 更为一般的分析，详见本文第三部分。

③Abadie 等（2010）使用干预后与干预前预测均方误差之比作为安慰剂检验的统计量：$MRatio=MSPE^{Post}_{j}\big/MSPE^{Pre}_{j}$。而 Abadie 等（2015）使用干预后与干预前预测均方误差根之比作为安慰剂检验统计量：$RMRatio=RMSPE^{Post}_{j}\big/RMSPE^{Pre}_{j}$。

计显著性，而本文的方法则能够动态展现干预后各时点的显著性水平，避免了“预测均方误差比”指标易受个别时点上的离群值影响的局限，以及政策效果方向差异导致的无效比较。同时，本文的方法也可以视为对前期文献中提到的基于Bootstrap构造置信区间的方法（如Rudholm等，2022）的修正：经由准标准化转换后的数据包含较少的噪音，更符合Bootstrap抽样的假设条件。

后文结构安排如下：第二部分为文献综述；第三部分为准标准化安慰剂检验模型的设定；第四部分为准标准化转换有效性检验；第五部分为置信区间的构造；最后总结全文。

二、文献综述

在政策评价领域，经常遇到实验组中只有一个或少数几个实验对象的情形，此时双重差分法（DID）和断点回归（RDD）等检验方法不再适用，而合成控制法则独具优势。然而，小样本下的统计推断却成为一个棘手问题。以Abadie等（2010）为代表的一系列文献均采用安慰剂检验进行统计推断。为了避免控制组样本在干预前拟合欠佳导致的偏误，Abadie等（2010）建议删除这些欠佳的样本。具体过程是以加州（California，简称CA）（实验组）在干预前时段内的预测均方误差为判断基准，记为 $MSPE_{CA}$，设定一个临界倍数 k（取值为20，10，5，2等），若控制组第 j 个州的预测均方误差满足 $MSPE_j > k \times MSPE_{CA}$，则认为其噪音过大，并予以删除。最终基于上述经过删减处理后的剩余样本进行统计推断。显然，k 的取值对统计推断结果有实质性影响，但其选取依据却非常主观，研究者可以任意选择一个适当 k 值，以便结果显著，这将导致假设检验有失客观性。同时，随着 j 的取值不断降低，当剩余样本数小于10时，会出现安慰剂检验与经验 p 值检验的结论相互矛盾的情况①。

对此，现有文献主要从两个角度进行改进。一部分学者尝试通过对现有统计量进行修正，以提高纠偏后的统计量在小样本排序检验中的表现；另一部分学者则通过修正政策效应估计值的分布形式来构造置信区间，回归到传统的大样本统计推断框架下。

鉴于Abadie等（2010）在删除拟合欠佳样本时过于主观，Hollingsworth和Wing（2020）借鉴匹配方法中衡量协变量在实验组和控制组之间均衡度的 $Cohen'D$ 指数，来衡量安慰剂检验中干预前的拟合程度，并基于经验法则将 $Cohen'D > 0.25$ 设定为拟合欠佳。Abadie等（2015）则采用干预后与干预前的预测均方根误差之比，作为检验统计量，通过构造经验分布来进行统计推断。其基本思路在于，若干预前拟合较差（噪音成分多），则干预后的噪音也会较多，采用比值的方法可以在很大程度上降低偏误。该方法的好处是无需删除拟合欠佳的样本，但局限在于只能从总体上判断，会受到个别年份离群值的影响。

如何获取政策效果估计值的置信区间也是学者们着力解决的问题。Li（2020）通过对干预前样本进行二次抽样（subsampling），进而使用Bootstrap获得平均处理效应的渐进分布，以构造置信区间。但是，该方法要求干预前的时间跨度要足够长，否则容易引发高维权重问题，导致平均处理效应偏离 t 分布或者正态分布。对此，Chernozhukov等（2020）建议使用K折交叉拟合法（K-fold cross-fitted）予以修正，以便利用 t 分布构造置信区间。上述两种方法仅适用于估计平均处理效应的置信区间，也有部分文献尝试构造时点政策效果的置信区间，如Kim等（2020）和Rudholm等（2022）分

①在附录 Part II中，本文提供了 Abadie 等（2015）关于两德统一的案例。在子图 A2（b）中，尽管安慰剂检验认为两德统一显著降低了西德的人均 GDP 水平，但是经验 p 值=1/5，大于传统显著性水平，如 0.10、0.05 和 0.01，认为并不存在显著的政策效果，因此面临显著性判断尴尬的情形。

别使用了贝叶斯估计法和Bootstrap方法，而Firpo和Possebom（2018）则将排序检验结果引入处理效应方程中。

不同于上述直接修正假设检验过程的做法，也有文献通过间接提升控制组干预前的拟合程度，在一定程度上解决样本主观删除的问题。基本观点是，合成控制法干预前拟合欠佳主要源于因子模型误设和权重取值范围限制。对于前者，模型中因子变量的不可观测性及其对结果变量的异质性影响都可能导致干预前拟合欠佳。解决方法是改进非线性模型的设定方式和估计方法，如在预测变量中加入干预前的结果变量（Abadie等，2015）、采用组内变换（demean）消除因子载荷中不可观测的个体效应（Ferman和Pinto，2021）、利用分位数回归的残差来修正由因子载荷异质性所导致的拟合欠佳问题（Chen，2020）。然而，当干预前的结果变量无法替代因子载荷时，可以采用时间多项式拟合的方式纠正结果变量中的随机冲击因素，使得结果变量成为因子载荷变量的近似替代，其过程类似于两阶段最小二乘法（2SLS）（Powell，2018）。当存在变量测量误差时，可以采用主成分分析法（PCA）降低偏误（Agarwal等，2020）。当因子模型具有非线性特征时，需要在权重估计中引入惩罚项来控制可观测变量的非线性形式导致的插值偏差问题（interpolation bias）（Abadie等，2010；Kellogg等，2020；Abadie和L'Hour，2021）。如果进一步放松权重估计过程中的线性映射假设，则需要使用非参合成控制法（nonparametric synthetic control method，non-SCM）来估计最优权重（Cerulli，2019）[①]。对于后者，需要引入正则化条件（如弹性网、Lasso或岭回归），以允许实验组可以映射到控制组凸组合以外，从而提升干预前的拟合程度[②]。Doudchenko和Imbens（2016）提出使用弹性网作为目标函数的惩罚项，尽可能减少对权重取值的限制（如允许权重之和不为1）。而作为该模型的特例，Lasso合成控制法（synthetic control using lasso，SCUL），以及基于岭回归的扩展合成控制法（ridge-augmented synthetic control method）则放松了权重非负的限制（Hollingsworth和Wing，2020；Ben-Michael等，2021）。

综上所述，尽管通过修正因子模型设定和放松权重设定能够在一定程度上解决控制组干预前拟合欠佳的问题，但往往需要引入一些新的假设条件或是采用更为复杂的估计方法，致使相应的检验方法仅适用于特定场景。相比之下，Abadie等（2015）引入的干预后与干预前的预测均方根误差之比反而是一种简单明了的处理方式[③]。其本质是采用干预前的RMSPE的倒数作为权重，对干预后的RMSPE进行标准化处理，从而保证不同个体之间的可比性。本文借助这一思想，进一步提出准标准化安慰剂检验的模型设定。

三、准标准化安慰剂检验模型设定

（一）合成控制法

本文首先回顾合成控制法政策效果的估计过程，便于读者更好地理解安慰剂检验中的误差来源，以及准标准化处理的思路。假设样本中包含 $J+1$ 个个体，共 T 期数据。实验个体标记为 $j=1$，而控制组个体标记为 $j=2,\cdots,J+1$。政策干预发生在第 T_0 期，记干预前（Pre）时间集合为 $T_0 \subseteq \{1,\cdots,T_0\}$，干预后（Post）时间集合为 $T_1 \subseteq \{T_0+1,\cdots,T\}$。假设 Y_{jt}^I 和 Y_{jt}^N 分别表示“受到”和“未受到”政策干预两种情况下的“潜在结果”，则观察到的结果变量可以表示为 $Y_{jt}=(1-D_{jt})Y_{jt}^N+D_{jt}Y_{jt}^I$，其中，

①非参数合成控制法对于带宽和核函数的选择具有一定的敏感性，可能会出现过拟合的问题。

②本文使用一个案例简单说明这一过程。假设研究对象是某一事件冲击对美国 GDP 的影响，那么在权重非负且和为 1 的条件下，任何国家加权平均都无法拟合事件发生前美国 GDP 的变动轨迹，因而造成干预前拟合状况较差（或者是存在严重的拟合欠佳问题）。

③该统计量定义为 $Ratio=RMSPE_1/RMSPE_0$，其中，$RMSPE_0$ 和 $RMSPE_1$ 分别为干预前和干预后时段的预测均方根误差。

$D_{jt}=1\{t\in T_1, j=1\}$。因此，实验个体（$j=1$）在干预后第 t 时点（$t\in T_1$）的处理效应如下[①]：

$$Gap_{1t}^{Post}=\alpha_{1t}=Y_{1t}^{I}-Y_{1t}^{N} \tag{1}$$

显然，Y_{1t}^{N} 不可观测，Abadie等（2010）建议使用控制组个体的加权平均值估计：

$$\hat{Y}_{1t}^{N}=\sum_{j=2}^{J+1}\omega_j^{*}Y_{jt} \tag{2}$$

由于实验组和控制组个体可能来源于异质性分布，最优权重 ω_j^{*} 难以完全复现实验组“潜在结果”Y_{1t}^{N} 的变动轨迹，使得政策效果 Gap_{1t}^{Post} 存在偏误问题。为了更为直观呈现政策效果的偏误形式，本文参照Abadie等（2010），采用如下因子模型刻画结果变量 Y_{jt}^{N} 的数据生成过程：

$$Y_{jt}^{N}=\delta_t+\boldsymbol{\theta}_t\boldsymbol{Z}_j+\boldsymbol{\lambda}_t\boldsymbol{\mu}_j+\varepsilon_{jt} \tag{3}$$

其中，δ_t 为时间趋势项，$\boldsymbol{Z}_j$ 和 $\boldsymbol{\mu}_j$ 分别为可观测和不可观测解释变量构成的向量，$\boldsymbol{\theta}_t$ 和 $\boldsymbol{\lambda}_t$ 为对应的时变参数向量[②]，ε_{jt} 为不可观测的随机冲击。

将式（3）和式（2）带入式（1），得到干预后各时点（$\forall t\in T_1$）的政策效果：

$$\begin{aligned}\widehat{Gap}_{1t}^{Post}=&\alpha_{1t}-\sum_{j=2}^{J+1}\omega_{jt}^{*}\alpha_{jt}\\&+\underbrace{\underbrace{\boldsymbol{\theta}_t\left(\boldsymbol{Z}_1^{Post}-\sum_{j=2}^{J+1}\omega_j^{*}\boldsymbol{Z}_j^{Post}\right)}_{\text{A}}+\underbrace{\boldsymbol{\lambda}_t\left(\boldsymbol{\mu}_1^{Post}-\sum_{j=2}^{J+1}\omega_j^{*}\boldsymbol{\mu}_j^{Post}\right)}_{\text{B}}+\underbrace{\left(\varepsilon_{1t}^{Post}-\sum_{j=2}^{J+1}\omega_j^{*}\varepsilon_{jt}^{Post}\right)}_{\text{C}}}_{\eta_{1t}^{Post}}\end{aligned} \tag{4}$$

其中，α_{1t} 为真实的政策效果，$\sum_{j=2}^{J+1}\omega_{jt}^{*}\alpha_{jt}$ 为控制组政策效果干扰项，A、B和C分别为干预后可观测的预测变量、不可观测的因子载荷和随机冲击所造成的估计偏误。不妨设 η_{1t}^{Post}（=A+B+C）表示干预后时段的噪音成分[③]。

实验组在干预前（$\forall t\in T_0$）未受政策影响，其政策效果 $\alpha_{1t}=\sum_{j=2}^{J+1}\omega_{jt}^{*}\alpha_{jt}=0$，换言之，干预前的政策效果 $\widehat{Gap}_{1t}^{Pre}$ 仅包含噪音成分：

$$\widehat{Gap}_{1t}^{Pre}=\underbrace{\underbrace{\boldsymbol{\theta}_t\left(\boldsymbol{Z}_1^{Pre}-\sum_{j=2}^{J+1}\omega_j^{*}\boldsymbol{Z}_j^{Pre}\right)}_{\text{A}'}+\underbrace{\boldsymbol{\lambda}_t\left(\boldsymbol{\mu}_1^{Pre}-\sum_{j=2}^{J+1}\omega_j^{*}\boldsymbol{\mu}_j^{Pre}\right)}_{\text{B}'}+\underbrace{\left(\varepsilon_{1t}^{Pre}-\sum_{j=2}^{J+1}\omega_j^{*}\varepsilon_{jt}^{Pre}\right)}_{\text{C}'}}_{\eta_{1t}^{Pre}} \tag{5}$$

其中，A′、B′和C′的含义与A、B和C相同，三者之和为干预前的噪音成分，记为 η_{1t}^{Pre}（=A′+B′+C′）[④]。

为了使式（4）中的政策效果估计值 $\widehat{Gap}_{1t}^{Post}$ 是真实的政策效果 α_{1t} 的无偏且有效估计量，Abadie等（2010）、Firpo和Possenbom（2018）提出以下三个假设条件。

假设1：干预前、后时段的噪音成分不存在显著差异，即 $\eta_{1t}^{Post}=\eta_{1t}^{Pre}$。

如果假设1不满足，即干预前、后的噪音成分存在显著差异，那么即使选取的最优权重能够使干预前的噪音影响 $\eta_{1t}^{Pre}\to 0$，也无法保证干预后的噪音干预成分 $\eta_{1t}^{Post}\to 0$，此时政策效果估计值就存在较为严重的外推预测偏误（Abadie，2021）。

假设2：政策效果并不存在溢出效应。

在式（4）中，如果控制组中的个体也直接或间接受到了政策干预，那么控制组政策效果干扰项

---

①在式（1）中，为了与前期文献保持一致，在不至于混淆的情况下，同时用 Gap_{1t}^{Post} 和 α_{1t} 表示政策效应。

② $\boldsymbol{\lambda}_t\boldsymbol{\mu}_j$ 的作用在于允许每个个体可以有不同的时间趋势，若 $\boldsymbol{\lambda}_t$ 为常数，式（3）就是一个常规的 DID 模型。

③合成控制法中的噪音干预与 Rubin（1974）因果模型中的选择性偏差均来源于实验组与控制组之间的个体差异。

④为了与式（4）表述相一致，本文仍使用 $\widehat{Gap}_{1t}^{Pre}$ 与 η_{1t}^{Pre} 分别表示政策效果和噪音成分，但二者数值上相同。为了便于在下文进行区分，仅在 η_{1t}^{Post} 出现的地方对应使用 η_{1t}^{Pre}，其余部分均使用 $\widehat{Gap}_{1t}^{Pre}$ 表示干预前噪音。

$\sum_{j=2}^{J+1}\omega_j^*\alpha_{jt}\neq 0$。此时，即使假设1成立且存在最优权重使得$\eta_{1t}^{Post}=0$，政策效果为$\alpha_{1t}-\sum_{j=2}^{J+1}\omega_j^*\alpha_{jt}$，而非真实值$\alpha_{1t}$。

假设3：样本集中每个个体受到政策干预的概率相同。

若实验组是出于特定原因被选中的，即实验组被政策干预的概率明显高于控制组，则会产生样本选择偏误问题，致使干预后政策效果的分布难以满足费舍尔精准检验法的对称性要求，导致实验组与控制组之间不具可比性（Hahn和Shi，2017；Firpo和Possebom，2018）①。

（二）安慰剂检验模型的不足

合成控制法主要用于估计单个或少数几个实验组的政策效果，样本量通常都很小，这对于统计推断提出了挑战。为了获得政策效果的显著性水平，Abadie等（2010）和Abadie等（2015）提出了基于小样本条件下的安慰剂检验，本质上类似于费舍尔精准检验（Fisher，1936；Imbens和Rubin，2015）。通过迭代的方式对控制组个体与实验组个体进行相同的合成控制（Abadie等，2010；Abadie等，2015），获得干预后各时点实验组真实政策效果$\widehat{Gap}_1^{Post}=\left(\widehat{Gap}_{1,T_0+1}^{Post},\widehat{Gap}_{1,T_0+2}^{Post},\cdots,\widehat{Gap}_{1,T}^{Post}\right)'$，以及控制组“虚拟”政策效果$\widehat{Gap}_j^{Post}=\left(\widehat{Gap}_{j,T_0+1}^{Post},\widehat{Gap}_{j,T_0+2}^{Post},\cdots,\widehat{Gap}_{j,T}^{Post}\right)'$，其中$j=2,3,\cdots,J+1$。在满足假设3的前提下，通过“图形可视化”或“经验 p 值”的方式进行显著性判断。如果干预后$\widehat{Gap}_{1t}^{Post}$大于任意$\widehat{Gap}_{jt}^{Post}$，那么认为该政策对于实验组产生了显著影响，否则，则认为该政策可能是一个偶然事件。

但是，受制于权重估计过程中，控制组使用了与实验组几乎相同的预测变量矩阵（$\boldsymbol{Z}$）和样本组合（donor pool），因而权重（ω_j^*）可能无法使其实现与实验组相似的干预前拟合效果②。这导致控制组的噪音成分η_{jt}^{Post}和η_{jt}^{Pre}不为零，致使干预后控制组的“虚拟”政策效果（$\widehat{Gap}_{jt}^{Post}$）与实验组的真实政策效果（$\widehat{Gap}_{1t}^{Post}$）之间不具可比性③。

为了解决这一问题，本文提出使用准标准化转换的方式降低干预后噪音干预。其核心思想是根据实验组和控制组干预前的拟合状况，对拟合欠佳的个体施加严格的惩罚（赋予较小权重），而对拟合较好个体降低惩罚力度，以期实现降低干预后噪音成分对显著性水平的扭曲。

（三）准标准化模型构建

在合成控制法中，Abadie等（2010）使用干预前的预测均方根误差（*RMSPE*）来衡量式（5）中的$\widehat{Gap}_{1t}^{Pre}$是否接近于零，表示如下：

$$RMSPE_1^{Pre}=\sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\widehat{Gap}_{1t}^{Pre}\right)^2}=\sqrt{MSPE_1^{Pre}} \tag{6}$$

作为干预前拟合程度的统计量，$RMSPE_1^{Pre}$最大限度地包含了引起干预前偏差的因素。如果使用$RMSPE_1^{Pre}$对干预前各时点的噪音成分（$\widehat{Gap}_{1t}^{Pre}$）进行惩罚，则能够很好地降低干预前噪音成分的影响。

在满足假设1的前提下，本文对式（4）和式（5）同时进行$RMSPE_1^{Pre}$的惩罚，以便修正干预前的拟合程度，例如，对实验组的准标准化转换如下：

---

①Firpo 和 Possenbom（2018）提出使用敏感性系数分析法，检验安慰剂检验过程是否满足均匀分配的特征。

②协变量非平衡性问题导致实验组的个体特征无法通过控制组的个体特征加权平均获得，因而使得其干预前的拟合程度降低。在第四部分蒙特卡洛模拟分析部分会对此问题做更深入的讨论。

③附录中的子图 A1（a）和子图 A1（b）便呈现了这种情形。

$$\widehat{Gap}_{1t}^{Pre,S}=\frac{\widehat{Gap}_{1t}^{Pre}}{RMSPE_1^{Pre}}=\varphi_1\times\widehat{Gap}_{1t}^{Pre} \tag{7}$$

更为一般的表示如下：

$$\widehat{Gap}_{jt}^{Pre,S}=\frac{\widehat{Gap}_{jt}^{Pre}}{RMSPE_j^{Pre}}=\varphi_j\times\widehat{Gap}_{jt}^{Pre} \tag{8}$$

若令惩罚因子 $\varphi_j=\dfrac{1}{RMSPE_j^{Pre}}\in(0,1)$[①]，则其反映的是对个体的惩罚力度（或权重）。当干预前拟合较好时，φ_j 取值较大，代表惩罚力度较小；反之亦然。因此，经过准标准化转换后，干预前各时期噪音成分 $\widehat{Gap}_{jt}^{Pre,S}$ 会大幅度降低，而其拟合程度则会相应提升。

为了更为直观地说明准标准化转换的作用，本文使用Ferman和Pinto（2016，2017）文中的“标准化均方误差指数”来衡量干预前的拟合效果，定义如下：

$$\tilde{R}^2=1-\frac{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt}^{Pre}-Y_{jt}^{Pre,N}\right)^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt}^{Pre}-\bar{Y}_j^{Pre}\right)^2}=1-\frac{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\widehat{Gap}_{jt}^{Pre}\right)^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt}^{Pre}-\bar{Y}_j^{Pre}\right)^2} \tag{9}$$

其中，$\left(Y_{jt}^{Pre}-Y_{jt}^{Pre,N}\right)$ 表示个体 j 在干预前的噪音部分，$\left(Y_{jt}^{Pre}-\bar{Y}_j^{Pre}\right)$ 表示个体 j 在干预前各时点上结果变量的离差。易于看出，$\tilde{R}^2$ 的定义和作用都非常类似于传统回归分析中的拟合优度。因此，若 $\tilde{R}^2\to1$，则表示干预前实验组与“反事实”估计结果近似完美拟合，而 $\tilde{R}^2$ 越低，则表示噪音越严重。需要说明的是，当干预前拟合状况很差时，$\tilde{R}^2$ 可能为负值（Ferman和Pinto，2016；Ferman和Pinto，2017）。因此，控制组的 $\tilde{R}^2$ 越接近于1、与实验组的 $\tilde{R}^2$ 越接近，表示控制组与实验组之间的可比性越强。

用式（8）中的 $\widehat{Gap}_{jt}^{Pre,S}$ 替代式（9）中的 $\widehat{Gap}_{jt}^{Pre}$，可得准标准化转换下的 $\tilde{R}_S^2$：

$$\tilde{R}_S^2=1-\frac{1}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt}^{Pre}-\bar{Y}_j^{Pre}\right)^2}\times\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\widehat{Gap}_{jt}^{Pre,S}\right)^2 \tag{10}$$

由于 $\widehat{Gap}_{jt}^{Pre,S}<\widehat{Gap}_{jt}^{Pre}$，可知 $\tilde{R}_S^2>\tilde{R}^2$。因此，利用准标准化转换可以减小样本干预前的噪音（$\widehat{Gap}_{jt}^{Pre}$），以便有效降低干预后噪音成分对政策效果估计值的影响。

四、准标准化转换的有效性检验

为了检验准标准化转换的有效性，本文以Abadie等（2010）文中使用的加州禁烟法案为例进行对比分析，进而采用蒙特卡洛模拟，分析本文方法的稳健性[②]。由于实验组与控制组样本均满足 $RMSPE_j^{Pre}>1$，本文选择的惩罚因子表达式为 $\varphi_j=\dfrac{1}{RMSPE_j^{Pre}}$。

---

①若结果变量是相对指标(如，GDP 增长率或股票收益率)，$RMSPE_j^{Pre}$ 很有可能小于 1，对应的惩罚因子 $\varphi_j>1$，导致干预后的政策效果被放大。为此，在使用准标准化转换时，需先查验样本个体的 $RMSPE_j^{Pre}$ 是否存在小于 1 的情形，若有，则可将惩罚因子设定为 $\varphi_j=\dfrac{1}{RMSPE_j^{Pre}}$，在不改变 $RMSPE_j^{Pre}$ 相对大小的条件下，保证惩罚因子 $\varphi_j\in(0,1)$ 恒成立。

②需要说明的是，鉴于嵌套法在估计最优权重过程中可能存在不收敛的现象，因此本文的所有分析均未采用嵌套法。这会导致本文的权重估计结果与 Abadie 等（2010）存在一些细微的差异，但对本文的结论不存在实质性影响。

（一）模型拟合效果分析

如前文所述，为了应对干预前拟合欠佳对安慰剂检验的影响，Abadie等（2010）通过删除干预前拟合欠佳的控制组个体来保证控制组与实验组的可比性。表1中列（1）~（6）重现了这一过程，而作为对比，列（7）则呈现了本文提出的准标准化转换方法的拟合效果。

表1　　**禁烟法案干预前拟合效果对比：准标准化转换前后的 $\tilde{R}^2$**

	$MSPE^{Pre}$	$\tilde{R}^2$					$\tilde{R}_S^2$
	（1）	（2）	（3）	（4）	（5）	（6）	（7）
			$k>20$	$k>10$	$k>5$	$k>2$	
Alabama	7.345	0.904	0.904	0.904	0.904	0.904	0.987
Arkansas	6.311	0.929	0.929	0.929	0.929	0.929	0.989
California	**4.153**	**0.968**	**0.968**	**0.968**	**0.968**	**0.968**	**0.992**
Colorado	36.501	0.731	0.731	0.731	–	–	0.993
Connecticut	27.216	−0.723	−0.723	−0.723	–	–	0.937
Delaware	101.00	−1.770	–	–	–	–	0.973
Georgia	1.926	0.948	0.948	0.948	0.948	0.948	0.973
Idaho	7.352	0.950	0.950	0.950	0.950	0.950	0.993
Illinois	10.940	0.829	0.829	0.829	0.829	–	0.984
Indiana	40.640	0.674	0.674	0.674	–	–	0.992
Iowa	15.954	0.798	0.798	0.798	0.798	–	0.987
Kansas	16.221	0.814	0.814	0.814	0.814	–	0.989
Kentucky	681.460	−0.299	–	–	–	–	0.998
Louisiana	5.019	0.953	0.953	0.953	0.953	0.953	0.991
Maine	17.323	0.660	0.660	0.660	0.660	–	0.980
Minnesota	20.014	0.630	0.630	0.630	0.630	–	0.982
Mississippi	6.351	0.905	0.905	0.905	0.905	0.905	0.985
Missouri	4.445	0.883	0.883	0.883	0.883	0.883	0.974
Montana	5.551	0.959	0.959	0.959	0.959	0.959	0.993
Nebraska	5.934	0.869	0.869	0.869	0.869	0.869	0.978
Nevada	59.568	0.888	0.888	–	–	–	0.998
New Hampshire	3485.457	−1.963	–	–	–	–	0.999
New Mexico	5.878	0.897	0.897	0.897	0.897	0.897	0.982
North Carolina	115.303	0.854	–	–	–	–	0.999
North Dakota	20.815	0.864	0.864	0.864	–	–	0.993
Ohio	12.921	0.251	0.251	0.251	0.251	–	0.942
Oklahoma	10.702	0.935	0.935	0.935	0.935	–	0.994
Pennsylvania	8.764	0.778	0.778	0.778	0.778	–	0.975
Rhode Island	162.627	−0.849	–	–	–	–	0.989
South Carolina	3.184	0.960	0.960	0.960	0.960	0.960	0.987
South Dakota	9.949	0.872	0.872	0.872	0.872	–	0.987
Tennessee	7.065	0.914	0.914	0.914	0.914	0.914	0.988
Texas	7.173	0.920	0.920	0.920	0.920	0.920	0.989
Utah	593.764	−14.211	–	–	–	–	0.974
Vermont	89.153	0.622	–	–	–	–	0.996
Virginia	35.102	0.657	0.657	0.657	–	–	0.990
West Virginia	10.296	0.814	0.814	0.814	0.814	–	0.982
Wisconsin	8.747	0.699	0.699	0.699	0.699	–	0.966
Wyoming	114.169	0.630	–	–	–	–	0.997

注：a. $MSPE^{Pre}$ 示干预前的预测均方根误差，见式（6），$\tilde{R}^2$ 和 $\tilde{R}_S^2$ 分别为准标准化转换前和转换后的标准化均方误差指数，对应式（9）和式（10）；b.临界倍数 $k=MSPE_j/MSPE_1$ $(j\geqslant 2)$，参考Abadie等（2010），k 的取值为20、10、5和2；c.列（1）描述的是安慰剂检验中39个州在干预前的 $MSPE$ 值，列（2）是基于列（1）结果估计的 $\tilde{R}^2$ 值，列（3）~（6）是以列（2）为基础，依次删除 $k>20$ 、$k>10$ 、$k>5$ 和 $k>2$ 的控制组个体后，剩余州的 $\tilde{R}^2$，列（7）描述的是经过准标准化转换后的各州标准化均方误差指数（ $\tilde{R}_S^2$ ）。

具体而言，列（1）呈现了各州在干预前时段的预测均方误差 $MSPE^{Pre}$，对比基准是实验组——加州，其 $MSPE^{Pre}$ 为4.153[①]。列（2）呈现了基于式（9）计算而得的 $\tilde{R}^2$。对比列（1）和列（2）可知，若某个实验组个体 $MSPE^{Pre}$ 较大，则其干预前的拟合程度较差，即 $\tilde{R}^2$ 较小。这意味着，两种评价方法本质上具有一致性。列（3）~（6）列呈现了基于Abadie等（2010）给出的筛选标准（删除大于加州 $MSPE$ 20倍、10倍、5倍和2倍后的控制组）选取的控制组样本的 $\tilde{R}^2$[②]。显然，筛选标准越严格，最终保留的控制组个体与实验组在干预前的拟合程度越相近。例如，在列（6）中，控制组个体的平均 $\tilde{R}^2$ 为0.922，已经非常接近加州的0.968。总体而言，虽然Abadie等（2010）对 k 值的设定存在一定的主观性，但这一处理方法确实能够在很大程度上确保控制组和实验组在干预前的拟合程度尽可能相近。同时，使用 $\tilde{R}^2$ 指标能够很好地刻画干预前的拟合程度。

那么，能否在不进行任何人为删除的情况下达到上述效果呢？从列（7）的结果来看，本文提出的准标准化转换可以实现上述目标。这里的 $\tilde{R}_S^2$ 是基于式（10）计算而得。整体来看，控制组的平均 $\tilde{R}_S^2$ 为0.984，与实验组（加州）的 $\tilde{R}_S^2$（0.992）非常接近，表明准标准化转换可以保证两组样本在干预前的拟合程度高度可比。由单个控制组个体的 $\tilde{R}_S^2$ 来看，那些在干预前拟合程度非常差的州（如Kentucky、New Hampshire、Utah等），经由准标准化转换后，得以大幅提高。这是因为在标准化转换过程中，本文对这些州进行了更为严格的惩罚。换个角度来看，Abadie等（2010）所采用的删除法其实也可以视为一种加权惩罚：以 $k=2$ 为例，那些 $MSPE_j^{Pre} > 2MSPE_1^{Pre}$ 的州被赋予的权重为零（删除），而其他控制组个体则被赋予相同的权重（保留）。这其实是一种非此即彼的0/1权重赋值法，而本文的准标准化转换则是连续型权重赋值法：给予每个控制组个体不同的权重。

图1更为直观地展示了上述对比分析结果。其中，子图（a）、（b）、（d）、（e）分别对应Abadie等（2010）文中的图4~7，而子图（f）则是在不删除任何样本的情况下经由准标准化转换后的数据绘制的。显然，能够在不损失样本信息的前提下，得到与Abadie等（2010）几近相同的检验结果。

（二）蒙特卡洛模拟分析

下面将通过两组蒙特卡洛模拟（MC）分析验证本文方法的稳健性。在数据生成过程部分，本文借鉴Flannery和Hankins（2013），以及Chernozhukov等（2020），采用“经验MC”法。该方法能够最大限度地保留现有观测样本的特征，仅生成本文需要重点研究的那些变量对应的随机数，将其添加到现有观测数据中，从而形成模拟数据。与传统的MC方法相比，“经验MC”方法仍然以加州禁烟法案结果为基准，便于比较方法之间的优劣性。本文也使用传统MC方法进行了模拟，其结论并未发生实质性变化。

1.干预前的噪音成分。

由式（5）可知，即使存在最优权重使得A′与B′部分均为零，但若实验组与控制组所受到的随机冲击不同，便会导致干预前时段内的C′不为零。这也是导致部分控制组个体在干预前拟合欠佳使Abadie等（2010）被迫采用删除法的主要原因。本文通过如下方式构造模拟数据来刻画这种情形。首先，生成一个服从标准正态分布的随机变量 $\nu_{1it} \sim N(0,1)$；其次，生成均值为零且服从正态分布的随机数 $\nu_{2it} \sim N(0,2^2)$ 和 $\nu_{5it} \sim N(0,5^2)$；最后，将 ν_{1it} 添加到加州的实际香烟消费量数据中，而随机数 ν_{2it} 和 ν_{5it} 则分别被添加到其他控制组州的实际香烟消费量中，以反映噪音成分的差异。由于本文在控制组个体中额外增加了干扰因素，若按照Abadie等（2010）的处理思路，则需要删除更多的控制

①由式（6）可知，预测均方根误差和预测均方误差两者对实验组干预前噪音大小的衡量标准是一致的。为了与 Abadie 等（2010）文中结果进行比较，本文此处采用预测均方误差进行呈现。

②表 1 中列（2）、（3）、（5）、（6）的结果与 Abadie 等（2010）文中的图 4~7 中使用的筛选标准是一致的。

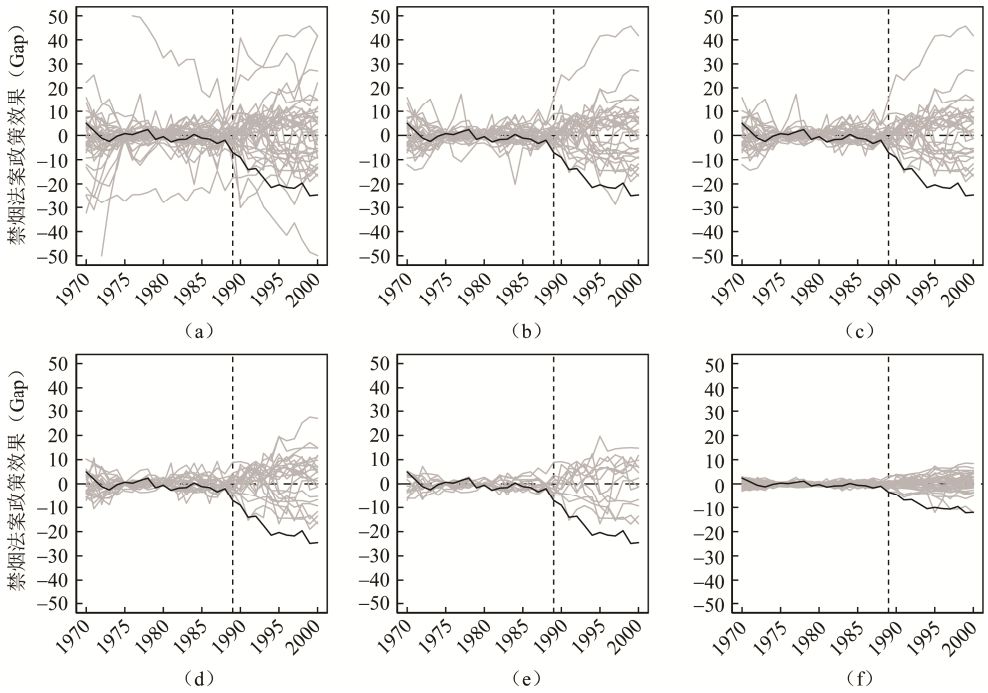

图1　法案安慰剂检验结果

注：图中（a）表示传统的安慰剂检验结果，对应于Abadie等（2010）中的图4，（b）~（e）是在（a）的基础上，依次删除临界倍数 $k>20$、$k>10$、$k>5$ 和 $k>2$ 的控制组个体后的结果，（f）表示准标准化转换后的安慰剂检验结果。图中纵向虚线代表禁烟法案发生时间（1989年），横向虚线代表无政策效果，即 $Gap_{1t}=0$。黑色实线代表加州，灰色实线代表其他州。

组个体才能满足前文提到的可比性要求。而若采用本文的准标准化处理，则不需要删除任何样本，只需对噪音成分较为严重的个体分配较小的权重即可。

图2的模拟结果证实了上述理论预期。其中，子图（a）~（c）是采用Abadie等（2010）文中引入噪音后的传统安慰剂检验结果（未删除任何样本）。对比子图（c）和（a）可知，随着噪音成分的增大，干预前的拟合欠佳问题加重，干预后安慰剂检验结果中的“虚拟”政策效果所包含的噪音成分亦同样增大，导致安慰剂检验结果失效。与上述结果形成鲜明对比的是子图（d）~（f），它们是基于准标准化转换后的数据执行安慰剂检验绘制的。显然，干预前的拟合程度大幅提高，因而实验组和控制组干预后时期的禁烟效果估计值近似满足可比性需求。与此同时，子图（e）和（f）的结果与子图（d）相似，表明准标准化转换对噪音成分的异方差并不敏感，具有很好的稳健性。

2.权重估计偏差。

根据Abadie等（2010）可知，最优权重是通过极小化实验组预测变量矩阵（$\boldsymbol{Z}_1$）与控制组预测变量矩阵 $\boldsymbol{Z}_0$ 之间的距离计算而得的。若部分预测变量存在衡量偏误必然会导致最优权重估计值有偏，这是导致控制组在干预前时段拟合欠佳的另一个来源。为了验证准标准化转换能否应对这种偏差，本文尝试在某些预测变量中人为加入噪音，以模拟不同程度的衡量偏误。具体做法为：生成均值为零且服从正态分布的随机数 $\nu_{2it}\sim N\left(0,2^2\right)$ 和 $\nu_{5it}\sim N\left(0,5^2\right)$，将它们分别添加到其中一个预测变量，如对数人均收入（$lnincome$），构造出新的对数人均收入变量 $lnincome_2$ 和 $lnincome_5$，进而分别用这两个变量替代 $lnincome$，进行合成控制法估计和安慰剂检验[①]。

图3呈现了模拟结果。与图2结果相似，随着对数人均收入水平的衡量偏误逐渐增加，子图（a）~

①模型设定中共包含 7 个预测变量，这里仅展示了基于对数人均收入变量的分析结果。本文也对其他预测变量进行了分析，其结论并未发生实质性变化。

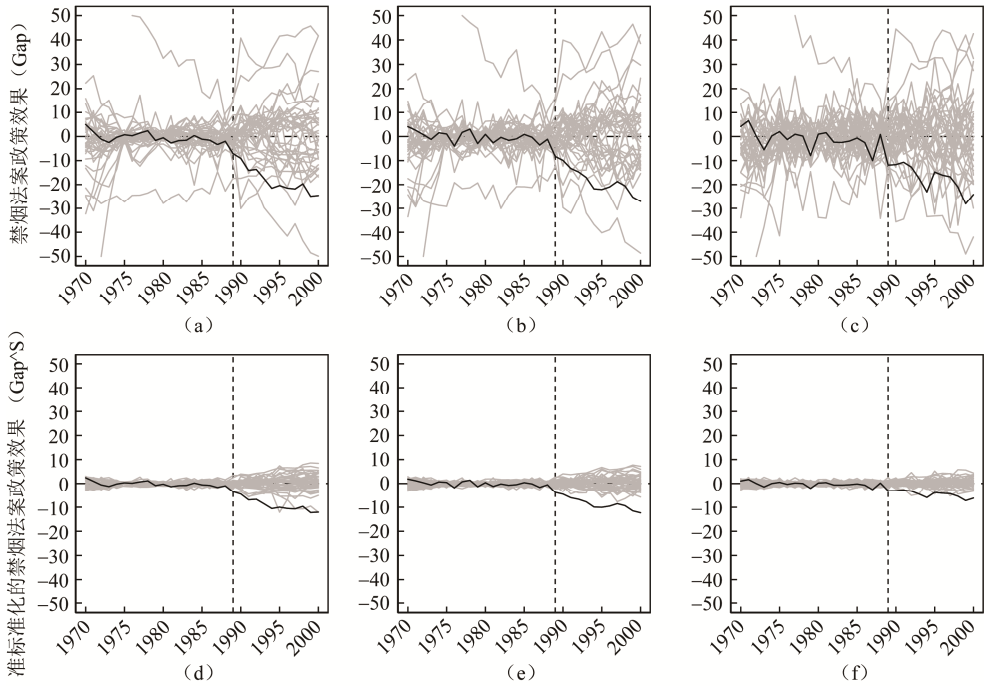

图2　异方差条件下稳健性检验

注：图中（a）、（b）和（c）是未经准标准化转换的安慰剂检验结果。其中（a）是标准的安慰剂检验结果，对应于Abadie等（2010）中的图4结果，作为分析的基准图形。（b）和（c）是控制组州人均香烟消费量分别增加标准差为2和标准差为5的噪音结果。相比而言，（d）、（e）和（f）分别是（a）、（b）和（c）对应经过准标准化转换后的安慰剂检验结果。图中纵向虚线代表禁烟法案发生时间（1989年），横向虚线代表无政策效果，即 $Gap_{1t}=0$ 。黑色实线代表加州，灰色实线代表其他州。

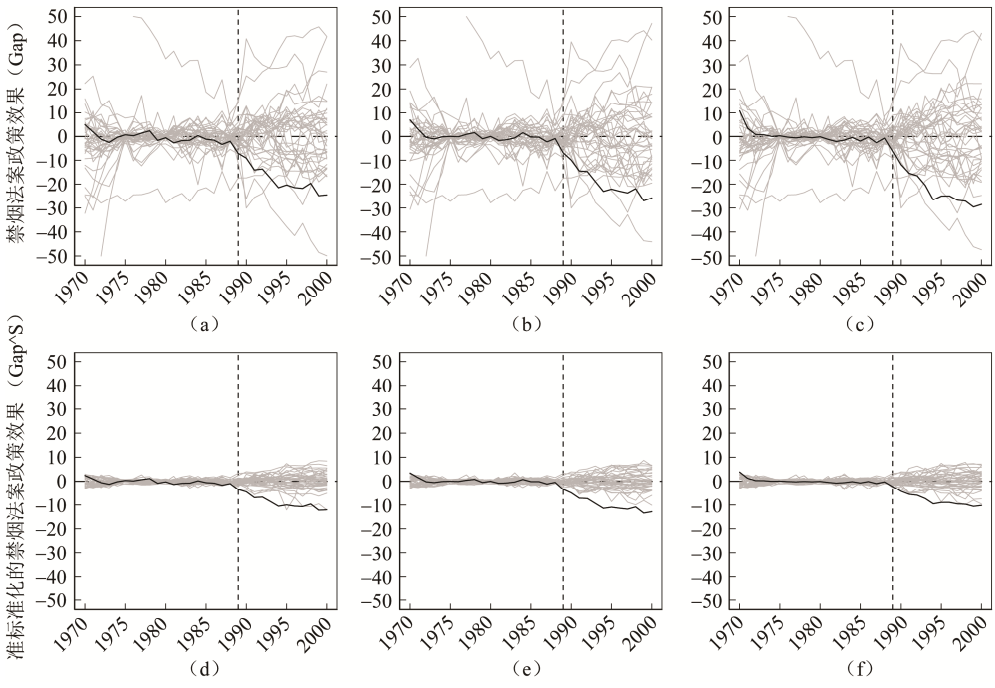

图3　权重估计分布偏差检验

注：图中（a）、（b）和（c）是未经准标准化转换的安慰剂检验结果。其中（a）是标准的安慰剂检验，对应Abadie等（2010）中的图4结果，作为基准分析图形。（b）和（c）是对数人均收入分别增加标准差为2和标准差为5的噪音结果。相比而言，（d）、（e）和（f）分别是（a）、（b）和（c）对应进行准标准化转换后的安慰剂检验结果。图中纵向虚线代表禁烟法案发生时间（1989年），横向虚线代表无政策效果，即 $Gap_{1t}=0$ 。黑色实线代表加州，灰色实线代表其他州。

（c）的干预前拟合欠佳现象有逐渐增强，所以传统安慰剂检验无效。相比而言，子图（d）~（e）是其对应进行准标准化转换后的结果，其干预前时期权重分布和大小偏差所导致的噪音干预得到有效的控制，使得其干预后时期禁烟效果估计值近似满足理想的安慰剂检验条件。子图（e）和（f）的显著性检验结果与子图（d）相似，表明准标准化转换在克服权重估计偏差所导致的噪音影响同时，兼顾满足稳健性需求。

综上所述，准标准化转换后的安慰剂检验方法能够有效地纠正控制组噪音成分过大，或部分预测变量存在衡量偏误导致的干预前拟合欠佳问题，提升了控制组在干预前的拟合程度，因而增强了传统安慰剂检验结果的有效性。

五、政策效果置信区间的构造

自Abadie等（2010）以来，多数文献都采用“图形可视化”或“经验 p 值”的方式来呈现安慰剂检验结果，以确定政策效应是否显著。近期的文献开始尝试估计政策效果的置信区间，以便使用传统的模型评价方法进行分析，如Chernozhukov等（2020），Kim等（2020）等。若样本中的个体在干预前都不存在拟合欠佳问题，则使用去一法（leave-one-out）或自抽样（Bootstrap）即可构造经验样本，进而得到置信区间。根据前文的分析，经由准标准化转换的数据相对更为干净，基本满足这一条件。

因此，本文继续使用Abadie等（2010）禁烟法案的数据，通过准标准化转换和Bootstrap方式获得加州各时点的置信区间。与回归模型不同的是，合成控制法模型无法获得加州各时点的残差估计值，因此无法使用参数Bootstrap的方法，通过抽取残差样本的方式重新构造各州香烟消费量（Hansen，1996；Xu，2017）。为此本文使用非参数的Bootstrap方式进行样本抽样，首先，通过对38个控制组州进行再抽样的方式获取经验样本[①]，进而使用经验样本进行合成控制法估计以及准标准化转换，得到加州各时期政策效果估计值经验分布（$\widehat{Gap}_{it}^{Pre,S}$）。其次，在时间 $t\in(1,T)$ 上，本文取经验样本 $100(\alpha/2)$ 分位点作为置信区间下界，而 $100(1-\alpha/2)$ 分位点作为置信区间上界（α 为选定的显著性水平），以此来获得各时间点的政策效果置信区间（Efron和Tibshirani，1993），用以检验原假设 $H_0:Gap_{CA,t}^{S}=0$。

表2描述的是禁烟法案95%的置信区间。其中干预后各时点95%的平均置信区间为[–8.969，–4.143]，而对应干预前平均置信区间为[–0.983，0.774]。具体而言，干预后各时点加州人均香烟消费量下降程度均在5%的显著性水平下显著，表明随着禁烟法案的实施，加州人均香烟消费量呈现逐渐下降趋势[②]。这与第四部分中，Abadie等（2010）检验结果（子图1（e））， 以及准标准化后的安慰剂检验结果（子图1（f））一致，表明本文的置信区间很好地揭示了禁烟法案对于加州香烟消费量的政策效果。与此同时，在政策干预前，1988年禁烟效果在5%的显著性水平下显著，这主要是“禁烟法案”于该年正式通过所导致结果。

然而，出于研究过程严谨性的需求，对于在1970年、1975年和1981年禁烟效果显著的情况，研究者可以参考Abadie等（2015）进行时间安慰剂检验，亦或是发生事件的随机性分析，以期进一步证明这些时点对于禁烟法案的政策效果不存在显著影响。

---

①本文的抽样过程是：首先将加州（实验组）从样本中取出；其次对于剩余的 38 个控制组州进行随机抽样，每次抽取 38 个州，并重复抽取 500 次（$i=500$）；最后将加州与随机抽取的控制组州混合，进行合成控制法分析。

②附录表 A1 中，本文详细提供了分位数和正态分布法构造的置信区间。

表2　　禁烟法案的置信区间

干预前（1970—1988年）		干预后（1989—2000年）	
$\widehat{Gap}^{S}_{CA,t}$	Bootstrap500次置信区间（95%）	$\widehat{Gap}^{S}_{CA,t}$	Bootstrap500次置信区间（95%）
2.433	[1.253, 3.378]	**−3.481**	[−4.231, −0.700]
0.949	[−0.000, 1.893]	**−4.452**	[−4.622, −1.494]
−0.555	[−1.292, 1.511]	**−6.907**	[−6.939, −3.363]
−1.226	[−1.550, 1.159]	**−6.701**	[−6.743, −3.504]
−0.163	[−0.298, 1.055]	**−8.586**	[−8.664, −3.996]
0.350	[0.065, 0.923]	**−10.551**	[−10.646, −4.712]
0.153	[−0.669, 0.755]	**−9.995**	[−10.073, −5.263]
0.705	[−0.758, 0.786]	**−10.526**	[−10.623, −5.096]
1.173	[−0.479, 1.222]	**−10.695**	[−10.883, −5.058]
−0.807	[−0.973, 0.110]	**−9.640**	[−9.730, −5.278]
−0.256	[−0.799, 0.083]	**−12.208**	[−12.315, −5.683]
−1.320	[−1.675, −0.156]	**−12.042**	[−12.159, −5.572]
−0.846	[−1.679, 0.117]		
−0.765	[−1.685, 0.312]		
0.196	[−1.800, 1.338]		
−0.558	[−1.366, 0.126]		
−0.750	[−1.322, 0.004]		
−1.640	[−1.916, 0.107]		
−0.976	[−1.744, −0.020]		

注：$\widehat{Gap}^{S}_{CA,t}$ 表示准标准化后的禁烟效果。

六、结论

尽管合成控制法已成为政策干预的重要研究工具，但其统计推断方法却颇受质疑。本文提出使用准标准化转换后的安慰剂检验，以有效降低由于控制组在干预前拟合欠佳而引入的噪音对统计推断的影响。新方法的最大优势在于，在安慰剂检验过程中无需人为删除样本，从而避免了显著性水平主观选择问题。与此同时，本文可以用转换后的“修正数据”，通过Bootstrap来构造各时点政策效果的置信区间，实现统计检验由小样本检验向大样本检验的过度。

准标准化安慰剂检验方式从干预前拟合角度进行处理，保障安慰剂检验结果的有效性，丰富了安慰剂检验修正方法，为进一步推广合成控制法在因果识别的自然实验中的使用，提供有力的保障。值得注意的是，由于数据结构经过了准标准化转换，虽然置信区间的估计实现了小样本推断向大样本推断的转化，但是其作用仍然仅限于政策效果显著性水平的统计推断。因此后续研究将进一步扩展置信区间模型，使之能够实现沟通实际政策效果估计与显著性推断的目的[①]。

参考文献

[1] Abadie A, Diamond A, Hainmueller J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program[J]. Journal of the American Statistical Association, 2010, 105(490): 493−505.

[2] Abadie A, Diamond A, Hainmueller J. Comparative Politics and the Synthetic Control Method[J]. American Journal of Political Science, 2015, 59(2): 495−510.

[3] Abadie A. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects[J]. Journal of Economic Literature, 2021, 59(2): 391−425.

[4] Abadie A, L’hour J. A Penalized Synthetic Control Estimator for Disaggregated Data[J]. Journal of the American Statistical Association, 2021,

①本文的 Stata 实现代码存放于 https://gitee.com/arlionn/synthsd。

1–34.
[5] Agarwal A, Alomar A, Cosson R, et al. Synthetic Interventions[R]. Working Paper, 2020.
[6] Athey S, Imbens G W. The State of Applied Econometrics: Causality and Policy Evaluation[J]. Journal of Economic Perspectives, 2017, 31(2): 3–32.
[7] Ben-Michael E, Feller A, Rothstein J. The Augmented Synthetic Control Method[J]. Journal of the American Statistical Association, 2021: 1–34.
[8] Cerulli G. A Flexible Synthetic Control Method for Modeling Policy Evaluation[J]. Economics Letters, 2019, 182: 40−44.
[9] Chen Y T. A Distributional Synthetic Control Method for Policy Evaluation[J]. Journal of Applied Econometrics, 2020, 35(5): 505−525.
[10] Chernozhukov V, Wuthrich K, Zhu Y. Practical and Robust t-test Based Inference for Synthetic Control and Related Methods[R]. Working Paper, 2020.
[11] Doudchenko N, Imbens G W. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis[R]. Working Paper, 2016.
[12] Efron B, Tibshirani R J. An Introduction to the Bootstrap[M]. CRC press, 1993.
[13] Ferman B, Pinto C. Revisiting the Synthetic Control Estimator[R]. Working Paper, 2016.
[14] Ferman B, Pinto C. Placebo Tests for Synthetic Controls[R]. Working Paper, 2017.
[15] Ferman B, Pinto C. Synthetic Controls with Imperfect Pre-Treatment Fit[J]. Quantitative Economics, 2021, 12(4): 1197−1221.
[16] Firpo S, Possebom V. Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets[J]. Journal of Causal Inference, 2018, 6(2): 1−26.
[17] Fisher R A. Design of Experiments[J]. British Medical Journal, 1936, 1(3923): 554.
[18] Flannery M J, Hankins K W. Estimating Dynamic Panel Models in Corporate Finance[J]. Journal of Corporate Finance, 2013, 19: 1−19.
[19] Hahn J, Shi R. Synthetic Control and Inference[J]. Econometrics, 2017, 5(4): 52.
[20] Hansen B E. Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis[J]. Econometrica, 1996, 64(2): 413−430.
[21] Hollingsworth A, Wing C. Tactics for Design and Inference in Synthetic Control Studies: An Applied Example Using High−Dimensional Data[R]. Working Paper, 2020.
[22] Imbens G W, Rubin D B. Causal Inference in Atatistics, Aocial, and Biomedical Sciences[M]. Cambridge University Press, 2015.
[23] Kellogg M, Mogstad M, Pouliot G, et al. Combining Matching and Synthetic Controls to Trade off Biases from Extrapolation and Interpolation[R]. Working Paper, 2020.
[24] Kim S, Lee C, Gupta S. Bayesian Synthetic Control Methods[J]. Journal of Marketing Research, 2020, 57 (5): 831–852.
[25] Li K T. Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods[J]. Journal of the American Statistical Association, 2020, 115(532): 2068−2083.
[26] Powell D. Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?[R]. Working Paper, 2018.
[27] Rubin D. B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies[J]. Journal of Educational Psychology, 1974, 66(5): 688–701.
[28] Rudholm N, Li Y, Carling K. How Does Big-Box Retail Entry Affect Labor Productivity in Durable Goods Retailing? A Synthetic Control Approach[J]. The Annals of Regional Science, 2022: 1−29.
[29] Xu Y. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models[J]. Political Analysis, 2017, 25(1): 57−76.

作者简介

连玉君，中山大学岭南学院副教授、博士生导师。研究方向为公司财务和金融计量。

李鑫（通讯作者），云南大学经济学院博士研究生。研究方向为人口经济学和环境经济学。电子邮箱：lixin_scholar@163.com。

(责任编辑：张　亮)