The Perils of the MoB. The Challenges in the Use of Big Data in Causal Investigations

Roberto Leombruni* and Sonia Della Monica

University of Turin, Department of Economics and Statistics

Abstract

Month of birth (MoB) is often used as an instrumental variable (IV) in studying the causal patterns linking work careers and health. A recent literature, however, is questioning its validity: since parents may manipulate the timing of births there is a potential correlation e.g. with mothers' characteristics which may be also relevant for the outcome. In this paper we consider a related but different matter. The use of the MoB as an IV rests in its empirical association with several later outcomes, such as e.g. educational attainment and age at marriage. This implies that there may be multiple pathways going from the IV to the outcome, threatening the validity of a straightforward application of an IV estimator. We discuss the issue considering the relationship between work careers and health, exemplifying it with a Monte Carlo simulation and an application to Italian data.

Keywords

Month of Birth, Instrumental Variables, Work and Health relations.

1. Introduction

Besides its impact on businesses and the economy at large, the advent of Big Data is more and more characterising also the production of official statistics and academic research. Already a decade ago, in leading economic journals such as the American Economic Review and the Quarterly Journal of Economics the share of articles based on traditional statistical surveys was fading out in favor of the ones based on Big Data of administrative source [1, 2]. The high level of detail and the huge sample size they provide, in particular, are revealing particularly fit for the implementation of causal modelling, for instance in the case of regression discontinuity designs, where the identification strategy rests on the availability of a mass of individual closely around, e.g, a legal age requirement set by a policy. Another interesting setting is the case of instrumental variables estimators, which require a first stage in which a potentially endogenous explicative is modelled as a function of other characteristics exogenous to the model. Here the leverage provided by Big Data is twofold: along the "wide" side, the availability of high-dimensional data allows the implementation of ML predictive models in the first stage to more effectively control for bias, as in double or de-biased machine learning [3]. On the other side, instrumental variables are typically based on some "weird" and unexpected causal connection between the endogenous variable and a characteristic apparently unrelated with the phenomenon of interest. Here, the leverage of sample size and the accuracy of information is essential to turn these correlations into strong instruments to identify the causal relationship of interest.

An interesting example of the latter case is the diffusion of causal research exploiting detailed information on individuals' month of birth (MoB). Since the seminal paper by Angrist and Krueger [4] on the returns to education, the exact moment of the year when individuals are born has more and more been used as an instrumental variable in investigating also other causal patterns, thanks

2nd Workshop "New frontiers in Big Data and Artificial Intelligence" (BDAI 2025), May 29-30, 2025, Aosta, Italy

*Corresponding author.

roberto.leombruni@unito.it (R. Leombruni); sonia.dellamonica@unito.it (S. Della Monica)

to its empirical association with several later outcomes besides educational attainment, such as age at marriage and maternity [5, 6, 7], and work-career transitions [8]. A recent literature, however, is raising questions about its use, arguing about the validity (exogeneity) of the instrument. As an example, parents may actually manipulate the timing of birth for several reasons, creating a potential correlation e.g. with individual characteristics of the mother or the socio-economic position of the family which may in turn be relevant for the outcome of interest [9, 10].

In this paper we focus on a related but different matter, connected with the statistical power granted by large administrative datasets. The IV identification strategy, such as in Angrist and Krueger's paper, rests on the identification of some unexpected causal pathway going from the instrument (season of birth) to an endogenous variable (education) to the outcome of interest (wages). A potential paradox, here, is that the larger and more detailed are the data, the more probable is that other causal pathways may be revealed too. In other words, the more an instrumental variable is successfully associated with different characteristics, the more it is probable that there are other pathways going from the selected instrument to the outcome of interest, threatening the validity of IV also assuming no parenthood manipulation of the timing of birth. As an example, the season of birth may have a relation with contemporary factors such as climate conditions or air pollution, which may be important when health is a relevant dimension of the study. In this contribution we discuss the issue sketching various causal settings using Directed Acyclic Graphs (paragraph 2); illustrating the proper use of IV with a Monte Carlo simulation (paragraph 3); discussing the relevance of the issue considering the relationship between work careers and health in the case of the Italian context.

2. A graphical representation of the causal pathways

In Figure 1 we use two Directed Acyclic Graphs (DAGs) to represent Angrist and Krueger example (left panel) and the issue of parenthood manipulation (right panel). The interest is on the returns to education, where the identification issue is due to a confounding, unobserved factor (ability), which exerts an influence on both education and wage. An OLS regression of wage on education will overestimate the returns to education, since education – which is positively correlated with ability – will tend to capture also the positive effect of ability on the wage. In this case, however, since the MoB has no other direct or indirect connections with ability and wage, it can be used as an instrumental variable, and the returns to education may be identified regressing wage on E(education|MoB).

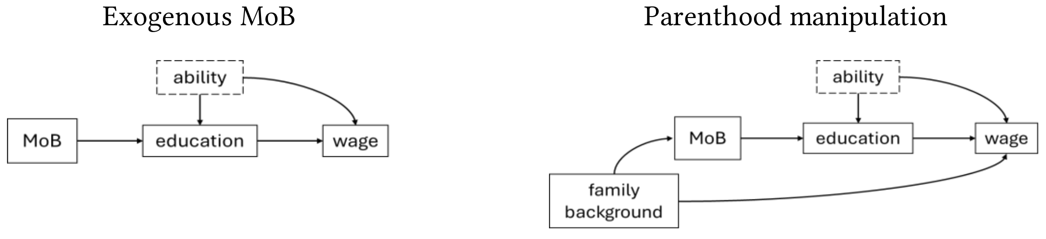


Figure 1: Directed Acyclic Graphs representing the causal pathways of Angrist and Krueger's 1991 paper on the estimation of the return of education using the season of birth as an instrumental variable (left panel) and the role of instrument manipulation (right panel). Dotted boxes indicate unobserved variables.

The issue with parenthood manipulation (right panel) is due to another confounding factor, the family background, which exerts an impact on both the MoB and wage. This issue, however, does not necessarily hamper the identification of the return to education. When family background is observable, it is sufficient to add it as a control in both stages of the IV estimator. When unobservable, one can exploit the fact that births' manipulation is not perfect, comparing those born in December of year *t* with those born in January of year *t+1*, i.e., individuals born in the same season but in different age cohorts.

In Figure 2 we represent the issue we are focusing on, i.e., the investigation of the relationship between career features and health. The identification issue here is reverse causality: working conditions entail many risk- and protective factors for health, and at the same time health exert an influence on individuals' work career. A classic example is the "healthy worker effect", where employed individuals are healthier than the general population – pointing to an apparent beneficial impact of work on health – but this is due to a selection into employment, since healthier people are more likely to enter and remain in the workforce.

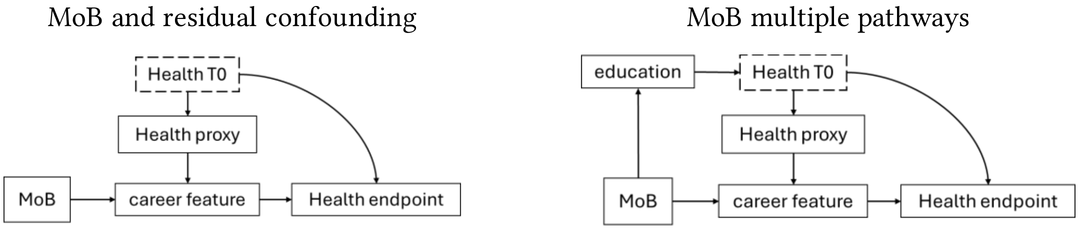


Figure 2: Directed Acyclic Graphs on the causal pathways between work careers and health with measurement error in health at the baseline. Left panel represents a situation in which MoB can be used to instrument the endogenous variable. Right panel represents the case in which there are two different pathways from the instrument to the outcome. Dotted boxes indicate unobserved variables.

In the left panel we represent a situation in which a researcher is interested in the impact of a work-career feature (e.g. work exposure) on a health endpoint (e.g. stroke), while controlling for general health at baseline (Health T0). The latter however is measured with error, and a proxy for general health is used. This strategy is prone to the residual confounding produced by the impact of the residual (unobserved) variability in health at baseline on both the work career and the health outcome. Here, the use of an instrumental variable such as the MoB may serve as a way of blocking the causal path from Health T0 to the work career and hence to identify the effect of interest regressing the health endpoint on E(career feature | MoB).

The right panel highlights the fact that the MoB has actually an influence also on the education level, which in turn has a causal connection with general health at the baseline. Note that in this DAG education level has no direct effect on the health endpoint, nor a direct effect on the career feature. Regarding the first point, the literature is largely unanimous in the presence of a causal connection between the level of education and general health. Here we simply do not assume a further, direct effect of education on the specific endpoint of interest, once controlling for health at the baseline. Regarding the second point, education has indeed many potential connections with several career features, but not necessarily all career events are causally linked to the level of education (see e.g. [12] on Job-hopping). What we are highlighting here is that, even in a highly simplified situation in which there are no direct links between education and work-career, there are two causal pathways going from MoB to the health endpoint, threatening the validity of regressing the outcome on E(career feature | MoB). In other words, even though education is technically not a confounder, its connection with the exogenous instrumental variable "restore" the residual confounding due to the imperfect measure of general health.

3. A Monte-Carlo illustration

In this section we illustrate the impact of multiple pathways in the DAG on the application of an IV strategy. We consider a data generating process closely mimicking the DAG in Figure 2, with the following steps:

- Month of birth is sampled from a discrete U[1, 12]
- Education is sampled from a discrete U[8, 24], with up to 2 years for those born later in the year
- General health is sampled from a U(0, 1), with up to 0.16 for individuals with higher education

- The health proxy adds to general health a classical measurement error sampled from a U(-0.3, 0.3)
- The career feature, interpreted as the age at retirement, is sampled from a discrete U[55, 68], with up to 2 years for healthier individuals and up to 2 years for those born earlier in the year
- The DGP of the health outcome is $Y = 100 + 100{*}\text{health\_T0} - \text{age\_at\_retirement} + \varepsilon$, with $\varepsilon \sim N(0, 20)$

In Table 1 we present the results of three different models. The first three columns report the Monte-Carlo coefficient estimates using an OLS model with various specification, averaged over 1,000 iterations with a sample size of 10,000 simulated individuals. Column (a) reports the unbiased average estimates obtained with full data, i.e. regressing the outcome on total work exposure and a correct measure of baseline health. Using the health proxy (column b) it is apparent the residual confounding due to the share of general health variability not captured by the proxy (-0.56 with respect to a true value of -1). Adding education to the specification, as a further control for general health at baseline, does not notably correct the bias (Column c).

Using the MoB as an instrument for total work exposure in order to control for endogeneity does not correct the bias, either; in our setting, it leads to an overestimation of the effect size (Column d). Bias correction is achieved only adding education as an instrument in the first stage of IV estimation (Column e).

Table 1
Average estimates of the causal impact of work exposure on health using OLS and IV, various specifications. Monte-Carlo simulation with N = 10.000, 1.000 iterations.

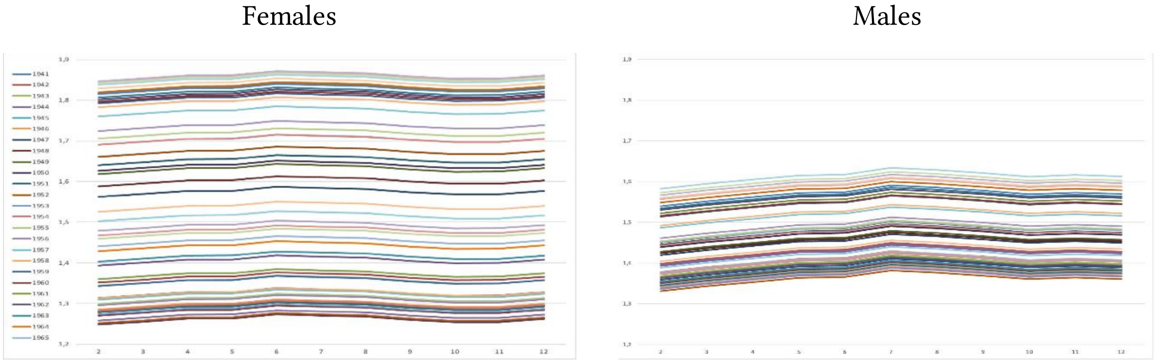
	OLS	OLS	OLS	IV	IV
	(a)	(b)	(c)	(d)	(e)
Work exposure	-1.00 *-1.13 -0.87*	-0.56 *-0.72 -0.40*	-0.55 *-0.71 -0.39*	-1.35 *-2.23 -0.48*	-0.99 *-1.82 -0.13*
Controls					
Health T0	✓				
Health proxy		✓	✓	✓	✓
Education			✓		
Instruments					
Month of Birth				✓	✓
Education					✓

4. Discussion

In the last decade there has been a surge in academic research and policy interest on social determinants of health and health inequalities, both internationally and in Italy. Due to the bi-directional nexus between work- and health biographies, however, the correct assessment of causal relationships is still an open question. IV estimators, one of the gold standards for causal inference, are increasingly used also in the specialized literature on our matter, thanks also to the availability of large datasets with very granular information on crucial individuals' and work career's characteristics.

In this contribution we considered an instrument – the month of birth – which due to its connection with many life course decisions may end up in many causal pathway from the instrument to the outcome. This situation, also in a very simplified setting where education does not exert any direct influence on the explicative of interest, may lead to highly biased estimates when the IV strategy is applied without properly controlling for education.

The relevance of the matter depends on the actual correlation between the instrument and the variables of interest. In the case of Italy, a link between MoB and various work careers features, among which labour market entry and age at retirement, has been already documented in [8]. As regards the connection between MoB and education, in Figure 3 we represent an original estimate using administrative data of the Ministry of Welfare. The figure plots the partial correlation between MoB and the average education level (average ISCED level, 4 categories), as estimated with a linear regression model for individuals born in the years from 1940 to 1995 stratified by gender. Birth cohoorts differences are controlled for with a set of dummies (reference category is 1940), so that the plots represent the average effect of eleven MoBs on education level (reference category is January). We find a statistically significant increase in educational attainment for all months up to july, coherently with the seminal idea by A&K, which is particularly strong for males, less strong and less statistically significant for women.



Females

Males

Figure 3: Linear regression model of 4-level ISCED education attainment on MoB dummies (reference: January), controlling for birth cohorts (reference: 1940), stratified by gender. **Data**: Work and Health Italian Panel, cohorts from 1940 to 1995, workers with a job transitions in the years 2008 to 2019.

The third "ingredient" of the DAG of Figure 2, the measurement error on health at baseline, is also relevant for Italian studies, since a thorough representation of individuals' health is not present in currently available socio-economic data on work- and health biograpies. As a consequence, to properly investigate causal links between work exposures and health outcomes using MoB as an IV, the level of education – even in an ideal situation in which it does not play any role for the work dimension of interest – should be included in the first- and second-stage of an IV estimation. Its exclusion could induce a confounding through the causal pathway from MoB to education to health.

The general point regards the fact that the availability of high frequency and highly granular socio-economic and spatial data allows the detection of several statistically significant correlations between apparently unconnected phenomena. From the one side this is allowing a broader application of causal modelling using an instrumental variable strategy, but exposes also to the risk of multiple causal chains going from the instrument to the outcome. This is a point which has recently been posed also by Mellon [13] in the case of climate conditions, who proposed a strategy to test the sensitivity of the estimates to possible violations of the exclusion restrictions. In the case here exemplified, we tested a direct solution of the issue using DAGs and a Monte-Carlo simulation of it. The somewhat counter-intuitive result is that, in our simplified illustration, the level of education – which is neither an explicative of the outcome, nor a mediator, nor a confounder – has to be controlled for in the first stage of the estimator to achieve bias correction. As exemplified with the DAG, this is necessary in order to block a secondary pathway going from the month of birth to the health outcome, in line with Brito & Pearl conditions for identification using DAGs [14].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. Chetty, Time trends in the use of administrative data for empirical research. 34th Annual NBER Summer Institute. Cambridge, Mass., July 9–27, 2012.

[2] R. Leombruni, Interviewing administrative records. A conceptual map for the use of big data for economic research. Italian Journal of Applied Statistics, 36(3), 2024, 295–326.

[3] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, Volume 21, Issue 1, 1 February 2018, C1–C68.

[4] J. D. Angrist, A. B. Krueger, Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics, 106(4), 1991, 979–1014.

[5] V. Skirbekk, H. P. Kohler, A. Prskawetz, Birth month, school graduation, and the timing of births and marriages. Demography 41, 2004, 547–568.

[6] L. Cavalli, Does the Month of Birth Influence the Timing of Life Course Decisions? Evidence from a Natural Experiment in Italy. Open Journal of Social Sciences, 2, 2014, 101-118.

[7] M. G. Kirdar, M. Dayioğlu, İ. Koç, The effects of compulsory-schooling laws on teenage marriage and births in Turkey. Journal of Human Capital, 12.4, 2018, 640-668.

[8] C. Ardito, R. Leombruni, D. Blane, and A. d'Errico, To Work or Not to Work? The Effect of Higher Pension Age on Cardiovascular Health. Industrial Relations, 59, 2020, 399-434.

[9] K. S. Buckles, D. M. Hungerman, Season of Birth and Later Outcomes: Old Questions, New Answers. The Review of Economics and Statistics, 95(3), 2013, 711–724.

[10] H. Torun, S. Tumen. The empirical content of season-of-birth effects: An investigation with Turkish data. Demographic Research 37, 2017, 1825–60.

[11] E. Fan, J.-T. Liu, e Y.-C. Chen. Is the Quarter of Birth Endogenous? New Evidence from Taiwan, the US, and Indonesia. Oxford Bulletin of Economics and Statistics, 79(6), 2017, 1087–1124.

[12] K. Steenackers, M. A. Guerry, Determinants of job-hopping: an empirical study in Belgium. International Journal of Manpower, 37(3), 2016, 494-510.

[13] J. Mellon, Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable. American Journal of Political Science, 2024, 1–18.

[14] C. Brito, J. Pearl. Generalized Instrumental Variables. arXiv, 12 december 2012.