arXiv:2302.05131v1 [stat.ME] 10 Feb 2023

样本外 R^2：估计与推断

Stijn Hawinkel*根特大学植物生物技术与生物信息学系、VIB植物系统生物学中心；Willem Waegeman 根特大学数据分析与数学建模系；Steven Maere 根特大学植物生物技术与生物信息学系、VIB植物系统生物学中心

2023年2月13日

摘要

样本外预测是检验预测模型的试金石，但通常无法获得独立测试数据集来评估预测误差。因此，样本外性能通常使用交叉验证或自助法等数据拆分算法进行估计。对于定量结果变量，解释方差与总方差之比可以用决定系数或样本内 R^2进行概括，这一指标易于解释，也便于在不同结果变量之间进行比较。与样本内 R^2 不同，样本外 R^2 尚未得到明确定义，样本外ˆ R^2 的变异性也在很大程度上被忽略。通常只报告其点估计，从而妨碍了对不同结果变量可预测性的正式比较。本文明确将样本外 R^2 定义为两个预测模型的比较，提供无偏估计量，并利用关于数据拆分估计不确定性的最新理论进展，为ˆ R^2提供标准误。我们通过模拟研究考察了 R^2 及其标准误的估计量性能。最后，我们通过构建置信区间，并基于基因表达数据比较用于预测定量甘蓝型油菜 和玉米 表型的模型，展示了这一新方法。

关键词：预测、决定系数、标准误、交叉验证、自助法

*S.H 谨此感谢 InariAgriculture NV 提供的资助；W.W. 获得了佛兰德政府根据“佛兰德人工智能研究计划”提供的资助。

1 引言

预测模型的性能通过损失函数进行评估，损失函数用于量化预测值与观测值之间的差异。对于定量结果，最常用的损失函数是平方误差损失，其定义为 $l(y_i,\ m(\mathbf{x}_i)) = (y_i - m(\mathbf{x}_i))^2$, 其中，$y_i$ 表示样本 i=1,. .. ,n 中的观测结果，n 表示样本量，而 $m(\mathbf{x}_i)$ 表示模型 m 使用一组预测变量 $\mathbf{x}_i$ 得出的结果预测值。通常会报告损失分布的汇总统计量，例如均方误差（MSE），其估计值为 $n^{-1}\sum_{i=1}^{n} l(y_i,\ m(\mathbf{x}_i))$（Hastie 等，2009）。然而，与针对分类结果的误分类损失不同，MSE 不易解释，因为它取决于所研究数据集的变异程度以及所用的测量单位。因此，MSE 常与样本中的总方差进行比较：$n^{-1}\sum_{i=1}^{n} l(y_i,\ \bar{y})$，其中 $\bar{y}$ 为结果变量的平均值；这里将其称为总平方均值（MST），以与 MSE 相对应。由此得到决定系数，即 R^2，其定义如下（Kvålseth，1985）

$$R^2 = 1 - \frac{MSE}{MST} \tag{1}$$

R^2 没有单位，因此可用于比较单位不同的结果变量（Valbuena 等，2019年）。它常被用作评估模型对给定样本拟合优度的统计量。在这种情况下，值m（$\mathbf{x}_i$）和 $\bar{y}$ 由使用整个样本（包括观测值 i）训练的模型进行预测，因此 R^2的取值范围为 0 到 1，可解释为模型所解释的方差比例（Anderson- Sprecher，1994）。人们提出了多种修正方法，以惩罚一般线性模型的复杂度，例如调整后的 R^2 （Wherry，1931），从而更好地指导模型构建。 R^2 也已扩展至广义线性模型（Zhang，2017；Cameron 和 Windmeijer，1997；Nagelkerke 等，1991）和生存分析（Verweij 和 Houwelingen，1993）。对于线性模型，研究者已推导出样本内 R^2 的标准误和置信区间（Cohen、West 和 Aiken，2014）。

然而，这些 R^2 值旨在作为单个数据集内 的拟合优度诊断和模型构建辅助工具。现代预测模型，例如随机森林（Breiman，2001），灵活性更高，并且与线性模型相比，对结果变量与预测变量之间关联类型的假设更少。此外，线性模型也可应用于预测变量远多于观测值的高维组学数据集。由于这两种情况都可能导致过拟合，因此在这种情况下，样本内损失无法很好地衡量预测性能；预测性能应在未用于训练模型的数据上评估。理想情况下，这些数据是独立测试数据集，但更常见的是，使用交叉验证（CV）（Bates、Hastie 和 Tibshirani，2021）或自助法（Efron 和 Tibshirani，1997）等数据拆分算法，在同一数据集上估计样本外损失。这样，损失是在数据集的另一部分上估计的，而这部分数据并未用于构建模型。因此，前述关于样本内 R^2 及其标准误和置信区间的结果不再成立。例如，样本外 R^2 值落在区间$]-\infty,\ 1]$内，而非样本内 R^2的 $[0,\ 1]$ 。不过，要进行统计推断，就需要对样本外 R^2 统计量作出形式化定义，并为点估计^R^2提供某种不确定性度量。在此，我们将样本外 R^2 定义为对当前预测模型与基准模型进行比较；

该基准预测模型忽略协变量信息；我们为其提供无偏估计量，并利用数据拆分算法领域的最新进展（Bates等，2021），给出样本外$\hat{}$ R^2的标准误（SE）。我们通过模拟研究验证样本外 R^2 及其标准误的估计量。随后，我们展示如何利用这些估计量比较结果变量的可预测性，并基于甘蓝型油菜 和玉米 田间试验的真实组学数据集构建置信区间。

2 样本外 R^2：定义与推断

2.1 将样本外 R^2 用作模型比较

我们主张将样本外 R^2 视为两个样本外预测模型的比较，正如常规样本内 R^2 是两个样本内模型的比较一样（Anderson-Sprecher，1994；Campbell 和 Thompson，2008）。设 m_d 为在数据集 $\mathbf{d} = (\mathbf{y}, \mathbf{x})$ 上训练得到的预测模型，该模型可根据 $\mathbf{x}$ 对新的 $\mathbf{y}$ 进行预测：$m_d(\mathbf{x})$。为了评估这些预测模型 m，我们关注假设样本外观测 Y_{oos} 相对于其预测值 $m_D(\mathbf{X}_{oos})$ 的期望平方误差损失，并对所有可能的、固定大小为 n 且用于估计 m 的数据集 $\mathbf{D} = (\mathbf{Y}, \mathbf{X})$ 求平均。更正式地说，我们要寻找

$$E_{\mathbf{D}}\left[E_{(Y,\mathbf{X})_{oos}}\left(\left[Y_{oos} - m_D(\mathbf{X}_{oos})\right]^2 \mid \mathbf{D}\right)\right], \tag{2}$$

内层期望是对所有样本外观测求取，外层期望则是对所有数据集求取。我们考虑一种常见情形：没有可用的独立样本外观测测试集，因此所有估计都必须基于唯一一个大小为 n 的观测数据集 $\mathbf{d}$ 进行。在此处及下文中，所有未带下标 oos 的变量都属于 $\mathbf{d}$。

比较中的第一个模型称为零模型（Anderson-Sprecher，1994），它简单地使用观测数据的$\bar{}$结果变量平均值 $\bar{Y}$ 作为所有样本外观测的预测，而忽略可用的预测变量。该模型的预期损失即样本外均方总误差，可根据如下推导，利用结果变量值向量 $\mathbf{y}$ 从$\bar{}$解析地估计，其依据是等式 $\mathrm{E}(Y) = \mathrm{E}(\bar{Y}) = \mathrm{E}(Y_{oos})$：

$$\begin{aligned}
MST &= E_{\mathbf{D}}\left[E_{Y_{oos}}\left(\left[Y_{oos} - \bar{Y}\right]^2 \mid \mathbf{D}\right)\right] \\
&= E_{\mathbf{D}}\left[E_{Y_{oos}}\left(\left[(Y_{oos} - E(Y)) + (E(Y) - \bar{Y})\right]^2 \mid \mathbf{D}\right)\right] \\
&= E_{\mathbf{D}}\left[E_{Y_{oos}}\left([Y_{oos} - E(Y)]^2 + \left[E(Y) - \bar{Y}\right]^2 + 2(Y_{oos} - E(Y))(E(Y) - \bar{Y}) \mid \mathbf{D}\right)\right] \\
&= E_{\mathbf{D}}\left[\mathrm{Var}(Y) + \left(E(\bar{Y}) - \bar{Y}\right)^2 + 2E_{Y_{oos}}\left((Y_{oos} - E(Y))(E(Y) - \bar{Y}) \mid \mathbf{D}\right)\right] \\
&= E_{\mathbf{D}}\left[\mathrm{Var}(Y) + \left(E(\bar{Y}) - \bar{Y}\right)^2 + 2(E(Y) - \bar{Y})E_{Y_{oos}}\left(Y_{oos} - E(Y_{oos}) \mid \mathbf{D}\right)\right] \\
&= \mathrm{Var}(Y) + \mathrm{Var}(\bar{Y}) + 0 = \frac{n+1}{n}\mathrm{Var}(Y).
\end{aligned} \tag{3}$$

这一结果很好地说明，期望损失是估计量围绕

期望值的波动与观测值围绕期望值的波动之和。总体方差 Var(Y) 的无偏估计量为 $(n-1)^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$。随后，样本外均方总误差的估计量将这一估计量扩大 (n+1)/n 倍，以考虑通过 Y 估计 E(Y)¯ 时所产生的波动。

比较中的第二个模型是利用协变量信息进行预测的预测模型。由于这是一个更复杂的模型，其期望样本外损失（均方误差）没有解析表达式，因此在缺少独立测试数据的情况下，需要通过数据拆分算法进行估计。这里我们讨论交叉验证和 .632自助法（Efron等，1997），但也可以采用其他方法，例如将可用数据一次性拆分为训练数据集和测试数据集。

在交叉验证（交叉验证）中，将 d = (y, x) 的样本划分为 K 个大小相等、各含 n/K 个观测值的折。为简化起见，假设 K 能整除 n。第 $k=1$，.....，K 折中的样本集合记为 $\mathbf{d}_k$，其他观测值记为 $\mathbf{d}_{\backslash k}$。对于所有 k，在 $\mathbf{d}_{\backslash k}$ 上训练模型，得到模型 $m_{k\circ}$ 该折的平方误差损失估计为 $\frac{K}{n}\sum_{i\in\mathbf{d}_k}(y_i-m_k(\mathbf{x}_i))^2,$，总体均方误差估计则为$n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathbf{d}_k}(y_i-m_k(\mathbf{x}_i))^2$。重复折叠划分可降低估计值 $\widehat{\quad}$ $\widehat{\quad}MSE$的变异性；最终估计值为 MSE各次重复划分所得结果的平均值。上述过程即为我们所说的简单交叉验证。关于嵌套交叉验证，可参见 Bates 等人（2021）；他们给出了 Var($\widehat{MSE}$) 的估计，并针对 m_k 是在大小为 $n\frac{K-1}{K}$ 而非 n 的数据集上估计这一事实提供了校正。

.632自助法是另一种估计均方误差的方法（Efron等，1997）。在这种情况下，通过对 n 样本进行有放回重抽样来划分样本。一个样本被包含在该自助样本中的概率为 $1-\exp(-1)\approx 0.632$，因此得名。此重抽样步骤重复 B 次，其中$\mathbf{d}_b$ 表示纳入样本的集合（同一个样本可能出现多次），m_b 表示在该集合上训练的模型，而$\mathbf{d}_{\backslash b}$表示包含未纳入样本的唯一值的集合，其中 $b=1,\ldots,B_{\circ}$ 令N_i^b为样本 i被纳入$\mathbf{d}_b$的次数，并令 $J_i^b=I(N_i^b=0)$，其中 I() 为指示函数。那么，均方误差的 .632 自助法估计为 $n^{-1}\exp(-1)\sum_{i=1}^{n}\left((y_i-m_d(\mathbf{x}_i))^2\right)+n^{-1}(1-\exp(-1))\sum_{i=1}^{n}\left[\frac{\sum_{b=1}^{B}(J_i^b(\quad y_i-m_b\quad i\quad ^2}{\sum_{b=1}^{B}J_i^b}\right](\mathbf{x})))$，其中 m_d是在所有样本上训练得到的。因此，自助法 .632 估计是样本内误差和样本外误差的加权平均，但也可以写成对所有样本的加权求和。Efron等人（1997）给出了 .632 估计的标准误，以及该估计量的若干变体。Bates 等人（2021）表明，交叉验证和 .632 自助法确实估计了式（2）中的量；众所周知，交叉验证是无偏的，而自助法在均方误差估计中存在轻微偏差（Braga-Neto和Dougherty，2004；Jiang和Simon，2007；Kohavi，2001；Molinaro、Simon和Pfeiffer，2005）。样本外

ple R^2 随后是一个 population parameter 定义为

$$R^2=1-\frac{E_{\mathbf{D}}\left[E_{Y_{oos}}\left([Y_{oos}-m_D(\mathbf{X}_{oos})]^2\mid\mathbf{D}\right)\right]}{E_{\mathbf{D}}\left[E_{Y_{oos}}\left([Y_{oos}-\bar{Y}]^2\mid\mathbf{D}\right)\right]},\tag{4}$$

并取决于用于训练预测模型的数据的样本量 n $\mathbf{D}$、预测模型，以及结果变量与预测变量的联合分布。样本外 R^2 的值反映了

零模型与更复杂的预测模型之间的比较：当其小于 0 时，零模型的样本外预测效果最佳；当其大于 0 时，更复杂的预测模型实现了最低的样本外平方误差损失。在后一种情况下， R^2 可解释为零模型的平方误差损失中由更复杂模型所解释的比例。为估计样本外 R^2，我们将前文所述基于观测数据得到的样本外平方误差 MST 和均方误差估计量代入：

$$\hat{R}^2 = 1 - \frac{\widehat{MSE}}{\widehat{MST}} = 1 - \frac{\widehat{MSE}}{(n+1)/(n(n-1))\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5}$$

这一表达式可视为 Campbell 等人（2008）提出的样本外预测 R^2 的预测等价形式。

2.2 R^2 的合并与平均估计量

在上文的公式（5）中，先计算所有观测值的平方误差损失并求和，再将其合并为最终的 R^2 估计。这在多变量损失函数估计领域（例如曲线下面积（AUC）估计）中称为合并策略，与平均方法相对；在后者中，分别计算每个留出折的性能度量，然后对各折的结果取平均，以得到最终估计（Bradley 1997年）。后一种方法不适用于均方误差的 .632 自助法估计，但可用于交叉验证。在这种情况下，分别对每个留出折的均方误差和均方总和进行估计，具体如下。每折的均方误差按第 2.1 节的方法估计。对于均方总和，我们考虑以下两种估计策略：一种是按照（3）基于训练折进行估计，并将 n 调整为训练折的样本量；另一种是将留出折 $\left[\sum_{i=1}^{n} I\ (i\in\mathbf{d}_k)\right]^{-1}\sum_{i\in\mathbf{d}_k}(y_i - \bar{y}_{\mathbf{d}_k})^2$ 的经验方差作为估计，其中 $\bar{y}_{\mathbf{d}_k}$ 为留出折$\mathbf{d}_k$的平均值，而 $I()$ 为指示函数（Valbuena等人 2019年）。两种情况下，都会将这些估计分别合并为每折的 R^2 估计，然后再取平均。我们将这两种估计量分别称为“使用训练集均方总和的平均 R^2 ”和“使用测试集均方总和的平均 R^2 ”。

ˆ2.3 R^2的标准误

作为估计量ˆR^2不确定性的一种度量，我们推导其标准误（SE）$\left(\text{SE}(\hat{R}^2) = \text{Var}(\hat{R}^2)^{1/2}\right)$的表达式。为此，我们依赖估计量ˆ R^2的渐近正态性。根据一阶Delta方法（Gauss 1823年），

$$\text{Var}(\hat{R}^2) \approx \left(\nabla R^2\right)^T \begin{bmatrix} \text{Var}(\widehat{MSE}) & \text{Cov}(\widehat{MSE}, \widehat{MST}) \\ \text{Cov}(\widehat{MSE}, \widehat{MST}) & \text{Var}(\widehat{MST}) \end{bmatrix} \nabla R^2. \tag{6}$$

梯度表示为 $\left(\nabla R^2\right)^T = \left(\frac{\partial R^2}{\partial MSE}, \frac{\partial R^2}{\partial MST}\right) = \left(-\frac{1}{MST}, \frac{MSE}{MST^2}\right)$，并在估计值 $\widehat{MSE}$和 $\widehat{MST}$处进行计算。对于均方误差的交叉验证估计，MSE 的方差估计由

Bates 等人（2021）给出；对于均方误差的 .632 自助法估计，则由 Efron 等人（1997）给出。$\widehat{MST}$ 的方差估计为 $^2{}_{(n-1)}\widehat{MST}^2$（Harding、Tremblay 和 Cousineau，2014）。

协方差 Cov($\widehat{MSE}$, $\widehat{MST}$) 通常相当大且为正，因为均方误差和 MST 是在同一个结果向量 **y** 上估计的。它可以分解为 Cov($\widehat{MSE}$, $\widehat{MST}$) =Cor($\widehat{MSE}$, $\widehat{MST}$) $\left(\text{Var}(\widehat{MSE})\text{Var}(\widehat{MST})\right)^{1/2}$。Cor($\widehat{MSE}$, $\widehat{MST}$) = ρ无法通过解析方法推导，因此可通过非参数或参数自助法，或者如下所述通过刀切法进行估计。大小为 n 的自助样本可以通过非参数方式从数据 $\mathbf{d}$ 中有放回地抽取条目获得，也可以通过假设一个参数模型，利用数据估计相应参数，再从由这些估计值参数化的模型中抽取 $\mathbf{Y}$ ，同时固定 $\mathbf{x}$ ，以参数方式获得。随后，按照与原始样本相同的方式，对每个自助样本估计均方误差和 MST。在刀切法中，依次删除每个观测值，并在剩余的样本（样本量为 $n-1$）上估计均方误差和 MST，从而得到数量等于样本量 n 的刀切估计。对于自助样本和刀切样本，为减少计算时间，均方误差估计采用简单交叉验证而非嵌套交叉验证，但交叉验证折的划分仍按照在观测数据集上进行均方误差估计时的方式重复进行。均方误差和 MST 的自助法或刀切法估计量之间的经验相关系数，被用作均方误差与 MST 估计量之间相关系数的估计值。当使用 .632自助法估计均方误差，并结合自助法估计均方误差与 MST 之间的相关系数时，我们将前者称为内层自助样本，将后者称为外层自助样本。

2.4 关于 R^2 的推断

给定标准误估计，可以使用单侧近似 **z** 检验来检验零假设 $R^2 \leq 0$，即判断预测模型是否显著优于零模型。标准误的另一种常见用途是构造置信区间，这同样依赖于估计量 R^2 的渐近正态性。置信区间构造如下：

$$\hat{R}^2 \pm \widehat{\text{Var}}(\hat{R}^2)^{1/2}\Phi^{-1}\left(\frac{\alpha}{2}\right) \tag{7}$$

其中， α 为显著性水平， Φ^{-1} 为标准正态分布的逆累积分布函数。置信区间的上界截断为 1。

作为式（6）中德尔塔法标准误的替代方法，可以将 R^2 估计值在非参数或参数自助样本中的标准差（以与估计 ρ相同的方式获得）用作标准误的估计值（下文称为自助法标准误）。作为基于德尔塔法标准误的置信区间的替代方法，我们考虑基于自助法 R^2 估计值分布构造的百分位数法和偏差校正加速（BCa）自助法置信区间（DiCiccio和Efron 1996），以及以与式（7）中德尔塔法标准误相同的方式，利用自助法标准误构造的置信区间。我们选择后一种方法，而非bootstrap-t方法（DiCiccio等人

1996），因为后者需要为每个自助样本实例计算标准误，因此计算量过大。

3 模拟研究

我们开展了一项模拟研究，在一维预测模型（OLS）和高维预测模型（EN）上应用所提出的 R^2 估计方法。我们研究估计量对 MST 和均方误差（MSE）的表现，并比较用于 R^2的合并与平均估计量。随后，我们将针对ˆ R^2 的德尔塔法标准误（SE）估计量及其相应置信区间，与基于自助法标准误以及百分位数法和偏差校正与加速法置信区间的竞争方法进行比较。

3.1 设置

3.1.1 数据生成

在模拟研究中，观测值 Y_i 取自以下模型:

$$Y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2). \tag{8}$$

$\mathbf{X}$是 $n \times p$设计矩阵， $\boldsymbol{\beta}$是长度为p的系数向量； $\sigma^2 = 1$是残差方差。我们研究一维情景（p=1）和高维情景（p=1,000）。$\mathbf{X}$ 的所有元素均独立取自标准正态分布。对于一维情景，我们考虑以下样本量n: 20、30、50 和 100，并将 β 设为 0、0.5、1 或 1.5。对于高维情景，我们考虑样本量 30、50、75 和 100，并将 $\boldsymbol{\beta}$ 的前 10 个元素均设为 0、0.5 或 1；其余元素均为 0。在一维情景中，生成 1,000 个蒙特卡洛数据集实例；在高维情景中生成 100 个。

3.1.2 预测模型

在一维情景中，使用$stats$ R 包中的$lm.fit$ 函数，通过普通最小二乘法（OLS）预测结果变量。在高维情景中，使用$glmnet$ R 包中的$glmnet$ 函数，通过弹性网（EN）（Zou和Hastie，2005）预测结果变量，混合参数固定为 0.5。惩罚参数通过内层循环中的 10折交叉验证进行调整，具体实现使用$cv.glmnet$ 函数。

3.1.3 均方误差及其标准误

样本外均方误差采用 Bates 等人（2021）的方法，通过10折交叉验证估计，或采用 Efron 等人（1997）的方法，通过.632自助法估计。相应的标准误分别通过嵌套交叉验证（含9个内层折）（Bates 等，2021）或经验影响函数（Efron 等人（1997））计算。按照 Bates 等人（2021）的建议，重复划分交叉验证折，划分次数为：一维情景下1、25，

100或200次，高维情景下5、25或100次。.632自助法估计均方误差所用的自助样本数分别设为25、100和200。通过自助法估计均方误差与MST估计量之间的相关性时，使用了10、50、100或500个自助法实例。

3.1.4 真值与诊断

仅凭生成模型 (8) 无法得知真实的样本外均方误差，因为它取决于对 m 的估计精度。我们通过蒙特卡洛模拟对其进行近似：根据 (8) 生成 5,000 个数据集，这些数据集使用与第 3.1.1 节相同的样本量数组 n 和系数 β；随后，将预测模型拟合到这些数据集，并在从 (8) 抽取、样本量为 10,000 的独立测试数据集上评估其预测性能。由此得到 5000 个精确的均方误差估计值，其平均值可作为真实样本外均方误差的近似值。随后，结合近似得到的真实样本外均方误差与由 $\sigma^2\ (n+1)\ n$ 给出的真实 MST，使用 (4) 近似该参数设置和预测方法下真实的样本外 R^2（见表 S1 和 S2）。对于所有近似真实 R^2 值小于 0 的情形，使用单侧近似 z 检验检验 $R^2 \leq 0$；将零假设被拒绝的比例作为第一类错误率的估计值。显著性水平设为 5%。95% 置信区间的覆盖率近似为置信区间包含近似真实 R^2的蒙特卡洛实例所占的比例。此外，还计算了置信区间的平均宽度。

$\hat{}$ R^2 的真实标准误采用不同方式近似，因为它也取决于数据拆分算法及其设置（交叉验证拆分次数或自助抽样次数）。对于第 3.1.1 节中提到的所有蒙特卡洛实例，取相关算法和参数设置下 R^2 值的标准差$\hat{}$作为近似值，见补充图 S1-S3、S20-S21。 MSE与 MST之间的真实相关性 ρ则以相同方式近似，即计算这两个量在相同蒙特卡洛实例中的经验相关系数。通过计算估计标准误与近似真实标准误之比，并对所有蒙特卡洛实例取几何平均数，来评估不同标准误估计量的准确性。此外，还计算了估计标准误相对于近似真实标准误的均方误差，该指标同时反映估计量的偏差和方差。通过绘制估计值$\hat{}\rho$的箱线图，以图形方式评估 ρ 的估计准确性。数据生成、 R^2 估计和评估的完整流程见图 1。

3.2 结果

3.2 R^2 的组成部分：MST 和均方误差

图S7通过数值结果展示了MST估计量(3)的无偏性。在一维情景中，通过.632自助法进行的均方误差估计存在向下偏差，而Bates 等人（2021）提出的经过偏差校正的交叉验证估计是无偏的（补充图S5和S22）。该

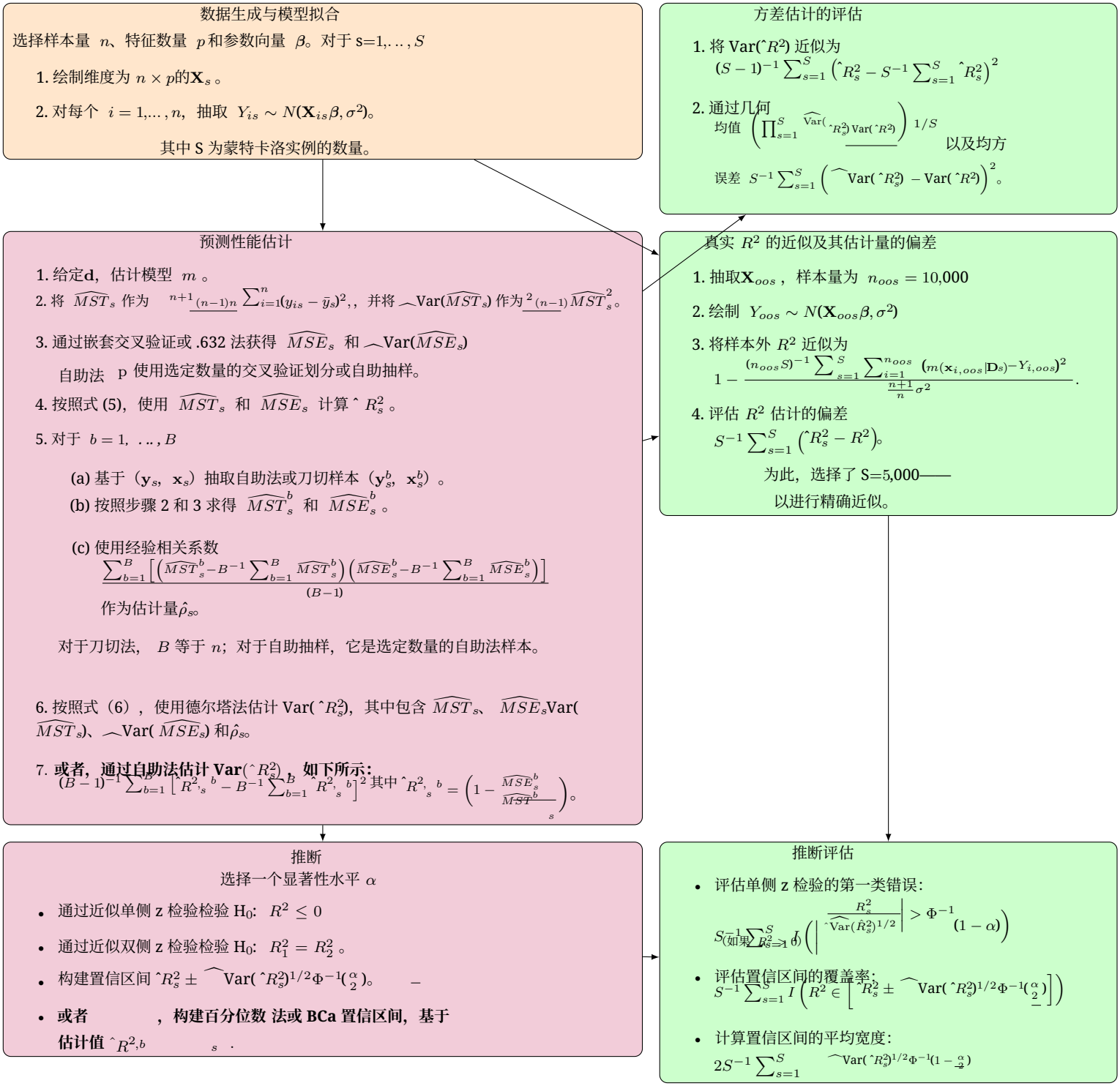

图 1：模拟研究的示意图。左上方的面板（橙色）描述重复的数据生成过程。中央和左下方的面板（紫色）描述对 R^2 及 Var($\hat{R}^2$) 的估计，以及基于模拟数据或真实数据的统计推断。右侧面板（绿色）描述对真实 R^2 及 Var($\hat{R}^2$) 的近似，以及性能评估；只有在已知真实值的模拟数据中才能进行性能评估。箭头表示一个模块的输出作为另一个模块的输入。粗体文字表示偏离本文所提方法的竞争方法。

$\widehat{\text{通过交叉验证或 .632自助法估计 Var}}(MSE)$ 时，在样本量较小时会有轻微的向上偏差，但随着样本量增大，这种偏差会减小（补充图 S8-S9）。

对于高维情景，补充图 S22 显示，在样本量较大且效应量较大时，均方误差估计存在向上偏差。这种偏差是因为在嵌套交叉验证方案中，k 折交叉验证使用大小为 $n\frac{k-2}{k}$ 的数据集训练模型，而不是使用样本量为 n 的数据训练模型；交叉验证试图估计的正是后者的性能。Bates 等人（2021）在附录 C 中提出的偏差校正方法在一维情形下效果良好，如图 S5 所示，但在高维情形下失效；这可能是因为正如 Bates 等人（2021）所提到的，我们处于比例稀疏情形，且 n 和 p 均趋于无穷。增加折数或许可以减小这种偏差（Bates 等，2021）。

3.2.2 R^2 的合并估计与平均估计

一维情景下，不同备选估计量对 R^2 的估计表现如补充图 S4 所示。合并 R^2 是对式（4）所定义真实 R^2 的无偏估计量，且方差较低。采用训练集均方总和的平均 R^2 方差也较低，但在信号强度较弱时存在向下偏差。采用测试集均方总和的平均 R^2 估计量波动很大，在样本量较小时存在显著的向下偏差；即使样本量为 100，仍有部分偏差。因此，我们选择使用合并估计量（5）。补充图 S36-S39 评估了该合并估计量 $\hat{R}^2$ 的正态性假设；德尔塔法对标准误的近似以及置信区间的构造均要求这一假设。可以看到一些偏离正态性的情况，尤其是在样本量较小时。

3.2.3 一维情景

$\widehat{\text{一维情景下交叉验证的 SE}}(R^2)$ 估计准确性见图 2 和补充图 S10。德尔塔法标准误的偏差大多为正（即偏保守），且随效应量增大而减小。非参数自助法在估计 $\widehat{\text{相关性 }}\rho$（$MSE$ 与 MST 之间）方面表现最佳（补充图 S11）。在零假设场景下，自助法标准误比德尔塔法标准误更准确，但随着信号强度增强，自助法标准误往往出现向下偏差（即偏宽松），参数自助法尤为明显。对于采用非参数自助法或刀切法估计 ρ的德尔塔法标准误，置信区间的覆盖率（图 2）接近名义水平；但基于自助法（自助法标准误、百分位数法和偏差校正与加速法）的区间在某些场景下覆盖不足。当预测变量完全不能预测结果变量时，除偏差校正加速置信区间外，所有置信区间的覆盖率均高于 95% 的名义水平。德尔塔法标准误置信区间通常比自助法置信区间更宽（图 S12）。所有方法均能将近似单侧 z 检验的第一类错误控制在 5% 的显著性水平以下（图 S13）。.632自助法也得出了类似结论（图 S14-S19），尽管其在 R^2 估计中略有向上偏差（图 S6）。

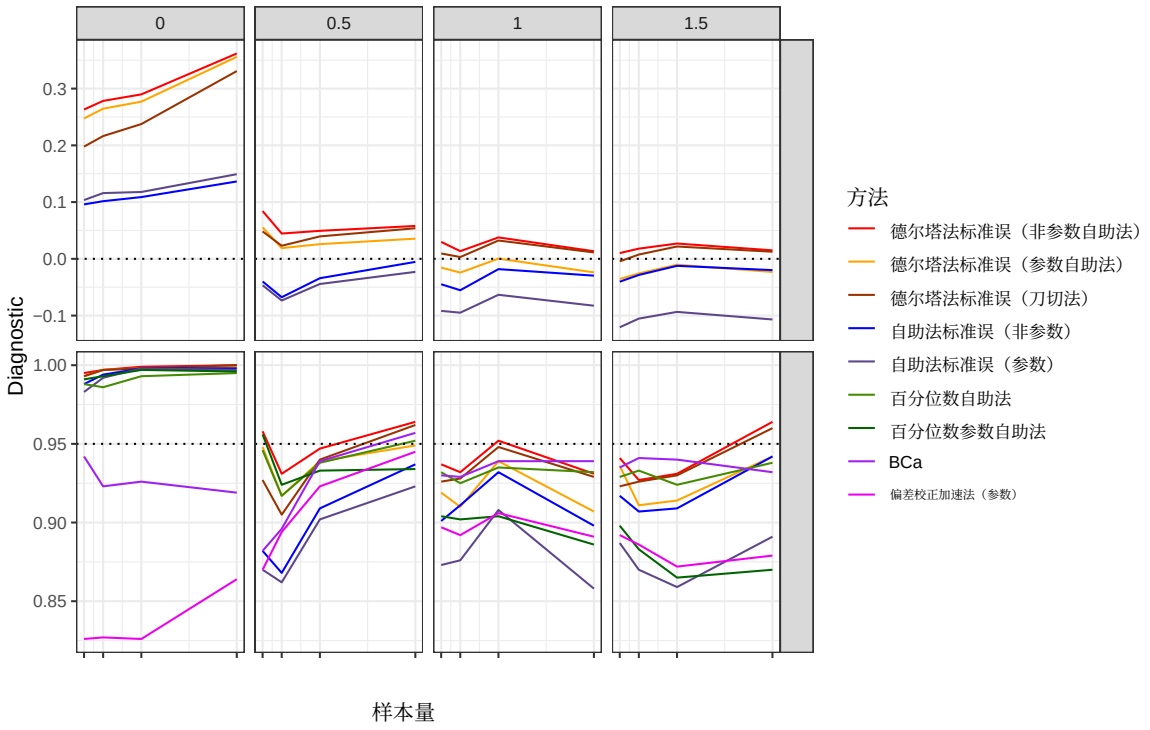

图 2：使用交叉验证的一维模拟场景的诊断： R^2 估计值与近似真实标准误（SE）之比的几何平均值的 log10（上方各图），以及置信区间的覆盖率（下方各图），分别随估计方法（颜色）、样本量（x 轴）和效应量（各列）变化；共进行了 500 次自助抽样，并对交叉验证划分重复了 200 次。虚线分别表示无偏 R^2 估计和 95% 的名义覆盖率。

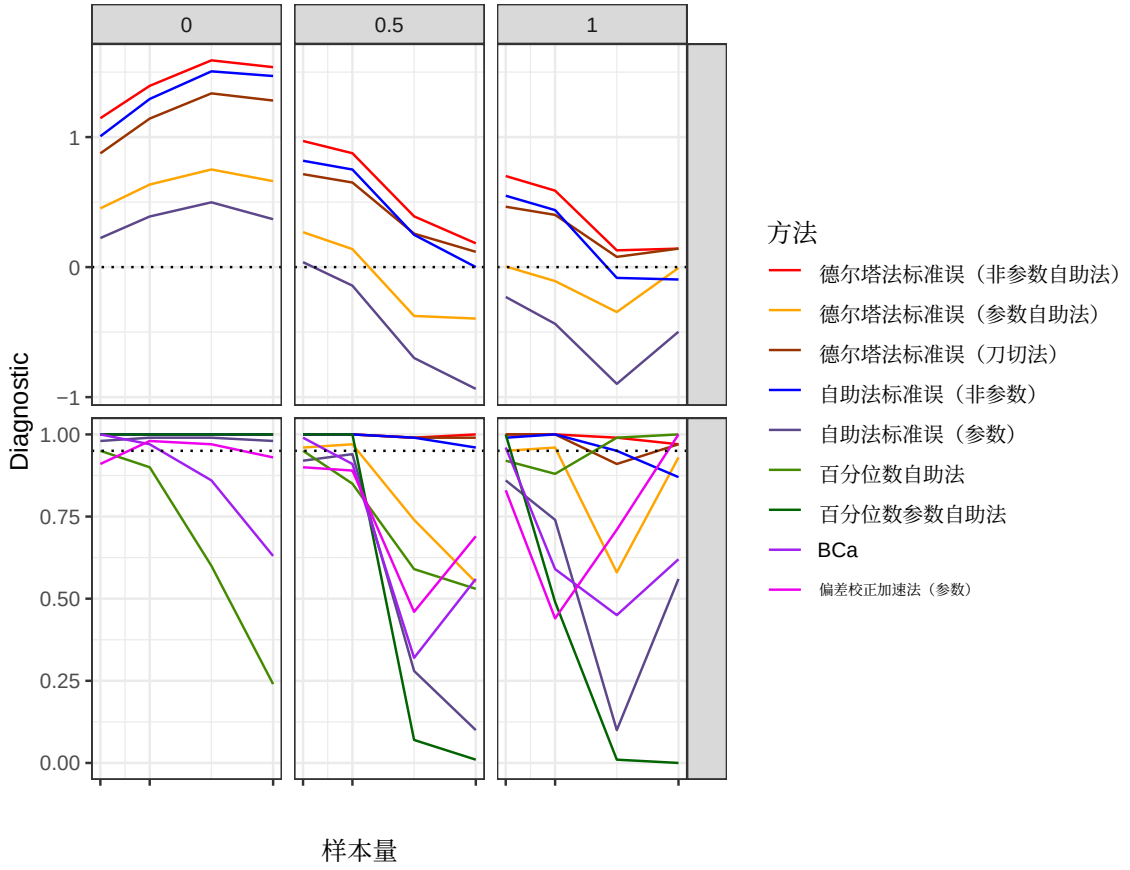

图3：使用交叉验证进行高维模拟场景的诊断：估计得到的 R^2 标准误（SE）与近似真实标准误之比的几何平均数的 log10 值（上方各图），以及置信区间的覆盖率（下方各图），随估计方法（颜色）、样本量（横轴）和效应量（各列）变化的情况；基于100次自助抽样和100次交叉验证划分重复。虚线分别表示无偏 R^2 估计和95%的名义覆盖率。

3.2.4 高维情景

在高维情景中使用交叉验证时，对于预测价值为零的零假设场景，参数自助法对 R^2 的标准误估计最佳；但对于信号强度较高的情况，尤其是样本量较大时，它会低估标准误（图3和补充图 S24）。随着样本量和效应量增加，采用非参数自助法或刀切法估计 ρ 的德尔塔法标准误，以及非参数自助法标准误表现最佳，并逐渐收敛于真实值（补充图 S26）。这些发现也反映在置信区间的覆盖率上：除百分位数自助法和偏差校正加速置信区间外，所有方法在零假设场景下的覆盖率均高于 95% 的名义水平。随着信号强度增加，只有采用非参数自助法或刀切法进行 ρ 估计的德尔塔法标准误置信区间，以及非参数自助法标准误置信区间，能够保持接近名义水平的覆盖率（图3）；不过，其他方法覆盖率较低，部分原因是上述样本量较大时均方误差估计存在偏差（补充图 S22–S23）。对于每个样本量，置信区间都远宽于一维情景下的置信区间（补充图 S28）。自助法标准误方法无法将第一类错误控制在显著性水平内（补充图 S29）。对 R^2 进行 .632 自助法估计时也发现了类似结果（补充图 S30–S35）。

在低维和高维情景中，当预测变量不具备预测价值时，MSE 与 MST 之间的相关性接近 1；但随着效应量增大，相关性会减弱，有时甚至变为负值（补充图 S11、S16、S26 和 S32）。负相关可能是设计矩阵随机性所致：预测变量变异性更大的设计会使结果变量的方差增大（MST较高），但也会带来更好的参数估计值，从而提升预测效果（MSE较低）。

4 案例研究：甘蓝型油菜和玉米表型的可预测性

对 62 株甘蓝型油菜 植株（De Meyer 等，2022）和 60 株玉米 植株（Cruz 等，2020）测量了基因表达和表型。每种作物均选取 5 种表型进行预测：甘蓝型油菜的第8叶宽度、分枝数、叶片数、根宽和种子数，以及玉米的叶片长度、叶片宽度、苞叶长度、穗长和株高。分析前，对基因表达测量值进行了$rlog$ 转换（Love、Huber 和 Anders，2014）。仅保留表达量最高的 5,000 个基因用于模型拟合。使用与模拟研究相同的设置（见第 3.1.2 节），通过弹性网根据这 5,000 个基因预测结果变量表型，并采用 10 折交叉验证，将数据随机划分为折并重复 100 次，以估计相应的均方误差。根据第 2.3 节所述的德尔塔法，并结合对 ρ的刀切法估计，计算了所得 R^2 的标准误。另使用 50 个自助样本，通过非参数自助法估计标准误。进行了单侧近似 z 检验，以检验 $R^2 \leq 0$，并构建了置信区间。

甘蓝型油菜和玉米数据的估计 R^2 值、标准误、p值和置信区间分别见表1和表2。对于甘蓝型油菜，第8叶宽度和种子数的 R^2

根据基于德尔塔法标准误的单侧近似 z 检验，其与 0 的差异显著；对于玉米，刀片16宽度根据该检验也具有显著性。与模拟结果类似，自助法标准误大多较小，因此置信区间更窄、P值更小，但相应显著性检验的结论相似。

	$\hat{R}^2$	德尔塔法				自助法			
		SE	P值	2,5%	97,5%	SE	P值	2,5%	97,5%
第8叶宽度	0.72	0.07	8.06e-25	0.58	0.86	0.09	4.06e-15	0.54	0.91
总分枝数	0.11	0.17	2.61e-01	-0.22	0.44	0.16	2.58e-01	-0.22	0.43
叶片数	0.19	0.25	2.29e-01	-0.30	0.67	0.15	1.16e-01	-0.12	0.49
根系宽度	-0.02	0.17	5.46e-01	-0.35	0.31	0.22	5.35e-01	-0.45	0.41
种子数	0.42	0.20	1.77e-02	0.03	0.82	0.11	3.97e-05	0.21	0.63

表 1：5 个甘蓝型油菜表型的估计 R^2，以及相应的德尔塔法和非参数自助法标准误（SE）、单侧 p 值和 95% 置信区间上下限。

	$\hat{R}^2$	德尔塔法				自助法			
		SE	P值	2,5%	97,5%	SE	P值	2,5%	97,5%
刀片16长度	0.30	0.23	9.60e-02	-0.15	0.74	0.16	3.15e-02	-0.02	0.61
刀片16宽度	0.49	0.21	1.08e-02	0.07	0.90	0.09	1.44e-07	0.30	0.67
苞叶长度	0.30	0.29	1.49e-01	-0.27	0.87	0.19	5.86e-02	-0.08	0.68
穗长	-0.02	0.18	5.41e-01	-0.38	0.34	0.24	5.32e-01	-0.48	0.44
株高	-0.01	0.15	5.36e-01	-0.30	0.27	0.19	5.28e-01	-0.38	0.35

表2：5个玉米 表型的估计 R^2、相应的德尔塔法标准误和非参数自助法标准误（SE）、单侧P值以及95% 置信区间的下限和上限。对于刀片16长度，采用自助法标准误时，P值显著，但置信区间包含0；这是因为P值为单侧检验，而置信区间为双侧。

一个相关的科学问题是比较同一数据集内不同表型的 R^2 估计值。如果设计矩阵是固定的，不同估计值将相互独立，差值的方差便等于两个 R^2 估计值方差之和。然而，如果设计矩阵是随机的（如这里的基因表达测量值），两个表型的均方误差估计值之间就会相关。两个不同表型 a和 b的 R^2 之差的估计量的方差因此等于 $\mathrm{Var}(R_a^2 - R_b^2) = \mathrm{Var}(R_a^2) + \mathrm{Var}(R_b^2) - 2\mathrm{Cor}(R_a^2, R_b^2)\sqrt{\mathrm{Var}(R_a^2)\mathrm{Var}(R_b^2)}$。相关性 $\mathrm{Cor}(R_a^2, R_b^2)$ 通过自助法估计：从基因表达矩阵中有放回地抽取样本50次，并使用表型 a 和 b 的相应条目估计 R^2。随后，将这50个自助法估计值的经验相关系数用作 $\mathrm{Cor}(R_a^2, R_b^2)$ 的估计值。表3展示了使用德尔塔法标准误计算的检验统计量和P值，用于近似双侧 z 检验；该检验的零假设是甘蓝型油菜 的 R^2 值之差为零。第8叶宽度^ 是唯一一个 R^2 与其他某些表型显著不同的表型：总分枝数、叶片数和根系宽度。

	第8叶宽度	总分枝数	叶片数	根系宽度
总分枝数	-2.98 (0.0029)			
叶片数	-2.39 (0.017)	0.27 (0.79)		
根系宽度	-2.45 (0.014)	-0.39 (0.7)	-0.61 (0.54)	
种子数	-1.59 (0.11)	1.69 (0.091)	0.88 (0.38)	1.41 (0.16)

表3：列和行中甘蓝型油菜表型 R^2 差异的近似双侧 z 统计量，括号内为使用德尔塔法标准误计算的相应 P 值。例如，左上角的数值为总分枝数 R^2 减去第8叶宽度 R^2，再除以该差异的标准误估计。

另一个值得研究的问题是，不同物种的某些表型的可预测性是否存在差异。由于这些估计量是基于独立数据集计算的，因此两个 R^2估计量之差的方差就是它们方差之和。因此，使用德尔塔法标准误，甘蓝型油菜第8叶宽度与玉米叶片宽度的 R^2 值之差的近似双侧 z 统计量为 $\frac{0.72-0.49}{\sqrt{0.07^2+0.21^2}} = 1.06$，结果不显著（P值为 $= 0.28$）。然而，甘蓝型油菜第8叶宽度与玉米株高之间可预测性的标准化差异$\frac{0.72-(-0.01)}{\sqrt{0.07^2+0.15^2}} = 4.52$ 确实显著（P值为 $=6.2 \times 10^{-6}$）。

5 讨论

对类似于 R^2的度量的研究通常聚焦于样本内表现，但 R^2 也经常被用于评估样本外预测。我们在此正式将样本外 R^2 定义为 1 减去特定预测模型的预测损失与忽略协变量信息的零模型预测损失之比。由此得到的 R^2 具有清晰的解释：当其大于 0 时，

该预测模型对样本外预测是有用的。当其小于 0 时，观测数据的平均值是更好的预测模型，原因可能是预测变量未包含足够的结果向量信息、样本量过小而无法进行准确的模型拟合，或模型不适合该预测任务。我们展示了如何使用数据拆分算法估计样本外 R^2 。我们发现，合并 R^2 估计量先分别估计零模型和预测模型的平方ˆ误差损失，然后再将二者合并为最终估计值 R^2，其是ˆ无偏的。因此，与在每个交叉验证折中分别计算 R^2 值、再对各折取平均且会产生偏差的平均估计量相比，应优先使用这种合并估计量。不同于 .632自助法，交叉验证结合合并估计量能够提供几乎无偏的 R^2估计，这与先前的研究结果一致（Braga-Neto 等，2004；Jiang 等，2007；Kohavi，2001；Molinaro 等，2005），因此应作为估计预测性能的首选方法。然而，在具有强信号的高维情境中，仍会出现一定程度的向下估计偏差；如果计算条件允许，使用足够高的折数可以抵消这种偏差。最后，需要注意的是，R^2 估计并不专门适用于在当前数据集上训练的模型，而是适用于在相似的随机抽取数据集上训练的所有模型。

与任何根据有限样本估计的参数一样，通过数据拆分算法得到的估计值也存在不确定性。然而，由于缺乏针对 R^2 类指标的标准误估计量，人们往往只报告其点估计。仅凭这些点估计，无法回答一些简单的统计问题，例如预测模型是否显著降低了相对于零模型的损失。回答这一问题的一种方法是反复置换结果变量，每次都重新拟合预测模型并计算预测性能，从而构建性能估计量的零分布（Cruz 等，2020；De Meyer 等，2022）。这类置换方法的缺点是计算开销较大，而且没有为 R^2 指明明确的零值。这些置换方法检验的零假设是预测变量无法预测结果变量，这意味着存在某个小于 0 的未知 R^2 值（见补充材料第 3 节）。此外，置换方法只能得出 p 值，却无法得出标准误。

作为另一种选择，我们基于均方误差损失标准误估计方面的最新进展，为通过交叉验证或 .632自助法估计的样本外 R^2 提供了标准误。我们的方法也很容易扩展到基于独立测试数据估计的 R^2 值。ˆ R^2 的标准误可用于检验零假设 H_0: $R^2 \leq 0$，与之相对应的、易于解释的备择假设 H_a: $R^2 > 0$ 表明，该预测模型显著优于不含预测变量的零模型ˆ。 R^2 的标准误也可用于构造置信区间，以及比较单位可能不同的结果变量的可预测性。另一方面，对于同一结果变量，直接比较不同预测模型的均方误差值更为合适，因为这样无需通过德尔塔法近似比值的方差。

我们提供的德尔塔法标准误估计量在某些情况下存在向上偏差。可能的

解释包括：样本量较小时，估计量的正态性偏离最为明显，德尔塔法给出的近似较差（见补充材料第 2 节）；以及难以估计 MSE 与 MST 之间的相关性 ρ。然而，使用 ρ 的非参数估计或刀切估计得到的德尔塔法标准误，仍优于自助法标准误；后者可能存在向下偏差，导致无法控制第一类错误，且置信区间的覆盖率低于名义覆盖率。当预测变量几乎没有预测价值时，样本外 R^2 的估计量方差较小，但所有考察的方法都高估了该方差。值得庆幸的是，随着预测价值提高，这种偏差会减小，从而有望提高检出真正可预测结果变量的能力。尽管如此，在高维情景下， R^2 的估计量方差仍然相当大，这可能是因为高维惩罚模型的模型拟合具有较大的变异性。因此，对于此类模型估计出的 R^2 值（及其细微差异），应避免过度解读。因此，我们建议报告预测模型诊断的标准误和置信区间，以便了解结果的可重复性、指导后续研究设计，并进行模型比较。

补充材料

补充材料 详尽模拟结果、证明和软件版本。（pdf）

R代码 用于计算 R^2 及其标准误，以及运行所有模拟和分析的 R 代码，可在 https://github.com/maerelab/Rsquared 获取。（url）

参考文献

Anderson-Sprecher, R. (1994). “模型比较与 R²”。载于：《美国统计学家》48.2，第 113 –117 页。Bates, S., T. Hastie, and R. Tibshirani (2021). “交叉验证：它估计什么，效果如何？”载于：*arXiv* 预印本 *arXiv:2104.00673*。Bradley, A. P. (1997). “ROC 曲线下面积在评估机器学习算法中的应用”。载于：*Pattern Recognit.* 30.7，第 1145 –1159 页。Braga-Neto, U. M. and E. R. Dougherty (2004). “交叉验证对小样本微阵列分类是否有效？”载于：*Bioinformatics (Oxford, England)* 20.3，第 374 –380 页。Breiman, L. (2001). 载于：*Mach. Learn.* 45.1，第 5 –32 页。Cameron, A. C. and F. A. G.Windmeijer (1997). “若干常见非线性回归模型的拟合优度 R 平方度量”。载于：《计量经济学杂志》77.2，第 329 –342 页。Campbell, J. Y. and S. B. Thompson (2008). “样本外预测超额股票收益：有什么方法能胜过历史平均值吗？”载于：《金融研究评论》21.4，第 1509 –1531 页。Cohen, P., S. G. West, and L. S. Aiken (2014). 行为科学中的应用多元回归/相关分析。心理学出版社。

Cruz, D. F., S. D. Meyer, J. Ampe, H. Sprenger, D. Herman, T. V. Hautegem, et al. (2020). “在田间利用单株组学，将玉米基因与功能和表型联系起来”。载于：分子系统生物学16.12，e9667。De Meyer, S.,D. F.Cruz, T. D. Swaef, P.Lootens, J. D. Block, K. Bird, et al. (2022). “根据莲座期叶片基因表达预测单株田间种植的甘蓝型油菜产量性状”。载于：*bioRxiv*。DiCiccio, T. J. and B. Efron (1996). “自助法置信区间”。载于：统计科学11.3，第 189 –228 页。Efron, B. and R. Tibshirani (1997). “交叉验证的改进： $632+$ 自助法”。载于：美国统计学会会刊92.438，第 548 –560 页。Friedman, J., T. Hastie, and R. Tibshirani (2010). “通过坐标下降求解广义线性模型的正则化路径”。载于：统计软件杂志33.1，第 1 –22 页。Gauss, C. F. (1823). “关于受最小误差约束的观测组合理论”。载于：哥廷根：*Dieterich*。Harding, B., C. Tremblay, and D. Cousineau (2014). “标准误：使用蒙特卡罗模拟对标准误估计量的综述与评估”。载于：心理学定量方法10.2，第 107 –123 页。Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). 统计学习要素：数据挖掘、推断与预测。第 2 卷。施普林格。Jiang, W. and R. Simon (2007). “用于估计微阵列分类预测误差的自助法与调整自助法的比较”。载于：*Stat. Med.*26.29，第 5320 –5334 页。Kohavi, R. (Mar. 2001). “交叉验证和自助法在准确度估计与模型选择中的研究”。载于：14。Kvålseth, T. O. (1985). “关于 R^2 的注意事项”。载于：《美国统计学家》39.4，第 279 –285 页。Love, M. I., W. Huber, and S. Anders (Dec. 2014). “使用 DESeq2 对 RNA测序数据的倍数变化和离散度进行调节估计”。载于：基因组生物学15.12，第 550 页。Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005). “预测误差估计：重抽样方法比较”。载于：《生物信息学》21.15，第 3301 –3307 页。Nagelkerke, N. J. D. et al. (1991). “关于决定系数一般定义的注记”。载于：*Biometrika* 78.3，第 691 –692 页。Valbuena, R., A. Hernando, J. A. Manzanera, E. B. Görgens, D. R. A. Almeida, C. A. Silva, et al. (2019). “评估观测森林生物量与预测森林生物量：R平方、一致性指数还是最大信息系数？”载于：《欧洲遥感杂志》52.1，第 345 –358 页。Verweij, P. J. M. and H. C. V. Houwelingen (1993). “生存分析中的交叉验证”。载于：*Stat. Med.*12.24，第 2305 –2314 页。Wherry, R. J. (1931). “预测多重相关系数收缩的新公式”。载于：《数学统计年刊》2.4，第 440 –457 页。Zhang, D. (2017). “广义线性模型的决定系数”。载于：《美国统计学家》71.4，第 310 –316 页。

Zou, H. 和 T. Hastie（2005）。“通过弹性网络进行正则化和变量选择”。载于：*J. R. Stat.Soc. Series B Stat. Methodol.* 67.2，第301–320页。