突破经济增长跨国实证研究的困境

——基于机器学习方法的新视角*

刘岩†　　谢天‡

武汉大学　　上海财经大学

第一版：2019 年 3 月 15 日

最新版：2019 年 7 月 5 日

（初稿，请勿引用）

摘　要：自上世纪 90 年代开始，随着跨国经济增长数据数据库的建立与普及，跨国回归分析成为了检验经济增长理论与假设的基准实证方法。经过近 30 年的发展，跨国回归分析的方法论困境日益明显：样本有限性、模型不确定性、结论因果性等 3 个方面的挑战，对传统跨国实证分析方法的进一步发展形成了显著的制约。针对这一困境，本文提出可以充分利用新近的机器学习方法，有效缓解或克服传统分析方法存在的局限。本文系统阐述了机器学习方法的三个优势：适应小样本问题；处理变量选择问题；克服样本非线性问题。与此同时，机器学习方法的这些特性又能够有效缓解传统分析中普遍存在的内生性问题。我们在标准的跨国经济增长数据集上，对 LASSO、回归树、套袋法、随机森林、神经网络等 5 种机器学习方法的实证效果进行了检验，并与传统分析方法进行了对比，从而进一步说明机器学习方法的优势。研究结果显示，套袋法与随机森林法在跨国经济增长及类似实证问题中具有突出的适用性、有效性、稳健性。

关键词：跨国经济增长；实证方法；机器学习；模型不确定性

---

* 本研究受到国家自然科学基金（项目号：71661137003、71503191、71701175、91646206）资助。

† 武汉大学经管学院金融系；联系电话：18062100728；电子邮箱：yanliu.ems@whu.edu.cn。

‡ 上海财经大学国际商学院院；联系电话：18659256281；电子邮箱：xietian@xmu.edu.cn。

一、 引言

解释世界上不同国家与地区经济增长表现的差异，预测一个国家或地区未来的经济增长状况，寻求恰当的经济增长政策，一直以来都是经济学的核心问题。随着上世纪 80 年代末、90 年代初，覆盖全球的跨国经济增长与国民经济核算数据库的建立与完善，经济学家得以对二战后长期发展的各类经济增长理论进行有效的实证检验，并提出进一步的理论改进意见或经济增长政策建议。以跨国回归分析为核心的经济增长研究纲领，迄今已经推进了近 30 年，取得了异常丰硕的研究成果，并且目前仍然是经济增长实证研究的最主要方法[①]。相应的研究焦点，初期以 Solow 模型资本积累机制与经济增长收敛性为核心；接下来在 90 年代蓬勃发展的内生经济增长理论指引下，迅速拓展到教育与人力资本积累方面；在 2000 年前后，延伸到地理、制度、全球一体化等“深层次”决定因素中（Spolaore and Wacziarg 2013，Johnson and Papageorgiou 2019）。一系列的研究成果，极大的丰富了学术界对经济增长浅层（proximate）与深层（deep）决定因素的理解。相应结论，特别是制度等深层决定因素等的作用，在 2000 年初与“华盛顿共识”（Washington consensus）高度融和，一度被认为是完成了对经济增长过程的客观规律性认识（Rodrik 2006）。

然而，随着跨国实证研究范围与深度的快速增长，学术界也很快意识到这一研究范式实际上存在着诸多不易克服的问题与缺陷，进而极大制约了跨国实证研究的进一步发展。过去 30 年来，经济增长跨国实证研究的标准范式是跨国增长回归（cross-country growth regression）：以一国人均实际产出 y_i 为被解释变量，根据一定的理论引入一系列解释变量作为回归变量 $\boldsymbol{x}_i$（黑体表示列向量），进而建立回归方程

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad (*)$$

并可根据样本情况，将上述截面回归拓展为面板回归模型。每个经济增长理论，都对特定解释变量的系数有对应的理论判断，研究者因此可以通过对回归模型系

[①] 最新的跨国回归分析代表为 Acemoglu et al.（2019）等人在 JPE 的待刊文章，用跨国面板数据分析民主化对经济增长的促进作用。尽管使用了最新的因果推断技术，但该文的基本模型仍然是跨国线性回归。

数估计的统计检验，得出支持或不支持相应理论观点的结论。由于跨国增长回归通常呈现为一个约化（reduced form）的线性回归模型，因而此类模型所固有的问题无一例外，均会对跨国增长回归结果的意义产生影响。

在 Brock and Durlauf（2001）的综述中，总结了主流的增长回归实证研究面临的 3 方面突出挑战：

1. 理论开放性（theory open-endedness）——研究者可以想象多种多样的增长理论，选择越来越多的解释变量加入增长回归 $(*)$ 中；
2. 参数异质性——不同国家各异的增长经验，可能难以通过 $(*)$ 中一组各国同质的参数来解释；
3. 因果性——约化线性回归模型中，解释变量 $\boldsymbol{x}_i$ 普遍存在与残差项 ϵ_i 的相关性，导致系数估计的内生性问题。

上述挑战 1 的实质，在于跨国样本的有限（全球只有 200 余个国家和地区）与可能选取的解释变量庞大数量这一基本事实。Durlauf et al.（2005）在 *Handbook of Growth Economics* 的综述章节中（p. 608 及 Appendix B），总结了 43 类共计 145 个文献中使用过的回归变量。因此每一个跨国实证研究，首先要解决的就是变量选择问题。上述挑战 2 可以等价的解释为潜在增长决定因素与经济增长表现之间可能存在显著的非线性性，变量 $\boldsymbol{x}$ 在一个国家对经济增长有促进作用，在另一个国家可能具有抑制作用。大量的实证文献，也反复验证了跨国增长数据样本中存在显著的非线性性（见 Durlauf 2009 的评述）。对于挑战 3，在传统回归分析范式下，上述变量选择问题与非线性问题，更是导致跨国回归出现内生性的一个基本原因。由于样本有限，每个跨国回归模型中不可能吸纳所有的潜在解释变量[①]，但这些变量之间很容易存在相关性，因此每个特定的跨国回归总是可能出现遗漏变量问题，从而产生内生性。类似的，忽略解释变量的可能具有的非线性效应，以及解释变量之间的交叉（非线性）效应，也同样会导致约化的线性回归模型出现遗漏变量导致的内生性问题。

[①] Sala-i-Martin et al.（2004）指出，跨国回归分析中解释变量的数量很少超过 20 个，平均而言在 10 个以内。

Brock and Durlauf（2001）将上述挑战 1 与 2 并称为模型不确定性（model uncertainty）问题[①]。上述讨论表明，模型不确定性是传统的跨国回归分析范式的一个根本性制约因素。过去十余年来，学术界不乏克服这一问题的尝试，最具代表性的是 Bayes 模型平均（Bayesian model averaging，简写为 BMA）方法[②]。BMA 方法的基本思路，是给定一组潜在的回归变量，将不同变量组合定义为一个模型，然后考虑大量模型回归估计结果的平均。然而，BMA 方法的基础仍然是线性回归模型，因此并不能有效克服非线性性带来的挑战。对于非线性性，传统研究范式的主要应对方法是诉诸于非参数（包括半参数）方法[③]，但由于估计过程的复杂性，非参方法又无法有效解决大量潜在解释变量带来的变量选择问题。

如果我们跳出传统线性增长回归模型 $(*)$ 的局限，考虑更一般的增长决定因素模型，我们就可以从全新的角度来理解上述模型不确定问题给传统增长回归带来的困境，同时也便于我们引入新的、基于机器学习的研究范式来突破这一困境。具体而言，我们可以将经济增长表现 y 与其决定因素通过下述方程联系起来：

$$y = f(\boldsymbol{x}), \qquad (**)$$

其中 $\boldsymbol{x}$ 是一组研究者事前确定的、对经济增长有潜在影响的变量；函数 $f(\cdot)$ 可以具有任意的函数形式。与传统跨国回归 $(*)$ 式相比，$(**)$ 式并不要求 $\boldsymbol{x}$ 包括的变量数小于样本数，同时也可以涵盖 $\boldsymbol{x}$ 对 y 的任意非线性影响。当然，在传统计量分析框架及特定的样本性质下，我们也可以通过非参数方法来估计 $f(\cdot)$。但正如上面讨论已经指出，当样本有限而 $\boldsymbol{x}$ 又包含很多变量时，传统的非参数

① Brock 与 Durlauf 进一步将该问题的理论根源，归结到实证模型是否能够使得样本联合分布满足可交换性（exchangeability）这一重要概率性质上来。简而言之，可交换性可理解为对于一组国家增长经验适用的模型，是否对其他国家的增长经验同样具有适用性。

② 经典文献包括 Fernandez et al.（2001），Sala-i-Martin et al.（2004），以及 Durlauf et al.（2008）等。

③ 经典文献包括 Liu and Stengos（1999），Henderson et al.（2012）等。

估计存在很多技术障碍；此外，非参数模型中变量选择问题带来的模型不确定性如何处理，目前仍是尚待解决的问题（Henderson et al. 2012）。

与传统分析范式不同，我们可以将（$**$）式看做是在给定样本 $\mathbb{S}=\{y_i,\boldsymbol{x}_i\}$ 上的预测（prediction）问题：给定 $\mathbb{S}$，研究者的目标是掌握变量 $\boldsymbol{x}$ 的变动如何预测 y 的变动。实现这一目标的关键，在于充分挖掘出样本集合所包含的 $\boldsymbol{x}$ 与 y 的关联信息，包括任何可能存在的非线性关联。对于这一类预测问题，机器学习提供了一套全新的方法论[①]，可以灵活、有效的处理多变量、非线性的预测问题。在本文中，我们具体考虑了 5 个较为常用的监督学习算法：LASSO，回归树（Regression Tree，简写为 RT），套袋回归树（RT with Bagging，简写为 BG），随机森林法（Random Forest，简写为 RF），以及神经网络法（Neural Network，简写为 NN）。其中，LASSO 主要应用于线性回归形式的变量选择问题，而其他 4 种方法理论上可以深入、灵活的刻画数据样本存在的非线性特征，因此我们将分析重点放在后 4 种方法上。RT、Bagging 与 RF 三种方法都是非常主流的处理小样本数据集的算法，其中 RT 是后两种方法的基础[②]。与线性回归模型相比，RT 可以解决多变量与非线性性两个问题，但其自身存在较强的过度拟合倾向。与此不同，Bagging 和 RF 两个方法，通过更有效的样本间交叉学习，可以有效克服 RT 的过度拟合问题，让预测结果具有更强的稳健性。与这 3 种方法不同，NN 在处理小样本预测问题时，容易出现线性化过度拟合的问题，从而忽略了样本中可能存在的非线性关系。下文的跨国增长实证分析，也反映出这一特征。

为了详细说明机器学习在方法论上的对传统跨国实证分析困境的突破意义，本文使用 Sala-i-Martin et al.（2004）构建的一个标准跨国增长数据集，对上述 5 种机器学习方法的实证表现进行了系统对比，并进一步阐述了样本数据存在的非线性性对传统分析方法的制约。该截面数据集包括 12 类共 67 个经济增长的解

---

① 具体而言，我们考虑的都是监督学习（supervised learning）方法；另外一大类机器学习方法称为非监督学习，主要用途是对数据进行分类，见第三节的介绍。关于机器学习在社会科学中的应用与前景，Varian（2014）、Athey and Imbens（2017）、Mullainathan and Spiess（2017）提供了全面的综述；中文综述可见陈硕与王宣艺（2018）、黄乃静与于明哲（2018）。

② RT 在跨国增长回归问题中的应用最早可以追溯到 Durlauf and Johnson（1995），但长期以来经济增长实证研究领域的学者并未将 RT 作为机器学习方法的一种来加以考虑。

释变量[①]。通过对比这几种方法在增长变量选择、增长表现预测（解释）效能以及增长现象非线性特征刻画等方面的差异，我们力图说明机器学习方法可以有效克服传统方法在应对模型不确定性方面的困境，捕捉经济增长经验中广泛存在的非线性性，从而改进我们对经济增长经验机理的认识，有助于实现更准确的因果关系推断。

本文的贡献集中在方法论的层面。就我们掌握的研究现状而言，本文是第一篇系统探讨机器学习方法对跨国增长实证研究意义的文献。我们从机器学习——特别是本文重点考虑的套袋法与随机森林法——的理论特性出发，系统论述了这一新的方法论对于突破传统跨国回归研究范式之困境的原因与意义。在应用层面，本文也是首篇将套袋法与随机森林法引入跨国增长实证研究的文献，并突出了这些方法在处理多变量与非线性性方面的灵活性与有效性。从更宽泛的角度而言，本文意图阐述机器学习方法对解决其他社会科学领域存在的模型不确定性问题所具有的普适性意义[②]。

本文的结构安排如下：第二节中，我们对跨国增长实证研究的方法论进行简要的回顾，并对涉及机器学习的相关文献进行梳理；第三节中，我们对机器学习的一般特征，以及本文所应用的四个具体算法，进行较为细致的介绍，特别突出其对于克服模型不确定性问题的原理与直观解释；第四节中，我们汇报上述方法在 Sala-i-Martin et al.（2004）跨国增长数据集上的应用；第五节是简要的结论。

二、 文献综述

跨国增长实证领域最初均使用截面回归模型，如 Kormendi and Meguire（1985）、Barro（1991）、Barro and Sala-i-Martin（1992）、Barro and Lee（1994）。其潜在假设是模型设定的解释变量已经可以考虑各国经济增长差异的所有异质性特征，但显然这样的假设是过于严苛的。Islam（1995）将跨国回归由截面拓展到面板模型，从而可以通过固定效应来捕捉国家层面不随时间改变的

① 这一数据集仍然是目前最常用的、包括变量最多的跨国增长数据集。

② 在商业管理领域，参见 Liu and Xie（2019）、Lehrer and Xie（2019）应用机器学习技术提升电影票房预测的准确性与稳健性。

异质性因素。与此同期，Durlauf and Johnson（1995）强调了样本分组的重要性，并具体说明了广泛存在的组间异质性，从而引领一系列后续文献分析跨国增长样本存在的门限效应，包括 Hansen（2000）、Canova（2004）、Masanjala and Papageorgiou（2004）、Kourtellos et al.（2007）、Minier（2007a/b）及 Tan（2010）等文章。针对广泛存在的非线性特征，一系列文献使用了非参数（包括半参数）方法，如 Liu and Stengos（1999）、Durlauf et al.（2001）及 Henderson et al.（2012）等文章。但这类非线性方法并未讨论变量选择问题带来的模型不确定性。

几乎从跨国增长回归研究的兴起之时，研究者对通过线性回归模型所得结果的稳健性就持谨慎态度。Levine and Renelt（1992）应用极值界限（extreme bound）方法，指出当时大部分跨国回归系数均不显著。Sala-i-Martin（1997）随后开始利用 Bayes 方法评估跨国回归系数的稳健性，并最终形成了以 Fernandez et al.（2001）与 Sala-i-Martin et al.（2004）为代表的 Bayes 模型平均法；后续研究还包括 Durlauf et al.（2008）、Ley and Steel（2009/2012）。但 BMA 方法几乎都只在线性回归框架下，讨论变量选择带来的模型不确定性问题，并未涉及非线性特征与变量选择并存的模型不确定性。

在机器学习方法的应用方面，就我们所知，仅有 Varian（2014）在其述评文章中讨论了应用 LASSO 等方法进行跨国回归变量选择的可能性（pp. 16-17）。Brock and Durlauf（1995）使用了 RT 的方法，作为其分样本估计的稳健性说明；Johnson and Takeyama（2001）利用 RT 方法来说明初始条件对经济增长的影响；Tan（2010）使用了 RT 方法的一个变种来分析地理、制度等因素对经济增长俱乐部效应的影响[①]。但这些文献均为从机器学习的角度来考察 RT 方法的特征[②]，

[①] 其使用的 RT 算法变种为 GUIDE。

[②] 这些文献在使用 RT 方法时，几乎都考虑分段线性回归作为预测函数，从而继承了线性回归模型对解释变量数目的限制，无法在跨国有限样本条件下考虑大量的回归变量。唯一的例外是 Tan（2010）；该文在分段线性预测之外，考虑了分段常数预测，这与机器学习文献中对 RT 的设定类似。但该文并未深入利用这一特点，从而克服多变量问题对传统分析的局限。

并将其与更一般的模型不确定性问题相联系。此外，这些文献无一例外没有考虑 RT 方法近期在机器学习领域的重要拓展，如 Bagging 与 RF。

三、 机器学习：新的方法论

在本节中，我们首先对机器学期的一般原理与类型做一个基本的背景介绍。接下来，以保证读者对后续各个方法结果及比较的直观理解为原则，我们对本文考虑的五个具体方法做一个稍微详细些的介绍，更多的具体技术细节引用参考文献。之后，针对机器（监督）学习方法，我们引入一套较为完整的评估与处理模型不确定性问题的方法论，从而实现对具体方法所得结果的直观理解与系统评价。首先，我们解释五个具体机器学习方法下，如何衡量各个预测变量的重要性，以处理备选解释变量过多导致的模型变量选择不确定问题。其后，我们阐述如何对机器学习获得的具体预测函数 $\hat{f}$ 进行定性刻画，从而获得对样本非线性性特征的系统度量，建立对机器学习方法预测优势的直观理解。

1. 机器学习的基本逻辑

根据 Mitchell（1997）的定义，机器学习是指从数据中识别出规律并以此完成预测、分类及聚类的算法总称。从被解释变量 y 是否已知的角度来看，机器学习方法可以被分为监督学习（supervised learning）和非监督学习（unsupervised learning）；前者中 y 是已知的，而后者中是未知的。经济学数据特性决定了我们所研究的被解释变量往往是已知的，因此主要考虑监督学习方法。解释变量的常见类型包括类别型变量（categorical variable）和数值型变量（numerical variable）两种[①]。其中，数值变量在经济学中应用范围更加广阔。和类别型变量的离散性不同，数值型变量往往被假设在连续的实数域之中，因此更加适合回归类型的机器学习方法，如 Breiman et al.（1984）提出的回归决策

[①] 类型变量，如颜色、种群、国家等，在机器学习中有很广泛的应用。与之对应的典型机器学习方法包括 Breiman et al.（1984）提出的分类决策树（Classification Tree），Vapnik 等人提出的支持向量机（Support Vector Machine）等。SVM 的理论最早由 Vapnik 与 Chervonenkis 在 1963 年提出；现在广泛使用的 SVM 算法，建立在 Boser, Guyon, and Vapnik（1992）与 Cortes and Vapnik（1995）等后续工作上。

树（Regression Tree）[①]，Tibshirani（1996）所提出的 LASSO（Least Absolute Shrinkage and Selection Operator）方法等。

数值型变量的监督学习算法，可视作在给定样本 $\{y_i, \boldsymbol{x}_i\}_{i=1,\dots,N}$ 之下，对被解释变量 y 与解释变量（向量）$\boldsymbol{x}$ 间（未知）函数关系 $y = f(\boldsymbol{x})$ 寻找最佳近似函数 $\hat{y} = \hat{f}(\boldsymbol{x})$ 的问题。其中，解释变量 $\boldsymbol{x}_i$ 包括 K 个分量，可写为 $\boldsymbol{x}_i = (x_{1i}, \dots, x_{Ki})'$ 的分量形式，并且解释变量个数 K 可以接近或超过样本量 N。不同的机器学习算法使用不同的最优近似标准，亦选取不同的近似函数 $\hat{f}$ 构造方式。不同算法之间的比较，一般是通过比较近似函数所给出的预测精度，如均方预测误差，或其他风险测度。监督学习方法关注的重点，在于如何快速、有效获取近似函数，实现针对特定样本数据集的优良预测表现。正是因为其数据导向特征，机器学习方法较少受到（先验）模型设定限制，从而可以灵活应对数据本身的特性，如变量不确定性与模型非线性性等。在本文讨论的跨国经济增长实证研究问题上，机器学习方法的这些特性，正好可以用来处理传统方法所面对的模型不确定性困境。

2. 五个具体方法

本文考虑 5 个具体的机器学习算法：LASSO，回归树，套袋法，随机森林，以及神经网络。这 5 个方法也是目前数值型监督学习领域应用最广泛的方法。

2.1. LASSO

LASSO 方法与传统的线性回归模型最为接近，其近似预测函数同样是解释变量的线性函数 $\hat{y} = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$。不同之处在于，传统线性回归确定回归系数 $\widehat{\boldsymbol{\beta}}$ 是直接通过最小化残差平方和 $\min \sum_i (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ 来实现；而 LASSO 方法在上述目标函数上再添加了系数的绝对值之和：

$$\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,N} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda(|\beta_1| + \cdots + |\beta_K|),$$

[①] Breiman et al.（1984）同时提出了分类和回归决策树（Classification and Regression Tree，CART）。尽管 CART 经常被同时提及，但它其实包含了两个应用场景截然不同的决策树方法。

其中 K 为解释变量的数目[①]，$\lambda > 0$ 为权重系数[②]。LASSO 方法的最大特征，在于系数绝对值之和项倾向于减少非零回归系数的个数。因此，当备选解释变量较多，需要确定变量有效性时，LASSO 可以实现便捷的变量选择功能。但由于 LASSO 产生的近似方程仍然是解释变量的线性形式，因此当数据样本可能存在显著的非线性特征时，LASSO 的预测表现会收到很大限制。在我们下文的实证分析中，我们将主要侧重于说明 LASSO 的变量选择功能。

2.2. 回归树

Breiman et al.（1984）首先提出了分类和回归决策树（Classification and Regression Tree，CART）的概念。决策树的结构类似于流程图，其中每个内部结点表示对数据属性的测试与衡量，每个分支（或叶子）代表衡量的结果，并且每个叶结点代表类标签（即在考虑了所有属性特征之后做出的决定）。从 root（起始）到 leaf（末端）的路径则代表了分类规则。在本文的应用中，我们专注于使用回归树，因为我们实证应用使用的都是实数数据而非分类数据。

回归树递归地将数据划分为不同的本地数组，并将每个组的平均值作为最终预测值。在回归树中，结点 τ 包含 n_τ 个观测值（视作观测值指标 i 的集合）。每个结点只能分为两个叶子，由 τ_L 和 τ_R 分别代表。每片叶子分别包含 n_L 和 n_R 个观测值，并且 $n_L + n_R = n_\tau$。我们将结点内的平方和定义为

$$\mathrm{SSR}(\tau) = \sum_{i=1}^{n_\tau} (y_i - \bar{y}_\tau)^2,$$

其中 $\bar{y}_\tau$ 为平均值。我们将结点 τ 的 n_τ 个观察值拆分为 τ_L 和 τ_R 两个部分，拆分点的选择遵循最大化以下残差平方和（SSR）差值的原则[③]

---

① 与 LASSO 紧密联系的还有岭回归（ridge regression），其差别在于目标函数添加的不是系数绝对值之和，而是平方和。岭回归主要性质在于克服线性回归系数估计的不稳定性，但不具有 LASSO 方法的变量选取特性。

② 通过将样本随机划分为学习组与校验组并，进行多次交叉校验（cross validation），可以得到 LASSO 惩罚项系数的最优值 λ^*。下文中所分析的 LASSO 预测函数，就是基于 λ^* 得到的。

③ 具体而言，拆分点的选择是按照如下步骤。首先，对每一个变量 x_k，将结点 τ 所包括的样本取值 $X(\tau,k) \equiv \{x_{ki}\}_{i\in\tau}$ 从小到大排序，然后依次考虑一系列样本分组 $\tau_S^j(k), \tau_B^j(k), j = 1, \dots, |\tau|$，其中 $\tau_S^j(k)$ 为 $X(\tau,k)$ 中前 j 小取值 x_{ki} 所对应的指标集合，$\tau_B^j(k) = \tau \setminus \tau_S^j(k)$。其次，对所有分组，计算 $\Delta^j(\tau,k) =$

$$\Delta = \mathrm{SSR}(\tau) - \mathrm{SSR}(\tau_L) - \mathrm{SSR}(\tau_R), \tag{1}$$

这也可以理解为最小化拆分后两结点 SSR 之和。结点 τ_L 或者 τ_R 可以视为新的结点并继续拆分过程。我们从树的顶部开始（完整样本）并将相同的方法应用于所有后续结点。

一个完成的回归树会包含很多叶子，而每个叶子包含样本数据的子集，并且所有叶子的合集是完整样本。也就是说，与逐步回归类似，回归树每一次拆分类似于模型变量的选择。回归树在每个子集内继续拆分，直到触发某些预先设置好的停止规则。而这很可能导致过度拟合。因此，在实践中我们往往会使用某些标准（或信息准则等）去修剪回归树。这些标准往往会在拟合优度以及模型复杂性（结点数量、叶子数量等）之间寻求一个均衡，以降低过度拟合带来的危害。

使用回归树做预测主要依赖于每个叶子中因变量子集的平均值。这些平均值也构成了回归树的整体拟合值。Hastie et al.（2009）提供的证据表明，在实践中，来自回归树的预测具有较低的偏差但是方差往往较大，原因在于回归树的不稳定性，即观测数据中的较小的变化可能导致差异很大的分割序列，并因此导致迥然不同预测值。文献中广泛使用了两种方法纠正这个问题：套袋回归树法（RT with Bagging）和随机森林法（Random Forest）。

2.3. 套袋法

Breiman（1996）提出的用 bootstrap 聚合（bootstrap aggregation，简称为 Bagging）方法，通过构造随机生成的训练集来达到改进 CART 等方法预测精度和稳健性的目的。给定一个原始数据集 $\{y_i, \boldsymbol{x}_i\}$，套袋法首先生成 B 个新的训练数据集 $\left\{y_i^b, \boldsymbol{x}_i^b\right\}_{b=1,\ldots,B}$，其中每一组新训练集都包含有 N 个样本，且每一个样本都是原始数据的“有放回”采样。因此某些观测值会有重复，大样本下的 $\left\{y_i^b, \boldsymbol{x}_i^b\right\}$ 预计会有 $(1-1/e) \approx 63.2\%$ 的独特样本。我们将回归树法应用到每个训练数据集，并构建一个（不剪枝）回归树。在套袋法的预测中，我们首先使用每棵回归树计算出一个预测值，最终预测值是所有 B 个预测值的简单平均值。

$\mathrm{SSR}(\tau) - \mathrm{SSR}\left(\tau_S^j(k)\right) - \mathrm{SSR}\left(\tau_B^j(k)\right)$。最后，找到实现 $\max_{j,k} \Delta^j(\tau, k)$ 的分组 $\tau_S^{j^*}(k^*), \tau_B^{j^*}(k^*)$，进而将前者记为 τ_L，后者记为 τ_R，相应的 $X(\tau, k^*)$ 中第 j 小的取值，记为该结点的拆分点。

许多研究发现，结合了数百或数千棵树的套袋法相比只使用单一棵树的回归树法，在预测表现方面有大幅改进。

对套袋法来说，其包含的每棵树都是相同分布的。特别是当 $\boldsymbol{x}$ 中包含有解释力极强的变量，所有的树都将重点使用这些强变量，进而导致所有的树都很相似，并且有较大的相关性。这会导致所有树的偏差与单个树木的偏差相差无几，然而相关树的数量 B 增加会导致方差下降。

2.4. 随机森林法

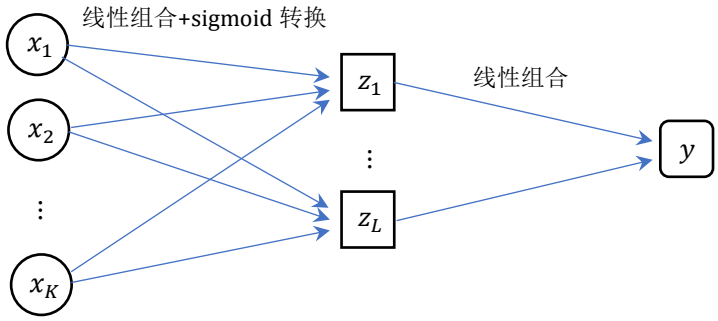
为了进一步减弱套袋法中，各树之间的相关性，Breiman（2001）进一步提出了随机森林法。随机森林法类似于套袋法，因为两者都依靠对原始数据使用 bootstrap 方法，进而构建 B 个随机训练集。但对于随机森林，在构建每棵树时，每一个结点上的拆分，我们会从总共 K 个预测变量中随机抽取 q 个预测变量（不放回采样，$q < K$）。其中，q 的默认值为 $K/3$；如果 $q = K$，则等同于套袋法。依照此法，我们最终生成了 B 棵回归树，而最终的预测值仍然是每棵树预测值的简单平均值。

研究发现随机森林在预测变量数较大时，预测表现提升很明显。如果预测变量中存在许多无关变量，则有相关变量参与分割的概率会变得较低。同时，通过随机选择预测变量，随机森林生成的 B 棵树的相关性要低于套袋法。

2.5. 神经网络法

神经网络法包括异常多的变种，并且是近年来大放异彩的人工智能深度学习（deep learning）方法的基础。由于我们需要处理的预测问题较为简单，因此我们也仅考虑最基本的单层神经网络方法，以及通用的转换函数设定[①]；算法示意图如下：

[①] 神经网络方法更详尽的介绍，见 Hastie et al.（2009）。我们使用的默认通用设置，来自于 MATLAB 自带函数 feedforwardnet；其中，中间层结点数设为 10 个。



其工作流程如下：首先，根据确定好的中间神经元层结点数目 L，将输入的 K 个预测变量组合为 L 个不同的线性组合 $\boldsymbol{\theta}_\ell^\top \boldsymbol{x}$，$\ell = 1, \ldots, L$；接下来通过 sigmoid 函数将这些线性组合变量转换为中间变量 $z_\ell = \sigma(\boldsymbol{\theta}_\ell^\top \boldsymbol{x})$[①]；最后，再由 z_ℓ 的线性组合得到最终的预测值 $y = \boldsymbol{\phi}^\top \boldsymbol{z}$。算法会根据样本 $\{y_i, \boldsymbol{x}_i\}$ 自动学习并确定相关参数，从而得到预测函数。

得益于以中间层神经元为数据转换中介，神经网络法在构造预测函数上具有极大的灵活性[②]；并可以通过增加神经元层级数，使其具有网络结构进而实现前后神经元之间的反馈特征，进一步增强其预测适应能力。但这一特征也容易造成神经网络的过度拟合问题，降低其预测稳健性。此外，由于神经元结点同时考虑所有预测变量的作用，造成神经网络方法本身不具有一个自然的变量选取结构。事实上，当数据本身具有较强的空间或时间维度相关性时，神经元起到的作用是提取原始数据在局部区域的总体特征[③]。但对我们要处理的跨国增长问题而言，数据本身没有特定的排序结构，因此限制了神经网络方法的突出优势。

2.6. 方法小结

总结上述介绍，我们可以明显看出，LASSO 方法主要着眼于变量选择，对变量间的非线性关系不予考虑，也并不着眼最大程度提高预测准确性。与此不同，回归树方法着眼于获得尽可能好的样本内预测准确性，同时在预测树的构造过程中，也连带考虑了变量选择问题。回归树的过度拟合倾向，是套袋法与随机森林法所希望解决的问题。因此，后两种方法在继承回归树方法优异的非线性预测性

① 其具体函数形式为 $\sigma(v) = 1/(1 + e^{-v})$。

② 从中间层神经元 $\{z_\ell\}$ 到最终预测值 y 同样可以考虑非线性映射，进一步增强其预测函数灵活性。

③ 典型的例子如人脸图像识别中神经元起到了局部抽象概括（五官）的作用，或在语音识别中起到词组识别的作用。

质的同时，也能够有效解决近似函数的不稳定性，从而实现预测的稳健性。与之对比，近年来流行的神经网络方法在具有极强的预测灵活性特征下，容易出现过度拟合与变量选择能力较弱的问题。此外，与前 4 种方法相比，神经网络法更适合具有固定排序及相关性结构的数据。虽然跨国增长的数据本身没有这样的结构，但由于神经网络方法近年来是机器学习方法的一个典型代表，我们仍然将其纳入本文的考察范围。

3. 预测变量重要性的衡量与筛选

LASSO 方法在衡量变量重要性时，通常采用以下标准（如 Varian 2014）。随着权重系数 λ 的增大，LASSO 在计算系数向量 $\boldsymbol{\beta}$ 的取值时，会将越来越多的分量设为 0。这提供了一种衡量变量重要次序的方法，即将系数最先达到 0 的变量，认定为最不重要，而系数越晚变为 0 的变量，其重要性越高。因此，LASSO 对变量重要性的衡量是一种排序，而不是一个连续取值的标准。这与其他方法具有较大差别。

对于回归树、套袋法、随机森林三个方法，由于其预测函数的构建均是基于决策树，因此各分支结点对应的预测变量自然的具有一个重要性排序，越重要的变量，越靠近树的根部。由于一棵决策树中，同一个变量可能出现在多个结点，以及套袋法和随机森林法中同一个变量在各（随机）样本下所生成决策树中对应的位置可能不同，因此我们仍然需要一个量化的指标来衡量各变量的重要性。为此，我们首先在单棵树上，计算出每个预测变量的重要性得分，然后在所有树（若适用）上对分数进行平均，以获得变量的全局重要性分数。理念上来讲，最重要的变量即是对（预测）精度变化影响最大的变量。

对于回归树，我们计算每一个变量在所有拆分结点上 Δ 值（1 式）的总和并除以结点的数量，以此来衡量变量的重要性（Breiman et al. 1984）。更高的值意味着所对应的预测因子更加重要。

对于套袋和随机森林，每棵树都由各自随机抽取的样本生成，并且每一个自助法样本中都会有部分观测数据被排除。这部分被排除的样本（Out-of-Bag，简写为 OOB）可用于对回归树进行评估，并且由于没有参加树的构造，不存在过度

拟合的风险。这一特征也为我们提供了相比回归树而言更好的变量重要性测度基础（Breiman 1996、2001）。具体而言，我们首先计算第 b 棵树产生的预测函数在其 OOB 样本中的预测误差。然后，对我们所关心的变量 x_k，在 OOB 样本中随机置换该变量观测值的顺序，并根据新生成的样本计算出新的预测误差。我们比较两次预测误差的差值，定义为 Δ_k^b。对于每一个变量 $k=1,\dots,K$，我们计算出所有树 $b=1,\dots,B$ 的平均值 $\overline{\Delta}_k$，并以 $\overline{\Delta}_k$ 的由大到小的排列顺序决定 K 个变量的重要性。这一测度方法的理念是，如果预测变量 k 更加重要，那么通过置换其观测值顺序，这个预测变量会因失去正确的样本对应关系（特别是 x_{ki} 与 y_i 之间的对应关系），而带来预测能力更大的下降，进而带来更高的 $\overline{\Delta}_k$。

如上一小节所述，神经网络法并不具有一个自然的变量选择结构，所有预测变量通过中间层神经元都进入到最终预测目标中。因此，该方法下的变量选择标准具有一定的任意性。注意到我们考虑的单层神经网络方法中，预测变量到各个神经元需经过一次线性组合（忽略转换函数 σ），而从各个神经元到最终预测值，又需要经过一次线性组合，因此一个简便的变量重要性测度方法，就是将一个变量在两次线性变换中所涉及的系数（标准化后）相乘并求和。具体而言，我们使用如下公式计算一个变量的重要性：$w_k=\sum_{\ell=1,\dots,L}\varpi_{k\ell}\cdot\varpi_\ell$，其中$\varpi_\ell=|\phi_\ell|/\sum_m|\phi_m|$ 为 $\boldsymbol{\phi}$ 中各系数标准化得到，而 $\varpi_{k\ell}=|\theta_{k\ell}|/\sum_m|\theta_{m\ell}|$ 为 $\boldsymbol{\theta}_\ell$ 中各系数标准化得到。

一旦我们可以对预测变量的重要性进行排序，自然也就可以在需要的情况下进行变量选择。下一节跨国经济增长的实证分析中，我们将报告应用上述变量选择方法所得的前十大的预测变量，确实表现出稳健、高效的增长预测能力。

4. 预测函数关键特征的度量

上一小节我们系统描述了如何对 5 种方法进行变量重要性测算与排序。注意，这些测算方式并没有着眼于预测变量的样本内解释力。换言之，我们并没有用各个变量对样本内目标变量变动（如方差）的解释份额，来做为变量重要性测度及选择的依据。这样操作的理由很简单，我们希望变量选择具有良好的稳健性，因此要避免以样本内解释力度为出发点的过度拟合问题。

不过，一旦我们得到了一个具体的预测函数 $\hat{f}$，我们就希望从直观角度尽可能全面、充分的理解这个预测函数的特性，从而对该机器学习方法的特性进行更全面的了解。考虑到 $\hat{f}$ 一般而言是一个高维、非线性函数，我们需要采取一定的分析学思想，对各个预测变量通过 $\hat{f}$ 所反映出的预测效能进行一个分解、测算，从而更好的掌握 $\hat{f}$ 本身的特性。我们将这一笼统的目标，具体分拆为 3 个方面：i. 单个变量的样本内解释效力的测算；ii. 多个变量组合的联合解释效力测算；iii. 解释效力中线性趋势与非线性特征贡献的分解。由于这部分内容技术性较强，我们在正文中直接给出相关测算公式，将更详细的推导与讨论放到附录 A 中展开说明[①]。

对单个变量解释效力的测算，我们采取一个“减法”思维，从样本预测值 $\hat{f}(\boldsymbol{x}_i)$ 中减去变量 x_{ki} 的作用，再求和度量总的预测效能损失程度，从而获得变量 x_k 的解释效力。具体而言，我们使用如下公式测算单个变量解释力度的水平值（average explained prediction variation）：

$$AEPV(\hat{f};k)=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_{1i},\dots,x_{k-1,i},x_{kj},x_{k+1,i},\dots,x_{Ki})\right)^2. \quad (1)$$

其中 $\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_{1i},\dots,x_{k-1,i},x_{kj},x_{k+1,i},\dots,x_{Ki})$ 剔除了 x_k 在样本点 i 处对预测值 $\hat{f}(\boldsymbol{x}_i)$ 的贡献，故两者之差反映了 x_k 在该点样本值对预测的贡献。进一步定义所有变量联合的解释力度（average prediction variation）：

$$APV(\hat{f})=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_{j=1}^{N}\hat{f}(\boldsymbol{x}_j)\right)^2, \quad (2)$$

则可计算单个变量的百分比解释力度（fraction of EPV）：

$$FEPV(\hat{f};k)=\frac{AEPV(\hat{f};k)}{APV(\hat{f})}. \quad (3)$$

① 附录 A.4，从线性预测函数出发，对下面讨论的各个预测解释力度测算指标进行了直观说明。

在此基础上，我们进一步定义两个变量的联合解释力度：

$$AEPV(\hat{f};k,m)=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_{j=1}^{N}\hat{f}(\dots,x_{kj},\dots,x_{mj},\dots)\right)^2, \tag{4}$$

并通过减去 x_k,x_m 两个变量各自单独的解释力度，得到其交互作用解释力度（interactive EPV）：

$$IEPV(\hat{f};k,m)=AEPV(\hat{f};k,m)-AEPV(\hat{f};k)-AEPV(\hat{f};m). \tag{5}$$

与单变量类似，我们可以计算交互作用的百分比解释力度：

$$FIEPV(\hat{f},k)=\frac{IEPV(\hat{f};k,m)}{APV(\hat{f})}. \tag{6}$$

上式反映了两个变量的交互效应，对预测值的贡献比例。

最后，我们将 $\Delta\hat{y}_{[k]i}\equiv\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_j\hat{f}(\dots,x_{kj},\dots)$ 视作变量样本值 x_{ki} 对预测值 $\hat{y}_i=\hat{f}(\boldsymbol{x}_i)$ 的贡献，并通过下述辅助回归

$$\Delta\hat{y}_{[k]i}=\alpha+\beta_k x_{ki}+\xi_i,\qquad i=1,\dots,N, \tag{7}$$

来刻画变量 x_k 对预测值的贡献中，线性趋势项与非线性项各自的贡献。上述回归的 R^2 捕捉了线性趋势的解释贡献，对应的 $1-R^2$ 则测算了变量非线性性特征的解释贡献。

四、 实证分析

为了具体说明机器学习方法在经济增长跨国实证分析中的优势，我们使用跨国增长研究领域的标准数据，系统对比上述 5 种广泛使用的数值型监督学习算法及传统分析方法的实证表现，以此说明机器学习方法可以有效突破传统方法范式所面对的困境。

1. 数据样本

我们的数据样本是 Sala-i-Martin et al.（2004）AER 文章所采用的样本[①]，包括 88 个国家的人均实际 GDP 增速（被解释变量）与 67 个潜在解释变量。这一数据集也是后续 BMA 方法文献所使用的标准数据集，如 Ley and Steel（2009）等。表 1 包括了所有变量的代码及释义，详细说明及来源请见 Sala-i-Martin et al.（2004）。注意，表中第一个变量 GR6096 是被解释变量。

表 1：样本变量

变量代码	释义	分组	变量代码	释义	分组
GR6096	1960-96 年人均 GDP 增速	Na	LIFE060	1960 年预期寿命	2
ABSLATIT	纬度绝对值	4	LT100CR	通航水域土地占比	7
AIRDIST	距大城市的空中距离	7	MALFAL66	60 年代疟疾虚拟变量	4
AVELF	民族语言分化	Na	MINING	采矿业的 GDP 比例	11
BRIT	英国殖民地虚拟变量	10	MUSLIM00	穆斯林比例	6
BUDDHA	佛教徒比例	6	NEWSTATE	独立时间	10
CATH00	天主教徒比例	6	OIL	石油生产国虚拟变量	11
CIV72	公民自由	8	OPENDEC1	1965-74 开放度	3
COLONY	殖民地虚拟变量	10	ORTH00	东正教比例	6
CONFUC	儒教比例	6	OTHFRAC	外语人口比例	na
DENS60	人口密度 1960	7	P60	1960 小学教育	1
DENS65C	60 年代沿海人口密度	7	PI6090	1960-90 通胀率	3
DENS65I	60 年代内陆人口密度	7	SQPI6090	1960-90 通胀率平方	3
DPOP6090	1960-90 人口增长率	1	PRIGHTS	政治权利	8
EAST	东亚虚拟变量	5	POP1560	15 岁以下人口比例	2
ECORG	资本主义	8	POP60	1960 年人口	7
ENGFRAC	英语人口	na	POP6560	65 岁以上人口比例	2
EUROPE	欧洲虚拟变量	5	PRIEXP70	1970 年初级产品出口	11
FERTLDC1	60 年代的生育率	2	PROT00	新教徒比例	6
GDE1	国防开支	3	RERD	实际汇率扭曲	3
GDPCH60L	60 年人均 GDP (log)	1	REVCOUP	革命和政变	9
GEEREC1	60 年代公共教育支出占比	1	SAFRICA	非洲虚拟变量	5
GGCFD3	公共投资份额	3	SCOUT	经济外向性	3
GOVNOM1	60 年代政府名义支出占比	3	SIZE60	经济规模	7
GOVSH61	60 年代政府实际支出占比	3	SOCIALIST	社会主义虚拟变量	8
GVR61	60 年代政府消费占比	3	SPAIN	西班牙殖民地虚拟变量	10
H60	1960 高等教育水平	1	TOT1DEC1	60 年代的贸易增长	12
HERF00	信教程度	6	TOTIND	贸易条件排名	12
HINDU00	印度教徒比例	6	TROPICAR	热带地区面积占比	4
IPRICE1	投资品价格	3	TROPPOP	热带地区人口占比	4
LAAM	拉丁美洲虚拟变量	5	WARTIME	1960-90 战争花费比例	9
LANDAREA	土地面积	7	WARTORN	1960-90 战争虚拟变量	9

[①] AER 官网页面：http://www.aeaweb.org/articles?id=10.1257/0002828042002570。

LANDLOCK	内陆国家虚拟变量	7	YRSOPEN	1950-94 年开放年数	3
LHCPC	油气储量 1993	11	ZTROPICS	热带气候虚拟变量	4

注：数据来源为 Sala-i-Martin et al.（2004）；分组代码见表 2，部分变量无分组，记为 na

同时，为了能够更充分的讨论后续的实证结果及其经济意义，我们借鉴 Ciccone and Jarocinski（2010）的分类方法，将上述 67 个解释变量，依据其理论归属，分为 12 组。具体分组及编号见表 3[①]。

表 2：变量分组及编号

编号	分组名称	编号	分组名称	编号	分组名称
1	新古典增长理论	5	地区异质性	9	战争及冲突
2	人口结构	6	宗教	10	殖民地历史
3	宏观政策	7	区位及贸易	11	自然资源
4	自然地理特征	8	制度	12	贸易条件

2. 实证结果

2.1. 变量选择

我们首先汇报 5 种机器学习方法的变量选择结果，结果见表 3。为简洁起见，我们只列出了按照第三节 2.2 中介绍的重要性测度排序前 10 位的变量。注意，不同的方法其重要性衡量标准是不同的。作为对比，我们同时列出了 Sala-i-Martin et al.（2004）使用 BMA 方法得到的前 10 位显著解释变量。BMA 方法中，变量显著性的通用衡量指标是后验纳入概率（posterior inclusion probability）；Sala-i-Martin et al.（2004）一共发现 18 个变量具有显著性[②]。

表 3：变量选取对比

	LASSO	**回归树**	**套袋法**	**随机森林**	**神经网络**	**BMA**
1	<u>YRSOPEN</u>	**MALFAL66**	**MALFAL66**	**MALFAL66**	SPAIN	**EAST**
2	TROPPOP	<u>BUDDHA</u>	**EAST**	**EAST**	GOVNOM1	**P60**
3	**P60**	ABSLATIT	<u>BUDDHA</u>	<u>BUDDHA</u>	LANDAREA	**IPRICE1**
4	**EAST**	LANDAREA	**P60**	<u>YRSOPEN</u>	MUSLIM00	**GDPCH60L**
5	**CONFUC**	GDE1	ABSLATIT	**LIFE060**	SOCIALIST	**TROPICAR**

① Ciccone and Jarocinski（2010）的“理论”变量分组中，忽略了 3 个变量：民族语言分划（AVELF），英语人口（ENGFRAC），外语人口比例（OTHFRAC）。由于这 3 个变量在后续实证分析中并没有表现出很强的预测解释力，因此我们也没有对其进行补充分组。

② Sala-i-Martin et al.（2004）将变量显著性定义为后验纳入概率超过先验纳入概率，亦即数据样本的信息提高了该变量的纳入概率。其文章中具体选用的先验纳入概率为 10.4%。故这 18 个显著变量的后验纳入概率均大于等于 10.4%。

6	**MALFAL66**	GVR61	**LIFE060**	TROPICAR	OPENDEC1	**DENS65C**
7	RERD	**IPRICE1**	YRSOPEN	ABSLATIT	**EAST**	**MALFAL66**
8	**IPRICE1**	AIRDIST	AIRDIST	**CONFUC**	TOT1DEC1	**LIFE060**
9	BUDDHA	SIZE60	GGCFD3	**P60**	**IPRICE1**	**CONFUC**
10	GVR61	POP6560	**DENS65C**	FERTLDC1	REVCOUP	**SAFRICA**

注：代码涵义见表 1；加黑表示该变量出现在 BMA 选取的前十显著变量；下划线表示该变量是 BMA 选取的非前十但显著的变量

为了说明机器学习方法变量选择的效能，我们在表 3 中，将标准的 BMA 方法变量选取结果作为对比基础。第 2 到 6 列中，我们将机器学习方法选择出来的前十个重要变量中，与 BMA 前十变量重复的变量均加黑标识；此外，如果该变量是 BMA 方法选取的前十之外的显著变量，我们以下划线标识。在对比结果之前，我们再次指出，BMA 方法从设计上能够有效处理变量选择带来的模型不确定性，识别出不同模型设定下反复出现、具有显著预测效应的变量，但其基础是线性模型，无法应对样本的非线性特征。

从该表结果可见，回归树选取的前十变量中，只有 4 个是 BMA 显著变量。这反映了回归树方法过度拟合数据的倾向，使其在有限样本下容易关注预测效果并不特别稳健的变量。套袋法在设计时有意增强模型预测效能的稳健性，表 3 的结果也显示出这一点，一共识别出 7 个 BMA 显著变量，且其中有 5 个都是 BMA 前十显著的。随机森林与套袋法接近，理论上在数据呈现高截面相关性时有更强的稳健性，使其选取的解释变量能在更多的情况下显示出良好的预测能力；但当截面相关性有限时，预测效能未必一定超过套袋法。表 3 的结果清晰的显示了这一点，选取的前十变量中有 7 个都是 BMA 显著变量，且 5 个是 BMA 前十显著变量。表 3 中 LASSO 的结果同样在意料之中。与其他 3 个机器学习方法相比，LASSO 是为一个以线性预测函数为基础的，最接近 BMA 的基础设定；同时系数绝对值之和的增加，大幅提高了其变量选取的稳健性。与随机森林方法表现类似，LASSO 选取的前十变量中有 8 个是 BMA 显著变量，其中 5 个是 BMA 前十显著变量。与这 4 个方法相对比，神经网络方法结果有较大区别，所选择的前十变量中，仅有 2 个是 BMA 显著变量。如第三节中所述，神经网络方法的特征是深入挖掘样本间横向或纵向的固定关联特性，从而获取优良的预测能力。但是在跨国经济增长的样本中，样本本身具有的横向固定相关性有限，因此其数据挖掘的能力无法凸显出来。

图 1 进一步绘制了回归树、套袋法与随机森林 3 种方法选择前十变量时各自的量化标准，即第三节 2.2 中描述的重要性指标。

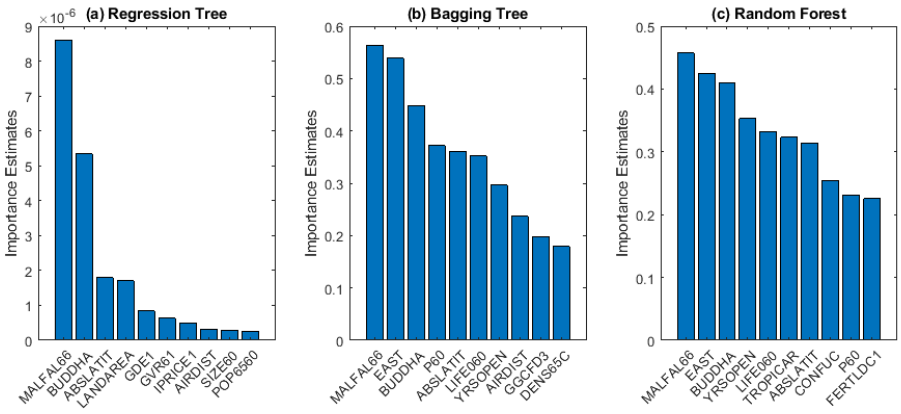


图 1：回归树、套袋法与随机森林法的前十变量选取结果

由图可见，回归树倾向于对不同变量给出差异明显的重要性判断。与此不同，套袋法与随机森林法对各自前十大解释变量的重要性评估较为一致，且后者较前者更为均衡。这反映出套袋法与随机森林法通过 Bootstrap 进行模拟交叉验证后，其产生的近似预测函数稳定性更高，变量选择过程不会由于微弱的样本异常导致结果的大幅改变[①]。

2.2. 预测有效性

为了评估机器学习方法给出的近似模型的有效性，我们系统对比不同方法预测值对应的 R^2。具体而言，给定样本数据 $\{y_i, \boldsymbol{x}_i\}_{i=1,\ldots,N}$ 以及一个机器学习方法所得的近似预测函数 $\hat{f}$，我们首先计算相应的样本内中心 R^2：

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2}{\sum_i (y_i - \bar{y})^2},$$

其中 $\bar{y}$ 为样本均值。注意，由于回归树及其衍生的套袋法与随机森林，近似函数（预测树）的构造过程中，均是以样本估计 SSR 为判断基础，因此本质上均是最小化 L^2-风险的预测方程。相应的，这些方法计算出的 R^2 应当在 0 和 1 之

① 注意，套袋法与随机森林法均依赖于通过 Bootstrap 方法随机模拟生成新的样本，因此计算结果会带有一定的随机性。在我们的实证分析中，Bootstrap 随机抽样的次数设为 200。表 3 及以下的结果中，涉及这两种方法的，严格说来均是特定随机抽样后的结果。不过，我们反复进行过多次抽样，对应的计算结果没有实证性改变。另一方面，我们可以通过增加随机抽样次数，消除计算结果中的随机性。但通过与随机抽样 10000 次的结果做对比，随机抽样 200 次的结果已经具有较好的稳健性。

间。与此不同，LASSO 考虑的不是纯粹的 L^2-风险，因此其产生的预测方程 $\hat{f}$ 可能导致 R^2 小于 0。注意，对于 LASSO 之外的 4 个机器学习方法，其产生的近似预测函数 $\hat{f}$ 均可能会呈现出高度非线性的特征。

注意上述中心 R^2 反映的是对数据样本整体的拟合情况，从模型不确定性问题角度看，主要着眼于样本中的非线性特征。为了更加突出模型不确定性的另一个来源——变量选择问题——我们暂时将非线性特征搁置，利用 5 个机器学习方法所挑选出的 5 组前十预测变量，在线性回归框架下，考察其样本内预测功效。这样做的一个好处，在于我们可以有效对比回归树、套袋法、随机森林与 LASSO 四个方法在线性约束下的预测能力，避免 LASSO 方法本身 R^2 评价的不适用性。此外，我们还可以对比包括所有 67 个变量的全样本线性回归模型的预测表现，从而更全面的考察机器学习方法的有效性。

表 4：模型有效性（样本内预测）评估

	LASSO	**回归树**	**套袋法**	**随机森林**	**神经网络**	**GUM**
	全变量非线性样本内预测					
MAE	0.0207	0.0047	0.0047	0.0051	0.0046	
中心 R^2	-0.4262	0.8696	0.8818	0.8690	0.8862	
	前十大变量线性样本内预测					
中心 R^2	0.5449	0.6215	0.6671	0.6016	0.5766	0.9135
调整 R^2	0.4858	0.5723	0.6239	0.5499	0.5216	0.6236

注：GUM 表示一般无约束线性预测模型，此处指包含全部 67 个变量的线性回归模型

表 3 汇报了按照上述两种方式所进行的模型预测效力评估。其中上半部分报告了使用各个机器学习方法产生的原始（非线性）近似预测函数的 R^2 表现以及预测残差绝对值的平均值（mean absolute error）。总体而言，除 LASSO 之外的 4 种方法，均能较好的捕捉 88 个国家样本数据的总体特征，其中心 R^2 没有明显差别。GUM 所代表的包含 67 个解释变量的线性回归模型产生了高于机器学习方法的中心 R^2，但这仅只是由于解释变量数目接近样本量本身这一事实。如果将参数个数考虑进来，GUM 调整 R^2 自然大幅下降。

表 3 下半部分的结果将预测模型约束到线性范围内，并且只考虑 5 个机器学习方法各自挑选的前十大变量。如此我们可以暂时忽略样本非线性性，而只聚焦变量选择不确定性问题。此时中心 R^2 已经呈现出明显差别。回归树因为过度拟合数据局部特征，而忽略了数据整体的线性趋势特征，造成其中心 R^2 相比无约束非线性情况大幅下降。套袋法与随机森林法由于更加稳健的预测特征，能够较好的捕捉数据的线性趋势特征，因此对应的中心 R^2 表现优越。LASSO 由于模型本身的线性特征，其约束预测效果也具有优势。比较 5 个方法在受约束线性预测时的调整 R^2，可以完全消除参数数目的影响，并且可以直接与 GUM 的调整 R^2 进行对比。结果显示，套袋法所选变量呈现出最好的线性预测能力，甚至大幅超过基准模型为线性的 LASSO 方法。同时，考虑到参数个数后，套袋法、随机森林与 LASSO 所选变量的线性预测能力，均高于或接近包含所有 67 个解释变量的 GUM，进一步说明了机器学习方法较传统线性模型在克服变量不确定性问题时的明显优势。

2.3. 预测函数特征分解

为了进一步理解机器学习方法在跨国经济增长预测效能方面的突出优势，我们利用第三节中论述的预测函数关键特征测度方法体系，对 5 种机器学习方法所得的预测函数进行分析，特别突出各方法选取前十变量的关键特征。

第一，我们讨论 5 种机器学习方法所产生的 5 个预测函数中，单个预测变量的解释力度。正文中，我们仅考虑各个方法选取的前十预测变量的解释力度，结果见表 5[①]。表中 FEPV 一列，报告了每个方法下前十变量所具有的解释力度，以百分比表示，测算方法见上一节(3)式。变量代码下出现下划线的，说明该变量位列该方法 67 个变量解释力度排序前十。显然，除了神经网络法之外，其他 4 种方法所选取的前十预测变量，基本上同时也是解释力度排名前十的变量。各个方法选取的前十预测变量所具有的单一变量平均解释力度相差较大，从回归树方法的最高值 4.73%，到神经网络法的最低值 0.33%，说明各种方法的预测函数对变

① 67 个变量的完整结果见附录 B 表 B2。

量间的交互作用捕捉程度不尽相同[①]。套袋法与随机森林法下，前十预测变量的平均解释力度分别为 0.85%与 0.47%，但联系到表 4 中这两个方法前十变量的良好预测效能，这说明套袋法与随机森林法通过充分捕捉变量间的交互预测效能，大幅提高了各自前十变量的总体预测解释力。

与表 1 中各个变量所属理论分组类别相对照可见，LASSO、回归树、套袋法与随机森林法所识别出的主要预测解释变量，集中在自然地理环境、宗教、人口结构、区位和贸易等方面，更多的反映经济增长的深层次根源（Spolaore and Wacziarg 2013）。归属于新古典增长理论类别的变量，主要为 P60，即 60 年代的小学教育程度。特别的，1960 年的人均 GDP（GDPCH60L）不但没有进入各方法选取的前十预测变量行列，甚至都不属于各方法下具有前十解释力度的变量序列[②]。这说明新古典增长模型的主要理论预测，即经济增长的绝对收敛或者俱乐部收敛，在样本数据中很难获得支持。

表 5：5 种方法前十预测变量特征测度

LASSO	**FEPV**	**NL**	**回归树**	**FEPV**	**NL**	**套袋法**	**FEPV**	**NL**
YRSOPEN	1.54	0.00	MALFAL66	14.63	72.65	MALFAL66	2.26	26.10
EAST	6.49	0.00	BUDDHA	12.68	24.68	EAST	1.55	5.88
TROPPOP	0.57	0.00	ABSLATIT	7.19	99.93	BUDDHA	2.01	22.24
P60	4.48	0.00	LANDAREA	4.02	99.61	P60	0.55	29.56
MALFAL66	0.90	0.00	GDE1	2.98	93.88	ABSLATIT	0.65	40.76
CONFUC	4.00	0.00	GVR61	0.97	95.97	LIFE060	1.25	46.01
RERD	0.96	0.00	IPRICE1	1.68	98.52	YRSOPEN	0.12	22.37
IPRICE1	3.21	0.00	AIRDIST	1.26	71.61	AIRDIST	0.05	51.41
BUDDHA	1.80	0.00	SIZE60	1.03	85.36	GGCFD3	0.01	84.48
GVR61	0.54	0.00	POP6560	0.88	97.64	DENS65C	0.05	85.40
平均	2.45	0.00	平均	4.73	83.99	平均	0.85	41.42
随机森林			**神经网络**					
MALFAL66	0.90	28.64	SPAIN	0.51	2.38			
EAST	0.68	4.12	GOVNOM1	0.18	11.04			

[①] 五种方法下，67 个变量的累计单一变量解释力度也呈现较大差异：LASSO 为 27.81%，回归树最高为 48.51%，套袋法为 9.44%，随机森林法最低为 5.77%，神经网络法为 17.59%。根据第三节第 4 小节的测算框架，不能由单一变量解释的预测变动，均归因于变量间的多重交互作用。这说明，各个方法的预测结果中，归为多重交互效应的部分存在较大差异。

[②] 详见附录 B 表 B1。GDPCH60L 仅在神经网络方法下进入解释力前十行列，在回归树方法中位列 12，在其他 3 种方法中均在解释力度前 20 之外。

BUDDHA	1.10	24.70	LANDAREA	0.06	25.02
YRSOPEN	0.30	22.34	MUSLIM00	0.02	91.19
LIFE060	0.48	51.05	SOCIALIST	0.24	21.59
TROPICAR	0.09	27.14	OPENDEC1	0.02	34.83
ABSLATIT	0.26	43.02	EAST	1.54	0.58
CONFUC	0.17	32.15	TOT1DEC1	0.03	31.25
P60	0.67	24.07	IPRICE1	0.58	3.90
FERTLDC1	0.06	62.89	REVCOUP	0.08	16.66
平均	0.47	32.01	平均	0.33	23.84

注：表中数字单位为百分比；FEPV 由(3)式定义，NL 表示非线性性测度，由(7)式辅助回归进行测算；下划线表示该变量解释力度排名为该方法下 67 个变量的前十位

第二，表 5 中 NL 的一列，报告了各个变量在相应预测函数中非线性特征，单位同为百分比，测度方法为 100%减去(7)式辅助回归所得的R^2。除去 LASSO 方法总是线性预测函数，故非线性性总为 0 之外，其他 4 个方法均呈现出一定的非线性特征。其中回归树的非线性性最高，平均而言其前十预测变量的预测效果中 84%来自非线性变动。与之相比，套袋法与随机森林法通过使用 Bootstrap 抽样方法，平滑了回归树方法频繁出现的局部非线性拟合，因此变量的平均非线性程度有明显下降。比较令人意外的是神经网络法产生的预测函数并未显示出突出的非线性特征，变量的平均非线性性仅 20%有余[①]。在下一小节中，我们选择套袋法和随机森林法，用图示方法进一步说明预测函数的非线性性。

第三，我们用热力图的形式报告 5 种机器学习方法下，前十变量的两两交互作用对预测值的贡献。具体结果见图 2。其中，每个方法所得结果单独以子图（热力图）形式呈现。每个热力图的上三角区域空白，两两变量组合在下三角区域已经出现过。热力图矩阵中，每个方块的颜色代表两两变量组合交互预测解释力（FIEPV，见 6 式定义），颜色越深表示解释力越大；方块中的数字代表解释力数值，单位为百分比。注意，每个子图所用的色温尺度均不相同。

① 尽管神经网络方法中神经元层的 sigmoid 转换函数可以灵活拟合各类非线性函数，但是在跨国经济增长的样本数据下，神经网络学习的结果倾向于平均使用各个变量，导致总体预测函数更接近一个无约束（无变量选择）的线性预测形式。这一点在附录图 B3 中有更明确的呈现。

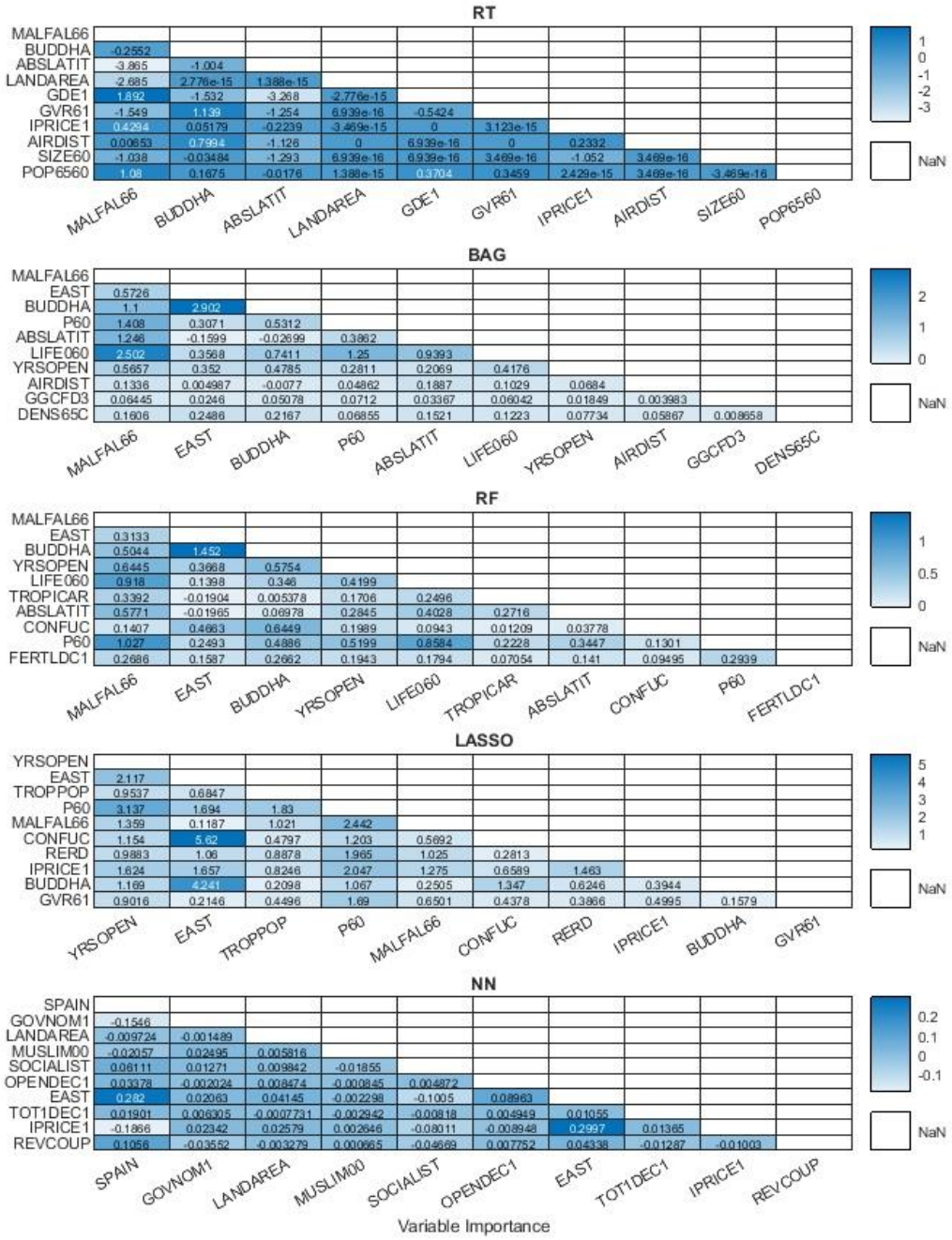


图 2：变量交互作用解释力

从图 2 的整体特征来看，可以发现神经网络方法下，变量间的双重交互作用大小比较一致，且幅度均较为有限；而回归树方法下，存在比较突出的正、负向双重交互作用。与此不同，LASSO、套袋法与随机森林法下，双重交互作用均为正向，且两两变量组合件交互作用差异较大。从幅度上看，套袋法与 LASSO 均存在交互作用大于 2%的情况，联系表 5 给出的单一变量解释力度，可见交互作用对预测值的贡献可以非常显著。

观察套袋法、随机森林法与 LASSO 三种方法，一个突出的特征是显著的双重交互作用更容易出现在同类别或相近类别变量之间。如儒家文化（CONFUC）、佛教比例（BUDDHA）与东亚地区（EAST）三者之间，就呈现出显著的两两正向交互作用。类似的，疟疾流行程度（MALFAL60）与预期寿命（LIFE60）之间，也存在很强的正向交互作用。这些特征说明，在经济增长的深层次决定因素中，不同类别深层次因素或同类别不同因素之间的交互作用，对经济增长具有重要的预测意义。同时，这也说明拘泥于单纯线性模型的跨国实证研究，会容易忽略掉深层次决定因素间通过交互效应带来的重要影响。

2.4. 非线性特征图示

最后，我们着重说明机器学习方法在应对数据中可能存在的非线性特征时的显著优势。由于 LASSO 本身是线性模型，而前面汇报的结果已经说明回归树方法本身的局限性，以及神经网络法所呈现出的近似线性特征，因此下面只考虑套袋法与随机森林两个方法。

如第三节第 4 小节所述，为反映变量 x_k 对预测值的影响，我们可以聚焦在该变量带来的预测值变动 $\Delta\hat{y}_{[k]}$ 上。图 3 和图 4，分别对套袋法和随机森林法所选取的前十预测变量，绘制 $\Delta\hat{y}_{[k]}$ 与 x_k 的样本关系图。同时，在每个变量的子图中，我们还添加了一条红色的线性趋势线——实质上，这条线就代表了辅助回归(7)式的结果。注意，每个子图中的蓝色圆点，代表的该样本点处的预测值（变动）$\Delta\hat{y}_{[k]i}$。

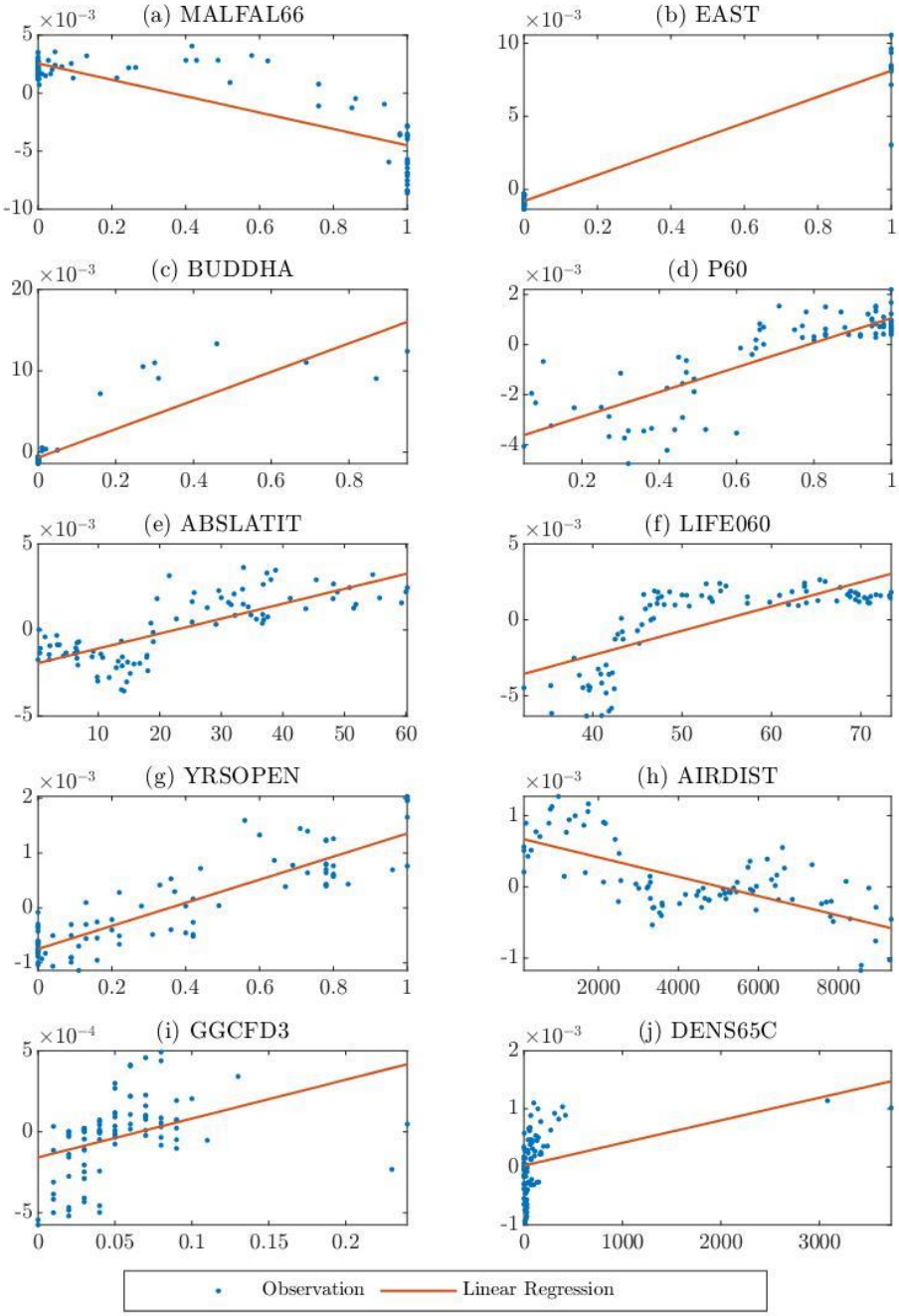


图 3：套袋法预测值及趋势

注意这些图中的纵坐标均为经济增长速度的水平值，因此相应解释变量对经济增长的边际作用反映在图中趋势线的斜率和预测样本点的整体走势上。以图 3 中子图（a）为例，解释变量 1960 年代疟疾流行程度对经济增长的作用，仅在其水平上升到一定程度（0.6 以上）才开始显现，在此之前疟疾流行程度并不影响

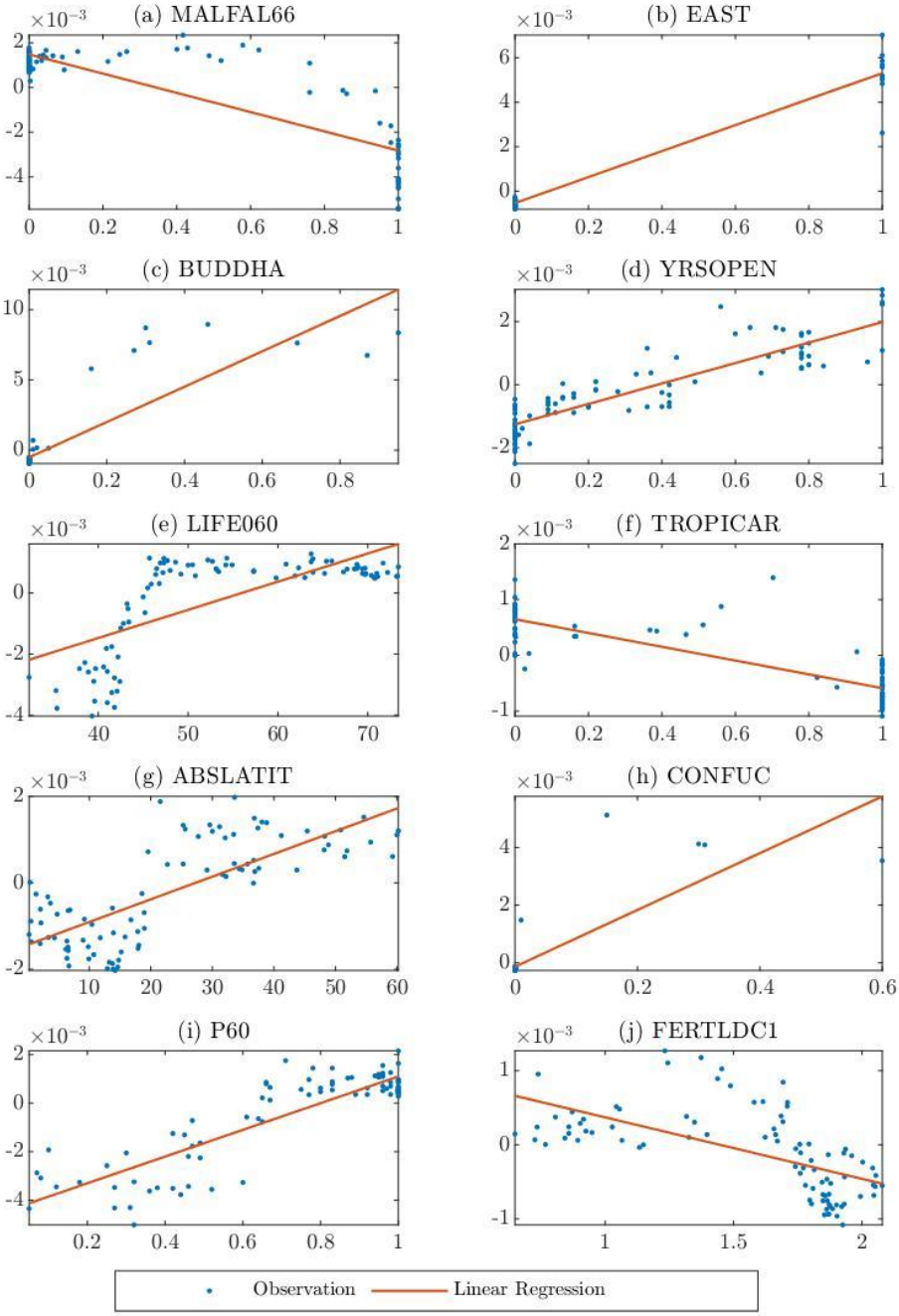


图 4：随机森林法预测值及趋势

经济增长。类似的情况还出现在子图（d） 1960 年代小学教育普及水平与子图（f）1960 年代人均预期寿命等解释变量中。对照图 3 和 4，也可以观察到，部分预测变量呈现出较为一致的增长预测作用，非线性性程度较低。如经济开放程

度（YRSOPEN），不论在哪种方法与何种样本区间下，总体的线性趋势是其预测效能的主要部分。

五、 结论

经过近 30 年的发展，跨国实证研究中最主流的回归分析范式，日益显示出方法论困境。样本有限性、模型不确定性、结论因果性等 3 个方面的挑战，对传统跨国实证分析方法的进一步发展形成了显著的制约。针对这一困境，本文提出可以充分利用新近的机器学习方法，有效缓解或克服传统分析方法存在的局限。本文系统阐述了机器学习方法的三个优势：适应小样本问题；处理变量选择问题；克服样本非线性问题。与此同时，机器学习方法的这些特性又能够有效缓解传统分析中普遍存在的内生性问题。利用 Sala-i-Martin et al.（2004）标准的跨国经济增长数据集，我们对 LASSO、回归树、套袋法、随机森林及神经网络法等 5 种最广泛使用的数值型监督学习方法进行了实证测试，以说明其相对传统分析方法的优势。研究结果显示，套袋法与随机森林法在跨国经济增长及类似实证问题中具有突出的适用性、有效性、稳健性。

参考文献

[1] Athey S, Imbens G W. The State of Applied Econometrics: Causality and Policy Evaluation[J]. Journal of Economic Perspectives. 2017, 31(2): 3-32.

[2] Barro R J. Economic Growth in a Cross Section of Countries[J]. Quarterly Journal of Economics. 1991, 106(2): 407-443.

[3] Barro R J, Lee J W. Sources of Economic Growth[J]. Carnegie-Rochester Conference Series on Public Policy. 1994, 40: 1-46.

[4] Barro R J, Sala-I-Martin X. Convergence[J]. Journal of Political Economy. 1992, 100(2): 223-251.

[5] Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[C]. ACM, 1992.

[6] Breiman L. Bagging Predictors[J]. Machine Learning. 1996, 24(2): 123-140.

[7] Breiman L. Random Forests[J]. Machine Learning. 2001, 45(1): 5-32.

[8] Breiman L, Friedman J, Olshen R A, et al. Classification and Regression Trees[M]. Roca Raton: Chapman and Hall/CRC, 1984.

[9] Brock W A, Durlauf S N. What Have We Learned from a Decade of Empirical Research on Growth? Growth Empirics and Reality[J]. World Bank Economic Review. 2001, 15(2): 229-272.

[10] Canova F. Testing for Convergence Clubs in Income Per Capita: A Predictive Density Approach[J]. International Economic Review. 2004, 45(1): 49-77.

[11] Ciccone A, Jarociński M. Determinants of Economic Growth: Will Data Tell?[J]. American Economic Journal: Macroeconomics. 2010, 2(4): 222-246.

[12] Cortes C, Vapnik V. Support-vector Networks[J]. Machine Learning. 1995, 20(3): 273-297.

[13] Durlauf S N. The Rise and Fall of Cross-Country Growth Regressions[J]. History of Political Economy. 2009, 41(Suppl_1): 315-333.

[14] Durlauf S N, Johnson P A. Multiple Regimes and Cross-country Growth Behaviour[J]. Journal of Applied Econometrics. 1995, 10(4): 365-384.

[15] Durlauf S N, Johnson P A, Temple J R W. Growth Econometrics[M]. Aghion P, Durlauf S N, Elsevier, 2005: 1A, 555-677.

[16] Durlauf S N, Kourtellos A, Minkin A. The Local Solow Growth Model[J]. European Economic Review. 2001, 45(4): 928-940.

[17] Durlauf S N, Kourtellos A, Tan C M. Are Any Growth Theories Robust?[J]. Economic Journal. 2008, 118(527): 329-346.

[18] Easterly W, Kremer M, Pritchett L, et al. Good Policy or Good Luck?[J]. Journal of Monetary Economics. 1993, 32(3): 459-483.

[19] Fernandez C, Ley E, Steel M F J. Model Uncertainty in Cross-country Growth Regressions[J]. Journal of Applied Econometrics. 2001, 16(5): 563-576.

[20] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Prediction, Inference and Data Mining[M]. 2 ed. New York: Springer-Verlag, 2009.

[21] Henderson D J, Papageorgiou C, Parmeter C F. Growth Empirics without Parameters[J]. Economic Journal. 2012, 122(559): 125-154.

[22] Johnson P A, Papageorgiou C. What Remains of Cross-Country Convergence?[J]. Journal of Economic Literature. 2019, forthcoming.

[23] Johnson P A, Takeyama L N. Initial Conditions and Economic Growth in the US States[J]. European Economic Review. 2001, 45(4): 919-927.

[24] Kormendi R C, Meguire P G. Macroeconomic Determinants of Growth: Cross-country Evidence[J]. Journal of Monetary Economics. 1985, 16(2): 141-163.

[25] Kourtellos A, Tan C M, Zhang X. Is the Relationship between Aid and Economic Growth Nonlinear?[J]. Journal of Macroeconomics. 2007, 29(3): 515-540.

[26] Lehrer S F, Xie T. The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success[Z]. 2018.

[27] Levine R, Renelt D. A Sensitivity Analysis of Cross-Country Growth Regressions[J]. American Economic Review. 1992, 82(4): 942-963.

[28] Ley E, Steel M F J. On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression[J]. Journal of Applied Econometrics. 2009, 24(4): 651-674.

[29] Ley E, Steel M F J. Mixtures of g-priors for Bayesian Model Averaging with Economic Applications[J]. Journal of Econometrics. 2012, 171(2): 251-266.

[30] Liu Y, Xie T. Machine Learning Versus Econometrics: Prediction of Box Office[J]. Applied Economics Letters. 2019, 26(2): 124-130.

[31] Liu Z, Stengos T. Non-linearities in Cross-country Growth Regressions: A Semiparametric Approach[J]. Journal of Applied Econometrics. 1999, 14(5): 527-538.

[32] Masanjala W H, Papageorgiou C. The Solow Model with CES Technology: Nonlinearities and

Parameter Heterogeneity[J]. Journal of Applied Econometrics. 2004, 19(2): 171-201.
[33] Minier J. Nonlinearities and Robustness in Growth Regressions[J]. American Economic Review. 2007, 97(2): 388-392.
[34] Minier J. Institutions and Parameter Heterogeneity[J]. Journal of Macroeconomics. 2007, 29(3): 595-611.
[35] Mitchell T M. Machine Learning[M]. McGraw-Hill, 1997.
[36] Mullainathan S, Spiess J. Machine Learning: An Applied Econometric Approach[J]. Journal of Economic Perspectives. 2017, 31(2): 87-106.
[37] Sala-I-Martin X, Doppelhofer G, Miller R I. Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach[J]. American Economic Review. 2004, 94(4): 813-835.
[38] Spolaore E, Wacziarg R. How Deep Are the Roots of Economic Development?[J]. Journal of Economic Literature. 2013, 51(2): 325-369.
[39] Tan C M. No One True Path: Uncovering the Interplay between Geography, Institutions, and Fractionalization in Economic Development[J]. Journal of Applied Econometrics. 2010, 25(7): 1100-1127.
[40] Tibshirani R. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological). 1996, 58(1): 267-288.
[41] Varian H R. Big Data: New Tricks for Econometrics[J]. Journal of Economic Perspectives. 2014, 28(2): 3-28.
[42] 陈硕，王宣艺. 机器学习在社会科中的应用：回顾与展望[Z]. 复旦大学, 2018.
[43] 黄乃静，于明哲. 机器学习对经济学研究的影响研究进展[J]. 经济学动态. 2018(07): 115-129.

附录

A. 测算框架的详细说明

A.1.单个变量解释效力的测算

给定预测函数 $\hat{f}$，我们希望衡量某个具体的预测变量 x_{ki} 的样本变动，最终带来了多少样本内预测值 $\hat{y}_i$ 的变动，从而得到该变量对预测值内样本内变动的贡献份额。这里的难点，在于如何将 x_{ki} 样本变动的同时，其他变量 $x_{ji}, j \neq k$ 样本变动带来的预测值 $\hat{y}_i$ 变动，纳入考虑。在通常的“加法”思维下，我们希望固定住其他变量，只让 x_{ki} 变化并观察预测值 $\hat{f}(\dots, x_{ki}, \dots)$ 出现了多少变化。但这样一来，我们所衡量的 x_{ki} 对预测值变动的贡献，本质上都是“条件”评估——测算结果取决于其他变量被固定住的取值水平；与此同时，给定其他变量样本取值的多样性，并没有任何一组特定的取值具有明显的“合理性”。这样一来，上述条件评估本身又变成一个高维度的对象，从而失去了对单一变量解释力贡献测度的简明性。与此相反，我们可以采取一个“减法”思维，直接获得一个“无条件”测度。具体而言，我们定义预测函数 $\hat{f}$ 关于变量 x_k 的平均可解释预测变动（average explained prediction variation）如下：

$$AEPV(\hat{f}; k) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\hat{f}\left(x_{1i}, \dots, x_{k-1,i}, x_{kj}, x_{k+1,i}, \dots, x_{Ki}\right)\right)^2. \quad \text{(A8)}$$

上式中的第 i 个求和项解释如下：首先，给定样本点 $\boldsymbol{x}_i$ 及其对应的预测值 $\hat{f}(\boldsymbol{x}_i) = \hat{f}(x_{1i}, \dots, x_{ki}, \dots, x_{Ki})$，固定 x_k 之外的变量样本取值不变，但让 x_k 取遍样本取值 $\{x_{k1}, \dots, x_{kN}\}$ 并计算对应预测值的平均，则这一平均预测值反映了除 x_k 之外所有其余变量对该样本点处预测值的贡献，同时仅保留 x_k 自身在整个样本取值范围内预测贡献的平均说明；其次，$\hat{f}(\boldsymbol{x}_i)$ 与该预测均值的离差，刻画了该样本点处，x_k 取对应样本值 x_{ki} 时，预测值的改进程度，从而反映了变量 x_k 在该样本点对于预测值的贡献大小；最后，将所有样本点处的离差取平方求和再取平均，即得到变量 x_k 在所有样本点处，对样本内预测值变动的总体平均

解释度。下面第 A.4 小节，我们会进一步用线性预测函数的例子，说明上述公式的意义。

我们可以进一步定义平均预测变动（average prediction variation）如下：

$$APV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\hat{f}(\boldsymbol{x}_j)\right)^2, \tag{A9}$$

换言之，$APV(\hat{f})$ 就是样本内预测值的方差。在此基础上，我们定义预测函数 $\hat{f}$ 关于变量 x_k 的可解释预测变动占比（fraction of EPV）：

$$FEPV(\hat{f};k) = \frac{AEPV(\hat{f};k)}{APV(\hat{f})}. \tag{A10}$$

这一占比变量，简洁的反映了 x_k 对预测函数 $\hat{f}$ 样本内预测的贡献比例。

A.2. 两个变量联合解释效力的测算

有了单变量情形作为铺垫，我们下面讨论两个变量的情形。首先，沿用 $AEPV(\hat{f};\cdot)$ 的记号，我们定义两个变量 x_k, x_m 的联合可解释预测变动如下：

$$AEPV(\hat{f};k,m) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\hat{f}(\dots,x_{kj},\dots,x_{mj},\dots)\right)^2, \tag{A11}$$

上式求和项中 $\hat{f}(\dots,x_{kj},\dots,x_{mj},\dots)$ 的省略号表示除 x_k, x_m 外所有其他变量均为样本点 i 处的取值。这里定义的 $AEPV(\hat{f};k,m)$ 包括了 x_k, x_m 两个变量各自对预测值的独立贡献，以及两个变量对预测值的交互贡献。为测量这后一部分预测贡献，我们可以进一步定义交互可解释预测变动（interactive EPV）：

$$IEPV(\hat{f};k,m) = AEPV(\hat{f};k,m) - AEPV(\hat{f};k) - AEPV(\hat{f};m). \tag{A12}$$

同样，我们将在第 A.4 节中用线性预测函数为例，进一步说明 $AEPV(\hat{f};k,m)$ 及 $IEPV(\hat{f};k,m)$ 的涵义。与单变量情形类似，可以计算上述两个特征量与

$AEPV(\hat{f})$ 的比例，用来简洁测度两个变量的联合解释贡献占比与交互解释贡献占比：

$$FEPV(\hat{f};k,m)=\frac{AEPV(\hat{f};k,m)}{APV(\hat{f})},\qquad FIEPV(\hat{f},k)=\frac{IEPV(\hat{f};k,m)}{APV(\hat{f})}.\qquad \text{(A13)}$$

A.3. 预测函数非线性性的度量

正如前文所述，机器学习方法较传统计量方法——特别是以线性回归为基础的线性预测方法——的一个突出优势是可以灵活的捕捉、处理数据样本中存在的非线性性。但如何刻画机器学习方法所生成预测函数的非线性性，却并没有一套现成的技术。不过，我们下面马上会说明，在上述变量解释力测算框架下，我们可以很自然的度量预测函数的非线性性。

参照 A.1 小节，我们可以将 $\Delta\hat{y}_{[k]i}\equiv\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_j\hat{f}(\dots,x_{kj},\dots)$ 视作变量样本值 x_{ki} 对预测值 $\hat{y}_i$ 的预测贡献。若以 $\Delta\hat{y}_{[k]}$ 为分析对象，并将其视作仅与 x_k 有关，那么我们可以进一步将 x_k 对目标变量的预测效果——即 $\Delta\hat{y}_{[k]}$ 体现的部分——分解为线性趋势项与非线性项两个部分。为此，我们只需要用 $\Delta\hat{y}_{[k]}$ 对 x_k 进行如下的辅助性线性回归[①]：

$$\Delta\hat{y}_{[k]i}=\alpha+\beta_k x_{ki}+\xi_i,\qquad i=1,\dots,N,\qquad \text{(A14)}$$

进而可以将回归所得的线性部分 $\alpha+\beta_k x_{ki}$ 解释为 x_k 预测贡献中的线性趋势项，残差项 ξ_{ki} 则可视作相应的非线性部分。同时，该回归的 R^2 自然可以解释为 x_k 预测能力中线性趋势部分的比例，而 $1-R^2$ 则表示非线性变动部分的比例。

类似的，我们还可以衡量多个变量的联合非线性性。以两个变量为例说明，我们仅需定义 $\Delta\hat{y}_{[km]i}\equiv\hat{f}(\boldsymbol{x}_i)-\frac{1}{N}\sum_j\hat{f}(\dots,x_{kj},\dots,x_{mj},\dots)$，并通过 2-元辅助回归

$$\Delta\hat{y}_{[km]i}=\alpha+\beta_k x_{ki}+\beta_m x_{mi}+\xi_i,\qquad i=1,\dots,N$$

① 下一小节会说明，如果在辅助回归中忽略常数项，则预测函数非线性性可能会造成回归结果的偏误。

进行分析即可。

A.4. 测算框架的直观说明：三个示例

我们首先考察线性预测函数的例子。假设预测函数由如下形式：

$$\hat{f}(\boldsymbol{x}_i) = \boldsymbol{\beta}^\top \boldsymbol{x}_i = \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}.$$

直接计算可知，$\Delta\hat{y}_{[k]i} = \hat{f}(\boldsymbol{x}_i) - \frac{1}{N}\sum_j \hat{f}\left(\ldots, x_{kj}, \ldots\right) = \beta_k x_{ki} - \beta_k \bar{x}_k$，其中 $\bar{x}_k$ 表示样本均值。因此，(1)式定义的 $AEPV\left(\hat{f}; k\right) = \frac{\beta_k^2}{N}\sum_i (x_{ki} - \bar{x}_k)^2$，即线性回归中变量 x_k 的可解释样本方差（explained sample variance）。类似的，两个变量 x_k, x_m 联合时，$AEPV\left(\hat{f}; k, m\right) = \frac{1}{N}\sum_i (\beta_k(x_{ki} - \bar{x}_k) + \beta_m(x_{mi} - \bar{x}_m))^2$；此时存在两个变量交互作用对预测的贡献，相应的 $IEPV\left(\hat{f}; k, m\right) = \frac{2\beta_k\beta_m}{N}\sum_i (x_{ki} - \bar{x}_k)(x_{mi} - \bar{x}_m)$。此外，由于 $\Delta\hat{y}_{[k]i}$ 此时是线性依赖于 x_{ki}，故非线性性分解所用的辅助回归 (7) 可实现完美拟合，$R^2 = 1$；同时 $\Delta\hat{y}_{[k]i} = \beta_k x_{ki} - \beta_k \bar{x}_k$ 也说明辅助回归方程 (7) 中包括常数项的必要性。

接下来，我们考虑一般的可加性预测函数：

$$\hat{f}(\boldsymbol{x}_i) = g_1(x_{1i}) + \cdots + g_K(x_{Ki}),$$

上式中各个 $g_k(\cdot)$ 均假设为非线性函数。直接计算可知，$\Delta\hat{y}_{[k]i} = g_k(x_{ki}) - \bar{g}_k$，其中 $\bar{g}_k$ 表示 $g_k(x_{ki})$ 的样本均值。此时，单变量或双变量下 AEPV 和 IEPV 的表达式与线性预测函数时类似，不再赘述。但此时 $\Delta\hat{y}_{[k]i}$ 不再是线性的依赖于 x_{ki}，因此辅助回归 (7) 将通过 $R^2 < 1$ 捕捉到 $g_k(\cdot)$ 反映的非线性特征。

最后，我们考虑一个含有交叉项的（拟）线性预测模型：

$$\hat{f}(\boldsymbol{x}_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \lambda x_{1i} x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki}.$$

此时，对应的

$$\begin{aligned}
\Delta\hat{y}_{[1]i} &= (x_{1i} - \bar{x}_1)(\beta_1 + \lambda x_{2i}), \\
\Delta\hat{y}_{[2]i} &= (x_{2i} - \bar{x}_2)(\beta_2 + \lambda x_{1i}), \\
\Delta\hat{y}_{[12]i} &= \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \lambda(x_{1i}x_{2i} - \overline{x_1 x_2}) \\
&= (x_{1i} - \bar{x}_1)(\beta_1 + \lambda x_{2i}) + (x_{2i} - \bar{x}_2)(\beta_2 + \lambda x_{1i}) \\
&\quad + \lambda(\bar{x}_1 x_{2i} + x_{1i}\bar{x}_2 - x_{1i}x_{2i} - \overline{x_1 x_2}),
\end{aligned}$$

其中 $\overline{x_1x_2}$ 表示 $x_{1i}x_{2i}$ 的样本均值。由此可算得

$$
\begin{aligned}
IEPV(\hat{f};1,2) &= AEPV(\hat{f};1,2) - AEPV(\hat{f};1) - AEPV(\hat{f};2) \\
&= \frac{\lambda^2}{N}\sum_i (\bar{x}_1x_{2i} + x_{1i}\bar{x}_2 - x_{1i}x_{2i} - \overline{x_1x_2})^2 \\
&\quad + \frac{2\lambda}{N}\sum_i (x_{1i} - \bar{x}_1)(\beta_1 + \lambda x_{2i})\,(\bar{x}_1x_{2i} + x_{1i}\bar{x}_2 - x_{1i}x_{2i} - \overline{x_1x_2}) \\
&\quad + \frac{2\lambda}{N}\sum_i (x_{2i} - \bar{x}_2)(\beta_2 + \lambda x_{1i})\,(\bar{x}_1x_{2i} + x_{1i}\bar{x}_2 - x_{1i}x_{2i} - \overline{x_1x_2}) \\
&\quad + \frac{2}{N}\sum_i (x_{1i} - \bar{x}_1)(\beta_1 + \lambda x_{2i})(x_{2i} - \bar{x}_2)(\beta_2 + \lambda x_{1i})\,.
\end{aligned}
$$

B. 预测函数关键特征度量

B.1. 单一变量预测解释力度结果

表 B1：单一变量解释力度排序

LASSO		RT		BAG		RF		NN	
EAST	6.49	MALFAL66	14.63	MALFAL66	2.26	BUDDHA	1.10	SAFRICA	1.76
P60	4.48	BUDDHA	12.68	BUDDHA	2.01	MALFAL66	0.90	EAST	1.54
CONFUC	4.00	ABSLATIT	7.19	EAST	1.55	EAST	0.68	YRSOPEN	1.13
IPRICE1	3.21	LANDAREA	4.02	LIFE060	1.25	P60	0.67	P60	0.99
TROPICAR	2.18	GDE1	2.98	ABSLATIT	0.65	LIFE060	0.48	LIFE060	0.93
BUDDHA	1.80	IPRICE1	1.68	P60	0.55	YRSOPEN	0.30	CIV72	0.81
YRSOPEN	1.54	AIRDIST	1.26	DENS60	0.16	ABSLATIT	0.26	MALFAL66	0.78
RERD	0.96	SIZE60	1.03	YRSOPEN	0.12	CONFUC	0.17	PROT00	0.66
MALFAL66	0.90	GVR61	0.97	CONFUC	0.11	DENS65C	0.15	GDPCH60L	0.62
DENS65C	0.84	POP6560	0.88	IPRICE1	0.10	H60	0.15	IPRICE1	0.58
TROPPOP	0.57	RERD	0.52	TOTIND	0.08	TROPICAR	0.09	BRIT	0.58
GVR61	0.54	GDPCH60L	0.32	H60	0.06	DENS60	0.07	AVELF	0.56
MINING	0.12	DENS60	0.21	LANDAREA	0.06	TROPPOP	0.07	SPAIN	0.51
SPAIN	0.09	CATH00	0.15	RERD	0.05	FERTLDC1	0.06	LAAM	0.51
MUSLIM00	0.08	AVELF	0.00	POP6560	0.05	ZTROPICS	0.06	COLONY	0.36
OTHFRAC	0.00	BRIT	0.00	DENS65C	0.05	IPRICE1	0.06	OTHFRAC	0.34
ABSLATIT	0.00	CIV72	0.00	AIRDIST	0.05	TOTIND	0.05	LANDLOCK	0.32
AIRDIST	0.00	COLONY	0.00	DENS65I	0.03	AIRDIST	0.04	SCOUT	0.26
AVELF	0.00	CONFUC	0.00	SIZE60	0.02	DENS65I	0.04	LT100CR	0.25
BRIT	0.00	DENS65C	0.00	SAFRICA	0.02	POP6560	0.04	POP6560	0.24
CATH00	0.00	DENS65I	0.00	GDPCH60L	0.02	GDPCH60L	0.03	DPOP6090	0.24
CIV72	0.00	DPOP6090	0.00	PRIGHTS	0.01	RERD	0.03	SOCIALIST	0.24
COLONY	0.00	EAST	0.00	ZTROPICS	0.01	PRIEXP70	0.03	PRIEXP70	0.23
DENS60	0.00	ECORG	0.00	OPENDEC1	0.01	LANDAREA	0.02	RERD	0.21
DENS65I	0.00	ENGFRAC	0.00	HERF00	0.01	SAFRICA	0.02	TROPPOP	0.19
DPOP6090	0.00	EUROPE	0.00	TROPPOP	0.01	CATH00	0.02	GOVNOM1	0.18
ECORG	0.00	FERTLDC1	0.00	LHCPC	0.01	LAAM	0.02	CONFUC	0.16
ENGFRAC	0.00	GEEREC1	0.00	FERTLDC1	0.01	LHCPC	0.01	DENS60	0.14

EUROPE	0.00	GGCFD3	0.00	PRIEXP70	0.01	AVELF	0.01	PRIGHTS	0.14
FERTLDC1	0.00	GOVNOM1	0.00	REVCOUP	0.01	PRIGHTS	0.01	MINING	0.14
GDE1	0.00	GOVSH61	0.00	GGCFD3	0.01	GDE1	0.01	FERTLDC1	0.13
GDPCH60L	0.00	H60	0.00	GEEREC1	0.01	DPOP6090	0.01	BUDDHA	0.13
GEEREC1	0.00	HERF00	0.00	GDE1	0.01	SPAIN	0.01	EUROPE	0.12
GGCFD3	0.00	HINDU00	0.00	LT100CR	0.01	OPENDEC1	0.01	WARTORN	0.11
GOVNOM1	0.00	LAAM	0.00	GOVSH61	0.01	LT100CR	0.01	ZTROPICS	0.11
GOVSH61	0.00	LANDLOCK	0.00	SPAIN	0.01	REVCOUP	0.01	OIL	0.10
H60	0.00	LHCPC	0.00	CATH00	0.01	PROT00	0.01	ECORG	0.10
HERF00	0.00	LIFE060	0.00	CIV72	0.01	HERF00	0.01	GOVSH61	0.10
HINDU00	0.00	LT100CR	0.00	TROPICAR	0.00	SIZE60	0.01	CATH00	0.10
LAAM	0.00	MINING	0.00	PROT00	0.00	POP1560	0.01	GGCFD3	0.09
LANDAREA	0.00	MUSLIM00	0.00	LAAM	0.00	COLONY	0.01	POP1560	0.08
LANDLOCK	0.00	NEWSTATE	0.00	AVELF	0.00	GEEREC1	0.01	NEWSTATE	0.08
LHCPC	0.00	OIL	0.00	POP1560	0.00	CIV72	0.01	REVCOUP	0.08
LIFE060	0.00	OPENDEC1	0.00	GOVNOM1	0.00	GOVSH61	0.00	LANDAREA	0.06
LT100CR	0.00	ORTH00	0.00	TOT1DEC1	0.00	GGCFD3	0.00	LHCPC	0.06
NEWSTATE	0.00	OTHFRAC	0.00	GVR61	0.00	POP60	0.00	GEEREC1	0.05
OIL	0.00	P60	0.00	MUSLIM00	0.00	GOVNOM1	0.00	SIZE60	0.05
OPENDEC1	0.00	PI6090	0.00	PI6090	0.00	SQPI6090	0.00	DENS65C	0.05
ORTH00	0.00	SQPI6090	0.00	DPOP6090	0.00	MINING	0.00	H60	0.05
PI6090	0.00	PRIGHTS	0.00	POP60	0.00	TOT1DEC1	0.00	TROPICAR	0.04
SQPI6090	0.00	POP1560	0.00	MINING	0.00	ENGFRAC	0.00	PI6090	0.04
PRIGHTS	0.00	POP60	0.00	OTHFRAC	0.00	GVR61	0.00	ENGFRAC	0.04
POP1560	0.00	PRIEXP70	0.00	COLONY	0.00	OTHFRAC	0.00	TOT1DEC1	0.03
POP60	0.00	PROT00	0.00	HINDU00	0.00	PI6090	0.00	HERF00	0.03
POP6560	0.00	REVCOUP	0.00	ENGFRAC	0.00	MUSLIM00	0.00	GDE1	0.03
PRIEXP70	0.00	SAFRICA	0.00	SCOUT	0.00	WARTIME	0.00	TOTIND	0.03
PROT00	0.00	SCOUT	0.00	ECORG	0.00	NEWSTATE	0.00	AIRDIST	0.03
REVCOUP	0.00	SOCIALIST	0.00	WARTIME	0.00	ECORG	0.00	OPENDEC1	0.02
SAFRICA	0.00	SPAIN	0.00	BRIT	0.00	HINDU00	0.00	MUSLIM00	0.02
SCOUT	0.00	TOT1DEC1	0.00	NEWSTATE	0.00	SCOUT	0.00	ABSLATIT	0.01
SIZE60	0.00	TOTIND	0.00	ORTH00	0.00	BRIT	0.00	WARTIME	0.01
SOCIALIST	0.00	TROPICAR	0.00	LANDLOCK	0.00	WARTORN	0.00	DENS65I	0.01
TOT1DEC1	0.00	TROPPOP	0.00	WARTORN	0.00	ORTH00	0.00	GVR61	0.01
TOTIND	0.00	WARTIME	0.00	EUROPE	0.00	LANDLOCK	0.00	SQPI6090	0.00
WARTIME	0.00	WARTORN	0.00	OIL	0.00	EUROPE	0.00	POP60	0.00
WARTORN	0.00	YRSOPEN	0.00	SQPI6090	0.00	OIL	0.00	HINDU00	0.00
ZTROPICS	0.00	ZTROPICS	0.00	SOCIALIST	0.00	SOCIALIST	0.00	ORTH00	0.00

注：每种方法对应变量按照解释力度从大到小排序；每个变量右侧数字为百分比解释力度，单位为百分比

B.2. 单一变量预测非线性性测度结果

表 B2：单一变量预测非线性性排序

LASSO		RT		BAG		RF		NN	
IPRICE1	0.00	BUDDHA	24.68	EAST	5.88	EAST	4.12	DENS65C	0.55
DENS65C	0.00	AIRDIST	71.61	ENGFRAC	14.91	SAFRICA	10.32	EAST	0.58
YRSOPEN	0.00	MALFAL66	72.65	SPAIN	15.82	SPAIN	12.65	PROT00	1.50

P60	0.00	GDPCH60L	84.08	BUDDHA	22.24	LAAM	16.04	GDPCH60L	1.93
RERD	0.00	EAST	84.75	YRSOPEN	22.37	TROPPOP	19.72	BUDDHA	2.22
GVR61	0.00	SIZE60	85.36	SAFRICA	23.62	COLONY	20.21	SPAIN	2.38
EAST	0.00	CONFUC	85.47	MALFAL66	26.10	YRSOPEN	22.34	YRSOPEN	2.50
CONFUC	0.00	DENS65C	88.79	LAAM	27.98	P60	24.07	SAFRICA	2.59
TROPICAR	0.00	ENGFRAC	90.68	RERD	29.44	BUDDHA	24.70	LIFE060	2.82
BUDDHA	0.00	SCOUT	93.78	P60	29.56	TROPICAR	27.14	CIV72	3.33
MALFAL66	0.00	GDE1	93.88	CONFUC	29.94	ENGFRAC	27.60	RERD	3.44
MINING	0.00	DENS60	94.40	COLONY	35.90	MALFAL66	28.64	CONFUC	3.72
TROPPOP	0.00	CATH00	94.57	ABSLATIT	40.76	CONFUC	32.15	P60	3.87
MUSLIM00	0.00	PROT00	95.00	REVCOUP	41.22	RERD	35.46	IPRICE1	3.90
SPAIN	0.00	DENS65I	95.60	LIFE060	46.01	REVCOUP	40.35	MINING	5.49
OTHFRAC	0.00	GVR61	95.97	TROPPOP	47.17	ABSLATIT	43.02	AVELF	5.50
SOCIALIST	96.00	RERD	95.97	ZTROPICS	47.31	EUROPE	43.52	COLONY	6.23
OPENDEC1	96.85	EUROPE	96.22	CIV72	48.51	CIV72	43.59	DPOP6090	6.26
WARTORN	97.62	TOT1DEC1	96.32	TROPICAR	50.25	PRIEXP70	45.17	MALFAL66	7.13
WARTIME	97.97	COLONY	96.87	SCOUT	50.58	AVELF	46.07	H60	7.42
GDPCH60L	98.34	MINING	97.06	AIRDIST	51.41	LIFE060	51.05	GGCFD3	7.75
GOVSH61	98.38	LAAM	97.26	PROT00	52.16	PROT00	51.57	LAAM	8.86
ZTROPICS	98.58	ZTROPICS	97.27	PRIEXP70	52.25	ZTROPICS	51.78	PI6090	9.12
DENS65I	98.67	OPENDEC1	97.54	ECORG	55.42	NEWSTATE	52.42	GOVNOM1	11.04
GOVNOM1	98.76	H60	97.63	SIZE60	55.52	AIRDIST	52.86	POP6560	11.72
TOT1DEC1	98.82	POP6560	97.64	HINDU00	59.26	PRIGHTS	53.36	PRIEXP70	11.96
SCOUT	99.06	SPAIN	97.67	PRIGHTS	59.51	GDPCH60L	53.90	FERTLDC1	12.02
POP6560	99.27	OIL	97.90	NEWSTATE	63.27	DENS65C	54.77	ECORG	12.04
AIRDIST	99.28	LHCPC	97.90	AVELF	64.72	IPRICE1	60.70	EUROPE	12.96
ORTH00	99.37	TOTIND	98.08	LANDLOCK	65.03	GOVSH61	62.76	SIZE60	12.98
LANDAREA	99.39	HINDU00	98.20	OTHFRAC	71.59	FERTLDC1	62.89	POP1560	14.02
GGCFD3	99.40	PRIGHTS	98.20	DENS60	72.35	LT100CR	63.72	GOVSH61	14.79
HERF00	99.40	PRIEXP70	98.35	GVR61	73.38	DPOP6090	64.18	ORTH00	15.77
PI6090	99.51	HERF00	98.50	POP1560	73.84	GVR61	66.57	REVCOUP	16.66
GEEREC1	99.52	POP1560	98.51	DPOP6090	73.90	BRIT	66.79	TROPPOP	16.89
LANDLOCK	99.53	IPRICE1	98.52	FERTLDC1	74.06	GGCFD3	68.31	LT100CR	17.26
GDE1	99.60	NEWSTATE	98.58	IPRICE1	74.31	LANDLOCK	70.12	DENS60	17.27
DENS60	99.60	POP60	98.71	GDE1	74.32	GDE1	74.43	CATH00	19.78
DPOP6090	99.62	TROPICAR	98.80	OPENDEC1	75.60	TOTIND	77.51	SOCIALIST	21.59
COLONY	99.68	PI6090	98.90	WARTORN	78.88	ECORG	77.52	OTHFRAC	22.01
POP1560	99.68	ORTH00	98.97	MUSLIM00	79.11	DENS60	78.57	BRIT	24.36
PRIEXP70	99.75	GEEREC1	99.09	TOTIND	79.78	TOT1DEC1	78.82	LANDAREA	25.02
TOTIND	99.76	GGCFD3	99.40	GOVSH61	82.46	MUSLIM00	78.88	GEEREC1	25.69
ABSLATIT	99.80	LT100CR	99.40	H60	83.95	OTHFRAC	79.29	PRIGHTS	26.64
OIL	99.82	AVELF	99.40	GGCFD3	84.48	HINDU00	81.60	LANDLOCK	30.19

LHCPC	99.84	WARTORN	99.47	DENS65C	85.40	WARTORN	82.08	TOT1DEC1	31.25
HINDU00	99.86	TROPPOP	99.47	DENS65I	88.29	GOVNOM1	83.36	TOTIND	31.40
PROT00	99.87	YRSOPEN	99.55	POP6560	91.18	H60	84.92	GDE1	34.58
SQPI6090	99.88	CIV72	99.58	PI6090	95.27	ORTH00	86.79	OPENDEC1	34.83
CATH00	99.89	DPOP6090	99.58	GDPCH60L	95.61	OPENDEC1	86.96	SCOUT	36.61
ECORG	99.90	LANDAREA	99.61	GOVNOM1	96.56	DENS65I	87.95	OIL	36.99
LAAM	99.92	WARTIME	99.64	CATH00	96.80	PI6090	93.07	WARTIME	37.17
SIZE60	99.93	LIFE060	99.65	HERF00	97.11	LANDAREA	97.10	ENGFRAC	39.10
LT100CR	99.93	GOVSH61	99.72	LHCPC	97.77	POP1560	97.15	AIRDIST	48.19
EUROPE	99.95	GOVNOM1	99.75	BRIT	98.05	OIL	98.00	ZTROPICS	50.50
LIFE060	99.96	LANDLOCK	99.82	GEEREC1	98.13	SQPI6090	98.57	ABSLATIT	58.54
POP60	99.96	SQPI6090	99.83	ORTH00	98.32	MINING	98.68	HERF00	60.34
ENGFRAC	99.97	ECORG	99.92	LANDAREA	98.33	POP6560	98.86	LHCPC	61.86
PRIGHTS	99.98	ABSLATIT	99.93	MINING	99.22	LHCPC	99.35	POP60	69.33
FERTLDC1	99.98	OTHFRAC	99.94	POP60	99.48	SIZE60	99.53	WARTORN	69.45
NEWSTATE	99.99	P60	99.95	EUROPE	99.50	SOCIALIST	99.55	DENS65I	73.23
CIV72	99.99	SAFRICA	99.96	SQPI6090	99.50	WARTIME	99.65	NEWSTATE	79.76
AVELF	99.99	SOCIALIST	99.99	WARTIME	99.63	GEEREC1	99.74	SQPI6090	90.91
H60	100.00	BRIT	99.99	LT100CR	99.81	HERF00	99.78	MUSLIM00	91.19
REVCOUP	100.00	REVCOUP	100.00	OIL	99.85	SCOUT	99.83	TROPICAR	93.02
SAFRICA	100.00	MUSLIM00	100.00	SOCIALIST	99.93	CATH00	99.94	GVR61	99.18
BRIT	100.00	FERTLDC1	100.00	TOT1DEC1	100.00	POP60	99.98	HINDU00	100.00

注：每种方法对应变量按照非线性性从小到大排序；变量右侧数字表示其非线性性程度，单位为百分比，依照正文(7)式的辅助回归进行测算

B.3. 前十变量预测非线性性图示

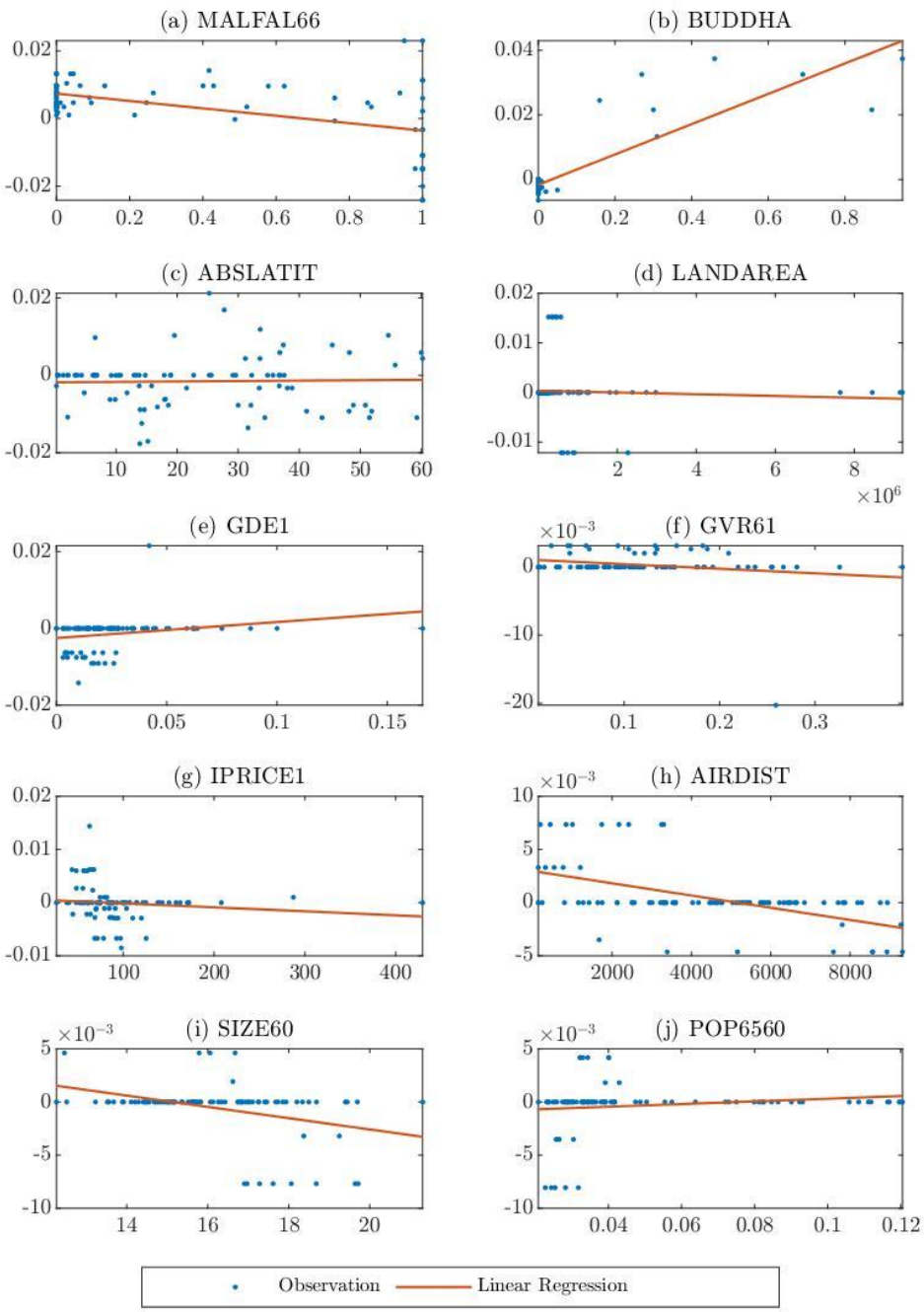


图 B1：LASSO 方法下变量预测非线性性及趋势特征

注：图中蓝色圆点代表该样本点处的预测值 $\hat{y}$

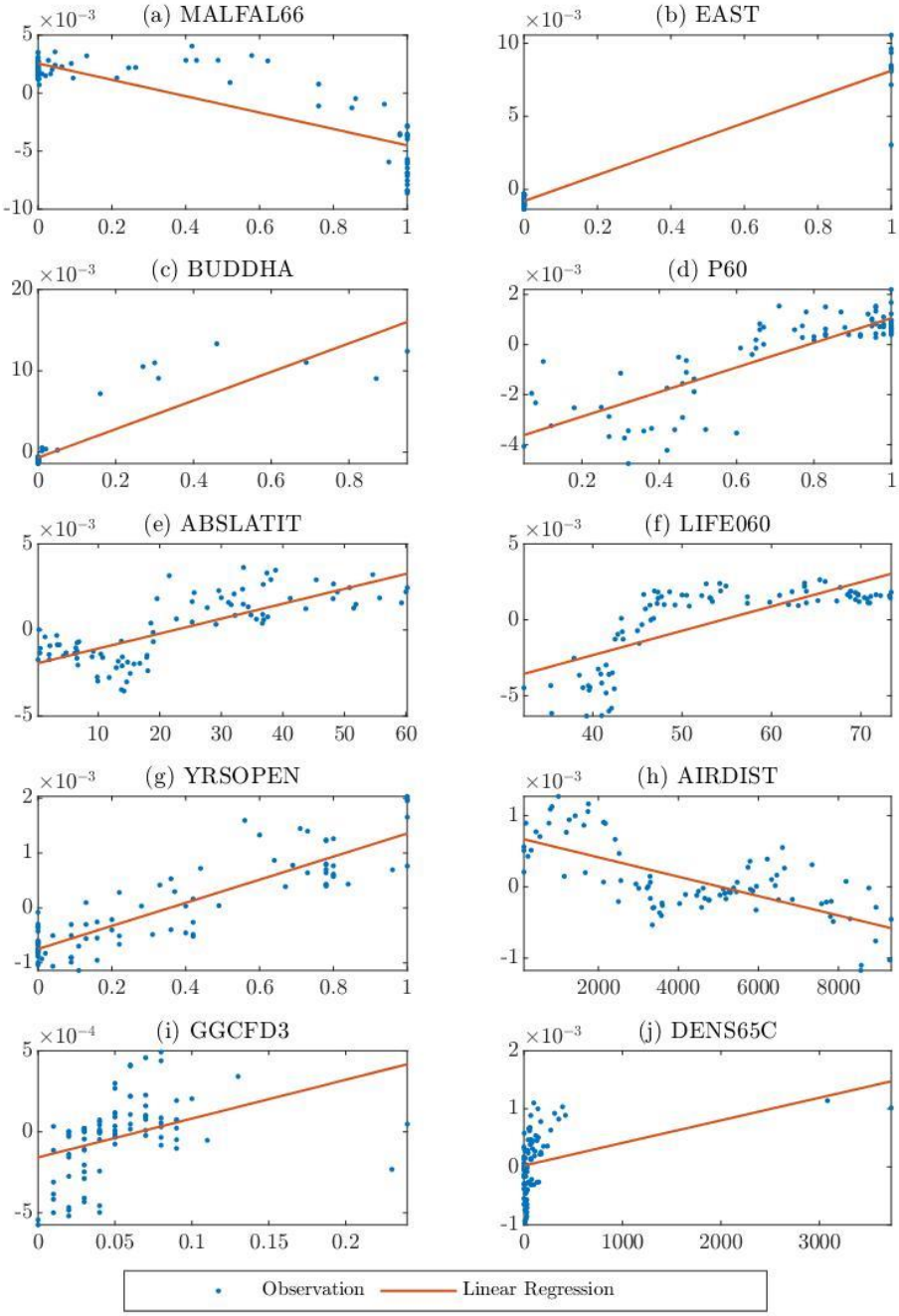


图 B2：回归树方法下变量预测非线性性及趋势特征

注：图中蓝色圆点代表该样本点处的预测值 $\hat{y}$

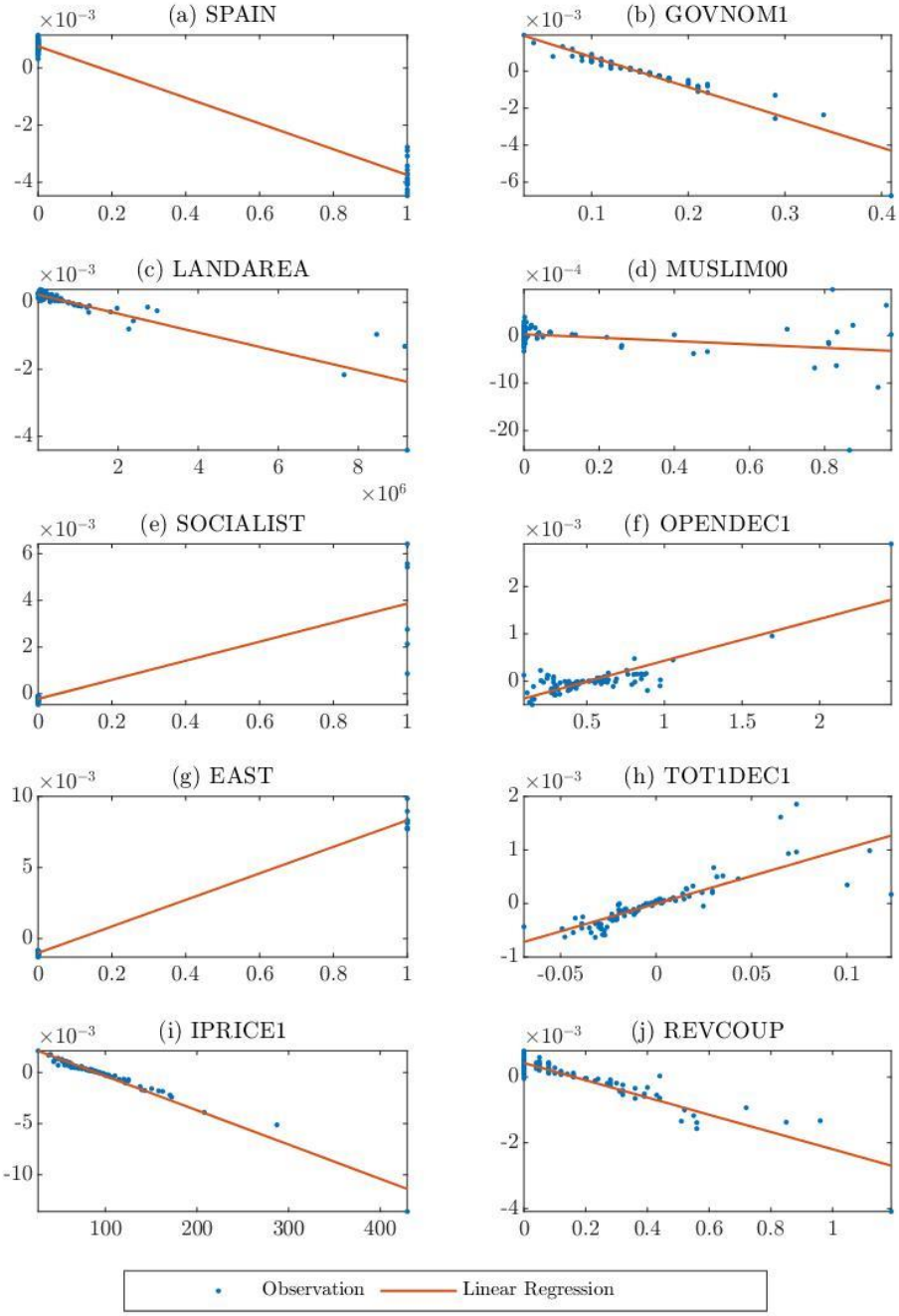


图 B3：神经网络方法下变量预测非线性性及趋势特征

注：图中蓝色圆点代表该样本点处的预测值 $\hat{y}$