Selection of instrumental variables for casual interference study: the theorem and application

Yueting Wu[1, *], Guanghua Ren[2, *]

[1]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, Shanghai, 201100, China

*Corresponding author's e-mail: guanghua.ren@gecacdemy.cn

Keywords: causal inference, instrumental variable, prescribing preference, gender, SNP

Abstract: Causal inference is the process of determining the actual, independent effects of a given phenomenon (cause) within a larger system, which is getting more and more attention in the area of sociology, economics and medicine. Judea Pearl said that causal inference is the foundation of scientific research. However, under the framework of counterfactual causality, it is extremely difficult to make causal inferences based on quantitative analysis of survey data on account of endogeneity. Confounders affect both cause and effect, which leads us to draw false conclusions. Fortunately, instrumental variables provide us with a solution. When we find an instrumental variable that only affects the cause but is independent of confounders, we can use this variable to make causal inferences. The choice of instrumental variables determines the validity of causal inference, so wise choice is very important. In this review, we summarize three widely used instrumental variables. Their applications have been analyzed and explained in detail. Their use reveals the most appealing characteristic of instrumental variables: all of them seem to be irrelevant to what we try to explore, but they play a crucial role in finding out the causality among objects of the study. So far, few articles have summarized the three instrumental variables, which has made this review contributory.

1. Introduction

Under the framework of counterfactual causality, it is extremely difficult to make causal inferences based on quantitative analysis of survey data. The main reason is that researchers face an eternal challenge when they want to prove that a cause they are interested in has an effect: endogeneity. It means that if potential and unobserved disturbance effect "cause" and "effect" at the same time, the estimators obtained by regression analysis using the least square model (OLS) will be biased and not infer causality properly.

Instrumental variables (IV) were first put forward by Philip G Wright[1]. The simplest linear regression model could introduce the basic understanding of IV.

$$y = \beta_0 + \beta_1 x_1 + \beta X + \varepsilon \quad (1)$$

Here y is the dependent variable(effect); x_1 is the independent variable, namely the explanatory variable(cause). X is the exogenous control vector, and ε is the error term. If ε is independent of x_1, we can use the OLS model to objectively estimate the equation. However, if another unobserved variable x_2 is omitted from the model (1) and x_1 and x_2 are also correlated, then the OLS estimate for β_1 is biased, and x_2 is called "cofounder". At this point, x_1 is referred to as the "endogenous" explanatory variable. To address this problem, the instrumental variable Z should be introduced. This variable needs to be not only relevant with the endogenous explanatory variable but also irrelevant to the error term, that is, the instrumental variable is strictly "exogenous". In another word, Z affects y only by affecting x_1. Based on the prerequisites of the tool variable and the exogenous feature of X, we can know that:

$$\mathrm{Cov}(Z, x_1) \neq 0; \mathrm{Cov}(Z, \varepsilon) = 0; \mathrm{Cov}(Z, X) = 0 \quad (2)$$

From equation (1) and (2) We can derive that

$$\mathrm{Cov}(\mathrm{Z},\mathrm{y}) = \beta_1 \mathrm{Cov}(\mathrm{Z},\mathrm{x}_1) \tag{3}$$

Therefore, we can carry out an unbiased estimation of β_1:

$$\hat{\beta}_1 = \frac{\sum(Z_i - \bar{Z})(y_i - \bar{y})}{\sum(Z_i - \bar{Z})(x_{1i} - x_1)} \tag{4}$$

In this way, by introducing an instrumental variable Z, we can derive the true relationship between x and y. As long as the effect of Z on y is significant, we can infer that y and x_1 has a causal relationship. Figure 1. provides a clearer idea of the role of the instrumental variable.

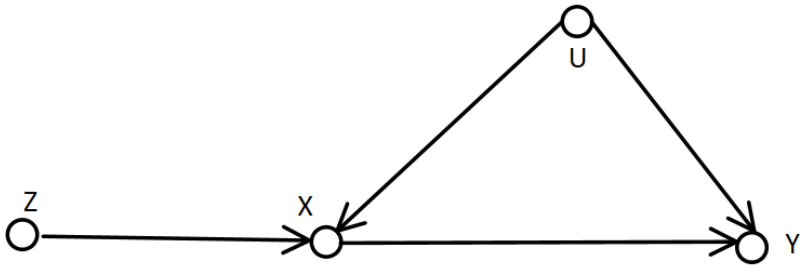

Figure 1. Diagram of the Causal Relationships Using Instrumental Variable Analysis. X is U is the cofounders, which is irrelevant to instrumental variable Z. In this case, the instrumental variable Z can only affect the dependent variable Y indirectly by affecting the independent variable X. If the instrumental variable Z is closely related to the independent variable X, then any incremental change in the instrumental variable Z will inevitably produce an impact on the independent variable X from outside the model.

2. Applications of instrumental variables for casual inference Study

2.1. The use of gender variable

In sociological or biological research, gender is a very interesting instrumental variable. When a group is selected, gender is a random variable for the individual. At the same time, some factors are affected by gender, such as preferences. Still, its randomness has nothing to do with other objective variables, such as age and living area, which often cause bias.

Many studies focusing on the effects of children on mothers always use gender as an instrumental variable. Children have a direct biological relationship with the mother. For example, the number of pregnancies can change the mother's physical function. Children also influence the mother's lifestyle, which is quite obvious for the existence of a baby will cost some time of its mother. Here are some cases which use gender cleverly. As have been reported by T van den Broek et al. [2], the gender of the two oldest children was used as instrumental variables. In their study, the number of children was regarded as a factor leading to mother's obesity. It has been considered that there is a potential biological link between the number of children and a woman's weight. The pregnant woman's progesterone is elevated during pregnancy, leading to the accumulation of body fat, which may be retained after pregnancy; Having more children also gives mothers more responsibility, resulting in fewer opportunities to exercise. But the link between the number of children and a woman's weight could have confounding factors, such as age (mothers with more children are older and therefore more likely to be obese) and so on. In this case, an instrumental variable independent of confounding factors is needed. Since it's more likely that mothers would like to have a third child when having two children of the same sex, the sex of the two oldest children and the number of children is relevant.

Meanwhile, the gender of the first two children has no relationship with mothers' objective characteristics such as age, which indicates that gender is a wise choice as an instrumental variable. In another study, F Teufel et al. used identical models to extrapolate the effects of childbirth and child-rearing on a mother's blood pressure in India. They used the sex of their first-born child as an instrumental variable. In countries such as India, where boys are preferred, they will be motivated to continue the pregnancy if women can't give birth to a boy. Thus, the sex of the first child can be a very reasonable instrumental variable when researches are done in the social context of India.

Gender also has a certain influence on personality. Women are comparatively more empathetic than men and more likely to care for children due to maternity. As reported, mothers allocate more resources to improve their children's survival chances than fathers do [4]. Furtherly, different ratios of male to female will affect social policies' tendency. One example is that female politicians tend to choose policies providing more benefits and public goods for children [5]. As have been reported by D Güvercin [6], the relationship between women's political participation and child labor is analyzed. The OLS regression of child labor on the female seat share in the national parliament is: $\text{Child labor} = \alpha + \beta\ \text{female set share} + \gamma X + \varepsilon$. "Child labor" represents the fraction of child labor, "female seat share" represents the seat share of female representatives in the national parliament, X represents other explanatory variables, and ε represents the random error term. This formula is expected to describe the influence of female set share to the fraction of child labor. However, due to the limited data, some variables such as changes in policies, economy and people's value can't be observed or evaluated sufficiently but can affect both female seat share and the fraction of child labor. This will cause bias when estimating the effect. Therefore, the instrumental variable is needed. The interaction between gender electoral quotas and gender index is used as an instrumental variable. The gender index is designed to represent the intensity of linguistic gender marking. Females with a high gender index have certain language habits, like preferring to use exaggerated, emphatic words. It's reported that these people are more likely to be a relatively vulnerable group in society, enjoying higher gender electoral quotas in politics. Therefore, the interaction of the gender index and gender quota can show gaps between gender, which affects female set share. If this interaction influences the fraction of child labor, female set share can also be regarded to affect it.

So far, gender-related instrumental variables have been widely used, which provides some ideas for future research.

2.2. Prescribing preference

Instrumental variable analysis is suggested as a possible alternative to traditional analysis when unmeasured confounding effects are present in observational studies. For example, physician prescription preference is often used as an instrumental variable in assessing the effects of drugs. Since it is associated with the patient's treatment but not with (or only weakly associated with) unobservable patient risk factors such as their BMI values, it provides a good balance of measured patient characteristics and remains consistently strong across the time.

First, M A Brookhart et al., prescription preference was first used as an instrumental variable when comparing the effect of exposure to COX-2 inhibitors with non-selective, non-steroidal anti-inflammatory medications on gastrointestinal complications [7](Figure 2). A tough nut of using physician preference is that it will change with drug companies' marketing campaigns and new information about drug safety and effectiveness and physicians' own evolving clinical experience. On the other hand, the dynamic aspect of preference was particularly relevant with COX-2 inhibitors, which were aggressively marketed, adopted quickly by some physicians and not so quickly by others, and then affected by various safety issues that might impede their use. All of these factors increase the difficulty of measuring it. In this case, it is proposed to estimate a physician's current preference for COX-2 inhibitors over non-selective NSAIDs by using the physician's last NSAIDS. Under this approach, if a physician's last new NSAID prescription was for a COX-2 inhibitor, that physician is classified as a "COX-2 prescriber" for the next patient. Otherwise, he would be classified as a non-selective prescriber of NSAIDS. Based on the work of Brookhart et al. [7], J A Rassen et al. evaluated the strength of instrumental variables and the reduction of imbalance resulting from the instrumental

variable's application. The advantage of using prior patient's treatment to estimate preferences is that any changes in preferences can be recorded quickly. Still, two problems arise: first, the prior patient's treatment may not reflect the physician's true preferences, and second, the simple IV prescribed may not have the strength and effectiveness required. Therefore, to better represent physicians' prescribing preferences, J A Rassen et al. did three things. Firstly, they extended the length of time using the four most recent prescriptions. The similarity of the four prescriptions created a more stable estimate of preference. Secondly, physicians' skills and patients' expectations also lead to misestimates of physicians' prescribing preferences. So they classify physicians by rank (primary care, speciality, year of graduation, and classify patients by cofounders (age, age relative to the average in the physician's practice). Subgroups that hold similar IV assumptions are expected to be isolated, thus improving efficiency and balance. Finally, prescribed patients were stratified to better share the main characteristics such as age or gender. By this rearrangement, they hoped that the treatment given to previous patients reflects overall preferences and preferences within a particular patient subgroup. These three measures help to make the right use of prescribing preference as an instrumental variable.

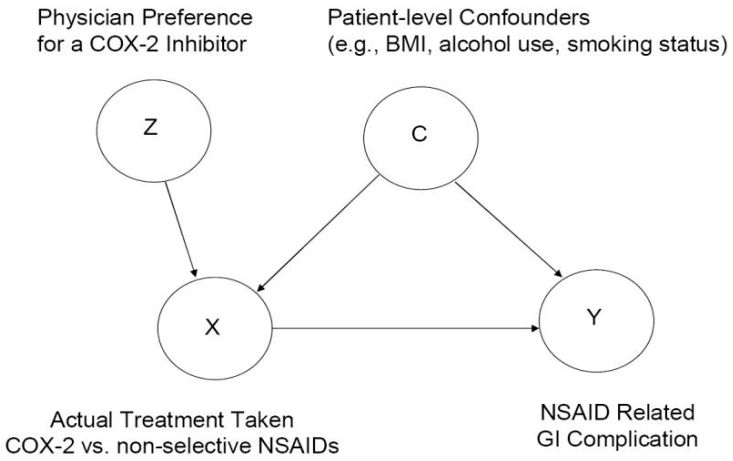

Figure 2. Diagram of the Causal Relationships Using Instrumental Variable Analysis in M A Brookhart et al.'s research[7]: Physician Preference is the instrumental variable which affect the actual treatment but has no relationship with patient-level confounders

Physician's prescribing preferences are increasingly being used as an instrumental variable in subsequent therapeutic efficacy studies, such as antipsychotics on death in the elderly [9]. However, differences in prescribing patterns among physicians may reflect differences in preferences or case combinations. In addition, the possible assumption of using physician preferences as a tool for point estimation is controversial. Although, some studies have discussed the applicability of prescription preferences [10], the conclusion is that deterministic monotony is often unreasonable in favor of the physician as an instrumental variable. Depending on the definition of the instrumental variable, random monotonicity may be reasonable.

2.3. Single-nucleotide polymorphism (SNP)

In alcohol-related studies, researchers had to rely on participants' own reports, which required participants to assess their own levels of alcohol consumption. This means that there is potential for reverse causality and confusion. Furthermore, it is difficult to separate the effects of alcohol consumption from confounding factors, for drinking may be caused by simple love or complex social factors. In this case, instrumental variables are a good way to solve the problem. In many IV analyses, genetic variants are used as proxies for exposure status (Figure 3). Unlike the risk factors of interest,

the genetic variations were randomly assigned at conception and thus were not associated with potential confounding factors.

In recent years, many studies have focused on the harmful effects of alcohol consumption. Most of the studies on this used a single functional single-nucleotide polymorphism(SNP) as a genetic instrument. The other studies have used combinations of multiple SNPs as instrumental variables [11]. The selection of SNP is the key to causal analysis. For alcohol, functional variations in genes encoding alcohol dehydrogenase (ADH) were associated with alcohol intake: people with fast alcohol degrading alleles *ADH1B*2* and *ADH1C*1* always consumed less alcohol than those with slow alcohol degrading alleles *ADH1B*1* and *ADH1C*2*. To our delightment, no pleiotropic effect was found for the alcohol dehydrogenase *ADH1B* and *ADH1C* genotypes, which means that these genotypes do not directly cause illness. In conclusion, the ADH genotype is a valid candidate for an unbiased tool for lifelong drinking.

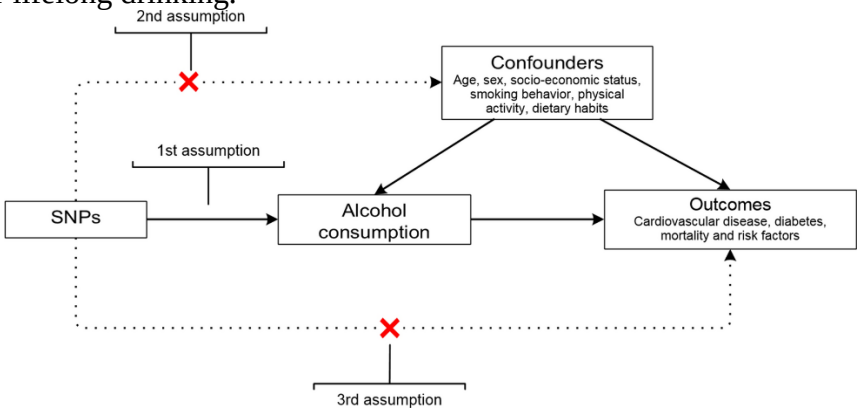

Figure 3. Diagram of the Causal Relations by Using SNPs as Instrumental Variables. To use SNPs as instrumental variables, it needs to satisfy three assumptions. 1st assumption: SNPs affect one's alcohol consumption. 2ed assumption: SNPs are irrelevant to confounders, 3rd assumption: SNPs have no direct relation with outcome.

Many studies look into the detail of alcohol consumption using SNPs as instrumental variables. For example, A I Christensen et al.[12] paid attention to a report saying that light to moderate alcohol consumption was associated with reduced cardiovascular risk compared to non-drinkers. However, the apparent cardio-protective effect associated with light to moderate alcohol consumption could be explained by lifestyle: people consuming light to moderate alcohol prefer a healthy lifestyle and pay more attention to control their desire to drink. In the absence of viable randomized trials to confirm or disprove the cardio-protective effects of light to moderate drinking, they used genetic variations to proxy for alcohol consumption. This approach avoids some limitations of observational studies because the distribution of genetic variations is random in terms of potential confounders, and genotypes cannot be affected by illness. In their research, a non-synonymous single nucleotide polymorphism (rs1229984) in the alcohol dehydrogenase 1B gene (*ADH1B*) encodes the ADH1B enzyme and provides a major pathway for alcohol metabolism, is associated with lower alcohol dependence in adult drinkers and adolescents. Therefore, rs1229984 is selected as an instrumental variable to investigate the role of alcohol in high blood pressure and various cancers. The findings suggest that even for light to moderate drinkers, reducing alcohol consumption can benefit cardiovascular health.

Other alleles of ADH have also been studied, such as using variants in *ADH1B* and *ADH1C* to estimate the causal effect of long-term alcohol consumption on BMI, SBP, DBP, HDLc, non-HDLc, triglycerides, fibrinogen, and glucose[13] and using five variants located in *ADH1B*, *ADH1C* and *ADH4* genes as genetic tools and combined into unweighted genetic scores[14].

3. Conclusions

The selection of instrumental variables is an art. This review shows the selection of instrumental variables from the individual level to the gene level from this review. The choice of these three instrumental variables is representative and intelligent. Essentially, an instrumental variable is a variable found outside the model's scope that is related to an explanatory variable. A good tool variable can greatly simplify our problem. So far, the research on instrumental variables has been relatively mature. Appropriate instrumental variables can be found in all areas. What we need to do is to find a better way to use instrumental variables based on predecessors.

Currently, there is a lack of evaluation criteria for the use of instrumental variables, which is expected to occur in the future.

References

[1] Stock, J. H., & Trebbi, F. (2003). Retrospectives: who invented instrumental variable regression?. Journal of Economic Perspectives, 17(3), 177-194.

[2] van den Broek, T., & Fleischmann, M. (2021). The causal effect of number of children on later-life overweight and obesity in parous women An instrumental variable study. Preventive Medicine Reports, 101528.

[3] Teufel, F., Geldsetzer, P., Sudharsanan, N., Subramanyam, M., Yapa, H. M., De Neve, J. W., ... & Bärnighausen, T. (2021). The effect of bearing and rearing a child on blood pressure: a nationally representative instrumental variable analysis of 444611 mothers in India. International Journal of Epidemiology.

[4] Benería, L., Berik, G., & Floro, M. S. (2015). Gender, development, and globalization: Economics as if all people mattered. Routledge.

[5] Miller, G. (2008). Women's suffrage, political responsiveness, and child survival in American history. The Quarterly Journal of Economics, 123(3), 1287-1327.

[6] Güvercin, D. (2020). Women in politics and child labor: An instrumental variable approach. The European Journal of Development Research, 32(4), 873-888.

[7] Brookhart, M. A., Wang, P., Solomon, D. H., & Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology (Cambridge, Mass.), 17(3), 268.

[8] Rassen, J. A., Brookhart, M. A., Glynn, R. J., Mittleman, M. A., & Schneeweiss, S. (2009). Instrumental variables II: instrumental variable application—in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. Journal of clinical epidemiology, 62(12), 1233-1241.

[9] Pratt, N., Roughead, E. E., Ryan, P., & Salter, A. (2010). Antipsychotics and the risk of death in the elderly: an instrumental variable analysis using two preference based instruments. Pharmacoepidemiology and drug safety, 19(7), 699-707.

[10] Boef, A. G., le Cessie, S., Dekkers, O. M., Frey, P., Kearney, P. M., Kerse, N., ... & den Elzen, W. P. (2016). Physician's prescribing preference as an instrumental variable. Epidemiology, 27(2), 276-283.

[11] van de Luitgaarden, I. A., van Oort, S., Bouman, E. J., Schoonmade, L. J., Schrieks, I. C., Grobbee, D. E., ... & Beulens, J. W. (2021). Alcohol consumption in relation to cardiovascular diseases and mortality: a systematic review of Mendelian randomization studies. European journal of epidemiology, 1-15.

[12] Christensen, A. I., Nordestgaard, B. G., & Tolstrup, J. S. (2018). Alcohol intake and risk of ischemic and haemorrhagic stroke: results from a Mendelian randomisation study. Journal of stroke, 20(2), 218.

[13] Lawlor, D. A., Nordestgaard, B. G., Benn, M., Zuccolo, L., Tybjaerg-Hansen, A., & Davey Smith, G. (2013). Exploring causal associations between alcohol and coronary heart disease risk factors: findings from a Mendelian randomization study in the Copenhagen General Population Study. European heart journal, 34(32), 2519-2528.

[14] Silverwood, R. J., Holmes, M. V., Dale, C. E., Lawlor, D. A., Whittaker, J. C., Smith, G. D., ... & Dudbridge, F. (2014). Testing for non-linear causal effects using a binary genotype in a Mendelian randomization study: application to alcohol and cardiovascular traits. International journal of epidemiology, 43(6), 1781-1790.