Discussion:

Assumption Smuggling in Tests of Causal Mechanisms
&
Evaluating the Validity of the Exclusion Restriction in Instrumental Variables Designs Using Automated Partial Identification

Polmeth 2024 Summer Meeting

July 18, 2024

Throat clearing

- Thanks
- Disclaimer: Confusion, misunderstandings, and bad advice are likely
- Hope: My misunderstandings suggest misreadings others may have

Paper 1:

Assumption Smuggling in Tests of Causal Mechanisms, by
Matthew Blackwell, Ruofan Ma, and Aleksei Opacics

Overall

I think this is great and you have a very thorough treatment here. Useful to read, and to teach from.

Overall

I think this is great and you have a very thorough treatment here. Useful to read, and to teach from.

Mainly I want to talk about framing:

- Just helping the reader organize the main points
- The connection between what errors you are pointing out and what people are doing in practice, or should be doing
- Relatedly: are there evidentiary value of IOTs that may persist despite “the indirect effect estimate includes zero”

Organizing main points

As a first time reader I found it took some time and pages to pull out the main things you were concerned with.

Organizing main points

As a first time reader I found it took some time and pages to pull out the main things you were concerned with.

One possibility: Organize around the several reasons why the IOT falls short of providing the information we would hope for. Enumerate them, e.g.

Organizing main points

As a first time reader I found it took some time and pages to pull out the main things you were concerned with.

One possibility: Organize around the several reasons why the IOT falls short of providing the information we would hope for. Enumerate them, e.g.

- If you get a detectable (average) effect of A on M, you still know nothing about effect of M on Y, so you know nothing about indirect effect through M

Organizing main points

As a first time reader I found it took some time and pages to pull out the main things you were concerned with.

One possibility: Organize around the several reasons why the IOT falls short of providing the information we would hope for. Enumerate them, e.g.

- If you get a detectable (average) effect of A on M, you still know nothing about effect of M on Y, so you know nothing about indirect effect through M
- If you come up null on (average) effect of A on M, then you can't know there aren't individual effects in different directions.

Organizing main points

As a first time reader I found it took some time and pages to pull out the main things you were concerned with.

One possibility: Organize around the several reasons why the IOT falls short of providing the information we would hope for. Enumerate them, e.g.

- If you get a detectable (average) effect of A on M, you still know nothing about effect of M on Y, so you know nothing about indirect effect through M
- If you come up null on (average) effect of A on M, then you can't know there aren't individual effects in different directions.
- Indirect effect is not product of coefficients (though I'm not sure we need that here in context of what can go wrong with IOT-based claims)

How are IOTs getting used, and is it wrong?

I was uncertain about what practices you are arguing against, and how common they are, or how often IOTs are correctly used and stated.

How are IOTs getting used, and is it wrong?

I was uncertain about what practices you are arguing against, and how common they are, or how often IOTs are correctly used and stated.

In short, you need to smuggle a lot of assumptions to estimate the indirect effect, or to argue it is non-zero—*but are people really relying on that or getting something else out of these*?

Are there different, valid, aims that are fulfilled?

Put another way, you ask "what is the point of mechanism tests", noting

> *"there are many instances where all approaches to establishing indirect effects, including IOTs, rely on shaky assumptions that are difficult to justify".*

Are there different, valid, aims that are fulfilled?

Put another way, you ask “what is the point of mechanism tests”, noting

> *“there are many instances where all approaches to establishing indirect effects, including IOTs, rely on shaky assumptions that are difficult to justify”.*

On the other hand, “inability to rule out zero indirect effect” is not the same as “IOTs are uninformative” (see next)

Are there different, valid, aims that are fulfilled?

Put another way, you ask "what is the point of mechanism tests", noting

> *"there are many instances where all approaches to establishing indirect effects, including IOTs, rely on shaky assumptions that are difficult to justify".*

On the other hand, "inability to rule out zero indirect effect" is not the same as "IOTs are uninformative" (see next)

I had trouble figuring out whether you were objecting to the remaining evidentiary value, or suggesting that people don't understand these limitations, or something else.

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Still, IOT may be useful:

- If there is no apparent (average) effect of A on M, then either

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Still, IOT may be useful:

- If there is no apparent (average) effect of A on M, then either
  - there really is little effect of A on M and it doesn't look good for the mediation story, or

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Still, IOT may be useful:

- If there is no apparent (average) effect of A on M, then either
  - there really is little effect of A on M and it doesn't look good for the mediation story, or
  - the effects of A on M are exist but are so heterogeneous you can't see it, at least ruling out the simple monotonic $A \rightarrow M \rightarrow Y$ that may have been posed.

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Still, IOT may be useful:

- If there is no apparent (average) effect of A on M, then either
    - there really is little effect of A on M and it doesn't look good for the mediation story, or
    - the effects of A on M are exist but are so heterogeneous you can't see it, at least ruling out the simple monotonic $A \rightarrow M \rightarrow Y$ that may have been posed.
- If you do see an average effect of A on M then there:
    - M might still not be a mediator (we don't know $M \rightarrow Y$
    - But the simple $A \rightarrow M \rightarrow Y$ story is still in the game, worth continuing to investigate by other means.

Evidentiary value of IOT?

Suppose you have an indirect effect interval estimate that includes zero, as you often will.

Still, IOT may be useful:

- If there is no apparent (average) effect of A on M, then either
  - there really is little effect of A on M and it doesn't look good for the mediation story, or
  - the effects of A on M are exist but are so heterogeneous you can't see it, at least ruling out the simple monotonic $A \to M \to Y$ that may have been posed.
- If you do see an average effect of A on M then there:
  - M might still not be a mediator (we don't know $M \to Y$
  - But the simple $A \to M \to Y$ story is still in the game, worth continuing to investigate by other means.

Is this kind of "plausibility probe" approach,

- still scientifically relevant?
- something some or many investigators correctly understand and state?

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \rightarrow M$), you know blood pressure *might* be on the path to reduced heart disease.

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \to M$), you know blood pressure *might* be on the path to reduced heart disease.
    - you don't know if blood pressure → heart disease is there, but

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \rightarrow M$), you know blood pressure *might* be on the path to reduced heart disease.
    - you don't know if blood pressure → heart disease is there, but
    - the simple mediation story lives to fight on, suggesting further investigation, like randomizing blood pressure pharma in a separate study.

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \to M$), you know blood pressure *might* be on the path to reduced heart disease.
  - you don't know if blood pressure → heart disease is there, but
  - the simple mediation story lives to fight on, suggesting further investigation, like randomizing blood pressure pharma in a separate study.
- If instead you see no (averge) effect of exercise on blood pressure,

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \to M$), you know blood pressure *might* be on the path to reduced heart disease.
    - you don't know if blood pressure → heart disease is there, but
    - the simple mediation story lives to fight on, suggesting further investigation, like randomizing blood pressure pharma in a separate study.
- If instead you see no (averge) effect of exercise on blood pressure,
    - At best, "its complicated"; you have to give up the mechanism, or entertain differently signed individual effects

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \rightarrow M$), you know blood pressure *might* be on the path to reduced heart disease.
  - you don't know if blood pressure → heart disease is there, but
  - the simple mediation story lives to fight on, suggesting further investigation, like randomizing blood pressure pharma in a separate study.
- If instead you see no (averge) effect of exercise on blood pressure,
  - At best, "its complicated"; you have to give up the mechanism, or entertain differently signed individual effects
  - the case for a *beneficial, average* effect flowing through blood pressure remains possible but requires certain arrangements of the signs of the component effects.

Example

You know exercise reduces risk of cardiovascular events, but you want to know if this is partly through blood pressure.

- In an exercise experiment, if you see an (average) effect of exercise on blood pressure ($A \rightarrow M$), you know blood pressure *might* be on the path to reduced heart disease.
    - you don't know if blood pressure → heart disease is there, but
    - the simple mediation story lives to fight on, suggesting further investigation, like randomizing blood pressure pharma in a separate study.
- If instead you see no (averge) effect of exercise on blood pressure,
    - At best, "its complicated"; you have to give up the mechanism, or entertain differently signed individual effects
    - the case for a *beneficial, average* effect flowing through blood pressure remains possible but requires certain arrangements of the signs of the component effects.
    - may suggests a within-person design or in-lab manipulation to look at different directions of the exercise-blood pressure effect.

You note something similar:

Citing Green, Ha, & Bullock:

> *[T]his kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation.*

You note something similar:

Citing Green, Ha, & Bullock:

> *[T]his kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation.*

But you are concerned:

> *"We have not found researchers using IOTs to either acknowledge these assumptions or be as cautious in the interpretation of IOTs as Green, Ha and Bullock (2010) recommend."*

You note something similar:

Citing Green, Ha, & Bullock:

> *[T]his kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation.*

But you are concerned:

> *"We have not found researchers using IOTs to either acknowledge these assumptions or be as cautious in the interpretation of IOTs as Green, Ha and Bullock (2010) recommend."*

But again I wonder, if the question is "can you get some evidence from IOTs" rather than "cane you rule out indirect effect of zero", are so many authors really getting it wrong?

You note something similar:

Citing Green, Ha, & Bullock:

> *[T]his kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation.*

But you are concerned:

> *"We have not found researchers using IOTs to either acknowledge these assumptions or be as cautious in the interpretation of IOTs as Green, Ha and Bullock (2010) recommend."*

But again I wonder, if the question is "can you get some evidence from IOTs" rather than "cane you rule out indirect effect of zero", are so many authors really getting it wrong?

Maybe the upshot is: your analysis gives us the opportunity to help encourage careful interpretation in light of the assumptions involved, and aids in thinking up the "further experimental investigations" you need.

Paper 2:

Evaluating the Validity and Robustness of Instrumental-Variable Analyses, by Cooper, Duarte, Keele, Knox, Mattes, Mummolo

Overall

Really interesting, and I've been looking forward to seeing work on the instrument inequalities

Overall

Really interesting, and I've been looking forward to seeing work on the instrument inequalities

That alone is sufficient for important, albeit theoretical publication.

Overall

Really interesting, and I've been looking forward to seeing work on the instrument inequalities

That alone is sufficient for important, albeit theoretical publication.

The next question for me is about practicality: will people understand it, can it improve IV practice?

Overall

Really interesting, and I've been looking forward to seeing work on the instrument inequalities

That alone is sufficient for important, albeit theoretical publication.

The next question for me is about practicality: will people understand it, can it improve IV practice?

I'll comment first on the first part of the paper, i.e. the instrumental inequalities.

First, on falsification

- I think the paper gets the notion of "falsification tests" completely right
- But I think you'll save yourself a lot of headaches if very early and clearly in the paper you describe what you mean by falsification and it's contrast to validation.
- There were some places where I fear readers might misunderstand at first, e.g.

  > *"[these tests] are known to have indirect observable implications that allow for necessary tests of validity"*

Second, fighting for intuition?

The non-intuitiveness of these inequalities, as you acknowledge, is annoying!

Second, fighting for intuition?

The non-intuitiveness of these inequalities, as you acknowledge, is annoying!

To spitball irresponsibly,

- Are these some tables where you can play joint probabilities of Z, D, and Y and gain some experience with how things can fail to add up without violating assumptions?
- I've seen some efforts (on YouTube!) to give intuition to Bell's theorem…I can't say they were successful but I wonder if you'd have any luck adapting these.

Power?

As falsification tests, these inequalities are helpful only in proportion to $Pr(alarm|notvalid)$, which you might call the “power” of the test.

Power?

As falsification tests, these inequalities are helpful only in proportion to $Pr(alarm|notvalid)$, which you might call the "power" of the test.

- I don't suppose there are any insights you can offer about this probability? And does it depend on sample size?

Power?

As falsification tests, these inequalities are helpful only in proportion to $Pr(alarm|notvalid)$, which you might call the "power" of the test.

- I don't suppose there are any insights you can offer about this probability? And does it depend on sample size?
- I wonder if some simulation in the first part of the paper can be demonstrative: vary the severity of a violation in some sense we can understand, then see how often it trips the falsification alarm.

Power?

As falsification tests, these inequalities are helpful only in proportion to $Pr(alarm|notvalid)$, which you might call the “power” of the test.

- I don’t suppose there are any insights you can offer about this probability? And does it depend on sample size?
- I wonder if some simulation in the first part of the paper can be demonstrative: vary the severity of a violation in some sense we can understand, then see how often it trips the falsification alarm.
- Worth checking loads of IV studies and see how often theres evidence of a violation? (I see you have a rejection on monotonicity in one of your applications – curious though if exclusion is tougher.)
    - You could subset to those where there is really good reason to think exclusion and/or monotonicity are violated and see if falsification rate is higher in these.

These tests versus others?

Should we understand the instrumental inequality tests to differe in status in any way from conventional falsification tests?

These tests versus others?

Should we understand the instrumental inequality tests to differe in status in any way from conventional falsification tests?

- They have a kind of magic in that they get at something you would think is untestable

These tests versus others?

Should we understand the instrumental inequality tests to differe in status in any way from conventional falsification tests?

- They have a kind of magic in that they get at something you would think is untestable
- But are they "better"? Relates to the above question on power, in part.

These tests versus others?

Should we understand the instrumental inequality tests to differe in status in any way from conventional falsification tests?

- They have a kind of magic in that they get at something you would think is untestable
- But are they "better"? Relates to the above question on power, in part.
- Of course better to have more clever ways to falsify...unless power is basically 0.

Moving to the second part

I’m a bit confused about the connection between the two parts of the paper, and whether this ought to be one paper or two.

Moving to the second part

I'm a bit confused about the connection between the two parts of the paper, and whether this ought to be one paper or two.

Simulation: does sample size matter here...or is it all identification related and you just pick a large sample size (N=1e6!?!) to skip to the end of the story?

Moving to the second part

I'm a bit confused about the connection between the two parts of the paper, and whether this ought to be one paper or two.

Simulation: does sample size matter here...or is it all identification related and you just pick a large sample size (N=1e6!?!) to skip to the end of the story?

This seems intimately related to similar questions, inequality, and bounds that arise in recent work on Probabilities of Necessity and Sufficiency... see work by Scott Mueller with Pearl.

Applications and comparisons

I'd be curious about some of the basic statistics on the applications — sample size, strength of first stage, strength of reduced form. A bit hard to take in the results without that background.

Applications and comparisons

I'd be curious about some of the basic statistics on the applications — sample size, strength of first stage, strength of reduced form. A bit hard to take in the results without that background.

I wonder how the things you can learn here compare to what you can learn and reason about by other approaches including (of course) IV sensitivity anlysis in the OVB framework (Cinelli & Hazlett 2024+).

- One selling point of yours is the non-parametric setting...
- But when you have binary instrument and treatment (as in your first example), that's less restrictive, you could compare what we can learn.
- I suspect you'll find these applications fragile using the OVB-IV approach as well, in different terms.
- Though your approach also considers the montonicity violations, which is a plus.