Benchmarking the Unobserved Coverage Failure in Omitted-Variable Sensitivity Bounds

Created by: Viktor Popdonchev
Supervisors: Matej Havelka, Jesse Krijthe

TU Delft

1. Motivation

Causal inference requires an untestable assumption: all confounders are measured. Sensitivity analysis bounds potential hidden bias but requires a confounder strength, often supplied by benchmarking against an observed covariate. This assumes the covariate is a valid reference. We ask: if this assumption fails, does the bound still hold, and do standard statistics warn the analyst?

2. Background

The Cinelli-Hazlett bound describes a hidden confounder with two numbers: how much residual treatment variance and how much residual outcome variance it would explain. The practitioner cannot observe the confounder, so the bound benchmarks it against an observed covariate. This bundles two distinct claims:

1. The hidden confounder is no stronger than the observed covariate
2. The chosen covariate is an appropriate structural reference for the hidden confounder.

Our research isolates what happens to the mathematical bounds when the second condition, the benchmarking assumption, fails.

3. Research Questions

- At what point does the omitted-variable bound collapse when the benchmarking assumption is violated?
- Is the failure point a property of the bound itself, or an artefact of which covariates were measured?
- Is the benchmarking violation what breaks the bound?
- Do the standard reported statistics warn of the failure, or does the bound fail silently?

4. Method

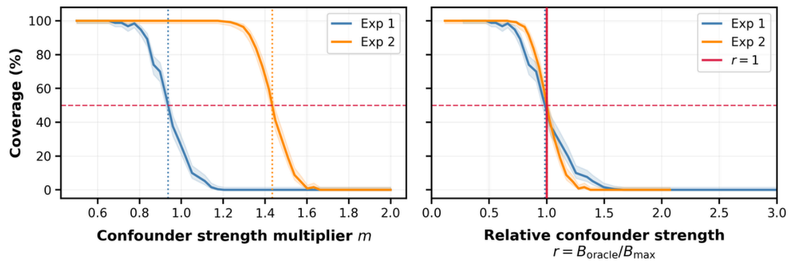
Because unobserved confounding cannot be evaluated in real-world datasets, we rely on a Monte Carlo simulation (partially linear data-generating process, binary treatment) where the true hidden confounder is entirely controlled. We experiment with

- Confounder Strength
- Resemblance (how the covariates proxy the confounder)
- Effect Alignment.

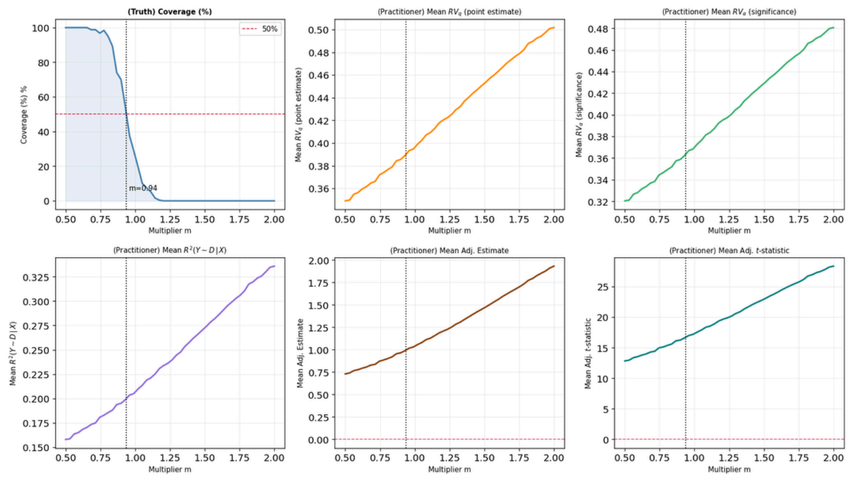
For each trial we measure two things:

1. Does the LOO bound cover the real bias?
2. What do the practitioner-used diagnostic metrics show?

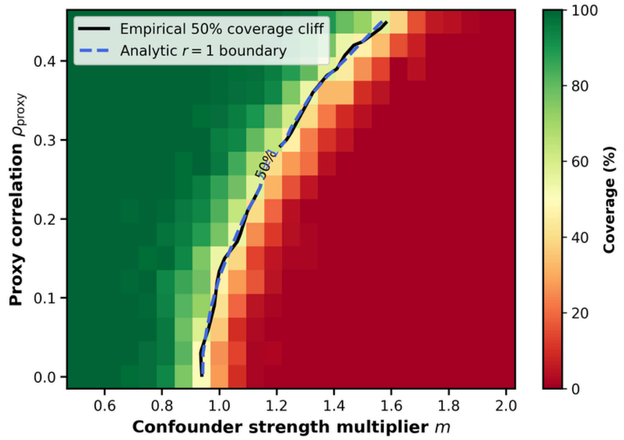
5. Experiments



Does the failure point depend on which covariates were measured? Left: against raw strength m, the two covariate sets fail at different points. Right: against relative strength r = B_oracle / B_max, both fail at r ≈ 1. The gap is an artefact of the axis



Do the reported statistics warn of the failure? The five statistics across the sweep, failure point marked. Every one moves toward greater apparent robustness as coverage collapses. None signals it.



Is a benchmarking violation what breaks the bound? Coverage over confounder strength m and proxy quality ρ_proxy. The bound fails (red) only in the corner where both levers break. The r = 1 contour traces the failure line.

6. Results

1. The bound holds until the confounder reaches the benchmark's strength, then fails over a narrow range, not gradually.
2. The failure sits at a single threshold, r ≈ 1, across structurally different covariate sets: the point where the confounder overtakes the benchmark.
3. None of the five reported statistics warns. Every one moves toward greater apparent robustness as coverage collapses.
4. Two levers set the failure: the confounder's strength and how well the covariates proxy it. The alignment of the confounder's two effects sets how far past r = 1 the bound survives.

7. Limitations

- Results establish a best-case baseline.
- Covariates here are independent, the most favourable case for benchmarking; real ones are correlated, and a confounder is usually proxied by several at once.
- Design assumes a partially linear process, OLS, one confounder.
- Not shown to hold for nonparametric estimation, panel or time-series data, or IV designs.

8. Conclusion

The sensemakr template is reliable when the benchmarking assumption holds and silently misleading when it does not. The default k = 1 sits exactly at r = 1, the edge of failure, so a confounder even slightly stronger than assumed removes the reported robustness with no sign that it has. We recommend two additions to the standard report: a subject-matter defence of which covariate the confounder is expected to resemble, and adjusted estimates across a range of multipliers, for example k in {1, 2, 3, 5}, rather than the default alone.

References

Altonji, Elder & Taber (2005). JPE. — Angrist & Pischke (2009). Mostly Harmless Econometrics. — Baitairian et al. (2025). HAL preprint. — Chernozhukov et al. (2018). Econometrics Journal. — Chernozhukov et al. (2024). arXiv. — Cinelli & Hazlett (2020). JRSS-B. — Cornfield et al. (1959). JNCI. — Frank (2000). SMR. — Imbens (2003). AER. — Imbens & Rubin (2015). Cambridge UP. — Oster (2019). JBES. — Pearl (2009). Causality. — Rosenbaum (2002). Observational Studies. — Rosenbaum & Rubin (1983). Biometrika. — Tan (2006). JASA. — VanderWeele & Ding (2017). Annals of Internal Medicine.