Mining Causality: AI-Assisted Search for Instrumental Variables

Sukjin Han

University of Bristol
(Visiting Stanford GSB)

November 2024

Causal Data Science Meeting

Instrumental Variables Method

Endogeneity is the major challenge in conducting causal inference in observational settings

Instrumental variables (IVs) method is a leading empirical strategy for causal inference

- Imbens & Angrist 94, Blundell & Powell 03; Heckman & Vytlacil 05; Hernán & Robins 06; Mogstad & Torgovitsky 24

Finding IVs is challenging and is mainly a heuristic process

- relying on human researcher's creativity
- justifying its validity is largely rhetorical

IV Discovery Using Large Language Models

Large language models (LLMs) have sophisticated language processing abilities

We propose using LLMs to search for IVs

- through narratives and counterfactual reasoning
- similar to how a human researcher would do

LLMs can accelerate this process and explore an extremely large search space

IV Discovery Using Large Language Models

Benefits of AI-assisted approach to IV discovery:

1. search at a speedy rate, while adapting to the particularities of their settings
2. interacting with AI can inspire ideas for possible domains for novel IVs
3. increase the possibility of obtaining multiple IVs, which would then enable formal testing of validity via over-identifying restrictions
4. having a list of candidate IVs would increase the chances of finding actual data or guide the construction of such data (e.g., design of experiments)

Prompting Strategies

We demonstrate how to construct prompts to guide LLMs to search for IVs

- text representation of IV assumptions is the main component

1. multi-step prompting
   - divides search task into multiple subtasks
   - each stage's prompt focuses only on portion of the IV assumptions
   - separates counterfactual statements of different complexities
2. role-playing prompting
   - suitable for mimicking the endogenous decisions of economic agents
   - perspective of the agents in realistic scenarios
   - not of researchers searching for IVs

$\Rightarrow$ minimize likelihood that LLM perceives IV search task

Applications

We apply our method to well-known examples in economics using OpenAI's ChatGPT-4 (GPT4):

1. returns to schooling
2. supply and demand
3. peer effects

GPT4 produced list of candidate IVs and provided rationale for validity

- the list contains IVs that are popularly used in the literature
- and new IVs

Extensions

We extend our strategy to...

1. finding control variables in regression and difference-in-differences and
2. finding running variables in regression discontinuity designs

Related Social Science Literature

Using AI to assist heuristic parts of human research processes:

- Ludwig & Mullainathan 24, Mullainathan & Rambachan 24: hypothesis generation
- **this paper**: new variables discovered implicitly maintain hypotheses on validity

Using LLMs in economic research:

- Du, Kanodia, Brunborg, Vafa & Athey: fine-tuned LLM to predict job transition
- Manning, Zhu & Horton 24: automating entire research process
- **this paper**: LLMs in causal inference
  - incorporating structure from econometric assumptions and
  - allowing for human intervention in discovery processes

IV Assumptions

Let Y be outcome of interest; D be potentially endogenous treatment; X be covariates

Let $\mathcal{Z}_K \equiv \{Z_1, ..., Z_K\}$ be the list of IVs Z_k's

- K is the desired number of IVs to discover

Let $Y(d, z_k)$ be counterfactual outcome given (d, z_k)

We say Z_k is a valid IV if it satisfies the following assumptions:

Assumption REL (Relevance)

Conditional on X, the distribution of D given $Z_k = z_k$ is a nontrivial function of z_k.

Assumption EX (Exclusion)

For any (d, z_k), $Y(d, z_k) = Y(d)$.

Assumption IND (Independence)

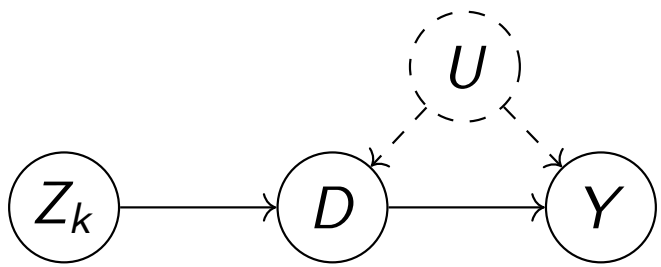
For any d, $Y(d) \perp Z_k$ conditional on X.

IV Assumptions

The goal is to search for IVs that satisfy Assumptions REL, EX and IND

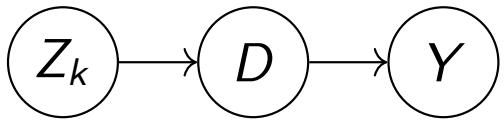
Causal direct acyclic graph (DAG) that implies the assumptions

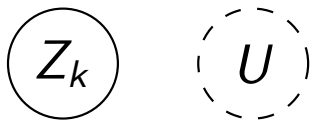
- $Y(d)$ being transformation of latent confounders U
- X suppressed



Prompt Construction

We propose a multi-step approach for IV discovery

- Step 1: prompt LLM to search for IVs that satisfy a verbal description of REL and EX
  - i.e., 

- Step 2, among the IVs in Step 1, prompt LLM to refine the search by selecting those that satisfy a verbal description of IND
  - i.e., 

- In both steps, the prompts will involve counterfactual statements
- In each step, we ask LLM to provide rationale for its responses
- (three-stage prompting in the paper)

Step 1: Prompt to Search for IVs

Prompt 1 (Search for IVs)

```
you are [agent] who needs to make a [treatment]
decision in [scenario].  what are factors that can
determine your decision but do not directly affect
your [outcome], except through [treatment] (that is,
factors that affect your [outcome] only through
[treatment])?  list [K_0] factors that are
quantifiable.  explain the answers.
```

Step 2: Prompt to Refine the Search

Prompt 2 (Refine IVs)

```
you are [agent] in [scenario], as previously
described.  among the [K_0] factors listed above,
choose [K] factors that are most likely to be
unassociated with [confounders], which determine your
[outcome].  for each factor chosen, explain your
reasoning.
```

Extension: Prompts to Search and Refine with Covariates

Prompt 2_x (Refine IVs with Covariates)

```
suppose you are [agent] in [scenario] with
[covariates].  among the [K_0] factors listed above,
choose [K] factors that are most likely to be
unassociated with [confounders], which determine your
[outcome].  for each factor chosen, explain your
reasoning.
```

Application: Returns to College

Prompt 1-1 (Example: Returns to College)

```
you are a high school graduate.  you need to make a
college attendance decision.  what would be factors
(factors of schools and factors of yourself) that can
determine your decision but that do not directly
affect your future earnings, except through college
attendance (that is, that affect your earnings only
through college attendance)?  list forty factors that
are quantifiable, twenty for school factors and twenty
for factors of yourself.  explain the answers.
```

Application: Returns to College

Prompt 2_x-1 (Example: Returns to College)

```
suppose you are a student with family income $10K per
year, who is asian female from california, whose
parents have college education, who is catholic.
among the forty factors listed above, choose all
factors that are not associated with your innate
ability and personality and school quality, which
determine earnings.  create separate lists for school
factors and factors of yourself.  for each factor
chosen, explain the reasoning.
```

Application 1: Returns to College

Discovered IVs:

- Distance from home to college (Card 95 (1955); Card 01 (3134))
- Personal savings for education
- Number of siblings attending college (Griliches 77 (2393))
- Parents' educational background
- Religious affiliations

Application 1: Returns to College

Discovered IVs (continued):

- Campus crime rates
- Diversity of student body
- Library size (volumes held)
- Availability of online courses
- Campus facilities quality (gyms, labs, etc.)
- Campus housing capacity
- Proximity to urban centers
- Environmental sustainability rating
- Technology integration in classrooms

Application 2: Returns to Years of Schooling

Discovered IVs:

- Credits Required for Graduation
- State Education Laws (Angrist & Krueger 91 (3951); Staiger & Stock 97 (10656))
- Distance to Nearest College/University (Card 95 (1955); Card 01 (3134))
- Transportation Availability
- Local Education Funding per Student
- Local College Enrollment Rates
- Local Unemployment Rates

Application 3: Peer Effects

Discovered IVs:

- Engagement Rate
- Frequency of Posts
- Content Reach
- Number of Platforms Used
- School Policies on Smoking
- Local Regulations on Tobacco Advertising
- Access to Smartphones Among Peers
- Visibility at School Events
- History of Disciplinary Actions at School
- Academic Performance
- Extracurricular Leadership Roles

Application 4: Demand Estimation

Discovered IVs:

- Fuel Costs
- Fishing Equipment Depreciation
- Weather Conditions (Angrist, Graddy & Imbens 00 (454))
- Regulatory Costs
- Interest Rates
- Insurance Costs
- Utility Costs at Sales Points

Application 4: Demand Estimation

Discovered IVs:

- Economic Conditions
- Technological Advances
- Government Subsidies
- Tariffs on Imports
- Employee Training Costs

Extension: Conditional Independence

Consider a conditional independence (CI) assumption

- more crucial role of the vector of control variables $X \equiv (X_1, ..., X_L)$

Assumption CI (Conditional Independence)

For any d, $D \perp Y(d)|X$.

- common introduced in causal inference
- esp. when combined with machine learning (e.g., debiased/double machine learning)
- closely related to matching and propensity score matching

Extension: Conditional Independence

Prompt C1 (Search for Control Variables)

```
you are [agent] who needs to make a [treatment]
decision in [scenario]. what factors determine your
decision? list [L_0] factors that are quantifiable.
explain the answers.
```

Prompt C2 (Search for Control Variables)

```
among the [L_0] factors listed above, choose all
factors that directly determine your [outcome], not
only indirectly through [treatment]. the chosen
factors can still influence your [treatment]. for
each chosen factor, explain the reasoning.
```

Parallel Trend in Difference-in-Differences

Assumption PT (Parallel Trend)

$E[\Delta Y(0)|D, X] = E[\Delta Y(0)|X]$ where
$\Delta Y(0) \equiv Y_{after}(0) - Y_{before}(0)$.

- PT can be viewed as a mean independence version of CI

Therefore, Prompts C1–C2 can be directly used to search for X that satisfy PT

- by inputting "`average temporal changes in [outcome_t] during the time of no [treatment]`" for `[outcome]` in Prompt C2

Application: Effects of Minimum Wage

Discovered control variables:

- Inflation Rates
- Consumer Price Index (CPI)
- Job Vacancy Rates
- Labor Productivity Growth
- Employment Growth Rates
- Labor Force Participation Rate
- Union Membership Rates
- Turnover Rates
- Corporate Profit Trends
- Economic Diversity Score
- Percentage of Workforce in Gig Economy

Discussions

Possible ways for sophisticating the prompts:

1. using previously known IVs in the literature to guide LLMs
   - evoke *few-shot learning* in LLMs (Brown et al 20)
   - "orthogonalize" the search
2. elaborated search toward finding IVs that are more policy-relevant (Imbens & Angrist 94; Heckman & Vytlacil 05)
3. incorporating detailed contextual information
   - e.g. via system messages
4. aggregation of findings across sessions
   - account for and leverage the stochastic nature of LLMs' responses

Concluding Remarks

How to validate LLM's performance?

- horse race among LLMs? fine-tuning (Du et al 24)?
- ground truth of valid IVs?
- very reason we propose using LLMs from the first place!

LLMs can be collaborative tools in conducting causal inference research!

- other applications?
- feel free to reach out!

Thank You! ☺