Instrumental Variable Analysis for Estimation of Treatment Effects with Dichotomous Outcomes

Ehsan Karim

March 14, 2010

Abstract

This report based on is organized in sections in accordance with the given objectives:

1. Explaining the ideas of instrumental variables.
2. For continuous outcome variable, explaining how and why instrumental variable methods work.
3. Explaining the problems caused by the use of a dichotomous outcome.
4. Showing the performance of any two instrumental variable methods using simulation, when the outcome variable is binary.

1 Instrumental Variable

Instrumental variables are commonly used in the estimation of the causal effects, especially when random experiments are not feasible. The use of instrumental variables are very popular because the required assumptions are often plausible to justify, compared to other regression approaches. Another place of usefulness of these instrumental variables is in the analysis of measurement error.

By definition, an instrument is a variable that

- is correlated with (endogenous[1]) explanatory variable $\mathcal{T}$ or predictor of $\mathcal{T}$,
- but itself unassociated with the disturbance term ϵ of the prediction equation,

[1] In the system of simultaneous equations, the endogenous variables are the dependent variables that are determined by the system. However, they may act as explanatory variables in some other equation in the same system. On the other hand, exogenous variables are determined outside of the system, which means the system is not really interested about their cause - treating them as predetermined variables.

- also unassociated with the outcome $\mathcal{Y}$, except via the effect of endogenous explanatory variable $\mathcal{T}$.

For sake of simplification, it is also assumed that there is no treatment effect heterogeneity. However, if desired, one can deal with the treatment effect heterogeneity.

However, researcher must be cautious in choosing instrument variable $\mathcal{Z}$. Choosing weak instrument, i.e., poor predictor of endogenous explanatory variable $\mathcal{T}$ will not have much success in predicting the outcome variable of interest, $\mathcal{Y}$. Efficiency will go down as a result of poor choice of instruments. This will have even more detrimental effects in small sample situations. This will be discussed in details in the later sections.

2 Instrumental Variable Methods

2.1 Structural Equation Model

Structural equation models are used to model instrumental variables. This model is easier to conceptualize using two separate equations. Usually the first equation is concerned with the treatment assignment mechanism $\mathcal{T}$, that is modelled as a function of some observed variables. These variables are the observed confounders, $\mathcal{C}$ and instrumental variables or instruments, $\mathcal{Z}$. Then, the next model defines outcome $\mathcal{Y}$ as a function of the treatment $\mathcal{T}$ and the observed confounders $\mathcal{C}$.

There are several methods of analyzing structural variable models. Linear structural models[2] are very popular due to their simplest form and easy interpretation. The above models can be easily expressed as follows:

$$\mathcal{T} = \tau_{00} + \tau_{01}\,\mathcal{Z} + \tau_{02}\,\mathcal{C} + \epsilon_0. \tag{1}$$

$$\mathcal{Y} = \tau_{10} + \tau_{11}\,\mathcal{T} + \tau_{12}\,\mathcal{C} + \epsilon_1. \tag{2}$$

Here τ_{11} has the interpretation of causal treatment effect. However, in presence on unmeasured confounders $\mathcal{U}$, the disturbance terms ϵ_0 and ϵ_1 from these two models (1 and 2) can be serially correlated. As a consequence of that, the estimate of τ_{11} is usually biased.

In the above equations, $\mathcal{T}$ is an endogenous explanatory variable, since it is correlated with the disturbance term ϵ_0 through model (1). But $\mathcal{Z}$ is a valid instrumental variable

[2]Nonlinear models, such as marginal models can also be used.

since it is correlated with $\mathcal{T}$, but uncorrelated with the disturbance term ϵ_0. For identification purpose, it is required that there is at least as many instrumental variable as endogenous explanatory variable.

2.2 Two-stage least squares

2.2.1 Motivation

One of the key assumption of ordinary least squares method is that the disturbance term and covariates are uncorrelated. In econometric term, this is called exogeneity assumption. However, there are several circumstances, when this assumption does not hold (in econometric term, this phenomenon is known as endogeneity):

- if some important covariate is omitted (bias due to omitted or unmeasured variable). The generalized form of this is model mis-specification.
- if outcome variable is also dependent on some of the covariates (bias due to simultaneity), or if the independent variable is not fixed.
- if the covariates are subject to measurement error.
- if there exists sample selection bias.

The immediate consequence due to this assumption violation is to get estimates from the regression which are biased and inconsistent.

The simplest illustration can be as follows: let the regression model under consideration is $\mathcal{Y} = \tau_{11}\,\mathcal{T} + \epsilon$, where the disturbance term ϵ and covariate $\mathcal{T}$ are correlated. Then the estimate of τ_{11} by ordinary least squares method will be

$$\begin{aligned}
\hat{\tau_{11}} &= (\mathcal{T}'\mathcal{T})^{-1}(\mathcal{T}'\mathcal{Y}) \\
&= (\mathcal{T}'\mathcal{T})^{-1}\big(\mathcal{T}'(\tau_{11}\,\mathcal{T} + \epsilon)\big) \\
&= (\mathcal{T}'\mathcal{T})^{-1}\big(\mathcal{T}'(\tau_{11}\,\mathcal{T})\big) + (\mathcal{T}'\mathcal{T})^{-1}\big(\mathcal{T}'\epsilon\big) \\
&= \tau_{11} + (\mathcal{T}'\mathcal{T})^{-1}(\mathcal{T}'\epsilon) \qquad (3)
\end{aligned}$$

Due to the fact that ϵ and $\mathcal{T}$ are correlated, the estimate of τ_{11} from ordinary least square is going to be biased. Even after controlling for some measured confounder $\mathcal{C}$, the unbiasedness is not achieved due to the assumption violation. Also, with the correlation, since the last term of (3) does not converge to zero even with increasing sample size, the estimate is also inconsistent. This is the general consequence of endogeneity.

2.2.2 Procedure

There are several approaches used in case of endogeneity, but Two-stage least squares is the most popular approach that addresses this problem. What is does is as follows:

1. In the first stage it predicts $\hat{\mathcal{T}}$ from model (1) after controlling for $\mathcal{C}$ and $\mathcal{Z}$, which is uncorrelated with unmeasured covariates $\mathcal{U}$ (ϵ_0 is a function of $\mathcal{U}$).

2. Then in the second stage, this unconfounded prediction of $\mathcal{T}$ is used in (2) to model $\mathcal{Y}$. This helps removing bias in estimating τ_{11}.

That way, since $\hat{\mathcal{T}}$ used in second stage is uncorrelated with ϵ_0, there will be no systematic bias due to unmeasured covariates. Also, by removing ϵ_0 from the process, residual sum of squares will be minimized, and hence a better fit is achieved. The higher the R-square in the first model, the better - this means that the instruments $\mathcal{Z}$ used in the analysis are good predictors of endogenous predictors $\mathcal{T}$, which will be in turn used for prediction of final outcome $\mathcal{Y}$.

For simple illustration, let us forget about $\mathcal{C}$ for now, and then the simplest formula would be $\mathcal{T} = \tau_{01}\,\mathcal{Z} + \epsilon_0$. After fitting this, we get $E(\mathcal{T}|\mathcal{Z}) = \hat{\mathcal{T}} = \mathcal{Z}(\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}\mathcal{T}$. Therefore, in the second stage, if we use this conditional values, we get $E(\mathcal{Y}|\mathcal{Z}) = \tau_{11}\,E(\mathcal{T}|\mathcal{Z}) + E(\epsilon_1|\mathcal{Z})$. The estimate of τ_{11} is then

$$\begin{aligned}
\hat{\tau_{11}} &= (\hat{\mathcal{T}}'\hat{\mathcal{T}})^{-1}\hat{\mathcal{T}}'\mathcal{Y} \\
&= (\mathcal{Z}'\mathcal{T})^{-1}(\mathcal{Z}'\mathcal{Y}) \\
&= (\mathcal{Z}'\mathcal{T})^{-1}\big(\mathcal{Z}'(\tau_{11}\,\mathcal{T} + \epsilon_1)\big) \\
&= (\mathcal{Z}'\mathcal{T})^{-1}\big(\mathcal{Z}'(\tau_{11}\,\mathcal{T})\big) + (\mathcal{Z}'\mathcal{T})^{-1}\big(\mathcal{Z}'\epsilon_1\big) \\
&= \tau_{11} + (\mathcal{Z}'\mathcal{T})^{-1}(\mathcal{Z}'\epsilon_1) \qquad (4)
\end{aligned}$$

Here, since $\mathcal{Z}$ and ϵ_1 are uncorrelated, the last term of (4) vanishes, providing an unbiased estimate of τ_{11}. By including $\mathcal{C}$ in the above equation, similar result can be obtained.

Also, by not imposing any distributional assumption on the disturbance terms, the results of this approach are more robust. However, relative measure of effects are not directly achievable since two stage least squares method is basically an additive model. Therefore, risk difference is obtained from such approaches.

2.2.3 Explaining using Graphs

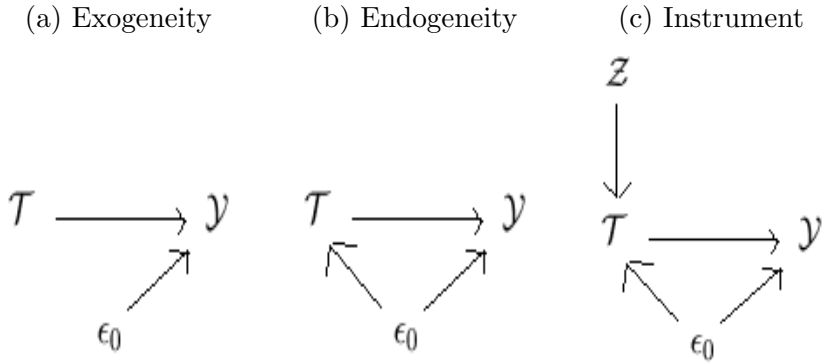
The ordinary least square makes the assumption of exogeneity, that is, in the model of regressor $\mathcal{T}$ predicting outcome $\mathcal{Y}$, the regressor $\mathcal{T}$ and the disturbance term ϵ_0 are un-

correlated. Both regressor and the disturbance term are independent causes of outcome $\mathcal{Y}$. Therefore, the effect of $\mathcal{T}$ on $\mathcal{Y}$ can be estimated without any bias. This scenario is graphically shown in Figure 1 (a).

However, when the regressor $\mathcal{T}$ and the disturbance term ϵ_0 are correlated, then in presence of endogeneity, it is not so clear how much $\mathcal{T}$ is affecting $\mathcal{Y}$, since ϵ_0 is now affecting $\mathcal{T}$ as well, and hence the contribution of only $\mathcal{T}$ is not separable any more. As a result, the bias is introduced in the estimation of treatment effect τ_{11}. This scenario is graphically shown in Figure 1 (b).

Randomization is the standard tool to solve the problem of endogeneity, but in practice, which might not be feasible in many situations. Therefore, to estimate the treatment effects τ_{11} unbiasedly from the non-randomized data, one resorts to instrumental variable $\mathcal{Z}$ that is causally associated with treatment or predictor $\mathcal{T}$, but not with the disturbance term ϵ_0, since there is no direct path between instrument $\mathcal{Z}$ and disturbance term ϵ_0.

Figure 1: Graphical Depiction



Although instrument $\mathcal{Z}$ and outcome $\mathcal{Y}$ both are in the same path, the source of association goes through treatment $\mathcal{T}$. That means, the only way $\mathcal{Z}$ affects $\mathcal{Y}$ is through $\mathcal{T}$. Since instrument $\mathcal{Z}$ is not directly connected with outcome $\mathcal{Y}$, according to backdoor criterion, there is no possibility of confounding due to instrument $\mathcal{Z}$. And hence, causal inference of treatment are not hindered by $\mathcal{Z}$. However, disturbance term ϵ_0 is a cause of $\mathcal{Y}$, and so is $\mathcal{T}$, making $\mathcal{Y}$ a collider and there is a backdoor open for the path between treatment $\mathcal{T}$ and outcome $\mathcal{Y}$. Thus confounding is present in the relationship. If we can somehow break the association path between treatment $\mathcal{T}$ and disturbance term ϵ_0, then the causal inference of treatment effect τ_{11} of treatment $\mathcal{T}$ on outcome $\mathcal{Y}$ will be free from any confounding.

From the relationship of instrument $\mathcal{Z}$ and endogeneous explanatory variable $\mathcal{T}$, we can orthogonalize $\mathcal{T}$ by conditioning on $\mathcal{Z}$. This will block the path of confounding since $\mathcal{Z}$ is neither the collider, nor any descendent of the collider. This will make the predicted treatment $\hat{\mathcal{T}}$ independent of the disturbance term ϵ_0 - hence making the path between treatment $\mathcal{T}$ and $\mathcal{Y}$ free from any confounding. Then we can use this $\hat{\mathcal{T}}$ in estimating its effect on the outcome $\mathcal{Y}$. That way, we can get unbiased estimate of treatment effect τ_{11}, by the help of instrumental variable $\mathcal{Z}$. This is shown graphically in Figure 1 (c).

2.3 Generalized Method of Moments

The generalized method of moment is established based on assumptions about moments, rather than distributional assumptions regarding disturbance terms. Because of this minimal assumption requirement, generalized method of moments are applicable to many practical problems. Also, many estimates such as least squares, maximum likelihood, instrumental variables methods can be seen as special cases of unified framework of generalized method of moments.

Before discussing generalized method of moments (GMM), first a special case - method-of-moment needs to be explained:

2.3.1 Method of Moments

This method dates back to Pearson. The idea is to estimate the mean of a distribution by the sample mean, and the variance by the sample variance, and so on for the higher moments, without much distributional assumption.

For simple illustration, let us consider that μ is the object of interest. Let us solve the sample moment condition

$$\big(\sum_{i=1}^{n}(y_i - \mu)\big)/n = 0,$$

which is sample analog of the population moment condition

$$E\big(y - \mu\big) = 0.$$

This will yield solution as follows:

$$\hat{\mu}_{MM} = \big(\sum_{i=1}^{n} y_i\big)/n,$$

which is the average obtained from the sample. Such estimators that solves sample moment equations to produce estimated are termed as method-of-moments estimators.

One special case can be ordinary least squares, having the moment condition

$$\begin{aligned} E\big(x_i\epsilon_i\big) &= 0 \\ \Rightarrow E\big(x_i'(y_i - x_i\beta)\big) &= 0 \end{aligned}$$

from the assumption of exogeneity in the model $y_i = x_i\beta + \epsilon_i$. Replacing it by sample moment condition, we have

$$\begin{aligned} \frac{\sum_{i=1}^n x_i'(y_i - x_i\hat{\beta})}{n} &= 0 \\ \Rightarrow \frac{\sum_{i=1}^n x_i'y_i}{n} - \frac{\sum_{i=1}^n x_i'x_i\hat{\beta}}{n} &= 0 \\ \Rightarrow \sum_{i=1}^n x_i'y_i = \sum_{i=1}^n x_i'x_i\hat{\beta} \\ \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^n x_i'y_i}{\sum_{i=1}^n x_i'x_i} \\ \Rightarrow \hat{\beta} = (\sum_{i=1}^n x_i'x_i)^{-1}(\sum_{i=1}^n x_i'y_i) \end{aligned}$$

At least as many moment condition equations are required as many parameters to be estimated by this approach. Otherwise the solution is not obtainable[3]. Lack of this condition is called under-identification. The method-of-moments only works if the number of moment conditions is equal to number of parameters of interest. This condition is known as exact-identified condition, which yields unique solution.

If there are more moment conditions than estimable parameters, then the system of equation becomes over-identified and the method of moment can not solve such system. The cause of over-identification is basically attributed to sampling error, because in the method of moments, we assume that in the population level, all the conditions will hold perfectly. Of course, it is possible to obtain estimated simply by throwing away the additional equations, but then the estimates will not satisfy all the thrown away conditions. Therefore, we need to use other methods for estimation, such as generalized method of moments.

[3]However, if instrumental variables are introduced in the system, then a solution might be obtainable with desirable properties. Then the sample moment condition will be $\big(\sum_{i=1}^n z_i'(y_i - x_i\hat{\beta})\big)/n = 0$, with z_i being the instrument. The idea is analogous to what was discussed in §2.2.

2.3.2 Details of Generalized Method of Moments

The generalized method of moments has long been an important tool for estimating various econometric and finance models. Generalized method of moments estimators obtains the estimates that minimizes a quadratic form of sample moment condition so that it can get as close as possible to solving the over-identified system of sample moment conditions. However, when the system is exact-identified, then the generalized method of moments estimate reduces to method of moments estimate.

Procedure: To explain how it works, let us suppose that there are q population moment conditions, that is, set of moment functions (does not have to be linear) with expectations zero as follows:

$$E\big(m(y_i, \theta)\big) = 0. \tag{5}$$

Here, m is $q \times 1$ vector of functions (say, $y_i - \theta$) whose expected values are zero in the population, y_i is the observation from subject i and θ is a $k \times 1$ vector of parameters. If the system is over-identified, we will have $k < q$, or if the system is exact-identified, we will have $k = q$. Notice that, if we knew the distribution, we could solve the expectation and get an estimate of θ. Since we do not know the distribution, we replace it with sample averages to obtain the analogous sample moments.

The sample moments corresponding to the population moments will be

$$\bar{m}_n(\hat{\theta}) = \big(\sum_{i=1}^{n} m(y_i, \theta)\big)/n.$$

From the exact-identified system ($q = k$), generalized method of moment will reduce to method of moment, and $\bar{m}_n(\hat{\theta})$ will be obtained directly from the equation $\bar{m}_n(\hat{\theta}) = 0$[4]. However, from the over-identified system ($q > k$), there is no solution to $\bar{m}_n(\hat{\theta}) = 0$ in general. The generalized method of moment will choose the estimate such that

$$\begin{aligned} \hat{\theta}_{GMM} &= \arg\min_{\theta} \; Q \\ &= \arg\min_{\theta} \; \bar{m}_n(\hat{\theta})' W \bar{m}_n(\hat{\theta}). \end{aligned} \tag{6}$$

The term $Q = \bar{m}_n(\hat{\theta})' W \bar{m}_n(\hat{\theta})$ measures the distance from the moment condition to zero, W being the weight matrix (explained later).

[4]Considering the distance function, equivalent solution can be obtained by solving $\bar{m}_n(\hat{\theta})' \bar{m}_n(\hat{\theta}) = 0$.

Why It Works: Let us explain why this would be a reasonable estimate. According to the law of large numbers, under the weak conditions,

$$\begin{aligned} \big(\sum_{i=1}^{n} m(y_i,\theta)\big)/n &\rightarrow E\big(m(y_i,\theta)\big) \qquad \text{for } n\rightarrow\infty \\ \Rightarrow \bar{m}_n(\hat{\theta}) &\rightarrow \bar{m}(\theta) \qquad \text{for } n\rightarrow\infty. \end{aligned}$$

Assuming this equality under large sample size, if we can obtain an estimate $\hat{\theta}$ as a solution from the sample moment condition equation $m_n(\hat{\theta}) = 0$, then that should be a reasonably good estimate for parameter θ. But for over-identified case, such solution is not obtainable. Generalized method of moment then searches for an estimate $\hat{\theta}$ in the parameter space, that minimizes the distance between the vector $m_n(\hat{\theta})$ and 0, which is obtained by minimizing quadratic distance function Q in equation (6), and finds a solution as close as possible to zero. Nonetheless, $\hat{\theta}$ has to be an unique solution to equation (5), for the estimate to be consistent. For that, some other boundary conditions on higher moments $m(y_i,\theta)$ has to satisfy as well.

The Weight Matrix: W in the equation is a symmetric positive definite weight matrix. The generalized method of moment estimator depends on this weight matrix. To allow more flexibility, W expands the Euclidean distance to more general distance.

A simple illustration of this weight matrix using two moment condition will be: $\bar{m}_n(\hat{\theta}) = [m_1 \;\; m_2]'$. Therefore, putting the simplest weight $W = I$, i.e., the indentity matrix will result in

$$\bar{m}_n(\hat{\theta})'W\bar{m}_n(\hat{\theta}) = \bar{m}_n(\hat{\theta})'I\bar{m}_n(\hat{\theta}) = \begin{pmatrix} m_1 & m_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = m_1^2 + m_2^2,$$

which is the square of the Euclidean distance from $\bar{m}_n(\hat{\theta})$ to 0. Similarly, by changing W, the coordinates can be manipulated. Also, since sample moments are not usually independent from each other, using a diagonal weighting matrix W may not be appropriate. The optimal W, which produces efficient estimators, is $W = \big(\lim\; n\rightarrow\infty\; var\{\sqrt{n}\bar{m}_n(\hat{\theta})\}\big)^{-1}$. However, for practical reasons, some other choices of W are often used as well (see §3.2.2).

Solution Approaches: The solution comes analytically or by numerical optimization, depending on the approach chosen:

I. Two-step Hansen [1982] introduced this method, where we need to go through two steps to eventually obtain a solution out of the equation (6). As an initial value of

W in the simplest case, we set $W = I$ and get

$$\hat{\theta}_{GMM_1} = \arg\min_{\theta} \ \bar{m}_n(\hat{\theta})' \bar{m}_n(\hat{\theta}).$$

Using this $\hat{\theta}_{GMM_1}$, we can obtain an estimate of $cov(\bar{m}_n(\hat{\theta}))$ and set $\hat{W} = \{cov(\bar{m}_n(\hat{\theta}))\}^{-1}$, and solve the next equation:

$$\hat{\theta}_{GMM_2} = \arg\min_{\theta} \ \bar{m}_n(\hat{\theta})' \hat{W} \bar{m}_n(\hat{\theta}).$$

Various researches also suggest its poor performances, especially in small sample cases.

II. Iterative To overcome the limitations of two-step procedure, Hansen et al. [1996] suggested this approach. Here, from the estimator $\hat{\theta}_{GMM_2}$, we can keep on updating $\hat{W}$ and then $\hat{\theta}_{GMM_3}$ and repeat the process through updating $\hat{W}$ until convergence is achieved for a given tolerance level. The smaller the tolerance level, more precise the estimate becomes. This approach has the advantage of being independent of the initial W. However, asymptotically, not much difference is achieved.

III. Continuous Updating Another approach, also suggested by Hansen et al. [1996], is simultaneously updating $\hat{W}$ and $\hat{\theta}$, by recognizing the dependence of W on θ:

$$\hat{\theta}_{GMM} = \arg\min_{\theta} \ \bar{m}_n(\hat{\theta})' \left(\overline{\{m\, m'\}}_n(\theta)\right)^{-1} \bar{m}_n(\hat{\theta}).$$

The solution has to come from numerical optimization for this approach, since the method is highly non-linear. The obvious problem is to find a starting value which is not too far from the minimum. However, no initial estimate of W is required for this method.

IV. Generalized Empirical Likelihood This approach was suggested by Smith [1997], which is not very popular in the literature yet.

Since the generalized method of moments estimator converges to normality as sample size is large, i.e., $\sqrt{n}(\hat{\theta}_{GMM} - \theta) \to N(0, \sigma)$ in distribution, one can perform inferential procedures under the assumption that $\hat{\theta}_{GMM} \sim N(\theta, \hat{\sigma}/n)$. Usually J test is performed to check whether the moment condition holds or not.

Although the estimates from generalized method of moments are more flexible and robust than the usual methods, their efficient depends on the choice of moment conditions. Also, as mentioned before, use of weak instrument has even more detrimental effect on efficiency.

3 Dichotomous Outcomes

Dichotomous outcome variables are of particular interest in the field of epidemiology. Measurement error and confounding are major issue in causal inference in context of epidemiological researches. Confounding due to measured covariates can be adjusted with various design or analysis tools, but confounding due to unmeasured covariates needs to be handled with more sophisticated tools. This is especially true for observational studies, which is a very common scenario for epidemiological research. Econometricians have used instrumental variable methods to deal with similar problems of estimating treatment effects in presence of unmeasured confounder, and therefore, use of instrumental variable method is gaining popularity among epidemiologists as well.

The instrumental variables methods can be used for controlling unmeasured confounding, especially from non-randomized data. Note that, the instrumental variables are not the same as proxy variables, but we can purify the proxy variables T by the use of instruments $\mathcal{Z}$ under certain conditions. For using a variable $\mathcal{Z}$ as an instrument of a proxy variable T, both instrument $\mathcal{Z}$ and proxy variable T have to be correlated, but instrument variable $\mathcal{Z}$ has to be uncorrelated with the measurement errors in proxy variable T. However, one possibility can be to obtain a second measurement of the proxy variable, which might then satisfy the conditions, but describing that method or logic is beyond the scope of this report.

3.1 Instruments in Epidemiological Setting

3.1.1 Definition

Just like in our previous definition, if we assume the treatment is denoted by $\mathcal{T}$, measured confounder by $\mathcal{C}$, unmeasured confounder by $\mathcal{U}$ (where $\epsilon \equiv f(\mathcal{U})$, to mach up with the previous definition), outcome by $\mathcal{Y}$; then, a variable $\mathcal{Z}$ is called an instrument, if

- treatment $\mathcal{T}$ and variable $\mathcal{Z}$ are correlated (stronger correlation will indicate stronger instrument),
- but itself unassociated with the unmeasured confounder $\mathcal{U}$,
- also unassociated with the outcome $\mathcal{Y}$, except via the effect of treatment $\mathcal{T}$, after controlling for confounders $\mathcal{C}$ and $\mathcal{U}$. That is, if we condition on $\mathcal{T}$, $\mathcal{C}$ and $\mathcal{U}$, outcome $\mathcal{Y}$ and variable $\mathcal{Z}$ will be unassociated.

In the presence of confounding, it is assumed that above assumptions holds at least with in each strata of the confounder $\mathcal{C}$.

3.1.2 Practical Issues

Just like our original problem, identifying an instrument is not as easy as it sounds. One major reason is that the required assumptions can not be tested empirically. Still, researcher from the specific problem might have a hunch about which variable could be the instrument. Therefore, obtaining a strong instrument might not be a possibility in all situations. However, once we identify an instrument, it is straightforward to determine whether it is strong or weak, by regressing and checking R^2. Also, use of multiple instruments is not very common in epidemiological researches due to the problem of identifying them properly. Moreover, in presence of hight confounding, the possibility of finding a strong instrument will be less, that is, in a regression setting, if treatment is a function of confounder and instrument, then more contribution of confounder will usually lead to a small contribution for the instrument.

3.2 Instrumental Variable Methods in Epidemiology

3.2.1 Pseudo-Two-Stage Logistic

Two-stage least squares method is an elementary method of analyzing instrumental variables. However, epidemiologist usually deals with binary outcome data or count data - and modelling these kind of data requires non-linear or generalized linear models. But the instrumental variable methods were originally developed for linear models. Therefore, for analyzing dichotomous or count treatments and outcomes, these instrumental variable methods need to be modified, so that generalized linear models such as logit or probit[5] models can be suitably used for the purpose. Pseudo-Two-stage method is one of those, where linear regression models are replaced by either logistic or probit models treating them just the same way.

However, there are several theoretical impediments in context of analyzing dichotomous outcome and exposure variables using pseudo-Two-stage least squares method. Model misspecification, specially formulating the first stage model with all the required confounders is a major issue in this setting. Otherwise, the estimated from the second stage model will not be unbiased. Inconsistency of the estimates and boundary problems[6] are consequences of this. The estimates of standard errors are often not reliable too.

[5] Although Probit model coefficient of treatment is not directly interpretable as a logarithms of odds ratio, logistic model coefficient can be approximately achieved by multiplying probit model coefficient with 1.6.

[6] Probit models deals with probabilities, and can be used to solve such issue of predicted values going beyond 0-1 range.

3.2.2 Generalized Method of Moments

To construct estimator through generalized method of moments, we need to make assumption about moment functions, that results in zero after taking expectation in a population level. Usually, the moments of the disturbance term ϵ or the unmeasured confounders can be used in this purpose, since we make several assumptions about them. For example, for generalized regression models, we have -

$$\begin{aligned} \mathcal{Y} &= \mu(\mathcal{T}, \mathcal{C}) + \epsilon \\ &= \left(1 + \exp\left\{-(\alpha + \beta\mathcal{T} + \gamma\mathcal{C})\right\}\right)^{-1} + \epsilon \qquad \text{for logistic} \\ &= \Phi(\alpha + \beta\mathcal{T} + \gamma\mathcal{C}) + \epsilon \qquad\qquad\quad \text{for probit,} \end{aligned} \tag{7}$$

where Φ is the cumulative distribution function of the standard normal distribution. In the above equation, $\mu()$ is a non-linear function to model the dichotomous outcomes, as shown in the later equations. This forces us to use the numerical optimization techniques.

Notice that, we cannot use the assumption $E(\mathcal{T}\,\epsilon) = 0$ due to existence of unmeasured confounders, and hence instrument (i.e., $\mathcal{Z}$) has to be used to solve this problem with some specific assumptions about the disturbance term ϵ, such as,

- The mean of the disturbance term is zero, i.e., $E(\epsilon) = 0$
- The instrument and disturbance term are uncorrelated, i.e., $E(\mathcal{Z}\,\epsilon) = 0$
- The measured confounder and disturbance term are uncorrelated, i.e., $E(\mathcal{C}\,\epsilon) = 0$.

Replacing the population moment conditions with sample moment conditions, one can estimate β using generalized method of moments, leaving other parameters as nuisance. When we have more moment conditions than the number of parameters to be estimated, generalized method of moments have to be used. Then, as usual, equation (6) is used to obtain a solution.

For most of the solution approaches, an initial W is required, which is usually unknown. For convenience, some uses either $W = I$ (two-step approach) or $\hat{W} = \sum_{i=1}^{n} \hat{\mathcal{Z}}_i \hat{\mathcal{Z}}_i' \hat{\epsilon}^2$ (this can be obtained from the sample), as suggested in epidemiologic literature. But none of these offers efficient estimates. Use of continuous updating approach may be one way to bypass this dilemma. Nonetheless, this approach requires reasonable initial values of the parameters.

4 Simulation Study

From previous research with continuous data, we know that the strength of association between instrument $\mathcal{Z}$ and the treatment $\mathcal{T}$ plays an important role in getting various quality of prediction. Same is true for the association between measured confounder $\mathcal{C}$ and the treatment $\mathcal{T}$. A simulation study with various levels of association can be useful to show whether that is also true for dichotomous outcomes or not.

4.1 The Simulation Design

The design is similar to Johnston et al. [2008] simulation. The key steps are as follows[7]:

Generate Variables At first, measured confounder $\mathcal{C}$, disturbance term ϵ and instrument $\mathcal{Z}$ is generated from independent standard normals $N(0, 1)$. The sample size is set to 1000, since the properties of generalized method of moments can be properly assessed under large sample conditions.

Set conditions to generate Treatment Choose α_1 and α_2 such that the association between T (defined by the equation $T = \alpha_1\mathcal{Z} + \alpha_2\mathcal{C} + \epsilon_1$) and $\mathcal{Z}$ belongs within a defined the range (so that the association can be grouped as weak or strong for instruments). Same rule is applied for fixing the association of T and $\mathcal{C}$ (high or low confounding is defined the same way). Once the data with specified correlation structure is obtained, T is redefined to $\mathcal{T}$ such that $\mathcal{T} = I_{T>0}$, to make it dichotomous.

Generate Outcome To generate the outcome, fist Y is generated from a poisson distribution, where the mean λ is defined as a function of treatment $\mathcal{T}$, measured confounder $\mathcal{C}$ and unmeasured confounder $\mathcal{U}$ (which is generated independently from standard normal $N(0, 1)$), and some small ϵ_2, combined in an additive structure with an intercept ν (to make sure λ is not negative). Here, $\beta = \log(3) = 1.1$, the parameter of interest for this study, with the relative risk associated with $\mathcal{T}$ being 3. Then to make the outcome dichotomous, Y is redefined to $\mathcal{Y}$ such that $\mathcal{Y} = I_{Y>\nu}$.

4.2 Methods in Use

With the generated data, three methods are implemented in the current report, using the same generated datasets[8] so that the results out of these three methods are comparable:

[7] R [R Development Core Team, 2010] implementation of this data generating process and all the analysis codes for 'logistic', 'two-stage logistic' and 'generalized method of moments with instrumental variables' can be obtained from the author of this report.

[8] The program is controlled by setting same seed values.

I. Logistic model with unmeasured covariate Usual logistic model is applied having treatment $\mathcal{T}$, measured confounders $\mathcal{C}$ and unmeasured confounders $\mathcal{U}$. Of course, in real life situation, unmeasured confounders are not available. This is used for obtaining gold-standard estimates so that the performance of the other methods can be evaluated.

II. Two-Stage Logistic with instrumental variable As mentioned before, two-stage least squares is theoretically sound for continuous outcomes. However, for epidemiological data, usually the generalized linear models with discrete outcome data are more appropriate. Therefore, some replaces the linear models by the logistic regression, which is named 'Pseudo-2SL' or simply '2SL' in this report. Researches showed that this method is inconsistent, but for practical use, the bias may be small, and the application of this method is straightforward. Therefore this method is frequently used in the mentioned context.

III. Generalized method of moment with instrumental variable For the generalized method of moments, logistic model was used, with parameters α (intercept), β (slope for treatment) and γ (slope for measured confounder). Here only β is the parameter of ultimate interest, making others nuisance parameters. For this, the used moment conditions are the ones described in §3.2.2. To elaborate, based on the orthogonality assumptions, the moment conditions can be re-written in the following form:

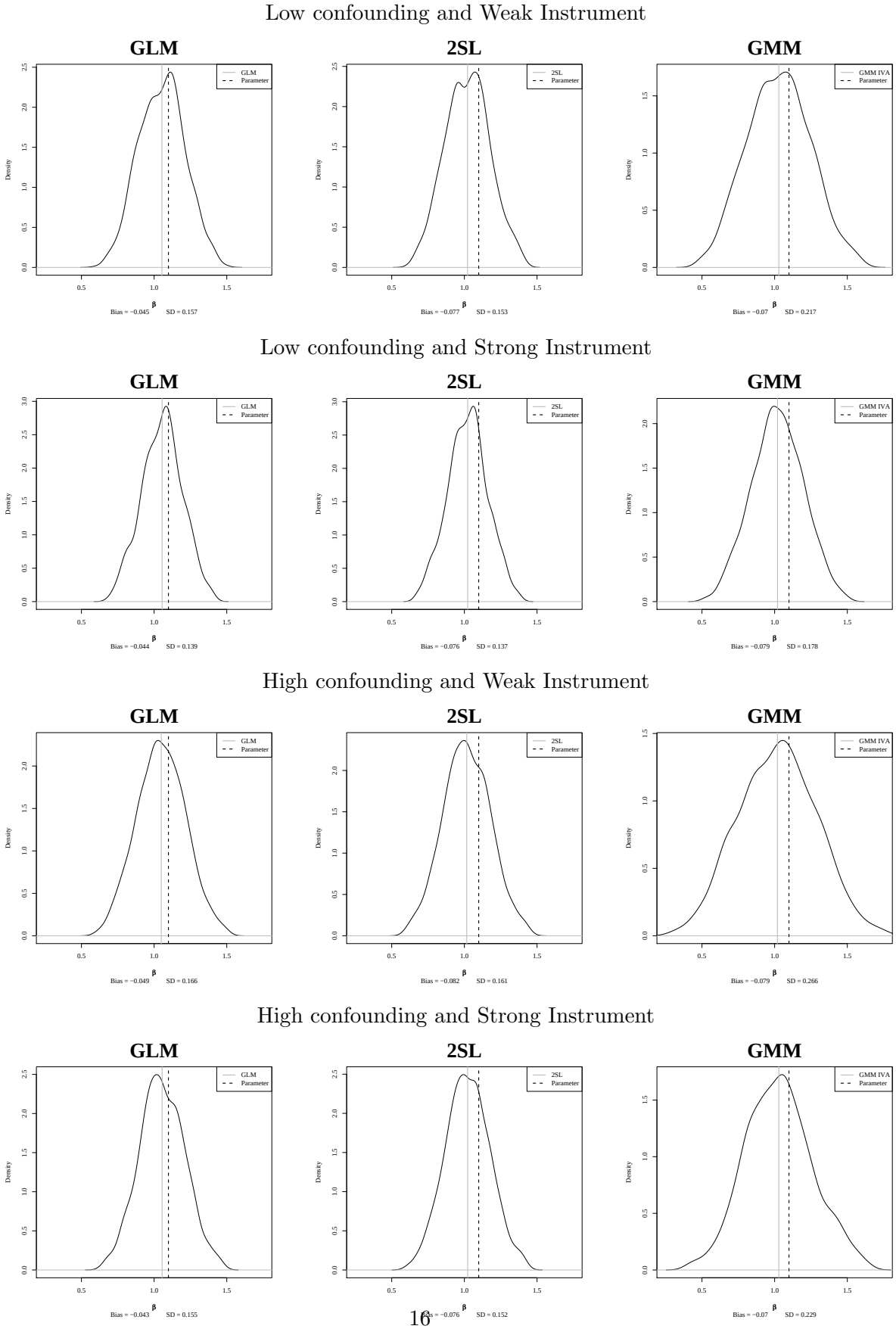
$$m(y_i, \theta) \quad = \quad \begin{pmatrix} \mathcal{Y} - \mu(\mathcal{T}, \mathcal{C}) \\ \mathcal{Z}(\mathcal{Y} - \mu(\mathcal{T}, \mathcal{C})) \\ \mathcal{C}(\mathcal{Y} - \mu(\mathcal{T}, \mathcal{C})) \end{pmatrix}$$

under the assumption that the model is correctly specified in $\mu(.)$ and $\mathcal{Y} - \mu(\mathcal{T}, \mathcal{C}) = \epsilon$ as mentioned in equation (7). Therefore, for estimated residual $\hat{\epsilon}_i = e_i$, the corresponding sample moment conditions would be:

$$\bar{m}_n(\hat{\theta}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\{e_i\} \\ \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathcal{Z}}_i e_i\} \\ \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathcal{C}}_i e_i\} \end{pmatrix} = 0.$$

Since the gradient of $\bar{m}_n(\hat{\theta})$ is in a complicated format, the 'continuous updating approach' (described in §2.3.2) for solving generalized method of moments was used. Hence providing an initial $\hat{W}$ was not necessary. Nelder and Mead optimization algorithm is used which works for even non-differentiable functions. For initial values, a logistic model was fitted with treatment $\mathcal{T}$ and confounder $\mathcal{C}$ as covariates.

Figure 2: Estimated treatment coefficients using GLM, Pseudo Logistic and GMM from specified levels of confounding and instrument strength.



4.3 Results and Discussion

The data generation was performed for specific requirement of correlation among T, $\mathcal{Z}$ and also T, $\mathcal{C}$. Once the data that satisfies the requirement is found, the respective methods were applied to obtain corresponding estimates. This was iterated until 1000 estimates were obtained from different datasets. The respective density plots of estimates are shown in Figure 2.

Since the logistic model (named as 'GLM' in Figure 2) includes both measured and unmeasured covariates, the estimates are reasonably well in all cases. The biases for the estimates of this method are the least. Also, the estimates are not affected by the level of instruments.

Generalized method of moments with instrumental variables (named as 'GMM' in Figure 2) depends on the level of instruments, since with stronger instrument, the estimates from this method are in general better in terms of variability. The variances are higher, especially in the case of weak instruments. In terms of bias or variability, they can not out perform the logistic model. Over all the variability is always greater than the other models.

On the other hand, for two-stage logistic model (named as '2SL' in Figure 2), variability is relatively low. However, the bias is almost always greater than any other methods.

4.4 Practical Problems and Solutions

Some of the outputs being extreme, made the analysis troublesome, especially because a lot of results had to be summarized from the simulation. Researchers have been dealing with this problem subjectively. In this analysis, a point is defined outlier if it is 1.5 times interquartile range (IQR) more than the third quartile Q_3, or if it is 1.5 times interquartile range less than the first quartile Q_1, which is a common practice in detecting outliers in statistical literature. Following this definition, all outliers were excluded from the analysis. This is why the shown density plots in Figure 2 are not skewed in any case. The reasons for generating such extreme results are explained next.

An initial value was required for initialization process. Fortunately, the log likelihood function for logistic regression is globally concave, hence we do not have to deal with problems associated with local maxima while using optimization algorithm. Still, we have some other numerical optimization problems while dealing with dichotomous outcomes. Altman et al. [2004] describes some of the numerical issues that occurs in the binary

regression situation. Some of them are:

- Non-convergence is a problem in general.
- When Hauck–Donner phenomenon occurs, that is, when fitted values are very close to 0 or 1, usually programming language rounds up after certain decimal places, and as a consequence, numerically 0 or 1 occurs for the fitted values, as mentioned in Venables and Ripley [2002].
- Theoretically there can be situations where the likelihood function does not have any maxima – hence no maximum likelihood estimates. It is known as the case of complete separation. A related problem is quasi–complete separation, which is also a possibility.

The estimates and standard errors under non–convergence or when Hauck–Donner phenomenon occurs are usually very different from the converged estimates. Hence such outputs were detected by checking each and every analysis results with the given identifying conditions and omitted from further analysis[9].

5 Bibliographic Notes

To understand the basic concepts of instrumental variables, and the associated methods of estimation, some econometric books were useful, such as Hall [2005], Hamilton [1994], Wooldridge [2002], Verbeek [2008], Damodar [1999] and Heij et al. [2004]. While implementing some of the instrumental variable methods, a few books and manuals were consulted, such as Fox [2002], Chausse [2010], Iacus [2008]. Among epidemiologic literature, Johnston et al. [2008], Martens et al. [2006], Greenland [2000] and Rassen et al. [2009] were very helpful.

References

Altman, M., Gill, J., and McDonald, M. (2004). *Numerical issues in statistical computing for the social scientist*. Wiley-IEEE.

Chausse, P. (2010). *gmm: Generalized Method of Moments and Generalized Empirical Likelihood*. R package version 1.3-0.

Damodar, G. (1999). *Basic econometrics*. McGraw-Hill Companies.

[9] Out of 1000 iterations, at most 14 samples were discarded in the current analysis based on the criteria mentioned above.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Sage Pubns.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Hall, A. (2005). *Generalized method of moments*. Oxford University Press, USA.

Hamilton, J. (1994). *Time series analysis*. Princeton Univ Pr.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, L., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Heij, C., de Boer, P., Franses, P., Kloek, T., and Van Dijk, H. (2004). *Econometric methods with applications in business and economics*. Oxford University Press, USA.

Iacus, S. (2008). *Simulation and inference for stochastic differential equations: with R examples*. Springer Verlag.

Johnston, K., Gustafson, P., Levy, A., and Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in medicine*, 27(9):1539–1556.

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. (2006). Instrumental variables: application and limitations. *Epidemiology*, 17(3):260.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rassen, J., Schneeweiss, S., Glynn, R., Mittleman, M., and Brookhart, M. (2009). Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American journal of epidemiology*, 169(3):273.

Smith, R. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *The Economic Journal*, 107(441):503–519.

Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Springer verlag.

Verbeek, M. (2008). *A guide to modern econometrics*. Wiley.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. The MIT press.