Instrumental variable analysis in the context of dichotomous outcome and exposure with a numerical experiment in pharmacoepidemiology

Babagnidé François Koladjo, Sylvie Escolano, Pascale Tubert-Bitter

Koladjo *et al. BMC Medical Research Methodology* (2018) 18:61
https://doi.org/10.1186/s12874-018-0513-y

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Instrumental variable analysis in the context of dichotomous outcome and exposure with a numerical experiment in pharmacoepidemiology

Babagnidé François Koladjo*, Sylvie Escolano and Pascale Tubert-Bitter

Abstract

Background: In pharmacoepidemiology, the prescription preference-based instrumental variables (IV) are often used with linear models to solve the endogeneity due to unobserved confounders even when the outcome and the endogenous treatment are dichotomous variables. Using this instrumental variable, we proceed by Monte-Carlo simulations to compare the IV-based generalized method of moment (IV-GMM) and the two-stage residual inclusion (2SRI) method in this context.

Methods: We established the formula allowing us to compute the instrument's strength and the confounding level in the context of logistic regression models. We then varied the instrument's strength and the confounding level to cover a large range of scenarios in the simulation study. We also explore two prescription preference-based instruments.

Results: We found that the 2SRI is less biased than the other methods and yields satisfactory confidence intervals. The proportion of previous patients of the same physician who were prescribed the treatment of interest displayed a good performance as a proxy of the physician's preference instrument.

Conclusions: This work shows that when analysing real data with dichotomous outcome and exposure, appropriate 2SRI estimation could be used in presence of unmeasured confounding.

Keywords: Instrumental variable, Nonlinear least squares, Logistic regression, Physician's prescription preference, Pharmacoepidemiology, Observational studies, Simulation study

Background

In observational studies, unobserved confounding may biase the estimation of target effect. Over the last decade, this issue has recieved a growing attention in the field of epidemiological studies attempting to assess adverse effects of drugs with a few works focusing on instrumental variable (IV) approaches. Instrumental variable estimation is a well known approach for assessing endogeneity in statistical modelling [1, 2]. Endogeneity often arises when a causal model is poorly specified therby introducing a structural bias in the estimation of its parameters. This may result from a measurement error in variables [3], an unobserved variable [4] or an inverse causality between the outcome and some regressors. The general IV method of estimation attempts to remove this bias by using structural equations which incorporate instrumental variables in the model. Several theoretical approaches have been developed to build estimators of parameters and to study the properties of these estimators in causal models with endogeneity (see [5]). A well-known example is the case of linear models in which IV estimation leads to estimators with good properties of convergence such as consistency discussed in [6]. The structural bias can be completely removed in this case.

In pharmacoepidemiology, we most often deal with binary covariables (drug exposure), binary responses (adverse event indicator) and confounding variables

*Correspondence: francois.koladjo@gmail.com
[1]Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, 16 Avenue Paul Vaillant-Couturier, 94807 Villejuif, France

which are variables that are correlated with exposure and response. A nonlinear model should be the first choice in this context to match with the specific nature of the variables. However, to quantify the risk of the adverse effect of a treatment in the presence of unobserved confounding, researchers investigating the IV-based estimation often model the probability of dichotomous events as a linear function of covariables, thus ignoring the basic features of a probability. Terza and colleagues [7] investigated the influences of mispecification on the estimation when a linear IV model is used in an inherently nonlinear regression setting with endogeneity. A substantial bias was demonstrated in their results. In the context of pharmacoepidemiology modelling, endogeneity is often due to unobserved confounding and various nonlinear IV methods such as the Generalized Method of Moment (GMM;[8, 9]) or the two-stage residual inclusion (2SRI; [10]) can be used to solve this issue. However in a review of IV methods, Klungel and colleagues [11] claimed that the GMM estimator with the logistic regression model is not consistent for causal Odd Ratio (OR) estimation owing to the non-collapsibility of the OR. Consistency is also not guaranteed for the 2SRI when the regression models are nonlinear in both stages. For these nonlinear IV methods, theoretical results exist under very restricted assumptions which do not cover the possible frameworks of real data. Overall, in the context of binary outcome several simulation studies investigate mainly 2-stage IV methods with the first step being linear and the second step being logistic as in [12]. A few articles concern double probit models (Chapman et al. [13]). Very few address the comparison of GMM and 2-stage approaches and none study GMM, 2-stage double logistic using the prescription preference-based instrumental variables.

In a simulation study, we compare the IV-based GMM and the 2SRI methods to the conventional method which does not account for endogeneity. These comparisons are based on the estimation of the regression coefficient of the exposure varible in nonlinear logistic model. Our numerical comparison of the methods involves several scenarios with different confounding levels and different instrument strengths for which computation formulas are established in the context of dichotomous outcome and exposure. We recall the general formula of the covariance matrix for the two-step estimation methods and give the corresponding expression for two-stage nonlinear least squares method in the context of logistic regressions.

The paper is organized as follows: we specify the model and describe the methods of estimation that will be analysed. Then we describe the simulation design, the criteria for evaluating the performances of the methods and the results of our simulations. The final sections discuss the results and make some concluding remarks. Details on the computation of the covariance matrix of the 2SRI method, the instrument's strength and the confounding level are to be found in the appendices, as well as a detailed description of the simulation model and supplementary results.

Methods

Model

We consider a general model with dichotomous outcome and exposure that can be written as

$$Y = F(\beta_0 + T\beta_t + X_1\beta_1 + X_2\beta_2 + X_u\beta_u) + e \quad (1)$$

$$\mathbb{E}(T|Z, X_1, X_2, X_u) = r(\alpha_0 + Z\alpha_z + X_1\alpha_1 + X_2\alpha_2 + X_u) \quad (2)$$

with Y and T as binary outcome (event or not) and treatment (T_1 or T_2) respectively, X_1, X_2 some covariables and X_u an unobserved confounder of the outcome and treatment. The function $F(.) = r(.)$ denotes the logistic distribution function also known as $expit(.)$: $expit(.) = exp(.)/(1+exp(.))$ and e denotes the error term. The parameter $\beta = (\beta_0, \beta_t, \beta_1, \beta_2)$ denotes the vector of unknown parameters to be estimated. Without a confounding variable, all observed regressors are exogenous. In this case, the true model is written

$$Y = F(\beta_0 + T\beta_t + X_1\beta_1 + X_2\beta_2) + e \quad (3)$$

and conventional regression methods are suitable for estimating the parameters β. We will denote (3) *the conventional model*. If this model is adjusted to data in the presence of an unobserved confounder, the estimated coefficients would lead to a bias with a level depending on the confounding level. As the confounder X_u is not independent of treatment, the residuals of the conventional model are associated with the treatment. This causes endogeneity so a single regression of the outcome on observed covariables will fail to estimate β_t efficiently. A common strategy is to consider another regression model that links the endogenous variable with others. Equation (2) defines the auxiliary model that predicts treatment T as a function of covariables X_1, X_2, the confounder X_u and another variable Z. Variable Z denotes the instrumental variable (or instrument) related to the treatment, i.e. a variable correlated with the treatment and which has no direct association with the outcome.

The bias due to the unobserved confounder can significantly be reduced by means of the two-stage regression model using a valid instrument. As defined by Johnston and colleagues [14] and Greenland [15], a valid instrument must not be correlated with an unobserved confounder or with the error term in the true model (1). Formally, we assume that the instrument Z meets the following assumptions:

- $\mathbb{C}ov(Z, Y|T, X_u, X_1, X_2) = 0$,
- $\mathbb{C}ov(Z, T) \neq 0$,
- $\mathbb{C}ov(Z, X_u) = 0$, $\mathbb{C}ov(Z, X_1) = 0$ and $\mathbb{C}ov(Z, X_2) = 0$.

We also assume that the confounder X_u is not associated with the covariables X_1 and X_2, that is $\mathbb{C}ov(X_u, X_1) = 0$ and $\mathbb{C}ov(X_u, X_2) = 0$. The main goal is to estimate β_t, the treatment coefficient which is the basis of risk evaluation; however an estimation of β_t is obtained in general by estimating vector β which is discussed below.

As already proved in the simple case of a linear model (for which the functions F and r are equal to identity in Eqs. (1) and (2)), a high association between the treatment and an instrument should improve the IV estimation of β. Finding a strong instrument is then a crucial step in all procedures of instrumental variable estimation.

In what follows, we first present some specific instrumental variables often used in pharmacoepidemiology, then discuss some IV estimators of β to obtain an estimate of β_t before addressing the properties of these estimators.

Instrument in pharmacoepidemiology

An instrumental variable can be determined in many ways, provided that it meets the assumptions listed above. One of the problems is to find a valid instrument with a reasonable strength. The strength of an instrumental variable can be defined as resulting from the level of its association with the endogenous treatment. As such, it could be quantified by using the correlation coefficient between the treatment and the related instrument. In the wide range of pharmacoepidemiologic applications, we can summarize the various instrumental variables in three categories:

Geographical variation. Proximity to the care provider can positively influence access to treatment of a patient compared to others who live far away from health services. To account for this difference between patients, some researchers (see [16]) consider the distance between a patient and a care provider as an instrumental variable. Although this seems realistic as there is no direct association between this distance and the occurrence of disease, the presence or absence of health services can be associated with some socioeconomic characteristics. The latter are often considered as unmeasured confounders that call into question the suitability of using this instrument.

Calendar time. The use of calendar time as an instrument in pharmacoepidemiology often relies on the occurrence of an event that could change the attitude of the physician or patient regarding a treatment. This could be a change in guidelines for example or a change due to the arrival of a new drug on the market. The time from that event to the date of treatment defines the calendar time which clearly affects the outcome of the treatment since the change in physician or patient attitude will be more pronounced immediately after the event has occurred than later. An example of use of calendar time as an instrument can be found in [17].

Physician's prescription preference. The most often used instrumental variables in pharmacoepidemiology are preference-based [18–20]. The issue is to compare the effectiveness of two treatments T_1 and T_2 when the assignment of treatment to the patient is not randomized. This is the case in observational studies where the prescriber of the treatment (the physician) introduces an effect that influences the outcome via the prescribed treatment. This effect results in the instrumental variable that reflects the influence of care-providers on the patient-treatment relationship. Brookhart and colleagues [21] define this instrumental variable as the "Physician's Preference" (PP) and propose to use the treatment prescribed by a physician to its previous patient as a proxy of this IV for his/her new patient.

The instrumental variable $\left(Z_i^*\right)$ of the i^{th} patient will then be the treatment prescribed to the previous patient of the same physician. As a physician's preference could change over time, Abrahamowicz and colleagues [22] introduce a new procedure to detect the change point and build a new proxy of PP that includes not only the treatment prescribed to the previous patient but all the previous prescriptions since the change point.

Those instrumental variables and some others are presented in a more detailed form in [23], [24, 25] or in [26] with enlightening discussion on their validity. In this work, we carry out a simulation study to examine how the proxy Z^* of physician's preference performs in the context of logistic regression. We also examine the physician's preference-based IV in the continuous form (see [22]), i.e. the proportion (pr) of all previous patients of the same physician who were prescribed the treatment of interest. This corresponds to the empirical estimator of the probability for a physician to prefer the treatment of interest.

IV Estimation of β_t

Estimating β_t in model (1) with an unobserved confounder X_u amounts to estimating vector β and taking the corresponding component of treatment T. Below, we present two methods that can provide consistent estimation of β and then that of β_t: the two-stage residual inclusion (2SRI) method and the generalized method of moment (GMM).

The two-stage residual inclusion method is a modified version of the two-stage least squares (2SLS) method used to estimate the parameters in linear models with instrumental variables. As mentioned by Greene [5], the first stage of the 2SLS method predicts the endogenous variable (the treatment here) using the instruments and other covariables (Eq. (2)). In the second stage, the endogenous

variable is just replaced by its prediction from the first stage. This method is called two-stage predictor substitution (2SPS) when the first stage is nonlinear. Unlike the 2SLS, the residuals of the first regression serve as a regressor in the second stage of 2SRI method. This method also generalizes to the nonlinear models i.e. when first and second stages are nonlinear. The rationale of this approach can be intrinsically related to the form of the true model: sometimes, the prediction equation of the outcome includes the error term of the auxiliary regression. An example is the case when the confounder is the only source of error in the auxiliary regression as considered in [27]. In a linear model, both 2SPS and 2SRI approaches are equivalent.

The GMM is an alternative method for obtaining a reliable estimator of parameter β in a model with an endogenous variable. It is based on the classical assumption

$$\mathbb{E}(e|T, X_1, X_2, X_u) = 0 \tag{4}$$

of the error term in Eq. (1). This assumption does not hold in general because the confounder X_u is unobserved (i.e. $\mathbb{E}(e|T, X_1, X_2) \neq 0$). In this case, the treatment is endogenous and its coefficient β_t cannot be consistently estimated. One suppose in general that there exists some observable instruments w_1 such that $\mathbb{E}(e|w) = 0$ with $w = (w_1, X_1, X_2)$. Typically, w corresponds to the vector of exogenous and endogenous variables with endogenous regressors replaced by their corresponding instruments. Using the law of iterated expectation, the last condition implies

$$\mathbb{E}(e.w) = 0. \tag{5}$$

The method of moment solves the empirical version of (5) in β, i.e. $\frac{1}{n}\sum_{i=1}^{n} e_i w_i = 0$ where n is the sample size, e_i is the i^{th} component of e and w_i the i^{th} row of w. In turn, the GMM minimizes the quadratic form

$$q(\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} e_i w_i\right)' \Omega \left(\frac{1}{n}\sum_{i=1}^{n} e_i w_i\right) \tag{6}$$

where Ω denotes a weighting matrix. As discussed in [28], there are several choices for matrix Ω leading to different estimators of β. The optimal approach is to define Ω as the inverse of the asymptotic covariance matrix (depending on β) of the estimator. Some alternative procedures are also suggested by Hansen and colleagues [29].

Properties

The properties of the 2SRI are addressed by Terza in [27] when the residual also acting as unobserved confounder in the first-stage regression (at Eq. (2)) is additive. Under this assumption and using nonlinear least squares regression in each stage, they show the consistency of the estimator $\hat{\beta}$ from the second stage. Since the confounder is not additive in the model (2), the first stage estimate of a 2SRI will not be consistent, nor will $\hat{\beta}$. The residual from the first-stage is indeed an unknown function of an unobserved confounder. Then there is a bias that depends on the form of this unknown function when one applies the 2SRI method to the structural model at Eqs. (1) and (2). The derivation of the covariance matrix of 2SRI estimator follows a two-step regression covariance matrix of the form

$$\begin{aligned}\mathbb{V}ar(\hat{\beta}) = &\left(A_{22}^{-1} S_2 A_{22}^{-1'}\right)/n + \left(A_{22}^{-1} A_{21} A_{11}^{-1} S_1 A_{11}^{-1'} A_{21}' A_{22}^{-1'}\right)/n \\ &- \left(A_{22}^{-1} S_{21} A_{11}^{-1'} A_{21}' A_{22}^{-1'}\right)/n,\end{aligned} \tag{7}$$

where the computation of matrices A and S is given in Appendix A.

The GMM is a well documented estimation procedure. Both in linear and nonlinear models, several results on the estimator have already been established. In the literature on econometric analysis, the nonlinear GMM with an instrumental variable has received particular attention. In the pioneering work by Amemiya [30], the author demonstrated the consistency and derived the asymptotic distribution of the nonlinear two-stage least squares estimator (NL2SLS).

This result provided an important insight into how to handle nonlinear models with endogeneity. Later, Hansen [31] showed the asymptotic properties (consistency and asymptotic distribution) of the GMM to be a kind of generalization of the NL2SLS. More recently, Cameron and Trivedi [28] reviewed the method and gave details on the computation of the estimator's covariance matrix in some specific cases. Despite the different results on the GMM with an instrumental variable, its performances in terms of bias and variance depend on the validity and strength of the instrument and the nature of the variables in the model. The computation of the covariance matrix of GMM is given in [28] and implemented in dedicated softwares of which [32] is a good example.

Below, we investigate the performances of these methods with numerical experiments.

Simulation design and data generation

A numerical experiment was conducted to investigate and compare several methods of IV estimation in the context of a dichotomous outcome and exposure with endogeneity in pharmacoepidemiology. In this experiment, we cover a wide range of possible scenarios. We choose values of parameter $\alpha = (\alpha_0, \alpha_z, \alpha_1, \alpha_2)$ corresponding to some values of correlation between the variable $T^* = \alpha_0 + PP\alpha_z + X_1\alpha_1 + X_2\alpha_2 + X_u$ and the Physician's prescribing Preference instrument PP. In fact, we keep α_0, α_1 and α_2 fixed and only α_z varies from a scenario to the other. The

computation of this correlation is given in Appendix B. It somewhat reflects the strength of the instrument when the confounder and other covariables are kept fixed. For each value of the instrument's strength, there are three levels of confounding measured by the standard deviation σ_u of the confounding variable X_u, $\sigma_u \in \{0.5, 1, 1.5\}$ which leads to a set of correlations between T^* and X_u. We then have nine scenarios of strengths and confounding level.

For each scenario, we generate $ns = 1000$ Monte Carlo samples of size n, $n = 10000, 20000$ and 30000. The number of patients per physician is kept fixed and equals 100; the confounder X_u and covariates X_1 and X_2 are assumed to have the normal distributions $N(0, \sigma_u)$, $N(-2, 1)$ and $N(-3, 1)$ respectively and the physician's prescribing preference has the Bernoulli distribution $B(0.7)$. We first simulate the covariables X_1 and X_2, the confounder X_u and the physician's prescribing preference which is the same for all the patients of the same physician. Using the already fixed values of parameters α, the probability p_i that patient i will be prescribed the drug of interest is calculated by inverting the logit function, i.e. $p_i = F(\alpha_0 + PP_i\alpha_z + X_{1i}\alpha_1 + X_{2i}\alpha_2 + X_{ui})$. Treatment T_i of patient i is then generated as a Bernoulli realisation with parameter p_i. The same procedure as for the treatment is used to simulate for each patient i, the corresponding outcome y_i. We fixed the parameter α such that the proportion of exposed patients ranges between 2 and 6% and the prevalence of the event of interest is chosen to be smaller than 5% to reflect a real-life situation of a new treatment (not frequently prescribed) and rare adverse event. With the fixed value of β, we compute the probability $F(\beta_0 + T_i\beta_t + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{ui}\beta_u)$ of the event for patient i and then simulate his outcome y_i. We also explored a more balanced situation in term of exposure frenquencies (between 26 and 45%).

Finally, proxy Z^* of PP is the treatment given to the previous patient and the continuous instrument pr is the proportion of patients of the same physician who were previously prescribed the treatment of interest. More details on the simulation model and the data generating R code are given in Appendix D.

Estimation methods

For the true and conventional models which do not assume endogeneity, the classical one-step regression method without instrumental variable is used. The estimations are performed with the existing regression functions (glm and nls) implemented in R statistical software (R Development core team 2008) in the R package stats.

For the 2SRI method, a two-step regression is used following the procedure outlined in the section dedicated to IV estimation procedure. Recall that the covariance matrices of $\hat{\beta}$ from the second step regression for both methods retrieved from the software results will not be valid since their calculation ignores the fact that some estimated parameters from the first-stage regression are included in the second stage. Then, we re-evaluated these covariance matrices using the sequential two-step estimation procedure taking into account the fact that an estimated variable is used in the second step. The computation of these covariance matrices is given in Appendix A.

For the GMM, the R package gmm proposed by Chaussé [32] is a very helpful tool for computating parameters and estimating covariance. The user needs to implement the sample version of the moment condition function at Eq. (5) and its gradient if possible; if not, a numerical approximation of the gradient function will be used by the gmm function to perform the estimation.

From these estimations, we calculate the asymptotic covariance matrices and the corresponding confidence intervals whose levels are evaluated below.

Evaluation criteria

To evaluate the performances of the various methods, we consider several criteria including the percentage of relative bias (rB in %) defined as

$$\text{rB} = 100 * \frac{1}{ns}\sum_{j=1}^{ns}\left(\frac{\hat{\beta}_t^{(j)}}{\beta_t} - 1\right);$$

the asymptotic standard deviation estimated by the square root of the Monte-Carlo mean of variance $\bar{\hat{\sigma}}^2 = \frac{1}{ns}\sum_{j=1}^{ns}\hat{\sigma}_j^2$, with $\hat{\sigma}_j^2$ the asymptotic variance of $\hat{\beta}_t^{(j)}$ and the Monte-Carlo estimator of the true variance $\mathbb{V}ar(\hat{\beta}_t) = \frac{1}{ns-1}\sum_{j=1}^{ns}\left(\hat{\beta}_t^{(j)} - \bar{\hat{\beta}}_t\right)^2$. We also consider the square root of the mean squares error rMSE given by

$$\text{rMSE} = \sqrt{\frac{1}{ns}\sum_{j=1}^{ns}\left(\hat{\beta}_t^{(j)} - \beta_t\right)^2},$$

and the lower and upper non-coverage probabilities (in %) defined as

$$Er_{inf} = 100 * \frac{1}{ns}\sum_{j=1}^{ns} 1_{\left[\beta_t < IC_{inf}^{(j)}\right]}$$

and

$$Er_{sup} = 100 * \frac{1}{ns}\sum_{j=1}^{ns} 1_{\left[\beta_t > IC_{sup}^{(j)}\right]}$$

where $IC^{(j)} = \left[IC_{inf}^{(j)}; IC_{sup}^{(j)}\right]$ denotes the confidence interval of β_t from the j^{th} Monte-Carlo sample using the asymptotic distribution of $\hat{\beta}_t^{(j)}$. The nonparametric bootstrap based estimate of variance and non-coverage probabilities are also investigated in these simulations and the results are analysed below. We complete all these criteria by the equivalent of the first-stage F-statistics in linear

regression (see [33]) testing instrument exclusion in the treatment choice model.

Results

Table 1 summarizes the performances of each method in terms of relative bias (rB), the standard deviation (sd), the square root of mean squares error (rMSE) and the non-coverage probabilities (pval $= Er_{inf} + Er_{sup}$). It presents the results related to instrument *pr*. There were only slight differences between the results with instrument *pr* and those with Z^*, so we omitted the results related to instrument Z^*. For some samples, the GMM fails to converge owing to singularity problems in the covariance matrix. The estimations from these samples are simply removed (cases marked '—' in Table 1) for all methods. For the other scenarios, infinite variance estimate or outlier coefficient estimate may be observed; the corresponding samples were dropped for the calculation of the criteria. Table 2 shows the number of samples leading to an outlier estimation (rB $>$ 100% or infinite variance) among the 1000 simulated samples. Figure 1 complements the results in Table 1 by displaying the boxplot distributions of rB. Each series of letters a,b,c and d corresponds to the results related to an instrument strength with each letter corresponding to a method as detailed in the legend of Fig. 1. In Table 1 the sd values refer to the Monte-Carlo-based standard deviation. Except for the GMM estimator where outlier values for the asymptotic variance were observed, the Monte-Carlo-based standard deviation, the bootstrap-based estimate (not shown) and the asymptotic estimate were very close.

As expected, the relative bias shows that the estimation from the true and conventional models are insensitive to instrument strength but the confounding level affects the estimation in the conventional model: the relative bias increases with level of confounding. In the presence of a strong instrument, the 2SRI tends to improve the estimate when the level of confounding increases. This trend is reversed when the instrument is weak, i.e. the relative bias and the confounding level have the same direction of variation. The percentage of relative bias of the GMM does not seem too sensitive to instrument strength: it just changes slightly when the strength of the instrument grows. However, this bias increases with the magnitude of confounding, which shows the impact of endogeneity on this method (See Fig. 1).

For the standard deviation (sd) and the square root of the mean squares error (rMSE), the asymptotic results ($n = 30000$) show that both criteria decrease when the level of confounding or instrument strength grows, this being the case for all methods except the GMM for which the rMSE decreases very slowly or remains almost constant in some cases. This trend confirms the already obverved low sensitivity of the GMM to the strength of the instruments used in this simulation. Even though the 2SRI has the larger sd than the other methods in all scenarios with an impact on rMSE in several cases, rMSE for 2SRI method seems improved with high confounding and a strong instrument.

Concerning the non coverage probabilities (pval), the true model estimation and the 2SRI displayed an estimated non-coverage probability around the nominal level of 5% in almost all scenarios. Their values ranged from 4 to 6% and reached 7% in rare cases. The non-coverage probability was very large for the other methods, even with large samples: the results showed an overestimation of the coefficient of treatment for the GMM and conventional approaches. Even at low confounding levels, the conventional method yields very poor coverage probabilities which is coherent with what was observed for the relative bias.

We observe that the metric of the instrument we use retains the same direction of variation with the F-statistics eequivalents (Tables 3 and 4 of Appendix C). Finally, Table 5 in Appendix D summarizes the performances of each method for more balanced exposure frequencies and large sample size ($n = 30000$). In general, performances are less good in comparison with small exposure frequency situation. One could also note that numerical problems arise more often (see Table 6 of Appendix D). Nevertheless, 2SRI is better in term of relative bias and non coverage probability than the conventional method and GMM.

Overall, the results are satisfactory for the 2SRI approach which achieves a similar level of performance to the true model regarding estimation of the confidence interval in the imbalanced situation. We close this section by pointing out the strong numerical instability observed when computing the GMM estimator during these simulations. This could explain the modest performance displayed by the GMM in the simulation results.

Discussion

In this paper, we focus on the effectiveness of regression coefficient estimation in a context of endogeneity, particularly the endogeneity due to unobserved confounding. We are interested in the coefficient of an endogenous treatment which is the basis of risk assessment in pharmacoepidemiology. Linear models are often used in this context to model the probability of dichotomous events (see [7]). Through a simulation study, we investigate the behavior of parameter estimation in nonlinear models specifically logistic regression using some IV-based methods that could potentially be used to overcome the endogeneity issue. The simulation study also made it possible to assess two preference-based instrumental variables in pharmacoepidemiology.

Table 1 Performances of methods using instrument *pr*

		Instrument strength											
		Weak				Mod				Strong			
Level	Method	rB	sd	rMSE	pval	rB	sd	rMSE	pval	rB	sd	rMSE	pval
30000													
High	Tr	0.19	0.12	0.12	0.05	0.21	0.10	0.10	0.06	0.14	0.08	0.08	0.04
	Conv	29.28	0.12	0.89	1.00	27.53	0.10	0.83	1.00	25.10	0.08	0.76	1.00
	2SRI	− 11.44	0.86	0.92	0.06	− 3.44	0.73	0.74	0.04	1.92	0.58	0.58	0.04
	GMM	29.37	0.23	0.91	0.38	28.39	0.19	0.87	0.77	26.69	0.13	0.81	0.93
Med	Tr	0.23	0.13	0.13	0.06	0.12	0.10	0.10	0.05	0.19	0.09	0.09	0.05
	Conv	15.95	0.13	0.50	1.00	14.90	0.11	0.46	1.00	13.10	0.10	0.40	1.00
	2SRI	− 9.75	0.97	1.01	0.07	− 6.35	0.80	0.82	0.05	− 3.40	0.66	0.66	0.05
	GMM	15.35	0.18	0.49	0.29	15.72	0.17	0.50	0.56	15.28	0.15	0.48	0.89
Low	Tr	0.10	0.14	0.14	0.06	− 0.13	0.11	0.11	0.04	0.18	0.10	0.10	0.04
	Conv	4.92	0.15	0.21	0.72	4.39	0.12	0.18	0.67	3.95	0.12	0.17	0.56
	2SRI	− 6.32	1.04	1.06	0.06	− 5.49	0.86	0.87	0.04	− 4.23	0.73	0.75	0.04
	GMM	3.80	0.14	0.18	0.30	4.15	0.14	0.19	0.39	4.61	0.12	0.18	0.84
20000													
High	Tr	0.27	0.14	0.14	0.04	0.05	0.12	0.12	0.05	0.30	0.10	0.10	0.06
	Conv	29.53	0.14	0.90	1.00	27.51	0.11	0.83	1.00	25.21	0.11	0.76	1.00
	2SRI	− 11.03	1.05	1.10	0.05	−5.51	0.91	0.92	0.04	1.34	0.70	0.70	0.04
	GMM	29.01	0.22	0.90	0.32	28.03	0.17	0.86	0.72	27.28	0.19	0.84	0.92
Med	Tr	0.09	0.16	0.16	0.06	0.22	0.13	0.13	0.06	0.19	0.11	0.11	0.07
	Conv	15.94	0.16	0.50	0.99	15.06	0.14	0.47	1.00	13.31	0.13	0.42	0.99
	2SRI	− 10.08	1.19	1.23	0.06	−6.69	1.00	1.02	0.05	− 3.67	0.79	0.80	0.04
	GMM	15.01	0.18	0.48	0.25	15.64	0.19	0.50	0.53	15.40	0.17	0.49	0.85
Low	Tr	− 0.24	0.12	0.12	0.06	0.11	0.14	0.14	0.06	–	–	–	–
	Conv	3.58	0.15	0.19	0.42	4.71	0.16	0.21	0.61	–	–	–	–
	2SRI	0.04	0.88	0.89	0.04	−5.92	1.09	1.10	0.04	–	–	–	–
	GMM	4.41	0.16	0.22	0.79	4.15	0.16	0.20	0.39	–	–	–	–
10000													
High	Tr	0.03	0.21	0.21	0.05	0.49	0.16	0.16	0.05	0.43	0.14	0.14	0.06
	Conv	29.69	0.20	0.91	1.00	28.22	0.16	0.86	1.00	25.70	0.15	0.79	1.00
	2SRI	− 14.47	1.64	1.70	0.07	−2.71	1.27	1.27	0.04	1.19	1.04	1.04	0.05
	GMM	28.86	0.25	0.90	0.30	28.61	0.24	0.89	0.68	27.45	0.22	0.85	0.90
Med	Tr	–	–	–	–	0.58	0.18	0.18	0.04	0.48	0.15	0.16	0.05
	Conv	–	–	–	–	15.83	0.19	0.51	0.97	13.78	0.17	0.45	0.93
	2SRI	–	–	–	–	−7.42	1.41	1.43	0.05	−3.57	1.13	1.13	0.03
	GMM	–	–	–	–	15.74	0.23	0.52	0.50	15.66	0.21	0.51	0.80
Low	Tr	–	–	–	–	0.08	0.20	0.20	0.05	−0.36	0.17	0.17	0.07
	Conv	–	–	–	–	5.27	0.22	0.27	0.55	3.88	0.20	0.23	0.37
	2SRI	–	–	–	–	−6.12	1.61	1.62	0.05	−4.37	1.31	1.31	0.04
	GMM	–	–	–	–	3.77	0.24	0.27	0.36	3.85	0.21	0.24	0.74

Legend: Tr = True model, Conv = Conventional model, 2SRI = Two-Stage Residual Inclusion, GMM = Generalized Method of Moment. Low, Med (Medium), High denote the level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength. For the criteria, rB = relative bias (%), sd = standard deviation, rMSE = root Mean Squares Error and pval = non-coverage probabilities. The numbers 10000, 20000 and 30000 stand for different sample sizes

The results reported from the simulation study show that the 2SRI using nonlinear regression at each stage is an interesting alternative for estimating the coefficient of the endogenous treatment in a logistic regression model. It is very simple to implement and yields satisfactory results regarding the bias and the confidence

Table 2 Number of samples among 1000 leading to outliers in GMM estimation

n	Level	Weak	Mod	Strong
30000				
	High	155	61	31
	Med	78	65	43
	Low	41	149	65
20000				
	High	149	70	64
	Med	64	72	43
	Low	25	124	–
10000				
	High	95	58	37
	Med	–	78	55
	Low	–	110	44

Legend: Low, Med (Medium), High denote level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength. The number *n* with values 10000, 20000 and 30000 stands for the sample size

interval estimate. It was found to yield the most accurate estimate of non-coverage probabilities and thus a more accurate estimate of confidence intervals among the IV-based methods that were compared. However, the conventional approach behaved better than the 2SRI in some cases, especially when the confounding level was weak. We believe that in these cases, the level of confounding is not sufficiently high to require the use of an instrument in the estimation. However, to our knowledge there is still no way to assess the level of unmeasured confounding. For the GMM, the estimation procedure was remarkably unstable. That instability may be attributable to the dichotomous nature of the variables (outcome and exposure) in the context of pharmacoepidemiology with the preference-based instrument. This situation makes the GMM approach is not to be recommended in this context unless another instrument has proved to behave satisfactorily with this method.

Concerning the instruments under investigation in this study, the proportion of all previous patients of the same physician who were prescribed the treatment of interest proved to be a good proxy of the physician's preference instrument. This instrument was previously considered by Abrahamowicz and colleagues [22] for investigating the detection of a possible change point in the physician's preference and their results also seemed satisfactory. This proxy of the physician's preference instrument is thus a credible alternative to the well known other proxy based only on the single patient of the same physician who was prescribed the treatment of interest.

Even though this work throws light on the performances of IV estimators in the context of a nonlinear model with endogeneity, more work is needed to explore the behavior of these estimators in other contexts when the prevalence of exposure and/or outcome varies.

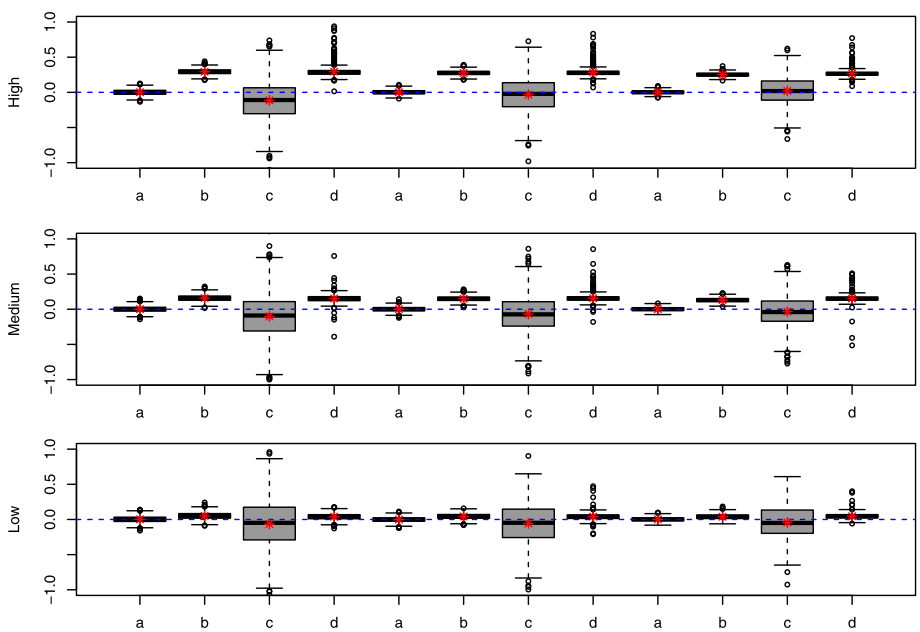

Fig. 1 Relative bias (rB) of the methods. a: True model b: conventional model; c : 2SRI with instrument pr; d: GMM with instrument pr. Low, Medium and High indicate the corresponding level of confounding and the instrument strength grows from a, b, c, d sequence to the next (from left to right)

Conclusions

In observational studies, when assessing the effect of drug exposure on a dichotomous outcome, investigators could use appropriate 2SRI estimation to account for unmeasured confounding. This work showed that two logistic regressions as well as a physician's preference proxy for IV yeald satisfactory results.

Appendix A: Asymptotic variance of the nonlinear 2SRI

Using the sequential two-step estimation procedure, the nonlinear 2SRI estimator minimizes the least squares criterion

$$Q(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - F_{\hat{\alpha}}\left(X_{i\hat{\alpha}}^{'}\beta\right)\right)^2 \tag{8}$$

where $\hat{\alpha}$ denotes the nonlinear least squares estimator of α from the first stage. If we set $\eta_{i\beta} = X_{i\hat{\alpha}}^{'}\beta$, the first-order condition in β is given by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\hat{\alpha}}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\left(y_i - F_{\hat{\alpha}}(\eta_{i\beta})\right) = 0. \tag{9}$$

Given that $\hat{\alpha}$ is a consistent estimator of α_0, the Taylor-lagrange expansion of (9) arround the true value (α_0, β_0) gives

$$\begin{aligned}
0 = {} & \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\left(y_i - F_{\alpha}(\eta_{i\beta})\right)_{(\alpha_c;\beta_c)} + \\
& \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 F_{\alpha}}{\partial \eta_{i\beta}\partial \eta_{i\beta}^{'}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \alpha}\frac{\partial \eta_{i\beta}}{\partial \beta'}\left(y_i - F_{\alpha}(\eta_{i\beta})\right)\right)_{(\alpha_c;\beta_c)}(\hat{\alpha} - \alpha_0) + \\
& \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial^2 \eta_{i\beta}}{\partial \alpha \partial \beta'}\left(y_i - F_{\alpha}(\eta_{i\beta})\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \alpha}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_c;\beta_c)}(\hat{\alpha} - \alpha_0) + \\
& \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 F_{\alpha}}{\partial \eta_{i\beta}\partial \eta_{i\beta}^{'}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\left(y_i - F_{\alpha}(\eta_{i\beta})\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_c;\beta_c)}(\hat{\beta} - \beta_0)
\end{aligned}$$

for some $(\alpha_c; \beta_c)$ between $(\hat{\alpha}; \hat{\beta})$ and $(\alpha_0; \beta_0)$. Under the assumption $\mathbb{E}\left(\frac{\partial Q}{\partial \beta}\right)_{\beta_0} = 0$ the terms involving the residuals $\left(y_i - F_{\alpha}(\eta_{i\beta})\right)$ in the above expansion, except the first, all tend in probability to zero and we obtain

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_0) \approx {} & \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)^{-1}_{(\alpha_0;\beta_0)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\left(y_i - F_{\alpha}(\eta_{i\beta})\right)\right)_{(\alpha_0;\beta_0)} \\
& - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)^{-1}_{(\alpha_0;\beta_0)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \alpha}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_0;\beta_0)}(\hat{\alpha} - \alpha_0).
\end{aligned}$$

The quantity $\sqrt{n}(\hat{\beta} - \beta_0)$ may then be written

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_0) \approx {} & \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)^{-1}_{(\alpha_0;\beta_0)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\left(y_i - F_{\alpha}(\eta_{i\beta})\right)\right)_{(\alpha_0;\beta_0)} \\
& - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)^{-1}_{(\alpha_0;\beta_0)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \alpha}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_{\alpha}}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_0;\beta_0)} \\
& \times \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\frac{\partial \eta_{i\alpha}}{\partial \alpha}\frac{\partial \eta_{i\alpha}}{\partial \alpha'}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\right)^{-1}_{\alpha_0}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\frac{\partial \eta_{i\alpha}}{\partial \alpha}\left(T_i - r(\eta_{i\alpha})\right)\right)_{\alpha_0}
\end{aligned}$$

with $\eta_{i\alpha} = w_i^{'}\alpha$. The latter is derived from the asymptotic approximation of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ using a similar Taylor-Lagrange expansion for the first-stage nonlinear least squares regression.

The previous expansion leads to the covariance matrix of $\hat{\beta}$ of the form

$$\mathbb{V}ar(\hat{\beta}) = \frac{1}{n}\left(A_{22}^{-1}S_2A_{22}^{-1'} + A_{22}^{-1}A_{21}A_{11}^{-1}S_1A_{11}^{-1'}A_{21}^{'}A_{22}^{-1'} - A_{22}^{-1}S_{21}A_{11}^{-1'}A_{21}^{'}A_{22}^{-1'}\right). \tag{10}$$

Under the assumption of independence between observations, the matrices involved in this covariance matrix are given by

$$A_{11} = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\frac{\partial \eta_{i\alpha}}{\partial \alpha}\frac{\partial \eta_{i\alpha}}{\partial \alpha'}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\right)_{\alpha_0}$$

$$A_{21} = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \alpha}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_0;\beta_0)}$$

$$A_{22} = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_0;\beta_0)}$$

$$S_1 = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\frac{\partial \eta_{i\alpha}}{\partial \alpha}U_{i\alpha}^2\frac{\partial \eta_{i\alpha}}{\partial \alpha'}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\right)_{\alpha_0}$$

$$S_2 = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}U_{i\beta}^2\frac{\partial \eta_{i\beta}}{\partial \beta'}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\right)_{(\alpha_0;\beta_0)}$$

$$S_{21} = p\lim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial F_\alpha}{\partial \eta_{i\beta}}(\eta_{i\beta})\frac{\partial \eta_{i\beta}}{\partial \beta}U_{i\beta}U_{i\alpha}\frac{\partial \eta_{i\alpha}}{\partial \alpha'}\frac{\partial r}{\partial \eta_{i\alpha}}(\eta_{i\alpha})\right)_{(\alpha_0;\beta_0)}$$

where p lim denotes the limite in probability, $U_\alpha = T - P_\alpha$, $U_\beta = y - P_\beta$, with $P_\alpha = r(w\alpha)$ and $P_\beta = F_\alpha(X\beta)$.

An estimation of this covariance matrix can be obtained using the plug-in estimator of each matrix involved in its expression. Let $P_{\hat{\alpha}} = r(w\hat{\alpha})$, $P_{\hat{\beta}} = F_{\hat{\alpha}}(X\hat{\beta})$, $U_{\hat{\alpha}} = T - P_{\hat{\alpha}}$ and $U_{\hat{\beta}} = y - P_{\hat{\alpha}}$, then the corresponding estimator of matrices at equation (10) are such that

$$\hat{A}_{11} = \frac{1}{n}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})w\right]^{'}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})w\right],$$

$$\hat{A}_{21} = \frac{\hat{\beta}_u}{n}\left[P_{\hat{\beta}}^2\left(1-P_{\hat{\beta}}\right)^2X\right]^{'}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})w\right]$$

$$\hat{A}_{22} = \frac{1}{n}\left[P_{\hat{\beta}}\left(1-P_{\hat{\beta}}\right)X\right]^{'}\left[P_{\hat{\beta}}\left(1-P_{\hat{\beta}}\right)X\right],$$

$$\hat{S}_1 = \frac{1}{n}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})U_{\hat{\alpha}}w\right]^{'}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})U_{\hat{\alpha}}w\right]$$

$$\hat{S}_2 = \frac{1}{n}\left[P_{\hat{\beta}}(1-P_{\hat{\beta}})U_{\hat{\beta}}X\right]^{'}\left[P_{\hat{\beta}}\left(1-P_{\hat{\beta}}\right)U_{\hat{\beta}}X\right],$$

$$\hat{S}_{21} = \frac{1}{n}\left[P_{\hat{\beta}}\left(1-P_{\hat{\beta}}\right)U_{\hat{\beta}}X\right]^{'}\left[P_{\hat{\alpha}}(1-P_{\hat{\alpha}})U_{\hat{\alpha}}w\right].$$

Appendix B: Instrument strength and confounding level

We give bellow the computation of instrument strength and confounding level

- **Instrument strength**

The strength of an instrument results from the correlation between it and the corresponding endogenous variable. In the model considered here, the strength of an instrument Z is given by $\mathbb{C}orr(T,Z) = \frac{\mathbb{C}ov(T,Z)}{\sqrt{\mathbb{V}ar(T)}\sqrt{\mathbb{V}ar(Z)}}$. As the treatment has a causal link with other covariables Z, X_1, X_2 and X_u, we have

$$\mathbb{V}ar(T) = \mathbb{V}ar_Z[\mathbb{E}(T|Z,X_1,X_2,X_u)] + \mathbb{E}_Z[\mathbb{V}ar(T|Z,X_1,X_2,X_u)].$$

If we consider only the explanatory effect of the instrument in treatment T and replace other covariables and the confounder by their average effect, we have $\mathbb{V}ar(T) = \mathbb{V}ar_Z\left(\frac{1}{1+A_Z}\right) + \mathbb{E}_Z\left(\frac{A_Z}{(1+A_Z)^2}\right)$ with $A_Z = exp(-(\alpha_0 + Z\alpha_z + \mu_1\alpha_1 + \mu_2\alpha_2 + \mu_u))$, $\mu_1 = \mathbb{E}(X_1)$, $\mu_2 = \mathbb{E}(X_2)$ and $\mu_u = \mathbb{E}(X_u)$. This variance may then be written

$$\begin{aligned}\mathbb{V}ar(T) &= \mathbb{E}_Z\left(\frac{1}{1+A_Z}\right)^2 - \left(\mathbb{E}_Z\left(\frac{1}{1+A_Z}\right)\right)^2 + \mathbb{E}_Z\left(\frac{A_Z}{(1+A_Z)^2}\right) &(11)\\ &= \mathbb{E}_Z\left(\frac{1+A_Z}{(1+A_Z)^2}\right) - \left(\mathbb{E}_Z\left(\frac{1}{1+A_Z}\right)\right)^2 &(12)\\ &= \mathbb{E}_Z\left(\frac{1}{1+A_Z}\right) - \left(\mathbb{E}_Z\left(\frac{1}{1+A_Z}\right)\right)^2. &(13)\end{aligned}$$

Considering a dichotomous instrument Z having the Bernoulli distribution $B(p)$, we have $\mathbb{E}_Z\left(\frac{1}{1+A_Z}\right) = \frac{p}{1+A_1} + \frac{1-p}{1+A_0}$, where $A_j, j = 0,1$ is A_Z with Z replaced by j. We finally obtain

$$\mathbb{V}ar(T) = \left(\frac{p}{1+A_1} + \frac{1-p}{1+A_0}\right)\left(1 - \left(\frac{p}{1+A_1} + \frac{1-p}{1+A_0}\right)\right).$$

We also have

$$\begin{aligned}\mathbb{E}(T) &= \mathbb{E}_Z(\mathbb{E}(T|Z,X_1,X_2,X_u)) &(14)\\ &= \mathbb{E}_Z\left(\frac{1}{1+A_Z}\right) &(15)\\ &= \frac{p}{1+A_1} + \frac{1-p}{1+A_0} &(16)\end{aligned}$$

and then $\mathbb{E}(Z)\mathbb{E}(T) = \frac{p^2}{1+A_1} + \frac{p(1-p)}{1+A_0}$. Furthermore, $\mathbb{E}(ZT) = \mathbb{E}_Z(\mathbb{E}(ZT|Z,X_1,X_2,X_u)) = \mathbb{E}_Z(Z\mathbb{E}(T|Z,X_1,X_2,X_u))$ which leads to $\mathbb{E}(ZT) = \mathbb{E}_Z\left(\frac{Z}{1+A_Z}\right) = \frac{p}{1+A_1}$. The covariance between Z and T is then given by

$$\mathbb{C}ov(Z,T) = \mathbb{E}(ZT) - \mathbb{E}(Z)\mathbb{E}(T) \quad (17)$$

$$= \frac{p}{1+A_1} - \frac{p^2}{1+A_1} - \frac{p(1-p)}{1+A_0} \quad (18)$$

$$= p(1-p)\left(\frac{1}{1+A_1} - \frac{1}{1+A_0}\right). \quad (19)$$

The correlation between Z and T is

$$\mathbb{C}orr(T,Z) = \frac{\left(\frac{1}{1+A_1} - \frac{1}{1+A_0}\right)\sqrt{p(1-p)}}{\sqrt{\left(\frac{p}{1+A_1} + \frac{1-p}{1+A_0}\right)\left(1 - \left(\frac{p}{1+A_1} + \frac{1-p}{1+A_0}\right)\right)}}. \quad (20)$$

Then for a given instrument Z with p fixed in (20), $\alpha_0, \alpha_z, \alpha_1$ and α_2 can be chosen to reach a value of A_Z that leads to a desired value of $\mathbb{C}orr(T,Z)$.
Another criterion that could be used to quantify an instrument's strength is the correlation between $T^* = \alpha_0 + Z\alpha_z + X_1\alpha_1 + X_2\alpha_2 + X_u$ and the instrument Z. Since $\mathbb{C}ov(Z,T^*) = \mathbb{C}ov(Z, \alpha_0 + Z\alpha_z + X_1\alpha_1 + X_2\alpha_2 + X_u) = \alpha_z\mathbb{V}ar(Z)$ and $\mathbb{V}ar(T^*) = \alpha_z^2\mathbb{V}ar(Z) + \alpha_1^2\mathbb{V}ar(X_1) + \alpha_2^2\mathbb{V}ar(X_2) + \mathbb{V}ar(X_u)$, under the assumptions on these variables, we have

$$\mathbb{C}orr(Z,T^*) = \frac{\alpha_z\sqrt{p(1-p)}}{\left(\alpha_z^2 p(1-p) + \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \sigma_u^2\right)^{1/2}} \quad (21)$$

where $\sigma_i^2 = \mathbb{V}ar(X_i)$ and $\sigma_u^2 = \mathbb{V}ar(X_u)$.

- **Confounding level**
  A straightforward calculation as above leads to the following correlation between T^* and X_u that expresses the level of confounding.

$$\mathbb{C}orr(X_u,T^*) = \frac{\sigma_u}{\left(\alpha_z^2 p(1-p) + \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \sigma_u^2\right)^{1/2}}. \quad (22)$$

Appendix C: F-statistics for each scenario

The following tables display the equivalent of the first-stage F-statistics in linear regression (see [33]) testing instrument exclusion in the treatment choice model.

Appendix D: Description of simulation model and parameters values, results of the second scenario

For all scenarios, the model generating the binary outcome is the index function

$$Y_i = 1(Y_i^* - \varepsilon_i > 0),$$

with $Y_i^* = \beta_0 + T_i\beta_t + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{ui}\beta_u$ and ε_i the standart logistic distribution. T_i is the observed binary treatment of the individual i, X_{1i} and X_{2i} some characteristics of patient i and X_{ui} the unmeasured confounding factor. Besides β_0 the intercept, β_0, β_t, β_1, β_2 and β_u are parameters related to T, X_1, X_2 and X_u respectively. We fixed these parameters $\beta_0 = -0.6$, $\beta_t = 3$, $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_u = 1$ to keep the prevalence of event less than 5%.

The treatment choice for the i^{th} patient was generated from a Bernoulli model with success probability p_i which depends on the patient's characteristics $X_1 \sim N(-2,1)$

Table 3 Monte-carlo mean of F-statistics in each scenario using the proportion of patients who received the same treatment as proxy of instrument

n	Level	Weak	Mod	Strong
30000				
	High	14.14	101.32	432.00
	Med	16.95	105.93	458.32
	Low	17.65	108.05	465.88
20000				
	High	11.47	69.15	290.29
	Med	13.49	71.53	302.10
	Low	15.93	75.59	315.68
10000				
	High	8.96	36.64	147.29
	Med	11.76	39.26	152.53
	Low	12.48	42.80	161.04

Legend: Low, Med (Medium), High denote level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength. The number *n* with values 10000, 20000 and 30000 stands for the sample size

Table 4 Monte-carlo mean of F-statistics in each scenario using the treatment prescribed to the last patient as proxy of instrument

n	Level	Weak	Mod	Strong
30000				
	High	8.13	13.99	45.30
	Med	12.18	15.66	49.45
	Low	13.50	16.18	50.83
20000				
	High	8.47	10.94	32.26
	Med	11.98	12.11	32.99
	Low	13.35	13.79	31.93
10000				
	High	8.25	8.07	18.48
	Med	10.69	9.80	18.88
	Low	14.17	10.38	19.76

Legend: Low, Med (Medium), High denote level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength. The number *n* with values 10000, 20000 and 30000 stands for the sample size

Table 5 Performances of methods using instrument *pr*

		Instrument strength											
		Weak				Mod				Strong			
Level	Method	*rB*	*sd*	*rMSE*	*pval*	*rB*	*sd*	*rMSE*	*pval*	*rB*	*sd*	*rMSE*	*pval*
High	Tr	1.62	0.23	0.24	0.06	0.52	0.26	0.24	0.04	1.93	0.40	0.32	0.06
	Conv	46.14	0.23	1.40	1.00	42.40	0.26	1.29	1.00	39.62	0.39	1.23	1.00
	2SRI	14.98	0.66	0.85	0.13	9.41	0.47	0.53	0.09	8.49	0.44	0.53	0.10
	GMM	72.01	315.65	5.06	0.27	75.50	109.91	4.70	0.32	71.32	60.00	4.25	0.40
Med	Tr	0.58	0.25	0.23	0.05	1.25	0.31	0.30	0.07	1.61	0.48	0.32	0.04
	Conv	25.40	0.25	0.80	0.96	23.98	0.31	0.78	0.85	21.37	0.48	0.72	0.63
	2SRI	9.93	0.66	0.70	0.07	6.90	0.54	0.61	0.08	5.45	0.54	0.54	0.09
	GMM	36.76	47.90	43.89	0.40	32.85	32.04	2.17	0.51	44.82	41.87	3.36	0.35
Low	Tr	0.66	0.26	0.26	0.07	1.51	0.33	0.28	0.04	2.20	0.55	0.35	0.05
	Conv	7.86	0.26	0.35	0.15	7.96	0.33	0.37	0.11	7.72	0.55	0.41	0.10
	2SRI	6.43	0.71	0.74	0.07	5.76	0.64	0.62	0.06	5.91	0.66	0.67	0.10
	GMM	22.04	31.69	2.42	0.49	26.15	31.68	3.49	0.64	43.12	25.42	4.14	0.46

Legend: Tr = True model, Conv = Conventional model, 2SRI = Two-Stage Residual Inclusion, GMM = Generalized Method of Moment. Low, Med (Medium), High denote the level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength. For the criteria, *rB* = relative bias (%), *sd* = standard deviation, *rMSE* = root Mean Squares error and *pval* = non-coverage probabilities

and $X_2 \sim N(-3, 1)$, on a binary instrument $Z \sim b(0.7)$ and on the confounding factor $X_u \sim N(0, \sigma_u)$. The probability p_i is given by $p_i = F(\alpha_0 + Z_i\alpha_z + X_{1i}\alpha_1 + X_{2i}\alpha_2 + X_{ui})$, where F denotes the logistic distribution function. The standard deviation σ_u of the confounding factor takes values 0.5, 1 and 1.5 corresponding to Low, Medium and high level of confounding respectively. For a fixed level of confounding, we varied only α_z value over $\{1, 2, 3\}$ of which each element corresponds to an instrument strength: 1 for "Low" instrument, 2 for "Moderate" and 3 for "High" instrument. All other parameters in the treatment choice model remain constant ($\alpha_0 = 0.2$, $\alpha_1 = 2$ and $\alpha_2 = 1.2$). The data are generated using the R function sim2Logit2().

We investigated the performances of methods in the context of rare exposures (2 to 6%), and rare events (less than 5%). To check whether the results obtained remain valid in the context of higher exposure (near 50%), we design new simulations in which only the intercepts α_0 and β_0 are modified in the previous design. We held all other parameters constant and fixed $\alpha_0 = 5$ and $\beta_0 = -2.3$. The prevalence of exposure ranged then between 26% and 45% whereas that of event is maintained lower than 6%. We present in Table 5 the results from this second scope over 500 Monte Carlo samples of size 30000. Table 6 displays the number of Monte Carlo samples with outliers.

Table 6 Number of samples among 500 leading to outliers in GMM estimation

n	Level	Weak	Mod	Strong
	High	241	208	155
	Med	220	188	185
	Low	124	127	145

Legend: Low, Med (Medium), High denote level of confounding whereas Weak, Mod (Moderate), Strong stand for instrument strength

This function allows to simulate a compound logistic model with covariates

```
sim2Logit2 <- function(n,m,sigma,alpha,beta,
k=100) {
 # This function allow to simulate a
 compound logistic model with covariates
 # Input
 #  n : The sample size (an integer
 multiple of k)
 #  sigma : the standart deviation of the
 confounder
 #
 #  alpha: coefficients of variables in
 auxiliary model
 #  beta : coefficients of variables in
 causal model

 # Output: The data frame with all
 variables

 # Physician's labels : k patients per
 physician
 Ph    <- rep(1:(n/k),each=k)
 nPh   <- length(unique(Ph))
```

```
pP    <- rbinom(nPh,size=1,prob=m[2])

# Generating the treatment TT
# PP  <- rbinom(n,size=1,prob=m[2])
# pass? de 0.5 ? 0.7
PP  <- rep(pP,each=k)
x1  <- rnorm(n,mean = m[3],sd=1) # pass?
de 3 ?
x2  <- rnorm(n,mean = m[4],sd=1) # pass?
de 1.5 ?
xu  <- rnorm(n,mean = m[5],sd=sigma)
ZZ  <- as.matrix(cbind(rep(1,n),PP,x1,
x2,xu))
probTT    <- inv.logit(ZZ%*%alpha)
TT        <- rbinom(n,size=1,prob=probTT)
tt <- TT[1:k]
fs   <- function(i) return(sum(tt[1:
(i-1)])/(i-1))
prop <- c()
J  <- 1:k
while(max(J)<=n) {
  tt <- TT[J]
  prop <- c(prop,NA,sapply(2:k,FUN=fs))
 J <- J + k
}
# Generating the outcome y
XX  <- as.matrix(cbind(rep(1,n),TT,x1,
x2,xu))
y   <- 1*(XX%*%beta - rlogis(n)>0)
z   <- rep(NA,n)
for(i in 2:n)
{if(Ph[i]==Ph[i-1]) z[i] <- TT[i-1]}
data <- data.frame(XX,PP,z,prop,y)
colnames(data) = c('const','TT','x1','x2',
'xu','PP','z','prop','y')
return(data)
}
```

Abbreviations

2SLS: Two-Stage Least Squares; 2SPS: Two-Stage Predictor Substitution; 2SRI: Two-Stage Residual Inclusion; GMM: Generalized Method of Moment; Inserm: Institut national de la santé et de la recherche médicale; IV: Instrumental variable; OR: Odd Ratio; PP: Physician Preference; pval: non coverage probability; rB: Relative Bias; rMSE: root Mean Square Error; sd: standard deviation; UVSQ: Université de Versailles Saint-Quentin-en-Yvelines

Funding

This work was supported by grants from the Fondation pour la Recherche Médicale ("Iatrogénie des médicaments 2013") and the Agence Nationale pour la Recherche ("Appel à projet générique 2015"). The funding bodies had no role in the contents and the writing of the manuscript.

Availability of data and materials

The R programm for simulation are available from the corresponding author.

Authors' contributions

BFK and PTB conceived the study. BFK performed the simulation study and drafted the manuscript. BFK, PTB and SE interpreted the results, read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 21 November 2016 Accepted: 23 May 2018
Published online: 22 June 2018

References

1. Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol. 2009;169(3):273–84. https://doi.org/10.1093/aje/kwn299. http://aje.oxfordjournals.org/content/169/3/273.full.pdf+html.
2. Gowrisankaran G, Town RJ. Estimating the quality of care in hospitals using instrumental variables. J Health Econ. 1999;18(6):747–67. https://doi.org/10.1016/S0167-6296(99)00022-3.
3. Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Nonlinear Models. Monographs on Statistics and Applied Probability. London: Chapman & Hall; 1995.
4. Nestler S. Using instrumental variables to estimate the parameters in unconditional and conditional second-order latent growth models. Struct Equ Model A Multidiscip J. 2015;22(3):461–73.
5. Greene WH. Econometric Analysis, 7th edition, international edition. Boston: Pearson; 2012. pp. 259–89. Previous edition: 2008.
6. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J Am Stat Assoc. 1995;90(430):443–50.
7. Terza JV, Bradford WD, Dismuke CE. The use of linear instrumental variables methods in health services research and health economics: A cautionary note. Health Serv Res. 2008;43(3):1102–1120.
8. Foster EM. Instrumental Variables for Logistic Regression: An Illustration. Soc Sci Res. 1997;26(4):487–504. https://doi.org/10.1006/ssre.1997.0606. Accessed 02 Oct 2015.
9. Terza JV. Estimation of policy effects using parametric nonlinear models: a contextual critique of the generalized method of moments. Health Serv Outcome Res Methodol. 2006;6(3):177–98. https://doi.org/10.1007/s10742-006-0013-0.
10. Cai B, Small DS, Have TRT. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. Stat Med. 2011;30(15):1809–1824. https://doi.org/10.1002/sim.4241.
11. Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BHC, Leufkens HGM, de Boer A. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol. 2004;57:1223–1231. https://doi.org/10.1016/j.jclinepi.2004.03.011.
12. Palmer TM, Sterne JAC, Harbord RM, Lawlor DA, Sheehan NA, Meng S, Granell R, Smith GD, Didelez V. Instrumental variable estimation of causal risk ratios and causal odds ratios in mendelian randomization analyses. Am J Epidemiol. 2011;173(12):1392.
13. Chapman CG, Brooks JM. Treatment effect estimation using nonlinear two-stage instrumental variable estimators: Another cautionary note. Health Serv Res. 2016;51(6):2375–394. https://doi.org/10.1111/1475-6773.12463.
14. Johnston KM, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. Stat Med. 2008;27(9):1539–1556. https://doi.org/10.1002/sim.3036.
15. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000;29(4):722–9. https://doi.org/10.1093/ije/29.4.722. http://ije.oxfordjournals.org/content/29/4/722.full.pdf+html.
16. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: Analysis using instrumental variables. JAMA. 1994;272(11):859–66. https://doi.org/10.1001/jama.1994.03520110039026.

17. Cain LE, Cole SR, Greenland S, Brown TT, Chmiel JS, Kingsley L, Detels R. Effect of highly active antiretroviral therapy on incident aids using calendar period as an instrumental variable. Am J Epidemiol. 2009;169(9): 1124–1132. https://doi.org/10.1093/aje/kwp002. http://aje.oxfordjournals.org/content/169/9/1124.full.pdf+html.
18. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat. 2007;3(1):14. https://doi.org/10.2202/1557-4679.1072.
19. Brooks JM, Chrischilles EA, Scott SD, Chen-Hardee SS. Was breast conserving surgery underutilized for early stage breast cancer? instrumental variables evidence for stage ii patients from iowa. Health Serv Res. 2003;38(6p1):1385–1402. https://doi.org/10.1111/j.1475-6773.2003.00184.x.
20. Johnston SC. Combining ecological and individual variables to reduce confounding by indication:: Case study–subarachnoid hemorrhage treatment. J Clin Epidemiol. 2000;53(12):1236–1241. https://doi.org/10.1016/S0895-4356(00)00251-1.
21. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology. 2006;17(3):268–75.
22. Abrahamowicz M, Beauchamp ME, Ionescu-Ittu R, Delaney JAC, Pilote L. Reducing the variance of the prescribing preference-based instrumental variable estimates of the treatment effect. Am J Epidemiol. 2011;174(4): 494–502. https://doi.org/10.1093/aje/kwr057. http://aje.oxfordjournals.org/content/174/4/494.full.pdf+html.
23. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med. 2014;33(13):2297–340. https://doi.org/10.1002/sim.6128.
24. Uddin MJ, Groenwold RHH, de Boer A, Gardarsdottir H, Martin E, Candore G, Belitser SV, Hoes AW, Roes KCB, Klungel OH. Instrumental variables analysis using multiple databases: an example of antidepressant use and risk of hip fracture. Pharmacoepidemiol Drug Saf. 2016;25: 122–31. https://doi.org/10.1002/pds.3863. PDS-14-0390.R2.
25. Uddin MJ, Groenwold RHH, de Boer A, Afonso ASM, Primatesta P, Becker C, Belitser SV, Hoes AW, Roes KCB, Klungel OH. Evaluating different physician's prescribing preference based instrumental variables in two primary care databases: a study of inhaled long-acting beta2-agonist use and the risk of myocardial infarction. Pharmacoepidemiol Drug Saf. 2016;25:132–41. https://doi.org/10.1002/pds.3860. PDS-14-0383.R2.
26. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf. 2010;19(6):537–54. https://doi.org/10.1002/pds.1908.
27. Terza JV, Basu A, Rathouz PJ. Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. J Health Econ. 2008;27(3):531–43. https://doi.org/10.1016/j.jhealeco.2007.09.009. Accessed 02 Oct 2015.
28. Cameron AC, Trivedi PK. Microeconometrics: Methods and Applications. Cambridge: Cambridge University Press; 2005. pp. 166–220. https://doi.org/10.1017/CBO9780511811241.
29. Hansen LP, Heaton J, Yaron A. Finite-sample properties of some alternative gmm estimators. J Bus Econ Stat. 1996;14(3):262–80.
30. Amemiya T. The nonlinear two-stage least-squares estimator. J Econ. 1974;2(2):105–10. https://doi.org/10.1016/0304-4076(74)90033-5.
31. Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50(4):1029–1054.
32. Chausse P. Computing generalized method of moments and generalized empirical likelihood with r. J Stat Softw. 2010;34(1):1–35.
33. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat. 2002;20(4):518–29.