Instrumental variable methods under dependent censoring

A Data Management Plan created using DMPonline.be

Creator: Gilles Crommen

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 11PKA24N

ID: 203026

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Instrumental variable methods are a popular set of econometric techniques to identify and estimate causal effects. They can however fail when the outcome variable of interest is a duration. For example, consider evaluating the causal effect of a job training service on unemployment duration, where job seekers often stop answering follow-up surveys before they find new employment. In that case, the unemployment duration is said to be right censored by some censoring time. To solve this issue, it is customary to assume that the duration and censoring time are statistically independent of each other. However, this assumption is likely to be violated in the example on job seekers, that is there is dependent censoring. The goal of this fellowship is to develop new instrumental variable methods that are valid under dependent censoring. The work that follows from this research proposal aims to (i) suggest an estimator for the complier causal hazard ratio under a semiparametric copula model for dependent censoring; (ii) develop a model to recover a quantile treatment effect under dependent censoring; (iii) provide a review paper on instrumental variable methods in duration analysis; and (iv) explore other avenues such as double machine learning or local distribution regression. This research will fill an important gap in the literature on instrumental variable methods for duration outcomes, where the independent censoring assumption is often taken as a given without proper reasoning.

Last modified: 19-06-2024

Instrumental variable methods under dependent censoring
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
(JTPA)	The National Job Training Partnership Act Study	Reuse	Digital	Observational	DTA,SD2 and CSV	<1GB	
HIPB-2	Health Insurance Plan of Greater New York Breast Cancer Screening Trial	Reuse	Digital	Observational	CSV	<1GB	
R-scripts	Code written to apply new methods to data	Generate	Digital	Software	.R scripts	Estimated <100MB	
Numerical results	Results from simulations	Generate	Digital	Simulation data	text (.txt) and Excel (.xlsx) files	Estimated around 5GB	
Manuscripts	text to describe new methods, methodologies and results	Generate	Digital	Other	Stored on Overleaf and on my laptop as .tex files and PDF files	Estimated to be around 10GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

JTPA: https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study
HIPB-2: https://cdas.cancer.gov/login/?next=/projects/hipb/2/agreements/

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will process pseudonymous data from the HIPB-2 data set. The Health Insurance Plan of Greater New York (HIP) Breast Cancer Screening Trial was a randomized screening trial (1963–1986) designed to test whether breast cancer mortality rates could be reduced through the use of periodic screening. The periodic screening consisted of an initial examination followed by 3 subsequent annual examinations. The screening consisted of a clinical breast examination and mammography. 60,695 observations, 1 per participant. This dataset includes study demographic (race and age), screening, cancer, death, and follow-up variables. PRET reference number: G-2024-8228

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

HIPB-2:
RECIPIENT must not use DATA for any study other than the approved Research Planunless RECIPIENT obtains the written consent of NCI by way of a new approved application through CDAS or by written and signed amendment to this Agreement. RECIPIENT grants NCI the right to publicly disclose the Research Plan, including titles, summaries or any other information contained therein as well as the names and contact information for the investigators conducting the research. The DATA will be used solely by RECIPIENT LEAD INVESTIGATOR and RECIPIENT's faculty, employees, fellows, students, and agents that have a need to use, or provide a service in respect of, the DATA in connection with the Research Plan and whose obligations for using the DATA are consistent with the terms of this Agreement. This Agreement shall be in effect for five (5) years from the Effective Date. At the end of these five (5) years, if RECIPIENT is still using DATA for the approved Research Plan, RECIPIENT may seek an amendment to extend the term of this Agreement. This Agreement may be terminated by either Party for any reason by providing written notice to the other Party at least thirty (30) days prior to the desired termination date. Upon expiration or earlier termination of this Agreement or if RECIPIENT's use of DATA is complete, RECIPIENT must destroy DATA and upon NCI's request, confirm in writing as to such destruction. The RECIPIENT may retain one (1) copy of the DATA to the extent necessary to comply with the records retention requirements under any law, and for the purposes of research integrity and verification.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

HIPB-2:
The DATA will not be further distributed to others without NCI's written consent. The RECIPIENT shall refer any request for the DATA to NCI. DATA are the property of NCI and are made available as a service to the research community. RECIPIENT will not claim, infer, or imply ownership of DATA or any endorsement of RECIPIENT'S activities or products by the U.S. Government, DHHS, NIH, NCI, or NCI employees. RECIPIENT will acknowledge NCI as the source of DATA in all publications and presentations by including language

substantially similar to the following: "The authors thank the National Cancer Institute for access to NCI's data collected by the HIP Breast Cancer Screening Trial (HIPB)". Each publication and presentation should reference the CDAS Project Number. RECIPIENT must submit a description of each publication resulting from its use of DATA to the following website: https://cdas.cancer.gov/projects/hipb/2/HIPB-2/discussion/. RECIPIENT agrees that NCI may publicly disclose this description.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

R-scripts:
Every R file will have some comments added to the script.
Numerical results:
All the generated data sets and numerical results are stored locally, with the file name including the time, date and specific setting it was generated from.
Manuscripts:
They are stored locally and on Overleaf. Back-ups are made from time to time so that there is also some version control.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

All publically available data will be stored in the cloud on the OneDrive of KU Leuven of Gilles Crommen, which has a capacity of 2TB. The HIPB-2 data will be stored on a secure hard drive.

How will the data be backed up?

Backups are made on a secure hard drive every week by Gilles Crommen.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

We will not need more capacity than the 2TB available on the cloud. Extra storage can always be bought when necessary for hard drive backups.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data on the OneDrive of KU Leuven are password protected. The hard drive with the backup is safely kept in Gilles' office and also

password protected.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

No costs are expected.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 10 years.

Where will these data be archived (stored and curated for the long-term)?

The hard drive will be stored by Gilles Crommen. Additionally, all finished manuscripts will be made publically available on arXiv. All finished R-scripts and numerical results will be publicly available on Gilles' GitHub account. The JTPA data is publicly available. A full copy of all the data will be given to Ingrid Van Keilegom and Jad Beyhum at the end of the project.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No costs are expected.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

All finished manuscripts will be made publically available on arXiv. All finished R-scripts and numerical results will be publicly available on Gilles' GitHub account. The JTPA data is publicly available.

If access is restricted, please specify who will be able to access the data and under what conditions.

HIPB-2:
RECIPIENT must not use DATA for any study other than the approved Research Planunless RECIPIENT obtains the written consent of NCI by way of a new approved application through CDAS or by written and signed amendment to this Agreement. RECIPIENT grants NCI the right to publicly disclose the Research Plan, including titles, summaries or any other information contained therein as well as the names and contact information for the investigators conducting the research. The DATA will be used solely by RECIPIENT LEAD INVESTIGATOR and RECIPIENT's faculty, employees, fellows, students, and agents that have a need to use, or provide a service in respect of, the DATA in connection with the Research Plan and whose obligations for using the DATA are consistent with the terms of this Agreement. This Agreement shall be in effect for five (5) years from the Effective Date. At the end of these five (5) years, if RECIPIENT is still using DATA for the approved Research Plan, RECIPIENT may seek an amendment to extend the term of this Agreement. This Agreement may be terminated by either Party for any reason by providing written notice to the other Party at least thirty (30) days prior to the desired termination date. Upon expiration or earlier termination of this Agreement or if RECIPIENT's use of DATA is complete, RECIPIENT must destroy DATA and upon NCI's request, confirm in writing as to such destruction. The RECIPIENT may retain one (1) copy of the DATA to the extent necessary to comply with the records retention requirements under any law, and for the purposes of research integrity and verification.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

HIPB-2:
RECIPIENT must not use DATA for any study other than the approved Research Planunless RECIPIENT obtains the written consent of NCI by way of a new approved application through CDAS or by written and signed amendment to this Agreement. RECIPIENT grants NCI the right to publicly disclose the Research Plan, including titles, summaries or any other information contained therein as well as the names and contact information for the investigators conducting the research. The DATA will be used solely by RECIPIENT LEAD INVESTIGATOR and RECIPIENT's faculty, employees, fellows, students, and agents that have a need to use, or provide a service in respect of, the DATA in connection with the Research Plan and whose obligations for using the DATA are consistent with the terms of this Agreement. This Agreement shall be in effect for five (5) years from the Effective Date. At the end of these five (5) years, if RECIPIENT is still using DATA for the approved Research Plan, RECIPIENT may seek an amendment to extend the term of this Agreement. This Agreement may be terminated by either Party for any reason by providing written notice to the other Party at least thirty (30) days prior to the desired termination date. Upon expiration or earlier termination of this Agreement or if RECIPIENT's use of DATA is complete, RECIPIENT must destroy DATA and upon NCI's request, confirm in writing as to such destruction. The RECIPIENT may retain one (1) copy of the DATA to the extent necessary to comply with the records retention requirements under any law, and for the purposes of research integrity and verification.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All finished manuscripts will be made publically available on arXiv. All finished R-scripts and numerical results will be publicly available on Gilles' GitHub account. The JTPA data is publicly available at https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study

When will the data be made available?

When manuscripts are submitted for publication.

Which data usage licenses are you going to provide? If none, please explain why.

None.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

No costs are expected.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Gilles Crommen

Who will manage data storage and backup during the research project?

Gilles Crommen

Who will manage data preservation and sharing?

Gilles Crommen (Ingrid Van Keilegom and Jad Beyhum after the project is finished)

Who will update and implement this DMP?

Gilles Crommen