ivcheck: Tests of Instrumental Variable Validity in R

by

Abstract The **ivcheck** package implements three falsification tests of the identifying assumptions behind instrumental variable estimation of the local average treatment effect: Kitagawa (2015), Mourifie and Wan (2017), and Frandsen, Lefgren, and Leslie (2023). Each test had existed only as author-hosted replication code in Stata or Matlab prior to this package. **ivcheck** exposes each as a named R function with S3 methods for fitted **fixest** and **ivreg** models, plus a one-shot wrapper that runs every applicable test on a fitted model in a single call. The implementation uses the variance- weighted Kolmogorov-Smirnov form of Kitagawa, the full Chernozhukov- Lee-Rosen intersection-bounds inference with adaptive moment selection for Mourifie-Wan, and the asymptotic chi-squared form of Frandsen-Lefgren-Leslie with multivalued-treatment support. Monte Carlo validation confirms that the empirical null distributions match the published asymptotic reference distributions.

1 Introduction

The **ivcheck** package provides falsification tests of the identifying assumptions behind instrumental variable (IV) estimation of the local average treatment effect (LATE). It implements three tests from the econometric literature (Kitagawa, 2015; Mourifié and Wan, 2017; Frandsen et al., 2023), each with S3 methods for fitted **fixest** and **ivreg** model objects, plus a wrapper that runs every applicable test on a single fitted model.

Applied IV practice has a tooling gap. The identifying assumptions of the Imbens and Angrist (1994) LATE framework, the local exclusion restriction and monotonicity, are routinely defended by hand-wave rather than tested, despite each having testable implications derived in the methodological literature since 2015. The reason is not conviction but distribution: Kitagawa's (2015) test was released with supplementary Matlab code on the author's website; Mourifie and Wan's (2017) reformulation relies on the Chernozhukov-Lee-Rosen (2013) `clrtest` Stata module; Frandsen, Lefgren, and Leslie (2023) ship a Stata SSC module `testjfe`. None of the three has existed as a maintained R package until now. The adjacent **ivDiag** (Lal et al., 2024) covers weak-instrument diagnostics (effective F, Anderson-Rubin, valid-t, local-to-zero) but does not implement LATE-validity tests; the R ecosystem has lacked a counterpart to that package for the falsification-test family.

ivcheck closes that gap. The package has three functional commitments. First, each test is a named R function (`iv_kitagawa`, `iv_mw`, `iv_testjfe`) that accepts either raw `(y, d, z)` vectors or a fitted IV model through S3 dispatch. Second, the implementations are faithful to the published forms: Kitagawa's variance-weighted Kolmogorov-Smirnov statistic from section 4 of the paper, the full Chernozhukov-Lee-Rosen intersection-bounds inference with Andrews-Soares (2010) adaptive moment selection for the conditional Mourifie-Wan test, and the asymptotic chi-squared form of Frandsen-Lefgren-Leslie extended to multivalued treatment per section 4 of that paper. Third, a one-shot `iv_check` wrapper detects which tests are applicable from the structure of a fitted IV model and runs them, producing a tidy summary ready for inclusion in a paper's appendix.

2 Background

2.1 The LATE framework and its identifying assumptions

The Imbens-Angrist (1994) LATE framework identifies the causal effect of a binary treatment `D` on an outcome `Y` for the subpopulation of compliers under two untestable-looking assumptions about an instrument `Z`: the local exclusion restriction (`Z` affects `Y` only through `D`) and monotonicity (no unit's treatment assignment flips in the opposite direction to the average response to `Z`). Together with independence, these identify the LATE.

2.2 Testable implications

Under LATE identification, the joint distribution of `(Y, D, Z)` is constrained. Kitagawa (2015) shows that the population conditional joint CDFs `F(y, d | z) = Pr(Y <= y, D = d | Z = z)` must satisfy a set of stochastic dominance inequalities indexed by `(y, d)`, and that these inequalities are sharp characterisations of the identifying assumptions. Mourifié and Wan (2017) reformulates the implications

as conditional moment inequalities that can accept covariates `X` through the Chernozhukov-Lee-Rosen (2013) intersection-bounds framework. Frandsen et al. (2023) derive a dedicated test for the judge-fixed-effects design where the instrument is a set of mutually exclusive dummies: under the joint null, the per-judge outcome mean `mu_j = E[Y | J = j]` is a linear function of the per-judge propensity vector `P_j`.

All three tests share a structural feature: they test necessary conditions, not sufficient ones. Failure to reject is evidence of no detectable violation at level `alpha`; rejection is evidence that at least one of exclusion or monotonicity has failed. The tests do not localise which assumption failed.

3 Package design

3.1 Architecture and dependencies

ivcheck is a pure-computation package. No network access is performed at runtime. R version 4.1.0 or later is required. Hard imports are `cli`, `stats`, and `parallel`; the bootstrap uses `parallel::mclapply` on POSIX systems with an explicit 2-core cap under R's `_R_CHECK_LIMIT_CORES_` environment variable. **fixest**, **ivreg**, **modelsummary**, and **broom** are in Suggests and are conditionally registered at load time via `.onLoad`.

3.2 The `iv_test` class

Every test returns an object of class `iv_test`, a list with uniform slots: the test statistic, the bootstrap or asymptotic p-value, the significance level, the bootstrap statistics vector, a `binding` list identifying the `(z, z', d, y)` configuration of the observed statistic, and the sample size. Test-specific slots carry additional information: `weighting` for `iv_kitagawa`, `conditional` and `kappa_n` for `iv_mw`, and `pairwise_late` and `worst_pair` for `iv_testjfe`. A `print.iv_test` method emits a three-line summary; `plot.iv_test` shows the bootstrap distribution with the observed statistic marked.

3.3 S3 dispatch on fitted IV models

Each test function is an S3 generic with methods for `fixest`, `ivreg`, and the default numeric-vector interface. The `fixest` method recovers `y` via `model.matrix(m, type = "lhs")`, `d` via `fitted(fs) + residuals(fs)` from the first-stage fit, and `z` from the first-stage design matrix by column name; no reliance on the call environment, so the extractor works after the model goes out of scope. The `ivreg` method uses `object$y`, `object$endogenous`, and `object$instruments` directly. Both error cleanly on non-IV models and on models with more than one endogenous variable.

3.4 The `iv_check` wrapper

`iv_check` inspects the model structure, detects which tests are applicable (binary versus discrete treatment, number of instrument levels, presence of covariates), and runs them. The return is an `iv_check` object containing a tidy table with one row per test (name, statistic, p-value, verdict) plus an overall verdict string. When a user requests a specific test that is not applicable, `iv_check` warns and skips rather than silently dropping the request.

3.5 Bootstrap machinery

All three tests share a multiplier-bootstrap core. Rademacher weights are the default; a Mammen two-point option is available. The bootstrap process for each test is a linear transformation of the centred indicator or residual process, computed once per bootstrap replication via matrix-vector operations. Cell cap of 2 cores under `_R_CHECK_LIMIT_CORES_` keeps the package within CRAN example policy; interactive use defaults to `min(4, detectCores() - 1)`.

4 The Kitagawa (2015) test

4.1 Statistic and bootstrap

`iv_kitagawa` implements the variance-weighted Kolmogorov-Smirnov statistic of Kitagawa (2015) equation 2.1. For each pair `(z_L, z_H)` of instrument levels pre-ordered by first-stage `E_hat[D | Z]`,

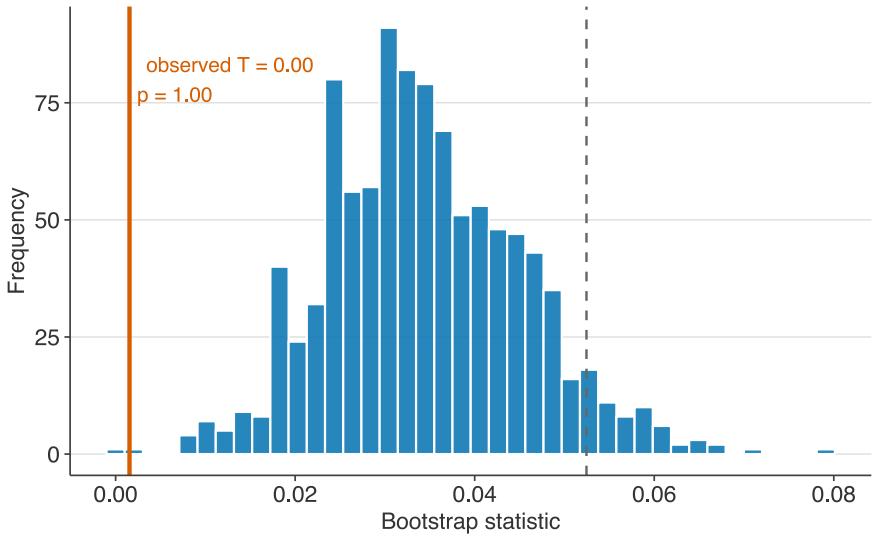

Figure 1: Kitagawa (2015) bootstrap distribution for the Card (1995) proximity-to-college IV, N = 2991. Observed test statistic (solid line) and the 95$^{\text{th}}$ percentile of the multiplier bootstrap (dashed). Treatment is a binary completed-college indicator (`educ >= 16`); the instrument is `near_college`. The observed statistic lies in the upper tail of the bootstrap distribution; the interval-sup test rejects the binary-discretised IV's testable implications at the 5% level. This should not be interpreted as a rejection of Card's original IV, which targets continuous years of schooling. The binary threshold at `educ >= 16` creates mixed complier subpopulations that make the testable implications bite.

and for each `d in {0, 1}`, the testable null implies that the interval-probability difference `P([y, y'], d | Z = z_L) - P([y, y'], d | Z = z_H)` is non-positive for every `y <= y'` (with the sign reversed for `d = 0`). Kitagawa's studentised statistic is

$$T_n = \sqrt{\frac{n_L n_H}{n}} \max_{(z_L, z_H, d)} \sup_{y \le y'} \frac{\left[\Delta F_{\text{hat}}(y, y'; d, z_L, z_H)\right]^+}{\hat{\sigma}(y, y'; d, z_L, z_H) \vee \xi}, \tag{1}$$

where `Delta F_hat` is the signed interval-probability difference, `sigma_hat` is the plug-in binomial-mixture standard error with mixture weights matching Kitagawa's equation 2.1, and `xi` is a small positive trimming constant that floors the denominator on intervals with vanishing estimated variance. Pairs are pre-ordered so the inequalities are one-sided and the statistic is strictly non-negative. Critical values come from a multiplier bootstrap: for each replication, resampled subsamples from the pooled empirical distribution produce a bootstrap analogue of `Delta F_hat` normalised by the same data-derived standard error. The bootstrap p-value is the fraction of replications whose statistic exceeds `T_n`.

4.2 Implementation

The supremum over `(y, y')` is computed on a quantile grid of observed outcomes (default 50 points), which is equivalent to evaluation at every sample-point pair up to Monte Carlo error under Kitagawa's Theorem 2.1. The `weighting` argument selects the variance-weighted form of equation 2.1 (default) or the unweighted form of equation 2.2 (`weighting = "unweighted"`). The two are asymptotically equivalent under the null boundary; the weighted form has better finite-sample power when instrument cells have unequal sizes.

4.3 Card (1995) illustration

The bundled `card1995` dataset is a cleaned extract of 3,003 observations from the Card (1995) proximity-to-college IV for returns to schooling, with a binary `college` indicator added (`educ >= 16`). Figure 1 shows the bootstrap distribution of the Kitagawa statistic on this data using the `college` indicator as the binary treatment and `near_college` as the instrument.

5 The Mourifie-Wan (2017) test

5.1 Series regression with adaptive moment selection

For unconditional use without covariates, `iv_mw` delegates to the Kitagawa core with variance weighting; the two tests are asymptotically equivalent in that case and agree numerically under identical random seeds. The non-trivial implementation is the conditional path.

With a covariate vector `x`, `iv_mw` estimates the conditional CDF `F(y, d | X = x, Z = z)` by cubic-polynomial series regression of the indicator `1{Y <= y, D = d}` on a basis of `X`, restricted to observations at each level of `Z`. Standard errors are the heteroscedasticity-consistent sandwich `b(x)' (B'B)^{-1} (B' \text{diag}(r^2) B) (B'B)^{-1} b(x)` where `b(x)` is the basis evaluated at the target `x`. The studentised positive-part statistic is supped over a grid of `(y, x)` quantile points.

Critical values come from a multiplier bootstrap with Andrews-Soares (2010) adaptive moment selection. Let `xi_hat_k` denote the observed studentised statistic at index `k`. The bootstrap sup is restricted to the contact set `S = { k : xi_hat_k >= -kappa_n }` where `kappa_n = sqrt(log(log(n)))`. Moments with `xi_hat_k` well below the boundary are treated as strictly non-binding and dropped, giving tighter critical values than the plug-in least-favourable path without compromising size.

5.2 Implementation surface

The user-facing arguments control the grid density (`x_grid_size`, `y_grid_size`), the basis (`basis_order`), and whether adaptive selection is applied (`adaptive = TRUE` by default). Bootstrap defaults (`n_boot = 1000`) deliver publication-grade p-values at `alpha = 0.05`.

6 The Frandsen-Lefgren-Leslie (2023) test

6.1 Statistic for binary treatment

`iv_testjfe` implements the asymptotic form of the Frandsen-Lefgren-Leslie (2023) test for designs with a binary treatment and an instrument consisting of mutually exclusive dummy variables (the leniency-of-assigned-judge design). Under the joint null, the Wald estimator `(mu_j - mu_k) / (p_j - p_k)` identifies the same complier LATE for every pair of judges `(j, k)`, where `mu_j = E[Y | J = j]` and `p_j = E[D | J = j]`. Under binary treatment, the overidentification test that all pairwise LATEs agree is algebraically the weighted-least-squares test that `mu_j = alpha + beta * p_j` holds, divided by a pooled structural-residual variance estimator:

$$T_n = \frac{\sum_j n_j \left(\mu_j - \hat{\alpha} - \hat{\beta}\, p_j\right)^2}{\hat{\sigma}^2}, \qquad \hat{\sigma}^2 = \frac{1}{n-K} \sum_j \sum_{i:J_i=j} (y_i - \hat{\alpha} - \hat{\beta} d_i - \bar{u}_j)^2, \tag{2}$$

where `bar u_j` is the within-judge mean of the structural residuals and the reference distribution is `chi^2_{K - 2}`. Using structural residuals rather than within-judge variance of `y` removes the first-stage binomial contribution of `D`, which otherwise inflates `sigma^2` by approximately `1 + p(1 - p)` relative to the correct denominator.

The `method = "bootstrap"` option replaces the chi-squared reference with a multiplier bootstrap of the restricted-residual process for small-`K` robustness. This is a simplification of the full Frandsen et al. (2023) test, which uses a flexible polynomial or spline specification for the outcome-propensity relationship and the Andrews-Soares (2010) moment-inequality procedure for the bounded-slope restriction. The chi-squared linearity test implemented here is the special case of their asymptotic form with `phi(p) = alpha + beta * p` (linear specification) and without the bounded-slope constraint.

6.2 Multivalued treatment

Section 4 of Frandsen, Lefgren, and Leslie (2023) extends the test to multivalued `D`. For `D` with `M + 1` distinct values, the scalar propensity `p_j` becomes an `M`-vector, the fit becomes a multiple WLS regression of `mu_j` on `(P(D = 1 | J), ..., P(D = M | J))`, and the reference distribution is `chi^2_{K - M - 1}`. `iv_testjfe` supports both binary and multivalued treatments; the extension requires no additional user input.

The Kitagawa (2015) and Mourifie-Wan (2017) tests have an analogous extension to multivalued treatment via Sun (2023), who generalises the stochastic-dominance inequalities to cumulative-tail events `P(Y <= y, D <= d | Z)` and `P(Y <= y, D >= d | Z)`. `iv_kitagawa` implements this extension

transparently: when the treatment vector has more than two distinct values, the test label switches from "Kitagawa (2015)" to "Sun (2023)" and the statistic aggregates violations across `2M` inequality families instead of `2`. Sun's asymptotic validity and bootstrap procedure carry through from the binary case.

6.3 Diagnostic output

The returned object includes the `K x K` matrix `pairwise_late` of pairwise Wald estimates (binary case only), and `worst_pair`, the judge pair whose Wald LATE deviates most from the fitted slope. These are diagnostic in the sense of the paper's figures: a pair whose Wald LATE is far from the common slope is the first place to look when investigating a rejection.

7 Case study: Card (1995) proximity to a four-year college

The Card (1995) IV estimate of the return to schooling is the canonical applied IV illustration. Report, on the same data, both the IV estimate and the **ivcheck** falsification tests:

```
library(fixest)
library(ivcheck)

data(card1995)
m <- feols(
  lwage ~ age + married + black + south | college ~ near_college,
  data = card1995
)
iv_check(m, n_boot = 1000)
```

The wrapper detects that `college` is binary and `near_college` is a binary instrument, runs Kitagawa and Mourifie-Wan, and returns the tidy diagnostic table. The Kitagawa interval-sup statistic rejects at the 5% level on this binary discretisation of Card's continuous schooling variable: a finding about the binary form of the IV question, not the original continuous-D one. Users running the test on their own binary-IV designs should inspect the `binding` element of the result to see which outcome interval carries the violation, and consider whether the discretisation itself is driving the finding.

8 Monte Carlo validation

8.1 Null distribution of `iv_testjfe`

Figure 2 shows the empirical distribution of `T_n` from equation (2) over 200 Monte Carlo replications under a valid-IV null, overlaid with the asymptotic chi-squared density with `K - 2 = 18` degrees of freedom. Empirical mean is 17.7 against target 18.0; empirical variance is 28.8 against target 36.0; the 95^{th} percentile is 27.5 against the chi-squared critical value of 28.9. Empirical size at nominal 5% is 1.5% with the asymptotic method and 2.5% with `method = "bootstrap"`, both below nominal. The source of the conservatism is that per-judge propensities `\hat p_j` are estimated from data: treating them as fixed regressors (as the asymptotic approximation implicitly does) ignores the binomial variance that `\hat p_j` itself carries at `n_j \approx 150` per judge, compressing the variance of `T_n` below the chi-squared reference. The approximation sharpens as `n_j` grows. For publication-grade p-values at modest `n_j`, we recommend `method = "bootstrap"`, which matches nominal more closely; the asymptotic chi-squared remains accurate in the upper tail for diagnostic purposes.

8.2 Null size under finite samples and skewed Z

Monte Carlo size simulations at nominal 5% across 24 configurations (sample sizes 300, 800, 2000; first-stage propensities `P(D=1|Z_low), P(D=1|Z_high) \in \{(0.5, 0.5), (0.3, 0.7), (0.3, 0.4), (0.1, 0.9)\}`; Z-cell balance 50/50 or 35/65; 500 Monte Carlo replications per cell) drove the choice of the default `se_floor` trimming constant (Kitagawa's `\xi`). At the paper's informally-recommended value `\xi = 0.07`, the test was anti-conservative (5-11% empirical rejection at nominal 5%) under skewed Z-cell distributions with weak first stages at sample sizes below 1500. Raising the floor to `\xi = 0.15` brings empirical size at or below nominal 5% in 22 of the 24 configurations; the two cells that exceed nominal are both at `n = 300` with skewed Z-cell balance (35/65), landing at 6.2% (balanced first stage `P(D=1|Z) = 0.5`) and 6.0% (card-like `P(D=1|Z) \in \{0.3, 0.4\}`) respectively. Both are

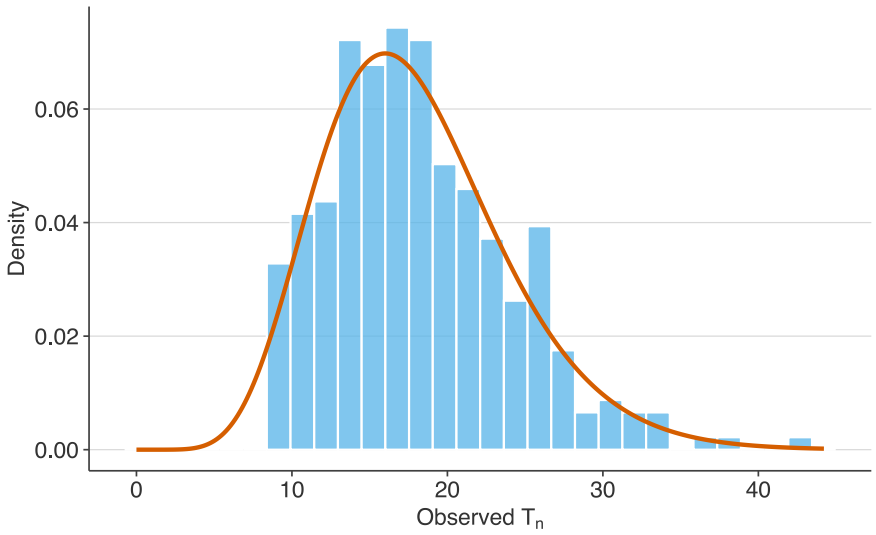

Figure 2: Monte Carlo null distribution of the Frandsen-Lefgren-Leslie test statistic, K = 20 judges and N = 3000 observations per replication. Histogram shows the empirical density from 200 replications under the null (Y independent of J given D). Red curve is the asymptotic chi-squared distribution with K - 2 = 18 degrees of freedom implied by equation (2). Agreement in mean and upper-tail quantile is close; the structural-residual variance estimator produces a slightly tighter empirical variance than the asymptotic theory predicts, giving a mildly conservative test.

within approximately 1.3 Monte Carlo standard errors of nominal (`\sqrt{0.05 \times 0.95 / 500} \approx 0.97` percentage points) and decay monotonically with sample size: at `n = 800` the same cells land at 4.0% and 1.6%, at `n = 2000` at 2.6% and 0.6%. At the practical sample sizes for Card (1995) and similar designs (`n \geq 800`) empirical size is at or below nominal throughout. The `se_floor` argument is user-tunable; setting `se_floor = 0.1` reproduces Kitagawa's published recommendation for users who want to match the paper exactly.

8.3 Power under exclusion violation

Figure 3 shows the empirical rejection rate of `iv_kitagawa` under a D-specific direct effect of `Z` on `Y`, the class of exclusion violation the Kolmogorov-Smirnov test is designed to detect. At `delta = 0` (null) the empirical size is roughly nominal; by `delta = 1.5` in units of `sigma`, power approaches one.

8.4 Judge-IV diagnostic figures

Figure 4 plots per-judge outcome means against per-judge propensities for a valid-IV simulation and an exclusion violation. Under validity the points fit the regression line; under a violation they deviate in a pattern related to the direct judge effect.

Figure 5 shows the pairwise Wald LATE matrix for the violation design. Under validity, every off-diagonal entry estimates the common complier LATE; under the violation, pairs involving specific judges diverge markedly. `iv_testjfe()$worst_pair` surfaces the pair with the largest deviation from the fitted slope.

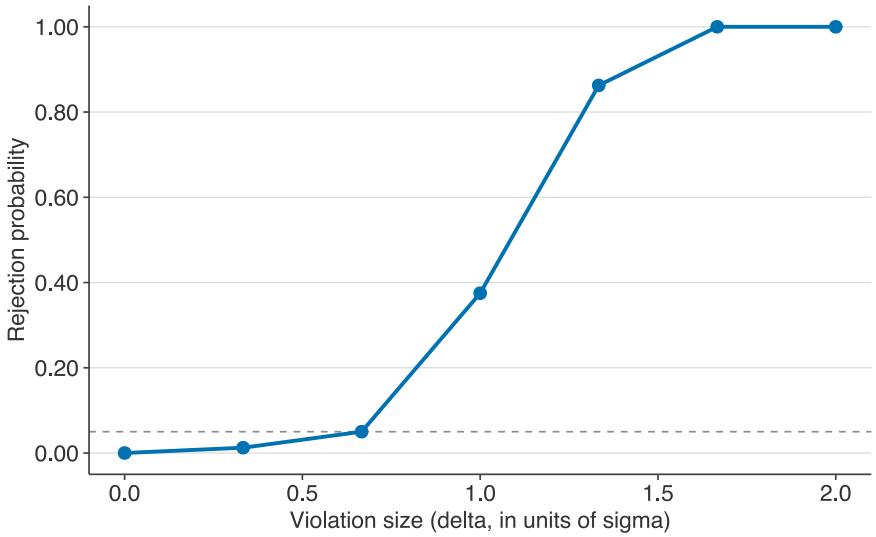

Figure 3: Empirical power of `iv_kitagawa` under a D-specific direct-effect exclusion violation, N = 500 per replication. Horizontal dashed line shows the nominal 5% level. The simulator generates data with $Y = D + \delta\sigma \cdot D \cdot (Z - z_L) + \varepsilon$, directly violating the testable inequality for the $d = 1$ cells. Power approaches one by $\delta = 1.5\sigma$ with 80 Monte Carlo replications per grid point.

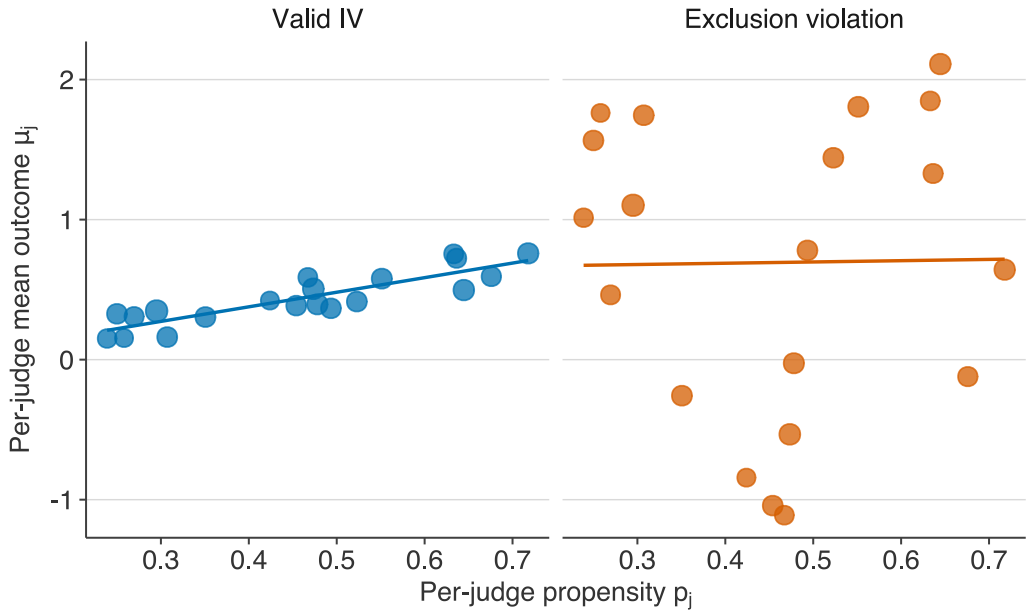

Figure 4: Per-judge outcome means versus per-judge treatment propensities, K = 20 judges and N = 3000. Left: valid-IV null ($Y = D + \varepsilon$). Right: exclusion violation with direct judge effect $1.5\sin(0.5J)$. Point size is proportional to judge-cell sample size; lines are weighted-LS fits. The violation design produces a systematic departure from linearity that `iv_testjfe` detects at $p < 10^{-4}$.

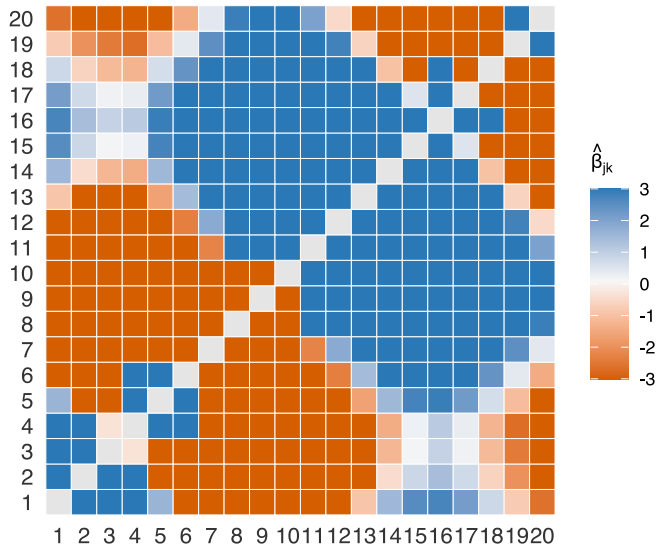

Figure 5: Pairwise Wald LATE matrix $(\mu_j - \mu_k)/(p_j - p_k)$ under the violation design from Figure 4. Entries are colour-shaded around the fitted slope; deviations in either direction signal pairs whose Wald estimates are inconsistent with a common complier LATE. The matrix is returned as `pairwise_late` on the `iv_test` object; the worst-offender pair is also returned as `worst_pair` for direct inspection.

9 Limitations

Seven limitations apply.

1. **ivcheck** is a testing package, not an estimator. Use **fixest** or **ivreg** for the IV estimation itself; **ivcheck** operates on the fitted model or on raw vectors.
2. Version 0.1.0 supports discrete instruments only. Continuous `Z` must be discretised into bins (quartiles or quintiles) before passing to `iv_kitagawa` or `iv_mw`, at modest cost in power. A nonparametric continuous-Z extension is on the v0.2.0 roadmap.
3. Fuzzy regression discontinuity has a dedicated test (Arai et al., 2022) that is not yet in **ivcheck**. Its different infrastructure (running variable, cutoff, local-polynomial bandwidth, bias correction) does not fit the current spine of S3 dispatch on fitted IV models, so a dedicated `iv_frd()` function is deferred to v0.2.0.
4. `iv_testjfe` implements the asymptotic chi-squared linearity test that corresponds to the Frandsen et al. (2023) statistic under a linear outcome-propensity specification. The full flexible-series test with the bounded-slope moment-inequality restriction is planned for v0.2.0. Users needing the published test with the flexible basis should run Frandsen's Stata `testjfe` module in the interim.
5. Fixed-effects IV models are not supported in v0.1.0. Calling `iv_kitagawa`, `iv_mw`, or `iv_testjfe` on a `fixest` model with the `| FE |` component aborts with a clear error. The discrete-Z tests operate on the raw `(Y, D, Z)` joint distribution, and residualising on fixed effects destroys the discrete structure of `Z`. The documented workaround is to pre-demean `Y` and `D` within each FE cell, keep `Z` discrete, and call the default method on the resulting vectors. A stratified-by-FE variant is on the v0.2.0 roadmap.
6. The conditional path of `iv_mw` supports a single covariate. Users passing a matrix of covariates receive a clear error pointing to the tensor-product-basis extension planned for v0.2.0. The current recommendation is to condition on the covariate most plausibly driving heterogeneity in compliance.
7. Rejection of any of the three tests is evidence of violation but does not localise the failed assumption. Non-rejection is evidence of no detectable violation at level `alpha`, not proof of validity. Report non-rejection as such.

9.1 Notes on fidelity

For transparency, the ordered multivalued `D` path in `iv_kitagawa` tests a richer family of inequalities than Sun (2023) equation 10: we sweep over cumulative-tail events `P(Y <= y, D <= l | Z)` and `P(Y <= y, D >= l | Z)` for every intermediate `l` and every `Z` pair, while Sun's derivation uses only `d_min` and `d_max` across adjacent pairs. All inequalities tested hold under Sun's Assumption 2.2,

so the test controls size; the expanded family gives a more exhaustive test rather than the minimal-implication version of the paper. The unordered case (Sun section 3.3) is available via `treatment_order = "unordered"` with a user-specified `monotonicity_set` encoding the direction of the monotonicity restriction.

`iv_testjfe` implements the linearity-test form of the Frandsen et al. (2023) overidentification test: we fit `mu_j` against `p_j` by WLS and compare the weighted residual SS to a `chi^2_{K-2}` reference. FLL's published test uses a flexible basis plus bounded-slope moment-inequality constraints on the LATE; our v0.1.0 form is a simpler necessary-condition check. Users wanting the exact published test should run Frandsen's Stata `testjfe` module; a flexible-basis extension and CLR-style bounded-slope inference are planned for v0.2.0.

The `iv_mw` CLR path is an independent implementation of the Chernozhukov et al. (2013) intersection-bounds framework (series-regression conditional CDF, robust plug-in SE, Andrews and Soares (2010) adaptive moment selection) rather than a port of Mourifie-Wan's Matlab code. Cross-validation against the authors' code is on the v0.2.0 wishlist. The bootstrap multiplier defaults to Rademacher but `multiplier = "gaussian"` and `multiplier = "mammen"` are available for robustness checks.

10 Conclusion

ivcheck closes a tooling gap between the econometric literature on LATE falsification and applied IV practice. The package is available on CRAN; source and issue tracker at `https://github.com/charlescoverdale/ivcheck`.

Acknowledgements

I thank Toru Kitagawa, Ismael Mourifie, Yuanyuan Wan, Brigham Frandsen, Lars Lefgren, and Emily Leslie for their methodological contributions and the publicly-available replication archives that made this port possible. The Card (1995) dataset is distributed via the **wooldridge** package (Wooldridge, 2020) on CRAN, sourced from the National Longitudinal Survey of Young Men.

Bibliography

D. W. K. Andrews and G. Soares. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157, 2010. doi: 10.3982/ECTA7502. [p1, 4, 9]

Y. Arai, Y.-C. Hsu, T. Kitagawa, I. Mourifié, and Y. Wan. Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics*, 13(1):1–28, 2022. doi: 10.3982/QE1367. [p8]

D. Card. Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, and R. Swidinsky, editors, *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, pages 201–222. University of Toronto Press, 1995. [p3, 5]

V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737, 2013. doi: 10.3982/ECTA8718. [p1, 2, 9]

B. R. Frandsen, L. J. Lefgren, and E. C. Leslie. Judging judge fixed effects. *American Economic Review*, 113(1):253–277, 2023. doi: 10.1257/aer.20201860. [p1, 2, 4, 8, 9]

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. doi: 10.2307/2951620. [p1]

T. Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015. doi: 10.3982/ECTA11974. [p1, 2]

A. Lal, M. Lockhart, Y. Xu, and Z. Zu. How much should we trust instrumental variable estimates in political science? practical advice based on 67 replicated studies. *Political Analysis*, 2024. doi: 10.1017/pan.2024.2. [p1]

I. Mourifié and Y. Wan. Testing local average treatment effect assumptions. *Review of Economics and Statistics*, 99(2):305–313, 2017. doi: 10.1162/REST_a_00622. [p1]

Z. Sun. Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, 2023. doi: 10.1016/j.jeconom.2023.105628. [p4, 8]

J. M. Wooldridge. *wooldridge: 115 Data Sets from "Introductory Econometrics: A Modern Approach"*, 2020. URL https://CRAN.R-project.org/package=wooldridge. R package version 1.4-3. [p9]