《机器学习及R应用》，高等教育出版社，2020年11月

机器学习在经济学的应用：以决策树与随机森林为例

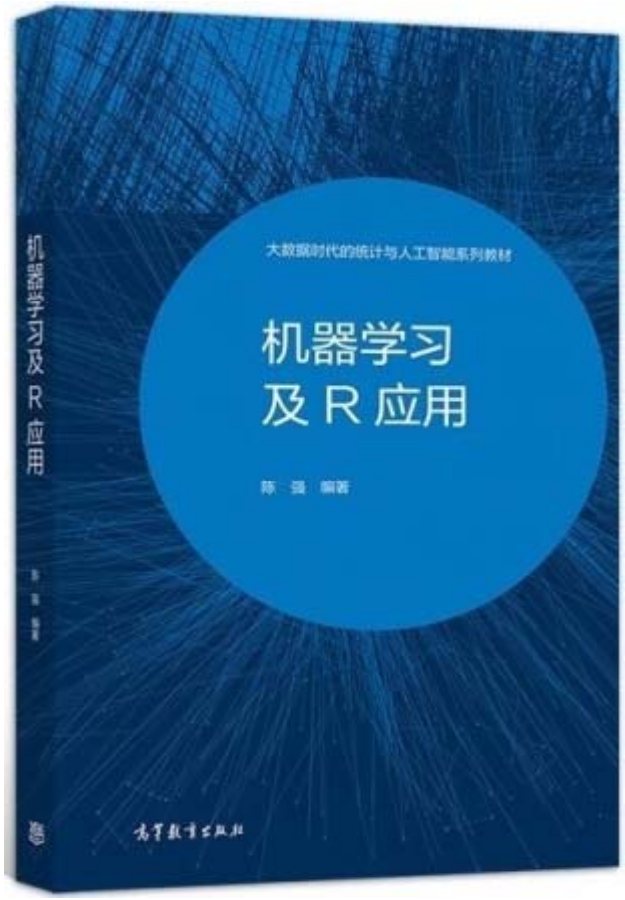
陈强

山东大学经济学院

www.econometrics-stata.com

Outline

1. 决策树（Decision Tree）
2. 决策树的应用
3. 装袋法（Bagging）
4. 随机森林（Random Forest）
5. 随机森林的应用



陈强，《机器学习及R应用》，高等教育出版社，2020年11月，458页，双色印刷

配套数据、R程序：
www.econometrics-stata.com

配套课程(详见网站)：
机器学习及R应用现场班 (北京，2021.1.20-24，经管之家主办)

机器学习的分类

- 机器学习主要分为两类

- 监督学习（Supervised Learning）：用 $\boldsymbol{x}$ 预测 y

- 非监督学习（Unsupervised Learning）：寻找 $\boldsymbol{x}$ 本身的规律或模式，包括主成分分析、聚类分析等

监督学习的分类

- 如果 *y*为连续变量，则称为“回归问题”（Regression）

- 如果 *y*为离散变量，称为“分类问题”（Classification）

监督学习的算法

- 传统算法：OLS，Logit，Multivariate Logit，线性判别分析，K近邻法，决策树，朴素贝叶斯等

- 现代算法：惩罚回归（Lasso, Ridge），装袋法（Bagging），随机森林（Random Forest），提升法（Boosting），支持向量机（Support Vector Machine，简记SVM），人工神经网络（Artificial Neural Network，简记ANN）

(线性)参数回归的困境

- OLS： $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$
- Logit： $\mathrm{P}(y_i = 1 | \boldsymbol{x}_i) = \dfrac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$
- 函数形式？解决方案1：非参数回归
  （K近邻法，决策树等）

  解决方案2：人工神经网络（复杂的参数回归：非线性激活函数的嵌套）
- 高维数据？解决方案：惩罚回归（Lasso，Logistic Lasso）；SVM；ANN；决策树等

*K*近邻法 (KNN)

- 考虑离 $\boldsymbol{x}$ 最近的*K*个邻居
- 记 $N_K(\boldsymbol{x})$为最靠近 $\boldsymbol{x}$ 的 *K*个 $\boldsymbol{x}_i$观测值的集合，则 *K* 近邻估计量 (K-nearest neighbor estimator)为：

$$\hat{f}(\boldsymbol{x}) \equiv \frac{1}{K} \sum_{\boldsymbol{x}_i \in N_K(\boldsymbol{x})} y_i$$

- 如果*K* =1，则为最近邻法。

KNN for Classification

- 对于分类问题(Y 为离散变量)，则采取“多数票规则”(majority vote rule)。

- 以K 近邻中最常见的类别作为预测

- 如果K 近邻中最常见的类别有两个(并列)，则随机选一个最常见类别作为预测

KNN的缺点

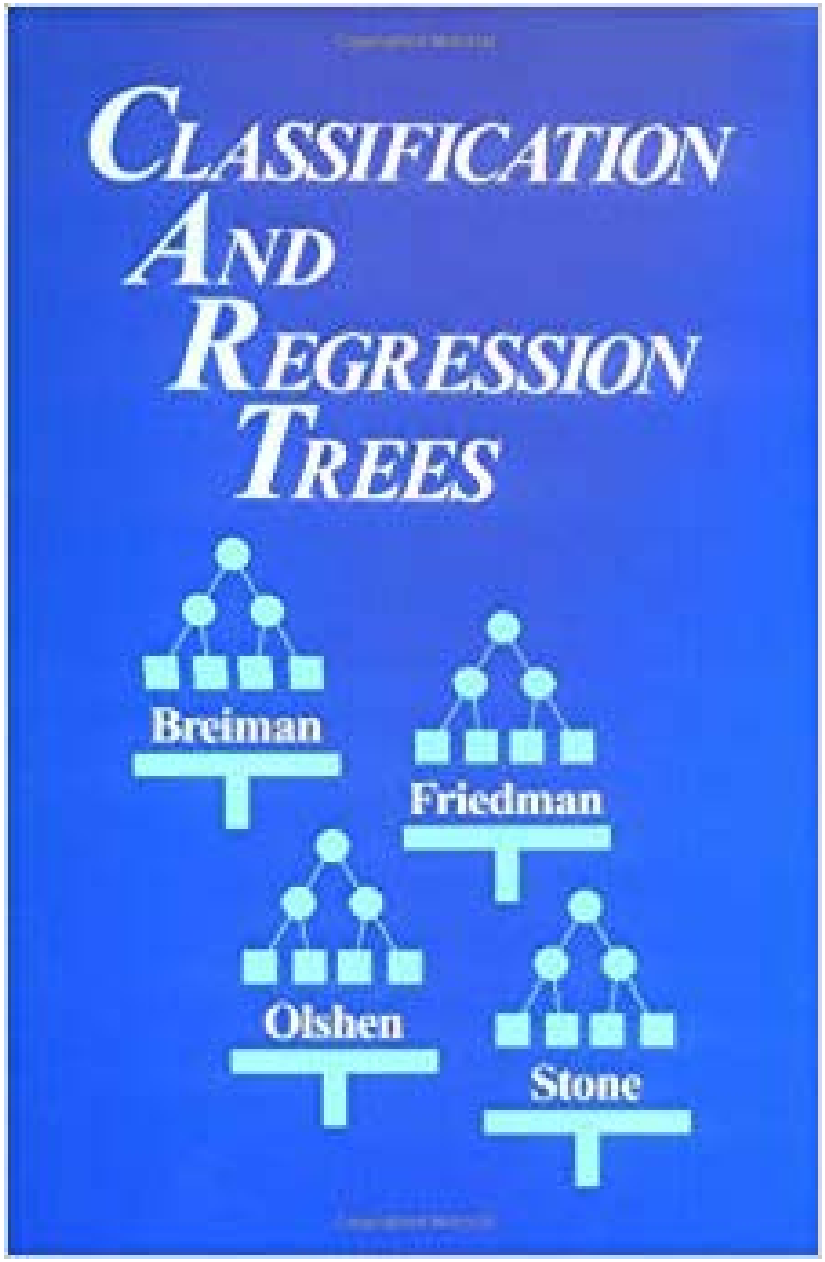
- 懒惰学习(lazy learning)：平时不学习，预测时才去找邻居，导致预测较慢，不适用于“在线学习”(online learning)

- 维度灾难 (curse of dimensionality)：高维空间很难找邻居。KNN一般要求 $p \ll n$

- KNN易受噪音变量的影响

决策树(Decision Trees)的发展

- 决策树的思想与雏形形成于1960年代
- Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees*. Wadsworth. 提出CART算法
- Quinlan, J. R. (1986) "Induction of Decision Trees," Machine Learning, 1(1), 81-106. 提出ID3算法(Iterative Dichotomiser 3)，后演变为C4.5，C5.0算法 (C表示Classifier)


CLASSIFICATION
AND
REGRESSION
TREES
Breiman
Friedman
Olshen
Stone

决策树的分类

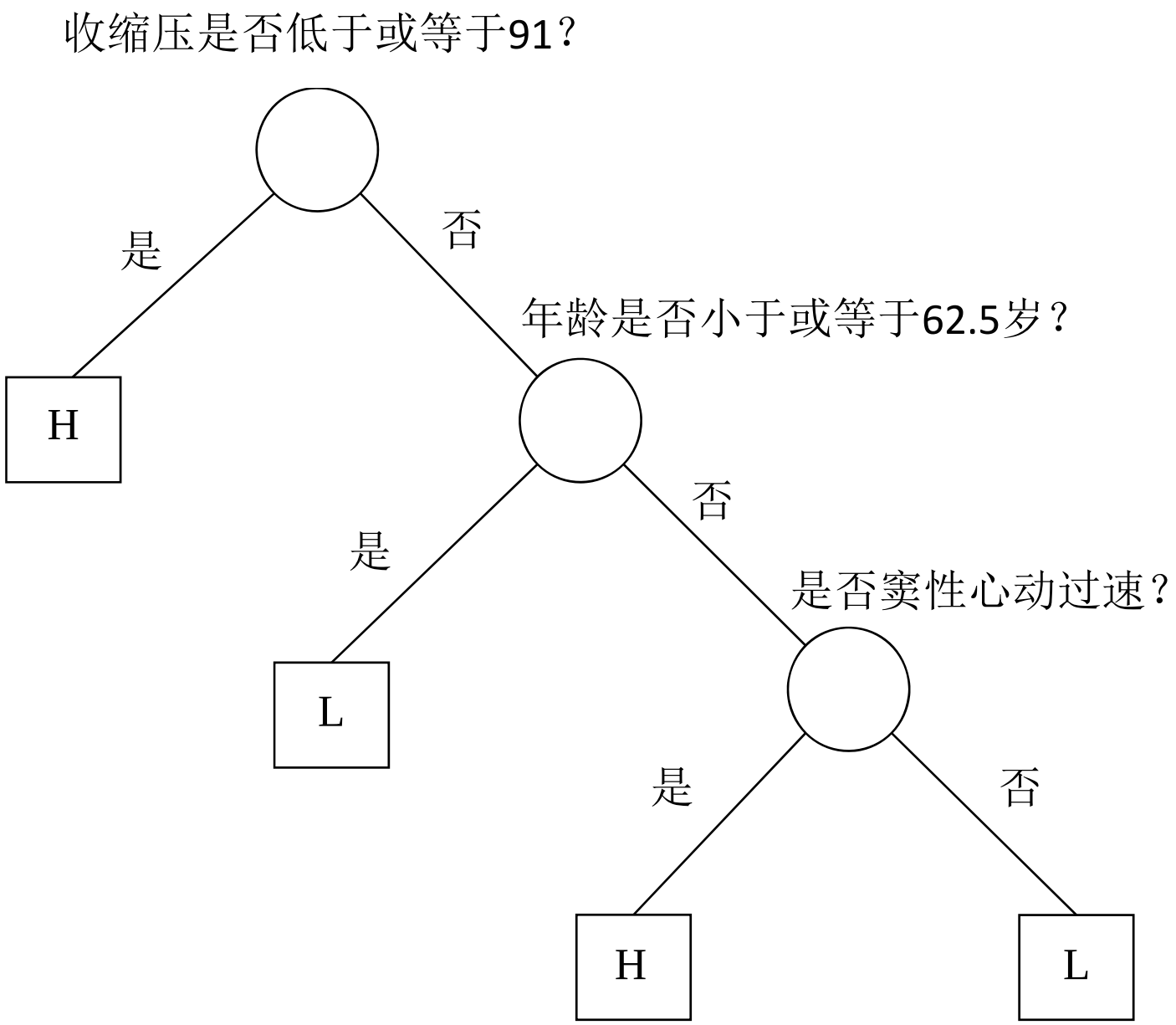
- 分类树(classification tree)：用于分类问题
- 回归树(regression tree)：用于回归问题

例：识别高危心梗病人

- Breiman et al. (1984)研究了UCSD医学中心的一个案例。当心梗病人进入UCSD医学中心后24小时内，测量19个变量，包括血压、年龄以及17个排序或虚拟变量。共215个病人(其中37人高危)。

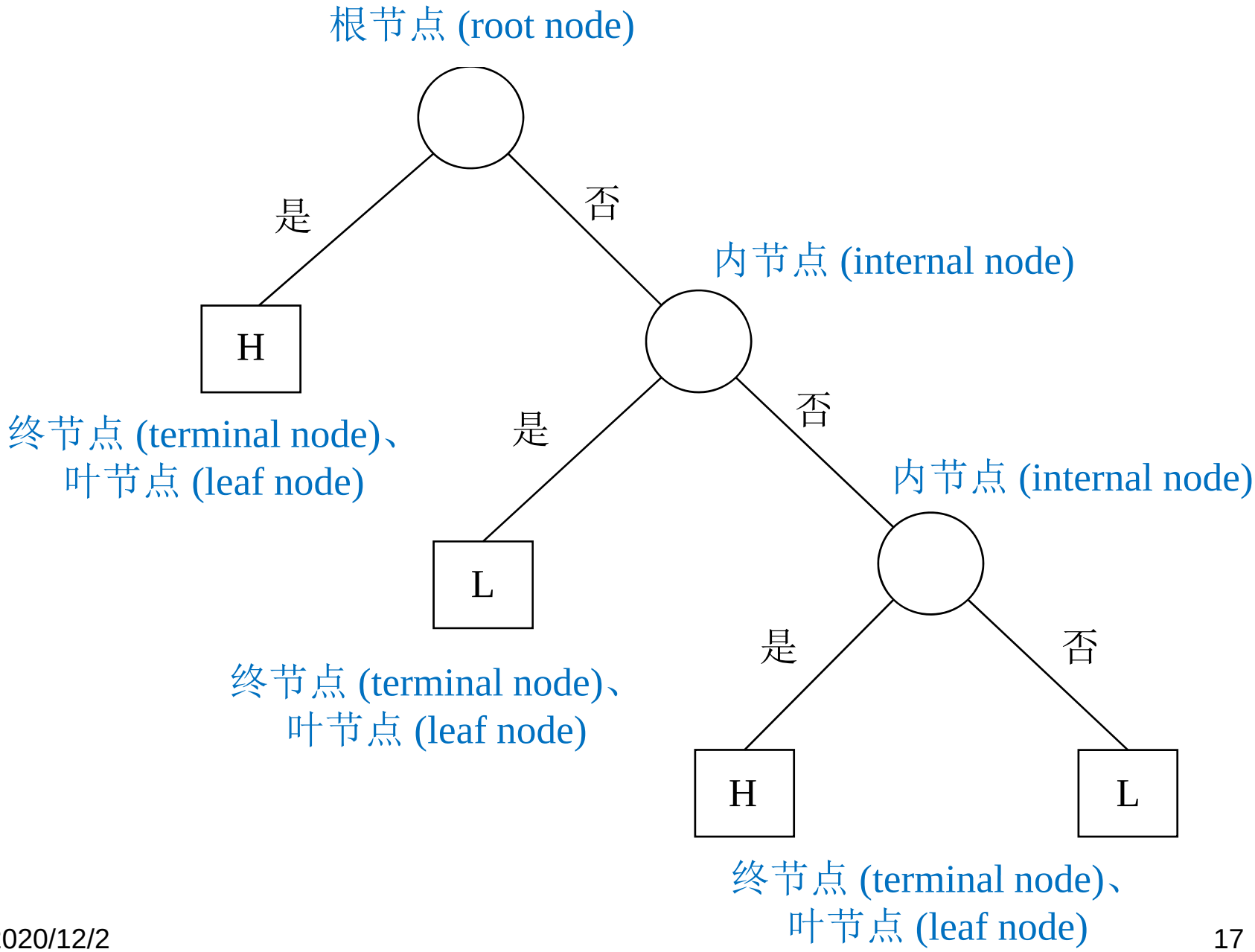
- 研究目的是预测哪些“高危病人”（High risk，记为H，无法活过30天），哪些是“低危病人”（Low risk，记为L，可活过30天）。

- Breiman et al. (1984)建立了如下的分类树


收缩压是否低于或等于91？
是
否
H
年龄是否小于或等于62.5岁？
是
否
L
是否窦性心动过速？
是
否
H
L

决策树的优点

- Breiman et al. (1984)发现，决策树的预测效果可能优于参数模型(判别分析，逻辑回归)
- 分类树的预测非常简单，just drop an observation down the tree (回答一系列的是或否问题)，使用“多数票规则” (majority vote rule)
- 决策树的可解释性较强(甚至不用方程！)


根节点 (root node)
是
否
H
内节点 (internal node)
终节点 (terminal node)、
叶节点 (leaf node)
是
否
L
内节点 (internal node)
终节点 (terminal node)、
叶节点 (leaf node)
是
否
H
L
终节点 (terminal node)、
叶节点 (leaf node)

递归分割 (Recursive Partitioning)

- CART算法使用“二叉树”(binary tree)，本质上将“特征空间”(feature space)进行递归分割，每次总是沿着与某个变量 x 轴平行的方向进行切割，切成矩形区域(boxes)

- 在数学上，决策树为“分段常值函数”(piecewise constant function)

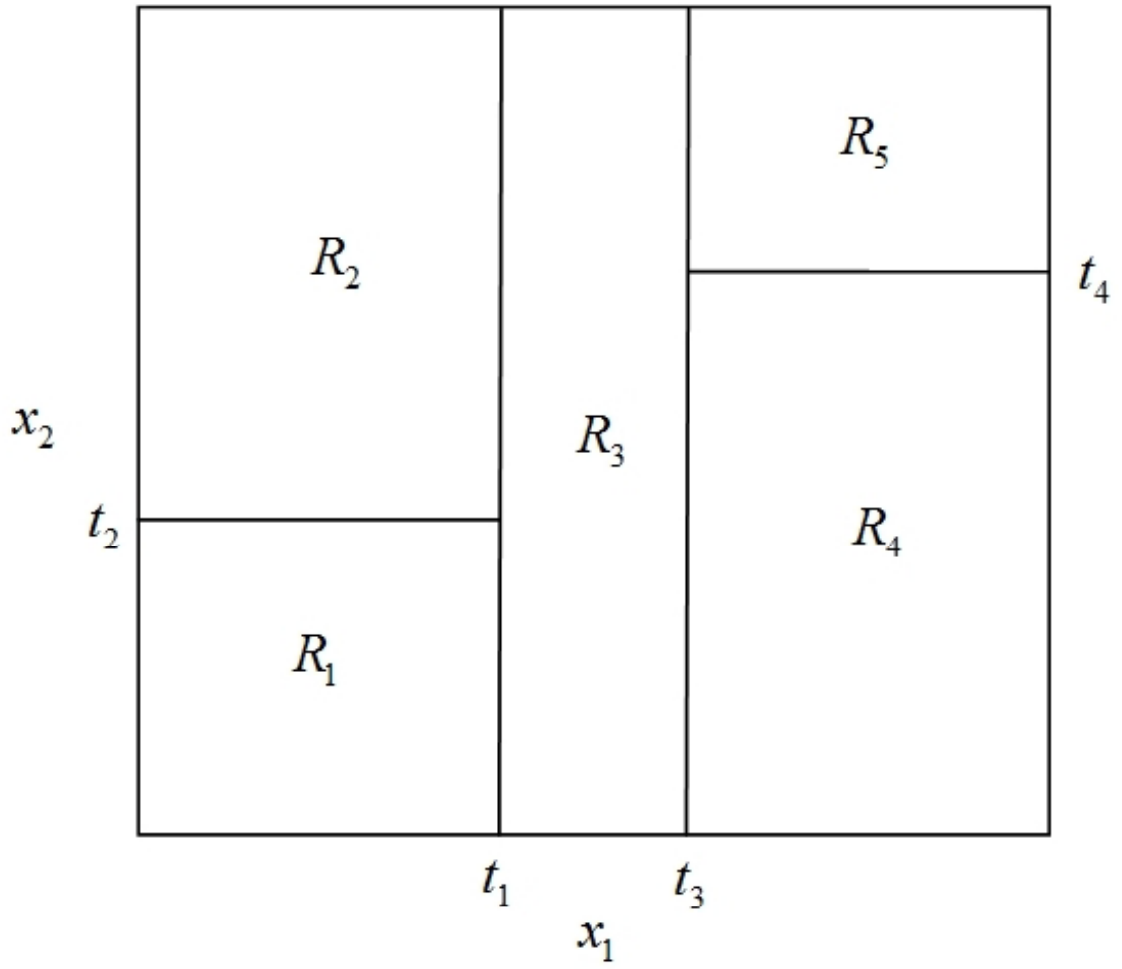


图 11.2 决策树对特征空间的递归分割

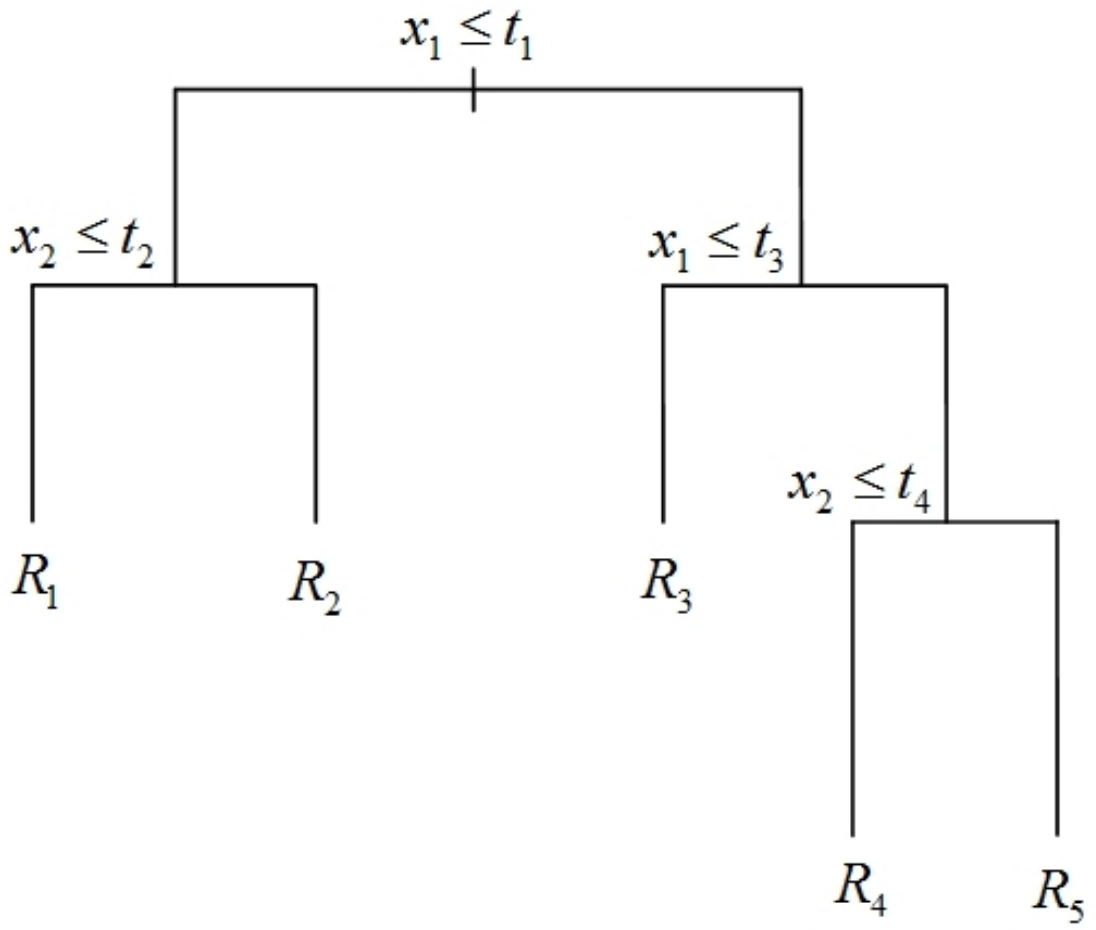


图 11.3 两个特征变量的决策树

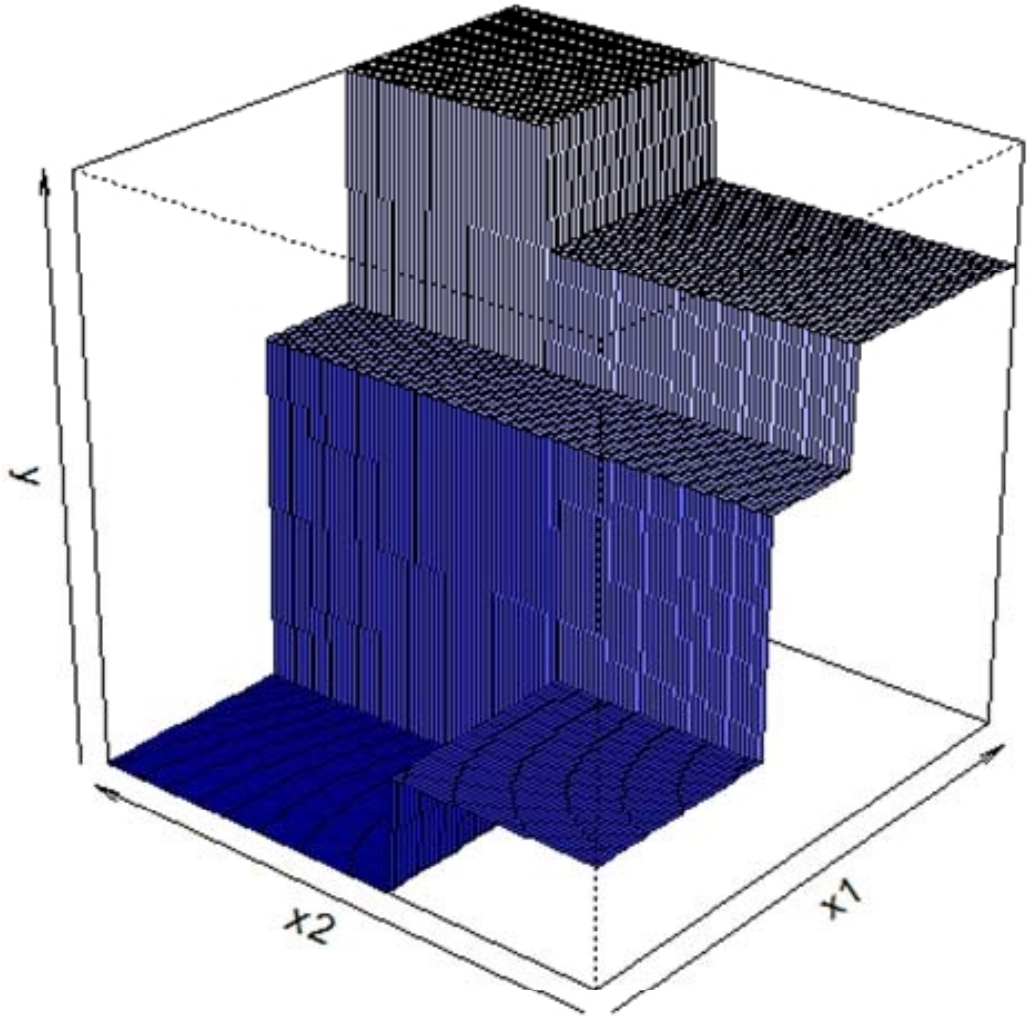


图 11.4 作为分段常值函数的决策树

分裂准则 (splitting criterion)

- 选择什么变量(split variable)进行分裂？
- 在该变量的什么取值(cut)进行分裂？
- 目标：使得分裂之后的两个**子节点** 内部的“纯度”(purity)最高。换言之，分裂之后，数据的“不纯度”(impurity)下降最多

Node Impurity Function

- 假设响应变量 y 共分 K 类，取值为 $1,\cdots,K$ 。在节点 t (node t)，记 y 不同取值的相应概率(频率)为 $p_1,\cdots,p_K$ ，其中 $p_k \geq 0, \sum_{k=1}^{K} p_k = 1$

- 作为“分裂准则”(splitting criterion)，首先定义一个“节点不纯度函数”(node impurity function) $\varphi(p_1,\cdots,p_K) \geq 0$ 。该函数应具备以下性质：

节点不纯度函数的性质

① 当 $p_1 = \cdots = p_K = \frac{1}{K}$ 时，$\varphi(p_1, \cdots, p_K)$ 达到最大值

① 当且仅当 $(p_1, p_2 \cdots, p_K) = (1,0,\cdots,0), (0,1,\cdots,0)$，…，或 $(0,0,\cdots,1)$ 时，$\varphi(p_1, \cdots, p_K)$ 达到最小值 0

② $\varphi(p_1, \cdots, p_K)$ 关于其自变量是对称的。

- 满足这些性质的函数并不唯一

错分率

- 一个自然的选择是“错误率”(error rate) 或“错分率”(misclassification rate)：

$$\mathrm{Err}(p_1,\cdots,p_K)\equiv 1-\max\{p_1,\cdots,p_K\}$$

- 不难验证，错分率满足以上三条性质。在二分类问题中，错分率简化为

$$\mathrm{Err}(p_1,p_2)\equiv 1-\max\{p_1,1-p_1\}=\begin{cases}p_1 & if\ \ 0\le p_1<0.5\\ 1-p_1 & if\ \ 1\ge p_1\ge 0.5\end{cases}$$

基尼指数

- 基尼指数度量的是，从概率分布$(p_1,\cdots,p_K)$中随机抽取两个观测值，则这两个观测值的类别不一致的概率为

$$\text{Gini}(p_1,\cdots,p_K)=\sum_{k=1}^{K}p_k(1-p_k)=\sum_{k=1}^{K}p_k-\sum_{k=1}^{K}p_k^2=1-\sum_{k=1}^{K}p_k^2$$

- 其中，$\sum_{k=1}^{K}p_k^2$ 可视为随机抽样的两个观测值之类别一致的概率

信息熵 (information entropy)

- 如果将 y 的每个可能取值的信息量，以相应概率 p_k 为权重，加权求和即可得“信息熵”(Shannon, 1948)：

$$\text{Entropy}(p_1, \cdots, p_K) \equiv -\sum_{k=1}^{K} p_k \log_2 p_k$$

- 其中，$\log_2(\cdot)$ 为以2为底的对数 (也可使用自然对数)，其单位称为“比特”(bit，即binary digits)；并定义 $0 \cdot \log_2(0) \equiv 0$ ，因为根据洛必达法则，

$$\lim_{p \to 0} p \cdot \log_2(p) = 0$$

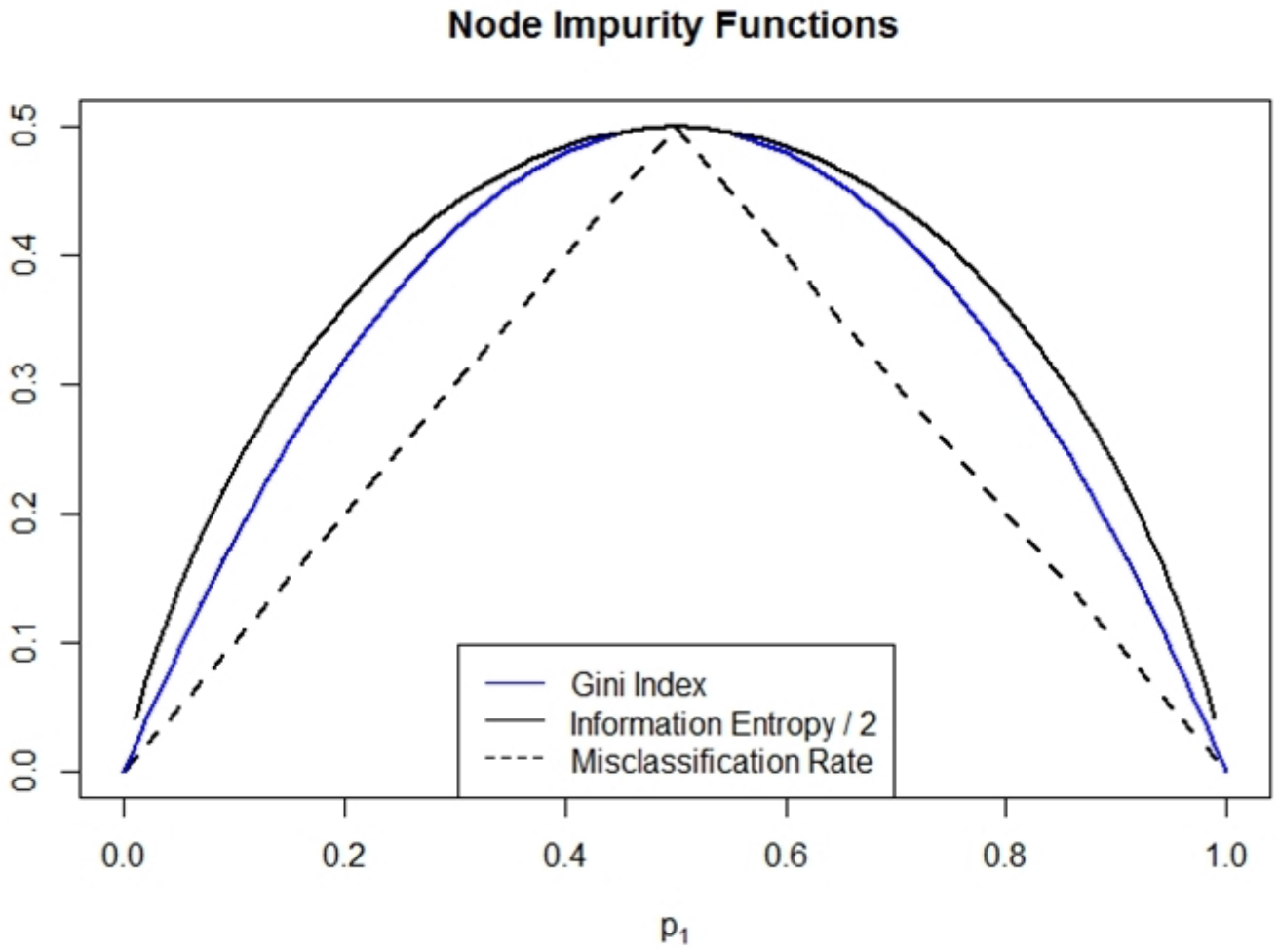


图 11.8 三种节点不纯度函数的比较

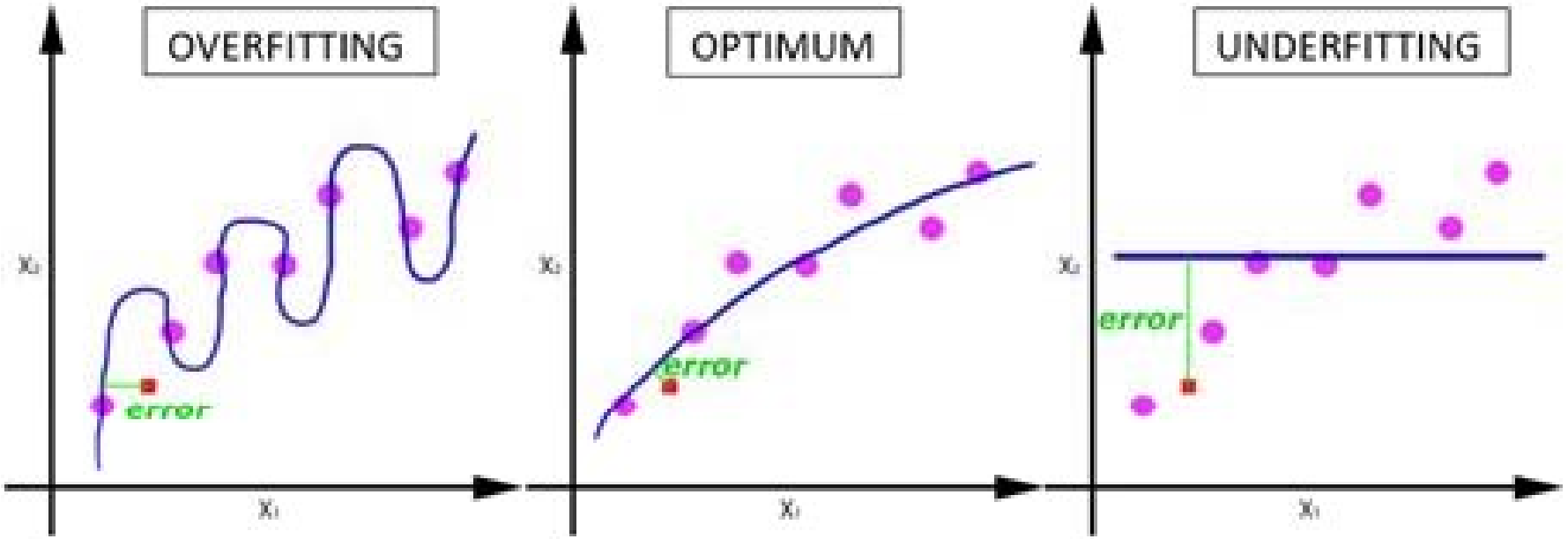
何时停止分裂？

- 如果不停地进行分裂，将使得最后每个叶节点只有一个观测值，导致过拟合（overfit）

- 何时停止分裂？

- 解决方法：先让决策树尽情生长，记最大的树为$T_{\max}$，再进行“修枝”(pruning)

Overfit的图示



成本复杂性修枝 (Cost-complexity Pruning)

- 对于任意子树 $T \leq T_{\max}$，定义其“复杂性”(complexity) 为子树*T*的终节点数目，记为 $|T|$

- 为避免过拟合，不希望决策树过于复杂，故惩罚其规模 $|T|$：

$$\min_T \underbrace{R(T)}_{cost} + \alpha \cdot \underbrace{|T|}_{complexity}$$

- 其中，$R(T)$ 为原来的损失函数(比如 0-1损失函数)，α 为调节参数(通过交叉验证确定)

决策树 vs. KNN

- 二者的共同点：都采取“分而治之”(Divide and Conquer)策略，将特征空间分割为若干子区域，进行预测

- KNN在分区域时，不考虑 y，高维空间易遇到维度灾难，且易受噪音变量影响

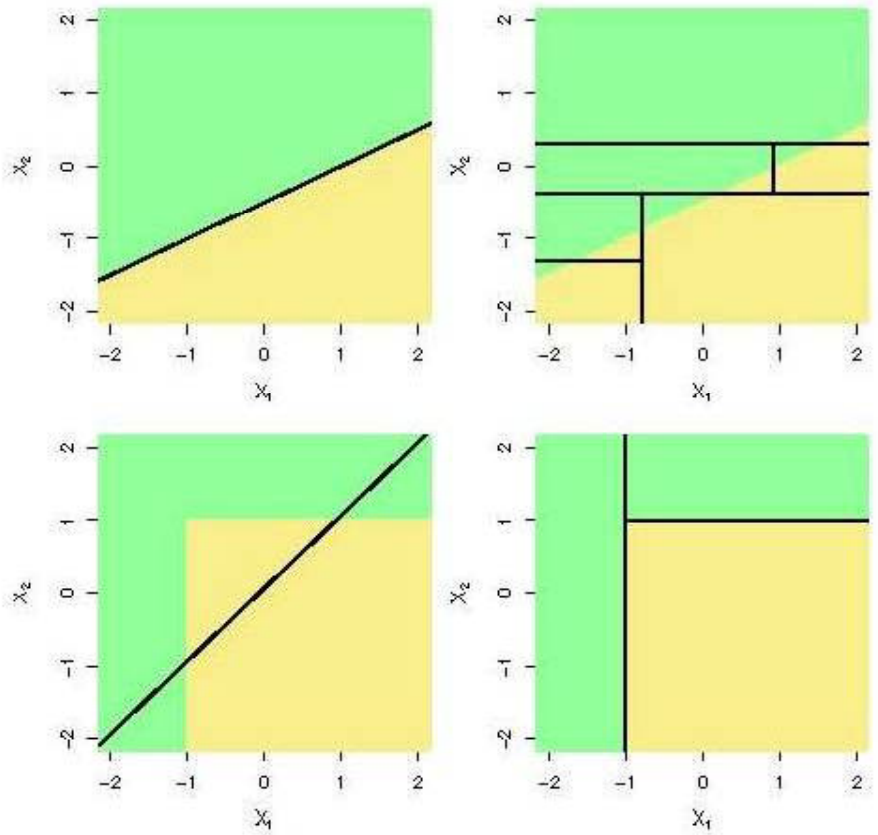
- 决策树在分区域时，考虑 $\boldsymbol{x}$ 对 y 的影响，分区更具智慧，可视为“adaptive nearest neighbor”，容易推广到高维空间，且不受噪音变量的影响

决策树的缺点

- 决策树在递归分裂时，仅考虑“超矩形体”(hyperrectangle)。如果真实的决策边界与此相差较远或不规则，可能导致较大误差

- 基于决策树的集成学习(装袋法、随机森林、提升法)，可得到光滑的决策边际，大幅提高预测准确率。

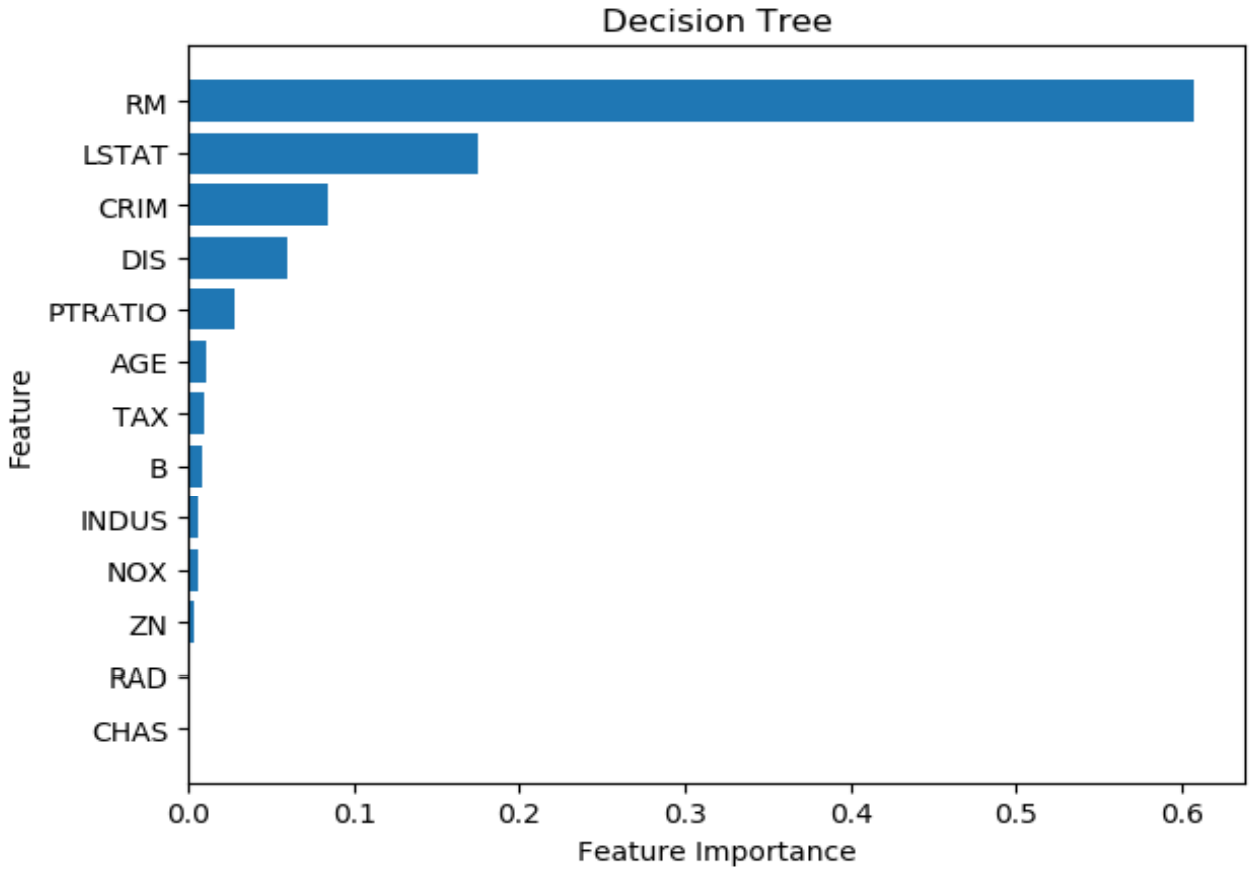
决策树 vs 线性模型



Top Row: True linear boundary; Bottom row: true non-linear boundary.

Left column: linear model; Right column: tree-based model

特征重要性图
（Feature Importance Plot）



决策树的应用

- Kelly, Morgan and Cormac O Grada, 2000. Market Contagion: Evidence from the Panics of 1854 and 1857, *American Economic Review*, 90(5), 1110-1124.

- Athey, Susan and Guido Imbens, 2016. Recursive Partitioning for Heterogeneous Causal Effects, *PNAS*, 113(27), 7353-60.

  （因果树：Causal Tree）

Market Contagion: Evidence from the Panics of 1854 and 1857

By MORGAN KELLY AND CORMAC O GRADA*

To test a model of contagion—where individuals hear some bad news and communicate it to their acquaintances, who then pass it on, leading to a market panic—requires a knowledge of the information networks of participants, something hitherto unavailable. For two panics in the 1850's this paper examines the behavior of Irish depositors in a New York bank. As recent immigrants, their social network was determined largely by their place of origin in Ireland, and where they lived in New York. During both panics this social network turns out to be the prime determinant of behavior. (*JEL* G21, N21)

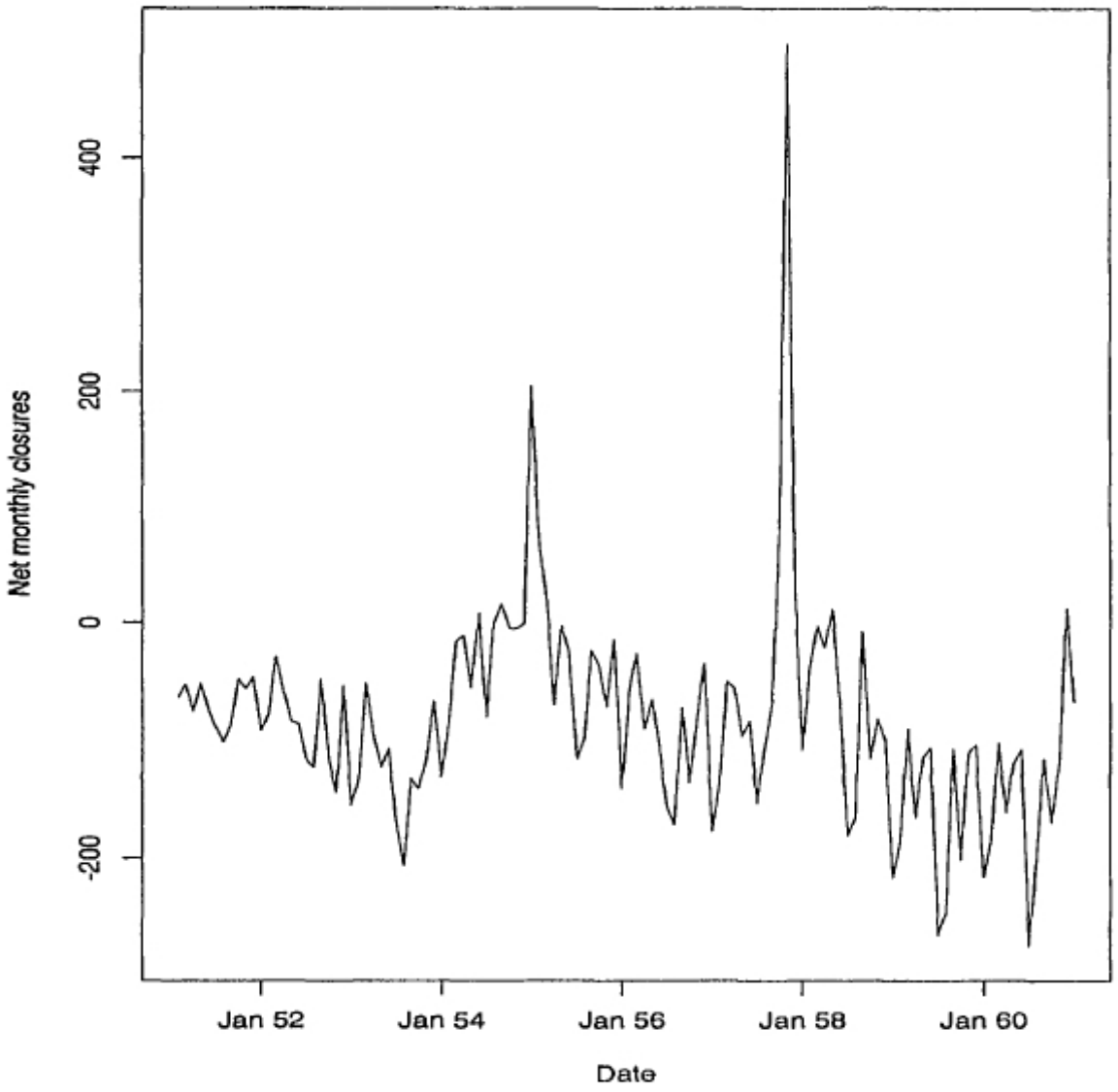


FIGURE 1. CHANGE IN NUMBER OF ACCOUNTS (NET CLOSURES) BY MONTH, 1851–1860

TABLE 2—DEFINITIONS OF VARIABLES

Panicked	Account closed during panic
Previous deposits	Number of deposits made into account at annualized rate, excluding initial deposit
Previous withdrawals	Number of withdrawals from account prior to panic
Closing balance	Closing balance if panicked, balance at end of panic otherwise
Length open	Number of months the account had been open prior to panic
Years in United States	Number of years the depositor had lived in the United States
Occupation	Occupation: laborer (l), professional (p), or other (o)
Sex	Female or male
District	Depositor's address given by grid coordinate of Phelps's 1857 "New York City Street and Avenue Guide" (3b–6d); otherwise Downtown (dt), Midtown (mt), Uptown (ut), Long Island (li), Brooklyn (bn), Staten Island (si), New Jersey (nj), Upstate (us), or other (oth)
County	Depositor's county of origin in Ireland

Note: Panic is defined as the period from December 11 to December 30 for the 1854 data, and from September 28 to October 13 for the 1857 data.

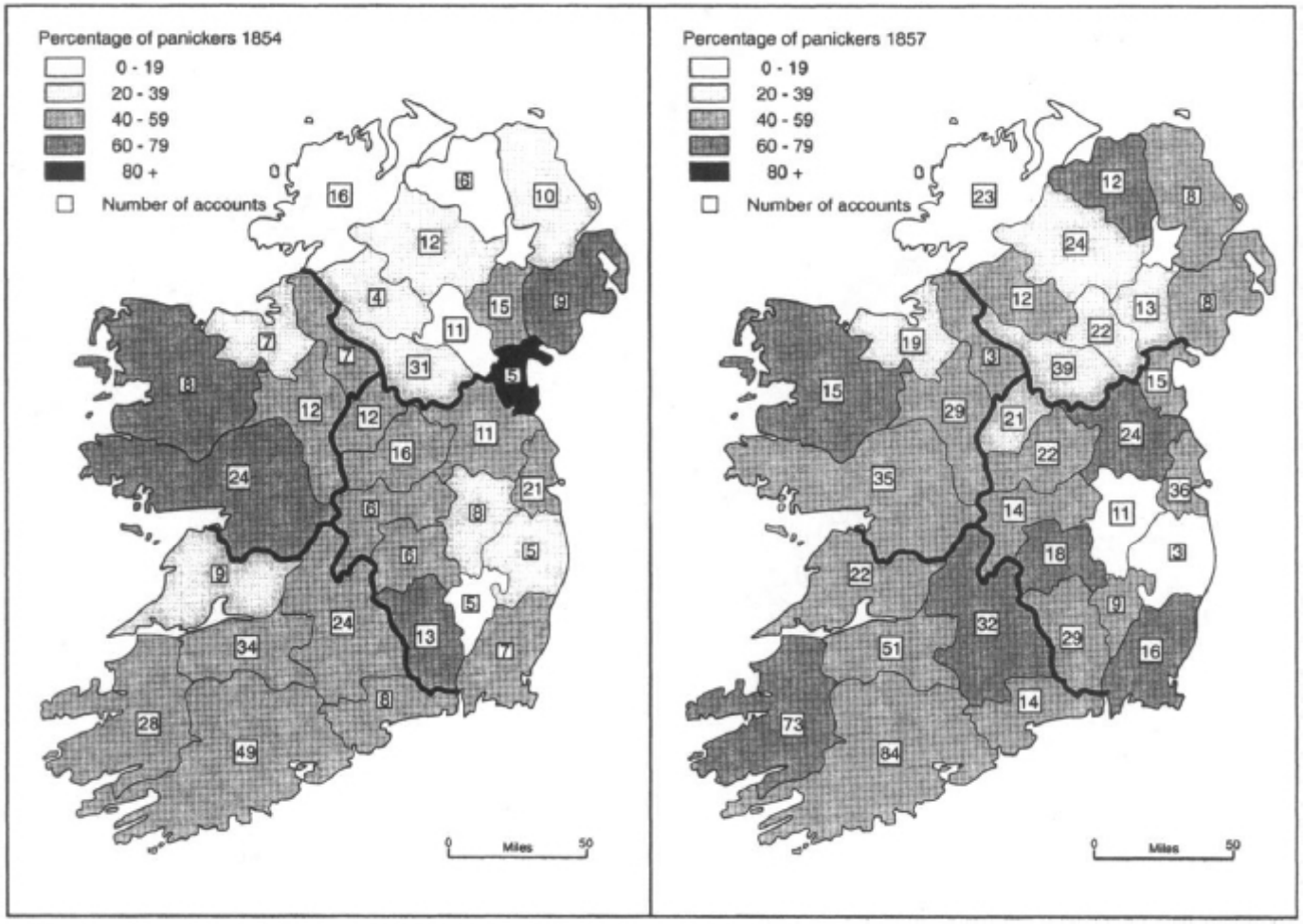


FIGURE 3. NUMBER OF DEPOSITORS AND PERCENTAGE OF PANICKERS BY COUNTY OF ORIGIN IN IRELAND, 1854 AND 1857

Note: Dark lines are boundaries of provinces.

TABLE 4—CHARACTERISTICS OF PANICKERS: LOGISTIC REGRESSION

	1854	1857	1854	1857
Intercept	0.4976	0.5951**	0.5747	0.6118*
	(0.2758)	(0.2178)	(0.3314)	(0.2543)
Previous deposits	0.0139	0.0059	0.0062	−0.0037
	(0.021)	(0.0261)	(0.0208)	(0.0264)
Previous withdrawals	0.0225	0.043	0.0281	0.0389
	(0.0173)	(0.0303)	(0.0182)	(0.0306)
Closing balance	−0.0018*	−0.001*	−0.001*	−0.0008
	(0.0008)	(0.0005)	(0.0008)	(0.0005)
Length open	−0.0324**	−0.0116**	−0.034**	−0.0117**
	(0.0102)	(0.0044)	(0.0105)	(0.0045)
Years in United States	−0.0456*	−0.0459**	−0.04*	−0.0396**
	(0.0201)	(0.0126)	(0.0202)	(0.0128)
Female	−0.1171	0.3581*	−0.1263	0.3627*
	(0.2273)	(0.1618)	(0.2319)	(0.1635)
Laborer	0.452*	0.2565	0.4945*	0.2124
	(0.2133)	(0.1572)	(0.2182)	(0.1591)
Professional	−0.4297	0.299	−0.2398	0.3737
	(0.7098)	(0.4267)	(0.7138)	(0.4303)
Ulster			−0.8247**	−0.6116**
			(0.2951)	(0.219)
Connacht			0.4871	0.025
			(0.338)	(0.2405)
Munster			−0.0295	0.2065
			(0.2617)	(0.1916)
Density	0.0648	0.0882	0.0641	0.0882
Null deviance	602	1061	602	1061
Residual deviance	553	1001	537	987
Percent misclassified	35	39	32	36

Logit with county fixed effects?

When county-of-origin dummies were added to the regressions in the first two columns of Table 4, only Galway and Wexford in 1854 were individually significant. However, the county dummies are jointly strongly significant: the chi-squared statistic for the hypothesis that their coefficients are jointly zero is 54 for 1854 and 72 for 1857. Using district-of-residence instead of county-of-origin dummies, districts 4d and 4e were individually significant in 1854, and district 4b in 1857. Testing joint significance of all districts gave a borderline significant chi-squared statistic of 30 for 1854, but a highly significant 43 for 1857. If county of origin is already included, the district-of-residence dummies are jointly insignificant for 1854, but significant at 2 percent for 1857. While these results suggest that social network factors may play an important role in panics, we need a technique that can handle factors with many levels, and possible nonlinear interactions

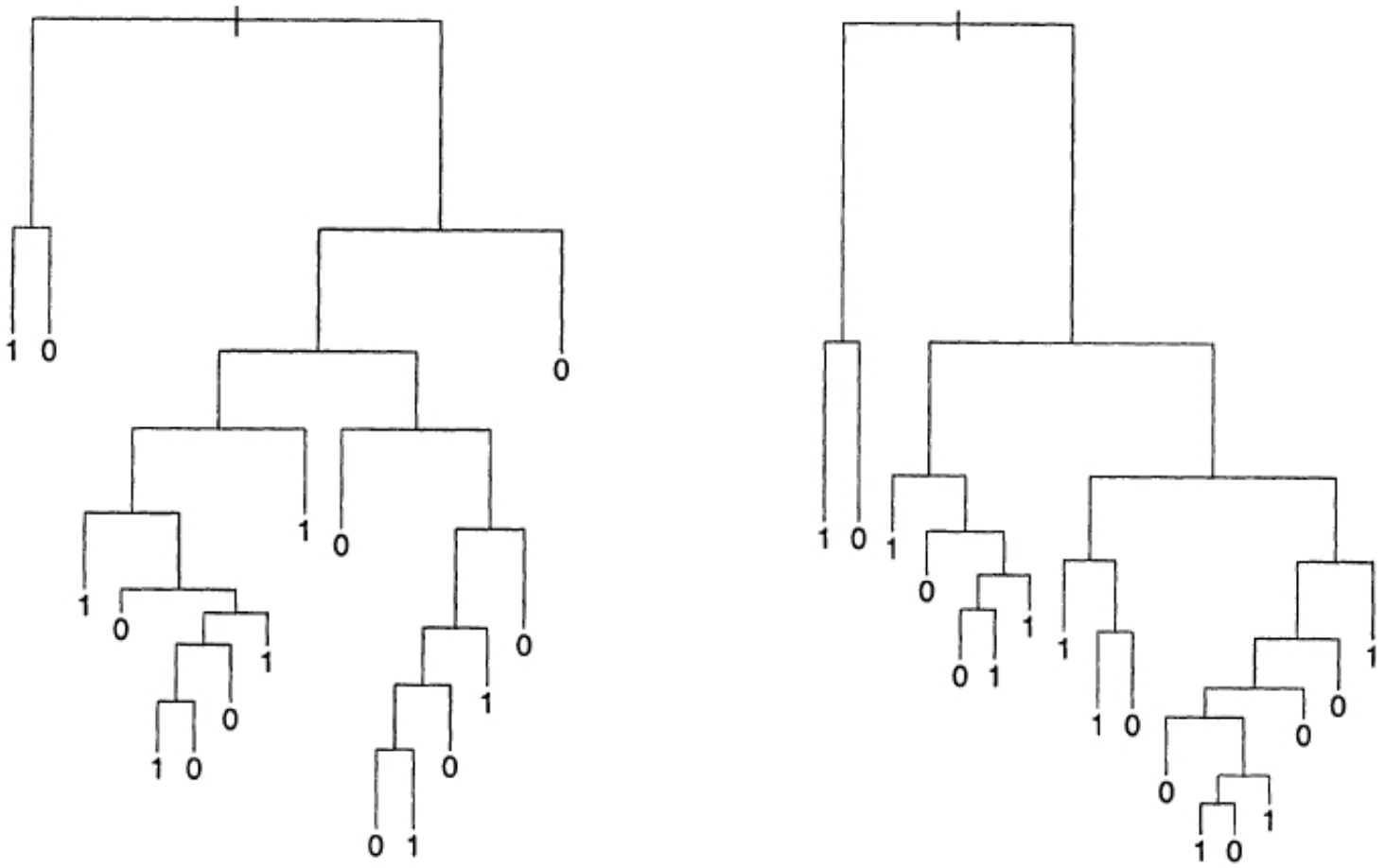
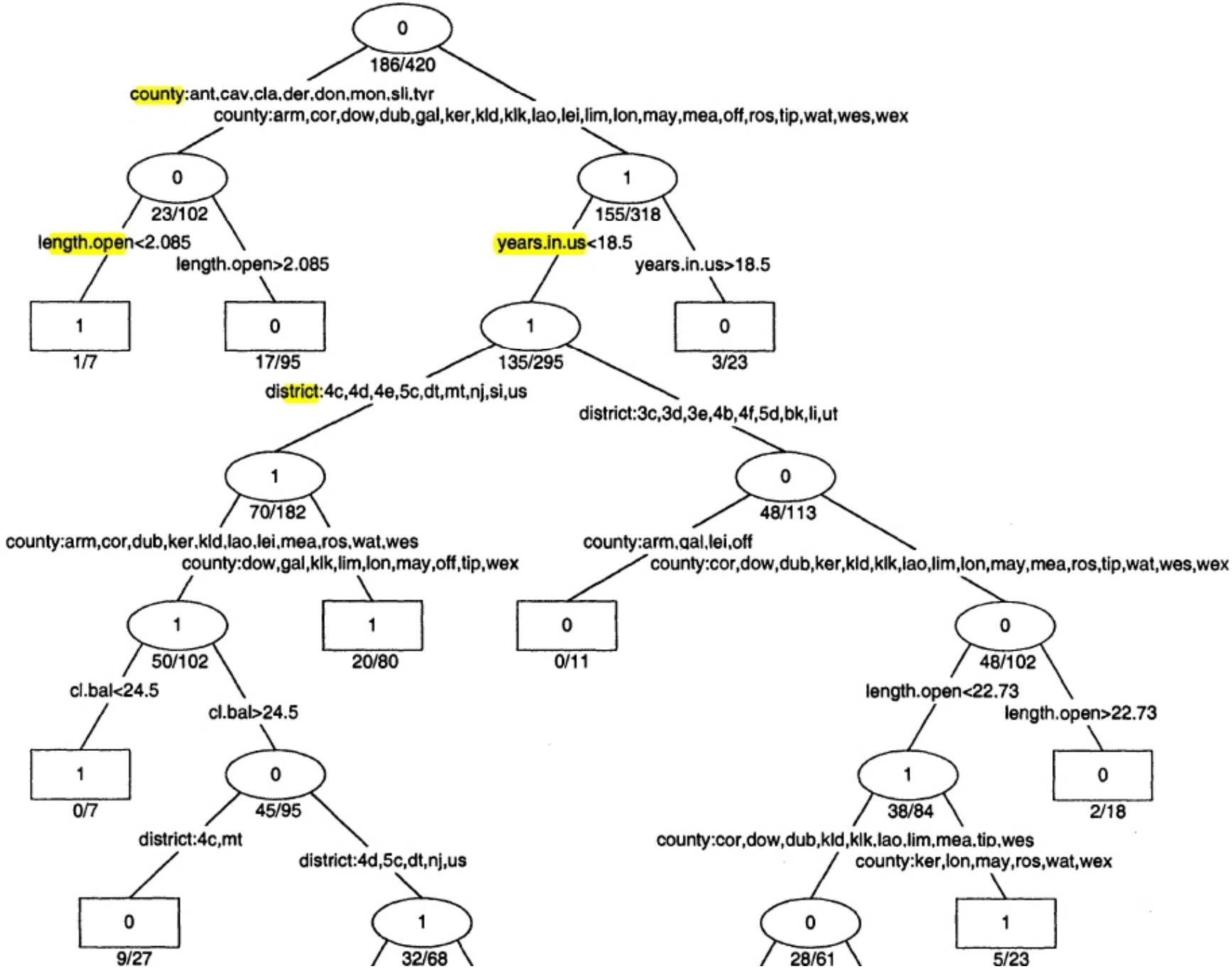
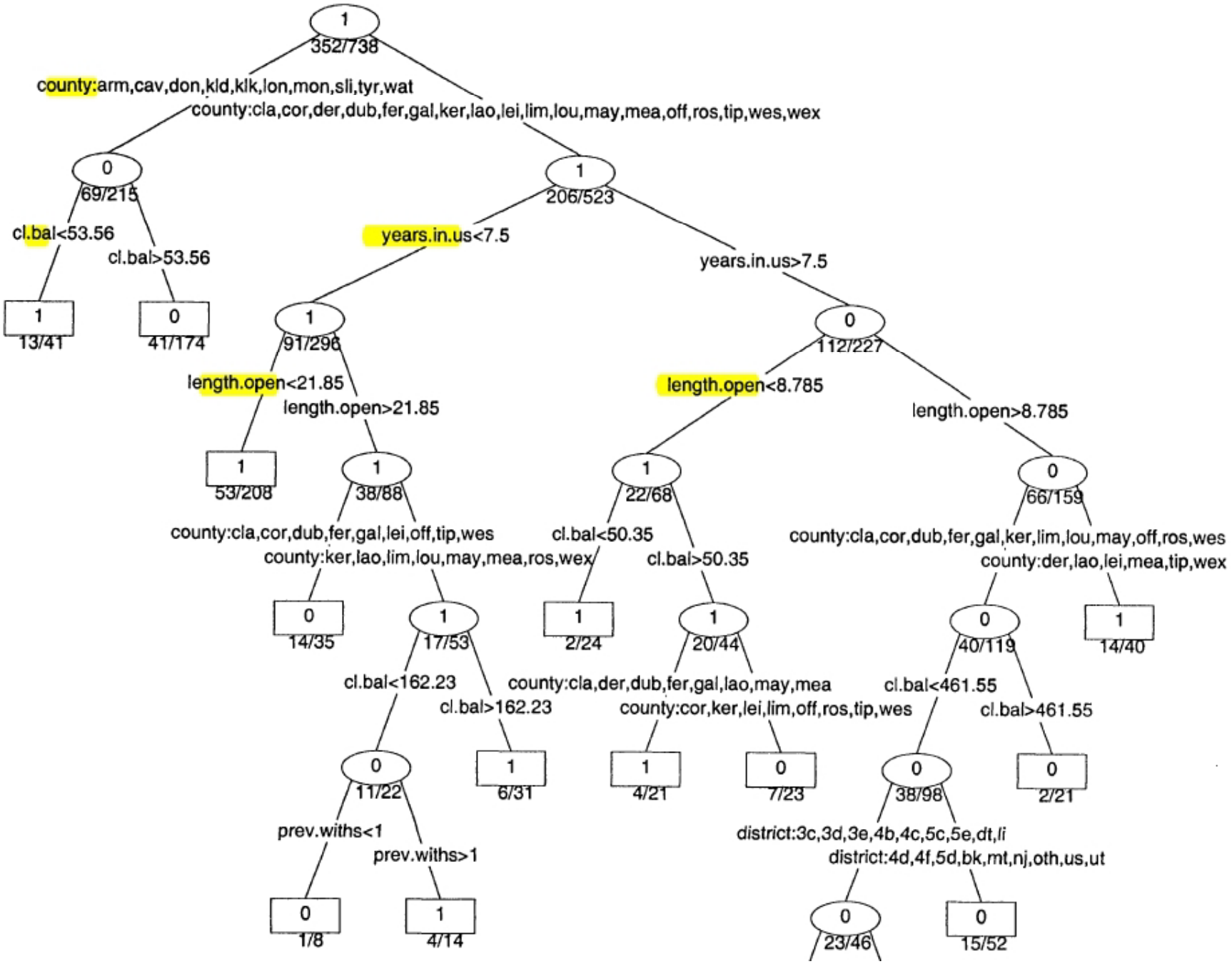


FIGURE 6. SCHEMATIC VERSION OF CLASSIFICATION TREES IN FIGURES 4 (1854, LEFT) AND 5 (1857, RIGHT)

Note: The depth of the branches below each node indicates the relative importance of each split in reducing the misclassification rate.


0
186/420
county:ant,cav,cla,der,don,mon,sli,tyr
county:arm,cor,dow,dub,gal,ker,kld,klk,lao,lei,lim,lon,may,mea,off,ros,tip,wat,wes,wex
0
23/102
1
155/318
length.open<2.085
length.open>2.085
years.in.us<18.5
years.in.us>18.5
1
1/7
0
17/95
1
135/295
0
3/23
district:4c,4d,4e,5c,dt,mt,nj,si,us
district:3c,3d,3e,4b,4f,5d,bk,li,ut
1
70/182
0
48/113
county:arm,cor,dub,ker,kld,lao,lei,mea,ros,wat,wes
county:dow,gal,klk,lim,lon,may,off,tip,wex
county:arm,gal,lei,off
county:cor,dow,dub,ker,kld,klk,lao,lim,lon,may,mea,ros,tip,wat,wes,wex
1
50/102
1
20/80
0
0/11
0
48/102
cl.bal<24.5
cl.bal>24.5
length.open<22.73
length.open>22.73
1
0/7
0
45/95
1
38/84
0
2/18
district:4c,mt
district:4d,5c,dt,nj,us
county:cor,dow,dub,kld,klk,lao,lim,mea,tip,wes
county:ker,lon,may,ros,wat,wex
0
9/27
1
32/68
0
28/61
1
5/23


1
352/738
county:arm,cav,don,kld,klk,lon,mon,sli,tyr,wat
county:cla,cor,der,dub,fer,gal,ker,lao,lei,lim,lou,may,mea,off,ros,tip,wes,wex
0
69/215
1
206/523
cl.bal<53.56
cl.bal>53.56
years.in.us<7.5
years.in.us>7.5
1
13/41
0
41/174
1
91/296
0
112/227
length.open<21.85
length.open>21.85
length.open<8.785
length.open>8.785
1
53/208
1
38/88
1
22/68
0
66/159
county:cla,cor,dub,fer,gal,lei,off,tip,wes
county:ker,lao,lim,lou,may,mea,ros,wex
cl.bal<50.35
cl.bal>50.35
county:cla,cor,dub,fer,gal,ker,lim,lou,may,off,ros,wes
county:der,lao,lei,mea,tip,wex
0
14/35
1
17/53
1
2/24
1
20/44
0
40/119
1
14/40
cl.bal<162.23
cl.bal>162.23
county:cla,der,dub,fer,gal,lao,may,mea
county:cor,ker,lei,lim,off,ros,tip,wes
cl.bal<461.55
cl.bal>461.55
0
11/22
1
6/31
1
4/21
0
7/23
0
38/98
0
2/21
prev.withs<1
prev.withs>1
district:3c,3d,3e,4b,4c,5c,5e,dt,li
district:4d,4f,5d,bk,mt,nj,oth,us,ut
0
1/8
1
4/14
0
23/46
0
15/52

Conclusion of Market Contagion Paper

When we examined the behavior of these depositors in the panics of 1854 and 1857, we found that whether an individual panicked or not depended strongly on how long they had lived in America, and how long they had been with the bank. The most important factor in whether they panicked, however, was county of origin. Depositors from one set of counties tended to close their accounts in both panics, while otherwise identical individuals from other counties tended to stay with the bank. Our results show that individual behavior depends not only on private information but on access to the information and opinions of other group members, and raises the possibility that a handful of influential individuals can have a lot of power over group opinion.

Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] and Guido Imbens[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015)

In this paper we propose methods for estimating heterogeneity in causal effects in experimental and observational studies and for conducting hypothesis tests about the magnitude of differences in treatment effects across subsets of the population. We provide a data-driven approach to partition the data into subpopulations that differ in the magnitude of their treatment effects. The approach enables the construction of valid confidence intervals for treatment effects, even with many covariates relative to the sample size, and without "sparsity" assumptions. We propose an "honest" approach to estimation, whereby one sample is used to construct the partition and another to estimate treatment effects for each subpopulation. Our approach builds on regression tree methods, modified to optimize for goodness of fit in treatment effects and to account for honest estimation. Our model selection criterion anticipates that bias will be eliminated by honest estimation and also accounts for the effect of making additional splits on the variance of treatment effect estimates within each subpopulation. We address the challenge that the "ground truth" for a causal effect is not observed for any individual unit, so that standard approaches to cross-validation must be modified. Through a simulation study, we show that for our preferred

Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction. In this paper, we focus on the analogous goal of deriving a partition of the population according to treatment effect heterogeneity, building on standard regression trees (5, 6). Whether the ultimate goal in an application is to derive a partition or fully personalized treatment effect estimates depends on the setting; settings where partitions may be desirable include those where decision rules must be remembered, applied, or interpreted by human beings or computers with limited processing power or memory. Examples include treatment guidelines to be used by physicians or even online personalization applications where having a simple lookup table reduces latency for the user. We show that an attractive feature of focusing on partitions is that we can achieve nominal coverage of confidence intervals for estimated treatment effects even in settings with a modest number of observations and many covariates. Our approach has applicability even for settings such as clinical trials of drugs with only a

集成学习(ensemble learning)

- 是否可能将一些“弱学习器”(weak learner，比如决策树)组合起来，构成一个“强学习器”(strong learner)？

- 这种方法称为“集成学习”(ensemble learning)或“组合学习”

- “三个臭皮匠，顶一个诸葛亮”

基学习器

- 用于集成学习的弱学习器也称为“基学习器”(base learner)
- 决策树是最常见的基学习器
- 数据 + 算法 = 结果
- 如何得到不同的决策树？

搅动 + 组合

- 给定算法(比如CART)，搅动数据，得到不同的决策树模型，再组合在一起

- Perturb + Combine

- Breiman(1996)提出装袋法，使用“自助法”(bootstrap)来搅动数据（Breiman于1993年从伯克利统计系退休）

Leo Breiman 1928-2005

Professor of Statistics, UC Berkeley
Verified email at stat.berkeley.edu - Homepage

Data Analysis Statistics Machine Learning

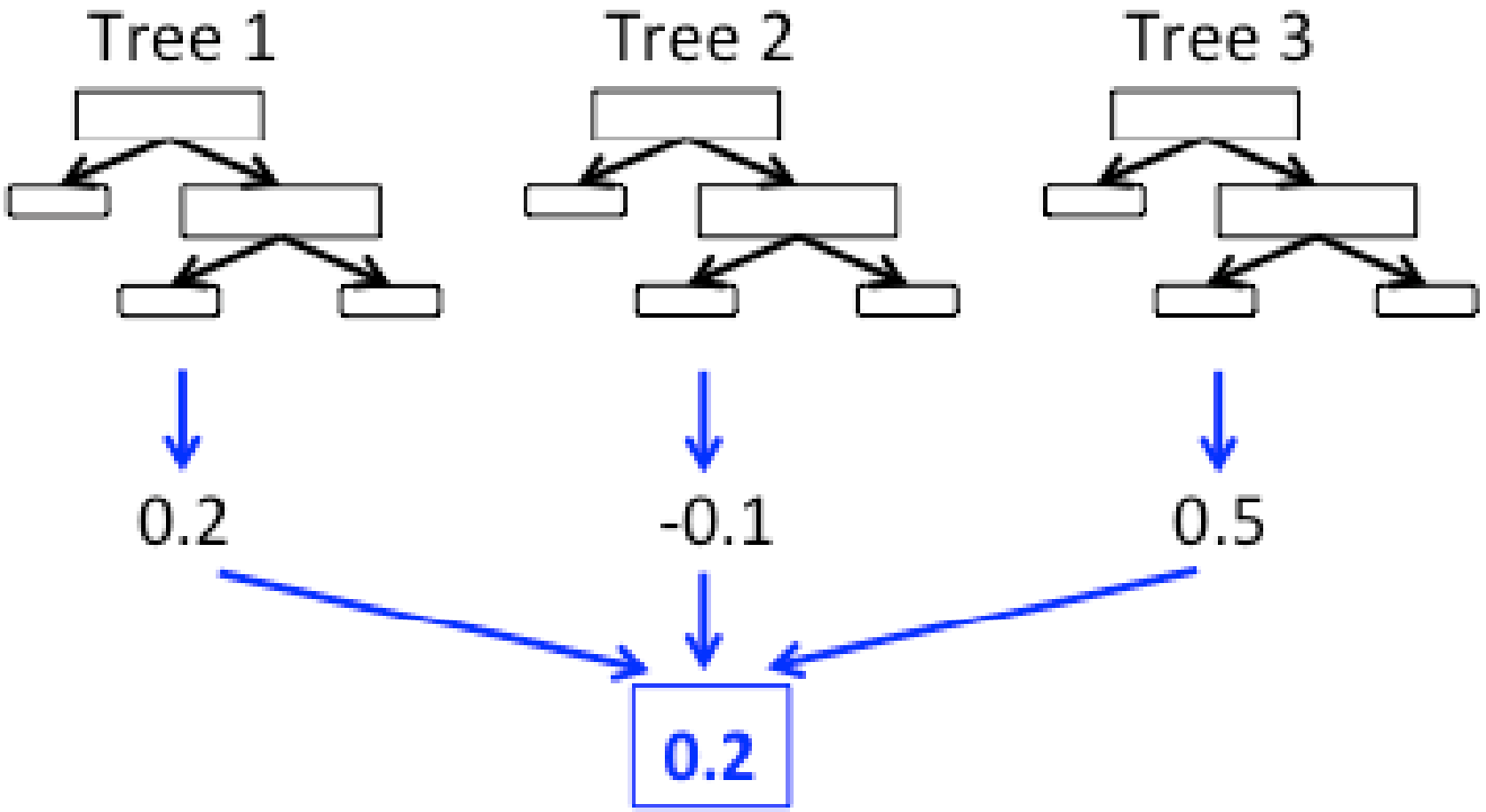
ARTICLES CITED BY

TITLE	CITED BY
Random forests L Breiman Machine learning 45 (1), 5-32, 2001	52451
Classification and Regression Trees L Breiman, JH Friedman, RA Olshen, CJ Stone CRC Press, New York, 1999	~~43480~~ *
Classification and regression trees L Breiman Chapman & Hall/CRC, 1984	43480 *
Bagging predictors L Breiman Machine learning 24 (2), 123-140, 1996	22445

装袋法

- 对训练数据进行有放回的再抽样(resampling with replacement)，得到 *B*个自助样本(bootstrap samples).
- 估计 *B* 棵决策树(不修枝)，比如 *B*=500
- 对于回归树，将*B* 棵决策树的预测结果进行简单算术平均
- 对于分类树，将*B* 棵决策树的预测结果进行多数投票

Ensemble Model:
example for regression



Bagging

- 由于将自助样本(bootstrap samples)的估计结果汇总(aggregating)，故名“bootstrap aggregating”，简记 bagging。

- Bagging的主要功能在于降低方差，以提高模型的预测准确率(accuracy)

通过平均降低方差

- 如果 $Y_1, Y_2, \cdots, Y_n$ 为 i.i.d.，且方差为 σ^2 ，则其样本均值的方差可缩小n倍：

$$\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}$$

- 让决策树尽量生长(不修枝)，使得决策树的偏差(bias)最小

- 通过bagging来控制方差(variance)

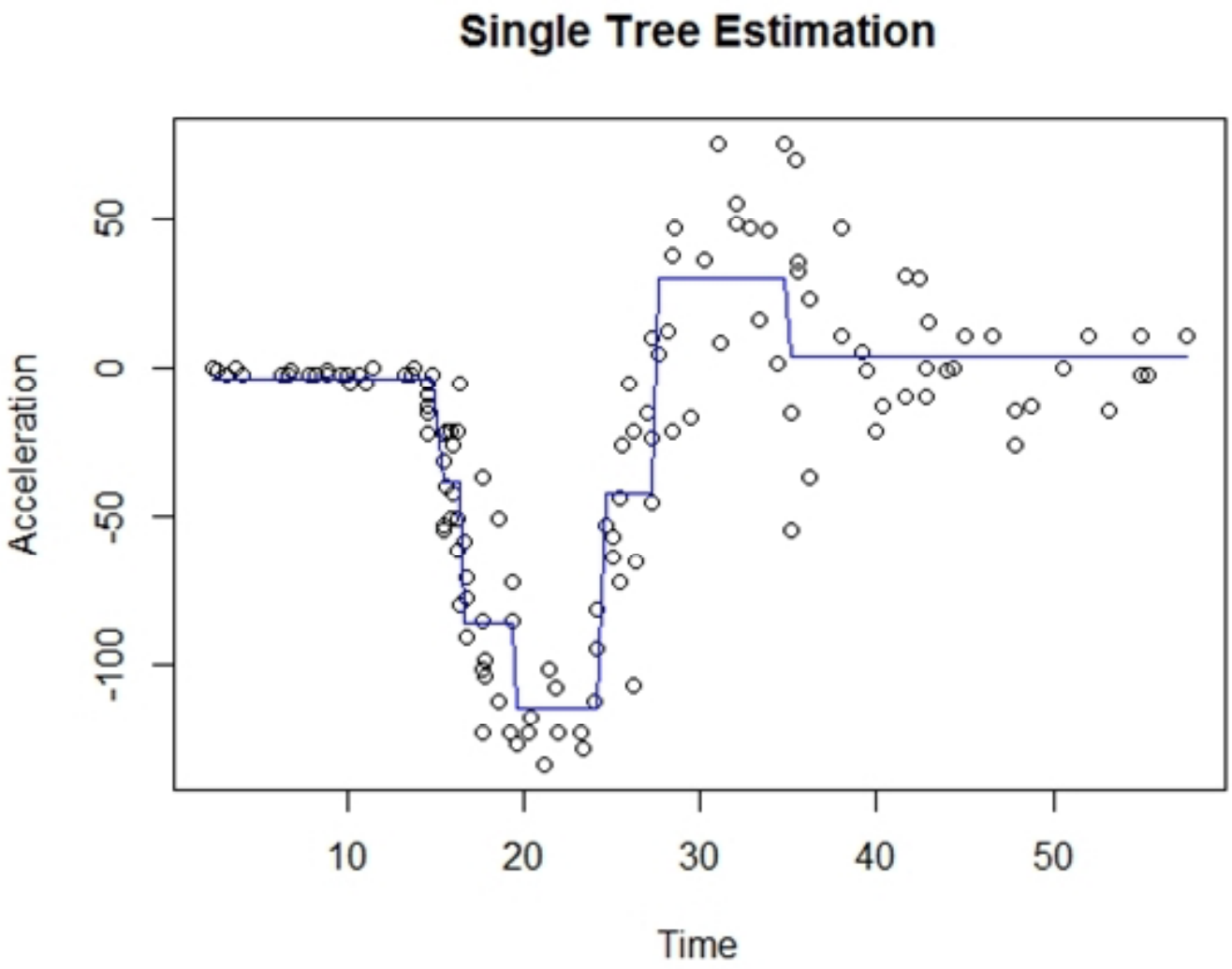


图 12.1 单棵决策树所估计的回归函数

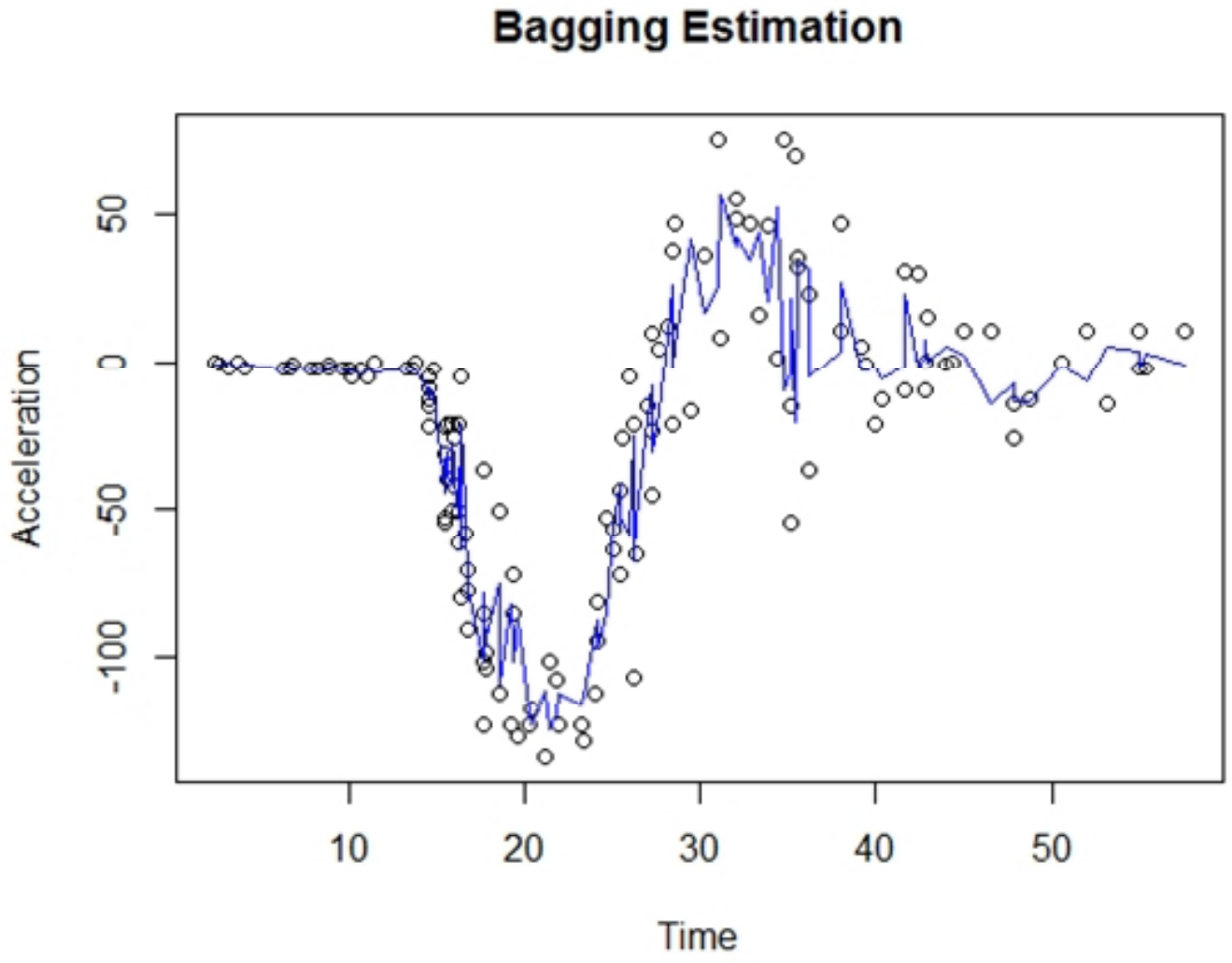


图 12.2 装袋法所估计的回归函数

分类问题：单棵树 vs. 装袋树的错分率

Table 2. Misclassification Rates (%)

Data Set	$\bar{e}_S$	$\bar{e}_B$	Decrease
waveform	29.1	19.3	34%
heart	4.9	2.8	43%
breast cancer	5.9	3.7	37%
ionosphere	11.2	7.9	29%
diabetes	25.3	23.9	6%
glass	30.4	23.6	22%
soybean	8.6	6.8	21%

Table 5. Test Set Misclassification Rates (%)

Data Set	e_S	e_B	Decrease
letters	12.6	6.4	49%
satellite	14.8	10.3	30%
shuttle	.062	.014	77%
DNA	6.2	5.0	19%

- 注：$\overline{e}_S$ 为单棵树的平均测试误差，而 $\overline{e}_B$ 为装袋树的平均测试误差。来源：Breiman (1996)

回归问题：单棵树 vs. 装袋树的均方误差

Table 8. Mean Squared Test Set Error

Data Set	$\bar{e}_S$	$\bar{e}_B$	Decrease
Boston Housing	20.0	11.6	42%
Ozone	23.9	18.8	21%
Friedman #1	11.4	6.1	46%
Friedman #2	31,100	22,100	29%
Friedman #3	.0403	.0242	40%

- 注：$\overline{e}_S$ 为单棵树的平均测试误差，而 $\overline{e}_B$ 为装袋树的平均测试误差。来源：Breiman (1996)

如何使树之间更不相关

- 如果$Y_1, Y_2, \cdots, Y_n$为 i.i.d.，则其样本均值的方差可缩小 n 倍
- 但如果 $Y_1, Y_2, \cdots, Y_n$ 相关，则样本均值的方差一般不会缩小 n 倍
- “三个臭皮匠，顶一个诸葛亮”：成立条件？
- 如何使树之间更不相关？How to decorrelate？

随机森林

- Bagging的决策树之间相关性强，是因为使用了相同的解释变量
- Breiman (2001)提出“随机森林”(Random Forest)
- 在Bagging的基础上(依然使用bootstrap samples)，在每个决策树的节点进行分裂时，仅随机选取部分变量(m个变量)作为候选的分裂变量

随机选择变量(column subsampling)

- “随机特征选择”(random feature selection)的目的是降低决策树之间的相关性 (decorrelate)

- 对于回归树，建议 $m = p/3$ (p为变量个数)

- 对于分类树，建议 $m = \sqrt{p}$

- 最优 m (也记为 *mtry*) 取决于数据，应视为“调节参数”(tuning parameter)

偏差与方差的权衡

- 在每个节点，仅随机选择一小部分变量
($m=\sqrt{p}$ 或 $p/3$)，无疑会导致偏差(未使用全部信息)

- 但这使得决策树之间更不相关，从而降低随机森林的方差

- 总效果可使均方误差(MSE)下降

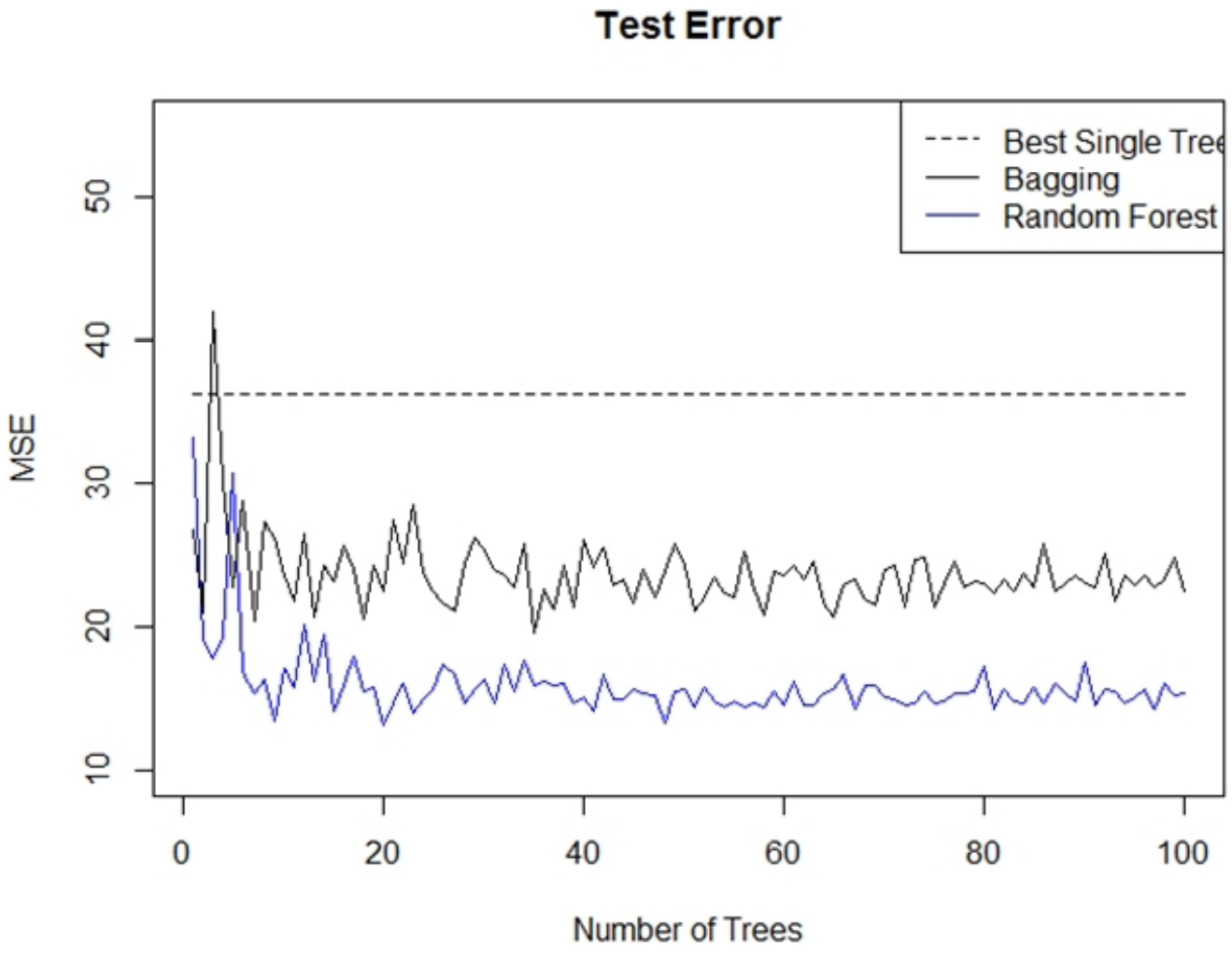


图 12.8 测试误差的比较

Bagging为RF的特例

- Bagging为Random Forest的特例
- 对于随机森林，令 $m = p$ ，则为装袋法
- 在每个节点，装袋法均将所有变量作为候选的分裂变量

随机森林的调节变量

- *mtry* (随机选择的变量个数)
- *number of trees*
- *node size* (minimum size of terminal nodes)

图 12.11 随机森林的测试集误差与 mtry

变量重要性

- 随机森林包含很多决策树，无法像单棵决策树那样进行解释。
- 如何度量“变量重要性”(Variable Importance)？
- 由于每次节点分裂仅使用一个变量，故对于某个变量，可度量随机森林中每棵树由于该变量所导致的基尼指数(或残差平方和)的下降幅度。
- 针对此下降幅度，对每棵树进行平均，以此度量该变量的重要性

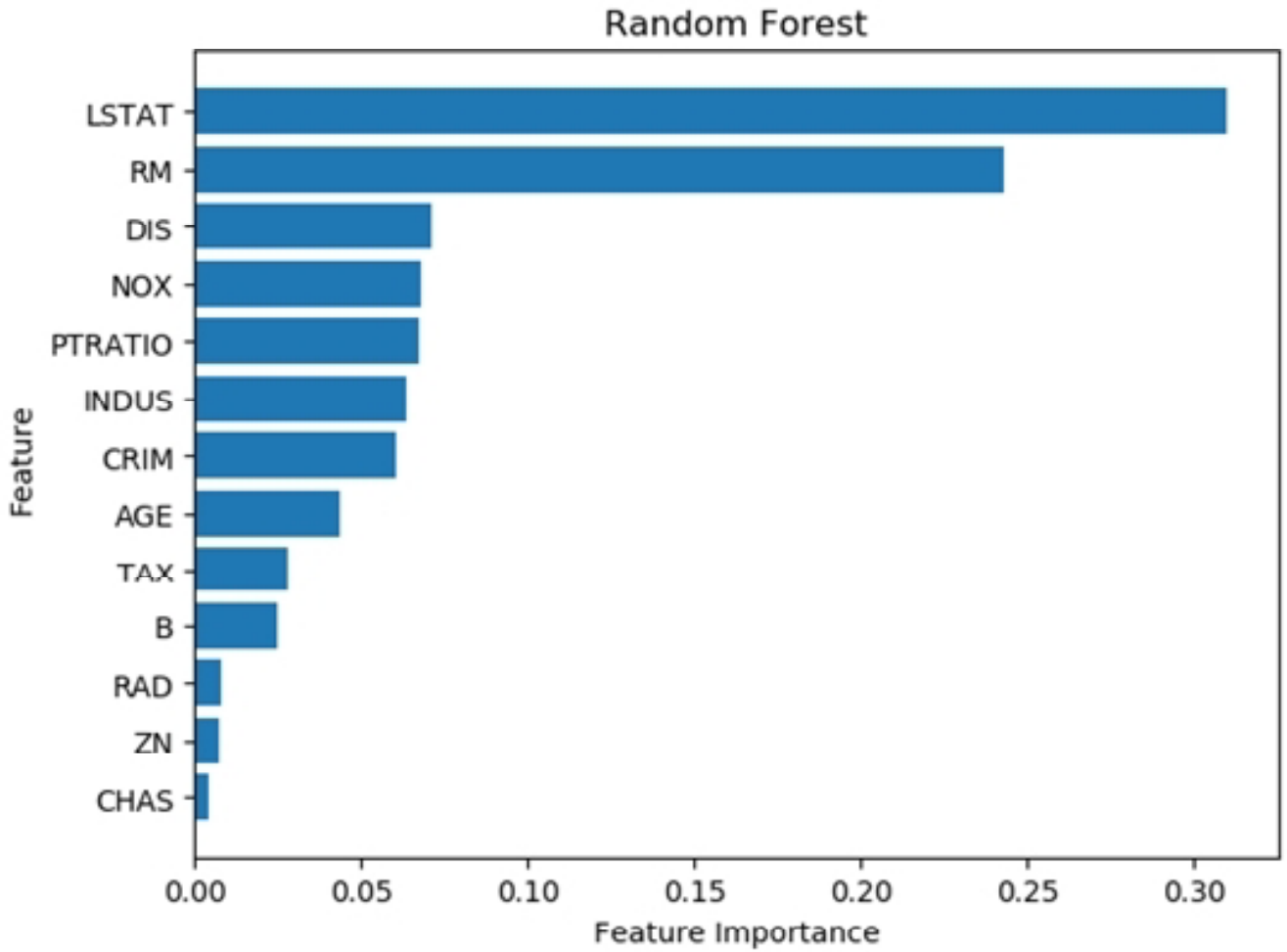


图 12.5 随机森林的变量重要性图

随机森林的应用

- Wager, Stefan and Susan Athey, 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests, *Journal of the American Statistical Association*, 113(523), 1228-1242.

- Meinshausen, Nicolai, 2006. Quantile Regression Forests, *Journal of Machine Learning Research*, 7, 983-999.

- Chen, Qiang and Zhijie Xiao, 2020. Robust Nonparametric Confidence Intervals for Treatment Effects in Panel Data Using Quantile Regression Forests, working paper.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1228–1242, Theory and Methods
https://doi.org/10.1080/01621459.2017.1319839

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager and Susan Athey

Stanford University, Stanford, CA

ABSTRACT

Many scientific and engineering challenges—ranging from personalized medicine to customized marketing recommendations—require an understanding of treatment effect heterogeneity. In this article, we develop a nonparametric *causal forest* for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. In the potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect and have an asymptotically Gaussian and centered sampling distribution. We also discuss a practical method for constructing asymptotic confidence intervals for the true treatment effect that are centered at the causal forest estimates. Our theoretical results rely on a generic Gaussian theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference. In experiments, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially in the presence of irrelevant covariates.

ARTICLE HISTORY
Received December 2015
Revised March 2017

KEYWORDS
Adaptive nearest neighbors matching; Asymptotic normality; Potential outcomes; Unconfoundedness

Quantile Regression Forests

Nicolai Meinshausen NICOLAI@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
8092 Zürich, Switzerland

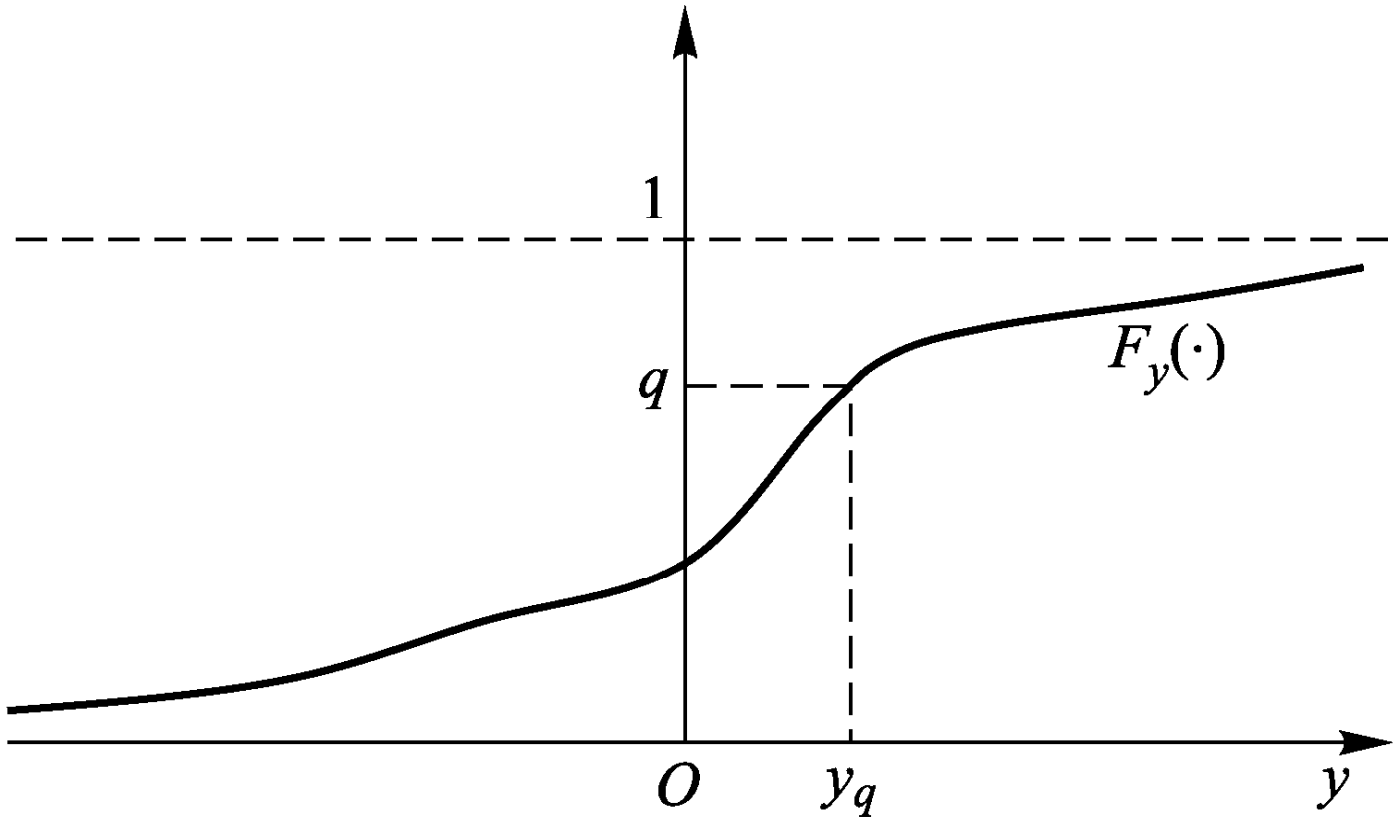
Editor: Greg Ridgeway

Abstract

Random forests were introduced as a machine learning tool in Breiman (2001) and have since proven to be very popular and powerful for high-dimensional regression and classification. For regression, random forests give an accurate approximation of the conditional mean of a response variable. It is shown here that random forests provide information about the full conditional distribution of the response variable, not only about the conditional mean. Conditional quantiles can be inferred with quantile regression forests, a generalisation of random forests. Quantile regression forests give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. The algorithm is shown to be consistent. Numerical examples suggest that the algorithm is competitive in terms of predictive power.

Keywords: quantile regression, random forests, adaptive neighborhood regression

总体分位数



总体q分位数与累积分布函数

总体分位数的定义

- 假设*Y*为连续型随机变量，其累积分布函数为 $F_y(\cdot)$，则*Y*的“总体*q*分位数”(population q^{th} quantile，0<*q*<1)，记为 y_q ，满足以下定义式：

$$q = \mathrm{P}(Y \le y_q) = F_y(y_q)$$

- 总体*q*分位数 y_q 正好将总体分布分为两部分，小于或等于 y_q 的概率为*q*，而大于 y_q 的概率为 1-*q*

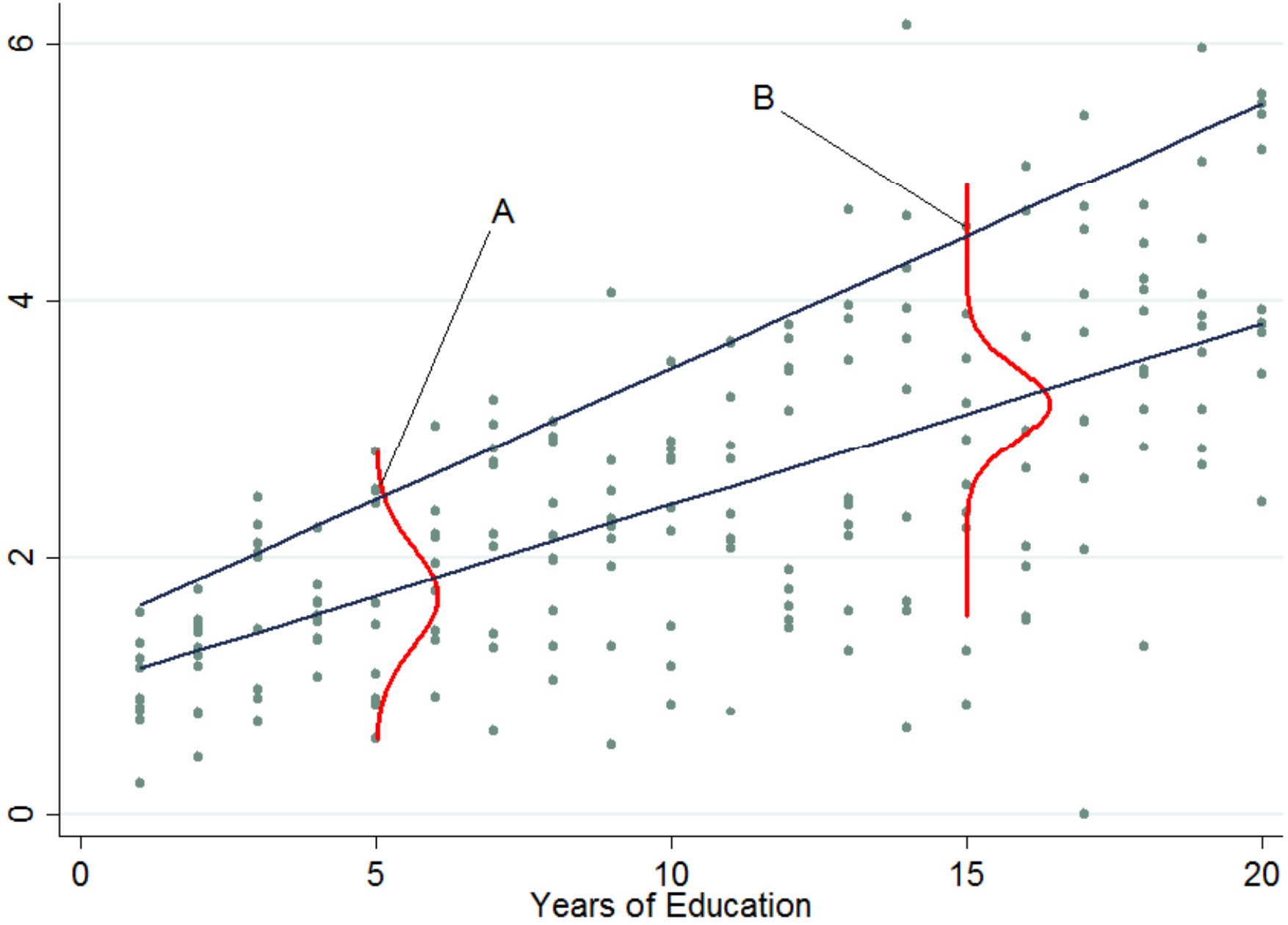
- 如果 $F_y(\cdot)$ 严格单调递增，则有 $y_q = F_y^{-1}(q)$

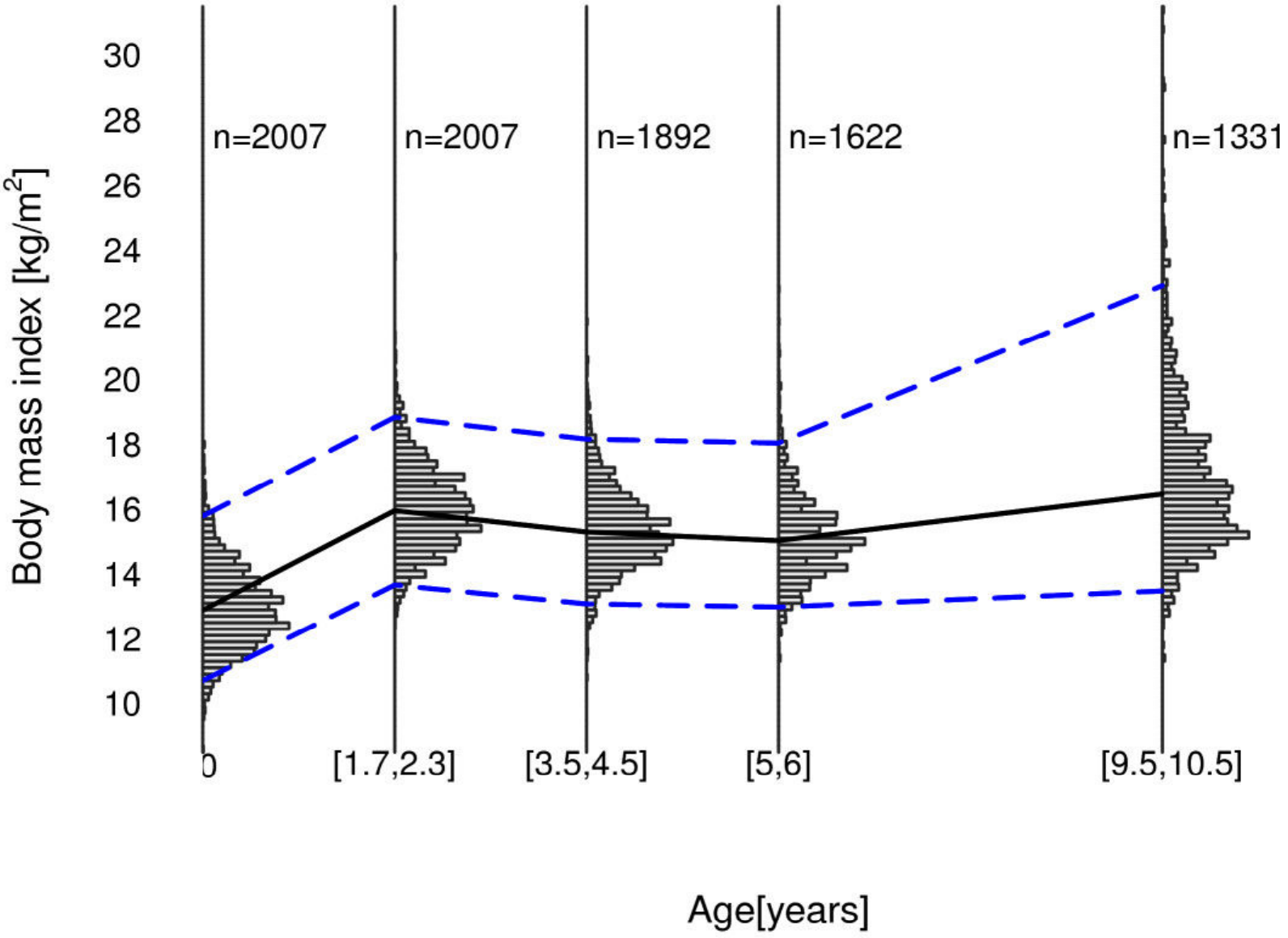
条件分布的总体分位数

- 条件分布 $y|\mathbf{x}$ 的总体 q 分位数，记为y_q ，满足以下定义式：

$$q = F_{y|x}(y_q)$$

- 由于条件累积分布函数$F_{y|x}(\cdot)$ 依赖于$\boldsymbol{x}$，故条件分布 $y|\mathbf{x}$ 的总体 q 分位数也依赖于$\boldsymbol{x}$，可写为 $y_q(\boldsymbol{x})$，称为“条件分位数函数”(conditional quantile function)。


6
4
2
0
B
A
0
5
10
15
20
Years of Education


Body mass index [kg/m^2]
30
28
26
24
22
20
18
16
14
12
10
n=2007
n=2007
n=1892
n=1622
n=1331
0
[1.7,2.3]
[3.5,4.5]
[5,6]
[9.5,10.5]
Age[years]

Random Forest as Weighted Average

The prediction of a single tree $T(\theta)$ for a new data point $X = x$ is obtained by averaging over the observed values in leaf $\ell(x, \theta)$. Let the weight vector $w_i(x, \theta)$ be given by a positive constant if observation X_i is part of leaf $\ell(x, \theta)$ and 0 if it is not. The weights sum to one, and thus

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{\ell(x,\theta)}\}}}{\#\{j : X_j \in R_{\ell(x,\theta)}\}}. \tag{4}$$

The prediction of a single tree, given covariate $X = x$, is then the weighted average of the original observations $Y_i, i = 1, \ldots, n$,

$$\text{single tree:} \quad \hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta)\, Y_i.$$

Using random forests, the conditional mean $E(Y|X = x)$ is approximated by the averaged prediction of k single trees, each constructed with an i.i.d. vector θ_t, $t = 1, \ldots, k$. Let $w_i(x)$ be the average of $w_i(\theta)$ over this collection of trees,

$$w_i(x) = k^{-1} \sum_{t=1}^{k} w_i(x, \theta_t). \tag{5}$$

The prediction of random forests is then

$$\text{Random Forests:} \quad \hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

Quantile Regression Forest

It was shown above that random forests approximates the conditional mean $E(Y|X = x)$ by a weighted mean over the observations of the response variable Y. One could suspect that the weighted observations deliver not only a good approximation to the conditional mean but to the full conditional distribution. The conditional distribution function of Y, given $X = x$, is given by

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x).$$

The last expression is suited to draw analogies with the random forest approximation of the conditional mean $E(Y|X = x)$. Just as $E(Y|X = x)$ is approximated by a weighted mean over the observations of Y, define an approximation to $E(1_{\{Y \leq y\}}|X = x)$ by the weighted mean over the observations of $1_{\{Y \leq y\}}$,

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)\, 1_{\{Y_i \leq y\}}, \tag{6}$$

using the same weights $w_i(x)$ as for random forests, defined in equation (5). This approximation is at the heart of the quantile regression forests algorithm.

Quantile Control Method (QCM)

- Chen, Qiang, and Zhijie Xiao, "Robust Nonparametric Confidence Intervals for Treatment Effects in Panel Data Using Quantile Regression Forest," 2020, working paper

- 使用“随机森林”（random forest）进行分位数回归，估计回归控制法的置信区间

- Forthcoming R package qcm

Introduction

- Approaches to estimate treatment effects in panel data with only one treated unit have become popular in applied works, which include synthetic control method (Abadie and Gardeazabal, 2003, Abadie et al., 2010), and regression control method (Hsiao et al., 2012).

- However, no pointwise standard errors or confidence intervals for the treatment effects have been provided in the literature yet.

Contributions

- We propose a direct nonparametric construction of pointwise robust confidence intervals using quantile regression forest (QRF), and exploits cross-sectional correlation to construct counterfactuals as in Hsiao et al. (2012).

- An advantage of this approach is that it does not require the post-treatment period to be large for inference.

- Monte Carlos simulations show good coverage probability for the confidence intervals, which are robust to heteroskedasticity, autocorrelation, and model misspecification.

A PANEL DATA APPROACH FOR PROGRAM EVALUATION: MEASURING THE BENEFITS OF POLITICAL AND ECONOMIC INTEGRATION OF HONG KONG WITH MAINLAND CHINA

CHENG HSIAO,[a,b,c] H. STEVE CHING[b] AND SHUI KI WAN[d]*

[a] *University of Southern California, Los Angeles, CA, USA*
[b] *City University of Hong Kong, Kowloon, Hong Kong*
[c] *WISE, Xiamen University, Xiamen, China*
[d] *Hong Kong Baptist University, Kowloon, Hong Kong*

SUMMARY

We propose a simple-to-implement panel data method to evaluate the impacts of social policy. The basic idea is to exploit the dependence among cross-sectional units to construct the counterfactuals. The cross-sectional correlations are attributed to the presence of some (unobserved) common factors. However, instead of trying to estimate the unobserved factors, we propose to use observed data. We use a panel of 24 countries to evaluate the impact of political and economic integration of Hong Kong with mainland China. We find that the political integration hardly had any impact on the growth of the Hong Kong economy. However, the economic integration has raised Hong Kong's annual real GDP by about 4%. Copyright © 2011 John Wiley & Sons, Ltd.

Case Study from Hsiao et al.(2012):

Political Integration of HK_CN with mainland China in 1997

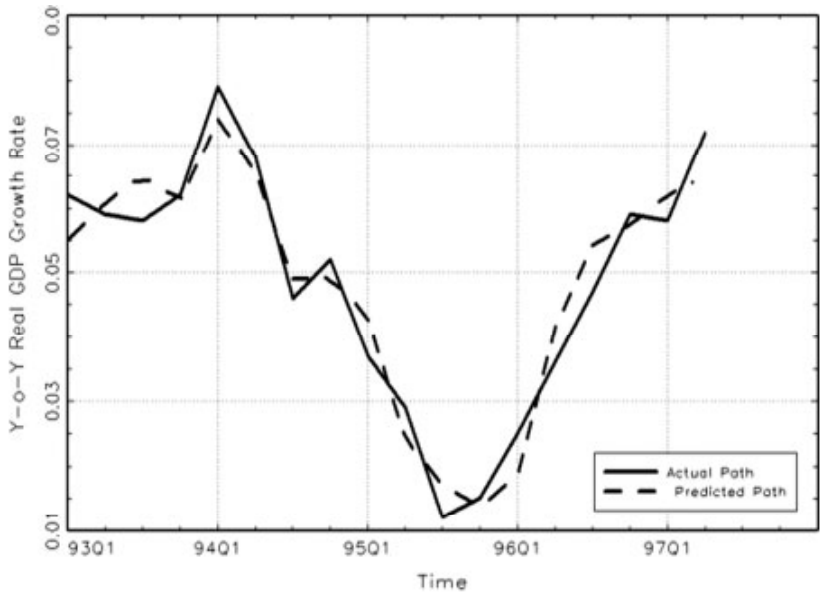


Figure 1. AICC: actual and predicted real GDP from 1993:Q1 to 1997:Q2

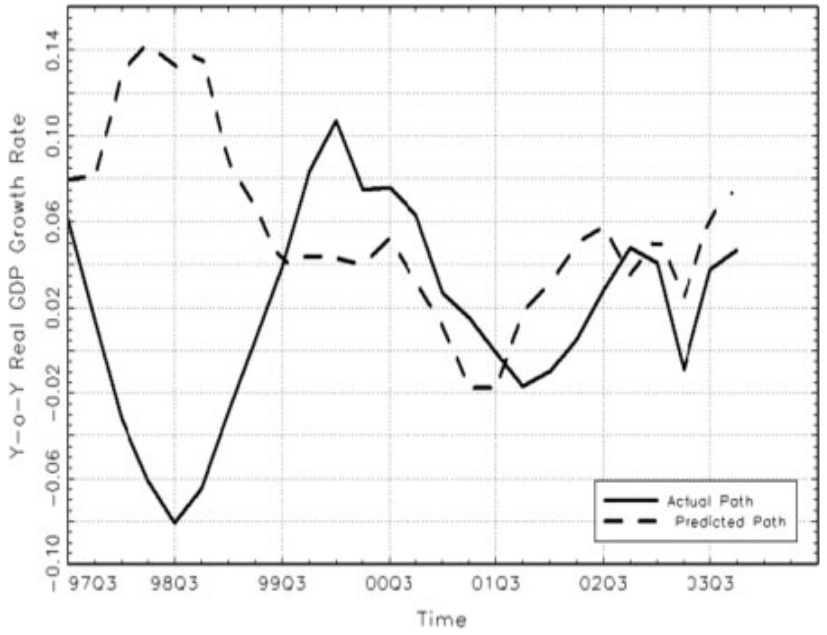


Figure 2. AICC : actual and counterfactual real GDP from 1997:Q3 to 2003:Q4

Case Study from Hsiao et al. (2012):

Economic Integration of HK_CN with mainland China in 2004

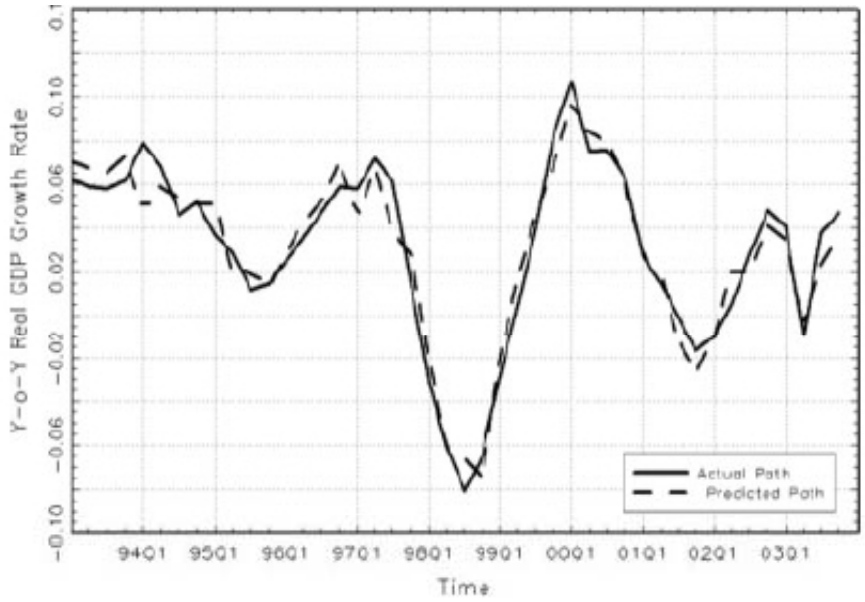


Figure 7. AICC: actual and predicted real GDP from 1993:Q1 to 2003:Q4

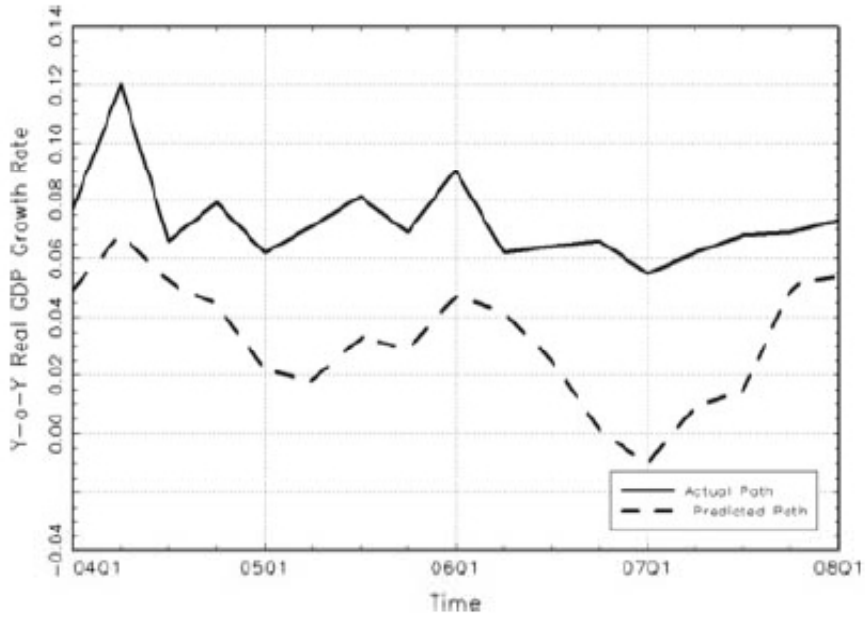


Figure 8. AICC: actual and counterfactual real GDP from 2004:Q1 to 2008:Q1

A Quantile Regression Approach

- We use regression quantiles of the posttreatment counterfactual outcome to directly construct valid confidence intervals of the treatment effects:

$$\mathrm{P}\left(Q_{y_{1t}^0}(\alpha/2) \le y_{1t}^0 \le Q_{y_{1t}^0}(1-\alpha/2)\right) = 1-\alpha$$

- Since $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$, we could plug in the expression $y_{1t}^0 = y_{1t}^1 - \Delta_{1t}$ to get

$$\mathrm{P}\left(Q_{y_{1t}^0}(\alpha/2) \le y_{1t}^1 - \Delta_{1t} \le Q_{y_{1t}^0}(1-\alpha/2)\right) = 1-\alpha$$

Confidence Interval for Treatment Effects

- Since $y_{1t}^1 = y_{1t}$ during post-treatment period, rearranging ,

$$\mathrm{P}\left(y_{1t}^1 - Q_{y_{1t}^0}(1-\alpha/2) \le \Delta_{1t} \le y_{1t}^1 - Q_{y_{1t}^0}(\alpha/2) \right) = 1-\alpha$$

- With consistent estimators:

$$\mathrm{P}\left(y_{1t}^1 - \widehat{Q_{y_{1t}^0}(1-\alpha/2)} \le \Delta_{1t} \le y_{1t}^1 - \widehat{Q_{y_{1t}^0}(\alpha/2)} \right) = 1-\alpha$$

Consistency of Quantile Regression Forests

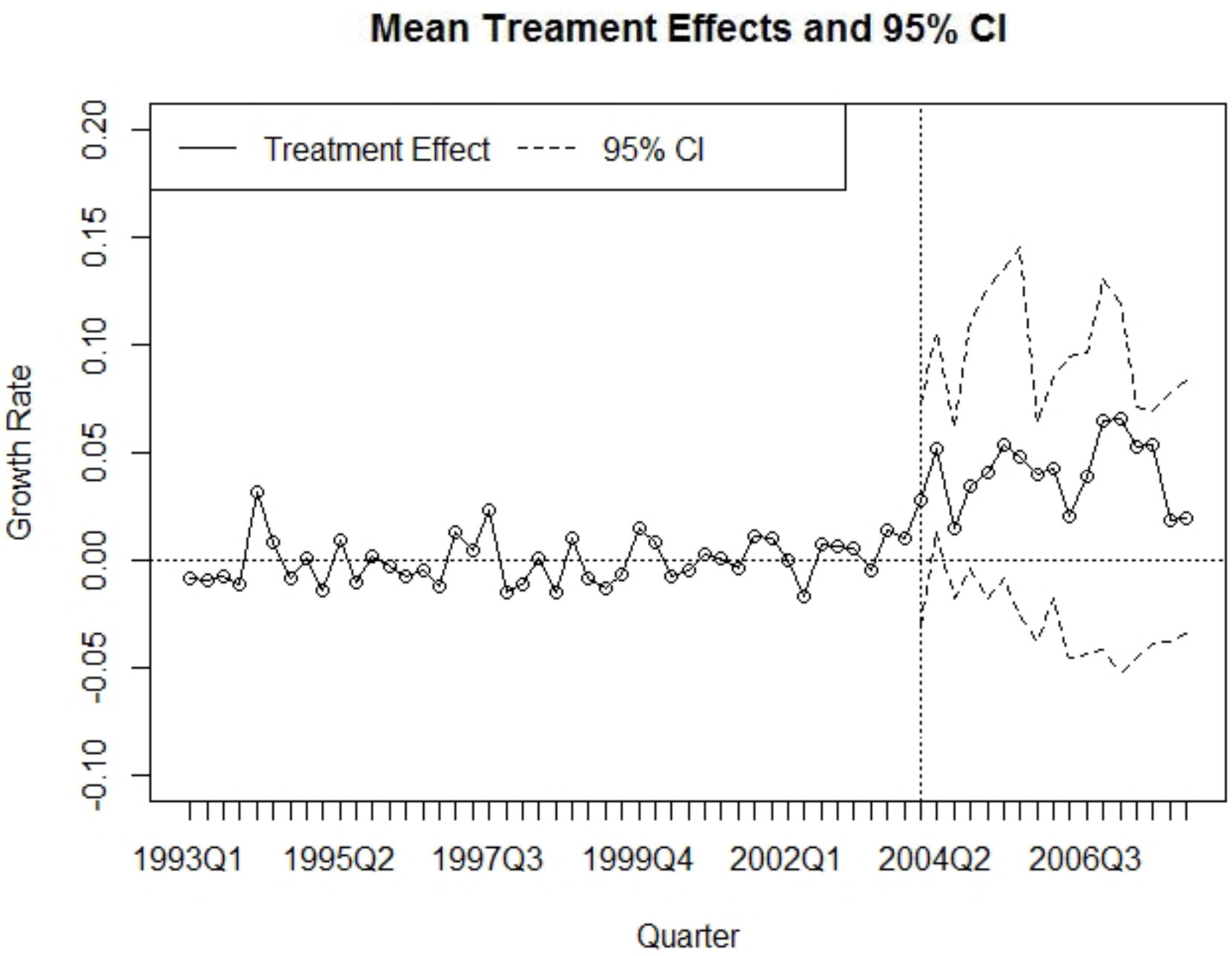
- Meinshausen (2006) uses random forest to estimate conditional CDF, then invert it to get QR.
- Meinshausen (2006) provides a proof of consistency for QRF under i.i.d. assumptions
- We extend the proof to the time series case

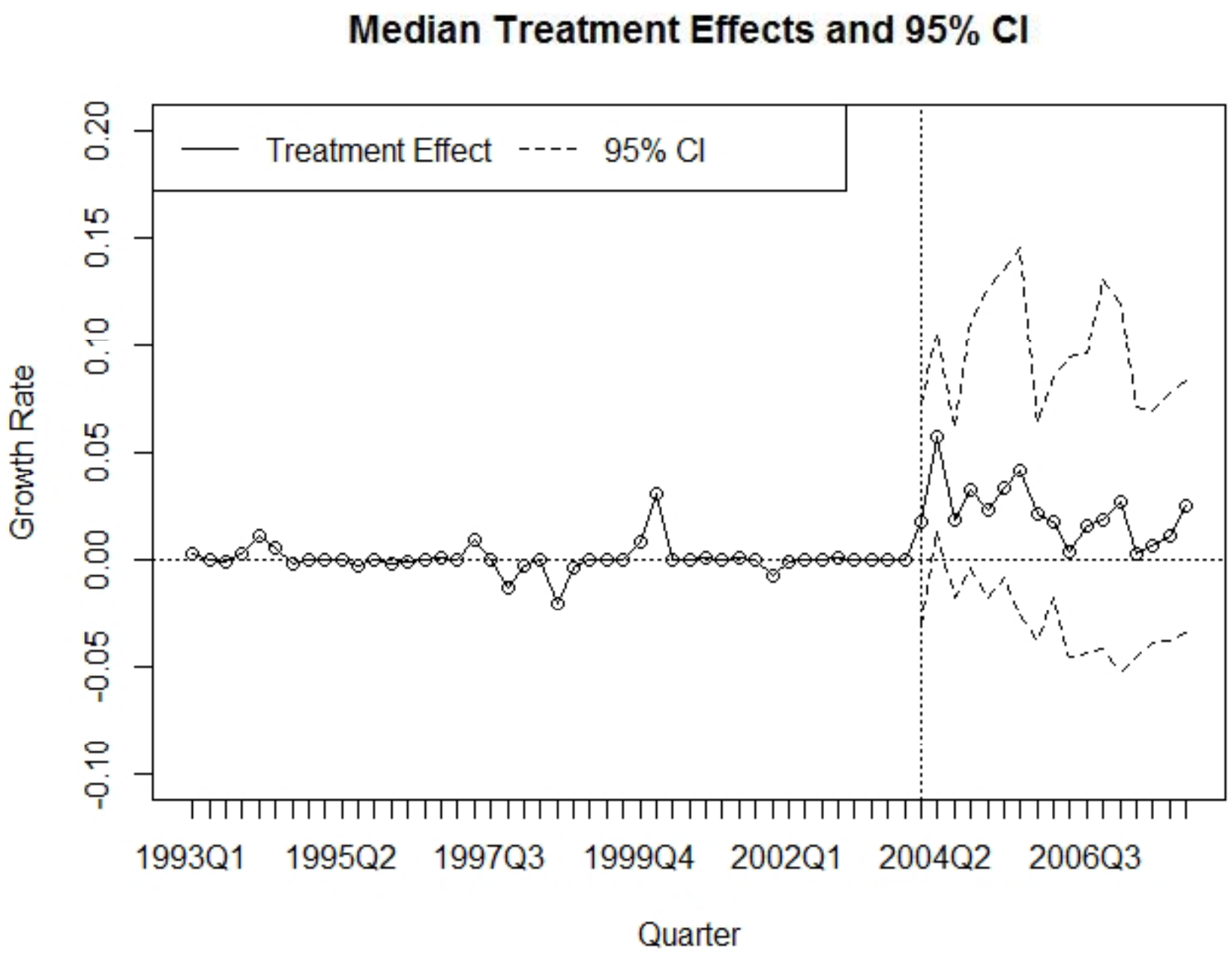
Application: Economic Integration of HK_CN with Mainland China in 2004

- 《内地与香港关于建立更紧密经贸关系的安排》（简称CEPA）的主要内容包括：允许众多中国香港产品零关税进入内地，放宽内地对中国香港服务业的准入领域以及贸易便利化三方面。

- Pretreatment period: 1993Q1 - 2003Q4 (44 quarters)

- Treatment period: 2004Q1 - 2008Q1 (17 quarters)


Mean Treament Effects and 95% CI
Treatment Effect
95% CI
Growth Rate
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
1993Q1
1995Q2
1997Q3
1999Q4
2002Q1
2004Q2
2006Q3
Quarter


Median Treatment Effects and 95% CI
Treatment Effect
95% CI
Growth Rate
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
1993Q1
1995Q2
1997Q3
1999Q4
2002Q1
2004Q2
2006Q3
Quarter

Simulations

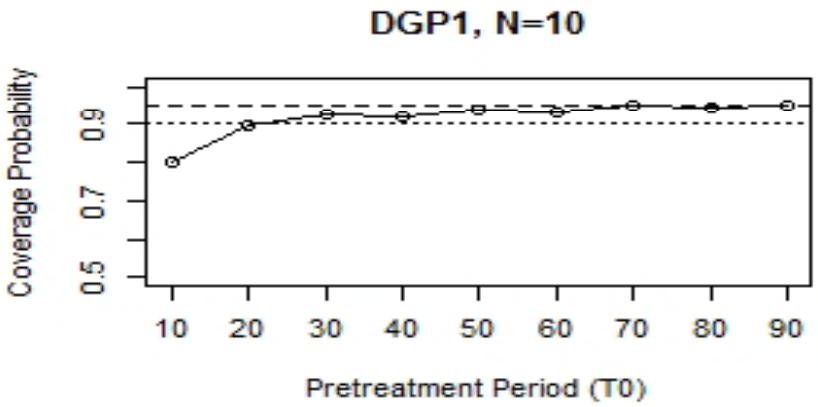
- We conduct simulations for the four DGPs in Hsiao et al. (2012)
- Also simulations for DGPs under heteroskedasticity, autocorrelation, and model misspecification.
- Good coverage probability in finite samples

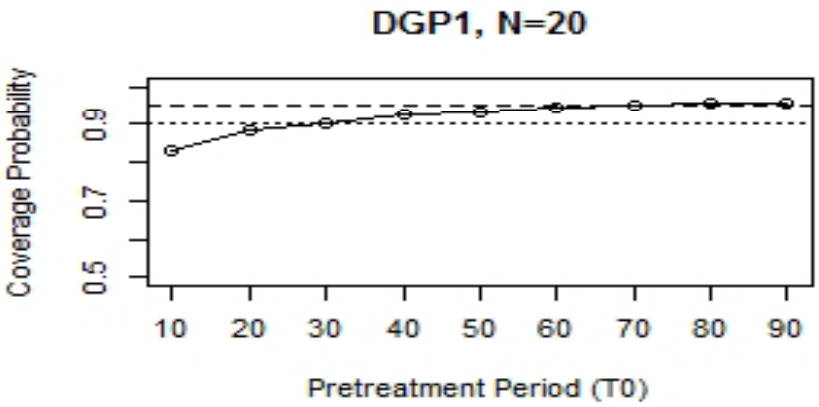
Table 1. Coverage Probability for DGP 1

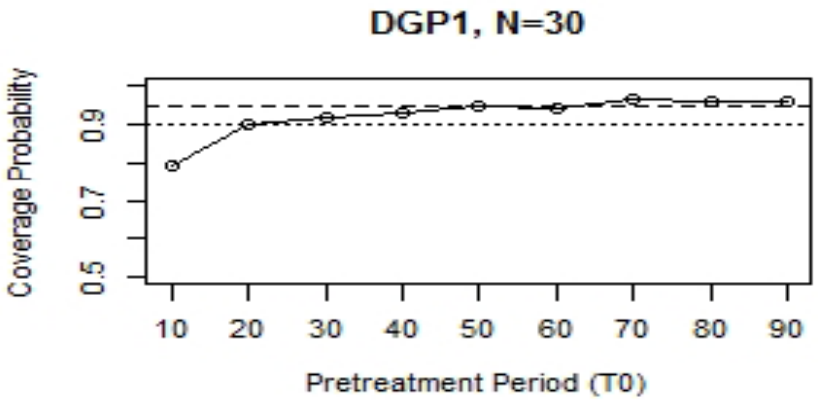
$$y_{it}^{0} = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + u_{it} \qquad f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t} \qquad f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t}$$

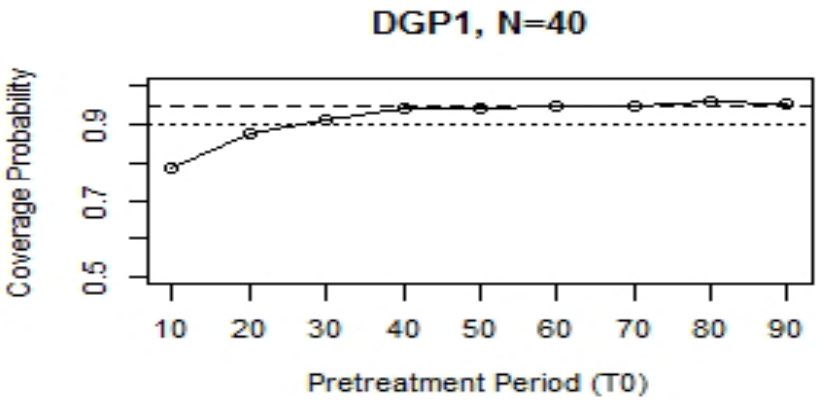
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.803	0.897	0.926	0.922	0.936	0.934	0.948	0.943	0.951
$N = 20$	0.828	0.883	0.902	0.924	0.931	0.946	0.949	0.954	0.955
$N = 30$	0.794	0.9	0.919	0.931	0.951	0.943	0.966	0.96	0.962
$N = 40$	0.786	0.875	0.912	0.941	0.945	0.951	0.947	0.962	0.952
$N = 50$	0.811	0.904	0.911	0.933	0.947	0.957	0.935	0.957	0.955
$N = 60$	0.794	0.89	0.925	0.944	0.929	0.945	0.944	0.962	0.95

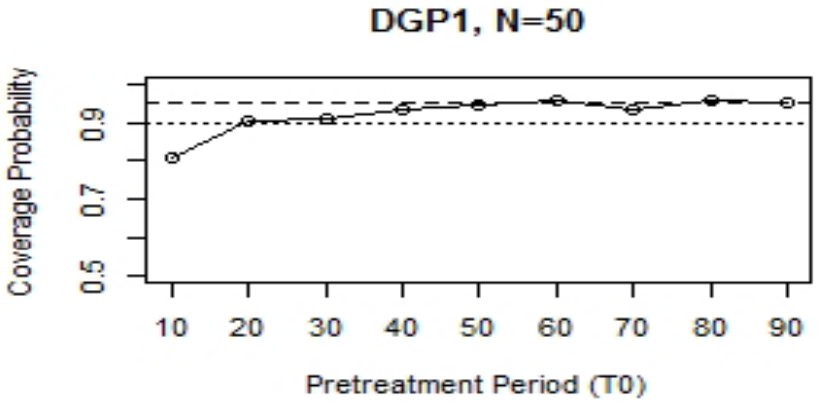
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP1, N=10
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)


DGP1, N=20
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)


DGP1, N=30
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)


DGP1, N=40
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)


DGP1, N=50
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)

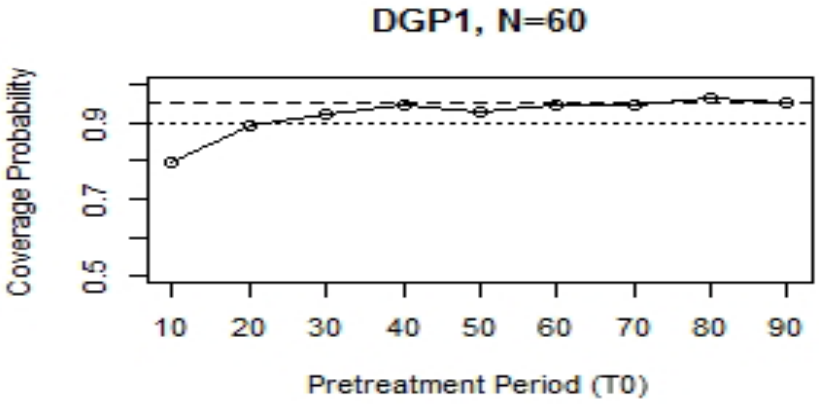

DGP1, N=60
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)

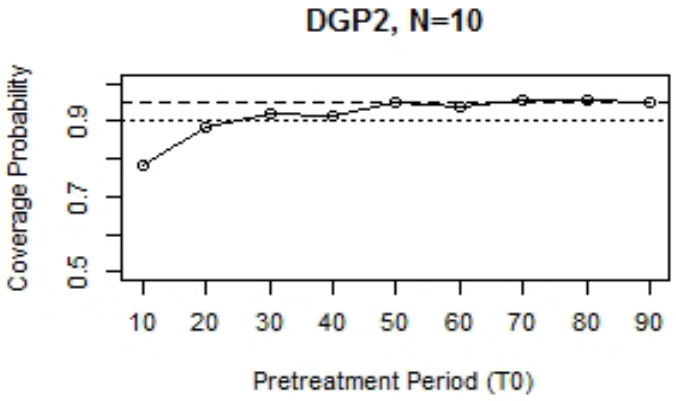
Table 2. Coverage Probability for DGP 2

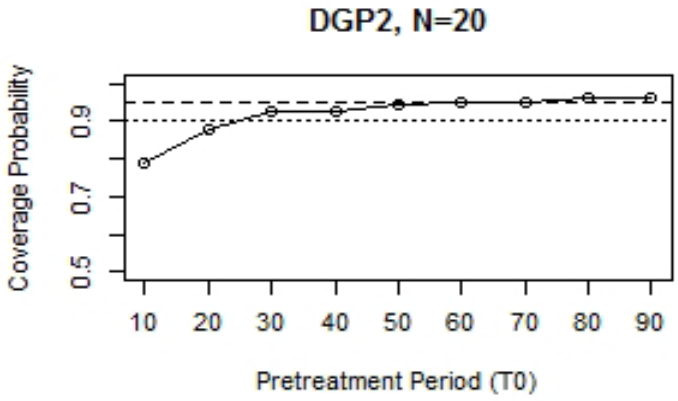
$$y_{it}^{0} = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + b_{i3} f_{3t} + u_{it} \qquad f_{1t} = 0.8 f_{1,t-1} + \varepsilon_{1t}$$

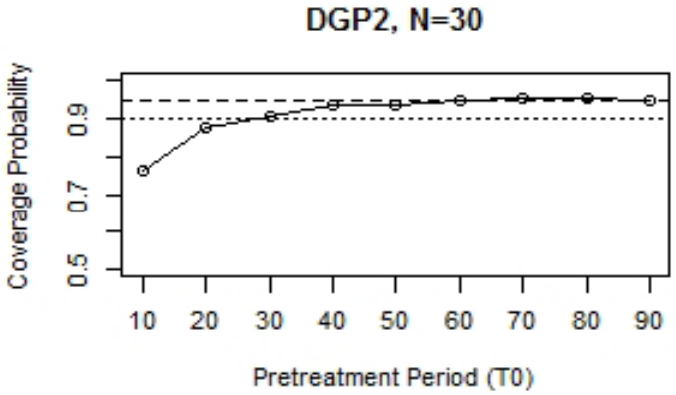
$$f_{2t} = -0.6 f_{2,t-1} + \varepsilon_{2t} + 0.8\varepsilon_{2,t-1} \qquad f_{3t} = \varepsilon_{3t} + 0.9\varepsilon_{3,t-1} + 0.4\varepsilon_{3,t-2}$$

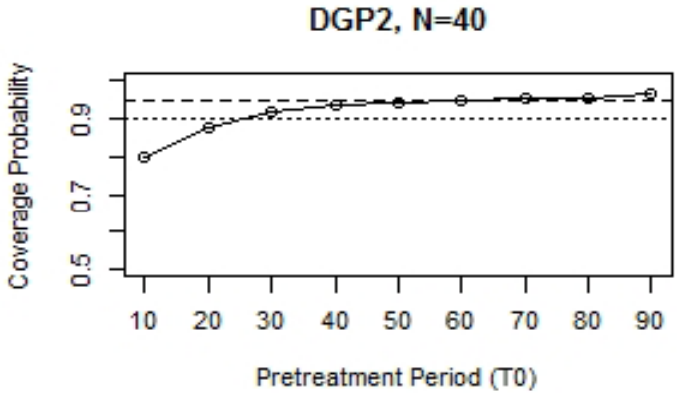
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.781	0.883	0.921	0.917	0.948	0.938	0.955	0.957	0.949
$N = 20$	0.789	0.878	0.924	0.927	0.947	0.951	0.95	0.965	0.963
$N = 30$	0.765	0.875	0.907	0.938	0.936	0.951	0.956	0.956	0.95
$N = 40$	0.799	0.875	0.917	0.934	0.945	0.946	0.954	0.953	0.969
$N = 50$	0.786	0.875	0.918	0.941	0.949	0.947	0.963	0.967	0.955
$N = 60$	0.756	0.874	0.931	0.936	0.938	0.953	0.955	0.96	0.955

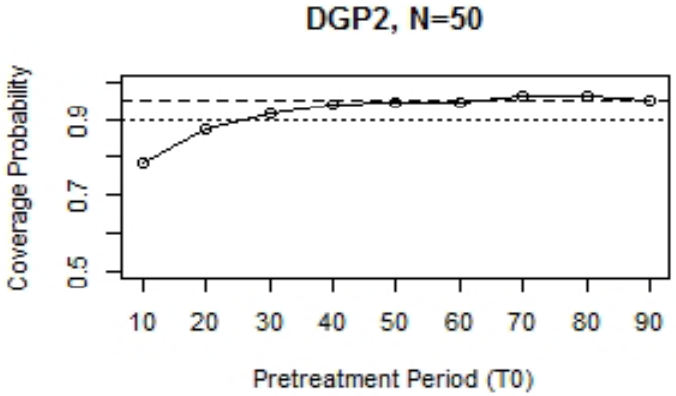
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP2, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP2, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP2, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP2, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP2, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

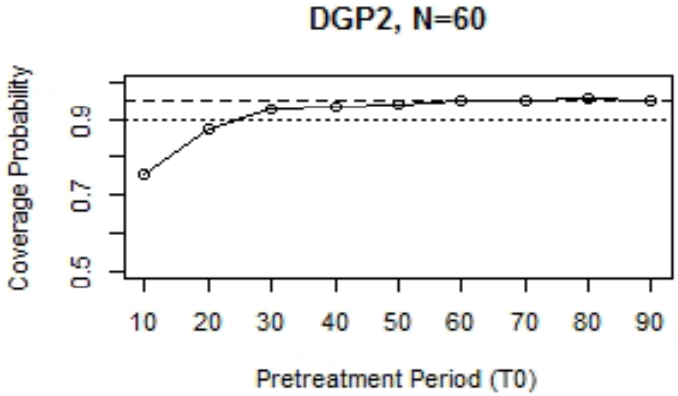
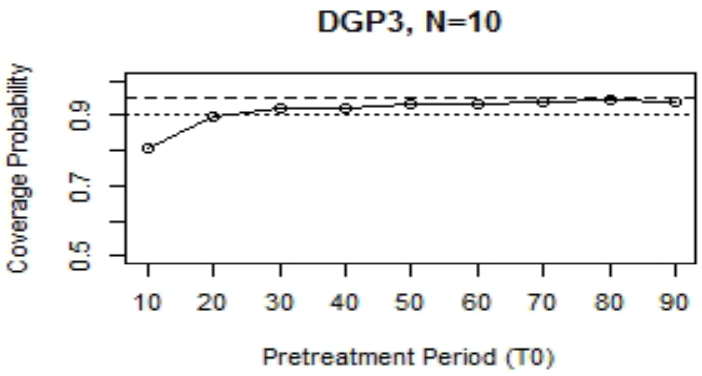

DGP2, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

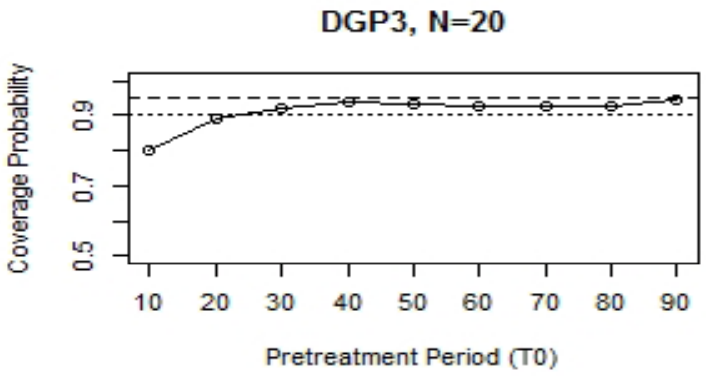
Table 3. Coverage Probability for DGP 3

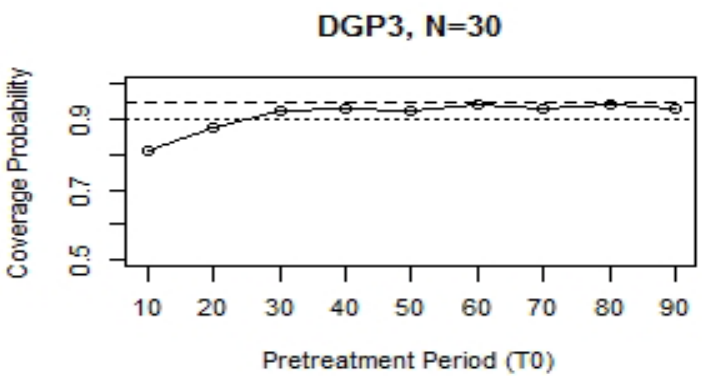
$$y_{it}^0 = \alpha_i + b_i f_t + u_{it} \qquad f_t = i.i.d.\ N(0,1)$$

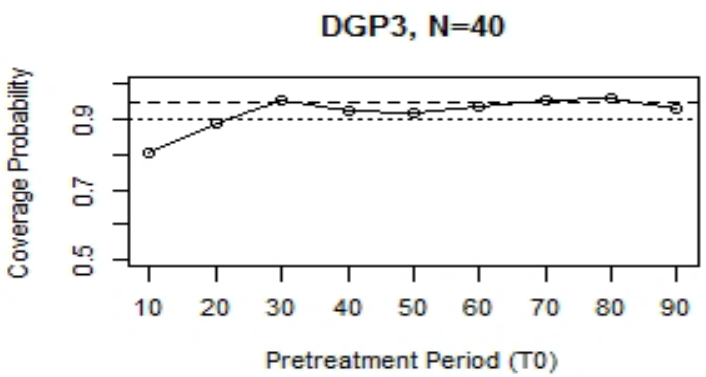
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.804	0.897	0.921	0.921	0.935	0.934	0.94	0.942	0.938
$N = 20$	0.802	0.892	0.919	0.939	0.93	0.928	0.929	0.925	0.945
$N = 30$	0.813	0.876	0.927	0.929	0.923	0.941	0.933	0.945	0.929
$N = 40$	0.804	0.887	0.952	0.924	0.917	0.935	0.956	0.962	0.93
$N = 50$	0.827	0.896	0.92	0.926	0.935	0.948	0.941	0.938	0.951
$N = 60$	0.814	0.895	0.916	0.938	0.935	0.947	0.937	0.932	0.947

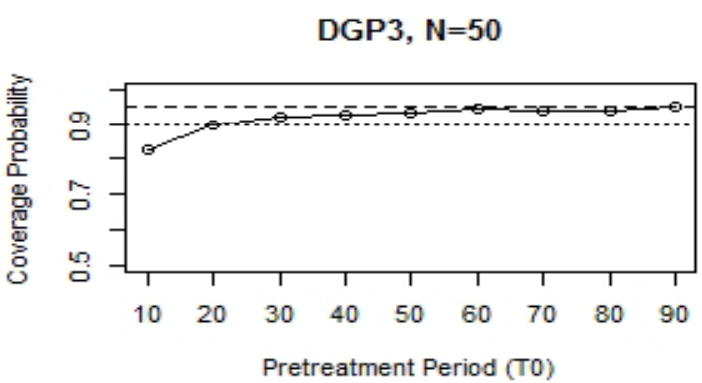
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP3, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

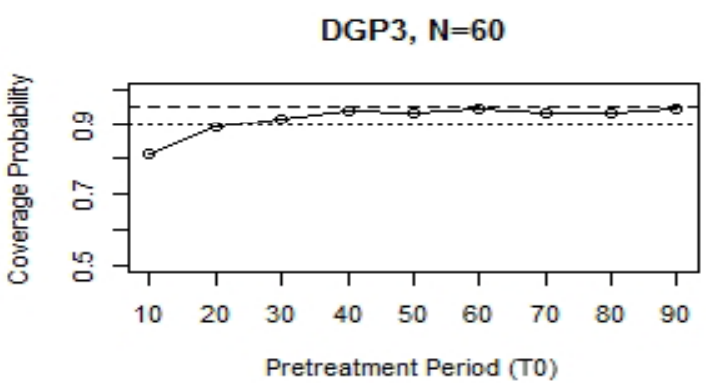
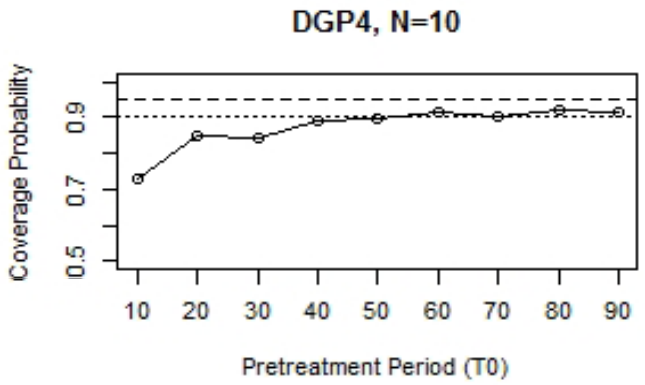

DGP3, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

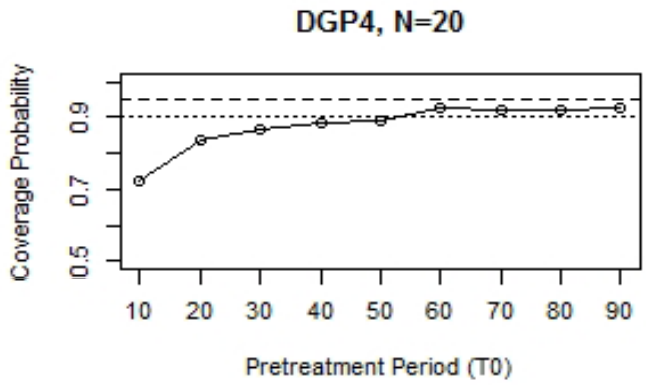
Table 4. Coverage Probability for DGP 4

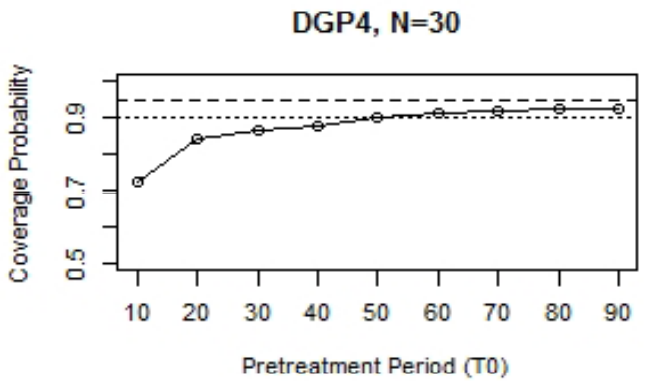
$$y_{it}^{0} = \alpha_i + b_i f_t + u_{it} \qquad f_t = 0.95 f_{t-1} + \varepsilon_{1t}$$

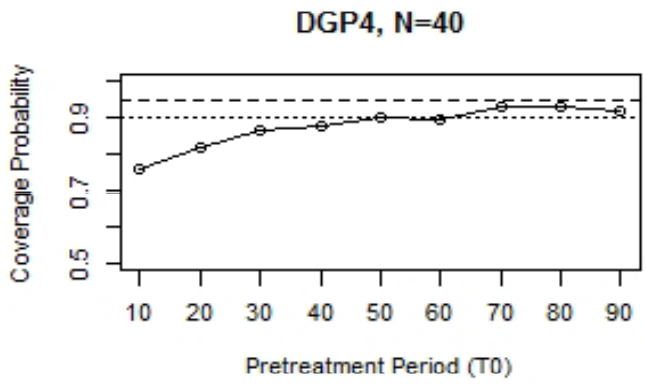
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.727	0.848	0.843	0.893	0.895	0.915	0.901	0.921	0.914
$N = 20$	0.723	0.835	0.868	0.887	0.889	0.926	0.919	0.921	0.925
$N = 30$	0.72	0.842	0.863	0.874	0.902	0.913	0.92	0.925	0.924
$N = 40$	0.758	0.817	0.863	0.877	0.899	0.892	0.933	0.931	0.92
$N = 50$	0.761	0.831	0.874	0.882	0.893	0.902	0.923	0.924	0.923
$N = 60$	0.748	0.816	0.875	0.889	0.899	0.925	0.907	0.913	0.939

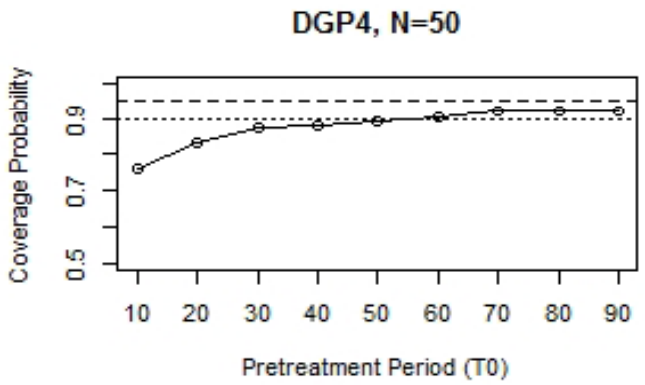
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP4, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP4, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP4, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP4, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP4, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

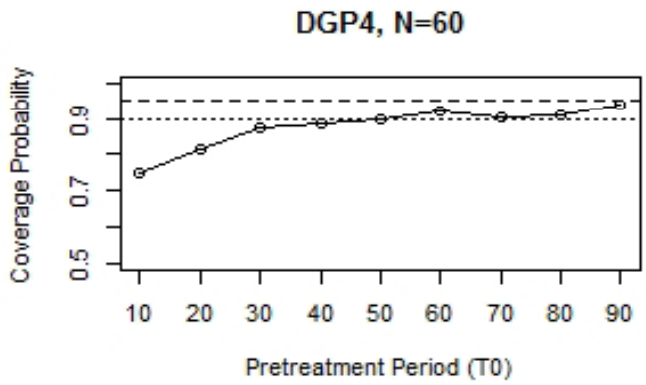
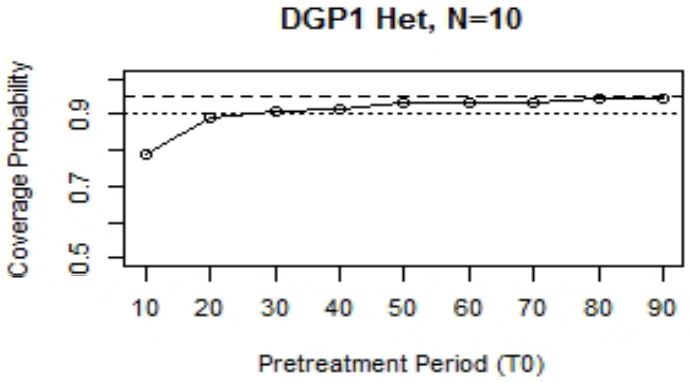
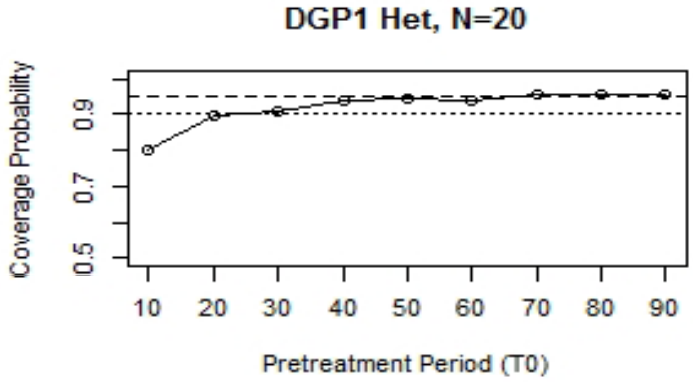

DGP4, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

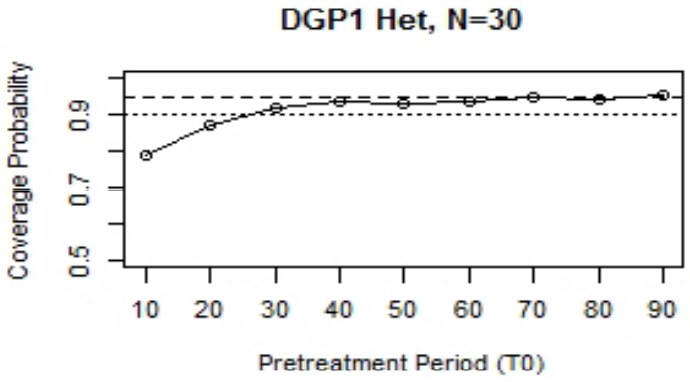
Table 5. CP for DGP 1 with Cross-sectional Heteroskedasticity

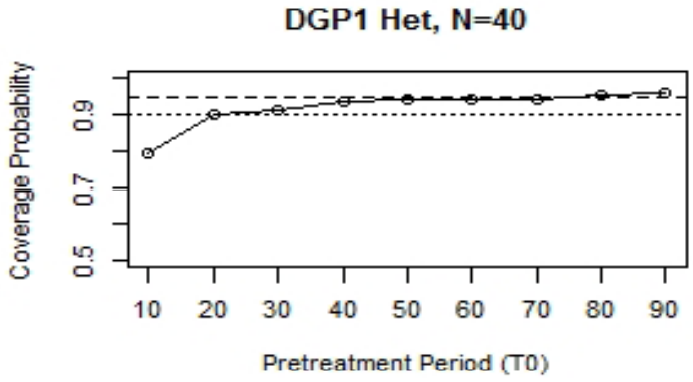
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.788	0.888	0.906	0.912	0.933	0.933	0.932	0.943	0.944
$N = 20$	0.798	0.897	0.91	0.939	0.946	0.939	0.955	0.954	0.955
$N = 30$	0.784	0.869	0.917	0.939	0.933	0.938	0.948	0.945	0.952
$N = 40$	0.792	0.903	0.914	0.937	0.942	0.944	0.944	0.956	0.96
$N = 50$	0.805	0.879	0.93	0.935	0.929	0.936	0.95	0.945	0.96
$N = 60$	0.8	0.9	0.919	0.938	0.941	0.938	0.939	0.944	0.967

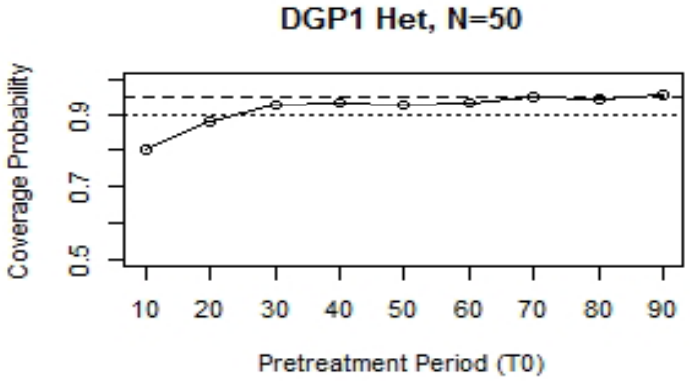
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP1 Het, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 Het, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 Het, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 Het, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 Het, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

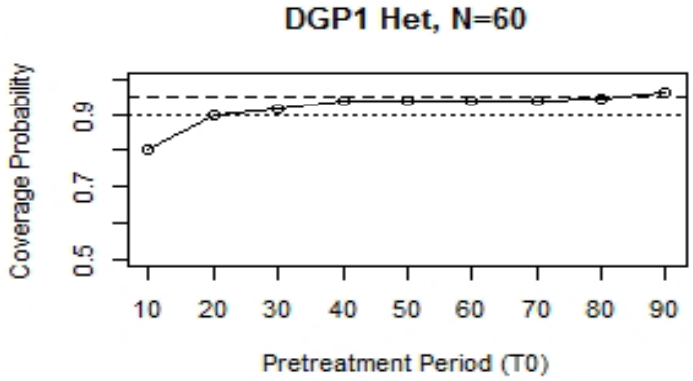
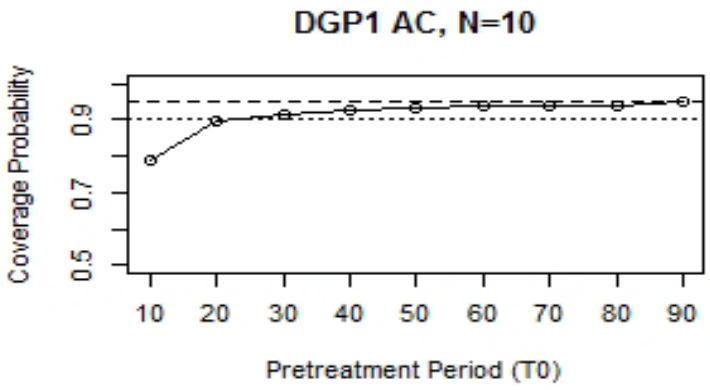
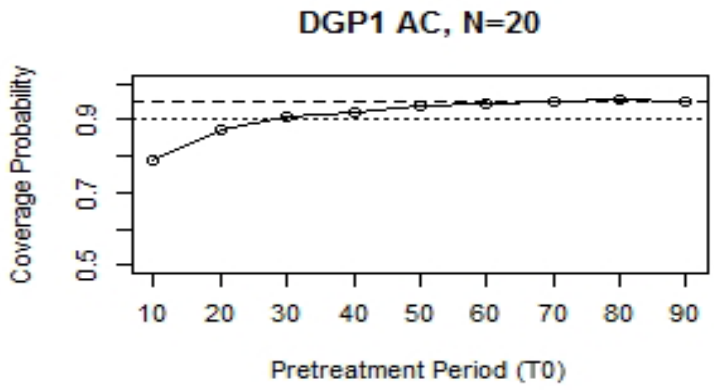

DGP1 Het, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

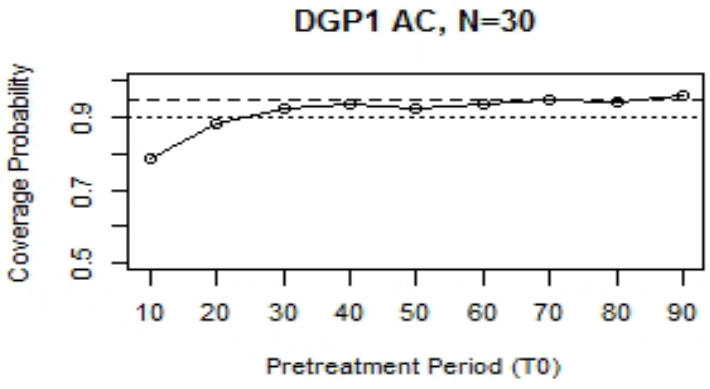
Table 6. CP for DGP 1 with Within-panel Autocorrelation

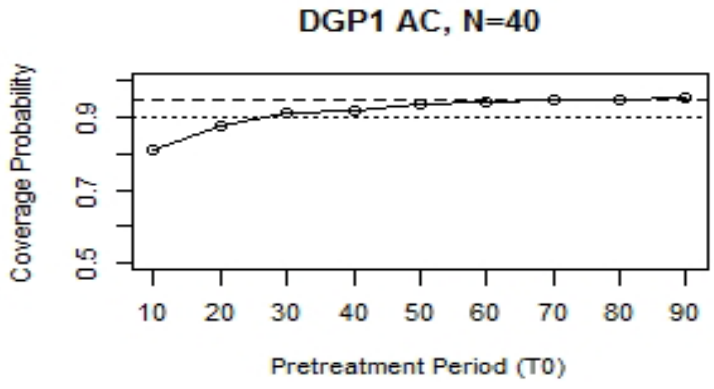
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.791	0.897	0.915	0.926	0.934	0.94	0.941	0.941	0.948
$N = 20$	0.79	0.87	0.909	0.923	0.939	0.942	0.953	0.957	0.949
$N = 30$	0.789	0.885	0.924	0.936	0.926	0.938	0.95	0.945	0.961
$N = 40$	0.808	0.879	0.915	0.921	0.938	0.945	0.95	0.947	0.957
$N = 50$	0.772	0.887	0.917	0.918	0.952	0.949	0.937	0.953	0.959
$N = 60$	0.801	0.898	0.919	0.935	0.944	0.937	0.939	0.954	0.95

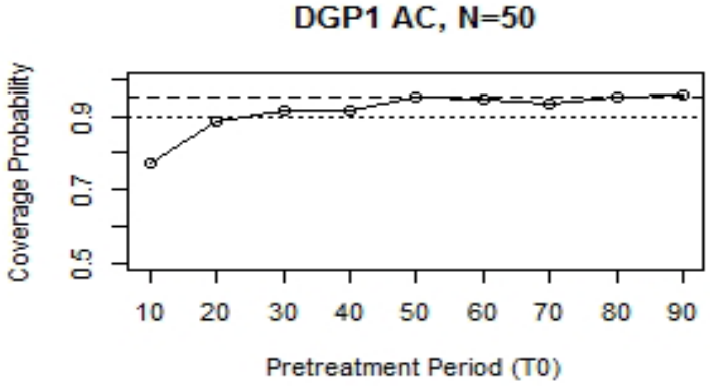
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP1 AC, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 AC, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 AC, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 AC, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 AC, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

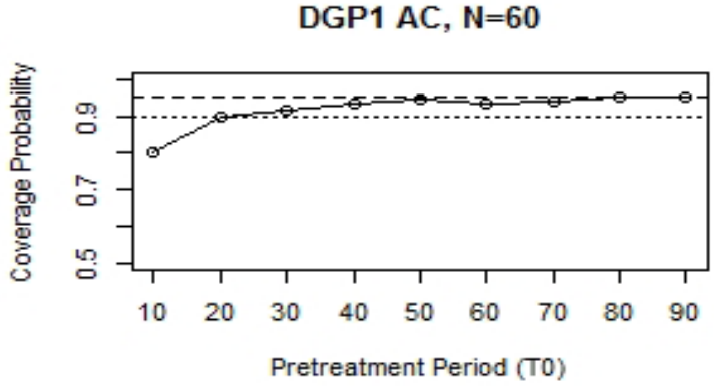
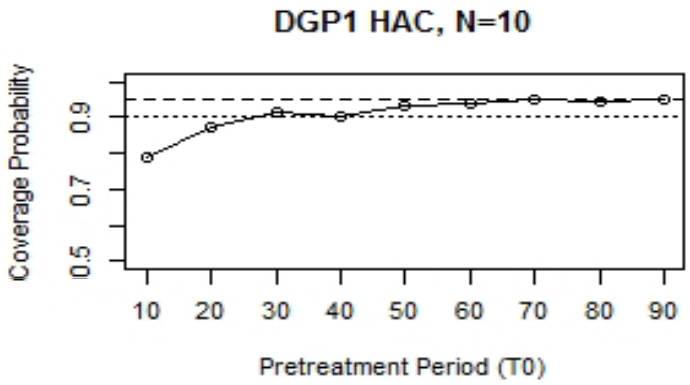
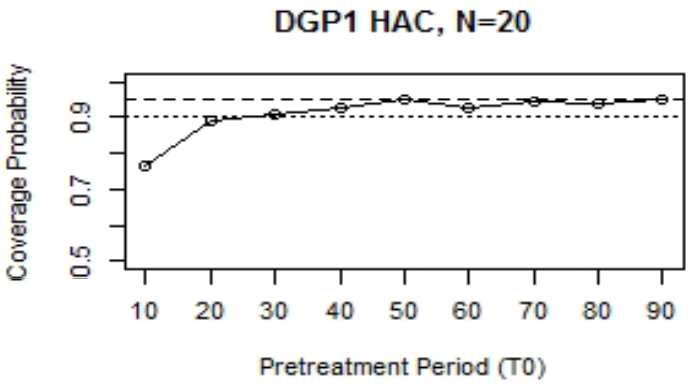

DGP1 AC, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

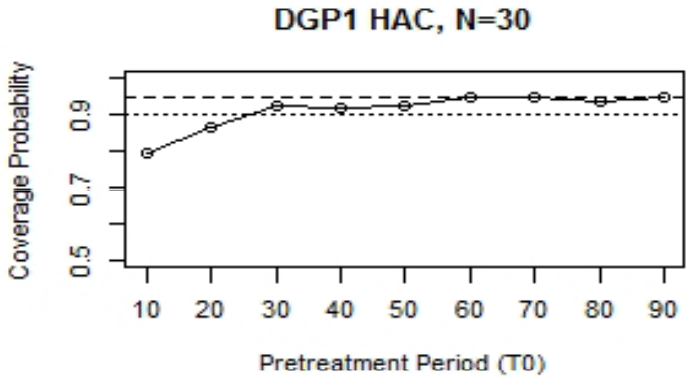
Table 7. CP for DGP 1 with Heteroskedasticity and Autocorrelation

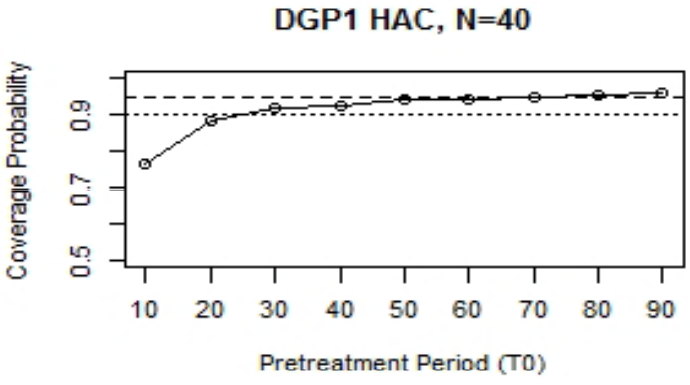
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.791	0.87	0.915	0.905	0.93	0.937	0.953	0.944	0.948
$N = 20$	0.767	0.89	0.906	0.927	0.95	0.928	0.947	0.94	0.949
$N = 30$	0.791	0.863	0.927	0.921	0.922	0.948	0.95	0.937	0.947
$N = 40$	0.765	0.883	0.92	0.926	0.945	0.943	0.947	0.952	0.958
$N = 50$	0.757	0.88	0.927	0.929	0.923	0.939	0.951	0.961	0.949
$N = 60$	0.783	0.895	0.896	0.919	0.94	0.936	0.945	0.952	0.959

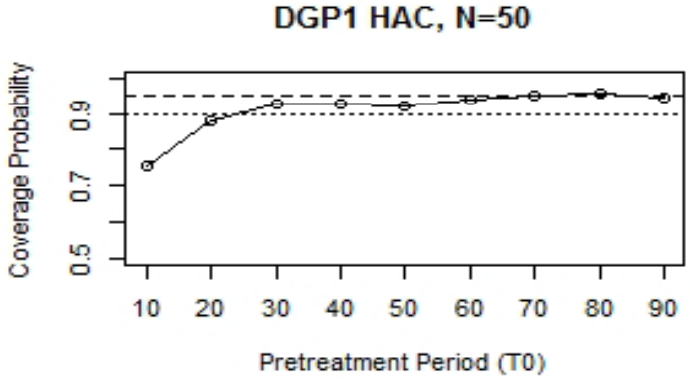
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP1 HAC, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 HAC, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 HAC, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 HAC, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP1 HAC, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

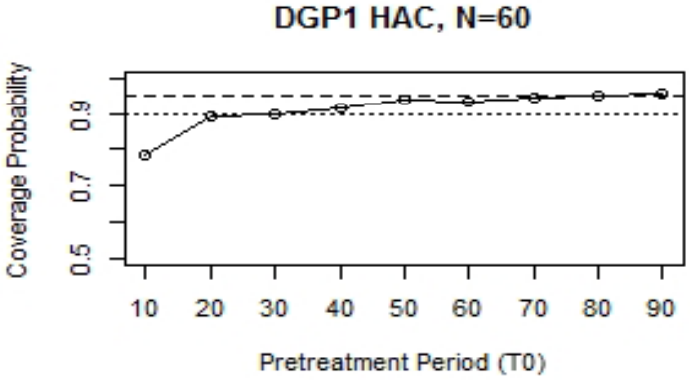
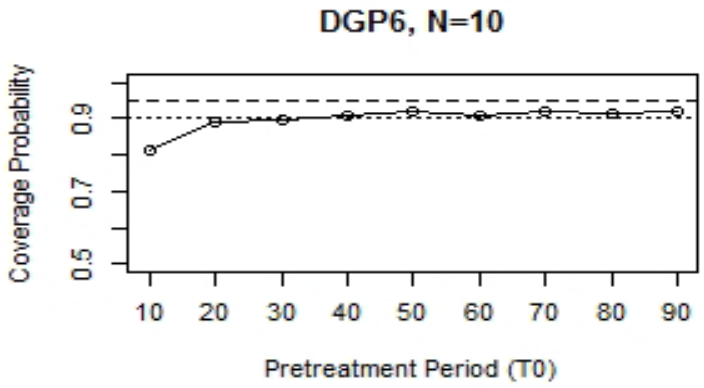
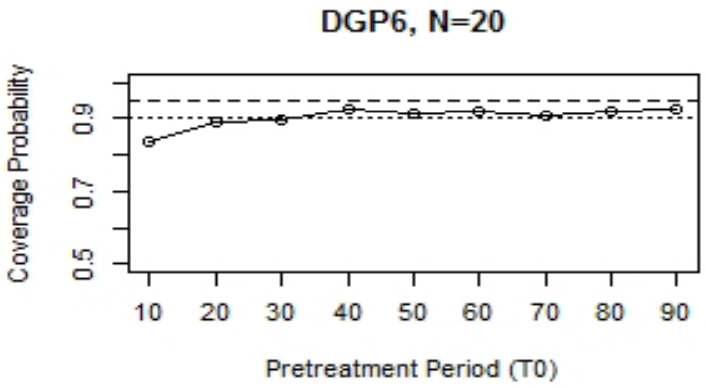

DGP1 HAC, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

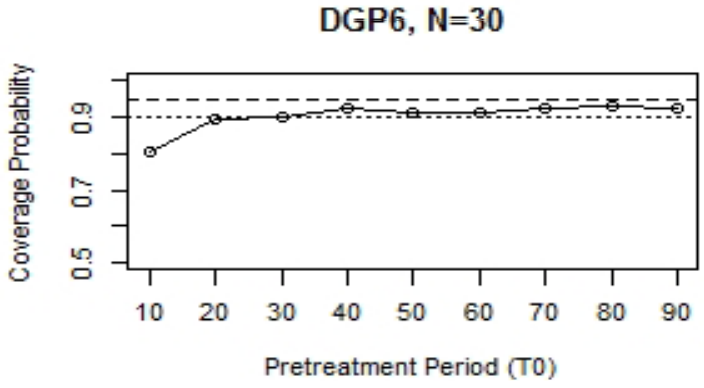
Table 8. CP for DGP 1 with a nonlinear logit transform

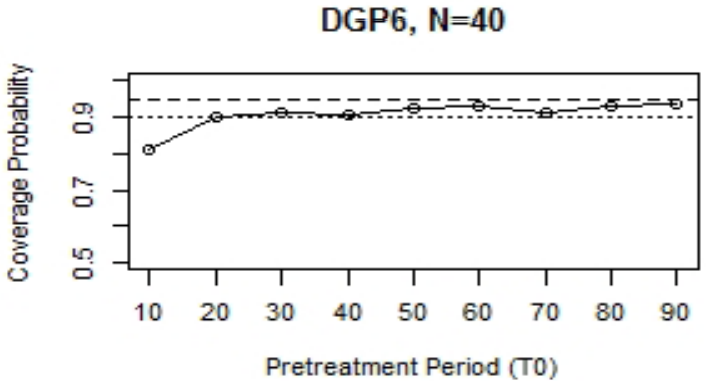
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.814	0.888	0.896	0.907	0.919	0.909	0.918	0.913	0.92
$N = 20$	0.834	0.888	0.894	0.924	0.914	0.921	0.908	0.92	0.927
$N = 30$	0.807	0.892	0.901	0.927	0.915	0.915	0.924	0.933	0.924
$N = 40$	0.812	0.903	0.913	0.909	0.922	0.932	0.915	0.928	0.938
$N = 50$	0.84	0.879	0.902	0.913	0.921	0.934	0.921	0.93	0.933
$N = 60$	0.821	0.903	0.91	0.924	0.921	0.931	0.92	0.924	0.938

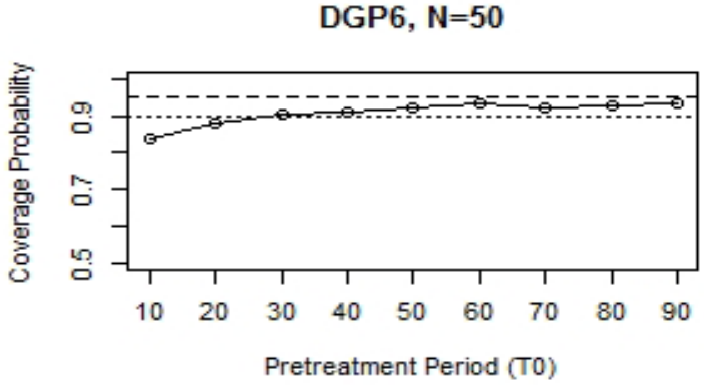
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.

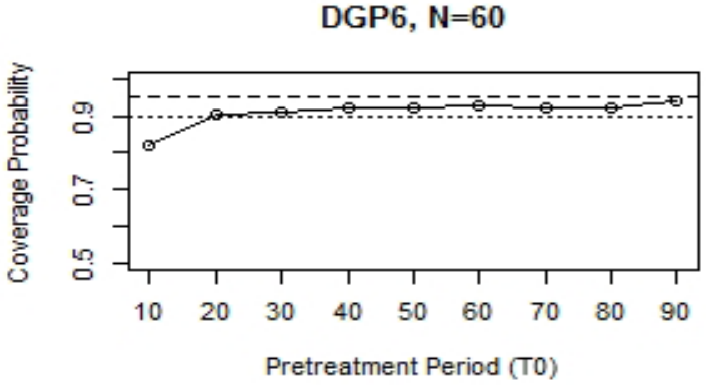

DGP6, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

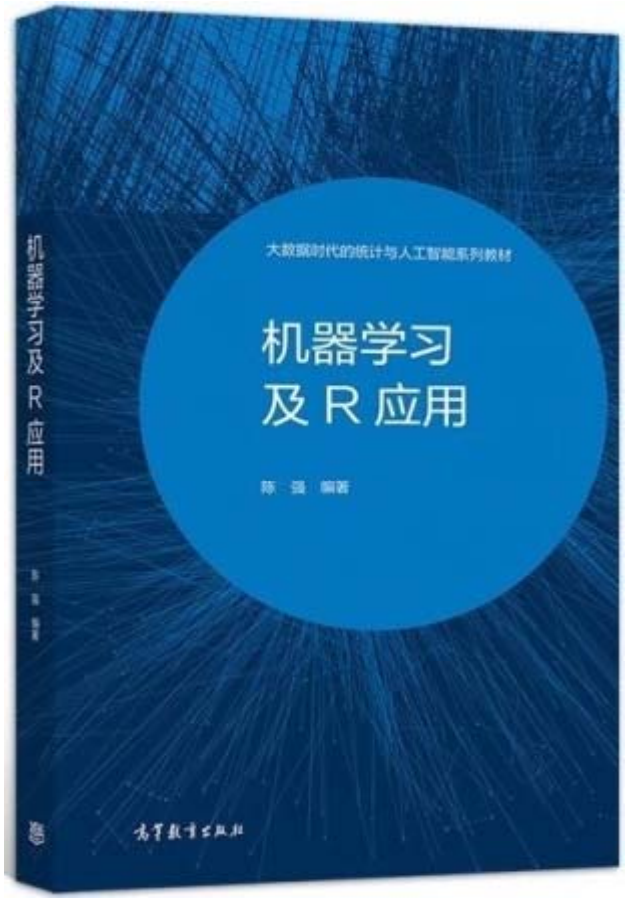

DGP6, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP6, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP6, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP6, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

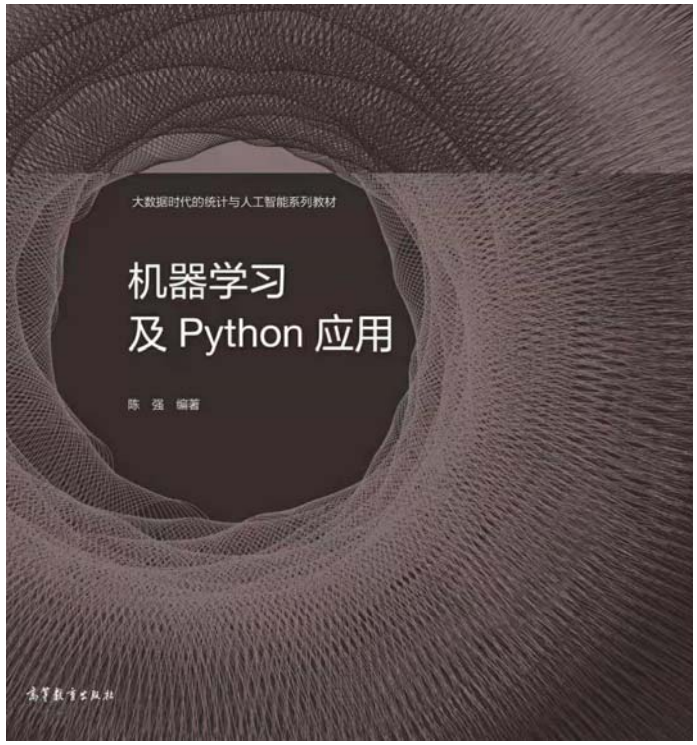

DGP6, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)



陈强，《机器学习及R应用》，高等教育出版社，2020年11月，458页，双色印刷

配套数据、R程序：
www.econometrics-stata.com

配套课程(详见网站)：
机器学习及R应用现场班 (北京，2021.1.20-24，经管之家主办)



陈强，《机器学习及Python应用》，高等教育出版社，2021年1月，即将出版

配套数据、Python程序：
www.econometrics-stata.com
（coming soon）

Thank You ☺

陈强
山东大学经济学院
www.econometrics-stata.com