`tesensitivity`: A Stata package for assessing the unconfoundedness assumption

Matthew A. Masten
Duke University

Alexandre Poirier
Georgetown University

Linqi Zhang
Boston College

Stata Conference Chicago
July 12, 2019

References

Based on two papers:

Masten and Poirier (2018) "Identification of Treatment Effects under Conditional Partial Independence," *Econometrica*

- Gives the identification theory

Masten, Poirier, and Zhang (2019) "Assessing Sensitivity to Unconfoundedness: Estimation and Inference," Working paper

- Gives the estimation and inference theory

The standard treatment effects model

$X \in \{0, 1\}$ is a binary treatment

(Y_1, Y_0) are unobserved potential outcomes. We observe

$$Y = XY_1 + (1 - X)Y_0$$

along with X and a vector of covariates W

Goal: Identify parameters like

$$\text{ATE} = \mathbb{E}(Y_1 - Y_0) \qquad \text{and} \qquad \text{QTE}(\tau) = Q_{Y_1}(\tau) - Q_{Y_0}(\tau)$$

The standard treatment effects model

Baseline assumptions:

1. Unconfoundedness:

$$Y_1 \perp\!\!\!\perp X \mid W \qquad \text{and} \qquad Y_0 \perp\!\!\!\perp X \mid W$$

2. Overlap:

$$0 < \mathbb{P}(X = 1 \mid W = w) < 1$$

   for all $w \in \text{supp}(W)$

Under these assumptions, ATE and QTE(τ) are point identified

The standard treatment effects model

Baseline assumptions:

1. Unconfoundedness:

$$Y_1 \perp\!\!\!\perp X \mid W \qquad \text{and} \qquad Y_0 \perp\!\!\!\perp X \mid W$$

2. Overlap:

$$0 < \mathbb{P}(X = 1 \mid W = w) < 1$$

   for all $w \in \text{supp}(W)$

Under these assumptions, ATE and QTE(τ) are point identified

Thus just go to the data and compute your treatment effects

Huge literature on how to do this: `teffects`

The standard treatment effects model

Problem: Our treatment effect estimates are only as good as the assumptions behind them...

...so what if our assumptions don't hold?

The standard treatment effects model

Problem: Our treatment effect estimates are only as good as the assumptions behind them...

...so what if our assumptions don't hold?

Overlap: This assumption is solely about X and W. Hence it's refutable

- Many ways to check this in finite samples, and it's commonly done (`teffects overlap`)

The standard treatment effects model

Problem: Our treatment effect estimates are only as good as the assumptions behind them...

...so what if our assumptions don't hold?

Overlap: This assumption is solely about X and W. Hence it's refutable

- Many ways to check this in finite samples, and it's commonly done (`teffects overlap`)

But what about unconfoundedness?

- Unlike overlap, it's *not* refutable—It's an assumption on unobservables

  ⇒ Much less clear how to "assess" this assumption

Assessing unconfoundedness

Lots of approaches, including Rosenbaum and Rubin (1983), Mauro (1990), Robins, Rotnitzky, and Scharfstein (2000), Imbens (2003), Altonji, Elder, and Taber (2005, 2008), Hosman, Hansen, and Holland (2010), Krauth (2016), Oster (2019), among others

These approaches rely on strong auxiliary assumptions, like

- Potential outcome functions which are linear in all variables
- Homogeneous treatment effects

Arguably goes against the spirit of sensitivity analysis

Assessing unconfoundedness

Nonparametric options in the literature:

1. Ichino, Mealli, and Nannicini (2008)
   - Requires all variables to be discrete
   - Uses lots of sensitivity parameters
   - `sensatt`, discussed in Nannicini (2008) "A simulation-based sensitivity analysis for matching estimators," *The Stata Journal*
2. Rosenbaum (1995, 2002) and subsequent work
   - Uses randomization inference
   - `mhbounds`, discussed in Becker and Caliendo (2007) "Sensitivity analysis for average treatment effects," *The Stata Journal*
3. Our approach:
   - Large population version of Rosenbaum's approach
   - Allows us to split the identification analysis from the estimation and inference theory (don't have to commit to a specific testing procedure)

Relaxing unconfoundedness

Unconfoundedness says $Y_1 \perp\!\!\!\perp X \mid W$. That is,

$$\mathbb{P}(X = 1 \mid Y_1 = y_1, W = w) = \mathbb{P}(X = 1 \mid W = w)$$

for all w. Likewise for Y_0

Relaxing unconfoundedness

Unconfoundedness says $Y_1 \perp\!\!\!\perp X \mid W$. That is,

$$\mathbb{P}(X = 1 \mid Y_1 = y_1, W = w) - \mathbb{P}(X = 1 \mid W = w) = 0$$

for all w. Likewise for Y_0

Relaxing unconfoundedness

Unconfoundedness says $Y_1 \perp\!\!\!\perp X \mid W$. That is,

$$\mathbb{P}(X = 1 \mid Y_1 = y_1, W = w) - \mathbb{P}(X = 1 \mid W = w) = 0$$

for all w. Likewise for Y_0

We relax it by supposing

$$\left| \mathbb{P}(X = 1 \mid Y_1 = y_1, W = w) - \mathbb{P}(X = 1 \mid W = w) \right| \leq c$$

for all w, for some known $c \in [0, 1]$. Likewise for Y_0

We call this conditional c-dependence

Identification

In the papers, we derive sharp bounds on ATE, ATT, QTEs, and other parameters

We provide sample analog estimators, estimation theory, and inference theory

Estimation

The bounds all depend on two objects:

1. The quantile regression $Q_{Y|X,W}(q \mid x, w)$
2. The propensity score $\mathbb{P}(X = 1 \mid W = w)$

You can use anything you'd like to estimate these

We start with probably the simplest approach:

1. Linear quantile regression of Y on $(1, X, W)$
2. Logistic regression of X on $(1, W)$

Empirical illustration

We use the classic National Supported Work (NSW) demonstration dataset (MDRC 1983), as analyzed by LaLonde (1986) and reconstructed Dehejia and Wahba (1999)

Used by other sensitivity analysis papers—allows for direct comparison

In particular, we will compare our nonparametric results with the parametric ones obtained in Imbens (2003)

Empirical illustration

The NSW experiment randomly assigned participants to either...

- (treatment) receive a guaranteed job for 9 to 18 months along with frequent counselor meetings or
- (control) be left in the labor market by themselves

Outcome of interest is earnings in 1978

Empirical illustration

We use two subsamples:

1. Experimental data: The Dehejia and Wahba (1999) subsample of all males in LaLonde's NSW data where earnings are observed in 1974, 1975, 1978
   - 445 people: 185 treated, 260 control
2. Observational data: The 185 NSW treatment group combined with 2490 people in a control group constructed from the PSID, and then dropping anyone with earnings above $5,000
   - 390 people: 148 treated, 242 control

These two subsamples were considered by Imbens (2003)

Empirical illustration: Baseline results

Table: Baseline treatment effect estimates (in 1978 dollars).

	ATE	ATT	Sample size
Experimental dataset	1633	1738	445
	(650)	(689)	
Observational dataset	3337	4001	390
	(769)	(762)	

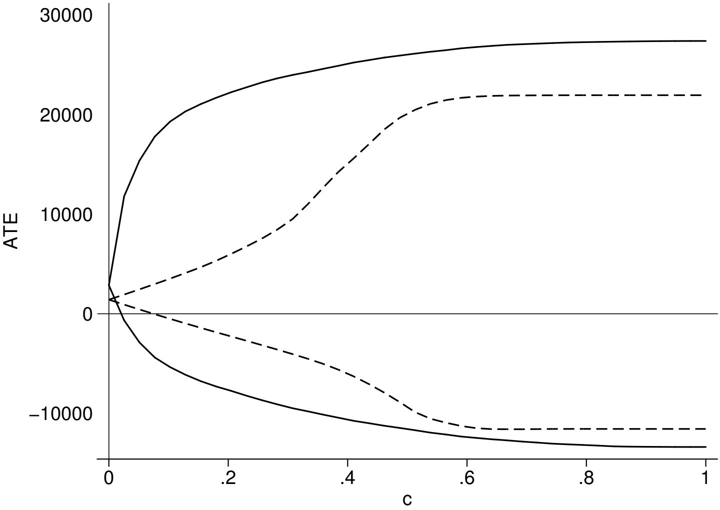
Standard errors in parentheses.

```
teffects ipw (‘Y’) (‘X’ ‘W’)
teffects ipw (‘Y’) (‘X’ ‘W’), atet
```

Empirical illustration: Sensitivity analysis

```
tesensitivity ‘Y’ ‘X’ ‘W’, ate atet breakdown
```

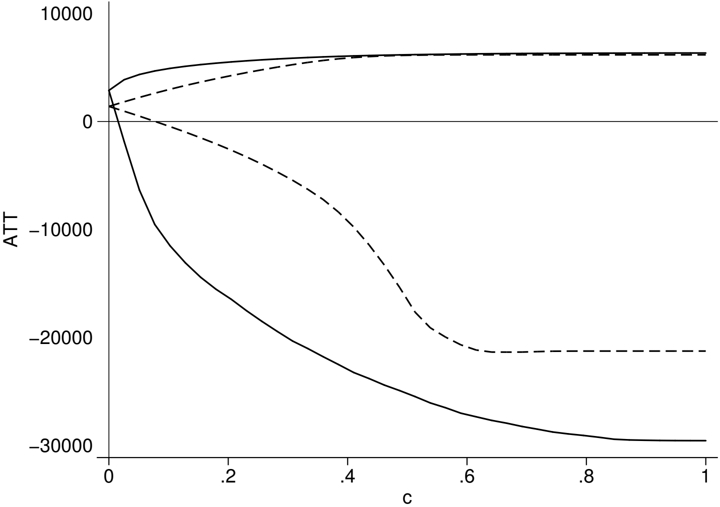
Empirical illustration: Bounds on ATE



Estimated breakdown points: 0.075 (experimental) 0.02 (observational)

```
tesensitivity 'Y' 'X' 'W', ate atet breakdown
```

Empirical illustration: Bounds on ATT



Estimated breakdown points: 0.08 (experimental) 0.01 (observational)

```
tesensitivity ‘Y’ ‘X’ ‘W’, ate atet breakdown
```

Calibrating c

How to determine what values of c are ‘large’ and which are ‘small’?

This is a key question for any sensitivity analysis—and it’s very difficult!

Two approaches:

1. Relative comparisons: Compare bounds *across* datasets or studies
2. Absolute comparison: Calibrate c within a single dataset

Calibrating c

To do an absolute comparison, we use a classic idea (Cornfield et al 1959, Imbens 2003, Altonji, Elder, and Taber 2005, 2008, Oster 2019):

> Use selection on observables to calibrate our beliefs about selection on unobservables

Important caveat: We only provide a rule of thumb

- Not (yet) theoretically justified!
- Lots of research left to do before we have a fully satisfactory approach

Calibrating c

Say $W = (W_1, W_2)$. Define

$$\overline{c}_1 = \sup_{w_2} \sup_{w_1} \left| \mathbb{P}(X = 1 \mid W_1 = w_1, W_2 = w_2) - \mathbb{P}(X = 1 \mid W_2 = w_2) \right|$$

This is a measure of the impact on the propensity score of adding W_1 given that we already included W_2

Calibrating c

Say $W = (W_1, W_2)$. Define

$$\overline{c}_1 = \sup_{w_2} \sup_{w_1} \left| \mathbb{P}(X = 1 \mid W_1 = w_1, W_2 = w_2) - \mathbb{P}(X = 1 \mid W_2 = w_2) \right|$$

This is a measure of the impact on the propensity score of adding W_1 given that we already included W_2

Can do the same, but swapping roles of W_1 and W_2; yields $\overline{c}_2$

Calibrating c

Say $W = (W_1, W_2)$. Define

$$\overline{c}_1 = \sup_{w_2} \sup_{w_1} \left| \mathbb{P}(X = 1 \mid W_1 = w_1, W_2 = w_2) - \mathbb{P}(X = 1 \mid W_2 = w_2) \right|$$

This is a measure of the impact on the propensity score of adding W_1 given that we already included W_2

Can do the same, but swapping roles of W_1 and W_2; yields $\overline{c}_2$

Idea: c-dependence is the same thing, except we're adding the unobservable Y_1 given that we already included W

Calibrating c

Might expect the impact of adding Y_1 in addition to W is smaller than $\overline{c}_1$ and $\overline{c}_2$, so can also look at the distribution of

$$|\mathbb{P}(X = 1 \mid W_1, W_2) - \mathbb{P}(X = 1 \mid W_2)|$$

For example, the 50th, 75th, and 90th quantiles

Empirical illustration: Calibrating c

```
tesensitivity ‘Y’ ‘X’ ‘W’, ate atet breakdown ckvector ckdensity
```

Empirical illustration: Calibrating c

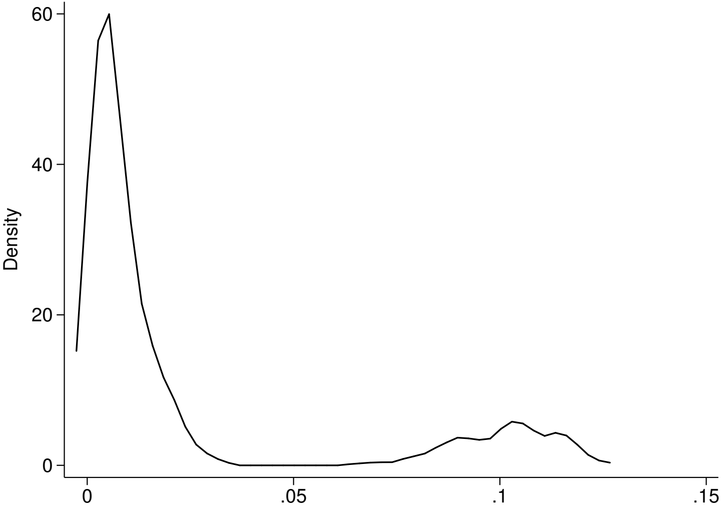
Variation in $|p_{1|W}(W_{-k}, W_k) - p_{1|W_{-k}}(W_{-k})|$ (experimental data)

	p50	p75	p90	$\bar{c}_k$
Earnings in 1975	0.001	0.004	0.008	0.053
Black	0.007	0.009	0.014	0.082
Positive earnings in 1974	0.002	0.010	0.018	0.034
Education	0.012	0.022	0.031	0.087
Married	0.006	0.012	0.032	0.042
Age	0.015	0.024	0.034	0.099
Earnings in 1974	0.002	0.011	0.035	0.209
Positive earnings in 1975	0.013	0.017	0.062	0.082
Hispanic	0.007	0.017	0.099	0.124

Estimated breakdown point: 0.075

Empirical illustration: Calibrating c

Kernel density estimate of $|p_{1|W}(W_{-k}, W_k) - p_{1|W_{-k}}(W_{-k})|$ for $k =$ hispanic indicator (experimental data)



Empirical illustration: Calibrating c

Variation in $|p_{1|W}(W_{-k}, W_k) - p_{1|W_{-k}}(W_{-k})|$ (observational data)

	p50	p75	p90	$\bar{c}_k$
Earnings in 1974	0.000	0.001	0.009	0.065
Hispanic	0.003	0.011	0.024	0.214
Education	0.006	0.017	0.042	0.127
Earnings in 1975	0.002	0.010	0.057	0.276
Positive earnings in 1975	0.007	0.019	0.076	0.295
Positive earnings in 1974	0.012	0.028	0.099	0.423
Married	0.028	0.079	0.172	0.314
Age	0.035	0.093	0.205	0.508
Black	0.053	0.143	0.266	0.477

Estimated breakdown point: 0.02

Empirical illustration: Overall findings

Relative comparisons:

- The experimental dataset is relatively less sensitive to relaxations of unconfoundedness than the observational dataset
    - For most c's, the observational bounds are wider than the experimental bounds, often substantially wider

Absolute comparisons:

- For the experimental dataset, most variation in leave-out-variable-k propensity scores is smaller than the ATE and ATT breakdown points.
- But not for the observational dataset

Empirical illustration: Takeaways

Imbens (2003) found that this observational dataset was relatively robust

Our conclusion differs because our bounds do not impose the strong parametric assumptions he made; in particular,

- homogeneous treatment effects
- normally distributed outcomes
- all violations occur solely through a single binary confounder

Ironic that many methods for sensitivity analyses themselves rely on strong auxiliary assumptions

The conclusions of the sensitivity analysis may themselves be sensitive to changing these auxiliary assumptions, as we see here

$\Rightarrow$ Use nonparametric methods for sensitivity analysis!

Conclusion

Estimates from `teffects` rely on two assumptions:

1. Unconfoundedness
2. Overlap

Overlap is easier to assess, but unconfoundedness is important too!

`tesensitivity` is a tool for assessing unconfoundedness which does not require strong auxiliary assumptions

- Package will be online in the next few months
- We are *very* interested in feedback from practitioners, so please email us if you have questions or problems, or use our (future) github issues page!