基于杠杆分数的大数据回归最优子数据选择

Vasilis Chasiotis 和 Dimitris Karlis

D希腊雅典经济与商业大学统计系

arXiv:2305.01597v2 [stat.ME] 5 Jul 2023

摘要

随着数据集规模的扩大，建模过程对计算资源的需求也随之增加，因为传统的回归方法需要对大型数据矩阵求逆。主要问题在于数据规模庞大，因此一种标准做法是进行子抽样，以获取大数据中信息最丰富的部分。本文探讨了一种现有的、基于杠杆分数的方法，该方法最初用于线性模型判别中的子数据选择。我们的目标是提出将上述方法用于选择信息最丰富的数据点，以估计一阶线性模型和含交互作用的模型中的未知参数。我们通过模拟实验和真实数据应用得出结论：基于杠杆分数的方法优于现有方法。

关键词：D-最优设计；试验设计；子数据；线性回归；信息矩阵

1 引言

由于大型数据集规模庞大且复杂，往往很容易超出传统计算资源的处理能力，甚至可能导致某些分析完全无法进行。此外，一些统计模型可能需要更多内存或更强的处理能力。然而，由于标准机器无法提供这些资源，分析过程可能会非常复杂。

应对这些挑战的一种方法是从大数据中选取子数据来开展分析。数据缩减是在从完整数据中选取的较小样本上进行分析。这种方法可以减少分析所需的计算资源，从而能够针对较小的数据集进行分析。

应谨慎选择子数据，以确保其能够代表大数据，从而使基于子数据分析得出的结论能够推广到大数据。总体而言，由于计算资源有限，大数据会给统计分析带来重大挑战，而数据缩减是克服这些挑战的有效方法。然而，在

实施子抽样之前，充分考虑其潜在缺点和局限性非常重要。

Drineas 等人(2011年）提出随机抽取一部分数据。他们的思路是对数据进行随机哈达玛变换，然后通过均匀子抽样随机选取子数据。他们的目标是近似线性回归模型中的普通最小二乘估计量。然而，他们的方法存在固有的随机性问题。

因此，近年来迅速发展起来的一种替代方法，着重于确定性地选择数据点，使得一小部分完整数据能够保留完整数据中的大部分信息。由于最优设计问题依赖于数据选择，这类方法与试验设计的概念密切相关。因此，最优设计理论对于建立从完整数据中选取信息量最大子数据的框架非常有用。

在本文余下部分，我们将用 n、 p 和 k 分别表示完整数据规模、协变量数量和子数据规模。

作为首次尝试，Wang 等人（2019年）提出了基于信息的最优子数据选择（IBOSS）方法，其灵感源于最优试验设计的概念。他们着重于从完整数据中选择信息量最大的子数据，以估计未知参数。他们的思路是“最大化”信息矩阵，这正是最优试验设计理论的核心目标。总体而言，他们得出结论：使未知参数协方差矩阵的逆矩阵行列式最大的子数据是 D-最优的，因此它包含完整数据中信息最丰富的数据点。此外，他们还基于未知参数协方差矩阵逆矩阵行列式的一个上界，开发了一种用于选择 D-最优子数据的算法。更具体地说，IBOSS 方法的算法依次选取所有协变量取值最小和最大的那些数据点，并排除之前已选取的数据点；其时间复杂度为 O（$np+kp^2$），或者在 O（np）时为 $n > kp$。

Wang 等人（2021年）提出了用于选择子数据的正交子抽样（OSS）方法，该方法基于二水平正交表的最优性。需要指出的是，二水平正交表既能使估计参数的平均方差最小化，又能提供最佳预测（Dey 和 Mukerjee，1999年），因此它是线性回归的最优设计。Wang 等人开发的序贯添加算法（2021年）基于二水平正交表的组合正交性。此外，他们通过消除数据点来避免算法耗时过长，因此其计算复杂度为 O（np对数k）。该算法基于一个差异函数，用于衡量在同时保留两个与正交数组最优性相关的特征时，数据点所产生的偏差。第一个特征是选择极端数据点；第二个特征是使所选数据点的符号尽可能不同（组合正交性）。此外，在从完整数据中选择信息丰富的子数据时，OSS 方法优于 IBOSS 方法。

Ren 和 ZhaoRen 和 Zhao（2021年），受Wang 等人（2021年）的启发，着重于选择接近 $k \times p$ 强度 2 的二水平正交阵列的子数据。然而，Chasiotis 和 Karlis（2023年）提到了 Ren 和 Zhao （2021年）的算法 3 在实现方面存在的问题，因此他们的方法未纳入考量。

此外，Chasiotis 和 Karlis（2023年）的方法旨在识别并交换未被选中的数据点与某种方法（即 OSS 或 IBOSS 方法）已选中的数据点，以提高 D-最优性准则的值。Chasiotis 和 Karlis（2023年）提出的两种算法被视为对现有算法的扩展，因此评估时应权衡 D-最优性准则值的提升与由此产生的额外计算时间。

Wang 等人 Wang 等人 (2019年) 的方法已扩展到其他情况，例如多项逻辑回归（Yao 和 Wang，2019年）、分位数回归（Wang 和 Ma，2021年）以及逻辑回归（Cheng 等人，2020年）。此外，关于更多相关研究，请读者参阅Wang (2019年）、Lee 等人 (2021年）以及Deldossi 和 Tommasi(2022年）。关于基于设计从大数据中进行子数据选择或子抽样的更多信息，可参阅Yao 和 Wang(2021年）和Yu 等人(2023年的综述论文，这些论文全面概述了该领域当前的研究状况。

当前论文中的方法基于杠杆分数（LEVSS），从完整数据中确定性地选择信息最丰富的数据点，Yu 和 Wang (2022年) 已经提出该方法来选择用于线性模型判别的子数据。然而，有一些动机促使我们进一步研究杠杆分数在选择信息最丰富的数据点以估计未知参数中的作用。首先，由于 IBOSS 和 OSS 方法获得的子数据是 D-最优的，Yu 和 Wang (2022年) 中的定理2促使我们进一步研究基于 LEVSS 方法选择信息最丰富的数据点以估计未知参数的问题。此外，Chasiotis 和 Karlis(2023年) 的研究也给了我们启发，他们专注于选择凸包尽可能接近由完整数据生成的凸包的数据点；也就是说，在子数据下，信息矩阵的行列式将会很大。Chasiotis 和 Karlis (2023年) 还证明，最大化信息矩阵的行列式可以转化为最大化所选子数据下协变量的广义方差。需要注意的是，所选数据点云所占空间的体积与广义方差的平方根成正比。

为了进一步说明我们的研究动机，请看图1中的数据。我们有 $n = 5000$ 和 $p = 2$。观测值 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 服从多元正态分布，即 $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma} = (\Sigma_{ij})$，$i$，$j = 1$，2 是协方差矩阵。此外，对于 $i = j = 1$，2，$\Sigma_{ij} = 1$；对于 $i \neq j = 1$，2，$\Sigma_{ij} = 0.5$。假设我们有兴趣选择 $k = 50$ 个数据点。IBOSS 方法选择两个协变量的极端数据点，而 OSS 方法选择尽可能靠近数据域各个角的数据点。从某种意义上说，LEVSS 方法似乎会选择凸包较大的数据点。可以看出，LEVSS 方法选出的数据点能够更精确地反映完整数据的结构属性。因此，根据 Yu 和 Wang（2022）以及 Chasiotis 和 Karlis（2023）的理论结果，LEVSS 方法似乎确实很有希望用于选择信息量最大的观测值。

在 Yu 等人 （2023年）发表的综述论文中，作者评估了众多算法从大数据中选取最具信息量的子数据以估计未知参数的能力，并提供了模拟实验以及真实数据应用。其中评估的算法之一是 LEVSS 方法的算法。

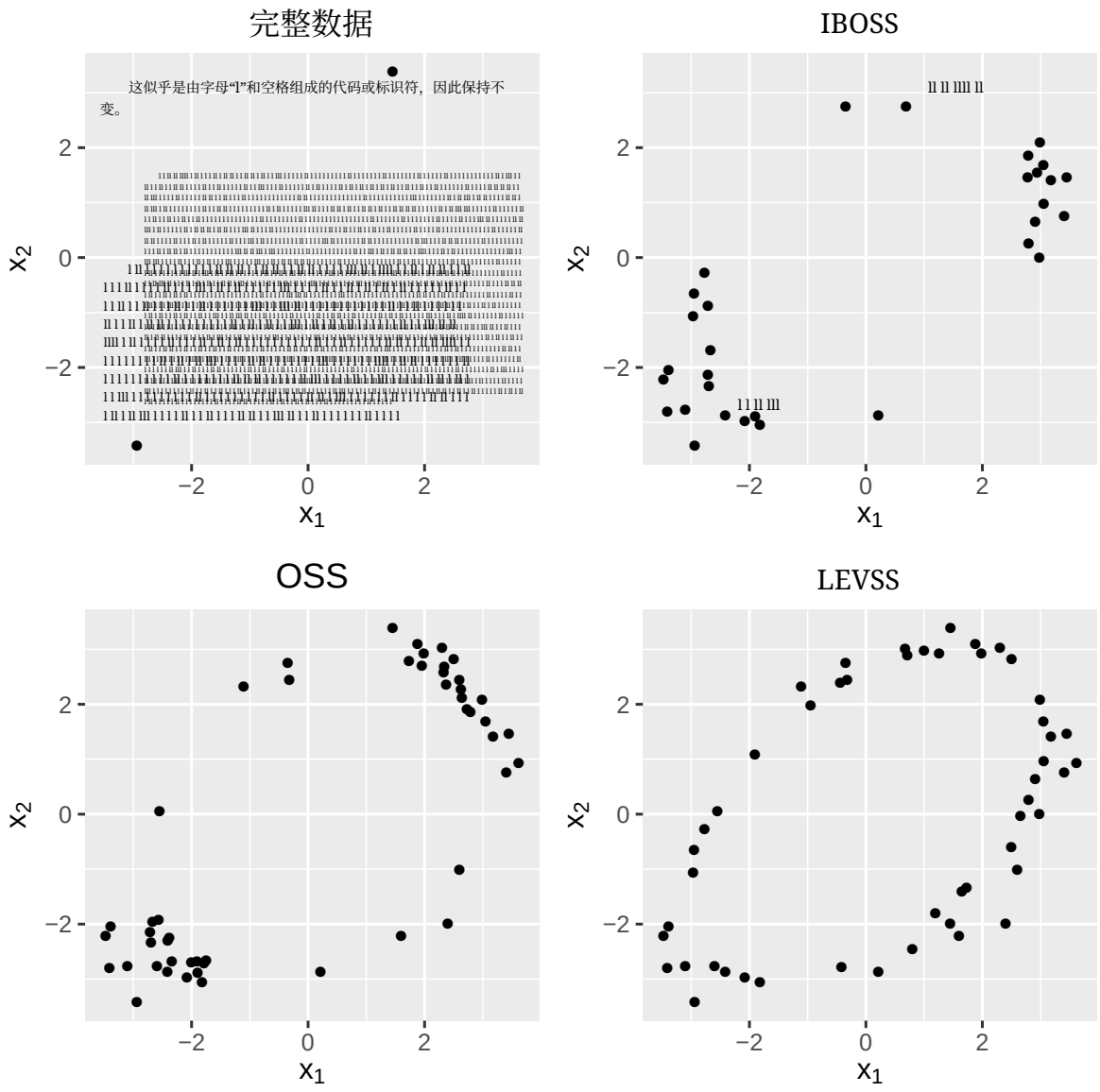

图1：不同方法的示例。生成了一个包含 $n = 5000$ 和 $p = 2$的数据集。使用不同的方法选取了 50 个数据点。第一行中，第一幅图展示了完整数据，第二幅图展示了 IBOSS 方法。第二行中，第一幅图和第二幅图分别展示了 OSS 和 LEVSS 方法。

然而，在当前论文中，我们通过模拟实验和真实数据应用提供了进一步证据，表明 LEVSS 方法能够选取最具信息量的数据点，以便估计未知参数。此外，我们不仅考虑一阶线性模型，还考虑了含交互作用的模型。此外，我们还应指出，在针对线性模型的情况下，我们有兴趣将基于 LEVSS 方法选取的子数据与基于 IBOSS 和 OSS 方法选取的子数据进行比较，也就是说，我们假设预测变量和响应变量遵循一个预先设定的模型。这意味着我们不考虑无模型子抽样方法，

旨在假设一个既定模型的情况下阐明其潜在过程。

本文其余部分的组织结构如下。在第 2 节中，我们简要回顾线性回归模型下基于子数据的最佳线性无偏估计量。在第 3 节中，我们介绍 LEVSS 方法的算法。在第 4 节中，我们提供模拟证据以支持 LEVSS 方法，包括与现有方法（IBOSS 和 OSS）的比较。在第 5 节中，我们使用一个真实数据集进行说明；在第 6 节中，我们对本文进行总结并展开讨论。

2 预备知识

假设完整数据表示为（$\mathbf{x}_i$， y_i）、……、（$\mathbf{x}_n$， y_n）。设线性回归模型为：

$$y_i = \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

其中，β_0 为截距参数，$\mathbf{x}_i$ =（x_{i1}，x_{i2}，……，x_{ip}）^{T} 为协变量向量，$\boldsymbol{\beta}_1$ =（β_1，β_2，……，β_p）^{T} 为 p维未知斜率参数向量，y_i 为响应变量，ϵ_i 为误差项。在给定协变量$\mathbf{x}_i$、$i = 1$、2、……、n的条件下，y_i彼此不相关，并且 ϵ_i满足 E(ϵ_i) $= 0$ 和 V(ϵ_i) $= \sigma^2$。

考虑模型（1）下的完整数据，$\boldsymbol{\beta}$ =（β_0，$\boldsymbol{\beta}_1^{\mathrm{T}}$）^{T}的最小二乘估计量，即其最佳线性无偏估计量，为

$$\hat{\boldsymbol{\beta}}_{\mathrm{Full}} = \left(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \mathbf{z}_i y_i,$$

其中，$\mathbf{z}_i = (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$。

其逆矩阵

$$\mathbf{Q}_{\mathrm{Full}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}$$

$\hat{}$ 等于 $\boldsymbol{\beta}_{\mathrm{Full}}$ 的协方差矩阵，其中$\mathbf{Q}_{\mathrm{Full}}$ 是误差项 ϵ_i'服从正态分布时，完整数据下 $\boldsymbol{\beta}$ 的观测费舍尔信息矩阵。$\mathbf{Q}_{\mathrm{Full}}$ 仍称为信息矩阵，即使不要求正态性假设。

如果完整数据的样本量 n 过大，则对全部数据进行完整分析可能不可行。因此，受计算资源限制，我们希望通过选择完整数据的一个子集，从完整数据中获取有用信息。

令 δ_i 为一个指示变量，用于表示 ($\mathbf{x}_i$， y_i) 是否被纳入子数据。因此，$\delta_i = 0$ 若 ($\mathbf{x}_i$， y_i) 未被纳入子数据，则取值为 $\delta_i = 1$；否则取值为相反值。另外，我们假设要选取大小为 k 的子数据，即 $\sum_{i=1}^{n} \delta_i = k$。因此，$\boldsymbol{\beta}$ 的最小二乘估计仍然是基于子数据的最佳线性无偏估计，即，

$$\hat{\boldsymbol{\beta}}_{\mathrm{Sub}} = \left(\sum_{i=1}^{n} \delta_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \delta_i \mathbf{z}_i y_i.$$

大小为 k 的子数据对应的信息矩阵可写为

$$\mathbf{Q}_{\text{Sub}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \delta_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}. \tag{2}$$

所选子数据应当以某种方式达到最优。根据 D-最优性准则，当相应的 $\mathbf{Q}_{\text{Sub}}$的行列式达到最大值时，大小为 k 的子数据是 D-最优的。

此外，我们还应提及，Chasiotis 和 Karlis （2023年）证明了$\mathbf{Q}_{\text{Sub}}$ 在2 中的行列式是协变量Wilks，1932年）在所选子数据下的广义方差，因此，他们将最大化 $\mathbf{Q}_{\text{Sub}}$ 在2 中的行列式问题转化为最大化所选子数据下协变量的广义方差问题。

3 基于杠杆值的算法

本节介绍Yu 和 Wang（2022年）提出的用于线性模型判别的确定性杠杆值选择（LEVSS）算法。

按照 Yu 和 Wang（2022年）给出的符号，令 # (Γ) 表示集合 Γ 的基数，令 κ (B) := λ_{max} (B) /λ_{min} (B) 表示矩阵 B 的条件数，其中 λ_{max}(B) 和 λ_{min} (B) 分别是平方矩阵 B 的最大和最小特征值。此外，当 B 为奇异矩阵时，则有 κ(B) $= \infty$。

LEVSS 算法见算法 1。

算法 1LEVSS

输入：设计矩阵 X $=(\mathbf{x}_i^{\mathrm{T}})$ $i = 1, 2, \ldots, n$、目标样本量 ($k > p$) 和阈值 T (≥ 1)。输出：选定的索引集 Γ 以及子数据对应的设计矩阵。初始化： $\Gamma = \emptyset$, $\mathbf{U}_\Gamma = \emptyset$, $\kappa\left(\mathbf{U}_\Gamma^{\mathrm{T}} \mathbf{U}_\Gamma\right) = \infty$。对 X 进行奇异值分解，得到 X $=\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$，计算杠杆分数 $h_{ii} := \|U_{i\cdot}\|^2$，其中 $U_{i\cdot}$ 表示 U 的第 i 行，并对 h_{ii}'s 进行排序，使其满足$h_{(11)} \geq \ldots \geq h_{(nn)}$。对于 i，依次取 $1, \ldots, n$，执行以下操作：若 # (Γ) $\leq k$或 $\kappa\left(\mathbf{U}_\Gamma^{\mathrm{T}} \mathbf{U}_\Gamma\right) \geq T$，则将与 $h_{(ii)}$ 对应的数据点的索引添加到集合 Γ。将 $\mathbf{U}_\Gamma$ 更新为 U 中对应于 Γ的选定行。否则，跳出循环。结束 if。结束循环。

Yu 和 Wang (2022年）提到，LEVSS 算法关于条件数的停止准则旨在确保子数据下的设计矩阵不存在病态，即防止多重共线性。此外，他们还提到，从几何角度来看，条件数的阈值 T 可以防止子数据落入低秩子空间。而且，如果 LEVSS 算法选择了超过

k 个数据点，例如 k^*，那么他们建议采用简单随机抽样，从中选取 k 个。k^*。

此外，他们指出，当协变量来自椭圆等高分布族时（Fang 等，1990年），LEVSS 算法关于条件数的停止准则并不重要，因为子数据的协变量空间会迅速扩展到完整数据的协变量空间。

LEVSS 算法的时间复杂度为 $O(np^2)$。

4 模拟实验

本节基于模拟数据评估 LEVSS 算法的性能，同时给出 IBOSS 和 OSS 方法的算法结果，以便进行比较。

4.1 一阶线性模型

在模型（1）下，协变量 $\mathbf{x}_i$按照以下情形生成。

- 情形1。$\mathbf{x}_i$彼此独立，且服从定义在$[0,1]^p$ 上的多元均匀分布，所有协变量均相互独立。
- 情况 2。 $\mathbf{x}_i$ 服从多元正态分布，即 $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$，wit h

$$\boldsymbol{\Sigma} = \left(0.5^{I(i,j)}\right), i, j = 1, 2, \ldots, p, \tag{3}$$

W此处 $I(i, j) = 0$对 $i = j = 1, 2, \ldots\ldots, p$，以及 $I(i, j) = 1$对 $i \neq j = 1, 2, \ldots\ldots, p$.

- 案例3。 $\mathbf{x}_i$服从截断的多元正态分布，其定义域为 $[-5, 5]^p$，即 $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$，协方差矩阵为 $\boldsymbol{\Sigma}$ 见（3）。

响应数据由（1）中的线性模型生成，其中$\boldsymbol{\beta}$ 的真实值是一个所有元素均为 1 的 51 维向量，且 $\sigma^2 = 9$。我们包含一个截距，因此 $p = 50$。

重复进行 1000 次模拟，并计算通过 IBOSS、OSS 和 LEVSS 方法选取的子数据的经验均方误差（均方误差）。我们使用调整后的估计量 $\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1^{Sub}$ (Wang 等人，2019年) 估计截距，其中 $\bar{y}$ 是完整数据中响应变量的均值，$\bar{\mathbf{x}}$ 是完整数据中所有协变量均值构成的向量，而 $\hat{\boldsymbol{\beta}}_1^{Sub}$ 是基于子数据得到的 $\boldsymbol{\beta}_1^{Sub}$ 的普通最小二乘估计量。因此，我们将第 $\beta_0^{(r)} - \beta_0$ 次重复中的 ($\hat{}^2$和 $||\hat{}\boldsymbol{\beta}_1^{(r)} - \boldsymbol{\beta}_1||^2$ 分别视为截距估计量和斜率估计量的均方误差，其中 $\hat{}\ r$和 $\hat{}\beta_0^{(r)}$ 分别是第 $\boldsymbol{\beta}_1^{(r)}$ 次重复中的 $\hat{}\ \beta_0$ 和 $\hat{}\boldsymbol{\beta}_1^{Sub}$ 。 r

我们研究完整数据规模分别为 $n = 5 \times 10^3$、10_4、10_5 和 10^6,，且子数据规模固定为 $k = 1000$的情形。图2、3 和4 分别展示了案例1、2和3中，采用不同方法选取的子数据所估计的斜率参数的均方误差。我们还给出了均值（♦）。

在情形2和情形3中，LEVSS算法始终优于IBOSS和OSS算法。此外，在情形1中，LEVSS算法始终优于IBOSS算法，

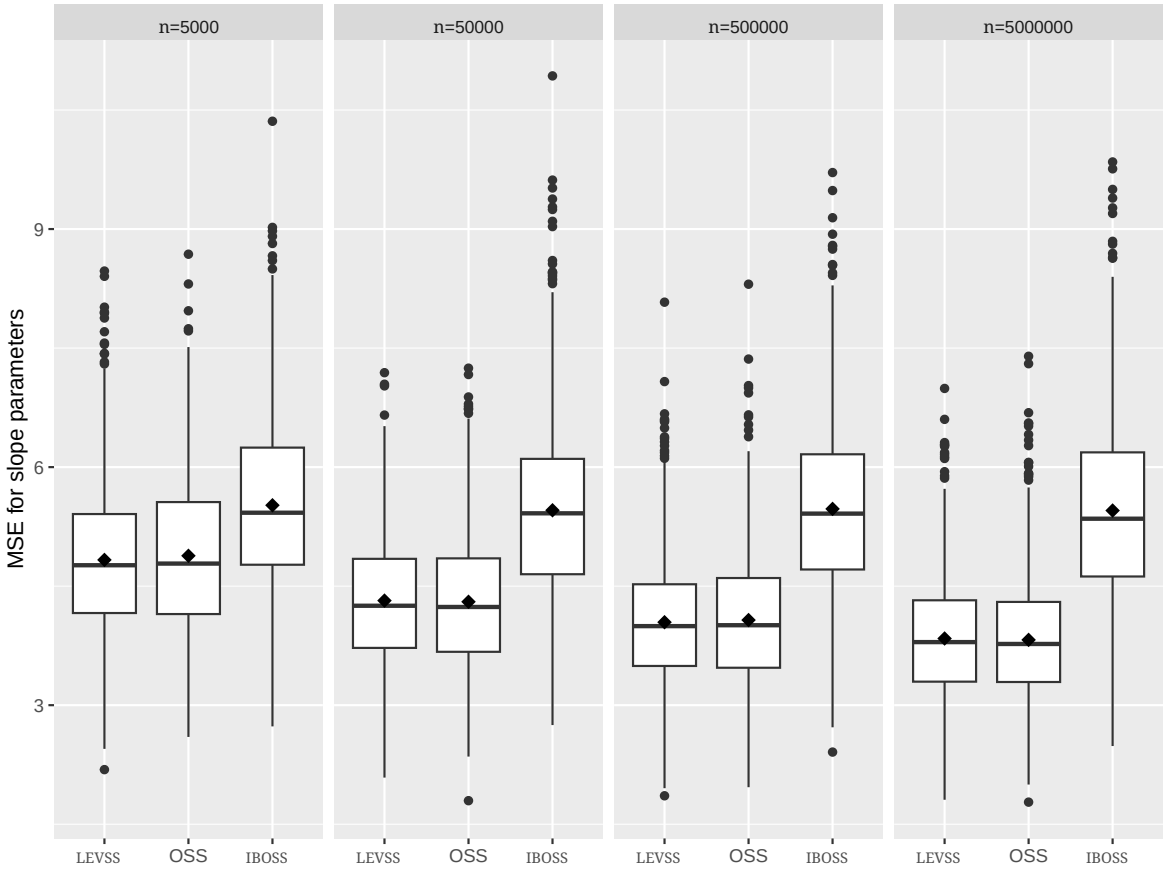

图2：当完整数据规模为 $n = 5 \times 10^3, 10^4, 10^5$ 且 10^6，子数据规模为 $k = 1000$时，针对情形1的协变量，采用不同方法选取的子数据所估计的斜率参数的均方误差。

并且略优于 OSS。这是因为当观测值 $\mathbf{x}_i$'s, $i = 1, 2, \ldots, 50$ 相互独立且服从协变量均相互独立的多元均匀分布时，LEVSS 和 OSS 算法所选子数据的结构非常相似。此外，尽管子数据规模固定为 n，随着完整数据规模 $k = 1000$的增加，LEVSS 方法估计的斜率参数的均方误差会降低。这表明，随着完整数据规模增大，LEVSS 方法能够从完整数据中识别出更多信息丰富的数据点。此外，无论协变量是否有界，随着 n 增大，LEVSS 方法中估计的斜率参数的均方误差都会快速降低，与 OSS 方法中的情况相同。总的来说，与 IBOSS 和 OSS 方法相比，LEVSS 方法能够更准确地估计模型参数，如图 2、3 和 4所示。三种方法的截距均方误差不变，因此为简洁起见，省略相关结果。关于 IBOSS 和 OSS 方法的结果，请参阅补充材料中 Wang 等人 (2021年) 的图 1。

4.2 含交互作用的模型

我们感兴趣的是评估 IBOSS、OSS 和 LEVSS 的方法，同时考虑协变量之间存在交互作用的情况。响应数据的生成方式如下：

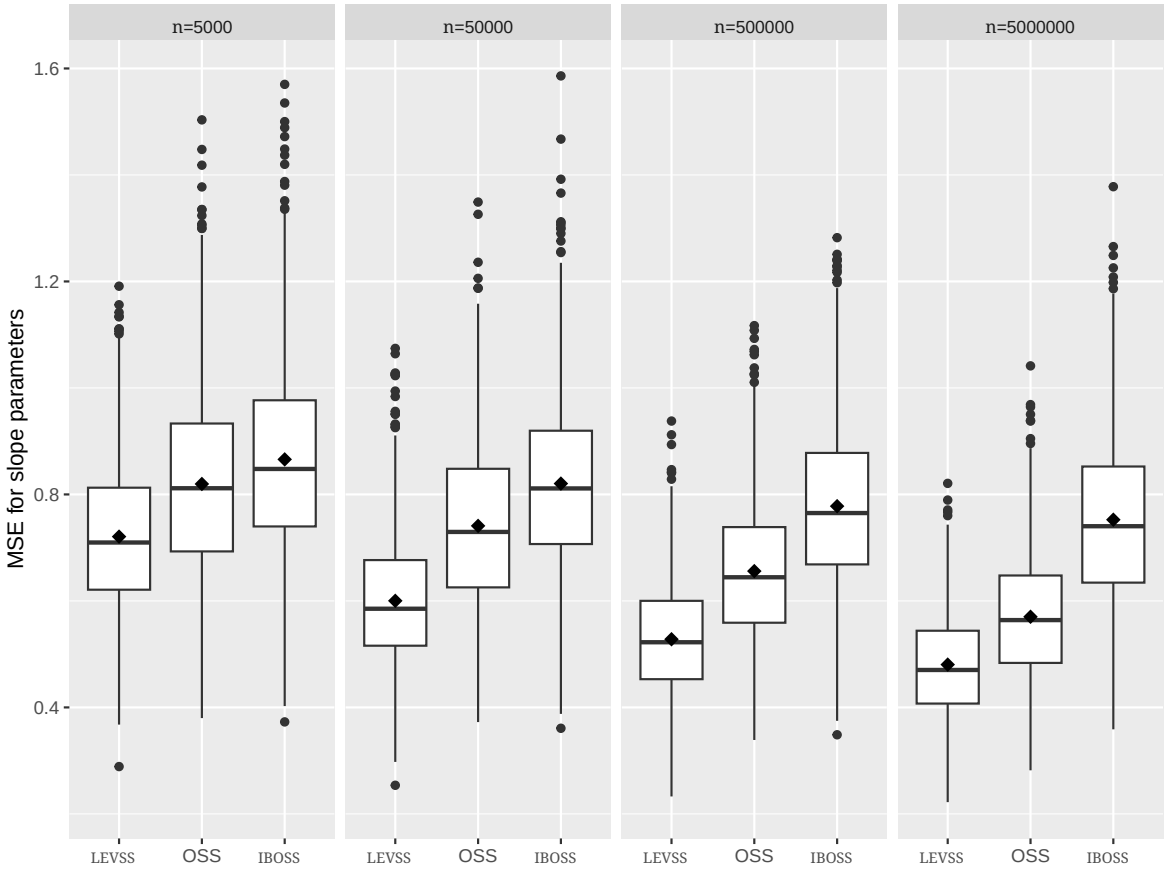

图 3：当完整数据规模为 $n = 5 \times 10^3, 10^4, 10^5$ 和 10^6，且子数据规模为 $k = 1000$ 时，对于案例 2 的协变量，采用不同方法选取的子数据所估计的斜率参数的均方误差。

针对小节4.1中的每种协变量情况，均基于模型

$$\mathbf{y} = \beta_0 + \mathbf{X}_m\boldsymbol{\beta}_m + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon},$$

其中，$\mathbf{X}_m$ 的每一列均为一个协变量，$\mathbf{X}_\gamma$ 的每一列均为$\mathbf{X}_m$中两列的逐元素乘积，$\boldsymbol{\beta}_m = (\beta_1, \beta_2, \ldots\ldots, \beta_p)^{\mathrm{T}}$ 是一个 p维的主效应向量，$\boldsymbol{\beta}_\gamma = (\beta_{p+1}, \beta_{p+2}, \ldots\ldots, \beta_{p(p-1)/2})^{\mathrm{T}}$ 是一个 $p(p-1)/2$维的交互效应向量，$\boldsymbol{\epsilon}$ 为误差项。

我们令 $p = 10$, $\beta_0 = 1$, $\boldsymbol{\beta}_m$ 的真实值和 $\boldsymbol{\beta}_\gamma$ 分别为元素全为 1 的 10 维和 45 维向量，并令 $\sigma^2 = 9$。

模拟重复进行 1000 次。我们分别计算由 IBOSS、OSS 和 LEVSS 方法选出的子数据中估计量$\boldsymbol{\beta}_m$ 和 $\boldsymbol{\beta}_\gamma$的均方误差$\beta_m$ 和均方误差β_γ。我们研究完整数据规模为 $n = 5 \times 10^3$、10_4、10_5 和 10^6,，且子数据规模固定为 $k = 1000$的情形。图5 和 6 展示了案例3中不同方法所选子数据的均方误差β_m 和均方误差β_γ 。我们还给出了均值（⧫）。案例1和案例2的结果相似，因此为简洁起见不再列出。

首先，采用 LEVSS 方法选出的子数据的均方误差β_m 低于采用 IBOSS 和 OSS 方法选出的子数据的均方误差，因此 LEVSS 提供了更准确的估计

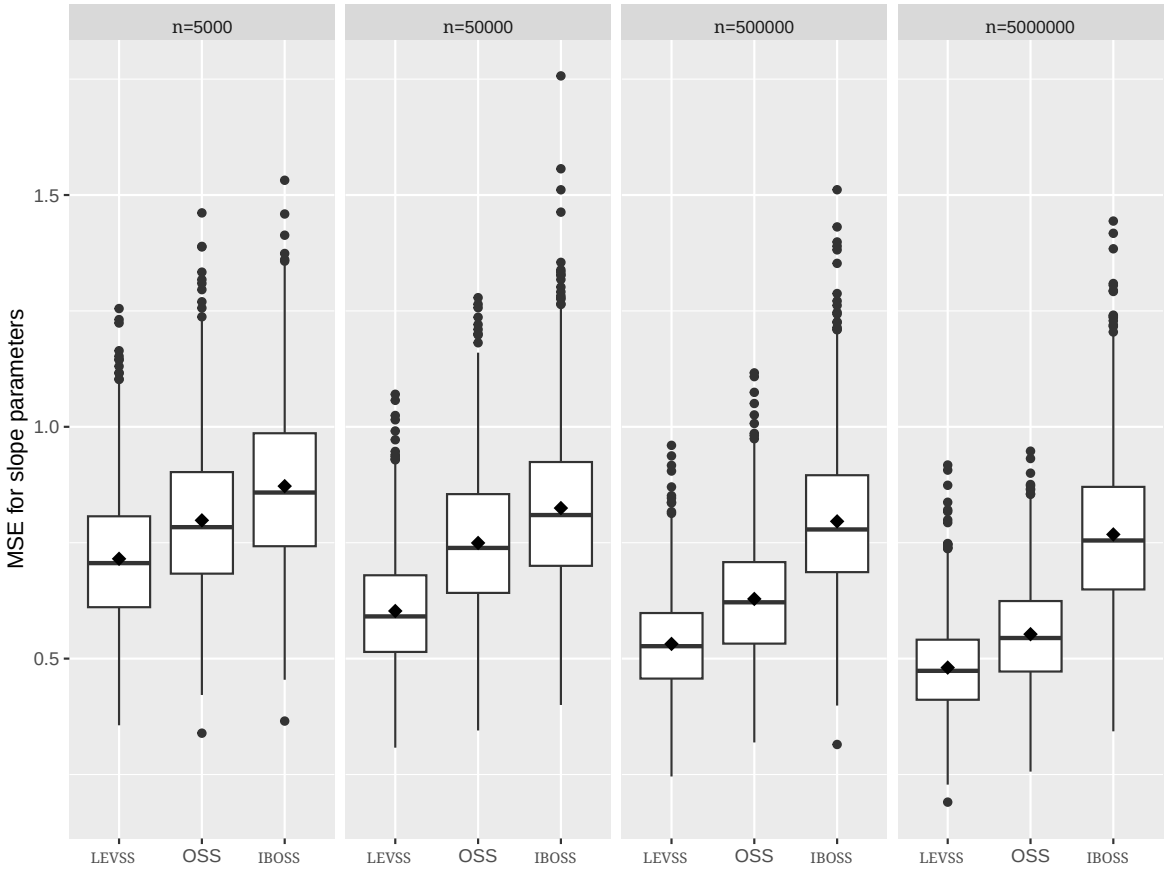

图4：当完整数据规模为 $n = 5 \times 10^3, 10^4, 10^5$ 和 10^6 且子数据规模为 $k = 1000$时，针对情形3中的协变量，采用不同方法选取的子数据所估计的斜率参数的均方误差。

与其他方法相比，主要效应的均方误差更小。此外，与 IBOSS 和 OSS 方法相比，LEVSS 方法能够识别显著的交互效应，因为 LEVSS 方法的均方误差β_m 最小。值得一提的是，LEVSS 和 OSS 方法均根据协变量选择数据点，而 IBOSS 方法还依赖交互项。

4.3 计算时间

我们重点考察在子数据规模等于 $k = 1000$且协变量数量等于 $p = 50$时，针对不同完整数据规模 n，情形1下IBOSS、OSS和LEVSS方法的计算时间。所有计算均在配备3.6 GHz英特尔8核I7处理器和16GB内存的个人计算机上进行。

在表1中，我们列出了IBOSS、OSS和LEVSS方法的平均计算时间（单位：秒）。

对于任意完整数据规模 n，LEVSS 方法的算法都比 OSS 方法的算法更快。此外，LEVSS 与 IBOSS 方法的算法在计算时间上的差异非常小。

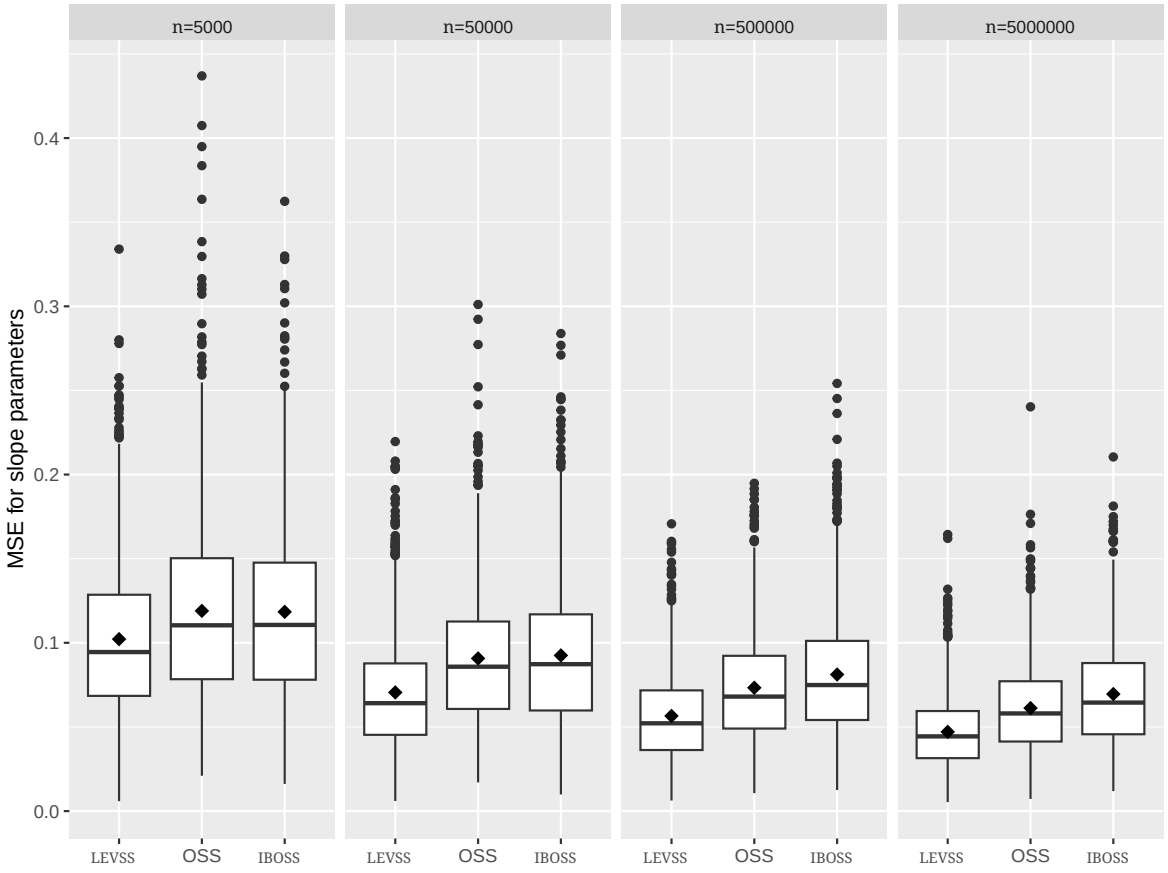

图5：当完整数据规模为 $n = 5 \times 10^3$，$\mathbf{10}_4$，$\mathbf{10}_5$ 且 10^6、子数据规模为$k = 1000$时，针对案例3协变量，通过不同方法选取的子数据的均方误差β_m 。

n	5×10^3	5×10^4	5×10^5	5×10^6
IBOSS	0.175	0.758	6.858	73.14
OSS	3.205	6.886	18.351	150.06
LEVSS	1.205	1.812	7.911	83.16

表 1：当子数据规模等于 $k = 1000$ 且协变量数量等于 $p = 50$ 时，IBOSS、OSS 和 LEVSS 方法在不同完整数据规模 $n = 5 \times 10^3$、$\mathbf{10}_4$、$\mathbf{10}_5$ 和 10^6, 下的平均执行时间（秒）。

5 真实数据应用

在本节中，我们通过一个真实数据示例评估 LEVSS 方法的性能，并考察模型（1）中斜率参数的普通最小二乘估计量的准确性。

真实数据示例所用的数据集包含澳大利亚南部海岸四种真实波浪情景（悉尼、阿德莱德、珀斯和塔斯马尼亚）中波浪能转换器的位置及其吸收功率。完整数据包含 $n = 288{,}000$个数据点，记录了32个位置变量和16个吸收功率变量，因此

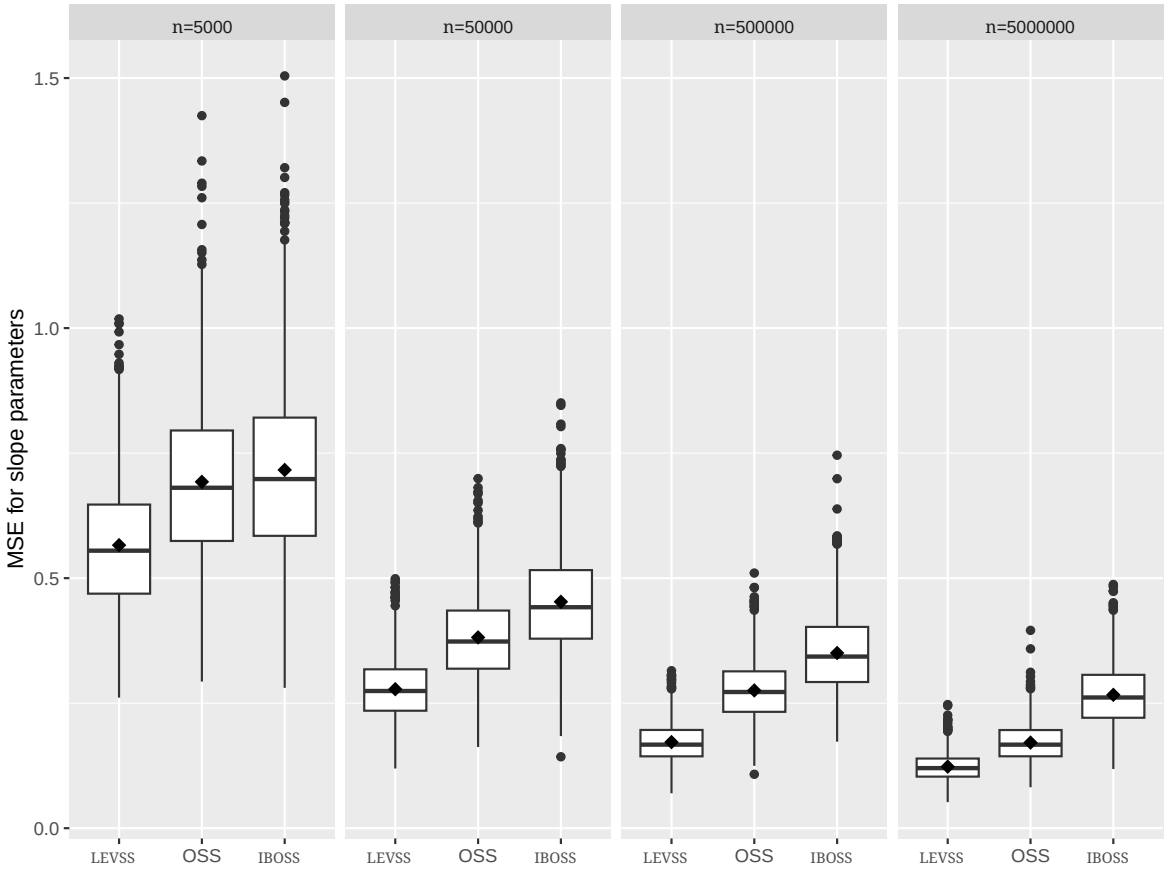

图6：完整数据规模为 $n = 5 \times 10^3$、10_4、10_5 且 10^6，子数据规模为$k = 1000$时，不同方法为案例3的协变量选取子数据所得的均方误差β_γ。

模型中的协变量数量为 $p = 48$。响应变量是农场的总发电量；在分析中，我们对其进行对数变换。有关该数据集的更多信息，请参阅“UCI机器学习资料库”Dua和Graff(2019年)。

我们比较 LEVSS 方法的算法与 IBOSS 和 OSS 方法的算法的表现。具体而言，对于每种算法，我们使用 100 个自助样本计算斜率参数向量的均方误差，方法与 Wang 等人(2019) 及 Wang 等人(2021相同。每个自助样本都是从完整数据中有放回抽取的随机样本，样本规模为n。对于每个自助样本，分别运行每种算法以获得子数据，然后根据选出的子数据估计模型参数。对于 LEVSS 方法的算法，我们采用不同的条件数阈值 T，即 $T = 25$、20 和 15，并且也考虑不存在条件数停止准则的情形。这些对 LEVSS 方法算法的调整用于真实数据示例，而不用于第4节的模拟实验，因为当 k 已经选出时，条件数已很小；也就是说，子数据的协变量空间迅速扩展到完整数据的协变量空间。

图7 展示了不同方法在 $k = 5p$、 $10p$、 $20p$ 和

$30p$的自助法均方误差。此外，为了更清晰地展示图7，我们对每个均方误差取以 10 为底的对数。我们还提供了均值（♦）。在最小化自助法均方误差方面，对 LEVSS 方法算法的所有修改都优于 IBOSS 和 OSS 算法。此外，正如Wang 等人（2021年）所指出的，IBOSS 方法在这个真实数据示例中的表现不佳，可能是因为并非所有变量都很重要。另一方面，可以说，LEVSS 方法尽可能逼近由完整数据生成的凸包，因此表现出色。讨论对 LEVSS 方法算法所做的修改很重要。随着子数据规模 k 增大，忽略条件数停止准则时，自助法均方误差最小。此外，需要注意的是，在某种意义上，当 T 变小时，LEVSS 方法的行为类似于对完整数据进行简单随机子抽样，因为此时选取的数据点超过 k 。随着子数据规模 k 增大，这一判断似乎是合理的，因为当 k较小时，例如 $k = 240$，当 T 在所用取值中最小时，自助法均方误差也最小。然而，取 T的另一个值（小于 15），则会使任何子数据规模 k下的自助法均方误差都增大。此外，我们还应注意，随着T 的值增大，LEVSS 方法的算法运行得更快。忽略条件数停止准则时，LEVSS 方法算法的修改版本运行速度最快。

6 结论

我们评估了 LEVSS 方法的算法，以便从大型数据集中以最优方式选择数据点，从而能够进行回归并尽可能获得信息量最大的系数。此外，我们还将 LEVSS 方法的算法与 IBOSS 和 OSS 方法的算法进行了比较，以展示其带来的改进。

此外，对 LEVSS 方法算法的修改为后者提供了非常有用的信息。阈值 T取值较大，甚至不采用针对条件数的停止准则，似乎都能选出信息量更大的数据点。然而，在这些情况下应用 LEVSS 方法的算法时，应考虑由此导致的多重共线性程度，因为正如 Yu 和 Wang(2022年) 所指出的，条件数较大可能导致矩阵病态，进而引发多重共线性。

此外，根据Yu 等人(2023年) 给出的结果，一些无模型子抽样方法（SPARTAN、SP）在部分模拟实验中表现优于 LEVSS 方法。因此，不仅需要开展进一步且更全面的研究，还需要开发能够适应大数据时代真实数据的新方法。

需要说明的是，我们没有以任何方式优化所使用的 `R` ，因此还可能进一步节省时间。

参考文献

Chasiotis, V. 和 Karlis, D.（2023年）。大数据回归的子数据选择：一种改进方法。doi.org/10.48550/arXiv.2305.00218。

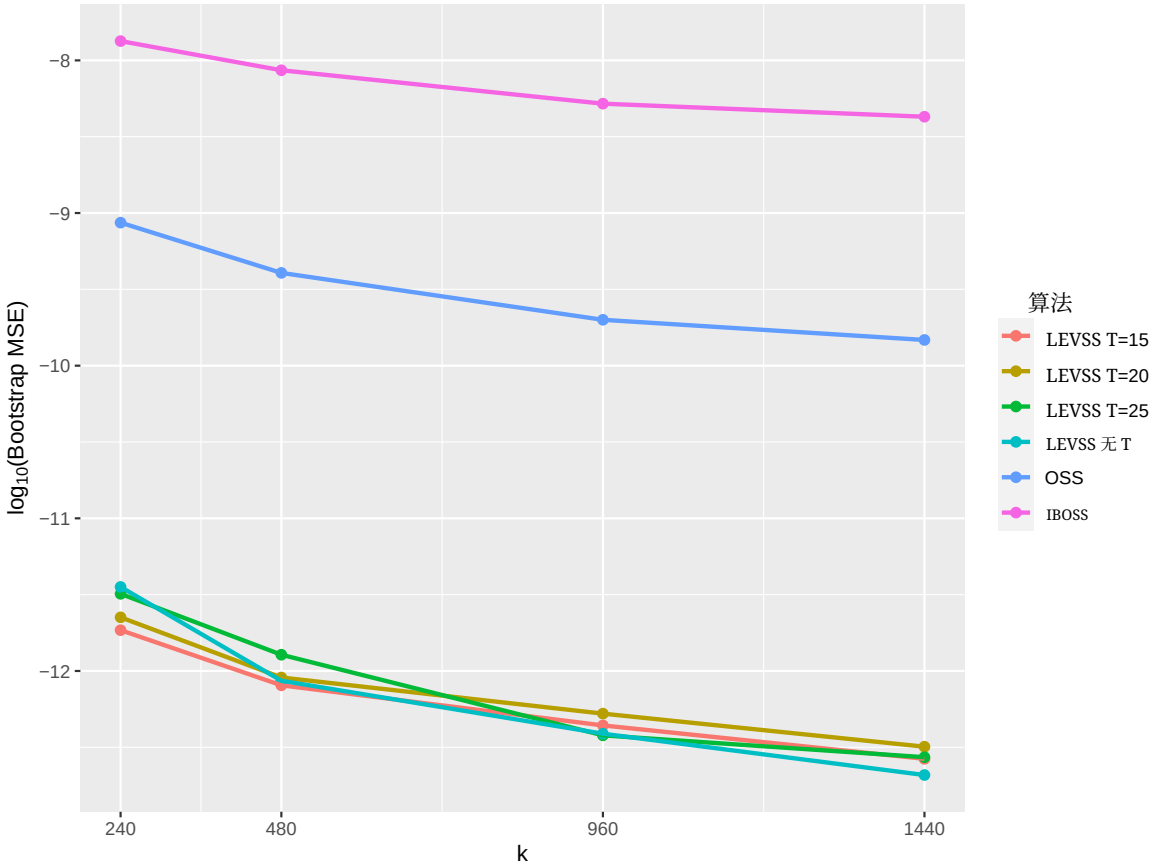

图7：当子数据规模为 $k = 5p$、$10p$、$20p$ 和 $30p$ 时，不同方法所选子数据的自助法均方误差。条件数阈值 T 的不同取值分别为 25（LEVSS $T = 25$）、20（LEVSS $T = 20$）和 15（LEVSS $T = 15$）。此外，LEVSS 方法的算法在不采用条件数停止准则的情况下实现（LEVSS no T）。

Cheng, Q.、Wang, H. 和 Yang, M.（2020年）。基于信息的最优子数据选择用于大数据逻辑回归。*Journal of Statistical Planning and Inference*，209：112–122。Deldossi, L. 和 Tommasi, C.（2022年）。从大型数据集中进行最优设计子抽样。*Journal of Quality Technology*，54(1)：93–101。Dey, A. 和 Mukerjee, R.（1999年）。分数因子设计。John Wiley & Sons。Drineas, P.、Mahoney, M. W.、Muthukrishnan, S. 和 Sarl´os, T.（2011年）。更快速的最小二乘近似。*Numerische mathematik*，117(2)：219–249。Dua, D. 和 Graff, C.（2019年）。UCI机器学习资料库。 http://archive.ics.uci.edu/ml。Fang, K.-T.、Kotz, S. 和 Ng, K. W.（1990年）。对称多元分布及相关分布。Springer。

Lee, J., Schifano, E. D. 和 Wang, H.（2021年）。广义线性模型的快速最优子抽样概率近似。 doi.org/10.1016/j.ecosta.2021.02.007。Ren, M. 和 Zhao, S.-L.（2021年）。基于正交数组的大数据子数据选择。
doi.org/10.1080/03610926.2021.2012196。Wang, H.（2019年）。分治信息型最优子数据选择算法。 *Journal of Statistical Theory and Practice*, 13(3):1–19。Wang, H. 和 Ma, Y.（2021年）。大数据分位数回归的最优子抽样。*Biometrika*, 108(1):99–112。Wang, H.、Yang, M. 和 Stufken, J.（2019年）。大数据线性回归的信息型最优子数据选择。 *Journal of the American Statistical Association*, 114(525):393–405。Wang, L.、Elmstedt, J.、Wong, W. K. 和 Xu, H.（2021年）。大数据线性回归的正交子抽样。 *Annals of Applied Statistics*, 15(3):1273–1290。Wilks, S. S.（1932年）。方差分析中的若干推广。 *Biometrika*, 24(3-4):471–494。Yao, Y. 和 Wang, H.（2019年）。Softmax回归的最优子抽样。 *StatisticalPapers*, 60(2):585–599。Yao, Y. 和 Wang, H.（2021年）。海量数据集最优子抽样方法综述。*Journal of Data Science*, 19(1):151–172。Yu, J. 、 Ai, M. 和 Ye, Z. （ 2023 年）。大数据设计启发式子抽样综述。doi.org/10.1007/s00362-022-01386-w。Yu, J. 和 Wang, H.（2022年）。用于线性模型区分的子数据选择算法。 *Statistical Papers*, 63:1883–1906。