

- [lwdid](#)

- 研究需求声明
- 概述
- 环境要求
  - 软件
  - 数据
- 安装
- 快速开始
  - 1) Common Timing (小 N 或大 N)
  - 2) Staggered Adoption
  - 3) Repeated Cross-Section → Panel Aggregation
- 核心 API
- 方法选择指南
  - 变换方法
  - 何时使用 `detrend` 而非 `demean`
  - 估计器
- 推断行为
  - 最小样本要求
- 模式参数说明
  - Common Timing 模式 (`gvar=None`)
  - Staggered 模式 (提供 `gvar`)
  - 预处理期动态效应与平行趋势检验
  - 无预期效应稳健性检验
- 数据要求与校验规则
  - 非平衡面板
- 返回对象 (`LWDIDResults`)
- 诊断工具包
  - 趋势诊断
  - 敏感性分析
  - 聚类诊断
  - 选择机制诊断
- 当前边界与注意事项
- 文档
- 版本说明
- 获取帮助
- 贡献
- 引用

- 参考文献
- 许可协议
- 作者

# lwdid

面向面板数据的 Lee-Wooldridge 双重差分方法（Difference-in-Differences）

version 0.2.0 python 3.8+ license AGPL-3.0 docs passing



`lwdid` 实现了 Lee & Wooldridge (2025, 2026) 的滚动变换 (rolling-transformation) DiD 工作流，并在源码层面支持：

- 共同处理时点 (common timing) 设计 (`d + post`)；
- 错位处理时点 (staggered adoption) 设计 (`gvar`)；
- 季节性调整 (`Q`、`season_var`)，支持季度、月度或周度数据；
- 面向 OLS 路径的 HC0–HC4 稳健/聚类推断；
- 基于倾向得分流程的 IPW / IPWRA / PSM 估计器；
- 预处理期动态效应与平行趋势检验；
- 随机化推断 (`bootstrap` / `permutation`)；
- 诊断工具包：趋势评估、敏感性分析与聚类诊断；
- 事件研究图与结果导出工具。

## 研究需求声明

双重差分法 (DiD) 是社会科学中最广泛使用的因果推断方法之一。然而，现有实现通常假设较大的横截面样本量，在处理组或控制组单位数量较少时表现不佳。Lee & Wooldridge (2026) 提出了一种滚动变换方法，将面板 DiD 转化为截面处理效应问题，在经典线性模型假设下实现精确的 t 分布推断——即使仅有  $N_0 \geq 1$  个控制单位和  $N_1 \geq 1$  个处理单位（总计  $N \geq 3$ ）。Lee & Wooldridge (2025) 将该框架扩展至错位处理时点设计，支持 cohort-time 特定效应、多种估计器 (RA、IPW、IPWRA、PSM) 以及灵活的聚合策略。

`lwdid` 提供了两篇论文的统一 Python 实现，填补了现有生态系统中的空白——目前没有 Python 包同时提供：

- DiD 的精确小样本推断，
- 个体特异的滚动变换（去均值与去趋势），
- 共同处理时点与错位处理时点的集成支持，
- 高频面板数据的季节性调整，
- 用于平行趋势评估的预处理期动态效应，
- 内置的趋势异质性、敏感性分析与聚类诊断工具包。

## 概述

---

`lwdid` 遵循 Lee-Wooldridge 的滚动变换策略：先利用处理前信息在个体层面对面板结果变量进行变换，再在截面框架中估计处理效应。

该包面向三类实际 DiD 工作流设计：

- 使用精确/稳健 OLS 推断的共同处理时点 (`d`、`post`)；
- 支持 cohort-time 效应与聚合的错位处理时点 (`gvar`)；
- 在 DiD 估计前先将重复横截面聚合为面板格式的重复横截面聚合。

相较于单一固定估计器流水线，`lwdid` 提供了多种变换与估计器组合 (`ra`、`ipw`、`ipwra`、`psm`)，同时通过 `LWDIDResults` 统一了推断与输出 API。

## 环境要求

---

## 软件

- Python `>=3.8,<3.13`

- `numpy>=1.20,<3.0`
- `pandas>=1.3,<3.0`
- `scipy>=1.7,<2.0`
- `statsmodels>=0.13,<1.0`
- `scikit-learn>=1.0`
- `matplotlib>=3.3`
- `openpyxl>=3.1`

## 数据

估计前请确保：

- 使用长格式面板结构（每行对应一个 unit-time 观测）；
- (`ivar, time`) 键唯一；
- 所选模式下的处理变量编码有效：
  - 共同处理时点：`d + post`；
  - 错位处理时点：`gvar` 中 `0/NaN/inf` 表示从未处理；
- 所选 `rolling` 方法有足够的处理前观测。

更详细的校验规则见下文**数据要求与校验规则**。

## 安装

从 PyPI 安装：

```
pip install lwdid
```

## 快速开始

### 1) Common Timing (小 N 或大 N)

```
import pandas as pd
from lwdid import lwdid

df = pd.read_csv("data/smoking.csv")
```

```

res = lwddid(
    data=df,
    y="lcigssale",
    d="d",
    ivar="state",
    tvar="year",
    post="post",
    rolling="detrend",
    estimator="ra",
    vce="hc3",           # None / hc0 / hc1(robust) / hc2 / hc3 / hc4 /
cluster
    alpha=0.05,
)

print(res.summary())
print(res.att, res.se_att, res.pvalue)

```

## 2) Staggered Adoption

```

import pandas as pd
from lwddid import lwddid

df = pd.read_csv("data/castle.csv")
df["gvar"] = df["effyear"].fillna(0) # 0 / NaN / inf => never treated

res = lwddid(
    data=df,
    y="lhomicide",
    ivar="sid",
    tvar="year",
    gvar="gvar",
    rolling="demean",
    estimator="ra",
    control_group="not_yet_treated",
    aggregate="none",
    vce="hc3",
    include_pretreatment=True,
)

res.plot_event_study(aggregation="weighted", title="Castle Law Event Study")

```

## 3) Repeated Cross-Section → Panel Aggregation

```

from lwddid import aggregate_to_panel

agg = aggregate_to_panel(
    data=raw_df,
    unit_var="state",
    time_var="year",
    outcome_var="outcome",
    weight_var="survey_weight",      # 可选
    treatment_var="treated",        # 可选一致性检查
)
panel_df = agg.panel_data

```

## 核心 API

```

lwddid(
    data, y, d=None, ivar=None, tvar=None, post=None, rolling="demean", *,
    gvar=None, control_group="not_yet_treated", estimator="ra",
    aggregate="cohort",
    balanced_panel="warn", ps_controls=None, trim_threshold=0.01,
    return_diagnostics=False, n_neighbors=1, caliper=None,
    with_replacement=True,
    match_order="data", vce=None, controls=None, cluster_var=None,
    alpha=0.05,
    ri=False, rireps=1000, seed=None, ri_method="bootstrap",
    graph=False, gid=None, graph_options=None,
    season_var=None, Q=4, auto_detect_frequency=False,
    include_pretreatment=False, pretreatment_test=True,
    pretreatment_alpha=0.05,
    exclude_pre_periods=0
)

```

## 方法选择指南

## 变换方法

方法	数据类型	解决的问题	实际优势
demean	年度/有序面板	个体间水平差异	rolling DiD 基线实现最直接
detrend	年度/有序面板	个体特异线性趋势 (CHT 风格关键)	当 cohort 间前趋势不同更稳健；小样本推荐 (Lee &

方法	数据类型	解决的问题	实际优势
		注)	Wooldridge, 2026)
demeanq	季节性面板（通过 Q 支持季度/月度/周度）	季节性水平效应 + 个体异质性	无需显式重写 FE 模型即可处理周期结构
detrendq	季节性面板	个体趋势 + 季节结构	当趋势与季节性同时存在时最稳健

四种变换方法均同时支持共同处理时点和错位处理时点设计。

## 何时使用 **detrend** 而非 **demean**

在条件异质性趋势 (CHT) 假设下 (Lee & Wooldridge, 2025)，每个处理 cohort 可能有独立的线性趋势。**detrend** 方法移除个体特异线性趋势，放松了标准平行趋势假设。以下情况建议使用 **detrend**：

- 处理前结果趋势在 cohort 间存在明显差异，
- 怀疑存在个体特异线性趋势（如差异化增长率），
- 小样本场景下 HC3 + detrending 提供最佳覆盖率 (Lee & Wooldridge, 2026)。

## 估计器

估计器	最适配的数据/设定	主要用途	优势	推断分布
ra	小样本或大样本、偏好线性调整流程	基线 ATT 估计	OLS 路径透明 + 完整 vce 支持	t 分布
ipw	存在协变量不平衡且可建 PS 模型	对照组重加权到处理组支持集	加权估计器更简洁	正态分布 (z)
ipwra	担心模型设定错误	双重稳健 ATT	PS 模型或结果模型任一正确即可一致	正态分布 (z)
psm	处理组与对照组可匹配	最近邻匹配 ATT	匹配样本解释直观	正态分布 (z)

说明：IPW、IPWRA 和 PSM 当前使用基于正态分布的推断。**ra** 估计器使用具有适当自由度 ( $df = N_1 + N_0 - 2$ ) 的 t 分布推断，根据 Lee & Wooldridge (2026) 推荐用于小样

本。未来版本将把 IPW、IPWRA 和 PSM 也迁移至 t 分布推断。

## 推断行为

---

- OLS 路径 (`ra`) 支持的 `vce` 选项: `None`、`hc0`、`hc1/robust`、`hc2`、`hc3`、`hc4`、`cluster`。
- 小样本推荐使用 HC3 (Lee & Wooldridge, 2026)。
- 聚类 OLS 使用 `df = G - 1` (`G` 为聚类数)。当分析单位细于政策层级时 (如县级数据、州级政策)，应在政策层级聚类 (Lee & Wooldridge, 2026)。
- `ra` 的 p 值与置信区间基于 t 分布临界值 ( $df = N_1 + N_0 - 2$ )。
- `ipw`、`ipwra` 和 `psm` 的 p 值与置信区间基于正态临界值。
- `ri=True` 可启用随机化推断以获得精确 p 值：
  - `ri_method="bootstrap"` (默认)；
  - `ri_method="permutation"`。
- 事件研究的加权聚合 (`plot_event_study(aggregation="weighted")`) 通过事件时间聚合工具使用 t 分布构造置信区间，采用保守 `df` 策略 (取各 cohort 中最小 `df`)。

## 最小样本要求

根据 Lee & Wooldridge (2026)，精确推断要求：

- $N_0 \geq 1$  (至少 1 个控制单位)，
- $N_1 \geq 1$  (至少 1 个处理单位)，
- $N = N_0 + N_1 \geq 3$  (至少 3 个总单位)。

## 模式参数说明

---

### Common Timing 模式 (`gvar=None`)

必需参数: `d`、`post`、`ivar`、`tvar`。

- 该模式下 `control_group` 与 `aggregate` 会被忽略。
- `d` 必须在个体内时间不变。
- `post` 必须单调 (不允许处理撤销)。

# Staggered 模式 (提供 gvar)

必需参数: `gvar`、`ivar`、`tvar`。

- 若同时提供 `gvar` 与 `d/post`, 则 `d` 和 `post` 会被忽略。
- `gvar` 编码规则: 正值表示首次处理期; `0 / NaN / np.inf` 表示从未处理。
- `control_group`: `not_yet_treated`、`never_treated`、`all_others`。
- `aggregate`: `none`、`cohort`、`overall`。
- 当 `aggregate in {"cohort", "overall"}` 时必须有 never-treated 单位; 控制组策略会自动切换为 `never_treated`。

## 预处理期动态效应与平行趋势检验

设置 `include_pretreatment=True` (staggered 模式) 可计算预处理期变换后结果, 用于平行趋势评估。变换使用未来预处理期  $\{t+1, \dots, g-1\}$  作为参考 (Lee & Wooldridge, 2025) :

- **去均值 (Demeaning)**:  $\hat{y}_{itg} = Y_{it} - \text{mean}(Y_{i,t+1}, \dots, Y_{i,g-1})$
- **去趋势 (Detrending)**:  $\ddot{Y}_{itg} = Y_{it} - \text{对未来预处理期结果关于时间回归的拟合值}$

时期  $t = g-1$  作为锚点 (参考基准), 用于事件研究图的可视化。设置 `pretreatment_test=True` 可运行各期单独 t 检验和联合 F 检验 ( $H_0$ : 所有预处理期 ATT = 0)。

## 无预期效应稳健性检验

设置 `exclude_pre_periods=k` 可排除处理前紧邻的  $k$  个时期, 使其不参与变换所用的预处理样本。当单位可能在正式处理前调整行为时, 此选项可用于检验无预期 (no-anticipation) 假设的敏感性。

## 数据要求与校验规则

- 长格式面板 (一行 = 一个单位-时期观测)。
- (`ivar, time`) 组合必须唯一。
- 最小样本量校验: `N >= 3` 个单位。

- `controls` 按时间不变回归变量处理。
- 季节性模式要求 `season_var` (或兼容旧接口的 `tvar=[year, quarter]`)，且取值必须在 `1..Q`。
- 原始输入中不应出现保留的内部列名：
  - `d_`、`post_`、`tindex`、`tq`、`ydot`、`ydot_postavg`、`firstpost`。

## 非平衡面板

`balanced_panel` 参数控制非平衡面板的处理方式：

- `"warn"` (默认)：发出警告并提供选择机制诊断信息。
- `"error"`：面板不平衡时抛出错误。
- `"ignore"`：静默处理。

选择机制可依赖于未观测的时间不变异质性，但不能系统性地依赖于  $Y_{\{it\}(\infty)}$  的冲击 (Lee & Wooldridge, 2025)。最小预处理期观测要求：

- `demean`：每个单位至少 1 个预处理期。
- `detrend`：每个单位至少 2 个预处理期。
- `demeanq`：每个单位至少  $Q + 1$  个预处理期。
- `detrendq`：每个单位至少  $Q + 2$  个预处理期。

## 返回对象 (`LWIDResults`)

---

核心字段：

- `att`、`se_att`、`t_stat`、`pvalue`、`ci_lower`、`ci_upper`
- `nobs`、`n_treated`、`n_control`、`df_resid`、`df_inference`
- `att_by_period` (共同处理时点)
- `att_by_cohort_time`、`att_by_cohort`、`att_overall` (错位处理时点，取决于 `aggregate`)
- `att_pre_treatment`、`parallel_trends_test` (当 `include_pretreatment=True`)

关键方法：

- `summary()`
- `plot(...)` (共同处理时点风格的变换后结果图)

- `plot_event_study(...)` (错位处理时点事件研究图)
- `to_excel(...)`、`to_csv(...)`、`to_latex(...)`

# 诊断工具包

`lwdid` 内置了针对常见方法论问题的诊断模块：

## 趋势诊断

```
from lwdid import test_parallel_trends, diagnose_heterogeneous_trends,
recommend_transformation

# 检验平行趋势假设
pt_result = test_parallel_trends(data, y="outcome", ivar="unit",
tvar="year", gvar="gvar")

# 诊断 cohort 间的异质性趋势
ht_diag = diagnose_heterogeneous_trends(data, y="outcome", ivar="unit",
tvar="year", gvar="gvar")

# 获取变换方法推荐 (demean vs detrend)
rec = recommend_transformation(data, y="outcome", ivar="unit", tvar="year",
gvar="gvar")
```

## 敏感性分析

```
from lwdid import robustness_pre_periods, sensitivity_no_anticipation,
sensitivity_analysis

# 预处理期数量的稳健性检验
rob = robustness_pre_periods(data, y="outcome", ivar="unit", tvar="year",
d="d", post="post")

# 无预期假设违反的敏感性分析
na_sens = sensitivity_no_anticipation(data, y="outcome", ivar="unit",
tvar="year", d="d", post="post")

# 综合敏感性分析
comp = sensitivity_analysis(data, y="outcome", ivar="unit", tvar="year",
d="d", post="post")
```

# 聚类诊断

```
from lwdid import diagnose_clustering, recommend_clustering_level

# 诊断聚类结构
clust_diag = diagnose_clustering(data, ivar="county",
potential_cluster_vars=["state", "region"], gvar="gvar")

# 获取聚类层级推荐
rec = recommend_clustering_level(data, ivar="county", tvar="year",
potential_cluster_vars=["state", "region"], gvar="gvar")
```

## 选择机制诊断

```
from lwdid import diagnose_selection_mechanism

# 诊断非平衡面板的选择机制
sel_diag = diagnose_selection_mechanism(data, ivar="unit", tvar="year",
y="outcome")
```

## 当前边界与注意事项

- 错位处理时点模式下 `graph=True` 不会直接执行绘图；请在估计完成后使用 `results.plot_event_study()`。
- `balanced_panel="error"` 为严格模式；`warn/ignore` 不会触发同样的硬性检查。
- 当重叠较弱时，倾向得分估计器可能修剪大量观测 (`trim_threshold`)，有效样本量可能显著减少。
- IPW、IPWRA 和 PSM 当前使用基于正态分布的推断；未来版本将迁移至与 `ra` 一致的 t 分布推断。

## 文档

完整 API 文档请访问 [lwdid.readthedocs.io](https://lwdid.readthedocs.io)。

# 版本说明

---

- 当前包版本: `0.2.0` (`pyproject.toml`)。
- 包含广义季节性支持 (`Q`、`season_var`) 及错位处理时点季节性变换 (staggered 模式下的 `demeanq`、`detrendq`)。
- 预处理期动态效应与平行趋势检验 (`include_pretreatment`、`pretreatment_test`)。
- 无预期效应稳健性检验 (`exclude_pre_periods`)。
- 诊断工具包: 趋势诊断、敏感性分析、聚类诊断、选择机制诊断。
- Wild cluster bootstrap 推断。

## 获取帮助

---

- 通过 [GitHub Issues](#) 报告 bug 或提出功能建议。
- 方法论问题请参阅引用的论文, 或在仓库中发起讨论。

## 贡献

---

欢迎贡献。请参阅 [CONTRIBUTING.md](#) 了解如何报告 bug、建议功能和提交 pull request。

本项目遵循 [Contributor Covenant 行为准则](#)。

## 引用

---

如果您在研究中使用了 `lwddid`, 请同时引用软件和相关论文。机器可读的引用文件见 [CITATION.cff](#)。

```
@software{cai_lwddid_2025,
  author = {Cai, Xuanyu and Xu, Wenli},
  title = {lwddid: Lee--Wooldridge Difference-in-Differences for Panel
Data},
  version = {0.2.0},
  year = {2025},
  url = {https://github.com/gorgeousfish/lwddid-py}
}
```

# 参考文献

---

- Lee, Soo Jeong and Wooldridge, Jeffrey M., *Simple Approaches to Inference with Difference-in-Differences Estimators with Small Cross-Sectional Sample Sizes* (January 03, 2026). Available at SSRN: <https://ssrn.com/abstract=5325686> or <http://dx.doi.org/10.2139/ssrn.5325686>
- Lee, Soo Jeong and Wooldridge, Jeffrey M., *A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data* (December 24, 2025). Available at SSRN: <https://ssrn.com/abstract=4516518> or <http://dx.doi.org/10.2139/ssrn.4516518>

# 许可协议

---

本项目采用 **GNU Affero 通用公共许可证 v3.0 或更高版本 (AGPL-3.0-or-later)** 授权。完整许可文本见 **LICENSE**。

# 作者

---

蔡炫宇 (Xuanyu Cai)、许文立 (Wenli Xu)