# 评估精准医学中预测个体治疗效应的预测建模策略

Pamela Solano[a]、M Lee Van Horn[b]、Thomas Jaki[a, c]

[a]信息学与数据科学学院，雷根斯堡大学，德国[b]新墨西哥大学，美国[c]*MRC*生物统计学部，剑桥大学，英国

arXiv:2602.06210v1 [stat.AP] 5 Feb 2026

**摘要**

采用结构化模拟框架，揭示预测个体治疗效应时模型准确度、方向性和稳健性方面的新差异。

精准医学旨在为患者匹配能够带来最大获益的治疗。预测个体治疗效应（PITE）——治疗与对照条件下预测结局之差——量化了这种获益，但由于反事实不可观测、维度高且交互作用复杂，因此难以估计。我们采用结构化模拟框架，对比了 30+ 种建模策略，包括惩罚回归和基于投影的方法、灵活学习器及树集成，并改变样本量、维度、多重共线性和交互作用复杂度。我们使用均方根误差（RMSE）衡量预测准确度，使用方向准确率（DIR）衡量能否正确区分获益与伤害。内部验证得出的估计偏乐观，而在存在分布偏移和高阶交互的情况下，外部验证更清楚地揭示了模型的不足。惩罚回归和基于投影的方法——岭回归、套索回归、弹性网络、偏最小二乘法（PLS）和主成分回归（PCR）——始终表现出较强的RMSE和DIR性能。灵活学习器仅在信号强且样本量充足时表现出色。结果凸显了稳健的线性／投影默认方法的价值，以及严格开展外部验证的必要性。

关键词：精准医学，个体治疗效应，异质性治疗效应，基于模拟的评估，正则化回归，临床决策支持，高维数据，生物标志物，模型可解释性，预测模型稳定性

## 1. 引言

治疗效应的估计已不再局限于群体层面的汇总，而是逐渐转向个体化方法，这与精准医学的目标相一致。传统上，随机对照试验（RCT）侧重于平均治疗效应（ATE），但这一指标假设所有个体的治疗反应相同。然而在实践中，生物学、临床和社会人口学因素常常导致结局存在显著差异，从而造成有实际意义的治疗反应异质性 [32]。

为解决这一问题，预测个体治疗效应（PITEs）的概念日益受到关注，可用于指导患者特异性治疗推荐 [38, 8, 30]。与概括群体平均获益的平均治疗效应（ATE）不同，PITE估计值是在给定个体协变量的条件下，该个体接受治疗与对照时预测结局之间的预期差异。从平均效应估计转向个体化效应估计，是因果推断和临床决策领域的一项根本性进步 [1, 26]。

形式化地，在 PITE 框架下，令 $f_T(x_i)$ 表示基于个体 $i$ 的协变量向量 $x_i \in \mathbb{R}^p$ 的预测函数，并令 $T_i \in \{t, c\}$ 为二元指示变量，其中 $t$ 表示治疗组，$c$ 表示对照组。个体 $i$ 的 PITE 定义为:

$$\widehat{\mathrm{PITE}}_i = \hat{f}_t(x_i) - \hat{f}_c(x_i), \quad 1 \le i \le n,$$

^^其中，$f_t(\cdot)$ 和 $f_c(\cdot)$ 分别是针对治疗组和对照组单独拟合的预测模型。

在这一构造下，假设这些预测函数的前两个矩已定义，则 $\mathrm{PITE}_i$ 的方差为

$$\mathrm{Var}(\mathrm{PITE_i}) = \mathrm{Var}\left(\hat{f}_t(x_i)\right) + \mathrm{Var}\left(\hat{f}_c(x_i)\right)$$

^^其中协方差项为零，即 $\mathrm{Cov}(f_t, f_c) = 0$，因为两组的模型是独立估计的。

估计 PITE 可以采用多种方法，其唯一动机是使不相关的^^$f_t$ 和 $f_c$。然而，还会面临维度、异质性和非线性结局等外部挑战。惩罚回归 [70], 、基于树的集成方法 [6, 60], 和贝叶斯方法 [9, 22] 等数据自适应技术日益受到欢迎，尤其是在样本量有限且协变量结构复杂的早期临床试验和真实世界研究中。在 PITE 设定下，既有研究 [3]已强调了若干挑战。归根结底，如何界定PITE估计的良好性能高度依赖具体情境，不存在一种可被普遍视为最优的方法。

尽管方法学取得了显著进展，准确估计个体化治疗效果（PITEs）仍然困难，对于哪些方法或假设最可靠，目前尚未形成共识。一个关键难点在于，在结局预测方面表现出色的模型，未必能提供有效或稳定的治疗效应估计。最小化预测损失（例如观测结局的均方误差）并不能直接最小化治疗效应估计误差 [10]。如命题1所示，类似的推理同样适用于PITE估计。

一个常见但具有误导性的假设是：只要准确估计了治疗组和对照组的潜在结果，得到的PITE估计值也必然准确。尽管这一假设直观上似乎合理，但无论从代数角度还是实证角度来看，它都不成立。由于 PITE 被定义为两个预测值之差，两个模型各自的误差可能以难以预料的方式相互作用——放大噪声、引入偏差或扭曲校准。多项研究 [52, 69, 12] 表明

即使结果模型各自表现良好，推导出的处理效应也可能不稳定、存在偏差或校准不佳。重要的是，无法根据结局层面性能指标直接推断任何标准的 PITE 性能指标（例如均方根误差、平均绝对误差、决定系数、校准斜率）。

我们在一般的基于均方误差的框架中形式化了这一区分，并表明，即使模型对结局 $Y$ 的预测准确性很高，仍可能产生不稳定或具有误导性的个体化治疗效果（第 2.2 节）。

鉴于这些问题，本文并不旨在推崇某一种最佳方法。相反，我们为可靠估计个体化治疗效果提供了一份有原则且实用的指南，重点在于识别临床试验场景中对治疗效应建模时涉及的统计特性、局限性和权衡。我们着重探讨估计选择如何影响临床试验场景中的变异性、偏差、校准和决策相关性。

为实现这一目标，我们对五类模型中的 30 多种代表性方法进行了全面且系统的评估：（i）基于正则化的模型（如套索回归、弹性网络、贝叶斯收缩）；（ii）基于树的集成学习器（如随机森林、梯度提升、BART [22]）；（iii）贝叶斯学习方法（如尖峰与板先验、贝叶斯神经网络）；（iv）传统参数模型（如含交互项的广义线性模型）；（v）非线性和基于核的方法（如神经网络、核正则化最小二乘、投影寻踪回归）。

尽管人们对个体化治疗效应推断的热情日益高涨，但关于如何可靠地估计和验证PITE，仍有一些基础科学问题尚未解决。正则化、投影或非参数灵活性等建模选择，如何与样本量、维度和治疗效应异质性相互作用，从而影响所得估计值的偏差、方差和校准？在什么样的数据条件下，包括树集成和贝叶斯加性模型在内的灵活学习器能够真正优于传统的惩罚回归或基于投影的方法；又在何时会面临过拟合或不稳定的风险？同样重要的是，当估计对象本身（即反事实对比）无法观测时，应如何量化预测性能；现有指标（如均方根误差和方向准确率）在多大程度上能够反映具有临床意义的可靠性?

本文通过在 PITE框架下对三十多种当代建模策略进行系统的、基于模拟的评估，探讨了这些问题。我们的目标不仅是根据性能对方法进行排名，更是揭示决定某些模型在表示个体化治疗效应时何以成功或失败的规律。通过明确建立模型适应性与患者层面复杂性之间的联系，我们旨在厘清患者间预测处理效应的差异在何种条件下能够转化为可指导临床实践的洞见。通过明确建立模型适应性与患者层面复杂性之间的联系，我们旨在厘清患者间预测处理效应的差异在何种条件下能够转化为可指导临床实践的洞见。这些问题构成了更广泛研究议程的基础：开发稳健、可解释[1] 且

[1]可解释性是指模型统计结构的透明度——包括其参数化方式、收缩行为以及方差–偏差权衡特征——而不是对所预测的个体治疗

ge可推广的工具，用于在精准医学中预测个体治疗效应 .

为回答这些问题，我们设计了一个全面且严谨的模拟框架，以反映真实世界临床数据的复杂性。评估涵盖了多种情景，包括治疗效应异质性程度各异、高维且相关的协变量、样本量有限以及有意设置的模型设定错误——在这些条件下，传统估计量往往表现不佳。我们不仅考察各建模策略的预测准确性，也评估其推断稳定性和临床可解释性。我们采用互补指标量化性能，以捕捉可靠性的不同维度：PITE的期望均方误差（MSE）用于衡量总体估计质量，校准和调整后的 $R^2$ 用于评估模型拟合，平均绝对误差（MAE）用于衡量对离群值的稳健性，方向准确性准则则用于评估模型能否正确区分治疗获益与伤害，而不受效应幅度影响。

我们认识到，反事实无法直接验证，因此也强调稳健性能评估技术的重要性。这些技术包括基于匹配的验证框架 [16, 34] 以及用于评估校准和区分度的指标 [59, 26], ，它们对于监管、临床和外部可迁移性评估至关重要 [55]。此外，我们还利用多重插补 [38], 、贝叶斯数据增广和正则化技术等工具，处理缺失数据、离群值和复杂生物标志物交互作用等常见的真实世界复杂问题；这些工具有助于在高维场景下稳定PITE估计值 [3]。

本研究主要在三个方面作出了贡献。首先，对用于估计个体化治疗效果（PITEs）的多种估计策略进行分类并加以说明，将其与各自的基本假设和适用场景联系起来。其次，深入揭示了这些估计策略在不同数据生成条件下的表现。第三，提供了实用建议和注意事项，包括强调何时应采用或避免采用某些方法的原则，并指出以结局为导向的损失函数（如均方误差（MSE））可能导致无效或不稳定的治疗效应估计值。

## 2. 方法学

### *2.1.* 评估方法

我们评估了一系列广泛且具有代表性的模型，用于估计预测的个体治疗效应（PITEs），涵盖经典统计方法、现代机器学习以及贝叶斯方法。线性模型与正则化线性模型——包括普通最小二乘法 [53],广义线性模型（GLMs） [44], 套索回归 [57], 岭回归 [24], 弹性网络 [70], 以及主成分回归 [31]——提供了成熟可靠的基础，尤其因其可解释性和计算效率而受到重视。这些模型在低到中等维度下较为有效，但在处理非线性和复杂特征交互作用时可能力不从心。为应对多重共线性和高维度问题，我们纳入了惩罚回归和基于投影的方法，例如偏最小二乘法 [67], 逐步选择 [11], 以及通过非负最小二乘法施加的非负约束

效应本身。

[39]。这些技术能够在保持可解释性的同时实现简约建模与特征选择。

为了对非线性和交互丰富的关系进行建模，我们纳入了基于树的方法与集成学习器（例如，决策树、随机森林 [6], 提升树 [14], 分位数森林 [46]），它们对高维数据尤为有效，且无需严格的参数假设。贝叶斯变体，如贝叶斯加性回归树（BART） [9] 和尖峰-平板回归 [7] ，能够实现不确定性量化与自适应正则化；而神经网络——包括贝叶斯神经网络 [43]——能够捕捉复杂模式并提供稳健的预测性能，但计算成本更高。我们还考虑了灵活的非参数与基于核的方法，如核岭回归 [58] 和投影寻踪回归[15], ，以及约束模型与潜变量模型，如分位数回归 [33] 和监督主成分分析 [2]。这些模型在捕捉分布细微差异和处理具有挑战性的数据结构方面具有优势。总体而言，这一建模谱系体现了灵活性、可解释性与计算可行性之间的平衡，适用于PITE估计中遇到的多样化临床数据情景。这些方法名称的进一步细节与缩写见补充材料附录A。

在定义 PITE 评估指标之前，我们首先从解释不确定性和可解释性的角度对其进行考察。该分析凸显了**均方根误差** 和**方向**的互补作用；二者是我们主要分析中采用的诊断指标。

### *2.2.* 诊断性预测性能

我们使用两个互补指标评估*PITE*估计值 的预测性能，它们分别反映准确性和临床效用的不同方面。

第一个指标是**均方根误差（RMSE）**，它概括了预测的个体治疗效应与真实（未观测）个体治疗效应之间的总体差异：

$$\mathrm{RMSE} = \sqrt{\mathbb{E}\left[(\widehat{PITE}(x_i) - PITE(x_i))^2\right]}. \tag{1}$$

均方根误差反映了 PITE 估计值中的方差和偏差，可作为衡量估计准确性的全局指标。

第二个指标称为**方向**，用于量化针对每个个体正确识别治疗效应符号的能力：

$$\mathrm{Direction} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\Big(\mathrm{sign}(\widehat{PITE}(x_i)) = \mathrm{sign}(PITE(x_i))\Big). \tag{2}$$

其中， $\mathbb{I}(\cdot)$ 表示指示函数。该指标衡量的是效应方向的一致性，而非效应大小；在精准医学等决策场景中，这一点尤为重要，因为判断治疗是有益还是有害，可能比精确估计效应更重要。

虽然这两个指标都很直观，但要理解其表现，最好对 PITE 误差进行形式化分解。具体而言，PITE 的均方误差可以用处理组与对照组结果模型对应的均方误差表示：

**命题1.** 假设 $f_t(\cdot)$ 和 $f_c(\cdot)$ 是期望值和方差均可良好定义的预测变量。则*PITE*均方误差（均方误差$_{\text{PITE}}$）可分解为

$$\text{MSE}_{\text{PITE}} = \text{MSE}_{\text{t}} + \text{MSE}_{\text{c}} - 2\,\text{bias}_t \cdot \text{bias}_c, \tag{3}$$

其中均方误差$_{\text{t}} = Var[\hat{f}_t(X)] +$ 偏差$_t^2$和均方误差$_{\text{c}} = Var[\hat{f}_c(X)] +$ 偏差$_c^2$分别表示治疗模型和对照模型的均方误差，其中偏差$_t = \mathbb{E}[f_t(X)] - \mathbb{E}[Y(t)]$ 针对治疗，偏差$_c = \mathbb{E}[f_c(X)] - \mathbb{E}[Y(c)]$ 针对对照。

该命题将如下直觉形式化：最小化结果模型$\hat{f}_t$ 和 $f_c$ 的预测损失并不必然最小化PITE估计中的误差。该分解突出表明PITE准确度不仅取决于 $f_t$ 和 $f_c$各自的预测性能，还取决于其偏差之间的相互作用。特别是，方向相反的偏差可能叠加并放大PITE估计值中的误差，从而凸显出需要能够考虑两个结果模型联合行为的诊断。

为求完整，我们还评估了其他性能指标，以补充主要误差度量。令 PITE$_i$ 表示受试者 $i$ 的真实预测个体治疗效应，$\hat{}$PITE$_i$ 表示使用给定模型得到的相应估计值。

**平均绝对误差（MAE）**概括了估计的个体化治疗效果与真实个体化治疗效果之间的平均绝对偏差：

$$\text{MAE}_{\text{PITE}} = \frac{1}{n}\sum_{i=1}^{n} \left|\hat{\text{PITE}}_i - \text{PITE}_i\right|.$$

**决定系数**（$R^2$）量化了模型估计值对真实 PITE 值变异的解释比例：

$$R^2_{\text{PITE}} = 1 - \frac{\sum_{i=1}^{n}(\text{PITE}_i - \hat{\text{PITE}}_i)^2}{\sum_{i=1}^{n}(\text{PITE}_i - \bar{\text{PITE}})^2},$$

其中，PITE 是真实 PITE 值的样本均值。

最后，我们通过对真实效应与估计效应进行简单线性回归来评估**校准**：

$$\text{PITE}_i = \alpha + \beta\,\hat{\text{PITE}}_i + \eta_i,$$

其中，$\alpha$ 和 $\beta$ 分别表示截距和斜率。完美校准对应于 $\alpha = 0$和 $\beta = 1$，这表明估计效应在水平和尺度上均无偏。与这些值的偏离反映了系统偏差（$\alpha$ 非零）或效应幅度校准不良（斜率 $\beta \neq 1$）。

这些指标的详细定义和分解见补充材料

附录B部分还提供了实证依据，说明为何我们将均方根误差和方向作为主要评估标准，而非其他指标。

## 3. 模拟研究

### *3.1.* 模拟设计

我们生成合成数据，以在受控条件下评估所提出的 PITE框架的性能。对于每次重复实验，我们模拟 $N$ 名独立受试者，其协变量向量为 $p$维， $X_i = (X_{i1}, \ldots\ldots, X_{ip})^\top$ ，这些向量来自多元正态分布：

$$X_i \sim \mathcal{N}_p(0, \Sigma), \qquad i = 1, \ldots, N,$$

其中， $\Sigma$ 是一个 $p \times p$ 协方差矩阵，用于规定协变量之间的相关结构。除非另有说明，为简单起见，将 $\Sigma$ 设为 $\rho I_p$ ，其中 $\rho$控制相关性大小。

治疗分配与协变量相互独立，并将一半受试者分配至治疗组，其中$T_i = 1$ 表示分配至治疗组，$T_i = 0$ 表示分配至对照组。

我们规定对照和治疗条件下的潜在结果如下。对于每名受试者，对照组平均响应为

$$f_0(X_i) = X_i^\top \beta_0,$$

其中， $\beta_0 \in \mathbb{R}^p$是一个从 $\mathcal{N}(0, 0.1_2)$中抽取的基线系数向量。

个体治疗效应建模如下

$$\Delta(X_i) = X_i^\top \beta_\Delta,$$

其中， $\beta_\Delta \in \mathbb{R}^p$ 是一个治疗效应系数向量，用于确定协变量如何调节个体治疗获益。在我们的模拟中， $\beta_\Delta$ 从以下分布中抽取

$$\beta_\Delta \sim \mathcal{N}(\mu_{\beta_\Delta}, 0.01^2 I_p),$$

其中，$\mu_{\beta_\Delta}$ 表示治疗效应系数的均值。较小的方差 $0.01^2$ 只会引入轻微的异质性，从而确保PITE估计值的差异主要由平均治疗效应的变化驱动，而非个体间的高变异性。这种构造使预期个体获益能够围绕目标效应保持适度且对称的扰动。

观测到的结局由潜在结果框架生成，并带有加性高斯噪声：

$$Y_i(c) = f_c(X_i) + \varepsilon_{ci}, \qquad Y_i(t) = Y_i(c) + \Delta(X_i),$$

其中 $\varepsilon_{ci} \sim \mathcal{N}(0,1)$。最后，观测到的响应变量为

$$Y_i^{\text{obs}} = Y_i(c) + T_i \cdot \Delta(X_i).$$

因此，模拟数据集包括协变量（$X_i$）、治疗指示变量（$T_i$）、观测结局（$Y_i^{\text{obs}}$）以及真实的个体期望治疗效应 $\Delta(X_i)$，后者作为评估的真实值。模拟通过 27 个计算线程并行运行，重复 135 次。有关软件和模型实现的详细信息，请参阅补充材料的附录 C。具体算法详见算法1。

### *3.2.* 内部验证

从具有指定相关结构的多元正态分布中抽取协变量，以表示不同程度的多重共线性：低（$\rho = 0$）、中等（$\rho = 0.5$）和高（$\rho = 0.95$）。治疗分配独立于协变量，半数受试者被分配至治疗组，另一半被分配至对照组。结局由协变量的线性函数加上高斯噪声生成。

在不同模拟情景中，我们改变了治疗效应系数的均值（$\mu_{\beta_\Delta} \in \{0, 0.25, 0.5\}$）、协变量的数量（$p \in \{5, 15, 45\}$）以及样本量（$n \in \{250, 500, 750\}$）。

在同一总体中，我们按照 Rolling 和 Yang [52] 的方法，将样本按 50:50 的比例拆分，生成训练集和测试集。随后，使用训练集训练预测模型，并在测试集上进行评估。

在具有相关协变量、不同样本量和一系列治疗效应的内部验证中，与存在交互作用的外部验证相比，模型表现总体较强且更加稳定。均方根误差（RMSE）的均值约为 $= 0.88$，中位数约为 $= 0.51$，取值范围为 (0.03, 22)；总体而言，方向准确率（DIR）较高（均值 $= 0.76$，中位数 $= 0.84$），其中若干模型的方向准确率接近完美（DIR $> 0.95$）。

图 1 显示，各模型的表现存在明显差异，衡量指标为 RMSE 和 DIR。若干模型，尤其是 krlsPoly 和 rpart2，反复落入灰色危险区域，表明其 RMSE 较高且 DIR 较低。在高维协变量（$p = 45$）以及治疗效应较低或为零的情景中，这种模式尤为明显：预测准确性下降，方向一致性未能达到可接受的阈值（DIR $< 0.6$）。相反，表现良好的模型则未落入这些危险区域，在估计的个体治疗效应与真实值之间实现了更好的一致性。

即使在相关协变量的内部验证中，各方法在估计 PITE 时也表现出较大的预测误差和较差的方向可靠性。值得注意的是，krlsPoly 和 rpart2 经常名列表现最差的方法，RMSE 最高达 22.0，MAE 超过 17.5，最大决定系数达到 79.7%，覆盖率较差 $\alpha$ & $\beta$，且方向准确率始终接近或低于 0.55。这些失效在高维场景（$p = 45$）中最为明显，尤其是在中度至高度相关的情况下（$\rho = 0.5$˘$0.95$）；这一现象出现在所有处理效应水平下，见表1。

**算法1** PITE评估的模拟–验证–匹配流程

**输入：** 模拟参数 $\{n, p, \text{ete}\}$；预测模型列表 $\mathcal{M} = \{m_1, \ldots, m_K\}$；数据生成器 sim_data()；交叉验证折数数量 $= 10$。

**输出：** $\widehat{\Delta}_{ij}$，观察到的个体化治疗效果 $O_{ij}$，以及训练和验证人群中的真实获益。

**步骤 1：数据生成（训练数据集）**

使用 sim_data($p, n,$ ete) 模拟训练数据集 $\mathcal{D}_{\text{train}} = \{(\boldsymbol{X}_i, T_i, Y_i)\}_{i=1}^{n}$。将 $\mathcal{D}_{\text{train}}$ 划分为两个治疗子集：

$$\mathcal{D}_0 = \{(\boldsymbol{X}_i, Y_i) : T_i = 0\} \text{ and } \mathcal{D}_1 = \{(\boldsymbol{X}_i, Y_i) : T_i = 1\}$$

**步骤 2：外部数据集（生成）**

使用相同的数据生成机制但采用不同的随机种子，独立生成一个新数据集 $\mathcal{D}_{\text{val}} = \{(\boldsymbol{X}_j^*, T_j^*, Y_j^*)\}_{j=1}^{n}$，从而生成一个独立的外部总体；该总体具有相同的结构参数，但协变量、$\mu_{\beta_\Delta}$、治疗分配和结局的具体取值不同。

将 $\mathcal{D}_{\text{val}}$ 划分为治疗组（$T^* = 1$）和对照组（$T^* = 0$）。

**步骤 3 ：用于预测的外部验证数据集匹配程序**

通过 MatchIt，使用马氏距离对 $\mathcal{D}_{\text{val}}$ 中的治疗组和对照组受试者进行最近邻匹配。获取与匹配对相对应的索引集合（$\mathcal{I}_A$，$\mathcal{I}_B$）：

$$\mathcal{I}_A = \{i : T_i^* = 0\} \text{ and } \mathcal{I}_B = \{j : T_j^* = 1\}$$

这些索引用于计算匹配受试者之间成对的观测治疗效果差异和预测治疗效果差异。

**步骤 4：模型训练与预测**

每个模型 $m_k \in \mathcal{M}$ 分别在 $\mathcal{D}_0$ 和 $\mathcal{D}_1$ 上进行训练，以分别预测对照和治疗条件下的潜在结果。

$$\widehat{f}_0^{(k)} = m_k(Y \sim \boldsymbol{X}, \mathcal{D}_0), \quad \widehat{f}_1^{(k)} = m_k(Y \sim \boldsymbol{X}, \mathcal{D}_1).$$

对模型使用 10 折交叉验证。

按照**步骤 3**，预测每个验证对象的潜在结果：

$$\widehat{Y}_{0j}^{(k)} = \widehat{f}_0^{(k)}(\boldsymbol{X}_j^*), \quad \widehat{Y}_{1j}^{(k)} = \widehat{f}_1^{(k)}(\boldsymbol{X}_j^*).$$

计算匹配对的估计个体化治疗效果：

$$\widehat{\Delta}_{ij}^{(k)} = \widehat{Y}_{1,\mathcal{I}_B}^{(k)} - \widehat{Y}_{0,\mathcal{I}_A}^{(k)}.$$

计算观察到的个体化治疗效果：

$$O_{ij} = Y_{\mathcal{I}_B}^* - Y_{\mathcal{I}_A}^*.$$

存储 $\widehat{\Delta}_{ij}^{(k)}$ 和 $O_{ij}$ 以供后续评估。

**备注：**

- 训练和验证在独立模拟总体上进行（该步骤仅用于外部验证）。
- 匹配仅应用于验证数据集，以形成可比的治疗组–对照组配对用于评估。

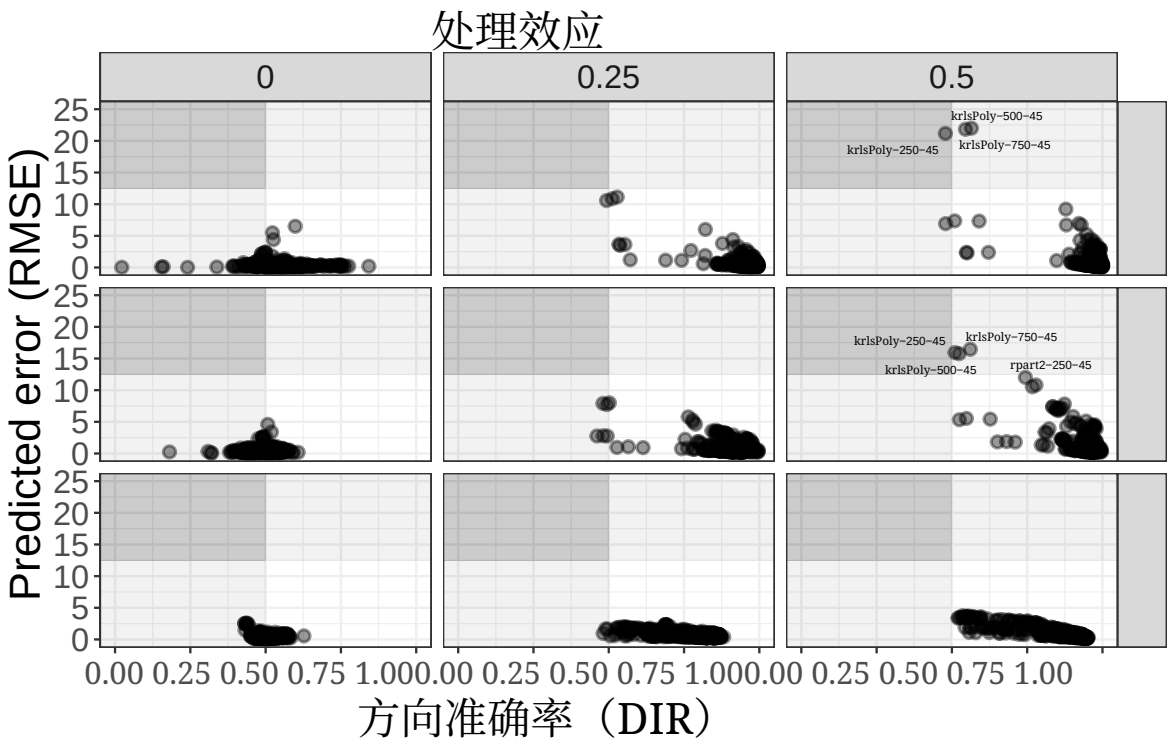


图 1：内部验证：各模型的预测性能均方根误差（RMSE）与方向准确率（DIR）之间的关系，以及处理效应影响与协变量之间的相关值。每个点代表一个特定的模型参数配置，并标注了模型名称、样本量（n）和预测变量数量（p）。深灰色区域表示不理想区域

| $\mu_{\beta_\Delta}$ | $\rho$ | 模型 | p | n | RMSE | $R^2$ | MAE | DIR | $\alpha_0$ | $\alpha_1$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.95 | Bayenet | 45 | 250 | 6.507 | 0.148 | 6.233 | 0.598 | 0.400 | 0.200 |
| 0.00 | 0.95 | Bayenet | 45 | 500 | 5.492 | 0.384 | 5.314 | 0.523 | 0.000 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 250 | 8.002 | 0.706 | 6.367 | 0.502 | 0.000 | 0.000 |
| 0.25 | 0.50 | rpart2 | 45 | 250 | 5.784 | 0.494 | 4.571 | 0.764 | 1.000 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 500 | 7.904 | 0.681 | 6.286 | 0.481 | 0.000 | 0.000 |
| 0.25 | 0.50 | rpart2 | 45 | 500 | 5.304 | 0.565 | 4.147 | 0.778 | 0.400 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 750 | 7.729 | 0.735 | 6.227 | 0.494 | 0.000 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 250 | 11.134 | 0.341 | 8.945 | 0.528 | 0.000 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 500 | 10.574 | 0.638 | 8.523 | 0.493 | 0.000 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 750 | 10.861 | 0.611 | 8.545 | 0.511 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 250 | 5.327 | 0.616 | 4.284 | 0.525 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 500 | 5.537 | 0.653 | 4.422 | 0.548 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 750 | 5.423 | 0.674 | 4.328 | 0.627 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 250 | 15.935 | 0.759 | 12.688 | 0.510 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 250 | 12.009 | 0.498 | 9.797 | 0.744 | 0.800 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 500 | 15.762 | 0.801 | 12.466 | 0.524 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 500 | 10.535 | 0.568 | 8.431 | 0.767 | 1.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 750 | 16.459 | 0.797 | 13.197 | 0.561 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 750 | 10.822 | 0.577 | 8.613 | 0.779 | 0.600 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 15 | 250 | 7.327 | 0.541 | 5.832 | 0.590 | 0.000 | 0.000 |

内部验证下不同模拟情景中的成功模型表现。图2表明，在具有相关协变量的情景中进行内部验证时，模型表现虽令人安心，但也呈现出一定规律。在表现出色的情景中（RMSE $< 1$ 和 DIR $> 0.95$）：简单方法、惩罚回归和基于投影的线性方法，凭借较低的 RMSE 和较高的 DIR，成为表现最稳定的方法。广义线性模型系列（广义线性模型、贝叶斯广义线性模型）、惩罚回归（套索回归、岭回归、弹性网络、L1/L2glmnet/L2glmnet）、逐步/选择变体（基于AIC的广义线性模型逐步选择、lmStepAIC、leapForward/Backward/Seq）以及低维投影方法（偏最小二乘、主成分回归、非负最小二乘、惩罚回归）在各种条件下反复表现出色，涵盖从小样本、协变量相关性较弱到大样本、协变量相关性较强的广泛情形（代表性条件索引： $[1, 6, 12, 18,24, 30, 36]$）。

相比之下，更灵活的非参数树模型/集成方法（BART/SoftBart、随机森林系列、QRF、Cubist，以及 blackboost/bstTree 等提升树）仅在较狭窄的条件下表现良好——通常要求样本量、预期治疗效应信号和/或协变量相关性有利。例如，BART 似乎在高相关性、中等信号的设定下表现良好（[7,19] 条件——即 $\rho = 0.95$ 预期治疗效应中等），而随机森林系列（rf、rangerRF、qrf）则主要在高相关性分层中出现在成功方法之列（例如 $[7, 19, 31]$）。

### *3.3.* 外部验证

第二种模拟情景用于评估在一个总体中训练的模型应用于外部总体时的泛化程度——这是实际应用中一项关键却常被忽视的挑战。模拟设计遵循Kuhlemeier等人 [34] 所述的流程，并在算法1中详细说明。协变量结构相关的模拟参数，包括 $p$、 $n$、 $\rho$和$\mu_{\beta_\Delta}$，均按照前一种模拟情景进行变动。

均方根误差（RMSE）最小为 1.34，最大为 22.23，中位数为 1.56。方向（DIR）指标介于 0.16 至 0.98 之间（均值为 $= 0.69$，中位数为 $= 0.70$），表明在大多数情况下，方向预测表现良好且具有临床可解释性。

#### *3.3.1.* 相关协变量

图3展示了不同模型配置、治疗效应大小和协变量相关结构下预测误差（RMSE）与方向准确率（DIR）之间的权衡。随着协变量间相关性增强，预测性能总体有所下降（$\rho$）；在 $\rho = 0.95$ 条件下，模型的 RMSE 较高而 DIR 较低。治疗效应系数的均值从 0 增至 0.5 时，信号增强使 DIR 提高，但在某些情况下 RMSE 也有所上升，这反映出精确估计的难度增加。样本量较大且维度适中的模型（例如 $\mathrm{n} = 750$、$\mathrm{p} = 45$）始终表现较好（RMSE 低、DIR 高），尤其是在协变量相关性较弱时。相比之下，高维度且样本量较小的模型往往无法准确识别个体层面的治疗效应。具体而言，采用多项式基展开的核正则化最小二乘（krlsPoly）在协变量间相关性增强时准确率下降，其中 $\rho$

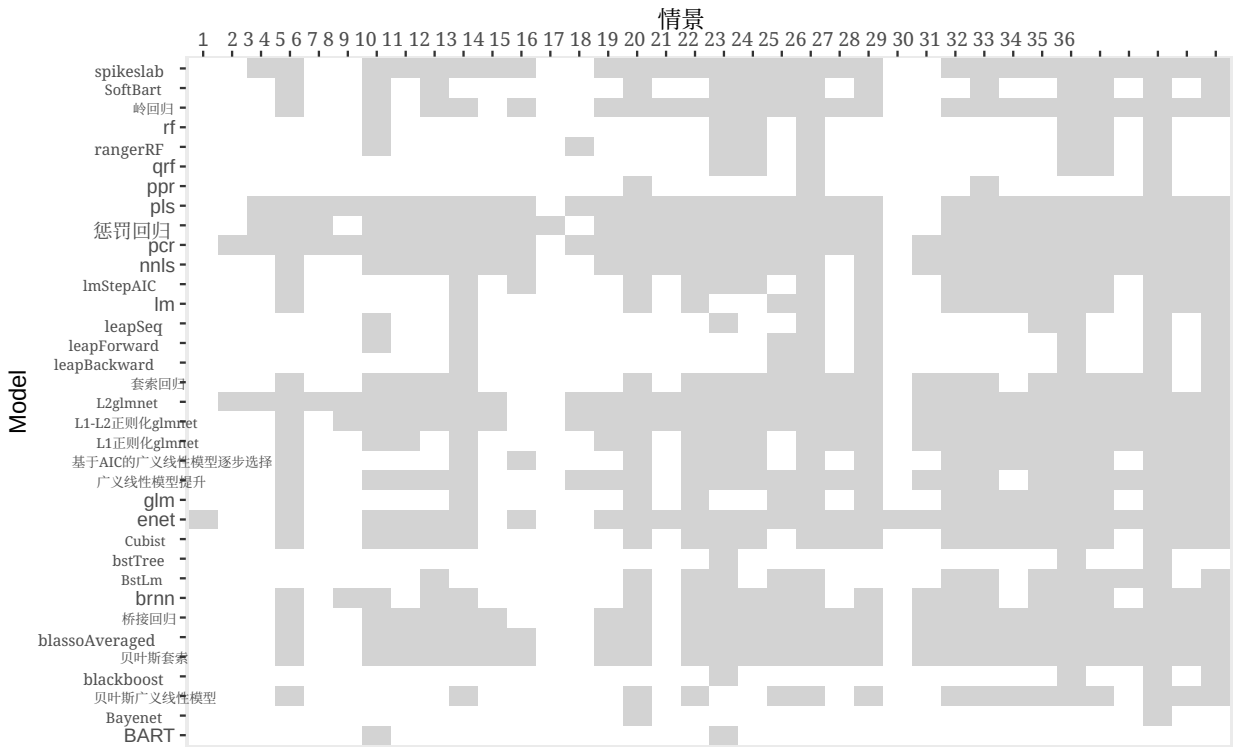


图2：内部验证：在具有协变量交互的外部验证情景中实现高PITE准确度（均方根误差 $< 1$ 和方向准确率 $> 0.95$）的模型

设定$n$ = 样本量； $\rho$ = 协变量相关性； $\mu_{\beta_\Delta}$ = 治疗效应系数的均值； $p$ =协变量数量。

| ID | $n$ | $\rho$ | $\mu_{\beta_\Delta}$ | $p$ | ID | $n$ | $\rho$ | $\mu_{\beta_\Delta}$ | $p$ | ID | $n$ | $\rho$ | $\mu_{\beta_\Delta}$ | $p$ | ID | $n$ | $\rho$ | $\mu_{\beta_\Delta}$ | $p$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 250 | 0 | 0.50 | 5 | 10 | 250 | 0.95 | 0.50 | 15 | 19 | 500 | 0.95 | 0.25 | 15 | 28 | 750 | 0.5 | 0.50 | 15 |
| 2 | 250 | 0.5 | 0.25 | 15 | 11 | 250 | 0.95 | 0.50 | 45 | 20 | 500 | 0.95 | 0.25 | 45 | 29 | 750 | 0.5 | 0.50 | 45 |
| 3 | 250 | 0.5 | 0.25 | 45 | 12 | 250 | 0.95 | 0.50 | 5 | 21 | 500 | 0.95 | 0.25 | 5 | 30 | 750 | 0.5 | 0.50 | 5 |
| 4 | 250 | 0.5 | 0.50 | 15 | 13 | 500 | 0 | 0.50 | 5 | 22 | 500 | 0.95 | 0.50 | 15 | 31 | 750 | 0.95 | 0.25 | 15 |
| 5 | 250 | 0.5 | 0.50 | 45 | 14 | 500 | 0.5 | 0.25 | 15 | 23 | 500 | 0.95 | 0.50 | 45 | 32 | 750 | 0.95 | 0.25 | 45 |
| 6 | 250 | 0.5 | 0.50 | 5 | 15 | 500 | 0.5 | 0.25 | 45 | 24 | 500 | 0.95 | 0.50 | 5 | 33 | 750 | 0.95 | 0.25 | 5 |
| 7 | 250 | 0.95 | 0.25 | 15 | 16 | 500 | 0.5 | 0.50 | 15 | 25 | 750 | 0 | 0.50 | 15 | 34 | 750 | 0.95 | 0.50 | 15 |
| 8 | 250 | 0.95 | 0.25 | 45 | 17 | 500 | 0.5 | 0.50 | 45 | 26 | 750 | 0.5 | 0.25 | 15 | 35 | 750 | 0.95 | 0.50 | 45 |
| 9 | 250 | 0.95 | 0.25 | 5 | 18 | 500 | 0.5 | 0.50 | 5 | 27 | 750 | 0.5 | 0.25 | 45 | 36 | 750 | 0.95 | 0.50 | 5 |

$= 0$0.95，导致均方根误差升高且方向准确率降低。即使在样本 $n = 750$下，高维配置（$p \in c$(15，45)）在高处理效应水平下也经常表现不佳。详见表2。

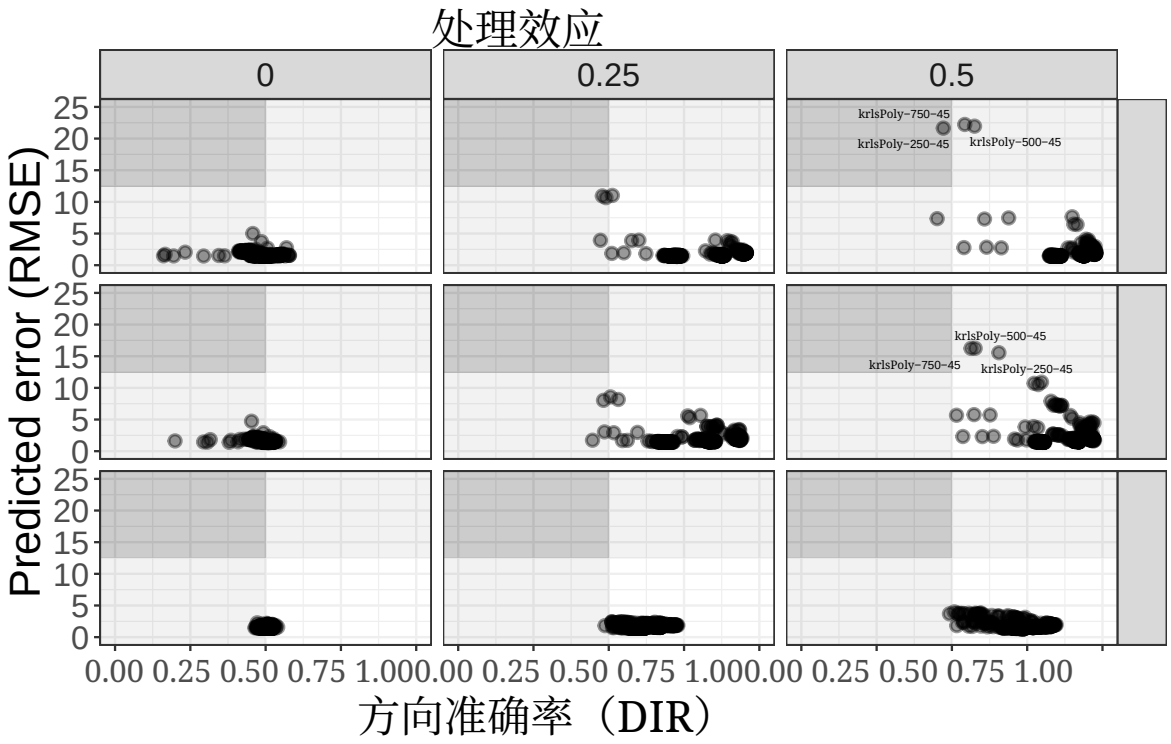


图3：外部验证（相关协变量）：展示不同模型下预测性能均方根误差（RMSE）与方向准确率（DIR）之间的关系，以及处理效应影响和协变量之间相关值的关系。每个点代表一个特定的模型参数配置，并标注了模型名称、样本量（n）和预测变量数量（p）。较深的灰色区域表示不理想区域

不同外部验证情景下表现最佳的模型。 图4总结了在不同模拟条件下表现最佳的模型，这些模型能够在最小化PITE均方根误差的同时保持方向准确率——即正确识别治疗对每位个体是有益还是有害。

总体而言，PLS 和 PCR 在全部九种模拟条件下均表现出稳健的性能。这些模型似乎尤其适用于中到高度多重共线性（$\rho = 0.5$–0.95）且样本量为 $n \in \{500, 750\}$ 的设定；在包含 $n = 250$ 的三种情景中，性能退化也较为有限。对于不同水平的平均治疗效应系数 $\mu_{\beta_\Delta}$，它们均保持了较低的均方根误差值和较强的方向一致性。相比之下，套索回归、投影寻踪回归、SoftBart 和广义线性模型提升等更复杂的模型仅在一到两种条件下表现良好，这表明它们的性能可能对数据生成过程的具体特征更为敏感。岭回归、spikeslab、桥接回归和贝叶斯套索等模型表现出中等稳健性，在九种设定中约有七种表现良好。综合来看，这些

| $\mu_{\beta_\Delta}$ | $\rho$ | 模型 | p | n 均方根误差 | | $R^2$ | MAE | DIR | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.95 | Bayenet | 45 | 500 | 5.004 | 0.190 | 4.579 | 0.457 | 0.200 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 250 | 8.580 | 0.620 | 7.003 | 0.506 | 0.000 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 500 | 8.165 | 0.635 | 6.469 | 0.532 | 0.000 | 0.000 |
| 0.25 | 0.50 | rpart2 | 45 | 500 | 5.579 | 0.544 | 4.443 | 0.762 | 0.800 | 0.000 |
| 0.25 | 0.50 | krlsPoly | 45 | 750 | 8.028 | 0.659 | 6.382 | 0.483 | 0.000 | 0.000 |
| 0.25 | 0.50 | rpart2 | 45 | 750 | 5.288 | 0.575 | 4.224 | 0.769 | 0.800 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 250 | 11.027 | 0.300 | 8.659 | 0.512 | 0.000 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 500 | 10.671 | 0.457 | 8.595 | 0.491 | 0.000 | 0.000 |
| 0.25 | 0.95 | krlsPoly | 45 | 750 | 10.943 | 0.527 | 8.717 | 0.479 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 250 | 5.676 | 0.452 | 4.555 | 0.515 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 500 | 5.723 | 0.474 | 4.553 | 0.627 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 15 | 750 | 5.773 | 0.497 | 4.603 | 0.574 | 0.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 250 | 15.548 | 0.787 | 12.345 | 0.656 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 250 | 10.522 | 0.574 | 8.483 | 0.787 | 1.000 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 500 | 16.248 | 0.812 | 13.088 | 0.577 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 500 | 10.902 | 0.568 | 8.620 | 0.798 | 0.800 | 0.000 |
| 0.50 | 0.50 | krlsPoly | 45 | 750 | 16.236 | 0.760 | 12.905 | 0.563 | 0.000 | 0.000 |
| 0.50 | 0.50 | rpart2 | 45 | 750 | 10.711 | 0.574 | 8.558 | 0.773 | 0.800 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 15 | 250 | 7.357 | 0.410 | 5.827 | 0.451 | 0.000 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 15 | 500 | 7.468 | 0.627 | 5.955 | 0.688 | 0.000 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 15 | 750 | 7.295 | 0.665 | 5.823 | 0.608 | 0.000 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 45 | 250 | 21.652 | 0.606 | 17.159 | 0.471 | 0.000 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 45 | 500 | 21.960 | 0.602 | 17.486 | 0.575 | 0.000 | 0.000 |
| 0.50 | 0.95 | krlsPoly | 45 | 750 | 22.229 | 0.728 | 17.676 | 0.543 | 0.000 | 0.000 |

表2：外部验证（相关协变量）：均方根误差（RMSE）超过5且方向准确率（DIR）低于0.8的情景，表明PITE估计效果不佳。

结果表明，在结构化模拟设置中，较简单的正则化线性模型往往能在预测准确性与可解释性之间取得较好的平衡，从而用于估计 PITE。

### *3.3.2.* 协变量间交互作用

本模拟研究旨在纳入协变量间的高阶交互作用，以反映治疗效应源于复杂潜在结构的情景。具体而言，我们生成样本量为 $n \in \{500, 750, 1000\}$ 的样本，其中包含六个基线协变量，这些协变量独立地取自多元正态分布。基于这些协变量，我们构造所有最高达六阶的交互项，形成一个包含63个候选预测变量的完整设计矩阵：六个主效应以及所有高阶交互项，直至六重交互项。

从这组协变量中，随机选取一个大小为 $p \in \{5, 15, 45\}$ 的子集，以定义协变量集 $W$，随后将其用于个体治疗获益的生成。

$$\Delta(X_i) = W_i^\top \beta_\Delta,$$

对照潜在结局由基线协变量的线性函数生成 $X$,

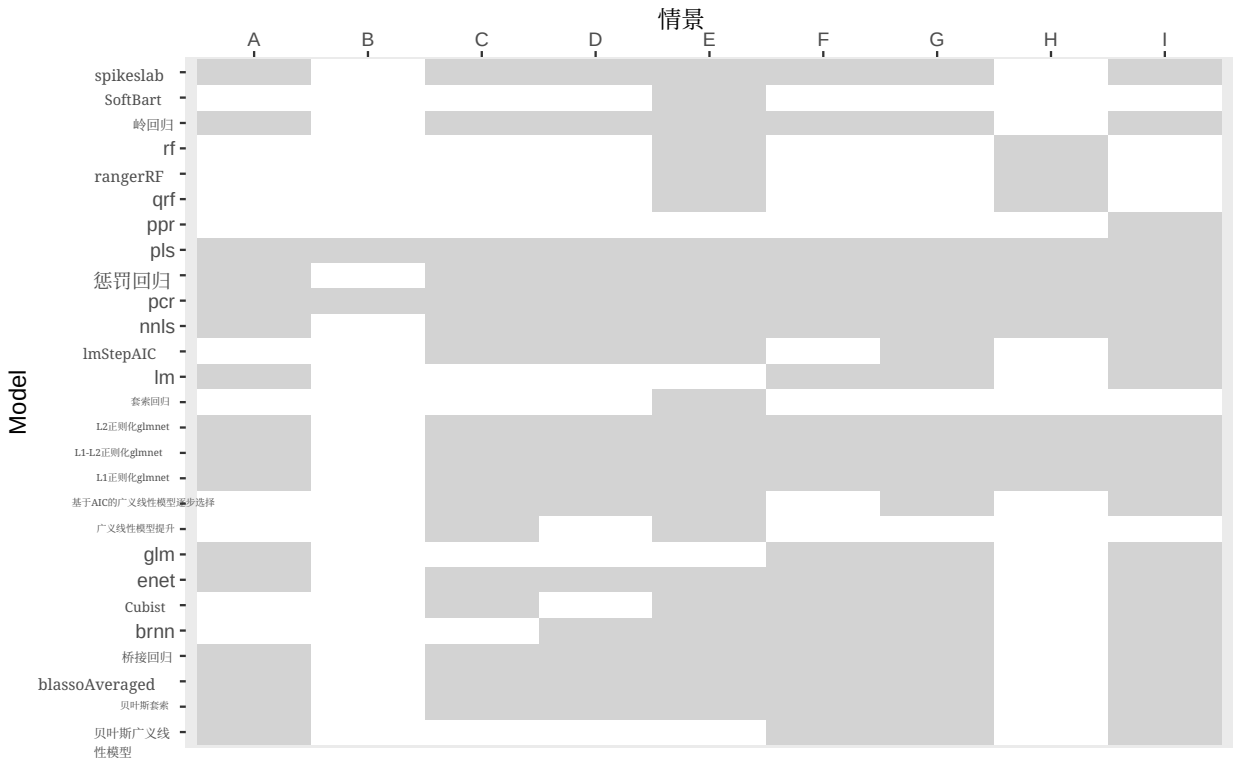


图4：外部验证（相关协变量）：RMSE 最低 $< 2$ 且 DIR 较高 $> 0.95$的模型。模拟条件 [A–I] 及维度 $p = 45$。

设定： [A] $n = 250$, $\rho = 0.5$, ete$=0.5$; [B] $n = 250$, $\rho = 0.95$, ete$=0.25$; [C] $n = 250$, $\rho = 0.95$, ete$=0.5$; [D]$n = 500$, $\rho = 0.5$, ete$=0.5$; [E] $n = 500$, $\rho = 0.95$, ete$=0.25$; [F] $n = 500$, $\rho = 0.95$, ete$=0.5$; [G] $n = 750$,$\rho = 0.5$, ete$=0.5$; [H] $n = 750$, $\rho = 0.95$, ete$=0.25$; [I] $n = 750$, $\rho = 0.95$, ete$=0.5$。

也从多元正态分布中独立抽取。个体治疗效应建模为 $W$ 的线性函数。对于接受治疗的个体，将治疗效应加到对照结局上，得到每个个体的观测结局。

在这一情境下，模型表现出中等的准确性，平均 RMSE 为 1.77（范围： $1.42^{\sim}4.64$）。方向准确率（DIR）是识别治疗有益还是有害的一项关键指标，其平均值为 0.59，最小值为 0.46，最大值为 0.73——这表明其在指导个体化治疗决策方面的可靠性有限。

图5展示了纳入交互效应的外部验证中整体的变异情况，包括预测准确性和方向可靠性如何随治疗效应量变化。在所有设定中，落入左上方灰色区域（RMSE 高、DIR 低）的模型尤其令人担忧，因为它们不仅预测效果差，还会错误判定治疗有益还是有害；若将其用于个体化决策，可能带来临床风险。随着交互项数量增加，我们观察到模型表现的变异性增大，更多模型偏移至灰色区域，尤其是在治疗效应较小或为零时（例如最左侧的面板）。

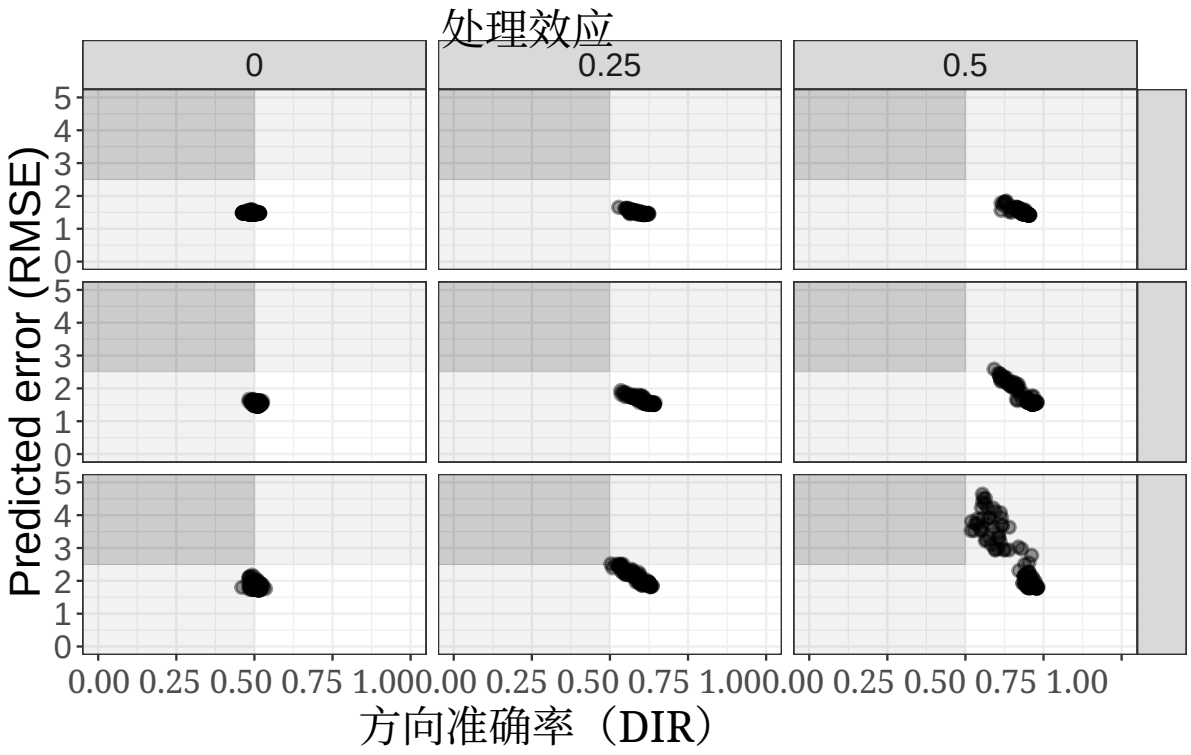


图5：外部验证（协变量间交互）：各模型的预测性能均方根误差（RMSE）与方向准确率（DIR）之间的关系，以及处理效应影响和协变量间相关值。每个点代表一种特定的模型参数配置，并标注了模型名称、样本量（n）和交互作用数量。深灰色区域表示不理想的区域

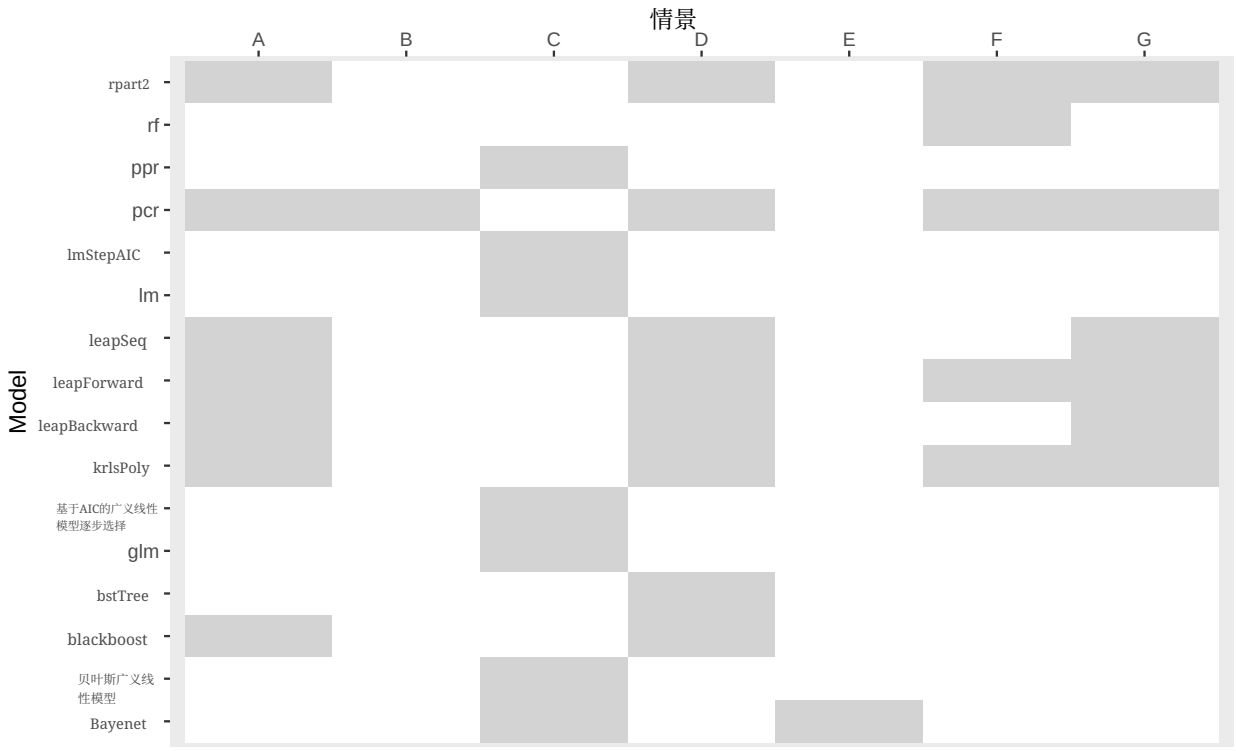


图6：外部验证（协变量间交互）：均方误差高 $\geq 2$ 且方向准确率低 $\leq 0.55$的模型。在模拟条件 [A–G]下，交互维度为45。

设定： [A] $n = 1000$, $\mu_{\beta_\Delta}$=0.25; [B] $n = 1000$, $\mu_{\beta_\Delta}$=0.5; [C] $n = 500$, $\mu_{\beta_\Delta}$=0; [D] $n = 500$, $\mu_{\beta_\Delta}$=0.25; [E] $n = 750$, $\mu_{\beta_\Delta}$=0; [F] $n = 750$, $\mu_{\beta_\Delta}$=0.25; [G] $n = 750$, $\mu_{\beta_\Delta}$=0.5;

协变量交互下的最差模型表现。 在包含协变量间交互效应和高维交互（45）的外部验证情景中，若干模型表现不佳，其定义为 RMSE $\geq 2$ 且 DIR $\leq 0.55$。图6显示，这些最差情形最常出现在治疗效应的均值$\mu_{\beta_\Delta}$为零或较低时，尤其是在样本量为 500 至 1000 的条件 [A] 至 [G], 下。这些模型中，主成分回归、rpart2 和 krlsPoly 的失败次数最多，在多达五种条件下表现不佳；而 leapForward、leapSeq 和 blackboost 等模型则在三种或更多条件下表现不佳。这些模型在存在交互效应时持续表现不佳，表明它们并不适合在复杂的高维场景下进行个体化治疗效应估计。

协变量交互下的最佳预测表现。 图7显示，在包含交互效应和强治疗信号（$\mu_{\beta_\Delta} = 0.5$）的外部验证条件下，多种模型均取得了较高表现，定义为 RMSE $< 1.6$ 且 DIR $> 0.65$。这些成功案例在条件 [A] 至 [F], 下均稳定出现，这些条件包含中到大的样本量（$n = 500$˘1000）和较低的交互维度（交互作用数量 =5 或交互作用数量 15）。多个模型——包括贝叶斯广义线性模型、贝叶斯套索、Cubist、广义线性模型提升、岭回归、spikeslab 以及所有 glmnet 变体——在全部六种条件下均取得了完美表现，展现出较高的预测准确性和可靠的方向推断。更复杂或自适应的模型，如 BART、贝叶斯正则化神经网络和 bstTree，也保持了稳定表现，凸显了它们在交互丰富情景下的稳健性。

### *3.4.* 示例：不同模型在不同个体病例复杂度下的表现

为考察预测准确性在不同个体观测间的变化，我们基于方向准确率指标（公式 2）分析了一个具有代表性的模拟数据集，其中包含 250 名验证患者。该指标量化了估计的个体预测治疗效应（PITE）与真实值之间的符号一致性。对于每位患者 $i$，我们计算了正确预测治疗方向的模型所占比例，

$$C_i = \frac{1}{K} \sum_{k=1}^{K} D_i^{(k)},$$

其中， $K$ 表示评估模型的总数。因此， $C_i \in [0, 1]$ 表示正确识别治疗对患者 $i$ 有益还是有害的方法所占的比例，并可作为预测共识或复杂度的度量。我们将患者分为十二个按序排列的复杂度类别，从 $C_i = 1$ （所有模型均正确）到 $C_i = 0$ （没有模型正确），每档相差 10%。

从概念上讲，这一谱系的两端具有参考意义。属于**100%类别** 的患者对应于最简单、最可预测的情形，所有评估模型均正确推断了治疗效应方向——表明存在清晰、一致的治疗反应模式。相比之下，**0%类别**代表最复杂或最模糊的情形，没有任何模型能够正确

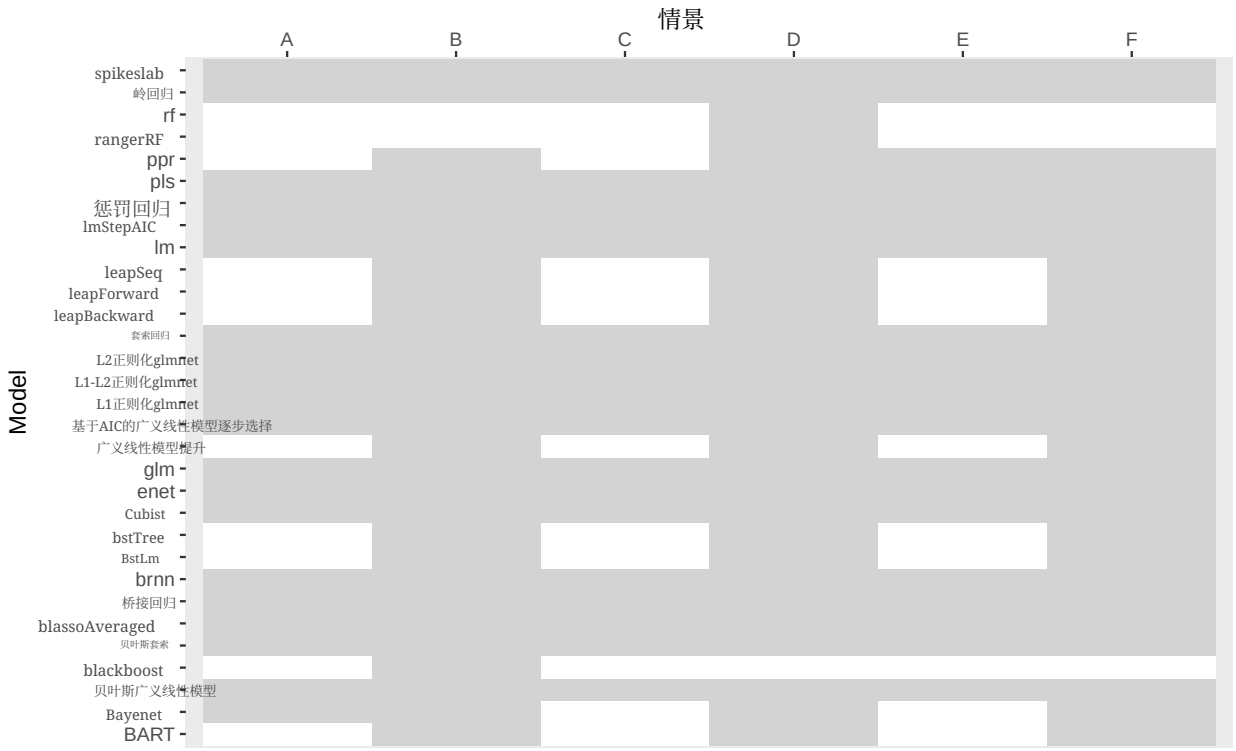


图7：外部验证（协变量间交互）：在存在协变量交互的外部验证情景中达到高 PITE 准确度的模型（RMSE $< 1.6$ 和 DIR $> 0.65$）
设定： [A] $n = 1000$，交互作用数量15； [B] $n = 1000$，交互作用数量5； [C] $n = 500$，交互作用数量15； [D] $n = 500$，交互作用数量5； [E] $n = 750$，交互作用数量15；[F] $n = 750$，交互作用数量5。

预测了治疗方向。这些患者可能表现出特异性的治疗反应、较低的信噪比，或任何模型都难以充分捕捉的交互结构。中间类别（例如，40–70%）反映了不同程度的预测共识，即只有部分模型正确判断了治疗方向。

对于每个类别，我们都计算了各建模方法正确分类的患者比例。结果汇总于表3。这个基于复杂度的框架能够灵活地呈现患者层面异质性不断增加时模型的稳健性，便于从可靠性和适应性两方面直接比较预测策略。

表3报告了一个代表性模拟数据集的结果，该数据集包含250名验证患者（$n = 500$, $p = 45$, $ete = 0.5$），并将个体划分为12个有序复杂度类别。复杂度定义为：对于每位患者，能够正确预测真实预测个体治疗效应（PITE）方向的方法所占的比例。

在复杂度最低的分层（90–100%）中，所有模型的方向准确率都接近完美（$> 95\%$）。然而，随着复杂度增加，模型表现出现了显著差异。BART、BRNN、岭回归和主成分回归等灵活学习器，即使面对复杂度中等的

表3：按复杂度分层的患者中，11种预测建模策略的方向准确率。

| 类别 | 数量 | pls | brnn | rf | 桥接回归 | enet | pcr | 岭回归 | spikeslab | 贝叶斯套索 | 套索回归 | BART |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0% | 27 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| (0-10] | 8 | 0.0 | 12.5 | 25.0 | 0.0 | 0.0 | 25.0 | 12.5 | 0.0 | 0.0 | 0.0 | 25.0 |
| (10-20] | 10 | 0.0 | 10.0 | 30.0 | 0.0 | 10.0 | 80.0 | 20.0 | 10.0 | 0.0 | 0.0 | 40.0 |
| (20-30] | 5 | 20.0 | 20.0 | 60.0 | 20.0 | 0.0 | 60.0 | 40.0 | 40.0 | 0.0 | 0.0 | 40.0 |
| (30-40] | 5 | 20.0 | 20.0 | 60.0 | 40.0 | 0.0 | 80.0 | 40.0 | 20.0 | 20.0 | 20.0 | 80.0 |
| (40-50] | 8 | 12.5 | 62.5 | 50.0 | 62.5 | 37.5 | 50.0 | 62.5 | 12.5 | 50.0 | 37.5 | 62.5 |
| (50-60] | 9 | 44.4 | 77.8 | 0.0 | 33.3 | 77.8 | 22.2 | 66.7 | 44.4 | 100.0 | 100.0 | 33.3 |
| (60-70] | 10 | 90.0 | 100.0 | 40.0 | 60.0 | 80.0 | 20.0 | 70.0 | 70.0 | 70.0 | 80.0 | 20.0 |
| (70-80] | 21 | 95.2 | 90.5 | 19.0 | 90.5 | 95.2 | 14.3 | 90.5 | 90.5 | 95.2 | 100.0 | 19.0 |
| (80-90] | 21 | 95.2 | 95.2 | 33.3 | 100.0 | 100.0 | 38.1 | 90.5 | 95.2 | 95.2 | 95.2 | 61.9 |
| (90-100) | 45 | 97.8 | 100.0 | 86.7 | 100.0 | 100.0 | 44.4 | 93.3 | 95.6 | 97.8 | 100.0 | 84.4 |
| 100% | 81 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

**注：** 复杂度类别反映患者层面的难度，定义为正确预测真实 PITE 方向的方法所占的比例。数量表示在每个类别中，各方法正确识别治疗方向的患者人数。研究纳入了来自代表性方法家族的 11 种广泛使用的方法：惩罚回归（套索回归、贝叶斯套索、弹性网络、岭回归）、基于投影的方法（主成分回归、偏最小二乘）、基于树的模型（随机森林、BART）、贝叶斯方法（桥接回归、spikeslab）以及神经网络（贝叶斯正则化神经网络）。

患者（30–60%），仍表现稳健；而纯惩罚线性模型（套索回归、贝叶斯套索、弹性网络）的表现则迅速恶化，在高复杂度设定（具有挑战性的患者）中无法恢复正确的方向。树集成（随机森林）和贝叶斯收缩方法（桥接回归、尖峰-平板）表现出中等程度的稳健性，在中等范围（40–70%）内保持较强表现，但此后开始下降。

这些结果凸显了预测学习器各自不同的工作特性。BART 和 BRNN 等方法兼具非线性灵活性与强正则化能力，因此能够随着信号结构变得更加复杂而平稳地适应。岭回归和主成分回归凭借偏差—方差稳定性表现出稳健性，在中等程度的异质性下始终表现稳定。相比之下，Lasso 型估计量虽具有可解释性且较为简约，却对特征相关性和交互效应高度敏感，因此在复杂的治疗反应面下实用性有限。

从实践角度来看，本文提出的基于复杂度的框架提供了一种结构化诊断方法，用于评估 PITE 估计值的稳定性。在 100% 时几乎普遍一致、而在 0% 时完全不一致的现象，凸显了当前统计学习方法的局限：前者反映了具有可预测结构的治疗反应，后者则揭示了 PITE 估计值存在根本性的不确定性。

在实际应用中，这些发现表明，对于以复杂度较低至中等的患者为主的群体，灵活的贝叶斯模型或集成学习器（如 BART、BRNN）可能最能可靠地估计治疗方向。当可解释性或特征选择至关重要时，惩罚回归仍可作为集成或堆叠框架中的补充工具。预测共识这一解读视角由此将方法学评估与临床洞见联系起来，为异质性患者群体提供有原则的模型选择指导。

## 4. 结论

我们的模拟证实，内部验证——尤其是在存在相关性但没有交互效应的情况下——为评估预测的个体治疗效应（PITE）提供了更平稳、波动更小的环境。在这些条件下，许多方法都能实现较低的估计均方根误差（RMSE）和较高的方向准确率（DIR），从而掩盖了潜在弱点。相比之下，外部验证通过 $\mu_{\beta_\Delta}$ 引入分布偏移，并在某些情景中引入复杂的高阶交互作用，从而揭示模型的真实稳健性。这些交互作用会显著降低若干常用方法的表现，凸显了在能够检验模型普适性的条件下评估模型的必要性。外部验证仍是评估模型的内部表现能否转化为临床决策的黄金标准。

在不同模型和多种验证方案中，正则化线性方法和基于投影的方法——包括岭回归、套索回归、弹性网络、glmnet 变体、偏最小二乘法（PLS）和主成分回归（PCR）——在不同样本量、多重共线性水平和治疗效应大小下，始终实现了较低的均方根误差（RMSE）和较高的方向准确率（DIR）。这些方法表现出显著的稳定性，避免了核方法（如 krlsPoly）和递归分割（rpart2）中观察到的严重性能下降；后两者的 RMSE $\geq 5$ 且 DIR $< 0.8$。贝叶斯加性回归树（BART）、提升算法及某些贝叶斯惩罚回归等灵活学习器表现不俗，但仅限于有利的条件下——样本量中等至较大、治疗信号较强（$\mu_{\beta_\Delta} \approx 0.5$）且维度适中（$p \leq 15$）。考虑到严格的阈值（RMSE $< 1$，DIR $> 0.95$），即使在内部验证中，也只有少数模型在高复杂度条件下反复取得成功。

这些结果为在精准医学和临床试验设定中选择和审核模型提供了路线图。首先，惩罚回归线性方法和基于投影的方法可作为稳健的默认选择，在广泛的现实条件下保持稳定。其次，灵活的非参数方法——包括基于树的集成方法和贝叶斯学习器——应被视为有条件的策略：当数据特征符合其优势时，这些方法效果良好；否则则容易出现不稳定。第三，在外部验证中评估 PITE 有效性时，估计均方根误差和方向准确率都至关重要，因为模型即使能够准确预测效应幅度，却错误判断治疗方向，仍可能误导临床决策。最后，仅靠内部验证不足以确认某种方法已具备临床应用准备度；必须进行严格的外部验证，以确保其安全且具有普适性的部署。随着该领域迈向可解释人工智能（XAI），平衡预测性能、可解释性和稳健性，将是把 PITE 估计转化为可操作、可信赖的临床工具的关键。

## 5. 讨论

我们的模拟表明，在复杂的高维场景中，正则化线性方法与基于投影的方法始终能提供最可靠的个体治疗效应预测（PITE）。

在 $p = 45$ 情景中——这些情景反映了基因组学、影像和基于电子健康记录的照护等现代生物医学应用——偏最小二乘和主成分回归在保留预测结构的同时有效降低了维度，即使在极端多重共线性（$\rho = 0.95$）下也能获得较低的均方根误差和较高的方向准确率。相比之下，krlsPoly和rpart2等模型频繁出现在表现较差的表格中，产生了偏高的均方根误差（最高达22.2）和低方向正确率（$< 0.6$），尤其是在相关性和治疗效应量均较高时——而这恰恰是准确治疗分配最为关键之处。

外部验证结果进一步表明，需要保持谨慎。一些模型在样本内表现良好，但应用于独立数据时性能急剧下降，说明其泛化能力较差。在这些变化下，正则化方法和基于投影的方法仍能保持治疗效应的大小和方向，表明当稳健性至关重要时，这些方法值得采用。对交互作用不敏感的模型——例如逐步回归的变体、PCR 和 rpart2——尤其容易受到影响；而套索回归、岭回归、弹性网络、贝叶斯方法（blasso、bayesglm、spikeslab）和灵活学习器（BART、brnn、Cubist）则在不同条件下表现稳定且准确。

仅靠内部验证不足以确保稳定性。即使不包含交互项，表现较差的模型也集中在低方向准确率、高均方根误差区域，这凸显了正则化和交互感知诊断的必要性。尽管 glmnet、blasso、岭回归和 PLS 等方法在内部验证中表现出色，但其最佳结果出现在治疗效应较大且样本量较多的设定中，说明有必要针对较弱信号和外部变化进行压力测试。

这些结果与既往研究一致 [52], ，强调应采用 PITE 特异性验证（使用均方根误差和方向准确率等指标），而非依赖间接的基于结局的诊断。尽管我们的模拟涵盖了高维性、多重共线性和复杂交互作用，未来研究仍应关注校准、领域自适应以及真实世界临床数据集。

稳健且易于解释的方法，尤其是正则化方法或基于投影的方法，为治疗分层和个性化决策提供了有前景的基础。通过结合内部验证、外部验证和方向性评估，这些方法能够提供具有临床实践指导价值的个体化治疗效果，同时避免过拟合和不稳定性。

需要考虑的一个重要问题是，在我们的模拟中，数据生成机制遵循线性回归假设，因此预计大多数模型都能取得较好的表现。然而，通过在线性、非线性和基于树的方法中评估 PITE，我们明确量化了不同结构假设之间的预测差异。

最后，表3表明，PITE框架中的模型表现会受到患者层面复杂性的显著影响。虽然大多数方法在低复杂度（“容易”）病例中表现良好，但随着复杂度增加，只有一部分方法——其中尤以BART、BRNN、岭回归和主成分回归为代表——能够保持较高的方向准确率。这些总体稳健的学习器在90–100%类别中表现出近乎完美的一致性，即使在中度异质性下也保有较强的预测能力，说明它们能够抵御噪声，并捕捉非线性效应或交互效应。桥接回归、尖峰-平板和随机森林等灵活的贝叶斯模型及集成模型在低至中等复杂度层中同样表现出色，但在高度复杂的病例中稳定性较差，这与它们对微弱或相互交织的治疗信号较为敏感相一致。相比之下，

套索回归、贝叶斯套索和弹性网络等惩罚线性方法虽然有助于提高可解释性和进行变量选择，但随着复杂度增加，其表现会急剧下降，反映出它们在捕捉高阶依赖关系方面灵活性有限。偏最小二乘和主成分回归等基于投影的方法在低复杂度设定下表现尚可，但当潜在结构与治疗反应机制不匹配时，性能会迅速恶化。值得注意的是，在最复杂的分层（0–10%）中，所有受评估的模型都未能恢复正确的治疗方向，这凸显了此类病例固有的难度，也说明需要结合灵活学习器与先验知识的混合或集成框架。总体而言，这些结果强调，PITE估计中的模型选择不应仅依赖汇总性能指标，还应考虑患者复杂性与模型适应性之间的相互作用。在实践中，当极复杂患者较少时，BART和BRNN是精准医学应用中可靠的默认策略；而将惩罚回归与更灵活的学习器结合，则可能在可解释性与稳健性之间实现有效平衡，适用于异质性人群。

**利益声明**

未声明任何利益冲突。

**参考文献**

[1] Joshua D. Angrist. 理论与实践中的治疗效应异质性. 《经济学杂志》, 114(494):C52–C83, 2004年. ISSN 00130133, 14680297. URL http://www.jstor.org/ stable/3590307.[2] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. 通过监督主成分进行预测. 《美国统计学会会刊》, 101(473):119–137, 2006年3月. ISSN 1537-274X. doi: 10.1198/016214505000000628. URL http://dx.doi.org/10. 1198/016214505000000628.[3] Nicolás M. Ballarini, Gerd K. Rosenkranz, Thomas Jaki, Franz König, and Martin Posch. 通过预测个体治疗效应识别临床试验中的亚组. *PLOSONE*, 13(10):e0205971, 2018年10月. doi: 10.1371/journal.pone.0205971. URL https://doi. org/10.1371/journal.pone.0205971.[4] Thomas Lumley（基于 Alan Miller 的 Fortran 代码）. ***leaps***：回归子集选择, 2024年. URL https://CRAN.R-project.org/package=leaps. R软件包版本 3.2.[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 使用 lme4 拟合线性混合效应模型. 统计软件杂志, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.[6] Leo Breiman. 随机森林. 机器学习, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/a:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

[7] P. J. Brown, M.Vannucci, and T. Fearn. 多变量贝叶斯变量选择与预测.皇家统计学会期刊：*B*辑（统计方法学）, 60(3):627–641, 1998. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/2985935.

[8] Chi Chang、Thomas Jaki、Muhammad Saad Sadiq、Alena Kuhlemeier、Daniel Feaster、Natalie Cole、Andrea Lamont、Daniel Oberski、Yasin Desai 和 M. Lee Van Horn。评估治疗效应中是否存在个体差异的置换检验。《医学研究中的统计方法》，30(11):2369–2381，2021年9月。doi：10.1177/09622802211033640。URL：https://doi.org/10.1177/09622802211033640。

[9] Hugh A. Chipman、Edward I. George 和 Robert E. McCulloch。BART：贝叶斯加性回归树。应用统计学年刊，4(1)，2010年3月。ISSN 1932-6157。doi：10.1214/09-aoas285。URL：http://dx.doi.org/10.1214/09-AOAS285。

[10]Vincent Dorie、Masataka Harada、Nicole Bohme Carnegie 和 Jennifer Hill。用于评估对未测量混杂的敏感性的灵活且易于解释的框架。医学统计学，35(20):3453–3470，2016年5月。ISSN 1097-0258。doi: 10.1002/sim.6973。URL http://dx.doi.org/10.1002/sim.6973。

[11] Bradley Efron、Trevor Hastie、Iain Johnstone 和 Robert Tibshirani。最小角回归。《统计年鉴》，32(2)，2004年4月。ISSN 0090-5364。doi: 10.1214/009053604000000067。URL http://dx.doi.org/10.1214/009053604000000067。

[12] Orestis Efthimiou、Jeroen Hoogland、Thomas P.A. Debray、Michael Seo、Toshiaki A.Furukawa、Matthias Egger 和 Ian R. White。衡量预测模型个性化治疗选择的性能。医学统计学，42(8):1188–1206，2023年1月。ISSN 1097-0258。doi: 10.1002/sim.9665。URL http://dx.doi.org/10.1002/sim.9665。

[13]Jeremy Ferwerda、Jens Hainmueller 和 Chad J. Hazlett。R 中的基于核的正则化最小二乘法（核正则化最小二乘）及 Stata（krls）。统计软件杂志，79(3):1–26，2017年。doi: 10.18637/jss.v079.i03.

[14] Jerome H. Friedman。贪心函数逼近：一种梯度提升机。《统计年鉴》，29(5)，2001年10月。ISSN 0090-5364。doi: 10.1214/aos/1013203451。URL http://dx.doi.org/10.1214/aos/1013203451。

[15] Jerome H. Friedman 和 Werner Stuetzle。投影寻踪回归。《美国统计学会杂志》，76(376):817–823，1981年12月。ISSN 1537-274X。doi: 10.1080/ 01621459.1981.10477729。URL http://dx.doi.org/10.1080/01621459.1981.10477729。

[16] Zijun Gao、Trevor Hastie 和 Robert Tibshirani。通过匹配评估异质性治疗效应估计准确性。医学统计学，40(17):3990–4013，2021年4月。ISSN 1097-0258。doi: 10.1002/sim.9010。URL http://dx.doi.org/10.1002/sim.9010。

[17] Andrew Gelman 和 Jennifer Hill。使用回归和多层次/分层模型进行数据分析。剑桥大学出版社，2006年12月。ISBN 9780511790942。doi: 10.1017/ cbo9780511790942。URL http://dx.doi.org/10.1017/CBO9780511790942.[18]Andrew Gelman 和 Yu-Sung Su。*arm*：使用回归和多层次/分层模型进行数据分析，2024年。URL https://CRAN.R-project.org/package=arm。R软件包版本 1.14-4。[19] Jelle J. Goeman。Cox比例风险模型中的L1惩罚估计。生物统计学杂志，52(1):70–84，2010年2月。ISSN 1521-4036。doi: 10.1002/bimj.200900028。URL http://dx.doi.org/10.1002/bimj.200900028。[20] Jelle J. Goeman。*penalized*：*L1*惩罚估计，2022年。R软件包版本 0.9-52。[21] Robert B. Gramacy。*monomvn*：单调缺失数据下*MVN*和*Student-t*数据的估计，2024年。URL https://CRAN.R-project.org/package=monomvn。R软件包版本 1.9-21。[22] Jennifer L. Hill。用于因果推断的贝叶斯非参数建模。计算与图形统计学杂志，20(1):217–240，2011年1月。ISSN 1537-2715。doi: 10.1198/jcgs.2010.08162。URL http://dx.doi.org/10.1198/jcgs.2010.08162。[23] Daniel E. Ho、Kosuke Imai、Gary King 和 Elizabeth A. Stuart。MatchIt：用于参数因果推断的非参数预处理。统计软件杂志，42(8):1–28，2011年。doi: 10.18637/jss.v042.i08。[24] Arthur E. Hoerl 和 Robert W. Kennard。岭回归：针对非正交问题的有偏估计。计量学，42(1):80–86，2000年2月。ISSN 1537-2723。doi: 10.1080/ 00401706.2000.10485983。URL http://dx.doi.org/10.1080/00401706.2000.10485983。[25] Benjamin Hofner、Andreas Mayr、Nikola Robinzonov 和 Matthias Schmid。*mboost*：*R*中的基于模型的提升，2014年。R软件包。[26] J. Hoogland、O. Efthimiou、T. L. Nguyen 和 T. P. A. Debray。评估个体化治疗效应预测：基于模型的区分度与校准评估视角。医学统计学，43(23):4481–4498，2024年8月。ISSN 1097-0258。doi: 10.1002/sim.10186。URL http://dx.doi.org/10.1002/sim.10186。[27] Torsten Hothorn、Kurt Hornik 和 Achim Zeileis。无偏递归分割：一种条件推断框架。计算与图形统计学杂志，15(3):651–674，2006年。doi: 10.1198/106186006X133933。[28] H. Ishwaran、J.S. Rao 和 U.B. Kogalur。*spikeslab*：使用尖峰与板回归进行预测和变量选择，2022年。URL https://cran.r-project.org/package=spikeslab。R软件包版本 1.1.6。

[29] Friedman J、Hastie T 和 Tibshirani R. 通过坐标下降为广义线性模型生成正则化路径。统计软件杂志，33(1):1–22，2010。doi: 10.18637/jss.v033.i01.[30] Thomas Jaki、Chi Chang、Alena Kuhlemeier 和 M. Lee Van Horn。预测个体治疗效应：机器学习和人工智能面临的挑战与机遇。*KI* - 人工智能，2024年1月。ISSN 1610-1987。doi: 10.1007/s13218-023-00827-4。URL http://dx.doi.org/10.1007/s13218-023-00827-4.[31] Ian T Jolliffe。主成分分析。Springer-Verlag，2002。ISBN 0387954422。doi: 10.1007/b98835。URL http://dx.doi.org/10.1007/b98835。

[32] David M Kent、Ewout Steyerberg 和 David van Klaveren。《个性化循证医学：异质性治疗效应的预测方法》。*BMJ*，第 k4245 页，2018年12月。ISSN 1756-1833。doi：10.1136/bmj.k4245。URL：http://dx.doi.org/10.1136/bmj. k4245。

[33] Roger Koenker。《分位数回归》。剑桥大学出版社，2005年。ISBN 9780511754098。doi：10.1017/cbo9780511754098。URL：http://dx.doi.org/10.1017/CBO9780511754098。

[34] Alena Kuhlemeier、Tomas Jaki、Katie Witkiewitz、Elizabeth A. Stuart 和 M Lee Van Horn。对样本外受访者的个体治疗效应预测值进行验证。医学统计学，2024年。doi：10.1002/sim.10187。

[35] Kuhn 和 Max。使用 caret 软件包在 r 中构建预测模型。统计软件杂志，28(5):1–26，2008年。doi：10.18637/jss.v028.i05。URL https://www.jstatsoft.org/index.php/jss/article/view/v028i05。

[36] Max Kuhn 和 Ross Quinlan。*Cubist*：基于规则和实例的回归建模，2025年。URL https://CRAN.R-project.org/package=Cubist。R 软件包版本 0.5.0。

[37]Max Kuhn 和 Ross Quinlan。*Cubist*：基于规则和实例的回归建-模，2025年。URL https://topepo.github.io/Cubist/。R 软件包版本 0.5.0.9000, https://github.com/topepo/Cubist。

[38]Andrea Lamont、Michael D Lyons、Thomas Jaki、Elizabeth Stuart、Daniel J Feaster、Kukatharmini Tharmaratnam、Daniel Oberski、Hemant Ishwaran、Dawn K Wilson 和 M Lee Van Horn。在随机临床试验中识别预测的个体治疗效应。《医学研究中的统计方法》，27(1):142–157，2016年3月。ISSN 1477-0334。doi：10.1177/0962280215623981。URL http://dx.doi.org/10.1177/0962280215623981。

[39] Charles L. Lawson 和 Richard J. Hanson。《最小二乘问题求解》。工业与应用数学学会，1995年1月。ISBN 9781611971217。doi：10.1137/1. 9781611971217。URL http://dx.doi.org/10.1137/1.9781611971217。

[40] Andy Liaw 和 Matthew Wiener。使用 randomForest 进行分类和回归。*R News*，2（3）：18–22，2002年。URL https://CRAN.R-project.org/doc/Rnews/。[41] Kristian Hovde Liland、Bjørn-Helge Mevik 和 Ron Wehrens。*pls*：偏最小二乘法和主成分回归，2024年。URL https://CRAN.R-project.org/package=pls。R软件包版本 2.8-5。[42] Xi Lu 和 Cen Wu。*Bayenet*：稳健贝叶斯弹性网络，2025年。URL https://CRAN. R-project.org/package=Bayenet。R软件包版本 0.3。[43] David J. C. MacKay。用于反向传播网络的实用贝叶斯框架。*NeuralComputation*，4（3）：448–472，1992年5月。ISSN 1530-888X。doi：10.1162/neco.1992.4.3.448。URL http://dx.doi.org/10.1162/neco.1992.4.3.448。[44] Peter McCullagh 和 John A Nelder。《广义线性模型》。Chapman and Hall/CRC，1989年。[45] Nicolai Meinshausen。*quantregForest*：分位数回归森林，2024年。URL https://CRAN.R-project.org/package=quantregForest。R软件包版本 1.3-7.1。[46] Nicolai Meinshausen 和 Greg Ridgeway。分位数回归森林。机器学习研究杂志，7（6），2006年。[47] E. A. Nadaraya。论回归估计。*Theory of Probability & Its Applications*，9（1）：141–142，1964年1月。ISSN 1095-7219。doi：10.1137/1109020。URL http://dx.doi.org/10.1137/1109020。[48] Hong Ooi。*glmnetUtils*：*“Glmnet”*实用工具，2023年。URL https://CRAN.R-project.org/ package=glmnetUtils。R软件包版本 1.1.9。[49] Trevor Park 和 George Casella。贝叶斯套索。《美国统计学会会刊》，103（482）：681–686，2008年6月。ISSN 1537-274X。doi：10.1198/016214508000000337。URL http://dx.doi.org/10.1198/016214508000000337。[50] Paulino Perez Rodriguez 和 Daniel Gianola。*brnn*：前馈神经网络的贝叶斯正则化，2025年。URL https://CRAN.R-project.org/package=brnn。R软件包版本 0.9.4。[51] R核心团队。*R*：统计计算语言与环境。R统计计算基金会，奥地利维也纳，2023年。URL https://www.R-project.org/。R版本 4.3.1（2023年6月16日，“Beagle Scouts”）。[52] Craig A. Rolling 和 Yuhong Yang。用于估计治疗效应的模型选择。英国皇家统计学会会刊：*B*辑（统计方法学），76（4）：749–769，2014年。[53] George AF Seber 和 Alan J Lee。《线性回归分析》。John Wiley & Sons，2004年。

[54] Rodney Sparapani、Charles Spanbauer 和 Robert McCulloch。使用贝叶斯加性回归树进行非参数机器学习与高效计算：BART R软件包。统计软件杂志，97(1):1–66，2021年。doi: 10.18637/jss.v097.i01。[55] Ewout W. Steyerberg。《临床预测模型：开发、验证与更新的实用方法》。施普林格国际出版，2019年。ISBN 9783030163990。doi: 10.1007/978-3-030-16399-0。URL http://dx.doi.org/10.1007/978-3-030-16399-0。[56] Nathan Teetor。*zeallot*：多重赋值、解包与解构赋值，2018年。URL https://CRAN.R-project.org/package=zeallot。R软件包版本 0.1.0。[57] Robert Tibshirani。通过套索回归进行回归收缩与变量选择。皇家统计学会期刊。*B*辑（方法学），58(1):267–288，1996年。ISSN 00359246。URL http://www.jstor.org/stable/2346178。[58] Michael E. Tipping。稀疏贝叶斯学习与相关向量机。机器学习研究杂志，1:211–244，2001年。URL http://jmlr.csail.mit.edu/papers/v1/ tipping01a.html。[59] Ben Van Calster、Laure Wynants、Jan F.M. Verbeek、Jan Y. Verbakel、Evangelia Christodoulou、Andrew J. Vickers、Monique J. Roobol 和 Ewout W. Steyerberg。决策曲线分析的报告与解读：研究人员指南。《欧洲泌尿学》，74(6):796–804，2018年12月。ISSN 0302-2838。doi: 10.1016/j.eururo.2018.08.038。URL http://dx.doi.org/10.1016/ j.eururo.2018.08.038。[60] Stefan Wager 和 Susan Athey。使用随机森林估计和推断异质性治疗效应。《美国统计学会会刊》，113(523):1228–1242，2018年6月。ISSN 1537-274X。doi: 10.1080/01621459.2017.1319839。URL http://dx.doi.org/ 10.1080/01621459.2017.1319839。[61] Zhu Wang。*bst*：梯度提升，2022年。URL https://CRAN.R-project.org/package=bst。R软件包版本 0.3-24。[62] Hadley Wickham。*ggplot2*：数据分析的优雅图形。纽约 Springer-Verlag，2016年。ISBN 978-3-319-24277-4。URL https://ggplot2.tidyverse.org。[63] Hadley Wickham 和 Jennifer Bryan。*readxl*：读取 *Excel* 文件，2023年。URL https://CRAN. R-project.org/package=readxl。R软件包版本 1.4.3。[64] Hadley Wickham、Mara Averick、Jennifer Bryan、Winston Chang、Lucy D'Agostino McGowan、Romain François、Garrett Grolemund、Alex Hayes、Lionel Henry、Jim Hester、Max Kuhn、Thomas Lin Pedersen、Evan Miller、Stephan Milton Bache、Kirill Müller、Jeroen Ooms、David Robinson、Dana Paige Seidel、Vitalie Spinu、Kohske Takahashi、Davis Vaughan、Claus Wilke、Kara Woo 和 Hiroaki Yutani。欢迎使用 tidyverse。开放源代码软件杂志，4(43):1686，2019年。doi: 10.21105/joss.01686。

[65] Hadley Wickham、Romain François、Lionel Henry、Kirill Müller 和 Davis Vaughan。*dplyr*：数据操作语法，2023年。URL https://CRAN.R-project.org/package=dplyr。R软件包版本 1.1.4。[66] Hadley Wickham、Davis Vaughan 和 Maximilian Girlich。*tidyr*：整理杂乱数据，2024年。URL https://CRAN.R-project.org/package=tidyr。R软件包版本 1.3.1。[67] Svante Wold、Kim Esbensen 和 Paul Geladi。主成分分析。化学计量学与智能实验室系统，2(1–3):37–52，1987年8月。ISSN 0169-7439。doi: 10.1016/ 0169-7439(87)80084-9。URL http://dx.doi.org/10.1016/0169-7439(87)80084-9。[68] Marvin N. Wright 和 Andreas Ziegler：ranger：在 C++ 和 R 中实现随机森林的快速方法。统计软件杂志，77(1):1–17，2017年。doi: 10.18637/jss.v077.i01。[69] Yan Zhao、Xiao Fang 和 David Simchi-Levi。使用多重处理和一般响应类型的增益建模，2017年。URL https://arxiv.org/abs/1705.08492。[70] Hui Zou 和 Trevor Hastie。通过弹性网络进行正则化与变量选择。皇家统计学会期刊。*B*辑（统计方法学），67(2):301–320，2005年。ISSN 13697412，14679868。URL http://www.jstor.org/stable/3647580。[71] Hui Zou 和 Trevor Hastie。*elasticnet*：用于稀疏估计和稀疏主成分分析的弹性网络，2020年。URL https://CRAN.R-project.org/package=elasticnet。R软件包版本 1.3。

## 附录A。方法

### 线性模型与正则化线性模型

**线性模型**是统计分析的标准方法，以可解释性著称。**线性模型（lm）**假设预测变量与响应变量之间存在直接的线性关系，并通过最小化残差平方和来估计系数 [53]。虽然线性模型具有较高的可解释性且计算成本低，但其对线性假设、残差正态性和多重共线性较为敏感。在此基础上，**广义线性模型（glm）**通过各种链接函数来适应不同的结果分布（例如，二元数据使用二项分布，计数数据使用泊松分布），从而提供了更高的灵活性 [44]。这种灵活性同样能够以较低的计算负担实现，因此对于非正态因变量而言具有重要价值。

为应对复杂数据集固有的挑战，尤其是多重共线性和高维度问题，我们进一步采用了**正则化线性模型**。**套索回归（lasso）**采用L1惩罚项，通过将部分系数精确设为零来促进模型稀疏性，从而实现内在的特征选择 [57]。**岭回归（ridge）**引入L2惩罚项，通过有效压缩系数来处理多重共线性并降低方差，同时不会强制将系数设为零 [24]。这两种正则化方法的计算成本都相对较低。结合二者的优势，**弹性网络（enet）**同时利用L1和L2惩罚项，为处理高度相关的预测变量并有效进行变量选择提供了稳健方案 [70]；其计算需求仍然较低。对于需要高度定制化正则化的情形，**惩罚回归（penalized）**为自定义惩罚结构提供了灵活框架 [19], ，能够以最低的计算成本实现对正则化的具体控制。此外，当系数必须从根本上保持非负时，例如在某些物理或经济应用中，**非负最小二乘法（nnls）**可提供受约束的线性解，确保系数为正 [39], ，同时具有较低的计算成本。

对于降维或变量选择至关重要的数据集，我们还采用了在建模前转换或筛选特征的方法。**主成分回归（pcr）**首先将预测变量正交变换为主成分，然后进行线性回归 [31]。这种方法适用于存在多重共线性的高维数据，但计算成本中等，且会降低直接可解释性。类似地，**偏最小二乘法（pls）**通过构建能够最大化预测变量与响应变量之间协方差的潜变量，在优先考虑预测准确度的高维场景中表现出色 [67]。其训练的计算成本也为中等。最后，我们采用了**逐步回归（ leapFor-ward、leapBackward、lmStepAIC）**，这是一种依据赤池信息准则等标准进行迭代变量选择的技术，用于识别重要的预测变量 [11]。该方法的计算成本通常较低，但其迭代特性有时会导致局部最优。

### 基于树的方法与集成方法

我们的分析还借助了**基于树的方法和集成方法**的优势，这些方法以能够捕捉复杂的非线性关系并增强预测稳健性而著称。

基础**决策树（rpart2）**通过基于规则的结构提供较高的可解释性，且计算成本低，但容易出现过拟合和高方差。为解决这一问题，作为一种集成方法，**随机森林（rf）**通过装袋法（自助聚合）构建多棵决策树，显著提升预测准确性，并增强对过拟合的稳健性，尤其适用于高维和非线性场景 [6]。随机森林模型的训练通常需要中等的计算成本。其扩展方法**分位数随机森林（qrf）**能够进一步加深我们对数据的理解：它通过预测响应变量的条件分位数，提供详细的不确定性量化结果，并揭示分布尾部的特征 [46]；这一额外的细节也需要中等的计算成本。

除了装袋法之外，还实现了**提升技术**。**提升树（blackboost、bstTree、gamboost、glmboost、BstLm）**通过依次构建弱学习器来纠正之前的错误，始终实现较高的预测性能 [14]。由于其迭代特性，这些模型通常会产生中高等计算成本。将集成方法与贝叶斯原理相结合的**贝叶斯树（bartMachine）**基于贝叶斯加性回归树（BART），通过许多简单树的求和，在提供出色预测能力的同时，还能为预测提供至关重要的不确定性估计 [9]。其贝叶斯推断固有的复杂性使其计算成本较高。对于需要在可解释性与准确性之间取得平衡的模型，**基于规则的树（cubist）**生成可解释的规则，每条规则都与一个局部线性模型相关联，以中等计算成本结合了决策规则与线性回归的优势[37]。

贝叶斯模型

为了明确量化不确定性并纳入先验知识，采用了一套**贝叶斯模型**。**尖峰-平板（spikeslab）**通过使用鼓励稀疏性的特定先验，为同时进行变量选择和参数估计提供了系统的贝叶斯框架[7]。依赖马尔可夫链蒙特卡罗（MCMC）方法进行推断，使其计算成本较高，尤其是在处理高维数据时。同样，**贝叶斯套索（blasso）**通过采用 L1先验，为套索回归提供了概率解释，能够给出系数的完整后验分布，并从根本上纳入正则化参数不确定性 [49]。其计算成本为中高等，具体取决于 MCMC 采样的复杂度。为提高稳定性，**平均贝叶斯套索（blassoAveraged）**通过对多个贝叶斯套索模型进行平均来降低变异性；正如预期，这会因多次采样运行而带来较高的计算负担。

进一步扩展我们的贝叶斯工具集，**贝叶斯岭回归（桥接回归）**为使用 L2先验进行系数收缩提供了概率框架，能够有效处理多重共线性，同时量化参数不确定性 [24, 19]。基于 MCMC 的推断使其计算成本处于中高水平。**贝叶斯广义线性模型（贝叶斯广义线性模型）**将广义线性模型扩展到贝叶斯设定，可针对各种结果分布对参数进行概率推断 [17]。尽管灵活性很高，但由于依赖抽样方法，其计算成本通常处于中等水平。对于非线性关系，**核贝叶斯回归**

（**rvmPoly**）是一种稀疏贝叶斯核方法，它通过学习对预测至关重要的一组“相关向量”，为复杂的非线性情景提供概率输出 [58]。该模型的计算成本也处于中高水平。

神经网络

对于要求高度灵活性与复杂模式识别的任务，我们纳入了**神经网络**。具体而言，**贝叶斯神经网络（贝叶斯正则化神经网络）**通过在网络权重上设置先验分布来整合贝叶斯推断 [43]。该方法不仅增强了对过拟合的稳健性，还为预测提供了极具价值的不确定性估计，这对于在敏感应用中做出明智决策至关重要。神经网络中贝叶斯推断固有的复杂度带来了较高的计算成本。

非参数方法、核、潜变量、投影与约束模型

最后，我们还采用了其他若干专门模型，以获得特定的分析优势。**投影寻踪（投影寻踪回归）**是一种非参数方法，旨在通过搜索预测变量“有趣的”低维投影来揭示隐藏的非线性关系 [15]。尽管该方法在高维数据的探索性分析中功能强大，但其计算成本通常为中等。对于有监督降维，**基于主成分分析（superpc）**根据主成分与响应变量的相关性选择主成分，随后将其用于回归模型 [2]。该方法在基因组学等高维领域尤为有用，计算成本为中等。当存在非线性关系但其形式未知时，**核回归（krlsPoly）**通过对局部数据点加权，提供一种灵活的非参数方法 [47]。这种灵活性通常伴随中等计算成本，具体取决于数据集大小和核选择。最后，为了解预测变量如何影响响应变量分布的不同部分（而不仅仅是均值），我们采用了**分位数回归（rqlasso）**，该方法还结合 Lasso 正则化，以便在特定分位数上进行变量选择 [33]。这一方法能提供更丰富的分布特征洞见，计算成本通常为中等。

潜变量模型关注的是假定用于生成数据的不可见结构。约束模型关注的是模型拟合过程中所施加的规则或限制。

表 A.4：模型及其缩写列表

| 全称 | 缩写 |
|---|---|
| 线性回归 | lm |
| 广义线性模型 | glm |
| 岭回归 | 岭回归 |
| 套索回归 | 套索回归 |
| 弹性网络 | enet |
| 惩罚线性回归 | 惩罚回归 |
| 带前向选择的线性模型 | leapForward |
| 带后向选择的线性模型 | leapBackward |
| 带逐步选择的线性回归 | lmStepAIC |
| 主成分回归 | pcr |
| 偏最小二乘法 | pls |
| 非负最小二乘法 | nnls |
| 尖峰与板回归 | spikeslab |
| 贝叶斯套索 | 贝叶斯套索 |
| 贝叶斯岭回归（模型平均） | blassoAveraged |
| 贝叶斯岭回归 | 桥接回归 |
| 贝叶斯广义线性模型 | 贝叶斯广义线性模型 |
| 贝叶斯正则化神经网络 | brnn |
| 随机森林 | rf |
| 分位数随机森林 | qrf |
| CART | rpart2 |
| Cubist | Cubist |
| 提升树 | blackboost |
| 提升树 | bstTree |
| 提升广义加性模型 | gamboost |
| 提升广义线性模型 | 广义线性模型提升 |
| 提升线性模型 | BstLm |
| 贝叶斯加性回归树 | bart-softbart |
| 带多项式核的相关向量机 | rvmPoly |
| 投影寻踪回归 | ppr |
| 多项式核正则化最小二乘 | krlsPoly |
| 带有 LASSO惩罚的分位数回归 | rqlasso |

Linear), Neural Net (Neural
(Boosted Additive Models),

## 附录 B. PITE 指标分解

附录 *B.1.* 命题*1*

证明。 根据定义,

$$\mathrm{Var}[\widehat{\mathrm{PITE}}(x)] = \mathrm{Var}[\hat{f}_1(x)] + \mathrm{Var}[\hat{f}_0(x)] - 2\ \mathrm{Cov}[\hat{f}_1(x), \hat{f}_0(x)].$$

由于均方误差是方差与偏差平方之和（偏差-方差权衡），因此可写为:

$$\mathcal{R}_{\mathrm{PITE}} = \mathrm{Var}[\widehat{\mathrm{PITE}}(x)] + \mathrm{bias}^2_{\mathrm{PITE}},$$

其中

$$\mathrm{bias}_{\mathrm{PITE}} = \mathrm{bias}_1 - \mathrm{bias}_0\,.$$

展开偏差$^2_{\mathrm{PITE}}$ 并重新整理各项，即可得到所述结果。 □

我们引入以下记号，以分解 PITE 指标和抽样变异性:

$$e_{1i} = f_1(x_i) - \hat{f}_1(x_i), \qquad e_{0i} = f_0(x_i) - \hat{f}_0(x_i),$$

其中， $e_{1i}$ 和 $e_{0i}$ 分别表示治疗组预测模型和对照组预测模型的估计误差。同样，定义

$$\epsilon_{1i} = f_1(x_i) - \bar{f}_1(x_i), \qquad \epsilon_{0i} = f_0(x_i) - \bar{f}_0(x_i),$$

其中， $f_1(x_i)$ 和 $f_0(x_i)$ 表示期望预测函数（例如，在重复抽样下）。因此， $\epsilon_{1i}$ 和 $\epsilon_{0i}$ 刻画了其相对于期望值的偏离，从而分离出由抽样导致的变异性。

*PITE-$R^2$*分解。考虑

$$R^2_{\mathrm{PITE}} = 1 - \frac{\sum_{i=1}^n (PITE_i - \hat{PITE}_i)^2}{\sum_{i=1}^n (PITE_i - \bar{\mathrm{PITE}})^2},$$

an ˆ ˆ ˆ d 代入 $PITE_i = f_{1i} - f_{0i}$ 和 $PITE_i = f_{1i} - f_{0i}$ 后，分子变为 mes

$$\sum_{i=1}^n (e_{1i} - e_{0i})^2 = \sum_{i=1}^n \left(e_{1i}^2 + e_{0i}^2 - 2e_{1i}e_{0i}\right),$$

以及分母

$$\sum_{i=1}^n (\epsilon_{1i} - \epsilon_{0i})^2 = \sum_{i=1}^n \left(\epsilon_{1i}^2 + \epsilon_{0i}^2 - 2\epsilon_{1i}\epsilon_{0i}\right).$$

Thus

$$R^2_{\mathrm{PITE}} = 1 - \frac{\mathrm{MSE}_1 + \mathrm{MSE}_0 - 2\,\mathrm{Cov}(e_1, e_0)}{\mathrm{VAR}_1 + \mathrm{VAR}_0 - 2\,\mathrm{Cov}(\epsilon_1, \epsilon_0)},$$

其中 $\mathrm{MSE}_t = \frac{1}{n}\sum e_{ti}^2$ 和 $\mathrm{VAR}_t = \frac{1}{n}\sum \epsilon_{ti}^2$ 分别对应 $t \in \{0,1\}$。

证明。 假设解释方差（PITE-$R^2$）定义如下：

$$R^2_{PITE} = 1 - \frac{\sum_{i=1}^n (PITE_i - \hat{PITE}_i)^2}{\sum_{i=1}^n (PITE_i - \bar{PITE})^2}$$

其中，$PITE_i = f_{1i} - f_{0i}$（真实个体治疗效应）、$PITE_i = f_{1i} - f_{0i}$（预测治疗效应）以及 $PITE = f_1 - f_0$。

利用恒等式 $PITE_i = f_{1i} - f_{0i}$ 和 $PITE_i = f_{1i} - f_{0i}$，我们展开分子：

$$(PITE_i - \hat{PITE}_i)^2 = \left[(f_{1i} - \hat{f}_{1i}) - (f_{0i} - \hat{f}_{0i})\right]^2 = (e_{1i} - e_{0i})^2 = e_{1i}^2 + e_{0i}^2 - 2e_{1i}e_{0i},$$

其中，$e_{1i} = f_{1i} - f_{1i}$ 和 $e_{0i} = f_{0i} - f_{0i}$。因此，分子变为：

$$\sum_{i=1}^n (PITE_i - \hat{PITE}_i)^2 = \sum_{i=1}^n (e_{1i}^2 + e_{0i}^2 - 2e_{1i}e_{0i}). \tag{B.1}$$

类似地，我们对.....进行处理

$$(PITE_i - \bar{PITE}_i)^2 = \left[(f_{1i} - \bar{f}_{1i}) - (f_{0i} - \bar{f}_{0i})\right]^2 = (\epsilon_{1i} - \epsilon_{0i})^2 = \epsilon_{1i}^2 + \epsilon_{0i}^2 - 2\epsilon_{1i}\epsilon_{0i},$$

其中 $\epsilon_{1,i} = f_{1,i} - f_{1,i}$, $\epsilon_{0,i} = f_{0,i} - f_{0,i}$

最后，我们得到

$$R^2_{PITE} = 1 - \frac{\mathrm{MSE}_1 + \mathrm{MSE}_0 - 2\,\mathrm{Cov}(\mathrm{e}_1, \mathrm{e}_0)}{\mathrm{VAR}_1 + \mathrm{VAR}_0 - 2\,\mathrm{Cov}(\epsilon_1, \epsilon_0)}$$

其中，均方误差$_1 = \frac{1}{n}\sum_{i=1}^n e_{1,i}^2$、均方误差$_0 = \frac{1}{n}\sum_{i=1}^n e_{0,i}^2$、协方差$_e = \frac{1}{n}\sum_{i=1}^n e_{1,i}e_{0,i}$、方差$_1 = \frac{1}{n}\sum_{i=1}^n \epsilon_{1,i}^2$、方差$_0 = \frac{1}{n}\sum_{i=1}^n \epsilon_{0,i}^2$、协方差$_f = \frac{1}{n}\sum_{i=1}^n \epsilon_{1,i}\epsilon_{0,i}$。 □

注意，PITE-$R^2$ 不仅取决于个体结果模型 $R^2$ 的值，还取决于协方差$(e_1, e_0)$：误差正相关可能改善 $R^2_{\mathrm{PITE}}$。

*PITE-MAE*分解。根据

$$\mathrm{MAE}_{\mathrm{PITE}} = \frac{1}{n}\sum_{i=1}^n |e_{1i} - e_{0i}|,$$

三角不等式和反三角不等式可得

$$|\mathrm{MAE}_{f_1} - \mathrm{MAE}_{f_0}| \le \mathrm{MAE}_{\mathrm{PITE}} \le \mathrm{MAE}_{f_1} + \mathrm{MAE}_{f_0},$$

其中平均绝对误差$_{f_t} = \frac{1}{n}\sum_{i=1}^n |e_{ti}|$。

证明。

$$MAE_{PITE} = \frac{1}{n}\sum_{i=1}^{n}\left|(f_{1i} - f_{0i}) - (\hat{f}_{1i} - \hat{f}_{0i})\right| = \frac{1}{n}\sum_{i=1}^{n}|e_{1i} - e_{0i}|. \tag{B.2}$$

由三角不等式得出的上界

$$|e_{1i} - e_{0i}| \le |e_{1i}| + |e_{0i}|,$$

由反三角不等式得出的下界:

$$|e_{1i} - e_{0i}| \ge ||e_{1i}| - |e_{0i}||,$$

将两个界应用于所有观测值:

$$\frac{1}{n}\sum_{i=1}^{n}||e_{1i}| - |e_{0i}|| \le \frac{1}{n}\sum_{i=1}^{n}|e_{1i} - e_{0i}| \le \frac{1}{n}\sum_{i=1}^{n}(|e_{1i}| + |e_{0i}|),$$

由此得到

$$\frac{1}{n}\sum_{i=1}^{n}||e_{1i}| - |e_{0i}|| \le \mathrm{MAE}_{PITE} \le \mathrm{MAE}_{f_1} + \mathrm{MAE}_{f_0},$$

利用詹森不等式和三角不等式

$$||e_{1i}| - |e_{0i}|| \ge |\,\mathrm{MAE}_{f_1} - \mathrm{MAE}_{f_0}\,|$$

我们得到

$$|\,\mathrm{MAE}_{f_1} - \mathrm{MAE}_{f_0}\,| \le \mathrm{MAE}_{PITE} \le \mathrm{MAE}_{f_1} + \mathrm{MAE}_{f_0},$$

so

$$MAE_{PITE} \le MAE_{f_1} + MAE_{f_0}, \tag{B.3}$$

其中 $MAE_{f_t} = \frac{1}{n}\sum_{i=1}^{n}|f_{ti} - \hat{f}_{ti}|$ 适用于 $t \in \{0,1\}$。 □

这表明，当 $e_1$ 和 $e_0$ 符号相同时，上界是紧的；如果误差相互抵消，误差可能小得多。

$\hat{}$***PITE*** 校准分解。校准回归 $PITE_i$ 对 $PITE_i$:

$$PITE_i = \alpha + \beta\, \hat{PITE}_i + \eta_i.$$

如果每个结果模型都满足

$$f_{ti} = \alpha_t + \beta_t \hat{f}_{ti} + \varepsilon_{ti}, \quad t \in \{0,1\},$$

then

$$PITE_i = (\alpha_1 - \alpha_0) + \beta_1 \hat{f}_{1i} - \beta_0 \hat{f}_{0i} + (\varepsilon_{1i} - \varepsilon_{0i}).$$

PITE 校准对应于 $\alpha = \alpha_1 - \alpha_0$、 $\beta_1 = \beta_0 = \beta$ 和 $\eta_i = \varepsilon_{1i} - \varepsilon_0$ $_i$.

证明。 如果两个结果模型都经过良好校准：

$$f_{1i} = \alpha_1 + \beta_1 \hat{f}_{1i} + \varepsilon_{1i},$$
$$f_{0i} = \alpha_0 + \beta_0 \hat{f}_{0i} + \varepsilon_{0i},$$

那么处理效应为：

$$PITE_i = f_{1i} - f_{0i} = (\alpha_1 - \alpha_0) + \beta_1 \hat{f}_{1i} - \beta_0 \hat{f}_{0i} + (\varepsilon_{1i} - \varepsilon_{0i}).$$

要使其等价于：

$$PITE_i = \alpha + \beta \hat{PITE}_i + \eta_i,$$

我们需要：

$$\boxed{\alpha = \alpha_1 - \alpha_0, \quad \beta_1 = \beta_0 = \beta, \quad \eta_i = \varepsilon_{1i} - \varepsilon_{0i}.}$$

□

请注意， $f_1$ 和 $f_0$ 中的完美校准并不能确保 PITE 校准，除非斜率相匹配且截距之差恰当。

使用某种关系。令

$$e_{1i} = f_1(x_i) - \hat{f}_1(x_i), \quad e_{0i} = f_0(x_i) - \hat{f}_0(x_i), \quad \epsilon_{1i} = f_1(x_i) - \bar{f}_1(x_i), \quad \epsilon_0(x_i) = f_0(x_i) - \bar{f}_0(x_i).$$

实证充分性诊断指标评估。为分析诊断指标的主要贡献，主成分分析显示，两个成分解释了 67% 的变异性。其中，主成分1（标准差 $=$ 1.67）主要受决定系数、方向准确率和$\alpha$的影响，反映了它们对诊断指标信息主要变异的显著贡献。主成分2（标准差 $=$ 1.12）主要由均方根误差和平均绝对误差定义，表明这是一个与误差指标相关的维度。

**附录 C. 模拟**

软件与模型实现

分析使用 Windows 平台（x86_64-w64-mingw32，ucrt）上的 R 4.3.1 [51] 版本完成。用于数据处理、建模和可视化的 R 软件包包括 zeallot [56],readxl [63], MatchIt [23], dplyr [65], tidyr [66], caret [35], 偏最小二乘 [41], lme4 [5], mboost [25],bst [61], monomvn [21], randomForest [40], party [27], Cubist [36], elasticnet [71], leaps [4],

惩罚回归 [20], 核正则化最小二乘 [13], quantregForest [45], spikeslab [28], arm [18], 贝叶斯正则化神经网络 [50], ranger [68],Bayenet [42],BART [54],glmnet [29], glmnetUtils [48], ggplot2 [62],和 tidyverse [64]。

总体训练流程：分别在处理组和对照组子集上训练模型。对于通过 caret::train 实现的方法，我们采用 10 折交叉验证（trainControl(method= "cv", number = 10)），并使用 predict(fit, newdata=test) 获取测试集上的预测结果。

贝叶斯加性回归树（*BART*）：使用 wbart 拟合 BART 模型，进行 50 次预热迭代（nskip = 50）和 200 次后验抽样（ndpost = 200）。采用后验预测均值（yhat.test.mean）进行结局预测，从而灵活地建模非线性和交互作用。

随机森林（*RF*）：使用 ranger 实现随机森林模型，包含 500 棵树，并采用 3 个并行线程（num.threads = 3L）。预测结果取所有树预测值的均值。

贝叶斯套索网络（*BayeNet*）：使用 max.steps = 500 and penalty= "lasso" 拟合 BayeNet 模型。通过 predict(fit, X, clin) 获取后验预测均值作为预测结果。

弹性网络及其变体（*glmnet*）：使用 cva.glmnet 拟合模型，并通过交叉验证选择正则化参数（$\lambda$）。采用了 L1（套索回归，alpha = 1.0）、L2（岭回归，alpha = 0）和混合（弹性网络，alpha = 0.2）惩罚项。通过 predict(fit, newx) 获得预测结果。

附录 *C.1.* 外部相关性

均方根误差的范围为1.34至22.23（均值 $=$ 1.84，中位数 $=$ 1.56），表明误差总体较低，但偶有较大离群值。方向准确率的范围为0.16至0.98（均值 $=$ 0.69，中位数 $=$ 0.70），表明在大多数情况下方向表现良好。决定系数的取值范围为-0.007至0.995（均值 $=$ 0.43，中位数 $=$ 0.35），证实在某些设定下表现较强，但在其他设定下拟合较差。平均绝对误差的范围为1.05至17.68，中位数为1.25。校准指标同样较为分散： $a_0$ 的范围为0至1（中位数 $=$ 0.60）， $a_1$ 的范围为0至0.8（中位数 $=$ 0），表明治疗效应普遍被低估。这些模式证实各指标存在显著变异性以及离群值。

表 C.7：模型性能指标的汇总统计量。

| | 均方根误差 | 调整决定系数 | 平均绝对误差 | DIR | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|
| **Min.** | 1.337 | -0.00667 | 1.046 | 0.1610 | 0.0000 | 0.00000 |
| 第一四分位数 | 1.468 | 0.03796 | 1.175 | 0.5103 | 0.4000 | 0.00000 |
| **中位数** | 1.564 | 0.35256 | 1.247 | 0.6983 | 0.6000 | 0.00000 |
| **Mean** | 1.835 | 0.42946 | 1.463 | 0.6941 | 0.5976 | 0.04766 |
| 第三四分位数 | 1.870 | 0.76893 | 1.495 | 0.8550 | 0.8000 | 0.00000 |
| **Max.** | 22.229 | 0.99507 | 17.676 | 0.9783 | 1.0000 | 0.80000 |
| **NA's** | | 14 | | | | |

附录 *C.2*：外部交互

平均绝对误差为1.37。调整后$R^2$总体较低（均值 $= 0.22$），且在某些情况下为负，反映出较差的模型泛化。此外，个性化组件较弱，$a_1$ 值可忽略不计，均值为 $a_0 = 0.17$，表明对治疗效应异质性的捕捉有限。

表 C.8：模型性能指标的汇总统计量。

| | 均方根误差 | 调整决定系数 | 平均绝对误差 | DIR | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|
| **Min.** | 1.416 | -0.003479 | 1.134 | 0.4600 | 0.0000 | 0 |
| 第一四分位数 | 1.498 | 0.002966 | 1.199 | 0.5108 | 0.0000 | 0 |
| **中位数** | 1.582 | 0.147590 | 1.261 | 0.5916 | 0.0000 | 0 |
| **Mean** | 1.774 | 0.221663 | 1.374 | 0.5901 | 0.1688 | 0 |
| 第三四分位数 | 1.851 | 0.370098 | 1.461 | 0.6612 | 0.2000 | 0 |
| **Max.** | 4.640 | 0.756603 | 2.697 | 0.7324 | 1.0000 | 0 |
| **NA's** | | 1 | | | | |

附录 *C.3*：内部相关

调整后 $R^2$ 的中位数为 0.81，平均绝对误差的中位数较低（0.40），这表明大多数模型都能在这些受控的内部条件下准确捕捉个体治疗效应。此外，$\alpha = 0.33$ 和 $\beta = 0.22$ 的平均值适中，反映出一定程度的个性化，表明部分模型利用了治疗分配中的患者层面异质性。

表 C.9：模型性能指标的汇总统计量。

| | 均方根误差 | 调整决定系数 | DIR | MAE | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|
| **Min.** | 0.02567 | -0.006055 | 0.0224 | 0.0206 | 0.0000 | 0.0000 |
| 第一四分位数 | 0.32037 | 0.200436 | 0.5376 | 0.2590 | 0.0000 | 0.0000 |
| **中位数** | 0.50757 | 0.806061 | 0.8448 | 0.4048 | 0.2000 | 0.2000 |
| **Mean** | 0.87941 | 0.622523 | 0.7635 | 0.6974 | 0.3281 | 0.2175 |
| 第三四分位数 | 0.87909 | 0.980559 | 0.9560 | 0.6934 | 0.6000 | 0.4000 |
| **Max.** | 21.99891 | 0.999999 | 1.0000 | 17.5477 | 1.0000 | 1.0000 |
| **NA's** | | 16 | | | | |

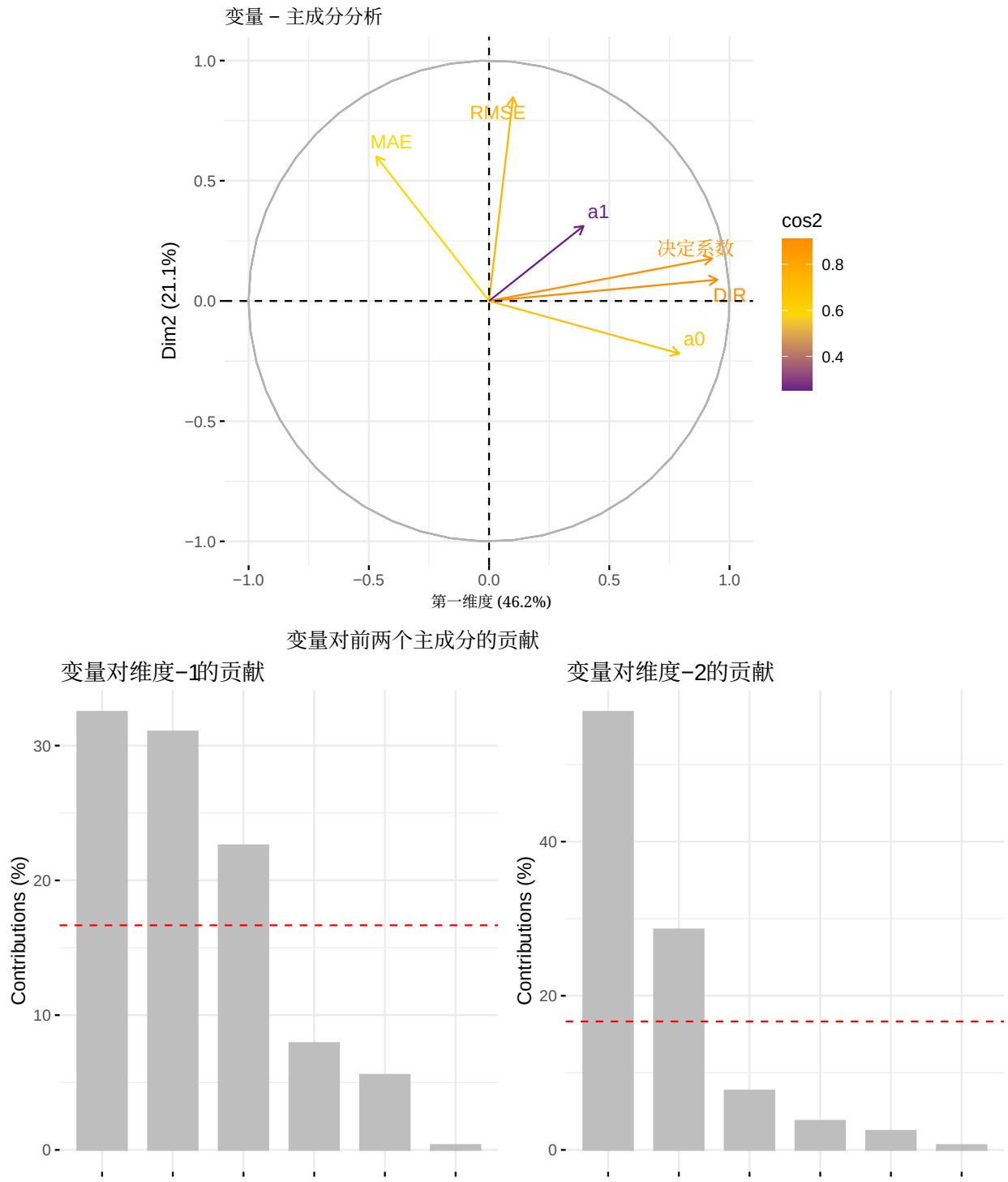
变量 – 主成分分析
RMSE
MAE
a1
决定系数
D_R
a0
cos2
0.8
0.6
0.4
1.0
0.5
0.0
−0.5
−1.0
Dim2 (21.1%)
第一维度 (46.2%)
变量对前两个主成分的贡献
变量对维度−1的贡献
变量对维度−2的贡献
Contributions (%)
30
20
10
0
40

图 B.8：输入标题