Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods

Uddin MJ¹, Groenwold RH¹,², Ton de Boer¹, Belitser SV¹, Roes KC² and Klungel OH¹,²*
¹Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Utrecht, The Netherlands
²Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Abstract

Instrumental variables (IV) analysis seems an attractive method to control for unmeasured confounding in observational epidemiological studies. Here, we provide an overview of the estimation methods of IV analysis and indicate their possible advantages and limitations. We found that two-stage least squares is the method of first choice if exposure and outcome are both continuous and show a linear relation. In case of a nonlinear relation, two-stage residual inclusion may be a suitable alternative. In settings with binary outcomes as well as nonlinear relations between exposure and outcome, generalized method of moments (GMM), structural mean models (SMM), and bivariate probit models perform well, yet GMM and SMM are generally more robust. The standard errors of the IV estimate can be estimated using a robust or bootstrap method. All estimation methods are prone to bias when the IV assumptions are violated. Researchers should be aware of the underlying assumptions of the estimation methods as well as the key assumptions of the IV when interpreting the exposure effects estimated through IV analysis.

Keywords: Instrumental variables; Estimation method; Unobserved confounding; Epidemiology; Statistical methods; Observational studies; Causal inference

Introduction

Instrumental variable (IV) analysis has primarily been used in economics and social science research, as a tool for causal inference, but has begun to appear in epidemiologic research over the last decade to control for unmeasured confounding [1-6]. An IV is a variable that can be considered to mimic the treatment assignment process in a randomized study [7-10]. IV analysis generally involves a two-stage modelling approach to estimate the exposure effects. In the first stage, the effect of the IV on exposure is estimated, whereas in the second stage, outcomes are compared in terms of predicted exposure rather than the actual exposure [11]. To value the estimates obtained through IV analysis, it is important to understand the underlying methodology of the estimation methods in the IV analysis.

Over the last decade, several reviews of IV analysis were published, covering various aspects including the key assumptions, estimating parameters, possible IVs, estimation methods, reporting of the results, and the use of IVs in comparative effectiveness research [3,4,12-23]. We summarized these reviews in Table 1. However, none of these articles included all possible estimation methods of IV analysis. Hence, we aimed to provide an overview of the estimation methods and to indicate their possible advantages and limitations. After a general introduction to the assumptions underlying IV analysis, we will describe the methods that have been used in IV studies in medical research.

Instrumental variables

The IV is an observed variable, which is related to exposure and only related to the outcome through exposure. This resembles a randomized trial, in which treatment allocation typically almost perfectly coincides with the actual treatment received and (in case of a double blind trial) treatment assignment only affects the outcome through the received treatment (hence the term pseudo-randomisation that is used for IV methods). This implies that an IV is neither directly nor indirectly (e.g. through observed or unobserved confounders) associated with the outcome [6,18,24]. Therefore, all observed and unobserved confounders should on average be equally distributed among different levels of the IV (similar to a randomized trial). These assumptions are illustrated in Figure 1. Along with these basic assumptions, there are other assumptions (i.e., homogeneous treatment effects, monotonicity) that are needed for point identification of IV estimates [14,19].

Notation

Throughout this article, we use the following notation: Y denotes the outcome, X denotes exposure, and Z denotes the IV. C and U denote the (one or more) observed and unobserved confounding variables, respectively. X denotes the predicted value of exposure. Finally, β denotes the IV estimator, i.e., the estimator of the causal relation between exposure and outcome.

Estimation method of IV analysis

Ratio estimator (RE)

In a study with a single binary IV, the RE (also called Wald [25] or grouping estimator) can be applied and which is expressed as:

\[
\hat{\beta}_{IV} = \frac{\hat{p}(Y=1|Z=1) - \hat{p}(Y=1|Z=0)}{\hat{p}(X=1|Z=1) - \hat{p}(X=1|Z=0)}
\]

(1)

*Corresponding author: Klungel OH, Division of Pharmacoepidemiology and Clinical Pharmacology, University of Utrecht, Utrecht, Netherlands, Tel: +31685384692, Fax +31-30 253 9166; E-mail: O.H.Klungel@uu.nl

Received January 06, 2015; Accepted March 08, 2015; Published March 15, 2015


Copyright: © 2015 Uddin MJ, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
<table>
<thead>
<tr>
<th>Author</th>
<th>Publication year</th>
<th>Journal name</th>
<th>Title</th>
<th>Main features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenland</td>
<td>2000</td>
<td>International Journal of Epidemiology</td>
<td>An introduction to instrumental variables for epidemiologists</td>
<td>- basic introduction with an empirical example</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- link with randomized studies with non-compliance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- estimated bound for the exposure effects</td>
</tr>
<tr>
<td>Martens et al.</td>
<td>2006</td>
<td>Epidemiology</td>
<td>Instrumental variables: application and limitations</td>
<td>- fundamental issues are described with several practical details using graphical representation</td>
</tr>
<tr>
<td>Hernan and Robins</td>
<td>2006</td>
<td>Epidemiology</td>
<td>Instruments for causal inference: an epidemiologists dream?</td>
<td>- overview of IV analysis with explanation of several key assumptions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- highlights limitations and emphasis on estimating parameters of IV analysis</td>
</tr>
<tr>
<td>Rassen et al.</td>
<td>2009</td>
<td>Journal of Clinical Epidemiology</td>
<td>Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships</td>
<td>- demonstrates how IV analysis arises from an analogous but potentially impossible RCT design</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- shows estimation of effects with an empirical example</td>
</tr>
<tr>
<td>Rassen et al.</td>
<td>2009</td>
<td>Journal of Clinical Epidemiology</td>
<td>Instrumental variables II: instrumental variable application—in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance</td>
<td>- assesses the overall relationship between strength and imbalance of confounders between IV categories with an empirical example</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- assesses several possible IVs</td>
</tr>
<tr>
<td>Rassen et al.</td>
<td>2009</td>
<td>American Journal of Epidemiology</td>
<td>Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes</td>
<td>- reviews commonly used IV estimation methods for binary outcome and compared them in empirical examples</td>
</tr>
<tr>
<td>Brookhart et al.</td>
<td>2010</td>
<td>Pharmacoepidemiology and Drug Safety</td>
<td>Instrumental variable methods in comparative safety and effectiveness research</td>
<td>- guidance on reporting of IV analysis with an empirical example</td>
</tr>
<tr>
<td>Chen and Briesacher</td>
<td>2011</td>
<td>Journal of Clinical Epidemiology</td>
<td>Use of instrumental variable in prescription drug research with observational data: a systematic review</td>
<td>- review of practice of IV analysis in epidemiology</td>
</tr>
<tr>
<td>Palmer et al.</td>
<td>2011</td>
<td>American Journal of Epidemiology</td>
<td>Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization</td>
<td>- overview of commonly used IV estimation methods for continuous exposure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- empirical example of Mendelian randomization study</td>
</tr>
<tr>
<td>Davies et al.</td>
<td>2013</td>
<td>Epidemiology</td>
<td>Issues in the reporting and conduct of instrumental variable studies: a systematic review</td>
<td>- review of practice of IV analysis in epidemiology - focus on target parameter (e.g. RD, OR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- reviews methods used to estimate standard errors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- proposes a checklist of information to be reported by studies using instrumental variables</td>
</tr>
<tr>
<td>Swanson and Hernan</td>
<td>2013</td>
<td>Epidemiology</td>
<td>Commentary: How to report instrumental variable analyses (suggestions welcome)</td>
<td>- provided flow chart for reporting of IV analyses</td>
</tr>
</tbody>
</table>
Instrumental variable methods for causal inference

Table 1: Introductory and review articles of instrumental variable analysis in Epidemiologic studies (2000-2014)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baiocchi et al.</td>
<td>2014</td>
<td>Statistics in Med</td>
</tr>
<tr>
<td>Garabedian et al.</td>
<td>2014</td>
<td>Annals of Internal Medicine</td>
</tr>
</tbody>
</table>

The first model estimates the effect of the IV on exposure, whereas in the second model outcomes are compared in terms of predicted exposure rather than the actual exposure. The latter model yields the estimated parameter, \( \beta \), which is the IV estimator. For a single IV, the \( \beta \) is equivalent to the estimators in the equations (1), (2), and (3). In case of multiple IVs, information on these IVs can be simultaneously incorporated in model (4). Then, \( \beta \) is the weighted average of the ratio estimators [30]. For multiple IVs, 2SLS provides biased estimates [30-32] and another method, e.g., limited information maximum likelihood (LIML), [33] can be an alternative. One of the conditions of this method is that the error term should be homoscedastic (homogeneity of variance). However, in case of heteroscedasticity, other methods (e.g., generalized method of moments) can be considered [34]. Moreover, the 2SLS may produce biased results in the case of binary variables or non-linear relationship between exposure and outcome (Table 2).

Linear probability model (LPM)

This method is a particular form of the 2SLS in which the outcome, exposure, and IV are binary and provides exposure effects on the risk difference scale. When there is a single binary IV, the estimator can be expressed as in equation (3) [13,35-37].

LPM is a simple technique to estimate the parameter and interpret as the regression coefficients based on linear regression. However, in linear IV analysis, LPM may provide ambiguous results because the common technique of linear IVs designed for a continuous response [38]. It should be noted that the LPM of binary exposure and outcome may produce predicted values outside of the 0–1 range [28]. Hence, for rare binary outcomes, some predicted probabilities may become negative [39]. In addition, the probability of success increases linearly with exposure, that is, the marginal or incremental effect of exposure remains constant [37], which is logically impossible for binary outcomes [14].

Two-stage predictor substitution (2SPS)

The two-stage predictor substitution is an extension of the 2SLS to nonlinear models, which targets a marginal (population-averaged) odds ratio [36,40-42]. In the first-stage, a nonlinear least squares method (NLS) or any other consistent estimation technique is used to estimate the relationship between the IV and exposure [43]. Then, the predicted exposure status from the first-stage model replaces the observed exposure as the principal covariate in the second-stage model on the outcome [43,44]. For a continuous exposure and outcome, 2SPS and 2SLS show similar results [24,36].

Two-stage residual inclusion (2SRI)

2SRI (also called control function estimator) [45] is another two-stage method and was first suggested by Hausman [46]. The general notion of the 2SRI is to include the error terms (residuals) from the first-stage model as an additional variable along with the exposure in the second-stage model [47]. The models in the first and second-stage can be either linear or nonlinear models. In case of linear models, the 2SRI estimate is equivalent to the 2SLS and 2SPS estimates [44,48].
<table>
<thead>
<tr>
<th>Method</th>
<th>Basic notion</th>
<th>Exposure effects</th>
<th>Strength</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio estimator (RE)</td>
<td>- the RE is appropriate when only one IV</td>
<td>RD, RR, OR</td>
<td>- simple estimation method</td>
<td>- not suitable for multiple IVs</td>
</tr>
<tr>
<td></td>
<td>- linear models without making parametric assumptions on the error terms</td>
<td></td>
<td>- with a single binary IV and no other confounders, 2SLS = RE</td>
<td>- does not allow to adjust confounders</td>
</tr>
<tr>
<td>Two-stage least squares (2SLS)</td>
<td>- for multiple IVs, IV estimator is the weighted average of the ratio estimators</td>
<td></td>
<td>- natural starting point of IV analysis</td>
<td>- may not consistent for the causal OR</td>
</tr>
<tr>
<td></td>
<td>- estimator similar as classical regression</td>
<td></td>
<td>- the estimate asymptotically unbiased</td>
<td>- show biased results in binary cases or in the case of non-linear models</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- widely used for binary exposure and outcome</td>
<td>- for multiple IVs, 2SLS estimator is biased and hence</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- provides the exposure effect on risk difference scale</td>
<td>limited information of maximum likelihood method would be an alternative</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- unlike RE, it is able to adjust the possible measured confounders</td>
<td>- for smaller sample sizes, limited information maximum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>likelihood estimator is more efficient and consistent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2SLS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- IV and 2SLS are a special case of GMM; however both yield the same</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>results in the case of homoscedastic errors variance</td>
</tr>
<tr>
<td>Linear probability models (LPM)</td>
<td>- applied for binary outcome, exposure, and IV, the data are modelled using linear functions</td>
<td>RD</td>
<td>- simple to estimate and interpret as the regression coefficients</td>
<td>- sometimes predicted probabilities outside of the 0–1 range and for</td>
</tr>
<tr>
<td></td>
<td>- for a single binary IV, the estimator equivalent to the RE</td>
<td></td>
<td>- the RD is consistent for the ACE</td>
<td>rare outcomes this may become negative</td>
</tr>
<tr>
<td>Two-stage predictor substitution (2SPS)</td>
<td>- the rope extension to nonlinear models of the linear IV models - targets a marginal (population-averaged) odds ratio</td>
<td>RD, RR, OR</td>
<td>- suitable for non-linear association between exposure and outcome</td>
<td>- assumes the marginal/incremental effect of exposure remains constant</td>
</tr>
<tr>
<td></td>
<td>- it is the mimic of 2SLS</td>
<td></td>
<td></td>
<td>which is logically impossible for binary outcome</td>
</tr>
<tr>
<td></td>
<td>- non-linear least squares is used to estimate the parameter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- for a linear model, 2SPS = 2SLS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-stage residual inclusion (2SRI)</td>
<td>- include the estimated unobservable confounder (residual) from the first-stage as an additional variable along with the exposure in the second-stage model - also called control function estimator - under a linear model, 2SRI = 2SLS</td>
<td>RD, RR, OR</td>
<td>- yields consistent estimates for linear and non-linear models</td>
<td>- in practice, 2SPS in non-linear model does not always yield consistent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>exposure effects on the outcome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- parameter estimation process is more difficult than 2SLS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- under a logistic regression model, 2SPS may not provide causal OR</td>
</tr>
<tr>
<td>Two-stage logistic regression (2SLR)</td>
<td>- when outcome and exposure are binary and interest to estimate OR - fully parametric, maximum likelihood technique is used to estimate the parameters</td>
<td>OR</td>
<td>- parallel to 2SLS using LRM in both stages instead of linear models</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>if the first-stage logistic model is not correctly specified then</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>second-stage parameter estimates might be biased</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- estimator does not provide COR</td>
</tr>
<tr>
<td>Three-stage least squares (3SLS)</td>
<td>- an extension of 2SLS but unlike the 2SLS, all coefficients are estimated simultaneously, requires three steps - in 2SLS, if the errors in the two equations are correlated, the 3SLS can be an alternative</td>
<td>RD, RR</td>
<td>- more information is used and hence the estimators are likely to be</td>
<td>- more vulnerable to a misspecification of the error terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>more efficient than 2SLS</td>
<td>- very rarely applied in epidemiologic studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- estimation process is more complicated than 2SLS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- 3SLS becomes inconsistent if errors are heteroskedastic</td>
</tr>
<tr>
<td>Structural mean models (SMM)</td>
<td>- SMMs use IVs via G-estimation and involves the assumption of conditional mean independence - additive SMMs use continuous outcome and multiplicative SMMS use positive-valued outcomes - MSMM assumed log-linear model to measure the risk ratio - LSMM assumes logistic regression model which is fitted by maximum likelihood technique</td>
<td>RD, RR, OR</td>
<td>- the assumption of no effect modification is impossible to verify</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- with a binary outcome, additive SMMs and MSMM suffer from the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>limitations of linear and log-linear models</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(e.g., predicted response probabilities may outside of the interval [0,1])</td>
</tr>
</tbody>
</table>
However, for logistic regression model (LRM), 2SRI estimator may not provide causal odds ratio due to non-collapsibility of the odds ratio. 2SRI yields consistent estimates for both linear and nonlinear models [49,50]. The advantage of 2SRI over 2LS is that 2SLS is only consistent when the second-stage model is linear, whereas this restriction does not hold for 2SRI [43,51]. Moreover, this method shows more precise estimates than 2SPS [52].

**Two-stage logistic regression (2SLR)**

When both the outcome and exposure are binary and the interest is to use IV to estimate odds ratios, 2SLR can be applied. It is similar to 2LSL, but instead of linear models using logistic models in both stages [4,53]. This method is fully parametric and maximum likelihood estimation is used to estimate the parameters. If the first-stage logistic model is not correctly specified, the estimates from the second-stage can be biased [54,55]. Also, note that this method may not provide the causal odds ratio due to the non-collapsibility of the OR [19].

**Three-stage least squares method (3SLS)**

The 3SLS generalizes the 2LSL. Possible correlation of the errors ($\varepsilon_2$ and $\varepsilon_3$) in equations (4) and (5) is not taken into account by 2LSL. 3SLS accounts for the possible correlations between errors and may improve the efficiency of the estimator [56,57]. Unlike 2LSL, in which the coefficients of the two equations are estimated separately, in 3SLS all coefficients are estimated simultaneously. This requires three steps. The first-stage is similar to the 2LSL, i.e., a linear regression of X on Z to get X. In the second-stage, the residuals of the second-stage 2SLS model are obtained to estimate the cross-model correlation matrix (correlation between error terms in both models). Finally, in the third-stage the estimated correlation matrix is used to obtain the IV estimator. When there is no correlation between the error terms of the 2SLS models, the 3SLS reduces to a 2SLS. However, 3SLS is more vulnerable to misspecification error since misspecification of one of the models in the first or second will affect the third stage model [58].

**Structural mean models (SMMs)**

SMMs explicitly use counterfactuals or potential outcomes [52], which were originally proposed by Robins [59] in the context of randomized trials with non-compliance to estimate the causal effects for the treated (exposed) individuals. SMMs are semi-parametric models and use IV’s via G-estimation for identification and estimation of the causal parameter. This method involves the assumption of a conditional mean independence [14,19,60-62] and does not make distributional assumptions about the exposure [19]. SMMs with an identity link is sometimes called additive SMMs and can be used for continuous outcomes and multiplicative SMMs with log-linear model can be used for positive-valued/binary outcomes in order to estimate the causal risk ratio [19,63]. Additionally, the logistic structural mean model (LSMM) developed by Vansteelandt and Goetghebeur [64] and Robins and Rotnitzky [65] can also be used for binary outcome in order to estimate causal odds ratio [19,63].

To handle continuous outcome data, the IV estimator from the additive SMMs can be expressed as equation (2) given that the assumptions of CMI and no effect modification by Z are fulfilled [14,62,66,67]. This estimator provides the average treatment effect (ATT) for the treated individuals [19,68].

The advantage of this method is that it relaxes several of the modelling restrictions such as homogeneous treatment effects required by more classical methods such as RE/two-stage IV methods [14,19]. One of the key assumptions of this method is no effect modification, which is difficult to verify in practical situations [67].

SMMs have been extended by Robins [60] to a general setting of structural nested mean models (SNMM) for repeated measures at multiple time points. The SMMs are a subclass of the SNMM [59,69]. When instruments, exposures, and confounders are time-dependent,
SNMM can be used to estimate causal effects of exposure on the outcome [14]. Details and mathematical formulations of SMMs are described elsewhere [14,19,63].

**Generalized method of moments (GMM)**

When applying the GMM a system of equations is set up, which is then solved numerically using computer algorithms. This technique was formalized by Hansen [70] and is a broad class of estimation methods that allow for a larger number of equations (moment conditions) than parameters [4,53,71] that are not possible in the MSMM and LSMM [19]. More clearly, the GMM allows for estimation of parameters in an over-identified model (number of IVs greater than the number of exposures). GMM with linear model can be similar to the ones used in 2SLS [72] but GMM is also a non-linear analogue of 2SLS [17], which is called multiplicative GMM. Detailed explanations can be found elsewhere [4,19,53].

In general, the nonlinear optimum GMM estimator is asymptotically more efficient than 2SLS [73]. Since GMM is a moment based method without parametric assumptions, it is less prone to model misspecification than 2SLR or bivariate probit models when exposure and outcome are binary [4]. In case of a linear model and single IV, the GMM estimator is equivalent to 2SLS, additive SMM, and LIML [53,66,74]. On the other hand, with log-linear model, (i.e., MGMM) [19], it is equivalent with MSMM and provides the population causal risk ratio [19]. However, this estimator with logistic regression model is not consistent for the causal odds ratio due to non-collapsibility of the odds ratio [17].

In case of a binary or count outcome, Palmer et al. [75] suggested a two-stage IV method where the first-stage is a linear regression and the second-stage model is a logistic or log-linear model [19]. Since IV analysis with logistic regression may not provide a consistent exposure effect, in order to estimate causal risk ratio, GMM with log-linear model is preferable. Moreover, 2SRI [48] is also applicable in the setting of count outcome.

**Bivariate probit models (BPM)**

When the outcome of interest is binary, so-called probit models can be applied for IV analysis. In contrast to 2SLS, probit models directly model probabilities (i.e., are restricted on (0, 1)) [4,30]. BPM can be applied in two-stages, but unlike common two-stage estimation methods, this method is estimated via full-information maximum likelihood, which takes into account the correlation between the error terms in the two equations [24]. A more detailed model description can be found elsewhere [4,30].

The interpretation of BPM parameters is not like those of ordinary regression model parameters (e.g., logarithm of odds ratio from a logistic model). However, by multiplying a probit coefficient by approximately 1.6 or 1.8, probit coefficients approximate the coefficients obtained through logistic regression [4].

In case of binary outcome, linear IV methods may yield biased results and BPM may be preferable [30,47]. Furthermore, the estimates are more efficient than 2SLS, whereas 2SLS models are more robust to incorrect modelling assumptions regarding the bivariate normal distribution of the error terms [76,77]. However, when the distribution of error terms is not normal or the average probability of the outcome variable is close to one or zero, or if there is more than one exposure, the estimates from the BPM are generally not consistent for the average causal effect [30,77].

**Other estimation methods**

Apart from the methods discussed above, the outcome variable in epidemiologic research may also be a time-to-event. Also in case of these outcome variables, IV analysis has been applied with two-stage method. In that case, the second-stage model could be a Cox proportional hazards model [78-80]. However, Brookhart et al. [3] stated that this approach for IV analysis is not motivated by a theoretical model and, therefore, parameters that are obtained from this approach may not be causally interpretable. Examples of this approach are a study of the effect of rosiglitazone on (time to) cardiovascular hospitalization and all-cause mortality using facility-prescribing patterns as an IV [78], and a study of the effect of adjuvant chemotherapy on (time to) breast cancer recurrence using physician preference as an IV [79].

**Standard error and characteristics of IV estimators**

Consider two-stage models for IV analysis, in which the predicted value of exposure from the first-stage model is included in the second-stage model. The uncertainty around this prediction is not taken into account in the latter model, which therefore may result in incorrect precision. Typically, standard errors (SEs) of the IV estimate from the second-stage model are too small [24,30,44,45]. An alternative method to estimate a correct SE is the so-called sandwich variance estimator (robust SE), which involves cross products of the predicted treatment and a dispersion factor based on the observed treatment [49]. Most statistical software packages provide this sandwich variance estimate [10]. Angrist and Krueger [10] noticed that these SEs are asymptotically valid, but in practice (with finite sample size) they are only approximately valid.

An alternative way of estimating SEs is the bootstrap method [81]. Here bootstrap samples of the original data can be used to estimate the variation in the IV estimates and hence its SE [4,6,82-84]. It should be stressed that one of the weaknesses of the IV estimator is that it tends to display large SEs relative to the conventional regression estimator [13,85]. It is also noted that the IV estimator can perform poorly in finite samples and show biased results [31] and this bias is amplified when the IV is weak [14,31].

**Interpretation of exposure effects from IV analysis**

Researchers may be interested to estimate the average treatment effects over the entire study population [27]. However, it has been argued that the basic assumptions of IV analysis are not sufficient to achieve point estimates for the causal effect of exposure on the outcome, but only estimate upper and lower bounds of this parameter [14,86,87]. To achieve a point estimate of the average causal effect (ACE) over the entire study population, the additional strong assumption of homogeneity of exposure across levels of the IV should be satisfied [52]. Moreover, IV analysis captures the ATT under the assumption of no effect modification by IV [52]. When exposure effects are not homogeneous across IV levels, under the monotonicity assumption (i.e., the IV affects the treatment deterministically in one direction), the IV estimate quantifies the local average treatment effect (LATE) [88], which is only informative for a subset of the study population, namely those who comply with the IV [27,89-91].

**Assessment of IV assumptions**

As noted, IV analysis must satisfy three basic assumptions and if these assumptions do not hold, results may be severely biased [3,13]. The first assumption (i.e., the IV is related to exposure) is generally easier to check using available statistical methods than the other two
assumptions. The second (IV has no direct effect on outcome) and third (IVs independent of confounders) assumptions are unverifiable or not directly testable as they involve unobservable variables [1,13,18,19,68,76,92]. Some authors proposed circumstantial evidence to support these assumptions [2,5,93,94]. Alternatively, for the third assumption a falsification test based on the standardized difference can be applied [95].

In order to check the first assumption, the F-statistic value from the first-stage linear regression model is widely used although this statistic highly affected by sample size [76,83,85]. There is a rule of thumb that if the F-statistic value is greater than 10, the first assumption holds [13,96,97]. Other measures for the strength of the association between IV and exposure include the first-stage regression coefficient of the IV [50,98] or the R² of a linear first-stage model [15,78,83], the odds ratio [6,93], or pseudo-R-squared of the first-stage model [76]. When the correlation between IV and exposure is not strong enough, IV analysis is likely to be biased (weak IV bias, which increases with the weakness of the IV). A weak IV will provide large SEs for the IV estimator [3,13,31,47,99].

Discussion

We provided an overview of estimation methods of IV analyses, highlighting their strengths and limitations for epidemiological research. These methods share aspects, yet also have some particularities. However, when the IV assumptions are violated, the sample size is small, or IV models are not correctly specified, all methods tend to perform poorly and show biased results.

The methods can be categorized as moments based and semi-parametric (e.g. 2SLS, GMM, SMM) or likelihood based (e.g. BPM, 2SRL, LSMM) methods. The moment based methods or semi-parametric method are in general less efficient than likelihood based methods. However, likelihood methods are more vulnerable to incorrect modelling assumptions, in which case moment based methods are more robust. In empirical data, although several IV methods can be applicable in the same combination of IV, exposure, and outcome, considering different methods’ assumptions, target parameters being estimated are different, so the interpretations of exposure effects appear different [45]. Therefore, choosing an appropriate IV method requires attention [76].

In order to obtain ACE or LATE or ATT, along with basic assumptions, extra assumptions such as homogeneous exposure effect or monotonicity in case of heterogeneous exposure effect or no effect modification by IV, respectively should be fulfilled. These different assumptions result in the estimation of different causal effects, and hence, researchers should be aware for interpretations of the IV estimates [14].

In randomized trials, the IV of treatment assignment satisfies the assumptions by design, but in observational studies, this is not the case. In the latter situation, subject matter knowledge and theoretical motivations (why is an IV related to treatment and unrelated to patients’ characteristics and outcome?) should be given especially regarding the second and third condition underlying the IV method. If the IV is weakly related to exposure and correlated with unmeasured variables, IV methods may yield biased results [100]. In addition, the main critique of any IV analysis is that the IV may affect the outcome through some pathway other than through the exposure of interest [32]. This condition cannot be verified empirically.

From a methodological perspective, the IV method is a powerful statistical tool, given that a valid IV is present and IV analysis correctly applied. In that case, it can provide a valid estimate in the presence of measured and unmeasured confounding. However, if there is strong confounding effect, it is difficult to find an appropriate IV [13].

A limitation of our study is that we restricted ourselves to IV methods that are commonly used in epidemiologic research. We did not discuss nonparametric and Bayesian IV methods. We refer to the literature for examples of the methods [12,38,86,101-104]. Because of limited space, we did not describe mathematical models with detailed derivation of IV estimators for all methods.

In conclusion, IV analysis is a powerful method to control for confounding (both measured and unmeasured). Some estimation methods (e.g., 2SLS, 2SRI) can be applied in many situations, whereas others (e.g., RE, BPM, 2SLR) can only be applied in a limited number of situations. Irrespective of the methods that are used in a particular study, in order to provide valid interpretation of the exposure effect on the outcome, researchers should be aware of the underlying methodology of the estimation method as well as key assumptions of the IV.

Running Head: Methods for IV estimation

FUNDING

The PROTECT project is supported by the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement no 115004, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution. In the context of the IMI Joint Undertaking (IMI JU), the Department of Pharmacoepidemiology, Utrecht University, also received a direct financial contribution from Pfizer. The views expressed are those of the authors only and not of their respective institution or company.

Conflicts Of Interest: “Olaf Klungel had received unrestricted funding for pharmaco epidemiological research from the Dutch private-public funded Top Institute Pharma (TI Pharma Grant T6.101 Mondriaan).”

References


25. Wald A (1940) The fitting of straight lines if both variables are subject to error. Annals Math Stat 11: 824-300.


