Mining Causality: AI-Assisted Search for Instrumental Variables

Sukjin Han

University of Bristol

World Congress 2025

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. 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)
4. increase the possibility of obtaining multiple IVs, which would enable formal testing of validity via over-identifying restrictions

Prompting Strategies

We demonstrate how to construct prompts to guide LLMs

- 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
2. role-playing prompting
   - mimic the endogenous decisions of economic agents
   - perspective of the agents in realistic scenarios
   - not of researchers searching for IVs

⇒ reduce likelihood that LLM perceives IV search task and suggests existing IVs

Prompting Strategies

3. multi-agent prompting
   - adversarial LLM reviews the proposed IVs
   - feedback to defender LLM to refine the proposal

⇒ generate more sophisticated responses

Applications

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

1. returns to schooling
2. demand estimation
3. production function estimation
4. 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 IVs that appear novel

Expert evaluation from survey: some IVs are novel and at least as valid as established IVs

Extensions

From a broader perspective, the proposal is to systematically "search for exogeneity"

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

Related Social Science Literature

Using LLMs in causal discovery:

- Ban et al 23, Cohrs et al 24, Jiralerspong et al 24, Le et al 24, Long et al 23, Takayama et al 24
- **this paper**: discover variables with particular causal structure rather than finding causal links among *pre-determined* variables

IV Assumptions

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)

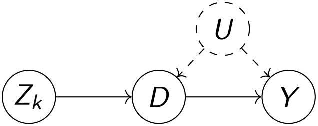
For any d, $Y(d) \perp Z_k$ conditional on X.

IV Assumptions

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

Causal direct acyclic graph (DAG) that implies the assumptions:

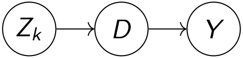
- $Y(d)$ being transformation of latent confounders U
- X suppressed

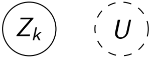


Prompt Construction

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., 

- (three-stage prompting in the paper)
- additionally, each step's prompt is given in the form of a role play
- in each step, we ask LLM to provide rationale for its responses

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]? list [K_0] factors that are quantifiable. explain the answers.

Step 2: Prompt to Refine the Search

Prompt 2_x (Refine IVs)

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.

Why Multi-Step Approach

1. better performance when handling subtasks step-by-step
    - chain of thought (CoT)
2. separates counterfactual statements of different complexities
    - more room for user inspection
3. intermediate outputs can offer insights
4. significantly reduce the likelihood that LLMs recognize IV discovery
    - esp. when running each step in independent session

Why Role-Playing Approach

1. natural as D represents an economic agent's decision
2. LLMs gather better contextual information and generate more tailored and unique responses
3. more effective in guiding LLMs to respond as the relevant economic agent rather than as a researcher searching for IVs

Discovered IVs

Application 1: 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 1: 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))
  - “This is purely geographical and independent of personal attributes.”
- 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
  - "This is a regulatory requirement and does not depend on your personal qualities or the inherent quality of your school but can determine the length of time you stay in school."
- 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: 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 3: Demand Estimation

Discovered IVs (continued):

- Economic Conditions
- Technological Advances
- Government Subsidies
- Tariffs on Imports
  - "Tariffs imposed on imported fish are a matter of international trade policy and do not change based on daily fluctuations in market demand or customers characteristics."
- Employee Training Costs

Application 4: Peer Effects

Discovered IVs:

- Engagement Rate
  - "While the number of followers might be initially influenced by shared backgrounds, the engagement rate depends more on the content quality and how it resonates with the audience at any given time, rather than the reasons why the audience initially formed."
- Frequency of Posts
- Content Reach
- Number of Platforms Used

Application 4: Peer Effects

Discovered IVs (continued):

- School Policies on Smoking
- Local Regulations on Tobacco Advertising
- Access to Smartphones Among Peers
  - "This might vary widely even within similar socio-economic backgrounds due to individual family decisions or priorities."
- Visibility at School Events
- History of Disciplinary Actions at School
- Academic Performance
- Extracurricular Leadership Roles

Expert Evaluation

Validation through Expert Survey

Conducted survey of experts in various fields of empirical economics

Elicit evaluation of the IVs discovered by LLMs

1. validity score: 1 = least valid; 5 = most valid
2. non-novelty score: 1 = least likely to exist in the literature; 5 = most likely to exist
   - lower non-novelty scores = *higher* IV novelty

We report:

1. $E[\text{validity}|\text{non-novelty} = k]$ for each $k = 1, \ldots, 5$
2. IV names s.t.

$$E[\text{validity}|\text{non-novelty} = 1 \text{ or } 2,\ \text{IV name}] \geq \text{benchmark}$$

   - benchmark = $E[\text{validity}|\text{non-novelty} = 5]$

Evaluation Results

Average Validity Across Novelty
($k = 1$: most novel; $k = 5$: least novel)

Domain[†]	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$ (benchmark)
1	1.87	2.23	2.72	2.17	2.63
2	3.00	2.10	2.39	2.97	3.19
3	1.92	2.22	2.70	3.19	3.72
4	1.92	2.39	2.63	3.76	3.00
5	2.36	2.26	2.96	2.39	3.67
6	2.10	2.03	2.53	3.00	2.00

[†] Domain ID: 1. Returns to Schooling; 2. Returns to Years of Education; 3. Production Functions; 4. Demand Estimation; 5. Peer Effects (technology adoption); 6. Peer Effects (teenage smoking). $N = 10$.

Evaluation Results

IVs with $k = 1$ or 2 Outperforming Benchmark
($k = 1$: most novel; $k = 5$: least novel)

Domain[†] (Benchmark)	Novel IVs Suggested
1 (2.63)	Campus Housing Capacity (+0.70) Technology Integration in Classrooms (+0.37) Campus Crime Rates (+0.03)
2 (3.19)	Credits Required for Graduation (+0.31) Transportation Availability (+0.31)
4 (3.00)	Fuel Costs (+1.00) Tariffs on Imports (+1.00) Technological Advances (+0.00) Government Subsidies (+0.00)

Evaluation Results

IVs with $k = 1$ or 2 Outperforming Benchmark
($k = 1$: most novel; $k = 5$: least novel)

Domain[†] (Benchmark)	Novel IVs Suggested
6 (2.00)	Access to Smartphones Among Peers (+1.50)
	Content Reach (+1.33)
	Engagement Rate (+1.00)
	Local Regulations on Tobacco Advertising (+1.00)
	Number of Platforms Used (+0.40)
	Frequency of Posts (+0.33)
	Visibility at School Events (+0.25)
	School Policies on Smoking (+0.00)
	History of Disciplinary Actions at School (+0.00)
	Academic Performance (+0.00)

Adversarial LLM

Refinement with Adversary

Request another LLM to play the role of adversary

Procedure:

1. the defender LLM finds IVs
2. the adversarial LLM review the responses and produces counter-arguments (without using jargon)
   - full disclosure of the IV discovery task in the adversarial stage
3. the counter-arguments are given to the defender LLM
4. the defender LLM refine the previous responses

Overall, this procedure generates more sophisticated responses

Refinement with Adversary

Returns to college:

- same: “geographical proximity” or “transportation”
- “school facilities” are refined
  - previous: “library size”, “technology integration in classrooms”
  - “percentage of renewable energy used on campus”, “number of green spaces on campus”, “campus medical facilities”
- refined variables appear to be weak IVs

Returns to year of schooling:

- “state laws” is refined to “mandatory minimum years of schooling required by state”

Refinement with Adversary

Demand estimation:

- “weather condition” is refined to “global climate patterns (e.g., El Niño)”
- because the counter-argument was “bad weather can also discourage customers from coming to the market”

Discussions

Comparison to Direct Approach

Alternatively, one can directly “teach” IVs to LLM and ask to find ones in specific setting

- this LLM generates mostly known IVs
- when asked “`did you have sources for your answers?`” it responds by citing well-known references on IVs
- in contrast, with the proposed method, LLM responds that it “didn’t have specific sources”
- when pressed to provide sources (in the returns to education application) it cites academic references from behavioral and social sciences on students’ decision-making, including qualitative and case studies

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, Turnover Rates
- Labor Productivity Growth
- Employment Growth Rates
- Labor Force Participation Rate
- Union Membership Rates
- Corporate Profit Trends
- Economic Diversity Score
- Percentage of Workforce in Gig Economy

Conclusions

Concluding Remarks

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

Human validation of the LLM's performance

- ground truth of valid IVs?
- very reason we propose using LLMs from the first place!

LLMs can be useful tools in conducting causal inference and economic research

- other applications?
- more to come, stay tuned!

Thank You! ☺