具有许多协变量的线性回归的野生自助推断

Wenze Li*

2025 年 6 月

摘要

我们提出了一种对野自助法的简单修改，并建立了其在含有大量协变量和异方差误差的线性回归模型中的渐近有效性。蒙特卡罗模拟显示，与基于标准正态临界值的其他方法相比，改进后的野自助法具有优异的小样本性能，尤其是在样本量较小和/或控制变量的数量与样本量同数量级时。

关键词：自助法，多个协变量，异方差性。

JEL 分类代码：C12, C21

1 介绍

估计某一变量的因果效应的一个重要计量经济学方法是基于线性回归模型，假设在控制了一组足够大的协变量（例如固定效应面板数据模型或部分线性模型）后，感兴趣的变量是外生的。然而，在存在（条件）异方差的情况下，使用许多协变量进行推断可能会很具有挑战性。Cattaneo et al. (2018a) 提供了一个统一的非标准分布近似框架，该框架容纳了文献中研究的各种渐近情况，包括小带宽、大量工具变量和大量协变量的渐近情况。此外，

*Division of Economics, School of Social Sciences, Nanyang Technological University. Email: wenze001@e.ntu.edu.sg.

I am sincerely grateful to Wenjie Wang (Nanyang Technological University) for his generous guidance and insightful comments, which have substantially strengthened the arguments and analysis presented in this paper.

如 Cattaneo et al. (2018b) 所示，在协变量数量 q_n 随样本大小 n 增长的渐近框架下，传统的异方差稳健方差估计量的一致性将需要 $q_n/n \to 0$。在其开创性研究中，Cattaneo et al. (2018b) 和 Jochmans (2022) 提出了新的方差估计量，这些估计量即使在 $q_n/n \nrightarrow 0$ 情况下也保持一致，并基于正态临界值提供渐近有效的推断。

然而，众所周知，渐近正态逼近可能在有限样本中产生扭曲，尤其是在观测值数量较少时。自举法被发现可以改进许多情况下的渐近逼近；例如，参见 Hall and Horowitz (1996)，Horowitz (2001)，Davidson and MacKinnon (2010)，Djogbenou et al. (2019) 以及其中的参考文献。然而，在高维设定下它可能会失效；例如，参见 El Karoui and Purdom (2018) 及其参考文献。对于具有异方差误差的线性回归，Mammen (1993) 建立了非参数自举和野自举的有效性需要 $q_n^{1+\delta}/n \to 0, \delta > 0$。类似地，对于具有许多工具变量（IV）和同方差误差的线性工具变量模型，Wang and Kaffo (2016) 显示了当工具变量的数量与样本量 n 同数量级时标准残差引导法的不一致性，并提出了一种有效的替代引导程序。然而，他们的方法不能直接扩展到既有许多工具变量又有异方差的情况，即使控制变量的数量是固定的。

据我们所知，在存在 $q_n/n \nrightarrow 0$ 和异方差的情况下，没有现有的有效的自助法适用于线性回归模型。在 Cattaneo et al. (2018b) 中的模拟还表明，非参数自助法似乎在其设置中失败（例如，请参见他们论文中的备注 2）。在这篇论文中，根据 Jochmans (2022) 中构造的（近似）交叉拟合方差估计量，我们提出了一种对标准残差自助法的简单修改，并证明了在 $q_n/n \nrightarrow 0$ 条件下它的渐近有效性。蒙特卡洛实验展示了我们的方法具有良好的有限样本性能。

2 模型和设置

考虑线性回归模型

$$y_i = x_i\beta + w_i'\gamma + u_i,\ i = 1, ..., n, \tag{1}$$

其中 y_i 是一个标量结果变量，x_i 是一个标量解释变量，如某种治疗，w_i 是 $q_n \times 1$ 个协变量的向量，u_i 是误差项。我们允许 q_n 是样本大小 n 的一个不可忽略的部分。普通最小二乘（OLS）估计量 β 定义为

$$\hat{\beta}_n = \left(\sum_{i=1}^n \hat{v}_i^2\right)^{-1} \left(\sum_{i=1}^n \hat{v}_i y_i\right),$$

其中 $\hat{v}_i = \sum_{j=1}^n (M_n)_{ij} x_j$，$(M_n)_{ij} = \{i = j\} - w_i'(\sum_{k=1}^n w_k w_k')^{-1} w_j$，以及 $\{\cdot\}$ 表示指示函数。

最近，Cattaneo et al. (2018b) 提出了一种新的异方差性稳健的方差估计量，当 $\limsup_n q_n/n < 1/2$ 时该估计量是一致的。此外，Jochmans (2022) 提出了一个替代的方差估计量，只要 $\limsup_n q_n/n < 1$ 成立就保持一致，并且其相应的 t 比率可以定义为 $t_n = (\hat{\beta}_n - \beta_0)/\sqrt{\acute{\Omega}_n}$，其中

$$\acute{\Omega}_n = \left(\sum_{i=1}^n \hat{v}_i^2\right)^{-1} \left(\sum_{i=1}^n \hat{v}_i^2 (y_i \acute{u}_i)\right) \left(\sum_{i=1}^n \hat{v}_i^2\right)^{-1}, \tag{2}$$

与 $\acute{u}_i = \hat{u}_i/(M_n)_{ii}$ 和 $\hat{u}_i = \sum_{j=1}^n (M_n)_{ij}(y_j - x_j\hat{\beta}_n)$。然后，基于正态临界值的 t-检验在大样本中具有正确的大小。然而，模拟表明渐近正态逼近可能在有限样本中的表现不满意（第 5 节），特别是在样本量小和/或 q_n 等于 n 的大比例时。在第 3 节中，我们提出了一种修改的野自助法，该方法遵循（近似）交叉拟合方差估计器 Jochmans (2022)。

3 修改后的野自助法程序

我们的程序定义如下：

步骤 1： 给定零假设 $H_0 : \beta = \beta_0$，生成残差 $\{\tilde{u}_i(\beta_0)\}_{i=1}^n$，其中 $\tilde{u}_i(\beta_0) = y_i - x_i\beta_0 - w_i'\tilde{\gamma}_n$，和 $\tilde{\gamma}_n$ 是零限制的 OLS 估计量 γ。

步骤 2： 生成 $\{u_i^*\}_{i=1}^n$，其中 $u_i^* = a_n(\beta_0)\omega_i^*\tilde{u}_i(\beta_0)$，$\{\omega_i^*\}_{i=1}^n$ 是独立同分布的随机权重样本，这些权重与数据无关，均值为零且方差为单位，而

$$a_n(\beta_0) = \sqrt{\max\{\acute{\Sigma}_n(\beta_0), 1/n\}/\hat{\Sigma}_n(\beta_0)} \tag{3}$$

是一个调整因子，考虑到了协变量的高维性。具体来说，$\acute{\Sigma}_n(\beta_0) = \sum_{i=1}^n \hat{v}_i^2 y_i \acute{u}_i(\beta_0)$，$\hat{\Sigma}_n(\beta_0) = \sum_{i=1}^n \hat{v}_i^2 \tilde{u}_i^2(\beta_0)$，$\tilde{u}_i(\beta_0)$ 在第一步中定义，以及 $\acute{u}_i(\beta_0) = \tilde{u}_i(\beta_0)/(M_n)_{ii}$。请注意，$\acute{\Sigma}_n(\beta_0)$ 类似于 (2) 中 $\acute{\Omega}_n$ 的“核心”部分，但在这里我们强加了 H_0，因为我们是在第一步中施加零假设生成自助样本的。此外，$max\{\cdot, 1/n\}$ 防止在有限样本中 $\acute{\Sigma}_n(\beta_0)$ 可能取非正值的情况。[1]

步骤 3： 生成 $y_i^* = x_i\beta_0 + w_i'\tilde{\gamma}_n + u_i^*, i = 1, ..., n$。然后，计算 bootstrap OLS 估计量 $\hat{\beta}_n^* = (\sum_{i=1}^n \hat{v}_i^2)^{-1}(\sum_{i=1}^n \hat{v}_i y_i^*)$，以及 bootstrap 残差 $\hat{u}_i^* = \sum_{j=1}^n (M_n)_{ij}(y_j^* - x_j\hat{\beta}^*)$。

步骤 4： 计算 $t_n^* = (\hat{\beta}_n^* - \beta_0)/\sqrt{\acute{\Omega}_n^*}$，其中 $\acute{\Omega}_n^* = (\sum_{i=1}^n \hat{v}_i^2)^{-1}(\sum_{i=1}^n \hat{v}_i^2(y_i^*\acute{u}_i^*))(\sum_{i=1}^n \hat{v}_i^2)^{-1}$，与 $\acute{u}_i^* = \hat{u}_i^*/(M_n)_{ii}$ 即，$\acute{\Omega}_n^*$ 具有与 $\acute{\Omega}_n$ 相同的公式，但使用 bootstrap 样本。**步骤 5：** 重复步骤

[1] 这也可能发生在 Cattaneo et al. (2018b) 和 Jochmans (2022) 的方差估计量上。

2-4 B 次，并计算 bootstrap$p$ 值 $p_n^* = B^{-1}\sum_{b=1}^{B} 1\{|t_n| > |t_n^{*(b)}|\}$。如果 p_n^* 小于名义水平 α，则拒绝 H_0。

若干评论是必要的。

评论 1。在第一步中，我们在计算残差时施加了零，这是由 Cameron et al. (2008)，Davidson and Flachaire (2008)，Roodman et al. (2019) 等所倡导的。因此，在第二步中我们还对调整因子 $a_n(\beta_0)$ 施加了零。当 $q_n/n \to 0$，$\acute{\Sigma}_n(\beta_0)$ 和 $\hat{\Sigma}_n(\beta_0)$ 渐近等价时，我们的程序简化为标准的（零假设强加）野自助法。

评论 2. 相比上述第 t 型程序，可以考虑一种第百分位数型程序，其引导 p 值等于 $B^{-1}\sum_{b=1}^{B} 1\{|\sqrt{n}(\hat{\beta}_n - \beta_0)| > |\sqrt{n}(\hat{\beta}_n^{*(b)} - \beta_0)|\}$。然而，按照文献中的建议（例如，请参阅上面引用的论文），我们在模拟中发现第百分位数 t 表现更好，因此也推荐使用它。

评论 3. 在这里我们关注的是 x_i 是一个标量变量的情况。可以通过对得分自助法进行修改（例如，Kline and Santos (2012)）将分析扩展到 x_i 是一个固定 d_x 维向量的情况。具体来说，对于测试 $H_\lambda : c'\beta = \lambda$ 的过程，其中 $c \in R^{d_x}$ 和 $\lambda \in R$，可以定义如下：

(1) 获得无约束的 OLS 估计量 $\tilde{\beta}_n$ 和 $\tilde{\gamma}_n$，并将其用于得分贡献 $\{S_i(\tilde{\beta}_n)\}_{i=1}^n$，其中 $S_i(\tilde{\beta}_n) = \hat{v}_i\tilde{u}_i(\tilde{\beta}_n)$ 和 $\tilde{u}_i(\tilde{\beta}_n) = y_i - x_i'\tilde{\beta}_n - w_i'\tilde{\gamma}_n$。

(2) 计算扰动得分贡献 $\{S_i^*(\tilde{\beta}_n)\}_{i=1}^n$，其中 $S_i^*(\tilde{\beta}_n) = \omega_i^*\hat{v}_i\tilde{u}_i(\tilde{\beta}_n)$;

(3) 计算引导统计量 $T_n^* = c'\left(n^{-1}\sum_{i=1}^n \hat{v}_i\hat{v}_i'\right)^{-1}\hat{A}_n(\tilde{\beta}_n)(n^{-1/2}\sum_{i=1}^n S_i^*(\tilde{\beta}_n))$，其中 $\hat{A}_n(\tilde{\beta}_n) = \acute{\Sigma}_n^{1/2}(\tilde{\beta}_n)\hat{\Sigma}_n^{-1/2}(\tilde{\beta}_n)$ 是一个调整矩阵，包含 $\acute{\Sigma}_n(\tilde{\beta}_n) = \sum_{i=1}^n \hat{v}_i\hat{v}_i'y_i\acute{u}_i(\tilde{\beta}_n)$ 和 $\hat{\Sigma}_n(\tilde{\beta}_n) = \sum_{i=1}^n \hat{v}_i\hat{v}_i'\tilde{u}_i^2(\tilde{\beta}_n)$;

(4) 使用数据条件下 T_n^* 的分布作为 $T_n = \sqrt{n}(c'\hat{\beta}_n - \lambda)$ 零分布的估计。

当 x_i 是一个标量时，这种修改后的得分引导程序等同于评论 2 中的百分位类型过程。我们推荐当前步骤 1-5 中的程序，因为它具有更好的有限样本性能，并且可能对实践者来说也更友好（例如，只需在标准野生引导的残差中添加一个调整因子）。

4 渐近理论

下面介绍一些用于建立自助法有效性的常规条件。假设 1-4 与 Cattaneo et al. (2018b) 和 Jochmans (2022) 中的那些非常相似。假设 5 包含自助法程序的条件。附录 A 对这些假设进行了进一步讨论。

令 $\mathcal{W}_n$ 表示一组随机变量，使得 $E[w_i|\mathcal{W}_n] = w_i$。令 $\epsilon_i = u_i - e_i$，其中 $e_i = E[u_i|X_n, \mathcal{W}_n]$，和 $V_i = x_i - E[x_i|\mathcal{W}_n]$。令 $\sigma_i^2 = E[\epsilon_i^2|X_n, \mathcal{W}_n]$，和 $\tilde{V}_i = \sum_{j=1}^n (M_n)_{ij} V_j$。

假设 1. 误差 ϵ_i 在给定 X_n 和 $\mathcal{W}_n$ 的条件下，相对于 i 是不相关的，并且集合 $\{\epsilon_i, V_i : i \in N_g\}$ 在给定 $\mathcal{W}_n$ 的条件下，相对于 g 是独立的，其中 $N_1, ..., N_G$ 表示将 $\{1, ..., n\}$ 分割为 G 个集合，使得 $\max_g |N_g| = O(1)$。

假设 2. 以概率趋于一，$\sum_{i=1}^n w_i w_i'$ 具有满秩，

$$\max_i \left(E[\epsilon_i^4|X_n, \mathcal{W}_n] + \sigma_i^{-2} + E[V_i^4|\mathcal{W}_n]\right) + \left(\lambda_{min}\left(n^{-1}\sum_{i=1}^n E[\tilde{V}_i^2|\mathcal{W}_n]\right)\right)^{-1} = O_p(1),$$

和 $\limsup_n q_n/n < 1$，其中 $\lambda_{min}(\cdot)$ 表示其参数的最小特征值。

假设 3. $\chi_n = O(1), \eta_n + n(\eta_n - \rho_n) + n\chi_n\eta_n = o(1)$，和 $\max_i |\hat{v}_i|/\sqrt{n} = o_p(1)$，其中 $\eta_n = n^{-1}\sum_{i=1}^n E[e_i^2]$，$\rho_n = n^{-1}\sum_{i=1}^n E[E[e_i|\mathcal{W}_n]^2]$，$\chi_n = n^{-1}\sum_{i=1}^n E[Q_i^2]$，$Q_i = E[v_i|\mathcal{W}_n]$，和 $v_i = x_i - (\sum_{j=1}^n E[x_j w_j'])(\sum_{j=1}^n E[w_j w_j'])^{-1} w_i$。

假设 4. $n\eta_n = O(1)$，$P\left[\min_i (M_n)_{ii} > 0\right] \to 1$，$(\min_i (M_n)_{ii})^{-1} = O_p(1)$，$n^{-1}\sum_{i=1}^n \tilde{Q}_i^4 = O_p(1)$，和 $\max_i |\mu_i|/\sqrt{n} = o_p(1)$，其中 $\mu_i = E[y_i|X_n, \mathcal{W}_n]$ 和 $\tilde{Q}_i = \sum_{j=1}^n (M_n)_{i,j} Q_j$。

假设 5. $\{\omega_i^*\}_{i=1}^n$ 是一个独立同分布的随机权重样本，与数据无关，并且它满足 $E[\omega_i^*] = 0$，$E[\omega_i^{*2}] = 1$ 和 $E[\omega_i^{*4}] < \infty$。此外，在原假设下 $\max_i |\hat{\mu}_i(\beta_0)|/\sqrt{n} = o_p(1)$，其中 $\hat{\mu}_i(\beta_0) = x_i\beta_0 + w_i'\tilde{\gamma}_n$。

定理 1. 设 F_n 和 F_n^* 分别是样本条件下 t_n 和 t_n^* 的累积分布函数。假设条件 *1-5* 成立。如果 H_0 为真，则

$$\sup_{c \in R} |F_n(c) - F_n^*(c)| = o_p(1).$$

定理 1 表明了在许多协变量和异方差性条件下修改后的野自助法的渐近有效性。对于自助中心极限定理（CLT）的证明，我们应用条件乘数 CLT（例如，van der Vaart and Wellner (1996) 的第 2.9 章）。

5 蒙特卡罗模拟

模型类似于在 Cattaneo et al. (2018b) 和 Jochmans (2022) 中考虑的模型：

$$y_i = x_i\beta + w_i'\gamma + \epsilon_i,$$

其中 $x_i \sim i.i.d.N(0,1), w_i$ 包含一个常数项和一组 $q_n - 1$ 零/一哑变量，以及 $\epsilon_i \sim i.i.d.N(0,1)$。虚拟变量是独立绘制的，成功概率为 π 和 $\gamma = 0$。样本量固定为 $n = 100$，和 $q_n/n \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$。蒙特卡罗重复次数设置为 $10,000$，且在整个模拟过程中引导重复次数设置为 199。参照 Jochmans (2022)，我们考虑三种设计，它们在 β 和 π 上有所不同：$\beta = 1$ 和 $\pi = 0.02$（设计 A），$\beta = 1$ 和 $\pi = 0.01$（设计 B）以及 $\beta = 2$ 和 $\pi = 0.02$（设计 C）。设计 A 的结果如表 1 所示。双边 t 检验的实证零假设拒绝频率给出了 Eicker-White 方差估计器（“HC0”）、Cattaneo et al. (2018b) 的方差估计器（“HCK”）、Jochmans (2022) 的方差估计器（“HCA”）以及分别采用高斯和 Rademacher 随机权重的提出的野自助法（“Wild-G” 和 “Wild-R”）。

我们突出以下几点发现。基于 “HC0”、“HCK” 和 “HCA” 的测试都倾向于随着 q_n 的增加而过度拒绝原假设，其中 “HCA” 的扭曲最小（例如，当 $q_n/n = 0.9$ 时，它们的零假设拒绝频率分别为 58.1%、58.1% 和 17.2%）。相比之下，两种自助法在各种 q_n 值下都能控制大小。此外，当 q_n 增加时，两者都变得更加保守，并且使用高斯权重的那个似乎比使用 Rademacher 权重的略微不那么保守。设计 B 和设计 C 的结果分别报告在表 2 和表 3 中。整体模式与表 1 中观察到的非常相似。在补充附录的第 C 节中，我们展示了面板数据模型的进一步模拟结果。

q_n/n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
HC0	0.071	0.097	0.116	0.150	0.186	0.243	0.316	0.407	0.581
HCK	0.059	0.069	0.071	0.083	0.095	0.116	0.165	0.228	0.581
HCA	0.067	0.075	0.076	0.084	0.084	0.087	0.100	0.121	0.172
Wild-G	0.051	0.046	0.047	0.048	0.045	0.044	0.044	0.040	0.035
Wild-R	0.038	0.044	0.036	0.041	0.040	0.039	0.041	0.039	0.033

表 1: 设计 A 的零假设拒绝频率

注意：“HC0”, “HCK”, 和 “HCA” 分别表示 Eicker-White 的方差估计量、Cattaneo et al. (2018b) 和 Jochmans (2022)，而 “Wild-G” 和 “Wild-R” 分别表示带有高斯权重和 Rademacher 权重的提议 wild bootstrap 方法。

q_n/n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
HC0	0.075	0.097	0.113	0.152	0.189	0.242	0.314	0.410	0.574
HCK	0.062	0.069	0.070	0.081	0.092	0.117	0.168	0.238	0.574
HCA	0.072	0.075	0.075	0.082	0.084	0.090	0.107	0.127	0.176
Wild-G	0.050	0.050	0.045	0.048	0.044	0.045	0.042	0.040	0.037
Wild-R	0.037	0.041	0.036	0.041	0.036	0.040	0.037	0.036	0.033

表 2: 设计 B 的零拒绝频率

q_n/n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
HC0	0.073	0.095	0.122	0.141	0.195	0.240	0.310	0.413	0.583
HCK	0.063	0.067	0.072	0.075	0.095	0.120	0.161	0.233	0.583
HCA	0.097	0.105	0.114	0.104	0.120	0.124	0.136	0.151	0.175
Wild-G	0.044	0.041	0.043	0.036	0.042	0.042	0.041	0.040	0.036
Wild-R	0.037	0.034	0.034	0.031	0.034	0.037	0.041	0.036	0.031

表 3: 设计 C 的零拒绝频率

6 结论

本文提出了一种对含有大量协变量和异方差误差的线性回归模型的标准残差引导程序进行简单的修改。我们证明了当协变量的数量与样本量处于同一数量级时，其渐近有效性。我们在引导程序中构建调整因子的方法遵循 Jochmans (2022) 提出的（近似）交叉拟合方差估计器。蒙特卡洛模拟表明，相对于基于标准正态临界值的替代方法，修改后的残差引导法在小样本量和/或协变量数量与样本量处于同一数量级的情况下具有出色的有限样本性能。对于未来研究可能的方向，我们注意到存在越来越多关于许多（弱）工具变量和非同方差误差下的稳健推断文献，也可能包含大量控制变量。[2] 由于数据结构的复杂性，在这种情况下，渐近正态逼近可能会表现不佳。另一方面，发现当适当实施时，引导方法可以显著提高 IV 模型的推理准确性，包括工具变量可能相当弱的情况。因此，考虑新的基于引导的方法，这些方法能同时对许多工具变量、大量控制变量和异方差误差具有鲁棒性可能是有

[2]例如，参见 Evdokimov et al. (2018),Crudu et al. (2021),Mikusheva and Sun (2022),Matsushita and Otsu (2024),Lim et al. (2024a),Dovì et al. (2024),Boot and Ligtenberg (2023),Navjeevan (2023),Boot and Nibbering (2024),Lim et al. (2024b) 和 Yap (2024) 等。

趣的。[3]

参考文献

Boot, T. and J. W. Ligtenberg (2023): "Identification- and many instrument-robust inference via invariant moment conditions," *arXiv:2303.07822*.

Boot, T. and D. Nibbering (2024): "Inference on LATEs with covariates," *arXiv preprint arXiv:2402.12607*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90, 414–427.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a): "Alternative asymptotics and the partially linear model with many regressors," *Econometric Theory*, 34, 277–301.

——— (2018b): "Inference in linear regression models with many covariates and heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350–1361.

Crudu, F., G. Mellace, and Z. Sándor (2021): "Inference in instrumental variable models with heteroskedasticity and many instruments," *Econometric Theory*, 37, 281–310.

Davidson, R. and E. Flachaire (2008): "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146, 162–169.

Davidson, R. and J. G. MacKinnon (2008): "Bootstrap inference in a linear equation estimated by instrumental variables," *The Econometrics Journal*, 11, 443–477.

——— (2010): "Wild bootstrap tests for IV regression," *Journal of Business & Economic Statistics*, 28, 128–144.

——— (2014): "Bootstrap confidence sets with weak instruments," *Econometric Reviews*, 33, 651–675.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393–412.

Dovì, M.-S., A. B. Kock, and S. Mavroeidis (2024): "A Ridge-Regularized Jackknifed Anderson-Rubin Test," *Journal of Business & Economic Statistics*, 1–12.

[3]例如，参见 Moreira et al. (2009)，Davidson and MacKinnon (2008, 2010, 2014)，Wang and Liu (2015)，Wang and Kaffo (2016)，Kaffo and Wang (2017)，Wang and Doko Tchatoka (2018)，Finlay and Magnusson (2019)，MacKinnon (2023) 和 Wang and Zhang (2024)。

El Karoui, N. and E. Purdom (2018): "Can we trust the bootstrap in high-dimensions? The case of linear models," *The Journal of Machine Learning Research*, 19, 170–235.

Evdokimov, K. et al. (2018): "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects," Tech. rep., Princeton University. Economics Department.

Finlay, K. and L. M. Magnusson (2019): "Two applications of wild bootstrap methods to improve inference in cluster-IV models," *Journal of Applied Econometrics*, 34, 911–933.

Hall, P. and J. L. Horowitz (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64, 891–916.

Horowitz, J. L. (2001): "The bootstrap," in *Handbook of econometrics*, Elsevier, vol. 5, 3159–3228.

Jochmans, K. (2022): "Heteroscedasticity-robust inference in linear regression models with many covariates," *Journal of the American Statistical Association*, 117, 887–896.

Kaffo, M. and W. Wang (2017): "On bootstrap validity for specification testing with many weak instruments," *Economics Letters*, 157, 107–111.

Kline, P. and A. Santos (2012): "A score based approach to wild bootstrap inference," *Journal of Econometric Methods*, 1, 23–41.

Lim, D., W. Wang, and Y. Zhang (2024a): "A conditional linear combination test with many weak instruments," *Journal of Econometrics*, 238, 105602.

——— (2024b): "A Dimension-Agnostic Bootstrap Anderson-Rubin Test For Instrumental Variable Regressions," *arXiv preprint arXiv:2412.01603*.

MacKinnon, J. G. (2023): "Fast cluster bootstrap methods for linear regression models," *Econometrics and Statistics*, 26, 52–71.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, 21, 255–285.

Matsushita, Y. and T. Otsu (2024): "A jackknife Lagrange multiplier test with many weak instruments," *Econometric Theory*, 40, 447–470.

Mikusheva, A. and L. Sun (2022): "Inference with many weak instruments," *Review of Economic Studies*, forthcoming.

Moreira, M. J., J. Porter, and G. Suarez (2009): "Bootstrap validity for the score test when instruments may be weak," *Journal of Econometrics*, 149, 52–64.

Navjeevan, M. (2023): "An Identification and Dimensionality Robust Test for Instrumental Variables Models," *arXiv preprint arXiv:2311.14892*.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb (2019): "Fast and wild: Bootstrap inference in Stata using boottest," *The Stata Journal*, 19, 4–60.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.

Wang, W. and F. Doko Tchatoka (2018): "On bootstrap inconsistency and Bonferroni-based size-correction for the subset Anderson–Rubin test under conditional homoskedasticity," *Journal of Econometrics*, 207, 188–211.

Wang, W. and M. Kaffo (2016): "Bootstrap inference for instrumental variable models with many weak instruments," *Journal of Econometrics*, 192, 231–268.

Wang, W. and Q. Liu (2015): "Bootstrap-based selection for instrumental variables model," *Economics Bulletin*, 35, 1886–1896.

Wang, W. and Y. Zhang (2024): "Wild bootstrap inference for instrumental variables regressions with weak and few clusters," *Journal of Econometrics*, 241, 105727.

Yap, L. (2024): "Inference with Many Weak Instruments and Heterogeneity," *arXiv preprint arXiv:2408.11193*.

“具有许多协变量的线性回归的野生自助法推断”的补充附录

第 A 节进一步讨论了第 4 节中的假设。第 B 节包含了定理 1 的证明。第 C 节展示了单向固定效应模型的模拟结果。

以下符号用于引导渐近性：对于任何引导统计量 T^*，如果对于任意的 $\delta > 0$，$\epsilon > 0$，$\lim_{n\to\infty} P[P^*[|T^*| > \delta] > \epsilon] = 0$，即 $P^*[|T^*| > \delta] = o_p(1)$，我们写 $T^* = o_{p^*}(1)$ 概率。为了简洁起见，我们用简写 $T^* = o_{p^*}(1)$ 来表示在概率上 $T^* = o_{p^*}(1)$。另外，如果在给定样本的条件下，T^* 在 P^* 下弱收敛到 T，对于包含在一个概率趋近于一的集合中的所有样本，我们写成在概率上 $T^* \to^{d^*} T$。具体地说，如果对于任何有界且一致连续的函数 f，$T^* \to^{d^*} T$ 在概率意义上成立当且仅当 $E^*[f(T^*)] \to E[f(T)]$ 在概率意义上成立，其中 E^* 表示由自助法诱导的概率测度下的期望。

A 关于假设的讨论

假设 1-3 与假设 1-3 在 Cattaneo et al. (2018b) 和 Jochmans (2022) 中紧密相关，在这些假设下，他们展示了高维情况下 OLS 估计量的渐近正态性，其中 $q_n/n \nrightarrow 0$ 作为 $n \to \infty$。

具体来说，假设 1 不仅涵盖了标准的随机抽样数据，还涵盖了诸如短期面板数据这样的重复测量数据。$\{N_1, ..., N_{G_n}\}$ 表示将 $\{1, ..., n\}$ 划分为 G_n 个层，这些层彼此独立，但在层内的观测值之间允许存在依赖性。然而，这一假设不允许误差项在不同单元之间存在相关性，因此排除了样本中的聚类、空间或时间序列依赖性。

假设 2 包含标准秩和矩条件。此外，它允许 q_n 以与样本量相同的速率增长。

此外，设置在 Cattaneo et al. (2018b) 和 Jochmans (2022) 允许线性回归模型 (1) 成为条件期望 $\mu_i = E[y_i|X_n, \mathcal{W}_n]$ 的参数线性均方逼近的情况。假设 3 提供了这种逼近应如何快速改进的正规条件。我们建议感兴趣的读者参阅 Cattaneo et al. (2018b) 的第 3.2.3 节和 Jochmans (2022) 的第 2.1 节以获得对该假设的更详细讨论。

假设 4 与假设 4 的 Jochmans (2022) 紧密相关，这是为了保证其方差估计量的一致性。特别地，条件 $\max_i \mu_i^2 = o_p(n)$ 用于控制 $\sum_{i=1}^n \hat{v}_i^2(y_i\acute{u}_i)$ 的方差，因为 $y_i\acute{u}_i$ 的方差依赖于 μ_i^2。感兴趣的读者可以参考 Jochmans (2022) 的第 2.2 节和补充附录 A.2 以获取更多讨论及相关充分条件。

假设 5 包含了修改后的野自举程序的条件。具体来说，它要求野自举中的随机权重与样本

独立，其均值为零，方差为一，并且具有有限的四阶矩。类似于假设 4，我们需要在假设 5 中的最后一个条件来控制 $\sum_{i=1}^{n} \hat{v}_i^2(y_i^* \acute{u}_i^*)$ 的（条件）方差。

B 定理 1 的证明

我们首先证明在原假设下，$n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(y_i^* \acute{u}_i^* - \sigma_i^{*2}) = o_{p^*}(1)$。为此，我们遵循定理 1 在 Jochmans (2022) 中的证明步骤。注意

$$n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(y_i^* \acute{u}_i^* - \sigma_i^{*2}) = n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(u_i^{*2} - \sigma_i^{*2}) + n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(y_i^* \acute{u}_i^* - u_i^{2*}), \tag{4}$$

其中 $\sigma_i^{*2} = E^*[u_i^{*2}] = a_n^2(\beta_0)\tilde{u}_i^2(\beta_0)$。

对于（4）右侧的第一项，我们有 $E^*\left[n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(u_i^{*2} - \sigma_i^{*2})\right] = 0$. 此外，对于方差（条件于数据）

$$\begin{aligned} &E^*\left[\left(n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(u_i^{*2} - \sigma_i^{*2})\right)^2\right] \\ &= n^{-2}\sum_{i=1}^{n} \hat{v}_i^4\left(E^*[u_i^{*4}] - \sigma_i^{*4}\right) = n^{-2}\sum_{i=1}^{n} \hat{v}_i^4\left(a_n^4(\beta_0)E^*[\omega_i^{*4}]\tilde{u}_i^4(\beta_0) - (a_n^2(\beta_0)\tilde{u}_i^2(\beta_0))\right) = o_p(1), \end{aligned}$$

这可以从 $a_n^2(\beta_0) = O_p(1)$，$\tilde{u}_i^2(\beta_0) = O_p(1)$，$(\max_i |\hat{v}_i|/\sqrt{n})^2 = o_p(1)$，$n^{-1}\sum_{i=1}^{n} \hat{v}_i^2 = O_p(1)$，以及假设 5 得出。特别地，$n^{-1}\sum_{i=1}^{n} \hat{v}_i^2 \le n^{-1}\sum_{i=1}^{n} v_i^2 \le 2(n^{-1}\sum_{i=1}^{n} Q_i^2) + 2(n^{-1}\sum_{i=1}^{n} V_i^2) = O_p(\chi_n) + O_p(1) = O_p(1)$。

因此，我们有

$$n^{-1}\sum_{i=1}^{n} \hat{v}_i^2(u_i^{*2} - \sigma_i^{*2}) = o_{p^*}(1). \tag{5}$$

然后，对于公式 (4) 右边的第二项，我们使用以下分解：

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\hat{v}_i^2(y_i^*\acute{u}_i^*-u_i^{2*})\\
&=n^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\hat{v}_i^2u_i^*(A_n)_{ij}u_j^*+n^{-1}\sum_{i=1}^{n}\hat{v}_i^2((A_n)_{ii}-1)u_i^{*2}+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{v}_i^2\hat{\mu}_i(\beta_0)(A_n)_{ij}u_j^*, \quad (6)
\end{aligned}
$$

这来自于 $(A_n)_{ij}=(H_n)_{ij}/(M_n)_{ii}$,$(H_n)_{ij}=(M_n)_{ij}-(n^{-1}\sum_{k=1}^{n}\hat{v}_k^2)^{-1}(n^{-1}\hat{v}_i\hat{v}_j)$，和 $\acute{u}_i^*=\sum_{j=1}^{n}((H_n)_{ij}/(M_n)_{ii})u_j^*=\sum_{j=1}^{n}(A_n)_{ij}u_j^*$。我们旨在证明（6）右边的所有三项都是 $o_{p^*}(1)$。

对于第一个右端项，我们注意到 $E^*\left[n^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\hat{v}_i^2u_i^*(A_n)_{ij}u_j^*\right]=0$ 和

$$
\begin{aligned}
&E^*\left[\left(n^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\hat{v}_i^2u_i^*(A_n)_{ij}u_j^*\right)^2\right]\\
&=n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1\neq i_1}\sum_{j_2\neq i_2}\hat{v}_{i_1}^2(A_n)_{i_1,j_1}E^*[u_{i_1}^*u_{j_1}^*u_{i_2}^*u_{j_2}^*](A_n)_{i_2,j_2}\hat{v}_{i_2}^2\\
&\leq\left(\max_i\sigma_i^{*2}\right)^2n^{-2}\sum_{i=1}^{n}\hat{v}_i^4\sum_{j\neq i}(A_n)_{ij}^2+\left(\max_i\sigma_i^{*2}\right)^2n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}\hat{v}_i^2\hat{v}_j^2(A_n)_{ij}(A_n)_{ji}=o_p(1),
\end{aligned}
$$

这可以从 $E^*[u_{i_1}^*u_{j_1}^*u_{i_2}^*u_{j_2}^*]=0$ 得出，除了以下情况：(1)$i_1=i_2$ 和 $j_1=j_2$，或 (2)$i_1=j_2$ 和 $i_2=j_1$，$\sum_{j\neq i}(A_n)_{ij}^2\leq\sum_{j=1}^{n}(A_n)_{ij}^2=(H_n)_{ii}(M_n)_{ii}^{-2}\leq(M_n)_{ii}^{-2}$，$\sum_{j\neq i}(A_n)_{ij}(A_n)_{ji}\leq\sum_{j=1}^{n}(A_n)_{ij}(A_n)_{ji}\leq(\min_i(M_n)_{ii})^{-2}$,$(\min_i(M_n)_{ii})^{-1}=O_p(1)$,$\max_i\sigma_i^{*2}=O_p(1)$,$(\max_i|\hat{v}_i|/\sqrt{n})^2=o_p(1)$，和 $n^{-1}\sum_{i=1}^{n}\hat{v}_i^2=O_p(1)$。

使用类似的论证,我们发现 (6) 中的第二个右边项通过使用 $n^{-1}\sum_{i=1}^{n}\hat{v}_i^2=O_p(1)$,$(\max_i|\hat{v}_i|/\sqrt{n})^2=o_p(1)$,$\max_i\sigma_i^{*2}=O_p(1)$ 和 $(\min_i(M_n)_{ii})^{-1}=O_p(1)$ 具有（条件）均值

$$
E^*\left[n^{-1}\sum_{i=1}^{n}\hat{v}_i^2((A_n)_{ii}-1)u_i^{*2}\right]=\left(n^{-1}\sum_{i=1}^{n}\hat{v}_i^2\right)^{-1}n^{-2}\sum_{i=1}^{n}\hat{v}_i^4\sigma_i^{*2}(M_n)_{ii}^{-1}=o_p(1),
$$

。类似地，我们通过使用柯西-施瓦茨不等式和假设 5 得到其（条件）方差为

$$
\begin{aligned}
&E^*\left[\left(n^{-1}\sum_{i=1}^{n}\hat{v}_i^2((A_n)_{ii}-1)u_i^{*2}\right)^2\right]\\
&=\left(n^{-1}\sum_{i=1}^{n}\hat{v}_i^2\right)^{-2}\left(n^{-4}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{v}_i^4\hat{v}_j^4(M_n)_{ii}^{-1}(M_n)_{jj}^{-1}E^*[u_i^{*2}u_j^{*2}]\right)=o_p(1),
\end{aligned}
$$

。

公式 (6) 中的第三个右侧项在原假设下，利用 $\max_i \sigma_i^{*2} = O_p(1)$,$(\min_i (M_n)_{ii})^{-1} = O_p(1)$,$\max_i |\hat{\mu}_i(\beta_0)|/\sqrt{n} = o_p(1)$ 和 $n^{-1}\sum_{i=1}^n \hat{v}_i^4 = n^{-1}\sum_{i=1}^n (\tilde{Q}_i + \tilde{v}_i)^4 \leq 4(n^{-1}\sum_{i=1}^n \tilde{Q}_i^4) + 4(n^{-1}\sum_{i=1}^n \tilde{v}_i^4) = O_p(1)$ 的条件下，具有条件均值 $E^*\left[n^{-1}\sum_{i=1}^n \sum_{j=1}^n \hat{v}_i^2 \hat{\mu}_i(\beta_0)(A_n)_{ij} u_j^*\right] = 0$, 和条件方差

$$
\begin{aligned}
E^*\left[\left(n^{-1}\sum_{i=1}^n\sum_{j=1}^n \hat{v}_i^2\hat{\mu}_i(\beta_0)(A_n)_{ij}u_j^*\right)^2\right] &= \sum_{j=1}^n \sigma_j^{*2}\left(n^{-1}\sum_{i=1}^n \hat{v}_i^2\hat{\mu}_i(\beta_0)(A_n)_{ij}\right)^2 \\
&\leq \left(\max_i \sigma_i^{*2}\right)\left(\min_i (M_n)_{ii}\right)^{-2}\left(\max_i |\hat{\mu}_i(\beta_0)|/\sqrt{n}\right)^2\left(n^{-1}\sum_{i=1}^n \hat{v}_i^4\right) = o_p(1),
\end{aligned}
$$

，这源自假设 1,2 和 4。

因此，我们得到方程 (6) 右侧的所有三项都是 $o_{p^*}(1)$，从而是 $n^{-1}\sum_{i=1}^n \hat{v}_i^2(y_i^* \acute{u}_i^* - u_i^{*2}) = o_{p^*}(1)$。

结合 (5)，我们得到在原假设下，$n^{-1}\sum_{i=1}^n \hat{v}_i^2(y_i^*\acute{u}_i^* - \sigma_i^{*2}) = o_{p^*}(1)$，这与 σ_i^{*2} 的定义一起进一步暗示了

$$
n^{-1}\sum_{i=1}^n \hat{v}_i^2 y_i^* \acute{u}_i^* - n^{-1}\acute{\Sigma}_n(\beta_0) = o_{p^*}(1). \tag{7}
$$

现在，设 $\bar{t}_n^* = (\hat{\beta}_n^* - \beta_0)/\sqrt{\acute{\Omega}_n(\beta_0)}$，其中 $\acute{\Omega}_n(\beta_0) = (\sum_{i=1}^n \hat{v}_i^2)^{-1}\acute{\Sigma}_n(\beta_0)(\sum_{i=1}^n \hat{v}_i^2)^{-1}$。然后，在原假设下，根据 van der Vaart and Wellner (1996) 的引理 2.9.5 和第 3 节中的自助数据生成过程，我们有

$$
\bar{t}_n^* = \left(n^{-1}\acute{\Omega}_n(\beta_0)\right)^{-1/2}\left(\sum_{i=1}^n \hat{v}_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{v}_i u_i^* \to^{d^*} N(0,1),
$$

依概率成立。进一步地，由 (7) 和斯卢茨基定理，我们有

$$
t_n^* = \left(n^{-1}\acute{\Omega}_n^*\right)^{-1/2}\left(n^{-1}\acute{\Omega}_n(\beta_0)\right)^{1/2}\bar{t}_n^* \to^{d^*} N(0,1),
$$

依概率成立。最后，结论由波利亚定理得出。■

C 固定效应模型的模拟结果

第二个考虑的模拟模型是面板数据的一元固定效应模型，类似于在 Jochmans (2022) 中考虑的那个。对于双下标数据 $(y_{(g,m)}, x_{(g,m)})$，可以写为

$$y_{(g,m)} = x_{(g,m)}\beta + \alpha_g + \epsilon_{(g,m)}, \quad g = 1, ..., G, \quad m = 1, ..., M,$$

其中 α_g 是一个特定组的截距项。对于这个模型，固定效应估计量等于 OLS 估计量在 $y_{(g,m)}$ 上使用 $x_{(g,m)}$ 和 G 组虚拟变量。根据 Jochmans (2022)，我们令 $x_{(g,m)} \sim i.i.d.N(0,1)$，$\epsilon_{(g,m)} \sim i.i.d.N(0,1)$，$\beta = 2$，以及 $\alpha_g = 0$ 对于所有 g。总样本量 $n = G \times M$ 设置为 100，我们设组数为 $G \in \{5, 10, 20, 25, 50\}$。模拟结果见表 A1。我们观察到基于 $HC0$、HCK 和 HCA 的检验都倾向于过度拒绝，尤其是在 $G = 50$ 情况下，而两种野性自助法程序也表现出更好的小样本性能。

G	5	10	20	25	50
HC0	0.069	0.081	0.093	0.109	0.191
HCK	0.064	0.067	0.066	0.071	0.191
HCA	0.107	0.102	0.105	0.110	0.123
Wild-G	0.046	0.045	0.043	0.042	0.042
Wild-R	0.038	0.037	0.037	0.035	0.036

表 A1: 面板数据的零假设拒绝频率