Instrumental variable analysis in the presence of unmeasured confounding

Zhongheng Zhang[1], Md Jamal Uddin[2,3], Jing Cheng[4], Tao Huang[5]

[1]Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China; [2]Department of Statistics, Shahjalal University of Science and Technology, Sylhet, Bangladesh; [3]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark; [4]Division of Oral Epidemiology and Dental Public Health, University of California, San Francisco, USA; [5]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100000, China

Correspondence to: Zhongheng Zhang. Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, No. 3 East Qingchun Road, Hangzhou 310016, China. Email: zh_zhang1984@zju.edu.cn.

Abstract: Observational studies are prone to bias due to confounding either measured or unmeasured. While measured confounding can be controlled for with a variety of sophisticated methods such as propensity score-based matching, stratification and multivariable regression model, the unmeasured confounding is usually cumbersome, leading to biased estimates. In econometrics, instrumental variable (IV) is widely used to control for unmeasured confounding. However, its use in clinical researches is generally less employed. In some subspecialties of clinical medicine such as pharmacoepidemiological research, IV analysis is increasingly used in recent years. With the development of electronic healthcare records, more and more healthcare data are available to clinical investigators. Such kind of data are observational in nature, thus estimates based on these data are subject to confounding. This article aims to review several methods for implementing IV analysis for binary and continuous outcomes. R code for these analyses are provided and explained in the main text.

Keywords: Instrumental variable (IV); confounding; probit regression; two-stage least square

Submitted Jan 29, 2018. Accepted for publication Mar 20, 2018.
doi: 10.21037/atm.2018.03.37
View this article at: http://dx.doi.org/10.21037/atm.2018.03.37

Introduction

While randomized controlled trials (RCTs) are able to balance both measured and unmeasured confounders between comparison groups by the mechanism of randomization, observational studies usually suffer from confounding effects. Confounders are variables associated with both the assignment of treatment and the outcome (1). When the distributions of confounders are not balanced between treated and control groups in an observational study, the estimated treatment effect can be biased. Various methods have been commonly used to account for measured confounder such as matching, multivariable regression adjustment, stratification and so on (2). However, these methods cannot address the problem of unmeasured confounding, which is not uncommon in clinical researches.

Instrumental variable (IV) analysis is a method widely used in econometrics and social sciences, to account for unmeasured confounding (3). The IV is a variable associated with the treatment assignment, it affects the outcome only through the exposure and it is independent of confounders (4,5). Angrist and Krueger (1991) and others (3,5) provided a good review of applications of the IV method. One challenge in IV analysis is to choose a good IV in a real clinical study. In this paper, we will suggest a systematic strategy for addressing this challenge and introduce easily implemented step-by-step practical advice to perform IV analyses in real studies. Simulations, R codes and real examples in clinical research with the IV approach will be discussed and compared with regular analyses.

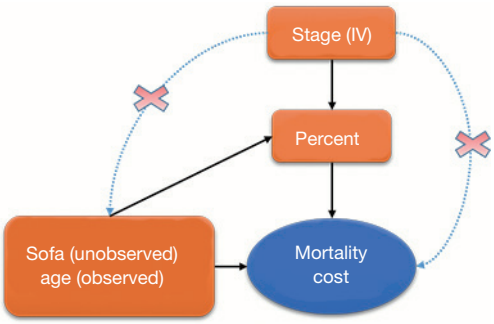

Figure 1 Schematic representation of the relationship between variables. IV, instrumental variable.

Working example

This study is motivated by a real multicenter study we have. The real study investigated the effectiveness of early enteral nutrition (EN) on recovery (measured by mortality) and medical costs in 410 critically ill patients (age: 64.71±16.93 years; male 64%) enrolled in ten tertiary care hospitals (6,7). Traditionally, critically ill patients start EN feeding gradually and after 48 hours since admission. However, the late adoption of EN feeding may delay the recovery of patients leading to higher mortality and longer hospital stays. Therefore, the multicenter study aimed to evaluate if early EN would improve patients' recovery and reduce mortality and costs. In stage I of the study between April 2016 and July 2016, the attending physicians at the ten hospitals provided usual care or EN based on their own preference and local guidance. Two hundred and thirty-six patients were enrolled in the first stage. In the training period between August 2016 and September 2016, all the physicians, nurses, and dieticians at the ten hospitals received 2-month training of using standardized EN feeding protocol within 48 hours after admission to the hospital. Then in stage II between September 2016 and January 2017, all the ten hospitals fully implemented the standardized EN feeding protocol. 147 patients were enrolled and received the standardized EN protocol in the second stage.

Early EN feeding was measured as the percentage of actual daily EN feeding accounting for the total daily requirement target. Age and severity of patients' illness are important predictors known to patients' recovery and can confound the effect of EN feeding if they are not balanced between the comparison groups (*Figure 1*). In the EN feeding study, severity of illness was measured by the sequential organ failure assessment (SOFA) score, which ranged from 0–24 points with higher values indicating severer illness. To illustrate the IV analysis, we suppose that patients' age was known but the severity of illness (SOFA) was not measured. When the important confounder illness severity is not measured, IV methods are very helpful for obtaining consistent estimates for effects of early EN feeding on recovery when a valid and strong IV can be found. In an IV analysis, we first would like to have a variable that satisfies the three key features as an IV: relevance, effective random assumption and exclusion restriction (ER) (8). The stage of the EN feeding study is a change in feeding guideline over time and can serve as a natural IV in this study. Baiocchi *et al.* [2014] discussed some examples using calendar time as the IV, such as the sharply decreased use of hormone replacement therapy (HRT) in 2002 (8). In the EN feeding study, the stage directly affected the use of standardized EN feeding protocol such that the standardized EN feeding (treatment) was sharply increased in stage II (relevance). The change in the feeding practice occurred in a relatively short period of time after a 2-month training period, and there were no notable changes in other medical practices and medical coding systems during the same time period, so the stage of the study seems independent of unmeasured confounders (effective random assumption) and we don't expect direct effects of the stage time on recovery other than through its effect on the EN feeding practice (ER).

Given that the stage seems a valid IV, we would like to assess its strength on the choice of different feeding practices (treatment). An IV is weak if it only has a slight impact on the treatment choice. A weak IV may lead to a treatment estimate with large variance and sensitive to a slight departure from the three IV assumptions. In the EN feeding study, approximately 40% and 50% of patients took the EN feeding within 48 hours in stages I and II, respectively. The stage had a big impact on the choice of treatment and should work well as a strong IV.

In this study, mortality and medical cost are binary and continuous outcomes of interest, respectively. We will model the two outcomes with appropriate models.

To illustrate the IV analysis, we generated the data in R (version 3.3.2). As discussed above, we will use stage as the IV and evaluate the effects of percentage of actually EN feeding ("percent") on mortality ("mort") and medical costs ("cost") with age as a measured confounder and severity ("sofa") as an unmeasured confounder (*Figure 1*).

Table 1 Effect on mortality: comparison between true model and model without sofa

Variables	True model, mean (95% CI)	Model without sofa, mean (95% CI)	Bias
Intercept	−0.85 (−1.63, −0.07)	−0.03 (−0.57, 0.52)	0.96
Percent	−0.97 (−1.48, −0.46)	−1.22 (−1.70, −0.75)	0.26
Age	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)	0
Sofa	0.07 (0.02, 0.12)	NA	NA

```
> set.seed(123)
> n <-1000
> stage <- rbinom(n,1,0.5)
> age <- rnorm(n,68,20)
> sofa <- round(rnorm(n,10,3))#0-24
> percent <- 0.5*stage+ rnorm(n,0.7,0.01)-
  0.001*age-0.03*sofa
> linpred <- 0.1*sofa-percent+0.01*age-1.1
> prob <- exp(linpred)/(1+exp(linpred))
> mort<-ifelse(runif(n,0,1)<prob,1,0)
> cost<-10*sofa-40*percent+age+rnorm(n,100,20)
> dt<-data.frame(stage=stage,percent=percent,
   age=age,mort=mort,
cost=cost)
```

Assuming that *sofa* was not measured in the study, we could estimate the treatment effect with a regular (generalized) linear model with sofa omitted and compare this ordinary estimated effect with the true effect for bias.

```
> mod.true<-glm(mort~percent+age+sofa,
  family = "binomial",dt)
> mod.true.conf<-confint(mod.true)
> mod.biased<-glm(mort~percent+age,
  family = "binomial",dt)
> mod.biased.conf<-confint(mod.biased)
> tab<-cbind(coef(mod.true), mod.true.conf,
  c(coef(mod.biased),NA),
  rbind(mod.biased.conf,c(NA,NA)))
> colnames(tab)<-c("true","loCI","hiCI",
  "biased","loCI","hiCI")
> round(tab,2)
             true  loCI  hiCI biased  loCI  hiCI
(Intercept) -0.85 -1.63 -0.07  -0.03 -0.57  0.52
percent     -0.97 -1.48 -0.46  -1.22 -1.70 -0.75
age          0.01  0.00  0.02   0.01  0.00  0.02
sofa         0.07  0.02  0.12     NA    NA    NA
```

The above code fit two models. The *mod.true* model considers the unmeasured confounder and gives the true effect of *percent* on *mort*. The exponentiation of the coefficient of treatment "percent" (-0.97) gives an odds ratio of 0.38, indicating that the increase of one unit of the percentage of EN is associated with lower odds of death. The second model *mod.biased* omits the unmeasured *sofa* variable, which is commonly done in observational studies when there are unmeasured confounders. It is shown that the second model overestimates the treatment effect by approximately 20% in linear predictor scale (*Table 1*). The effect of *percent* on *cost* is estimated in the same way:

```
> mod.cost<-lm(cost~percent+age+sofa,dt)
> mod.cost.conf<-confint(mod.cost)
> mod.cost.biased<-lm(cost~percent+age,dt)
> mod.cbiased.conf<-confint(mod.cost.biased)
> tab1<-cbind(coef(mod.cost), mod.cost.conf,
  c(coef(mod.cost.biased),NA),
  rbind(mod.cbiased.conf,c(NA,NA)))
> colnames(tab1)<-c("true","loCI","hiCI",
  "biased","loCI","hiCI")
> round(tab1,2)
             true   loCI   hiCI biased   loCI   hiCI
(Intercept) 93.99  86.38 101.59 211.90 202.84 220.96
percent    -36.61 -41.57 -31.65 -75.23 -83.05 -67.40
age          1.08   1.01   1.14   1.11   1.01   1.22
sofa         9.82   9.37  10.27     NA     NA     NA
```

The *cost* is a continuous dependent variable and thus is fitted with linear regression model. While the full model estimated an effect size of -36.61, the biased model gives a coefficient of -75.23. The biased model overestimated the

Table 2 Effect on cost: comparison between true model and model without sofa

Variables	True model, mean (95% CI)	Model without sofa, mean (95% CI)	Bias
Intercept	93.99 (86.38, 101.59)	211.90 (202.84, 220.96)	1.25
Percent	−36.61 (−41.57, −31.65)	−75.23 (−83.05, −67.40)	1.05
Age	1.08 (1.01, 1.14)	1.11 (1.01, 1.22)	0.03
Sofa	9.82 (9.37, 10.27)	NA	NA

saved cost with increasing EN feeding percentage (*Table 2*).

Coefficients in the true model deviate from the true value because of random errors in the simulation process. Since only one set of data is observed in real clinical research, one randomly generated dataset will be employed for illustration purpose. Ideally, the true coefficient values can be obtained from the following R code with 1,000 simulations. If the following chunk is run, the subsequent results will not be exactly the same as that shown in this article, because the randomly generated sample which would be used in subsequent sections is not the same as that generated in the working example section. Thus, we do not suggest to run this chunk.

```
> set.seed(123)
> n <-1000
> sim <- 1000
> true.coef.mort   <- matrix(NA,sim,4)
> biased.coef.mort  <- matrix(NA,sim,3)
> true.coef.cost    <- matrix(NA,sim,4)
> biased.coef.cost  <- matrix(NA,sim,3)
> for (i in 1:sim){
  stage <- rbinom(n,1,0.5)
  age   <- rnorm(n,68,20)
  sofa  <- round(rnorm(n,10,3))#0-24
  percent <- 0.5*stage  - 0.001*age -
   0.03*sofa + rnorm(n,0.7,0.01)
  linpred <- 0.1*sofa - percent +
   0.01*age - 1.1
  prob <- exp(linpred)/(1+exp(linpred))
  mort <- ifelse(runif(n,0,1) < prob,1,0)
  cost <- 100 + 10*sofa -
   40*percent + age + rnorm(n,0,1)
  dt <- data.frame(stage=stage,
              percent=percent,
              age=age,
              mort=mort,
              cost=cost)
  mort.true <- glm(mort ~ percent + age + sofa,
              family = "binomial",dt)
  true.coef.mort[i,] <- coefficients(mort.true)
  mort.bias <- glm(mort ~ percent + age ,
              family = "binomial",dt)
  biased.coef.mort[i,] <- coefficients(mort.bias)
  true.lm.cost <- lm(cost ~ percent +
               age + sofa, data=dt)
  true.coef.cost[i,] <- coefficients(true.lm.cost)

  bias.lm.cost <- lm(cost ~ percent + age, data=dt)
  biased.coef.cost[i,] <- coefficients(bias.lm.cost)
  }

> cat("Coef. for the true mortality model:,\n",
    colMeans(true.coef.mort),"\n",
  "Coef. for the biased mortality model:,\n",
  colMeans(biased.coef.mort),"\n",
  "Coef. for the true cost model:,\n",
  colMeans(true.coef.cost),"\n",
  "Coef. for the biased cost model:,\n",
  colMeans(biased.coef.cost))
```

Manual IV analysis for continuous dependent variable

In this section, we present manual calculation of treatment effect in the presence of unmeasured confounding using IV analysis.

Multiple regression model is commonly used to investigate the effect of a predictor on an outcome, controlling for other measured covariates. Consider a linear regression function:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_P x_{Pi} + \varepsilon_i \quad [1]$$

where i=1,2, ..., n. The equation can be written in matrix

format:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1\, x_{11} \ldots x_{p1} \\ 1\, x_{12} \ldots x_{p2} \\ \vdots \quad \vdots \ldots \vdots \\ 1\, x_{1n} \ldots x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad [2]$$

and can be simply written as: Y =Xβ + ε, where X is a (p+1)*n matrix, Y is a n*1 column vector, β is a (p+1)*1 column vector, and ε is a n*1 column vector. The matrix X and vector β are multiplied using the techniques of matrix multiplication. The vector $X\beta$ is added to the vector ε with matrix addition. The parameter β can be estimated using the following matrix equation:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = (X'X)^{-1}X'Y \quad [3]$$

where X' is the transpose of the X matrix, and $(X'X)^{-1}$ is the inverse of the $X'X$ matrix. The inverse of a square matrix A is a matrix A^{-1} such that the multiplication AA^{-1}= I, where I is an identity matri1x (e.g., a matrix with 1's on the diagonal and 0's elsewhere). In case β is uncorrelated with ε, the estimation in Equation [3] can be unbiased. However, when there is unmeasured confounding, the estimation is biased and we need to introduce instrumental variables Z. As described previously, Z is (I) correlated with X, but (II) not directly correlated with outcome y except for that via the effect of X. Suppose the relationship between X and Z is given by:

$$X = Z\delta + errors$$

The IV analysis is specified with the following equation:

$$\widehat{\beta}_{IV} = (Z'X)^{-1}Z'Y$$

Direct IV analysis can be performed with the following R code:

```
> X<-cbind(1,percent,age)
> Z<-cbind(1,age,stage)
> betaIV<-solve(t(Z)%*%X) %*% t(Z)%*%cost
> betaIV
                    [,1]
                    188.337494356
percent             -39.741849274
age                 1.156065365
```

Two-stage least square (2SLS) method is commonly used for IV analysis (9), which involves two stages. The first stage regress X on Z: $X = Z\delta + errors$. The estimated coefficients $\widehat{\delta} = (Z'Z)^{-1}Z'X$, and the predicted value is estimated as:

$$\widehat{X} = Z\widehat{\delta} = Z(Z'Z)^{-1}Z'X = P_Z X. \quad [4]$$

The second stage is to regress Y on the predicted values from the first stage: $Y = \widehat{X}\beta + \text{errors}$. The estimated 2SLS coefficient can be computed as:

$$\beta_{2SLS} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y = (P_Z'X'P_ZX)^{-1}P_Z'X'Y, \quad [5]$$

Note that P_Z is a symmetric and idempotent matrix that P_ZP_Z' =P_ZP_Z = P_Z. Equation [5] can be written as:

$$\widehat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_ZY = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y$$

In the artificial example, X = [*age*, *percent*] and Z = [*stage*, *age*]. Recall that confounders other than the IV should be included in the first stage model on treatment in addition to the IV, so the covariate age was included in the vector Z in the first stage model. The 2SLS estimation can be performed using the following R code:

```
> X<-cbind(1,percent,age)
> Z<-cbind(1,age,stage)
> beta2sls<-solve(t(X)%*%Z %*%
  solve(t(Z)%*%Z) %*%
    t(Z)%*%X)  %*%
    t(X)%*%Z %*%
    solve(t(Z)%*%Z) %*%
    t(Z)%*%cost
> beta2sls
                    [,1]
                    188.337494356
percent             -39.741849274
age                 1.156065365
```

The solve() function is used to solve an equation a %*% x = b for x, where b can be either a vector or a matrix. The inverse of a matrix is to solve an equation with b = identity matrix. In the solve() function, b is taken to be an identity matrix by default (e.g., b argument is missing), which is the case in the above example. The result shows that the coefficient for percent is –39.7, which approximates the true effect size of –36.6. In case the number of instruments is equal to the number of endogenous predictors (e.g., an endogenous variable is correlated with both the independent variable in the model, and with the error term), direct IV estimation is

equivalent to the 2SLS method (the model is just identified).

Analysis with ivreg() function

IV analysis with continuous outcome can be easily performed using the ivreg() function in the AER package (version 1.2-5) (10).

```
> library(ivpack)
> ivmod<-ivreg(cost~percent+
  age|stage+age,
  data=dt,x=T)
> sum.stats<-summary(ivmod,vcov = sandwich,
      diagnostics = TRUE)
```

The ivreg() function fits IV regression by using 2SLS, which is equivalent to direct IV estimation when the number of instruments is equal to the number of predictors. The first argument specifies a formula with the regression relationship and the instruments. The formula has three parts in the form of y ~ x1 + x2 | z1 + z2 + z3, where y is the outcome variable, *x*s are endogenous variables and zs are instruments. Exogenous variables (e.g., a variable which is unaffected by other variables within an model) such as "age" in our example should be included in both sides of "|" symbol. The returned values can be extracted with the summary() function. Diagnostics of the model is returned by setting diagnostics = TRUE.

```
> round(sum.stats$coefficients,3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	188.337	4.918	38.294	0
percent	−39.742	4.386	−9.061	0
age	1.156	0.055	21.043	0

The results show that the coefficient for the variable percent is −39.7, which is exactly the same with that obtained by manual calculation.

```
> sum.stats$ diagnostics
```

	df1	df2	Statistic	P value
Weak instruments	1	997	8055.882818	0.000000000e+00
Wu-Hausman	1	996	1714.383775	9.676714954e-219
Sargan	0	NA	NA	NA

The diagnostic statistics of IV analysis are shown in the above output. The strength of the IV can be evaluated with the partial first-stage F statistic. An IV is considered as a weak IV if the partial F statistic is less than 10 in the first stage model, that is, its impact on the choice of treatment is weak. A weak IV could lead to a larger variance in the coefficient, and severe finite-sample bias (11). The partial first-stage F statistic of 8056 indicates that stage is a strong IV in the study. Durbin-Wu-Hausman test compares the ordinary least square (OLS) estimate versus the IV estimate of the treatment effect assuming homogeneous treatment effects, that is, the treatment effect is the same at different levels of covariates (12,13). The rejection of the null hypothesis can be due to unmeasured confounding or heterogeneous treatment effects. Alternatively, Guo *et al.* [2014] proposed a test with an IV for unmeasured confounding, which distinguishes from treatment effect heterogeneity (14). There was no evidence of unmeasured confounding in the EN feeding study ($P<0.001$ for Durbin-Wu-Hausman test). Sargan is a test of instrument exogeneity using overidentifying restrictions, called the J-statistic in Stock and Watson. In case when there are more instruments than endogenous variables, the model is overidentified, and we have some excess information. To have consistent treatment estimates, all the IV should be valid. The test examines if all the IVs are in fact exogenous, that is, uncorrelated with the model residuals. The rejection of the null hypothesis of this global test indicates that at least one IV is invalid (15). Sargan test works when there are more IVs than endogenous variables, thus it is not applicable in our example.

Logistic regression for binary outcome

In case when the outcome variable is binary, IV analysis can be performed by fitting generalized model to the binary outcome. Probit and Logistic regression models are most commonly used (16). In this section, we will introduce the logistic regression model approach and probit model will be introduced in the next section. In analogy with the 2SLS method, two-stage predictor substitution (2SPS) method can be applied when the outcome is a binary variable (17). 2SPS works by first regressing the treatment X on the IV Z and observed exogenous covariates X_e, obtaining predicted $\widehat{X}$ and then fit a logistic regression of Y on $\widehat{X}$ and X_e. R code for performing the procedure is as follows:

```
> s1<-lm(percent~stage+age,data = dt)
> phat<-predict(s1)
> s2.2sps<-glm(mort~phat+age,#2SPS
```

```
    data=dt,
    family = "binomial")
```

The first line regresses variable "percent on stage" (Z) and "age" (exogenous covariate X_e). Then the predict function is employed to estimate the predicted percent ($\widehat{X}$). In the second stage, the binary variable "mort" is regressed on the predicted "percent" and "age" in a logistic model. By default, the logit link function is used for binomial distribution outcome.

```
> round(summary(s2.2sps)$coefficients,4)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	–0.1831	0.2825	–0.6481	0.5169
phat	–0.9726	0.2549	–3.8161	0.0001
age	0.0112	0.0033	3.4455	0.0006

The variable "phat" in the above table is the predicted value of "percent" by the model *s1*. The results show that the coefficient of percent is –0.97, which is consistent with the true value of –0.97. However, this method is biased due to the non-collapsibility of the logistic regression (18).

Alternatively, the two-stage residual inclusion (2SRI) method can be used to estimate the true effect of an endogenous treatment. The first stage is the same as that in 2SPS. The second stage fits a logistic regression model for Y on X_e, X and the residual from the first stage. The estimated coefficient for X in second stage is the estimate of treatment effect.

```
> s2.2sri<-glm(mort~percent+#2SRI
    age+residuals.lm(s1),
    data=dt,
    family = "binomial")
> round(summary(s2.2sri)$coefficients,4)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	–0.1865	0.2854	–0.6537	0.5133
percent	–0.9915	0.2575	–3.8503	0.0001
age	0.0115	0.0033	3.4762	0.0005
residuals.lm(s1)	–2.2468	0.7793	–2.8833	0.0039

The residual of the first stage linear regression is obtained using the residuals.lm() function. The results show that the effect of treatment deviates a little from the true effect (–0.99 *vs.* –0.97), but is much better than the naïve logistic regression model (–1.22). Similar to 2SPS, 2SRI has been shown to be asymptotically biased except when there is no unmeasured confounding (18).

Control function approach for binary outcome

Blundell and Powell proposed a control function approach to deal with endogeneity (19). Similar to the 2SRI method, the first step of the control function approach regresses treatment X on the IV Z and observed exogenous covariates X_e, then collects the residuals $\widehat{v}$. The second step estimates the probit model of interest, by including the first stage residuals $\widehat{v}$ as an additional regressor. This method is termed the control function approach, as the inclusion of $\widehat{v}$ controls for the correlation between v and ε, where ε is the structural error term in the $Y = X\beta + \varepsilon$ equation.

Average structural function (ASF) is the probability of response variable given values of regressors, in the absence of endogeneity.

$$P(y \mid \overline{X}, v) = \widehat{ASF} = \frac{1}{N}\sum_{i=1}^{N}\phi(\overline{X}\hat{\beta} + \rho\hat{v}_i)$$

Where $\overline{X}$ is the mean of a vector of covariates including the treatment variable, $\widehat{\beta}$ is the estimated coefficient in the second stage, ρ is the estimated coefficient for $\widehat{v}$ in the second stage. $\widehat{v}_i$ is the residual obtained from the first stage for each patient (i). Φ() is a function that transforms the linear predictor into probability scale. Then, ASF is the average of predicted probability. In the example, the treatment variable is allowed to vary across a range so that its effect on the probability of response variable can be shown. Suppose the variable "sofa" is known, then the true ASF can be obtained.

```
> dat <- data.frame(cbind(mean(age),
    mean(sofa),
    seq(min(percent),
    max(percent),length.out=50)))
> names(dat) <- c("age","sofa","percent")
> lprd<- -1.1 + 0.01*dat$age -
  dat$percent+0.1*dat$sofa
> dat$asf <- exp(lprd)/(1+exp(lprd))
```

The above code holds covariates at their mean values and the treatment variable "percent" is allowed to vary between maximum and minimum values. A total of 50 values were generated. The coefficients used for data generating are used here to compute linear predictor (*lprd*). The linear predictor is then transformed into probability with logit

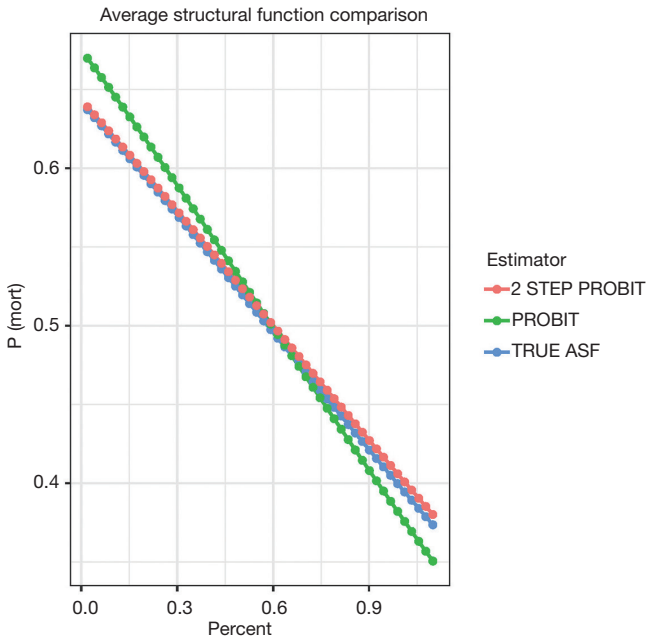

Figure 2 The probability of mortality against "percent" obtained with three methods. The result shows that while the two-step probit model is consistent with the true model, the naïve probit model is biased. ASF, average structural function.

transformation. Suppose the "sofa" is not known in the study, and we perform naïve probit regression model. The estimated ASF can be obtained with the following code.

```
> r1 <- glm(mort~percent+age,
  binomial(link="probit"))
> dat1 <- data.frame(cbind(mean(age),
  seq(min(percent),
  max(percent),length.out=50)))
> names(dat1) <- c("age","percent")
> asf1 <- cbind(dat1$percent,
      predict(r1,dat1,
      type = "response"))
```

The above computation is the same as the estimation of the true ASF, except that the variable "sofa" is unknown. The "percent" varies in the same range.

Then we proceed to estimate treatment effect with the two-step control function approach.

```
> s1<-lm(percent~stage+age,data = dt)
> v1 <- (residuals(s1))/
  sd(residuals(s1))
> r2 <- glm(mort~age+percent+v1,
  binomial(link="probit"))
```

The two-stage control function approach is similar to 2SRI method except that the residual is scaled in the example. The following code is to estimate ASF:

```
> asf2 <- cbind(seq(min(percent),
     max(percent),length.out = 50),NA)
> for(i in 1:dim(asf2)[1]){
  dat2 <- data.frame(cbind(mean(age),asf2[i,1],v1))
  names(dat2) <- c("age","percent","v1")
  asf2[i,2] <- mean(predict(r2,dat2,
       type = "response"))
}
```

The for() function is used to loop through all 50 values of percent. Within each loop, there is 1,000 v1 values and thus 1,000 predicted response values. We need to take the mean of these response values to obtain the mean probability for each given value of "percent".

```
> plotdat <- data.frame(rbind(cbind(dat$percent,
   dat$asf,"TRUE ASF"),
   cbind(dat$percent,asf1[,2],"PROBIT"),
   cbind(dat$percent,asf2[,2],"2 STEP PROBIT")))
> names(plotdat) <- c("percent","mort","Estimator")
> plotdat$percent <- as.numeric(as.character(plotdat$percent))
> plotdat$mort <- as.numeric(as.character(plotdat$mort))
> library(ggplot2)
> ggplot(plotdat, aes(x=percent, y=mort,
  colour = Estimator, group=Estimator)) +
  geom_line(size=0.8) + geom_point()+
  scale_x_continuous('percent') +
  scale_y_continuous('P(mort)') +
  theme_bw() +
  labs(title = "Average Structural Function Comparison")
```

Figure 2 plots the probability of mortality against "percent". The result showed that while the two-step probit model is consistent with the true model, the naïve probit model is biased. Also note the three lines are straight, which is attributable to the range of percent is restricted between 0 and 1. If the range of percent is extended, between –10 and

10, for example, the shape of these curves will be sigmoid.

Local average response function (LARF)

Theoretical basis of the LARF method was developed by Abadie (20). The method involves two steps: the first step constructs pseudo-weights according to the probability of receiving the treatment instrument; and the second step involves the computation of LARF conditional on covariates and treatment by using pseudo-weights. Mathematical details of the LARF is not reviewed here and readers can consult the references (20,21). In this section, we focus on how to implement the LARF method to estimate treatment effect in the presence of endogeneity. The LARF package (version: 1.4) is employed for this purpose (21).

```
> library(LARF)
> mod.cont<-larf(cost~age,treatment = dt$percent,
   instrument = dt$stage,data = dt)
> summary(mod.cont)$coefficients
```

	Estimate	SE	P
Treatment	−75.225913	3.70236834	4.54361e-77
(Intercept)	211.899912	4.54796896	1.59897e-252
age	1.111813	0.05371983	1.91816e-79

The result is not as expected. The value −75.2 is the estimate consistent with the naïve biased estimate. Since the LARF requires treatment variable to be binary, we simulate the sample in another way.

```
> set.seed(123)
> trtstar<-4*stage-
 0.01*age-0.2*sofa+
 rnorm(n,0.6,0.01)
> dt$trt<-rbinom(n,1,pnorm(trtstar))
> mortstar <-0.2*sofa-
 4*dt$trt+
 0.01*age-rnorm(n,0.7,0.1)
> dt$mort.trt<-rbinom(n,1,pnorm(mortstar))
> dt$cost.trt<-30*sofa-
 40*dt$trt+age+
 rnorm(n,100,2)
> lm.trtT<-lm(cost.trt~age+trt+sofa,dt)
> lm.trtB<-lm(cost.trt~age+trt,dt)
> tab.trt<-cbind(coef(lm.trtT),
   c(coef(lm.trtB),NA))
> tab.trt
```

	[,1]	[,2]
(Intercept)	100.2314586	392.725127
age	0.9982993	1.234915
trt	−40.0556142	−59.944512
sofa	29.9892341	NA

```
> mod.cont<-larf(cost.trt+trt~age|stage,
   data = dt)
> summary(mod.cont)$coefficients
```

	Estimate	SE	P
Treatment	−45.288400	6.1535971	3.850071e-13
(Intercept)	392.687192	11.2744649	1.553136e-174
age	1.077041	0.1577494	1.498830e-11

The use of LARF method appears to adjust some of the biases induced by ignoring a confounding factor *sofa*. The true effect is −40.5, the naïve biased estimate is −59.9, and the LARF result is −45.3. The result is reasonable for continuous outcome *cost.trt*. However, it is not the case for binary outcome *mort.trt*.

```
> glmT<-glm(mort.trt~trt+age+sofa,
   data=dt,
   family = 'binomial'(link = "probit"))
> glmB<-glm(mort.trt~trt+age,
    data=dt,
    family = 'binomial'(link = "probit"))
> tabmort<-cbind(coef(glmT),
          c(coef(glmB),NA))
> tabmort
```

	[,1]	[,2]
(Intercept)	0.160385237	1.42456752
trt	−4.063437158	−3.82182534
age	0.006896202	0.00775861
sofa	0.146833455	NA

```
> mod.bi<-larf(mort.trt~age,treatment = dt$trt,
  instrument = dt$stage,data = dt)
> round(summary(mod.bi)$coefficients,4)
```

	Estimate	SE	P	MEM	M E M - SE	MEM-P
Treatment	−3.7122	0.1630	0.0000	−0.9365	0.0116	0.0000

(Intercept)	1.5689	0.2057	0.0000	0.4530	0.0436	0.0000
age	0.0045	0.0021	0.0381	0.0018	0.0009	0.0385

The results of LARF method seems to bias the effect size more than the naïve regression model. Possibly, estimation with one random dataset may cause bias and simulation with 1000 or more times may be necessary in order to obtain the true estimates.

Discussion

In this article, we discussed some methods for IV analysis and showed R code for the performance of them. The manual analysis is complex and requires some knowledges on matrix manipulation. This is not suitable for research practice but can help to understand how IV analysis works. 2SLS method is commonly used for IV analysis. It is a natural starting point of IV analysis, and the estimate is asymptotically unbiased. However, 2SLS estimate can be biased in binary cases or in the case of non-linear models. 2SPS can be applied when the outcome is a binary variable. However, 2SPS in non-linear model does not always yield consistent exposure effects on the outcome, and parameter estimation process is more difficult than 2SLS. 2SPS may not provide causal OR under a logistic regression model. 2SRI is able to yield consistent estimates for both linear and non-linear models. It performs better than 2SPS. 2SPS is suitable in the case of a binary exposure with a binary or count outcome. LARF is suitable for estimating treatment effect when both the treatment and its instrument are binary.

Acknowledgements

Funding: This study was supported by funding from Zhejiang Provincial Natural Science Foundation of China (LGF18H150005) to Z Zhang.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Grobbee DE, Hoes AW. Clinical Epidemiology: Principles, Methods, And Applications For Clinical Research. 1st edition. Jones & Bartlett Learning, 2008.
2. Klungel OH, Martens EP, Psaty BM, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004;57:1223-31.
3. Angrist JD, Imbens GW. Identification and Estimation of Local Average Treatment Effects. Cambridge, MA: National Bureau of Economic Research, 1995.
4. Uddin MJ, Groenwold RH, Ali MS, et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. Int J Clin Pharm 2016;38:714-23.
5. Klungel OH, Jamal Uddin M, de Boer A, et al. Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods. Pharm Anal Acta 2015;6:353.
6. Zhang Z, Li Q, Jiang L, et al. Effectiveness of enteral feeding protocol on clinical outcomes in critically ill patients: a study protocol for before-and-after design. Ann Transl Med 2016;4:308-8.
7. Li Q, Zhang Z, Xie B, et al. Effectiveness of enteral feeding protocol on clinical outcomes in critically ill patients: A before and after study. PLoS One 2017;12:e0182393.
8. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med 2014;33:2297-340.
9. James LR, Singh BK. An introduction to the logic, assumptions, and basic analytic procedures of two-stage least squares. Psychol Bull 1978;85:1104-22.
10. Kleiber C, Zeileis A. Applied Econometrics with R. Springer, 2008.
11. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 2002;20:518-29.
12. Wu DM. Alternative Tests of Independence between Stochastic Regressors and Disturbances: Finite Sample Results. Econometrica 1974;42:529-46.
13. Hausman JA. Specification Tests in Econometrics. Econometrica 1978;46:1251-71.
14. Guo Z, Cheng J, Lorch SA, et al. Using an instrumental variable to test for unmeasured confounding. Stat Med 2014;33:3528-46.
15. Small DS. Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions. J Am Stat Assoc 2007;102:1049-58.
16. Clarke PS, Windmeijer F. Instrumental Variable Estimators for Binary Outcomes. J Am Stat Assoc 2012;107:1638-52.
17. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric

modeling. J Health Econ 2008;27:531-43.
18. Cai B, Small DS, Have TR. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. Stat Med 2011;30:1809-24.
19. Blundell RW, Powell JL. Endogeneity in Semiparametric Binary Response Models. Rev Econ Stud 2014;71:655-79.
20. Abadie A. Semiparametric instrumental variable estimation of treatment response models. J Econom 2003;113:231-63.
21. An W, Wang X. LARF: Instrumental Variable Estimation of Causal Effects through Local Average Response Functions. J Stat Softw 2016;71:1-13.

Cite this article as: Zhang Z, Uddin MJ, Cheng J, Huang T. Instrumental variable analysis in the presence of unmeasured confounding. Ann Transl Med 2018;6(10):182. doi: 10.21037/atm.2018.03.37

