Exploring Complications With Instrumental Variable Selection and Their Solutions

Kieran Douglas,

Literature Review

Undergraduate Research Opportunities Consortium, University of Arizona

Dr. Anna Josephson

July 15th, 2024

Preface

The dynamic and complex nature of the world can make it difficult to create models that accurately and precisely measure its contents in consistent and valid ways. There are a variety of models which do a reasonably good job at improving predictive power and thus the understanding of researchers about the world, but all have their share of issues. Flaws in models and methods can be pervasive, and long go unnoticed, which can lead to a large volume of research that may have severe validity problems, corrupting the scientific record.

Emblematic of this are critiques of the Instrumental Variables Regression (IV). These criticisms emerged in the 1980s and have since spawned skepticism related to their application throughout the social sciences. This has given rise to a rich area of literature which investigates the use of IVs in economics and develops certain criteria and techniques by which researchers can select their IVs in hopes of avoiding validity problems. This review provides an overview of the existing literature on IV selection and their validity problems. Specifically, I investigate the use of weather as an IV. Additionally, this review will explore a new method for more efficient data collection, which employs a Large Language Model that is trained to extract specific information from academic papers.[1]

[1] It is important to note that in the field of Economics publishing can often be a cumbersome and drawn-out process, relative to other social science and STEM fields (Önder et al., 2011). This is largely the result of a broad shift towards more empirical applied work which coincides with papers that are much longer and denser than what may be common in other disciplines (Josephson & Michler, 2023). Several of the works cited in this review are still under revision at the time of its creation. In economics it is not unusual to cite such papers, including papers in review, working papers, etc. Similarly, the method used for multi-paper data collection is based on a recent thesis which aims to provide guidance on rainfall variable selection for IV regression (Agme et al., 2024).

Literature on Instrumental Variables

The first use of IV regression in economics appeared in Wright (1928), in which the author demonstrated that IV regression could be used to estimate coefficients on endogenous regressors. This technique provided a method to address endogeneity among explanatory variables. Endogeneity is a problem in which the explanatory variable (X) and the outcome (Y) are correlated with some confounder (U) in the error term leading to biased estimates. IVs offered a way around this problem by assuming that the IV (W) affects the outcome (Y) only through the explanatory variable (X) and is not correlated with the error term (U). This is called the exclusion restriction and is a key assumption to IV regression. If an instrument is correlated with the error term, or if there is some other explanatory variable (J) that is impacted by the IV (W) and the error term (U), there is a violation to the exclusion restriction. Such a violation may produce biased estimates or outright invalid causal inference.

At the time of its advent, the problems and pitfalls of IVs were not evident. Awareness began to grow in the late 1980s and 1990s, with work from Bound et al. (1995) who suggested that finding valid instruments could be much more of a challenge than initially thought, as well as from Angrist et al. (1996) who formalized IV assumptions, making it easier to understand their potential for violations.[2] This increased scrutiny towards the use of IV regression led researchers to reexamine previous work. Staiger & Stock (1997), for example, suggest specific guidelines for papers that use IVs to avoid the potential for a weak instrument problem.[3] They also use this development to explore Angrist and Krueger (1991) that uses an IV to estimate the returns to

[2] For those interested, the two most notable IV assumptions include relevance and exogeneity. Relevancy holds that the instrument must be correlated with the endogenous explanatory variable of interest. Exogeneity states that the instrument must not have an independent effect on the outcome, known as an exclusion restriction (Jagsi and Yu, 2014). If either of these core assumptions are violated the instrument may face severe validity problems.
[3] When an instrument has a weak correlation with the endogenous explanatory variable.

education. They find that with their updated methodology, estimates on returns to education are higher than initially reported.

Morck & Yeung (2011) went as far as to suggest that economists would be better off viewing historical information[4] not as a tool shed of IVs, but as a complex chain of causality that can often be deciphered better through detailed context, the plausibility of alternative narratives, external consistency, and the recognition that human decision making is intrinsically exogenous. This argument may seem regressive, but Morck & Yeung observe several strict limitations on the validity of IVs can greatly affect their utility to researchers. They posit that this reduced utility is due to a tragedy of the commons that exists when it comes to the use of specific IVs. Each additional study that employs the same IV has the chance of revealing new correlations that run the risk of weakening the overall exogeneity assumption for that variable. Intuitively, increased use coupled with growing scrutiny opens the door for a whole slew of previously unobserved effects on outcomes or realizations of a weak instrument problem or violation of the exclusion restriction. Thus, Morck & Yeung additionally argue that each successful use of a particular instrument can potentially create additional latent variable bias problems for all other uses of that instrument.

The arguments put forth in Morck & Yeung (2011) address a commonly held question on the state of economics research: how we effectively bridge the gap between abstract mathematical models and complex real-world phenomena. Though the profession could likely benefit from incorporating more interdisciplinarity, the shift towards empirical procedures is quite self-correcting. Growing skepticism tends to pave the way for more enhanced precision in methodologies. This is exactly what has been demonstrated more recently in the use of weather

[4] Historical information is often used as an IV as decisions, actions, or events can be viewed as "random" to modern phenomena (e.g., Nunn, 2008; Donaldson, 2018; Dube & Harish, 2020).

as an IV. Consider the seminal paper is this domain: Mellon (2023), which examines 289 studies to reveal 195 variables that were previously linked to weather. These variables all represent potential exclusion restriction violations to the instrument used. Through sensitivity analysis, he found that the magnitude of the violations present to the exclusion restriction (necessary for a valid IV) were sufficient to overturn several existing results and raise serious questions of validity for several others. Mellon demonstrates that the use of weather as an instrument is fundamentally flawed because it violates the exclusion restriction assumption for most plausible applications. His findings suggest that the current set of criteria for accepting something as an IV tends to be quite unsystematic, which can lead to serious validity problems.

In a similar vein, Agme et al. (2024) explore a more specific application of weather as an instrument by looking at the use of rainfall data to explain a wide variety of outcomes. The ambiguity becomes evident when considering what rainfall data looks like when being used as an IV. Different researchers use an unsystematic method to choose among a diverse selection of metrics to represent rainfall data (e.g., mean annual rainfall, deviations in total rainfall, total rainfall) to predict agricultural productivity. This raises the question of whether researchers would be better off using one metric over another. Agme et al. (2024) gather remote sensing data and calculated fourteen common rainfall metrics used in economics literature. They examine the predictive power of each metric-remote sensing pair on agricultural productivity, specifically looking for metrics that are: (a) consistently significant across remote sensing products and countries; (b) have a consistent sign; and (c) are similar in significance and sign to other metrics (suggesting substitutability). Agme et al. (2024) find that there are few metrics that are consistent in predicting agricultural productivity, as certain instruments will flip (e.g., from positive to negative) the sign and statistical significance (e.g., from statistically significant to statistically

insignificant) of its predictive power. Additionally, while rainfall is external to some degree, there are within-household decision making factors that may affect outcomes and tend to be correlated with characteristics that do not change over time. Without controlling for these fixed effects[5] rainfall as an IV often fails to satisfy the exclusion restriction, but after controlling for them, it appears to develop a weak instrument problem. This implies that rainfall may have a direct impact on both the endogenous explanatory variable and agricultural productivity. Agme et al. (2024) suggest that rather than abandoning rainfall as an IV altogether, researchers should work to better justify their choices and pick from a smaller set of metrics that give more consistent results.

Using a Large Language Model to Investigate Instrumental Variables

Unlike the process used by Mellon (2023) in which data were collected through the manual examination of each of the 289 studies to identify relevant information (e.g., instrument used, explanatory variable, outcome), Agme et al. (2024) attempt a new method using artificial intelligence with the goal of improving efficiency and broadening the scope of what is feasible. To start, Agme et al. use a comprehensive catalog of scholarly works called OpenAlex to generate a dataset containing all available economic papers that use rainfall as an IV. Then, they use the Large Language Model (LLM) ChatGPT to “read" through each of the papers with the goal of categorization and identification of variables. Agme et al. (2024) instruct the model to generate a CSV file containing the relevant information to be used for further analysis in the paper. Due to the novelty of this method and the relatively young model, the process of getting it to behave was a bit tedious.[6] The LLM does not process information like humans, and due to the

[5] These are unobserved variables that don't change with time like race, which if left unaccounted for, could lead to biased estimates given its relation to other variables. In the case of Agme et al. (2024) household fixed effects include things like past experience with weather shocks and geographic location.

[6] Though he did not elaborate on this in his paper, we had several conversations throughout the course of my research project in which we discussed some of the complications that came up in getting the LLM to behave. He trained and tested the model through its web interface,

often-complex structure of academic papers, it was difficult for the model to return consistent and accurate results. Additionally, Agme et al. (2024) did not formally document the steps necessary to achieve the desired output (which I will do). This provides an opportunity for expanding upon their work to create a reproducible and well cataloged method for the use of LLMs as part of the data generating and literature review process, in hopes that it can be applied elsewhere to enhance future research.

Method and Contribution to the Literature

Based on Agme et al. (2024) and their OpenAlex output of academic works containing rainfall as an IV, I am writing a Stata script that calls a ChatGPT API to analyze each paper. Ideally, the LLM should "read" through each paper in its entirety and output a CSV that highlights prespecified information including the rainfall metric used, the paper title, and the outcome of interest. Additionally, the code calls on an API that matches the paper contents to its DOI via the internet to improve accessibility. The approach that I am taking in this project is like that which may be used with other types of machine learning. Random forests, for example, is an ensemble method that uses smaller training subsets of data to learn patterns that it can then use to make predictions for an entire dataset. Similarly, I am training the model on a subset of papers, some of which are confirmed to have specific information that I want the model to identify, to ensure that it is providing me with accurate results.

Due to the LLM data collection process being a very new approach to research, there is not much literature surrounding its application. Aydın & Karaarslan (2022) note that we are just scratching the surface when it comes to the advances that will be made possible by this novel technology,

manually feeding it each PDF and refining the script with each step. Since my project involved writing code with the same purpose but in a more efficient and repurposable way, I used some variation of several of the natural language techniques that he recommended to help me get around similar hurdles.

and in the near future, the research and data gathering processes will likely be far less time consuming. In the field of economics especially, few researchers have begun to explore its use. This makes it even more exciting, as this proof-of-concept may be one of the first steps towards a completely new strategy to be used in the research process. Agme et al. (2024) struggled to acquire consistent and reliable data from the LLM, but it seemed to improve as it was exposed to greater quantities of training data. I have run into similar issues with the model and have found myself relying heavily on backwards induction to teach the model what I want it to look for. In theory, with enough data the model could provide near perfect information regarding each paper[7] and its contents, which could massively accelerate the process of data collection and broaden the scope of what is achievable when it comes to questions like IV selection. Ideally this work will permit data collection from a larger body of literature which may provide a stronger and more definitive argument regarding rainfall metric selection for IV regression models. This is particularly relevant for systematic reviews and meta-analyses which rely on the analysis of an entire body of literature, disparate from a literature review of the present sort (Deeks et al., 2023). Rather than reading through hundreds of papers like Mellon (2023) or spending hours trying to convince a LLM what to look for like Agme (2024), the hope is that eventually the process will be near if not fully automated, providing researchers with reliable data at a lower cost.

[7] LLMs are capable of processing data and generating outputs, but it is not based on a true comprehension of what the model is exposed to. They provide information based on statistical patterns, so there will likely always exist some degree of uncertainty (Mitchell & Krakauer, 2023)

Conclusion:

The growing body of literature and skepticism surrounding the use of IVs has raised many important questions surrounding their validity. Given the importance and wide applicability of IV regression in economics and beyond, the argument is not that it should be outright abandoned, but rather examined and improved. Much of the recent literature has both highlighted flaws in IV use to the extent that they may completely invalidate the findings, in addition to solutions and recommendations for updated methodology to avoid violating the methods' validity assumptions (Staiger & Stock, 1997; Morck & Yeung, 2011; Mellon, 2023; Agme et al., 2024). If the newly developed criteria are solidified and widely adopted, the quality and accuracy of results among papers that use IVs will be drastically improved. An important part of this methodological evolution involves increasing the quantity of data collected so that metric strength can be tested, and a consensus can be reached for what makes it reliable. This is largely what my project aims to do, acting as both a proof of concept for a new data collection method fueled by artificial intelligence, and increasing the quantity of information available for testing. Ideally, this method can be utilized for future IV research that looks to explore and improve upon specific application and metric use.

When Morck & Yeung (2011) refer to a "Tragedy of the Commons" that is present with IVs due to weakened instruments from overuse, a negative association comes to mind. The observed weak instrument problem created is, however, a marker of good science. Complacency is dangerous, and skepticism keeps the world moving forward. To find previous research invalid may be frustrating or feel like a personal attack to many researchers, but a vital aspect of good science is that it builds upon itself. What was true yesterday may not hold today, but that does not mean that yesterday's findings were without value.

References:

Agme, C. D., Josephson, A., & Michler, J. D. (2024). Variable selection in economic applications of remotely sensed weather data: Evidence from the LSMS-ISA [Master's thesis, University of Arizona]. Department of Agricultural and Resource Economics.

Arizona Ali Önder, Mario Crucini , Robert Driskill , John Conley. (2011, October 24). *Publication lags and young economists' research output*. CEPR. https://cepr.org/voxeu/columns/publication-lags-and-young-economists-research-output

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996a). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.2307/2291629

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996b). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.2307/2291629

Angrist, J. D., & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Aydın, Ö., & Karaarslan, E. (2022). *OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare* (SSRN Scholarly Paper 4308687). https://doi.org/10.2139/ssrn.4308687

Bound, J., Jaeger, D. A., & Baker, R. M. (1995a). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.2307/2291055

Bound, J., Jaeger, D. A., & Baker, R. M. (1995b). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.2307/2291055

Bound, J., Jaeger, D. A., & Baker, R. M. (1995c). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.1080/01621459.1995.10476536

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2023). Chapter 10: Analysing data and undertaking meta-analyses. Cochrane Handbook for Systematic Reviews of Interventions (Version 6.4). Cochrane. https://training.cochrane.org/handbook/current/chapter-10

Donaldson, D. (2018). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *American Economic Review*, *108*(4–5), 899–934. https://doi.org/10.1257/aer.20101199

Imbens, G. W. (2014). Instrumental Variables: An Econometrician's Perspective. *Statistical Science*, *29*(3), 323–358.

Josephson, A., & Michler, J. D. (2023). *Research Ethics in Applied Economics: A Practical Guide*. Routledge. https://doi.org/10.4324/9781003025061

Mellon, J. (2023). *Rain, Rain, Go Away: 195 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable* (SSRN Scholarly Paper 3715610). https://doi.org/10.2139/ssrn.3715610

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Morck, R., & Yeung, B. (2011). Economics, History, and Causation. *The Business History Review*, *85*(1), 39–63.

Nunn, N. (2008). THE LONG-TERM EFFECTS OF AFRICA'S SLAVE TRADES. *QUARTERLY JOURNAL OF ECONOMICS*.

Reshma Jagsi and James B. Yu. (2014). *Instrumental Variable Analysis—An overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/medicine-and-dentistry/instrumental-variable-analysis

Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, *65*(3), 557–586. https://doi.org/10.2307/2171753

Stock, J. H. (1997). *Testing for Weak Instruments in Linear IV Regression*.

Stock, J. H., & Yogo, M. (2002). *Testing for Weak Instruments in Linear IV Regression* (SSRN Scholarly Paper 346941). https://papers.ssrn.com/abstract=346941